Keyword Extraction from RCV1 dataset

Godfrey John Dias
18002701

Auckland University of Technology

Email- jjt4398@autuni.ac.nz

Anvay Ajit Karambelkar
18012461

Auckland University of Technology

Email- sst6311@autuni.ac.nz

***Abstract*— The Keyword extraction is highly essential in performing text analytics. The meaningful keywords from the entire text gives the user an overall idea of the text, rather than reading all the information from the text. This research paper describes the text analysis of RCV1 dataset. The objective of this case study is to determine the top five frequently appearing names of the organizations in a set of 2500 text files. The mining of text was done in different ways to clean the noise from the text. The model was trained by applying POS tagging and Named Entity Recognition (NER) techniques. The model performed very well and the results concluded that Reuters, Standard & Poor's, U.S Securities and Exchange Commision, New Stock Exchange, Merrill Lynch appeared at the top of the list.**

Keywords—POS tagging, NER, NLP, Keyword Extraction

I. Introduction

Natural language processing is an efficient way of analysing a text document by applying computational techniques (Weerasooriya et. al,2016). The source of text can be in the form of language or videos used by individuals. Because of increase in social platform or social networking, there is an unprecedented amount of textual information online(Hammar,2018). Thus, there is need of text mining technique to interpret vital information from the text for any organization. One of the most important parts of text mining is extracting keywords from a plain text, which can in the form of blogs, new articles, or online videos. Keywords are represented as a sequence of one or more words, this can in the form of text document. There are huge number of documents available online, so it's not possible to analyse the test data manually. Thus, many researchers are keen to retrieve useful information from the textual data by implementing text mining algorithms or NLP. Text mining is most effective approach to compressed information this can be useful in the applications of text summarization, text clustering, parts-of-speech tagging, sentiment analysis etc. (Bordoloi & Biswas, 2018). Moreover, there are methods based on machine to extract the textual information. Due to growth of textual documents from the news or articles online, the analyses of that text are necessary to summarize the information in a meaningful way, this approach of collecting information is called sentiment analysis.

The keyword extract is the most crucial task in sentiment analysis. The decisions can only be made only if the keyword extraction process is implemented properly. Humans use semantic and syntactic approach to deeply understand the topic and solve the problem by emphasizing relevant keyword. In such a way NLP arranges all the words from the sentences in a syntactic order and assigns a part-of-speech (POS) tags (Weerasooriya et. al,2016). POS tagging is a function in NLP that extracts the essential information even though if it has grammatical errors. The common representation of the text is by using Vector space model (VSM). There arises a problem when it is treated computationally. If it is applied on online social networking text rather than normal text document than the problem become worse(Bordoloi & Biswas, 2018). The factors that creates this problem are casualness, grammatical errors, bad words or slang words even because of short length of text. Thus graph based technique is implemented to perform keyword extraction can be the best way to do analysis. Sometimes the graph of made from the sentences of the text is disconnected because of diversity in the text. Therefore, an appropriate graph based approach should be used to solve most of the graph-based models.

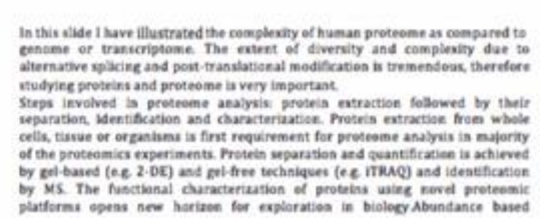
Thus, the keyword extraction process is described in this research paper by analyzing RCV1 dataset. The objective of this research is to perform text mining techniques to extract top 5 most frequent organization in the news. The text was prepossed using Named entity recognition (NER) method on given text files.Entire text analysis was done in R software. The description of data and analysis is explained in the following sections.

.

II. Related Work

A. Keyword extraction from a video transcript

This proposed research paper explains the extraction of keywords from the transcripts of lectures from online video. The text dataset chosen for this analysis was extracted from NPTEL transcripts available for different courses. All the transcript was in form of video format. A simple example of the video transcript is shown in Figure 1 below,



In this slide I have illustrated the complexity of human proteome as compared to genome or transcriptome. The extent of diversity and complexity due to alternative splicing and post-translational modification is tremendous, therefore studying proteins and proteome is very important.
Steps involved in proteome analysis: protein extraction followed by their separation, identification and characterization. Protein extraction from whole cells, tissue or organisms is first requirement for proteome analysis in majority of the proteomics experiments. Protein separation and quantification is achieved by gel-based (e.g. 2-DE) and gel-free techniques (e.g. iTRAQ) and identification by MS. The functional characterization of proteins using novel proteomic platforms opens new horizon for exploration in biology.Abundance based

Figure **1**

The convertor that was used to extract the metadata file was PDF to XML converter. To handle that file in simplest way the data was stored in XML file. Further this text was entered the sentence tokenizer that gives the output of the tokenized sentences. Then these tokenized sentences were transformed into the list of the words with the help on the tokenizer, this process is done in third step. The tokens that are obtained were processed by POS tagger in the next step. The POS tagger attaches the part-of-speech tagger to each token. The next step is NP chucking, the researching of text is done to find chucks that includes individual nouns. The Np chuck grammar is needed to perform NP chucking. The formation of the chucks is decided by that NP chuck grammar. this is the frequently used NP chunk grammar (DT) is the determiner, (JJ) is the adjective and last is (NN) which is the noun. The NP chunking process described by its output in Figure 2. The S is called as root node and then chunks are further spited into NP chunks in the form of sentences. Then those NP tagged leaves are stored as a final keyword after their extraction. The experimented tag pattern used in the research approach are.

