UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais

Dissertação de Mestrado

Robert Anderson Nogueira de Oliveira

São Cristóvão – Sergipe

2017

UNIVERSIDADE FEDERAL DE SERGIPE

CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

Robert Anderson Nogueira de Oliveira

Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Orientador(a): Prof. Dr. Methanias Colaço Rodrigues Júnior

São Cristóvão – Sergipe

2017

Robert Anderson Nogueira de Oliveira

Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais/ Robert Anderson Nogueira de Oliveira. – São Cristóvão – Sergipe, 2017-
77 p. : il. (algumas color.) ; 30 cm.

Orientador: Prof. Dr. Methanias Colaço Rodrigues Júnior

Dissertação de Mestrado – UNIVERSIDADE FEDERAL DE SERGIPE
CENTRO DE CIÊNCIAS EXATAS E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO, 2017.

1. Engenharia de software experimental. 2. Textos jurídicos. I. Prof. Dr. Methanias Colaço Rodrigues Júnior. II. Universidade Federal de Sergipe. III. Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais.

CDU 02:141:005.7

Robert Anderson Nogueira de Oliveira

Análise e Avaliação Experimentais de Técnicas para Recuperação de Documentos Jurisprudenciais

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Sergipe como requisito parcial para a obtenção do título de mestre em Ciência da Computação.

Trabalho aprovado. São Cristóvão – Sergipe, 14 de Agosto de 2017:

Prof. Dr. Methanias Colaço Rodrigues Júnior
Orientador

Prof. Dr. Jugurta Rosa Montalvão Filho
Convidado 1

Prof. Dr. Leonardo Nogueira Matos
Convidado 2

Prof. Dr. Renato Lima Novais
Convidado 3

São Cristóvão – Sergipe
2017

À minha amada esposa Lívia, minha filha Giovanna e
meu filho Henrique, que está a caminho.

Agradecimentos

À misteriosa e singular cadeia de eventos que me trouxe até o presente momento.

Aos meus pais, por terem me ensinado desde cedo a importância da educação e por não terem poupado esforços para me proporcioná-la.

Aos meus irmãos Kyllbert e Albert, pelos momentos incríveis vividos desde a infância.

A todos os verdadeiros mestres, com ou sem diploma, que tive o prazer de encontrar durante minha caminhada.

Ao meu orientador, Prof. Dr. Methanias Colaço Rodrigues Júnior, por ter despertado a minha paixão pela engenharia de software experimental, por sua disponibilidade e pelas valiosas orientações.

Ao Prof. Dr. Jugurta Rosa Montalvão Filho, por ter aceitado prontamente o convite para participar da minha banca de avaliação e pelas sugestões de suma relevância fornecidas para aprimoramento do presente trabalho.

À Prof.ª Dr.ª Daniela Barreiro Claro, pelas críticas construtivas durante a avaliação do projeto de pesquisa.

Aos Professores Dr. Leonardo Nogueira Matos e Dr. Renato Lima Novais, por terem aceitado fazer parte da banca examinadora, contribuição ímpar para o enriquecimento deste trabalho.

Aos professores Dr. Rogério Patrício Chagas do Nascimento e Dr Michel dos Santos Soares, pela incomensurável ajuda que tornou possível a apresentação do artigo em Portugal.

A todos os demais professores do PROCC/UFS que tive a satisfação de conhecer durante minha estadia no Mestrado, pelo conhecimento transmitido.

Ao Tribunal de Justiça do Estado de Sergipe, por ter disponibilizado a base jurisprudencial para realização dos experimentos.

Aos meus amigos da Diretoria de Produção de Suporte, em especial da Divisão de Banco de Dados, por fazerem do meu local de trabalho um lugar prazeroso e que emana criatividade.

A minha esposa Lívia, por sempre acreditar em mim e me fazer a cada dia um ser humano melhor do que jamais pensei ser. Sem o seu esforço, paciência, abdicação e compreensão, esse trabalho não seria possível.

Resumo

Os algoritmos de radicalização são normalmente utilizados na fase de pré-processamento textual, com o intuito de reduzir a dimensionalidade dos dados. No entanto, essa redução apresenta diferentes níveis de eficácia, a depender do domínio sobre o qual eles são aplicados. Desse modo, a título ilustrativo, há relatos na literatura que evidenciam o efeito da radicalização quando aplicada a dicionários ou bases textuais de notícias. Por outro lado, não foi encontrado qualquer estudo que analisasse o impacto da radicalização sobre a jurisprudência judicial brasileira, composta por decisões da magistratura, instrumento fundamental para que os profissionais do Direito possam exercer o seu papel. Assim, o presente trabalho apresenta os resultados obtidos por meio da análise e avaliação experimentais da radicalização aplicada sobre documentos jurisprudenciais verídicos, oriundos do Tribunal de Justiça do Estado de Sergipe. Os resultados mostraram que, dentre os algoritmos analisados, o RSLP possuiu a maior capacidade de redução de dimensionalidade dos dados. Outrossim, mediante avaliação extrínseca dos algoritmos de radicalização sobre a recuperação de documentos jurisprudenciais, os algoritmos RSLP-S e UniNE, radicalizadores menos agressivos, apresentaram o melhor relação custo-benefício, haja vista que reduziram a dimensionalidade dos dados e tiveram eficácia similar, ou até mesmo superior, à do grupo de controle.

Palavras-chave: Engenharia de software experimental, textos jurídicos, redução de dimensionalidade, jurisprudência.

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. Thus, for instance, there are reports in the literature that show the effect of radicalization when applied to dictionaries or textual bases of news. On the other hand, we have not found any studies analyzing the impact of radicalization on Brazilian judicial jurisprudence, composed of decisions handed down by the judiciary, a fundamental instrument for legal professionals to play their role. Thus, this work presents the results obtained through the analysis and experimental evaluation of the stemmers applied on real jurisprudential documents, originating from the Court of Justice of the State of Sergipe. The results showed that, among the analyzed algorithms, the RSLP possessed the greatest capacity of dimensionality reduction of the data. The RSLP-S and UniNE algorithms, the less aggressive stemmers, presented the best cost-benefit ratio, due to the fact that they reduced the dimensionality of the data and had similar efficacy or, in some cases, superior to the control group.

Keywords: Experimental software engineering, judicial documents, dimensionality reduction, jurisprudence.

Lista de ilustrações

Lista de tabelas

Lista de códigos

Lista de abreviaturas e siglas

AP	*Average Precision*
ASG	Acórdãos do Segundo Grau[1]
ATR	Acórdãos da Turma Recursal[2]
CSV	*Comma Separeted Values*
DSG	Decisões Monocráticas do Segundo Grau[3]
DTR	Decisões Monocráticas da Turma Recursal[4]
IR	*Information Retrieval*
MAP	*Mean Average Precision*
NIST	*National Institute of Standards and Technology*
OLTP	*Online Transaction Processing*
PR	Percentual de redução
Pr@10	Média da Precisão com corte no décimo resultado
RAM	*Random Access Memory*
RP	*R-Precision*
RSLP	*Removedor de Sufixos da Língua Portuguesa*
TJSE	Tribunal de Justiça do Estado de Sergipe
TREC	*Text Retrieval Conference*
TUD	Termos únicos por documento
XML	*Extensible Markup Language*

[1] Em inglês, esse termo foi traduzido como *Judgements of Appeals Court* (JAC)
[2] Em inglês, esse termo foi traduzido como *Judgements of Special Courts* (JSC)
[3] Em inglês, esse termo foi traduzido como *Monocratic Decisions of Appeals Court* (MAC)
[4] Em inglês, esse termo foi traduzido como *Monocratic Decisions of Special Courts* (MSC)

Sumário

1

Introdução

Todos os dias, os tribunais, por meio de seus magistrados, julgam os mais variados temas das esferas do Direito, gerando um grande corpo de conhecimento jurídico que norteia novas decisões e serve como base argumentativa para as partes envolvidas pleitearem seus interesses. Assim, a partir do corpus formado pelo conjunto de decisões uniformes proferidas pelo judiciário a respeito de um determinado assunto (MAXIMILIANO, 2011), emerge o conceito de jurisprudência, instrumento fundamental para que os profissionais do Direito possam exercer o seu papel. Para Santos (2001, p. 137), a jurisprudência é a "ciência do Direito e dos princípios de Direito seguidos num país, numa dada época ou em certa e determinada matéria legal".

Diante da necessidade de busca nessas bases jurisprudenciais, cada órgão do judiciário acaba desenvolvendo sua própria solução, tanto para recuperar quanto para exibir os resultados encontrados. Segundo Magalhães (2008), a maioria dessas ferramentas faz uso de palavras-chave, sem aplicar qualquer algoritmo de radicalização — em inglês, *stemming* — de tal forma que pesquisar por "fatal" retorna somente documentos que contenham exatamente esse termo, ignorando aqueles que possuam apenas "fatalidade", por exemplo. Assim, muitos documentos que poderiam ser relevantes para a necessidade do usuário, acabam não sendo achados durante as buscas.

A qualidade dos algoritmos de *stemming* pode ser mensurada de duas maneiras (FLORES; MOREIRA, 2016):

- Intrínseca: avalia se os radicais gerados pelos algoritmos mantêm os mesmos aspectos semânticos e morfológicos das palavras que os originaram; e
- Extrínseca: analisa o impacto da utilização dos algoritmos de radicalização no domínio da aplicação.

Desse modo, a avaliação extrínseca está ligada diretamente ao domínio no qual os algo-

ritmos serão aplicados, já que a mesma técnica pode apresentar impactos positivos ou negativos, a depender do contexto (WEISS et al., 2010). Apesar de tal importância, não foram encontrados na literatura estudos que avaliassem a eficácia da radicalização quando aplicados a bases jurídicas. O seguinte trecho ilustra esta carência:

> A MinerJur permite aos seus usuários optar entre o Porter Stemming ou Stemmer Portuguese (RSLP) para efetuar a normalização morfológica. No entanto, como o objetivo do estudo de caso não era analisar qual desses algoritmos apresentava melhor desempenho, foi necessário optar por um deles [...].(MAGALHÃES, 2008, p. 109).

Além disso, a autora sugere como possível contribuição para a evolução de sua dissertação, fazer um "estudo comparativo entre os algoritmos de normalização morfológica, assim como entre os algoritmos de comparação de *string*, a fim de analisar quais apresentam melhores resultados." (MAGALHÃES, 2008, p. 127).

Por isso, a proposta deste trabalho foi conduzir um processo experimental, seguindo as diretrizes de Wohlin et al. (2012), para realizar uma avaliação extrínseca dos efeitos da radicalização sobre a redução de dimensionalidade de bases jurisprudenciais, bem como analisar seu impacto sobre a eficácia de recuperação de documentos jurídicos.

1.1 Análise do Problema

Os algoritmos de radicalização podem ser utilizados na fase de pré-processamento textual em técnicas de mineração e aprendizagem de máquina como (ZHAI; MASSUNG, 2016): obtenção, classificação e clusterização de documentos, não se limitando a estas, com a finalidade de reduzir a dimensionalidade dos dados envolvidos (JAMES et al., 2013). No entanto, essa redução tem o potencial de impactar diretamente em métricas como precisão e revocação (*recall*) (ALVARES; GARCIA; FERRAZ, 2005).

No domínio legal, espera-se dos sistemas de recuperação um alto índice de revocação (KONTOSTATHIS; KULP, 2007), portanto, após uma busca, devem ser retornados todos os documentos que sejam relevantes para uma determinada necessidade de informação. Por outro lado, existem relatos na literatura que evidenciam o nível de subjetividade inerente ao julgamento de relevância para documentos obtidos por meio de um sistema de recuperação, haja vista que resultados experimentais mostram níveis de concordância que variam entre 28% e 50%, na média (CARTERETTE; VOORHEES, 2011; WEBBER, 2011).

