Topic Modelling Brazilian Supreme Court Lawsuits

Pedro Henrique Luz de Araujo Teófilo Emidio de Campos

Department of Computer Science, University of Brasília, Brasília – DF, Brazil

pedrohluzaraujo@gmail.com, *t.decampos@st-annes.oxon.org*

Sponsored by FAPDF and CNPq

JURIX 2020

10 December 2020

Overview

Introduction

Motivation

- The Brazilian Judiciary is burdened with a huge amount of lawsuits.
- About 80 million suits awaited judgement in 2017.
- Average processing times reaching more than seven years in some cases.
- R$ 90.7 billion spent in 2017 (about 28 billion dollars).
- ML and NLP can contribute—quicker, cheaper more efficient document analysis.

Topic Models

- Unsupervised ML algorithms.
- Objective: discover topics that occur in a collection of documents.
- How: Statistical analysis of the words that comprise the documents.
- Use: organize, explore and index massive amounts of data.
- Advantage: cheap—no need for data annotation.

Objectives

- Employ Latent Dirichlet Allocation (LDA) [Blei et al., 2003] to model Extraordinary Appeals (*Recursos Extraordinários*—RE) received by Brazil's Supreme Court.
- Label and analyse obtained topics (qualitative evaluation).
- Use topic distribution vectors as input for a multi-label classification task (quantitative evaluation).

Contributions

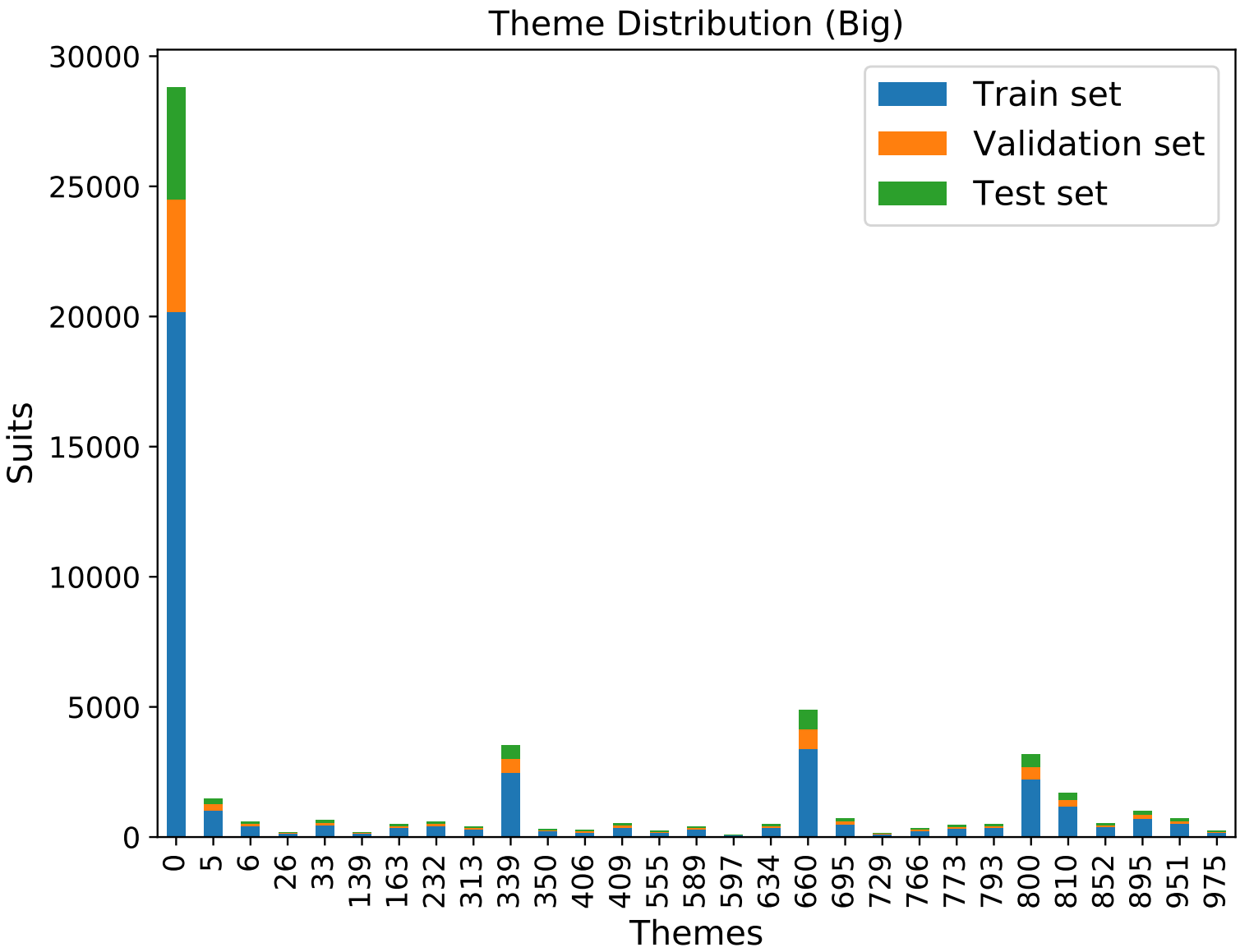
- The qualitative analysis of the semantics of each topic from models with 10 and 30 topics trained on the STF data..
- The quantitative analysis of topic relevance by using topic distribution vectors as input for general repercussion theme (*tema de repercussão geral*) classification. We experiment with models of 10, 30, 100, 300 and 1,000 topics.

