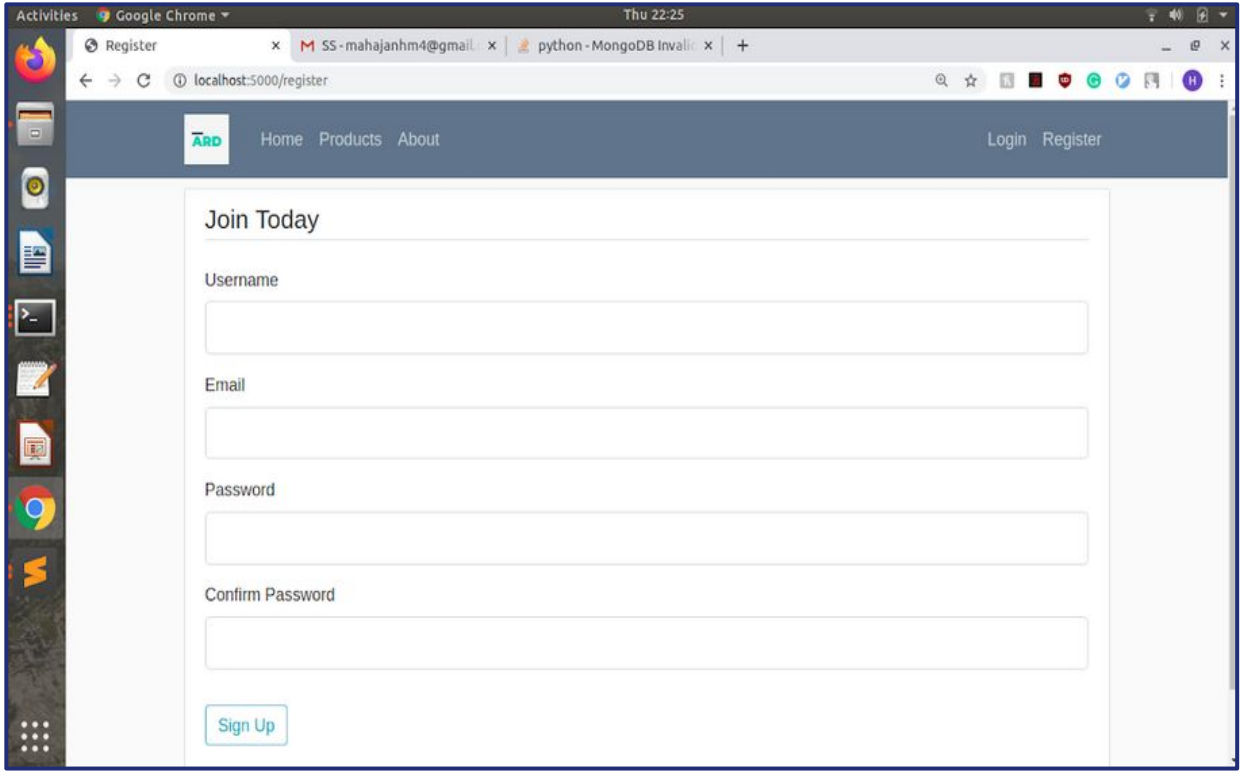
Big Data Analytics
Mini Project - Sentiment Analysis

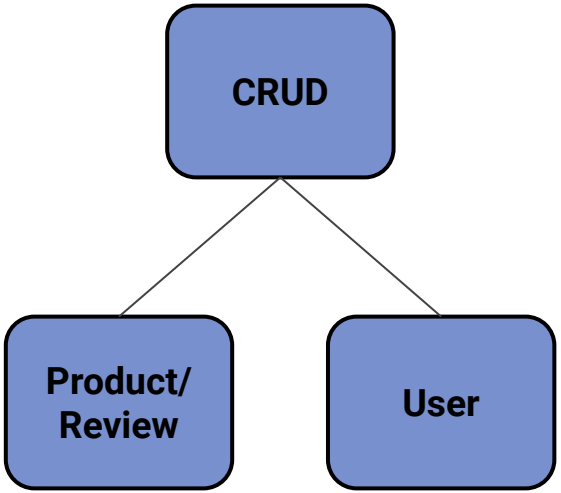
PA - 35 Raghav Gaggar
PC - 34 Siddharth Srivastava
PC - 39 Yash Sharma
PC - 42 Yash Shekhadar
PC - 45 Hrishikesh Mahajan

