Professional and Executive Development Program

Ashoka University

Assignment A3.3: Text processing using Python
(Ungraded + Bonus)

Course 3: Text processing using Python *Faculty: Prof. Lipika Dey*
Assignment Type: Individual *Bonus Marks: 60*
Date: September 23, 2023, Saturday *Bonus Deadline: September 25, 2023, Monday (11:59 pm)*

Instructions

Submission (if you choose to submit the solutions)

- Copy-paste the answers (values, tables, images etc., as applicable) in a separate doc / ppt file with corresponding task numbers. You may instead choose to submit your Google Collab/Jupyter/IPython Notebooks with all your code and responses in it directly.

- Add brief notes on how you arrived at the answers wherever necessary. You may instead choose to add comments to your Google Collab/Jupyter/IPython Notebooks directly.

- You need to submit your Google Collab/Jupyter/IPython Notebook (along with the intermediate files that may have been generated, if necessary) on Moodle. Name the files with the prefix A3.3_<Your Full Name> and submit it on Moodle.

Marks

- This assignment is completely ungraded and your marks for A3.3 will be calculated based on the average of your A3.1 and A3.2 marks.

- Nonetheless, there is an additional bonus task worth 60 marks that will be awarded to you if you choose to submit the solutions for Bonus Question only i.e. Q3. The deadline to submit your responses to the bonus question (only) will be as specified above.

About Dataset 1: Cafe Reviews Dataset

This dataset is a comprehensive dataset that captures customer reviews, along with a some essential information about a cafe such as an index, cafe name, overall rating, cuisine, average cost for two, city, and insightful reviews. The dataset has been shared on Moodle here. The columns are described below:

- **Index:** A unique identifier for each review entry.

- **Name:** The name of the cafe being reviewed.
- **Overall Rating:** The overall rating provided by the reviewer.
- **Cuisine:** The type of cuisine offered by the cafe.
- **Rate for Two:** The average cost for two people dining at the cafe.
- **City:** The city where the cafe is located.
- **Review:** A detailed review written by the customer, capturing their experience.

About Dataset 2: Adverse Drug Effect (ADE) Dataset

This dataset contains labelled bio-medical statements that can be used for the classification task - whether a statement is ADE-related or not. The dataset has been shared on Moodle here. The columns are described below:

- **ID:** These may not be unique IDs.
- **LABEL:** 'POS' is used for the statements which indicate an "adverse drug effect" and 'NEG' is used to for statements that do not an "adverse drug effect".
- **Text:** The bio-medical staement that we are interested in classifying.

Tasks to be completed during the lab hour

1. **Text classifier with Dataset 1.** **[Practice - 0 marks]**

   a. Load the file into appropriate data structures so that the subsequent operations can be done.

   b. Generate a new file "cafe_sentiment.csv" which contains only the reviews and a label - where the labels are generated as follows:

   if (rating > 3.5), then "GOOD" else "BAD"

   c. Develop a text classifier using any model using the SciKit Learn library. Use 70 to 80% of data as training and remaining as test data.

   (Hint: https://scikit-learn.org/stable/supervised_learning.html#supervised-learning).

d. What is the performance of your model? Which classifier did you choose and why?

2. **Named Entity Recognition with Dataset 2.** **[Practice - 0 marks]**

    a. Write code to perform Named Entity Recognition using Spacy and SciSpacy libraries to detect the Drugs from the statements.

    b. Compile the output of NERs for each statement and generate an aggregate table as follows:

    For each unique output of the NERs - find the total of number times it occurs in the dataset, number of times it occurs in statements with POS label and number of times it occurs in statements with NEG label.

3. **Bonus Task: Clustering with Dataset 2:**

    a. Write code to run k-means clustering algorithm over the bio-medical statements in the above-mentioned file. Starting with k=2, you can go up to k=10 or more. **[20 marks]**

    b. Compute the silhouette score for each run and plot a graph to show how they change with k. **[10 marks]**

    c. Write a program that can determine automatically the value for "k" that finds the best segregation of POS and NEG statements.

    (Hint: k may be greater than 2. Use Purity measures.) **[30 marks]**

Study Material

1. Top 5 NLP Tools in Python for Text Analysis Applications
2. NumPy: the absolute basics for beginners
3. Pandas tutorial: 10 minutes to pandas
4. Scikit-learn User Guide
5. spaCy 101: Everything you need to know
6. scispaCy