Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence
An Experimental Analysis

Robert A. N. de Oliveira and Methanias Colaço Júnior
UFS - Universidade Federal de Sergipe, São Cristóvão, SE, Brazil

Keywords: Dimensionality Reduction, Experimental Analysis, Jurisprudence, Stemming.

Abstract: Stemming algorithms are commonly used during textual preprocessing phase in order to reduce data dimensionality. However, this reduction presents different efficacy levels depending on the domain that it's applied to. Hence, this work is an experimental analysis about dimensionality reduction by stemming a real database of judicial jurisprudence formed by four subsets of documents. With such document base, it is necessary to adopt techniques that increase the efficiency of storage and search for such information, otherwise there is a loss of both computing resources and access to justice, as stakeholders may not find the document they need to plead their rights. The results show that, depending on the algorithm and the collection, there may be a reduction of up to 52% of these terms in the documents. Furthermore, we have found a strong correlation between the reduction percentage and the quantity of unique terms in the original document. This way, RSLP algorithm was the most effective in terms of dimensionality reduction, among the stemming algorithms analyzed, in the four collections studied and it excelled when applied to judgments of Appeals Court.

1 INTRODUCTION

Every day, the courts, through their magistrates, judge various themes of the Law, generating a large base of legal knowledge that guides new decisions and works as an argumentative base to the related parties that plead their interests. Hence, from the corpus formed by a uniform set of decisions handed down by the judiciary on a particular subject (Maximiliano, 2011), emerges the concept of jurisprudence, fundamental tool for legal professionals to exercise their role.

This way, those decisions generate three types of documents (Santos, 2001):

- Trial Court sentence: when the judge utters a procedural trial in first instance;
- Monocratic Decision: when a magistrate decides alone, in second instance, a lawsuit that has uniform interpretation;
- Judgment: when collegiate organ, composed by one rapporteur and at least two magistrates, utters sentence in second instance.

A decision in second instance may be the result of an appeal from a sentence uttered by an Appeals Court judge or by Special Courts judge, creating specifics documents for each one of them.

With such document base, it is necessary to adopt techniques that increase the efficiency of storage and search for such information, otherwise there is a loss of both computing resources and access to justice, as stakeholders may not find the document they need to plead their rights.

In this scenario, according to (Flores and Moreira, 2016; Orengo et al., 2007), stemming algorithms can reduce the texts dimensionality, thereby improving the use of computing resources, and increase the relevancy of the results returned by retrieval systems. In fact, these algorithms are commonly used during textual preprocessing phase in order to reduce data dimensionality. However, this reduction presents different efficacy levels depending on the domain it is applied. The legal universe has its own jargon and we have not found reports in the literature showing that the same benefits are obtained when stemming is applied to jurisprudential bases.

Therefore, the objective of this study was to analyze, following an experimental process, using quantitative metrics, the effectiveness of stemming on the dimensionality reduction of real jurisprudential bases. The results showed that, depending on the algorithm and the collection, there may be a reduction of up to 52% of these terms in the documents. Furthermore,

Oliveira, R. and Júnior, M.
Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence - An Experimental Analysis.
DOI: 10.5220/0006317100990105
In *Proceedings of the 19th International Conference on Enterprise Information Systems (ICEIS 2017) - Volume 1*, pages 99-105
ISBN: 978-989-758-247-9

we have found a strong correlation between the reduction percentage and the quantity of unique terms in the original document. This way, RSLP algorithm was the most effective in terms of dimensionality reduction in the four collections analyzed and it was excelled when applied to judgments of Appeals Court.

The rest of the paper is structured as follows. Section 2 presents the related work. Section 3 conceptualizes stemming and describes the algorithms used in this research. In Section 4, we present the definition and planning of the experiment. In Section 5, we show the experiment execution. Section 6 contains the results of the experiment. Finally, Section 7 presents the conclusion and future work.

2 RELATED WORK

This paper analyses the impact of stemming on dimension reduction of jurisprudence texts in Brazilian Portuguese, therefore this section will present articles that had a similar approach.

(Alvares et al., 2005) carried out an assessment of vocabulary reduction, along with *overstemming* and *understemming* errors described in the following section , by stemming 1,500 words available in dictionaries of Brazilian Portuguese language. This approach differs from ours, since they propose a new stemming algorithm, StemBR, and compares it to two different ones. On the other hand, here we will use algorithms available in (Lucene, 2005).

