Topic Modeling

A cura di:

Davide Locci

Alessandro Piroddi

Web Analitycs e Analisi Testuale

Prof. Marco Ortu

Indice

Obiettivi Progetto

- *Analizzare le tendenze della stampa italiana e statunitense nell'ultimo triennio*
- *Utilizzare come fonte i tweet dei loro account ufficiali*
- *Impatto sulle notizie del verificarsi di due grandi eventi covid-19 e invasione russa in Ucraina*

Testate Giornalistiche

Italiane: la Repubblica, CORRIERE DELLA SERA, il Giornale.it

Inglesi: WSJ, The New York Times

Periodo temporale

Pre-covid: 01-07-19 al 31-12-19

Covid: 01-01-20 al 31-01-22

Guerra: 01-02-22 al 17-05-22

Tweets Scraping

- *Rate limits*
- *Richiesta formale*
- *7 giorni antecedenti l'estrazione*

- *No Rate limits*
- *Nessuna richiesta formale di accesso ai token*
- *periodo antecedente in linea con il progetto*

File.csv "grezzi" composti da 6 colonne:

- *testo*
- *data*
- *nome giornale*
- *lingua*
- *numero likes*
- *numero retweets*

Tools

pandas

SNSCRAPE

Tweets Cleaning

Le fasi

- ***Pulizia generale***
- ***Tokenizzazione***
- ***Rimozione Stopwords***
- ***Lemmatizzazione***

Tools

pandas

spaCy

NumPy

Tweets Cleaning

Lemmatizzazione

- Singolari, plurali
- maschili, femminili

→ Lemma

Verbi → Infinito

Esempi

Amico, amica, amici, amiche	Amico
Friend, friends	Friend
Vinci	Vincere
Wins	Win

Stemming

Parola → Radice sintattica

Esempi

Doing	Do
National	Nation
Witnesses	Witness

Tools

pandas

spaCy

NumPy

Preliminar Analysis

- *Individuazione top words*
- *Trend temporale occorenze di parola*
- *Trend temporale tweet,retweet,like*

Tools

pandas

NLTK

plotly

Prettytable

matplotlib

Preliminar Analysis

Individuazione top words

Qualche esempio

Word	Count
trump	3073
president	3040
write	1839
people	1794
company	1253
house	1047

Word	Count
coronavirus	12009
president	10225
people	9745
trump	8631
pandemic	7368
write	7344

Word	Count
ukraine	2786
russia	2141
russian	1752
president	1485
people	1285
war	1050

Word	Count
salvini	318
pd	183
conte	179
m5s	138
migrante	124
dimaio	103

Word	Count
italia	2294
covid	1590
conte	1409
italiano	1291
vaccino	1052
presidente	1031

Word	Count
ucraina	866
russo	780
guerra	594
russia	579
putin	574
italia	533

Word	Count
trump	3073
president	3040
write	1839
people	1794
company	1253
house	1047

Word	Count
coronavirus	12009
president	10225
people	9745
trump	8631
pandemic	7368
write	7344

Word	Count
ukraine	2786
russia	2141
russian	1752
president	1485
people	1285
war	1050

Tools

pandas

NLTK

plotly

Prettytable

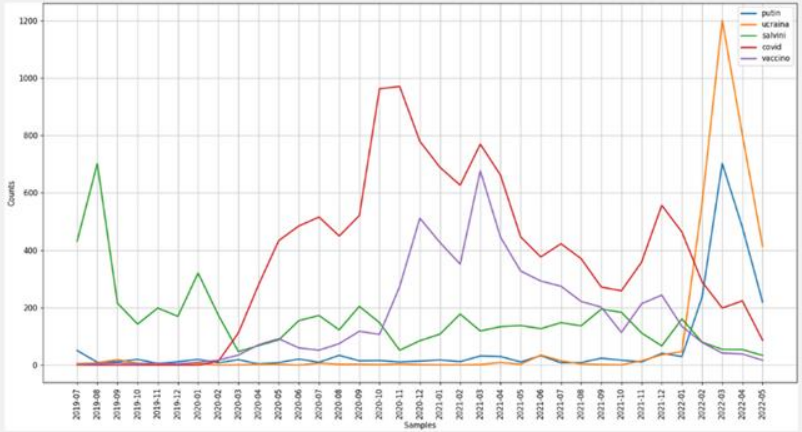
matplotlib

Preliminar Analysis

Trend temporale occorenze di parola

Parole target

- putin
- ucraina
- salvini
- covid
- vaccino



Tools

pandas

NLTK

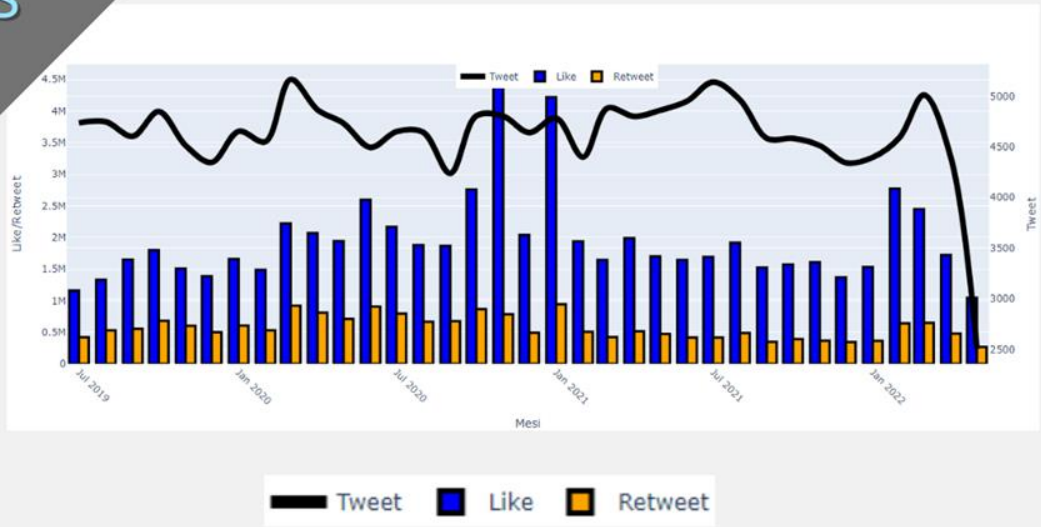
plotly

Prettytable

matplotlib

Preliminar Analysis

Trend temporale tweet, retweet, like



Tweet Like Retweet

Tools

pandas

NLTK

plotly

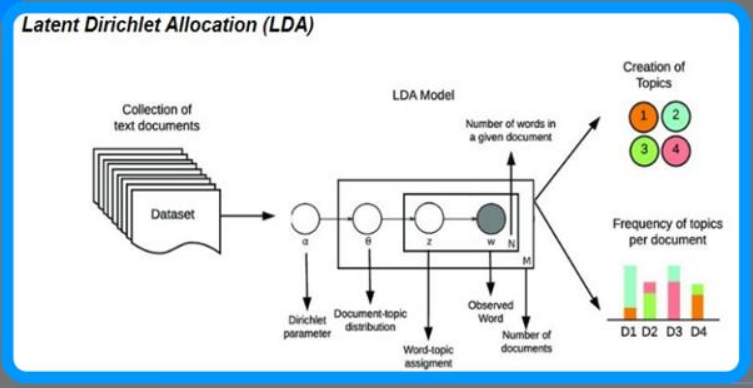
Prettytable

matplotlib

Topic Modeling
Definizione

Nel natural language processing, e più in generale nell'apprendimento statistico, gli algoritmi di topic modeling sono modelli statistici non supervisionati che cercano di associare un argomento ad un documento appartenente ad una collezione di documenti. Un problema di topic modeling non è nient'altro che un problema relativo a determinare quali sono gli argomenti trattati all'interno di un corpus di documenti

