Last time: Vectors as knowledge representation

- Vectors can represent complex knowledge; and real-world entities
  - When vectors are used as KR, reasoning means to apply vector mathematics, including statistics and machine learning.
  - Choices in using vectors as KR include: How to represent an entity as a vector; and how to compute similarity between vectors, classify vectors, group vectors, etc.
- Cosine similarity is a widely used similarity measure between vectors; it corresponds to the cosine of the angle between the vectors. Intuitively, the more the two vectors show in the same direction, the more similar we interpret the two represented entities to be.

WISSEN
TECHNIK
LEIDENSCHAFT

Information Retrieval and Natural Language Processing

Viktoria Pammer-Schindler

Introduction to Data Science and Artificial Intelligence

www.staff.tugraz.at/viktoria.pammer-schindler

Learning Goals

- Define the computational task of information retrieval
- Understand the rationale of using the TFIDF measure for representing query and documents in information retrieval.
- Compute the dictionary vector, term frequency vector, TFIDF vector, and (query-)document similarity for textual documents (without any natural language pre-processing)
- List and describe challenges in natural language processing, in particular as pertaining to the creation of reasonable vector representations of documents in information retrieval.
- Explain imperfections in the above document similarity if no natural language pre-processing is done.

TU Graz

SCIENCE
PASSION
TECHNOLOGY

Information Retrieval

ISDS
INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Information Retrieval Definitions

„Information Retrieval is a field concerned with the structure, analysis, organization, storage, searching, and retrieval of information.“ [Salton 1968]

Finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need, from within large collections (usually stored on computers). [Manning et al., 2008]

Information Retrieval

Given

- Set of documents (items) D
- Query q

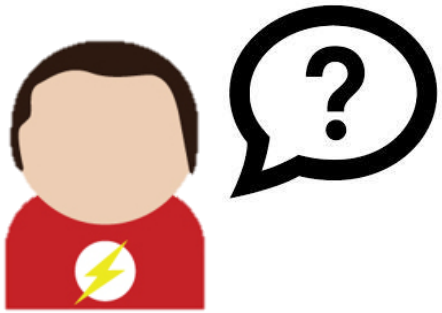
Do:

- Assign a ranking score to each document in D that represents relevance

$$r = f(q,D)$$

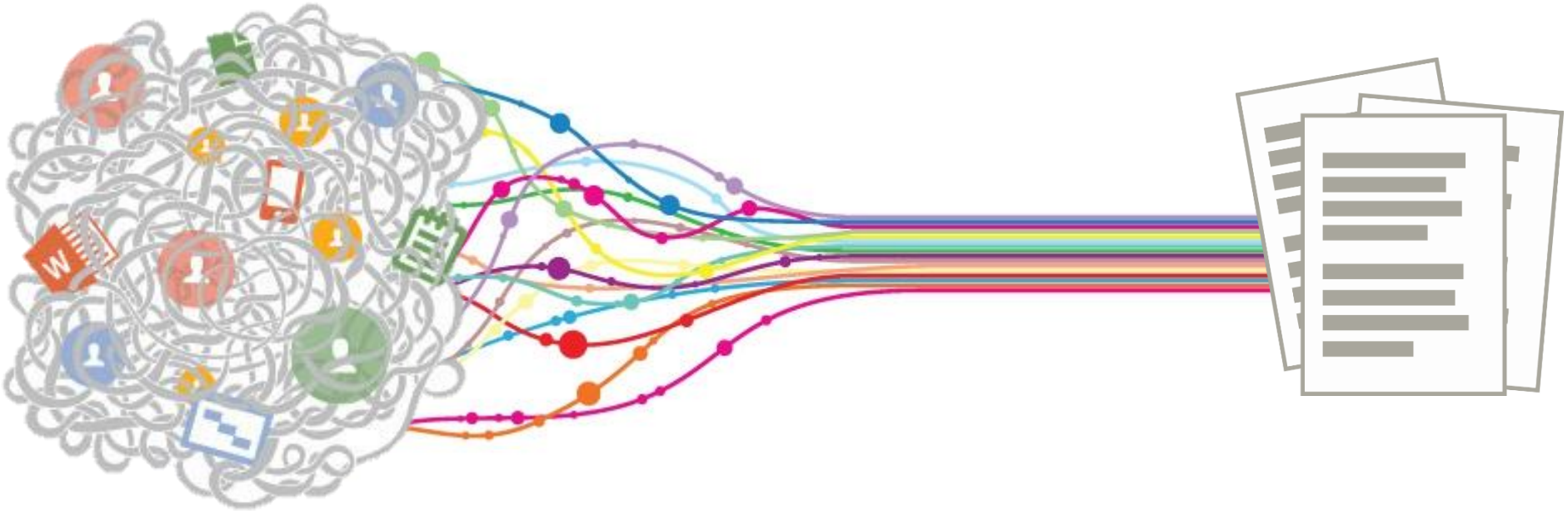
- Decide which of the documents are relevant enough to include in the return set

Information Retrieval

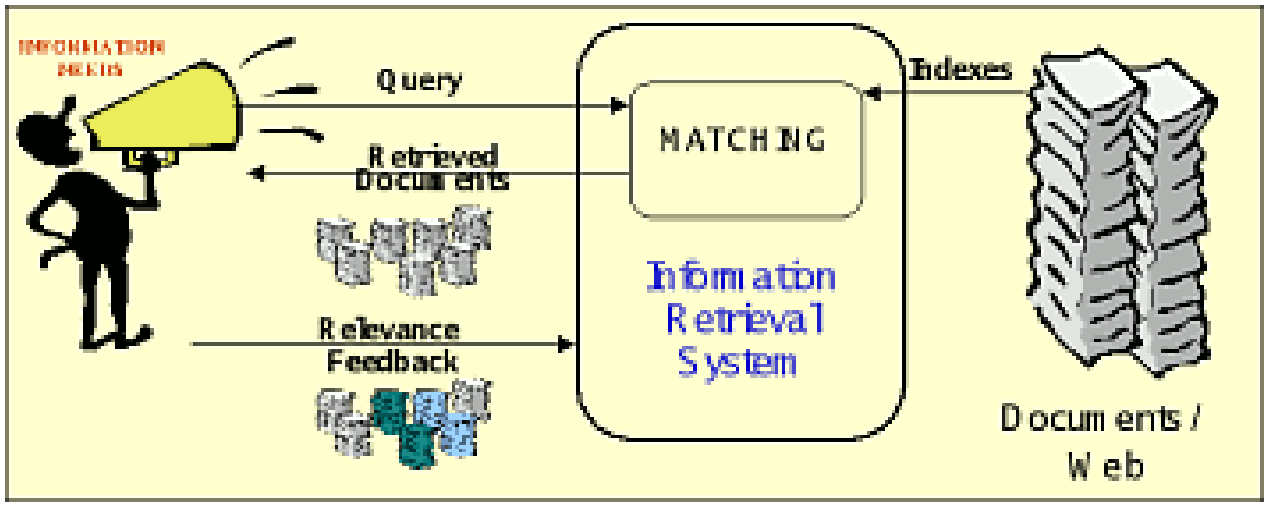


Query

How do I train for climbing?



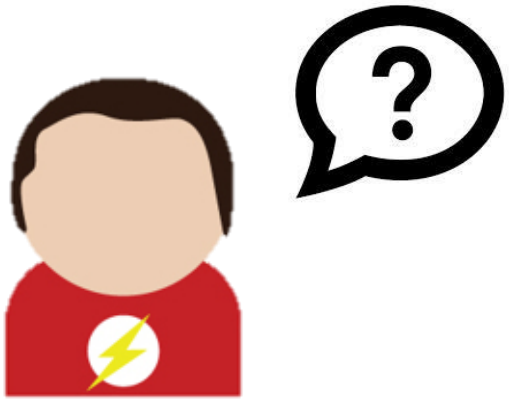
Information Retrieval



Retrieved from: https://commons.wikimedia.org/wiki/File:InformationRetrieval.png

- What are typical information retrieval use cases?
- Have you used information retrieval functionality lately?

Data Retrieval



Query

Select Name from myFriends
where RoomMate = Yes

Name	Birthday	RoomMate
Leonard	12.05.1975	Yes
Raj	22.07.1980	No
Howard	16.10.1983	No
Penny	05.05.1984	No

answer: Leonard

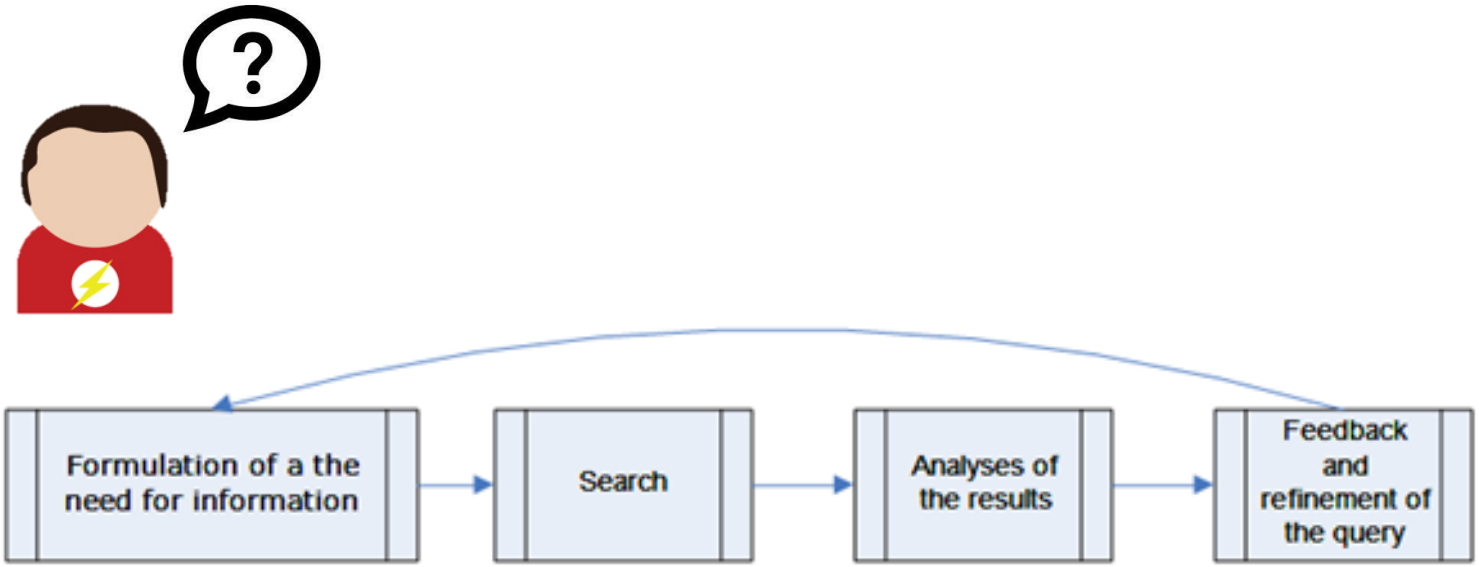
Data Retrieval vs Information Retrieval

	Data Retrieval	Information Retrieval
Content	Structured	Unstructured
Matching	Exact match	Partial match, best match
Model	Deterministic	Probabilistic
Query language	Artificial	Natural
Query specification	Complete	Incomplete
Retrieved objects	Matching query	Relevant
Error response	Sensitive	Insensitive

WISSEN
TECHNIK
LEIDENSCHAFT

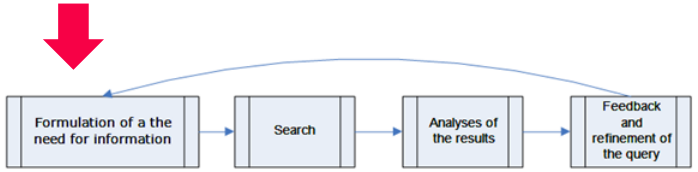
Information Retrieval Process

Information Retrieval Process – high level perspective



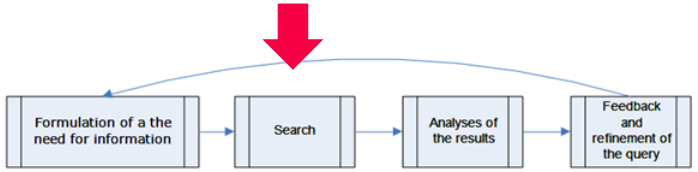
(Manning et al., 2008)

Query Formulation



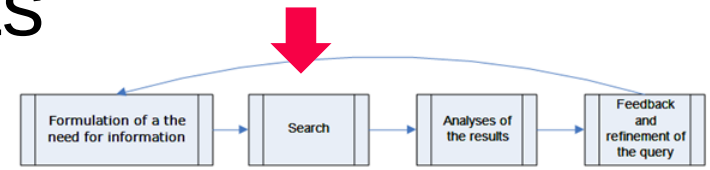
- Different types of queries
  - content-related queries
  - query by example
  - context-extended queries

Search



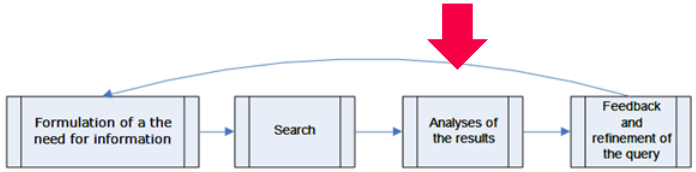
- Find data representations which allows direct comparison of two data objects
  - Text Document : represent documents as term vectors
  - Image : represent images as color schemas
  - Audio File : represent audio tracks as amplitude streams
- Add structured data (Metadata) to the data objects

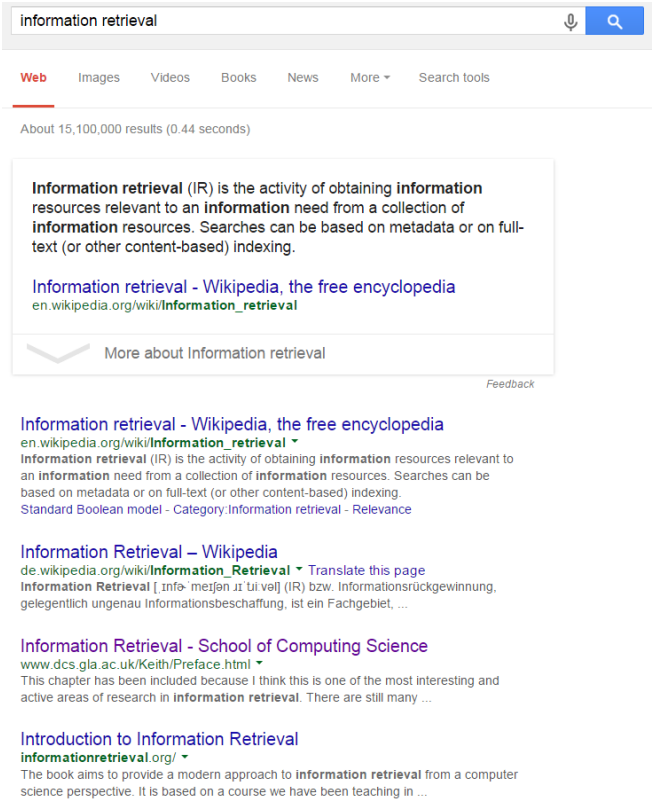
Search – ranking the results



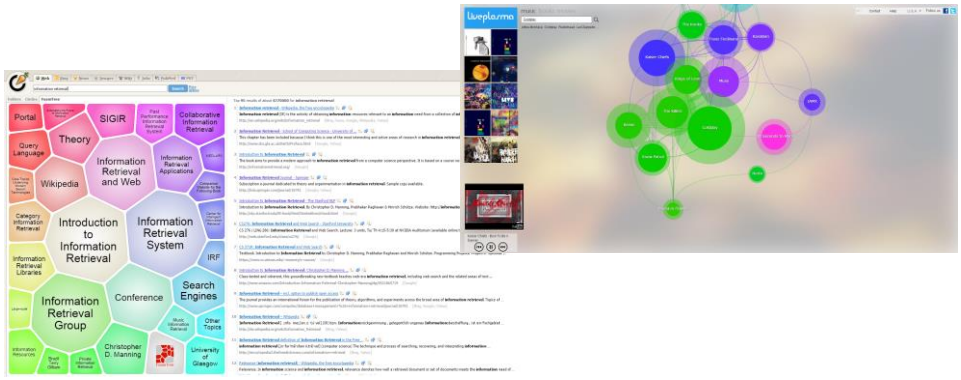
“The fundamental value created by Google is the ranking!” (John Battelle)

Analysis of Search Results

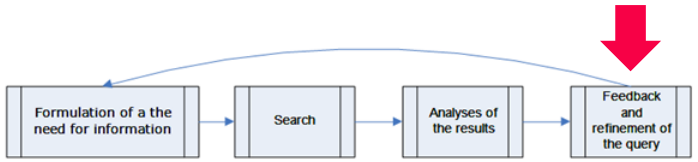




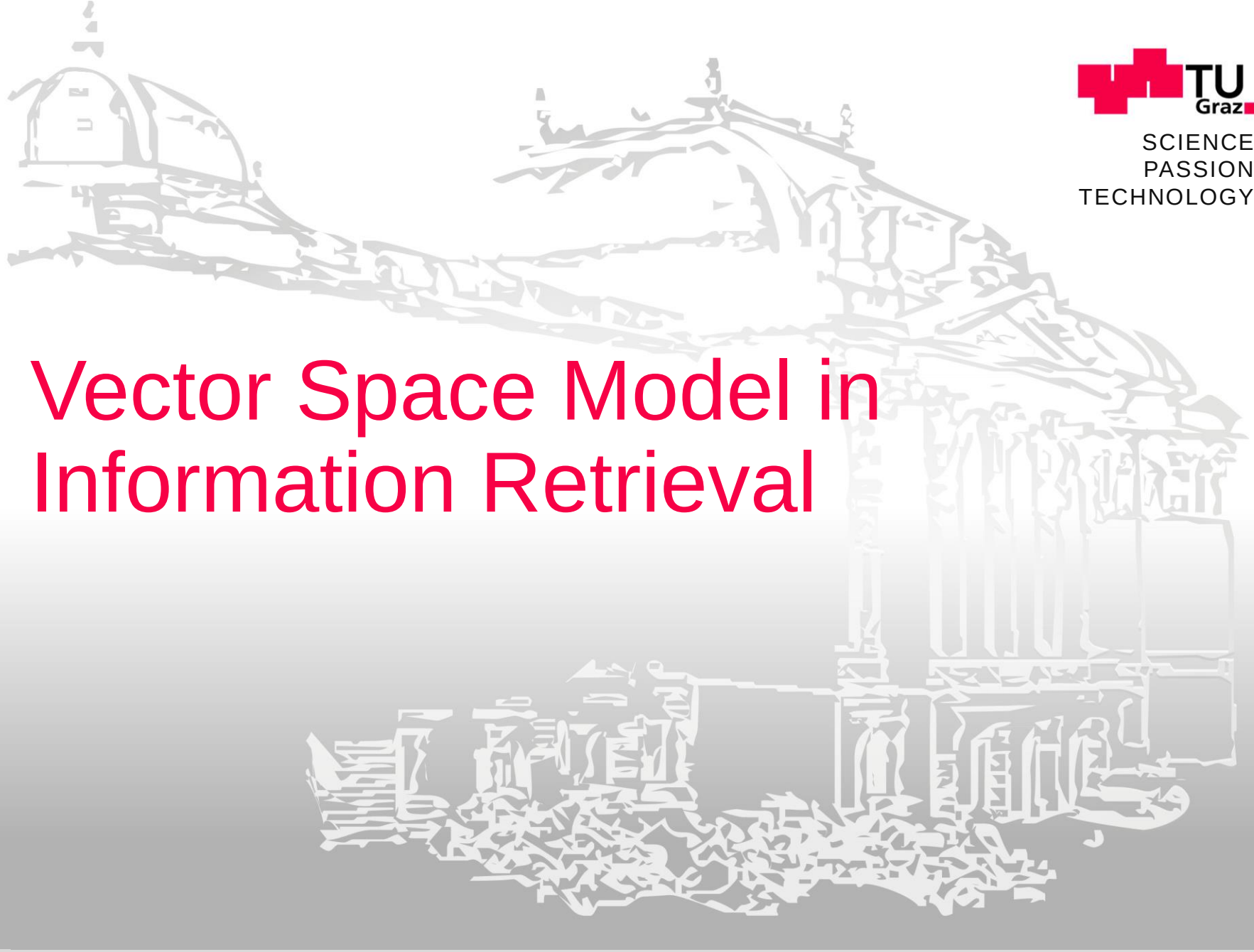
Usage of **visualization techniques** to get a graphical overview of the results



User Feedback



- Information Retrieval is an iterative process
- Refinement of the need of information
  - Approximation of the „optimal“ search
  - Increasing of precision and recall
  - Adaption of the similarity measurement



TU Graz

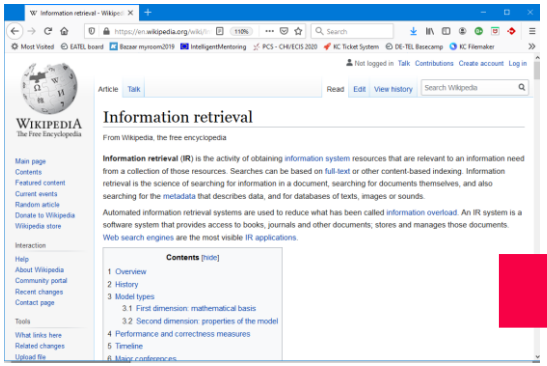
SCIENCE
PASSION
TECHNOLOGY

Vector Space Model in Information Retrieval

ISDS
INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Idea: Vectors as mathematical, non-symbolic representation of ~~complex entities~~ **items**

$$\begin{pmatrix} 0{,}7 \\ 0{,}5 \\ 0{,}3 \\ 0{,}42 \\ 0{,}8 \\ 0{,}33 \\ \dots \end{pmatrix}$$



$$\begin{pmatrix} 0{,}1 \\ 0{,}2 \\ 0{,}4 \\ 0{,}44 \\ 0{,}1 \\ 0.33 \\ \dots \end{pmatrix}$$

Core idea 1 of Vector Space Model in Information Retrieval

Given a set of documents D = (d_1, d_2, …d_n)

Given a query q

Represent everything as vectors

Use cosine similarity to assign a ranking score to each document $d_i \in D$

- R = f(q,D) = d_1, d_2, …d_n: sim(q,d_i)>sim(q,d_{i+1})

Core idea 2 of Vector Space Model in Information Retrieval

- Elements in the vectors should correspond to *content elements* from the documents and the query (words, word-stems, concepts, phrases, …)

➢ A dictionary vector is needed

  - Contains all the selected content elements from the documents and query
  - Assigns each content element an index.
  - Constitutes a CHOICE (which elements are we interested in)

Festival Example – Dictionary Vector

Query q = „*festival in Graz*“

Document d_1 = „*The Elevate festival in Graz is a highly visible, public event that draws international audience*“.

Document d_2 = „*All festivals in Austria, have been cancelled due to the ongoing CoVid-19 situation. This regulation of course also applies to Graz. It is currently not forseeable when public festivals will be able to take place again*.“

The dictionary vector to the right already “hides” a few choices in which words to choose and which not. Can you spot them?

$$\begin{pmatrix} festival \\ graz \\ elevate \\ highly \\ visible \\ public \\ event \\ draws \\ international \\ audience \\ Austria \\ cancelled \\ ongoing \\ CoVid-19 \\ situation \\ regulation \\ applies \\ forseeable \\ take \\ place \\ again \end{pmatrix}$$

Remaining question (for NOW): Which value to use for corresponding index in query and document vectors?

Term frequency-Inverse document frequency (TFIDF) – the discriminative value for a term t_i in a corpus D

How often does each term occur in a document d_i? But normalize with document length - Normalised term frequency vector tf_i for document d_i

$$tf_i = \frac{\#occurrences\ of\ term\ in\ d_i}{\#terms\ in\ d_i}$$

How well can each term differentiate between documents? - Inverse document frequency vector t_i in corpus D: $idf = log_{10} \frac{|D|}{\#documents\ with\ term\ in\ it}$

TFIDF: $tfidf_i = tf_i \circ idf$
(element-wise multiplication of vectors t_i and idf

TFIDF Example

Document 1: „I love sun!“
Document 2: „I hate sun!“
Document 3: „I love rain!“
Query: „Does someone else love the sun?“

Step 1: Dictionary vector – Remember it‘s a choice (and the below one is more basic than the choice made for the festival example on slide 13)

$$dict = \begin{pmatrix} i \\ love \\ hate \\ sun \\ rain \end{pmatrix}$$

TFIDF Example continued

Document 1: „I love sun!“

Document 2: „I hate sun!“

Document 3: „I love rain!“

Query: „Does someone else love the sun?“

$$dict = \begin{pmatrix} i \\ love \\ hate \\ sun \\ rain \end{pmatrix}$$

Step 2: Inverse document frequency

$$idf = \begin{pmatrix} 0 \\ 0.18 \\ 0.48 \\ 0.18 \\ 0.48 \end{pmatrix}$$

Step 3: normalised term frequency vectors for document and query

$$t_1 = \begin{pmatrix} 0.33 \\ 0.33 \\ 0 \\ 0.33 \\ 0 \end{pmatrix} \quad t_2 = \begin{pmatrix} 0.33 \\ 0 \\ 0.33 \\ 0.33 \\ 0 \end{pmatrix}$$

$$t_3 = \begin{pmatrix} 0.33 \\ 0.33 \\ 0 \\ 0 \\ 0.33 \end{pmatrix} \quad q_t = \begin{pmatrix} 0 \\ 0.17 \\ 0 \\ 0.17 \\ 0 \end{pmatrix}$$

TFIDF Example continued

Step 4: TFIDF vectors for document and query

$$idf = \begin{pmatrix} 0 \\ 0.18 \\ 0.48 \\ 0.18 \\ 0.48 \end{pmatrix}$$

$$tfidf_1 = \begin{pmatrix} 0 \\ 0.0594 \\ 0 \\ 0.0594 \\ 0 \end{pmatrix} \quad tfidf_2 = \begin{pmatrix} 0 \\ 0 \\ 0.1584 \\ 0.0594 \\ 0 \end{pmatrix} \quad tfidf_3 = \begin{pmatrix} 0 \\ 0.0594 \\ 0 \\ 0 \\ 0.1584 \end{pmatrix} \quad q_{tfidf} = \begin{pmatrix} 0 \\ 0.0306 \\ 0 \\ 0.0306 \\ 0 \end{pmatrix}$$

Vector space model is a ranked information retrieval model

… because results are ranked, in the pure vector space model based on (cosine) similarity between query and document vectors.

TU Graz

SCIENCE
PASSION
TECHNOLOGY

Natural Language Processing in Information Retrieval

ISDS
INSTITUTE OF INTERACTIVE SYSTEMS AND DATA SCIENCE

Natural Language Processing

Is a field at the intersection of
- Linguistics
- Computer science
- Electrical engineering (speech processing)

That studies how computers can understand natural (=human) language.

Broad challenges are:
- Speech processing
- Natural language understanding
- Natural language generation

NLP is language-depending – most progress has been made for English.

Natural Language Processing in Information Retrieval

Above, we have skimmed inelegantly over how to create the dictionary vector.

Typically, the following would be pre-processing steps in order to create the dictionariy vector, and before computing the TFIDF vectors for documents, and the query (or other item representations):

- Remove unwanted characters and markup (e.g., HTML tags, punctuation)
- Break into tokens (e.g., on whitespace)

- Detect common phrases (e.g., using a dictionary)
- Remove common words (a, an, and, the, it, …)
- Stem tokens to word roots (e.g., computational -> comput)

Challenges

- Consider semantic information (word sense disambiguation, synonyms)
- Consider structural syntactic information (phrases, word order, proximity of words)

More advanced natural language processing techniques useful for Information Retrieval

Word sense disambiguation – determine the sense of an ambiguous word based on the context.

- Examples: Apple (fruit or company?), Jaguar (animal or car brand?)

Information extraction: Identify specific pieces of information in a document

- Examples: Well known persons or concepts (CoVid-19, Gesundheitsminister Rudi Anschober, Wuhan, …)

Question answering – return answers instead of documents that (ideally) contain the answer.

SCIENCE
PASSION
TECHNOLOGY

Discussion

Discussion

- **Imperfection:** Queries imperfectly represent the information need of a user
- **Relevance:** The relevance of items is imperfectly described *only* by content-related characteristics (~ vector similarity). Additionally important are:
  - Being timely (recent information)
  - Being authoritative (from a trusted source, or otherwise plausible)
  - Sometimes location is a good differentiator
  - ➢ Such characteristics can be considered as weight in the ranking function - in this lecture we have used as ranking function $R = f(q,D) = d_1, d_2, \ldots d_n: sim(q,d_i) > sim(q,d_{i+1})$

Recommended Reading

- Information Retrieval and Web Search – Course Materials from Raymond J. Mooney, Univ. of Texas, US
  - https://www.cs.utexas.edu/~mooney/ir-course - Units 1 (Intro) and 2 (Vector Space Retrieval Models)

Exercise 10

From slide 16 – which document is more similar to the query, using cosine similarity?

Hints

- The exercise is asking you to compute sim(q_t,t_1) and sim(q_t,t_2) using the cosine similarity
- … and to find out which of the two values is the larger one.

$$sim(a,b) = \cos(\theta) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$

Exercise 11

TFIDF Example continued – cosine similarity: Which document matches best the given query?

Take the vectors from slide 22

Hint: Do the calculations, but before you do them, have a close look at the vectors, and try to understand or remember how the cosine similarity works – what do you notice?

$$sim(a,b) = \cos(\theta) = \frac{a \cdot b}{\|a\| \cdot \|b\|}$$