Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica per il Management

Big-data e Turismo: Analisi delle recensioni utente sulla piattaforma Expedia

Relatore:
Chiar.mo Prof.
MARCO DI FELICE

Presentata da:
PARIDE MARTINELLI

Correlatore:
Chiar.mo Prof.
MARCELLO MARIANI

Sessione III
Anno Accademico 2014/2015

Introduzione

Il progressivo aumento della dimensioni dei dataset, derivante dall'evoluzione tecnologica e dall'enorme volume e varietà di informazioni diffusa in rete attraverso i social media, ha determinato l'esigenza di definire una nuova tipologia di dati, i cosiddetti big data. Questo termine viene utilizzato per descrivere una mole di dati così estesa in termine di volume, velocità e varietà da richiedere particolari tecnologie per il loro immagazzinamento e per la loro gestione. L'aumento della dimensione dei dataset ha portato, inoltre ad un nuovo obiettivo, ovvero quello di estrarre informazioni aggiuntive rispetto a quelle che si potrebbero ottenere analizzando piccole serie di dati.

I tradizionali DBMS, come ad esempio gli RDBMS, ovvero i database relazionali, non sono in grado si contenere una quantità di dati così estesa, in termini di volume; oltre che per il volume i database relazionali non riuscirebbero a gestire i big data per la loro rigidità di schema, infatti altra caratteristica di questa nuova tipologia di dati è la loro varietà, intesa anche come assenza di uno schema fisso. Nascono quindi i database NoSQL, acronimo Not Only SQL, appunto a significare che esistono diverse situazioni per le quali il modello relazionale risulta inappropriato, ma tante altre per le quali tale modello è ancora la soluzione migliore.

A proposito del nuovo obiettivo associato ai big data, ovvero quello di estrarre informazioni aggiuntive per ottenere risultati più precisi, di un livello più elevato e di maggior interesse, ad esempio per il busiess, molto si parla

del tema: big data e turismo. Questo grande tema ha l'obiettivo principale di analizzare dati ricavati dalle piattaforme online di tipo booking e studiare il loro rapporto con il turismo. Molti articoli infatti sono stati scritti per descrivere questo tipo di analisi, alcuni concentrati sulle recensioni in generale, altri più focalizzati sulle valutazioni, e altri ancora sulle risposte date dagli albergatori alle recensioni che ricevono.
Questi studi, analizzando le esperienze descritte dai turisti in rete, cercano anche di anticipare trend e ricavare informazioni con cui indirizzare il settore turistico e l'industria alberghiera.

Big data e turismo è appunto il tema principale di questa tesi: esamina l'utilizzo dei social media come mezzo di interazione tra turisti e albergatori, prendendo come riferimento la piattaforma di booking Expedia.com e come campione di hotel tutti gli hotel della penisola italiana. In particolare si vogliono studiare i due tipi di utenti che utilizzano questo tipo di piattaforma, ovvero turisti e albergatori. Dei turisti si vuole capire il loro livello di attività su social media, calcolando l'andamento delle recensioni postate; poi si passa ad uno studio della singola recensione e infine si è cercato di stabilire le preferenze dei turisti in base alle valutazioni. Anche per quanto riguarda l'altra tipologia di utenti, ovvero gli albergatori, si è voluto capire quanto utilizzino Expedia, studiando le risposte alle recensioni.Infine, tramite un confronto tra la distribuzione delle recensioni suddivise per lingua tra le varie regioni italiane, e i dati ENIT del turismo, si è voluto rilevare se vi è una correlazione tra recensioni e densità di turismo.

Prima di fare questo tipo di analisi, e di trarre delle conclusioni, è stato redatto un primo capitolo che parla dei big data, cosa sono, come nascono, le loro caratteristiche e il loro utilizzo; viene anche menzionato un articolo, scritto da Accenture che testimonia il grande successo dei big data. Dopo aver parlato di questa tipologia di dati, vengono descritti i DBMS, prima in generale e successivamente più in particolare su una tipologia di DBMS,

ovvero quelli non relazionali, chiamati NoSQL. Il primo capitolo si conclude con la descrizione dei uno dei più famosi database NoSQL, quale MongoDB.

Il secondo capitolo invece descrive lo stato dell'arte sul tema big data e turismo, infatti vengono discussi vari articoli, già scritti sul tema in due importanti riviste del settore, ovvero la *Cornell Hospitality Quarterly* e l'*International Journal of Hospitality Management*, mettendoli in relazione tra di loro e confrontando i risultati ottenuti dalle loro analisi.

Un terzo capitolo descrive le varie fasi di progettazione e di implementazione dell'applicazione di estrazione ed analisi dei dati. Vengono quindi descritte le specifiche del progetto, vengono elencate le tecnologie utilizzate per l'estrazione, il salvataggio e la gestione dei dati e viene descritto il database realizzato. Si parlerà anche delle APIs Expedia studiate per un corretto utilizzo delle query di estrazione dei dati relativi a hotel e recensioni.

Il quarto ed ultimo capitolo descrive in dettaglio l'analisi svolta, sezionandola per tutte le sue varie fasi; di ogni fase viene spiegata la metodologia di estrazione, vengono elencati i dati ottenuti, i quali verranno poi rappresentati graficamente, e infine verranno tratte specifiche conclusioni in merito ai risultati ottenuti.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Big Data, DBMS e NoSQL

In questo capitolo verranno introdotti i Big Data, in particolare verrà spiegato cosa sono e le loro caratteristiche principali; si parlerà dei *dbms* e ne verranno elencati i principali con le loro caratteristiche; e infine verranno trattati i database NoSQL, soffermandosi in modo particolare su *MongoDB*.

1.1 Introduzione ai Big Data

I big data sono dati che superano i limiti degli strumenti di database tradizionali. Il termine big data è poi utilizzato, per estensione, anche per definire le tecnologie volte a estrarre conoscenza e valore da questa tipologia di dati [1].

I big data, a partire dal 2012, stanno riscontrando grandissimo successo nel campo informatico, anche se grandi aziende, del calibro di Google, utilizzano tecnologie in grado di elaborare dati da diverso tempo, investendo su di esse moltissime risorse. Questo grande successo dei big data degli ultimi anni nel campo informatico è dovuto alla disponibilità di tecnologie open source che utilizzano hardware a prezzi contenuti e alla disponibilità di piattaforme cloud, entrambi fattori che concorrono decisamente all'abbattimento dei co-

sti .

Big data è infatti un termine usato per descrivere una mole di dati così estesa in termini di volume, velocità e varietà da richiedere tecnologie e metodi analitici, non semplici e immediati, per l'estrazione di un dato. Queste moli di dati si caratterizzano anche per il fatto di essere eterogenee e destrutturate, come immagini, email, dati GPS o informazioni prese dai Social Network, quindi anche per questo difficili da gestire attraverso le tecnologie tradizionali.

L'esigenza di definire questa tipologia di dati è nata dal progressivo aumento della dimensione dei dataset, derivante dall'evoluzione tecnologica. L'aumento della dimensione dei dataset ha portato, inoltre, ad un nuovo obiettivo, ovvero quello di estrarre informazioni aggiuntive rispetto a quelle che si potrebbero ottenere analizzando piccole serie di dati. Un esempio di utilizzo dei big data potrebbe essere quello dell'analisi dei dati del commercio e dei mercati per ottenere un trend complessivo della società tramite il flusso di informazioni che viaggiano e transitano attraverso internet; un altro esempio potrebbe essere l'analisi dei dati delle piattaforme online di tipo booking per ottenere informazioni sull'andamento del turismo, proprio come è stato fatto per la mia tesi.

Con i big data la mole dei dati è dell'ordine degli Zettabyte, ovvero di miliardi di Terabyte; per questo si richiede una potenza di calcolo parallelo e massivo apposita, eseguita su decine, centinaia migliaia di server.

1.1.1 Definizione di Big Data

Si parla di big data quando si ha un dataset talmente grande da richiedere strumenti non convenzionali per estrapolare, gestire, e processare informazioni entro un tempo ragionevole. Non esiste una dimensione di riferimento,

per definire la dimensione dei dati e il tempo di cui stiamo parlando, poiché questa cambia sempre, in relazione al progresso tecnologico, infatti la potenza delle macchine è in continuo aumento, così come la loro velocità e di conseguenza i dataset sono sempre più grandi.

Secondo uno studio, condotto nel 2001 dall'analista Doug Laney, venne definito il modello di crescita dei dataset come tridimensionale; questo modello venne chiamato modello delle 3"V". Il modello afferma che con il passare del tempo aumentano: volume e varietà dei dati, e la loro velocità di generazione; modello che tutt'ora viene usato per definire le principali caratteristiche dei big data. Infatti in molti casi questo modello è ancora valido, nonostante nel 2012 il modello sia stato esteso ad una quarta variabile, la veridicità [2].

1.1.2 Caratteristiche principali

I big data sono disponibili in enormi volumi, si presentano con formati destrutturati e caratteristiche eterogenee e, spesso, sono prodotti con estrema velocità. Volume, varietà e velocità (volume, variety, velocity) sono dunque i fattori che li identificano.

Volume: uno degli aspetti che caratterizzano i big data, come suggerisce il nome, è la loro quantità. Questa grandissima quantità di dati viene generata ad esempio dall'utente attraverso l'utilizzo di piattaforme del Web 2.0 [1], oppure automaticamente da macchine industriali o da transazioni bancarie

[1]Il termine Web 2.0, apparso nel 2005, indica genericamente la seconda fase di sviluppo e diffusione di Internet, caratterizzata da un forte incremento dell'interazione tra sito e utente: maggiore partecipazione dei fruitori che spesso diventano anche autori (blog, chat, wiki); più efficiente condivisione delle informazioni, che possono essere più facilmente recuperate e scambiate (YouTube); affermazione dei social network (Facebook). http://www.treccani.it/enciclopedia/web-2-0/

[1].

L'ampio volume di dati che è possibile raccogliere al giorno d'oggi, potrebbe apparentemente rappresentare un problema. In realtà, quello del volume dei big data, è un falso problema, in quanto cloud e virtualizzazione aiutano nella gestione del grosso volume di dati disponibili, semplificando il processo di raccolta, immagazzinamento e accesso ai dati.

L'IDC [2] stima che nel 2020 l'insieme di tutti i dati in formato digitale sarà pari a 40 zettabyte, circa 5,2 exabyte per ogni uomo, donna, bambino presente sulla terra [3].

Velocità: la velocità con cui i dati si rendono disponibili è il secondo fattore che identifica i big data. Questa caratteristica è un altro fattore, oltre al volume, che rende necessario l'utilizzo di strumenti in grado di tenerne il passo. Per le aziende la sfida è cercare di effettuare un'analisi dei dati in tempo reale, o quasi; infatti esse cercano di sfruttarli con altrettanta rapidità, attingendo da essi le informazioni utili per il business e minimizzando i tempi di elaborazione. Da questa esigenza, di ottenere una risposta di calcolo molto veloce, sono nati i database non relazionali [1].

Varietà: la diversità di formati e, spesso, l'assenza di una struttura che possa essere rappresentata attraverso una tabella in un database relazionale, sono la terza caratteristica dei big data.

La caratteristica di varietà puó essere associata alla tipologia di dato, che puó essere ad esempio TXT, CSV, PDF, Word, ma anche alla provenienza del dato, ovvero alle fonti diverse come ad esempio i social media, quali Facebook o Twitter, i DBMS operativi, o un qualsiasi sito web; questi dati vengono co-

[2]IDC, International Data Corporation,è un'azienda di ricerche di mercato, analisi e consulenza, specializzata nell'Information Technology [1].

munemente chiamati multi-sorgente, e quindi aventi strutture di diverso tipo.

L'eterogeneità di formati, fonti e strutture rende difficoltoso il processo di utilizzo dei big data con gli strumenti tradizionali. Per il salvataggio di dati semistrutturati, la scelta ricade spesso sui database NoSQL, database che forniscono i meccanismi adatti a organizzare i dati ma non impongono una rigidità nella struttura logica. [1].

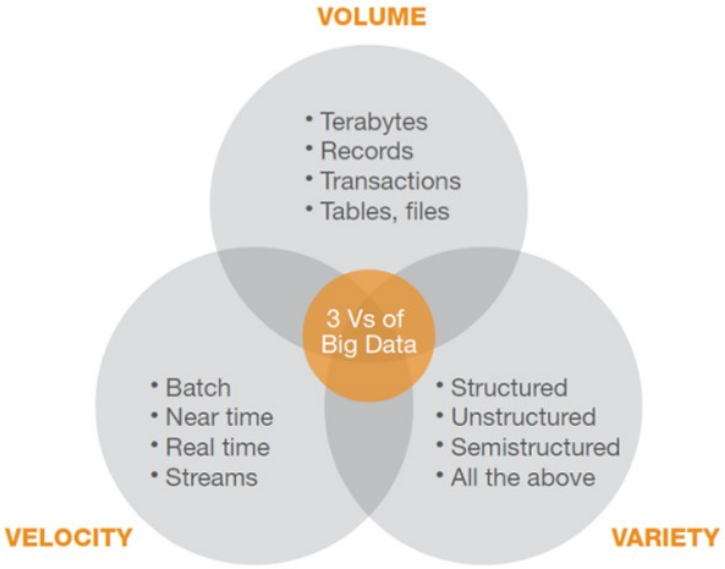

Figura 1.1: Le 3V dei big data

1.1.3 Opportunità e rischi dei Big Data

Le opportunità che i big data portano con sé possono essere viste sotto tre aspetti:

1. Il business: i big data danno la possibilità di perseguire nuovi modelli di business o di ottenere sensibili vantaggi competitivi sul business tradizionale dell'azienda.

2. La tecnologia: la dimensione e la complessità dei dati richiedono tecnologie adeguate, al fine di trarre valore dai big data.

3. L'aspetto finanziario: diversi casi di utilizzo dei big data dimostrano che essi portano ad indubbi vantaggi economici alle aziende che hanno adottato soluzioni di questo tipo; bisogna peró tenere anche conto degli ingenti costi che occorre sostenere per implementare un progetto che utilizza i big data.

I big data, ovviamente, non racchiudono in se soltanto caratteristiche positive e grandi opportunità: essi infatti presentano alcuni aspetti critici che potrebbero vanificare i vantaggi, come ad esempio la qualità e l'affidabilità dei dati, le questioni legate alla privacy e alla proprietà dei dati.

Qualità e affidabilità: la qualità è determinata da un insieme di caratteristiche che devono essere rispettate per un corretto utilizzo dei big data. Le caratteristiche sono le seguenti:

- *Completezza*: presenza di tutti i dati necessari per descrivere un'entità, una transazione o un evento;
- *Consistenza*: assenza di contraddizione nei dati;
- *Accuratezza*: conformità ai valori reali, cioè correttezza dei valori;
- *Assenza di duplicazione*: campi, record o tabelle devono essere presenti soltanto una volta, sia nello stesso sistema che in sistemi diversi;
- *Integrità*: caratteristica usata in riferimento ai database relazionali. Essi infatti garantiscono che i dati rispettino alcuni vincoli.

Anche se queste caratteristiche, vengono almeno in parte rispettate, spesso in azienda la qualità complessiva non è elevata a causa di errori quali: errori nelle operazioni di data entry manuale; errori nei software di gestione dei dati; o errori di progettazione delle basi di dati. La qualità dei dati deve

essere controllata e verificata, e il processo di data entry deve indicare come output quali sono i dati con livelli di integrità, completezza, consistenza e accuratezza ritenuti accettabili e quali invece sono da migliorare.

Possiamo, inoltre, distinguere tre tipi di dati nei confronti dei quali possiamo riscontrare diverse problematiche sulla qualità:

- *Dati provenienti da sistemi operazionali*: i sistemi operazionali possono essere ad esempio sistemi legati al mondo della finanza o della grande distribuzione. Il problema della qualità sorge nei casi in cui questi tipi di sistemi producano una vasta quantitù di dati; per fare fronte a questo problema esistono molti strumenti per il controllo e la pulizia dei dati.

- *Dati provenienti da sensori o strumenti scientifici*: come è ovvio questi dati generati automaticamente da macchine non sono soggetti ad errori di immissione, ma possono presentare problemi di qualità dovuti a difetti nei sensori o negli strumenti di misura.

- *Dati provenienti dal Web*: nel caso dei dati provenienti dai Social Network, essi si presentano in un formato semistrutturato: i metadati sono più affidabili, invece il testo è spesso soggetto ad errori e imprecisioni, ne sono un esempio lampante i commenti, i tweet o i post contenti errori di battitura, errori grammaticali, ma anche abbreviazioni e modi di dire.

Un'altra questione importante riguarda la *caratterizzazione delle informazioni*: non sempre è possibile distinguere significati diversi di una stessa parola o sigla, come ad esempio la parola "cucina" può riferirsi sia all'arte culinaria, sia all'insieme di mobili ed elettrodomestici. La sfida che i big data pongono è dunque legata alla rilevanza e all'attinenza che essi hanno rispetto allo scopo dell'analisi.

Infine le fonti del web potrebbero presentare il *problema della veridicità*: non sempre le notizie o i documenti contengono affermazioni e dati veritieri.

Privacy e proprietà dei dati: queste problematiche, e quindi di conseguenza anche il problema della possibilità di utilizzo da parte i terzi, riguardano sia alcune tipologie di dati, sia le informazioni che è possibile estrarre attraverso l'analisi. Un esempio potrebbe essere appunto il Web: il fatto che sul web circolino molti dati e che siano accessibili a tutti, non significa che sia etico utilizzarli; infatti dai social network è possibile estrarre dati sensibili, quali orientamento politico e credo religioso degli utenti che potrebbero essere usati in modo inappropriato e discriminatorio; dalle banche dati di aziende ospedaliere è possibile estrarre dati relativi alla salute, se essi non sono adeguatamente protetti; e infine, ormai, è praticamente impossibile non lasciare le cosiddette tracce elettroniche dei propri spostamenti, infatti telefoni, smartphone e sistemi elettronici di pagamento sono alcuni esempi di come gli spostamenti di una persona possano essere monitorati.

1.1.4 Testimonianza di successo dei Big Data

Le aziende che già utilizzano i big data sono enormemente soddisfatte dei propri risultati di business, e affrontano sfide per mantenersi competitive e diventare imprese digitali, come risulta da uno studio effettuato da Accenture [3].
Infatti il sondaggio *Big Success with Big Data* di Accenture Analytics dimostra che i big data stanno decollando; secondo questo sondaggio il 92% di coloro che hanno portato a termine progetti basati sui Big Data è soddisfatto dei risultati di business ottenuti, e il 94% riferiscono che l'applicazione soddisfa le loro esigenze. Sempre dallo stesso sondaggio risulta che per l'89% degli intervistati i big data sono molto importanti per la transizione dell'or-

[3] Accenture è una multinazionale di consulenza aziendale, servizi tecnologici e outsourcing. Accenture è attualmente la società di consulenza aziendale più grande al mondo. https://it.wikipedia.org/wiki/Accenture

ganizzazione verso il digitale.
Big Success with Big Data si è anche occupata di stabilire in quali attività vengono struttati i big data: 54% per l'identificazione di nuove fonti di reddito; sviluppo di nuovi prodotti o servizi (50%).
Tuttavia, per molti, l'implementazione dei big data è anche connessa a sfide impegnative in termini di sicurezza, budget, talento e integrazione tecnologica [4].

Il sondaggio Big Success with Big Data condotto da Accenture Analytics mette in luce molte informazioni importanti, in particolare:

C'è molto da imparare quando si avviano iniziative e progetti con l'utilizzo dei big data, in particolare per quanto riguarda le matrici di dati e tecniche analitiche. Infatti le implementazioni di big data sono impegnative, ma non impossibili, esse costringono gli utenti a rimanere flessibili e quindi capaci di adattarsi e imparare man mano che crescono.

Le aziende più grandi ottengono risultati migliori dai big data, questo perché le grandi aziende hanno una concezione più ampia del significato del termine big data, e utilizzano un numero maggiore di tipi e di fonti di dati per una gamma più vasta di obiettivi e un ventaglio più esteso di funzioni. Le organizzazioni di maggiori dimensioni cominciano con iniziative focalizzate in ambiti concreti, quali relazioni con la clientela, sviluppo prodotti e attività operative, anziché cercare di fare tutto contemporaneamente.

Acquisire talenti nel campo degli analytics non è facile. Competenze e talenti sono un problema per molti e continuano a scarseggiare. Le aziende di maggior successo si procurano talenti ovunque riescano a trovarli, affidandosi pesantemente a risorse esterne ed esperte, quali consulenti, dipendenti a contratto e risorse di vendor di tecnologia. Le organizzazioni stanno anche cercando soluzioni per assumere e sviluppare talenti al proprio interno.

I Big Data hanno un potenziale di trasformazione. L'89% delle aziende che utilizzano i Big Data sono convinte che essi rivoluzioneranno l'operatività dell'azienda esattamente come fece Internet e ritengono che siano molto importanti per la trasformazione digitale della propria azienda. Alla domanda su quale sia l'ambito in cui prevedono che i Big Data avranno l'impatto maggiore in azienda nei prossimi cinque anni, i dirigenti intervistati hanno indicato: relazioni con la clientela (63 %); sviluppo prodotti (58 %); e attività operative (56 %). L'opinione diffusa è chiara: i Big Data portano con sé una trasformazione dirompente, anche se le aziende non sempre concordano su cosa sia compreso nei "Big Data".

1.2 I DBMS

DBMS è la sigla di data base management system, che tradotto vuol dire sistema di gestione di basi di dati. Le principali funzioni del DBMS sono quelle di garantire il mantenimento della corretta strutturazione dei dati nei diversi database gestiti e di facilitare l'accesso delle applicazioni ai dati, tramite opportune istruzioni impartite dal sistema operativo. A queste funzionalità di base si aggiungono quelle di interrogazione, le cosiddette *query* , e di modifica del database.

A seconda del modello di organizzazione dei dati sul quale questi sistemi si basano si avranno DBMS relazionali, gerarchici e così via. Lo standard che si è rivelato vincente tra gli utenti e che caratterizza la maggior parte dei DBMS oggi esistenti è quello del DBMS relazionale (anche abbreviato come RDBMS) che utilizza SQL (*structured query language*) come linguaggio di interrogazione dei dati. L'obiettivo per il quale i DBMS si sono originariamente affermati e diffusi è quello di fornire all'utente una interfaccia opportuna per gestire in modo "astratto" i dati, svincolandoli in tal modo dalla loro collocazione fisica e permettendo di agire in modo relativamente facile sugli stessi,

garantendo al contempo che la struttura sottostante rimanesse fisicamente integra.

Negli anni più recenti, la diffusione crescente dei sistemi informativi nelle organizzazioni, la complessità dell'insieme di utenti ai quali si indirizzano le applicazioni e la sempre più ingente mole di dati che i database gestiscono abitualmente hanno richiesto lo sviluppo di quelle funzionalità dei DBMS che abilitano l'efficienza e l'efficacia delle applicazioni che li utilizzano; in particolare la gestione degli accessi concorrenti e il controllo delle transazioni, cioè delle operazioni di creazione, modifica e cancellazione dei dati [5].

In fine, si puó affermare, che i DBMS in generale svolgono un ruolo fondamentale in numerose applicazioni informatiche, dalla contabilità, alla gestione delle risorse umane e alla finanza, fino a contesti tecnici come la gestione di rete o la telefonia [6].

1.2.1 Caratteristiche principali dei DBMS

Un DBMS è un sistema software che è in grado di gestire collezioni di dati di grandi dimensioni, condivisi e persistenti, in maniera efficace e sicura.

Le principali funzionalità sono:

- Creazione di una base di dati e memorizzazione di essa su una memoria secondaria;
- Possibilità di accesso di lettura e scrittura dei dati in qualsiasi momento, da parte del creatore;
- Possibilità di condivisione dei dati tra diversi utenti o tra diverse applicazioni;

- Possibilità di implementare un paradigma di separazione di dati e applicazioni, infatti le applicazioni non necessitano di conoscere la struttura fisica dei dati (come devono essere memorizzati su disco) ma solo la struttura logica (cosa rappresentano).

Componenti di un DBMS, suddivisi per caratteristiche li contraddistinguono:

Efficienza nella gestione dei dati: i DBMS forniscono adeguate strutture dati per organizzare i dati all'interno dei file, e per supportare le operazioni di ricerca/aggiornamento. Le strutture dati di cui parliamo di solito sono strutture ad albero o tabelle hash[4]. L'*indice* è quel componente che contiene le informazioni sulla posizione di memorizzazione delle tuple sulla base del valore del campo chiave; permette quindi un accesso diretto più performante alla risorsa.

Concorrenza: in molti sistemi è fondamentale gestire operazioni concorrenti di accesso ai dati, come ad esempio PayPal ha un processing di oltre 7.7 milioni di pagamenti al giorno. Per fare fronte a questo problema la maggior parte dei DBMS forniscono un livello di locking molto più elevato rispetto a quello convenzionale; allo stesso tempo, un DBMS deve garantire che non ci siano interferenze tra accessi provenienti da diverse applicazioni. Il *Lock Manager* è quel componente responsabile di gestire i lock alle risorse del DB e di rispondere alle richieste delle transazioni; è quindi quell'elemento che consegna i permessi di lettura e di scrittura alle transazioni per le risorse condivise.

Affidabilità: alcune operazioni sui dati sono particolarmente delicate, e devono essere gestite in maniera opportuna, secondo la regola del tutto o niente, ad esempio durante un operazione di trasferimento di denaro non è

[4]In informatica una hash table, in italiano tabella hash è una struttura dati usata per mettere in corrispondenza una data chiave con un dato valore. Viene usata per l'implementazione di strutture dati astratte associative come Map o Set.

accettabile che il software si blocchi a metà della transizione, con il rischio di non trasferire l'intera somma di denaro. Per questo motivo i DBMS devono fornire appositi strumenti per annullare operazioni non complete e fare rollback dello stato del sistema, ovvero tornare allo stato di partenza. In molti DBMS esistono quindi degli strumenti e degli algoritmi che garantiscono persistenza dei dati anche in presenza di malfunzionamenti, ne sono un esempio i *log*, nei quali vengono indicate tutte le operazioni svolte dal DBMS; tramite i log è quindi possibile fare do/undo delle operazioni.

Sicurezza: la maggior parte dei DBMS implementa politiche di controllo degli accessi ai dati mediante *sistemi* di permessi che permettono di identificare quali sono le operazioni consentite ad un determinato utente e quali sono i dati che appartengono ad un determinato utente.

1.2.2 Un po' di storia sui DBMS

Information Management System (IMS) è il nome di un Software sviluppato da IBM nel 1968 utilizzato come supporto alle missioni di Apollo [5] per la gestione dei dati tecnici e amministrativi e delle forniture dei materiali. Si trattava già di un modello gerarchico di gestione dei dati con un motore transazionale per la concorrenza.

In seguito nel 1970, un ricercatore della IBM, Edgar Codd [6], pubblica la sua visione di modello "relazionale" dei dati, basato sul concetto matematico di relazione tra insiemi. Negli stessi anni IBM lavora allo sviluppo di un linguaggio basato sul modello relazionale, quello che oggi chiamiamo SQL,

[5] Si parla di Apollo 7, che fu la prima missione con equipaggio nel programma di Apollo ad essere lanciata dopo il tragico incidente dell'Apollo 1. Fu una missione orbitale di 11 giorni e la prima missione spaziale americana con tre uomini. la concorrenza.
https://it.wikipedia.org/wiki/Apollo

[6] Edgar Frank "Ted" Codd, nato a Portland il 23 agosto 1923 e morto a Williams Island il 18 aprile 2003, è stato un informatico britannico, fondatore della teoria delle basi di dati relazionali.

e all'implementazione di un RDBMS sperimentale, ma contemporaneamente continua a lavorare sul vecchio IMS.

Qualche anno dopo, nel 1979, una piccola startup, chiamata Relational Software Inc, produce un primo esempio di RDBMS commerciale. Questa startup, in pochi anni divenne una vera e propria azienda, quella che oggi conosciamo come Oracle Corporation.

Negli anni '80 compaiono i primi DBMS basati sul modello ad oggetto, i cosiddetti ORDBMS, che cercano di emulare il successo del paradigma di programmazione ad oggetti e facilitare l'integrazione tra DBMS e i linguaggi ad alto livello, del calibro di C++ o di Java. Sempre in parallelo agli ORDBMS, viene anche sviluppato un vero e proprio linguaggio, utilizzato per questo modello di DBMS, un linguaggio chiamato OQL, che non è altro che l'omologo di SQL per il paradigma ad oggetti. Contrariamente a quanto si potesse pensare, data l'importanza dei linguaggi di programmazione orientata agli oggetti, questo tipo di DBMS è sempre stato poco diffuso.

Avvicinandosi agli anni 2000, abbiamo che, solo nel 2011, il mercato degli RDBMS ha avuto una crescita del 16.5% con ricavi complessivi pari a 24 miliardi di dollari. Fino ad oggi il mercato dei RDBMS è dominato da quattro vendor, che da soli occupano una percentuale di ricavi pari al 75% del totale. Questi quattro vendor sono: Oracle, con un incasso record nel 2011 di 10 bilioni di dollari, IBM, Microsoft e SAP.

Oggi, una delle nuove linee evolutive dei DBMS è rappresentata dall'approccio NoSQL. Questo nuovo approccio è dominato da un'idea di base, cioè quella di superare la rigidità del modello relazionale nella definizione dello schema, consentendo una più facile espansione del DB in termini di dati, e di computazione distribuita. Alcuni esempi di DBMS NoSQL sono ad esempio Apache Cassandra, Apache Couch, e MongoDB.

Un commento storico da sottolineare è il fatto che se in passato i DBMS erano diffusi principalmente presso le grandi aziende e istituzioni (che potevano permettersi l'impegno economico derivante dall'acquisto delle grandi infrastrutture hardware necessarie per realizzare un sistema di database efficiente), oggi il loro utilizzo è diffuso praticamente in ogni contesto.

Un altro fatto storico rilevante dei DBMS, è il loro utilizzo, il quale risale agli inizi della storia dell'informatica, anche se la grande maggioranza di questi erano programmi specializzati per l'accesso di un singolo database. Oggi, invece, i moderni sistemi possono essere utilizzati per compiere operazioni su un gran numero di basi di dati differenti. Questa "specializzazione" era dovuta alla necessità di guadagnare in velocità di esecuzione pur perdendo in flessibilità [6].

1.2.3 I principali tipi di DBMS

Al giorno d'oggi esiste una grandissima gamma di DBMS, ma prima di elencarne i principali è bene notare che un DBMS puó essere visto come un'architettura software a 3 livelli:

1. Schema esterno: descrive come si presenta il db;
2. Schema logico: descrive cosa rappresenta il db;
3. Schema fisico: descrive come e dove sono memorizzati i dati.

Ed è proprio in base al livello logico che differiscono i vari tipi di DBMS: [7]

Modello Relazionale: chiamato anche con l'acronimo RDBMS, è il sistema di gestione di database relazionali. In questo modello i dati sono registrati in tabelle a due dimensioni, ovvero composte da sole righe e colonne e la manipolazione di questi dati si fa secondo la teoria matematica delle relazioni.

Modello Gerarchico: i dati sono classificati gerarchicamente, secondo un arborescenza discendente. Questo modello utilizza dei puntatori tra le diverse registrazioni. Si tratta del primo modello di DBMS.

