Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica

TARO-TZ: confronto tra testate giornalistiche online in diversi fusi orari

Relatore:
Chiar.mo Prof.
ANGELO DI IORIO

Presentata da:
DANIELE ROMANELLA

I Sessione
Anno Accademico 2023/2024

Ai giganti

Indice

1 Introduzione **1**

2 Related works **3**
2.1 Background: cos'è "TARO" 3
2.1.1 Reperimento delle notizie 4
2.1.2 Snapshots 4
2.1.3 Fonti eterogenee 5
2.2 Altri studi riguardo la similarità testuale 6
2.2.1 News clustering based on similarity analysis 6
2.2.2 An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts 6
2.2.3 Building hypertext links in newspaper articles using semantic similarity 7

3 Taro-TZ: un' estensione di TARO per gestire fusi orari diversi **8**
3.1 Estensione del Modello e idea delle modifiche 9
3.1.1 Aree di interesse dello studio 11
3.2 Analisi dei giornali: aggiornamento delle notizie 11
3.2.1 Aggiornamento delle home page 11
3.2.2 Aggiornamento delle pagine tematiche 14

4 Implementazione **17**
4.1 Lettura dei feed RSS 18
4.1.1 Problemi relativi ai feed RSS 18
4.2 Scraping 18
4.2.1 Problemi relativi allo scraping 19
4.2.2 Utilizzo di Scrapy 19
4.2.3 Estensione delle informazioni delle notizie all'interno degli snapshot 22
4.2.4 Introduzione di nuovi scraper 22

4.2.5 Salvataggio delle notizie 23
4.3 Pipeline 23
4.3.1 Traduzione 24
4.4 Confronto delle notizie 26
4.4.1 Creazione degli insiemi da comparare 26
4.4.2 Calcolo della similarità 29
4.4.3 Creazione dei grafici 30
4.4.4 Rilevamento delle notizie in home page e nelle aree tematiche 30

5 Esperimenti **32**
5.1 Legende e letture dei grafici 33
5.1.1 Struttura dei grafici 34
5.2 Analisi delle home page 35
5.2.1 Analisi delle home page, esclusione di alcuni giornali 36
5.2.2 Comparazione binaria di AGI e ANSA 39
5.3 Analisi delle aree tematiche 39
5.3.1 Economia 39
5.3.2 Esteri 41

6 Conclusioni **44**
6.1 Sviluppi futuri 44

Riferimenti bibliografici **46**

Elenco delle figure

2.1 Funzionamento del modello di TARO, le linee blu tratteggiata e quella rossa a puntini indicano la presenza di due notizie distinte all'interno degli snapshot.
Immagine ispirata all'articolo riguardante TARO [CDlB23] 4
2.2 Struttura di uno snapshot. Immagine copiata dalla tesi di Giuseppe Carrino [Car] 5

3.1 **Ipotesi A**. Le notizie vengono pubblicate quasi simultaneamente nel mondo, rendendo non ostacolanti le differenze dei fusi orari fra i vari luoghi del mondo 9
3.2 **Ipotesi B**. Le notizie subiscono rallentamenti nella pubblicazione a causa dei fusi orari. 9
3.3 **Ipotesi A**- Si suppongano intervalli orari di 3 ore. Iniziando dalle ore 00 alle ore 02:59:59, dalle ore 03:00:00 alle ore 05:59:59 e così via fino all'ultimo intervallo falle ore 21:00:00 alle ore 23:59:59. Nell'immagine viene mostrato, attraverso i colori, quali sono le notizie che verrebbero confrontate in base all'orario di rilevamento 10
3.4 **Ipotesi B**- Si suppongano gli stessi intervalli di prima. È mostrato come cambierebbe la situazione se considerassimo l'ora locale. 10
3.5 Home page di ANSA alle ore 9:30 del giorno 20/06/2024 12
3.6 Home page di ANSA alle ore 11:30 del giorno 20/06/2024, sono evidenziate le notizie nuove 13
3.7 Pagina economica di ANSA alle ore 9:30 del giorno 20/06/2024 14
3.8 Pagina economica di ANSA alle ore 11:30 del giorno 20/06/2024 sono evidenziate le notizie nuove in rosso ed in blu quelle presenti precedentemente 15
3.9 Pagina economica di ANSA alle ore 11:30 del giorno 20/06/2024 sono evidenziate le notizie nuove in rosso ed in blu quelle presenti precedentemente 16

4.1 Visualizzazione grafica di come tutti i passaggi devono essere eseguiti 17

4.2 Lista degli URL utilizzati dallo scraper per iniziare il rilevamento delle notizie . . . 20
4.3 Struttura di una notizia in uno snapshot dopo l'estensione degli spiders. 23
4.4 Struttura del filesystem . . . 24
4.5 Una notizia salvata in formato JSON. Sono stati mantenuti solo i campi originali e quelli tradotti per questioni di brevità, per la stessa ragione il contenuto di alcuni campi è stato troncato. . . . 25
4.6 Passaggi logici da compiere per confrontare le notizie e raggiungere l'obiettivo prefissato da questa tesi. . . . 26
4.7 Associazione dei temi con le loro parole chiave . . . 31

5.1 Una mappa che indica tutti i luoghi in cui sono pubblicati i giornali . 33
5.2 Analisi con un unico frame temporale di ventiquattro ore. . . . 34
5.3 Analisi con frame temporali di sei ore. . . . 35
5.4 Analisi con frame temporali di sei ore, visualizzazione dei rapporti . . 35
5.5 Analisi con frame temporali di sei ore, escludendo ANSA . . . 36
5.6 Analisi con frame temporali di sei ore, escludendo ANSA, visualizzazione dei rapporti . . . 37
5.7 Analisi con frame temporali di sei ore, escludendo tutti i giornali italiani. 38
5.8 Analisi con frame temporali di sei ore, escludendo tutti i giornali italiani, visualizzazione dei rapporti. . . . 38
5.9 Comparazione delle home page di AGI ed ANSA. . . . 39
5.10 Analisi della sezione di economia dei giornali con frame temporali di sei ore. . . . 40
5.11 Analisi della sezione di economia dei giornali con frame temporali di sei ore, visualizzazione dei rapporti . . . 41
5.12 Analisi della sezione esteri dei giornali con frame temporali di sei ore. 42
5.13 Analisi della sezione esteri dei giornali con frame temporali di sei ore, visualizzazione dei rapporti. . . . 42
5.14 Analisi della sezione esteri con frame temporali orari. . . . 43
5.15 Analisi della sezione esteri con frame temporali orari, visualizzazione dei rapporti. . . . 43

Frammenti di codice

4.1 Metodo scrapy.Spider::parse di cui si è fatto l'override nello scraper in esame. 20
4.2 Definizione del metodo parseArticle 21
4.3 Funzionamento del controllo che consente di saltare alcuni file per analizzare solo quelli effettivamente utili. 27
4.4 Controllo degli articoli per verificarne la pre-elaborazione dalla pipeline 27
4.5 Controllo degli articoli per verificarne presenza nel range indicato . . 28
4.6 Trasformazione di una stringa in uno spacy Doc 29
4.7 Calcolo della similarità tra due notizie dato un threshold. 29

Capitolo 1

Introduzione

Questa tesi vuole estendere un lavoro precedente già sviluppato, chiamato "TARO". TARO è un software che si occupa di confrontare testate giornalistiche, e per fare ciò, svolge un'operazione di mappatura degli articoli per capire come i vari giornali trattano le notizie. Nel modello utilizzato le analisi sono effettuate esclusivamente attraverso l'ora di rilevamento delle notizie. Da questo, la necessità di espandere il modello già esistente dando una nuova dimensione allo spazio delle analisi, questa estensione sarà utile per capire se la diffusione delle notizie è influenzata dai fusi orari dei vari paesi oppure no, in particolare saranno analizzati i timestamp di rilevamento delle notizie e l'ora nel fuso orario della nazione[1] in cui sono pubblicate.
Nel capitolo 2 - Related Works, sarà analizzato prima il modello di TARO e successivamente saranno analizzati altri modelli di similarità creati da altri ricercatori. Sono presenti poche ricerche che studiano di come i fusi orari impattino con la pubblicazione di contenuti, tuttavia, è stata rilevata una ricerca che spiega in maniera abbastanza chiara di come, in realtà, i fusi orari, per studiare i comportamenti delle community online, siano un limite [Wie11].
Nel capitolo 3 - Una possibile soluzione, sarà mostrata un'idea di come il problema può essere affrontato con l'ausilio di grafici ed immagini estratte dal giornale "ANSA", inoltre questo capitolo è utile per mostrare i due approcci che saranno utilizzati per gli esperimenti con l'ausilio di immagini e linee temporali.
Nel capitolo 4 - Implementazione, sarà mostrata e spiegata una parte significativa del codice e di come questo è stato strutturato per raggiungere l'obiettivo che ci si è posti, inoltre, saranno analizzati anche diversi problemi riguardanti il reperimento delle notizie via internet, inoltre saranno illustrate quali sono state le fonti scelte e il relativo fuso orario.

[1] Oppure una parte della nazione, come, per esempio, negli Stati Uniti d'America sono presenti diversi fusi orari

Nel capitolo 5 - Esperimenti saranno visualizzati i risultati dell'esecuzione del codice scritto precedentemente, il capitolo, corredato di grafici e spiegazioni di questi, mostrerà quanto in realtà il problema sia significativo.
Infine, nel capitolo 6 - Conclusioni, saranno presenti le considerazioni finali sul lavoro svolto, spiegando le motivazioni dietro i risultati ottenuti e dando spunti per eventuali sviluppi futuri.

Capitolo 2

Related works

2.1 Background: cos'è "TARO"

TARO (Tons of Articles Ready to Outline) è un modello sviluppato da **un gruppo di ricerca dell'Università di Bologna** [CDlB23] [Car] che nasce con l'intento di analizzare le testate giornalistiche, per capire quali notizie sono sono uguali ad altre e quali notizie sono "saltate" da alcuni giornali piuttosto che da altri.
Il modello originale analizza i giornali relativamente al momento in cui le notizie sono rilevate in degli *snapshot* e per le analisi non tiene conto dell'ora locale in cui un articolo è pubblicato. Inoltre, per individuare le notizie simili, il programma, sfrutta la nozione di "similarità", ovvero quanto le notizie sono tra loro somiglianti.
La **Figura 2.1** spiega il funzionamento del vecchio modello di TARO, mettendo in evidenza, come sia considerata solo la data di rilevamento della notizia, detta anche *Scraping time*.
Il software risolve il problema di trattare notizie provenienti da nazioni con lingue tra loro difformi. Essendo tra di loro indipendenti, spesso le testate giornalistiche pubblicano i propri contenuti nei momenti che ritengono più opportuni. In base alla regolarità delle notizie è possibile distinguere fra testate ad **Edizione** e testate a **Flusso**:

- Le testate a **Edizione** sono tutte quelle testate giornalistiche che pubblicano i propri contenuti ad orari prefissati. Ad esempio ogni giorno alle 20 oppure alle 14 e tipicamente si tratta di notiziari pensati per essere trasmessi in TV.
- Le testate a **Flusso** sono tutte quelle testate giornalistiche che pubblicano i propri contenuti in momenti arbitrari, spesso la loro pubblicazione avviene sul proprio sito WEB o su eventuali feed RSS.
  In questo caso il software sviluppato, effettua degli "snapshot" ad intervalli

cadenzati per memorizzare la situazione del giornale in determinati momenti della giornata.

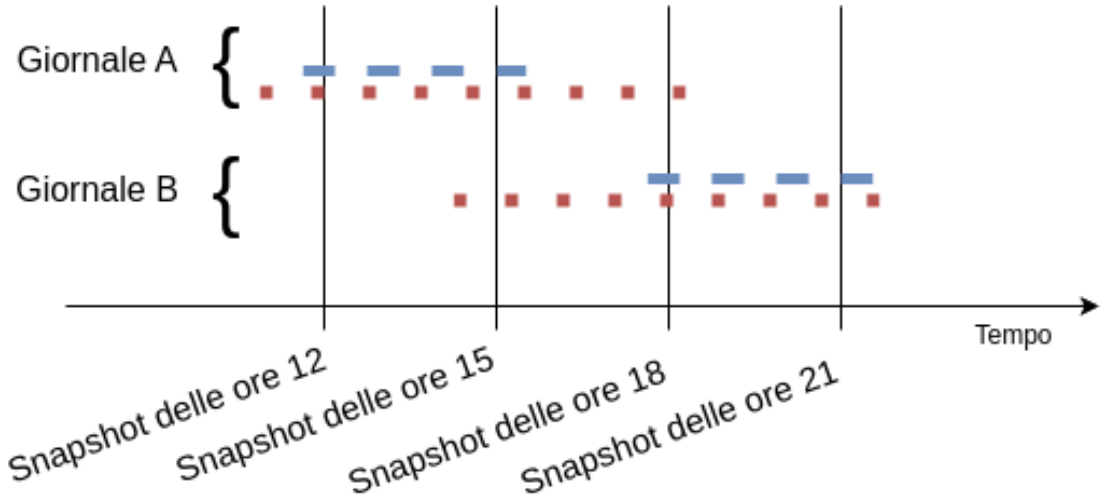

Figura 2.1: Funzionamento del modello di TARO, le linee blu tratteggiata e quella rossa a puntini indicano la presenza di due notizie distinte all'interno degli snapshot.
Immagine ispirata all'articolo riguardante TARO [CDlB23]

2.1.1 Reperimento delle notizie

Le notizie sono state reperite attraverso:

- Scraping: una tecnica che sfrutta software per estrarre delle informazioni, tipicamente del testo, da una pagina web.
  Spesso i software utilizzati per questa tecnica simulano la navigazione sul documento che contiene le informazioni da estrarre.
- Feed RSS: un flusso di informazioni in formato XML utilizzato per scambiare le informazioni fra diverse piattaforme o applicazioni.
  Questo approccio ha un problema: spesso capita che i feed non vengano aggiornati per tanto tempo. Quindi, è preferibile effettuare lo scraping, se possibile[1].

2.1.2 Snapshots

Gli snapshots sono insiemi di notizie catturate in un dato momento e memorizzate in un file. I file hanno come nome il datetime del rilevamento e, sono divisi in cartelle. Ogni notizia all'interno dello snapshot contiene molte informazioni, tra cui:

[1]Lo scraping non sempre è permesso.