Fora deste domínio, Flores e Moreira (2016) avaliaram o impacto dos algoritmos de radicalização na melhoria de sistemas de recuperação de documentos escritos em inglês, francês, espanhol e português. Para tal, aplicaram a radicalização sobre coleções de testes disponíveis para cada um dos idiomas escolhidos. No caso do português, utilizando a base da Folha de São Paulo, concluíram que os algoritmos reduziram em até 31,59% o número de termos distintos

no índice e pelo menos seis deles apresentaram um incremento estatisticamente significativo na MAP *(Média das Precisões Médias, em inglês, Mean Average Precision*).

Contudo, o universo jurídico possui jargão próprio, enfatiza a não repetição de palavras e faz uso de um vocabulário mais rebuscado. Desse modo, essas características podem ter influência direta sobre a efetividade dos algoritmos (CÂMARA JÚNIOR, 2007).

Assim, sabendo que em bases não jurídicas a aplicação da radicalização reduziu a dimensionalidade dos dados e, ainda, aumentou a relevância dos resultados retornados pelo sistema de buscas, o judiciário poderia beneficiar-se desses achados, caso houvesse evidências experimentais mostrando que o mesmo se aplica às bases judiciais. Além disso, essa redução de dimensionalidade contribuiria para aumentar a economia de recursos computacionais, uma vez que índices menores ocupam menos espaço em disco e na RAM, fazendo com que eles possam ser realocados para outras atividades relacionadas à prestação de serviço jurisdicional ao cidadão.

Nesse sentido, utilizamos neste trabalho a base jurisprudencial do Tribunal de Justiça do Estado de Sergipe, formada por quatro coleções de documentos: decisões monocráticas do Segundo Grau, acórdãos do Segundo Grau, decisões monocráticas da Turma Recursal e acórdãos da Turma Recursal (detalhadas na seção 2.1). Já para redução de dimensionalidade, fizemos uso dos algoritmos Porter, RSLP, RSLP-S e UniNE (explicados na seção 2.2).

Diante de tal cenário, foram elencadas as seguintes questões:

1. Q1: No contexto jurisprudencial, a aplicação de algoritmos de radicalização reduz de forma significativa a quantidade de termos únicos por documento?
2. Q2: A eficácia dos algoritmos de radicalização é a mesma em todas as coleções judiciais?
3. Q3: A radicalização tem efeito sobre os resultados obtidos mediante as buscas jurisprudenciais?

Para nortear o trabalho, sintetizamos essas três questões sob a hipótese de pesquisa segundo a qual, dentre os algoritmos Porter, RSLP, RSLP-S e UniNE, existe pelo menos um deles que possui uma redução de dimensionalidade estatisticamente significativa sem, contudo, causar prejuízo perante a obtenção de documentos jurisprudenciais.

Levando em consideração que este estudo tem por alicerce as diretrizes de Wohlin et al. (2012), esta questão de pesquisa será ponderada por meio de métricas (detalhadas na seção 2.3.1):

- Valor médio de termos únicos por documento (μ_{TUD});
- Média do percentual de redução de dimensionalidade de cada algoritmo na coleção (μ_{PR});

- Média da *Average Precision* (MAP);
- Média da Precisão com corte no décimo resultado (MPC(10));
- Média da *R-Precision* (MRP).

Assim, a hipótese de pesquisa poderá ser testada estatisticamente por meio das seguintes hipóteses experimentais:

- Hipótese 1 (Q1)
  - $H0^{TUD}$: A média de termos únicos por documento sem radicalização é igual à média de termos únicos por documento com radicalização, para cada um dos algoritmos analisados.
  - $H1^{TUD}$: A média de termos únicos por documento sem radicalização é maior que a média de termos únicos por documento com radicalização, para pelo menos um dos algoritmos analisados.
- Hipótese 2 (Q2)
  - $H0^{PR}$: As médias do percentual de redução de termos únicos por documento são iguais nas quatro coleções.
  - $H1^{PR}$: As médias do percentual de redução de termos únicos por documento não são iguais nas quatro coleções.
- Hipótese 3 (Q3)
  - $H0^{MAP}$: As MAPs com e sem radicalização são iguais para todos os algoritmos.
  - $H1^{MAP}$: As MAPs com e sem radicalização são diferentes, para pelo menos um dos algoritmos.
- Hipótese 4 (Q3)
  - $H0^{MPC(10)}$: A MPCs(10) com e sem radicalização são iguais para todos os algoritmos.
  - $H1^{MPC(10)}$: A MPCs(10) com e sem radicalização são diferentes, para pelo menos um dos algoritmos.
- Hipótese 5 (Q3)
  - $H0^{MRP}$: As MRPs com e sem radicalização são iguais para todos os algoritmos.
  - $H1^{MRP}$: As MRPs com e sem radicalização são diferentes, para pelo menos um dos algoritmos.

Nesse cenário, a hipótese nula é representada por $H0$ e a hipótese alternativa, sobre a qual ela será testada, é denominada $H1$.

1.2 Justificativa

Esta dissertação realizou uma avaliação experimental que permitiu mensurar a efetividade dos algoritmos de radicalização, quando aplicados ao contexto de buscas jurisprudenciais, no que tange à redução de termos indexados e ao reflexo desta sobre os resultados retornados pelas consultas. Assim, trabalhos posteriores poderão utilizar os resultados relatados aqui para escolher o *stemmer* mais adequado para o objetivo pretendido, uma vez que não encontramos análise semelhante aplicada ao universo jurídico.

1.3 Objetivos da Pesquisa

Esta seção descreve os objetivos geral e específicos pretendidos para a realização desta dissertação.

O objetivo deste projeto foi avaliar o uso de técnicas de radicalização sobre a base jurisprudencial do Tribunal de Justiça do Estado de Sergipe.

1.3.1 Objetivos Específicos

Os objetivos norteadores do presente trabalho foram os expostos abaixo:

- Gerar uma massa de teste com documentos jurisprudenciais;
- Mensurar o impacto da radicalização sobre a redução de dimensionalidade das bases;
- Avaliar a recuperação dos documentos radicalizados.

1.4 Método

O método adotado pelo trabalho envolveu, inicialmente, uma revisão da literatura, com abordagens sistemáticas (KITCHENHAM, 2004), tendo por finalidade encontrar pesquisas sobre algoritmos de radicalização aplicados à língua portuguesa no contexto de recuperação de informações — em inglês, *information retrieval* (IR).

Para operacionalizar a revisão, acessamos a base Scopus por meio do portal de periódicos da CAPES (`https://www.periodicos.capes.gov.br`), pois ele permite fazer download dos artigos sem restrições.

Nesse sentido, procuramos ser bem restritivos durante a busca, de forma que somente fossem retornados artigos que aplicaram a radicalização em documentos na língua portuguesa e

no contexto de sistemas de recuperação. Para tal, a *string* de busca utilizada durante a pesquisa foi:

TITLE-ABS-KEY (portuguese AND stem AND information AND retrieval) AND (LIMIT-TO (DOCTYPE , ”ar”))

Após a execução da busca, um único artigo foi encontrado e lido na íntegra. Nele, Flores e Moreira (2016) mensuraram o impacto da radicalização sobre sistemas de recuperação de documentos escritos em inglês, francês, espanhol e português. Como para a língua portuguesa eles utilizaram a base de notícias da Folha de São Paulo, ficamos interessados em fazer uma análise semelhante aplicada a bases jurisprudenciais.

Em relação ao processo, o trabalho seguiu um processo experimental, de acordo com as diretrizes de Wohlin et al. (2012), iniciando-se pela obtenção dos documentos relativos à jurisprudência judicial do Tribunal de Justiça do Estado de Sergipe (TJSE), haja vista que o autor do estudo é funcionário do referido órgão desde 2005 e exerce a função de Chefe da Divisão de Banco de Dados desde 2011, possuindo, assim, acesso aos arquivos que são indexados pelo sistema de busca atual.

Como dito anteriormente, o TJSE possui quatro coleções de documentos jurisprudenciais. Observando a Figura 1, podemos perceber que a quantidade de documentos de cada uma das coleções varia consideravelmente por ano. A base mais antiga, formada pelos acórdãos do Segundo Grau, vem sendo populada desde meados de 1994.

Em setembro de 2016, coletamos toda a base jurisprudencial indexada pelo atual sistema de buscas da instituição (Figura 2). Cada base é formada por XMLs, gerados a partir de uma base de dados transacional (Código 1).

