

**MGN801: Business Analytics**

**Assignment-3 Text Analytics**

**Submitted By-**

**Shikhar Srivastava (11701485)**

**Roll No- RQ1752B33**

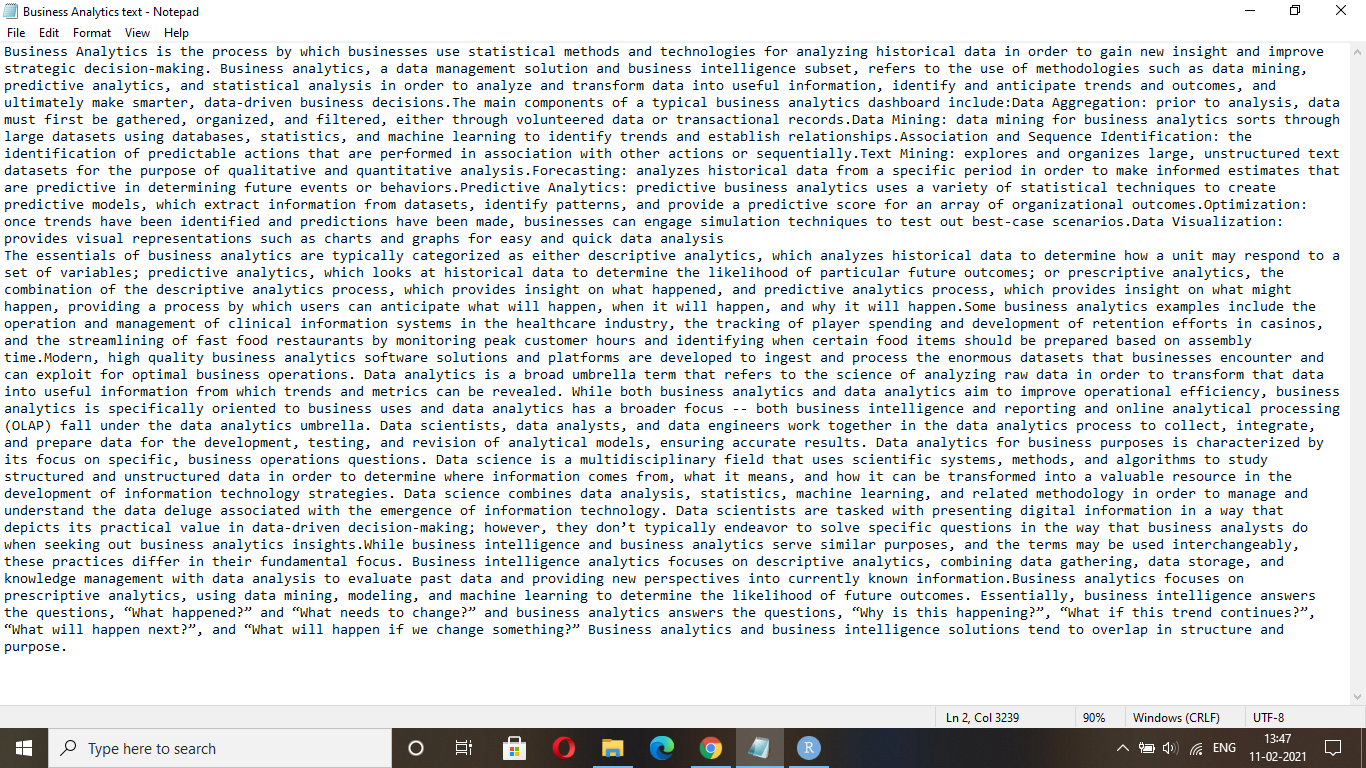
**Submitted To-**

**Dr Mohit Jamwal**

**Assistant Professor, School of Business**

**Screenshot of the text (Text Description- Business Analytics)**

<https://www.omnisci.com/technical-glossary/business-analytics#:~:text=Business%20Analytics%20is%20the%20process,and%20improve%20strategic%20decision%2Dmaking>.



