MediCrawl - A Web Search Engine For Diseases

Devharsh Trivedi,[1] Vaishnavi Gopalakrishnan[2] and Wendy Hui Wang[3]

***Abstract*— We present in this paper MediCrawl - A web search engine made with open source tools for searching diseases. In this project we target both the common and the rare diseases and try to beat the performance of popular search engines like FindZebra and MayoClinic.**

I. INTRODUCTION

A. Web Search Engine

Internet is our everyday source for all kinds of information, like news, broadcast, music, pictures, or even movies. We are immersed in the flood of big data every day. However, the information we need is only a small portion of it and for different individuals the importance of information is hardly the same. A search engine will make it convenient for us to absorb information efficiently.

B. Crawling and Indexing

Search Engines have automated robots called crawlers that use links to scour the Internet, find web pages, and decipher page data that are indexed to be included in search engine results.

C. Ranking

Each search engine has different ranking factors, but they all have a few factors in common: keywords, content, and links. Keywords and content are arguably two of the most important factors that search engines look for when ranking pages. Because of this, it is imperative that you know what keywords are in highest demand within your market and incorporate those keywords into the content on your website. All your website's content will naturally create a collection of links; search engines use link analysis algorithms that look at the sources, number, and anchor texts of links to help determine their relevance in search queries. TF-IDF is one such method for ranking.

II. SCOPE OF WORK

A. Problem

Design a web search engine (Medi Crawl) for diseases.[2]

B. Vision

We are going to use open-source libraries to build our web crawler to download the data from the internet, and we are using No SQL database to store the data, to perform Map Reduce (if required) on the database, and we are going to build a simple web interface as front-end layer for users to search text using Node & Express JS.

C. Accessibility and Indexing

For your website to be visible in search results, search engines must be able to find it. They do this by crawling the web and looking for relevant and index-able content, such as link structures and HTML features like alt tags for images.

D. Ranking Factors

The different factors that we will be looking at are; 1) contents, to make sure that your content is relevant and utilizes keywords without being duplicated on multiple pages, 2) keywords, and 3) HTML markup, to make sure that your source code is relevant to crawlers.

E. Data Management

The data management plan to get the data for our prototype. We are planning to implement a search engine by building our own indexed database to collect web pages via Web crawling.

F. Query Implementation

- Query should be returned and displayed within 2s.
- Relevance (confidence of results) should be more than 75%.

III. RESOURCES

A. Prerequisites

- Python: https://www.anaconda.com/distribution/#download-section
- Git: https://git-scm.com/
- Git-LFS: https://com.puter.tips/2020/04/install-git-lfs-in-macos.html
- Solr: https://lucene.apache.org/solr/downloads.html
- Nutch: http://nutch.apache.org/downloads.html

TABLE I: Components of web search engine

Functionality	Module Type	Software
Crawling	Web Crawler	Nutch
Indexing	Database	Solr
Ranking	Database	Solr
Searching	User Interface	Node JS

[1]Devharsh: He is a Ph.D. student in the Department of Computer Science, Stevens Institute of Technology, NJ Email: dtrived5@stevens.edu

[2]Vaishnavi: She is a Masters student in the Department of Computer Science, Stevens Institute of Technology, NJ Email: vgopalak@stevens.edu

[3]Wendy: She is an Associate Professor in the Department of Computer Science, Stevens Institute of Technology, NJ Email: hwang4@stevens.edu

