LC029 정보검색

Information Retrieval

- Information Retrieval (IR) is **finding material** (usually documents) of an **unstructured** nature (usually text) that satisfies an **information need** from within **large collections** (usually stored on computers).

Information Retrieval Model

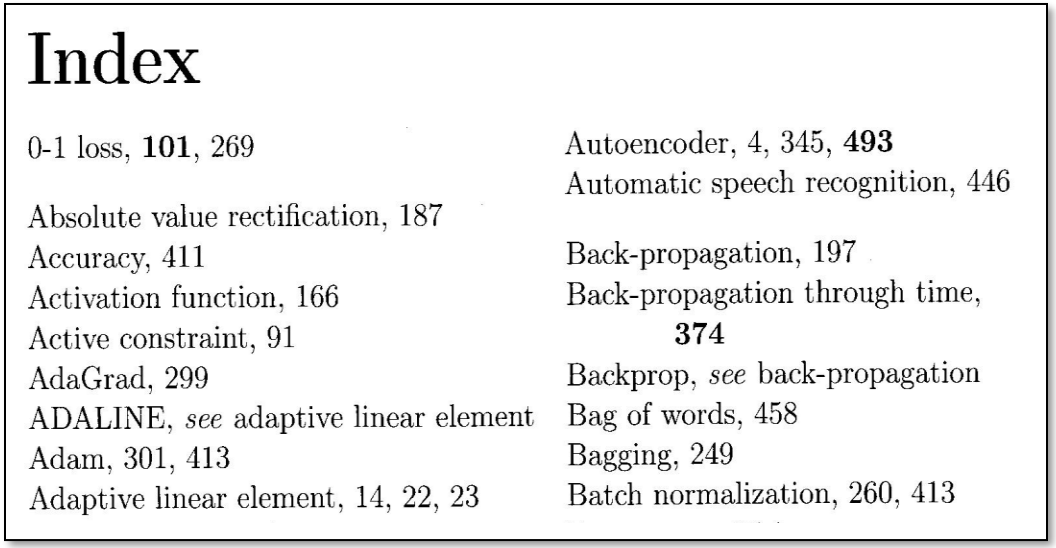
- Boolean Retrieval Model
- Vector Space Model

Boolean Retrieval Model

- Boolean Retrieval
- The Term Vocabulary and Postings Lists
- Dictionaries and Tolerant Retrieval
- Index Construction
- Index Compression

