ML1010

APPLIED MACHINE LEARNING & LIFECYCLE

PROJECT PROPOSAL

AMAZON FINE FOODS REVIEWS

By

Neha Panchal, Sagnik Adusumilli

Certificate in Machine Learning

York University

School of Continuing Studies

Toronto, Ontario, Canada

Instructor

[Annie En-Shiun Lee](https://learn.continue.yorku.ca/user/view.php?id=25014&course=1)

Date of Submission: Sunday, April 7, 2019

PROJECT PROPOSAL:  Problem, Dataset, and Features

**1.PROBLEM:** Classic Sentiment Analysis

**2.PROBLEM DEFINITION:** Determine whether the given review is positive (Rating of 4 or 5) or negative (rating of 1 or 2).

[Q] How to determine if a review is positive or negative?  
  
[Ans] We use the Score/Rating. A rating of 4 or 5 can be considered a positive review. A review of 1 or 2 can be considered negative. A review of 3 is neutral and ignored. This is an approximate and proxy way of determining the polarity (positivity/negativity) of a review.

**3. DATASET:**

**Dataset name**:Amazon Fine Food Reviews Analysis

**Data Source:** <https://www.kaggle.com/snap/amazon-fine-food-reviews>

The Amazon Fine Food Reviews dataset consists of reviews of fine foods from Amazon.

Number of reviews:568,454

Number of users:256,059

Number of products:74,258

Timespan: Oct 1999-Oct 2012

Attributes/Columns in data: 10

**Attribute Information:**

1. Id
2. Productid - unique identifier for the product
3. User id - unique identifier for the user
4. Profile Name
5. Helpfulness Numerator - number of users who found the review helpful
6. Helpfulness Denominator - number of users who indicated whether they found the review helpful or not
7. Score - rating between 1 and 5
8. Time - timestamp for the review
9. Summary - brief summary of the review
10. Text - text of the review

-Dataset is available in two formats.

1.Reviews.CSV

2.database.sqlite

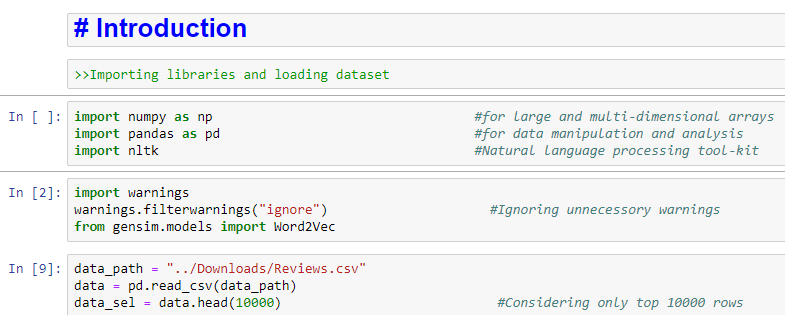
Reviews

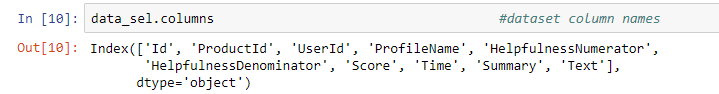
-We are using top 10,000 reviews only for this project.

**4. Data Preparation and Data Exploration:** clean, prepare and explore your data.

4.1  **Importing libraries and** Loading the dataset

we are loading the dataset using pandas.





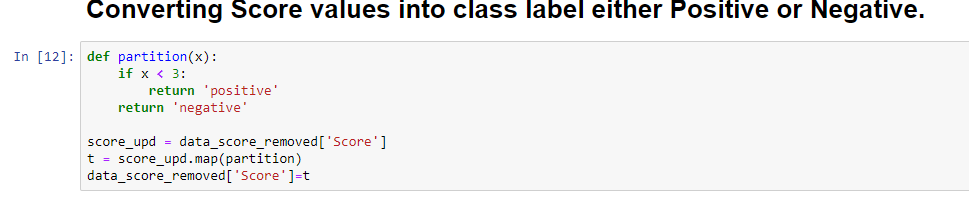
4.2 Exploratory Data Analysis:

We have followed Exploratory data analysis (EDA) approach to analyzing data sets to summarize their main characteristics with visual methods.

In this we have removed duplicate values and we will focus on ‘text’ and ‘score’ columns because these two columns play crucial role to predict the reviews.

