**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’.

**Works Done:**

* Data Munging:(6 hours)

The data is in the form of group of reviews in json files where each json file have some list of reviews 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: (2 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: (3 hours)

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

* Applying Classifier: (7 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 as a column to the dataframe.

* Data Visualization: (2 hours)

1. 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>

1. Review of products based on time and date range. Also, a statistical analysis is done based on the counts, ratings and make level.

* Building the system: (9-10 hours)

1. Check with the best classifier other than Naive Bayes based on the accuracies and apply to the data.
2. This classifier is used to analyze the n-grams of the cleaned words based on the input given by the user.
3. At the end, we the best product based on the classification and then-grams of the cleaned words, is identified.

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
    1. 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
    1. 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.
* Implemented Multinomial Naive Bayes on the text data. (10 hours)
* Implemented the recommender system by using the Multinomial Naive Bayes classifier. (20 hours)
* About the recommender system
  + Input – One Feature (like Screen, camera, gpu, processor, etc)
  + Output – Name of the Phone recommended
  + Steps
    1. Takes the feature word as an input.
    2. Pulls out the ngrams based on the given word and builds a new dataframe with ProductID and the ngrams.
    3. Previously trained classifier is applied on these ngrams and classifies them into good or bad.
    4. Come up with one productID for which the number of goods are high.
    5. Fetches the name of the product using the productID which becomes the final recommendation.
* Worked with the report documentation.

***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)

Link for Jupyter Notebook -

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