

AC50002 Programming Languages for Data Engineering Python Assignment

Name: Ahmad Zia Bahrami



December 18, 2023

Dundee University

DD1 4HN

# Introduction:

This concise report pertains to the initial assignment within the Python for the Programming Language for Data Engineering Module. The assignment's primary objective is to fulfil the specified task, which involves reading a list of names from input files, generating three-letter abbreviations for each name, and computing a score for each abbreviation according to predefined rules. To accomplish this task, the following steps have been undertaken:

1. **Read Values Function:**

Through this function I would like to read letter-values pairs from the values.txt file which already provided by the professor for the fulfilment of this assignment and return a dictionary containing letter-value pairs.

1. **Tokenize Name Function:**

According to the task’s rules, this function will tokenize the input name into words, removing apostrophes and non-alphabetic characters. This function will receive a name string as input and generate a list of words without apostrophes and split based on non-alphabetic characters as an output.

1. **Calculate Score Function:**

According to the rules which defined in the assignment task, this function will Calculates the score for a given letter based on its position in a word and predefined values. The function will receive letter, position, and values dictionary as argument and will return a score based on letter position and values.

1. **Generate Abbreviations Function:**

According to the assignment task’s rules, as well as the example which provided in the task explanation this function is going to generates three-letter abbreviations for the input name and calculates the total score. The input for the function as name string and values dictionary and after the process the function will return a tuple containing the abbreviation and its score.

1. **Choose Best Abbreviation Function:**

Generate Abbreviations function will return as list of abbreviation for each name. The purpose of this function is to select the best abbreviation from a list based on the lowest score. The input for this function is a list of abbreviation-score tuple. The function after the process will return the best abbreviation with the lowest score.

1. **Main Function (main):**

The main function serves as the orchestrator of the entire program. Its primary role is to seamlessly guide users through the input process, ensuring accurate information is provided. This includes collecting the user's surname and the filename containing a list of names.

Furthermore, the function validates the file extension, specifically checking for the required '.txt' extension. Once confirmed, it extracts essential information, such as the input file name (excluding the extension) and retrieves letter values from an auxiliary file named 'values.txt'.

Upon successful validation and information extraction, the main function initiates the core processing of the input file. For each line (name) in the input file, the program generates abbreviations based on predefined rules and calculates a score for each abbreviation.

The results, including the original name and its corresponding abbreviation, are then meticulously written to a new file. The robust error-handling mechanism implemented in the function ensures that any unforeseen issues are promptly addressed, providing users with informative messages for effective troubleshooting.

In summary, the main function acts as a user-friendly guide, facilitating a smooth and error-resilient execution of the program. Its ultimate objective is to enhance user experience by delivering clear feedback throughout the entire process.

# Testing

Regarding testing, various names have been input into the program to evaluate its performance. Notably, the input file includes names such as "Object-Oriented Programming," "magnetic resonance imaging," "urinary tract," "chronic obstructive pulmonary disease," and "body mass index." The anticipated output, based on the defined rules, should consist of three-letter abbreviations for each name, accompanied by their respective scores.

Upon testing, the program has demonstrated functionality close to the task rules and provided examples. For instance:

"Object-Oriented Programming" yielded the abbreviation "OOP."

"Magnetic Resonance Imaging" resulted in the abbreviation "MRI."

"Urinary Tract" generated the abbreviation "UT."

"Chronic Obstructive Pulmonary Disease" produced the abbreviation "COPD."

"Body Mass Index" led to the abbreviation "BMI."

These outputs align with the expected outcomes based on the prescribed rules. However, it's acknowledged that variations may occur depending on the interpretation of the rules and specific test cases. Further insights or adjustments can be made based on additional expectations or clarifications regarding the desired outcomes.