

FIT9133 Assignment #2

Week11, May 2019

**Roopesh Kumar Ramesh**

Student ID: 30344565

Email: rram0019.student.monash.edu

Master of Data Science

Monash University

Contents

|  |  |  |
| --- | --- | --- |
| Serial No. | Title | Page No |
| 1 | Introduction | 3 |
| 2 | Task1: Handling with File Contents and Pre-processing | 3 |
| 3 | Task2: Building a Class for Data Analysis | 4 |
| 4 | Task 3: Analysing the File for Data Visualization | 5 |

Introduction

Assignment to implement a basic parser to investigate the natural-language posts from Q&A (Question and Answering) site. The parser can perform basic data extraction, statistical analysis on several linguistic features and also to present the analysis results using some form of visualization. For all three tasks, the provided template source code files have been used according to instructions. NumPy, Matplotlib and re libraries have been used.

Task1: Handling with File Contents and Pre-processing

In the first task, I began by reading in all the posts of the given dataset. I have then conducted several pre-processing tasks to clean the post content (Body) needed for analysis in the subsequent tasks (Task 2 and 3) in this assignment. Upon completing the pre-processing tasks, the content of questions and answers have been saved as two individual output files.

# How to run:

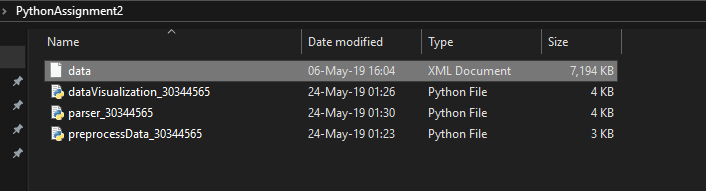


Figure 1: Files required to run program successfully

Open file ‘preprocessData\_30344565’ in PyCharm. Make sure file ‘data.xml’ is in the same directory as said file. Run file. There should be two additional files after the completion of the program.

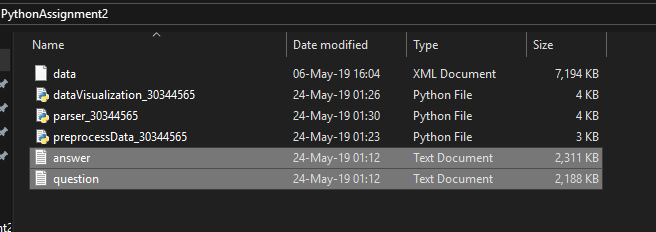


Figure 2: Folder structure after execution of program

### Output of Task 1

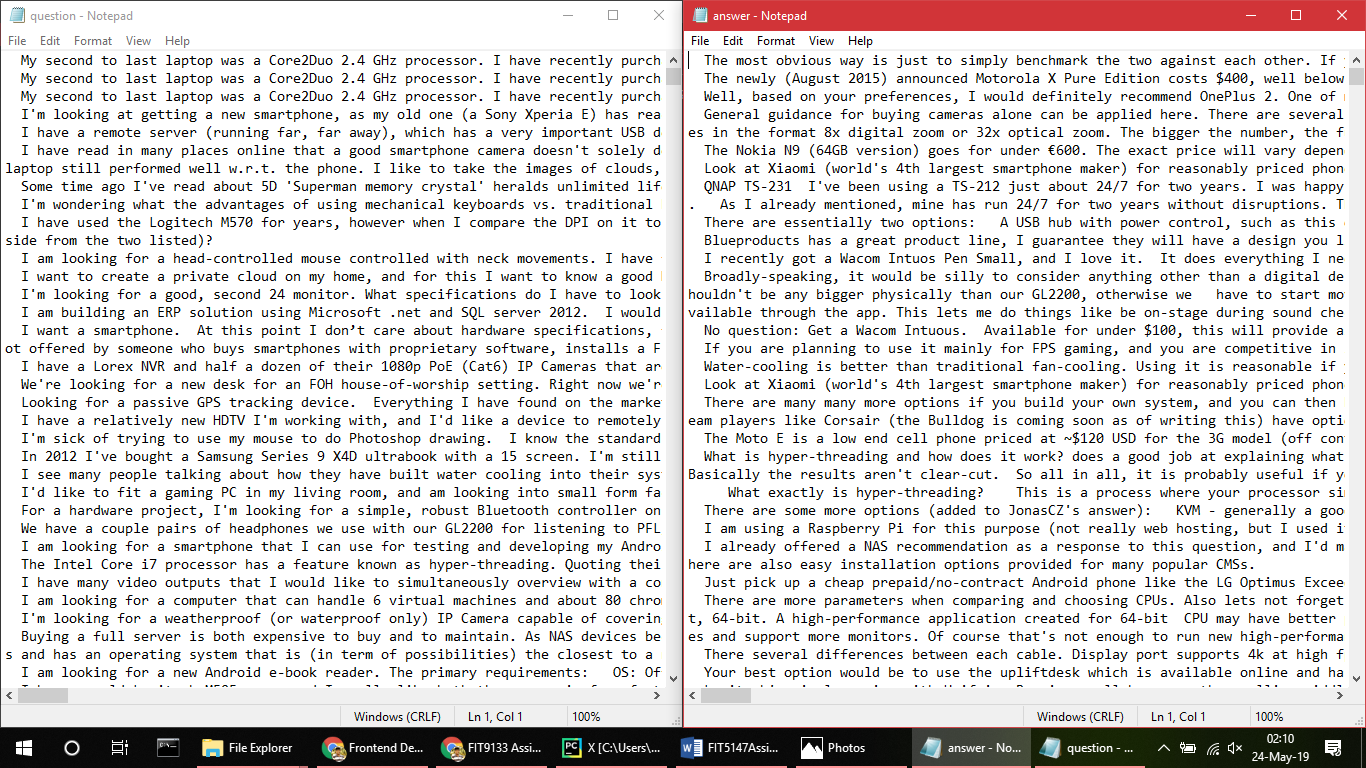
. 

