Topic: Text Mining (NLP)

One:

1) Extract reviews of any product from ecommerce website like Snapdeal and Amazon

2) Perform sentimental analysis

Amazon Reviews & Sentiment Analysis for iPhone 11 Pro 256 GB

Reviews extraction from website

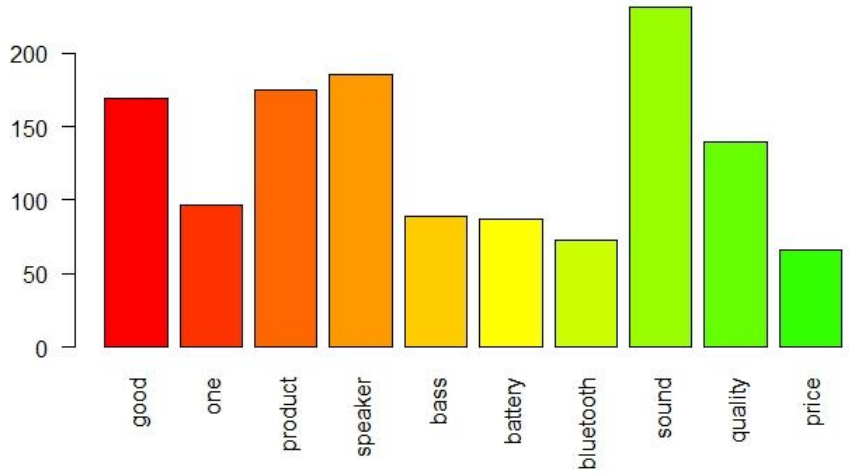
1. Loading the reviews of a product from the website url of the review page.
2. Need to create a null object to insert the reviews.
3. Now, the loop is created to extract the reviews with the pages numbers where we need to use the html nodes of the page, so that all the reviews are now assigned to the null object.
4. The reviews text files id ready for further analysis in text mining.

Steps involved in Sentiment Analysis:

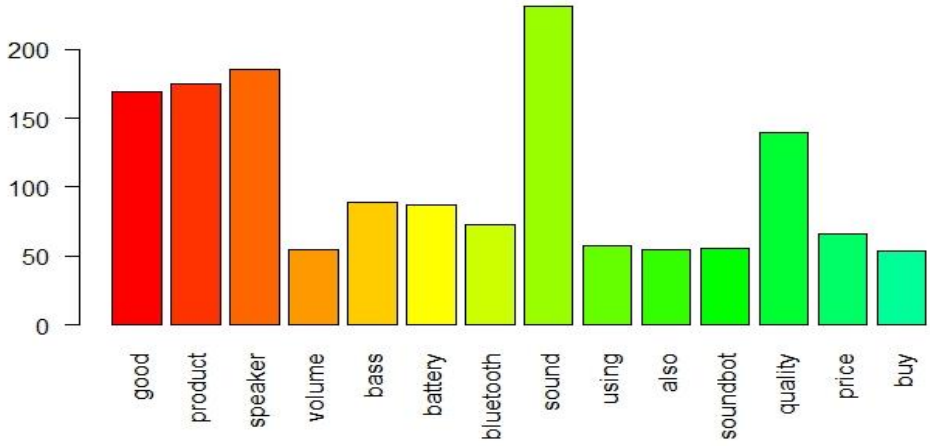
5. Now using “tm” library we do sentiment analysis by converting character data into corpus type.
6. Inspect is the function which is used to check the individual review with the help of the review number.
7. Now the corpus has to be cleaned using the different data cleansing techniques.
8. Converting the whole corpus in to lower case letters using the function called “tolower” and later the “punctuations” and “numbers” are removed from the corpus.
9. To make the data more sensible, “stop words” are removed from the corpus for further analysis.
10. As the stop words are removed, it forms a white space which need to be removed by using “stripWhitespace”.

- As the corpus is in unstructured format now by using “Term Document Matrix” or “Document Term Matrix” , the data is converted into structured format.

- Now take first 20 reviews and creating a TDM and a Bar plot is taken to see which words are repeated at the most.
- Taking the words which are repeated >= 60 are used to build a bar plot

