Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica

Uso di dati bibliografici aperti per l'Abilitazione Scientifica Nazionale: un esperimento sui settori bibliometrici

Relatore:
Prof.
Di Iorio Angelo

Presentata da:
Aleotti Federico

Sessione III
Anno Accademico 2018/2019

Introduzione

L'obiettivo di questa tesi è studiare le potenzialità e i limiti dell'uso di dati bibliometrici aperti per l'Abilitazioni Scientifica Nazionale in Italia. Il lavoro nasce dall'articolo "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation" [2] e si pone l'obiettivo di realizzare un software in grado di replicare ed automatizzare l'analisi descritta nell'articolo.
Gli autori hanno studiato la completezza delle sorgenti open che forniscono dati bibliografici sulle pubblicazioni scientifiche rispetto alle sorgenti closed. Esistono infatti diverse sorgenti pubbliche che permettono di ottenere i dati relativi ai paper prodotti ed alle citazioni che questi ultimi hanno ricevuto, come Crossref [7] e COCI [6], che si contrappongono a sorgenti private come Scopus [10] e Web of Science [11] che vendono le informazioni in merito alle pubblicazioni dei candidati.

Un possibile metodo che permetta di comparare le sorgenti open a quelle closed è quello di effettuare uno stesso calcolo partendo da due sorgenti dati differenti: questo approccio è stato adottato nell'articolo dove sono stati calcolati gli indici dell'Abilitazione Scientifica Nazionale, un concorso pubblico che ha come obiettivo quello di valutare l'autorevolezza dei candidati che desiderano ricoprire ruoli in ambito universitario.
I tre indici riguardanti il numero di pubblicazioni su journal, la somma delle citazioni ottenute e l'h-index del candidato vengono calcolati da ANVUR sulla base dei dati forniti da Scopus e Web of Science.

Nell'articolo sono stati calcolati i medesimi indici, unicamente sul settore informatico, basandosi su Crossref e COCI ed i risultati sono stati confrontati con quelli reali prodotti da ANVUR, in modo da verificare il grado di completezza delle sorgenti usate rispetto a quelle private.

Il programma oggetto di questa trattazione riprende il metodo del paper ampliandolo a tutti settori scientifico-disciplinari mediante uno script che automatizzi i calcoli necessari per ottenere i risultati desiderati e fornisca in output un set di informazioni che permettano di eseguire l'analisi in maniera immediata. L'output infatti contiene i valori in percentuale di corrispondenze tra i risultati ottenuti partendo dalle due sorgenti rispetto al totale dei candidati esaminati, utili per avere un riscontro di quale sia il livello di agreement tra le due sorgenti usate.
Come verrà discusso in seguito, il programma produce i risultati attesi, in linea con quelli riportati nell'articolo per il settore di Informatica, ed è in grado di fornire gli strumenti necessari per stabilire quanto le sorgenti usate siano complete rispetto a Scopus e WoS e quali settori siano più compatibili con le sorgenti open.
I risultati ottenuti sugli altri settori scientifico-disciplinari sono simili a quelli ottenuti per il settore 01-B1: nessun settore è quindi presente sulle sorgenti open in modo preponderante rispetto agli altri. Questi risultati portano alla medesima conclusione trattata nel paper, e cioè che ad oggi le sorgenti open in questione non sono ancora sufficientemente complete per sostituire quelle closed per calcolare gli indici dell'ASN.

In seguito saranno trattati e descritti i vari aspetti che compongono il background su cui il progetto si basa, l'architettura ed il workflow, i risultati ottenuti mediante una breve analisi in merito ai settori che non sono stati considerati nel paper e l'implementazione dettagliata degli algoritmi usati negli script.

Indice

Elenco delle figure

Elenco delle tabelle

Capitolo 1

Contesto

In questo capitolo verranno trattate le fondamenta su cui è stato costruito il progetto: l'Abilitazione Scientifica Nazionale, il movimento I4OC, Open Citations e COCI ed infine un precedente risultato che costituisce il punto di partenza del progetto. Infatti il progetto nasce come continuazione dell'articolo "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation"[2] che tratta il confronto tra diverse sorgenti di dati bibliografici open e closed. Per farlo è stata usata l'Abilitazione Scientifica Nazionale come benchmark: i risultati prodotti sulla base dei dati open sono stati confrontati con i risultati reali che sono invece prodotti sulla base dei dati closed in modo da ottenere un riscontro quantitativo sull'effettiva differenza tra le due sorgenti. Lo scopo di questo studio è applicare i medesimi criteri usati nell'articolo a tutti i settori scientifico-disciplinari bibliometrici mediante la creazione di uno strumento che permetta di automatizzare l'analisi. Nelle seguenti sezioni verranno approfonditi i vari aspetti che formano il background del progetto.

1.1 Abilitazione Scientifica Nazionale

Per poter accedere ai concorsi per la qualifica di professore di I o II fascia è necessario aver ottenuto l'Abilitazione Scientifica Nazionale (ASN), che co-

stituisce un requisito non sufficiente per l'accesso a cariche all'interno delle università. L'ASN viene riconosciuta ai candidati che hanno un numero di pubblicazioni sufficientemente elevato ed una quantità di citazioni sugli articoli tali da rendere il candidato autorevole e quindi più adatto a ricoprire un ruolo all'interno di un'università.

Il metodo usato per stabilire se il candidato rispetti i suddetti canoni ed i parametri, o soglie, che definiscono quantitativamente i requisiti minimi che i candidati devono soddisfare per ottenere l'abilitazione sono stati fissati dalla legge 240 del 30 dicembre 2010 [9].
Nella normativa vengono descritte nel dettaglio le procedure da seguire per l'ottenimento dell'abilitazione a seconda del settore e sotto-settore per il quale il candidato intende presentarsi, in particolare per ottenere l'abilitazione su settori bibliometrici è necessario che almeno due dei tre indici calcolati sulle pubblicazioni dei candidati siano superiori rispetto alle soglie prefissate.
I settori scientifico-disciplinari sono una categorizzazione degli insegnamenti usata in ambito universitario in Italia, si raggruppano in settori concorsuali, macro-settori ed aree. Nel caso di Informatica, il cui codice è 01/B1, 01 si riferisce all'area e B1 al settore concorsuale. In questa trattazione viene genericamente fatto riferimento ai settori concorsuali come settori ed alle aree come aree o macro-settori
Per settori bibliometrici si intendono, secondo la normativa, quelli facenti parte delle aree disciplinari dalla 1 alla 9 (ambito scientifico), che si contrappongono ai settori non bibliografici contenuti nelle aree disciplinari dalla 10 alla 14 (ambito umanistico).
Gli indici sono tre:

- numero di articoli pubblicati su journal
- numero di citazioni ricevute
- h-index

Il primo indice si riferisce al numero di pubblicazioni di articoli su riviste scientifiche, il secondo indice si riferisce alla somma di tutte le citazioni ricevute da una qualunque produzione del candidato ed il terzo indice è dato dal numero di articoli che hanno ricevuto un numero di citazioni almeno pari al numero di articoli stesso.
I dati relativi al candidato utili per il calcolo dei tre indici devono rispettare dei vincoli temporali: a seconda del settore e della fascia per cui il candidato fa richiesta infatti esiste una finestra temporale prefissata che definisce il periodo entro il quale gli articoli e le citazioni devono essere state prodotte per essere considerate valide.

I tre indici vengono calcolati dall'Agenzia Nazionale di Valutazione del sistema Universitario e della Ricerca (ANVUR) sulla base dei dati forniti da Scopus e Web of Science, due servizi privati che si occupano di raccogliere, gestire ed infine rivendere i dati relativi alle pubblicazioni scientifiche. La candidatura viene validata se almeno due indici su tre superano le soglie fissate dalla legge 240, le quali dipendono dal settore per il quale è stata effettuata la candidatura, dal livello (professore ordinario o associato) al quale il candidato vuole accedere, dal periodo di attività ed altri fattori correttivi, quale, ad esempio, periodi di indisposizione all'attività di ricerca.
Di seguito sono riportati due esempi di soglie organizzate per settore e fascia per evidenziare come, da settore a settore, i requisiti possano variare notevolmente:

Nel triennio dal 2016 al 2018 si sono tenute 5 sessioni per le quali sono state presentate complessivamente 41684 candidature. Poiché è possibile per uno stesso candidato fare domanda per settori o fasce diverse il dato non tiene conto delle ripetizioni dei candidati, ma considera ogni richiesta separatamente.

Criteri simili vengono usati, anche per altre applicazioni, in altri paesi oltre all'Italia, quali Germania, Finlandia e Norvegia [2]. Da ciò si evince un'im-

Informatica (01/B1)					
I Fascia			II Fascia		
9	304	10	4	157	7

Tabella 1.1: Soglie per il settore 01-B1

Logica Matematica e Matematiche Complementari (01/A1-Mat/04)					
I Fascia			II Fascia		
4	4	2	2	4	1

Tabella 1.2: Soglie per il settore 01-A1

portanza sempre maggiore legata alla produzione scritta in tutto il mondo scientifico.

1.2 Initiative for Open Citations e COCI

Questa crescente necessità di poter accedere liberamente ai dati relativi alle pubblicazioni scientifiche ha portato nel 2017 alla formazione dell'Initiative for Open Citations(I4OC) [5], ovvero un progetto il cui scopo è quello di raccogliere, catalogare e rendere pubblici e liberamente accessibili i dati relativi alle citazioni delle produzioni scientifiche come Linked Open Data usando tecniche di Semantic Web. I4OC ha come obiettivo quello di rendere i dati strutturati secondo formati standard per fare si che siano accessibili in maniera programmatica; separabili, ovvero utilizzabili singolarmente, senza quindi dover ottenere anche tutti i riferimenti relativi agli articoli stessi; ed infine open, ovvero liberamente accessibili ed usabili senza vincoli legati alla proprietà.
Molti publisher, tra cui American Geophysical Union, Association for Computing Machinery, BMJ, Cambridge University Press, Cold Spring Harbor Laboratory Press, EMBO Press, Royal Society of Chemistry, SAGE Publishing, Springer Nature, Taylor Francis, and Wiley hanno deciso di collaborare alla popolazione degli achivi di Open Citations portandoli nel giro di

due anni a coprire più della metà delle pubblicazioni presenti su Crossref [8].

