ASSIGNMENT -2

Building a Child Language Analyser

SUBMITTED BY-Suyogya Taneja(28961897)

Suyogya Taneja

[Email address]

TABLE OF CONTENTS

INTRODUCTION

In this assignment, a basic language analyser is implemented to investigate the linguistic characteristics of children with some form of language disorders has been implemented. The analyser is able to perform basic statistical analysis on a number of linguistic features and also to present the analysis results using some form of visualisation.

TASK-1 :HANDLING WITH FILE CONTENTS AND PRE-PROCESSING

In this task the the data sets of SLI and TD are refined.In this task **glob module** has been used to find the path names of the .txt files.

The filtering of the symbols has been done on the following basis:

- Removed those words that have either '[' as prefix or ']' as suffix1 but retained these three symbols: [//], [/], and [* m:+ed]
- Retained those words that have either '<' as prefix or '>' as suffix but these two symbols have been removed
- Removed those words that have prefixes of '&' and '+'
- Retained those words that have either '(' as prefix or ')' as suffix but these two symbols have been removed.

The output of the cleaned dataset is stored in the following two files:

1. SLI_CLEANED.txt
2. TD_CLEANED.txt

TASK-2: BUILDING A CLASS FOR DATA ANALYSIS

In this task the cleaned data of both the groups of children transcripts are analysed.
The class for data analysis is named as **STATISTICS**. Following are the methods in the class Statistics.

1. len_of_transcript(self,path)------This method is used to calcualate the length of the transcript i.e is indicated by the number of statements.

2. size_of_vocabulary(self,path)------This method is used to calculate the size of the vocabulary i.e is indicated by the number of unique words.
3. repitition(self,path)--------This method is used to calculate the number of repetition for certain words or phrases.- indicated by the chat symbol[/].
4. retracing(self,path)---------- This method is used to calculate the number of retracing for certain words or phrases.- indicated by the chat symbol[//]
5. grammar_error(self,path)--------This method is used to calculate the number of grammatical errrors that are detected.- indicated by the chat symbol[* m:+ed]
6. pauses(self,path)----------This method is used to calulate the number of pauses made- indicated by the chat symbol (.).
7. __init__(self)----This is a constructor that creates the intances of the class.This method uses dictionary to store all the statistics. The statistics values are stored in a list as a value in the dictionary.

8. __str__(self)--------This method presents the output in readable format and returns a formatted string .
9. Analyse_script(self , cleaned_file)----------This method performs analysis on the cleaned scripts.It accepts the cleaned script as the argument and extracts the required data for analysis.

FOLLOWING ARE THE STATISTICS OF THE TWO SCRIPTS:

	SLI_CLEANED.txt	TD_CLEANED.txt
Length of the transcript	939	1106
Size of the vocabulary	710	805
Number of Repetition of words	218	166
Number of retracing of certain words	127	154
Number of Grammatical errors	4	1
Number of pauses made	236	333

```
C:\Users\Suyogya\Anaconda3\python.exe C:/Users/Suyogya/ASSIGNMENT-2/task2_28
          CLEANED FILE    LENGTH OF THE TRANSCRIPT    SIZE OF THE VOCABULARY
       SLI_CLEANED.txt                         939                       710
        TD_CLEANED.txt                        1106                       805

Process finished with exit code 0
```

Figure 1

```
61897.py
   REPITITION OF WORDS    NUMBER OF RETRACING OF CERTAIN WORDS    NUMBER OF GRAMMATICAL ERRORS DETECTED    NUMBER OF PAUSES MADE:
      218                                    127                                    4                                    236
      166                                    154                                    1                                    333
```

Figure 2

TASK-3: BUILDING A CLASS FOR DATA VISUALISATION

In this task a class named Visualise has been implementated. The data set of both the groups has been used to create suitable graphs using **matplotlib**.

The methods used in this method are as follows:

1. __init__(self)----------This method uses **PANDAS** dataframe to store the data that is to be ploted.

2. __str__(self)--------This method presents the output in readable format and returns a formatted string.

3. Compute_averages(self)--------- This method returns the average or mean of the six statistics for each child group (i.e.both the SLI and TD groups).

4. Visualise_statistics(self)---------- This method constructs the BAR graph to demonstrate the mean difference between the two groups i.e SLI and TD for each of the six statistic for comparison purpose.

