Vector Space Model
Conclusions and Summary

COMP3009J: Information Retrieval

Dr. David Lillis (david.lillis@ucd.ie)

UCD School of Computer Science
Beijing Dublin International College

TF-IDF Advantages

- The advantages of the TF-IDF weighting scheme over others are:
  - Using term frequency gives more influence to a document that contains many occurrences of a particular term.
  - Using inverse document frequency lessens the impact of terms that appear very often in the collection. These terms are regarded as being less informative.

Vector Space Model: Advantages and Disadvantages

- There are three main advantages of the vector space model:
  - The term weighting schemes can improve retrieval performance.
  - Partial matching strategy allows for retrieval of documents that match part of the query.
  - The cosine similarity can be used to return a ranked list of documents .
- The main disadvantage of the model is the assumption that terms are independent, which can sometimes harm performance.

Vector Space Model: Implementation

- The main steps in implementing the vector space model are:
  1. Read in the documents.
  2. Divide each document into its constituent terms.
  3. Remove stopwords.
  4. Stem terms.
  5. Add documents to an index.
  6. Calculate TF-IDF.
     - Each term has one IDF score.
     - Each term has a separate TF score in each document.
  7. Receive a query.
  8. Retrieve query results.

Summary

- The **Vector Space Model** of Information Retrieval is based on concepts from vector algebra.
- Documents and queries are represented as vectors.
- Cosine similarity is used to decide which documents most closely match a query.
- Terms can be weighted in multiple ways.
- A common weighting scheme is TF-IDF (Term Frequency-Inverse Document Frequency).