- Titolo della notizia
- Data raw: data di pubblicazione ottenuta dal documento WEB
- Data: data di pubblicazione
- Url: contiene un link HTTP alla pagina WEB che punta alla pagina dell'articolo
- News url: contiene un link HTTP all'articolo
- Sottotitolo
- Contenuto
- Lingua: codice della lingua in formato ISO 639
- Sorgente della notizia


```
[
    {
        "title": "Comment expliquer le coup de chaleur sans précédent des régions polaires ?",
        "date_raw": "Wed, 23 Mar 2022 16:38:47 GMT",
        "date": "2022-03-23",
        "url": "https://www.france24.com/fr/plan%C3%A8te/rss",
        "news_url": "https://www.france24.com/fr/%C3%A9co-tech/
        20220323-comment-expliquer-le-coup-de-chaleur-sans-pr%C3%A9c%C3%A9dent-des-r%C3%A9gions-polaires",
        "subtitle": "En fin de semaine dernière, les températures ont battu tous les records de chaleur aussi bien en Antarctique qu'en Arctique. ...",
        "content": "Des températures allant jusqu' et entre 20 et 30 °C de plus que d'habitude à certains endroits en Arctique. ...",
        "ranked": 2,
        "placed": "Abroad",
        "epoch": 1648319356.3339121,
        "language": "FR",
        "source": "France24",
        "en_title": "How can we explain the unprecedented heat cut in polar regions?",
        "en_content": "Temperatures up to and between 20 and 30 °C more than usual in some locations in the Arctic. ...",
        "en_subtitle": "In the end of last week, temperatures broke all heat records in both Antarctica and the Arctic. ...",
    },
    {
        "title": "Le combat des ONG pour faire exister le climat dans la campagne présidentielle",
        "date_raw": "Sat, 12 Mar 2022 08:56:50 GMT",
        "date": "2022-03-12",
        "url": "https://www.france24.com/fr/plan%C3%A8te/rss",
        "news_url": "https://www.france24.com/fr/france/20220312-le-combat-des-ong-pour-faire-exister-le-climat-dans-la-campagne-pr%C3%A9sidentielle",
        "subtitle": "Sujet de préoccupation majeur chez les Français et enjeu planétaire, les questions climatique et environnementale sont absentes ...",
        "ranked": 8,
        "placed": "Abroad",
        "epoch": 1648319356.4149504,
        "language": "FR",
        "source": "France24",
        "en_title": "The fight of NGOs to make the climate exist in the presidential campaign",
        "en_content": "and the media space. In the midst of the debates on Vladimir Putin on NATO, economic sanctions or European defence, ...",
        "en_subtitle": "Subject of major concern among French and global issues, climate and environmental issues are missing ...",
    }
]
```


Figura 2.2: Struttura di uno snapshot. Immagine copiata dalla tesi di Giuseppe Carrino [Car]

2.1.3 Fonti eterogenee

Nel lavoro svolto precedentemente erano presenti un certo numero di fonti eterogenee, sia per modalità di pubblicazione[2], sia per lingua.
In particolare sono presenti: quattro testate in lingua inglese, tre in tedesco, due

[2]Testate ad edizione ed a flusso

in italiano, due in francese ed una in spagnolo. Per normalizzare le notizie è stata utilizzata una libreria di python chiamata argostranslate[3][4].

2.2 Altri studi riguardo la similarità testuale

Di seguito una raccolta di studi che riguardano la similarità fra notizie testuali.

2.2.1 News clustering based on similarity analysis

News clustering based on similarity analysis [BA17] è uno studio effettuato come ausilio per raggiungere un altro obiettivo, quello di effettuare la ricerca sull'analisi della guerra psicologica su Internet.
Nello studio viene spiegato come viene calcolata la similarità per poi creare dei *Cluster* che trattano di notizie simili. I ricercatori hanno utilizzato WordNet che contiene synsets: raggruppamenti di parole con lo stesso significato semantico. Inoltre, prima di eseguire qualunque confronto, vengono rimosse le stopword [5] e successivamente si procede ad assegnare ad ogni parola un significato semantico (soggetto, complemento oggetto, ...) ed un significato funzionale (verbo, nome, ...). I risultati mostrati nell'articolo sono tutto sommato accettabili anche se nei cluster mostrati è possibile notare alcune criticità.

2.2.2 An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts

An Architecture for Efficient Document Clustering and Retrieval on a Dynamic Collection of Newspaper Texts, [AFSea98], è un articolo nel quale i ricercatori mostrano come è possibile creare dei cluster che contengono notizie che trattano dello stesso argomento. L'idea è relativamente semplice: si tratta di cluster organizzati secondo una struttura ad albero. Il software, usando una matrice di similarità, collega innanzitutto le notizie che ritiene più simili creando dei minicluster. Durante questo processo alcune notizie potrebbero rimanere non associate quindi è possibile che alcune notizie non vengano inserite nessun insieme. Nell'articolo, inoltre, viene mostrato come sia possibile creare matrici di similarità di dimensione pari a $N \times k$,

[3] `https://github.com/argosopentech/argos-translate`

[4] Argostranslate è una libreria che utilizza OpenNMT per le traduzioni e può essere utilizzata in molti modi diversi: sia come libreria python, sia da riga di comando e anche come applicativo con GUI.

[5] Stop words: Parole che sono presenti nella frase ma non apportano significato semantico, come ad esempio, gli articoli e le congiunzioni

dove, N è il numero degli articoli e k costante tale che $k << N$. Questo consente di evitare la creazione di una matrice $N \times N$ così da poter risparmiare molte risorse computazionali. I risultati di questo studio si mostrano soddisfacenti.

2.2.3 Building hypertext links in newspaper articles using semantic similarity

L'articolo Building hypertext links in newspaper articles using semantic similarity [G+97] propone di risolvere un problema ben preciso, utilizzando la nozione di similarità come nel caso di questa tesi. L'idea è quella di creare un link ipertestuale fra due paragrafi di articoli diversi nel caso in cui si rilevi che questi sono semanticamente vicini. L'analisi viene effettuata per mezzo delle catene lessicali che, come spiegate nell'articolo, sono sequenze di parole semanticamente legate fra loro. Quindi, una volta pre-elaborate le catene lessicali, si procede al calcolo della similarità fra due catene estratte da due documenti distinti. Il calcolo della similarità viene effettuato elaborando i density vectors[6] e, successivamente, dopo aver normalizzato i vettori, si procede a calcolare una matrice di similarità fra tutti i paragrafi utilizzando la similarità fra le catene lessicali che li compongono. Infine, posto un threshold, si decide se è opportuno creare un link fra i due paragrafi oppure no.

[6]L'importanza di una certa catena lessicale all'interno dell'articolo da cui è estratta

Capitolo 3

Taro-TZ: un' estensione di TARO per gestire fusi orari diversi

L'idea alla base di questa tesi è molto semplice: si vuole capire se i fusi orari incidono sul numero di notizie. In particolare, dato un orario prestabilito globale[1], si vuole capire che cosa accade in termini numerici al numero di notizie ritenute simili. Innanzitutto si vogliono illustrare due possibili ipotesi con l'ausilio delle linee del tempo, come mostrato in **Figura 3.1** e **Figura 3.2**. Queste due immagini vogliono farci capire cosa potrebbe accadere ad una notizia nel momento in cui viene riportata da più testate giornalistiche site in fusi orari diversi.
Ipotesi A: Il flusso delle notizie non è influenzato dai fusi orari.
Ipotesi B: Le notizie sono influenzate dai fusi orari e la differenza dei tempi di pubblicazione, a seconda del fuso orario, è considerevole. Questo approccio potrebbe essere più appropriato per notizie che, nonostante il giornale sia a flusso, sono pubblicate all'incirca allo stesso orario locale per tutte le testate giornalistiche, come gli articoli riguardanti lo spettacolo, commenti politici e commenti economici.

[1] L'orario scelto è stato il tempo di scraping, ovvero, il datetime del server che effettua le analisi. Per tutte le analisi effettuate, lo scraping time appartiene alla timezone CEST (Central European Summer Time)

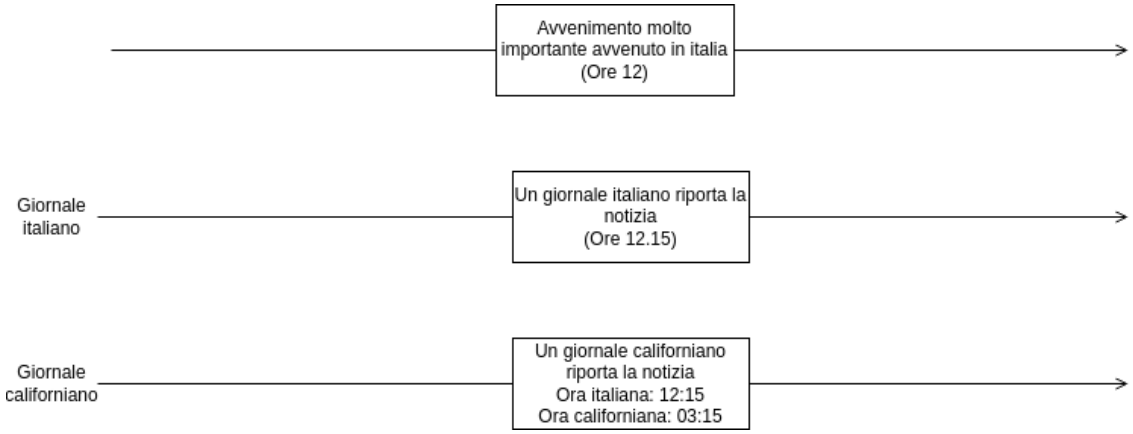

Figura 3.1: Ipotesi A. Le notizie vengono pubblicate quasi simultaneamente nel mondo, rendendo non ostacolanti le differenze dei fusi orari fra i vari luoghi del mondo

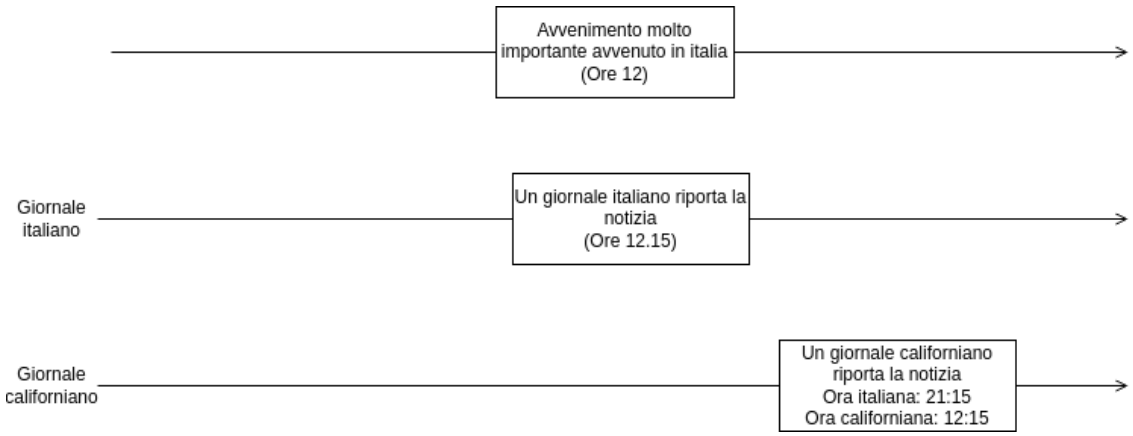

Figura 3.2: Ipotesi B. Le notizie subiscono rallentamenti nella pubblicazione a causa dei fusi orari.

3.1 Estensione del Modello e idea delle modifiche

L'obiettivo di questa Tesi è estendere il modello di TARO aggiungendo una nuova dimensione al modo di effettuare la comparazione delle notizie. I nuovi gruppi di notizie saranno analizzati in base a due orari, quindi l'ora locale del giornale e l'ora prefissata globale di cui si è discusso discusso precedentemente. Viene mostrato, nella **Figura 3.3** e nella **Figura 3.4** un esempio giocattolo che **riporta notizie inventate** al fine di mostrare l'idea dietro le due ipotesi esplicitate precedentemente, inoltre le figure di cui sotto, mostrano come le notizie vengono confrontate, non si sta ponendo il focus sulla similarità tra le notizie. Inoltre, l'ora locale è l'orario nel fuso orario della zona in cui viene pubblicato il giornale nel momento in cui la notizia viene rilevata, da questo momento in poi ci riferiremo a questo concetto come " *Ora*

locale".

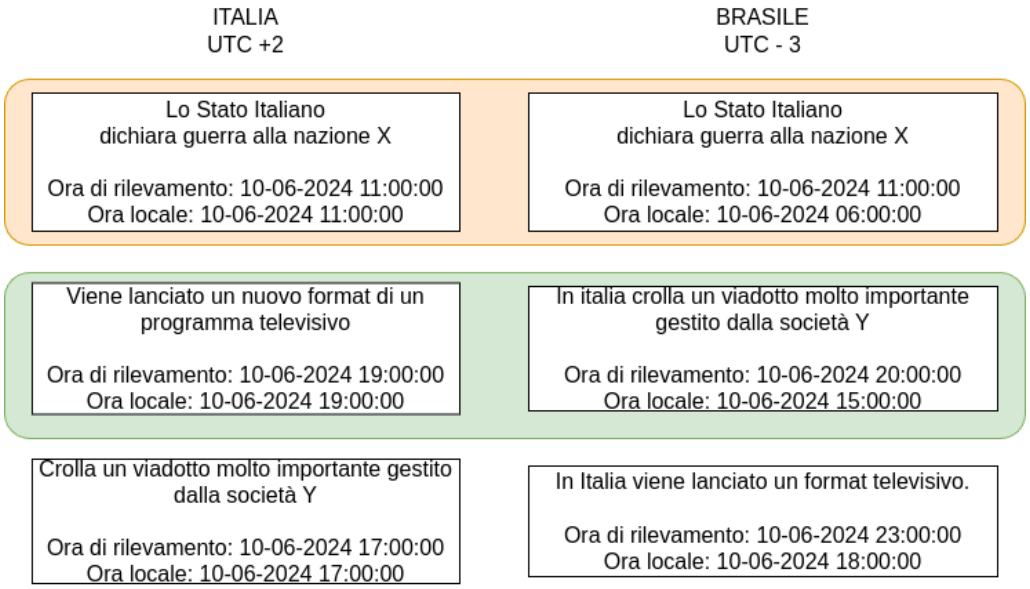

Figura 3.3: Ipotesi A- Si suppongano intervalli orari di 3 ore. Iniziando dalle ore 00 alle ore 02:59:59, dalle ore 03:00:00 alle ore 05:59:59 e così via fino all'ultimo intervallo falle ore 21:00:00 alle ore 23:59:59. Nell'immagine viene mostrato, attraverso i colori, quali sono le notizie che verrebbero confrontate in base all'orario di rilevamento

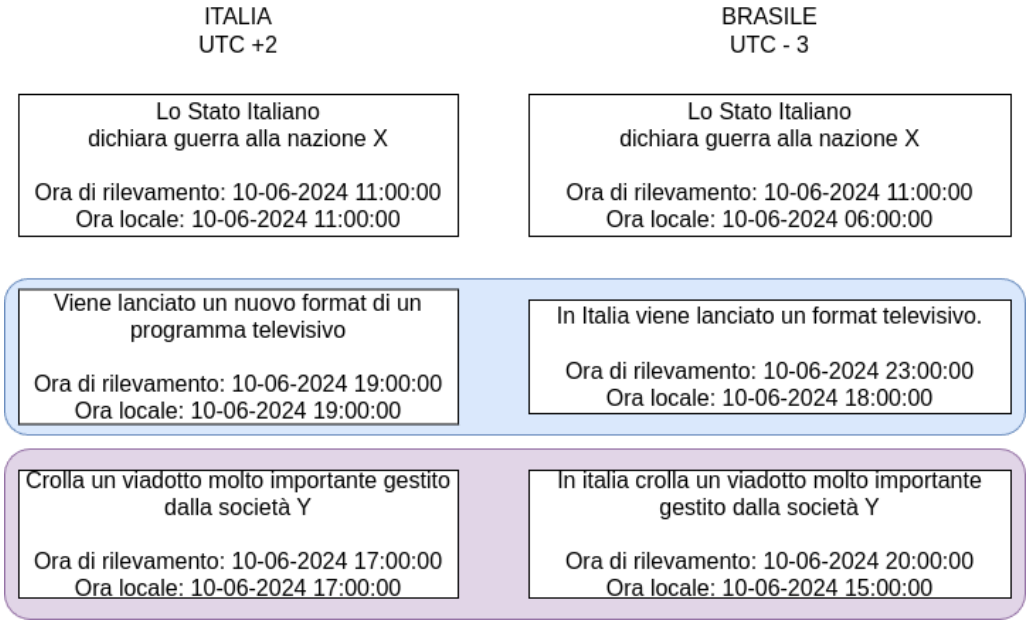

Figura 3.4: Ipotesi B- Si suppongano gli stessi intervalli di prima. È mostrato come cambierebbe la situazione se considerassimo l'ora locale.

