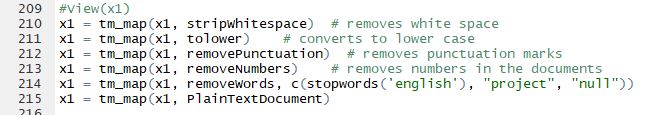
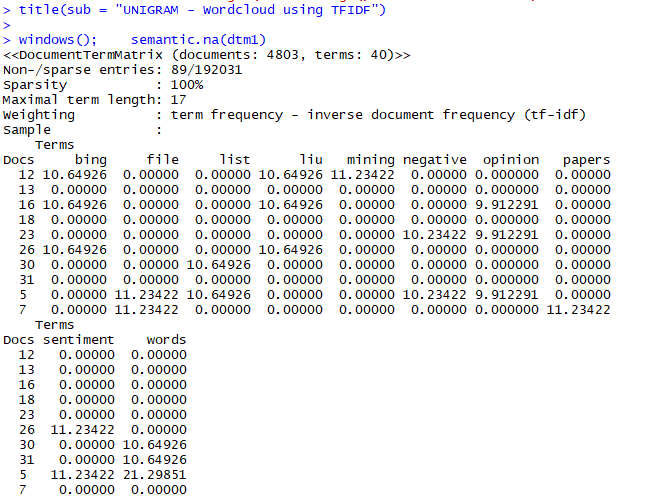
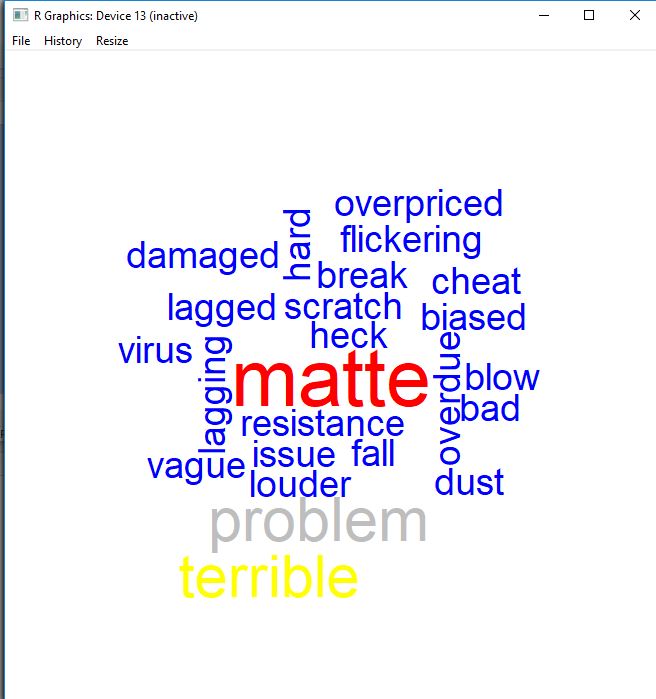
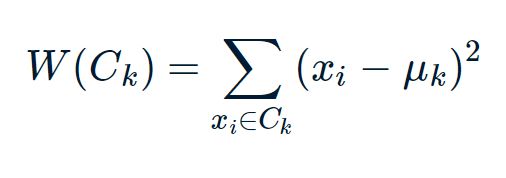
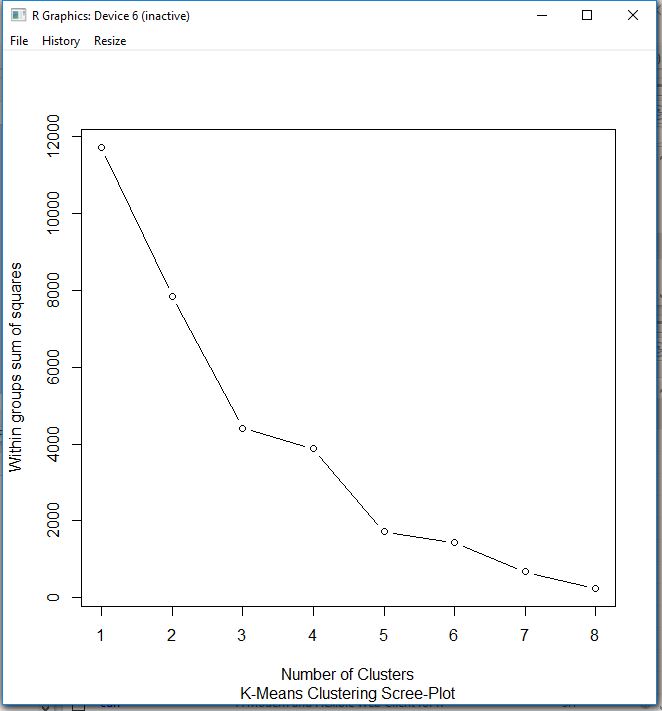
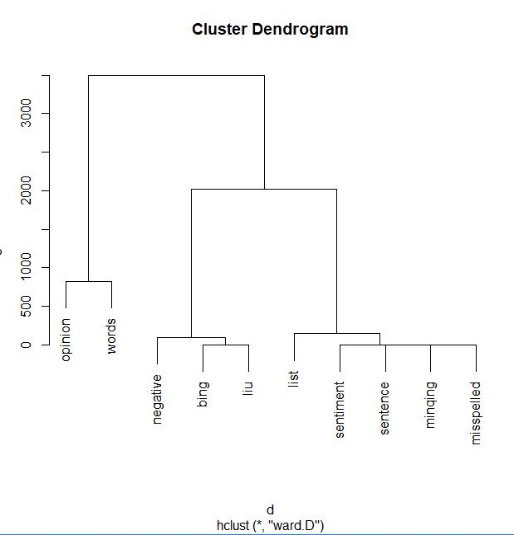
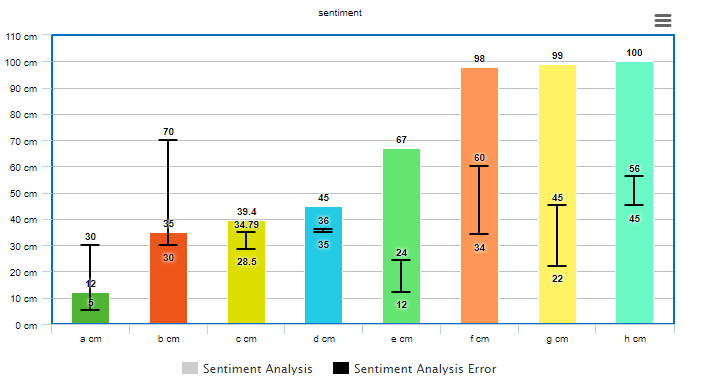
Sentiment Analysis on Amazon Review  
  
