Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE

Corso di Laurea in Informatica per il Management

VALUTAZIONE DELLE PRESSIONI DI GOOGLE SUL GOVERNO DEGLI U.S.A.

DALLA PROPOSTA DI UNA METODOLOGIA PER L'ANALISI DEI LOBBYING REPORT AL TOPIC MODELING

Relatore:
Chiar.mo Prof.
EDOARDO VINCENZO
EUGENIO MOLLONA

Presentata da:
MARTINO SIMONETTI

II Sessione
Anno Accademico 2019/2020

Abstract

Alphabet Inc., la holdling a cui fra le altre fa capo Google, è una delle Big Tech, le più grandi e importanti companies che operano nel mondo delle tecnologie dell'informazione. Vista la posizione preminente che occupano nei mercati in cui operano e la carenza di una legislazione specifica, queste grandi aziende sono spesso coinvolte in indagini delle autorità antitrust. Allo scopo di carpire quale sia la strategia aziendale di Alphabet, in un primo momento, sono stati analizzati i dati relativi alle operazioni di lobbying svolte dal 2010 al 2020 e in un secondo momento sono stati analizzati dati testuali contenenti informazioni su Google tramite l'algoritmo di Topic Modeling LDA. A tal proposito è stato possibile fare un confronto fra due differenti tool per il Topic Modeling: Mallet e Gensim. È inoltre stato possibile utilizzare le API di Twitter per scaricare dati testuali, da utilizzare per l'indagine.

Indice

Elenco delle figure

Introduzione

Nell'ultimo decennio le grandi aziende tecnologiche hanno aumentato sempre di più i servizi rivolti ai clienti, permeando così, sempre più, le vite di milioni di persone. Google in particolare offre decine di servizi, partendo dal Motore di ricerca, la Gmail, il sistema operativo Android, fino ad arrivare a servizi come GPay. L'azienda di Mountain View, nel corso dell'ultimo ventennio, ha ottenuto una posizione di primo piano, congegnando il proprio business attraverso uno scambio implicito fra servizi gratuiti offerti agli utenti e la cessione di dati da parte degli stessi.

Da questa situazione è nata la necessità di provare a carpire quali siano la ragioni di questo successo.

Questo lavoro di tesi, quindi, tende ad analizzare alcuni aspetti dell'attività di Google, in particolare l'attività di lobbying verso il Congresso degli Stati Uniti, allo scopo di individuare le strategie utilizzate. L'analisi svolta fa riferimento all'arco temporale 2010 - 2020. Al fine di ampliare la mole di informazioni a disposizione si è deciso di procedere con un'analisi dei bilanci societari, e un'analisi testuale di articoli di stampa e tweet, tramite algoritmi di text mining, aventi per oggetto l'attività di influenza sul sistema politico legislativo.

La parte più significativa del lavoro è consistita nel sviluppare delle procedure che permettessero di estrapolare dati significativi da materiali strutturati diversamente.

Capitolo 1

Alphabet Inc. e la pressione sul governo U.S.A.

Alphabet Inc. è una holding multinazionale statunitense con sede a Mountain View, California. È stata costituita, il 2 ottobre 2015, in occasione di una ristrutturazione del gruppo Google ed è diventata la capofila di Google e di diverse ex consociate.[3]

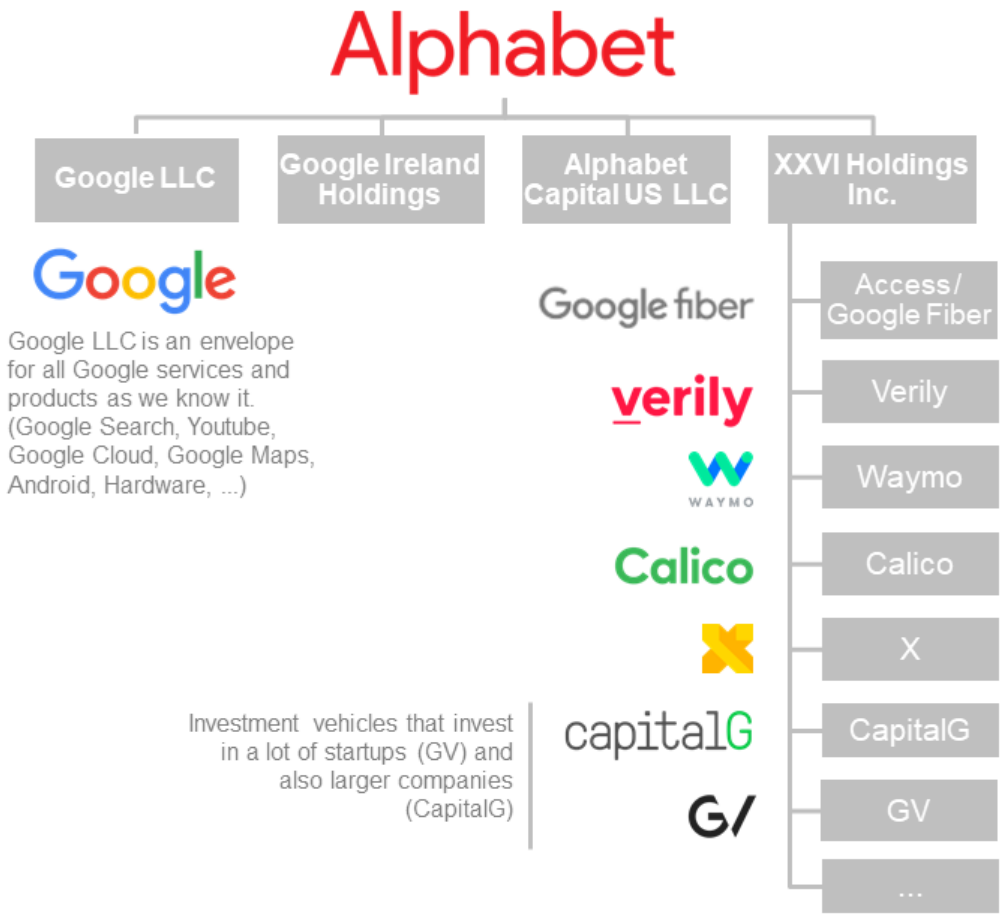


Figura 1.1: Struttura aziendale di Alphabet Inc. al 2018 [1]

Google LLC invece nasce nel 1998 quando Larry Page e Sergey Brin, studenti dell'Università di Stanford, mettono in rete il loro motore di ricerca. "Dal 2001 Google ha acquistato oltre 160 aziende, provenienti dai più svariati settori, per poi essere integrate in Google, o lasciate almeno in parte indipendenti."[4]

Google dichiara che: "La nostra missione è organizzare le informazioni a livello mondiale e renderle universalmente accessibili e utili.".[5]

Google opera in molti settori, i principali servizi che offre ai propri clienti sono: il motore di ricerca Google, il sistema operativo Android, il sistema operativo Chrome OS e servizi web quali YouTube, Gmail, Play Store, Google Maps e molti altri.

Dal 2010 al 2020 negli U.S.A. sono stati spesi oltre 2 miliardi di dollari l'anno per spese di lobbying, circa l'1% di questa spesa è stata sostenuta da Google prima e da Alphabet poi.

	2019	2018	2017	2016	2015	2014	2013	2012	2011	2010
Spesa Google	$ 16,565	$ 32,153	$ 22,773	$ 19,636	$ 26,192	$ 25,900	$ 21,090	$ 24,045	$ 13,503	$ 6,970
Spesa Totale	$ 2,614,548	$ 2,613,110	$ 2,484,193	$ 2,384,912	$ 2,423,368	$ 2,433,729	$ 2,406,593	$ 2,468,775	$ 2,472,491	$ 2,639,241
Google/Tot	0.6%	1.2%	0.9%	0.8%	1.1%	1.1%	0.9%	1.0%	0.5%	0.3%

(Dati in migliaia di Dollari Americani)

Figura 1.2: Spesa Lobbying Google su toltale USA, Fonte dati opensecrets.org [Elaborazione dell'autore]

1.1 Lobby

Si può definire Lobby come "Termine usato negli Stati Uniti d'America, e poi diffuso anche altrove, per definire quei gruppi di persone che, senza appartenere a un corpo legislativo e senza incarichi di governo, si propongono di esercitare la loro influenza su chi ha facoltà di decisioni politiche, per ottenere l'emanazione di provvedimenti normativi, in proprio favore o dei loro clienti, riguardo a determinati problemi o interessi" [6]. In particolare i lobbyisti sono solitamente persone ben collegate agli organi decisionali dello Stato, spesso sono avvocati o ex parlamentari, che utilizzano le loro relazioni con il sistema politico legislativo per esercitare, dietro compenso, influenze al fine di ottenere vantaggi normativi a favore dei loro committenti.[7]

Negli Stati Uniti una lunga tradizione di lobbying, circa due secoli, ha portato, già dal 1946, a regolamentare il fenomeno. [8]

Il Federal Regulation of Lobbying Act del 1946 è uno statuto emanato dal Congresso degli Stati Uniti per ridurre l'influenza delle lobby. Lo scopo principale della legge era quello di fornire informazioni ai membri del Congresso su coloro che esercitavano pressioni su di essi. La legge del 1946 è stata abrogata dal Lobbying Disclosure Act (LDA[1]) del 1995. [9]

[1]Disambiguazione: in questo capito l'acronimo LDA verrà usato per indicare il Lobbying Disclosure Act. Nei capitoli sucessivi l'acronimo LDA sarà utilizzato riferendosi al modello statistico Latent Dirichlet Allocation.

L'LDA è integrata nal 2007 dal Honest Leadership and Open Government Act. Il nuovo quadro normativo tende a rafforzare il grado di trasparenza nell'attività dei lobbisti.

Le norme definiscono **cliente**, qualsiasi persona o entità che impieghi o retribuisca un'altra persona, con compensi finanziari o di altro tipo, per condurre attività di lobbying per conto del cliente stesso. È considerato cliente anche un ente che utilizza i propri dipendenti per attività dirette lobbying. Quindi l'attività di lobbying si può svolgere sia indirettamente, incaricando terzi, sia utilizzando i propri dipendenti.

Le norme citate definiscono inoltre il **lobbista**, come qualsiasi individuo o ente impiegato o retribuito da un **cliente**, per servizi che includono più di un'attività di lobbismo. Non viene considerato lobbista un soggetto le cui attività di lobbismo costituiscono, nell'arco di un trimestre, meno del 20% della sua attività professionale.

Nella legislazione sono definiti gli interventi che devono essere divulgati alle autorità competenti.

Vengono inoltre, all'interno della sezione 4 del LDA, definite le regole per la registrazione dei lobbisti al Congresso degli U.S.A.: ogni organizzazione che si mette in contatto con un lobbista lo deve registrare entro 45 giorni alla Camera dei Rappresentanti e al Senato. La sezione 5 del LDA definisce le modalità di pubblicazione e il contenuto dei report trimestrali, che i lobbisti sono tenuti a counicare alle autorità.[10] Questi report sono l'oggetto dell'analisi descritta in questo lavoro.

1.2 Contesto politico

Al fine di comprendere meglio il contesto politico in cui si sviluppano le pressioni di Google viene fornita la seguente tabella.

		Senate		*House of Representatives*	
		Conservatori	Democratici	Conservatori	Democratici
2011/12	112° Congresso	47	53	241	191
2013/14	113° Congresso	45	55	234	201
2015/16	114° Congresso	54	44	247	187
2017/18	115° Congresso	51	49	236	196
2019/20	116° Congresso	53	47	199	235

(fonte dati: Wikipedia)

Figura 1.3: Spese di Lobbying Google Inc. [Elaborazione dell'autore]

Le celle evidenziate in rosso rappresentano la maggioranza del Partito Repubblicano, mentre le celle evidenziate in blu rappresentano la maggioranza del Partito Democratico.

Interessante è notare come durante il 115° Congresso, con la presidenza di Donald Trump, il partito Repubblicano aveva anche la maggioranza in entrambe le Camere.

1.3 Spese in Lobby nel bilancio di Google

Il seguente grafico, ottenuto dall'analisi dei lobbying report (nel Capitolo 2 verrà approfondito come sono stati raccolti i dati), rappresenta gli investimenti annuali di Google in lobbying dal 2010 al 2019.

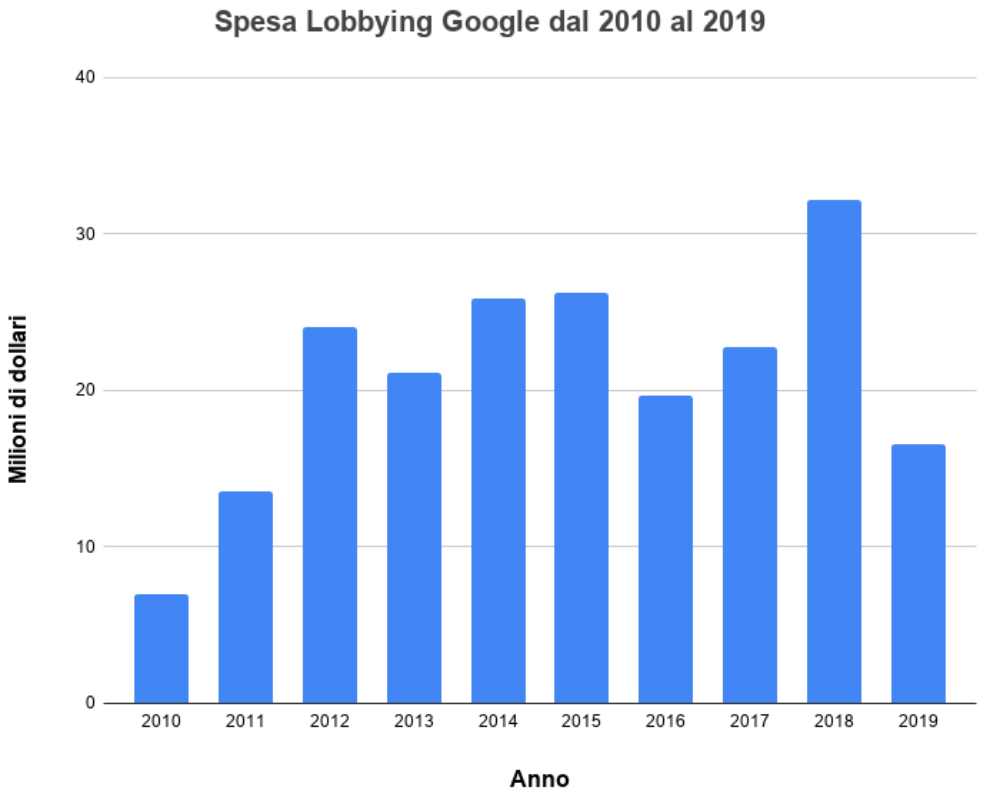


Figura 1.4: Spese di Lobbying Google Inc. [Dati raccolti dai lobbying report][Elaborazione dell'autore]

Come si può apprezzare dal grafico (Figura 3), dal 2010 al 2012 Google ha progressivamente aumentato il proprio investimento in lobbying, dal 2012 in poi tale spesa si è assestata intorno ai venti, trenta milioni di dollari l'anno. Nel 2019 si nota, invece, una

riduzione, in questo tipo di spese, del 50% circa. Tale tendenza viene confermata dai dati relativi al 2020, in cui si registra una spesa inferiore ai dieci milioni, nei primi tre trimestri.

È stato deciso di confrontare i dati delle spese annuali in lobbying con quelli relativi ai bilanci di Google-Alphabet dello stesso periodo. Tali dati sono stati reperiti dai "FISCAL YEAR RESULTS" che si possono trovare sul sito di Alphabet Inc. (2015-2020) e sul sito della SEC[2] (2010-2014).[2]

Sono stati analizzati i principali indicatori del conto economico.

[2]La SEC, Securities and Exchange Commission (Commissione per i Titoli e gli Scambi) è l'ente federale statunitense preposto alla vigilanza della borsa valori. [11]

(Dati in milioni di Dollari Americani)

	Google Inc.					Alphabet Inc.				
	2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**	**2017**	**2018**	**2019**
Revenues	$ 29,321	$ 37,905	$ 50,175	$ 59,825	$ 66,001	$ 74,989	$ 90,272	$110,855	$136,819	$161,857
Cost of revenues	$ 10,417	$ 13,188	$ 17,176	$ 21,993	$ 25,691	$ 28,164	$ 35,138	$ 45,583	$ 59,549	$ 71,896
Research and development	$ 3,762	$ 5,162	$ 6,793	$ 7,952	$ 9,832	$ 12,282	$ 13,948	$ 16,625	$ 21,419	$ 26,018
Sales and marketing	$ 2,799	$ 4,589	$ 6,143	$ 7,253	$ 8,131	$ 9,047	$ 10,485	$ 12,893	$ 16,333	$ 18,464
General and administrative	$ 1,962	$ 2,724	$ 3,845	$ 4,796	$ 5,851	$ 6,136	$ 6,985	$ 6,872	$ 8,126	$ 9,551
European Commission fines	$ -	$ -	$ -	$ -	$ -	$ -	$ -	$ 2,736	$ 5,071	$ 1,697
Total costs and expenses	$ 18,940	$ 25,663	$ 37,415	$ 45,859	$ 49,505	$ 55,629	$ 66,556	$ 84,709	$110,498	$127,626
Reddito operativo	$ 10,381	$ 12,242	$ 12,760	$ 13,966	$ 16,496	$ 19,360	$ 23,716	$ 26,146	$ 26,321	$ 34,231
delta RO	-	18%	4%	9%	18%	17%	23%	10%	1%	30%
Utile netto	$ 8,505	$ 9,737	$ 10,737	$ 12,920	$ 14,444	$ 16,348	$ 19,478	$ 12,662	$ 30,736	$ 34,343
delta Utile		14%	10%	20%	12%	13%	19%	-35%	143%	12%
Spese lobbying	$ 7.00	$ 13.50	$ 24.00	$ 21.00	$ 26.00	$ 26.20	$ 19.60	$ 22.80	$ 32.15	$ 16.56
delta Lobbying	-	93%	78%	-13%	24%	1%	-25%	16%	41%	-48%
lobbying/costi totali	0.04%	0.05%	0.06%	0.05%	0.05%	0.05%	0.03%	0.03%	0.03%	0.01%
lobbying/RO	0.07%	0.11%	0.19%	0.15%	0.16%	0.14%	0.08%	0.09%	0.12%	0.05%
lobbying/utile netto	0.08%	0.14%	0.22%	0.16%	0.18%	0.16%	0.10%	0.18%	0.10%	0.05%

Figura 1.5: Gestione caratteristica, utile netto, spese lobbying e indici ricavati. [Elaborazione dell'autore]

A partire dai valori estrapolati dai bilanci sono stati calcolati i seguenti indicatori:

- Il reddito operativo (RO), il margine derivante dall'attività tipica dell'azienda, così calcolato:

$$RO = Revenues\ (ricavi) - Total\ costs\ and\ expenses\ (costi\ totali)$$

- Delta RO, Delta Utile Netto e Delta Spese Lobby vanno a definire la variazione percentuale fra il valore dell'anno corrente rispetto all'anno precedente:

$$\Delta = \left(\frac{valore_{anno}}{valore_{anno-1}} - 1 \right) \%$$

- L'incidenza delle spese di lobbying rispetto al RO, all'utile netto e ai costi totali.