Uno dei componenti di Open Citations è l'Open Citations Index of Crossref Open Doi-to-Doi Citations (COCI). COCI è nato nel 2010 ad opera di David Shotton presso l'Università di Oxford come progetto annuale per le creazione di un repository open contenente i dati delle citazioni accademiche e nel 2015 l'Università di Bologna, nella fattispecie Silvio Peroni, ha preso parte al progetto ospitando il progetto presso il DISI.
Nel corso degli anni COCI è diventato mano a mano più evoluto e completo, tanto da arrivare a raccogliere 624 milioni di citazioni[3] accessibili liberamente attraverso vari mezzi. Proprio per via del continuo sviluppo e ampliamento di COCI è interessante monitorare il suo stato rispetto ad altri servizi che gestiscono i medesimi dati, in particolare rispetto a servizi ritenuti affidabili per la loro completezza come Scopus e Web of Science.

1.3 Risultati precedenti

Il progetto di questa tesi nasce da uno studio pubblicato nell'articolo "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation" nel 2019 che si è occupato di analizzare il grado di maturità e completezza degli archivi di COCI. Per farlo sono stati confrontati i risultati dell'Abilitazione Scientifica Nazionale con i risultati calcolati sulla base delle citazioni presenti in COCI, con la finalità di ottenere un dato quantitativo sulla discrepanza tra sorgenti proprietarie come Scopus e Web of Science (usati per calcolare i dati reali da ANVUR) e COCI.
L'esperimento è stato svolto sul settore 01-B1 (Informatica) ed i dati necessari sono stati raccolti da Crossref e DBLP per quanto concerne gli articoli e COCI per le citazioni. DBLP è un archivio online che si occupa di raccogliere ed indicizzare le risorse bibliografiche relative al settore di Computer Science, rendendole disponibili anche mediante l'invocazione di API come nel caso di Crossref.

I risultati hanno messo in evidenza come sia effettivamente presente una notevole differenza tra le sorgenti dati, favorendo quelle proprietarie, e portando alla conclusione che Open Citations non sia ancora pronta per sostituire Scopus e Web of Science. Infatti su 518 candidati per il ruolo di professore ordinario solo il 59,07% ha ottenuto un risultato congruente a quello calcolato da ANVUR, mentre su 757 candidati per il ruolo di professore associato solamente il 57,60%.
Il risultato riportato fa riferimento ai dati relativi ai soli CV dei candidati, per i risultati ottenuti mediante l'impiego di DBLP o unitamente CV e DBLP si fa riferimento all'articolo stesso.
Di seguito sono riportate alcune tabelle estratte dall'articolo che riassumono il risultato:

Full Professor (518 candidates)	
Overall	59.07%
Journals	89.77%
Citations	50.77%
h-index	59.65%

Tabella 1.3: Risultati per la prima fascia

Associate Professor (757 candidates)	
Overall	57.60%
Journals	80.58%
Citations	49.14%
h-index	62.35%

Tabella 1.4: Risultati per la seconda fascia

Come si può osservare, nonostante il dato relativo ai journal sia alto, globalmente si presenta un forte distacco tra COCI ed i dati usati da ANVUR.
Tenendo però in considerazione il fatto che Open Citation è in evoluzione continua e che l'esperimento è stato svolto sul solo settore 01-B1, si è resa

necessaria la creazione di uno strumento che permetta di replicare l'esperimento in maniera più immediata su tutti i settori, in modo tale da monitorare lo stato di Open Citation e, più in generale, delle sorgenti open con l'intento di verificare se e quando le sorgenti open potranno sostituire le sorgenti closed.

Un altro risultato rilevante per il progetto è quello riportato nell'articolo "Evaluating the Availability of Open Citation Data" [4] del 2019 in cui è stata misurata la quantità di citazioni open presenti in COCI rispetto alla globalità delle produzioni presenti in Crossref. Dallo studio è emerso che solo il 24,22% delle pubblicazioni presenti su Crossref possedevano dati sulle citazioni open, mentre il 16.62% avevano citazioni private e le restanti non avevano citazioni registrate.

Un altro articolo degno di nota è "Funders should mandate open citations"[1] del 2018 in cui viene spiegata l'importanza del rendere accessibili i dati delle citazioni liberamente e del perché sia un processo che presenta numerose difficoltà e ostacoli, tra gli altri l'esistenza di gruppi che vendono questi dati e la scarsa collaborazione degli autori nella pubblicazione dei dati su Crossref.

Capitolo 2

Descrizione del progetto

In questo capitolo verranno trattate in maniera discorsiva e descritte le varie fasi che portano ai risultati in output, senza però addentrarsi nei dettagli implementativi che verranno invece discussi nel capitolo successivo.

2.1 Dati in input

2.1.1 Sorgenti open

Lo scopo del progetto è quello di calcolare gli indici dell'Abilitazione Scientifica Nazionale su tutti i settori disciplinari usando i dati raccolti da fonti open ed incrociare i dati con i risultati effettivi dell'ASN che sono stati calcolati da ANVUR con i dati forniti da Scopus e WoS.
Le sorgenti di dati pubbliche usate sono Crossref e COCI: i dati raccolti da Crossref riguardano gli articoli, mentre quelli raccolti da COCI riguardano le citazioni. In particolare i dati relativi agli articoli sono utili per distinguere le produzioni pubblicate su journal da quelle non pubblicate su journal e per ottenere la data di pubblicazione degli articoli, mentre i dati relativi alle citazioni sono utili per calcolare il numero di citazioni ricevute dalle produzioni e l'h-index del candidato.

2.1.2 File in input

Al programma viene fornito in input un file tsv generato a partire dai CV dei candidati contenente tutte le informazioni che è stato possibile ottenere dal curriculum e dalla sottoscrizione della domanda, in particolare i seguenti dati:

- sessione a cui il candidato ha preso parte
- la fascia per la quale il candidato ha sottoscritto la domanda
- il settore disciplinare
- l'id univoco del candidato
- l'elenco dei doi delle produzioni del candidato
- i risultati reali ottenuti sui tre indici
- le soglie relative ai tre indici

Tutti i CV in formato PDF vengono parsati da un tool in grado di ricostruire i dati ivi presenti in un formato tabellare, è possibile quindi che alcune informazioni presenti nei CV non vengano riconosciute dal parser e quindi non siano riportate nel file di input.

Di seguito verrà descritto il workflow organizzato in 4 fasi, ciascuna dipendente dalle precedenti, come riportato nello schema seguente:

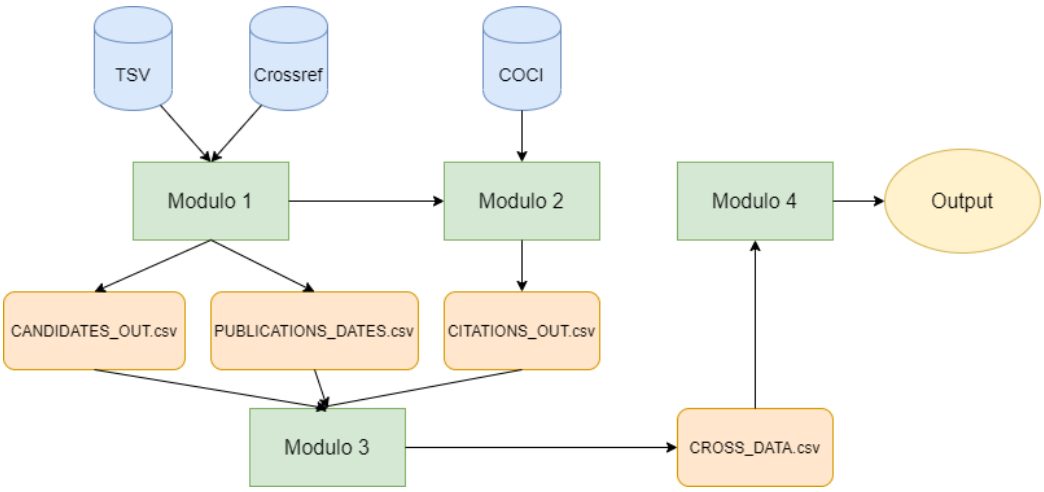

Figura 2.1: Schema dell'architettura e del workflow

2.2 Fase 1: analisi dei candidati

La prima fase si occupa di elaborare il file tsv in input per ottenere i dati relativi agli articoli pubblicati: per farlo viene interrogata un'API di Crossref che restituisce tutte le informazioni presenti sul database di Crossref relative al DOI presente nella richiesta. Quindi, candidato per candidato, viene invocata l'API Crossref su ogni DOI presente nel file tsv al fine di sapere se il DOI in questione sia o meno collegato ad un articolo pubblicato su journal e di ottenere la data di pubblicazione di quel DOI.
Dopodiché le informazioni ottenute vengono memorizzate in un file csv intermedio in cui vengono riportate anche gli altri dati utili presenti nel tsv iniziale che verranno usati negli step successivi.In aggiunta al csv intermedio

dei candidati viene anche generato un csv contenente la data di pubblicazione degli articoli che lo script è stato in grado di recuperare.

2.3 Fase 2: analisi delle citazioni

La seconda fase si occupa di calcolare i dati relativi alle citazioni e di produrre un secondo file intermedio che contiene i risultati strutturati in maniera tale da essere fruibili in maniera immediata. I dati presenti in COCI vengono recuperati mediante un file csv contenente il dump del database COCI in un certo istante, l'ultimo ad essere stato reso disponibile è aggiornato a Gennaio 2020, ma per i test è stato usato il dump che contiene lo stato di Novembre 2018.
Siccome il dump è un file di dimensioni notevoli (diverse decine di GB) si è deciso di fornire il suddetto in input, quindi previo download, piuttosto che scaricare dinamicamente la versione più aggiornata.
Alternativamente COCI mette a disposizione a sua volta delle API pubbliche che restituiscono i dati relativi al singolo DOI, ma poichè l'interrogazione avrebbe riguardato circa un milione e mezzo di DOI la soluzione di accedere direttamente ai dati di COCI mediante csv è risultata più veloce.

