Progress Report - Chicken and Rice

Group Members:

Nanyi Yang (nanyiry2)

Ivan Zhong (ninghan2)

Cody Wang (yaohuiw2)

1) Which tasks have been completed?

Currently we have completed a barebones web scraper similar to the one in MP2.1. However, this scraper currently does not utilize Selenium, and so it is not very robust and is not able to scrape websites that render their content using Javascript.

As for our search engine, we have implemented a general structure using the BM25 model from metapy, similar to MP2.2. However, our current implementation could not load any corpus data as we need to use the corpus that we collected from our scrapper. We will incorporate the BM25 model with our collected corpus from our database in future steps.

2) Which tasks are pending?

Some next steps for the web scraper will be to figure out how to use Selenium with a headless browser to simulate a user. We will also potentially need to figure out how to store scraped data in some sort of database.

For the tasks on the search engine, we are currently choosing what ranker to apply and how to train the ranker. We are thinking about utilizing metapy.index.OkapiBM25(k1, b, k3) but we are still confused on how to optimize the coefficient that needs to be inserted in. One way of optimization that we've thought about is to apply the algorithm of Support Vector Machine which takes each pair of (k1, b, k3) as input vector and outputs true/false as return value for us to adjust it.

3) Are you facing any challenges?

On the web scraper, we are running into a few issues with the way the content of the page is delivered. Because we want the web scraper to be robust, we will probably need to use Selenium and a headless browser to read text data that is rendered with Javascript.

Besides the challenge above, we are also worried about the actual implementation steps of the optimization algorithm that we thought of. We should find a line/hyperplane to do the classification steps, but we recently have no ideas on how to set it.