

Estrazione Temi e Parole Chiave

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

Centro Interdisciplinare di Ricerche per la Computerizzazione
dei Segni dell'Espressione (CIRCSE)



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

COSA FAREMO

6 MARZO: Introduzione teorica e uso di una pipeline (demo online)

13 MARZO: Uso di una pipeline per l'italiano + focus su NER

20 MARZO: Estrazione di temi e parole chiave

27 MARZO: Estrazione del sentiment

TOPIC MODELING

“Topic models are a suite of algorithms that **uncover the hidden thematic structure** in document collections. These algorithms help us develop new ways to search, browse and summarize **large archives of texts**”

(<http://www.cs.columbia.edu/~blei/topicmodeling.html>)

- Quali argomenti sono contenuti in un corpus?
- Algoritmi NON SUPERVISIONATI applicabili a grandi collezioni di documenti
- Algoritmo più usato: latent Dirichlet allocation (LDA)

TOPIC MODELING: Latent Dirichlet Allocation

- Processo inferenziale e iterativo che cerca di rispondere alla domanda: qual è la struttura nascosta di topic che probabilmente ha generato il corpus?
- Intuizioni di base:
 - i topic sono generati **prima** dei documenti e sono strutture nascoste nei documenti stessi
 - un topic è una **distribuzione** di probabilità su un insieme di parole: il topic “IMMIGRATION” contiene parole relative all’immigrazione con alta probabilità
 - ogni documento contiene **più topic**
 - tutti i documenti del corpus contengono lo stesso gruppo di topic ma ciascuno in **proporzioni differenti**

TOPIC MODELING: LDA - COME FUNZIONA

- 1) viene generato un dizionario di tutte le parole del corpus
- 2) per ogni documento vengono indicate le frequenze assolute delle parole effettivamente in esso presenti
- 3) all'inizio ogni parola viene assegnata a uno o più topic in modo semi-casuale: più è alta la frequenza della parola e più probabilmente viene assegnata a più topic
- 4) l'assegnazione iniziale viene aggiornata in maniera iterativa: per ciascuna parola la sua assegnazione ad un certo topic in un certo documento viene aggiornata in base a:
 - la frequenza relativa con cui la parola appare nel topic
 - la frequenza assoluta delle altre parole appartenenti allo stesso topic nel documento

TOPIC MODELING

- Quali argomenti sono contenuti in un corpus?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

TOPIC MODELING: LDA

- Quali argomenti sono contenuti in un corpus?

But to fix our immigration system, we must change our leadership in Washington and we must change it quickly. Sadly, sadly there is no other way. The truth is our immigration system is worse than anybody ever realized. But the facts aren't known because the media won't report on them. The politicians won't talk about them and the special interests spend a lot of money trying to cover them up because they are making an absolute fortune. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the immigration system in our country is that it serves the needs of wealthy donors, political activists and powerful, powerful politicians.

Trump, 31 August 2016

- IMMIGRATION

- POLITICS

- ECONOMY

TOPIC MODELING: LDA

| | | | | |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| But to fix our must change Washington and quickly. Sadly other way. immigration s anybody ever aren't known report on their talk about interests spent to cover the making an abso way it is. complicated subject, you fundamental immigration sy that it serve donors, poli powerful, powe | As secretary Clinton allo criminal alie because their to take them. They were too them back. Who would do this? this? A weak a would do thi described Hill most radical in United States summary of wh support sanctu Security, Med welfare for al by making them which will die immigrants. | Social Secu lifetime we immigrants b citizens. And being treated veterans. Reme going to all illegal immigr visa overstay release on the hey, go ahead It's called Expanding unconstitution including ins millions of i even more crim Obama's non- And she wants in Syrian refu country . | All Americans country, in wonderful, p immigrants are jobs and wag totally protect our nation are people living everybody. An erased -- it lawful immigr if you look a the borders, a are erased, borders, we r And that's r And I have t endorsed by th 16,500. By IC First time anybody for pr | As I mentioned, Pueblo is filled with wonderful, hard-working immigrants. It's these hard-working immigrants who stand to lose the most from our open border immigration policy. Illegal immigration and broken Visa programs take jobs directly from Latino and Hispanic workers living here lawfully today -- you know that. They're taking your jobs. Illegal immigration also brings with it massive crime and massive drugs, including a terrible heroin problem right here in Colorado -- you have a big problem. So we're going to build the border wall and we are not -- what? We're going to build the wall and we're going to stop the drugs, the gangs, the violence from pouring into Colorado. |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

“That’s how topic modeling works in practice. You assign words to topics randomly and then just keep improving the model, to make your guess more internally consistent, until the model reaches an equilibrium that is as consistent as the collection allows.”

