Your Name: Raghav Sharma

Your Andrew ID: rvsharma

Homework 5

Collaboration and Originality

1. Did you receive help of any kind from anyone in developing your software for this assignment (Yes or No)? It is not necessary to describe discussions with the instructor or TAs.

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.
   No
2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.
   No
3. Are you the author of every line of source code submitted for this assignment (Yes or No)? It is not necessary to mention software provided by the instructor.

   If you answered No:
   a. identify the software that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

   Yes
4. Are you the author of every word of your report (Yes or No)?

   If you answered No:
   a. identify the text that you did not write,
   b. explain where it came from, and
   c. explain why you used it.

Yes

Your Name: Raghav Sharma

Your Andrew ID: rvsharma

Homework 5

Instructions

1 Experiment: Diversity and relevance baselines

1.1 Experimental results

		Indri	Indri + PM2	Indri + xQuAD	BM25	BM25+ PM2	BM25+ xQuAD
Diversity	**P-IA@10**	0.08000	0.15666	0.08666	0.09583	0.24633	0.24966
	P-IA@20	0.12458	0.18500	0.13666	0.09166	0.21850	0.20816
	αNDCG@20	0.23923	0.32017	0.23637	0.23942	0.46675	0.48006
Relevance	**P@10**	0.1000	0.1300	0.1100	0.1200	0.2900	0.2700
	P@20	0.1550	0.1900	0.1600	0.1350	0.2550	0.2600
	P@30	0.1600	0.2167	0.1800	0.1433	0.2433	0.2400
	MAP	0.0634	0.0451	0.0336	0.0630	0.0758	0.0735

1.2 Parameters

retrievalAlgorithm=BM25
Indri:mu=2500
Indri:lambda=0.4
BM25:b=0.75
BM25:k_1=1.2
BM25:k_3=0.0
diversity=true
diversity:maxInputRankingsLength=100
diversity:maxResultRankingLength=50
diversity:algorithm=PM2
diversity:lambda=0.5

1.3 Discussion

Relevance discussion: The precision or relevance for Indri PM2 and xQuAD improves when compared to the baseline Indri. This is because with PM2 and xQuAD different intents are covered when compared to only baseline Indri. This improves chances of satisfying the information need of the user and improves the precision metrics for P@n and MAP for PM2 and xQuAD.

The xQuAD brings up moderate ranking for the other intents and picks best document for the first intent. Due to this it impacts precision metrics and harms the precision for xQuAD. This is why the precision for xQuAD is lower than PM2 for Indri and BM25.

The MAP scores for diversity algorithms is lower than for Baseline Indri and BM25. The coverage of other intents means that many relevant documents to user will rank lower in the ranking and the user will have to go further down the ranking to find the relevant intent for his information need. This causes a drop in the MAP ranking for the algorithms.

Diversity metrics Discussion: The intent aware precision (IA@k) are a strong indicator of intent coverage than traditional metrics because they take different intents into consideration and produce a ranking based on them. As we notice the baseline Indri and BM25 score low on diversity metrics than does the PM2 and xQuAD, since baseline do not take intents into consideration and the metrics rank results which do not cover different intents lower and give a lower score to them.

On the other hand as the diversity algorithms cover different intents and place them on higher ranks, PM2 and xQuAD perform better.

BM25 being a probabilistic framework performs better in terms of precision ranking than Indri which is language based. This trend is retained in diversity algorithms as we see that Indri precision is lower compared to BM25 precision scores for Diversity P-IA@n and alpha-NDCG@20.

Intent Aware precision for BM25 decreases as we go from IA@10 to IA@20, while it increases for Indri. For PM2 and xQuAD, PM2 is a better ranking algorithm because it covers all the intents properly. While xQuAD will select moderate document for other intents and strong documents for the first intent. This is evident by the higher diversity score for PM2 compared to xQuAD across retrieval algorithms of BM25 and Indri. And this trend also retains when we vary the lambda and retrieval and ranking parameters in the next two experiments.