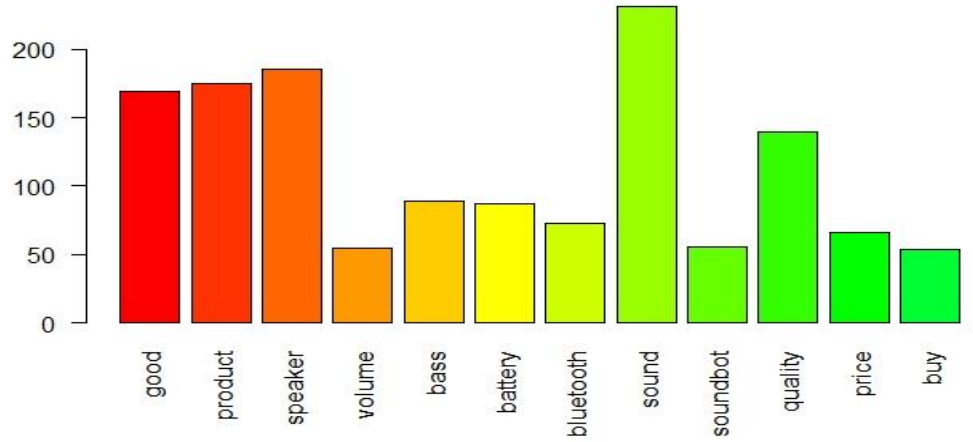


- Now still further data cleansing needs to be done by removing words which doesn't make any sense / influence in the analysis.
- Here the terms "one", which is repeating more than 50 times, so these words are now removed from the corpus.
- After removing the terms and white spaces formed, once again a tdm is constructed to checking weather any further cleansing of data is required by taking the reviews from 100 to 109 with the matrix of 1:20

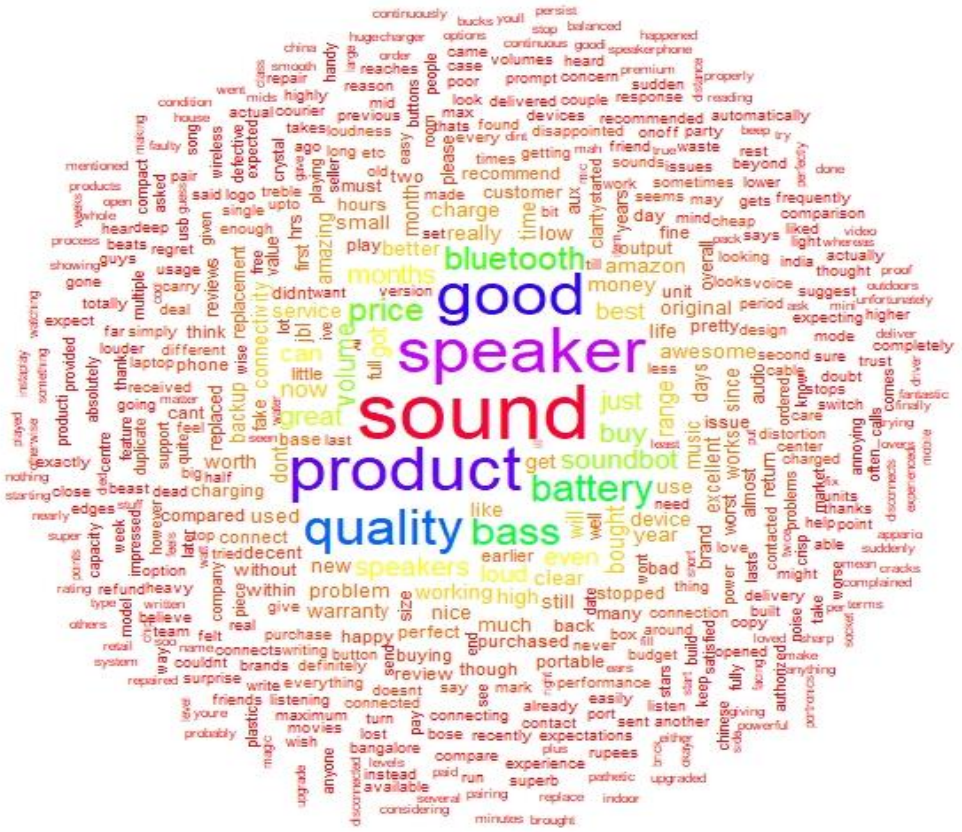


- Now still further data cleansing needs to be done by removing words which doesn't make any sense / influence in the analysis.

- Here the terms "using" and "also" are repeating more than 50 times, so these words are now removed from the corpus.
- After removing the terms and white spaces formed, once again a tdm is constructed to checking weather any further cleansing of data is required by taking the reviews from 100 to 109 with the matrix of 1:20



- The graph shows that there are no words that doesn't add any value to analysis.
- Using the "wordcloud" library the word could be built using the corpus after the data cleansing.



- Using the wordcloud, it shows the most common words , average ratings and sentiment scores associated with each word.

