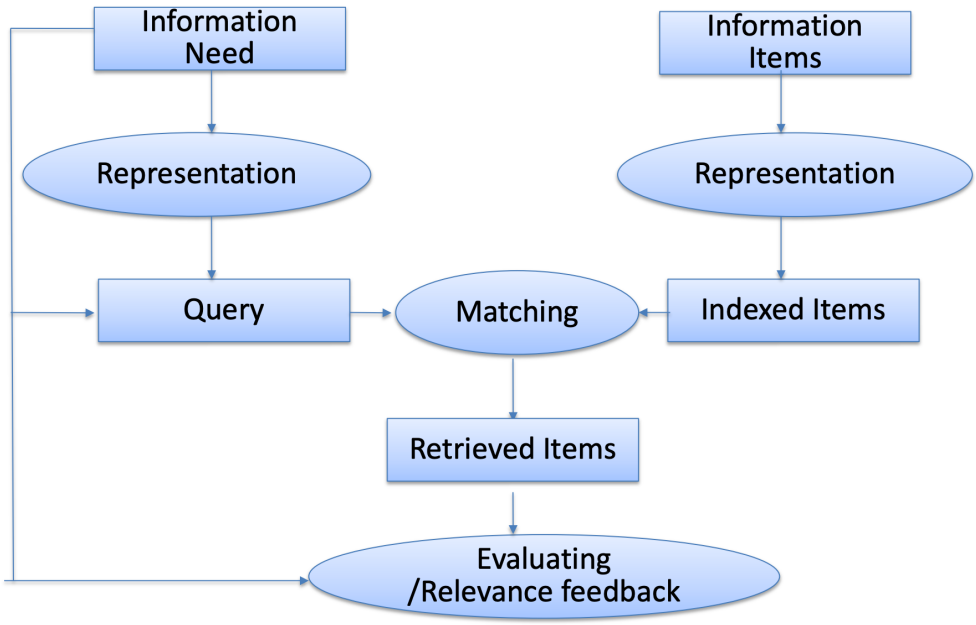
Week 1 - Introduction to Information Retrieval ¶

Information Retrieval



Example of **information need**:

> Who is the writer responsible for the line “Where there’s life, there’s threat.” and is this a play on the Roman “Where there’s life there’shope”?

Representation of Documents

A document is an ordered sequence of words (i.e. a **time series**).

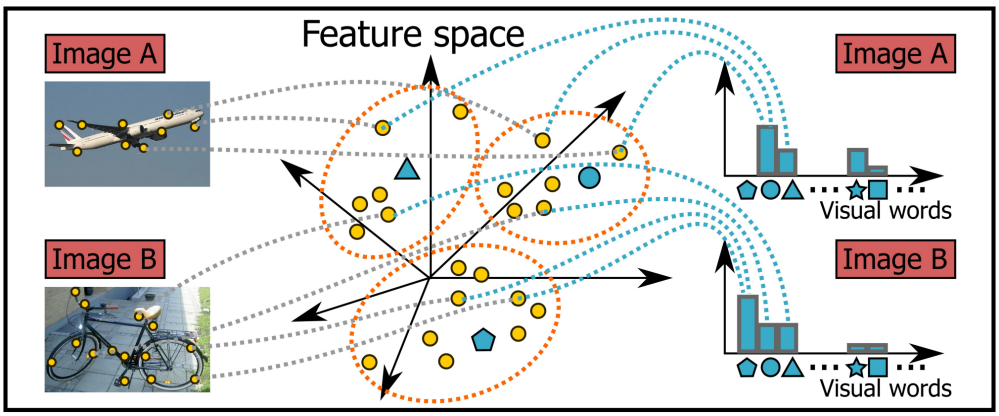
How to represent documents?

> **Bag of words**: list of unique words in a document, does not capture meaning, purely statistical, yet works surprisingly well

"The brown dog and the black dog jumped over the brown fox"

Word	Frequency
and	1
brown	2
dog	2
fox	1
jumped	1
over	1
the	3

Bag of features: same approach used for image and audio search, geometric and temporal information is discarded



Representation of User Need: Query

- Explicit
  - "best restaurant"
- Implicit – context
  - Location (local restaurants)
  - Time (breakfast, lunch, dinner)
  - User profile (John likes French cuisine)

Relevance

Relevance is the "correspondence" between information needs (queries) and information items (documents).

However, the exact meaning of relevance depends on applications:

- useful
- topical
- interesting
- ...

Predicting relevance is the central goal of inforamtion retrieval.

Predicting Relevance

What information can be used to predict relevance?

- Query
- Query context
- User profile
- Content-dependent information
- Content-independent information

Retrieval Models

A retrieval model:

- abstracts away from the real world
- is a mathematical representation of the essential aspects of a retrieval system
- aims at computing relevance and retrieving relevant documents
- thus, either explicitly or implicitly, defines relevance

Evaluation

1. **Precision**: The proportion of retrieved documents that are relevant

$$precision = \frac{\text{number of retrieved docs that are relevant}}{\text{number of retrieved docs}}$$

2. **Recall**: The proportion of relevant documents actually retrieved

$$recall = \frac{\text{number of retrieved docs that are relevant}}{\text{total number of relevant docs}}$$