ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

DIPARTIMENTO DI SCIENZE STATISTICHE

"PAOLO FORTUNATI"

Corso di Laurea in Scienze Statistiche

TEXT MINING SU SORGENTI DI TESTO COMPLESSE: ANALISI DELLE CONSIDERAZIONI FINALI DEL GOVERNATORE SULLE RELAZIONI ANNUALI DELLA BANCA D'ITALIA DAL 2008 AL 2017

Tesi di Laurea in Utilizzo Statistico di Banche Dati Economiche Online

Presentata da:

Paolo Dalena

Matr. 0000800074

Relatore:

Prof. Ignazio Drudi

Correlatori:

Prof. Fabrizio Alboni

Dott. Corrado Lanera

APPELLO I

ANNO ACCADEMICO 2018/2019

A Mariella Mastrangelo e Adriano Dalena,

perché senza di loro non avrei né potuto né voluto essere a questo punto.

INDICE

1 Introduzione

Nell'era della rivoluzione digitale, della frenetica informatizzazione e della crescente e quasi fastidiosa presenza delle parole *Intelligenza* e *Artificiale* nella nostra quotidianità, l'applicazione di tecniche automatiche per accompagnare la comprensione e l'interpretazione di testi ufficiali risulta indiscutibilmente attuale. Infatti, tali metodi presentano potenzialità pressoché smisurate e l'impiego, benché non approfondito, di questi ultimi non può che costituire una buona pratica verosimilmente interessante.

Si è scelto di studiare le Considerazioni Finali del Governatore sulle Relazioni Annuali della Banca d'Italia riferite agli anni dal 2008 al 2017, analizzandone lo stile e il linguaggio, i temi trattati e le opinioni espresse. La descrizione dei dati e delle motivazioni alla base di questa scelta, assieme alla presentazione delle tecniche di *text mining*, vengono affrontate nei successivi paragrafi di questo capitolo. In quello seguente vengono illustrati gli obiettivi dello studio. Nel terzo vengono presentati nel dettaglio i metodi mediante cui l'analisi è stata svolta. Il quarto capitolo contiene l'esposizione dei risultati ottenuti correlata all'interpretazione di quest'ultimi. Infine, nella sezione relativa alle conclusioni si tirano le somme dello studio effettuato, fornendo spazio anche alla descrizione delle criticità.

Tutte le funzioni create e utilizzate, assieme ai dati analizzati, ai grafici realizzati, ai software adoperati, a tutto il codice necessario all'analisi organizzato per argomento, e tutto ciò che è stato utile per portare a termine questo lavoro è facilmente e gratuitamente reperibile attraverso il mio *GitHub,* nella cartella tesi[1].

1.1 Relazioni annuali della Banca d'Italia

La Banca d'Italia è la banca centrale della Repubblica Italiana, con sede centrale a Roma e sedi secondarie e succursali in tutta Italia. È un istituto di diritto pubblico, regolato da norme nazionali ed europee. È parte integrante dell'Eurosistema, composto dalle banche centrali nazionali dell'area dell'euro e dalla Banca centrale europea. L'Eurosistema e le banche centrali

[1] accessibile via *https://github.com/PaoloDalena/tesi*.

degli Stati membri dell'Unione europea che non hanno adottato l'euro compongono il Sistema europeo di banche centrali.

Persegue finalità d'interesse generale nel settore monetario e finanziario: il mantenimento della stabilità dei prezzi, la stabilità e l'efficienza del sistema finanziario e gli altri compiti ad essa affidati dall'ordinamento nazionale.

L'assetto funzionale e di governo della Banca riflette l'esigenza di tutelarne rigorosamente l'indipendenza da condizionamenti esterni, presupposto essenziale per svolgere con efficacia l'azione istituzionale. Le normative nazionali ed europee garantiscono l'autonomia necessaria a perseguire il mandato; a fronte di tale autonomia sono previsti stringenti doveri di trasparenza e pubblicità. L'Istituto rende conto del suo operato al Governo, al Parlamento e ai cittadini attraverso la diffusione di dati e notizie sull'attività istituzionale e sull'impiego delle risorse.[i]

Le pubblicazioni della Banca d'Italia riflettono le attività svolte dall'Istituto, sono a carattere economico-finanziario, storico e giuridico e sono tutte gratuite e disponibili online. Tra queste, quella che più fornisce una presentazione sintetica dell'intera situazione economica del Paese è senza dubbio la *Relazione annuale.* Quest'ultima viene pubblicata ogni anno alla fine del mese di maggio e contiene un'analisi approfondita dei principali sviluppi dell'economia italiana e internazionale nell'anno precedente e nei primi mesi di quello in corso ed è corredata di un'appendice statistica diffusa solo online. È inoltre oggetto, in una riunione pubblica non limitata ai Partecipanti, di Considerazioni da parte del Governatore della Banca d'Italia.

Tale pubblicazione può definirsi come una vera e propria consulenza analitica e informativa sullo stato dell'economia che la Banca d'Italia offre agli organi costituzionali in materia di politica economica e finanziaria. In virtù di ciò, è facile comprendere l'importanza della Relazione annuale e quanto il contenuto di quest'ultima costituisca un perfetto quadro sintetico da porre alla base di uno studio sulla situazione economica italiana e sulla sua evoluzione nel tempo.

1.2 Considerazioni finali del Governatore

Come già specificato, la Relazione annuale della Banca d'Italia è oggetto di discussione all'interno di una riunione pubblica. In occasione della diffusione di quest'ultima, dunque, il Governatore della Banca d'Italia presenta le cosiddette *Considerazioni finali*.

Il Governatore della Banca d'Italia ha il compito di rappresentare l'istituto bancario con terzi, di presiedere l'assemblea e di informare il governo italiano in materia di finanza estera o interna. Fino a prima dell'introduzione dell'Euro si occupava anche della politica monetaria nazionale. Tale funzione viene esercitata collegialmente insieme alle altre banche centrali dell'area Euro.

La nomina del Governatore è disposta con decreto del presidente della Repubblica, su proposta del Presidente del Consiglio dei ministri, previa deliberazione del Consiglio dei Ministri, sentito il parere del Consiglio Superiore della Banca d'Italia. Il procedimento si applica anche per la revoca del Governatore. La sua carica, che fino al 2005 non prevedeva limite di mandato, dura sei anni ed è rinnovabile una sola volta.[ii]

Soffermandoci sul periodo di nostro interesse, il Governatore in carica per il periodo 2005-2011 è stato Mario Draghi, che attualmente riveste il prestigioso ruolo di Presidente della Banca centrale europea, a cui è succeduto il 1° novembre 2011 Ignazio Visco, attualmente in carica.

Le Considerazioni finali riferite all'anno 2008, 2009 e 2010, dunque, sono state redatte da Draghi, mentre le successive sette in analisi (pubblicate fino a maggio 2018, quindi riferite al periodo 2011-2017) sono frutto del lavoro dell'attuale Governatore.

Tali pubblicazioni, disponibili gratuitamente online, forniscono un parere sintetico e più "umano" del quadro economico complessivo del Paese. Si tratta, infatti, di veri e propri commenti che il Governatore è chiamato a fare per tirare le somme dell'anno passato.

Per queste e altre ragioni pratiche, correlate alla necessità di utilizzare testi il più possibile privi di contenuti multimediali, si è scelto di analizzare le Considerazioni finali del Governatore della Banca d'Italia riferite ai dieci anni tra il 2008 e il 2017. Ciò fornirà un'idea chiara dell'evoluzione della situazione economica dell'Italia nel periodo di riferimento a partire da come quest'ultima viene presentata nelle Relazioni annuali.

1.3 Text mining su sorgenti di testo complesse

Il *data mining* (dall'inglese estrazione di dati) è l'insieme di tecniche e metodologie che hanno come obiettivo l'estrazione di informazioni utili da grandi quantità di dati, attraverso metodi automatici o semi-automatici (come ad esempio il *machine learning*).

Il *text mining*, anche detto *text data mining* o (in maniera per certi versi errata) *text analytics*, è una forma particolare di data mining nella quale i dati consistono in testi in lingua naturale, in altre parole documenti "destrutturati". Il text mining unisce la tecnologia della lingua con gli algoritmi del data mining. L'obiettivo è il medesimo: l'estrazione di informazione implicita contenuta in un insieme di documenti.

Tale disciplina, conosciuta anche con l'acronimo TM, dunque, si occupa della ricerca, dell'analisi e della classificazione tematica delle informazioni contenute nei documenti. A differenza della maggior parte dei dati con cui lavora la statistica, nei documenti testuali le informazioni sono presenti in forma di testo libero (frasi e parole) e soltanto in minima parte come testo strutturato (tabelle, grafici, ecc.). Va detto anche che gran parte delle comunicazioni, quindi dello scambio di informazioni, tra esseri umani avviene mediante documentazioni non strutturate (libri, giornali, conversazioni).