Modello Reticolare: come il modello gerarchico questo modello usa dei puntatori verso le registrazioni. Tuttavia la struttura non è più necessariamente arborescente in senso discendente.

Modello ad Oggetti: chiamato anche con l'acronimo ODBMS, è il sistema di gestione di database oggetto. In questo modello i dati sono registrati sotto forma di oggetti, cioè di strutture chiamate classi che presentano dei dati membri. I campi sono istanze di queste classi.

Approcci NoSQL: che vedremo in dettaglio nel capitolo successivo.

Viene elencata di seguito una lista dei principali DBMS: [7]

- Microsoft SQLServer 2008 offre una piattaforma dati affidabile, produttiva ed efficiente per eseguire le più esigenti applicazioni, abbattere i tempi e costi di sviluppo e di gestione di applicazioni e fornire informazioni traducibili in azioni a tutti i livelli dell'organizzazione. SQL Server è alla base di BI software, ossia di Business Intelligence software come il Data Warehouse (archiviazione e immagazzinamento) utile per produrre relazioni e analisi all'interno di un'organizzazione.
- Nasce come SyBase SQL Server, nel 1996 prende il nome di Adaptive Server Enterprise. Viene utilizzato specialmente per l'allocazione dinamica della memoria, su piattaforme che offrono servizi Java, XML, SSL.
- MySQL è un DBMS relazionale inserito in diverse piattaforme come LAMP, acronimo di Linux, Apache, MySQL, o MAMP, acronimo di

Mac, Apache, MySQL. Serve per la creazione di siti e applicazioni Web dinamiche. I siti di Wikipedia sono gestiti dal software MediaWiki che è basato su un database MySQL.

- Access, prodotto dalla Microsoft, è il più diffuso sistema di gestione di basi di dati per l'ambiente Microsoft Windows. Si puó usare in due modalità: come gestore di basi dati autonomo su pc e come interfaccia verso altri sistemi. Un esempio di modalità come interfaccia è servirsi di Access come strumento che permette di evitare di scrivere in SQL, in quanto acquisisce schemi e semplici interrogazioni tramite una rappresentazione grafica facilmente comprensibile; questi input vengono tradotti in comandi SQL in modo trasparente.

- Oracle è uno dei più famosi database management system (DBMS), scritto in linguaggio C. Esso fa parte dei cosiddetti RDBMS, ovvero dei sistemi di database basati sul modello relazionale. La società informatica che lo produce è la Oracle Corporation [7] , e rilasció la prima versione di Oracle nel 1977.

- PostgreSQL è un completo DBMS ad oggetti rilasciato con licenza libera. Spesso viene abbreviato con "Postgres", sebbene questo sia un nome vecchio dello stesso progetto. PostgreSQL è una reale alternativa, sia ad altri prodotti liberi come MySQL, sia quelli a codice chiuso come Oracle ed offre caratteristiche uniche nel suo genere che lo pongono per alcuni aspetti all'avanguardia nel settore dei database.

- SQLite è una libreria software scritta in linguaggio C che implementa un DBMS SQL di tipo ACID [8] incorporabile all'interno di applicazioni mobile, utilizzato soprattutto in App implementate con linguaggio

[7] La Oracle Corporation è una delle società informatiche più grandi del mondo, fondata nel 1977 ed ha la sua sede centrale in California. Il fondatore, nonché Chief Executive Officer ed importante azionista è Lawrence J. Ellison. https://it.wikipedia.org/wiki/Oracle

[8] Nell'ambito dei database, ACID deriva dall'acronimo inglese Atomicity, Consistency, Isolation, Durability (Atomicità, Coerenza, Isolamento e Durabilità) ed indica le proprietà

Android. Il suo creatore, D.Richard Hipp, lo ha rilasciato nel pubblico dominio, rendendolo utilizzabile quindi senza nessuna restrizione. Permette di creare una base di dati incorporata in un unico file come nel caso dei moduli Access. SQLite non è un processo stand alone, utilizzabile di per se, ma puó essere incorporato all'interno di un programma. è utilizzabile con il linguaggio C/C++/Java e molti altri linguaggi.

1.3 Database NoSQL

Il termine NoSQL identifica tutti quei database che si discostano dalle regole che caratterizzano i database relazionali, detti RDBMS. Secondo questa definizione, rientrano nella categoria dei database non relazionali tecnologie e strumenti molto diversi tra di loro ma con un fattore in comune: essi sono spesso utilizzati per immagazzinare grandi quantità di dati e sono altamente scalabili [9] [1].

La nascita di questa tipologia di database risale agli anni '60, con alcuni software Multi Value (implementato nel 1965 nell'azienda TRW) e IMS (sviluppato da IBM per il programma spaziale Apollo, di cui abbiamo già parlato nei capitoli precedenti). Tuttavia il nome NoSQL, con l'attuale significato, è comparso nel 2009 quando Eric Evans, dipendente della Rackspace [10] , lo utilizzò per definire la branchia dei database non relazionali [13].

Come possiamo notare dalla Figura 1.1, al giorno d'oggi, MongoDB, che è il maggior rappresentante di questa tipologia di database, copre una grande fetta di mercato del mondo delle basi di dati.

logiche che devono avere le transizioni. https://it.wikipedia.org/wiki/ACID

[9] Il termine scalabilità si riferisce alle capacità di un sistema di "crescere" o di diminuire di scala in funzione delle necessità e delle disponibilità. Un sistema che gode di questa proprietà viene detto scalabile. https://it.wikipedia.org/wiki/Scalabilit%C3%A0

[10] Rackspace Inc. è una società di cloud computing gestito con sede a Windcrest, Texas, Stati Uniti d'America. https://en.wikipedia.org/wiki/Rackspace

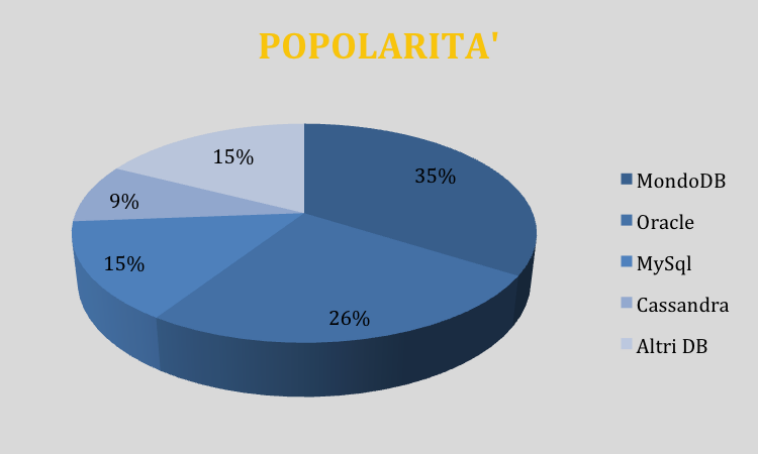

Figura 1.2: Grafico rappresentante la percentuale di popolarità dei maggiori DBMS

1.3.1 Origine del nome NoSQL

Il termine NoSQL fu usato per la prima volta nel 1998 per una base di dati relazionale open source che non usava l'interfaccia SQL. L'autore, Carlo Strozzi, dichiarò che "come movimento, NoSQL diparte in modo radicale dal modello relazionale e quindi andrebbe chiamato in modo più appropriato NoRELL, o qualcosa di simile" [12].

Contrariamente a quanto si potrebbe pensare, questo tipo di movimento non è contrario ai database relazionali, infatti il termine NoSQL è acronimo di Not Only SQL, appunto a significare che esistono diverse situazioni per le quali il modello relazionale risulta inappropriato, ma tante altre per le quali tale modello è ancora la soluzione migliore.
NoSQL viene definito come "Next generation databases mostly addressing some of the points: being non-relational, distributed, open source and horizzontally scalable"; che tradotto sta a significare "Una banca dati di nuova generazione caratterizzata da: non essere relazionale, distribuita, open source e scalabile orizzontalmente" [15].

1.3.2 Caratteristiche principali

Le proprietà principali dei sistemi NoSQL sono le seguenti:

Essi sono database distribuiti, ovvero una collezione di dati logicamente appartenente allo stesso sistema e distribuiti su più server collegati in rete [16].

Sono strumenti generalmente open-source, ovvero sono software di cui gli autori rendono pubblico il codice sorgente, favorendone il libero utilizzo e permettendo ai programmatori indipendenti di apportarvi modifiche ed estensioni.

Non dispongono di uno schema, ovvero di una struttura fissa del database.

Non supportano le operazioni di join, ovvero un operazione che permette di selezionare dati da più tabelle, evidentemente correlate tra di loro.

Non implementano le proprietà ACID delle transazioni, ma delle proprietà chiamate BASE. Esse sono state introdotte da Eric Browers, autore anche del teorema di CAP di cui parleremo nel capitolo successivo. L'acronimo BASE sta ad identificare:

- *Basically Available*: ad ogni richiesta vi è una garanzia di risposta, anche nel caso in cui il sistema distribuito sia soggetto a guasti;
- *Soft State*: la consistenza dei dati non è garantita in ogni istante;
- *Eventually Consistent*: il sistema diventa consistente dopo un certo intervallo di tempo, se le attività di modifica dei dati cessano.

Queste proprietà sono dovute principalmente al fatto che questa tipologia di database è nata per essere veloce, flessibile e distribuita.

I sistemi NoSQL sono scalabili orizzontalmente, un database si dice scalabile orizzontalmente quando aggiunge nuovi nodi a quelli esistenti; invece si dice che un database è scalabile verticalmente quando aumenta la potenza del singolo nodo, potenziando quindi RAM, CPU o hard disks.[17].
La scalabilità orizzontale dei database NoSQL consente di poter fare a meno delle prestazioni fornite da harware ad alto costo, utilizzando invece commody harware [11] . I nodi di un cluster [12] su cui è installato un database NoSQL possono essere aggiunti o rimossi senza particolari problematiche di gestione, realizzando così una piena scalabilità orizzontale a costi moderati [1].

Essi sono in grado di gestire grandi moli di dati.

E infine, supportano le repliche dei dati, cosa impossibile per gli altri tipi di DBMS.

1.3.3 Fattori che hanno portato alla sua diffusione

Le motivazioni che hanno portato alla diffusione del movimento NoSQL sono sostanzialmente raggruppabili i tre grandi fattori:

1. Gestione dei Big-data : i big data, come spiegato nel primo capitolo, sono moli di dati eterogenei, distribuiti e difficili da gestire attraverso le tecnologie tradizionali, come gli RDBMS. I sistemi NoSQL riescono a fare fronte al problema dei big data su tutti loro aspetti, quali volume, velocità e varietà.

2. Limitazione del modello relazionale: il modello NoSQL supera quelli che sono i tre grandi limiti del modello relazionale.

[11] Commody hardware è un componente periferico o un dispositivo che è relativamente poco costoso, essi sono ampliamente disponibili e più o meno intercambiabili tra di loro. http://whatis.techtarget.com/definition/commodity-hardware

[12] Con cluster si indica un agglomerato di oggetti dello stesso tipo; nei dispositivi come le memorie di massa, indica l'unità logica di memorizzazione di un file. http://www.pc-facile.com/glossario/cluster/

Il *primo* grande *limite* è quello del vincolo della forma tabellare dei dati, superato dal modello NoSQL in quanto esso accetta una struttura e una forma dei dati molto più libere, non si parla più di tabelle con righe e colonne, ma si parla di collezioni e insiemi di documenti.

La *seconda limitazione* è quella relativa alle operazioni implementabili in SQL, molto limitate rispetto al modello NoSQL. Ad esempio in SQL non è possibile memorizzare un grafo e calcolare il percorso minimo tra due punti.

La *terza* ed ultima *limitazione* dei DBMS relazionali sta nella scalabilità. Il loro tipo di scalabilità comporta una serie di problemi, come l'obbligo di gestione dei vincoli, l'impossibilità di replicare i dati, una difficile gestione delle transazioni e la necessità di soddisfare le proprietà ACID; tutti problemi risolti dal modello NoSQL grazie all'adozione della scalabilità orizzontale.

3. Teorema CAP: questo teorema afferma che è impossibile, per un sistema informatico distribuito, ovvero un sistema gestito da un cluster, garantire contemporaneamente tutte e tre le seguenti proprietà, ma al massimo due alla volta:

   Coerenza: tutti i nodi vedono gli stessi dati nello stesso tempo;

   Disponibilità (Availability): garanzia che ogni richiesta riceva una risposta, sia che la query sia andata a buon fine, sia che la richiesta non abbia avuto successo;

   Tolleranza di partizione: il sistema continua a funzionare correttamente anche in presenza di perdita di messaggi o di partizionamenti della rete.

1.3.4 Principali modelli NoSQL

Il termine NoSQL identifica una moltitudine di DBMS, basati principalmente sui quattro modelli logici che seguono.

Database chiave/valore: questi database sono basati sul concetto di associative array, cioè una semplice struttura in grado di concentrare un insieme di coppie chiave/valore. [1].

La *chiave* in questa tipologia di modello rappresenta quindi un valore unico utilizzato per le operazioni di ricerca; invece il *valore* è qualsiasi cosa che rappresenti la chiave. La Tabella 1.1 rappresenta un esempio di quello cha che potrebbe essere una rappresentazione del modello chiave valore:

CHIAVE	VALORE
1	{ 2171077, Hotel Villa Nacalua, Via Dell'Autostrada 5, Citta Sant'Angelo, PE, ITA, 42.522192, 14.133172 }
2	{ 8706482, Hotel Royal, StreetAddress, Viale Dalmazia 132, Vasto, CH, ITA, 42.090313, 14.730133 }
3	{ 1207496 , La Réserve Hotel Terme , Via Santa Croce sn , Caramanico Terme , PE , ITA , 42.159294 , 14.012636 }

Tabella 1.1: Esempio di tabella corrispondente alla collezione di Hotel Expedia, utilizzando il modello chiave valore.

I principali DBMS che utilizzano questo modello sono ad esempio BerkeleyDB, Project Voldemort.

Database document-oriented: questo modello è simile al modello chiave valore, tranne che per il fatto che il valore non è trasparente per il database ma è un formato che il sistema può interpretare e interrogare. I formati più usati per la memorizzazione del valore sono XML e JSON. JSON, essendo

semplicemente un oggetto JavaScript serializzato, può essere molto più utile in ambiente web. La porzione di codice seguente dimostra come potrebbe essere un esempio di oggetto json.

```
{
"_id" : ObjectId("56b23e68783b8ea0b30e7f41"),
"HotelID" : "2171077",
"Name" : "Hotel Villa Nacalua",
"Location" : {
  "StreetAddress" : "Via Dell'Autostrada 5",
  "City" : "Citta Sant'Angelo",
  "Province" : "PE",
  "Country" : "ITA",
  "GeoLocation" : {
    "Latitude" : "42.522192",
    "Longitude" : "14.133172"
    }
  }
}
```

Generalmente i DBMS document-oriented utilizzano una o più proprietà degli oggetti per indicizzarli ed è possibile effettuare delle interrogazioni basate sulle proprietà dell'oggetto [18].

I principali DBMS che utilizzano questo modello sono ad esempio MongoDB, utilizzato per la mia tesi, e di cui ne parleremo in modo più approfondito nel prossimo capitolo, e CouchDB.

Database column-oriented: questo modello è caratterizzato dal fatto che i dati sono organizzati su colonne, al contrario di quanto avviene su i tradizionali RDMS, nei quali i dati vengono memorizzati sulle righe. Un insieme di colonne viene chiamata Column family, che rappresenta quindi un contenitore di colonne; ogni Column family è scritta su un file diverso e ogni riga dispone di una chiave primaria, chiamata row key.

I vantaggi del modello column-oriented stanno nel fatto che utilizza uno schema abbastanza flessibile, ha una grande efficienza nello storage, ovvero nella memorizzazione dei dati, e infine nel fatto che vi è una maggiore possibilità di compressione dei dati.

Un esempio di come potrebbe essere memorizzato un hotel Expedia su un database column-oriented, viene rappresentato dalla Tabella 1.2.

HotelID	
ID	Value
1	2171077
2	8706482
3	1207496

Name	
ID	Value
1	Hotel Villa Nacalua
2	Hotel Royal
3	La Réserve Hotel Terme

StreetAddress	
ID	Value
1	Via Dell'Autostrada 5
2	Viale Dalmazia 132
3	Via Santa Croce sn

Tabella 1.2: Esempio di una parte di hotel Expedia su un db column-oriented.

I principali DBMS che utilizzano questo modello sono ad esempio HBase e Cassandra.

Database graph-oriented: il modello chiave valore e il modello orientato agli oggetti hanno il problema che non sono adatti a contenere dati molto interconnessi: un fattore molto limitante in un'applicazione complessa come ad esempio un social network. Un database a grafi [13] può essere visto come un caso particolare di un database orientato ai documenti in cui alcuni particolari documenti rappresentano le relazioni.

Questo tipo di database è sicuramente molto potente, se consideriamo che il modello permette un'operazione molto interessante: l'attraversamento. Rispetto ad una normale query su database chiave-valore, l'attraversamento

[13] Un grafo è un insieme di elementi detti nodi o vertici che possono essere collegati fra di loro da linee chiamati archi o spigoli. https://it.wikipedia.org/wiki/Grafo

stabilisce come passare da un nodo all'altro, utilizzando le relazioni tra i nodi [18].

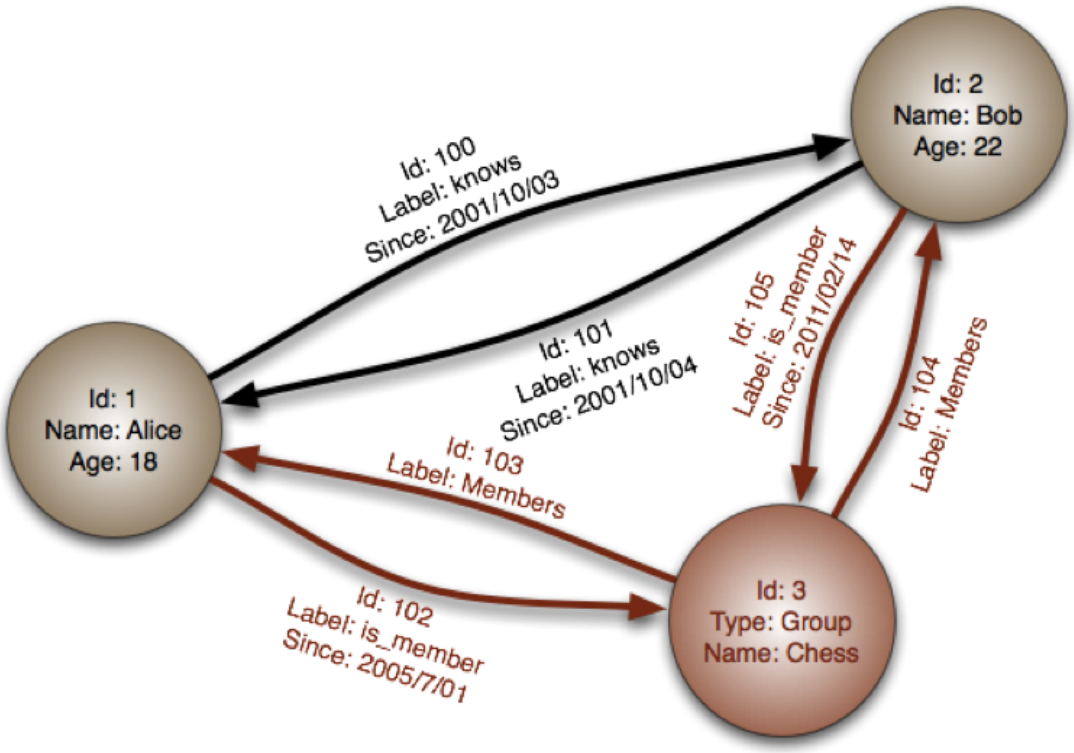

Figura 1.3: Database graph-oriented.

Quindi possiamo affermare che in questo modello i dati sono strutturati a forma di grafi; in particolare i nodi corrispondono agli attribuiti, quindi le righe di un RDBMS, e gli archi rappresentano le relazioni tra gli attributi.

I principali DBMS che utilizzano questo modello sono ad esempio Neo4J e Titan.

1.4 MongoDB

MongoDB, come detto in precedenza, è una tipologia di database NoSQL, basato sul modello document-oriented, quindi caratterizzato dalla possibilità di gestire dati eterogenei, quindi non per forza omogenei, e complessi. MongoDB supporta un database organizzato in collezioni, le collezioni contengono liste di documenti e ogni documento è un insieme di campi.

In un database RDBMS le collezioni corrispondono alle tabelle, i documenti alle righe, e i campi alla colonna di una riga.

1.4.1 Introduzione

La parola MongoDB, deriva dalla parola humongous che significa enorme. Esso è un DBMS non relazionale orientato ai documenti, come già detto più volte. Classificato come database di tipo NoSQL, MongoDB si allontana dalla struttura tradizionale basata su tabelle, come avviene nei database relazionali, lasciando spazio a documenti BSON [14] con schema logico dinamico, rendendo l'integrazione di dati di alcuni tipi di applicazioni più facile e più veloce. Rilasciato sotto una combinazione della GNU Affero General Public License [15] e dell'Apache License [16] , MongoDB è un software libero e open

[14]BSON è un formato binario per rappresentare strutture dati semplici e array associativi (chiamati oggetti o documenti in MongoDB). Il nome "BSON" è basato sul termine JSON e significa "JSON Binario" dall'inglese "Binary JSON". https://it.wikipedia.org/wiki/BSON

[15]La GNU Affero General Public License (AGPL) è una licenza di software libero. Essa si caratterizza dal fatto che si riferisce all'utilizzo del software su una rete di calcolatori, e non su uno singolo. Questa licenza richiede che il codice sorgente, se modificato, sia reso disponibile a chiunque utilizzi l'opera sulla rete.

[16]La Licenza Apache è una licenza di software libero con copyleft scritta da Apache Software Foundation che obbliga gli utenti a preservare l'informativa di diritto d'autore e d'esclusione di responsabilità nelle versioni modificate.

source.

MongoDB viene sviluppato inizialmente dalla società software 10gen, che oggi prende il nome di MongoDB Inc. , nell'Ottobre 2007, ma solo nel 2009 l'azienda si sposta verso un modello open source. Da allora MongoDB è stato adottato come back-end da un alto numero di grandi siti web e società di sevizi. MongoDB è oggi il più popolare database NoSQL [19].

1.4.2 Caratteristiche ed elementi principali

MongoDB si identifica rispetto agli altri DBMS per alcune particolarità come: [1].

- La possibilità di gestire dati complessi;
- Privilegia le performance rispetto alle funzionalità fornite;
- Portabilità, ovvero la possibilità di eseguire MonogDB su molteplici sistemi operativi;
- Alta disponibilità, attraverso le repliche;
- Scalabilità attraverso lo sharding, cioè il processo con cui è possibile partizionare una collection, suddividendo i documenti in essa contenuti tra più istanze di MongoDB.

Gli elementi principali di MongoDB sono invece: [1]

Documenti: MongoDB, come già detto, utilizza un formato, chiamato BSON, per rendere persistenti i documenti. Questi documenti possono contenere coppie chiave/valore, dove il valore può, a sua volta, contenere un altro documento o un array di documenti, oltre che a tipi di dato base, come stringhe, date, numeri interi etc. Ogni documento ha un campo predefinito, il campo "id", che viene assegnato in fase di inserimento dall'utente o dal

sistema. I documenti BSON posso avere una dimensione massima di 16MB.

Collezioni: Gli oggetti BSON sono raggruppati in collection. Anche se assomigliano molto alle tabelle di RDBMS, le collezioni non richiedono uno schema fisso, infatti gli oggetti BJON che vi appartengono non devono necessariamente avere lo stesso insieme di campi.

Database: le collezioni a loro volta sono contenuti in un database.

1.4.3 Utilizzo

Utilizzare MongoDB è molto semplice, basta inserire comandi in linguaggio JavaScript tramite shell o driver immersi in linguaggio ospite, come ad esempio Java. È necessario l'utilizzo del linguaggio JSON come input/output delle query di aggiornamento e selezione e l'utilizzo del linguaggio BSON per rappresentare i documenti internamente.

Lo strumento da riga di comando consente, tra le altre funzionalità, di aprire una shell di comandi, oppure di lanciare script salvati su un file .js. Le istruzioni sono molto semplici e ne vengono riportate le principali nella Tabella 1.3 [20].

SQL	MongoDB	DESCRIZIONE
INSERT INTO hotels VALUES (...)	db.hotels.insert(...)	Inserimento di un documento nella collezione
SELECT * FROM hotels	db.hotels.find()	Ricerca di tutti i documenti nella collezione
SELECT * FROM hotels WHERE {HotelID = 2171077}	db.hotels.find({"HotelID":"2171077"})	Ricerca di un documento specificando un valore
SELECT Nome FROM hotels WHERE {HotelID = 2171077}	db.hotels.find({"HotelID":"2171077"}, {"Name": 1})	Ricerca di un solo valore del documento
SELECT DISTINCT HotelID FROM hotels	db.hotels.distinct("HotelID")	Ricerca di tutti i documenti senza duplicati
SELECT COUNT(*) FROM hotels	db.hotels.find().count()	Conta il numero dei risultati della query
SELECT * FROM hotels LIMIT 5	db.hotels.find().limit(5)	Restituisce solo i primi n, in questo caso 5, risultati della query
SELECT StarRating FROM hotels ORDER BY StarRating DESC;	db.hotels.find().sort({"StarRating":-1})	Ordina i risultati in modo decrescente, come nel nostro caso, o crescente
DELETE FROM hotels WHERE {HotelID = 2171077}	db.hotels.remove({"HotelID":"2171077"})	Elimina dalla collezione un determinato oggetto

Tabella 1.3: Confronto operazioni SQL e MongoDB.

Capitolo 2

Big-data e turismo

Nel corso degli ultimi anni i social media [1] stanno assumendo un ruolo sempre più rilevante nello scambio di informazioni e valutazioni di prodotti e servizi, influenzando le scelte dei consumatori con conseguente impatto sui risultati economici di interi settori.
Diventa pertanto importante avere appropriati strumenti di analisi di questa enorme massa di informazioni, quali solo big data possono fornire, da parte delle aziende, per poter analizzare le informazioni scambiate dai loro clienti o potenziali clienti ed utilizzare queste analisi nelle loro strategie di sviluppo, ad esempio prevedendo trend di preferenze.
Uno dei settori di maggiore impatto è il settore alberghiero, ove specifici social network e piattaforme di booking [2] sono sempre più utilizzate (come il nostro studio dimostrerà) per scambiare esperienze tra turisti e influenzare le intenzioni di prenotazione. Per questi motivi sono stati fatti parecchi studi per mettere in correlazione i dati estraibili da queste piattaforme ed ottenere informazioni sul turismo e sui turisti.

[1] Social media, in italiano media sociali, è un termine generico che indica tecnologie e pratiche online che le persone adottano per condividere contenuti testuali, immagini video e audio.

[2] Le piattaforme di tipo booking sono quei siti web dove l'utente ha la possibilità di prenotare voli, alberghi e molto altro.

In questo capitolo verranno quindi riportati diversi studi aventi come tema principale i big data e le relative analisi di informazioni estratte da piattaforme di booking/social neteork nel settore del turismo. Questi studi sono stati presi da diversi articoli e verranno raggruppati in base alle loro conclusioni ai particolari aspetti delle tese sostenute.

2.1 Correlazione tra valutazione e volume delle recensioni

In corrispondenza dell'aumento della popolarità dei social media, utilizzati dai turisti per la prenotazione dei loro viaggi, è aumentato anche il numero delle recensioni postate dagli utenti per descrivere i loro viaggi e le loro esperienze di soggiorno negli hotel. Con l'aumento del numero delle recensioni è aumentato anche il trend delle valutazioni riferite al singolo hotel, e proprio per questo le valutazioni dei clienti sul web sono cresciute di importanza per le imprese turistiche.

Questo aumento è dimostrato nell'articolo intitolato "Online Customer Reviews of Hotels. As Participation Increases, Better Evaluation Is Obtained" , il quale riporta uno studio basato su un campione ampio e variegato di 16680 hotel in 249 zone turistiche. Lo studio ha rilevato una relazione tra valutazione e volume, ovvero che mentre le prime recensioni erano praticamente tutte negative, con l'aumentare del loro numero nel tempo, quelle positive hanno iniziato ad assumere sempre più rilevanza, bilanciando l'eccessivo numero di critiche rispetto agli apprezzamenti. Attualmente il numero delle recensioni positive ha superato quello delle recensioni negative.
Questo studio dovrebbe indurre ogni hotel a cercare di acquisire sempre un numero più elevato di recensioni, in modo tale da avere un giusto equilibrio tra recensioni positive e recensioni negative; infatti minore è il numero delle recensioni, per un determinato hotel, e maggiore è il numero di recensioni negative per quell'hotel; a confermare la tesi sta anche la scoperta che le

prime recensioni che ricevono gli hotel, sono per lo più negative, non solo per gli hotel di bassa categoria, ma anche per gli hotel considerati più belli [22].

2.2 Come il management degli hotel utilizza i social media

Un'altra tipologia di studio, sempre in riferimento alle recensioni, si preoccupa di definire il rapporto che hanno gli albergatori con le recensioni e con i social media; ovvero si parla di studi che hanno l'obiettivo di definire l'utilizzo dei social media da parte degli hotel, delle loro risposte alle recensioni e di come queste risposte sono considerate dai turisti.

Un articolo in particolare, che si intitola *"Responding to Online Reviews: Problem Solving and Engagement in Hotels"* tratta di un'analisi avente come obiettivo lo studio di quattro hotel di fascia alta nella zona occidentale degli Stati Uniti, in particolare sul loro comportamento in relazione alle recensioni che ottengono su TripAdvisor. Questi quattro hotel sono stati scelti come campione, proprio perché scelgono due approcci completamente differenti in merito alla scelta di risposta del management dell'hotel alle recensioni.

Dall'analisi si evince che due hotel rispondono regolarmente ai commenti dei loro clienti, mentre gli altri due non rispondono quasi mai. In primo luogo, gli hotel che hanno risposto frequentemente alle recensioni dei loro clienti, considerano questo tipo di interazione uno scambio di fiducia reciproco, mentre gli hotel che non hanno dato risposte ritengono che le loro recensioni abbiano una visione estremamente positiva o estremamente negativa, e quindi non necessitano di risposta. In secondo luogo, da questa analisi, si è ricavato che gli hotel che rispondevano alle recensioni in modo frequente avevano anche uno stile di risposta collaborativa, che ha comportato una regolare consultazione della pagina TripAdvisor dell'hotel e un sempre mag-

giore scambio di opinioni; dall'altro canto, per quanto riguarda gli hotel poco attivi in questo campo, anche i clienti si sono rilevati sempre meno attivi e la loro pagina veniva consultata solo se necessario. Altro aspetto rilevante di questo studio è stata la scoperta che le pagine di TripAdvisor, degli hotel attivi nel rispondere alle recensioni, vengono gestite da impiegati interni all'hotel, al contrario degli hotel poco attivi, la cui pagina viene gestita da enti esterni all'hotel, quindi non dai dipendenti [23].

Sempre per quanto riguarda lo studio avente come obiettivo quello di definire il rapporto che hanno gli albergatori con i social media. è interessare riportare un articolo che si preoccupa di definire come dovrebbero essere le risposte degli albergatori alle recensioni negative.