Código 1 – Estrura XML dos documentos indexados pelo TJSE.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE gsafeed
PUBLIC "-//Google//DTD GSA Feeds//EN" "UTF-8">
<gsafeed>
    <header>
        <datasource>sg-acordaos</datasource>
        <feedtype>incremental</feedtype>
    </header>
    <group action="add">
        <add>
            <doc>
                <field name="NomeDoCampo1">
                    valorDoCampo1
                </field>
                <field name="NomeDoCampo2">
                    valorDoCampo2
                </field>
                <!-- Outros campos adicionados -->
                <field name="NomeDoCampoN">
                    valorDoCampoN
                </field>
            </doc>
        </add>
        <!-- Outros documentos -->
        <!--
        <add>
            <doc>
            ...
            </doc>
        </add>
        -->
    </group>
</gsafeed>
```

Nesse contexto, a Tabela 1 exibe o número de documentos (N), a média de caracteres por documento (μ) e o desvio padrão (σ) para cada uma das coleções que estão sendo analisadas pelo presente trabalho.

De posse desses dados, foi conduzido um experimento para avaliar o impacto dos algoritmos de radicalização [1] analisados por Flores e Moreira (2016) na redução da quantidade de termos únicos sobre os quatro corpus disponíveis (Q1 e Q2): a) acórdãos do Segundo Grau (ASG);

[1] Para este trabalho, foram somente considerados os algoritmos de radicalização baseados em regra que apresentaram um aumento estatisticamente significativo da MAP: Porter, RSLP, RSLP-S e UniNE.

b) decisões monocráticas do Segundo Grau (DSG); c) acórdãos da Turma Recursal (ATR); e d) decisões monocráticas da Turma Recursal (DTR). Para tal, adotou-se um nível de confiança de 95% ($\alpha = 0{,}05$) para comparação das médias de termos únicos por documento obtidas a partir da aplicação dos algoritmos de radicalização sobre uma amostra dos quatro tipos de jurisprudência disponíveis.

Tabela 1 – Propriedades das Coleções.

Coleção	N	μ	σ
ASG	181.994	11.625,59	8.270,06
DSG	37.044	8.416,16	6.935,75
ATR	37.161	9.507,41	5.718,97
DTR	23.149	6.567,67	4.009,46

Fonte: Elaborada pelo autor.

Figura 1 – Quantidade de documentos por período.

Fonte: Elaborada pelo autor.

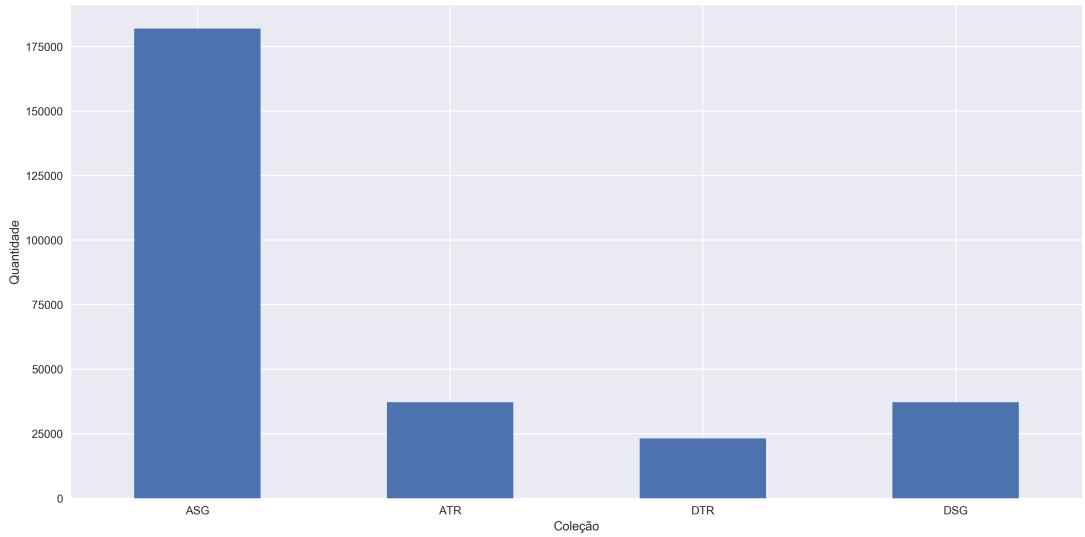

Figura 2 – Total de documentos por corpus.

Fonte: Elaborada pelo autor.

Em seguida, para responder a Q3, precisamos criar uma coleção de testes formada por três componentes (MANNING; RAGHAVAN; SCHÜTZE, 2009):

1. Um conjunto de documentos;
2. Uma lista de consultas que representem necessidades de informação; e
3. Julgamentos de relevância, geralmente binários (relevante ou não relevante), para cada par consulta-documento.

Contudo, a criação de tais coleções possui um custo elevado de tempo e dinheiro, podendo ficar em torno de milhões de dólares (ROITBLAT; KERSHAW; OOT, 2009), já que é preciso alocar especialistas tanto para elaborar as necessidades de informação quanto para julgar a relevância de cada par consulta-documento.

Apesar de não terem sido encontrados estudos que mostrem o custo relacionado à criação de coleções de testes jurisprudenciais, foi possível fazer uma estimativa.De acordo com o último edital, disponível em `https://goo.gl/NqYDuQ`, um magistrado em início de carreira recebe R$ 26.125,16. Dessa maneira, considerando a jornada de 6h por dia e 22 dias úteis por mês, um juiz do Tribunal de Justiça do Estado de Sergipe ganha R$ 197,92 por hora.

Messias et al. (2008) conduziram um experimento para mensurar quantos caracteres por minuto uma pessoa saudável consegue ler em média. Utilizando "[...] textos [...] de uma enciclopédia escrita para crianças entre 9 e 11 anos"(MESSIAS et al., 2008, p. 2), chegaram ao resultado de 1.100 caracteres por minuto. Textos jurídicos possuem uma complexidade linguís-

tica consideravelmente maior que textos infantis, mas, para fins de estimativa, consideremos este valor.

Observando a Tabela 1, cada acórdão do Segundo Grau possui em média 11.627 caracteres. Desse modo, um juiz levaria cerca de 10 minutos para ler cada documento. Considerando que para avaliação do segundo experimento utilizássemos 100 consultas, submetidas para sistemas indexados sem radicalização e indexados com os quatro algoritmos analisados, e fizéssemos uso de um *pool* com 100 documentos para efetuar o julgamento de relevância, um único magistrado teria que ler 10.000 documentos — 100 documentos para cada consulta —, ou seja, o equivalente a 1.667h de leitura, com o custo de R$ 329.929,10.

Diante desta questão, utilizamos amostras dos logs de buscas jurisprudenciais do TJSE para elaboração das necessidades de informação (consultas), tendo em vista que elas representam necessidades reais de usuários do sistema.

Já para o julgamento de relevância, Sakai e Lin (2010) realizaram um experimento comparando seis métodos, dois dos quais proposto por eles, que não necessitam de intervenção humana para avaliar o desempenho de sistemas de recuperação de informação. Apesar da simplicidade dos métodos propostos pelo autor, eles conseguiram uma precisão de cerca de 80% quando comparada à do julgamento realizado por seres humanos. Frente a esse resultado, resolvemos utilizar o método mais eficiente proposto pelos autores, o *nruns*, sobre a base de jurisprudência.

Dessa maneira, o julgamento de relevância deu-se da seguinte forma:

1. O mesmo conjunto de documentos foi indexado utilizando cada um dos algoritmos escolhidos;
2. Uma consulta foi disparada para cada uma das diferentes indexações;
3. Um *pool* de documentos foi composto pelos 30 primeiros resultados obtidos a partir de cada algoritmo;
4. Este *pool* foi ordenado em ordem decrescente pelo número de vezes que um mesmo documento apareceu nos resultados;
5. Os primeiros 30% foram marcados como relevantes;
6. O processo foi repetido para 100 consultas em cada uma das quatro coleções.

Assim, documentos mais populares foram marcados como relevantes, já que "sistemas que recebem documentos populares não são necessariamente bons; no entanto, os que não recebem são provavelmente ruins" (SAKAI; KANDO et al., 2008, p. 4, tradução nossa). De posse dos julgamentos de relevância, efetuou-se um experimento para mensurar o impacto da radicalização sobre as métricas MAP, MRP e MPC(10) dos resultados encontrados pelas consultas, respondendo a Q3.

1.5 Organização da Dissertação

As bases conceitual e experimental da presente dissertação serão fornecidas por meio de 5 capítulos, conforme descrição a seguir:

- O capítulo 1 apresentou a introdução;
- O capítulo 2 apresenta o referencial teórico referente ao detalhamento dos conceitos de jurisprudência, bem como dos algoritmos de radicalização e sistemas de buscas;
- O capítulo 3 descreve a condução do Experimento I, *Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence — An Experimental Analysis*, publicado no ICEIS 2017 (*The 19th International Conference on Enterprise Information Systems*). Tal previsão encontra-se disciplinada no §2 do artigo 1º da IN 02/2015/PROCC/UFS, que regulamenta a estrutura do documento da dissertação de Mestrado;