ABSTRACT

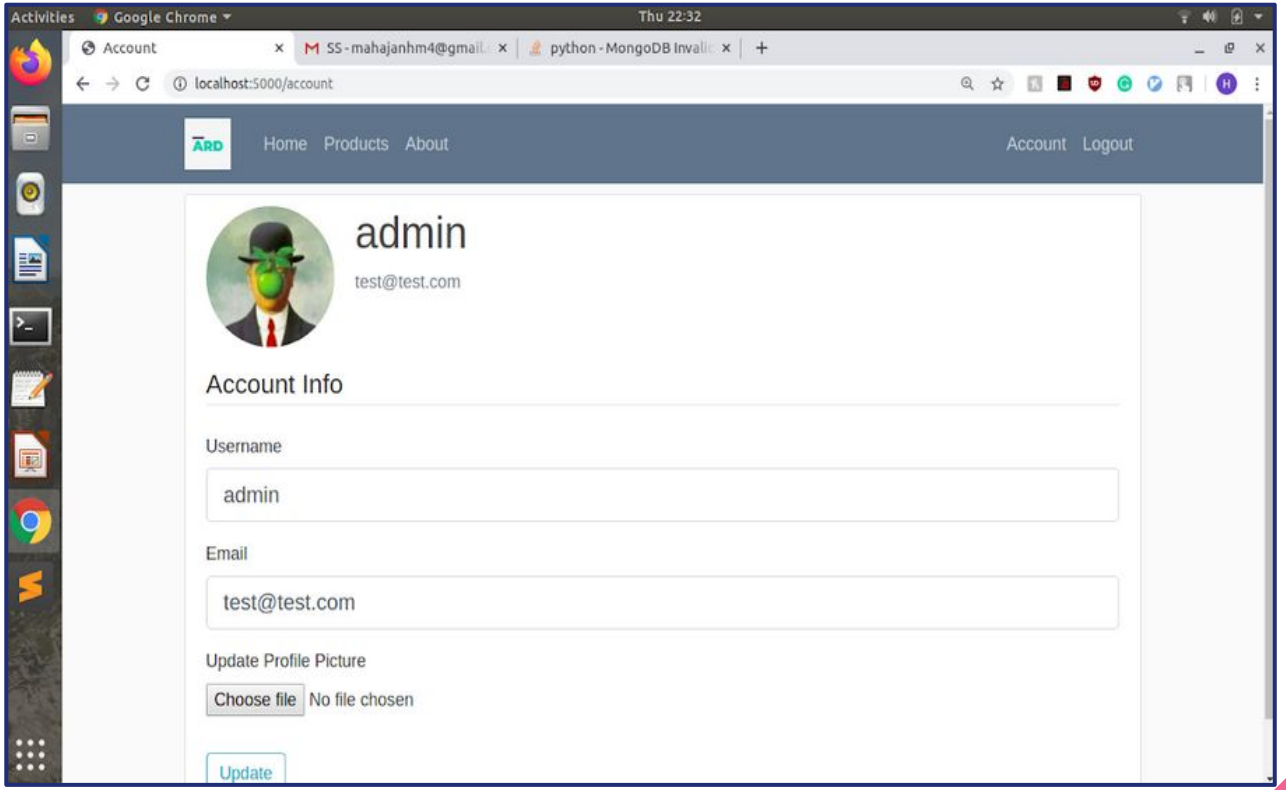
We present a pipeline suitable for big data and showcase the framework on an example of sentiment analysis as a machine learning task. We use the Amazon Appliances Review Dataset (http://deepyeti.ucsd.edu/jianmo/amazon/index.html) and do much of our pre-processing using the MapReduce framework. We present a dashboard-based front-end, through which we will demo the CRUD operations and present our results from the algorithms that we applied on the data. We conclude our work by reporting some visualizations engendered from our analysis, and delineate a future line of work.

CRUD Operations





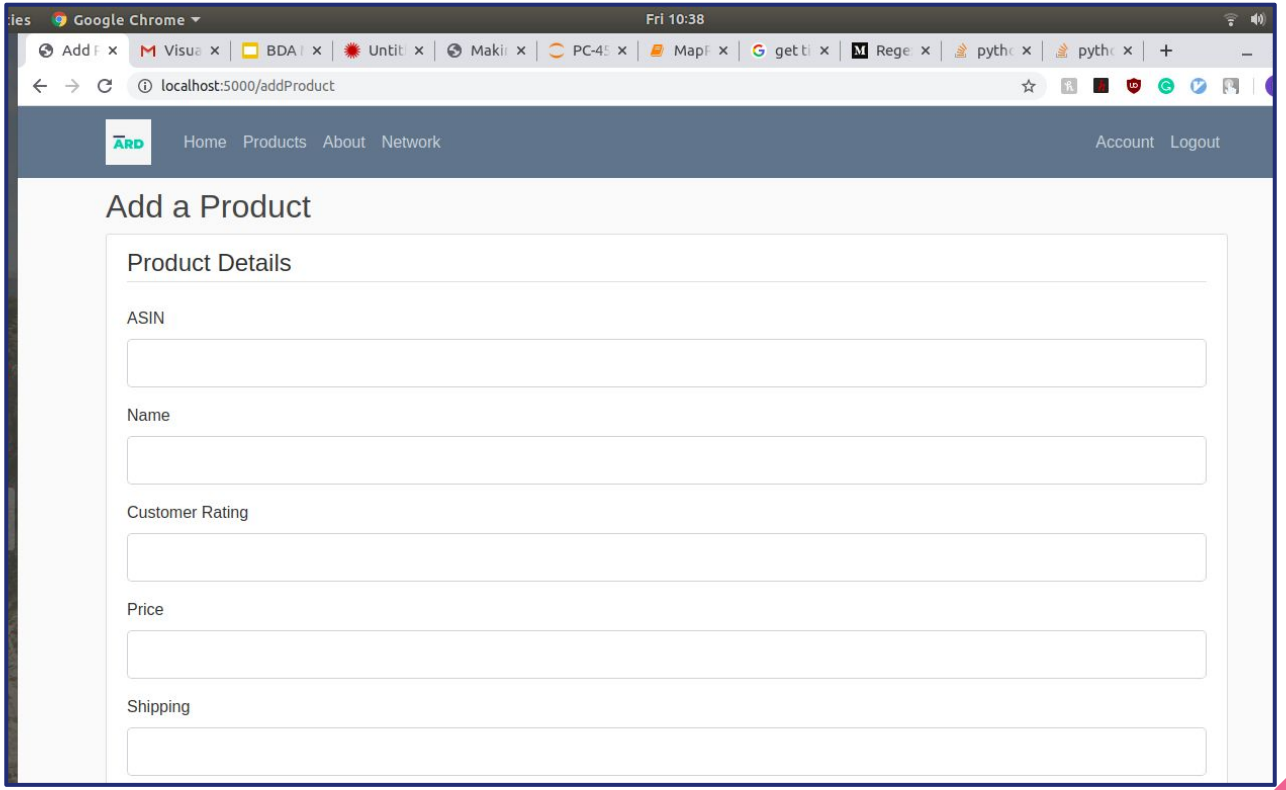
User profile facilitated by CRUD



Flask

Once the new user is inserted into database, he/she can view and/or update his/her profile.

