Sentiment Analysis for Ipad 2

Sida Ye

Prof. Aldous *Department of Statistics*

Introduction to sentiment analysis

Whenever we need to make a decision, we often seek out the opinions of others. In the past, we seek opinions from friends and family or companies would use surveys, focus groups, opinion polls, and consultants. Now, customer reviews on the Internet has risen exponentially over the last decade. It is an important resource for buying products or attending events. In these situations, we would like to see what others are saying about them. Also it is a significant aspect for companies making decisions about their products or services.

Sentiment analysis, the computational study of how opinions, attitudes and emotions are expressed in nature language, provides techniques for extracting some emotional words, subjective information from large datasets of customer reviews and summarizing it. It can thus be vital to service providers, product producers, moviemakers, allowing them to quickly assess how new products and features are being received.

This graph is an example for a sentiment word cloud

Objectives

Sentiment analysis involves cross-fields backgrounds. It is one type of data science. We can use it on several different fields such as movie industry, electronic product industry and some services industry.

- ✓ Learn the conception of sentiment analysis. Doing a good sentiment analysis need to prepare well in a lot fields. Dealing with the natural language is very hard. First is the text normalization. Then it should be the whitespace tokenizer. Third is the Bayes classifier.
- ✓ Gain real understanding about programming in R on data collection and implement sentiment analysis. By utilizing statistical software R to retrieve data from html or xml format.
- ✓ The objective, while working on this independent study, will be to understand how to conduct research, do research survey and think through the problems and figuring out the path to the best result. Through this project I am also going to learn some Machine Learning techniques and practices.

Methods to implement sentiment analysis

- ➢ First step is collection data from amazon. I used the statistical software, R, to write down the function, which can automatically retrieve customer reviews and rates from Amazon. I only need to input the production ID in amazon and the function will output the customer reviews and rates in a csv format file.

```r
data collection function.R
####################### Function for collecting reviews
datacollection <- function(ProdID){
  library(XML)
  url <- paste0("http://www.amazon.com/product-reviews/",ProdID)
  url <- paste0(url,"/ref=cm_cr_pr_top_link_2?ie=UTF8&pageNumber=")
  url <- paste0(url,"1")
  doc <- htmlParse(url)
  numreview <- as.numeric(gsub(",","",strsplit(xpathSApply(doc,'//b',xmlValue)[1]," ")
  numpage <- numreview%/%10 + 1
  url <- gsub("1","",url)
  url <- paste0(url,1:numpage)
  reviews <- list()
  for(i in 1:numpage){
    doc<-htmlParse(url[i])
    reviews[i]<-list(xpathSApply(doc,'//div[@class="reviewText"]',xmlValue))
    finalreviews<-unlist(reviews)
  }
  rates<-list()
  for(i in 1:numpage){
      doc<-htmlParse(url[i])
      rates[i]<-list(xpathSApply(doc,'//span[@class="swSprite s_star_5_0 "]|//span[@cl
    }
  numericrates<-vector()
  for (i in 1:numreview){
      numericrates[i]<-as.numeric(strsplit(unlist(rates)," ")[[i]][1])
    }
  dfreview<-data.frame(ReviewID=c(1:length(finalreviews)),customer_reviews=finalreview
  write.csv(dfreview,ProdID)
}
```

- ➢ Second is going to find the list of words, which contains both positive words and negative words. Then, the next step is to tokenization. Tokenizing (splitting a string into its desired constituent parts) is fundamental to all Neuro-linguistic programming tasks. It is very complex if you want to do it extremely accurate, because sometime it is unusually represented a single cluster of punctuation like :-(might already represent the whole opinion. Hence, I did in a sample way. I use the function strsplit in R in order to split every word with a space. For example, "I love ipad. It is awesome." By applying the function, we get eight units.

