Mod6-CaseStudy1-Solution

Edureka

Text Processing

Load the required packages and libraries library(tm)

Loading required package: NLP library(SnowballC)

library(stringr) library(readr)

library(DT)

Laod the data 'tweets.csv', don't forget to add the argument 'stringsAsFactors = FALSE'.

Length:1181 Min. :-2.0000 ## Class:character 1st Qu.:-0.6000 ## Mode :character Median : 0.0000 ## Mean :-0.1931 ## 3rd Qu.: 0.2000 ## Max. : 2.0000 Now, extract the relevant variable, the one containing the text.

r1 = as.character(tweets\$Tweet)

Now, remove punctuations

Next, change the case of the word to lowercase so that same words are not counted as different because of lower or upper case corpus = tm_map(corpus, content_transformer(tolower))

Now, please carry out the process of stemming, motivated by the desire to represent words with different endings as the same word. corpus = tm_map(corpus, stemDocument)

frequencies = DocumentTermMatrix(corpus) Now create the data frame from the output of the above line

allTweets = as.data.frame(as.matrix(frequencies)) Create a word cloud and set random.order = TRUE:

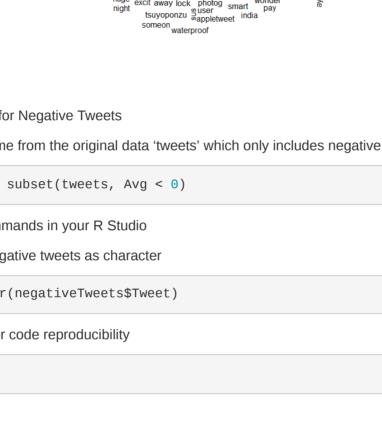
deal hasnt worth better omg gonna absolut httpbittyzedk poor real wasnt httpbittyzedk poor real wasnt httpbittyychyvour option is two yooo continu to download continu to option total appropriate the continuation of mishiza sound total appropriate the continuation of mishiza sound total appropriate the continuation of the con

alarm adambain unlock brand touch ive condom joke perfect secur condescens hard photographi hard photographi sham photographi hard promose p

In the above word cloud, adjust the frequency level with min.freq parameter set at 5

wordcloud(colnames(allTweets), colSums(allTweets), random.order=FALSE, min.freq=5)

biometrsaid first thing iphonec will innovnowthank know doesnt start messag service mac para emiss service maching iphonec will ipodplayerpromo service maching para emiss right and right



Perform Stemming

corpus2 = tm_map(corpus2, stripWhitespace) corpus2 = tm_map(corpus2, removeWords, stopwords('english')) corpus2 = tm_map(corpus2, stemDocument)

Create a word cloud and set random.order = FALSE WordCloud 3 - In the above word cloud, adjust the frequency level with min.freq parameter set wordcloud(colnames(negTweets), colSums(negTweets), random.order=TRUE)

copi gyour sinc s motorola want app miss disappoint today screen away twitter quiet enough samssung samsung

today screen away twitter quiet enough put today screen away twitter quiet enough samssung samsung pleas mad pleas mad pleas think anyon put to changbecom problem samsung samsung samsung pleas mad pleas think anyon put to changbecom problem samsung samsu

releas jonyiv phone protectimessag done plastic ichooseblackberri evenstarz evenstarz switch product serious

iphon

facebook

wordcloud(colnames(negTweets), colSums(negTweets), random.order=FALSE)



support charger

Jarn knowtwitter help last knowtwitter quiet devic cheap phon get well isnt wtf go lock color color for the samsung color color and the sell refus gold amazon aint yooo aanyon googleven everi black muchsuck right to sell refus gold amazon aint yooo aanyon googleven everi black muchsuck right to sell refus gold amazon android pictur switch stand disappoint instead drop pay wideo announc pleas scanner short email miss chang photog sus recommend event got longer

Now, run the follwoing commands in your R Studio Extracting relevant positive tweets

Creating Word Cloud for Positive Tweets

positiveTweets = subset(tweets, Avg > 0)

p1 = as.character(positiveTweets\$Tweet)

Set the seed to 100 for code reproducibility

sample3 = sample(p1, (length(p1)))

Now create the data frame from the output of the above line

posTweets = as.data.frame(as.matrix(frequencies3))

Create three word clouds using the following three instructions

data frame as 'positiveTweets'

set.seed(100)

Sample

watch emoji china innov problem freeever noth screen nokia problem freeever noth screen poke problem free preorder wait microsoft gold back take finger ipod facebook good hour problem freed pod hour problem freed problem freed

day carbon fix custom theyr

create a new dataframe from the original data 'tweets' which only includes positive tweets, where the Avg Value is greater than zero, and name the

WordCloud 1 - Create a word cloud and set random.order = TRUE. WordCloud 2 - Create a word cloud and set random.order = FALSE WordCloud

Bag of Words - Run the above codes 1 - Create a Corpus 2 - Remove punctuations 3 - Convert to lowercase 4 - Remove Numbers 5 - Remove whitespaces 6 - Remove stopwords 7 - Perform Stemming corpus3 = Corpus(VectorSource(list(sample3))) corpus3 = tm_map(corpus3, removePunctuation) corpus3 = tm_map(corpus3, content_transformer(tolower)) corpus3 = tm_map(corpus3, removeNumbers) corpus3 = tm_map(corpus3, stripWhitespace) corpus3 = tm_map(corpus3, removeWords, stopwords('english')) corpus3 = tm_map(corpus3, stemDocument) Create a document term matrix from the resultant corpus frequencies3 = DocumentTermMatrix(corpus3)

