STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY

MediCrawl

A web search engine for medical diagnosis

Devharsh & Vaishnavi
Department of Computer Science
May 4th 2020

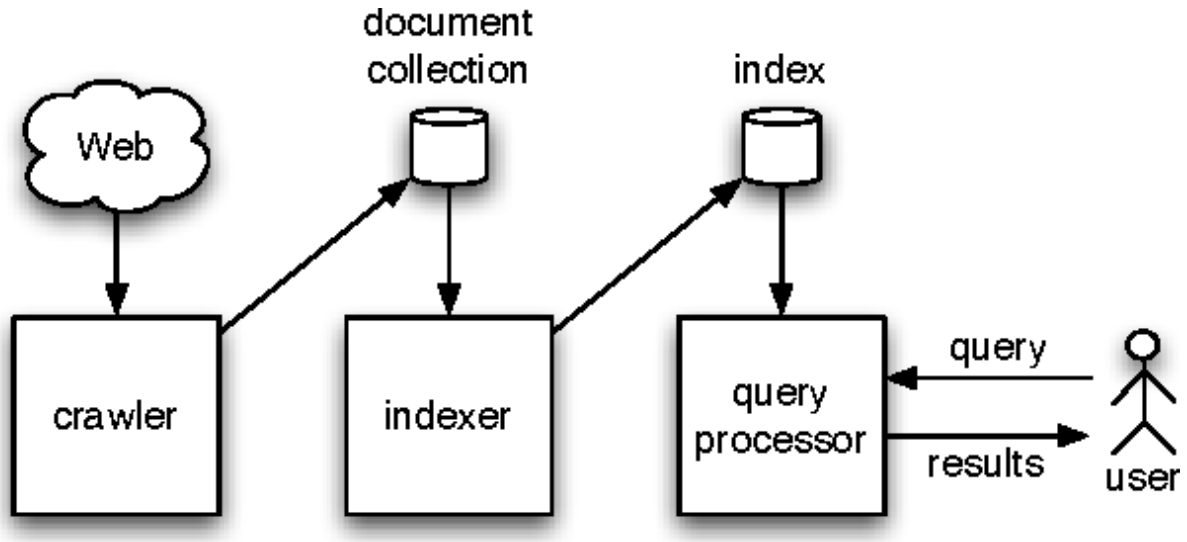
Introduction

Web Search Engine

- Components of a web search engine:

  1. Crawling
  2. Indexing
  3. Ranking
  4. Searching

- Web search engine for diseases:
  — Mayo Clinic
  — FindZebra



Introduction

Mayo Clinic

Find Diseases & Conditions

Find a disease or condition by its first letter

A B C D E F G H I J K L M N O

P Q R S T U V W X Y Z #

Check your symptoms to find a possible cause

Try the Symptom Checker

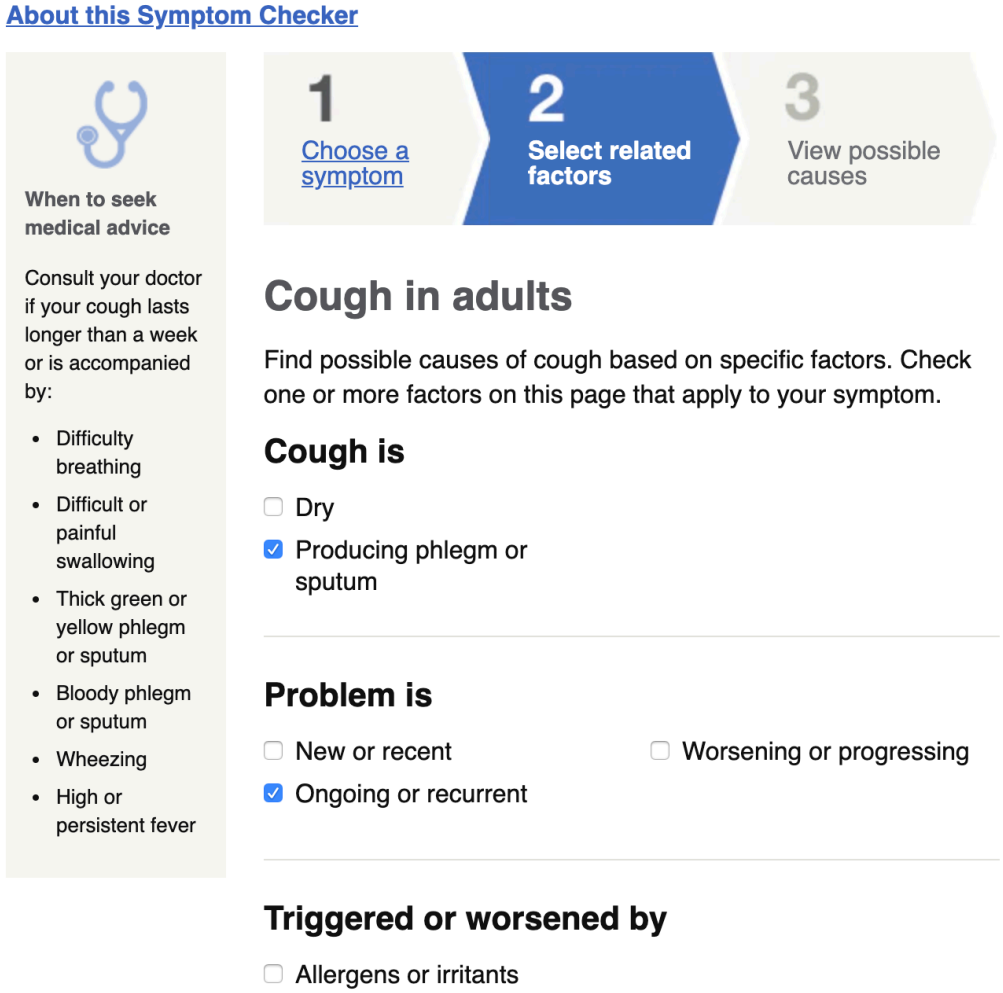
Featured topics

- Bone marrow transplant
- Brain aneurysm
- Brain tumor
- Breast cancer
- Colon cancer
- Congenital heart disease
- Glioma
- Heart arrhythmia
- Heart valve disease
- Living donor transplant
- Lung transplant
- Rectal cancer