	words	scores	word
1	abandon	-2	abandon
2	abandoned	-2	abandoned
3	abandons	-2	abandons
4	abducted	-2	abducted
5	abduction	-2	abduction
6	abductions	-2	abductions
7	abhor	-3	abhor
8	abhorred	-3	abhorred
9	abhorrent	-3	abhorrent
10	abhors	-3	abhors
11	abilities	2	abilities

```r
> strsplit(a[2]," ")
[[1]]
 [1] "Exactly"    "what"      "shipper"
 [7] "The"        "cord"      "connecting"
[13] "is"         "crimped"   "so"
[19] "to"         "attach"    "but"
[25] "connected." "I"         "won't"
[31] "if"         "at"        "all"
```

- ➢ Then we are moving to the next step. I try to match the words from word list with words in the reviews. Here, I use a Bayesian classifier, which means one positive match counts one point, and one negative match counts negative one point.

```r
scores = laply(sentences, function(sentence, pos.words, neg.words) {

  # clean up sentences with R's regex-driven global substitute, gsub():
  sentence = gsub('[[:punct:]]', '', sentence)
  sentence = gsub('[[:cntrl:]]', '', sentence)
  sentence = gsub('\\d+', '', sentence)
  # and convert to lower case:
  sentence = tolower(sentence)

  # split into words. str_split is in the stringr package
  word.list = str_split(sentence, '\\s+')
  # sometimes a list() is one level of hierarchy too much
  words = unlist(word.list)
```

- ➢ Split all the sentence into single, lower letter words. This is a preparation for the further sentiment analysis.

```r
# compare our words to the dictionaries of positive & negative terms
pos.matches = match(words, pos.words)
neg.matches = match(words, neg.words)

# match() returns the position of the matched term or NA
# we just want a TRUE/FALSE:
pos.matches = !is.na(pos.matches)
neg.matches = !is.na(neg.matches)

# and conveniently enough, TRUE/FALSE will be treated as 1/0 by sum():
score = sum(pos.matches) - sum(neg.matches)
```

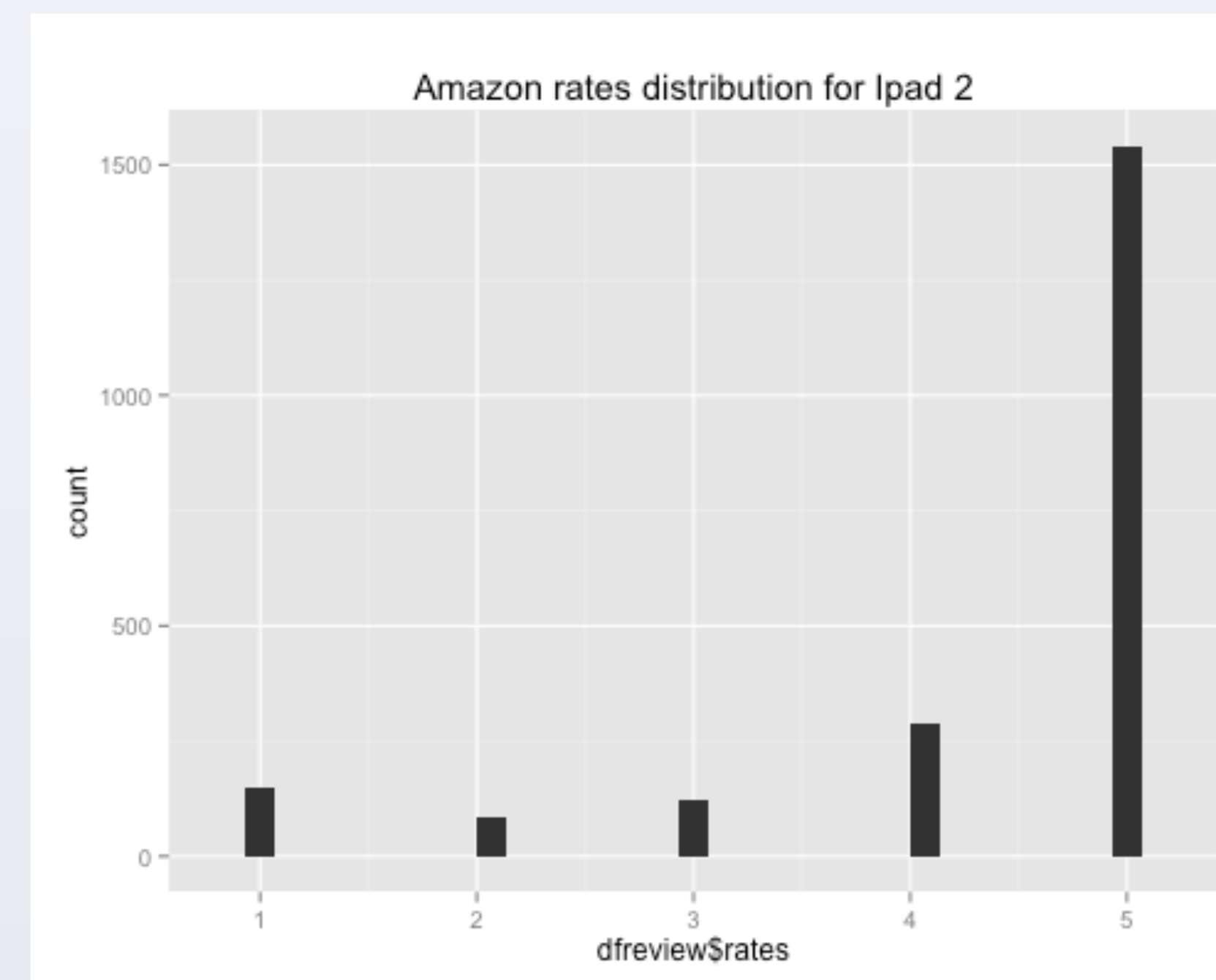
- ➢ Based on the method above, I can get a sentiment score finally.