develop time your applety natzarticl compani hope year devic champagn alreadillog sak nokia suppose your applety natzarticl compani hope year develop amaz your applety natzarticl compani hope year hope year develop sak nokia suppose your applety natzarticl compani hope year hope your applety natzarticl compani hope year develop sak nokia suppose your applets your app develop time bring seem fashion coolbeauti bigsupport technolog ipad devic champagn alreadi ol day worklaunch

news dont love perfect now the market busi introduc ipod make move

3 - In the above word cloud, adjust the frequency level with min.freq parameter set at 5

wordcloud(colnames(posTweets), colSums(posTweets), random.order=TRUE)

system support smartphon

thehottspot stick

technolog appletweetgenerat yes simpl

technolog amaz year screen interest samsungsa

whitesilv god thatperfect upgrad quickerimpress

world barato futur seem loanyon android macbook ever without microsoft level news

iphoto call done busi next pleas announc

care askideasay pleas microsoft level news

improv book book improved book significant pleas announce colour mishiza

care askideasay pleas microsoft level news

samsung may what microsoft level news

colour mishiza googl doesnttouchid

support smartphon

the stick simples samsungsa

repair wish

repair wish

microsoft level news

colour mishiza

googl doesnttouchid

support smartphon

technolog appletweetgenerat yes simpl

repair wish

microsoft level news

colour mishiza

googl doesnttouchid

support smartphon

therest samsungsa

repair wish

microsoft level news

colour mishiza

googl doesnttouchid

support smartphon

therest samsungsa

repair wish

microsoft level news

colour mishiza

googl doesnttouchid

support smartphon

therest samsungsa

repair wish

microsoft level news

colour mishiza

googl doesnttouchid

support smartphon

file level news

repair wish

repa sure made sure made gonna tech well champagn tech well champagn better back itun now via love app good dear ipod get ipad thank want clet brand someth homebring hope iphonec make mobil genius promoipodplayerpromo innov twitter bubblemania nokia peopl evenstarz green work genius prowid natzglad appletv aroundmani move servic mac bubblemania nokia peopl evenstarz green work genius promoipodplayerpromo innov twitter genius promoipodplayerpromoin innov twitter genius promoipodplayerpromoin promoipodplayerpromoin

wordcloud(colnames(posTweets), colSums(posTweets), random.order=FALSE)

best lipad best lipad best lipad control thank on the second control thank on the second control thank on the second devic phone ten lol make can featur make can new wellthing phon right sure back innov come macbook iphonec time good need wark letter good fingerprint sawsom tech work upgrad compani your give pleas

nokia 🎖 gold 👨 helpwant E =

still^{via} ipoditun best ipad

wordcloud(colnames(posTweets), colSums(posTweets), random.order=FALSE, min.freq=5)

improv hope done microsoft wow way got market witter apromoipodplayer promo perfect impodplayer promo alreadi look store

store solor signature will brooklynk

one dear peop

Creating Word Cloud for Negative Tweets Create a new dataframe from the original data 'tweets' which only includes negative tweets, where the Avg Value is less than zero negativeTweets = subset(tweets, Avg < 0)</pre> Run the follwoing commands in your R Studio Extracting relevant negative tweets as character n1 = as.character(negativeTweets\$Tweet) Set the seed to 100 for code reproducibility set.seed(100) Sample sample2 = sample(n1, (length(n1)))Bag of Words - Run the above codes 1 - Create a Corpus 2 - Remove punctuations 3 - Convert to lowercase 4 - Remove Numbers 5 - Remove whitespaces 6 - Remove stopwords 7 corpus2 = Corpus(VectorSource(list(sample2))) corpus2 = tm_map(corpus2, removePunctuation) corpus2 = tm_map(corpus2, content_transformer(tolower)) corpus2 = tm_map(corpus2, removeNumbers)

Create a document term matrix from the resultant corpus frequencies2 = DocumentTermMatrix(corpus2) Now create the data frame from the output of the above line negTweets = as.data.frame(as.matrix(frequencies2)) Create three word clouds using the following three instructions WordCloud 1 - Create a word cloud and set random.order = TRUE. WordCloud 2 -

tweets <- read.csv("tweets.csv", stringsAsFactors = FALSE)</pre> View(tweets) summary(tweets) Tweet Avg Set the seed to 100 for code reproducibility set.seed(100) Run the following command, 'sample = sample(r1, (length(r1)))', in your RStudio, now you are ready for Bag of Words sample = sample(r1, (length(r1)))Data Preprocessing using Bag of Words Technique Create a Corpus - which, in simple terms, is nothing but a collection of text documents. corpus = Corpus(VectorSource(list(sample))) corpus = tm_map(corpus, removePunctuation) Next, remove numbers corpus = tm_map(corpus, removeNumbers) Next, remove whitespaces corpus = tm_map(corpus, stripWhitespace) Now, remove unhelpful terms, also referred as stopwords corpus = tm_map(corpus, removeWords, stopwords('english')) Creat a document term matrix from the corpus wordcloud(colnames(allTweets), colSums(allTweets), random.order=TRUE) joconfino your option download poletweet gramatic your option download your option of your option your option of your option your option of your option your your option your option your option your option your your option your Create a word cloud and set random.order = FALSE: wordcloud(colnames(allTweets), colSums(allTweets), random.order=FALSE) tablet probabl onlin match teamiphon system excel consum sec sound song consum sec power strategi samssungijordan baratol anoth harden stav order tig park will gasjer httpothyrhyru dorza bforev accion omg simpli swagaal is achievatur ja beratol soon way to the program of the park will be soon om way to the program of the park will be soon of the park will gasjer httpothyrhyru dorza bforev accion omg simpli swagaal is achievatur ja beratol soon way to the program of the park will be soon om way to the park will be soon on way to the park w office occupied anoth sound song constant seep power step of the sound office of cloud notable gone burt enginneither strategi keep power step of the sound of the sound standard sound standard sound standard sound so