Infatti complessivamente il collo di bottiglia sul piano temporale è dato dall'interrogazione a Crossref che, nonostante proceda in parallelo su più DOI contemporaneamente, richiede molto più tempo rispetto a tutte le altre fasi del progetto.
Optando per le API di COCI piuttosto che per il dump si otterrebbe l'effetto di raddoppiare i tempi necessari all'elaborazione delle prime due fasi poichè per ogni DOI sarebbe necessario interrogare l'API Crossref e COCI.
Lo script parsa il file csv contenente il dump di COCI e per ogni citazione verifica se il DOI dell'articolo citato sia o meno presente tra i DOI dei candidati e, nel primo caso, la citazione viene registrata in un dizionario contenente il numero di citazioni associate a ciascun DOI per ogni sessione e fascia.

Quando il parsing termina il dizionario generato viene trascritto in un file csv intermedio che contiene l'associazione DOI-numero di citazioni.

2.4 Fase 3: calcolo degli indicatori

Il passo successivo usa i file generati nelle fasi precedenti per calcolare gli indici dell'Abilitazione Scientifica Nazionale: per ogni candidato si procede a contare i DOI degli articoli pubblicati su journal, il numero di citazioni ricevute da tutti i DOI ed a calcolare l'h-index del candidato.
Per ogni DOI viene verificata la validità temporale dell'articolo secondo quanto stabilito dalla legge 240 e se si tratta di un articolo journal viene incrementato un contatore, dopodiché vengono recuperate le citazioni per quell'articolo dal file csv intermedio contenente le citazioni per il DOI e si procede ad incrementare il numero di citazioni ricevute dal candidato, il valore delle citazioni ricevute dal doi viene poi registrato in un array che viene usato per il calcolo dell'h-index.
L'h-index si ottiene ordinando il suddetto array in ordine decrescente ed iterando sullo stesso fino a quando il valore dell'i-esimo elemento non è minore del valore dell'iteratore: quando si verifica la condizione d'arresto il valore dell'h-index corrisponde al valore dell'iteratore.
I dati ottenuti vengono registrati all'interno di un quarto file csv intermedio che permette di incrociare i dati reali con i dati ottenuti in maniera semplice

2.5 Fase 4: produzione ed analisi dei risultati

L'ultimo passo consiste nell'incrociare i risultati dell'ASN con i risultati generati dai dati open: sia gli indici relativi ai dati reali che quelli relativi ai dati calcolati vengono confrontati con le soglie e nel caso in cui entrambi i set di indici soddisfino o meno le soglie si ha un match tra i due, altrimenti se i due risultati sono differenti si ha un mismatch. Alla fine dell'elaborazione si hanno a disposizione i dati necessari per calcolare la percentuale di candidati

che hanno ottenuto lo stesso risultato partendo da dati closed e dati open, viene quindi elaborata una percentuale di corrispondenza tra i risultati reali e quelli calcolati sui tre indici e sul risultato globale, sia divisi settore per settore sia raggruppati insieme, in modo da avere un quadro complessivo sul match e mismatch tra i risultati ottenuti.
Per semplificare la lettura dei risultati vengono anche prodotti dei grafici che riassumono l'esito dell'esperimento: per ogni settore viene creato un istogramma che mostra le percentuali dei tre indici e dell'overall, ovvero il risultato ottenuto all'ASN, sia per la prima che per la seconda fascia.
Di seguito è riportato il grafico precedentemente descritto ottenuto sul settore 01-B1:

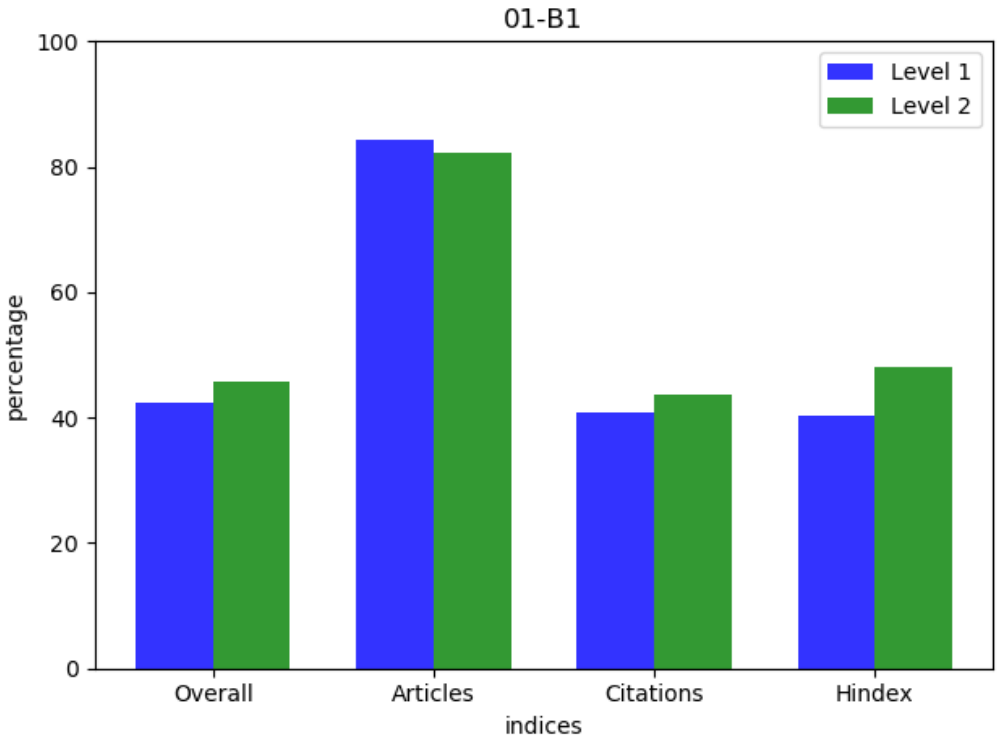

Figura 2.2: Esempio di grafico prodotto

Viene anche generato un grafico che contiene i dati dell'overall in percentuale per ogni area ordinate in ordine crescente, che mette in evidenza quali siano i macro settori che più si prestano ad un utilizzo delle sorgenti open per questo tipo di analisi: più la percentuale è alta infatti più è stato possibile recuperare informazioni simili a quelle presenti su Scopus e WoS, quindi i dati e le informazioni relative a quel settore sono molto abbondanti tra le sorgenti open rispetto a macro settori che ottengono percentuali inferiori.

Il progetto è strutturato in modo tale da essere modulare: non si tratta di un processo a cascata unico, ma di quattro fasi distinte che possono operare atomicamente dati i file in input. In questo modo è per esempio possibile ripetere le fasi 2, 3 e 4 in poco tempo sullo stesso set di dati o su set di dati differenti eventualmente cambiando i parametri di configurazione relativi alle finestre temporali. La modularità rende anche possibile un approccio divide-et-impera nella fase 1: invece che dare in pasto ad un'unica macchina un file tsv monolitico si può suddividere il tsv in più file separati e lanciare in parallelo su più calcolatori facendo poi una merge dei risultati ottenuti alla fine dell'elaborazione. In questo modo i tempi necessari all'elaborazione della fase 1 viene ridotto linearmente alla quantità di calcolatori impiegati.
La fase che richiede più tempo infatti è la prima poichè è necessario effettuare una chiamata http per ogni DOI presente nel file tsv di partenza. Mediamente l'interrogazione e l'elaborazione dell'output impiegano circa 1 secondo e, nonostante le chiamate vengano effettuate in parallelo, complessivamente l'elaborazione per il concorso ASN di riferimento per questo progetto richiede giorni. Per questo motivo la possibilità di dividere il job su più calcolatori potrebbe risolvere il problema del bottleneck generato dalla prima fase abbattendo i tempi necessari al suo completamento.

Per eseguire l'analisi discussa in questa trattazione che ha riguardato circa 45 mila candidati sono stati necessari circa 3 giorni per completare il primo modulo, 20 minuti per completare il secondo e pochi secondi per il terzo

ed il quarto; risulta quindi evidente che il tempo necessario per completare l'esecuzione dell'intero workflow è dettato dalla fase 1.

Capitolo 3

Risultati

In questo capitolo verranno discussi i risultati ottenuti dallo script sui candidati che hanno presentato domanda per l'ultima Abilitazione Scientifica Nazionale, ovvero quella svoltasi nel triennio 2016-2018. La discussione dei risultati verterà sul corretto funzionamento del programma mediante un confronto con i risultati ottenuti dall'articolo "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation" [2] e su una prima analisi dei risultati ottenuti per tutti i settori che non sono stati considerati nel suddetto articolo.

3.1 Correttezza e differenze rispetto ai risultati precedenti

Il settore 01-B1 è particolarmente rilevante in questa trattazione perché è stato l'unico settore sul quale sono stati eseguiti i calcoli nell'articolo [2], pertanto è fondamentale che i risultati ottenuti dal programma sul settore 01-B1 siano ragionevolmente comparabili a quelli ottenuti precedentemente per comprovare la correttezza del software.

Di seguito sono riportate le tabelle contenenti i risultati dell'articolo e quelli generati dal programma del programma raggruppate per fascia:

Full Professor (518 candidates)		
Indici	Risultati paper	Risultati programma
Overall	59.07%	42.36%
Journals	89.77%	84.33%
Citations	50.77%	40.81%
h-index	59.65%	40.43%

Tabella 3.1: Risultati per la fascia di Professore Ordinario

Associate Professor (757 candidates)		
Indici	Risultati paper	Risultati programma
Overall	57.60%	45.73%
Journals	80.58%	82.12%
Citations	49.14%	43.60%
h-index	62.35%	48.13%

Tabella 3.2: Risultati per la fascia di Professore Associato

Come si può vedere i risultati ottenuti dal programma sono molto differenti da quelli riportati nel paper. Questo discostamento è dovuto a molteplici ragioni che verranno elencate ed analizzate di seguito.
La prima e maggiore differenza rispetto all'articolo è insita nei dati in ingresso che sono stati usati per eseguire il calcolo: i dati presenti nel tsv usato per questo esperimento infatti sono differenti in quanto molti DOI non sono stati correttamente riportati, con conseguente riduzione dei risultati in output.
L'analisi è stata quindi ristretta ad un insieme molto limitato di candidati per cui i DOI in input sono stati controllati manualmente.

Di seguito sono riportati i dati ottenuti:

Script Paper			
ID	Articles	Citations	H-Index
6754	12	129	8
32924	23	4467	12
26344	5	355	10

Tabella 3.3: Risultati ottenuti dallo script usato nel paper

Progetto			
ID	Articles	Citations	H-Index
6754	12	129	7
32924	23	4475	12
26344	5	356	9

Tabella 3.4: Risultati ottenuti dal programma

Un altro elemento che contribuisce al distacco tra i risultati è l'utilizzo che viene fatto della sorgente DBLP da parte dello script usato per l'articolo: i dati presenti in DBLP infatti sono stati usati per completare quelli ricavati dai CV dei candidati, in particolare se i dati ottenuti da DBLP contenevano DOI non riportati nel tsv, allora questi venivano aggiunti alla lista dei DOI del candidato. Poiché questa meccanica non è stata implementata del programma, nei dati usati per questo esperimento sono presenti meno DOI rispetto a quelli usati per il paper.

Infine va menzionato il fatto che le sorgenti usate sono in aggiornamento continuo. Questo fa sì che a seconda del momento nel quale vengono invocati i servizi di Crossref e COCI i dati forniti possano variare, per quanto ai fini dell'analisi questo comporti un impatto di magnitudo inferiore rispetto ai precedenti punti.

Poiché, tenendo conto delle precisazioni fatte, i risultati del paper e quelli ottenuti dal programma sono comparabili, allora si può ipotizzare che gli algoritmi usati nello script siano corretti ed i risultati prodotti siano attendibili anche negli altri settori, ma poiché l'analisi ha riguardato un insieme ristretto di candidati è necessario approfondire con ulteriori esperimenti.
Di seguito è riportato il grafico generato dal programma relativamente al settore 01-B1:

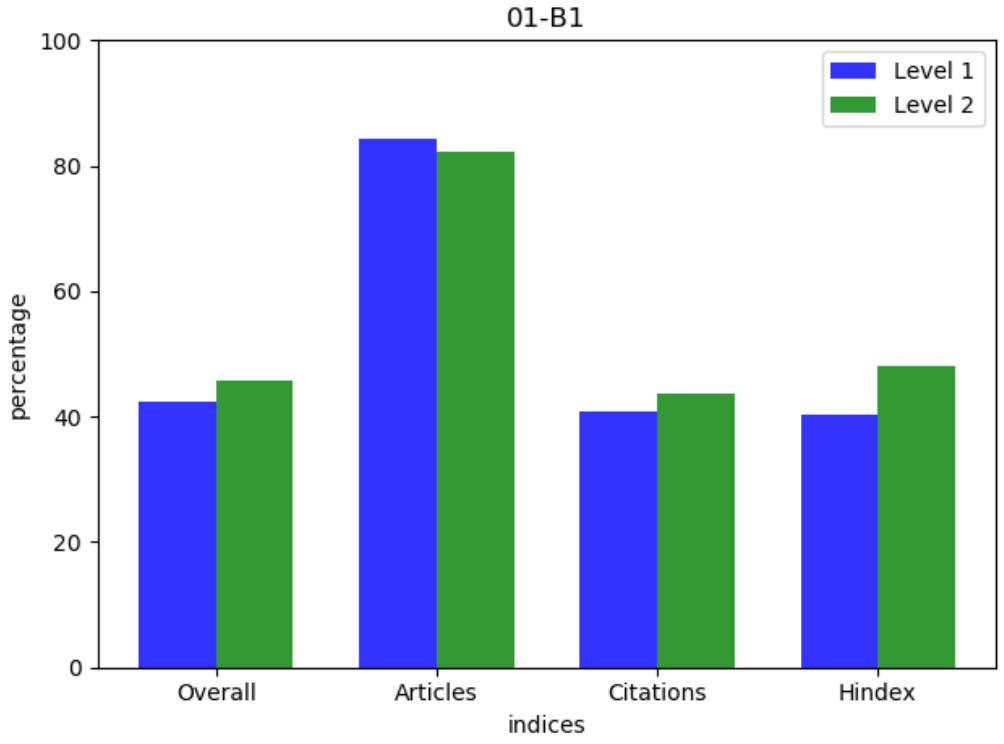

Figura 3.1: Grafico del settore di Informatica

Come si può vedere le percentuali dell'Overall, Citations e h-index sono molto basse e non raggiungono il 50% del totale dei candidati elaborati per quel settore; invece la percentuale degli Articles è decisamente maggiore. Da ciò si potrebbe ipotizzare che Crossref sia più completa per quanto riguarda i

dati sugli articoli rispetto a COCI per quanto concerne le citazioni: per molti candidati infatti sono presenti i dati delle pubblicazioni e non quelli delle citazioni, per questo motivo è presente un divario tra l'indice Articles e l'indice Citations. La percentuale dei matching sull' h-index è in linea con quella delle citazioni, il secondo indice infatti è correlato al primo, così come l'indice Overall dipende dagli altri tre.
Si evidenzia inoltre una leggera differenza tra i risultati dei candidati di fascia 1 e 2, ma non sufficientemente elevata da portare a conclusioni di maggiore predisposizione di una fascia rispetto all'altra.

3.2 Esposizione dei risultati ottenuti sugli altri settori

Come è stato ampiamente discusso in questa trattazione, l'obiettivo è incentrato sulla possibilità di analizzare tutti i settori scientifico-disciplinari, non solo il settore 01-B1. Di seguito sono riportati alcuni grafici relativi ai risultati ottenuti sui settori di un sottoinsieme di aree disciplinari:

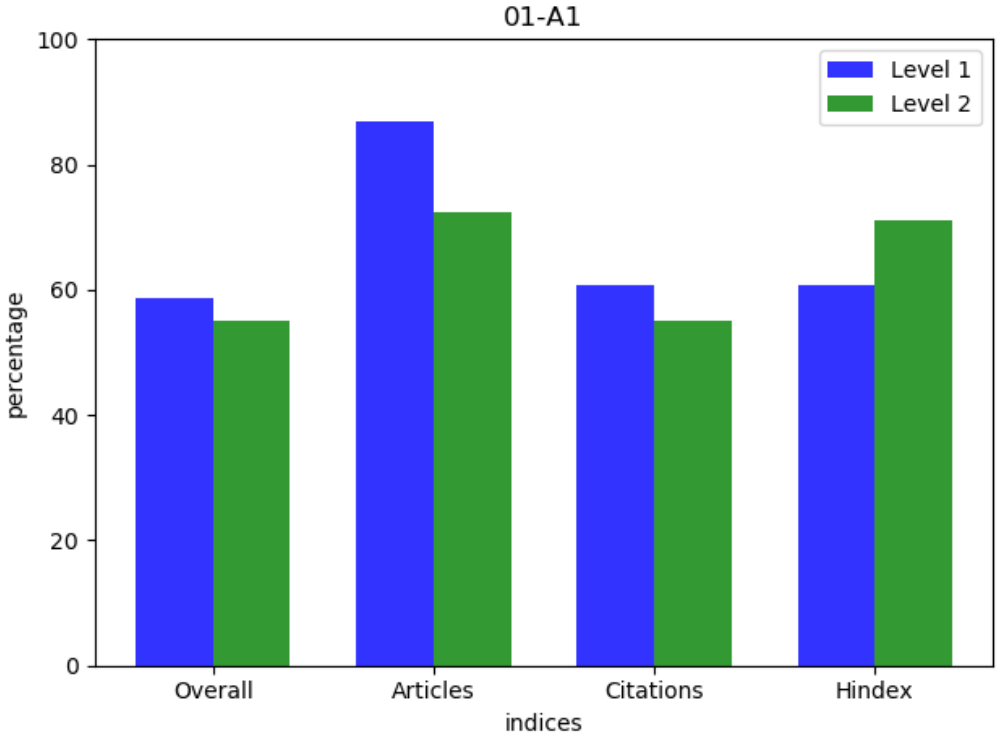

Figura 3.2: Grafico del settore "Logica matematica e matematiche complementari"

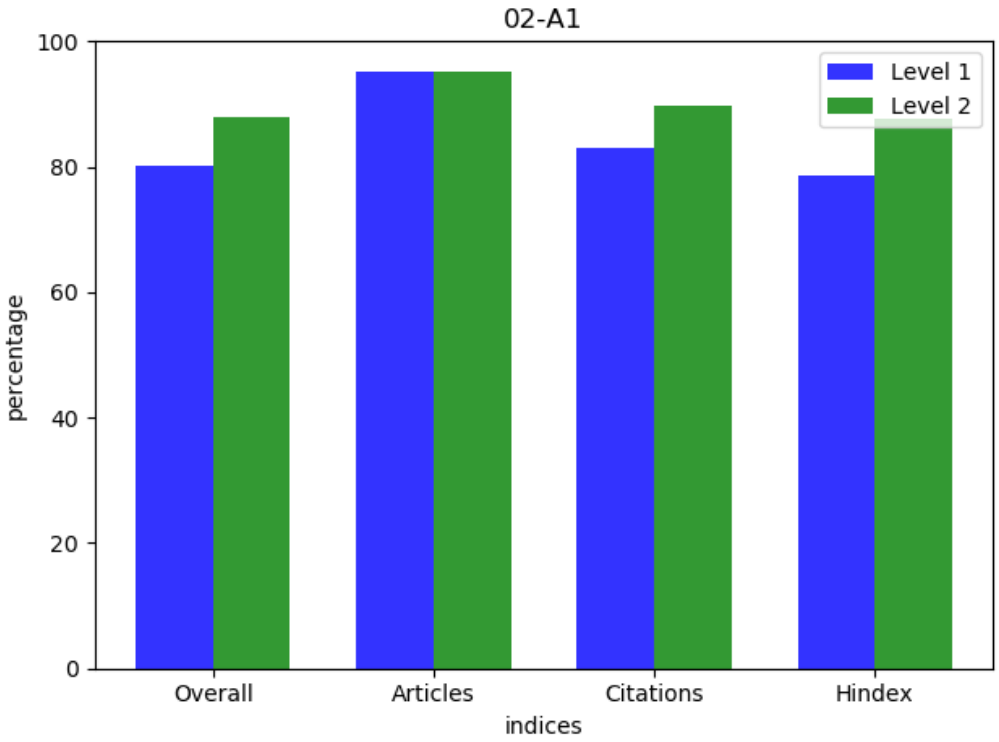

Figura 3.3: Grafico del settore "Fisica sperimentale delle interazioni fondamentali"

