UN'ANALISI TESTUALE SUGLI ARTICOLI DEL NEW YORK TIMES

Giulia Lumia, Giuliano De Blasi, Marco D'Antoni

Abstract

L'analisi in oggetto ha come scopo principale l'estrazione di informazioni utili da dati di tipo testuale, nello specifico dagli articoli pubblicati sul giornale New York times nell'arco della prima settimana del maggio 2016.
Applicando le principali tecniche di text mining è stato possibile ottenere una visione sia globale sia riferita ai singoli giorni di pubblicazione, con il vantaggio, tipico di questo tipo di analisi, di non dover leggere ogni singolo articolo per comprenderne l'argomento o il sentiment.
E' possibile ricondurre l'analisi a tre macrofasi, ciascuna delle quali caratterizzata da specifici obiettivi guida:

- ANALISI ESPLORATIVA, finalizzata ad ottenere informazioni strutturali sugli articoli, determinandone lunghezza e leggibilità ed evidenziandone le parole e i bigrammi più rilevanti;
- ANALISI DEL SENTIMENT, finalizzata all'individuazione del sentiment prevalente negli articoli, utilizzando appositi dizionari;
- ANALISI DEL TOPIC MODEL, finalizzata all'individuazione degli argomenti principalmente trattati negli articoli considerati.

1. Analisi esplorativa

Durante la fase esplorativa dell'analisi l'attenzione è stata incentrata su:

- Monogrammi, ossia le singole parole che compongono gli 845 articoli considerati;
- Bigrammi, ossia le coppie di parole consecutive;
- Lunghezza e leggibilità dei singoli articoli.

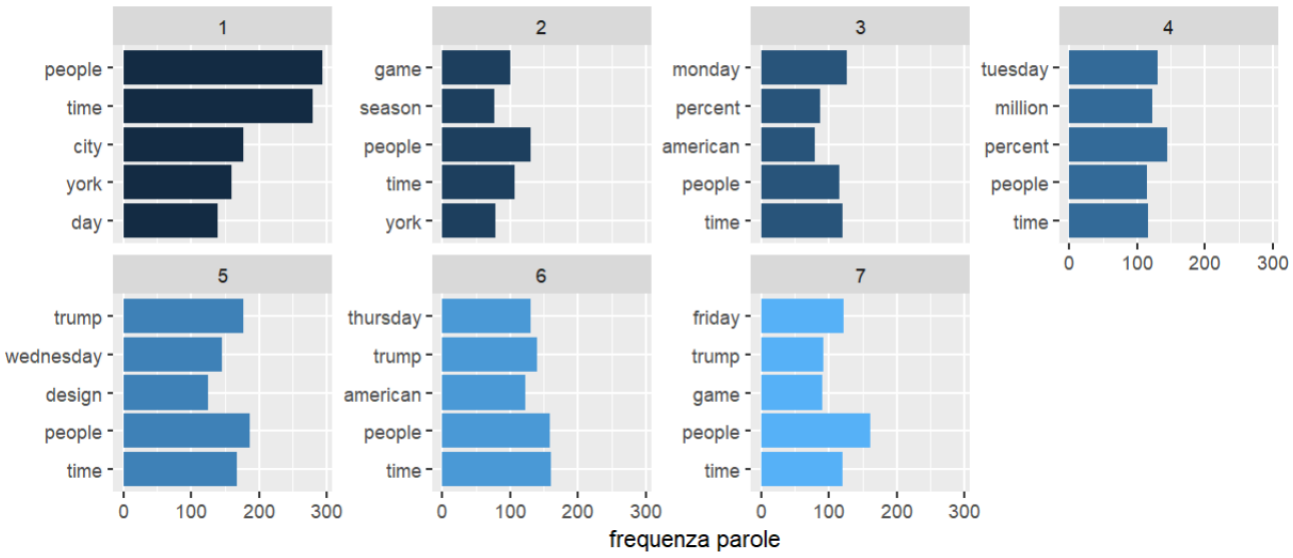
Per procedere all'analisi di monogrammi e bigrammi si è rivelato utile ignorare le cosiddette stop-words: parti del discorso come articoli, congiunzioni e preposizioni, presenti in maniera rilevante in qualsiasi testo e quindi incapaci di fornire informazioni d'interesse.

Il GRAFICO N. 1.1 è un wordcloud: una rappresentazione grafica "a nuvola" che permette di individuare le parole più spesso utilizzate in un dato testo, in questo caso nell'insieme degli articoli pubblicati.
La grandezza con cui le parole vengono rappresentate è proporzionale alla frequenza d'uso.
Osservando questo grafico è possibile capire quali siano le parole più rilevanti e intuire che i principali argomenti affrontati riguardano politica ed economia.

