Software name: Visual XML

Person in charge: Kai Xu

Version number: Not released

Student number: 32891059

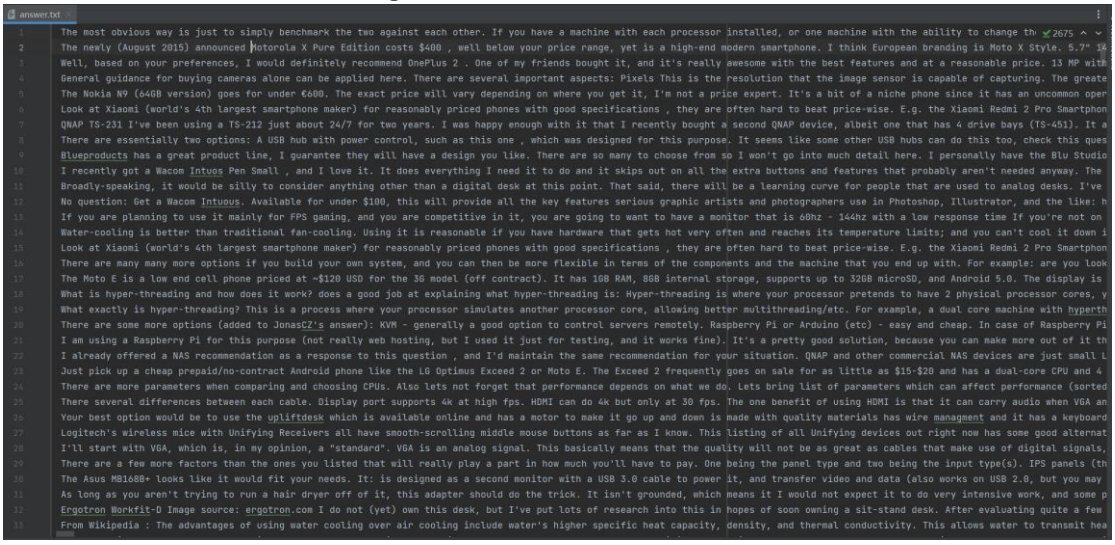
Directory

Task1 file name is parser_32891059.py. It is responsible for processing the data.xml content and filtering the collated data into articles of question and answer type, and storing the filtered corresponding content in the answer document and question document, respectively.

Operation process:

1. Place an xml file in the same folder as the preprocessData_32891059 file.
2. Open the Pycharm software
3. Install the pandas library (Install the Pandas library if you do not have one)
   Node: Ignore this if the Pandas library is installed
4. Run the preprocessData_32891059 file
5. question.txt and answer.txt are automatically generated and stored
6. question.txt and answer.txt are saved in the current folder by default

The result should look like the figure:



Task2 file name is preprocessData_32891059.py. It is responsible for cleaning up all the information such as id date type and content of the data.xml content. And output the cleaned data.

Operation process:

1. Place an xml file in the same folder as the preprocessData_32891059 file.
2. Open the Pycharm software
3. Add print output to the dataVisualization_32891059.py file by adding print (line_clean) at line 30

The result should look like the figure:

Task3 file name is dataVisualization_32891059.py. It is responsible for cleaning up all the information such as id date type and content of the data.xml content. And output the cleaned data.

Operation process:

1. Place an xml file in the same folder as the preprocessData_32891059 file.
2. Open the Pycharm software
3. Run the dataVisualization_32891059.py file
4. By default, the drawing box pops up and is saved in the current folder.

The result should look like these figures:

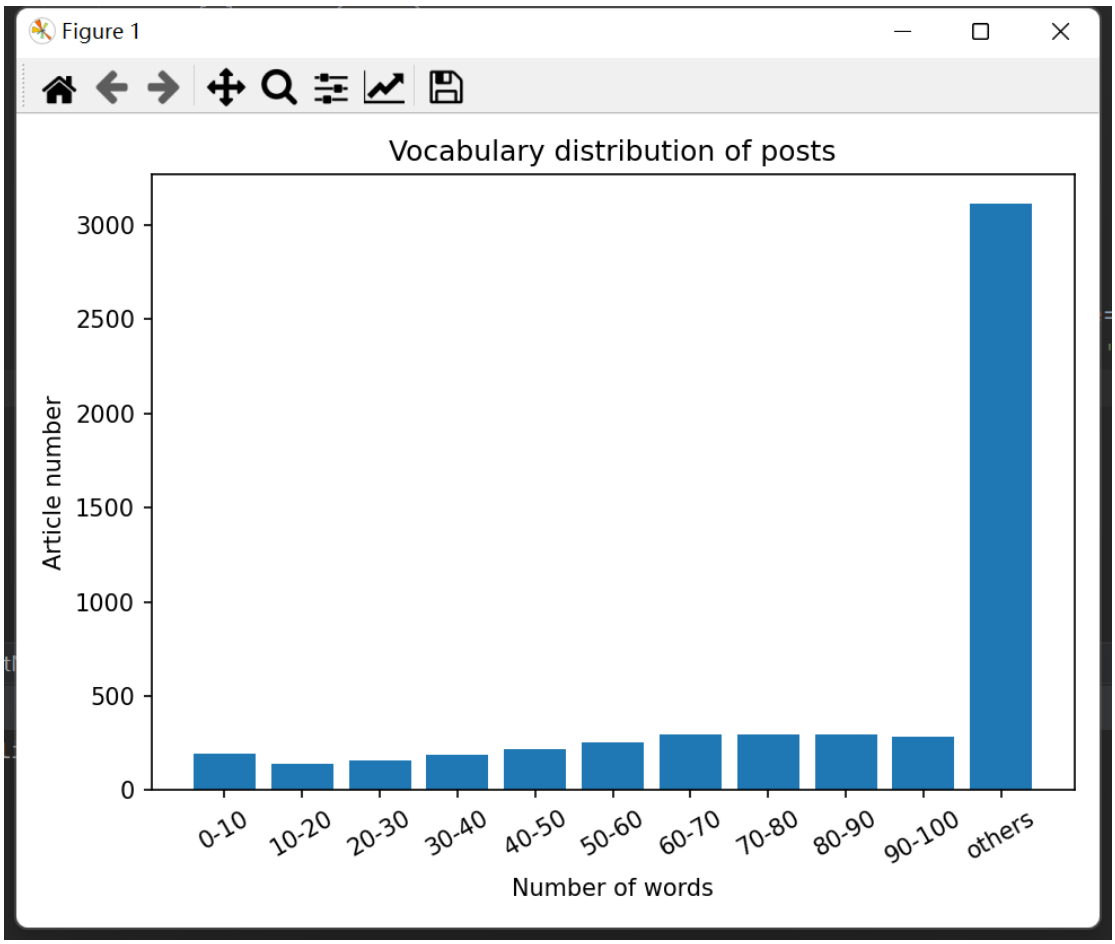


Figure 1: Vocabulary distribution of posts

Data source: data.xml

Analysis: According to the figure 1, we can see that the number of posts with more than 100 words is the largest, while the number of posts with between 10 and 20 words is the least. There is no significant difference in the number of posts between 0 and 100.

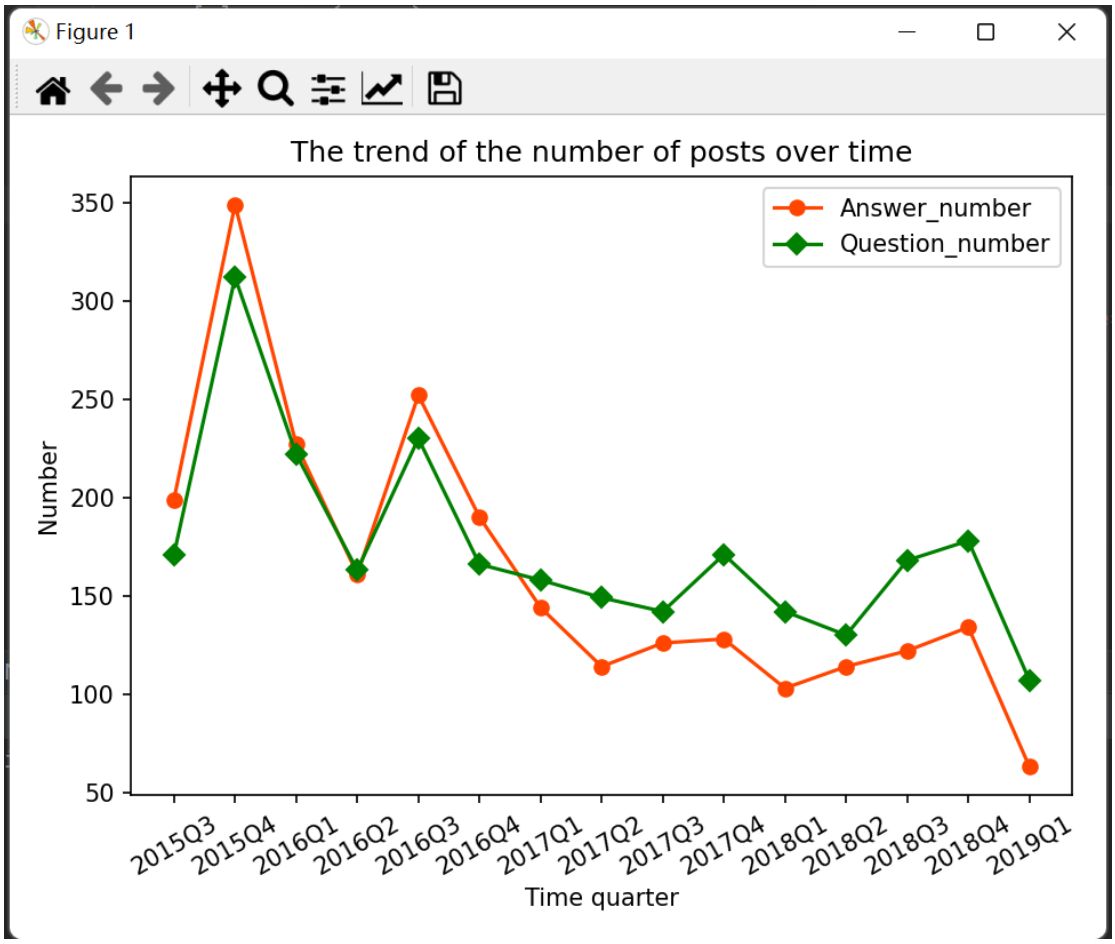


Figure 2: The trend of the number of posts over time

Data source: data.xml

Analysis: According to the figure 2. We can see the number of posts answering and asking questions in each quarter. In chronological order, the total number of posts answering and asking questions is decreasing. After the fourth quarter of 2016, the number of posts answering was significantly less than the number of posts asking questions. The number of posts answering and asking questions was the highest in the fourth quarter of 2015, at about 350 and 320, respectively.