Fundamental of Big Data Analytics (DS2004)

Assignment #0

Deadline: 15-02-2024

Course Team		
Dr. Muhammad Ishtiaq	Course Coordinator	m.ishtiaq@nu.edu.pk
Ms. Kainat Iqbal	Course Instructor	kainat.iqbal@nu.edu.pk
Ibrahim Bin Umair	Teaching Assisant	i200567@nu.edu.pk
Muhammad Huzaifa Khan	Teaching Assisant	i212689@nu.edu.pk
Hashim Muhammad Nadeem	Teaching Assisant	i211675@nu.edu.pk

Assignment Guidelines:

- This is an individual assignment.
- The entire code and explanations must be included in a *single* executed **Jupyter Notebook file (.ipynb)** – no other file format will be accepted.
- To ensure timely and accurate grading, please follow the file naming convention "**RollNumber_FullName_A0.ipynb**" (for instance, "**i211234_JohnDoe_A0.ipynb**"). Failure to comply with the correct naming convention will result in a deduction of 5 marks.
- To enhance the readability of your submission, please use Markdown elements (such as headings) to organise the code and explanations/findings.
- Any portion of the code with an error, even minor ones, will not be accepted.
- You may refer to online sources such as websites and/or ChatGPT, but make you can explain the code you have written.
- To earn bonus marks, please use comments and adhere to correct **PEP 8** coding conventions.

Plagiarism Policy:

Plagiarism is a grave academic offense that can severely undermine your academic integrity and reputation. Any instance of a student found to have plagiarised their assignment, whether from a peer or external source, will be subject to strict consequences. This may result in a zero score for the current or all assignments, or in severe cases, even a failure in the course. Furthermore, all instances of plagiarism will be promptly reported to the Disciplinary Committee for further action.

Text-Based Sentiment Analysis Using Python

[40 Marks]

Sentiment analysis is essential in data analysis as it helps businesses interpret the emotional tone in textual data, such as customer reviews and social media posts. By discerning sentiments, organisations can gain valuable insights into customer opinions, manage brand reputation, refine marketing strategies, enhance customer experiences, and proactively address potential issues. It plays a crucial role in market research, social media monitoring, and financial analysis, providing a qualitative dimension to data analytics that is instrumental in making informed decisions and optimising overall business performance.

For this task, your objective is to create a simple rule-based sentiment analysis algorithm using pure Python. You are given a JavaScript Object Notation (JSON) file containing Amazon.com reviews for cell phones and accessories. External libraries like pandas and NumPy are *prohibited*, but you can utilise built-in modules such as json, re, and statistics. The following provides a comprehensive outline of the implementation:

1. **Data Loading & Preprocessing [8 Marks]:**
   - Read and parse the JavaScript Object Notation (JSON) data into a suitable Python data structure. **[1 Mark]**
   - Explore the dataset to understand its structure, size, and characteristics. Filter the dataset to retain only the necessary columns to minimise unnecessary computations. **[2 Marks]**
   - Apply the text preprocessing methods covered in the introductory course, like removing punctuation and stop words. Utilise this list (https://gist.github.com/sebleier/554280) for a pre-made set of stop words. Avoid using any text processing libraries for this step. **[5 Marks]**
2. **Thematic Analysis [10 Marks]:**
   - Create a technique to identify key phrases or words that are commonly found in positive and negative reviews. This may include basic frequency analysis, such as Cumulative Frequency Distribution (CDF), or more advanced methods. Use the ratings as a reference point to distinguish between good and bad reviews. **[10 Marks]**
3. **Sentiment Analysis [20 Marks]:**
   - Once you have performed the thematic analysis, you will design a simple rule-based sentiment analysis algorithm. Further information on rule-based sentiment analysis can be found at: https://vtiya.medium.com/rule-base-sentiment-analysis-adfad898470b
   - Define rules or heuristics to assess the sentiment (positive, negative, or neutral) of a given review by analysing the review text. For instance, assign weights to commonly used words, and accumulate these weights based on their frequency in the review text. **[15 Marks]**

- **Example:** If the word “bad” is frequently found in negative reviews, give it a weight of 0.05. Similarly, assign higher weights like 0.65 to positive words such as “good” and “amazing.”
- This process may require some manual input and thoughtful selection. Your rules or heuristics may differ depending on your approach, but they should be supported by your thematic analysis.

Consumer comment	Meaningful words	Sentiment weight	Sentiment score
“Employees are **always** so **polite** and **helpful**. They have almost everything larger stores have, which is **good** enough for me! They even have white sweet potatoes which you usually only find at specialty grocers. **Great** experience!”	Always Polite Helpful Good Great	1 2 2 1 2	8
“**Nice** store. Seems **clean** and **well organized**. It is a bit on the *smaller* size as far as other stores go, so prepare for it to perhaps feel a little *crowded*, even if there aren't that many people in the store. This also means that their stock of certain items may be a bit *smaller* than the **larger** locations, but that's to be expected.”	Nice Clean Well Organized *Smaller* *Crowded* Larger	2 2 1 2 -1 (x2) -1 1	5
“Since a **new** manager took over this once **best** supermarket around has *fallen* way off. They *barely* have anything, produce is always *expired*, and the store is not kept up very well.”	New Best *Fallen* *Barely* *Expired*	1 2 -1 -1 -2	-1

FIG. 1: A model demonstration of a basic rule-based sentiment analysis algorithm relying on commonly occurring words.

- After obtaining a cumulative score for the review text, establish a threshold to categorise each review as positive, negative, or neutral. For instance, scores below 0.5 may be labelled as negative, while those above 0.5 as positive (the specific lower and upper bounds will be determined by your specified rules). **[5 Marks]**

4. **Storage [2 Marks]:**
   - Save the end result in a text file, putting the review text next to its sentiment. **[2 Marks]**

Note:

This assignment aims to introduce you to fundamental data structures and techniques in Python that will be essential throughout the course. View it as a chance to learn and grasp important methods. There’s flexibility in how you implement the assignment, so feel free to explore creative approaches, including advanced statistical techniques and data structures. While you *can* use generative tools like ChatGPT, be prepared to explain your work during the demonstration.