Deliverable 3

Data606

**Text Tagging using NLP**

- Srinivasa Akhil Vutukuri

**Data Exploration**

I have discussed some of the data exploration in Deliverable 2 like text cleaning, trends in the data, major category and optimizing null values. So, I started working on cleaning the text data such as removing punctuation, changing everything into single font and removing stop words as this mainly reduces the distortion as those are more used words like “ an, the , but, then “. So, as per text preprocessing i started with lemmatized the words to remove multiple copies of singular- plural like rocks – rock , better -good . After that I use stemming which brings words to its root form which decreased my wordcount with a huge margin. After preprocessing I started to look out for most frequently used word as I divided all the sentences into word tokens and individual tokens for each category.

I tried extracting the data from the html link provided in my data column weblink. 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, I used BeautifulSoup to parse the html page and retrieve the data which I would like to have 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 thought of retrieving top 10 magazines but some of the top magazine stopped support for the articles as the data is from 2014. So, I would like to discuss more about this in my deliverable 4 as I need some to figure out the optimized way to retrieve information.

**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**

As my data is ready for my model, as my data has multiclass labels which means I cannot work with usual models 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, I started using onevsrestclassifier which helps normal regression models to be compatible with multiclass data. So, I decided to go with logistic regression for my initial analysis as my data is categorical which is perfect for logistic. So, using onevsrestclassifier on logistic which I thought would give me some results which could work upon.

**Accuracy of the models**

Upon applying logistic on the data, I got 94.4 % accuracy which gave me some joy and shock as I got good accuracy on the first try but I thought this might me overfitting. So, I checked on the confusion matric which also showed me some positive result. For some cross validation I used shuffle split from sklearn which helps to get know the actual accuracy. Upon validation I got 94.3 % which is in the same range as my actual result. So, after that I used naïve Bayes which I got 92 % and I checked the confusion matrix but to be sure I cross validated that using sklearn which gave me a result in same zone. So, I tried working on decision tree, but I got low accuracy around 42 which is quite shocking as I thought I would get much better result than logistic.

**Conclusion**

I found the trends of my text data and the related topics as the most of my initial assumptions are proved like I thought most of the business articles would be about stocks while stock is second most used word whereas US which refers to United States is the most used word where as in medical articles study, virus, Ebola , cure has toped the articles as these are probably from a similar article as the Ebola is epidemic which shook the world. Moreover, retrieving the data from html pages to work upon took a lot of time as many pages doesn’t have the support as well as obtaining the data from all the articles from a single magazine took me around 1 and half hour. So, on the brighter side I got better accuracies than I expected from some of the models like logistic and naïve Bayes around 94 % while working to improve the existing models and working on the new model for this data for the final submission (deliverable 4) and I hope I get sufficient data from the comparison of the articles and their similarity.