

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.

Number of
people who
found the
review helpful

Number of people
who indicated
whether or not the
review was helpful

The screenshot shows an Amazon review interface. Red circles and lines highlight specific elements: a circle around '129 of 134' points to the 'Number of people who found the review helpful'; a circle around the 5-star rating points to the 'Rating'; a line from the text 'What a great TV. When the decision came down to either ...' points to the 'Summary'; a line from the main review text points to the 'Review'; and a line from the 'Was this review helpful to you?' text points to the '-Product ID' and '-Reviewer User ID' labels.

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?

Rating

-Product ID

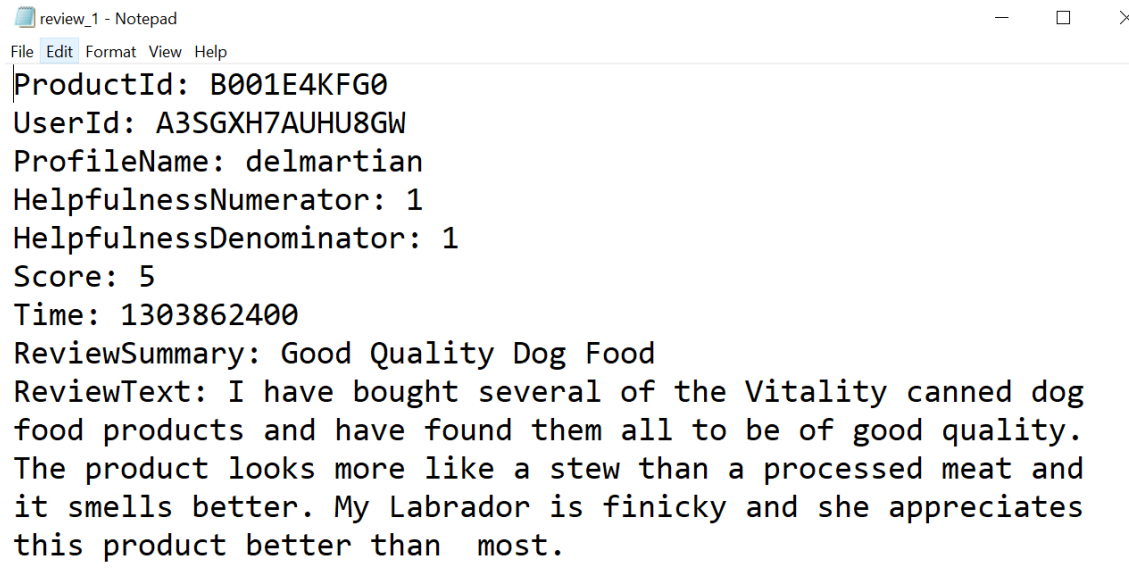
-Reviewer User ID

Summary

Review

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:

A screenshot of a Notepad window titled 'review_1 - Notepad'. The window contains a text file with the following content:

```
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.
```

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) = $\text{HelpfulnessNumerator} / \text{HelpfulnessDenominator}$

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.