PROGETTO TEXT MINING 2020/2021

Massaro Leonardo, 27000/410

l.massaro4@lumsastud.it

Prodotto: Playstation 5
Dataset: 701 recensioni
E-commerce: Amazon

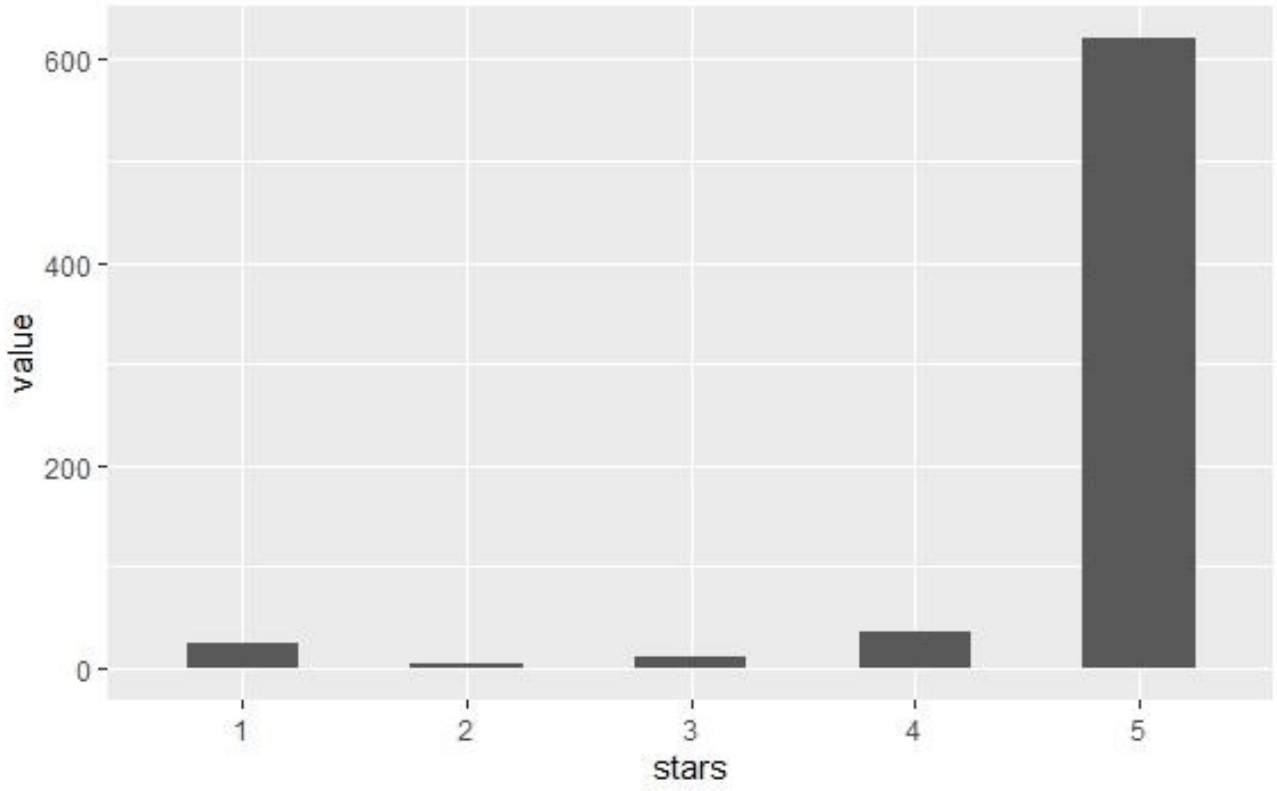
Attraverso Rstudio, linguaggio di programmazione per il calcolo statistico e la grafica, ho cercato di analizzare e di individuare il sentiment delle persone che lo hanno recensito, attraverso l'analisi delle parole usate, dalla metodologia di recensione delle stelle e individuando la percentuale di recensione negative e positive.

ANALISI RECENSIONE PER STELLE

Come primo passo ho analizzato la metodologia di recensione a stelle.

Per calcolare le valutazioni con stelle di un prodotto, Amazon non utilizza semplici medie, ma si avvale di modelli di apprendimento automatico.

