Università degli Studi di Torino
Dipartimento di Filosofia e Scienze dell'Educazione

RELAZIONE FINALE

Sul lavoro svolto per la borsa di studio di ricerca per il progetto REPOSUM

Reponsabile del progetto:
Prof. Guido Bonino

Borsista:
Dott. Giulio Carducci

14 Giugno 2019

1 Introduzione

L'obiettivo del progetto REPOSUM è quello di rafforzare l'ecosistema piemontese delle tecnologie semantiche, attraverso la collaborazione tra il gruppo DR2 dell'Università di Torino e l'azienda torinese Synapta. Tale collaborazione ha lo scopo di fornire al gruppo DR2 gli strumenti necessari ad automatizzare, almeno parzialmente, l'estrazione e l'analisi delle informazioni contenute nei metadati di corpora di tesi di dottorato di filosofia. Partecipa al progetto anche il Prof. Daniele Radicioni del Dipartimento di informatica, i cui interessi di ricerca, focalizzati su Text Mining, Information Extraction ed elaborazione del linguaggio naturale, sono particolarmente rilevanti in questo contesto. Il progetto, dunque, ha anche la funzione di mediazione tra la componente umanistica e quella informatica, in accordo con gli interessi dello stesso gruppo DR2 verso l'istituzione di un nuovo centro di Digital Humanities presso l'Università di Torino.

Le attività di ricerca previste dal progetto rientrano in due principali gruppi:

1. **Raccolta e gestione dei dati**: catalogazione, annotazione e rappresentazione dei dati con tecniche statistiche, di machine learning e di analisi del linguaggio naturale.

2. **Ricostruzione delle carriere accademiche**: identificazione e studio di possibili approcci alla parziale automazione della ricostruzione delle carriere accademiche dei filosofi, e implementazione di queste soluzioni in flussi di elaborazione che permettano l'analisi e validazione dei risultati.

Il resto del documento è strutturato come segue: la sezione 2 descrive i metadati su cui è basato l'intero progetto, utilizzando statistiche e metriche numeriche; le sezioni 3 e 4 riportano i risultati ottenuti nelle due principali categorie di attività elencate sopra, i quali vengono infine discussi nella sezione 5.

2 Corpora di riferimento

Il progetto si basa su due corpora di metadati riguardanti tesi di dottorato inglesi e statunitensi discusse prevalentemente dal 1900 a oggi. I metadati più significativi tra tutti quelli disponibili sono i seguenti:

- Titolo
- Autore
- Argomento
- Università/Istituzione
- Anno di pubblicazione
- Abstract
- Relatore

Tabella 1: Statistiche sul numero di documenti per ogni corpus.

Corpus	Con Abstract	Senza Abstract	Totale
UK	201,718	273,665	475,383
US	20,944	9,399	30,343

Di particolare interesse per questo studio è l'abstract, in quanto permette di applicare diverse tecniche di elaborazione del linguaggio naturale (o Natural Language Processing - NLP). Nel corso della relazione verrà posta particolare enfasi su questo metadato. Proprio sulla base della presenza o meno dell'abstract, il corpus di tesi inglesi è ulteriormente suddiviso in due dataset, uno con abstract e uno senza. Come si vedrà più avanti, questa suddivisione è molto importante e verrà tenuta a mente nel corso di tutto il lavoro. La differenza principale tra i due corpora sta nel fatto che quello statunitense contiene solamente tesi filosofiche, mentre quello inglese contiene tesi appartenenti a svariati campi di studio, il che ha richiesto lo studio e l'implementazione di una soluzione per individuare le tesi di filosofia tra tutte quelle a disposizione (si veda la sezione 3.3).

Verranno adesso presentate alcune informazioni e metriche numeriche riguardanti i corpora. La tabella 1 riporta le statistiche sul numero di documenti per ogni corpus, mentre la figura 1 mostra la distribuzione dell'anno di pubblicazione a partire dal 1900. Infine, la tabella 2 riporta alcune statistiche sul contenuto testuale dei documenti.

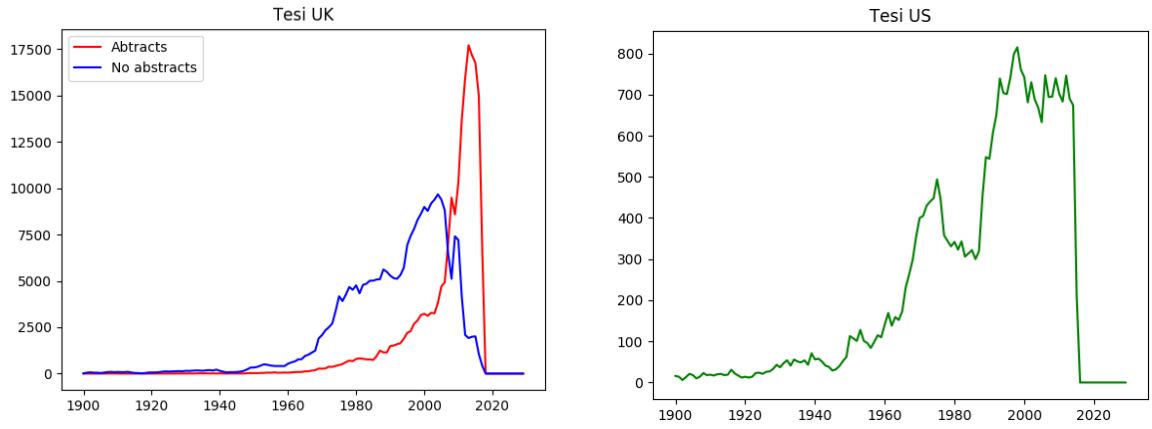

Figura 1: Distibuzione delle tesi per anno di pubblicazione.

3 Raccolta e gestione dei dati

In questa sezione verranno riportate tutte le analisi eseguite sui corpora per catalogazione, annotazione semantica e rappresentazione dei dati. Queste comprendono topic modeling, Entity Recognition e Named Entity Recognition, Classificazione, Rappresentazione semantica dei documenti. È importante però fare la distinzione tra i due corpus, inglese e statunitense. Poichè non è possibile sapere quali tesi inglesi siano effettivamente filoso-

Tabella 2: Statistiche sul contenuto testuale dei corpora. I valori riportati sono calcolati sui campi titolo, argomento e (se presente) abstract. Il preprocessing consiste in operazioni di pulizia del testo per rimuovere punteggiatura, rumore e parole poco significative (es. articoli, congiunzioni).