Ted Underwood, 2012

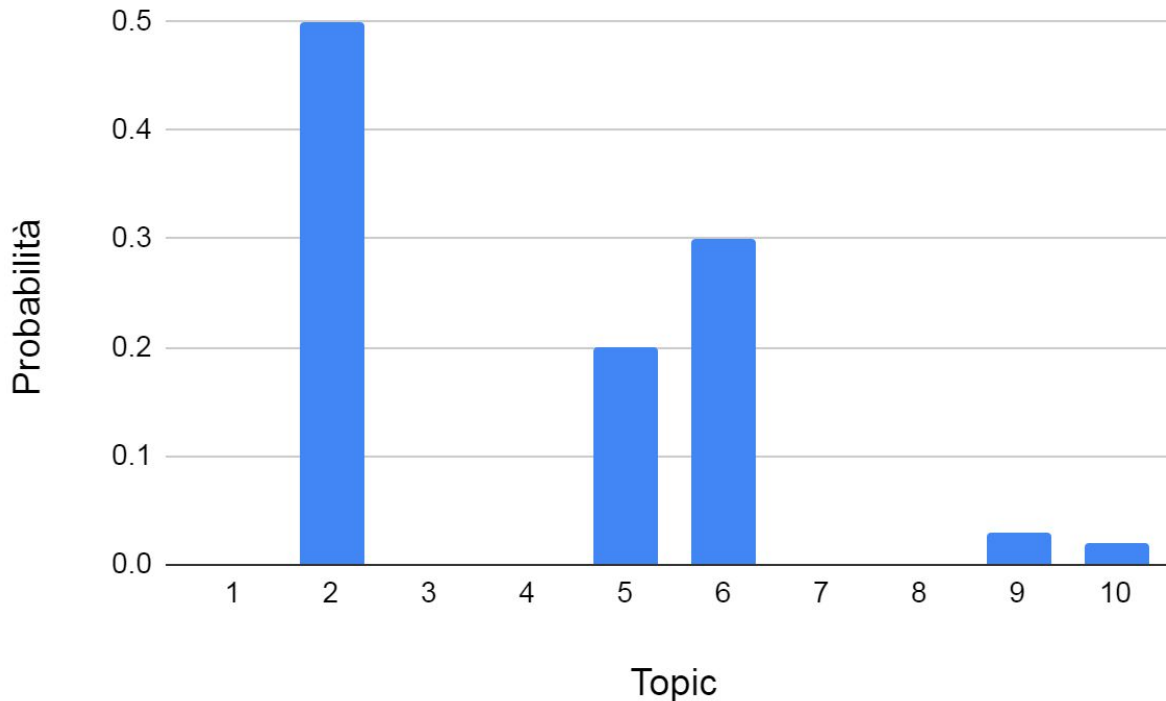
TOPIC MODELING: LDA

Ci dobbiamo aspettare 2 output principali:

- 1) una lista di topic (cioè di gruppi di parole)
- 2) una lista dei documenti che sono fortemente associati a ciascun topic

Idealmente, ogni topic dovrebbe essere ben distinto da tutti gli altri

TOPIC MODELING: LDA



2 - immigrants, immigration, border, Mexico, wall → IMMIGRATION

5 - politicians, Washington, party, democrats, activists → POLITICS

6 - money, GDP, fortune, economy, donors, banks, wealthy → ECONOMY

TOPIC MODELING

“Essentially, all models are wrong, but some are useful.”

George Box, 1987



- Non ci sono metodi facili di valutazione
- Non ci sono metodi certi e facili per determinare il numero migliore di topic
- Molto ambiguo e “troppo” configurabile



- Buon punto di partenza per esplorare i dati
- Genera nuovi modi per guardare a grosse quantità di dati

APPLICAZIONE: SCIENZE POLITICHE

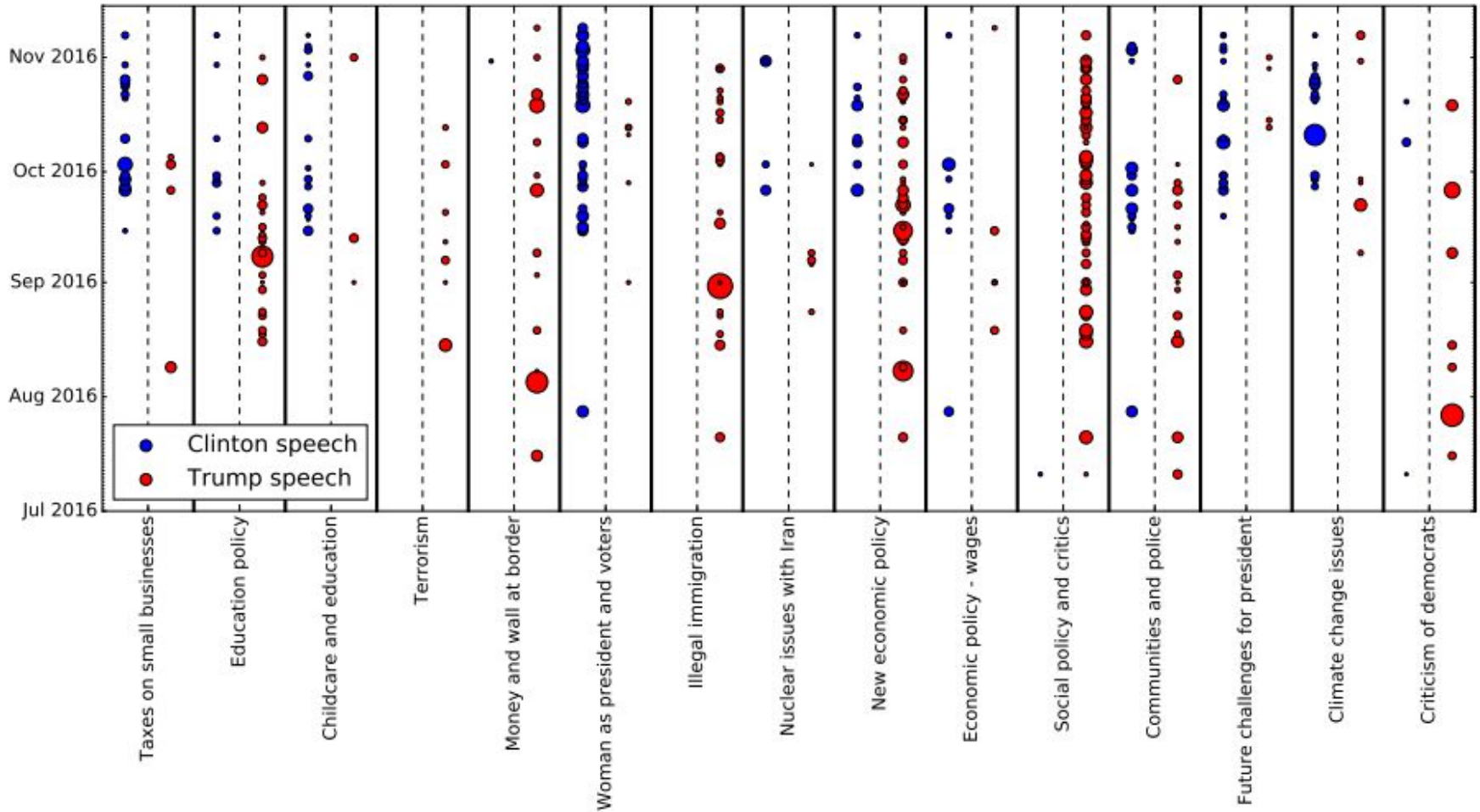
Topic Signatures in Political Campaign Speeches

