INFO300 Term Project – Due November 30, 2022

This is the final project for INFO300, Information Retrieval Systems. The goal of the project is to let you apply the IR theories and methods that you learn in this class to a practical IR issue through implementing and experimenting some features of Elasticsearch.

You may form a team of 3 to 4 people to work together in this project, following these steps:

Step 1. Prepare a test collection for the experiment

Identify a specific area of interest and use a Web scraping tool such as Scrapy to collect a group of webpages. You may choose a specific topic such as "Colleges in the Philadelphia area," "Data Science Master's programs in U.S and China," "Data science Jobs & Careers", or some sports or other topics of your interest. The collection you collect should have at least 200 items and generally no more than 2000 items. Check and clean up the output from Scrapy to make sure that the format and data are indexable by ElasticSearch.

Step 2. Implement a baseline IR system – System 1

1) Index the test collection using the default setting of ElasticSearch
2) Learn how to use the ElasticSearch Python API to create a search interface
3) Create a simple web page to allow searching for the test collection
4) Test and make sure the interface and search work correctly.

Step 3. Implement new IR Features to enhance the base system – System 2 & 3

There are many ways to enhance IR systems and evaluate their performances. In ElasticSearch, for example, you may

1) configure different text analysis & indexing methods
2) use a different similarity scoring
3) apply different ranking algorithms

Study and choose two different ways to reindex the test collection (through reconfigure the mapping) to create **System 2** and **System 3.** Your system 2 & system 3 can focus on one area (such as two different ways of calculating similarities or two different rankings, etc.) You can also implement other enhancement not listed in the above.

Step 4. Evaluate and compare the performances of the three systems

1) Run at least 6 different queries and make relevant judgments for the top 10 hits
2) compare their results across 3 systems
3) Calculate MAP for each of the queries on each system
4) Discuss and report the comparison results.

Final product: Your team should submit a final report and video that include:

1) A video of 5 minutes or less showing how your systems work
2) how the team worked together (who works on which parts, etc.)
3) codes and results for every steps above
4) What you have learned from the comparison results and from this project.

A team needs to submit only one final report.