(Orengo et al., 2007) conducted a comparative study of stemming algorithm related to reduction of terms in a collection of tests formed by the Folha de São Paulo newspaper and evaluated its impact on the results returned by a retrieval system. Different from this proposal, there are no further details on the dimensionality reduction per document, considering that they focused on an analysis of the metrics taken from the search system.

Similar to the article mentioned above, (Flores and Moreira, 2016) measured the impact of stemming on testing collections available in different languages (English, French, Portuguese and Spanish). This way, they collected dimensionality reduction metrics, *overstemming*, *understemming* and also measured the reflection on the application of these algorithms in precision and recall of information retrieval systems. However, due to its scope, the paper did not go into detail on any of the analyzes.

It is worth mentioning that, until now, papers that run a detailed analysis of dimensionality reduction per document, like the one presented, were not found. In addition, related work used collections that do not reflect the documents found in the legal universe.

3 STEMMING

The stemming process consists of grouping different words connected by a common stem, based on a set of rules which act by removing suffixes and prefixes (Figure 1). Table 1 shows the application of five stemming algorithms used during this experiment with six distinct words, in which NoStem is the control group, i.e., it generates no reduction of terms.

Except for the control group, the other algorithms used in the experiment are based on rules and act by removing suffixes (Flores and Moreira, 2016):

- Porter: originally written in English, in 1980, and adapted to Portuguese language later;
- RSLP (*Removedor de Sufixos da Lingua Portuguesa*): published in 2001, contains approximately 200 rules and an exception list to almost each one of them;
- RSLP-S: a lean version of RSLP that uses only plural reduction;
- UniNE: contains less rules than Porter and RSLP, however it is more aggressive than RSLP-S.

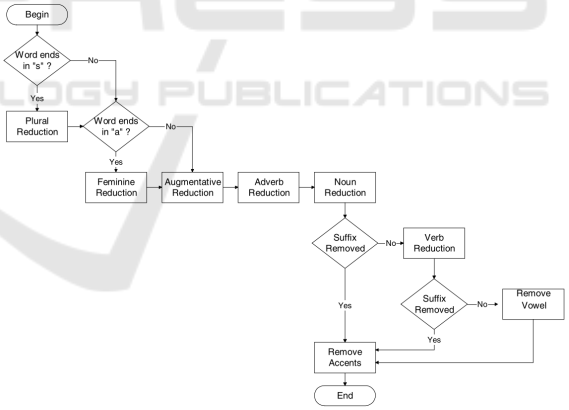

Figure 1: Sequence of steps for the RSLP algorithm(Orengo et al., 2007).

Thus, considering semantic and morphological aspects, a stemming algorithm can commit two error types: a) *overstemming*, when the part removed is not a suffix, instead it is part of the word stem; and b) *understemming*, when the suffix removal does not take place fully. In this study, such errors will not be evaluated.

Table 1: Example of stemming using five algorithms of the experiment.

NoStem	constituições	limitações	regimento	considerando	anuência	estelionato
Porter	constituiçõ	limit	regiment	consider	anuênc	estelionat
RSLP	constitu	limit	reg	consider	anu	estelionat
RSLP-S	constituição	limitação	regimento	considerando	anuênc	estelionato
UniNE	constituica	limitaca	regiment	considerand	anuenci	estelionat

4 DEFINITION AND EXPERIMENT PLANNING

In this and next two sections, this paper will be presented as an experimental process according to Wohlin et al. guidelines, described in (Wohlin et al., 2012). Therefore, initially, we will explain planning and definition of the experiment. After that, we will refer to its execution and data analysis.

4.1 Goal Definition

The goal of this work is to analyze the impact of stemming algorithms in the dimensionality reduction of jurisprudential documents.

In order to achieve it, we will conduct an experiment, in a controlled environment, in which the reduction of unique terms per document will be measured, inside each collection, along with an analysis of statistically significant differences of effectiveness of the same algorithm, among four documentary bases adopted by the study.

The following is the goal formalization, according to GQM model proposed by Basili (Basili et al., 1994): **Analyze** stemming algorithms **with the purpose of** evaluating them **with respect to** dimensionality reduction and effectiveness **from the point of view of** data analysts **in the context of** jurisprudential documents.

