Analisi degli Evidenziatori su amazon

Damiano Caputo | matricola: 27469/410| Tecniche Informatiche per la gestione dei dati

Contenuto del file zip:

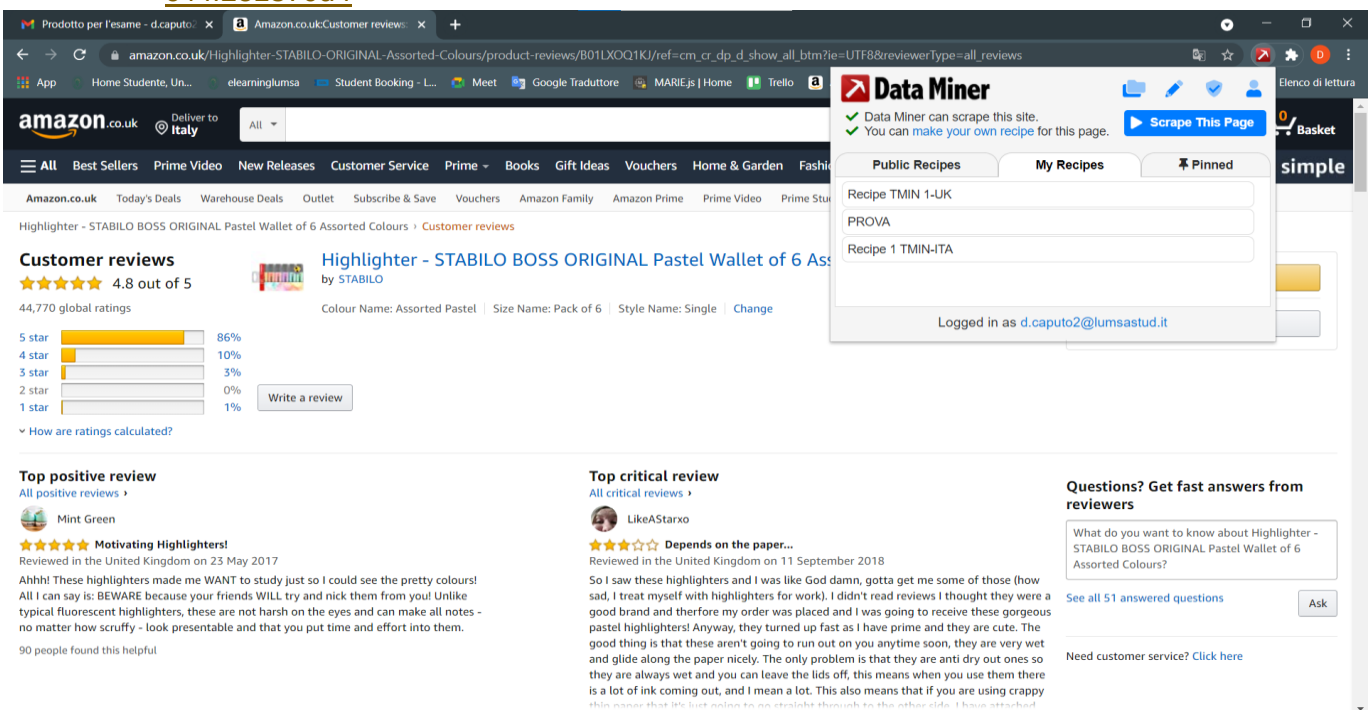
1. Pdf della relazione;
2. Script di R, per la corretta visualizzazione del codice e per la replica dei risultati ottenuti;
3. Il dataset in formato csv: EVIDENZIATORI.csv

Ambiente di Lavoro:

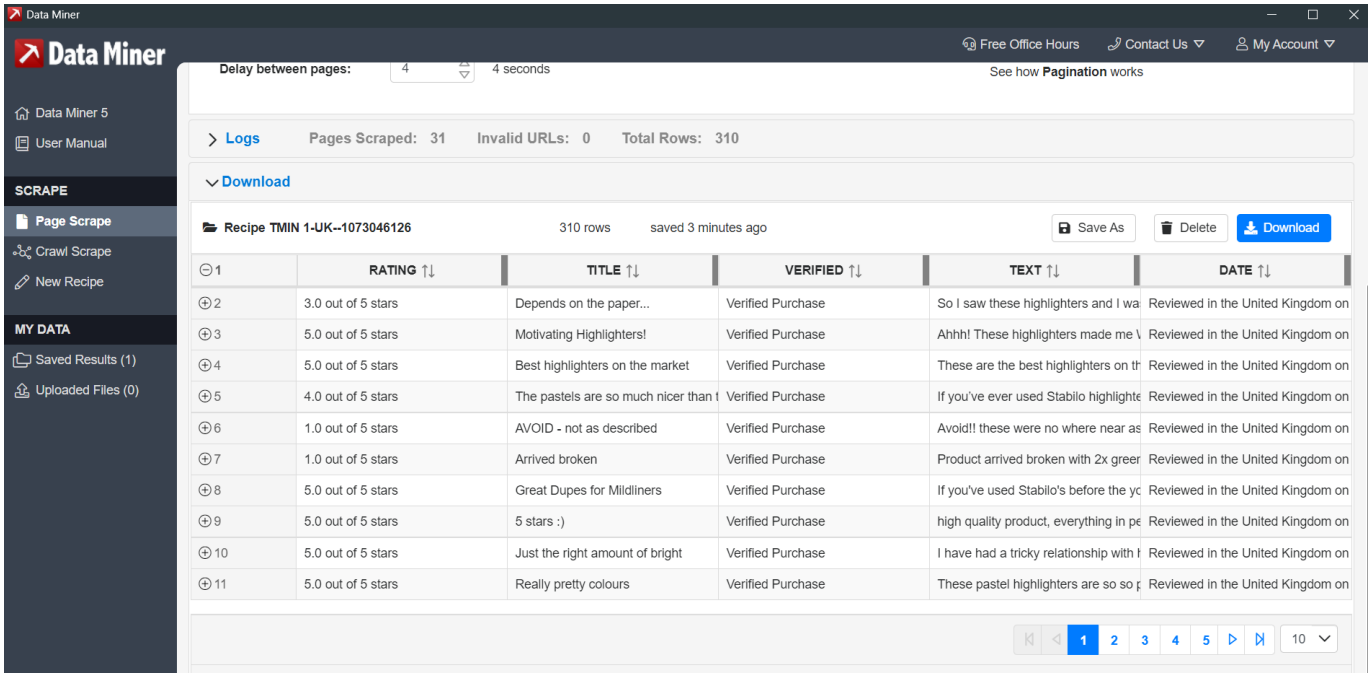
- Windows 10
- Rstudio: versione 1.4.1103

