Mini-project 2

1. Overview

There are two approaches to finish mini project 2, you can choose one of them.

i. You are asked to complete a project on an NLP topic. Your project should include the following three components:
   - `Collect data`
   - `Pre-process data`
   - `Analyze data`

ii. You are asked to choose a topic in NLP such as Machine Translation, Handwriting Recognition, Text Alignment and Matching, or Text Similarity Calculation. Select an NLP-related topic and develop a practical application.

Your final grade is based on code and a report.

2. Project details

For the first approach (i):

(1) Topics

Select a topic that is related to Natural Language Processing (NLP). Here are some examples:

- Douban movie review collection and analysis
- Keyword analysis of corporate annual reports
- Weibo trending topics Tracking and Analysis

You are not required to choose a very challenging topic; instead, you can choose something within your reach. The choice of topic will **NOT** affect the final grade.

(2) Collect data

There are serval ways to collect raw data from the Internet:

- Download pdf/doc documents directly (recommended)
- Fetch data from website API (recommended if you are familiar)
- Build a crawler to crawl data (challenging)

Raw data should primarily be in text format. Other formats (such as images or audio) are beyond the scope of this course, but you are welcome to use them provided you can process them effectively. Note that the chosen method will **NOT** affect your final grade.

(3) Pre-process data

This step aims to transform the data into a useful and efficient format using Python techniques taught in this course. Besides basic grammar, you may also refer to the following modules to process data:

- **jieba**: The most used Chinese word-splitting tools
- **numpy**: The fundamental package for scientific computing

(4) Analyze data

In this step, you are asked to output processed data in a user-friendly way (like Wordcloud). You may refer to the following modules:

- **pandas**: A powerful data analysis and manipulation tool
- **matplotlib.pyplot**: A plotting library for creating static, animated, or interactive visualizations
- **WordCloud**: A visualization tool to represent word frequency

The output in this step **GREATLY** impacts the final score. The grading criteria are:

- Intuitive
- User-friendly
- Logical
- Comprehensive

For the second approach (ii):

You are not required to implement a highly optimized or complex algorithm. It is sufficient to implement basic functionality of an algorithm. For instance, if you choose machine translation, your algorithm needs to handle words or sentences correctly; it does not need to perfectly translate all types of inputs. Additionally, you are required to explain the core things of the algorithm in your report.

3. Report

For both approaches, you need attach a report:

- Why did you choose this topic?
- How did you implement it (e.g. which Python module did you use and how you use it)?
- The results (attach screenshots and explain them)?
- (approach I) How the results may contribute to further research?
- (approach II) What is the principle or core components?

No page or word limitation. You should submit it in PDF.

4. Grading Criteria

Score	Requirement
0-60	Partially done
60-80	Fully done with unsatisfactory output
80-100	Fully done with exemplary output

5. Submission

You need to rename your program and report file as `project2.py` and `project2.pdf`, respectively. Then, you should compress all files into a rar or zip file and name it using your id, for example, `projct2_Xiaoli_12345.zip`.