When we are looking at MAP and P@n, we notice that diversity algorithms do not rank high

2 Experiment: Effect of λ

2.1 Experimental results

	λ=0.0	**λ=0.25**	**λ=0.50**	**λ=0.75**	**λ=1.0**
Indri + PM2					
P-IA@10	0.17916	0.17250	0.15666	0.14916	0.13333
P-IA@20	0.18875	0.19125	0.18500	0.17666	0.15750
αNDCG@20	0.32740	0.32890	0.32017	0.30460	0.28085
Indri + xQuAD					
P-IA@10	0.08000	0.24383	0.08666	0.11666	0.16333
P-IA@20	0.12458	0.22225	0.13666	0.14541	0.19375
αNDCG@20	0.23923	0.23493	0.23637	0.22249	0.32264

	λ=0.0	λ=0.25	λ=0.50	λ=0.75	λ=1.0
BM25 + PM2					
P-IA@10	0.26633	0.24383	0.24633	0.24883	0.16250
P-IA@20	0.22475	0.22225	0.21850	0.21725	0.15916
αNDCG@20	0.45663	0.45126	0.46675	0.46833	0.41201
BM25 + xQuAD					
P-IA@10	0.09583	0.21716	0.24966	0.25883	0.25133
P-IA@20	0.09166	0.19983	0.20816	0.21691	0.21608
αNDCG@20	0.23942	0.47038	0.48006	0.48187	0.46811

2.2 Parameters

The constant parameters that were used to generate the results are given below.
indexPath=outlen/index-cw09
queryFilePath=Experiments/HW5/Input/q.qry.txt
retrievalAlgorithm=BM25
Indri:mu=2500
Indri:lambda=0.4
BM25:b=0.75
BM25:k_1=1.2
BM25:k_3=0.0
diversity=true
diversity:maxInputRankingsLength=100
diversity:maxResultRankingLength=50
diversity:intentsFile=Experiments/HW5/Input/q.intents.txt

2.3 Discussion

Since the lambda parameter has different meaning for PM2 and xQuad, their effect on both algorithms are discussed separately.

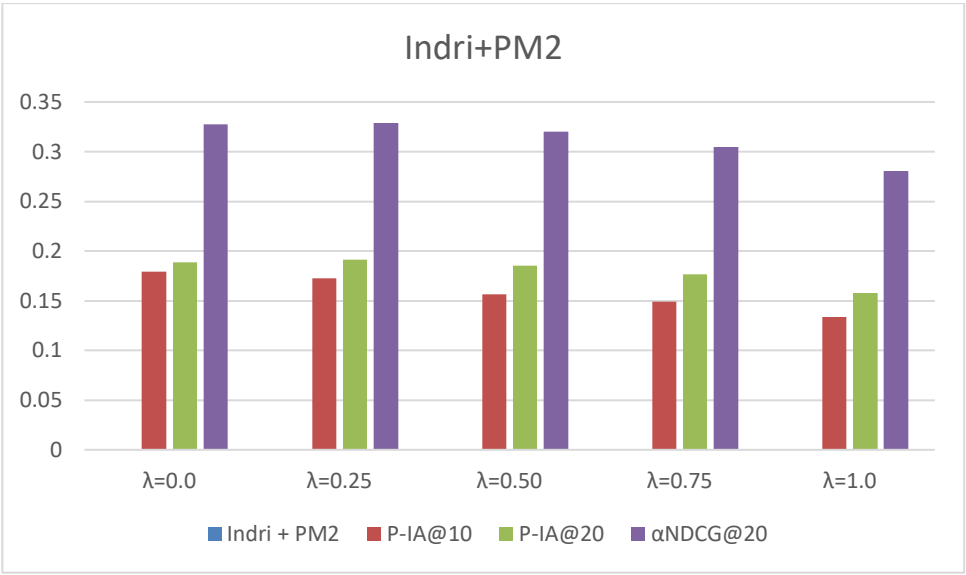
Effect of λ on PM2:

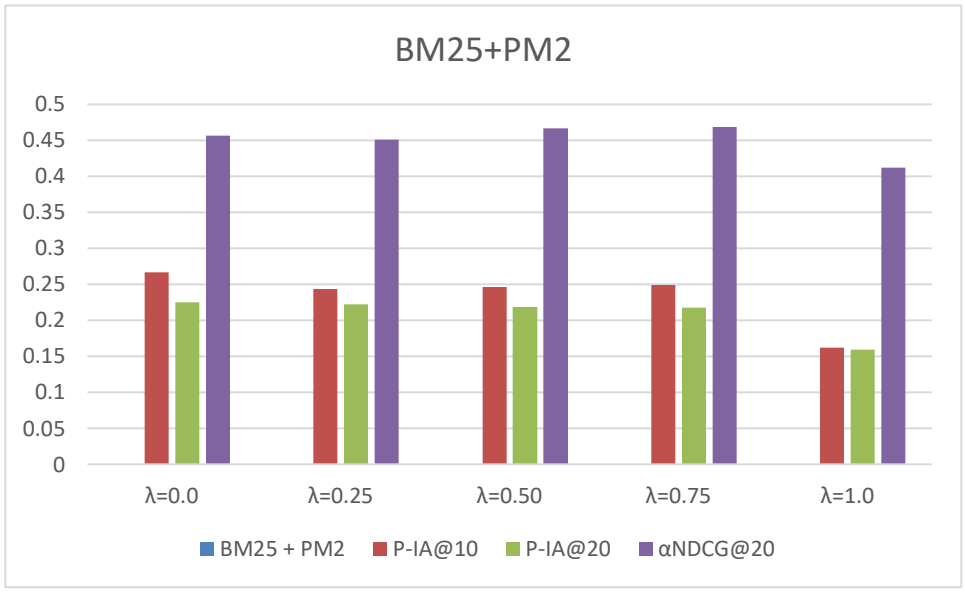
The **λ** parameter for PM2 indicates the priority to the intent that is being considered in the current iteration. Higher the lambda higher the priority for the current intent being considered, and other intents are also given priority, but it is lower than the current priority being considered.

This property ensures that documents which cover other intents too are given a higher weightage compared to a document that covers just one of these intents. This also ensures that when we are covering other intents, we consider documents that cover the current intent well while also taking other intents into consideration when ranking for diversity.

When lambda is 1.0, importance is only given to the current intent and other intents are not considered for the documents. This becomes almost similar to xQuAD and we notice that alpha-NDCG and intent aware precision for PM2 become same to xQuAD. As we reduce the value of lambda to 0.75 and 0.5 we notice an increase in a-NDCG and P-IA. During this setting, other intents are also being considered when ranking the documents in diversified ranking. This leads to the higher a-NDCG and P-IA score and results

in a better balance of ranking for PM2. We observe that a-NDCG is highest for lambda = 0.25 for Indri. And for Lambda.= 0.75 for BM25. The best lambda is than a mean of these two parameter setting when we are considering PM2. So to achieve maximum a-NDCG we choose lambda = 0.5. Intent aware precision fluctuates for different lambda values, but the general trend is a decrease in P-IA@n when the lambda value is increase from 0.0 to 1.0.