3.1.1 Aree di interesse dello studio

Lo studio in questione è stato effettuato principalmente sulle pagine principali dei giornali e su alcune aree tematiche trattate da questi, come ad esempio:

- Esteri
- Economia
- Sport
- Cultura
- Politica
- Tecnologia

Inoltre, in questa tesi sono state analizzate solamente le testate giornalistiche a **Flusso**.

3.2 Analisi dei giornali: aggiornamento delle notizie

I giornali a flusso, aggiornano le loro notizie con una frequenza non sempre costante. Può capitare che i giornali modifichino più frequentemente la home page delle loro pagine tematiche.

3.2.1 Aggiornamento delle home page

Le home page sono le pagine principali dei siti web, nel caso della tesi in esame, sono le pagine web principali dei vari giornali. Si vuole porre particolare attenzione a come le notizie cambiano, questo sarà utile quando vengono effettuati confronti fra le varie notizie. Nella **Figura 3.5** si mostra l'home page di ANSA e nella **Figura 3.6** è rimostrata ancora la stessa pagina ma dopo che sono trascorse due ore. Nella seconda immagine vengono messe in evidenza le notizie che non sono presenti nella prima figura.

Figura 3.5: Home page di ANSA alle ore 9:30 del giorno 20/06/2024

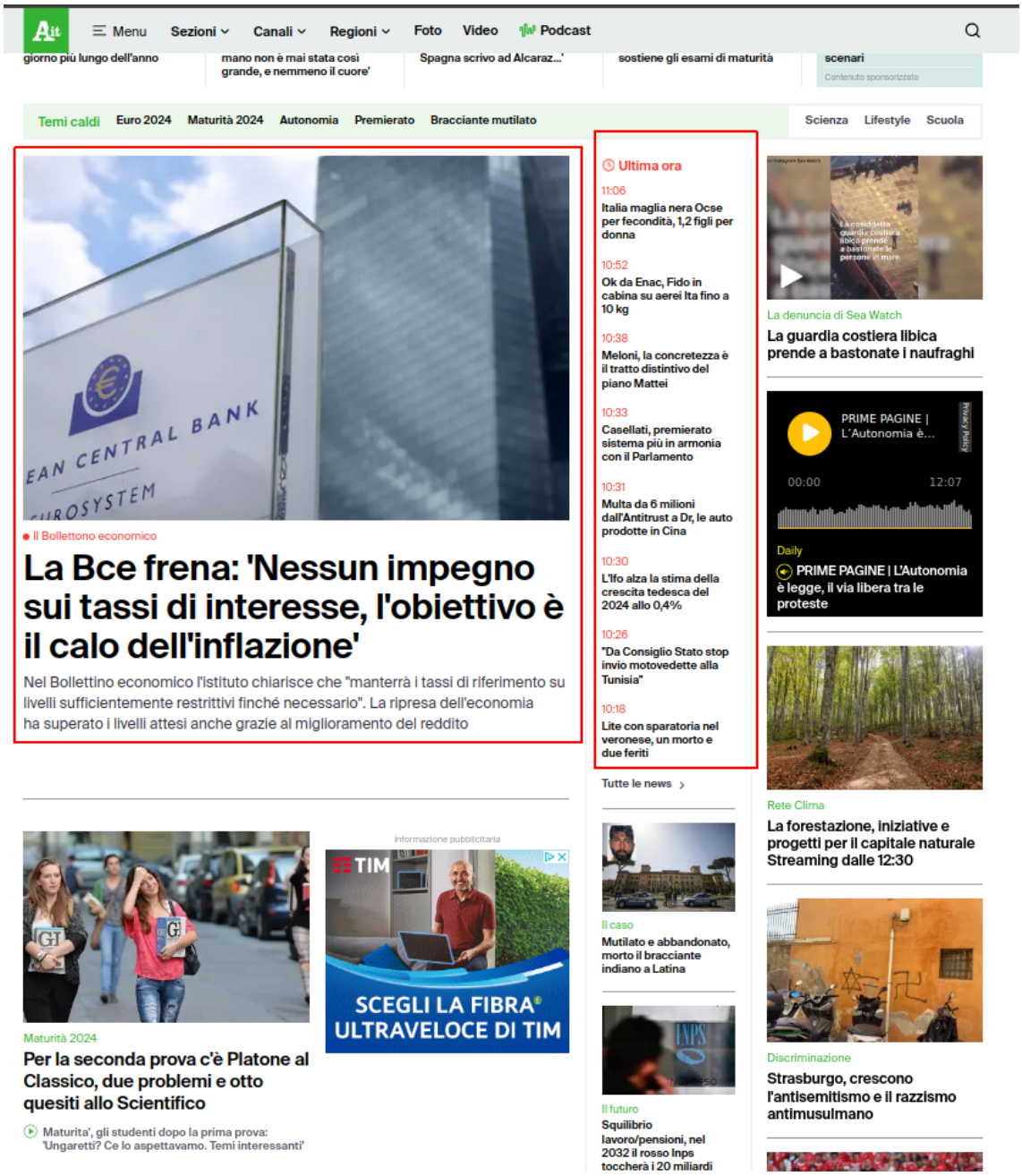

Figura 3.6: Home page di ANSA alle ore 11:30 del giorno 20/06/2024, sono evidenziate le notizie nuove

3.2.2 Aggiornamento delle pagine tematiche

Le pagine tematiche sono delle pagine dei siti web in cui si parla specificatamente di un solo tema, come ad esempio: economia, sport, politica, esteri, eccetera. È anche interessante notare come le varie aree tematiche differiscano anche per il numero di notizie ritenute simili. Analogamente per quanto fatto per la home page, si prenda in esame la sezione economia di ANSA. Nella **Figura 3.7** viene mostrata la pagina economica di ANSA alle ore 9:30, successivamente nella **Figura 3.8** viene la stessa pagina ma alle ore 11:30. In questo caso si nota che facendo scrolldown ci sono molte notizie precedentemente pubblicate che ancora sono presenti nella pagina e questo è visibile nella **Figura 3.9**

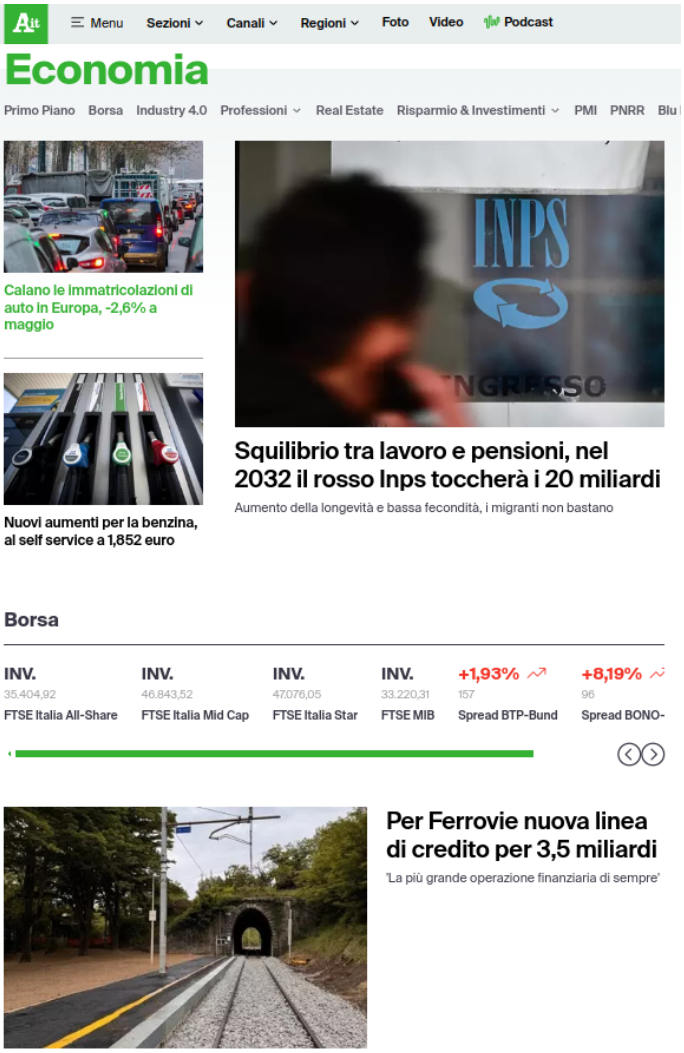

Figura 3.7: Pagina economica di ANSA alle ore 9:30 del giorno 20/06/2024

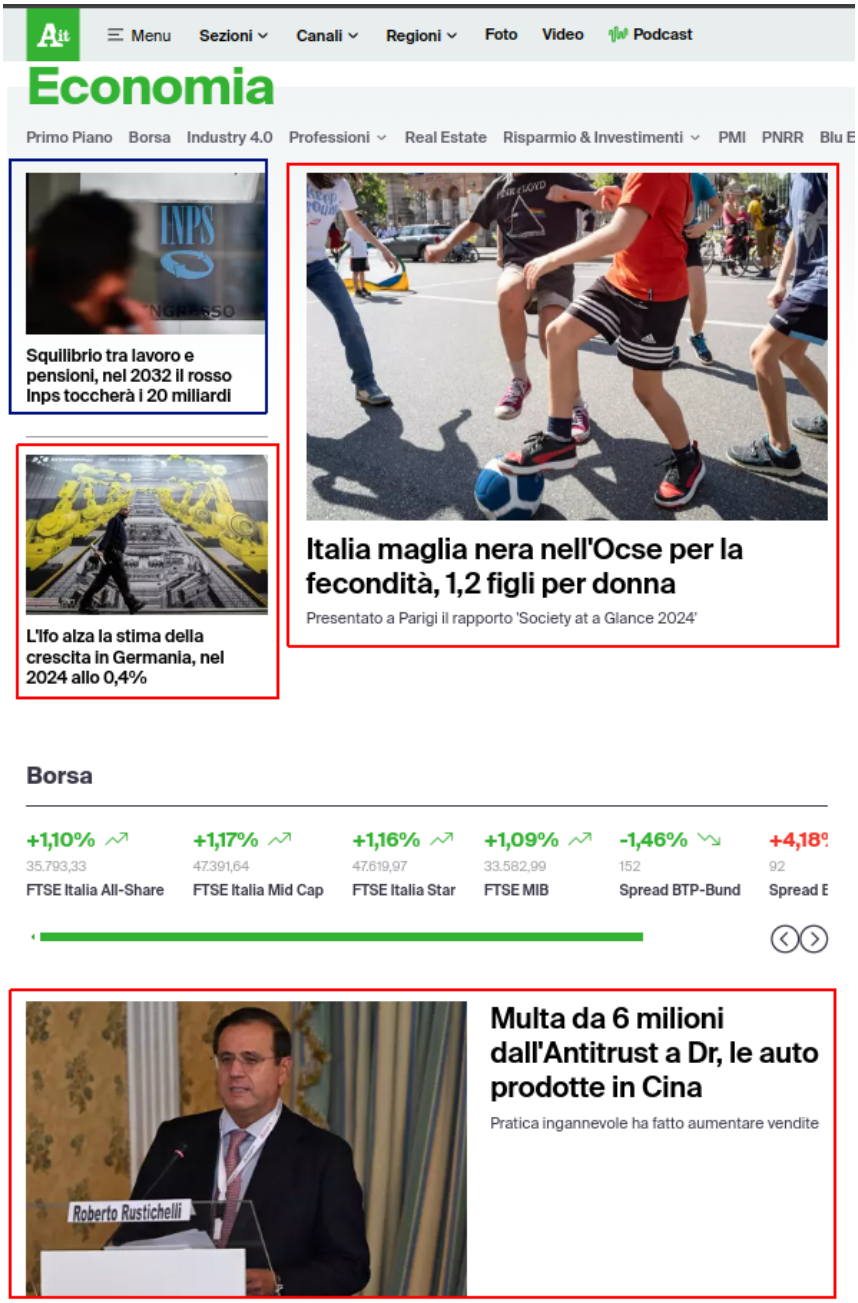

Figura 3.8: Pagina economica di ANSA alle ore 11:30 del giorno 20/06/2024 sono evidenziate le notizie nuove in rosso ed in blu quelle presenti precedentemente

Figura 3.9: Pagina economica di ANSA alle ore 11:30 del giorno 20/06/2024 sono evidenziate le notizie nuove in rosso ed in blu quelle presenti precedentemente

Capitolo 4

Implementazione

L'esecuzione del progetto prevede diversi step che sono fra loro consecutivi, i passi logici sono i seguenti:

- **Reperimento delle notizie**: Le notizie vengono reperite attraverso la lettura dei feed RSS o lo scraping. Questa è la fase in cui le notizie vengono salvate in file, quindi il software crea gli snapshot in formato JSON. Tutte le fasi successive si occupano di modificare questi file o leggerli.
- **Uniformazione delle notizie**: Le notizie devono essere uniformate per essere confrontate. Questo processo può essere effettuato da una **pipeline**.
- **Confronto delle notizie**: Le notizie vengono raggruppate e confrontate fra loro, successivamente sono generati dei grafici e salvati in immagini per essere poi visualizzati.
- **Visualizzazione dei risultati**: L'unica fase che ha bisogno dell'intervento umano, vengono visualizzati i grafici e si traggono conclusioni da questi.

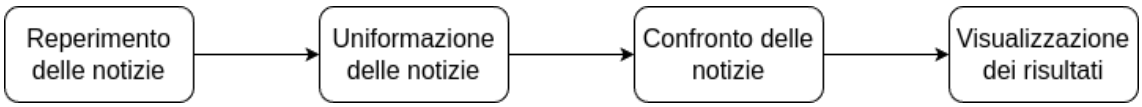

Figura 4.1: Visualizzazione grafica di come tutti i passaggi devono essere eseguiti

4.1 Lettura dei feed RSS

I feed RSS[1] sono dei flussi di informazione che servono a distribuire contenuti in formato XML. I contenuti in questo formato, in genere, hanno una struttura molto rigida che, non tende a variare nel tempo. I feed RSS sono stati inventati per permettere agli utenti di rimanere aggiornati su temi o giornali che seguono, tuttavia per essere utilizzati occorre disporre di un lettore di notizie RSS. I vantaggi di questo sistema sono molteplici, in quanto, per più fonti viene utilizzato un solo lettore si avranno a disposizione tutte le notizie a cui si è interessati, senza sforzarsi di visitare tutti i siti web dei vari autori, per cui consente di risparmiare tempo.

