CS 4250 – Assignment #1

Maximum Points: 100 pts.

Bronco ID: |__|__|__|__|__|__|__|__|__|

Last Name: ______________________

First Name: ______________________

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:

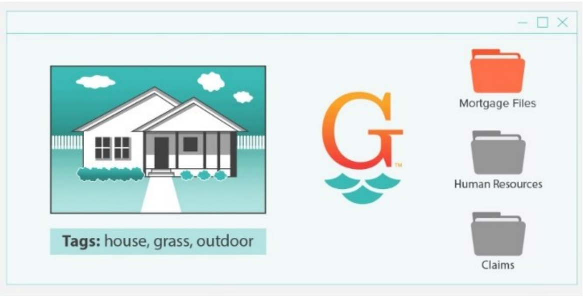
   a. [4 points]. Why is querying a database table easier compared to querying text documents? For full marks, **list** and **explain** at least **two factors** to elaborate your answer.
   b. [2 points]. Explain how **text** has been **used** by Information Retrieval researchers to compare multimedia documents and how this **scenario** is currently being **changed**.

2. [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.

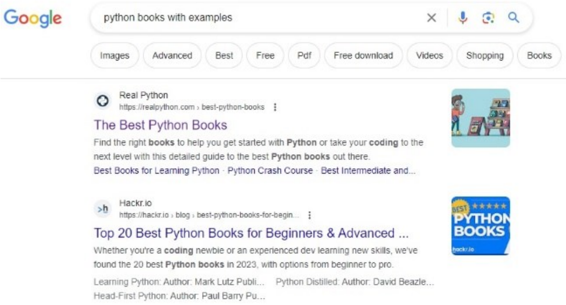
   a. Web search engine.
   b. Vertical search engine.
   c. Enterprise search engine.
   d. Desktop search engine.
   e. Finally, **explain** how peer-to-peer search engines differ from the other previous types.

3. [8 points – 2 points each]. **Identify** and **explain** the following Information Retrieval tasks.

a.



b.



python books with examples

Real Python
https://realpython.com › best-python-books

The Best Python Books

Find the right **books** to help you get started with **Python** or take your **coding** to the next level with this detailed guide to the best **Python books** out there.

Best Books for Learning Python · Python Crash Course · Best Intermediate and...

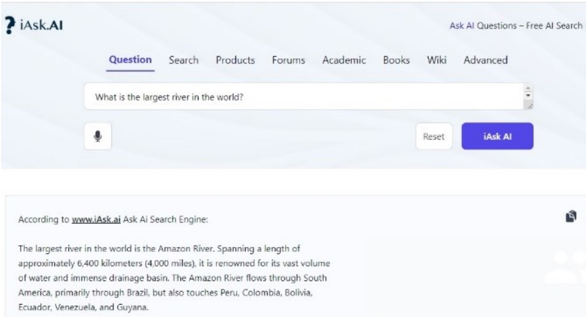
Hackr.io
https://hackr.io › blog › best-python-books-for-begin...

Top 20 Best Python Books for Beginners & Advanced ...

Whether you're a **coding** newbie or an experienced dev learning new skills, we've found the 20 best **Python books** in 2023, with options from beginner to pro.

Learning Python: Author: Mark Lutz Publi... Python Distilled: Author: David Beazle...
Head-First Python: Author: Paul Barry Pu...

c.

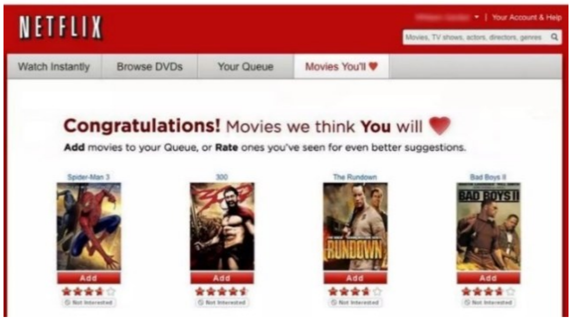


What is the largest river in the world?

According to www.iAsk.ai Ask Ai Search Engine:

The largest river in the world is the Amazon River. Spanning a length of approximately 6,400 kilometers (4,000 miles), it is renowned for its vast volume of water and immense drainage basin. The Amazon River flows through South America, primarily through Brazil, but also touches Peru, Colombia, Bolivia, Ecuador, Venezuela, and Guyana.

d.



Congratulations! Movies we think **You** will ♥

Add movies to your Queue, or **Rate** ones you've seen for even better suggestions.