"Factors Affecting Customer Satisfaction in Responses to Negative Online Hotel Reviews" parla infatti di un'analisi delle tipologie di risposte degli hotel alle recensioni negative che essi ottengono, e ha portato alla conclusione che la risposta dovrebbe avere gli stessi principi di una risposta fatta ad un reclamo orale del cliente.

L'analisi, basata su una serie di ipotetiche risposte fatte a recensioni negative, ha dimostrato che una risposta empatica da parte dell'hotel migliora la sua valutazione. Allo stesso modo, un gruppo di 176 potenziali clienti valuta positivamente una risposta, se in essa è presente un riferimento specifico alla lamentela della recensione, rendendo in questo modo la risposta più personale e meno generica. È interessante notare anche che la tempistica con cui un hotel risponde ad una recensione negativa non influenza la valutazione attribuita alla risposta. Questa tesi porta alla conclusione che la risposta ad una recensione negativa dovrebbe avere lo stesso tono di una risposta ad una critica orale. La principale differenza tra una critica orale e una recensione negativa è che, nella critica orale un lungo tempo di attesa di una adeguata risposta di motivazione influenza la valutazione in modo negativo, contra-

riamente per la recensione online ove in realtà l'utente non resta veramente ad attendere una risposta, quindi il tempo d'attesa non influenza sulla valutazione. Questo studio suggerisce inoltre che i gestori dell'hotel dovrebbero includere risposte empatiche o riferimenti specifici alla critica ricevuta [24].

2.3 Recensioni e risposte degli hotel: positive o negative, quali sono le più numerose?

Come si può evincere anche dagli articoli precedentemente citati, le recensioni si possono suddividere in due grandi categorie, le recensioni positive e le recensioni negative. E qui entra un altro importantissimo caso di studio, ovvero quello che si occupa di calcolare il volume delle recensioni positive in relazione a quello delle recensioni negative.

Un primo studio viene descritto dall'articolo *"What can big data and text analytics tell us about hotel guest experience and satisfaction?"* . Questo studio si propone di esplorare e dimostrare l'utilità di Big Data Analytics per comprendere meglio importanti questioni sull'ospitalità, vale a dire il rapporto che intercorre tra l'esperienza degli ospiti degli hotel e la loro soddisfazione. In particolare, questo studio applica un approccio Text Analytics su una grande quantità di recensioni di consumatori estratte da Expedia.com per decomporre l'esperienza in hotel degli ospiti ed esaminare la sua associazione con indici di soddisfazione. Lo studio qui citato, porta alla considerazione che la soddisfazione dei clienti tende ad essere più sul lato positivo [25].

Un altro articolo, intitolato *"Customer engagement behaviors and hotel responses"*, riporta invece uno studio che porta a definire come i potenziali clienti percepiscono le due tipologie di recensioni e le risposte del management degli hotel alle recensioni.
I risultati dello studio sperimentale hanno dimostrato che le recensioni po-

sitive sono maggiormente considerate rispetto a quelle negative. Per quanto riguarda le risposte degli Hotel alle recensioni negative, si è riscontrato una maggiore efficacia nelle risposte specifiche rispetto a quelle generiche. I potenziali clienti percepiscono le recensioni positive come più utili e credibili rispetto a quelle negative. Tali risultati sono ulteriormente confermati da risultati qualitativi: le recensioni positive facilitano il processo decisionale di altri clienti attraverso la condivisione di esperienze positive, mentre il posting negativo di commenti può essere considerato una ritorsione verso l'hotel dopo una sgradevole esperienza.
Questo studio ha trovato che i potenziali clienti interpretano le risposte positive degli hotel come un apprezzamento per i loro clienti o parte della strategia di gestione della relazione del Cliente. Per risposte a recensioni negative, i clienti percepiscono che le risposte possono avere tre motivazioni: gestione della relazione del cliente, gestione della reputazione online, e ripristino del disservizio. Mentre le risposte degli hotel a recensioni positive non ha influenzato la valutazione delle risposte da parte dei potenziali clienti, le risposte specifiche e non generiche a recensioni negative fanno guadagnare fiducia e più alta qualità di comunicazione rispetto a risposte generiche [26].

2.4 Il dibattito sui fattori che influiscono maggiormente le scelte dei potenziali consumatori

L'ultimo articolo citato dimostra che uno dei criteri che maggiormente influenza il turista nella scelta dell'hotel è la valutazione. Un altro parametro che influenza questo tipo di scelta, è, come spiegato nell'articolo *"Compliance with eWOM : The influence of hotel reviews on booking intention from the perspective of consumer conformity"* il numero totale delle recensioni.
Questo articolo riporta infatti uno studio che ha determinato l'influenza del

rating delle recensioni, la quantità di recensioni, e gli effetti di interazione tra loro (tra rating e quantità) per i consumatori classificati come conformisti e anticonformisti.
In una prima fase dello studio, è stato rilevato che il rating delle recensioni ha una significativa influenza sulla intenzione di prenotazione. In altre parole, quando un potenziale cliente legge una recensione positiva, aumenta significativamente la sua propensione a prenotare, e vice versa; l'influenza della recensione positiva è rafforzato dal numero di recensioni, come pure l'esposizione ripetuta a recensioni negative è particolarmente dannoso. Pertanto l'aumento del numero di recensioni rafforza l'influenza della valutazione (sia se positive che negative) sulle intenzioni di prenotazione.
In una seconda fase dello studio, è stato scoperto che i consumatori conformisti sono più propensi a farsi influenzare dal passaparola online, indipendentemente da rating o quantità. La persuasività di recensioni positive è stata mostrata essere più pronunciata tra conformisti. Inoltre, un piccolo numero di recensioni è sufficiente per convincere i conformisti, mentre i non conformisti richiedono un maggiore numero per essere persuasi. Così, quantità di recensioni e ripetuta esposizione sono fondamentali quando si tratta di non-conformisti [39].

A testimoniare che non c'è solo il rating della valutazione, ma anche il volume come fattore di influenza è il seguente articolo, intitolato "Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry". Lo studio riportato su questo articolo esplora l'impatto del rating e numero di recensioni sul valore generato attraverso transazioni on-line di un hotel. Attraverso la collaborazione con società di consulenza di viaggio Click, il team di ricerca ha raccolto un campione di 178 hotel in rappresentanza di varie catene negli Stati Uniti.
I risultati della ricerca dimostrano che il rating di TripAdvisor, nonché il numero di recensioni hanno avuto una relazione positiva con il valore medio di ogni prenotazione online; ovvero più alta è la valutazione in stelle di

TripAdvisor , maggiore è il valore della prenotazione; analogo impatto ha il numero di recensioni. La presente ricerca dimostra l'impatto delle recensioni sulla posizione finanziaria di un hotel [33].

Questo dimostra che i criteri che influenzano la scelta del turista possono essere due, il volume delle recensioni, ovvero il numero totale, e il rating, ovvero la valutazione attribuita con la recensione. Parecchi articoli parlano di queste due caratteristiche e molti di questi hanno pareri discordanti.

2.4.1 Rating

Ad accreditare la tesi che sono le valutazioni delle recensioni a definire la scelta del turista vengono qui proposti due articoli.

Il primo, che ha come titolo *"The effectiveness of managing social media on hotel performance"* indaga su come le recensioni online influiscano sulle performances economiche degli hotel. Una catena alberghiera internazionale ha fornito i dati di performance ed i dati delle recensione online. Una delle principali società di social media per il settore alberghiero ha raccolto i dati di recensioni online. I risultati indicano che le valutazioni complessive sono il fattore predittivo più importante delle performance di un hotel, seguito dalla risposta ai commenti negativi. Migliori sono le valutazioni complessive e più alto è il tasso di risposta ai commenti negativi, più alto è il rendimento economico dell'albergo. Pertanto, recensioni online e social media, in particolare il punteggio complessivo e risposta ai commenti negativi, devono essere gestiti come una parte fondamentale del marketing alberghiero.
Contrariamente alle nostre aspettative, il volume delle recensioni e la deviazione standard delle valutazioni non hanno avuto effetti significativi. E' stato inoltre rilevato che ogni hotel ha ricevuto, in media, 32 recensioni; il tasso di risposta medio per i commenti negativi è stato di circa il 7%; il tasso di risposta ai commenti negativi era superiore al tasso di risposta ai commenti

positivi. Un contributo rilevante di questo studio è che la risposta ai commenti negativi è un fattore determinante della performance albergo [29].

L'articolo *"Web reviews influence on expectations and purchasing intentions of hotel potential customers"* propone uno studio sperimentale, che prende come campione 349 giovani e adulti i quali sono stati coinvolti in un sondaggio online che ha chiesto di immaginare la ricerca di un hotel e leggere le recensioni di altri clienti di un albergo ipotetico prescelto. I risultati mostrano una correlazione positiva tra l'intenzione di acquisto e l'aspettativa del cliente con il rating della recensione. Al contrario, la presenza di risposte dell'hotel a recensioni di ospiti ha un impatto negativo sulle intenzioni di acquisto.
Lo studio dimostra come l'intenzione di prenotazione nell'industria alberghiera sia influenzata dal rating (positivo o negativo) delle recensioni. Dimostra inoltre che la presenza di risposte dell'hotel alle recensioni dei clienti non è considerata un fattore chiave dagli intervistati. Al contrario, ha un impatto negativo sulla intenzione di acquisto. La natura delle informazioni in questo caso è probabilmente considerata come non spontanea e di parte [30].

Questo articolo, come il precedente, dimostra che il rating delle valutazione è l'elemento che maggiormente influenza la scelte del turista. A differenza del primo peró, questo articolo, afferma che le risposte degli hotel non influenzano positivamente la scelta dell'utente.

Un ultimo articolo, da citare a testimonianza che il rating è la cosa che influisce maggiormente sulla scelta, ha come titolo *"The business value of online consumer reviews and management response to hotel performance"*. L'articolo parla di uno studio che identifica gli impatti sul business delle recensioni dei consumatori e delle risposte del management degli hotel. L'articolo presenta una analisi su un insieme di dati di recensioni online di consumatori e le risposte del management di 843 hotel su un sito web di recensioni.

Lo studio evidenzia che il punteggio complessivo, la varianza ed il volume delle recensione sono positivamente associati ai risultati economici dell'albergo. Le valutazioni complessive sono percepite come il fattore più importante che influenza le prestazioni albergo, seguito da varianza e quantità delle recensioni. Tuttavia, le risposte dell'Hotel sono negativamente collegate alle prestazioni dell'Hotel. Questo studio indica che il numero di risposte dell'Hotel non è efficace per migliorare i risultati economici. I risultati evidenziano anche l'impatto delle valutazioni specifiche (posizione, pulizia) e delle rispettive varianze sulle performances. Si è rilevato che la associazione tra volume delle recensioni e valutazione complessiva, tra volume delle recensioni e giudizio sulla posizione e sulla pulizia, hanno una relazione positiva con le performances dell'Hotel: il volume delle recensioni rafforza la valutazione complessiva e le valutazioni di posizione e la pulizia [34].

2.4.2 Volume

Altri studi invece dimostrano come il volume delle recensioni sia il criterio che maggiormente influenza i turisti nella losco scelta di prenotazione.

Infatti, l'articolo *"Please, talk about it! When hotel popularity boosts preferences"* dimostra che la preferenza dei consumatori aumenta con il numero di recensioni, indipendentemente che la valutazione media sia alta o bassa. Questo articolo esamina l'impatto della popolarità (misurata come numero di recensioni) e della qualità (misurata come la reputazione online) fornite di ex consumatori. In particolar modo si testa se i consumatori tendono a preferire alternative popolari anche se quelle alternative sono qualificate di povera qualità e se ció varia con i dati demografici (es. età) dei consumatori. I risultati riportati risultati svelano che la presenza di molti recensioni (e dunque essere popolare), a prescindere dalle valutazioni, fa aumentare la preferenza soprattutto tra femmine ed anziani. La maggioranza preferisce hotel valutati negativamente da molti consumatori rispetto a hotel con stessa valutazione

negativa da meno consumatori. Questo comportamento va contro il valore delle informazioni perché la valutazioni di molti dovrebbe fornire maggiore certezza [27].

Il volume delle recensioni non influisce positivamente solo sulla scelta del turista nel decidere l'hotel in cui alloggiare, ma influisce anche sulla scelta di inserire o meno la sua esperienza vissuta durante il periodo di soggiorno nell'hotel. Quindi maggiore è il numero delle recensioni già presenti per quell'hotel e maggiormente l'utente viene invogliato ad aggiungerne un'altra.

Infatti l'articolo intitolato *"Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach"* propone un'analisi fatta prendendo come campione delle recensioni provenienti da diversi siti, fatte in periodi di tempo diversi, e riferite ad hotel localizzati in diverse zone turistiche. Per questo studio è stata utilizzata la tecnica della sentiment analysis, cercando di ottenere come risultato la qualità del servizio dell'hotel e le sue prestazioni.
Il risultato di questa analisi dimostra il vantaggio di utilizzare il testo scritto per misurare in modo più accurato ed efficiente le opinioni dei clienti, e quindi di non basarsi solo sulla valutazione numerica o sulla quantità di recensioni. In secondo luogo, i risultati indicano che le diverse dimensioni delle recensioni hanno un diverso impatto sulla valutazione dei clienti che le leggono. Infine questa tesi a portato alla conclusione che più un hotel è recensito, e più l'utente è invogliato a scrivere anche lui un commento [28].

2.4.3 Altri fattori influenzano il turista

Altri studi, hanno dimostrato che esiste un terzo fattore che influenza la scelta del turista. Questo peró è un fattore che determina il grado di affidabilità che il turista pone su una determinata recensione; stiamo parlando delle informazioni personali dell'utente che posta la sua recensione, se sono pre-

senti questa recensione assume più valore rispetto ad una recensione anonima.

L'articolo che parla di questo studio è intitolato *"Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition"*. Lo studio ha esaminato come la presenza di informazioni di identificazione personale di chi effettua recensioni online possa influenzare la considerazione dei potenziali clienti che leggono le recensioni e le loro intenzioni di prenotazione. I risultati di un esperimento su un campione di 274 studenti universitari indicano come la presenza di informazioni personali influenzi positivamente la credibilità percepita delle recensioni online.
Per giungere alle loro conclusioni, sono state fornite informazioni di identificazione personale dei recensori (Nome, stato di residenza e data del soggiorno) nelle loro recensioni e si è rivelato che la presenza di informazioni di identificazione personale ha una positivo effetto sulla credibilità percepita delle recensioni online, che a sua volta ha un significativo effetto sulla intenzione di prenotazione. In particolare si è notato che la presenza di recensioni ambivalenti trasmettono complessivamente un messaggio negativo, e che se includono informazioni personali riducono l'intenzione di prenotazione sia tra chi aveva in precedenza una predisposizione negativa o neutra, sia se l'aveva positiva [32].

2.4.4 Camere di lusso e camere di fascia bassa

Per far fronte a queste discordanze sui diversi risultati ottenuti dagli studi precedentemente elencati, ovvero tra coloro che sostengono che siano le valutazioni il maggior criterio di influenza, e coloro che sostengono che sia il volume delle recensioni il criterio che maggiormente influenza le scelte dei turisti, è stato scritto un articolo che da una motivazione a questi risultati non conformi. Infatti l'articolo dimostra che è la tipologia di camera a diver-

sificare il criterio di scelta dell'interessato.

L'articolo di cui stiamo parlando, ha come titolo *"The Differential Effects of the Quality and Quantity of OnlineReviews on Hotel Room Sales"* e dimostra che le due caratteristiche principali delle recensioni, ovvero volume e valenza, hanno effetti diversi per gli hotel di diversi catene. Diversi studi hanno dimostrato che le recensioni online influenzano molto sulla scelta dell'hotel, e soprattutto sul guadagno per ogni stanza disponibile. Infatti, per quanto riguarda la valenza di queste recensioni, ovvero per quanto riguarda la valutazione, ha un effetto maggiore per le camere di lusso, invece per le camere di hotel di fascia inferiore la valutazione ha poco impatto. Mentre il numero delle recensioni ha un effetto maggiore per gli hotel meno lussuosi, e un effetto negativo per gli hotel di fascia superiore. Sulla base di uno studio di 319 hotel di Londra, è stato possibile dimostrare che questo effetto è valido sia per le zone urbane, che per le zone extraurbane, ma allo stesso modo è valido anche per le catene di hotel e per gli hotel indipendenti [31].

2.5 Studi correlati

Sempre prendendo in considerazione le recensioni dei turisti è stato possibile determinare alcuni dei loro gusti e alcune delle loro preferenze.

Infatti l'articolo *"A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently"* che si avvale di oltre 86.000 recensioni di clienti di hotel stellati di Hong Kong vuole esplorare la distribuzione e la differenza di comportamento sulle valutazioni online di ospiti di lingua inglese e non di lingua inglese.
Lo studio riportato su questo articolo ha calcolato che i clienti di lingua inglese tendono a dare voti più alti rispetto ai non-inglesi. Questi ultimi sembrano essere più soddisfatti con hotel di classe media, mentre gli inglesi

preferiscono hotel di alta classe. In secondo luogo, negli hotel a cinque stelle gli ospiti inglesi preferiscono un entusiasmo generico da parte del personale degli hotel e non desiderano essere disturbati, mentre quelli non di lingua inglese gradiscono un maggior entusiasmo. Negli hotel a tre, quattro stelle gli Inglesi preferirebbero stanze più grandi; fornendo stanze più grandi questi hotel attirerebbero più ospiti inglesi. Inoltre, gli hotel con un maggior numero di ospiti non di lingua inglese hanno una minore valutazione media rispetto a quelli con un numero minore di ospiti non inglesi, perché quest'ultimi tendono a dare valutazioni minori, particolarmente ad hotel di alta classe [38].

Invece sempre per quanto riguarda i fattori che incidono sulla scelta dell'utente dei social media per la prenotazione dell'hotel, è stato fatto un altro studio, ma che peró non considera più le recensioni, quindi ne la loro valenza e ne il loro volume, ma considera un'altri fattori interessanti.

Questo articolo, intitolato *"The Complex Matter of Online Hotel Choice"*, parte dal presupposto che parecchi studi hanno dimostrato che solitamente i primi risultati di ricerca ottengono molta più attenzione da parte degli utenti rispetto gli ultimi risultati della lista, la stessa cosa accade per quanto riguarda la ricerca degli hotel; ma questi non sono gli unici fattori di interesse per quanto riguarda la scelta di un hotel. Infatti questo studio dimostra che la complessità del processo decisionale del consumatore va oltre al semplice posizionamento dell'hotel nella lista dei risultati. Altri fattori che influenzano la scelta dei consumatori sono il numero di opzioni, la presenza o meno di immagini, il prezzo e la presenza o meno di una descrizione che accompagna le immagini.
Lo studio ha infatti dimostrato i seguenti risultati: se la scelta era compresa tra un campione di 5 hotel, allora gli utenti controllavano tutte le opzioni; se il campione di hotel da prendere in considerazione era di 20 unità, gli utenti iniziavano a scartare hotel in base al prezzo e in base alle immagini piuttosto che in base alla descrizione dell'hotel. Ma ció che ha contribuito

maggiormente ai criteri di scelta è stato il posizionamento dell'hotel tra la lista dei risultati ottenuti. Pertanto apparire tra i primi posti dei risultati di ricerca è un ottimo modo per garantirsi un maggiore successo e una più elevata possibilità di essere notati. A confermare questa testi è il fatto che la metà dei risultati ottenuti, non sono nemmeno stati presi in considerazione dell'utente, perché troppo in basso nella lista. Questo studio ha portato anche conclusioni importanti per quanto riguarda gli effetti delle immagini nel processo decisionale dei soggetti. Quando le immagini erano presenti, i soggetti hanno trascorso molto più tempo sulle pagine web rispetto ad altre pagine senza immagini. Questo avviene di norma per tre motivazioni: la prima sta nel fatto che spesso l'utente è pigro e si annoia a leggere informazioni testuali e preferisce osservare foto e immagini; la seconda motivazione è che la presenza di immagini incrementa il livello di sicurezza dell'acquirente che vede con i propri occhi il prodotto che deve comprare, in questo caso la camera dell'hotel; per ultimo aiuta gli utenti a prendere in considerazione alcuni hotel che solo dalla loro descrizione non avrebbero preso in considerazione, perché poco esaustiva o poco comprensibile [35].

2.5.1 Social media, agenzie di viaggio o altro ancora?

Infine, ci sono altri due articoli, sempre aventi come tema i big data e il turismo, e sempre riguardanti le scelte dei turisti, ma non sono basati su analisi delle recensioni, ma studiano i canali con cui viene effettuata la prenotazione.

Il primo articolo, è *"The Influence of Embedded Social Media Channels on Travelers' Gratifications, Satisfaction, and Purchase Intentions"* , descrive uno studio, il cui scopo è quello di esaminare l'efficacia dei canali di social media utilizzati per le prenotazioni e la loro influenza sul comportamento dei viaggiatori. Per fare questo studio è stata esaminata la relazione tra gli apprezzamenti fatti dai visitatori, il livello di soddisfazione e l'intenzione di prenotare, mettendo a confronto due tipi di esperienze degli utenti, ovve-

ro quelli che usano social media per la decisione e la prenotazione e quelli che non li usano. I risultati hanno indicato che i viaggiatori che hanno utilizzato social media avevano livelli più elevati di soddisfazione in relazione alle informazioni ricevute, coinvolgimento, interazione e hanno influenzato positivamente il viaggiatore nella prenotazione; questo vuol dire che le informazioni più che esaurienti dei social media, la loro capacità di suggerire e di catturare l'attenzione del viaggiatore, e l'interazione diretta con l'utente moderno influenzano indirettamente il viaggiatore e le sue intenzioni di acquisto. L'articolo dimostra quindi, che le piattaforme online di prenotazione influenzano, positivamente, molto di più che una semplice agenzia di viaggi o un qualsiasi altro metodo di prenotazione offline. I risultati offrono nuove conoscenze riguardanti l'influenza diretta di interazione sociale percepita in relazione alla soddisfazione del viaggiatore e alle intenzioni di acquisto; l'analisi suggerisce che gli hotel dovrebbero incorporare dei canali social media integrati con il loro sito web per aumentare il numero dei clienti. Un'altra analisi fatta e documentata su questo articolo dimostra che il 94% dei principali siti di hotel o catene utilizzano social media, quali Facebook, Twetter e YouTube come mezzo per farsi pubblicità. Di questo 94%, solo il 55% ha uno staff che si dedica solo ed esclusivamente alla gestione dei social network, il restante 45% utilizza personale interno per la loro gestione. Una tecnica utile per la gestione dei social media, sarebbe quella di spingere il cliente, quindi il viaggiare alla "co-creazione" di informazioni coinvolgendolo a commentare a lasciare valutazioni e recensioni [36].

L'altro articolo è invece intitolato *"Travel Planning: Searching for and Booking Hotels on the Internet"* e parla di un sondaggio fatto su un campione di 249 turisti in un hotel a Seattle, Washington; otto su dieci intervistati hanno utilizzato siti web per la ricerca della camera in cui alloggiare, i restanti hanno utilizzato ancora il classico metodo della chiamata per chiedere informazioni sulle disponibilità. Di coloro che hanno cercato la camera online, il 67% ha continuato online anche con l'operazione di prenotazione, il

26% ha chiamato direttamente l'hotel per prenotare la stanza, e il restante 7% si è affidato ad un'agenzia di viaggi.
I risultati della ricerca hanno dimostrato che la motivazione principale per cui molti utenti utilizzino un contatto diretto per la prenotazione, quale appunto la chiamata tramite numero telefonico, sta nel tentativo di negoziare un prezzo inferiore a quello trovato online. Invece per quanto riguarda coloro che prenotano con mezzo elettronico, si è calcolato che il 37% utilizzano il sito ufficiale dell'hotel per prenotare la camera, il 30% utilizzano siti di terze parti, e il 25% utilizzano altri tipi di siti ancora, come aste.
Altri due risultati di questa ricerca sono che: contrariamente di quanto dimostrava lo stesso studio fatto nel 1990, le donne hanno superato gli uomini nelle attività di ricerca di informazioni online; e secondo, coloro che hanno acquistato camere alberghiere online hanno sempre una fascia di età di più bassa, con un numero di notti prenotate sempre maggiori [37].

2.6 Summary

Nell'appendice A sono presenti tabelle che riassumono tutti gli articoli sopra citati, in modo tale da poterli confrontare meglio e analizzare le varie conclusioni in rapporto tra di loro.

Capitolo 3

Progettazione e implementazione

In questa sezione si parlerà per sommi capi del progetto realizzato. Verranno descritte le tecnologie utilizzate, la tipologia di database per la memorizzazione dei dati e le operazioni svolte per ottenere i dati necessari alla ricerca. Infine verrà proposto un elenco delle funzioni e degli algoritmi utilizzati per la creazione della struttura dati finale.

3.1 Specifiche del progetto

Il progetto consiste in un'analisi statistica sull'utilizzo delle piattaforme online di tipo booking e sull'effetto che hanno in relazione all'andamento dell'economia del turismo. In particolare questa analisi utilizza i dati relativi a tutti gli hotel italiani, e alle loro informazioni prese dal database di Expedia.com, preso come esempio di piattaforma di questo tipo.

Questo progetto si pone doversi obiettivi, tutti incentrati sulle recensioni che gli hotel ottengono dai loro clienti. In particolare si vuole studiare l'andamento temporale delle recensioni con il passare degli anni, per determinare

se effettivamente è vero che queste piattaforme stanno prendendo sempre più piede, ovvero se sempre più utenti e hotel utilizzano questo sistema per interagire tra di loro.

In una seconda fase più specifica dello studio, analizzeremo l'utilizzo di Expedia da parte dei turisti andando a calcolare le dimensioni delle recensioni e quali sono i criteri di valutazione. studieremo anche come viene utilizzato Expedia dal management dell'hotel, calcolando quante recensioni ottengono e a quante di queste effettivamente rispondono. Infine scopriremo la distribuzione tra turisti italiani e stranieri nelle diverse aree geografiche.

3.1.1 Expedia.com

Expedia è un sito web di viaggi statunitense lanciato nel 2001 da Expedia, azienda fondata nel 1996 dalla Microsoft, da cui si scorporerà del 1999. Expedia supporta trenta versioni di lingue, per trenta nazioni. Tramite Expedia è possibile prenotare biglietti d'aereo, hotel, automobili a noleggio e crociere, pacchetti vacanza e vari servizi attraverso internet o telefono. Il sito utilizza vari sistemi di prenotazione come Amadeus, il sistema di prenotazioni per voli ed aerei del Sabre [1] , oltre al proprio sistema di prenotazioni. Grazie a questa piattaforma l'utente può inoltre scrivere recensioni su alberghi, appartamenti, b&b, ecc. ma solo se è stata fatta una prenotazione. [21].

La Figura 3.1 mostra la home page di Expedia.com ed è possibile accedervi dall'URL ww.expedia.com .

[1]SABRE, è l'acronimo di Semi-Automated Business Research Environment, è un sistema informatico di prenotazioni utilizzato da compagnie aeree, ferroviarie, catene di hotel ed agenzie di viaggi. Questo sistema è nato negli Stati Uniti negli anni sessanta da uno studio di IBM per l'American Airlines.

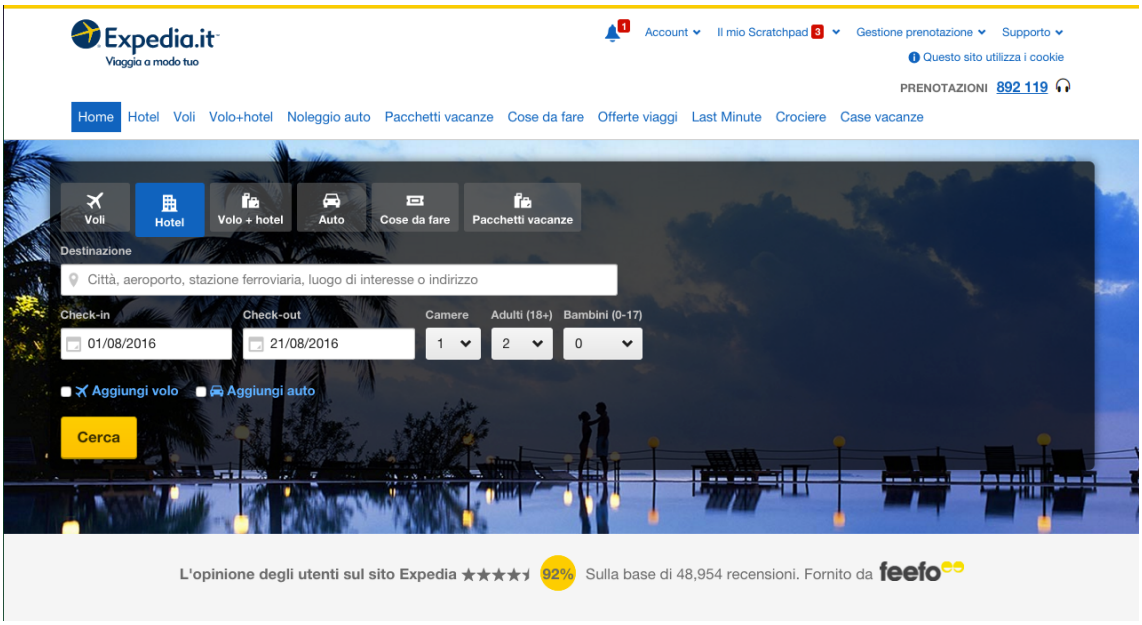

Figura 3.1: Home page di Expedia.com

3.2 Tecnologie utilizzate

Per la creazione del progetto sono state utilizzate diverse tecnologie, ognuna delle quali per uno specifico compito. Infatti è stato utilizzato il PHP per le richieste al server [2] Expedia.com, per il recupero dei dati utili; JavaScript e Node.js per l'iterazione con il database, ovvero con MongoDB; e come editor di testo per scrivere algoritmi e funzioni nei diversi linguaggio è stato utilizzato Komodo Edit 8.

[2]Un server in informatica è un componente o sottosistema informatico di elaborazione e gestione del traffico di informazioni che fornisce, a livello logico e fisico, un qualunque tipo di servizio ad altri componenti che ne fanno richiesta. https://it.wikipedia.org/wiki/Server

3.2.1 PHP

PHP è l'acronimo ricorsivo di Hypertext Processor, ed è un linguaggio di scripting general-purpose [3] , open source molto utilizzato. Questo linguaggio è specialmente indicato per lo sviluppo web e può essere integrato nell'HTML. Ciò che distingue PHP da altri linguaggi di scripting del tipo client-side JavaScript è che il codice viene eseguito nel server, generando HTML che sarà poi inviato al client [4]. La cosa più interessante dell'uso di PHP è che si tratta di un linguaggio estremamente semplice per neofita, ma che, tuttavia, offre molte prestazioni avanzate al programmatore di professione [40].

Nel progetto descritto in questa tesi il PHP è stato utilizzato per effettuare tutte le richieste al server Expedia.com. In particolare è stato usato il metodo *file_get_contents()*, passandogli in input un determinato URL [5] , ricavato dalle API Expedia, per recuperare tutte le informazioni relative a tutti gli hotel italiani e tutte le recensioni che hanno ottenuto. Inoltre è stato utilizzato il metodo *file_put_contents()* per la memorizzazione in locale dell'output json ottenuto, in un file .json per poi poterli importare direttamente sul database.

