

A Study on Lemma vs Stem for Legal Information Retrieval Using R Tidyverse.

IMS UniPD @ AILA 2020 Task 1

Giorgio Maria Di Nunzio^{a,b}

^a*Department of Information Engineering, University of Padova, Italy*

^b*Department of Mathematics, University of Padova, Italy*

Abstract

In this paper, we describe the results of the participation of the Information Management Systems (IMS) group at AILA 2020 Task 1, precedents and statutes retrieval. In particular, we participated in both subtasks: precedents retrieval (task a) and statutes retrieval (task b). The goal of our work was to compare and evaluate the efficacy of a simple reproducible approach based on the use of either lemmas or stems with a tf-idf vector space model and a plain BM25 model.

The results vary significantly from one subtask/evaluation measure to another. For the subtask of statutes retrieval, our approach performed well, being second only to a participant that used BERT to represent documents.

Keywords

Legal IR, BM25, TF-IDF, Text Pipelines, R Tidyverse

1. Introduction

The FIRE Artificial Intelligence for Legal Assistance (AILA) is an evaluation challenge of series of shared tasks aimed at developing datasets and methods for solving variety of legal problems by means of search engine approaches [1]. This year, AILA proposed two different legal documents tasks: precedent and statute retrieval task, and semantic segmentation task. In this paper, we report the results of our participation to the first task, precedent and Statute Retrieval [2]. This task investigates the problem of identifying the relevant statutes and prior cases given a description of a situation (i.e. the query, in traditional IR terms).

The contribution of our experiments to this task can be summarized as follows:

- the implementation of a reproducible pipeline for text analysis;
- an evaluation of basic rankers based on lexical levels with a tf-idf approach and a BM25 approach.

This work follows a series of reproducible experiments that has origin in the CLEF eHealth Task [3], and the source code for replicating all our experiments will be online.¹

Forum for Information Retrieval Evaluation 2020, December 16-20, 2020, Hyderabad, India

✉ giorgiomaria.dinunzio@unipd.it (G. M. Di Nunzio)

🌐 <http://github.com/gmdn> (G. M. Di Nunzio)

🆔 0000-0001-9709-6392 (G. M. Di Nunzio)



© 2020 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

📄 CEUR Workshop Proceedings (CEUR-WS.org)

¹<http://github.com/gmdn>

The remainder of the paper will introduce the methodology and a brief summary of the experimental settings that we used in order to create the official runs that we submitted for this task.

2. Method

In this section, we summarize the pipeline for text pre-processing which has been developed in the last years [4]. In general, our method follows the principles described by [5] where the idea is to mine textual information from large text collections in an efficient and effective by means of organized workflows named pipelines. Pipelines are an effective way to manage the sequential process of text analysis by splitting the source code into steps, where the output of one step is the input for the subsequent step. The R programming language has an interesting set of packages that follow this idea, named tidyverse,² that we will use in our experiments.

Apart from being a tidy way of organizing software, an important advantage in working with pipelines is that this practice promotes shareability and reproducibility in research workflows which is one of the main pillars in the European Open Science Cloud (EOSC).³

2.1. Pipeline for Data Cleaning

In order to produce the clean dataset, we followed the same pipeline for data ingestion and preparation for all the experiments:

- split text into words;⁴
- remove stopwords;
- remove words with less than two characters;
- lemmatize/stem words;⁵
- compute tf-idf for each word;
- compute relevance score (BM25) for each word.

3. Experiments

In this section, we briefly describe the setting of official runs that we submitted for this task and the preliminary results sent by the organizers before the workshop.

3.1. Dataset

The datasets of the two subtasks consisted in:

1.
 - Corpus: 3,257 casedocs
 - Queries: 50 description of situations
2.
 - Corpus: 197 statutes
 - Queries: 50 description of situations (same as subtask 1)

²<https://www.tidyverse.org>

³<https://www.eosc-portal.eu>

⁴<https://www.tidytextmining.com>

⁵<https://cran.r-project.org/web/packages/corpus/vignettes/stemmer.html>

3.2. Run Settings

For subtask 1, we split each casedocs into the following parts (we add an example for casedoc number 1):

- who: Masud Khan v State Of Uttar Pradesh;
- where: Supreme Court of India;
- when: 26 September 1973;
- what: Writ Petition No. 117 of 1973;
- delivered: The Judgment was delivered by : A. Alagiriswami, J.;
- text: 1. Petitioner Masud Khan prays for his release on the ground that he, an Indian citizen ...