  Similarly using the wordcloud2 library, we can use the same common words is different shapes as shown below.

  wordcloud2(w1, size=0.5, shape = 'circle') ← Circle Shape

wordcloud2(w1, size=0.5, shape = 'triangle') ← Triangle Shape

wordcloud2(w1, size=0.5, shape = 'star') ← Star Shape

Bigram Analysis:

- Using the libraries “rjava” and “RWeka” a 2 words wordcloud is built.
- With minimum frequency 2 and NgramTokenizer a two-word data frame is formed.

- The bigram wordcloud shows the most common words and the average ratings and sentiment scores associated with each word

- Many of these words are divided in two clusters: one with a positive sentiment score and one with a negative sentiment score. The quantity of words with positive Amazon ratings but negative sentiment scores is concerning, so we will look into the effect this has on sentiment by further analysis using the positive and negative word clouds

Positive Word Cloud:

Negative Word Cloud:

1) Extract movie reviews for any movie from IMDB and perform sentimental analysis

IMDB Reviews and Sentiment Analysis on Movie (Forrest Gump)

Reviews extraction from website

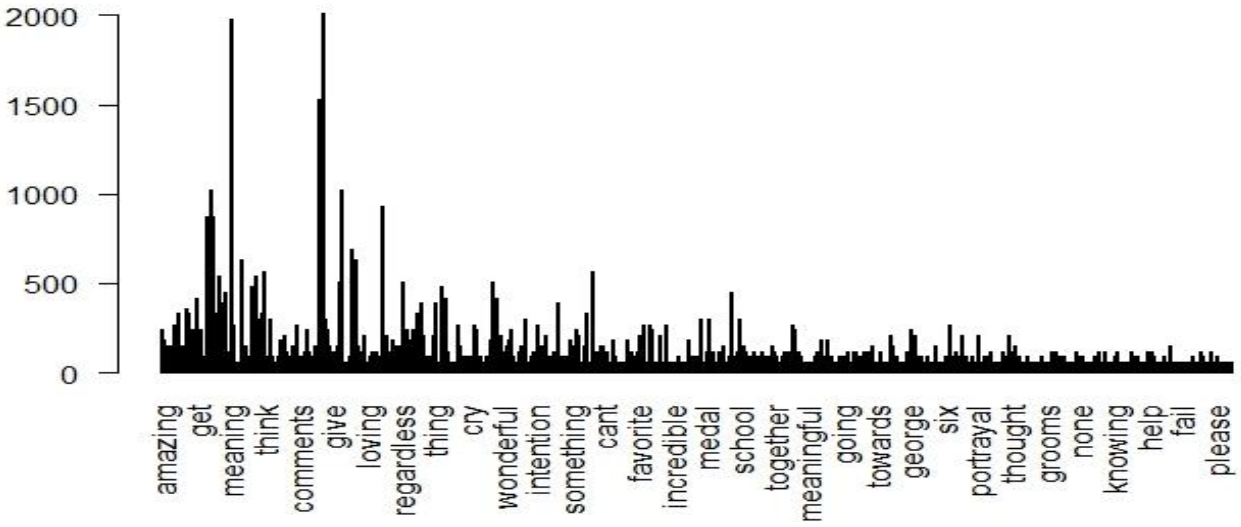
1. Loading the reviews of a movie from the website with the help of url.
2. Need to create a null object to insert the reviews.
3. Now, the loop is created to extract the reviews with the pages numbers where we need to use the html nodes of the page, so that all the reviews are now assigned to the null object.
4. The reviews text files id ready for further analysis in text mining.

