Text Analysis and Retrieval

3. Basics of Information Retrieval

Prof. Jan Šnajder

University of Zagreb
Faculty of Electrical Engineering and Computing (FER)

Academic Year 2022/2023

 v3.0

Intro

Reminder: What is IR?

Information retrieval *(Wikipedia)*

The activity of obtaining information resources relevant to a user's information need from a collection of information resources.

- **Information needs** (expressed by users in the form of queries)
- **Information resources** (typically unstructured – text, images, video, audio, etc.)

Information needs

Information need

Information need is an individual or group's desire to locate and obtain information to satisfy a conscious or unconscious need. Needs and interests call forth information.
(Robert S. Taylor: "The process of asking questions", 2007)

- (Un)conscious needs for information are expressed via queries
  - In text retrieval: words and phrases
    (e.g., "ISIS attacks", "coronavirus pandemic")
  - In image content retrieval: images

Information retrieval problem

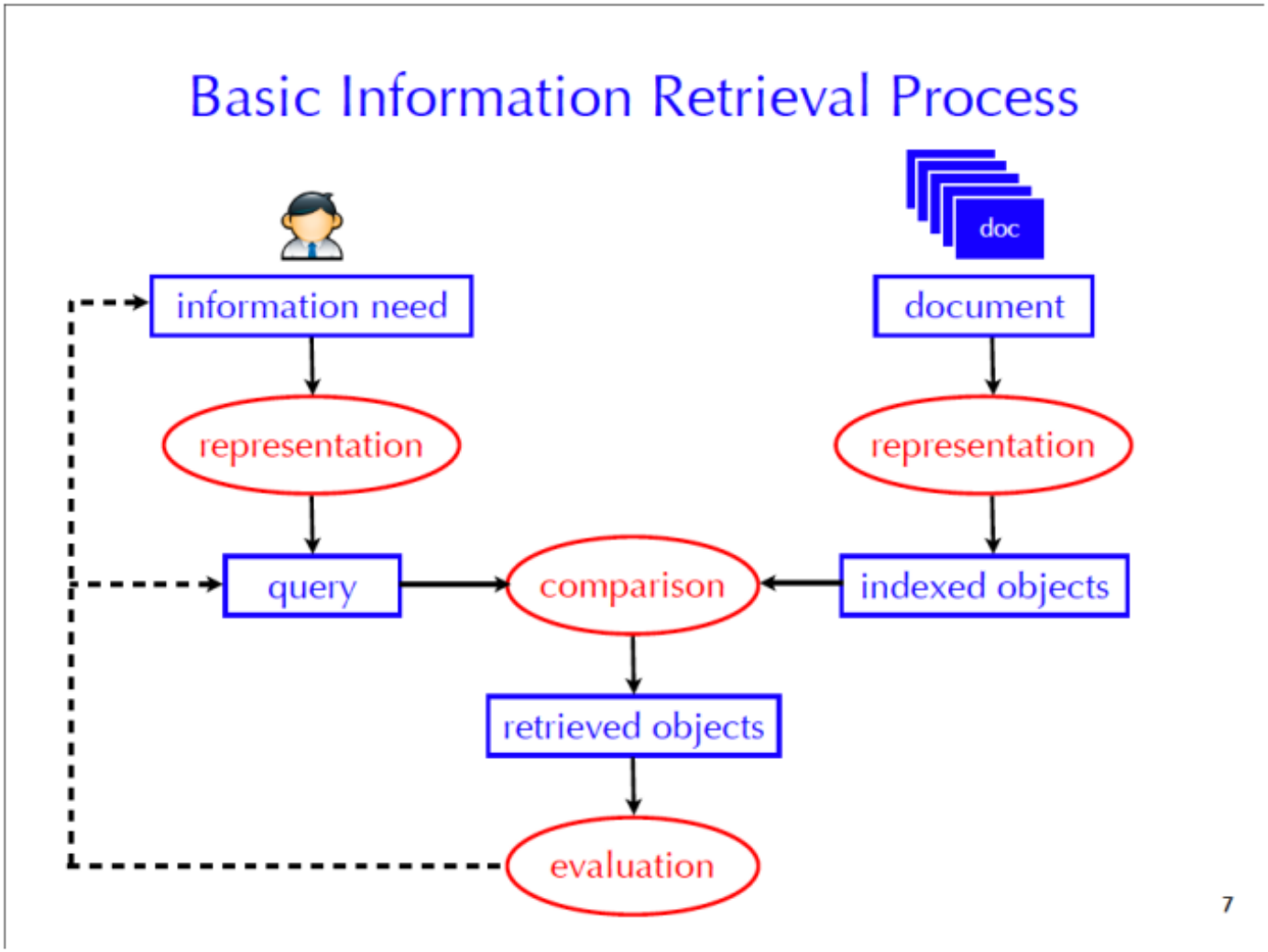


Diagram from Jaime Arguello, UNC-Chapel Hill

Unstructured document representation

Document snippet

One evening Frodo and Sam were walking together in the cool twilight. Both of them felt restless again. On Frodo suddenly the shadow of parting had falling: the time to leave Lothlorien was near.

Bag of words

{(One, 1), (evening, 1), (Frodo, 2), (and, 2), (Sam, 1) (were, 1), (walking, 1), (together, 1), (in, 1), (the, 3), (cool, 1), (twilight, 1), (Both, 1), (of, 2), (them, 1), (felt, 1), (restless, 1), (again, 1), (On, 1), (suddenly, 1), (shadow, 1), (parting, 1), (had, 1), (falling, 1), (time, 1), (to, 1), (leave, 1), (Lothlorien, 1), (was, 1), (near,1)}

Weakly-structured document representation

Document snippet

One evening Frodo and Sam were walking together in the cool twilight. Both of them felt restless again. On Frodo suddenly the shadow of parting had falling: the time to leave Lothlorien was near.

Bag of nouns

{(evening, 1), (Frodo, 2), (Sam, 1), (twilight, 1), (shadow, 1), (parting, 1), (time, 1), (Lothlorien, 1)}

Bag of named entity terms

{(Frodo, 2), (Sam, 1), (Lothlorien, 1)}