4.1.1 Problemi relativi ai feed RSS

Purtroppo i feed RSS possono presentare diversi problemi che lo rendono poco pratico o non particolarmente conveniente, specie per raggiungere l'obiettivo di questa tesi. Alcuni problemi sono riguardano:

- **Aggiornamenti**: Può capitare che il gestore del feed non sia particolarmente interessato a mantenerlo sempre aggiornato, dunque, può capitare che le notizie siano molto in ritardo rispetto il sito dell'autore o nei casi più gravi, alcuni contenuti non vengano pubblicati.
- **Replica dei contenuti**: Essendo strutturati, i feed RSS si prestano particolarmente bene per duplicare i contenuti del sito. È abbastanza semplice, infatti, creare un crawler[2] per poi ripubblicare automaticamente tutte le notizie presenti su un altro sito web, senza l'autorizzazione dell'autore originale.
- **Assenza dei feed**: Talvolta i creatori di contenuti potrebbero decidere di non adottare i feed RSS. Questo problema porta ad analizzare una nuova metodologia per il reperimento delle notizie: lo scraping.

4.2 Scraping

Lo scraping è una tecnica di estrazione delle notizie da un sito web in forma automatica. L'attività principale di un web scraper, detto anche crawler o spider, è quello di simulare la navigazione di un browser su un sito web.

[1] RSS è l'acronimo di "Really Simple Syndication".

[2] Un crawler, anche detto spider, è un software che analizza ed eventualmente colleziona i contenuti in rete, in maniera automatizzata.

4.2.1 Problemi relativi allo scraping

Lo scraping presenta una serie di problemi che lo rendono a volte difficoltoso o lo vietano. Anzitutto, lo scraping non è sempre permesso dagli autori di alcuni siti web, in quanto facilita la ripubblicazione di notizie, come per i feed RSS.
Inoltre, le strutture delle pagine web possono cambiare molto rapidamente ed in questo caso occorre aggiornare gli spider per poter continuare l'attività. Tuttavia, nonostante si effettui un aggiornamento tempestivo, a meno di non conoscere anticipatamente la nuova struttura delle pagine web, si incorre in una perdita di scansione dei dati, quindi, una perdita di uno o più snapshot.

4.2.2 Utilizzo di Scrapy

Per la realizzazione degli scraper è stata mantenuta la stessa libreria utilizzata per TARO, ovvero Scrapy[3].
Scrapy è un software open-source utilizzato per estrarre automaticamente informazioni da siti web utilizzando degli spiders, che sono utilizzati per creare gli snapshot delle notizie. Analizzando uno dei crawler creati è possibile capire come sono stati strutturati, di seguito, si fa riferimento allo scraper del giornale "Los Angeles Times".
Anzitutto tutti gli scraper hanno uno o più URL da cui iniziare il loro lavoro.

Dopo aver analizzato le pagine indicate come *start_urls*, si procede a identificare gli articoli per poi ottenere un riferimento ipertestuale a questi. Nel **codice 4.1** si fa notare il metodo parse dello scraper in esame. Occorre fare override del metodo *scrapy.Spider::parse* per modificare il comportamento dello spider e fare in modo che collezioni informazioni dal campo che si desidera. Inoltre, è interessante notare che quando viene chiamata l'istruzione *yield* il software segue il nuovo link che sarà analizzato usando il metodo *parseArticle*, la cui implementazione è visibile nel **codice 4.2**. I metadati passati sono utili per tenere traccia del riferimento che ha portato al rilevamento di quella notizia, questi link, insieme a quelli degli articoli saranno utilizzati per classificare gli articoli nelle varie sezioni dei giornali.

[3]Repository di Scrapy: `https://github.com/scrapy/scrapy`.

```
start_urls = [
        "https://www.latimes.com/",
        "https://www.latimes.com/california",
        "https://www.latimes.com/topic/california-law-politics",
        "https://www.latimes.com/topic/earthquakes",
        "https://www.latimes.com/topic/education",
        "https://www.latimes.com/topic/fires",
        "https://www.latimes.com/business",
        "https://www.latimes.com/business/technology",
        "https://www.latimes.com/business/real-estate",
        "https://www.latimes.com/entertainment-arts/business",
        "https://www.latimes.com/business/autos",
        "https://www.latimes.com/sports",
        "https://www.latimes.com/sports/soccer",
        "https://www.latimes.com/food",
        "https://www.latimes.com/environment",
        ...
]
```

Figura 4.2: Lista degli URL utilizzati dallo scraper per iniziare il rilevamento delle notizie

Codice 4.1: Metodo scrapy.Spider::parse di cui si è fatto l'override nello scraper in esame.

```
def parse(self, response):
    super().parse(response)

    articles = response
        .css("h2.promo-title a.link::attr(href)").getall()
    for article_link in articles:
        if (response.url == self.start_urls[0]
            and response.url not in self.home_page):
            self.home_page.append(article_link)

        if (self.base_url in article_link
            and article_link not in self.captured):
            self.captured.append(article_link)
            yield response.follow(article_link,
```

```
                self.parseArticle,
                meta={'parent': response.url})
```

Codice 4.2: Definizione del metodo parseArticle

```
def parseArticle(self, response):
    parent_url = response.meta['parent']
    title = response.css("h1.headline::text").get()
    if title is None:
        return
    today = date.today()
    date_raw = response.css("time.published-date::attr(datetime)").get(
    news_url = response.request.url
    content_paragraph = response
        .xpath('//div[contains(@data-element,"story-body")]//p')
    content = ''
    self.ranked = self.ranked + 1
    timestamp = time.time()
    for p in content_paragraph:
        p_content = p.xpath('./text()').get()
        if (isinstance(p_content, str) == False
            or p_content is None):
            continue
        content = content + "\n" + p_content

    if news_url in self.home_page:
        parent_url = self.start_urls[0]

    new = {
        'title': title,
        'date_raw': date_raw,
        'date': today,
        'url': parent_url,
        'news_url': news_url,
        'subtitle': '',
        'content': content,
        'ranked': self.ranked,
        'placed': 'Abroad',
        'epoch': timestamp,
        'language': 'EN',
        'source': 'LosAngelesTimes',
        'local_time': self.calculate_local_time(),
```

```
        'timezone': self.timezone,
        'scraping_time': datetime.now().strftime("%Y-%m-%dT%H.%M.%S")
    }
    self.edition.append(new)
```

4.2.3 Estensione delle informazioni delle notizie all'interno degli snapshot

Al fine di perseguire gli obiettivi di questa tesi, occorre, estendere le notizie salvate negli snapshot, in quanto devono essere aggiunte nuove informazioni che poi saranno elaborate, in particolare, torneranno utili per creare dei gruppi di confronto di notizie. Per ogni notizia, sono stati aggiunti tre nuovi campi:

- **Local time**: Ora del fuso orario nella zona in cui viene pubblicato il giornale al momento del rilevamento della notizia.
- **Timezone**: Riferimento del fuso orario nell'area geografica in cui ha sede il giornale.
- **Scraping time**: Ora del server nel momento in cui viene rilevata la notizia.

Nella **Figura 4.3** è possibile notare la nuova struttura delle notizie in formato JSON, all'interno di uno snapshot. Nell'immagine proposta il contenuto della notizia è stato troncato in quanto troppo lungo.

4.2.4 Introduzione di nuovi scraper

Oltre ad estendere le informazioni relative alle notizie che vengono catturate, è stato importante anche introdurre scrapers per i nuovi giornali.
In questa tesi sono stati analizzati tutti i giornali presenti nella tabella sotto con la nazione di pubblicazione indicata secondo lo standard ISO 3166-1 alpha-2.

Nome	Timezone	Offset
Expresso (PT)	Western European Summer Time (WEST)	UTC +1
ANSA (IT)	Center European Summer Time (CEST)	UTC +2
Agi (IT)	Center European Summer Time (CEST)	UTC +2
Sowetanlive (ZA)	South African Standard Time (SAST)	UTC +2
Brasil 247 (BR)	Brasília Time (BRT)	UTC -3
Los Angeles Times (US)	Pacific Daylight Time (PDT)	UTC -7
9News (AU)	Australian Eastern Standard Time (AEST)	UTC +10

```
{
    "title": "At a Cannes Film Festival of big swings and faceplants, real life takes a
      back seat",
    "date_raw": "2024-05-19T10:00:17.363Z",
    "date": "2024-05-20",
    "url": "https://www.latimes.com/entertainment-arts/movies",
    "news_url": "https://www.latimes.com/entertainment-arts/movies/story/2024-05-19/cannes
      -2024-yorgos-lanthimos-emma-stone-kinds-of-kindness-on-becoming-a-guinea-fowl-bird
      -andrea-arnold-richard-gere-oh-canada",
    "subtitle": "",
    "content": "“Is it too real for ya?” snarls the Gang of Four-soundalike punk band Ever
      the optimist...",
    "ranked": 174,
    "placed": "Abroad",
    "epoch": 1716177760.083001,
    "language": "EN",
    "source": "LosAngelesTimes",
    "local_time": "2024-05-19 21:02:40",
    "timezone": "America/Los_Angeles",
    "scraping_time": "2024-05-20T06.02.40"
}
```


Figura 4.3: Struttura di una notizia in uno snapshot dopo l'estensione degli spiders.

4.2.5 Salvataggio delle notizie

Tutte le notizie sono organizzate in dizionari Python e successivamente sono inserite in una lista che conterrà tutti gli articoli dello lo snapshot. Quando il rilevamento delle notizie è completo gli scrapers creano un nuovo file denominato con il giorno, l'orario e il timestamp nel momento in cui è salvato. I file contenenti gli snapshot sono organizzati attraverso il filesystem in una gerarchia di cartelle, visibile nella **Figura 4.4**. Alla radice è presente una cartella che ne contiene altre due, gli snapshot delle testate ad edizione e quelle a flusso, in entrambe le cartelle sono presenti i codici ISO delle lingue in cui sono pubblicate le notizie, a loro volta, queste cartelle contengono i nomi dei giornali che hanno pubblicato i vari snapshot e, infine, sono presenti gli snapshot codificati in formato JSON.

4.3 Pipeline

Per rendere più veloce lo scraping è stata introdotta una pipeline che traduce le notizie raccolte e le uniforma, la pipeline è stata eseguita ogni giorno alle ore 01 da un server appositamente predisposto.

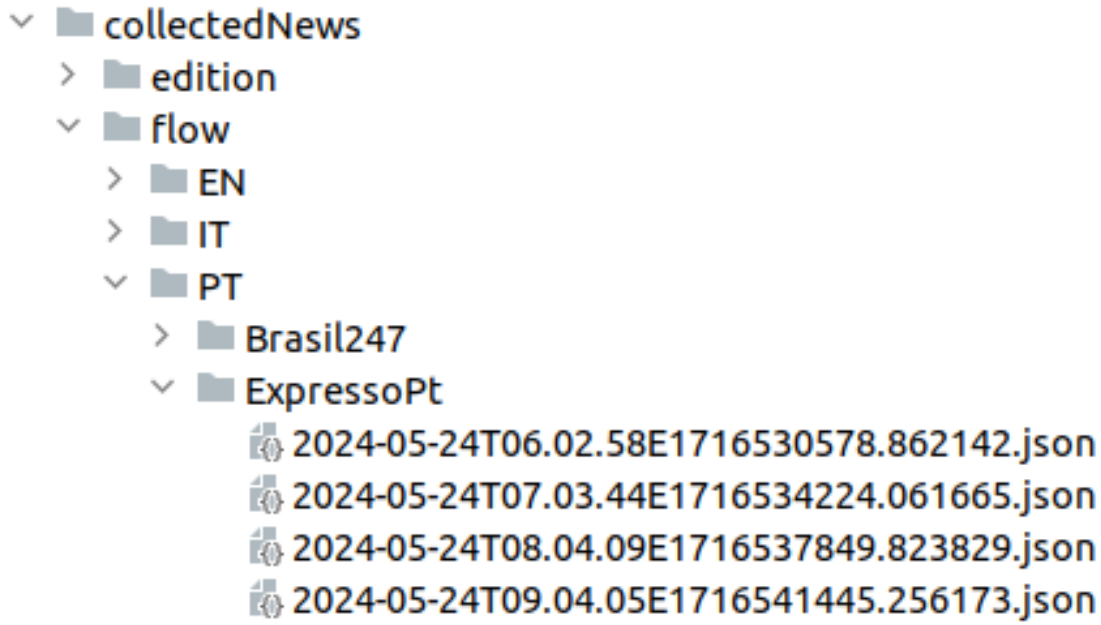

Figura 4.4: Struttura del filesystem

4.3.1 Traduzione

Le varie testate giornalistiche che sono state prese in esame sono pubblicate in Paesi diversi e quindi le notizie hanno lingue diverse tra loro.
Di seguito sono esposti i vari giornali con la loro lingua di pubblicazione.

Nome	Lingua
Expresso (PT)	Portoghese
ANSA (IT)	Italiano
Agi (IT)	Italiano
Sowetanlive (ZA)	Inglese
Brasil 247 (BR)	Portoghese
Los Angeles Times (US)	Inglese
9News (AU)	Inglese

Per eseguire le traduzioni, ancora una volta, è stata mantenuta scelta per TARO, quindi è stata adottata la libreria di python *argos-translate*[4], tutte le notizie sono tradotte in inglese. Argos-translate è una libreria che utilizza OpenNMT per le traduzioni, oltre ad essere utilizzato come libreria può essere usato come programma con GUI e da linea di comando.
La libreria è stata utilizzata come per estendere gli oggetti salvati in JSON. In

[4]Link al repository di argos-translate: `https://www.argosopentech.com/`

particolare vengono tradotti i campi: *title*, *subtitle* e *content*. Una volta effettuate le traduzioni saranno inseriti nei nuovi campi negli oggetti. Nello specifico, sono inseriti i nomi dei campi preceduti da *"en_"*.


```
{
  "title": "Operação Marquês: dez anos depois há um novo imbróglio jurídico para
    resolver e crimes prestes a prescrever",
  "subtitle": "Uma juíza não aceita refazer a decisão instrutória de Ivo Rosa e
    argumenta que só este juiz o pode fazer. (...)",
  "content": "A juíza Sofia Marinho Pires tomou o lugar de Ivo Rosa no Tribunal Central
    de Instrução Criminal de Lisboa depois de o colega ter sido promovido a
    desembargador da Relação de Lisboa. (...)",
  "en_title": "Operation Marquis: ten years later there is a new legal imbroglio to
    resolve and crimes about to prescribe",
  "en_content": "Judge Sofia Marinho Pires took the place of Ivo Rosa at the Lisbon
    Central Criminal Instruction Court after the colleague was promoted to
    disembargator of the Lisbon Relations. (...)",
  "en_subtitle": "A judge does not accept to redo the instructional decision of Ivo
    Rosa and argues that only this judge can do so. (...)"
}
```


Figura 4.5: Una notizia salvata in formato JSON. Sono stati mantenuti solo i campi originali e quelli tradotti per questioni di brevità, per la stessa ragione il contenuto di alcuni campi è stato troncato.

4.4 Confronto delle notizie

Una volta uniformate le notizie occorre effettuarne il confronto per generare successivamente dei grafici.
In questa sezione discuteremo dei passaggi che vengono effettuati per poter effettuare le dovute comparazioni.

1. Creazione degli insiemi da comparare
2. Calcolo della similarità tra le notizie appartenenti ad insiemi diversi
3. Creazione dei grafici

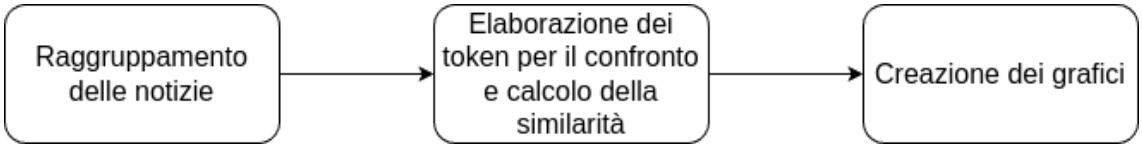

Figura 4.6: Passaggi logici da compiere per confrontare le notizie e raggiungere l'obiettivo prefissato da questa tesi.

4.4.1 Creazione degli insiemi da comparare

Prima di effettuare qualunque confronto è necessario creare degli insiemi, uno per ogni giornale, in modo tale che poi possano essere comparate notizie appartenenti a giornali diversi. Quando si comparano due insiemi, ogni notizia nel primo viene comparata con ogni altra nel secondo gruppo.
La creazione degli insiemi è uguale sia che si sta analizzando lo scraping time sia se si considera l'ora locale. Per l'estrazione degli articoli è stata definita la funzione *get_newspaper_time_articles*. La funzione ha, fra gli altri, alcuni parametri molto utili tra cui:

- La path del giornale che si sta considerando.
- La data a cui si è interessati ad ottenere le notizie[5].
- L'orario di inizio del range a cui si è interessanti.
- L'orario di fine del range di cui prima.
- Il nome del campo che contiene il datetime considerato.

Per evitare di scansionare file dove non è possibile individuare notizie utili e quindi, ridurre le operazioni di accesso al disco, è stata definita una lista che contiene il

[5]Nel codice: *current_date*

giorno prima, il giorno in cui interessa estrarre le notizie e il giorno successivo[6]. Quanto detto sopra è realizzato nel codice che segue

Codice 4.3: Funzionamento del controllo che consente di saltare alcuni file per analizzare solo quelli effettivamente utili.

```
...
current_date = datetime.strptime(date, "%Y-%m-%d")
check_days = [
    (current_date - timedelta(days=1)).strftime('%Y-%m-%d'),
    date,
    (current_date + timedelta(days=1)).strftime('%Y-%m-%d')
]
...
for directory_entry in os.scandir(newspaper_dir):
    dir = f"{newspaper_dir}/{directory_entry.name}"
    should_check = False
    for tmp_d in check_days:
        if tmp_d in directory_entry.name:
            should_check = True

    if not should_check:
        continue
...
```

Se i file superano il controllo precedente saranno caricati in memoria e convertiti in dati utilizzando la libreria *json*.
Ovviamente i file devono essere stati elaborati dalla **pipeline** per confrontare le notizie al loro interno, quindi viene estratto il primo articolo e si procede a verificare la presenza dei campi che testimoniano l'avvenuta lavorazione all'interno di questo.

Codice 4.4: Controllo degli articoli per verificarne la pre-elaborazione dalla pipeline

```
...
with open(dir, "r", encoding="utf-8") as f:
    try:
        snapshot = json.load(f)
        if len(snapshot) == 0:
            continue
        first_article = snapshot[0]
```

[6]Il giorno precedente e quello successivo sono stati inseriti in quanto è possibile che le notizie si trovino salvate uno snapshot che ha come nome giorno diverso da quello preso in esame ma la loro data sia effettivamente quella cercata, a causa dei cambi orari dati dal fuso orario.

```
    except:
        continue
if ("scraping_time" not in first_article
        or "timezone" not in first_article
        or "local_time" not in first_article
        or "en_title" not in first_article): continue
...
```

Una volta effettuato quest'ulteriore controllo, si verifica che l'articolo sia effettivamente nel range orario desiderato. Occorre calcolare prima gli oggetti di tipo *datetime* da usare per il confronto, inoltre, è stato introdotto anche un meccanismo, implementato attraverso una lista, per evitare di estrarre più volte lo stesso articolo, se presente in più snapshot. Il funzionamento è alquanto banale, è sufficiente memorizzare i titoli delle notizie e verificare che la notizia che si sta analizzando abbia un titolo che non è presente all'interno della lista. Il tutto è presentato dal seguente frammento di codice.

Codice 4.5: Controllo degli articoli per verificarne presenza nel range indicato

```
...
start_time = datetime.strptime(f"{date} {start_time}",
    "%Y-%m-%d %H:%M:%S")
end_time = datetime.strptime(f"{date} {end_time}",
    "%Y-%m-%d %H:%M:%S")
current_date = datetime.strptime(date, "%Y-%m-%d")
...
for article in snapshot:
    checking_time = article[time]
    checking_time = checking_time.replace("T", " ")
    checking_time = checking_time.replace(".", ":")
    checking_time = datetime.strptime(checking_time,
        "%Y-%m-%d %H:%M:%S")
    if not start_time <= checking_time <= end_time:
        continue
    en_title = article["en_title"]
    if en_title not in titles_captured:
        titles_captured.append(en_title)
        output.append(article)
...
return output
```

4.4.2 Calcolo della similarità

Il calcolo della similarità è stato effettuato utilizzando un modello già allenato messo a disposizione dalla libreria di Python, Spacy[7], una libreria open-source che propone modelli già addestrati per il calcolo della similarità tra testi e contenuti testuali.
Per calcolare la similarità occorre trasformare il contenuto della notizia in un documento di Spacy come mostrato nell'esempio sotto

Codice 4.6: Trasformazione di una stringa in uno spacy Doc

```
nlp = spacy.load("en_core_web_lg")

doc = nlp("Hi! I like hamburgers!")
```

Questa operazione è core intensive e richiede del tempo, si è provato ad inserirla nella **pipeline** con successo, tuttavia lo spazio richiesto per salvare tutti i nuovi snapshot era eccessivo quindi in poco tempo si sarebbe saturata la memoria a disposizione del calcolatore. Per questo le notizie sono trasformate in Spacy Doc[8] nel momento in cui occorre effettuare dei confronti, anche in questo caso le notizie per le quali è già stato calcolato lo Spacy Doc vengono inserite in cache e riutilizzate se si rende necessario effettuare nuovamente dei confronti.
Una volta elaborate tutte le notizie, è possibile, calcolare la similarità fra i documenti ottenuti, per fare questo la classe *Doc* di Spacy mette a disposizione il metodo *Doc::similarity*[9] che calcola l'algoritmo del coseno fra i word vectors dei documenti. L'algoritmo del coseno ritorna un valore di tipo *float* compreso fra 0 e 1, valori più vicini ad 1 indicano una maggiore similarità, valori vicini a 0 indicano una scarsa somiglianza fra le notizie.
Tuttavia si è ritenuto necessario sapere se una notizia è simile ad un'altra, non importa il fattore di somiglianza, per questo, è stato posto un threshold, se il fattore di similarità è almeno questa soglia, allora le notizie sono ritenute simili, altrimenti, no.

Codice 4.7: Calcolo della similarità tra due notizie dato un threshold.

```
def calculate_similarity(art_a, art_b, threshold):
    similarity = art_a.similarity(art_b)
    if similarity >= threshold:
        return (True, similarity)
    else:
        return (False, similarity)
```

[7] Link alla documentazione di Spacy: `https://spacy.io/usage/spacy-101`
[8] Un oggetto della libreria Spacy che contiene dei Word Vectors calcolati usando Word2Vec
[9] `https://spacy.io/api/doc#similarity`

Per tutte le analisi date è stato posto un threshold fisso, ovvero, $threshold = 0.9875$. È stato osservato che effettuando esperimenti, questo valore, garantiva il giusto compromesso tra notizie realmente simili e falsi positivi.

4.4.3 Creazione dei grafici

Una volta che il software ha determinato quali sono le notizie simili occorre contare le notizie negli insiemi e quali, fra queste sono state ritenute non uniche, dopo aver effettuato ciò, attraverso semplici operazioni di calcolo[10], si procede alla creazione dei grafici.
Per crearli è stata utilizzata la libreria di Python *matplotlib*. Sono stati realizzati due tipi di grafici:

- Grafici relativi al numero di notizie
- Grafici relativi al rapporto di $\frac{Notizie\ simili}{Totale\ notizie}$

Inoltre, grazie al livello di parametrizzazione di tutti gli script per effettuare le analisi, è stato possibile effettuare le analisi in vari frame temporali di dimensioni:

- un'ora
- tre ore
- sei ore
- dodici ore
- ventiquattro ore

Nei grafici si vuole mettere in evidenza come cambia il numero di notizie ed il loro rapporto nel caso in cui si scelga di analizzarle seguendo l'orario di *scraping* o l'*ora locale*.

4.4.4 Rilevamento delle notizie in home page e nelle aree tematiche

. Fino ad ora è stato analizzato come le notizie sono raggruppate, analizzate e confrontate, senza dare particolarmente attenzione al fatto che queste si trovino sulla home page o nelle aree tematiche. Questa fondamentale distinzione è realizzata attraverso gli URL della notizia o della pagina web che la conteneva, in particolare, la presenza di alcune **parole chiave** all'interno di questi fa comprendere se un articolo è pubblicato in un'area tematica piuttosto che in un'altra.

[10]Gli articoli di un insieme sono salvati in una lista e anche le notizie ritenute simili ad altre. Attraverso la funzione Python *len(...)* è possibile effettuare il conteggio di queste strutture dati

Tuttavia, rimane vero, che alcune notizie in home page possono appartenere ad alcune aree tematiche per questo come è possibile notare nel *Codice 4.2* se il link che ha generato una notizia è il primo elemento dell'array *start_urls*[11] allora il titolo sarà salvato all'interno di un array speciale che contiene tutti i titolo delle notizie in home page e, successivamente, quando la notizia sarà salvata verrà attribuito al campo *url*, il link della home page.
Per quanto riguarda, invece, l'analisi delle aree tematiche, ad ogni categoria analizzata è stata associata una lista di parole chiave. Quando si effettua l'analisi per una qualunque area tematica, viene si verifica la presenza delle parole chiave associate all'interno degli URL della notizia. Se sono presenti allora viene considerata, altrimenti la notizia viene ignorata.

```
categories = {
    "world": ["world", "mondo", "estero", "internacional", "mundo", ...],
    "economy": ["economy", "finance", "economia", "business", "empreender"],
    "sport": ["sport", "esporte", "sports"],
    "tech": ["tech", "future_tech", "technology"],
    "culture": ["culture", "cultura", "environment"],
    "politics": ["politics", "politica"],
}
```

Figura 4.7: Associazione dei temi con le loro parole chiave

[11] Che corrisponde in tutti gli scraper al link in homepage.

Capitolo 5

Esperimenti

Di seguito sono mostrati alcuni dei grafici generati dagli script.
Per tutti i grafici è stato analizzato il giorno 20 maggio 2024, per cui, tutti i grafici sono originati a partire dalle stesse notizie. Tuttavia, i vari grafici possono analizzare finestre orarie diverse.
Attraverso la seguente tabella è possibile ricordare quali sono stati i giornali analizzati nel seguente esperimento.

Nome	Timezone	Offset
Expresso (PT)	Western European Summer Time (WEST)	UTC +1
ANSA (IT)	Center European Summer Time (CEST)	UTC +2
Agi (IT)	Center European Summer Time (CEST)	UTC +2
Sowetanlive (ZA)	South African Standard Time (SAST)	UTC +2
Brasil 247 (BR)	Brasília Time (BRT)	UTC -3
Los Angeles Times (US)	Pacific Daylight Time (PDT)	UTC -7
9News (AU)	Australian Eastern Standard Time (AEST)	UTC +10

L'ordine dei giornali non è casuale. Sono ordinati in ordine crescente per differenza dell'offset rispetto all'UTC[1], grazie a questa scelta è facile capire che visualizzando i grafici da sinistra verso destra ci si allontanerà dal meridiano di Greenwich e quindi aumenterà conseguenzialmente la distanza di pubblicazione dei vari giornali, la **Figura 5.1** dà le informazioni per collocare geograficamente le aree di pubblicazione dei giornali.
I seguenti grafici vogliono rispondere a una serie di domande, ovvero, cosa succede se si confrontasse la vecchia selezione di notizie di TARO con quella nuova, cosa accade in termini numerici alle notizie simili e al totale delle notizie e che rapporto

[1]Differenza rispetto all'UTC considerata in valore assoluto

c'è tra queste due quantità. Ci si aspetta che i giornali che pubblicano in fusi orari molto distanti dallo *Scraping time* abbiano dati molto differenti da quelli che sono, invece, più vicini.

Figura 5.1: Una mappa che indica tutti i luoghi in cui sono pubblicati i giornali

I grafici mostrati sono quelli ritenuti più significativi e sono presenti molti altri grafici, creati analizzando diverse finestre temporali e anche altre aree tematiche, che per questioni di brevità non sono mostrati.

5.1 Legende e letture dei grafici

Per ogni analisi presentata, sono presenti due grafici consecutivi, che rappresentano rispettivamente il numero di notizie e il rapporto tra queste. La legenda è la seguente:

- Per quanto riguarda i grafici che indicano il numero di notizie, sono presenti quattro colonne che hanno i seguenti colori:

  - **Rosso**: Notizie ritenute simili ad altre considerando l'ora locale dei giornali.

  - **Azzurro**: Notizie totali considerando l'ora locale dei giornali.

  - **Verde**: Notizie ritenute simili ad altre considerando lo *Scraping time*.

  - **Giallo**: Notizie totali considerando lo *Scraping time*.

- Per quanto riguarda, invece, i grafici che indicano i rapporti tra notizie simili e totale notizie, sono presenti due colonne, che hanno i seguenti colori:

  - **Azzurro**: Rapporto di notizie ritenute simili e totale notizie considerando l'ora locale dei giornali.

- **Verde**: Rapporto di notizie ritenute simili e totale notizie considerando lo *Scraping time*.

5.1.1 Struttura dei grafici

I grafici sono stati strutturati in un certo modo per rendere la loro lettura chiara.

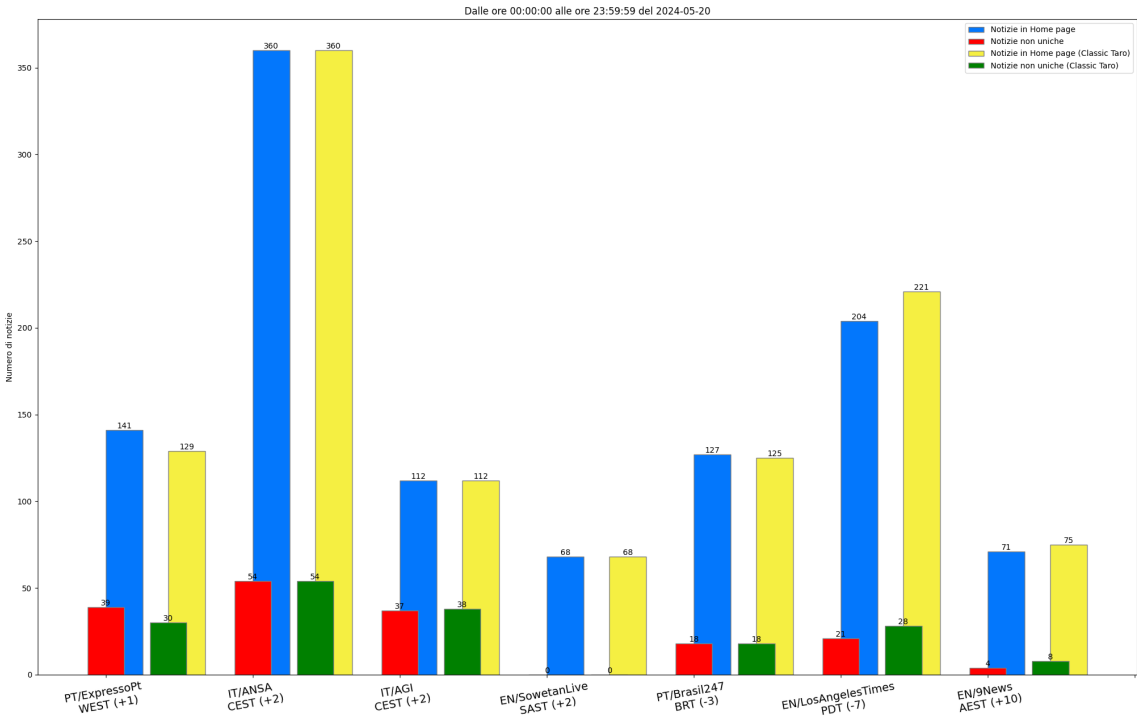

Figura 5.2: Analisi con un unico frame temporale di ventiquattro ore.

