Topic 2: IR Models

CAIM: Cerca i Anàlisi d’Informació Massiva

Exercise list, Fall 2025

Basic Comprehension Questions. Make sure you can answer them before proceeding.

1. True or false: The boolean model does not rank documents in the answer, while the vectorial model allows for ranking.
2. Suppose you are given the frequency of every term in a given document. What other information do you need to compute its representation in tf-idf weights?
3. Hide the course slides. Write down the formula of the cosine measure of document similarity. Now look at the slides. Check your answer. Repeat until correct.
4. Same for the tf-idf weight assignment scheme.
5. Write down the definitions of recall, precision, coverage, and novelty. Explain them in words in a way that you think your classmates would understand.
6. Explain to yourself how to compute a precision/recall graph.
7. True or false or criticize: To maximize user satisfaction, aim at a balance between recall and precision.
8. Write down Rocchio’s formula for user relevance feedback.

Exercise 1

Documents:

- D1: *Shipment of gold damaged in a fire*
- D2: *Delivery of silver arrived in a silver truck*
- D3: *Shipment of gold arrived in a truck*

Term set: `T = {fire, gold, silver, truck}`

1. Compute, using the boolean model, what documents satisfy the query

```
(fire OR gold) AND (truck OR NOT silver)
```

and justify your answer.

2. Do the same with the query

```
(fire OR NOT silver) AND (NOT truck OR NOT fire)
```

3. Argue whether it is possible to rewrite these queries using only the operators `AND`, `OR`, and `BUTNOT` (Boolean minus) in a logically equivalent way for *all* possible document collections.

Exercise 2

Document collection (five documents):

- Doc1: *we wish efficiency in the implementation for a particular application*
- Doc2: *the classification methods are an application of Li's ideas*
- Doc3: *the classification has not followed any implementation pattern*
- Doc4: *we have to take care of the implementation time and implementation efficiency*
- Doc5: *the efficiency is in terms of implementation methods and application methods*

Assume every word with 6 or more letters is a term, and terms are ordered by first appearance.

1. Give the representation of each document in the boolean model.
2. Give the representation in the vector model using tf-idf weights of documents Doc1 and Doc5. Compute the cosine similarity between Doc1 and Doc5.

(Hint: one expected answer gives similarity ≈ 0.162.)

Exercise 3

Indexed terms and document frequency percentages (percent of documents where term appears):

Term	% docs
computer	10
software	10
bugs	5
code	2
developer	2
programmers	2

Query: `Q = "computer software programmers"`.

Use tf-idf weights for documents, binary weights for the query, and cosine measure. Compute similarity between Q and:

- D1 = "programmers write computer software code"
- D2 = "most software has bugs, but good software has less bugs than bad software"
- D3 = "some bugs can be found only by executing the software, not by examining the source code"

(Expected similarities: ≈ 0.966, 0.436, 0.244.)

Exercise 4

Suppose terms A, B, C, D appear in 10,000, 8,000, 5,000, and 3,000 documents respectively, out of a collection of 100,000.

1. For boolean query `(A AND B) OR (C AND D)`, what is the worst-case upper bound on the size of the answer set?
2. For `(A AND B) OR (A AND D)`, how large can the answer be in the worst case? Think carefully about overlaps.
3. Compute the similarity between documents represented by term lists `"A B B A C C"` and `"D A D B B C C"` using tf-idf and cosine.

(Answers: 1) 11,000 2) 10,000 3) ≈ 0.736.)

Exercise 5

Collection size: 1,000,000 documents. Term document frequencies:

Term	# docs
computing	300,000
networks	200,000
computer	100,000
files	100,000
system	100,000
client	80,000
programs	80,000
transfer	50,000
agents	40,000
p2p	20,000
applications	10,000

1. Compute cosine similarity (tf-idf) between:

   - D1 = "p2p programs help users sharing files, applications, other programs, etc. in computer networks"
   - D2 = "p2p networks contain programs, applications, and also files"

   (Expected similarity $\approx$ 0.925.)

2. Give a document with exactly two different terms that **maximizes** similarity with the document

   > "p2p networks contain programs, applications, and also files"

   Compute that maximum similarity and justify optimality among two-term documents.

Exercise 6

Collection of 4 documents:

- Doc1: *Shared Computer Resources*
- Doc2: *Computer Services*
- Doc3: *Digital Shared Components*
- Doc4: *Computer Resources Shared Components*

1. Boolean representation of Doc3.

2. Which documents retrieved by boolean query `"Computer BUTNOT Components"`?
3. Compute idf(`Computer`) and idf(`Components`) (use log-based idf, state the base and smoothing you choose).
4. Compute vector (tf-idf) for Doc4.
5. Compute similarity between query `"Computer Components"` (binary weights) and Doc4 (tf-idf), using cosine.

(Expected answer for #5: $\approx$ 0.6534.)

Exercise 7

A user found 10 relevant documents at positions: `2, 6, 12, 18, 20, 22, 30, 36, 40, 50`. Assume there are no other relevant documents.

1. Draw a precision–recall graph at 10 recall levels and provide the table of numbers used to plot it.
2. Draw the *interpolated* version of the precision-recall graph.

Exercise 8

We have a small corpus with 100 documents (IDs 1..100). Relevant documents for a given query are 1..20.

Two information retrieval systems give the following ranked answers:

> S1 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 28, 29, 30, 31, 32, 33, 40, 41, 42, 43, 44, 45, 50, 51, 52, 53, 54, 60, 62, 63, 64, 70, 78, 80, 81, 82, 83, 85, 90, 91, 92, 93, 94, 95, 96, 98]
>
> S2 = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 25, 26, 27, 28, 29, 30, 35, 36, 37, 38, 40, 42, 45, 46, 48, 50, 51, 60, 61, 64, 70, 72, 78, 79, 90]

For each system:

a) Compute recall, precision, and F-measure (with $\alpha = 1/2, 1/4, 3/4$).

b) Compute novelty and coverage assuming the user already knew documents with odd IDs and did not know even IDs.

Exercise 9

Prove or disprove: Any boolean query that uses `NOT` can be transformed into an equivalent query that uses only `AND`, `OR`, and `BUTNOT` (where `BUTNOT` is equivalent to `A AND NOT B`). Give a constructive method or a counterexample.

Exercise 10

Given documents:

- D1: "machine learning algorithms for classification"
- D2: "deep learning methods and neural networks"
- D3: "classification metrics and evaluation"

Query: Q = "learning classification"

1. Compute tf-idf vectors for D1, D2, D3 (use log-frequency tf and idf = log(N/df) with N=3). Use binary weights for Q. Compute cosine similarities and rank.
2. Now suppose the user indicates D1 is relevant and D3 is non-relevant. Apply one iteration of Rocchio relevance feedback with α=1, β=0.75, γ=0.15. Show the updated query vector and re-rank.

Exercise 11

A system returns a ranked list of 10 documents, with relevant docs appearing at positions `1, 3, 4, 7`. Compute:

- Precision@1, Precision@3, Precision@5
- Average Precision (AP) for this query

Exercise 12

You are evaluating two information retrieval (IR) systems, System A and System B. For a specific query, you have been given the ranking of the top 10 documents returned by each system. An expert has assessed the relevance of these documents. The expert found 4 relevant documents in the entire collection that were placed in the following positions in each system:

- System A returns relevant documents at positions 1, 2, 9, and 10.
- System B returns relevant documents at positions 4, 5, 6, and 7.

Which one do you think is a better system?