DATA MINING PROJECT

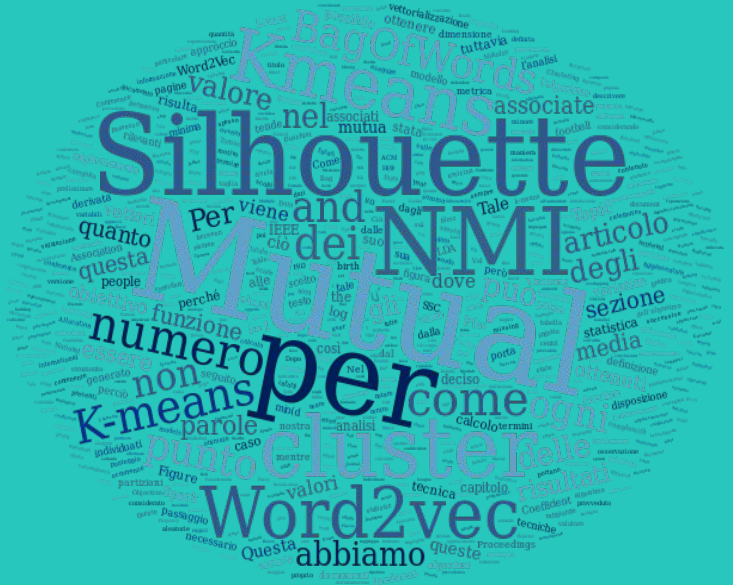
SUGGESTED PROJECT GOAL B

Cazzaro Davide	1138635
Dalla Cia Massimo	1153092
Lovisotto Enrico	1156704
Vianello Enrico	1153101

Introduzione

Obiettivi

- rappresentazione dell’input
- clusterizzabilità degli articoli
- bontà di varie tecniche di clustering
- numero ottimale di cluster
- legame tra cluster e categorie



