Project Report on Review Based Recommendation System

Submitted By:
Pooja Bavishi (1421008)
Riddhi Patel(1421013)
Saurabh Chauhan(1421015)

Guided By:
Prof. Sanjay Chaudhary

Department of Computer Science [M.Tech (CSE)]

BIG DATA ANALYTICS

Institute of Engineering & Technology

Ahmedabad University

May 13, 2015

Contents

List of Figures

1 Project Definition

Mobile phones are becoming a primary platform for information access. More and more people use these communication and information access tools, and the functionalities and the challenges provided by these devices are growing.Hence it is important to understand the capabilities of this channel and the information access behavior of mobile users. Recommender Systems (RS) are information filtering and decision support tools aimed at addressing these problems, providing product and service recommendations personalized to the user's needs and preferences at each particular request.

Consumer reviews, opinions and shared experiences in the use of a product is a powerful source of information about consumer preferences that can be used in recommender systems. Despite the importance and value of such information, there is no comprehensive mechanism that formalizes the opinions selection and retrieval process and the utilization of retrieved opinions due to the difficulty of extracting information from text data.

Product Ratings: Product ratings are usually visualized on a scale of one-to-five stars. They allow users to get an at-a glance assessment of a product. Ratings can be made based on specific predefined criteria (e.g. price performance ratio and quality) or as an expression of the overall satisfaction with a product.

Product Reviews: Product reviews allow users to describe their experience with products as continuous text. In this context, different levels of details can be allowed. At Opinions, for example, users can submit short reviews (up to 100 words) and regular reviews.

