Search and Data Mining (Spring 2020)

Assignment #2: Retrieval Models Competition

(Individual assignment, due on April 6, 11:59:59pm)

Introduction

The goal of this assignment is to implement basic traditional retrieval algorithms and experiment with them on multiple data collections. You will be using our newly developed system, i.e., VIRLab, for this assignment.

The system for this class is available at: https://virlab.eecis.udel.edu/UDEL (The URL is case sensitive). You need to register for the system first. You could do that by clicking the *Register* button on the home page and then fill out the registration form. After you submit your registration form, please allow 24 hours for the TA to approve your registration. If you can not login the system after the 24 hours time period, please contact the TA.

The assignment has two parts: (1) implementation of two basic retrieval functions (i.e., Okapi and Pivoted) and evaluation on one data collection; and (2) search engine competitions.

Part I. Comparing and Implementing Two Basic Retrieval Models [40 points]

1. [5 points] Compare the TF-IDF pivoted normalization formula and Okapi formula analytically.

tf	is the term's frequency in document
qtf	is the term's frequency in query
N	is the total number of documents in the collection
df	is the number of documents that contain the term
dl	is the document length (in bytes), and
avdl	is the average document length

Okapi weighting based document score:

$$\sum_{t \in Q,D} ln\frac{N - df + 0.5}{df + 0.5} \cdot \frac{(k_1 + 1)tf}{k_1((1-b) + b\frac{dl}{avdl}) + tf} \cdot \frac{(k_3 + 1)qtf}{k_3 + qtf}$$

k_1 (between 1.0–2.0), b (usually 0.75), and k_3 (between 0–1000) are constants.

Pivoted normalization weighting based document score:

$$\sum_{t \in Q,D} \frac{1 + ln(1 + ln(tf))}{(1-s) + s\frac{dl}{avdl}} \cdot qtf \cdot ln\frac{N+1}{df}$$

s is a constant (usually 0.20).

Both formulas are given in the figure above. What are the common statistical information about

documents and queries that they both use? How are the two formulas similar to each other, and how are they different?

2. [20 points] Implement the TF-IDF pivoted normalization method and Okapi retrieval formula above. For the Okapi formula, set `k3=1000` and `k1=1.2` (you can hard code it, if you want), so that you have only one parameter `b` to vary. In this way, both algorithms have precisely one parameter to tune.

   Please read the tutorial provided by the VIRLab system, and learn about how to implement the retrieval functions.

3. [15 points] Test both algorithms with the provided collections. You should be able to do it by clicking on the "Manage functions" after you have saved your functions.
   - Change the values of the parameter in each algorithm (between 0 and 1), and evaluate the performance. Plot how the MAP changes as the parameter value changes. Compare the two algorithms' performance and behavior. Is the performance sensitive to the setting of the parameter in each of them? Do they perform similarly or differently? When optimally tuned, which method gives the best performance?
   - Do you have different observations for different collections?

Part II. Retrieval Model Competition [60 points]

We set up a retrieval model competition for each collection, and the top-performed systems will be displayed in the "Leaderboard". You are encouraged to try different retrieval functions either from the provided papers or **proposed by yourself**.

You are expected to implement at least THREE different retrieval functions other than the Okapi and Pivoted. The top 3 ranked retrieval functions for each collection would receive bonus points. Moreover, extra bonus points will be given to the well-performed retrieval functions that are proposed by the students. Note that after implementing the retrieval functions, you need to evaluate them through "manage functions" to make sure that the functions will be considered for ranking on the leaderboards.

- Exploring the similarity space (SIGIR forum)
- Word Document Density and Relevance Scoring (SIGIR 2000)
- A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval (SIGIR2001)
- An Exploration of Axiomatic Approaches to Information Retrieval (SIGIR 2005)
- A Study of Poisson Query Generation Model for Information Retrieval (SIGIR 2007)
- Generalized Inverse Document Frequency (CIKM 2008)
- Lower-Bounding Term Frequency Normalization (CIKM 2011)

What to turn in

- Please submit the following files on canvas.
  - Your answers to Part I (1), (2) and (3).
  - Retrieval functions implemented in Part II.
    - For each function, please describe the name used in the system, your implemented code as well as their mathematic formulas.
    - If you proposed a function, please describe how they are derived and the

motivation behind them.

- For each function and each collection, please report the performance in a table.

~~More information about the VIRLab server:~~

~~All the students can use the VIRLab server for your assignments anytime between now and March 21. Starting from March 22, we will limit the access to the server to avoid the server overload. More specifically, between March 22 and March 27, only a subset of students can access the server every day as follows:~~

- ~~March 22 and March 23: Only students whose (StudentID mod 3) ==0 can access the server.~~
- ~~March 24 and March 25: Only students whose (StudentID mod 3) ==1 can access the server.~~
- ~~March 26 and March 27: Only students whose (StudentID mod 3) ==2 can access the server.~~

The limited access period is moved to :

- **April 1 and April 2: Only students whose (StudentID mod 3) ==0 can access the server.**
- **April 3 and April 4: Only students whose (StudentID mod 3) ==1 can access the server.**
- **April 5 and April 6: Only students whose (StudentID mod 3) ==2 can access the server.**

Please note that during the limited access period, when you could access the server, instead of using your original password, the password you should be using is CPEG657.

Moreover, we will give bonus points (5 points) to the first ten students who finish the homework early (based on the submission time on Canvas). Again, the main goal is to reduce the server load so that the server can be up running for all of you to finish the assignment.