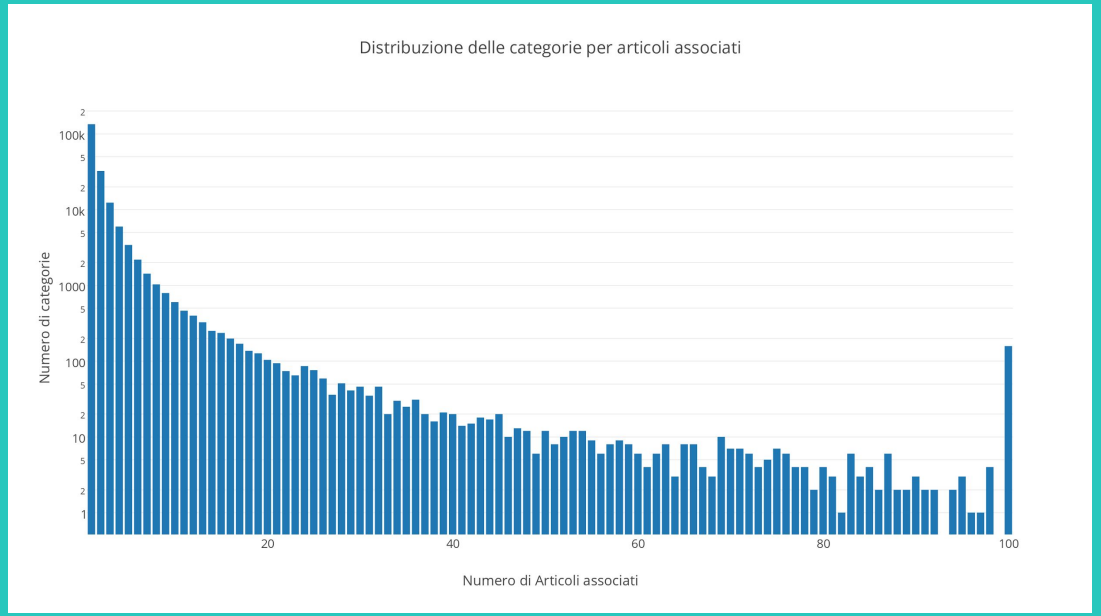
Dataset in input

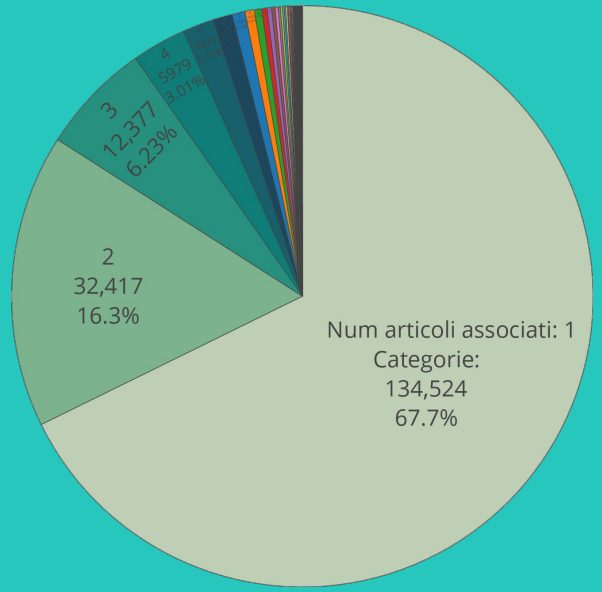
Dump di Wikipedia di 100k articoli

Per ogni articolo un JSON con

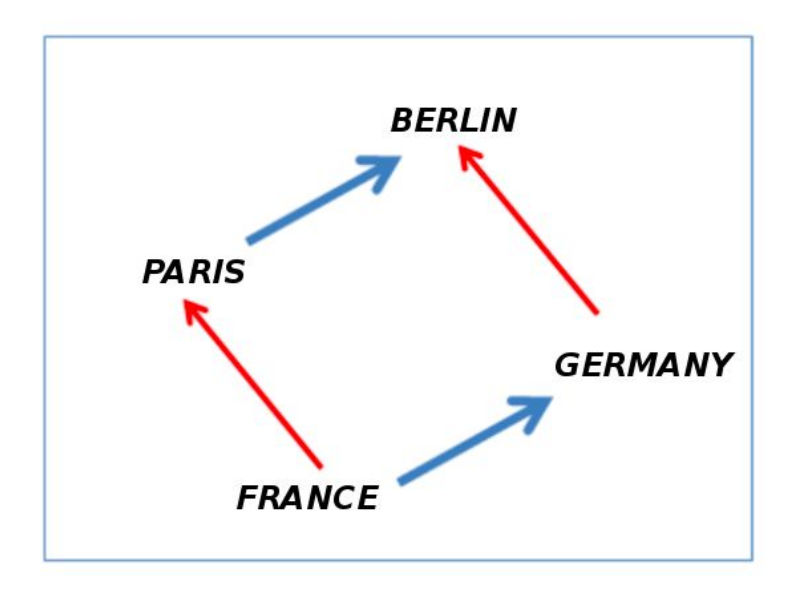
- titolo
- testo
- id
- categorie

Categorie


Distribuzione delle categorie per articoli associati
Numero di categorie
Numero di Articoli associati
100k
10k
1000
100
10
1
20
40
60
80
100


3
12,377
6.23%
4
5979
3.01%
2
32,417
16.3%
Num articoli associati: 1
Categorie:
134,524
67.7%