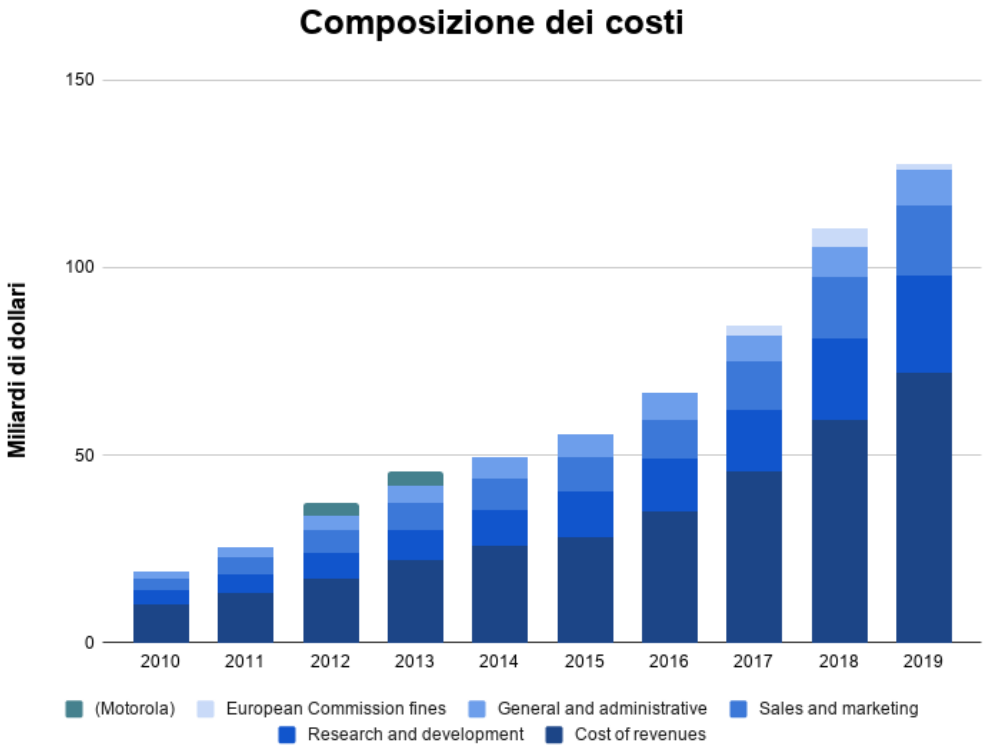


Figura 1.6: Composizione dei costi totali risultanti dal conto economico di Alphabet dal 2010 al 2019. [dati bilanci [2]] [Elaborazione dell'autore]

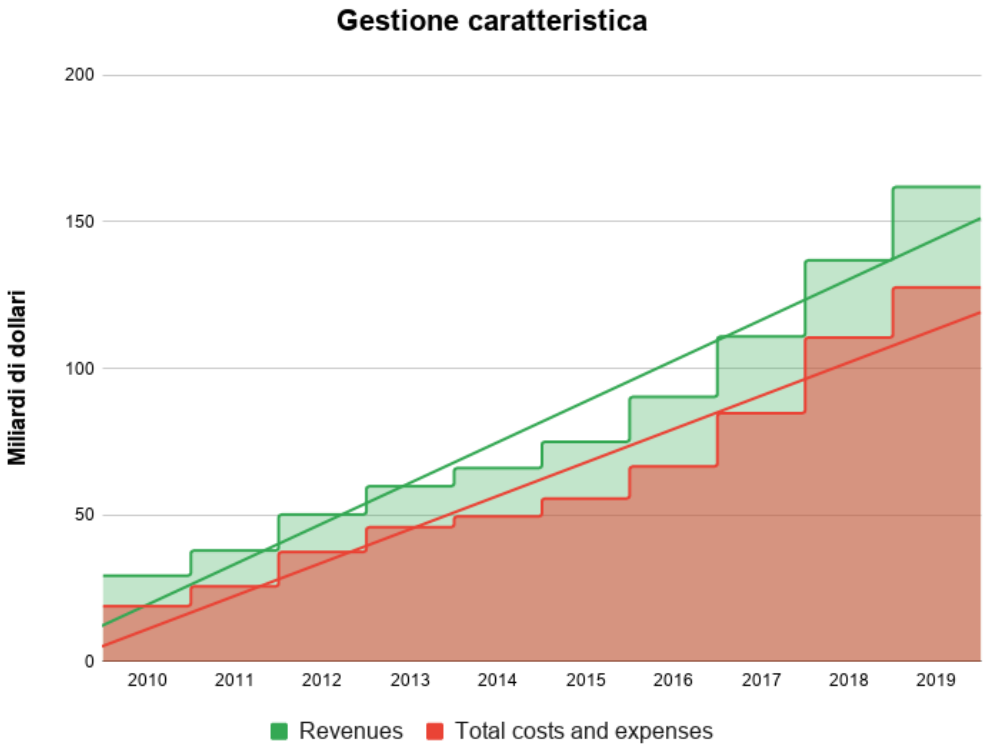


Figura 1.7: L'area verde fra la curva rossa e quella verde corrisponde al RO di Alphabet dal 2010 al 2019. [dati bilanci [2]] [Elaborazione dell'autore]

Si può notare come col passare del tempo ricavi e costi aumentino in modo sostanzialmente lineare. Tuttavia la retta interpolante i ricavi ha una pendenza maggiore rispetto a quella dei costi, di conseguenza anche il RO aumenta linearmente.

Interessante è l'incidenza che le spese di lobbying hanno, per Google, rispetto ai costi totali, all'utile netto e al RO. Il grafico che segue (Figura 1.8) rappresenta tali incidenze percentuali.

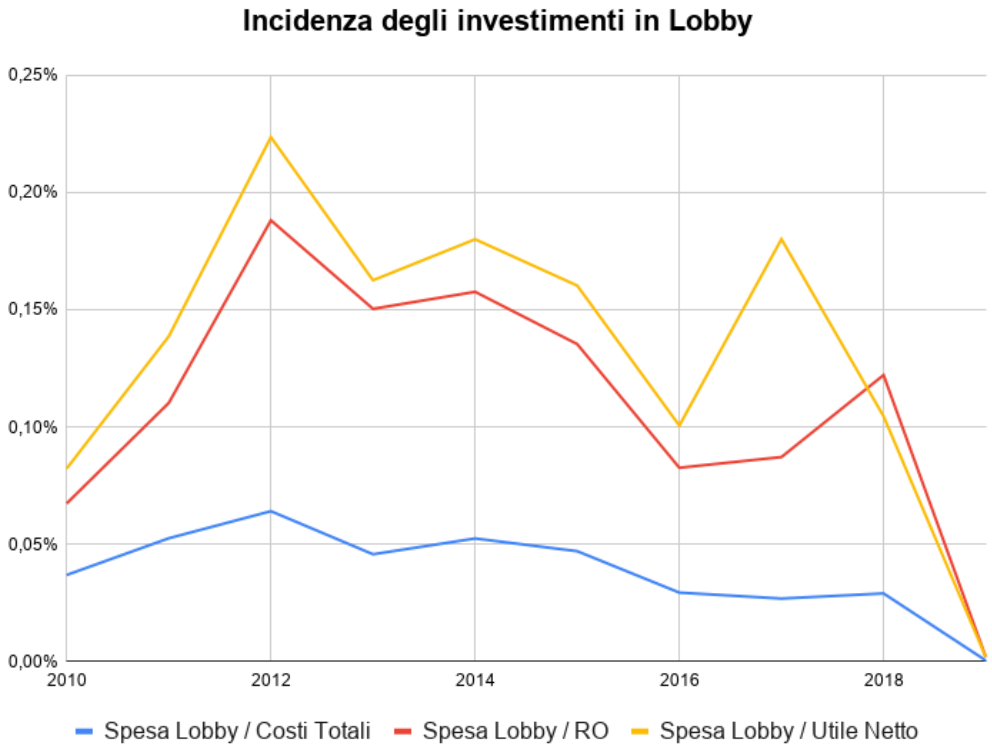


Figura 1.8: Incidenza percentuale al variare del tempo. [dati bilanci [2]] [Elaborazione dell'autore]

Si noti come le spese di lobbying, pur essendo elevate in termini assoluti, non incidano in modo significativo rispetto ai costi e agli utili di Google.

Per confrontare significativamente i dati relativi ai bilanci con quelli riguardanti gli investimenti in lobby, si possono considerare gli incrementi (decrementi) percentuali dei valori sopra citati.

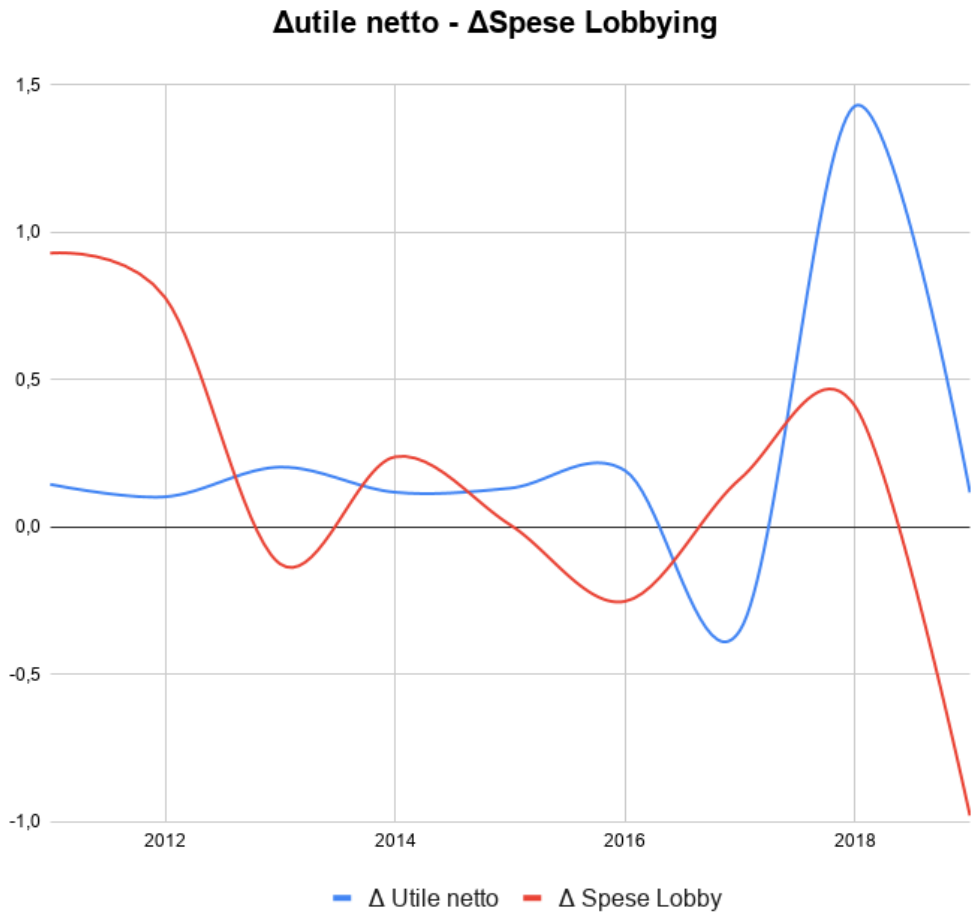


Figura 1.9: Andamento al variare del tempo dei valori Delta Utile Netto e Delta Spese Lobby. [dati bilanci [2]] [Elaborazione dell'autore]

Si può apprezzare come, in alcuni casi, l'andamento della curva blu (delta Utile Netto) venga anticipato dall'andamento della curva rossa (delta Spese Lobby).

In conclusione, dai dati utilizzati in questo contesto e dall'elaborazione degli stessi, si può evincere che, benché le spese di lobbying di Google siano importanti in valore assoluto, non sembrano incidere significativamente sugli altri valori di bilancio.

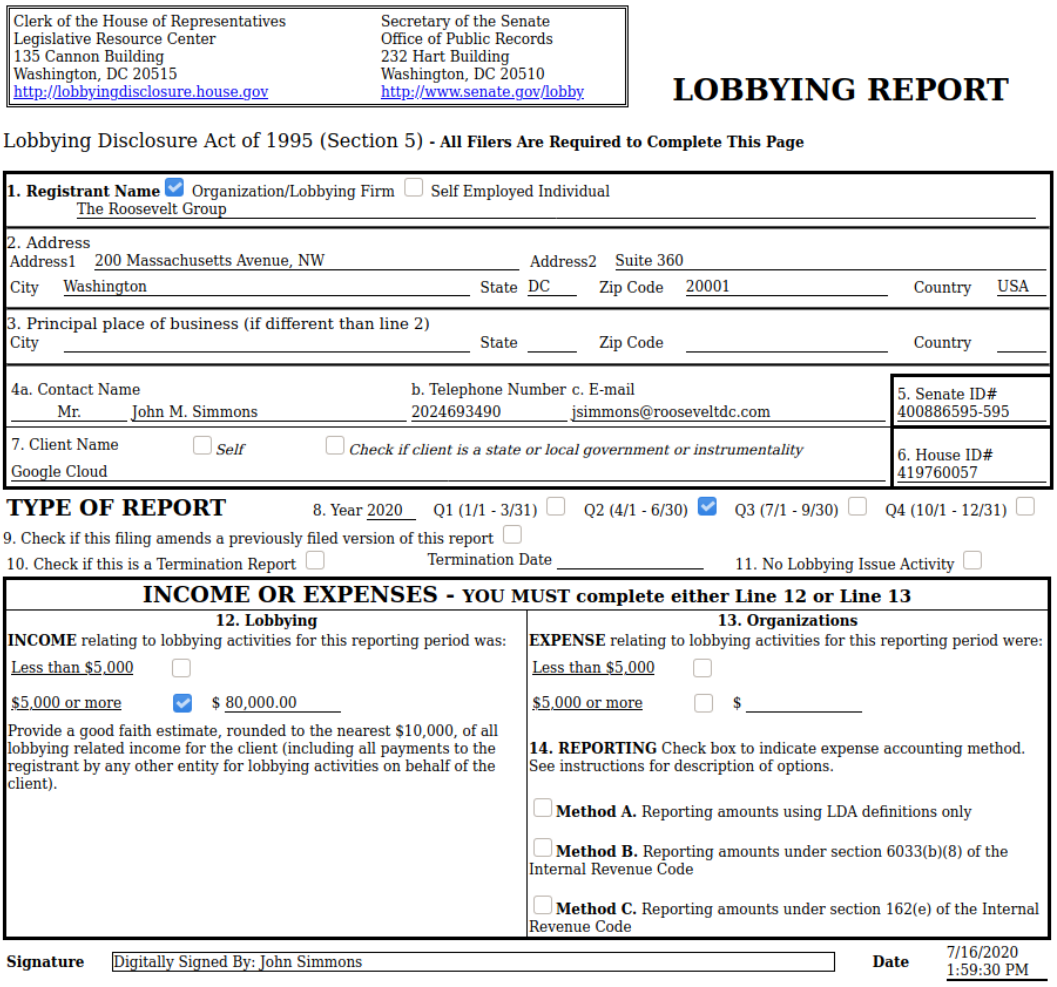
Capitolo 2

Raccolta dati Lobbying Report

Quanto descritto in questo paragrafo è stato sviluppato per l'analisi dei lobbying report di Alphabet Inc., la stessa metodologia (senza alcuna modifica sugli script) potrà essere usata per analizzare altre aziende statunitensi.