In the runs of subtask1, we used only the last field (that we named ‘text’) for the retrieval of precedents.

The goal of our experiments is to compare the effectiveness of the different lexical choices (stem or lemma) with a baseline (BM25).

We submitted three runs for each subtask, and we followed the same procedure for each subtask:

- bm25_lemma: this run uses a BM25 retrieval model with lemmas;
- tfidf_lemma: this run uses a tfidf document representation and a cosine similarity score on lemmas;
- tfidf_stem: this run uses a tfidf document representation and a cosine similarity score on stems;

3.3. Results

The organizers of this task provided the results (averaged across topics) achieved by many baselines compared to the runs of each participant. In Table 1 and Table 2, we show a summary of these results.

A preliminary analysis of the results shows that, in terms of standard evaluation measures such as MAP, BPREF, reciprocal rank, and P@10 the use of the BM25 with lemmas performed worse compared to the other two runs.

For subtask 1 (Table 1), the performance was poor compared to the median values for all the measures. This result, compared with the performance of the other subtask, requires a failure analysis to understand how the choice of just on field affected the performance, as well as how additional information in the statistics of the word may be useful or not (see for example team HLJIT2019-AILA of AILA 2019 [6]).

For subtask 2, the use of a vector space model with tf-idf on stems was the second best run overall. It is interesting to see that, despite the type of query (which is a complex description of the situation), and the result in the previous subtask, this approach performed very well compared to the other systems.

Table 1

Summary of the results for subtask 1-a, precedents retrieval. Runs are ordered by MAP.

participant	Run_ID	MAP	BPREF	recip_rank	P @ 10
UB	UB-3	0.16	0.11	0.24	0.08
double_liu_2020	double_liu_2020_3	0.14	0.10	0.19	0.07
fs_hu	fs_hu_task1a	0.14	0.09	0.20	0.10
double_liu_2020	double_liu_2020_1	0.13	0.07	0.20	0.07
TUW_informatics	basic	0.13	0.07	0.19	0.07
f s_hit_1	fs_hit_1_task1a_01	0.13	0.09	0.19	0.07
LAWNICS	LAWNICS_2	0.13	0.09	0.16	0.10
TUW_informatics	word_count	0.13	0.07	0.19	0.06
SSNCSE_NLP	task_1a_1	0.13	0.09	0.20	0.08
f s_hit_2	fs_hit_2_task1a_01	0.12	0.07	0.19	0.07
double_liu_2020	double_liu_2020_2	0.12	0.06	0.20	0.08
UB	UB-1	0.12	0.07	0.20	0.09
f s_hit_1	fs_hit_1_task1a_02	0.12	0.07	0.21	0.09
UB	UB-2	0.12	0.08	0.20	0.07
TUW_informatics	false_friends	0.11	0.07	0.19	0.05
LAWNICS	LAWNICS_1	0.11	0.08	0.16	0.08
Uottawa_NLP	run3_TFIDF	0.08	0.04	0.12	0.05
f s_hit_1	fs_hit_1_task1a_03	0.07	0.03	0.11	0.07
SSNCSE_NLP	task_1a_2	0.07	0.04	0.10	0.05
IMS_UNIPD	tfidf_lemma	0.06	0.03	0.11	0.02
IMS_UNIPD	tfidf_stem	0.06	0.03	0.11	0.03
IMS_UNIPD	bm25_lemma	0.04	0.01	0.15	0.03
f s_hit_2	fs_hit_2_task1a_02	0.01	0.00	0.04	0.02
Uottawa_NLP	run1_Glove	0.01	0.00	0.02	0.00
f s_hit_2	fs_hit_2_task1a_03	0.01	0.00	0.04	0.02
Uottawa_NLP	run2_Doc2Vec	0.00	0.00	0.01	0.00

4. Final remarks and Future Work

The aim of our participation to the FIRE AILA 2020 Task 1 was to test the effectiveness of a reproducible baseline without any learning strategy. The initial results show a completely different scenario according which subtask is considered. The approach seems very promising, but a failure analysis and a topic-by-topic comparison is needed to understand when and how the different combination in the retrieval pipeline are significantly better/worse than other models.