Boolean Retrieval Model

- Boolean Retrieval



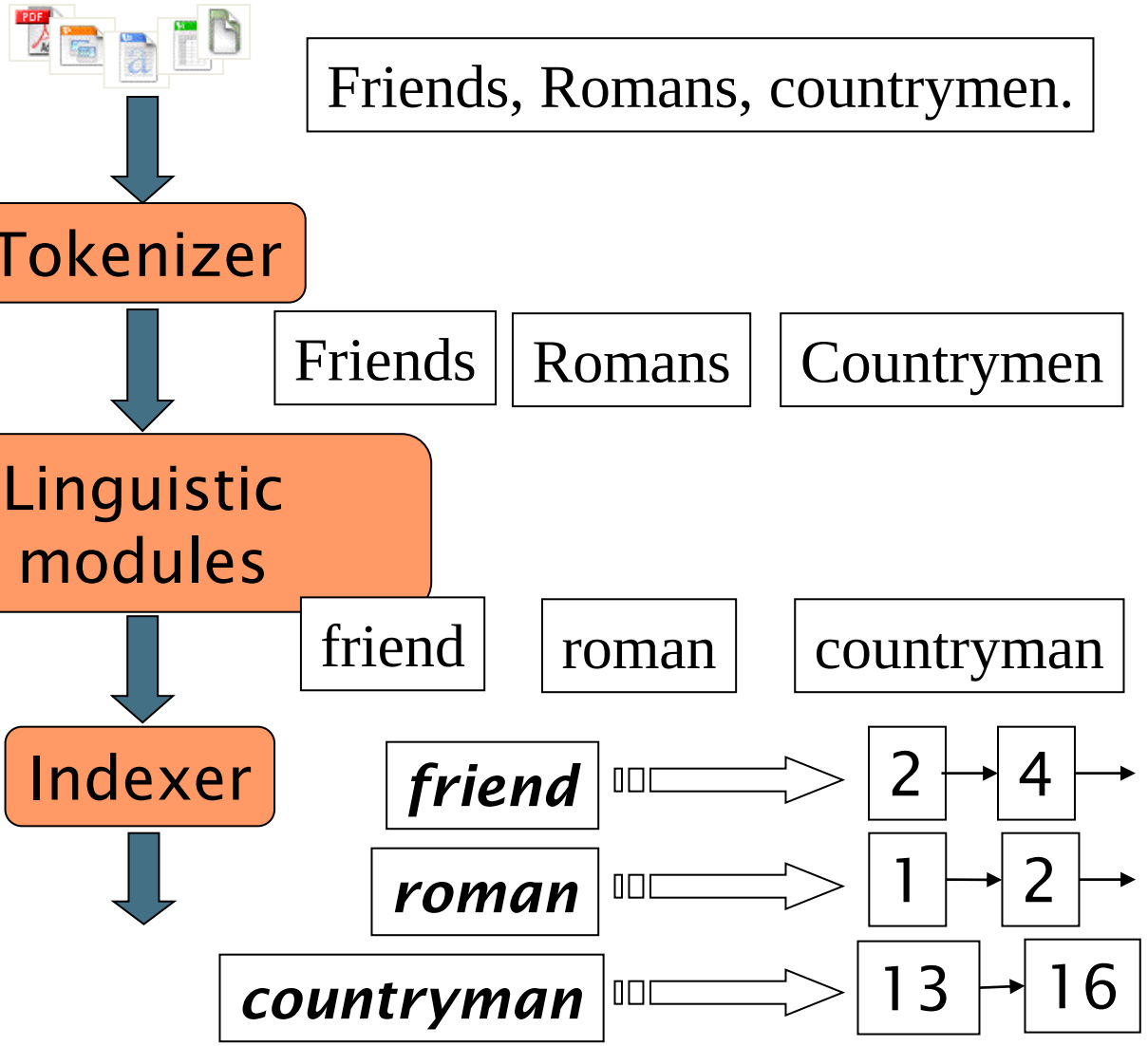
Index

- The Term Vocabulary and Postings Lists
- Dictionaries and Tolerant Retrieval
- Index Construction
- Index Compression

Boolean Retrieval Model

- Boolean Retrieval
- The Term Vocabulary and Postings Lists
  - Tokenization
  - Stop Words
  - Normalization
- Dictionaries and Tolerant Retrieval
- Index Construction
- Index Compression

Tokenization



Stop Words

- Most frequent words considered insignificant
- Significantly reduces the size of index

Normalization

query

USA

car

document

In **U.S.A.** bla bla …

Best **SUV** in the world bla bla …

Boolean Retrieval Model

- Boolean Retrieval
- The Term Vocabulary and Postings Lists
- Dictionaries and Tolerant Retrieval
  - Dictionary Structures
  - Wild-card queries
  - Spelling correction
- Index Construction
- Index Compression

Dictionary Structures

- How do we store a dictionary in memory efficiently, so that we could quickly look up elements at query processing time?
  - Hash table
  - Tree