See More Diseases and Conditions

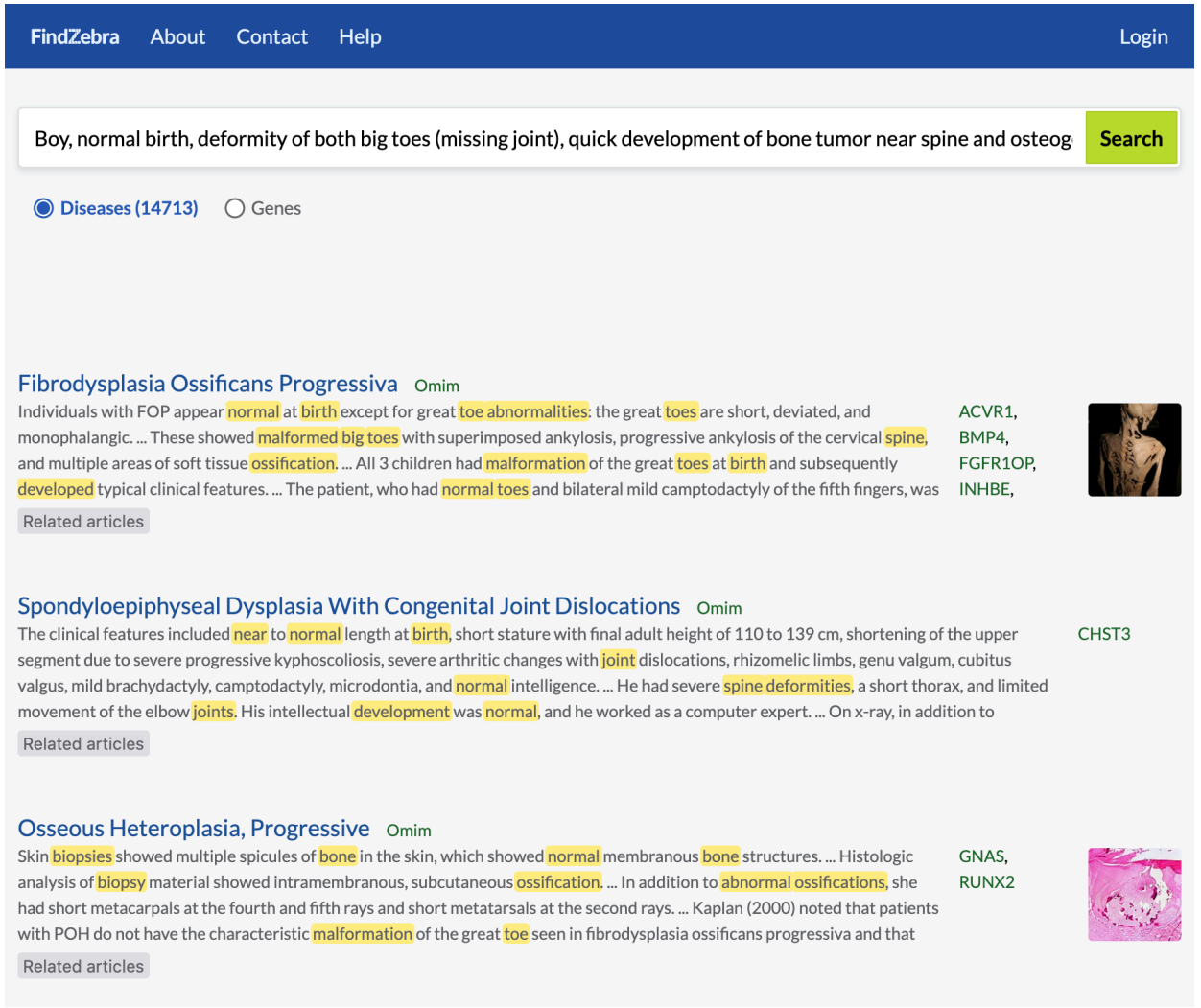
Introduction

Mayo Clinic



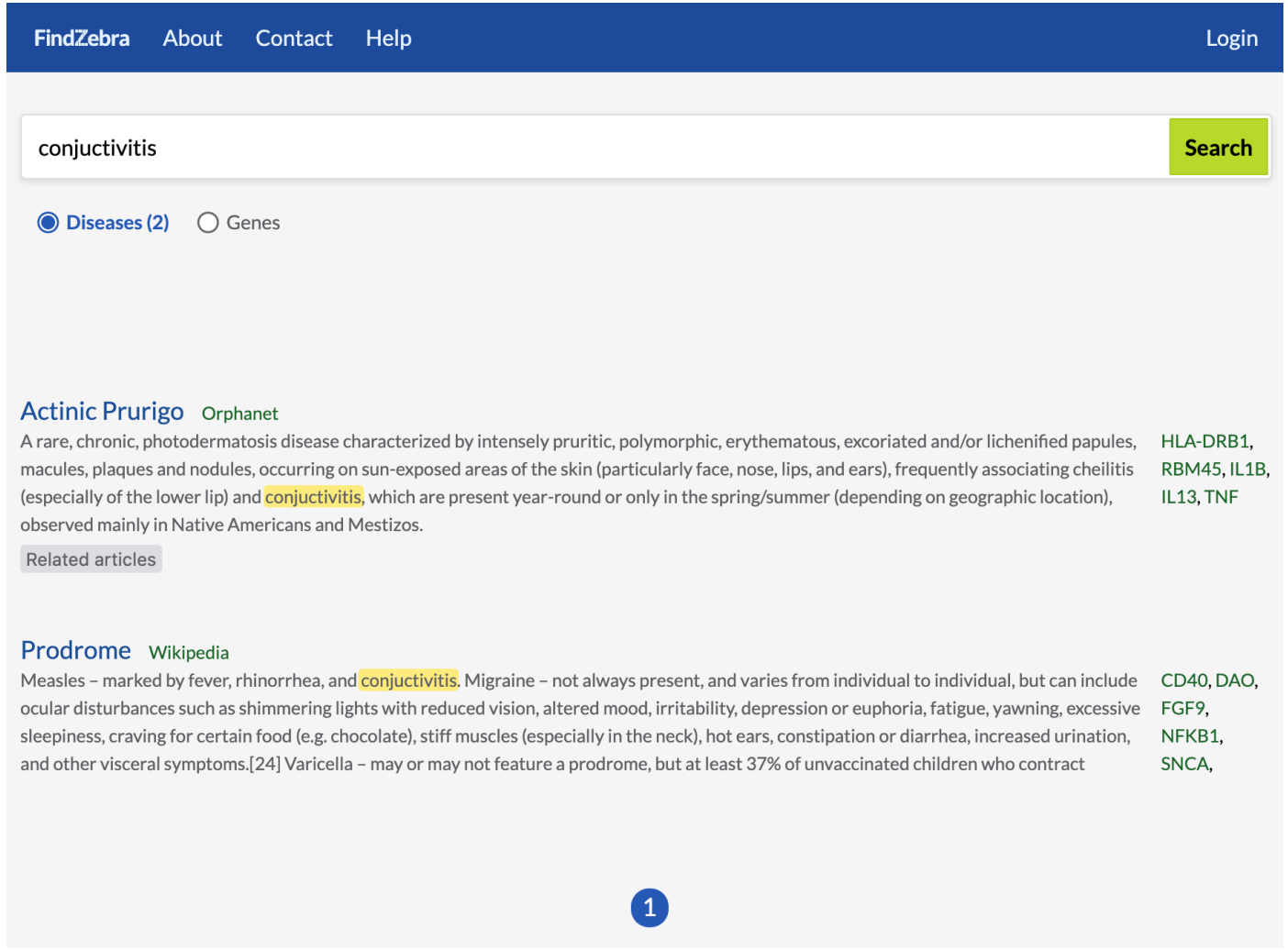
Introduction

FindZebra



Introduction

FindZebra



Motivation – Information Retrieval

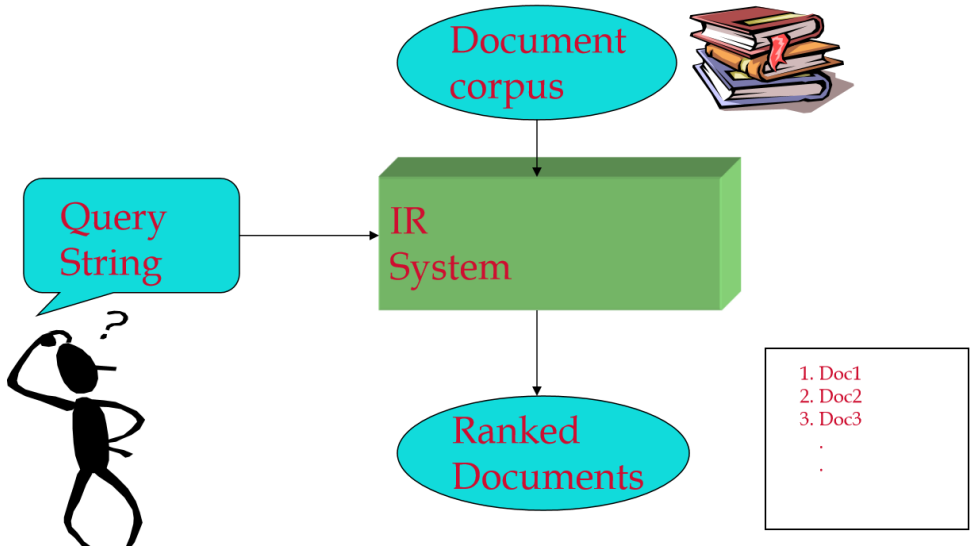
Goals: To find relevant documents to the query from large set of unstructured data

How do we represent text?

- Bag of words
- Represented as vectors

How to get terms?

- Stop words
- Normalization
- Case folding
- Stemming
- Thesauri
- Human factor
- Other languages



Information retrieval process

Motivation

How to organize terms in index structure?

- Inverted Index

How to support phrase queries?

- Positional index

How to reduce search time?

- Binary tree

How to retrieve relevant answers?