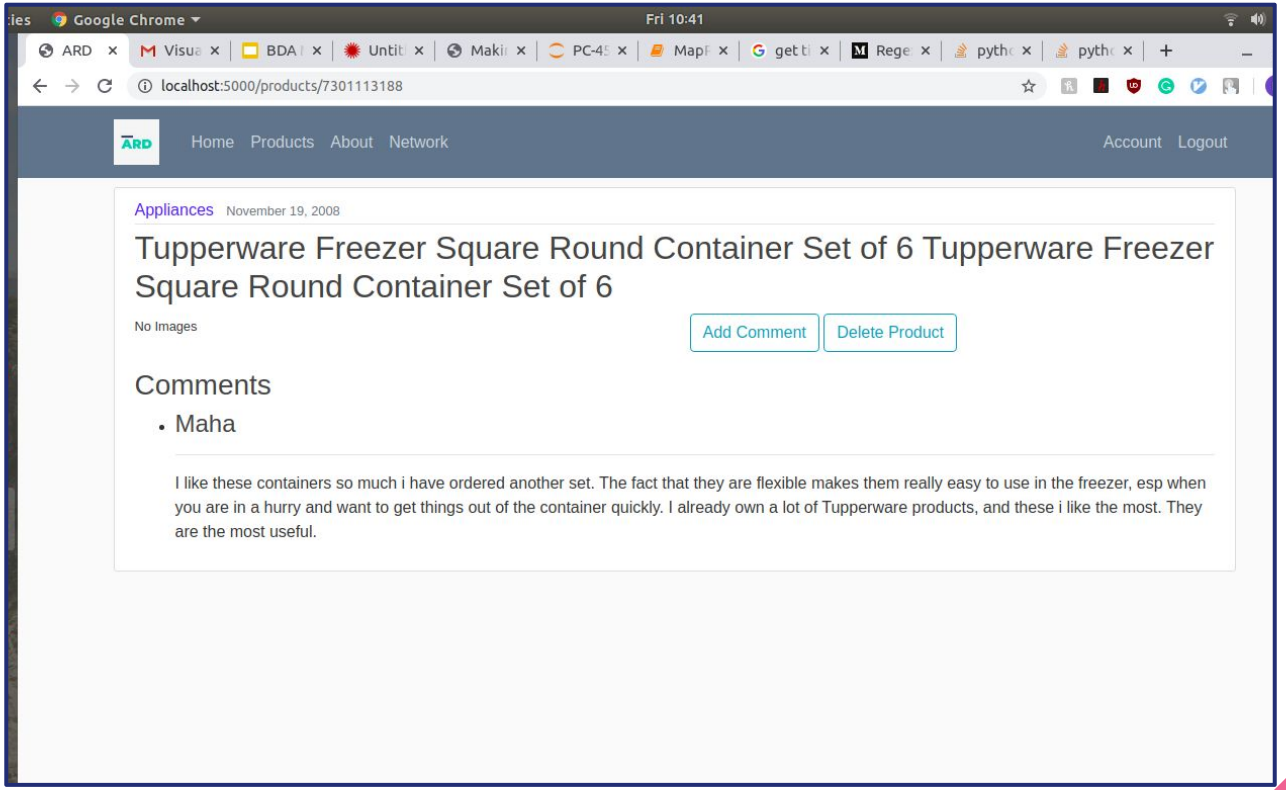
Adding new Products



Flask

Users can add new products by filling up the forms, after submission user will be directed to the product page.

