

Data Mining e Business Intelligence per le aziende

Business Intelligence per Big Data (01RLBNG)

Lorenzo Bonasera (269702)
Simona Maria Borrello (277789)

July 4, 2020



**POLITECNICO
DI TORINO**

- Introduzione
- Obiettivi
- Descrizione e caratterizzazione del dataset
- Pre-processing dei dati testuali
- Clustering
- Regole di associazione
- Applicazioni business

Introduzione

Gli strumenti forniti dalla *Data Science* oggiogiorno giocano un ruolo fondamentale nel supporto delle aziende in ambito analitico e decisionale, a prescindere dal campo di applicazione. Attraverso il *data mining* è possibile estrarre informazioni e conoscenza fruibili ai fini di *business intelligence*, fornendo assistenza alla decisione strategica aziendale. Tra gli strumenti a disposizione dell'analista sono presenti quelli relativi al *text mining*, descritto come il processo di estrazione di informazioni qualitative e quantitative a partire da dati testuali.

Nel caso in oggetto, questi strumenti vengono applicati per derivare informazioni utili da un dataset contenente documenti di testo, in particolare notizie giornalistiche rese disponibili dalla compagnia *Thomson Reuters* riguardanti stock di aziende quali *Amazon*, *Google*, ecc. che hanno influito sull'andamento delle quotazioni in borsa delle stesse.



Obiettivi

- Pre-processing e caratterizzazione del dataset
- Confronto tra le varie tecniche di *data mining*
- Suddivisione del dataset in cluster
- Estrazione e selezione dei contenuti più rilevanti
- Applicazione business della conoscenza estratta

Descrizione e caratterizzazione del dataset

Il dataset in analisi è formato da 1965 documenti di tipo testuale, codificati tramite codifica *UTF-8* e scritti in diverse lingue. Ogni documento è provvisto dei seguenti attributi:

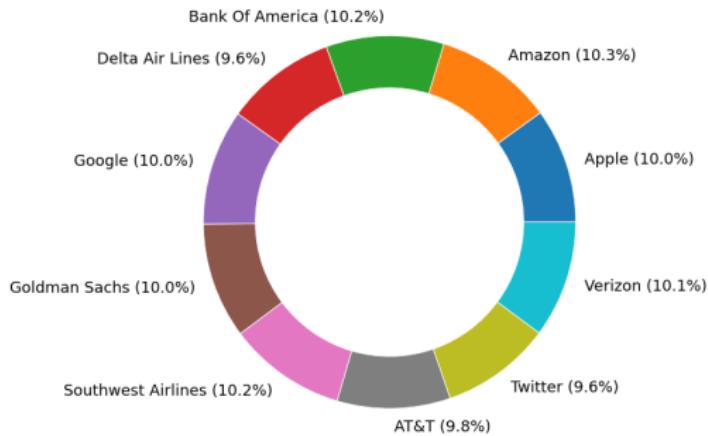
- *Data* nel formato *AAAA-MM-GG*
- Numero di parole *positive* per le quotazioni in borsa
- Numero di parole *negative* per le quotazioni in borsa
- *Stock* di riferimento

Ai fini delle analisi è di fondamentale importanza che i documenti riportino una sola lingua: essendo l'inglese la più diffusa, sono filtrati tutti i testi contenenti lingue diverse attraverso l'operatore di *RapidMiner Text Vectorization*, ottenendo 1933 documenti. L'attributo *data* è utilizzato per suddividere i documenti appartenenti al dataset in base all'anno e al mese, riducendone la granularità.



Descrizione e caratterizzazione del dataset (1)

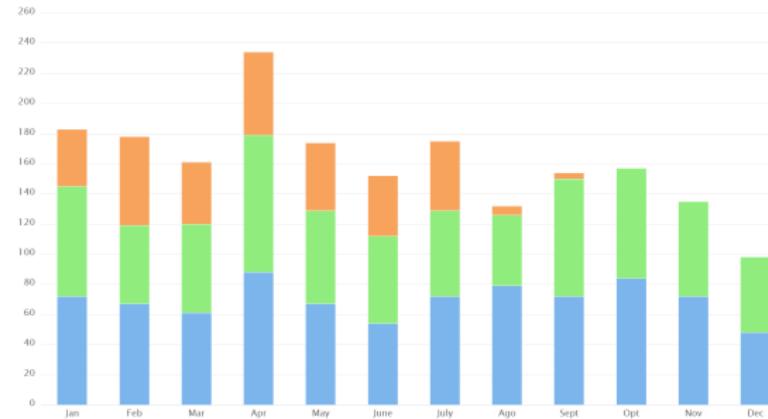
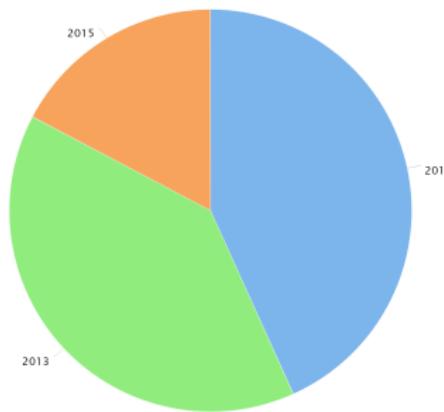
Il seguente grafico mostra la distribuzione dell'attributo *stock* all'interno del dataset:



Circa il 20% dei records riguarda news del settore *Finance* (Bank of America, Goldman Sachs), un ulteriore 20% concerne l'ambito *Transportation* (Delta Air Lines, Southwest Airlines), mentre il restante interessa il settore *Computer and Technology*. [4]

Descrizione e caratterizzazione del dataset (2)

Le news fanno riferimento agli anni 2011, 2013 e 2015. I documenti riferiti all'anno 2015 costituiscono il $\approx 17.3\%$ del dataset, mentre quelli relativi agli altri anni sono distribuiti quasi equamente.



