# Assignment Report: Feature Extraction from Text

Name :Muhammad Shaheer(20p-0480)

Instructor:Sir Shahzaib

## 1. Introduction

This report outlines the process and methodology used to complete the assignment on feature extraction from text, specifically calculating the Automated Readability Index (ARI) for given text files. The ARI score helps gauge the readability of text, which can be used for various applications, such as text authorship analysis.

## 2. Objective

The objective of this assignment was to implement a program that reads text files from a given directory, computes the number of characters, words, and sentences in each file, and calculates the ARI score based on these counts.

## 3. Methodology

To solve the assignment, the following steps were taken:  
1. **Reading the Text Files**: Python was used to read the content of the text files from the specified directory.  
2. **Counting Sentences**: A function was created to count the number of sentences using punctuation marks such as periods, question marks, and exclamation points.  
3. **Counting Words**: A function was developed to count words as sequences of alphanumeric characters separated by spaces or sentence terminators.  
4. **Counting Characters**: Another function was implemented to count only alphanumeric characters while excluding punctuation and spaces.  
5. **Calculating ARI**: The ARI score was computed using the provided formula:  
 ARI = (4.71 × (characters / words)) + (0.5 × (words / sentences)) - 21.43.  
 Non-integer scores were rounded up to the nearest whole number.

## 4. Implementation

The program was implemented in Python using functions to structure the code efficiently. Each of the steps mentioned in the methodology was encapsulated in separate functions to ensure modularity and readability. The final program reads the files, processes the content to count sentences, words, and characters, and calculates the ARI score for each file.

## 5. Example Results

The following are examples of the output generated by the program:  
- **Text File 1**: Number of sentences: 2, Number of words: 7, Number of characters: 23, ARI score: 0  
- **Text File 2**: Number of sentences: 3, Number of words: 31, Number of characters: 126, ARI score: 3  
- **Text File 3**: Number of sentences: 3, Number of words: 126, Number of characters: 499, ARI score: 19

## 6. Conclusion

The assignment was successfully completed by implementing a Python program that accurately reads, processes, and analyzes text files to compute the ARI score. This process demonstratesd the importance of text processing in machine learning and AI applications.