- No capítulo 4 é apresentado o processo experimental conduzido para realização do Experimento II, responsável por avaliar a recuperação dos documentos radicalizados;
- Por fim, o capítulo 5 apresenta as conclusões da dissertação.

2

Fundamentação Teórica

Este capítulo descreve uma visão geral dos principais conceitos pertinentes ao assunto tratado, objetivando dar um embasamento teórico para o mesmo. Assim, aqui são definidos jurisprudência, os algoritmos de radicalização e as métricas para avaliação de sistemas de recuperação de informação que serão adotadas para execução deste projeto.

2.1 Jurisprudência

As decisões proferidas pelos magistrados geram três tipos de documentos (SANTOS, 2001):

- Sentença: quando o juiz de direito profere um julgamento processual em primeira instância;
- Decisão Monocrática: quando um magistrado decide sozinho, em segunda instância, sobre uma causa que já possui interpretação uniforme;
- Acórdão: quando um órgão colegiado, formado por um relator e pelo menos dois magistrados, profere uma sentença em segunda instância.

Uma decisão de segunda instância pode ser fruto de um recurso interposto a partir de uma sentença proferida por um juiz de Primeiro Grau ou por um juiz dos Juizados Especiais, dando origem a documentos do Segundo Grau e das Turmas Recursais, respectivamente.

Diante de tamanha base documental, é imprescindível a adoção de técnicas que aumentem a eficácia do armazenamento e da busca de tais informações, pois, do contrário, há um prejuízo tanto de recursos computacionais quanto de acesso à Justiça, já que as partes interessadas podem não encontrar o documento de que precisam para pleitearem os seus direitos.

Nesse cenário, segundo a literatura (FLORES; MOREIRA, 2016; ORENGO; BURIOL; COELHO, 2007), os algoritmos de radicalização podem reduzir a dimensionalidade dos textos, melhorando, com isso, o uso dos recursos computacionais, e, ainda, podem aumentar a relevância dos resultados retornados pelos sistemas de buscas. No entanto, o universo jurídico possui características singulares — jargão próprio, baixa repetição de palavras, vocabulário rebuscado — e não foram encontrados relatos na literatura demonstrando que os mesmos benefícios são obtidos quando a radicalização é aplicada a bases jurisprudenciais.

2.2 Algoritmos de Radicalização

O processo de radicalização consiste em agrupar diferentes palavras em função de um radical (em inglês, *stem*) comum. A Tabela 2 mostra a aplicação dos cinco algoritmos de radicalização utilizados durante a realização do experimento sobre quatro palavras distintas, ressaltando que o NoStem é o grupo de controle, ou seja, ele não gera redução dos termos.

Tabela 2 – Exemplo da radicalização utilizando os cinco algoritmos do experimento.

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

Fonte: Elaborada pelo autor.

Com exceção do grupo de controle, os demais algoritmos utilizados no experimento são baseados em regras e atuam por meio da remoção de sufixos (FLORES; MOREIRA, 2016):

- Porter: escrito originalmente para o inglês, em 1980, e adaptado para o português, posteriormente (PORTER, 1980);
- RSLP: publicado em 2001, possui cerca de 200 regras e uma lista de exceções para quase cada uma delas (ORENGO; HUYCK, 2001);
- RSLP-S: uma versão enxuta do RSLP, que utiliza somente a redução de plural (ORENGO; BURIOL; COELHO, 2007);
- UniNE: possui menos regras que o Porter e RSLP, porém mais que o RSLP-S (FLORES; MOREIRA, 2016).

Por meio da Figura 3, podemos observar o funcionamento de um algoritmo de radicalização baseado em regra. Apesar de o conjunto de passos ser referente ao algoritmo RSLP, o

mais agressivo em termos de redução de dimensionalidade dos algoritmos analisados, a lógica dos demais *stemmers* estudados é similar, havendo somente variações nas regras.

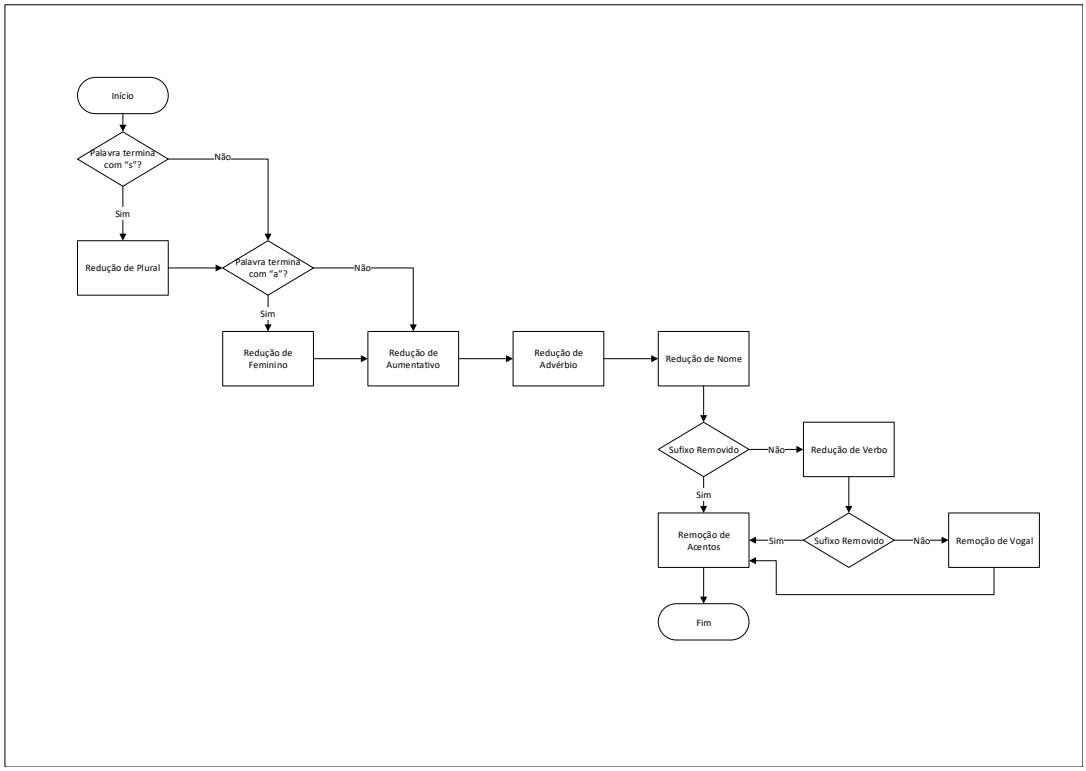

Figura 3 – Sequência de passos do algoritmo RSLP.

Fonte: Adaptado de Orengo, Buriol e Coelho (2007).

Assim, considerando a questão linguística, um algoritmo de radicalização pode cometer dois tipos de erros: a) *overstemming*, quando a parte removida não é um sufixo, mas parte do radical da palavra; e b) *understemming*, quando a remoção do sufixo não se dá de forma completa.

Neste estudo, a análise desses erros não contribuiria para o cálculo das métricas, uma vez que o algoritmo utilizado para efetuar o julgamento de relevância, o *nruns*, leva em consideração somente a quantidade de vezes em que um determinado documento aparece nos resultados das buscas, não importando os aspectos semânticos dos termos envolvidos.

2.3 Recuperação de Documentos Jurisprudenciais

Em virtude do número de documentos gerados diariamente pelos Tribunais, torna-se essencial mensurar a efetividade dos sistemas de recuperação que operam sobre esta base jurisprudencial. Nesta seção, forneceremos a base conceitual necessária para a compreensão das métricas avaliadas pelos experimentos conduzidos neste trabalho.

2.3.1 Métricas para Avaliação de Sistemas de Recuperação

Em meados da década de 60, foi publicado um trabalho com os resultados da avaliação experimental sobre uma coleção contendo 1400 documentos e 225 consultas na área de aerodinâmica, avaliando a efetividade de diferentes linguagens de indexação (CLEVERDON, 1967). Assim, para cada consulta, dividiu-se a coleção de documentos em dois grupos: relevantes e não relevantes.

Tabela 3 – Matrix de contingência 2 x 2.

	Relevante	Não Relevante	
Documento Encontrado	a	b	a + b
Documento Não Encontrado	c	d	c + d
	a + c	b + d	a + b + c + d = N

Fonte: Adaptado de Cleverdon (1967).

De acordo com a classificação do documento e se ele foi encontrado ou não pelo sistema de buscas, conforme ilustrado pela Tabela 2, foram calculadas três métricas: precisão (2.1), *recall* (2.2) e *fallout* (2.3).

$$\text{precisão} = 100\frac{a}{a+b} \tag{2.1}$$

$$recall = 100\frac{a}{a+c} \tag{2.2}$$

$$fallout = 100\frac{b}{b+d} \tag{2.3}$$

Ainda hoje, o método utilizado pelo trabalho de Cleverdon (1967) representa o alicerce sobre o qual a eficácia dos sistemas de recuperação é aferida (VOORHEES; HARMAN, 2005). No entanto, com o aumento do poder computacional ocorrido nas décadas seguintes, o volume de documentos das coleções digitais cresceu vertiginosamente, tornando inviável analisar, na íntegra, a relevância de milhões de documentos. Sendo assim, como comparar diferentes sistemas de buscas que atuam sobre estas grandes coleções?

Pensando nisso, em 1990, a Agência de Projetos de Pesquisas Avançadas (DARPA) iniciou uma parceria com o *National Institute of Standards and Technology* para construção de uma coleção de testes composta por milhões de documentos, centenas de vezes maior que as coleções não-proprietárias existentes na época, para avaliar o projeto TIPSTER (VOORHEES; HARMAN, 2005). No ano seguinte, essa coleção foi disponibilizada para a comunidade e deu origem à primeira *Text Retrieval Conference* (TREC), objetivando encorajar a pesquisa em sistemas de recuperação sobre grandes coleções.