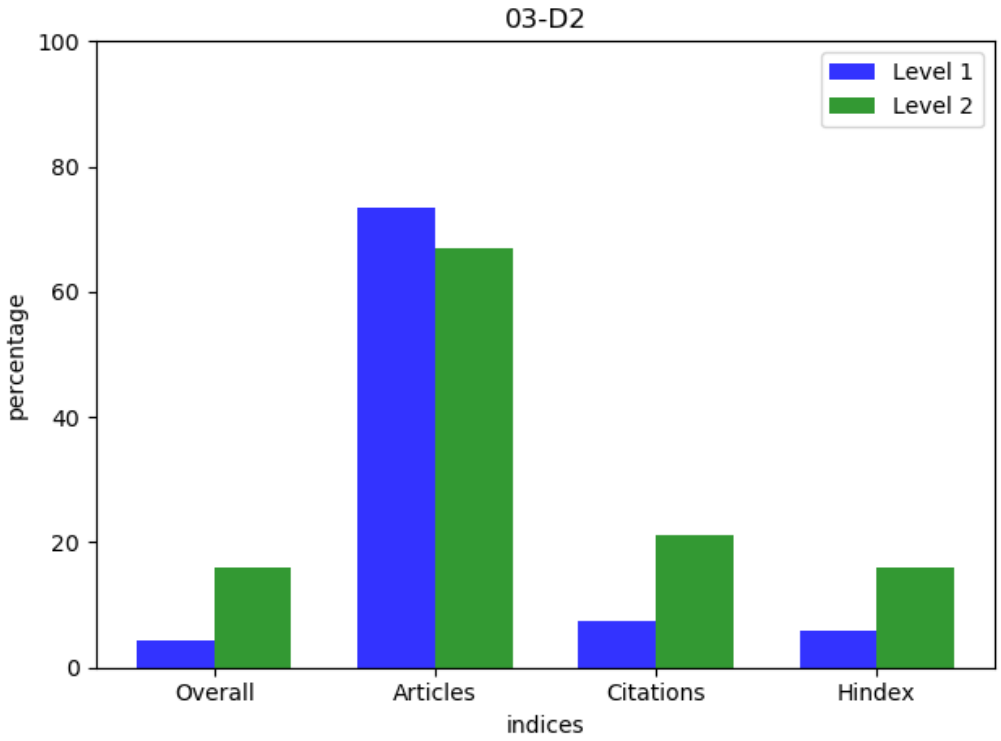

Figura 3.4: Grafico del settore "Tecnologia, socioeconomia e normativa dei medicinali"

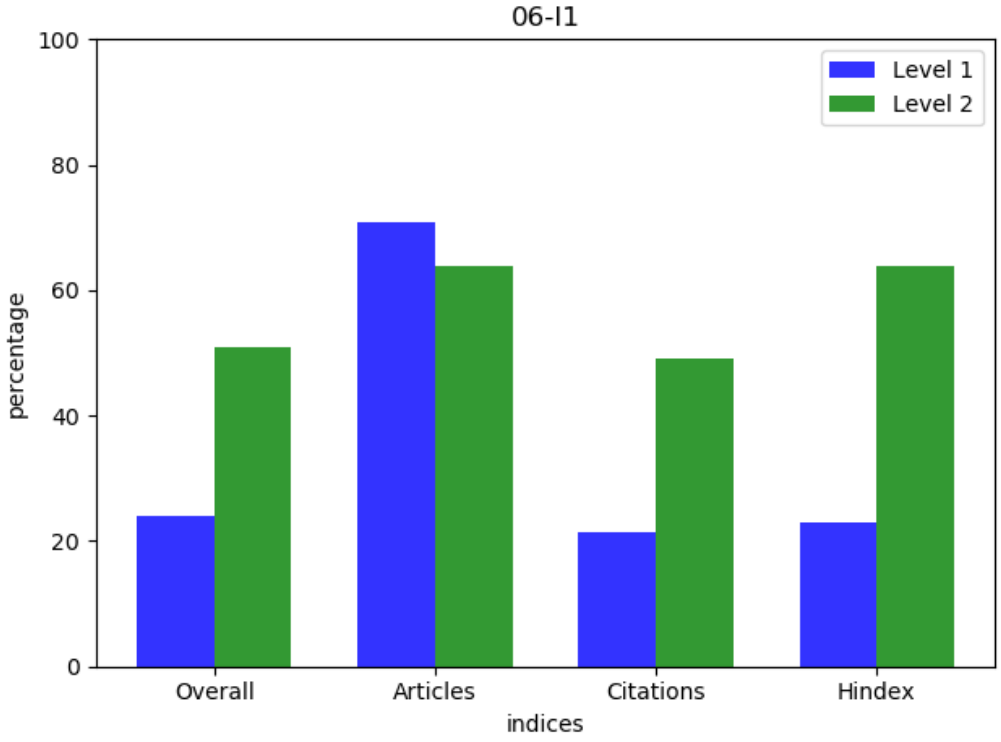

Figura 3.5: Grafico del settore "Diagnostica per immagini, radioterapia e neuroradiologia"

Il seguente grafico invece rappresenta la media dei risultati ottenuti su tutti i settori analizzati, utile per esemplificare il risultato globale dell'esperimento:

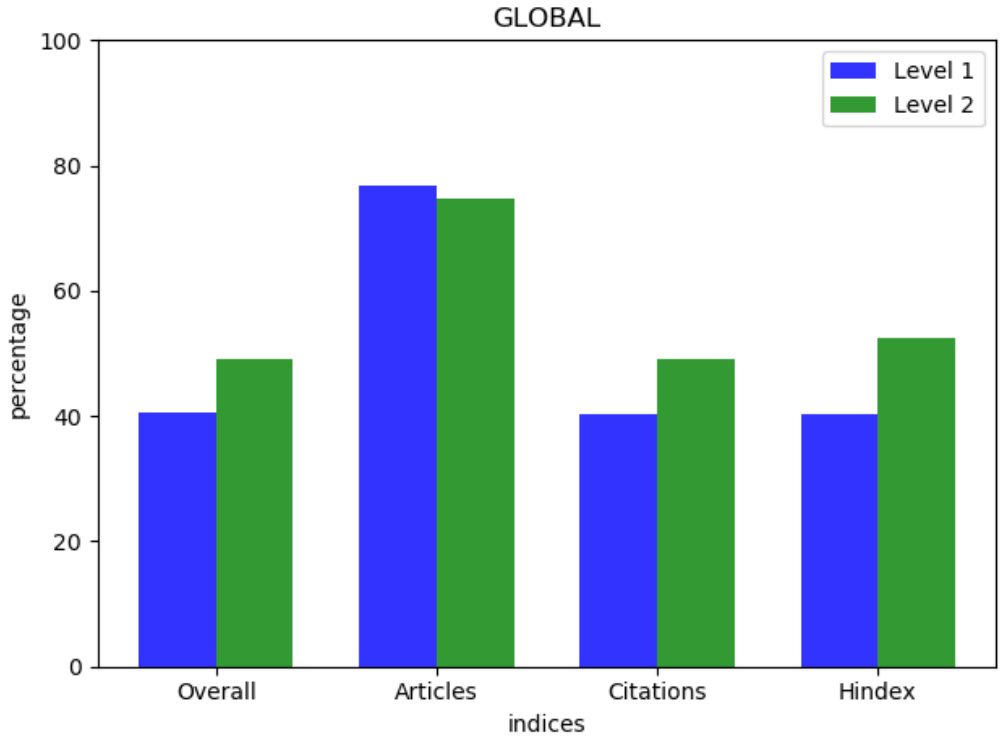

Figura 3.6: Grafico globale

In tutti i settori sono state ottenute delle percentuali di matching molto basse per quanto concerne le citazioni e l'h-index, mentre per l'indice degli articoli pubblicati su journal i risultati sono migliori. Questi dati portano alla conclusione che, sebbene Crossref permetta di ottenere dei risultati che tendono ad avvicinarsi a quelli reali, COCI sia ancora troppo incompleto rispetto alle sorgenti closed. Infine gli ultimi due grafici presentati raffigurano le aree organizzate in ordine crescente divise per fasce: sull'asse delle ordinate è presente la media dei risultati in percentuale dell'overall ottenuti su ogni settore appartenente all'area in oggetto.

