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**.

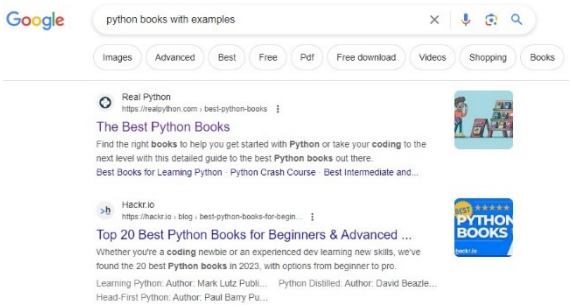
2. [10 points. 2 points each]. A search engine is the practical **application** of Information Retrieval techniques to large-scale text collections. **Explain** the scope of the different search engine applications and give **one practical example**.

   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.

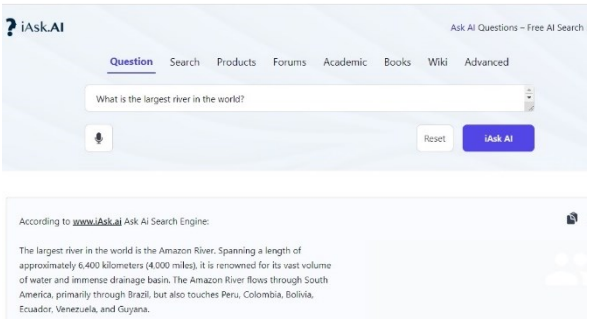
3. [8 points – 2 points each]. **Identify** and **explain** the following **tasks** that involve Information Retrieval.

   a.

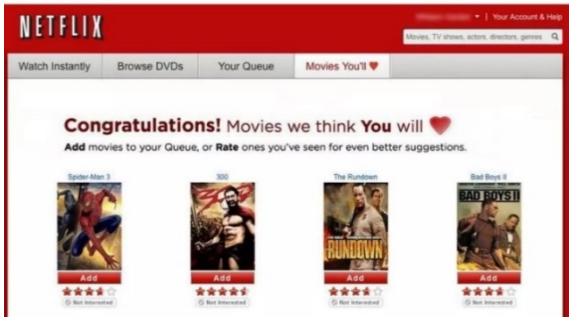
   b.



c.



d.

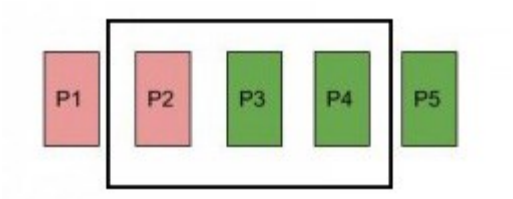


4. [8 points. 2 points each]. A retrieval model is a formal representation of the process of matching a query and a document, forming the basis of ranking algorithms that sort documents according to their relevance. Considering that **relevance** is one of the big issues for Information Retrieval research,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

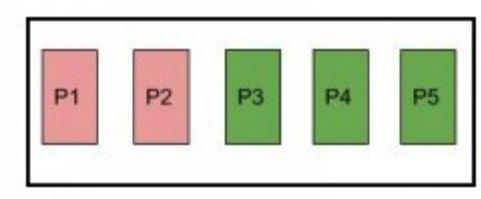
   a. Explain why **topical relevance** and **user relevance** should be considered during search.
   b. Considering **only topical** but **not user relevance**, give a hypothetical example of a good search engine output based on a query. Provide the **user profile**, **query,** and **document returned**.
   c. 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**.
   d. 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**.

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, and the black rectangles show the retrieved documents. Show your **math** (fraction and final value) for full marks.

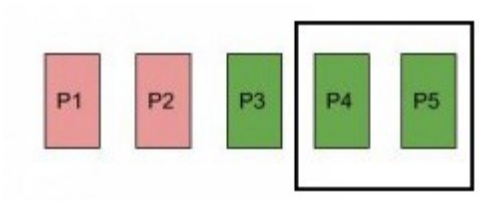
   a.



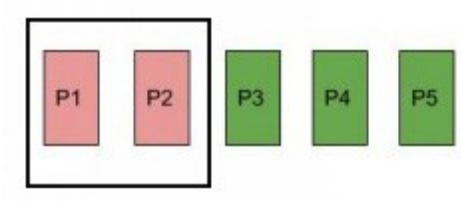
   b.



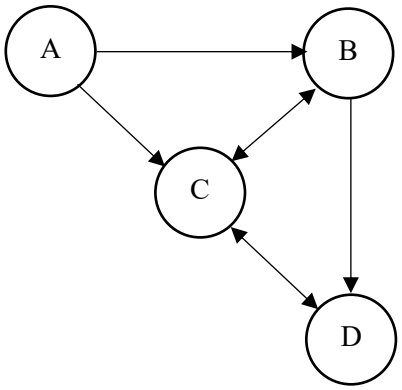
   c.



   d.



6. [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.



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 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 d_1 to d_3, 3) show your **math** for tf, idf, and tf-idf for full marks.

   d_1 = "I love cats and cats".

   d_2= "She loves her dog".

   d_3= "They love their dogs and cat".

8. [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.

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!