Rappresentazione: Word2Vec

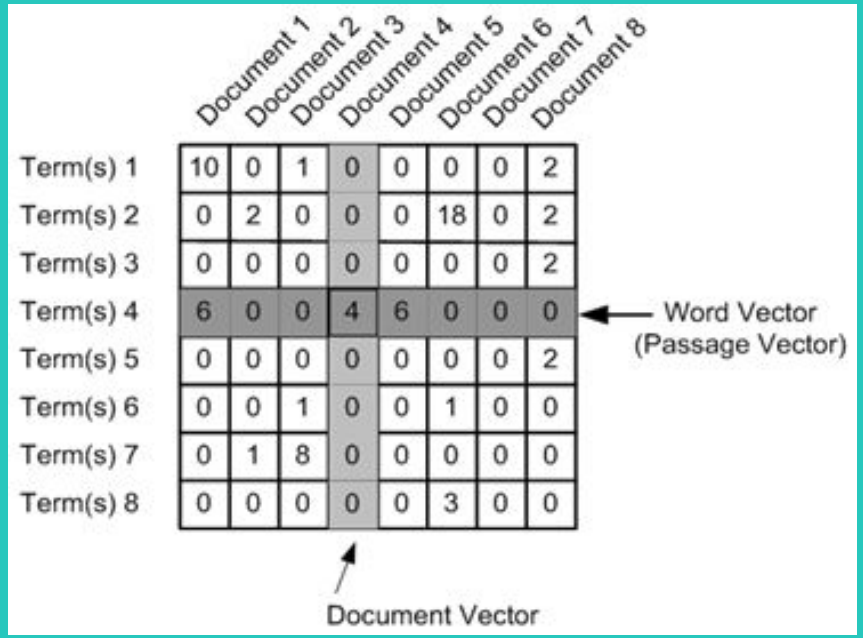


Rappresentazione: Bag of Words

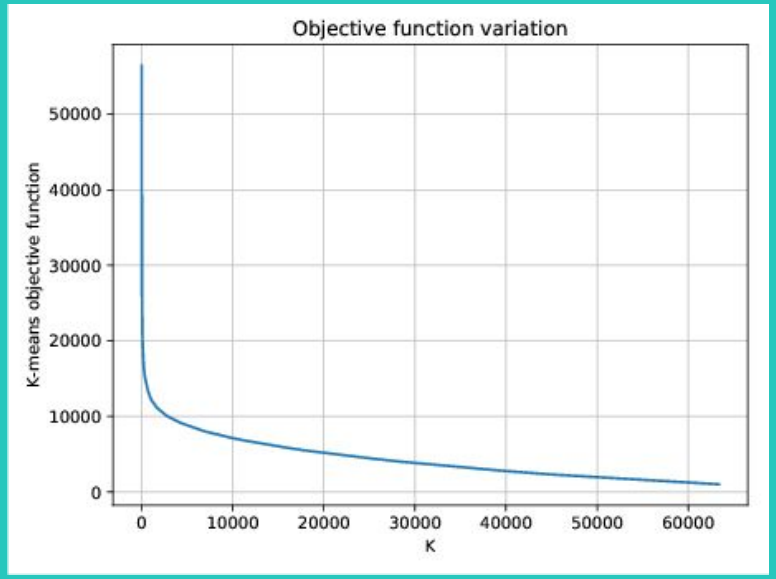
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

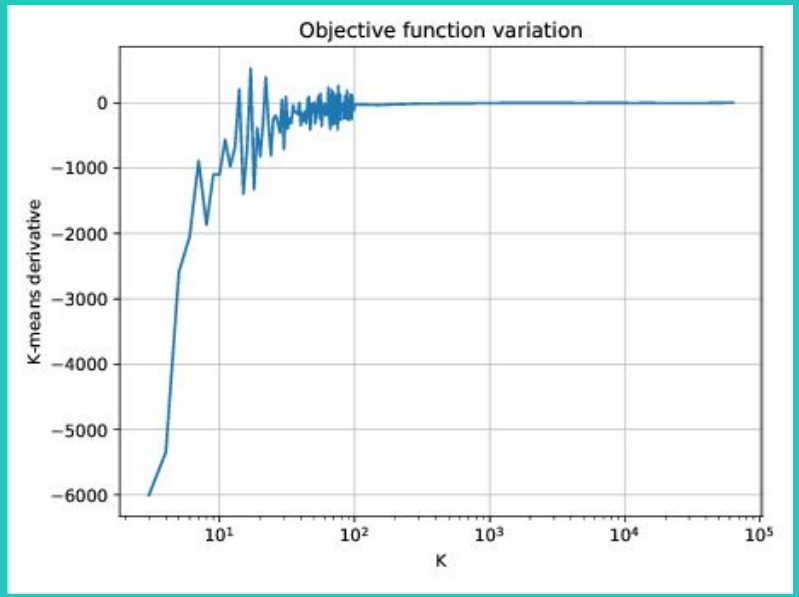
$tf_{i,j}$ = numero di occorrenze di i in j
df_i = numero di documenti che contengono i
N = numero totale di documenti

Matrice 3000 x N



Clustering: K-means





Clustering: LDA

Modello statistico generativo

- Documenti considerati mixture di topic
- Topic caratterizzati da insiemi di termini

