

These answers serve as examples and may include additional details. For more accurate grading, please adhere to the grading rubrics.

Q1 (1 point).

“tf idf” is a useful quantity that expresses the significance of term ‘i’ in document ‘j’ - it’s commonly expressed as (where tf is the term frequency, idf is the inverse document frequency):

$$w_{ij} = tf_{ij} \cdot idf_i$$

Why include idf in the calculation at all, what is the intuition/reasoning behind this? In your answer, please do NOT ‘define’ idf, that won’t get you any points.

Solution:


Including IDF in TF-IDF boosts the significance of rare terms while reducing the influence of common ones, improving the representation of each term's importance within a document. In other words, we discount (minimize) terms that aren’t ‘special’ [by virtue of them occurring in too many documents].

Rubrics:

- Full credits if the reasoning is correct
- -1 if they have only defined the terms tf or idf, and not the reasoning.
- -0.5 if the reasoning does not make sense or is partially correct.


Q2 (1 point).

In the context of computing the 'f score', we studied 'means' (averages) between numbers:

 Medium < > ⋮

If a and b are positive numbers, then

$$\text{Arithmetic Mean (AM)} = \frac{a + b}{2}$$
$$\text{Geometric Mean (GM)} = \sqrt{ab}$$
$$\text{Harmonic Mean (HM)} = \frac{2ab}{a + b}$$



Given 'a' and 'b' like above, what is yet another mean you could invent? Note - it doesn't need to have a practical purpose (though you might invent one later) :) But, it does need to lie between 'a' and 'b' [otherwise it won't qualify as a mean!].

Solution:

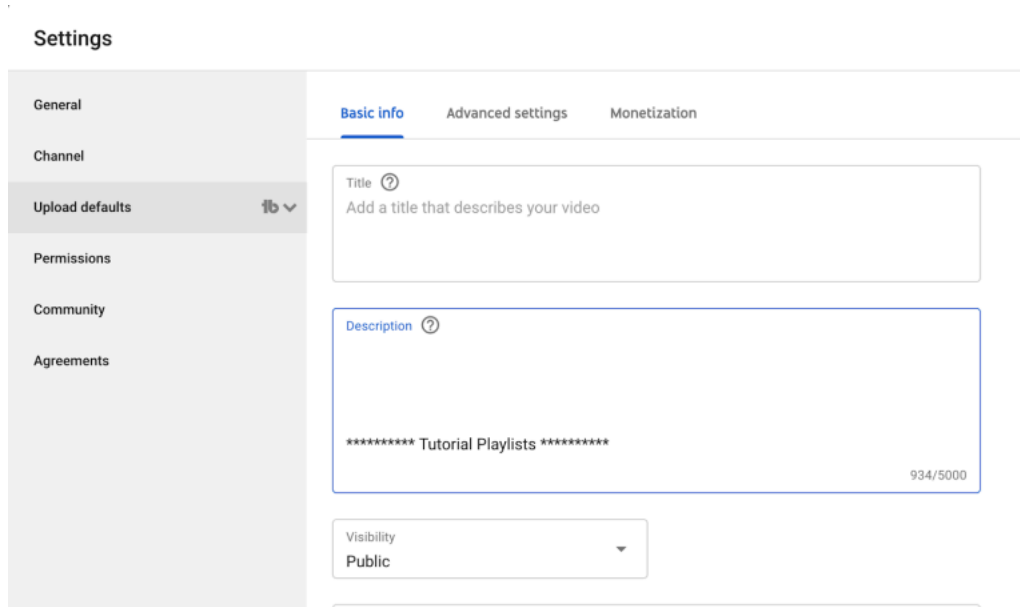
- Any other mean equation that calculates to a value between 'a' and 'b'. (Test on 2 cases (small number, big number) and (number, number+1)).
- SPECIFICALLY, an AM, GM, or HM formula -of- the existing means, eg. $\frac{1}{2}*(AM+GM)$, or $\sqrt{GM*HM}$, etc.