GRAFICO N. 1.1: wordcloud monogrammi

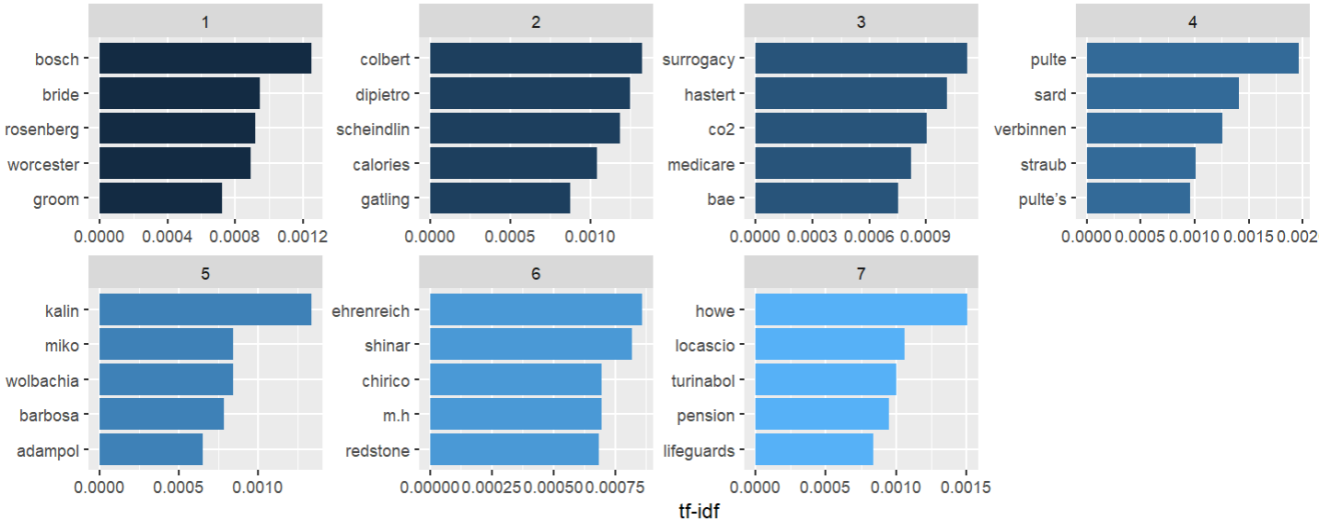
Dall'analisi del word barplot (GRAFICO N. 1.2), è possibile notare che "people" e "time", tra le parole più diffuse a livello globale considerando l'insieme degli articoli, sono presenti con una frequenza molto elevata anche negli articoli pubblicati in ciascuno dei sette giorni.
A partire dal terzo giorno di pubblicazione, evidente è la presenza della parola riferita al giorno della settimana (“Monday”, “Tuesday”, “Wednesday”, “Thursday”, “Friday”).
Anche in questo caso dall’individuazione delle parole più frequenti è possibile avere un’idea degli argomenti trattati: è verosimile pensare che in ciascun giorno siano stati affrontati più argomenti, ma che la domenica l’argomento principale sia stato lo sport (“game”, “season”, “time”).

GRAFICO N. 1.2: word barplot delle 5 parole più utilizzate giornalmente



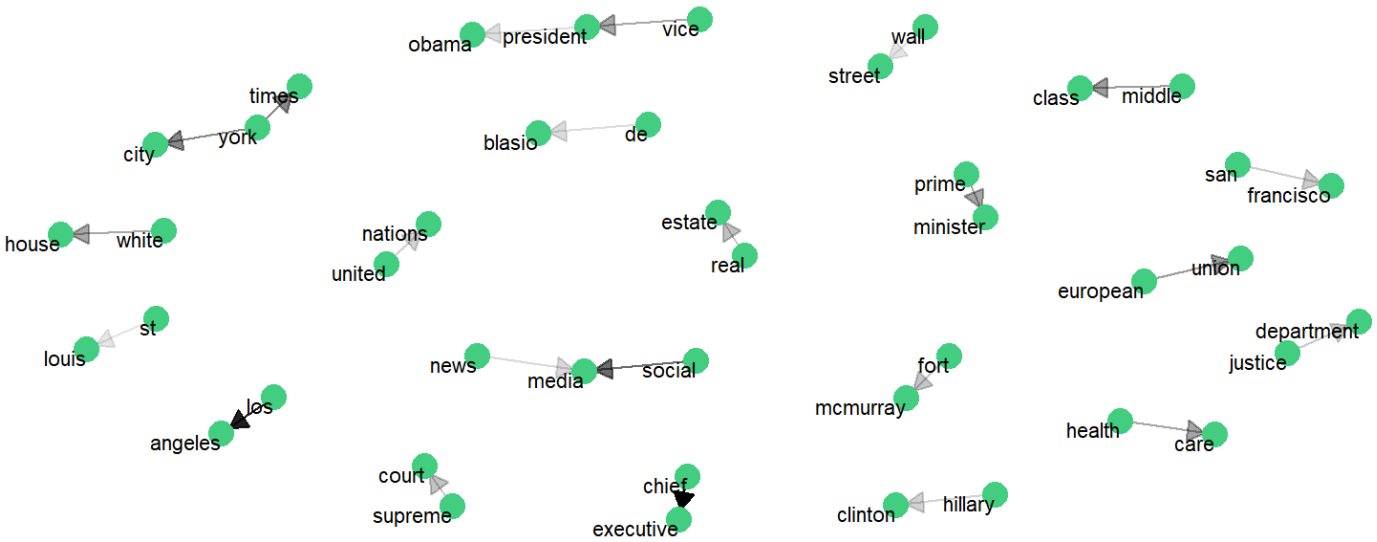
La frequenza con cui una parola viene utilizzata in un articolo può essere considerata indicatore della sua importanza. Quando però si dispone di più articoli è utile considerare quelle parole che caratterizzano ogni articolo o gli articoli di un determinato giorno.
Il calcolo della Term Frequency - Inverse Document Frecuency (TF – IDF) ha permesso di individuare quelle parole presenti con una certa frequenza negli articoli di un certo giorno di pubblicazione, ma al tempo stesso poco utilizzate negli altri. Il GRAFICO N. 1.3 mostra le parole che meglio contraddistinguono gli articoli pubblicati in un certo giorno rispetto a quelli pubblicati in quelli rimanenti.
E’ ragionevole supporre che le persone il cui cognome compare tra le seguenti parole siano state le protagoniste degli argomenti del giorno.

