Assignment 5: TF-IDF Representation

phuonglh@gmail.com

April 6, 2024

1 Problem

In this assignment, you will implement a ranked retrieval method which uses the tf–idf representation and the cosine proximity to sort the relevance of a document with respect to a a query. This is a continued work of the previous assignments. Now your search application should has two options for searching: boolean (exact) search and ranked search. Suppose that a user selects the ranked search option.

Given an input query, your program should:

1. Compute the tf–idf vector of the query and thoses of the documents in your indexed collection;
2. Compute the proximity scores using the cosine values;
3. Sort and present results to the user.

2 Guide

See the lecture slides for the tf–idf and cosine proximity formula. Think about how to compute vectors and values efficiently. Sparse vectors? Caching? You are free to choose the GUI components and layout so that your interface is convenient for users.

3 Submission

If you develop your program in Scala, Python, or Java, submit your source file *YourFullName.scala*, *YourFullName.py*, or *YourFullName.java*, respectively; for example *NguyenPhuTrong.scala*. Your data file has the name *NguyenPhuTrong.json*. If you have multiple source files then put them in a directory and zip it to *YourFullName.zip* and submit before due date.