OUTPUT-2:**AVERAGE OF THE STATISTICS OF TWO STATISTICS:**

```
C:\Users\Suyogya\Anaconda3\python.exe C:/Users/Suyogya/ASSIGNMENT-2/task3_28961897.py
Average of STATISTICS OF SLI_CLEANED.txt: 372.333333333
Average of STATISTICS TD_CLEANED.txt: 427.5
```

OUTPUT-3:**BAR GRAPH TO SHOW THE MEAN DIFFERENCE BETWEEN THE TWO STATISTICS:**

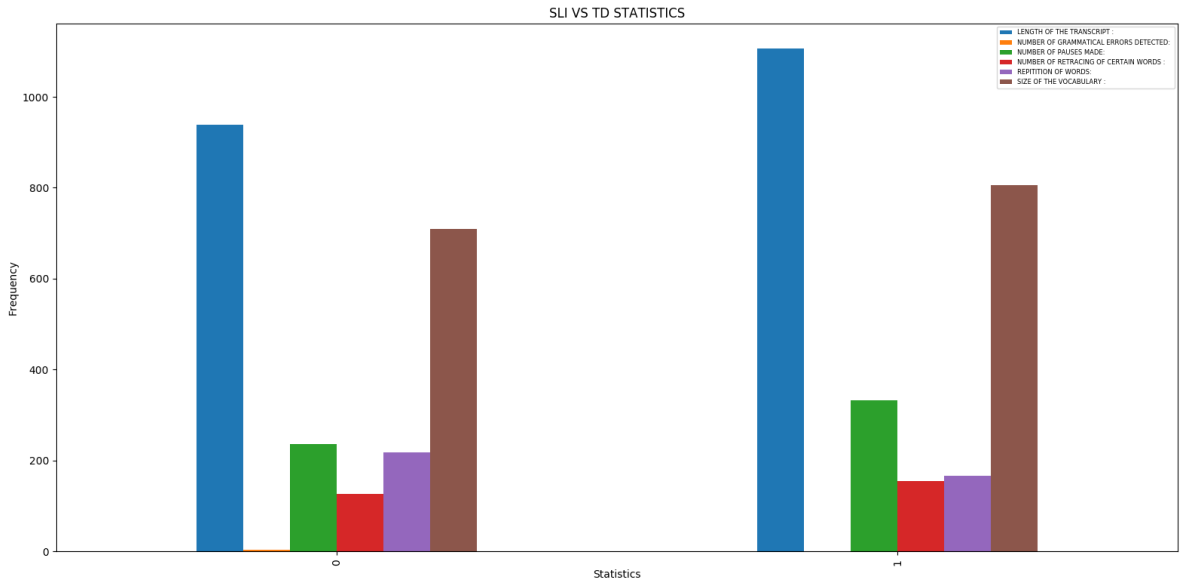

Figure 3

HOW TO RUN

There are following files are in the zipped foler named A2_28961897:

1. task1_28961897.py
2. task2_28961897.py
3. task3_28961897.py
4. SLI_CLEANED.txt
5. TD_CLEANED.txt
6. TD_DATASET
7. SLI_DATASET

All three files are in a zipped folder. FOLLOW the below mentioned steps to run the program .

1. Unzip the folder and extract all the files.
2. After extracting the files run the files in any python enviornment preferrably PYCHARM or Jupyter notebook.
3. **ALL the files except for SLI_CLEANED.txt and TD_CLEANED.txt should be in the same directory.**
4. The files should run in order i.e task1_28961897, task2_28961897, task3_28961897 .
5. When you run the **task1_28961897,** the dataset folder of SLI and TD should be in the same folder i.e the current working directory.
6. After you run the **task1_28961897** , you get two text files **i.e SLI_CLEANED.txt** and **TD_CLEANED.txt** which is in the same folder which your code is running**.**
7. After first task, run the **task2_28961897 .** When you run this task you will get all the required statistics as the output.
8. After task2, run the file **task3_28961897.** When you run this task you will get two outputs:
    1. AVERAGE OF the statistics of SLI_cleaned and TD_cleaned
    2. BAR GRAPH that shows the mean difference between the two groups SLI vs TD.

CONCLUSION

A basic language analyser has been implemented by using the given data sets . The datasets SLI and TD have been cleaned and then different statistics have been calculated . After calculating the differnet statistics their average mean has been calculated. At the end , on the basis of the statistics calculated for both the groups a BAR graph has been plotted that shows the mean difference between the two groups SLI and TD.