Rubrics:

- Full credits if any relevant mean involving 'a' and 'b' in some mathematical combination that follows the above point
- -1 if the mean does not fall in the range [a, b]
- -1 if the equation doesn't contain both a and b.

Q3 (1 point).

YouTube still relies on text [meta]data that the content creator provides, like so:



Such data is used to update the video (inverted) search index to make it be searchable by users.

What about the future - how (else) would be videos be made searchable?

Solution:

In the future, it will be possible to use content-based promptings to ask a foundation model to search and retrieve videos based on a combination of modality-specific criteria (e.g., audio, image, text, graph, object, video, and more). Search will *transition from metadata-based to content-based*.

In other words, the use of multimodal ML, specifically image and audio-based 'deep' search, will help bypass exclusive reliance on text metadata.

Rubrics:

- Full credit for mentioning the above or any other valid answer (there are many possible answers here)
- -0.5 if the reasoning is not "futuristic." or just meta-data based

Q4 (1+1=2 points).



The bottom-left in the above photo of the old Embarcadero Freeway (in San Francisco!), reminds you of what data structure used in info retrieval?

Solution:

- Skip lists / skip pointers

Rubrics:

- Give full credits for mentioning skip list / skip pointers data structure
- -0.5 for only mentioning “express lane”
- -1 for mentioning any other data structure

How does the data structure work (why is it useful)? Explain very briefly.

Solution:

- A skip list is a data structure that allows fast search, insertion, and deletion of elements. It works by layering multiple linked lists, where each level skips over a number of elements. The top-level list provides a faster/express route across the entire set, dropping down to lower levels to get closer to the target element.
- Skip lists are useful because they maintain sorted sequences with high efficiency and simplicity, supporting average-case time complexities comparable to balanced trees, but with easier implementation and maintenance. [As a side note, in IR, it's used to merge postings lists, in an 'AND' Boolean search.]

Rubrics:

- Full credits for correct working (0.5) + usefulness (0.5)
- -1 if explained this question for any data structure other than skip-list
- -0.5 for not mentioning how skip lists work
- -0.5 for not mentioning why it is useful

Q5 (1 + 1*4 = 5 points). What is the fundamental measure that is used in a variety of information retrieval algorithms/applications (that helps fetch high quality results)?

Solution:

- Similarity!

Rubrics:

- Full credits for stating 'similarity'
- -1 for mentioning any other measure

Discuss four applications of this measure, very briefly (fit them all in this page).

Solution:

- Rocchio
- kNN
- Page Rank
- Recommendation System
- similarity search using vector embeddings
- etc...

Rubrics:

- Full credits for writing **four** applications that briefly explain how similarity measure is being used
- -0.5 for each application if the student has written the application but not used the similarity measure
- -1 for each missing application

Q6 (1*5 = 5 points).

The 'search box' isn't only found in web search engines such as Google, Bing, Yahoo etc. Pick 5 other websites that you use, where you enter search terms - what does each site index and store? List and explain briefly.

Solution:

- Amazon (Product Listings)
- Netflix (Movies & TV Shows)
- Spotify (Songs)
- Instagram (Photos)
- Twitter (Text)
- OfferUp
- USPTO
- eBay
- LinkedIn
- etc...

NEED also to list WHAT is indexed+stored [eg. products, patents, auctions, songs, playlists...]

Rubrics:

- Full credits for mentioning 5 websites and what is indexed + stored
- -1 for each missing website with a search bar and corresponding index
- -0.5 for each example where a chosen site has a relevant search box but has an incorrect/vague explanation of the index and what is being stored at each website

Q7 (2+2+1 = 5 points).

As a toy example related to video recommendations (eg. on YouTube-like sites), imagine there are (just!) 4 videos that visitors can watch, for which we collect viewing stats to use in recommending videos.

Fill the following matrix with viewing stats (counts) for 100 views total, of videos watched together, where viewing order does not matter:

	1	2	3	4
1				
2				
3				
4				

Solution:

- the matrix should be 'symmetric.': the sum of the upper and lower triangle should both be 100, eg:

	1	2	3	4
1	0	30	28	24
2		0	10	6
3			0	2
4				0

Rubrics:

- Full credits if the matrix fulfills the above requirements
- 1 if the sum of the matrix is not 100
- 1 if the matrix is not symmetric

Fill in stats again, 100 views total again, where order of viewing does matter:

	1	2	3	4
1				
2				
3				
4				

Solution:

- the sum of the matrix should be 100
- the numbers do NOT need to be mirrored across the diagonal, eg. (1,2) could be 30, but (2,1) could be 28 (or 32)

Rubrics:

- Full credits if the matrix fulfills the above requirements
- -2 if the sum of the matrix is not 100

What is this type of recommendation algorithm called?

Solution:

Collaborative filtering or Co-visitation or Associative rule mining

Rubrics:

- Full credits for writing any above solution
- -1 if incorrect

Q8 (1*5 = 5 points).

This diagram has been hangin' out on the front page of our course:



Now that we are done with about 1/2 of the course, explain the above diagram, in terms of the topics we studied: web crawling, deduplication, text processing, inverted indexing, querying (ordered chronologically here, following our lecture schedule).

Sample Solution:

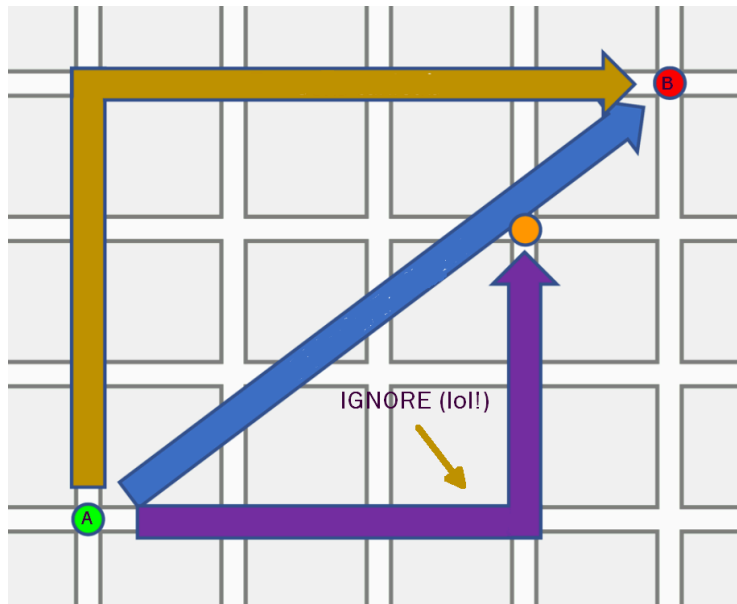
- **Web Crawling:**
 - It is also known as a spider that systematically browses the internet to index content. It begins with a set of URLs (also known as seeds) and traverses to the links present on these pages to add them to a BFS queue.
 - It exists between the agent and content and the main purpose of a web crawler is to index content and save it based on the information present on the page.
- **Deduplication:**
 - It is the process of identifying and avoiding identical/near identical content so that a single copy of the information is indexed.
 - It exists between the agent and content and helps in building the index efficiently.
- **Text Processing:**
 - Text processing in search engines involves analyzing and organizing the text found on web pages to make it searchable.
 - Text processing is used to analyze the queries entered by the user and is also used to extract information and index content.

- Inverted Indexing:
 - Inverted indexing is the process of linking words to where they are found in a database, turning the usual way of indexing around. Instead of keeping a list of pages and what's on them, an inverted index makes a list of every single word that shows up in any document and shows which documents contain each word.
 - It exists between the agent and the content.
- Querying:
 - Querying is the process of requesting information from a search engine.
 - It connects the users with the agents so that they can retrieve relevant information from the content database.