Metrica	**Tesi UK con abstract**	**Tesi UK senza abstract**	**Tesi US**
Parole totali	64,946,071	3,480,858	9,491,124
Parole totali dopo il preprocessing	37,875,285	5,454,769	40,424,273
Parole uniche	1,496,488	258,045	380,315
Parole uniche dopo il preprocessing	435,825	124,650	146,217
Media parole per documento	321.96	12.72	312.79
Media parole per documento dopo il preprocessing	187.76	9.31	179.77

fiche (a meno di analizzare a mano tutti i 475,000 documenti), alcune analisi riportare in questa sezione verranno condotte solamente sul corpus di tesi americane, avendo la certezza che siano tutte filosofiche. Più in dettaglio, le analisi riportate nelle sezioni 3.1, 3.2, 3.4, 3.5 sono condotte sul corpus statunitense, mentre quelle nella sezione 3.3 sono applicate solamente sul corpus inglese.

3.1 Topic Modeling

Il Topic Modeling è un tipo di analisi statistica per individuare argomenti astratti all'interno di una collezione di documenti, allo scopo di scoprire una possibile struttura semantica degli stessi. Nella pratica, dato un dataset di $\mathcal{N}$ documenti, viene studiata l'occorrenza delle parole nei testi per individuare le parole usate più spesso insieme, o in contesti simili, basandosi sull'intuizione che l'utilizzo di termini specifici determini il concetto astratto a cui tali termini si riferiscono. Se condotto su un gran numero di documenti, la statistica sulla co-occorrenza delle parole si rivela particolarmente efficace al fine di caratterizzare gli argomenti più ricorrenti.

L'algoritmo utilizzato per derivare i diversi topic dal corpus di tesi statunitensi è Latent Dirichlet Allocation (LDA), che nella configurazione più semplice richiede di specificare solamente il numero di argomenti, o topic, che si vogliono individuare nel testo. In generale, non esiste un numero ottimale, quindi è consigliato calcolare numeri diversi di topic con esecuzioni ripetute dell'algoritmo, e analizzare i risultati manualmente per verificare il numero più adatto al particolare dataset e agli interessi di ricerca. In questo senso, lo strumento **LDAvis** si rivela molto utile per la visualizzazione e l'analisi dei risultati; la figura 2 ne mostra un esempio, dove viene analizzato il contenuto del topic 1 su un totale di 20. La grandezza del cerchio sulla sinistra indica la frequenza del topic selezionato nel corpus, mentre quelle sulla destra sono le sue parole più significative e caratterizzanti. La barra blu indica la frequenza totale della parola tra tutti i topic,

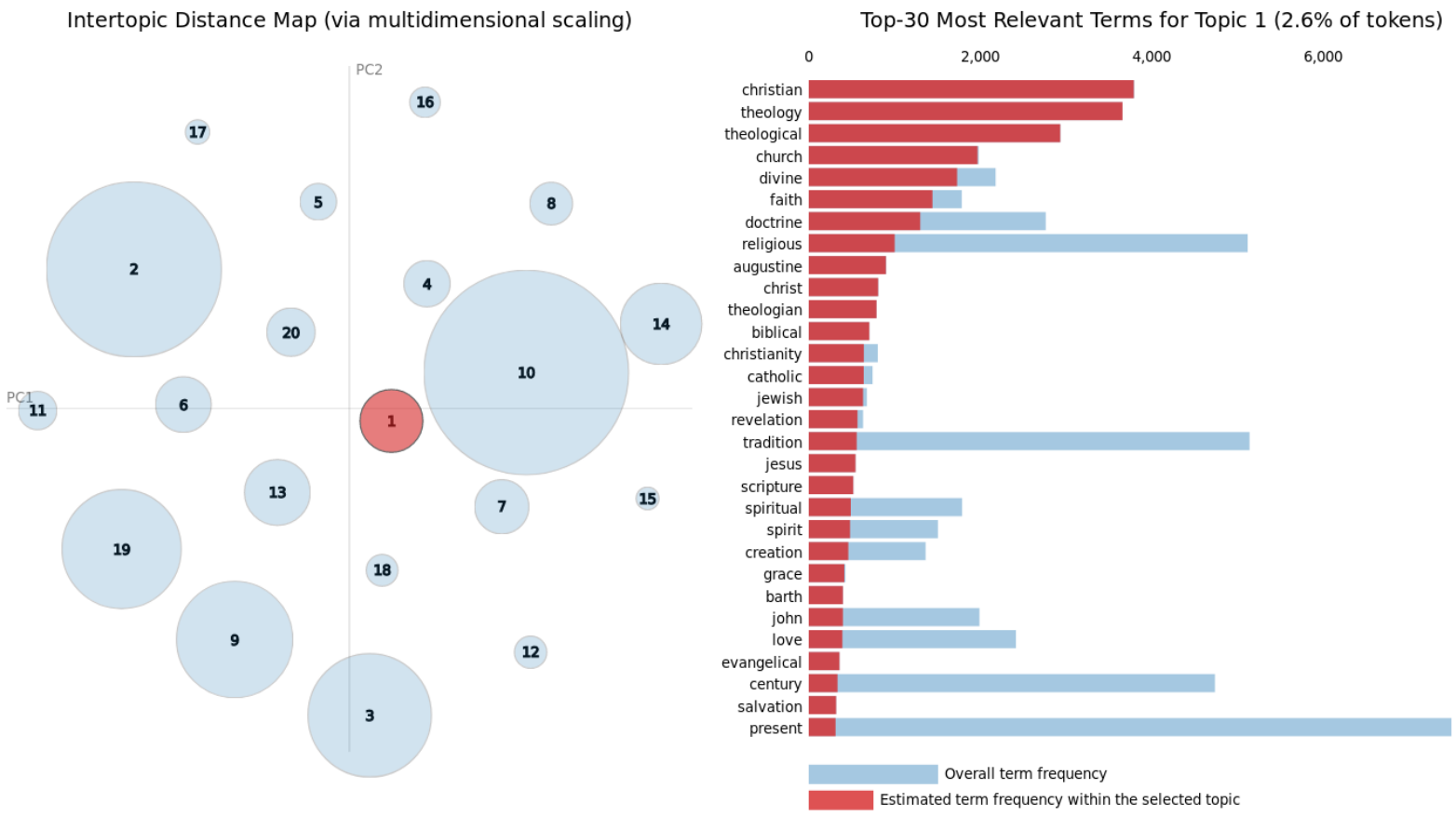

Figura 2: Visualizzazione del topic numero 1 in un topic model con 20 argomenti. Osservando le parole risulta evidente che tratti di religione e teologia.

mentre quella rossa, sovrapposta alla blu, indica la frequenza della parola solamente nel topic in oggetto. In un modello ottimale, la differenza tra le due barre dovrebbe essere minima o assente (come in 'christian' o 'theology'), stando a indicare che il termine è molto caratterizzante per l'argomento in questione e non per gli altri. Le parole 'religious', 'tradition' o 'present', ad esempio, sono poco caratterizzanti per il topic 1. Va specificato che l'algoritmo indentifica solamente un numero scelto da noi di argomenti ricorrenti nel testo che riceve in input, è poi compito nostro osservare le parole più caratterizzanti per un dato topic e assegnargli un'etichetta. Per esempio, i termini in figura si riferiscono chiaramente alla religione e alla teologia.

