Homework #4 – Task 2 Evaluation Report

Yoke Kai Wen, 2020280598

April 4, 2021

1 Introduction

In this report, I analyse annotation consistency between different annotators on the relevance of search query results, analyse search engine (SE) performance, analyse my VSM and BM25 ranking models' performance, and comment on difficulties faced during evaluation.

2 Annotation consistency

2.1 Quantitative analysis of annotation consistency

Figure 1 shows the Kappa scores for query relevance judgements from different annotators. Cohen's Kappa coefficient measures inter-judge (dis)agreement - for complete agreement, a Kappa score of 1 is achieved; if there is no agreement among raters other than what would be expected by chance, the Kappa score goes below 0.

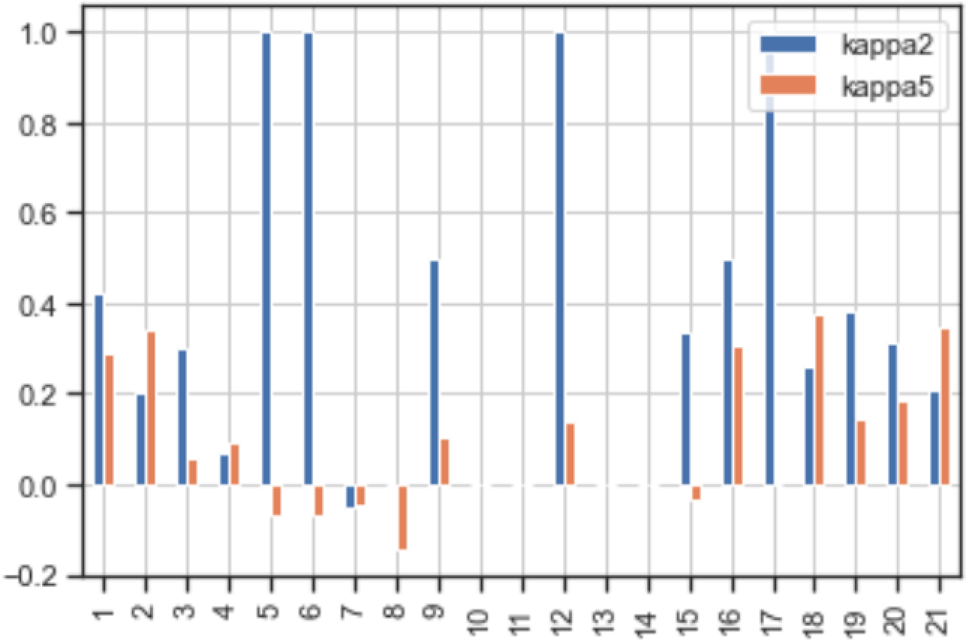


Figure 1: Bar chart showing 2 and 5-levels Kappa scores across 21 queries

Figure 1 shows varying Kappa scores for each query. Furthermore, Kappa2 and Kappa5 scores differ significantly in most cases. For queries 10, 11, 13, 14, there are no scores because there is only one annotator. After analysing the Kappa scores and the actual queries, I find the Kappa measure not a reliable and suitable measure of annotation consistency in this case, particularly for Kappa5, for the following reasons:

1. Number of query results annotated is too small and cannot reflect accurately the agreement by chance

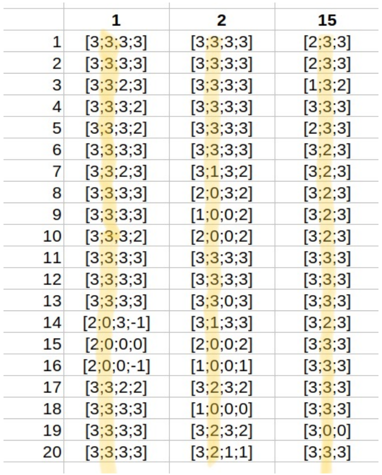
   Firstly, the number of query results for each query is too small (only 20), hence the calculation of $P(E)$ which represents what agreement would be by chance is too inaccurate. For instance, for query 8, there were 19 agreements and 1 disagreement, but using the Kappa2 formula, the Kappa2 score = 0, because the agreement by chance is also very high, which negates the high proportion of agreement. If there were more query results that are annotated, the agreement by chance could be computed with greater reliability.

