INFO 7390 PROJECT – Spring 2017

BLOG ANALYSIS

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. 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.

Sample Dataset:


```
{
  • organizations: [],
  • uuid: "543889fbd6dd983908e53232ace0b88f35822ca8",
  • thread: {
      ◦ site_full: "www.tripadvisor.com",
      ◦ main_image: https://media-cdn.tripadvisor.com/media/photo-s/02/52/38/75/guest-room.jpg,
      ◦ site_section: https://www.tripadvisor.com/Hotel_Review-g187791-d316644-Reviews-Raffaello_Hotel-Rome_Lazio.html,
      ◦ section_title: "Raffaello Hotel - UPDATED 2017 Reviews &amp;amp; Price Comparison (Rome, Italy) - TripAdvisor",
      ◦ url: https://www.tripadvisor.com/ShowUserReviews-g187791-d316644-r463592325-Raffaello_Hotel-Rome_Lazio.html,
      ◦ country: "US",
      ◦ domain_rank: 189,
      ◦ title: "Terrible Staff, Tiny Rooms",
      ◦ performance_score: 0,
      ◦ site: "tripadvisor.com",
      ◦ participants_count: 1,
      ◦ title_full: "Terrible Staff, Tiny Rooms - Review of Raffaello Hotel, Rome, Italy - TripAdvisor",
      ◦ spam_score: 0,
      ◦ site_type: "discussions",
      ◦ published: "2017-02-28T02:00:00.000+02:00",
      ◦ replies_count: 0,
      ◦ uuid: "543889fbd6dd983908e53232ace0b88f35822ca8"
    },
  • author: "Nicholas W",
  • url: https://www.tripadvisor.com/ShowUserReviews-g187791-d316644-r463592325-Raffaello_Hotel-Rome_Lazio.html,
  • ord_in_thread: 0,
  • title: "Terrible Staff, Tiny Rooms",
```

End User:

Marketing analysts would be our end users, who will have access to accurate data insights about different domains. This would help him understand prospective audience.

We would also be reporting spam blogs to the marketing analyst.

Tasks:

- Downloading the data and converting it into structured form
- Data Wrangling and Exploratory Data Analysis
- Classify your data
- Clustering
- Building User Interface and Web Services
- Dockerizing and Scheduling the pipeline
- Documentation

https://app.scrumdo.com/projects/training486222/#/iteration/207509/board

Download the data and convert it into structured form:

Down the historic data from the site manually and read these JSON files individually(Each domain has a different structure of data). Convert this data from different files into one dataframe and form a single CSV/Excel file.

Repeat same steps for live data coming from webhose.io API.

Data Wrangling and Exploratory Data Analysis:

Data Concatenation: The downloaded historical data of different domains needs to be concatenated to generate a summarized data set which can be used to follow further steps.

Missing Data Analysis: Handle missing data using various techniques.

Exploratory Data Analysis: Find patterns and relations between multiple aspects of data and get some meaningful insights.

Visualization: Visualize the analysis using Tableau and PowerBI.

Classification:

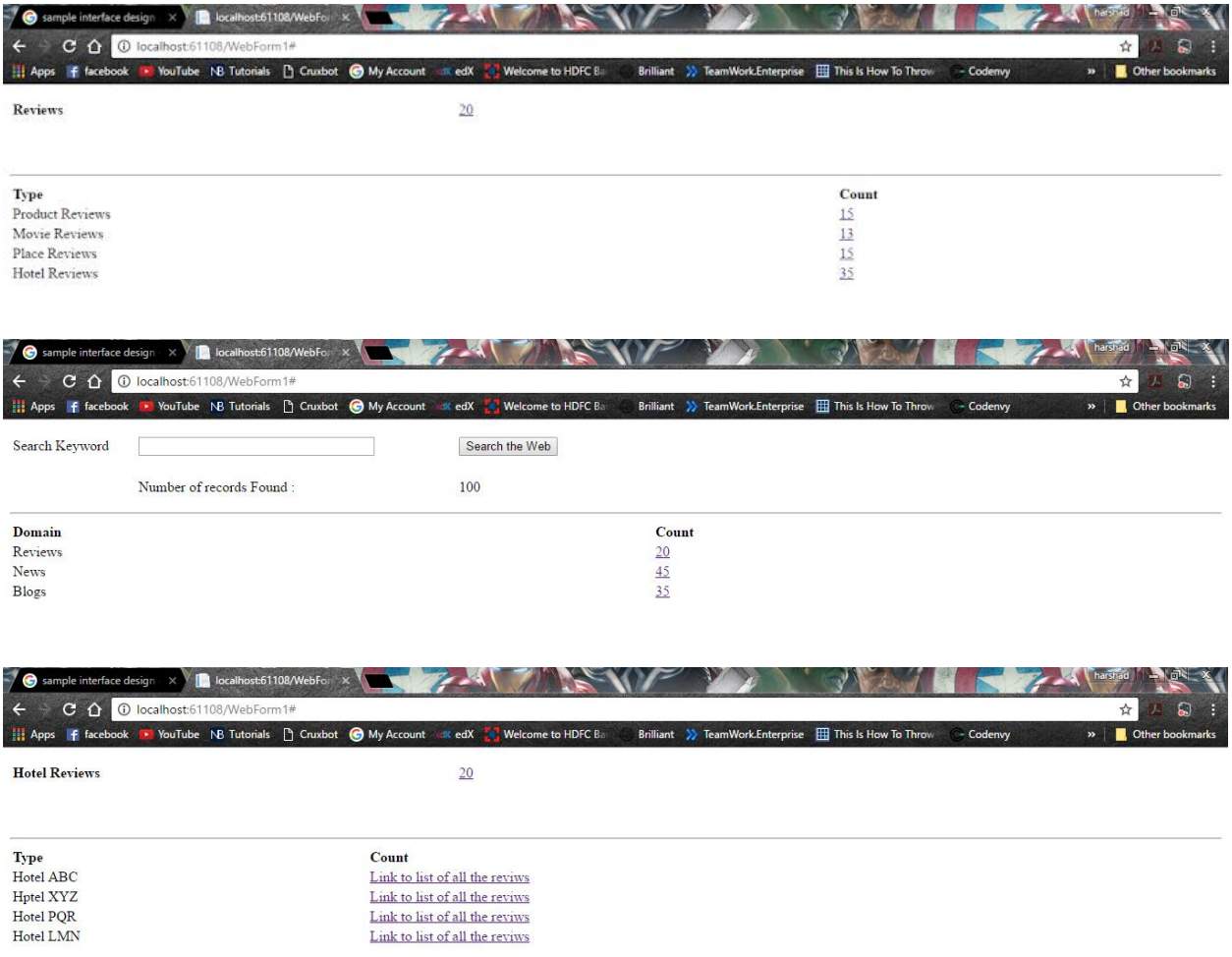
Based on title of the article and origination site, build a classification model to classify live data coming from the API into groups like News article, Reviews, Blogs, etc. We can build and evaluate algorithm like KNN, Logistic Regression, Random forest, Neural Network.

Clustering:

Apply Natural Language processing techniques to read text fields from the article/reviews and determine which cluster it belongs to. We can use word2vec, doc2vec to achieve this.

Building User Interface and Web Services:

We will be creating web services by deploying out best models on Microsoft Azure Machine Learning Studio. We shall design and build customer friendly GUI which can be easily used by various market analysts to observe trends thereby helping them understand the prospective target audience.



Dockerize and scheduling the pipeline:

Pipeline all the steps included above using Luigi and dockerize this pipeline by creating a docker image. The entire process will be scheduled on Amazon. Clean and output file would be stored on Amazon S3 bucket.

Tools/Software Packages:

Jupyter Notebook

RStudio

Tableau

PowerBI

Docker

Luigi

AWS S3

Word2vec

Tsne

Deliverables:

- Working UI
- Comprehensive Report
- Pipelined Docker Image
- .ipymb/ .r files