3.2 Entity Recognition

Entity Recognition, Named Entity Recognition e Entity Linking sono applicazioni del campo di studio dell'Information Extraction, cioè l'estrazione di informazioni utili da gruppi di documenti, nel nostro caso testuali. Queste tre applicazioni ruotano intorno alle **entità**, ovvero concetti reali o astratti menzionati nel testo, come ad esempio persone, luoghi, oppure organizzazioni. Il primo passo, *Entity Recognition* (ER), consiste nell'individuare all'interno del testo questi riferimenti. Dopodichè, con *Named Entity Re-*

cognition (NER) si può determinare il tipo di queste entità, ad esempio persone. Infine, l'*Entity Linking* consiste nel collegare tali entità ad una base di conoscenza esterna (come **Wikidata** o **DBpedia**) in modo da risolvere possibili ambiguità. Per esempio, l'entità di tipo persona *Hegel* può far riferimento al filosofo Georg Wilhelm Friedrich Hegel, alla pattinatrice Idora Hegel o al cantante Rob Hegel.

Solitamente, i modelli di ER/NER sono basati su modelli statistici e/o matematici pre-addestrati su grandi quantità di dati, già annotati con rispettive entità e loro tipi, e che in base alla posizione della parola e sua funzione grammaticale nella frase, sono in grado di individuare le entità con una buon precisione. Tra le varie librerie esistenti in letteratura si è scelto di usare **spaCy**, per via della sua accuratezza, rapidità, facilità di utilizzo e possibilità di visualizzare i risultati in modo chiaro. SpaCy dispone di modelli pre-addestrati in varie lingue in che sono adatti agli usi più generali. In figura 3 è mostrato il risultato del riconoscimento di entità sull'abstract di una delle tesi. Si può subito notare che la fase di ER è piuttosto accurata, mentre quella di NER produce decisamente più errori, in quanto più complessa.

Una volta individuate, le entità possono essere utilizzate semplicemente come rappresentazione semantica di un documento, oppure come features in altri contesti di machine learning, rendendo ER/NER un passo intermedio tra il documento originale e il l'obiettivo finale. Tuttavia, i modelli pre-addestrati non sono abbastanza accurati per i nostri scopi, e per un risultato ottimale è necessario ampliarli con informazioni aggiuntive in modo da aumentare la loro precisione. Non disponendo di queste informazioni (è possibile generarle manualmente ma l'operazione richiederebbe un tempo davvero notevole) si è deciso di non procedere in questa direzione.

3.2.1 Entity Linking con TellMeFirst

Un particolare tipo di strumento per l'entity linking consiste in **TellMeFirst**[1] (TMF), un software proprietario di Synapta che in questi mesi ho rinnovato e reimplementato da zero usando tecnologie più recenti. Tellmefirst è un classificatore di testo e motore di ricerca semantico che sfrutta il contenuto delle pagine di Wikipedia per individuare all'interno di una frase o breve periodo un numero arbitrario di concetti principali, come ad esempio persone, luoghi, opere d'arte o eventi, espressi come entità di Wikidata. Si tratta di un vero e proprio sistema di riconoscimento di entità, che comprende anche la fase di disambiguazione e linking, cambia solo il procedimento con cui è effettuato.

TMF non è altro che un motore di ricerca testuale, che cerca un breve testo all'interno di un database costruito a partire dal contenuto dell'intero corpus di Wikipedia, e restituisce i titoli degli articoli che sono più vicini semanticamente al testo dato come input. La particolare intuizione che sta dietro questo algoritmo consiste nel fatto che un un articolo non è definito dal contenuto della sua corrispondente pagina su Wikipedia, bensì dai paragrafi di tutte le altre pagine in cui compare un link a quell'articolo. Per esempio, il primo paragrafo della pagina Wikipedia di *Hegel*[2] contiene vari link, tra cui

