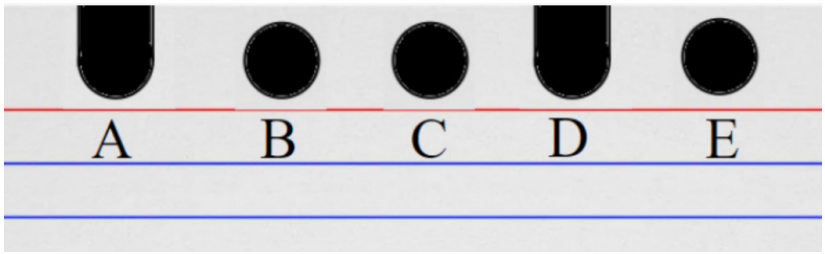
CSCI 572 Midterm Rubrics

Total Marks	25 points
Bonus	5 points
Late Penalty	< 30 mins - 2.5 marks >= 30 mins - 5 marks

Q1: Being the **search 'geek' that you are, you decide to 'index' your collection of 1000 books, using 1000 index cards and punching holes in them and cutting out some holes, like so:**



In the above, which shows the index card for one of your 1000 books, your book belongs to categories B,C,E [not punched out], and not to A,D [punched out]. Simple enough. You have 999 more such cards, where each card has 0,1,2,3 or 4 holes punched out [but not all 5, duh!!].

> **a (2 points). You are given a single long knitting needle (or chopstick!) that can pass through holes in the entire stack. Using it, would you do NOT((NOT A) OR (NOT B))?**

[SOLUTION]
Applying De Morgan's Law:
NOT((NOT A) OR (NOT B)) => A AND B
To pick out index cards for books in category A and B:
- Stack all the index cards
- Slide the knitting needle through the hole A
- Pull the needle outside (sideways), all cards that are stuck to needle are of category A
- Put these cards aside now (call these cards **Set1**)
- Now, slide the knitting needle through the hole B of **Set1**
- Pull the needle outside, all cards that are stuck to needle are of category B (as well as A)
- Put these cards aside now (call these cards Set2)
- cards in Set2 include categories A and B

In other words, we do this sequentially - first we do NOT A by letting the NOT A cards drop. Then we do NOT B on what remains, letting those NOT B cards drop. The dropped pile is now, NOT A OR NOT B because it is a merge of NOT A, NOT B. The NOT of that is what is stuck to the needle, ie NOT (NOT A OR NOT B) [which is actually, A AND B - shhh, lol].

[RUBRICS]

- +1 point: for correct expression conversion
- 0 point: for incorrect or intermediate expression
- +1 point: correct or similar explanation
- 0 point: for incorrect or intermediate explanation

b (1 point). Given two rods, how would you do 'A and B'?

[SOLUTION]
To pick out index cards for books in category A and B:
- Stack all the index cards
- Place two rods in holes A and B
- Pull both the rods outside together, all cards that are stuck in rods are of category A and B

[RUBRICS]

- +1 point: correct or similar explanation
- 0 point: for incorrect or intermediate explanation

c (1 point). What is the following Boolean expression equivalent to? NOT((NOT A) AND (NOT B))

[SOLUTION]
Applying De Morgan's Law:
NOT((NOT A) AND (NOT B)) => A OR B

[RUBRICS]

- +1 point: for final correct expression
- 0 point: for incorrect or intermediate expression

Q2: Searching a database (eg. using SQL, to find a list of students with GPA > 3.9) is both similar, and different, compared to a web **search (eg. 'high-performing schools in my area'). How are they similar, how are they different? (3 points)**

[SOLUTION]

Similarities:

Indexing: Both rely on indexing mechanisms to improve search efficiency and speed. Databases and search engines create indexes to quickly locate relevant data.

Filtering and Sorting: Both offer options to filter, sort, and refine search results to match user preferences.

Differences:

1. Structure of Data:

Database Search: Involves search on structured data with well-defined schemas.
Web Search: Involves search on unstructured or semi-structured data from the web.

2. Query Type
Database Search: Utilizes structured query languages for precise data retrieval.
Web Search: Uses natural language queries to find relevant content on the internet.

3. Index Type
DB search: uses B-Tree index, hash index, bitmap index, quadtree index, etc.
Web search: uses inverted (TF-IDF) index.