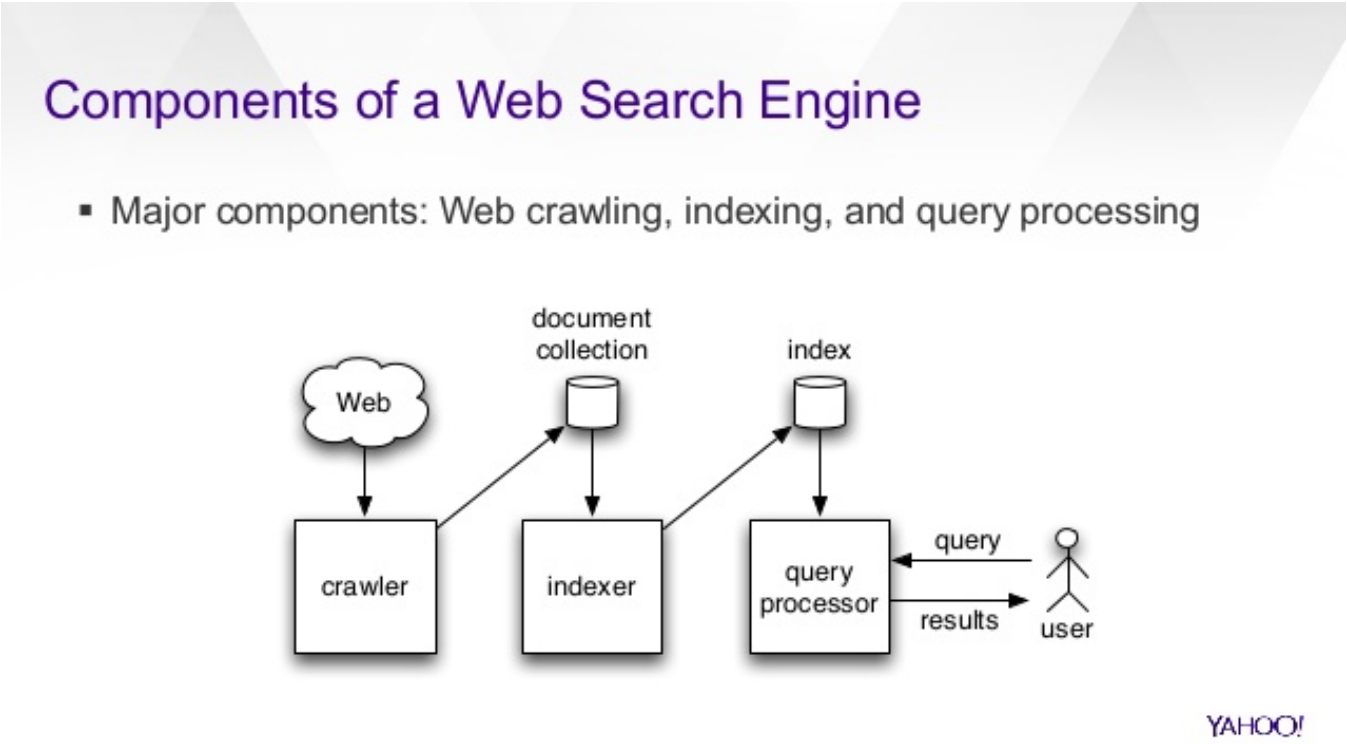

Fig. 1: Components of a Web Search Engine[1]

IV. Role of members

Devharsh is leading this project and he will be working on back-end to build Web Crawler for the engine and indexing.

- Design you own Web ”spider” to go through several levels deep starting from a specified page.
- Results displayed should be ranked.

Vaishnavi will be working on MongoDB and Node JS to store and process the data.

- A simple query interface should provide the capability to perform simple Boolean queries.
- Retrieved documents should be automatically indexed (inverted file format) and stored on a server.

In terms of other work like graphical web interface, improving user experience and efficiency or some unpredictable small problems to fix, we will be working together.

TABLE II: Project Schedule

Module	Planned	Progress	Member
Planning & Design	March 15	Complete	Devharsh
Implement Crawling	April 5	Complete	Devharsh
Implement Indexing	April 10	Complete	Vaishnavi
Implement Ranking	April 20	Work in progress	Devharsh
Implement Searching	May 4	Work in progress	Vaishnavi
Testing & Improvements	May 8	On track	Both
Documentation	May 12	On track	Both

V. Implementation

We have implemented a web crawler using Nutch for medical diseases and indexed them in Solr. Remaining work is to increase the database of indexed pages and improve the accuracy (MAP, MRR) of results. Also a clean and easy to use UI needs to be implemented as mentioned in the Project Schedule above.

A. Crawling - Completed

a) Apache Nutch: [3]

- Nutch is an open source search engine implemented in Java.
- Nutch implements “Map Reduce” distributed processing model.
- Nutch installations typically operate at one of three scales: local file system, intranet, or whole web.
- Nutch is built on top of Lucene, which is an API for text indexing and searching.

b) Apache Lucene: [4]

- NOT a crawler
- NOT an application
- NOT a library for doing Google pageRank
- An open source Java-based IR library enabling text based search

c) Nutch v/s Lucene: [3]

- Use Lucene if a web crawler is not needed.
- A common scenario is that you have a web front end to a database that you want to make search-able. The best way to do this is to index the data directly from the database using the Lucene API, and then write code to do searches against the index, again using Lucene.

