Milestone 3-
COL-764 Project
Precedent Retreival of Legal Cases

Pratik, Prawar, Rohan Debbarma,
Indian Institute of Techonology, Delhi
cs5180415@iitd.ac.in, cs5180417@iitd.ac.in

Abstract

This report evaluates the precedent retreival of legal cases with reference to FIRE-2017 dataset which has citation markers. It provides results related to experimenting with BM-25 and TF-IDF based approaches for retreival task. In the context of retrieval, sub-query approach which models each query case as a set of sub-queries was utilised to improve the accuracy (MAP) of the model. Preliminary results indicate that well-tuned BM-25 model using Citation Context as queries outperform TF-IDF based approaches.

1 Introduction

In this report, we have used variants of classical IR Models such as BM-25 and VSM (using TF-IDF) to retrieve precedent cases.

The approaches are briefly discussed in the sections below.

2 Approaches Used

In this section, a brief description of the approaches used (BM-25 and TF-IDF) in this part along with the results is provided.

2.1 BM-25 based models

2.1.1 Citation Context + BM25

As we had already mentioned in the previous report, the FIRE-2017 dataset had citation markers which helped us to identify the regions of query document which are most relevant to a paticular citation. In other words, we can view each query document as a collection of sub-queries, each referring to one citation. Each of the sub-queries could then be treated as an individual query in BM-25 and be used to rank the documents.

To combine the rankings of each of the sub-queries in a query, we take the maximum BM-25 score obtained for a document for all the sub-queries. The documents are then sorted in descending order by their BM-25 score.

The procedure is as follows:

1. The sections of text corresponding to each citation marker namely, $[?CITATION?]$ was identified and the text on either side of it upto 100 spaces on both sides each was captured

as the sub-query for each query. We experimented with different text fragments and found moderately long queries gave better results.

As an example, taking 40 spaces around the marker on both sides gave maximum MAP of 0.40 whereas taking 100 spaces corresponded to MAP of 0.47 which is a significant improvement. This points to the fact that for this dataset, longer queries which provide larger context performed better.

2. The query was then tokenized with $word\_tokenizer$ of nltk and experimentation was done with stemming (with PorterStemmer) and stopword removal (Nltk.stopwords).
3. Each of the prior cases which need to be ranked were then tokenized using nltk $word\_tokenizer$.
4. The BM-25 model was then trained on the set of prior cases.
5. Each of the sub-queries of a query case were then fed as queries to obtain a ranking and BM-25 scores for each of the prior cases.
6. For each prior case, we took the maximum of the BM-25 scores it obtained from all the sub-queries.
7. So, using all the subqueries, we now have a final ranking of all the prior cases in terms of their BM-25 scores.
8. Using these scores, a ranking was generated and the metrics such as MAP, MRR and P@10 evaluated for them.

Note- For BM-25, the value of k1 and b used after optimisation was 1.75 and 0.95

2.1.2 Citation Context + BM25 + IDF Screening

In this variant, instead of using the whole query text as the query in a subquery, top 50 $\%$ of the terms in the text ranked by their IDF score are used for rankings. This is done to check whether some of the terms in that query which are specific to the document can be a better representative of the query text.

Table 1: Results of BM-25 and variants

Method Used	Stopword	Stemming	MAP	MRR	$P@10$
Citation Context + BM-25	Yes	No	0.477	0.821	0.281
Citation Context + BM-25	No	No	0.477	0.819	0.280
Citation Context + BM-25	Yes	Yes	0.411	0.768	0.249
Citation Context + BM-25	No	Yes	0.419	0.777	0.249
Citation Context + BM-25 + IDF Screening	Yes	No	0.467	0.787	0.276
Citation Context + BM-25 + IDF Screening	No	No	0.466,	0.797	0.271
Citation Context + BM-25 + IDF Screening	Yes	Yes	0.413	0.345	0.253
Citation Context + BM-25 + IDF Screening	No	Yes	0.402	0.720	0.252

2.1.3 Discussion on results

From the above table, it is clear that BM-25 model applied on sub-queries of each query and then using the max BM-25 score of each document gives good result and it has outperformed best method mentioned in [1] by a margin of 0.08 which is a significant improvement. For faster processing, IDF screening variant of BM-25 described above also performs well and gives a MAP of 0.467.

Applying stemming on the query text has yielded lower MAP because in many legal scenarios, some nouns such as named entities and verbs which are generally used in judgements have been degraded to their root forms due to stemming. This has resulted in lower MAP scores. The best MAP achieved with stemming is 0.419 which is around 0.05 less than optimal MAP found.

Stop word Removal has negligible impact on MAP. This is may be because the IDF scores in BM-25 have already made their weights close to zero in an optimally tuned case.