[1] `https://tellmefirst.synapta.io/`

[2] `https://it.wikipedia.org/wiki/Georg_Wilhelm_Friedrich_Hegel`

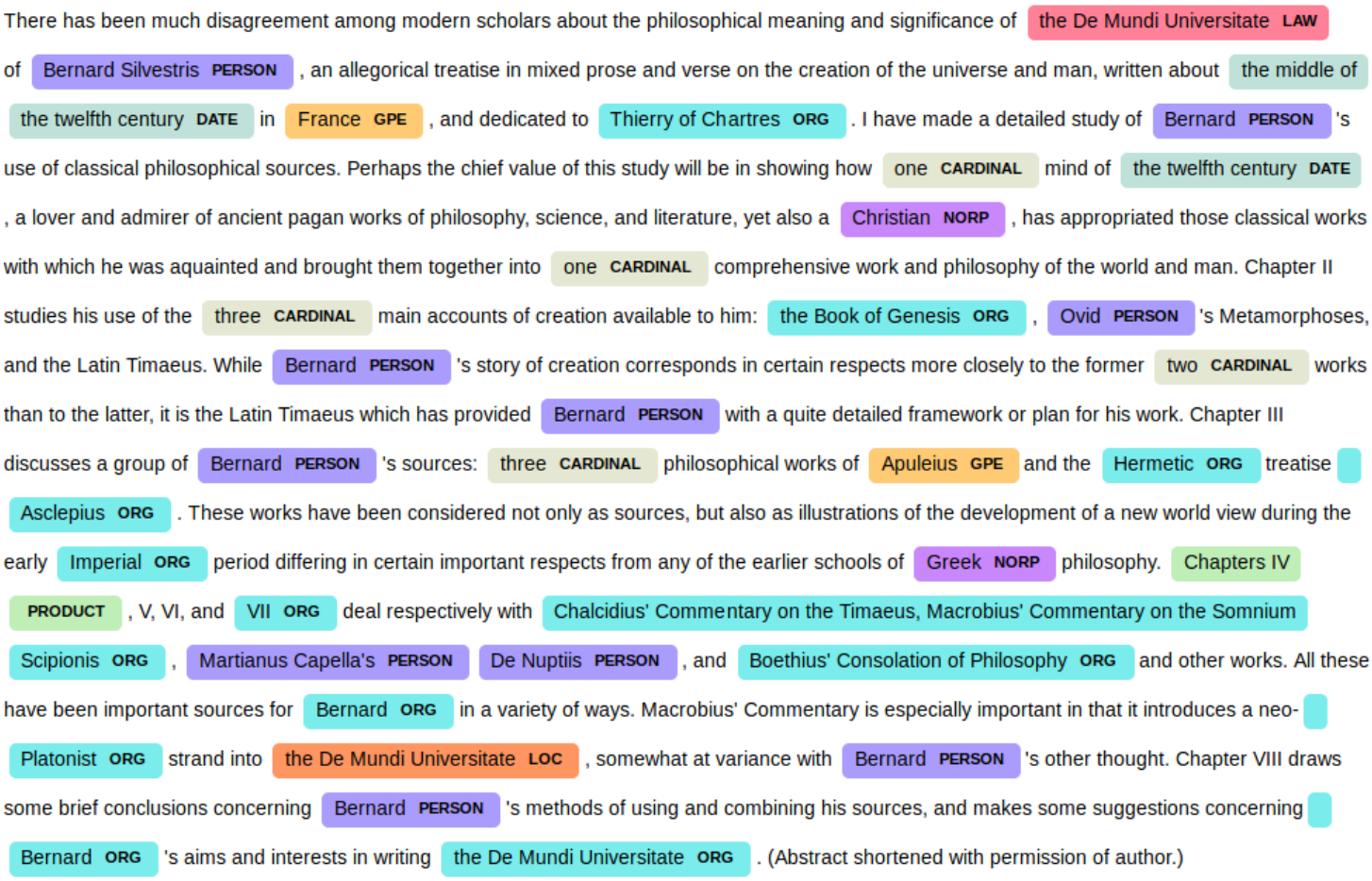

Figura 3: Risultato del riconoscimento di entità sull'abstract di una tesi. Le entità individuate vengono mostrate direttamente sul testo originale in modo chiaro e intuitivo. Il testo in maiuscolo e grassetto a fianco delle entità riconosciute ne indica il tipo, ad esempio persone (PERSON), luoghi (GPE), organizzazioni (ORG).

uno all'*idealismo tedesco*. La voce dell'idealismo tedesco nel motore di ricerca di TellMeFirst conterrà quindi, tra i vari paragrafi, quello della pagina di Hegel. In questo modo una determinata entità (in questo caso una corrente filosofica) viene descritta da tutti i paragrafi nella base di conoscenza di Wikipedia che ne parlano. Il risultato è che TMF può riuscire ad inferire delle entità anche se quetse non sono espressamente menzionate nel testo. Nel corso dell'attività di ricerca documentata in questa relazione TellMeFirst è stato usato più di una volta; nelle sezione che seguono sono riportate le applicazioni che hanno fornito i risultati più promettenti.

3.3 Classificazione

La parte più corposa del lavoro sulle tesi è senza dubbio quella della classificazione. Come abbiamo già detto, il corpus inglese non contiene solamente tesi filosofiche, come quello americano, ma di svariati campi di studio. Inoltre, i dati non sono annotati, il che vuol

dire che non sappiamo con certezza quali tesi tra le tante siano effettivamente filosofiche. Una necessità dei membri del gruppo DR2 era quindi sapere quali e quante fossero le tesi filosofiche in questo corpus, o almeno averne una stima. In questa sezione verrà illustrata brevemente la soluzione sviluppata per rispondere a questi interrogativi. Il lavoro di classificazione è stato presentato alla conferenza IRCDL 2019 che si è tenuta a Pisa a fine gennaio, con il titolo *Semantically Aware Text Categorisation for Metadata Annotation* ed è stato pubblicato nei proceedings della conferenza[3].

Per individuare le tesi filosofiche è stato deciso di addestrare un classificatore binario, che dica cioè per ogni tesi se sia filosofica o meno; un prerequisito indispensabile tuttavia consiste nell'avere un insieme di dati già classificati correttamente, in modo da imparare da essi ed estendere la classificazione a quelli rimanenti. Il primo passo è stato dunque quello di estrarre il maggior numero di documenti ritenuti possibilmente filosofici con un'euristica piuttosto semplice: selezionare tutti i record che contengono la stringa **philoph** nel campo **subject**; non è un approccio completamente privo di errori, ma si è dimostrato funzionale. In questo modo sono state individuate circa $3,500$ tesi, da usare come esempi positivi, a cui è stato aggiunto un campione casuale di 30000 tesi estratto da quelle rimanenti, da usare come esempi negativi. Da entrambi i gruppi sono stati poi rimossi 500 elementi, che sono stati analizzati a mano dagli esperti di dominio del gruppo DR2 per avere l'assoluta certezza che fossero filosofici o meno. I circa $33,000$ record rimanenti costituiscono il training set, mentre i $1,000$ annotati a mano costituiscono il test set. Una volta ottenuti i dati per l'addestramento e la verifica dei risultati, i record vanno trasformati in un formato facile da comprendere per un algoritmo, ovvero in *features*. A seconda del tipo di trasformazione scelto e del tipo di algoritmo da utilizzare per la classificazione, il procedimento è diverso.

3.3.1 Classificazione con Bag-of-Words

In questo caso, nonostante per alcune tesi fosse disponibile l'abstract, si è scelto di usare solamente il titolo come informazione su cui basare la classificazione, sia per uniformare tutti i record, sia per ottenere un lower bound delle prestazioni dell'algoritmo.

L'approccio **Bag-Of-Words**[4] (BOW) mappa un documento testuale su uno spazio vettoriale di dimensione arbitraria, in cui ogni dimensione di questo spazio corrisponde a una parola distinta. Il valore del vettore per la dimensione i è il numero di volte che la parola corripondente alla posizione i occorre nel testo. In questo modo un qualsiasi testo di lunghezza arbitraria può essere mappato su uno spazio vettoriale (= lista di numeri) di dimensione fissa.

Dopo aver trasformato tutti i record, il training set set è stato usato per addestrare un classificatore binario con l'algoritmo **Random Forest**, con cui poi sono stati classificati i record del test set e i risultati paragonati a quelli dell'annotazione manuale per verificare l'accuratezza del classificatore, il quale ha mostrato buona **precision** ma basso **recall**[5]. Questo significa che le tesi classificate come filosofiche sono per lo più corrette,

[3] `https://link.springer.com/book/10.1007/978-3-030-11226-4`
[4] `https://en.wikipedia.org/wiki/Bag-of-words_model`
[5] `https://en.wikipedia.org/wiki/Precision_and_recall`

ma rimangono molti record, anch'essi filosofici, che vengono tralasciati dal classificatore (e quindi classificati come non filosofici). Per ovviare a questo problema è stato introdotto un cosiddetto **modulo semantico**, il quale interroga la base di conoscenza **BabelNet** per individuare dei filosofi o concetti filosofici menzionati nel testo, classificando il record come filosofico se dovesse trovare un riscontro. Questo modulo viene eseguito solo per quei record classificati come non filosofici (per scelta implementativa), e la sua introduzione ha aumentato il recall del sistema del 22%.

