F033583 Introduction to Web Search & Mining
Group Project: Building A Search System
Final Report, Code and Demo Due: **June 14, 2019**

Introduction

This is a group-based research project. Each group should contain 2-4 students. In this project, there are three options in building a search system. Each group must email TA (wsmproject@163.com) your choice. There should be two choices in the email (First choice and Second choice). If a project idea is chosen by too many groups (each option can be selected about 1/3 groups), it will be allocated on a first-come, first-serve basis.

Specifications

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 (2016-2019) worth of questions and their corresponding answers from Stack Overflow (https://stackoverflow.com/). In a Q&A page, there's usually one question and several replies from other users. Each reply may have some comments. Questions are labelled in different tags and similar questions are also linked to.

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

1. User asks a question about a technique or model like those in the Stack Overflow, and the system should return a ranked list of all linked questions with their answers and a ranked list of related questions with their answers. The answers should contain their comments.
2. User asks a keyword which may be a machine learning framework, programming language 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 tag, the question or the answer. 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. As each question in Stack Overflow could have different answers, you are required to recommend a ranked list of answers. You may consider whether the answer is accepted by the question owner, the number of times it is considered useful or the number of its comments.

OPTION B

In this project, you are asked to crawl Baidu Baike (no less than 10,000 entries) and Baidu Zhidao (no less than 10,000 questions). The number of entries or questions under each category in Baidu Baike or Baidu Zhidao should be similar. For Baidu Baike, each entry includes title, subtitles and corresponding descriptions. For Baidu Zhidao, each question has different replies. There are always some images in Baidu Baike or Baidu Zhidao.

Data Source:
Baidu Baike https://baike.baidu.com/
Baidu Zhidao https://zhidao.baidu.com/

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

1. User asks a question like those in the Baidu Zhidao, and the system should return a ranked list of all similar questions along with their answers. If there are any Baidu Baike entries or images related to the question, the system also should return ranked lists of these entries and images.
2. User asks a keyword which may be any key words that might appear in the page of Baidu Baike or Baidu Zhidao, and the system returns a ranked list of related entries and a ranked list of all related questions along with their answers. If there are any related images, the system also should return a ranked list of these images.
3. Advanced search: let user enter search keyword for a particular region in the page, e.g., search in the Baidu Baike, the question of Baidu Zhidao, the answer of Baidu Zhidao or the images. 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. As each question in Baidu Zhidao could have different answers, you are required to delete unrelated answers and recommend a ranked list of answers. You may consider whether the answer is accepted by the question owner or the number of times the answer is considered useful.

OPTION C

In this project, all your data should be crawled from the following websites. Note that for each website we give, you need to follow 100 bloggers about media, e.g. @中国日报 (Weibo) or @BBCNews (Twitter), and crawl all of their blogs with comments starting from year 2016 (including 2016) up to now. You need to extract all the hashtag in the blogs and the complete dialogues in the comments.

Data Source (Choose one):
Xinlang Weibo (Chinese) https://www.weibo.com
Twitter (English) https://twitter.com

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

1. User asks a keyword which may be name, event or any key words that might appear in the page of Weibo or Twitter, and the system returns ranked lists of related blogs with comments and all the dialogues in these comments. The system also should return a ranked list of related hashtag.
2. Advanced search: let user enter search keyword for a particular region in the page, e.g., search in the blogs, the hashtag, the comments or the dialogues. 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.
3. User asks for an event, you are required to recommend a ranked list of bloggers who can best report this event. You may consider the number of followers of this blogger, the average number of comments and so on.

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 TA 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 wsmproject@163.com.
In addition, every member should send a confidential peer review form to wsmproject@163.com 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:	10%

Each group member will receive the same score for the first 5 part of the scores except for the peer review.