Data

Data I

- VICTOR dataset [Luz de Araujo et al., 2020]: 45,532 Extraordinary Appeals.
- Legal proceedings received by the STF (before processing and judging).
- Labels for 28 general repercussion themes (plus one class for other themes)—multi-label classification task.
  - General Repercussion themes: each lawsuit must relate to relevant economic, political, social or legal issues that exceed the interests of the parties.
- Pre-processing: lower-casing, removal of stop words, email and URL tokenizations, identification of simple citations to legislation.

Data II


Theme Distribution (Big)
Train set
Validation set
Test set
Suits
30000
25000
20000
15000
10000
5000
0
Themes
0
5
6
26
33
139
163
232
313
339
350
406
409
555
589
597
634
660
695
729
766
773
793
800
810
852
895
951
975

Latent Dirichlet Allocation

LDA

- Generative model: each document is represented as a random mixture over latent topics.
  - Each topic is represented as a distribution over words.
- The LDA procedure assigns:
  1. A topic distribution for each document.
  2. A topic for each word.
  3. A word distributions for each topic.

Experiments

Model Training

- 10 and 30-topic models.
- Gensim library [Řehůřek and Sojka, 2010]. Online LDA [Hoffman et al., 2010].
- Vocab: words that appear in more than 50 documents and in no more than half of them. 81,418 entries.
- Hyperparameters:
  - Mini-batches de 4,096 lawsuits.
  - 400 iterations per mini-batch.
  - 4 epochs.

Topic Distribution as Text Representation

- Quantitative analysis of detected topics: topic distributions as text representation for classification.
- 10, 30, 100, 300 and 1,000 dimensional vectors.
- Classifier: XGBoost [Chen and Guestrin, 2016].
- Multi-label: train a classifier for each theme.
- Comparison: word count and tf-idf weighted bags-of-words.
- Hyperparameter tuning using the validation set.

Results

Topic Analysis I

- We assign labels by examining the most relevant words from each topic.

$$r(\mathbf{w}, \mathbf{z}|\lambda) = \lambda \log P(\mathbf{w}|\mathbf{z}) + (1-\lambda) \log \frac{P(\mathbf{w}|\mathbf{z})}{P(\mathbf{w})} \tag{1}$$

Table: Topic labels and their respective four most relevant words (10 topics).

Topic	λ	Assigned label	Words
1	0.6	Public servant remuneration	servants, servant, limitation, remuneration
2	0	Criminal Law	narcotic, hydrometer, clandestine, interrogation
3	0.6	Pension Law	benefit, event, retirement, pension
4	0.6	Civil Law	bank, contract, consumer, *projudi*
5	0.6	Right to health	health, city, municipal, medication
6	0.4	OCR errors	*ento*, no, *ro*, *co*
7	0.6	Tax Law	*icms*, *ipi*, tax, income
8	0	Entities	*econorte*, *rcte*, *pieter*
9	0.4	Labor Law	*fgts*, *pss*, hours, payroll
10	0.6	Document access	original, site, access, report

Topic Analysis II

Table: Topic labels and their respective four most relevant words (30 topics).

