The Information Retrieval Task: Ranking of Studies for Systematic Reviews - Proposed Method

Jawaher A. Alghamdi
j.alghamdi@uqconnect.edu.au

The University of Queensland, Australia
Engineering and Information Technology Faculty

Abstract. In this work, we have proposed a method out of many ranking methods (namely, **PLM**). The ultimate goal is to achieve better results than the baseline methods that we have implemented in the previous part of this project and to achieve this, *PLM* method has been proposed in this part. As summarised results, final findings calculated for our method (**PLM**) along with other functions include TF-IDF, BM25. By applying evaluation measures on the aforementioned method, we have come to the conclusion that **PLM** method outperforms baseline BM25 on training data sets. Comparing the obtained results between **PLM** and baseline BM25, **PLM** has higher mean average precision than BM25 on training datasets while baseline BM25 surpass our method on testing data-sets. In this report, we firstly, discuss our proposed method, then we present a review of such method; section 2 is devoted for the implementation of the proposed method and their findings that we obtained are discussed in section three; some details about the empirical evaluation settings and analysis which have been carried out on our experiment were also employed and stated in section 4 ;section 5 is the conclusion of this report.

1 Parsimonious Language Models(PLM)

1.1 Review

So far we have, in part 1 and 2, seen how to score the documents and fuse these scores using some types of fusion algorithms and we have concluded that BM25 is the best ranking method compared to TF-IDF; CombMNZ and CombSUM were effective and comparable to each other as opposed to Borda method.In addition, we have investigated two types of query reduction methods which are so-called IDF-r and KLI, and we come to the conclusion that KLI outperform all other methods. In this chaper we have proposed another method which is so-called **PLM**. In[1], they have investigated this method and applied it at three stages. A less storage and CPU time consumed-models are the results concluded by[1]. They also stated that this method might perform as or better than the standard models. The work of [4] investigated the generation of queries, one of

the most studied subjects in Natural Language Processing (NLP). The work compared different automated methods for common keyword extraction, such as the Kullback-Leibler divergence for informativeness (KLI), explored in the work of [5], the **parsimonious language models (PLM)** , proposed in the work of [2], as well as the proportional inverse document frequency (IDF-r), proposed in the work of [3]. The work compared those automated methods with collected queries obtained by legal experts, through the Boolean method of comparison and best-match method. PLM is known to be a parametric model due to the parameter λ it is used. In [4], they explored the smoothing parameter λ between 0.1 and 1 and they applied it utilizing retantion rates. These works presented the accuracy and robustness of the **PLM** method in prediction text-based queries from different applications. For that reason, the above mentioned method will be employed in our present work.

1.2 Contribution

Being inspired by the work of [4], the major contribution of this proposed method is towards achieving better results when comparing to the baseline methods we have implemented in the previous part of the project. We chose to use PLM method for two practical reasons. Firstly, as mentioned previously, we are motivated by the findings of [4] and therefore we would like to adapt this method on our dataset in order to obtain similar results. Moreover, deciding which terms should be added to our array of terms based on the importance level was also an advantage for this method to be used.

2 Implementation

In this experiment, we will apply **PLM** (see the formula below) method using the same retantion rates we applied in previous experiments which are 0.3, 0.5, and 0.85. We first of all computed the E-step, which is the term frequency

$$E-step: \quad e_t = tf(t,D)\frac{\lambda P(t|D)}{(1-\lambda)P(t|C)+\lambda P(t|D)}$$

$$M-step: \quad P(t|D) = \frac{e_t}{\sum_{t' \in D} e_{t'}}$$

Fig. 1: PLM formula

multiplying by the division of (λ multiplying by the probability of the term in the document (the number of times this term appears in the document divided by the document length)) by the ((1-λ) multiplied by the probability of the term in the collection plus λ multiplied by the probability of the term in the document) for each term, then we could alternatively sum up the scores based on a certain threshold that appends the terms that are most important to our scoredTerm list. In this experiment, we have utilized retantion rates instead of the threshold. We assumed λ equals to 0.5. Indeed, we ran our experiments multiple times to get statistically meaningful results utilizing the idea of the threshold but we eventually inferred that using retantion rates provide better results .

3 Results

3.1 Baseline methods findings

To compare the results with baseline methods easily, we decided to state the results of previous baseline methods in this section.