Pre-processing dei dati testuali

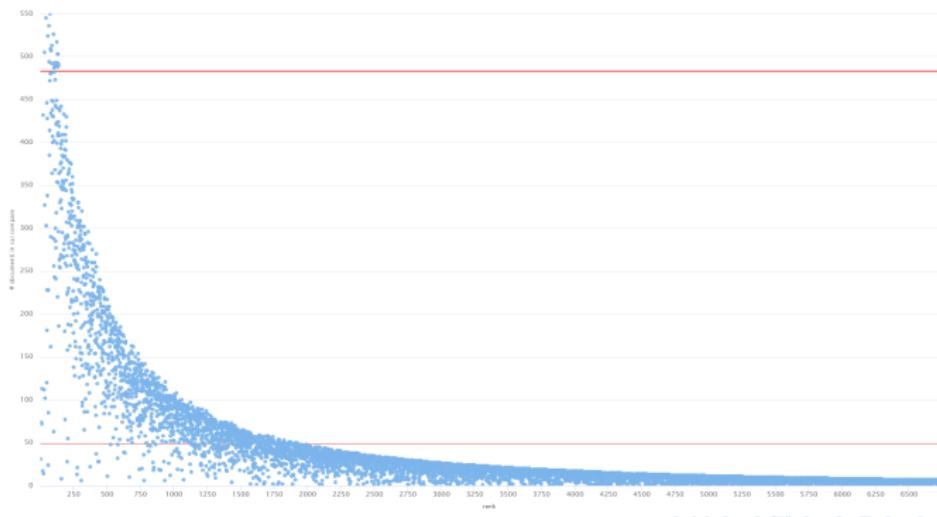
La fase di *pre-processing* è necessaria per poter trasformare i dati testuali tramite una rappresentazione chiamata *feature vector*, adatta per poter applicare i metodi e gli algoritmi di *data mining*. Essa è stata realizzata mediante l'applicazione della seguente sequenza di operatori in *RapidMiner*:



Il numero di token che si ottiene è pari a **27041**: si è dunque ritenuto opportuno applicare una strategia di *pruning*, effettuata rimuovendo i token presenti in meno del **2.5%** e più del **25%** del totale dei testi elaborati. L'analisi è dunque limitata a **1447** token. Sono successivamente assegnati dei pesi tramite la TF-IDF, funzione di peso adatta a contenuti di carattere eterogeneo.

Pre-processing dei dati testuali (1)

La scelta dei valori di *pruning* è motivata dalla distribuzione delle parole nei documenti in questione. Nel seguente grafico sono riportati i token ordinati in senso decrescente in base alla frequenza nell'insieme totale dei testi, assegnando come valore il numero di documenti diversi in cui ogni token è presente. Si nota come l'andamento sia inversamente proporzionale:



Pre-processing dei dati testuali (2)

Dal precedente grafico si evince come la maggior parte delle parole sia contenuta in meno di 50 documenti: il numero di token presenti sotto la soglia inferiore sono più di **25000**, mentre quelli sopra la soglia superiore sono pari a **75**. La seguente tabella riporta i primi dieci token ordinati per occorrenza totale, dei quali solo *bmo* viene incluso dal pruning:

word	# documenti	Occorrenze	rank
said	1606	8533	1
inc	955	6494	2
conf	31	6131	3
reuter	1928	5268	4
compani	1267	4998	5
percent	1008	4971	6
bmo	74	4520	7
year	1352	4392	8
corp	661	3883	9
bank	655	3416	10

Pre-processing dei dati testuali - Analisi del linguaggio

Definiamo, per ogni documento i , l'estensione lessicale ε_i come il rapporto tra il numero di token **distinti** e il numero di token **totali** nel documento. Questo indice riflette la ricchezza del lessico contenuto in un testo: più alto è il valore, maggiore è la varietà del vocabolario usato. Sono riportati i valori di ε_i per i primi 11 documenti :

0	1	2	3	4	5	6	7	8	9	10
0.676976	0.558824	0.80597	0.638182	0.707006	0.609977	0.594306	0.729927	0.570815	0.587444	0.60177

Calcolando la media tra tutti i documenti, pesata nel seguente modo:

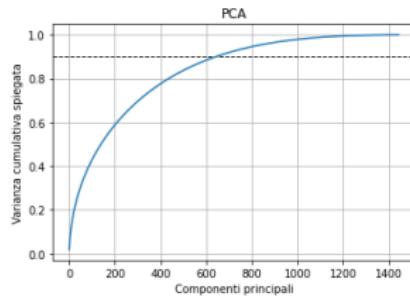
$$\varepsilon_{mean} = \frac{\sum_{i=1}^n \varepsilon_i \cdot \ell_i}{\sum_{i=1}^n \ell_i}$$

dove ℓ_i è la lunghezza totale dell' i -esimo documento. L'estensione lessicale media su tutto il dataset è $\varepsilon_{mean} \approx 0.57$.