References

- [1] P. Bhattacharya, P. Mehta, K. Ghosh, S. Ghosh, A. Pal, A. Bhattacharya, P. Majumder, Overview of the FIRE 2020 AILA track: Artificial Intelligence for Legal Assistance, in: Proceedings of FIRE 2020 - Forum for Information Retrieval Evaluation, 2020.
- [2] L. Goeuriot, H. Suominen, L. Kelly, Z. Liu, G. Pasi, G. S. Gonzales, M. Viviani, C. Xu, Overview of the CLEF eHealth 2020 task 2: Consumer health search with ad hoc and spoken queries, in: Working Notes of Conference and Labs of the Evaluation (CLEF) Forum, CEUR Workshop Proceedings, 2020.

Table 2

Summary of the results for subtask 1-b, statutes retrieval. Runs are ordered by MAP.

participant	Run_ID	MAP	BPREF	recip_rank	P @ 10
scnu	scnu_1	0.39	0.31	0.56	0.18
SSNCSE_NLP	task_1b_2	0.34	0.14	0.34	0.07
IMS_UNIPD	tfidf_stem	0.34	0.28	0.53	0.17
IMS_UNIPD	tfidf_lemma	0.32	0.26	0.53	0.17
UB	UB-2	0.31	0.26	0.58	0.15
UB	UB-1	0.31	0.26	0.57	0.14
SSN_NLP	R1	0.30	0.25	0.48	0.15
LAWNICS	LAWNICS_1	0.30	0.28	0.46	0.13
TUW_informatics	basic	0.26	0.20	0.49	0.13
TUW_informatics	word_count	0.26	0.21	0.39	0.14
Uottawa_NLP	run3_TFIDF	0.25	0.19	0.31	0.12
fs_hu	fs_hu_task1b	0.23	0.20	0.36	0.08
TUW_informatics	false_friends	0.23	0.19	0.38	0.10
IMS_UNIPD	bm25_lemma	0.23	0.16	0.46	0.15
f_s_hit_1	fs_hit_1_task1b_03	0.21	0.16	0.34	0.13
f_s_hit_2	fs_hit_2_task1b_01	0.20	0.16	0.35	0.10
LAWNICS	LAWNICS_2	0.20	0.15	0.48	0.09
f_s_hit_2	fs_hit_2_task1b_03	0.19	0.13	0.28	0.10
UB	UB-3	0.19	0.15	0.25	0.09
f_s_hit_2	fs_hit_2_task1b_02	0.18	0.12	0.25	0.12
f_s_hit_1	fs_hit_1_task1b_01	0.17	0.09	0.22	0.12
f_s_hit_1	fs_hit_1_task1b_02	0.17	0.09	0.22	0.12
Uottawa_NLP	run1_Glove	0.15	0.08	0.34	0.10
scnu	scnu_3	0.13	0.05	0.15	0.12
SSNCSE_NLP	task_1b_1	0.12	0.07	0.27	0.07
nlpninjas	nlpninjas_st1	0.09	0.02	0.12	0.07
Uottawa_NLP	run2_Doc2Vec	0.04	0.01	0.07	0.02
scnu	scnu_2	0.03	0.00	0.02	0.00

- [3] G. M. D. Nunzio, A study on a stopping strategy for systematic reviews based on a distributed effort approach, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association*, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, 2020, pp. 112–123. URL: https://doi.org/10.1007/978-3-030-58219-7_10. doi:10.1007/978-3-030-58219-7_10.
- [4] G. M. D. Nunzio, Classification of Animal Experiments: A Reproducible Study. IMS Unipd at CLEF eHealth Task 1, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), *Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum*, Lugano, Switzerland, September 9-12, 2019, volume 2380 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_104.pdf.
- [5] H. Wachsmuth, *Text Analysis Pipelines - Towards Ad-hoc Large-Scale Text Mining*, volume 9383 of *Lecture Notes in Computer Science*, Springer, 2015. URL: <https://doi.org/10.1007/978-3-319-25741-9>. doi:10.1007/978-3-319-25741-9.
- [6] Z. Zhao, H. Ning, L. Liu, C. Huang, L. Kong, Y. Han, Z. Han, Fire2019@aila: Legal information retrieval using improved BM25, in: *Working Notes of FIRE 2019 - Forum for Information Retrieval Evaluation*, Kolkata, India, December 12-15, 2019, 2019, pp. 40–45. URL: <http://>

[//ceur-ws.org/Vol-2517/T1-7.pdf](http://ceur-ws.org/Vol-2517/T1-7.pdf).