CS5481: Data Engineering - Assignment1

Instructions

1. Due on Tuesday, Oct. 8, 2024, 18:00:00 PM;
2. You can submit your answers by **a single PDF with the code package** or **a single jupyter notebook** containing both the answers and the code;
3. For the coding questions, besides the code, you are encouraged to additionally give some descriptions of your code design and its workflow. Detailed analysis of the experimental results are also preferred;
4. Total marks are 100;
5. Plagiarism or unjustified late submission will result in the assignment being invalid or points being deducted correspondingly.
6. If you have any questions, please post your questions on the Canvas-Discussion forum or contact TA Ms. Yuxuan Yao (email: yuxuanyao3-c@my.cityu.edu.hk).

Question 1 - Data Acquisition

(**20 marks**) Social media content, including blogs, articles, news, and Twitter posts, provides valuable insights for data science. Acquiring high-quality social media data is essential yet challenging. While crowdsourcing is an option, it can be costly; thus, we prefer to collect data through web scraping.

Please collect **30 pieces of social media content** from designated websites, ensuring that the data meets the following criteria:

1. Include articles, blogs, news, or posts along with their comments.
2. Focus solely on the textual content.
3. Ensure the data is formatted in a structured way (e.g., JSON or CSV).

We provide some social media websites that you can take a try.

- https://english.news.cn
- https://www.bbc.com/news
- https://medium.com
- https://twitter.com

Please submit your code and the obtained social media data.

Question 2 - Data Preprocessing

(**30 marks**) Regular Expressions, abbreviated as Regex or Regexp, are a string of characters created within the framework of Regex syntax rules. You can easily manage your data with Regex, which uses commands like finding, matching, and editing. Regex is an important tool during the data preprocessing stage.

We take some exercises about regular expressions in Python,

1. Write a pattern to check if a string contains only letters (both uppercase and lowercase).
   - Test cases: *Hello, world, 123abc*
2. Write a pattern to find all words that start with a vowel.
   -Test cases: *apple, banana, orange, grape*
3. Write a pattern to validate an email address.
   -Test cases: *test@example.com, invalid-email*
4. Write a pattern to extract all digits from a string.
   -Test cases: *The price is 100 dollars and 50 cents.*
5. Write a pattern to match a URL.
   -Test cases: *https://www.example.com, ftp://example.com*
6. Write a pattern to validate a US phone number format (e.g., (123) 456-7890)
   -Test cases: *(123) 456-7890, 123-456-7890*
7. Write a pattern to find a string that starts and ends with the same character.
   -Test cases: *radar, hello, level*
8. Write a pattern to validate a complex password. The password must contain at least one uppercase letter, one lowercase letter, one digit, one special character, and be at least 8 characters long.
   -Test cases: *Password1!, PASSWORD,1!, Pass1!*
9. Write a regex pattern to identify and extract all instances of dates in the format dd-mm-yyyy or yyyy/mm/dd from a given text. The pattern should handle both formats in a single regex.
   -Test cases: *I12-05-2023, 2023/06/15, and 01-01-2024.*
10. Create a regex pattern that matches a valid IPv4 address. The address must consist of four octets separated by dots, where each octet is a number between 0 and 255.
    -Test cases: *192.168.0.1, 256.100.50.25, 172.16.254.1*

Question 3 - Data Processing

(**20 marks**) The source files of Workshop on Statistical Machine Translation (WMT) are usually xml files. Before we train a model using these data, we should convert them from XML formt to line-based text. Please solve the following questions:

1. Please convert the data in this file [1] to the line-based text with your own Python codes. You need to remove all punctuation and convert all text to lowercase. You should submit your runnable codes and output file.
2. After you obtain the line-based text file, please create a BPE vocabulary (save each BPE token line by line) with subword-nmt [2]. You should submit your runnable codes and output file.

[1] https://github.com/wmt-conference/wmt-format-tools/tree/main/test/sample-data/sample-hyp.xml
[2] https://github.com/rsennrich/subword-nmt.git

Question 4 - Data Visualization

(**30 marks**) Data visualization is an effective method to overall evaluate the quality of the data. Generally, the conventional visualizations include column histogram/chart, pie chart, venn diagram, scatter plot, heatmap, etc.

1. Assume we have a set of employee records containing employee **ID** (Integer; 1-500), **department** (Categorical; HR, IT, Sales), **sex** (Binary; Male/Female), and **years of experience** (Integer; 0-40), we intend to analyze these attributes by visualization. Which visualization technique should be selected for different attributes?

2. Write a Python Program to randomly generate 500 employee records based on the above descriptions and visualize the generated data using your selected techniques.

3. Calculate the number of employees per department and visualize the results using a bar chart for the generated data.

4. Attention[1] is a classic and popular technique in natural language processing. Given two vectors $\mathbf{Q} \in \mathbb{R}^{5\times 10}$ and $\mathbf{K} \in \mathbb{R}^{5\times 10}$, the attention score of $\mathbf{Q}$ and $\mathbf{K}$ are calculated as:

$$\text{Attention\_Score}(\mathbf{Q}, \mathbf{K}) = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}),$$

   where d_k is the hidden dimension (10 in this case).

   Please randomly initialize $\mathbf{Q}$ and $\mathbf{K}$ vectors and visualize the attention score via **heatmap**.

Reference [1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems.