Assingment 02: Evaluation

Performance summary

✓ Assessed

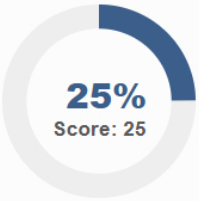
Success status	Score	Attempts
Undefined	25 of 100 points	1

Results

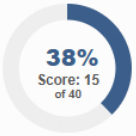
Course	**Web Retrieval SoSe 2025** ID: 4853531344 / 109642196056321
Test	**Assignment 02: Evaluation** ID: 4548625303

This are your test results

Duration	3h 38m 56s 5/2/2025, 5:23 PM - 5/7/2025, 6:12 PM
Answered	11 of 11 questions (100%)
Your score	25 of 100 points (25%)

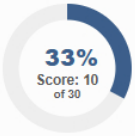


1. Knowledge Tasks (40 points) 5



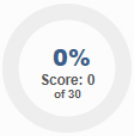
go to section

2. Practice Tasks (30 points) 4



go to section

3. Programming Tasks (30 points) 2



go to section

1. Knowledge Tasks (40 points) 15 of 40 points (38%)

True/false (15 points)

Status	Answered	
Your score	14 / 15	93%

Response

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

Please note:

- Maximum Overall Score -> 15 points
- Minimum Overall Score -> 0 points
- Incorrect Answer -> -0.5 point
- Unanswered -> 0 points

Unanswered	Right	Wrong	
☐	☐	☑	Web information resources are smaller than traditional information resources.
☐	☑	☐	Data in information retrieval is mostly semi-structured or unstructured.
☐	☐	☑	In information retrieval, the quality of retrieved documents does not matter.
☐	☑	☐	Indexing is used to avoid linearly scanning the texts for each query.
☐	☑	☐	Replacing human relevance assessment with LLMs can lead to biases specific to the LLM used.

☐	☐	☑	LLMs always agree with human assessments in relevance judgment tasks.
☐	☐	☑	LLMs can be used only for generating responses but not for training other models.
☐	☑	☐	The TREC initiative uses large test collections to promote research in improving search algorithms and information retrieval systems.
☐	☐	☑	Information Retrieval (IR) systems solely focus on retrieving textual data.
☐	☐	☑	Crowdsourcing relevance judgments is a practice that can completely replace expert evaluation due to its scalability.
☐	☐	☑	LLMs can perfectly handle the detection of misinformation within the documents they assess for relevance.
☐	☐	☑	User satisfaction and system usability are irrelevant in the evaluation of IR systems.
☐	☑	☐	Relevance feedback is a feature that adjusts the IR system's operations based on user input on the relevance of retrieved documents.
☐	☐	☑	The 'model in the loop' approach in IR completely eliminates the need for human judgment in relevance assessments.
☐	☐	☑	An LLM's training data does not influence its performance in relevance judgment tasks.
☐	☐	☑	The Cranfield Model is a system-oriented evaluation methodology using test collections and is considered outdated for modern IR systems.
☐	☑	☐	Query refinement is a process where the IR system modifies the user's original query to improve search results based on the system's understanding.
☐	☐	☑	LLMs that generate relevance judgments cannot learn from their mistakes or adjust over time.
☐	☐	☑	User-oriented evaluation focuses exclusively on the algorithmic efficiency of the IR system.
☐	☐	☑	All IR systems use a uniform model for data representation and query processing.
☐	☑	☐	Human-AI collaboration models suggest that LLMs should be used to support, not replace, human assessors.
☐	☐	☑	LLMs can currently understand and process user feedback to improve their relevance judgments.
☐	☐	☑	Relevance feedback from users is used to directly adjust the indexing of documents in an IR system.
☐	☑	☐	Automated relevance judgments by LLMs can surpass human performance in terms of speed and scalability.
☐	☐	☑	Human factors do not significantly influence the accuracy of labels in crowdsourced relevance judgments.
☐	☐	☑	Crowdsourcing relevance judgments is more expensive and slower than using trained professional assessors.
☐	☑	☐	The use of Amazon Mechanical Turk for crowdsourcing provides a platform for rapid and inexpensive relevance judgments.
☐	☑	☐	Crowdsourcing can be used to create training data for LLMs in relevance judgment tasks.
☐	☐	☑	The evaluation of IR systems never considers the user's satisfaction with the search results.
☐	☐	☑	The effectiveness of an IR system is solely based on its speed of retrieving documents.

▼ Solution

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

Please note:

- Maximum Overall Score -> 15 points
- Minimum Overall Score -> 0 points
- Incorrect Answer -> -0.5 point
- Unanswered -> 0 points

Unanswered	Right	Wrong	
☐	☐	☑	Web information resources are smaller than traditional information resources.
☐	☑	☐	Data in information retrieval is mostly semi-structured or unstructured.
☐	☐	☑	In information retrieval, the quality of retrieved documents does not matter.

☐	☐	☐	In information retrieval, the quality of retrieved documents does not matter.
☐	☑	☐	Indexing is used to avoid linearly scanning the texts for each query.
☐	☑	☐	Replacing human relevance assessment with LLMs can lead to biases specific to the LLM used.
☐	☐	☑	LLMs always agree with human assessments in relevance judgment tasks.
☐	☐	☑	LLMs can be used only for generating responses but not for training other models.
☐	☑	☐	The TREC initiative uses large test collections to promote research in improving search algorithms and information retrieval systems.
☐	☐	☑	Information Retrieval (IR) systems solely focus on retrieving textual data.
☐	☐	☑	Crowdsourcing relevance judgments is a practice that can completely replace expert evaluation due to its scalability.
☐	☐	☑	LLMs can perfectly handle the detection of misinformation within the documents they assess for relevance.
☐	☐	☑	User satisfaction and system usability are irrelevant in the evaluation of IR systems.
☐	☑	☐	Relevance feedback is a feature that adjusts the IR system's operations based on user input on the relevance of retrieved documents.
☐	☐	☑	The 'model in the loop' approach in IR completely eliminates the need for human judgment in relevance assessments.
☐	☐	☑	An LLM's training data does not influence its performance in relevance judgment tasks.
☐	☐	☑	The Cranfield Model is a system-oriented evaluation methodology using test collections and is considered outdated for modern IR systems.
☐	☑	☐	Query refinement is a process where the IR system modifies the user's original query to improve search results based on the system's understanding.
☐	☐	☑	LLMs that generate relevance judgments cannot learn from their mistakes or adjust over time.
☐	☐	☑	User-oriented evaluation focuses exclusively on the algorithmic efficiency of the IR system.
☐	☐	☑	All IR systems use a uniform model for data representation and query processing.
☐	☑	☐	Human-AI collaboration models suggest that LLMs should be used to support, not replace, human assessors.
☐	☑	☐	LLMs can currently understand and process user feedback to improve their relevance judgments.
☐	☐	☑	Relevance feedback from users is used to directly adjust the indexing of documents in an IR system.
☐	☑	☐	Automated relevance judgments by LLMs can surpass human performance in terms of speed and scalability.
☐	☐	☑	Human factors do not significantly influence the accuracy of labels in crowdsourced relevance judgments.
☐	☐	☑	Crowdsourcing relevance judgments is more expensive and slower than using trained professional assessors.
☐	☑	☐	The use of Amazon Mechanical Turk for crowdsourcing provides a platform for rapid and inexpensive relevance judgments.
☐	☑	☐	Crowdsourcing can be used to create training data for LLMs in relevance judgment tasks.
☐	☐	☑	The evaluation of IR systems never considers the user's satisfaction with the search results.
☐	☐	☑	The effectiveness of an IR system is solely based on its speed of retrieving documents.

☑ Evaluation measures of WIR (2 points)

Status	Answered	
Your score	1 / 2	50%

Response

Response

Which of the following is considered one of the evaluation measures of WIR systems?

Please note:

- Maximum Overall Score -> 2 points
- Minimum Overall Score -> 0 points
- Incorrect Answer -> -1 points
- Unanswered -> 0 points

☑ Relevance

☑ Speed of response

☑ Completeness

☐ Separation of concerns

☐ Maintainability

▼ Solution

Which of the following is considered one of the evaluation measures of WIR systems?

Please note:

- Maximum Overall Score -> 2 points
- Minimum Overall Score -> 0 points
- Incorrect Answer -> -1 points
- Unanswered -> 0 points

☑ Relevance

☑ Speed of response

☐ Completeness

☐ Separation of concerns

☐ Maintainability

Crowdsourcing Relevance Judgements (7 points)

Status	Answered	
Your score	0 / 7	0%

Response

Discuss the implications of label accuracy issues in crowdsourced tasks and strategies to enhance label quality.

```
Label accuracy in crowdsourced relevance judgment tasks is a major concern; if labels are inaccurate, any
machine learning models or information retrieval systems built on top are almost rendered invalid. A
crowdsourcing platform typically draws from a large pool of workers with a vast range of expertise,
motivation levels, and conscientiousness. In turn, label quality varies immensely: from really high to
really low to downright noisy and misleading. This problem is tenfold in tasks that require any nuance,
e.g., semantic similarity or subjective relevance, where each worker may interpret things differently.

Label inaccuracy may result in model degradation due to incorrect ground truth being used in training or
evaluation. In turn, such a model would generally be less accurate, yield biased results, and provide a
really poor user experience. Moreover, low-quality labels hinder the fair evaluation of algorithmic
improvements, hence blocking research progress and further development.

There are several ways to ensure label quality in crowdsourced tasks. These include the use of gold
questions, which are questions with an a priori known answer to test worker credibility; qualification
tests for workers; and training materials provided to workers to ensure they understand the task
requirements. Another mechanism is redundancy: collecting multiple judgments per item and aggregating them
with majority voting or weighted methods, such as expectation-maximization.
```

209 words (max. 250)

▶ Solution

Evaluation Methods (7 points)

Status	Answered	
Your score	0 / 7	0%

Response

Explain the differences between system-oriented and user-oriented evaluation methods in IR systems. What are the advantages and disadvantages of each?

```
Information Retrieval (IR) systems can be evaluated through system-oriented versus user-oriented evaluation
methods, the two differing in focus, benefits, and limitations.

The system-oriented evaluation looks at performance through some objective, quantifiable metrics-the well-
known ones being precision, recall, F1-score, MAP, etc. It tests the IR system on some predefined set of
```

queries and relevance judgments (test collection), without involving any real users. Within these usually benchmark tasks like TREC, this method allows fair comparisons between systems, is inexpensive, and can be easily automated not to mention that it can be repeated.

Nonetheless, the downside comes from this abstraction from the real world of use. System-oriented evaluations tend to disregard anything concerning how users behave or feel, disregarding the context of use. That is to say, a system might turn out very well for the metrics but will fail in satisfying the actual requirements of its users because of an interface that doesn't support users well or queries meaningfully misunderstood.

The user-oriented evaluation investigation, on the other hand, concerns itself largely with actual persons or groups of persons working with the system. The assessment leans on a variety of qualitative and quantitative methods and metrics, ranging from user studies and A/B testing to questionnaires and interviews-for user satisfaction, task completion time, and perceived relevance.

Some of the advantages of this approach reveal the real-world use context, which is essential for grasping the meaning behind how the system serves human information needs.

246 words (max. 250)

▸ Solution

Model in the Loop (9 points)

Status	Answered	
Your score	0 / 9	0%

Response

Explain the concept of "model in the loop" as it pertains to LLMs and human assessors. What are the potential benefits and drawbacks of this approach?

The term "model in the loop" refers to the integration of a large language model (LLM) into the human evaluation pipeline, whereby the model buttresses or interacts with human evaluators in some annotation, labeling, or decision-making task. Instead of data being entirely manually evaluated or solely dependent on the model's outputs, a collaborative loop is formed, with humans guiding the model and the model supporting or augmenting human judgment.

The LLM can issue candidate answers, explanations, summaries, or labels, while the human may accept, modify, or reject the suggestions. The interaction enhances scalability for evaluation tasks while maintaining some supervision from human quality control.

Benefits:
Efficiency: A labeling model can speed up the procedure, suggesting the likely answers so that human assessors have less time and cognitive burden.
Consistency: LLMs can ensure a predetermined level of consistency since they reduce variation in judgments across different annotators.
Skill amplification: Lesser-skilled or less-experienced assessors may achieve higher levels with less model assistance, thereby improving the overall quality of annotations.

Drawbacks:
Bias propagation: If the suggestions made by the model are not great, the human assessors might acquiesce to an erroneous suggestion without spotting its error, hence inducing confirmation bias.
Over-trust: Humans become complacent and replay outputs of the models instead of critically engaging with them and exercising their own judgment.
Transparency and accountability: This creates difficulty in[interruption]

226 words (max. 250)

▸ Solution

‹ go back to overview

2. Practice Tasks (30 points) 10 of 30 points (33%)

••• Query Expansion (5 points)

Status	Answered	
Your score	0 / 5	0%

Response

Given the original query "AI in medicine", manually generate an expanded query using synonyms and related terms.

Expanded Query: ("artificial intelligence" OR "

▾ Solution

Given the original query "AI in medicine", manually generate an expanded query using synonyms and related terms.

Expanded Query: gap

••• Synonym Expansion (5 points)

Status	Answered	
Your score	0 / 5	0%

Response

Generate a synonym list for the query "smartphone features" to use in query expansion.

Synonyms for "smartphone": mobile phone, cell phone,

Synonyms for "features": specifications, specs, capa

▼ Solution

Generate a synonym list for the query "smartphone features" to use in query expansion.

Synonyms for "smartphone": cell phone | mobile phone, iphone

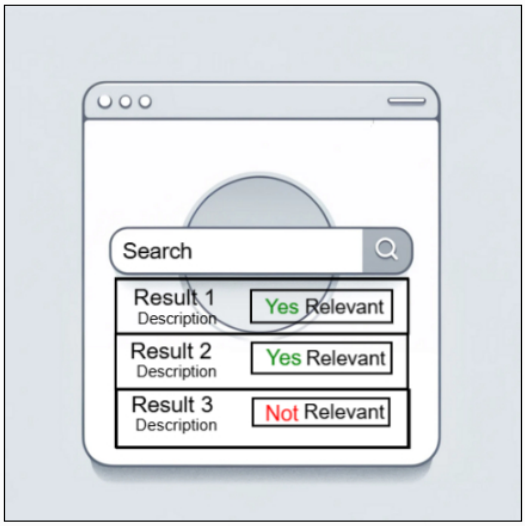
Synonyms for "features": specifications | specs, characteristics

Relevance Feedback (10 points)

Status	Answered	
Your score	0 / 10	0%

Response

Sketch a simple user interface for a search engine that includes relevance feedback options.



▶ Solution

Boolean Expression (10 points)

Status	Answered	
Your score	10 / 10	100%

Response

Translate the natural language query "Find documents that discuss 'data mining' but exclude 'deep learning' unless 'neural networks' is also mentioned" into a Boolean logic expression using AND, OR, NOT.

- ○ ("data mining") AND NOT ("neural networks") OR ((deep learning) AND (neural networks))
- ☑ ("data mining") AND (NOT "deep learning" OR "neural networks")
- ○ ("data mining") OR (NOT "deep learning" OR "neural networks")
- ○ ("data mining") AND (NOT "neural networks" OR "deep learning")
- ○ ("data mining") AND NOT ("deep learning") OR (("deep learning") AND ("neural networks"))

▶ Solution

‹ go back to overview

3. Programming Tasks (30 points) 0 of 30 points (0%)

Plot Document Frequencies (15 points)

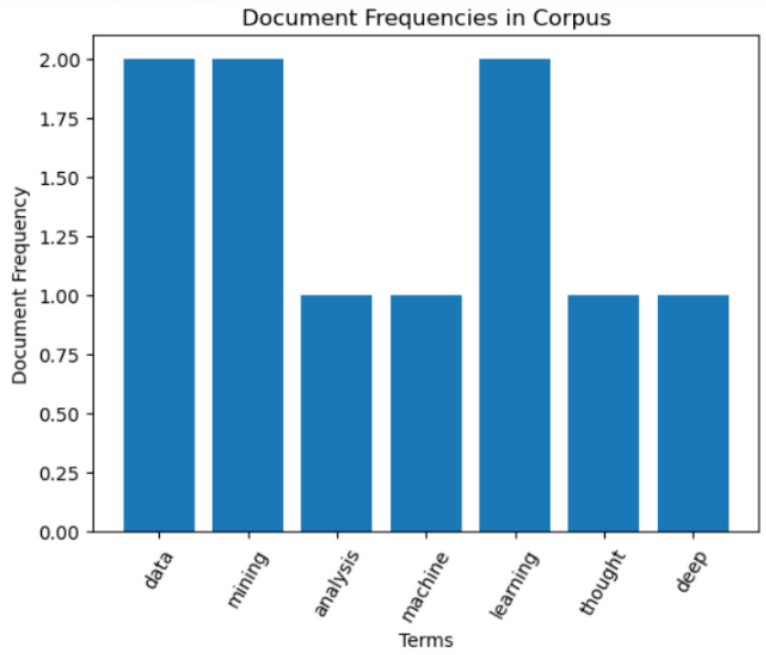
Status	Answered	
Your score	0 / 15	0%

Response

Write a Python script to plot the document frequencies (the number of documents containing a particular term) of terms in a corpus. Use as an example of documents the following:

```
# Example documents
documents = ["data mining data analysis", "machine learning data mining", "deep learning deep thought"]
```

Hint: you should plot something like:




```
import matplotlib.pyplot as plt
from collections import defaultdict

# Example documents
documents = ["data mining data analysis", "machine learning data mining", "deep learning deep thought"]

doc_freq = defaultdict(int)

for doc in documents:
    unique_terms = set(doc.lower().split())
    for term in unique_terms:
        doc_freq[term] += 1

terms = [term for term, freq in doc_freq.items()]
frequencies = [freq for term, freq in doc_freq.items()]

plt.figure(figsize=(10, 6))
plt.bar(terms, frequencies, color='skyblue')
plt.title("Document Frequencies in Corpus")
plt.xlabel("Terms")
plt.ylabel("Document Frequency")
plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```

89 words

▸ Solution

Word cloud (15 points)

Status	Answered		
Your score	0 / 15		0%

Response

Generate a word cloud from a collection of documents to visualize the most common terms.

Hint: Use the *wordcloud* package

```
from wordcloud import WordCloud
```

and the following example of documents to test your script:

```
docs = ["Python programming language", "Python and data analytics", "Programming in Python"]
```

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

# Example documents
docs = ["Python programming language", "Python and data analytics", "Programming in Python"]

text = " ".join(docs)
wordcloud = WordCloud(width=800, height=400, background_color='white').generate(text)
```

```
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.title("Word Cloud of Document Terms")
plt.show()
```

56 words

▸ Solution

go back to overview

Test execution

Information

Availability: Expired at 5/8/2025, 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