Text Processing

CISC489/689-010, Lecture #3

Monday, Feb. 16

Ben Carterette

Indexing

- An *index* is a list of things (keys) with pointers to other things (items).
  - Keywords → catalog numbers (→ shelves).
  - Concepts → page numbers.
  - Terms → documents.
- Need for indexes:
  - Ease of use.
  - Speed.
  - Scalability.

Manual vs. Automatic Indexing

- Manual:
  - An "expert" assigns keys to each item.
  - Example: card catalog.
- Automatic:
  - Keys automatically identified and assigned.
  - Example: Google.
- Automatic as good as manual for most purposes.

Text Processing

- First step in automatic indexing.
- Converting documents into *index terms.*
- Terms are not just words.
  - Not all words are of equal value in a search.
  - Sometimes not clear where words begin and end.
    - Especially when not space-separated, e.g. Chinese, Korean.
  - Matching the exact words typed by the user doesn't work very well in terms of effectiveness.

Text Processing Steps

- For each document:
  - Parse it to locate the parts that are important.
  - Segment and tokenize the text in the important parts to get *words*.
  - Remove *stop words*.
  - *Stem* words to common roots.
- Advanced processing may included phrases, entity tagging, link-graph features, and more.

Parsing

- Some parts of a document are more important than others.
- Document parser recognizes structure using *markup* such as HTML tags.
  - Headers, anchor text, bolded text are likely to be important.
  - JavaScript, style information, navigation links less likely to be important.
  - Metadata can also be important.