Mubashir Ali
Muhammad Abdullah

Business Understanding

Background:

This dataset is taken from the 'TrustPilot' website.

Business Objectives:

The objective is to find the most common reviews that the users has given to the Apple IPhone Company, the reviews are taken from the recent days.

Data Understanding

Collect Initial Data:

The data provided is consists of Data folder with a corpus of 18 documents included in it. All these documents were taken from 'Trustpilot' website.

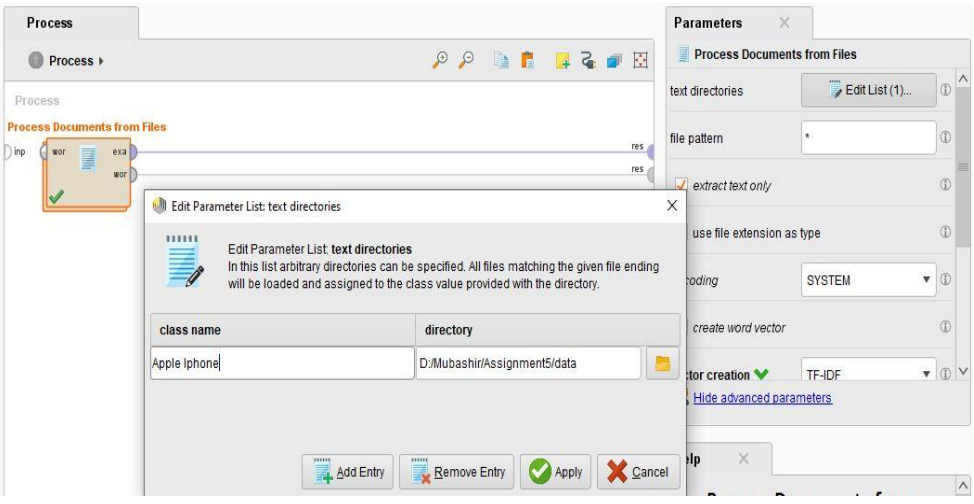
Description of the data:

The folder contains the 18 different documents with every document has a review in it and the respective document was saved by the name of the reviews giver.

Data Preparation

Clean Data

This process was performed using the RapidMiner tool, and was executed as follows: First a process was started by adding the operator 'Process Documents from Files' from the operator. Now in the parameters of that operator we found 'text directories' where we add our corpus of 18 documents, and the name of the class was specified by



'Apple Iphone'. In vector creation we selected 'TF-IDF' this generates word vectors that are stored in the multiple files.

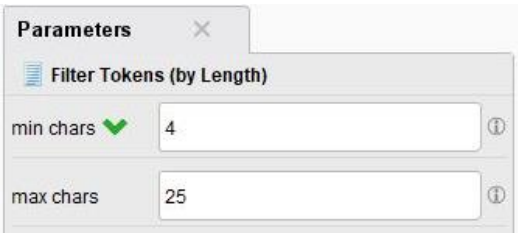
Preprocessing activities:

By double clicking on the operator 'Process Documents from Files' we get into a preprocessing activities.