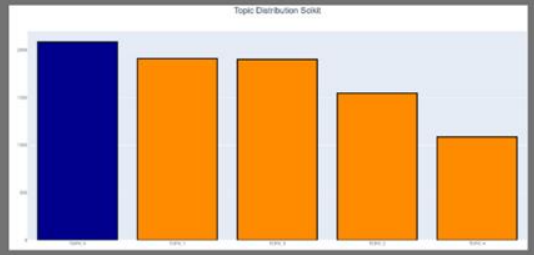
Topic Modeling LDA



L'algoritmo scelto per implementare la topic modeling è la LDA.
E' un algoritmo probabilistico che estrae a partire da un corpus di documenti, i vari topic di quei documenti. Definiti un corpus e il numero di topic si genererà una lista di parole chiave ed ogni topic sarà la combinazione chiave delle parole chiave individuate

Topic Modeling risultati Scikit

	Topic0	Topic1	Topic2	Topic3	Topic4	dominant_topic
Doc0	0.100000	0.100000	0.100000	**0.600000**	0.100000	3
Doc1	0.030000	0.030000	0.030000	**0.730000**	**0.180000**	3
Doc2	**0.200000**	0.030000	**0.700000**	0.030000	0.030000	2
Doc3	0.030000	**0.370000**	0.030000	**0.370000**	**0.200000**	1
Doc4	**0.800000**	0.050000	0.050000	0.050000	0.050000	0
Doc5	0.040000	0.040000	**0.240000**	**0.640000**	0.040000	3
Doc6	0.050000	0.050000	**0.800000**	0.050000	0.050000	2
Doc7	0.050000	0.050000	**0.800000**	0.050000	0.050000	2
Doc8	0.070000	0.070000	0.070000	**0.730000**	0.070000	3
Doc9	0.070000	**0.730000**	0.070000	0.070000	0.070000	1
Doc10	**0.550000**	0.050000	**0.300000**	0.050000	0.050000	0



	text	topic	weight	top words
Doc 0	congedo mestruale discutere spagna	3	0.599994	putin ucraina guerra russo morire russia kiev figlio zelensky mosca
Doc 1	padre figlio campo olimpiadi sordo buone notizie edicola	3	0.734044	putin ucraina guerra russo morire russia kiev figlio zelensky mosca
Doc 2	mascherina scuola lega pressing speranza decidere scienza	2	0.700081	uccidere diretta donna covid morto italia roma presidente green pass
Doc 3	colonna autobus lasciare acciaieria azovstal video	1	0.367069	milano tornare video gas parlare sanzione segreto rischio gara russia
Doc 4	mariupol irriducibile azovstal assedio dilemma resa	0	0.798245	ucraino auto europa inter mariupol italiano vincere attacco famiglia russo
Doc 5	vergognare padre riconoscere scelta giusto	3	0.640041	putin ucraina guerra russo morire russia kiev figlio zelensky mosca
Doc 6	addio giorgio chiellini stadium abbraccio emozione standing ovation	2	0.799998	uccidere diretta donna covid morto italia roma presidente green pass
Doc 7	coming out calciatore jake daniels gay inglese attività dire	2	0.799998	uccidere diretta donna covid morto italia roma presidente green pass
Doc 8	barricati sotterraneo acciaieria resistente azovstal	3	0.732997	putin ucraina guerra russo morire russia kiev figlio zelensky mosca
Doc 9	spiagge intesa ipotesi aumentare indennizzo	1	0.733332	milano tornare video gas parlare sanzione segreto rischio gara russia
Doc 10	sparatoria california odiare taiwan immigrato cinese uniti	0	0.550317	ucraino auto europa inter mariupol italiano vincere attacco famiglia russo
Doc 11	matteo vincitore certamen pensavo impreciso farcee	3	0.733330	putin ucraina guerra russo morire russia kiev figlio zelensky mosca
Doc 12	torino gallerie italia fotografia tesore	0	0.399961	ucraino auto europa inter mariupol italiano vincere attacco famiglia russo
Doc 13	comandante battaglione azov obbediremo ordine evacuazione	1	0.733330	milano tornare video gas parlare sanzione segreto rischio gara russia
Doc 14	berlusconi sorpresa treviglio convention fi comunismo	4	0.733332	italia cambiare usa draghi euro russo chiedere storia vedere succedere