	MAP	nDCG	Rprec	P_5
BM25	0.1629	0.5393	0.1843	0.2667
TF_IDF	0.1625	0.5389	0.1843	0.2667

Table 1: Evaluation measures for 2017 Testing Title Queries

	MAP	nDCG	Rprec	P_5
BM25	0.1462	0.5415	0.1704	0.2524
TF_IDF	0.1458	0.5412	0.1692	0.2619

Table 2: Evaluation measures for 2018 Training Title Queries

	MAP	nDCG	Rprec	P_5
BM25	0.1700	0.5933	0.1959	0.2200
TF_IDF	0.1705	0.5936	0.1976	0.2200

Table 3: Evaluation measures for 2018 Testing Title Queries

	MAP	nDCG	Rprec	P_5
BM25	0.1138	0.4817	0.1513	0.2200
TF_IDF	0.1134	0.4813	0.1514	0.2400

Table 4: Evaluation measures for 2017 Training Title Queries

3.2 Proposed method results

	MAP			Rprec			nDCG		
	0.3	0.5	0.85	0.3	0.5	0.85	0.3	0.5	0.85
BM25	0.0999	0.1071	0.1476	0.1173	0.1208	0.1754	0.4930	0.5081	0.5432
TF_IDF	0.0935	0.0986	0.1406	0.1103	0.1092	0.1600	0.4848	0.4960	0.5350

Table 5: 2018 Training Title Queries using PLM method.

	MAP			Rprec			nDCG		
	0.3	0.5	0.85	0.3	0.5	0.85	0.3	0.5	0.85
BM25	0.1068	0.1160	0.1585	0.1173	0.1376	0.1891	0.5385	0.5490	0.5879
TF_IDF	0.1046	0.1135	0.1616	0.1082	0.1285	0.1799	0.5329	0.5411	0.5866

Table 6: 2018 Testing Title Queries using PLM method.

The observations, see the above tables, indicate that, surprisingly, our method does not outperform the baseline BM25 which we implemented in the previous part on testing dataset. To support such claim, by looking at table 1 and 7 for instance, Baseline BM25 achieved 0.1629 compared to our method's result which

	MAP			Rprec			nDCG		
	0.3	0.5	0.85	0.3	0.5	0.85	0.3	0.5	0.85
BM25	0.1155	0.1216	0.1605	0.1312	0.1337	0.1733	0.4903	0.5032	0.5373
TF_IDF	0.1005	0.1040	0.1552	0.1233	0.1145	0.1688	0.4712	0.4813	0.5307

Table 7: 2017 Testing Title Queries using PLM method.

	MAP			Rprec			nDCG		
	0.3	0.5	0.85	0.3	0.5	0.85	0.3	0.5	0.85
BM25	0.0826	0.0882	0.1174	0.1120	0.1188	0.1565	0.4506	0.4641	0.4864
TF_IDF	0.0791	0.0848	0.1133	0.1083	0.1120	0.1517	0.4469	0.4574	0.4804

Table 8: 2017 Training Title Queries using PLM method.

is 0.1605-"taking into account the best rate we obtained which is 0.85". Fortunately, our method outperformed baseline BM25 in 2017-2018 training datasets which drives us to the fact that **PLM** method in this experiment performs very well on training data as opposed to the testing ones. As, for instance, it can be seen from table 4 and table 8, our method have achieved 0.1174 compared to only 0.1138 obtained by the baseline BM25.

As it can be seen by the following gain/loss figures, our method gain/loss are slightly different from the baseline BM25. We chose the best value "0.85" to compare with.

Looking at Fig. 2 and Fig. 5, it can be inferred that our method has slightly more gains than the baseline BM25 on training data.

Same consideration can be applied on box plots where PLM somehow performs better than baseline BM25.

It should be noted that evaluation measures along with statistical analysis are applied on the resulted files, we evaluated the retrieval methods with respect to the mean average precision (MAP) using trec-eval; we sat the cut-off value to (-M ,i.e. the maximum number of documents per topic) the number of documents to be re-ranked for each of the queries.

3.3 Fusion

	MAP		*Rprec*	
	Baseline-BM25	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**
Borda	0.1629	0.1591	0.1843	0.1802
CombSUM	0.1628	0.1629	0.1843	0.1837
CombMNZ	0.1628	0.1629	0.1843	0.1837

Table 9: Fusion Results for 2017 Testing Title Queries data-set.