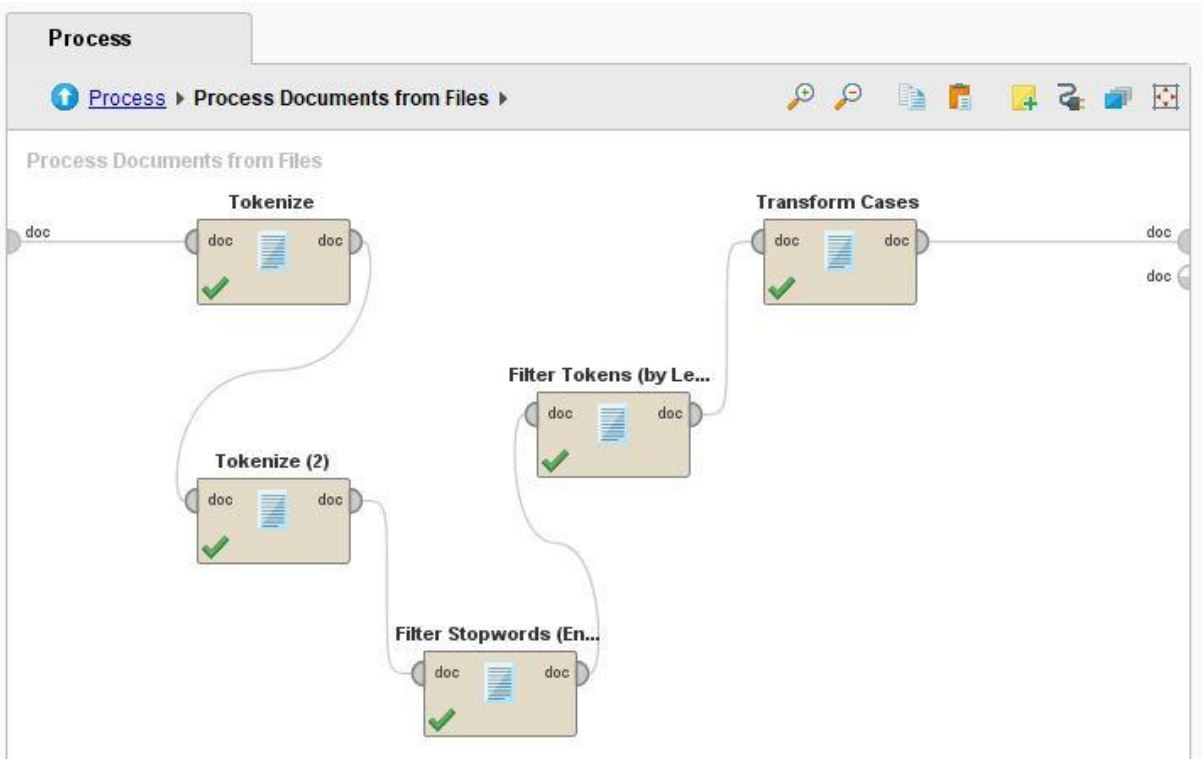
We use 2 tokenize operators in which we used 'non letters' and 'linguistic tokens' respectively.

After that we use 'Filter Stopwords (English)' operator to remove unnecessary conjunctions and English stopwords from our documents.

The next operator used was 'Filter Tokens (by length)', this operator is used to filter the tokens. The minimum and maximum numbers are 4 and 25 respectively.



We use 'Transform cases' operator to address the Case Sensitivity of our corpus.



This above figure shows the complete process of the 'Process Documents from Files'

After running the model we have a following results;

Word	Attribute Name	Total Occurences	Document Occurences	Apple Iphone ↓
apple	apple	17	6	17
phone	phone	15	9	15
iphone	iphone	12	8	12
cable	cable	7	1	7
cables	cables	7	1	7
service	service	7	3	7
charger	charger	6	3	6
cost	cost	5	2	5
love	love	5	5	5
told	told	5	2	5
android	android	4	2	4
made	made	4	2	4
products	products	4	3	4
said	said	4	2	4
amazon	amazon	3	2	3

The most common words are apple, phone, iphone, cable, service,charger are the ones with the most occurences as shown above in the figure.

After that model we add 'n-Grams' operator with a size of 2 terms to the model, we obtained a result (shown below in the figures).

Word	Attribute Name	Total Occurences	Document Occurences	Apple Iphone
absolutely	absolutely	1	1	1
absolutely_satisfied	absolutely_satisfied	1	1	1
accommodating	accommodating	1	1	1
accommodating_helped	accommodating_helped	1	1	1
amazon	amazon	3	2	3
amazon_apple	amazon_apple	1	1	1
android	android	4	2	4
android_ebay	android_ebay	1	1	1
android_league	android_league	1	1	1
android_techy	android_techy	1	1	1
android_user	android_user	1	1	1
anti	anti	1	1	1
anti_shatter	anti_shatter	1	1	1
anticipating	anticipating	1	1	1
anticipating_iphone	anticipating_iphone	1	1	1

Word	Attribute Name	Total Occurences	Document Occurences	Apple Iphone ↓
apple	apple	17	6	17
phone	phone	15	9	15
iphone	iphone	12	8	12
cable	cable	7	1	7
cables	cables	7	1	7
service	service	7	3	7
charger	charger	6	3	6
cost	cost	5	2	5
love	love	5	5	5
told	told	5	2	5
android	android	4	2	4
made	made	4	2	4
products	products	4	3	4
said	said	4	2	4
amazon	amazon	3	2	3

After that we added a new operator named 'sterm (dictionary)' to find the tokens that shares the same root and reduce it into a single common form. In the below figures we highlighted them with a blue color.