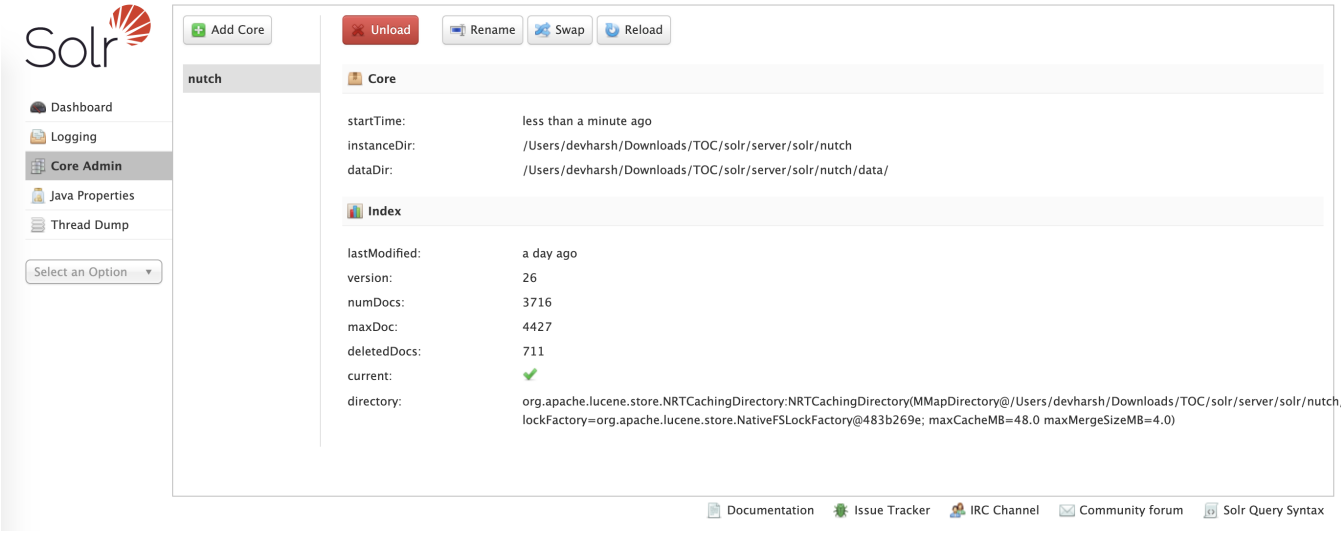

Fig. 2: Web pages crawled using Nutch and Indexed in Solr

- Nutch is a better fit for sites where you don't have direct access to the underlying data, or it comes from disparate sources.

d) Nutch Architecture: [3]

- Nutch is divided into two pieces: the crawler and the searcher.
- The crawler fetches pages and turns them into an inverted index.
- This inverted index is used by the searcher to resolve user's queries.
- Searcher and crawler components can be scaled independently of each other. The crawler system is driven by:
  - Nutch crawl tool.
  - Several types of data structures, including the web database, a set of segments, and the index.
- The web database, or WebDB stores two types of entities: pages and links.
- A page represents a page on the Web, and is indexed by its URL and the MD5 hash of its contents.
- A link represents a link from one web page (the source) to another (the target). A segment is a collection of pages fetched and indexed by the crawler in a single run.
- The fetch-list for a segment is a list of URLs for the crawler to fetch, and is generated from the WebDB.
- The fetcher output is the data retrieved from the pages in the fetch-list. The fetcher output for the segment is indexed and the index is stored in the segment.
- The index is the inverted index of all of the pages the system has retrieved, and is created by merging all of the individual segment indexes.
- Nutch uses Lucene for its indexing.

e) Steps performed by a Crawler: [3]

1) Create a new WebDB.
2) Inject root URLs into the WebDB.
3) Generate a fetch-list from the WebDB in a new segment.
4) Fetch content from URLs in the fetch-list.
5) Update the WebDB with links from fetched pages.
6) Repeat steps 3-5 until the required depth is reached.
7) Update segments with scores and links from the WebDB.
8) Index the fetched pages.
9) Eliminate duplicate content (and duplicate URLs) from the indexes.
10) Merge the indexes into a single index for searching.