4.2 Planning

Context Selection. The experiment will be *in vitro* and will use the entire judicial jurisprudence database of Supreme Court of the State of Sergipe, formed by four collections: a) judgments of Appeals Court (181,994 documents); b) monocratic decisions of Appeals Court (37,142 documents); c) judgments of Special Courts (37,161 documents); and d) monocratic decisions of Special Courts (23,151 documents).

Dependent Variables. The average of unique terms per document (UTD) and the average percentage of reduction of unique terms per document (RP) taken from the stemmer application.

- Unique Terms: UTD_S = Frequency of unique terms after document stemming.
- Average of unique terms: $\mu = (UTD_{S1} + UTD_{S2} + + UTD_{Sn})/n$
- Reduction percentage: $RP_R = 100 - (UTD_S * 100)/UTD_{NoStem}$
- Average of reduction percentage: $\mu = (RP_{S1} + RP_{S2} + ... + RP_{Sn})/n$

Independent Variables. Document collection of judgments of Appeals Court (JAC), monocratic decisions of Appeals Court (MAC), judgments of Special Courts (JSC) monocratic decisions of Special Courts (MSC); the stemming algorithms (NoStem, Porter, RSLP, RSLP-S and UniNE).

Hypothesis Formulation. The research questions for this experiment are: do stemming algorithms reduce the dimensionality of jurisprudential documents? Is the effectiveness of each algorithm the same for all four collections studied?

For the first research question, we considered the quantity of unique terms per document as a metric to evaluate the dimensionality reduction. For the second question, we adopted the reduction percentage of each algorithm, considering that the comparison was made among documents of a different nature, making the use of absolute values inadequate. In this scenario, the following assumptions will be verified:

Hypothesis 1 (For each of the four collections).

- **Null Hypothesis $H0^{UTD}$:** The stemming algorithms have the same average of unique terms per document ($\mu_{NoStem^{UTD}} = \mu_{Porter^{UTD}} = \mu_{RSLP^{UTD}} = \mu_{RSLP-S^{UTD}} = \mu_{UniNE^{UTD}}$).
- **Alternative Hypothesis $H1^{UTD}$:** The stemming algorithms have different averages of unique terms per document ($\mu_{i^{UTD}} \neq \mu_{j^{UTD}}$ for at least one pair(i,j)).

Hypothesis 2 (For each of the stemming algorithms).

- **Null Hypothesis $H0^{RP}$:** The percentage averages of reduction of unique terms per document are the same in all four collections ($\mu_{JAC^{RP}} = \mu_{MAC^{RP}} = \mu_{JSC^{RP}} = \mu_{MSC^{RP}}$).

- **Alternative Hypothesis H1^{RP}:** The percentage averages of reduction of unique terms per document are different in all four collections ($\mu_{iRP} \neq \mu_{jRP}$ for at least one pair(i,j)).

Selection of Participants and Objects. The documents of each collection were chosen randomly taking into consideration their number of characters. So, the quantity of documents were determined by the sample calculation of a finite population:

$$n = \frac{z^2.\sigma^2.N}{e^2.(N-1).z^2.\sigma^2} \quad (1)$$

Where, n is the sample size, z is the standardized value (we adopted 1.96, i.e., 95% of trust level), σ is the standard deviation of population, e is the margin of error (we adopted 5% of σ) and N is the population size. Table 2 shows the number of selected documents after sample calculation, along with size, mean and standard deviations of the population.

Table 2: Sample size per collection.

Coll.	N	μ	σ	n
JAC	181,994	11,626.65	8,270.08	1,524
MAC	37,142	8,396.27	6,940.01	1,476
JSC	37,161	9,509.41	5,718.97	1,476
MSC	23,151	6,569.90	4,009.80	1,442

Experiment Project. The jurisprudential documents have a great variability in terms of number of characters, thus, in order to ensure confidence on hypothesis tests, we will utilize a *randomized complete block design* (RCBD) (Wohlin et al., 2012) , this way, each algorithm will be applied to the same document and those documents will be randomly taken from each collection, increasing the experiment precision. Furthermore, before applying stemming, a preprocessing for textual standardization will be performed in which the content of documents will be shifted to small caps and punctuation characters will be removed. NoStem represents the unique terms of the document with no stemming, therefore, it acts as a control group.

Instrumentation. We developed a Java application in order to iterate on each document of the sample, applying stemming algorithms and counting the frequency of unique terms after the execution. In the end, the application will store the observations performed in a CSV file (Comma Separated Values) for each collection.

