Developing an Information Retrieval System with Spelling Correction and Wildcard Queries

About

This project aims to enhance the Information Retrieval (IR) system developed in the first assignment by handling Spelling Correction and Wildcard Queries. This assignment can be completed independently of Project 1. You can find the data here.

Project Overview

In this project, you will implement spelling correction and wildcard queries using Python. Key components and functionalities are as follows:

- **Document Preprocessing:** Your project will begin by reading and preprocessing a collection of text documents. You only need to refer to the dataset as a word list.
- **Spelling Correction:** You will implement a function for isolated spelling correction. Your function needs to correct an input query using Levenshtein distance based on the words in the list. As the word list derived from the data is not complete, your function does not work flawlessly for all input queries.
- **Wildcard Queries:** You will implement a function that handles wildcard queries using Permuterm or K-gram (K=2) indexing. Your function needs to be able to support queries with one or two * symbols. To this end, you might need to post-filter false positive outcomes. You cannot use Regex.
- **Information Retrieval System (Optional):** You may enhance an Information Retrieval System that you might have developed in Project 1 using the implemented functions for spelling correction and wildcard queries.

Sample Documents

- Document 1:

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

- Document 2:

```
Another example document with different content. Spelling correction is important for retrieval.
```

- Document 3:

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

Input – Spell Checking

query = "festivsl funders"

Output – Spell Checking (Expected Results)

expected_results = "festival founders"

Input – Wildcard Queries

wildcard_query = "n*b*y"

Output (Expected Results)

expected_results = ["nobody"]

Input – Information Retrieval System

enhanced_query = "exa*le AND contrnt"

Output (Expected Results)

expected_results = [2, 3]

Deliverables

- Two functions for handling spelling corrections and wildcard queries based on the dataset.
- Comprehensive documentation describing the functions' architecture and components.
- A report summarizing key findings, challenges faced, and any enhancements made to the IR system during the project's development (you may append to your first report).

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.