# Information mining and visualization of data from the Brazilian Supreme Court (STF): a case study

Flávio Codeço **COELHO**[1], Renato Rocha **SOUZA**[2] , Daniel Magalhães **CHADA**[3] and Pablo de Camargo **CERDEIRA**[4]

[1,2]Professor - Applied Mathematics School - Fundação Getúlio Vargas – Praia de Botafogo, 190, Rio de Janeiro, RJ, Brazil
[3]Phd Student – Business Administration School – Fundação Getúlio Vargas – Praia de Botafogo, 190, Rio de Janeiro, RJ, Brazil
[4]Under Secretary of Consumer Defense – Rio de Janeiro City Hall – Rio de Janeiro, RJ, Brasil

## 1. Introduction

Large collections of text data represent a substantial challenge for the extraction of relevant bits of information to feed subsequent statistical analysis and visualization pipelines.
The peculiarities of the knowledge domain of such a task often requires the implementation of custom natural language processing pipelines, along with specific knowledge organization systems, to describe the relevant terminology. This paper describes a joint research made by the Law School (Direito Rio) and the Applied Math School (EMAp) from Getulio Vargas Foundation (FGV), Brazil. After initial contacts, it was established a joint venture between researchers from EMAp and Direito Rio to analyze the information from judges activities, in some of the Brazilian courts. Initially, the Law School intended to analyze the behavior of the Brazilian Supreme Court (STF) to support public policy-making, and to identify bottlenecks in the judicial processes at the STF. The task was to analyze the text the entire set of recorded judicial decisions, normally accessible through the STF institutional site.

This data had never been analyzed ate this scale before, so a great deal of exploratory analyses was expected in order to reveal hidden patterns in the data. A number of a priori questions were proposed, basically aimed at determining if the STF was performing according to its constitutionally defined role, and if not, in what way it could be changed to better serve its purpose.

Some of the methodology described herein was applied to the generation of the results published in the Project first technical report(“I Relatório – abril/2011 – O Múltiplo Supremo”, 2011). The results presented here go in the more general direction of exploratory data analysis.

## 2. Methodology

## 2.1 Data Retrieval

The data for the analysis consisted of the entire collection of judicial decisions from 1988 to the present. This data although public is only available to the public through a simple web based query interface. Given the dimension of the dataset manual retrieval was out of the question. Therefore, a bespoke software tool was developed to fetch the documents. This process resulted in a collection of approximately 1.2 million cases (HTML documents). Due the fragility of the STF web servers, the data retrieval phase lasted approximately four months.

In order to facilitate information extraction, the original documents, in HTML format were parsed and stored on a relational database for further analysis. We sought to preserve as much structure as possible when analyzing the documents retrieved. Figure 1 has an example of the kind of relational structure we manage to elicit from the raw documents. From these initial tables a large number of analytical table were derived for analytical purposes.

*Figure 1: Main tables of the resulting relational database.*

The next big challenge, was to extract specific information from the large collection of texts in “t_decisoes.decisao”. This variable contains the full text of each decision.

## 2.2 Extracting information

Most of the time, it is not difficult to deal with the quantitative aspects of the information to be extracted in the documents. Identifying specific words, its frequencies and collocations are the basis of the quantitative textual analysis. Techniques like identifying Term frequency / Inverse Document frequency and the extraction of noun phrases allowed us to identify the subject of the cases and their most used expressions. We were also able to identify the main litigants, the numeric evolution of the legal cases throughout the last decades, the lifespan of each case and so on. Information like names, dates, locations, legal citations, case subjects, etc. were all analyzed, aiming to characterize STF processes and the judges work. After identifying the law citation signature - using a regular expressions - we were able to associate the set of laws that were cited in each legal case (figure 2).

From the sets of citations extracted from each decision, we proposed some a metric for law usage based on Shannon's entropy(Bommarito II & Katz, 2010). The rationale behind it was that the entropy be greater for cases in which the decision was based on wider jurisprudence while decision based on a single citation to a law, would give zero entropy.

