**Text Tagging using NLP**

- Srinivasa Akhil Vutukuri

**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. 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. 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**

I have facing quite some problems while importing my dataset as it imports in different way. So, after the using of separator I could get some proper data frame. I had some columns with non-significant parameters and after dropping them, I has to find a way to replace my missing values . Instead of dropping I found some replacement through some methods. Then I broke down Unix time into readable format where I have done exploratory analysis with month and time where I found patterns for most type of industry which is posted. Then I started on my text cleaning where I removed punctuations , stop words, removed numbers from the text data, changed the data into single case format. Then I had done lemmatizations where I faced some issues with the results. Then I tokenized the words and created a data frame ready for some analysis.

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/>

Presentation link - <https://youtu.be/uOlSEQPaVIs>