University of Macau

CISC3025 – Natural Language Processing

Project#1, 2024/2025

(Due date: **21st March, 2025**)

Project Rule

This is an **individual** course project. You are strongly recommended to commence work on each assignment task of the project soon after it is announced in class/UMMoodle. Students are free to discuss the project, but they are not permitted to share any code and report.

Project #1: Write a spider to collect & preprocess the data

What should I do in this project?

In this project, you are required to use python to crawl Books to Scrape website,[1] extract structured data from their pages and preprocess the extracted data. If you are new to spider/crawler, please search tutorials on the internet. We may use requests, BeautifulSoup, and some other packages to deal with simple crawling.

Here go the basic project requirements,

1. In the Books to Scrape website, each page contains 20 books. Extract the URL of each book from the website pages 1-10, save the tuple (title, url) to ./url.json.
2. Extract the description of the books on pages 1-10. The description content needs to be saved as text files (.txt) in the folder "./test." The name of the saved files should be formatted as "BookName.txt," from which "BookName" infers the name of the book. Each file contains a description of the corresponding book. For example, "A Fierce and Subtle Poison.txt" refers to the title of the book "A Fierce and Subtle Poison."
3. Preprocess the extracted description. Use the regular expression and string operation to strip the useless symbols and blanks in the raw text, tokenize the description and stem the tokenized word using the package in NLTK. Building your own tokenizer and stemmer is also **strongly** encouraged. The corpus needs to be saved as "corpus.json," in JSON format. The JSON file contains a list. Each element consists of the title and description of the book. You can also output the preprocessed text in the folder "./corpus."

[1] http://books.toscrape.com/index.html

Where can I find the starter code?

There are three starter code files:

- **spider-url.py**: The code implements a simple spider to extract the URL of each book from webpage 1-5. Modify the code to accomplish requirement 1.
- **spider-books.py**: The code implements a simple spider to extract the description of each book from the URL. You need to use the BeautifulSoup package to look for the book's description in the HTML text.
- **preprocess.py**: The code forms a template for data preprocessing. Implement the details of the functions to achieve the data preprocessing pipeline. The "output_to_file" function may help you output the preprocessed text to files.

Note:

- The starter code implements some basic functions of a spider/crawler. You can follow the instructions in the code to finish the project. You can also write your own code from scratch, but the code structure needs to be explained in the report.
- Using regular expression to extract content is **strongly** encouraged.

What do you need to submit?

- **Runnable source code**
  - You need to make sure your code functions well according to the project requirements.
  - Of course, any other optimization designs are allowable. You can describe your genius designs in your report.
- **Text records files and JSON files**
  - Provide at least 200 records of book data.
  - The url.json file is saved as JSON format, containing a list. Each tuple element records the title and URL of each book.
  - The test folder contains the raw text extracted from the website.
  - The corpus folder contains the preprocessed text data. The text data needs to be saved as text files (.txt) in folder "./corpus." The name of saved files should be formatted as "BookName.txt," from which "BookName" infers the name of the book. Each file contains the preprocessed description of the corresponding book.
  - The corpus.json file should contain the structured preprocessed book description text data. The JSON file contains a list of title and description of the book.

- **The folder structure:**

```
project1/
  test/
    bookName.txt
  corpus/
    bookName.txt
  spider-url.py
  spider-books.py
  preprocess.py
  corpus.json
  url.json
```

- **Report**
  - You need to submit a report to introduce your work. The report should contain the following information:
    - An introduction about your project, what is it about? What did you achieve?
    - Description of the project, what you did? What is your approach, design, or method? What did/didn’t work? How did you solve it? What did you achieve? What are the results?
    - Conclusions, what was accomplished/learned? What would you have done differently? What would you suggest for future work?
    - (Optional) References to the tools, online resources, books, or papers that were useful to your project.