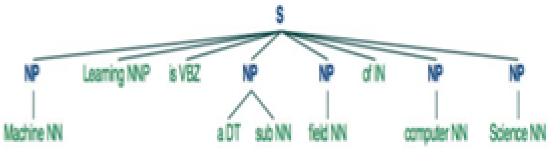
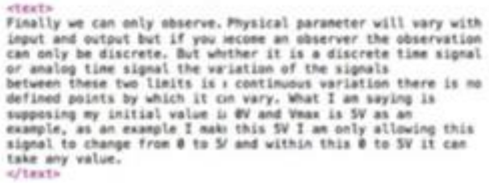


Figure 2

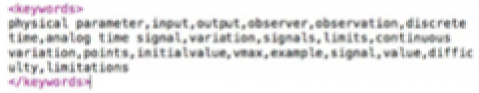
This methodology was performed on Python's NLTK because this interface is very much convenient and around 50 corpora and lexical resources are present in this interface. The word tokenization, NP chunking and parts-of-speech tagging were done by this toolkit. The identification of grammar-based chucks was done by grammar rule based chunker. The main feature of the chunker is that it forms 'chunks' of the word instead of forming the sentences. After executing the code, the pattern that gave the best output was the last pattern,

When the chuck of the code was executed in the python natural language toolkit, the keyword were produced as displayed in Figure 3 and Figure 4.



```
<text>
Finally we can only observe. Physical parameter will vary with
input and output but if you become an observer the observation
can only be discrete. But whther it is a discrete time signal
or analog time signal the variation of the signals
between these two limits is a continuous variation there is no
defined points by which it cn vary. What I am saying is
supposing my initial value is 0V and Vmax is 5V as an
example, as an example I make this 5V I am only allowing this
signal to change from 0 to 5V and within this 0 to 5V it can
take any value.
</text>
```

Figure 3



```
<keywords>
physical parameter,input,output,observer,observation,discrete
time,analog time signal,variation,signals,limits,continuous
variation,points,initialvalue,vmax,example,signal,value,diffic
ulty,limitations
</keywords>
```

Figure 4

In such a way the keywords were extracted by implementing part-of-speech tagging and Np chunker. This methodology was very much efficient is extracting keywords or noun phrases. Although it does extract the keyword but still it has its limitations. The drawback of this is it works only on noun phrases; any other phrases cannot be executed in this chunker.

B. *NLP parsing in biomedicine*

This research describes the application of test mining to extract information based on phosphorylation from the text. This methodology is implemented in three parts. The text was pre-processed by applying the method of substituting the entity names with entity symbols. The Sandford lexical parser was used for parsing sentences. The model used was probabilistic context free grammar model (PCFG). The accuracy of the parser was improved by entity symbols. To process the text further the list of sentences from the text were maintained. To successfully mine the text, PPK should be identified. PPK can be noun or verb phrase, it is important to recognize before mining the information. The pattern used in this study was divided into two types, Base form and sub form. They have defined a BIO which is a noun that represents an entity. In between there is a token that which indicates the number of words between BIO and PPK. Thus, there are several base-forms and sub-forms. Thereafter, the extraction of entity was done by using single/pair/triplet. This extraction was important before applying classification algorithm. There are two methods to extract entity single/pair/triplet in this research.

Method 1: This method is for more than one substrate/kinase/site. This method givens the entity names and their frequency. The words are allocated to the sentences beginning from 0. In this method the hash table is formed to put the entities, PPK and the position of the word. Those entities appear more than once in the hash table are recognized as a group. In such a way all the entities are created by clubbing the entities in the hash table.

Method 2: The first step of this method is the same as that of method 1. This method is applied when the doesn't lie inside the entity pair. Further the PPK is recognized when more than one the has table are recognized as a group. And the last step is like method 1, all the entities are created by clubbing the entities from the parse tree.

SVM classification algorithm: After completion of extraction process the phosphorylation information was processed for support vector machine algorithm. 9 features were used to train machine learning models. PPK, height, parts-of-speech tagging, entity count, distance between substrate and PPK, distance between kinase and PPK, distance between site and PPK, distance between entities and order were those features. Support vector machine is considered to powerful machine learning algorithm. It has excellent ability to fit the model using an extra dimension. This model was trained in WEKA software by altering few parameters. The dataset was divided into training and testing to apply 10-fold cross validation. Finally, the performance was measured by comparing the evaluation metrics. The values of True positive rate, False positive rate, precision, recall and F-value were analyzed to get the best results. In such a way a good F-score was recorded from the analysis.

