CS5481: Data Engineering - Assignment1

Instructions

1. Due at Tuesday, Oct. 3, 2023, 12:59:59 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. If you have any questions, please post your questions on the Canvas-Discussion forum or contact TA Mr. Han Wu (email: weishao4-c@my.cityu.edu.hk).

Question 1 - Data Acquisition

(**20 marks**) Social media data, such as blogs, articles, news or Twitter posts, is much valuable for data science. However, how to obtain high-quality social media data becomes an important and challenging problem. Alternatively, we can collect the data by crowdsourcing, but it might be expensive. Therefore, we prefer to gather social media data by web scraping.

Please try to crawl **20 pieces of social media data** from social media websites. The data should satisfy following requirements:

1. Types of the article, blog, news or posts with its comments;
2. We just need the textual information;
3. Try to clean the data, i.e., removing all HTML tags.

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 the pattern to check that a string only contains a certain set of characters (in this case a-zA-Z and 0-9).

   - Test cases: *ABCDEFabcdef123450* and *ABCD@Fabcdef123450*

2. Write the pattern that matches a string that has an ‘a’ followed by one or more ‘b’.

   - Test cases: *bab*, *abbbbb* and *baaaa*

3. Write the pattern to check whether a string starts and ends with a specific number (in this case 6).

   - Test cases: *65117896*, *78238936* and *56666665*

4. Write the pattern to search the number (0-9) of length between 2-4 in a given string.

   - Test cases: *Exercises number 1, 23, 345, and 45678 are important*

5. Write the pattern to remove leading zeros from an IP address.

   - Test cases: *210.08.090.194* and *010.01.010.100*

6. Write the pattern to replace whitespaces with an underscore and vice versa.

   - Test cases: *Python Exercises Of Regular_Expression*

7. Write the pattern to convert the date of yyyy-mm-dd format to dd-mm-yyyy format.

   - Test cases: *2022-09-10*

8. Write the pattern to find all words starting with ‘a’ and ‘e’.

   - Test cases: *The following example creates an ArrayList with a capacity of 50 elements. Four elements are then added to the ArrayList and the ArrayList is trimmed accordingly.*

9. Write the pattern to extract values between quotation marks of a string.

   - Test cases: *Regex can be used in programming languages such as “Python”, “SQL”, “Javascript”, “R”, “Google Analytics”, “Google Data Studio”, and throughout the coding process.*

10. Write the pattern to find urls in a string.

   - Test cases: *Find more Examples at Github https://www.github.com or W3School https://www.w3schools.com/.*

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 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-src.xml

[2] https://github.com/rsennrich/subword-nmt.git

Question 4 - Data Visualization

(**20 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 user profiles, including **user_id** (Integer;1-200), **sex** (Binary;Male/Female), **age** (Integer;18-100), **height** (Float;100.0-200.0) and **weight** (Float;30.0-100.0), 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 200 user profiles following above descriptions and visualize the generated data using your selected techniques.
3. 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 **Q** and **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.