5. **Steps involved in Sentiment Analysis:**

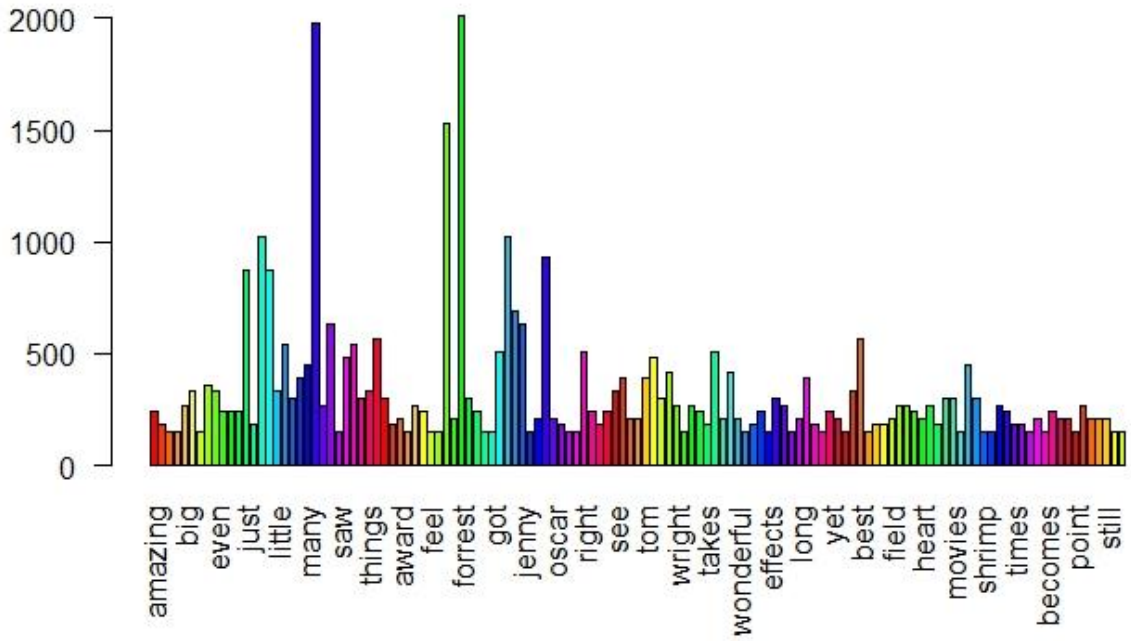
6. Now using “tm” library we do sentiment analysis by converting character data into corpus type.
7. Inspect is the function which is used to check the individual review with the help of the review number.
8. Now the corpus has to be cleaned using the different data cleansing techniques.
9. Converting the whole corpus in to lower case letters using the function called “tolower”.
10. Now the punctuations and numbers are removed from the corpus.
11. To make the corpus more sensible stop words are removed from the corpus to make further analysis.
12. As the stop words are removed there forms a white space which need to be removed by using “stripWhitespace”.
13. As the corpus is in unstructured format now by using “Term Document Matrix” or “Document Term Matrix” , the data is converted into structured format.

14. Now take first 20 reviews and creating a TDM and a Bar plot is taken to see which words are repeated at the most.

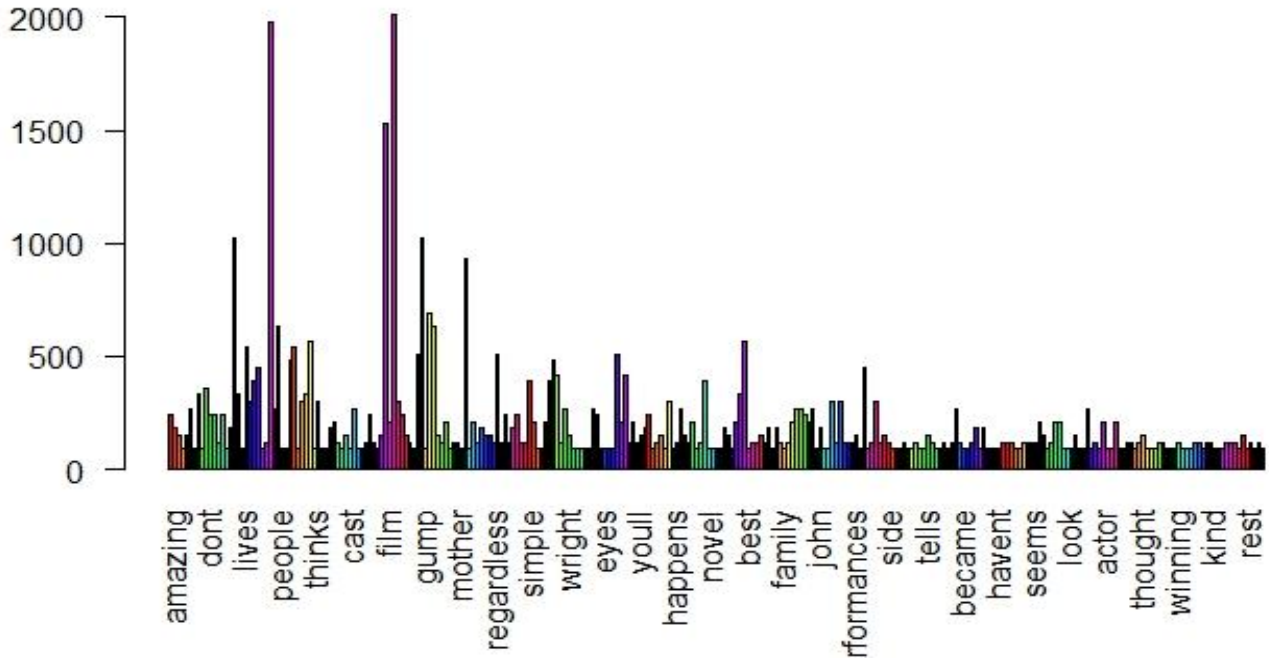
15. Taking the words which are repeated >= 60 are used to build a bar plot.