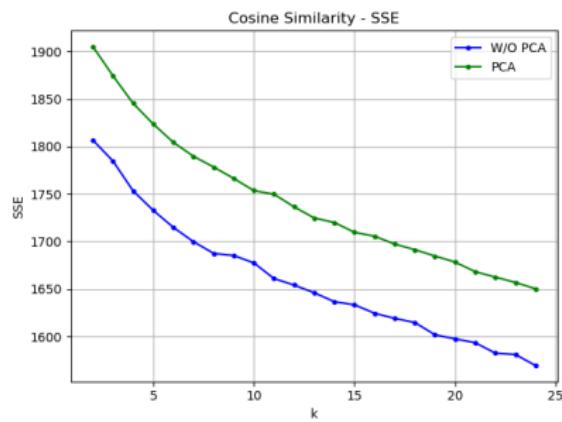
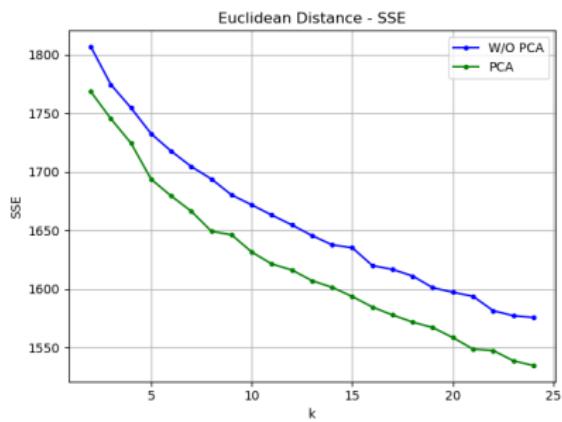
Clustering

Trattandosi di testi di carattere eterogeneo, risulta di fondamentale importanza raggruppare i documenti in *cluster*, in modo da poter analizzare separatamente ogni categoria. Tra i metodi di clustering sono considerati e confrontati i seguenti algoritmi: **K-Means**, **DBSCAN** e **Agglomerative Clustering**. Per studiare gli effetti della *dimensionality reduction* sulla qualità dei cluster, ogni sessione di analisi prevede il confronto tra l'utilizzo o meno di uno step di **PCA** che limiti il numero di attributi a **1000**: questa scelta è motivata dal seguente grafico, che mostra come, riducendo il numero di token al $\approx 71.43\%$, la varianza spiegata sia del $\approx 98\%$. Si è scelto di includere soltanto i token come attributi per il clustering.



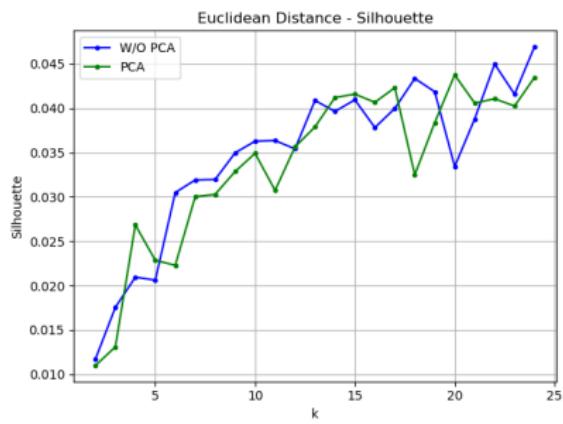
Clustering - K-Means

Applicando l'algoritmo **K-Means**, è necessario stimare il miglior valore di K : il primo approccio utilizzato è quello di tracciare un grafico dell'andamento del *SSE* al variare di K , in cerca di un ginocchio (knee approach). Come metrica viene presa in considerazione **euclidean distance** e come similitudine **cosine similarity**, indicata in caso di dati testuali. I risultati ottenuti tramite Python sono rappresentati dai seguenti grafici:



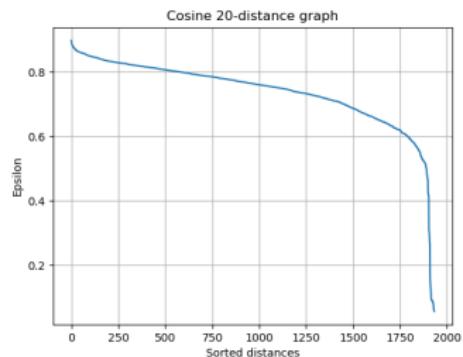
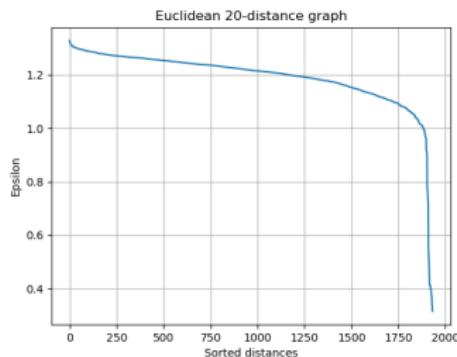
Clustering - K-Means (1)

Dai grafici ottenuti, è evidente come nessun valore di K riporti, in entrambi i casi, l'andamento desiderato: l' SSE diminuisce quasi linearmente all'aumentare del numero di cluster formati. Questo comportamento rende l' SSE poco adatto come metrica per individuare il valore di K , pertanto viene scelta la *Silhouette* come metrica di preferenza, ottenendo i seguenti grafici:



Clustering - DBSCAN

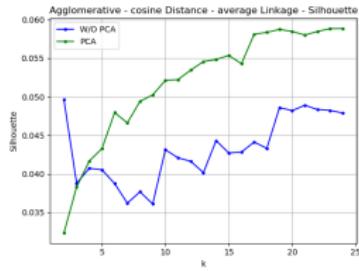
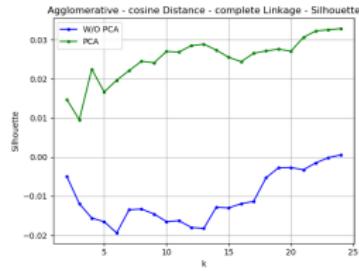
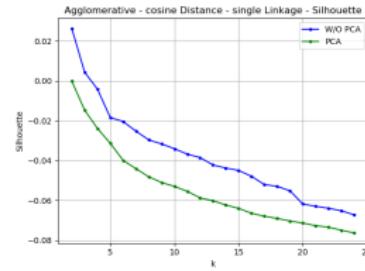
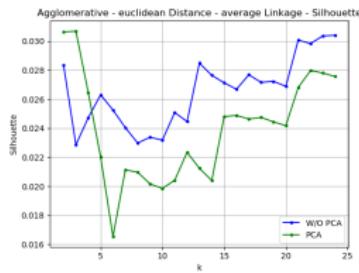
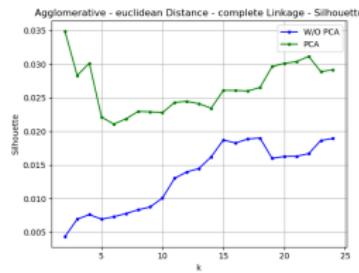
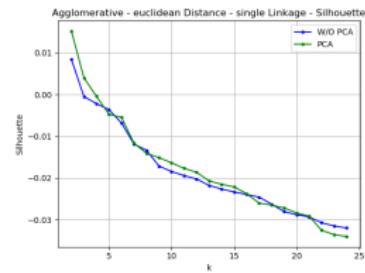
Applicando **DBSCAN** è necessario stimare i valori del parametro *MinPoints* e del rispettivo ϵ : la ricerca è effettuata tramite **grid search**, separatamente per la metrica euclidea e la similitudine coseno, cercando un possibile gomito ad ogni valore di *MinPoints* da testare. La migliore coppia è scelta in base ai valori ottenuti dalla Silhouette. Per *MinPoints* = 20:



scegliendo rispettivamente $\epsilon = 1.0$ e $\epsilon = 0.5$. Si nota come la distanza euclidea necessiti di un valore di ϵ maggiore.

Clustering - Agglomerative

Per applicare l'algoritmo di tipo gerarchico **Agglomerative Clustering** nelle sue tre forme (*Single link*, *Complete Link*, *Average Link*), si cerca il miglior valore di cluster formati in base alla *Silhouette* ottenuta:

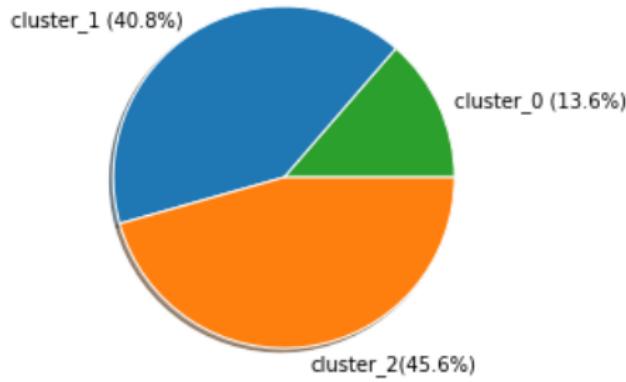


Clustering (1)

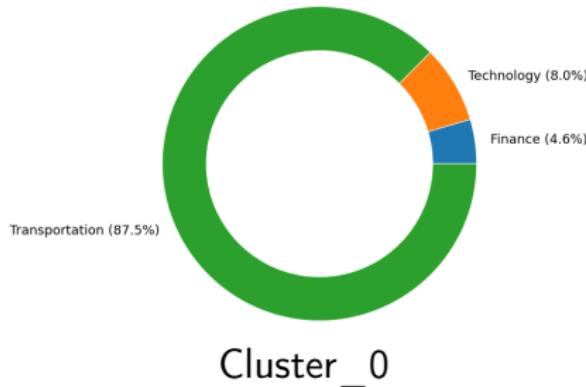
- I risultati ottenuti tramite **DBSCAN** risultano essere insufficienti, con meno di 4 cluster creati di cui uno ricoprente più del 98% dei dati: l'algoritmo si è mostrato inefficace nel contesto in analisi
- Applicando l'**Agglomerative Clustering** si osserva come *Single Linkage* abbia le peggiori performance, a causa della sua sensibilità al rumore, mentre *Average Linkage* riporta prestazioni mediamente superiori, essendo un buon compromesso tra robustezza al rumore e bias.
- Si decide dunque di optare per il **K-Means** con la **cosine similarity**, **senza** uno step di PCA e scegliendo **9** come miglior valore di K in base al grafico della **Silhouette**.