GRAFICO N. 1.3: word barplot delle 5 parole caratteristiche di ciascun giorno



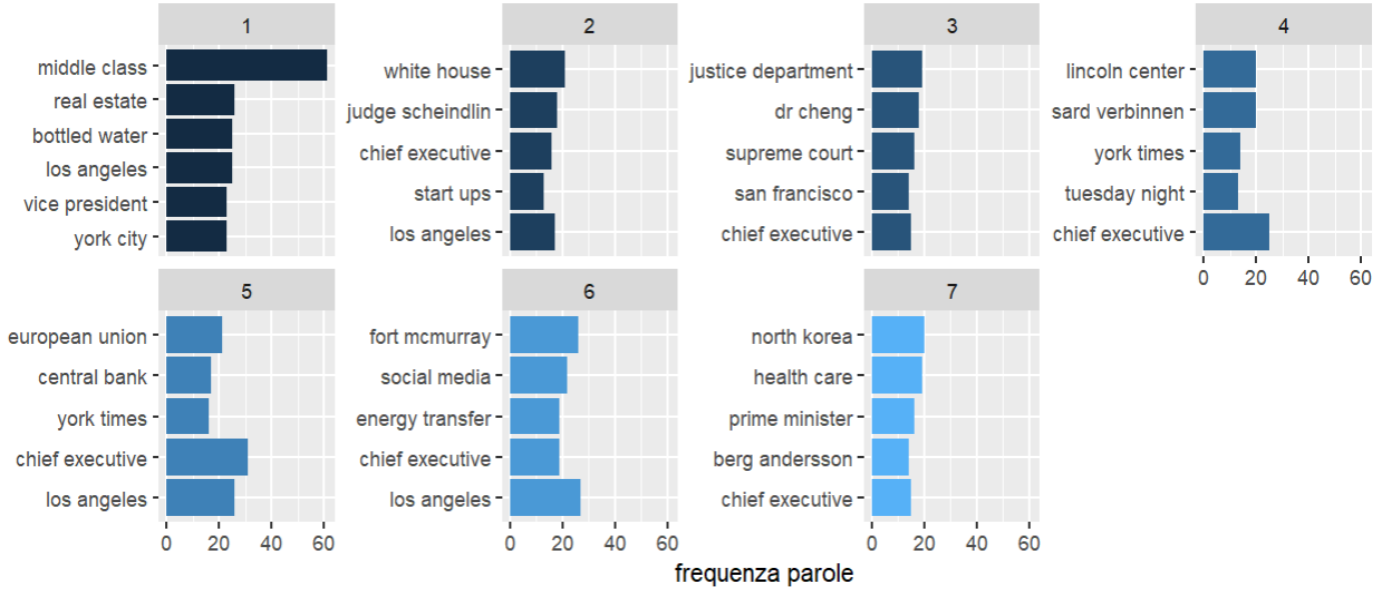
Il GRAFICO N. 1.4 è un text network: una rappresentazione costituita da nodi (parole) e ponti (legami), che permette di mostrare i bigrammi più rilevanti e la relazione asimmetrica esistente tra le parole che li compongono, ossia l'ordine secondo il quale si presentano negli articoli.
I bigrammi in figura sono presenti almeno 40 volte negli articoli considerati.
L'edge, ossia il collegamento tra i nodi, assume una trasparenza proporzionale alla rarità del bigramma per cui a frecce più scure corrispondono bigrammi più frequenti:
tra i bigrammi più rilevanti dal punto di vista della frequenza troviamo "los angeles", "chief executive", "european union" e "social media".

GRAFICO N. 1.4: text network dei bigrammi più frequenti



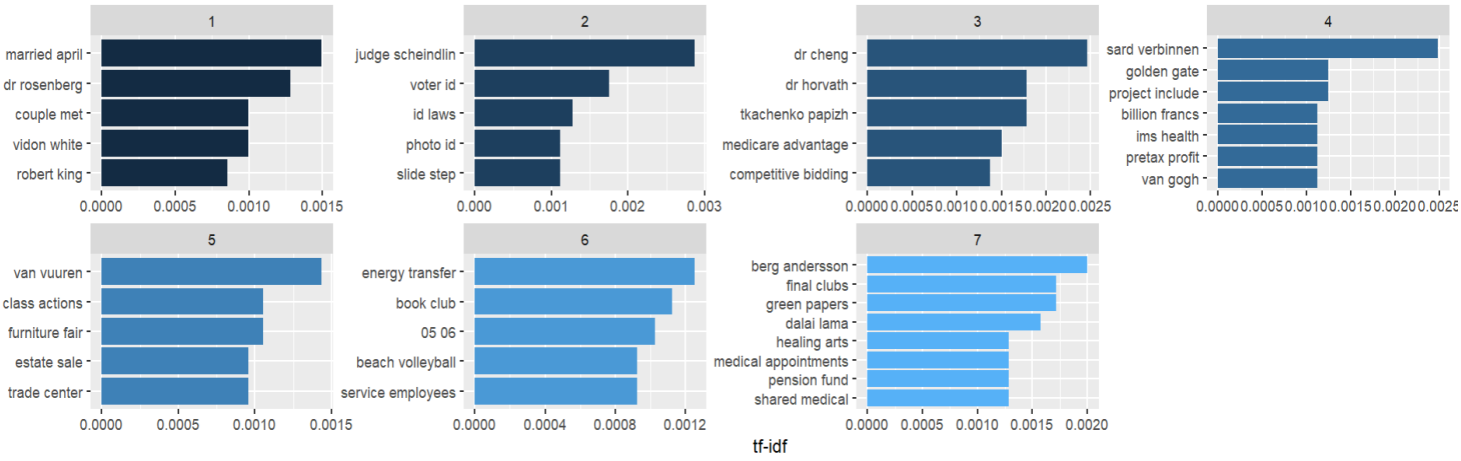
Dall'analisi del GRAFICO N. 1.5 è possibile individuare i bigrammi più utilizzati, condizionatamente a ciascun giorno di pubblicazione: da notare il fatto che il bigramma "chief executive" è quasi sempre presente mentre il bigramma "middle class" si presenta più del doppio delle volte rispetto agli altri .