<https://www.aclweb.org/anthology/D17-1249.pdf>

“We propose to identify in political speeches the favourite topics considered by each candidate as well as how and when they evolve throughout the campaign. In our opinion, this gives critical clues to identify and to explain each candidate’s main ideas and their evolution.”

(Gautrais et al., 2017)

APPLICAZIONE: SCIENZE POLITICHE



APPLICAZIONE: MODA E SOCIETÀ

Robots Reading Vogue - Data Mining is in Fashion

<http://dh.library.yale.edu/projects/vogue/>

“Few magazines can boast being continuously published for over a century, familiar and interesting to almost everyone, full of iconic pictures — and also completely digitized and marked up as both text and images. What can you do with over 2,700 covers, 400,000 pages, 6 TB of data?”

APPLICAZIONE: MODA E SOCIETÀ

“Women’s Health”

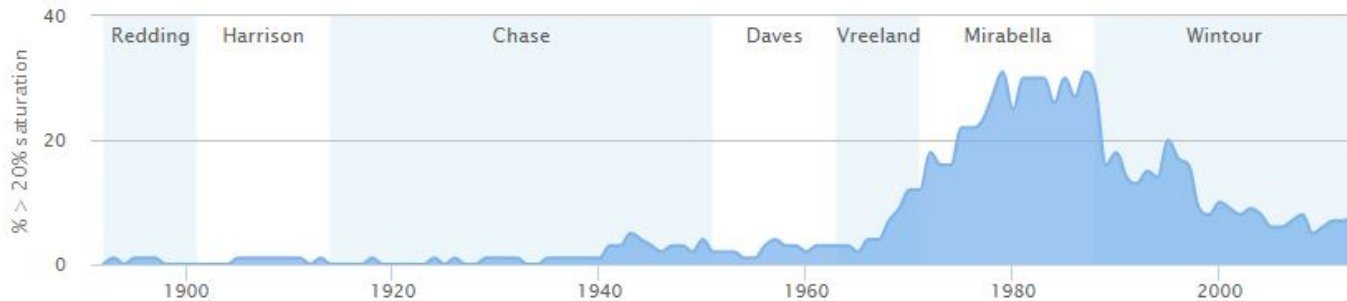
Women’s Health Words



Women’s Health Phrases



Women’s Health over Time



Articles

Click the timeline to the left.

APPLICAZIONE: STORIA

Language Resources for Historical Newspapers: the Impresso Collection

<https://www.aclweb.org/anthology/2020.lrec-1.121.pdf>

“The impresso web application supports faceted search with respect to language-specific topics (French, German, Luxembourgish). We use the well-known MALLET toolkit, which allows the training and inference of topic models with Latent Dirichlet Allocation.”

(Ehrmann et al., 2020)

APPLICAZIONE: STORIA

← TOPICS

fr "beurre · viande · pomme · fruit · sucre ..."

beurre · viande · pomme · fruit · sucre · pain · légume · fromage · sel · plat · huile · lait · crème · soupe · pâte · repas · four · sauce · jus · morceau · salade · recette · café · oignon · tomate · poisson · chocolat · menu · farine · préparation · citron · graisse · vea

MORE ...

572,651 ARTICLES

ORDER BY

TOPIC RELEVANCE ▾



Recettes de cuisine

Journal de Genève, FRIDAY, APRIL 29, 1938 (p. 10; 10; 10)

Recettes de cuisine Riz royal à la mandarine Pour 7 à 8 personnes : 125 gr. de riz ; 1 litre et quart de lait environ ; 150 gr. de sucre en poudre environ ; 3 œufs (3 jaunes et un blanc) ; 2 feuilles de gélatine ; 125 gr...

VIEW

ADD TO COLLECTION ... ▾



NOS RECETTES - NOS RECETTES - NOS RECETTES - NOS RECETTES

L'Express, WEDNESDAY, FEBRUARY 28, 1979 (p. 11)

NOS RECETTES-NOS RECETTES-NOS RECETTES-NOS RECETTES Rôti de bœuf braisé Pour quatre personnes : 1 kg 500 de rôti de bœuf braisé (ce morceau suffit pour deux fois), 1 cuillerée à soupe de moutarde 1 cuillerée à soupe de graisse, 2 gros oignons, 1 poireau, 1 carotte, 1 cuillerée à café...

<https://impresso-project.ch/app/>

APPLICAZIONE: STUDI LETTERARI

Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama

<http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html>

“The data used in this study comes from the Théâtre classique collection maintained by Paul Fièvre (2007-2015). At the time of writing, this continually-growing, freely available collection of French dramatic texts contained 890 plays published between 1610 and 1810, thus covering the Classical Age and the Enlightenment.”