- Now still further data cleansing needs to be done by removing words which doesn't make any value in the analysis.
- Here the terms 'going','cant','get', 'give', 'going', 'none' are repeating more than 150 times, so these words are now removed from the corpus
- After removing the terms and white spaces formed, once again a tdm is constructed to checking whether any further cleansing of data is required by taking the reviews from 100 to 120 with the matrix of 1:20
- The bar plot visualization after removing the above terms and now the bar plot is constructed with the words >= 150, below is the bar plot representation.



- Now still further data cleansing needs to be done by removing words which doesn't make any value in the analysis.
- Here the terms 'even','just','feel', 'got', 'see', 'takes', 'yet', 'still', 'didnt', 'like', 'saw', 'well', 'along', 'also', 'told', 'goes' are repeating more than 80 times, so these words are now removed from the corpus
- After removing the terms and white spaces formed, once again a tdm is constructed to checking whether any further cleansing of data is required by taking the reviews from 100 to 120 with the matrix of 1:20
- The bar plot visualization after removing the above terms and now the bar plot is constructed with the words >= 80, below is the bar plot representation.



- The graph shows that there are no words that doesn't add any value to analysis.
- Using the "wordcloud" library the word could be built using the corpus after the data cleansing

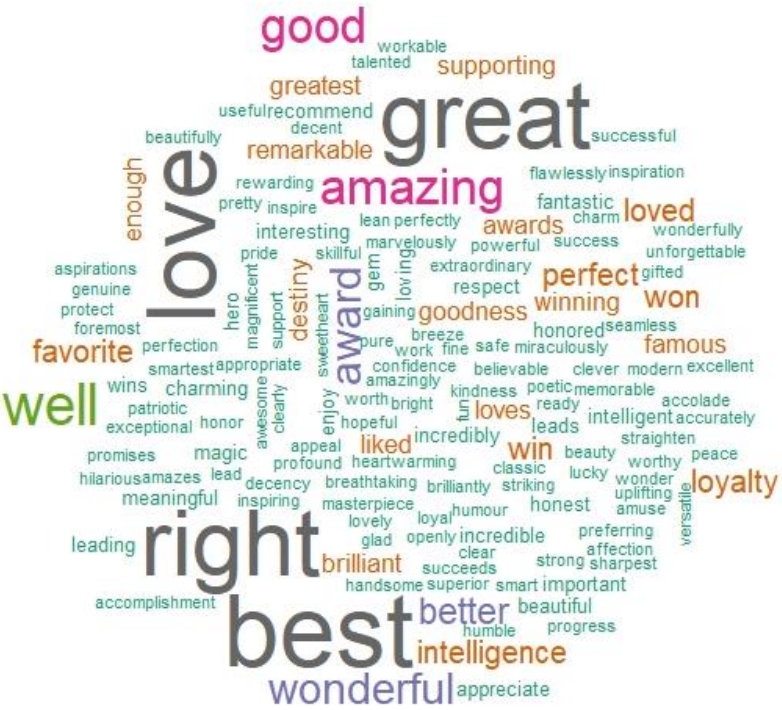
- The wordcloud shows the most common words and the average ratings and sentiment scores associated with each word

Bigram Analysis:

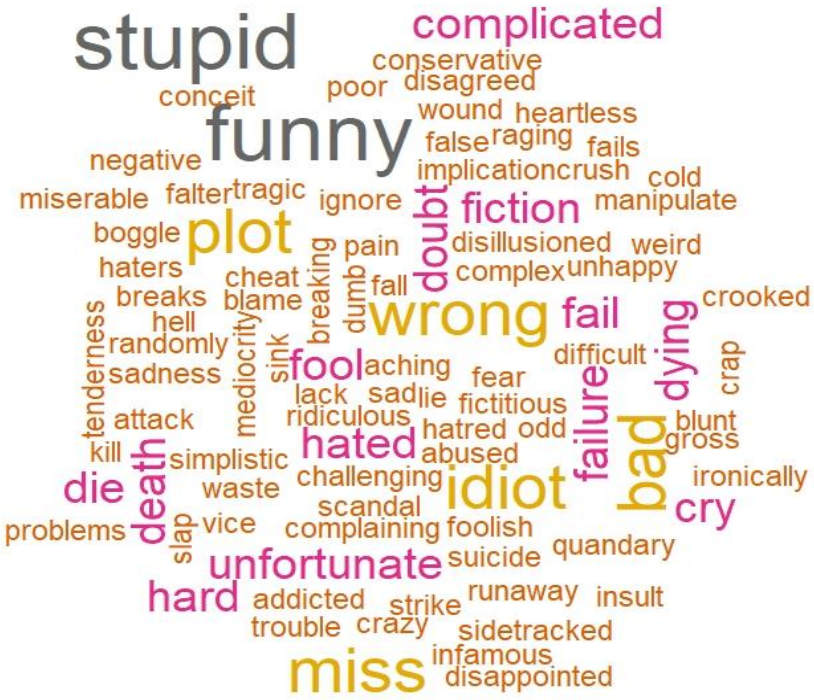
- Using the libraries “rjava” and “RWeka” a 2 words word cloud is built
- With minimum frequency 2 and NgramTokenizer a two-word data frame is formed