Word ↑	Attribute Name	Total Occurences	Document Occurences	Apple Iphone
anymore_told	anymore_told	1	1	1
apple	apple	17	6	17
apple_cables	apple_cables	1	1	1
apple_care	apple_care	1	1	1
apple_certified	apple_certified	1	1	1
apple_charging	apple_charging	1	1	1
apple_lightning	apple_lightning	1	1	1
apple_make	apple_make	1	1	1
apple_makes	apple_makes	1	1	1
apple_product	apple_product	1	1	1
apple_products	apple_products	3	2	3
apple_said	apple_said	1	1	1
apple_security	apple_security	1	1	1
apple_suck	apple_suck	1	1	1
apple_take	apple_take	1	1	1

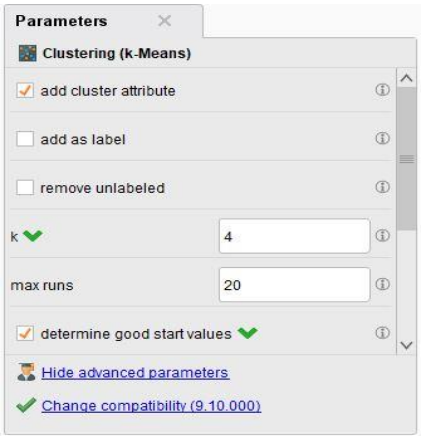
Word	Attribute Name	Total Occurences	Document Occurences	Apple Iphone
anymore_told	anymore_told	1	1	1
app	app	2	2	2
appl	appl	17	6	17
apple_c	apple_c	1	1	1
apple_car	apple_car	1	1	1
apple_certifi	apple_certifi	1	1	1
apple_charg	apple_charg	1	1	1
apple_lightn	apple_lightn	1	1	1
apple_mak	apple_mak	2	1	2
apple_product	apple_product	4	2	4
apple_said	apple_said	1	1	1
apple_secur	apple_secur	1	1	1
apple_suck	apple_suck	1	1	1
apple_tak	apple_tak	1	1	1
apple_w	apple_w	1	1	1

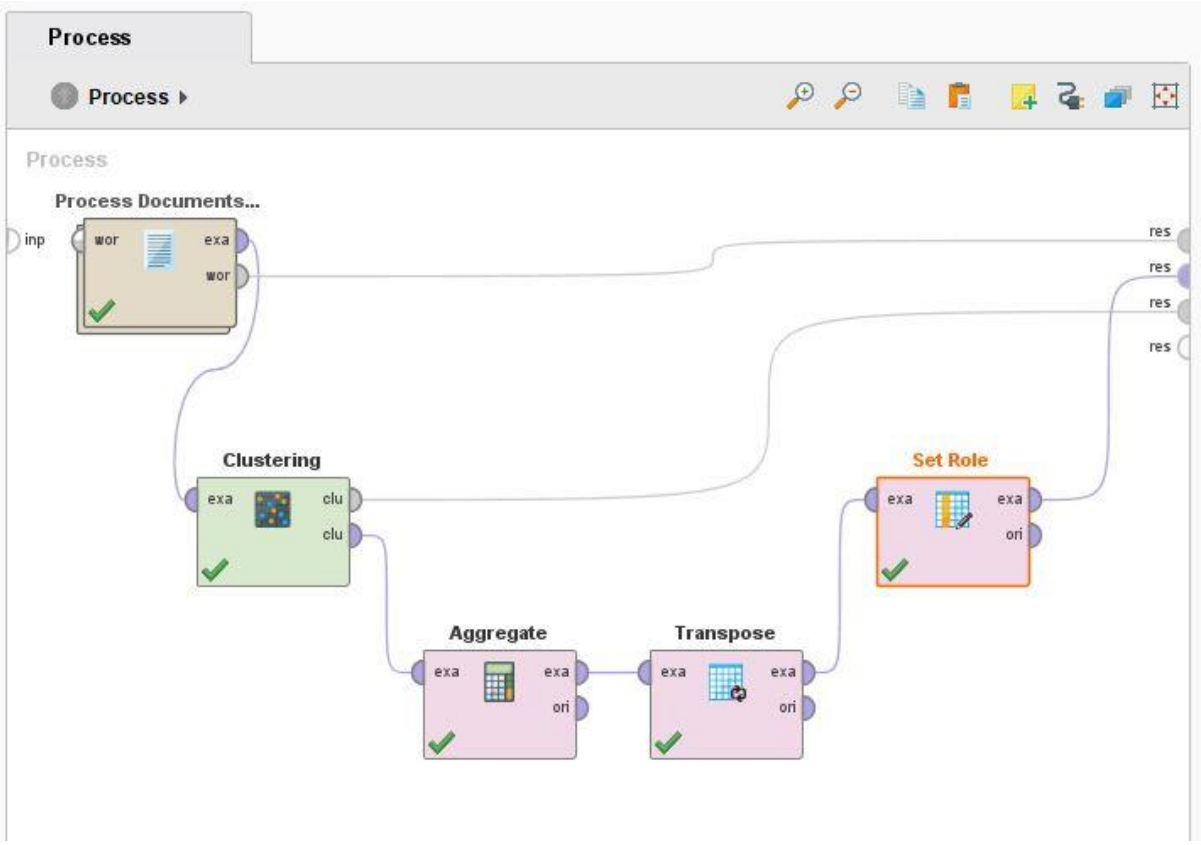
The above figures show that the apple_mak word is combined after the stemmer operator we seen the first figure (above) that shows that apple_make and apple_makes are two different words.

