CS572 - Spring 25 Midterm Rubrics

Q1 (1*3 = 3 points).
Increasingly, the distinction between ‘search’ and ‘database access’ is becoming blurred - after all, the end-user doesn’t care how the information they want is being retrieved.

What are three examples where searching (via a web browser) “actually” involves a database query/search? For each example, provide both of these: sample search query, what data might be retrieved to answer it.

Sample Solution:

Example 1: Flight Search on Google

Sample search query: "Flights from Los Angeles to New York on March 20"
data retrieved: Google aggregates flight information from various airline databases, pulling data like:

- Available flights
- Departure and arrival times
- Ticket prices
- Airlines operating the route
- Layovers and total travel time

In the background, Google's search system translates this query into structured database queries to retrieve flight schedules from airline databases or aggregators like Skyscanner.

Example 2: Product Search on Google (E-commerce Data)

Sample search query: "Best budget gaming laptops under $1000"
data retrieved: Google accesses e-commerce databases (Amazon, Best Buy, Walmart, etc.) to retrieve:

- Laptop names, prices, and specifications
- Product reviews and ratings
- Seller details
- Availability and shipping options

Google likely queries structured product catalogs indexed from multiple e-commerce sites and filters results based on price constraints.

Example 3: Local Business Search (Google Maps or Yelp Integration)

Sample search query: "Best coffee shops near Times Square"
data retrieved: The search system queries a structured location database to retrieve:

- Business names, addresses, and ratings
- Customer reviews
- Operating hours
- Photos and menu details
- Distance from the searcher's location

Google's search integrates with Google Maps, which queries geospatial databases and business directories to provide accurate, location-based results.

——————

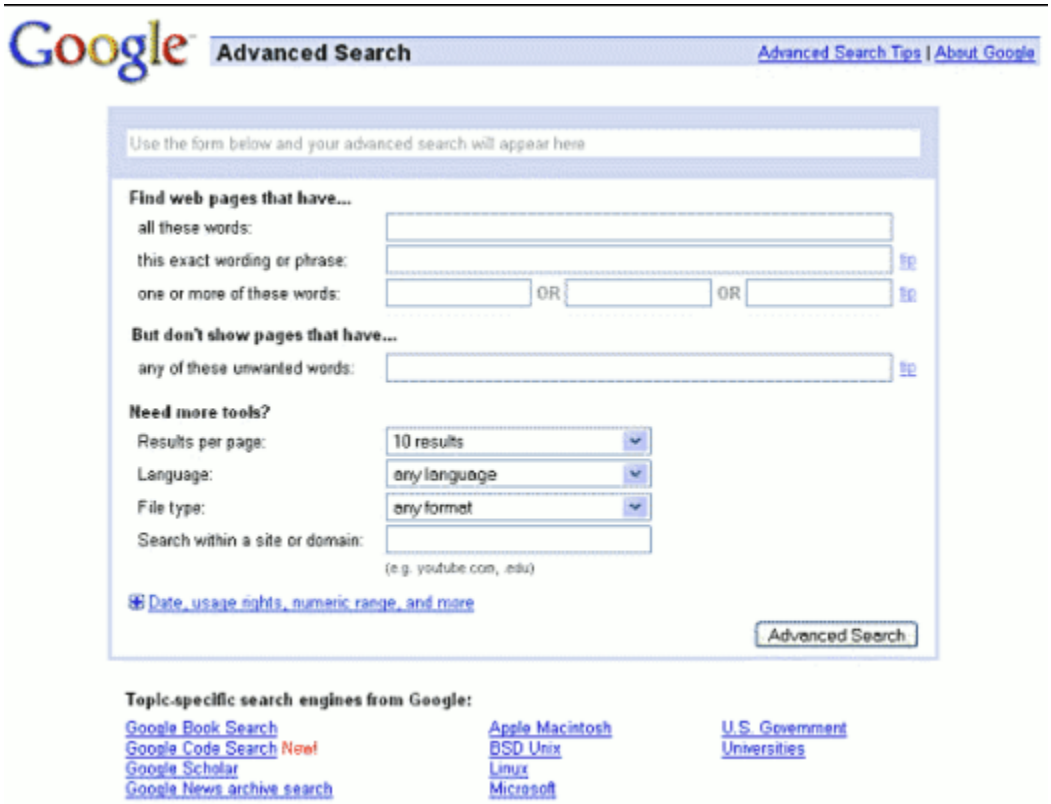
** Other acceptable examples (but not limited to):