Topic	λ	Assigned label	Words
1	0.6	Civil liability	damage, damages, compensation, non-material
2	0.22	Expiration of social security benefit	benefit, expiration, limit, social security (*previdenciário*)
3	0.6	Tax Law	treasury, tax, revenue, taxation
4	0.1	Miscellaneous - Legal vocabulary, enttities and laws	serial number, *pet*, stamp, *itaperuna*
5	0.4	Public servant bonus	bonus, performance, inactive, evaluation
6	0.4	Rural social security	rural, contribution, LEI_8212, pension
7	0.6	Public servant remuneration readjustment	readjustment, servants, remuneration, *urv*
8	0.4	OCR errors	*ento*, no, *ro*, *ffl*
9	0.6	Members of the military	military, servant, servicemen, servants
10	0	Criminal Law	clandestine, *sepetiba*, semi-open, narcotic
11	0.4	Contract law	contract, contracts, fee, accounts
12	0.05	Technical Councils	*confea*, *crea*, agronomy, LEI_6496
13	0.2	Public tender	tender, candidate, notice, openings
14	0.4	Anticipation of remuneration readjustment	*upag*, *pccs*, labor, LEI_8460
15	0.6	Right to health	health, medication (plural), treatment, medication (singular)

Topic Analysis III

Table: Topic labels and their respective four most relevant words (30 topics).

Topic	λ	Assigned label	Words
16	0.9	Savings account, interest and monetary correction	correction, monetary, savings account, delay
17	0.6	Document access	original, site, acesse, report
18	0.6	labor complaints	*estran*, *tst*, entity, claimant
19	0.4	Miscellaneous - Consumer Law and Bahia (Brazilian state)	consumer, *salvador*, *bahia*, *pdf*
20	0	Entities - names	*lauxen*, *tainá*, *heloise*, *soeli*
21	0.7	Qualification	*num*, normal, internment, *foz*
22	0.5	insurance	insurance, *previd*, institute, *dpu*
23	0.4	Payroll	hours, *fgts*, payroll, overtime
24	0	Miscellaneous - Organisations, charters and non-Portuguese words	*andaterra*, *peixer*, funds, market
25	0.5	Fiscal documents	*ltda*, *ipi*, *nfe*, *icms*
26	0.4	Rio Grande do Sul (Brazilian state)	*sul* , *grande*, *alegre*, *paese*
27	0.4	Income tax	updated, months, *rra*, *irpf*
28	0.2	Tax Law - circulation of goods	compatible, *issqn*, exit, *eireli*
29	0.2	Miscellaneous - Procedure and Paraná (Brazilian state)	*paraná*, *arq*, *curitiba*, *mov*
30	0.4	Payments	*jam*, *vlr*, received, credit