Team no: 09  
  
Team Member: Reetayan Das  
  
contact:reetayd@clemson.edu  
  
  
  
  
  
  
  
  
  
  
  
  
  
  
Contents  
  
1. Introduction  
  
2. What is sentiment analysis and why we need it?  
  
3. The Key Components in Sentiment Analysis   
  
4.Types of method  
  
5.Data Pre-processing  
  
6.Classifiers: a)BAG of WORDS(tf\_tdf model)  
   
 a)K means clustering  
  
 b)Hierarchical Classification  
  
7.Performance Analysis  
  
8.Limitations of Sentiment Analysis  
  
9.Conclusions   
  
10.References  
  
  
  
  
  
  
  
  
  
  
  
  
 Introduction  
  
  
Being the consumers we always expect the best quality of product from the vendors. The stakeholders from the product based companies also care about the customer satisfaction and they really do want to know about the products behavior in the current market. They need to know that really how much the product retains the potential on its own, what is its pros and cons. The consumers are saying their opinions and statements by posting their reviews in the customer forum. To examine the product’s goodness we need to analyze the reviews carefully. From the stakeholder’s point of view if we really gather the good points and bad points together from a review it will at least give us an insight which will help us to take decisions about the product whether we need to launch a new upgraded version or we need to downgrade the version. Consumers will be able to think whether they will recommend or buy the product later or not.  
  
What is sentiment analysis and why we need it?  
  
It’s the procedure of picking up the feelings, emotions and ideas from where we want to gain an insight. We do that by the use of natural language processing (NLP) and computational techniques to automate the extraction or classification of sentiment form unstructured text. The Text Analysis techniques and applying of Natural Language processing’s cover a major role here. Picking up an impression of the product from the reviews make a people known about that product and make others aware too. In this “others” a lot of people exists. We all are consumers but few people are stakeholders who really care about the customer satisfaction. So, the review or opinions on which the product’s existence and rating depends gives us a polarity of the text .It can be positive , negative or neutral.  
  