Preprocessing

1. **Morphological normalization (stemming/lemmatization)**
   - Conflating various forms of the same word to a common form
   - Important for morphologically rich languages such as Croatian
   - Stemming (e.g., kućom → kuć) more often used than lemmatization (e.g., kućom → kuća)
2. **Stop words removal**
   - Stop words: semantically void terms such as determiners, prepositions, conjunctions, pronouns, etc.
   - Content words: nouns, verbs, adjectives, adverbs
   - Stop words removal: removes stop words and retains only the content words
   - English stop words: `https://www.ranks.nl/stopwords`

⇒ both methods reduce the size of the bag-of-words representation and generally improve retrieval performance

Three components of an IR model

The **basic retrieval model** is a triple (f_d, f_q, r):

1. f_d is a function that maps documents to retrieval representations

$$f_d(d) = x_d$$

2. f_q is a function that maps queries to retrieval representations

$$f_q(q) = x_q$$

3. r is a ranking function that produces a relevance score: a real-valued number that indicates the potential relevance of the document d for query q based on x_d and x_q

$$relevance(d, q) = r\big(f_d(d), f_q(q)\big) = r(x_d, x_q)$$

IR paradigms

- Information retrieval models roughly fall into three paradigms:
  1. **Set-theoretic models**
     - Boolean model
     - Extended Boolean model
  2. **Algebraic models**
     - Vector space model
     - Latent semantic indexing
  3. **Probabilistic models**
     - BM25
     - Language model
- Additionally, there are IR models that utilize **link analysis algorithms** (e.g., PageRank, HITS)

Vector space model

- Documents and queries are represented as vectors of index terms
- Weights are non-negative real numbers

$$\mathbf{d_j} = [w_{1j}, w_{2j}, \ldots, w_{tj}]$$
$$\mathbf{q} = [w_{1q}, w_{2q}, \ldots, w_{tq}]$$

- The relevance of the document for the query is estimated by computing some **distance or similarity metric** between the two vectors
  - Distance metrics – Euclidean, Manhattan, etc.
    - More relevant when distance is lower
  - Similarity metrics – Cosine, Dice, etc.
    - More relevant when similarity is larger