3.3.2 Classificazione con TellMeFirst

Un secondo approccio per la classificazione consiste nell'utilizzo di **TellMeFirst** (vedi sezione 3.2.1), per associare a ogni record le entità wikidata più *vicine semanticamente*, per poi mappare ogni record su uno spazio vettoriale definito dalle entità, invece che dalle parole come nel caso precedente. Le tesi filosofiche dovrebbero dunque risiedere in una regione dello spazio vettoriale più o meno circoscritta, e lontana da altre tesi relative a campi di studio diversi. Per limitare la dimensione dello spazio delle entità Wikidata è stata introdotta un'ulteriore trasformazione, che è quella di riduzione della dimensionalità. Questa serve a mappare uno spazio di dimensione $\mathcal{N}$ su un nuovo spazio di dimensione $\mathcal{M}$, con $\mathcal{M} \leq \mathcal{N}$, senza perdita di informazioni (o con perdita minima). In questo modo si cerca di ovviare al problema detto *Curse of Dimensionality*[6]: il bisogno di dati per l'addestramento cresce in modo esponenziale rispetto alla dimensione dello spazio vettoriale; con uno spazio notevolmente ridotto, infatti, a parità di dati dovrebbero migliorare le prestazioni[7].

L'algoritmo di riduzione della dimensionalità scelto è **Singular Value Decomposition** (SVD), e la dimensione del nuovo spazio 100 (per l'approccio bag-of-words, la dimensione dello spazio vettoriale era circa $40,000$). L'algoritmo usato per addestrare il classificatore è **Bernoulli Naive Bayes**, il quale è particolarmente efficace per i testi brevi. Questo approccio si rivela particolamente promettente, in quanto da solo (senza il 'modulo semantico') riesce ad ottenere valori di precision e recall poco al di sotto di quelli della soluzione precedente.

La tabella 3 riassume brevemente i risultati delle diverse soluzioni adottate per la classificazione.

3.4 Rappresentazione semantica dei documenti

L'obiettivo degli studi riportati in questa sezione è quello di trovare una rappresentazione significativa delle tesi filosofiche che ne riesca a catturare le proprietà e caratteristiche, per poi sfruttarla per implementare uno strumento di inferenza, predizione, reasoning o altro.

[6] `https://en.wikipedia.org/wiki/Curse_of_dimensionality`

[7] Va specificato che questo procedimento vale anche per l'approccio Bag-Of-Words, ma in quel caso non c'è stato un miglioramento dei risultati in seguito dalla sua applicazione e quindi si è scelto di non usarlo.

Tabella 3: Riepilogo delle prestazioni delle 3 diverse soluzioni pensate per risolvere il problema della classificazione. La soluzione con Bag-Of-Words + modulo semantico è quella con le prestazioni migliori.

Implementazione	**Accuracy**	**Precision**	**Recall**
BOW	72.20%	**82.27%**	56.60%
BOW + modulo semantico	**79.20%**	79.44%	**78.80%**
TMF	72.60%	71.40%	75.40%

Dopo diversi incontri per definire il tipo di analisi da effettuare e la strategia con cui applicarla, la strada che è risultata più promettente è stata quella dell'estrazione di entità con TellMeFirst, in modo da individuare concetti e persone rilevanti di cui 'parlano' i documenti. Si è scelto poi di rappresentare queste informazioni in un grafo non orientato, contenente due tipi di nodi:

- **Nodi di tipo tesi**
- **Nodi di tipo entità**

Per ognuna delle $30,343$ tesi nel corpus statunitense viene quindi creato un nodo di tipo tesi, 10 nodi di tipo entità a partire dal titolo e 20 nodi di tipo entità a partire dall'abstract (se presente), estraendo rispettivamente le 10 e le 20 entità più significative da titolo e abstract con TMF. Ciascun nodo entità è connesso al rispettivo nodo tesi da un cammino che ha un peso proporzionale al punteggio di vicinanza restituito da TMF. L'intuizione dietro questo approccio sta nel fatto che se due o più tesi diverse condividono la stessa entità, allora nel grafo ci sarà un cammino di lunghezza 2 che collega tutti i nodi tesi, passando per quel nodo entità. Su un grafo così costruito è possibile sviluppare e applicare strumenti di inferenza semantica e reasoning, così come algoritmi e misure più proprie dei grafi come betweenness centrality, shortest paths o community detection.

Per aumentare ulteriormente il numero di possibili connessioni tra i nodi delle tesi, si è pensato di estendere il grafo in questo modo:

1. Recuperare la lista di tutti i filosofi presenti in Wikidata e aggiungerli al grafo come nodi ti tipo **entità wikidata**;
2. A partire da questa lista, per ogni filosofo esplorare la base di conoscenza di Wikidata traversando proprietà che possono essere di interesse per il contesto filosofico (es. influenced, influencedBy, educatedAt, author, doctoral_student, field_of_work, movement, member_of, instance_of, subclass_of e altre), in modo ricorsivo finchè almeno una di queste è disponibile;
3. Aggiungere le nuove entità al grafo man mano che vengono individuate da questa tecnica di esplorazione ricorsiva, nella forma di triple (soggetto, relazione, oggetto). Ogni entità è collegata a quella 'padre' da un cammino che ha come etichetta il nome della proprietà traversata per arrivare dal padre all'entità corrente.

Per esempio, la pagina Wikidata di Immanuel Kant[8] contiene la proprità *movement: German idealism*, quindi verrà aggiunta al grafo la tripla (Immanuel_Kant, movement, German_idealism). German idealism ha a sua volta due proprietà significative, *instance of: philosophical movement* e *facet of: idealism*. Dunque verranno aggiunte al grafo anche le triple (German_idealism, instance_of, philosophical_movement) e (German_idealism, facet_of, idealism). Si proseguirà poi esplorando le entità philosophical movement e idealism per cercare altre proprietà significative, e così via fino a trovare un'entità che non ha proprietà che ci interessano.

Questo procedimento è stato possibile perchè anche quelle estratte con TellMeFirst sono entità Wikidata, quindi dello stesso dominio, e ha permesso di aggiungere di $93,149$ triple al grafo delle tesi e entità TMF preesistente, il quale contava già circa $730,000$ triple, per un totale di circa $843,000$ triple. Trattandosi di entità strettamente legate al campo filosofico, è ragionevole supporre che un certo numero di esse si trovi già nel grafo perchè estratto con TMF, e che quindi i due insiemi di triple siano parzialmente sovrapposti; in questo modo va quindi a crescere il grado di connessione all'interno del grafo; riuscendo potenzialmente a raggiungere anche documenti che prima potevano essere isolati dagli altri.

3.5 Inferenza semantica sui documenti

L'analisi più immediata che è possibile fare sul grafo delle tesi è quella del cammino minimo tra due nodi. Date due tesi, quanto è lungo il percorso minimo che collega i rispettivi nodi, e quali nodi intermedi attraversa ?