**#Using tm and pdf tools package, we can analyze data from pdf files**

Step-1 Get the data from it’s location

getwd()

Step-2 Loading R packages

library(tm)

library(SnowballC)

library(topicmodels)

library(wordcloud)

library(stringr)

library(reshape2)

library(sentimentr)

library(scales)

library(RCurl)

library(syuzhet)

Step-3 Importing the text

filenames=list.files(getwd(),pattern="\*.txt")

filenames

class(filenames)

Step-4 Read the files

files=lapply(filenames,readLines)

files

class(files)

str(files)

Step-5 Create corpus form vector (#tm package is required)

library(tm)

articles.corpus=Corpus(VectorSource(files))

class(articles.corpus)

**Step-6 Data Cleaning- Text Pre-processing (tm\_map function does data cleaning)**

Data cleansing or data cleaning is the process of detecting and correcting corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

#Make each letter lowercase

articles.corpus=tm\_map(articles.corpus,tolower)

#Remove punctuation

articles.corpus=tm\_map(articles.corpus,removePunctuation)

#Remove numbers

articles.corpus=tm\_map(articles.corpus,removeNumbers)

#Remove generic and custom stopwords

stopwords() #To check the number of stopwords

articles.corpus=tm\_map(articles.corpus,removeWords,stopwords()) #To remove the stopwords

articles.corpus=tm\_map(articles.corpus,removeWords,c("and","Digital"))

**Step-7 Visualization- Word Cloud**

A tag cloud is a novelty visual representation of text data, typically used to depict keyword metadata on websites, or to visualize free form text. Tags are usually single words, and the importance of each tag is shown with font size or color.

library(wordcloud)

wordcloud(articles.corpus,randomorder=F,colors=brewer.pal(8, "Dark2")) #random order=F means more significant words appear in the centre and bigger



Hence, from the word cloud it is clear that the text mainly talks about the data, business and analytics with their highest frequencies in the text, which is why I named the text title as business analytics as seen in the screenshot above.

Step-8 #Create TDM (Term document matrix)

#A term document matrix is a way of represnting the words in the text

tdm=TermDocumentMatrix((articles.corpus))

class(tdm)

tdmfreq=as.matrix(tdm)

class(tdmfreq)

termFreq=rowSums(as.matrix(tdm))

termFreq

Step-9 Subsetting TDM

termFreqsubset=subset(termFreq,termFreq>=1)

class(termFreqsubset)

termFreqsubset

Step-10 Creating a dataframe

tdmf=data.frame(term=names(termFreqsubset),freq=termFreqsubset)

View(tdmf)

row.names(tdmf)=NULL #Remove the rownames

Step-11 Creating a barplot (Top 9 most occurring words)

library(ggplot2)

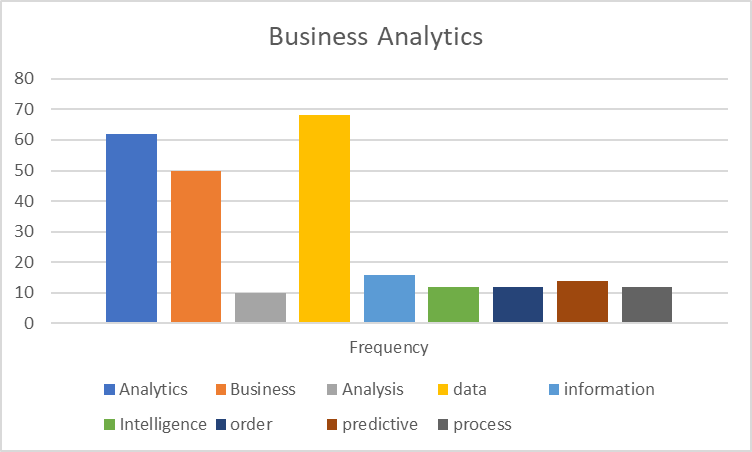
tdmplot=ggplot2(tdmf,aes(x=term,y=freq))+

geom\_bar(stat="identity")+xlab("Terms")+ylab("Count")+

coord\_flip()+

theme(axis.text=element\_text(size=9))

tdmplot



It is clear from the graph as well that the word data, business and analytics occurs maximum times as 68, 50 and 62 times respectively.