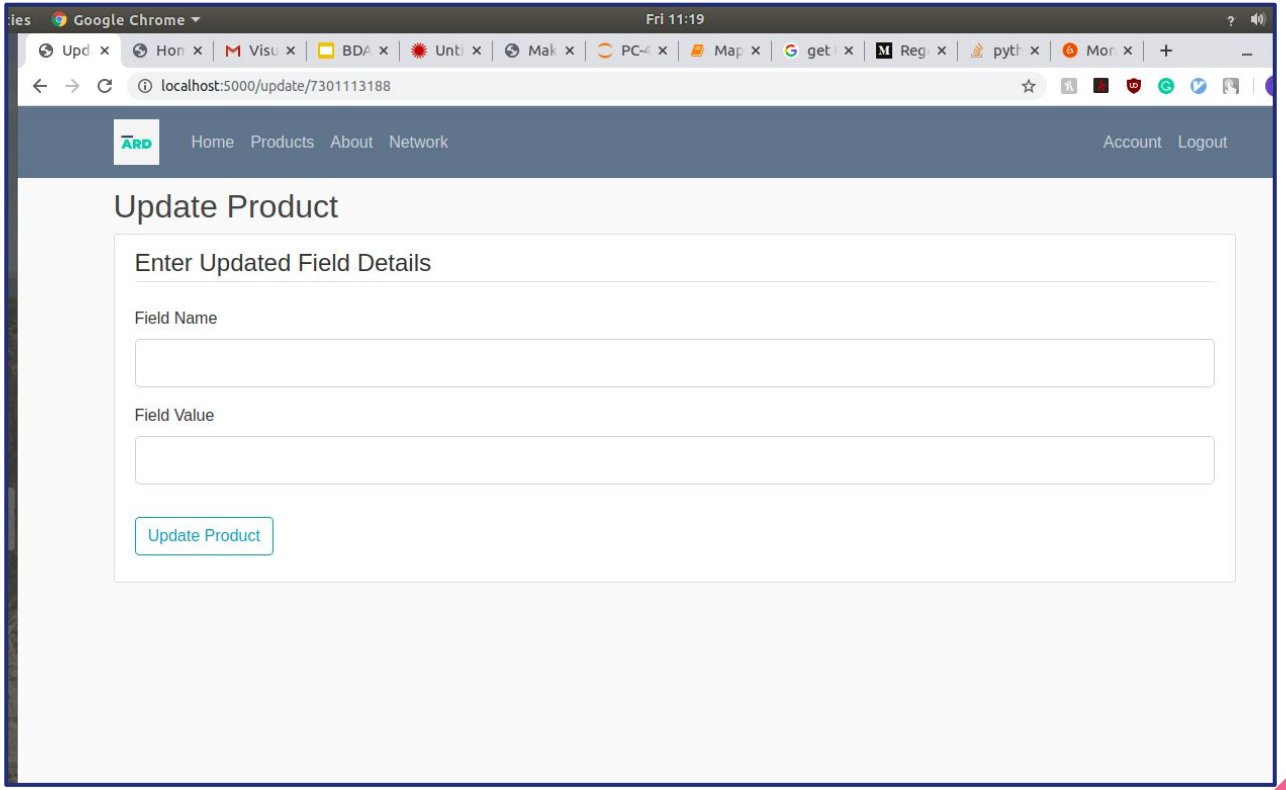
Adding Comments, Deleting Products



Every product page will allow the user to add comments to the product and those comments will be listed below.

Products can be deleted from the database after a small confirmation message.

Updating Product Values



Flask

Users can update the various value fields present in the database to their liking.

MapReduce

- MapReduce component was used to query 10 highest occuring words in the user submitted review.
- The algorithm scanned textual data of around 602k reviews.

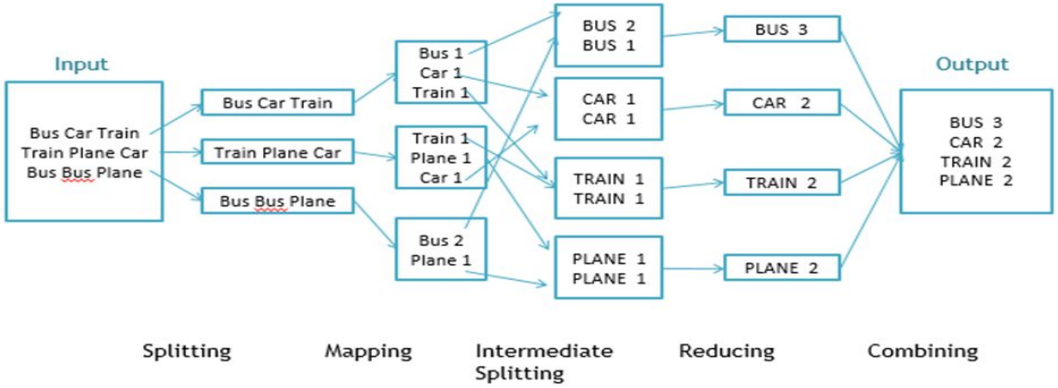
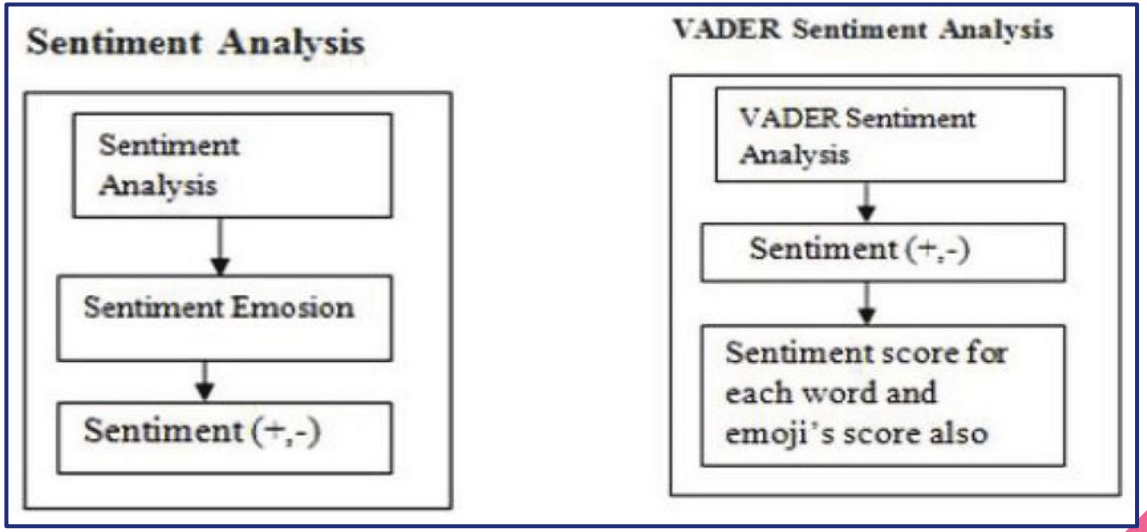