Results

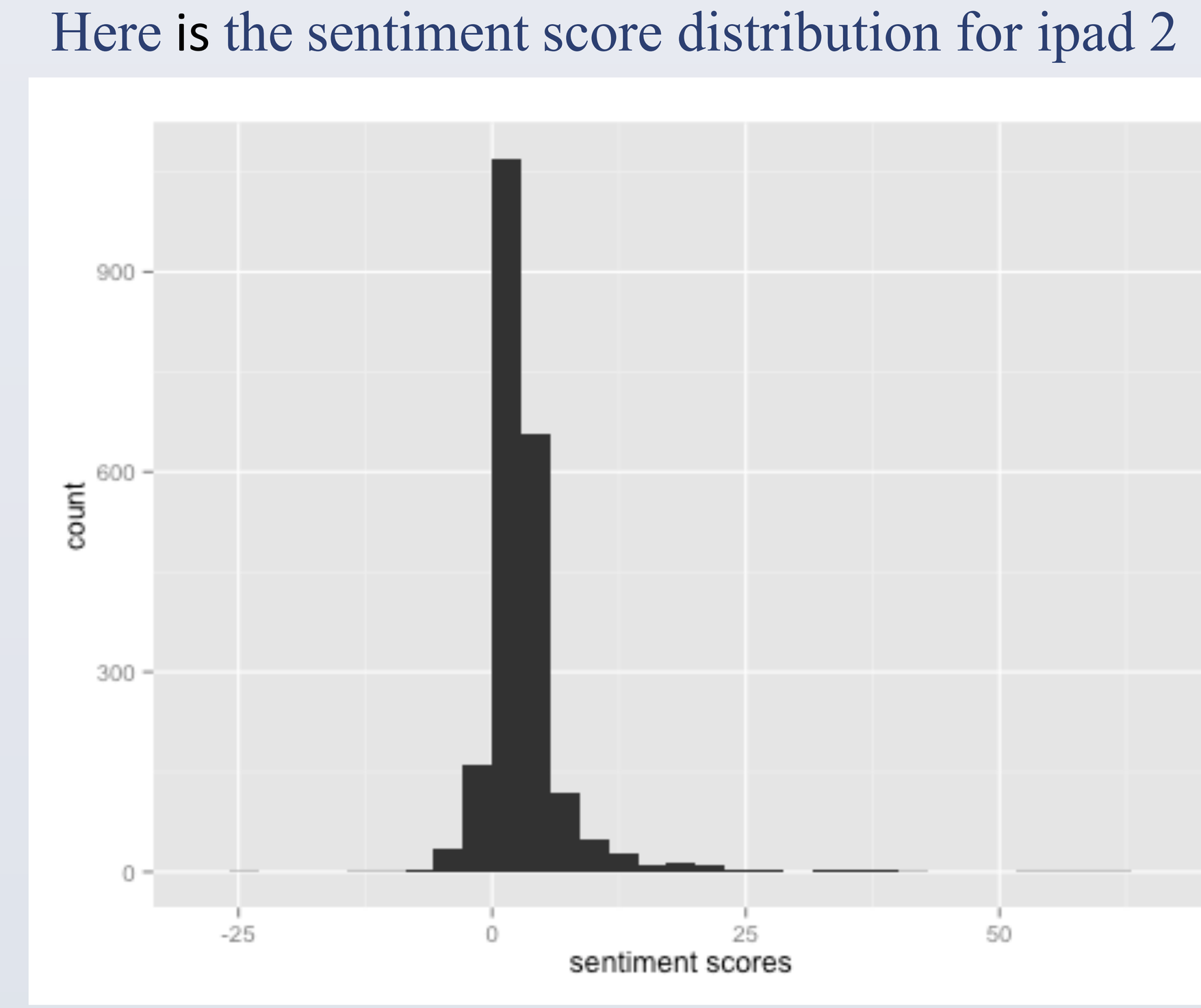
This data frame contains the review ID, amazon star rate, sentiment score, number of positive words, number of negative words, length of reviews and the context of customer reviews

	ReviewID	rates	sentiment_scores	numofpos	numofneg	length_of_words	customer_reviews
1	1	5	0	2	2	47	I was recently reminded how cool i
2	2	4	-1	0	1	34	Exactly what shipper described. Pr
3	3	5	-1	0	1	20	I bought this for my husband for h
4	4	3	-2	1	3	27	When I received the item, it was d
5	5	5	5	5	0	20	good product woks very well and it
6	6	5	4	4	0	53	Nextworth was very prompt and gave
7	7	5	4	5	1	46	Love love love! Came super fast, a
8	8	4	3	3	0	36	This unit was in great shape, it l
9	9	1	-1	0	1	57	this item its used!!!! it was alre
10	10	5	3	3	0	62	The IPad that I purchased was not
11	11	5	4	4	0	23	Love its flexibility and portabili
12	12	5	5	5	0	61	Bought used iPad 2 for family use.
13	13	5	0	1	1	32	Got the 64GB refurbished model for