2. Degree of relevance is not exactly categorical and hence Kappa5 is not a suitable measure

   There are 5 categories of relevance: -1, 0, 1, 2, 3. However, these are not separate and unrelated categories, because category 2 and 3 are more similar than category 0 and 3. However, the Kappa5 formula treats all categories as unrelated and independent from each other, resulting in Kappa5 score being very low even when the relevance annotations are subjectively quite close to each other. For instance, in queries 5 and 6, all the queries had annotations of 1/2/3 resulting in perfect Kappa2 scores, but negative Kappa5 scores, which is unintuitive.

2.2 Comparing my annotations (from HW2) with other annotators

Figure 2 shows my relevance annotations compared to other annotators for the three queries I annotated. Query 15 is ”best lasagna recipe how to make” , Query 1 is ”benefits of running regularly” and Query 2 is ”buy computer screen”.



	1	2	15
1	[3;3;3;3]	[3;3;3;3]	[2;3;3]
2	[3;3;3;3]	[3;3;3;3]	[2;3;3]
3	[3;3;2;3]	[3;3;3;3]	[1;3;2]
4	[3;3;3;2]	[3;3;3;3]	[3;3;3]
5	[3;3;3;2]	[3;3;3;3]	[2;3;3]
6	[3;3;3;3]	[3;3;3;3]	[3;2;3]
7	[3;3;2;3]	[3;1;3;2]	[3;2;3]
8	[3;3;3;3]	[2;0;3;2]	[3;2;3]
9	[3;3;3;3]	[1;0;0;2]	[3;2;3]
10	[3;3;3;2]	[2;0;0;2]	[3;2;3]
11	[3;3;3;3]	[3;3;3;3]	[3;3;3]
12	[3;3;3;3]	[3;3;3;3]	[3;3;3]
13	[3;3;3;3]	[3;3;0;3]	[3;3;3]
14	[2;0;3;-1]	[3;1;3;3]	[3;2;3]
15	[2;0;0;0]	[2;0;0;2]	[3;3;3]
16	[2;0;0;-1]	[1;0;0;1]	[3;3;3]
17	[3;3;2;2]	[3;2;3;2]	[3;3;3]
18	[3;3;3;3]	[1;0;0;0]	[3;3;3]
19	[3;3;3;3]	[3;2;3;2]	[3;0;0]
20	[3;3;3;3]	[3;2;1;1]	[3;3;3]

Figure 2: Comparing my annotations (highlighted in yellow) with other annotators for queries 1, 2, 15

My annotations are mostly consistent with other annotators for query 15, having more inconsistencies for query 1, and having many inconsistencies for query 2. Some of these inconsistencies can be attributed to the lack of clarity in the queries, especially for query 2 which is extremely unclear. Query 2 has only three words, and as a human annotator I am not even sure what the query user was intending to get - recommendations for the best computer screens? Prices and shops for computer screens? Is there a particular brand or desired property like 'touchscreen'? In general, I find myself less generous than other annotators in awarding relevance scores. Below I investigate the query results that elicited vastly different opinions from different annotators:

1. Query 15, BING Rank 3, [1;**3**;2]

   https://www.tasteofhome.com/collection/how-to-make-the-best-lasagna-recipe-ever/

   I gave a significantly higher rating than the other annotators. I believe it is because this website does not give a step-by-step recipe for making lasagne, but rather offers several detailed tips to improve the taste of lasagne such that it will be the best lasagne ever. In my opinion, this is completely relevant to the query which is trying to find out how to make the best lasagna, since there is the implication that the query user already knows how to make lasagana and is merely trying to find out how to improve. Therefore, even though there was no step-by-step recipe, I thought the tips were fully relevant.

2. Query 15, BAIDU Rank 9, [3;**0**;0]

   http://www.baidu.com/link?url=UeVJOlh0ZWmwBC4iEPOsQV4bDtB8DjbCgarVJHrKtyUsnunQLrfS5Sx

   I gave a lower rating than the other annotators. I found this search result mostly irrelevant, because it is a recipe for lasagna soup rather than lasagna which are completely two different things. However, I did not give a -1 score because the result was technically still a lasagna recipe. I am not sure why one of the annotators would give a score of 3, perhaps the annotator is not aware of what a lasagna dish is.