2.1 Struttura di un Lobbying Report

Allo scopo di raccogliere dati relativi alle operazioni di lobbying di Google sono stati analizzati i lobbying report presentati dall'azienda al Congresso degli U.S.A. Il LDA del 1995, successivamente modificato dal "Honest Leadership and Open Government Act" del 2007, prevede che tutte le aziende di lobbisti presentino rapporti trimestrali delle loro attività al Clerk della Camera dei Rappresentanti e al Segretario del Senato degli Stati Uniti. [12]



Clerk of the House of Representatives
Legislative Resource Center
135 Cannon Building
Washington, DC 20515
http://lobbyingdisclosure.house.gov

Secretary of the Senate
Office of Public Records
232 Hart Building
Washington, DC 20510
http://www.senate.gov/lobby

LOBBYING REPORT

Lobbying Disclosure Act of 1995 (Section 5) - **All Filers Are Required to Complete This Page**

1. Registrant Name ☑ Organization/Lobbying Firm ☐ Self Employed Individual
The Roosevelt Group

2. Address
Address1 200 Massachusetts Avenue, NW Address2 Suite 360
City Washington State DC Zip Code 20001 Country USA

3. Principal place of business (if different than line 2)
City State Zip Code Country

4a. Contact Name: Mr. John M. Simmons
b. Telephone Number: 2024693490
c. E-mail: jsimmons@rooseveltdc.com
5. Senate ID# 400886595-595

7. Client Name ☐ *Self* ☐ *Check if client is a state or local government or instrumentality*
Google Cloud
6. House ID# 419760057

TYPE OF REPORT 8. Year 2020 Q1 (1/1 - 3/31) ☐ Q2 (4/1 - 6/30) ☑ Q3 (7/1 - 9/30) ☐ Q4 (10/1 - 12/31) ☐
9. Check if this filing amends a previously filed version of this report ☐
10. Check if this is a Termination Report ☐ Termination Date 11. No Lobbying Issue Activity ☐

INCOME OR EXPENSES - YOU MUST complete either Line 12 or Line 13

12. Lobbying
INCOME relating to lobbying activities for this reporting period was:
Less than $5,000 ☐
$5,000 or more ☑ $ 80,000.00
Provide a good faith estimate, rounded to the nearest $10,000, of all lobbying related income for the client (including all payments to the registrant by any other entity for lobbying activities on behalf of the client).

13. Organizations
EXPENSE relating to lobbying activities for this reporting period were:
Less than $5,000 ☐
$5,000 or more ☐ $
14. REPORTING Check box to indicate expense accounting method. See instructions for description of options.
☐ **Method A.** Reporting amounts using LDA definitions only
☐ **Method B.** Reporting amounts under section 6033(b)(8) of the Internal Revenue Code
☐ **Method C.** Reporting amounts under section 162(e) of the Internal Revenue Code

Signature Digitally Signed By: John Simmons **Date** 7/16/2020 1:59:30 PM

Figura 2.1: Esempio di Lobbing Report: parte iniziale

All'interno dei lobbing report troviamo molteplici informazioni, le più interessanti ai fini di questa ricerca sono le seguenti:

- Il nome dell'azienda lobbista, sezione "1. Registrant Name";
- Il nome del cliente, sezione "7. Client Name";
- Il costo delle operazioni svolte nel trimestre, sezione "12. Lobbying" (se il cliente ha ingaggiato un'azienda esterna per esercitare pressioni sulla politica) oppure sezione "13. Organizations" (se il lobbista coincide con il cliente, in questo caso infatti sarà l'azienda stessa ad esercitare pressioni).

Tutte le informazioni che si trovano in questa prima parte dei lobbying report sono disponibili sul sito https://www.opensecrets.org. OpenSecrets.org permette, nella sezione "Lobbying", di filtrare i Report per cliente e anno, sarà quindi possibile visualizzare una tabella con tutte le informazioni sopra citate.

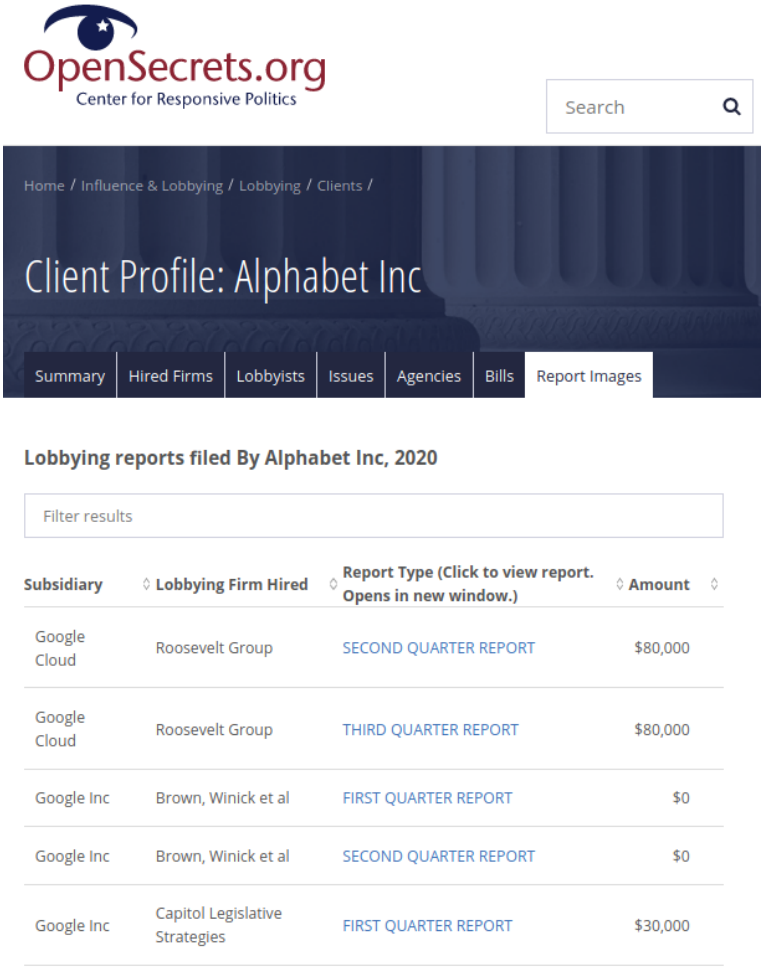


Figura 2.2: opensecrets.org: Report Images

I Lobbying Report hanno una seconda sezione chiamata "LOBBYING ACTIVITY". Uno stesso report, di norma, ha più d'una lobbying activity.

LOBBYING ACTIVITY. Select as many codes as necessary to reflect the general issue areas in which the registrant engaged in lobbying on behalf of the client during the reporting period. Using a separate page for each code, provide information as requested. Add additional page(s) as needed.

15. General issue area code TEC

16. Specific lobbying issues

Privacy, competition tech, and small business policy issues
CDA Sec. 230
H.R.6800 - The Heroes Act
Issues related to COVID-19 relief

17. House(s) of Congress and Federal agencies ☐ Check if None

U.S. SENATE, U.S. HOUSE OF REPRESENTATIVES

18. Name of each individual who acted as a lobbyist in this issue area

First Name	Last Name	Suffix	Covered Official Position (if applicable)	New
Oscar	Ramirez			☐
Dana	Thompson			☐
Aaron	Trujillo			☐
Miya	Patel			☐

Figura 2.3: Esempio di Lobbing Report: seconda parte

In questa seconda sezione si possono ricavare informazioni relative alle tematiche specifiche delle operazioni di lobbying. In particolare:

- Nella sezione "15. General issue area code" dev'essere inserito un codice che identifichi un'area tematica, tutti i possibili "General issue area code" sono disponibili nell'Appendice A. Sono dei codici che identificano delle aree di interesse politico nelle quali si può sviluppare un'operazione di lobbying.
- Nella sezione "16. Specific lobbying issues" possono essere presenti codici di leggi specifiche sulle quali il lobbyista sta esercitando pressione.

Per accedere alle informazioni presenti in questa seconda sezione è necessario aprire il report. Nella tabella disponibile su openSecrets.org (vedi Figura 2.2) sono presenti i link alle pagine web che permettono di visualizzare i lobbying report.

2.2 Raccolta dati Lobbying Activity

Ai fini della ricerca è interessante capire quali "General Issue Area Code" sono presenti in ogni report, sarà poi possibile stimare a quanto ammonta la spesa destinata ad

ogni area d'interesse:

$$costoSingoloIssueAreaCode = \frac{\text{costo Totale Report}}{numeroTotaleIssueAreaCodePresenti}$$

Si potrà poi sapere in quanti report sarà stato utilizzato ogni singolo Issue Area Code in un anno e si potrà stimare quanto è stato speso per ogni area. Per effettuare il conteggio è stato utilizzato un particolare procedimento, che viene schematizzato di seguito:

1. preparazione di un CSV[1] con i link dei report e il relativo costo;
2. conteggio tramite script Python;
3. elaborazione dei dati ricavati dallo script.

2.2.1 Preparazione CSV

A partire dalla tabella "Report Images" del sito opensecrets.org (vedi Figura 2.2) sono state copiate per intero le colonne "Report Type" che contengono i link ai report e la colonna "Amount" dove sono contenuti i costi totali dei report. Le colonne sono state incollate in un foglio di calcolo elettronico che supporta le Macro VBA (è stato utilizzato LibreOffice Calc). A questo punto tramite una macro vengono estrapolati gli URL dai collegamenti ipertestuali, col seguente codice VisualBasic:

```
sub SHOWURL
   foglio = ThisComponent.CurrentController.activeSheet
   celleSelezionate = ThisComponent.Currentselection

   if celleSelezionate.columns.count > 1 then
       msgbox ("Errore: seleziona solo la colonna con i link",16,"Error")
       exit sub
```

[1] Il comma-separated values (abbreviato in CSV) è un formato di file, basato su file di testo utilizzato per l'importazione ed esportazione (ad esempio da fogli elettronici o database) di una tabella di dati.

```
    end if

    colonna = celleSelezionate.Rangeaddress.Startcolumn

    for i = 0 to celleSelezionate.rows.count - 1
        cellaCorrente = celleSelezionate.getcellbyPosition(0,i)
        rigaCorrente = cellaCorrente.celladdress.row
        cellaScrittura = foglio.getcellbyposition(colonna+1,rigaCorrente)

        if cellaCorrente.Textfields.count = 1 then

            if cellaCorrente.Textfields(0).URL > "" then
                stringaURL = cellaCorrente.Textfields(0).URL
                cellaScrittura.String = sringaURL
            else
               goto 100
            endif

        else
            100:
            msgbox (cellaCorrente.AbsoluteName+
            " non contiene un link",64,"Information")
        endif

    next i

End sub
```

Una volta estratti gli URL bisognerà creare una tabella che, come prima riga, abbia tutti i possibili General Issue Area Code e che nelle prime due colonne abbia rispettivamente gli URL e gli investimenti per ogni report.

		ACC	ADV	. . .	WEL
https://soprweb.senate.gov/index...	80000				
https://soprweb.senate.gov/index...	100000				
https://soprweb.senate.gov/index...	90000				
. . .	. . .				

Tabella 2.1: Esempio formato CSV da esportare.

Una volta preparata la tabella, come descritto, la si può esportare come file in formato CSV.

2.2.2 Conteggio tramite script Python

È stato sviluppato il seguente script Python per il conteggio degli Issue Area Code all'interno dei report.

```
#!/usr/bin/python3
# -*- coding: utf-8 -*-
from urllib.request import urlopen
import urllib.request
from bs4 import BeautifulSoup
from io import StringIO
from numpy import genfromtxt, shape, savetxt

# salva in una matrice il contenuto del file csv dove sono presenti i
# link dei report e tutti gli issue code possibili
my_data = genfromtxt('/path/esempio/lobbyProva.csv', delimiter=',',
dtype=None, encoding=None)

# altezza matrix --> ci sono i link dei report
height = my_data.shape[0]
# larghezza --> ci sono tutti i code
weight = my_data.shape[1]
```

```
opener = urllib.request.build_opener()
opener.addheaders = [('User-agent', 'Mozilla/5.0')]

for x in range(height):
    if x > 0 :
        link = str(my_data[x][0])
        # il testo presente nell'URL viene salvato nella variabile
        textLink = BeautifulSoup(opener.open(link).read())

        for y in range(weight):
            if y > 1 :
                cod = str("15. General issue area code " + my_data[0][y])
                if cod in textLink.text:
                    # se il codice è presente nel report salvo 1 nella matrice
                    my_data[x][y] = "1"
                else:
                    # altrimenti 0
                    my_data[x][y] = "0"

# salvo la matrice "piena" in un nuovo csv
savetxt("/path/esempio/lobbyRisultatoProva.csv", my_data, delimiter=',',
header='string',comments='', fmt='%s')
```

Lo script restituisce una tabella di risultato contenete valori o 0 o 1, come ad esempio:

		ACC	ADV	. . .	WEL
https://soprweb.senate.gov/index...	80000	1	0	. . .	0
https://soprweb.senate.gov/index...	100000	0	0	. . .	0
https://soprweb.senate.gov/index...	90000	0	1	. . .	0
. . .	. . .	. . .	. . .	. . .	. . .

Tabella 2.2: Esempio tabella di risultato.

Lo script Python salva nella variabile *textLink* tutto il testo di ogni report e controlla la presenza di ogni General Issue Area Code all'interno di *textLink*. Se il codice, è presente la matrice sarà aggiornata con un "1" se invece non è presente verrà aggiornata con uno "0".

N.B. La ricerca avviene controllando che in *textLink.text* sia presente la seguente stringa:

```
str("15. General issue area code " + my_data[0][y])
```

(dove *my_ data[0][y]* è il General Issue Area Code); Questo perchè in sezioni del report, non rilevanti ai fini di questa ricerca, possono essere citati i codici. Quindi se il controllo fosse fatto solo con:

```
str(my_data[0][y])
```

avremmo dei risultati falsati.

2.2.3 Conteggio delle leggi

Ogni legge che è in discussione alla Camera dei Rappresentanti (United States House of Representatives) o al Senato (United States Senate) viene identificata da un codice (esempio: H.R.3010, S.709). Per effettuare il conteggio dei codici relativi alle leggi che potrebbero trovarsi, o meno, nella sezione "16. Specific lobbying issues" dei Lobbying Report si segue un procedimento pressoché analogo a quello appena descritto. Ci sono solo due criticità che rendono questo procedimento leggermente più complesso rispetto al precedente:

- non c'è una lista definita di tutti i codici delle leggi;
- il codice non ha una collocazione precisa come il General Issue Area Code e nemmeno una codifica precisa, per esempio la legge "H.R.207" potrebbe essere scritta pure "HR207" o "H. R. 207" o "H R207" etc. con lettere minuscole o maiuscole.

Il primo problema si risolve andando nella sezione Home/Influence & Lobbying/Lobbying/Clients/Bills di opensecrets.org dove si trova la lista completa dei disegni di legge per cui è stata fatta una qualche pressione in un determinato anno, dal cliente preso in considerazione. I codici si possono copiare e sostituire ai General Issue Area Code nella fase di costruzione del CSV.

Il secondo problema viene risolto modificando leggermente lo script Python, in modo che possa gestire tutte le casistiche sopracitate.

(la parte antecedente al ciclo for è uguale allo script precedente)

```
for x in range(height):
 if my_data[x][0] != "/":
  if x > 0 :
    link = str(my_data[x][0])
    soup = BeautifulSoup(opener.open(link).read())
    c = str(soup.text)
    # uniformo il tutti i caratteri -> sono resi tutti maiuscoli
    # sono rimossi i "."
    a = c.upper().replace(".", "")
    # sono rimossi i "\n"
    a = a.replace("\n", "")
    # sono rimossi tutti i possibili doppi spazzi
    while '  ' in a:
        a = a.replace('  ', ' ')

    for y in range(weight):
        if y > 1 :
            codi = str(my_data[0][y])
            # rimossi tutti i punti e i doppi spazi anche dai codici
            cod = codi.replace(".", "")
            cod = cod.replace(" ", "")
            # a seconda dei casi creo le variabili spaz e spaz2 per gestire
            # le casistiche degli spazzi possibili spazzi
```

```
# togliere tutti gli spazzi da problemi di falsi conteggi
if cod[0] == "S":
    spaz = cod[:1] + " " + cod[1:]
    spaz2 = "una stringa che sicuramente non è presente nel report"
elif cod[0] == "H" and len(cod)>4 and cod[3] == "S":
    spaz = cod[:4] + " " + cod[4:]
    spaz2 = cod[:1] + " " +  cod[1:4] + " " + cod[4:]

elif cod[0] == "H":
    spaz = cod[:2] + " " + cod[2:]
    spaz2 = cod[:1] + " " +  cod[1] + " " + cod[2:]

if (cod in a) :
    nuova = a.partition(cod)[2]
    if (nuova[0].isdigit()) or (nuova[0].isalpha()):
         my_data[x][y] = "0"
    else :
        my_data[x][y] = "1"

elif spaz in a:
    nuova = a.partition(spaz)[2]
    if (nuova[0].isdigit()) or (nuova[0].isalpha()):
         my_data[x][y] = "0"
    else :
        my_data[x][y] = "1"

elif  spaz2 in a:
    nuova = a.partition(spaz2)[2]
    if (nuova[0].isdigit()) or (nuova[0].isalpha()):
         my_data[x][y] = "0"
    else :
```

```
                my_data[x][y] = "1"

        else:
            my_data[x][y] = "0"
```

Come il precedente questo script restituisce una tabella di 0 e 1.