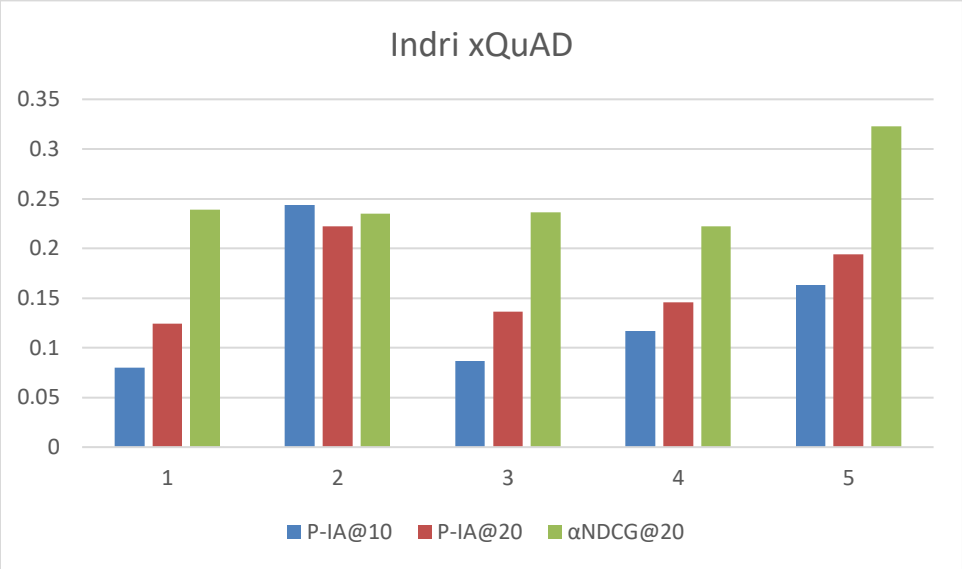


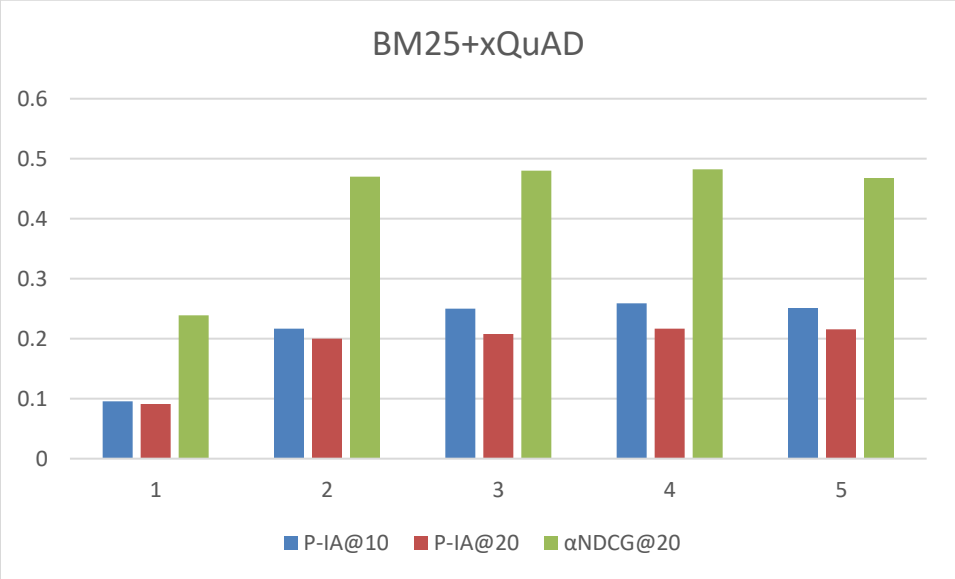


Effect of Lambda on xQuAD:

Lambda of xQuAD is the importance of the diversity of the document being selected. The scoring for xQuAD considers the score of document, considering both its relevance and how different it is from previously selected documents. When lambda = 0.0, only the relevance of the document is considered and the ranking becomes same as Baseline Indri or BM25. However, as we increase lambda from 0.0 to 0.5 or

0.75, we notice an improvement in a-NDCG and P-IA@n. The diversified ranking then considers other documents while ranking and gives a higher weightage to a document if it hasn't been covered. However, xQuAD is not perfectly balanced and selected mediocre documents when considering other intents. Thus it gets a lower A-NDCG and P-IA@n score than PM2 for Indri. However for BM25, xQuAD does better than PM2, this might be due to the nature of BM25, since it is a probabilistic framework and Indri is language model framework.





3 Experiment: The effect of the re-ranking depth

3.1 Experimental results

	25 / 25	50 / 25	100 / 50	200 / 100
Indri + PM2				

P-IA@10	0.14750	0.16083	0.15666	0.17333
P-IA@20	0.13000	0.17425	0.18500	0.20125
αNDCG@20	0.26448	0.31810	0.32017	0.35489
Indri + xQuAD				
P-IA@10	0.09250	0.09833	0.08666	0.08000
P-IA@20	0.12625	0.13458	0.13666	0.13625
αNDCG@20	0.21411	0.21679	0.23637	0.23563

