INFO 7390 PROJECT – Spring 2017

Web traffic using Forums and Message Board

PROJECT PROPOSAL

Overview:

Websites and Internet in today's world has become a huge part of all our lives. Every single word on the Internet is an important data for someone and utilizing this data in a good way is a major challenge for all the Data Scientist across the Globe. Among all the things going on the Internet, we need to understand which way the traffic is going. What is that thing which is pulling crowd and keeping all these things Marketers can come up with a strategy to improve upon their business.

The subject of our project is getting data from online discussions, blogs, news and message boards about an Organization. This data will have various details about the site where the blog is posted, it's title, url, publish date, performance score and various other features.

Goals:

- Data Wrangling and Pre-processing
- Data Analysis and key Insights
- Pipelining the entire flow using Luigi / Airflow
- Dockerizing the process
- Building Classification and prediction Algorithms
- Deploying the outcome using an API

Approach:

Our first challenge is to programmatically download the data from

https://webhose.io/datasets

A zipped folder will be downloaded in the current path. The folder contains different files for each record in json format.

Once the folder is available, we will perform the following:

1) Data Wrangling and Exploratory Data Analysis:

- **Data Concatenation**: The folder contains n number of files (n represents the number of records). We need to pick each file and concatenate in a single large json file containing the entire data.
- **Missing Data Analysis**: The way to handle missing values

- **Feature Engineering**: The variable we might require to predict the best website or the best blog for an Organization to publish their ads and invest more on the sites which is most viewed by people or site with most blogs.
- **Visualization:** We are planning to get some meaningful statistical and analytical insights on the data we have and present it by publishing it on either powerBI or Tableau.
- **Pipeline**: We are planning to automate the entire process using Luigi or Airflow.
- **Dockerizing the entire workflow**: We will use docker to automate the whole process and run the docker image which will give us the predicted value at the end of the process.

2) Building and Evaluating models:

We will use our cleaned data and the features which we engineered in part 1 as an input to this section of the project and start building models. The main reason to build various Machine Learning Algorithms is to train the data as to understand which website or the blog is the best for a kind of category in an Organization, the date when it was mentioned and how the performance has fared over the time.

- **Classification:**

  - Using the data, based on the reviews and the post we will build classification models to predict which section is the user talking about in his post or blog. Since, the sections are more than 2, we will be using multi-class classification Algorithm.
  - We'll repeat this using various other models and choose the best model based on the Accuracy.

- **Prediction:**

  - We'll write a prediction script using R / Python to build a Regression model for training the data on predicting the number of hits or number of views a blog might get based on the site it is posted.
  - This will help the Company to advertise or invest more on advertisement on a site to obtain maximum outcome and analyze the internet traffic.

- **Deployment:**

  - We are not sure regarding the website part, should we be creating or will we require or not. We will try and implement it and present it using API but haven't got an idea on it.