2.3 Elaborazione finale

A partire dai CSV ottenuti tramite un foglio di calcolo elettronico si possono calcolare con facilità i dati che interessano questa analisi. Per quanto riguarda gli Issue:

- **Issue Report**: il numero di volte che è oggetto di lobbying una specifica issue in un anno (n° di report in un anno specifico che citano la specifica issue):

$$\sum(report\ che\ citano\ una\ specifica\ issue\ in\ un\ anno)$$

- **Issue Valore**: la somma del valore totale dei report che nell'anno specifico citano una specifica issue:

$$\sum(report_1 * valoreTotale_{report1} + ... + report_n * valoreTotale_{report\ n})$$

- **Issue Valore Ponderato**: la somma del valore medio dei report che nell'anno specifico citano una specifica issue:

$$\sum(report_1 * \frac{valoreTotale_{report1}}{n\ issue\ citate\ in\ report_1} + ... + report_n * \frac{valoreTotale_{report\ n}}{n\ issue\ citate\ in\ report\ _n})$$

Invece per quanto riguarda le proposte di legge:

- **Bill Report**: il numero di volte che è oggetto di lobbying una specifica legge in un anno (n° di report in un anno specifico che citano la specifica legge):

$$\sum(report\ che\ citano\ una\ specifica\ legge\ in\ un\ anno)$$

- **Bill Valore**: La somma del valore totale dei report che nell'anno specifico citano una specifica legge:

$$\sum(report_1 * valoreTotale_{report1} + ... + report_n * valoreTotale_{report\ n})$$

- **Bill Valore Ponderato**: La somma del valore medio dei report che nell'anno specifico citano una specifica legge:

$$\sum(report_1 * \frac{valoreTotale_{report1}}{n\ leggi\ citate\ in\ report_1} + ... + report_n * \frac{valoreTotale_{report\ n}}{n\ leggi\ citate\ in\ report\ _n})$$

- **Bill valore ponderato 2**: si tratta di una variazione di Bill Valore Ponderato, ottenuto utilizzando come fattore di ponderazione del peso del report, il numero delle issue, invece che il numero delle leggi. Si ottiene come:

$$\sum(report_1 * \frac{valoreTotale_{report1}}{n\ issue\ citate\ in\ report_1} + ... + report_n * \frac{valoreTotale_{report\ n}}{n\ issue\ citate\ in\ report\ _n})$$

2.3.1 Elaborazione specifica per le leggi

Importante è considerare, nell'analisi delle leggi, come negli U.S.A. ogni due anni venga rieletta metà dei parlamentari, il Congresso quindi si ricostituisce e le sigle delle leggi cambiano. Per esempio lo stesso codice legge H.R.3010 è relativo a delle leggi diverse: 115th Congress (2017-2018): Promoting Good Cyber Hygiene Act; 116th Congress (2019-2020): Honoring All Veterans Act. Quindi nell'elaborazione finale dei dati, relativi alle leggi, bisogna necessariamente considerare il conteggio di due anni in due anni.

Capitolo 3

Text Analysis

Una valida fonte di informazioni sulle aziende, in particolare sulle loro azioni, utili per carpire la loro strategia, è data dai testi che ne parlano: trascrizioni di interviste o dichiarazioni, comunicati stampa, articoli, post, tweet. Una volta ottenuto un cospicuo numero di testi è possibile cercare di estrarne delle informazioni utili, utilizzando degli algoritmi di text mining.

"Il text mining è una tecnica che utilizza l'elaborazione del linguaggio naturale per trasformare il testo libero, non strutturato, di documenti/database, in dati strutturati e normalizzati."[13]

3.1 Raccolta dei dati

I testi che vanno a formare il corpus da analizzare tramite text mining possono provenire da diverse fonti:

- Testi rilasciati direttamente dal soggetto analizzato, come ad esempio comunicati stampa oppure trascrizioni di discorsi fatti da soggetti rilevanti (esempio: membri del CDA);
- Articoli di cronaca o analisi;
- Testi tratti da social, come ad esempio tweet.

Un'importante accortezza che bisogna osservare riguarda la lingua dei testi: tutti i testi che vanno a formare il corpus devono essere scritti nella stessa lingua. Questo perché gli algoritmi riconoscono le parole (lo specifico argomento verrà approfondito in seguito).

3.1.1 Twitter

Un'interessante fonte di dati testuali è Twitter. Attraverso le API ufficiali si possono scaricare i tweet filtrandoli secondo diversi parametri.

Per scaricare i tweet per prima cosa si deve accedere o creare il proprio Developer Account Twitter al sito https://developer.twitter.com/en. Una volta creato il proprio "Bearer Token" seguendo la documentazione fornita da Twitter (https://developer.twitter.com/en/docs/authentication/oauth-2-0/bearer-tokens) sarà possibile scaricare i tweet attraverso il termiale del proprio pc.

Una API che permette di scaricare dati significativi è: **user_timeline** (vedi Figura 3.1).

```
❯ curl -H 'Authorization: Bearer AAAAAAAAAAAAAAAAAAAAAAUiIAEAAAAAxjZs2QF%2ByyyOmTs9
                                                                    ' 'https://api.twit
ter.com/1.1/statuses/user_timeline.json?user_id=93957809&count=200' > ~/Scrivania/p
rova.json
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100  712k  100  712k    0     0  1483k      0 --:--:-- --:--:-- --:--:-- 1480k
```

Figura 3.1: Esempio di download tweet utilizzando l'API "user_timeline"

L'API *user_timeline* permette di scaricare un massimo di 200 tweet (*count=200*) pubblicati da un dato utente del quale di deve indicare l'*user_id*. L'*user_id* si può ricavare attraverso diversi siti che offrono questo servizio, come ad esempio: http://gettwitterid.com.

Un'altra API che permette di scaricare tweet interessanti è *search fullarchive* , essa permette di scrivere una query che va a filtrare tutti i tweet presenti sulla piattaforma (vedi Figura 3.2). La documentazione per questa API la si può trovare al seguente indirizzo: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/quick-start/premium-full-archive.

```
❯ curl --request POST \
  --url https://api.twitter.com/1.1/tweets/search/fullarchive/dev.json \
  --header 'authorization: Bearer AAAAAAAAAAAAAAAAAAAAAAUiIAEAAAAAxjZs2QF%2ByyyOmTs9[redacted]' \
  --header 'content-type: application/json' \
  --data '{
                "query":"(#alphabet OR #ericschmidt OR #SundarPichai)(#immigration OR
#ImmigrationMatters)",
                "maxResults": "100",
                "fromDate":"201801010000",
                "toDate":"201812312359"
                }' > ~/Scrivania/tesi/tweet/prova.json

  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 21303  100 21054  100   249  47312    559 --:--:-- --:--:-- --:--:-- 47764
```

Figura 3.2: Esempio di download tweet facendo una query

In questo caso i parametri da inserire sono molteplici:

- "**query**" permette di scrivere una vera e propria query inserendo dei parametri di ricerca. Nel caso esposto nella Figura 3.2 viene richiesta la presenza di alcuni hashtag all'interno dei tweet perché possano essere scaricati, in particolare deve esserci un hashtag fra #alphabet, #ericschmidt e #SundarPichai assieme a un hashtag fra #immigration e #ImmigrationMatters.
- "**maxResults**" il numero massimo di tweet da scaricare, la versione gratuita del Developper Account di Twitter permette di scaricare al massimo 100 tweet per richiesta.
- "**fromDate**" e "toDate" permettono di inserire un range temporale.

La documentazione per questa API la si può trovare al seguente indirizzo: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/overview/premium. I tweet saranno scaricati sotto forma di file Json, per estrarre il solo testo si può utilizzare uno script come questo (bisognerà adattarlo leggermente rispetto al file Json che restituisce l'API):

```
#!/usr/bin/python3
# -*- coding: utf-8 -*-
import json
```

```
# Apertura JSON file
f = open('/home/prova.json',)

data = json.load(f)
fileTesto = open("/home/tweet.txt", "w")

for i in range(len(data)):
    print(data[i]['text'],"\n")
    stringa_iniziale = data[i]['text']+'\n\n'
    fileTesto.write(stringa_iniziale)
    if 'retweeted_status' in data[i]:
        print('\t',data[i]['retweeted_status']['text'],'\n')
        stringa = '\t\t'+data[i]['retweeted_status']['text']+'\n\n'
        fileTesto.write(stringa)
fileTesto.close()

f.close()
```

A questo punto si avrà un file di testo utilizzabile per il text minig.

3.2 Topic Modeling - Algoritmo LDA

Il Topic Modeling è un metodo di data mining utilizzato per comprendere e classificare grandi quantità di dati testuali. "Un topic model è un tipo di modello statistico per scoprire gli "argomenti" (topic) astratti che si verificano in una raccolta di documenti."[14] In generale, un modello è una rappresentazione di un oggetto che cattura ciò che è importante di quell'oggetto in modo da renderlo più facilmente utilizzabile. Ad esempio, un modello architettonico ridimensiona un edificio in modo che possa essere più facilmente osservato senza sacrificare le proporzioni e il dettaglio dell'originale. Un Topic Model è simile; modella un corpus di documenti, non ridimensionandoli, ma generando argomenti rappresentativi del contenuto dei documenti. In particolare, un Topic Model

è un modello probabilistico utilizzato per scoprire argomenti o strutture latenti in una raccolta di documenti.[15] Il presupposto di fondo è che ogni dato documento contenga al suo interno argomenti latenti. Questi argomenti sono latenti perché non segnalati esplicitamente dall'autore.[16] In particolare nella ricerca sarà approfondito l'algoritmo LDA, che è stato utilizzato nell'analisi.

Latent Dirichlet Allocation (LDA) è un topic model che genera argomenti basati sulla frequenza di parole all'interno di una serie di documenti. LDA è particolarmente utile per trovare un mix ragionevolmente accurato di argomenti all'interno di un determinato set di documenti. [17] È un modello probabilistico generativo per raccolte di dati discreti, basato su classificatori Bayesiani.[18]

Allo scopo di applicare questo modello sui dati raccolti, esistono due principali soluzioni: il software Mallet e il pacchetto Gensim per Python. Prima di vedere nello specifico questi due possibili strumenti, è utile capire quali siano i passaggi da svolgere prima che l'algoritmo processi i dati. Questa operazione, a cui viene sottoposto il corpus (l'insieme di tutti i documenti che verranno processati) prende il nome di *preprocess*. Lo scopo del preprocess, detto anche "cleaning up", è quello di massimizzare i risultati ottenibili dall'algoritmo di topic modeling. Le principali operazioni che vengono svolte sono le seguenti:

- Uniformare tutte le lettere al *lower case* (o al *upper case*) ciò permette di non avere differenze fra maiuscole e minuscole. Questa operazione, ad esempio, permette che le parole "casa" e "Casa" le quali nel testo di partenza sarebbero state considerate come due parle distinte, vengano considerate una sola parola.

- Lemmatizzare tutte le parole del corpus. La lemmatizzazione è il processo di riduzione di una forma flessa di una parola alla sua forma canonica, detta lemma.[19] Questa operazione consente di uniformare tutte le parole che hanno lo stesso significato. Le parole "librone", "libretto" e "libraccio" verranno tutte trasformate nella parola "libro" che è il loro lemma.

- Rimozione delle *stop-words*. Le stop-words sono parole che non attribuiscono nessuna informazione aggiuntiva al testo e quindi inficerebbero il processo dell'algoritmo. Tipicamente le stop-words sono congiunzioni, avverbi, preposizioni, articoli. All'in-

terno di un corpus di articoli su FCA (Fiat Chrysler Automobiles) oltre a parole come "e", "il", "bene" etc. sono stop-words anche le parole "fiat" e "auto" in quanto è ragionevole supporre che queste parole siano presenti in tutti i documenti del corpus con una discreta frequenza, non portando quindi un reale surplus di informazioni. Le parole "fiat" e "auto" normalmente non sono considerate stop-words, ma lo diventano nel contesto di un corpus di articoli su FCA. In generale le stop-words sono tutte quelle parole che si ripetono con una frequenza molto elevata nel testo.

- Tokenizzare il testo. Significa scomporre il testo in tokens, quindi dividere il testo in singole parole. In questa fase viene anche eliminata la punteggiatura.

Vediamo un esempio di come verrebbe preprocessata la frase "Alpabet è la holding di Google.":

- Uniformare al lower case: "alphabet è la holding di google.";
- Lemmatizzare: "alphbet essere la holding di google."
- Rimozione delle stop-words: "alpahbet holding google."
- Tokenizzazione: ["alphabet", "holding", "google"]

Questo processo di uniformazione e selezione delle parole è fondamentale in quanto ogni parola sarà sostituita da numeri, molto più facili da processare dal calcolatore.

3.2.1 Mallet (MAchine Learning for LanguagE Toolkit)

Mallet è un pacchetto sviluppato in Java che può essere utilizzato per svolgere analisi statistiche, analisi tramite topic modeling, clustering ed estrazione di informazioni da documenti e altre applicazioni di machine learnig su dati testuali. Il software open surce Mallet è sviluppato dall'University of Massachusetts Amherst in particolare dal docente Andrew McCallum[20]. Tutte le informazioni necessarie per l'utilizzo di Mallet si possono trovare al sito:

http://mallet.cs.umass.edu/index.php.

Dopo aver scaricato la cartella di Mallet è possibile costruire, tramite Apache ANT (http://ant.apache.org/), l'ambiente di lavoro Mallet 2.0. Mallet non presenta un'interfaccia grafica, si dovrà quindi utilizzare tramite terminale. Bisogna da prima entrare nella cartella nella quale è stato costruito Mallet 2.0, in seguito sarà possibile eseguire specifici comandi utilizzando il seguente pattern:

```
bin/mallet [command] --option value --option value ...
```

I comandi per importare file su Mallet sono:

- `--input [path documents]`

  Permette di importare la cartella con all'interno il corpus, al posto di "path documents" si deve inserire il path della directory, possono essere inseriti anche più path. (COMANDO OBBLIGATORIO)

- `--output [nome_file.mallet]`

  Permette di nominare il file mallet di output. Il file *.mallet* sarà il file che verrà processato dall'algoritmo di topic modeling. (COMANDO OBBLIGATORIO)

- `--stoplist-file [path stopwords]`

  Consente di inserire il path di un file che contiene le stop-words da rimuovere.

- `--keep-sequence`

  Permette di preservare il documento come una sequenza di parole, piuttosto che come vettore di conteggio di parole. Per utilizzare il toolkit di topic modeling bisogna utilizzare questo comando.

Una volta ottenuto il file .mallet esso potrà essere processato dall'algoritmo di topic modeling. Il comando da utilizzare sarà:

```
./bin/mallet train-topics --input [nome_file].mallet
--num-topics [n] --num-top-words [n] --num-iterations [n]
--output-state output.gz  --output-topic-keys output-keys.txt
--output-doc-topics output-composition.txt --diagnostics-file output.xml
```

All'interno del comando si possono settare diverse variabili, come il numero di topics che dovranno essere ricercate, il numero di parole da visualizzare per ogni topic e il numero di iterazioni che il programma deve eseguire.

Mallet si è dimostrato un software di facile utilizzo.

3.2.2 Topic modeling con Python: Gensim e Spacy

Una valida alternativa a Mallet è rappresentata dalla libreria open source Gensim sviluppata in Python. La libreria permette di importare nel proprio script le funzioni necessarie ad eseguire algoritmi di topic modeling. La documentazione relativa a Gensim è disponibile al seguente indirizzo https://radimrehurek.com/gensim/index.html.

Una volta importata la libreria l'utilizzo del metodo LdaModel(...) permette l'elaborazione del corpus. Esempio:

```
ldamodel = LdaModel(corpus=corpus, num_topics=10, id2word=dictionary)
```

L'utilizzo del modello è molto semplice.