Questi modelli prendono in considerazione molti fattori, come ad esempio il tempo trascorso dalla pubblicazione della valutazione o della revisione e lo stato di acquisto verificato. I modelli utilizzano vari criteri che stabiliscono l'autenticità del commento. Il sistema apprende e migliora costantemente con il tempo.



Attraverso ggplot2 ho inserito i feedback ottenuti all'interno di un grafico a barre.

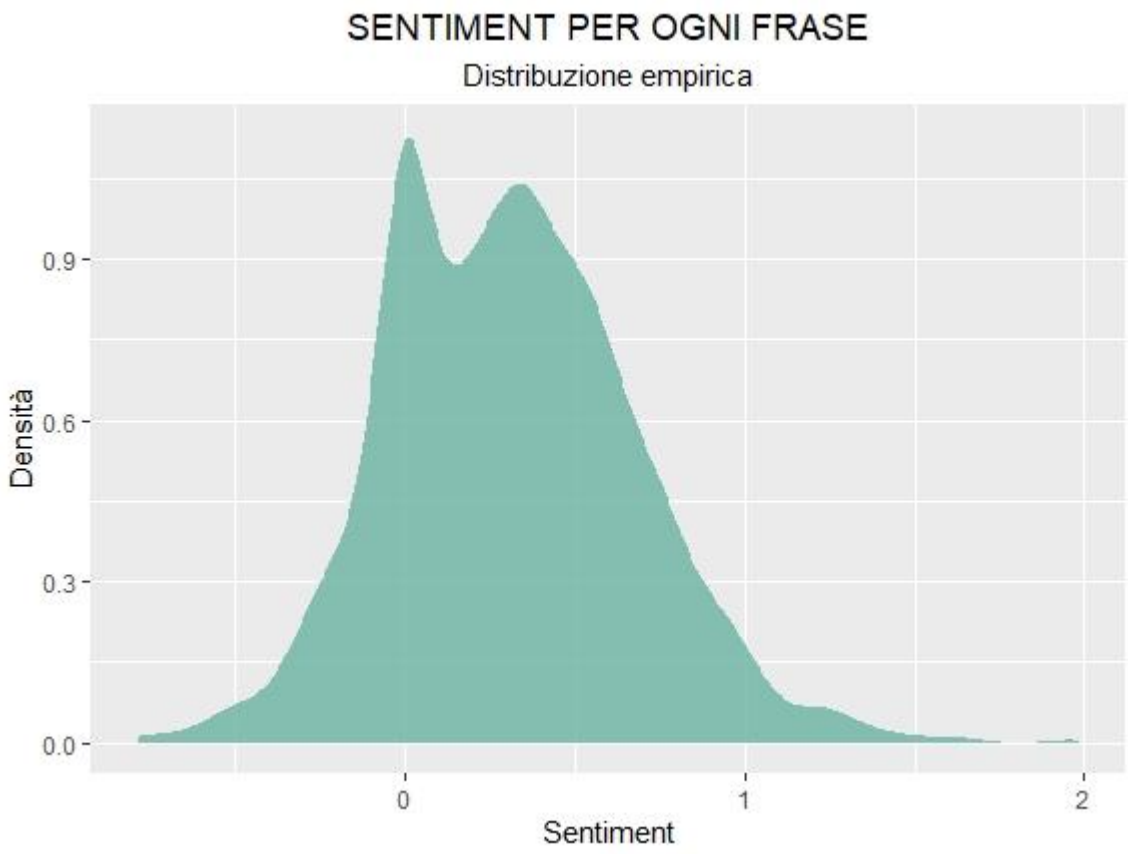
Come si può evincere si evidenzia che la gran parte delle persone hanno espresso un giudizio positivo assegnando 5 stelle su 5 ovvero testimonia un parere positivo sul prodotto recensito.

SENTIMENT ANALYSIS

L'analisi del sentiment o sentiment analysis è un campo dell'elaborazione del linguaggio naturale che si occupa di costruire sistemi per l'identificazione ed estrazione di opinioni dal testo. Si basa sui principali metodi di linguistica computazionale e di analisi testuale. La Sentiment Analysis può essere definita anche come un'attività concentrata ad analizzare ed ascoltare il web, con l'obiettivo di capire quello che gli utenti dicono e pensano del proprio brand.

