

Big Data Analytics

Mini Project - Sentiment Analysis

PA - 35 Raghav Gaggar
PC - 34 Siddharth Srivastava
PC - 39 Yash Sharma
PC - 42 Yash Shekhar
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



Flask

Activities Google Chrome Thu 22:25

Register x SS-mahajanhm4@gmail x python - MongoDB Invalid x +

localhost:5000/register

ARD Home Products About Login Register

Join Today

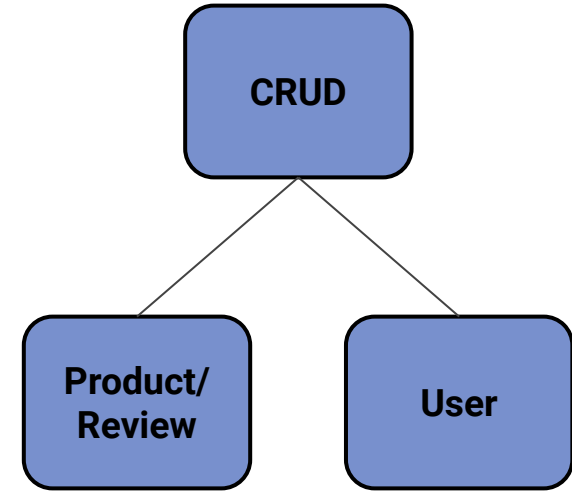
Username

Email

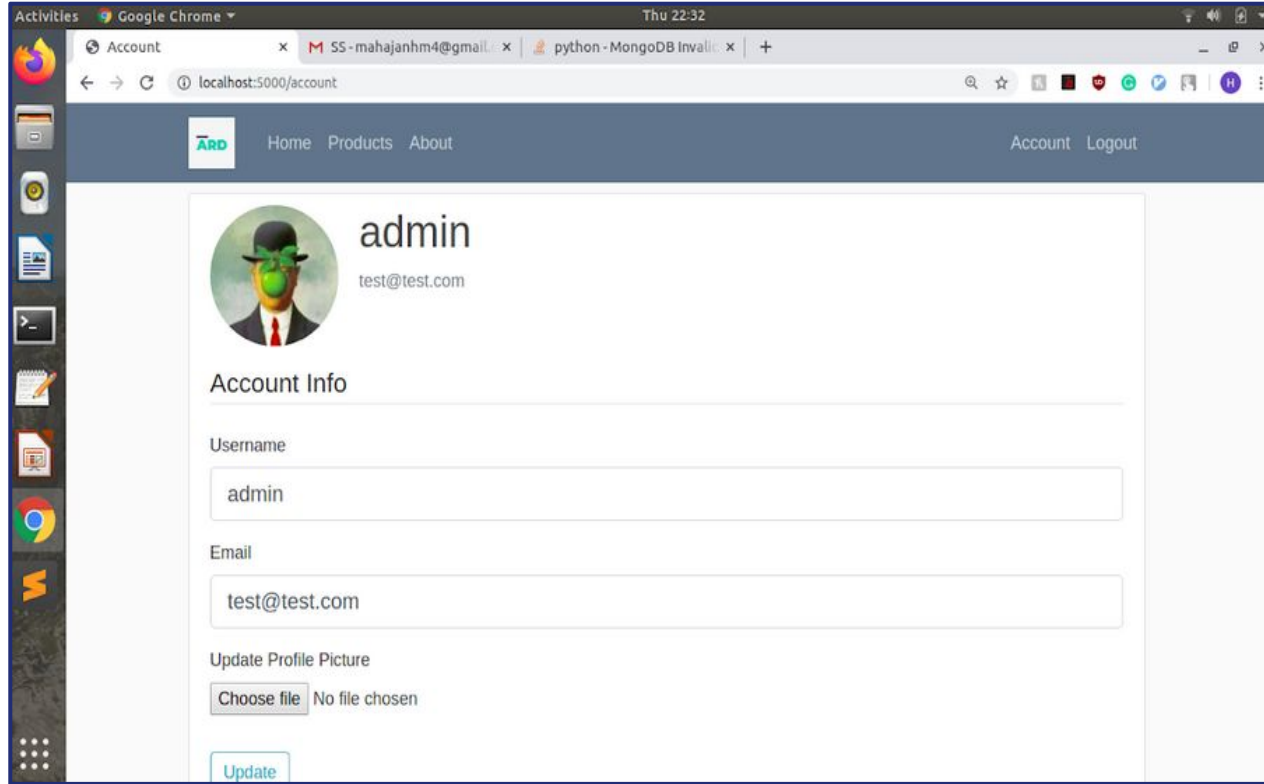
Password

Confirm Password

Sign Up

