MONASH University

FIT9133 Programming foundations in python

Assignment - 2

Building a Child Language Analyser
with Python Programming

Semester 2 2018

Dinesh Karthikeyan
Student Id: 29592461
Course: Master of Data Science

Table of Contents

Program Execution Requirement

The project is developed with **python 3.6.0** and **the PyCharm Edu 2018.1.3** is the **IDE used to develop.** To execute the project using command line, ensure that python 3.6.0, pandas, NumPy, matplotlib packages are installed in the system. Open command prompt and use **cd-Current Directory** command to change to the directory where the project details are kept and place the ENNI Dataset folder with sub folders SLI and TD that holds SLI and TD files respectively. Now check for python package installation by typing **python** and pressing **Enter key**. Once verification is done and if **python** is present then the project can be executed by **python file_name.py** (file_name should be the name of the project file i.e. task1_29592461.py,task2_29592461.py and task3_29592461)

Assumptions:

The following are the assumptions considered from assignment 2 specification while developing the project:

1. ENNI Dataset is placed in the Project folder.
2. Length of Transcript is the count of statements that ends with ".", "!" or "?"'.
3. File read first is the 10th file (i.e.) SLI-10.txt or TD-10.txt.
4. Using Pandas for Data Frame Creation.
5. Using MatplotLib for displaying bar graph.
6. [/] represents Repetition.
7. [//] represents Retracing.
8. (.) represents Pauses.
9. [*m] and [*m:+ed] represents grammatical errors
10. Each task can be run separately , so importing one task in another.

Task Description

Task 1: FileParser Class:

The purpose of this task is to clean the files in the ENNI Dataset/SLI and ENNI Dataset/TD folders so as to have only the child statements with the CHAT symbols by removing the examiner comments and the expressions made by the child. The method parse_clean does this task by reading the files into a list using OS package and each file is opened and the lines are read into a list and then lines starting with "*CHI:" are taken (i.e.) both single line and multi-line child statements are taken and written into cleaned file and placed in cleaned folder (SLI_Cleaned or TD_Cleaned) based on the group of the child, then the cleaned file is opened and lines are read into a list, now the cleaning of symbols "<",">" is done at the line level, then the line is split into words based on the spaces between and "(",")" are cleaned retaining the symbol "(.)" and then words that are enclosed within "[" and "]" are cleaned retaining the symbols "[/]", "[//]", "[*m]" and "[*m:+ed]". After cleaning all these the cleaned content is written into the cleaned file again. Now the Cleaned files in the output path holds fully cleaned child statements.

Task 2: DataAnalyser Class:

The purpose of this class is to analyse Length of Transcript, Number of Repetition, Number of Retracing, Number of Grammatical Errors, Number of Pauses and Size of Child's Vocabulary. This class has 6 lists (i.e.) one for each statistics to be analysed. The analyse_script method takes one cleaned file as input and analyses the file and populates the list. Length of Transcript is calculated by the count of lines ending with ".", "!" or "?". Number of repetition is calculated with count of chat symbol "[/]", Number of retracing is calculated with count of chat symbol "[//]", Number of Grammatical Errors is calculated with the count of chat symbol "[*m]" and "[*m:+ed]", Number of pauses is calculated with the count of chat symbol "(.)" and Size of Vocabulary is calculated by count of unique words in the file ignoring the chat symbols and punctuations. This is run for all the files in SLI_Cleaned and TD_Cleaned folder with 2 objects one for SLI and one for TD respectively. The __str__ method is overloaded so as to display the statistics' list for entire group. To help Task 3 these objects are populated in a list and returned in the main function. Task 1 is imported and main of Task 1 is called so as to enable running of Task 2 separately.

Task 3:DataVisualisation Class:

The purpose of this class is to show the comparison of mean statistics from Task 2 between SLI and TD Group. This class has a Pandas Data Frame as an instance variable and the data from Task 2 is the input for the constructor, the lists in the Task 2 objects are retrieved and populated into local dictionaries with the name of the statistics as the keys and the lists as the values for those keys. These dictionaries are added into a list and passed into the Data Frame constructor to form the Data Frame (instance variable). The compute_average method is used to calculate mean of the statistics and store in the same Data Frame.
The visualise_statistics method is used to save and display a graph that compares SLI and TD with the means of those 6 statistics with bar chart. The __str__ method is overloaded so as to display the Data Frame. Pandas Package, NumPy Package and Matplotlib package are imported for displaying graph and for using DataFrame. Task 2 is imported so as to run Task 1 and Task 2 and main of Task 2 is called which enables Task 3 to run independently.

Output Screenshots

Task 1

```
File is cleaned and saved in SLI_Cleaned as SLI-10_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-1_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-2_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-3_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-4_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-5_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-6_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-7_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-8_cleaned.txt
File is cleaned and saved in SLI_Cleaned as SLI-9_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-10_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-1_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-2_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-3_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-4_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-5_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-6_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-7_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-8_cleaned.txt
File is cleaned and saved in TD_Cleaned as TD-9_cleaned.txt
```

Figure. 1

Figure. 1 shows the display after Task 1 is run and the cleaned files being written into respective cleaned folder.

Task 2

```
                                STATISTICS
                                ____________________

SLI :
  Length of Transcript : [57, 67, 70, 106, 68, 77, 61, 68, 72, 70]
  Size of Vocabulary : [123, 126, 114, 148, 135, 160, 103, 149, 148, 137]
  Number of Repetition : [14, 47, 5, 39, 21, 9, 14, 28, 8, 45]
  Number of Retracing : [13, 10, 11, 5, 44, 18, 10, 12, 13, 10]
  Number of Grammatical Errors : [1, 1, 2, 0, 1, 0, 0, 0, 0, 0]
  Number of Pauses : [22, 12, 40, 16, 45, 36, 11, 7, 40, 22]

TD :
  Length of Transcript : [91, 94, 90, 81, 86, 90, 84, 76, 90, 81]
  Size of Vocabulary : [206, 116, 182, 200, 177, 165, 183, 179, 170, 194]
  Number of Repetition : [6, 14, 8, 21, 48, 9, 18, 21, 10, 23]
  Number of Retracing : [20, 11, 11, 22, 7, 21, 23, 22, 11, 15]
  Number of Grammatical Errors : [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
  Number of Pauses : [28, 24, 53, 39, 18, 38, 41, 52, 25, 41]
```

Figure 2

Figure. 2 shows the Lists for the statistics printed by overloaded __str__ method in Task 2. These lists hols the values of all the files from the cleaned folders from like 10,1 to 9 for both SLI and TD.

Task 3

```
                                TABLE OF STATISTICS
                                ___________________

                  Length of Transcript   Number of Grammatical Errors                  Number of Pauses            Number of Repetition                Number of Retracing
                Size of Vocabulary
SLI  [57, 67, 70, 106, 68, 77, 61, 68, 72, 70]  [1, 1, 2, 0, 1, 0, 0, 0, 0, 0]  [22, 12, 40, 16, 45, 36, 11, 7, 40, 22]  [14, 47, 5, 39, 21, 9, 14, 28, 8, 45]  [13, 10, 11, 5, 44, 18, 10, 12, 13, 10]  [123, 126
, 114, 148, 135, 160, 103, 149, 148, ...
TD    [91, 94, 90, 81, 86, 90, 84, 76, 90, 81]  [0, 0, 0, 1, 0, 0, 0, 0, 0, 0]  [28, 24, 53, 39, 18, 38, 41, 52, 25, 41]  [6, 14, 8, 21, 48, 9, 18, 21, 10, 23]  [20, 11, 11, 22, 7, 21, 23, 22, 11, 15]  [206, 116
, 182, 200, 177, 165, 183, 179, 170, ...
```

Figure 3

Figure. 3 shows Table of statistics printed from the Data Frame using overloaded __str__ method in Task 3. Columns represent the statistics type and rows represent the values for the statistics for the corresponding group of children.