- Flight search → Queries airline databases for schedules, prices, and availability.
- Stock price search → Retrieves real-time stock data from financial databases.
- Job search → Queries job listings from recruitment databases like LinkedIn or Indeed.
- Real estate search → Fetches property listings from housing databases.
- Weather search → Retrieves forecasts from meteorological databases.
- Movie showtimes → Queries theater schedules from entertainment databases.
- Restaurant search → Fetches business hours, ratings, and menus from Yelp or Google Business.

Rubric: +1 point for each logical example (0.5 for providing the query + 0.5 for providing the data retrieved)

Q2 (1.5 + 1.5 = 3 points):

The 'version 1' implementation of Google's 'Advanced Search' interface looked like this [the current version does not]:



It looks like there is a limitation with the ‘OR’ search - the user could only put in 3 terms (or just 2). Is that really the case? How would you overcome the ’3 terms max’ limit [what is the ‘hack’]?

Hint#1: the ‘advanced search’ page is simply a UI over ‘raw’/usual search interface. Hint #2: *everything* in computing is either data or code.

Solution: No, the limitation exists only in the user interface (UI) and is not a fundamental restriction of the search engine. To overcome this UI limitation, you can manually enter the OR terms in the search box (e.g., term1 OR term2 OR term3 OR term4 OR term5), as Google Search does not impose a limit on the number of OR terms you can include in a query

Rubrics: +1.5 points for answering no ; +1.5 points for the correct explanation.

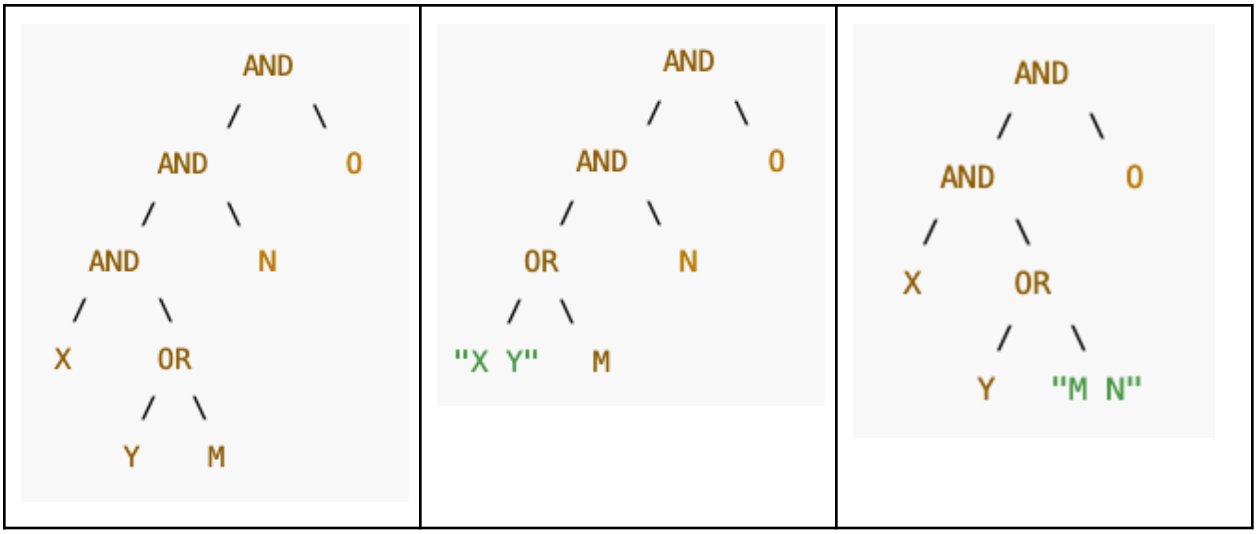
Q3 (1 + 1 + 1 = 3 points):
Draw parse trees for the following Boolean searches [that we’d enter in the regular search box].
a. X Y OR M N O

b. “X Y” OR M N O

c. X Y OR “M N” O

Solution

(a) X Y OR M N O	**(b) “X Y” OR M N O**	**(c) X Y OR “M N” O**
X AND (Y OR M) AND N AND O	**(“X Y” OR M) AND N AND O**	**X AND (Y OR “M N”) AND O**



Rubrics: +1 for each example with a correct parse tree.