B. Indexing - Completed

We are using Solr along with Nutch to index the crawled web pages. Solr is the third most popular search engine after ElasticSearch and Splunk.[5]

a) Apache Solr: [4]

- An open source enterprise search server
- Based on the Lucene Java search library
- A web based application that processes HTTP request and returns HTTP responses
- Completed with XML/HTTP APIs, caching, replication, and web administration interface.

b) Solr as NoSQL: [4]

- NoSQL : not only SQL
- Characteristics:
  - Non-traditional data stores
  - Not designed for SQL type query
  - Document oriented, data format agnostic (JSON, XML, CSV, binary)
- Versioning and optimistic locking with Real Time GET, allows read/write/update with/without conflict.
- Atomic updates can add/remove/change and increment a field in existing index with/without re-indexing

C. Ranking - Work in Progress

In information retrieval systems, Learning to Rank is used to re-rank the top N retrieved documents using trained machine learning models. The hope is that such sophisticated models can make more nuanced ranking decisions than standard ranking functions like TF-IDF or BM25. We are currently using LinearModel which is a scoring model that computes scores using a dot product.[6] We are planning to experiment with popular ranking algorithms like TF-IDF, BM25 and Neural Network Model to determine which works best for our case.

```
{
  "class" : "org.apache.solr.ltr.model
  .LinearModel",
  "name" : "myModel",
  "features" : [
    { "name" : "contentLength" },
    { "name" : "titleLength" },
    { "name" : "originalScore" }
  ],
  "params" : {
    "weights" : {
      "contentLength" : 1.0,
      "titleLength" : 0.1,
      "originalScore" : 0.5
    }
  }
}
```

D. Searching - Work in Progress

Planning to develop a web based user interface in Node JS.

a) Searching in Solr: [4]

- The search query is processed by a Request Handler:
  - Request Handler calls a query parser
  - Query parser interprets query's term & parameters
  - Input to a query parser can include:
    - Search strings – common terms
    - Parameters for fine tuning, eg. Boolean logic
    - Parameters for controlling the presentation of the query response, eg. Specifying the order in which results are diplayed.
- Searching in Solr can be done by sending HTTP Get or Post requests e.g. http://localhost:8983/solr/select?q=covid-19
- Solr provides a simple method to sort on 1 or more indexed fields. Use the sort parameter e.g. . . . ?q=lcd&sort=price asc

E. Procedure

Perform the following steps to setup the web search engine from scratch. Make sure to install prerequisites software.

1) Download and extract Nutch and Solr in nutch and solr directories
2) Modify nutch/conf/nutch-site.xml: set http.agent.name and indexr-solr property
3) Modify nutch/conf/schema.xml
4) Create a URL seed list: nutch/urls/seed.txt
5) cp -r solr/server/solr/configsets/_default solr/server/solr/configsets/nutch
6) cp nutch/conf/schema.xml solr/server/solr/configsets/nutch/conf
7) rm solr/server/solr/configsets/nutch/conf/managed-schema
8) Modify solr/server/solr/configsets/nutch/conf/solrconfig .xml to enable LTR and change _text_ to text
9) start solr with LTR: solr/bin/solr start -Dsolr.ltr.enabled=true
10) set JAVA_HOME environment variable
11) Crawl and Index simultaneously: nutch/bin/crawl -i -D solr.server.url=http://localhost:8983/solr/nutch -s nutch/urls/ Crawl 3
12) Test if it succeeded: curl 'http://localhost:8983/solr/nutch/query?q=breast+cancer &fl=url,title'
13) Create /path/features.json and /path/model.json
14) Upload features to Solr: curl -XPUT 'http://localhost:8983/solr/nutch/schema/feature-store' –data-binary "@./path/features.json" -H 'Content-type:application/json'
15) View uploaded features: curl 'http://localhost:8983/solr/nutch/schema/feature-store/ _DEFAULT_'
16) Test feature extraction: curl 'http://localhost:8983/solr/nutch/query?q=test&fl=id, score,[features]'
17) Upload model to Solr: curl -XPUT 'http://localhost:8983/solr/nutch/schema/model-store' –data-binary "@./path/model.json" -H 'Content-type:application/json'
18) View uploaded model: curl 'http://localhost:8983/solr/nutch/schema/model-store'
19) Test model: curl -g 'http://localhost:8983/solr/nutch/query?q=breast+cancer &rq={!ltr%20model=myModel%20efi.query=breast+ cancer}&fl=url,title,[features]'