2 Logical Diagram

2.1 Architecture

2.1.1 Abstract Architecture

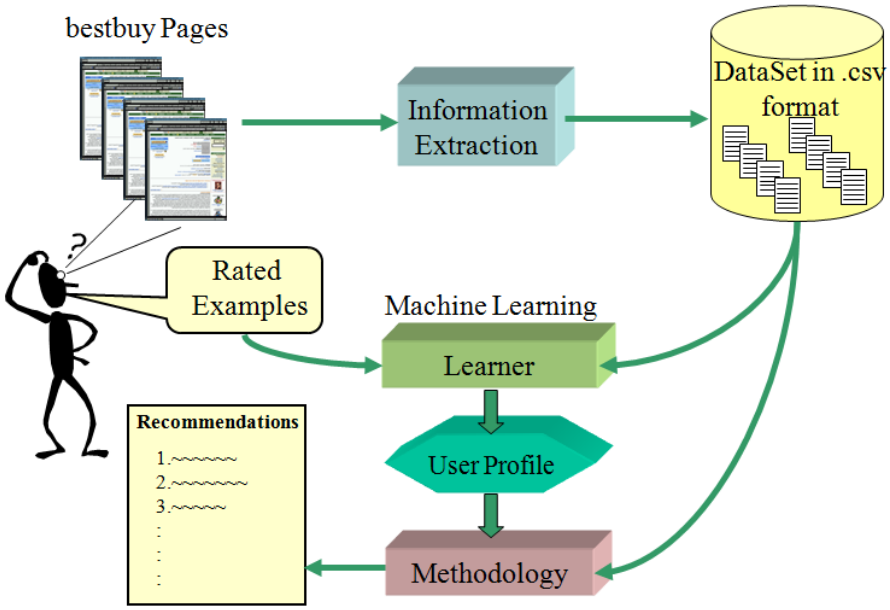


Figure 1: Abstract Architecture of Proposed System

2.1.2 Concrete Architecture of Our Proposed System

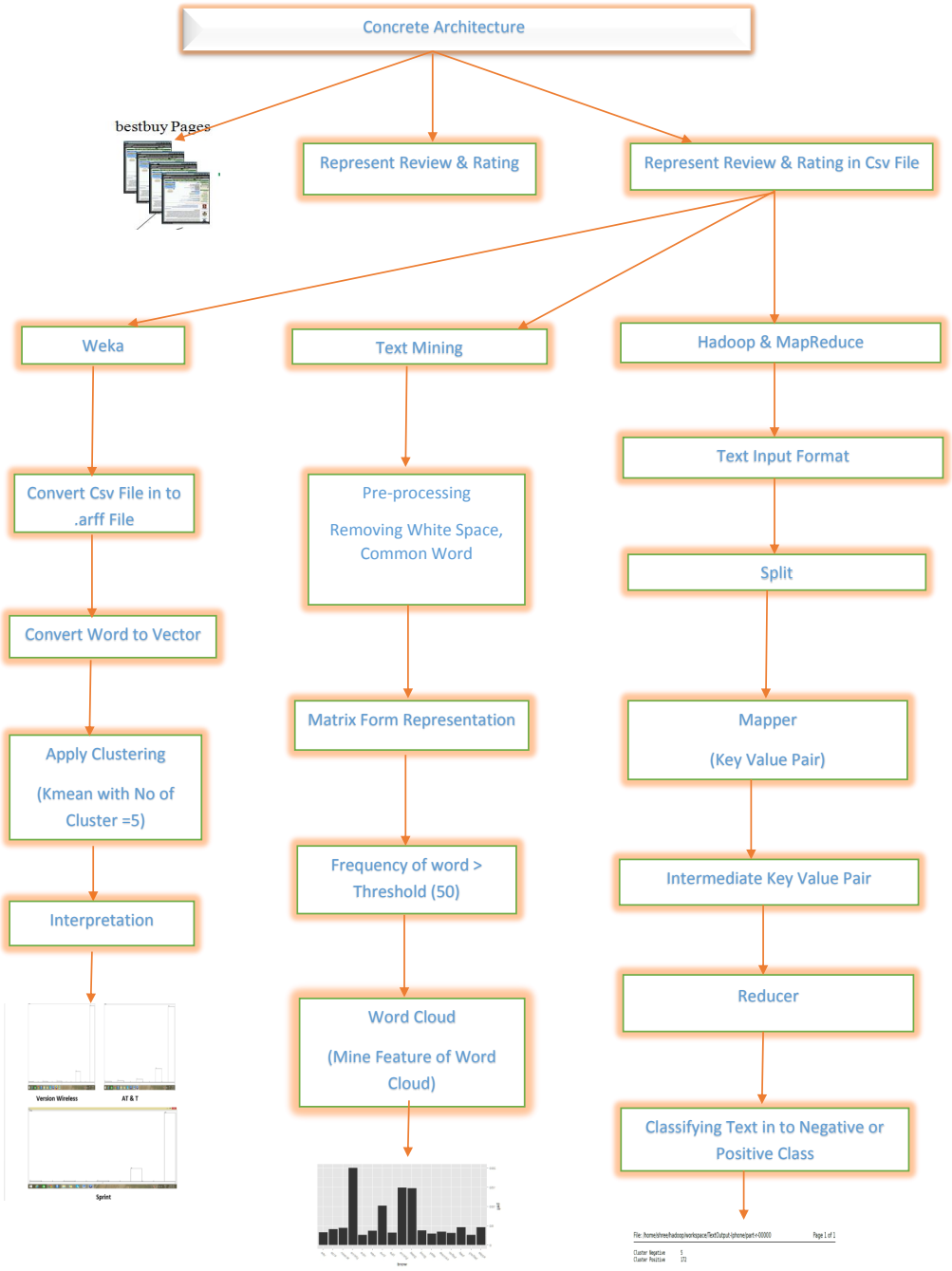


Figure 2: Concrete Architecture of Proposed System

2.2 Flowchart/Algorithm

Data analysis is a process of obtaining raw data and converting it into information useful for decision-making by users.Data is collected from a website http://www.bestbuy.com/.Data is in form of customers review and rating given in scale of 1 to 5 stars. Data initially obtained must be processed or organized for analysis. For instance, these may involve placing data into rows and columns in a table format for further analysis, such as within a spreadsheet.

The need for data cleaning will arise from problems in the way that data is entered and stored. Data cleaning is the process of preventing and correcting these errors. Common tasks include record matching, avoid duplication, and column segmentation.

After the text pre-processing is done, some machine learning technique needs to applied for extracting the information and discovering the knowledge.In clustering we have clusters on bases of rating which is given by users,Service provider of phone,colors of the phone,etc. In Features extraction we extracted the features which has frequency greater than 50 from the review which is given by users and formed word cloud.In classification we have divide the user's review into the Negative and Positive Class.

From results of above techniques we recommend to users which phone they should buy or which phone is better.

The Data Flow Diagram of our proposed system is shown as below.