Q4. (3 points)

The Mean Reciprocal Rank is a measure used to the quality of ranked answers. From our slides:

- The mean reciprocal rank is the average of the reciprocal ranks of results for a sample of queries

$$\text{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\text{rank}_i}.$$

- For example, suppose we have the following three sample queries for a system that tries to translate English words to their plurals. In each case, the system makes three guesses, with the first one being the one it thinks is most likely correct:

Query	Results	Correct response	Rank	Reciprocal rank
cat	catten, cati, **cats**	cats	3	1/3
tori	torii, **tori**, toruses	tori	2	1/2
virus	**viruses**, virii, viri	viruses	1	1

the mean reciprocal rank as (1/3 + 1/2 + 1)/3 = 11/18 or about 0.61.

What is an alternative to the above, that avoids inverses? In other words, suppose we don't want to perform 1/x where x is the rank, what else can we have, for x? Hint: it's quite popular online!

Mean Average Precision (MAP) at K is a quality metric that helps evaluate the ability of the recommender or ranking system to return relevant items in the top-K results while placing more relevant items at the top. We can express it as the following:

MAP = 1/|Q| (sum_1_|Q| (AP@K))

Where:

- K is a chosen cutoff point.
- Q is the total number of queries in the evaluated dataset.
- AP is the average precision for a given ranking list.

AP@K = (# of relevant results at K) / K

Rubrics:

- +1 points if mentions the alternative
- +1 points if explains the alternative
- +1 point if mentions the formula

Q5 (1+2 = 3 points)

We discussed 'agentic AI' quite a lot during lectures.

a. How would you describe an 'agent' [eg what is it comprised of?]

Possible solutions or high level ideas (1 point):

- A system that senses the environment and acts upon it.
- A system with a goal, a form of decision making, or internal model of the world.

b. How would agents be useful in information retrieval? Be sure to provide an example.

If the student uses one aspect provided above (i.e. sensing, decision making, or goal driven) in the context of information retrieval (2 points). For example: 1.) A personalized recommendation system uses the **model of the user's intents and interests** to provide correct search suggestions or 2.) A web-crawling agent conditions the searched URLs under the **goal of finding only files relevant to the task.**

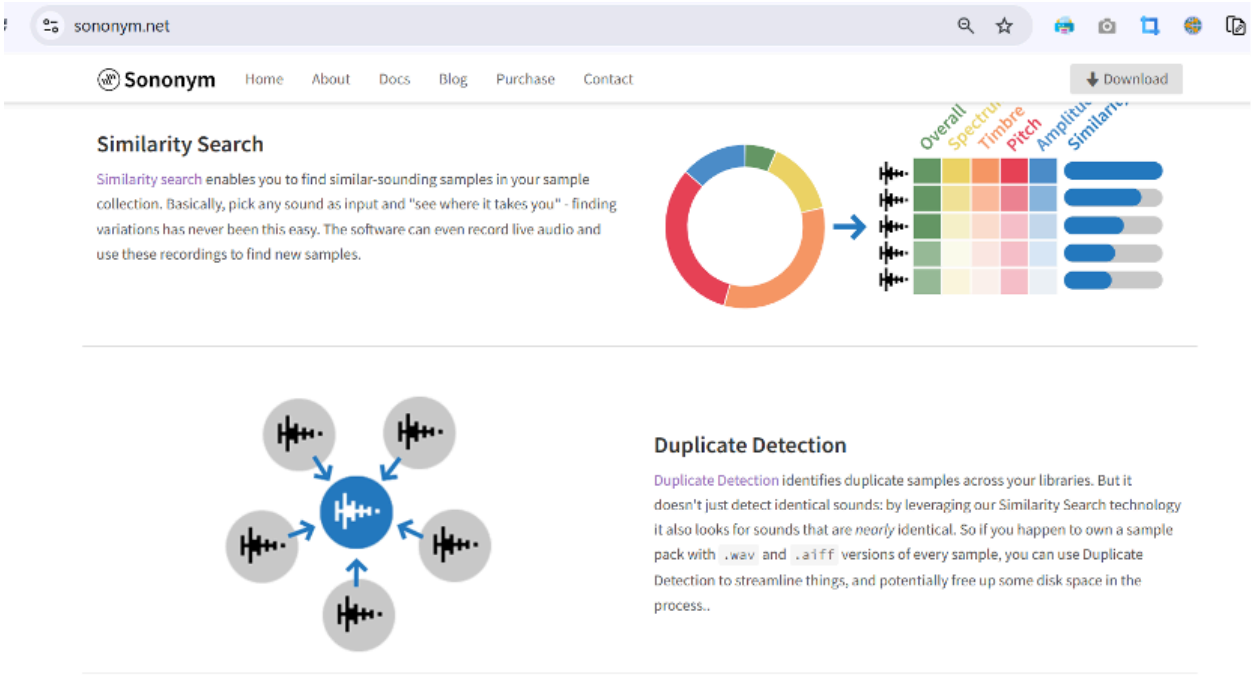
1 point for using the correct attribute. 1 point for giving a correct example in the context of information retrieval.