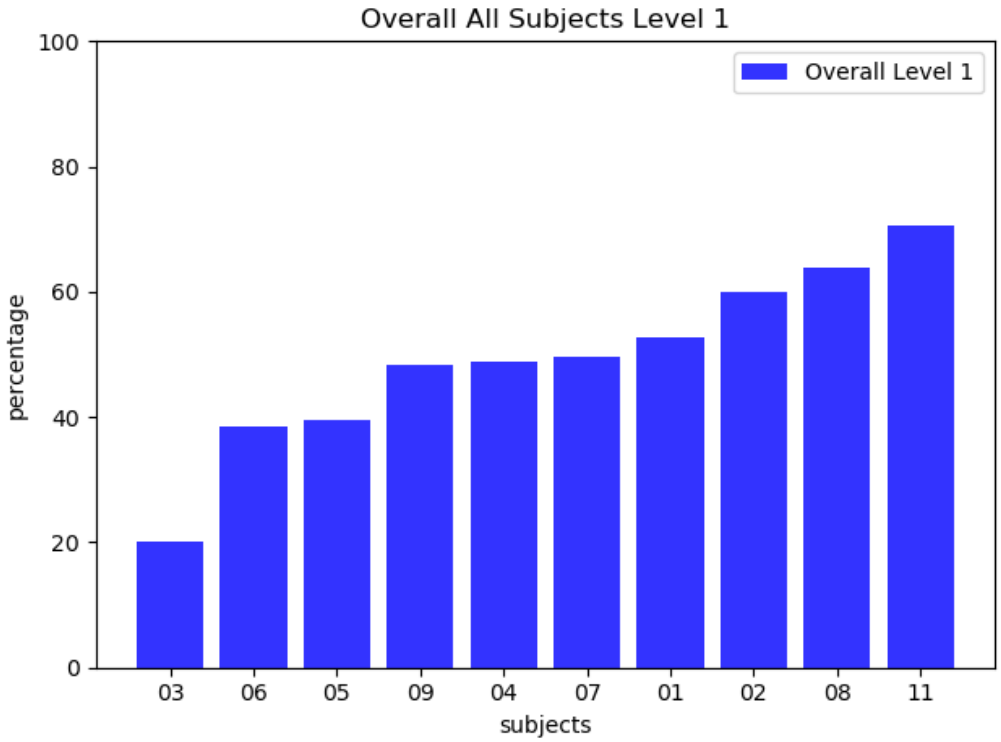

Figura 3.7: Grafico delle aree per la prima fascia

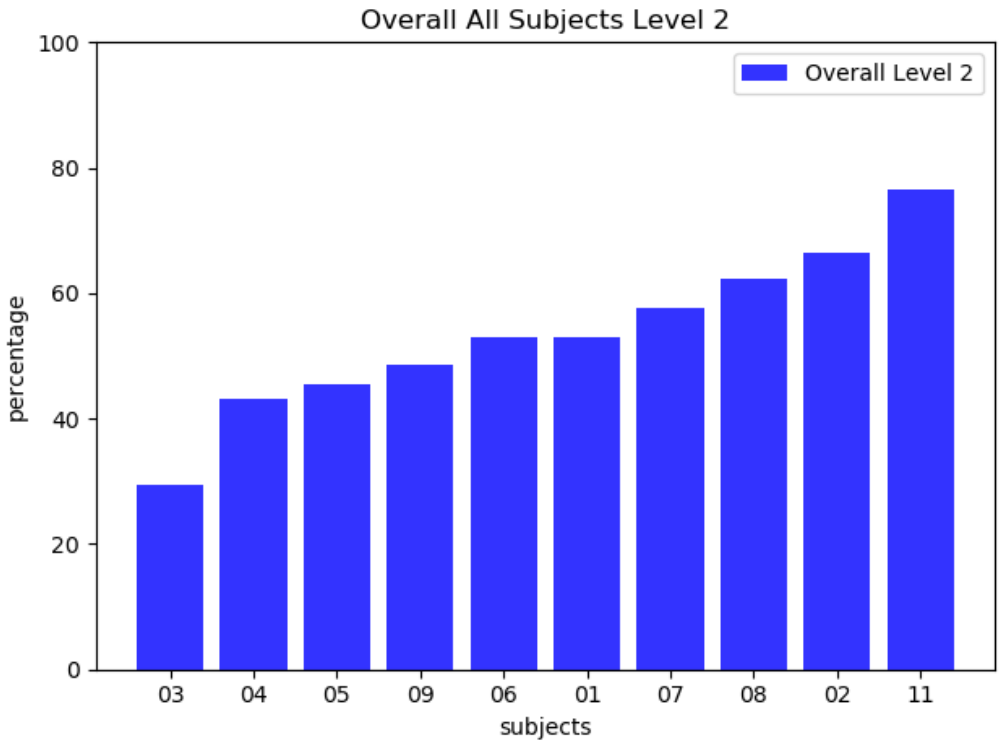

Figura 3.8: Grafico delle aree per la seconda fascia

Da questi grafici emerge che l'area i cui settori sono più presenti sulle sorgenti open è la 02, ovvero l'area delle Scienze Fisiche. I risultati, sebbene migliori rispetto alle altre aree, non sono sufficientemente buoni da affermare che le sorgenti closed possano essere sostituite da sorgenti open per quell'area.

3.3 Discussione dei risultati ottenuti

I risultati riportati nella precedente sezione, per quanto siano traslati rispetto a quelli riportati nel paper, sono comunque in linea con la conclusione descritta nell'articolo, e cioè che al momento non è plausibile una sostituzione delle sorgenti closed con sorgenti aperte poiché i dati derivanti da questa sostituzione sarebbero eccessivamente incompleti rispetto alla reale produzione dei candidati. Fatta eccezione per pochi settori, come ad esempio il settore 02/A1 - "Fisica sperimentale delle interazioni fondamentali" che ha ottenuto ottimi risultati, per tutti gli altri l'agreement tra i risultati prodotti con sorgenti open ed i risultati reali è molto basso.
La conseguenza della sostituzione sarebbe quindi un sistematico peggioramento degli indici bibliometrici calcolati per l'Abilitazione Scientifica Nazionale che porterebbe in molti casi a non superare le soglie in maniera immeritata.

Però poiché uno degli scopi principali del progetto era quello di costruire uno strumento che permettesse un'analisi programmatica delle sorgenti open in relazione all'ASN, scopo che, come è stato dimostrato, è stato raggiunto, in futuro sarà possibile usare il medesimo strumento per ripetere le analisi in modo automatico con la finalità di monitorare lo stato di completezza delle sorgenti dati in questione.

Capitolo 4

Descrizione degli aspetti implementativi

Nel capitolo precedente è stata descritta una panoramica del workflow senza la trattazione dell'implementazione delle meccaniche che costituiscono il progetto. In questo capitolo verranno discussi i dettagli implementativi del workflow precedentemente esposto.

Il progetto è interamente scritto in Python (3.7.4), linguaggio general purpose che ben si presta ad eseguire analisi di questo tipo per via della sua semplicità e duttilità. Sono state usate svariate librerie standard di Python per la gestione dei file csv, dei formati data, delle chiamate http e della gestione I/O su file system oltre a librerie specifiche per assolvere a task particolari come la libreria crossref.restful per gestire le invocazioni a Crossref in maniera sostenibile e la libreria matplotlib per la generazione dei grafici.

L'albero dei file che compongono il progetto a verrà fatto riferimento nelle sezioni successive è la seguente:

- asn
  - candidates.py
  - citations.py
  - cross_data.py
  - graphs.py
  - menu.py
  - tools.py
- configuration.py
- data
  - images
  - output
  - tmp
  - CANDIDATES.tsv
  - CANDIDATES_OUT.csv
  - CITATIONS_OUT.csv
  - COCI_DATA.csv
  - CROSS_DATA.csv
  - PUBLICATIONS_DATES.csv
- images
- main.py
- OUTPUT.txt

4.1 Configurazione

Il progetto è impostato su 4 moduli, ciascuno slegato dagli altri in modo da permettere un'esecuzione atomica dei vari blocchi logici. I moduli però sono dipendenti dai risultati intermedi generati dai moduli precedenti, si è quindi resa necessaria la creazione di un file di configurazione che permettesse di definire con un punto di accesso unico tutti i parametri legati al file system. Questo viene fatto nel file configuration.py dove sono definiti, oltre alle locazioni dei file intermedi, i parametri statici globali che vengono usati all'interno degli script come le finestre temporali di validità degli articoli e delle citazioni, le date di riferimento delle sessioni dell'ASN e l'insieme dei settori scientifico-disciplinari che si vuole usare come filtro rispetto all'insieme di tutti i settori.
In questo modo è possibile configurare i parametri del progetto attraverso una modifica a questo file, avendo la possibilità di ripetere l'analisi usando parametri deversi in maniera immediata.

4.2 Il Main

L'esecuzione dello script viene lanciata attraverso un main che renderizza a terminale una scelta multipla su quale azione si voglia effettuare. Le opzioni riguardano i 4 moduli ad ognuno dei quali è associato un numero. Una volta effettuata la scelta viene eventualmente richiesto, a seconda del modulo scelto, di inserire il path relativo o assoluto al file che bisogna fornire in input. Per i file intermedi generati dallo script non viene richiesto il path poichè è definito all'interno del file di configurazione.
Il main quindi funge da controller per i vari moduli, che vengono richiamati a seconda della scelta effettuata dall'utente, e si occupa anche di istanziare le variabili globali ricavate dal file di configurazione per poi passarle alle funzioni invocate.

4.3 L'elaborazione dei candidato

Il primo modulo si occupa di gestire l'elaborazione dei candidati: viene fornito in input un file tsv contenente tutte le informazioni relative ai candidati (vedi 2.1) il quale deve essere processato per ricavare solo i dati che saranno poi usati e per recuperare le informazioni mancanti. In particolare è necessario ottenere per ogni DOI presente tra i DOI del candidato la tipologia di produzione e la data di pubblicazione: la prima informazione viene usata per il conteggio degli articoli pubblicati su journal, mentre la seconda è necessaria per verificare la validità rispetto alla finestra temporale dell'articolo stesso e delle eventuali citazioni ad esso correlate.
Nel caso in cui non fosse possibile recuperare la tipologia dell'articolo questo verrebbe escluso dagli articoli journal, ma sarebbe comunque potenzialmente valido per il calcolo delle citazioni e dell'h-index; invece nel caso in cui non si riuscisse ad ottenere la data di pubblicazione il DOI in esame non verrebbe usato nelle fasi successive in quanto sarebbe impossibile stabilire se l'articolo sia valido o meno temporalmente.

