

**CS 4250 – Assignment #1**

**Maximum Points: 100 pts.**



Bronco ID: | 0 | 1 | 6 | 2 | 0 | 7 | 8 | 1 | 5 |

Last Name: Thai

First Name: Nhan

**Note 1:** Your submission header must have the format as shown in the above-enclosed rounded rectangle.

**Note 2:** Homework is to be done individually. You may discuss the homework problems with your fellow students, but you are NOT allowed to copy – either in part or in whole – anyone else’s answers.

**Note 3:** Your deliverable should be a .pdf file submitted through Gradescope until the deadline. Do not forget to assign a page to each of your answers when making a submission. In addition, source code (.py files) should be added to an online repository (e.g., github) to be downloaded and executed later.

**Note 4:** All submitted materials must be legible. Figures/diagrams must have good quality.

**Note 5:** Please use and check the Canvas discussion for further instructions, questions, answers, and hints. The bold words/sentences provide information for a complete or accurate answer.

1. [6 points]. Despite the current advances in the field, the primary focus of Information Retrieval is still on text and text documents. Based on this information, answer the questions below:

* 1. [4 points]. Why is querying a database table easier compared to querying text documents? For full marks, **list** and **explain** at least **two** **factors** to elaborate your answer.

Answers: The first reason that querying a database table is easier than querying a text document is because the text document is unstructured. We easily find that through articles or news where there is lots of information contained within a paragraph. On the other hand, databases have well defined attributes and indexes which we clearly distinguish the data types and entries. For example, a list of students in CS4250 courses will have each column define last names, first names, GPA, scores, student ID, etc… Because of that, database tables make things simpler to search and obtain information. Another reason is that in text paragraph, we might find that many words have synonymous or meaning the same thing but different words. This factor makes it difficult to compare a query. On the contrary, database tables are easier to compare to a query because they have clear and well-defined semantics to find matches.

* 1. [2 points]. Explain how **text** has been **used** by Information Retrieval researchers to compare multimedia documents and how this **scenario** is currently being **changed**.

Answers: Researchers extract text from multimedia documents through captions, keywords, tags, relevance to index, search and query for contents. Besides, advanced technologies nowadays such as AI and Machine Learning models will aid humans in extracting many contents from multimedia like video features, text, image, music, speech. Thanks to them, comparing multimedia documents becomes faster without relying only on text representations.

1. [10 points. 2 points each]. A search engine is the practical **application** of Information Retrieval techniques to large-scale text collections. **Explain** the scope of the different search engine applications and give **one practical example**.

* 1. Web search engine.

Answers: The most widely used type of search engine, designed to index and retrieve information from the World Wide Web.

Example: Google is a prominent web search engine that provides users with links to websites, images, and videos in response to their queries.

* 1. Vertical search engine.

Answers: Focuses on specific sectors or industries, limiting its search scope to a particular topic or niche.

Example: Yelp is a vertical search engine that specializes in helping users find local businesses, particularly restaurants, by providing reviews and ratings

* 1. Enterprise search engine.

Answers: Enable efficient search across internal documents, databases, emails, and other proprietary resources for organizations and businesses.

Example: Microsoft SharePoint includes enterprise search functionality that allows employees to search across internal resources like reports, presentations, and emails within an organization.

* 1. Desktop search engine.

Answers: A tool that helps users find files, documents, emails, or applications stored locally on their computers.

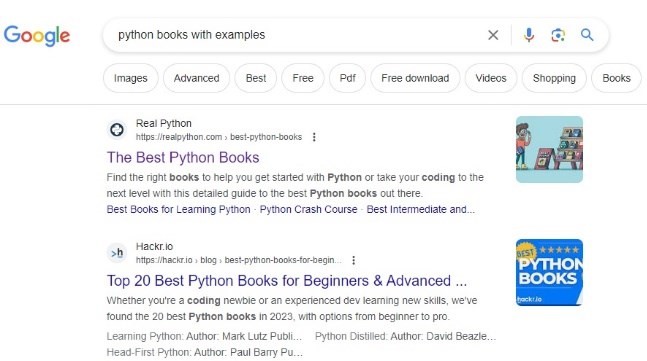
Example: Windows Search is a desktop search engine that enables users to find files, folders, and applications on their local computers through a keyword search.

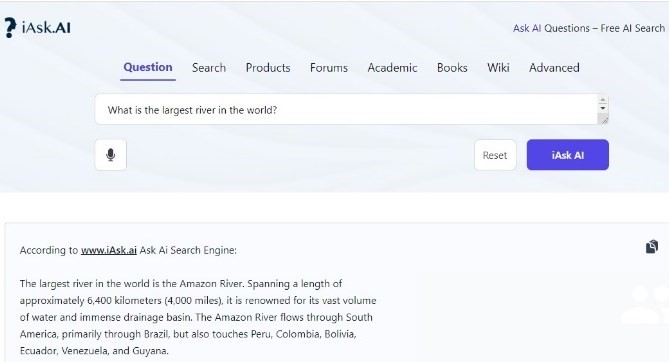
* 1. Finally, **explain** how peer-to-peer search engines differ from the other previous types.

