User Documentation

FIT9133 Assignment3

28453093, Zhengxin Tang

1. Instruction

This program contains five python files, six tok files for Stylometry Analysis and stopword_list.txt. The program analyzes six corpuses for William Shakespeare and Christopher Marlowe from many aspects.

- preprocessor_28453093.py
  This file contains a class Preprocessor. This class have one instance and is used to split an input text into individual tokens, then save them in the instance.
- character_28453093.py
  This file contains a class CharacterAnalyser. This class is used for analyzing characters from the given token list, including saving occurrences of each characters and punctuations into instance character_df and total number of characters in the given token list.
- word_28453093.py
  This file contains a class WordAnalyser. This class analyse words from many aspects. It counts the occurrences of different words, stop words and word length.
- visualiser_28453093.py
  This file contains a class AnalysisVisualiser. This class is mainly for drawing figure and show them for users.
- main_28453093.py
  This file contains the main function, which drives the flow of execution of the program. It has two menus: Stylometry Analysis Main Menu is for choosing different corpus to analyze or compare two corpuses or analyze all corpuses together; Frequency Analysis Menu is a child menu and can lead users to analyze for different frequencies, and them show the figures.
- Six tok files
  These six tok files contain all corpuses for analysis.
- Stopword_list.txt
  The stopword_list.txt contains stopwords which will be used in word analysis.

2. Screen shots

- Main menu:

```
----------Stylometry Analysis Main Menu----------
Please select the corpus for analysis:
[1]Edward_II_Marlowe
[2]Hamlet_Shakespeare
[3]Henry_VI_Part1_Shakespeare
[4]Henry_VI_Part2_Shakespeare
[5]Jew_of_Malta_Marlowe
[6]Richard_II_Shakespeare
[7]Analyse and compare two corpus
[8]Analyse all together
[9]Quit
```

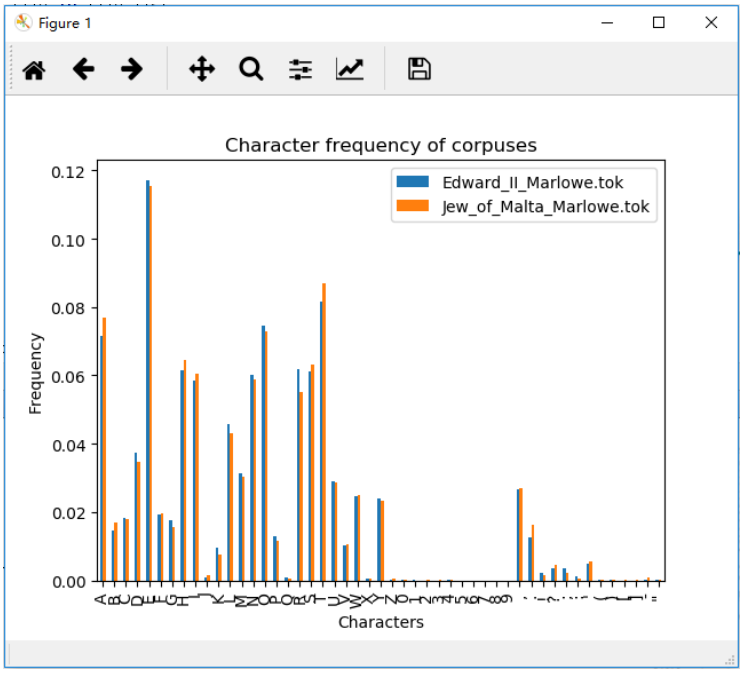
- Input 7 and then input 1 and 5, the child menu shows:

```
7
Please input the first corpus(from 1-6)1
Please input the second corpus(from 1-6)5
----------Frequency Analysis Menu(for two or all corpuses)----------
Visualising from which aspect?
[1]Character frequency
[2]Punctuation frequency
[3]Stopword frequency
[4]Word length frequency
[5]Quit
```

- Input 1 to analyze character frequency:

```
1
Analysing...this may cost some time...
```

- After analysis complete, the figure shows:



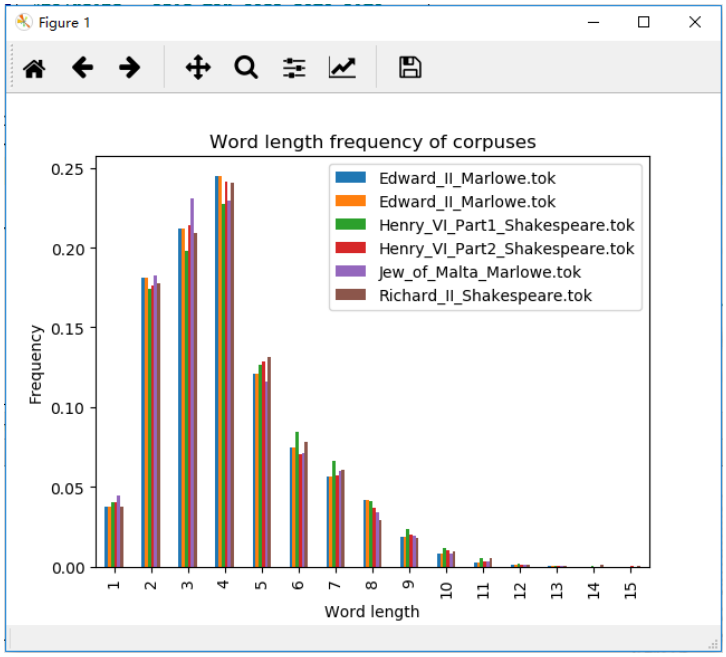
- Close figure, input 5 to quit and input 8 to analyze all:

```
[5]Quit
5
Quit visualising menu.
----------Stylometry Analysis Main Menu----------
Please select the corpus for analysis:
[1]Edward_II_Marlowe
[2]Hamlet_Shakespeare
[3]Henry_VI_Part1_Shakespeare
[4]Henry_VI_Part2_Shakespeare
[5]Jew_of_Malta_Marlowe
[6]Richard_II_Shakespeare
[7]Analyse and compare two corpus
[8]Analyse all together
[9]Quit
8
```

- Input 4 to analyze word length frequency:

```
----------Frequency Analysis Menu(for two or all corpuses)----------
Visualising from which aspect?
[1]Character frequency
[2]Punctuation frequency
[3]Stopword frequency
[4]Word length frequency
[5]Quit
4
Analysing...this may cost some time...
```

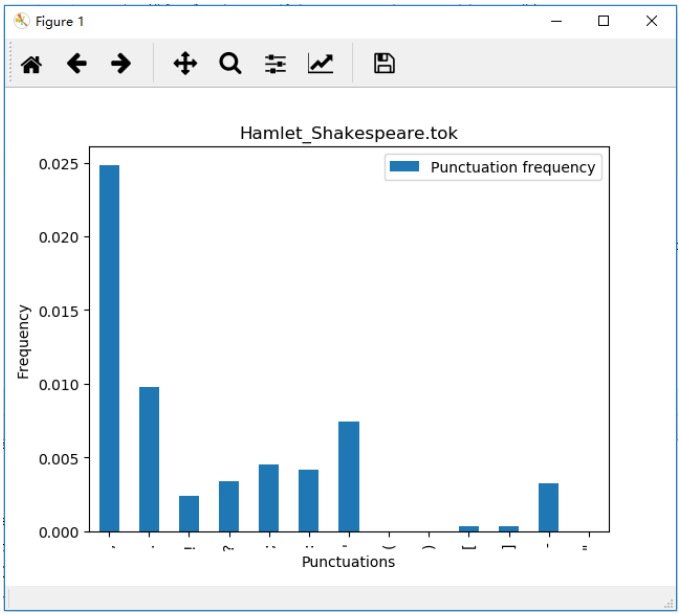
- After analysis complete, the figure shows:



- The same to other analysis selections. We can also go back to the main menu and analyze any single corpus:

```
----------Stylometry Analysis Main Menu----------
Please select the corpus for analysis:
[1]Edward_II_Marlowe
[2]Hamlet_Shakespeare
[3]Henry_VI_Part1_Shakespeare
[4]Henry_VI_Part2_Shakespeare
[5]Jew_of_Malta_Marlowe
[6]Richard_II_Shakespeare
[7]Analyse and compare two corpus
[8]Analyse all together
[9]Quit
2
36457tokens in the list.
Analysing...this may cost some time...
```

- Input 2 to analyze punctuations:



- After all analysis, input 9 in the main menu to stop the program.

3. Limitations

- All related python files need to be in the same directory, including six tok files and stopword_list.txt so that the program can run correctly.
- When analyzing, it may cost some time to get the output and plot the figure (approximately 30s to 1 minute).
- Because the stopwords are too many, it may be difficult to distinguish different stopwords in the figure.
- Users have to close the figure to input next command in the menu.