(Schöch, 2017)

APPLICAZIONE: STUDI LETTERARI

topic 32 (1/60)

oseraimer
seulâme
peine craindre
doux
secret vain
offrir
laisser
hymen
souffrir
espoir
connaître
croire
plaire
aveu
prix
princesse
plandre
voeu
intérêt
forcer ardeur mériter
madame gloire
choix flamme

topic 3 (6/60)

être mêmeexpliquer
attendre
surprendre
chercher
esprit
trouver
effet
passer
aimer
seul crois
entendre moins
penser
besoin
mystère
peine
doute
ignorer
suspçon
temps
part
soin
oser
paraître
taire
tenir
croire
connaître
ami
avoir
cacher
apprendre
découvrir
avouer
sembler

topic 30 (53/60)

poète
seul
sujet
génie
muse
mauvais
goût
art
talent
rôle
scène
nom
merveille
premier
public
lire
pièce
sonnet
plaisir
nommer
commencer
prose
écrit
esprit
beau
trouver
nouveau
représenter
rime
acteur
théâtre
jouer
œuvre
comédie
temps

topic 34 (57/60)

remède
chose
vif
malade
savant
monseigneur
science
fou
statue
connaître
homme
maladie
corps
folie
cause
avis
guérison
médecine
art
jours
effet
seul
santé
malade
docteur
vapeur
charlatan
médecin
guérir
habile
mourir
sentir
raison
soutenir
goutte
secours
vrai
manquer

(Schöch, 2017)

APPLICAZIONE: STUDI LETTERARI

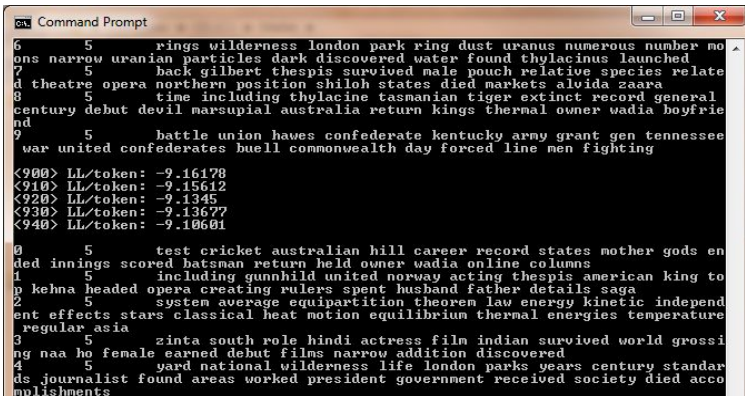
PROBLEMI INTERPRETATIVI

- Il concetto di topic (o thema) nella **linguistica funzionalista**?
- La nozione di isotopia in **strutturalismo e semiotica**?
- I concetti di tema e motivo nella **critica tematica**?
- La **nozione foucaultiana** di «discorso»?

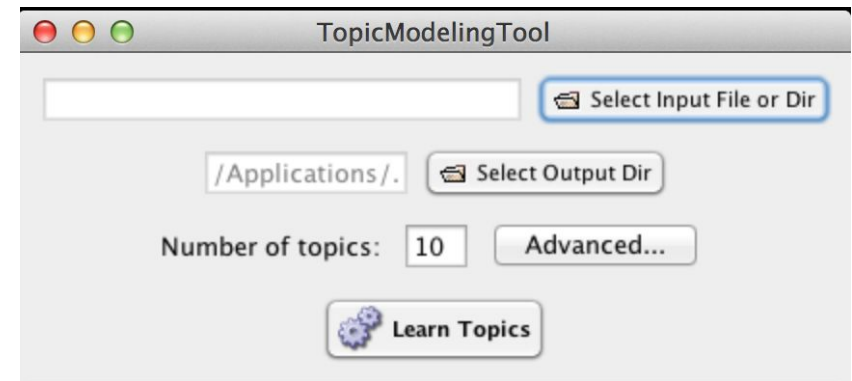
“La discussione sulle possibili interpretazioni semiotico-letterarie della nozione di topic model e la constatazione della difficoltà teoriche che esse presentano ci porta ad affermare che in effetti **non è possibile trovare un unico e soddisfacente correlato teorico-letterario** dei risultati di questi metodi di analisi quantitativa”
(Ciotti, 2017)

TOPIC MODELING: STRUMENTI

- Online Demo:
<https://mimno.infosci.cornell.edu/jsLDA/>
- MALLET:
<http://mallet.cs.umass.edu/>
- Topic-modeling-tool:
<https://nlp.stanford.edu/software/tmt/tmt-0.4/>



```
6      5      rings wilderness london park ring dust uranus numerous number mo
ons narrow uranian particles dark discovered water found thylacinus launched
7      5      back gilbert thespis survived male pouch relative species relate
d theatre opera northern position shiloh states died markets alvida zaara
8      5      time including thylacine tasmanian tiger extinct record general
century debut devil marsupial australia return kings thermal owner wadia boyfrie
nd
9      5      battle union haves confederate kentucky army grant gen tennessee
war united confederates buell commonwealth day forced line men fighting
<900> LL/token: -9.16178
<910> LL/token: -9.15612
<920> LL/token: -9.1345
<930> LL/token: -9.13677
<940> LL/token: -9.10601
0      5      test cricket australian hill career record states mother gods en
ded innings scored batsman return held owner wadia online columns
1      5      including gunnhild united norway acting thespis american king to
p kehna headed opera creating rulers spent husband father details saga
2      5      system average equipartition theorem law energy kinetic independ
ent effects stars classical heat motion equilibrium thermal energies temperature
regular asia
3      5      zinta south role hindi actress film indian survived world grossi
ng naa he female earned debut films narrow addition discovered
4      5      yard national wilderness life london parks years century standar
ds journalist found areas worked president government received society died acco
mplishments
```

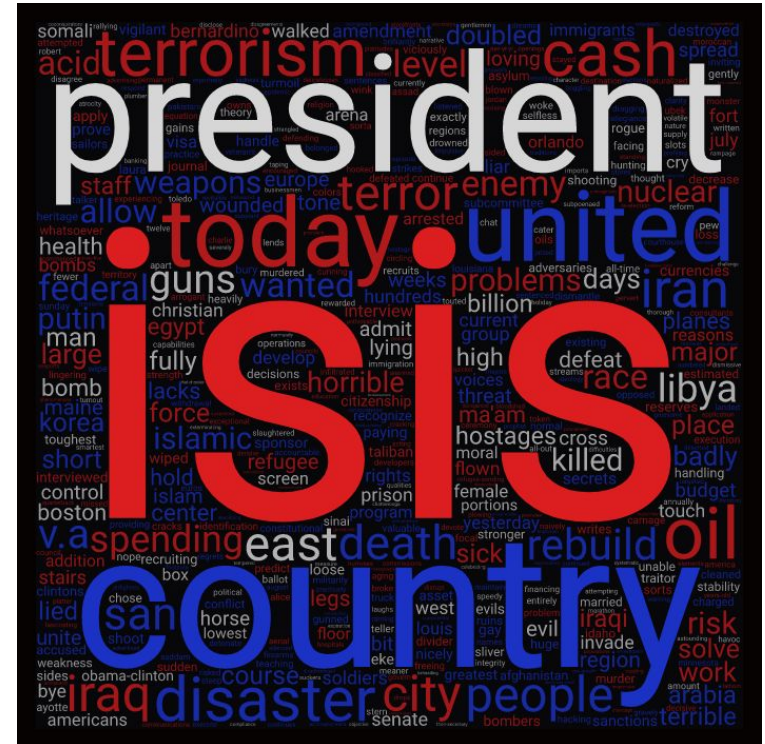


TOPIC MODELING: DEMO ONLINE

1. Aprire [Trump.txt](#) con un editor di testo e vedere come è formattato
2. Aprire [stopwords-en.txt](#) con un editor di testo
3. Aprire un browser (no Explorer) e andare su:
<https://mimno.infosci.cornell.edu/jsLDA/jslda.html>
4. Cliccare su “Choose File” per l’opzione [Document Upload](#) e caricare [Trump.txt](#)
5. Cliccare “Choose File” per l’opzione [Stoplist Upload](#) e caricare [stopwords-en.txt](#)
6. Cliccare su [Load](#)
7. Cliccare su [Vocabulary](#) per pulire il dizionario
8. Cliccare su [Run 50 iterations](#)
9. Quando i topic sembrano stabili andare nella sezione [Download](#) per vedere i possibili file dei risultati scaricabili
10. Non chiudere la pagina web

TOPIC MODELING: VISUALIZZAZIONE

- Word Cloud:
 - Scaricare il file “Topic words” nella sezione *Download*
 - Scegliere un topic (controllando sulla demo il contenuto) e selezionare tutte le parole con peso diverso da 0
 - Copiare parole e pesi in un altro foglio di calcolo
 - Andare su <https://wordart.com/create> e creare la propria world cloud



TOPIC MODELING: VISUALIZZAZIONE

- Network:
 - Scaricare il file “Doc-topic graph file (for Gephi)” nella sezione *Download*
 - Andare su <https://vistorian.net/> e cliccare su “Demo”
 - Cliccare su “New” accanto a “Networks”
 - Cliccare su “Upload a new link table” e scegliere il file [gephi.csv](#) o [gephi-time.csv](#) che vi ho fornito
 - Selezionare il significato delle singole colonne (id è già ok):
Source → Source Node; Target → Target Node; Weight → Weight;
Type → Link Type; (per il file gephi-time.csv anche Time → Time
(formato: YYYY/MM/DD))
 - Cliccare in alto su “Node Link”
 - Provare anche le visualizzazioni: “Adjacency Matrix” e “Time Arcs”

ESTRAZIONE PAROLE-CHIAVE

- Keyphrases (or key-concepts) = ngrammi che catturano i concetti principali di un documento

But to fix our **immigration system**, we must change our **leadership in Washington** and we must change it quickly. Sadly, sadly there is no other way. The truth is our **immigration system** is worse than anybody ever realized. But the facts aren't known because the **media** won't report on them. The **politicians** won't talk about them and the special interests spend a **lot of money** trying to cover them up because they are making an absolute **fortune**. That's the way it is. Today, on a very complicated and very difficult subject, you will get the truth. The fundamental problem with the **immigration system** in our **country** is that it serves the needs of **wealthy donors**, **political activists** and powerful, **powerful politicians**.

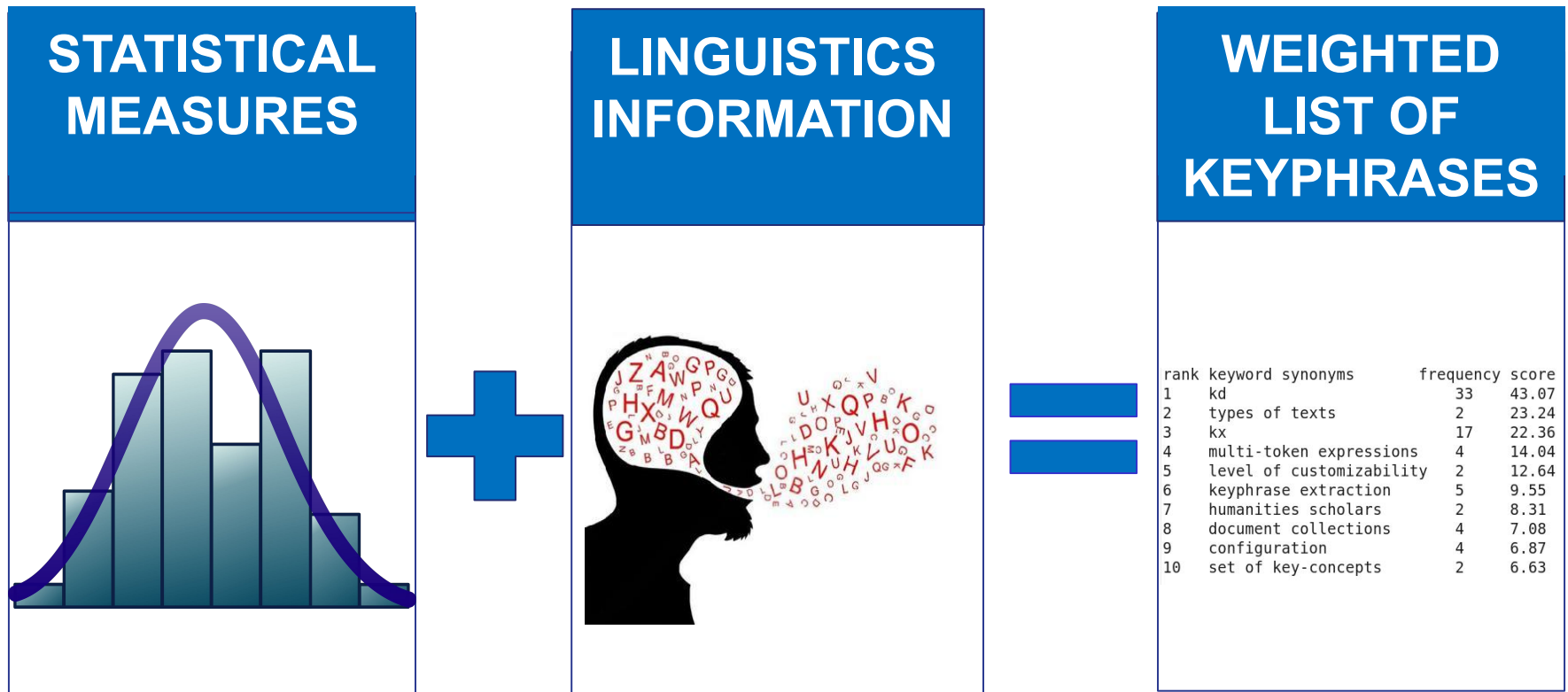
Sia espressioni singole che
multi-token

Sia documenti singoli che
collezioni di documenti

ESTRAZIONE PAROLE-CHIAVE

- KD = Keyphrase Digger

http://celct.fbk.eu:8080/KD_KeyDigger/



KD = COME FUNZIONA

Fasi di analisi:

- 1) l'input è diviso in “fette” processate parallelamente per aumentare la velocità di estrazione
- 2) vengono individuati gli ngrammi in base ai pattern di PoS definiti per ogni lingua
EN: A + S = Great War
IT: A + S = Grande Guerra / S + A = Guerra Mondiale
- 3) gli ngrammi estratti dalle varie “fette” sono riuniti in un'unica lista
- 4) vengono calcolate le frequenze assolute e rimossi gli ngrammi sotto la soglia definita dall'utente
- 5) viene calcolato un punteggio di rilevanza in base a vari parametri definibili dall'utente

KD = COME USARLO

1. Demo online: http://celct.fbk.eu:8080/KD_KeyDigger/ (vedi dopo)
2. Tint: incluso nella versione usata il 13/10
Nuove opzioni di configurazioni aggiunte al file *default-config.properties*

```
keyphrase.language = ITALIAN
```

```
keyphrase.prefer_specific_concept = MEDIUM
```

```
keyphrase.local_frequency_threshold = 2
```

```
keyphrase.numberOfConcepts = 5
```

```
keyphrase.max_keyword_length = 4
```

```
keyphrase.boost_acronyms = false
```

```
keyphrase.skip_keyword_with_proper_noun = true
```

```
keyphrase.skip_proper_noun = true
```

```
keyphrase.no_abstract = true
```

```
keyphrase.rerank_by_position= false
```

KD = COME USARLO

3. Codice su Github: <https://github.com/dhfbk/KD>
4. Pacchetto Java: scaricare la cartella **kd** condivisa su Teams
 - Vedere come si struttura la cartella **kd**
 - Andare sul Terminale e usare il comando `cd` per entrare nella cartella **kd**
 - Scrivere il seguente comando poi premere Invio (output sul Terminale)

```
java -Dfile.encoding=UTF-8 -jar KD.jar -lp conf -lang  
ITALIAN -c TOKEN_LEMMA_POS -p MEDIUM -s -sk -n 20  
-STDOUT Cap31-kd.txt
```
 - Scrivere il seguente comando poi premere Invio (output in un file)

```
java -Dfile.encoding=UTF-8 -jar KD.jar -lp conf -lang  
ITALIAN -c TOKEN_LEMMA_POS -p MEDIUM -s -sk -n 20  
Cap31-kd.txt
```

N.B Per i file inglesi serve l'opzione `-us` che lancia Stanford, per quelli italiani bisogna dare in input un file già processato

KD = COME USARLO

- Scrivere il seguente comando poi premere Invio (output sul Terminale)
`java -Dfile.encoding=UTF-8 -jar KD.jar -lp conf -lang ENGLISH -us -p MEDIUM -s -sk -n 20 -g ALL_LEMMA frankenstein`
- Scrivere il comando per vedere le (tante) opzioni disponibili
`java -jar KD.jar -h`
- I file di configurazione sono in:
`conf → ITALIAN → configuration_files`
e modificabili con un editor di testo
- È possibile aggiungere nuove lingue

ESTRAZIONE PAROLE-CHIAVE: APPLICAZIONE

<http://alcidedigitale.fbk.eu/>

The screenshot displays the Alcidè Digitale application interface, which is divided into two main sections: 'Tempo e contesto' (Time and context) on the left and 'Risultati' (Results) on the right.