Il file tsv in input viene parsato e su ogni riga vengono estratti i dati relativi a sessione, fascia, settore, id del candidato, elenco dei DOI, indice reale degli articoli, indice reale delle citazioni, indice reale dell'h-index, la soglia degli articoli, la soglia delle citazioni e la soglia dell'h-index.
Dalla stringa contenente la lista dei DOI delle pubblicazioni del candidato viene poi ricavato un array contenente i suddetti DOI e viene lanciata l'esecuzione in parallelo sugli elementi dell'array appena generato di una funzione che si occupa di invocare i servizi di Crossref.
La funzione, attraverso la chiamata all'API di Crossref mappata dentro la libreria crossref.restful, ottiene tutte le informazioni open presenti in Crossref relative al DOI richiesto e restituisce la tipologia e la data di pubblicazione dell'articolo. La funzione ritorna anche, se presente, la lista degli autori che viene memorizzata in un dizionario a parte.

Quando l'esecuzione parallela sui DOI del candidato termina viene lanciata un'altra funzione che si occupa di riconoscere il nome e cognome del candidato tra quelli restituiti dalle invocazioni a Crossref: il dizionario degli autori contiene una serie di liste composte dagli autori che hanno preso parte alle singole produzioni del candidato, verosimilmente le sue generalità sono quelle che compaiono più frequentemente all'interno del dizionario. Vengono quindi conteggiate le occorrenze dei nomi e cognomi all'interno delle liste e, alla fine dell'elaborazione, la coppia nome-cognome con più occorrenze viene associata al candidato.
L'algoritmo usato presenta problemi nel caso limite in cui sia presente un solo DOI per il candidato, ma essendo che il nome e cognome non rappresentano informazioni essenziali per l'elaborazione e che si tratta di una casistica improbabile è lecito ritenere che questo sia trascurabile.

Successivamente i dati recuperati da Crossref unitamente a quelli già presenti nel tsv vengono trascritti su un csv intermedio (data/CANDIDATES_OUT.csv) che verrà usato per le elaborazioni successive. In CANDIDATES_OUT sono presenti anche, oltre ai dati precedentemente elencati, il nome e cognome dell'autore e la lista dei DOI journal, mentre le date di pubblicazioni vengono trascritte in un file csv a parte (data/PUBLICATIONS_DATES.csv) che associa ad ogni DOI la sua data di pubblicazione.
La scrittura sui file intermedi avviene alla fine dell'elaborazione del singolo candidato: questo da la possibilità di interrompere l'esecuzione all'occorrenza e di riprenderla successivamente. Quando lo script viene rilanciato vengono conteggiate le righe presenti nel csv intermedio ed il parsing del tsv riprende dal valore ottenuto nel conteggio. Ovviamente potrebbe succedere che non ci sia un rapporto uno a uno tra il numero di righe parsate ed il numero di righe riportate nel csv dei candidati, per questo motivo alla fine dell'elaborazione del tsv viene lanciata una funzione il cui scopo è quello di ripulire i due csv prodotti da tutti gli elementi ripetuti in modo da avere due file intermedi puliti.

4.4 L'elaborazione delle citazioni

Il secondo modulo si occupa di gestire l'elaborazione delle citazioni ricevute dai DOI dei candidati, per farlo viene usato il dump in formato csv del database di COCI come sorgente dati e la lista di tutti i DOI dei candidati come filtro. In COCI sono infatti presenti oltre 624 milioni di citazioni, pertanto costruire un dizionario che contenga l'associazione DOI-citazioni su un numero così alto di DOI generebbe problemi di memoria nel calcolatore. Pertanto vengono considerati solamente i DOI dei candidati, riducendo enormemente l'occupazione in memoria del dizionario generato.
Lo script parsa il csv di COCI riga per riga e se il DOI dell'articolo citato figura tra i DOI dei candidati quella riga contiene una citazione potenzialmente utile al calcolo. Poichè uno stesso DOI può figurare tra i DOI di due o più candidati che hanno sottoscritto candidature per fasce o sessioni differenti è necessario considerare la citazione in modo separato per ogni sessione e fascia, ovvero verificare che a seconda della sessione e della fascia che si sta considerando l'articolo rientri o meno all'interno della finestra temporale definita ed incrementare contatori separati per ogni coppia sessione-fascia. In questo modo l'elaborazione produce un output immediatamente usabile su ogni combinazione di sessione-fascia del candidato.
Per ogni citazione compatibile con i DOI dei candidati vengono iterate le sessioni presenti nel file di configurazione e per ogni sessione viene calcolata la finestra temporale su entrambe le fasce, se la data di pubblicazione della citazione rientra all'interno della finestra temporale della seconda fascia, che è più stringente, vuol dire che la citazione è valida sia per la prima che per la seconda fascia ed i loro contatori vengono incrementati, altrimenti viene verificata la validità per la prima fascia.

Viene quindi generato un dizionario strutturato come il seguente:

```
{
    doi_1: {
        sessione_1: {
            fascia_1: numero_citazioni,
            fascia_2: numero_citazioni
        },
        sessione_2: {
            fascia_1: numero_citazioni,
            fascia_2: numero_citazioni
        },
        ...
    },
    doi_2: {
        sessione_1: {
            fascia_1: numero_citazioni,
            fascia_2: numero_citazioni
        },
        sessione_2: {
            fascia_1: numero_citazioni,
            fascia_2: numero_citazioni
        },
        ...
    },
    ...
}
```

Alla fine dell'elaborazione il dizionario viene trascritto in un file csv intermedio (data/CITATIONS_OUT.csv) le cui righe contengono il DOI seguito da una colonna per sessione, all'interno della cui cella è presente il valore re-

gistrato dal counter sulla prima e sulla seconda fascia separate dal carattere "/". Di seguito è riportato un esempio del file CITATIONS_OUT.csv:

CITATIONS_OUT.csv				
doi	session_1	session_2	session_3	session_4
10.1038/nbt.1411	166/164	171/169	172/170	180/178
10.1002/cpe.993	222/219	226/216	228/218	232/222

Tabella 4.1: Esempio del file CITATIONS_OUT.csv

4.5 Il calcolo degli indici

Il terzo modulo si occupa di incrociare i dati prodotti dai primi due, calcolare i tre indici ASN per i candidati e generare un file csv che racchiude tutti i dati utili per il calcolo finale.
Vengono inizialmente generati tre dizionari a partire dai file intermedi CANDIDATES_OUT.csv, CITATIONS_OUT.csv e PUBLICATIONS_OUT.csv e per ogni candidato viene poi conteggiato il numero dei DOI journal: per ogni DOI presente nella lista dei DOI journal viene verificata la validità temporale attraverso la sessione e la fascia del candidato e la data di pubblicazione presente nel dizionario delle date di pubblicazione. Se la data di pubblicazione non è presente il DOI viene scartato, invece se la data è presente e rientra all'interno della finestra temporale il contatore degli articoli journal viene incrementato.

Successivamente viene calcolato l'indice delle citazioni: per farlo vengono iterati tutti i DOI del candidato e ad uno ad uno viene verificato che l'articolo sia valido temporalmente usando il medesimo procedimento descritto in precedenza per gli articoli journal, se lo è vengono ricavate le citazioni ricevute da quell'articolo dal dizionario delle citazioni usando come chiavi il DOI, la sessione e la fascia del canditato. Il numero di citazioni ricevute dal DOI in esame va ad incrementare un contatore globale per il candidato e

viene successivamente inserito in un array utile per il calcolo del terzo indice. Una volta che tutti i DOI del candidato sono stati iterati il secondo indice equivale al valore del contatore globale del candidato ed è stato prodotto un array contenente il numero di citazioni ricevute da ciascun DOI, il quale viene poi ordinato in ordine decrescente. In seguito lo script calcola l'h-index del candidato iterando sull'array ordinato fino a quando il valore dell'i-esima cella dell'array non contiene un valore minore al valore dell'iteratore: quando si verifica questa condizione significa che le successive iterazioni andrebbero a valutare articoli che hanno ricevuto meno citazioni rispetto al valore attuale dell'h-index, quindi il valore dell'h-index corrisponde al valore dell'iteratore stesso.

Quando il procedimento descritto è stato ripetuto su tutti i candidati presenti nel dizionario lo script ha prodotto i risultati dei tre indici per tutti i candidati, quindi ha terminato l'elaborazione. I dati ottenuti e le informazioni iniziali vengono riportati nel quarto ed ultimo file intermedio (data/CROSS_DATA.csv) che contiene, oltre ai dati del candidato, i valori dei tre indici appena calcolati, dei tre indici reali forniti nel tsv e delle soglie con cui gli indici devono essere confrontati.

4.6 Il calcolo dei risultati

Il quarto ed ultimo modulo si occupa di confrontare gli indici reali dell'ASN con quelli prodotti usando sorgenti open, produrre i dati in output in percentuale e generare i grafici relativi ai risultati. La descrizione della generazione dei grafici verrà affrontata nel capitolo seguente, in questo capitolo vengono descritte la prima e seconda parte del quarto modulo.
Dopo aver eseguito i primi tre moduli, lo script ha prodotto in output un file intermedio che racchiude al suo interno tutte le informazioni utili per verificare l'effettivo grado di convergenza tra i risultati ottenuti da ANVUR ed i risultati che si otterrebbero usando, invece che sorgenti closed, COCI e Cros-

sref: infatti nel file CROSS_DATA.csv sono presenti per ogni candidato i tre indici calcolati dal programma, i tre indici reali e le soglie che il candidato deve soddisfare per ottenere l'abilitazione. Per ottenere il dato atteso in output non resta quindi che confrontare i risultati ottenuti dai candidati a seconda della sorgente dati: per farlo si verifica su ciascun indice se i valori calcolati e reali superino o meno la soglia fissata, si confrontano i risultati tra loro e se c'è una corrispondenza è presente un match su quell'indice, altrimenti no. Dopo aver verificato i tre indici separatamente è necessario verificare che il candidato soddisfi le soglie, ovvero che due indici su tre superino le soglie fisssate. Questo viene fatto sia sui dati reali che sui dati calcolati e, come per gli indici, i risultati vengono confrontati per controllare se sia presente o meno un match.