Iterazioni: 100

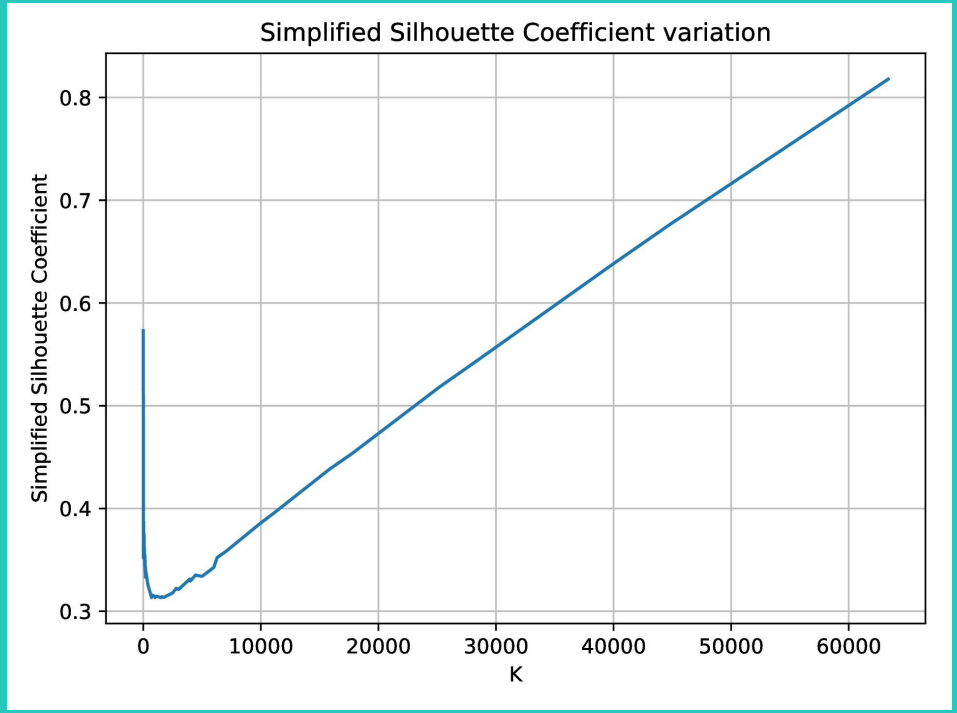
k = 80

Topic 1		Topic 2		Topic 3		Topic 4	
word	**weight**	**word**	**weight**	**word**	**weight**	**word**	**weight**
king	0.0779	attack	0.0446	bank	0.0392	system	0.0259
prince	0.0333	kill	0.0403	company	0.0367	computer	0.0239
queen	0.0325	fight	0.0402	purchase	0.0284	user	0.0224
emperor	0.0295	battle	0.0313	sell	0.0271	software	0.0220

Simplified Silhouette

$$SSC = \frac{b-a}{max(a,b)} = \frac{b-a}{b}$$

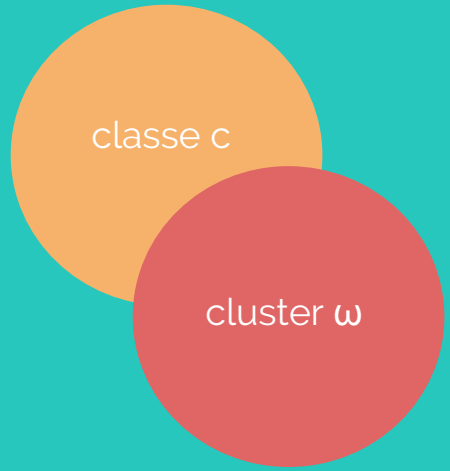
$$S = \frac{min(d) - d}{max(d,\ min(d))} \xrightarrow[K \to n]{} 1$$



