Bronco ID: |0|1|3|4|8|4|6|7|9|

Last Name: Francisco

First Name: Serrano

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.
        1. With querying a database, you already have predefined features that you can use to compare against other records in the database— it makes comparisons between two records much easier than an analogous comparison using text documents. Another reason would be because of text documents are much less structured than if you were querying a database, knowing where to find the information you are looking for becomes more complicated in a text document.
    2. [2 points]. Explain how text has been used by Information Retrieval researchers to compare multimedia documents and how this scenario is currently being changed.
        1. The handling of multimedia documents within IR meant we either needed some kind of text alongside multimedia documents or we needed specialized algorithms to deal with the multimedia to facilitate the indexing/searching of it. Text has played a larger role in the former, researchers can opt to translate the multimedia documents into some text-based version or there can be text-based features that allow us to approach it as we would a text document. There is a larger push toward the development of more specialized algorithms that directly work with the multimedia content alongside text-based approaches.
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.
    1. Web search engine.
        1. With a web search engine, you need to be able to catalog, index, and provide search capabilities for all the visible part of the internet. Need to be able to provide quick responses, or users will flee to a competitor.
    2. Vertical search engine.
        1. They need to be able to index content focused on a specific topic, it might be content related to travel or real-estate, but it is search that is focused in on one kind of umbrella topic.
    3. Enterprise search engine.
        1. It needs to be able to take in a lot of information from different sources/topics and then use that when doing a search.
    4. Desktop search engine.

        1. It needs to be able to use a variety of emails, new documents, and webpages to provide an interface for the user to search.
    5. Finally, explain how peer-to-peer search engines differ from the other previous types.
        1. It is a decentralized search engine that employs nodes to index, store, and retrieve data. Instead of using a centralized server, like larger search engines would, it is usually more of a community-oriented effort to facilitate web search.
3. Identify and explain the following Information Retrieval Tasks
    1. Classification because the image has the tags with house, grass, outdoor so they have classified this image.
    2. Adhoc search because the user is looking for specific documents on the topic of python.
    3. Question Answering since the user would be asking a question and receiving a response in return.
    4. Filtering since they have identified a user profile that would enjoy those multimedia documents.
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.
    1. Explain why topical relevance and user relevance should be considered during search.
        1. Topical relevance means that the content that we end up presenting to the user is directly related and covering the overall topic of their query, this is of upmost importance because without it we aren’t in the realm of answering their query. User relevance is importance as well because while something might be topical, it might not be relevant to a specific user because of something cultural, geographic, gender related, or other attributes of their user profile. Both go together for providing relevant documents for the given query.
    2. Considering only topic relevance but not user relevance, give a practical example of a good search engine output based on a query.
        1. Query: Apple MacBook Air
        2. Results/Output: A list review/articles on the laptop, Apple’s sales page ranking at the top, and other online retailer’s (not any specific physical site, just online shopping pages) sites.

            Output:
            Apple's Official Site: MacBook Air - Apple's revolutionary laptop with the M1 chip.

            www.apple.com/macbook-air

            TechReviews: Apple MacBook Air Review 2023 - Unboxing, Performance and Verdict www.techreviews.com/macbook-air-2023-review

BuyMac: MacBook Air Deals and Discounts - Order now and get 20% off.

www.buymac.com/macbook-air-deals

GizmoGeek: How does MacBook Air compare to MacBook Pro?

www.gizmogeek.com/macbook-air-vs-pro

   3. It is a good output because it is directly topical to the query, the engine is finding the most relevant documents which would firstly be what we consider a source (apple’s page itself), other online retailers might have their site ranking as well (like BestBuy, target, etc.), and it covers reviews which are what most people would search with tech device related queries.

3. Considering only user relevance but not topic relevance, give a practical example of a good search engine output based on a query.
   1. Query: Good bars
   2. Results/Output: A list of the best bars near their current location, using past search history (where they have shopped at, what restaurants they frequent, etc.) to provide something more tailored to the user.

Output:

Ranked Search Results for "Good bars":

LocalCraft Pub: Your favorite craft beer joint is just around the corner!

2 miles away from you

Wine & Dine: Fine wines and tapas that you'll love based on your last trip to Italy!

4 miles away from you

LiveJazz Bar: You listened to Miles Davis last week; you'll love the live jazz here.

5 miles away from you

VeganBar: Craft cocktails with vegan options, considering your recent vegan restaurant visits.

6 miles away from you

CollegePub: You searched for college football last week; this bar has great game day specials.

4 miles away from you

    3. Good because it factors in almost entirely the user profile, if it was more topical it would be more along the lines of what makes a bar good instead of something that is suggesting good bars.
4. Considering both topic relevance and user relevance, give a practical example of a good search engine output based on a query.
    1. Query: Latest mobile phones
    2. Results/Output: A list of the most recently released phones, has some topical diversity (covers different brands/models), while also considering the kind of features that the user has shown a preference for (camera type/quality, number of cameras, what OS, brand preferences, etc.).

Output: Ranked Search Results for "Latest mobile phones":

Apple iPhone 14: Official site showcasing features and specs. You've only owned iPhones before.

www.apple.com/iphone-14

Samsung Galaxy S23: All you need to know. Based on your search for high-quality cameras and good speakers.

www.samsung.com/galaxy-s23

Google Pixel 7: Full review and specs. Noting your interest in Google Assistant and Android OS/ecosystem.

www.google.com/pixel-7

Good because it is topical, it focuses on the latest phones while filtering the content down to fit the user’s preferences that they have shown in past search history.

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

⑤ (from AI 4250)

percision : $\frac{\text{\# of relevant retrieved}}{\text{\# of retrieved}}$

vs

recall : $\frac{\text{\# of relevant retrieved}}{\text{\# of relevant}}$

(green = relv. & red= non relev., box= retrieved)

a.) percision = $\frac{2}{3}$

recall = $\frac{2}{3}$

b.) percision = $\frac{3}{5}$

recall = $\frac{3}{3}$

c.) percision = $\frac{2}{2}$

recall = $\frac{2}{3}$

d.) percision = $\frac{0}{2}$

recall = $\frac{0}{3}$

6. We would need to crawl webpages, so we would need to identify which sites and then visit them, and once we got to them, we would need to begin the text acquisition phase. At this point, systems will save the content of the webpage and associated metadata in a Document data store so that they don't need to re-do this part of the process. This leads us to the next phase, the text transformation phase where you are getting the text document and parse it into terms (words, phrases, names, links, etc.). After that, we go into the index creation phase where the index terms are used to create data structures (indexes) to allow for modern relevant quick searches. To create a good index, you need to employ some document statistics and assign weights to terms. Once the index is built, you can use it to implement the ranking algorithm for a given query. During the user interaction phase, you provide an interface for the user to enter queries and click on the most relevant one. To evaluate a search engine, we can use precision and recall. Precision measures the portion of retrieved documents that were relevant, and recall measures the portion of retrieved documents that were retrieved.
7. 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:

a. you must conduct stop word removal and stemming before indexing the terms,
b. place the terms in the matrix following the sequence of their occurrences in the documents from $d1$ to $d3$
c. show your math for full marks.

⑦ Derive tf.idf term weights according to the data

→ need to remove stop words and do stemming b4 indexing

→ does stemming remove the plural?

d_1 = "I love cats and cats"

P S C S

after removing stopping words/stemming

d_1 = "love cat cat"

d_2 = "she loves her dog"

P S P

after removing stopping words/stemming

d_2 = "love dog"

d_3 = "they love their dogs and cat"

P P S C

after removing stopping words/stemming

d_3 = "love dog cat"

$$tf = \begin{array}{c|ccc} & t_1 & t_2 & t_3 \\ \hline d_1 & \frac{1}{3} & \frac{2}{3} & 0 \\ d_2 & \frac{1}{2} & 0 & \frac{1}{2} \\ d_3 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}$$

note: I assumed this is what you meant in requirement #2.

(love and cat appear first in doc #1, then dog appears in doc #2)

ve convert and end up w/

d_1 = "love cat cat"

d_2 = "love dog"

d_3 = "love dog cat"

where $t = [\overset{t_1}{\text{"love"}}, \overset{t_2}{\text{"cat"}}, \overset{t_3}{\text{"dog"}}]$

where $|D| = 3$, $D = [d_1, d_2, d_3]$

$$idf(\overset{t_1}{\text{"love"}}, D) = \log\left(\frac{3}{3}\right)$$

$$= 0$$

$$idf(\overset{t_2}{\text{"cat"}}, D) = \log\left(\frac{3}{2}\right)$$

$$= 0.1760$$

$$idf(\overset{t_3}{\text{"dog"}}, D) = \log\left(\frac{3}{2}\right)$$

$$= 0.1760$$

$w_{1,2} = 2 * 0.176 \cdot \frac{1}{3}$

$= 0.352 \cdot \frac{1}{3} = 0.1173$

$w_{1,1} = 0$ (bc $\text{idf}(t_1, D) = 0$)

$w_{1,3} = 0$ (bc $\text{idf}(t_3, D) = 0$)

$w_{2,1} = 0$ bc $\text{idf}(t_1, D) = 0$

$w_{2,2} = 0$ bc $\text{tf}(t_2, d_2) = 0$

$w_{2,3} = \frac{1}{2} * \text{idf}(t_3, D) = \frac{1}{2} * 0.1760$

$= 0.088$

$w_{3,1} = 0$ bc $\text{idf}(t_1, D) = 0$

$w_{3,2} = w_{3,3} = \frac{1}{3} * 0.1760 = 0.05867$

$$\text{tf-idf} = \begin{array}{cc|c|c} & t_1 & t_2 & t_3 \\ d_1 & 0 & .1173 & 0 \\ d_2 & 0 & 0 & .088 \\ d_3 & 0 & .0587 & 0.0587 \end{array}$$

8. complete the program search_engine.py that will read the file collection.csv

https://github.com/franserr99/4250Assignments/tree/main/a1