Data Visualization - PA 2 - Mohit Surana

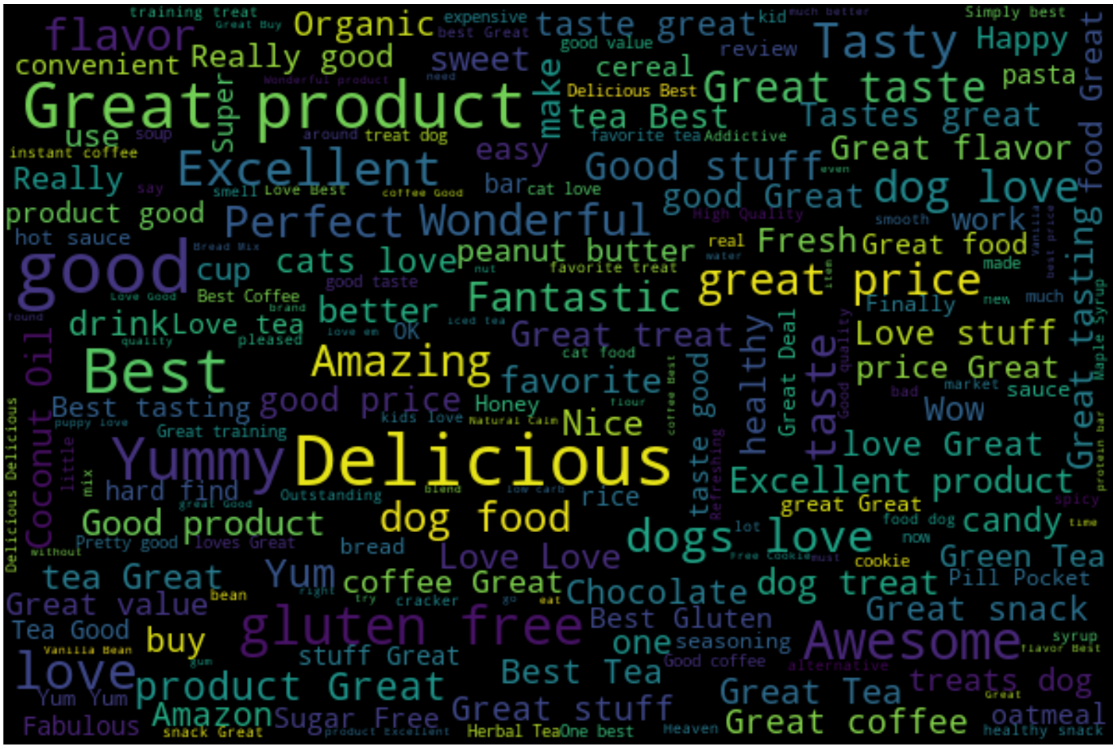
Introduction

The dataset I chose was the Amazon Fine Foods reviews which is a dataset of about 500,000 reviews with information like productId, review score, summary, text and a few other fields. I picked this because of the availability of the additional metadata in the form of review text. Most of the other interesting datasets (eg. California roads, Google Web Pages) only contained edges in the form of nodeIds which would not allow us to do any meaningful interpretation without a mapping to the original road names or webpages.

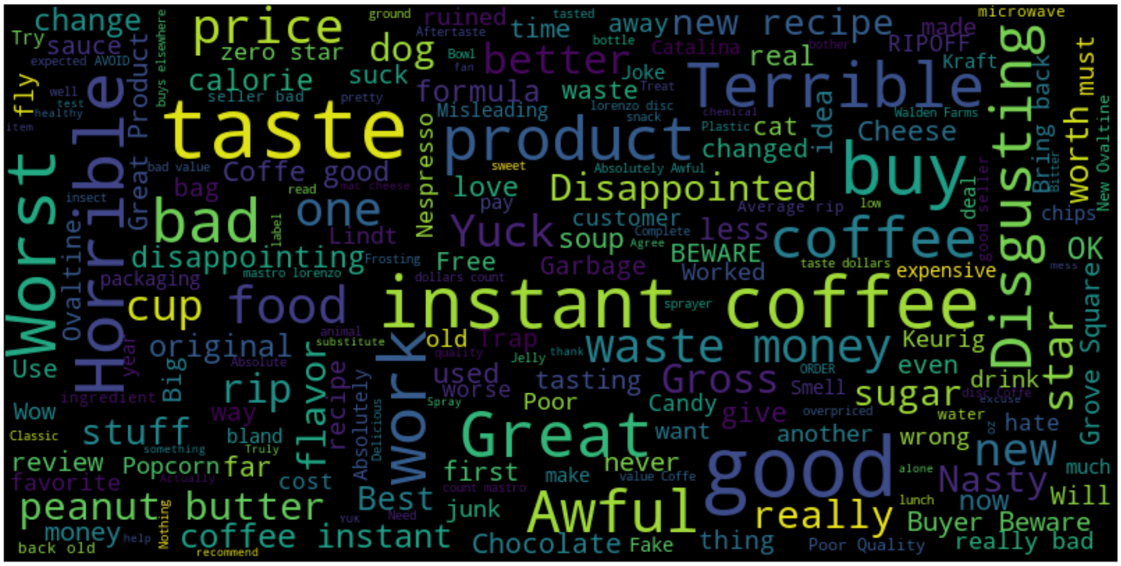
Data sample

```
product/productId: B001E4KFG0
review/userId: A3SGXH7AUHU8GW
review/profileName: delmartian
review/helpfulness: 1/1
review/score: 5.0
review/time: 1303862400
review/summary: Good Quality Dog Food
review/text: I have bought several of the Vitality canned dog food products and have
found them all to be of good quality. The product looks more like a stew than a
processed meat and it smells better. My Labrador is finicky and she appreciates this
product better than most.
```

Top words from Good Reviews



Top words from Bad Reviews



Methodology & Observations

To get good quality data, I plotted the distribution of the number of reviews per product, and decided to filter out all the products that had very few reviews (less than 5).

After that, I move on to my primary task: Insights through text patterns in items reviews

I calculated the average review score and picked the best and worst reviewed items. Then I transformed their corresponding reviews and summaries into a WordCloud. On studying the WordCloud, it seems that a lot of words are overlapping or generic which indicate that the use of WordCloud directly is not very meaningful. We have the productId only. In case we had the product name as well, we could have run the word cloud on the name so that the words in the WordCloud would mainly be nouns/names of products.

The data does show certain points that have business value:
Eg. The high appearance of Gluten free, dog food and peanut butter in the good review WordCloud indicate strong markets and customer satisfaction.
In the bad review WordCloud, if you look past the overlapping items, you notice that instant coffee (especially Keurig) is causing people to be unhappy and that could be pitched as an area of improvement.

Scope for future work

The perfect follow up to this would be a way to look at a drill down version of the data.
Lot of people commented **Disgusting** in the reviews. What products were responsible for this observation? In other words, the visualization currently highlights areas people ought to look at, and another tool would help to find out exactly what was happening under the hood.

Thank you!