Text Mining: articoli di politica e sport della BBC

Adolfo Luna Nunez, Elena Signori, Simone Sironi,

Novembre 20,2020

Abstract

L'obiettivo del presente report è di elaborare ed analizzare una collezione di 928 articoli riguardanti politica e sport tratti dal sito della BBC.
Inizialmente, sotto l'ipotesi di non conoscerne la reale natura, abbiamo raggruppato gli articoli (in base ai termini presenti in essi) e i termini con metodi di clustering partizionale e di clustering gerarchico. Il fatto di conoscere la reale natura dei documenti analizzati ci ha permesso di trarre immediate conclusioni riguardanti la qualità dei nostri risultati.
In una seconda fase, questa volta sotto l'ipotesi di conoscere la reale natura degli articoli, abbiamo classificato i documenti utilizzando un algoritmo non parametrico: il KNN (K-Nearest Neighbours).

1 Introduzione

Si può affermare con assoluta certezza che tra le grandi rivoluzioni nel mondo informatico vi è Internet. Inizialmente sia le persone che vi accedevano sia le pagine contenute all'interno della rete erano un numero limitato, adesso con i ripetuti abbassamenti dei costi dei calcolatori, delle connessioni internet e grazie al protocollo http (Hyper Text Transfer Protocol), definito da Tim Berners-Lee nel 1991, si è reso possibile far consultare a milioni di persone una grandissima quantità di risorse.
È quindi necessario lo studio di una tecnica che ci permetta di estrarre informazioni senza dover leggere tutti i testi a nostra disposizione; l'analisi tradizionale dei dati che si basa sul modello relazionale in cui i dati sono memorizzati in tabelle (i cosiddetti dati strutturati) non è applicabile in questo caso dato che ci troviamo di fronte a dati non strutturati.
Uno dei metodi per trarre beneficio da questa enorme quantità di dati testuali è utilizzare tecniche di text mining.
Il text mining si pone l'obiettivo di applicare metodi e algoritmi per estrarre automaticamente conoscenza dal testo e per classificare o raggruppare documenti in base ai contenuti.
Con il presente report proponiamo un'analisi di alcuni articoli della BBC attraverso l'applicazione di metodi di clustering e di classificazione. Lo scopo è quello di ricavare informazioni non note e potenzialmente utili circa il contenuto dei testi e di individuare il metodo che permetta di assegnare ai documenti la categoria corrispondente all'argomento trattato.

2 Materiali e metodi

2.1 Materiali

Il Corpus oggetto dell'analisi contiene 928 articoli estratti dal sito web della BBC (British Broadcasting Corporation, il più grande e autorevole editore radiotelevisivo del Regno Unito con sede a Londra) riguardanti il periodo 2004-2005. Originariamente il corpus era diviso in due parti: la prima composta da 417 articoli che trattano di politica, mentre la seconda da 511 che trattano di sport. Con l'obiettivo di non perdere il campo di appartenenza originario di ciascun articolo, i documenti sono stati preventivamente rinominati come: p(numero).txt se trattavano di politica e s(numero).txt se trattavano di sport.

2.2 Metodi

Il primo passo è stato eseguire una fase di pre-processing del corpus, cioè una pulizia del documento. Sono stati quindi eliminati i caratteri speciali, i segni di punteggiatura, i numeri e sono state trasformate le lettere maiuscole in minuscole. Infine, abbiamo rimosso le stop words della lingua inglese, ossia parole del vocabolario inglese ritenute non utili allo scopo dell'analisi che rappresentano circa il 20-30% della totalità del testo, a cui abbiamo aggiunto altre parole come aggettivi, verbi comuni e avverbi.
Attraverso una funzione personalizzata abbiamo eseguito sul Corpus contemporaneamente le operazioni di stemDocument e stemCompletion. La funzione stemDocument ci permette di eliminare prefisso/suffisso delle parole, mentre la funzione stemCompletion ci permette di non perdere l'interpretabilità delle parole ricompletando ciascun termine con il prefisso/suffisso che appariva più frequentemente nel testo originale.
Terminate le operazioni di pulizia del Corpus, abbiamo costruito la Document-Term Matrix, una matrice che ha sulle righe gli identificativi dei documenti (Document: termine generico che indica un documento) e sulle colonne le singole parole contenute in tutti i documenti (Term: tipicamente una parola singola, può essere però formato da brevi composizioni). All' interno della matrice troviamo le term frequency, ossia le frequenze di apparizione di ogni termine in ciascun documento.
Abbiamo individuato i 100 termini più frequenti e li abbiamo visualizzati graficamente attraverso un barplot (un grafico a barre che presenta sull'asse delle ascisse i termini e sull'asse delle ordinate la loro frequenza) e un word cloud (una "nuvola di parole" che permette di rappresentare i termini scrivendoli a grandezze diverse in relazione alla loro frequenza). Abbiamo poi svolto una breve analisi delle associazioni tra i termini che ci ha permesso di fare alcune considerazioni.
Tra queste parole abbiamo eliminato quelle che risultavano poco rilevanti per l'analisi o di disturbo. Abbiamo anche provveduto ad eliminare i termini in comune tra le due categorie di articoli in modo da aumentare l'efficienza dei metodi di clustering e di classificazione e diminuire lo sforzo di tali metodi a livello computazionale.
L'analisi del Corpus è stata svolta attraverso le fasi seguenti:

1. Clustering dei documenti
2. Clustering dei termini
3. Classificazione dei documenti

1. Clustering dei documenti

Supponendo di non conoscere la reale natura degli articoli, abbiamo cercato di raggruppare i documenti attraverso metodi di clustering. Questa tecnica fa parte dei metodi non supervisionati, vale a dire quei metodi che nei dati di addestramento non prevedono la presenza di informazioni relative alla caratteristica di interesse (variabile target, nel nostro caso la reale natura degli articoli). Il clustering si divide in due diverse tipologie: gerarchico, che crea una decomposizione gerarchica degli oggetti, e partizionale, che crea una serie di partizioni e ne valuta la qualità. Nella nostra analisi abbiamo scelto di applicare sia il clustering gerarchico che quello partizionale.

Clustering gerarchico

Per il clustering gerarchico abbiamo preso in considerazione tre metriche che ci permettono di misurare la distanza tra le osservazioni e di stabilire quindi una similarità/dissimilarità tra di esse: in questo caso tra gli articoli.

Abbiamo scelto:

- Distanza euclidea:

$$d_e(D_i, D_j) = \sqrt{\sum_{t=1}^{m} (f_{it} - f_{jt})^2}$$

- Distanza di manhattan:

$$d_m(D_i, D_j) = \sum_{t=1}^{m} |f_{it} - f_{jt}|$$

- Distanza basata sul coefficiente di correlazione di Pearson:

$$d_p = \frac{1 - \rho(D_i, D_j)}{2} \qquad \text{dove} \quad \rho(D_i, D_j) = \frac{\sum_{t=1}^{m} (f_{it} - \bar{f}_t)(f_{jt} - \bar{f}_t)}{\sqrt{\sum_{t=1}^{m} (f_{it} - \bar{f}_t)^2} \sqrt{\sum_{t=1}^{m} (f_{jt} - \bar{f}_t)^2}}$$

Nel caso in esame possiamo considerare tali metriche come misura della dissimilarità tra i documenti. Questo significa che maggiore è il valore della distanza meno i due articoli considerati sono simili tra loro.

Nel clustering gerarchico è necessario specificare anche il legame che si vuole utilizzare. Il linkage (legame) definisce il metodo che viene impiegato per misurare la distanza tra un'osservazione (articolo) e un gruppo, oppure tra due gruppi.

Per la nostra classificazione abbiamo scelto di considerare il linkage completo e il linkage di Ward:

- Complete linkage: la distanza tra due clusters (oppure tra un'osservazione e un cluster) è determinata dalla distanza maggiore tra ogni coppia di osservazioni nei due differenti clusters.

$$d(C_1, C_2) = \max(\, d(c_{1i}, c_{2i}) \in D \,)$$

(Dove D è una matrice di dissimilarità)

- Ward linkage: la distanza tra due clusters è determinata dall'incremento della devianza intra-gruppo che si otterrebbe se venissero uniti i due clusters considerati. L'obiettivo di tale legame è minimizzare la distanza intra-gruppo finale.

$$d_{in}(C_1, C_2) = \sum_{k=1}^{G} \sum_{s=1}^{p} \sum_{i=}^{n} (f_{is} - \overline{f_{sk}})$$

Nel processo del clustering gerarchico ad ogni step vengono agglomerati i gruppi che minimizzano la distanza definita dalla metrica specificata.
Per ciascuna coppia di distanza-linkage applicati alla Document-Term Matrix, abbiamo costruito un dendrogramma: una rappresentazione grafica che permette di visualizzare il processo agglomerativo delle osservazioni nei clusters e la loro similarità individuata dall'altezza del più basso nodo interno condiviso.
La scelta del punto di taglio del dendogramma, e quindi del numero di gruppi, è stata fatta sulla base della silhouette media per il clustering totale. La silhouette viene calcolata per ogni osservazione ed è un indicatore che, tenendo conto sia della separazione sia della coesione, ci fornisce il livello di corretta appartenenza al gruppo assegnato; assume valori compresi tra -1 e 1, dove il valore 1 indica che l'osservazione è stata assegnata al gruppo corretto.
Inoltre, abbiamo usato questo indicatore anche per valutare quale coppia di distanza e legame restituisse il raggruppamento migliore, ma abbiamo tenuto conto anche della rappresentazione grafica.

Clustering partizionale

Per il clustering partizionale abbiamo scelto di applicare il metodo delle k-medie.
Si tratta di un algoritmo iterativo nel quale, ad ogni step, vengono definiti k centroidi e ogni osservazione viene assegnata ad uno dei k gruppi in base alla distanza dal centroide corrispondente, che deve essere la minore tra tutte le distanze dai centroidi disponibili. Il numero dei clusters deve essere fissato in partenza e, poiché la prima determinazione dei centroidi è casuale, è opportuno ripetere l'algoritmo più volte.
Anche per questo metodo di raggruppamento abbiamo utilizzato la silhouette come indicatore per valutare la qualità del clustering e abbiamo cercato il numero di gruppi che la massimizzasse.
Una volta individuati i gruppi abbiamo scelto di rappresentarli attraverso un clusplot: un grafico che permette di visualizzare i clusters in uno spazio bidimensionale attraverso le prime due componenti principali.
Oltre al metodo k-means abbiamo provato ad applicare anche il metodo k-medoids. Si tratta di una variante del metodo k-means, nel quale al posto dei centroidi (vettori di medie delle variabili per le osservazioni del cluster) vengono utilizzati i medoidi; un medoide è un'osservazione per la quale la dissimilarità media con tutte le altre osservazioni del cluster è minima. Per valutare la qualità del metodo abbiamo usato la silhouette media, come nei metodi precedenti,

e informazioni aggiuntive restituite dalla funzione pamk (funzione specifica per l'utilizzo del metodo k-medoids).

2. Clustering dei termini

In questa fase abbiamo utilizzato la Term-Document Matrix, vale a dire la trasposta della matrice Document-Term che quindi presenta i termini come intestazioni delle righe e i documenti come intestazioni delle colonne.
Anche per i termini abbiamo deciso di applicare dei metodi non supervisionati di clustering, sia di tipo gerarchico che partizionale, per cercare di capire se all'interno delle due macrocategorie a noi note (sport e politica) si potessero individuare anche dei sotto-argomenti.

Clustering gerarchico

Nel clustering gerarchico abbiamo scelto di considerare come metriche di distanza: distanza euclidea, distanza di manhattan e distanza basata sul coefficiente di correlazione di Pearson. Come linkage abbiamo utilizzato il legame di Ward e il legame completo.
Per ogni coppia distanza-linkage abbiamo rappresentato un dendrogramma e abbiamo individuato il punto di taglio, che determina il numero di gruppi, che massimizzasse la silhouette media del clustering totale.

Clustering partizionale

Nel clustering partizionale abbiamo applicato l'algoritmo k-means valutando il numero di gruppi da realizzare sulla base della silhouette media. Abbiamo poi cercato di interpretare i risultati ottenuti.

3. Classificazione dei documenti

Poiché possiamo sapere la tipologia originale degli articoli, abbiamo deciso di provare a classificare i documenti. Prima di tutto è necessario confermare la conoscenza della variabile target dato che la classificazione fa parte dei metodi supervisionati. Un metodo si dice supervisionato se i dati di addestramento contengono informazioni sulla caratteristica di interesse, cioè sulla variabile risposta (nel nostro caso è la tipologia degli articoli).
Per poter avere i dati nella forma richiesta, abbiamo aggiunto alla Document-Term Matrix una colonna che costituisce una variabile dicotomica chiamata "type" che indica la natura dell'articolo (0=politica, 1=sport).
Tra i metodi di classificazione troviamo: regressione logistica, analisi discriminante (lineare o quadratica) e k-nearest neighbour. Abbiamo escluso l'applicazione della regressione logistica perché i termini, che costituirebbero le variabili esplicative, sono troppo numerosi.
Inoltre l'analisi discriminante, sia lineare che quadratica, è stata esclusa perché per la sua applicazione è necessario che i dati rispettino le assunzioni di normalità delle variabili esplicative e nel caso della LDA anche della varianza comune tra le classi, che non sono soddisfatte dai nostri dati.

Ci siamo quindi concentrati sul metodo K-Nearest Neighbour. Si tratta di un algoritmo non parametrico, cioè che non richiede assunzioni sulle distribuzioni delle esplicative, definito lazy learning algorithm perché non prevede la stima di un modello. Per utilizzare questo metodo è sufficiente definire il parametro di tuning (k) che definisce il numero di "vicini" che deve essere considerato per classificare un'osservazione. Per classificare un'osservazione si calcola la sua distanza, generalmente quella euclidea, da tutte le altre e si identificano le k osservazioni più vicine a quella di interesse. Successivamente si individua la classe di appartenenza di ogni vicino e all'osservazione d'interesse verrà attribuita la classe presente in misura maggiore nel "vicinato".
L'applicazione del KNN si è svolta a partire dalla matrice Document-Term a cui abbiamo aggiunto la colonna "type". Abbiamo trasformato questa matrice in un dataframe e abbiamo diviso quest'ultimo in training e test set. Nel training set abbiamo usato un ciclo for per trovare il valore di k che ottimizza il metodo. Per la valutazione della classificazione ci siamo basati sull'accuracy, un indicatore che calcola la proporzione di osservazioni che sono state previste correttamente e che si ricava dalla Confusion Matrix. Per valutare le performance del KNN abbiamo usato anche la curva di ROC, un grafico che traccia la probabilità di ottenere un true positive in funzione della probabilità di un false positive per una serie di valori di cut-off. L'obiettivo è quello di massimizzare il true positive rate (proporzione di osservazioni classificate correttamente) e di minimizzare il false positive rate (proporzione di osservazioni classificate erroneamente).

3 Risultati

Per prima cosa abbiamo caricato i dati in R tramite il comando "Corpus", ottenendo un oggetto di tipo "Simple Corpus" costituito da una collezione di 928 articoli.
Per rendere più efficiente l'analisi degli articoli e dei termini abbiamo eseguito una fase di pulizia dei documenti (pre-processing). Per prima cosa abbiamo rimosso i caratteri speciali trovati aprendo articoli casualmente (ad esempio "£", "-"), sostituendoli con uno spazio bianco. Abbiamo poi trasformato le lettere maiuscole in minuscole e rimosso numeri e punteggiatura attraverso le apposite funzioni (tolower, removeNumbers, removePunctuation). Successivamente, abbiamo rimosso anche le stop words della lingua inglese proposte dal software a cui abbiamo aggiunto altre parole non utili ai fini della nostra analisi, come: numeri espressi in lettere, alcuni verbi, avverbi e aggettivi. Come ultima fase prima dello stemming abbiamo rimosso tutti gli spazi bianchi in eccesso tramite la funzione "stripWhitespace". Per effettuare l'operazione di stemming e il successivo completamento delle parole, abbiamo utilizzato la funzione "text_tokens" in modo che ogni articolo fosse trasformato in un vettore di singoli termini. Abbiamo poi utilizzato un ciclo "for" per applicare ad ogni articolo le funzioni stemDocument e stemCompletion e, infine, abbiamo riconvertito i vettori di parole in testi.
Terminata la fase di pre-processing, abbiamo costruito la Document-Term Matrix, di dimensioni 928x320, dimensionalità ottenuta dai 928 articoli tenendo conto solo dei termini di lunghezza compresa tra 3 e 30 caratteri e apparsi in almeno 80 articoli (≈9%) e al massimo in 840 (≈90%). Abbiamo limitato i termini nella matrice per evitare di trattare termini molto rari o troppo frequenti.

Costruita la matrice abbiamo visualizzato i 100 termini più frequenti e ci siamo resi conto della necessità di allungare la lista delle stop words data la presenza di molte parole non informative (ad esempio also, year, week); abbiamo quindi aggiornato la lista e ripetuto i passi precedenti.