GRAFICO N. 1.5: word barplot dei 5 bigrammi più utilizzati giornalmente



E’ possibile quantificare la capacità di ciascun bigramma di caratterizzare gli articoli pubblicati in un certo giorno da quelli pubblicati negli altri. Il GRAFICO N. 1.6 mostra le coppie di parole consecutive che negli articoli di ciascun giorno hanno maggiore potere discriminante: è possibile notare che bigrammi come “chief executive”, presenti negli articoli di quasi ciascun giorno, non sono stati rappresentati in quanto aventi un basso potere discriminante.

GRAFICO N. 1.6: word barplot delle 5 parole caratteristiche di ciascun giorno



Per ciascuno degli 845 articoli pubblicati nei 7 giorni considerati è stata determinata la lunghezza, data dal numero di parole in essi contenute (stop words comprese).
Ciò ha permesso di visualizzare sia la distribuzione complessiva (GRAFICO N. 1.7) sia la distribuzione condizionata rispetto al giorno di pubblicazione (GRAFICO N. 1.8).
La distribuzione delle lunghezze degli articoli è asimmetrica positiva:
la quasi totalità degli articoli ha una lunghezza di al più 2000 parole, ma sono presenti comunque dei valori nelle code dovuti ad articoli con lunghezza notevolmente elevata, maggiore anche a 4000 parole.

La distribuzione delle lunghezze condizionatamente ai giorni di pubblicazione permette di confermare quanto detto prima: la lunghezza degli articoli si mantiene quasi sempre sotto i 2000, raggiungendo talvolta dei picchi con degli articoli particolarmente lunghi, responsabili della coda della distribuzione complessiva.