Estrazione dei Dati:

– Attraverso l'uso dell'estensione dataminer, compatibile con google chrome, ho effettuato lo scraping web sulla pagina degli evidenziatori stablio di amazon del regno unito: https://www.amazon.co.uk/Highlighter-STABILO-ORIGINAL-Assorted-Colours/dp/B01LXOQ1KJ/ref=sxin_9_ac_d_rm?ac_md=0-0-aGlnaGxpZ2h0ZXJz-ac_d_rm&crid=Q89DRMIABL2L&cv_ct_cx=highlighters&dchild=1&keywords=highlighters&pd_rd_i=B01LXOQ1KJ&pd_rd_r=4b43eebc-1b58-4860-8c5f-fab59afd9145&pd_rd_w=7LK6k&pd_rd_wg=GLSPP&pf_rd_p=73573abc-9548-43f0-87cb-a185286cee4c&pf_rd_r=WFGRQ2TZQ0PTMBN8HEJH&psc=1&qid=1617887411&sprefix=high%2Caps%2C200&sr=1-1-fe323411-17bb-433b-b2f8-c44f2e1370d4



Da questa prima schermata ho creato una ricetta che mi permettesse di estrarre i dati nel modo corretto e che possano risultare più utili. Una volta scelta la ricetta ho eseguito lo scraping di 50 pagine di recensioni, ottenendo in questo modo un dataset di 500 righe e di 5 colonne:



	RATING	TITLE	VERIFIED	TEXT	DATE
2	3.0 out of 5 stars	Depends on the paper...	Verified Purchase	So I saw these highlighters and I wa	Reviewed in the United Kingdom on
3	5.0 out of 5 stars	Motivating Highlighters!	Verified Purchase	Ahhh! These highlighters made me \	Reviewed in the United Kingdom on
4	5.0 out of 5 stars	Best highlighters on the market	Verified Purchase	These are the best highlighters on th	Reviewed in the United Kingdom on
5	4.0 out of 5 stars	The pastels are so much nicer than t	Verified Purchase	If you've ever used Stabilo highlighte	Reviewed in the United Kingdom on
6	1.0 out of 5 stars	AVOID - not as described	Verified Purchase	Avoid!! these were no where near as	Reviewed in the United Kingdom on
7	1.0 out of 5 stars	Arrived broken	Verified Purchase	Product arrived broken with 2x greer	Reviewed in the United Kingdom on
8	5.0 out of 5 stars	Great Dupes for Mildliners	Verified Purchase	If you've used Stabilo's before the yc	Reviewed in the United Kingdom on
9	5.0 out of 5 stars	5 stars :)	Verified Purchase	high quality product, everything in pe	Reviewed in the United Kingdom on
10	5.0 out of 5 stars	Just the right amount of bright	Verified Purchase	I have had a tricky relationship with h	Reviewed in the United Kingdom on
11	5.0 out of 5 stars	Really pretty colours	Verified Purchase	These pastel highlighters are so so p	Reviewed in the United Kingdom on

come si può vedere le colonne estratte e che comporranno il dataset sono: il rating delle stelle, il titolo della recensione, il verificato, il contenuto in forma testuale ed infine la data in cui è stata scritta la recensione. Di conseguenza il dataset utilizzato è composto da tutti caratteri testutuali, quindi andando avanti con le analisi modificheremo il contenuto in maniera tale da poter lavorare con più facilità.

Librerie Utilizzate:

Le librerie utilizzate per condurre le analisi sono:

1. **readr**: per importare il dataset in Rstudio
2. **ggplot2**: per effettuare i grafici delle analisi condotte
3. **textclean**: per la funzione mgsub(), utilizzata per pulire in punti specifici delle colonne del dataset
4. **tm e ggthemes**: utilizzate entrambe per eseguire la pulizia totale della colonna delle recensioni (TEXT)
5. **wordcloud2**: per realizzare la word cloud
6. **sentimentr e tidyverse**: utilizzate per l'analisi del sentiment delle recensioni.
7. **topicmodels e tidytext**: per conseguire il topic modelling

Analisi preliminare sul Dataset:

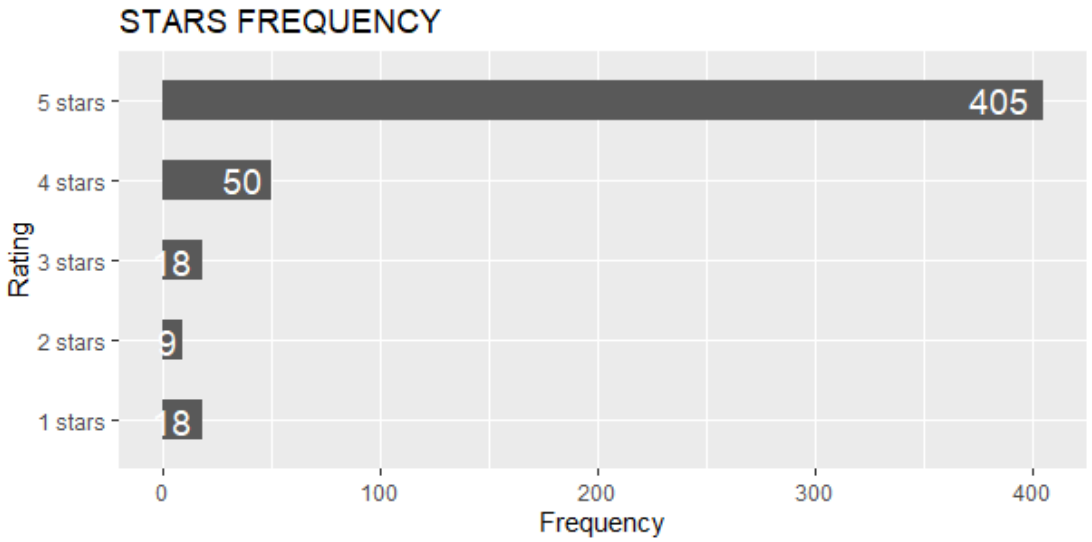
Una volta importato il dataset, sopra spiegato, ho notato che le colonne del rating e delle date presentavano, nel contenuto delle righe, delle parole che non interessano nelle analisi condotte. Per la colonna delle stelle ho eliminato la parte ".0 out of" riuscendo quindi ad ottenere il numero e la parola "stars"; mentre per le date ho eliminato "Reviewed in the United Kingdom on" e di conseguenza ho cercato di ottenere come prima cosa un formato classico per la data ovvero: gg/mm/aaaa, per poi passare al formato di R: aaaa-mm-gg, infine impostando il formato solo annuale sono rimasto con solo l'anno in cui è stata scritta la recensione permettendomi di poter conseguire un analisi annuale. Come ultima valutazione sui dati posso affermare che la colonna del verificato è inutile, perché gli acquisti sono stati effettuati da persone con l'account verificato.

Analisi della Frequenza delle Stelle:

Sono partito attuando un'analisi preliminare sulla colonna del rating delle stelle, per vedere, attraverso la frequenza di esse, com'è l'opinione della gente su questo prodotto.

Una volta costruita la tabella (con table), la trasformo di conseguenza in un subset contenente 2 colonne: in una la classificazione del rating delle stelle, nell'altra quante volte appaiono nel dataset e con una leggera pulizia per renderle più leggibili nel grafico.

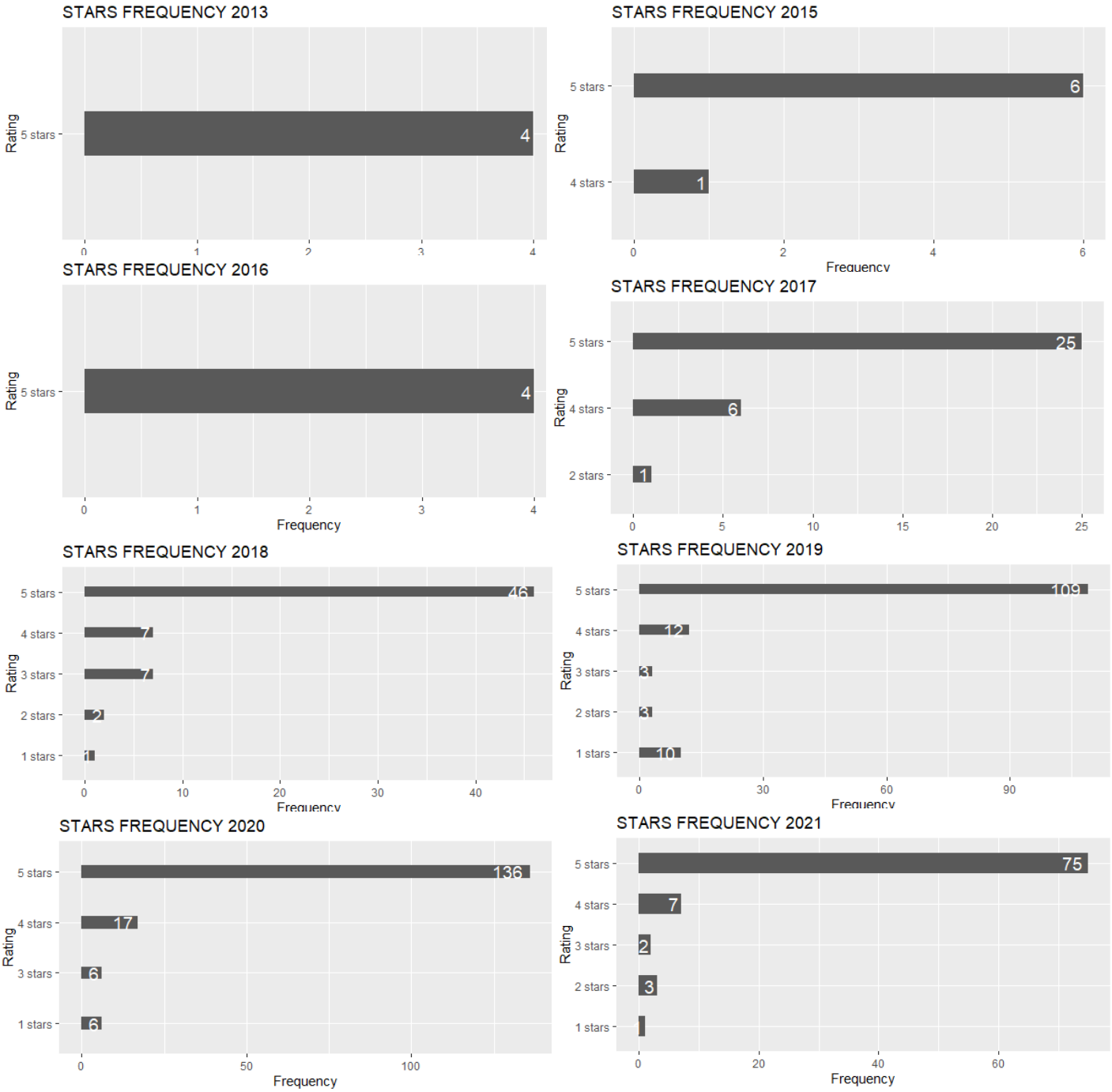
	Rating	Frequency
1	1 stars	18
2	2 stars	9
3	3 stars	18
4	4 stars	50
5	5 stars	405



Attraverso la libreria ggplot2 ho realizzato il grafico, come possiamo notare dal risultato ottenuto è che le persone hanno scelto maggiormente le 5 stelle per questo prodotto.