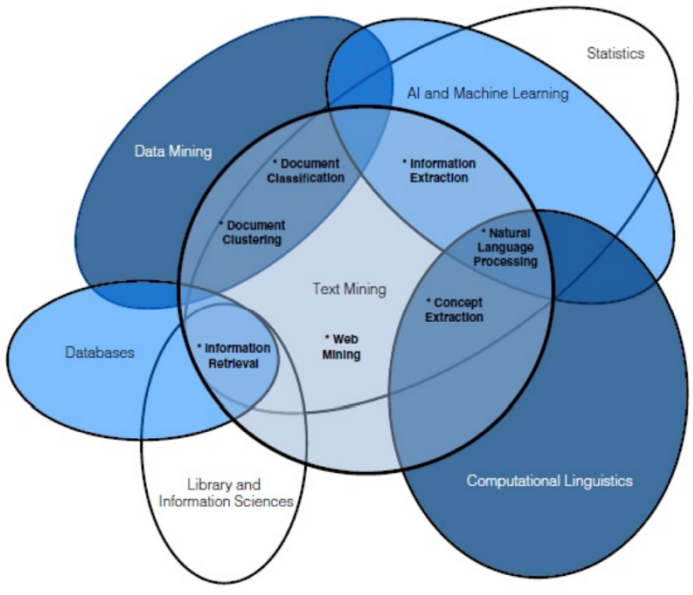

Figura 1: Interazione del Text Mining con gli altri campi di ricerca

Il TM è un campo multidisciplinare basato su diversi insiemi di tecniche, riunite sotto il nome di *information retrieval* (recupero delle informazioni), *data mining*, *machine learning* (apprendimento automatico), statistica e *computational linguistics* (linguistica computazionale). La Figura 1 mostra le interazioni tra il *text mining* e alcuni degli ambiti presentati.[iii]

Data la vastità di possibili applicazioni della tecnica in questione, fornire una descrizione esaustiva e universale di come avvenga il processo di TM risulta un compito particolarmente arduo. Tuttavia, è possibile riconoscere alcuni step:

- Collezione dei dati: prima fase che prevede la raccolta e la selezione di documenti che possono essere utili per l'analisi.
- Pre-processamento del testo: in cui si adatta il testo grezzo in testo analizzabile. In particolare, le operazioni di pre-elaborazione e pulizia sono eseguite per rilevare e rimuovere le anomalie, in modo da poter catturare la vera essenza del testo disponibile e anche semplicemente per ridurre la dimensione dei dati. In questa fase si applicano procedure di:
  - *Tokenizzazione*, che permette di rompere una sequenza di caratteri in unità (di solito parole o frasi) chiamati *token*.
  - Filtraggio, al fine di rimuovere parti del testo non necessarie.
  - Lemmatizzazione, che fa sì che le varie forme flesse di una parola vengano raggruppate in modo tale da essere analizzate come una singola entità.
  - Derivazione, ossia il processo mediante il quale si crea una forma (tema o parola) da una radice o da una parola preesistente.
- Applicazione delle tecniche di *text mining*: questa è la fase di maggior interesse, in cui i dati testuali (parole chiave, concetti, verbi, nomi, aggettivi, ecc.) sono estratti tramite tecniche basate su diversi algoritmi. Tra questi, i più diffusi sono:
  - Categorizzazione dei testi: rappresenta l'inizio del processo di analisi del testo tramite l'assegnazione di categorie predefinite ai token;
  - Estrazione dell'informazione: è una tecnica che estrae informazioni significative da una grande quantità di testo. Solitamente queste informazioni vengono prese da documenti non strutturati e/o semi-strutturati leggibili da una macchina e tramutate in informazioni strutturate.
  - Recupero delle informazioni: rappresenta l'insieme delle tecniche utilizzate per gestire la rappresentazione, la memorizzazione, l'organizzazione e l'accesso ad oggetti contenenti informazioni quali documenti, pagine web, cataloghi online e oggetti multimediali. È utilizzato anche dai motori di ricerca di *Google* e *Yahoo* per estrarre documenti da una ricerca sul web.

- *Clustering*: è un processo non supervisionato per classificare i documenti di testo in gruppi simili definiti cluster. In un cluster sono raggruppati insiemi di testo che si riferiscono ad uno stesso argomento o parole chiave identiche.
- Riepilogo di testo: è il problema di creare un riassunto breve, accurato e scorrevole di un documento di testo più lungo.
- Analisi del sentimento: nota anche come *opinion mining*, questo metodo viene utilizzato per estrarre informazioni soggettive dal contenuto. Proprio come suggerisce il termine, ha a che fare con l'emozione, il sentimento. Fondamentalmente, viene applicata per comprendere la risposta emotiva di un soggetto in un contesto.[iv]

I software e i linguaggi di programmazione che permettono l'attuazione delle tecniche di TM sono diversi e con differenti caratteristiche. In questa trattazione, come vedremo in dettaglio più avanti, l'analisi è stata condotta quasi interamente su R.

In conclusione, è necessario aggiungere che le Considerazioni finali del Governatore sono disponibili online[v] solo in formato PDF *(Portable Document Format)*. Questo formato è senza dubbio il più utilizzato per la diffusione di pubblicazioni disponibili in rete, in quanto permette di organizzare le parole in colonne, grafici e tabelle facilitando la lettura da parte degli esseri umani. Tuttavia, ciò che per gli umani semplifica le operazioni, comporta l'impossibilità di utilizzo diretto per le macchine.

Gli algoritmi di TM, infatti, sono direttamente applicabili solo se la sorgente dei dati è semplice. Ciò avviene quando, ad esempio, si hanno a disposizione file di testo (formato *.txt*), ossia documenti che contengono solo e soltanto lettere, numeri, segni di punteggiatura, spazi e altri simboli stampabili. Con i PDF, data la presenza di testo formattato in modi differenti, immagini, grafici, tabelle o qualunque altra tipologia di contenuto multimediale, la situazione si complica non poco (per questo si parla di sorgenti di testo *complesse*). È necessaria, dunque, una attenta estrazione e trasformazione dei dati, di cui si tratterà esaurientemente in seguito.

2 Obiettivi

2.1 Differenze lessicali

Nello studio delle Considerazioni finali del Governatore un primo approccio per comprendere l'evoluzione della situazione può partire dall'analisi delle differenze lessicali nel tempo. Ha senso, infatti, studiare come e se sono cambiate nei dieci anni di riferimento le modalità con cui ci si esprime e con cui vengono costruite le frasi.

Per osservare queste peculiarità è necessario studiare la distribuzione delle diverse parti del discorso (sostantivi, aggettivi, verbi, avverbi, pronomi, ecc.), ma sarebbe utile disporre anche di misure più sintetiche, come la lunghezza media delle frasi o il numero di termini non ripetuti all'interno del documento. Questi ultimi due dati, infatti, potrebbero rispettivamente fornirci informazioni rilevanti circa la complessità della trattazione, sotto il punto di vista della subordinazione delle proposizioni (a frasi mediamente più lunghe corrisponde una maggiore subordinazione) e del lessico (tante più sono le parole che non si ripetono all'interno di un testo, tanto più quest'ultimo sarà lessicalmente ricco).

2.2 Termini più utilizzati

Risulta interessante osservare la presenza di termini ricorrenti all'interno dei singoli documenti o degli stessi raggruppati per periodi, al fine di studiare eventuali correlazioni tra la situazione economica del momento in esame e le parole più presenti. Ad esempio, è naturale aspettarsi che negli anni successivi alla crisi del 2008-2009 il termine "crisi" sia fortemente presente nei documenti.

Un ulteriore obiettivo di interesse risiede nell'individuare le parole comuni tra i diversi documenti negli anni, così da osservare un trend oppure dei termini ricorrenti nel tempo.

2.3 Approccio positivo o negativo

Mediante l'*opinion mining*, è possibile studiare se le parole presenti nei documenti sono prevalentemente collegati ad emozioni positive o negative.

In questo modo si può comprendere come viene descritta la situazione generale dell'economia italiana e se i toni di questo approccio sono cambiati o meno nel tempo.

2.4 Argomenti ricorrenti

Grazie alle potenzialità del *topic modeling*, una tecnica che permette di creare modelli probabilistici che attraverso l'analisi delle parole caratterizzanti i testi individuano gli argomenti trattati in un documento, sarà possibile osservare eventuali temi ricorrenti negli anni.

Di conseguenza, si potrà esaminare come la discussione presente nelle Relazioni Annuali sia cambiata nel tempo e in quale verso, ma anche quali siano gli oggetti di analisi più importanti, in quanto questi ultimi coincideranno con quelli di cui è necessario parlare più spesso, ovvero i più presenti.

3 Metodi

3.1 Esportazione del testo

Come già accennato, i dati oggetto di questa analisi sono reperibili solo in forma complessa, ovvero in formato PDF. È stato dunque necessario, in primo luogo, ricercare il software che offrisse le performance migliori circa l'esportazione accurata del testo e, in secondo luogo, automatizzarne il processo.

Al fine di comprendere le particolari funzionalità dei diversi metodi per l'importazione di testo da PDF così da poter comprendere quale sia il più accurato e, in particolare, il più adatto alle nostre esigenze, è stata condotta una revisione della letteratura sull'argomento.

Una volta trovato il software, se ne è automatizzato il funzionamento mediante la creazione di un pacchetto *R* contenente le funzioni utili alla causa.

3.1.1 Revisione della letteratura

La revisione della letteratura, condotta sugli archivi di *arXiv*[2] e *ACM (Association of Computer Machinery) Digital Lybrary*[3]*,* ha permesso di studiare una pubblicazione[vi] che offre una perfetta sintesi complessiva dei software esistenti fino ad allora (giugno 2017) suddivisi in base alle loro diverse caratteristiche e capacità.

Sono stati valutati 13 diversi programmi, scelti in modo da escludere quelli con funzionamento simile, più uno (*PdfAct*[vii]*,* inizialmente chiamato *Icecite*) presentato nell'articolo. Di questi sono state analizzate le caratteristiche in termini di identificazione dei confini dei paragrafi, del corretto ordine di lettura e dei ruoli semantici dei termini; di capacità di traduzione delle legature, dei segni diacritici e delle parole con il trattino; di possibili formati di output.

Una volta comprese le capacità dei singoli software, l'analisi nella pubblicazione ha spostato la sua attenzione sulle valutazioni delle prestazioni. Sono stati, dunque, presi in esame i possibili

[2] Archivio di articoli scientifici in fisica, matematica, informatica, finanza quantitativa e biologia, accessibile via *https://arxiv.org/*

[3] Collezione di tutte le pubblicazioni della ACL, accessibile via *https://dl.acm.org/*

errori di riconoscimento delle parole o dei paragrafi sugli output di un archivio costituito da circa dodicimila articoli scientifici presi casualmente da *arXiv*, in formato PDF.