Cosine similarity

Salton, 1983

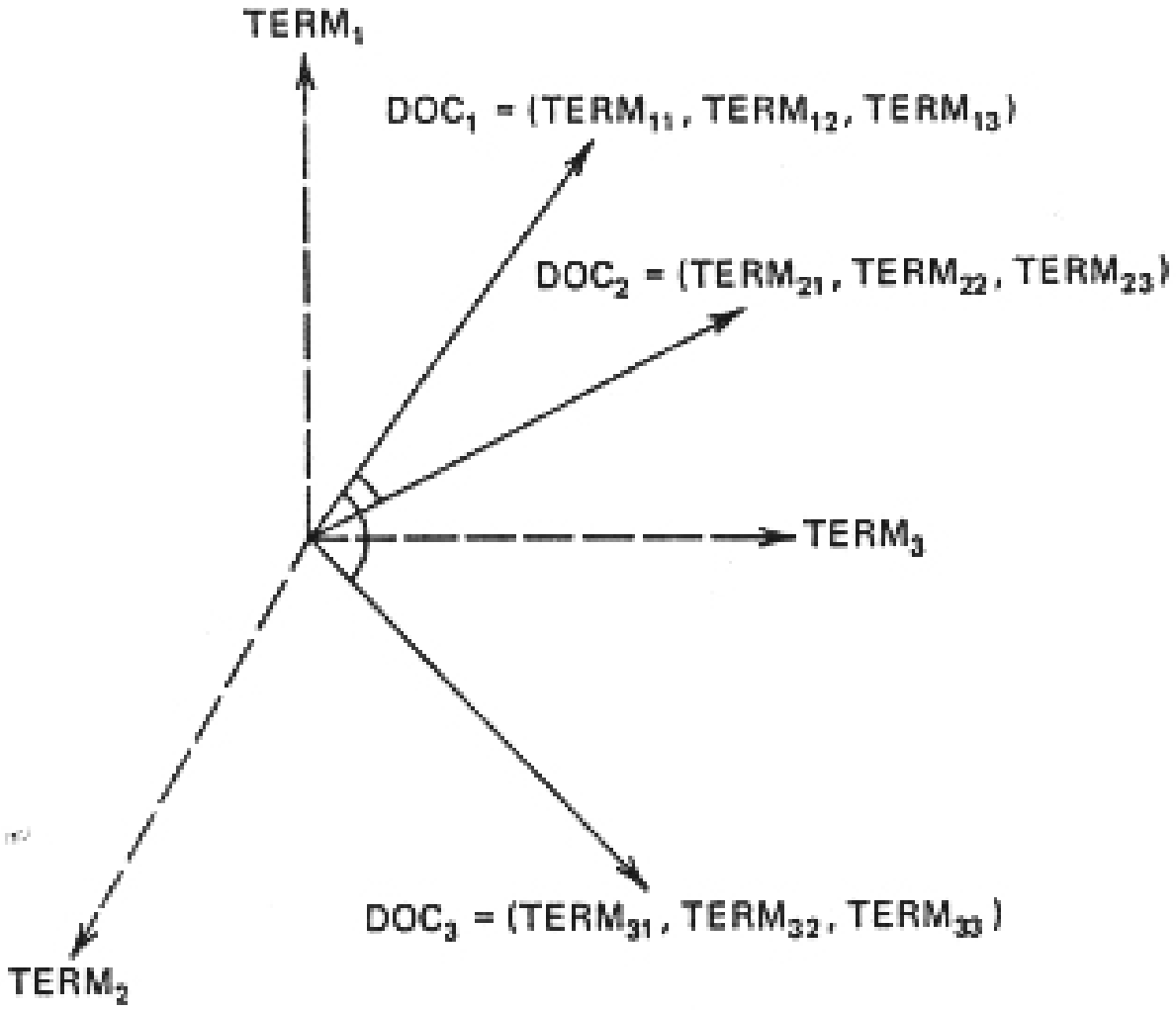


Figure 4-2 Vector representation of document space.

Vector space model – term weighting

- How are the weights w_{ij} of index terms for documents computed?
- Two intuitive assumptions:
  1. The relevance of an index term for the document is proportional to its frequency in the document (**term frequency** component)
     - i.e., more frequent $\Rightarrow$ more relevant
  2. The relevance of an index term for any document is inversely proportional to the number of documents in the collection in which it occurs (**inverse document frequency** component)
     - i.e., more common across documents $\Rightarrow$ less relevant (e.g., stopwords such as "*the*")