Un po' più complessa invece è la fase di preprocessing. Esistono più librerie che forniscono degli strumenti per il NLP ("L'elaborazione del linguaggio naturale, detta anche NLP (dall'inglese Natural Language Processing), è il processo di trattamento automatico mediante un calcolatore elettronico delle informazioni scritte o parlate in una lingua naturale." [21]), fra le più diffuse ci sono NLTK (https://www.nltk.org/) e spaCy (https://spacy.io/). È stato scelto di utilizzare la libreria spaCy; in rete numerosi articoli evidenziano le differenze fra le due librerie[22][23]. Una volta importata la libreria spaCy si può procedere col cleaning up dei dati. SpaCy supporta più lingue è quindi necessario importare la lingua nella quale sono scritti i testi del corpus, nel caso di questa analisi è stata importata la lingua inglese:

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

Viene poi messa a disposizione una lista di stop-words alla quale se ne possono aggiungere delle proprie. [24]

```
# importre le stop-words
from spacy.lang.en.stop_words import STOP_WORDS

# aggiungere le proprie stop-words
my_stop_words = [u'google', u'say', u'company', u'ciccio', u'mr', u'new']
for stopword in my_stop_words:
    lexeme = nlp.vocab[stopword]
    lexeme.is_stop = True
```

Di seguito uno script che mostra come implementare un modello di topic modeling su python.

```
import matplotlib.pyplot as plt
import gensim
import numpy as np
import spacy
from spacy.lang.en.stop_words import STOP_WORDS
from gensim.models import CoherenceModel, LdaModel, LsiModel, HdpModel
from gensim.models.wrappers import LdaMallet
from gensim.test.utils import common_texts
from gensim.corpora import Dictionary
import pyLDAvis.gensim
import os, gensim
from imp import reload

def clean(text):
    return str(''.join([i if ord(i) < 128 else ' ' for i in text]))

text = open("/home/martino/Scrivania/tesi/wsj.txt").read()
text = text.lower()

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

```
while '\n\n\n' in text:
    text = text.replace('\n\n\n', '\n\n')

while '@' in text:
    text = text.replace('@', '')

while '\t' in text:
    text = text.replace('\t', '')

while '  ' in text:
    text = text.replace('  ', ' ')

my_stop_words = [u'google', u'say', u'company', u'ciccio', u'mr', u'new']

for stopword in my_stop_words:
    lexeme = nlp.vocab[stopword]
    lexeme.is_stop = True

doc = nlp(clean(text))

texts, article = [], []
for w in doc:
    # se la parola non è una stop-word
    if w.text != '\n\n' and not w.is_stop and not w.is_punct and
    not w.like_num and not w.like_url and not len(w) == 1 and
    not w.lemma_ in my_stop_words and w.pos_ != 'VERB':
        # aggiungo al vettore article il lemma della parola
            article.append(w.lemma_)
    # se uguale a \n\n significa che è finito un testo
    if w.text == '\n\n':
        texts.append(article)
```

```
        article = []

dictionary = Dictionary(texts)
corpus = [dictionary.doc2bow(text) for text in texts]

ldamodel = LdaModel(corpus=corpus, num_topics=10, id2word=dictionary)
print(ldamodel.show_topics())

