Text Representation
http://www.cse.iitb.ac.in/~soumen/mining-the-web/

Ahmed Rafea
Text Representation

- Document Preprocessing
- Vector Space Model for Document Storage
- Measure of Similarity
Document preprocessing (1/3)

- **Tokenization**
  - Filtering away tags
  - Tokens regarded as nonempty sequence of characters excluding spaces and punctuations.
  - Token represented by a suitable integer, $tid$, typically 32 bits
  - Optional: stemming/conflation of words
  - Result: document (did) transformed into a sequence of integers ($tid$, $pos$)
Document preprocessing (2/3)

- **Stopwords**
  - Function words and connectives
  - Appear in large number of documents and little use in pinpointing documents
  - **Issues**
    - Queries containing only stopwords ruled out
    - Polysemous words that are stopwords in one sense but not in others
      - E.g.; *can* as a verb vs. *can* as a noun
Document preprocessing (3/3)

- **Stemming**
  - Remove inflections that convey parts of speech, tense and number
  - E.g.: university and universal both stem to universe.
  - Techniques
    - morphological analysis (e.g., Porter's algorithm)
    - dictionary lookup (e.g., WordNet).
  - Stemming may increase the number of documents in the response of a query but at the price of precision
    - It is not a good idea to stem Abbreviations, and names coined in the technical and commercial sectors
    - E.g.: Stemming “ides” to “IDE”, the hard disk standard, “SOCKS” firewall protocol to “sock” worn on the foot, may be bad!
The vector space model (1/4)

- Documents represented as vectors in a multi-dimensional Euclidean space
  - Each axis = a term (token)
- Coordinate of document d in direction of term t determined by:
  - Term frequency TF(d,t)
    - number of times term t occurs in document d, scaled in a variety of ways to normalize document length
  - Inverse document frequency IDF(t)
    - to scale down the coordinates of terms that occur in many documents
The vector space model (2/4)

- **Term frequency**

  \[ \text{TF}(d, t) = \frac{n(d, t)}{\sum_{\tau} n(d, \tau)} \]

  \[ \text{TF}(d, t) = \frac{n(d, t)}{\max_{\tau} (n(d, \tau))} \]

- **Cornell SMART system uses a smoothed version**

  \[ TF(d,t) = 0 \quad \text{if} \quad n(d,t) = 0 \]

  \[ TF(d,t) = 1 + \log(1 + n(d,t)) \quad \text{otherwise} \]
The vector space model (3/4)

- Inverse document frequency
  - Given
    - D is the document collection and $D_t$ is the set of documents containing t
  - Formulae
    - mostly dampened functions of $\frac{D}{|D_t|}$
    - SMART

$$IDF(t) = \log\left(\frac{1+|D|}{|D_t|}\right)$$
Vector space model (4/4)

- Coordinate of document $d$ in axis $t$
  - $d_t = TF(d, t) \times IDF(t)$
  - Transformed to $\tilde{d}$ in the TFIDF-space

- Query $q$
  - Interpreted as a document
  - Transformed to $\tilde{q}$ in the same TFIDF-space as $d$
Measures of Similarity (1/2)

- Distance measure
  - Magnitude of the vector difference
    - $\cdot | \vec{d} - \vec{q} |$
  - Document vectors must be normalized to unit length ($L_1$ or $L_2$)
    - Else shorter documents dominate (since queries are short)

- Cosine similarity
  - cosine of the angle between $\vec{d}$ and $\vec{q}$
    - Shorter documents are penalized
Measures of Similarity (2/2)

- Jaccard coefficient of similarity between document $d_1$ and $d_2$
- $T(d) =$ set of tokens in document $d$
  - Symmetric, reflexive
  - Forgives any number of occurrences and any permutations of the terms.

\[
. r'(d_1, d_2) = \frac{|T(d_1) \cap T(d_2)|}{|T(d_1) \cup T(d_2)|}
\]

- Symmetric, reflexive
- Forgives any number of occurrences and any permutations of the terms.