Come seconda analisi ho eseguito un andamento annuale sulla frequenza delle stelle, ovvero ho creato un filtro che mi permettesse di racchiudere in vari subset le

stelle di quel determinato anno. I subset per anno sono: 2013, 2015, 2016, 2017, 2018, 2019, 2020 e 2021, quello che otteniamo sono i seguenti grafici:



Da questi risultati notiamo che anche annualmente le 5 stelle sono state la scelta maggiore in tutti gli anni. Quindi posso assumere preliminarmente che il prodotto è apprezzato dagli utenti e andando avanti con le analisi confermeremo o smentiremo quanto appena detto.

Frequenza delle Parole e Pulizia delle Recensioni e dei Titoli:

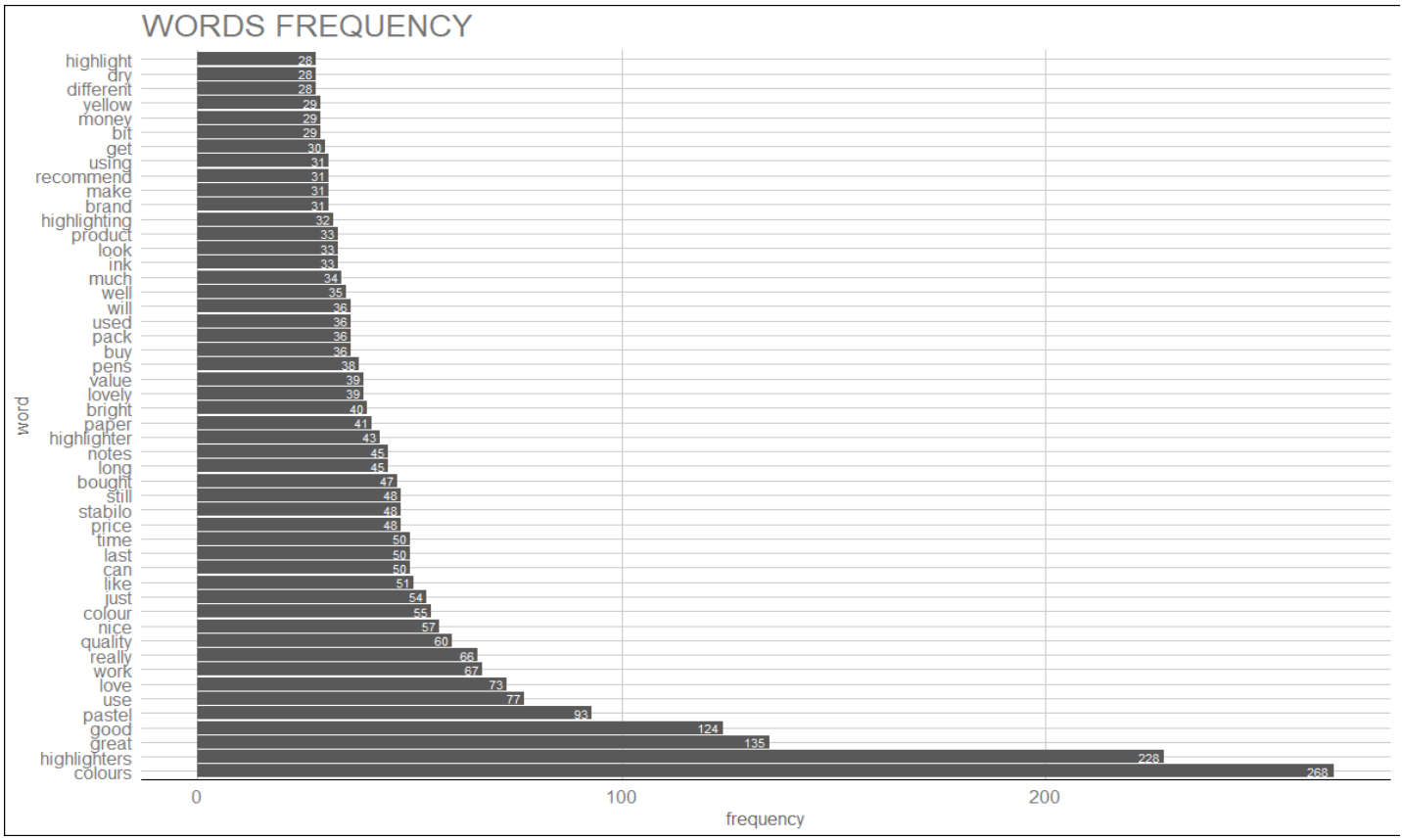
Prima di eseguire l'analisi della frequenza delle parole ho creato 2 funzioni che mi permettessero di pulire le recensioni senza errori. Le funzioni create sono: tryTolower(), che permette di rendere tutti i caratteri in minuscolo, e clean.corpus(), realizzata attraverso la libreria tm, permetterà di pulire le recensioni nel corpus togliendo: numeri, spazi in eccesso, stop words e la punteggiatura.

Una volta trasformato la colonna delle recensioni in un corpus ed utilizzato le funzioni sopra citate, ho realizzato un dataset che contenesse tutte le parole più frequenti presenti nelle recensioni in maniera tale di poter condurre un'analisi. Quello che ho ottenuto è mostrato qui a destra, la cosa che mi ha colpito di più è stata la frequenza delle parole "great" e "good" iniziando quindi a confermare una delle prime tesi esposte, ovvero quella che affermava che il prodotto è di gradimento ai clienti che lo acquistano.

	word	frequency
colours	colours	268
highlighters	highlighters	228
great	great	135
good	good	124
pastel	pastel	93
use	use	77
love	love	73
work	work	67
really	really	66
quality	quality	60
nice	nice	57
colour	colour	55
just	just	54
like	like	51
can	can	50
last	last	50
time	time	50
price	price	48

