Question	Rubrics
Q1) In IR, What are similarity measures for? Name and discuss 4 of them that we studied? (3 points)	**Possible answers:** Similarity measures are used to compare the content of two or more records and determine how similar they are , Used 1)to rank the retrieved the documents based on the relevance 2)To control the size of the retrieved set by enforcing a threshold 3)to detect near duplicates or to do deduplication for approximate match (**+1 for any close enough answer**) 4 similarity measures - for eg:- Jaccard, edit, cosine, hamming, Euclidean, Levenshtein (**+0.5 * 4 = 2....Any 4 with correct explanation**) **-0.25 point / similarity measure if no explanation is given.** **0 if the answer is wrong**
Q2)How is 'speedup' achieved by a search engine, in serving queries, and in crawling the web, in terms of these four aspects: data structures, computational machinery, disk space, bandwidth? In other words, provide four different ways/techniques, one related to each aspect. (2 points)	**Data structures:** use efficient data structures to store and index web pages, such as inverted indexes, hash tables, skip tables, tries, KD tries, bloom filters or B-trees (**+ 0.5 points for mentioning any one of these data structures**) **Computational machinery :** -use distributed computing techniques, such as parallel processing or MapReduce, to speed up the processing of large volumes of data. -use specialized hardware like GPUs or FPGAs to speed up certain operations, such as indexing or ranking. -use machine learning techniques, such as deep learning or natural language processing, to improve the accuracy and relevance of search results to automate certain tasks, such as query classification or content extraction (**0.5 points for a correct technique**) **Disk space :** -usage of compression techniques to reduce the amount of disk space required to store its index and other data. -use data deduplication techniques to identify and eliminate duplicate data across different pages and filesto reduce the amount of disk space -partition data across multiple disks or servers to improve performance and reduce the risk of data loss (**0.5 points for a correct technique**)

	Bandwidth : -usage of caching and content distribution networks (CDNs) to reduce the amount of bandwidth required to serve queries and crawl the web or browser caching, to reduce the amount of data that needs to be transmitted over the network -can use parallel fetching techniques to reduce the time required to retrieve web pages over the network (**0.5 points for a correct technique**) **Total 0.5 * 4 = 2 points** **0 if the answer is wrong**
Q3)How is a web browser and a web crawler similar? How are they different?(2 points)	**Similarity :** -A web browser and a web crawler are similar in that they both interact with web pages on the internet. -They both retrieve web pages from servers on the internet by communicating with them using HTTP. -They both access HTML content, and they can navigate between pages by following links. (**+1 point for any correct similarity**) **Difference** : -A web browser is designed for human users to access and view web pages, while a web crawler is designed for automated indexing and data collection. -Web browsers display web pages in a graphical user interface, allowing users to navigate between pages, search for content, and interact with web-based applications. -Web crawlers systematically browse the internet, usually for the purpose of indexing or collecting information. They automatically visit web pages, follow links, and extract data to create indexes or databases.(**+1 point for correct differences between each other i.e 0.5 for web crawler and 0.5 for web browser**) **0 if the answer is wrong**
Q4)Broadly speaking, pages in a site typically point to other pages in the site (eg. think CNN, Wikipedia, OfferUp, YouTube, usc.edu, etc).A web crawler can be written, to hold future (as of yet unvisited) URLs in a queue data structure, or, in a stack data structure. Which of these would site administrators prefer (if they had a say in the crawler architecture), and why? (2 points)	Correct answer: Queue datastructure(**+0.5 point for mentioning using queue data structure**) **Why??** -Queue data structure follows a First-In-First-Out (FIFO) order, which ensures that the URLs are processed in the order they were added to the queue This is important because it ensures that pages are crawled in a systematic and predictable way, which can help site