3.2.2 JavaScript

JavaScript , spesso abbreviato in js, è un linguaggio leggero, interpretato, funzionale e orientato agli oggetti, conosciuto per lo più come linguaggio di

[3] In elettronica e informatica per dispositivi general purpose si intendono dispositivi elettronici che non siano dedicati ad un solo possibile utilizzo, ma dispositivi versatili che di solito caricano componenti software che sono invece soluzioni specifiche per una particolare esigenza.

[4] Un client, in informatica, indica una componente che accede a servizi o alle risorse di un'altra componente detta server. https://it.wikipedia.org/wiki/Client

[5] URL, letteralmente Uniform Resource Locator, è un indirizzo che indica univocamente una risposta su internet. Le "risposte" sono pagine HTML, file, immagini e altro ancora. In base alle operazioni che si intendono fare, verranno utilizzati diversi protocolli come l'http. L'URL è una sottoclasse degli URI (Uniform Resource Identifier, conosciuti fino a poco tempo fa come Universal Resource Identifier). http://www.pc-facile.com/glossario/url/

script per pagine web, ma utilizzato in molti ambienti non browser così come node.js, Apache CouchDB, che è una tipologia di database citata nel primo capitolo, o MongoDB [41].

Infatti, per la memorizzazione dei dati ottenuti è stato utilizzato un particolare database, MongoDB, il quale utilizza il JavaScript come linguaggio per interagire con il database tramite shell [6] . Tramite l'uso di comandi JavaScript da shell mongo è stato possibile effettuare la creazione del database, la creazione delle collezioni e la richiesta di documenti appartenenti a determinate collezioni tramite specifiche query; vedremo in dettaglio la struttura del database nel capitolo di questa sezione dedicato a MongoDB.

Invece per quanto riguarda operazioni più specifiche o funzioni più elaborate per richiedere particolari dati o insiemi di dati è stato utilizzato un file esterno, che utilizza il framework Node.js, e successivamente importato nella mongo shell con il comando *load("/percorso/del/file/nomeFile.js")*.

3.2.3 Node.js

Node.js è un framework per realizzare applicazioni Web in JavaScript, tipicamente usato nella "client-side", ma anche per la scrittura di applicazioni "server-side".
La piattaforma è basata sul JavaScript Engine V8, creato da Google e utilizzato da Chrome e disponibile sulle principali piattaforme, anche se maggiormente performante su sistemi operativi UNIX-like.
La caratteristica principale di Node.js risiede nella possibilità di accedere alle risorse del sistema operativo in modalità event-driven e non sfruttando il classico modello basato su processi thred concorrenti, utilizzato dai classici

[6] La shell è l'interfaccia testuale tramite la quale l'utente può operare ed interagire con il sistema. La shell è un normale programma che interpreta ed esegue i comandi dell'utente, permettendogli di eseguire altri programmi che accedono all'hardware della macchina tramite le chiamate al sistema. http://openskill.info/infobox.php?ID=31

web server, ma sfruttando un tipo di programmazione orientata agli eventi. Questo approccio dovrebbe garantire una certa efficienza delle applicazioni grazie ad una sistema di callback [7] gestito a basso livello a runtime, ovvero quanto viene eseguito il programma [42].

Ai fini del progetto, Node.js è stato utilizzato per una migliore gestione delle operazioni di richiesta dati e interazione con il database. Infatti scrivendo tutte le operazioni in un file Node.js è stato possibile raggruppare più di una query e inserirle all'interno di specifici algoritmi per un più facile recupero di determinati dati, che da shell sarebbe risultato difficile o impossibile fare.
Sono infatti state scritte su un file Node.js gli algoritmi per la pulizia del database, per recuperare tutte le recensioni di tutti gli hotel, per calcolare la somma delle recensioni raggruppate per lingua, per determinare la lunghezza del testo delle recensioni, e molti altri che verranno descritti in dettaglio nei capitoli successivi.

3.2.4 MongoDB

Come spiegato più volte, come tipologia di database per la memorizzazione dei dati è stato utilizzato un particolare tipo di DBMS NoSQL, ovvero MongoDB.

Vediamo quindi più in specifico la struttura del db usato per questo progetto:

Il database prende il nome di *expediaTest* e occupa 0.135GB di memoria, ed è costituito da tre collezioni, chiamate rispettivamente *hotels*, *summary-*

[7] In programmazione, un callback(o, in italiano, richiamo) è, in genere, una funzione, o un blocco di codice, che viene passata come parametro ad un'altra funzione. https://it.wikipedia.org/wiki/Callback

Reviews e *textReviews*.

La collezione *hotels* è a sua volta costituita da 21424 collezioni, che corrispondo ai 21424 hotel italiani recuperati da Expedia; infatti questa collezione rappresenta l'insieme di tutti gli hotel italiani che è stato possibile recuperare tramite la query al server Expedia.

Ogni *hotel*, quindi ogni documento della collezione, viene caratterizzato da:

- "_id": l'id del documento creato automaticamente da MongoDB;
- "HotelID": l'identificativo dell'hotel;
- "Nome": il nome dell'hotel;
- "Location": che a sua volta è una collezione, composta da:
  - "StreetAddress": via in cui è collocato l'hotel;
  - "City": la città o il paese in cui risiede l'hotel;
  - "Province": la provincia del città o del paese dell'hotel;
  - "Country": la nazione, che nel campione di hotel presi in considerazione è sempre "ITA", perché sono stati recuperati solo hotel italiani;
  - "GeoLocation": un ulteriore collezione composta da:
    - "Latitude";
    - "Longitude".
- "Description": che è la descrizione testuale dell'hotel, sono spesso presenti informazioni relative alla posizione geografica e i principali confort che offre la struttura;
- "FeaturedOffer": una collezione composta da:

  - "Price": che a sua volta contiene i seguenti documenti:
    - "TotalRate": con:
      - "Value": prezzo della stanza;
      - "Currency": tipo di moneta a cui si riferisce il prezzo;
  - "CheckInDate": data di arrivo del soggiorno;
  - "LengthOfStay": durata del soggiorno;
  - "DetailsUrl": URL per il collegamento diretto alla pagina Expedia della prenotazione, per un eventuale conferma;

- "DetailsUrl": URL della pagina Expedia con le caratteristiche dell'hotel;
- "StarRating": numero di stelle dell'hotel;
- "ThumbnailUrl": URL della foto dell'hotel;
- "GuestRating": valutazione media ottenuta dalle varie recensioni;
- "GestReviewCount": numero di recensioni associate all'hotel;
- "AmenityList": collezione formata da:
  - "Amenity": che è un array costituito da tutti i confort offerti dall'hotel.

Non tutti gli hotel sono composti da tutti questi attributi, ad esempio l'attributo, nonché collezione, "FeaturedOffer" compare solo in hotel recuperati da una particolare query; invece l'attributo "GuestRating", compare solo se il "GuestReviewCount" è maggiore di 0, infatti dagli hotel senza recensioni non è possibile calcolare una valutazione media.

La collezione *summaryReviews*, è invece composta da 15903 documenti, numero che corrisponde esattamente al numero di hotel recensiti sul totale di 21424 hotel. Ogni documento di questa collezione rappresenta un sommario

dell'insieme delle singole recensioni ottenute da un determinato hotel.

Ogni sommario è caratterizzato da:

- "_id" : identificativo attribuito automaticamente dal sistema, per identificare il documento;
- "reviewSummaryCollection": collezione composta a sua volta dalla collezione:
  - "reviewSummary":
    - "id": identificativo del sommario attribuito da Expedia;
    - "hotelId": identificativo dell'hotel a cui si riferisce il sommario;
    - "totalReviewCnt": numero di recensioni ottenute dall'hotel;
    - "avgOverallRating": valutazione media complessiva;
    - "clanliness": valutazione media attribuita alla pulizia della stanza;
    - "serviceAndStaff": valutazione media attribuita al servizio e allo staff;
    - "roomComfort": valutazione media attribuita ai confort della stanza;
    - "hotelCodition": valutazione media attribuita alle condizioni dell'hotel in generale;
    - "convenienceOfLocation": valutazione media attribuita alla comodità dell'hotel;
    - "neighborhoodSatisfaction": valutazione media attribuita alla località dell'hotel;
    - "roomQuality": valutazione media attribuita alla qualità generale della stanza;
    - "targetedBrand": il marchio che corrisponde sempre ad Expedia;

  - "originSummary": è anche essa una collezione di :
    - "origin";
    - "reviewCnt";
    - "recommendedPercent";
    - "avgOverallRating";
    - "cleanlinses";
    - "serviceAndStaff";
    - "roomComfrot";
    - "hotelCondition";
    - "convenienceOfLocation";
    - "neighborhoodSatisfaction";
    - "valueForMoney";
    - "roomQuality";
    - "categoryCounts";
    - "languageCounts" ;
  - "recinnededPercent": percentuale di recensioni ricommentate dall'hotel;
  - "valueForMoney": rapporto qualità prezzo;
  - "categoryCunts": collezione composta da tutte le categorie di recensioni, con associata la quantità di recensioni riferite a quella categoria;
  - "languageCounts" : collezione composta da tutte le lingue in cui sono stati scritte le recensioni riferite a quell'hotel, associate alla qualità;
  - "featuredReview": collezione composta dalla recensione in primo piano, e comprende tutte le principali caratteristiche della recensione.

Infine, la collezione *textReviews* è composta da 11146 documenti, ognuno dei quali composti da altre collezioni, che rappresentano l'insieme delle recensioni in lingua inglese prese da un determinato hotel. La differenza con

la collezione precedente sta nel fatto che in questa collezione compaiono i riferimenti della singola recensione e non un sommario generale; quindi grazie a questa collezione è possibile recuperare tra le altre cose la valutazione e il testo scritto della singola recensione.

Una singola *review* viene rappresentata in questo modo:

- "tpid";
- "eapid";
- "hotelId": l'identificativo dell'hotel a cui si riferisce;
- "langId";
- "initId";
- "reviewId": l'identificativo della singola recensione;
- "ratingOveral": numero intero che corrisponde alla valutazione che l'utente ha attribuito all'hotel in generale;
- "contentLocale": provenienza dell'utente;
- "userDispalyName": nome dell'utente che ha postato da recensione;
- "brandType";
- "moderationStatus": stato della recensione, se è stata approvata o meno;
- "photos": è un array contenente un insieme di foto che potrebbero arricchire la recensione;
- "contentCodes";
- "title": titolo della recensione;
- "reviewText": testo della recensione;

- "featured": assume il valore booleano true o false e precisa se la recensione è stata messa in primo piano (true) o no (false);
- "recommended": assume il valore booleano true o false e precisa se la recensione è stata ricommentata dall'hotel (true) o no (false);
- "ratingsOnly";
- "userNickname";
- "ratingRoomCleanliness": valutazione attribuita alla pulizia della stanza;
- "ratingHotelCondition": valutazione attribuita alle condizioni generali dell'hotel;
- "ratingServide": valutazione attribuita al servizio dell'hotel;
- "ratingRoomComfort": valutazione attribuita alla stanza in generale;
- "positiveRemarks";
- "negativeRemarks";
- "locationRemarks";
- "lastInitial";
- "userLocation";
- "managementResponses";
- "totalPositiveFeedbacks";
- "totalThanks";
- "reviewSubmissionTime": anno, mese, giorno e ora in cui è stata postata la recensione;
- "incrementalThanks";

- ”reviewerCategories”: collezione che descrive la categoria in cui viene inserita la recensione:
  - ”categoryId”;
  - ”categoryLabel”;
- ”isFlaggable”;
- ”isUnverified”;
- ”isRecommended”: assume i valori di ”YES”, se la recensione ha avuta una risposta da parte dell’hotel, o ”NO”, in caso contrario.

In questa collezione, molti campi potrebbero essere vuoti, come ad esempio l’”userDisplayName” o il capo ”photo” perché non sono capi obbligatori nel momento dell’inserimento della recensione; oppure altri campi come ”positiveRemarks” potrebbero essere vuoti perché ancora nessun’utente ha valutato positivamente quella recensione.

3.3 Expedia API Documentation

Il sito http://hackathon.expedia.com/ mette a disposizione una documentazione abbastanza dettagliata su come interrogare il database Expedia e recuperare i dati di maggior interesse. La Tabella 3.1 riassume brevemente quali sono le APIs [8] che la piattaforma mette a disposizione.

[8]Con application programming interface (in acronimo API, e in italiano interfaccia di programmazione di un’applicazione), in informatica, si indica ogni insieme di procedure disponibili al programmatore, di solito raggruppate a formare un set di strumenti specifici, per l’esecuzione di un determinato compito all’interno di un certo programma. Spesso con tale termine si intendono le librerie software disponibili in un certo linguaggio di programmazione. https://it.wikipedia.org/wiki/Application_programming_interface

Di queste APIs, per la redazione di questa tesi, ne sono state utilizzate solo alcune, ovvero *Geography Search* per il recupero dell'ID delle regioni italiane, *Hotel* Reviews per ottenere le recensioni degli hotel e *Hotel Search* per ottenere le informazioni di tutti gli hotel italiani, che vedremo in modo approfondito più avanti.

Per poter utilizzare queste APIs bisogna ottenere un'API Key, registrandosi al sito e creando un progetto specificando quale delle APIs si è intenzionati ad utilizzare nella propria applicazione.

3.3.1 Sample Use Cases

Sempre lo stesso sito mette anche a disposizione degli esempi di caso d'uso in cui si possono utilizzare le APIs descritte precedentemente. La Tabella 3.2 mostra i principali casi d'uso messi a disposizioni dalle Expedia APIs.

3.3.2 Geography Search

Geography Search è l'API che permette di recuperare le informazioni di una determinata zona geografica; è stata utilizzata in questo progetto per ottenere gli id di tutte le regioni d'Italia, di alcune province e delle principali città turistiche.

Per recuperare le informazioni sulle regioni geografiche è possibile procedere in diversi modi, che si differenziano in base al parametro aggiuntivo inserito in input:

- *bbox* : serve per recuperare tutte le regioni che risiedono all'interno di un rettangolo di area determinato da due coppie di punti geografici, ovvero da due coppie di longitudine e latitudine.
  Esempio di query: http://terminal2.expedia.com/x/geo/features?bbox=-122.453269,37.777363,-122.395935,37.810462&apikey=INSERT_KEY_HERE

- *In.value*: serve per recuperare tutte le regioni geografiche chiamate in un certo modo, infatti basta inserire il nome di una regione, o di una città e verranno recuperare tutte le regioni e tutte le città con quel nome.
  Esempio di query: http://terminal2.expedia.com/x/geo/features?ln.op=cn&ln.value=Naples&type=region&apikey=INSERT_KEY_HERE

- *type*: modificando il type nella precedente query è possibile recuperare solo determinate tipologie di zone geografiche; in quel caso veniva richiesto di recuperare le regioni ("region"), ma è anche possibile richiedere le città ("city"), o un insieme di città, ("multi_city_vicinity"), o addirittura una nazione ("country") e tante altre tipologie.

Queste sono solo alcune delle principali funzioni che permettono di effettuare le Geography Search API.

3.3.3 Hotel Reviews

Le *Hotel Reviews* invece permettono di recuperare tutte le recensioni relative ad un determinato hotel. Utilizzando queste API è possibile ottenere due tipi di risposta: un sommario di tutte le recensioni di un determinato hotel; oppure tutte le recensioni in lingua inglese di un determinato hotel.

Esse si possono utilizzare in due modi:

- Inserendo l'id dell'hotel in input: vengono recuperate tutte le recensioni di quell'hotel, ed è possibile anche specificare altri valori in input:

  - *summary*: accetta due valori, o "true" o "false". Se viene specificato "true" la query ritorna solo il sommario delle recensioni di un determinato hotel; invece se viene specificato false, ritornano tutte le recensioni in lingua inglese di un determinato hotel;

  - *sortBy*: permette di ordinare le recensioni per un determinato parametro, come ad esempio la valutazione;
  - "start: permette di settare il numero della recensione da cui si vuole partire;
  - *items*: numero massimo di recensioni che si vuole ottenere con un'unica richiesta;
  - *categoryFilter*: permette di filtrare la recensione per il tipo, ad esempio "Families".

- Inserendo l'id della recensione: ritorna una singola recensione con tutte le sue caratteristiche.

3.3.4 Hotel Search

Hotel Search API offre la possibilità di cercare all'interno dell'inventario degli hotel disponibili, attraverso diverse metodologie di ricerca, tutte le informazioni accessibili per quell'hotel. Alcune di queste informazioni possono essere ad esempio l'id, il nome, la posizione geografica, la descrizione testuale e tanto altro ancora. In questa richiesta è possibile specificare la data di arrivo e la data di partenza, per ottenere solo gli hotel che hanno disponibilità per quei giorni; altrimenti, se queste date non vengono specificate, viene inserita una data di default che corrisponde alla data corrente, ovvero al giorno in cui viene effettuata la query. Il limite di questa query è che restituisce solamente un massimo di 1000 hotel alla volta; ma è anche possibile specificare un limite inferiore, ma non maggiore.

Come già accennato in questa breve descrizione, sono diverse le metodologie per recuperare gli hotel, esse si differiscono dai parametri inseriti in input:

- *location*: serve per specificare longitudine e latitudine, in modo da trovare un punto geografico, all'interno del quale determinare se è presente

o meno un hotel; è spesso associata a radius, che serve per specificare un determinato raggio, in km, per descrivere il cerchio dentro cui andare a ricercare gli hotel;
Esempio di query: http://terminal2.expedia.com/x/hotels?location=47.6063889,122.3308333&radius=5km&dates=2015-05-19,2015-05-22& apikey=INSERT_KEY_HERE

- *regionids*: serve per specificare l'id della regione all'interno della quale si vogliono recuperare gli hotel.
  Esempio di query: http://terminal2.expedia.com/x/hotels?regionids=178279&dates=2015-05-19,2015-05-22&adults=3&childages=6,9&api key=INSERT_KEY_HERE

- *hotelid*: specificando l'id dell'hotel è possibile recuperalo direttamente.
  Esempio di query: http://terminal2.expedia.com/x/hotels?hotelids=28082,11133&dates=2015-05-19,2015-05-22&apikey=INSERT_KEY_HERE

- *exclude*: parametro aggiuntivo, non obbligatorio che serve per escludere dall'oggetto json di ritorno alcuni parametri che descrivono l'hotel.
  Esempio di query: http://terminal2.expedia.com/x/hotels?location=47.6063889,-122.3308333&radius=5km&dates=2015-05-19,2015-05-22 &exclude=address,description,amenitylist&apikey=INSERT_KEY_HERE

3.4 Dettagli implementativi

Dopo una prima fasi di studio delle Expedia APIs, si è passati subito alla fase implementativa, per la creazione di un algoritmo efficiente da poter utilizzare per il recupero dei dati e il salvataggio degli stessi. Durante questa prima fase tecnica di estrazione dei dati sono sorti parecchi problemi, come il limite massimo di hotel che una singola richiesta restituiva e il recupero di quegli hotel in cui non comparivano informazioni precise sulla loro posizione geografica. Anche durante la seconda fase, ovvero quella di salvataggio, sono sorti alcuni problemi, infatti effettuare un'unica chiamata per il salvataggio

dei dati in locale sovraccaricava il server, mandandolo in time out.

Inizialmente, infatti, si è provato a recuperare tutti gli hotel italiani con un'unica query, inserendo come regionid l'id dell'Italia, ma essendoci più di 22mila hotel in Italia, la query restituiva solo 1000 risultati, troppo pochi per effettuare l'analisi prestabilita. Quindi si è deciso di suddividere l'Italia per le sue 20 regioni, facendo una query per recuperare l'id di ogni regione, e procedere così con lo stessa metodologia di prima. Anche in questo caso, per la maggior parte delle regioni, ci siamo trovati nella stessa situazione con regioni molto grandi, come ad esempio Lombardia, Lazio e Toscana, che superavano di gran lunga i 1000 hotel, addirittura si sono trovati più di 3000 hotel per la Toscana. A loro volta, queste regioni sono state quindi divise per province; a questo punto però anche le grandi città come Milano e Roma superavano i 1000 hotel. Queste province sono quindi state suddivise per longitudine e latitudine, e per città altrimenti sarebbe risultato un lavoro troppo dispendioso a livello di tempo.

Una volta recuperati tutti gli hotel, in questo modo, ci si è accorti che il numero totale degli hotel non rispecchiava esattamente il totale degli hotel per singola regione; ad esempio del Veneto si erano recuperati un totale di 1463 hotel su 1792, in Toscana 2872 su 3207, in Sicilia 2019 su 2144. Questo perché sul database Expedia, molti hotel non hanno indicazioni precise o completamente corrette sulla posizione geografica. Per risolvere il più possibile questo errore si è deciso di inserire, oltre a regioni e province, anche le maggiori città turistiche di ogni regione, come Cervia per l'Emilia Romagna, Cortina d'Ampezzo e la Valle di Cadore per il veneto; oppure le isole come ad esempio le isole di Ischia e Procida e l'Isola di Capri per la Campania e tante altre ancora.

Arrivando così ad un totale di 21424 hotel disponibili il giorno in cui è stata effettuata la query al database Expedia.

3.4.1 Recupero dei regionids

Come è stato descritto precedentemente per determinare gli hotel italiani è indispensabile conoscere i *regionids*, ovvero gli identificativi univoci che permettono di identificare una determinata regione; per recuperarli sono state utilizzate le *Geography* Search APIs, in particolare la richiesta con *In.value*, in modo tale da poter specificare testualmente il nome della regione, della provincia o della città di cui si volesse sapere l'id.

La seguente porzione di codice mostra il semplice script utilizzato per recuperare le varie regioni italiane, nell'esempio compare la Lombardia, ma allo stesso modo è stato è stato fatto per tutte le altre 19 regioni.

```
<?php

    //ALGORITMO PER RECUPERARE GLI ID DELLE REGIONI
    $url = "http://terminal2.expedia.com/x/geo/features?ln.op=cn
    &ln.value=Lombardia&type=region&apikey=6
    UOjgENOLK0BUYA6UwnsNtXZaBJxqIGY";
    $pagina = file_get_contents($url);
    $json_output = json_decode($pagina, true);
    print_r($json_output);

?>
```

Per recuperare l'id delle province invece è stato settato il type=multi_city_vicinity, invece che a *region*. In questo modo, inserendo l'id recuperato, nella query di ricerca degli hotel è stato possibile ottenere tutti gli hotel vicini a quella zona.

```
<?php

    //ALGORITMO PER RECUPERARE GLI ID DELLE REGIONI
```

```
4     $url = "http://terminal2.expedia.com/x/geo/features?ln.op=cn
      &ln.value=Milano&type=multi_city_vicinity&apikey=6
      UOjgENOLK0BUYA6UwnsNtXZaBJxqIGY";
5     $pagina = file_get_contents($url);
6     $json_output = json_decode($pagina, true);
7     print_r($json_output);
8
9 ?>
```

3.4.2 Hotels

Una volta ottenuti tutti i *regionids*, più di 70 per essere precisi, tra regioni italiane, province, maggiori città, isole e laghi, si è passati all'algoritmo per il recupero degli hotel veri e propri e per il loro salvataggio in locale.

Per fare questo si sono inseriti gli id delle regioni in un array [9] , tramite un primo ciclo for [10] è stato iterato tutto l'array, e per ogni id si è fatta la richiesta degli hotel; all'interno di questo primo ciclo è stato fatto una secondo ciclo in modo che per ogni richiesta venisse stampato un singolo hotel alla volta in uno stesso file.json; questo per avere, in un secondo momento, una collezione più ordinata nel database, composta da un documento per ogni hotel. Altrimenti la query avrebbe stampato un blocco unico, quindi un'unica collezione, composto da più hotel, quindi da più collezioni concatenate.
Questo procedimento è stato suddiviso per gruppi di 10-12 *regioids* alla volta, altrimenti con una richiesta unica la connessione con il server si sarebbe interrotta perché eccessivamente pesante. Questo è un esempio di chiamata al server Expedia per il recupero dei primi hotel, e del loro salvataggio su un file.json

[9] Un array è una struttura dati complessa, statica e omogenea; rappresenta un insieme di valori dello stesso tipo.

[10] Si chiama ciclo for quel ciclo che permettere di iterare una serie di elementi, partendo da una posizione di partenza, che può essere per esempio zero, fino ad arrivare ad una posizione massima, e per ogni iterazione il contatore utilizzato per controllare la posizione si incrementa.

```php
<?php

    $regions = array("6048566","6048561","6051441","6049594","
    6049591","6049587","6059421","6035254","6051788","6051795","
    6047212","6051792");


    for($i = 0; $i < count($regions); $i++) {

        //recuro le informazioni degli hotel
        $url = "http://terminal2.expedia.com/x/hotels?regionids=
    $regions[$i]&apikey=6UOjgENOLK0BUYA6UwnsNtXZaBJxqIGY";
        $page = file_get_contents($url);
        $json_output = json_decode($page, true);

        //creo/vado a recuprare un file esterono .json
        $hotelsTestFile = "/Users/paridemartinelli/Desktop/
    hotels.json";
        $contentHotelsFile = file_get_contents($hotelsTestFile);

        //ciglo solo gli hotel del risultato ottenuto
        for($a = 0; $a < count($json_output[HotelInfoList][
    HotelInfo]); $a++ ){

            //inserisco singolarmente gli hotel nel file .json
            $contentHotelsFile .= json_encode($json_output[
    HotelInfoList][HotelInfo][$a]);
            file_put_contents($hotelsTestFile,
    $contentHotelsFile);
        }
    }

    echo "fintio!";

?>
```

Invece per quanto riguarda il recupero degli hotel situati in province molto grandi, dove il numero dei risultati sarebbe stato maggiore di 1000, quali

Napoli, Roma e Milano, si è deciso di suddividere la regione in vare zone, determinando un punto tramite latitudine e longitudine e scegliendo un determinato raggio d'azione. Ad esempio Napoli è stata suddivisa in due zone, la prima avente come coordinate geografiche 40.843372 e 14.355362, e come raggio 30km, la seconda invece ha come coordinate geografiche 40.392337 e 14.992146, e 60km di raggio.

```
1  //RECUPERO GLI ID DEGLI HOTEL CHE SI TROVANO IN UNA
   DETERMINATA POSIZIONE (INDICATA DALLA LONGITUDIE, LATITUDINE
   E APIEZZA IN KM DELL'ARIEA DA CUI ESTRARLI)
2
3  /* Napoli:
4   * * Nord: 40.843372, 14.355362 con 30km h:939
5   * * Sud: 40.392337, 14.992146 con 60km h:883
6  */
7  $url = "http://terminal2.expedia.com/x/hotels?location
   =40.843372,14.355362&radius=30km&apikey=6
   UOjgENOLK0BUYA6UwnsNtXZaBJxqIGY";
8  $page = file_get_contents($url);
9  $json_output = json_decode($page, true);
10
11 //creo/vado a recuprare un file esterono .json
12 $hotelsTestFile = "/Users/paridemartinelli/Desktop/hotels3.
   json";
13 $contentHotelsFile = file_get_contents($hotelsTestFile);
14
15 //ciglo solo gli hotel del risultato ottenuto
16 for($a = 0; $a < count($json_output[HotelInfoList][HotelInfo
   ]); $a++ ){
17
18     //inserisco singolarmente gli hotel nel file .json
19     $contentHotelsFile .= json_encode($json_output[
   HotelInfoList][HotelInfo][$a]);
20     file_put_contents($hotelsTestFile, $contentHotelsFile);
21 }
```

3.4.3 Summary Reviews

Il punto di maggiore interesse di questo progetto sono le recensioni, ovvero l'insieme di commenti e valutazioni postate su Expedia.com dai turisti che hanno alloggiato nei diversi hotel.

Esistono due tipologie di richieste per ottenere informazioni in merito alle recensioni, la prima setta il parametro *summary=true*, e serve per ottenere solo un sommario delle recensioni, la seconda invece setta il parametro *summary=false*, e serve per ottenere le caratteristiche specifiche della singola recensione scritta in lingua inglese.
In entrambi i casi c'è un parametro obbligatorio da inserire per ottenere tutte le recensioni di un determinato hotel, ovvero l'hotelId, che è l'identificativo univoco dell'hotel da cui vogliamo estrarre le recensioni.
Per ottenere tutti gli id dei vari hotel in un unico array, è stato creato un file.txt tramite MongoDB, con tutti e soli gli id degli hotel memorizzati.

```
1 //COMANDO DA UTILIZZARE DA TERMINALE:
2 //mongo expediaTest /Users/paridemartinelli/Desktop/
    myFileMongoDb.js > output2.txt
```

Una volta letto il file.txt e inseriti gli id degli hotel in un array, è stato possibile scorrerli uno ad uno e recuperare tutte le recensioni dei singoli hotel.

```
1
2     $min = 0;
3     $max = 200;
4     for($i = $min; $i <= $max ; $i++){
5
6         //recupero il summary delle recensioni per il count
    delle recensioni suddivisi per lingua
7         $url = "http://terminal2.expedia.com/x/reviews/hotels?
    hotelId=$hotels2[$i]&summary=true&apikey=
    bvht8cwZ80VueNcFTFYczWRhqvw7jQpS";
8         $page = file_get_contents($url);
9         $json_output = json_decode($page, true);
```

```
10
11      if($json_output[reviewSummaryCollection][reviewSummary
    ][0][totalReviewCnt] > 0){
12
13          //creo/vado a recuprare un file esterono .json
14          $reviewsTestFile = "/Users/paridemartinelli/Desktop/
    reviews/reviewsSummary1/reviews".$max.".json";
15          $contentreviewsFile = file_get_contents(
    $reviewsTestFile);
16
17          //inserisco tutte le recensioni in un file .json
18          $contentreviewsFile .= $page;
19          file_put_contents($reviewsTestFile,
    $contentreviewsFile);
20
21      }
22  }
```

Dal codice è possibile notare che all'intero del ciclo for che scorre tutti gli id degli hotel, è presente un if [11] , che controlla se il parametro *totalReviewCnt* dell'hotel corrispondente all'id è maggiore di zero, in modo tale da prendere in considerazione solo gli hotel con recensioni.

Anche in questo caso, come per il recupero degli hotel il ciclo for è stato limitato a 200 id alla volta, sempre per non sovraccaricare il sistema.

Tutto questo per quanto riguarda la prima tipologia di richiesta delle recensioni.

[11] L'if è una condizione che permette di stabilire se la funzione che c'è al suo interno si può eseguire o no. Infatti vengono inserite come parametri dell'if determinate condizioni che devono essere verificate prima di poterci entrare e quindi prima di eseguire la porzione di codice che c'è al suo interno.