Dopo un'attenta valutazione delle caratteristiche e delle performance dei diversi software, si è concluso che il più adatto al nostro scopo è proprio *PdfAct*, ovvero quello proposto dagli autori dell'articolo. Quest'ultimo è una libreria *Java* capace di riconoscere e separare i *LTBs* (*Logical Text Blocks,* blocchi logici del testo) secondo un approccio *rule-based* che analizza le distanze, le posizioni e il font dei caratteri.

3.1.2 Automatizzazione dell'estrazione del testo su R

A seguito della revisione della letteratura, sono state create delle funzioni, raggruppate in un più ampio pacchetto R chiamato *corecage*[4], che permettono di eseguire automaticamente il processo di estrazione dei dati interamente attraverso R. Tali funzioni sono in grado di creare dei semplici file in formato *.txt* contenenti il testo accuratamente estratto dalle Considerazioni finali del Governatore (in formato PDF), fornendo semplicemente il *path* della cartella dove sono presenti i PDF e quello della cartella dove si vogliono organizzare i nuovi file di testo.

Nonostante il software utilizzato sia una libreria *Java* e dunque non si presti ad essere utilizzato direttamente in un ambiente R, è possibile eseguire tutti i comandi comodamente all'interno di quest'ultimo[5].

Ciò ha permesso di avere a disposizione dei file semplici contenenti solo e soltanto i termini presenti nel corpo e nelle intestazioni dei paragrafi delle Considerazioni finali del Governatore. Questi 10 file (uno per ogni anno di interesse) sono stati, dunque, utilizzati per la creazione del corpus di documenti su cui sono state operate le successive analisi, mediante la libreria *tm*[viii] per il *text mining* su *R*.

[4] Disponibile attraverso il mio account *GitHub*, al link *https://github.com/PaoloDalena/corecage*

[5] In quanto una stringa viene eseguita direttamente come comando di Sistema grazie alla funzione base *system*. Per ulteriori chiarimenti sulle funzioni utilizzate, si rimanda all'*help* di R fornito nelle documentazioni del pacchetto *corecage*.

3.2 Lemmatizzazione, categorizzazione e pulizia

Per eseguire accuratamente il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma, ci si è avvalsi di un *tool* esterno chiamato *TreeTagger*. Quest'ultimo è uno strumento che permette di annotare le parole contenute in testi di svariate lingue con la categoria grammaticale ed il lemma appropriati ed è stato sviluppato da Helmut Schmid nell'ambito del *TC project* presso *l'Institute for Computational Linguistics* della *University of Stuttgart.*[ix]

Per poter adoperare questo strumento in un ambiente R è stato utilizzata una libreria chiamata *koRpus*, disponibile sul *CRAN*, che offre diversi servizi utili alla *text analysis,* tra cui un *wrapper* per *TreeTagger*, e la libreria di supporto per la lingua italiana (*koRpus.lang.it*).[x]

A partire dall'output fornito da questo *tool* è dunque possibile ridurre le parole provenienti dallo stesso lemma ad un'unica entità e organizzarle in base alla loro categoria grammaticale. Saremo in grado, quindi, di osservare come si distribuiscono le diverse parti del discorso nei nostri documenti.

Una volta ottenuti i lemmi organizzati per categoria grammaticale, sono state rimosse dai dati le cosiddette *stopwords* (parole d'arresto). Quest'ultime sono le parole più comuni di una lingua (come gli articoli o le congiunzioni), che sono di norma maggiormente presenti in un testo e potrebbero creare problemi nelle analisi. Per ovvi motivi, inoltre, sono stati eliminati i termini che ricorrono in questi particolari testi in analisi e dunque, nello specifico, parole come "considerazione", "finale", "governatore", "banca", "Italia", etc. Per ottenere, infine, una lista esauriente delle parole d'arresto della lingua italiana è stata utilizzata la funzione *stopwords* implementata nel pacchetto *tm*, che fornisce un elenco di termini da rimuovere per diverse lingue, tra cui l'italiano.[xi]

3.3 Analisi delle differenze lessicali

Mediante la funzione *describe* della libreria *koRpus* è possibile osservare diverse statistiche descrittive sui dati risultanti dall'applicazione della lemmatizzazione con *TreeTagger*. Tra queste troviamo diversi indici che descrivono il numero di caratteri all'interno dei documenti (tutti i caratteri, senza spazi, solo lettere, etc), il numero di parole e di frasi e la loro lunghezza media.[xii]

Inoltre, mediante un'apposita funzione[xiii] dello stesso pacchetto, è stato possibile calcolare i diversi indici *MTLD* (*Measure of Textual Lexical Diversity*, letteralmente misura della diversità lessicale testuale). Questi ultimi sono dei chiari indicatori della ricchezza lessicale del testo, in quanto sono calcolati a partire dal rapporto tra il numero di termini unici presenti in un testo e il numero totale di parole al suo interno[6]. Infatti, l'aumentare del numero di parole che non si ripetono in un documento corrisponde all'utilizzo di un vocabolario più ampio, sinonimo di una maggiore ricchezza lessicale.

3.4 Creazione delle nuvole di parole

I dati lemmatizzati, categorizzati e puliti sono stati in seguito organizzati in base alla frequenza con cui compaiono all'interno dei diversi documenti, così da poter costruire delle nuvole di parole mediante l'utilizzo dei pacchetti R *wordcloud*[xiv] e *wordcloud2*[xv].

All'interno delle nuvole di parole, dunque, i termini appariranno tanto più grandi quanto più di frequente sono presenti nelle Considerazioni Finali del Governatore. Alla base di questa ponderazione c'è la semplice idea che in un documento tanto più delle parole sono frequenti, tanto più tendono ad essere importanti e significative per il contenuto.

La nuvola di parole è un tipo di grafico comunemente utilizzato per visualizzare in modo sintetico il contenuto di un discorso o di un set di documenti, e può fornire spunti per comprendere e interpretare i contenuti dei testi. Da un punto di vista statistico, è equivalente a un grafico a barre di frequenze univariate. Rispetto a questo tipo di grafico la *wordcloud* rende sicuramente più difficile quantificare la frequenza relativa delle parole, tuttavia ha il vantaggio di permettere una visualizzazione che consente di cogliere in modo immediato la rilevanza delle stesse.

---

[6] Tale descrizione di come avviene il calcolo è decisamente semplicistica. Per ulteriori chiarimenti, si rimanda a McCarthy, Philip M., e Scott Jarvis. «MTLD, Vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment». *Behavior Research Methods* 42, n. 2 (1 maggio 2010): 381–92.

Inoltre, per confrontare i documenti all'interno del corpus sono state utilizzate la *comparison cloud* e la *commonalty cloud*. Nella prima, la dimensione delle parole è definita in base ai differenti tassi di presenza delle parole all'interno di ogni documento: la *comparison cloud* evidenzia, quindi, le differenze. La *commonality cloud*, invece, evidenzia parole comuni a tutti i documenti. In questo secondo caso, la dimensione della parola è funzione della sua frequenza minima tra i documenti. Quindi se una parola manca da un qualsiasi documento ha dimensione nulla (cioè non viene visualizzata).[xvi]

Le nuvole di parole dei confronti sono state costruite mediante l'utilizzo delle funzioni specifiche *comparison.cloud* e *commonality.cloud* della libreria *wordcloud*, mentre quelle complessive sono state create grazie a funzioni contenute nel pacchetto *wordcloud2*, che offre una maggiore libertà in termini di stile grafico.

Per ragioni pratiche, correlate alla difficoltà di interpretazione di dieci nuvole diverse, si è scelto di riunire i documenti in tre periodi. Il primo comprende gli anni dal 2008 al 2010, il secondo quelli dal 2011 al 2014, mentre il terzo quelli dal 2015 al 2017. Anche i confronti sono stati costruiti a partire dai dati così suddivisi.

Infine, poiché gli elementi sono organizzati in categorie grammaticali, è stato possibile costruire le nuvole di parole limitandosi all'analisi di una sola di queste tipologie alla volta. Data la scarsa informazione contenuta in parti del discorso secondarie, si è più volte scelto di includere nelle *wordcloud* lemmatizzate i soli sostantivi.

3.5 Sentiment analysis

Per collegare alle parole contenute nei dati la loro polarità, positiva o negativa, è necessario utilizzare un dizionario di *opinion word* (detto *lexicon*), ovvero un vero e proprio elenco di aggettivi, nomi, verbi e avverbi a cui vengono associate le emozioni, e dunque le opinioni, che rispecchiano. Ad esempio, alla parola "sole" sarà associata l'emozione della gioia e un'opinione positiva, mentre alla parola "abbandono" un sentimento di tristezza e una polarità negativa.

Tuttavia, trovare un *lexicon* costituito adeguatamente per la lingua italiana è un'ardua impresa. È senza dubbio più semplice reperirne uno in lingua inglese. Per le successive analisi, dunque, le polarità dei termini sono state estratte a partire dalle traduzioni in lingua italiana di due dizionari differenti.

Il primo è il *subjectivity lexicon* di *Janyce Wiebe*[xvii], docente presso il *Department of Computer Science and Intelligent Systems Program* della *University of Pittsburgh,* ed è stato adoperato mediante alcuni adattamenti delle funzioni contenute nella libreria *sentiment.* Il secondo è il *NRC emotion lexicon*[xviii], fornito da *Saif M. Mohammad, Senior Research Scientist* presso il *National Research Council Canada.* Quest'ultimo è accessibile, in lingua inglese, mediante la libreria *syuzhet*[xix], dunque è stato utilizzato attraverso adattamenti delle funzioni contenute in questo pacchetto.

