Text Data - Sentiment Analysis

Dataset Link - txt_reviews.zip

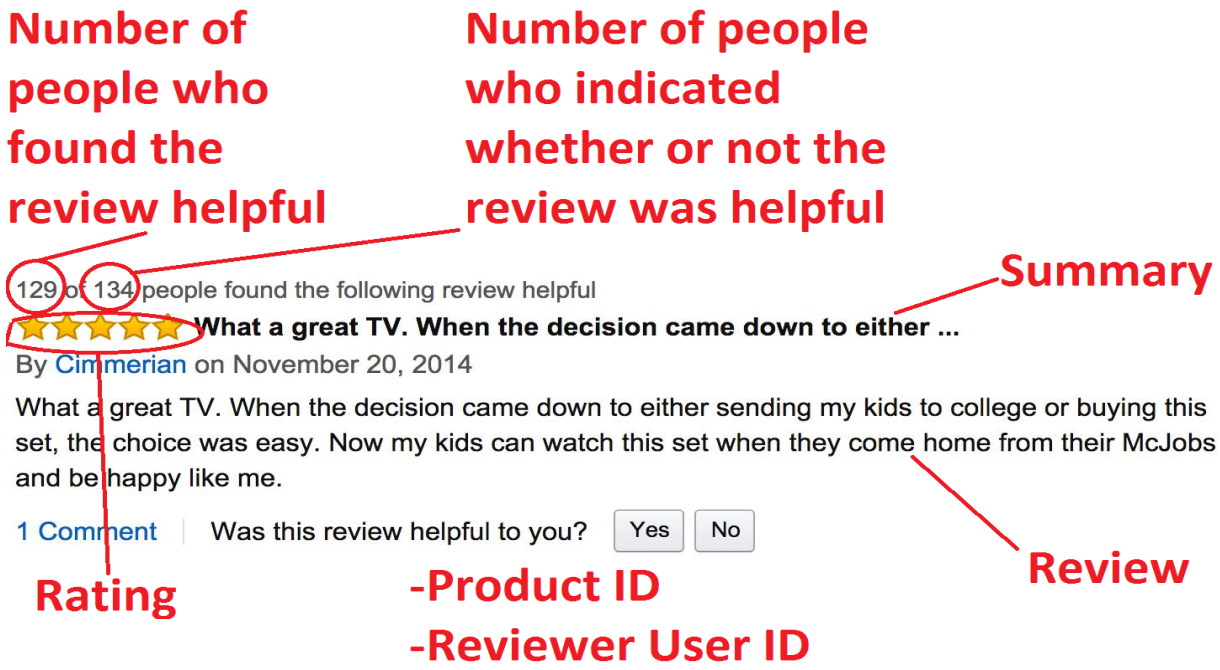
Data Description

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:

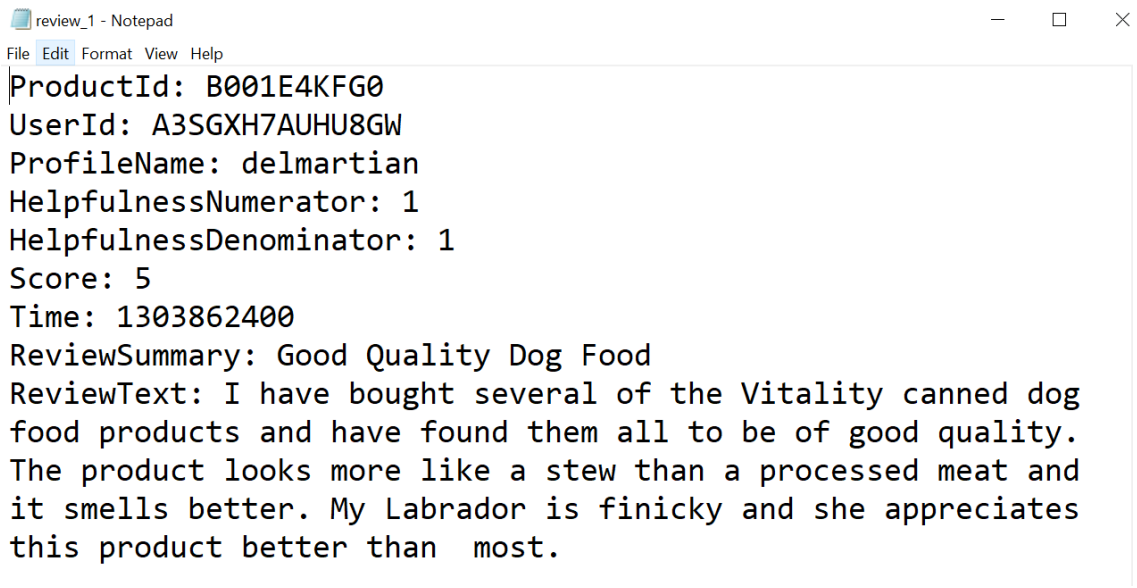
- Reviews from Oct 1999 - Oct 2012 - 568,454 reviews
- 256,059 Users and 74,258 products
- 260 users with > 50 reviews

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



SPRINT 1 - Create DataFrame from raw text files

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



Task - Your task here is to use your Data Engineering skills to transform the given data(i.e. Text files) to tabular format(i.e. csv file). The columns in this .csv file 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 - Build a model

Task A - Perform data preprocessing on the given text data and convert it into numerical vectors.

Task B - Build models to predict the Score of a given text review.

Client Expectations

1. Show me some nice analysis on the given data.
2. Show me the comparison of various ML models.
3. Model should be light for deployment.
4. Model should have very less latency.
5. Create a REST API to interact with the model.