3.4.4 Reviews

Per quanto riguarda la seconda tipologia di richiesta delle recensioni, grazie alla quale è stato possibile il recupero di informazioni più dettagliate, quali testo, data e ora di insermento, e l'utente, si è proceduto in modo del tutto analogo.

```
<?php

//ALGORITMO:
/* 1.Recuperare l'id dal file output2.txt
 * 2.Verificare il numero totale di recensioni in lingua inglese
    ($totRev)
 * 3.Utilizzare la query per il recupero delle recensioni
   specificando $totRev
    http://terminal2.expedia.com/x/reviews/hotels?hotelId
   =1406673&summary=false&sortBy=DATEDESCWITHLANGBUCKETS&items=
   $totRev&apikey=bvht8cwZ80VueNcFTFYczWRhqvw7jQpS
 *4.Salvare tutto in un file esterno (come al solito)
*/


    //1. Recupero gli id degli hotel
    $filename = "/Users/paridemartinelli/output2.txt";
    $handle = fopen($filename, "r");
    $contents = fread($handle, filesize($filename));
    fclose($handle);
    $hotels = explode(",",$contents);

    //2. Per ogni hotel recupero il numero totale di recensioni
   e salvo il valore nella variabile $totRev
    $min=4401;
    $max=4600;

    for($i = $min; $i<$max; $i++){

```

```
25          $url = "http://terminal2.expedia.com/x/reviews/hotels?
    hotelId=$hotels[$i]&summary=true&apikey=
    bvht8cwZ80VueNcFTFYczWRhqvw7jQpS";
26          $page = file_get_contents($url);
27          $json_output = json_decode($page, true);
28          $totRev = $json_output["reviewSummaryCollection"]["
    reviewSummary"][0]["originSummary"][0]["languageCounts"]["en"
    ];
29          if($totRev > 0){
30              //print($totRev);
31
32              //3. Recupero le recesioni di quell'hotel, inserendo
     il valore appena trovato nell'items
33              $url1 = "http://terminal2.expedia.com/x/reviews/
    hotels?hotelId=$hotels[$i]&summary=false&sortBy=
    DATEDESCWITHLANGBUCKETS&items=$totRev&apikey=
    bvht8cwZ80VueNcFTFYczWRhqvw7jQpS";
34              $page1 = file_get_contents($url1);
35
36              //4.Salvo le recensioni trovate in un file json
37              //creo/vado a recuprare un file esterono .json
38              $reviewsTestFile = "/Users/paridemartinelli/Desktop/
    reviews/reviewsWithText/textRev".$max.".json";
39              $contentreviewsFile = file_get_contents(
    $reviewsTestFile);
40
41              //inserisco tutte le recensioni in un file .json
42              $contentreviewsFile .= $page1;
43              file_put_contents($reviewsTestFile,
    $contentreviewsFile);
44
45
46          }
47      }
48      print("finito!");
49
50  ?>
```

In questo caso compare una variabile, *$totRev*, che va a verificare il numero delle recensioni scritte in lingua inglese, così da poter considerare solo gli hotel in cui questo parametro è maggiore di 0.

3.4.5 ExpediaTest db

Una volta recuperati tutti i dati in file esterni .json si è proceduto importandoli nelle rispettive collezioni.

```
//COMANDO PER IMPORTARE I FILE .json DA UTILIZZARE SU TERMINALE
//mongoimport --db "nomeDB" --collection "nomeCollezione" --file
    "percorso/del/file/nomeFile.json
```

Una volta popolate le collezioni, rispettivamente *hotels* con gli hotel, *summaryReviews* con i sommari delle recensioni e *textReviews* con le recensioni scritte in inglese, è stata fatta una pulizia delle occorrenze multiple, come mostra la porzione di codice seguente, estratta come esempio, per la pulizia della collezione *hotels*:

```
//ALGORITMO PER LA PULIZZIA DELLE OCCORRENZE MULTIPLE
//hotel
cursor = db.hotels.distinct("HotelID");
print(cursor.length)
for (i = 0 ; i < cursor.length ; i ++) {

    hotelCount = db.hotels.find({"HotelID":cursor[i]}).count()
    hotelToDelete = db.hotels.find({"HotelID":cursor[i]}, {"_id"
    :1}).map( function(u) { return u._id; } )
    if (hotelCount > 1) {

        for (a = 1 ; a < hotelToDelete.length; a ++) {

            db.hotels.remove( {"_id":hotelToDelete[a]} )
        }
    }
}
```

La pulizia delle collezioni si è resa necessaria poichè le query di estrazione degli hotel per posizione geografica con latitudine, longitudine e raggio (ad esempio per le città con oltre 1000 hotel) estraevano, in parte, stessi risultati. Queste estrazioni con possibili occorrenze doppie sono state effettuate appositamente, aumentando il raggio, in modo tale che due o più zone si intersecassero per non lasciare porzioni di città scoperte.

Un'ulteriore modifica alla collezione hotels è stata fatta nel parametro *GuestReviewCount*; infatti questo parametro era inserito come stringa, ma per effettuare tutte le operazioni necessarie è stato opportuno trasformarlo in intero.

```
//ALGORITMO PER LA CONVERSIONE DI STRINGE IN INTERI
//Usato per il numero delle recensioni totali:
var convert = function(document){
var intValue = parseInt(document.GuestReviewCount, 10);
  db.hotels.update(
    {_id:document._id},
    {$set: {"GuestReviewCount": intValue}}
  );
}
```

API	Summary
Car Search	Fornisce informazioni sul noleggio auto che Expedia mette a disposizione. L'utente può cercare il noleggio utilizzando il codice pickup location IATA con le date di prelievo e rilascio. Le auto restituite come risposta possono essere ordinate in base al prezzo; l'utente può limitare il numero di automobili e filtrare i risultati in basi ai fornitori.
Flight Search	Serve per ottenere disponibilità e prezzi dei voli per un determinato punto di partenza, destinazione e date del viaggio.
Flights Overview	Utilizzata per ottenere diversi set di risultati per un dato criterio di ricerca, sempre per ottenere i voli. In particolare restituisce il volo meno costoso, dato in input un calendario di date.
Flights Prices Trends And Predicions	Permette di recuperare l'andamento dei vecchi prezzi e di calcolare una previsione dei prezzi futuri dei voli.
Geography Search	Ricerca di determinate locazioni geografiche basata sulle regioni, sugli hotel o sui punti di interesse.
Hotel Reviews	Recupera tutte le recensioni verificate per un dato albergo. Tutte le recensioni vengono scritte da clienti Expedia che hanno soggiornato presso l'hotel che desiderano recensire.
Hotel Search	Offre la possibilità di cercare un hotel in diversi modi, ad esempio inserendo una posizione geografica. Di ogni hotel vengono restituite le informazioni principali, come l'id, il nome e l'indirizzo.
Natural Language Hotel Search	Utilizza il linguaggio naturale per ottenere determinati insiemi di hotel. Inserendo, infatti una stringa di testo, restituisce gli hotel che tra le informazioni possiedono i parametri ricercati.
Package Search	Questa query permette di trovare delle offerte comprese di volo e hotel presenti su Expedia.
Suggestions and Resoluzions API	Permette di aiutare l'utente nella ricerca dei voli e degli hotel. Recupera un numero limitato di soluzioni suggerite da Expedia per la prenotazione della vacanza.
Things to do!	Fornisce l'elenco delle attività che è possibile svolgere in una determinata area di interesse.
Travel Trends	Consente ai viaggiatori di scoprire quale sono le mete più ambite e gli hotel più popolari dove andare in vacanza.
Unreal Deals	Trova offerte speciali per pacchetti compresi di voli e hotel.

Tabella 3.1: Expedia APIs.

Caso d'uso	Descrizione
How to get Expedia RegionIDs?	Tutte le regioni geografiche presenti in Expedia sono identificate da un ID, chiamato RegionID. Molte delle API Expedia richiedono come input, per ottenere determinati valori, questo parametro. Per ottenerlo è sufficiente utilizzare ad esempio la seguente query: http://terminal2.expedia.com/x/geo/features?ln.op=cn&ln.value=Milan&type=region&apikey={INSERT_KEY_HERE} Per ottenere gli HotelIDs, ovvero gli identificativi degli hotel, ci possono essere diversi modi:
How to get Expedia HotelIDs?	1. Inserendo come parametro di ricerca la longitudine e la latitudine: http://terminal2.expedia.com/x/hotels?location=45.494928,9.338065&radius=25km&apikey={INSERT_KEY_HERE} 2. Inserendo un RegionIDs: http://terminal2.expedia.com/x/hotels?regionids=6048566&apikey={INSERT_KEY_HERE}
Find Package Deals For A Set of Hotels	1. Ottenere una lista di hotel per una determinata area; 2. Utilizzare la seguente query per ottenere le offerte Expedia degli hotel appena trovati: http://terminal2.expedia.com/x/deals/hotels?adultCount=2&hotelIds=554676,14917,1833&checkInDate=2015-04-15&checkOutDate=2015-04-29&apikey={INSERT_KEY_HERE}
Find Hotels Near A Point Of Interest (POI)	1. Cercare un punto di interesse, come ad esempio "Seattle Center": http://terminal2.expedia.com/x/geo/features?ln.op=cn&ln.value=seattle%20center&limit=5&apikey={INSERT_KEY_HERE} 2.Utilizzare gli id delle regioni geografiche trovati con la richiesta precedente e cercare gli hotel all'interno di quella zona tramite la seguente query: http://terminal2.expedia.com/x/geo/features/319476405484128016/features?within=1km&type=hotel&apikey={INSERT_KEY_HERE} 3. Ora che si hanno gli id degli hotel utilizzare la prossima query per trovare le migliori offerte degli hotel: http://terminal2.expedia.com/x/deals/hotels?adultCount=2&hotelIds=20194&checkInDate=2015-04-15&checkOutDate=2015-04-30&apikey={INSERT_KEY_HERE}
Search Using Natural Language	1. E' possibile utilizzare del testo in linguaggio naturale come parametro di richiesta, come ad esempio "My wife and I need a deal in Las Vegas this weekend" per ottenere gli hotel: http://terminal2.expedia.com/x/nlp/results?q=my%20wife%20and%20i%20need%20a%20deal%20in%20las%20vegas%20this%20weekend&apikey={INSERT_KEY_HERE}

Tabella 3.2: Sample Use Cases.

Capitolo 4

Analisi dei dati

Nel quarto ed ultimo capito vengono analizzati i dati estratti. In particolare verrà spiegata la metodologia di estrazione, la rappresentazione in forma tabellare e la rappresentazione in forma grafica; infine, questi risultati, verranno analizzati e commentati al fine di determinare conclusioni sulla piattaforma Expedia, in merito al suo utilizzo e al rapporto che hanno i suoi utenti con le recensioni.

4.1 Prima fase: la diffusione dei social media

Un buon 90% degli articoli trattati nel capitolo precedente sosteneva il progressivo aumento di popolarità delle piattaforme di tipo booking in generale. La prima fase di questa studio si pone appunto l'obiettivo di verificare se questa affermazione con i nostri dati; ovvero si vuole verificare se i social media utilizzati per la prenotazione di voli e hotel, e il loro utilizzo per lo scambio di opinioni tra turisti e albergatori si stia diffondendo sempre di più.

Per effettuare questo tipo di analisi si sono studiate le recensioni postate dai turisti e si è studiato il loro andamento con il passare degli anni. Come campione su cui basare la ricerca statistica si sono prese tutte le recensioni

di hotel italiani postate sulla piattaforma Expedia, presa come esempio dalla vasta gamma di piattaforme di booking.

Per la precisone l'analisi si è basata su un campione di 21424 hotel italiani (oltre i due terzi dei circa 33000 hotel italiani in base ai dati ISTAT, l'Istituto di statistica Italiano), dei quali solo 5293 non hanno recensioni, quindi solo il 24,7% non sono recensiti. Dei restanti 16185 si sono trovate un totale di 897806 recensioni.

```
cursor = db.hotels.find({},{"GuestReviewCount":1}).map( function
    (u) { return u.GuestReviewCount; } );
sum = 0;
for(i = 0; i < cursor.length; i++){
    sum = sum + cursor[i];
}
print(sum);
```

Di queste 897806 recensioni, ricavate dalla somma del numero di *"GuestReviewCount"* della collezione *hotels*, si sono calcolate un totale di 458719 recensioni scritte, ovvero comprese di testo e valutazione, andando a sommare, questa volta, i *"totalReviewCnt"* della collezione *summaryReviews*, perché solo da li era possibile ricavare questo valore:

```
cursor = db.summaryReviews.find({},{"reviewSummaryCollection.
    reviewSummary.totalReviewCnt":1})
count = db.summaryReviews.find({},{"reviewSummaryCollection.
    reviewSummary.totalReviewCnt":1}).count()
sum = 0;
for(i = 0; i < count; i++){
    sum = sum + cursor[i]["reviewSummaryCollection"]["
    reviewSummary"][0]["totalReviewCnt"];
}
print(sum);
```

Quindi si può dedurre che le restanti 439087 abbiano solamente una valutazione numerica che indica il grado di soddisfazione dell'utente.

Ora, ritornando all'obiettivo principale di questa prima fase di ricerca, ovvero determinare l'andamento del volume delle recensioni al passare del tempo, è necessario andare a recuperare il parametro *"reviewSubmissionTime"* della collezione *textReviews*, l'unica collezione che ci dia riferimenti temporali sulla singola recensione. Purtroppo, come già spiegato in precedenza, questa collezione contiene solamente le recensioni di lingua inglese. Ma è stato calcolato, dalla collezione *summaryReviews* che il totale delle recensioni scritte in lingua inglese sono 216929, quindi coprono il 47,3% del totale, una somma più che significativa per determinare conclusioni che potessero essere valide anche per quanto riguarda l'andamento totale di tutte le recensioni.

La Tabella 4.1 mostra un riassunto dei dati appena descritti.

TOTALI	#
Totale recensioni	897806
Totale recensioni scritte	458719
Totale recensioni solo valutazioni	439087
Totale recensioni in lingua inglese	216929

Tabella 4.1: Totali recensioni.

Una volta giunti alla conclusione che è possibile calcolare l'andamento generale delle recensioni, basandoci sul totale delle recensioni in lingua inglese, non ci resta che suddividere queste recensioni per anno.

```
1 //Scorro tutte le recensioni
2 for (i = 0; i < count; i ++) {
3
```

```
4     numeroRecensioni = cursor[i]["reviewDetails"]["
      numberOfReviewsInThisPage"];
5     //recuper l'anno e il mese
6     for (c = 0; c < numeroRecensioni; c++) {
7
8         //recuper l'intera data
9         data = cursor[i]["reviewDetails"]["reviewCollection"]["
      review"][c]["reviewSubmissionTime"];
10        arrayData = data.split("-");
11        annoData = arrayData[0].replace("\"","");
12        meseData = arrayData[1];
13
14        if (annoData == year) {
15
16            tot ++;
17
18            if (meseData == 01) {
19               gennaio ++;
20            } else if (meseData == 02) {
21                febbraio ++;
22            } else if (meseData == 03) {
23                marzo ++;
24            } else if (meseData == 04) {
25                aprile ++;
26            } else if (meseData == 05) {
27                maggio ++;
28            } else if (meseData == 06) {
29                giugno ++;
30            } else if (meseData == 07) {
31                luglio ++;
32            } else if (meseData == 08) {
33                agosto ++;
34            } else if (meseData == 09) {
35                settembre ++;
36            } else if (meseData == 10) {
37                ottobre ++;
38            } else if (meseData == 11) {
39                novembre ++;
```

```
40                } else if (meseData == 12) {
41                    dicembre ++;
42                }
43            }
44        }
45
46    }
```

Questa porzione di codice ci mostra che l'algoritmo non solo calcola il totale delle recensioni per anno, ma determina anche il numero delle recensioni ottenute al mese. Infatti, dopo aver estratto la *"reviewSubmissionTime"* dalla collezione *textReviews*, è stato possibile, incrementando una variabile per ogni mese dell'anno, andare a calcolare il totale relativo delle recensioni postate dagli utenti per ogni singolo mese dell'anno. La variabile "year" rappresenta l'anno da prendere in considerazione, e per questo studio si sono presi in esame le recensioni che vanno dal 2010 al 2015. Le Tabella 4.2 e 4.3 ci dimostrano i risultati ottenuti.

ANNO	GENNAIO	FEBBRAIO	MARZO	APRILE	MAGGIO	GIUGNO	LUGLIO	AGOSTO	SETTEMBRE	OTTOBRE	NOVEMBRE	DICEMBRE
2010	422	401	586	836	746	1138	1100	1199	1313	1378	836	424
2011	407	368	655	1025	1145	1336	1163	968	1619	1673	1071	461
2012	733	780	1078	1746	1927	2455	2533	2139	2686	2557	1501	843
2013	897	875	1546	2292	3374	3897	4050	3943	3788	4345	2392	1142
2014	1398	1172	1800	3345	4435	4852	5111	4528	4867	4881	2662	1451
2015	1637	1581	2503	4014	5324	6130	6024	5894	6457	5539	2830	1366

Tabella 4.2: Andamento mensile delle recensioni.

Si è deciso di calcolare anche l'andamento mensile delle recensioni, per capire, non solo se è vero che con l'aumentare degli anni il numero delle recensioni sta aumentando, ma anche per capire quali sono i mesi in cui gli utenti sono più attivi.

ANNO	TOTALE RECENSIONI
2010	20523
2011	23338
2012	20978
2013	32541
2014	40502
2015	49299

Tabella 4.3: Andamento annuale delle recensioni.

Il Figura 4.1 mostra una sintesi dei risultati ottenuti, dal quale è facile estrarne le conclusioni in merito alle domande precedentemente poste.

Dalla Figura 4.1 è possibile notare, infatti, che con il passare degli anni il numero delle recensioni è aumentato notevolmente, partendo da un totale di 20523 recensioni nel 2010, come ci suggerisce la Tabella 4.3, e arrivando ad un totale di 49299 nel 2015.

Inoltre da questo grafico è possibile affermare che l'attività degli utenti è abbastanza costante, infatti per tutti gli anni analizzati, i mesi come Gennaio, Febbraio e Dicembre, sono i mesi in cui gli utenti non postano commenti; questo perché a Gennaio e a Febbraio, la maggior parte delle persone lavora, e dicembre perché molti utenti sono in vacanza, quindi solitamente la recensione viene fatta alla fine del periodo di soggiorno dell'hotel.
Ma è anche possibile notare i mesi in cui vi è il picco dell'attività dei turisti, infatti nei mesi di Settembre e Ottobre, quando l'utente è appena tornato dalle ferie di Agosto, si sono calcolati fino a 6457 recensioni postate.

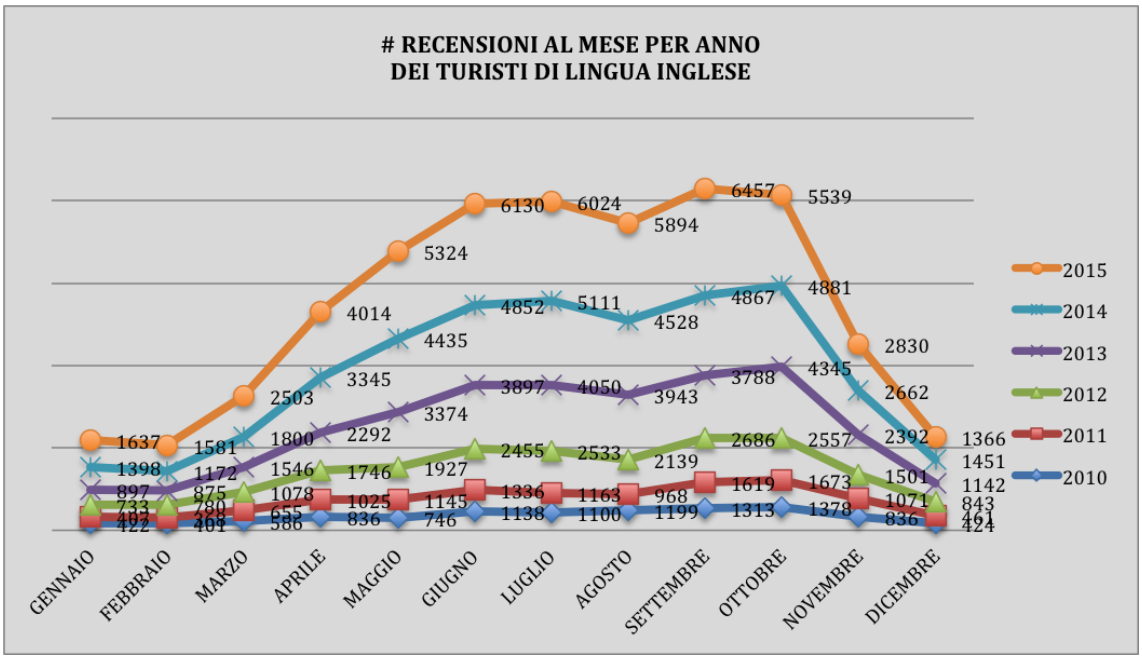

Figura 4.1: Andamento temporale delle recensioni.

4.2 Seconda fase: L'utilizzo dei social media

Dopo aver verificato l'aumento di popolarità dei social media negli anni, passiamo ad una seconda fase di studio, ovvero quella che ci permette di determinare come vengono utilizzati i social media dai suoi utenti, turisti ed albergatori.
Per quanto riguarda lo studio dell'utilizzo di Expedia, come campione per rappresentare l'insieme di tutti i social media, da parte dei turisti è stata calcolata in modo indiretto la loro attività di recensire gli hotel; infatti l'attività dei turisti sui social media è possibile determinarla dal numero di recensioni ottenute dai singoli hotel.

Sempre per quanto riguarda lo studio dei turisti su Expedia, è possibile determinare, attraverso la dimensione delle recensioni, "come" viene utilizzata la piattaforma.

E infine, studiando le valutazioni, attribuite in ogni singola recensione, agli hotel, sono state analizzate le loro preferenze, andando a vedere se c'è un relazione tra numero delle stelle e valutazione attribuita.

Dopo aver studiato il rapporto che hanno i turisti con Expedia, si è cercato di determinate che rapporto hanno invece gli hotel, in particolar modo gli albergatori, con la piattaforma; andando a studiare, in un primo momento, la percentuale media delle recensioni a cui rispondono e, in un secondo momento, si è cercato di stabilire a quale tipologia di recensioni rispondono maggiormente.

4.2.1 L'utilizzo da parte dei turisti

La prima domanda a cui vogliamo dare una risposta è: Quanto viene utilizzato Expedia da parte dei turisti?
Per dare una risposta a questa domanda, si è calcolato quindi il numero delle recensioni per ogni singolo hotel, in particolare si sono calcolati quanti hotel hanno meno di "n" recensioni, attraverso seguente funzione:

```
1 //db.hotels.find({"GuestReviewCount":{$lt:5}}).count()
```

La quale mostra, come esempio, la query utilizzata per calcolare il numero di hotel che hanno meno di 5 recensioni. Utilizzando la stessa query con "n" che va da 5 a 4000 è possibile determinare i seguenti valori:

Dalla Tabella 4.4 è possibile dimostrare che già più del 50% degli hotel ha meno di 10 recensioni, questo vuol dire che sono pochi gli hotel con molte recensioni, per la precisione quelli che superano le 100 recensioni sono solo circa il 10%.

#RECENSIONI (ricavate con "meno di")	#HOTEL	%HOTEL
5	9399	43,871
10	11997	55,998
15	13540	63,200
20	14590	68,101
30	16085	75,079
40	17009	79,392
50	17697	82,604
100	19323	90,193
200	20363	95,048
300	20771	96,952
400	20989	97,970
500	21137	98,660
1000	21367	99,734
2000	21419	99,977
3000	21423	99,995
4000	21424	100,000

Tabella 4.4: % hotel per numero recensioni (con meno di "n")

Queste conclusioni si possono leggere meglio su una curva CDF [1] , in cui sull'asse delle x compaiono i vari valori assunti da "n" e sulla y le percentuali degli hotel che hanno ottenuto meno di "n" recensioni.

Sull'asse delle x compaiono solo i valori da 0 a 500, per cercare di focalizzare l'attenzione sulla prima parte del grafico, questo perché dopo le 250 recensioni i valori crescono più o meno costanti fino ad arrivare al 100%.
Invece dalla prima parte del grafico si evince che la maggior parte degli hotel ha poche recensioni, infatti il grafico sale vertiginosamente da 0 a 50 recensioni, questo ad indicare che la maggior parte degli hotel ha meno di 50 recensioni, successivamente curva fino a 100 recensioni e dopodiché inizia a salire costantemente.

[1] CDF è l'acronimo di Cumulative Distribution Function, in italiano funzione di ripartizione, o funzione cumulati. In statistica è una funzione di variabile reale che racchiude le informazioni su un fenomeno (un insieme di dati, un evento casuale) riguardanti la sua presenza o la sua distribuzione prima o dopo un certo punto. https://it.wikipedia.org/wiki/Funzione_di_ripartizione

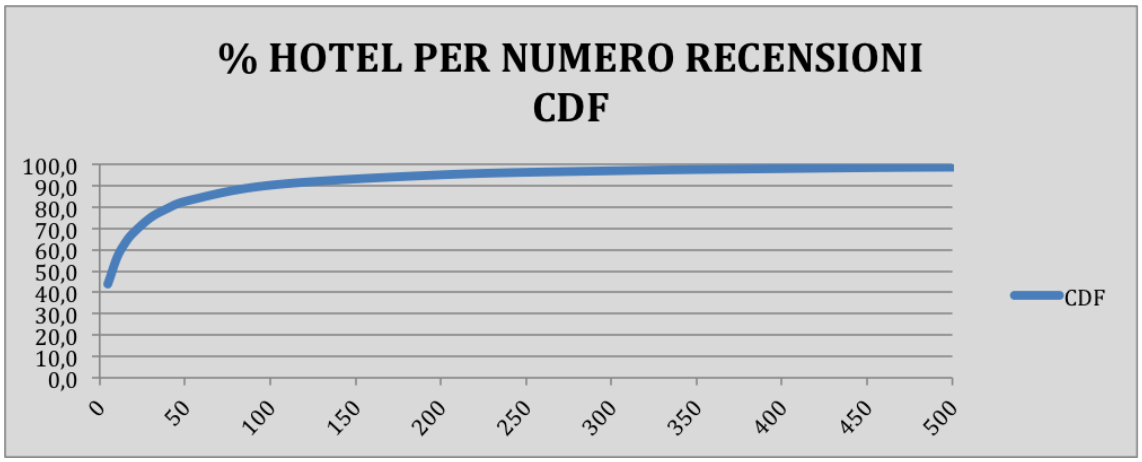

Figura 4.2: % hotel per numero recensioni.

Questo ci dice due cose in relazione all'obiettivo che volevamo portare a termine: la prima è che i turisti che popolano le regioni italiane, non recensiscono tutti gli hotel allo stesso modo, ma solamente determinati hotel; e la seconda è che quei pochi hotel che ricevono recensioni ne ricevono tantissime. Questo forse a confermare la tesi, letta nell'articolo *"Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach"* che sostiene che maggiore è il numero delle recensioni per un certo hotel, e maggiore è la possibilità che un altro utente aggiunga la sua esperienza di soggiorno.

La seconda domanda che ci poniamo, non è più quanto, ma come viene utilizzato Expedia dai turisti?
Per cercare di capire come utilizzano i turisti questo tipo di piattaforma abbiamo in un primo momento messo a confronto le recensioni aventi testo scritto con le recensioni aventi solo valutazioni numeriche.

Già nel capitolo precedente si era parlato del numero di recensioni con testo scritto, le quali raggiungevano un numero pari a 458719 su un totale di 897806 recensioni. Questo dimostra che esattamente il 51% degli utenti preferisce descrivere testualmente le esperienze vissute, e poco meno della

metà, invece, da più importanza alle valutazioni.

Successivamente invece si è posto il problema di determinare la dimensione delle recensioni postate dagli utenti, per capire più o meno la lunghezza media di una recensione.
Per ottenere questi risultati è necessario recuperare il valore *"reviewText"* dalla collezione *textReviews*.

Prima di passare alla fase di calcolo è opportuno fare due piccoli accorgimenti: il primo è che delle 216929 recensioni, in lingua inglese, di cui si parlava nella prima fase di ricerca, è stato possibile recuperarne solo 208941; ma questo non influisce sullo studio, in quanto vuol dire che se ne sono perse solamente 7988, quindi solo il 3,7%. Il secondo accorgimento è che delle 208941 recensioni rimanenti, se ne sono trovate 1948 con una lunghezza pari a zero; questo potrebbe accadere per diversi motivi, come ad esempio un errore da parte dell'utente nella scrittura della recensione inserendo uno spazio vuoto che il database Expedia non riesce ad interpretarlo nel modo corretto, andando quindi a memorizzare la recensione con spazio vuoto nell'insieme delle recensioni aventi testo scritto, quindi con data, ora e tutti i dati che le caratterizzano.

Possiamo ora andare a calcolare le dimensioni delle recensioni.

```
1 for(index = 0; index < count; index ++){
2
3     for (i = 0; i < cursor[index]["reviewDetails"]["
    numberOfReviewsInThisPage"]; i ++){
4
5         rev = cursor[index]["reviewDetails"]["reviewCollection"
    ]["review"][i]["reviewText"];
6
7         rev = rev.replace(/\n/g,'');
8         rev = rev.replace(/\r/g, '');
```

```
        revWithoutBlank = rev.replace(/ /g,'');

        if (revWithoutBlank.length < n) {
            arraySize[indexSize] = parola.length;
            arraySizeWithoutBlank[indexSize] = parolaSenzaSpazi.
    length;
            indexSize ++ ;
        }

    }
}
print(indexSize);
```

Come è possibile vedere dal codice, per determinare le dimensioni delle recensioni è stata semplicemente calcolata la lunghezza del valore *"reviewText"*. È stata calcolata anche la lunghezza della recensione senza spazi, questo perché spesso durante il salvataggio del commento scritto, sul database Expedia, "spazi" e "a capo" non venivano tradotti correttamente, lasciando qualche "\n" o "\r" in più, che andavano a incidere sulla lunghezza finale della recensione.
I risultati così ottenuti vengono elencati nella Tabella 4.5.

Dopo aver calcolato i suddetti valori, è stato possibile determinare anche quale fosse la dimensione massima e la dimensione minima delle recensioni.

```
//recupero dimensione massima e dimensione minima
arraySize.sort(function(a, b){return b-a});
arraySizeWithoutBlack.sort(function(a, b){return b-a});
print(arraySize[0])
print(arraySizeWithoutBlack[0]);
```

Ottenendo i valori ripostati nella Tabella 4.6.

Dalla Tabella 4.6 notiamo che esistono recensioni scritte con lunghezza uguale a uno, per essere più precisi ne sono state trovate 45. Il testo di queste

(MENO DI) LENGTH	#RECENSIONI CON SPAZZI	% RECENSIONI CON SPAZZI	#RECENSIONI SENZA SPAZZI	% RECENSIONI SENZA SPAZZI
5	2040	0,98%	2042	0,98%
10	2164	1,04%	2205	1,06%
15	2342	1,12%	2427	1,16%
20	2525	1,21%	2663	1,27%
30	3006	1,44%	3274	1,57%
40	3550	1,70%	4041	1,93%
50	4293	2,05%	8562	4,10%
60	9057	4,33%	14678	7,02%
70	14109	6,75%	20724	9,92%
80	19174	9,18%	26699	12,78%
90	24157	11,56%	32338	15,48%
100	28827	13,80%	37834	18,11%
200	71237	34,09%	86783	41,53%
300	106370	50,91%	125511	60,07%
400	134343	64,30%	152341	72,91%
500	154455	73,92%	170184	81,45%
1000	196782	94,18%	202586	96,96%
1500	208811	99,94%	208940	100,00%
2000	208941	100,00%	208941	100,00%

Tabella 4.5: Lunghezza delle recensioni.

recensioni corrisponde per la maggior parte dei casi a "." o a lettere singole come "a" o "b". Ma esistono anche 19 recensioni, con testo di lunghezza due, delle quali la maggior parte sono "ok" e una di esse è uno smile ":)". Tutte queste recensioni, ad eccezione di casi particolari, come lo smile, potrebbero rientrare nella categoria delle recensioni con solo valutazione, poiché il loro testo non è utile ai fini di nessuna analisi. Questo fino alle recensioni di lunghezza quattro, in cui compaiono i primi "good", "nice" e "fine".