3. Query 1, SOUGOU Ranks 4;5;6, [2;**0**;3;-1]; [2;**0**;0;0]; [2;**0**;0;-1]

   https://www.healthline.com/health/fitness-exercise/essential-runner-stretches

   https://www.healthline.com/health/essential-exercises-improve-running-technique

   https://www.healthline.com/health/osteoarthritis/knee/alternatives-to-high-impact-exercises

   I gave a lower score than other annotators, because query 1 was looking for the benefits of running, while ranked-4 result was talking about essential runner stretches, ranked-5 result was talking about improving running technique and ranked-6 result was talking about alternatives to high-impact exercises. These are all completely irrelevant. But since the topic was still focused on running, I did not give a score of -1. Some other annotators also gave a similarly low score. I do not understand why some annotators could rate 2 or 3 for these results.

4. Query 2, SOUGOU Ranks 7;8;9;10, [3;**1**;3;2]; [2;**0**;3;2]; [1;**0**;0;2]; [2;**0**;0;2]

   http://www.sogou.com/link?url=hedJjaC291OxPYOT4xv71e-FGWiFI7Ru0ve9Lv8XpMX-2pn0Ul8DNj2W
   .&query=buy+computer+screen

   http://www.sogou.com/link?url=hedJjaC291NhEGhDd4M8il3B1ggCd1T4DzG2x6G8xux7aDKB80Rprm5u
   &query=buy+computer+screen

   http://www.sogou.com/link?url=hedJjaC291NhEGhDd4M8il3B1ggCd1T4DzG2x6G8xux7aDKB80Rprm5u
   LDck-J9wlJ-&query=buy+computer+screen

http://www.sogou.com/link?url=hedJjaC291NhEGhDd4M8il3B1ggCd1T4o3qQK80dJyEN2VKz_
q1Xo8UlM5w3UnvL1fQNuKrxo03GknO8-hc370Kd6vFU6dcW2Nx_aT3w4NTdaP6doPzDEUKd6vFU6dcW2
aT3w4NQHMtTa4ZJ0RQ..&query=buy+computer+screen

I gave a lower score than the other annotators for these four search results, because I find them irrelevant. The query user, in his query description, says ”Plan to buy a computer screen in online shops”, so I used that as my relevance guideline. Ranked-7 result was a blog review post recommending best computer monitors; ranked-8 result was an online shop showing touch screen laptops on sale; ranked-9 result was an online shop showing laptop screen protectors on sale; ranked-10 result was for computer and tablet repair and other services. All irrelevant, but still on the topic of computer and screens, so I did not give a -1 score. I can understand why some of the annotators gave higher scores to the ranked-7 result, as a blog review of best computer monitors could be relevant to the query 'buy computer screen', but it is not what the user wanted according to his query description. For the ranked-8 result, I cannot understand why people could give high ratings - laptops are very different from laptop screens? Ranked-9 and ranked-10 results were even more far-fetched but surprisingly one annotator could still give a high score of 2.

5. Query 2, BING Ranks 3;4;5;10, [3;**3**;0;3]; [3;**1**;3;3]; [2;**0**;0;2]; [3;**2**;1;1]

   https://www.laptopscreen.com/

   https://www.bestbuy.com/site/shop/touch-screen-monitor

   https://www.techrepublic.com/blog/cracking-open/how-to-replace-a-broken-laptop-screen/

   https://www.bestbuy.com/site/searchpage.jsp?id=pcat17071&st=apple+computer+monitors

   Ranked-3 result was an online shop with a dubious interface selling laptop screens of different brands and types - I find this highly relevant to the query and gave a high score of 3, while another annotator might have been disturbed by the website design and thus gave a low score of zero; ranked-4 result was an online shop selling touch screen monitors - I find this partially relevant, but if the user was looking for touch-screen monitors he would have added that to his query, so I gave a moderate score of 1, but apparently other annotators found this result completely relevant; ranked-5 result was a blog article guide on how to replace a broken laptop screen, which is completely irrelevant so I gave a score of zero, but some other annotators still gave a score of 2 - perhaps they think that the user might be interested in looking at a repair option too; ranked-10 result was an online shop selling specifically Apple computer monitors - I found this partially relevant, but also too specific, so I gave a moderate score of 2, while other annotators had varying opinions.

