**Text Tagging using NLP**

Final Report

Data 606- Capstone Project

Spring 2020

Instructor : Dr. Ergun Simsek

Submitted by:

Srinivasa Akhil Vutukuri

**Overview :**

My primary goal in this project to create text tagging through headlines of the story on the data aggregated from various news websites from march to august in 2014 using Natural Language Processing. I would also like to do some analysis based on trends in the data like number of articles published on which days , posted time , which website keeps up to date news, most published website, Which type of articles are frequently posted on the news sites and create some visualizations based on the trends in the data and plot the graph using time where and which point of time the highest posts are recorded and the reasons. I will the create an algorithm for text tagging and find the best suitable one.

The Idea was to determine the similarity between various articles from various news sites by retrieving actual article though article link in the dataset

**Motivation:**

Whenever there’s a news about an issue I have always observed that there’s have been more than one point of views from different publishers/news networks. This was always intriguing as to what the actual truth might be. When I came across this dataset, I saw it as an opportunity to present all the published versions and the facts that they entailed. Adding to this the course work through the previous semesters enabled me to view this problem in a data science project perspective. Hence, I have decided to move forward through this project.

**Data Source and Dataset :**

The dataset which I am going to use for this project is the *News Aggregator Data Set* from *archive.ics.url.edu* . There are 417k rows and 8 columns

Dataset consist columns like:

ID

TITLE

URL

PUBLISHER

CATEGORY - It contains various fields like business, health, entertainment etc.,

STORY

HOSTNAME - Name of the news site

TIMESTAMP - Timestamp is the Unix time of the published article.

**Related Literature/Industry Research and Outcomes**

As we know there is a data which is hovering around in cloud like structured , unstructured and semi structured data. There is a huge amount of social media data and text data which needs to be processed. We can get valuable insights from text mining from product suggestions to fraud detections. We can get even make potential advertisements. So, text mining has a huge impact when it comes to big data industry.(1) We can look into patterns of search and work with suggestions. We can even make use of chatbots and assign better work allocation for customer care services. There are much more uses in warehouse managements where one can create tags of various layers of allotment including summarizing and organizing important data from spam using spam filtering.(2) Sentiment analysis is also a part of text mining where we can extract valuable options which can improve and boost the approach of branding.

There is various project on text mining but what I was looking to do different is to retrieve the related article from website article weblink provided from my dataset and get the inferences from different publishers and similarity in the articles as well as the relativity. It might be a challenge as the some of the data web links are inactive which can affect the model and analysis. There are quite a lot of project with quite less predictions. My project goal is to attain an accuracy around 70 to 80%.

**Transformation, Data Cleaning and Exploratory Data Analysis**

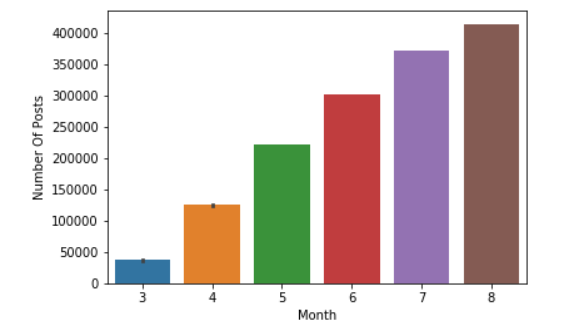
Initial cleaning the data is a huge task as the data is pretty clumsy and inconsistent. Importing the raw dataset created some problems as the data is not aligned in the dataset using pandas. So, after using the separator, data was imported into the dataset in a readable manner. Some of the columns are non-significant as the parameter are random and after dropping them, replacing missing values was the primary task. Instead of dropping them, missing values are replaced through some methods. Next step is to break down Unix time into readable format as it could be used for exploratory analysis to find patterns in articles published in terms of type of industry 

Figure 1 Total number of articles published per month

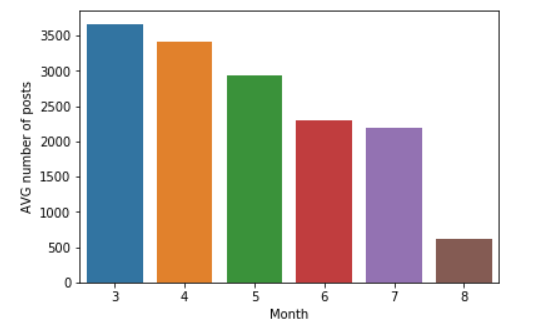


Figure 2 Avg number of articles published per day in every month

Fig 1 indicates that number of articles published by all the publishers in a span of 6 months. Fig 2 represents average numbers of posts per day in a span of 6 months as number of days varies from month to month which changes the whole perspective of the findings.