La disposizione delle colonne all'interno del grafico non è casuale, la colonna rossa è strettamente collegata a quella blu, come la colonna verde è strettamente collegata a quella gialla. Inoltre, la colonna rossa e la colonna verde sono leggermente spostate rispetto la colonna blu e la colonna gialla in quanto, se fossero state sovrapposte, il grafico sarebbe stato ambiguo: non sarebbe più stato chiaro quale significato semantico dare al valore riportato dalle colonne blu e giallo. In quanto avrebbe potuto riportare, ad una lettura poco attenta del grafico, la somma totale delle notizie.
Invece per quanto riguarda i grafici che rappresentano i rapporti tra notizie ritenute simili e il totale notizie, non sono state fatte scelte implementative particolarmente interessanti.

5.2 Analisi delle home page

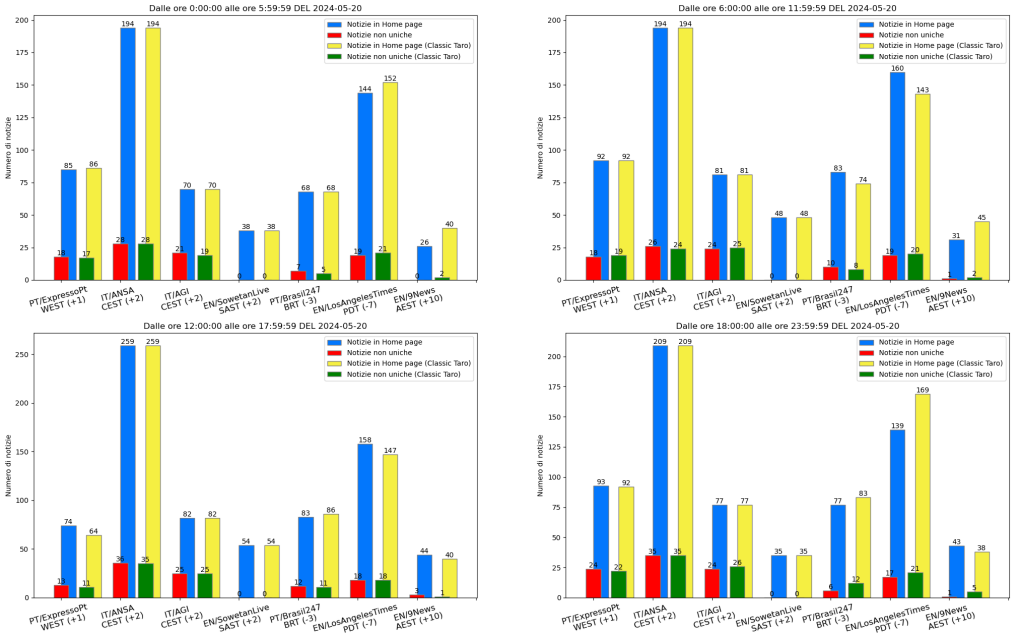

Figura 5.3: Analisi con frame temporali di sei ore.

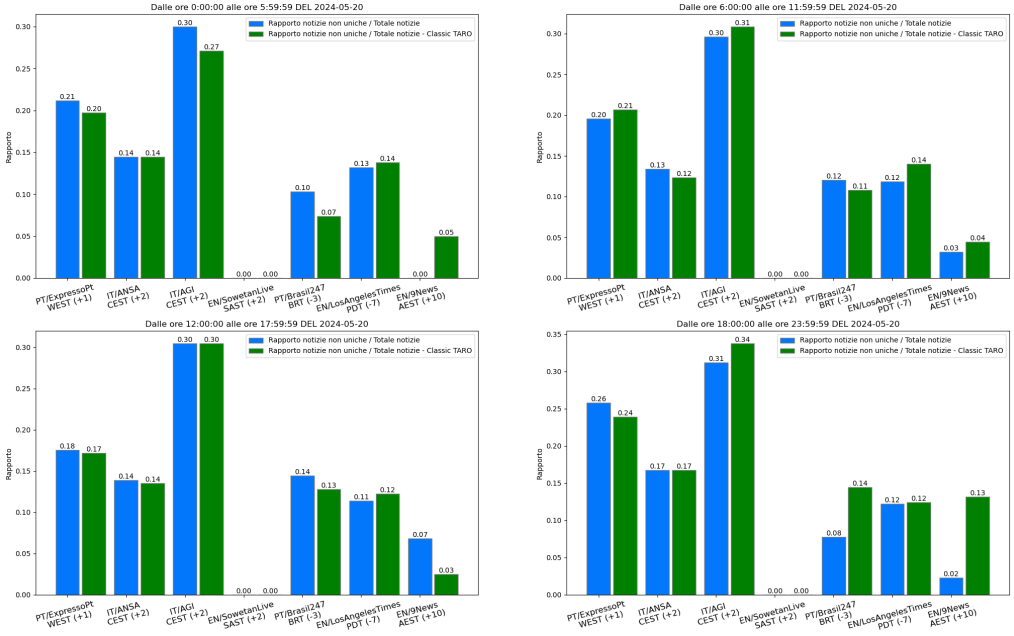

Figura 5.4: Analisi con frame temporali di sei ore, visualizzazione dei rapporti

Nella **Figura 5.3** è possibile notare che, essendo lo *Scraping time* uguale all'ora locale del giornale, per i giornali che hanno lo stesso offset del fuso orario uguale all'offset dello scraping time (UTC +2), il numero di notizie rilevato nelle home page è uguale. Inoltre, nel grafico rappresentante i rapporti dei numeri di notizie è particolarmente evidente che per il giornale "SowetanLIVE" non siano presenti notizie simili con altri giornali.
Inoltre, con l'aumentare della distanza dei fusi orari, è possibile notare anche delle variazioni dei rapporti rispetto i giornali che sono, tra loro, più vicini, questo è particolarmente evidente nella **Figura 5.4**, la variazione indicata potrebbe essere data da qualche perturbazione delle notizie in quanto il numero di queste non è particolarmente significativo.

5.2.1 Analisi delle home page, esclusione di alcuni giornali

Analizzando le home page ci si è chiesto cosa sarebbe cambiato se si fosse escluso il giornale "ANSA" dalle analisi delle home page e cosa sarebbe accaduto se oltre ad escludere "ANSA" si sarebbe escluso anche il giornale "AGI".
Si è deciso di escludere il giornale "ANSA" in quanto ha un maggior numero di notizie, inoltre, si vuole verificare che il numero di notizie ritenute simili per il giornale "AGI" cambi.

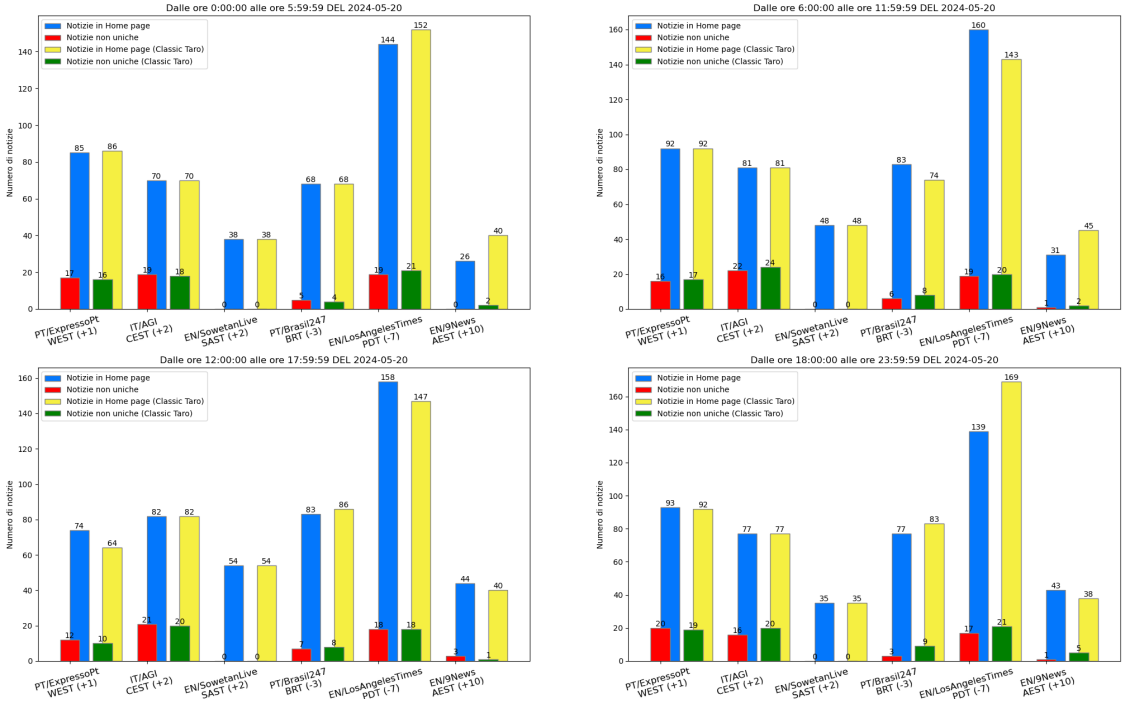

Figura 5.5: Analisi con frame temporali di sei ore, escludendo ANSA

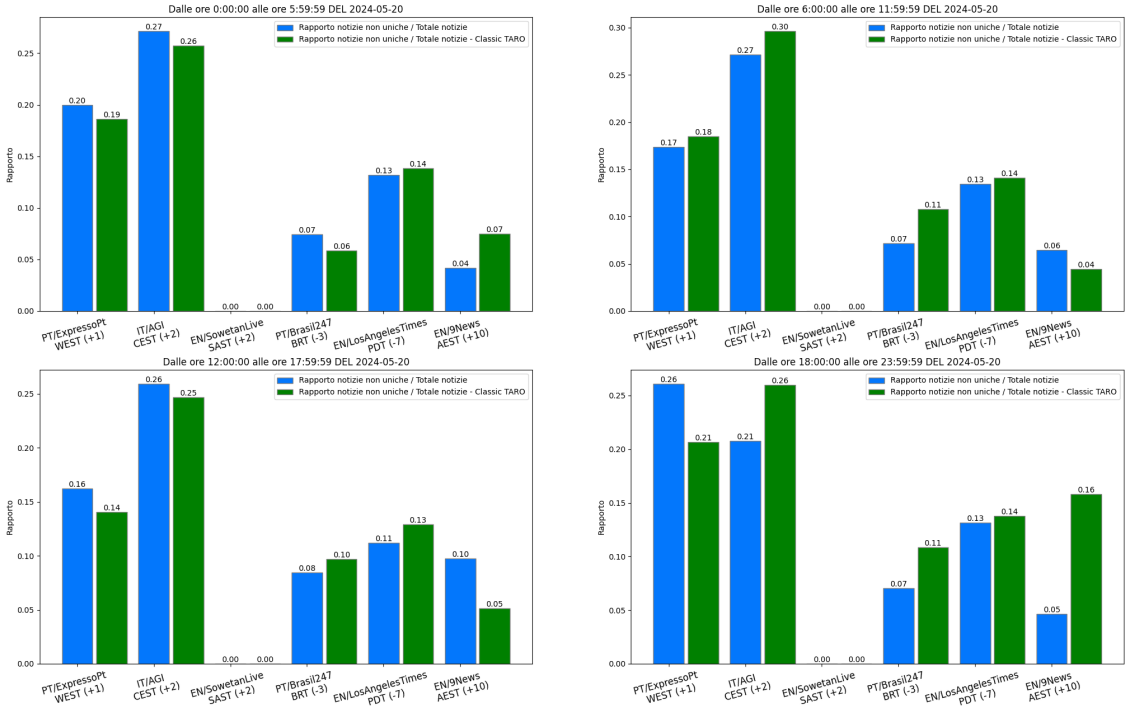

Figura 5.6: Analisi con frame temporali di sei ore, escludendo ANSA, visualizzazione dei rapporti

Come è possibile notare che escludendo il giornale "ANSA", il giornale italiano "AGI" ha un numero minore di notizie ritenute simili ad altri giornali, in quanto, alcune di queste notizie erano state ritenute simili nel gruppo di articoli del giornale escluso.
Per ragioni di completezza è mostrato anche la situazione escludendo tutti i giornali italiani, quindi anche "AGI", quindi il giornale "SowetanLIVE" è l'unico giornale rimasto nel grafico per cui vale che lo *scraping time* è uguale all'orario locale.

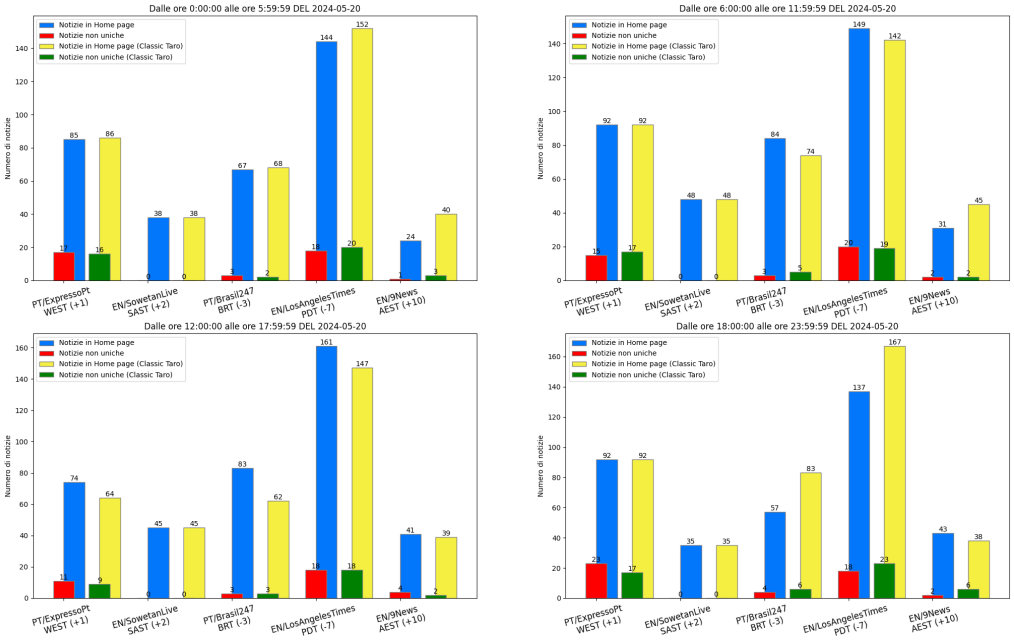

Figura 5.7: Analisi con frame temporali di sei ore, escludendo tutti i giornali italiani.