The Score column is ranged from 1 to 5, and we have removed all Scores that are equal to 3 because we assume these are neutral and did not provide us any useful information. We will add a new column called “Positivity”, where score above 3 is represented as a 1,we say it was positively rated. Otherwise, it’ll be represented as a 0, indicating it was negatively rated.



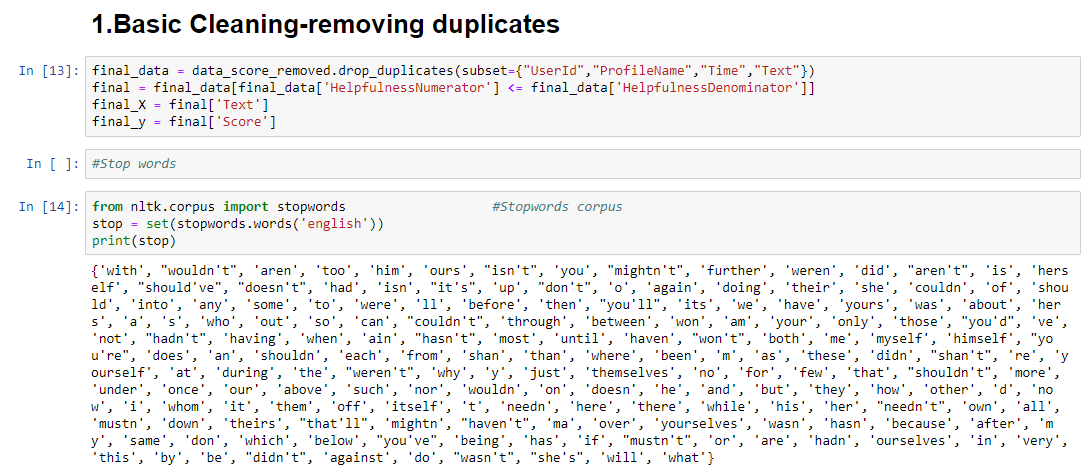


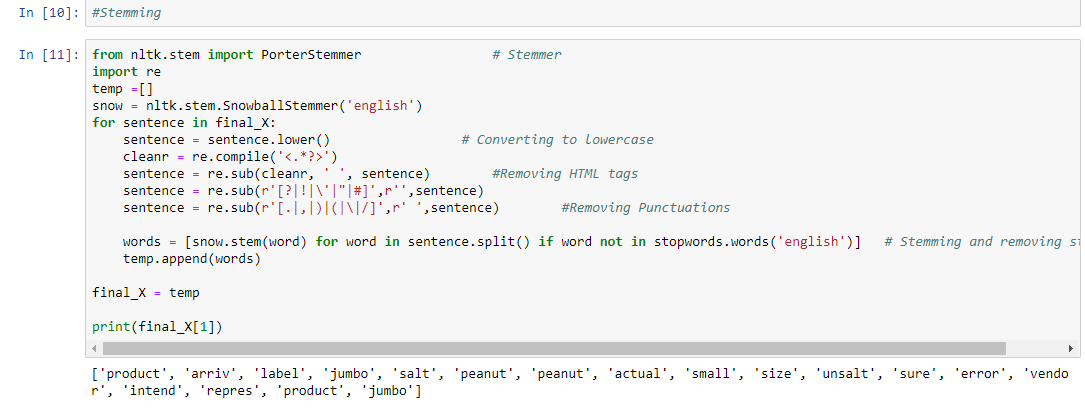
**4.3 Data Preprocessing**:

After duplication removal, we did some preprocessing before we go on further with analysis and to make the prediction model.

Hence in the Preprocessing phase we do the following :-

1.Begin by removing the html tags  
2. Remove any punctuations or limited set of special characters like , or . or # etc.  
3. Check if the word is made up of English letters and is not alpha-numeric  
4. Check to see if the length of the word is greater than 2 (as it was researched that there is no adjective in 2-letters)  
5. Convert the word to lowercase  
6. Remove Stop words  
7. Finally Snowball Stemming the word (it was observed to be better than Porter Stemming)



Stemming:

5. Feature Engineering:

Word2Vec

6. Submit a summary (1 paragraph) for your project milestone and a write-up (maximum 2 pages) of the above steps. .

1. Training and Testing

TBD

1. Modeling

TBD

1. Predictions

TBD