Answers: Unlike the previous types of search engines that rely on a centralized server to index and retrieve data, **peer-to-peer (P2P) search engines** operate in a decentralized manner. P2P search engines differ from other types in that there is no centralized index or database; instead, participants share and retrieve content directly from each other's systems.

Example: BitTorrent, which uses peer-to-peer technology, allows users to search for and download files shared by others in the network without using a central server.

1. [8 points – 2 points each]. **Identify** and **explain** the following **tasks** that involve Information Retrieval.

 a.

 c. d.

Answers:  
a) Classification: This process involves assigning appropriate labels to documents or files. In this example, images or files are being categorized and systematically organized.

b) Ad-hoc search: This method entails retrieving relevant documents based on a specific query. In this case, a user enters a custom query into Google, which generates a ranked list of pertinent results.

c) Question answering: This model focuses on delivering precise answers to user questions. Here, a user poses a query to an AI system, which responds with a concise and accurate answer.

d) Filtering: This approach involves identifying relevant user profiles for newly available content. For example, Netflix analyzes user preferences to recommend appropriate movies.

1. [8 points. 2 points each]. A retrieval model is a formal representation of the process of matching a query and a document, forming the basis of ranking algorithms that sort documents according to their relevance. Considering that **relevance** is one of the big issues for Information Retrieval research, answer the questions below.

* 1. Explain why **topical relevance** and **user relevance** should be considered during search.

Answers: Topical relevance guarantees that search results align with the subject or theme of the query, avoiding unrelated information. On the other hand, user relevance analyzes individual preferences, context, and requirements to deliver related personal results. Therefore, Ignoring these factor can lead to unsatisfactory search outcomes.

* 1. Considering **only topical** but **not user relevance**, give a hypothetical example of a good search engine output based on a query. Provide the **user** **profile**, **query,** and **document returned**.   
     Answers:  
     + User profile: Nhan, Software Developer interested in Machine Learning, Database System.

+ Query: “Optimizing database queries for machine learning applications”

+ Result: *Title*: "Best Practices for Optimizing Database Queries in Machine Learning Pipelines"  
*Summary*: This guide explores techniques for optimizing database queries specifically in machine learning workflows. It covers query indexing, efficient data retrieval, and reducing latency in large-scale Machine Learning models.

* 1. Considering **only user** but **not topical relevance**, give a hypothetical example of a good search engine output based on a query. Provide the **user** **profile**, **query,** and **document returned**.

Answers:  
+ User profile: Nhan, Software Developer interested in Machine Learning, Database System.

+ Query: “Machine Learning internships near me”

+ Result: “Popeyes Chicken Louisiana – 2.3 miles away”

* 1. Considering both **topical** and **user relevance**, give a hypothetical example of a good search engine output based on a query. Provide the **user** **profile**, **query,** and **document returned**.

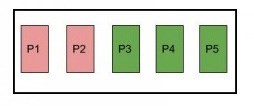
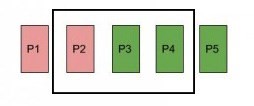
Answers:  
+ User profile: Nhan, Software Developer interested in Machine Learning, Database System.

+ Query: “Machine Learning internships near Cal Poly Pomona”

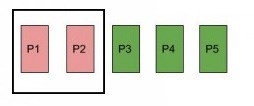
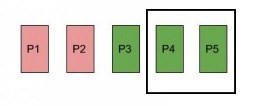
+ Result: “Lists of internships opportunities in Pomona”

1. [8 points. 2 points each]. Another core issue for information retrieval is evaluation. Two measures that have been extensively used for comparing search engines are precision and recall. Given the scenarios below, calculate the **precision** and **recall** of the corresponding search engines. Hint: green and red colors show the relevant and irrelevant documents respectively for a given query, and the black rectangles show the retrieved documents. Show your **math** (fraction and final value) for full marks.

* 1. b.



c. d.



Answers:

a)

b)

c)

d)

1. [20 points]. Derive the PageRank values of the four pages below **until the third iteration**. Show your **math** for full marks. Hint: **no Dumping Factor** needs to be considered.

A

B

C

D

Using matrix multiplication, we will have the initial Node

Node A = [ 0, 1/2, 1/2, 0 ]

Node B = [ 0, 0, 1/2, 1/2 ]

Node C = [ 0, 1/2, 0, 1/2 ]

Node D = [ 0, 0, 1/4, 0 ]

Using each node linked vectors as a column, we will have a 4x4 matrix looks like:

First Iteration: A = B = C = D = ¼

Second Iteration:  
Third Iteration:

Iteration Table:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Node | Initial | 1st iteration | 2nd iteration | 3rd iteration | Page Rank |
| A | 1/4 | 0 | 0 | 0 | **4** |
| B | 1/4 | 1/4 | 5/32 | 3/32 | **3** |
| C | 1/4 | 5/16 | 3/16 | 19/128 | **2** |
| D | 1/4 | 1/4 | 9/32 | 11/64 | **1** |

1. [20 points]. Index term weights reflect the relative importance of words in documents and are used to compute scores for ranking. One of the most common types used in retrieval models is known as tf-idf. Derive the tf-idf document-term matrix according to the data below. Requirements: 1) you must conduct **stopword removal** (pronouns/conjunctions) and **stemming** before indexing the terms, 2) place the terms in the matrix following the sequence of their occurrences in the documents from 𝑑1 to 𝑑3, 3) show your **math** for tf, idf, and tf-idf for full marks.