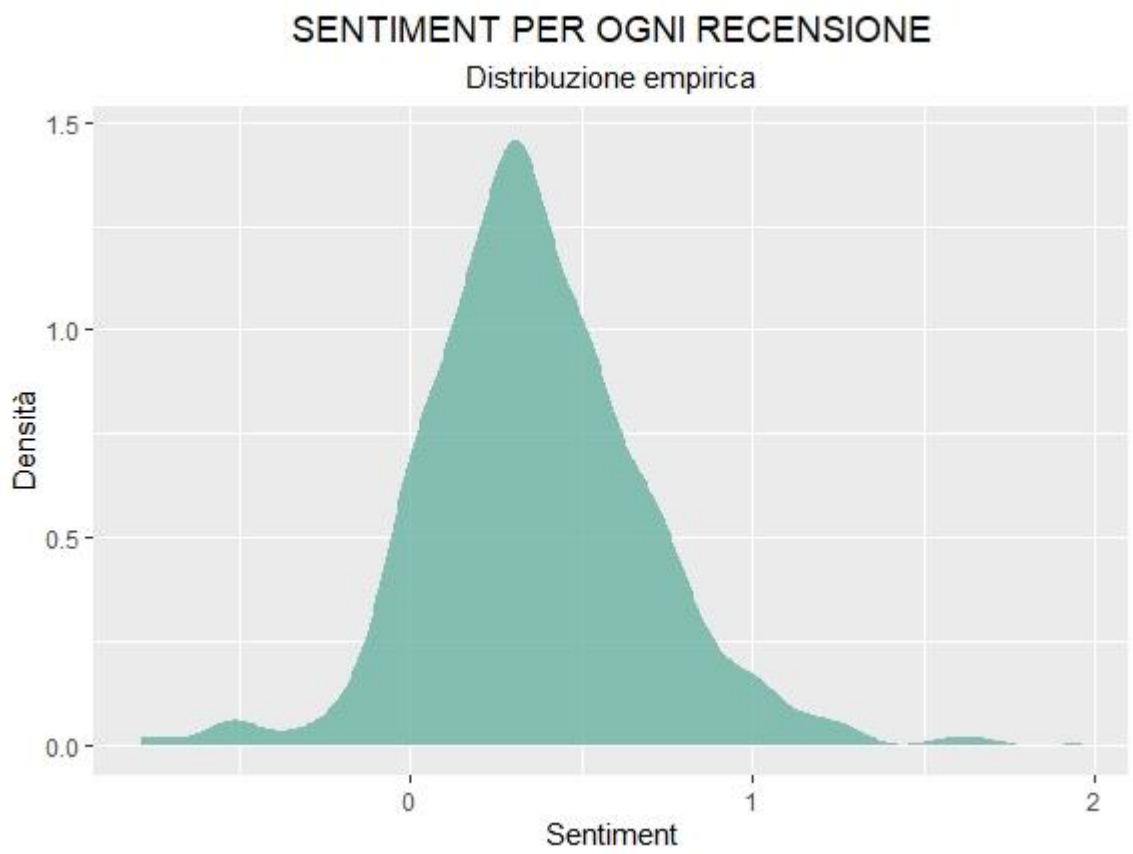
ANALISI SENTIMENT DELLE RECENSIONI

Attraverso la libreria sentimentr, che ci permette di calcolare il sentimento di polarità del testo a livello di frase e facoltativamente aggregato per righe o raggruppamento di variabili ho potuto calcolare il sentiment per ogni frase e il sentiment per ogni recensione ottenendo risultati diversi da quelli che mi aspettavo dopo aver analizzato le recensioni per stelle.



Questo grafico evidenzia la densità delle nostre recensioni per frase, se la densità massima è < -1 allora avremmo recensione negative al contrario > 1 avremmo recensioni positive.

Come possiamo vedere il picco di densità si trova tra lo 0 e lo 0.3 circa facendo intuire che le nostre recensioni siano neutre.