F. Sample Responses

Following are the results as JSON responses you will get based on the query performed.

a) A Basic Query for "Breast Cancer":

```
{
  "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
      "q":"breast cancer",
      "fl":"url,title"}},
    "response":{"numFound":1122,"start":0,
    "docs":[
      {
        "title":"Breast cancer - Symptoms and
```

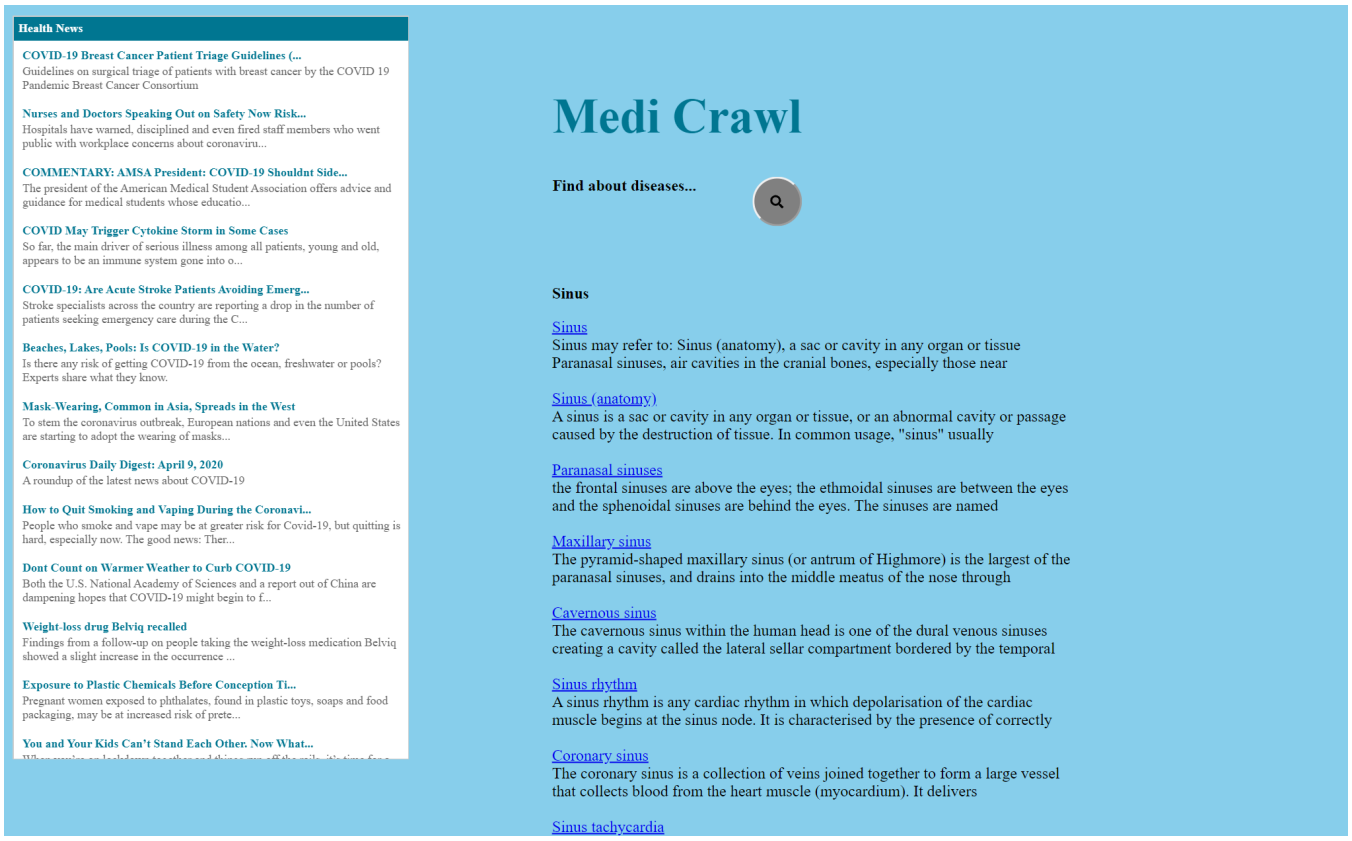

Fig. 3: Web User Interface


```
  causes - Mayo Clinic",
  "url":"https://www.mayoclinic.org/
  diseases-conditions/breast-cancer/
  symptoms-causes/syc-20352470"},
{
  "title":"Breast cancer - Diagnosis and
  treatment - Mayo Clinic",
  "url":"https://www.mayoclinic.org/
  diseases-conditions/breast-cancer/
  diagnosis-treatment/drc-20352475"},
{
  "title":"Breast cancer - Care at Mayo
  Clinic - Mayo Clinic",
  "url":"https://www.mayoclinic.org/
  diseases-conditions/breast-cancer/
  care-at-mayo-clinic/mac-20352479"},
{
  "title":"Breast cancer: Symptoms,
  causes, and treatment",
  "url":"https://www.medicalnewstoday.com/
  articles/37136"},
{
  "title":"Breast Clinic - Overview -
  Mayo Clinic",
  "url":"https://www.mayoclinic.org/
  departments-centers/breast-clinic/
  sections/overview/ovc-20459469"},
{
  "title":"Bring Your Brave.",
  "url":"https://bringyourbrave.tumblr.
  com/"},
{
  "title":"Breast Cancer News from
  Medical News Today",
  "url":"https://www.medicalnewstoday.com/
  categories/breast-cancer"},
{
  "title":"Breast Cancer | CDC",
  "url":"https://www.cdc.gov/cancer/
  breast/"},
{
  "title":"Breast cancer - Doctors and
  departments - Mayo Clinic",
  "url":"https://www.mayoclinic.org/
  diseases-conditions/breast-cancer/
  doctors-departments/ddc-20352478"},
{
  "title":"Breast Cancer | Disease of
  the Week | CDC",
  "url":"https://www.cdc.gov/dotw/
  breastcancer/index.html"}]
```


b) Query with Features score for "Breast Cancer":


```
{
```

```
"responseHeader":{
  "status":0,
  "QTime":8,
  "params":{
    "q":"breast cancer",
    "fl":"url,title,[features]",
    "rq":"{!ltr model=myModel efi.query=
    breast cancer}"}},
  "response":{"numFound":1122,"start":0,
  "docs":[
    {
      "title":"Science Clips - Volume 12,
      Issue 10, March 23, 2020",
      "url":"https://www.cdc.gov/library/
      sciclips/issues/index.html",
      "[features]":"originalScore=2.2747984,