5 EXPERIMENT EXECUTION

5.1 Preparation

The preparation phase consisted of obtaining collections referring to judicial jurisprudence. Thus, documents were extracted from an OLTP base (Online Transaction Processing) and converted to XML format (eXtensible Markup Language) facilitating the experiment packaging.

5.2 Execution

By the end of previous phases, the experiment started executing the Java application, in accordance with what was defined in the planning phase.

5.3 Data Collection

The application recorded, for each collection, the document identifier, the number of unique terms and the stemming algorithm adopted CSV format (Table 3).

Table 3: Input example in CSV file.

ID,UTD,Stemmer
20110020500144363266 2,679,NoStem
20110020500144363266 2,580,Porter
20110020500144363266 2,547,RSLP
20110020500144363266 2,651,RSLPS
20110020500144363266 2,636,UniNE

5.4 Data Validation

The Java application was built using Test Driven Development (TDD) (Agarwal and Deep, 2014) approach , therefore, we wrote unit test cases to validate if the frequency count of unique terms per document worked as expected.

Averages of unique terms per document were computed and the percentage averages of dimensionality reduction were obtained by applying stemming algorithms, considering control group.

To support this analysis, interpretation and results validation, we used five types of statistical tests: the Shapiro-Wilk test, the Friedman test, the Kruskal-Wallis test, the Wilcoxon test and the Mann-Whitney test. The Shapiro-Wilk test was used to verify sampling normality, as literature shows it has higher test power than other approaches (Ahad et al., 2011; Razali and Wah, 2011). Considering RCBD project of the experiment, with a factor and multiple treatments, the Friedman test (Theodorsson-Norheim, 1987) and the Kruskal-Wallis test (Wohlin et al., 2012) were

used to demonstrate the existence of different averages of paired and independent samples, respectively, that did not obtain data normality, verifying χ^2 (Chi-Square) magnitude. Finally, a post hoc analysis of the Friedman and Kruskal-Wallis tests was run using, respectively, the Wilcoxon and Mann-Whitney tests, to compare the averages of each treatment, applying the Benferroni adjustment in the significance level (Holm, 1979). As we perform multiple comparisons among different treatments, this adjustment is important, since it reduces the possibility of rejection of the null hypothesis when it is indeed true (Error Type I) (Dunn, 1961).

All statistical tests were performed using SPSS (SPSS, 2012) and re-evaluated with R (Team, 2008) and SciPy (Jones et al., 2001).

6 RESULTS

To answer experimental questions, CSV files generated by the Java application were analyzed. The results of stemming impact on the average of unique terms per document and on percentage average of dimensionality reduction per document, can be seen in Figure 2 and Figure 3, respectively.

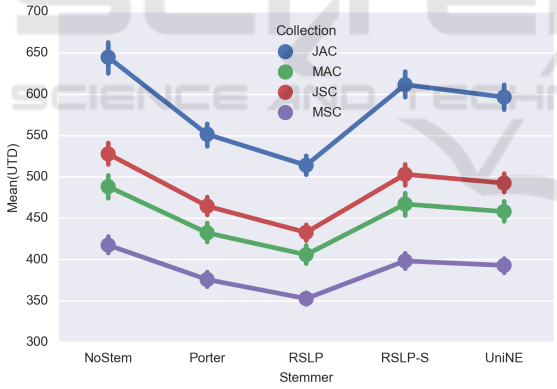

Figure 2: The average number of unique terms per document obtained by each stemmer.

6.1 Analysis and Interpretation

Visually, analyzing Figures 2 and 3, a stemming application seems to generate differences in both, the average of reduction of unique terms per document and in the average percentage of dimensionality reduction. However, it is not possible to claim that with no statistical evidences that confirm that.

Finally, we used 95% of trust level ($\alpha = 0.05$), to the entire experiment and, later on, we analyzed if the samples had normal distribution. However, this

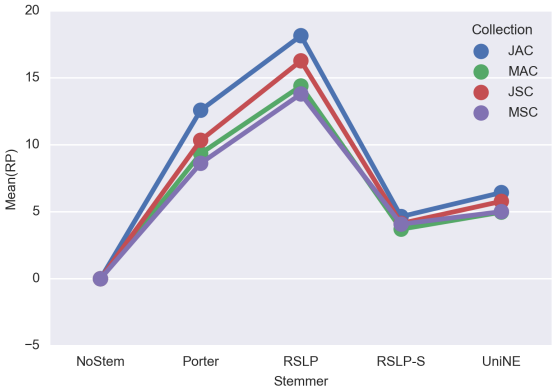

Figure 3: The average percentage of dimensionality reduction per document generated by stemming.

hypothesis was rejected, since the Shapiro-Wilk test obtained p-value below 0.001, lower than the significance level adopted, in every collection and algorithm. This way, considering data distribution and RCBD design adopted for the experiment, we performed the Friedman test to verify Hypothesis 1 (Table 4).