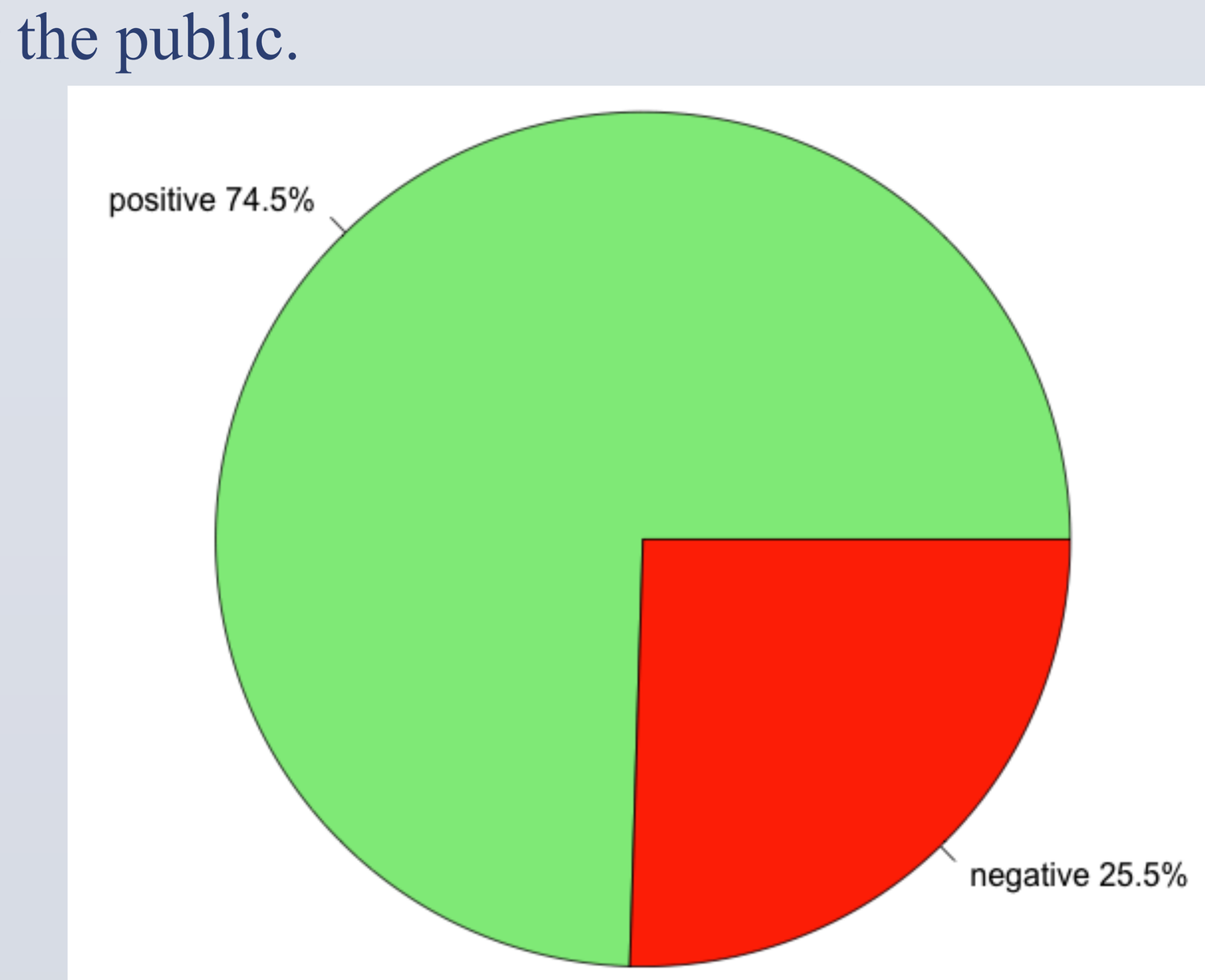
Here is the rate distribution for ipad2



Here is the sentiment score distribution for ipad 2

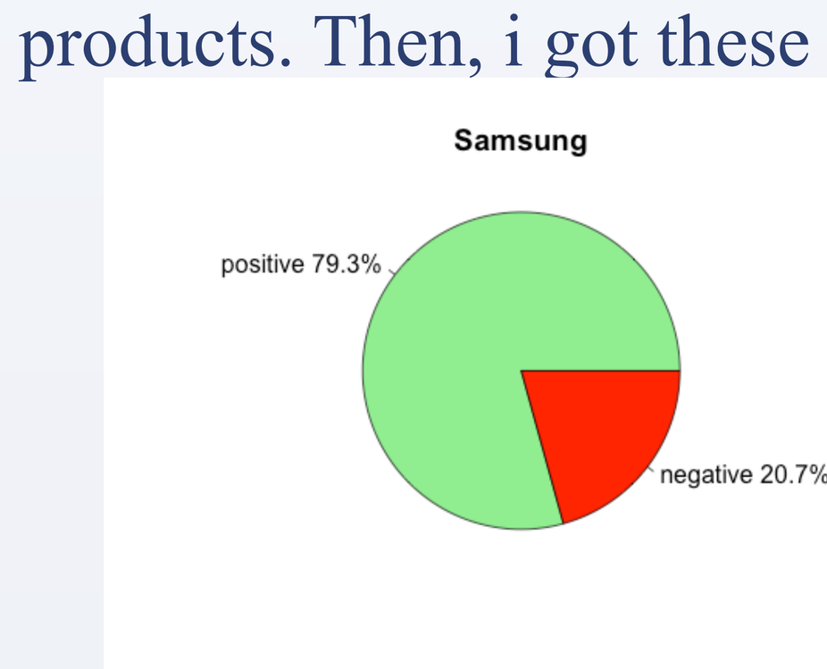


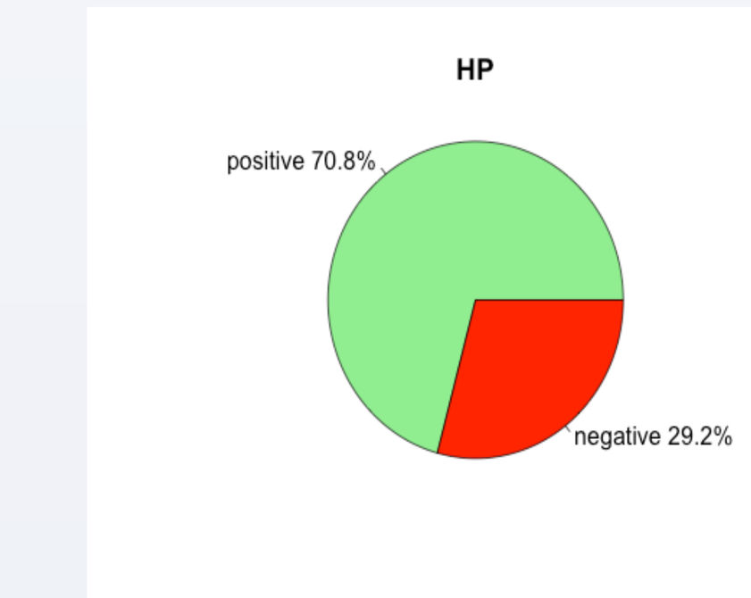
By comparing these two plots, we can see that the distribution of sentiment scores follows the distribution of amazon rates. As ipad2 has plenty of rates with 5 stars, the sentiment scores are also skew to left, which means that most of the sentiments in the reviews are positive. It implies that ipad2 is popular among the public.