graph = pyLDAvis.gensim.prepare(ldamodel, corpus, dictionary)
pyLDAvis.show(graph)
```

La libreria pyLDAvis (https://github.com/bmabey/pyLDAvis) permette di visualizzare graficamente i risultati ottenuti dal modello LDA.

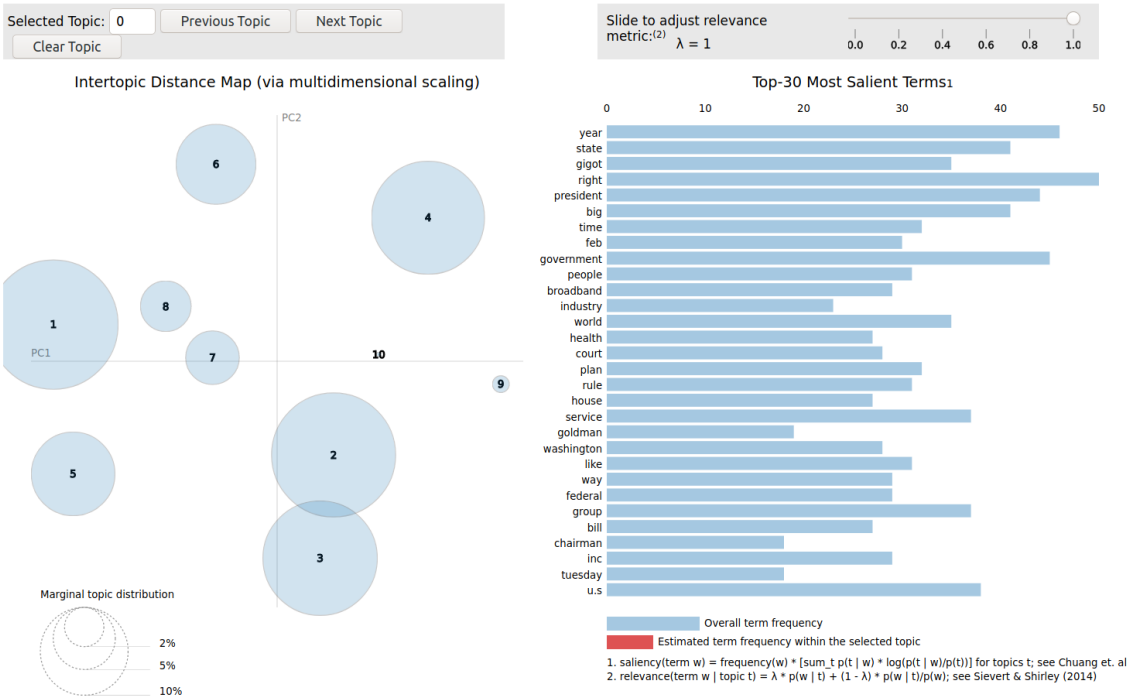


Figura 3.3: Esempio grafico pyLDAvis

3.2.3 Confronto: Mallet Python

Sia Mallet che Python (con librerie Gensim e spaCy) sono due valide alternative per modellare un corpus di testi. Fra i principali Pro di Mallet c'è sicuramente la facilità d'uso, Mallet può essere utilizzato senza avere alcuna competenza di programmazione. In Mallet tutto il *preprocessing* viene fatto automaticamente e la lista di stop-words è facile da inserire.

Al contrario se si vuole utilizzare Gensim il *corpus* andrà preparato manualmente, la libreria spaCy svolgerà il lavoro, ma sarà comunque necessario programmare. Un aspetto a favore dell'utilizzo di Python è la possibilità visualizzare graficamente i risultati.

La sostanziale differenza fra Mallet e Gensim è data dalla cleaning up dei dati. Il *preprocessing* spesso fa la differenza fra un modello che fornisce informazioni utili e un modello che restituisce dati sterili. Quindi si consiglia l'utilizzo di Gensim ad utenti esperti, mentre Mallet ad utenti meno esperti.

Durante il lavoro di questa tesi è stato utilizzato principalmente Gensim per due principali ragioni: in primo luogo per la possibilità di visualizzare i risultati graficamente, in secondo luogo per acquisire maggiore consapevolezza sul cleaning up dei dati.

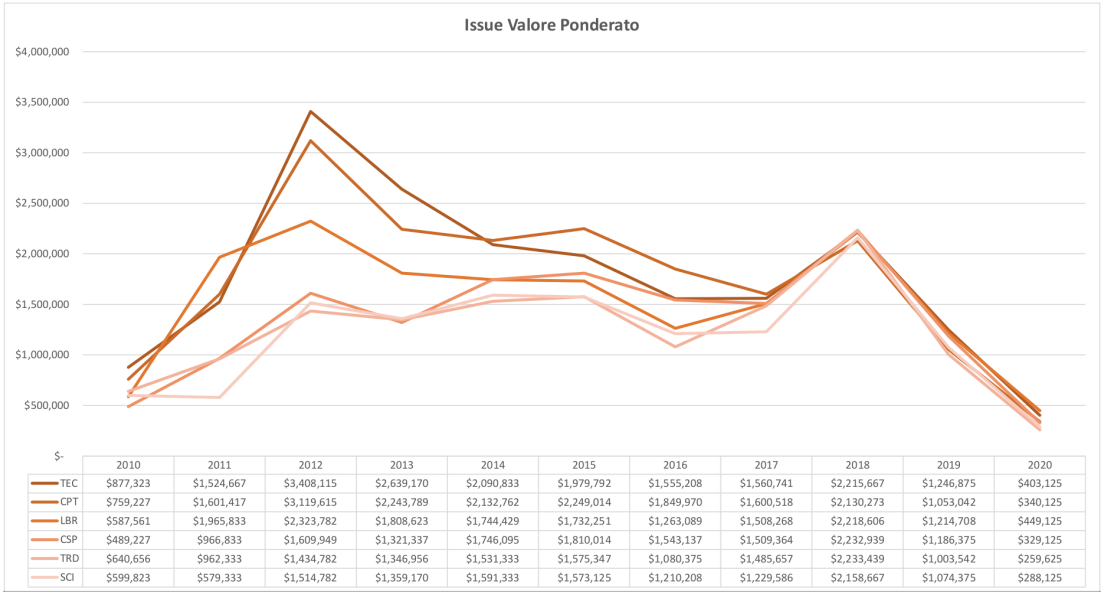
3.3 Applicazione Topic Modeling per analisi di lobbying

È stato deciso di sperimentare il topic modeling all'interno di un'analisi di lobbying allo scopo di testare la possibilità di raccogliere informazioni capaci di completare quelle derivanti dall'analisi dei lobbying report. Infatti all'interno dei lobbing report vengono raccolte informazioni sulla base dei "General Issue Area Code", i quali indicano un'area tematica, ma non danno ulteriori informazioni. Queste informazioni mancanti sono tipicamente colmate dalla presenza del codice di una legge, in assenza del codice legge non vi sono dati oggettivi che possano approfondire l'analisi. L'ipotesi fatta è che questi dati

possano essere recuperati tramite la modellazione di un corpus di testi che abbia come soggetto l'azienda analizzata.

A partire dai risultai dell'analisi dei report si andranno quindi ad individuare dei periodi dove l'attività di lobbying è più intensa per dei specifici General Issue Area Code.

Ad esempio, il seguente grafico rappresenta gli Issue valore ponderato calcolati nell'analisi dei report.



	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
TEC	$877,323	$1,524,667	$3,408,115	$2,639,170	$2,090,833	$1,979,792	$1,555,208	$1,560,741	$2,215,667	$1,246,875	$403,125
CPT	$759,227	$1,601,417	$3,119,615	$2,243,789	$2,132,762	$2,249,014	$1,849,970	$1,600,518	$2,130,273	$1,053,042	$340,125
LBR	$587,561	$1,965,833	$2,323,782	$1,808,623	$1,744,429	$1,732,251	$1,263,089	$1,508,268	$2,218,606	$1,214,708	$449,125
CSP	$489,227	$966,833	$1,609,949	$1,321,337	$1,746,095	$1,810,014	$1,543,137	$1,509,364	$2,232,939	$1,186,375	$329,125
TRD	$640,656	$962,333	$1,434,782	$1,346,956	$1,531,333	$1,575,347	$1,080,375	$1,485,657	$2,233,439	$1,003,542	$259,625
SCI	$599,823	$579,333	$1,514,782	$1,359,170	$1,591,333	$1,573,125	$1,210,208	$1,229,586	$2,158,667	$1,074,375	$288,125

Figura 3.4: Esempio di grafico Issue Valore Ponderato [dati lobbying report] [Elaborazione dell'autore]

In questo caso si può osservare come fra il 2011 e il 2013 ci sia un incremento importate di spesa nel area TEC, che sta per Telecommunications. Si andrà quindi a creare un corpus di testi che abbiano a che fare con le parole chiavi "Alphabet" o "Google", "Lobby" o "Politics" e "Telecommunications"; il corpus sarà poi modellato e se i risultati saranno significativi, verranno utilizzati nell'analisi.

Capitolo 4

Risultati

In questo capitolo verranno descritti i risultati più significativi ottenuti dall'analisi dei lobbying report e dall'analisi testuale. Sono stati analizzati in totale **934 lobbying report** consegnati al Congresso dal primo trimestre 2010 al secondo trimestre 2020 da lobbysti che hanno come cliente Google o associate. Sono stati spesi in totale **2'494'095'950 US Dollars**.

4.1 Risultati dell'analisi General Issue Area Code

In questa sezione verranno esposti i risultati ottenuti dalla raccolta ed elaborazione dei dati relativi ai Generic Issue Area Code. Le modalità di tale elaborazioni sono state descritte nel Capitolo 3. Dei settantanove codici previsti trentatre sono stati utilizzati in almeno un report.

Il grafico che segue (Figura 4.1) può dare una prima idea di quelli che sono i codici più rilevanti. Per comparare i dati relati a Issue Report, Issue Valore e Issue Valore Ponderato[1] è stato necessario calcolare un valore percentuale, infatti in termini assoluti i valori citati sono molto differenti. Importante sottolineare come tale valore percentuale non abbaia una particolare valenza al di là del confronto sopracitato.

[1] Vedi Capitolo 2.3

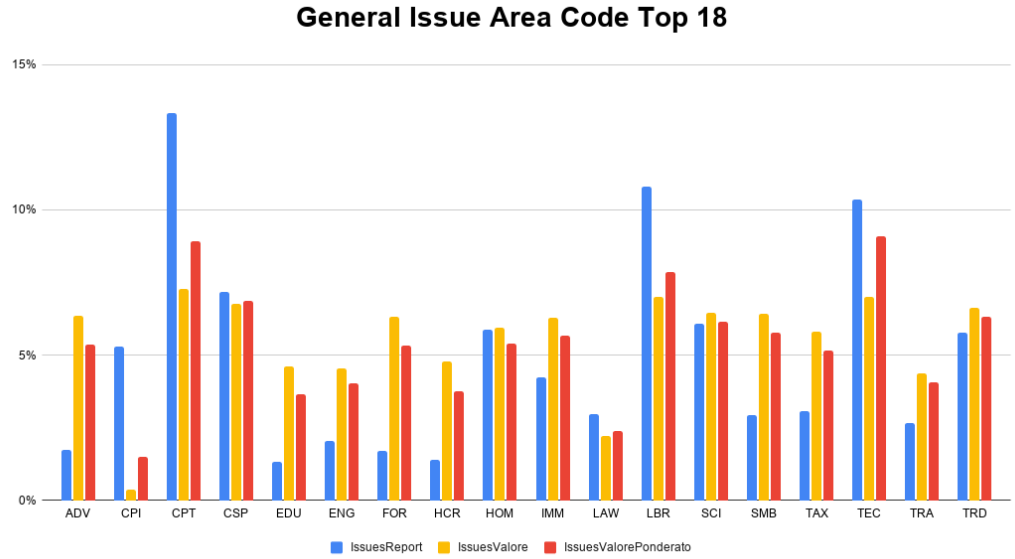


Figura 4.1: Confronto fra i diciotto codici più utilizzati. [dati lobbying report] [Elaborazione dell'autore]

4.1.1 Issue Report

I risultati relativi al conteggio degli Issue Report evidenzia come in particolare tre Issue Codes sono stati utilizzati in più del 35% dei report consegnati dal 2010 al 2020. Questi codici sono CPT (Copyright/Patent/Trademark), LBR(Labor Issues/Antitrust/-Workplace), TEC(Telecomunications).

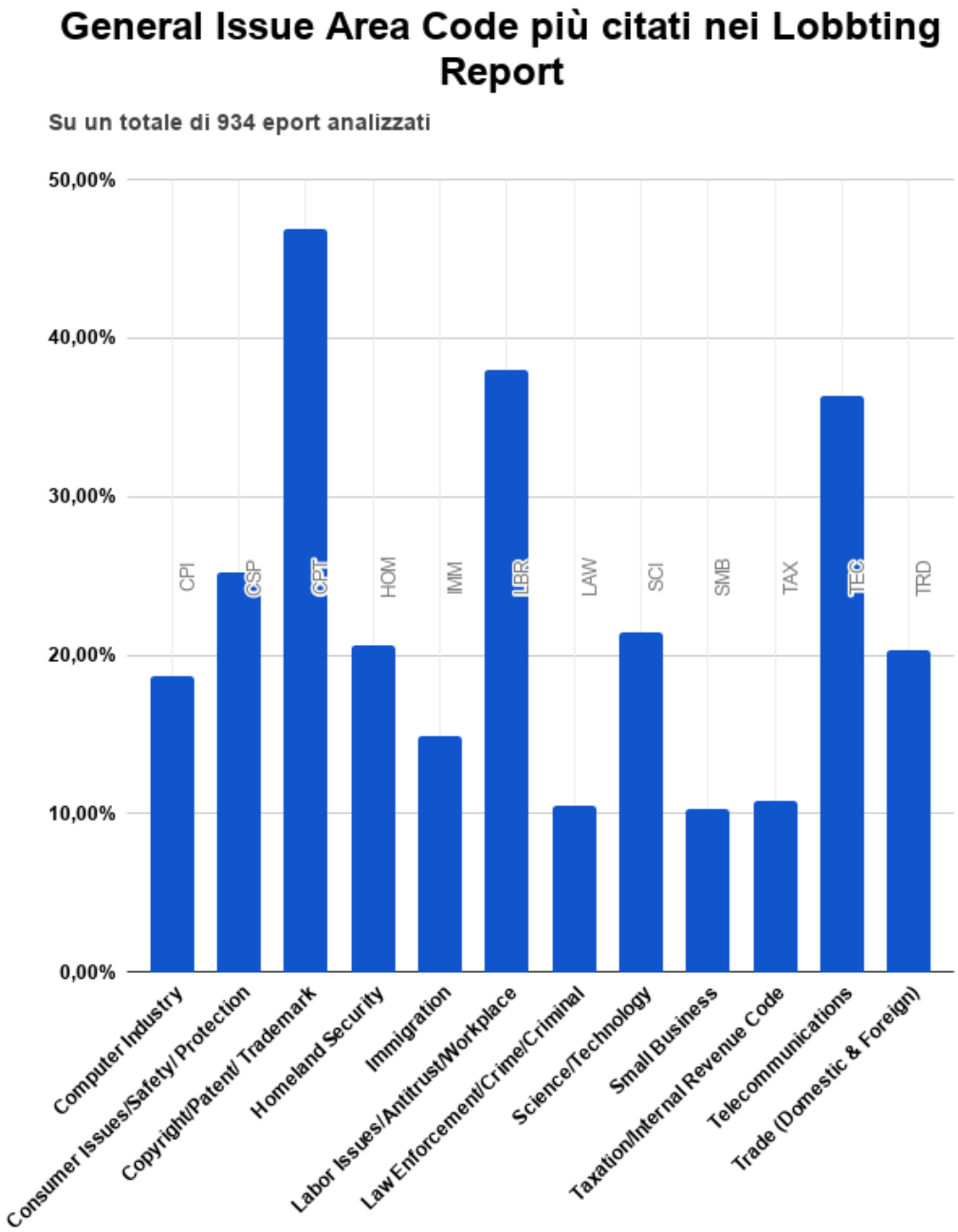


Figura 4.2: Grafico: Issue Area Code più citati nei lobbing reports. [dati lobbying report] [Elaborazione dell'autore]

General Issue Area Code più citati nei Lobbting Report			
Copyright/Patent/ Trademark	CPT	**438**	46,90%
Labor Issues/Antitrust/Workplace	LBR	**355**	38,01%
Telecommunications	TEC	**340**	36,40%
Consumer Issues/Safety/ Protection	CSP	**236**	25,27%
Science/Technology	SCI	**200**	21,41%
Homeland Security	HOM	**193**	20,66%
Trade (Domestic & Foreign)	TRD	**190**	20,34%
Computer Industry	CPI	**174**	18,63%
Immigration	IMM	**139**	14,88%
Taxation/Internal Revenue Code	TAX	**101**	10,81%
Law Enforcement/Crime/Criminal Justice	LAW	**98**	10,49%
Small Business	SMB	**96**	10,28%

Figura 4.3: Tabella: Issue Area Code più citati nei lobbing reports. [dati lobbying report] [Elaborazione dell'autore]

4.1.2 Issue Valore Ponderato

La stima "Issue Valore Ponderato" permette di approssimare la distribuzione della spesa di Lobbying fra le varie aree tematiche. Il grafico che segue permette di visualizzare graficamente tale distribuzione. I dati sono relativi alla sommatoria delle spese dal 2010 al 2020.

Si noti come ai primi posti si trovino gli stessi codici della sezione precedente.

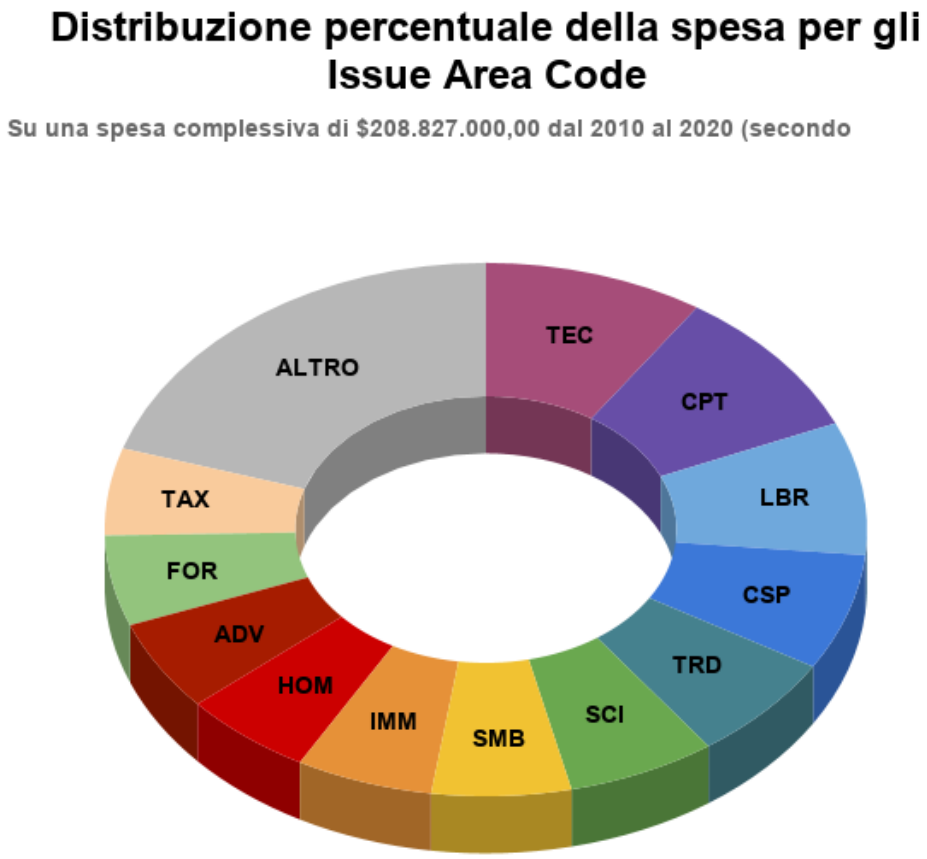


Figura 4.4: Tabella: Investimento stimato per ogni Issue Area Code. [dati lobbying report] [Elaborazione dell'autore]

Spesa stimata per ogni General Issue Area Code			
Telecommunications	TEC	$19.501.515,00	9,34%
Copyright/Patent/ Trademark	CPT	$19.079.752,00	9,14%
Labor Issues/Antitrust/Workplace	LBR	$16.816.274,00	8,05%
Consumer Issues/Safety/ Protection	CSP	$14.744.395,00	7,06%
Trade (Domestic & Foreign)	TRD	$13.554.046,00	6,49%
Science/Technology	SCI	$13.178.527,00	6,31%
Small Business	SMB	$12.336.765,00	5,91%
Immigration	IMM	$12.139.739,00	5,81%
Homeland Security	HOM	$11.568.859,00	5,54%
Advertising	ADV	$11.501.694,00	5,51%
Foreign Relations	FOR	$11.447.265,00	5,48%
Taxation/Internal Revenue Code	TAX	$11.036.649,00	5,29%
Altro	ALTRO	$41.921.520,00	20,07%

Figura 4.5: Tabella: Investimento stimato per ogni Issue Area Code. [dati lobbying report] [Elaborazione dell'autore]

4.1.3 Outsiders

Quasi tutti i codici rappresentati nella grafico e nella tabella rappresentano delle aree tematiche che sono strettamente connesse al *core-buisness* di Google. Le eccezioni sono HOM (Homeland Security), IMM (Immigration) e SMB (Small Business). Al fine di ottenere maggiori informazioni sulle implicazioni di Google in queste aree, apparentemente poco legate alle proprie attività principali, è utile osservare lo storico dei valori Isssue Report e Issue Valore Ponderato in relazione al totale.

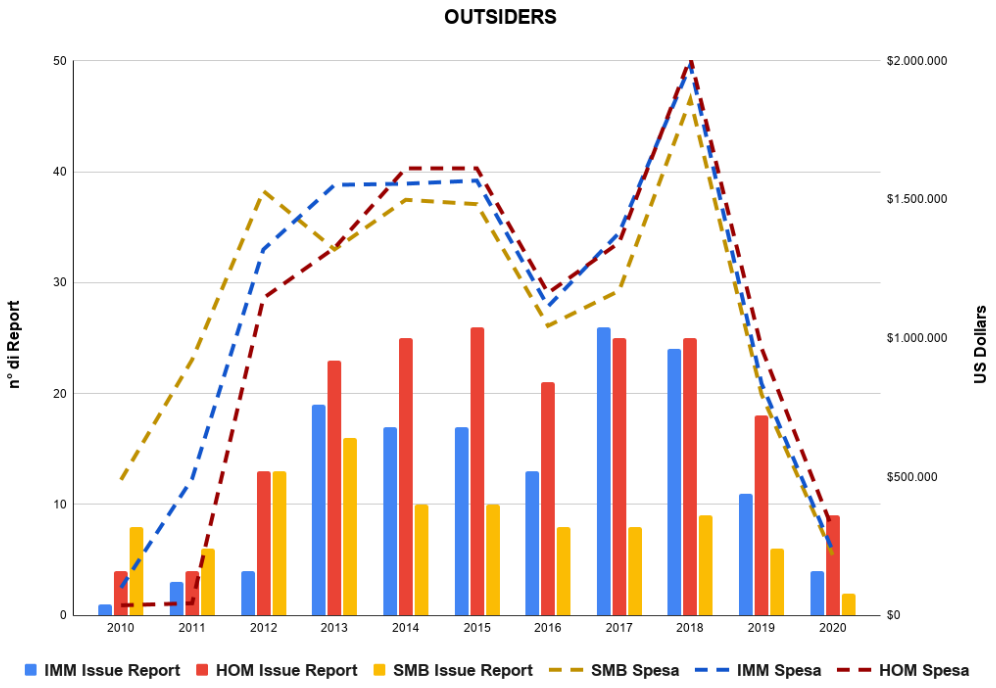


Figura 4.6: Issue Report e Issue Valore Ponderato degli issues outsider. [dati lobbying report] [Elaborazione dell'autore]

Nel grafico sopra riportato, l'istogramma che fa riferimento all'asse di sinistra rappresenta la distribuzione dei report in cui sono presenti i codici analizzati nel corso del tempo, mentre le linee tratteggiate vanno lette sull'asse di destra e indicano la spesa stimata per ogni codice durante il decennio.

4.2 Risultati dell'analisi Bill images

Per quanto riguarda l'elaborazione dei dati relativi alle leggi sono stati calcolati i seguenti valori: Bill Report, Bill Valore, Bill valore ponderato e Bill valore ponderato 2, calcolati come descritto nel capitolo 2.3.

L'analisi è stata sviluppata nel seguente modo: da prima sono state individuate le leggi più citate nei vari lobbying report, con un attenzione particolare al cambio di codice col passare del tempo, per poi andare ad osservare i relativi costi calcolati con Bill valore ponderato e Bill valore ponderato 2.

Le leggi più citate in assoluto, presenti in più di cinquanta report, sono state:

- **Electronic Communications Privacy (Act Amendments)**

  L'eletronic Coomunicatios Privacy Act, emanato nel 1986 dal Congresso degli Stati Uniti, fu teso ad estendere le restrizioni riguardanti le intercettazioni telefoniche governative per includere le trasmissioni di dati elettronici tramite computer, aggiungendo nuove disposizioni che vietano l'accesso alle comunicazioni elettroniche memorizzate e le cosiddette pen trap, disposizioni che consentono il tracciamento delle comunicazioni telefoniche [25].

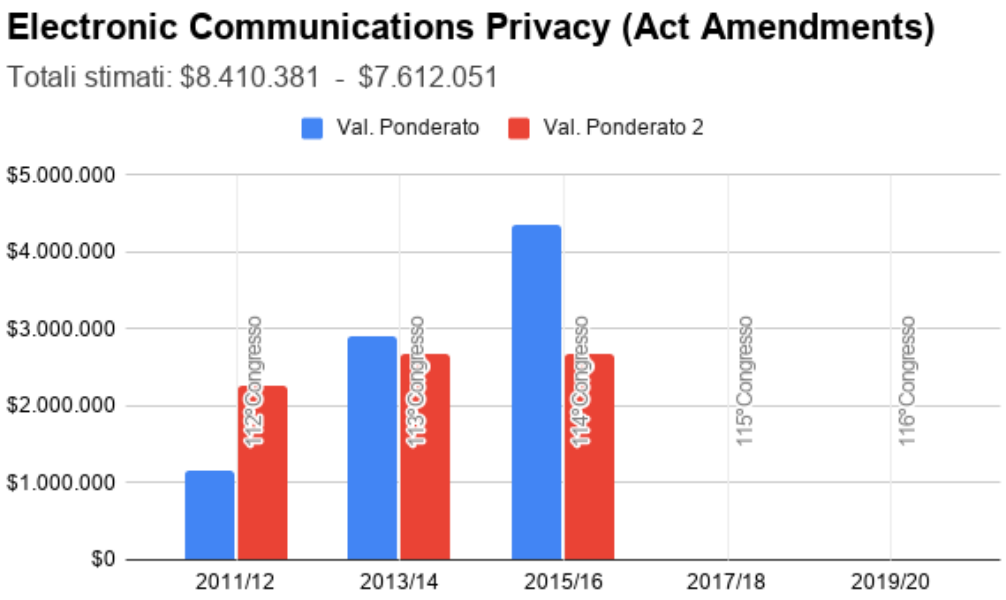


Figura 4.7: Spesa per: Electronic Communications Privacy [dati lobbying report] [Elaborazione dell'autore]

- **Innovation Act**

  L'innovation Act è un disegno di legge che modifica le norme e i regolamenti precedenti riguardanti le cause per violazione dei brevetti, questo nel tentativo di ridurre sensibilmente le cause riguardanti la violazione di brevetti [25].

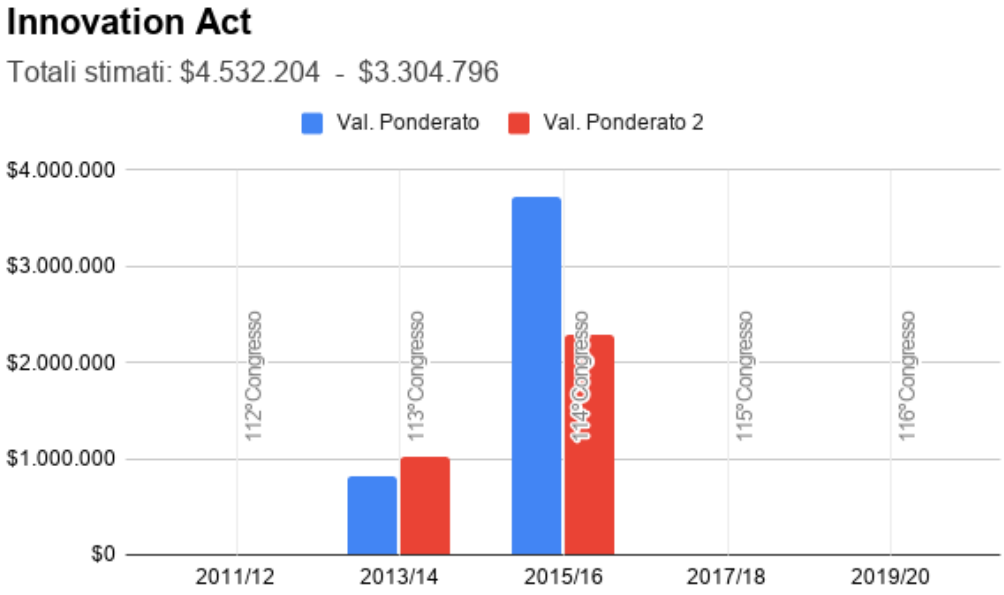


Figura 4.8: Spesa per: Innovation Act [dati lobbying report] [Elaborazione dell'autore]

- **Email Privacy Act**

  L'email Privacy Act, anch'esso disegno di legge, fu progettato per aggiornare e riformare la legge già esistente inerente alle comunicazioni online; in particolar modo l'Eletronic Communicatios Privacy Act del 1986. La legislazione, in breve, richiederebbe alle autorità di Giustizia degli Stati Uniti di ottenere un mandato di perquisizione per poter accedere a e-mail, dati nel cloud storage e altre comunicazioni [25].

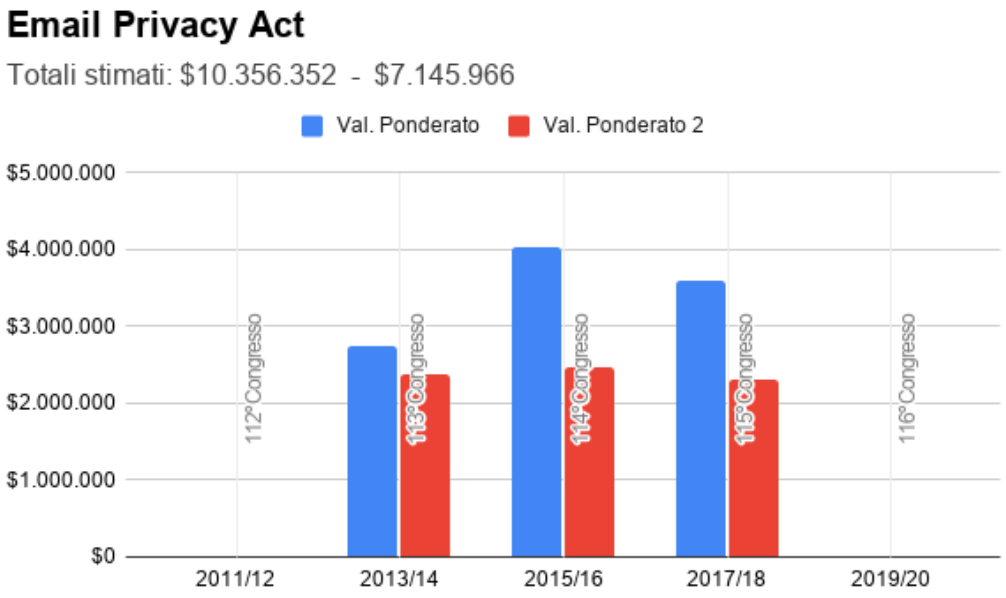


Figura 4.9: Spesa per: Email Privacy Act[dati lobbying report] [Elaborazione dell'autore]

- **Patent Quality Improvement Act**

  Questo disegno di legge va a modificare il Leahy-Smith America Invents Act per rimuoverne la clausola che prevedeva una scadenza di otto anni per un particolare tipo di brevetto ("Covered business method patent"). Inoltre si prefigge di far rientrare nella categoria "Covered business method patent", fra gli altri, metodi o apparecchiature per eseguire l'elaborazione dati [25].

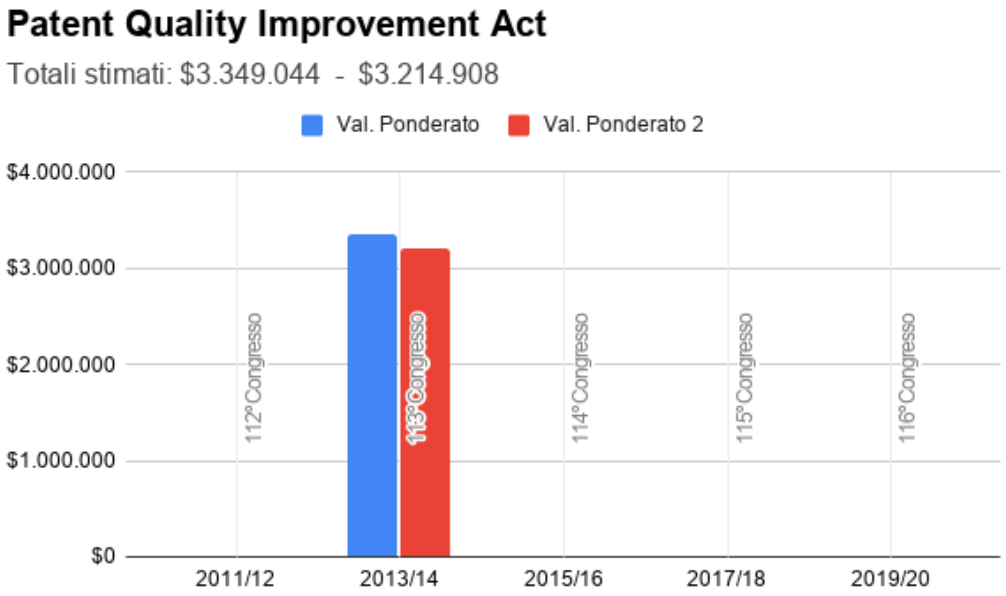


Figura 4.10: Spesa per: Patent Quality Improvement Act [dati lobbying report] [Elaborazione dell'autore]

- **Cyber Intelligence Sharing and Protection Act**

  Questo disegno di Legge ha come soggetto la condivisione di informazioni sul traffico internet tra Governo degli Stati Uniti e società tecnologiche e manifatturiere. L'obiettivo del disegno di legge era quello di aiutare il Governo ad indagare sulle minacce informatiche e garantire la sicurezza delle reti contro possibili attacchi informatici [25].

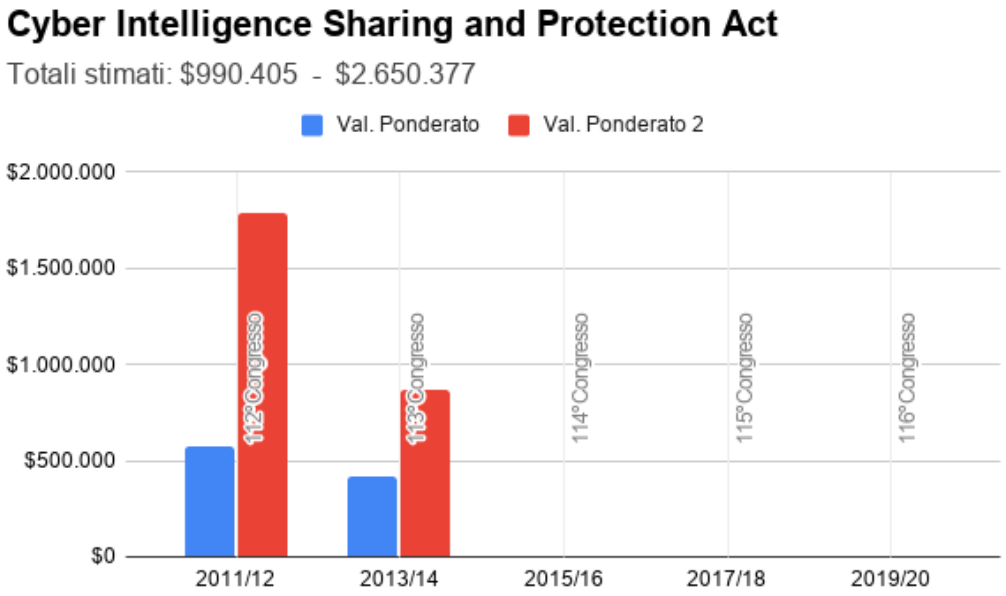