Clustering - 3-Means

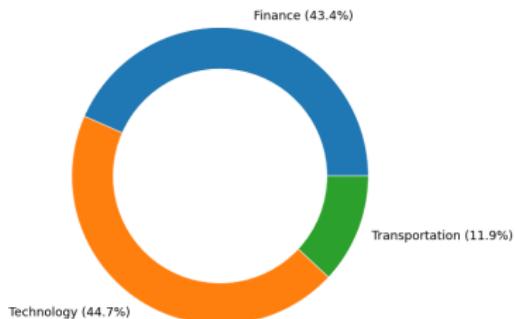
Viene analizzato il clustering ottenuto tramite **K-Means** con $K = 3$: tale scelta è indipendente dalle analisi valutate in precedenza, ed è dettata dal fatto che gli stock appartengono a 3 settori distinti (*Finance*, *Transportation*, *Computer & Technology*). Si nota come le dimensioni dei cluster ottenuti rispecchino le dimensioni dei settori che caratterizzano il dataset: ad esempio il cluster _0, il più piccolo dei tre, corrisponde al settore *Transportation*, che costituisce solo il 20% dell'intero dataset:



Clustering - 3-Means (1)



Cluster_0



Cluster_1



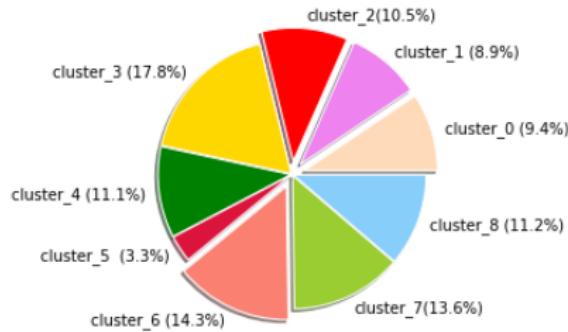
Cluster_2

Clustering - Caratterizzazione

Analizzando i cluster è possibile individuare un settore per ognuno di essi:

Cluster_0	Cluster_1	Cluster_2	Cluster_6
Telecomunicazioni	Tecnologia mobile	Trasporti aerei	Internet

Cluster_3	Cluster_4	Cluster_5	Cluster_7	Cluster_8
Politico-economico	Bancario	Meeting	Finanziario	Internazionale



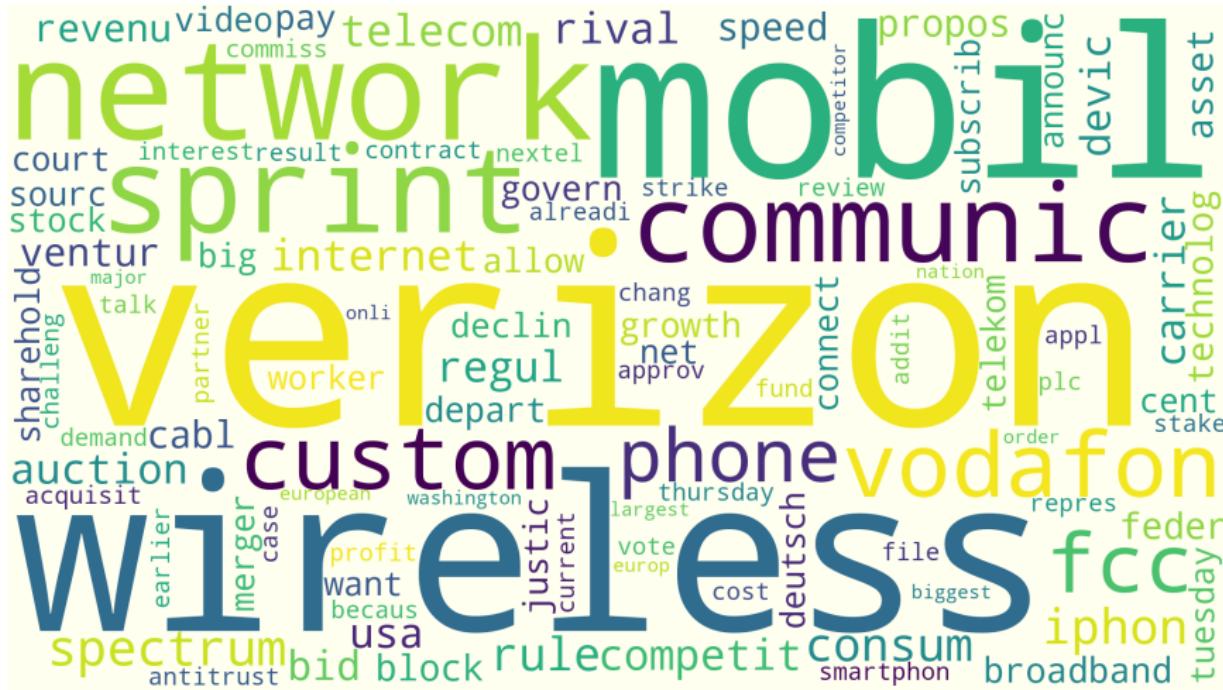
Clustering - Caratterizzazione (1)