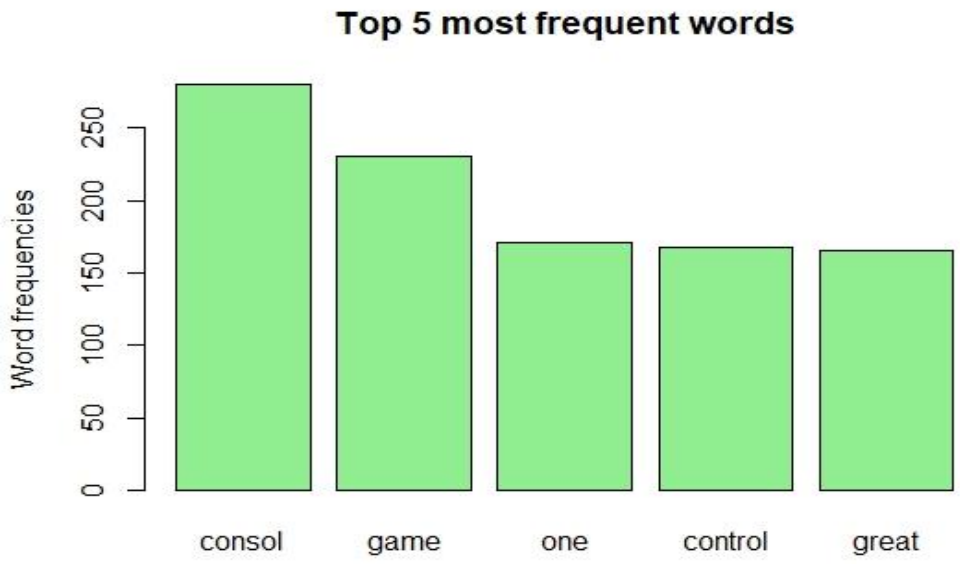


Il sentiment per ogni recensione ha un andamento neutro compreso tra lo 0 e 0.5. Come si evidenzia dai due grafici" sentiment per frase, sentiment per recensione" si riscontra un andamento neutro in entrambi i casi totalmente differente dall'analisi del grafico per stelle che evidenziava un picco positivo sulle 5 stelle, creando una discrepanza tra le due metodologie di recensione.

ANALISI FREQUENZA DELLE PAROLE

Come ulteriore analisi statistica ho analizzato le parole più frequenti all'interno delle recensioni e ho graficato i risultati grazie a ggplot2 e a WordCloud.

Per evitare problemi all'inizio dell'analisi del testo ho effettuato, tramite la libreria tm, una pulizia delle parole superflue, numeri, punteggiatura, spazi bianchi e carattere del testo ad esempio lettere maiuscole.



Come possiamo vedere dal grafico le parole più frequenti sono quasi tutte legate all'ambito del gaming, essendo il nostro prodotto l'ultima console uscita in commercio.

Grazie alla wordcloud possiamo visualizzare molte più parole utilizzate dalle persone per recensire il prodotto. Le parole più grandi corrispondono alle parole più frequenti:

- Console
- Game
- One
- Control
- great

Un altro modo di pensare alle relazioni tra le parole è con la funzione findAssocs() nel pacchetto tm. Per ogni parola, findAssocs() calcola la sua correlazione con ogni altra parola in un TDM o DTM. I punteggi vanno da 0 a 1. Un punteggio =1 identifica che due parole appaiono sempre insieme nei documenti, mentre un punteggio che si avvicina a 0 significa che i termini compaiono raramente nello stesso documento.

```
$consol
numeric(0)

$game
 play  load allow given  life  want
 0.42  0.32  0.27  0.27  0.26  0.25

$one
  get manag lucki
 0.46  0.29  0.26

$control
    feel   haptic    might  trigger download feedback      new
    0.38     0.33     0.28     0.28     0.27     0.26     0.25

$great
numeric(0)
```