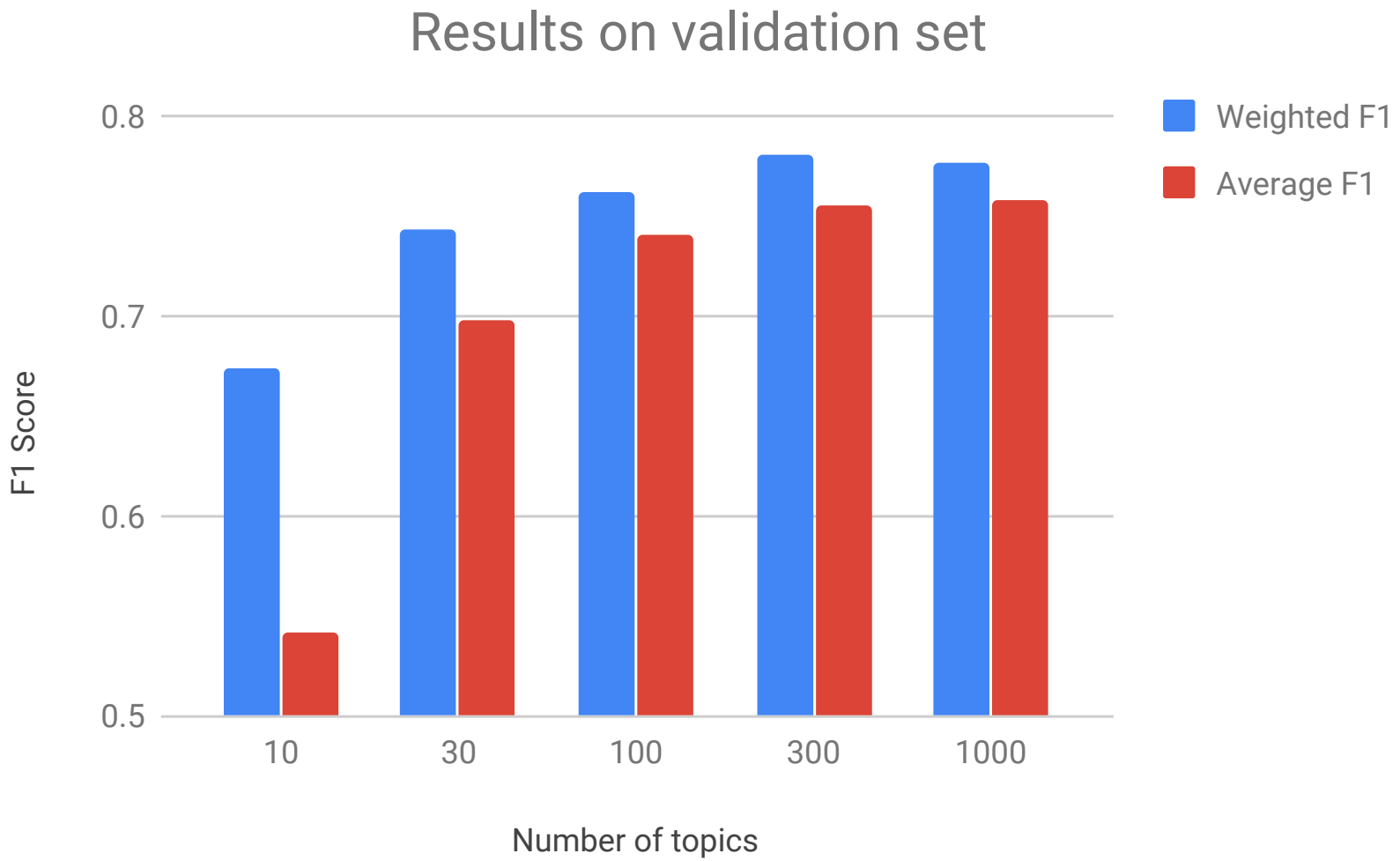
Labelling topics

- Most relevant words + most representative documents.
- Document with highest probability for topic 6: r cm emoi oit incm m t i o i m cofl inoioem oulfl tofl cmcmh co ffl ffl ffl a z a z ffl o t a o u ffl otoidtoaz d to a i o tn ffl em cmcocoulococm eo cocm [...].
- Most frequent words in the document with highest probability for topic 8: ‘mpf”, “distrib”, “dpu”, “instituto”,“seguro”, “previd”, “rafael”, “andrade”, “margalho”,“junior”, “bruyn”, “cornelio”, “herbert”, “pieter” and “sim”.

10 or 30 topics

- More fine-grained topics (4 topics related to tax law).
- More low-quality topics (Bahia + consumer law).
- Problem: document diversity—petitions, rulings, orders, statements, certificates and supporting documents.

Quantitative Analysis I



Quantitative Analysis II

Table: F1 scores (in %) on the test set of each text representation method. Assigning all themes to all samples yield a weighted (by class frequency) F1 score of 41.17 and an average F1 score of 5.48.

	Word counts	Tf-idf	300 topics
Weighted	**89.29**	89.22	78.07
Average	87.54	**88.37**	75.81

Discussion

- Does not outperform traditional methods—but much better results than a baseline that assigns all themes to all samples.
  - Detected topics are related to themes relevant to the Court and may aid with case management.
- Advantage of data compression: describes a lawsuit using 300 dimensions instead of 81,418, a relative reduction of 99.63%. Faster training and inference.

Conclusions

Conclusions

- Analysis of models with 10 and 30 topics found a correspondence between topics and legal matters.
- The quantitative analysis, using a classification task as a proxy for topic quality, supports the hypothesis that the detected topics may be useful for the Court staff.
- The interpretable, low-dimensional representations generated by LDA achieve good classification results.

References I


Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003).
Latent dirichlet allocation.
Journal of Machine Learning Research, 3:993–1022.
Acesso em 2018-11-21.


Chen, T. and Guestrin, C. (2016).
XGBoost: A scalable tree boosting system.
In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.


Hoffman, M., Bach, F. R., and Blei, D. M. (2010).
Online learning for latent dirichlet allocation.
In Lafferty, J. D., Williams, C. K. I., Shawe-Taylor, J., Zemel, R. S., and Culotta, A., editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc.

References II

Luz de Araujo, P. H., de Campos, T. E., Ataides Braz, F., and Correia da Silva, N. (2020).
Victor: a dataset for Brazilian legal documents classification.
In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 1449–1458, Marseille, France. European Language Resources Association.

Řehůřek, R. and Sojka, P. (2010).
Software Framework for Topic Modelling with Large Corpora.
In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.
http://is.muni.cz/publication/884893/en.