3.6 Topic modeling

L'applicazione di algoritmi di *topic modeling* permette di identificare gli argomenti di ogni singola sezione che va a costituire un intero documento. Fino ad ora nell'analisi abbiamo considerato i dati come un corpus costituito da dieci testi, ognuno contenenti i termini lemmatizzati delle Considerazioni finali del Governatore. Dato l'obiettivo di ricercare gli argomenti ricorrenti all'interno delle singole pubblicazioni, è stato necessario riorganizzare i testi.

Sono stati creati dieci corpora diversi costituiti da ogni singola pubblicazione. Dunque, se fino ad ora si è operato con un corpus di dieci documenti, d'ora in avanti si analizzeranno dieci corpora formati da un documento ciascuno.

Per riconoscere gli argomenti caratterizzanti le sezioni di ogni pubblicazione è stato utilizzato, mediante la funzione *LDA* presente nella libreria *topicmodels*[xx], un particolare modello probabilistico del testo chiamato *Latent Dirichlet Allocation* (allocazione latente di Dirichlet, *LDA*). Si tratta di una tecnica molto elaborata che si presta a innumerevoli applicazioni. Per citarne qualcuna, è il metodo con cui vengono ordinati per pertinenza i risultati di una ricerca su *Google,* ma anche quello utilizzato da *Amazon* per il *clustering* di clienti in base agli acquisti effettuati.[xxi]

L'*output* fornito da questa funzione è carico di informazioni, tra cui molte di complicata interpretazione. Per semplificarne la comprensione e l'utilizzo in linea con il nostro obiettivo, è stato adoperato in modo tale da ottenere i quattro lemmi che con più probabilità fanno parte di quattro *topics* che costituiscono il testo. La scelta del numero di parti in cui suddividere i documenti e di quanti termini studiare è arbitraria.

Queste parole, in quanto sono quelle che con maggiore probabilità si collocano congiuntamente di una sezione della pubblicazione piuttosto che in un'altra, daranno una precisa indicazione dell'argomento a cui fanno riferimento. Sarà, dunque, possibile ricostruire i temi trattati all'interno dei diversi testi e, di conseguenza, osservare se ce ne siano alcuni che vengono più volte ripresi negli anni.

4 Risultati e discussione

4.1 Distribuzione delle parti del discorso

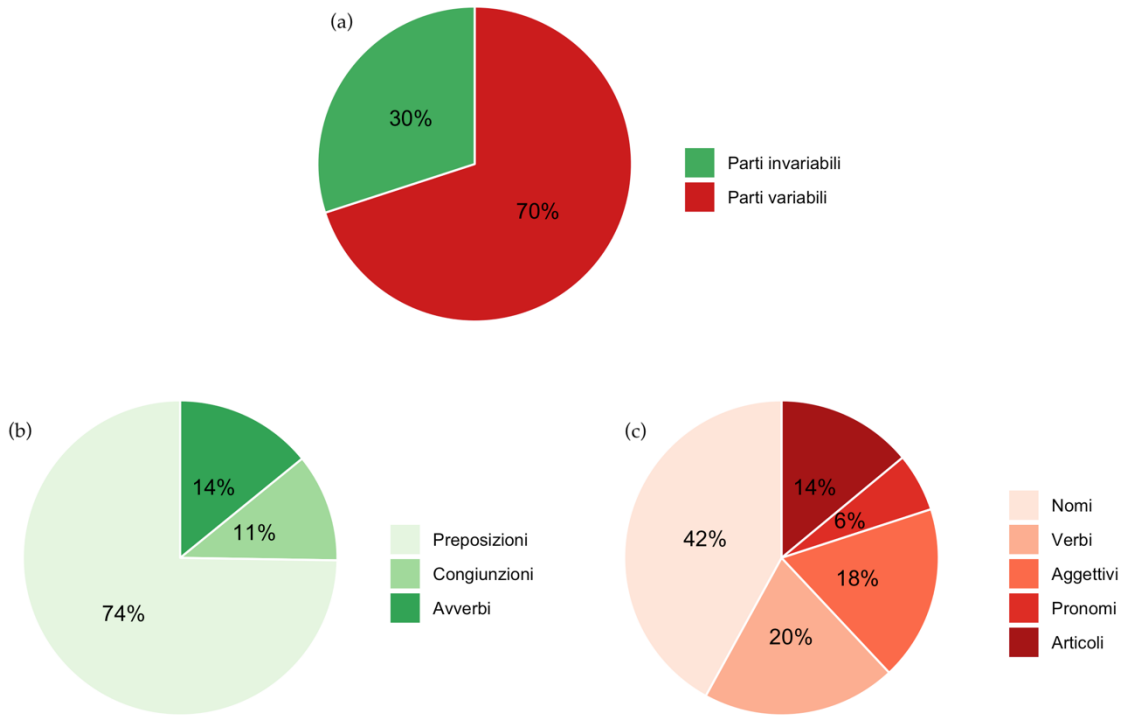

Figura 2: Distribuzione media delle parti del discorso (a), in particolare invariabili (b) e variabili (c)

Dato l'elevato numero di termini presenti all'interno dei documenti, le diverse distribuzioni per anno delle parti del discorso sono pressoché identiche. Dunque, analizzarne le variazioni nel periodo di tempo considerato ha poco senso. Risulta interessante, invece, osservare la distribuzione media delle nove parti del discorso della lingua italiana. Quest'ultima è rappresentata in Figura 2.

Il grafico *(a)* mostra come il 30% dei dati sia costituito da parti invariabili del discorso, ovvero preposizioni, avverbi, congiunzioni e interiezioni. In particolare, come si evince chiaramente dal grafico *(b)* che mostra la distribuzione limitatamente a questa sezione dei dati, la maggior parte delle parti invariabili del discorso è costituita da preposizioni (74%), seguita da avverbi (14%) e congiunzioni (11%). È importante rilevare la totale assenza di interiezioni nei testi osservati. Le ragioni di questo risultato risiedono sicuramente nel fatto che le esclamazioni esprimono un particolare atteggiamento emotivo dell'autore, che non può ritrovarsi in pubblicazioni ufficiali come quelle in analisi, a cui si addice un registro formale.

Prendendo in considerazione, invece, il grafico *(c)* della Figura 2, si può osservare la composizione media del 70% delle trattazioni, costituito dalle parti variabili del discorso. Anche in questo caso c'è una tipologia, quella dei nomi, che è molto più presente delle altre: costituisce quasi la metà (46%) del totale. Verbi, aggettivi e articoli, invece, sono presenti in proporzioni pressoché simili, pari rispettivamente al 20, 18 e 14 per cento del totale delle parti variabili. Per ultimi troviamo i pronomi, che contribuiscono per il solo 6%.

I risultati osservati, caratterizzati da una distribuzione ricca di nomi a discapito degli aggettivi, assieme all'assenza di interiezioni, sono in linea con la tipologia testuale analizzata. Infatti, queste particolarità rispecchiano quelle che, in linguistica, sono proprie di un testo espositivo con linguaggio scientifico, dunque oggettivo. Ci aspettiamo, inoltre, che quest'ultimo sia di tipo denotativo, dunque che si limiti al significato esplicito e referenziale della parola, senza alcuna libertà di interpretazione.

4.2 Caratteristiche lessicali

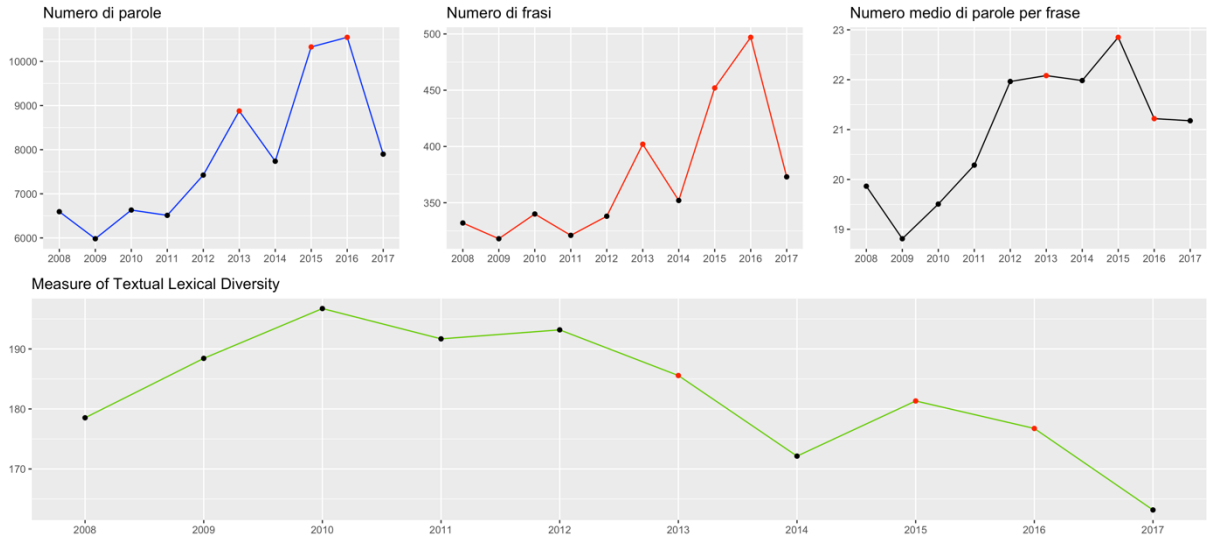

Figura 3: Panoramica delle caratteristiche lessicali dei dati nel periodo considerato

La Figura 3 offre una panoramica del numero di parole e di frasi totali, del numero medio di parole per frase e dell'indice MTLD delle pubblicazioni per ogni anno del periodo 2008-2017. Dai primi due grafici notiamo facilmente come con il passare del tempo le Considerazioni finali del Governatore tendano ad essere più lunghe ed esaurienti. Il numero totale di parole e frasi, infatti, dopo un minimo raggiunto del 2009 (5982 parole in 318 frasi), tende ad aumentare

nell'intervallo fino al 2017. Si osservano facilmente tre picchi, di cui uno meno significativo nel 2013 e due molto rilevanti correlati agli anni 2015 e 2016, che presentano rispettivamente un numero di parole pari a 8877, 10328 e 10546 (contro una media di dieci anni di 7853) e un numero di frasi pari a 402, 452 e 497 (media pari a 373). Le tre osservazioni di questi tre anni sono contrassegnate da un puntino rosso nei grafici della Figura 3.

