

Web Retrieval SoSe 2025

Module Overview

Lecture Materials

Exercise Materials

Tutorials Overview

Assignment 01: Introducti

Assignment 02: Evaluation

Assignment 03 - Internal

Assignment 04 - Underlyin

Assignment 05 - Language

Assignment 06 - Web Crawl

Assignment 07 - Search on

Assignment 08 - PageRank

Exam Eligibility Assignme

Forum

Course info

My course

Exam Eligibility Assignment

Performance summary

✓ Assessed

Success status



Score



Passed

88 of 100 points

Attempts



1

▾ Results

Course

Web Retrieval SoSe 2025

ID: 4853531344 / 111798300966131

Test

Exam Eligibility Assignment

ID: 4915462450

This are your test results

Duration

4h 24m 57s 5/27/2025, 11:13 AM - 6/1/2025, 11:40 AM

Answered

33 of 33 questions (100%)

Your score

88 of 100 points (88%)



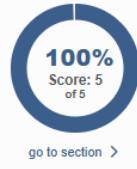
1. Knowledge Tasks (20 Points) 1



go to section >



1.2 Multiple Choice Questions (5 Points) 10



go to section >



2. Practical Tasks (50 Points) 1



go to section >



2.1 Vector-Based Ranking (22 Points) 3



go to section >



2.2 Query Likelihood Scoring (18 Points) 3



go to section >



2.3 Evaluation (8 Points) 4



go to section >



3.1 Document-Query Similarity (15 Points) 1



go to section >



3.2 Rocchio Feedback (15 Points) 10



go to section >



📚 1. Knowledge Tasks (20 Points) 15 of 15 points (100%)

1.1 True/False Questions (15 Points)

Status Answered

Your score 15 / 15

100%

Response

Each correct answer scores +0.5 points.

Each incorrect answer scores -0.5 points.

Unanswered questions score 0 points.

Maximum total score: 15 points

Minimum total score: 0 points			
Unanswered	Right	Wrong	
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	People can browse, navigate, or query to find information in web information retrieval (WIR) systems.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	The aim of an Information Retrieval (IR) system is to retrieve items that are relevant to the user's query and present them in a ranked list.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Database Management Systems (DBMS) primarily focus on freeform text rather than structured data.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	System-centered IR approaches focus on developing algorithms, while user-centered approaches focus on understanding user interactions.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	R-Precision uses a fixed cutoff value k regardless of the number of relevant documents.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Users rarely go beyond the first page of search results, influencing the use of Precision at k ($p@k$).
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	The ranking order of documents affects the precision and recall values in a Boolean retrieval system.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Fallout measures the fraction of non-relevant documents retrieved by the system.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Fallout is a very widely used metric for evaluating IR systems.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Term normalisation in document indexing converts all terms to the same case, typically lowercase.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Optimizing Boolean retrieval involves intersecting shorter posting lists first to reduce processing time.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Biword indexes can reliably verify that a document contains a long phrase without false positives.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Positional indexes allow for proximity queries, such as finding terms within a certain number of words of each other.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Terms that occur too frequently in a collection are often removed as stopwords because they are less informative.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	The Boolean model provides a ranked list of documents based on their relevance to the query.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Pseudo relevance feedback uses top- k results as positive feedback to adjust the query vector without user input.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	The sequence of query terms matters when constructing a query vector in the Vector Space Model.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	The cosine similarity measure in the Vector Space Model normalizes for document length.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	The Jaccard coefficient considers term frequency when computing query-document match scores.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Ranked retrieval models return a set of documents that strictly satisfy a query expression.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Boolean queries often result in either too few or too many results.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	The Boolean model is intuitive for the majority of users to express information needs.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Language models consider only the syntax of queries and not their semantics.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	A finite state automaton can be viewed as a deterministic language model.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	The sparseness issue makes it difficult to estimate probabilities for long phrases in language models.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Language models for IR do not normalize term frequencies by document length.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	The probability of a term in a unigram model is calculated as the number of occurrences of the term divided by the total number of word occurrences in the collection.
<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	Pre-trained large language models like BERT are primarily used for simple unigram-based IR tasks.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	Dirichlet smoothing incorporates the collection frequency of terms to adjust probabilities.
<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	The use of a global language model in smoothing is to ensure that all words have a non-zero probability of occurring.