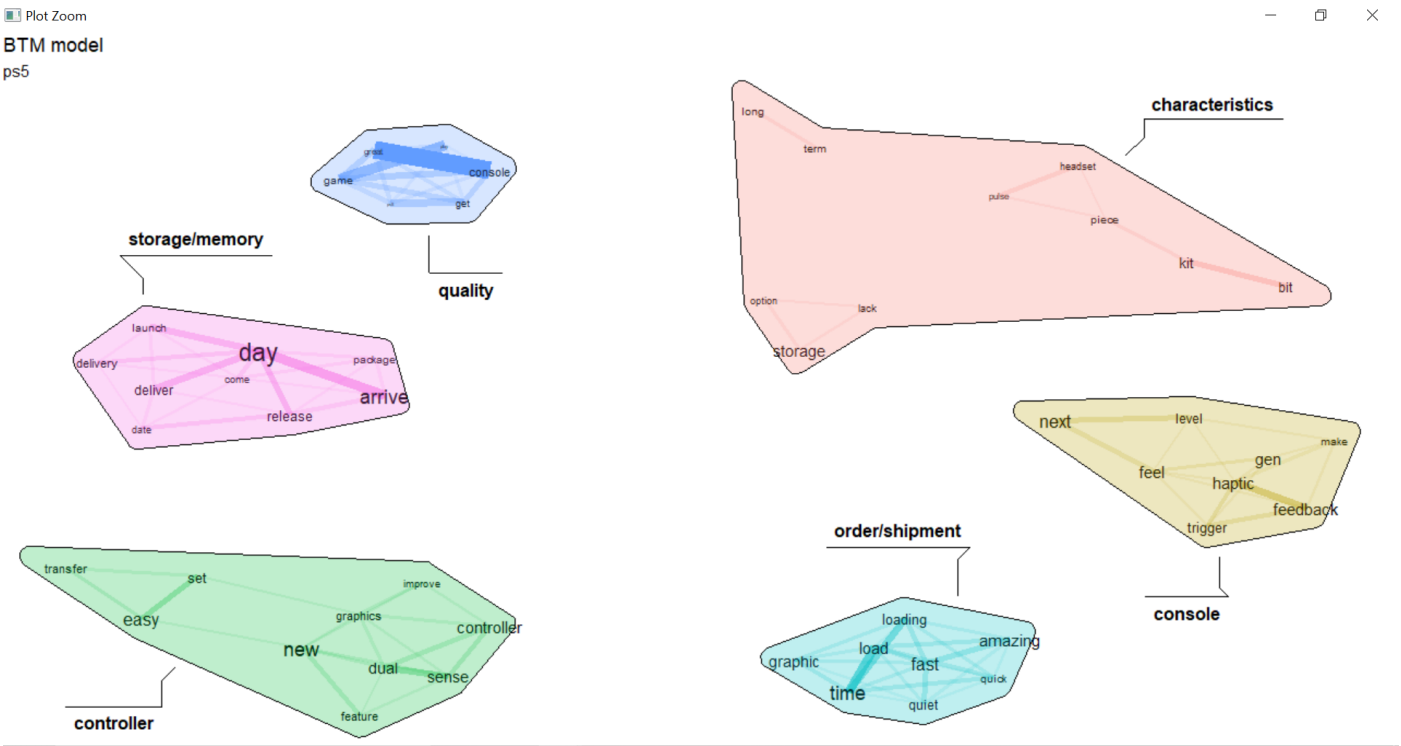
TOPIC MODELLING E BTM

Una volta analizzate le parole ho cercato di individuare il Topic modelling delle nostre recensioni.

Il Topic modelling è un tipo di modello statistico per scoprire gli "argomenti" (topic) astratti che si verificano in una raccolta di documenti, per individuare questi topic ho utilizzato la libreria BTM che ci permette di creare questi gruppi di parole con lo stesso campo semantico.

Prima di utilizzare la libreria BTM bisogna creare il dataset da inserire in input per la creazione dei gruppi.

Utilizzando la libreria Udpipe ho effettuato il Pos tagging di tutte le parole, cioè ho targhettato ogni parola con il suo specifico valore grammaticale (Nomi, Aggettivi…), così da poterlo inserire in input nella BTM evitando problemi nell'analisi del topic.



I Biterm Topic Models trovano argomenti in raccolte di brevi testi.

È un modello di argomento basato sulla co-occorrenza di parole che apprende gli argomenti modellando modelli di cooccorrenze parola-parola che sono chiamati biter.

Un biterm è costituito da due parole che si verificano insieme nella stessa finestra di testo breve.