GRAFICO N. 1.7: istogramma delle lunghezze suddivise in classi da 100

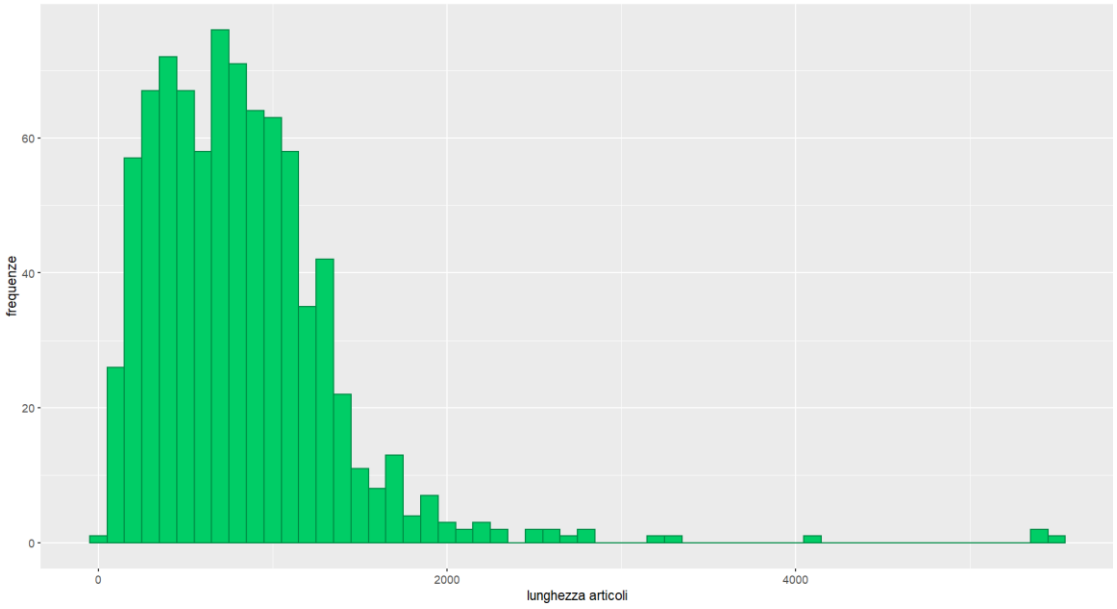
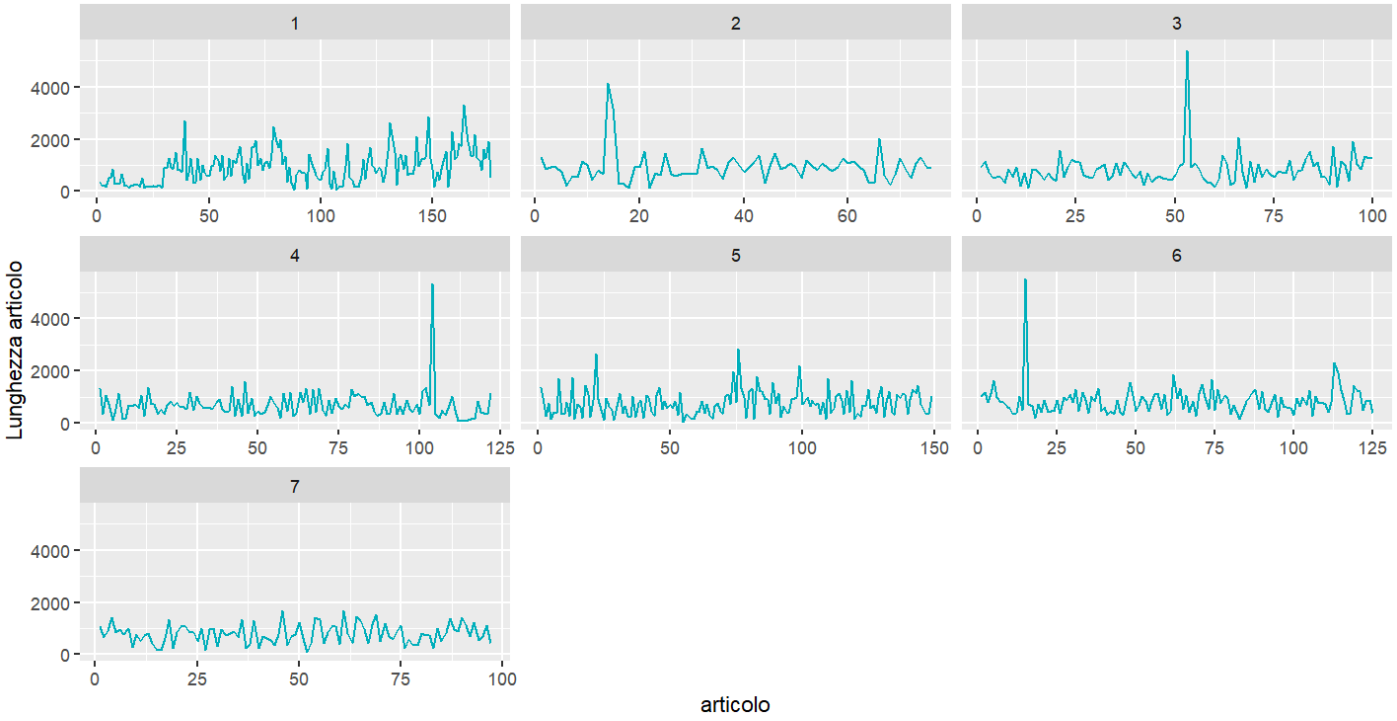
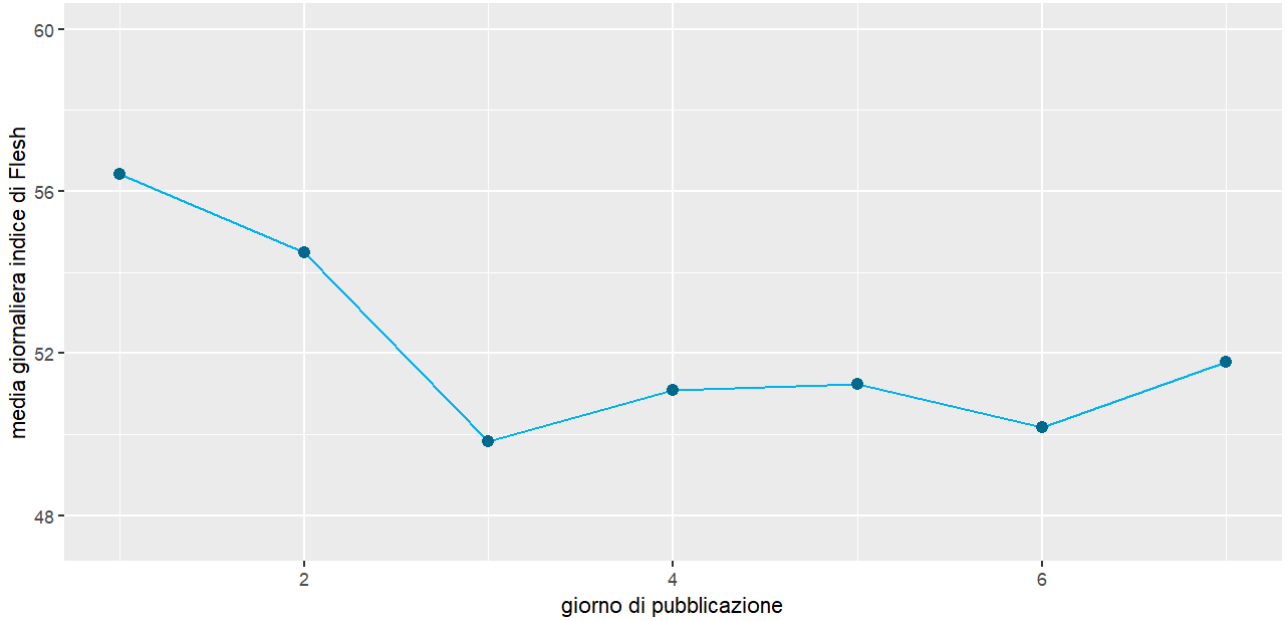


GRAFICO N. 1.8: andamento giornaliero della lunghezza degli articoli



Legata alla lunghezza di un articolo è la sua leggibilità, determinata tramite un apposito indice:
l'indice di leggibilità di Flesch è un indice che permette di quantificare la leggibilità di un testo, ossia la facilità con cui il lettore medio può riuscire a comprenderne il contenuto, tenendo conto del numero di frasi e parole utilizzate. Tale indice può assumere valori tra 0 e 100, suddivisibili in 10 classi:
a valori maggiori corrisponde una maggiore leggibilità.
Il GRAFICO N. 1.9 mostra il valore medio di leggibilità degli articoli pubblicati in ciascuno dei giorni di pubblicazione considerati.
I primi due giorni presentano in media degli articoli di maggiore comprensione, ma osservando il range di variazione nel corso della settimana, è possibile notare che il valore dell'indice di leggibilità si mantiene superiore a 50 punti e che pertanto gli articoli saranno in media comprensibili da coloro che hanno almeno un titolo di scuola superiore di secondo grado.

