Text Mining applicato ad articoli sportivi della BBC

Fratti Giorgio, 830518

27 Novembre, 2019

Abstract

In questo progetto, mi sono soffermato sull'analisi di 737 articoli di sport della BBC, divisi in 5 categorie: athletics, cricket, football, rugby e tennis, relativi al periodo 2004-2005. Ho svolto il lavoro in più fasi. La prima fase è la preparazione dei dati, che comprende la pulizia dei documenti (rimozione stopwords e stemming) e le analisi preliminari per verificare la qualità dei dati e le caratteristiche principali. Successivamente ho usato la tecnica del clustering gerarchico agglomerativo per cercare di raggruppare i documenti nelle 5 categorie principali senza, però, supporre di conoscere la reale natura dei dati. Nella fase successiva, sotto l'ipotesi di conoscere la reale classificazione degli articoli, ho provato a classificare i documenti attraverso il classificatore K-NN. Ho applicato nuovamente un metodo non supervisionato: il Latent Dirichlet Allocation, per classificare gli articoli in 5 macro-argomenti estratti dall'insieme dei documenti. Ho concluso il lavoro con una Sentiment Analysis per capire quali siano i sentimenti più ricorrenti negli articoli.

1 Introduzione

L'uomo ha sviluppato competenze relative alla comprensione dell'ambiente e la capacità di agire in relazione agli eventi circostanti. Nel data mining si vuole dare queste capacità a software e macchine. In questo caso in particolare, per estrarre informazioni riguardo il contenuto di un insieme di testi e per la loro categorizzazione, si è sviluppata la tecnologia del Text Mining. Nel lavoro attuale, si vuole cercare, proprio attraverso le tecniche del text mining, di estrarre informazioni dai testi che permettano la corretta classificazione e il raggruppamento di essi.

2 Materiali e metodi

2.1 Materiali

Il dataset prescelto per l'analisi consiste in un insieme di 737 documenti provenienti dal sito della BBC Sport reso disponibile per usi non commerciali e al solo scopo di ricerca. (D. Greene and P. Cunningham. "Pratical Solutions to the problem of Diagonal Dominance in Kernel Document Clustering", Proc. ICML 2006.). I documenti corrispondono ad articoli sportivi datati 2004-2005. Di tali articoli 101 sono relativi alla categoria Athletics, 124 parlano di Cricket, 265 trattano principalmente Football, 147 sono relativi al Rugby e gli ultimi 100 articoli parlano di Tennis. Al fine di non perdere l'informazione riguardante la reale natura degli articoli, i documenti sono stati salvati con l'iniziale dell'argomento seguito da un numero progressivo. ('a(numero).txt' se articolo di athletics, 'c(numero).txt' se parlano di cricket, ecc.).

2.2 Metodi

Dopo il caricamento iniziale del Corpus, prima di procedere con l'analisi degli articoli, ho provveduto alla pulizia dei documenti. Per prima cosa ho eliminato i simboli e li ho sostituiti con 'spazio' per evitare che si formassero nuove parole. Successivamente ho rimosso i segni di punteggiatura, ho trasformato tutte le lettere in minuscole, ho rimosso i numeri ed ho rimosso le stop-words utilizzando sia la lista standard "english" fornita da R, sia una lista di parole da me individuate grazie a delle indagini preliminari sulle parole più frequenti. Questa lista personalizzata contiene giorni, numeri, avverbi e aggettivi, che avrebbero potuto distorcere negativamente i risultati delle analisi. In seguito, ho provveduto allo stemming delle parole, che vengono troncate per recuperare la radice della parola e infine ho rimosso gli spazi bianchi in eccesso.

Dopo la fase di text cleaning ho creato la Document-Term Matrix, una matrice che ha come righe i documenti D e come colonne i termini T, presenti all'interno dei documenti, con all'interno la frequenza di apparizione dei termini di ciascun documento. Ho proseguito eseguendo le analisi preliminari come il conteggio delle frequenze di apparizione delle parole su tutti i documenti e il numero di termini/parole apparse in ogni documento. Ho visualizzato la frequenza di apparizione dei termini più popolari e le prime 200 parole con frequenza maggiore le ho visualizzati attraverso barplot e wordcloud. Infine, ho preferito creare una seconda Document-Term Matrix, contenente le parole con 3 o più caratteri e meno di 20, e che siano apparsi in almeno 10 documenti ed al massimo 700.