- Types of retrieval model

➢ Boolean model – Too rigid & Too few or too many results

➢ Vector space model – TF & IDF

➢Led to BM25

BM25

Ranking algorithm

- BM for Best Match
- Probabilistic term weighting model
- Larger documents – larger term frequency values
- BM25 – Normalize

BM25 ranking function

$$RSV^{BM\,25} = \sum_{i \in q} c_i^{BM\,25}(tf_i);$$

$$\sum_{i \in Q} \log \frac{(r_i+0.5)/(R-r_i+0.5)}{(n_i-r_i+0.5)/(N-n_i-R+r_i+0.5)} \cdot \frac{(k_1+1)f_i}{K+f_i} \cdot \frac{(k_2+1)qf_i}{k_2+qf_i}$$

$$K = k_1((1-b)+b \cdot \frac{dl}{avdl})$$

- k_1, k_2 and K are parameters whose values are set empirically
- dl is doc length
- Typical value for k_1 is 1.2, k_2 varies from 0 to 1000, b = 0.75

Explanation of BM25

- Query with two terms, "president lincoln", (*qf = 1)*
- No relevance information (*r and R are* zero)
- *N* = 500,000 documents
- *"president"* occurs in 40,000 documents (n_1 = 40, 000)
- *"lincoln"* occurs in 300 documents (n_2 = 300)
- "president" occurs 15 times in doc (f_1 = 15)
- *"lincoln"* occurs 25 times (f_2 = 25)
- document length is 90% of the average length (*dl/avdl* = .9)
- k_1 = 1.2, b = 0.75, and k_2 = 100
- $K = 1.2 \cdot (0.25 + 0.75 \cdot 0.9) = 1.11$

$$
\begin{aligned}
BM25(Q,D) &= \\
&\log \frac{(0+0.5)/(0-0+0.5)}{(40000-0+0.5)/(500000-40000-0+0+0.5)} \\
&\times \frac{(1.2+1)15}{1.11+15} \times \frac{(100+1)1}{100+1} \\
&+ \log \frac{(0+0.5)/(0-0+0.5)}{(300-0+0.5)/(500000-300-0+0+0.5)} \\
&\times \frac{(1.2+1)25}{1.11+25} \times \frac{(100+1)1}{100+1} \\
&= \log 460000.5/40000.5 \cdot 33/16.11 \cdot 101/101 \\
&\quad + \log 499700.5/300.5 \cdot 55/26.11 \cdot 101/101 \\
&= 2.44 \cdot 2.05 \cdot 1 + 7.42 \cdot 2.11 \cdot 1 \\
&= 5.00 + 15.66 = 20.66
\end{aligned}
$$

Implementation

Tools & Technologies

- Java (Runtime Environment)
- Web browser (HTML / CSS / JS)
- git
- Nutch
- Solr
- Node.js
- **ALL ARE FOSS!! (Free or Open Source Software)**

Implementation

Nutch

- Open source Web Crawler
- Written in Java
- Implements “Map-Reduce” distributing processing model
- Built on top of Apache Lucene
- Fetches pages and turn them into inverted index
- Configuration:
  — nutch/conf/**nutch-site.xml**
  — nutch/conf/**schema.xml**
  — nutch/urls/**seed.txt**
- Crawl and Index simultaneously:
  nutch/bin/crawl -i -Dsolr.server.url=http://localhost:8983/solr/nutch -s nutch/urls/ Crawl 3

