CS47300 Fall 2020 Assignment 5 Solutions

1 Data Privacy

1. How does personalization in search results impacts data privacy? If a search engine that tries to personalize your search results based on a profile you have provided wants to release search log data (queries and results) to the public, would user ID anonymization be enough? Explain why or why not. Hints: AOL Query Log Debacle.

   While personalization can improve search, it can also exacerbate privacy issues. For example, suppose you are working with your boss, and a query on python coding returns a lot of job opportunities. When your boss tries the same query, the results don't include job opportunities. What would your boss think you are interested in? If we think of the AOL query log problem, reporters identified someone from content of their queries. This could be made even easier if the results were personalized (for example, only child-safe results, suggesting the person issuing the query was underage.) This is just one example, there are many good answers to this.

2. What if users do not trust search engines and do not want to know/store more about users. However, you might already know enough in this class regarding how to form a user profile using context, search history, and query intent, improving search performance. Thus, there is a trade-off between protecting the personal information of users and giving better search results. Can we get the best of both? We want better search results without giving too much user information to search engines (not trusting search engines). Propose an idea of how this can be achieved? You need to describe the approach and how the expected trade-off can be achieved. Please be concise; half a page should be adequate. This is an active research area; you might find this SIGIR 2016 paper gives you some ideas.

   One approach to search privacy is based on "cover queries" - giving both real and fake queries to the search engine (see, for example, `http://trackmenot.io/` or `https://doi.org/10.1137/1.9781611972795.66` . This makes personalization more difficult. But as discussed in the paper above, we can make the "cover queries" similar enough to our interests to personalize results, but different enough that the potentially harmful or embarrassing results are likely a result of the cover queries than your actual interests.

2 Filter Bubbles

TikTok is a very popular platform where users can create, share and watch short videos. User's "information needs" are met through recommendations, based on profiles and collaborative filtering rather than explicit queries.

1. Have you used TikTok? How do TikTok recommend videos for you? How do you feel these are different from other recommender systems (for example, Amazon)?
   Yes, I have used TikTok. I think TikTok mostly used user profile to recommend videos. TicTok's recommendation mainly based on my follows, likes and browse history. It will recommend videos in the same topic to me. It's different from some other recommender systems like Amazon. TicTok generates recommendation when I slide on the APP and watch videos. It do not need me to search something. While Amazon's recommendation mostly based on search history and buying history.

2. Do you think TikToc is likely to produce filter bubbles? Give some reasons, and describe possible advantages and disadvantages of this.
   Yes. TicTok is likely to produce filter bubbles. Because it recommend videos based on our browsing history. When we watch a video, we'll find more similar videos on the interface,

and we may watch these similar videos, then it will recommend more such videos to us. This will cause filter bubbles. The advantage is that the recommended videos are more likely to be something we really are interested or like. However, the disadvantage is that the users are limited to watch on similar content.

3. How might you improve the recommender system to avoid generating filter bubbles?
   Besides recommend similar content all the time, the system can recommend some new content to users. For example, they can recommend some popular videos.

If you are not familiar with TikTok, you may describe and answer the above questions another system you are familiar with that is based on recommendations and/or profile-based filtering rather than ad-hoc queries.

3 Bias

Give an example of how an ad-hoc retrieval system might show bias in answering a query. Address: **A job recruiting platform**

1. What group is the system biased towards/against?
   It may biased against females and minorities.

2. What sort of harms might arise?
   The website may show more Caucasian males to the recruiter than females or minorities. So females and minorities may have less chance of being hired.

3. How might this have happened?
   It may because there are more male candidates on the position. It may also because the algorithm is trained by biased dataset.

4. How you might modify an information retrieval system to address this problem, and how might that affect scores for how "good" the ad-hoc retrieval system is. (E.g, precision/recall/rank-based metrics.)
   Firstly, I'll check if the training set is biased or not. If not, I'll try to train the algorithm with more fair dataset. I think this will increase both precision and recall of the system. I'll set up a fixed number of minority candidates and show that to the recruiter. I think this may decrease the precision and recall a little. But it will make the system become more fair.

4 Fake News Detection

Analyze the following statements, in terms of their likelihood of being fake news.

1. "Neural networks are a bubble, the next big learning algorithm will make them irrelevant, as happens with all things."

2. "Neural networks are a bubble, the next big learning algorithm will make them irrelevant. "

   You should discuss:

   - How would you go about determining the legitimacy of such a statement?
   - Would you expect a consensus on the statement?
   - What are some difficulties about identifying the truth of such a statement?

   For the first statement, I would consult authorities in the field, as that's the most reliable way to identify whether this statement is valid or not. For the most part, it is an opinion. However, not all news can be purely facts, and people who want to know what authorities in the field of AI and

machine learning are saying about neural networks would want to see this content. Also, analyzing it's word usage could be effective in determining whether it is "fake news" or not. There are a variety of good answers here, but the most reliable is consulting authorities or identifying the reliability of the source in this particular case. Part of the difficulty of identifying the truth of a statement like this is that consensus is NOT likely, but that doesn't necessarily mean it's fake news or not worth detecting. It's important to consider that fake news doesn't necessarily include all opinions, more that it should encompass things directly misinforming the public and harming the public because of it. Worth also noting, science isn't a democracy. Just because there may or may not be a consensus doesn't necessarily indicate the assertion is fact or fiction.