Another key observation is longer queries have provided better MAP in this model. This is may be because longer queries capture more information and relevant words related to the topic of each citation.

Therefore, using max ranking using BM-25 scores of sub-queries has provided significant improvement over the best method mentioned in [1].

2.2 TF-IDF + Pos-tagging based methods

In this particular method, we first created TF-IDF vectorizer and used this model to obtain a ranking for the whole dataset without taking the citation marker into consideration. We used three metrics as the evaluation metrics, Mean Average Precision(MAP), Mean Reciprocal Rank(MRR) and P@10. The model was then modified for several different strategies as discussed below:-

2.2.1 Citation Context

In the model developed above, we had considered the whole document as our dataset. In this part, we consider only a fixed number of tokens(in our case it is 400) around the citation markers in the document which helps us create multiple subqueries. Now, based on the cosine similarity of each subquery, we chose the subquery which gives us the maximum similarity score for a particular query. Using this subquery, we calculate the ranking, which is then further used to calculate the above given evaluation metrics.

2.2.2 Stopwords

Then we modify our model to include stopword removal as well. Hence, we test our model for both stopword containing and stopword removed content. The stopwords consist of the general 'english' language stopwords of nltk library.

2.3 Pos-Tagging

After testing for the above two strategies, another new strategy was tested which included the use of pos-tagging on top of the above models. Hence, in this model, we separate the terms into their particular parts of speech and then use only specific parts of speech for ranking the documents. The different combinations used in this method are as follows:-

1. No Pos-Tagging
2. Only using Nouns present in the text
3. Use nouns and verbs present in the text

In the last part, the nouns and verbs are seperately marked(so that similar words which can be both nouns and verbs can be seperated). This helps in improving the overall score.

2.3.1 Results

The scores using different metrics obtained using the different combinations of the above model are as follows:-

It can be observed that the best performance comes when using the vectorizer with citation context and stopword removal with no pos-tagging. Another interesting thing to note here is that when using citation context, the overall scores are higher compared to using full document as the context.

Also, stopword removal increases the overall accuracy of the model in all cases as can be seen from the above table. However, on introducing pos-tagging in the model, the scores decrease. Using no pos-tagging gives the maximum scores, followed by using only nouns as context, and lastly, using both nouns and verbs as context.

Table 2: Results of TF-IDF and Pos-Tagging

Method Used	Stopword	Pos-Tagging	MAP	MRR	$P@10$
TF-idf vectorizer + citation context	Yes	No	0.365	0.698	0.229
TF-idf vectorizer + citation context	No	No	0.256	0.618	0.164
TF-idf vectorizer + citation context	Yes	Nouns only	0.282	0.587	0.185
TF-idf vectorizer + citation context	No	Nouns only	0.244	0.486	0.173
TF-idf vectorizer + citation context	Yes	Nouns and verbs	0.224	0.450	0.153
TF-idf vectorizer + citation context	No	Nouns and verbs	0.204	0.401	0.148
TF-idf vectorizer + without citation context	Yes	No	0.284	0.556	0.190
TF-idf vectorizer + without citation context	No	No	0.173	0.424	0.124
TF-idf vectorizer + without citation context	Yes	Nouns only	0.247	0.466	0.171
TF-idf vectorizer + without citation context	No	Nouns only	0.204	0.386	0.144
TF-idf vectorizer + without citation context	Yes	Nouns and verbs	0.203	0.368	0.146
TF-idf vectorizer + without citation context	No	Nouns and verbs	0.175	0.331	0.119

3 Conclusion

From the above work, we can see that BM-25 with citation context has significantly outperformed TF-IDF based VSM Models which was expected.

Table 3: Overall Results of best models

Method Used	Stopword	Stemming	MAP	MRR	$P@10$
Citation Context + BM-25	Yes	No	0.477	0.821	0.281
TF-idf vectorizer + citation context	Yes	No	0.365	0.698	0.229

In this report, we looked at the technique of modeling each query case as a set of sub-queries which represent the query. This type of modeling has significant impact on the MAP since using the whole query document as a single query has not provided us good accuracy due to the fact that each query case is quite large and two different citations in a case are maybe on entirely different type of prior cases.

This report has been an attempt to comprehensively perform query retrieval using BM-25 and TF-IDF based methods and the method reported here has been able to ouperform the best method used in [1].

4 Future Work

In the subsequent part, we would experiment with other methods of retreival such as Language and Topic Modeling, Word2Vec. In addition, we will work on retrieval on FIRE2019 dataset which has no citation markers and presents a different kind of challenge during ad-hoc retrieval.

5 References

1. Mandal, A., Kripabandhu Ghosh, Arnab Bhattacharya, Arindam Pal and Saptarshi Ghosh. "Overview of the FIRE 2017 IRLeD Track: Information Retrieval from Legal Documents." FIRE (2017).