**Your Name: AKSHAT GAUR**

**Your Andrew ID:agaur**

# 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.

   No

   If you answered Yes, provide the name(s) of anyone who provided help, and describe the type of help that you received.

2. Did you give help of any kind to anyone in developing their software for this assignment (Yes or No)?

   No

   If you answered Yes, provide the name(s) of anyone that you helped, and describe the type of help that you provided.

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.

   Yes

   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.

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

   Yes

   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.

**Your Name: Akshat Gaur**

**Your Andrew ID: agaur**

# Homework 5

## Instructions

## 1 Experiment: Diversity baselines

### 1.1 Experimental results

| | Indri | Indri + PM2 | Indri + xQuAD | BM25 | BM25+ PM2 | BM25+ xQuAD |
|---|---|---|---|---|---|---|
| **P-IA@10** | 0.2621 | 0.3033 | 0.2887 | 0.1833 | 0.3743 | 0.3832 |
| **P-IA@20** | 0.2681 | 0.3257 | 0.3051 | 0.2450 | 0.3333 | 0.3428 |
| **αNDCG@20** | 0.4627 | 0.4971 | 0.4749 | 0.4204 | 0.6077 | 0.6101 |

### 1.2 Parameters

- Indri:
  - mu = 2500
  - lambda = 0.4
- BM25:
  - b = 0.75
  - $k_1$ = 1.2
  - $k_1$=0.0
- diversity:
  - maxInputRankingsLength = 100
  - maxResultRankingLength = 50
  - lambda = 0.5

### 1.3 Discussion

In this experiment, we will be using different metrics so that we can take into consideration different intents while scoring the documents. We are using these metrics instead of original metrics P@n or MAP because these metrics cannot be used to cover the different intents while ranking the documents

Using PM2 and xQuAD for performing diversification will take into consideration documents that cover different intents. xQuAD will prefer documents that cover multiple intents and PM2 will also prefer documents of different intents based on the popularity of that intent. Since both the approaches help in providing diversification of query so the above metric provides more score for the diversified approaches compared to the default algorithm being used.

The P-IA score for default Indri is comparatively high compared to the default BM25 algorithm. This observation might be because of flat score distribution of Indri compared to BM25. We can also say that might be the results produced by default Indri are more diversified and default BM25 score are more redundant.

From αNDCG also we can conclude that using diversification helps in improving the Indri and BM25 retrieval score. For Indri, PM2 works better compared to xQuAD while for BM25 xQuAD outperforms PM2.

We are adding diversification so the computational cost will increase. It depends on the number of query intents, the number of documents that we consider as input while preforming diversification (maxInputRankingsLength) and also on number of documents we finally need (maxResultRankingLength). For calculating diversified rank, we iterate for maxResultRankingLength and for each iteration we consider maxInputRankingsLength documents for calculating PM2 or xQuAD score.

From the experiment, we can also see that although default BM25 has lesser score because of more redundant documents but adding diversification improves BM25 much more than Indri from which we can deduce that BM25 might perform better for this corpus compared to Indri given we use diversification.

## 2 Experiment: The effect of diversification on relevance

### 2.1 Experimental results

| | **Indri** | **Indri + PM2** | **Indri + xQuAD** | **BM25** | **BM25+ PM2** | **BM25+ xQuAD** |
|---|---|---|---|---|---|---|
| **P@10** | 0.3800 | 0.4500 | 0.4300 | 0.3300 | 0.5700 | 0.5800 |
| **P@20** | 0.4400 | 0.4700 | 0.4650 | 0.3750 | 0.4650 | 0.5050 |
| **P@30** | 0.4100 | 0.4300 | 0.4267 | 0.3667 | 0.4333 | 0.4500 |
| **MAP** | 0.2267 | 0.1968 | 0.1806 | 0.1997 | 0.2068 | 0.2191 |