Assim, especialistas elaboraram uma lista de tópicos, cada qual contendo um identificador único, descrevendo uma necessidade de informação e quais seriam os critérios de julgamento utilizados para considerar um determinado documento da coleção como sendo relevante. Essa lista de tópicos foi disponibilizada para os participantes da conferência e eles desenvolveram seus sistemas de buscas para atender àquelas necessidades de informação. Em seguida, os participantes enviaram uma lista dos documentos retornados por tópico para os avaliadores da conferência. Para efetuar o julgamento de relevância, o seguinte procedimento foi adotado (VOORHEES; HARMAN, 2005):

- Para cada conjunto de resultados de um tópico, os primeiros X documentos foram escolhidos para compor um *pool*;
- Os resultados de cada sistema foram combinados, ordenados pelo identificador do documento e as duplicações foram removidas.

O próximo passo foi fornecer essa lista de documentos para assessores fazerem o julgamento de relevância. Para manter a consistência da avaliação, cada tópico era julgado por um único assessor. Ao final desta etapa, foi gerada uma lista com cada documento que fez parte do *pool* marcado como relevante ou não. Todos os documentos que não fizeram parte do *pool*, ou seja, não julgados, foram considerados como irrelevantes.

Nesse cenário, levando-se em conta que não houve um julgamento de todos os documentos que poderiam ser relevantes para determinado tópico, foi preciso criar novas métricas que permitissem fazer a comparação entre os sistemas. Tais métricas foram implementadas e publicadas no utilitário *trec_val* (NIST, 2016).

Por fim, de posse dos resultados das consultas retornados pelo seu sistema e da lista contendo o julgamento de relevância, cada participante da conferência poderia, finalmente, fazer o cálculo das métricas e publicar os achados.

Considerando a relevância desta conferência para o campo de pesquisa em IR, utilizaremos três de suas métricas, exemplificadas a partir de uma lista de documentos hipotéticos retornados por um sistema de buscas (Tabela 4).

Tabela 4 – Relação fictícia de resultados retornados por um sistema de buscas com os respectivos julgamentos de relevância.

Documento	Relevante
d1	Não
d2	Sim
d3	Sim
d4	Não
d5	Não
d6	Sim
d7	Sim
d8	Sim
d9	Sim
d10	Não
d11	Sim
d12	Sim
d13	Não
d14	Não

Fonte: Elaborada pelo autor.

Assim, descreveremos as métricas Pr@10, *R-Precision* e *Average Precision*, a partir de uma coleção de documentos $C = (d1, d2, ..., dn)$, ordenada pelo valor da semelhança entre o documento d e a consulta c, obtido por meio de uma função de *ranking*[1] r, tal que $r(d_i, c) >= r(d_j, c)$ para $i < j$, e do julgamento de relevância $f(c, d)$, que retorna "Sim", valor 1, caso o documento d seja relevante para a consulta c e "Não", valor 0, do contrário.

Precision at Document Cutoff λ. Esta métrica tem por objetivo mensurar a precisão de um sistema considerando um certo ponto de corte (λ):

$$Pr@\lambda = \sum_{i=1}^{\lambda} \frac{f(c, d_i)}{\lambda} \tag{2.4}$$

Neste trabalho, adotaremos $\lambda = 10$, valor comumente usado como ponto de corte para esta métrica (ORENGO; BURIOL; COELHO, 2007; VOORHEES; HARMAN, 2005; FANG; TAO; ZHAI, 2011). Chamaremos de MPC(10) quando calcularmos a média de Pr@10 dos n tópicos (consultas).

Ex: Pr@10 = 6 / 10 = 0,6 = 60% (Pr@10 da Tabela 4).

R-Precision. Esta métrica calcula a precisão em R, onde R é o número de documentos relevantes:

$$RP = \sum_{i=1}^{R} \frac{f(c, d_i)}{R} \tag{2.5}$$

[1] Para conhecer o estado da arte em funções de ranking, consultar Zhai e Massung (2016, p. 87–128).

Enquanto a Pr@λ é útil para medir a eficácia de um sistema com respeito à aplicação, esta é mais adequada para fazer comparação entre sistemas (VOORHEES; HARMAN, 2005). Vale ressaltar que no ponto R a precisão e a revocação (*recall*) possuem o mesmo valor. Denominaremos de MRP quando calcularmos a média *R-Precision* das n consultas.

Ex: R-Precision (R = 8) = 5/8 = 0,62% (R-Precision da Tabela 4).

Average Precision. De posse da quantidade de documentos julgados relevantes (R), esta métrica calcula a precisão e o *recall* para cada posição dos n documentos retornados pela consulta:

$$AP = \sum_{i=1}^{n} \frac{Pr@n_i f(c, d_i)}{R} \tag{2.6}$$

O princípio é o de que quanto mais cedo um resultado relevante aparecer, melhor. Quando é calculada a média de AP entre diferentes tópicos, essa métrica é chamada de *Mean Average Precision* (MAP). Ex. AP = (1/2 + 2/3 + 3/6 + 4/7 + 5/8 + 6/9 + 7/11 + 8/12) / 8 = 0,6 = 60% (AP da Tabela 4).

3

Experimento: Redução de Dimensionalidade Jurisprudencial

As seções a seguir descrevem o processo experimental utilizado para avaliar a redução de dimensionalidade obtida por meio da radicalização das bases jurisprudenciais. Conforme disciplinado pelo §2 do artigo 1º da IN 02/2015/PROCC/UFS, o texto foi extraído do artigo escrito por Oliveira e Colaço Júnior (2017a), respectivamente, autor e orientador desta dissertação. Além deste artigo, os autores publicaram a versão resumida dos resultados em Oliveira e Colaço Júnior (2017b).

3.1 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.

3.1.1 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 (V. R. BASILI; CALDIERA; ROMBACH, 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.

3.1.2 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 (37,044 documents); c) judgments of Special Courts (37,161 documents); and d) monocratic decisions of Special Courts (23,149 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$^{\text{UTD}}$:** The stemming algorithms have the same average of unique terms per document ($\mu_{NoStem^{UTD}} = \mu_{Porter^{UTD}} = \mu_{RSLP^{UTD}} = \mu_{RSLP\text{-}S^{UTD}} = \mu_{UniNE^{UTD}}$).
- **Alternative Hypothesis H1$^{\text{UTD}}$:** 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 ($\mu_{i^{RP}} \neq \mu_{j^{RP}}$ 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 unique terms. So, the quantity of documents were determined by the sample calculation of a finite population:

$$n = \frac{z^2.\sigma^2.N}{e^2.(N-1) + z^2.\sigma^2} \tag{3.1}$$

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 5 shows the number of selected documents after sample calculation, along with size, mean and standard deviations of the population.

Tabela 5 – Sample size per collection.

Coll.	N	μ	σ	n
JAC	181,994	638.45	322.15	1,524
MAC	37,044	488.63	276.56	1,476
JSC	37,161	520.07	247.05	1,476
MSC	23,149	419.54	192.39	1,442

Experiment Project. The jurisprudential documents have a great variability in respect to the number of unique terms, 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.

3.2 Experiment Execution

3.2.1 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.

3.2.2 Execution

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

3.2.3 Data Collection

The application recorded, for each collection, the document identifier, the number of unique terms and the stemming algorithm adopted CSV format (Table 6).

Tabela 6 – Input example in CSV file.

ID,UTD,Stemmer
201100205001443632662,679,NoStem
201100205001443632662,580,Porter
201100205001443632662,547,RSLP
201100205001443632662,651,RSLPS
201100205001443632662,636,UniNE

3.2.4 Data Validation