L’analisi dei documenti si svolge in sei fasi:

1. Clustering dei documenti;
2. Classificazione dei documenti;
3. Topic modelling;
4. Sentiment Analysis.

Clustering dei documenti

Tramite un metodo non supervisionato ho provato a suddividere i 737 articoli in modo sensato, provando a categorizzarli nelle loro 5 categorie base. Un metodo non supervisionato è principalmente un modello che non ha come input la variabile target. Un esempio è il clustering, che raggruppa le osservazioni che condividono caratteristiche simili sulla base di criteri precisi. Ogni gruppo (cluster) consiste in oggetti che sono simili tra loro e dissimili con oggetti di altri gruppi. Un algoritmo di clustering deve essere scalabile, capace di trattare differenti tipi di dati e di individuare gruppi di natura diversa e deve essere insensibile al numero di attributi. Nel mio caso ho deciso di usare il clustering gerarchico agglomerativo: si crea una decomposizione gerarchica degli oggetti. Ogni osservazione è considerata un cluster, ad ogni step la migliore coppia di osservazioni (minor distanza) finisce dentro ad un nuovo cluster. Questo procedimento è ripetuto fino a che tutti i cluster sono fusi assieme.

Ho provato 4 metriche diverse: 2 classiche matrici di distanza quali la distanza euclidea e la distanza di Manhattan e due matrici di distanza basate su indici di similarità quali la correlazione e la distanza cosine. Come metodo per calcolare la distanza tra clusters ho testato 4 diversi linkage: legame singolo, legame completo, legame medio e legame di Ward, e per ciascuna combinazione ho costruito il dendogramma. Ho tagliando il dendogramma creando 5 gruppi e ho verificato la qualità dell’analisi attraverso la silhouette, che è una misura di quanto gli elementi di un gruppo siano simili tra di loro e dissimili tra gli elementi degli altri gruppi. La silhouette assume valori compresi tra -1 e 1: a una osservazione ben assegnata corrisponde un valore di silhouette elevato.

Classificazione dei documenti

La classificazione rientra tra i metodi supervisionati. A differenza del metodo precedente, un metodo supervisionato utilizza dati in cui è nota la reale natura degli articoli, che nel mio caso ho inserito nella Document-Term matrix sotto forma di variabile che va da 0 a 4 (0=athletics, 1=cricket, 2=football, 3=rugby, 4=tennis). Come metodo di classificazione ho scelto il classificatore K – NN, un algoritmo non parametrico chiamato lazy-learning algorithm. L’unico

parametro presente nel K-NN è il parametro k, detto parametro di tuning. Questo algoritmo classifica una nuova osservazione sulla base della distanza euclidea minima dai suoi vicini. Prima si calcola la distanza rispetto a tutte le osservazioni del training, si individuano le k osservazioni più vicine e si classifica la nuova nella classe più frequente tra quelle vicine.

Prima di svolgere la classificazione basata su K-NN ho diviso la Document-Term matrix in train set e test set, rispettivamente l'80% e il 20% delle righe della matrice di partenza (contenente la colonna con la variabile target). Successivamente ho applicato l'algoritmo K-NN, ottimizzando il k in termini di accuracy.

Topic Modeling

Quello di modellare i topic è un metodo non supervisionato per raggruppare i documenti sulla base di argomenti specifici. Il modello che ho usato è LDA (Latent Dirichlet Allocation) e consiste in un metodo matematico che cerca di trovare k argomenti principali prefissati sulla base di termini che dovrebbero descrivere ciascun argomento. Prima di tutto, ho cercato 5 topic principali e, dopo aver verificato che potessero corrispondere alle 5 classi di sport in cui volevo suddividere gli articoli, ho verificato se effettivamente i documenti venissero classificati correttamente.