MEAN OF STATISTICS

	Length of Transcript	Number of Grammatical Errors	Number of Pauses	Number of Repetition	Number of Retracing	Size of Vocabulary
SLI	71.6	0.5	25.1	23	14.6	134.3
TD	86.3	0.1	35.9	17.8	16.3	177.2

Figure. 4

Figure. 4 shows Mean of the statistics printed from the Data Frame using overloaded __str__ method in Task 3. Columns represent the statistics type and rows represent the mean values for the statistics for the corresponding group of children.

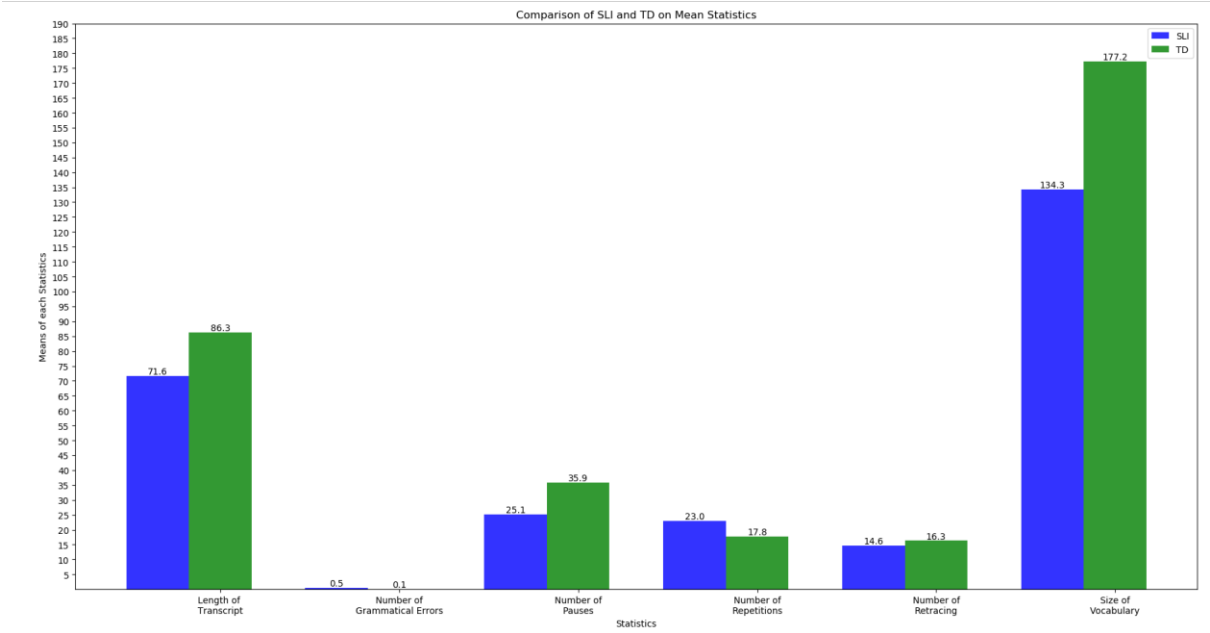

Figure. 5

Figure. 5 shows the graph depicting the statistics in a bar chart comparing SLI and TD based on the Mean values of those statistics.

Reference:

Matplotlib. (n.d.). Retrieved from https://matplotlib.org/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py

Matplotlib. (n.d.). Retrieved from https://matplotlib.org/api/_as_gen/matplotlib.pyplot.xticks.html

Matplotlib. (n.d.). Retrieved from https://matplotlib.org/api/_as_gen/matplotlib.pyplot.yticks.html

pandas.dataframe.values. (n.d.). Retrieved from https://pandas.pydata.org/pandas-docs/version/0.23/generated/pandas.DataFrame.values.html

Python,R and Linux Tips. (n.d.). Retrieved from https://cmdlinetips.com/2018/01/how-to-create-pandas-dataframe-from-multiple-lists/

PythonHow.com. (n.d.). Retrieved from PythonHow.com: https://pythonhow.com/accessing-dataframe-columns-rows-and-cells/

StackoverFlow. (n.d.). Retrieved from https://stackoverflow.com/questions/30228069/how-to-display-the-value-of-the-bar-on-each-bar-with-pyplot-barh