Fig. WorkFlow of MapReducing

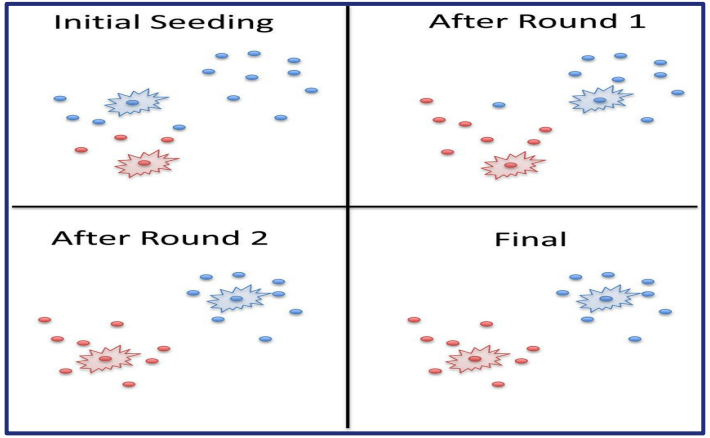
Machine Learning

- **Phase 1 of our pipeline**: we used VADER (Valence Aware Dictionary and sEntimenter Reasoner) for sentiment analysis.
- VADER gives a score which is amenable to sentiment analysis



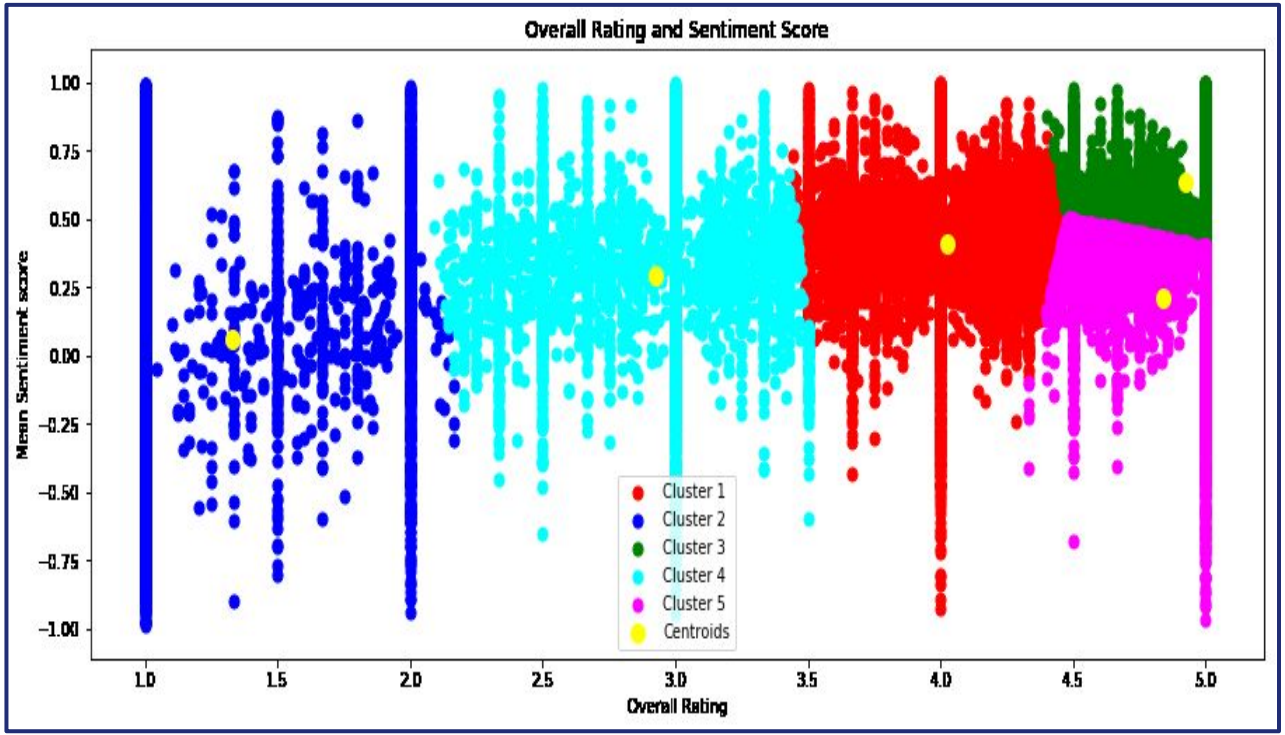
Phase 2 of pipeline: K-means

- We use the K-means clustering algorithm to form clusters of products.
- We used the overall rating of the product and the sentiment associated to form distinct clusters.



- We report the results of the cluster analysis in the next slide:

K-means results

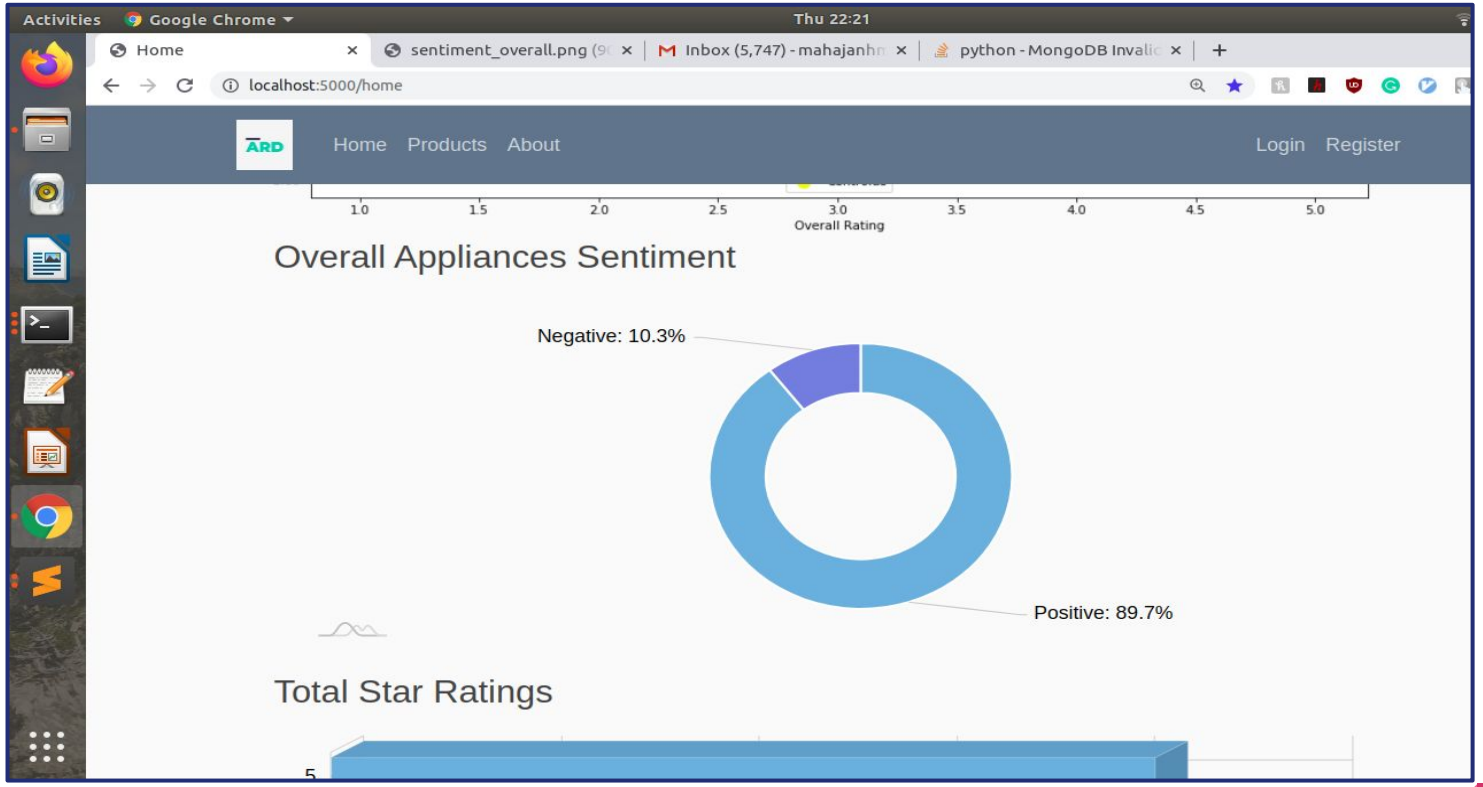


Output Clusters generated from K-Means Clustering algorithm.

Overall ratings of products and sentiments were used to form the clusters.