Figura 4.11: Spesa per: Cyber Intelligence Sharing and Protection Act [dati lobbying report] [Elaborazione dell'autore]

4.3 Risultati Topic Modeling

È stato svolto un approfondito lavoro di modellizzazione su diversi corpus (alcuni formati da tweet e altri formati da articoli di stampa), senza però aver ottenuto i risultati attesi. Ci si attendeva infatti che l'analisi testuale avrebbe permesso di completare le informazioni pregresse, l'esito non ha rispettato le attese nella misura preventivata. Ciò è probabilmente dovuto alla qualità del corpus di partenza. Gli strumenti e il tempo a disposizione, non hanno permesso di filtrare sufficientemente i testi, portando a modelli con molto rumore che non hanno permesso di entrare particolarmente nei dettagli.

In particolare questo approccio è stato utilizzato per cercare maggiori informazioni relative agli investimenti di Alphabet Inc. per quei settori che erano stati definiti Outsiders nel paragrafo 4.1.3.

Con riferimento al grafico in figura 4.6 è stato definito l'intervallo di massima spesa per il codice relativo all'immigrazione e sono stati creati due corpus di testi, uno formato da tweet e uno formato da articoli di stampa. Entrambi, tweet e articoli, sono stati filtrati attraverso parole chiave come "Alphabet Inc." e "immigration", considerando le date di pubblicazione dal 2012 al 2015 e dal 2017 al 2019 (picchi di spesa).

Di seguito i risultati ottenuti.

(*Tutti gli output che saranno presentati nelle prossime pagine sono stati ottenuti dall'autore.*)

Risultati dal Corpus di Tweet:

(ad alcuni topic, quelli meno generici, è stato dato un titolo)

1. (0, '0.034*"anti-#immigration" + 0.031*"dollar" + 0.028*"election" + 0.025*"gstephanopoulos" + 0.025*"t13" + 0.025*"savannahguthrie" + 0.025*"gma" + 0.025*"gmaday" + 0.025*"cnnchile" + 0.019*"reform"');

2. **Media d'informazione** (1, '0.052*"gma" + 0.030*"foxnews" + 0.030*"maszka" + 0.030*"deadly" + 0.030*"wds" + 0.030*"cnn" + 0.030*"illegal" + 0.030*"irs" + 0.030*"cnbc" + 0.029*"wsj"');

3. **migranti alla ricerca di lavoro** "ilo" è l'organizzazione internazionale del lavoro (2, '0.019*"refugee" + 0.019*"crisis" + 0.016*"ilo" + 0.014*"donaldtrump" + 0.014*"tool" + 0.014*"fundraising" + 0.013*"president" + 0.011*"decent" + 0.011*"worker" + 0.011*"economictimes"');

4. (3, '0.041*"refugee" + 0.038*"reform" + 0.028*"fraud" + 0.024*"maryherman2" + 0.021*"realdonaldtrump" + 0.021*"daca" + 0.017*"amp" + 0.015*"donation" + 0.015*"response" + 0.013*"potus"');

5. (4, '0.020*"crisis" + 0.019*"amp" + 0.016*"fundraising" + 0.016*"tool" + 0.016*"economictimes" + 0.012*"daca" + 0.009*"dem" + 0.008*"refugee" + 0.008*"trump" + 0.008*"maga"');

6. (5, '0.016*"crisis" + 0.012*"refugee" + 0.012*"law" + 0.012*"ilo" + 0.012*"fundraising" + 0.012*"tool" + 0.008*"today" + 0.008*"ceo" + 0.008*"german" + 0.008*"decent"');

7. (6, '0.035*"canada" + 0.023*"search" + 0.012*"dollar" + 0.012*"brexit" + 0.012*"nofear" + 0.012*"thank" + 0.006*"amp" + 0.006*"armored" + 0.006*"medium" + 0.006*"130th"');

8. (7, '0.023*"pour" + 0.020*"soutenir" + 0.020*"dons" + 0.020*"appel" + 0.020*"migratoire" + 0.020*"aux" + 0.020*"lance" + 0.020*"línvasion" + 0.020*"quand" + 0.017*"fabricerobert"');

9. (8, '0.018*"crisis" + 0.014*"reform" + 0.014*"trend" + 0.014*"news" + 0.014*"tech" + 0.009*"constitution" + 0.009*"online" + 0.009*"scammer" + 0.009*"today" + 0.009*"gunlaws"');

10. **Sundar Pichai** è il CEO di Alphabet ed ha origini indiane. La "fwd_us" è un grupod i lobbyoing a favore dell'immigrazione.

(9, ’0.040*"ceo" + 0.023*"india" + 0.023*"today" + 0.023*"launch" + 0.023*"congrats" + 0.023*"sundarpichai" + 0.023*"chennai" + 0.023*"exec" + 0.021*"fwd_us" + 0.021*"immigr"’)

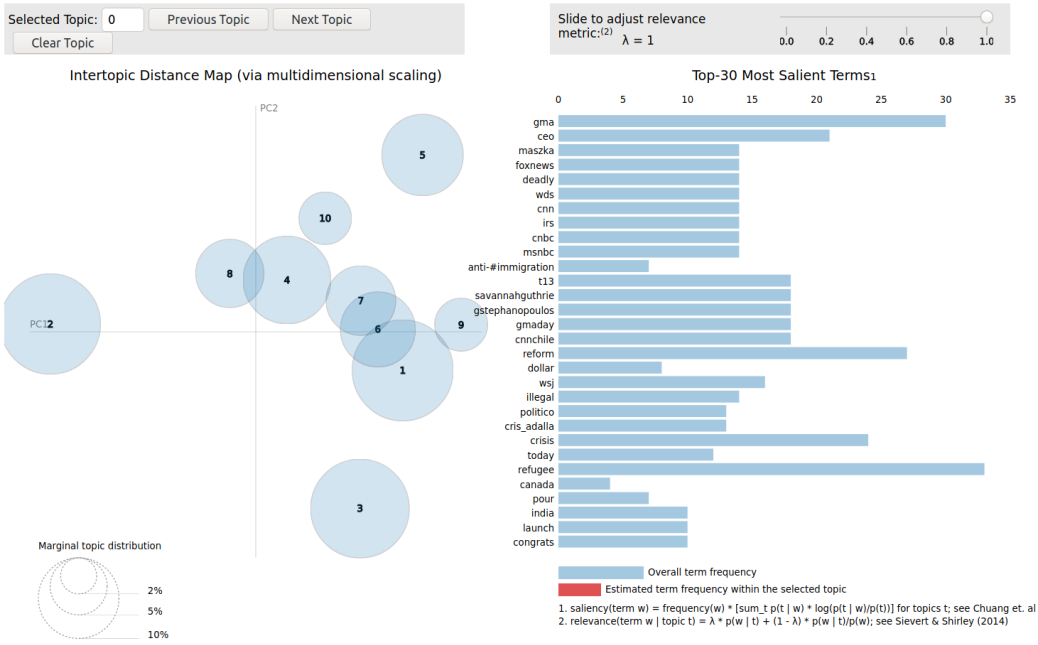


Figura 4.12: Visualizzazione grafica del topic model sul corpus ImmigrationTweet

Corpus : **Articoli di stampa, legati a Google e Immigration**
160 articoli circa Output:

1. (0, ’0.004*"account" + 0.004*"employee" + 0.004*"worker" + 0.003*"corp" + 0.003*"ad" + 0.003*"russian" + 0.003*"visa" + 0.003*"apple" + 0.003*"government"’),

2. (1, ’0.006*"startup" + 0.004*"program" + 0.003*"news" + 0.003*"user" + 0.003*"government" + 0.003*"internet" + 0.003*"plan" + 0.003*"time" + 0.003*"amazon"’),

3. (2, ’0.006*"ceo" + 0.004*"chief" + 0.004*"big" + 0.004*"order" + 0.003*"corp" + 0.003*"service" + 0.003*"policy" + 0.003*"datum" + 0.003*"country" + 0.003*"employee"’),

4. (3, '0.005*"chief" + 0.004*"employee" + 0.003*"government" + 0.003*"administration" + 0.003*"corp" + 0.003*"tax" + 0.003*"uber" + 0.003*"microsoft" + 0.003*"giant" + 0.003*"week"'),

5. (4, '0.007*"russian" + 0.005*"ad" + 0.005*"account" + 0.004*"election" + 0.004*"twitter" + 0.004*"public" + 0.004*"time" + 0.003*"week" + 0.003*"page" + 0.003*"chief"'),

6. (5, '0.004*"corp" + 0.004*"news" + 0.004*"russian" + 0.004*"ad" + 0.003*"large" + 0.003*"customer" + 0.003*"world" + 0.003*"visa" + 0.003*"order" + 0.003*"russia"'),

7. (6, '0.006*"administration" + 0.006*"order" + 0.005*"ceo" + 0.005*"employee" + 0.005*"policy" + 0.004*"donald" + 0.003*"rule" + 0.003*"election" + 0.003*"corp" + 0.003*"microsoft"'),

8. (7, '0.005*"policy" + 0.005*"corp" + 0.004*"chief" + 0.004*"white" + 0.004*"administration" + 0.004*"house" + 0.003*"monday" + 0.003*"market" + 0.003*"election" + 0.003*"apple"'),

9. (8, '0.008*"order" + 0.005*"chief" + 0.005*"program" + 0.005*"policy" + 0.004*"visa" + 0.004*"cio" + 0.004*"big" + 0.003*"datum" + 0.003*"uber" + 0.003*"news"'),

10. (9, '0.008*"employee" + 0.005*"country" + 0.005*"chief" + 0.004*"corp" + 0.004*"news" + 0.004*"uber" + 0.004*"week" + 0.004*"datum" + 0.003*"good" + 0.003*"time"')

Si può apprezzare come tali dati abbiano una valenza in termini assoluti: troviamo infatti delle informazioni interessanti, però relativamente alla proposta tali informazioni appaiono troppo generiche.

4.3.1 Risultati Topic modeling su corpus di testo non legati a Issue Area Code specifici

È stato deciso di ricercare informazioni utili in diversi corpus che potessero essere significativi.

Corpus : **Articoli di stampa, parole chiave: Google e Lobby (primi anni del decennio in analisi)**

250 articoli circa

Output:

1. (0, '0.007*"reform" + 0.005*"government" + 0.004*"year" + 0.004*"tort" + 0.004*"obama" + 0.004*"liberal" + 0.004*"service" + 0.004*"federal" + 0.004*"way" + 0.003*"big"');

2. (1, '0.007*"right" + 0.006*"gigot" + 0.005*"fcc" + 0.005*"time" + 0.005*"obama" + 0.004*"people" + 0.004*"president" + 0.004*"agency" + 0.004*"big" + 0.004*"broadband"');

3. (2, '0.006*"state" + 0.006*"president" + 0.005*"obama" + 0.005*"government" + 0.005*"gigot" + 0.005*"year" + 0.004*"right" + 0.004*"health" + 0.004*"goldman" + 0.004*"law"');

4. (3, '0.007*"right" + 0.005*"year" + 0.005*"agency" + 0.005*"obama" + 0.005*"president" + 0.004*"week" + 0.004*"government" + 0.004*"house" + 0.004*"fcc" + 0.004*"care"');

5. (4, '0.008*"state" + 0.005*"search" + 0.005*"world" + 0.004*"u.s" + 0.004*"conference" + 0.003*"chinese" + 0.003*"year" + 0.003*"time" + 0.003*"business" + 0.003*"big"');

6. (5, '0.005*"gigot" + 0.004*"right" + 0.003*"world" + 0.003*"health" + 0.003*"reform" + 0.003*"year" + 0.003*"liberal" + 0.003*"bill" + 0.003*"big" + 0.003*"time"');

7. (6, '0.006*"fcc" + 0.005*"year" + 0.005*"president" + 0.005*"state" + 0.004*"obama" + 0.004*"broadband" + 0.004*"big" + 0.004*"government" + 0.004*"u.s" + 0.004*"industry"');

8. (7, '0.026*"fcc" + 0.012*"provider" + 0.012*"broadband" + 0.011*"service" + 0.008*"genachowski" + 0.008*"traffic" + 0.008*"decision" + 0.008*"net" + 0.008*"rule" + 0.007*"court"'):

9. (8, ’0.010*"law" + 0.007*"group" + 0.006*"user" + 0.004*"fcc" + 0.004*"inc" + 0.004*"system" + 0.004*"government" + 0.004*"antitrust" + 0.004*"year" + 0.004*"rule"’):

10. (9, ’0.014*"feb" + 0.005*"world" + 0.004*"year" + 0.004*"big" + 0.003*"time" + 0.003*"system" + 0.003*"reform" + 0.003*"camera" + 0.003*"tort" + 0.003*"economic"’).

Questo corpus può dare delle indicazioni sulle attività di lobbying di Google nel periodo di massima crescita degli investimenti. In particolare il primo topic evidenzia parole legate alla politica riformista del presidente Obama. Il secondo, settimo e ottavo topic raccolgono gruppi di parole riguardanti l’ambito delle telecomunicazioni ("fcc" è la Federal Communications Commission e "genachowski" ne è il presidente). TEC (Telecomunications) è l’issue area code con il valore ponderato stimato più alto. Nel terzo, quarto e sesto topic ricorrono parole inerenti la riforma sanitaria. In fine, dal quinto topic, emergono parole come "business" e "chinese".

Corpus : **Tweet postati da Account rilevanti (Google, Eric Schmidt, Larry Page, Joshua Marcuse, Sundar Pichai)**

Output:

1. (0, '0.014*"step" + 0.012*"account" + 0.007*"digital" + 0.006*"lot" + 0.006*"message" + 0.006*"hand" + 0.006*"hmm" + 0.005*"team" + 0.004*"comms" + 0.004*"coach"');
2. (1, '0.021*"board" + 0.021*"innovation" + 0.019*"defense" + 0.015*"public" + 0.014*"meeting" + 0.012*"upcoming" + 0.011*"virtual" + 0.011*"tuesday" + 0.011*"step" + 0.010*"coach"');
3. (2, '0.010*"tech" + 0.010*"amp" + 0.010*"dod" + 0.009*"thank" + 0.008*"today" + 0.007*"ceo" + 0.006*"great" + 0.006*"email" + 0.006*"team" + 0.006*"address"');
4. (3, '0.010*"idea" + 0.009*"innovationboard" + 0.008*"innovation" + 0.008*"technology" + 0.008*"day" + 0.007*"dib" + 0.007*"team" + 0.007*"great" + 0.006*"video" + 0.006*"amp"');
5. (4, '0.007*"amp" + 0.006*"week" + 0.005*"episode" + 0.005*"conversation" + 0.005*"business" + 0.005*"china" + 0.005*"secretary" + 0.004*"proud" + 0.004*"defense" + 0.004*"regulation"');
6. (5, '0.014*"step" + 0.013*"amp" + 0.009*"great" + 0.009*"leadership" + 0.007*"team" + 0.007*"account" + 0.006*"joshuamarcuse" + 0.006*"ceo" + 0.006*"good" + 0.005*"dodjaic"');
7. (6, '0.017*"amp" + 0.013*"security" + 0.011*"tech" + 0.010*"national" + 0.007*"big" + 0.007*"today" + 0.005*"joshuamarcuse" + 0.005*"project" + 0.005*"googleorg" + 0.005*"software"');
8. (7, '0.036*"step" + 0.019*"people" + 0.012*"world" + 0.012*"password" + 0.009*"great" + 0.008*"page" + 0.008*"schmidt" + 0.007*"conversation" + 0.007*"power" + 0.007*"mobile"');

9. (8, '0.023*"people" + 0.020*"power" + 0.018*"mintpressnews" + 0.014*"today" + 0.010*"public" + 0.008*"time" + 0.008*"message" + 0.008*"account" + 0.008*"meeting" + 0.007*"way"');

10. (9, '0.011*"schmidt" + 0.008*"amp" + 0.008*"excited" + 0.007*"team" + 0.007*"chinese" + 0.007*"jjrosenberg" + 0.007*"tech" + 0.006*"able" + 0.006*"narendramodi" + 0.006*"vikasvirodhicongress"').

La modellazione di questo corpus, costituito dai tweet degli accout privati di alcune delle persone chiave di Google, ha restituito come risultato molte parole che hanno a che fare con l'area della Sicurezza Nazionale (General Issue Area Code: HOM). Inoltre sono presenti più volte parole che si riferiscono alla Cina. (*"dib": Defense Innovation Board; "dod": Department of Defense*)

Corpus : **Tweet scaricati con argomenti chiave: lobby e Google (2010 - 2020)**

Output:

1. (0, '0.017*"today" + 0.013*"step" + 0.010*"podcast" + 0.010*"hmm" + 0.009*"coach" + 0.009*"nscai" + 0.008*"recommendation" + 0.007*"team" + 0.007*"commissioner" + 0.006*"dollar"');

2. (1, '0.010*"way" + 0.008*"great" + 0.008*"amp" + 0.006*"excited" + 0.005*"change" + 0.005*"culture" + 0.005*"idea" + 0.004*"innovation" + 0.004*"public" + 0.004*"need"'),

3. (2, '0.014*"great" + 0.007*"program" + 0.007*"today" + 0.006*"health" + 0.006*"national" + 0.006*"innovationboard" + 0.006*"leadership" + 0.006*"team" + 0.005*"amp" + 0.005*"different"'),

4. (3, '0.017*"innovation" + 0.010*"defense" + 0.010*"upcoming" + 0.009*"board" + 0.009*"public" + 0.009*"meeting" + 0.008*"tuesday" + 0.008*"virtual" + 0.007*"people" + 0.007*"message"'),

5. (4, '0.010*"defcommunity" + 0.008*"joshuamarcuse" + 0.007*"defense" + 0.007*"force" + 0.007*"u.s" + 0.007*"board" + 0.006*"nscai" + 0.006*"commissioner" + 0.005*"team" + 0.005*"principle"'),

6. (5, '0.069*"step" + 0.025*"account" + 0.017*"password" + 0.011*"page" + 0.010*"today" + 0.008*"joshuamarcuse" + 0.007*"hmm" + 0.007*"amp" + 0.007*"digital" + 0.006*"business"'),

