

Topic Modeling

Rachele Sprugnoli – rachele.sprugnoli@unicatt.it

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



UNIVERSITÀ
CATTOLICA
del Sacro Cuore

COME SI SVILUPPA UN MODULO TAL

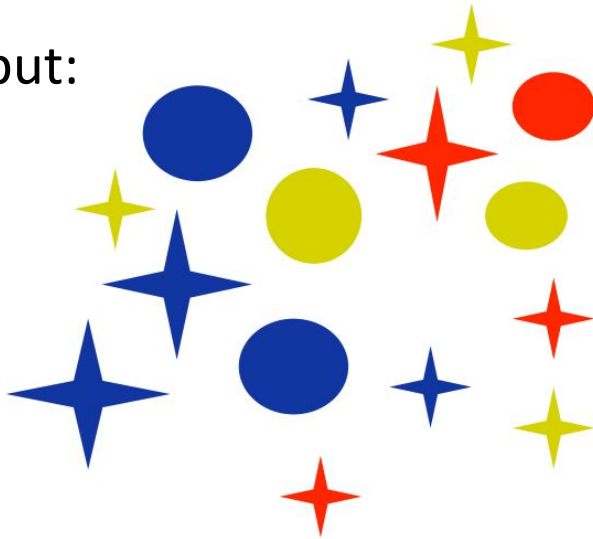
- **3 tipi principali di algoritmi di ML**

1. **NON SUPERVISIONATI:** non necessitano di un corpus annotato a mano per creare il modello
2. **SUPERVISIONATI:** utilizzano un corpus annotato a mano per la creazione dei modelli
3. **SEMI-SUPERVISIONATI:** combinano informazioni derivanti sia da corpora annotati che da dati non annotati

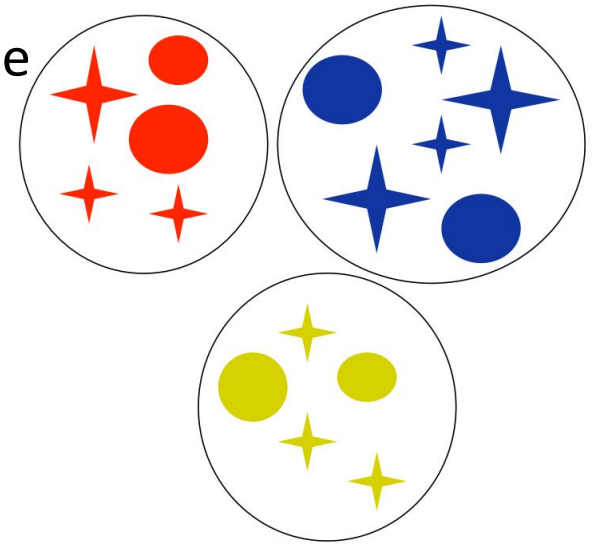
ML NON SUPERVISIONATO

- Esempio
 - **CLUSTERING**: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:



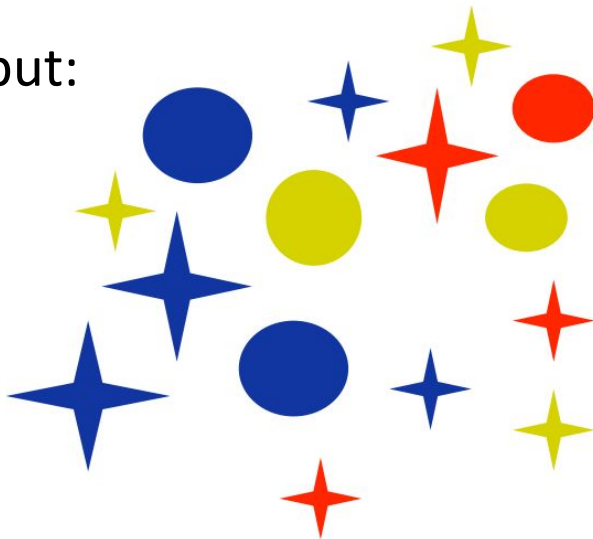
Output in base
al colore:



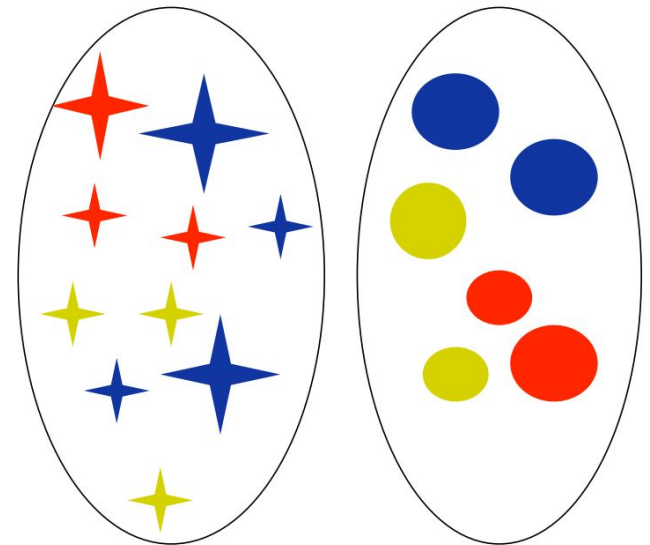
ML NON SUPERVISIONATO

- Esempio
 - **CLUSTERING**: raggruppamento dell'input in base a una qualche relazione di similitudine tra i dati

Input:



Output in
base alla
forma:



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

- 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

<p>But to fix our must change Washington and quickly. Sadly other way. immigration anybody ever aren't known report on their talk about interests spent to cover the making an abstract way it is. complicated subject, you fundamental immigration system that it serves donors, political powerful, power</p>	<p>As secretary Clinton allow criminal alien because their to take them. They were too them back. Who would do this? this? A weak would do this described Hill most radical in United States summary of which support sanctuary Security, Medicare welfare for all by making them which will die immigrants.</p>	<p>Social Security lifetime we immigrants citizens. And immigrants are being treated veterans. Remember going to all illegal immigrants visa overstay release on the hey, go ahead It's called Expanding unconstitution including ins millions of immigrants even more criminal Obama's non- And she wants in Syrian refugees country .</p>	<p>All Americans country, in wonderful, immigrants are jobs and wages totally protected our nation are people living everybody. And erased -- it lawful immigrants if you look at the borders, are erased, borders, we r And that's r And I have endorsed by the 16,500. By ICE First time anybody for pr</p>	<p>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.</p>
---	--	---	---	---

“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

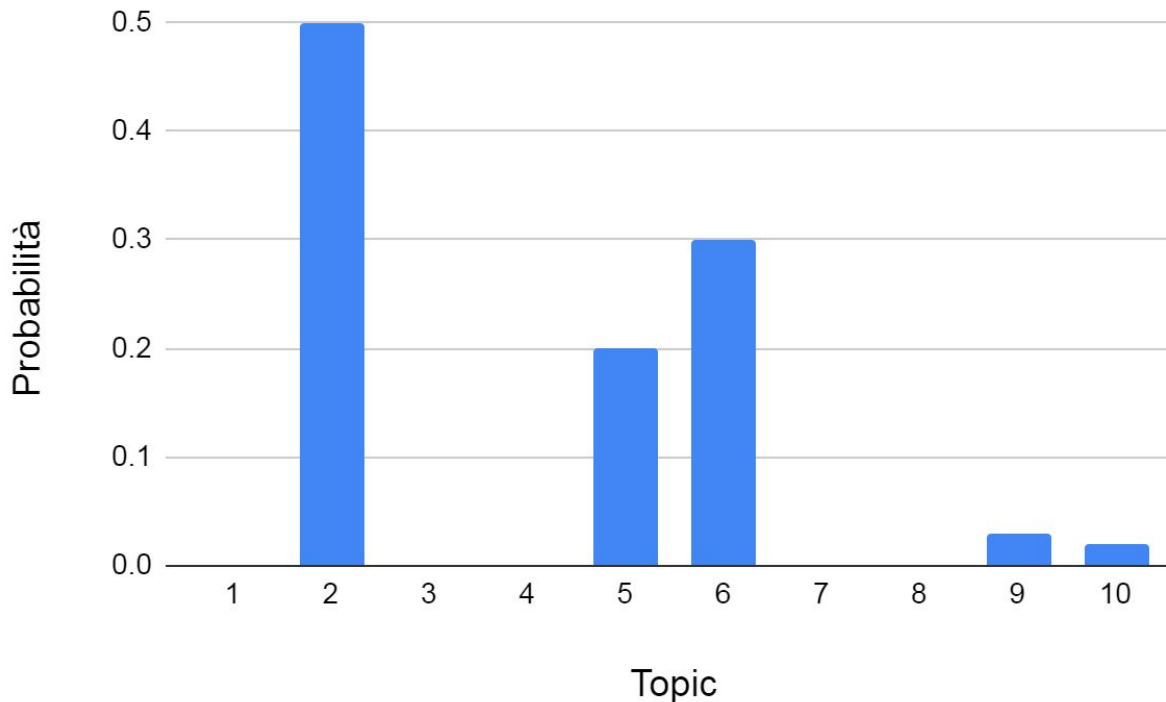
Ci dobbiamo aspettare 2 output principali:

- 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

- Esempio su un documento del corpus:



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

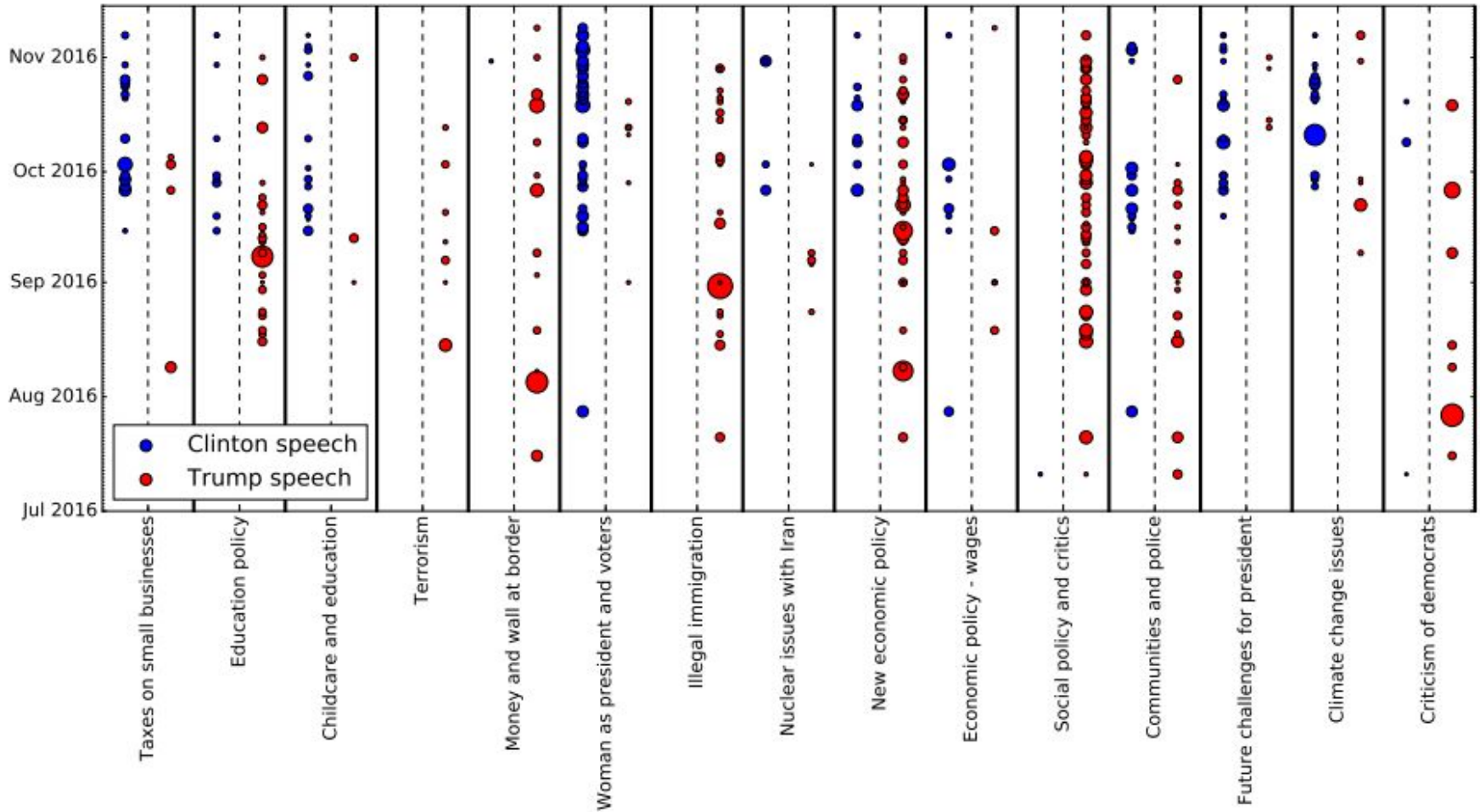
Topic Signatures in Political Campaign Speeches

<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”

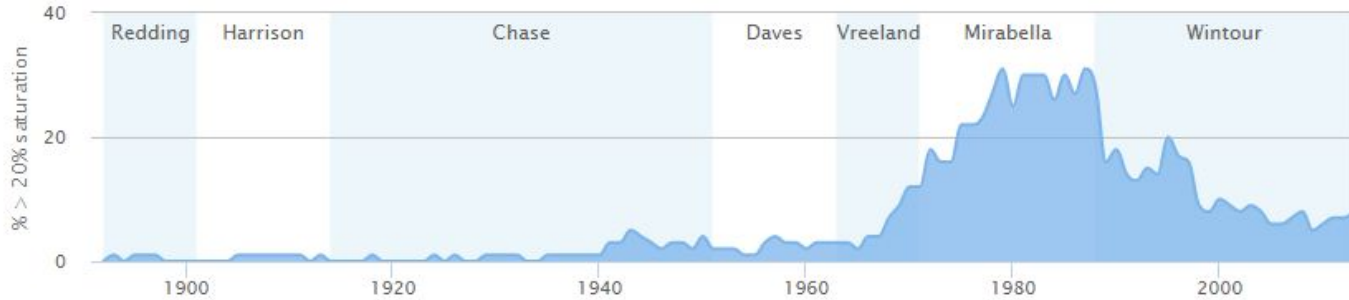
Women's Health Words



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: BENI CULTURALI

Trends in Contemporary Art Discourse: Using Topic Models to Analyze 25 years of Professional Art Criticism