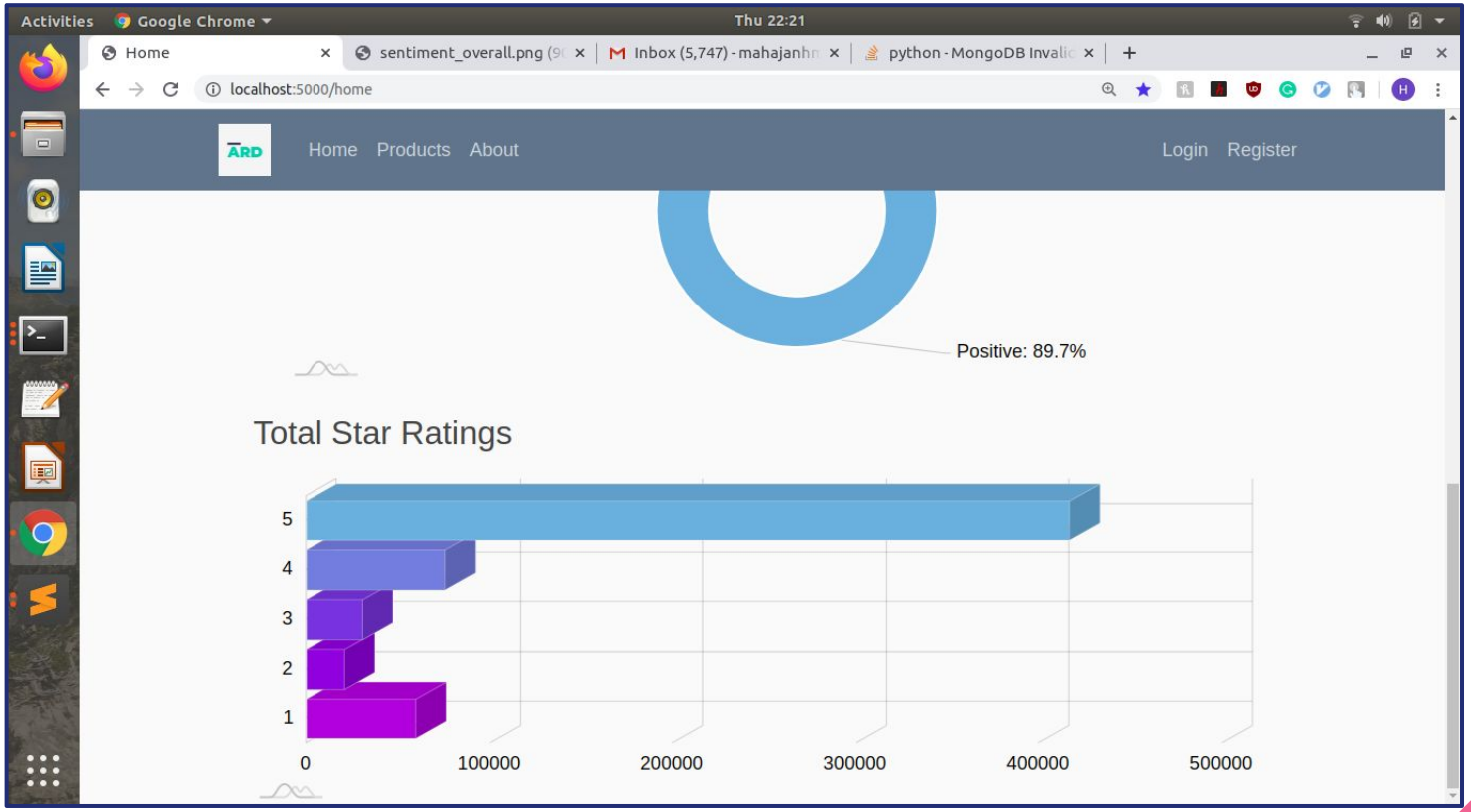
Visualization

- Visualization reports obtained were generated from either the outcome of machine learning algorithm or directly queried from the dataset.
- To ease out the process of visualization, **custom json** and **csv** files were generated by Python code snippets.