GRAFICO N. 1.9: andamento giornaliero dell'indice di Flesch di leggibilità



2. Analisi del sentiment

La sentiment analysis è una branca del text mining che permette di ottenere informazioni riguardo le emozioni suscitate dalle parole contenute in un testo, in questo caso gli articoli di un quotidiano.
Tra le varie applicazioni permette di capire se il loro contenuto fa riferimento a qualcosa di positivo o negativo.
Per determinare il sentiment di ciascun articolo sono stati presi in considerazione due dizionari:

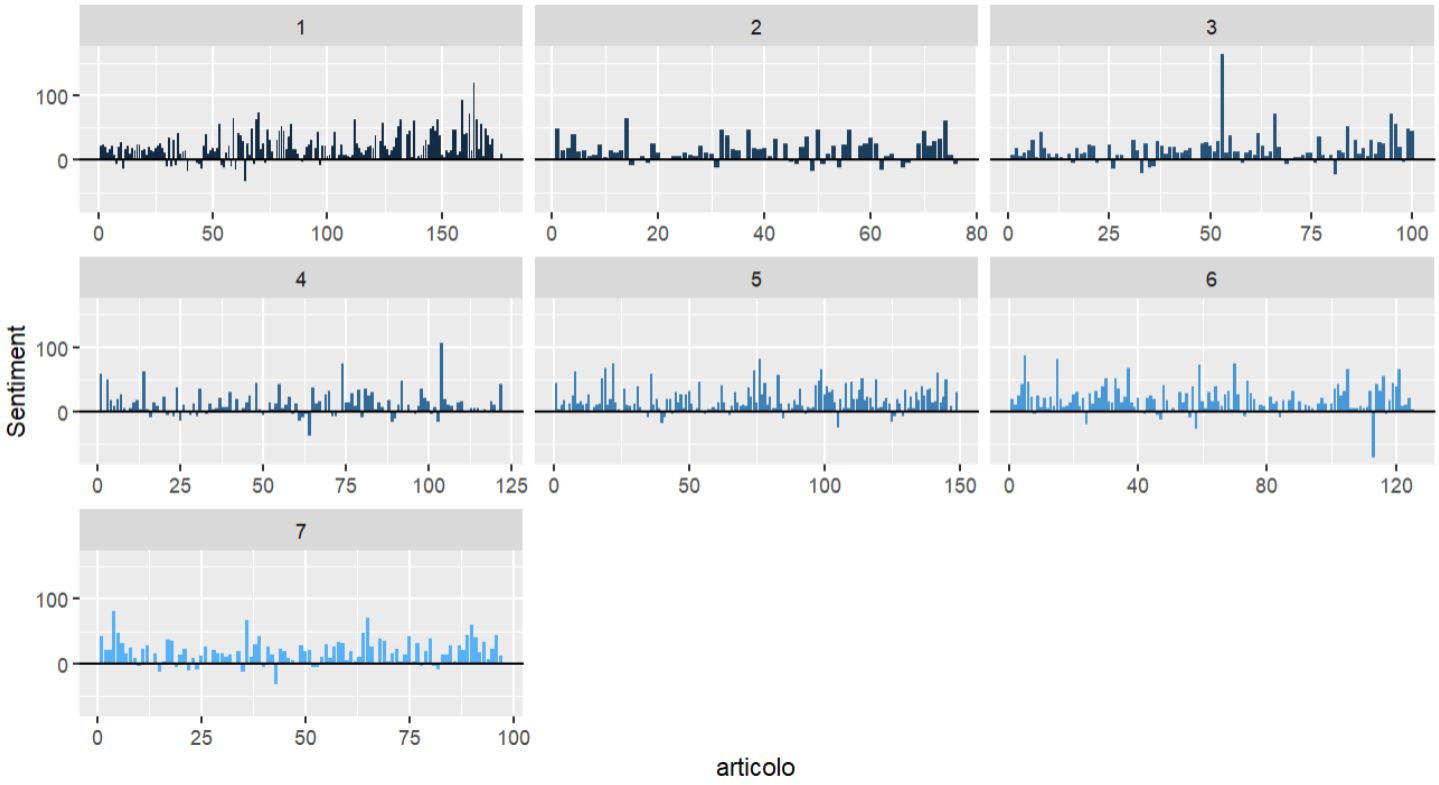
- NRC, contenente insiemi di parole riferiti a 10 sentiment;
- AFINN, contenente parole alle quali è stata associata l'intensità del sentiment espresso.

Nel primo caso il sentiment di ogni articolo è stato determinato come prevalenza delle parole positive o negative. Pertanto, dato un generico articolo, è possibile dire che esso ha un sentiment positivo quando il numero di parole positive prevale su quello delle negative.