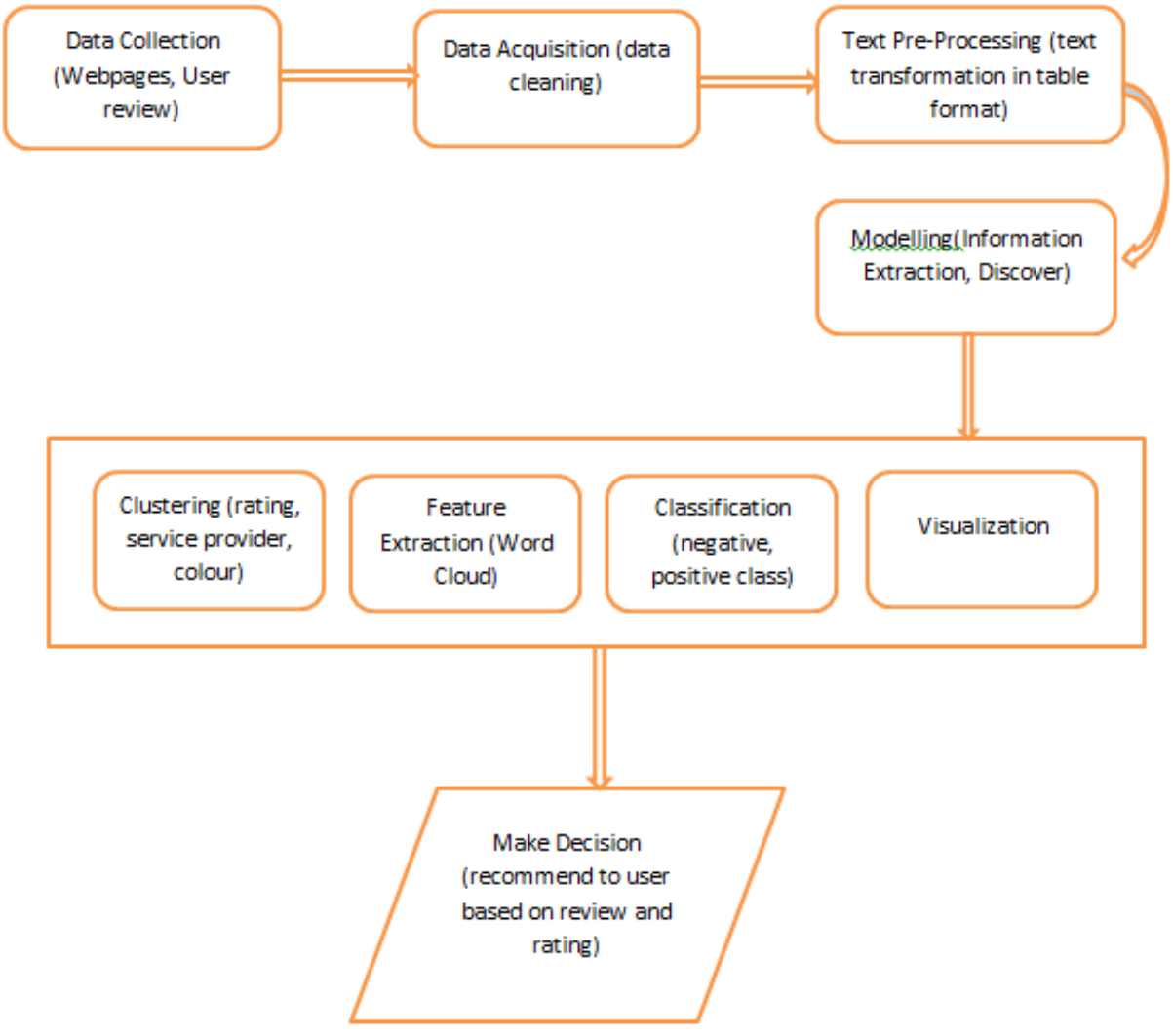


Figure 3: Flow Chart of Proposed System

2.3 Schema /Structure of DB

We have collected data from the website http://www.bestbuy.com.Following is the snapshot of sample dataset for iPhone 6.As shown in figure below data is unstructured format so we need to convert data into semi structured format.

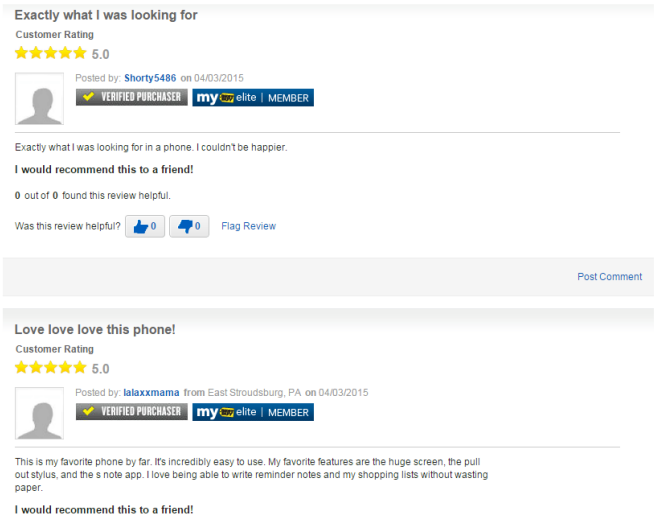
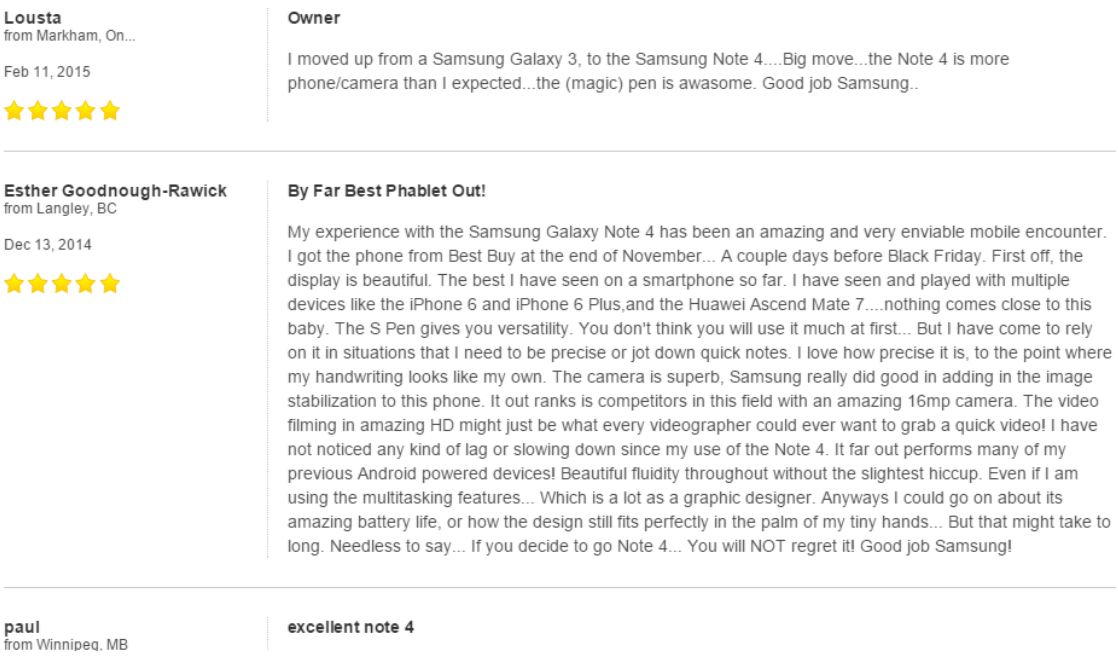


Figure 4: iPhone 6 Data Source Sample



Lousta
from Markham, On...

Feb 11, 2015

Owner

I moved up from a Samsung Galaxy 3, to the Samsung Note 4....Big move...the Note 4 is more phone/camera than I expected...the (magic) pen is awasome. Good job Samsung..

Esther Goodnough-Rawick
from Langley, BC

Dec 13, 2014

By Far Best Phablet Out!

