F033583 Introduction to Web Search & Mining

Group Project: Building A Search System

Final Report, Code and Demo Due: **Saturday, June. 23, 2017**

Specifications

In this project, you have two options in building a search system. The work includes crawling pages/media files, building the dataset, indexing data and creating a nice web interface for search.

OPTION A

In this project, you are asked to crawl at least 2-year worth of pages (2014-2017) from Xunyiwenyao (http://www.xywy.com/). Two types of pages, namely doctor's homepage and Q&A pages will be needed in this project. In a doctor homepage, you can get all the information about doctor himself/herself, the questions he/she has answered before and other similar doctors' link. In a Q&A page, there's usually one question and several replies from either doctors or other patients. Questions are classified in different category and similar questions are also linked to.

Examples of the pages are as follows:

Doctor homepage: http://club.xywy.com/doc_card/5866660

Q&A page: http://club.xywy.com/static/20170307/127336866.htm

You are required to build index and/or other data structures to support three kinds of queries.

1. User asks a question about a disease or symptom like those in the Q&A page, and the system should return a ranked list of all similar questions along with their answers.
2. User asks a keyword which may be a disease, name of the doctor or any key words that might appear in the page, and the system returns a ranked list of any pages you have indexed.
3. Advanced search: let user enter search keyword for a particular region in the page, e.g., search in the title, the question, the answer, the doctor's profile or a particular medical specialization, e.g. internal medicine or traditional Chinese medicine. The advance search should give the interface to allow the user to choose what region to search from. Appropriate regional/zonal indexes need to be build.
4. User asks for a disease or symptom, you are required to recommend a ranked list of doctors who can best treat this disease or symptom. You may consider what the doctor is good at or the reputation or the number of fans and so on.

OPTION B

In this project, you are asked to crawl a set of songs (no less than 10,000) including titles, names of the singer and songwriter, lyrics and the music mp3 files (average quality will do). Xiami(http://www.xiami.com/), Wangyiyun(http://music.163.com/) and other websites may be useful.

You are required to build indexes and/or other data structures to support three kinds of queries. You will need to take advantage of the search log to complete this project.

1. Free text queries, like search one song's title and return a list of ranked songs (There may be songs with duplicate titles), search a singer and return a list of songs according to the popularity or year posted or other properties. Also we may search several keywords hoping to get a list of songs whose titles/lyrics contain those words.
2. Record the user's information about clicking and music playing, choose or design the algorithm so that it will adjust the ranked list. For example, I searched a song named "love" and always click on the song by Jiaying Xu. The system should adjust the ranking next time I search the same keyword "love".
3. (Bonus) User sings or hums part of a song or tune into the mic, or upload a short audio clip containing part of a song, and the system returns a list of relevant songs. A similar service can be found here: http://www.midomi.com/

Deliverables

The final deliverables should include the following items:

- A well-written report to describe your ideas, design, implementation, example queries and results (with screenshots), conclusion, etc.
- A web demo deployed on any publically accessible web server (in case you can't find an accessible machine to host your code and data, you can deploy the server on your local computer and contact Yuchen for a personal demo in her office, before the due date).
- Source code of the whole search system
- Zipped archive of the entire crawled data

Each group submit all of the above electronically to Yuchen @ yukisha93@126.com. In addition, every member should send a confidential peer review form to Yuchen by the due date as well. The peer review form will be released on the course website for download.

Scoring Criteria

Your final score will consist of **six** parts.

Completeness of crawled pages:	20%
Precision/recall of keyword retrieval:	20%
Quality of ranking of the pages:	20%

Additional search features:	20%
Search system GUI and usability:	10%
Peer review:	20%

Each group member will receive the same score for the first 5 part of the scores except for the peer review. Note the total scores add up to more than 100% to give bonus for the extra features.