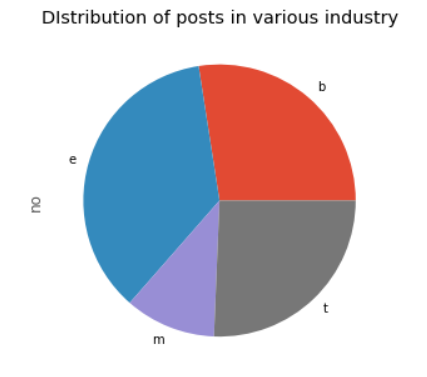


Figure 3 Different type of industry articles published

Fig 3 represents that entertainment(e) industry has dominance in the articles published followed by business(b) which are followed by Science and technology sector(t) and medical(m).

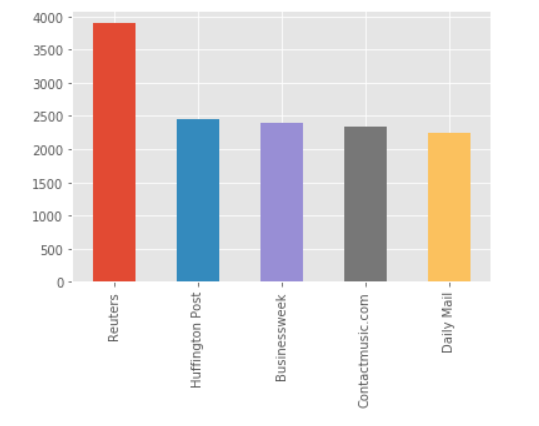


Figure 4 Most published magazines

Upon exploratory analysis, the results of the publishers with a greater number of articles in the span of 6 months are in fig 4. Reuters is leading with 3902 articles followed by Huffington post with 2496 articles. These trends are so useful as this can help in advertising, markets, pricing as the advertisements with more reach are charged more(3).