My experience with the Samsung Galaxy Note 4 has been an amazing and very enviable mobile encounter. I got the phone from Best Buy at the end of November... A couple days before Black Friday. First off, the display is beautiful. The best I have seen on a smartphone so far. I have seen and played with multiple devices like the iPhone 6 and iPhone 6 Plus,and the Huawei Ascend Mate 7....nothing comes close to this baby. The S Pen gives you versatility. You don't think you will use it much at first... But I have come to rely on it in situations that I need to be precise or jot down quick notes. I love how precise it is, to the point where my handwriting looks like my own. The camera is superb, Samsung really did good in adding in the image stabilization to this phone. It out ranks is competitors in this field with an amazing 16mp camera. The video filming in amazing HD might just be what every videographer could ever want to grab a quick video! I have not noticed any kind of lag or slowing down since my use of the Note 4. It far out performs many of my previous Android powered devices! Beautiful fluidity throughout without the slightest hiccup. Even if I am using the multitasking features... Which is a lot as a graphic designer. Anyways I could go on about its amazing battery life, or how the design still fits perfectly in the palm of my tiny hands... But that might take to long. Needless to say... If you decide to go Note 4... You will NOT regret it! Good job Samsung!

paul
from Winnipeg, MB

excellent note 4

Figure 5: Samsung Galaxy Note 4 Data Source Sample

From http://www.bestbuy.com website we have collected around 250 reviews and ratings of iPhone 6 and Samsung Galaxy Note 4 by Service Provider wise,Phone Colour wise and memory capability wise .All data are in unstructured format and so we need to convert into the semi structure format.

	A	B	C	D	E	F	G
1	Phone	Memory	service provi	Rating	Color	Date	Review
2	iPhone 6	16 GB	Sprint	4	Gold	3/14/2015	I have been using iPhones for a while and this is not their best. It is a very good phone but does feel as sturdy as the iPhone 5 or iPhone 4. The
3	iPhone 6	16 GB	Sprint	5	Gold	03-12-2015	The iPhone 6 is everything that was promised to Apple customers. I suggest this phone to upgraders and everyone else!
4	iPhone 6	16 GB	Sprint	4	Gold	03-10-2015	It's a great phone but a little too big for me faster then the 5 series, I'd prefer to have something smaller so I wont have to out it in my back pc
5	iPhone 6	16 GB	Sprint	5	Gold	ÿ03/10/2015	love it my new iPhone!!!!!!!!!!!,best purchase ever
6	iPhone 6	16 GB	Sprint	5	Gold	ÿ03/09/2015	I like the Gold iPhone 6 it's cool not too big and a lot faster then the older models! No problems with the phone do far!
7	iPhone 6	16 GB	Sprint	5	Gold	03-09-2015	I like the Silver iPhone 6 it's pretty good not too big and a lot faster then the older models!
8	iPhone 6	16 GB	Sprint	5	Gold	03-09-2015	Love the color first and for most. I have had every iPhone made in my possession by far I like this the best. Its very light weight and not to bul
9	iPhone 6	16 GB	Sprint	4	Gold	03-09-2015	Great Battery life.. Battery last for more than 2 day!
10	iPhone 6	16 GB	Sprint	5	Gold	03-07-2015	It was an iPhone purchase purchased online and picked up at store
11	iPhone 6	16 GB	Sprint	5	Gold	03-07-2015	I swore I would never buy another Apple product, but the sale and sprint plan was too good to pass up. We bought three! We are very happy
12	iPhone 6	16 GB	Sprint	5	Gold	ÿ03/06/2015	I wouldn't have purchased it if I didn't think it would live up to the high Apple standards I have come to expect. I haven't been disappointed.
13	iPhone 6	16 GB	Sprint	5	Gold	ÿ03/05/2015	Awesome phone, like every other apple product, this did not disappoint!
14	iPhone 6	16 GB	Sprint	5	Gold	03-05-2015	I bought it for my daughter and she just loves it.
15	iPhone 6	16 GB	Sprint	4	Gold	03-05-2015	We had a few minor issues at first, but once Apple last update/ fix everything is great.
16	iPhone 6	16 GB	Sprint	4	Gold	2/27/2015	I upgraded from an iPhone 5 to the iPhone 6. Only had it for 2 days, but so far so good. Pretty simple switch to the newest version iPhone. On
17	iPhone 6	16 GB	Sprint	5	Gold	2/25/2015	Great phone, looks great and lots of features, love it.
18	iPhone 6	16 GB	Sprint	5	Gold	2/25/2015	The larger screen size is a plus and mines is not the plus, but this phone rings louder and sounds clearer when you are speaking to someone.
19	iPhone 6	16 GB	Sprint	5	Gold	2/19/2015	Great Picture, awesome bigger size, longer battery life (much better than previous versions)
20	iPhone 6	16 GB	Sprint	5	Gold	2/19/2015	I have always been a apple customer (wouldn't have it any other way) But this phone is GREAT! so much bigger and slimmer than the 5s. I real
21	iPhone 6	16 GB	Sprint	5	Gold	2/19/2015	I have had an i-phone ever since it was introduced. This is faster, has a better camera and meets all my needs.
22	iPhone 6	16 GB	Sprint	5	Gold	2/19/2015	This phone takes it to the next level. The larger screen and streamlined profile are great, and the quality and access are as good as I expect ar
23	iPhone 6	16 GB	Sprint	5	Gold	2/19/2015	I am happy with the iphone6. I love the new features and the size.
24	iPhone 6	16 GB	Sprint	5	Gold	2/18/2015	I purchased the new IPhone 6 for my 15 yr old daughter. She absolutely love it.
25	iPhone 6	16 GB	Sprint	5	Gold	2/18/2015	Excellent, definitely recommend to everyone, perfect size and style.