TF-IDF weighting scheme

- The weight computed as the product of the term frequency component and the inverse document frequency component

$$w_{ij} = tf(t_i, d_j) \cdot idf(t_i, D)$$

where t_i is the i-th term from the index

- The most popular local and global schemes:

$$tf(t_i, d_j) = 0.5 + \frac{0.5 \cdot freq(t_i, d_j)}{\max_{t \in d_j} freq(t, d_j)}$$

$$idf(t_i, D) = \log \frac{|D|}{|\{d_j \in D \mid t_i \in d_j\}|}$$

Probabilistic retrieval models

- **Probabilistic retrieval models**
  - View retrieval as a problem of **estimating the probability of relevance** given a query, document, collection, etc.
  - Documents are ranked in decreasing order of this probability
- Rely on probability ranking principle

BM25

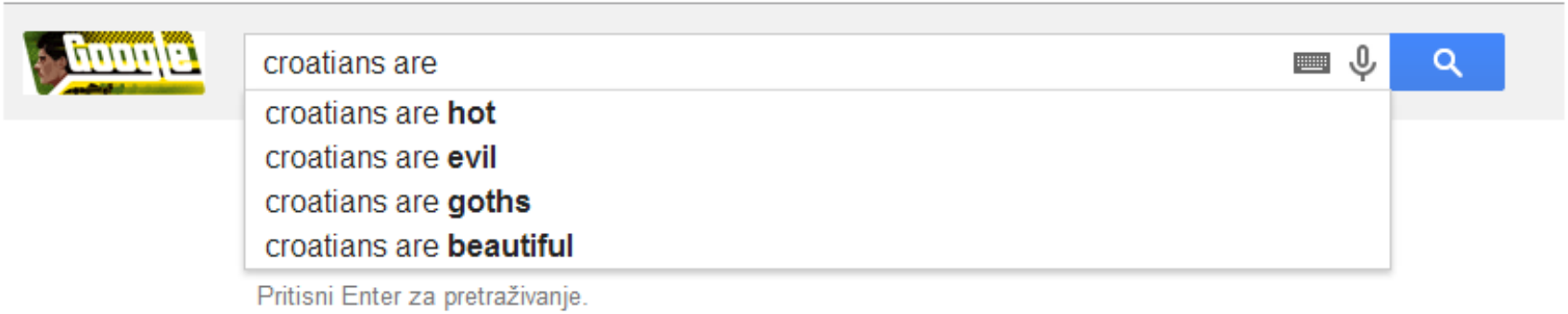
- The famous BM25 ranking function:

$$\sum_{t \in q} \frac{freq(t,d)(k_1+1)}{k_1(1-b)+k_1(l_d/l_{avg})b+freq(t,d)} \cdot idf(t,d)$$

- Often yields SOTA results!

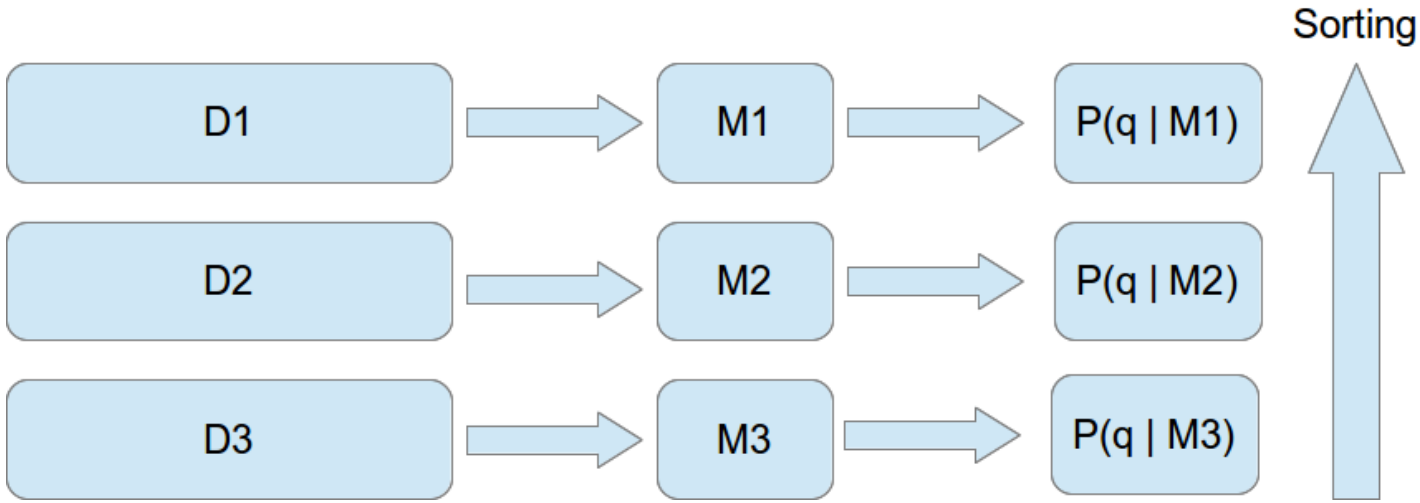
Language modeling for IR

- Approaching the probabilistic information retrieval problem from a different perspective
- Instead of modeling document probability given the query, we model the **query probability given the document**