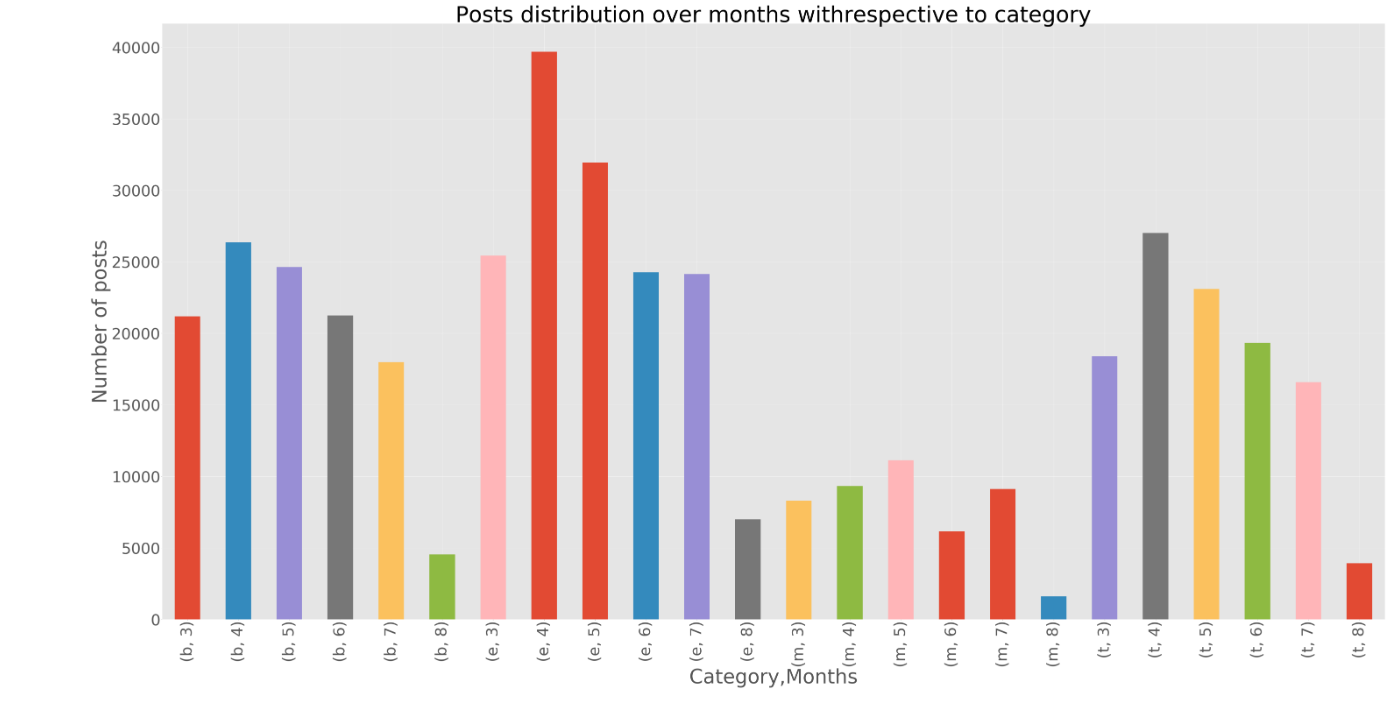


Figure 5 Articles published in every month with respective category

The articles published representing various industries with respective months is represented in fig 5. As we can see medical related few compared to entertainment. We can obtain some insights from this as articles related to entertainment in April are so high as Game of thrones has released their new episode.(4) whereas science and technology also has a hike in April as galaxy smart phone is released.(5) So, every insight can be useful if a meaning is derived from that.

Followed by exploratory analysis, text cleaning is one of the primary objectives as the whole dataset is about articles. Started with removing punctuations , stop words as this mainly reduces the distortion as those are more used words like “ an, the , but, then “, numbers from the text data, changed the data into single case format. Then the data is to be lemmatized which removes multiple copies of singular- plural like rocks – rock , better -good and stemming which brings words to its root form which decreased my wordcount with a huge margin. After preprocessing, started to look out for most frequently used word as I divided all the sentences into word tokens and individual tokens for each category.

Observed some trends in the news like in fig 7 , google, apple, Samsung, Microsoft are most used in science and technology industry as the new phone releases and top companies. In terms of entertainment industry new is most used words . it refers to new pop album, new videos, new season of a show etc., Coming to Business sector, US is most repeated word followed by stock which actually makes sense as stock market is one of the biggest news bulliten for the magazines. Coming on to medical Ebola , study, health is most used as Ebola occurred in early 2014 and study and vaccines are developed in that span of time which is obviously the source of the news.