It can be clearly observed from the tables that our method achieves better results on training data compared to the baseline methods using these three types of fusion algorithms. In contrast, baseline methods performed very well on testing as opposed to PLM method.

	MAP		*Rprec*	
	Baseline-BM25	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**
Borda	0.1702	0.1594	0.1970	0.1840
CombSUM	0.1703	0.1596	0.1973	0.1824
CombMNZ	0.1703	0.1596	0.1973	0.1824

Table 10: Fusion Results for 2018 Testing Title Queries data-set.

	MAP		*Rprec*	
	Baseline-BM25	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**
Borda	0.1461	0.1441	0.1696	0.1696
CombSUM	0.1461	0.1469	0.1699	0.1713
CombMNZ	0.1461	0.1469	0.1699	0.1713

Table 11: Fusion Results for 2018 Training Title Queries data-set.

	MAP		*Rprec*	
	Baseline-BM25	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**
Borda	0.1137	0.1164	0.1511	0.1565
CombSUM	0.1137	0.1173	0.1511	0.1562
CombMNZ	0.1137	0.1173	0.1511	0.1562

Table 12: Fusion Results for 2017 Training Title Queries data-set.

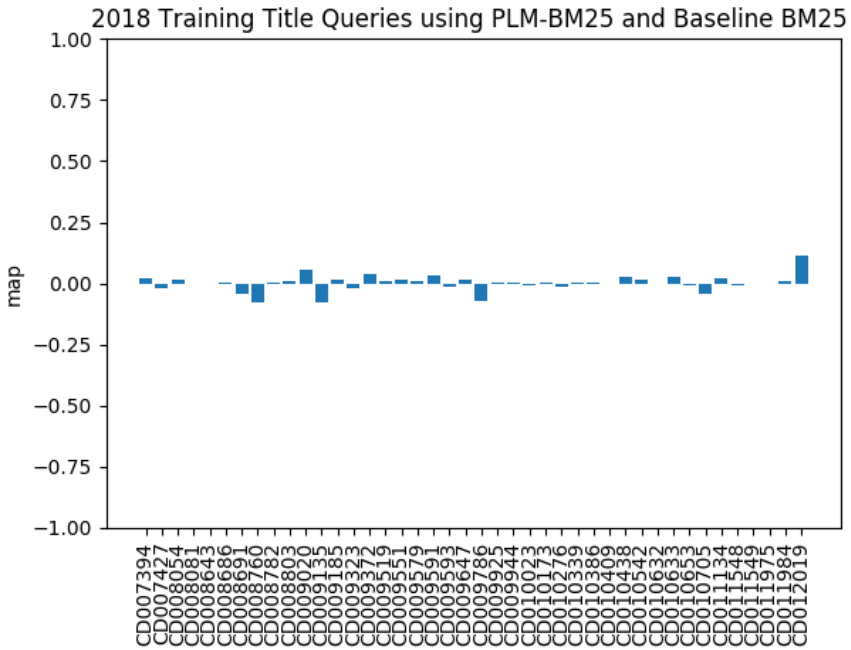

Fig. 2: Gain/Loss

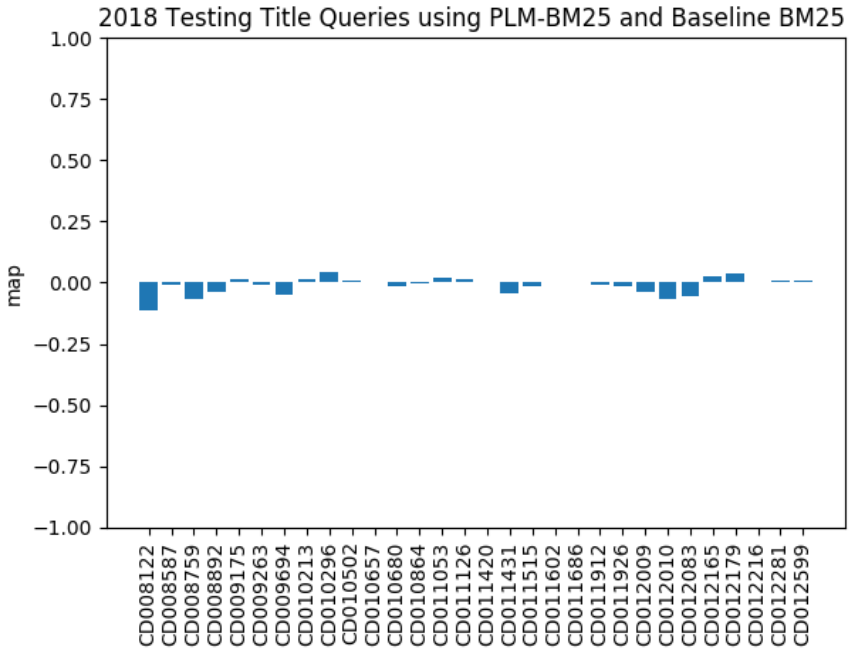