- The bigram wordcloud shows the most common words and the average ratings and sentiment scores associated with each word.
- Many of these words are divided in two clusters: one with a positive sentiment score and one with a negative sentiment score. The quantity of words with positive Amazon ratings but negative sentiment scores is concerning, so we will look into the effect this has on sentiment by further analysis using the positive and negative word clouds.

Positive Word Cloud:



Negative Word Cloud:



2) Extract anything you choose from the internet and do some research on how we extract using R Programming and perform sentimental analysis.

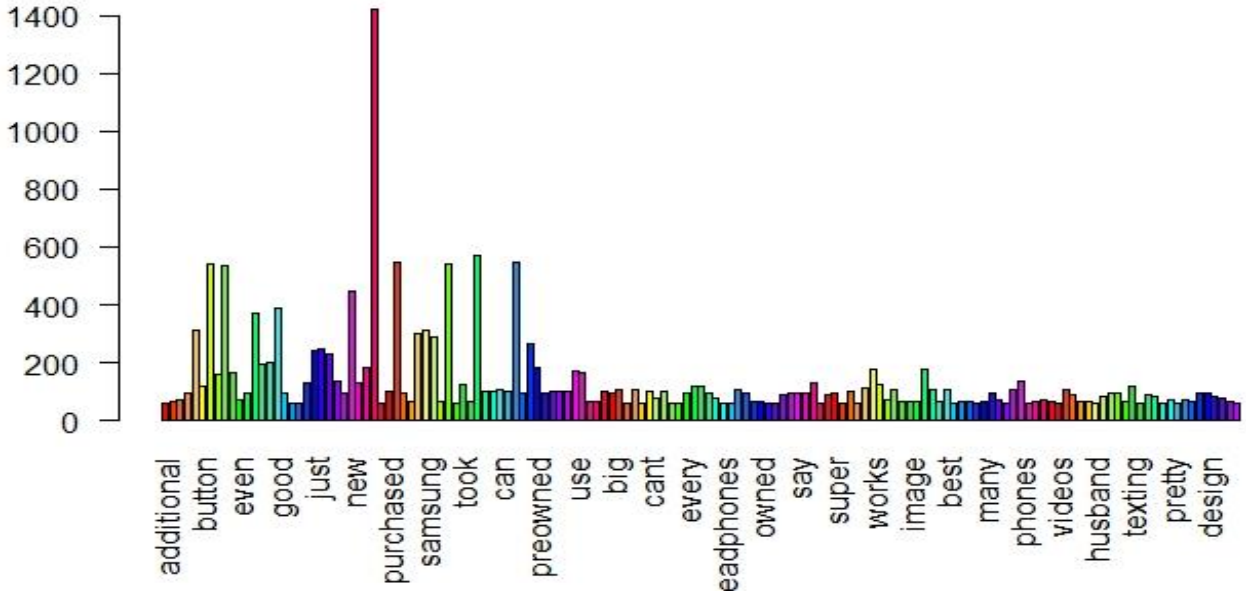
Ebay Reviews and Sentiment Analysis on Samsung Galaxy S8