Sentiment Analysis

La sentiment analysis si riferisce all'elaborazione del linguaggio e all'analisi del testo per identificare informazioni soggettive nelle fonti. In generale l'obiettivo della sentiment analysis è determinare le emozioni principali che traspaiono da un documento.

3 Risultati

Per prima cosa ho caricato i documenti grazie al comando Corpus. Prima di procedere con le prime analisi, al fine di renderle più efficienti, ho eseguito una pulizia dei documenti. Per eliminare i simboli ho usato una funzione che sostituisce determinati simboli con uno "spazio". Successivamente ho eliminato i segni di punteggiatura, ho trasformato tutte le lettere in lettere minuscole ed ho rimosso i numeri. Per la rimozione delle stop-words mi sono affidato inizialmente alla lista standard "english" fornita da R e, successivamente, ho usato anche una lista da me creata contenente mesi, giorni, numeri e i principali verbi, avverbi e aggettivi inglesi che comparendo

con maggiore frequenza avrebbero potuto distorcere negativamente le future analisi. Ho proceduto con lo stemming dei documenti ovvero con la riduzione delle forme flesse delle parole alla loro forma radice. Infine, ho provveduto ad eliminare gli spazi bianchi in eccesso.

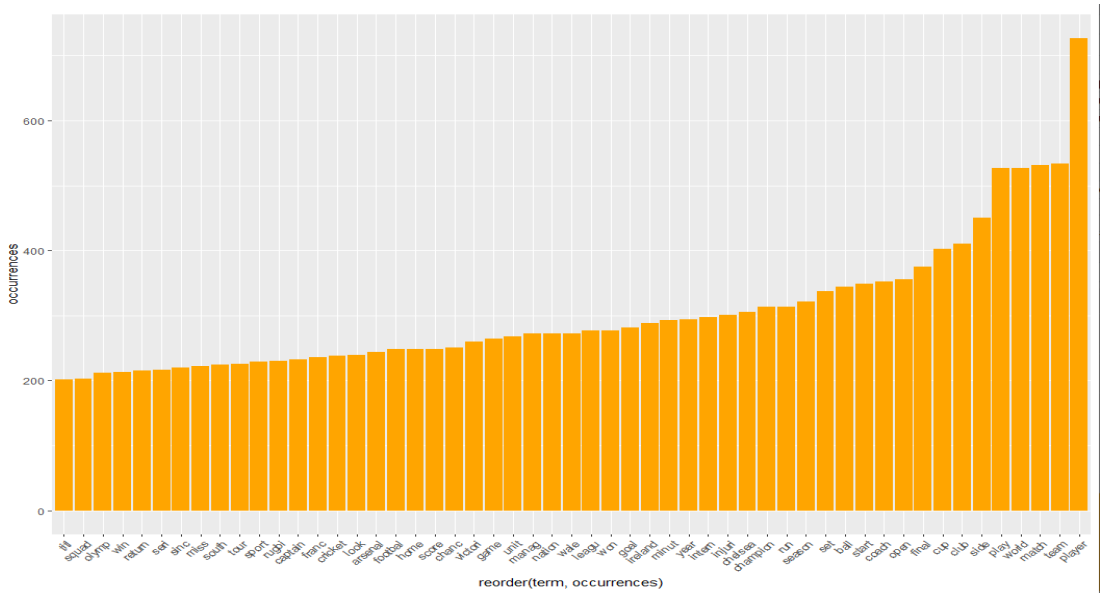
Terminata la fase di Text Cleaning del corpus, ho usato il comando DocumentTermMatrix per creare una matrice contenente la frequenza di tutti i termini presenti nel corpus per ogni documento, ottenendo una matrice di dimensioni 737x9648.

```
<<DocumentTermMatrix (documents: 737, terms: 9648)>>
Non-/sparse entries: 82237/7028339
Sparsity           : 99%
Maximal term length: 26
Weighting          : term frequency (tf)
Sample             :
```

Ho eseguito il conteggio delle frequenze di apparizione delle parole su tutti i documenti e ho calcolato il numero di termini/parole, apparse in ogni documento. Poi ho ordinato in maniera decrescente i termini in base alla loro frequenza di apparizione e ho verificato la frequenza di apparizione dei primi termini.

player	team	match	world	play	side
727	534	531	527	527	450

Ho visualizzato graficamente i 200 termini più frequenti tramite barplot e wordcloud.