III. Data Description

This research paper describes the text analytics of the Reuters Corpus Volume 1 (RCV1) dataset. The text data comprises of 2500 text files. It is an archive of more than 800000 newswire stories provided by reuters. To analyze this data there is a need to understand other constraint based on how it was produced. Reuters is one of the biggest televisions and news organizations. It creates around 1000 stories with 23 languages within a day. The quality control measures and coding policy used for producing RCV1 data was done by taking the interviews of Reuters personnel. Moreover, there was needed to access Reuters documentation to prepare data. This dataset contains English language stories written by their own journalists. The data was divided in CD-ROM and was in the form of XML type. To modify this dataset systamatical, the content was verified and validated. The duplicated and irrelevant documents was removed, and dateline and byline format were normalized. However, this dataset includes known errors, the list of the descriptions of the categories are inconsistent with categories that are assigned to the articles. This corpus of 2500 text files are categorized into 4 groups: Industrial, Government, Economics and market.

IV. Background

A. *Natural Language Processing(NLP)*

- “Computer systems that analyze, attempt to understand, or produce one or more human languages, such as English, Japanese, Italian, or Russian. The input might be text, spoken language, or keyboard input. The task might be to translate to another language, to comprehend and represent the content of text, to build a database or generate summaries, or to maintain a dialogue with a user as part of an interface for database/information retrieval” (Ralston et.al, 2004). Through NLP we can understand take-in, analyse, decode and make useful meaning of the data. NLP could be a difficult process since it deals with human language data. Therefore, the uncertainties in the data make it high level and conceptual. The learning of a language is easy for humans but implementing this into machines makes it demand. The paper by Collobert et. al (2011) suggest the four benchmark task. Part-Of-Speech tagging (POS), chunking (CHUNK), Named Entity Recognition (NER) and Semantic Role Labeling (SRL) for NLP which we will be further discussing in this paper. The objective of NLP is to understand language like humans. It paraphrases the input text. The next step is to change the text into some other language. Solve the questions with respective text and derive conclusions from the text. Although NLP possess the above features, its main goal is the NLU (Natural Language Understanding). Moreover, another important function of NLP is to understand query of the user along with the meaning of the proposed problem.

B. *Parts-OF-Speech tagging (POS tagging)*

The parts of speech tagging are the process that involves assigning parts of speech tags to each word (Kanakaraddi, & Nandyal,2018). It is considered as an important tool in analyzing text. There are two types of techniques in POS tagging those are supervised and unsupervised. The trained chunk of text is used and based on that text we get the output. It is much simpler to tag supervised text than unsupervised because in unsupervised text the entire corpus must be trained and then all the sentences are tagged. The supervised learning methods are then subdivided into statistical and hybrid. This is similar to rule-based techniques that takes the information from the experts and write the rules by hand. It is a tough task for a researcher to write those rules manually. But rule-based method is not effective because it does not produce the rule for unknown words. There are exceptional set of handwritten rules written by the researcher to improve the performance of rule-based system (Kanakaraddi, & Nandyal,2018). Parts-of-speech tagging is a part of the grammar of the text that contains verbs, adjectives, nouns, adverbs and determiners etc. To tag the text with respect to its grammar POS tagging is a popular method in NLP. There are certain POS tagger that are used for English language are Claw tagger, Tree tagger and Brill tagger. POS tagging is segregated into two classes, namely open and close class.

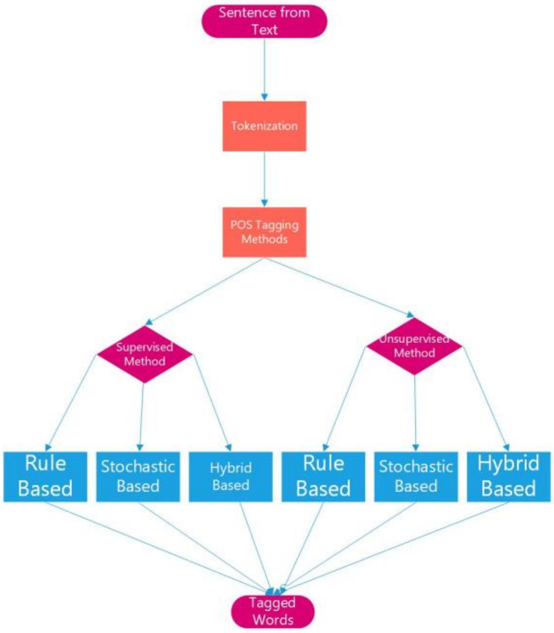


Figure 5

C. *Data pre-processing*

The given text data was pre-processed using the following steps.

- Normalization: In this step the text is normalized by converting upper case to lower case. Sometimes, the columns represented by the feature is not normalized, so transformation of raw text data is necessary. It removes the case sensitivity that exist in the raw data. Moreover, matching of the strings in done by normalizing the text. The association analysis becomes easier after normalization.