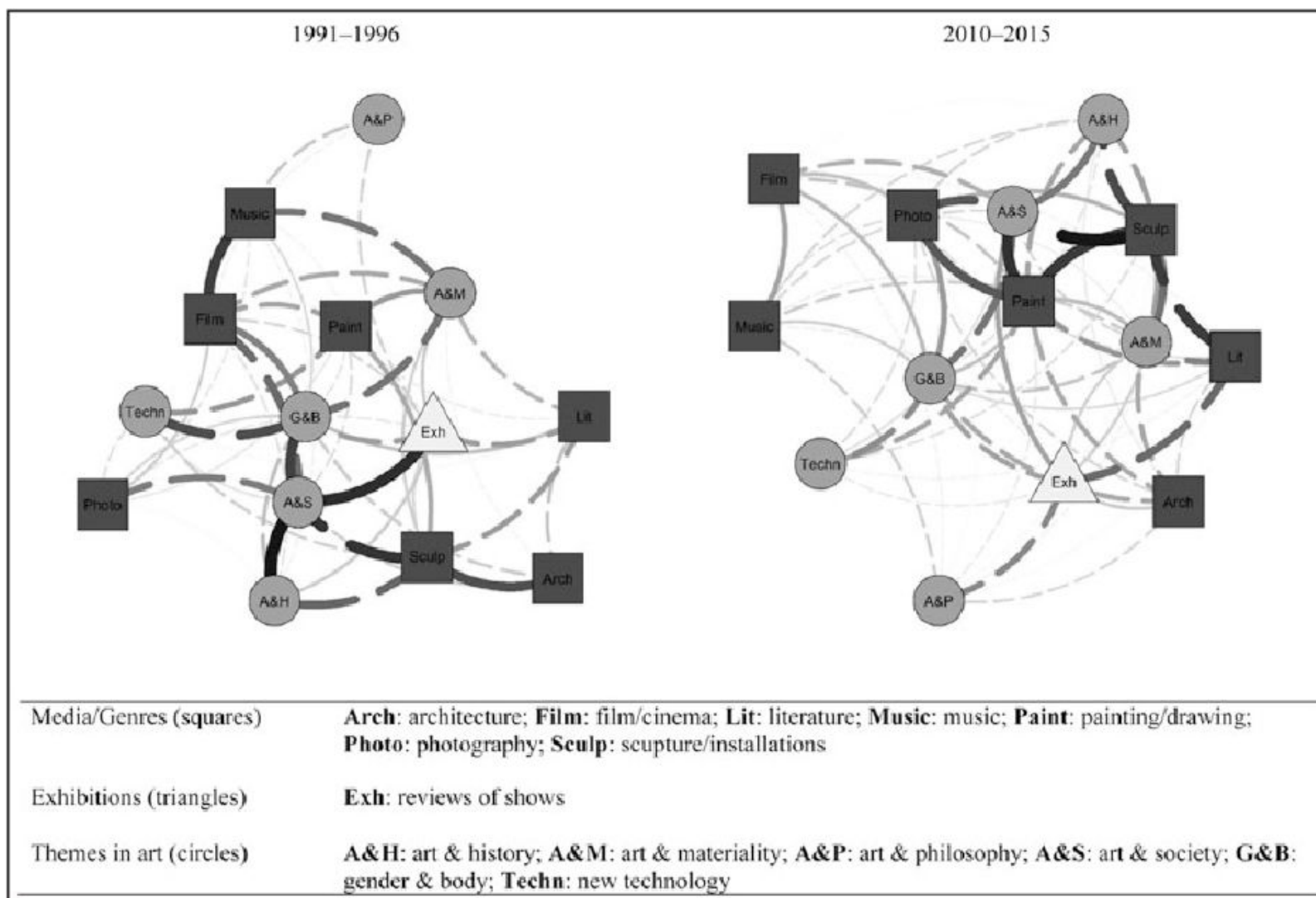
https://repub.eur.nl/pub/105624/2018_Roose-Roose-and-Daenekindt_Cultural-Sociology.pdf

“Our analysis shows that despite evolutions in the field of contemporary art – such as the ‘social turn’, in which contemporary art starts paying more attention to social forms and content – the prevalence of certain topics in contemporary art criticism has barely changed over the past 25 years.”

(Roose et al., 2018)

[ARCHEOLOGIA: <https://electricarchaeology.ca/>]