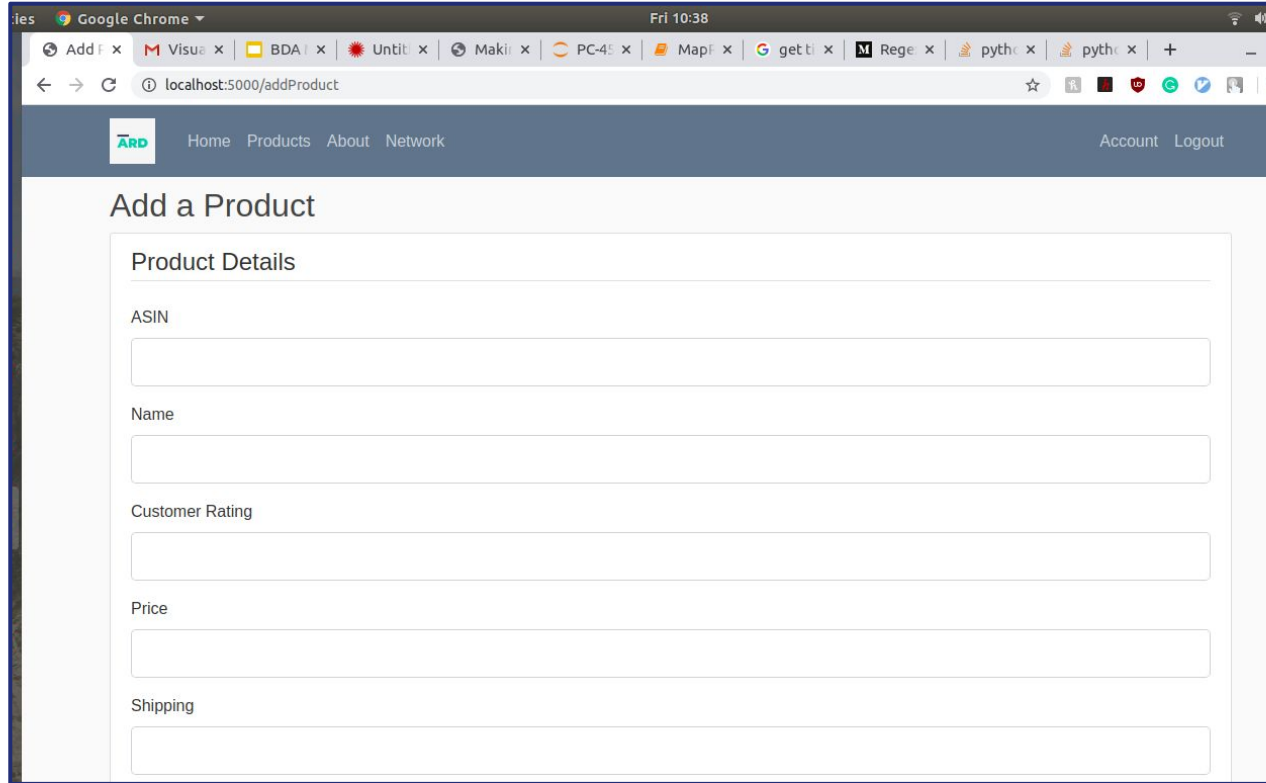


User profile facilitated by CRUD



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

Adding new Products



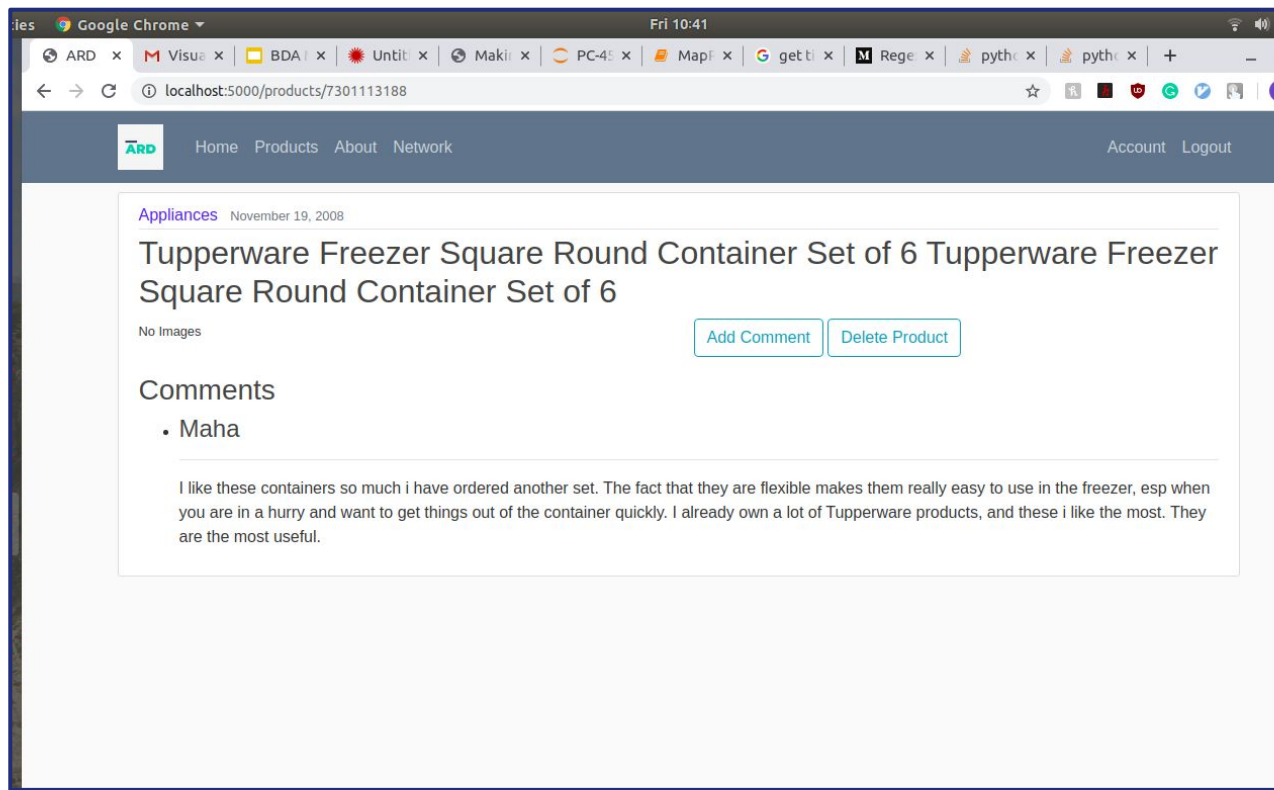
The screenshot shows a web browser window with the address bar at `localhost:5000/addProduct`. The page has a dark blue header with a logo on the left and navigation links: Home, Products, About, Network, Account, and Logout. The main content area is titled 'Add a Product' and contains a form with the following fields:

- Product Details
- ASIN
- Name
- Customer Rating
- Price
- Shipping



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

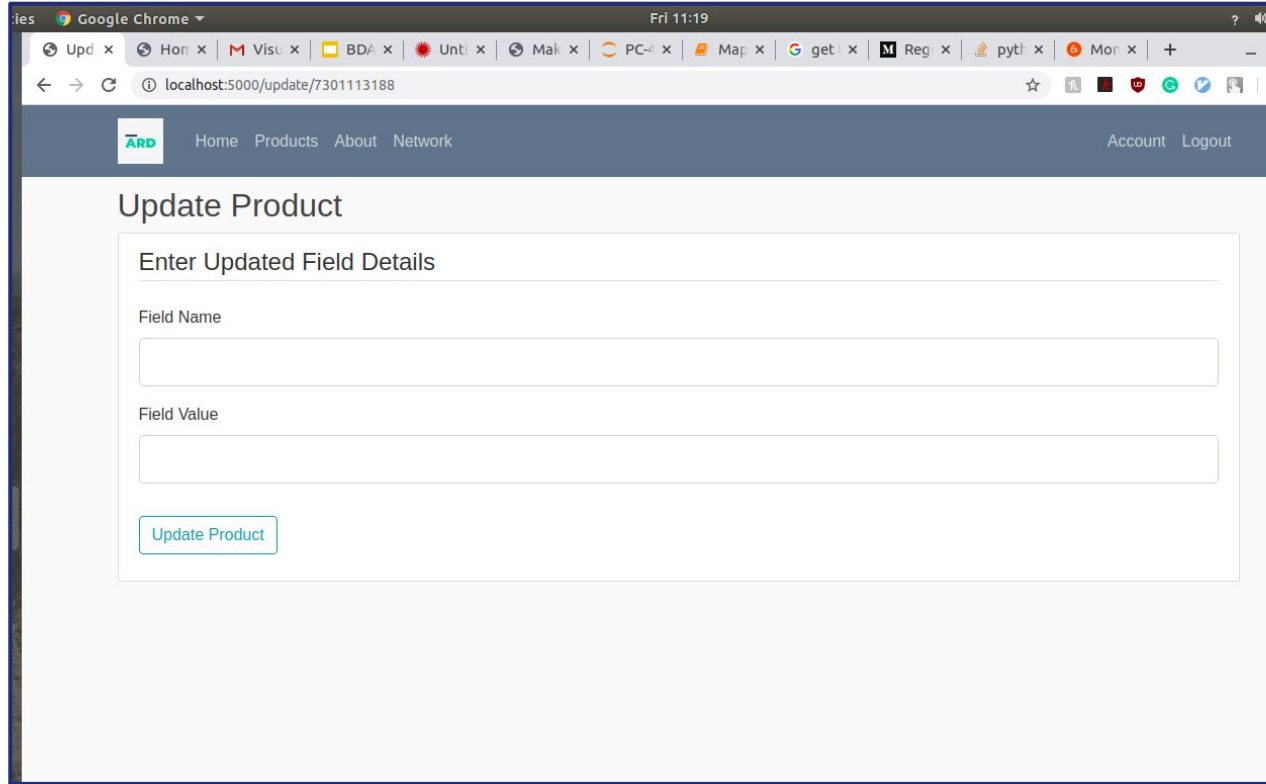
Adding Comments, Deleting Products



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



The screenshot shows a web browser window with the URL `localhost:5000/update/7301113188`. The page has a dark blue header with the ARD logo and navigation links: Home, Products, About, Network, Account, and Logout. The main content area is titled "Update Product" and contains a form with the heading "Enter Updated Field Details". The form has two input fields: "Field Name" and "Field Value". Below the fields is a blue button labeled "Update Product".

