Assignment 04 - Underlying Models (II) and Relevance Feedback

Performance summary

✓ Assessed

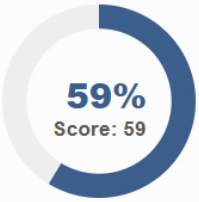
Success status	Score	Attempts
Undefined	59 of 100 points	1

Results

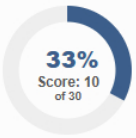
Course	**Web Retrieval SoSe 2025** ID: 4853531344 / 109697139206616
Test	**Assignment 04** ID: 4562682381

This are your test results

Duration	0h 53m 10s 5/19/2025, 11:14 AM - 5/19/2025, 12:08 PM
Answered	14 of 14 questions (100%)
Your score	59 of 100 points (59%)

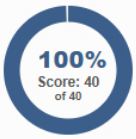


Knowledge Check Tasks (30 points) 4



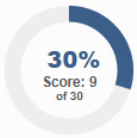
go to section

Practical Tasks (40 points) 7



go to section

Programming tasks (30 points) 3



go to section

Knowledge Check Tasks (30 points) 10 of 30 points (33%)

Vector Space Model (5 points)

Status	Answered	
Your score	5 / 5	100%

Response

Which of the following statements is correct? (unanswered/right/wrong)

Please note:

- Maximum Overall Score -> 5 points
- Minimum Overall Score -> 0 points
- Incorrect Answer -> -0.5 points
- Unanswered -> 0 points

Unanswered	Right	Wrong	
☐	☑	☐	Both a document and a query can be represented as vectors.
☐	☐	☑	The order of terms occurring in a query is considered by the vector space model.
☐	☑	☐	The frequency of terms occurring in a document is considered when representing the document as a vector (vector space model).
☐	☑	☐	The standard way of quantifying the similarity between two documents d1 and d2 is the cosine similarity of their vector representations.

☐	☐	☑	The pseudo relevance feedback is motivated by the fact that users are willing to give feedback for the retrieval results.

▸ Solution

☑ Inverse Document Frequency (5 points)

Status	Answered		
Your score	5 / 5		100%

Response

Which of the following sentences is true regarding the inverse document frequency (idf) of a term t, given a collection of N documents?

- ☐ The idf cannot be equal to 1.
- ☑ The idf of a rare term is high, whereas the idf of a frequent term is likely to be low..
- ☐ The idf counts the number of occurrences of a term in a document of the collection.
- ☐ The idf cannot be equal to 0.

▸ Solution

Jaccard Scoring (10 points)

Status	Answered		
Your score	0 / 10		0%

Response

Explain the limitations of using the Jaccard coefficient to score a query-document pair.

The Jaccard coefficient finds the degree of similarity between two sets, such as queries and documents, by evaluating the intersection and union of those two sets. Despite being quite straightforward to use, it is not helpful in scoring the relevance of a document to a user query for several reasons. First, it does not differentiate between the significance of terms; it equally weighs every term in the document. Consequently, highly relevant words alongside common phrases such as stop words reduce the effectiveness of the relevance scoring process. Second, the Jaccard coefficient only uses intersection cardinality as the set's membership which ignores how many times a word shows up in the document. This assumption discards useful information associated with term frequency which is essential to determining the document's relevance. The Jaccard measure also cannot account for semantic relationships; documents with different terminologies but using synonyms or related concepts will obtain lower Jaccard scores even when relevant. This metric also suffers from poor performance in documents with varying lengths. With longer documents, the quantity of terms in the union usually increases, so those documents tend to have higher term overlaps which will automatically drops the similarity score unfairly.

193 words (max. 200)

▸ Solution

Term Frequency (10 points)

Status	Answered		
Your score	0 / 10		0%

Response

One measure for quantifying the frequency of a term t in a document d is to use the *raw* term frequency (tf), which counts the number of occurrences of term t in the document d. Give a variant for measuring term frequency (write the formula and cite the reference).

A more sophisticated synonym of the raw term frequency (tf) which counts the number of times a term appears in a document is the logarithmic term frequency. This metric solves the problem of raw counts overestimating the relevance of documents that contain terms of high frequency to be smaller than the threshold, along edge-case and unimportant terms, which might almost always appear. The formula for logarithmic term frequency is: tf_log (t,d)= 1 + log(f_(t, d)) if f_(t, d) > 0; 0 otherwise.

In documents, such terms can dominate document representation and such variants are used to aid focus representation by infusing the approach which applies logarithm to reduce the effect of very high frequencies.

The primary goal of this strategy is to mitigate the overskewed bias caused by raw counts in a document that can be visualised as dominant, thereby improving precision. Without using raw counts, logarithmic smoothing brings aggressive checking of variance and index difference, adding strength of the term relative importance for boosted queries in automated systems.

This approach has been implemented widely in information retrieval systems as underlined in Salton and Buckley's (1988) driving term weighting work such that documents can utilize elaborated quantifying frameworks that improve efficiency.

191 words (max. 200)

▸ Solution

go back to overview

Practical Tasks (40 points) 40 of 40 points (100%)

Documents as Vectors (10 points)

Status	Answered
Your score	10 / 10 100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = "regression weak classification intelligence kernel"
- d1 = "network weights weak classification"
- d2 = "regression weak tangent"
- d3 = "weak classification artificial"

Model the four documents d0, d1, d2, d3 as vectors, by using the tf-idf for computing the term weights and considering the following term order:

< 'kernel', 'weight', 'classification', 'weak', 'tangent', 'artificial', 'network', 'intelligence', 'regression'>

Which of the following choices is the right VSM representation of the four documents? The T symbol indicates Transpose.

Unanswered	Right	Wrong	
☐	☐	☑	d0 = <0.602, 0, 0.125, 0, 0, 0, 0, 0.602, 0.301>T d1= <0, 0.602, 0.125, 0, 0, 0, 0.602, 0, 0>T d2= <0, 0, 0, 0, 0.602, 0, 0, 0.301, 0 >T d3= <0, 0, 0.125, 0, 0, 0.602, 0, 0, 0>T
☐	☑	☐	d0 = <0.602, 0, 0.125, 0, 0, 0, 0, 0.602, 0.301>T d1= <0, 0.602, 0.125, 0, 0, 0, 0.602, 0, 0>T d2= <0, 0, 0, 0, 0.602, 0, 0, 0, 0.301>T d3= <0, 0, 0.125, 0, 0, 0.602, 0, 0, 0>T
☐	☐	☑	d0 = <0.602, 0, 0.125, 0, 0, 0, 0, 0.602, 0.301>T d1= <0, 0.602, 0.125, 0, 0, 0, 0.602, 0, 0>T d2= <0, 0, 0, 0, 0.602, 0, 0, 0, 0.301>T d3= <0, 0, 0.125, 0, 0, 0, 0.602, 0, 0>T
☐	☐	☑	None of them

▸ Solution

☑ Query as a Vector (5 points)

Status	Answered
Your score	5 / 5 100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = "regression weak classification intelligence kernel"
- d1 = "network weights weak classification"
- d2 = "regression weak tangent"
- d3 = "weak classification artificial"

and q the query "regression weak tangent intelligence"

Model the query as a vector, by using the tf-idf for computing the term weights and considering the following term order:

< 'kernel', 'weights', 'classification', 'weak', 'tangent', 'artificial', 'network', 'intelligence', 'regression'>

Which of the following choices is the right VSM representation of the query? The T symbol indicates Transpose

- ☐ q= <0.602, 0, 0, 0, 0.602, 0, 0, 0.602, 0.301>T
- ☑ q= <0, 0, 0, 0, 0.602, 0, 0, 0.602, 0.301>T
- ☐ q= <0, 0, 0, 0.602, 0, 0, 0.602, 0.301>T
- ☐ None of them

▸ Solution

Cosine Similarity (q, d0) [5 points]

Status	Answered

Your score	5 / 5		100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = “regression weak classification intelligence kernel”
- d1 = “network weights weak classification”
- d2 = “regression weak tangent”
- d3 = “weak classification artificial”

and q the query “regression weak tangent intelligence”

Calculate the cosine similarity between the query q and the document d0 (use the document and query vectors that you calculated in the previous tasks). **Round up to two decimals.**

Cosine Similarity = 0.55

▸ Solution

Cosine Similarity (q, d1) [5 points]

Status	Answered		
Your score	5 / 5		100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = “regression weak classification intelligence kernel”
- d1 = “network weights weak classification”
- d2 = “regression weak tangent”
- d3 = “weak classification artificial”

and q the query “regression weak tangent intelligence”

Calculate the cosine similarity between the query q and the document d1 (use the document and query vectors that you calculated in the previous tasks). **Round up to two decimals.**

Cosine Similarity = 0.0

▸ Solution

Cosine Similarity (q, d2) [5 points]

Status	Answered		
Your score	5 / 5		100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = “regression weak classification intelligence kernel”
- d1 = “network weights weak classification”
- d2 = “regression weak tangent”
- d3 = “weak classification artificial”

and q the query “regression weak tangent intelligence”

Calculate the cosine similarity between the query q and the document d2 (use the document and query vectors that you calculated in the previous tasks). **Round up to two decimals.**

Cosine Similarity : 0.75

▸ Solution

Cosine Similarity (q, d3) [5 points]

Status	Answered		
Your score	5 / 5		100%

Response

Let D = {d0, d1, d2, d3} with

- d0 = “regression weak classification intelligence kernel”
- d1 = “network weights weak classification”
- d2 = “regression weak tangent”
- d3 = “weak classification artificial”

and q the query “regression weak tangent intelligence”

Calculate the cosine similarity between the query q and the document d3 (use the document and query vectors that you calculated in the previous tasks). **Round up to two decimals.**

Cosine Similarity = 0.0

▸ Solution

Jaccard coefficient [5 points]

Status	Answered		
Your score	5 / 5		100%

Response

Calculate the Jaccard coefficient for expressing the similarity between the following two documents D1 and D2:

D1 = "shall see sun soon"

D2 = "we shall see sea sun shine soon"

Round up to two decimals.

Jaccard coefficient : 0.57

▶ Solution

go back to overview

Programming tasks (30 points) 9 of 30 points (30%)

Consider the following documents **d0, d1, d2, d3,** and the query **q**:

- **d0 = "king palace jungle sleeps"**
- **d1 = "jungle lion fire king discovery"**
- **d2 = "king timon musafa"**
- **d3 = "timon lion simba"**
- **q = "lion king"**

word_bag = set(d0.split() + d1.split() + d2.split() +d3.split() +q.split())

Write a function in python named **Rocchio(alpha, beta, gamma)** that returns the Rocchio feedback, and use it for the following tasks.

••• Rocchio Feedback no. 1 (10 points)

Status	Answered		
Your score	3 / 10		30%

Response

A user characterizes as **relevant** the **d1** document and as **irrelevant** the **d3** document.

Calculate the Rocchio feedback using as **alpha =1**, **beta = 0.8** and **gamma = 0.1.**

Round up the results to two decimals.

fire 0.8

timon -0.1

discovery 0.8

sleeps 0.0

lion 1.7

king 1.8

jungle 0.8

simba -0.1

mufasa 0.0

palace 0.0

▼ Solution

A user characterizes as **relevant** the **d1** document and as **irrelevant** the **d3** document.

Calculate the Rocchio feedback using as **alpha =1**, **beta = 0.8** and **gamma = 0.1.**

Round up the results to two decimals.

fire 0.48

timon -0.03

discovery 0.48

sleeps 0.0

lion 0.51

king 0.22

jungle 0.24

simba -0.06

mufasa 0.0

palace 0.0

••• Rocchio Feedback no. 2 (10 points)

Status	Answered	
Your score	3 / 10	30%

Response

A user characterizes as **relevant** the **d1** document and as **irrelevant** the **d3** document.

Calculate the Rocchio feedback using as **alpha = 1**, **beta = 0.1** and **gamma = 0.9.**

Round up the results to two decimals.

fire 0.1
timon -0.9
discovery 0.1
sleeps 0.0
lion 0.1
king 1.0
jungle 0.1
simba -0.9
mufasa 0.0
palace 0.0

▼ Solution

A user characterizes as **relevant** the **d1** document and as **irrelevant** the **d3** document.

Calculate the Rocchio feedback using as **alpha = 1**, **beta = 0.1** and **gamma = 0.9.**

Round up the results to two decimals.

fire 0.06
timon -0.27
discovery 0.06
sleeps 0.0
lion 0.06
king 0.14
jungle 0.03
simba -0.54
mufasa 0.0
palace 0.0

••• Rocchio Feedback no. 3 (10 points)

Status	Answered	
Your score	3 / 10	30%

Response

A user characterizes as relevant the d1 document and as irrelevant the d3 document.

Calculate the Rocchio feedback using as alpha =1, beta = 1 and gamma = 1.

Round up the results to two decimals.

fire 1.0
timon -1.0
discovery 1.0
sleeps 0.0
lion 1.0
king 2.0
jungle 1.0
simba -1.0
mufasa 0.0
palace 0.0

▼ Solution

A user characterizes as relevant the d1 document and as irrelevant the d3 document.

Calculate the Rocchio feedback using as alpha =1, beta = 1 and gamma = 1.

Round up the results to two decimals.

fire	0.6
timon	-0.3
discovery	0.6
sleeps	0.0
lion	0.3
king	0.25
jungle	0.3
simba	-0.6
mufasa	0.0
palace	0.0

go back to overview

Test execution

Information

Availability: Expired at 5/22/2025, 1:59 PM

Max. attempts: Unlimited

Results of this test are visible to administrators and tutors of this course.

Start test

Change log

Go to top