Dalla Tabella 4.5 invece è possibile notare che anche questa volta, i dati sono stati raggruppati secondo il criterio "a meno" di, in modo tale da riuscire a capire quante recensioni hanno una lunghezza minore di "n". Dove "n" va da 5 fino a 2000, preso come riferimento per superare quello che è il valore massimo di 1562.
Come per il campione di dati precedenti, anche questi sono stati rappresen-

	LENGTH MIN	LENGTH MAX
CON SPAZI	1	1562
SENZA SPAZI	1	1500

Tabella 4.6: Lunghezza massima e minima delle recensioni.

tati su una curva CDF.

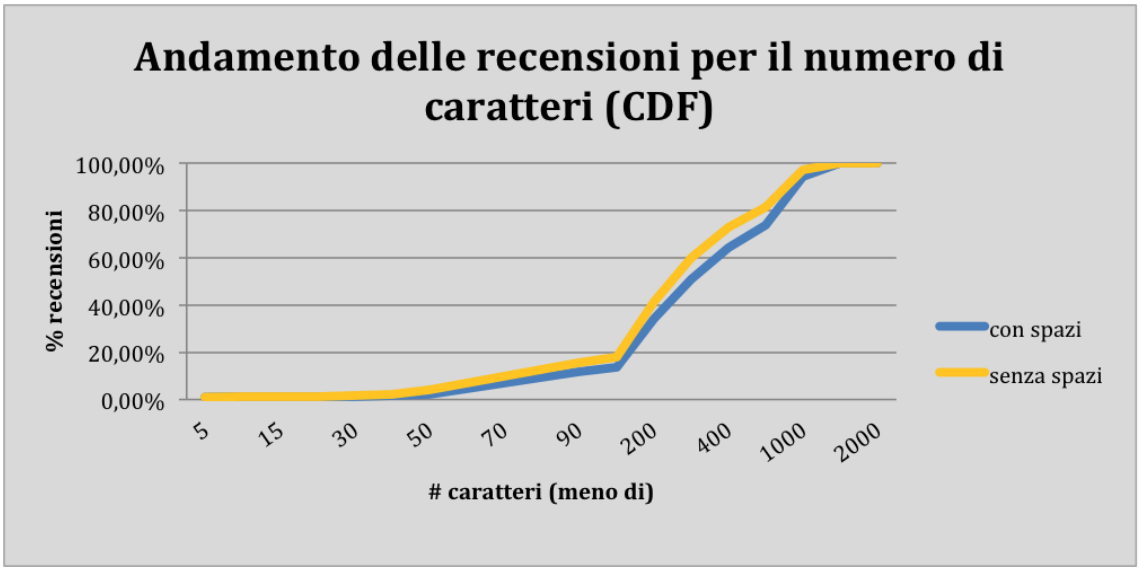

Figura 4.3: Lunghezza caratteri.

Da questo grafico si evince che circa il 60% degli utenti che scrive recensioni, utilizza un testo di lunghezza minore di 300 caratteri. Per intenderci 300 caratteri sono poco più di quattro righe di testo, come ad esempio *"Got booked into a deluxe room with twin beds despite a specific request for a queen or king bed. Staff attitude implied it was not their problem, but did offer an upgrade to a suite for 20 euros. With rooms going for 300 euros a night, it is laughable for them to nickel and dime a guest over mix up."*.
Un'altra cosa che è possibile notare, è che sono pochissimi gli utenti che scri-

vono recensioni di lunghezza inferiore ai 200 caratteri, ma sono anche pochi quelli che scrivono recensioni la cui lunghezza è maggiore di 1000; quindi possiamo affermare che gli utenti mediamente scrivono recensioni di una lunghezza che va da 200 a 500 caratteri.

Dopo uno studio del testo delle recensioni, si è passato ad uno studio delle valutazioni. Cercando di capire se vi è una relazione tra la valutazione e il numero delle stelle dell'hotel, per capire quali sono i gusti dei turisti e se gli hotel "stellati" sono valutati meglio dei quelli di categoria inferiore.

Per questa analisi si sono andati a recuperare, sempre dalla collezione *textReviews* , i valori attribuiti alla *"ratingOverall"*, che corrisponde alla valutazione complessiva che il turista ha dato all'hotel nella singola recensione; sempre con la stessa richiesta si è memorizzato l'*"hotelId"* per andare a recuperare l'hotel a cui si riferisce la recensione, e dalla collezione *hotels*, si è ricavato lo *"StarRating"* ovvero il numero di stelle.

```
1  //scorro tutte le recensioni
2  for(index = 0; index < count; index ++){
3
4      for (i = 0; i < cursorReviews[index]["reviewDetails"]["
       numberOfReviewsInThisPage"]; i ++){
5
6          //recupero la valutazione complessiva
7          val = cursorReviews[index]["reviewDetails"]["
       reviewCollection"]["review"][i]["ratingOverall"];
8
9          //recupero l'id dell'hotel corrispondente alla
       valutazione
10         idHotel = cursorReviews[index]["reviewDetails"]["
       reviewCollection"]["review"][i]["hotelId"];
11
12         //recupero il numero di stelle
13         cursorHotel = db.hotels.find({"HotelID":idHotel},{"
```

```
StarRating":1,"_id":0});
14        numeroStelle = cursorHotel[0]["StarRating"];
15
16        //controllo il numero di stelle per andare ad
    incrementare il contatore giusto
17        if (numeroStelle == "1.0" || numeroStelle == "1.5") {
18            if (val == 1) {
19                unaStellaValUno ++;
20            }else if ( val == 2) {
21                unaStellaValDue ++;
22            }else if ( val == 3) {
23                unaStellaValTre ++;
24            }else if ( val == 4) {
25                unaStellaValQuattro ++;
26            }else if ( val == 5) {
27                unaStellaValCinque ++;
28            }
29
30        } else if (numeroStelle == "2.0" || numeroStelle == "2.5
    ") {
31            if (val == 1) {
32                dueStelleValUno ++;
33            }else if ( val == 2) {
34                dueStelleValDue ++;
35            }else if ( val == 3) {
36                dueStelleValTre ++;
37            }else if ( val == 4) {
38                dueStelleValQuattro ++;
39            }else if ( val == 5) {
40                dueStelleValCinque ++;
41            }
42
43        } else if ( numeroStelle == "3.0" || numeroStelle == "
    3.5") {
44            if (val == 1) {
45                treStelleValUno ++;
46            }else if ( val == 2) {
47                treStelleValDue ++;
```

```
48              }else if ( val == 3) {
49                  treStelleValTre ++;
50              }else if ( val == 4) {
51                  treStelleValQuattro ++;
52              }else if ( val == 5) {
53                  treStelleValCinque ++;
54              }
55
56          } else if ( numeroStelle == "4.0" || numeroStelle == "
    4.5") {
57              if (val == 1) {
58                  quattroStelleValUno ++;
59              }else if ( val == 2) {
60                  quattroStelleValDue ++;
61              }else if ( val == 3) {
62                  quattroStelleValTre ++;
63              }else if ( val == 4) {
64                  quattroStelleValQuattro ++;
65              }else if ( val == 5) {
66                  quattroStelleValCinque ++;
67              }
68
69          } else if ( numeroStelle == "5.0" || numeroStelle == "
    5.5") {
70              if (val == 1) {
71                  cinqueStelleValUno ++;
72              }else if ( val == 2) {
73                  cinqueStelleValDue ++;
74              }else if ( val == 3) {
75                  cinqueStelleValTre ++;
76              }else if ( val == 4) {
77                  cinqueStelleValQuattro ++;
78              }else if ( val == 5) {
79                  cinqueStelleValCinque ++;
80              }
81
82          } else {
83              if (val == 1) {
```

```
84                  noStelleValUno ++;
85              }else if ( val == 2) {
86                  noStelleValDue ++;
87              }else if ( val == 3) {
88                  noStelleValTre ++;
89              }else if ( val == 4) {
90                  noStelleValQuattro ++;
91              }else if ( val == 5) {
92                  noStelleValCinque ++;
93              }
94          }
95      }
96  }
```

Dallo script è semplice capire che sono stati dichiarati 5 contatori per ogni tipologia di hotel, quindi 5 contatori, ognuna per le 5 differenti valutazioni (da 1 a 5), per le 6 tipologie di hotel, che corrispondono a hotel non stellati e hotel con una stella fino agli hotel con 5 stelle.

La Tabella 4.7 ci mostra una sintesi dei risultati ottenuti.

I risultati sono stati rappresentati a loro volta in diagrammi, inseriti nell'Appendice B, per poterli comprendere meglio.

Da questi diagrammi si può già iniziare a vedere che indipendentemente dalla tipologia di hotel, quindi indipendentemente dal numero delle stelle, gli utenti tendono a rilasciare principalmente recensioni positive. Questo va ad assecondare la tesi dell'articolo *"Online Customer Reviews of Hotels. As Participation Increases, Better Evaluation Is Obtained"* , il quale affermava che al giorno d'oggi il numero di recensioni positive è maggiore del numero quelle negative. Proprio come si può leggere da questi grafici, infatti il numero nelle valutazioni con rating 5 è nettamente maggiore delle altre.

Per capire ancora meglio questo risultato i dati sono stati raggruppati in

#STELLE	VALUTAZIONE	#RECENSIONI	# RECENSIONI TOT PER #STELLE	% SUL TOTALE RELATIVO AL # STELLE
0	1	684	26802	2,55%
	2	1097		4,09%
	3	2587		9,65%
	4	7986		29,80%
	5	14448		53,91%
1	1	87	1985	4,38%
	2	158		7,96%
	3	350		17,63%
	4	773		38,94%
	5	617		31,08%
2	1	281	6505	4,32%
	2	476		7,32%
	3	1167		17,94%
	4	2393		36,79%
	5	2188		33,64%
3	1	1366	53837	2,54%
	2	2799		5,20%
	3	7306		13,57%
	4	20439		37,96%
	5	21927		40,73%
4	1	1975	106495	1,85%
	2	4479		4,21%
	3	10524		9,88%
	4	35598		33,43%
	5	53919		50,63%
5	1	228	13317	1,71%
	2	498		3,74%
	3	1030		7,73%
	4	3011		22,61%
	5	8550		64,20%

Tabella 4.7: Correlazione valutazione-numero stelle.

un unico grafico (Figura 4.4), che rappresenta tutte le valutazioni delle varie tipologie di hotel raggruppate per rating.

Proprio come riportato precedentemente, in media il numero delle recensioni positive supera di gran lunga il numero di quelle negative; in particolare per tutte le categorie di hotel, le recensioni con rating 1 non superano mai il 10% del totale, invece per quanto riguarda le valutazioni positive, in particolare con rating 5, superano addirittura il 60% per gli hotel a 5 stelle.

Se invece raggruppiamo i risultati per tipologia di hotel, partendo da-

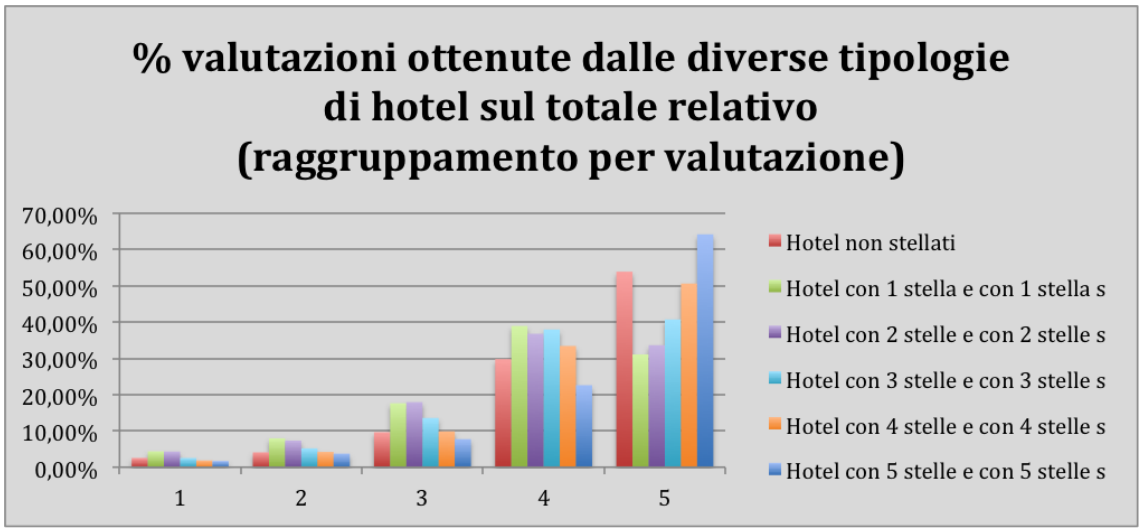

Figura 4.4: Sintesi delle valutazioni per tipologia di hotel raggruppate per rating.

gli hotel non stellati fino ad arrivare agli hotel di 5 stelle, otteniamo un diagramma come quello rappresentato nella Figura 4.5:

Da questo diagramma emergono diverse considerazioni: la prima è che, come si potrebbe pensare, gli hotel di fascia alta, ovvero quelli più lussuosi di 4-5 stelle ottengono, quasi tutti, recensioni con rating molto elevato, tra 4-5. Invece gli hotel di fascia medio bassa, 2-4 stelle, ottengono valutazioni sempre più bilanciate man mano che la fascia di hotel si abbassa; infatti a differenza degli hotel a 5 stelle in cui le recensioni con rating inferiore a 4 sono praticamente assenti, man mano che la fascia di hotel si abbassa le valutazioni negative iniziano progressivamente ad aumentare a discapito di quelle positive, ma senza mai superarle. Un'ultima considerazione è che, questa volta al contrario di quanto si possa aspettare, gli hotel non stellati, quindi di fascia più bassa in assoluto, hanno un numero di valutazioni positive molto più elevato di altri hotel di fascia superiore; infatti il 53,91% delle loro recensioni ha rating 5, superando addirittura i 50,63% degli hotel a 4 stelle.

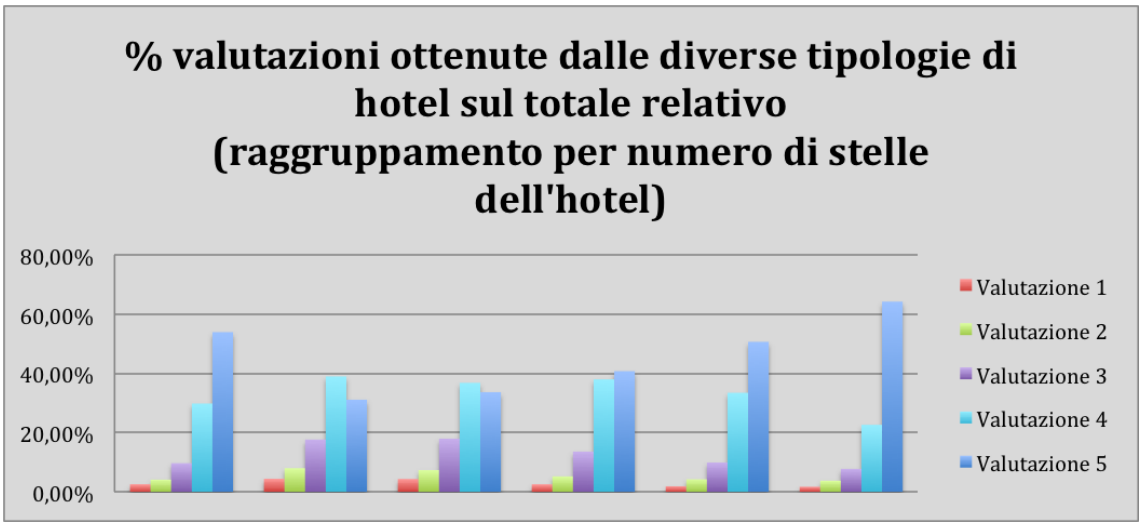

Figura 4.5: Sintesi delle valutazioni raggruppate per tipologia di hotel.

4.2.2 L'utilizzo da parte degli hotel

Dopo aver capito come viene utilizzato Expedia dai turisti, si è studiato come viene utilizzata la piattaforma dall'altra tipologia di utenti, gli albergatori.

Anche in questo caso la prima domanda che ci siamo posti è stata: Quanto viene utilizzato Expedia dagli hotel, in particolare appunto dagli albergatori?

Ma prima ancora bisogna tener presente che su un totale di 21424 hotel trovati, solo 15903 hanno almeno una recensione, quindi esattamente il 74,2% degli hotel è recensito. Di questi 15903 hotel, 7058 rispondono al 100% delle loro recensioni, quindi un buon 44,4%; invece 432 hotel, quindi solo il 2,7% non rispondono affatto alle recensioni.
Potrebbe sembrare, in questo modo, che una buona maggioranza degli albergatori sia molto attiva sui social media, data la stragrande maggioranza degli hotel che rispondono al 100% delle recensioni rispetto agli hotel che non rispondono. Ma questa visione potrebbe cambiare se consideriamo che

1402 hotel hanno solo una recensione e che quasi il 44% degli hotel ha meno di 5 recensioni, tutti dati che vanno ad incidere molto sul risultato finale, poiché gli hotel con poche recensioni incidono molto di più sul totale, invece gli hotel che dovrebbero incidere maggiormente dovrebbero essere gli hotel con molte recensioni.

Per ottenere queste informazioni è stato recuperata la percentuale delle recensioni che hanno avuto una risposta da parte dell'hotel, ovvero il *"recommendedPercent"* della collezione *summaryReviews*. Dopo questo analisi, si è passati ad una ricerca un po' più specifica, per raggruppare gli hotel a seconda delle percentuali di recensioni ri-commentate; viene riportato una porzione di codice che rappresenta l'algoritmo per determinare il numero di hotel che ha una percentuale di recensioni ri-commentate minore uguale a 5; e allo stesso modo è stato fatto per tutti i valori fino a 100.

```
1  //CALCOLO PERCENTUALI RISPOSTE CON <= a
2  cursor = db.summaryReviews.find();
3  count = db.summaryReviews.find().count();
4  countPercentuale = 0;
5  val = 5;
6  for( i = 0; i < count; i ++){
7
8      percentuale = cursor[i]["reviewSummaryCollection"]["
       reviewSummary"][0]["originSummary"][0]["recommendedPercent"];
9      if (percentuale >= val) {
10         countPercentuale ++;
11     }
12 }
13 print(countPercentuale);
```

Calcolando in questo caso quanti hotel hanno una percentuale di risposte minore di "n".
I risultati ottenuti sono stati riportati nella Tabella 4.8.

Questi dati sono stati rappresentati in un'altra curva CDF, dove sul-

#HOTEL	<= %RECOMMENTED	%HOTEL
432	5	2,0%
435	10	2,0%
441	15	2,1%
465	20	2,2%
502	25	2,3%
511	30	2,4%
637	35	3,0%
703	40	3,3%
753	45	3,5%
1281	50	6,0%
1340	55	6,3%
1469	60	6,9%
1750	65	8,2%
2421	70	11,3%
3137	75	14,6%
3603	80	16,8%
4994	85	23,3%
6517	90	30,4%
8061	95	37,6%
15903	100	74,2%

Tabella 4.8: Percentuale di risposte.

l'asse delle x ci sono i valori delle "recommendedPercent" invece sull'asse delle y ci sono le percentuali, corrispondenti al numero degli hotel che rispondono ad una percentuale di recensioni minore o uguale al numero di *"recommendedPercent"* presente sull'asse delle x.

Come si evinceva dai dati precedentemente commentati, anche da questo grafico si può vedere un assiduo utilizzo dei social media, in questo caso Expedia, anche da parte degli hotel. Infatti se meno del 20% di hotel risponde a meno dell'80% delle recensioni, vuol dire che circa l'80% degli hotel risponde ad almeno l'80% delle sue recensioni.

Infine, dopo aver calcolato la frequenza con cui gli albergatori rispondono alle recensioni, ci si è chiesti: con quale criterio rispondono a queste recensioni?

Ovvero con quest'ultima analisi relativa all'utilizzo di Expedia da parte

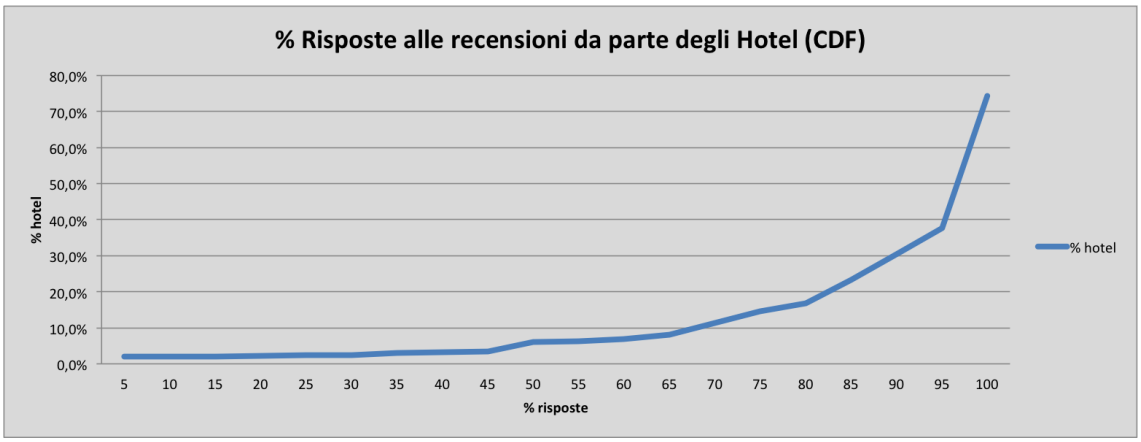

Figura 4.6: Percentuale di hotel per % di risposte alle recensioni

degli hotel, si vuole capire a quali tipologie di recensioni rispondono con più frequenza gli albergatori. Per ottenere una risposta a questa domanda si sono recuperati i "ratingOverall" dalla collezione "textReviews", a differenza di prima che per determinare la percentuale di recensioni ricommentate si era utilizzata la collezione "summaryReviews"; questo perché per sapere la valutazione attribuita alla singola recensione è necessario accedere alla collezione "textReviews". Prendendo quindi come riferimento solo le recensioni in lingua inglese si ha che su un totale di 208941 recensioni, 185643 di queste sono ri-commentate, quindi addirittura l'88,85% delle recensioni in lingua inglese viene ri-commentato.

Entrando più nello specifico siamo quindi andati a determinare a quali tipologie di recensioni rispondono maggiormente.

```
1 //scorro tutte le recensioni
2 for(index = 0; index < count; index ++){
3
4     for (i = 0; i < cursor[index]["reviewDetails"]["
    numberOfReviewsInThisPage"]; i ++){
5
```

```
6         //recupero la valutazione complessiva
7         val = cursor[index]["reviewDetails"]["reviewCollection"
    ]["review"][i]["ratingOverall"];
8         //controllo che se e' stata ricommentata o no
9         booleanRecommended = cursor[index]["reviewDetails"]["
    reviewCollection"]["review"][i]["recommended"];
10
11        if (val == 1) {
12            totUno ++;
13        }else if (val == 2) {
14            totDue ++;
15        }else if (val == 3) {
16            totTre ++;
17        }else if (val == 4) {
18            totQuattro ++;
19        }else if (val == 5) {
20            totCinque ++;
21        }
22
23        //Calcolo in lumero totale di risposte
24        if (booleanRecommended == true) {
25            if (val == 1) {
26                valUno ++;
27            } else if (val == 2) {
28                valDue ++;
29            } else if (val == 3) {
30                valTre ++;
31            } else if (val == 4) {
32                valQuattro ++;
33            } else if (val == 5) {
34                valCinque ++;
35            }
36        }
37    }
38 }
```

Ottenendo quindi questo insieme di valori:

Già dalla Tabella 4.9 è possibile affermare con certezza che gli albergato-

VALUTAZIONE MEDIA	#RISPOSTE	#RECENSIONI	%SUL TOTALE PER VAL
1	161	4621	0,09%
2	597	9507	0,32%
3	14267	22964	7,69%
4	69171	70200	37,26%
5	101447	101649	54,65%
TOT	185643	208941	100,00%

Tabella 4.9: Corrispondenza valutazione-risposta.

ri rispondono prevalentemente alle recensioni positive, ovvero alle recensioni con rating 4-5, piuttosto che quelle negative, con rating 1-3, proprio come riportato nella letteratura. Queste considerazioni si possono vedere ancora meglio dalla rappresentazione grafica del totale del numero delle risposte (Figura 4.7).

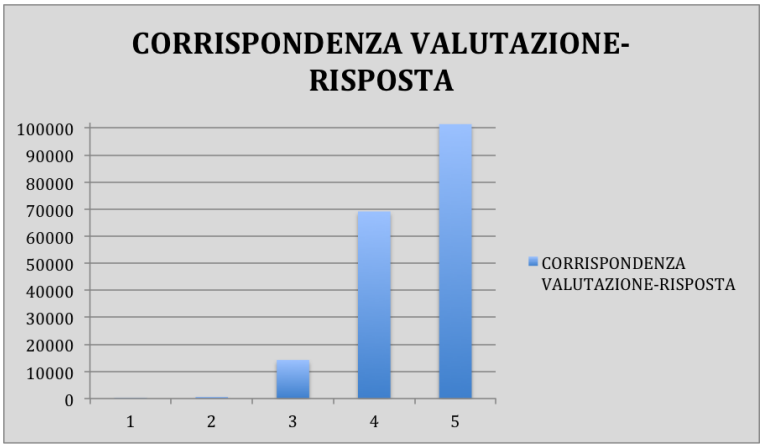

Figura 4.7: Corrispondenza valutazione-risposta con totali interi

Infatti su un totale di 185643 risposte, 101447, quindi esattamente il 54,65%, sono relative a recensioni con rating 5, e solamente 161 risposte, quindi lo 0,09%, sono relative a recensioni con rating 1. Graficamente, queste percentuali, possono essere rappresentate dal grafico rappresentato in

Figura 4.8.

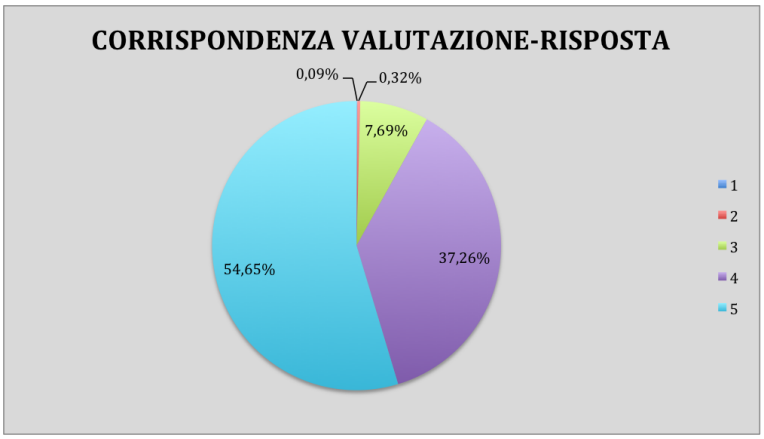

Figura 4.8: Corrispondenza valutazione-risposta con totali in percentuale.

In questo grafico vengono rappresentate rispettivamente con il colore azzurro, nettamente la fetta che copre la sezione più grande del grafico, la percentuale di risposte a recensioni con rating 5, invece con il colore blu e con il colore rosso, che addirittura fanno fatica a vedersi, vengono rappresentate le risposte a recensioni negative, con rating pari a 1-2.

4.3 Terza fase: Chi sono i turisti che visitano l'Italia?

La terza e ultima fase di questa ricerca si occupa di determinare chi sono gli utenti che recensiscono gli hotel italiani, per andare poi a confrontare i risultati ottenuti con la densità dei turisti italiani, per vedere se effettivamente il numero delle recensioni suddivise per lingua rispecchia la reale distribuzione di turisti nelle varie regioni dell'Italia.

Prima di tutto si è calcolato il numero delle recensioni in base alla lingua.

Prendiamo come esempio il codice per calcolare il numero delle recensioni inglesi, ma allo stesso modo si è fatto per parecchie altre lingue, come il tedesco, il francese e lo spagnolo, oltre che per l'italiano.

```
cursor = db.summaryReviews.find({},{"reviewSummaryCollection.
    reviewSummary.originSummary.languageCounts.en":1})
count = db.summaryReviews.find({},{"reviewSummaryCollection.
    reviewSummary.originSummary.languageCounts.en":1}).count()
enCount = 0
sum = 0;
for(i = 0; i < count; i++){
     if (cursor[i]["reviewSummaryCollection"]["reviewSummary"
    ][0]["originSummary"][0]["languageCounts"]["en"] > 0) {
          sum = sum + cursor[i]["reviewSummaryCollection"]["
    reviewSummary"][0]["originSummary"][0]["languageCounts"]["ru"
    ];
          enCount++
     }
}
print(enCount);
print(sum);
```

Come si può interpretare dal codice sopra riportato, il numero delle recensioni per lingua è stato ricavato dalla collezione *"languageCounts"* all'interno di *summaryReviews*. I risultati così ottenuti vengono riportati nella Tabella 4.10.

Si è quindi riusciti a recuperare un totale di 451563 recensioni, su un totale di 458719, questo vuol dire che vi sono altre 7156 recensioni in altre lingue. Ma già da questi risultati vi vede benissimo che la stragrande maggioranza delle recensioni è in lingua inglese, come già ribadito più volte, seguite da quelle scritte in italiano.

LINGUE	SIGLE	#RECENSIONI
Inglese	en	216929
Tedesco	de	63523
Coreano	ko	1230
Portoghese	pt	1529
Giapponese	ja	11644
Italiano	it	88372
Francese	fr	43304
Danese	da	3186
Olandese	nl	12030
Spagnolo	es	9175
Cinese	zh	574
Thailandese	th	64
Vietnamita	vi	3
Arabo	ar	0
Russo	ru	0

Tabella 4.10: Numero di recensioni per lingua.

Dalla Figura 4.9, notiamo che le quattro lingue più diffuse per la redazione delle recensioni su Expedia sono: inglese, con 216929 recensioni, che coprono il 47,3%; italiano, con 88372 recensioni, che coprono quindi il 19,3%; tedesco, con 63523 recensioni, che equivalgono al 13,8%; e infine, il francese con 43304 recensioni, che copre quindi il 9,4% del totale.

