Project 2: Automated Sentiment Analysis of Text Data with NLTK

COMM 155: Artificial Intelligence and New Media, Winter Quarter 2021, Prof. J. Joo, TA. Aakash Srinivasan

1. Sentiment Analysis

Sentiment analysis is a well known task in machine learning and its goal is to classify the attitude or tone of an author towards a product, a service, an event, or a person based on text content. In this project, you will use the NLTK's sentiment analysis function to analyze text sentiment using three datasets: 1) Amazon product review, 2) beer review, and 3) movie review. Each dataset provides a list of pairs of a review content and a numeric rating. For instance,

- Text: "I like this move"
- Rating: 5

For each dataset, you need to complete analysis as follows

- Import modules
- Open the input file (csv) using the csv module and read content (texts and ratings).
- Run the sentiment analysis function to each text review and retrieve a score. • Collect all the scores from the entire dataset.
- Evaluate correlation between user-generated ratings and NLTK-generated scores.
- Visualize the result using matplotlib.
- Answer to the questions asked in the project colab notebook.

Here is the link to the project colab notebook.

2. Data

You are given three csv files: amazon.csv, beer.csv, and movie.csv. Each csv file contains 5,000 samples of review and rating. Use the csv module to read content. The ranges of ratings differ in different files, e.g., 1-5 or 1-14. Here is the link to the datasets.

3. Functions and modules that you can use

Name	**Description**	**Inputs**	**Returns**
numpy.corrcoef(*x*, *y)*	Calculate Pearson product moment correlation coefficients.	Two lists containing numbers. The shape of x and y should be the same.	The correlation coefficient matrix of the variables.
polarity_scores*(x)*	Calculate floats for sentiment strength based on the input text.	A single string text data	A dictionary that has four fields, {'compound', 'neg', 'neu', and 'pos'}

Example:

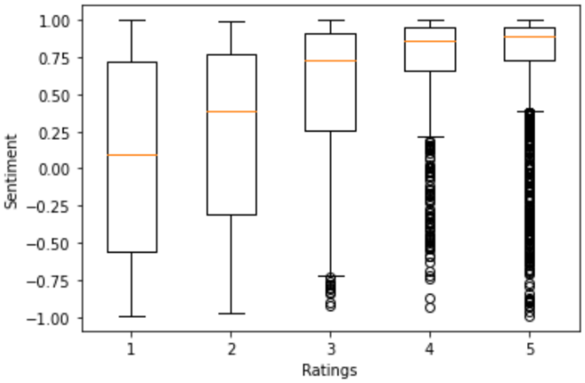
```
Import numpy
numpy.corrcoef([1,2,3], [1,3,2])
> array([[1. , 0.5],
 [0.5, 1. ]])
```

Example:

```
from nltk.sentiment.vader import SentimentIntensityAnalyzer
sid = SentimentIntensityAnalyzer()
sid.polarity_scores('I like you')
> {'compound': 0.3612, 'neg': 0.0, 'neu': 0.286, 'pos': 0.714}
```

4. Visualization

You need to plot the result obtained from each dataset using a box plot. Make three plots and answer the questions asked on the project notebook.



(example of the expected box plot)

5. Bonus Question

Create a Python function to repeat all the work done for a review category. It will be tested for robustness. For example, it should be able to run any of the three datasets provided for this homework.

Name	Description	Inputs	Returns
sentiment_analysis(x)	Repeat all the analysis done for a review category	A string with the csv file name, such as "amazon.csv"	The correlation coefficient and the boxplot

6. What to submit

The .ipynb file with problems completed to be submitted on CCLE before the due date. File name should be your UID. For example: 123456789.ipynb