Fig. 3: Gain/Loss

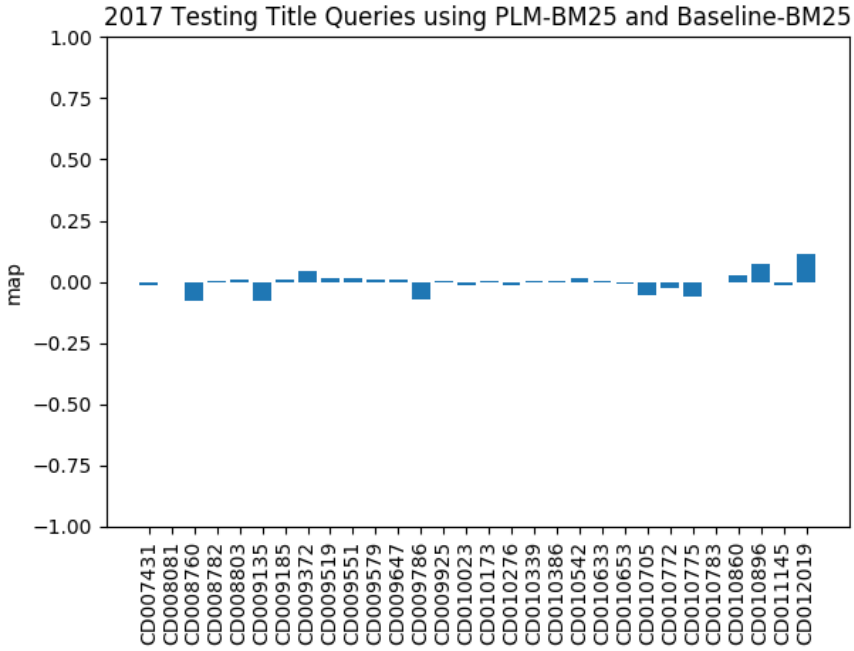

Fig. 4: Gain/Loss

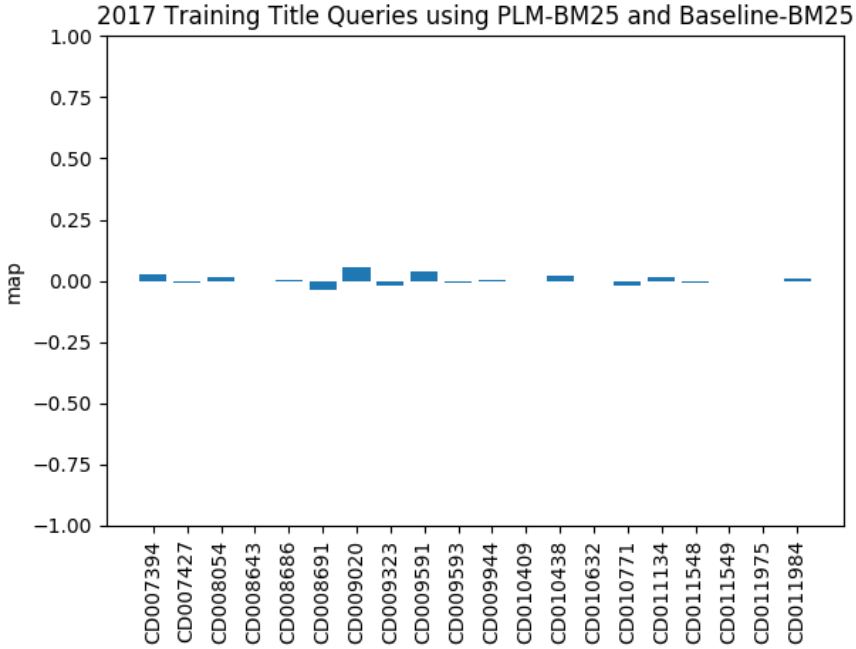

Fig. 5: Gain/Loss

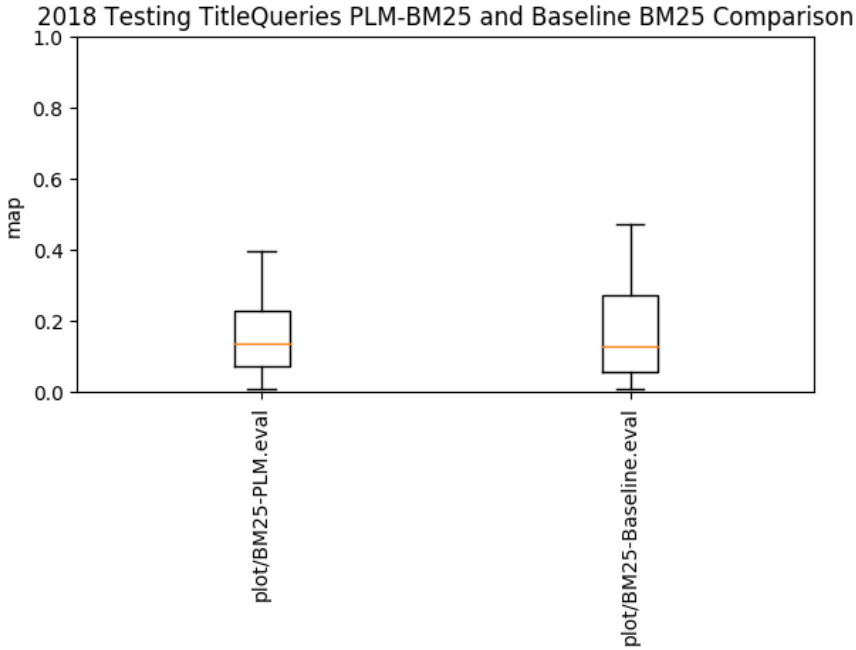

Fig. 6: Box-plot

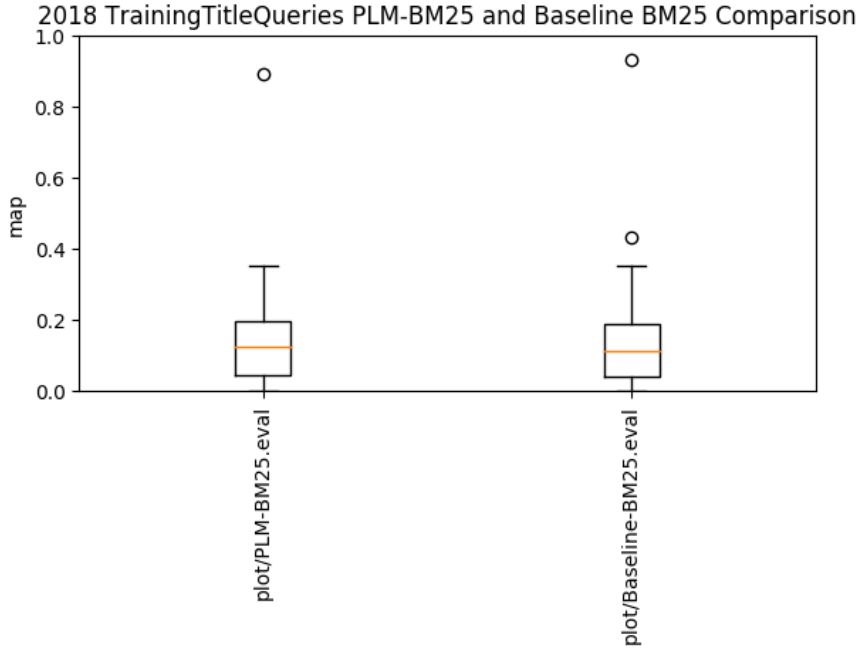

Fig. 7: Box-plot

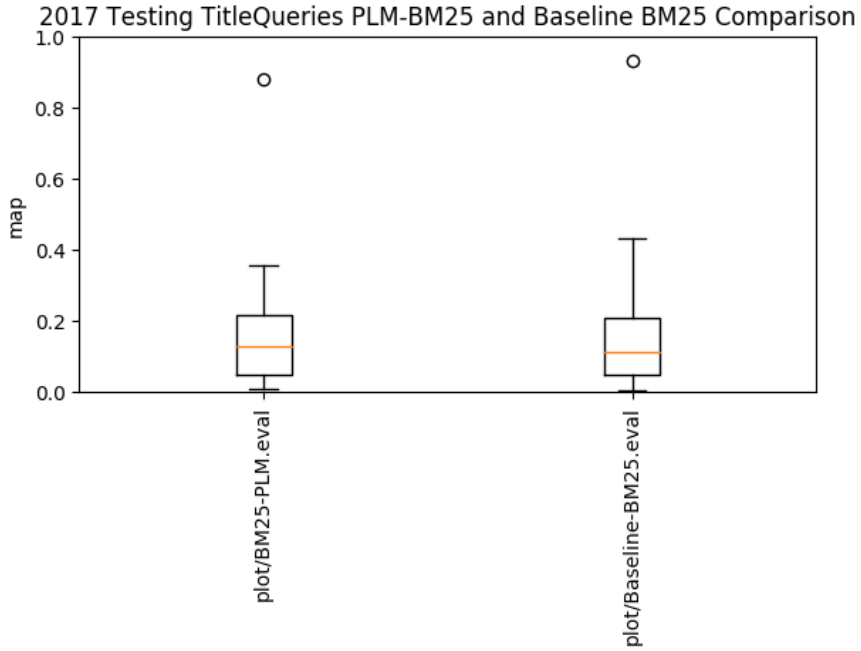

Fig. 8: Box-plot

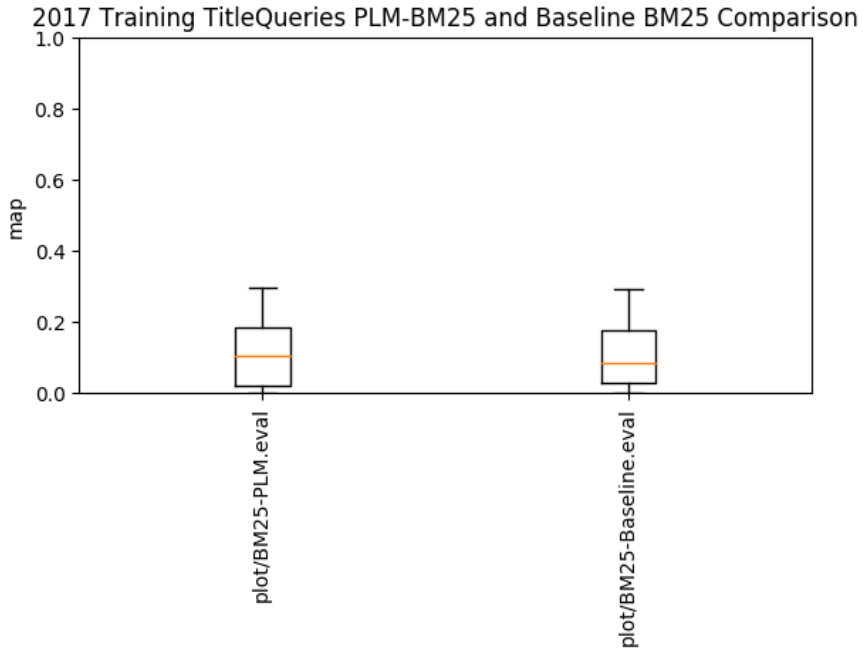

Fig. 9: Box-plot

4 Statistical Significance Tests

Statistical significance test is commonly used to evaluate the performance of two systems. Recalling that `(from the lecture)` the significance test is represented as a number between 0 and 1, where smaller p (when p less than 0.05 or 0.01) indicates stronger evidence to reject the null hypothesis,we can infer that there is a strong statistical significance.

4.1 Baseline methods findings-part1

According to the observations (see Table 13), it turns out that there is no a difference between the methods since p-value is larger than 0.05, 0.01 in both data-sets. Generally speaking, there was a dramatic increase in p value which means the chances of no significance difference occurs. It should be noted that we use TF-IDF along with BM25.