► Solution

[◀ go back to overview](#)

 **1.2 Multiple Choice Questions (5 Points)** 5 of 5 points (100%)

Select the correct answer. No negative points are deducted for incorrect answers.

Q1 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5 <div style="width: 100%; background-color: #005a7f; height: 10px; margin-left: 10px;"></div> 100%

Response

What is the primary role of a web crawler in an IR system?

- To rank search results based on relevance
- To collect and index web pages for retrieval
- To process user queries in real-time
- To refine user queries for better accuracy

► Solution

Q2 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5 <div style="width: 100%; background-color: #005a7f; height: 10px; margin-left: 10px;"></div> 100%

Response

Which component of an IR system is responsible for storing term-document relationships?

- Ranking Algorithm
- Inverted Index
- Web Crawler
- Query Processor

► Solution

Q3 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5 <div style="width: 100%; background-color: #005a7f; height: 10px; margin-left: 10px;"></div> 100%

Response

Which modern IR development focuses on generating answers directly from document corpora?

- Document Clustering
- Link Analysis
- Automated Indexing
- Generative IR

► Solution

Q4 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5 <div style="width: 100%; background-color: #005a7f; height: 10px; margin-left: 10px;"></div> 100%

Response

What is the primary role of a web crawler's frontier in web search?

- To store the list of URLs yet to be crawled
- To filter out spam pages during indexing
- To process user queries in real-time
- To rank web pages based on their relevance

► Solution

Q5 (0.5 Points)

Status	Answered	
Your score	0.5 / 0.5	<div style="width: 100%; background-color: #005a7a; height: 10px;"></div> 100%

Response

Which technique allows flexible matching like retrieving both "naïve" and "naive"?

- Boolean retrieval
- Positional index
- Biword indexing
- Wildcard search

► Solution

Q6 (0.5 Points)

Status	Answered	
Your score	0.5 / 0.5	<div style="width: 100%; background-color: #005a7a; height: 10px;"></div> 100%

Response

In language models for IR, what does the collection model $P(t | C)$ represent?

- The probability of a term occurring in a specific document
- The likelihood of a document being relevant to a query
- The frequency of a term in a query
- The probability of a term occurring across the entire collection

► Solution

Q7 (0.5 Points)

Status	Answered	
Your score	0.5 / 0.5	<div style="width: 100%; background-color: #005a7a; height: 10px;"></div> 100%

Response

What determines the size of the Term-Document (TD) matrix in an Information Retrieval system?

- Number of documents \times number of queries
- Number of users \times number of documents
- Number of unique terms \times number of documents
- Number of relevant documents \times number of terms

► Solution

Q8 (0.5 Points)

Status	Answered	
Your score	0.5 / 0.5	<div style="width: 100%; background-color: #005a7a; height: 10px;"></div> 100%

Response

Which of the following is a key advantage of using crowdsourcing for relevance judgments in Web Information Retrieval (WIR)?

- Greater use of advanced IR evaluation metrics
- Higher accuracy than expert judgments
- More consistent labeling across all topics
- Lower cost and faster data collection

► Solution

④ Q9 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5

Response

What is the purpose of a posting list in an inverted index?

- It records term frequencies
- It stores all stopwords
- It lists documents where a term occurs
- It compresses the entire document

► Solution

④ Q10 (0.5 Points)

Status	Answered
Your score	0.5 / 0.5

Response

In the context of IR evaluation, what does the Average Precision (AP) metric measure?

- The harmonic mean of precision and recall
- The fraction of non-relevant documents retrieved
- The average of precision values at each relevant document in the ranked list
- The total number of relevant documents retrieved

► Solution

[◀ go back to overview](#)

⌚ 2. Practical Tasks (50 Points) 2 of 2 points (100%)

[◀ go back to overview](#)

⌚ 2.1 Vector-Based Ranking (22 Points) 10 of 22 points (45%)

Given the following documents:

```
d0 = "learning vector representation classification"
d1 = "probabilistic inference language model"
d2 = "vector space model for retrieval"
```

and the Query:

```
q = "vector model retrieval"
```

⌚ 2.1.1 TF-IDF (8.8 Points)

Status	Answered
...	...

Your score

8.8 / 8.8

100%

Response

Compute the TF-IDF vectors for each document and the query. Each right answer scores 0.2 points.

Note: Give the numerical answers rounded to **three** decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0→ 0.0

	do	d1	d2	query
classification	0.477	0.0	0.0	0.0
for	0.0	0.0	0.477	0.0
inference	0.0	0.477	0.0	0.0
language	0.0	0.477	0.0	0.0
learning	0.477	0.0	0.0	0.0
model	0.0	0.176	0.176	0.176
probabilistic	0.0	0.477	0.0	0.0
representation	0.477	0.0	0.0	0.0
retrieval	0.0	0.0	0.477	0.477
space	0.0	0.0	0.477	0.0
vector	0.176	0.0	0.176	0.176

► Solution

2.1.2 Cosine Similarity Scores (12 Points)

Status

Answered

Your score

0 / 12

0%

Response

Compute **cosine similarity** scores between the query and each document. Each right answer scores 4 points. No negative points are deducted for wrong answers.

Note: Give the numerical answers rounded to **two** decimal places, e.g., 0.18. Round the last digit as follows: 0.182 → 0.18, 0.185 → 0.19, 0.186 → 0.19

d0: 0.068
d1: 0.068
d2: 0.624

▼ Solution

Compute **cosine similarity** scores between the query and each document. Each right answer scores 4 points. No negative points are deducted for wrong answers.

Note: Give the numerical answers rounded to **two** decimal places, e.g., 0.18. Round the last digit as follows: 0.182 → 0.18, 0.185 → 0.19, 0.186 → 0.19

d0: 0.07
d1: 0.07
d2: 0.62

2.1.3 Document Ranking (1.2 Points)

Status	Answered
Your score	1.2 / 1.2 100%

Response

Rank the documents by similarity. Which document has the highest rank?

d0

d1

d2

► Solution

[◀ go back to overview](#)

2.2 Query Likelihood Scoring (18 Points) 18 of 18 points (100%)

You are given a document:

d1 = "deep learning improves natural language understanding in many applications"

And a query:

q = "language applications"

If we assume the document itself is the collection, compute the **query likelihood** score:

2.2.1 Maximum Likelihood Estimation (6 Points)

Status	Answered
Your score	6 / 6 100%

Response

Under a **unigram** language model using Maximum Likelihood Estimation (MLE):

Under a **bigram** language model using Maximum Likelihood Estimation (MLE):

Note: Give the numerical answers rounded to **three** decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0 → 0.0

► Solution

2.2.2 Laplace Smoothing (6 Points)

Status	Answered
Your score	6 / 6 100%

Response

Under a **unigram** language model using Laplace smoothing (use '1' in the Laplace formula for alpha)

Note: Give the numerical answers rounded to **three** decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0 → 0.0

► Solution

2.2.3 Jelinek-Mercer Smoothing (6 Points)

Status	Answered
Your score	6 / 6 100%

Response

Under a **unigram** language model using Jelinek-Mercer smoothing ($\lambda = 0.8$, use uniform background model)

Note:

A **uniform background model** assumes that every term in the vocabulary has **equal probability** in the global collection:

$$P(w | C) = \frac{1}{|V|}$$

for all w in the vocabulary V.

Also: Give the numerical answers rounded to **three** decimal places, e.g., 0.185. Round the last digit as follows: 0.1856

Note: Give the numerical answers rounded to three decimal places, e.g., 0.186. Round the last digit as follows: 0.1866 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0→ 0.0

► Solution

[◀ go back to overview](#)

⌚ 2.3 Evaluation (8 Points) 8 of 8 points (100%)

Suppose you're building a **web search engine**. A user submits the query:

Mental health interventions using digital platforms

Your system returns the following ranked documents (top 10):

Rank	Document ID	Is Relevant?
1	D8	Yes
2	D3	No
3	D5	Yes
4	D6	Yes
5	D1	No
6	D7	Yes
7	D2	No
8	D9	No
9	D10	Yes
10	D4	No

Compute the following evaluation metrics at cut-off k=10:

- Precision@10
- Recall@10 (Assume total relevant documents in the collection = 6)
- F1-score@10
- Mean Average Precision (MAP)

⌚ 2.3.1 Precision@10 (2 Points)

Status	Answered
Your score	2 / 2

Response

Note: Give the numerical answers rounded to 3 decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0→ 0.0, 0.5→ 0.5

Precision@10: 0.5

► Solution

⌚ 2.3.2 Recall@10 (2 Points)

Status	Answered
Your score	2 / 2

Response

Note: Give the numerical answers rounded to 3 decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0→ 0.0, 0.5→ 0.5

Recall@10 (Assume total relevant documents in the collection = 6):

0.833

► Solution

⌚ 2.3.3 F1-score@10 (2 Points)

Status	Answered

Your score	2 / 2	<div style="width: 100%;">100%</div>
------------	-------	--------------------------------------

Response

Note: Give the numerical answers rounded to 3 decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0 → 0.0, 0.5 → 0.5

F1-score@10 (use the values for precision and recall, you computed in the previous questions):

0.625

► Solution

2.3.4 Mean Average Precision (MAP) (2 Points)

Status	Answered
--------	----------

Your score	2 / 2	<div style="width: 100%;">100%</div>
------------	-------	--------------------------------------

Response

Note: Give the numerical answers rounded to 3 decimal places, e.g., 0.185. Round the last digit as follows: 0.1856 → 0.186, 0.1851 → 0.185, 0.1855 → 0.186, 0 → 0.0, 0.5 → 0.5

Mean Average Precision (MAP) : 0.606

► Solution

2.4 Boolean Logic (2 Points)

Status	Answered
--------	----------

Your score	2 / 2	<div style="width: 100%;">100%</div>
------------	-------	--------------------------------------

Response

Translate the natural language query below into a Boolean logic expression using AND, OR, and NOT.

Find articles that mention “climate change” but exclude “global warming” unless “carbon emissions” is also mentioned.

(“climate change”) AND (NOT “carbon emissions” OR “global warming”)

(“climate change”) AND (NOT “global warming” OR “carbon emissions”)

(“climate change”) AND NOT (“carbon emissions”) OR ((“global warming”) AND (“carbon emissions”))

(“climate change”) AND NOT (“global warming”) OR ((“global warming”) AND (“carbon emissions”))

(“climate change”) OR (NOT “global warming” OR “carbon emissions”)

► Solution

◀ go back to overview

3. Programming Tasks (30 Points)

◀ go back to overview

3.1 Document-Query Similarity (15 Points) 15 of 15 points (100%)

Complete the following Python program that calculates the cosine similarity between a query and a set of documents using the Vector Space Model (VSM).

Fill in the blanks with the appropriate code snippets from the provided choices.

Each blank is worth 1.5 points.

▼ Gap with dropdown	
---------------------	--

Status	Answered
--------	----------

Your score	15 / 15	<div style="width: 100%;">100%</div>
------------	---------	--------------------------------------

Response

```

1 import pandas as pd
2 import numpy as np
3
4 corpus = [
5     "sun sun",
6     "cloud wind wind rain rain",
7     "sun cloud cloud wind",
8     "sun sun sun cloud cloud cloud wind wind wind rain",
9     "wind wind snow snow"
10]
11 query = "cloud wind"
12 vocabulary = ['sun', 'cloud', 'wind', 'rain', 'snow']
13 def build_tf_matrix(corpus, vocabulary):
14     tf_matrix = []

```

Line 16 doc.split() ▾

Line 17 words.count(term) ▾

Line 19 tf_matrix ▾

Line 24 tf_matrix ▾

Line 25 0 ▾

Line 26 N ▾

```

15     for doc in corpus:
16         words = [_____ for term in vocabulary]
17         tf_vector = [_____ for term in vocabulary]
18         tf_matrix.append(tf_vector)
19     return pd.DataFrame(_____, columns=vocabulary)
20
21 tf_matrix = build_tf_matrix(corpus, vocabulary)
22
23 def compute_idf(tf_matrix):
24     N = len(_____)
25     df = (tf_matrix > _____.sum(axis=0))
26     idf = np.log10(_____/df)
27     return idf
28
29 idf_vector = compute_idf(tf_matrix)
30 tfidf_matrix = _____ * idf_vector
31
32 def build_query_vector(query, vocabulary, idf_vector):
33     words = query.split()
34     tf = pd.Series([words.count(term) for term in vocabulary], index=vocabulary)
35     return tf * idf_vector
36
37 query_vector = build_query_vector(query, vocabulary, idf_vector)
38
39 def cosine_similarity(tfidf_matrix, query_vector):
40     similarities = {}
41     q_vec = _____ .to_numpy()
42     q_norm = np.linalg.norm(q_vec)
43     for i, doc_vector in tfidf_matrix.iterrows():
44         d_vec = doc_vector.to_numpy()
45         d_norm = np.linalg.norm(_____)
46         dot_product = np.dot(d_vec, q_vec)
47         sim = dot_product / (d_norm * q_norm) if d_norm > 0 and q_norm > 0 else _____
48         similarities[f'd{i+1}'] = sim
49
50 return similarities
51
52 similarities = cosine_similarity(tfidf_matrix, query_vector)
53 print("TF-IDF Cosine Similarities:", similarities)
54 print(tfidf_matrix)

```

Line 30: tf_matrix
 Line 41: query_vector
 Line 45: d_vec
 Line 47: 0.0
 ▶ Solution

[◀ go back to overview](#)

3.2 Rocchio Feedback (15 Points) 15 of 15 points (100%)

Read the Python code provided below, which implements a Rocchio-based query update using TF-IDF vectors and cosine similarity. Then, answer a set of multiple choice questions that test your understanding of the algorithm, its mathematical structure, and its behavior.

Tip: Rather than averaging all relevant documents equally, this version of Rocchio Algorithm uses pre-assigned weights to control how much each document influences the query.

```

1 import pandas as pd
2 import numpy as np
3 from sklearn.preprocessing import normalize
4 from sklearn.metrics.pairwise import cosine_similarity
5 from IPython.display import display
6
7 tfidf_df = pd.DataFrame([
8     [0.3, 0.0, 0.2, 0.0], # d1 (relevant)
9     [0.1, 0.4, 0.3, 0.0], # d2 (non-relevant)
10    [0.0, 0.1, 0.2, 0.5], # d3
11    [0.4, 0.4, 0.0, 0.0], # d4 (relevant)
12    [0.0, 0.0, 0.4, 0.6] # d5 (non-relevant)
13 ], index=['d1', 'd2', 'd3', 'd4', 'd5'], columns=['term1', 'term2', 'term3', 'term4'])
14
15 query_vector = pd.Series([0.6, 0.6, 0.3, 0.0], index=tfidf_df.columns)
16 query_vector = pd.Series(normalize(query_vector.values.reshape(1, -1))[0], index=tfidf_df.columns)
17
18 relevance_feedback = {
19     'relevant': {'d1': 1.0, 'd4': 0.8},
20     'non_relevant': {'d2': 1.0, 'd5': 0.6}
21 }
22
23 def rocchio_weighted(query_vector, tfidf_df, feedback_dict, alpha=1.0, beta=1.0, gamma=4.0, normalize_output=True):
24     relevant_docs = list(feedback_dict['relevant'].keys())
25     non_relevant_docs = list(feedback_dict['non_relevant'].keys())
26
27     rel_weights = np.array([feedback_dict['relevant'][d] for d in relevant_docs])
28     nonrel_weights = np.array([feedback_dict['non_relevant'][d] for d in non_relevant_docs])
29
30     rel_matrix = tfidf_df.loc[relevant_docs].to_numpy()
31     nonrel_matrix = tfidf_df.loc[non_relevant_docs].to_numpy()
32
33     rel_centroid = np.average(rel_matrix, axis=0, weights=rel_weights)
34     nonrel_centroid = np.average(nonrel_matrix, axis=0, weights=nonrel_weights)
35
36     updated_query = alpha * query_vector.to_numpy() + beta * rel_centroid - gamma * nonrel_centroid
37
38     if normalize_output:
39         updated_query = normalize(updated_query.reshape(1, -1))[0]
40
41     return pd.Series(updated_query, index=query_vector.index)
42
43 updated_query_vector = rocchio_weighted(query_vector, tfidf_df, relevance_feedback)
44
45 cos_sim_before = cosine_similarity(tfidf_df.values, query_vector.values.reshape(1, -1))
46 cos_sim_after = cosine_similarity(tfidf_df.values, updated_query_vector.values.reshape(1, -1))
47
48 ranked_docs_before = pd.Series(cos_sim_before.flatten(), index=tfidf_df.index).sort_values(ascending=False)
49 ranked_docs_after = pd.Series(cos_sim_after.flatten(), index=tfidf_df.index).sort_values(ascending=False)

```

Q1

Status	Answered
Your score	1.5 / 1.5
	100%

Response

Suppose the defined query vector emphasizes term1 and term2. How does this affect the cosine similarity before feedback?

- The similarity scores will all be equal
- Documents with higher values in term3 and term4 will rank higher

Documents emphasizing term1 and term2 will likely rank higher

Only relevant documents are matched

► Solution

● Q2

Status Answered

Your score 1.5 / 1.5  100%

Response

In the Rocchio update, which component is subtracted from the updated query?

- The average of relevant document vectors
- The centroid of the non-relevant documents, scaled by gamma and beta
- The original query vector
- The centroid of the non-relevant documents, scaled by gamma

► Solution

● Q3

Status Answered

Your score 1.5 / 1.5  100%

Response

What could happen if the gamma parameter is set too high?

- Cosine similarity will no longer work
- The updated query may move too far from the relevant subspace
- The TF-IDF values will be distorted
- The influence of relevant documents will be amplified

► Solution

● Q4

Status Answered

Your score 1.5 / 1.5  100%

Response

How does using `np.average(..., weights=...)` in the Rocchio function improve flexibility?

- It removes the need for the alpha parameter
- It converts TF-IDF to binary
- It allows for variable strength of user feedback on individual documents
- It normalizes the entire document matrix

► Solution

● Q5

Status Answered

Your score 1.5 / 1.5  100%

Response

Which condition would most likely lead to a **negative term weight** in the updated query vector?

- A term appears only in non-relevant documents with high TF-IDF

- The alpha parameter is set to zero
- The original query is zero
- A term is present in all documents equally

► Solution

④ Q6

Status	Answered
Your score	1.5 / 1.5 <div style="width: 100%; background-color: #337AB7; height: 10px; margin-top: 5px;"></div> 100%

Response

Why is cosine similarity preferred over Euclidean distance in this implementation?

- It penalizes longer documents
- It is independent of document length and focuses on vector direction
- It avoids the curse of dimensionality
- It works better with binary term frequencies

► Solution

④ Q7

Status	Answered
Your score	1.5 / 1.5 <div style="width: 100%; background-color: #337AB7; height: 10px; margin-top: 5px;"></div> 100%

Response

If document d5 has high TF-IDF for term4 and the query has zero weight for term4, what effect does that have before Rocchio is applied?

- d5 will be dropped from the matrix
- Rocchio will automatically increase the query's weight for term4
- d5 will have low similarity due to orthogonality with the query
- d5 will be ranked highest

► Solution

④ Q8

Status	Answered
Your score	1.5 / 1.5 <div style="width: 100%; background-color: #337AB7; height: 10px; margin-top: 5px;"></div> 100%

Response

How does Rocchio relevance feedback affect **recall** in a typical retrieval task?

- It always decreases recall
- It is designed to optimize only precision
- It may improve recall by moving the query toward the relevant document space
- It narrows the query to fewer terms, decreasing recall

► Solution

④ Q9

Status	Answered
Your score	1.5 / 1.5 <div style="width: 100%; background-color: #337AB7; height: 10px; margin-top: 5px;"></div> 100%

Response

What theoretical assumption does the Rocchio model make about the geometry of the document space?

- Relevant and non-relevant documents form orthogonal subspaces

- Documents follow a Gaussian distribution in each dimension
- All terms are equally informative
- The ideal query lies somewhere between the centroids of relevant and non-relevant documents

► Solution

⌚ Q10

Status	Answered
Your score	1.5 / 1.5  100%

Response

What would be the most likely effect of assigning higher weights to certain relevant documents in the feedback dictionary?

- The gamma value will override their influence
- Cosine similarity will not be computable
- These documents will be excluded from centroid calculation
- The updated query will shift more strongly toward the vocabulary of those documents

► Solution

[◀ go back to overview](#)

Test execution

Information

- ⌚ Availability: Expired at 6/2/2025, 6:00 PM
- ⌚ Max. attempts: Unlimited
- ⌚ Results of this test are visible to administrators and tutors of this course.

[Start test](#)

► Change log

[^ Go to top](#)

Logged in as Ravi Himmabhai Ramani (1447 People are online)



[Imprint](#)
[Datenschutzerklärung](#)

OpenOlat 19.1.14