### 2.2 Discussion

In our implementation, the approach that we are using for diversification is high precision. This is because when we are only considering documents retrieved from original query for performing re-ranking based on how it covers different intents. We are not considering the documents that are relevant for the intents but are not retrieved by original query so recall is not changed. Rather since we are reducing the result rank length to half we are reducing the recall value. We can conclude this from the P@n which is getting improved for both default Indri and BM25 when we apply diversification. From the result, we can also see that since the precision increased that means that might be the documents that are more relevant cater diverse intents and so using diversification algorithm ranks them higher which increases P@n. As we are reducing the number of documents while re-ranking the diversified result, the MAP score might give different results which depends on the initial retrieved documents. If the original result did not have the more relevant documents ranked higher and if the diversified result moves these documents higher in rank above the final number of documents required than the MAP score will

increase. If the original result did have relevant documents in the top but using diversification ranks them lower in order to cover different intents then MAP will get reduced. Like it might be that some document is not much relevant but very relevant for a particular intent so it will be pushed up thereby affecting the MAP. This will mostly happen after some top document when we might have covered the most relevant documents and so some irrelevant documents get pushed above some lesser relevant documents. So, we cannot arrive to any conclusion with much certainty using MAP metric in case of diversification algorithm that we use. Also, MAP considers both recall and precision and since we are reducing the final result length we might get lower recall so we cannot conclude much with certainty about the difference in the results of default algorithms and diversified algorithms.

As discussed in above experiment, the computational cost for diversified approach will be higher compared to default algorithms. So, if we want to make lesser people sad by covering more intents we will have higher computational cost or if we want to make more people happy by focusing on one and most commonly asked intent that we will have lesser computational cost.

## 3 Experiment: Effect of λ

### 3.1 Experimental results

| | **λ=0.0** | **λ=0.2** | **λ=0.4** | **λ=0.6** | **λ=0.8** | **λ=1.0** |
|---|---|---|---|---|---|---|
| **Indri + PM2** | | | | | | |
| **P-IA@10** | 0.2840 | 0.2898 | 0.3033 | 0.3108 | 0.3088 | 0.2983 |
| **P-IA@20** | 0.3153 | 0.3169 | 0.3282 | 0.3246 | 0.3183 | 0.2966 |
| **αNDCG@20** | 0.5061 | 0.5027 | 0.4961 | 0.4967 | 0.4997 | 0.5091 |
| **Indri + xQuAD** | | | | | | |
| **P-IA@10** | 0.2622 | 0.2622 | 0.2728 | 0.3020 | 0.2895 | 0.3085 |
| **P-IA@20** | 0.2681 | 0.2893 | 0.2959 | 0.3055 | 0.3164 | 0.3244 |
| **αNDCG@20** | 0.4627 | 0.4877 | 0.4674 | 0.4714 | 0.4543 | 0.4987 |

| | **λ=0.0** | **λ=0.2** | **λ=0.4** | **λ=0.6** | **λ=0.8** | **λ=1.0** |
|---|---|---|---|---|---|---|
| **BM25 + PM2** | | | | | | |
| **P-IA@10** | 0.3592 | 0.3658 | 0.3820 | 0.3693 | 0.3652 | 0.3037 |
| **P-IA@20** | 0.3272 | 0.3302 | 0.3318 | 0.3317 | 0.3388 | 0.2678 |
| **αNDCG@20** | 0.5970 | 0.6063 | 0.5987 | 0.6161 | 0.6206 | 0.6330 |
| **BM25 + xQuAD** | | | | | | |
| **P-IA@10** | 0.1833 | 0.3582 | 0.3867 | 0.3887 | 0.3753 | 0.3678 |
| **P-IA@20** | 0.2450 | 0.3239 | 0.3332 | 0.3480 | 0.3268 | 0.3346 |
| **αNDCG@20** | 0.4205 | 0.6030 | 0.6101 | 0.6117 | 0.6053 | 0.5960 |

### 3.2 Discussion

Lambda parameter decides how much diversification we want to add. It helps in controlling how much importance we want to give to diversify our results. The way xQuAD

works is that it gives preferences to documents that cover most of the intents and higher the value of lambda more is the weightage given to diversification. In PM2 we prefer documents based on the proportionality of their popularity and higher the value of lambda gives more importance to document which covers the intent which has been least represented i.e. high marginal relevance.

In xQuAD, using smaller value of lambda means we give lesser weightage to document that cover multiple intents which will lead to lesser precision-IA as its likely to be relevant. As we increase the value of lambda we perform higher diversification. Higher the diversification more are the chances of covering different intents. So, increasing lambda for xQuAD will improve P-IA (precision intent aware) to an extent for both the algorithms. But as we keep on increasing the lambda parameter we might end up giving more weightage to documents that are generic and cover more intents rather than the documents that cover some particular intent in depth. This might lead to reduced precision. We can also see that in case of Indri the α-NDCG@k score keeps fluctuating. The α-NDCG@k score for BM25 first increases and then falls which is as expected same as the precision-IA as we need to find trade-off between diversity and default algorithm results.