As more reviews are getting shared through online, more insights we are getting and the insights are changing every time. At a certain point of time we are considering an amount of review statements and after doing we are getting an impression of that product. It can be very possible that the next couple of statements can change the degree of insights within a short span of time. In a recent study carried out by Zendesk it was noted that 45 % of the negative and 30 % positive customer experiences are generated through the social media.  
  
  
The Key Components in Sentiment Analysis  
  
  
  
**Which Features to use**:  
  
The first problem here is which feature we should use in sentiment analysis.Having a body of text we need to know how to classify it. We normally cover here the text mining units from character level to semantic level. Most of the sentiment analysis we do based on the unigrams (words).Recently some study shows that n grams give the better results in this case. Sentences are also used as a unit for analysis in the field of opinion mining.  
  
Types of Methods  
  
  
Supervised Learning: The Data (observations, measurements etc) are labeled   
 with predefined classes. The analogy would be that a   
 supervisor gives classes(supervision) .Test Data are   
 classified into these classes too.  
  
Unsupervised learning: Here the class labels of the data are unknown. Given a   
 test Dataset, the task is to establish the existence of   
 classes or clusters in the data.  
  
Lexicon based Approaches: In this method a small set of opinion words collected   
 as a seed.Then Well known dictionaries are used to   
 expand the set of opinion words by adding their   
 synonyms and antonyms. The newly found words are added to seed list. The process continues until no more words found in the dictionary. In the end manual review is carried out to remove errors.  
  
  
   
  
  
  
  
  
Methods we followed   
  
  
BAG OF WORDS  
  
I was planning to develop a system which can predict the sentiment of a product   
review from the website <https://www.amazon.com>. I have gathered the trained datasets of a set of “positive” and “negative words” from the repository of University of Illinois , Chicago. As Test Data I have chosen a product random from the amazon for example ; Dell inspiron laptop. I have taken all their reviews in a single text file. Those reviews I wanted to get compared with the list of positive and negative words.  
  
Data Pre-Processing:  
  
For cleaning I removed the white space, punctuation marks and numbers.  
Then converted the upper case to lower cases and save it to plain text document.  
  
  
  
  
  
  
  
  
  
  
  
  
  
Term Document Matrix and Bag Of Words:  
  
I needed to create a term document matrix wherein this case our main objective was to make a word-cloud using TF-IDF score. Document Term matrix is normally used when each document represents as a row and each column is referred as a term. We observed the term frequency(tf) and find out the inverse document frequency(idf) and multiply both to get the TF-IDF score.  
Boolean  "frequencies": TF(*t*,*d*) = 1 if *t* occurs in *d* and 0 otherwise;

Term frequency adjusted for document length : *ft*,*d* ÷ (number of words in d)

Logarithm scaled frequency: TF (*t*,*d*) = log (1 + *ft*,*d*)  
  
IDF (t,D)=log(N/{d∈D;t∈d})   
where 1)N defines the total number of documents in the corpus .  
 2) {d∈D;t∈d}) defines the number of documents where t appears.  
  
TF\_IDF scoring : TF(t,d) \* IDF(t,D)  
  
So TF\_IDF scoring is a product of 2 matrices. For each entry in the matrix the term frequency measures the number of times that term (for example i)appears in a document(for example j), and the inverse document frequency measures the number of documents in the corpus   
which contain term i. When the term appears more frequently in the document then the TF\_IDF score will get increased but decreased when the term appears in other documents.  
Below is the picture of my sample output for tf idf.  
  
  
Based on the TF\_IDF score we get a representation of BAG OF WORDS where we observed the most frequent words are actually getting populated based on their score.

  
  
Above is the bag of words for a set of positive word ”**best**” mostly occurred and below is the example for the negative word “**matte**” and “**terrible**” mostly occurred.  
  
  
  
  
   
K Means Clustering   
  
I have followed another method named K means clustering where the main objective of the clusters is to find the similar words and put it them in the clusters.  
This procedure of the clustering is done by first initializing the data points each onto a cluster find out the centroid co-ordinate in cluster. The centroid keeps the average distances in the whole cluster. As long as we introduce data points in the cluster the co-ordinates of the centroid is getting changed .We will stop clustering until the centroid position is not getting changed. For Calculating the distance of centroid I needed the find out the summation of squared differences between the   
data points and the mean value assigned on the clusters.  
Here each data points represents the scoring for each word.  
  
