Question	Answer
	Priya and Kunal, **EXCELLENT answers overall** :) My comments/corrections are in **bold**. #3 and #17 are the only places where my answers override yours :) You can DELETE this paragraph before sending this to the other CPs/TAs, and, feel free to remove my bolding of the text in my responses, as well.
1.When we upload pics/clips/songs to social media, what specific mechanism do we use to help others find our content when they search? Why do we need this specific mechanism? (1 Mark)	**We use #hashtag(s) [one or more] so that anyone searching using a hashtag can find our content as well, eg. #MeToo.** **Hashtags are easy to index. They serve as textual descriptors for non-textual content even (eg songs), making search simple [because similarity search isn't widely implemented yet, it's brand new].** Indexing of pictures, clips, songs and other media is usually done using metadata. (+0.5) Pictures, clips, videos, etc media do not exactly have textual data that will help the search engine index it. Hence the metadata, like description, can provide more information about the actual content. (+0.5 for similar explanation)
2. What algorithms causes/leads to/results in/is implicated in …"filter bubbles"? How does it lead to this? (1 Mark)	"Filter bubbles" are caused by algorithms in personalized content recommendation systems. (+0.5) Recommendation systems suggest content that is tailored to the user's interests and preferences. Hence, the user is limited to seeing information that is restricted in the areas of their interests causing these "filter bubbles". (+0.5 for similar explanation)
3. What two other **items** can a search engine serve **us** (eg. via "snippets"), in addition to	**[The question asks for new items… Any new items not currently being served will**

what gets served already? Name each item, and briefly state why it would be of use to us. (1 Mark)	**qualify, the two answers below are examples]** **a. Walkthroughs inside a restaurant - so we can get a feel for the place** **b. First-person POVs of amusement rides, cruise ships, opera house seats, etc - again so we can get a preview** **[please don't use the answers below :)]** **Featured** snippets: Google's attempt to answer the query right on the search results page. It provides a brief summary of an answer to the search query displayed at the top of the search results. Eg. Displayed as paragraph, or list or table. (+1) **Rich** snippets : gives users a convenient summary information about their search results at a glance. It allows users to assess the relevance of a search result before going to the actual website. Eg. address, hours, directions, events, etc. (+1)
4. As you know, genAI(generative AI) is so-called because it can generate content (text, images, video, audio, more). How will this adversely affect search in the (near, even) future? Explain carefully (don't write a vague answer!). (1 Mark)	**Seach will get 'polluted' by genAI content, making the results unreliable and incorrect - eg. image-searching for 'Picasso' will pull up genAI images in the style of Picasso, in addition to genuine Picasso paintings, users will/might not know the difference.** ● The search engines may not be able to distinguish between genuine and fake content. ● Fake images/videos can cause spread of misinformation/disinformation if they get indexed by the search engines. ● AI generated content can sound or look unnatural which can make the users feel uncomfortable or distrustful about the source. (+0.5 for each effect with explanation)

5. ChatGPT (for example) is said to "hallucinate" sometimes (or a lot of times, depending on the type of questions) - an unfortunate term (because only minds can hallucinate!) used by companies who serve this kind of AI products. This means that the bot provides an incorrect answer(which we can verify using our own knowledge or experience or by doing a good old search!) WHAT mechanism (in the algorithm) causes this to happen? Please be specific. (1 Mark)	**The algorithm has no understanding of what underlies language. It merely computes words one by one based on what is in its embedding, without being to assess if its output is factual or not.** **Bias** : ChatGPT can provide incorrect answers to few questions because of an inherent bias caused by its limitation to understand real-world or/and lack of proper training data (imbalanced data). (+1 for similar explanation)
6. We typically write code (eg for your homeworks #2 to #5) to make use of IR algorithms. An alternative way is to use 'nodes' (a node is a box-like representation that encapsulates a specific task by executing task's code) and WIRE them visually, like so –, WHAT would be two specific (and different from each other) advantages of switching to this way of working (using nodes, as opposed to coding) (1 Mark)	• Simplicity : users without coding knowledge can also follow all the steps. • Language independent : coders from different backgrounds can interpret the flow of work and replicate it in the programming languages that they are comfortable in. • Less erroneous : syntactical errors caused by coding languages can be avoided. • Visual representation : small space can capture the flow of a program better compared to the more complex "class" structures enforced by programming. • **The visual representation is 'self documenting', making it easy to modify by others** (+0.5 for each unique advantage - can be beyond the list mentioned above)
7. Consider the diagram below (the bottom part is simply a zoomed-in portion of the top) What algorithms did we study that results in such a collection of polygons? Why is each polygon convex? (1 Mark)	The figure shows a Voronoi diagram that is obtained from a k-NN classification algorithm (+0.5) Each **item** is placed in a group such that it is closer to that group than any other group, hence each polygon is convex. More: If a polygon is concave, that means that that particular **item** could be closer to the center of a neighboring polygon compared to the one it is a part of currently, hence nearest neighbor based polygons cannot be concave. Another explanation: the polygon's edges are formed by pairwise perpendicular bisections