Q6 (3 points)
How would you explain the above increasing counts for uni- bi-, tri-, four and five grams [ignoring the fact that fourgram count > fivegram count]?
Solution:
Take bigrams and unigrams as an example, there are more possible bigrams than unigrams since a bigram contains more words than a unigram. For a sufficiently large corpus, unigrams repeat more frequently than bigrams. Hence, the number of distinct bigrams is larger than the number of distinct unigrams.

Rubrics:

- +2 points for mentioning there are more possible n-grams as n gets larger
- +1 points for mentioning n-grams repeat less frequently as n gets larger

Q7
The following is a piece of software for musicians:



A 'sample' is simply an audio file that plays for a few seconds, eg. 'blues_guitar_riff.wav', 'congas.mp3', etc. The "see where it takes you" is a recommendation system.

a. What type of recommendation system is the above? How does it work
Solutions:
Content similarity-based recommendation system.

Content similarity-based methods analyze the attributes of items (e.g., metadata, descriptions, or features) to recommend similar items to users. These methods typically use techniques such as TF-IDF, vector embeddings, or deep learning models to compute similarity scores between items. If a user interacts with an item, the system recommends other items with high similarity scores based on their content features.

Rubrics:

- +1 point if the answer correctly identifies the method as a content similarity-based approach or uses equivalent terminology.
- +1 point if the answer provides a reasonable explanation of how content similarity-based methods work (i.e., finding similar items based on item features).

b. what is another type of recommendation system (just name it)?

Solutions:

collaborative filtering, co-visit patterns, association rule mining, or similar methods. popularity-based recommendation is also acceptable.

Rubrics:

- +1 point if the answer includes any valid technique related to collaborative filtering, co-visit patterns, association rule mining, or similar methods.
- No points if the answer refers to content similarity-based algorithms.
- No points if the answer only mentions specific products (e.g., YouTube).

Q8 (2+1 = 3 points).
What is the big/seismic/tectonic/gigantic/massive/'yuuuu...ge' shift in IR?
a. explain in a paragraph or two.

Solutions:

The field of IR is undergoing a significant transformation with the integration of Large Language Models (LLMs) and generative AI technologies. Traditionally, IR systems focused on retrieving and ranking documents based on keyword matching and relevance metrics. However, the advent of LLMs has enabled a shift toward semantic-based search and generative information retrieval, where AI systems can understand and generate human-like text, providing more nuanced and contextually relevant responses to user queries. This paradigm shift allows for more interactive and conversational search experiences, moving beyond simple document retrieval to generating tailored information that aligns closely with user intent.

Rubrics:

- +1 for mentioning large language models and advancements in NLP or ML
- +1 for mentioning semantics and text understanding

b. why is the shift occurring now (why not 10 years prior, for ex)?

Solutions:

Advancements in

- Model architectures (specifically transformers)
- Computational power (GPUs and TPUs)
- Data availability (accumulation of vast amounts of digital content).

