TDT4117 Information Retrieval - Autumn 2025
Assignment 5

Important notes

Please carefully read the following notes and consider them for the assignment delivery. Submissions that do not fulfill these requirements will not be assessed and should be submitted again.

1. You may work in groups of maximum 2 students.
2. The assignment must be delivered in pdf format
3. This assignment is hand-written

Due date: 09.11.2025 23:59

Task 1: IR Models

Term-document matrix is a rudimentary data model for storing and querying a document collection. In such a matrix, the terms are derived from the entire vocabulary of the document collection and the entries correspond to whether the term is contained within the document. Answer the following questions in short with appropriate reasons based on the term-document representation of a document collection:

1. A user requires the capability of searching for phrases such that terms are matched in the documents with same ordering as contained within the query. Is this feasible with a term document matrix? If not, what additional information must be stored to provide such a search capability?

2. The term-document matrix described above records for each entry a boolean value indicating presence / absence of terms. Can we provide a ranking of documents matched for a query with this IR model?

If not, what changes would you make to the term-document matrix to support ranking?
3. You are considering to store a document collection containing one million documents with the term- document matrix representation. Is this a wise choice to consider? If not, what changes can you consider when implementing a solution using term-document matrix?

4. You are considering to use the term-document matrix representation for implementing search function- ality for files and folders in an operating system. Is this scenario a recall or a precision oriented task? Is it a wise choice to consider term-document matrix representation for this scenario?

5. Suggesting terms to complete a query is a important feature for modern search engines. Is it feasible to implement such a feature utilizing only term-document matrix? If not, what kind of a matrix would we require additionally?

Task 2: Boolean Retrieval

Consider the following toy document collection:

- d1 = *hbase is an open source key value store.*
- d2 = *dynamodb is a proprietary key value store.*
- d3 = *redis is popular.*
- d4 = *redis is not popular.*

Stopwords such as *is, an, a* may be ignored.

(a) Construct a Boolean term-document matrix for the collection.

(b) Using the matrix, find the result sets for the following queries:

(a) $q_1 = key \wedge value \wedge store$

(b) $q_2 = (key \wedge value \wedge store) \wedge (not \vee popular)$

Task 3: Evaluation Metrics

Two teams submitted their ranked lists for a query Q. "+" indicates relevant, "−" indicates non-relevant.

	1	2	3	4	5	6	7	8	9	10
Team 1	+	−	+	−	+	−	−	+	−	−
Team 2	−	+	−	−	+	+	−	−	−	−

There are 4 relevant documents in total.

Compute the following for both teams:

(a) Precision@5

(b) Recall@5

(c) Mean Average Precision (MAP)

(d) Mean Reciprocal Rank (MRR)

Task 4: Local Association and Query Expansion

Given the collection:

- d1 = *oslo to bergen. bergen to trondheim.*
- d2 = *stockholm to helsinki. helsinki to bergen.*
- d3 = *helsinki to oslo. oslo to stockholm.*

Let the correlation between terms w_u and w_v be:

$$c_{u,v} = \sum_{d_j \in D} f_{u,j} \cdot f_{v,j}$$

(a) Construct the unnormalized association matrix.

(b) Based on the matrix, identify which terms should be used for expanding the query *"oslo"*.

Task 5: Text Indexing

Given the sentence:
"The mind is wandering around the mind and does not mind that every mind does wander around."

(a) Construct an inverted list for the text, assuming stopwords *the, is, and* are removed.

(b) Briefly explain the difference between a **vocabulary trie** and a **suffix trie**.