Attraverso le prime tre componenti ottenute dalla PCA, è possibile rappresentare i documenti clusterizzati in tre dimensioni:



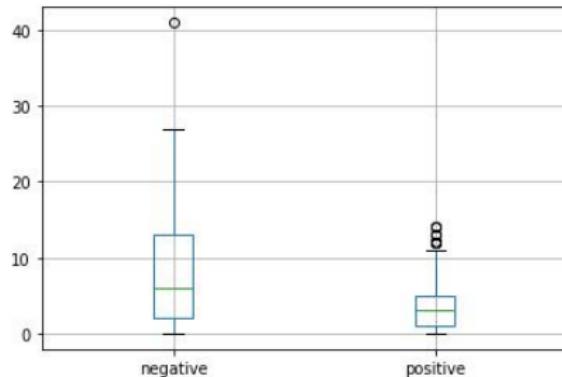
Si è deciso di restringere l'analisi ai cluster più interessanti ai fini di una possibile applicazione di *Business Intelligence*: 0, 1, 2 e 6.

Cluster 0 - Telecomunicazioni



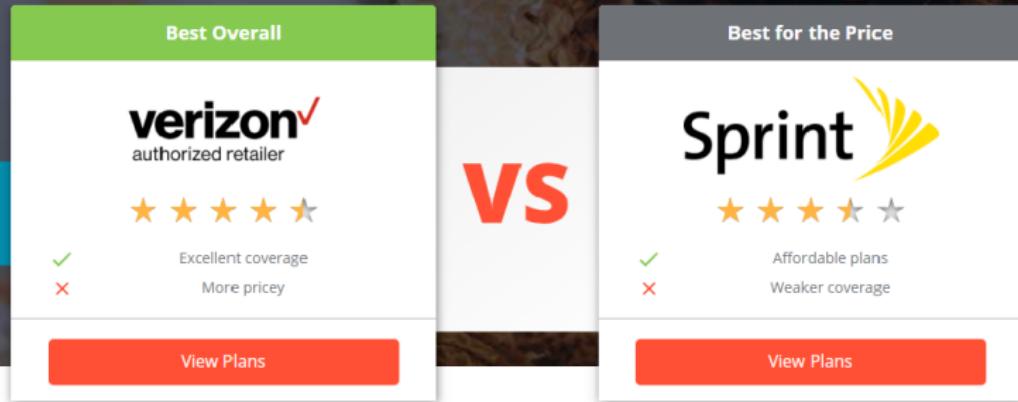
Cluster_0 - Regole di associazione

Premessa	Conclusione	Supporto	Confidenza	Lift
verizon, sprint	rival	0.214	0.650	1.714
vodafone, ventur	verizon, mobil, plc	0.253	0.708	2.477
deutsch, usa	communic, telekom	0.203	0.974	3.770

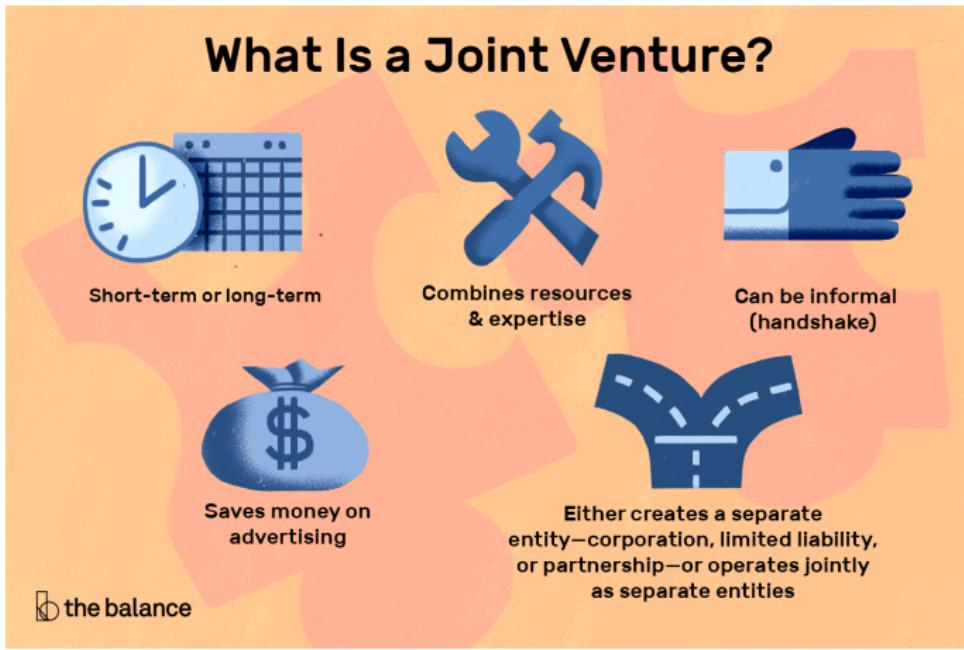


Sprint vs. Verizon Phone Plans Review 2020

The choice between Verizon or Sprint depends on whether you want reliable service or a low-cost bill every month.