La stragrande maggioranza delle recensioni in lingua inglese su Expedia, potrebbe essere dovuta tra le tante cose, anche al fatto che questa categoria di recensioni comprende turisti di ogni parte del mondo, infatti sono classificati come tali anche Americani e Australiani per esempio, la cui lingua madre è l'inglese, ma anche altri utenti provenienti da altre nazioni che usano l'inglese come lingua universale, quindi compresa da tutti, per postare i loro

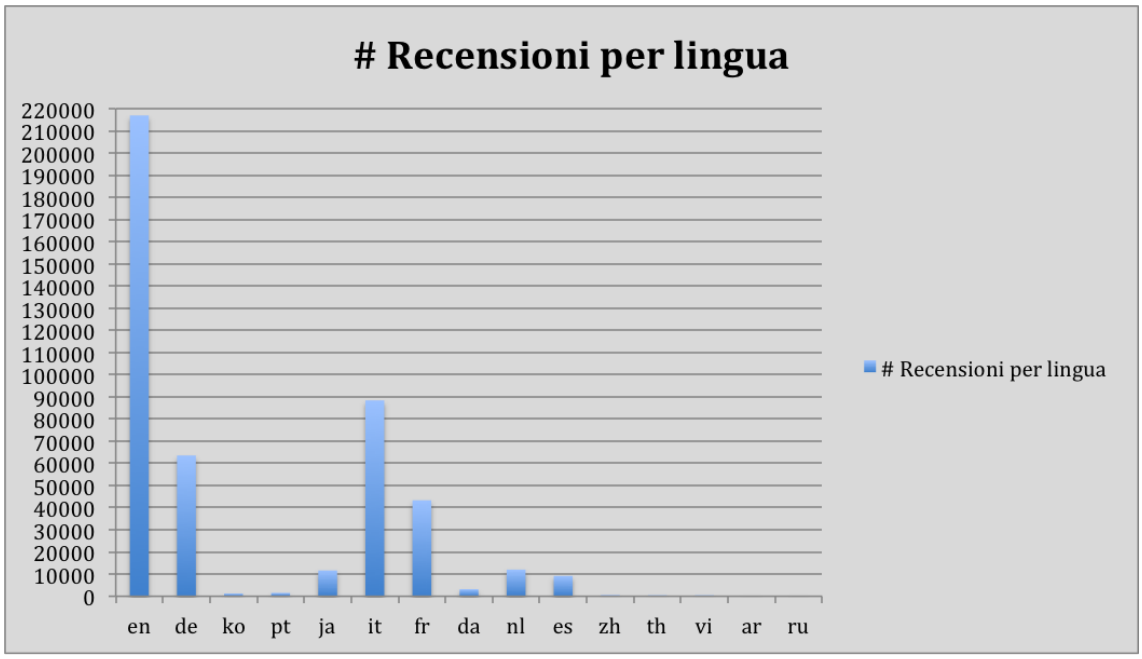

Figura 4.9: Numero di recensioni per lingua.

commenti.

Una volta capito chi sono gli utenti di Expedia, andiamo a studiare la loro distribuzione nelle varie regioni d'Italia, andando quindi a recuperare gli hotel in cui hanno alloggiato e a vedere come si distribuiscono i turisti in Italia che utilizzano questa piattaforma.

In una prima fase si sono suddivisi gli hotel per regione, raggruppandoli quindi per le province che costituiscono una determinata regione, controllando il parametro *"Province"* della collezione *"Location"* all'interno di *hotels*, e il parametro *"City"* recuperato dalla medesima collezione.
Non è bastato il parametro *"Province"* per il raggruppamento, poiché per alcuni hotel questo campo risultava vuoto, ad esempio per gli hotel situati a Milano, o a Roma; per questo motivo è stato aggiunto il parametro *"City"*, al quale per molte città, come Venezia e Firenze, si è dovuto aggiungere

anche la traduzione inglese del nome della città, quindi Venice e Florence, e allo stesso modo per tutte le città italiane che hanno una traduzione inglese, questo perché non tutte le città italiane sono registrate allo stesso modo.
Un altro problema, riscontrato durante la lettura del parametro *"Province"* è che non tutti questi valori rispecchiano la stessa struttura, infatti la maggior parte è scritta in sigle come "NA" per Napoli, "RM" per Roma, "BO" per Bologna", etc. ma altre sono scritte in parola, come ad esempio "Salerno", "Roma", che quindi compare sia in sigle che in parola ma compare anche in inglese "Rome" e tante altre ancora. Ma il problema principale è che alcune province sono addirittura memorizzate con il nome della regione, sono state trovate quindi delle province memorizzate come "Lombardia" e "Lombardy" o peggio ancora con il nome della nazione "Italia" o "Italy". Infine caso ancora più assurdo alcune regioni sono memorizzate con errori ortografici, ma per fortuna di queste se ne sono trovate pochissime.

Viene ad esempio riportata la query utilizzata per il recupero degli hotel della Lombardia:

```
cursorLombardia = db.hotels.find({$or:[{"Location.City":"Bergamo"},
                    {"Location.City":"Brescia"},
                    {"Location.City":"Como"},
                    {"Location.City":"Cremona"},
                    {"Location.City":"Lecco"},
                    {"Location.City":"Lodi"},
                    {"Location.City":"Mantova"},
                    {"Location.City":"Mantua"},
                    {"Location.City":"Milano"},
                    {"Location.City":"Milan"},
                    {"Location.City":"Monza"},
                    {"Location.City":"Pavia"},
                    {"Location.City":"Sandrio"},
                    {"Location.City":"Varese"},
                    {"Location.City":"Garda"},
                    {"Location.City":"Valtellina"},
                    {"Location.City":"Bellagio"},
```

```
18                          {"Location.Province":"BG"},
19                          {"Location.Province":"BS"},
20                          {"Location.Province":"CO"},
21                          {"Location.Province":"LC"},
22                          {"Location.Province":"LO"},
23                          {"Location.Province":"MN"},
24                          {"Location.Province":"MI"},
25                          {"Location.Province":"MB"},
26                          {"Location.Province":"PV"},
27                          {"Location.Province":"SO"},
28                          {"Location.Province":"VA"},
29                          {"Location.Province":"CR"},
30                          {"Location.Province":"Lombardia"},
31                          {"Location.Province":"Lombardy"},
32                          {"Location.Province":"Milano"}
33                          ]})
```

Quindi per colpa di tutte queste problematiche gli hotel recuperati e correttamente raggruppati sono 20334 su 21424, se ne sono quindi persi solo il 5%.

Per la precisione sono stati trovati i valori riportati nella Tabella 4.11, suddivisi per regioni.

Infine si è appunto calcolato il numero delle recensioni per le lingue con maggior numero di recensioni, quali inglese (en), tedesco (de), italiano (it) e francese (fr), per ogni regione italiana.

```
1 //Ricerca della densita dei turisti
2 sumEn = 0;
3 sumDe = 0;
4 sumIt = 0;
5 sumFr = 0;
6 cursorRegione = cursorEmiliaRomagna;
7 count = countEmiliaRomagna;
8 for ( i = 0; i < count; i++){
```

REGIONI	#HOTEL TROVATI
ABRUZZO	240
BASILICATA	105
CALABRIA	344
CAMPANIA	1390
EMILIA ROMAGNA	1258
FRIULI-VENEZIA-GIULIA	225
LAZIO	2854
LIGURIA	724
LOMBARDIA	2034
MARCHE	318
MOLISE	37
PIEMONTE	689
PUGLIA	1070
SARDEGNA	839
SICILIA	2012
TOSCANA	2701
TRENTINO ALTO ADIGE	985
UMBRIA	534
VALLE D'AOSTA	231
VENETO	1744

Tabella 4.11: Numero hotel trovati per regione.

```
9
10    cursorIdHotel = cursorRegione[i]["HotelID"];
11
12    if (cursorRegione[i]["GuestReviewCount"] > 0) {
13        cursorReviews = db.summaryReviews.find({"
reviewSummaryCollection.reviewSummary.hotelId":cursorIdHotel
});
14        countReviews = db.summaryReviews.find({"
reviewSummaryCollection.reviewSummary.hotelId":cursorIdHotel
}).count();
15
16        if (countReviews > 0){
17            if(cursorReviews[0]["reviewSummaryCollection"]["
reviewSummary"][0]["originSummary"][0]["languageCounts"]["en"
] > 0) {
18
19                sumEn = sumEn + cursorReviews[0]["
```

```
    reviewSummaryCollection"]["reviewSummary"][0]["originSummary"
    ][0]["languageCounts"]["en"];
20              }
21              if(cursorReviews[0]["reviewSummaryCollection"]["
    reviewSummary"][0]["originSummary"][0]["languageCounts"]["de"
    ] > 0) {
22
23                  sumDe = sumDe + cursorReviews[0]["
    reviewSummaryCollection"]["reviewSummary"][0]["originSummary"
    ][0]["languageCounts"]["de"];
24              }
25              if(cursorReviews[0]["reviewSummaryCollection"]["
    reviewSummary"][0]["originSummary"][0]["languageCounts"]["it"
    ] > 0) {
26
27                  sumIt = sumIt + cursorReviews[0]["
    reviewSummaryCollection"]["reviewSummary"][0]["originSummary"
    ][0]["languageCounts"]["it"];
28              }
29              if(cursorReviews[0]["reviewSummaryCollection"]["
    reviewSummary"][0]["originSummary"][0]["languageCounts"]["fr"
    ] > 0) {
30
31                  sumFr = sumFr + cursorReviews[0]["
    reviewSummaryCollection"]["reviewSummary"][0]["originSummary"
    ][0]["languageCounts"]["fr"];
32              }
33          }
34      }
35  }
```

Il codice riporta l'algoritmo utilizzato per determinare la densità dei turisti, in base alla lingua in cui sono state scritte le recensioni, per la regione Emilia Romagna, come possiamo notare dal *cursorRegione* che assume i valori di *cursorEmiliaRomagna*; infatti questa regione è stata presa come esempio, ma allo stesso modo è stato fatto per tutte le altre 19 regioni italiane.
Il procedimento è abbastanza semplice, tramite il ciclo for si scorrono tutti

gli hotel per una determinata regione, di questi hotel, viene preso *"HotelID"*, e si controlla che il *GuestReviewCount"* sia maggiore di 0, ovvero che l'hotel abbia recensioni, altrimenti non viene considerato; dopodiché dalla collezione *summaryReviews* si verifica il numero delle recensioni per una determinata lingua, e si vanno a sommare con il numero delle recensioni della stessa lingua dell'hotel precedente.

Ottenendo come output finale i seguenti valori:

REGIONI	# EN	# DE	# IT	# FR
ABRUZZO	249	90	720	30
BASILICATA	205	29	371	35
CALABRIA	315	123	612	40
CAMPANIA	18199	2082	5270	2327
EMILIA ROMAGNA	4819	1537	7442	788
FRIULI-VENEZIA-GIULIA	489	603	759	57
LAZIO	72579	23817	13959	14134
LIGURIA	4652	1153	3574	1477
LOMBARDIA	21296	8368	11282	3381
MARCHE	316	114	1096	44
MOLISE	29	3	62	2
PIEMONTE	2616	1115	3934	1224
PUGLIA	1232	408	2815	321
SARDEGNA	2511	1976	3042	976
SICILIA	8822	3181	9054	2836
TOSCANA	33450	4980	10573	5059
TRENTINO ALTO ADIGE	801	1693	1964	44
UMBRIA	1368	163	2292	107
VALLE D'AOSTA	253	36	632	111
VENETO	42201	12106	8225	10181

Tabella 4.12: Distribuzione dei turisti italiani in base alle recensioni.

La Tabella 4.12 si può rappresentare in un diagramma a colonne per avere una visione migliore di questa distribuzione.

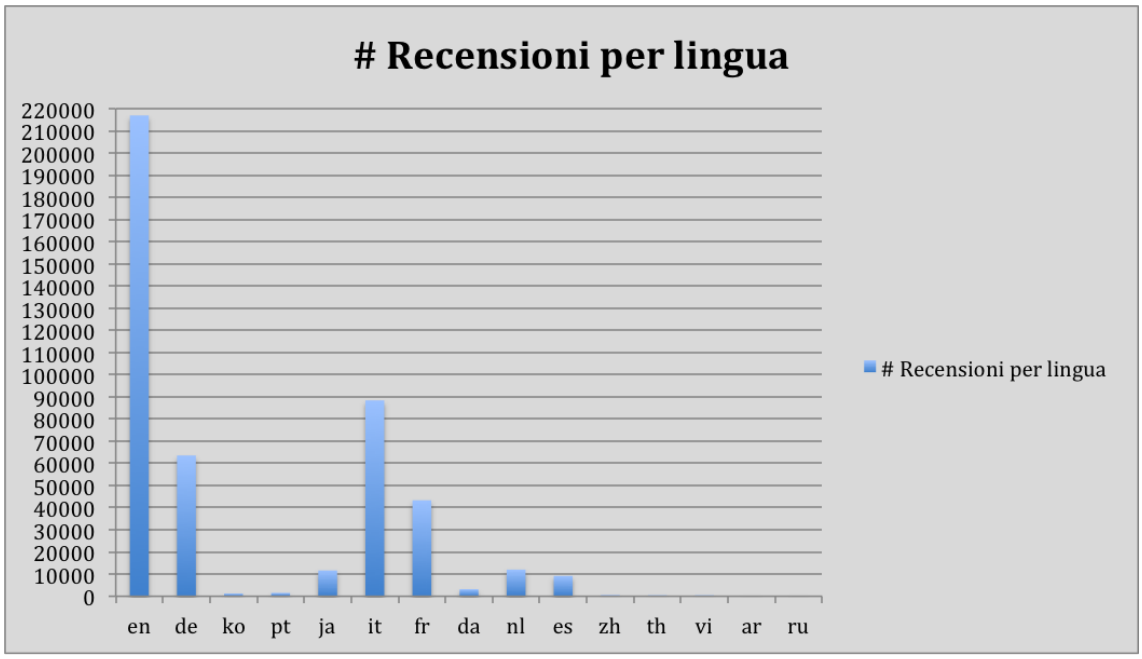

Figura 4.10: Numero di recensioni per lingua.

Dalla Figura 4.10 è possibile notare due cose: la prima è che senz'altro ci sono regioni in cui il numero delle recensioni è nettamente maggiore delle altre, come Lazio con 124489 recensioni, Lombardia con 44327 recensioni, Toscana, con 54062 recensioni, e Veneto con 72713 recensioni; ma questo dipende anche dal numero di hotel, infatti, ad eccezione del Veneto, sono anche le regioni con il numero più elevato di hotel. Il Veneto, stranamente, con molti meno hotel della Lombardia e della Toscana, si classifica al secondo posto per numero di recensioni ottenute. La seconda cosa che possiamo notare è che per queste grandissime regioni, con un elevato numero di hotel, i turisti inglesi sono nettamente superiori ai turisti di altre nazionalità, per quanto riguarda l'utilizzo di Expedia, arrivando fino ad un totale di 72579 recensioni in lingua inglese per la regione del Lazio. I turisti tedeschi superano gli italiani solo nel Lazio e nel Veneto, per il resto delle regioni, soprattutto nelle regioni più piccole, gli italiani si contendono spesso il primo e il secondo posto con gli inglesi. Per la precisione, nonostante le recensioni inglesi

superino quelle italiane, per 13 regioni su 20 le recensioni italiane superano di numero le recensioni inglesi; questo dovuto comunque al fatto che stiamo parlando di hotel italiani.

In percentuale, questi dati posso essere così raggruppati:

REGIONI	% EN	% DE	% IT	% FR
ABRUZZO	23%	8%	66%	3%
BASILICATA	32%	5%	58%	5%
CALABRIA	29%	11%	56%	4%
CAMPANIA	65%	7%	19%	8%
EMILIA ROMAGNA	33%	11%	51%	5%
FRIULI-VENEZIA-GIULIA	26%	32%	40%	3%
LAZIO	58%	19%	11%	11%
LIGURIA	43%	11%	33%	14%
LOMBARDIA	48%	19%	25%	8%
MARCHE	20%	7%	70%	3%
MOLISE	30%	3%	65%	2%
PIEMONTE	29%	13%	44%	14%
PUGLIA	26%	9%	59%	7%
SARDEGNA	30%	23%	36%	11%
SICILIA	37%	13%	38%	12%
TOSCANA	62%	9%	20%	9%
TRENTINO ALTO ADIGE	18%	38%	44%	1%
UMBRIA	35%	4%	58%	3%
VALLE D'AOSTA	25%	3%	61%	11%
VENETO	58%	17%	11%	14%

Tabella 4.13: Distribuzione dei turisti italiani in base alle recensioni in percentuale.

Questi dati in percentuali vengono rappresentati graficamente dal grafico della Figura 4.11.

Grazie a questo grafico possiamo vedere ancora meglio, quanto è stato affermato precedentemente, ovvero che per il 65% dei casi, ovvero, per 13

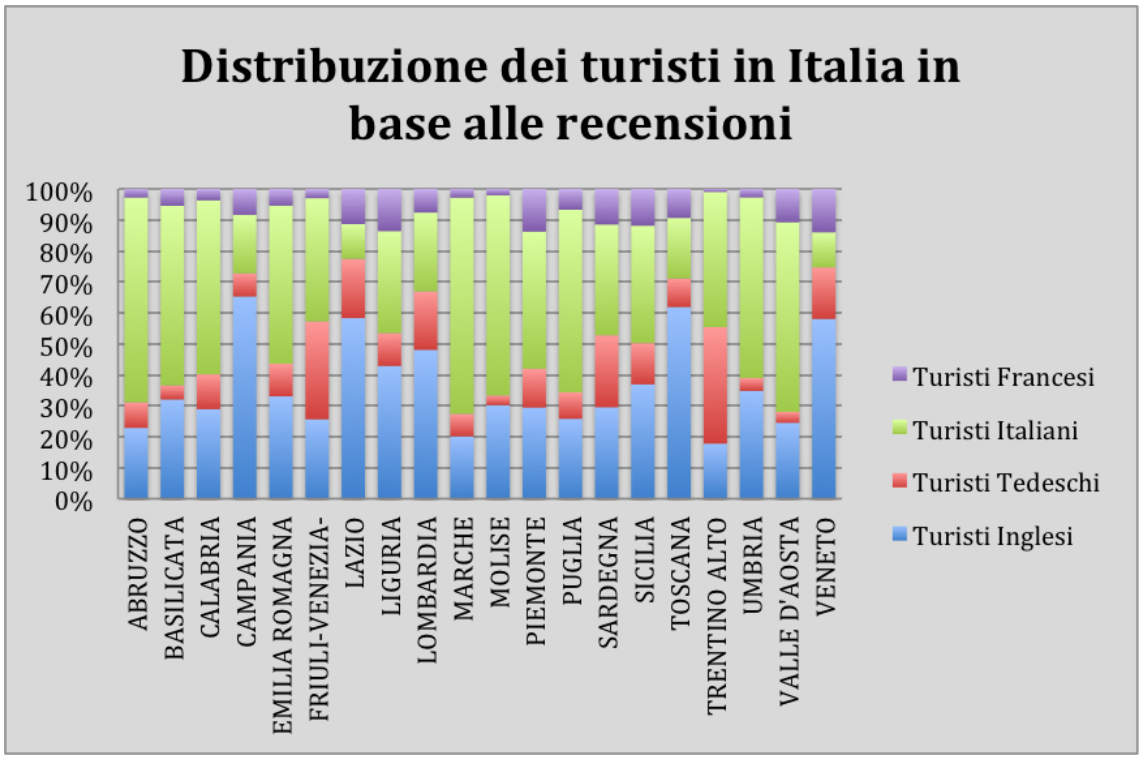

Figura 4.11: Distribuzione dei turisti italiani in base alle recensioni in percentuale.

regioni su 20, le recensioni italiane superano in percentuale le recensioni in lingua straniera, per la singola regione. In particolare nelle Marche si sono calcolati addirittura il 70% di recensioni in lingua italiana.

4.3.1 Le recensioni rispecchiano la realtà

La parte conclusiva di questa ricerca vuole capire se i dati appena raccolti, ovvero i dati relativi al numero delle recensioni suddivise per lingua, in particolare in base alle regioni italiane, rispecchia la reale distribuzione dei turisti che visitano le varie regioni d'Italia.

La prima fonte da cui sono stati ricavati i dati del turismo è ENIT, Ente

Nazionale Del Turismo, il quale ha ricavato questi dati da ISTAT. I dati, relativi all'anno 2014, raccolti da questa fonte sono stati rappresentati nel grafico della Figura 4.12.

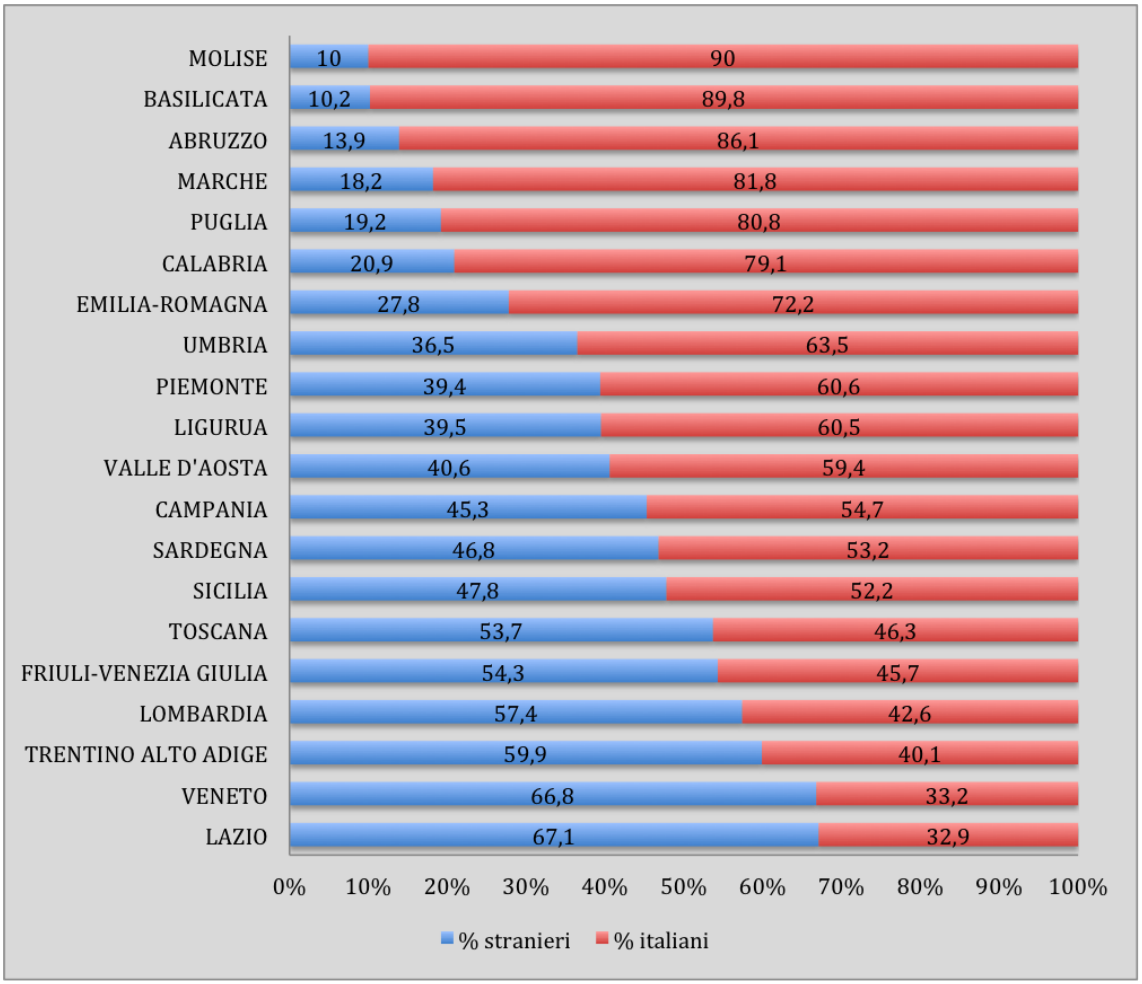

Figura 4.12: Distribuzione del turismo nelle regioni d'Italia, con dati ricavati da ENIT.

Da questo grafico, possiamo notare che la distribuzione dei turisti stranieri nelle varie regioni italiane, non di discosta molto dai risultati ottenuti dallo studio precedente. Infatti se per ogni singola regione andiamo a cal-

colare la percentuale di recensioni italiane e la andiamo a confrontare con la percentuale delle recensioni in lingua straniera in generale otteniamo un grafico molto simile a quello ottenuto dai dati ENIT.

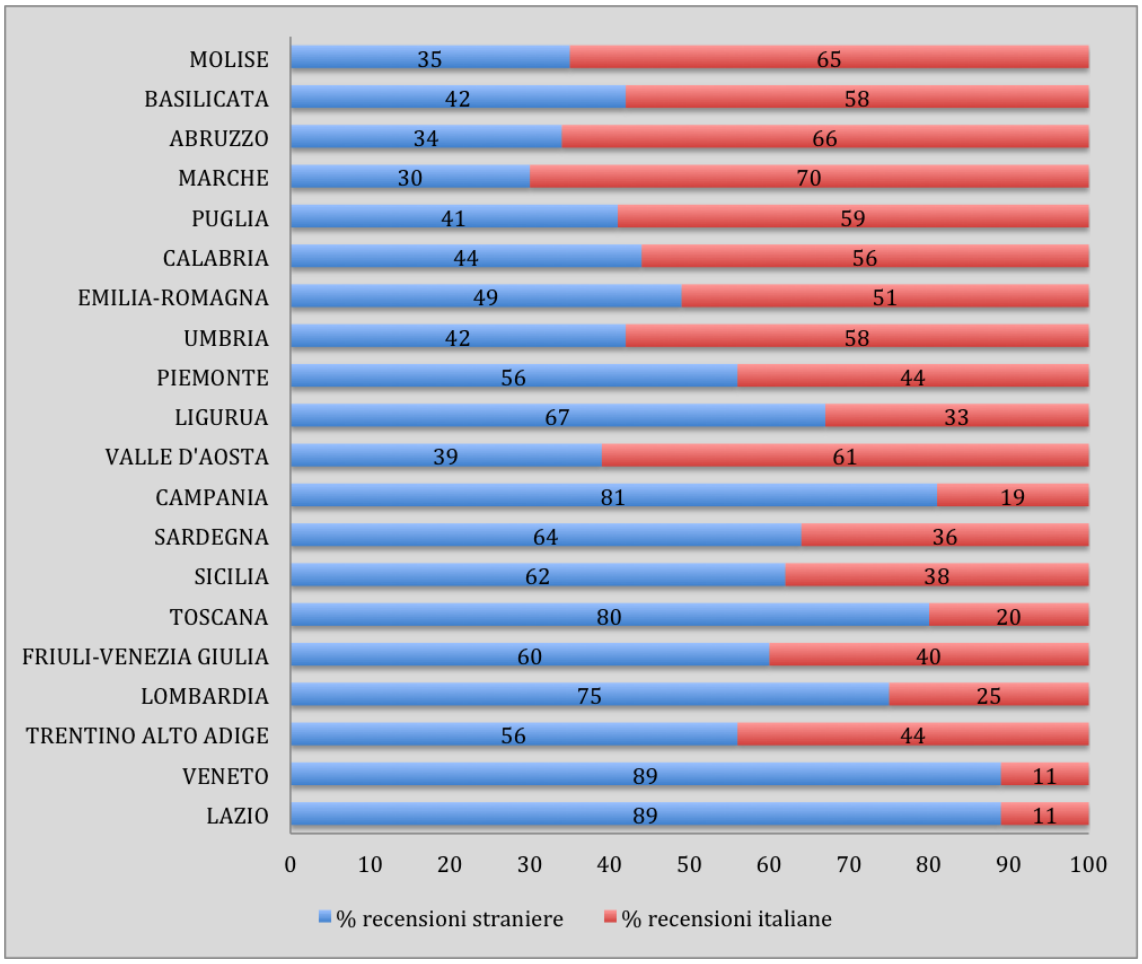

Figura 4.13: Distribuzione del turismo nelle regioni d'Italia, con i dati ricavati dalle recensioni Expedia.

Infatti, se andiamo a vedere per ogni singola regione il rapporto tra recensioni in lingua italiana e recensioni in lingua straniera rispecchiano quasi sempre il rapporto tra turisti italiani e turisti stranieri. Se andiamo ad esempio a vedere la percentuale di turisti stranieri in Molise, e la percentuale di

turisti italiani, notiamo che i turisti italiani superano di gran lunga il numero dei turisti stranieri; allo stesso modo, sempre considerano il Molise, notiamo che anche il numero delle recensioni in lingua italiana, supera di gran lunga il numero delle recensioni in lingua straniera. Stessa cosa per il Lazio, che ha una densità di turismo straniero molto più elevato rispetto a quello italiano e anche in questo caso il numero delle recensioni in lingua straniera supera di gran lunga il numero delle recensioni in italiano.
Quindi possiamo affermare che il rapporto di quantità tra recensioni italiane e recensioni straniere rispecchia per tutti i casi, tranne che per la Valle d'Aosta, la densità dei turisti italiani e stranieri nelle varie regioni italiane. Infatti un maggiore numero di recensioni in lingua italiana, corrispondono ad un maggiore numero di turisti italiani rispetto a quelli stranieri, in quella determinata regione; stessa cosa un maggior numero di recensioni in lingua straniera corrispondono ad un maggior numero di turisti stranieri in quella determinata regione. Come già citato, solo la Valle d'Aosta fa eccezione, infatti il numero delle recensioni italiane supera il numero di quelle straniere, ma la densità dei turisti ci dice l'opposto.

Queste conclusioni si possono notare ancora meglio nella Figura 4.14 che mette assieme i due grafici. Nella prima metà di grafico notiamo i dati provenienti da ISTAT e nella seconda metà, invece, ci sono i risultati ottenuti da questo studio; i quali .

Invece per quanto riguarda la percentuale media di turisti italiani su tutta l'Italia, si discosta molto dalla percentuale di recensioni italiane presenti su Expedia per gli hotel italiani. Infatti la percentuale di turisti italiani, calcolata da ENIT è di 50,6%, e quella di turisti stranieri è di 49,4%; invece per quanto riguarda la percentuale di recensioni con testo in italiano totale è di 21,34, quindi il restante 78,66% ha testo scritto in altre lingue.
Questo, forse, viene spiegato dal fatto che Expedia è più usata dai turisti stranieri rispetto quelli italiani, influenzando in parte i valori a livello Italia,

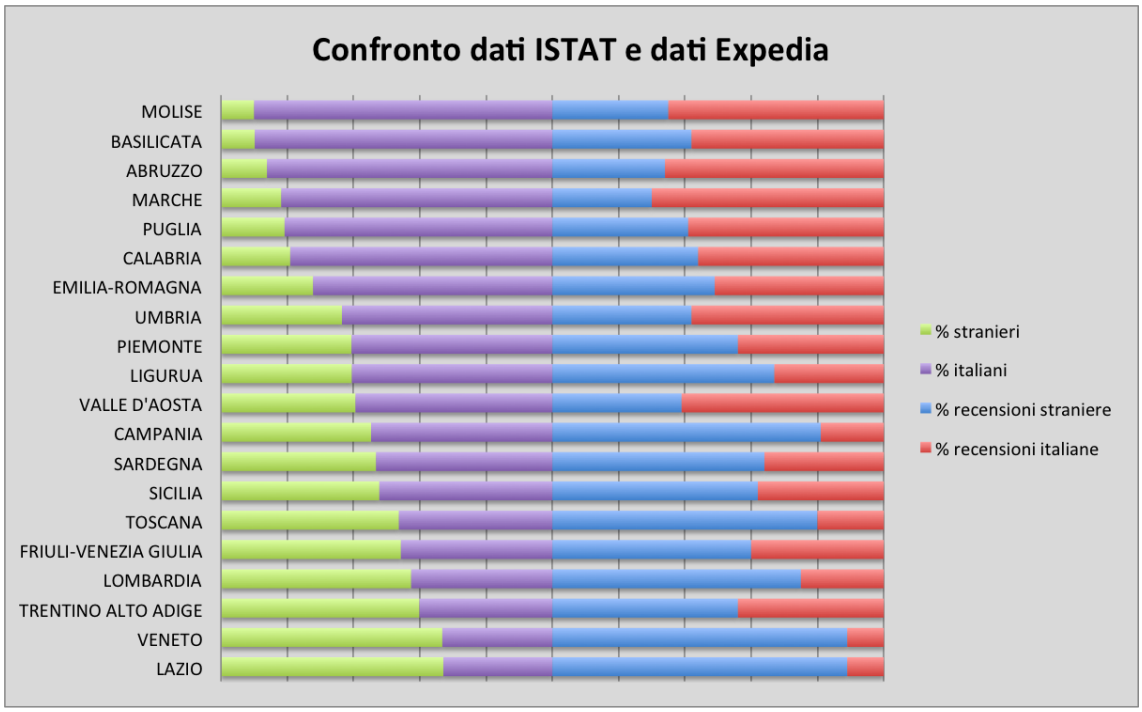

Figura 4.14: Confronto dati ISTAT con risultati Expedia

ma non significativamente la distribuzione (Stranieri/Italiani) per regione.