	of adjacent centroids, this will necessarily lead to a convex polygon. (+0.5 for similar explanation)
8. We revised the diagram on and off many times: Now that the course is over, how does it summarize the course? Pick 4 specific and different IR tasks we studied during the course (including “Assorted topics” pair of lectures) and explain (in a line or two) each, in terms of the three pieces of our diagram. (1 Mark)	Recommendation systems - The user’s previous interests are used as an input along with the current query by the agent to give personalized content suggestions. Question Answering - The input given by the user is usually in the form of a natural language question (Who/what) and maybe specific, the agent has to identify entities and find relevant answers associating the nature of the question and the entity to provide relevant content back to the user. Related search using Clustering - The user input is the main query, the agent has to identify queries similar to the input query that may help the user get more information. This additional content may be displayed separately as “related search”. Location-Based Search - In such specific searches, one major part of the query from the user would be the location and the proximity. The agent has to identify results specific to that proximity mentioned by the user to display content. LLM based agents - The user input is usually a query in natural language. The agent must use advanced techniques to identify relevant information and *answer* using “knowledge” and present result content as a natural language response. Basic (keyword) search - the agent is a classic search engine that uses its inverted index along with ranking algorithms (including PageRank) to return a ranked sequence of results to the user.

	(+0.25 for each task with relevant explanation - can be beyond this list)
9. How do recommendation engines work? (2 Marks) What are two different uses for them when we search? (2 Marks)	A recommendation system is any system which provides a recommendation/prediction/opinion to a user on items. ● Content-based filtering : uses item similarity/clustering to recommend items like ones you like. ● Collaborative filtering : uses user similarity i.e links between users and the item they chose as the basis of recommendation. (+2 for a similar explanation for content and collaborative explanation) Two different uses - ● Recommending content that is similar to the user's preference / promoting content may or may not be popular but similar to user taste / "sell items from long tail". ● Increase diversity by recommending content that the user hasn't discovered yet. ● (more generally) Recommending items bought together, nearby tourist attractions (that the user did not search for), similar songs… (+1 for a each use - can be beyond this list)
10. What is vector similarity search? For what two different IR tasks are vector DBs useful? Name and explain why we cannot do it without vector DBs (1 Mark)	The main idea is that if data and documents are alike, their vectors will be similar. In vector space, each object or data point is represented as a vector in a high dimensional space. The goal is to find objects that are similar to a given query vector based on some similarity measure. (+0.5 for a similar explanation) IR tasks where vector DBs are useful ● Image and Video Retrieval: By representing images or video frames as vectors, vector databases enable efficient search for visually similar content

	• Natural Language Processing (NLP): By representing text documents or sentences as vectors, vector databases allow efficient similarity search based on semantic or contextual similarities • Recommendation Systems: By representing items as vectors based on their attributes or features, vector databases enable efficient retrieval of similar items (0.5 for any two of the above two answers with correct explanation)
11. Summarize any two HWs #2 through #5 - what algorithm was used and what task did it help accomplish? (1 Mark)	HW2: Crawler4j, to develop a develop a 'crawler' to 'crawl' (visit, fetch contents of) a list of URLs, HW3: Uses Map Reduce to build an inverted index of a data HW4: Uses Lunr and Solr for indexing and searching documents HW5: Uses Weaviate to set up, load and query vectorized data (+1 for any two with correct explanation)
12. What does OPL stand for, in the OPL stack? What is its main use? (1 mark)	"OPL' is the name we could give to such a stack comprised of O(penAI)+P(inecone)+L(angChain) (+0.5 for the correct full form) It has increasingly become the industry solution to overcome the two limitations of LLMs: 1. LLMs hallucination: chatGPT will sometimes provide wrong answers with overconfidence. One of the underlying causes is that those language models are trained to predict the next word very effectively, or the next token to be precise. Given an input text, chatGPT will return words with high probability, which

	doesn't mean that chatGPT has reasoning ability. 2. Less up-to-date knowledge: chatGPT's training data is limited to internet data prior to Sep 2021. Therefore, it will produce less desirable answers if your questions are about recent trends or topics. (+0.5 for the any two valid reasons with correct explanation)
13. How do RDF triples make search better? (1 mark)	- RDF triples enable semantic search, which goes beyond simple keyword matching - RDF triples facilitate the integration of data from multiple sources through the use of unique identifiers (URIs) and linking - DF triples enable the expansion and enrichment of search queries - RDF triples provide a structured and standardized way to represent data - RDF triples can be used to represent ontologies and knowledge graphs, which enable reasoning and inference (+1 for any of the above two reasons with correct explanation)
14. Name four algorithms we looked at, for IR tasks, that reply to iteration or recursion. for each briefly explain how iteration or recursion helps? (4 Marks)	- Page Rank Algorithm: iteration is used to improve the accuracy and convergence of the ranking calculation - snippet generation algorithm: iteration plays a crucial role in generating concise and relevant snippets from a larger text document - Basic spelling correction algorithm uses iteration/recursion to improve the accuracy of the correction process. - Google's Query Processing Algorithm - K-Means - Hierarchical clustering algorithm - Rocchio Algorithm - KNN (+1 for a correct a algorithm with correct explanation)

	(Note: The answer is not limited to the above points, any other algorithms taught in class are valid)
15. Name and briefly discuss 4 ML based algorithms we looked in class (4 Marks)	- K-Means - Hierarchical clustering - Rocchio Algorithm [not ML :)] - KNN - **Neural networks (eg in Transformers that are in (chat)GPT)** (+1 for a correct a algorithm with correct explanation) (Note: The answer is not limited to the above points, any other algorithms taught in class are valid)
15. Consider the image, pick four algorithms that are useful in IR. Explain the algorithm and how they are useful in IR. (4 marks)	1. Naive Bayes: ● Text categorization ● Sentiment analysis 2. Support Vector Machines (SVM): ● Document classification ● Information extraction 3. Neural Networks: ● Convolutional neural networks (CNNs): ● Document classification ● Image retrieval ● Recurrent neural networks (RNNs): ● Text generation ● Machine translation ● Question answering 4. Decision Trees: ● Document classification ● Feature selection 5. Random Forests: ● Document classification ● Sentiment analysis 6. Gradient Boosting: ● Ranking and relevance prediction 7. K-Means or Hierarchical clustering ● document clustering ● query expansion ● text expansion **8. KNN - for classfication** (+1 for one correct algorithm with correct explanation)

	(Note: The answer is not limited to the above points, any other algorithms mentioned in the image that has a use case in IR are valid)
17. Tiktok's recommendation engine uses a specific data structure to optimize it. What is the name of the data structure? How does it work? (1 mark)	**The data structure is a 'cuckoo hash'. It works by using two hash tables/functions rather than just one so that if a collision occurs in an incoming key, the existing key is kicked out to a new location in the 'other' table. We can accept 'loose' explanations as long as the use of two tables (or functions) is mentioned.** TikTok uses Monolith DS to optimize its recommendation system. (+0.5 for the correct DS) (+0.5 for correct explanation)
18. given two vectors, what two similarity measures can we calculate? What do vectors have to do with IR? (1 mark)	Similarities we can calculate: - Cosine similarity - Euclidean Distance (0.5 for any two similarity measures taught in class) IR often involves representing documents, queries, or data points as vectors in a high-dimensional space. By representing textual or numerical features of documents as vectors, various similarity measures can be applied to compare and rank documents based on their similarity to a given query (+0.5 for a similar explanation)