      titleLength=9.0,contentLength=22552.0"},
    {
      "title":"Science Clips - Volume 12,
      Issue 10, March 23, 2020",
      "url":"https://www.cdc.gov/library/
      sciclips/issues/",
      "[features]":"originalScore=2.2747984,
      titleLength=9.0,contentLength=22552.0"},
    {
      "title":"Breast cancer - Diagnosis
      and treatment - Mayo Clinic",
      "url":"https://www.mayoclinic.org/
      diseases-conditions/breast-cancer/
      diagnosis-treatment/drc-20352475",
      "[features]":
      "originalScore=3.7579455,
      titleLength=7.0,
      contentLength=5144.0"},
    {
      "title":"The Topic Is Cancer | Blogs |
      CDC",
      "url":"https://blogs.cdc.gov/cancer/",
      "[features]":"originalScore=3.5015173,
      titleLength=6.0,
      contentLength=5144.0"},
    {
      "title":"All Issues - Mayo Clinic
      Health Letter",
      "url":"https://healthletter.mayoclinic.
      com/issues",
      "[features]":"originalScore=3.071907,
      titleLength=6.0,contentLength=5144.0"},
    {
      "title":"Menopause Treatment, Signs,
      Symptoms & Age",
      "url":"https://www.medicinenet.com/
      menopause/article.htm",
      "[features]":"originalScore=2.6457264,
      titleLength=5.0,contentLength=5144.0"},
    {
      "title":"Hormone Therapy for Women:
      Side Effects, Cancer Risks",
      "url":"https://www.medicinenet.com/
      hormone_therapy/article.htm",
      "[features]":"originalScore=3.3211832,
      titleLength=8.0,contentLength=4632.0"},
    {
      "title":"Cáncer de mama - Atención en
      Mayo Clinic - Mayo Clinic",
      "url":"https://www.mayoclinic.org/es-es/
      diseases-conditions/breast-cancer/
      care-at-mayo-clinic/mac-20352479",
      "[features]":"originalScore=3.6176624,
      titleLength=9.0,contentLength=4120.0"},
    {
      "title":"Hot Flashes Causes, Symptoms &
      Treatment Medicine for Men & Women",
      "url":"https://www.medicinenet.com/
      hot_flashes/article.htm",
      "[features]":"originalScore=2.8397057,
      titleLength=9.0,contentLength=3864.0"},
    {
      "title":"Breast cancer - Care at Mayo
      Clinic - Mayo Clinic",
      "url":"https://www.mayoclinic.org/
      diseases-conditions/breast-cancer/
      care-at-mayo-clinic/mac-20352479",
      "[features]":"originalScore=3.7574568,
      titleLength=8.0,contentLength=3608.0"}]
}}
```


c) Pagination - First 20 results:


```
{
 "responseHeader":{
    "status":0,
    "QTime":0,
    "params":{
    "q":"cancer",
    "fl":"url,title",
    "start":"0",
    "rows":"20"}},
    "response":{"numFound":1083,
    "start":0,"docs":[
    {
      "title":"Key Statistics for Prostate
      Cancer | Prostate Cancer Facts",
      "url":"https://www.cancer.org/cancer/
      prostate-cancer/about/key-statistics
      .html"},
    {
      "title":"Cancer | CDC",
      "url":"https://www.cdc.gov/cancer/"},
    {
      "title":"Cancer | CDC",
      "url":"https://www.cdc.gov/cancer/
      index.htm"},
    {
      "title":"Print Materials About Cancer
```

```
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/publications/"},
{
  "title":"Other Cancer Data Sources
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/data/other.htm"},
{
  "title":"Cancer Data and Statistics
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/data/"},
{
  "title":"Kinds of Cancer | CDC",
  "url":"https://www.cdc.gov/cancer/
  kinds.htm"},
{
  "title":"Cancer Center: Types,
  Symptoms, Causes, Tests, and
  Treatments, Including Chemo and
  Radiation",
  "url":"https://www.webmd.com/cancer/
  default.htm"},
{
  "title":"Risk Factors and Cancer
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  risk_factors.htm"},
{
  "title":"Cancer Resource Library
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/resources/"},
{
  "title":"Cancer Article Summaries
  | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/research/articles/index.htm"},
{
  "title":"Cancer Article Summaries
  Published in 2017 | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/research/articles/cancer-
  article-summaries-2017.html"},
{
  "title":"Educational Campaigns About
  Cancer | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/about/campaigns.htm"},
{
  "title":"National Programs to Prevent
  and Control Cancer | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/about/programs.htm"},
{
  "title":"Annual Report to the Nation
  on the Status of Cancer | CDC",
  "url":"https://www.cdc.gov/cancer/
  annual-report/index.htm"},
{
  "title":"How to Prevent Cancer or
  Find It Early | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/prevention/"},
{
  "title":"Cancer Research | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/research/index.htm"},
{
  "title":"Healthy People 2020 Targets |
  Annual Report to the Nation | CDC",
  "url":"https://www.cdc.gov/cancer/
  annual-report/healthy-people-2020-
  targets.htm"},
{
  "title":"Cancer Screening Tests | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/prevention/screening.htm"},
{
  "title":"Cancer Survival in the United
  States | CDC",
  "url":"https://www.cdc.gov/cancer/
  dcpc/research/articles/concord-2-
  supplement.htm"}]
}}
```


VI. Conclusions

Our source code is publicly available at https://github.com/devharsh/MediCrawl.

References

[1] "System and user aspects of web search latency."

[2] R. Dragusin, P. Petcu, C. Lioma, B. Larsen, H. L. Jørgensen, I. J. Cox, L. K. Hansen, P. Ingwersen, and O. Winther, "Findzebra: A search engine for rare diseases," *International Journal of Medical Informatics*, vol. 82, no. 6, p. 528–538, 2013.

[3] "Nutch search engine overview."

[4] "Introduction to lucene solr."

[5] "Db-engines ranking of search engines - 20 systems in ranking, april 2020."

[6] "Learning to rank — apache solr reference guide 8.5."