Query likelihood model

- Given a document collection D and a query q
- A language model M_d is built for **each document**
- Documents are scored according to the probability $P(q|M_d)$



- **Intuition:** language models corresponding to relevant documents should assign higher probability to the query

IR Tools

- **Terrier IR platform**
  - Open command-line IR platform
  - Popular in academia – used for IR evaluations in research
  - Stepwise usage – first indexing, then retrieval, and then evaluation
- **Lucene**
  - Open source information retrieval library written in Java (ports to many languages exist)
  - Describes a document as a set of (user definable) fields
  - Very flexible solution for applications requiring full text indexing and search
- **Elasticsearch** – RESTful search engine on top of Lucene
  - Docs: https://www.elastic.co/guide/index.html
  - Very limited NLP functionality:
    https://www.elastic.co/blog/text-classification-made-easy-with-elasticsearch

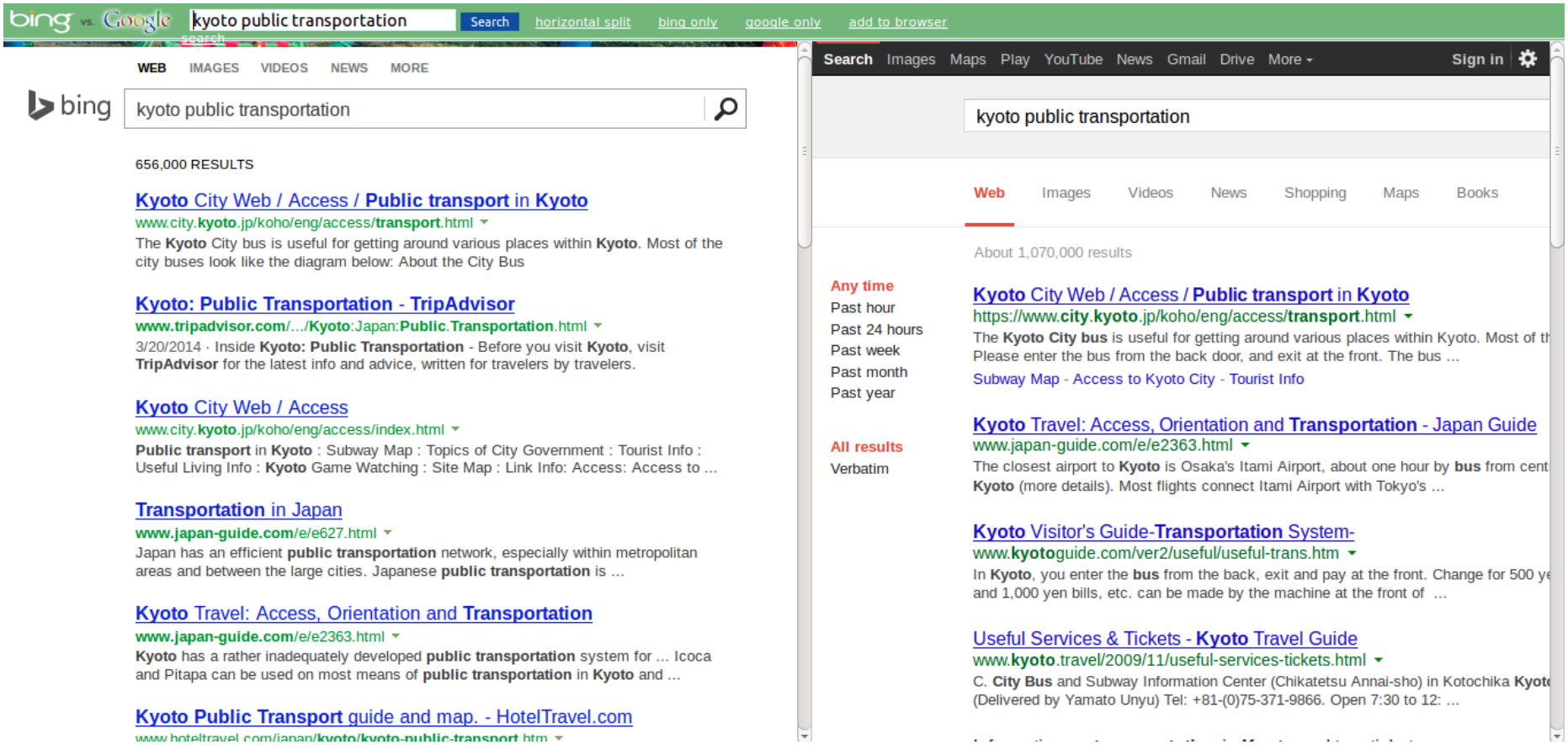
Learning outcomes 1

1. List three main components of an IR model
2. Describe the vector space model and the TF-IDF weighting scheme
3. Explain the probability ranking principle and BM25
4. Describe the LM information retrieval model
5. List the main IR tools available

Intro

IR evaluation

Clash of the giants: Which one is better?



IR test collection

- IR test collection is comprised of:
  1. Document collection
  2. Set of information needs (descriptions + queries)
     - At least 50 information needs
  3. Set of relevance judgments for each query–document pair
     - Binary relevance (document is relevant or not relevant) or graded relevance judgments (less common)
- Used for:
  - Evaluating retrieval effectiveness w.r.t. different settings (stemming, ranking model, etc.)
  - Comparing against other systems, typically in evaluation campaigns