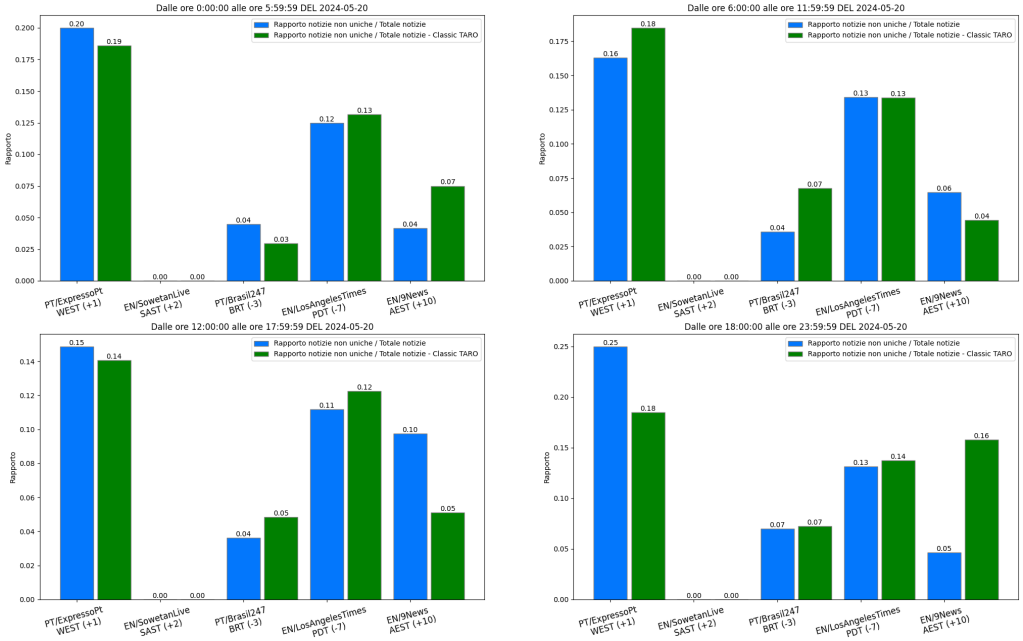

Figura 5.8: Analisi con frame temporali di sei ore, escludendo tutti i giornali italiani, visualizzazione dei rapporti.

5.2.2 Comparazione binaria di AGI e ANSA

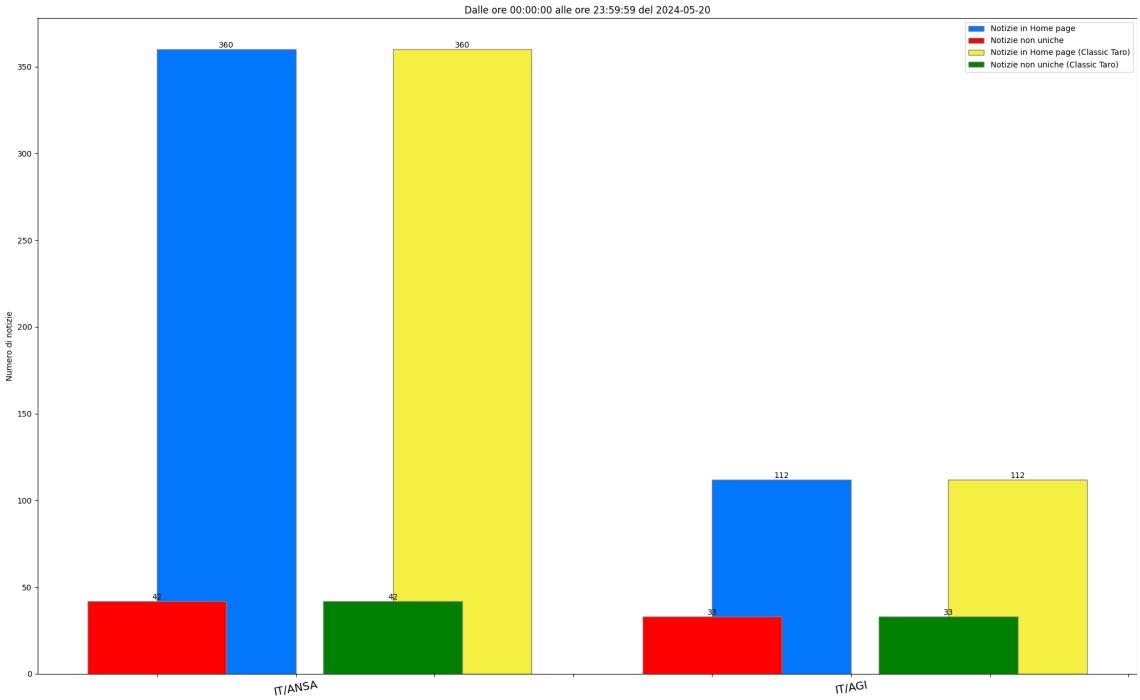

Figura 5.9: Comparazione delle home page di AGI ed ANSA.

Come è possibile notare nella **Figura 5.9** è possibile notare come, le colonne, considerando entrambi gli orari, riportano le stesse quantità: questo non è un caso, in quanto tutti i gruppi di notizie analizzati sono identici per entrambi gli approcci. Inoltre, il numero di notizie in più ritenute simili nel caso del giornale "ANSA" potrebbero essere falsi positivi.

5.3 Analisi delle aree tematiche

L'analisi delle aree tematiche è stata effettuata con gli stessi criteri utilizzati per quella delle home page. Tuttavia ci si sarebbe aspettata una differenza ancora minore, in quanto, nelle aree tematiche, mediamente sono pubblicate notizie meno frequentemente rispetto le home page.

5.3.1 Economia

Si analizzi la sezione economia di tutte le testate giornalistiche. È possibile notare che tutti i giornali hanno almeno una notizia ritenuta simile, sia per i nuovi metodi di analisi, sia per quelli di TARO senza tenere conto del fuso orario.

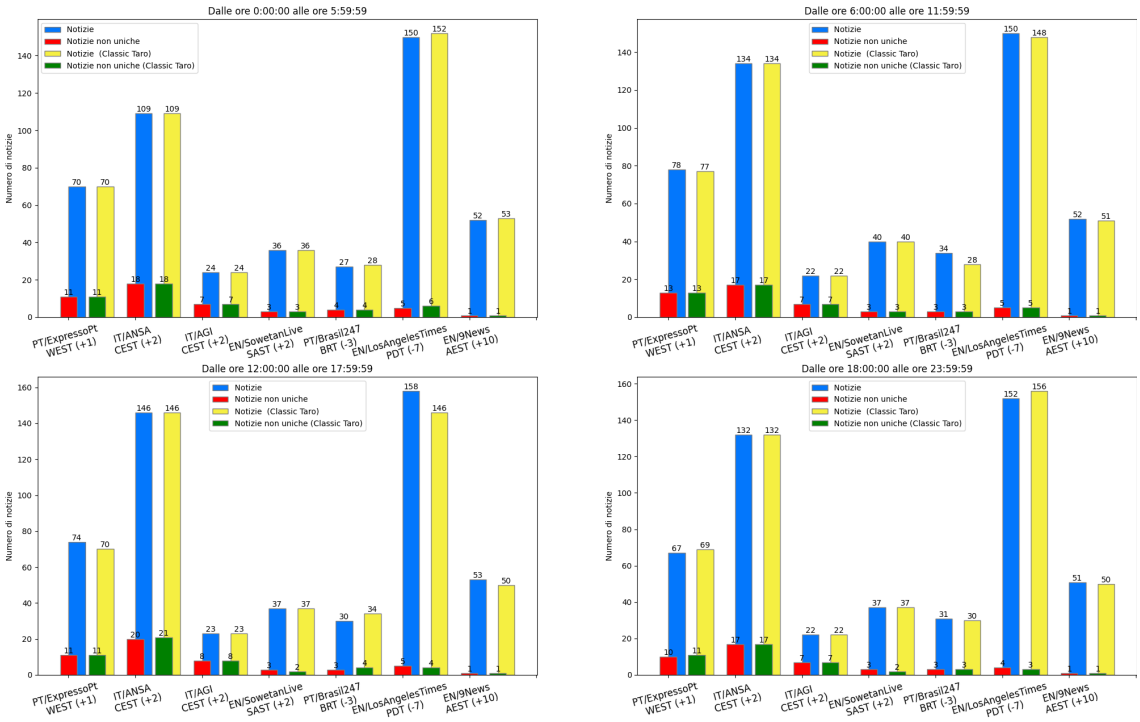

Figura 5.10: Analisi della sezione di economia dei giornali con frame temporali di sei ore.

Analizzando il grafico raffigurante i rapporti di notizie simili sul totale delle notizie (**Figura 5.11**) è possibile constatare che le differenze sono minime, questo a causa della scarsa frequenza di pubblicazione di notizie nuove nelle pagine tematiche.

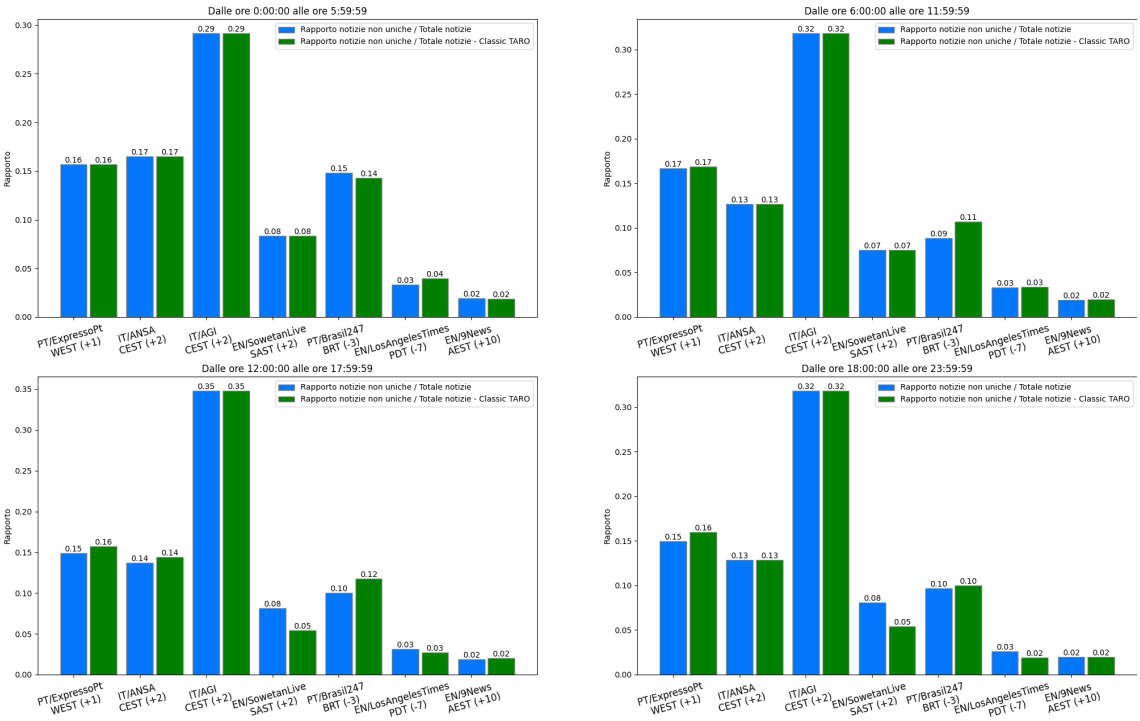

Figura 5.11: Analisi della sezione di economia dei giornali con frame temporali di sei ore, visualizzazione dei rapporti

5.3.2 Esteri

Ci si è chiesti che cosa sarebbe accaduto se si fosse analizzata l'area tematica riguardante gli esteri dei giornali.

Come ci si sarebbe aspettato, il grafico mostrante i rapporti ha valori molto alti. Inoltre, indica che in alcuni momenti della giornata si è arrivati a oltre il 60% di notizie ritenute simili sul totale, nel giornale "Expresso" pubblicato in Portogallo, questo dato è visibile nella **Figura 5.13**.

È possibile constatare, nella **Figura 5.14** e nella **Figura 5.15** che dalle ore 9:00 alle ore 9:59:59, analizzando le notizie per *Scraping time*, il numero di notizie simili, rispetto il totale è molto alto. La differenza con l'analisi con i dati relativi all'ora locale è data dal fatto che due notizie ritenute simili non sono state rilevate e questo ha peggiorato i risultati del nuovo approccio.

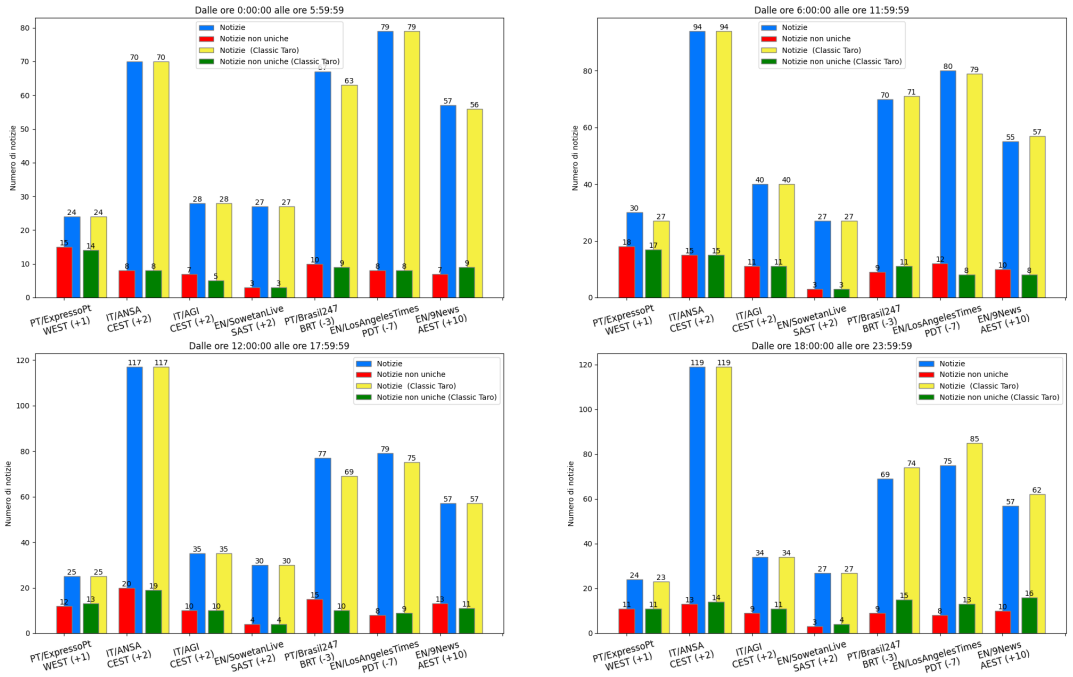

Figura 5.12: Analisi della sezione esteri dei giornali con frame temporali di sei ore.