Figure 6: iPhone 6 Data Sample

3 Implementation

3.1 Setup

- Software Configuration:
  - Data Mining Tool (Weka)
  - R
  - Ecllipse (Hadoop)
  - Notepad++
- Hardware Configuration
  - Intel(R) Core(TM) i5-2430M CPU @ 2.40Ghz
  - RAM 8.00 GB
  - OS 64 bit
- Methodology
  - Clustering
  - Text Mining
  - Map Reduce Programming

3.2 Source code

3.2.1 Text mining In R Source code

```
cname <- file.path("C:/Users/admin/Desktop/BDA_Project", "texts")
cname
dir(cname)
library(tm)
docs <- Corpus(DirSource(cname))
## Preprocessing
docs <- tm_map(docs, removePunctuation)   # Removing punctuation
docs <- tm_map(docs, removeNumbers)      # Removing numbers
docs <- tm_map(docs, tolower)   # Converting to lowercase
docs <- tm_map(docs, removeWords, stopwords("english"))   # Removing "stopwords"
library(SnowballC)
docs <- tm_map(docs, stemDocument)   # Removing common word endings (e.g., "ing", "es")
docs <- tm_map(docs, stripWhitespace)   #Stripping whitespace
docs <- tm_map(docs, PlainTextDocument)
## This is the end of the preprocessing stage.
### Stage the Data
dtm <- DocumentTermMatrix(docs)
tdm <- TermDocumentMatrix(docs)

### Explore  data
freq <- colSums(as.matrix(dtm))
length(freq)
ord <- order(freq)
m <- as.matrix(dtm)
dim(m)
write.csv(m, file="DocumentTermMatrix.csv")
#  Start by removing sparse terms
dtms <- removeSparseTerms(dtm, 0.1) # This makes a matrix that is 10% empty space, maximum.
### Word Frequency
head(table(freq), 20)
# The above output is two rows of numbers. The top number is the frequency with which
# words appear and the bottom number reflects how many words appear that frequently.

tail(table(freq), 20)
# Considering only the 20 greatest frequencies
#View a table of the terms after removing sparse terms, as above.
freq <- colSums(as.matrix(dtms))
```

```
freq
# This will identify all terms that appear frequently (in this case, 50 or more times).
findFreqTerms(dtm, lowfreq=50)   # Change "50" to whatever is most appropriate for your data.
#Plot words that appear at least 50 times.
library(ggplot2)
wf <- data.frame(word=names(freq), freq=freq)
p <- ggplot(subset(wf, freq>25), aes(word, freq))
p <- p + geom_bar(stat="identity")
p <- p + theme(axis.text.x=element_text(angle=45, hjust=1))
p
# Term Correlations.If words always appear together, then correlation=1.0.
findAssocs(dtm, c("question" , "analysi"), corlimit=0.98) # specifying a correlation limit of 0.98
### Word Clouds
#load the package word clouds.
library(wordcloud)
dtms <- removeSparseTerms(dtm, 0.15) # Prepare the data (max 15% empty space)
freq <- colSums(as.matrix(dtm)) # Find word frequencies
dark2 <- brewer.pal(6, "Dark2")
wordcloud(names(freq), freq, max.words=100, rot.per=0.2, colors=dark2)
```

3.2.2 Text Calssifying Source code in Map-Reduce Programming

1. Map-Reduce Driver File

```
package com.textminer;

import org.apache.hadoop.conf.Configuration;
import org.apache.hadoop.conf.Configured;
import org.apache.hadoop.fs.Path;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Job;
import org.apache.hadoop.mapreduce.lib.input.FileInputFormat;
import org.apache.hadoop.mapreduce.lib.input.TextInputFormat;
import org.apache.hadoop.mapreduce.lib.output.FileOutputFormat;
import org.apache.hadoop.mapreduce.lib.output.TextOutputFormat;
import org.apache.hadoop.util.Tool;
import org.apache.hadoop.util.ToolRunner;

public class TextDriver extends Configured implements Tool{
public static void main(String[] args) throws Exception {
int res = ToolRunner.run(new Configuration(), new TextDriver(), args);
System.exit(res);
}
@Override
public int run(String[] args) throws Exception {

// JobConf represents a MapReduce job configuration.
//JobConf is the primary interface for a user to describe a MapReduce job to the Hadoop
//framework for execution. The framework tries to faithfully execute the job as described by JobCo
Configuration conf = new Configuration();
Job job = new Job(conf, "Votcount");
//configure the input output key and value class
job.setOutputKeyClass(Text.class);
job.setOutputValueClass(IntWritable.class);
job.setMapperClass(TextMapper.class);
job.setReducerClass(TextReducer.class);
job.setInputFormatClass(TextInputFormat.class);
job.setOutputFormatClass(TextOutputFormat.class);
// add the input and out directory path
//FileInputFormat.addInputPath(job, new Path(("/home/shree/hadoop/workspace/TextInput-Samsung")));
```

```
FileInputFormat.addInputPath(job, new Path("/home/shree/hadoop/workspace/TextInput-Iphone/"));
//FileInputFormat.addInputPath(job, new Path("/home/shree/hadoop/workspace/TextOutput-Samsung/"));
FileOutputFormat.setOutputPath(job, new Path("/home/shree/hadoop/workspace/TextOutput-Iphone/"));
job.waitForCompletion(true);
return 0;
}
}
```