  
  
  
  
Below is the formula for finding out the distance for centroid.  
   
  
  
  
Below is the K means scree plot where in x axis the number of clusters is plotted   
and in y axis the sum of squares(w) mentioned above is plotted. If we see this graph we will understand that as long as the number of clusters increased(more data points are getting introduced) ,the W is getting down, but at a certain point k=3 the graph started going to parallel to x axis. That means the change of centroid is not getting affected much after k=3.  
  
  
  
  
  
Advantages and Disadvantages of K means Clustering:  
  
  
Hierarchical clustering  
  
  
The Hierarchical clustering method also I applied here .This another form of clustering followed agglomerative or bottom-up approach. The Leaf nodes on the branch are actually the base clusters where at the very initial. We then combine those base clusters and form a new cluster. In this case we need to find the nearness of the clusters. Nearness is being judged by the distance between the clustroid in clusters.   
  
clustroid = point “**closest**” to other points  
  
The possible meaning of the term “**closest**”:  
  
1)Smallest Maximum Distance to other points  
2)Smallest average Distance to other points  
3)Smallest sum of squares of distances to other points  
  
  
Termination condition:  
  
So I needed to pick a number of **K** upfront and stop when we have **K** clusters.  
It makes sense when we know the data naturally falls into k classes.  
Suppose we have two data points like “Equality” and “reliability” those belong to  
two different classes. When we got to know the two clusters then we stop.  
  
  
Above is the cluster dendrogram generated in the model.  
  
Performance Analysis  
  
  
For the above mentioned classifiers I used in in my project from the stakeholder point of view we now able to make a decision about the reviews of that product. It gives us an insight that which product is going to sustain in the market in future and also how much it serves the customer at it’s best. There are several statistical methods we can apply here.  
  
Sentiment rating can be shown in this   
  
  
  
Above is the sentiment analysis of the Dell Inspiron Laptop for different input reviews from different thread.  
  
  
  
  
  
  
  
  
Limitations of sentiment Analysis:  
  
Sentiment Analysis has its own limitations. The Text Mining methods are present   
to pick up the words and classify them. But for actual classification the emotion and sentiment depends on the mood of the language not the vocabulary completely.  
Means, for an example, its difficult to find the meaning of a sarcastic comment.  
Suppose there is a review “I was expecting it better to be delivered early”.  
The Classifier here will search for **better** word and infer the positive score from it.  
But The customer actually gives a negative impression on the review.  
  
Similarly, we are taking the analysis for a certain amount of time gathering a couple of reviews. The whole impression on the sentiment can be change on by getting a couple opposite reviews which might come further.   
  
  
Conclusion:  
  
As I have taken the paper of Sentiment Analysis on Amazon Review It gave me Satisfactory results to determine the sentiment of the review on a particular product by using the k means clustering, hierarchical clustering and tf\_idf model(bag of words).These methods gave me an idea on sentiment analysis, based on which we can take decisions on buying or selling the particular product and checking the sustainability in market also.  
  
  
  
References:  
  
[1] <https://machinelearningmastery.com/gentle-introduction-bag-words-model/>

[2] <http://ijcsit.com/docs/Volume%206/vol6issue06/ijcsit2015060652.pdf>

[3] <https://www.amazon.com/Dell-Inspiron-Touchscreen-Performance-Bluetooth/dp/B0778DVNHH/ref=sr_1_6?s=grocery&ie=UTF8&qid=1524071707&sr=8-6&keywords=dell+laptop>

[4] <https://idc9.github.io/stor390/notes/natural_language_processing/text_classification.html>

[5] <https://www.quora.com/What-is-a-term-document-matrix>

[6] <http://blog.aylien.com/why-is-sentiment-analysis-important-from-a/>  
  
  
[7] http://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

Notes: As my previous selected paper was “Matching structured with Unstructured Data” I decided to change my project topic on as per the reasons following  
  
1)Datasets were not available for the copy right issue and also I was not getting proper dataset.  
  
2)It was difficult to use word2vec and train a neural network model with smaller data.  
  
The unstructured data mostly used in the field of NLP(Natural Language processing) I decided to choose this paper for sentiment analysis which gives a proper vision for text mining and sentiment analysis.

---------------------------------