The Java application was built using Test Driven Development (TDD) (AGARWAL; 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; 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).

3.3 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 4 and Figure 5, respectively.

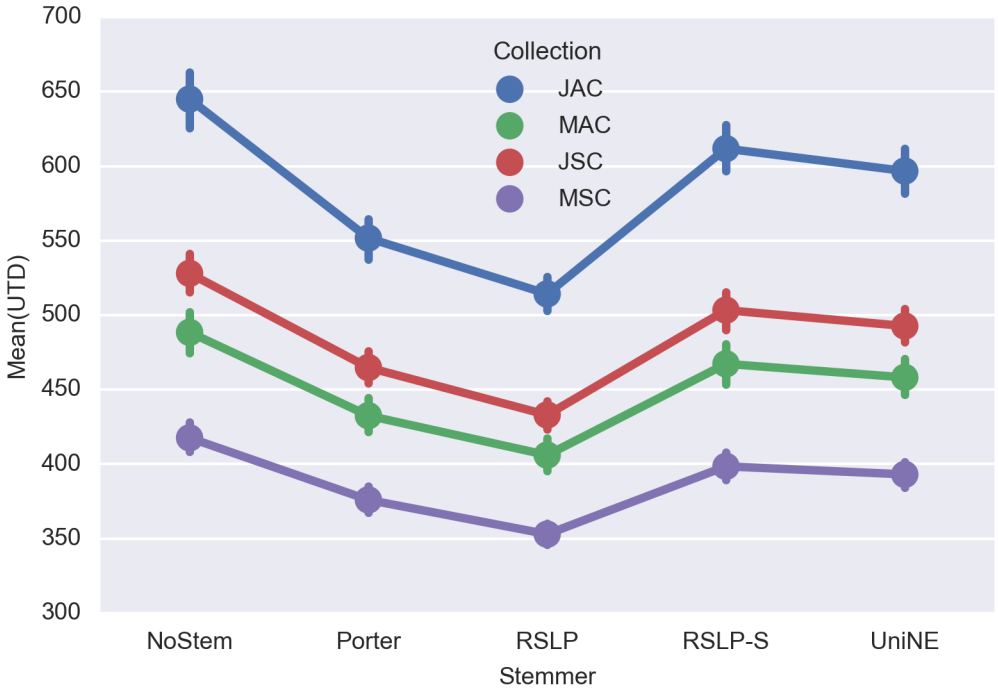

Figura 4 – The average number of unique terms per document obtained by each stemmer.

3.3.1 Analysis and Interpretation

Visually, analyzing Figures 4 and 5, a stemming application seems to generate differences in both, the average of reduction of unique terms per document and in the average percentage

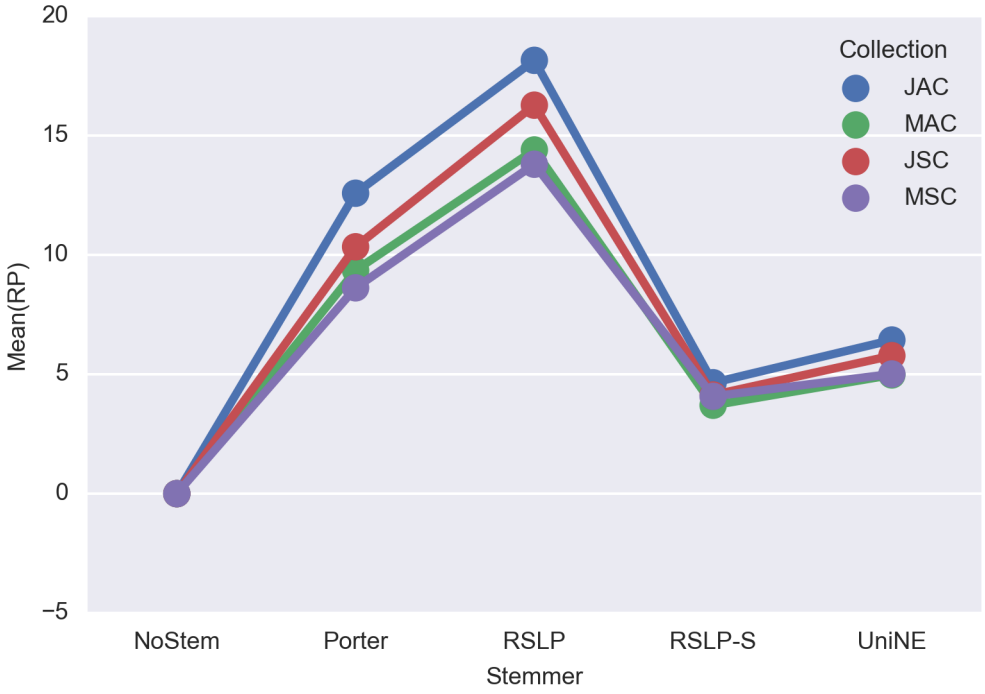

Figura 5 – The average percentage of dimensionality reduction per document generated by stemming.

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 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 7).

Tabela 7 – 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.

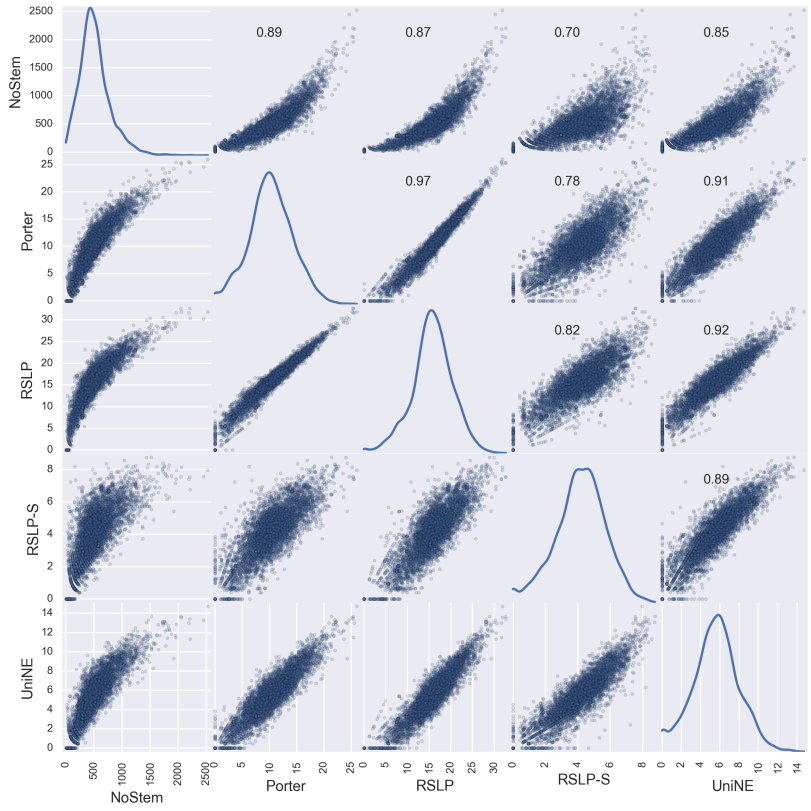

Figura 6 – Correlation matrix among stemming algorithms. NoStem unit is UTD and others are RP

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 8).

Tabela 8 – 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

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, 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 6, 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.

Tabela 9 – 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%

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.

In this scenario, shown in Table 9, 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.

3.3.2 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.

3.4 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.

4

Experimento: Radicalização X Recuperação de Documentos Jurisprudenciais

Neste capítulo, descrevemos a metodologia utilizada para realização do experimento, bem como os achados oriundos da análise dos resultados experimentais.

4.1 Definição e Planejamento do Experimento

Esta e as duas próximas seções relatam o processo experimental, de acordo com as diretrizes de Wohlin et al. (2012), utilizado para realização da análise de impacto da redução de dimensionalidade sobre a recuperação de documentos judiciais.

4.1.1 Definição do Objetivo

Conforme ilustrado pela Tabela 10, os documentos referentes aos acórdãos do Segundo Grau (ASG), decisões monocráticas do Segundo Grau (DSG), acórdãos da Turma Recursal (ATR) e decisões monocráticas da Turma Recursal (DTR) foram radicalizados utilizando os algoritmos Porter, RSLP, RSLP-S ,UniNE e NoStem, este último sendo o grupo de controle.

A coleção ASG, por exemplo, possui 408.336 termos únicos. Após a radicalização, esse número foi reduzido em 23%, 27%, 6% e 14%, respectivamente, por meio dos algoritmos Porter, RSLP, RSLP-S e UniNE.

Assim, o objetivo deste experimento é analisar o impacto desta redução de dimensionalidade, obtida por meio da radicalização, sobre a recuperação de documentos jurisprudenciais.

Para atingi-lo, conduziremos um experimento, em ambiente controlado, no qual serão calculadas as médias das métricas *Average Precision* (MAP), *R-Precision* (MRP) e *Precision at Document Cutoff 10* (MPC(10)) das informações recuperadas a partir da indexação dos documentos radicalizados. Nesse sentido, será verificado se há uma diferença estatisticamente significativa entre a recuperação de documentos radicalizados e o grupo de controle.

Segue a formalização do objetivo, segundo o modelo GQM (*Goal Question Metric*) proposto por V. Basili et al. (2014):

- **Analisar** sistemas de recuperação
- **com o propósito de** avaliá-los
- **com respeito às** métricas MAP, MRP e MPC(10)
- **do ponto de vista de** analistas de dados
- **no contexto de** documentos jurisprudenciais.

Tabela 10 – Redução de dimensionalidade nas coleções.

Coleção	NoStem	Porter	RSLP	RSLP-S	UniNE
ASG	408.336 (0%)	316.008 (23%)	295.822 (27%)	384.393 (6%)	350.679 (14%)
DSG	145.270 (0%)	110.378 (24%)	104.082 (28%)	135.851 (6%)	124.661 (14%)
ATR	188.266 (0%)	151.139 (20%)	144.213 (23%)	178.675 (5%)	165.836 (12%)
DTR	54.862 (0%)	39.640 (28%)	36.833 (33%)	50.944 (7%)	45.897 (16%)

Fonte: Elaborada pelo autor.

4.1.2 Planejamento

Seleção de Contexto. O experimento foi *in vitro* e utilizamos toda a base de documentos jurisprudenciais do Tribunal de Justiça do Estado de Sergipe[1], formada por quatro coleções: a) acórdãos do Segundo Grau (181.994 documentos); b) decisões monocráticas do Segundo Grau (37.044 documentos); c) acórdãos da Turma Recursal (37.161 documentos); e d) decisões monocráticas da Turma Recursal (23.149 documentos).

A Figura 7 exibe as variáveis dependentes e independentes envolvidas no experimento, detalhadas a seguir.

Variáveis independentes. Conjunto de consultas submetidas pelos usuários e registradas pelos logs do sistema de buscas do TJSE para cada uma das coleções (1). O motor de buscas com 20 bases indexadas (quatro coleções x cinco algoritmos) (2). Os resultados da submissão dessas consultas ao motor de buscas para cada uma das bases indexadas (3). Os julgamento de relevância desses resultados utilizando o algoritmo *nruns* (SAKAI; LIN, 2010) (4).

Variáveis dependentes. MAP, MRP e MPC(10), conceituadas no Capítulo 2 (5).

[1] Base coletada em setembro de 2016.

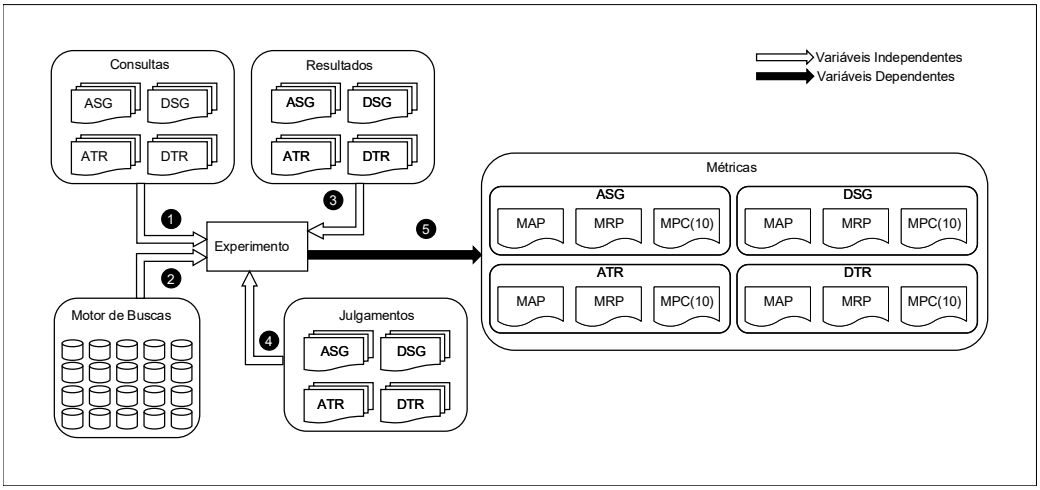

Figura 7 – Variáveis dependentes e independentes do experimento.

Fonte: Elaborada pelo autor.

Formulação de Hipótese. A questão de pesquisa para esse experimento é: a redução de dimensionalidade degrada a recuperação de documentos jurisprudenciais?

Para responder essa questão de pesquisa, analisaremos, por coleção, as métricas AP, RP e Pr@10. Portanto, as seguintes hipóteses serão verificadas:

Hipótese 1

- $H0^{MAP}$: As MAPs com e sem radicalização são iguais para todos os algoritmos ($NoStem^{MAP} = Porter^{MAP} = RSLP^{MAP} = RSLP\text{-}S^{MAP} = UniNE^{MAP}$).
- $H1^{MAP}$: As MAPs com e sem radicalização são diferentes, para pelo menos um dos algoritmos ($i^{MAP} \neq j^{MAP}$ para pelo menos um par (i,j)).

Hipótese 2

- $H0^{MPC(10)}$: As MPCs(10) com e sem radicalização são iguais para todos os algoritmos ($NoStem^{MPC(10)} = Porter^{MPC(10)} = RSLP^{MPC(10)} = RSLP\text{-}S^{MPC(10)} = UniNE^{MPC(10)}$).
- $H1^{MPC(10)}$: As MPCs(10) com e sem radicalização são diferentes, para pelo menos um dos algoritmos ($i^{MPC(10)} \neq j^{MPC(10)}$ para pelo menos um par (i,j)).

Hipótese 3

- $H0^{MRP}$: As MRPs com e sem radicalização são iguais para todos os algoritmos ($NoStem^{MRP} = Porter^{MRP} = RSLP^{MRP} = RSLP\text{-}S^{MRP} = UniNE^{MRP}$).
- $H1^{MRP}$: As MRPs com e sem radicalização são diferentes, para pelo menos um dos algoritmos ($i^{MRP} \neq j^{MRP}$ para pelo menos um par (i,j)).

Seleção de participantes e objetos. As consultas utilizadas durante o experimento foram obtidas por meio dos logs do sistema de consulta jurisprudencial do TJSE. Para cada uma das coleções, selecionaremos, aleatoriamente, 100 consultas, número considerado suficiente pela literatura para avaliação de sistemas de recuperação (MANNING; RAGHAVAN; SCHÜTZE, 2009; VOORHEES; HARMAN, 2005)