Tempo e contesto:

- At the top, there is a bar chart showing the frequency of key words over time, with a timeline slider below it ranging from 1901 to 1954.
- Below the chart, a list of document types is provided: Quotidiani / Periodici / Libri / Prefazioni / Documenti / Discorsi pubblici / Interventi istituzionali / Interventi di partito.
- The section is titled 'Parole chiave' (Key words) in a large green box.
- Below the title, there is a search bar with the placeholder text 'Cerca fra le parole chiave...' and a 'Cerca' button.
- A list of key words is displayed, each with a corresponding horizontal bar indicating its frequency: governo, italiana, nazione, italia, popolo, problemi, socialista, alto adige, partito, partito popolare, maggioranza, conferenza, and italiani.

Risultati:

- The section is titled 'Risultati' in a large red box.
- At the top right, it shows 'Totale Documenti: 2762'.
- Below the title, a list of search results is displayed, including: 'La Cassa centrale cattolica di mutuo soccorso', 'La Cassa centrale cattolica di mutuo soccorso in Trento (Cassa di soccorso registrata)', 'Il congresso straordinario del Comitato diocesano trentino per l'azione cattolica', 'Brevi note', 'La cultura presente e la riscossa cristiana. Discorso dello studente di filol. Alc. Degasperì al Congresso di Mezzocorona', 'I commenti nei circoli studenteschi - l'adunanza di ieri.', 'Comizio popolare per la questione universitaria.', 'Fra Bubbone ovvero sia la Matricola onesta.', 'Per l'Università italiana. Cattolici trentini!', 'Der Katholicismus und das XX Jahrhundert im Lichte Kirchlichen Entwicklung der Neuzeit - DR. Albert Ehrhard prof. ord. dell'Università di Vienna. - Stoccarda e Vienna 1902, presso G. Roth. -', 'Bancarotta dei darwinisti. Eugen Schmitt «Leo Tolstoi und seine Bedeutung» - Leipzig, 1901', 'La questione dell'università italiana', 'Al sig. prof. F. Pasini', and 'La causa Boera. (Note politiche)'.

KD: DEMO ONLINE

1. Tornare alla pagina web con la demo LDA
2. Cliccare sul topic relativo all'**immigrazione** e controllare l'ID del primo file sulla destra (i.e. il più rilevante): l'ID è il nome del file
3. Aprire la cartella Trump-Clinton e poi la cartella Trump: aprire il file *Trump_2016-08-31* con un editor di testo
4. Copiare il contenuto del file, andare su http://celct.fbk.eu:8080/KD_KeyDigger/ e incollarlo
5. Cliccare su **Run** e controllare il risultato
6. Possiamo lemmatizzare al volo usando: <http://textanalysisonline.com/spacy-word-lemmatize>
7. Scaricare i risultati in formato tsv cliccando su "Download Data" sotto la word cloud
8. È anche possibile scaricare le visualizzazioni

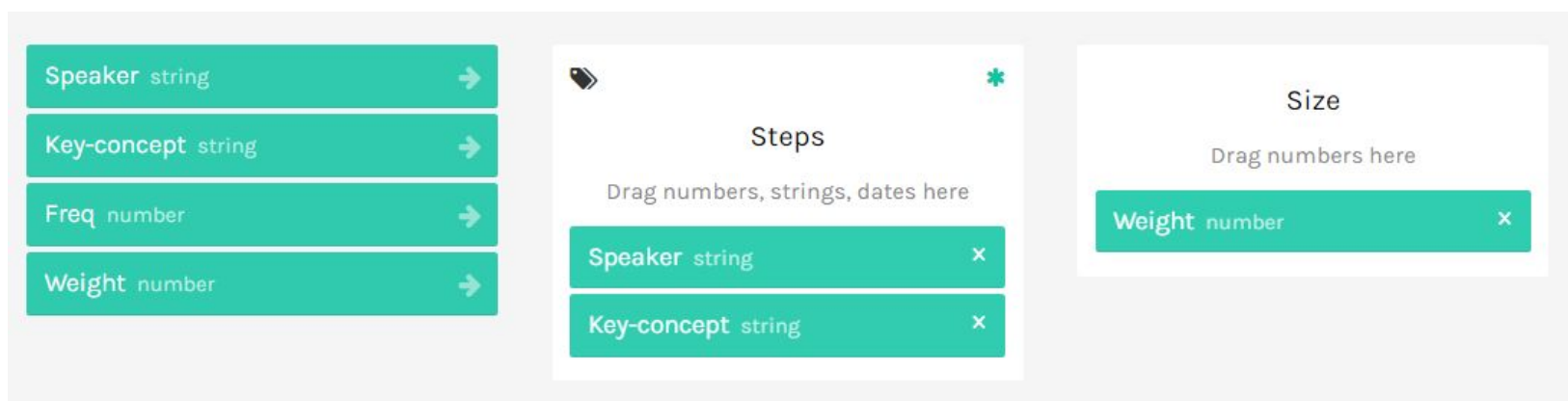
ESTRAZIONE PAROLE-CHIAVE: VISUALIZZAZIONE

1. Scaricare i risultati cliccando su “Download Data” sotto la word cloud
2. Estrarre le parole-chiave dal file più rilevante relativo al topic sull’immigrazione di Clinton: *Clinton_2016-08-05.txt*
3. Scaricare i risultati di Clinton: avremo due file tsv
4. Copiare il contenuto di ciascun tsv in un foglio di calcolo e creare un file come quello qui sotto