Spostando l'attenzione sul grafico riguardante il numero medio di parole per frase, osserviamo anche qui un trend positivo, ma più stabile. Si notano, infatti, un balzo tra il 2011 e il 2012, in cui si varia da frasi lunghe mediamente 20.29 parole a frasi mediamente composte da quasi 22 (21.96) parole, ma valori abbastanza stabili per i periodi precedenti e successivi al break. Anche in questo caso il minimo si raggiunge nel 2009 (18.81) e il massimo nel 2015 (22.85), a fronte di una media totale di 20.97. I documenti riferiti agli anni 2013 e 2016, nonostante siano quelli più lunghi, presentano dei valori medi relativi alla lunghezza delle frasi che poco si discostano dalla media del periodo 2012-2017, pari a 21.88.

Il quarto grafico della Figura 3 mostra i valori dell'indice MTLD, che rispecchia la ricchezza lessicale dei testi. Si osservano dei risultati molto diversi da quelli presentati fino ad ora: se in precedenza si notava una tendenza all'aumento col passare del tempo, adesso l'indice di interesse si distribuisce con trend discendente nel periodo di osservazione. Infatti, dopo un picco raggiunto nel 2010, dove si registra un indice MTLD pari a 196.73, con il passare degli anni il valore del dato è diminuito fino al minimo osservato nell'ultimo anno di interesse, pari a 163.17.

Focalizzando l'attenzione sul confronto complessivo tra i periodi 2008-2010, in cui era in carica Mario Draghi, e 2011-2017, in cui il Governatore era Ignazio Visco, possiamo trarre delle conclusioni interessanti. È chiaro, infatti, come nell'ultimo periodo di Draghi si prediligessero trattazioni più snelle, con meno parole e frasi mediamente più brevi, a favore di un lessico più ricco. I documenti redatti da Visco, invece, risultano più esaustivi, composti da frasi mediamente più lunghe e caratterizzati da una inferiore ricchezza lessicale.

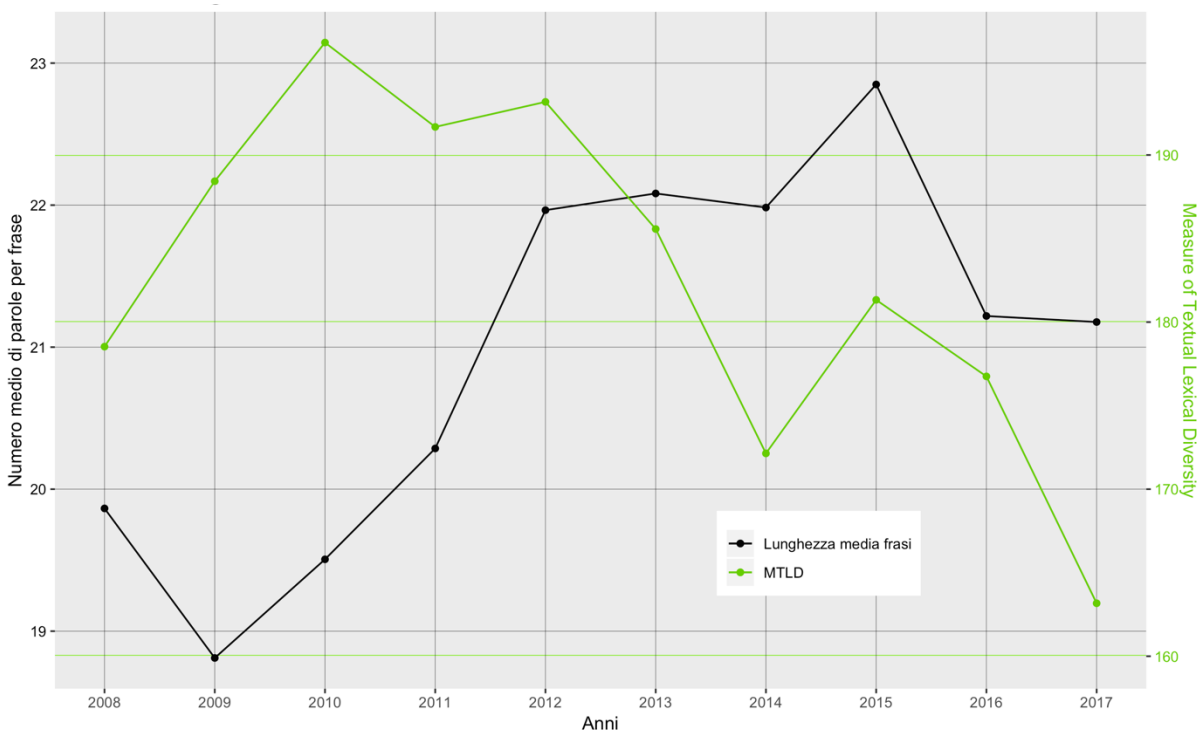

Figura 4: Confronto tra lunghezza media delle frasi e indice di ricchezza lessicale

Inoltre, se si studia il confronto tra la lunghezza media delle frasi e l'indice MTLD, disponibile in Figura 4, si giunge ad ulteriori interessanti risultati. Nel grafico vengono comparati l'andamento del numero medio di parole per frase, in nero, e quello dell'indice di ricchezza lessicale, in verde. Si nota facilmente come le due tendenze siano opposte. Con il passare del tempo, cioè, le frasi sono mediamente più lunghe e i termini presenti al loro interno mediamente meno ricercati. Adottando una visione linguistica dei dati, si può concludere che a frasi più lunghe e complicate, dunque verosimilmente a testi caratterizzati da più subordinazione, corrisponda l'utilizzo di un vocabolario meno ampio.

4.3 Wordcloud per periodi

Figura 5: wordcloud lemmatizzata riferita al periodo 2008-2010

In Figura 5 è riportata la nuvola di parole costituita dai lemmi presenti nelle Considerazioni Finali del Governatore degli anni 2008, 2009 e 2010. Si nota immediatamente come le parole più importanti, ovvero quelle che più ricorrono nei testi e dunque quelle che appaiono più grandi nel grafico, siano *crisi*, *sistema*, *mercato*, *potere* e *impresa*. Risulta chiaro come le pubblicazioni rispecchino la situazione dell'economia italiana del periodo di riferimento, in cui l'Italia era in tempi di assoluta crisi. È verosimile, dunque, che si parlasse di *crisi* che coinvolgesse il *mercato*, l'*impresa*, o addirittura l'intero *sistema*. Nella nuvola ritroviamo anche termini che si collegano alle condizioni alla base della situazione economica del tempo, ovvero l'alto livello del *debito* in rapporto con il *pil*, la scarsa o assente *crescita* economica e la scarsa credibilità di coloro che sono al *potere*.

Figura 6: wordcloud lemmatizzata riferita al periodo 2011-2014

Spostando l'attenzione, invece, sulla nuvola di parole riferita agli anni dal 2011 al 2014 raffigurata in Figura 6, possiamo trarre conclusioni differenti. Sebbene siano ancora molto presenti termini legati alla crisi, situazione che ritroviamo anche nel grafico precedentemente descritto, l'importanza che riveste il termine *europeo* ne fornisce un'interpretazione diversa. Si può osservare, infatti, come nel periodo di riferimento i finanziamenti forniti dall'Unione *Europea* abbiano rivestito un ruolo primario nel risollevare il Paese dalla *crisi*. Inoltre, aumenta la rilevanza dell'*attività* di *vigilanza* che la Banca Centrale *Europea* esegue nell'*area euro*, necessaria al mantenimento della stabilità finanziaria e, dunque, al miglioramento della situazione generale dell'*economia*. L'aiuto fornito dalla Comunità, infine, è sottolineato anche dall'importanza acquisita da termini come *credito* e *fondo*, dato che il contributo monetario viene fornito attraverso diversi Fondi Europei.

Figura 7: wordcloud lemmatizzata riferita al periodo 2015-2017

Anche la nuvola di parole riferita agli anni 2015, 2016 e 2017, riportata in Figura 7, rispecchia la situazione dell'economia italiana di quel periodo. Con il passare del tempo, infatti, si è continuato a usufruire degli aiuti provenienti dall'Unione Europea, ma la situazione non è migliorata. Questo ha portato il *debito pubblico* (parole più presenti nelle trattazioni considerate) ad un *aumento* fino ai massimi storici.[xxii] È verosimile, quindi, che costituisse un problema primario del Paese, dunque di ampia trattazione nelle Considerazioni finali del Governatore.