Projeto do experimento. Para assegurar a confiança dos testes de hipótese, foi adotado um modelo de blocos aleatórios — do inglês *randomized complete block design (RCBD)* (WOHLIN et al., 2012) —, assim, cada consulta foi escolhida de forma aleatória e submetida ao motor de buscas para cada versão radicalizada das coleções. Ressaltamos que o NoStem representa os termos do documento sem radicalização, portanto, ele atua como grupo de controle.

Instrumentação. O processo de instrumentação teve início com a preparação do ambiente para o experimento e o planejamento de coleta de dados. Utilizaremos o Apache Solr (ASF, 2011), versão 6.1.0, como motor de buscas para os documentos jurisprudenciais[2]. Em seguida, fizemos a indexação de toda a base jurisprudencial do TJSE, utilizando os algoritmos de radicalização e o grupo de controle. Por fim, foi desenvolvida uma aplicação Java para iterar sobre as consultas, submetê-las ao motor de buscas e calcular as métricas em função do julgamento de relevância proposto por Sakai e Lin (2010), registrando-as em arquivos CSV (*Comma Separated Values*).

4.2 Execução do Experimento

4.2.1 Preparação

A fase de preparação consistiu na obtenção das coleções referentes à jurisprudência judicial e dos logs do sistema de buscas do TJSE. Assim, a base documental foi a mesma utilizada em (OLIVEIRA; COLAÇO JÚNIOR, 2017a), arquivos XML (*eXtensible Markup Language*) extraídos a partir da base OLTP (*Online Transaction Processing*). De forma similar, os logs do sistema de buscas foram consolidados em um único arquivo e as consultas, selecionadas aleatoriamente, foram colocadas em arquivos XML, organizados por coleção.

4.2.2 Execução

Ao final das etapas anteriores, deu-se início ao experimento, com a execução da aplicação Java, seguindo o que foi definido no planejamento.

[2] Esta versão do Apache Solr utiliza o Okapi BM25 (ROBERTSON; ZARAGOVA, 2009) como métrica de similaridade para *rankear* os documentos em resposta às consultas.

4.2.3 Coleta de Dados

A aplicação registrou, para cada uma das bases indexadas pelo motor de buscas, os documentos retornados pela submissão das consultas, os julgamentos de relevância e o resultado das métricas obtidas. Ressaltamos que os registros do retorno das consultas e dos julgamentos foram realizados fazendo-se uso do mesmo formato do *trec_eval* (NIST, 2016), utilitário padrão da *Text REtrieval Conference* (`http://trec.nist.gov/`) para avaliação de sistemas de recuperação.

4.2.4 Validação dos Dados

A aplicação Java foi construída utilizando-se a abordagem de desenvolvimento dirigido a testes — do inglês *Test Driven Development* (TDD) (AGARWAL; DEEP, 2014) —, portanto, foram escritos casos de testes unitários para validar se o cálculo das métricas estavam de acordo com os padrões adotados por Voorhees e Harman (2005). Desse modo, os testes mostraram que os cálculos feitos pela aplicação estavam em conformidade com os elaborados pelo utilitário *trec_eval*.

Para assegurar a análise, interpretação e validação dos resultados, executamos a técnica de *bootstrap* (JAMES et al., 2013), visualizamos a distribuição dos dados conforme sugerido por Kitchenham et al. (2016) e utilizamos quatro testes estatísticos: teste de Shapiro-Wilk, teste de Levene, teste de Kruskal-Wallis e o teste de Mann-Whitney.

Utilizamos o *bootstrap* com mil iterações sobre o resultado das métricas, gerando, dessa forma, médias mais consistentes, pois levam em consideração as métricas que seriam obtidas repetindo-se o experimento mil vezes sobre diferentes resultados das consultas. Com o intuito de escolher o teste estatístico mais adequado para comparar os radicalizadores, utilizamos o teste de Shapiro-Wilk para verificar a normalidade das amostras, condição requerida para execução de testes paramétricos, já que a literatura mostra que ele possui um poder de teste superior ao das demais abordagens (AHAD et al., 2011; RAZALI; WAH, 2011). Fizemos uso do teste de Levene (LEVENE et al., 1960) para verificar a igualdade da variância (homocedasticidade) entre os grupos. Considerando o projeto RCBD do experimento, com um fator e múltiplos tratamentos, e que não houve homogeneidade na variância e na distribuição dos dados, executamos o teste de Kruskal-Wallis, não paramétrico, para validar as hipóteses experimentais. Por fim, conduzimos uma análise *post hoc* utilizando o teste de Mann-Whitney para comparar a diferença entre as médias de cada tratamento, aplicando a correção de Benferroni sobre o nível de significância (HOLM, 1979), reduzindo a possibilidade de rejeitarmos a hipótese nula quando ela for, de fato, verdadeira (Erro Tipo I) (DUNN, 1961).

Considerando ainda que "um resultado pode ser estatisticamente significativo e não ter relevância, sendo de reter que a substancialidade não se esgota nos valores de p obtidos" (LOUREIRO; GAMEIRO, 2011, p. 153), utilizamos o índice *Cohen's d* (COHEN, 1992) para mostrar a magnitude dos efeitos — em inglês, *effect size* — encontrados e seus respectivos intervalos de

confiança. Segundo Ellis (2010), existem, pelo menos, três razões para reportar o tamanho do efeito:

> First, doing so facilitates the interpretation of the practical significance of a study's findings. [...]. Second, expectations regarding the size of effects can be used to inform decisions about how many subjects or data points are needed in a study. [...] Third, effect sizes can be used to compare the results of studies done in different settings. [...]. (ELLIS, 2010, p. 24).

Usamos a linguagem R (TEAM, 2008) para execução de todos os testes estatísticos.

4.3 Resultados

Para responder a questão experimental, os arquivos CSV gerados pela aplicação Java foram analisados. Os resultados do impacto da radicalização sobre as métricas relevantes à recuperação de documentos jurisprudenciais estão descritos pela Tabela 11. Assim, podemos visualizar a eficácia dos algoritmos de radicalização agrupados por métrica e coleção. Além disso, a coluna *%* mostra a diferença percentual entre o tratamento e o grupo de controle, e a coluna $|d|$ exibe o índice *Cohen's d* com seu respectivo intervalo de confiança.

Tabela 11 – Métricas obtidas após avaliação dos algoritmos sobre as coleções.

Coleção	Métrica	Algoritmo	Valor	%	|d|	Coleção	Métrica	Algoritmo	Valor	%	|d|
ASG	MAP	NoStem	0,84	-	-	ATR	MAP	NoStem	0,77	-	-
		Porter	0,76	(-)9	3,07±0,13			Porter	0,74	(-)4	1,09±0,09
		RSLP	0,64	(-)24	8,29±0,27			RSLP	0,72	(-)6	1,79±0,10
		RSLP-S	0,81	(-)3	0,97±0,09			**RSLP-S**	**0,80**	**(+)4**	**0,90±0,09**
		UniNE	0,79	(-)6	2,29±0,11			**UniNE**	**0,82**	**(+)6**	**2,11±0,11**
	MPC	NoStem	0,84	-	-		MPC	NoStem	0,71	-	-
		Porter	0,72	(-)14	4,52±0,17			Porter	0,65	(-)8	1,89±0,11
		RSLP	0,57	(-)32	10,36±0,33			RSLP	0,61	(-)14	3,34±0,14
		RSLP-S	0,79	(-)6	1,75±0.10			**RSLP-S**	**0,72**	**(+)1**	**0,28±0,09**
		UniNE	0,74	(-)11	3,62±0,14			**UniNE**	**0,74**	**(+)4**	**0,92±0,09**
	MRP	NoStem	0,81	-	-		MRP	NoStem	0,76	-	-
		Porter	0,72	(-)11	3,87±0,15			Porter	0,69	(-)9	2,44±0,12
		RSLP	0,58	(-)28	9,86±0,32			RSLP	0,66	(-)13	3,69±0,14
		RSLP-S	0,78	(-)4	1,32±0.10			**RSLP-S**	**0,76**	**0**	**0,02±0,09**
		UniNE	0,74	(-)9	3,03±0,13			**UniNE**	**0,77**	**(+1)**	**0,41±0,09**