Wild-card queries

- When you are uncertain of the spelling

  Sydney vs. Sidney ? → **S*dney**

- When you are aware that the term has variants of spelling

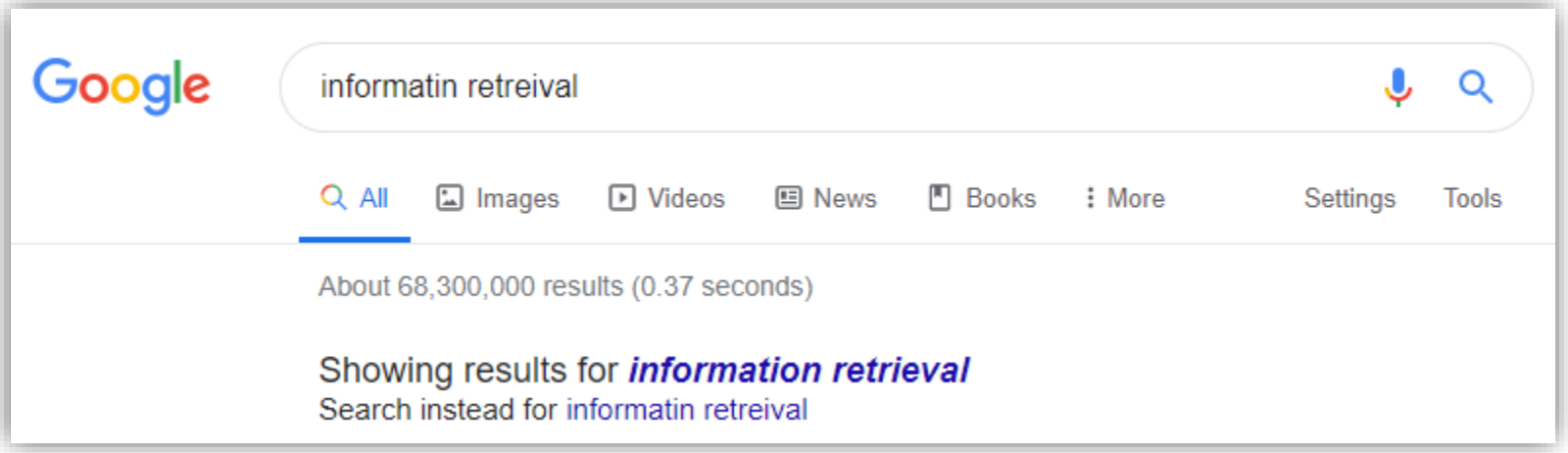
  color vs. colour → **colo*r**

- When you want documents containing variants of a term

  computation, computing, computer, computational, …
  → **comput***

Spelling correction

- Correcting documents being indexed
- Correcting user queries to retrieve “right” answers



Boolean Retrieval Model

- Boolean Retrieval
- The Term Vocabulary and Postings Lists
- Dictionaries and Tolerant Retrieval
- Index Construction
  - Sort-based Index Construction
  - Scalable Index Construction
    - BSBI: Blocked Sort-Based Indexing
    - SPIMI: Single-Pass In-Memory Indexing
    - Distributed Indexing
- Index Compression

Index Construction

symbol	statistic	value
N	documents	800,000
$\mathbf{L_{ave}}$	average number of tokens / doc	200
M	number of terms	400,000
	average number of bytes / token (including spaces/punctuation)	6
	average number of bytes / token (without spaces/punctuation)	4.5
	average number of bytes / term	7.5
	number of non-positional postings	100,000,000

Boolean Retrieval Model

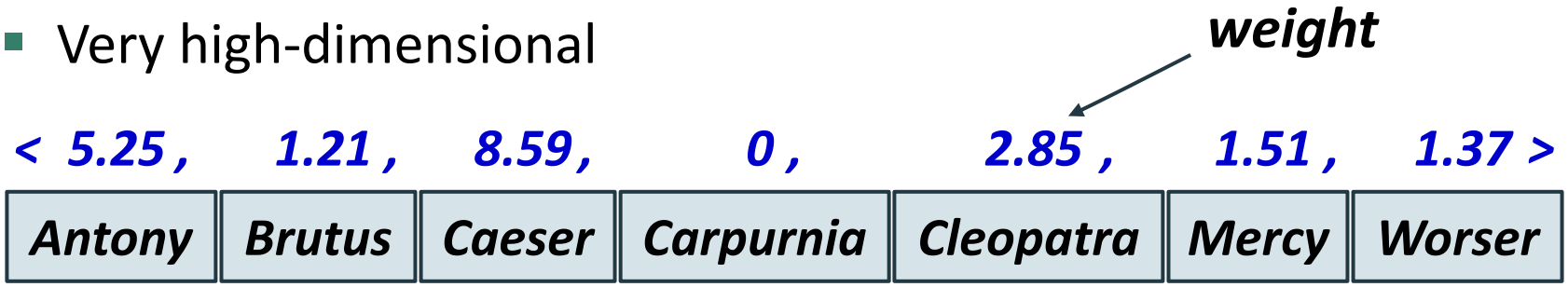
- Boolean Retrieval
- The Term Vocabulary and Postings Lists
- Dictionaries and Tolerant Retrieval
- Index Construction
- Index Compression
  - Dictionary Compression
  - Postings Compression
  - Huffman Code

Vector Space Model

- Why Vector Space Model?
  - Problems with Boolean Search
    - Documents either match or don't
    - Ranking the results is impossible
- How can we rank the documents in the collection with respect to a query?
  - Assign a *score*, say in [0, 1], to each document
  - This score measures how well document and query match