  - Fine-tuning of system parameters, done on a development test collection (to prevent overfitting)

IR test collections – topic examples

TREC topic 351

Title: Falkland petroleum exploration
Description: What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?
Narrative: Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.

TREC topic 409

Title: Legal, Pan Am, 103
Description: What legal sanctions have resulted from the destruction of Pan Am Flight 103 over Lockerbie, Scotland, on December 21, 1988?
Narrative: Documents describing any charges, claims, or fines presented to or imposed by any court or tribunal are relevant, but documents that discuss charges made in diplomatic jousting are not relevant.

Relevance judgments ("qrel file")

```
401 0 FBIS3-18916 0
401 0 FBIS3-18926 0
401 0 FBIS3-18943 1
401 0 FBIS3-18946 0
401 0 FBIS3-18972 0
401 0 FBIS3-18997 0
401 0 FBIS3-19003 0
401 0 FBIS3-19032 1
401 0 FBIS3-19037 0
401 0 FBIS3-19038 1
401 0 FBIS3-19042 0
401 0 FBIS3-19080 0
401 0 FBIS3-19103 0
401 0 FBIS3-19107 1
401 0 FBIS3-19110 0
401 0 FBIS3-19126 0
401 0 FBIS3-19133 0
401 0 FBIS3-19212 0
401 0 FBIS3-19213 0
401 0 FBIS3-19251 0
401 0 FBIS3-19290 0
401 0 FBIS3-19302 0
401 0 FBIS3-19303 0
401 0 FBIS3-19304 0
```

Evaluation metrics

- Compare retrieved documents against relevant documents
- Each document is either retrieved or not, and either relevant or not. This gives a 2×2 confusion matrix:

$$\begin{array}{lcc} & \text{relevant} & \text{not relevant} \\ \text{retrieved} & tp & fp \\ \text{not retrieved} & fn & tn \end{array}$$

We could compute accuracy as the fraction of correct decisions:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}$$

- **Q:** Why using accuracy is not a good idea?
  A: Given a query, most documents (say 99%) are irrelevant. A search engine that retrieves nothing will already be 99% accurate

