CSCI 4152/6509 — Natural Language Processing *30-Sep-2009*

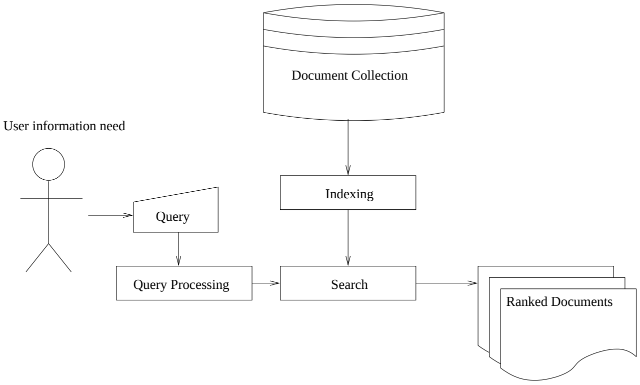
Lecture 9: Vector Space Model

Room: FASS 2176
Time: 11:35 – 12:25

Previous Lecture

- Aside: WordNet web site
- lexical semantics (continued):
  - metonymy,
  - WordNet resource,
  - semantic relations between words,
- semantic compositionality,
- semantic roles;
- Part II: Statistical approach to NLP,
- logical and plausible reasoning,
- two paradigms of NLP: logical and plausible,
- counting words and n-grams, Zipf's law,
- Elements of information retrieval,
- basic task definition of ad-hoc retrieval, typical IR system architecture

Typical IR System Architecture



Steps in Document and Query Processing

- stop-word removal
- rare word removal (optional)
- stemming
- optional query expansion
- document indexing
- document and query representation; e.g. vectors

Vector Space Model in IR

- after choosing a global set of terms $\{t_1, t_2, \dots, t_m\}$, documents and queries are represented as vectors of weights:

$$\vec{d} = (w_{1j}, w_{2j}, \dots, w_{mj}) \quad \vec{q} = (w_{1q}, w_{2q}, \dots, w_{mq})$$

- What are the weights? They could be binary (1 or 0), or term frequency, or something else.
- A standard choice is: *tfidf* — term frequency inverse document frequency weights

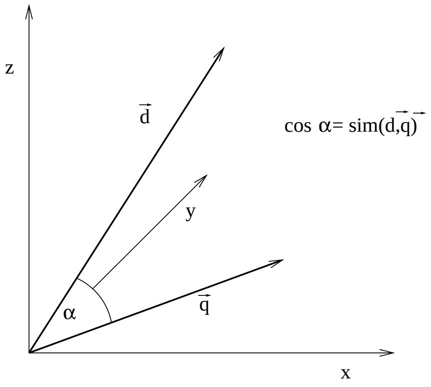
$$\mathit{tfidf} = \mathit{tf} \cdot \log\left(\frac{N}{\mathit{df}}\right)$$

- *tf* is frequency (count) of a term in document, which is sometimes log-ed as well
- *df* is document frequency, i.e., number of documents in the collection containing the term

Similarity Measure

- Cosine similarity measure

$$\mathit{sim}(q,d) = \frac{\sum_{i=1}^{m} w_{iq} w_{id}}{\sqrt{\sum_{i=1}^{m} w_{iq}^2} \cdot \sqrt{\sum_{i=1}^{m} w_{id}^2}} = \frac{\vec{q} \cdot \vec{d}}{|\vec{q}| \cdot |\vec{d}|}$$



Term-by-Document Matrix

- Term-by-Document matrix

	d_1	d_2	$\dots$	d_n
t_1	w_{11}	w_{12}	$\dots$	w_{1n}
t_2	w_{21}	w_{22}	$\dots$	w_{2n}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
t_m	w_{m1}	w_{m2}	$\dots$	w_{mn}

- reducing number of dimensions
  - stemming and stop-words
  - feature selection
  - Latent Semantic Analysis

Latent Semantic Analysis

- method for term-by-document dimensionality reduction
- singular value decomposition: $M_{m\times n} = U_{m\times m}\Sigma_{m\times n}V^T_{n\times n}$
- example with four terms and two documents
- closest by Frobenius norm matrix of rank $\le k$ is $M^{(k)}_{m\times n} = U_{m\times m}\Sigma^{(k)}_{m\times n}V^T_{n\times n}$
- concept and document representations