Vector Space Model

- Documents as Vectors
  - Very high-dimensional



- Query as a Vector

< 1 , 0 , 1 , 0 , 0 , 0 , 0 >

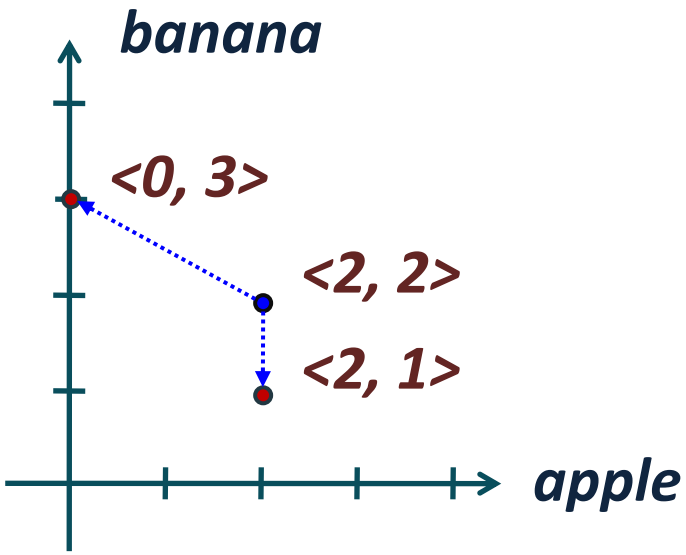
- Compute similarity to the query and then rank documents according to their **similarity** score

Vector Space Model

- How to get the *weight* of each term in a document?
  - Term Frequency
    - Term frequency is used to compute query-document match score
  - Document Frequency
    - Rare terms are more informative than frequent terms
  - *tf-idf* Weight
    - Best known weight scheme in information retrieval
    - Increases with the number of occurrences within a document
    - Increases with the rarity of the term in the collection

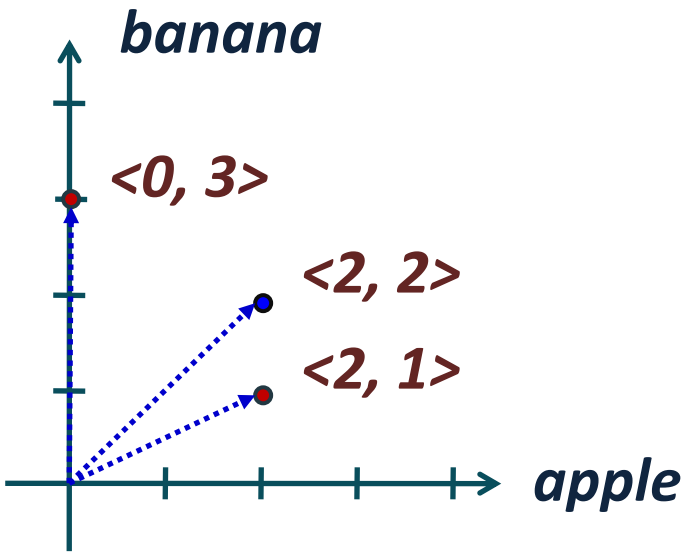
Vector Space Model

- How to get the **similarity** score



Vector Space Model

- How to get the **similarity** score



Cosine similarity

Cosine similarity

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\cdot\vec{d}}{\left|\vec{q}\right|\left|\vec{d}\right|} = \frac{\vec{q}}{\left|\vec{q}\right|}\cdot\frac{\vec{d}}{\left|\vec{d}\right|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

- The following two notions are equivalent
  - Rank documents in **increasing** order of the angle between query and document
  - Rank documents in **decreasing** order of ***cos*(*q*, *d*)**

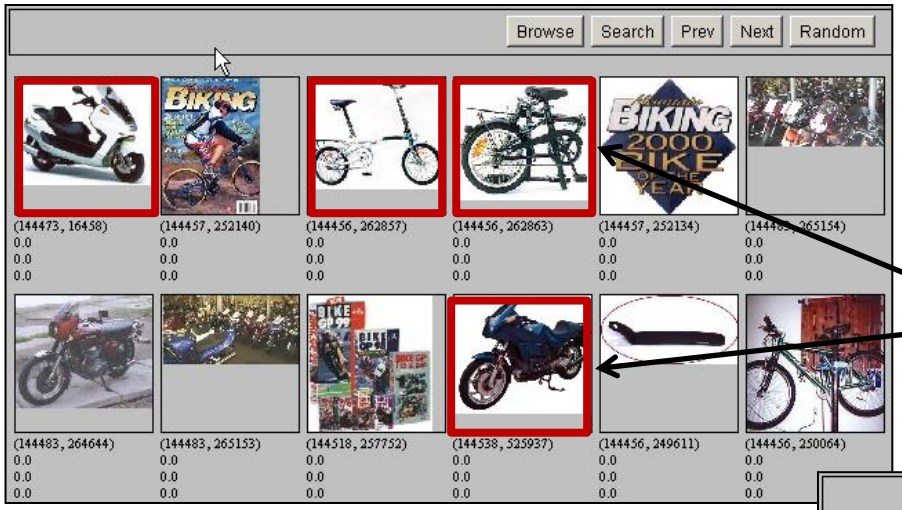
Vector Space Model

- Scoring, Term Weighting and the Vector Space Model
- Computing Scores in a Complete Search System
  - Efficient Cosine Ranking
  - Inexact top *K* document retrieval