Text mining methods

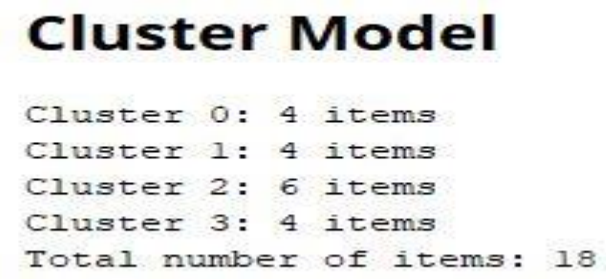
We create a K-means operator for the clustering as it was demanded in the assignment question file.

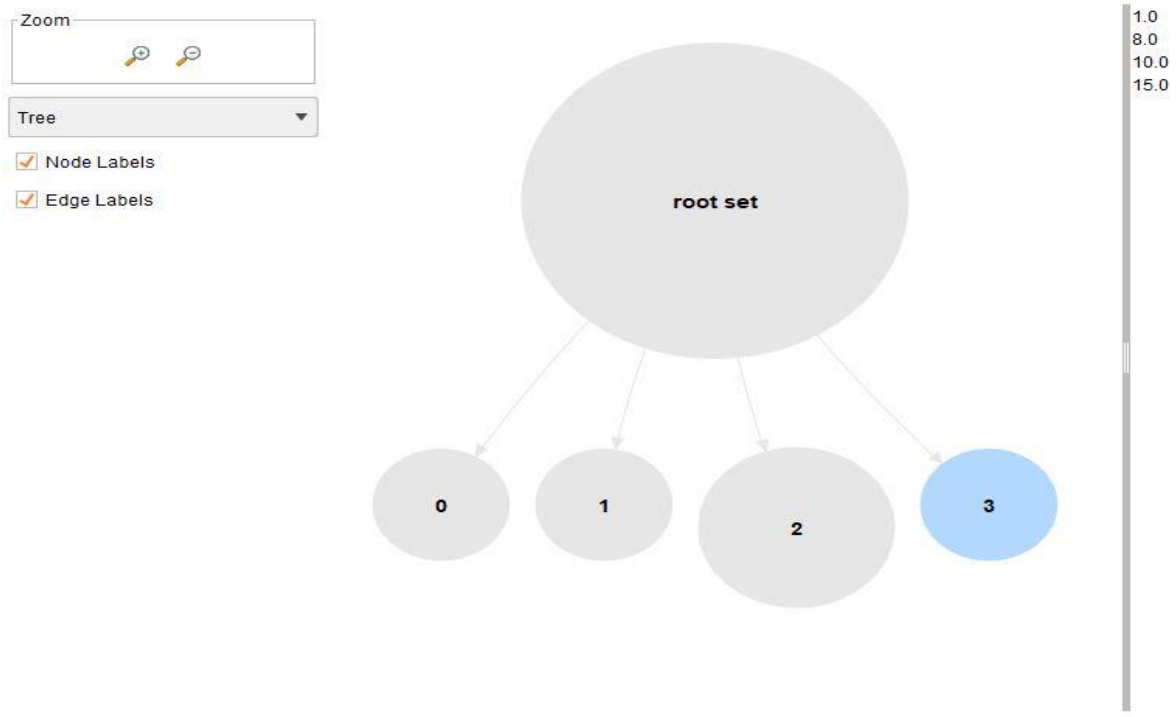
K-means cluster operator group tokens into clusters that have similar keywords, we use 4 as a value of 'k' with the 'max runs' as 20 after experimenting it for more than 5 times. We use the aggregate operator for the results to identify the top keywords in each cluster we transpose the result. All tokens are saved in the ID column in the clustering result, and change the role of ID to regular.





This shows the clusters we have made after the process run.