SENTIMENT = NUMERO DI PAROLE POSITIVE – NUMERO DI PAROLE NEGATIVE

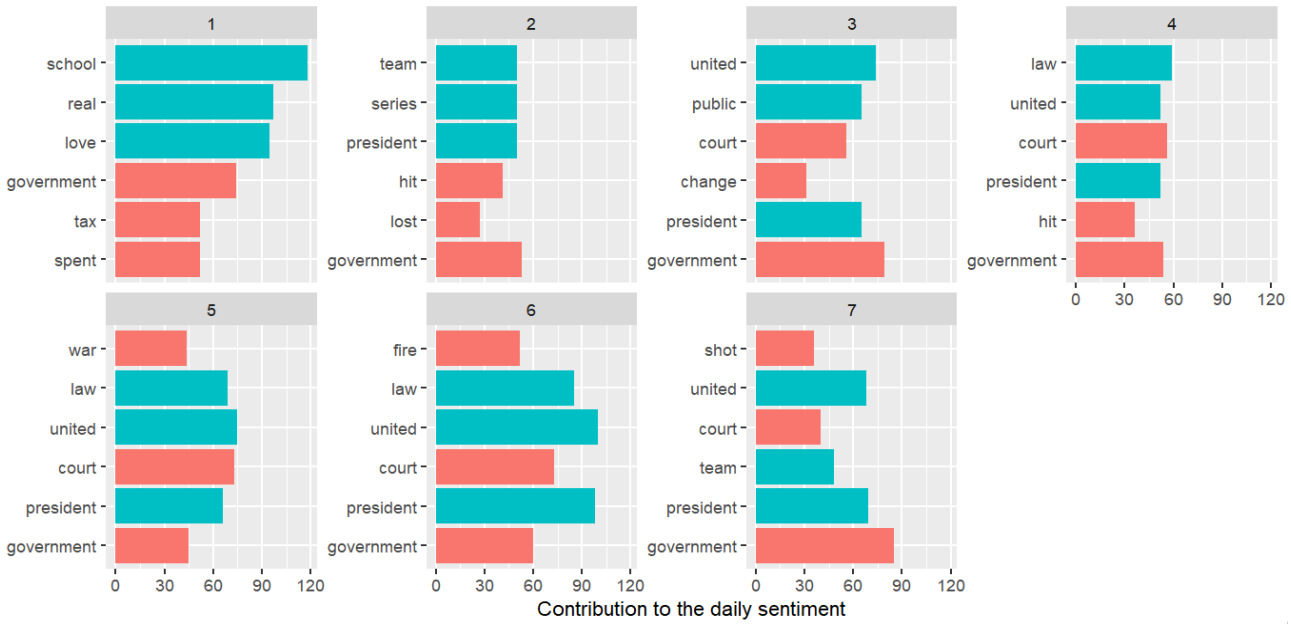
Sono state considerate **positive** tutte quelle parole associate ad un sentiment positivo, di gioia o fiducia mentre sono state considerate **negative** tutte quelle parole associate a sentiment negativi, di rabbia, tristezza, paura o disgusto.
Il GRAFICO N. 2.1 mostra l'andamento giornaliero del sentiment di ciascun articolo: valori maggiori di zero indicano che l'articolo di riferimento ha un sentiment positivo mentre valori minori di zero indicano l'opposto.
Osservando questo grafico è possibile notare che la quasi totalità degli articoli presenta un sentiment positivo in quanto la maggior parte dei valori del sentiment sono maggiori di zero.
Facilmente individuabili sono dei picchi dovuti ad articoli particolarmente positivi o negativi, come avviene ad esempio nei giorni tre e sei.

GRAFICO N. 2.1: andamento giornaliero del sentiment nei vari articoli (vocabolario NRC)



Il contributo di ciascuna parola alla determinazione del sentiment può essere differente ed è strettamente connesso al numero di volte in cui essa viene utilizzata.
Il GRAFICO N. 2.2 permette di visualizzare le parole che hanno maggiormente contribuito a determinare il sentiment negli articoli pubblicati nei sette giorni considerati, in quanto maggiormente utilizzate.

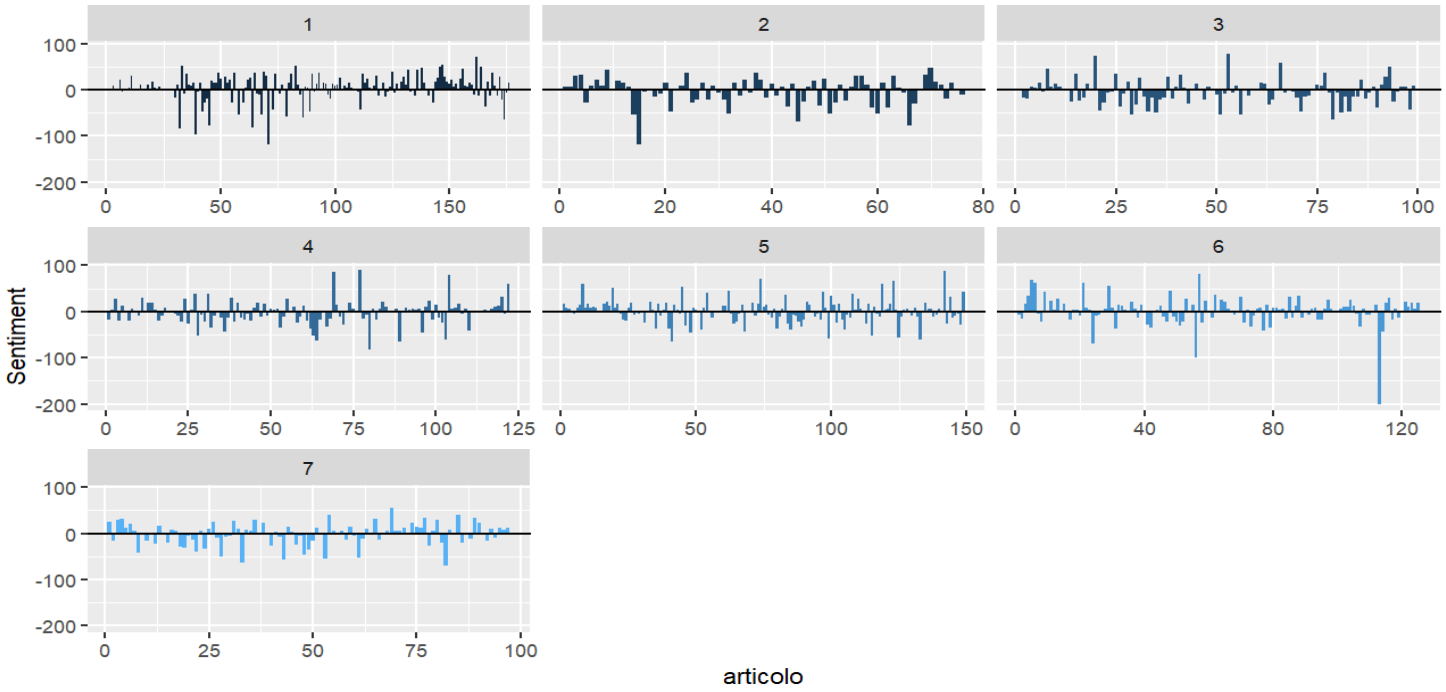
GRAFICO N. 2.2: parole che maggiormente contribuiscono al sentiment degli articoli nei 7 giorni (NRC)



Utilizzando il vocabolario AFINN, il sentiment di ciascun articolo è stato determinato considerando l'intensità dei sentiment espressi da ciascuna delle parole in esso contenute.

