MONASH University

FIT 9133 – Programming Foundations Python – S1 – 2018
Assignment #3
Building a Stylometric Analyser

Anil Gurbuz -29628792
agur0005@student.monash.edu

Start and Last Modification Dates

Start Date: 16.05.2018

Last Modification Date: 27.05.2018

Contents

1 Introduction **4**

2 Tasks **4**

2.1 Task1 4

2.2 Task2 4

2.3 Task3 5

2.4 Task4 6

2.5 Task5 7

1.Introduction

In this assignment it is asked to build a code structure in Python to make stylometroc analysis. There are 5 different steps to create this program and every step is defined by different tasks. User is guided for usage of every task one by one.

2.Tasks

2.1 Task1

In this task, it is asked to define a class to pre-process the input texts. Tokenise module accepts its argument as a string type object and splits it according to spaces in the text in order to create a tokenised list.

2.2 Task2

Task2 is done to analyse characters in the given texts. Analyse_characters module uses a loop to take every character of the every element of tokenised list and stores it into pre-defined instance variable's data frame.

It stores the accurences in the pre-defined data frame so if there is a character that doesn't occur in the tokenised list, Its index will be there with value of 0.

To make this module work, user needs to have python installed pandas and numpy libraries and module will import them to use in the code.

User's input texts should not contain punctuations other than the list below.
, . ? : ; ' ! – " ( ) … [ ] + &

To let this module work, first of all user needs to tokenise its text using pre-processor class

2.3 Task3

This task firstly retrieves the content of given URL which consists of stop words and some other HTML code.

To enable python to retrieve the content of web page, user needs to be connected to the internet otherwise program will not work.

To make this module work, user needs to have python installed urllib.request, pandas and numpy libraries and module will import them to use in the code.

This module creates a file named stop words in the directory that python project is opened.

Words that contains just one character counted as a word.

Word length of words like ‘Hamlet’s’ are counted as 8 characters.
‘s part is not counted as another word.

To let this module work, first of all user needs to tokenise its text using pre-processor class

2.4 Task4

In this task there are 4 modules to visualise the analysis done in the previous tasks.

One of them draws a graph that contains characters on the x-axis and the number of occurrences of each character on y-axis. If the user wants to use this module, he/she first needs to make a character analysis using CharacterAnalyser class.

One of them draws a graph that contains punctuations on the x-axis and the number of occurrences of each punctuation on y-axis. If the user wants to use this module, he/she first needs to make a character analysis using CharacterAnalyser class.

One of them draws a graph that contains word lengths on the x-axis and the number of occurrences of each word length on y-axis. If the user wants to use this module, he/she first needs to make a word analysis using WordAnalyser class.

One of them draws a graph that contains stop words on the x-axis and the number of occurrences of each stop word on y-axis. If the user wants to use this module, he/she first needs to make a word analysis using WordAnalyser class.

Each method creates a png file that stores the image of graph of character analysis, stop word analysis, punctuation analysis or word length analysis. These files are created in the same folder with Python Project and they are created when this module is run. Note that this .png files has a limited resolution so for stop word analysis, user needs to zoom the graph that comes as output of the program in order to make the graph readable.

2.5 Task5

In this task, all of the previous tasks are collected together to create the whole program.

First there is a get_input module to prompt user for entering the file locations that will be analysed.

The sequence of this input is important because the name of the file which is first written will be named as File1 at the output and last file entered will be named as File6 at the output.

Because some of the output graphs contains many labels on x-axis (for example the graph visualising stop word analysis), it is impossible to read the labels on x-axis. User needs to zoom in the graph and easily reads whole labels and their corresponding y value.

In order to make the program work, user must enter exactly 6 number of texts.

User must enter file names or paths with a space between each other.

Program gives a future warning after termination but it works fine and exit with exit code 0.

There is a command in this task to save the content of data frame as a text file that stores all the data of analysis. This file will be created on the same directory with python project when the program run. Name of the text file is All_data.txt