Reviews extraction from website

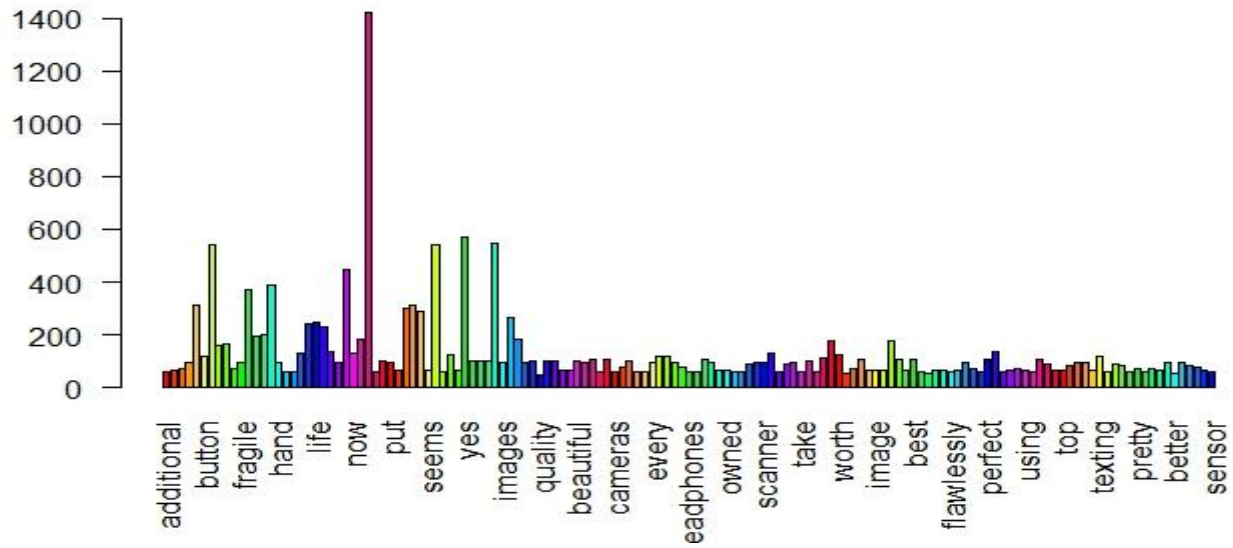
1. Loading the reviews of a product from the website using url.
2. Need to create a null object to insert the reviews
3. Now, the loop is created to extract the reviews with the pages numbers where we need to use the html nodes of the page, so that all the reviews are now assigned to the null object.
4. The reviews text files id ready for further analysis in text mining.

Steps involved in Sentiment Analysis:

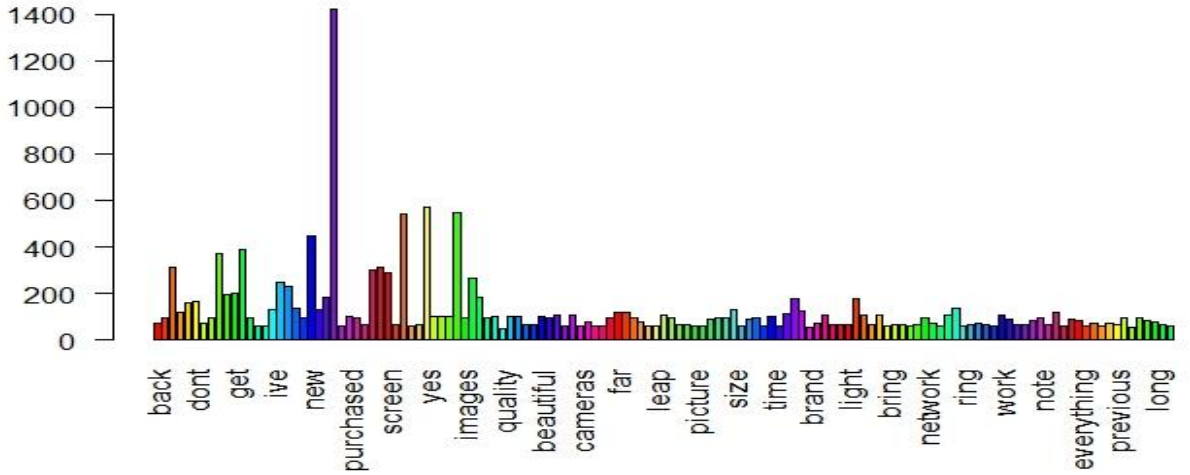
5. Now using “tm” library we do sentiment analysis by converting character data into corpus type
6. Inspect is the function which is used to check the individual review with the help of the review number.
7. Now the corpus has to be cleaned using the different data cleansing techniques.
8. Converting the whole corpus in to lower case letters using the function called “tolower”.
9. Now the punctuations and numbers are removed from the corpus.
10. To make the corpus more sensible stop words are removed from the corpus to make further analysis.
11. As the stop words are removed there forms a white space which need to be removed by using “stripWhitespace”.
12. As the corpus is in unstructured format now by using “Term Document Matrix” or “Document Term Matrix” , the data is converted into structured format.
13. Here we took first 20 reviews and creating a TDM and a Bar plot is taken to see which words are repeated at the most.
14. Taking the words which are repeated >= 50 are used to build a bar plot.



- Now still further data cleansing needs to be done by removing words which doesn't make any value in the analysis.
- Here the terms , 'use',say','can','cant','also','husband','conditionâ','purchaseâ' are repeating more than 50 times, so these words are now removed from the corpus
- After removing the terms and white spaces formed, once again a tdm is constructed to checking whether any further cleansing of data is required by taking the reviews from 100 to 109 with the matrix of 1:20
- The bar plot visualization after removing the above terms and now the bar plot is constructed with the words >= 50, below is the bar plot representation.