Precision and Recall

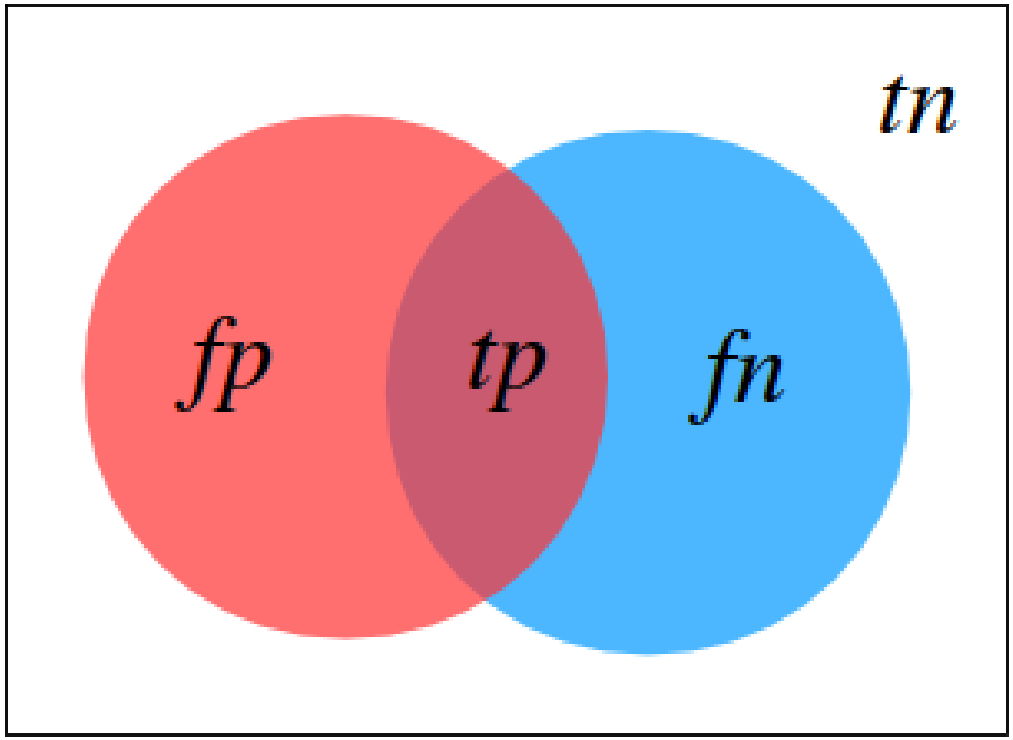
- **Precision (P)** is a fraction of retrieved documents that are relevant

$$P = \frac{\#(\text{relevant documents retrieved})}{\#(\text{retrieved documents})} = \frac{tp}{tp + fp}$$

- **Recall (R)** is a fraction of relevant documents that are retrieved

$$R = \frac{\#(\text{relevant documents retrieved})}{\#(\text{relevant documents})} = \frac{tp}{tp + fn}$$

Precision and Recall



$$P = \frac{tp}{tp + fp} \qquad R = \frac{tp}{tp + fn}$$

F-measure

- Combining P and R into a single number (the harmonic mean)

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \qquad \text{where } \beta^2 = \frac{1-\alpha}{\alpha}$$

- β controls the precision–recall trade-off:
  - $\beta = 1$ gives equal weight to precision and recall (**F1-score**):

$$F_{\beta=1} = \frac{2PR}{P+R}$$

  - $\beta = 0.5$ emphasizes precision twice as much as recall
  - $\beta = 2$ emphasized recall twice as much as precision

Evaluation of ranked results

- Modern search engines produce ranked results, but P, R, and F-score do not account for ranks
- Ideally, a search engine should rank all the relevant documents before the non-relevant ones
- Rank-based metrics:
  - Precision-recall curve
  - 11-point precision
  - MAP
  - P@k
  - R-precision
  - MRR