2. Mapper Class File

```
package com.textminer;
import java.io.IOException;
import java.util.StringTokenizer;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;

public class TextMapper extends Mapper<Object, Text, Text, IntWritable> {
private final static IntWritable one = new IntWritable(1);
int count=0;
@Override
public void map(Object key, Text value, Context output) throws IOException,
InterruptedException {
// use tokenizer to split the line into words
String word = new String("");
StringTokenizer itr = new StringTokenizer(value.toString());
System.out.println(value.toString());
// iterate through each word
while (itr.hasMoreTokens()) {
word = itr.nextToken();
// comapare the ech word with our specific word to classify onto positive and negative class and t
if(word.contentEquals("good")||word.contentEquals("perfect")||word.contentEquals("great")||word.co
{
output.write(new Text("Cluster Positive"), one);
}
else if(word.contentEquals("bad")||word.contentEquals("ok")||word.contentEquals("problem")||word.c
output.write(new Text("Cluster Negative"), one);
}
}
}
}
```

3. Reducer Class File

```
package com.textminer;
import java.io.IOException;
import java.util.Iterator;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Reducer;

public class TextReducer extends Reducer<Text, IntWritable, Text, IntWritable> {
@Override
public void reduce(Text key, Iterable<IntWritable> values, Context output)
throws IOException, InterruptedException {
//initialize the sum value with o
int Count = 0;

//iterate through the all the values with respect to key and sum up all of them
for(IntWritable value: values){
Count+= value.get();
}
// calculate the final count value
```

```
Count=(Count/2);
//pass the result to the output directory
output.write(key, new IntWritable(Count));
}
}
```

3.3 Experimental Results

1. Weka Results

   Customers rate phone after purchasing the phone based on functionality,look and feel,memory,service provider.Customers rate the phone in form of integer 1 to 5 stars where 5 star denotes highest rating and 1 star denotes lowest rating.On the basis of rating we have formed cluster like rating 5 star cluster,rating 4 star cluster,service provide wise cluster, color wise cluster ,etc.Result of clustering using the k-means algorithm in weka tool is shown below:

   - Comparison between iPhone 6 and Samsung Galaxy Note 4

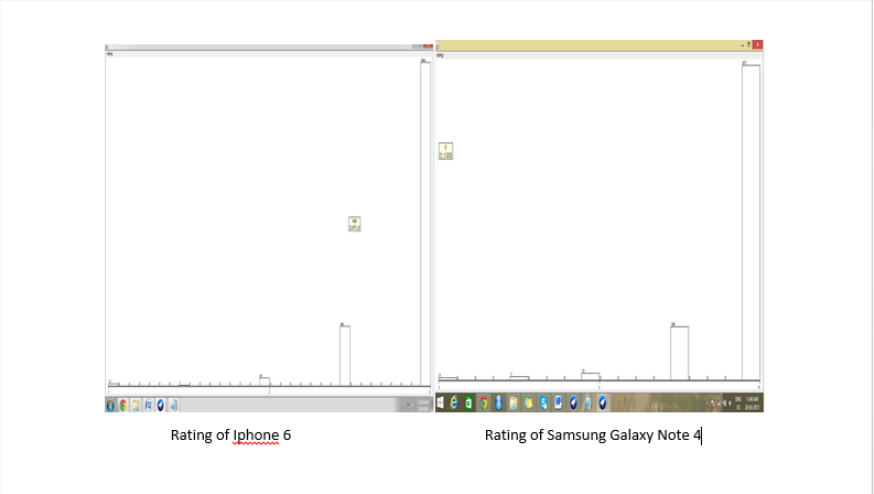


Figure 7: iPhone 6 and Samsung Galaxy Note 4 Rating Clustering

   - Colour wise comparison of iPhone 6

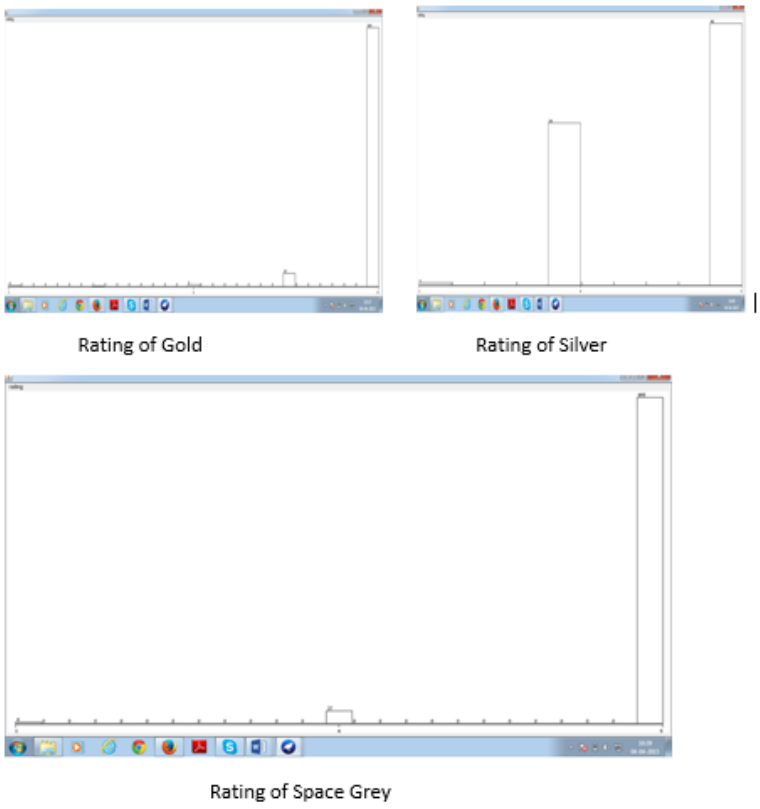


Figure 8: iPhone 6 Colour wise Clustering

- Service Provide Wise Comparison of iPhone 6

Figure 9: iPhone 6 Serive Provider wise Clustering

- Service Provide Wise Comparison of Samsung Galaxy Note 4

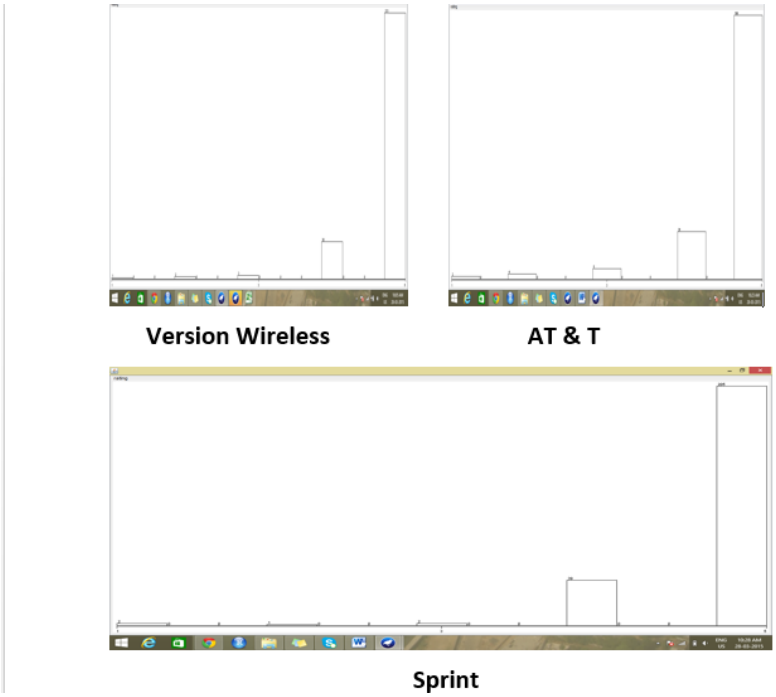


Figure 10: Samsung Galaxy Note 4 Service Provider wise Clustering

2. Text Mining (in R) Results

A text cloud or word cloud is collection of words arrange in random fashion.Word Cloud is visualization of word frequency in a given text as a weighted list.We formed word cloud from the words which has frequency greater that 50 to Extract the Features from the user's review.From the word cloud one can easy mine features , in word cloud the words appear more bigger words indicate that those words appear more frequently in review and small size words indicate that those words appear less frequently in review.

- Features Extraction From iPhone 6 Review (WordCloud)

Figure 11: WordCloud of iPhone6 Review

- Features Extraction From Samsung Galaxy Note 4 Review (WordCloud)

Figure 12: WordCloud of Samsung Galaxy Note 4 Review

- Words Frequency Plot for iPhone 6

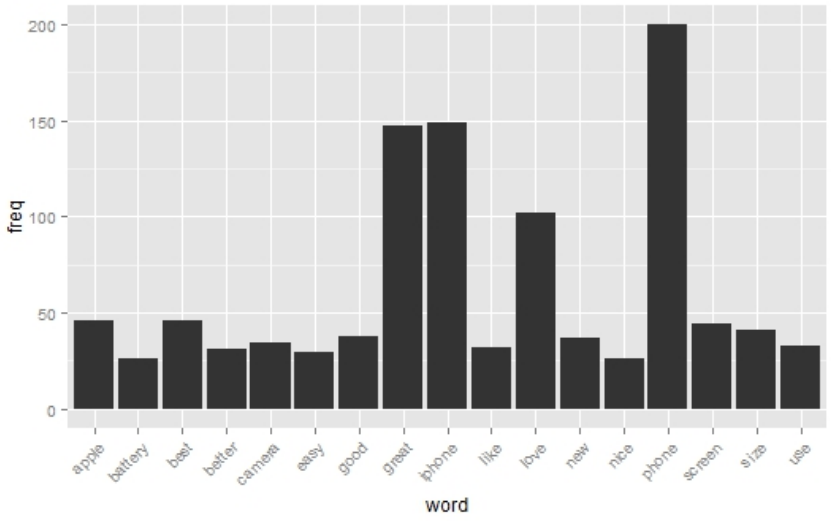


Figure 13: iPhone 6 word frequency plot

- Words Frequency Plot for Samsung Galaxy Note 4

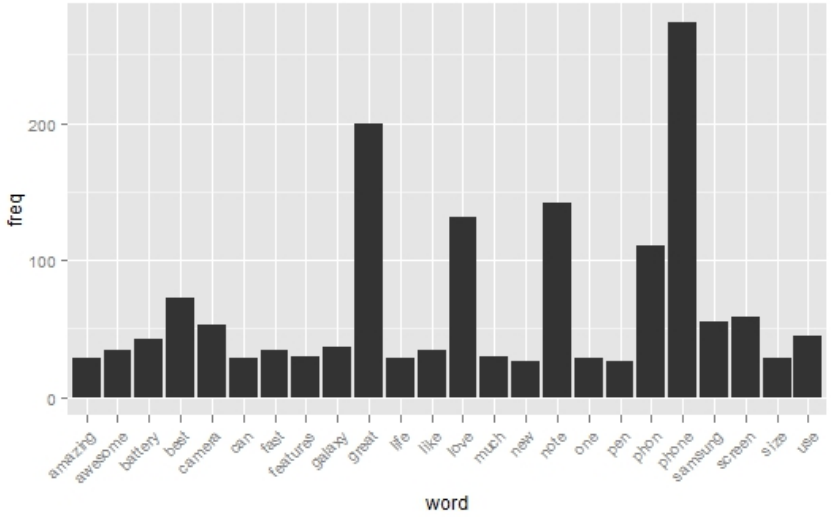


Figure 14: iPhone 6 word frequency plot

3. Map-Reduce Programming Result

Any Review of customer can be categorized into positive or negative class , we can treat it as binary classification problem(Two class classification problem) . For example, the sentence, "I am not excited by this product" expresses a negative sentiment about the product. The degree of the sentiment used is also taken into consideration. For example, "I love this phone" indicates a more positive sentiment than the sentence "I like this product". Apart from regular adjectives like "good", "bad" and "very good", conjunctions like "but", "although", "while" also have a say in the overall polarity of the sentence. Some sentences need not offer any particular opinion. For example, the sentence "I bought this phone two days back" is a subjective opinion on the product. So from this we have classified the reviews into positive and negative class for iPhone 6 and Samsung Glaxy Note 4.Result of this text classification is shown below.

- Text Classification for iPhone 6

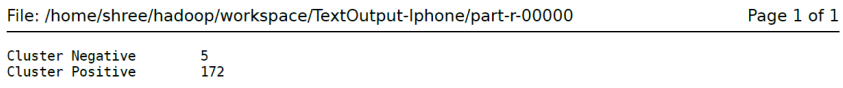


Figure 15: Classification of iPhone 6

- Text Classification for Samsung Galaxy Note 4

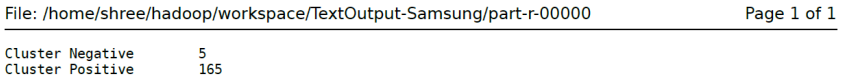


Figure 16: Classification of Samsung Galaxy Note 4

4 Interpretation of Result

After performing experiment, we interpreted the results using the data visualization with the help of google charts.So as per below graphs ,we got the almost same results because limitation of data for the iPhone 6 and Samsung Galaxy Note 4.

4.1 Visualization

- Visualization of iPhone 6 and Samsung Galaxy Note 4 Rating

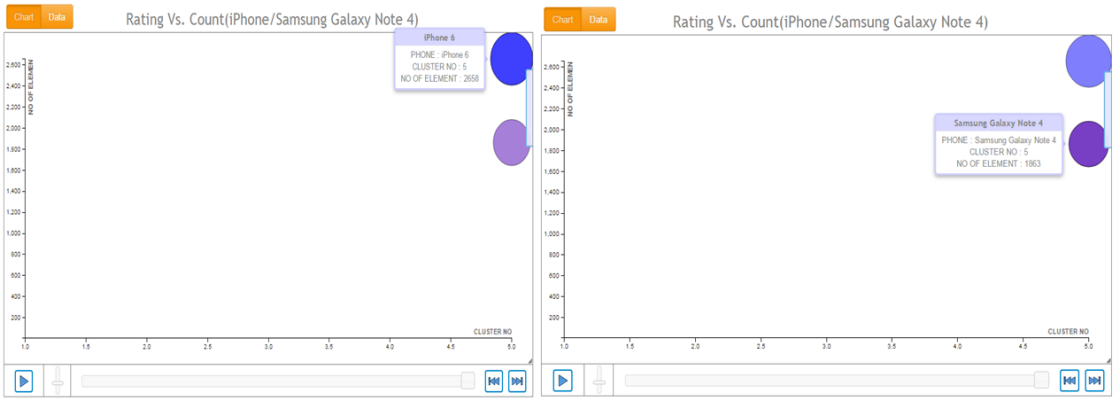


Figure 17: iPhone 6 and Samsung Galaxy Note 4 Rating

- Visualization of iPhone 6 and Samsung Galaxy Note 4 Features

Figure 18: iPhone 6 and Samsung Galaxy Note 4 Features Graph

5 Disadvantages of Our System

- We have limited data(around 250 rating and review for the iPhone 6 and Samsung Galaxy Note 4) so we can expand our horizon to collect more data so that we can provide proper recommendation.
- We are limited to particular web source http://www.bestbuy.com.Actually we should also collect review and rating from the trusted web sources like apple.com and etc.
- We are limited to only 2 product iPhone 6 and Samsung Galaxy Note 4.So we are not able to provide recommendation to user about other products.
- Result of Word Cloud in R is depends on the words frequency.So select those frequency value or threshold is challenging task for us.

6 Conclusion/Future Direction

In this Project we have compare two product iPhone 6 and Samsung Galaxy Note 4 on basis of customer ratings and review. For that we used weka tools,text mining,word cloud,hadoop map-reduce to analyze the customers rating and review,so that we can recommend users about product.

In future we will increase the number of reviews and ratings to provide the accurate recommendation.we will also use multi node clustering for the large number of reviews and ratings and implement generalized version of map-reduce program.

References

[1] G. Linden, B. Smith, and J. York, *"Amazon.com recommendations: item- to-item collaborative filtering"*,Internet Computing 7:1, pp. 76:80, 2003.

[2] For Data Source
http://www.bestbuy.com/

[3] For Weka Tutorials
http://sentimentmining.net/weka/

[4] Text Mining in R (word cloud)
https://sites.google.com/site/miningtwitter/questions/talking-about/wordclouds/wordcloud1.

[5] For Hadoop Map-Reduce Programming

`http://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html#Example\%3A+WordCount+v1.0`