4.4 *Comparison cloud e commonalty cloud*

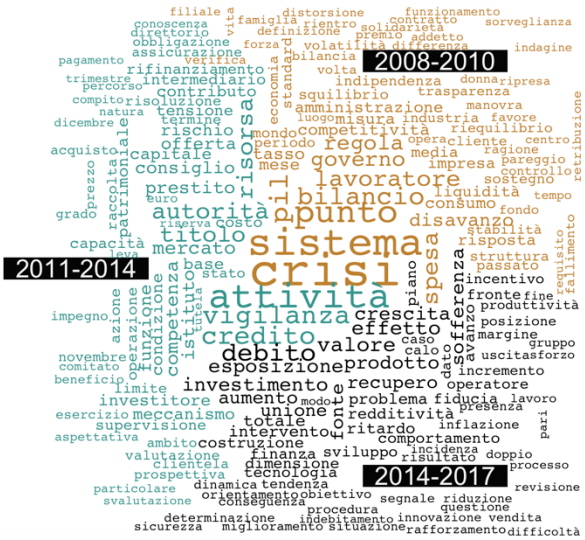

Figura 8: comparison cloud lemmatizzata dei tre periodi

La Figura 8 e la Figura 9 permettono di osservare le differenze e le similitudini nei termini presenti nei documenti.

La *comparison cloud* (Figura 8) descrive i lemmi che caratterizzano un periodo piuttosto che un altro, ovvero presenta un termine come appartenente ad uno dei gruppi solo se quest'ultimo è significativamente più presente nella classe di riferimento rispetto a quanto lo sia nelle altre. Riferendosi al nostro esempio, ciò vuol dire che il fatto che le parole *crisi* e *sistema* siano collegate al primo periodo riflette che, nonostante siano presenti in tutti e tre i gruppi di documenti, sono più frequenti negli anni 2008, 2009 e 2010. È d'interesse osservare l'assenza, rispettivamente nel secondo e terzo gruppo, dei termini *europeo* e *pubblico*, fondamentali per le conclusioni tratte in precedenza. Restano discriminanti, infine, lemmi legati all'*attività* di *vigilanza* e al *debito*.

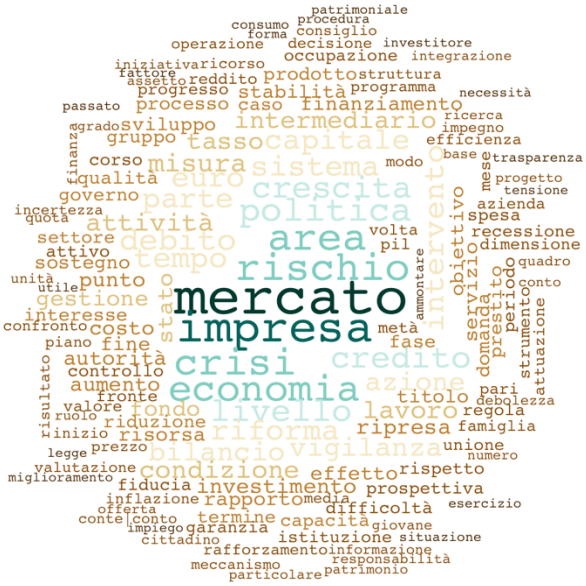

Figura 9: commonalty cloud lemmatizzata

La *commonalty cloud* (Figura 9), invece, presenta le parole più comuni a tutti i documenti, di dimensioni tanto più grandi quanto è maggiore la frequenza minima tra i documenti. Nel grafico, dunque, ritroviamo gli argomenti generali di cui tratta la pubblicazione, quali il *mercato*, l'*impresa* e l'*economia*, ma anche il *bilancio*, la *politica*, la *crescita* dell'*area* italiana.

Infine, è interessante notare come siano frequenti in tutti i documenti parole come *crisi*, *rischio, credito* o *debito*. Evidenza, questa, del fatto che il periodo analizzato è stato caratterizzato da una situazione economica turbolenta.

4.5 Positività dei lemmi

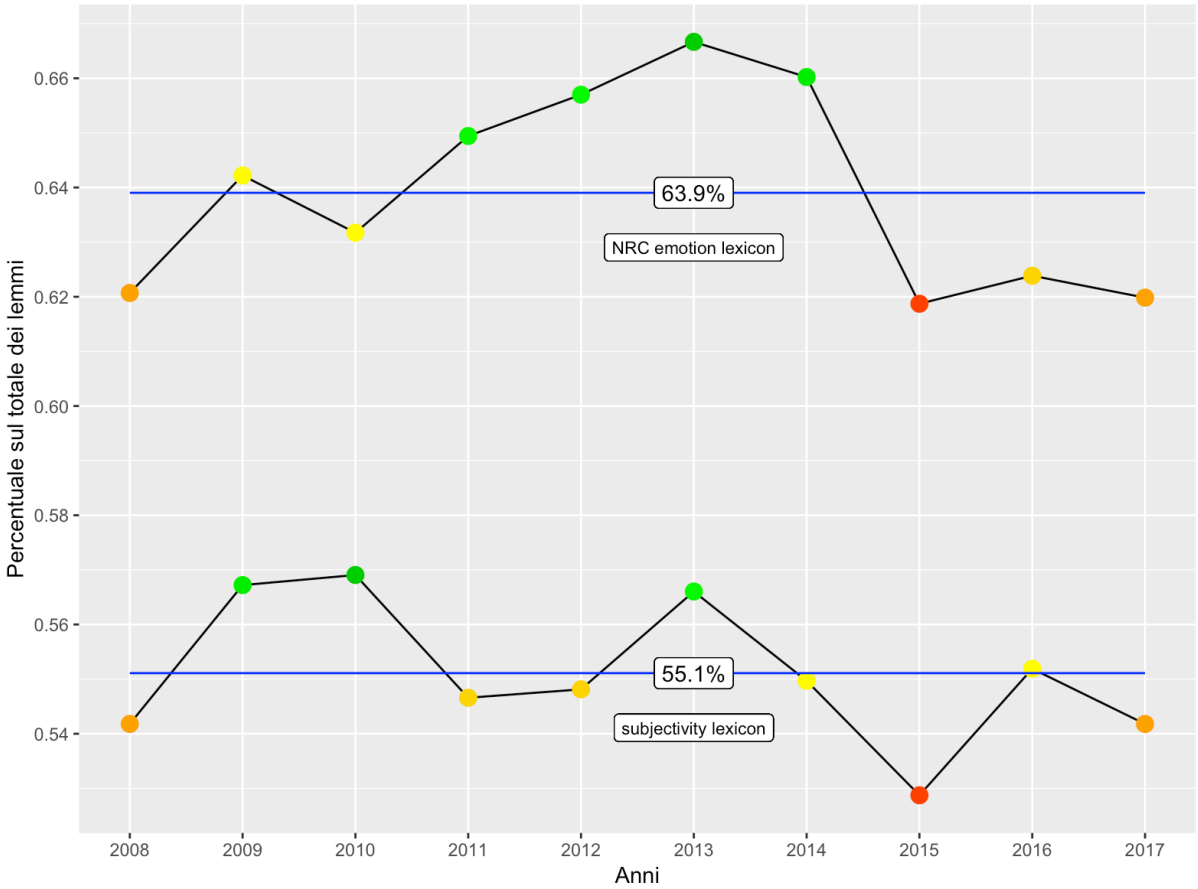

Figura 10: Concentrazione di lemmi positivi: confronto dei risultati ottenuti con i diversi lexicon

Come descritto nel capitolo precedente, le procedure di *opinion mining* sono state eseguite mediante l'utilizzo di due *lexicon* diversi. In Figura 10 è possibile osservare il confronto tra le concentrazioni dei lemmi positivi all'interno dei dieci documenti ottenute a partire dai due dizionari. È importante notare come tutte le osservazioni si dispongano sopra la soglia del 50%, ciò significa che la maggior parte dei lemmi presenti nella pubblicazione, nonostante vengano utilizzati in maniera oggettiva per discutere temi non sempre lieti, rispecchiano un'opinione positiva dell'autore. Tale osservazione è molto più rilevante nei risultati ottenuti attraverso l'utilizzo del *NRC emotion lexicon.* Questi ultimi, infatti, presentano concentrazioni di lemmi positivi nettamente superiori, come si evince dal confronto delle medie: 63.9% quella dei dati ottenuti con il *NRC emotion lexicon,* 55.1% quella con il *subjectivity lexicon.*

Inoltre, per quanto riguarda gli andamenti, si nota come questi siano simili e caratterizzati da un picco di polarità positiva, raggiunto nel 2013, e un minimo assoluto, registrato nell'anno 2015 (a cui corrisponderà, dunque, la massima concentrazione di opinione negativa).

