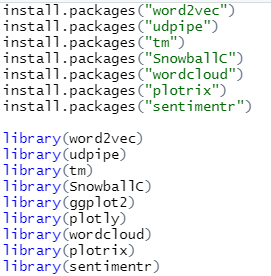
RELAZIONE PROGETTO “TEXT MINING”

Per la realizzazione del progetto ho usato il software “RStudio” e due estensioni di Google Chrome: Data Miner e Choicy. Ho utilizzato Data Miner per fare lo scraping delle recensioni del prodotto da me scelto, mentre Choicty l’ho utilizzato per cercare i prodotti con un maggior numero di recensioni.

Dopo aver importato le recensioni su RStudio, ho incominciato a installare i pacchetti e le librerie necessarie per analizzare le recensioni.

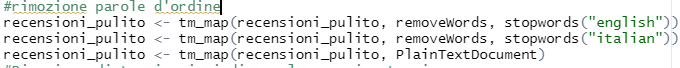


* “word2vec” e “wordcloud” servono per la rappresentazione della wordcloud.
* “udpipe” serve per la tokenizzazione indipendente del linguaggio
* “tm” è uno dei pacchetti più importanti dell’analisi testuale e permette di effettuare una serie di azioni preprocessing sui testi, come lo stemming.
* “SnowballC” serve quando si va a fare lo stemming di un testo.
* “ggplot2” è un pacchetto per la produzione di grafici a partire da un set di dati.
* “plotly” è un pacchetto per la produzione di grafici interattivi.
* “plotrix” è un pacchetto per la produzione di grafici, nello specifico grafici in 3D.
* “sentimentr” è un pacchetto che serve per l’analisi del sentiment.

Come prima cosa ho rinominato le “recensione\_pentola\_a\_pressione\_text\_mining” in “recensioni”.

Successivamente ho trasformato il carattere della colonna “Review Content” (recensioni) da maiuscolo a tutto minuscolo con la funzione “tolower”. Avendo le recensioni in minuscolo trasformato il dataset in corpus così da incominciare la pulizia del testo con il metodo stemming.

Ho rimosso la punteggiatura per avere un’analisi migliore, con la funzione tm\_map così da lasciare solo parole nelle recensioni. Successivamente ho rimosso le parole d’ordine come preposizione, congiunzioni ecc… sempre con la funzione tm\_map



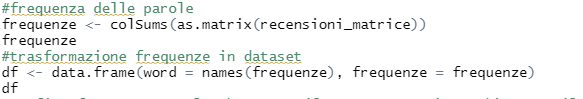
Ho rimosso sia quelle inglesi che quelle in italiano poiché le mie recensioni erano in entrambi le lingue.

Per pulire maggiormente il testo ho provato a levare i numeri, ma avendo preso le recensioni di un elettrodomestico, molte persone indicavano a numeri la vita di utilizzo dell’oggetto e quindi si andava a perdere il senso logico della recensione.

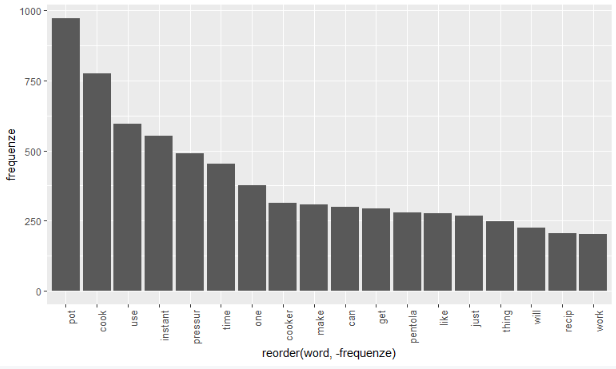
Dopo sono passato all’eliminazione degli spazi bianchi in eccesso

Per trasformare il dataset in matrice, ho prima ritrasformato il corpus in dataset e poi in matrice.

Ho trasformato il dataset in matrice per fare la frequenza delle parole e poi ritrasformato in dataset per fare il grafico.



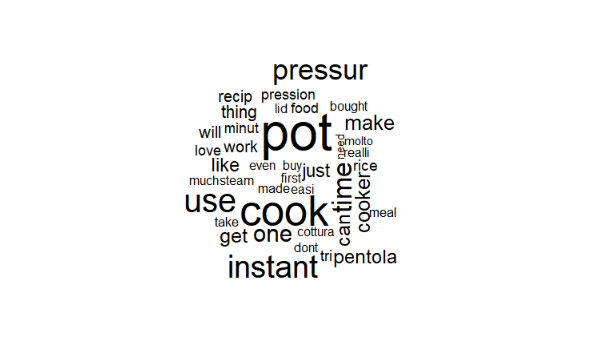
Successivamente ho fatto il grafico della frequenza delle parole.



Ho fatto questo grafico per capire la frequenza delle parole, essendo un aspetto molto importante dell’analisi delle recensioni.



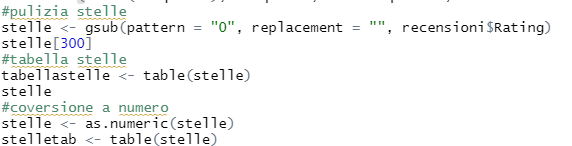
Per completare l’analisi delle frequenze delle recensioni ho fatto un Wordcloud. Il Wordcloud è una rappresentazione a “nuvola” delle frequenze usando il carattere delle parole in modo diverso per indicare la frequenza.



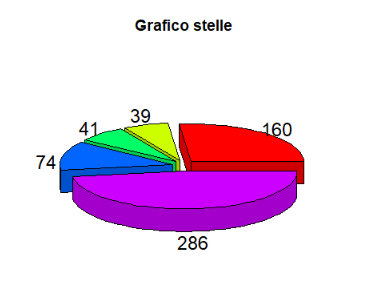
Finita l’analisi delle recensioni mi sono concentrato sull’analisi del “Rating” cioè della valutazione (stelle).

Come prima cosa quando ho trasformato la colonna delle stesse in numero con la funzione as.numeric.

Successivamente ho fatto una tabella di esse.



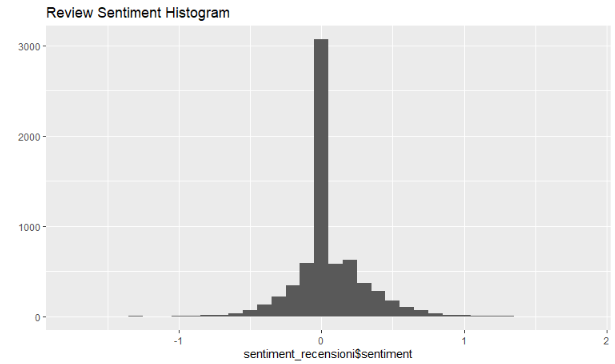
Per capire meglio la distruzione delle stelle da 1 a 5, ho creato un diagramma a torta.



Da questo diagramma possiamo notare che il prodotto è molto buono avendo una prevalenza di 286 stelle 5/5.

La sentiment analisys fondamentalmente serve a capire il grado di gradimento del prodotto da parte degli utenti.

Come Ultima cosa sono andato a fare l’analisi del sentiment, utilizzando un grafico.



Possiamo vedere dal grafico che l’analisi è molto positiva, confermando così il grafico delle stelle.