AC50002 PROGRAMMING LANGUAGES FOR DATA ENGINEERING
PYTHON-ASSIGNMENT

JITENDRANATH CHOWDARY KARUTURI (2557461)

Assignment overview:

- The Python assignment is to write a python program that generates three- letter abbreviations for each word based on specific rules. This program is designed to read names from an input file and outputs a file with a list of accepted abbreviations. The program calculates score for each abbreviation based on the defined rules and adheres to those abbreviations that scored the least.

Design components:

1. Reading Character Values:[1]
   - The **function load_values(value_file_path):** read character values from a file and returns a sorted dictionary based on numerical values. We take the file “values.txt” and create a dictionary based on the assigned numerical values for a word. We use this function to calculate a score for the abbreviation.
   - The program starts by reading character values from the "values.txt" file and creating a dictionary sorted by numerical values. This data is crucial for scoring each letter in the abbreviations.
2. Finding Least Letter in a Word[2]:
   - This program **find_least_char(word, sorted_char_values):** identifies the least letter in a word. The least letter is decided based on character values and score based on its position. This function finds the least letter and its score in a word based on the provided rules.
3. Generating Abbreviations and Scores[3]:
   - The core function, **generate_abbreviation(name, sorted_char_values)**, processes each name. It tokenizes names into words, and based on the number of words, it constructs abbreviations and calculates scores. Special attention is given to handling edge cases and applying the scoring rules. This function generates abbreviations and scores for names based on the number of words and specific rules.
4. Main Processing Function:
   - The main function adapts the entire process. It prompts the user for the input file and surname, calls the necessary functions, and writes the results to an output file. The function ensures proper formatting and handling of names with apostrophes.
5. Output -> output/surname _trees_abbrevs.txt
   - Creates the output filename as surname +'_'+ input_filename + '_abbrevs.txt'

Procedure to run the code: (save all the files in one folder)

1. Open **abbreviations assignment.py** and provide the path to values file in **def main():** function line 118.
2. The python code was written using the **visual studio** and executed in its terminal. After executing, it will prompt you to provide the **input file name** and **surname** by which the output needs to be saved.
3. Running the script will produce an output file in the folder with appropriate abbreviations to the given list of words.

---

1 https://openbookproject.net/thinkcs/python/english3e/ chapter 2,4,5

2 https://openbookproject.net/thinkcs/python/english3e/ chapter 15,16,21,22

3 https://openbookproject.net/thinkcs/python/english3e/ chapter 5,6, 20,22

Figure 1 prompts

Testing and Validation:

- The program can be tested by providing sample input files containing names. It allows users to verify the generated abbreviations and scores. For better testing accuracy two letter words and few words like name (Jitendranath Chowdary Karuturi), Python Programming Language has been added to the trees.txt file.
- The values.txt file needs to be correctly formatted with character values to ensure accurate scoring.

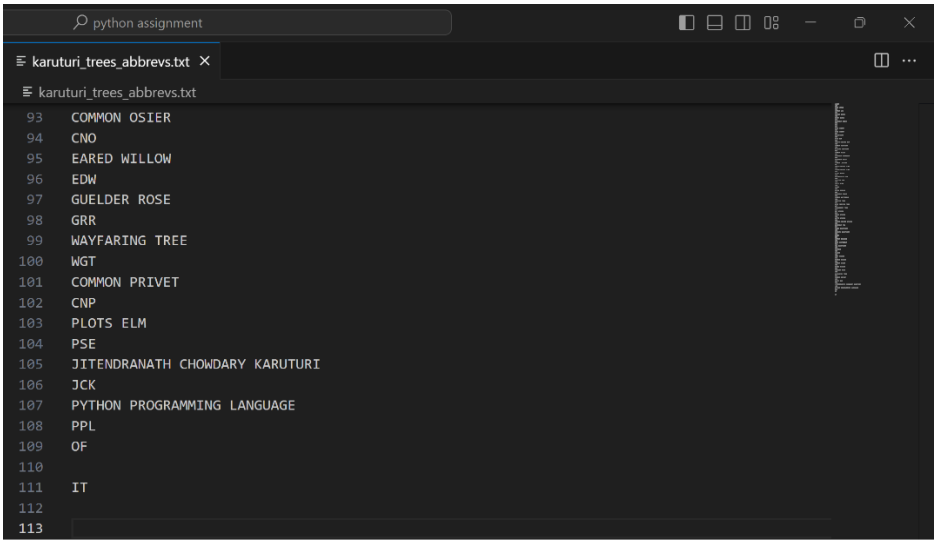

Figure 2 output file

- In conditions where the number of letters in a word is less than 3, it returns an empty string. This is because, the desired abbreviation for a word is supposed to have count of 3 letters and on that condition, given word has count of characters less than 3 cannot produce a valid abbreviation. so, for the given two letter words OF and IT the result has been returned as an empty string.
- For example, the final abbreviation for the word GUELDER ROSE is GRR.the score for G is 0 because it's the first letter, score for R is 5 because it's the last letter of the first word and the score of second R is 0 because it's the first letter of the word. So, the combined score is 5, which is less than all possible scores. Similarly for the remaining words the least score has been calculated for all the possible abbreviations and the least score abbreviation has been produced.
- Second example the input is Crab Apple, the output should be ADR based on the score, the score for A is 0 because it's the first letter, for the letter D score is 11 because the value for D is 9 and the index is 2 and for R score is 5 because it's the last letter. The combined least score is 16. The resulted abbreviation output is ADR.

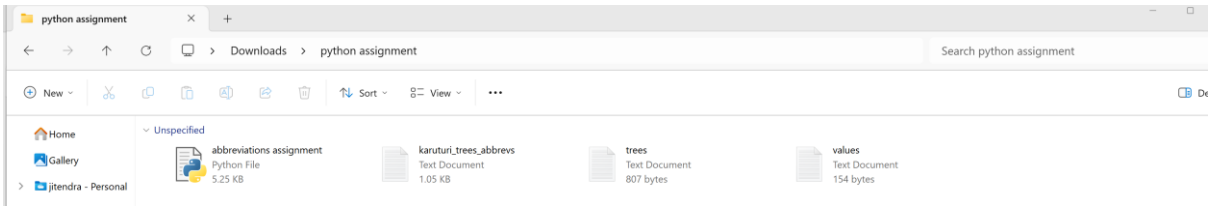

Figure 3 folder snip