𝑑1 = “I love cats and cats”.

𝑑2= “She loves her dog”.

𝑑3= “They love their dogs and cat”.

Answers:

Stopwords: pronouns, conjunctions

With stopword removal + stemming, we will have:  
d1 = “love”, “cat”, “cat”

d2 = “love, “dog”

d3 = “love”, “dog”, “cat”

Table for d1:

|  |  |
| --- | --- |
| Term | Term Count |
| love | 1 |
| cat | 2 |

tf(“love”, d1) = 1/3 = 0.333

df(“love”, D) = 3

idf(“love”, D) = log(|D| / df(“love”, D)) = log(|3| / 3) = 0

tf-idf(“love”, d1, D) = tf(“love”, d1) \* idf(“love”, D) = 0.333 \* 0 = 0

tf(“cat”, d1) = 2/3 = 0.667

df(“cat”, D) = 2

idf(“cat”, D) = log(|D| / df(“cat”, D) = log(|3| / 2) = 0.176

tf-idf(“cat”, d1, D) = tf(“cat”, d1) \* idf(“cat”, D) = 0.667 \* 0.176 = 0.117

c)

tf(“dog”, d1) = 0/3 = 0

df(“dog”, D) = 2

idf(“dog”, D) = log(|D| / df(“dog”, D)) = log(|3| / 2) = 0.176

tf-idf(“dog”, d1, D) = tf(“dog”, d1) \* idf(“dog”, D) = 0 \* 0.176 = 0

Table for d2:

|  |  |
| --- | --- |
| Term | Term Count |
| love | 1 |
| dog | 1 |

a)

tf(“love”, d2) = 1/2 = 0.5

df(“love”, D) = 3

idf(“love”, D) = log(|D| / df(“love”, D) = log(|3| / 3) = 0

tf-idf(“love”, d2, D) = tf(“love”, d2) \* idf(“love”, D) = 0.5 \* 0 = 0

b)

tf(“cat”, d2) = 0/2 = 0

df(“cat”, D) = 2

idf(“cat”, D) = log(|D| / df(“cat”, D) = log(|3| / 2) = 0.176

tf-idf(“cat”, d2, D) = tf(“cat”, d2) \* idf(“cat”, D) = 0 \* 0.176 = 0

c)

tf(“dog”, d2) = 1/2 = 0.5

df(“dog”, D) = 2

idf(“dog”, D) = log(|D| / df(“dog”, D) = log(|3| / 2) = 0.176

tf-idf(“dog”, d2, D) = tf(“dog, d2) \* idf(“dog”, D) = 0.5 \* 0.176 = 0.088

Table for d3:

|  |  |
| --- | --- |
| Term | Term Count |
| love | 1 |
| dog | 1 |
| cat | 1 |

a)

tf(“love”, d3) = 1/3 = 0.333

df(“love”, D) = 3

idf(“love”, D) = log(|D| / df(“love”, D) = log(|3| / 3) = 0

tf-idf(“love”, d3, D) = tf(“love”, d3) \* idf(“love”, D) = 0.333 \* 0 = 0

b)

tf(“cat”, d3) = 1/3 = 0.333

df(“cat”, D) = 2

idf(“cat”, D) = log(|D| / df(“cat”, D) = log(|3| / 2) = 0.176

tf-idf(“cat”, d3, D) = tf(“cat”, d3) \* idf(“cat”, D) = 0.333 \* 0.176 = 0.059

c)

tf(“dog”, d3) = 1/3 = 0.333

df(“dog”, D) = 2

idf(“dog”, D) = log(|D| / df(“dog”, D) = log(|3| / 2) = 0.176

tf-idf(“dog”, d3, D) = tf(“dog”, d3) \* idf(“dog”, D) = 0.333 \* 0.176 = 0.059

Final Table of tf-idf:

|  |  |  |  |
| --- | --- | --- | --- |
| Document-term matrix | love | cat | Dog |
| d1 | 0 | 0.117 | 0 |
| d2 | 0 | 0 | 0.0088 |
| d3 | 0 | 0.058 | 0.0058 |

1. [20 points]. Complete the Python program (indexing.py) that will read the file collection.csv and output the tf-idf document-term matrix following the same requirements defined in question 7. Add the link to an online repository as the answer to this question.

Github Link:

**Important Note:** Answers to all questions should be written clearly, concisely, and unmistakably delineated. You may resubmit multiple times until the deadline (the last submission will be considered).

**NO LATE ASSIGNMENTS WILL BE ACCEPTED. ALWAYS SUBMIT WHATEVER YOU HAVE COMPLETED FOR PARTIAL CREDIT BEFORE THE DEADLINE!**