Implementation

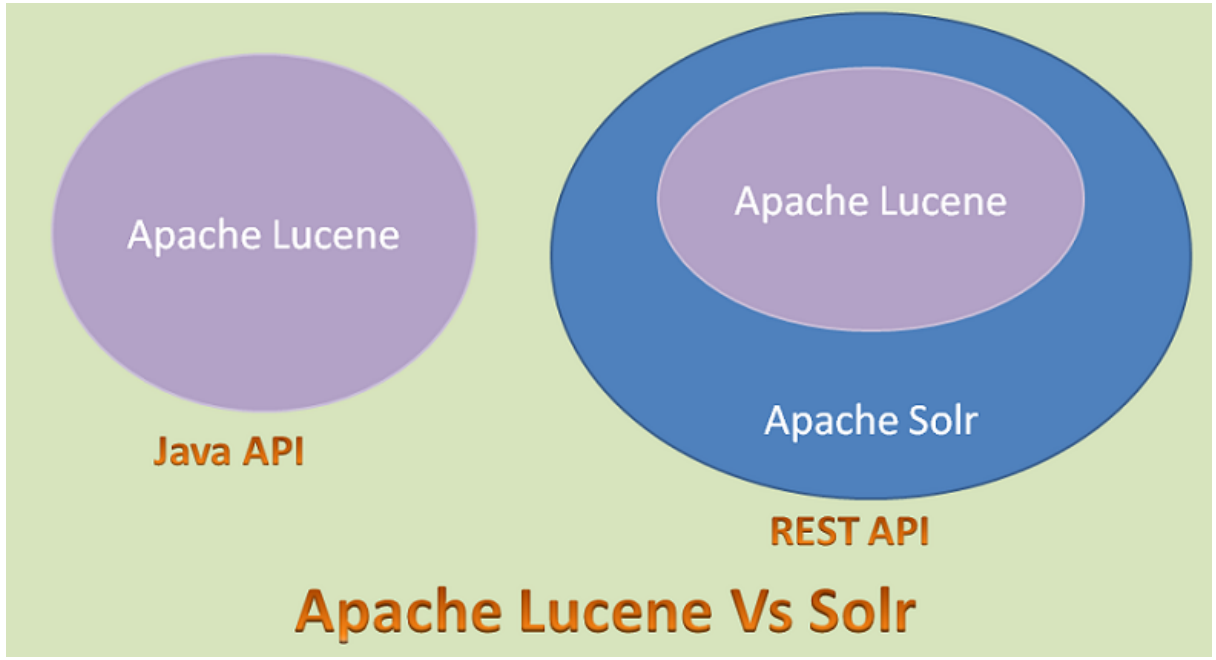
nutch/url/seed.txt

1. https://www.everydayhealth.com/infectious-diseases/
2. https://www.healthed.govt.nz/resource-table/table-infectious-diseases
3. https://www.livescience.com/36519-diseases-conditions-symptoms-treatments.html
4. https://www.healthline.com/health/infections#types
5. https://www.nhsinform.scot/illnesses-and-conditions/a-to-z
6. https://www.socialstyrelsen.se/rarediseases
7. https://www.hon.ch/en/
8. https://www.verywellhealth.com/rare-diseases-4014657
9. https://ghr.nlm.nih.gov/
10. https://rarediseases.org/
11. https://www.orpha.net/consor/cgi-bin/index.php
12. https://rarediseases.info.nih.gov/
13. https://www.ncbi.nlm.nih.gov/omim
14. https://www.mayoclinic.org/
15. https://www.who.int/en/
16. https://www.webmd.com/
17. https://www.adam.com/
18. https://www.medicalnewstoday.com/
19. https://www.cdc.gov/diseasesconditions/index.html
20. https://www.health.harvard.edu/

Implementation

Solr

- Open source enterprise search server
- No-SQL (non-traditional data store)
- Based on Lucene Java search library
- Wrapper:
  — HTTP request/response
  — XML API
  — caching / replication
  — web admin interface
- Configuration:
  solr/server/solr/configsets/nutch/conf/**solrconfig.xml**
- Run:
  solr/bin/solr start -Dsolr.ltr.enabled=true
- **Create a solr core before indexing from nutch!**



Implementation

Solr LTR (Learning to Rank model)

- 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.
- public class **LinearModel** extends LTRScoringModel
- A scoring model that computes scores using a dot product. Example models are RankSVM and Pranking.
- Example configuration:

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

Implementation

Solr LTR : Original Score Feature

- public class OriginalScoreFeature extends Feature
- This feature returns the original score that the document had before performing the reranking.
- Example configuration:

```
{
  "name":  "originalScore",
  "class": “org.apache.solr.ltr.feature.OriginalScoreFeature",
  "params": { }
}
```

Demonstration

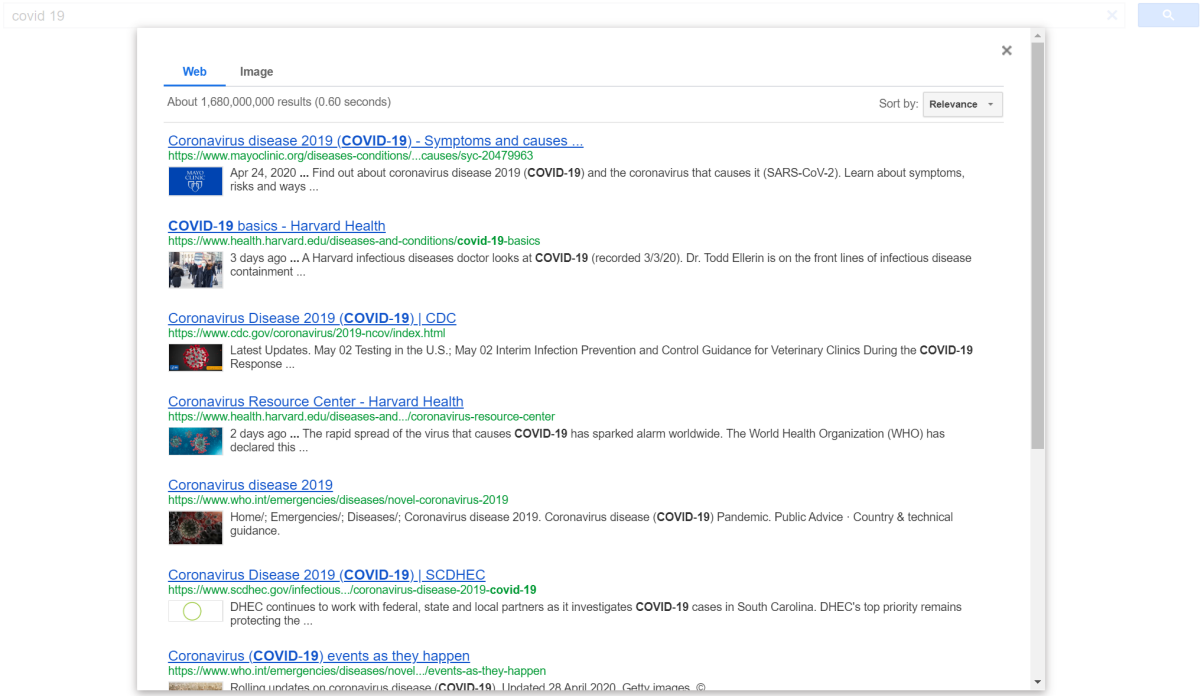
Disease	Symptoms
Covid-19	shortness of breath, lack of taste, fever
Schizophrenia	delusions, hallucinations, disorganized speech, lack of motivation or emotion
Mumps	swelling of salivary glands, usually the parotid glands
Toxic Epidermal Necrolysis	bullous skin conditions, respiratory failure, carbamazepine
Propionic Acidemia	girl, hypotonia, seizures, dehydration, polypnea, acidosis, massive ketonuria, hyperammonemia
Cushing (secondary to adrenal)	hypertension, adrenal mass