The same as the first answer essentially applies to the second, the slightly different wording would make it less likely to get flagged as fake news by a syntactic fake news detector. Lacking the last phrase would affect how a system analysing word use and phrase structure to determine fake news will predict this phrase.

3. For fake news, do you think false positives (wrongly identifying an article as fake) or false negatives (failing to identify a fake article) to be a worse problem. Briefly discuss why.

   I think that false positives are the bigger issue. Modern approaches to fake news take a fairly hardline stance on fake news, usually opting to simply remove said fake news media piece from circulation. This means if false positives are high, lots of articles may wrongly be removed from circulation, resulting in some very frustrated clientele. On the other hand though, fake news spreads very quickly and misinformation can be very publicly damaging as we've seen in recent events. There is a case to be made for both sides and dangers to an abundance of either.

4. How does this relate to using recall and precision as measures of the ability to detect fake news?

   An abundance of false positives would affect the precision score and not recall, so if you wanted a model that controlled for false positives and not false negatives, you would want to depend on precision to evaluate performance. The opposite is the case for recall; recall does not take into account false negatives, so if you only wanted to pay attention to false negatives then recall is the metric to evaluate with. You would expect high false positive totals to result in low precisions but high recalls, and high false negative totals to result in low recall scores. This connects back to the previous question in the sense that, if you're trying to please the posters on your media sharing platform, an abundance of false positives is a concern and so you'd probably want to pay more attention to precision. Alternatively, if you are more concerned with removing fake news from circulation, you'd be more preoccupied with recall as a statistic.

5 Federated Search

You are constructing a federated search based on the idea of constructing a sample dataset, and using query results on the sample to determine both which sources to run a full query on, and how to merge the results.

Part 1: Sample Dataset Construction You are given 100 sample documents from Server A (which has a total of 100,000 documents), 50 from Server B (which has a total of 1000), and 50 from Server C (which has a total of 10,000).

Server D tells you how many total documents contain a term. You issue a query for "Information", and retrieve the top 30 documents but are told there are 500 containing the word. You then issue a query for "Virus" and retrieve the top 30 documents, you are told 50 match. There is one document that appears in the top 30 results for both queries.

Estimate the total number of documents in Server D.

You then do the same two queries for Server E. Server E gives you 50 documents containing "Information" and 50 containing "Virus", there are 10 documents that appear in both top 50 lists. Estimate the total number of documents in Server E.

Answer:

Server D: $N = \frac{30*30}{1} = 900$

Server E: $N = \frac{50*50}{10} = 250$

Part 2: Source selection You run a query "Indiana Covid-19 statistics". the sample and get 1 matching document from Server B and 3 matching documents from Server C. The ranking is $D_{Cs1}, D_{Bs1}, D_{Cs2}, D_{Cs3}$.

What servers would you recommend querying, and what query would you use, if your goal is to get reasonable results for "Indiana Covid-19 statistics" without using too much bandwidth / computing resources.

Answer:

Option 1: Based on ReDDE, compute the Rel_Q, we get Server C has a higher ReDDE and we can conclude that Server C can help us to save more computing resources.

Option2: We have a 5% sample from Server B, and 1 matches. This means we would expect 20 to match in a full query of server B. From Server C, we have a 0.5% sample, and 3 matching, so we would expect 600 matching.

We have only a .1% sample of Server A. If there were 50 matching documents, the probability one would show up in the sample is less than 0.5. So to achieve high recall, Server A is probably a better choice than Server B. We have a much higher percentage from Server D and E, so the fact that we didn't get any in the sample suggests there are few if any.

If we assume the full query gets us 600 from Server C and 20 from Serer B, and the ranking is similar to the sample, then even if we assumed Server B and C were very close in which had "top ranked", and only 1/3 of the Server C results were top ranked, that would still mean 10 top results from server C for every one from Server B. So sending the query "Indiana Covid-19 Statistics" to only Server C is likely to get us a good "top 10", also sending the query to Servers A and B would get us a bit better recall.

Option 3: Remove the "statistics" word, can also help to reduce bandwidth.

Part 3: Results merging Suppose you issue the full query Indiana Covid-19 statistics to only Servers B and C. You get 10 documents from Server B, with the matching sample document D_{Bs1} 3rd on the list. From server C, you get 10 documents, D_{Cs1} is third on the list and D_{Cs2} is 5th on the list.

Come up with a means of merging the documents. You don't have scores, you only have rankings. Give your combined ranked list of 20 documents, and show how you determined the proper ordering. Note that this will require you to make some assumptions and estimates that combine different approaches we discussed in class.

Answer:

We know that D_{Cs1} is ranked higher than D_{Bs1}, so a most basic approach would be to assume that for two things at the same rank, the item from Server C should go higher. We could also make use of having the top 1/2 from server B, and only the top 1/60 from Server C, so this also suggests the top documents are from Server C - suggesting few from Server B should be high ranked. But since we know that the best three from Server B are better than the 5th from Server C, this also suggests at least 6 from Server B should go in the top 20.

There are many possible correct answers, but any correct answer should have the top three from Server B above the 5th from Server C, and the top three from Server C above the third from Server B.