CBS4958 FUNDAMENTALS OF COMPUTATIONAL LINGUISTICS

Programming/Written Assignment, September 2024

Student Name: 刘硕宙(LIU shuozhou)

Student ID: 24106164d

**Instructions**:

1. Download all files from the "Programming + Written Assignment 1" folder on Blackboard.

2. For each exercise, submit:

- An edited version of this document for textual answers.

- The Python file(s) for programming exercises. You may create a new Python script.

- If an exercise requires an output file, include it as well.

3. The total score for this assignment is 15 points. A minimum of 9 points is required to pass.

4. Ensure your name and student ID are clearly written at the beginning of this document.

5. Submit all required files to ***jinghang.gu@polyu.edu.hk*** by ***October 17, 2024***. Late submissions will incur a 4-point penalty.

6. If including multiple files, compress them into a single **.zip** file before sending.

7. If you find any errors or inconsistencies in the assignment, please contact me as soon as possible.

Declaration on Generative AI (GenAI) Usage:

You may use GenAI tools for this assignment. However, you must declare its usage as follows:

“I/We declare that Generative AI tools have been used to prepare the submitted work. The Generative AI tools used and the manner in which they were used are as follows:                                           ”

**Part 1: Programming Assignment**

Exercise 1 (3 points)

**Problem:** Write a function that takes a string as input and returns the frequency of each word in the string.

Example:

Input: "this is a test this is only a test"

Output: {'this': 2, 'is': 2, 'a': 2, 'test': 2, 'only': 1}

Exercise 2 (3 points)

**Problem**: Write a function that takes two textual files as inputs. The function should merge the contents into a new file. In the merged file, even-indexed lines come from the first file and odd-indexed lines from the second. Each line should have three columns: line number, original filename, and content, separated by tabs.

Example:

File1:

```

This is an apple.

That is a pear.

```

File2:

```

I have an apple watch.

They have apple phones.

```

Merged Output:

```

0 Filename1 This is an apple.

1 Filename2 I have an apple watch.

2 Filename1 That is a pear.

3 Filename2 They have apple phones.

```

Exercise 4 (4 points)

**Problem**: Download and decompress the file "*Dataset\_Xiang-Kuperberg\_2015.zip*". In the decompressed "*Dataset\_Xiang-Kuperberg\_2015.txt*" file, you'll find sentences for psycholinguistic experiments. Your task is to prepare these sentences for some deep neural models (e.g., BERT).

1. Write a function replacing the target word (fifth column) in the sentence (third column) with the tag '[MASK]'. Retain all other fields unchanged.
2. Write another function randomly select a word in the sentence (third column) and replace it with the tag '[UNK]'. Retain all other fields unchanged.

**Part 2: Written Assignment**

Exercise 1 (2 points)

**Questions**: Using your own words, explain what is text preprocessing in Computational Linguistics? List and explain at least two methods and their objectives.

Exercise 2 (3 points)

**Questions**: Describe the strategies and tools used to identify and handle programming bugs in Python.