- Removing Numbers : All the numbers from the text were removed because they were unimportant for the analysis. Our aim for this analysis dosen;t involve any numerical comparison. So, those numbers were in the form of noise, and they were removed. It completely depends on the user and his objective of the analysis whether to keep number or not.

- Removing punctuations : The punctuation symbols for our analysis were useless, so we removed those punctuations from the corpus. Punctuation provides a grammatical context of a sentence. Even though are helpful for tokenization and to create bag of words but for information extraction they are not useful.

- Removing stop words : The stop words such as a, an, the, etc. were removed from the analysis. In any language stop words are so common that their information value is almost zero. They do not carry much importance in text analysis. The frequency of stop words is always high. Thus, by removing those words improves efficiency of the model.

Moreover, we removed white space, bad words, irrelevant words to clean the corpus. Basically we pre-processed the text based on the research objective.

V. Research Design

The Research Design deals with finding out the first part of the objective i.e. to determine the top 5 most frequent organizations that appeared in the RCV1 Corpus. We used R programming through RStudio with the use of several for experimenting on the corpus. The approach that we have followed to attain this objection are as follows:

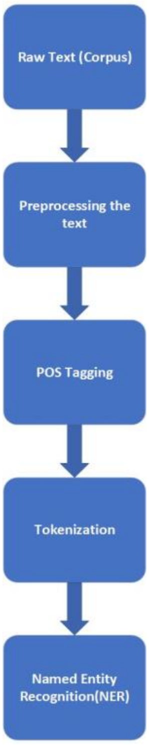


Figure 6

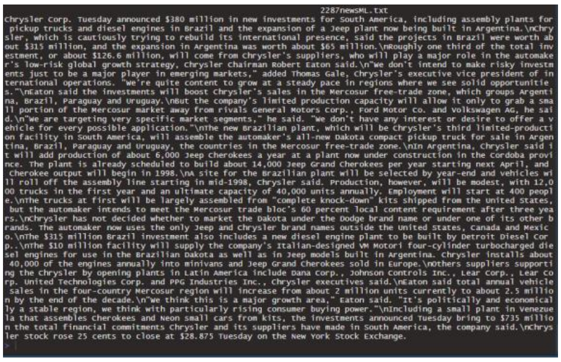
A. Preprocessing the corpus

The first step that we did was to clean the data before we could feed the could to RStudio. There are 7 steps carried out in the preprocessing step of the corpus:

- Removing punctuation
- Removing English Stop words
- Stripping white spaces
- Converting to lowercase
- Removing bad words
- Removing single letter
- Removing non- relevant words

These steps help to remove the noise in the text files of the corpus. These functions are made possible in R by using the “tm” library.

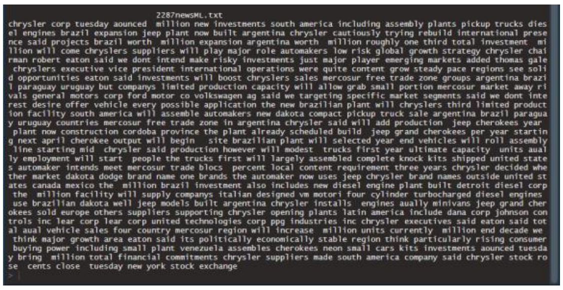
Before we ran the preprocessing functions the first text file was as follows:



```
2287newsML.txt
Chrysler Corp. Tuesday announced $380 million in new investments for South America, including assembly plants for
 pickup trucks and diesel engines in Brazil and the expansion of a Jeep plant now being built in Argentina.\nChry
sler, which is cautiously trying to rebuild its international presence, said the projects in Brazil were worth ab
out $315 million, and the expansion in Argentina was worth about $65 million.\nRoughly one third of the total inv
estment, or about $126.6 million, will come from Chrysler's suppliers, who will play a major role in the automake
r's low-risk global growth strategy, Chrysler Chairman Robert Eaton said.\n"We don't intend to make risky investm
ents just to be a major player in emerging markets," added Thomas Gale, Chrysler's executive vice president of in
ternational operations. "We're quite content to grow at a steady pace in regions where we see solid opportunitie
s."\nEaton said the investments will boost Chrysler's sales in the Mercosur free-trade zone, which groups Argenti
na, Brazil, Paraguay and Uruguay.\nBut the company's limited production capacity will allow it only to grab a sma
ll portion of the Mercosur market away from rivals General Motors Corp., Ford Motor Co. and Volkswagen AG, he sai
d.\n"We are targeting very specific market segments," he said. "We don't have any interest or desire to offer a v
ehicle for every possible application."\nThe new Brazilian plant, which will be Chrysler's third limited-producti
on facility in South America, will assemble the automaker's all-new Dakota compact pickup truck for sale in Argen
tina, Brazil, Paraguay and Uruguay, the countries in the Mercosur free-trade zone.\nIn Argentina, Chrysler said i
t will add production of about 6,000 Jeep Cherokees a year at a plant now under construction in the Cordoba provi
nce. The plant is already scheduled to build about 14,000 Jeep Grand Cherokees per year starting next April, and
 Cherokee output will begin in 1998.\nA site for the Brazilian plant will be selected by year-end and vehicles wi
ll roll off the assembly line starting in mid-1998, Chrysler said. Production, however, will be modest, with 12,0
00 trucks in the first year and an ultimate capacity of 40,000 units annually. Employment will start at 400 peopl
e.\nThe trucks at first will be largely assembled from "complete knock-down" kits shipped from the United States,
 but the automaker intends to meet the Mercosur trade bloc's 60 percent local content requirement after three yea
rs.\nChrysler has not decided whether to market the Dakota under the Dodge brand name or under one of its other b
rands. The automaker now uses the only Jeep and Chrysler brand names outside the United States, Canada and Mexic
o.\nThe $315 million Brazil investment also includes a new diesel engine plant to be built by Detroit Diesel Cor
p..\nThe $10 million facility will supply the company's Italian-designed VM Motori four-cylinder turbocharged die
sel engines for use in the Brazilian Dakota as well as in Jeep models built in Argentina. Chrysler installs about
 40,000 of the engines annually into minivans and Jeep Grand Cherokees sold in Europe.\nOthers suppliers supporti
ng the Chrysler by opening plants in Latin America include Dana Corp., Johnson Controls Inc., Lear Corp., Lear Co
rp. United Technologies Corp. and PPG Industries Inc., Chrysler executives said.\nEaton said total annual vehicle
 sales in the four-country Mercosur region will increase from about 2 million units currently to about 2.5 millio
n by the end of the decade.\n"We think this is a major growth area," Eaton said. "It's politically and economical
ly a stable region, we think with particularly rising consumer buying power."\nIncluding a small plant in Venezue
la that assembles Cherokees and Neon small cars from kits, the investments announced Tuesday bring to $735 millio
n the total financial commitments Chrysler and its suppliers have made in South America, the company said.\nChrys
ler stock rose 25 cents to close at $28.875 Tuesday on the New York Stock Exchange.
>
```

Figure 7

The following is the output from the code to clean the corpus text:

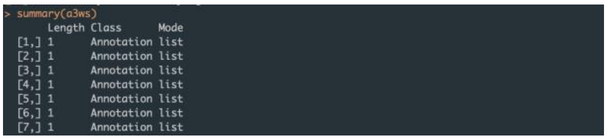


```
2287newsML.txt
chrysler corp tuesday aounced  million new investments south america including assembly plants pickup trucks dies
el engines brazil expansion jeep plant now built argentina chrysler cautiously trying rebuild international prese
nce said projects brazil worth  million expansion argentina worth  million roughly one third total investment  mi
llion will come chryslers suppliers will play major role automakers low risk global growth strategy chrysler chai
rman robert eaton said we dont intend make risky investments just major player emerging markets added thomas gale
 chryslers executive vice president international operations were quite content grow steady pace regions see soli
d opportunities eaton said investments will boost chryslers sales mercosur free trade zone groups argentina brazi
l paraguay uruguay but companys limited production capacity will allow grab small portion mercosur market away ri
vals general motors corp ford motor co volkswagen ag said we targeting specific market segments said we dont inte
rest desire offer vehicle every possible application the new brazilian plant will chryslers third limited product
ion facility south america will assemble automakers new dakota compact pickup truck sale argentina brazil paragua
y uruguay countries mercosur free trade zone in argentina chrysler said will add production  jeep cherokees year
 plant now construction cordoba province the plant already scheduled build  jeep grand cherokees per year startin
g next april cherokee output will begin   site brazilian plant will selected year end vehicles will roll assembly
 line starting mid  chrysler said production however will modest  trucks first year ultimate capacity  units aual
ly employment will start  people the trucks first will largely assembled complete knock kits shipped united state
s automaker intends meet mercosur trade blocs  percent local content requirement three years chrysler decided whe
ther market dakota dodge brand name one brands the automaker now uses jeep chrysler brand names outside united st
ates canada mexico the  million brazil investment also includes new diesel engine plant built detroit diesel corp
 the  million facility will supply companys italian designed vm motori four cylinder turbocharged diesel engines
 use brazilian dakota well jeep models built argentina chrysler installs  engines aually minivans jeep grand cher
okees sold europe others suppliers supporting chrysler opening plants latin america include dana corp johnson con
trols inc lear corp lear corp united technologies corp ppg industries inc chrysler executives said eaton said tot
al aual vehicle sales four country mercosur region will increase  million units currently  million end decade we
 think major growth area eaton said its politically economically stable region think particularly rising consumer
 buying power including small plant venezuela assembles cherokees neon small cars kits investments aounced tuesda
y bring  million total financial commitments chrysler suppliers made south america company said chrysler stock ro
se  cents close  tuesday new york stock exchange
>
```