- Now still further data cleansing needs to be done by removing words which doesn't make any value in the analysis.
- Here the terms , 'additional','just','took','doesnt','came','need','byâ','almost' repeating more than 50 times, so these words are now removed from the corpus
- After removing the terms and white spaces formed, once again a tdm is constructed to checking whether any further cleansing of data is required by taking the reviews from 100 to 109 with the matrix of 1:20
- The bar plot visualization after removing the above terms and now the bar plot is constructed with the words >= 50, below is the bar plot representation.

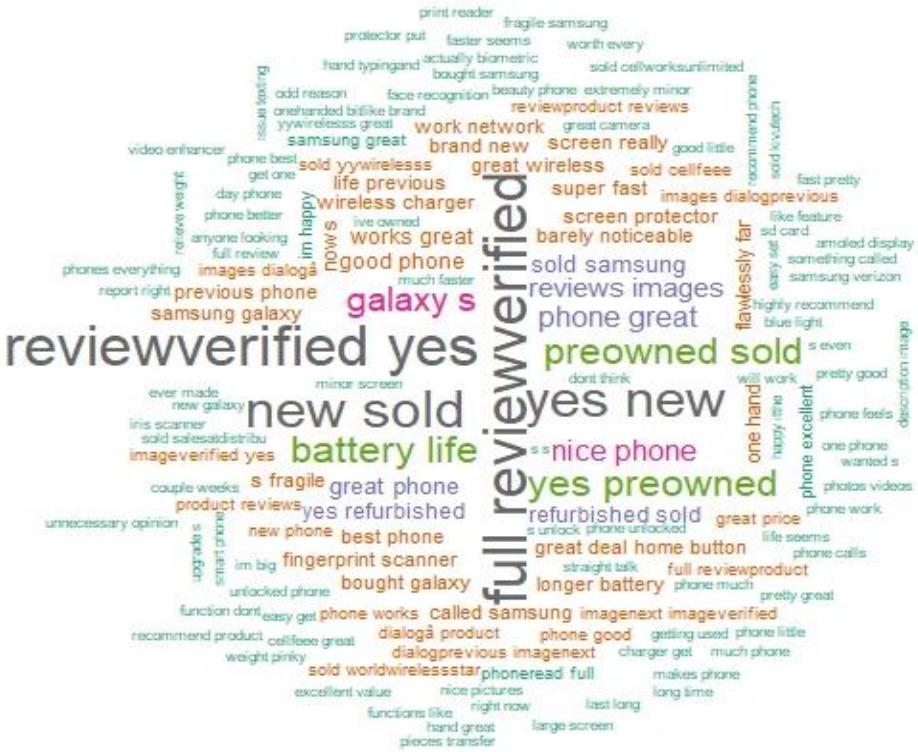


- The graph shows that there are no words that doesn't add any value to analysis.
- Using the "wordcloud" library the word could be built using the corpus after the data cleansing

- The wordcloud shows the most common words and the average ratings and sentiment scores associated with each word.
-

Bigram Analysis:

- Using the libraries "rjava" and "RWeka" a 2 words word cloud is built
- With minimum frequency 2 and NgramTokenizer a two-word data frame is formed



- The bigram wordcloud shows the most common words and the average ratings and sentiment scores associated with each word
- Many of these words are divided in two clusters: one with a positive sentiment score and one with a negative sentiment score. The quantity of words with positive Amazon ratings but negative sentiment scores is concerning, so we will look into the effect this has on sentiment by further analysis using the positive and negative word clouds

Positive Word Cloud:

Negative Word Cloud:

Hints:

1. Business Problem

    1.1. Objective
    1.2. Constraints (if any)

2. Data Pre-processing

    2.1 Data cleaning, Feature Engineering etc.

3. Model Building

    3.1 Partition the dataset
    3.2 Model(s) - Reasons to choose any algorithm
    3.3 Model(s) Improvement steps
    3.4 Model Evaluation
    3.5 Python and R codes

4. Deployment

    4.1 Deploy solutions using R shiny and Python Flask.

5. Result Share the benefits/impact of the solution - how or in what way the business (client) gets benefit from the solution provided.

Note:

1. For each assignment the solution should be submitted in the format
2. Research and Perform all possible steps for improving the model(s) accuracy

    Ex: Feature Engineering (unigram, bigram, trigram, word-cloud etc.

3. All the codes (executable programs) are running without errors
4. Documentation of the module should be submitted along with R & Python codes, elaborating on every step mentioned here.