**Text Tagging using NLP**

- 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.

I would also like to determine the similarity between various articles from various news sites. The

**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.

**Workflow :**

Initially cleaning the data is a huge task as my data is pretty clumsy and inconsistent. After that I would like to work on trends in the data and plotting them using suitable graphs. Then I would try Regression algorithms and based on the results on accuracy, I would then try on SVM , KNN , decision tree also. Then trying to change and optimize the attributes to obtain better accuracies. I will do text tagging using NLP ( Natural Language Processing) where the paragraphs are automatically categorized and segregated based to the tags. Based on the results of my initial analysis some of my data is not consistent as some of the URL’s are not active. After the filtering I will try to create an algorithm which analyses the similarity of the news articles by various news sites.

[**Data link**](https://archive.ics.uci.edu/ml/datasets/News+Aggregator)

[**Presentation Link**](https://youtu.be/QKlq2FQtCjE)