Conclusions

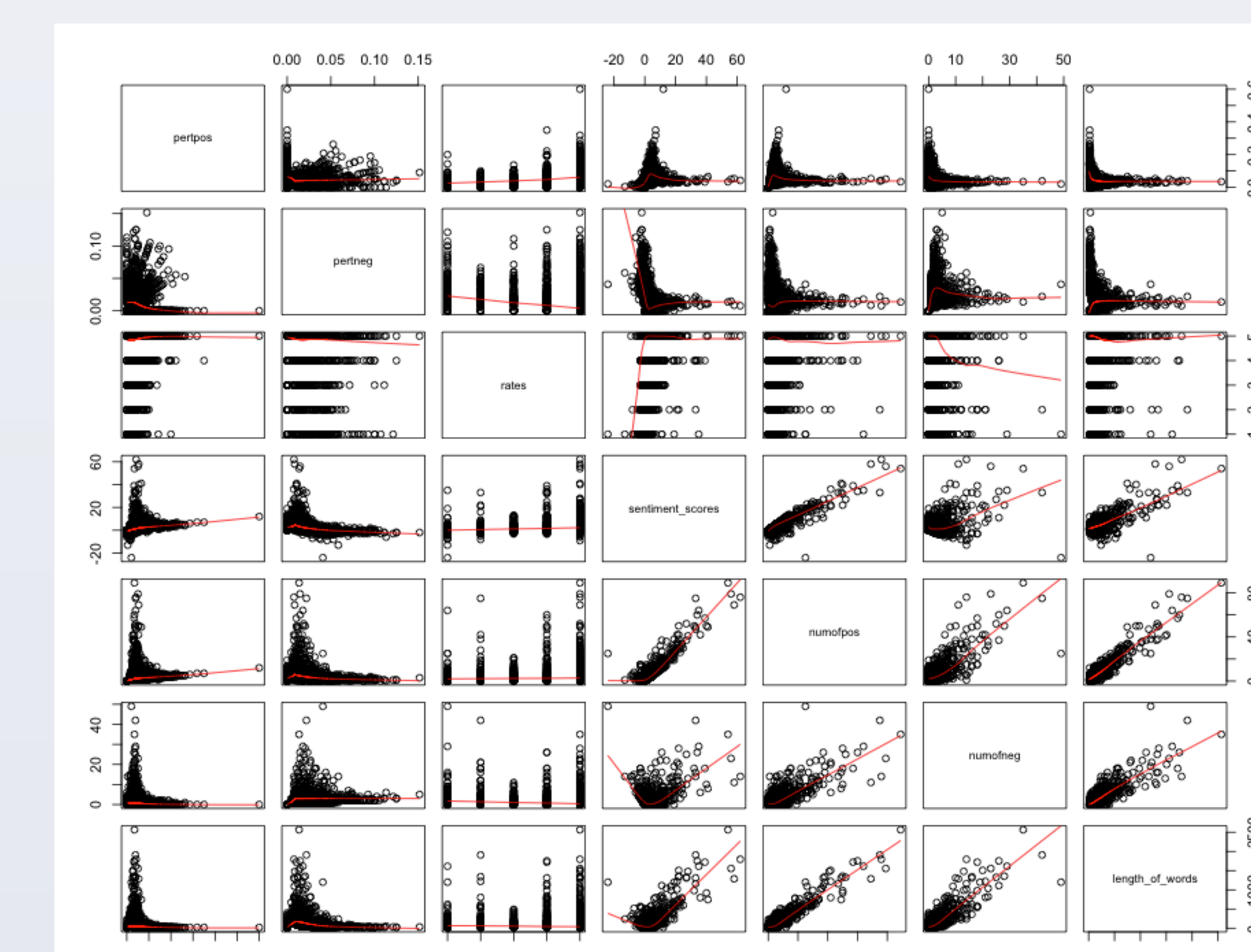
From the results, we can conclude that ipad 2 has a very positive customer reviews. But I still want to know how popular it is. Thus, I decided to compare it with its competitors, Samsung Galaxy Tab 10 and HP Touchpad. I did the same steps for the other two products. Then, i got these pie plots below.