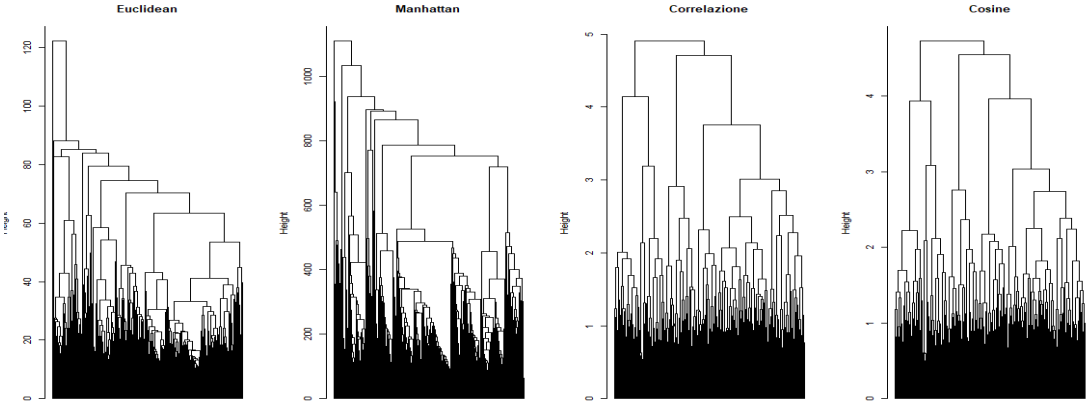


Prima di proseguire con le analisi ho preferito creare una seconda Document-Term Matrix dove ho considerato solo le parole con 3 o più caratteri ma meno di 20 e presenti in almeno 10 e massimo 700 documenti.

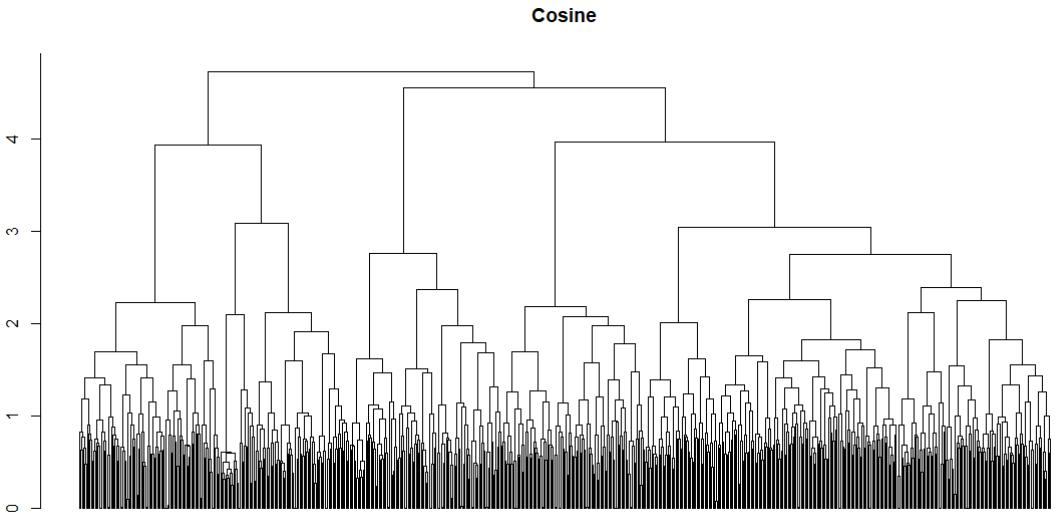
```
<<DocumentTermMatrix (documents: 737, terms: 1806)>>
Non-/sparse entries: 62801/1268221
Sparsity           : 95%
Maximal term length: 15
Weighting          : term frequency (tf)
Sample             :
```