Learning outcomes 2

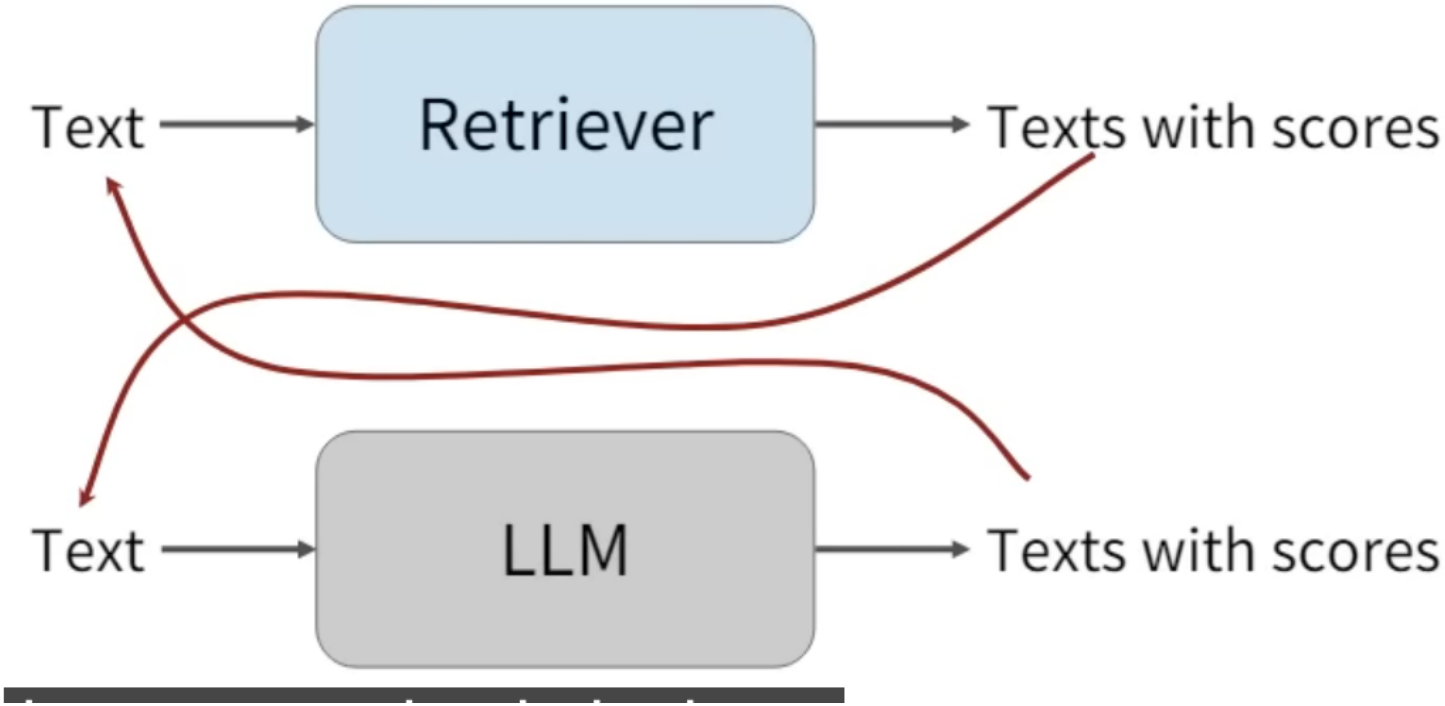
1. Explain what an IR test collection consist of and what it's used for
2. Define and calculate the standard IR evaluation metrics

Intro

Neural IR

- BERT+search:
  - https://jalammar.github.io/illustrated-retrieval-transformer/
  - https://blog.google/products/search/search-language-understanding-bert/
  - https://azure.microsoft.com/en-us/blog/bing-delivers-its-largest-improvement-in-search-experience-using-azure-g
- Generative LLMs for IR (also: conversational search):
  - https://www.technologyreview.com/2021/05/14/1024918/language-models-gpt3-search-engine-google/
  - Metzler et al. (2021). Rethinking search: making domain experts out of dilettantes. ACM SIGIR. https://arxiv.org/pdf/2105.02274.pdf
  - Shah & Bender. (2022). Situating search. ACM SIGIR. https://dl.acm.org/doi/abs/10.1145/3498366.3505816

Models can communicate in natural language



Christoper Potts: Stanford Webinar – GPT-3 & Beyond.

https://www.youtube.com/watch?app=desktop&v=-lnHHWRCDGk

Learning outcomes 3

1. Explain LLM-based search and what the advantages of this approach are over standard IR
2. Summarize the main critical points against LLM-based search and exemplify search scenarios in which such search fails

Study assignment

1. Watch TAR "Basics of IR 1/3" video:
   https://www.youtube.com/watch?v=MtzeLh99VHc&feature=youtu.be
2. Watch TAR "Basic of IR 2/3" video:
   https://www.youtube.com/watch?v=DeMc-OoGem8&feature=youtu.be
3. Read these three articles on using LLMs for search:
   - https://www.technologyreview.com/2021/05/14/1024918/language-models-gpt3-search-engine-google/
   - https://www.technologyreview.com/2023/02/16/1068695/chatgpt-chatbot-battle-search-microsoft-bing-google
   - https://blogs.bing.com/search/february-2023/The-new-Bing-Edge-%E2%80%93-Learning-from-our-first-week
4. Read (Shah & Bender, 2022) (focus on sections 1, 2, 4.2, and 4.3):
   https://dl.acm.org/doi/abs/10.1145/3498366.3505816
5. Self-check against learning outcomes!