Showing 1 to 18 of 1,570 entries, 2 total columns

1. La prima analisi che ho conseguito, su questo dataset, è stata la realizzazione del grafico delle parole più frequenti e non saranno mostrate tutte, perché da come si può leggere dall'immagine sopra sarebbero troppe parole e renderebbero il grafico illeggibile. Il risultato ottenuto è mostrato di seguito:



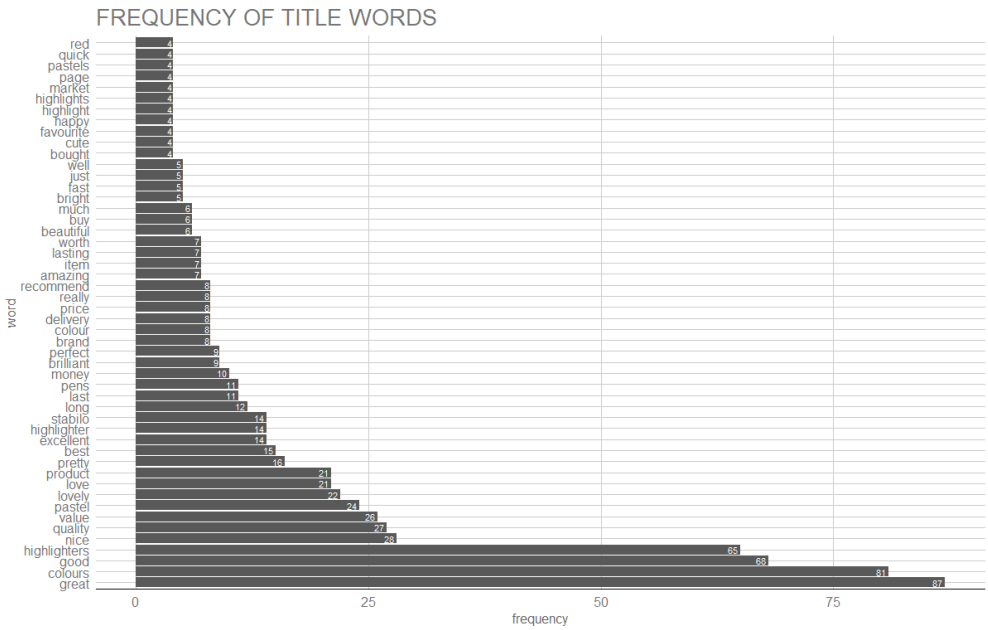
2. Attraverso la frequenza delle parole, spiegata sopra, ho creato una word cloud:

3. L'ultima analisi condotta sulla frequenza delle parole è stata eseguita sulla colonna dei titoli, per vedere quali sono le parole più frequenti. Di conseguenza ho utilizzato le stesse funzioni per la pulizia della frequenza delle parole delle recensioni, producendo un dataset mostrato qui a destra:

	word	frequency
great	great	87
colours	colours	81
good	good	68
highlighters	highlighters	65
nice	nice	28
quality	quality	27
value	value	26
pastel	pastel	24
lovely	lovely	22
love	love	21
product	product	21
pretty	pretty	16
best	best	15
excellent	excellent	14
highlighter	highlighter	14
stabilo	stabilo	14
long	long	12
last	last	11

Showing 1 to 18 of 360 entries, 2 total columns

Da questo dataset ho estratto il grafico ed il risultato ottenuto e questo:



notiamo come anche in questo una delle parole più frequenti è "good". Potendo assumere anche qui che il prodotto è apprezzato dai clienti.

Sentiment Analysis:

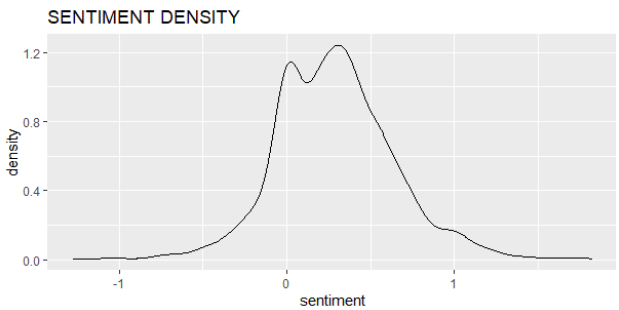
Un'altra analisi che ho condotto è stata la sentiment analysis, mi sono avvalso delle librerie sentimentr e tidyverse. Queste librerie mi consentono di attuarla senza un'ulteriore pulizia e di avere dei valori molto precisi e affidabili, impostando un range che va da -1, dove saranno classificate le recensioni negative, fino a +1 dove saranno presenti le positive. Quindi ho suddiviso questa valutazione in 4 parti:

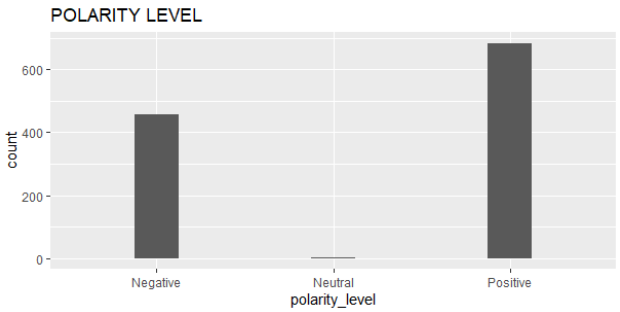
Nella prima parte sono partito realizzando un subset che contenesse tutte le recensioni, sia per lavorarci meglio sia per non andare ad intaccare il dataset originale. Successivamente ho estratto dal subset tutte le frasi facendo in modo che avessero comunque in id identificativo, per tenere traccia da quale recensione arrivi la frase, attraverso get_sentences. Con sentiment() estrapoliamo i valori di quest'ultimo e creeremo un dataset contente tutte le frasi e i valori corrispondenti tra cui: id della recensione, conteggio delle frasi, il conteggio delle parole, la deviazione standard di esse, il valore del sentiment e la polarità. Il dataset ottenuto è questo:

	Text	element_id	sentence_id	word_count	sentiment	polarity_level
1	So I saw these highlighters and I was like God damn, gotta ...	1	1	26	-0.196116135	Negative
2	I didn't read reviews I thought they were a good brand and...	1	2	27	0.240562612	Positive
3	Anyway, they turned up fast as I have prime and they are cu...	1	3	13	0.513097682	Positive
4	The good thing is that these aren't going to run out on you ...	1	4	25	0.426000000	Positive
5	The only problem is that they are anti dry out ones so they ...	1	5	42	-0.216024690	Negative
6	This also means that if you are using crappy thin paper that ...	1	6	23	-0.083405766	Negative
7	I have attached photos to show them on thick paper (nice a...	1	7	27	0.250185117	Positive
8	I have also attached a photo of the back of paper which sho...	1	8	20	0.000000000	Negative
9	Once it dries out it leaves my pages a bit bumpy from wher...	1	9	19	-0.114707867	Negative
10	Moral of the story, don't impulse buy cute things (I am a fe...	1	10	25	0.030000000	Negative
11	Other than that quite a decent set, but for the price, maybe ...	1	11	20	0.201246118	Positive
12	Ahhh!	2	1	1	0.000000000	Negative
13	These highlighters made me WANT to study just so I could s...	2	2	15	0.258198890	Positive
14	All I can say is: BEWARE because your friends WILL try and n...	2	3	16	-0.125000000	Negative
15	Unlike typical fluorescent highlighters, these are not harsh o...	2	4	31	0.359210604	Positive
16	These are the best highlighters on the market!	3	1	8	0.176776695	Negative
17	They have a good range of colours, which is great for me as...	3	2	23	0.417028828	Positive
18	They last absolutely ages before running out!	3	3	7	0.000000000	Negative

Showing 1 to 18 of 1,147 entries, 6 total columns

Considerando i valori del sentiment ho realizzato il grafico della densità permettendomi di capire se il prodotto è stato apprezzato o meno, dal risultato ottenuto il picco più alto determina la positività, il secondo picco sono le negative mentre il minimo sono le neutrali. Quindi posso assumere, dalle frasi singole, che il prodotto è molto apprezzato dagli utenti.





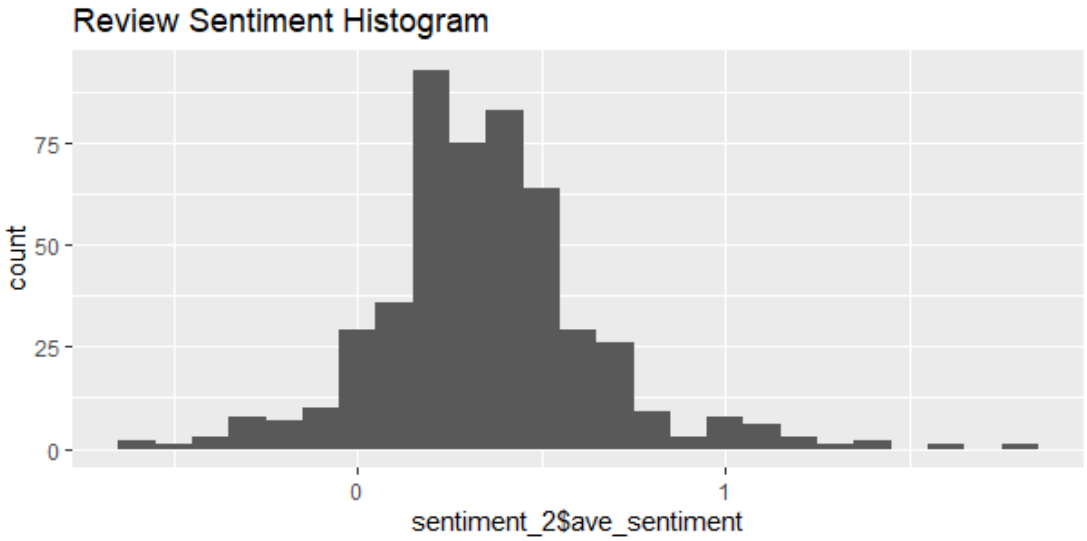
Confermando la mia tesi con il grafico della frequenza della polarità del sentiment, sopra mostrato, dove le positive superano di gran lunga le negative mentre le neutrali sono leggermente sopra lo 0, infatti sono 3.

Come seconda analisi ho eseguito il calcolo del sentiment sulle recensioni complete, per vedere se coincide con il sentiment delle frasi. Di conseguenza attraverso la funzione sentiment_by() possiamo raggruppare in un dataset tutte le recensioni e calcolare i vari valori che ci fornisce questa funzione. Il dataset che otteniamo è:

	element_id	word_count	sd	ave_sentiment
1	1	267	0.247277915	0.0970073181
2	2	63	0.224133255	0.1284807508
3	3	69	0.335692056	0.2832481466
4	4	147	0.308150737	0.1132605648
5	5	61	0.203911061	-0.0761945847
6	6	110	0.259750324	0.2230084611
7	7	21	*NA*	-0.1636634177
8	8	21	0.325970496	0.4309418796
9	9	188	0.319949348	0.1735550352
10	10	114	0.334085794	0.4351591662
11	11	154	0.204410529	0.2514721426
12	12	81	0.272793156	0.2656804702
13	13	164	0.339010619	0.1015386666
14	14	117	0.253819072	0.3069906878
15	15	54	0.166000452	0.4467827486
16	16	55	0.082304795	0.3300079936
17	17	64	0.525097397	-0.2539092337
18	18	53	0.499279832	0.2367687460

