**CORPUS**

**PROJECT PROGRESS REPORT**

**Team** – Balaram Remala, Mukund Nuthi, Pavan Teja Duggempudi

**Project End Goal**

To create a recommender system based on the review text at Make level.

***Balaram Remala:***

* Assigned task: To work with the reviews associated with ‘Laptops’.

**Achieved Goals:**

* Data Munging:(14 hours)

The data is in the form of group of reviews in json files where each json file have some list of review for each product along with their product specification. That way we have around 2000 products. Out of that 2000 products, few of them do not have proper product description, so I have removed those null values.

* Data Filtering: (4 hours)

Initially, the data is filtered from null values and a new column called “Make” is added to the dataframe for analysis based on brand name.

* Data Cleaning: (5 hours)

The data review content text is cleaned from short words and slang words using slang word dictionaries and word replacements.

* Applying Classifier: (10 hours)
  + 1. A Naïve Bayes Classifier is fed with a list of positive, negative and neutral words given from the txt file and trained to classify the content of our review and return if it is positive/negative/neutral as “Tag”.
    2. Once the Tag is obtained, it is added to the main dataframe and a review on counts of this column is done.
* Data Visualization: (2 hours)

Data is visualized based on makes with most positive review, most negative reviews and ratio of positive and negative counts are taken to get which brands are relatively better.

source: <https://textblob.readthedocs.io>

**Target Goals:**

* Check with the best classifier other than Naive Bayes based on the accuracies and apply to the data.
* Review of products based on time and date range.
* Data Visualization based on “Tag” and other features of the product.

All the above activities are done in a single ipython notebook linked below: <https://github.com/UNCG-CSE/CORPUS/blob/master/Project_Balaram/Project_Balaram.ipynb>

***Mukund Nuthi:***

* Assigned task: To work with the reviews associated with ‘Mobiles phones’.

**Achieved Goals:**

* Dataset understanding & creation - Worked with combining all the JSON files of our data and formed a single dataset which is later saved as pandas dataframe. (15 hours)
* This is the [link](https://github.com/UNCG-CSE/CORPUS/blob/master/Project_Mukund/JSON_READ.ipynb) for the jupyter notebook on dataset creation.
* Text Analytics (20 hours) (Follow this [Link](https://github.com/UNCG-CSE/CORPUS/blob/master/Project_Mukund/Text%20Analytics%20on%20Reviews.ipynb) for Jupyter notebook on Text Analytics )
  + Cleaning Text
    - Removed stopwords, punctuations and special characters from the text data.
  + Extracted the Make of the product and populated into a new column.
  + Created plots on Make vs Avg. ratings
  + Visualizations
    - Implementing the Wordcloud
      * Developed a function that can create wordclouds at Make level.
      * Created Fancy wordclouds which can render the given shape and colors while creating the wordcloud.
      * Wordclouds help to visualize the text which we are planning to add to the recommender system.

Worked with the report documentation.

**Target Goals:** (Since 2 weeks)

* Calculate the TF-IDF scores for the words and create some insightful visualizations
* Right now researching on the use of implementing different classifier for our use case and implement the better classifier. Using NLTK docs for researching (<http://www.nltk.org/book/ch06.html>)

***Pavan Teja Reddy Duggempudi***

* Assigned Task: To work with reviews associated with ‘Tablets’

**Achieved Goals:**

* Understanding the data and had discussions for setting up the goals. (6 hours)
* Worked on combining all the products (JSON files) and cleaning the data by deleting unwanted columns, handling the missing data and null values and merging the Product info with Product Reviews. (8 hours)
* After doing all this created a main required data frame to work on. (3 hours)
* *Visualization:*
  + Plotted a graph by taking price range and their mean ratings and tried to figure out how far the Price is affecting the ratings and came to the conclusion that Price doesn’t have noticeably effect on ratings. (8 hours)
* *Text Analytics:*
  + Next Step worked on getting the Product make and added a column to the main data frame. (5hours)

**Target Goals:**

* Now started working on text reviews and try to find what features are making it a good product and what features are making it a bad product. (6 hours so far)
* Later will be analyzing on make (brand) and what features are making it best and what are making it bad.

Link for Jupyter Notebook -

<https://github.com/UNCG-CSE/CORPUS/blob/master/Project_Pavan/Current_dataframe.ipynb>