7. (6, '0.011*"today" + 0.010*"great" + 0.009*"year" + 0.009*"conversation" + 0.006*"work" + 0.006*"people" + 0.006*"amp" + 0.005*"dib" + 0.005*"approach" + 0.005*"antonioguterres"'),

8. (7, '0.016*"public" + 0.015*"meeting" + 0.015*"innovation" + 0.013*"address" + 0.013*"email" + 0.011*"board" + 0.011*"defense" + 0.007*"language" + 0.007*"innovationboard" + 0.006*"time"'),

9. (8, ’0.011*"good" + 0.010*"dod" + 0.009*"amp" + 0.009*"team" + 0.008*"people" + 0.007*"step" + 0.007*"book" + 0.006*"great" + 0.006*"coach" + 0.006*"dollar"’),

10. (9, ’0.010*"amp" + 0.009*"innovation" + 0.009*"happy" + 0.007*"proud" + 0.006*"dod" + 0.006*"thank" + 0.006*"technology" + 0.005*"work" + 0.005*"world" + 0.005*"people"’).

Anche i risultati di questo corpus sottolineano l’argomento della Sicurezza Nazionale e quello della Sanità. (*"nscai": National Security Commission on Artificial Intelligence, "joshua marcuse": Google’s Head of Strategy Innovation, Global Public Sector*)

Corpus : **Articoli di stampa, parola chiave: Alphabet Inc. (2020)** *200 articoli circa*

Output:

1. (0, '0.014*"cent" + 0.013*"market" + 0.010*"tech" + 0.010*"year" + 0.008*"revenue" + 0.008*"new" + 0.007*"time" + 0.007*"business" + 0.007*"advertising" + 0.006*"stock"'),

2. (1, '0.014*"year" + 0.011*"tech" + 0.011*"big" + 0.010*"market" + 0.007*"time" + 0.007*"business" + 0.006*"new" + 0.006*"deal" + 0.006*"cent" + 0.005*"investor"'),

3. (2, '0.012*"cent" + 0.008*"advertising" + 0.008*"business" + 0.007*"tech" + 0.007*"market" + 0.006*"revenue" + 0.006*"year" + 0.005*"new" + 0.005*"twitter" + 0.005*"chief"'),

4. (3, '0.011*"tech" + 0.009*"cent" + 0.008*"year" + 0.008*"business" + 0.007*"market" + 0.007*"chief" + 0.007*"executive" + 0.006*"big" + 0.006*"advertising" + 0.006*"revenue"'),

5. (4, '0.015*"cent" + 0.013*"business" + 0.011*"year" + 0.008*"revenue" + 0.008*"advertising" + 0.007*"executive" + 0.007*"tech" + 0.007*"quarter" + 0.007*"time" + 0.007*"big"'),

6. (5, '0.010*"technology" + 0.008*"tech" + 0.007*"investor" + 0.007*"health" + 0.006*"datum" + 0.006*"year" + 0.006*"waymo" + 0.005*"business" + 0.005*"market" + 0.005*"right"'),

7. (6, '0.009*"datum" + 0.009*"tech" + 0.007*"year" + 0.006*"health" + 0.006*"new" + 0.005*"market" + 0.005*"technology" + 0.004*"law" + 0.004*"business" + 0.004*"investor"'),

8. (7, '0.007*"year" + 0.007*"market" + 0.005*"waymo" + 0.005*"tech" + 0.005*"big" + 0.005*"business" + 0.005*"investment" + 0.005*"executive" + 0.005*"health" + 0.005*"time"'),

9. (8, '0.009*"year" + 0.007*"new" + 0.007*"cent" + 0.006*"tech" + 0.005*"market" + 0.005*"bond" + 0.005*"executive" + 0.005*"project" + 0.004*"health" + 0.004*"time"'),

10. (9, '0.008*"year" + 0.006*"big" + 0.006*"market" + 0.006*"cent" + 0.006*"tech" + 0.006*"new" + 0.006*"business" + 0.005*"investor" + 0.004*"group" + 0.004*"right"')

Il corpus, del quale i risultati sono sopra citati, è relativo ad articoli pubblicati nel 2020 che parlano genericamente di Alphabet. Nel primo topic è presente un insieme di parole che fanno riferimento ad uno dei core business di Google: l'advertising (ADV). In generale anche negli altri topic troviamo parole relative allo stesso argomento.

Attraverso l'analisi testuale sono stati confermati alcuni risultati derivanti dalla valutazione dei lobbying report, come ad esempio l'interesse di Google nelle telecomunicazioni. In più è stato possibile ampliare le conoscenze sull'argomento: Sicurezza Nazionale in relazione a Google, che era stato definito, nel paragrafo 4.1.3, come un *outsider*, in quanto idealmente non legato al core business di Alphabet.

Capitolo 5

Conclusioni

Dall'analisi dei lobbying report appare chiaro come Google abbia investito ingenti somme allo scopo di consolidare il proprio vantaggio competitivo nei confronti di possibili nuovi entranti nei settori nei quali opera. Uno dei principali dati a sostegno di questa tesi è la rilevanza degli investimenti nell'area dei brevetti e del copyright. In tal senso si può supporre che le spese volte a tematiche che hanno a che fare con "small business" siano volte a creare un sistema dove Google abbia maggiori possibilità di "annettere" al proprio interno piccole start-up, per poterne utilizzare le tecnologie innovative e impedire che possano diventare in futuro temibili concorrenti. Un'altra voce che avvalora questa ipotesi è contenuta in un articolo di Naomi Klein; nel quale la giornalista sottolinea come Alphabet e le altre Big Tech facciano pressioni sul governo U.S.A. per ricevere fondi destinati ai propri settori di ricerca e sviluppo. Lo Stato dovrebbe contribuire alla ricerca di queste società, dal momento che rappresentano l'avanguardia tecnologica del Paese, al fine di mantenere questa supremazia, anche per questioni di sicurezza nazionale e impedire che i principali competitor, soprattutto cinesi, le possano sopravanzare[26]. Tale articolo conferma quanto è emerso nell'analisi testuale tramite topic modeling, dove molti cluster di parole hanno evidenziato connessioni fra: sicurezza nazionale, difesa, innovazione e Cina.

Google, quindi, utilizza il tema della Sicurezza Nazionale degli Stati Uniti per accrescere enormemente il proprio vantaggio competitivo nei confronti di nuove entranti: è arduo pensare di competere con un'azienda che oltre a profitti miliardari, da reinvestire

in ricerca, abbia anche finanziamenti pubblici.

Bisogna segnalare come l'analisi abbia evidenziato una riduzione consistente degli investimenti di Alphabet Inc. negli ultimi due anni. Si può supporre come tale contrazione possa essere ascritta ad una posizione di vantaggio alla quale è venuta meno, in una certa misura, l'esigenza di essere ulteriormente consolidata nel breve periodo.

Da subito, durante questo lavoro è emersa la difficoltà nel reperire i dati necessari per svolgere un'analisi basata sulla lobbying di una grande azienda statunitense.

Di questa tesi si vuole quindi sottolineare, soprattutto, la metodologia utilizzata per la raccolta dei dati, evidenziando come gli strumenti descritti possano essere facilmente riutilizzati nell'analisi di aziende diverse da Google stessa. In particolare la procedura sviluppata all'interno del Capitolo 2 permette un estrapolazione efficiente dei dati rilevanti all'interno di un lobbying report. Il confronto Mallet e Gensim, esposto nel Capitolo 3, evidenzia le modalità con cui affrontare la ricerca di informazioni rilevanti all'interno di diverse tipologie di testo.

Appendice A

Generic Issue Area Code

Nella pagina che segue è disponibile l'elenco di tutti i Generic Issue Area Code disponibili. Tale tabella è disponibile al seguente URL:

https://lda.congress.gov/ld/help/default.htm?turl=Documents%2FAppCodes.htm.

Code	Description	Code	Description
ACC	Accounting	HOM	Homeland Security
ADV	Advertising	HOU	Housing
AER	Aerospace	IMM	Immigration
AGR	Agriculture	IND	Indian/Native American Affairs
ALC	Alcohol & Drug Abuse	INS	Insurance
ANI	Animals	LBR	Labor Issues/Antitrust/Workplace
APP	Apparel/Clothing Industry/Textiles	INT	Intelligence and Surveillance
ART	Arts/Entertainment	LAW	Law Enforcement/Crime/Criminal Justice
AUT	Automotive Industry	MAN	Manufacturing
AVI	Aviation/Aircraft/Airlines	MAR	Marine/Maritime/Boating/Fisheries
BAN	Banking	MED	Medical/Disease Research/Clinical Labs
BNK	Bankruptcy	MIA	Media (Information/Publishing)
BEV	Beverage Industry	MMM	Medicare/Medicaid
BUD	Budget/Appropriations	MON	Minting/Money/Gold Standard
CAW	Clean Air & Water (Quality)	NAT	Natural Resources
CDT	Commodities (Big Ticket)	PHA	Pharmacy
CHM	Chemicals/Chemical Industry	POS	Postal
CIV	Civil Rights/Civil Liberties	RRR	Railroads
COM	Communications/Broadcasting/Radio/TV	RES	Real Estate/Land Use/Conservation
CPI	Computer Industry	REL	Religion
CSP	Consumer Issues/Safety/Protection	RET	Retirement
CON	Constitution	ROD	Roads/Highway
CPT	Copyright/Patent/Trademark	SCI	Science/Technology
DEF	Defense	SMB	Small Business
DOC	District of Columbia	SPO	Sports/Athletics
DIS	Disaster Planning/Emergencies	TAR	Miscellaneous Tariff Bills
ECN	Economics/Economic Development	TAX	Taxation/Internal Revenue Code
EDU	Education	TEC	Telecommunications
ENG	Energy/Nuclear	TOB	Tobacco
ENV	Environmental/Superfund	TOR	Torts
FAM	Family Issues/Abortion/Adoption	TRD	Trade (Domestic & Foreign)
FIR	Firearms/Guns/Ammunition	TRA	Transportation
FIN	Financial Institutions/Investments/Securities	TOU	Travel/Tourism
FOO	Food Industry (Safety, Labeling, etc.)	TRU	Trucking/Shipping
FOR	Foreign Relations	URB	Urban Development/Municipalities
FUE	Fuel/Gas/Oil	UNM	Unemployment
GAM	Gaming/Gambling/Casino	UTI	Utilities
GOV	Government Issues	VET	Veterans
HCR	Health Issues	WAS	Waste (hazardous/solid/interstate/nuclear)
		WEL	Welfare

Tabella A.1: Genric Issue Area Code

Bibliografia

[1] Kamil Franek. Google financial results: Guided overview analysis (2018).

[2] FISCAL YEAR RESULTS.

2010:
https://www.sec.gov/Archives/edgar/data/1288776/000119312511011026/dex991.htm
2011:
https://www.sec.gov/Archives/edgar/data/1288776/000119312512017008/d285258dex991.htm
2012:
https://www.sec.gov/Archives/edgar/data/1288776/000128877613000006/goog20121231exhibit991.htm
2013:
https://www.sec.gov/Archives/edgar/data/1288776/000128877614000013/googq42013exhibit991.htm
2014:
https://www.sec.gov/Archives/edgar/data/1288776/000128877615000004/googq42014exhibit991.htm
2015:
https://abc.xyz/investor/static/pdf/2015Q4_google_earnings_release.pdf
2016:
https://abc.xyz/investor/static/pdf/2016Q4_alphabet_earnings_release.pdf
2017:

https://abc.xyz/investor/static/pdf/2017Q4_alphabet_earnings_release.pdf
2018:
https://abc.xyz/investor/static/pdf/2018Q4_alphabet_earnings_release.pdf
2019:
https://abc.xyz/investor/static/pdf/2019Q4_alphabet_earnings_release.pdf.

[3] Wikipedia. Alphabet inc. — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[4] Wikipedia. Google (azienda) — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[5] Google. [Online; controllata il 21-novembre-2020].

[6] Treccani. [Online; controllata il 21-novembre-2020].

[7] Wikipedia. Lobbying in the united states — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[8] Wikipedia. Gruppo di pressione — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[9] Wikipedia. Federal regulation of lobbying act of 1946 — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[10] 104th Congress. Lobbying disclosure act of 1995.

[11] Wikipedia. Securities and exchange commission — wikipedia, l'enciclopedia libera. [Online; controllata il 21-novembre-2020].

[12] Lobbying disclosure web site, https://lobbyingdisclosure.house.gov/.

[13] Wikipedia. Text mining — wikipedia, l'enciclopedia libera. [Online; controllata il 20-novembre-2020].

[14] Wikipedia. Topic model — wikipedia, l'enciclopedia libera, 2020. [Online; controllata il 20-novembre-2020].

[15] Micah D. Saxton. A gentle introduction to topic modeling using python. *THEOLOGICAL LIBRARIANSHIP*, 2018.

[16] Megan R. Brett. Topic modeling: A basic introduction. *Journal of Digital Humanities*, 2013.

[17] Juan S. Munoz Arango Islam. Akef. Mallet vs gensim: Topic modeling for 20 news groups report. *University of Arkansas at Little Rock*, 2016.

[18] Michael I. Jordan David M. Blei, Andrew Y. Ng. Latent dirichlet allocation. *Journal of Machine Learning Research 3 (2003) 993-1022*, 2003.

[19] Wikipedia. Lemmatizzazione — wikipedia, l'enciclopedia libera, 2020. [Online; controllata il 23-novembre-2020].

[20] Andrew Kachites McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[21] Wikipedia. Elaborazione del linguaggio naturale — wikipedia, l'enciclopedia libera, 2020. [Online; controllata il 20-novembre-2020].

[22] Swaathi Kakarla. Natural language processing: Nltk vs spacy.

[23] Akanksha Malhotra. Introduction to libraries of nlp in python — nltk vs. spacy.

[24] Bhargav Srinivasa Desikan. text_analysis_tutorial.

[25] Congress of U.S.A. Congress web page https://www.congress.gov.

[26] Naomi Klein. I colossi del digitale stanno usando il virus per intrecciarsi con la politica e imporre un futuro a loro immagine e somiglianza. *L'espresso*, page 26/30, 07-06-2020.