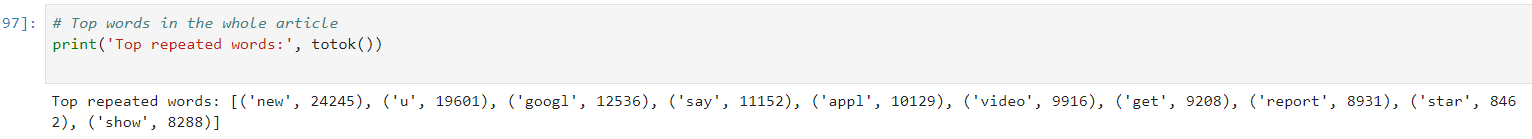


Figure 6 Most trending words

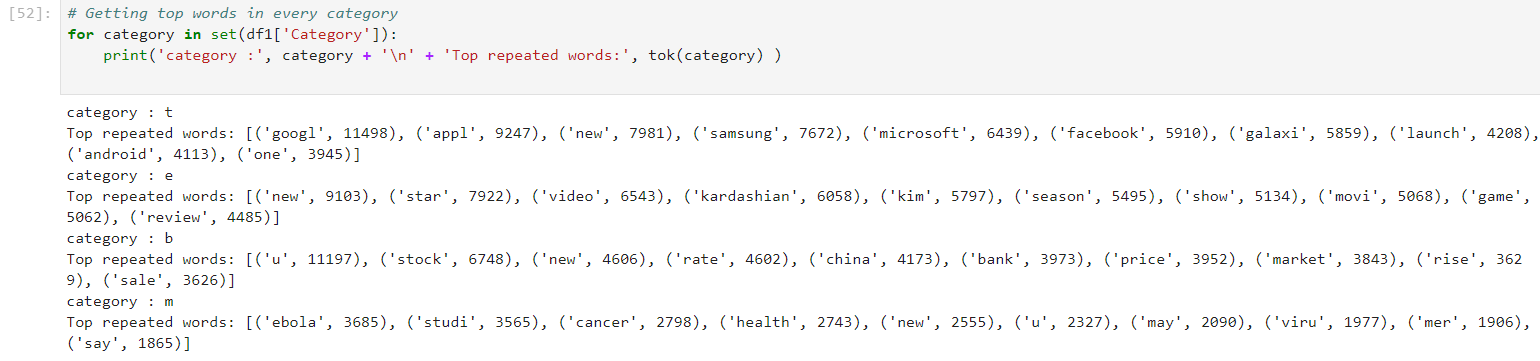


Figure 7 Most trending words in every category

After obtaining tokens from all the headlines , plotted a word cloud to represent the most important news and what they are based on in fig 8



Figure 8 Word cloud based on all words in article headline

Fig 8 shows that most used and important news like Samsung released galaxy mobile, a new season of game of thrones, news about Miley Cyrus, Kim Kardashian and Kanye west, Justin Bieber.

After preprocessing extracted the data from the html link provided in my data column weblink is done. As every html page is written with a unique code. After some analysis, I found out that all the website link from same magazine has same set of code which is quite different from other magazine websites. So, BeautifulSoup helped with parsing the html page and retrieve the data, which is used for analysis, but it is a huge time taking process as my data has 400,000 rows with more than 10,000 unique magazine publishers. To obtain the key from the website from all the magazine websites is quite impossible. So, I retrieved top 4 magazines as some of the top magazine stopped support for the articles as the data is from 2014.

**Transforming my data for the model**

Then I needed to feed data to my models which is quite difficult as scikit learn expects numerical values which is issue for text data input. So, I used label encoder on my category column to convert my text data to numbers from o to n-1 based on my features. I started working on my title column which is more likely like sentences which are also not an optimistic parameter for scikit learn. So, I vectorized the column which breaks down the text into number of words and into matrix form. This works with models as it is converted into numeric form. Then used train test split to create train and test data for the model inputs.

**Model and accuracy:**

As the data is ready for modelling, data has multiclass labels which means basic regression models will face a hard time with the data as regular models has 2 class like yes or no , true or false but when multiclass data is used in regular models it creates ambiguous decisions. So, onevsrestclassifier is used which helps normal regression models to be compatible with multiclass data. In classification linear will fail in this miserably as the data is categorical. Logistic regression is applied which gave an astounding 95.6 train and 94.4% test accuracy which might be overfitting as perfect accuracies in basic regression models are too good to be true. Upon Checking confusion matrix which also showed some positive result. Upon using cross validation shuffle split from sklearn which helps to get know that logistic regression is not over fitting shown as score is 94.3% as both in fig 9

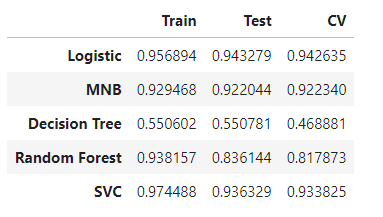


Figure 9 Test, train and Cross validation scores of all models

Other regression techniques like naïve Bayes, decision tree, random forest, svm are used to develop model. Naïve Bayes got a relatively good accuracy compared to decision tree and random forest. Random forest showed 93.8% train accuracy but got only 81% in testing. Whereas SVC results are also prominent with 93.6 % test accuracy from fig 7. Cross validation is done with various approaches like sklearn and shuffle split to obtain most certain cross validation scores.

A comparison chart is shown in fig 10 based on all the training models with accuracies and cross validation scores.

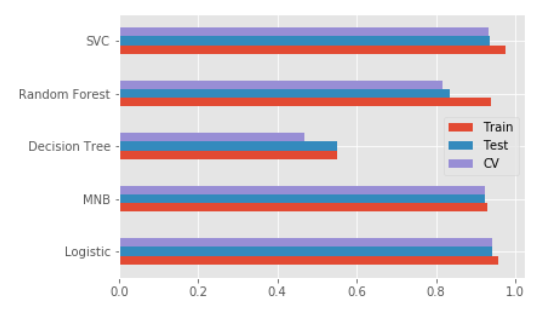


Figure 10 Comparison chart for all the score across all models

Major inferences found through this is all the regression models are to be boosted with multiclass classifier Onevsrestclassifier to obtain good results. Decision tree accuracies are quite low as expected as the data is text data with multi class labels. Random forest is boosted than decision tree but is not the best accuracy. Logistic has more accuracy and is most accurate predictive model which is quite wonderful as basic models like logistic giving results with top accuracy is remarkable.