	administrators better understand how their site is being indexed and improve their SEO strategy. -A queue data structure helps to ensure that the crawler does not waste resources by repeatedly visiting the same pages or getting stuck in loops. -A queue data structure ensures that the crawler explores all pages on the site, rather than only focusing on a small subset of pages. (+1.5 point for a correct explanation) **Possible other answer:** (1.5 points only in total, Reduce 0.5 points) Stack datastructure with a reasonable purpose Stack data structure is used in these cases: - -If the crawler encounters a dead end or a page with no useful information, it may need to backtrack to a previous page and explore a different branch. -A stack data structure can be used to keep track of the current branch being explored, and easily backtrack to the previous page once the branch has been fully explored. -In some cases, multiple crawlers may be used to explore a large website in parallel. In this case, a stack data structure can be useful to divide the workload between the crawlers, with each crawler responsible for exploring a different branch of the site.(+1.5 point for correct reason for usage of stack) +2 points if answer have stated both datastructures with reasons on when to use it (-0.5 point if the answer say the site administrators would prefer stack datastructure) -1 point for inconclusive statements on which datastructures to use 0 if the answer is wrong
Q5)We can use 'index cards' to index a small collection of books (THAT is why they are called that - duh!). We use a card for each book that we want to index. Shown below is a sample index card, for a book that belongs in categories B, C, E (intact holes), but not categories A,D (cut-out holes). Given a collection of such cards (eg. 1000 of them) where each card is for a book that could be in	To pick out index cards for books in category B and C: - Stack all the index cards - Place two rods in holes B and C - Pull both the rods outside together, all cards that are stuck in rods are of category B and C (+1 point for a similar explanation) To pick out index cards for books in category A and D: - Stack all the index cards

any of A, B,C,D,E categories (at least one, but can be 2, 3, 4 or all 5), how would you pick out all the books that are in B and C categories? Using the same 500 cards again, how would you select books that can be in A or D categories? You have access to several long rods (eg. knitting needles) that can pass through the holes.(2 points)	- Place one rod in hole A - Pull the rod outside together, all cards that are stuck in rod are of category A - Place another rod in hole D in the leftover cards - Pull the rod outside together, all cards that are stuck in rod are of category D - All the cards pulled out are of category A or D **(+1 point for a similar explanation)** **0 if the answer is wrong**
Q6)Discuss the role of the agent with respect to the user, and, with respect to the content? In other words, what goes on in the left pair of arrows, and in right pair? What is different about the agent's behavior, when the agent is based on a large pretrained language model, eg when it is ChatGPT integrated with Bing, or Google's Bard?(3 points)	**Role of the agent with respect to the user** W.r.t the user, the role of the agent is to understand the user's information needs and preferences, and to provide personalized and relevant search results based on those needs. -involves analyzing the user's search queries, interpreting the user's intent, and presenting the results in a way that is easy to understand and navigate. **+1 point** **Role of the agent with respect to the content** W.r.t the content, the role of the agent is to crawl, index, and analyze the content to identify relevant documents or data that match the user's search query. - involves using data structures and algorithms to efficiently search and retrieve information from a large corpus of data. **+1 point** **when the agent is based on a large pretrained language model** -Large pretrained language models have the ability to comprehend user queries in a manner that resembles human understanding since it is already trained on massive amounts of text data. -Large pretrained language models can understand the nuances of natural language, allowing them to interpret the content of web pages and other documents more accurately which can lead to more relevant search results and better content recommendations. -Language models can enable conversational interfaces, allowing the user to interact with the agent in a more natural and intuitive way. - Language models can potentially help SUMMARIZE results (**Any close enough answer can be awarded +1 point**) **Total 3 * (+1) = 3 points** **0 if the answer is wrong**

Q7) What is the purpose of discounting cumulative gains, when we rank search engines? what other rank based functions can you think of (which is also less harsh than 1/rank) What other functions can we use for this purpose? (3 points)	**Possible Answer of Purpose:** The purpose of discounting cumulative gains when ranking search engines is to account for the fact that users are more likely to click on results that appear higher up in the search results list. Simply summing up the relevance scores without discounting the effect of the ranking position can result in an overestimation of the actual relevance of the search results. **(+1 point for similar correct explanation of the purpose)** **Possible answers for other rank function that are less harsh:** - maybe a function that uses 1/square_root(rank) - **(+1 point if student gives a possible function that is less harsh than 1/root)** **Possible answers for other functions we can use:** - Binary relevance: This function assigns a relevance score of 1 to search results that are relevant to the user's search query and 0 to search results that are not relevant. - Graded relevance: This function assigns a graded relevance score to search results based on their degree of relevance to the user's search query. - Cumulative gain: This function calculates the total relevance score of a set of search results, with higher scores indicating greater relevance. - Normalized discounted cumulative gain (NDCG): This function is a variant of cumulative gain that takes into account the diminishing returns of relevance as the rank of search results increases. **(+1 point for a function with correct answer) (0.5 mark deduction if explanation is incorrect)** **0 if the answer is wrong**
Q8) youtube came up with a relatively simple way to create recommended videos. Rather than using a co-visitation count approach, it's also possible to group videos based on their content - what are two very different ways to do this? (2 points)	**Possible Answers:** - Text based approach - Visual Based Approach - Metadata analysis - Audio Analysis - Social Context - Machine Learning Algorithms **(any of the two approaches with correct explanation) (+1 point for each correct approach) (0.5 deduction if the explanation is not correct)** **We can also accept answers that mention 'content-based**