Rubrics:

- +1 mentioning or hinting to transformers

Q9 (1+1+1 = 3 points).

What could be an alternative to information retrieval, in certain (but not ALL) cases? Hint #1: read the question carefully - think of each word! Hint #2: this is related to Q8 above :) Also, be sure to provide two examples.

Solutions:

Alternative to Information Retrieval (IR) in Certain Cases

An alternative to **Information Retrieval (IR)** in certain (but not all) cases is **Information Generation (IG)**. Instead of retrieving pre-existing information from a database or indexed documents, **IG dynamically produces new content, insights, or responses** using advanced AI models. This approach is particularly valuable when **pre-existing data is insufficient, outdated, or too general**, and when users seek **context-aware, synthesized, or creative outputs** rather than exact matches from a repository.

Hint #1: Shift from "ALL Words" to Meaning & Context

The question hints at a fundamental shift from **keyword-based retrieval (exact word-matching)** to **semantic search and AI-driven understanding**. Traditional IR systems focus on retrieving documents based on keyword presence, requiring an exact or near-exact match between search terms and stored data. However, modern AI systems, such as **transformer-based models (e.g., BERT, GPT-4, and Mistral)**, interpret **context and meaning**, allowing them to retrieve or generate relevant responses **even if the exact words are not present**. This marks a move away from retrieving documents based solely on keywords toward **retrieving meaning** or even generating entirely new content.

Hint #2: The Big Shift in IR Due to Recent Breakthroughs

The **big shift in IR** is driven by recent advances in **Artificial Intelligence (AI) and Large Language Models (LLMs)**, enabling **Information Generation (IG)** as a viable alternative. Traditional IR has relied on **indexing and ranking algorithms** to fetch the most relevant documents. However, the introduction of AI-powered models means that rather than retrieving information verbatim, systems can **generate contextually relevant, synthesized, or creative responses** based on massive knowledge bases and real-time data processing. This shift is occurring now because of breakthroughs in:

- **Deep Learning & Transformers** (BERT, GPT, Gemini, Claude)
- **Neural Information Retrieval & Semantic Search**
- **Reinforcement Learning for Adaptive AI Responses**
- **Integration of Multi-Modal AI (text, image, audio synthesis)**

This transition redefines the role of IR, as AI-powered systems **don't just find information—they create it.**

10 Examples of Information Generation (IG) as an Alternative to Information Retrieval (IR) Across Different Domains

1. Legal & Compliance: AI-Generated Legal Contracts & Case Summaries

Description: Instead of retrieving standard contract templates, AI **generates customized legal documents, case law summaries, and compliance reports** tailored to specific client needs and jurisdictions.

2. Healthcare & Medicine: AI-Generated Personalized Treatment Plans

Description: Rather than retrieving past patient cases, AI **analyzes medical history, lab results, and genetics to generate individualized treatment recommendations.**

3. Finance & Investment: AI-Generated Market Predictions & Portfolio Strategies

Description: Instead of retrieving old financial reports, AI **analyzes real-time economic data and trends to generate investment insights and risk assessments.**

4. Journalism & Media: AI-Powered Automated News Reporting

Description: Instead of retrieving past news articles, AI **generates real-time news summaries, investigative reports, and financial market updates** based on live data streams.

5. Software Development: AI-Assisted Code Generation & Debugging

Description: Rather than retrieving code snippets from documentation, AI **generates new, optimized code, automates debugging, and suggests software improvements.**

6. Education & Training: AI-Created Personalized Learning Modules

Description: Instead of retrieving static textbooks or course materials, AI **dynamically generates customized lesson plans, quizzes, and interactive explanations** based on student progress.

7. Business & Marketing: AI-Generated Advertising Campaigns & Content

Description: Instead of retrieving pre-existing marketing templates, AI **creates targeted advertisements, personalized email campaigns, and SEO-optimized content** based on consumer data.

8. Scientific Research & Discovery: AI-Generated Hypotheses & Experiment Designs

Description: Rather than retrieving prior studies, AI **synthesizes research insights, proposes new scientific hypotheses, and suggests experimental methodologies.**

9. Customer Service & Support: AI-Generated Dynamic Chatbot Responses

Description: Instead of retrieving predefined FAQs, AI-powered assistants **generate context-aware, conversational responses tailored to user inquiries.**

10. Creative Arts & Entertainment: AI-Powered Storytelling & Music Composition

Description: Instead of retrieving past creative works, AI **generates original scripts, books, songs, and artwork** based on user prompts and stylistic preferences.