[RUBRICS]
- +1 point: for mentioning two valid similarities [½ mark for each similarity]
- 0 point: if valid similarities are not mentioned
- +2 point: for two of the three valid differences [1 mark for each difference]
- 0 point: if valid differences are not mentioned

Q3: Deduplication involves fingerprinting content so that we will know to avoid re-indexing it in our **search engine.**

a (1 point). Why not use MD5 (or SHA etc), a classic hashing algorithm, for this purpose?

[SOLUTION]
Documents that are nearly identical should have nearly similar fingerprints that differ only in a small # of bits. In other words, similar inputs should lead to similar outputs (hash values). But MD5 (or SHA), do not have this property, i.e even a tiny change in the input leads to a huge change in the output. Thus, we cannot use the classic hashing algorithms for this purpose.
[RUBRICS]
- +1 point: for any mention of acceptable answer similar to solution
- 0 point: for vague or incorrect explanation

b (2 point). What is the algorithm that actually does get used for this? Explain in a few lines how it works.

[SOLUTION]
SimHash is a dimension reduction technique - it maps a set of weighted features (contents of a document) to a low dimensional fingerprint, eg. a 64-bit word. The basic idea is to obtain a k-bit fingerprint for each document. A pair of documents are near duplicate if and only if fingerprints are at most k-bits apart. Instead of using permutations and probability we use Sim Hash.
Documents D1 and D2 are near duplicates iff

Hamming-Distance(Simhash (D1), Simhash(D2)) <= K

[RUBRICS]

- +1 point: for stating the algorithm
- 0 point: for incorrect algorithm
- +1 point: for explaining SimHash aptly with Hamming Distance
- 0 point: for vague or incorrect elaboration of SimHash

Q4: A 'rogue' **search engine crawler would simply disregard robots.txt [which specifies which parts of a site are off-limits]. Your webserver would end up serving such crawlers, possibly valuable resources that your site is hosting (that you did not mean to serve].**

How would you set up your server to 'roguely' respond to such calls? In other words, how to 'fight fire with fire'?

[POSSIBLE SOLUTION]

Acceptable methods include any mention of filters, **fake data (incorrect, including spam pages)**, delay responses, caching, IP addresses, crawler detection, authentication systems, firewalls, rate limiting, traps, ignoring requests, tags, and dynamic methods. There are other possible creative methods not listed here. Please use your judgment in determining those methods.

[RUBRICS]

- +2 point: for any mention of acceptable methods.
- -2 point: If the answer is vague or doesn't involve the above methods.

Q5: a (1 point). How did 'search** engine advertising' used to work, in the early days of the web?**

[POSSIBLE SOLUTION]

Acceptable answers include any mention of static, cost, banner, search engine, user search, per click, ad auction, Adsense, Google, website advertisement, affiliate marketing, social media marketing, cookies, brokers, bidding, and user profile. There are other possible reasonable answers not listed here. Please use your judgement in determining those answers. Note that students have very different interpretations of "early days", "now", and "future" which affect their answers in (a,b,c).

[RUBRICS]

- +1 point: for any mention of acceptable answers in (a).
- -1 point: If the answer is vague or doesn't involve the above answers.

b (2 points). How does it work now?

[POSSIBLE SOLUTION]

Acceptable answers include any mention of Google, website advertisement, affiliate marketing, social media marketing, cookies, brokers, bidding, user profile, dynamic, personalization, ChatGPT, BART, BARD, and recommendation. There are other possible reasonable answers not listed here. Please use your judgement in determining those answers. Note that students have very different interpretations of "early days", "now", and "future" which affect their answers in (a,b,c).

[RUBRICS]

- +2 point: for any mention of acceptable answers in (a).
- -2 point: If the answer is vague or doesn't involve the above answers.

c (1 point). Given the rise of AI-powered summarization, how is such advertising likely to work in the future (hint: think how various sites do this now!).

[POSSIBLE SOLUTION]

Acceptable answers include any mention of search engine, user search, Google, website advertisement, affiliate marketing, social media marketing, cookies, user profile, dynamic, personalization, ChatGPT, BART, BARD, and recommendation. There are other possible reasonable answers not listed here. Please use your judgement in determining those answers. Note that students have very different interpretations of "early days", "now", and "future" which affect their answers in (a,b,c).

A related answer would be, that the ad might likely be served as a video that the user is forced to watch, prior to the AI summarization being displayed - similar to ads on YouTube, CNN, etc.

[RUBRICS]

- +1 point: for any mention of acceptable answers in (a).
- -1 point: If the answer is vague or doesn't involve the above answers.

Q6: a (1 point). What is the purpose of 'discounting', in DCG for **search results ranking?**

[POSSIBLE SOLUTION]

(a) Acceptable answers include any mention of evaluation metric, penalize, relevant search, order, ranked low, and lower position.

The idea is to 'punish' results that are further down the line from the top, even if they are relevant - because we want the relevant results to be at/near the top.

[RUBRICS]

- +1 point: for any mention of acceptable answers.
- -1 point: If the answer is vague or doesn't involve the above answers.

b (1 point). The discounting factor is typically, 1/log2(rank). How would someone 'poison' the ranking system, to reward relevant results further down the list of retrieved documents?

[POSSIBLE SOLUTION]

(b) Acceptable answers include any mention of rewarding relevant search in lower position, rewarding non-relevant search in higher position, penalizing relevant search in higher position, and penalizing non-relevant search in lower position. Acceptable mathematical formulas include any variation of acceptable answers above.

Change the ranking function to be directly proportional to the rank value (not inversely proportional like DCG is), eg. rank, rank^2, n^rank (eg. 2^rank, e^rank etc).

[RUBRICS]

- +1 point: for any mention of acceptable answers.
- -1 point: If the answer is vague or doesn't involve the above answers.

c (1 point) What is the future of ranking?

[POSSIBLE SOLUTION]

(c) Acceptable answers include any mention of relevant search have a higher rank, consider other relevant factors, personalization, dynamic, human in the lop, reinforcement learning, deep learning, AI/ML system, NLP model, ChatGPT, BARD, BART, and user preference. There are other possible reasonable answers not listed here. Please use your judgement in determining those answers.

Ranking wouldn't matter as much, given that the raw links (search results) might not be displayed in the form of a list from top to bottom of a page - the results would be summarized by an LLM such as ChatGPT or Bard.

[RUBRICS]

- +1 point: for any mention of acceptable answers.
- -1 point: If the answer is vague or doesn't involve the above answers.

Q7: Google uses humans, to evaluate the quality of **search, why? Note that LLMs are trained this way too, loosely speaking.**

[POSSIBLE SOLUTION]

1. Human evaluators help search engine developers validate the effectiveness of new algorithms and updates. They compare the search results produced by different algorithms, assessing which ones are more relevant and useful to users.
2. They bring diverse perspectives and domain expertise to the evaluation process. They can help identify nuances and context that automated algorithms might miss, ensuring that the search results are more comprehensive and accurate.
3. Human evaluators can assess the quality of search results for complex and ambiguous queries that are challenging for algorithms. This feedback helps improve the handling of queries that involve context and intent.
4. Human evaluators can evaluate how well the search results meet user needs, taking into account factors like relevance, diversity, and user satisfaction.
5. Human evaluators can assess whether search results adhere to these guidelines, which can help filter out low-quality or inappropriate content.

In other words, humans are the ultimate judges, as opposed to algorithms that compute via data (because web page ie. document content, is a form of data].

[RUBRICS]

- +1 point: If there is ANY mention/interpretation of the above ways.
- -1 point: If the answer is vague or doesn't involve the above techniques.

Q8: Both precision (P) and recall (R) are useful measures, for characterizing the results of a **search. What are a couple of different ways of doing so? Which is better, and why?**

[POSSIBLE SOLUTION]

1. The F1 score is a single metric that combines both precision and recall into one value. It is the harmonic mean of precision and recall and is calculated using the formula: F1 = 2 * (Precision * Recall) / (Precision + Recall).

2. The precision-recall curve is a graphical representation that shows how precision and recall change as a function of a threshold or decision boundary used to classify items in a search result. Different thresholds can be applied to evaluate the trade-off between precision and recall.

We could average them (arithmetic mean, ie AM), or compute their harmonic mean (HM). HM is better because it punishes extreme values of P or R. Eg. P=0, R=1, AM will be 0.5, but HM would give us 0, which is what we'd want. We can accept almost any answer that states HM is better than AM, but with simpler/vague explanations.

[RUBRICS]

- +1 point: for each different way of using precision and recall.
- -1 point: If the answer is vague or employs contradicting use of precision and recall.

Q9: a (1 point). Indexing of videos (eg by YouTube), paradoxically doesn't involve the video itself! In what sense?

[POSSIBLE SOLUTION]

Video indexing is more about indexing the information and context around the video rather than the video itself. The metadata, transcriptions, and user interactions are used to make videos discoverable to users.

[RUBRICS]

- +1 point: If there is ANY mention of indexing using video transcript or meta data or recommendation algorithms based on user interaction.
- -1 point: If the answer is vague or doesn't involve the above techniques.

b (2 points). How would you index videos, using 'the video itself'? Note that this would lead to much better **search.**

[POSSIBLE SOLUTION]

"Content-based video indexing" involves analyzing the actual visual and audio content of the video to extract features and information. It can lead to more accurate search results. Possible ways to bring this about:

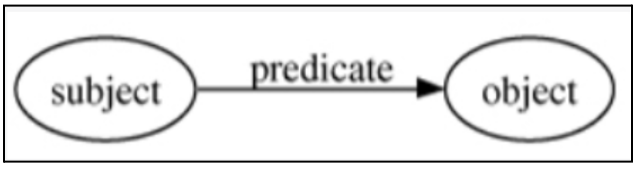
1. Audio and video feature extraction using computer vision
2. Categorizing using face recognition or object detection.
3. Using NLP and sentiment analysis for theme detection

So overall, 'ML-based content analysis' [examing the video itself] would be better.

[RUBRICS]

- +1 point: for each mention of complex data structure or AI/ML techniques for content analysis.
- -1 point: If the answer is vague or doesn't involve the above techniques.

Q10: An RDF triple, aka semantic triple, has a very simple form:



a (1 point). In the context of **search, where are these useful?**

[POSSIBLE SOLUTIONS]

1. RDF helps in expanding search results because they enable linking of data from various sources on the web. Meaning, It establishes connections between different datasets, making it easier to discover related information during a search.
2. Subjects, predicates, and objects can be treated as facets that users can use to filter and refine search results. This enables more precise and user-friendly search experiences.
3. **Search engines can infer relationships that are not explicitly stated in the data, leading to more comprehensive and accurate search results.**
4. RDF triples allows users to find information related to their query from different knowledge domains.
5. RDF triples are helpful for integrating data from different sources, even if they have varying structures and formats. This is particularly valuable for search engines that need to pull information from diverse data repositories.

[RUBRICS]

- +1 point: If there is ANY mention of tailored searches, cross-domain search results, linking datasets or search filtering.
- -1 point: If the answer is vague or doesn't involve search optimization using RDF.

b (1 point). Such triples are usually expressed as XML. XML syntax looks like this:

```
▼<note>
   <to>Tove</to>
   <from>Jani</from>
   <heading>Reminder</heading>
   <body>Don't forget me this weekend!</body>
 </note>
```

List an example RDF triple using XML.

[POSSIBLE SOLUTION]

```
<book>
        <title>Harry Potter and the Half-Blood Prince</title>
        <author>J K Rowling</author>
        <date>July 16 2005</date>
</book>
```

[NOTE]

Crowdmark automatically parses XML to text. So the XML tags of the student's solution would not be visible. Please DO NOT dock points for that. As long as a subject, predicate and an object exists in the text, award points

(Sample Text - Harry Potter and the Half-Blood Prince J K Rowling July 16 2005)

[RUBRICS]

- +1 point: Subject-predicate-object is established in the text.
- -1 point: Incorrect unrelated predicate between subject and object.

c (1 point). Alternatively, such triples can be in JSON as well. Sample JSON:

```
{
    "fruit": "Apple",
    "size": "Large",
    "color": "Red"
}
```

List a different sample RDF, as JSON.

[POSSIBLE SOLUTION]

```
{
        “Title”: “Harry Potter and the Half-Blood Prince”
        “Author”: “J K Rowling”
        “Date”: “July 16 2005”
}
```

[RUBRICS]

- +1 point: Subject-predicate-object is established in JSON format.
- -1 point: Incorrect unrelated predicate between subject and object.