APPLICAZIONE: BENI CULTURALI



APPLICAZIONE: STUDI LETTERARI

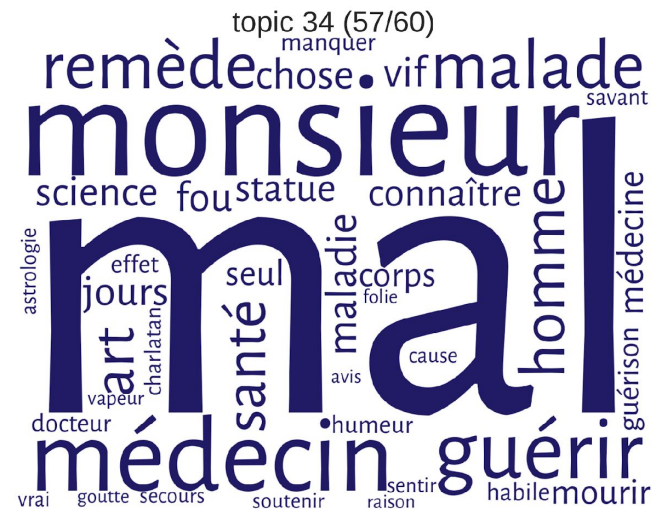
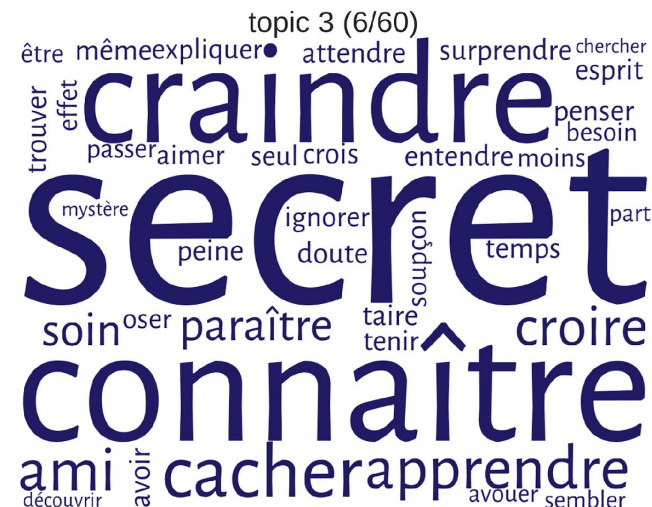
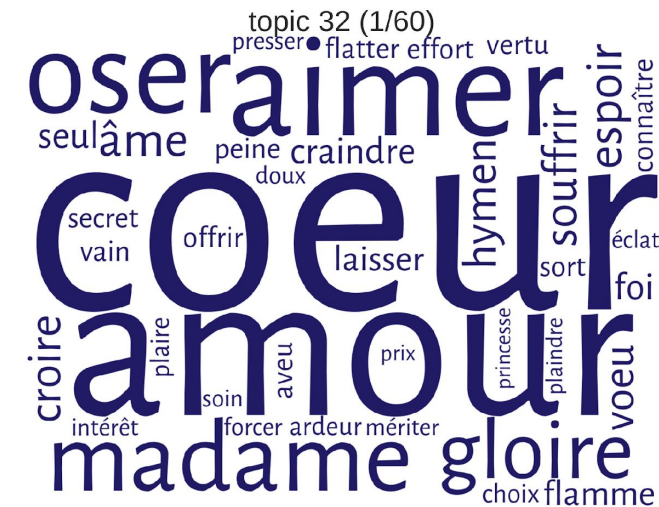
Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama

<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)

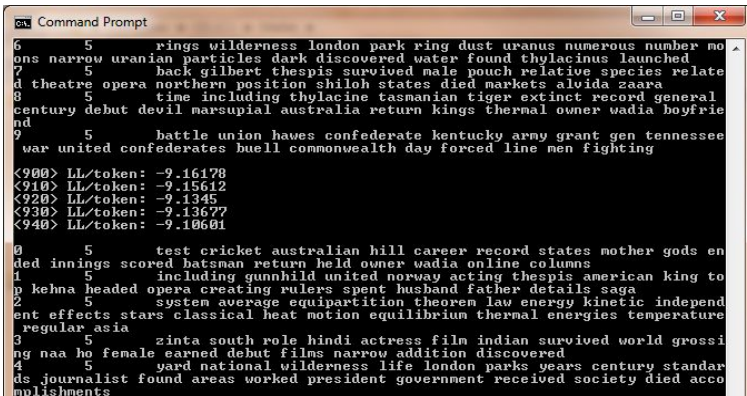
APPLICAZIONE: STUDI LETTERARI



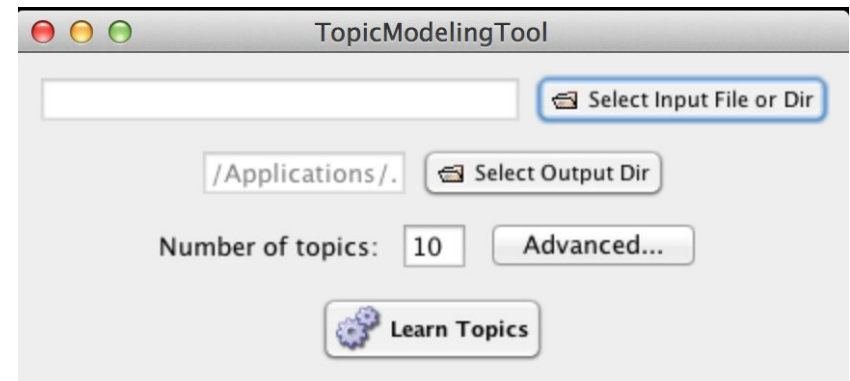
(Schöch, 2017)