Lo script genera un dizionario costruito con i dati presenti nel file CROSS_DATA.csv, dopodiché per ogni settore presente nel filtro del file di configurazione viene lanciata l'analisi dei candidati. Per ogni candidato appartenente al settore vengono confrontati i tre indici calcolati con le soglie, se due su tre superano le soglie il candidato viene validato, poi lo stesso procedimento viene applicato agli indici reali.
Successivamente vengono confrontati singolarmente gli indici ottenuti ed il risultato globale sulle due sorgenti dati: se gli indici o il risultato globale hanno ottenuto lo stesso risultato viene incrementato un contatore relativo ai singoli indici o all'overall. Il dizionario prodotto sul singolo settore viene poi inserito all'interno di un altro dizionario che raggruppa tutti i settori in modo tale da calcolare in un secondo momento le percentuali di matching.
Se nel file di configurazione non sono presenti settori da usare come filtro vengono presi in considerazione tutti i settori per cui è presente almeno un candidato.

Quando tutti i settori sono stati elaborati vengono calcolate per ogni settore le percentuali di matching sui singoli indici e sull'overall date dalla somma

dei candidati che hanno ottenuto un match su un indice o nel risultato globale e dalla somma totale dei candidati. Viene poi fatta anche una media dei risultati dei singoli settori per ottenere un risultato unico condiviso tra tutti i settori.
Infine i risultati ottenuti sui singoli settori ed il risultato della media vengono scritti in un file di testo (OUTPUT.txt) per essere facilmente accessibili. Di seguito viene riportato un esempio dell'output prodotto:

```
SUBJECT: 01-B1
LEVEL 1
OVERALL: 42.359767891682786 ARTICLES: 84.33268858800774
CITATIONS: 40.81237911025145 h-index: 40.42553191489362
LEVEL 2
OVERALL: 45.733333333333334 ARTICLES: 82.13333333333334
CITATIONS: 43.6 h-index: 48.13333333333333
```

4.7 La generazione dei grafici

Una volta completata l'elaborazione ed aver prodotto l'output il programma genera dei grafici per facilitare la lettura dei risultati ottenuti. Complessivamente viene prodotto un grafico per settore ed uno che racchiude i dati globali, in aggiunta vengono creati anche due grafici divisi per fascia dei macro-settori ordinati in ordine crescente contenenti la media degli overall dei sotto-settori. Quest'ultimo grafico è utile per individuare in maniera immediata quali siano i macro-settori che dispongono di più risorse liberamente accessibili sui database usati.
I grafici dei settori ed il grafico globale contengono i dati ottenuti sull'overall e sui tre indici divisi per fascia evidenziati con colori differenti.

Per ogni settore per cui è stato prodotto un output il programma lancia una funzione che si occupa di generare il grafico associato al settore: alla funzione vengono passati i valori percentuali dei matching sui tre indici e sull'overall

delle due fasce e, in base ai dati passati, usando la libreria matplotlib la funzione costruisce un istogramma che contiene sull'asse delle ascisse gli indici, per ogni indice una colonna per ogni fascia, e l'overall; mentre sull'asse delle ordinate vengono riportate le percentuali degli indici e dell'overall.
Lo stesso viene fatto per il grafico delle medie dei risultati su singoli settori; invece per i grafici dei macro-settori vengono prima calcolati i valori dell'overall come media degli overall dei settori appartenenti al macro-settore in questione, successivamente i risultati ottenuti vengono ordinati in ordine crescente e passati ad una funzione che genera un grafico per ciascuna fascia con i dati dei macro-settori. Nel capitolo relativo ai risultati sono presenti alcuni esempi di grafici trattati in questa sezione.

4.8 Specifiche Tecniche

L'esecuzione del progetto è stata lanciata unicamente su sistema operativo Microsoft Windows 10, ma il codice è stato scritto interamente in Python 3.7, pertanto il progetto è compatibile con tutti i sistemi operativi che supportino tale versione del linguaggio.
La macchina che è stata usata per lo sviluppo, i test e la produzione dei risultati riportati in questa tesi possiede le seguenti specifiche tecniche:

- Processore: Intel(R) Core(TM) i5-6600 CPU @ 3.30GHz
- Memoria installata (RAM): 8,00 GB
- Tipo sistema: Sistema operativo a 64 bit, processore basato su x64

Non sono quindi richiesti elevati requisiti minimi di sistema per poter utilizzare il programma. Un appunto però va dedicato allo spazio occupato dal dump di COCI che attualmente si attesta a 70GB che, sommato allo spazio occupato dal codice e dai file intermedi e di output, fa sì che per poter effettuare l'analisi siano necessari almeno 72 GB di spazio in memoria di massa.

Capitolo 5

Conclusioni

L'obiettivo del progetto era espandere e approfondire quanto riportato nell'articolo "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation" [2] mediante la creazione di un software in grado di replicare ed automatizzare l'analisi descritta nell'articolo. Il tema trattato riguarda la completezza e l'uso di risorse bibliografiche provenienti da sorgenti open e la loro effettiva diffusione rispetto alle risorse presenti su sorgenti closed.
Questo perché le risorse bibliografiche acquistano sempre più importanza in ambito accademico, anche come strumenti per verificare l'autorevolezza dei candidati per ruoli in ambito universitario, come nel caso dell'Abilitazione Scientifica Nazionale.
Pertanto non è accettabile che queste informazioni siano possedute unicamente da enti privati che amministrano un mercato su di esse, infatti già da alcuni anni è in atto un processo di liberalizzazione delle informazioni bibliografiche, sia relative ai metadati degli articoli che alle citazioni che gli articoli ricevono.

Per questo progetto sono state scelte come sorgenti open Crossref per quanto concerne i dati sugli articoli e COCI per i dati sulle citazioni e, per verificarne la completezza, sono stati calcolati gli indici dell'Abilitazione Scientifica

Nazionale tenutasi nel triennio 2016-2018 e confrontati con i risultati reali ottenuti dai candidati, che invece sono stati calcolati a partire dai dati presenti su Scopus e Web of Science, due sorgenti closed.

I risultati ottenuti forniscono due letture parallele: la prima relativamente al settore 01-B1 in quanto è l'unico settore ad essere trattato nell'articolo di riferimento e per questo fondamentale per verificare la correttezza del programma; la seconda lettura invece riguarda i risultati ottenuti su tutti gli altri settori esaminati e la loro predisposizione ed affinità con le sorgenti open. Come è stato dimostrato i risultati ottenuti sul settore 01-B1 sono in linea con quelli riportati nel paper, considerando opportunamente i threats to validity e le approssimazioni riportate nel capitolo sui risultati.
Dai risultati sui settori scientifico-disciplinari è emerso che, fatta eccezione per pochi settori come il settore 02-A1, in generale il grado di agreement tra i risultati prodotti basandosi su sorgenti open ed i risultati reali è molto basso.Il grafico relativo alle aree inoltre mostra che al momento non esistono aree nettamente più compatibili rispetto alle altre, per quanto le aree delle scienze matematiche e fisiche ottengano dei risultati migliori rispetto, per esempio, alle scienze chimiche. Anche per quanto concerne le fasce non si evidenzia un netto distacco tra la prima e la seconda, fatta eccezione per l'area delle scienze mediche, i cui settori ottengono risultati sensibilmente migliori sulla seconda fascia.
Le conclusioni quindi rispecchiano quelle trattate nell'articolo, confermando che le sorgenti open non sono ancora pronte a sostituire le sorgenti closed per calcolare gli indici dell'ASN e, più in generale, risultano ancora molto incomplete rispetto alle sorgenti closed Scopus e Web of Science.

5.1 Sviluppi futuri

Il primo aspetto che è possibile approfondire ulteriormente riguarda l'integrazione con altre sorgenti open, possibilmente relative a specifici settori, così come è stato fatto per l'articolo con DBLP. In questo modo si potrebbe ottenere una sinergia tra le diverse sorgenti open atta a migliorare i risultati in output in modo consistente. Secondariamente, inerentemente ad aspetti più legati all'implementazione, è possibile intervenire sulla velocità con la quale lo script produce i risultati implementando un meccanismo che parallelizzi ulteriormente l'esecuzione del primo modulo mediante la separazione del file in input in più chunks da distribuire a diversi calcolatori.

Bibliografia

[1] David SHOTTON. "Funders should mandate open citations." In: (2018). DOI: 10.1038/d41586-018-00104-7.

[2] Angelo Di Iorio, Silvio Peroni e Francesco Poggi. "Open data to evaluate academic researchers: an experiment with the Italian Scientific Habilitation". In: (feb. 2019).

[3] Ivan Heibi, Silvio Peroni e David Shotton. "Software review: COCI, the OpenCitations Index of Crossref open DOI-to-DOI citations". In: (2019). DOI: 10.1007/s11192-019-03217-6.

[4] Chifumi Nishioka e Michael Färber. "Evaluating the Availability of Open Citation Data". In: (2019).

[5] Silvio Peroni e David Shotton. "OpenCitations, an infrastructure organization for open scholarship". In: *Quantitative Science Studies* 1.1 (2020), pp. 428–444. DOI: 10.1162/qss_a_00023.

[6] *COCI*. URL: https://opencitations.net/index/coci.

[7] *Crossref*. URL: https://www.crossref.org/.

[8] *I4OC*. URL: https://i4oc.org/.

[9] *Legge 30 dicembre 2010, n. 240*. URL: https://www.camera.it/parlam/leggi/10240l.htm.

[10] *Scopus*. URL: https://www.elsevier.com/solutions/scopus.

[11] *Web of Science*. URL: https://clarivate.com/webofsciencegroup/solutions/web-of-science/.