Wiki search engine

PHASE II

Information Retrieval and Extraction
Mini Project

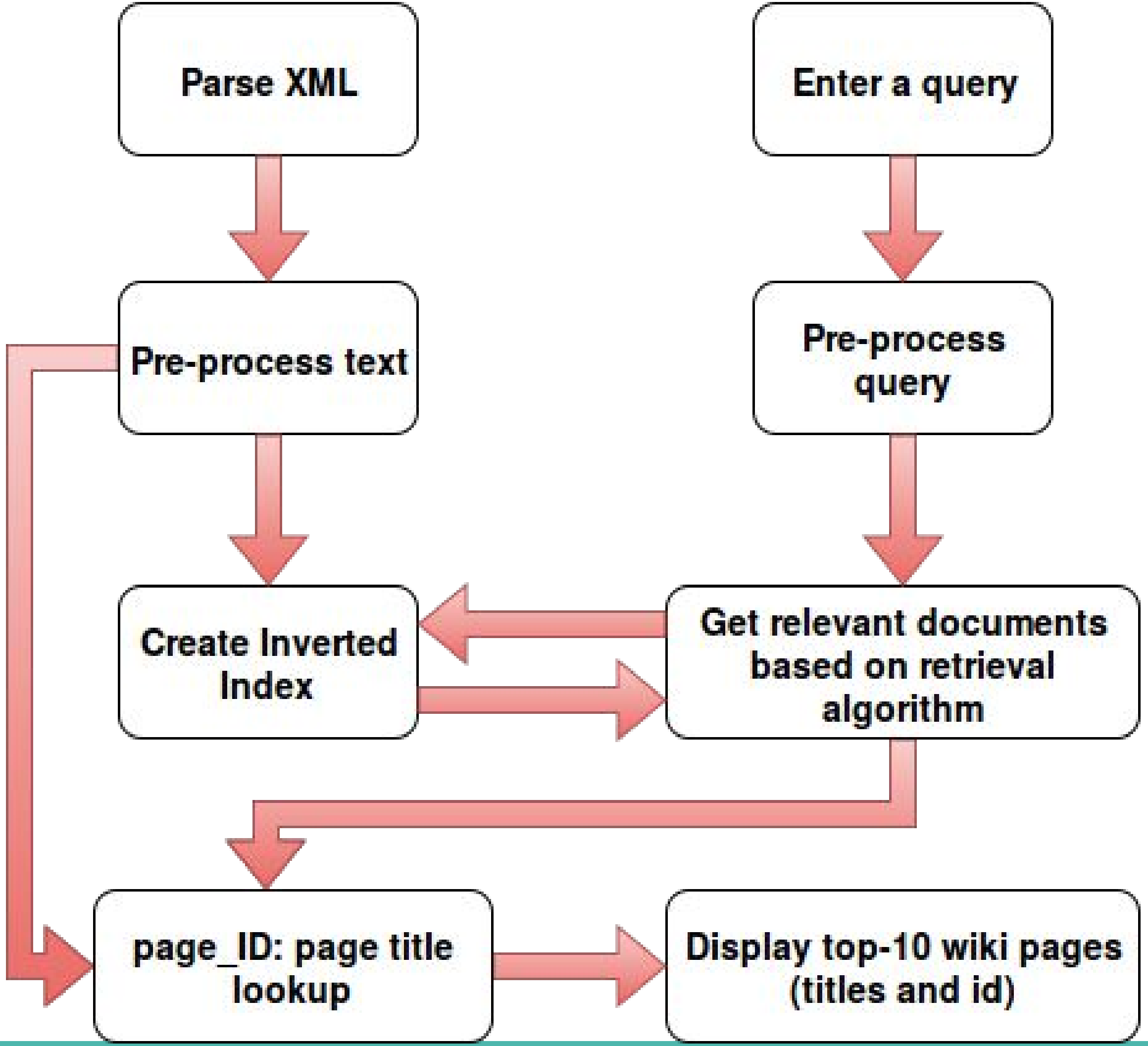
Indexing and Retrieval

FULL English WikiDump

~**46GB** uncompressed

http://10.2.4.182:8080/enwiki-latest-pages-articles.xml.tar.gz

https://dumps.wikimedia.org/enwiki/latest/enwiki-latest-pages-articles.xml.bz2


Parse XML
Pre-process text
Create Inverted Index
page_ID: page title lookup
Enter a query
Pre-process query
Get relevant documents based on retrieval algorithm
Display top-10 wiki pages (titles and id)

Ranking

- tf-idf weighting
- vector space ranking
  - jaccard similarity
  - cosine similarity
- probabilistic ranking
  - Okapi BM25

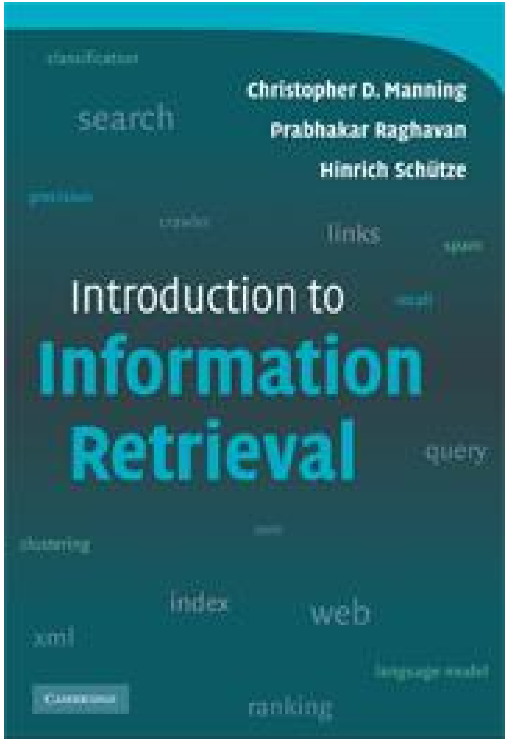
$$\mathbf{tf}\,(t,d) = \mathbf{f}\,(t,d)$$

$$\mathbf{idf}\,(t) = \mathbf{log}\,(\mathbf{N} \,/\, \mathbf{df}(t))$$

$$\mathbf{tf\text{-}idf}\,(t,d) = \mathbf{tf}\,(t,d) \times \mathbf{idf}\,(t)$$

$$\text{score}(q,d) = \sum_{t \in q} \text{tf-idf}_{t,d}$$

Ranking



classification

search

Christopher D. Manning

Prabhakar Raghavan

Hinrich Schütze

precision

crawler

links

spam

Introduction to

recall

Information Retrieval

query

clustering

index

web

xml

language model

Cambridge

ranking

Chapters: **6, 7, 11**

http://nlp.stanford.edu/IR-book/

Sample Queries

- he who must not be named
- t: the two towers i: 1954
- jon snow
- t: sachin b: e-commerce

Weighting Fields:

Decide your own ranking parameters / weights

Results should be displayed within **0 - 5 seconds** depending upon query type / length

Challenges

- **multi-level** indexing (retrieval speed)
- efficient use of **data-structures**, **algorithms** (indexing speed, memory)
- **threading** for long / multi-field queries (retrieval speed)
- index **compression** (index size reduction)
- arbitrary / long / multi-field queries (early search termination)
- efficient code debugging (it might take **~10 hours** to index full dump)

Thank You

Please stay back for doubts