Retrieving the actual article from magazine using article web html link is a huge task as parsing all the html pages takes a lot of time as in fig 11. Html code on every website differs from one another which makes it hard to get to parse the website with one code.

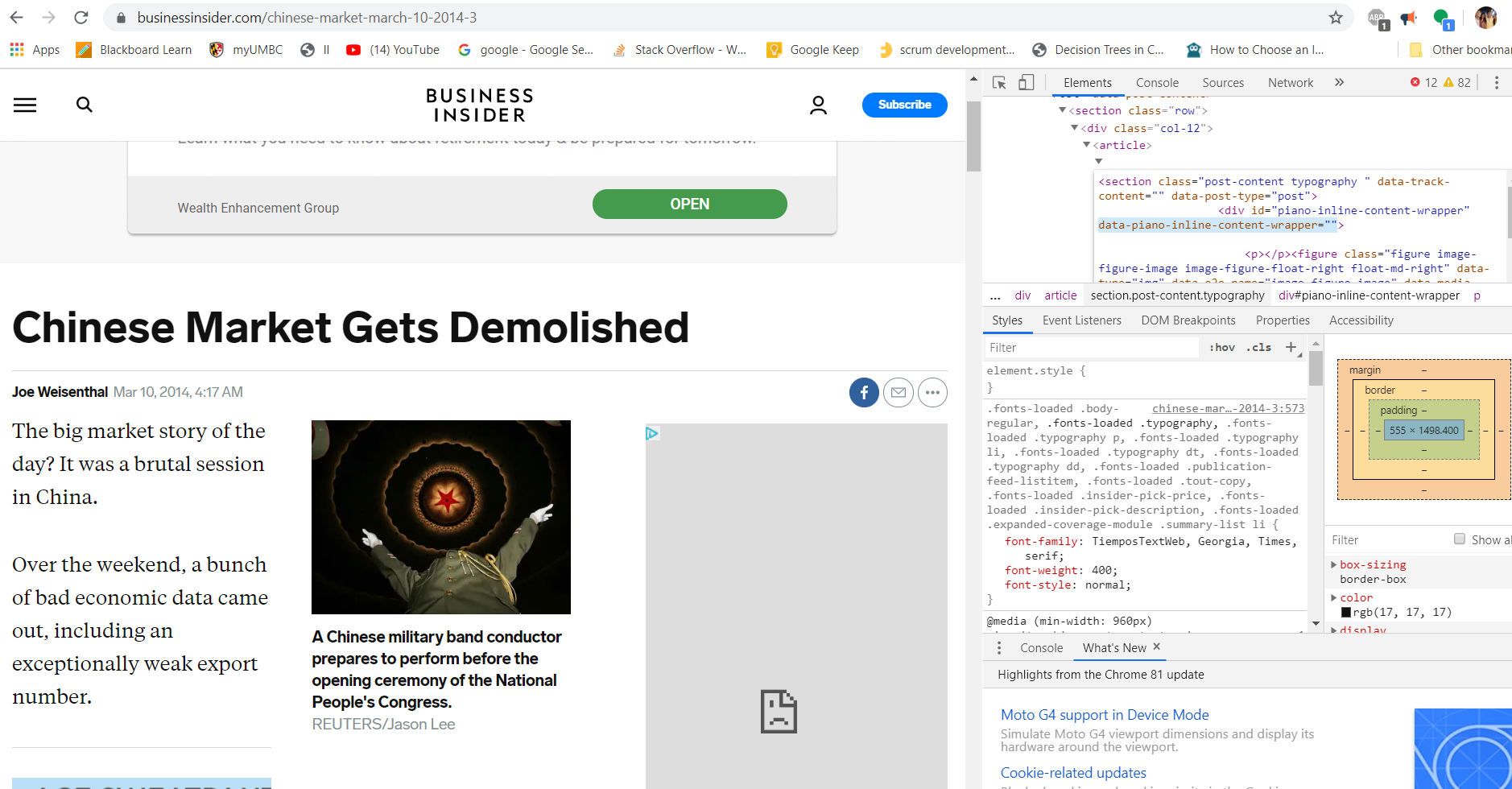


Figure 11 html website extraction source

Addition to that most of the pages put down the support of the articles as the articles are from 2014. Out of top 10 articles only 3 articles are actually active with retrievable articles. Upon retrieving them every article from 3 magazines, intersection article topics are quite low as this makes analysis quite difficult.

