MIS 5163 Data Mining and Text Analytics in Business

Exam Coding Questions

Please finish the following questions in Python, and upload the python code file (.py) to d2l dropbox to earn your points.

PLEASE ONLY USE THE DATA FILE PROVIDED FROM THE EXAM FOLDER, and THE MODEL MENTIONED

1. Write a python program with the given **review.json** data file, finish the following task: **(10 pts.)**
    1) Load the **review.json** file to pandas **DataFrame**, preprocess your dataset.
    2) Make a loop, for each reviews in the **DataFrame**, perform:
        a. Correct any **spelling errors** found in the reviews, print it out.
        b. Calculate the **total number of words**, print it out.
        c. Calculate the **total number of sentences**, print it out.
        d. Apply stemming to all words, print the resulting sentences containing the **stemmed words**.
        e. Perform lemmatization on all words, print the result sentences containing the **lemmatized words.**

2. Write a python program with the given **review.json** data file, finish the following task: **(10 pts.)**
    1) Load the **review.json** file to pandas **DataFrame**, preprocess your dataset.
    2) Make a loop, for each reviews in the **DataFrame**, perform:
        a. Calculate **total number of words, total number of sentences, total number of complex words,** print them out.
        b. Calculate the **Flesch-Kincaid Grade Level**, and print it out.
        c. Calculate the **Gunning Fog Index**, and print it out.
        d. Calculate the **Coleman-Liau Index**, and print it out.
        e. Calculate the **SMOG Index**, and print it out.
        f. Calculate the **Automated Readability Index**, and print it out.

3. Write a python program with the given **amazon_review.csv** data file, finish the following task: **(10 pts.)**
    1) Load the **amazon_review.csv** file to pandas **DataFrame**, preprocess your dataset.
    2) Make a loop, for each reviews in the **DataFrame**, perform:
        a. Use any available models, to make the **review summary**, keep it in **two sentences** or **less than 100 words**.
        b. Use any available models, to perform a **sentiment analysis**.
        c. Compare the **sentiment analysis result (compound)** with the **review ratings**. (Correlations or Similarities)
        d. Print each review summary along with sentiment analysis result (**Positive, Neutral, Negative**)
    3) Interpret your observation on the comparison between **sentiment analysis results** with **review ratings.**

4. Write a python program with the given **amazon_review.csv** and **subjectivity.csv** data files, finish the following task: **(10 pts.)**
    1) Load the **subjectivity.csv** file to pandas **DataFrame**, preprocess your dataset.
    2) Assign a correct **X** and **y**, vectorize the **X**.
    3) Spilt the data, train a classification model, evaluate it, and print the results.
    4) Load the **amazon_review.csv** file to pandas **DataFrame**, preprocess your dataset.
    5) Make a loop, for each reviews in the **DataFrame**, perform:
        a. Use your trained model, classify each reviews, print the review, along with your classification result (**Subjective** or **Objective**)
    6) Calculate the accuracy, precision, confusion matrix for the classification results of the **amazon_review.csv.**
    7) Split your subjectivity.csv data set into **subjective** dataset and **objective** dataset.
    8) Plot **word clouds** for **subjective** dataset and **objective** dataset.
    9) Interpret your observation of two different **word clouds**.
    10) Plot **bar chart** for top 10 words associate with **subjective** class and **objective** class. (Extra Credit)

5. Write a python program to finish the following task: **(10 pts.)**
    1) Find a **Steam Game Review Dataset**, you could either find it from **kaggle.com** (*Please provide me the link to download the dataset*), or you could use **Rapid API (https://rapidapi.com/psimavel/api/steam2)** App Reviews Endpoints (Extra Credit) to load 30 piece of reviews.
    2) Load the data set into pandas **DataFrame**, preprocess your dataset.
    3) **Summarize** your data, give some **brief statistic results** to explain this dataset.
    4) Visualize one or two columns, explain your observation.
    5) Vectorize your **review text** column into **TFIDF** vectors.
    6) Assign the word vectors to **X**, and assign an appropriate **y**. Explain the reasons about why you choose the column as y)
    7) Spilt the data, train a classification/prediction model, evaluate it, and print the results.