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
9      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

- URL: <https://mimno.infosci.cornell.edu/jsLDA/>
- Formato dati di input
 1. testo: doc_ID tab label tab text
 2. lista stopwords: una parola per riga

Esempi di testi:

```
1   introduzione   «L'Historia si può veramente
deffinire una guerra illustre contro il Tempo, perchè
togliendoli di mano gl' anni suoi prigionieri, anzi già
fatti cadaueri, li richiama in vita, li passa in
rassegna, e li schiera di nuouo in battaglia. Ma gl'
```

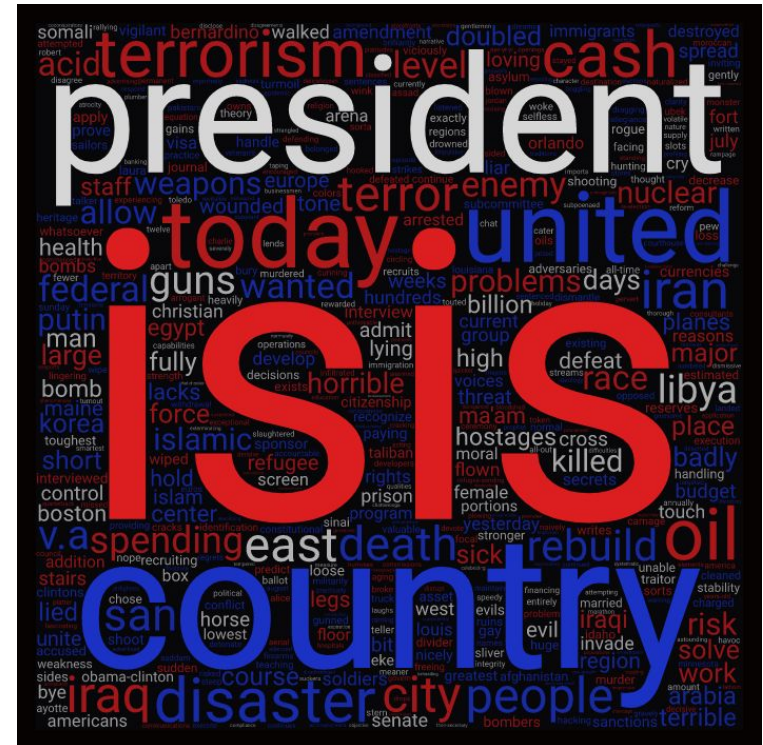
```
Trump_2016-07-22   2016-07-22   Thank you very much. We
had an amazing convention. That was one of the best. I
think it was one of the best ever. In terms -- in terms
of enthusiasm, in terms of I think what it represents,
getting our word out. Ivanka was incredible last night.
```


TOPIC MODELING: DEMO ONLINE

1. Aprire un browser (no Explorer) e andare su:
<https://mimno.infosci.cornell.edu/jsLDA/jslda.html>
2. Cliccare su “Choose File” per l’opzione [Document Upload](#) e caricare [Trump.txt](#)
3. Cliccare “Choose File” per l’opzione [Stoplist Upload](#) e caricare [stopwords-en.txt](#)
4. Cliccare su [Load](#)
5. Cliccare su [Vocabulary](#) per pulire il dizionario
6. Scegliere [15 topics](#) e cliccare su [Run 50 iterations](#)
7. Quando i topic sembrano stabili andare nella sezione [Download](#) e scaricare i veri file di output
8. 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



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)



GRAZIE!

Email: rachele.sprugnoli@unicatt.it

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

