FIT9133 Assignment 2
Building a Child Language Analyser

Name	Srikar Manthatti
Student_id	29803306
Student_email	sman0027@student.monash.edu
Course	Master Of Data Science

Contents:

Task 1: Handling with File Contents and Preprocessing:

In the given dataset, we have to extract the statements which have *CHI: at the starting. For this task, I have created two functions named first_cleaning, second_cleaning . These two functions will take two arguments frompath, topath.
Frompath: this argument holds the value of source location from the files will be picked.
Topath: this argument holds the values of destination location to which the files will be stored.
The first_cleaning function will pick up the raw files and will extract the statements which have *CHI: at the starting of the statement and will be stored in the Middle folder.
The Second_cleaning function will pick the files from middle folder which will further be used to clean the dataset and will remove the contents like '<','>','(','&','+'. After cleaning the files all the cleaned files will be stored in the a separate folder.

Task 2: Building a Class for Data Analysis:

In this task we will generate the statistics for the above cleaned transcripts. These statistics can be used as the good indicators for distinguishing between children between SLI and TD. A class named DataAnalyze has been created which has the following functions: __init__(self), __str__(self), analyse_script(self,cleaned_file). This init function will used to initialize the variables, and string function is an override function which will return the inbuilt return string. The analyse script is used to find the number of len-of_transcipt, repeating_words, retracing_words, grammatical_errors, pauses. The function will take two arguments, one is self which represents the object and the cleaned_file which will represent the file which has been cleaned in task 1. In this python script we have one more class named Count, in this count class we will declare the path of the files in different variables, respective objects for the different files will be created and the analyse_script will be called.

Task 3: Building a Class for Data Visualisation:

In task 2 after getting all the required statistics those values should be passed to task3 to visualize the data. In this task 3 python script, we have one class named Visualize which has 3 functions in it __init__(self,datalist), compute_averages(self), visualize_statistics(self). his initiator function is used to initialize the few variables. This init function will take two arguments, one is self which is th eobject itself.

The other argument is the datalist which will be returned in the DataAnalyse call of task2.py module. The compute_averages will be used to calculate the averages. The datalist will be used to calculate the averages. Since the return statement in the function will return the filename, statements_avg, vocabavg, words_repeat_avg, words_retrace_avg, grammar_errors_avg, Pauses_avg respectively. The calculated average values will be accessed by the visualize_statistics functions and will be represented as a bar graph for easy understanding.

Assumptions made:

1. The machine in which this script is running should have python installed in it.
2. The location of the input files mentioned in the code should be same as the location where the input files are located.
3. An output folder for storing the cleaned files should be created upfront.
4. Module os, numpy, matplotlib must have been present in the machine.

How to run the program:

1. Unzip the folder and save all the files at a specified location.
2. In all the python programs, please change the sli_file_path, td_file_path, sli_dest_path, td_dest_path accordingly.
3. In the task1 and extra variable which store a buffer folder named Middlefolder, please change the address of the variable also.
4. After making the required changes, right click on the program and select run the program.
5. Follow the same procedure for the task 2 also, the output values will the displayed on the console.
6. For task 3 also, please change the location values accordingly and run the program. An output will be generated in the other window which shows the bar graph plots.