So, instead of retrieving articles from top 10 magazines, magazines involved/posted articles on a single type of news is quite useful for the analysis. So, I selected a news from the dataset where Malaysian flight gone missing from Rader. As this is related to news industry and upon 30 articles available on this issue, found out that only 4 articles have active magazine e article which is retrievable as shown in fig 12

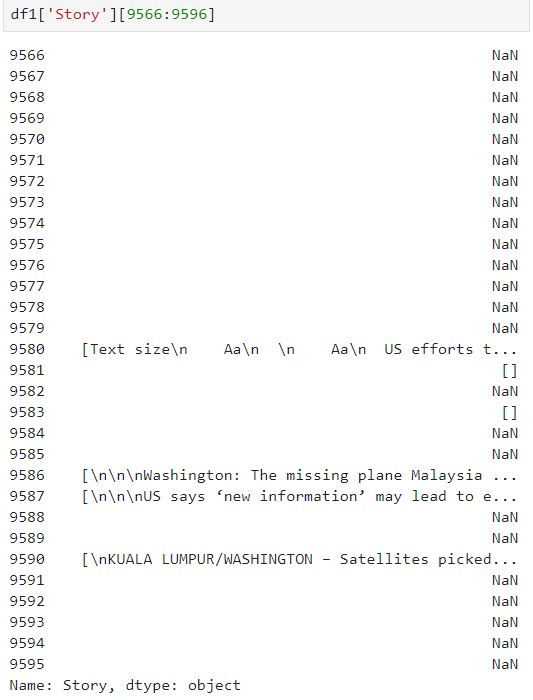


Figure 12 Number of articles which can be retrieved in terms for one topic

This actually made a hard time as the data is not sufficient to do a proper analysis and with 85% missing data it is highly unrealistic to work on this for meaningful observations.

Conclusion:

Most of the targets in the project are reached regarding exploratory analysis and observing trends, followed by text cleaning and most prominent news in that span of time but there are some let downs as the data is inconsistent as 85% of the news articles stories are not actively monitored by magazine publishes which makes it hard to retrieve the data shown in fig 12. This resulted in only 15% data which in based on 4 categories and having 10,000 different publishers which makes it impossible to track all the html pages as all the pages are coded with different indexes which is not possible to parse without the right information. On the bright side, regression models like logistic, SVM, naïve Bayes show more than 92% accuracy (shown in fig 9) which was much more than initially expected while decision tree and random forest didn’t show significant results as the data is multiclass and text data which made it hard for these models to recognize the patterns. Tokens are created and most popular words are found which results to trending news. Both trends of the text data and the related topics are proving the initial assumptions. For example, most of the business articles would be about stocks shown that stock is second most used word as per fig 7. Whereas US which refers to United States is the most used word where as in medical articles study, virus, Ebola , cure has topped the articles as these are from article on Ebola and clinical trials to create a vaccine as the Ebola is epidemic which shook the world.(6) News retrieval and comparison between various articles from various publishers is the only goal which is not quite achieved as working with inadequate data is highly unrealistic for meaningful observations.

References :

1. Full run of the Richmond Daily Dispatch from November 1860 to April 1865  - <http://dsl.richmond.edu/dispatch/>
2. Data mining over 400,000 pages of Vogue magazines - <http://dh.library.yale.edu/projects/vogue/>
3. Julie Meehan, principal; and Evert Gruyaert, senior manager, Deloitte Consulting LLP<https://deloitte.wsj.com/cmo/2018/03/02/6-disruptive-trends-in-pricing/>
4. <https://www.westeros.org/GoT/Features/Month/2014/12>
5. Rich McCormick , Samsung sales and profits down despite galaxy s5 launch <https://www.theverge.com/2014/7/30/5953603/samsung-sales-and-profits-down-q2-2014>
6. Infectious Disease News, October 2014, CDC: First Ebola case diagnosed in United States <https://www.healio.com/infectious-disease/emerging-diseases/news/print/infectious-disease-news/%7Bef9d6ac0-60bf-4063-b0d0-58cc65de44e7%7D/cdc-first-ebola-case-diagnosed-in-united-states>
7. Erik Gregerson, Malaysia Airlines flight 370 disappearance <https://www.britannica.com/event/Malaysia-Airlines-flight-370-disappearance>