Il calcolo del cammino minimo si basa sull'**agoritmo di Dijkstra**[9], e restituisce una serie di nodi intermedi tra sorgente e destinazione. Più precisamente, è possibile individuare la lista di tutti i cammini minimi, dal più corto al più lungo. Esaminando manualmente i risultati è possibile capire il tipo di relazione che lega due tesi, per esempio uno o più argomenti in comune.

In quest'ottica si colloca anche un successivo studio sulla somiglianza semantica dei documenti, che ha lo scopo di indivuare le tesi 'simili'. In questo contesto, la somiglianza si riferisce al contenuto semantico dei documenti, quindi si cercano tutte le tesi che trattano uno stesso argomento, o argomenti simili. Ad esempio, se due o più documenti affrontano il tema dell'idealismo tedesco, seppur trattandone autori o aspetti diversi, questi saranno simili per definizione. Il modo più efficace per implementare una soluzione del genere è usare un algoritmo di machine learning di tipo supervisionato; l'idea è quella di studiare la struttura del grafo (nodi e collegamenti) per capire quali sono le caratteristiche che definiscono la somiglianza tra due o più documenti. Questo vuol dire preparare una certa quantità di dati di cui si conosce già il valore di somiglianza, da cui l'algoritmo imparerà le caratteristiche più significative, per poi essere in grado di decidere, in base a queste stesse caratteristiche, se dei nuovi dati non analizzati in precedenza siano simili o meno.

Tutte le applicazioni più moderne di intelligenza artificiale usano un'architettura del genere, e se il problema di base è ben posto i risultati possono essere estremamente ef-

[8] `https://www.wikidata.org/wiki/Q9312`

[9] `https://it.wikipedia.org/wiki/Algoritmo_di_Dijkstra`

ficaci. Tuttavia, è necessario disporre di dati già annotati su cui basare lo studio. Il modo classico di ottenere questi dati è quello di analizzarne un campione significativo manualmente. Solitamente questa operazione viene fatta da un esperto di dominio, in modo da avere annotazioni quanto più possibile accurate, per questo motivo è stato chiesto ai membri del gruppo DR2 di eseguire questa operazione, che però è risultata troppo impegnativa per ottenere un numero di dati sufficiente. Si è quindi pensato di tentare un approccio alternativo: per identificare possibili gruppi di tesi simili è stato usato il topic modeling (vedi sezione 3.1), che altro non è che un tipo particolare di clustering, o raggruppamento, di documenti simili in un numero predefinito di insiemi. Per rendere più robusta questa soluzione, sono state eseguite più istanze di topic modeling, ognuna con un numero di topic diverso, e sono state idetificate le tesi che venivano raggruppate più spesso nello stesso insieme su tutte le istanze. Queste tesi avrebbero composto gli esempi positivi, mentre tutte le altre, campionate in modo casuale per estrarne un numero simile (o proporzionale) avrebbero composto quelli negativi. L'analisi manuale dei risultati ha dimostrato tuttavia che le tesi estratte in questo modo non erano poi così simili, probabilmente perchè i risultati di topic modeling sono piuttosto buoni su grandi numeri di documenti, ma molto variabili sul documento singolo.

A causa dei risultati iniziali poco promettenti questo percorso non è stato esplorato in modo più approfondito, ma resta un'alternativa potenzialmente valida all'identificazione di tesi o gruppi di tesi simili, nel caso in futuro dovesse essere disponibile un dataset annotato da usare per l'addestramento dell'algoritmo.

4 Ricostruzione delle carriere accademiche

La ricostruzione delle carriere accademiche dei filosofi è il secondo principale aspetto su cui il lavoro di questo progetto progetto di borsa di studio avrebbe dovuto concentrarsi. Nella pratica, già dai primi giorni è emersa la difficoltà di realizzare, anche solo parzialmente, un'automazione di questo compito. L'obiettivo iniziale del progetto consisteva nello sviluppo di una pipeline che permettesse la parziale automazione della ricostruzione delle carriere degli autori delle tesi presenti nel corpus statunitense, che comprendesse anche uno strumento per consentire l'analisi e validazione dei risultati.

I membri del gruppo DR2 avevano già ricostruito manualmente le carriere di qualche centinaio di filosofi, indicando, tra le varie informazioni, il rank dell'Università di partenza (quella in cui è stata discussa la tesi di dottorato), l'argomento della tesi, e, solo nel caso in cui la persona avesse trovato lavoro come professore, la fascia (professore, associato, ricercatore, ecc.), il dipartimento in cui insegna (o in cui ha insegnato), l'Università di arrivo e il suo rank, e infine, l'indice di successo accademico, un numero in una scala da 0 a 65 dove 65 è il successo massimo e 0 il minimo.

Queste carriere sono state ricostruite cercando manualmente i filosofi su Google con determinate query di ricerca, indentificandoli tra i possibili omonimi, e cercando le informazioni necessarie tra i risultati di ricerca più pertinenti, come siti web personali e accademici. Con questa strategia, la ricerca di un singolo filosofo richiede svariate query

sul motore di ricerca e un tempo non indifferente, senza avere alcuna certezza di poterlo individuare correttamente.

4.0.1 Ricerca dei filosofi su Bing

Un primo tenatativo di assistere i ricercatori in questo lavoro è stato quello di usare il motore di ricerca di **Bing**[10] per eseguire ricerche automatiche e analizzarne i risultati, con l'obiettivo di individuare qualche pattern sfruttabile per ripetere la ricerca su larga scala e indirizzare verso i risultati cercati (o almeno, cancellarne la maggior parte di quelli non pertinenti). Sono stati fatti vari tentativi con diverse keyword di ricerca:

- Nome del filosofo
- Nome del filosofo + "philosopher"
- Nome del filosofo + "professor"
- Nome del filosofo + Università in cui ha discusso la tesi
- Nome del filosofo + titolo della tesi

Nessuna di queste ricerche ha tuttavia restituito risultati soddisfacenti; si supponeva che le informazioni cercate dovessero trovarsi tra i primi 5 o 10 risultati di ricerca, ma solo in pochissimi casi (meno del 10%) era effettivamente così. Anche la ricerca totalmente manuale si è rivelata più complessa del previsto, in quanto nella maggior parte dei casi non si riuscivano a reperire informazioni certe sulla carriera del filosofo. In entrambi i casi, sono stati ricercati filosofi le cui tesi sono state pubblicate in anni diversi, dai meno recenti a più recenti, per diversificare la ricerca e così i risultati.