Tools

pandas

NLTK

plotly

Prettytable

matplotlib

scikit learn

Topic Modeling
risultati Gensim

```
[(0,
  '0.034*"votare" + 0.025*"stella" + 0.019*"diretta" + 0.015*"amadeus" + '
  '0.013*"italia" + 0.012*"salvini" + 0.012*"suv" + 0.010*"indicare" + '
  '0.009*"giro" + 0.008*"terzo"'),
 (1,
  '0.028*"appena" + 0.026*"morire" + 0.014*"italiano" + 0.014*"palco" + '
  '0.011*"mascherina" + 0.011*"fuga" + 0.010*"pass" + 0.010*"green" + '
  '0.010*"europeo" + 0.009*"regola"'),
 (2,
  '0.038*"gara" + 0.023*"covid" + 0.018*"vedere" + 0.010*"mattarella" + '
  '0.009*"pubblico" + 0.009*"milano" + 0.008*"cambiare" + 0.008*"febbraio" + '
  '0.008*"terra" + 0.007*"euro"'),
 (3,
  '0.054*"sanremo2022" + 0.026*"piacere" + 0.019*"esibire" + 0.016*"uccidere" '
  '+ 0.016*"serata" + 0.016*"valutazione" + 0.014*"indicare" + 0.014*"tornare" '
  '+ 0.013*"blanco" + 0.011*"mahmood"'),
 (4,
  '0.038*"sanremo" + 0.030*"canzone" + 0.017*"pandemia" + 0.014*"padre" + '
  '0.013*"festival" + 0.011*"storia" + 0.011*"voto" + 0.010*"ubriaco" + '
  '0.008*"lauro" + 0.007*"achille"')]
```

topics	n_documents
0	1735
1	2058
2	1944
3	1274
4	1521

Tools

pandas

NLTK

plotly

Prettytable

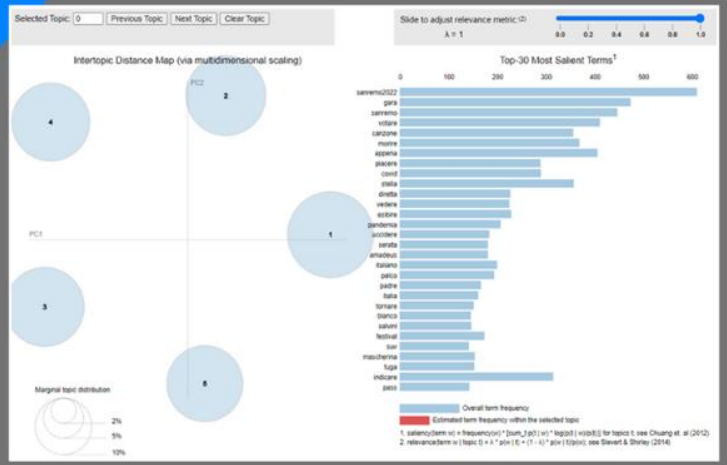
GENSIM
topic modelling for humans

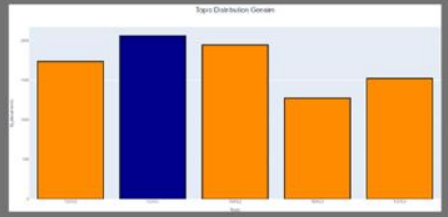
pyLDAvis

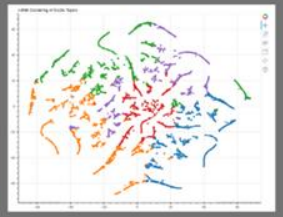
bokeh

IP[y]

Topic Modeling
risultati Gensim







Tools

pandas

NLTK

plotly

Prettytable

GENSIM
topic modelling for humans

pyLDAvis

bokeh

IP[y]

Bibliografia