Queries and Indexes

CISC489/689-010, Lecture #7

Wednesday, March 4

Ben Carterette

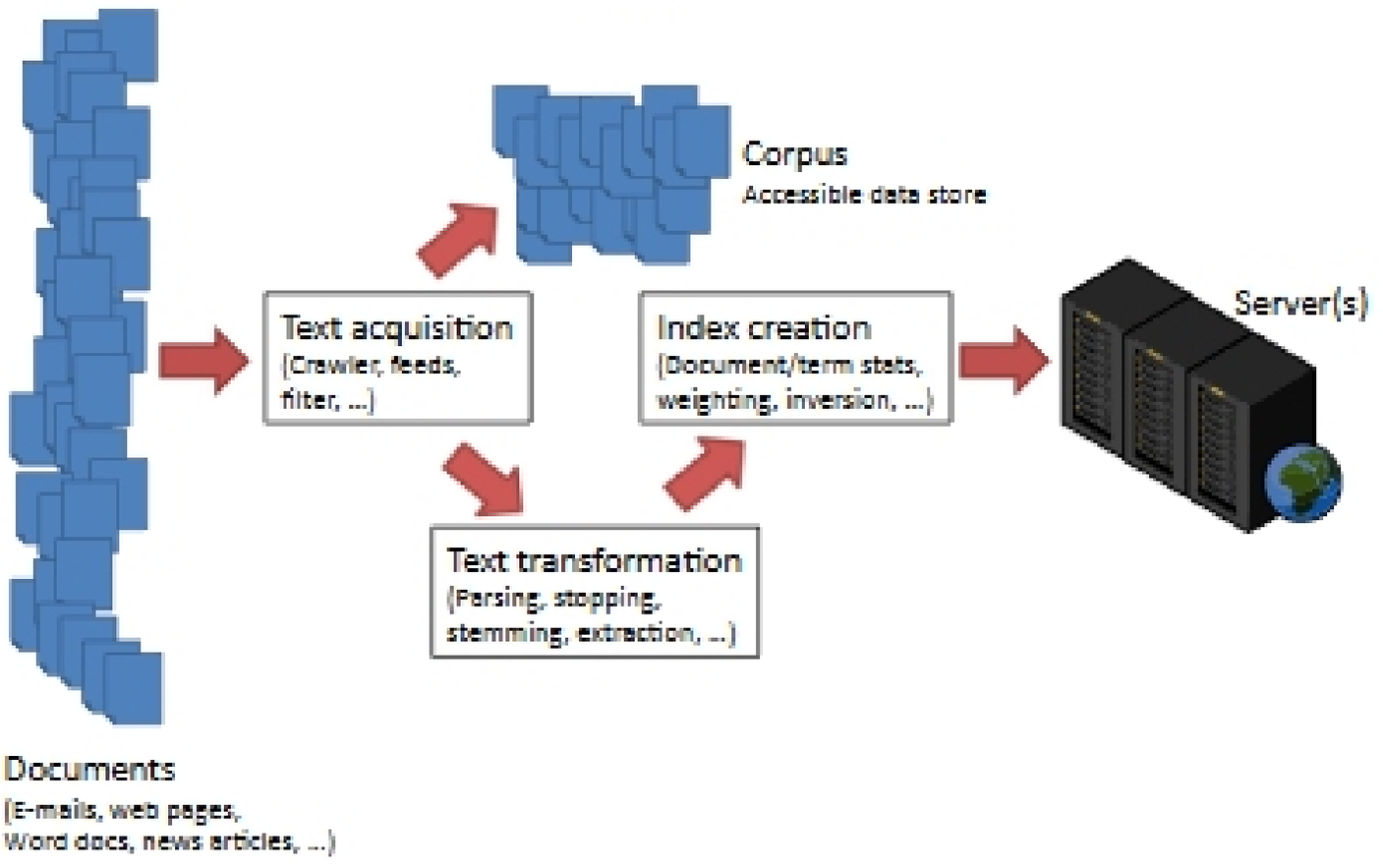
Project Notes

- Next worksheet:
  - Inverted lists for terms in the wiki000 documents.
  - For each term, store:
    - The list of document numbers it occurs in.
    - The term frequencies in those documents.
    - The document frequency (total number of documents it occurs in).
  - If inclined, you may store other information:
    - Term positions, field information, etc.

