Developing an Information Retrieval System with Advanced Boolean Search

About

This project aims to develop an Information Retrieval (IR) system that supports both standard Boolean queries and proximity queries. The system is designed to handle a collection of text documents, building an Inverted Index and a Positional Index to facilitate efficient document retrieval. You can find the data here.

Project Overview

In this project, you will build an Information Retrieval system from scratch using Python. Key components and functionalities are as follows:

- **Document Preprocessing:** Your project will begin by reading and preprocessing a collection of text documents.
- **Inverted Index:** You will create an Inverted Index, a vital data structure that maps terms to the documents in which they appear. This index is crucial for efficient document retrieval and supports standard Boolean queries.
- **Query Processing:** The system will handle two types of queries:
  - **Standard Boolean Queries:** Users can perform standard Boolean queries using the operators AND, OR, and NOT to retrieve relevant documents. For this project, queries are limited to two terms.
  - **Proximity Queries:** Users can also perform proximity queries, specifying a maximum distance between two terms in the documents they want to retrieve.
- **Index Optimization (Optional):** You have the flexibility to choose whether or not to implement index optimization. This feature can be included or excluded based on your project's goals. Index optimization may involve techniques to enhance query processing speed to improve search performance. For this project only, achieving successful index optimization will earn additional points.

Sample Documents

- Document 1:

```
This is a simple example document. It contains several words. The words should be processed and indexed.
```

- Document 2:

```
Another example document with different content. Document indexing is important for retrieval.
```

- Document 3:

```
Another example document to test Boolean search capabilities. This document contains relevant content.
```

Input Boolean Query

boolean_query = "example AND content"

Output (Expected Results)

expected_results = [2, 3]

Input proximity Query

proximity_query = "example NEAR/3 content"

Output (Expected Results)

expected_results_proximity = [2]

Deliverables

- A fully functional Information Retrieval system that effectively handles both Standard Boolean Queries and Proximity Queries.
- Comprehensive documentation describing the system's architecture, components, and any applied optimization strategies (if chosen to implement).
- A report summarizing key findings, challenges faced, and any enhancements made to the system during the project's development (including index optimization, if implemented).

References

Document and reference the sources, libraries, and tools used in the project. You can refer to the NLTK library for text processing and explore relevant literature on Information Retrieval techniques.