Clustering gerarchico agglomerativo

Ho analizzato 4 distanze diverse (Euclidean, Manhattan, Correlazione, Cosine) e 4 diversi legami: delle 16 combinazioni possibili ho visualizzato tutti i dendogrammi. Dei 4 diversi legami, il legame di Ward è quello che mi ha restituito i dendogrammi migliori. Nel passare da k+1 gruppi a k gruppi la devianza intra-gruppo aumenta mentre la devianza tra i gruppi diminuisce. Ad ogni passo del metodo di Ward si aggregano tra loro quei gruppi per la cui vi è il minor incremento della devianza intra-gruppo. (immagine seguente: dendogrammi ottenuti con legame di Ward e 4 metriche diverse)



Le metriche migliori invece, si sono rivelate essere quelle basate su indici di similarità, ovvero quella basata sulla correlazione e quella basata sul coseno di similitudine (una tecnica per la misurazione della similitudine tra i due vettori effettuata calcolando il coseno tra di loro), la mia scelta è ricaduta su quest'ultima metrica. (immagine seguente: dendogramma ottenuto con legame di Ward e metrica cosine similarity)



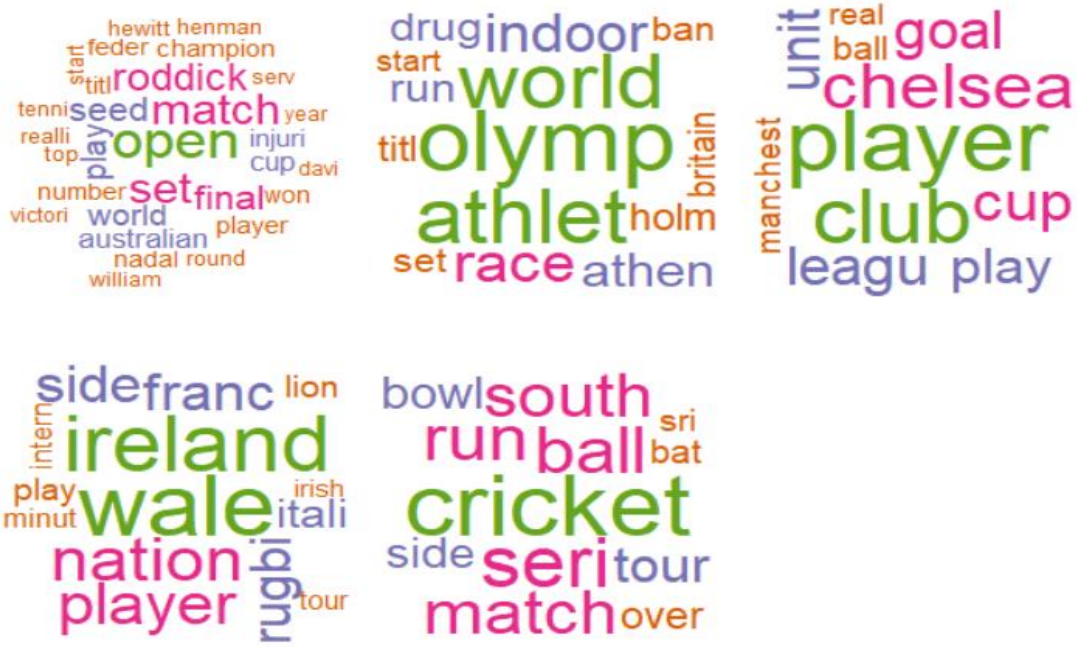
Per valutare la performance del metodo di cluster utilizzato ho deciso di controllare la silhouette, che ha valori da -1 a 1. Le osservazioni che hanno un'elevata silhouette, cioè con valore prossimo a 1, sono ben raggruppate. Se il valore della silhouette è compreso tra 0.26 e 0.5 la suddivisione perde di significatività. Nel nostro caso:

```
Silhouette of 737 units in 5 clusters from silhouette.default(x = clust, dist = dc) :
 Cluster sizes and average silhouette widths:
       107         98        308        108        116
0.07658162 0.09888372 0.02333740 0.08087853 0.09082337
Individual silhouette widths:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.10823  0.02684  0.06250  0.06017  0.09797  0.19048
```

