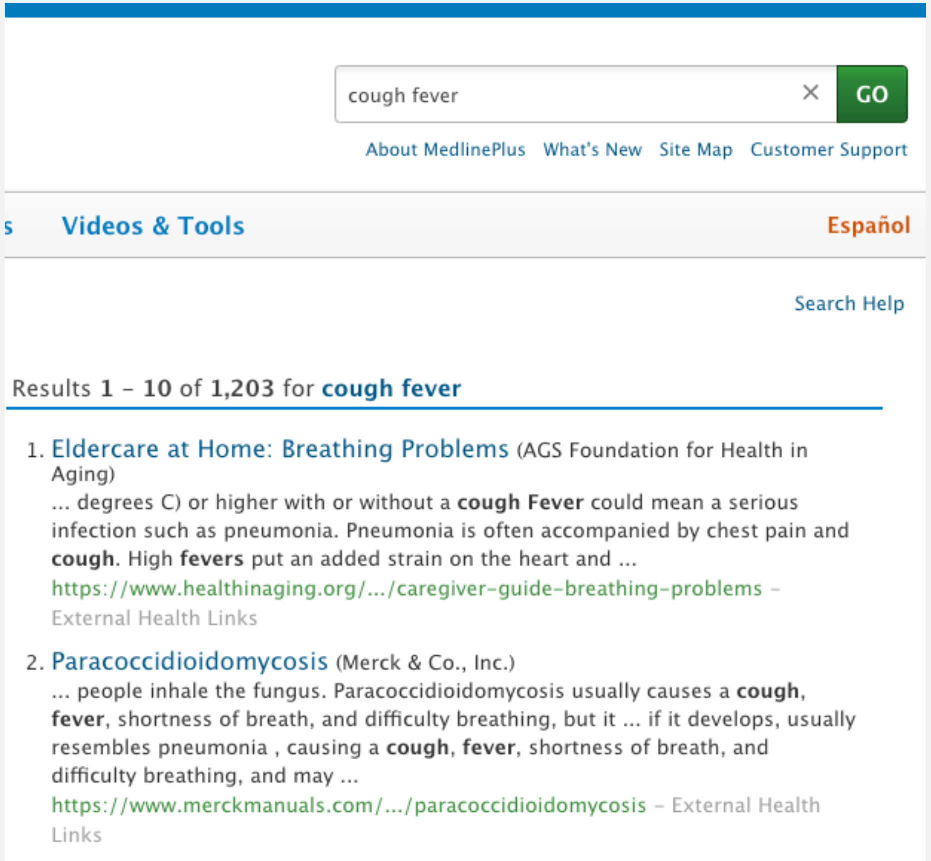
DISEASE-

A simple search engine, with a simple web crawler that focus on the disease

Rujun Yao

PROBLEMS

- Current pandemic of COVID-19
- It is hard to search for symptoms from search engine
- Low efficiency of Hospital
- "Shelter In Place" Order



Results **1 – 10** of **1,203** for **cough fever**

1. Eldercare at Home: Breathing Problems (AGS Foundation for Health in Aging)
   ... degrees C) or higher with or without a **cough Fever** could mean a serious infection such as pneumonia. Pneumonia is often accompanied by chest pain and **cough**. High **fevers** put an added strain on the heart and ...
   https://www.healthinaging.org/.../caregiver-guide-breathing-problems – External Health Links
2. Paracoccidioidomycosis (Merck & Co., Inc.)
   ... people inhale the fungus. Paracoccidioidomycosis usually causes a **cough**, **fever**, shortness of breath, and difficulty breathing, but it ... if it develops, usually resembles pneumonia , causing a **cough**, **fever**, shortness of breath, and difficulty breathing, and may ...
   https://www.merckmanuals.com/.../paracoccidioidomycosis – External Health Links

SOLUTION

- Web crawler to crawl a medical website (depth = 3)
- Html parser to parse links and text (a tag href attr)
- Store texts locally (offline)
- Indexed locally stored documents (inverted index term:docids)
- Query keywords
- Tf-idf cosine score rank system

SEED

- https://www.cdc.gov/
  - Many unrelated information
  - Pages in other language
  - Disease page, symptoms page are separated
- https://www.who.int/
  - Same problem as cdc
- https://medlineplus.gov/
  - Information are gathered together (related info)

CRAWLER

- Initialization: visited_set(), front_queue(), seed
- For each link in the queue:
  - Check if the link is valid. (external links? Spanish?)
  - Extract links, add all valid link into front_queue
  - Extract metadata and texts, save texts as local docs

INDEXER

- Remove pos of term in inverted index
- Form: {term1: {docID, docID, …}, term2: {docID, docID,…}, …}
- Other information: vector length, term frequency, document frequency and inverse document frequency, stop-word list
- Since there are up to 10,000 documents, The indexing process will take lot of time and memory, all of the information can be cached in disk storage.
- Use pickle.dump() and pickle.load() to save or load dictionary data structure in disk as binary file.

QUERY

- Cosine similarity and tf-idf
- Could load index from file system
- Use subprocess to open the retrieved documents.