Moogle: Advanced Document Search

Francisco Prestamo

2023

Overview

- Moogle can search within a group of .txt documents.
- It finds the documents most similar to a given query.
- Suggests alternatives if query words don't appear in the documents.
- Implements TF-IDF, Cosine Similarity, and Levenshtein Distance.

TF-IDF

- TF-IDF stands for Term Frequency-Inverse Document Frequency.
- It measures the importance of a word in a document within a corpus.
- The importance increases proportionally to the number of times a word appears in the document and is offset by the number of documents in the corpus that contain the word.

Cosine Similarity

- Cosine similarity is a measure of similarity between two non-zero vectors.
- In Moogle, it is used to determine the similarity between the query and documents.
- It considers the cosine of the angle between two vectors, giving a measure that is not affected by the magnitude of the vectors.

Levenshtein Distance

- Levenshtein Distance, or edit distance, is a string metric for measuring the difference between two sequences.
- Moogle utilizes this to suggest alternative queries if the original query words don't appear in the documents.
- This allows Moogle to handle typos and misspellings in the query.

Summary

- Moogle provides advanced search capabilities over a corpus of .txt documents.
- It uses TF-IDF, Cosine Similarity, and Levenshtein Distance to provide accurate results and helpful suggestions.

Thank You!

Questions?