CS5481 Assignment1 Report

Student Name: SONG Tao   Student ID: 56642520

Question 1 - Data Acquisition

For this question, I chose two data sources, Xinhuanet (news.cn) and BBC, to obtain data. In the submitted code, each website has its own corresponding code file. The code function is not only to obtain specific article content, but also to **use the website's search function to search for topics of interest first**, and then select articles from the most relevant article list. When searching, it supports turning pages on the search page to obtain more articles.

The saved articles are stored in JSON format, including the article title, metadata, and article body.

NOTE: For this question, I used the Selenium package and the Ubuntu distribution of the Google Chrome browser. For specific version numbers, see readme.md

Question 2 - Data Preprocessing

Regular Expression:

1. [a-zA-Z]
2. \b[aeiouAEIOU][a-zA-Z]*
3. ^\w+((.|-)|\+\w+)*@\w+(-|.\w+)*
4. \d+
5. ^\w*:\/\/\w+(-|.\w+)*
6. \(\d{3}\)\s\d{3}-\d{4}
7. ^(.).*\1
8. (?=.*[A-Z])(?=.*[a-z])(?=.*\d)(?=.*[^a-zA-Z\d]).{8,}
9. (?:\d{2}-\d{2}-\d{4}|\d{4}\/\d{2}\/\d{2})
10. \b(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})\.(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})\.(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})\.(?:25[0-5]|2[0-4][0-9]|[01]?[0-9]{1,2})\b

Question 3 - Data Processing

1. For the first sub-question, I used the XML toolkit that comes with Python to disassemble the dataset, remove punctuation and change it to lowercase, and finally saved two different files by language.
2. For the second sub-question, I used Python code to execute shell commands and created the BPE vocabulary using the subword-nmt toolkit.
   Shell command used:
   ```
       subword-nmt \
       learn-joint-bpe-and-vocab \
       --input {os.path.join(currentdir, infile+".txt")} \
       -s 10000 \
       -o {os.path.join(currentdir, "codes_file.code")} \
       --write-vocabulary {os.path.join(currentdir, "vocab_file_"+infile+".txt")}
   ```

Question 4 - Data Visualization

Please refer to the Jupyter notebook (main.ipynb) in the q4_data_visualization folder in the submitted project folder for details.