	filtering' [which is eqvt to the visual-based approach mentioned above] (Note: This answer is not limited to just these approaches, any other content based approaches that make sense can be correct) 0 if the answer is wrong
Q9) For 'power searching' Google, we use search modifiers such as :site, :filetype, :intext, etc. These help narrow down search. what three additional filters can you think of to narrow down searches to specific categories: (3 points)	**Possible Answers:** • inurl: This filter allows users to search for specific words or phrases within the URL of a webpage. • define: This filter allows users to find definitions of a specific word or phrase. • allintitle: This filter allows users to search for webpages that contain all of the specific words in their title. • allinurl: This filter allows users to search for webpages that contain all of the specific words in their URL. • location: This filter allows users to search for businesses or places of interest in a specific location. • book: This filter allows users to search for books and publications on a specific topic. • stocks: This filter allows users to view current stock prices and financial information for a specific company or ticker symbol. • date: This filter allows users to specify a date range for their search, such as "past week", "past month", or "past year". • Language: This search modifier would allow users to search for content in a specific language. • Author: This search modifier would allow users to search for content created by a specific author • Audience: This search modifier would allow users to narrow down their search results based on the intended audience of the content (any of these three filters with correct explanation) (+1 point for each correct point) (for incorrect explanation deduct 0.5 points) (Note: The answer is not limited to just these filters, any other filters that makes sense can be counted as correct)

	Yes! Mentioning 'new' ones [that Google doesn't use or document] is ok. 0 if the answer is wrong
Q10) Sites such as OfferUp, uspto.gov, ebay, ratemyprofessor etc offer specific services. Given that, how would the four sites mentioned above make it easy to search for what they offer? In other words, what might each index be so that we can search those indexes? (2 points)	**Possible Answers:** 1. OfferUp: OfferUp can index the listings created by users. This can include the title, description, location, and photos of each item. 2. USPTO.gov: The website can index information about each patent or trademark, including the title, abstract, inventors, and application number. 3. eBay: eBay can index the listings created by sellers. This can include the title, description, category, price, location, and photos of each item. 4. RateMyProfessor: RateMyProfessor can index information about each professor, including their name, school, department, and rating. (+0.5 points for correct relevant indexes for each website) (Note: The answer is not limited to just these indexes, any other indexes that makes sense for that particular website can be counted as correct) 0 if the answer is wrong

Q11) Characterize Rocchio, KNN and Nearest Neighbour technique for document classification in terms of aggregation, geometry and robustness.
(3 points)
	1. Aggregation
	    1. Rocchio: uses a centroid-based approach to aggregation.
	    2. KNN: uses aggregation based on the K most similar documents in the training set
	    3. Nearest Neighbor (K=1): Does not use aggregation, classify based on the document that is the closest
	2. Geometry
	    1. Rocchio: uses a geometry of points
	    2. KNN: uses a geometry of points
	    3. Nearest Neighbor (K=1): uses a geometry of points
	
	For this, we can also accept polygon (for Rocchio), point

	cluster (for kNN) and point (for nearest neighbor) 3. Robustness: 1. Rocchio: not very robust to noisy data or outliers 2. KNN: more robust to noisy data than Rocchio and NN 3. NN: not very robust to noisy data or outliers **(+1 point if the classification for a term is completely correct)** **(deduct 0.5 points if it is partially correct)** **0 if the answer is wrong**
Q12) Name two sites that you use, where 'approximate matches' are performed and results displayed, when a user enters a query (2 points)	**Possible Answers:** 1. Google 2. Amazon 3. YouTube 4. eBay 5. Bing 6. Yelp 7. TripAdvisor 8. Zillow 9. Spotify 10. LinkedIn **(+1 point for a correct website)** **(Note: the answer is not limited to these ten websites there can be others as well)** **0 if the answer is wrong**