3 SE and ranking model performance

Figure 3 shows the mean metric scores for the different search engines and ranking models. Note that each SE was used for different number of queries (BING 18; BAIDU 4; GOOGLE 8; SOGOU 10; DUCKDUCKGO 2; VSM 21; BM25 21), and hence the comparison is not very fair.

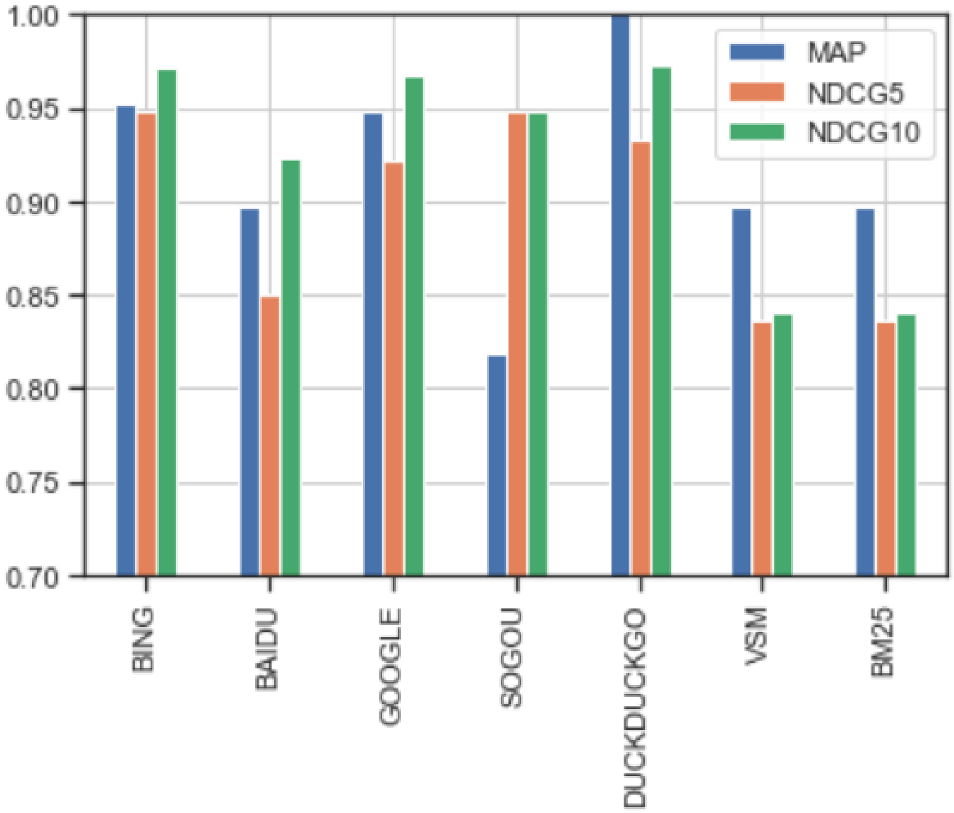


Figure 3: Bar chart showing mean MAP, NDCG@5, and NDCG@10 eveluation scores for all SEs and ranking models

I observe that BAIDU and SOGOU perform significantly worse than the other SEs, and I suspect it is because they are more optimised for Chinese search queries, while the input queries here are in English.
Also, the two ranking models I implemented also performed a lot worse than the SEs, perhaps because these two models merely considered the word occurrence in the html landing pages and ignored semantic meaning and other factors such as number of webpage links.

4 Evaluation difficulties and other comments

1. Firstly, the given data is not very clean. For instance, some query results were annotated by only one annotator, so we cannot calculate Kappa scores for it. For Query 13, the relevance scores of all the results are zero because the results do not match the search query at all - 'cherry red vs gateron red' leads to search results about china travel, clearly indicating that the wrong query was matched to the results. This also led to difficulties computing the NDCG scores, since normalisation involves dividing the scores by zero.

2. Secondly, it is difficult to judge whether we have coded the evaluation metrics correctly. The MAP and NDCG scores do not seem to correlate very well. It is also difficult to compare qualitatively the result ranking with the quantitative metrics.

3. Thirdly, some queries are more specific and of better quality than other queries, so the difficulty of retrieving relevant documents is different for each query. However, different SEs were used for each query, resulting in unfair evaluation.

Overall, this was a very tiring and tedious assignment, and very difficult to analyse and debug. Evaluation of ranking is really not trivial.