*Figure 2: Example of usage o regular expression to extract patterns of law citation from the decision text in HTML format.*

However, for the quintessential information to be mined, it was necessary to resort to more complex natural language processing (NLP) techniques to try to elicit and "understand" the "meaning" of some portions of the texts. Again, the development of bespoke software for NLP and a small set of STF specific, lightweight domain ontologies (see figure 3) we have built, allowed us to take steps further on the analysis of the material.

### 2.3 Information Visualization

*Figure 3: Light Domain Ontology of the stages of the life history a legal case. Names are maintained in the original language as their translation would convey little meaning. This diagram is here to illustrate the complexity of the taxonomy required to understand the data.*

A good strategy to identify hidden patterns on data is through visualization. Our tasks involved building graphs (networks) to connect various informations pieces extracted from the texts, depicting the established relationships. Due to the size and complexity of our dataset, graphs with hundreds of thousand (and sometimes many more) nodes were common. The graphs allowed us to identify, for example clusters of judges that were employing similar sets of laws, and animations permitted us to see how laws were used throughout the time. These graphs, though, not show here due to their complexity were instrumental in the identification of many patterns which led to other analyses.

## 3. Results

The application of the light ontology depicted on figure 3 allowed us to classify and analyze the stages, for example, with respect to duration (see figure 4). This dissection of the cases allowed us to pinpoint stages which contributed more to the total lifespan of a judicial case. Another important aspect revealed, was the multi-modal nature of the duration distributions, which indicates some severe bottlenecks for certain types of cases. Once identified, these bottlenecks can be further investigated and addressed.

*Figure 4: Distribution of durations of the "reautuação" stage. The left y-axis is the absolute number of this kind of stage in the database and the right y-axis is the normalize scale of the density plot (in green).*

The entropy of decisions, which was proposed as a proxy to the complexity of decisions as measured by the jurisprudence applied, showed a marked increase in time (figure 5). This may reflect the technological improvements in the accessibility of relevant jurisprudence.

*Figure 5: Entropy increase with Time. This effect may be due to the improvement in the accessibility of jurisprudence  with time. For this graph the entropy was calculated on the citations aggregated over an year.*

## 4. Conclusions

This work is pioneering in the large scale analysis of the dynamics of  the Brazilian supreme court. The project's initial technical report(“I Relatório – abril/2011 – O Múltiplo Supremo”, 2011), related important distortion in the work load of STF and has led the president of the STF, to present a constitutional amendment proposal, to fix these distortions.

One of the biggest challenges we have identified in this dataset, is the difficulty in asserting the outcome of a decision(Kastellec, 2010). The cryptic and vague language used to express the veredict makes it very hard even with the help of advanced NLP techniques, to know which part  the decision favors. Additionally, the frequent multiple appeals complicate matters even further.

The richness of this data is unquestionable, and to explore it we required advances mathematical and computational techniques to extract, analyze and visualize information. The initial stages of this work have taken advantage of some well established analytical tools which, nevertheless had to be adapted to work with large volumes of data. Distributed processing versions of these tools has been particularly lacking, leading our group to invest in the development of such tools. These developments take time, but  paint a bright future for the analyses of large text corpora.

In this paper, we sought to present some of the challenges in the analysis of the entire recorded history of the Brazilian supreme court. We hope that the preliminary analyses presented in this paper, are sufficient to illustrate the richness of this dataset, and its potential for further studies.

This is an ongoing project, and the tools and insights developed within it are already bearing fruit to other analytical projects targeting  other Brazilian courts, much larger in size.

## 5. References

Bommarito II, M. J., & Katz, D. M. (2010). A Mathematical Approach to the Study of the United States Code. *arXiv:1003.4146*. Recuperado de http://arxiv.org/abs/1003.4146

I Relatório – abril/2011 – O Múltiplo Supremo. (2011).*Supremo em Números*. Rio de Janeiro, RJ. Recuperado de http://www.supremoemnumeros.com.br/i-relatorio-abril2011-o-multiplo-supremo/

Kastellec, J. P. (2010). The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees. *Journal of Empirical Legal Studies, 7*(2), 202–230. doi:10.1111/j.1740-1461.2010.01176.x