Rubrics:

- Give full credit for mentioning a sound and reasonable explanation of its **connection to the provided diagram** for all 5 terms.
- -1 for each missing/incorrect explanation.
- -0.5 for each point where a student only explained the definition, function, or operation of the term but did not mention the connection to the diagram

Q9 (0.5+0.5+1 = 2 points).



In the diagram above, between A and B:

- what is the diagonal distance metric called?

Solution: Euclidean Distance

Rubrics: Full credits if the answer is correct, else -0.5

- what is the vertical-horizontal distance metric called?

Solution: Manhattan Distance or Taxicab Geometry

Rubrics: Full credits if the answer is correct, else -0.5

- WHY is the diagonal measure always shorter? You need to provide a simple but accurate explanation/reason/proof!

Solution: Triangle Inequality; in other words, **the shortest distance between two points is always a straight line** [in our case, the light-blue line] :)

Rubrics: Full credits if the answer is correct, else -1

Q10 (1*3 = 3 points). Using Augmented Reality (AR), we can overlay pixels (that display data, images, video, rendered 3D graphics) on top of live video (eg. think Pokémon GO). Your phone's camera captures video, which is then analyzed by the AR app, which computes the overlay in response.

Briefly discuss three search applications that can use AR. For each, talk about: what you would capture, what information can be retrieved about it, how can it be presented (overlaid).

Solution: (Examples, a question only asks for 3)

Art and History: Capture artworks or historical artifacts. Retrieve information about the artist, time period, or historical context. Present detailed information, audio guides, or 3D models overlaid on the artifacts.

Food and Dining: Capture restaurant menus or dishes. Retrieve information about ingredients, calories, or reviews. Present nutritional information, customer ratings, or recommended dishes overlaid on the menu or dishes.

Travel and Exploration: Capture landmarks or natural wonders. Retrieve information about location details, geological formations, or travel tips. Present historical facts, nearby attractions, or scenic views overlaid on the landmarks.

Fashion and Style: Capture clothing or accessories. Retrieve information about brands, prices, or styling tips. Present outfit suggestions, size guides, or virtual try-on overlays on the clothing.

Home Improvement: Capture home decor or furniture. Retrieve information about dimensions, materials, or customer reviews. Present room layout suggestions, color matching tips, or virtual placement of furniture in the room.

Sports and Fitness: Capture sports equipment or workout routines. Retrieve information about specifications, training tips, or exercise demonstrations. Present workout instructions, performance metrics, or equipment details overlaid on the equipment.

Nature and Wildlife: Capture plants or animals. Retrieve information about species, habitats, or conservation status. Present species identification, educational facts, or wildlife sounds overlaid on the plants or animals.

Music and Entertainment: Capture concert posters or movie posters. Retrieve information about artists, showtimes, or plot summaries. Present song previews, movie trailers, or concert ticket links overlaid on the posters.

Science and Technology: Capture scientific exhibits or tech gadgets. Retrieve information about innovations, functionalities, or research findings. Present interactive simulations, data visualizations, or product specifications overlaid on the exhibits or gadgets.

Fitness and Wellness: Capture yoga poses or workout equipment. Retrieve information about exercise benefits, proper form, or meditation techniques. Present guided workouts, breathing exercises, or posture corrections overlaid on the poses or equipment.

Education and Learning: Capture educational posters or textbooks. Retrieve information about concepts, definitions, or historical events. Present interactive lessons, quizzes, or additional resources overlaid on the posters or textbooks.

Events and Festivals: Capture event banners or festival stages. Retrieve information about performers, schedules, or ticket prices. Present event highlights, live streaming, or social media feeds overlaid on the banners or stages.

Rubrics:

- Full credits for mentioning any valid search applications
- -1 for each missing/incorrect search application
- -0.5 for each partial/invalid explanation