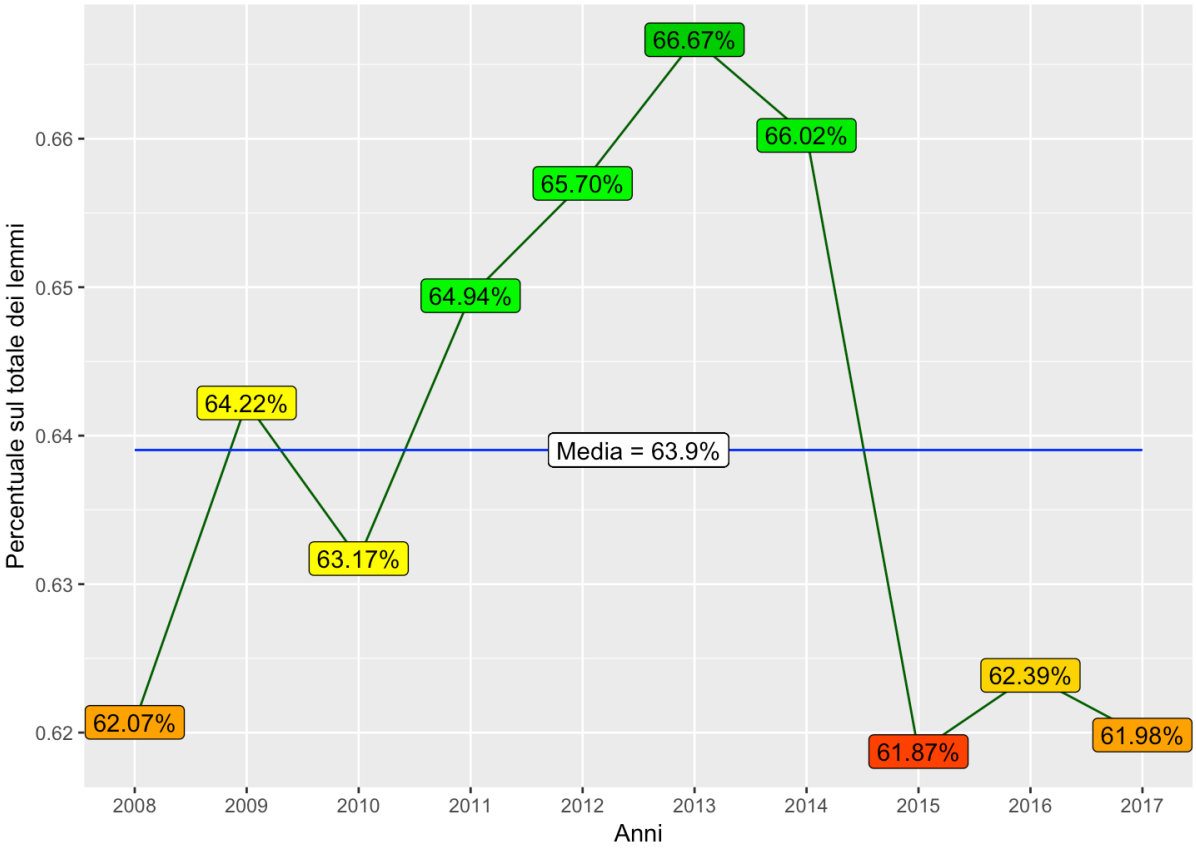

Figura 11: Andamento della concentrazione di lemmi positivi (NRC emotion lexicon)

Tra i due dizionari, quello fornito dal *National Research Council Canada* ha offerto una performance migliore in termini di riconoscimento della polarità dei termini. Ha più senso, dunque, analizzare in dettaglio i risultati ottenuti a partire da questo dizionario, riportati in Figura 11.

L'andamento risulta poco variabile, con un intervallo di valori compreso tra poco meno del 62 e il quasi 67 per cento, che si discostano di poco dalla media complessiva pari al 63.9%. Si possono, inoltre, facilmente distinguere tre periodi caratterizzati da una simile concentrazione di lemmi positivi. Il primo comprende gli anni 2008, 2009 e 2010 ed è caratterizzato da risultati vicini alla media complessiva, soprattutto per gli anni 2009 e 2010; la media delle osservazioni in questo periodo risulta, infatti, pari al 63.2%. Il secondo periodo, invece, corrisponde agli anni dal 2011 al 2014 e presenta concentrazioni decisamente al di sopra del valore intermedio. Difatti, la relativa media equivale al 65.8%. Infine, le pubblicazioni di 2015, 2016 e 2017 sono contraddistinte da una presenza significativamente più bassa di lemmi positivi: il valore medio riferito a questi ultimi tre anni risulta del 62.1%.

4.6 Argomenti delle parti del testo

Anni	Topic 1	Topic 2	Topic 3	Topic 4	Anni	Topic 1	Topic 2	Topic 3	Topic 4
2008	impresa	partecipante	italiano	finanziario	**2013**	pubblico	finanziario	credito	partecipante
	potere	finanziario	credito	mercato		potere	vigilanza	bancario	esercizio
	crisi	vigilanza	bancario	intervento		politica	nazionale	capitale	riserva
	pubblico	direttorio	crisi	internazionale		economico	autorità	impresa	funzione
2009	pubblico	finanziario	finanziario	mercato	**2014**	potere	mercato	attività	Risoluzione
	crisi	crisi	partecipante	bancario		inflazione	bancario	area	Crisi
	punto	dovere	filiale	intermediario		pubblico	vigilanza	rischio	Mercato
	pil	mercato	economia	europeo		crescita	finanziario	impresa	Europeo
2010	impresa	regola	primo	finanziario	**2015**	impresa	nuovo	potere	crisi
	pubblico	rischio	capitale	pil		potere	europeo	europeo	intervento
	italiano	politica	crescita	crisi		pubblico	vigilanza	finanziario	bancario
	spesa	crisi	grande	sistema		economia	finanziario	deteriorare	europeo
2011	mercato	primo	credito	mercato	**2016**	potere	lavoro	politica	debito
	finanziario	finanziario	rischio	finanziario		grande	occupazione	economia	pubblico
	potere	dovere	intermediario	europeo		pil	sistema	investimento	crisi
	pubblico	azione	bancario	bancario		europeo	produttivo	area	mercato
2012	pubblico	vigilanza	impresa	area	**2017**	impresa	impresa	finanziario	debito
	produttivo	attività	bancario	condizione		potere	crescita	potere	pubblico
	potere	nazionale	credito	euro		mercato	economia	deteriorare	finanziario
	attività	direttorio	dovere	europeo		spesa	attività	relazione	potere

Tabella 1: Quattro lemmi che con maggiore probabilità fanno parte dei quattro topics di ognuna delle dieci pubblicazioni

L'applicazione dell'algoritmo LDA ha permesso di ottenere i lemmi che con maggiore probabilità sono congiuntamente presenti all'interno di una sezione del testo piuttosto che in un'altra. Ciò vuol dire che queste quattro parole sono quelle che meglio definiscono l'argomento a cui si riferiscono. Dunque, da un'analisi di questi termini è possibile ricostruire i quattro *topic* affrontati in ogni documento e, di conseguenza, osservare se sono presenti delle ricorrenze.

Nella Tabella 1 sono riassunti i 16 termini di ogni pubblicazione, così da poter avere una visione d'insieme e interpretarli al meglio. Un esempio dell'output di R è invece fornito in Figura 12.

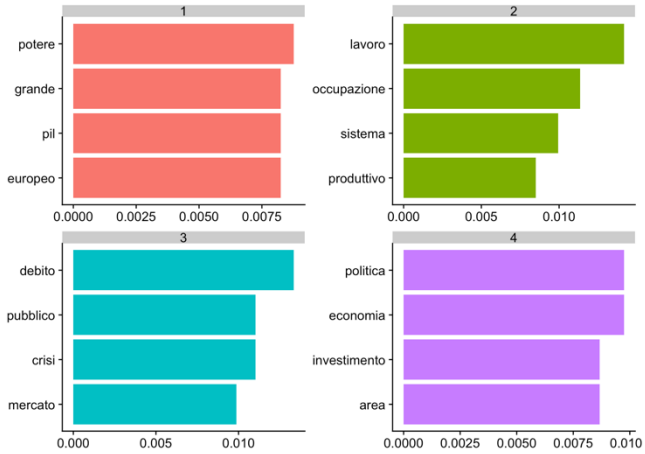

Figura 12: Esempio di output riferito all'anno 2016

Il grafico si riferisce ai risultati che si ottengono analizzando le Considerazioni Finali del Governatore riguardanti l'anno 2016. È possibile osservare i quattro lemmi che maggiormente caratterizzano un determinato argomento e la probabilità collegata agli stessi di appartenenza al tema in questione. Ciò vuol dire, ad esempio, che i lemmi *debito*, *pubblico*, *crisi* e *mercato* sono quelli che con maggiore probabilità si riferiscono congiuntamente ad uno dei quattro *topic* (il 3, nel grafico). Questo significa, verosimilmente, che nel documento del 2016 una sezione tratta il problema del debito pubblico e della crisi del mercato.

Per ogni anno, dunque, le parole chiave sono state riorganizzate nella Tabella 2. A seguito di un'attenta analisi, è possibile riconoscere diversi gruppi di parole ricorrenti. Il più presente è sicuramente quello riferito ai lemmi *impresa* e *pubblico,* spesso accompagnati da *potere* (in rosso in tabella). Si può sostenere, quindi, che in 9 documenti su 10 una sezione del testo sia adibita alla descrizione della situazione delle imprese all'interno mercato, oppure del problema del risanamento dei conti pubblici (ovvero il bilancio dello Stato). Un altro argomento ricorrente è quello riferito alle attività di *vigilanza* bancaria e *finanziaria* che la Banca d'Italia svolge a livello nazionale, presente in 5 trattazioni ed evidenziato in giallo. Ritroviamo spesso anche *topic* collegati al ruolo centrale rivestito dalle banche nel finanziamento dell'economia italiana. In 4 occasioni, infatti, possiamo osservare gruppi che contengono i lemmi *credito* e *bancario* (in viola). Inoltre, si osservano argomenti ricorrenti che caratterizzano solo pochi documenti. Ne sono dei chiari esempi quelli evidenziati in verde, arancione e blu, che ritroviamo solo in due degli ultimi quattro anni in analisi e si riferiscono rispettivamente al *mercato europeo* in relazione alla *crisi*, al *deteriorarsi* del *potere finanziario* e alla questione legata al *debito pubblico*. È interessante notare anche come nel 2016 acquisti importanza il tema del *lavoro* (in grigio), verosimilmente a causa delle politiche anti disoccupazione introdotte in quegli anni.

Infine, si osserva come i termini *crisi* e *debito pubblico* siano presenti soprattutto rispettivamente nel primo e terzo macroperiodo di trattazione. Parole legate al mercato europeo, invece, sono frequenti in tutta la tabella. Ciò rispecchia quasi perfettamente i risultati osservati mediante l'analisi delle nuvole di parole.

5 *Conclusioni*

A seguito della presentazione dei risultati, è necessario e doveroso ricostruire un filo conduttore all'interno di essi e discutere i limiti dell'analisi eseguita.