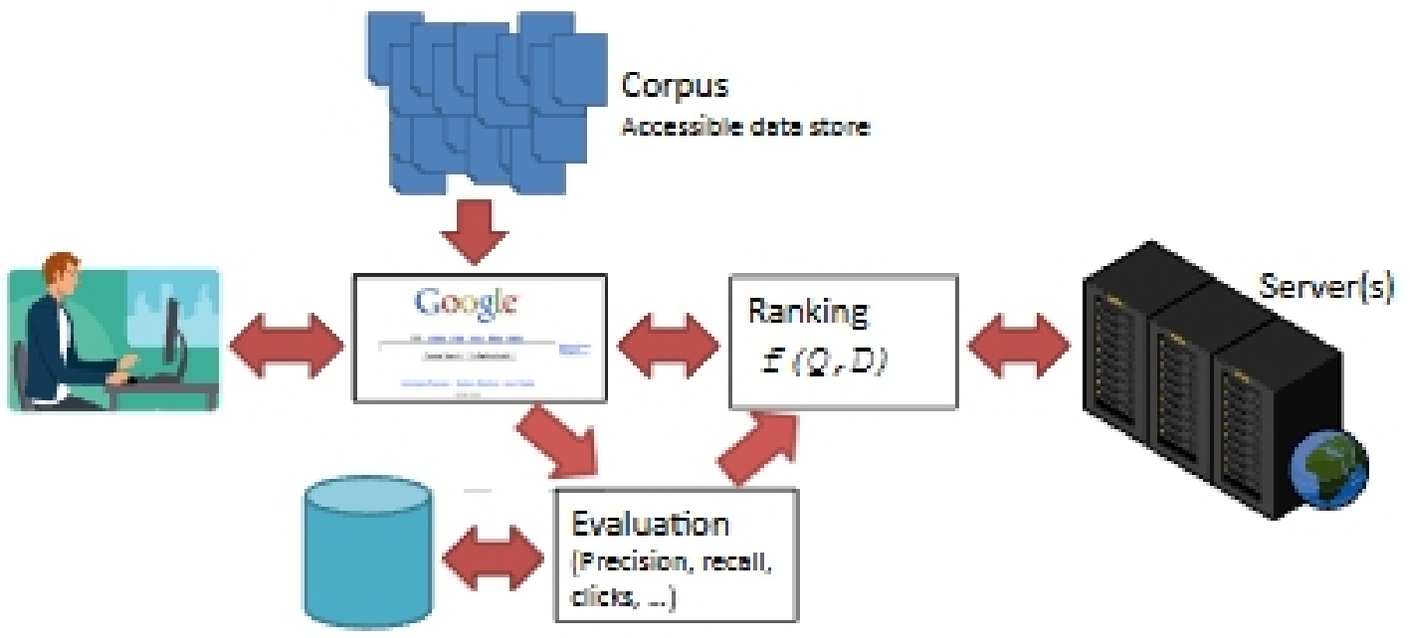
Project Notes

- Inverted list compression:
  - You should compress the inverted lists.
  - Use d-gaps for document numbers.
  - Compress integers using one of the methods discussed in class.
- Store everything in memory.
  - Writing to disk will be the next part of the project.
  - I strongly recommend using ir.cis to run your code.
    - It has a total of 128Gb of RAM (8 nodes, 16Gb per node).

Indexing Process



Query Process



Query Process

- We have an index stored on disk.
  - Inverted file, vocabulary, collection.
  - Contains features of terms and documents:
    - Term frequencies in documents, document frequencies, term positions, link-graph features, ...
- User inputs a query.
- Engine computes features of the query.
- Engine accesses index to respond to query.
  - Matches query features to document/term features in index to score each document.
- Returns a ranked list of documents.