Table 4: Results of the Friedman tests for the Hypothesis 1.

Coll.	χ^2	p-value
JAC	5,883.84	0.000
MAC	5,590.32	0.000
JSC	5,863.67	0.000
MSC	5,474.95	0.000

After applying the tests, we found a strong evidence for the hypothesis $H1^{UTD}$, showing that the averages of unique terms per document are not the same among the algorithms, since we verified a p-value below 0.001, to every collection, and χ^2 equal to 5,883.84; 5,590.32; 5,863.67 and 5,474.95, referred to collections JAC, MAC, JSC and MSC, respectively. After a post-hoc analysis with the Wilcoxon test, applying the Benferroni correction ($\alpha = \alpha / 10$), we found the following order related to the number of unique terms obtained after stemming: NoStem > RSLP-S > UniNE > Porter > RSLP, to every collection. In other words, RSLP algorithm was the most effective in the reduction of unique terms per document.

For Hypothesis 2, considering that the jurisprudential bases are independent, i.e., the same document does not appear in more than one collection, we adopted Kruskal-Wallis tests (Table 5).

According to the results, the percentage averages of reduction of algorithms are not the same for every collection, since p-value was less than 0.001 and χ^2 equal to 687.93; 711.83; 250.31 and 295.25, referred,

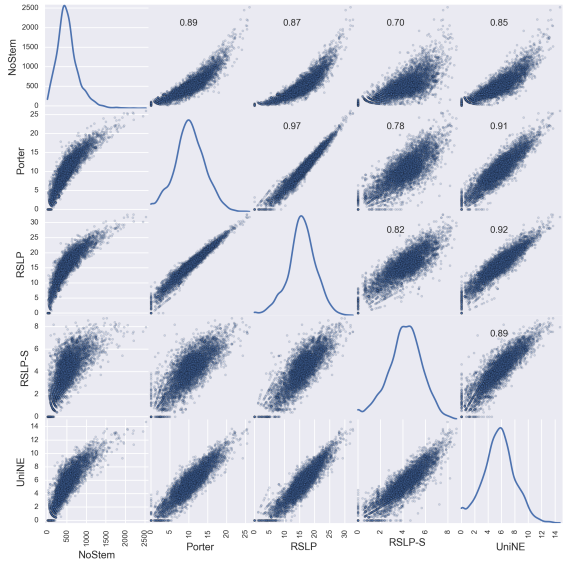

Figure 4: Correlation matrix among stemming algorithms. NoStem unit is UTD and others are RP.

Table 5: Results of the Kruskal-Wallis tests for the Hypothesis 2.

Stemmer	χ^2	p-value
Porter	687.93	0.000
RSLP	711.83	0.000
RSLP-S	250.31	0.000
UniNE	295.25	0.000

respectively, to Porter, RSLP, RSLP-S and UniNE algorithms, therefore, hypothesis $H0^{RP}$ was refuted. By conducting a post-hoc with the Mann-Whitney test, also applying the Benferroni adjustment ($\alpha = \alpha / 6$), we noticed that stemming algorithms reduced dimensionality more effectively in JAC collection.

As it can be seen in the first line of the correlation matrix showed by Figure 4, there is a strong positive correlation, ranging from 0.70 to 0.89, between the quantity of unique terms per document and the reduction percentage achieved by stemming algorithms. In other words, it suggests that the more words jurisprudential documents have, the better results the analyzed stemming algorithms will get. Furthermore, in the same figure, we noticed a linear relation between the algorithms, indicating that they maintain a proportionality related to the potential of dimensionality reduction of texts. Thus, the Porter and RSLP algorithms, for example, have a 0.97 correlation coefficient, indicating an almost perfect positive linear relationship.

To illustrate this correlation potential between quantity of unique terms and reduction percentage, we considered the entire sample of each collection as a single document. Then, we applied stemming algorithms to the collection.

Table 6: Sample dimensionality reduction.