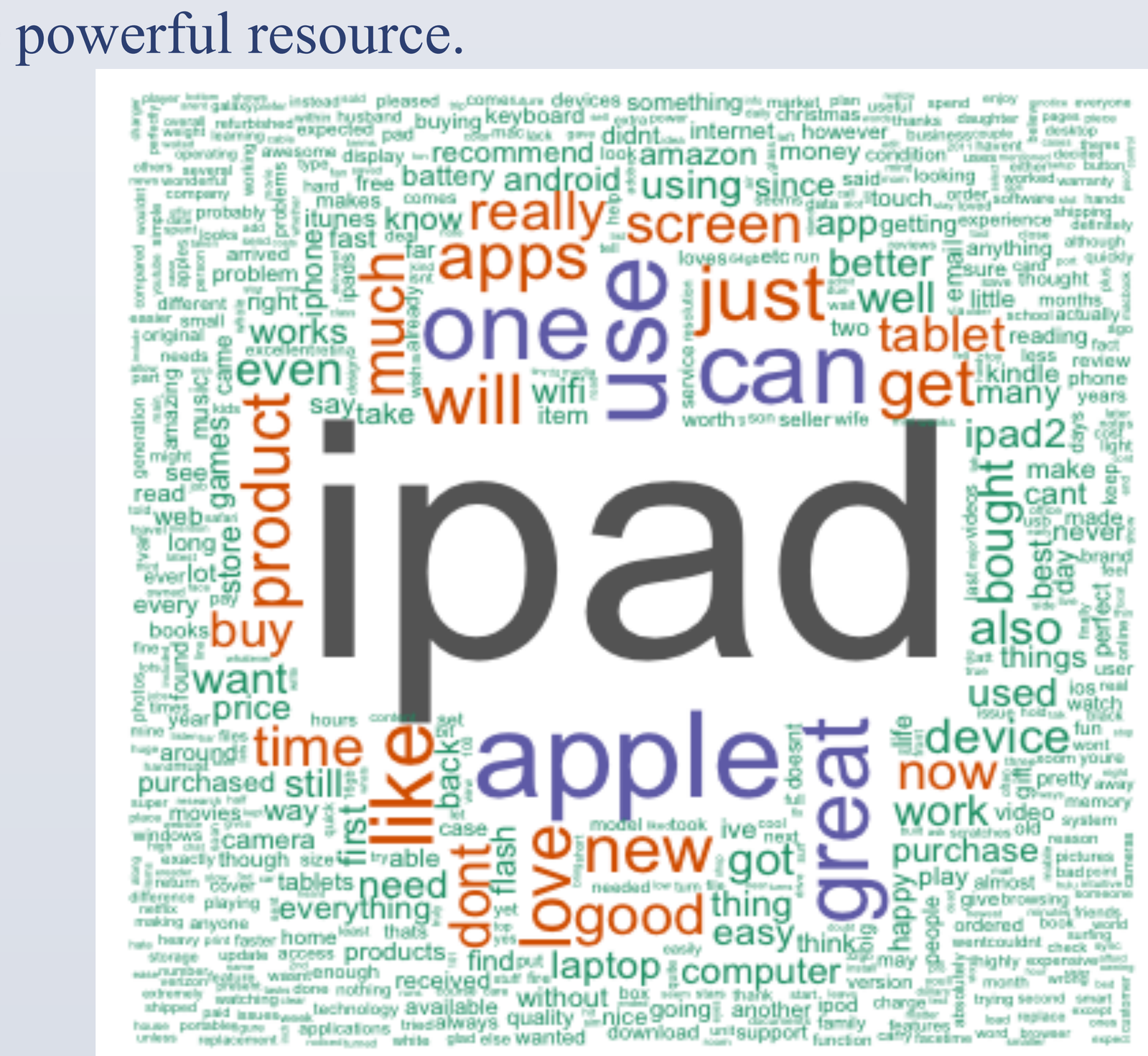


From the pie plots, we can see that the satisfaction of samsung > ipad > hp pad. By applying the sentiment analysis, though, it is not an accurate conclusion, but we could say that samsung tab 10 has the highest satisfaction rate among these three products.

Due to the limitation of the poster, I just post part of my further idea about the regression with the data of ipad 2.



We can use the sentiment analysis to check whether the customers satisfy with companies' products, instead of reviewing the rates from Amazon or other online shopping websites. By implementing the sentiment analysis with only the context of customer reviews, we can figure out the feedbacks from customers. Thus, for further usage, we can analysis the data from tweet, which represents a more powerful resource.



This is the word cloud for the customer reviews of ipad2 from Amazon

References

- Potts, Christopher. "Sentiment Symposium Tutorial." Sentiment Symposium Tutorial.

Acknowledgments

I want to appreciate Professor. Aldous for providing this opportunity to me to show my ideas about sentiment analysis and always providing encouraging guidance.