Figure 8

B. POS Tagging

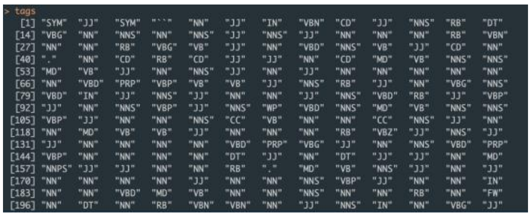
The cleaning of the data in the previous step would help the help the POS tagger to get a fully-text data which would make the identification of words as nouns, verbs, adjectives, adverbs, etc. easier. The “NLP”, “OpenNLP” and “RWeka” are the packages in R to run POS Tagging functions. The first step carried out in the POS tagging process is to tokenize the input for parsing and text analysis. Here is an output of the tokenized text :



```
> summary(a3ws)
     Length Class      Mode
[1,] 1      Annotation list
[2,] 1      Annotation list
[3,] 1      Annotation list
[4,] 1      Annotation list
[5,] 1      Annotation list
[6,] 1      Annotation list
[7,] 1      Annotation list
```

Figure 9

The Word tokens are then applied with POS tags :



```
> tags
  [1] "SYM"  "JJ"   "SYM"  "``"   "NN"   "JJ"   "IN"   "VBN"  "CD"   "JJ"   "NNS"  "RB"   "DT"
 [14] "VBG"  "NN"   "NNS"  "NN"   "NNS"  "JJ"   "NNS"  "JJ"   "NN"   "NN"   "NN"   "RB"   "VBN"
 [27] "NN"   "NN"   "RB"   "VBG"  "VB"   "JJ"   "NN"   "VBD"  "NNS"  "VB"   "JJ"   "CD"   "NN"
 [40] "."    "NN"   "CD"   "RB"   "CD"   "JJ"   "JJ"   "NN"   "CD"   "MD"   "VB"   "NNS"  "NNS"
 [53] "MD"   "VB"   "JJ"   "NN"   "NNS"  "JJ"   "NN"   "JJ"   "NN"   "NN"   "NN"   "NN"   "NN"
 [66] "NN"   "VBD"  "PRP"  "VBP"  "VB"   "VB"   "JJ"   "NNS"  "RB"   "JJ"   "NN"   "VBG"  "NNS"
 [79] "VBD"  "IN"   "JJ"   "NNS"  "JJ"   "NN"   "NN"   "JJ"   "NNS"  "VBD"  "RB"   "JJ"   "VBP"
 [92] "JJ"   "NN"   "NNS"  "VBP"  "JJ"   "NNS"  "WP"   "VBD"  "NNS"  "MD"   "VB"   "NNS"  "NNS"
[105] "VBP"  "JJ"   "NN"   "NN"   "NNS"  "CC"   "VB"   "NN"   "NN"   "CC"   "NNS"  "JJ"   "NN"
[118] "NN"   "MD"   "VB"   "VB"   "JJ"   "NN"   "NN"   "NN"   "RB"   "VBZ"  "JJ"   "NNS"  "JJ"
[131] "JJ"   "NN"   "NN"   "NN"   "NN"   "VBD"  "PRP"  "VBG"  "JJ"   "NN"   "NNS"  "VBD"  "PRP"
[144] "VBP"  "NN"   "NN"   "NN"   "NN"   "DT"   "JJ"   "NN"   "DT"   "JJ"   "JJ"   "NN"   "MD"
[157] "NNPS" "JJ"   "JJ"   "NN"   "NN"   "RB"   "."    "MD"   "VB"   "NNS"  "JJ"   "NN"   "JJ"
[170] "NN"   "NN"   "NN"   "NN"   "JJ"   "NN"   "NN"   "NNS"  "VBP"  "JJ"   "NN"   "NN"   "IN"
[183] "NN"   "NN"   "VBD"  "MD"   "VB"   "NN"   "NN"   "NNS"  "NN"   "NN"   "RB"   "NN"   "FW"
[196] "NN"   "DT"   "NN"   "RB"   "VBN"  "VBN"  "NN"   "JJ"   "NNS"  "IN"   "NN"   "VBG"  "JJ"
```

Figure 10

The different POS tags are plotted using the “ggplot2” library in R:

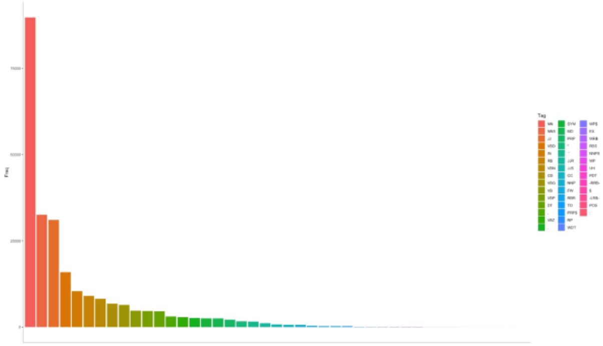
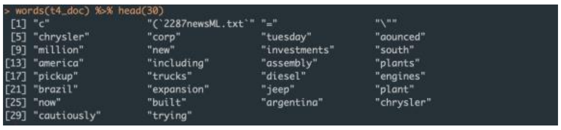


