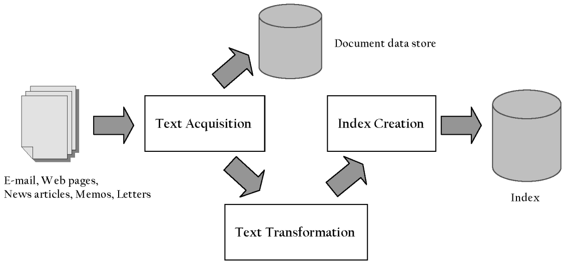
Information Retrieval
INFO 4300 / CS 4300

- Remaining Work for the Course
  - Project 3 – Programming (Fri)
    - » Hardcopy (Mon)
  - Critique 3 (Mon)
    - » Hardcopy (Tues)
  - Final exam
    - » Info on course web page
    - » Wednesday, December 18th, 7-9:30 PM, Barton Hall 100 East-Main Floor.

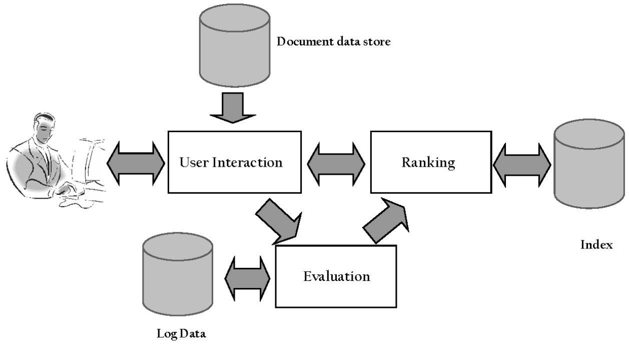
Quick Semester Review

- High-level view of search engine architecture
  - Indexing



Quick Semester Review

- High-level view of search engine architecture
  - Querying



Text Acquisition

- Web crawlers
  - How to retrieve web pages
  - Crawling the web
    - » Basic algorithm for web crawling

```
procedure CRAWLERTHREAD(frontier)
    while not frontier.done() do
        website ← frontier.nextSite()
        url ← website.nextURL()
        if website.permitsCrawl(url) then
            text ← retrieveURL(url)
            storeDocument(url, text)
            for each url in parse(text) do
                frontier.addURL(url)
            end for
        end if
        frontier.releaseSite(website)
    end while
end procedure
```

Text Acquisition

- Web crawlers
  - How to retrieve web pages
  - Crawling the web
    - Basic algorithm for web crawling
    - Politeness policies
    - Complications
      - Freshness (vs. Age) andrelevant HTTP protocol request options
      - Focused crawling (role of classification methods)
      - Deep web
    - Web crawls vs. desktop crawls
    - Usefulness of document feeds (RSS)

Text Acquisition

- Web crawlers
  - File conversion issues
  - Storing the documents
    - File compression
    - BigTable – Google’s document storage system
    - (Near)Duplicate detection

Text Transformation

- Word occurrence statistics
  - Zipf’s Law - distribution
  - Heap’s Law – vocabulary growth
- Issues for
  - Tokenization
  - Stopword removal
  - Stemming
  - Phrases
  - Document structure
  - Link analysis
  - Information extraction

Index Creation

- Inverted indexes
  - Word counts, proximity, fields and extents
- Index construction
  - Jon Park...
- Coding schemes for index compression

Retrieval Models

- Boolean retrieval
- Vector Space model
- Probabilistic Models
  - Binary independence model
  - BM25
  - Language models
    - Query likelihood model
    - Document likelihood model
    - Relevance model (compares the LMs representing the query and document topics)
- Learning to rank

User Interaction: Query Refinement

- Query transformation: stemming
- Query expansion
  - Thesaurus-based
  - Term association measures
  - Relevance feedback

Evaluation

- Covered throughout the semester
- Methods
  - Training, testing
  - Pooling
  - Query logs
- Metrics
  - Recall, precision, F-measure
  - Precision at rank R
  - Reciprocal rank
  - DCG, NDCG

Clustering and Classification

- Recently covered

Guest lectures

- Music
- Patent retrieval
- Topic modeling

(Jon covered material already mentioned.)

Other stuff

- Critique papers
- Analytical questions
- Programming

Course evaluations: 1%