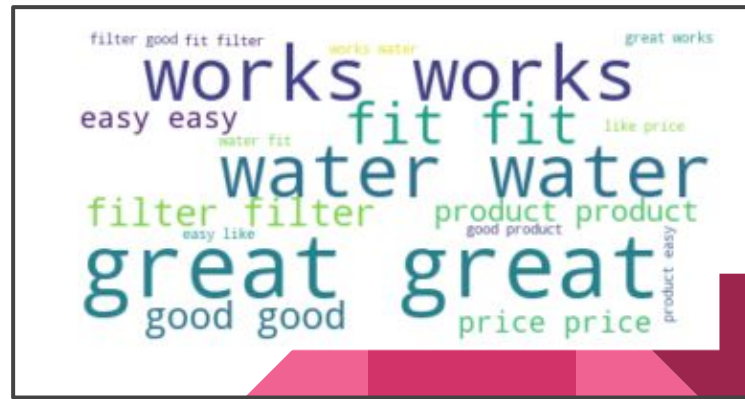
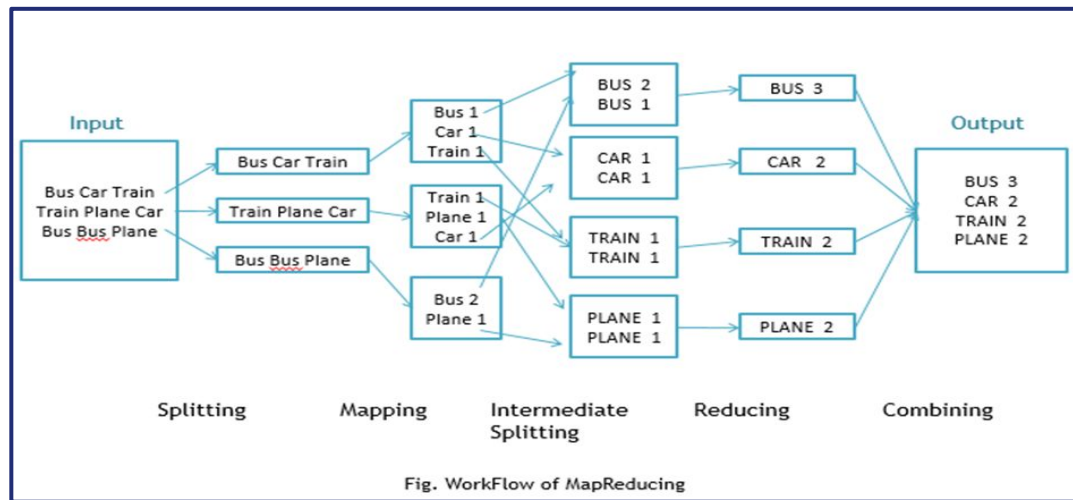


Users can update the various value fields present in the database to their liking.



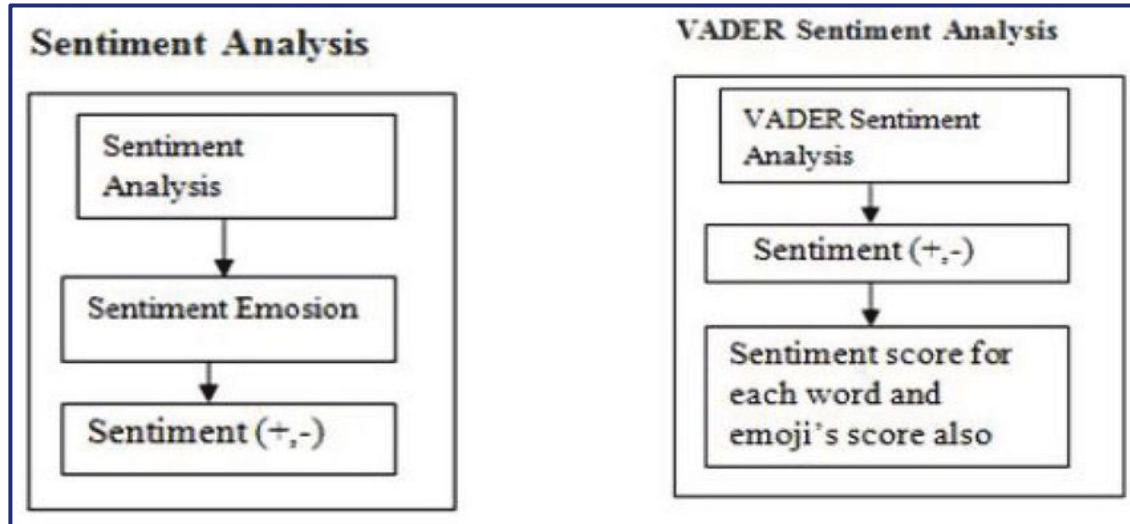
MapReduce

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



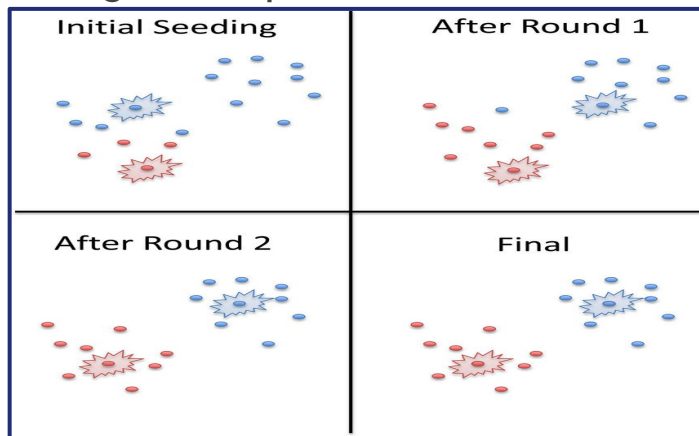
Machine Learning

- **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



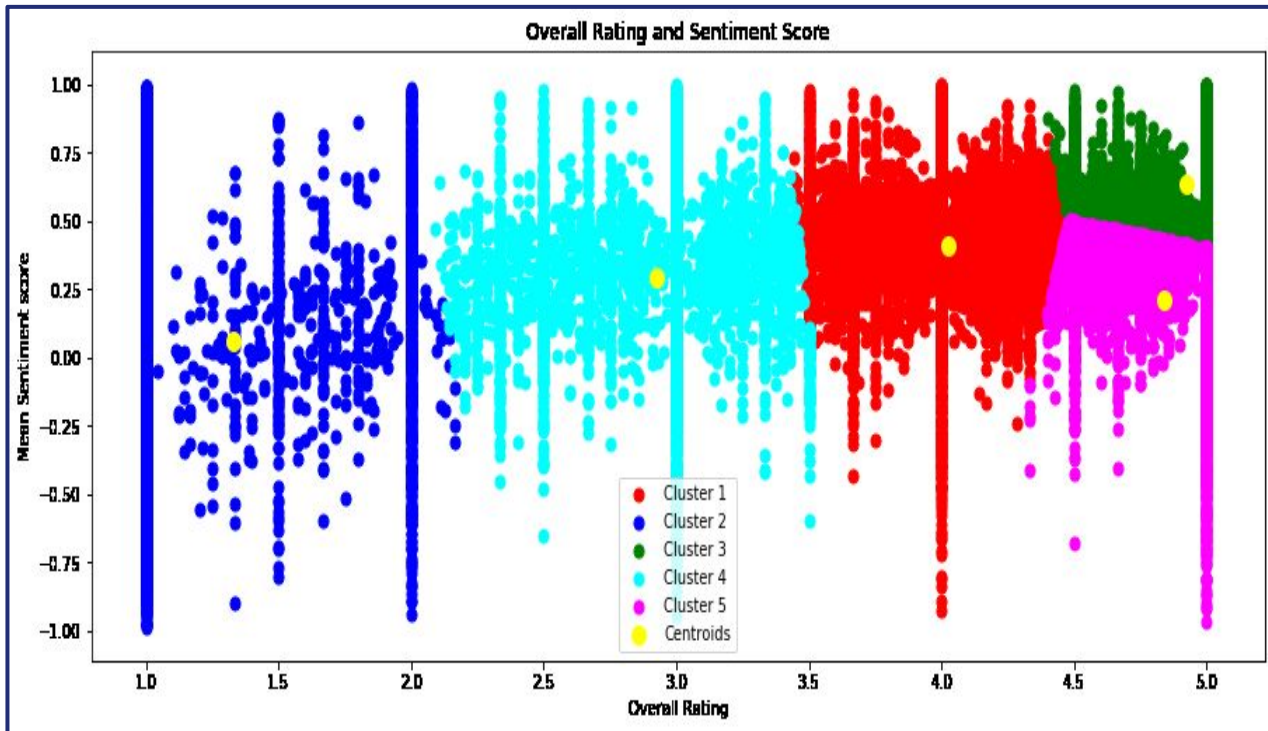
Phase 2 of pipeline: K-means

- 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



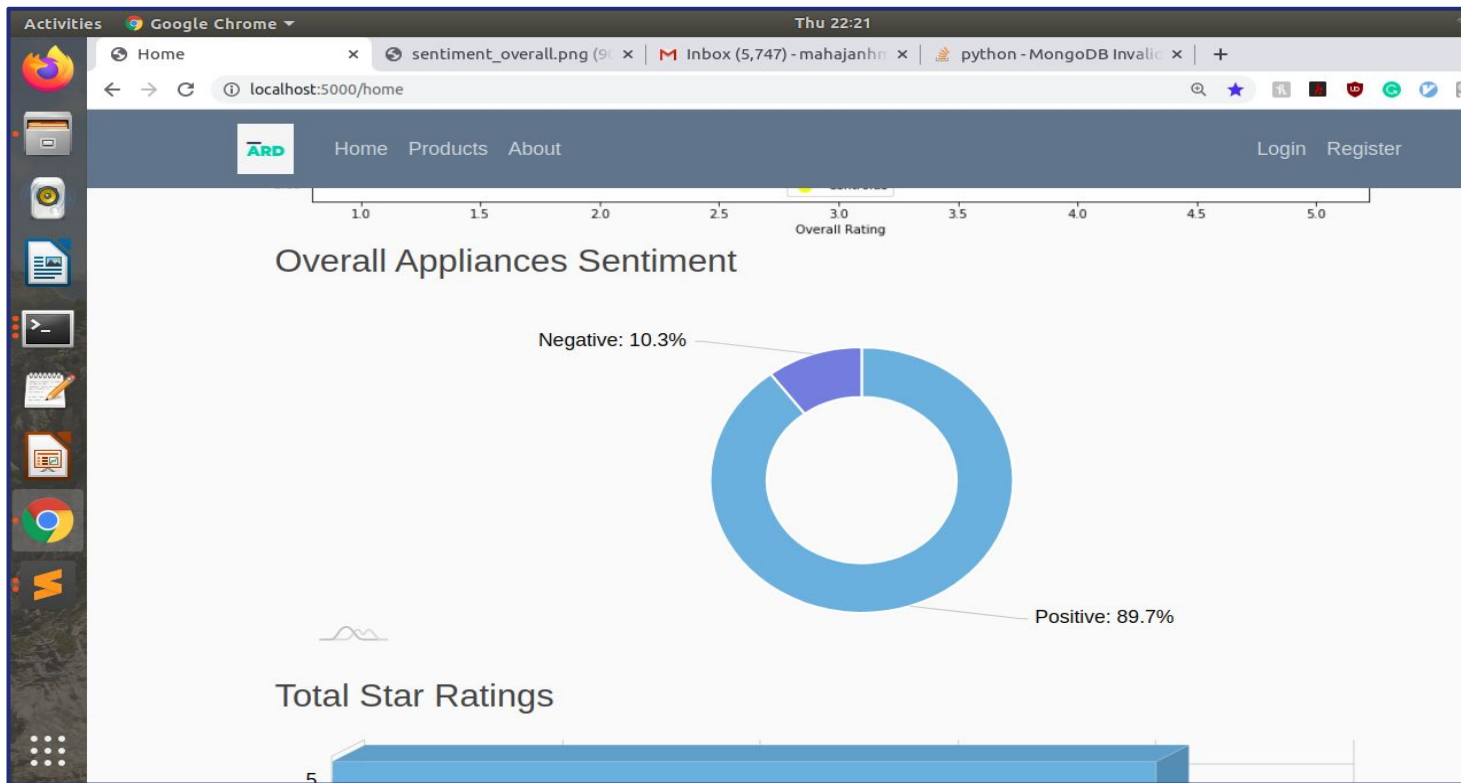
Output Clusters generated from K-Means Clustering algorithm.

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

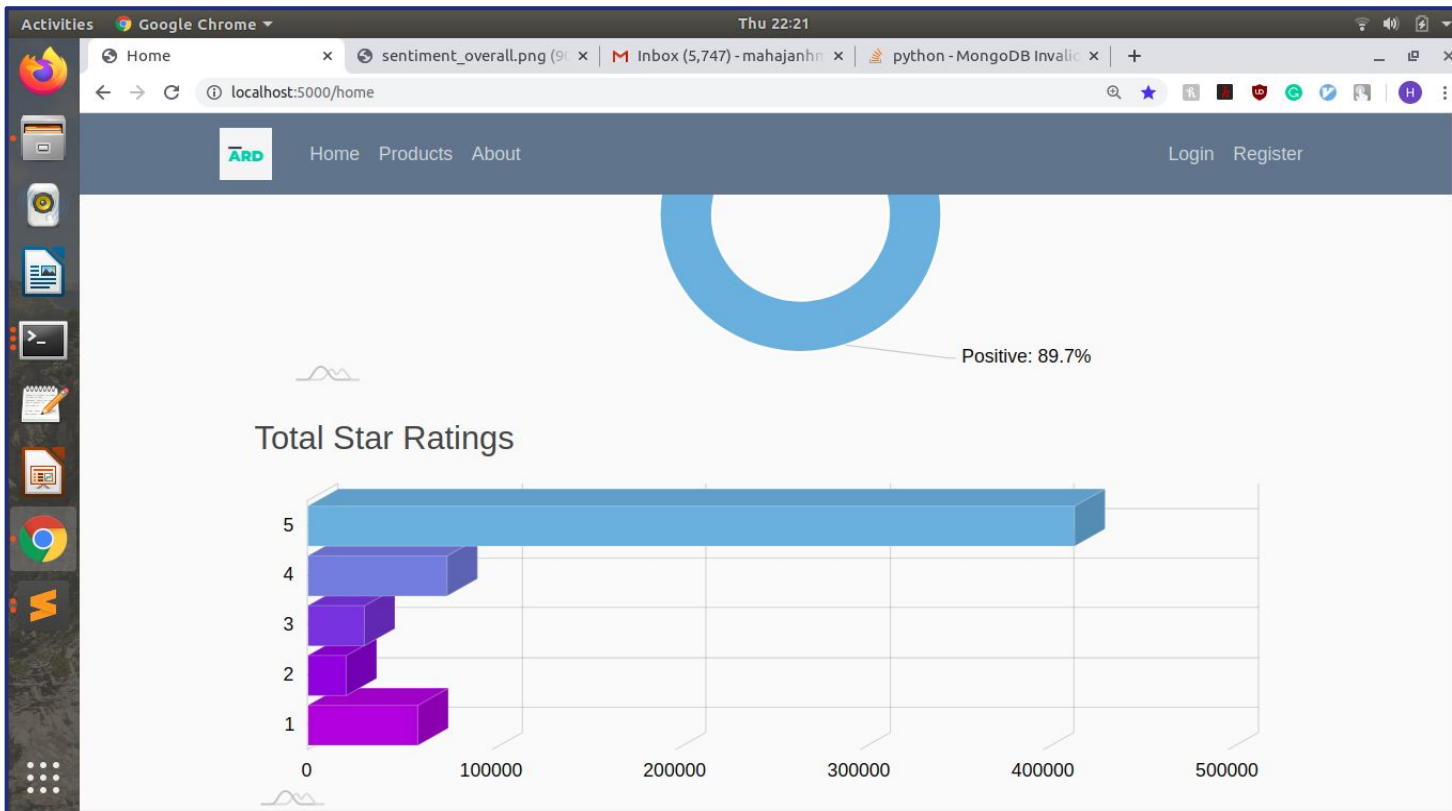
Visualization

- 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.



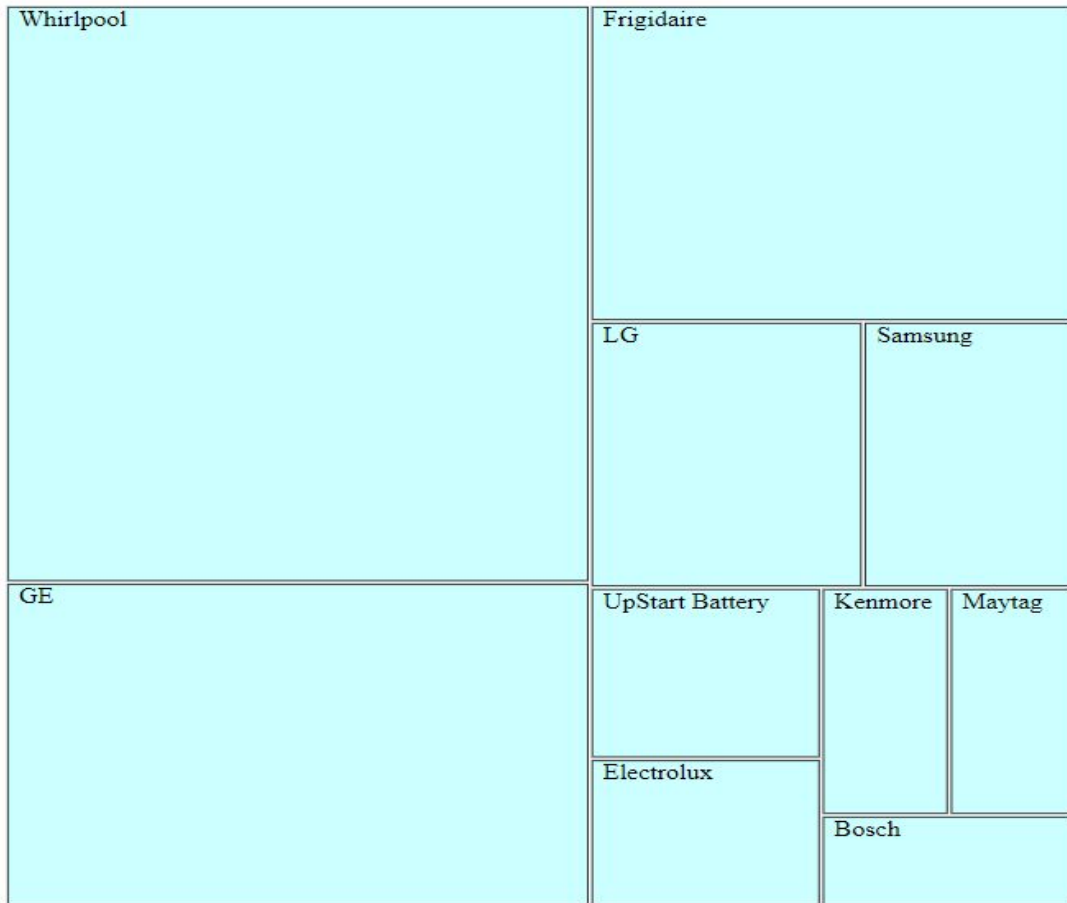


Pie Chart depicting Percentage of **Positive** and **Negative** Sentiment obtained from analysing the User reviews

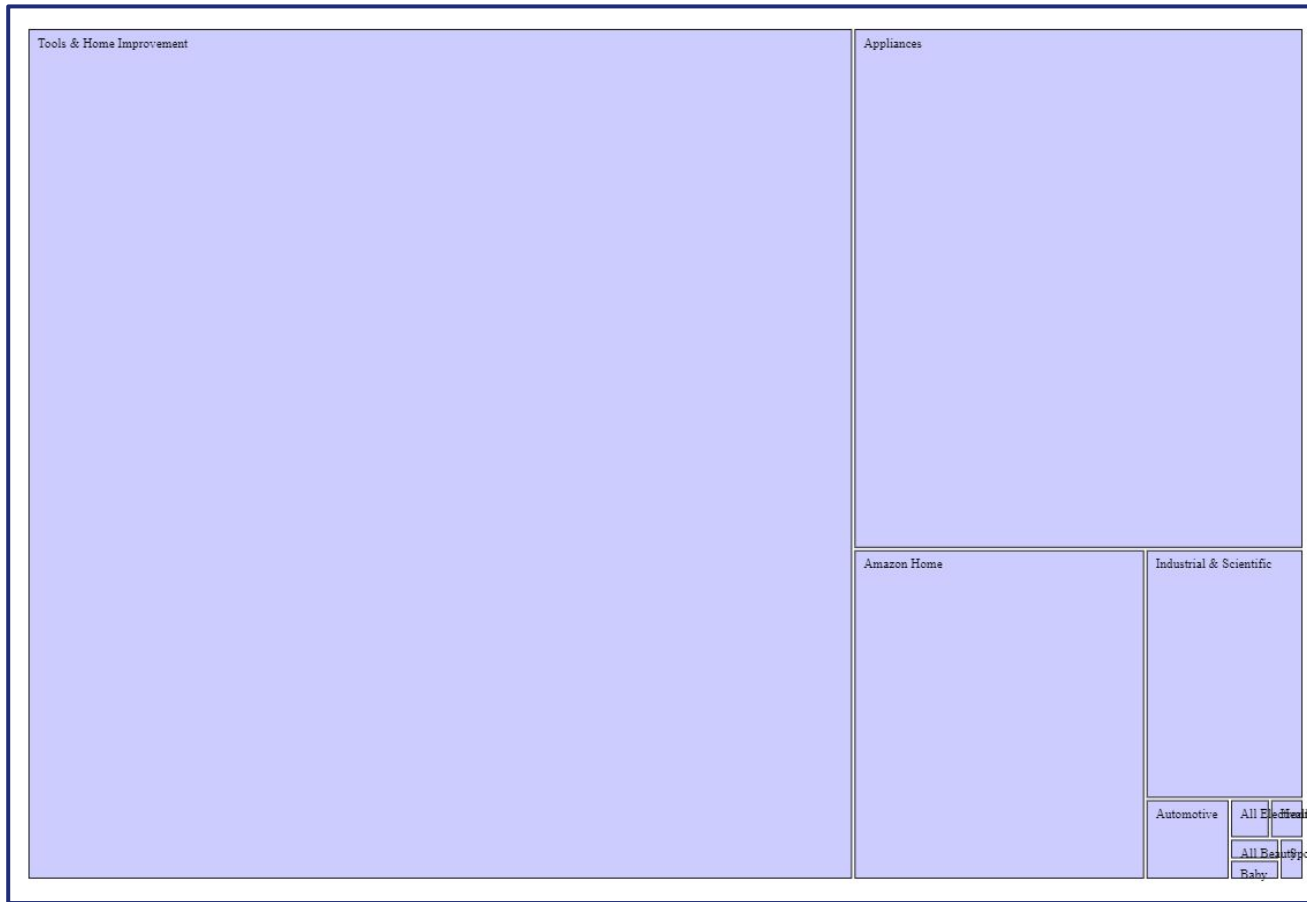


3-D Bar chart to visualize the overall customer rating of products

Top 10 Brand-wise Product count



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

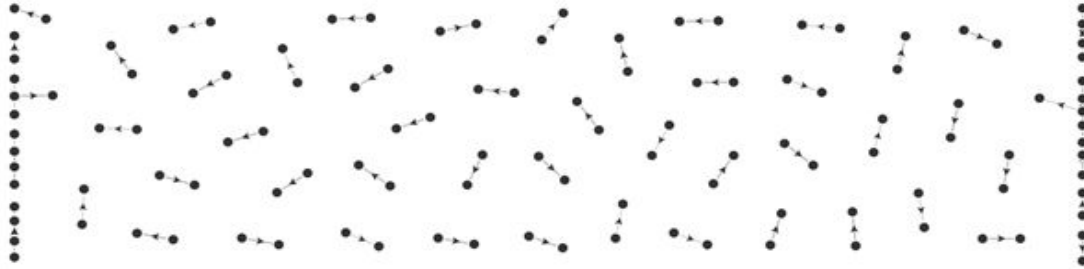


Treemap to visualize
**Top 10 most reviewed
category** of
appliances

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


```
[{"B009ERF1WM", "B0053F9EKE", "B075FDKDV4", "B01BRPCZKO", "B000NCTO9M", "B00I2XJGAW", "B0053F7UIM", "B0053F9HTI", "B002YT0CP4", "B000ZU9S3C"}  
B00WHPICFG
```


Directed network graph



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?