P-value			
	MAP	nDCG	P_5
2017 Training Title Queries	0.17392392246807886	0.4305460288403863	0.16254999902972722
2017 Testing Title Queries	0.2390969474748348	0.3001400326453135	NaN
2018 Training Title Queries	0.10731193007559892	0.33755901667724	0.1597858784099142
2018 Testing Title Queries	0.12139847744806853	0.2962209323980247	NaN

Table 13: P-value for both data-sets Title-Queries (BM25,TF_IDF)

4.2 The proposed method findings

p-value	MAP			Rprec			nDCG		
	0.3	**0.5**	**0.85**	**0.3**	**0.5**	**0.85**	**0.3**	**0.5**	**0.85**
2017 Training	0.0258	0.1167	0.1500	0.4028	0.0576	0.4758	0.1235	0.1270	0.1106
2017 Testing	0.0296	0.0102	0.3073	0.2721	0.1343	0.6747	0.0302	0.0058	0.2953
2018 Training	0.0011	0.0068	0.0449	0.0254	0.0061	0.0034	0.0053	0.0046	0.0732
2018 Testing	0.5925	0.4980	0.5597	0.1819	0.2433	0.2218	0.2875	0.0861	0.8045

Table 14: p-value for both data-sets using PLM method.

Table 14 shows how these methods differ from each other, the observations showed that there is somewhat a difference between the methods, so by looking at "2018 Training data-set",for instance,we can infer that there is a difference between the methods since the results are less than 0.01 whereas "2018 testing data sets p-value" shows that there is no a difference since the obtained results are larger than 0.05. We noticed that the higher the retantion rate (or threshold) is, the lower the difference is. In other words, the observations indicate that as the threshold increases, the difference is decreased. Thus, the methods are different w.r.t some data-sets whereas the difference doesn't exist in some data such as 2018 testing data. It should be noted that we use "map" measure to evaluate the results.

4.3 P-value Comparisons

p-value	*MAP*	*Rprec*	*nDCG*
	Baseline-BM25 and PLM-BM25	**Baseline-BM25 and PLM-BM25**	**Baseline-BM25 and PLM-BM25**
2017 Training	0.4437	0.5939	0.4108
2017 Testing	0.7490	0.3289	0.8296
2018 Training	0.7797	0.4561	0.7816
2018 Testing	0.0822	0.4067	0.3105

Table 15: p-value for both methods.

The table above presents the p-value for both methods. Comparing these p-values we obtained using baseline-BM25 and our proposed method, we have come to the conclusion that there is no significant difference, that is, there is a significant increase in p-values which indicates that there is no such difference (see Table 15). Generally speaking, it can be clearly seen from the table that the two methods are not significantly different.

p-value	MAP		Rprec		nDCG	
	Baseline-BM25	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**	**Baseline-BM25**	**PLM-BM25**
2017 Training	0.1739	0.1500	0.8363	0.4758	0.4305	0.1106
2017 Testing	0.2390	0.3073	NaN	0.6747	0.3001	0.2953
2018 Training	0.1073	0.0449	0.0539	0.0034	0.3375	0.0732
2018 Testing	0.1213	0.5597	0.0718	0.2218	0.2962	0.8045

Table 16: p-value for both methods along with TF-IDF.

Table 16 is computed using TF-IDF function along with BM25. As it can be seen that there is no significant difference as the observations obtained are greater than 0.05.

5 Conclusion

In this work, we have proposed **PLM** method. Throughout this experiment, we have analysed, evaluated, and compared our method against baseline BM25 and we have come to the conclusion that **PLM** method outperforms the baseline BM25 method, *which unsurprisingly*, on the training data-sets while Baseline BM25 surpass our method on the testing datasets. Whats more, regarding p-value, we concluded that there is no significant difference between both methods.

References

1. Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 178–185. ACM (2004)
2. Hiemstra, D., Robertson, S., Zaragoza, H.: Parsimonious language models for information retrieval. Proceedings of the 27th annual international conference on Research and development in information retrieval - SIGIR 04 (2004). https://doi.org/10.1145/1008992.1009025
3. Koopman, B., Cripwell, L., Zuccon, G.: Generating clinical queries from patient narratives. Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval - SIGIR 17 (2017). https://doi.org/10.1145/3077136.3080661
4. Locke, D., Zuccon, G., Scells, H.: Automatic query generation from legal texts for case law retrieval. Information Retrieval Technology Lecture Notes in Computer Science p. 181–193 (2017). https://doi.org/10.1007/978-3-319-70145-5$_1$4
5. Tomokiyo, T., Hurst, M.: A language model approach to keyphrase extraction. Proceedings of the ACL 2003 workshop on Multiword expressions analysis, acquisition and treatment - (2003). https://doi.org/10.3115/1119282.1119287