More on Information Retrieval

- Evaluation in Information Retrieval
- Relevance Feedback and Query Expansion

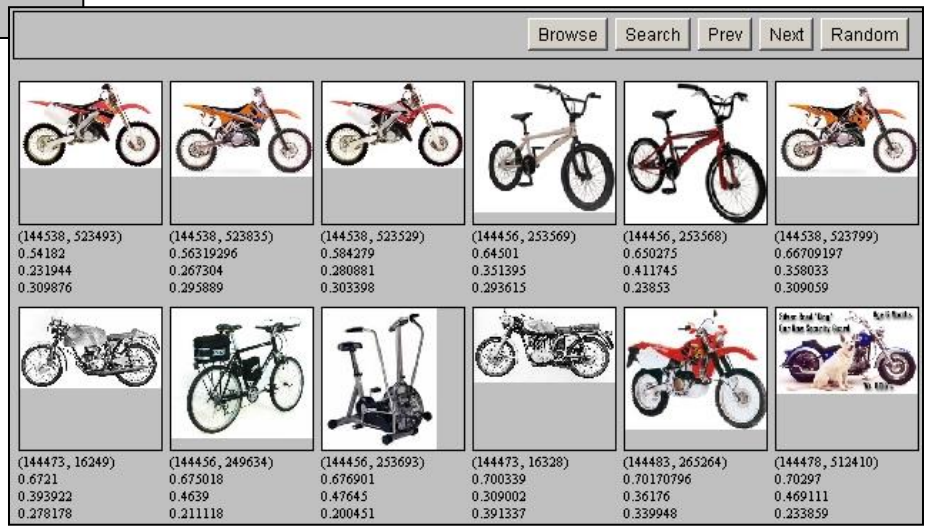
Relevance Feedback : Example 1



← initial search results

user feedback

improved search results →



Relevance Feedback : Example 2

Query: New space satellite applications

retrieved documents

relevance feedback

1. 0.539, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
2. 0.533, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
3. 0.528, 04/04/90, Science Panel Backs NASA Satellite Plan, But Urges Launches of Smaller Probes
4. 0.526, 09/09/91, A NASA Satellite Project Accomplishes Incredible Feat: Staying Within Budget

Improved results

1. 0.513, 07/09/91, NASA Scratches Environment Gear From Satellite Plan
2. 0.500, 08/13/91, NASA Hasn't Scrapped Imaging Spectrometer
3. 0.493, 08/07/89, When the Pentagon Launches a Secret Satellite, Space Sleuths Do Some Spy Work of Their Own
4. 0.493, 07/31/89, NASA Uses 'Warm' Superconductors For Fast Circuit

Relevance Feedback : Example 2

- Expanded query after relevance feedback

2.074	**new**	**15.12**	**space**
30.82	**satellite**	**5.660**	**application**
5.991	nasa	**5.196**	eos
4.196	launch	**3.972**	aster
3.516	instrument	**3.446**	arianespace
3.004	bundespost	**2.806**	ss
2.790	rocket	**2.053**	scientist
2.003	broadcast	**1.172**	earth
0.836	oil	**0.646**	measure

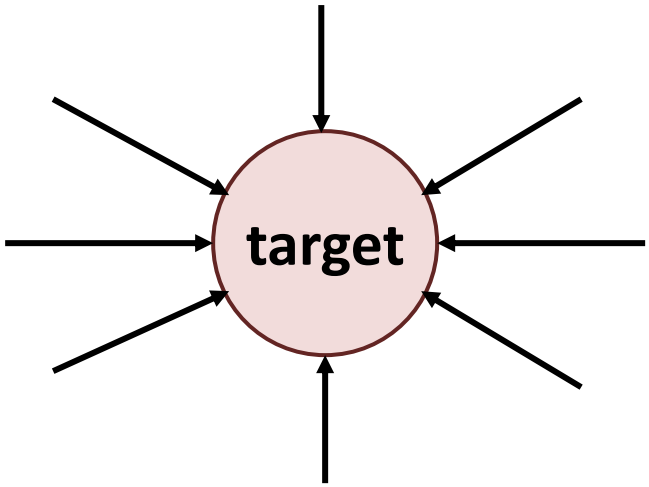
Query: New space satellite applications

More on Information Retrieval

- Web Search Basics
- Web Crawling and Indexes
- Link Analysis

Link Analysis

- Query Processing
  - Retrieve all pages satisfying the query
  - Order retrieved documents by their *link popularity*



- Problem : Link Spam
  - Set up multiple web pages pointing to a target web page