Normalized Mutual Information

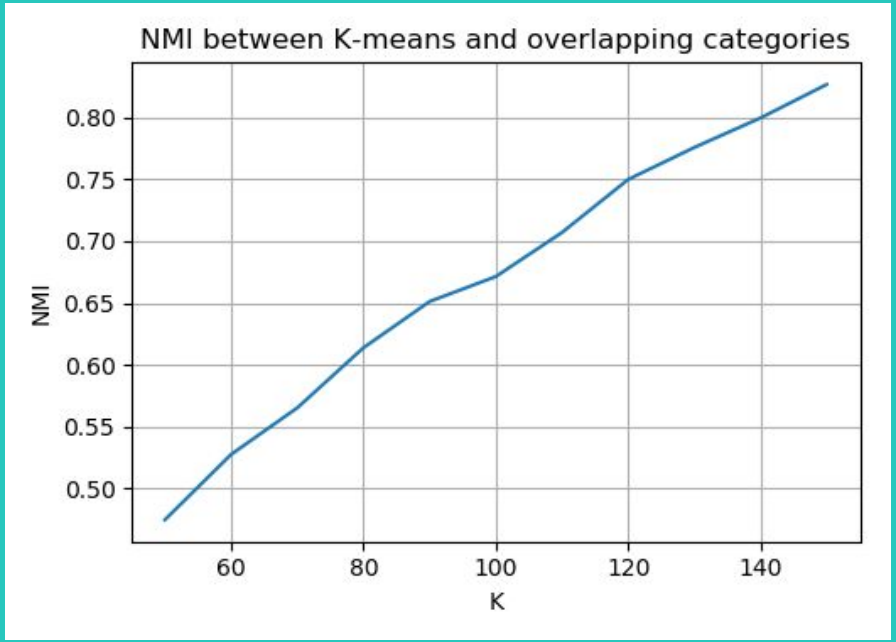
Quanto ci “dice” sulle categorie di un articolo la sua appartenenza ad un cluster?

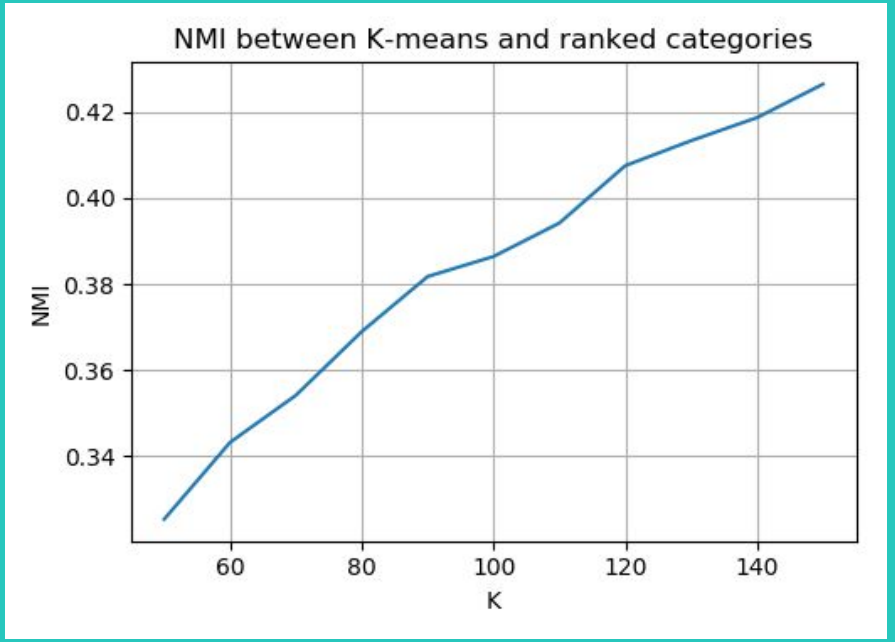
Quanto è stretto il legame tra c e ω?



$$NMI(\Omega, \mathbb{C}) = \frac{I(\Omega, \mathbb{C})}{[H(\Omega) + H(\mathbb{C})]/2}$$

Normalized Mutual Information





Conclusioni

- Word2Vec è rapida e descrive bene il dataset
- bag-of-words è molto più dispendiosa

- la maggioranza delle categorie non sono significative
- Hopkins rivela una buona clusterizzabilità del dataset
- K-means suggerisce K=100, confermata da NMI e Silhouette

- K-means è l'algoritmo più adatto al problema
- GMM, LDA, single-linkage hanno una complessità troppo elevata
- LDA permette tuttavia una veloce caratterizzazione dei cluster