Figure 11

The POS tags are now annotated to the text words in the data frame format :



```
> words(t4_doc) %>% head(30)
 [1] "c"                 "(`2287newsML.txt`" "="                 "\""
 [5] "chrysler"          "corp"              "tuesday"           "aounced"
 [9] "million"           "new"               "investments"       "south"
[13] "america"           "including"         "assembly"          "plants"
[17] "pickup"            "trucks"            "diesel"            "engines"
[21] "brazil"            "expansion"         "jeep"              "plant"
[25] "now"               "built"             "argentina"         "chrysler"
[29] "cautiously"        "trying"
```

Figure 12

C. *Using NER for mining the organization names*

We use NER for extracting the names of organizations that are predefined in the Maxent_Entity_Annotator. This step helps us in retrieving the organizations in the corpus and then we have ranked ordered the list in ascending order according for their frequency in the corpus. This has been plotted again using the “ggplot2” package in RStudio of the top 10 frequently appeared organisation names. We could also plot the same 5 organization but have stuck to 10 for better understanding of the outcome.

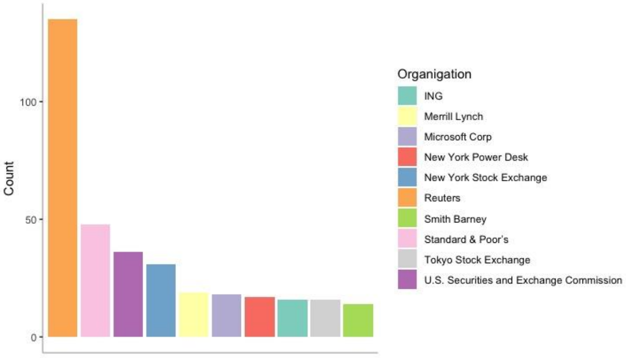


Figure 13

VI. Experimental Results

The above Research of the RCV1 corpus helps us to find out the top 5 frequently occurring organizations names in the text which is shown in Table 1

Table 1

Organization Name	***Frequency***
Reuters	135
Standard & Poor’s	48
U.S. Securities and Exchange Commission	36
New York Stock Exchange	31
Merrill Lynch	19

Thus the first objective of the research to extract the top five most frequent organizations in the news is obtained above. The investigation to infer “Why the organization were in the news?’ is done manually. We did a research about the top 5 company through the company website about each of them. We implied through this review that Reuters is an International News company, Standard & Poor’s is a financial research company for business intelligence, U.S. Securities and Exchange Commission is an agency of the government, New York Stock Exchange is an organization for trading of stocks and Merrill Lynch is an investment management company.

We then manually searched the words in the text and found out that since the RCV1 is news dataset and due to organization background it could be the one taking the news and hence the frequency in the corpus is the highest. Standard & Poor’s word frequency is second highest since it provided a periodic financial rating of the market. Thus putting it on number 2 in the top frequent organization list. The U.S. Securities and Exchange Commission were in the news due to the various organizations filling their documents to them. The New York Stock Exchange have a daily updates of the stocks in the market and since RCV1 is a news corpus, it will contain a daily update of the stock prices. Merrill Lynch similar to New York Stock Exchange gave out ratings of other companies in the market causing it to be in the most frequent list of organisations to appear in the corpus.

VII. Conclusion And Future Work

The research report analyses the RCV1 corpus to find out the top 5 most frequent organisations that appeared in the news. The reason as to why these organisations were in the news was reviewed. NLP was used to examine the corpus and make meaning out of it. The POS tagging and NER technique was used to achieve the objective of this paper.

The metrics evaluation of the NER output has to be considered for future to prove the authenticity of the results. The manually analysis of why the company is in the news can be contemplated to be done automatically by using the n-gram model.