Coll.	Porter	RSLP	RSLP-S	UniNE
JAC	46%	52%	12%	24%
MAC	39%	45%	11%	22%
JSC	35%	41%	10%	20%
MSC	35%	41%	10%	19%

In this scenario, shown in Table 6, one of the stemming algorithms achieved 52% of reduction (JAC-RSLP), confirming the linear relation mentioned above. We also noticed that the order of effectiveness was equivalent to the one found in the experiment using single documents (RSLP > Porter > UniNE > RSLP-S > NoStem).

Hence, due to the results found, it is possible to say that RSLP algorithm reduced judicial jurisprudence dimensionality more effectively than Porter, UniNE and RSLP-S. Besides, JAC collection showed higher reduction of unique terms, regardless which stemming algorithm was adopted.

6.2 Threats to Validity

Because the data was collected and analyzed by the authors, there happens to be a strong threat to internal and external validities. However, there is not conflict of interest. Thus, there are no reasons to privilege an algorithm over another. To mitigate any possible bias, documents were chosen randomly, according to RCBD guidelines.

7 CONCLUSION AND FUTURE WORK

This paper showed an important contribution related to application of stemming algorithms on jurisprudential bases. Indeed, data dimensionality reduction is used in a variety of text processing techniques, however, we have not found, so far, a quantitative study that analyzes its impact on Brazilian judicial real decisions.

According to experimental results, the use of stemming algorithms reduced the average of unique terms per document by 52%. Furthermore, we have found a strong correlation between the reduction percentage and the quantity of unique terms in the original document. This way, among the stemming algorithms analyzed, RSLP was the most effective in terms of dimensionality reduction in the four collections studied and it was excelled when applied to

judgments of Appeals Court.

Finally, for future work, we intend to analyze the reflection of the reduction from the perspective of a judicial information retrieval system, measuring its impact on MAP, R-Precision and Pr@10 metrics.

ACKNOWLEDGEMENTS

This study counted on Supreme Court of the State of Sergipe full support by sharing their database of judicial jurisprudence in text format.

REFERENCES

Agarwal, N. and Deep, P. (2014). Obtaining better software product by using test first programming technique. *Proceedings of the 5th International Conference on Confluence 2014: The Next Generation Information Technology Summit*, pages 742–747.

Ahad, N. A., Yin, T. S., Othman, A. R., and Yaacob, C. R. (2011). Sensitivity of normality tests to non-normal data. *Sains Malaysiana*, 40(6):637–641.

Alvares, R. V., Garcia, A. C. B., and Ferraz, I. (2005). STEMBR: A stemming algorithm for the Brazilian Portuguese language. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3808 LNCS:693–701.

Basili, V. R., Caldiera, G., and Rombach, H. D. (1994). The goal question metric approach. *Encyclopedia of Software Engineering*, 2:528–532.

Dunn, O. J. (1961). Multiple Comparisons Among Means. *Journal of the American Statistical Association*, 56(293):52–64.

Flores, F. N. and Moreira, V. P. (2016). Assessing the impact of Stemming Accuracy on Information Retrieval A multilingual perspective. *Information Processing & Management*, 0:1–15.

Holm, S. (1979). A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Jones, E., Oliphant, T., Peterson, P., and Others, A. (2001). SciPy: Open source scientific tools for Python.

Lucene, A. (2005). A high-performance, full-featured text search engine library. *URL: http://lucene.apache.org*.

Maximiliano, C. (2011). *Hermenêutica e Aplicação do Direito*. Forense, Rio de Janeiro, 20 edition.

Orengo, V. M., Buriol, L. S., and Coelho, A. R. (2007). A Study on the Use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval. pages 91–98.

Razali, N. M. and Wah, Y. B. (2011). Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. *Journal of Statistical Modeling and Analytics*, 2(1):21–33.

Santos, W. (2001). *Dicionário Jurídico Brasileiro*. Livraria Del Rey Editora LTDA.

SPSS, I. (2012). Statistical package for social science. *USA: International Business Machines Corporation SPSS Statistics*.

Team, R. D. C. (2008). R: A Language and Environment for Statistical Computing. Technical report, R Foundation for Statistical Computing, Vienna, Austria.

Theodorsson-Norheim, E. (1987). Friedman and quade tests: Basic computer program to perform nonparametric two-way analysis of variance and multiple comparisons on ranks of several related samples. *Computers in biology and medicine*, 17(2):85–99.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*. Springer Berlin Heidelberg, Berlin, Heidelberg.