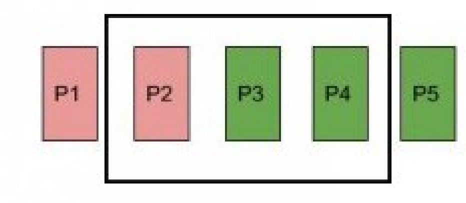
4. [8 points. 2 points each]. A retrieval model is a formal representation of the process of matching a query and a document, forming the basis of ranking algorithms that sort documents according to their relevance. Considering that relevance is one of the big issues for Information Retrieval research, answer the questions below.

   a. Explain why **topical relevance** and **user relevance** should be considered during search.
   b. Considering **only topic relevance** but **not user relevance**, give an example of a good search engine output based on a query.
   c. Considering **only user relevance** but **not topic relevance**, give an example of a good search engine output based on a query.
   d. Considering both **topic relevance** and **user relevance**, give an example of a good search engine output based on a query.

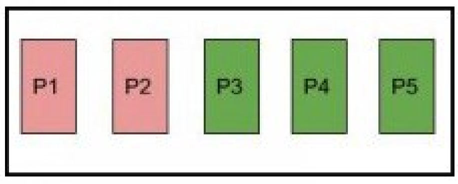
   Requirements: Elaborate **why** those examples are good and do not use the **same word** or **synonyms** for the **query/output**.

5. [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. Requirement: show your **math** for full marks.

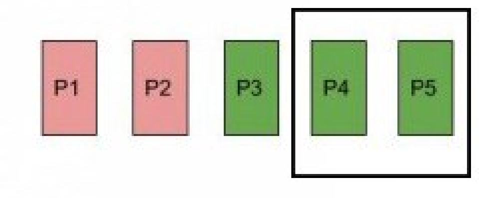
a.



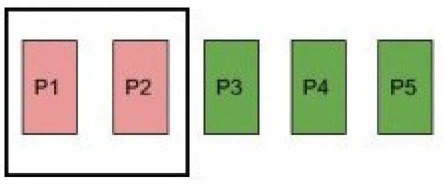
b.



c.

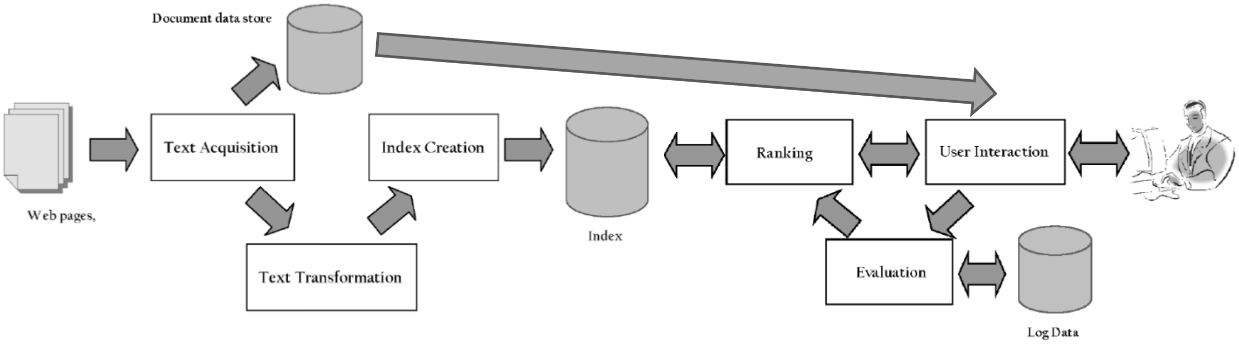


d.



6. [20 points]. Assume that you work for company A which wants to implement a competitive Web search engine. Use the words and the components in the image below to explain in one paragraph how those systems operate while executing the indexing and query processes to your supervisor. For full marks, **do not miss** a single **word** or **image component**.

   Words = {crawl, parse, terms, query, relevant, weights, precision & recall, click, user}



7. [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 term weights matrix according to the data below. Requirements: 1) you must conduct **stopword removal** and **stemming** before indexing the terms, 2) place the terms in the matrix following the sequence of their occurrences in the documents from d_1 to d_3, 3) show your **math** for full marks.

   d_1 = “I love cats and cats”.

   d_2= “She loves her dog”.

   d_3= “They love their dogs and cat”.

   Stopwords: pronouns, conjunctions.

8. [20 points]. Complete the Python program (search_engine.py) that will read the file collection.csv and output the precision/recall of a proposed search engine given the query q ={cat and dogs}. Add the link to an online repository as the answer to this question.

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!