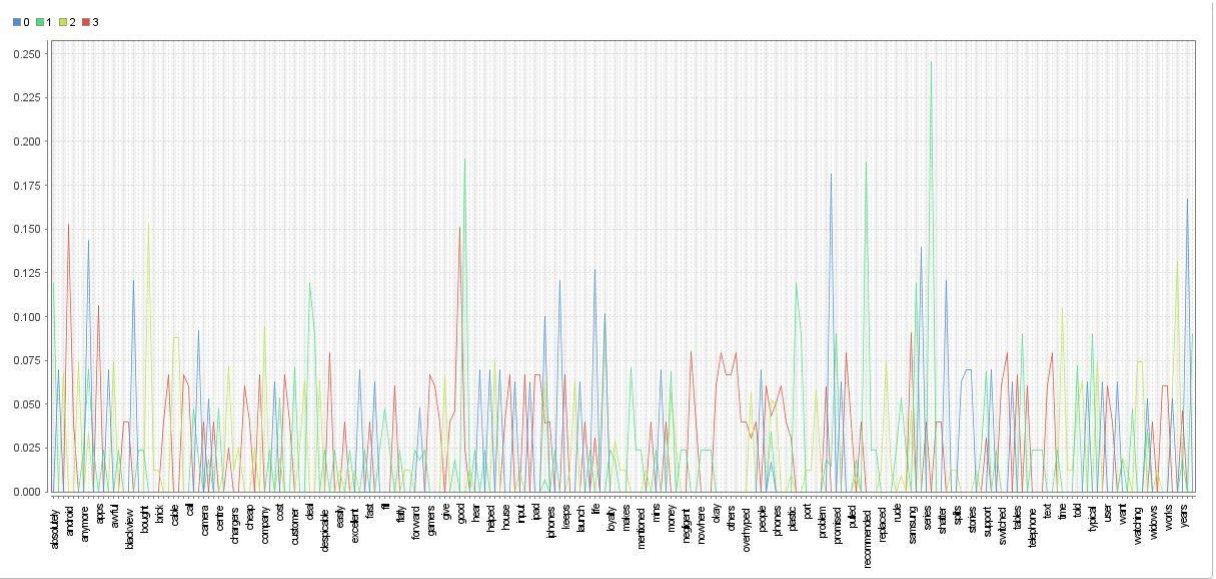
This figure shows the Tree of the clusters and on the right side we can 4 values while we selecting on the 3rd cluster because 3ed cluster has 4 items.

Centroid Table:

Attribute	cluster_0	cluster_1	cluster_2 ↓	cluster_3
brand	0	0	0.152	0
worst	0	0	0.132	0
time	0	0	0.105	0
love	0.102	0	0.101	0
company	0	0	0.094	0
cable	0	0	0.088	0
cables	0	0	0.088	0
anticipating	0	0	0.074	0
highly	0	0	0.074	0
reveal	0	0	0.074	0
watching	0	0	0.074	0
wednesday	0	0	0.074	0
awful	0	0	0.074	0
underwhelming	0	0	0.074	0
charger	0	0	0.072	0.025

Attribute	cluster_0	cluster_1 ↓	cluster_2	cluster_3
service	0	0.245	0	0
great	0	0.190	0	0.025
recommended	0	0.188	0	0
absolutely	0	0.119	0	0
deal	0	0.119	0	0
pleased	0	0.119	0	0
satisfied	0	0.119	0	0
delivered	0	0.090	0	0
poor	0	0.090	0	0
promised	0	0.090	0	0
take	0	0.090	0	0
typical	0	0.090	0	0
zero	0	0.090	0	0
told	0	0.072	0.048	0
customer	0	0.071	0	0

The Plot:



Report:

As we have been observing since analyzing the data, and it can be seen in the results shown above, the most common reviews are around apple and services.

We observe that words such as apple, phone, iphone, cable, service, and charger are the more frequent used in the analyzed corpus, and it is understandable that these 2 are the most common because people talk about them more than anything else.

Cluster 1 shows more words related to the service, which is not a surprise, since it is where users have more access problems, email viewing or interaction with email. Contrary to cluster 2 where words are related to brand.

The customer service manager can use these models to guide the buyers to whether they buy iphone or not and also the services of the Apple company.

References:

Dataset

https://www.trustpilot.com/review/iphone.com

Tutorial for Orange

https://www.bing.com/videos/search?q=text+mining+in+orange&docid=608055395170090738&mid=DC39F64A37F692A1F992DC39F64A37F692A1F992&view=detail&FORM=VIRE

Tutorial for Rapidminer

https://ertekprojects.com/ftp/ertek_et_al_Chapter_03_v22_RapidMiner.pdf