Rubrics: Discuss an alternative method to information retrieval (+1 Point). Provide the 1st example of this alternative method (+1 Point). Provide the 2nd example of this alternative method (+1 Point). Any alternative method discussion and examples are acceptable if they sound logical. Avoid a double penalty if the alternative method is wrong, but they provide two examples according to their alternative method; give them +2 points.

Q10 (1*3 = 3 points):

OMG - a ‘gimme’ question - name, and discuss, three ‘similarity’ measures that are used in IR.

Solution: Describe three methods among Euclidean Distance, Manhattan Distance, KL divergence, cosine similarity, Jaccard Index, Similarity Hash (Hamming Distance), or any other valid similarity measure.

Euclidean Distance: Measures the straight-line distance between two points in a multi-dimensional space; commonly used to find the shortest path in continuous space.

Manhattan Distance: Also known as "taxicab" distance, it calculates the distance between two points based on only vertical and horizontal paths, ideal for grid-based layouts.

KL Divergence: A measure of how one probability distribution diverges from a second, expected distribution; often used in information theory to compare distributions.

Cosine Similarity: Calculates the cosine of the angle between two non-zero vectors in a multi-dimensional space, quantifying the similarity in direction regardless of magnitude.

Jaccard Index: Measures the similarity between two sets by dividing the size of their intersection by the size of their union, useful in comparing binary data or categorical sets.

Similarity Hash: A hashing technique that maps similar inputs to the same or nearby hash values, helping in efficiently identifying near-duplicate data in large datasets.

Rubrics:

- 0.5 points for correctly naming each similarity measure (total: 1.5 points)
- 0.5 points for correct description of each measure (total: 1.5 points)

Q11 (2+1 = 3 points)

The idea of using a tree to metaphorically represent the structure of human knowledge is not new. The first metaphorical tree used to represent human knowledge dates back to sometime between AD 268 and 270.[13] This first metaphorical tree of knowledge is known as the Porphyrian tree, named after Greek philosopher and logician Porphyry. The Porphyrian tree reframed and developed Aristotle's classification scheme by presenting it in tree form.

Trees are not only useful, but also fundamentally critical in explaining the truth of reality. Charles Darwin writes, "The affinities of all the beings of the same class have sometimes been represented by a great tree. I believe this simile largely speaks the truth."[15] In fact, in Darwin's seminal book *On The Origin of Species*, the only illustration in the entire book is a tree structure.[16]

What (historic, but still used) IR scheme uses a 'tree', ie. a hierarchy? The tree scheme was employed for a while for organizing the web (as a form of IR), but was abandoned after.