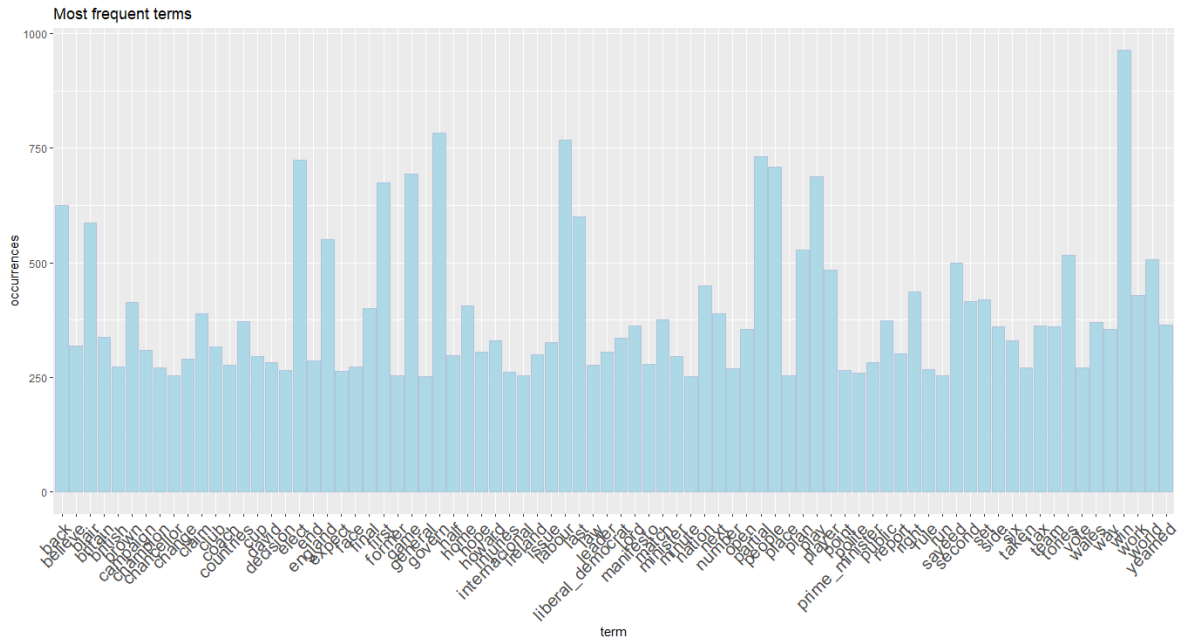

Figura 1: Barplot dei termini più frequenti (freq>250, 80 termini)

Dall'osservazione dei termini più frequenti abbiamo inoltre notato che molte parole potevano essere considerate insieme; questo ci ha portato, attraverso la funzione findAssocs, ad analizzare nel dettaglio la correlazione tra parole che potevano costituire un unico termine, ad esempio nome e cognome. Per risolvere questo problema abbiamo provveduto a sostituire queste combinazioni di termini in modo che fossero riconosciute come un'unica parola oppure, in altri casi, abbiamo considerato solo il temine secondo noi più significativo. Ad esempio: per nomi e cognomi abbiamo deciso di lavorare con il cognome, invece abbiamo considerato come "olympics" tutte le combinazioni di parole usate per indicare le olimpiadi ("Sidney olympics", "Athens olympics", "Sidney games"...). Abbiamo poi rappresentato le associazioni tra alcune delle parole che si presentano con maggiore frequenza negli articoli sulla base della loro correlazione. Per farlo abbiamo usato il comando plot, rappresentando i termini scelti sottoforma di una word network. Dal grafico possiamo osservare che si distinguono due gruppi di parole, uno per argomento, e sono uniti dalla parola "person" che non essendo specifica viene usata in entrambe le tipologie.

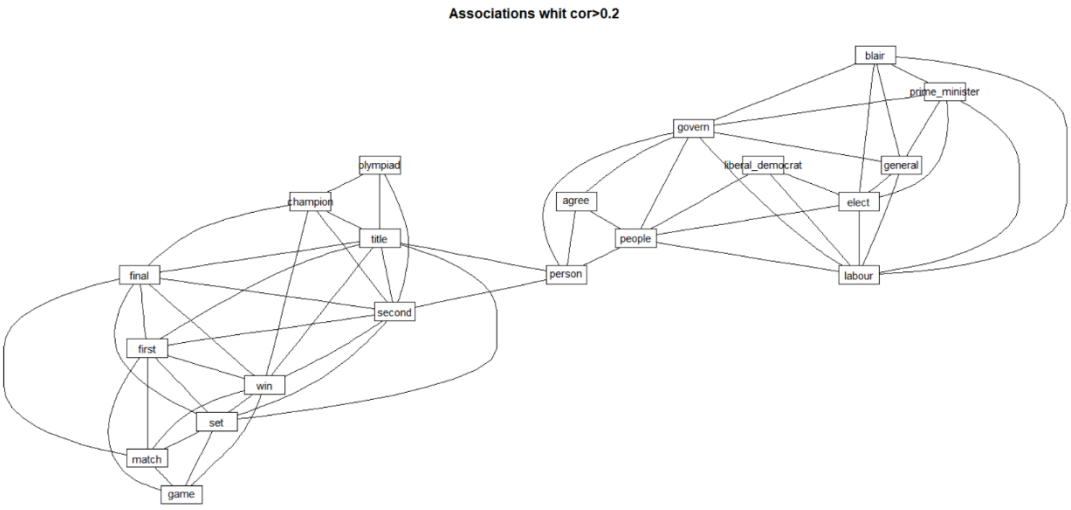

Figura 2: plot delle associazioni tra alcune parole (correlazione minima 0.2)

Prima di procedere con le altre fasi dell'analisi, abbiamo provato a visualizzare attraverso due grafici wordcloud i termini più frequenti divisi per categoria di articolo.

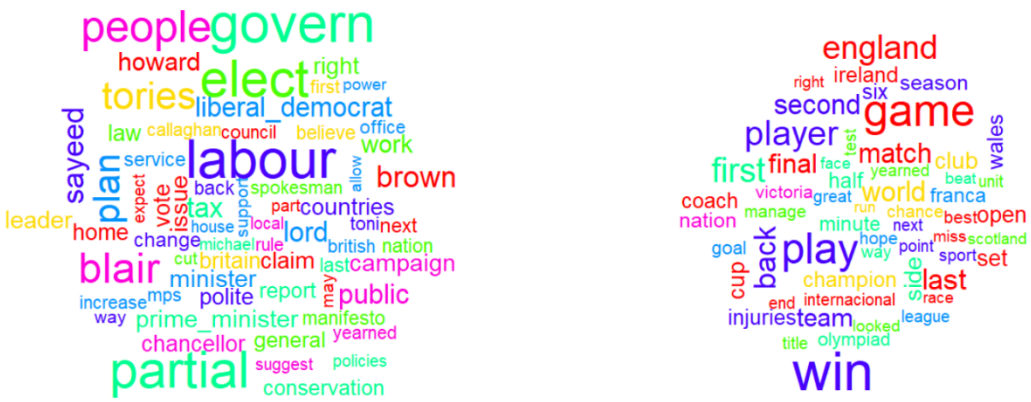

Figura 3: a sx wordcloud per articoli di politica, a dx wordcloud per articoli di sport (freq > 150)

Attraverso queste rappresentazioni (e anche alcune funzioni) abbiamo osservato la presenza di termini in comune tra le due tipologie di articoli e abbiamo deciso di eliminarli in modo da ottenere un clustering più efficiente nelle fasi successive.
Alcuni termini in comune: "david" "follow", "home", "move", "next", "right", "win".

1. Clustering dei documenti

Clustering gerarchico

Per il clustering gerarchico agglomerativo abbiamo considerato tre metriche di distanza differenti: distanza euclidea, distanza di manhattan e distanza basata sul coefficiente di correlazione di Pearson. Abbiamo inoltre deciso di confrontare due legami: Ward linkage e complete linkage.

Il numero ottimale di gruppi da realizzare è stato individuato sulla base della silhouette media. A questo scopo, abbiamo creato una funzione che calcolasse le silhouette medie per un numero di gruppi da 2 a 6 in modo da poterle confrontare.
Per ogni possibile combinazione distanza-linkage, grazie anche all'utilizzo del grafico clusplot, è stata evidenziata la presenza di alcuni outliers: "p (290).txt", "p (380).txt", "p (293).txt". Abbiamo deciso di eliminare tali articoli perché, disponendo di un elevato numero di documenti, abbiamo ritenuto che l'analisi non ne avrebbe risentito.
In ogni caso analizzato, il numero di clusters che massimizza la silhouette media è risultato essere pari a 2. Sapendo che originariamente gli articoli erano divisi in due categorie possiamo considerare questo risultato adeguato.
Tuttavia, nel caso del linkage completo, sia con distanza euclidea che di manhattan, abbiamo osservato che il clustering non viene effettuato in modo efficiente. Infatti, nel primo dei due gruppi vengono raggruppate la quasi totalità delle osservazioni mentre nel secondo sono presenti un numero molto ridotto di osservazioni; si tratta di ulteriori outliers rispetto a quelli già eliminati. Abbiamo agito eliminando anche questi articoli ma, nonostante questo, ogni volta che si eliminava un outlier se ne evidenziavano di nuovi impedendo un buon raggruppamento delle osservazioni. Inoltre, abbiamo notato che i valori della silhouette risultavano molto più alti con la presenza degli outliers, mentre rimuovendoli si sono ottenuti valori inferiori.
Possiamo quindi affermare che è importante non soffermarsi sui valori dell'indicatore quantitativo silhouette ma è bene approfondire i risultati ottenuti; inoltre abbiamo osservato che il linkage di Ward fornisce risultati migliori a livello rappresentativo perché permette di individuare in modo più definito i due clusters.
A sostengo di quanto appena detto riportiamo il dendogramma e il clusplot per la distanza di manhattan con linkage completo e di Ward. Si può osservare che con il legame completo pur avendo rimosso diversi outliers non è possibile distinguere chiaramente due gruppi.

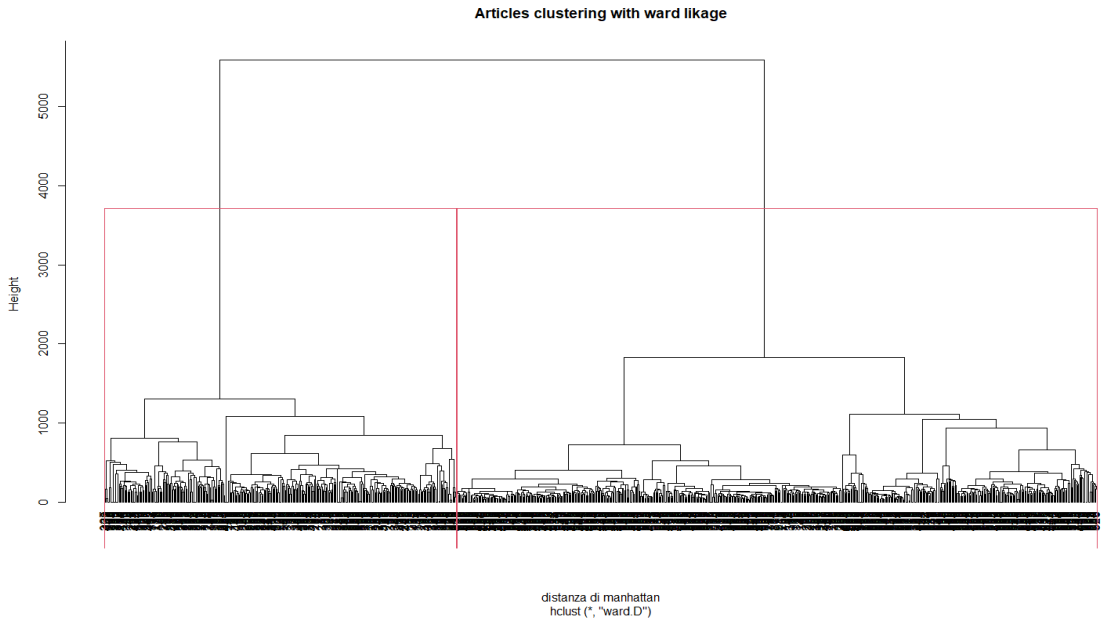

Figura 4: clustering degli articoli con distanza di manhattan e legame di Ward

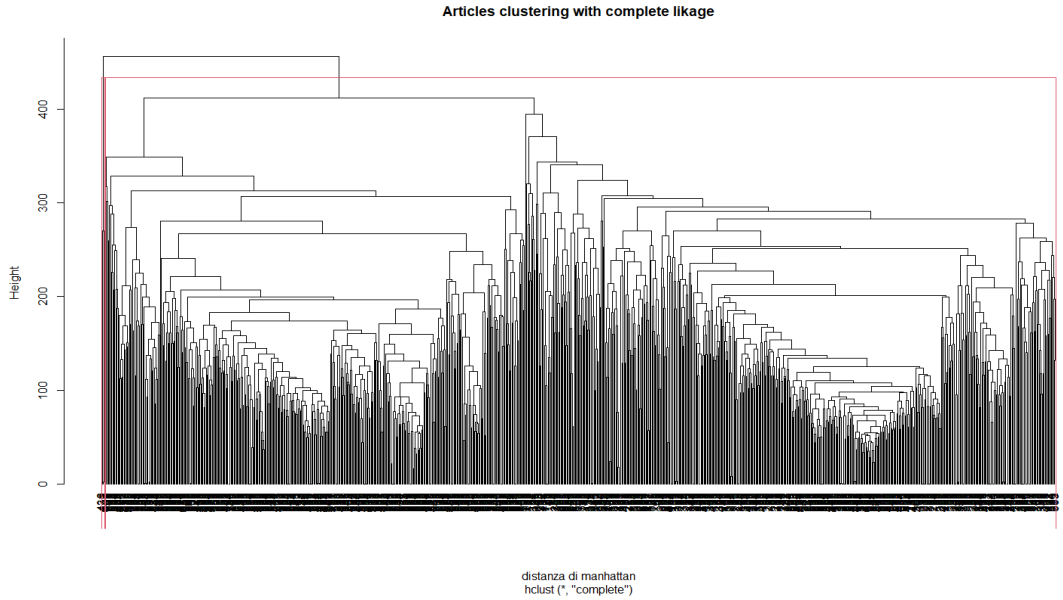

Figura 5: clustering degli articoli con distanza di manhattan e legame completo

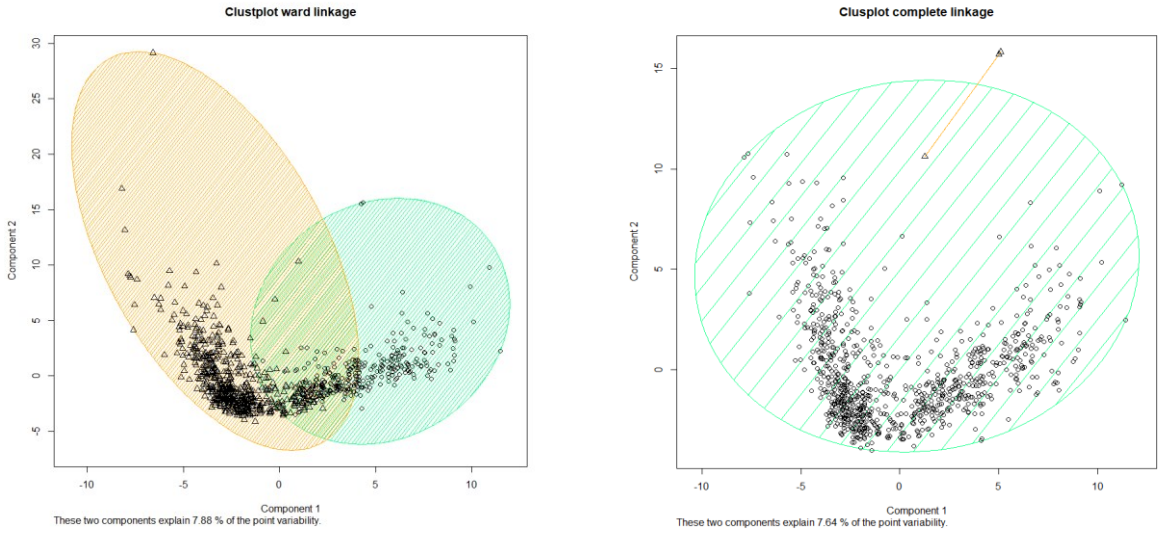

Figura 6: a sx clusplot con legame di Ward, a dx clusplot con legame completo (distanza di manhattan)

Per quanto riguarda le silhouette medie, con un numero di gruppi pari a 2 abbiamo ottenuto i valori riportati nella tabella seguente; abbiamo deciso di riportare i valori della silhouette media dopo aver rimosso alcuni outliers qualora fossero presenti.

Metrica di distanza	Linkage	Silhouette media
Euclidea	Ward	0.045
Euclidea	Complete	0.402
Manhattan	Ward	0.133
Manhattan	Complete	0.519
Coeff. di Pearson	Ward	0.121
Coeff. di Pearson	Complete	0.118

Figura 7: Tabella per confronto tra clustering

Dai dati è possibile osservare che il legame completo fornisce valori di silhouette superiori. Tuttavia, date le considerazioni appena fatte in merito alla rappresentazione, abbiamo deciso di soffermarci sul legame di Ward. Tra le tre metriche di distanza considerate quella che permette di massimizzare la silhouette media è la distanza di manhattan. Si tratta però di un valore che non può essere considerato ottimale perché risulta essere basso rispetto al valore massimo possibile, cioè 1 (solitamente si considera buono un valore superiore a 0.7).
Approfondendo il contenuto dei due cluster ottenuti con la distanza di manhattan e linkage di Ward, tenendo conto della nostra conoscenza a priori della natura degli articoli, sono emersi i seguenti risultati:

Cluster	Dimensione	N° art. Politica	N° art. Sport
1	328	326	2
2	597	88	509

Figura 8: tabella caratteristiche dei clusters con distanza di manhattan e linkage di Ward

Possiamo quindi dire che nonostante il valore molto basso della silhouette media il clustering non commette un errore eccessivo nel raggruppare gli articoli.

Clustering partizionale

Nel clustering partizionale abbiamo scelto di utilizzare il metodo k-means. Per prima cosa abbiamo scelto di applicare il metodo alla matrice DocumentTerm scalata completa di tutti gli articoli. Abbiamo testato diversi valori di k, poiché in questo metodo il numero di gruppi deve essere fissato in partenza, e abbiamo visto che il numero di clusters che massimizza la silhouette è 2. Rappresentando i due gruppi attraverso un grafico clusplot, ci siamo resi conto della presenza di alcuni possibili outliers; abbiamo quindi deciso di eliminarli e di ripetere l'analisi.

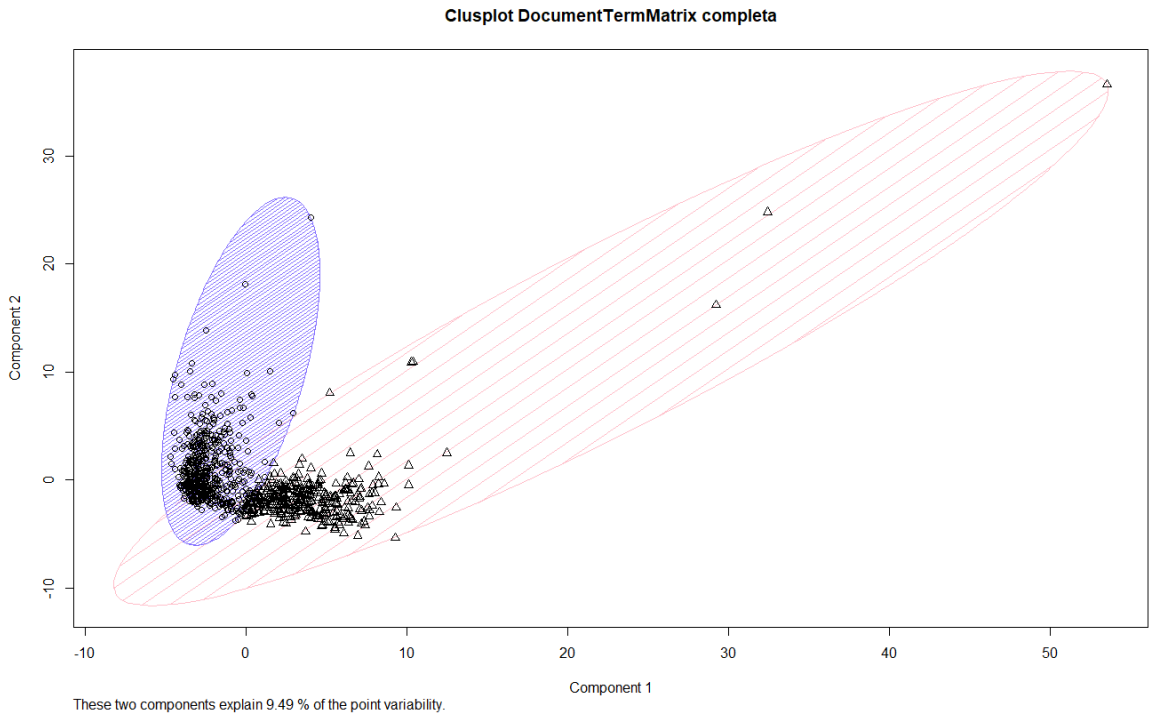

Figura 9: clusplot dei gruppi ottenuti con k-means considerando tutti gli articoli

Gli articoli che abbiamo rimosso dopo alcune ripetizioni dell'algoritmo sono: "p (290).txt", "p (380).txt", "p (293).txt", "s (491).txt", "s (371).txt", "s (58).txt".
La rimozione di questi documenti ha comportato un peggioramento del valore della silhouette media (da 0.088 a 0.074), ma ha permesso di ottenere una migliore rappresentazione grafica. Anche eliminando gli outliers la silhouette media è massimizzata con k=2.

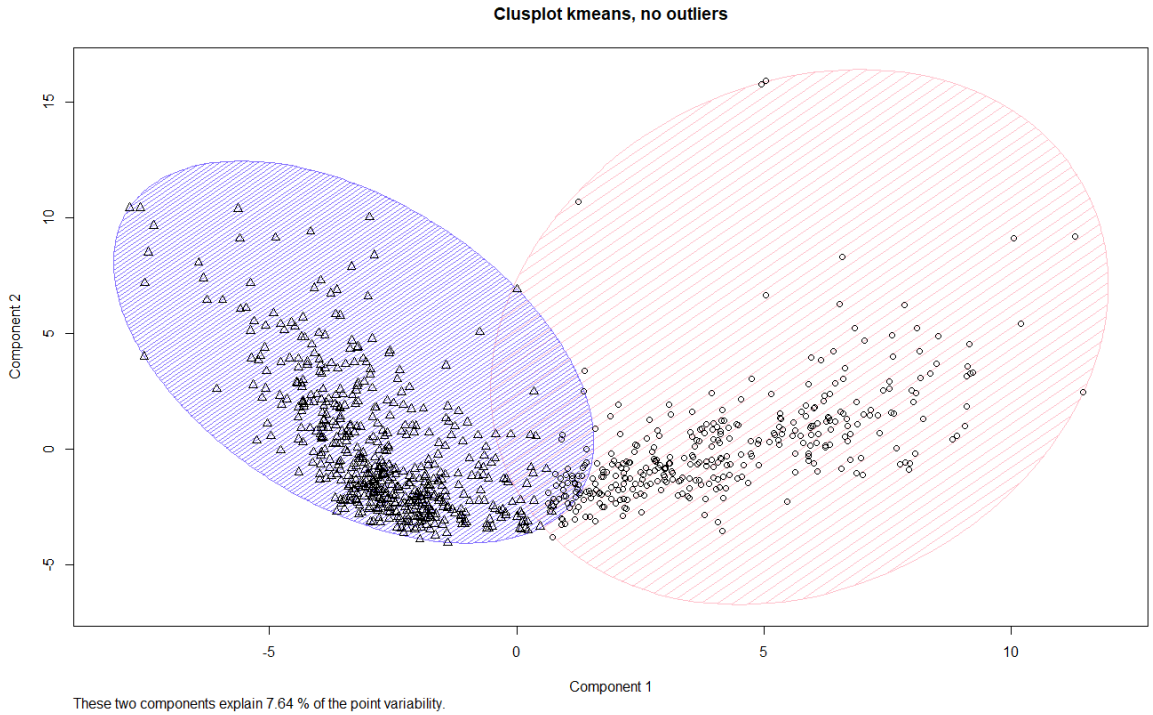

Figura 10: clusplot dei gruppi ottenuti con k-means senza outliers

Essendo a conoscenza della reale natura degli articoli possiamo dire che nonostante il valore basso della silhouette media, i due gruppi che vengono individuati attraverso l'algoritmo k-means distinguono abbastanza bene gli articoli di politica da quelli di sport. Nel primo gruppo sono presenti quasi esclusivamente articoli di politica ad eccezione di 3 articoli di sport, mentre nel secondo sono presenti prevalentemente articoli di sport e alcuni di politica ma non in misura eccessiva.

Cluster	Dimensione	N° art. Politica	N° art. Sport
1	370	367	3
2	552	47	505

Figura 11: tabella caratteristiche dei clusters con k-means

Come ultimo metodo di clustering applicato ai documenti, abbiamo voluto provare ad usare il k-medoids. Attraverso la funzione pamk abbiamo testato valori di k compresi tra 2 e 10 sulla matrice Document Term scalata completa di tutti gli articoli. La funzione pamk restituisce tra i suoi valori il k che ottimizza il processo di clustering che è risultato essere k=2. Rappresentando i due gruppi si osserva che ci sono alcuni outliers: "p (290).txt","p (380).txt","p (293).txt","s (491).txt","s (371).txt". Dopo averli rimossi abbiamo ripetuto il clustering ottenendo ancora un k ottimale pari a 2 e una migliore rappresentazione dei gruppi, mentre la silhouette media risulta inferiore dopo aver rimosso gli outliers (0.014 rispetto a 0.018).

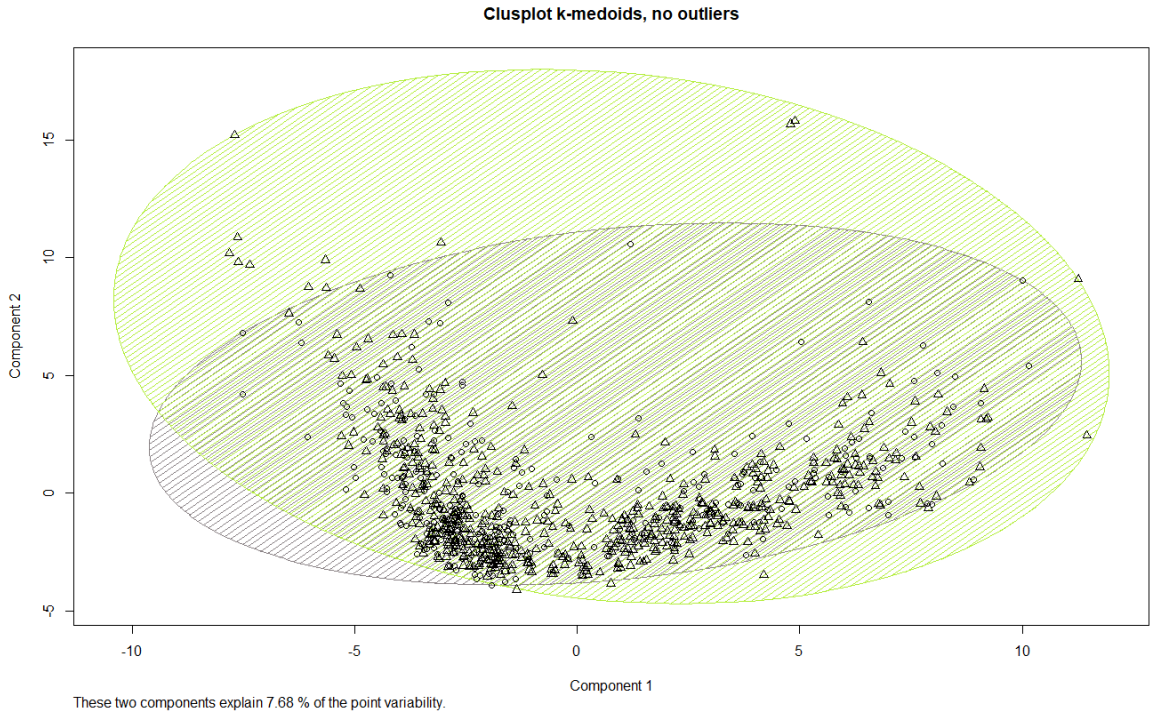

Figura 12: clusplot dei gruppi ottenuti con k-medoids senza outliers

Nonostante la rappresentazione dei clusters sia migliore rispetto a quella con gli outliers compresi, i due gruppi non sembrano essere ben distinti ma sembrano essere quasi sovrapposti. Questo significa che la distanza tra i gruppi è poca e quindi alcuni articoli saranno classificati in modo errato. Infatti, come si può vedere dalla tabella riportata, la distanza minima tra due osservazioni dei due gruppi è 7.976 (separation), che risulta essere un valore molto più basso rispetto alla distanza massima tra due osservazioni dello stesso cluster che assume per entrambi i gruppi valori vicini a 50 (diameter). Nella tabella abbiamo riportato anche la dissimilarità massima tra un'osservazione del cluster e il suo medoide (max dissimilarity) e la dissimilarità media tra le osservazioni e il medoide (average dissimilarity).

Cluster	Size	Max dissimilarity	Average dissimilarity	Diameter	Separation
1	325	35.132	18.122	46.452	7.976
2	598	39.838	17.168	52.090	7.976

Figura 13: tabella caratteristiche dei clusters con k-medoids

A conferma di quanto detto, abbiamo visualizzato gli articoli contenuti in ogni cluster e, poiché conosciamo la categoria originaria degli articoli, abbiamo visto che in entrambi i gruppi circa metà delle osservazioni sono articoli di politica e l'altra metà articoli di sport.

2. Clustering dei termini

Per applicare i metodi di clustering sui termini abbiamo creato la matrice Term Document considerando solo i termini di lunghezza compresa tra 3 e 30 caratteri e apparsi in almeno 80 articoli (≈9%) e al massimo in 840 (≈90%).
Successivamente abbiamo effettuato un'ulteriore selezione delle parole tenendo conto solo di quelle che si sono presentate con frequenza superiore a 180; in questo modo otteniamo una Term Document Matrix di dimensioni 114x928.

Clustering gerarchico

Per il clustering gerarchico abbiamo utilizzato e confrontato le stesse misure di distanza e gli stessi linkage scelti per il clustering gerarchico dei documenti. Per ogni coppia di distanza-linkage abbiamo rappresentato il relativo dendogramma e abbiamo scelto il punto di taglio in corrispondenza del numero di gruppi che massimizzasse la silhouette media. Per individuare il numero di clusters ottimale abbiamo usato la stessa funzione creata in precedenza che ci permette di confrontare le silhouette medie per diversi numeri di clusters (abbiamo testato da 2 a 10 clusters).
Nel caso della distanza euclidea con linkage completo il k che massimizzava la silhouette media era pari a 3; rappresentando i gruppi si è evidenziata la presenza di due outliers: "play" e "game". Eliminando entrambi gli outliers la silhouette risulta massimizzata per k=2 e il suo valore aumenta.

Anche in tutti gli altri casi esaminati il numero di clusters che massimizza la silhouette media è stato k=2.
Prendendo in considerazione la distanza euclidea si sono ottenuti i seguenti valori di silhouette media:

Linkage	Silhouette media
Ward	0.050
Complete	0.203

Figura 14: tabella della silhouette media per la distanza euclidea

Il linkage completo è quello che massimizza il valore dell'indicatore. Tuttavia, visualizzando i gruppi attraverso il dendogramma si osserva che il primo cluster contiene quasi tutti i termini ad eccezione di 12 che sono raggruppati nel secondo cluster. Da un punto di vista grafico, come era successo nel clustering dei documenti, il linkage di Ward restituisce una rappresentazione migliore pur avendo un valore inferiore della silhouette. Per questo abbiamo scelto di riportare questo secondo dendogramma.
Approfondendo il contenuto dei clusters notiamo che questa coppia di distanza-linkage permette di ottenere una buona discriminazione tra i termini perché in un gruppo sono presenti parole relative allo sport e nell'altro parole riguardanti la politica.

Cluster 1	Elect, liberal_democrat, manifesto, govern, conservation, prime_minister, campaign, countries, lord, chancellor, tories, vote, spokesman …
Cluster 2	Minute, olympiad, club, champion, player, league, game, internacional, injuries, point, title, sport, match, final, cup, season, lead …

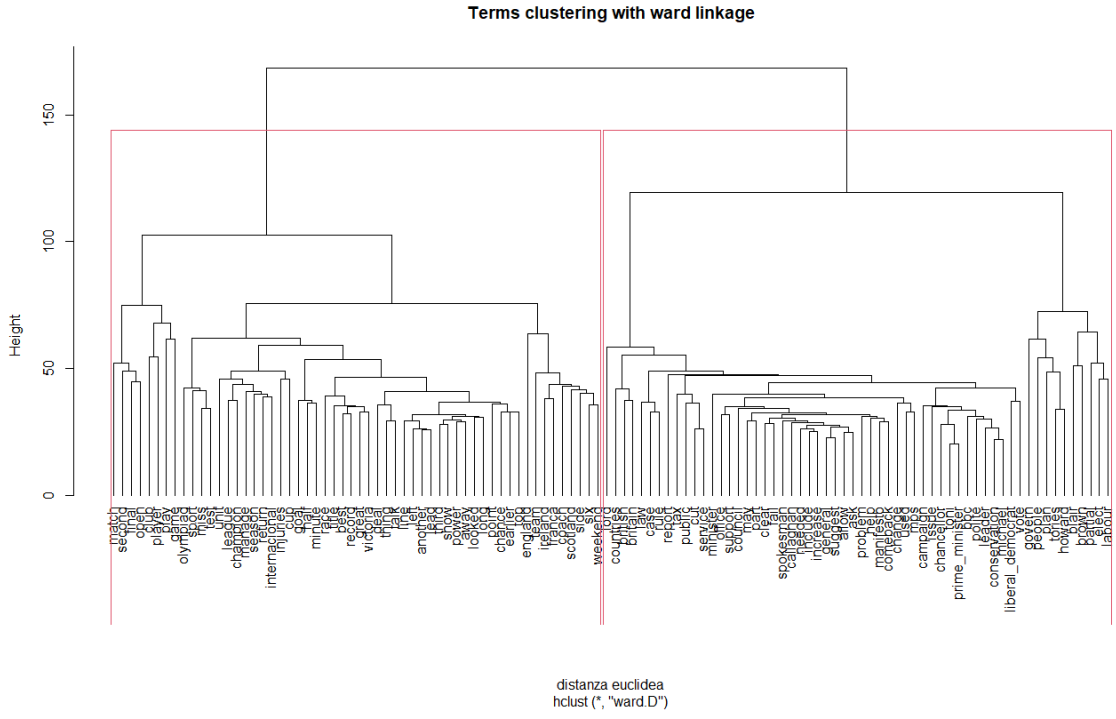

Figura 15: dendogramma con clusters per la distanza euclidea e linkage di Ward

Nel caso della distanza di manhattan i due legami considerati restituiscono valori quasi identici di silhouette media ed entrambi discriminano bene i termini in due gruppi ben definiti, uno relativo alla politica e uno relativo allo sport.

Linkage	Silhouette media
Ward	0.098
Complete	0.099

Figura 16: tabella della silhouette media per la distanza di manhattan

Nel caso del linkage di Ward, però, è emerso che un numero molto ridotto di termini relativi allo sport è contenuto nel gruppo di parole riguardanti la politica (ad esempio la parola "sport"). Questo non accade con il legame completo, riportiamo quindi il suo dendogramma.

Cluster 1	Elect, liberal_democrat, manifesto, govern, conservation, prime_minister, campaign, countries, lord, chancellor, tories, vote, spokesman, blair …
Cluster 2	Minute, olympiad, club, champion, player, league, game, internacional, injuries, point, title, sport, match, final, cup, season, lead …

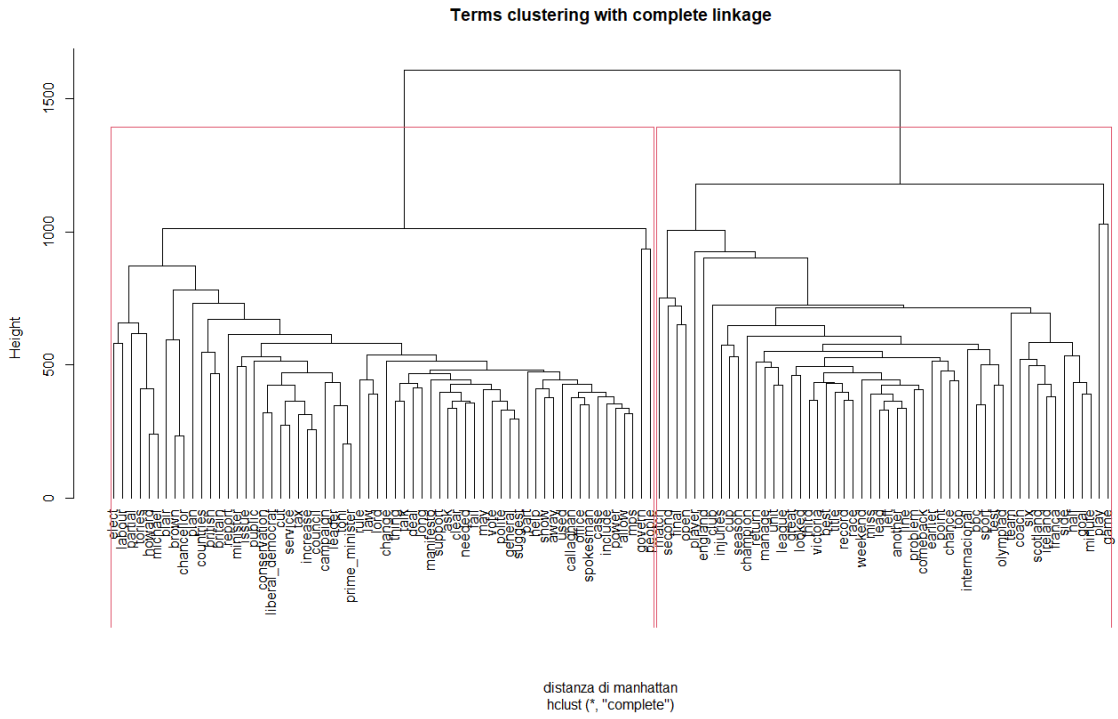

Figura 17: dendogramma con clusters per la distanza di manhattan con linkage completo

Considerando, infine, la distanza basata sulla correlazione di Pearson i due legami considerati con k=2 restituiscono valori della silhouette media molto vicini.

Linkage	Silhouette media
Ward	0.098
Complete	0.079

Figura 18: tabella della silhouette media per la distanza basata sul coeff. di correlazione di Pearson

In questo caso il linkage di Ward permette di ottenere una migliore discriminazione dei termini, distinguendo bene i due gruppi. Con il legame completo invece, alcune parole riguardanti lo sport vengono raggruppate insieme ad altre relative alla politica (ad esempio "sport", "olympiad", "club"). Riportiamo i clusters per il legame di Ward:

Cluster 1	Elect, liberal_democrat, manifesto, govern, conservation, prime_minister, campaign, countries, lord, chancellor, tories, vote, spokesman, blair …
Cluster 2	Minute, olympiad, club, champion, player, league, game, injuries, title, point, sport, match, final, cup, season, lead, goal, coach …

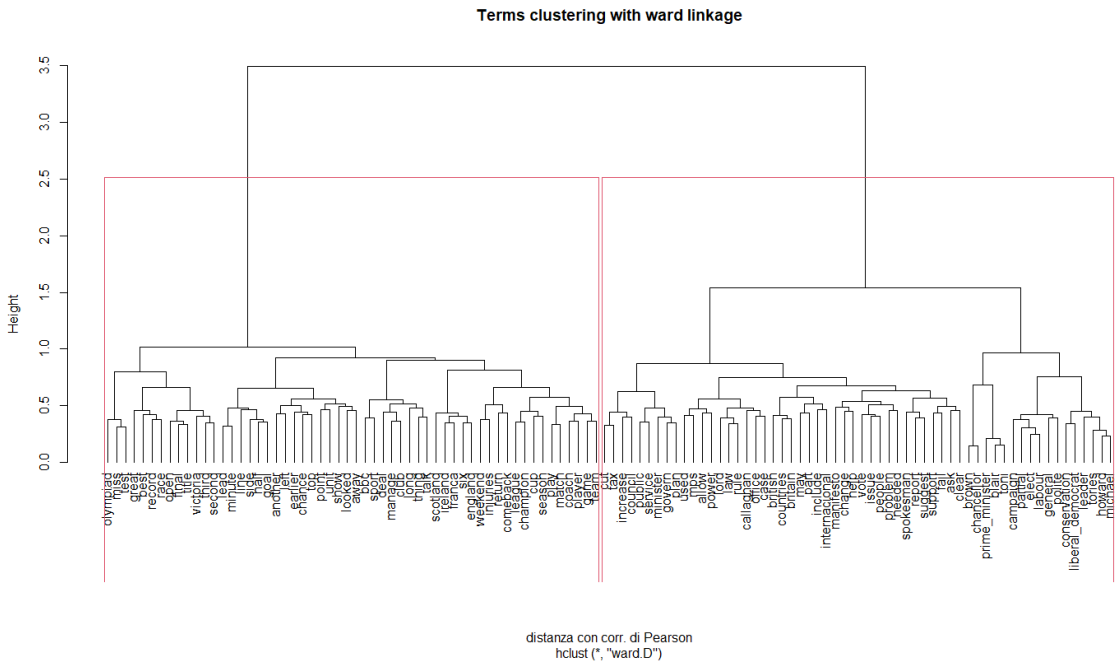

Figura 19: dendogramma con clusters per la distanza basata sul coeff. di correlazione di Pearson con linkage di Ward

Tenendo conto del legame che abbiamo scelto di considerare per ogni distanza, la silhouette media viene massimizzata dalla distanza di manhattan con legame completo. Tuttavia, è possibile osservare che il contenuto dei clusters che abbiamo ottenuto nei tre casi riportati è molto simile. Questo conferma quanto detto in precedenza, vale a dire che gli indicatori quantitativi non sempre sono adatti a valutare la qualità dei metodi applicati nel text mining.

Clustering partizionale

Nel clustering partizionale abbiamo applicato l'algoritmo k-means alla matrice Term Document completa di tutti i termini. Attraverso una funzione scritta da noi abbiamo calcolato la silhouette media per diversi valori di k per trovare il numero di gruppi che la massimizzasse. Come risultato abbiamo ottenuto un k ottimale pari a 4 con una silhouette media pari a 0.315, un valore abbastanza basso.
Abbiamo rappresentato i quattro gruppi attraverso un clusplot e abbiamo osservato che uno dei gruppi è composto solo da 4 parole che potrebbero essere degli outliers ("game", "play", "england", "player"). Abbiamo quindi provato a rimuoverli per vedere se si poteva ottenere una silhouette migliore o se cambiava il numero di cluster ottimale.
Ripetendo la funzione per trovare il valore di k che massimizza la silhouette media il risultato è stato k=3 con un valore dell'indicatore pari a 0.371, leggermente superiore a quella precedente. Successivamente, abbiamo rappresentato nuovamente i gruppi e ne abbiamo analizzato il contenuto.

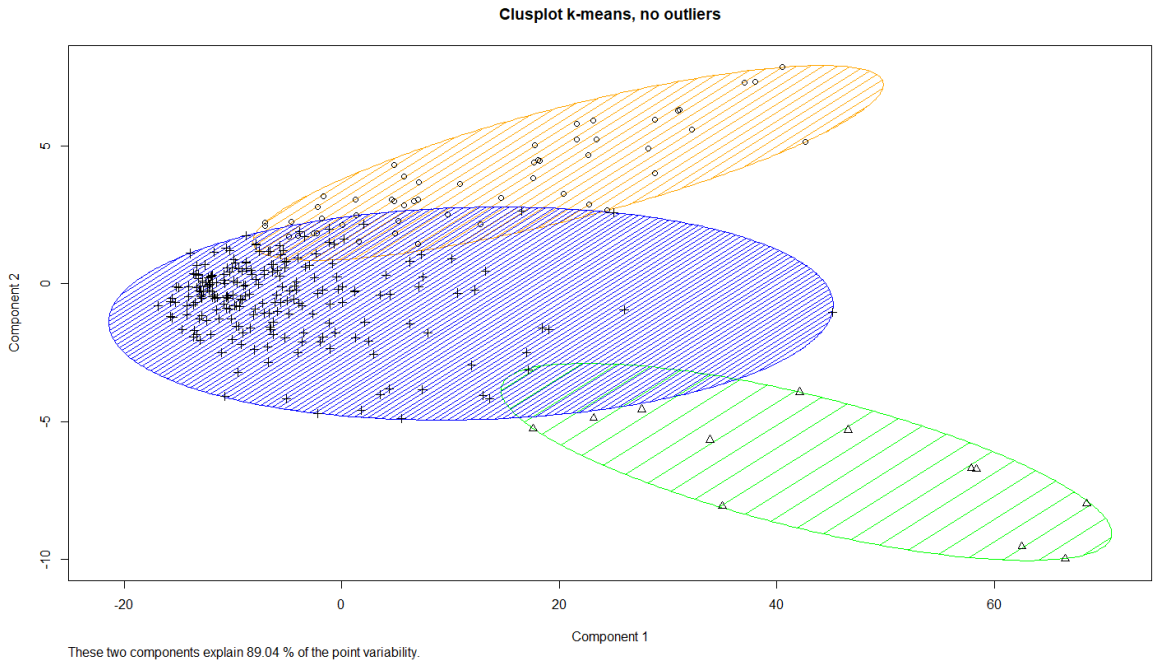

Figura 20: clusplot dei gruppi risultanti da k-means, senza outliers

Si osserva che un gruppo contiene molti più termini degli altri 219 rispetto a 12 e 52. Di seguito riportiamo alcuni dei termini contenuti in ogni gruppo.

Cluster 1 final, coach, champion, cup, team, olympiad, league, beat, kick, season, record, match, side, injuries, half, second, open, minute, club, championship, score …

Cluster 2 blair, brown, elect, govern, labour, liberal_democrat, partial, plan, tories, tax, people, howard

Cluster 3 prime_minister, programme, minister, spokesman, chancellor, council, football, train, lord, sport, campaign, support, bbc_radio, defend, perform, vote …

Notiamo che il cluster 1 contiene termini relativi allo sport, mentre il cluster 2 relativi alla politica. Il cluster 3, invece, contiene parole che riguardano entrambi gli argomenti. Questa struttura si può osservare anche dal grafico dove si vede che il gruppo 1 (giallo) e il gruppo 2 (verde) sono distanti tra loro, mentre il gruppo 3 (blu) si trova in mezzo agli altri due.

3. Classificazione dei documenti

Come ultimo punto della nostra analisi abbiamo deciso di provare a classificare i documenti; in questa fase abbiamo tenuto conto della conoscenza della reale natura degli articoli in quanto necessaria per l'applicazione dei metodi di classificazione.
Sono stati esclusi l'utilizzo della regressione logistica dato l'elevato numero di termini, che costituiscono le variabili, e anche l'uso dell'analisi discriminante perché sia quella lineare che quella quadratica richiedono, tra le loro assunzioni di base, che le variabili esplicative siano distribuite normalmente. Ciò non accade nel nostro caso in quanto le variabili sono discrete.
Il metodo scelto è quindi il K-nearest neighbour che risulta adatto in quanto non prevede assunzioni per l'applicazione (lazy learning algorithm); l'unica scelta da compiere è il valore del parametro di tuning.
Per applicare il metodo di classificazione abbiamo creato una nuova variabile chiamata "type" di tipo dicotomico; al valore 0 corrisponde la categoria "politica" e a 1 la categoria "sport". Questa variabile è stata aggiunta alla Document Term Matrix che abbiamo poi trasformato in un dataframe (928x288). Successivamente, abbiamo diviso il dataset in training set e test set (rispettivamente 70% e 30% delle osservazioni).
Nella fase di training per l'addestramento del modello abbiamo utilizzato un ciclo for per testare diversi valori di k (da 2 a 15) per individuare quale fosse il migliore in termini di accuratezza. Il numero di "vicini" da considerare per ottenere il massimo livello di accuratezza è k=2 con un'accuratezza pari al 96.31%. Il valore k=1 è stato escluso in quanto genera overfitting.
Una volta individuato il miglior valore di k, è possibile valutare il KNN nel test set per capire la sua capacità di generalizzare. Per comprendere meglio i risultati ottenuti riportiamo le confusion matrix del modello per il training e il test set e i valori degli indicatori principali.

TRAINING

		Predicted			
		Politica	Sport	Accuracy	0.9631
True class	Politica	263	24	Sensitivity	1.0000
	Sport	0	363	Specificity	0.9380

TEST

		Predicted			
		Politica	Sport	Accuracy	0.8921
True class	Politica	101	29	Sensitivity	0.9902
	Sport	1	147	Specificity	0.8352

Osserviamo che l'accuratezza si riduce dal training al test e quindi il training error è inferiore al test error. Generalmente si desidera la situazione opposta, ma anche provando ad usare un k superiore la situazione non cambia, l'accuracy diminuisce dal training al test.
Notiamo anche che i documenti di sport vengono classificati molto bene (come evidenziato dalla sensitivity) mentre la classificazione degli articoli di politica presenta qualche problema in più perché è maggiore il numero di osservazioni a cui è stata assegnata la classe sbagliata (specificity).
Per approfondire la valutazione del risultato ottenuto abbiamo utilizzato la curva di ROC e calcolato l'AUC (area sottesa alla curva). Come valore dell'AUC abbiamo ottenuto 0.9259, un valore abbastanza alto. Visti i valori degli indicatori, abbiamo ritenuto che il metodo KNN con k pari a 2 classifichi in modo adeguato le osservazioni.

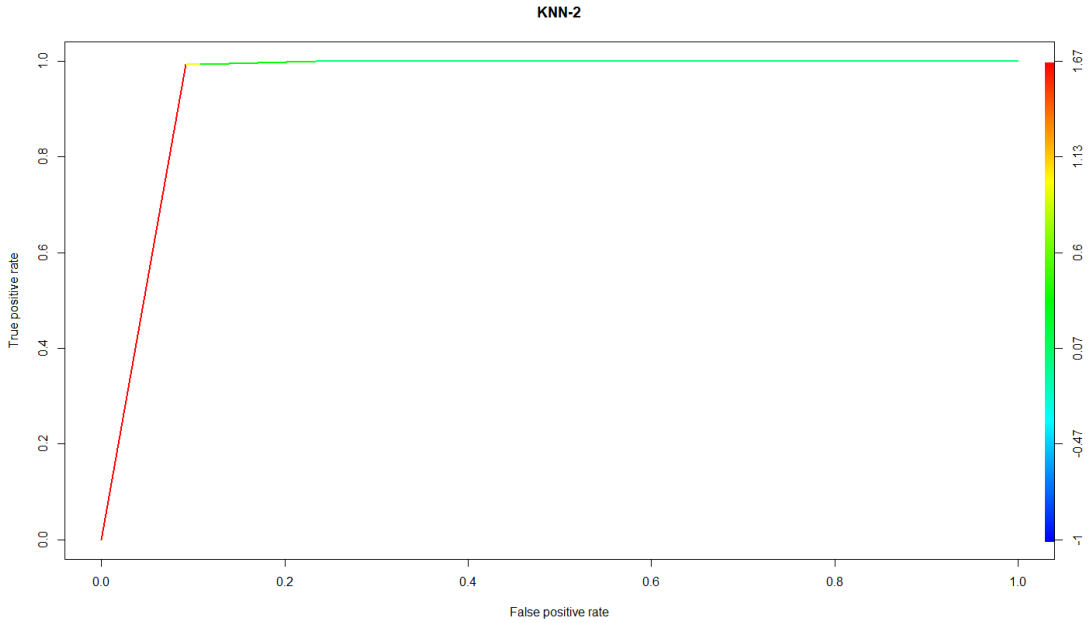

Figura 21: ROC curve per la valutazione del KNN con k=2

4 Discussioni

Abbiamo analizzato un problema di text mining relativo ad una serie di articoli della BBC relativi a due argomenti, sport e politica, con l'obiettivo finale di ricavare informazioni utili e di individuare un metodo che permettesse la discriminazione degli articoli sulla base del loro contenuto. Nel corso dell'analisi abbiamo osservato come i metodi di clustering, sia di tipo gerarchico agglomerativo che partizionale, se valutati con indicatori quantitativi risultino poco adeguati a raggruppare articoli, mentre in realtà, dal punto di vista grafico e contenutistico, i gruppi ottenuti distinguono bene le tipologie di articoli e i termini utilizzati. Questo dimostra come gli indicatori quantitativi non sempre siano adatti alla valutazione dei metodi di text mining. Alcuni metodi di clustering sono stati applicati anche ai termini per osservare se si potessero individuare gruppi che indicassero la presenza di sotto-argomenti rispetto alle due categorie generali. I risultati ottenuti hanno evidenziato che il metodo di clustering gerarchico individua solo due gruppi di termini, generalmente uno per argomento, mentre il metodo di clustering partizionale individua tre gruppi. Quest'ultimo risultato ci ha permesso di osservare che ci sono due gruppi che raccolgono un numero abbastanza contenuto di termini che riguardano ciascuno una delle due categorie principali. Il terzo gruppo, invece, racchiude il numero maggiore di parole, che sembrano potersi adattare per la maggior parte ad entrambi gli argomenti. Infine, dalla valutazione del metodo scelto per la classificazione dei documenti, il KNN, è emerso che è possibile ottenere una buona classificazione dei documenti seppur commettendo alcuni errori.