APPENDIX

Code in R programming

```
##Loading libraries
library(RXKCD);
library(tm);
library(SnowballC);
library(wordcloud)
library(RColorBrewer);
library(Rcpp)
getwd()

a<-Corpus(DirSource("Downloads/CCAT/"),
        readerControl = list(language="en"))
#specifies the exact folder where my text file(s) is for analysis with tm.

class(a)

x <- DirSource("Downloads/CCAT/")
#input path for documents
org_name = read.csv("Documents/organization.csv", header = FALSE)

YourCorpus <- Corpus(x, readerControl=list(reader=readPlain))
#load in documents

summary(YourCorpus)
#check what went in your corpus



docs <- Corpus(DirSource("Downloads/CCAT/"))
#load in documents
docs

summary(docs)

inspect(docs[1])

# get rid of html tags
pattern <- "</?\\w+((\\s+\\w+(\\s*=\\s*(?:\".*?\"|'"

for (j in seq(docs))
{
  docs[[j]] <- gsub("-", " ", docs[[j]])
  docs[[j]] <- gsub("@", "", docs[[j]])
  docs[[j]] <- gsub("nn|", "", docs[[j]])
  docs[[j]] <- gsub("pattern", "", docs[[j]])

}

inspect(docs[2])


###################
#Remove punctuation
###################

docs <- tm_map(docs, removePunctuation)

inspect(docs[1])



#########################
#Remove English Stopwords
#########################

length(stopwords("english"))

stopwords("english")

docs <- tm_map(docs, removeWords, stopwords("english"))

inspect(docs[1])

##################
#Strip whitespaces
##################

docs <- tm_map(docs, stripWhitespace)

inspect(docs[1])


################
#Remove numbers
################

docs <- tm_map(docs, removeNumbers)

inspect(docs[1])

###########################
#conversion to lower case
##########################

docs <- tm_map(docs, tolower)

inspect(docs[1])

###########################
#bad word list downloaded #
##########################


bads = readLines("Downloads/Terms-to-Block.csv")
docs <- tm_map(docs, removeWords, bads)

###############################################
# remove single letter and extar white spaces #
##############################################
docs <- tm_map(docs, removeWords, letters)

##########################
#remove non-relevant words
##########################
```

```
docs <- tm_map(docs, removeWords, c("per cent", "email",
"NA"))
inspect(docs[1])
docs

########################
#Part-of-speech tagging
########################
library(NLP)
library(openNLP)
library("tm")
library("SnowballC")
library("RWeka")
library("wordcloud")
library("reshape2")
library("ggplot2")
library(rJava)
library(magrittr)
library(dplyr)
s <- as.String(docs)
## Need sentence and word token annotations.
sent_token_annotator <- Maxent_Sent_Token_Annotator()
word_token_annotator <-
Maxent_Word_Token_Annotator()
pos_tag_annotator <- Maxent_POS_Tag_Annotator()

a3 <- NLP::annotate(s, list(sent_token_annotator,
word_token_annotator, pos_tag_annotator))
a3w <- subset(a3, type == "word")
tags <- sapply(a3w$features, `[[`, "POS")
tagtb <- table(tags)
summary(tagtb)
summary(a3ws)
a3ws <- annotations_in_spans(subset(a3, type == "word"),
                  subset(a3, type == "sentence")[3L])[[1L]]

## Determine the distribution of POS tags for word tokens.
tags <- sapply(a3w$features, `[[`, "POS")
tags
table(tags)

pof <- data.frame(table(tags))

colnames(pof) <- c("Tag", "Freq")
pof <- pof[order(pof$Freq, decreasing = TRUE),]
pof <- transform(pof, Tag = reorder(Tag, order(Freq,
decreasing = TRUE)))

ggplot(pof, aes(x = reorder(Tag, -Freq), y = Freq, fill=Tag))
+ geom_bar(stat="identity")+
 theme(axis.title.x=element_blank(),
     axis.text.x=element_blank(),
     axis.ticks.x=element_blank(),
     panel.background = element_blank(),
     axis.line = element_line(size = 0.5, colour = "gray"))

t3 <- NLP::annotate(s, list(sent_token_annotator,
word_token_annotator))
head(t3)

t3_doc <- AnnotatedPlainTextDocument(s, t3)
words(t3_doc) %>% head(10)

org <- Maxent_Entity_Annotator(kind = "organization")

t4 <- NLP::annotate(s, list(sent_token_annotator,
word_token_annotator, org))

t4_doc <- AnnotatedPlainTextDocument(s, t4)

library(ggplot2)
ggplot(data, aes(x = reorder(V1, -V2), y = V2, fill=V1)) +
geom_bar(stat="identity")+
 theme(axis.title.x=element_blank(),
     axis.text.x=element_blank(),
     axis.ticks.x=element_blank(),
     panel.background = element_blank(),
     axis.line = element_line(size = 0.5, colour = "gray"))+
 ylab("Count") +
guides(fill=guide_legend(title="Organigation")) +
 scale_fill_brewer(palette="Set3")
```

References

[1] Bordoloi, M., & Biswas, S. Kr. (2018). Keyword extraction from micro-blogs using collective weight. Social Network Analysis and Mining, 8(1), 58. https://doi.org/10.1007/s13278-018-0536-8

[2] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. Journal of machine learning research, 12(Aug), 2493-2537.

[3] Kanakaraddi, S.G., & Nandyal, S.S. (2018). Survey on Parts of Speech Tagger Techniques. 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 1-6.

[4] Ralston, A. (2004). Four Editions and Eight Publishers: A History of the Encyclopedia of Computer Science. IEEE Annals of the History of Computing, 26(1), 42.

[5] T. Weerasooriya, N. Perera and S. R. Liyanage, "A method to extract essential keywords from a tweet using NLP tools," 2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer), Negombo, 2016, pp. 29-34.
doi: 10.1109/ICTER.2016.7829895

[6] Hammar, Kim & Jaradat, Shatha & Dokoohaki, Nima & Matskin, Mihhail. (2018). Deep Text Mining of Instagram Data without Strong Supervision. 158-165. 10.1109/WI.2018.00-94.