Figure 3: Output of Task 1

Task2: Building a Class for Data Analysis

The second task is about collating the required data for analysis. Apart from extracting the clean body as achieved in Task 1, the main task here is to further parse the given row of the data in XML format with object-oriented programming.

My class “Parser” contains the methods to clean the data and get individual properties of the file.

# How to run:

Open file ‘parser\_30344565’ in PyCharm. Make sure file ‘preprocessData\_30344565’ is in the same directory as said file. Make sure to give input as string from data.xml file to main function.

### Output of Task 2:

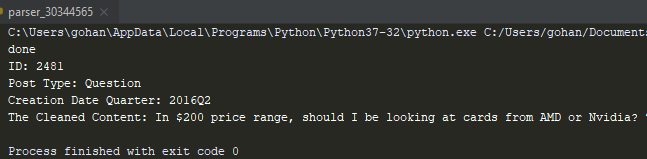


Figure 4: Output of Task 2

Task 3: Analysing the File for Data Visualization

In the last task, based on the class defined in Section 2.3 (Task 2), you will implement two functions to visualise the statistics as some form of graphs. The implementation of these two functions should make use of the external Python packages, including NumPy, SciPy, Pandas, and/or Matplotlib to create the suitable graphs for comparing the statistics collected for posts.

# How to run:

Have all 3 assignment files and data.xml file in the same directory. After successful execution, there should be two image files showing the plots.

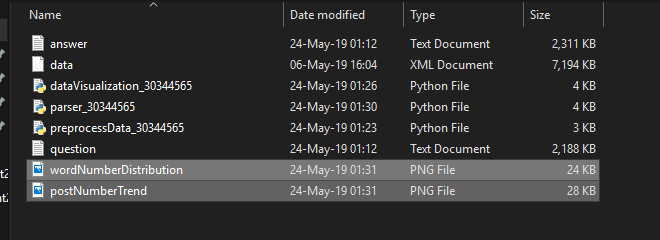


Figure 5: Folder structure after successful execution of visualization

## Output of Visualization

### Post Number Trend

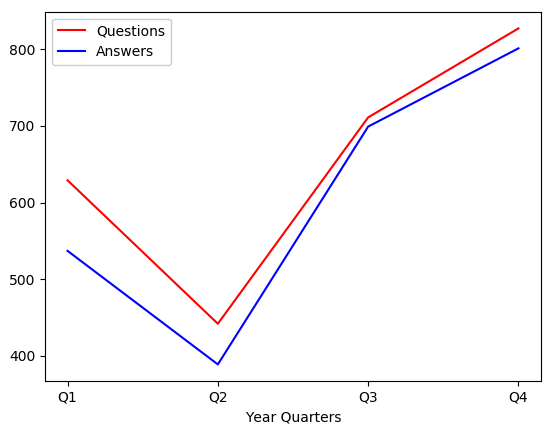


Figure 6: postNumberTrend

### Word Number Distribution

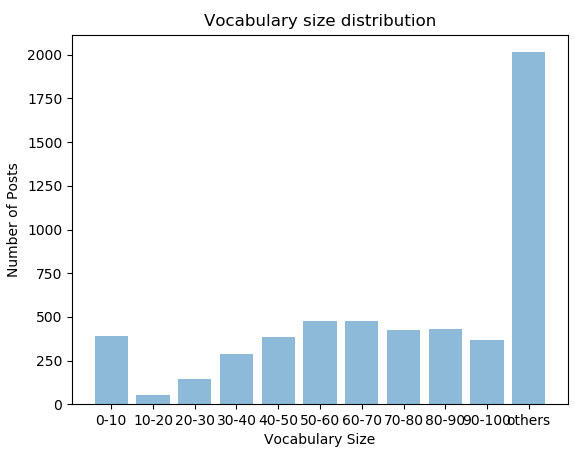


Figure 7: wordNumberDistribution