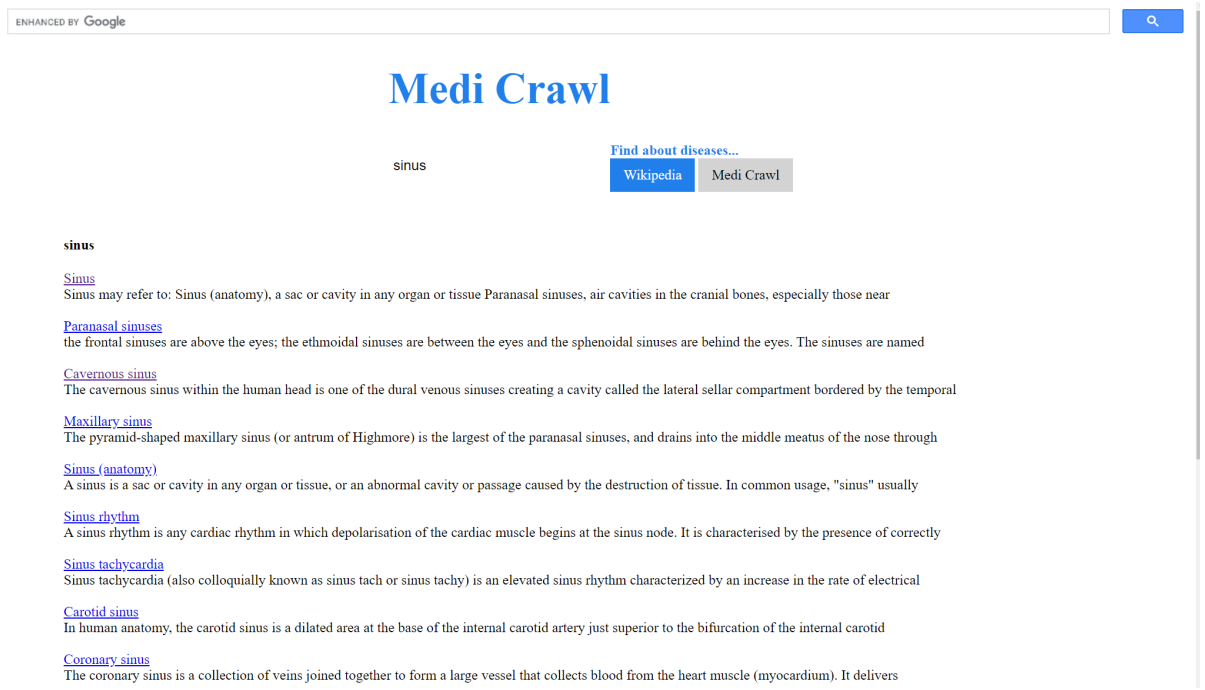
Demonstration

MediCrawl

Demonstration

MediCrawl : Google & Wikipedia API





Challenges

- Retrieve relevant documents to a query
- Retrieve from large set of documents efficiently
- Ranking and Searching issues with Long tailed queries
- Availability of dataset to test

Results

Document types (total = 54759)

text/html	49039
application/xhtml+xml	3948
application/pdf	1067
image/svg+xml	286
application/rss+xml	177
image/jpeg	78
text/plain	44
application/xml	20
application/vnd.openxmlformats-officedocument.wordprocessingml.document	16
image/png	14
audio/vorbis	13

video/mp4	10
image/gif	7
application/vnd.openxmlformats-officedocument.presentationml.presentation	3
application/vnd.openxmlformats-officedocument.spreadsheetml.sheet	3
application/rtf	2
image/vnd.microsoft.icon	2
text/csv	2
application/epub+zip	1
application/msword	1
application/vnd.ms-excel.sheet.macroenabled.12	1
audio/midi	1

Results

Ranking

- Sample Linear model ranking:


```
 "params":{
    "q":"cold",
    "fl":"title,url,[features]"}},
"response":{"numFound":520,"start":0,"docs":[
    {
      "title":"Common colds: Symptoms, causes, complications, and treatment",
      "url":"https://www.medicalnewstoday.com/articles/166606",
      "[features]":"originalScore=1.9142148,titleLength=7.0,contentLength=1560.0"},
    {
      "title":"Cold and Flu (Influenza) Center: Symptoms, Treatments, Causes, and Prevention",
      "url":"https://www.webmd.com/cold-and-flu/default.htm",
      "[features]":"originalScore=1.9114048,titleLength=10.0,contentLength=1048.0"},
    {
    ..
    ..
```

Conclusion

Conclusion & Future Work

- We present an efficient web search engine for medical diagnosis
- Works well for both common and rare diseases
- Easy to implement and portable
- Using open source / free softwares
- Future work could be implementing a Deep learning model that takes symptoms as inputs and outputs probabilities of diseases and probable remedies
- source code available at https://github.com/devharsh/MediCrawl

STEVENS

INSTITUTE *of* TECHNOLOGY

THE INNOVATION UNIVERSITY®

stevens.edu

Devharsh Trivedi (dtrived5@stevens.edu)

Vaishnavi Gopalakrishnan (vgopalak@stevens.edu)