Il percorso seguito ha permesso di approfondire tre aspetti fondamentali delle Considerazioni finali del Governatore: lo stile e il linguaggio adoperati, i toni e le opinioni mediante cui sono stati presentati i concetti, e i temi trattati nel corso degli anni.

Per quanto riguarda il primo punto, l'analisi della distribuzione media delle parti del discorso ha permesso di ricondurre la tipologia testuale di appartenenza dei dati a quella tipica di un testo espositivo con linguaggio scientifico e oggettivo. Inoltre, passando al vaglio l'evoluzione nel tempo della lunghezza delle parole e delle frasi e dell'indice MTLD, si è concluso che con il passare degli anni le trattazioni tendono ad essere sempre più esaustive e caratterizzate da un vocabolario più ristretto.

A proposito dei toni e delle opinioni, le operazioni di *opinion mining* hanno permesso di dimostrare come la maggior parte dei termini delle trattazioni siano collegate ad emozioni positive. A partire dall'andamento della polarità negli anni, inoltre, è stato possibile ricostruire tre periodi contrassegnati da caratteristiche simili. Il primo, costituito dagli anni 2008, 2009 e 2010, presenta positività in linea con la media totale. Quello intermedio, degli anni che vanno dal 2011 al 2014, è caratterizzato da una concentrazione di lemmi positivi significativamente maggiore. L'ultimo intervallo, composto da 2015, 2016 e 2017, è invece contraddistinto da termini mediamente meno positivi.

Ritroviamo queste considerazioni anche nell'analisi dei temi. Infatti, le *wordcloud* e i risultati dell'applicazione del *topic modeling* dimostrano come gli argomenti di maggiore trattazione nei tre periodi siano rispettivamente la crisi, il mercato europeo e il debito pubblico. Questi risultati sono stati confermati dall'osservazione degli argomenti ricorrenti nelle trattazioni nel tempo. Infine, lo studio dei lemmi presenti nella *commonalty cloud* ci ha permesso di ricostruire i macroargomenti di cui tratta in generale la pubblicazione, ovvero il mercato, le imprese e l'economia.

Tuttavia, va detto che le conclusioni descritte, benché decisamente rilevanti e in linea con quella che è stata la storia dell'economia italiana nel periodo di riferimento, potrebbero essere state oggetto di due diverse forme di distorsione.

La prima è legata all'inevitabile perdita di informazione a cui si va incontro durante i processi automatici di estrazione e riconoscimento del testo. Infatti, le funzioni, gli algoritmi e i software utilizzati, nonostante siano indiscutibilmente accurati, non sono impeccabili. Dunque, può capitare, e stando alla legge dei grandi numeri è sicuramente accaduto, che una parola venga riconosciuta in maniera errata o non venga riconosciuta affatto.

La seconda, purtroppo di maggiore rilevanza potenziale, è collegata all'interpretazione che è stata fornita dei risultati. Se prima, dunque, si trattava di imprecisione della macchina, ora si parla di errore umano. È naturale che il commento, nonostante venga eseguito a partire da dati oggettivi, sia soggettivo, dunque necessariamente diverso da quello che avrebbe offerto un altro individuo. D'altronde, è impossibile identificare un'interpretazione innegabilmente corretta o sbagliata, o tantomeno una più giusta di un'altra. Dal canto mio, posso ritenermi decisamente soddisfatto. E questo è l'importante.

Ringraziamenti

Grazie al professor Drudi, per aver sempre collaborato con il sorriso.

Grazie al professor Alboni, per il vitale e pronto supporto informatico che mi ha offerto.

Grazie a Corrado, per la sua infinita e costante disponibilità e per aver preso a cuore il volermi trasferire metodi e conoscenze, ma anche a Daniele, Elisa e tutti gli altri dell'UBESP di Padova, per avermi dato la possibilità di scoprire nuovi interessi divertendomi. Ringrazio anche il prof. Gregori e la famiglia di Sara, che hanno reso possibile l'esperienza di tirocinio.

Grazie ai miei genitori, per avermi sempre appoggiato e sostenuto sotto tutti i punti di vista e per avermi incoraggiato ad andare lontano a conoscere una nuova realtà, consci del fatto che per loro non sarebbe stato facile.

Grazie a Via del Porto, per avermi cullato e fatto sentire a casa in questi tre anni, andando a costituire quel ricordo che mi accompagnerà per tutta la vita. Grazie a Buba, Dere, Fabio, Giosuè, Giova e Mansu per avermi accolto come una famiglia accoglie un fratello e per avermi iniziato al meglio alla vita da studente. Grazie a Marx, Yoanna, Jack, Fede, Elisa, Nicoula e tutti quelli passati a fare compagnia.

Grazie ad Alex, Anna, Andrea, Bedin, Diego, Fonti, Goy, Luca, Nino, Sara e Vittoria, per essermi stati vicini come può solo chi si incontra quotidianamente e con cui si condivide tutto.

Grazie a Sara, per avermi accompagnato in esperienze che non avrei potuto nemmeno immaginare e per avermi sostenuto in qualsiasi momento, ma soprattutto per avermi insegnato ad apprezzare e amare la leggerezza e la semplicità.

Grazie ad Alberta, Barba, Bob, Caldi, Cassone, Dado, Fracchi, Giacovelli, Marcello, Ovo®, Pasquale, Simone, Will e tutti gli altri amici di Putignano, semplicemente per esserci ed esserci sempre stati.

Bibliografia e sitografia

[i] Banca d'Italia, «Banca d'Italia - Chi siamo», consultato 17 giugno 2019, https://www.bancaditalia.it/chi-siamo/index.html.

[ii] «Governatore della Banca d'Italia», in *Wikipedia*, 15 maggio 2019, https://it.wikipedia.org/w/index.php?title=Governatore_della_Banca_d%27Italia&oldid=104906596.

[iii] Ramzan Talib et al., «Text Mining: Techniques, Applications and Issues», *International Journal of Advanced Computer Science and Applications* 7, n. 11 (2016), https://doi.org/10.14569/IJACSA.2016.071153.

[iv] «Text mining: il processo di estrazione del testo», *Lorenzo Govoni* (blog), 16 luglio 2018, https://lorenzogovoni.com/text-mining/.

[v] Banca d'Italia, «Banca d'Italia - Interventi del Governatore», consultato 20 giugno 2019, https://www.bancaditalia.it/pubblicazioni/interventi-governatore/index.html.

[vi] Hannah Bast e Claudius Korzen, «A Benchmark and Evaluation for Text Extraction from PDF», in *Proceedings of the 17th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '17 (Piscataway, NJ, USA: IEEE Press, 2017), 99–108, http://dl.acm.org/citation.cfm?id=3200334.3200346.

[vii] «ad-freiburg/pdfact: A basic tool that extracts the structure from the PDF files of scientific articles.», consultato 12 febbraio 2019, https://github.com/ad-freiburg/pdfact.

[viii] Ingo Feinerer et al., *tm: Text Mining Package*, version 0.7-6, 2018, https://CRAN.R-project.org/package=tm.

[ix] «TreeTagger - a language independent part-of-speech tagger | Institute for Natural Language Processing | University of Stuttgart», consultato 20 giugno 2019, https://www.ims.uni-stuttgart.de/forschung/ressourcen/werkzeuge/treetagger.en.html.

[x] Meik Michalke et al., *koRpus: An R Package for Text Analysis*, version 0.11-5, 2018, https://CRAN.R-project.org/package=koRpus.

[xi] «stopwords function | R Documentation», consultato 20 giugno 2019, https://www.rdocumentation.org/packages/tm/versions/0.7-6/topics/stopwords.

[xii] «Using the koRpus Package for Text Analysis», consultato 21 giugno 2019, https://cran.r-project.org/web/packages/koRpus/vignettes/koRpus_vignette.html#accessing-data-from-korpus-objects.

[xiii] «MTLD: Lexical Diversity: Measure of Textual Lexical Diversity... in KoRpus: An R Package for Text Analysis», consultato 21 giugno 2019, https://rdrr.io/cran/koRpus/man/MTLD.html.

[xiv] Ian Fellows, *wordcloud: Word Clouds*, version 2.6, 2018, https://CRAN.R-project.org/package=wordcloud.

[xv] Dawei Lang e Guan-tin Chien, *wordcloud2: Create Word Cloud by «htmlwidget»*, version 0.2.1, 2018, https://CRAN.R-project.org/package=wordcloud2.

[xvi] Fabrizio Alboni e Ignazio Drudi, «Materiale didattico fornito per l'insegnamento Utilizzo Statistico di Banche Dati Economiche Online, Laurea in Scienze statistiche, Università di Bologna», 2018-2019.

[xvii] «Janyce Wiebe/Jan Wiebe», consultato 22 giugno 2019, https://people.cs.pitt.edu/~wiebe/.

[xviii] «NRC Emotion Lexicon», consultato 22 giugno 2019, https://saifmohammad.com/WebPages/NRC-Emotion-Lexicon.htm.

[xix] Matthew Jockers, *syuzhet: Extracts Sentiment and Sentiment-Derived Plot Arcs from Text*, version 1.0.4, 2017, https://CRAN.R-project.org/package=syuzhet.

[xx] Bettina Grün e Kurt Hornik, *topicmodels: Topic Models*, version 0.2-8, 2018, https://CRAN.R-project.org/package=topicmodels.

[xxi] «Come funziona LDA - Amazon SageMaker», consultato 28 giugno 2019, https://docs.aws.amazon.com/it_it/sagemaker/latest/dg/lda-how-it-works.html.

[xxii] «Debito pubblico: come, quando e perché è esploso in Italia», Il Sole 24 ORE, consultato 25 giugno 2019, https://www.ilsole24ore.com/art/debito-pubblico-come-quando-e-perche-e-esploso-italia-AEMRbSRG.