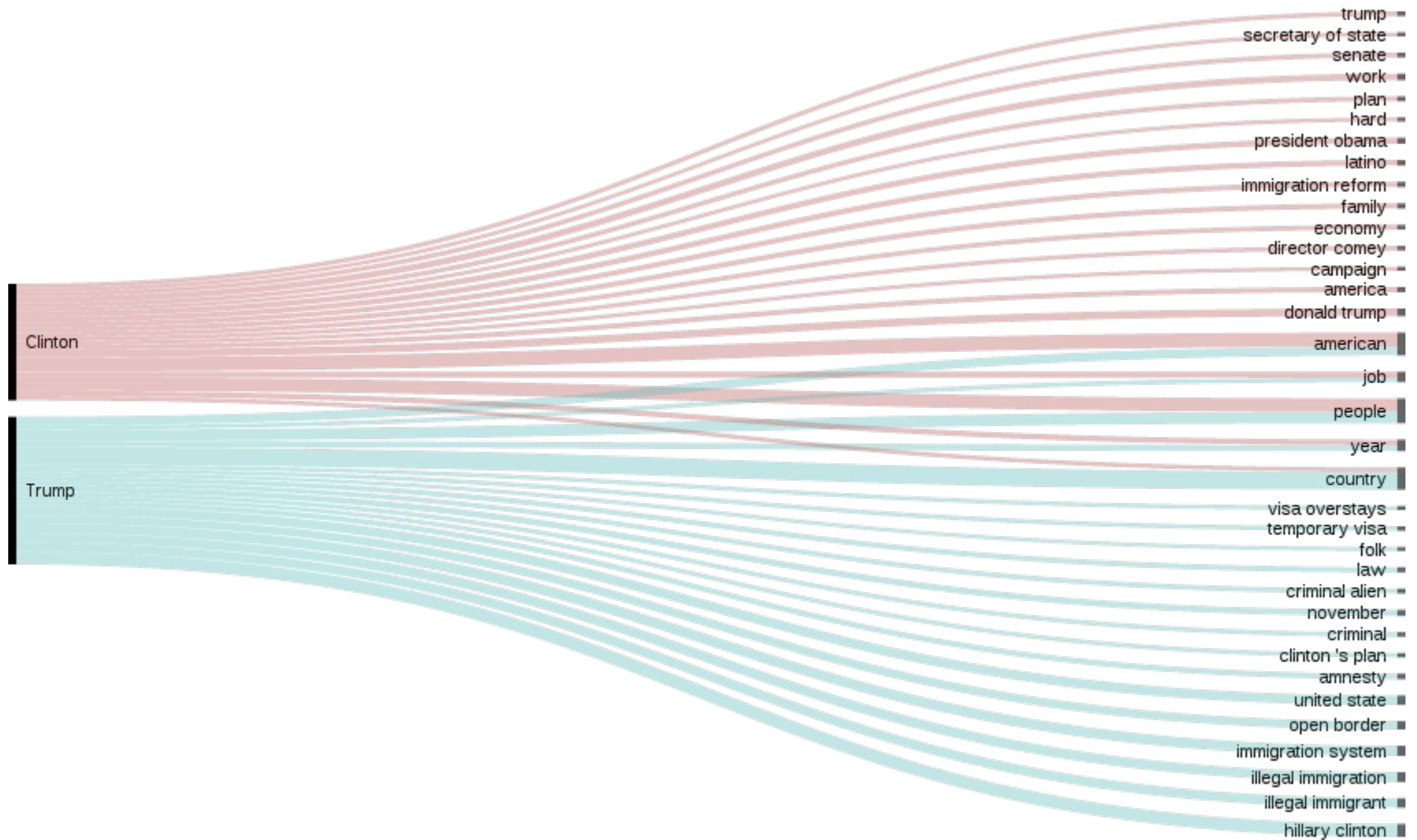
| Speaker | Key-concept | Freq | Weight |
|---------|------------------------|------|--------|
| Trump | <u>hillary clinton</u> | 19 | 9.475 |
| Trump | immigration system | 7 | 8.166 |
| Trump | illegal immigration | 7 | 7.317 |
| Trump | country | 54 | 6.751 |
| Trump | immigration law | 4 | 5.873 |

ESTRAZIONE PAROLE-CHIAVE: VISUALIZZAZIONE

5. Andare su RAW: <https://app.rawgraphs.io/>
6. Copiare il contenuto del foglio di calcolo appena creato e incollarlo in RAW
7. Scegliere sotto “Choose a Chart” la tipologia “Alluvial Diagram”
8. Scegliere le dimensioni sotto “Map your Dimensions” come di seguito:



ESTRAZIONE PAROLE-CHIAVE: VISUALIZZAZIONE



ALCUNI APPROFONDIMENTI

LDA:

<http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf>

<https://programminghistorian.org/en/lessons/topic-modeling-and-mallet>

http://testoesenso.it/article/download/462/pdf_227

<https://github.com/cpsievert/LDAvis> (visualizzazione avanzata)

<https://radimrehurek.com/gensim/> (libreria python)

ESTRAZIONE CONCETTI-CHIAVE:

https://iris.unito.it/retrieve/handle/2318/1532924/75495/Accademia_Univ_ersity_Press_978-88-99200-62-6.pdf#page=200

<http://ceur-ws.org/Vol-1749/paper38.pdf>

http://www.ai-ic.it/IJCoL/v2n2/5-sprugnoli_et_al.pdf

<https://github.com/dhfbk/KD>



GRAZIE!

Email: rachele.sprugnoli@unicatt.it

Twitter: [@RSprugnoli](https://twitter.com/RSprugnoli)