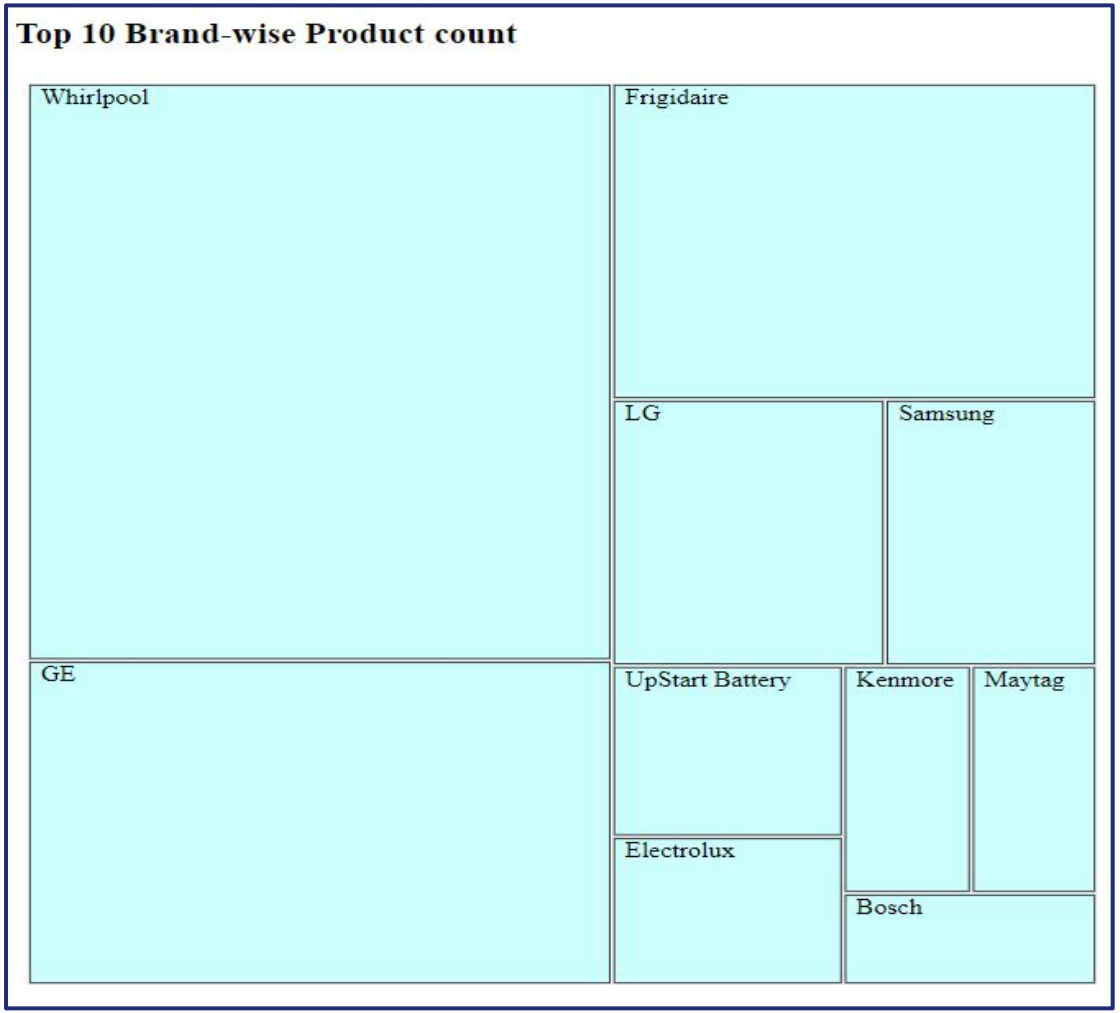


AMCHARTS

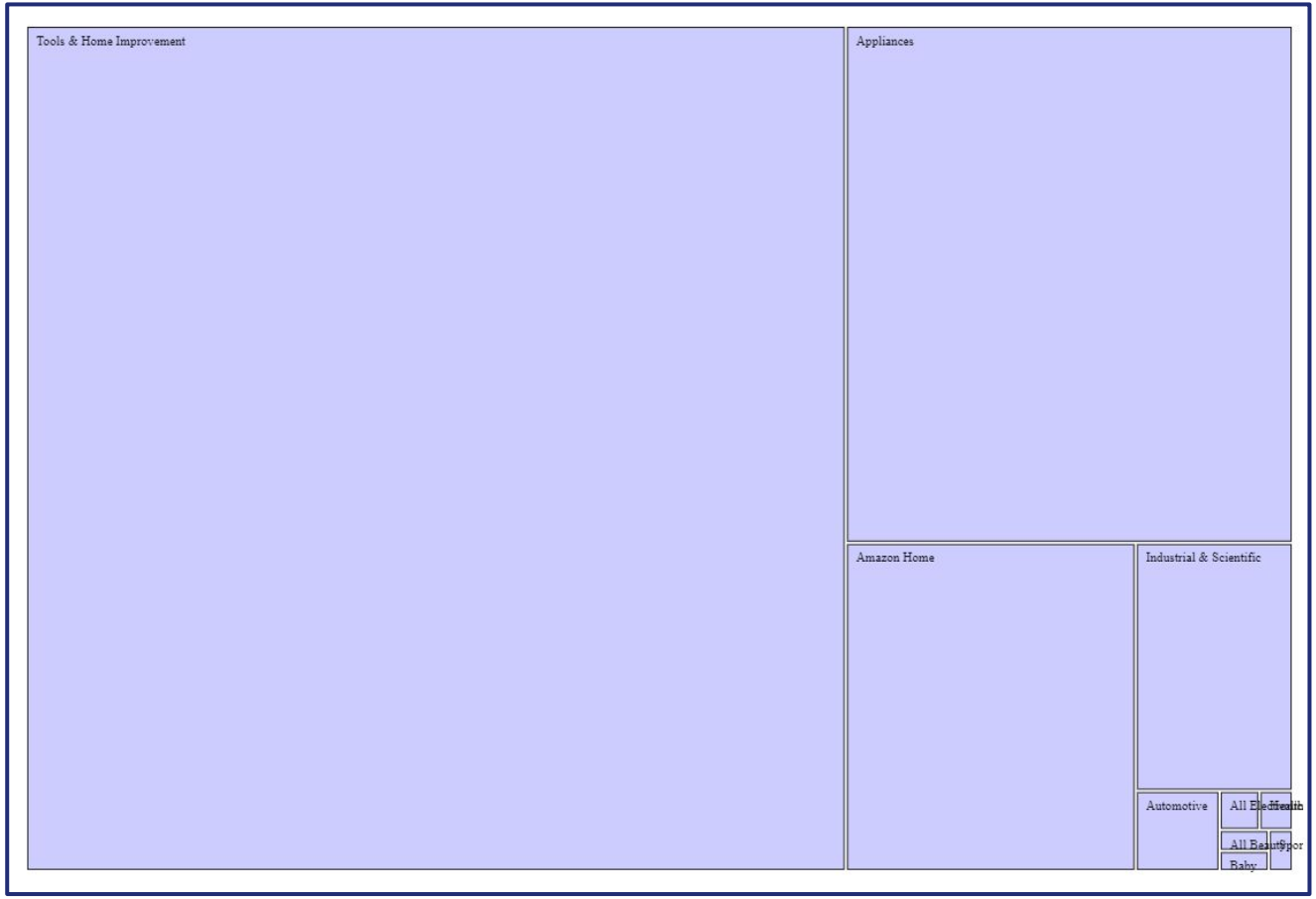
Pie Chart depicting Percentage of **Positive** and **Negative** Sentiment obtained from analysing the User reviews



3-D Bar chart to visualize the overall customer rating of products



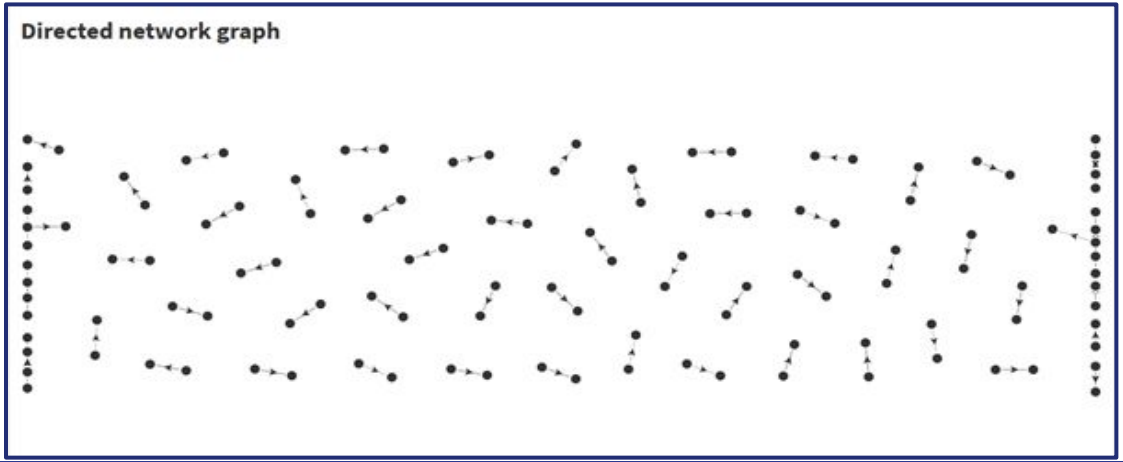
Treemap to visualize the **Top 10 brands** by **Product Count**



Treemap to visualize **Top 10 most reviewed category** of appliances

Flourish

Directed network graph to show relationship between a base product and the corresponding recommended products.



E.g. - Product:Universal Metal Industries Range Hood Grease Filter. Recommended:Broan BP58 Non-Ducted Charcoal Replacement Filter Pads for Range Hood

FUTURE WORK

- Disentangling big data to gain more meaning. Eg.: detect sarcasm accurately to circumvent spurious analysis.
- Visualize and track how useful our recommendations have been, and consequently provide better recommendations based on that.
- Exploratory Data Analytics can be performed on the initial dataset to curate it, if we collect raw data ourselves.
- Collecting opinions on the web will still require processing that can filter out un-opinionated user-generated content and also to test the trustworthiness of the opinion and its source.

Thank You!
Any Questions?