Feature Extraction from Text Report

MUHAMMAD SHAFEEN (P22-9278)

February 16, 2024

Précis

1 Introduction

The purpose of this program is to analyze text files from a given directory and compute the Readability Index for each file. The Readability Index is determined by counting the number of characters, words, and sentences in the text. Certain assumptions are made about what constitutes a character, word, and sentence for the ease of program implementation.

2 Assumptions and Rules for Counting

- Sentences are identified by occurrences of period ('.'), colon (:), semicolon (;), question mark (?), and exclamation mark (!).
- Each occurrence of the above sentence terminators is counted as a separate sentence.
- If a text has no sentence terminators, assume it has 1 sentence.
- A character is defined as any alpha-numeric character, excluding punctuation marks and spaces.
- Punctuation marks and spaces are not counted as characters.
- A word is defined as a sequence of one or more alpha-numeric characters delimited by white space or sentence terminators.

3 Program Implementation

The program reads text files from a specified directory and calculates the Readability Index for each file. The following steps are performed:

a. File Reading: The program reads each text file in the given directory.
b. Character Counting: Characters are counted, excluding punctuation marks and spaces.
c. Word Counting: Words are counted based on the defined rules.
d. Sentence Counting: Sentences are counted based on the specified sentence terminators.
e. Readability Index Calculation: The Readability Index is calculated using the formula:

4 Example Output

For each text file processed, the program provides the following information:

- Filename
- Number of characters.
- Number of words
- Number of sentences
- Readability Index

5 Usage

You will be given few options , you have to choose :

1. Read a file
2. Number of Sentences
3. Number of Words
4. Number of characters
5. Compue All of the tasks...
6. Exit

You can choose any of of the above it will compute for all the files provided with the code

6 Significance and Applications

Automated readability analysis has broad applications, ranging from educational materials assessment to content optimization for various audiences. This program provides a quick and automated way to evaluate the readability of textual content, aiding content creators and educators in producing material suitable for their target audience.

7 Conclusion

The program offers a valuable tool for assessing the readability of text files, providing insights into character, word, and sentence counts. While the rules and assumptions are heuristic and may not cover all linguistic nuances, they form a practical foundation for automated readability analysis. Future enhancements may involve incorporating more advanced natural language processing techniques to improve accuracy and broaden the scope of the program.

References

"GeeksForGeeks", "StackOverflow", "sparkbyexample", "RealPython".