**Step-12 Sentiment Analysis**

Sentiment analysis refers to the use of natural language processing, text analysis, computational linguistics, and biometrics to systematically identify, extract, quantify, and study affective states and subjective information.

library(sentimentr)

class(articles.corpus)

a=as.character(articles.corpus)

class(a)

#NRC is one of most popular dictionary for sentiment analysis

mysentiment=get\_nrc\_sentiment(a)

mysentiment

#Calculate the NRC scores

SentimentScores=data.frame(colSums(mysentiment[,]))

SentimentScores

#Give a name to data (since the variable names are not there)

#Give name to the scores column

names(SentimentScores)="Score"

#Give Row names

SentimentScores=cbind("sentiment"=rownames(SentimentScores),SentimentScores)

SentimentScores

#Removing Row names

rownames(SentimentScores)=NULL

SentimentScores

Step-13 Plotting the graphs for sentiment analysis

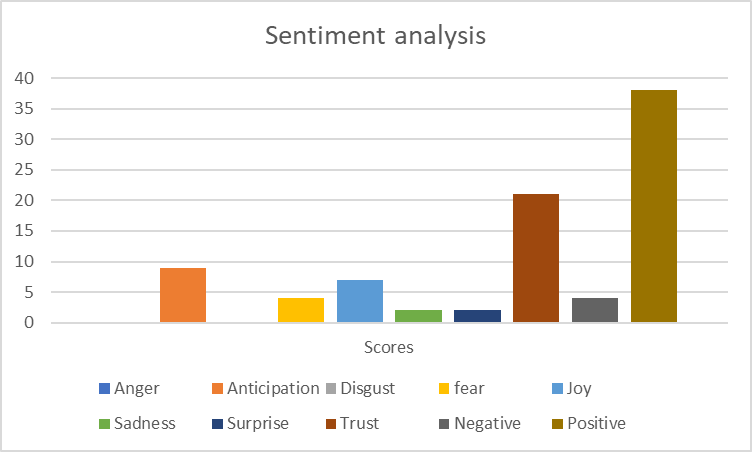
#Plotting the sentiment scores

ggplot2(SentimentScores,aes(x=sentiment,y=Score))+

geom\_bar(aes(fill=sentiment),stat="identity")+

xlab("Sentiment")+ylab("Score")+

ggtitle("Total Sentiment Score")



The topic of the text is just about the business analytics. So, the content is supposed to be more positive and trustful only with their sentiment scores as 38 and 21 respectively. Since they have talked about one of the most trending career in the filed of IT and other sectors, so there has been no negativity or disgust. The text regarding business analytics also talks about some sort of anticipation, surprise and joy may be due to the advantages of business and data analytics. Although there has been a little bit of anger, sadness and fear. So, they must have also talked about some negative points related to this profile.

**Step-14 Topic Modeling**

**Topic modeling** is an unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions that best characterize a set of documents.

#latent dirichlet allocation (LDA) models are a widely used topic model

#Create tdtm

library(topicmodels)

articleDtm=DocumentTermMatrix(articles.corpus,

control=list(minwordLength=3))

k=4 #If we need 4 topics to list out

SEED=1234

article.lda=LDA(articleDtm,k,method="Gibbs",control=list(seed=SEED))4

lda.topics=as.matrix(topics(article.lda))

lda.topics

lda.terms=terms(article.lda)

lda.terms

In topic modelling, they have talked about unsupervised machine learning technique that's capable of scanning a set of documents, detecting word and phrase patterns within them, and automatically clustering word groups and similar expressions.

**Links:**

The text is taken from the following website.

<https://www.omnisci.com/technical-glossary/business-analytics#:~:text=Business%20Analytics%20is%20the%20process,and%20improve%20strategic%20decision%2Dmaking>.

Now, you can view my project on

**Now, let us summarise the code through text mining is done**-