Conclusioni

Il contributo che vuole dare questa tesi è scoprire come vengono utilizzati i social media dai loro utenti; in particolare che rapporto c'è tra gli utenti delle piattaforme di tipo booking, ovvero turisti e hotel, tramite una analisi delle recensioni.
L'analisi si basa su un campione di 21424 hotel italiani, estratti dalla piattaforma Expedia.com, praticamente i 2/3 di tutti gli hotel che esistono in Italia. Lo studio si basa principalmente sulle recensioni ottenute da questi hotel, per la precisione si sono recuperate un totale di 897806 recensioni.

Il primo obiettivo della ricerca è stato determinare se effettivamente è vero che i social media stanno prendendo sempre più piede; e tramite Expidia.com possiamo affermare che, visto il progressivo aumento del numero delle recensioni dal 2010, con un totale di 20523 recensioni, al 2015, con un totale di 49299 recensioni, Expedia, ma in generale tutte le piattaforme di tipo booking si stanno diffondendo molto, almeno per quanto riguarda lo scambio reciproco di opinioni da parte degli utenti.
Dalla stessa analisi è emerso anche che l'attività degli utenti è concentrata in determinati periodi dell'anno, ovvero ci sono sempre gli stessi mesi dell'anno in cui vi è una maggiore attività nel postare recensioni, rispetto ad altri; ad esempio nei mesi di settembre e ottobre, in cui i turisti sono appena tornati dalle vacanze vi è un picco del numero delle recensioni postate, arrivando fino a 6457 recensioni postate nel mese di settembre del 2015; vi sono altri mesi invece, come gennaio e febbraio, in cui la maggior parte delle persona

lavora, dove il numero delle recensioni postate si riduce notevolmente.

Successivamente si è analizzata l'attività dei turisti nel postare recensioni, controllando il numero di recensioni per ogni hotel, giungendo alla conclusione che il 50% degli hotel ha meno di 10 recensioni; questo vuol dire che la maggior parte degli hotel ha poche recensioni, e quei pochi hotel che hanno tante recensioni ne hanno un numero elevatissimo.
Confermando pertanto la tesi che le recensioni rappresentano un "circolo virtuoso"per l'hotel: tanto più riceve recensioni e diventa popolare, tanto più i consumatori tendono a postare recensioni aumentando ulteriormente la popolarità.

Sempre nella fase di studio dei turisti, come prima tipologia di utenti che utilizzano Expedia, si sono suddivise le recensioni con testo scritto, da quelle con solo la valutazione, e si è giunti alla conclusione che il 51% dei turisti preferisce descrivere testualmente le esperienze vissute durante il soggiorno in hotel. Si è poi calcolato la dimensione media di una recensione, riscontrando che mediamente i turisti scrivono recensioni di 300 caratteri, per la precisione il 60% di questa categoria di utenti; praticamente nessuno scrive commenti di lunghezza inferiore ai 200 caratteri, ma allo stesso modo sono in pochissimi anche quelli che scrivono recensioni molto lunghe di oltre 500 caratteri. Ciò porta ad affermare che quando il cliente posta una recensione scritta non si limita ad un brevissimo giudizio, ma dedica un po' del suo tempo per condividere con altri consumatori la sua esperienza.

Infine ci siamo chiesti in base a cosa attribuiscono le valutazioni, andando a vedere se esiste una correlazione tra rating e numero di stelle dell'hotel; e come prima cosa si è notato che il numero delle recensioni positive supera il numero delle recensioni negative, infatti la maggior parte delle valutazioni ha rating molto alto, tra 4 e 5. La seconda cosa che è stato possibile scoprire è che, come ci si potrebbe aspettare, per gli hotel di fascia alta, quindi hotel

di 4-5 stelle, le valutazioni con rating negativo sono quasi assenti; questo rapporto, ovvero tra rating positivo e rating negativo, inizia leggermente a ristabilirsi mano a mano che la categoria dell'hotel si abbassa. Una cosa abbastanza strana è invece che gli hotel non stellati, quindi di fascia più bassa in assoluto, hanno un numero totale di recensioni con rating 5 addirittura maggiore degli hotel a 4 stelle.

Dopo aver studiato il rapporto che hanno i turisti con Expedia, si è passati ad analizzare la seconda categoria di utenti che utilizza la piattaforma, ovvero gli albergatori. Il primo risultato ottenuto è stato che la maggior parte di loro rispondono alle recensioni che ricevono; infatti circa l'80% degli hotel risponde a più dell'80% delle loro recensioni, questo vuol dire che l'attività degli albergatori sui social media è molto elevata. A concludere lo studio sugli albergatori si è cercato di capire a quale tipologia di recensioni rispondono con più frequenza; i risultati ci dicono che su un totale di 185643 risposte, 101447, quindi esattamente il 54,65%, sono relative a recensioni con rating 5, e solamente 161 risposte, quindi lo 0,09%, sono relative a recensioni con rating 1. Questo vuol dire che gli albergatori preferiscono rispondere alle recensioni positive rispetto a quelle negative.

La terza ed ultima fase della tesi ha lo scopo di determinare alcune caratteristiche dei turisti che soggiornano in Italia. Prima di tutto si sono raggruppate le recensioni in base alla lingua ed è emerso che le quattro lingue più diffuse per la scrittura di commenti su Expedia sono: inglese, con un totale di 216929 recensioni; tedesco, con un totale di 63523 recensioni; italiano, con 88372 recensioni e francese con 43304 recensioni. Questo vuol dire che i commenti in lingua inglese coprono addirittura il 47,3% del totale.

Successivamente, si è fatta un'analisi più specifica, andando a vedere come si distribuiscono queste recensioni, raggruppate per lingua, nelle varie regioni italiane; ed è emerso che per 13 regioni su 20 la lingua più diffusa per scriverle è l'italiano, e le restanti 7 hanno una prevalenza di recensioni

in lingua inglese. Ma si è anche notato che per queste ultime regioni, ovvero quelle in cui il numero delle recensioni in lingua inglese prevale, il numero di queste recensioni è altissimo, raggiungendo le 72579 recensioni, superando il record delle recensioni italiane, detenuto dalla stessa regione, di 13959. Questo potrebbe significare che i turisti stranieri condividono maggiormente le loro esperienze tramite recensioni rispetto a quelli italiani, o che la piattaforma Expedia è più utilizzata dai turisti stranieri.

Infine, si è andati a verificare se vi è una correlazione tra questa distribuzione di recensioni in basse alla lingua e alla densità di turisti stranieri nella varie regioni italiane. Confrontando i dati ottenuti dall'ENIT, recuperati dall'ISTAT, sulla densità di turisti italiani e stranieri in Italia, con il numero delle recensioni in lingua italiana e in lingua straniera, è possibile affermare che i dati ottenuti sono molto simili tra di loro. Infatti per ogni regione, ad eccezione della Valle d'Aosta, ad un maggior numero di recensioni in lingua straniera corrisponde un maggior numero di turisti stranieri per quella determinata regione; e allo stesso modo ad una maggior numero di recensioni in italiano, corrisponde un maggior numero di turisti italiani per quella regione. Ciò che differisce con i dati dell'ENIT sono però le percentuali che rappresentano il numero dei turisti italiani in rapporto ai turisti stranieri; infatti secondo l'ENIT i turisti italiani sono il 50,6%, e quelli stranieri sono il 49,4%. Invece per quanto riguarda la percentuale di recensioni con testo in italiano totale è di 21,34, quindi il restante 78,66% ha testo scritto in altre lingue. Questo, forse, viene spiegato dal fatto che Expedia, è più usata dai turisti di lingua inglese, rispetto che da quelli italiani.

Il limite di questa analisi, oltre alla piccola percentuale di errore di estrazione, ampliamente documentata all'interno della tesi, risiede nel prendere come campione di hotel solo quelli provenienti dalle regioni italiane. Un'analisi più approfondita, riuscendo ad ottenere gli strumenti giusti, si potrebbe fare prendendo come campione di dati non solo quelli relativi all'Italia ma

di tutta l'Europa, o addirittura di tutto il mondo. Oppure si potrebbe replicare lo studio per un'altra nazione e mettere a confronto i risultati ottenuti.

L'altro grande limite, dovuto alle APIs Expedia e non da altri fattori, è dovuto al fatto che gran parte di questo studio prende come campione di recensioni solo quelle di testo inglese. Se si fosse riusciti a recuperare anche le recensioni scritte in altre lingue si sarebbe potuto estendere l'analisi per un campione di dati più consistente e magari ottenere risultati più precisi.

Oltre ai limiti, si potrebbero fare altre considerazioni e altri possibili studi che non sono stati documentati in questa tesi.
Un esempio potrebbe essere il calcolo della regressione, per vedere quali sono i fattori che maggiormente influenzano le recensioni.
Oppure si potrebbe, prendendo ad esempio come campione di hotel quelli presenti nella regione Emilia Romagna, calcolare il numero delle recensioni e la valutazione media, per un determinato anno, ad esempio il 2013; e infine confrontare i valori ottenuti con i dati ISTAT sull'aumento del turismo tra il 2013 e il 2014, perché se vi sono tante recensioni con valutazioni medie molto alte e il turismo è aumentato, allora vuol dire che questi fattori, ovvero le recensioni e e le valutazioni influenzano notevolmente il turismo. Questo studio si potrebbe fare prendendo come campioni due o tre regioni italiane, e vedere se le conclusioni sono le stesse.
Infine, un'altro studio si potrebbe fare prendendo come riferimenti i ristoranti e non più gli hotel, per vedere se le considerazioni riportate in questa tesi, si potrebbe fare anche per le recensioni relative ai ristoranti.

Appendice A

Sommario articoli

In questa appendice compaiono diverse tabelle, nelle quali vengono riassunti gli articoli presenti nel capitolo 2. In ogni tabella è presente il titolo dell'articolo, l'elenco dei social media da cui hanno preso i dati, il campione dei dati analizzati, gli obiettivi che ogni articolo si è prefissato di analizzare e le conclusioni.

ARTICOLO	PIATTAFORMA	CAMPIONI	OBIETTIVI	CONCLUSIONI
Responding to Online Reviews: Problem Solving and Engagement in Hotels	TripAdvisor	4 hotel di fascia alta nella zona occidentale deigli Stati uniti	Comportamento degli hotel in base alle recensioni che ottengono.	2 hotel rispondono regolarmente alle recensioni, e 2 no.
Factors Affecting Customer Satisfaction in Responses to Negative Online Hotel Reviews	brand websites, agenzie online come Travelocity.com , global distribution systems	Gruppo di 176 potenziali clienti	Analisi della tipologia di risposta degli hotel a recensioni negative.	La risposta deve essere empatica e deve presentare un riferimento specifico alla lamentela.
Online Customer Reviews of Hotels. As Participation Increases, Better Evaluation Is Obtained	TripAdvisor	16680 hotel, in 249 zone turistiche, con un totale di oltre 1,28 milioni di commenti	Studio della valenza e del volume delle recensioni.	Su oltre 1,28 milioni di il 70% è risultato positivo, quindi con una valutazione compresa tra 4 e 5, su una scala di valori che va da 1-5, con 5 massimo.
The Differential Effects of the Quality and Quantity of OnlineReviews on Hotel Room Sales	Blogs, networks, chat rooms, online reviews, e social media in generale	319 hotel di Londra	Analisi degli effetti delle recensioni sugli hote, in termini di valenza e volume.	Ad influenzare l'economia degli hotel di fascia alta, sono le valutazioni e non più il numero di recensioni, contrariamente di quanto accade per gli hotel di fascia bassa.

Tabella A.1: Summary articoli

ARTICOLO	PIATTAFORMA	CAMPIONI	OBIETTIVI	CONCLUSIONI
Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach	TripAdvisor, Yelp, Facebook, Twitter, YouTube, Amazon e Travelocity	70,103 online user reviews	Sentiment analysis sulle recensioni scritte.	A seconda della dimensione dei commenti l'effetto è completamente diverso; importanza di utilizzare dati testuali per recensire la qualità dell'hotel; più un hotel è recensito, e più l'utente è invogliato a scrivere anche lui un commento.
The Influence of Embedded Social Media Channels on Travelers' Gratifications, Satisfaction, and Purchase Intentions	Facebook, Twitter, YouTube, blogs, Flickr, LinkedIn e Google+	Lista dei migliori U.S. hotel del 2011 classificati dall' Hotel Management	Esaminare l'efficacia dei canali di social media utilizzati per le prenotazioni e la loro influenza sul comportamento dei viaggiatori.	I viaggiatori che hanno utilizzato social media si sono rilevati più soddisfatti; i social media hanno influenzato positivamente il viaggiatore nella prenotazione;
Travel Planning: Searching for and Booking Hotels on the Internet	Ricerca tramite intervista	249 turisti in un hotel a Seattle, Washington	Ricerca sulla metodoliga utilizzata dai turisti per prenotare le camere degli hotel.	Degli 8/10 degli intervistati che utlizzano siti web per la ricerca di informazini sulle camere da prenotare, il 67% ha continuato online anche con l'operazione di prenotazione, il 26% ha chiamato direttamente l'hotel e il 7% si è affidato ad un'agenzia di viaggi.
The Complex Matter of Online Hotel Choice	Expedia.com, Travelocity, Orbitz e hotels.com	11 donne e 5 uomini di età compresa tra i 23 e i 42 anni	Analisi sul processo decisionale che porta l'utente a scegliere l'hotel.	I fattori che influenzano la scentla del consumatore sono: posizione dell'hotel nella lista dei risultati; presenza di immagini; il prezzo; descrizione testuale dell'immagine.

Tabella A.2: Summary articoli

ARTICOLO	PIATTAFORMA	CAMPIONI	OBIETTIVI	CONCLUSIONI
What can big data and text analytics tell us about hotel guest experience and satisfaction?	Expedia. com	10537 hotel negli Stati Uniti con un totale di 60648 recensioni	Verificare la utilità dei Big Data per comprendere meglio le esperienze degli ospiti di hotel.	Gli ospiti degli hotel tendono a dare principalmente valutazioni positive.
Customer engagement behaviors and hotel responses	TripAdvisor	101 intervistati	Analizzare come i potenziali clienti percepiscono le recensioni e le risposte degli albergatori.	Le recensioni positive influenzano maggiormente rispetto alle negative; Le risposte degli albergatori a recensioni negative se specifiche e non generiche sono percepite positivamente.
Please, talk about it! When hotel popularity boosts preferences	Simulazione del processo di prenotazione su un campione di persone	161 persone coinvolte	Valutare importanza del numero (popolarità) e rating (qualità) delle recensioni sulle intenzioni di prenotazione di potenziali consumatori.	Il volume influenza maggiormente del rating soprattutto per anziani e femmine.
The effectiveness of managing social media on hotel performance	TripAdvisor, Expedia.com, Priceline, Hotel.com e Yelp.	128 hotel negli Stati Uniti con un totale di 31930 recensioni.	Analizzare come i potenziali clienti percepiscono le recensioni e le risposte degli albergatori.	Le recensioni positive dei clienti e le risposte degli albergatori alle recensioni negative sono i fattori di maggior influenza sui potenziali consumatori.
Web reviews influence on expectations and purchasing intentions of hotel potential customers	TripAdvisor	Interviste a 349 potenziali consumatori italiani	Verificare come le recensioni e le risposte degli albergatori influenzano i potenziali consumatori.	Il rating influenza la scelta di potenziali consumatori mentre le risposte degli albergatori non sono considerate un fattore chiave .

Tabella A.3: Summary articoli

ARTICOLO	PIATTAFORMA	CAMPIONI	OBIETTIVI	CONCLUSIONI
Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry	TripAdvisor	178 hotel negli Stati Uniti	Studiare l'impatto sul business degli hotel di recensioni e risposte degli hotel.	Sia il rating che il volume delle recensioni hanno un impatto sui risultati economici degli hotel.
The business value of online consumer reviews and management response to hotel performance	TripAdvisor	843 hotel nel Texas con un totale di 4994 recensioni	Studiare l'impatto sul business degli hotel di recensioni e risposte degli hotel.	Il fattore più determinante sui risultati economici dell'hotel è la valutazione media delle recensioni, seguito dal volume; mentre non risultano determinanti le risposte degli albergatori.
A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently.	TripAdvisor	141 hotel in Hong Kong con un totale di 86239 recensioni	Studiare le differenze tra rating in lingua inglese ed in lingue non-inglesi.	I consumatori di lingua inglese preferiscono hotel di alta classe e danno valutazioni più positive rispetto a quelli non di lingua inglese.
Compliance with eWOM: The influence of hotel reviews on booking intention from the perspective of consumer conformity	TripAdvisor (recensioni simulate)	160 potenziali clienti	Studiare l'influenza di rating e il volume delle recensioni su potenziali consumatori conformisti e non.	Il rating delle recensioni influenza le scelte di potenziali clienti soprattutto per i conformisti, anche se è basso il volume; mentre i non conformisti richiedono un maggior numero.
Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition	Intervista generica	274 studenti	Verificare se la presenza di informazioni personali nelle recensioni influenzano la credibilità della recensione.	La presenza di informazioni personali influenza positivamente la credibilità delle recensioni.

Tabella A.4: Summary articoli

Appendice B

Grafici valutazioni

In questa appendice saranno presenti i grafici che rappresentano i trend delle valutazioni prese dagli hotel, raggruppati per categorie.

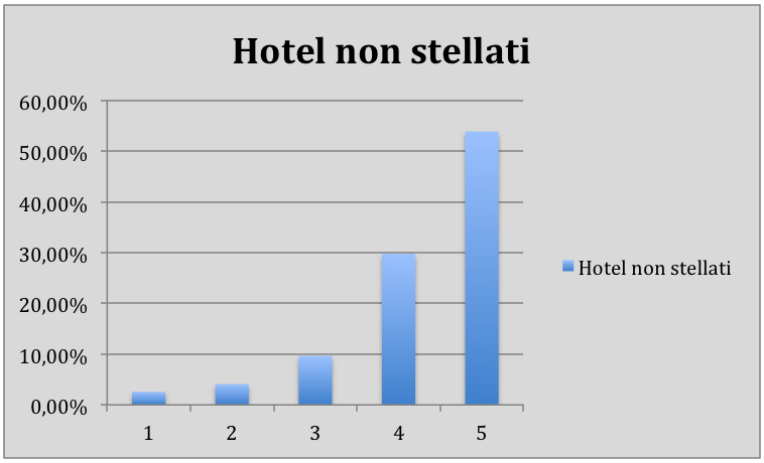

Figura B.1: Trend delle valutazioni degli hotel non stellati.

133

Figura B.2: Trend delle valutazioni degli hotel a 1 stella.

Figura B.3: Trend delle valutazioni degli hotel a 2 stelle.

Figura B.4: Trend delle valutazioni degli hotel a 3 stelle.

Figura B.5: Trend delle valutazioni degli hotel a 4 stelle.

Figura B.6: Trend delle valutazioni degli hotel a 5 stelle.

Bibliografia

[1] Alessandro Rezzani. *Big Data. Architettura, tecnologie e metodi per l'utilizzo di grandi basi di dati.* PerCorsi di Studio. Apogeo Education, Febbraio 2014. ISBN 8838789894.

[2] Wikipedia. Big data - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/Big_data.

[3] Michael de Waal-Montgomery. World's data volume to grow 40% per year 50 times by 2020: Aureus, Gennaio 2015. URL https://e27.co/worlds-data-volume-to-grow-40-per-year-50-times-by-2020-aureus-20150115-2/

[4] Accenture. *Big Success with Big Data.* Milano, 16 ottobre 2014. URL https://www.accenture.com/it-it/company-accenture-ricerca-big-data-big-success.aspx

[5] Treccani. DBMS - Treccani, Enciclopedie on line. URL http://www.treccani.it/enciclopedia/dbms/.

[6] Wikipedia. Database management system - wikipedia, l'enciclopedia libera, 2015. URL https://it.wikipedia.org/wiki/Database_management_system.

[7] CCM. I modelli di DBMS. Giugno 2014. URL http://it.ccm.net/contents/5-i-modelli-di-dbms.

[8] Zenas84's Blog. Principali DBMS presenti sul mercato, per l'archiviazione delle informazioni - Zenas84's Blog, log:"Software per la gestione dell'informazione, 2008/2009". URL https://zena84.wordpress.com/9-principali-dbms-presenti-sul-mercato-per-larchiviazione-delle-informazioni/.

[9] Wikipedia. Oracle - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/Oracle.

[10] Wikipedia. PostgreSQL - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/PostgreSQL.

[11] Wikipedia. SQLite - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/SQLite.

[12] Wikipedia. NoSQL - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/NoSQL.

[13] Knut Haugen. A brief history of nosql, 2010. URL https:/blog.knuthaugen.no/2010/03/a-brief-history-of-nosql.html.

[14] Matt Asay. Nosql databases eat into the relational database market, 2015. URL http://www.techrepublic.com/article/nosql-databases-eat-into-the-relational-database-market/.

[15] nosql-database.org. LIST OF NOSQL DATABASES. URL http://nosql-database.org/

[16] P. Atzeni, S. Ceri, S. Parabocchi, R. Torlone. Basi di Dati Distribuite. URL http://www.isa.cnr.it/dacierno/MaterialeDBUNISA1011/19_DBDISTRIBUITI.pdf

[17] HostingTalk.it. Introduzione alla scalabilità. http://www.hostingtalk.it/introduzione-alla-scalabilita_-c000000gN/

[18] Onofrio Panzarino. I database NoSQL. URL http://www.mokabyte.it/2011/03/nosql-1/

[19] Wikipedia. MongoDB - Wikipedia, l'enciclopedia libera. URL https://it.wikipedia.org/wiki/MongoDB

[20] Karl Seguin. Il Piccolo Libro di MongoDB. Seconda Edizione aggiornata a MongoDB 2.6. URL http://nicolaiarocci.com/mongodb/il-piccolo-libro-di-mongodb.pdf

[21] Wikipedia. Expedia - Wikipedia, l'enciclopedia libera. URL o https://it.wikipedia.org/wiki/Expedia.

[22] Santiago Melián-González, Jacques Bulchand-Gidumal, Beatriz González López-Valcárcel. *Online Customer Reviews of Hotels As Participation Increases, Better Evaluation Is Obtained*. Cornell Hospitality Quarterly, August 2013; vol. 54, 3: pp. 274-283. URL http://cqx.sagepub.com/content/54/3/274.full.

[23] Sun-Young Park, Jonathan P. Allen. *Responding to Online Reviews: Problem Solving and Engagement in Hotels*. Cornell Hospitality Quarterly, February 2013; vol. 54, 1: pp. 64-73. http://cqx.sagepub.com/content/54/1/64.full.

[24] Hyounae Min, Yumi Lim, Vincent P. Magnini. *Factors Affecting Customer Satisfaction in Responses to Negative Online Hotel Reviews*. Cornell Hospitality Quarterly, May 2015; vol. 56, 2: pp. 223-231. URL http://cqx.sagepub.com/content/56/2/223.full.

[25] Zheng Xiang, Muzaffer Uysal, Zvi Schwartz, John H. Gerdes Jr. *What can big data and text analytics tell us about hotel guest experience and satisfaction?*. International Journal of Hospitality Management 44 (2015) 120-130. URL http://www.sciencedirect.com/science/article/pii/S0278431914001698.

[26] Wei We, Li Miao, Zhuowei (Joy) Huang. *Customer engagement behaviors and hotel responses*. International Jour-

nal of Hospitality Management 33 (2013) 316-330. URL http://www.sciencedirect.com/science/article/pii/S027843191200134X.

[27] Giampaolo Viglia, Ladrón-de-Guevara, Roberto Furlan. *Please, talk about it! When hotel popularity boosts preferences.* International Journal of Hospitality Management 42 (2014) 155-164. URL http://www.sciencedirect.com/science/article/pii/S0278431914001194.

[28] Wenjing Duan, Yang Yu, Qing Cao, Stuart Levy. *Exploring the Impact of Social Media on Hotel Service Performance: A Sentimental Analysis Approach.* Cornell Hospitality Quarterly, 1938965515620483, first published on December 17, 2015. URL http://cqx.sagepub.com/content/early/2015/12/16/1938965515620483.full.

[29] Woo Gon Kim, Hyunjung Lim, Robert A. Brymer. *The effectiveness of managing social media on hotel performance.* International Journal of Hospitality Management 44 (2015) 165-171. URL http://www.sciencedirect.com/science/article/pii/S0278431914001704.

[30] Aurelio G. Mauri, Roberta Minazzi. *Web reviews influence on expectations and purchasing intentions of hotel potential customers.* International Journal of Hospitality Management 34 (2013) 99- 107.

[31] Inès Blal, Michael C. Sturman. *The Differential Effects of the Quality and Quantity of OnlineReviews on Hotel Room Sales.* Cornell Hospitality Quarterly, November 2014; vol. 55, 4: pp. 365-375. URL http://cqx.sagepub.com/content/55/4/365.full.

[32] Hui (Jimmy) Xie, Li Miao, Bo-Youn Lee. *Consumers' responses to ambivalent online hotel reviews: The role of perceived source credibility and pre-decisional disposition.* International Journal of Hospitality Management 30 (2011) 178-183. URL http://www.sciencedirect.com/science/article/pii/S0278431910000563

[33] Edwin N. Torres, Dipendra Singh, April Robertson-Ring. *Consumer reviews and the creation of booking transaction value: Lessons from the hotel industry.* International Journal of Hospitality Management 50 (2015) 77- 83. URL http://www.sciencedirect.com/science/article/pii/S0278431915001127.

[34] Karen L. Xie, Zili Zhang, Ziqiong Zhang. *The business value of online consumer reviews and management response to hotel performance.* International Journal of Hospitality Management 43 (2014) 1- 12. URL http://www.sciencedirect.com/science/article/pii/S027843191400125X.

[35] Bing Pan, Lixuan Zhang, Rob Law. *The Complex Matter of Online Hotel Choice.* Cornell Hospitality Quarterly, February 2013; vol. 54, 1: pp. 74-83., first published onOctober 30, 2012. URL http://cqx.sagepub.com/content/54/1/74.full.

[36] Ajay Aluri, Lisa Slevitch, Robert Larzelere. *The Influence of Embedded Social Media Channels on Travelers' Gratifications, Satisfaction, and Purchase Intentions.* Cornell Hospitality Quarterly, 1938965515615685, first published on December 28, 2015. URL http://cqx.sagepub.com/content/early/2015/12/25/1938965515615685.full.

[37] Rex S. Toh, Charles F. DeKay, Peter Raven. *Travel Planning: Searching for and Booking Hotels on the Internet.* Cornell Hospitality Quarterly, November 2011; vol. 52, 4: pp. 388-398., first published on September 1, 2011. URL http://cqx.sagepub.com/content/52/4/388.full.pdf+html.

[38] Markus Schuckert, Rob Law, Xianwei Liu. *A segmentation of online reviews by language groups: How English and non-English speakers rate hotels differently.* International Journal of Hospitality Management 48 (2015) 143- 149. URL http://www.sciencedirect.com/science/article/pii/S0278431914001935.

[39] Wen-Chin Tsao, Ming-Tsang Hsieh, Tom M.Y. Lin, Li-Wen Shih. *Compliance with eWOM: The influence of hotel reviews on boo-*

king intention from the perspective of consumer conformity. International Journal of Hospitality Management 46 (2015) 99-111. URL http://www.sciencedirect.com/science/article/pii/S0278431915000158

[40] php.net. Che cos'è il PHP?. URL http://php.net/manual/it/intro-whatis.php.

[41] MDM, Mozilla Developer Network. JavaScript. URL https://developer.mozilla.org/it/docs/Web/JavaScript.

[42] HTML.it. Introduzione a Node.js. URL http://www.html.it/pag/32814/introduzione-a-nodejs/.

Ringraziamenti

Sono parecchie le persone che vorrei ringraziare.
In primis il mio relatore, Marco Di Felice, che ha accettato la mia idea di tesi e ne ha fatto molto di più, dandomi svariati consigli e ampliando le mie idee di progetto fino a far diventare la mia tesi un elaborato di cui andar fiero. Grazie.
Subito dopo voglio ringraziare i miei genitori che per tutti questi anni mi hanno aiutato, mi sono stati vicini e mi hanno saputo consigliare sempre al meglio; ma soprattutto mi hanno fatto diventare quello che sono e se sono riuscito a superare tutti gli ostacoli fino ad arrivare a questo grande traguardo è solo grazie a loro. Grazie.
Un ringraziamento particolare va ai miei cugini Matteo, Mattia, Gaia e Vittoria che essendo figlio unico considero un po' come fratelli e molto di più. In particolare il maggiore, Matteo che fin da piccolo l'ho sempre visto come un modello da seguire, un modello a cui mai riuscirò ad eguagliare ma che mi da la forza di migliorare ogni giorno di più; e Mattia, tra i maschi il più giovane, ma dei tre penso il più forte, il più pazzo, lui è un amico, un compagno e so che ci sarà sempre per qualsiasi cosa. Grazie.
Ma ringrazio anche tutti i miei famigliari che in questi tre anni di università mi hanno sempre fatto capire quanto erano orgogliosi di me, congratulandosi per ogni risultato che ottenevo. Soprattutto la mia nonna, che nonostante l'età resiste e continua a dirmi che sono il suo "ciciu", proprio come mi chiamava il nonno quando ancora era tra di noi. Grazie.

Un altro grandissimo ringraziamento va alla mia fidanzata, Chiara, una persona meravigliosa, forse addirittura troppo per me; grazie perché in questo anno passato assieme mi hai insegnato tante cose, mi hai dato la giusta grinta per risolvere ogni problema, mi hai dato l'affetto per non sentirmi mai solo e sei sempre rimasta al mio fianco. Sei stata la giusta motivazione per concludere al meglio questo percorso di studi. Grazie per avermi fatto sentire importante.

Infine volevo ringraziare tutti i miei amici.
Grazie, a quelli che mi hanno alzato il morale tra una lezione e l'altra portandomi a mangiare il sushi.
Grazie, agli amici "lontani", che nonostante la distanza hanno fatto si che la nostra amicizia non sia mai venuta meno, ma anzi si è rilevata un amicizia sempre più forte, sempre più vera. Vi voglio bene.
E infine un grazie di cuore agli amici "vicini" che nonostante io fossi impegnato con gli studi, non si sono mai dimenticati di me, e mi hanno fatto sentire parte di loro anche quando non potevo esserci. Grazie.