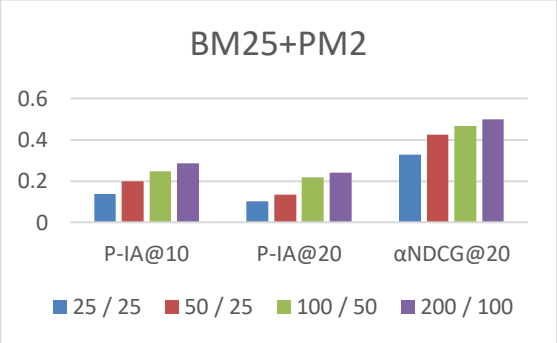
	25 / 25	**50 / 25**	**100 / 50**	**200 / 100**
BM25 + PM2				
P-IA@10	0.13583	0.19933	0.24633	0.28716
P-IA@20	0.10208	0.13466	0.21850	0.24125
αNDCG@20	0.32704	0.42371	0.46675	0.49783
BM25 + xQuAD				
P-IA@10	0.14083	0.19266	0.24966	0.27866
P-IA@20	0.10208	0.13633	0.20816	0.22783
αNDCG@20	0.32900	0.42487	0.48006	0.50846

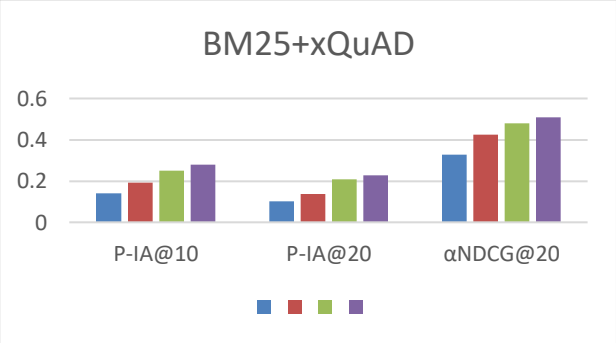
3.2 Parameters

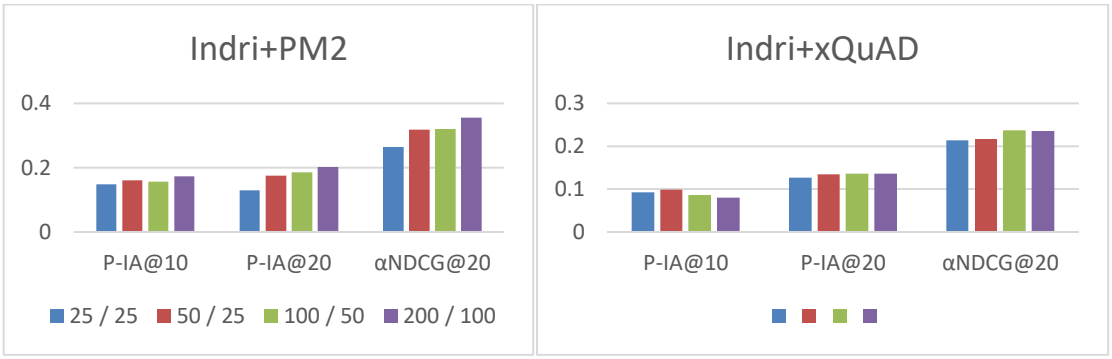
From the above experiment we found **λ = 0.5** to give the best scores for diversity metrics

3.3 Discussion

Effect of Reranking depth is similar for BM25 for both PM2 and xQuAD diversification algorithm. The scores are similar as we can see from the table and graph. When we see the a-NDCG metric, as the reranking depth and the number of initial document, considered increases, the a-NDCG increases as well. The more documents that we have to consider for diversification, the better will be the diversification for the resultant ranking. When only the top 25 documents are considered, the diversification algorithm does not have enough documents for diversifying, however as the number of initial ranking documents considered increase from 25-50, we see a significant jump (~20%) in the a-NDCG parameter. However, as we increase the input ranking documents from 50 to 100 and to 200, we see that the increase %age of the a-NDCG decreases from 20% to 8% and then to 6%. This is not a proportional increase in the a-NDCG and can be offset by the time taken for the algorithm to produce a ranking, since considering 200 document is more computationally costly than reranking 100 or 50 documents.







The trend for Indri PM2 and xQuAD follows a similar pattern except that we go further down the Intent aware precision ranking from top 10 to top 20 scores, we notice that P-IA increases unlike in BM25 where it decreased.

Also, the scores for PM2 are higher when compared to xQuAD which implies that for PM2 the effect of increasing the re ranking document depth has a larger impact than xQuAD. This can be reasoned based on the equation for PM2 which looks into the ranks for the initial document and choses the best document from the initial document scores. More the documents to be considered better is the a-NDCG score and we get a higher increase in a-NDCG.