Individuare possibile concorrenza [5]



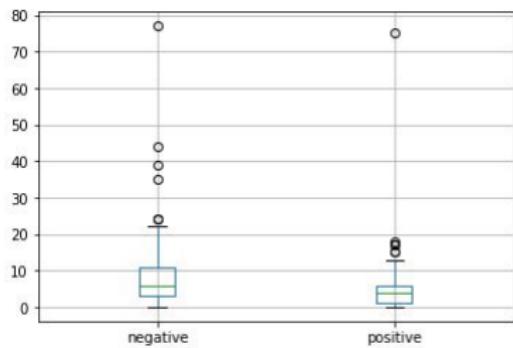
Individuare possibili joint ventures [3]

Cluster 1 - Tecnologia mobile

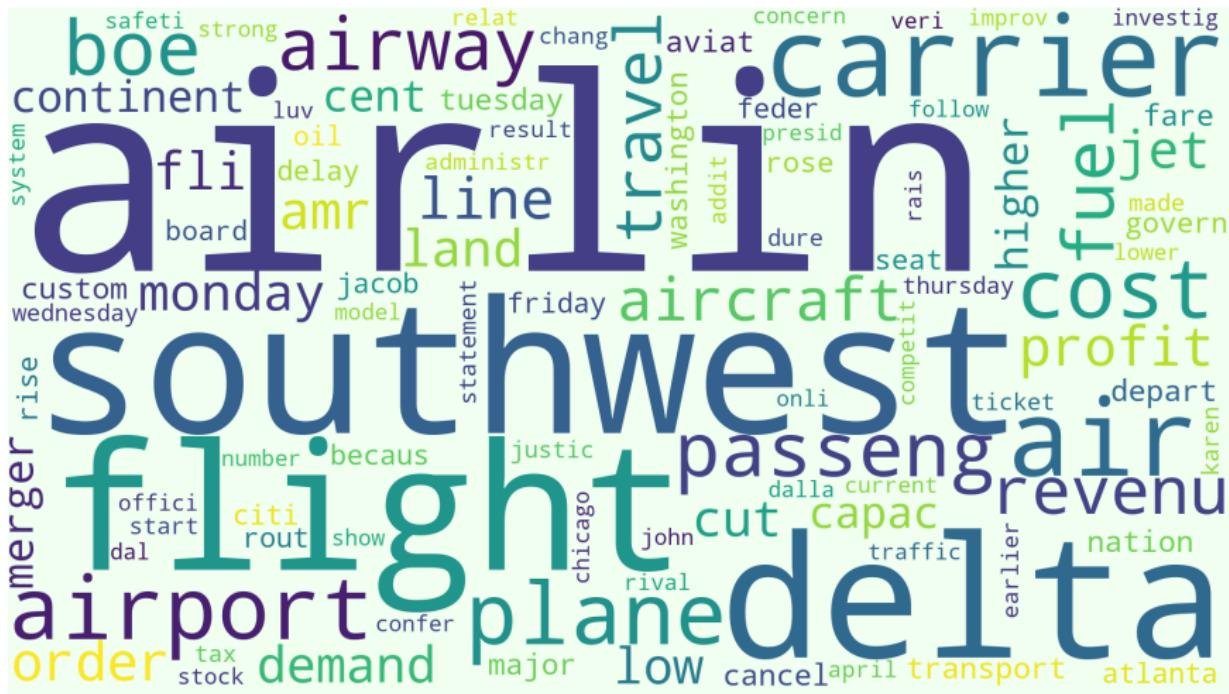


Cluster_1 - Regole di associazione

Premessa	Conclusione	Supporto	Confidenza	Lift
appl, android	iphon, googl	0.254	0.786	2.344
googl	android	0.358	0.697	1.883
android	googl	0.358	0.969	1.883

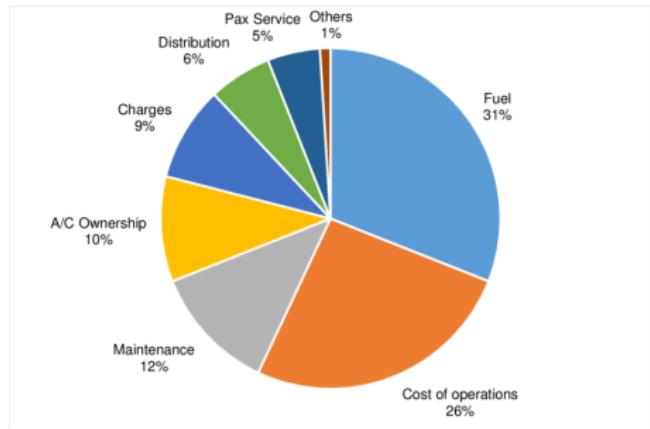
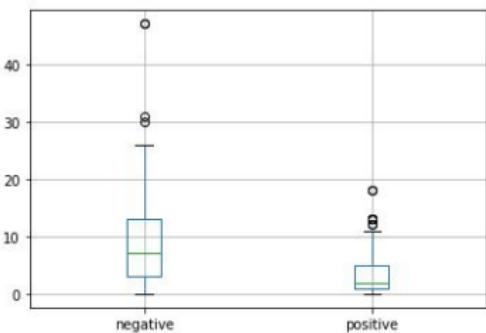


Cluster_2 - Trasporti aerei



Cluster_2 - Regole di associazione

Premessa	Conclusione	Supporto	Confidenza	Lift
cost, revenu	fuel	0.203	0.820	2.436
cost	fuel	0.282	0.633	1.881
fuel	cost	0.282	0.838	1.881



Documento 1435 - Incidente Aereo

Pilots Killed in UPS Cargo Jet Crash Near Birmingham, Ala., Airport

The Airbus A300 had two pilots on board when it crashed on final approach.