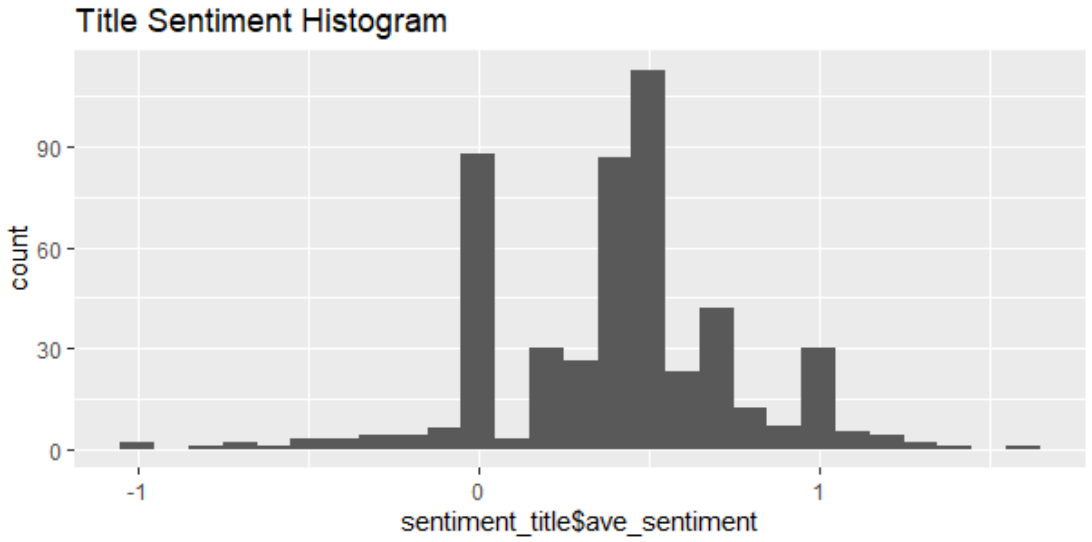
Showing 1 to 18 of 500 entries, 4 total columns

Ottenuto questo ho graficato l'istogramma che mi permetterà di capire se coincideranno o meno, il risultato mostra:



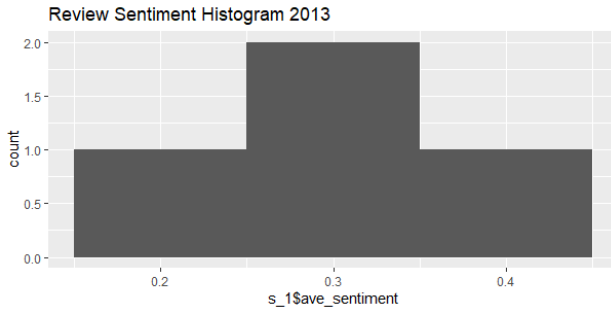
notiamo come l'andamento è maggiore nell'intervallo tra 0 e 1, quindi il sentiment certamente più che positivo. Di conseguenza il sentiment delle frasi singole e delle recensioni complete coincide nel risultato, quindi possiamo confermare che il prodotto è apprezzato dai clienti.

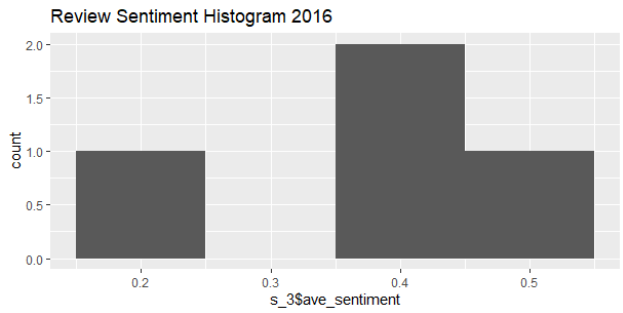
Continuando con la sentiment analysis ho analizzato la polarità dei titoli, mi aspetto che siano presenti meno parole ma che comunque siano più positive che negative. Eseguendo la stessa tecnica spiegata, ho utilizzato sentiment_by() per raggruppare le recensioni con i loro valori, grazie a quest'ultimi analizziamo i valori della colonna dei valori ponderati del sentiment per realizzare il grafico. Il risultato che otteniamo è:

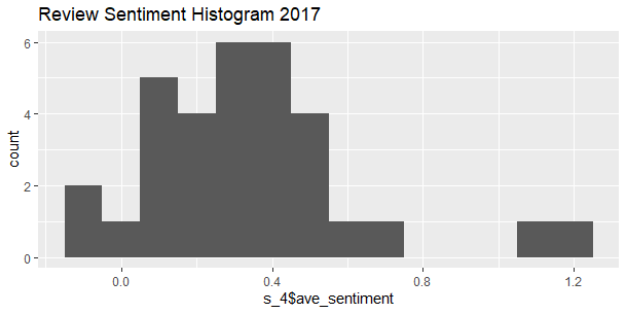


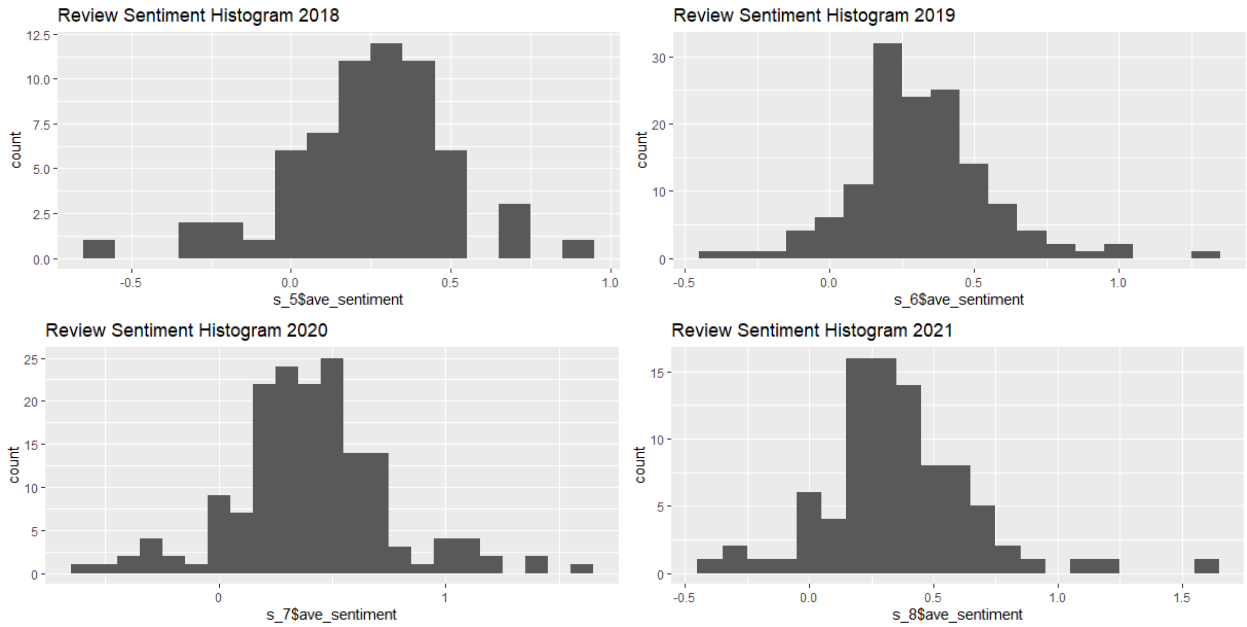
complessivamente possiamo dire che la polarità è positiva e quindi anche dai titoli si evince l'apprezzamento dei clienti, inoltre a differenza del testo delle recensioni molti dei titoli sono di carattere neutro lo possiamo anche comprendere dalle poche parole.

L'ultima valutazione che ho conseguito è un'analisi annuale del sentiment delle recensioni, per un'ulteriore conferma che quest'ultimo coincida con il sentiment delle stelle. Ho eseguito lo stesso metodo per realizzare l'analisi annuale delle stelle, quello che ho ottenuto sono i seguenti grafici:





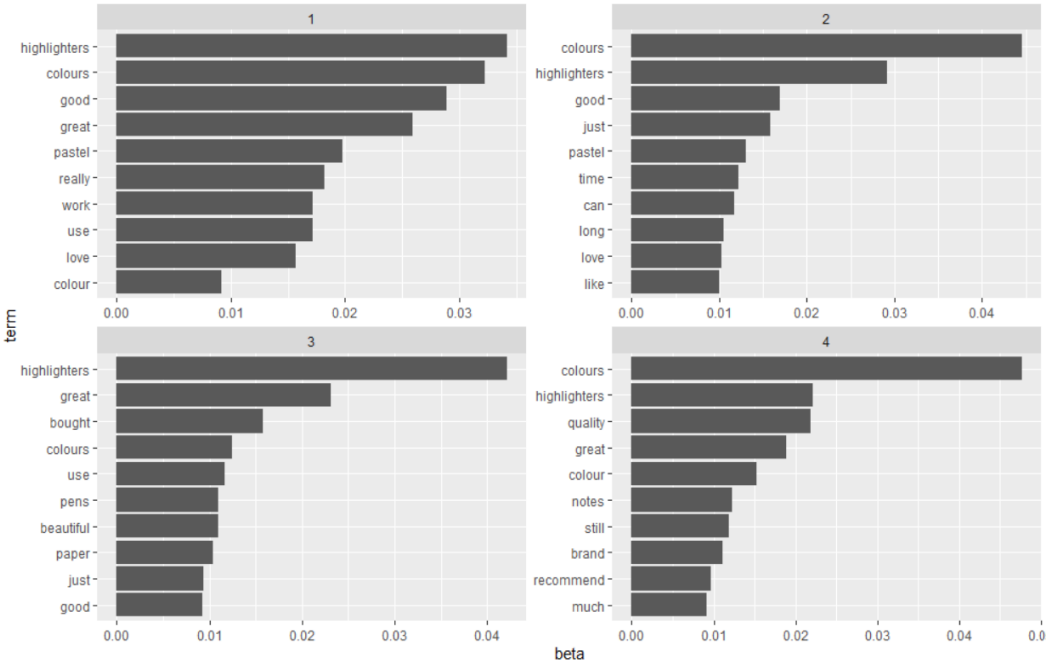




nell'intervallo dal 2013 al 2017 sono poche recensioni per questo hanno preso una forma dell'istogramma molto strana, pero sono comunque complessivamente positive. Nel resto degli intervalli sono presenti più recensioni quindi il grafico assume una distribuzione più normale, mantenendo una polarità più che positiva. In conclusione di questa analisi, come già detto, il prodotto che sto analizzando è più che apprezzato.

Topic Modelling:

L'ultima condotta è il topic modelling sulla colonna delle recensioni, per condurla ho utilizzato le librerie topicmodels e tidytext. Quello che mi aspetto, da questa analisi, è che più o meno i topic che realizzerò saranno molto simili, perché su 500 recensioni ci saranno pochi argomenti che si discostano dagli evidenziatori. Di conseguenza sono partito creando la struttura del LDA (Latent Dirichlet Allocation), attraverso la funzione LDA() di topicmodels, che permette di calcolare il valore della probabilità che una determinata parola ricaschi in uno dei topic, beta, questa probabilità dipende da quanti argomenti vogliamo analizzare. Analizzerò, in questo caso, 4 topic, immettendo quindi in LDA() il valore "k=4" così da poter poi graficare gli argomenti ottenuti. Infine attraverso la funzione topic_terms ho raggruppato i termini ottenuti in funzione del valore beta, spiegato poco sopra, e di conseguenza ho graficato gli argomenti ottenendo:



Possiamo notare che nei 4 topic creati le parole “highlighters” e “colours” appaiono in tutti e molto frequentemente. Concludiamo parlando degli argomenti che abbiamo realizzato:

- Nel topic 1: osserviamo che appare la parola “work” quindi deduco che il prodotto nell’ambito lavorativo è molto utilizzato e dalla parola “love” è apprezzato;
- Nel topic 2: è presente la parola “time” quindi possiamo pensare che si possa riferire allo studio e in questo campo durano molto e sono amati;
- Nel topic 3: si tratta dell’acquisto del prodotto, intuito dalla presenza della parola “bought” e di conseguenza è un acquisto approvato dai clienti;
- Infine, nel topic 4: appaiono le parole quality e recommend, si evince che l’argomento è un insieme di recensioni che trattano della qualità del prodotto ed è positiva anche sotto questo punto di vista.