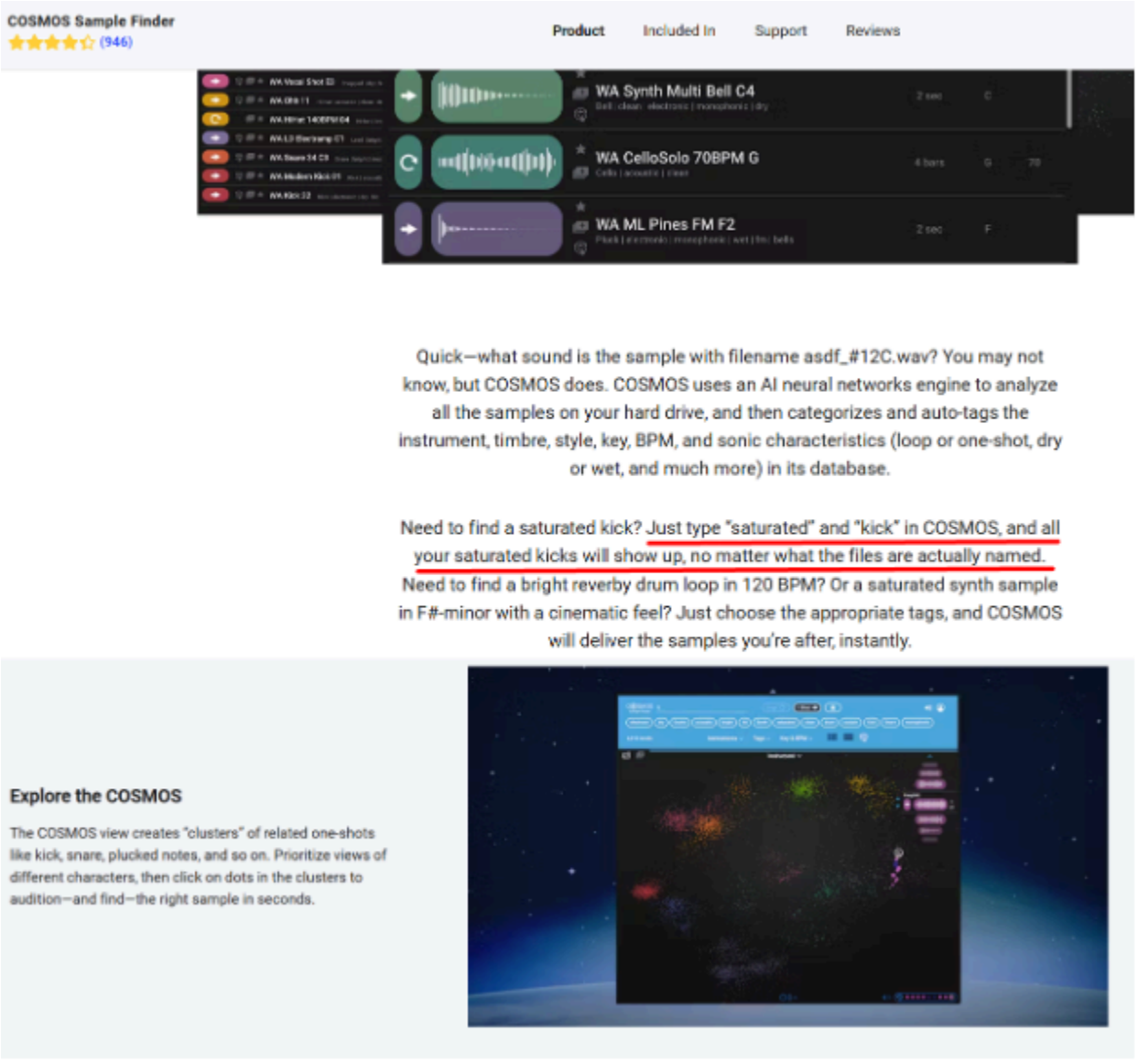
Solution: A historic but still-used information retrieval scheme that employs a tree or hierarchy is the Boolean model with a hierarchical classification system, organizing data into a tree-like structure, where each level represents a category, and classifiers at each level use Boolean logic to make decisions such as Yahoo! Directory (historically).

In the early days of the web, hierarchical directory structures were used to organize and retrieve information. As the web rapidly grew, maintaining and updating a manually curated directory became impractical, and search engines using automated indexing and ranking algorithms took over.
a. Incorrect. Tree/hierarchical classification scheme examples are: Boolean model, Directory, Dewey Decimal, or Library of Congress.
b. As the web rapidly grew, maintaining and updating a manually curated directory became impractical, and search engines using automated indexing and ranking algorithms took over.

Rubric:

- 2 pts for the scheme and explanation.
- 1 pts for the reasons.

Q12 (2+1 = 3 points).
The following is also (like in Q7), a piece of software for musicians:



a. what IR algorithm [that we covered in class] does it relate to? Explain.

b. what is the spatial data structure that is related to the algorithm in a.?

Solution

A: Inverted indexing. As is mentioned in the introduction, the system uses AI models to label each of the music, including the instrument, timbre, style and so on. When a user searches for a specific tag or combination of tags, the system quickly retrieves the files associated with those labels.

B: Term-document incidence matrix is related to the inverted indexing. In this case, a document is one sound sample in the hard drive, and this matrix is built based on AI model's labels.

Rubric:

A: 1 pts for the correct IR algorithm, 1 pts for correct explanation

B: 1 pts for the correct spatial structure.