By COLLEEN CURRY

14 August 2013, 14:22 • 5 min read



UPS Cargo Jet Crashes Near Birmingham International Airport

[1]

"about the cause of the fiery crash of the United Parcel Service Inc aircraft in which two pilots were killed"



Cluster_2 - Business Intelligence

DOCUMENTO 978

"Still, airlines predicted sizable disruption and hundreds of millions of dollars in lost revenue if delays happen as predicted and persist for a year."

DOCUMENTO 1035

"Airlines typically build winter weather delays and cancellations into their annual budgets, but large-scale disruptions that result in lost business and higher operational costs can hit quarterly earnings."

DOCUMENTO 1895

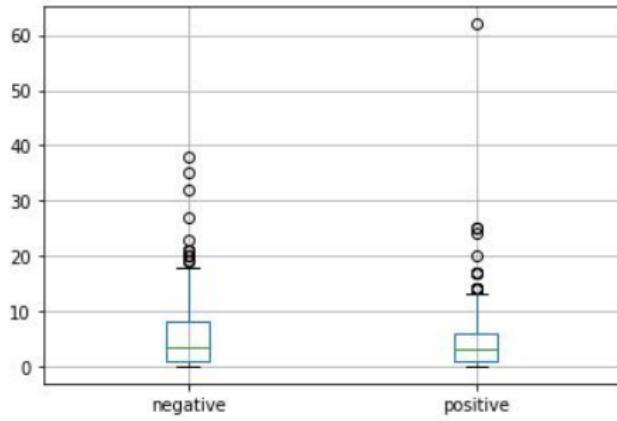
"Independent tracker FlightStats.com found higher numbers - 404 flights canceled on Monday with 7,027 delays, and 385 flights canceled on Tuesday with 6,396 delays. As soon as one major airport experiences significant delays it can have a knock-on impact across the system"

Cluster_6 - Internet

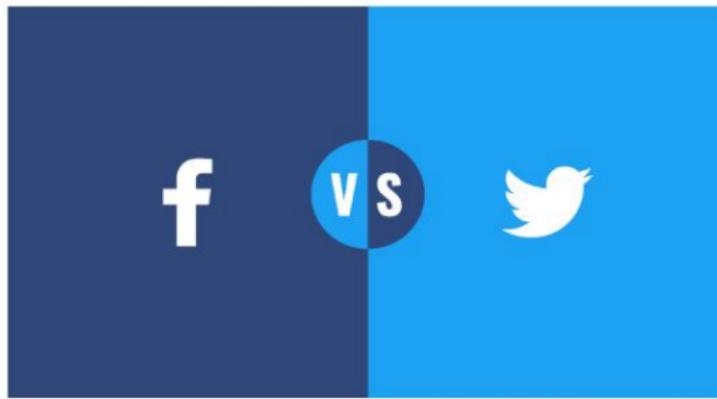


Cluster_6 - Regole di associazione

Premessa	Conclusione	Supporto	Confidenza	Lift
facebook	twitter	0.188	0.642	2.645
twitter	facebook	0.188	0.776	2.645
onlin	amazon	0.330	0.632	1.311
consum, retail	onlin	0.130	0.878	1.683



Facebook vs Twitter: Which is Best for Your Brand?



Scelta di spazi pubblicitari [2]



Regole di associazione

Non sono state trovate regole di associazione con lift negativo, mentre la maggior parte di esse riporta un lift poco superiore a 1, risultando di scarso interesse. I valori minimi di *support* e *confidence* sono stati scelti *ad hoc* per ogni cluster analizzato.

Considerando i cluster non presi in analisi, data l'appartenenza ad un settore con minor interesse ai fini di Business Intelligence, sono di seguito riportate le regole di associazione risultate più interessanti:

Premessa	Conclusione	Supporto	Confidenza	Lift
social	twitter	0.108	0.787	2.213
growth, chines	china	0.115	0.962	2.898

- Evidenziare principale **concorrenza**
- Evidenziare **punti di forza**
- Proporre possibili **joint ventures**
- Individuare possibile **pubblicità**
- Proporre potenziamento **componente online**

Sitografia

-  *Cargo Jet Crash.* URL: <https://abcnews.go.com/US/ups-plane-distress-calls-crash/story?id=19955672>.
-  *Facebook vs Twitter.* URL:
<https://sproutsocial.com/insights/facebook-vs-twitter/>.
-  *Joint Venture.* URL: <https://www.thebalancesmb.com/what-is-a-joint-venture-and-how-does-it-work-397540>.
-  *Stock Screener.* URL:
<https://www.macrotrends.net/stocks/stock-screener>.
-  *Verizon vs Sprint.* URL:
<https://www.reviews.org/mobile/sprint-vs-verizon-cell-phone-plans/>.

Grazie per l'attenzione!