In PM2, using smaller value of lambda means we give lesser weightage to cover least represented intents which will lead to lesser precision-IA. As we increase the parameter lambda we try to give more importance to document representing least represented intent. So, increasing lambda for PM2 will improve precision for some time for both the algorithms. But then after a certain value we will see decrease in precision which is because as we give more importance to intents which are least represented we may start considering some lesser relevant documents and pushing them up the ranking will reduce our precision-IA. We can also see that in case of BM25 the α-NDCG@k score keeps fluctuating. The α-NDCG@k score for Indri first decreases and then increases which is as expected same as the precision-IA as we need to find trade-off between diversity and default algorithm results.

## 4 Experiment: The effect of the re-ranking depth

### 4.1 Parameters

Diversity lambda = 0.6.

This value of parameter lambda can be concluded from the previous experiment we ran. Lambda 0.6 mostly gives best results with both the algorithms and the diversification approaches being used.

### 4.2 Experimental results

| | **25 / 25** | **50 / 25** | **100 / 25** | **100 / 50** | **200 / 100** |
|---|---|---|---|---|---|
| **Indri + PM2** | | | | | |
| **P-IA@10** | 0.3272 | 0.3167 | 0.3108 | 0.3108 | 0.3002 |
| **P-IA@20** | 0.2799 | 0.3032 | 0.3246 | 0.3246 | 0.3238 |
| **αNDCG@20** | 0.5059 | 0.4781 | 0.4967 | 0.4967 | 0.4940 |
| **Indri + xQuAD** | | | | | |

| | | | | | |
|---|---|---|---|---|---|
| **P-IA@10** | 0.2853 | 0.2922 | 0.3020 | 0.3020 | 0.2803 |
| **P-IA@20** | 0.2794 | 0.2976 | 0.3055 | 0.3055 | 0.3043 |
| **αNDCG@20** | 0.4712 | 0.4692 | 0.4714 | 0.4714 | 0.4652 |

| | **25 / 25** | **50 / 25** | **100 / 25** | **100 / 50** | **200 / 100** |
|---|---|---|---|---|---|
| **BM25 + PM2** | | | | | |
| **P-IA@10** | 0.3082 | 0.3685 | 0.3693 | 0.3693 | 0.3735 |
| **P-IA@20** | 0.2668 | 0.3223 | 0.3332 | 0.3332 | 0.3318 |
| **αNDCG@20** | 0.6162 | 0.6457 | 0.6161 | 0.6161 | 0.6045 |
| **BM25 + xQuAD** | | | | | |
| **P-IA@10** | 0.3290 | 0.3877 | 0.3887 | 0.3887 | 0.3837 |
| **P-IA@20** | 0.2668 | 0.3313 | 0.3480 | 0.3480 | 0.3447 |
| **αNDCG@20** | 0.6163 | 0.6769 | 0.6117 | 0.6117 | 0.6527 |

### 4.3 Discussion

**Input Document Length**

When we perform diversification, we try to cover all the relevant documents for each intent. As we increase the number of input documents we might first get relevant documents covering different intents so the precision-IA and αNDCG will increase as we cover more intents. As we keep on increasing the input documents we will later reach a threshold beyond which if we increase the input documents to be re-ranked we will now get more documents covering different intents which might be less relevant or irrelevant documents. Increasing input documents beyond this might reduce the precision value.

The value of input documents to be used depends on the query. In case of ambiguous query, we might have more query intents so we might have more number of relevant documents as we have more intents. So, the threshold value for such queries might be more. In case of queries which are more clear and have lesser intents might have lesser relevant documents so we have lesser threshold value for such queries.

Using higher value of input document length increases our computational cost as we now need to consider these many documents for each iteration over result re-ranking length to find most relevant document as per the diversification algorithm.

**Result re-Ranking Document Length**

The result re-ranking document length does not have much effect on the diversified results produced provided we keep the input document length same. This is because the document to be used for diversification is the same and we are just changing the length of output result. So, in such case we can see the result of precision-IA@n and αNDCG@n will be same for different result ranking length if the length is greater than n for each case.

Since result re-ranking document length does not have much effect on the result using bigger value of this parameter will only increase the computational cost. So, in the domains where we want faster retrieval and high precision and where user care about only top k documents we can choose smaller value of this parameter to improve the retrieval efficiency.