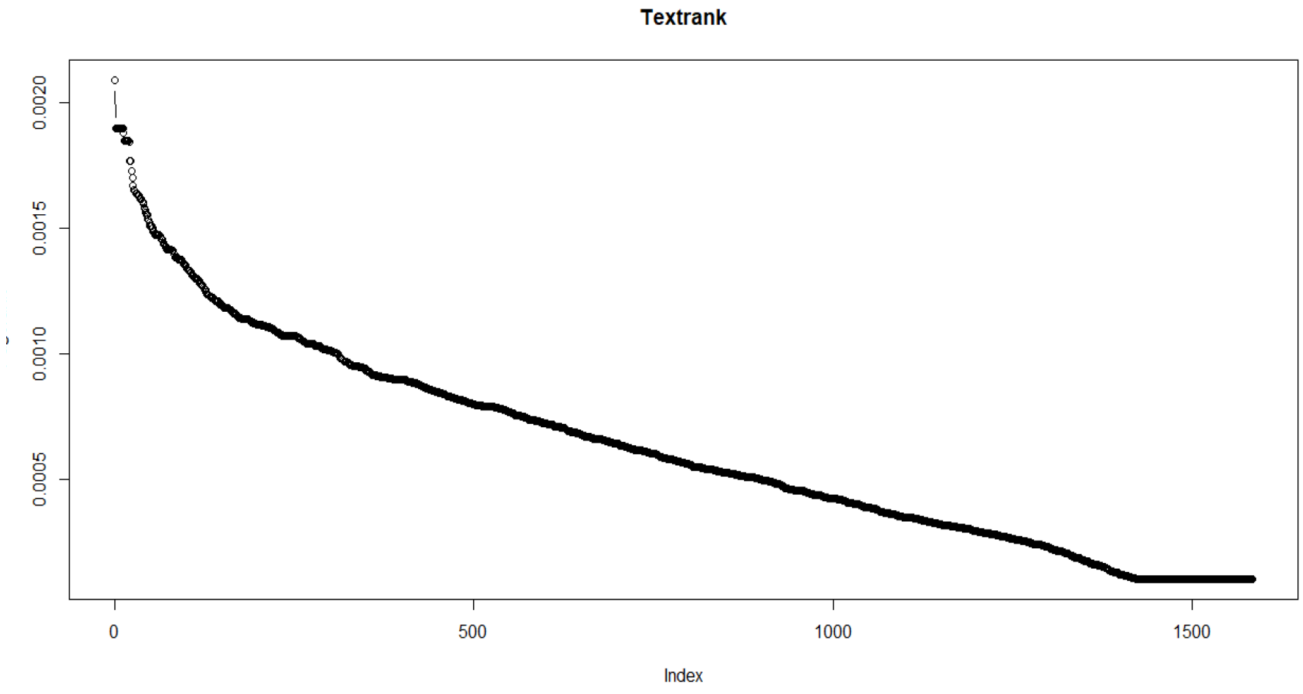
Questa finestra di contesto può essere ad esempio un messaggio di Twitter, una breve risposta a un sondaggio, una frase di un testo o un identificatore di documento.

Come possiamo vedere dal grafico abbiamo individuato 6 diversi gruppi di parole, tutte rientranti nel gruppo semantico del gaming, come ad esempio il gruppo dedicato al controller dove viene evidenziato il suo nome (dual sense), le feature e tutte le descrizioni effettuate dagli utenti.

SUMMARIZATION

Come ultima analisi ho effettuato la Summarization attraverso la libreria Textrank che consente di riassumere il testo calcolando come le frasi sono correlate tra loro. Questo viene fatto osservando la terminologia sovrapposta utilizzata nelle frasi per creare collegamenti tra le frasi. La rete di frasi risultante viene quindi collegata all'algoritmo "Pagerank" che identifica le frasi più importanti nel testo e le classifica.

In modo simile, 'textrank' può essere utilizzato anche per estrarre le parole chiave. Una rete di parole viene costruita cercando se le parole si susseguono. In cima a quella rete viene applicato l'algoritmo 'Pagerank' per estrarre le parole rilevanti dopo di che le parole rilevanti che si susseguono vengono combinate per ottenere le parole chiave.



Prima di effettuare la summarization bisogna riprendere il dataset utilizzato per la BTM avente il pos tagging, una volta ottenuto possiamo cominciare ad individuare le frasi e le parole più rilevanti all’interno del nostro testo.

Dopo aver completato l’analisi riusciamo a capire quali frasi dobbiamo analizzare, perché più significative, e quali invece sono inutili per il nostro scopo finale.

CONCLUSIONI

Grazie alle analisi effettuate possiamo constatare che il sentiment delle recensioni tende ad essere positivo, anche avendo qualche discrepanza tra i vari metodi di giudizio.