la silhouette media è di 0.06. Essendo minore di 0.26 significa che non è stato individuato nessun pattern tra i gruppi. Però, dato che siamo a conoscenza della reale natura dei dati ho provato a verificare la percentuale di documenti classificati correttamente, ottenendo i seguenti risultati:

	sport	risultati
1	Athletics	0.9306931
2	Cricket	0.9274194
3	Football	0.9735849
4	Rugby	0.6870748
5	Tennis	0.9200000
6	Tot	0.8955224

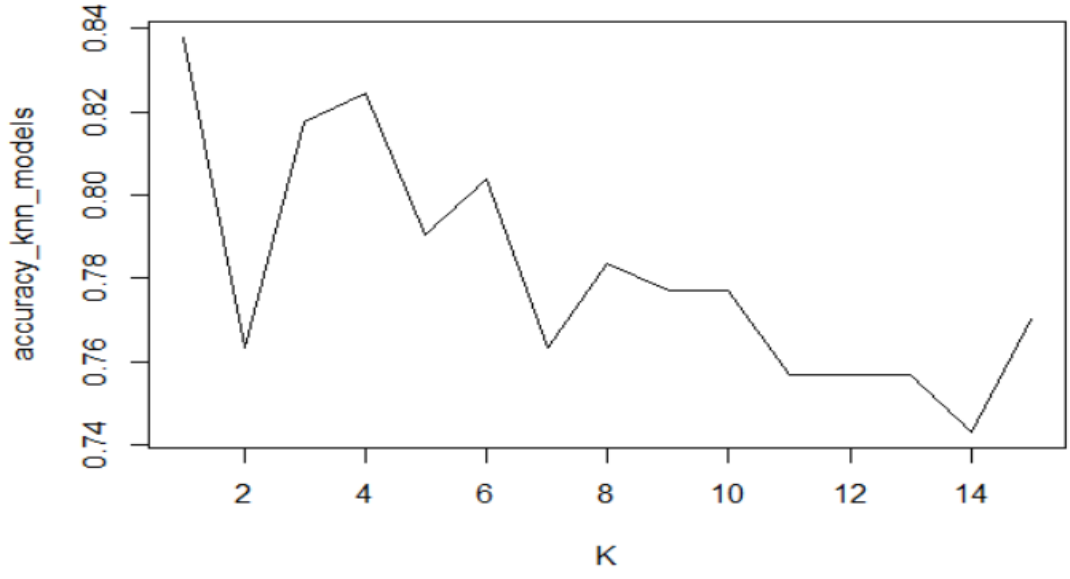
Ho provato anche a visualizzare i wordclouds suddivisi per gruppi:



Classificazione K – NN

Per prima cosa, ho aggiunto alla matrice usata precedentemente la colonna contenente la variabile target. Ho suddiviso la matrice in train set e test set, rispettivamente l'80% e il 20%.

Per scegliere il parametro di tuning (k) ho calcolato l'accuracy per k da 1 a 15:



Il parametro che mi consente di avere una accuratezza migliore è il parametro k=1, che considera quindi solo l'osservazione più vicina. Basandoci solo su quella però, si può incorrere in overfitting che potrebbe essere un grosso problema. Per evitare ciò è preferibile rilassare le linee di confine considerando più di un vicino. Per questa analisi ho considerato k=4, la scelta che mi consente di avere una accuracy migliore se si considera più di un vicino. Dopo avere scelto il parametro di tuning, il classificatore K-NN calcola la distanza tra un nuovo punto e tutti i punti del training, ordina le distanze e individua le k osservazioni più vicine e ne individua la loro categoria. A questo punto la classificazione sarà effettuata sulla base della classe più presente nell'intorno di dimensione k.

