COMM155

Project 2: Automated Sentiment Analysis of Text Data with NLTK

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.

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.

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 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'}

Ex)
Import numpy
numpy.corrcoef([1,2,3], [1,3,2])

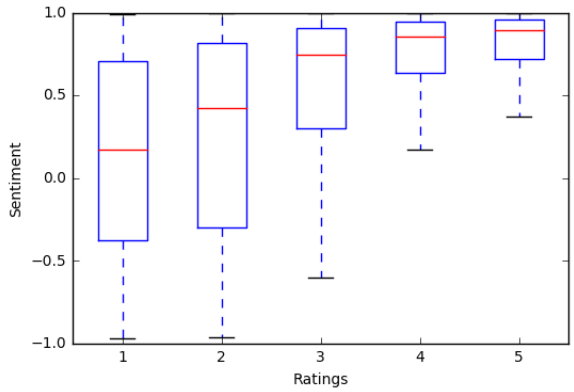
```
>  array([[1. , 0.5],
          [0.5, 1. ]])
```

Ex2)

```
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 include them in your report.



5. What to submit

a) Your code (.ipynb). You code must contain comments explaining important steps.
b) Report. The report must include following items. 1) an explanation of your code and algorithm, 2) results of execution, correlations and graphs, and 3) analysis of results (e.g., a few example texts drawn from each dataset and their results. Also explain which words are strongly associated with the sentiment score.)

6. Optional extra credit

Extra credit will be given if you can find or scrape any other text data and perform the same analysis. You will need to figure out how to process the raw text data into a csv format that your code can recognize.