4.0.2 Studio della correlazione tra metadati e successo accademico

A partire dalle carriere già ricostruite dai membri del gruppo DR2, sono stati fatte delle analisi sulla correlazione tra i diversi metadati associati ad ogni filosofo e il suo indice di successo accademico, per identificare dei possibili trend a partire dai dati. La figura 4 mette in relazione l'indice di successo accademico con i rank delle Università in cui è stata discussa la tesi; per semplicità e comodità di visualizzazione, i diversi inidici di successo sono raggruppati in 3 insiemi. Sull'asse delle ordinate è indicata la percentuale di distribuzione delle tesi per rank delle Università dove sono state discusse. Osservando in particolare il primo e l'ultimo grafico è evidente come gli indici di successo più bassi siano correlati a Università di rank bassi, e viceversa. Come è intuitivamente evidente quindi, studiare e discutere la tesi in un'Università di rank alto aumenta le possibilità di carriera e successo accademico.

Le figure 5 e 6 mostrano invece la relazione tra il filosofo trattato nella tesi di dottorato e, rispettivamente, il rank dell'Università e l'indice di successo accademico. Anche in questo caso risulta evidente una correlazione significativa tra questi aspetti, infatti la

[10] Sia Google che Bing offrono il servizio di ricerca automatica a pagamento, è stato scelto Bing perchè Synapta disponeva di un credito da utilizzare su questa piattaforma.

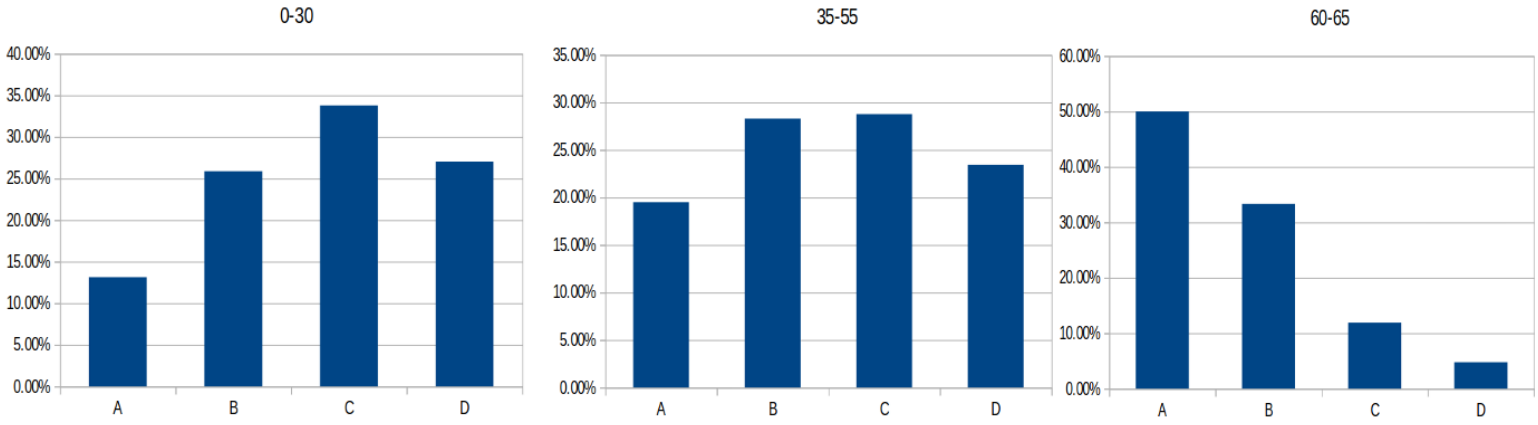

Figura 4: Distribuzione delle tesi per rank dell'Università in cui sono state discusse. Per evitare di visualizzare un grafico per ogni indice di successo diverso, 11 in totale, questi sono stati raggruppati in insiemi simili (associabili a: basso, medio, alto).

popolarità di alcuni filosofi (es. Dummett, Kripke, Lewis) è maggiore tra le Università di di rank alto, mentre quella di altri (es. Wittgenstein, Spinoza, Gadamer) è maggiore tra le Università di rank basso. Allo stesso modo, il successo accademico di dottorandi che hanno studiato i primi filosofi è generalmente più alto di quello di dottorandi che hanno studiato i secondi.

Le correlazioni che emergono da questi grafici sono piuttosto evidenti, tuttavia potrebbero essere influenzate dalla dimensione ridotta del campione o dal modo in cui è stata condotta l'analisi manuale delle carriere, che potrebbe aver preferito determinati argomenti e/o Università a discapito di altri.

Data la presenza di queste correlazioni, un semplice tentativo di testarne l'espressività è stato quello di addestrare un modello di machine learning che sfrutta le seguenti informazioni: filosofo trattato, Università e rank di partenza, Università e rank di arrivo per cercare di determinare il possibile indice di successo accademico. Sono stati testati vari modelli di classificazione non binaria (dato che gli indici di successo sono molteplici) o di regressione lineare (che determina un numero nell'intervallo continuo di possibili valori osservati nel training set), che però si sono rivelati piuttosto inefficaci, con valori di accuratezza al di sotto del 50%. Considerando invece come risultati corretti anche i due valori adiacenti a quello target (es. 40 e 50 oltre che 45), le prestazioni dei modelli aumentano leggermente, con valori di accuratezza, precision e recall che si aggirano sul 70%. Tuttavia, considerate le poche informazioni usate per addestrare i modelli e le correlazioni presenti tra i dati, risultati simili si possono facilmente inferire anche con un'analisi manuale.

La predizione dell'indice di successo accademico resta quindi un compito troppo complesso da automatizzare a partire dai pochi dati a disposizione; gli studi di correlazione effettuati però possono indirizzare i ricercatori verso la decisione, rimanendo comunque uno strumento utile per questo obiettivo.