A differenza dei risultati ottenuti utilizzando il vocabolario NRC è possibile notare che, in questo caso, in ciascuno dei giorni di pubblicazione, emerge una maggiore presenza di articoli dal sentiment complessivo negativo (GRAFICO N. 2.3). Il picco negativo precedentemente individuato in corrispondenza del sesto giorno conferma e rafforza la propria negatività.

GRAFICO N. 2.3: andamento giornaliero del sentiment nei vari articoli (vocabolario AFINN)



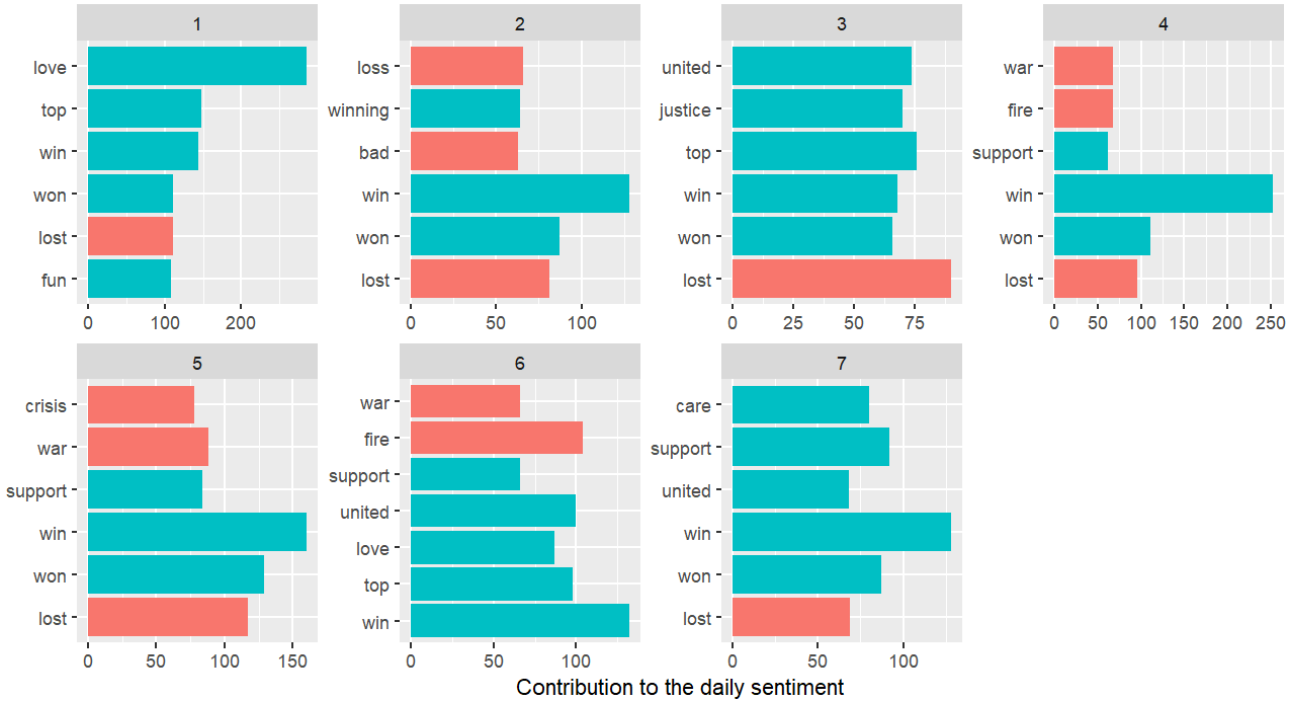
Il contributo di ciascuna parola alla determinazione del sentiment, in questo caso, è strettamente connesso non soltanto al numero di volte in cui essa viene utilizzata, ma anche all'intensità del sentiment che esprime: è necessario, pertanto, ponderare la frequenza per l'intensità.

CONTRIBUTO AL SENTIMENT = FREQUENZA * INTENSITA' SENTIMENT ESPRESSO

Il GRAFICO N. 2.4 permette di visualizzare le parole con contributo maggiore al sentiment complessivo degli articoli pubblicati in ciascuno dei sette giorni considerati.
E' possibile notare che alcune parole compaiono sempre o quasi:

- "win", "won", per quanto riguarda i contributi positivi;
- "lost", per quanto riguarda i contributi negativi.

GRAFICO N. 2.4: parole che maggiormente contribuiscono al sentiment degli articoli nei 7 giorni (AFINN)



3. Analisi del topic model

Il topic model è uno strumento statistico che, data una raccolta di documenti, permette di individuare gli argomenti principalmente trattati in essa, nonché l'argomento trattato da ciascun documento.

Il GRAFICO N. 3.1 permette di visualizzare le 20 parole maggiormente assimilabili ad un certo argomento, ossia le parole con maggiore probabilità (beta) di essere state generate da quell'argomento.
Dopo aver scelto il topic model ottimo ed averne analizzato i risultati, è possibile affermare con sufficiente fiducia che negli articoli pubblicati dal quotidiano New York times nella prima settimana di maggio siano stati principalmente affrontati i seguenti quattro argomenti:

- POLITICA/ELEZIONI ("Trump", "Clinton", "government", "president", "party", "campaign"...);
- VITA QUOTIDIANA ("art", "city", "house", "design", "home", "music", "museum", "family"...);
- SPORT ("game", "season", "team", "games", "league", "play", "players", "yankees", "won"...);
- ECONOMIA ("company", "business", "companies", "market", "executive", "chief", "quarter"...).

GRAFICO N. 3.1: