Task - Sentiment Analysis on Amazon Review Data

Dataset Link - txt_reviews.zip

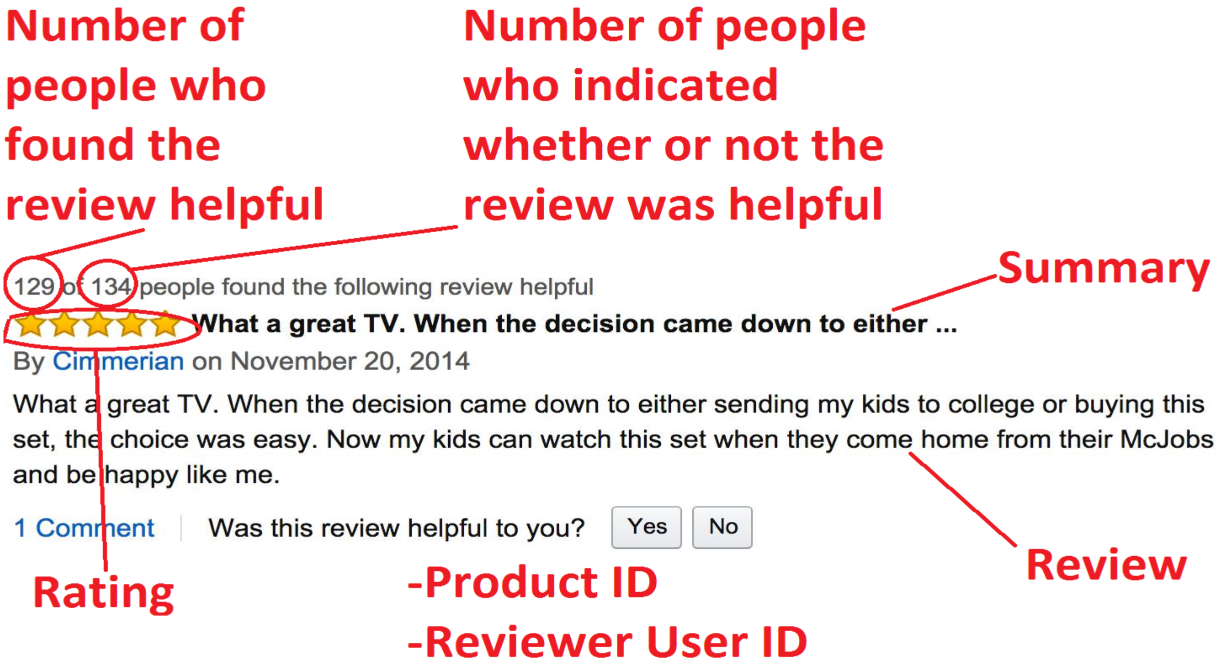
Data Description

This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories.

Data includes:

- Reviews from Oct 1999 - Oct 2012 - 568,454 reviews
- 256,059 Users and 74,258 products
- 260 users with > 50 reviews

Below attached is the screenshot of product review from Amazon Website.



Number of people who found the review helpful
Number of people who indicated whether or not the review was helpful
Summary
Rating
-Product ID
-Reviewer User ID
Review

129 of 134 people found the following review helpful

★★★★★ **What a great TV. When the decision came down to either ...**

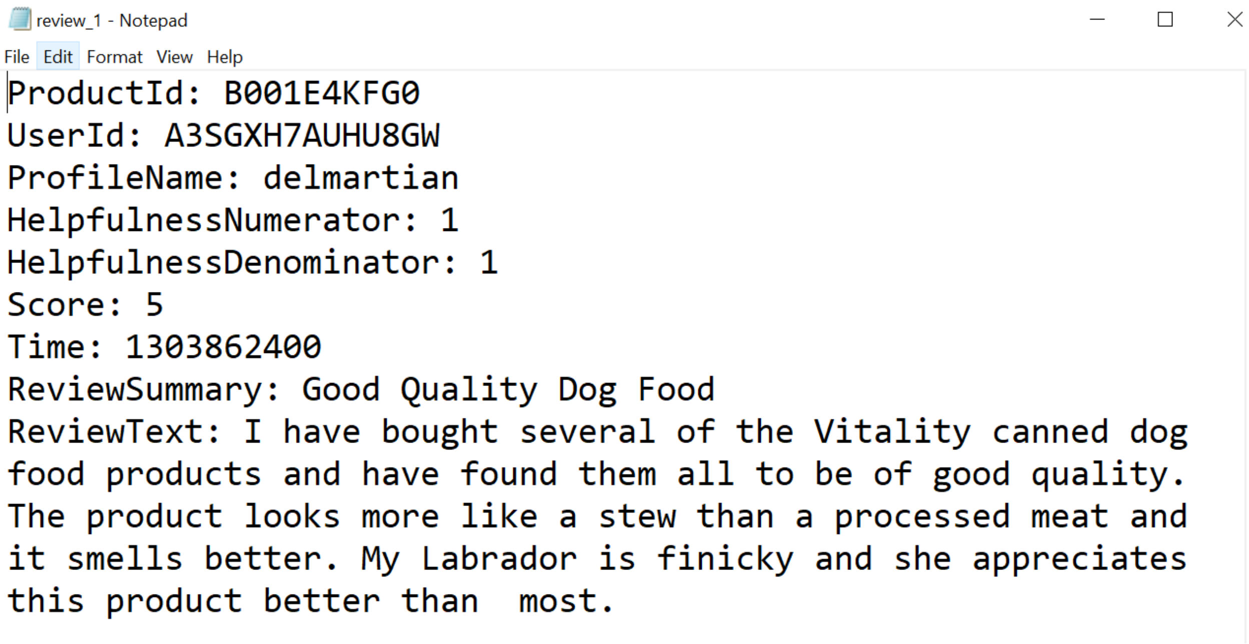
By Cimmerian on November 20, 2014

What a great TV. When the decision came down to either sending my kids to college or buying this set, the choice was easy. Now my kids can watch this set when they come home from their McJobs and be happy like me.

1 Comment | Was this review helpful to you? Yes No

SPRINT - 1

Given data consists of 568,454 text files. Each text file looks like the below attached image:



```
ProductId: B001E4KFG0
UserId: A3SGXH7AUHU8GW
ProfileName: delmartian
HelpfulnessNumerator: 1
HelpfulnessDenominator: 1
Score: 5
Time: 1303862400
ReviewSummary: Good Quality Dog Food
ReviewText: 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.
```

Your task here is to transform the given data(i.e. Text files) to tabular format(i.e. csv file). The columns in the table should be:

- Id - Unique row number
- ProductId - Unique identifier for the product
- UserId - Unique identifier for the user
- ProfileName
- HelpfulnessNumerator - Number of users who found the review helpful
- HelpfulnessDenominator - Number of users who indicated whether they found the review helpful
- Score - Rating between 1 and 5
- Time - Timestamp for the review
- ReviewSummary - Brief summary of the review
- ReviewText - Text of the review

NOTE - Helpfulness (fraction of users who found the review helpful) = HelpfulnessNumerator / HelpfulnessDenominator

SPRINT - 2

Work on the below mentioned Visualizations for exploratory data analysis:

1. Distribution of Ratings
2. Popular words in Positive Reviews (4-5 Rating)
3. Popular words in Negative Reviews (1-2 Rating)
4. Distribution of Helpfulness
5. How does rating affect Helpfulness?
6. How does word count vary by rating?
7. Etc…

Note - Use this blog written by Rob Castellano to understand the data analysis and how he generated insights (conclusion) from the visualizations.

SPRINT - 3

Build a model which takes the text review as input and predicts the rating of the review.