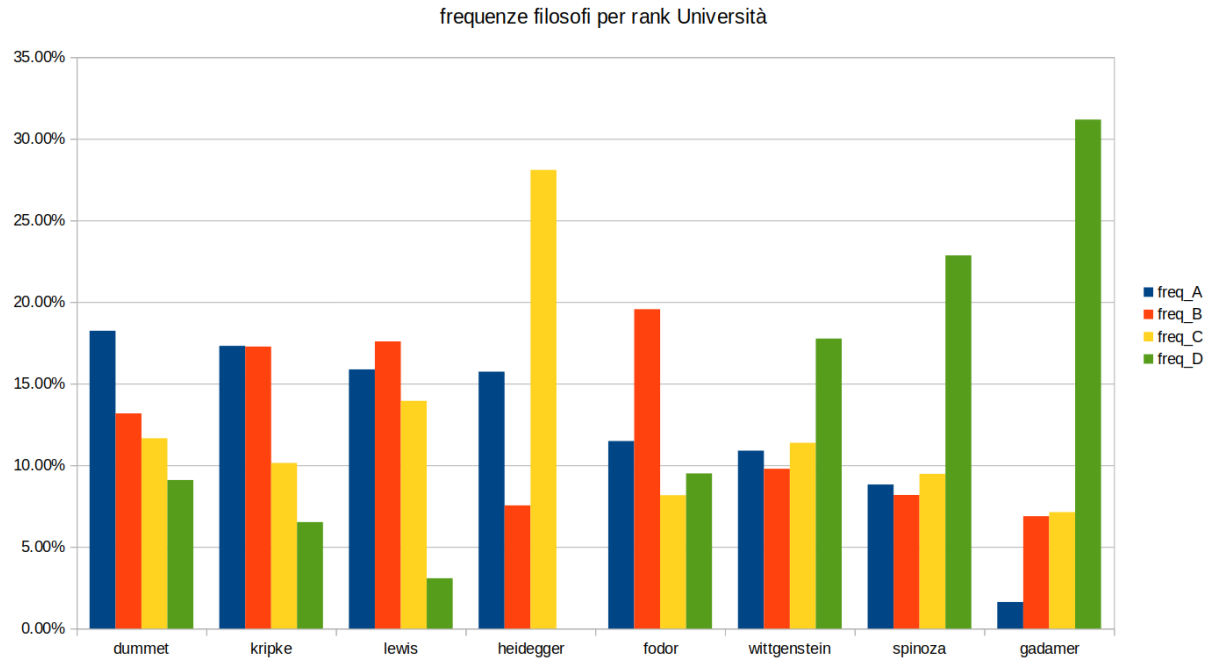

Figura 5: Distribuzione delle tesi per filosofo trattato e rank dell'Università. Sull'asse delle ascisse è riportato il filosofo discusso nella tesi e sull'asse delle ordinate la percentuale di tesi che trattano quel filosofo per un dato rank (A,B,C,D) dell'Università. I grafici dei 4 rank (colori diversi) sono indipendenti l'uno dallaltro, vengono riportati insieme solo per poterli confrontare.

5 Discussione dei risultati ottenuti

Nelle sezioni precedenti sono stati riportati i risultati principali ottenuti in quest'anno di lavoro. La classificazione delle tesi filosofiche è stata l'attività che ha richiesto più tempo e impegno, ma anche quella più fruttuosa in termini di obiettivi raggiunti. I valori di accuratezza ottenuti considerando solo le informazioni nel titolo delle tesi, quindi molto poche, sono notevoli. Questo lascia presupporre che un sistema più complesso, che considera una quantità maggiore di dati, potrebbe arrivare a risultati ancora più soddisfacenti.

Per quanto riguarda l'arricchimento semantico dei documenti, le tecniche di topic modeling e entity recognition/linking studiate e applicate in questo percorso hanno permesso di derivare informazioni aggiuntive dai singoli documenti, utili per migliorarne la descrizione e rappresentazione semantica, e che che sono state utilizzate per studiare e in parte implementare soluzioni di inferenza e ragionamento sui dati.

L'attività meno redditizia in termini di risultati è stata invece l'analisi delle carriere, che si è rivelata ben più complessa di quanto potesse sembrare all'inizio, e la cui automazione (o parziale automazione) a sostegno dell'attività di ricerca dei membri del gruppo DR2 non è stata possibile, a causa dei pochi dati a dispozione, della loro basso contenuto informativo, e delle difficoltà pratiche che stanno alla base dell'approccio. È stato comunque condotto uno studio sui dati disponibili e dimostrata la loro correlazione con

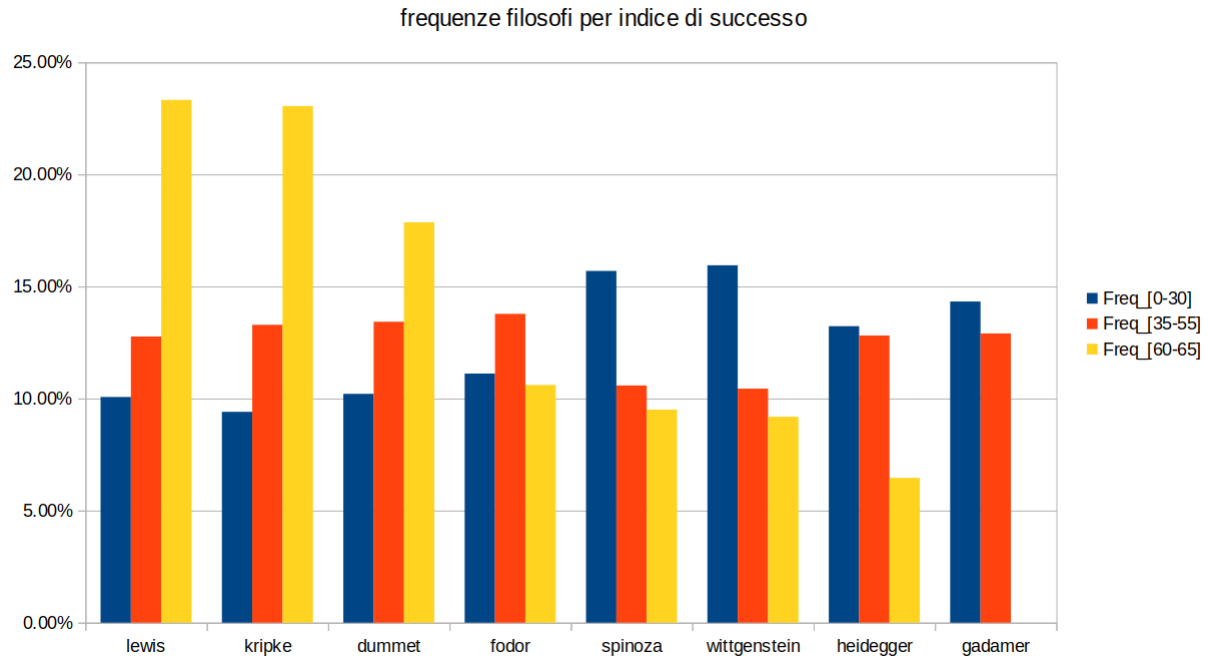

Figura 6: Distribuzione delle tesi per filosofo trattato e successo accademico. Come nel grafico precedente, ogni colore indica un grafico indipendente dagli altri.

il successo accademico dei filosofi.

6 Conclusione

L'attività di ricerca dell'ultimo anno per il progetto di ricerca REPOSUM ha portato diversi risultati significativi per quanto riguarda lo studio analitico e semantico dei metadati delle tesi, e fornito diversi spunti per proseguire il lavoro con tecniche o approcci diversi.

Ringrazio Guido Bonino, Paolo Tripodi, e tutti i membri del gruppo DR2 e non con cui ho avuto il piacere di collaborare in questi mesi; questa esperienza è risultata molto interessante e formativa, ha messo alla prova le competenze che già avevo nell'analisi di dati e sviluppo software orientati alla ricerca scientifica, e mi ha permesso di acquisire nuove conoscenze e abilità che mi mancavano o con cui ero poco familiare.