Per valutare le performance di questo classificatore, non avendo una variabile target dicotomica, ho deciso di affidarmi alla confusion matrix e sulle statistiche relative ad essa, in particolare sulla accuracy di ogni classe, ovvero (numero documenti classificati correttamente nella classe+numero documenti classificati correttamente non appartenenti alla classe)/numero totale dei documenti. (True Positive + True Negative)/Total population.

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1  2  3  4
         0 14  0  0  0  0
         1  1 23  0  0  7
         2  0  0 41  1  3
         3  2  0  7 18  7
         4  0  0  1  1 22
```

```
Overall Statistics

               Accuracy : 0.7973
                 95% CI : (0.7234, 0.8589)
    No Information Rate : 0.3311
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.7408

 Mcnemar's Test P-Value : NA
```

```
Statistics by Class:

                     Class: 0 Class: 1 Class: 2 Class: 3 Class: 4
Sensitivity           0.82353   1.0000   0.8367   0.9000   0.5641
Specificity           1.00000   0.9360   0.9596   0.8750   0.9817
Pos Pred Value        1.00000   0.7419   0.9111   0.5294   0.9167
Neg Pred Value        0.97761   1.0000   0.9223   0.9825   0.8629
Prevalence            0.11486   0.1554   0.3311   0.1351   0.2635
Detection Rate        0.09459   0.1554   0.2770   0.1216   0.1486
Detection Prevalence  0.09459   0.2095   0.3041   0.2297   0.1622
Balanced Accuracy     0.91176   0.9680   0.8982   0.8875   0.7729
```

Topic Modeling

Il topic model è una classe di tecniche di text mining non supervisionate; esistono diversi algoritmi di topic modeling, il più diffuso e quello che ho scelto per la mia analisi è quello noto come Latent Dirichlet Allocation (LDA). In LDA un testo è una distribuzione di probabilità su un insieme di topic, un topic è una distribuzione di probabilità su un insieme di parole. Questo modello generativo è in grado di estrapolare automaticamente le topic presenti in un insieme di documenti e la loro distribuzione statistica nel corpus. Nel mio caso, ho provato a ricavare dall'insieme di articoli, 5 argomenti principali e le principali parole che li descrivono.

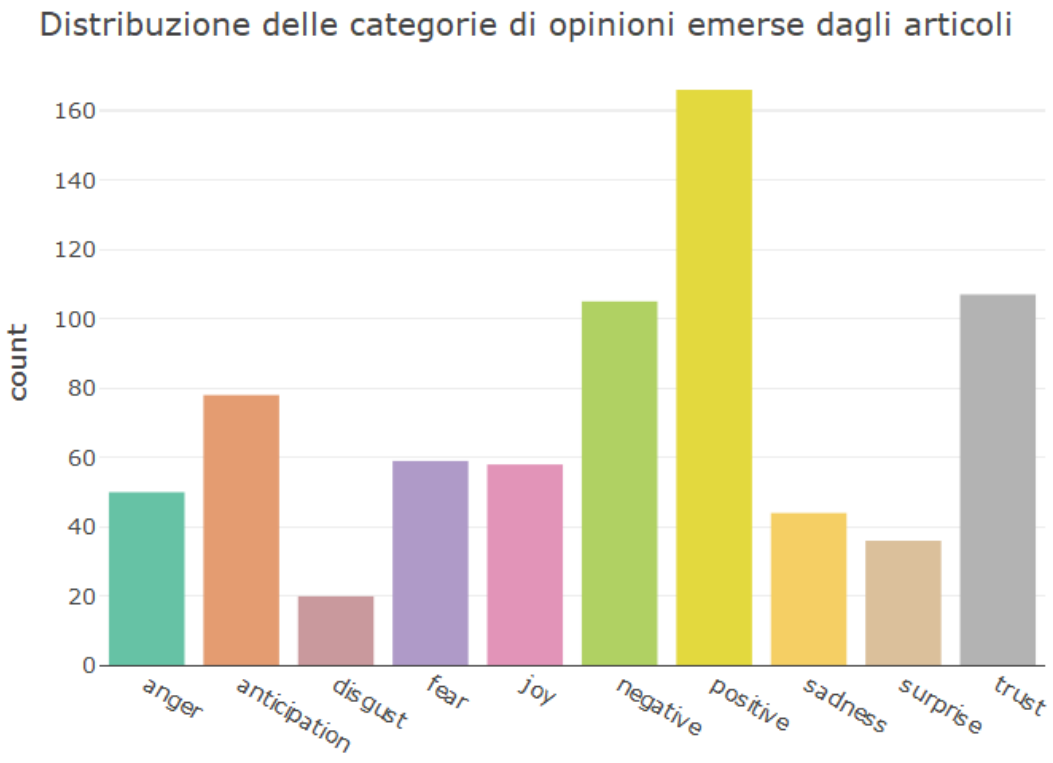
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	olymp	open	player	player	ball
2	minut	match	club	wale	run
3	athlet	set	chelsea	rugbi	seri
4	world	play	unit	team	south
5	race	ireland	leagu	cricket	pakistan
6	indoor	final	manag	coach	wicket

I 5 topic estrapolati sembrano corrispondere alle classi di interesse in cui siamo interessati a classificare i documenti. Topic 1 corrisponde all'argomento Athletics, Topic 2 corrisponde a Tennis, Topic 3 corrisponde a Football, Topic 4 corrisponde a Rugby e Topic 5 corrisponde a Cricket. Sempre grazie a LDA ho quindi assegnato ad ogni documento il topic che ha la maggior probabilità di descrivere l'articolo. Infine, essendo a conoscenza della reale natura degli articoli ho calcolato il numero di documenti classificati correttamente, ottenendo i seguenti risultati:

	sport	risultati
1	Athletics	0.8316832
2	Cricket	0.6693548
3	Football	0.7811321
4	Rugby	0.4965986
5	Tennis	0.9200000
6	Tot	0.7313433

Sentiment Analysis

L'analisi dei sentimenti è un nuovo campo del text mining, che cerca di estrarre opinioni dai testi. Ho svolto una rapida ed elementare analisi con lo scopo di verificare quali fossero le opinioni che emergono con più frequenza dagli articoli. Per l'analisi ho utilizzato il dizionario NRC, messo a disposizione da R, che contiene numerose parole inglesi a cui viene associato o meno la presenza di determinate emozioni (positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise e trust). Ho visualizzato i risultati tramite barplot.



Infine, ho deciso di visualizzare, attraverso Wordclouds, le parole associate ai quattro sentimenti che emergono con maggiore frequenza (positive, trust, negative, anticipation).

4 Discussioni

Nel presente elaborato, ho affrontato un problema di text mining, a partire da 737 articoli sportivi messi a disposizione dalla BBC sezione Sport. Ho utilizzato un metodo di clustering gerarchico agglomerativo per cercare di raggruppare gli articoli nei 5 gruppi principali in cui sono classificati. Attraverso questo metodo non è stato individuato alcun pattern tra i gruppi anche se in realtà essi non si discostano più di tanti dai gruppi reali di cui fanno parte. Ho deciso poi, di applicare un metodo supervisionato, sfruttando la conoscenza della reale natura degli articoli, per classificare i documenti. Ho utilizzato il classificatore K – NN che si è rivelato un buon stimatore dal punto di vista dell'accuracy del modello. Il metodo non supervisionato di modellare i topic mi ha restituito cinque topics riconducibili ad argomenti di atletica, cricket, calcio, rugby e tennis. Anche in questo caso, la conoscenza pregressa sulla reale classificazione degli articoli, mi è stata utile per valutare la performance del modello, da cui è emerso che la classificazione dei documenti di determinate categorie è risultata molto più precisa rispetto ad altre. Dall'analisi finale sui sentimenti che traspaiono dagli articoli sportivi della BBC si evidenzia il prevalere di sentimenti positivi, seguiti da sentimenti di fiducia e sentimenti negativi.