Relevant indexes not mentioned for each website.

Q1

3 / 3

In IR, what are similarity measures for? Name and discuss 4 of them that we studied.

Q2

2 / 2

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.

Q3

2 / 2

How is a web browser and a web crawler similar? How are they different?

Q4

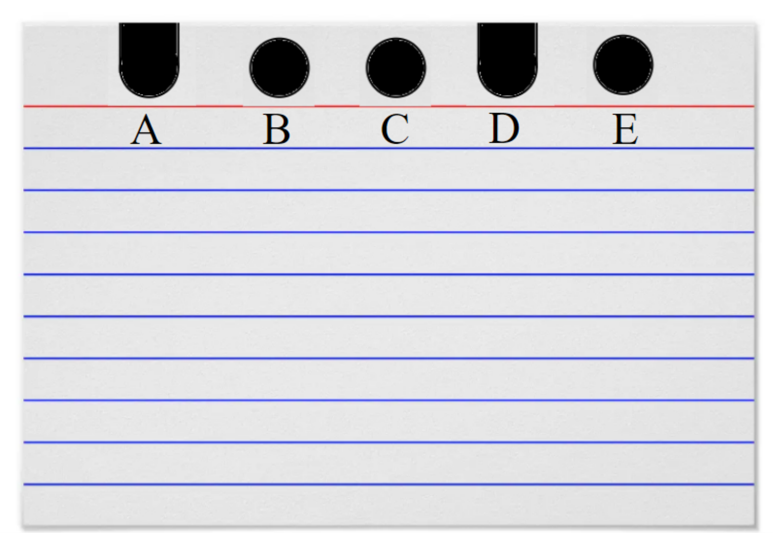
3 / 3

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?

Q5

2 / 2

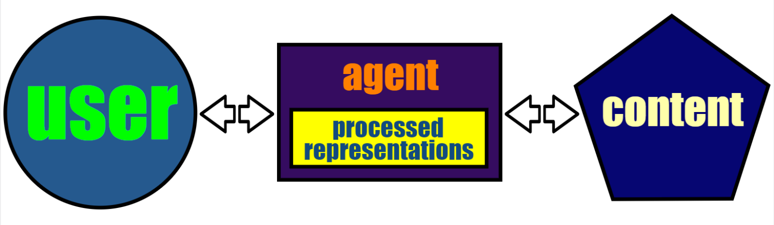
We can use ‘index cards’ to index a small collection of books (THAT is why they are called that - duh!). We use a card for each book that we want to index. Shown below is a sample index card, for a book that belongs in categories B, C, E (intact holes), but not categories A,D (cut-out holes). Given a collection of such cards (eg. 1000 of them) where each card is for a book that could be in 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.



Q6

2 / 3

The following diagram is on the front page of our site:

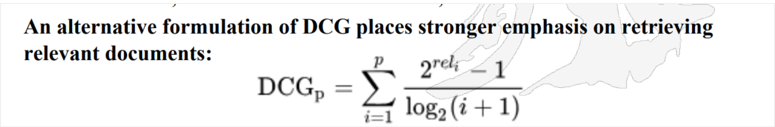


Discuss the role of the agent with respect to the user, and, with respect to the content? In other words, what goes on in the left pair of arrows, and, in right pair?
What is different about the agent's behavior, when the agent is based on a large pretrained language model, eg when it is ChatGPT integrated with Bing, or Google's Bard?

Q7

2 / 3

What is the purpose of (reason for) discounting cumulative gains, when we rank search engines?
We typically use 1/log2(rank+1), for discounting (because 1/rank is too 'harsh'). What other rank-based function can you think of (which is also less harsh than 1/rank)?
To boost relevance, use this alternative formula:



An alternative formulation of DCG places stronger emphasis on retrieving relevant documents:

$$\mathrm{DCG_p} = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{\log_2(i+1)}$$

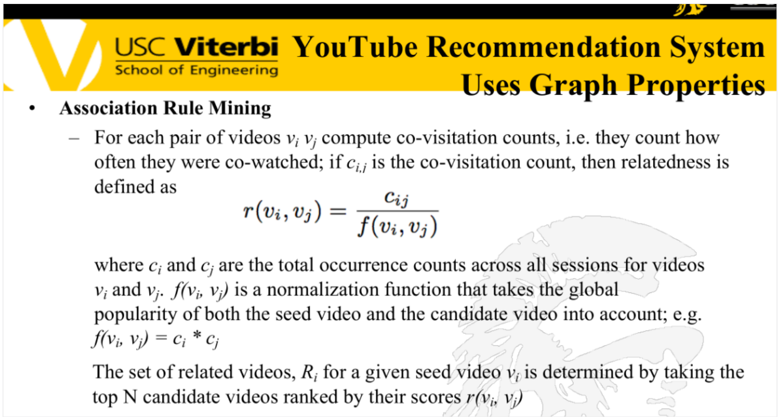
What other function (of relevance) can we use for this purpose?

Part 1 not answered

Q8

2 / 2

YouTube came up with a relatively simple way to create recommended videos:



USC Viterbi School of Engineering

YouTube Recommendation System Uses Graph Properties

- **Association Rule Mining**
  - For each pair of videos v_i v_j compute co-visitation counts, i.e. they count how often they were co-watched; if $c_{i,j}$ is the co-visitation count, then relatedness is defined as

$$r(v_i, v_j) = \frac{c_{ij}}{f(v_i, v_j)}$$

where c_i and c_j are the total occurrence counts across all sessions for videos v_i and v_j. $f(v_i, v_j)$ is a normalization function that takes the global popularity of both the seed video and the candidate video into account; e.g. $f(v_i, v_j) = c_i * c_j$

The set of related videos, R_i for a given seed video v_i is determined by taking the top N candidate videos ranked by their scores $r(v_i, v_j)$

Rather than such 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?

Q9

3 / 3

For 'power searching' Google, we use search modifiers such as :site, :filetype, :intext, etc. These help narrow down the search
What three additional keywords (search modifiers like the above) can you think of (that would be useful to people if Google added them), to narrow down searches to specific categories: "when I enter a search term/phrase X, I want it to be narrowed down to Y category"? Think broadly!

Q10

0.5 / 2

Sites such as OfferUp, uspto.gov, eBay, RateMyProfessor etc offer specific services (eg we don't search for an ML pdf on OfferUp). Given that, how would (or do) the four sites mentioned above, make it easy to search for what they offer? In other words, what might each one index (so that we can search those indexes)?

Q11

2.5 / 3

Characterize Rocchio, kNN and nearest-neighbor techniques for document classification (where we assign an incoming document, one of 'n' existing classes), in terms of

- aggregation (averaged or not)
- geometry (point, line, area...)
- robustness

In other words, for each technique, talk about the three aspects listed above.

Q12

2 / 2

Names two sites that you use, where 'approximate matches' are performed and results displayed, when a user enters a query (searches for something)?