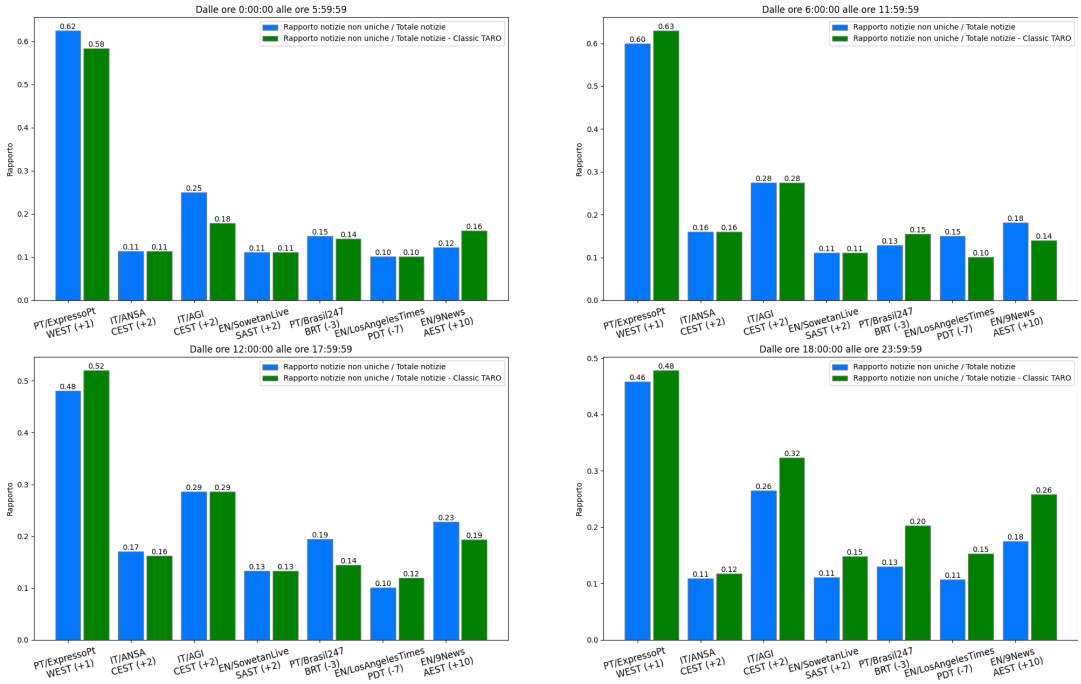

Figura 5.13: Analisi della sezione esteri dei giornali con frame temporali di sei ore, visualizzazione dei rapporti.

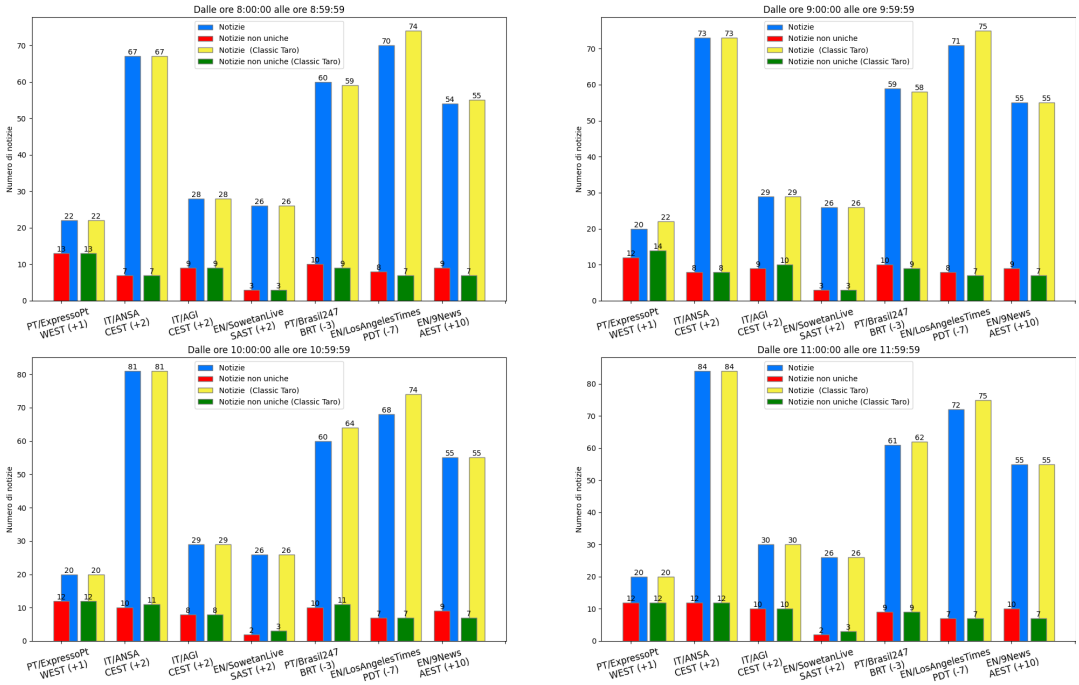

Figura 5.14: Analisi della sezione esteri con frame temporali orari.

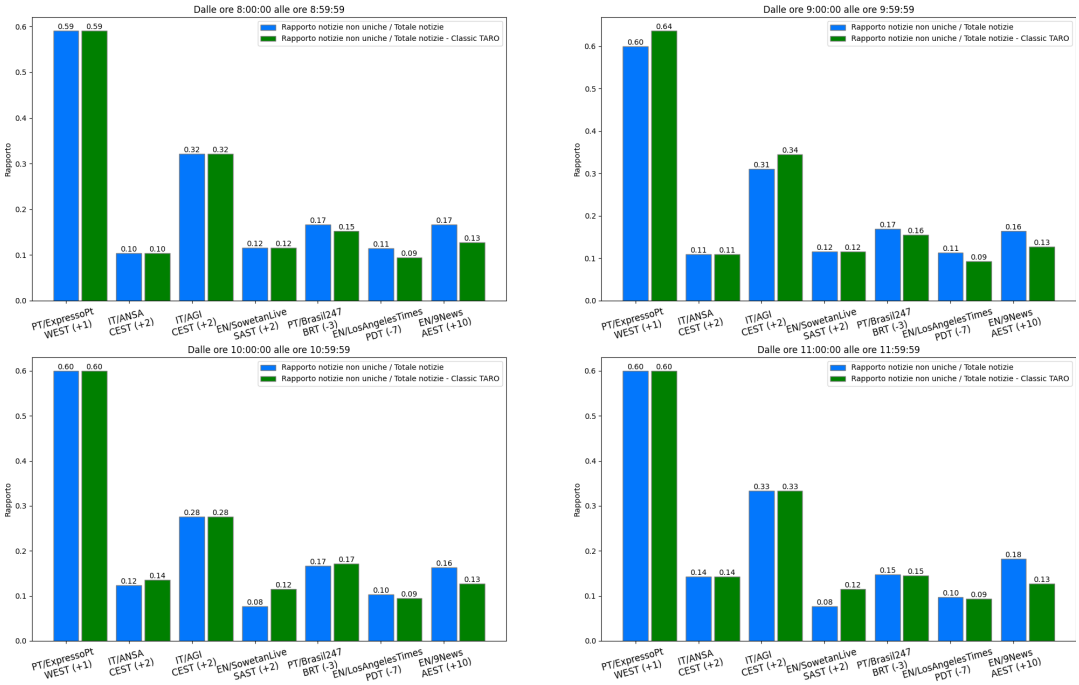

Figura 5.15: Analisi della sezione esteri con frame temporali orari, visualizzazione dei rapporti.

Capitolo 6

Conclusioni

Dagli esperimenti effettuati si può constatare che, contrariamente a quanto ci si sarebbe aspettato, la differenza di notizie, sia in termini assoluti, sia in termini relativi, non è significativa, in particolare, sono stati validati i risultati del vecchio modello di TARO.
Per quanto riguarda le home page, ci si sarebbe aspettata una maggiore differenza tra i due approcci illustrati, anche per le home page le notizie permangono molto tempo prima di essere sostituite da nuove. Inoltre, le notizie non sono particolarmente influenzate dai fusi orari, in quanto, per le fonti analizzate, quindi testate giornalistiche a **flusso**, si osserva che, in genere, le notizie sono pubblicate appena queste avvengono, quindi dato un evento con un'enorme risonanza mediatica, è molto probabile che i giornali, quasi simultaneamente pubblichino la notizia.
Per quanto riguarda, invece, le pagine tematiche dei giornali, è facilmente constatabile che, rispetto le home page dei giornali, la differenza in termini numerici tra i due metodi utilizzati per fare le analisi, è quasi nulla, quindi molto trascurabile e, ancora una volta, i risultati ottenuti dal vecchio modello di TARO sono stati validati.
Tuttavia, è stata confermata un'ipotesi che non era stata contemplata all'inizio della tesi. Come ci si aspettava le notizie pubblicate nella sezione "esteri" hanno, mediamente, in termini numerici, molte più notizie simili rispetto ad altre categorie. Questo a causa degli eventi che stanno succedendo nel mondo che hanno un'alta risonanza mediatica, come, le guerre.

6.1 Sviluppi futuri

Di seguito sono presenti diversi spunti che possono essere interessanti affinché possano essere studiati da altri studenti o curiosi, idee che possono essere sviluppare per arricchire ulteriormente il modello di TARO e quindi, successivamente, estenderlo

per altre scoperte interessanti.
Per continuare il lavoro svolto in questa tesi, è possibile estendere il modello proposto con altre fonti ed altri scrapers, in particolare sarebbe interessante se venissero inserite, in un primo momento, solamente delle testate giornalistiche ad edizione, quindi, che pubblicano i loro contenuti in tempi già schedulati.
Una volta fatto ciò sarebbe interessante anche capire come i giornali a flusso ed a edizione interagiscono fra di loro, ci si aspetterebbe che la maggior parte delle notizie pubblicate in un giornale a edizione sia presente anche nelle testate a flusso.
Un'altra idea potrebbe essere quella di analizzare una nazione molto estesa che comprende più fusi orari.
Se mettessimo da parte il modello che considera anche i fusi orari, sarebbe interessante:

- capire come altri modelli linguistici potrebbero migliorare o peggiorare il calcolo della similarità e il concetto di similarità stesso.
- analizzare in modo molto approfondito le aree tematiche, individuando anche altre sotto aree tematiche.
- creare un programma che data una notizia evidenza come questa sia presente nei vari paesi del mondo e utilizzando un planisfero si evidenzi quando questa notizia è stata pubblicata, ad esempio, con colori più forti i paesi che l'hanno pubblicata prima (o che l'hanno pubblicata più volte) e con colori sempre più chiari i paesi che l'hanno pubblicata dopo (o che l'hanno pubblicata meno volte).

Riferimenti bibliografici

[AFSea98] Mark Burnett Alan F. Smeaton and Francis Crimmins et al. An architecture for efficient document clustering and retrieval on a dynamic collection of newspaper texts. pages 1–9, 1998.

[BA17] Ilya Blokh and Vassil Alexandrov. News clustering based on similarity analysis. *Procedia Computer Science*, 122:715–719, 2017. 5th International Conference on Information Technology and Quantitative Management, ITQM 2017.

[Car] Giuseppe Carrino. *TARO: Infrastruttura per il Confronto di Testate Giornalistiche Internazionali*. PhD thesis.

[CDlB23] Giuseppe Carrino, Angelo Di lorio, and Gioele Barabucci. Comparison of news commonality and churn in international news outlets with taro. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[G+97] Stephen J Green et al. Building hypertext links in newspaper articles using semantic similarity. In *Third Workshop on Applications of Natural Language to Information Systems (NLDB'97)*, pages 178–190, 1997.

[Wie11] Wybo Wiersma. The invisible hands of time: How timezones shape online communities. *LogiLogi Foundation blog*, 2011.

Ringraziamenti

La stesura di questa tesi, come l'estensione del modello di TARO e la relativa implementazione non sono stati semplici, hanno richiesto una certa concentrazione e uno sforzo non indifferente, non solo per me ma anche per il professor Di Iorio e il ragazzo che ha iniziato questo progetto, il dottor Joseph Carrino, che hanno dedicato una buona parte del loro tempo nello spiegarmi l'obiettivo da raggiungere e a correggere eventuali scelte implementative non consone.
In particolare mi scuso con il professor Di Iorio, in quanto, in una prima fase non sono stato particolarmente attivo sul lavoro di tesi, in quanto sono sopraggiunti altri eventi.
Un ringraziamento particolare a mamma e papà, Cristina e Renato, che mi hanno sostenuto anche nei momenti pieni di ansia, instabilità emotiva e frustrazione, non solo durante la stesura di questa tesi ma soprattutto durante tutto il corso della mia vita, un grazie speciale a voi che sapete quanto possa essere complicato avere a che fare con me quando non sono dell'umore giusto.
Ringrazio Alessio e Manuel che sono stati i miei fratelli e mi hanno sempre sopportato, so di non essere il fratello migliore del mondo. Vi chiedo solo di tranquillizzare e ascoltare mamma e papà, so che sono un po' petulanti ma lo fanno perché si preoccupano e perché si aspettano grandi cose da noi.
Ringrazio Elisa che mi ha donato il suo affetto, il suo tempo e il suo amore. Senza di lei probabilmente mi sarei sentito molto più solo e smarrito nei momenti più difficili, non è scontato, inoltre, grazie per avermi accettato nonostante ti sentissi distante da me a causa del poco tempo trascorso insieme e scusami ancora.
Ringrazio i miei nonni, Angela, Vittorio e Antonietta ed i miei zii Michele e Lidia per essersi interessati al mio percorso di studi e per essersi preoccupati affinché io mangiassi e stessi bene, un ringraziamento anche a Domenico, che purtroppo non è più tra noi, ma so che anche lui si sarebbe preoccupato per me.
Ringrazio Guido per avermi dato la possibilità di auto finanziare i miei studi, un fatto non scontato, soprattutto nel mondo in cui ci troviamo, collaborare con te è sempre un piacere, non privo di qualche insidia (e di front-end da fare), da te ho imparato (e sto continuando ad imparare) a mantenere la calma e a trattare anche

con i caratteri più ostici (vedi me), ti ritengo un ottimo dirigente oltre che una gran persona.
Ringrazio Giovanni che è stato sempre presente e ha continuato a tenermi aggiornato sui componenti dei PC da gaming (molto importante anche se non sembra).
Ringrazio Gigi, Mattia, Roberta, Danilo, Ilaria, Samuele, Veronica, Flavia, Valerio, Alessia, Antonietta, Giovanni e Martina per essere rimasti miei amici durante tutto il percorso di studi, nonostante la distanza. Ringrazio il gruppo di regaz, ovvero, Francesco, gli Alessandri, Pietro, Giulia, Leonardo, Camilla, Mattia, Simone, Saverio, Lollo ed Elena per avermi aiutato a distrarmi dallo studio in sessione ed aver ascoltato le mie (inutili) paranoie, grazie per essere stati presenti.
Un ringraziamento anche a chi era presente all'inizio del percorso di studi ed adesso non è più nella mia vita, anche a loro auguro il meglio.

Chi sono i giganti?

> Siamo seduti sulle spalle di giganti, per questo possiamo vedere lontano.

Questa è una celebre frase di Bernardo di Chartres ripresa da Isaac Newton.
Ci fa riflettere sull'idea della cultura come un processo continuo dell'umanità, in cui i pensatori moderni, pur nani rispetto ai fondatori del sapere del passato, possono tuttavia sopravanzarli e progredire proprio grazie alle scoperte precedenti.
Mi sono rivisto nella persona sulle spalle di un'altra per guardare avanti, infatti grazie ai **miei giganti** ho potuto vedere questo traguardo, per poi raggiungerlo.