DSG	MAP	NoStem	0,90	-	-	DTR	MAP	NoStem	0,87	-	-
		Porter	0,80	(-)11	4,08±0,15			Porter	0,86	(-)1	0,40±0,09
		RSLP	0,71	(-)21	7,32±0,24			RSLP	0,77	(-)11	4.56±0,17
		RSLP-S	0,86	(-)4	1,63±0,10			**RSLP-S**	**0,87**	**0**	**0,11±0,09**
		UniNE	0,88	(-)2	0,70±0,09			UniNE	0,85	(-)2	0,85±0,09
	MPC	NoStem	0,79	-	-		MPC	NoStem	0,81	-	-
		Porter	0,70	(-)11	3,01±0,13			Porter	0,78	(-)4	1,32±0,10
		RSLP	0,61	(-)23	6,16±0,21			RSLP	0,67	(-)17	4,97±0,18
		RSLP-S	0,75	(-)5	1,28±0,10			RSLP-S	0,79	(-)2	0,99±0,09
		UniNE	0,77	(-)2	0,66±0,09			UniNE	0,77	(-)5	1,78±0,10
	MRP	NoStem	0,88	-	-		MRP	NoStem	0,86	-	-
		Porter	0,73	(-)17	6,23±0,21			Porter	0,82	(-)5	1,88±0,11
		RSLP	0,65	(-)26	9,22±0,03			RSLP	0,70	(-)19	6,27±0,21
		RSLP-S	0,83	(-)6	2,47±0,12			RSLP-S	0,82	(-)5	1,77±0,10
		UniNE	0,84	(-)4	2,01±0,11			UniNE	0,80	(-)7	2,62±0,12

Fonte: Elaborada pelo autor.

4.3.1 Análise e Interpretação

Olhando a Tabela 11, o uso da radicalização parece ter um impacto positivo somente nos acórdãos da Turma Recursal, com os algoritmos RSLP-S e UniNE causando um aumento das métricas em relação ao grupo de controle.

Contudo, é preciso que analisemos esta tabela sob a luz da estatística para encontrarmos evidências que corroborem, ou não, as aparentes diferenças descritas. Para tal, adotamos um nível de confiança de 95% ($\alpha = 0,05$) para todo o experimento.

Para melhorar a compreensão, separamos a análise por tipo de documento, facilitando a visualização do impacto da redução de dimensionalidade sobre as métricas estudadas.

4.3.1.1 Acórdãos do Segundo Grau

Iniciamos pela análise da normalidade dos dados da métrica MAP. Observando os gráficos, Figura 8, os cinco tratamentos parecem ter distribuição normal, uma vez que apresentam a maior parte dos valores ao redor da média (formato de sino). Além disso, os gráficos de probabilidade na parte inferior da mesma figura, mostram que os dados, representados pelos pontos, estão quase todos sobre a reta esperada para uma distribuição normal.

No entanto, a hipótese de normalidade dos dados foi rejeitada, pois o teste de Shapiro-Wilk apresentou um *p-value* inferior a 0,001 para o tratamento RSLP-S, abaixo, portanto, do nível de significância adotado para o experimento.

Em seguida, conduzimos o teste de Levene para validar a hipótese nula de homocedasticidade (igualdade das variâncias) entre os grupos. Contudo, essa hipótese foi rejeitada (*p-value* < 0,001).

Como nem todos os tratamentos apresentaram distribuição normal e há heterocedasticidade, conduzimos o teste de Kruskal-Wallis para validar a hipótese 1, igualdade da MAP entre os tratamentos ($H0^{AP}$). Uma vez conduzido, o teste mostrou evidência de diferença entre os algoritmos (*p-value* < 0.001).

Para visualizar essas diferenças, elaboramos a Figura 9 e conduzimos uma análise *post hoc* com testes de Mann-Whitney, aplicando o ajuste de Benferroni ($\alpha = \alpha/10$). Na figura, o ponto azul destaca o grupo de controle e as linhas verticais representam o intervalo de confiança. Assim, tanto por meio do gráfico quanto dos testes conduzidos, uma vez que as comparações entre os tratamentos apresentaram *p-value* inferior ao nível de significância adotado, foi possível evidenciar que os algoritmos de radicalização degradaram a obtenção de documentos com respeito à MAP.

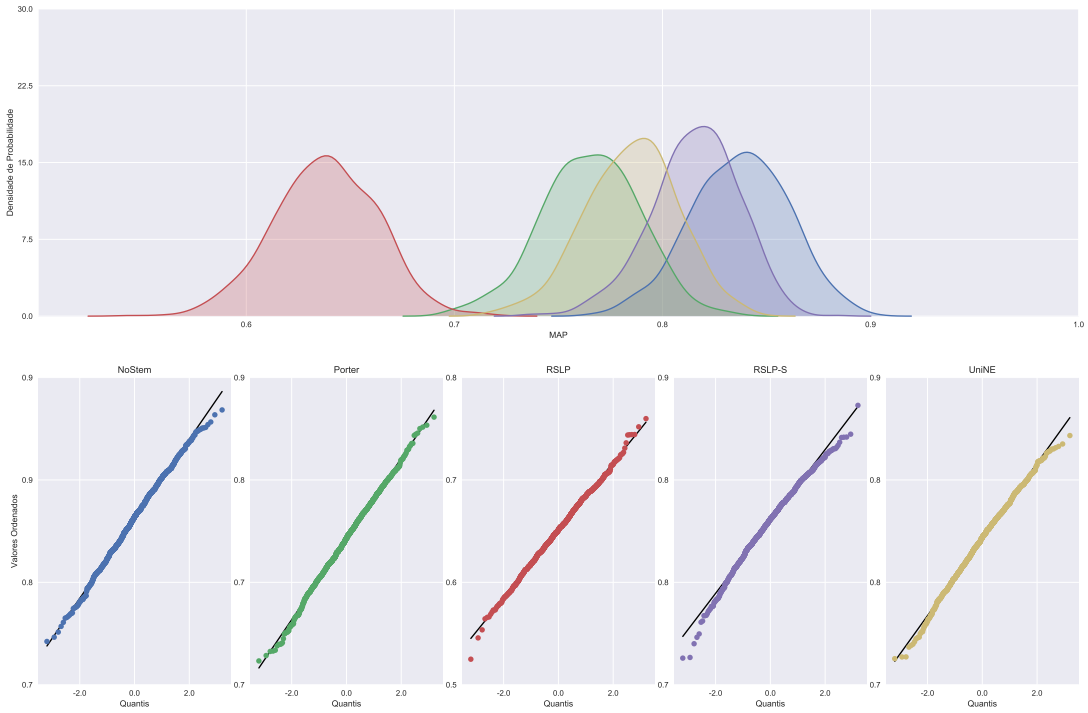

Figura 8 – Distribuição e gráficos de normalidade da métrica MAP dos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

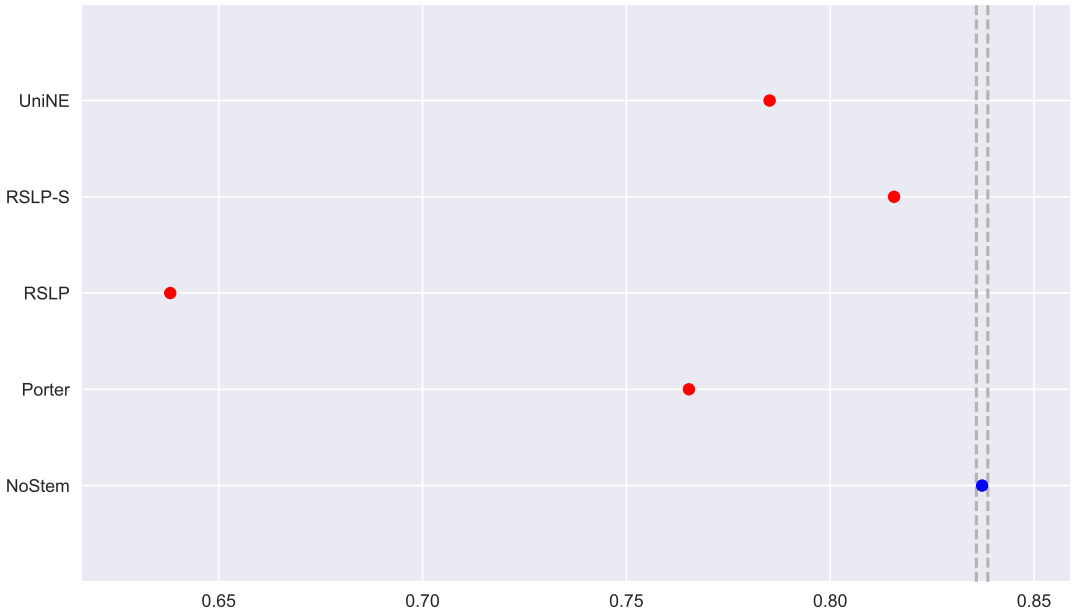

Figura 9 – Comparação da MAP nos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

Dando continuidade à análise das métricas, fizemos uma estudo da distribuição dos da-

dos referentes à MPC. Observando a Figura 10, temos que as distribuições parecem normais e a execução do teste de Shapiro-Wilk não rejeitou essa hipótese, exceto pelo algoritmo UniNE, pois ele ficou abaixo do nível de significância (*p-value* = 0,027).

O teste de Levene evidenciou heterocedasticidade dos dados (*p-value* < 0,001) e o teste de Kruskal-Wallis refutou a hipótese de igualdade da MPC entre os grupos ($H0^{Pr@10}$).

Após análise visual da métrica para cada tratamento (Figura 11) e estudo da significância estatística dessas diferenças usando o teste de Mann-Whitney, com todas as comparações apresentando *p-value* abaixo de 0,001, chegamos à conclusão de que a MPC, assim como a MAP, foi afetada negativamente pelo uso da radicalização.

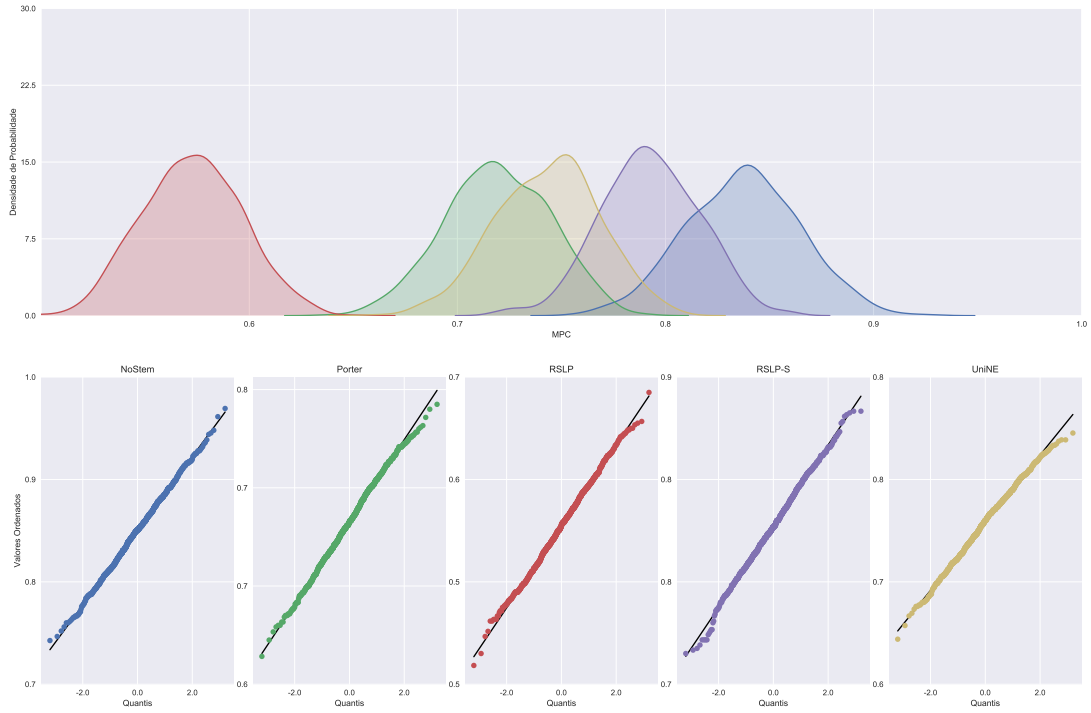

Figura 10 – Distribuição e gráficos de normalidade da métrica MPC dos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

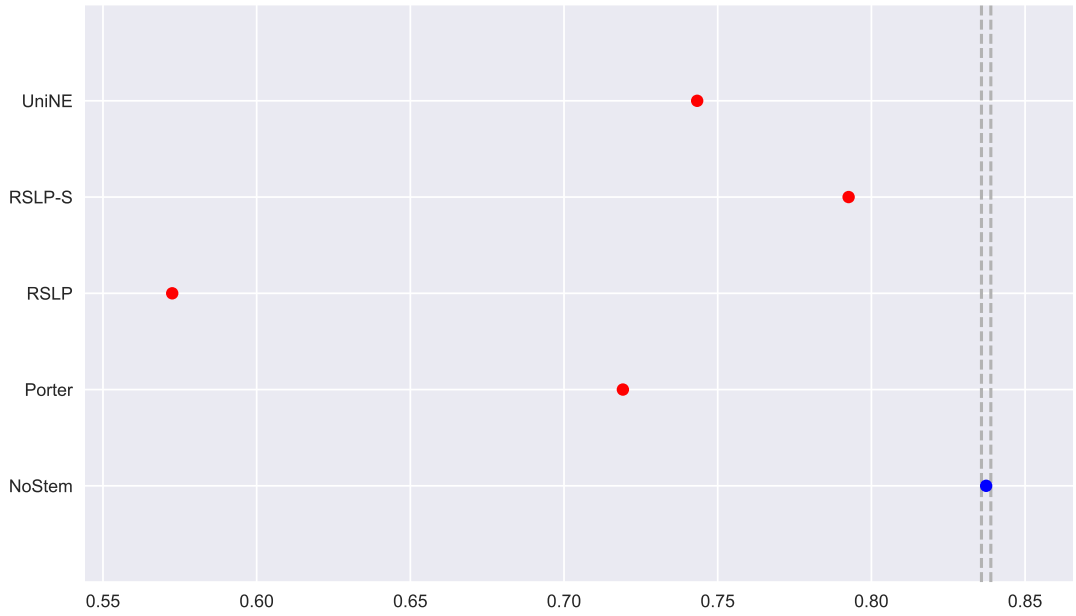

Figura 11 – Comparação da MPC nos Acórdão do Segundo Grau.

Fonte: Elaborada pelo autor.

Por fim, a distribuição dos dados da métrica MRP (Figura 10) comportou-se de forma análoga à da MPC, inclusive com o mesmo desvio de normalidade do algoritmo UniNE detectado pelo teste de Shapiro-Wilk (*p-value* < 0,001) e heterocedasticidade encontrada pelo teste de Levene (*p-value* < 0,001).

Assim como nas outras duas métricas, o teste de Kruskal-Wallis refutou a igualdade da MRP entre os algoritmos ($H0^{RP}$) e a análise *post hoc* com Mann-Whitney evidenciou que a diferença entre os tratamentos, ilustrada pela Figura 13, e o grupo de controle foi estatisticamente significativa (*p-value* < 0, 001).

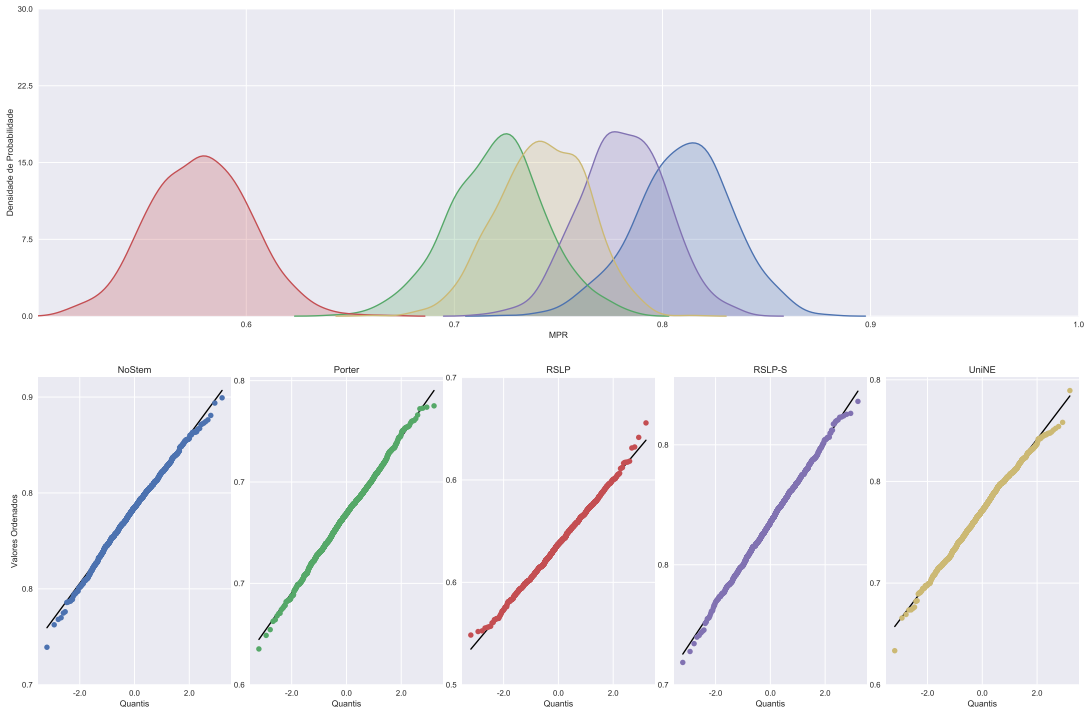

Figura 12 – Distribuição e gráficos de normalidade da métrica MRP dos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

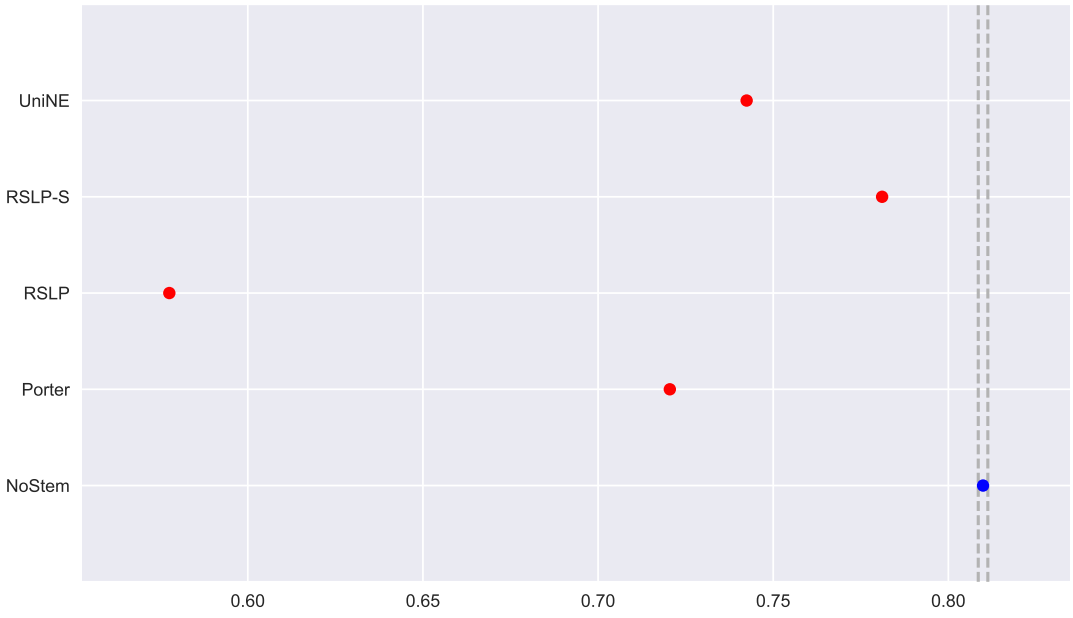

Figura 13 – Comparação da MRP nos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

O raio da Figura 14 mostra o valor percentual das métricas PR (redução de dimensi-

onalidade), MAP, MPC e MRP. Assim, conseguimos perceber, por exemplo, que apesar de o algoritmo RSLP reduzir em maior grau a dimensionalidade dos dados, ele destaca-se em relação aos demais pela diminuição da eficácia na recuperação de documentos.

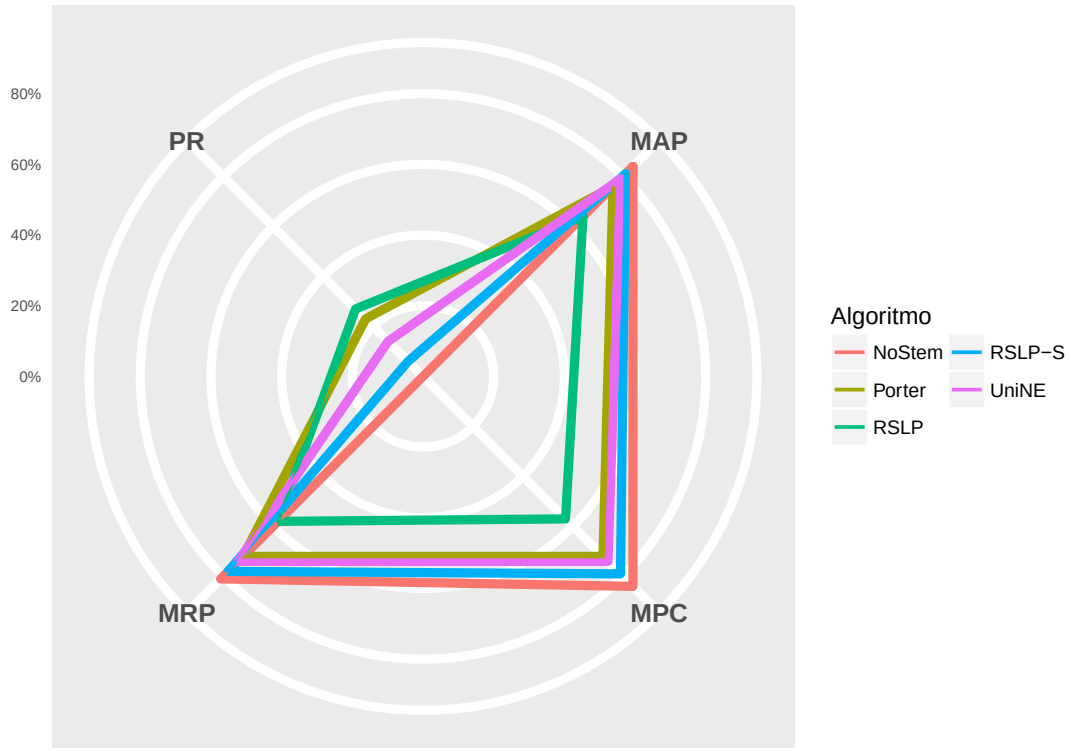

Figura 14 – Percentual de Redução (PR), MAP, MPC e MRP dos Acórdãos do Segundo Grau.

Fonte: Elaborada pelo autor.

4.3.1.2 Decisões Monocráticas do Segundo Grau

Nesta e nas duas próximas subseções, descreveremos os resultados encontrados de forma mais direta, tendo em vista que utilizamos o mesmo processo de análise descrito pelo tópico anterior.

Quanto à distribuição dos dados referentes à MAP (Figura 15), o teste de Shapiro-Wilk refutou a hipótese de normalidade dos algoritmos NoStem (*p-value* < 0,001) e Porter (*p-value* = 0,049). Na sequência, o teste de Levene refutou a hipótese de homocedasticidade entre os grupos (*p-value* < 0,001) e a igualdade da MAP, hipótese $H0^{AP}$, foi rejeitada pelo teste de Kruskal-Wallis com *p-value* inferior a 0,001. As diferenças ilustradas pela Figura 16 mostraram-se estatisticamente significativas (*p-value* < 0,001), por meio de uma análise *post hoc* com o teste de Mann-Whitney.

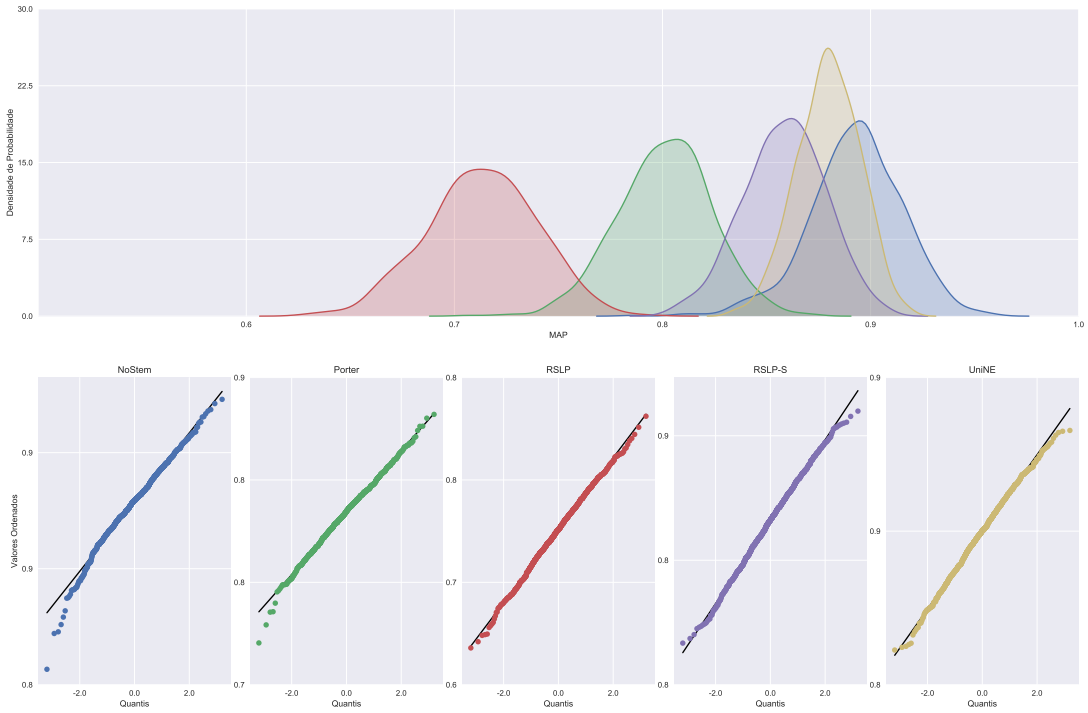

Figura 15 – Distribuição e gráficos de normalidade da métrica MAP das Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

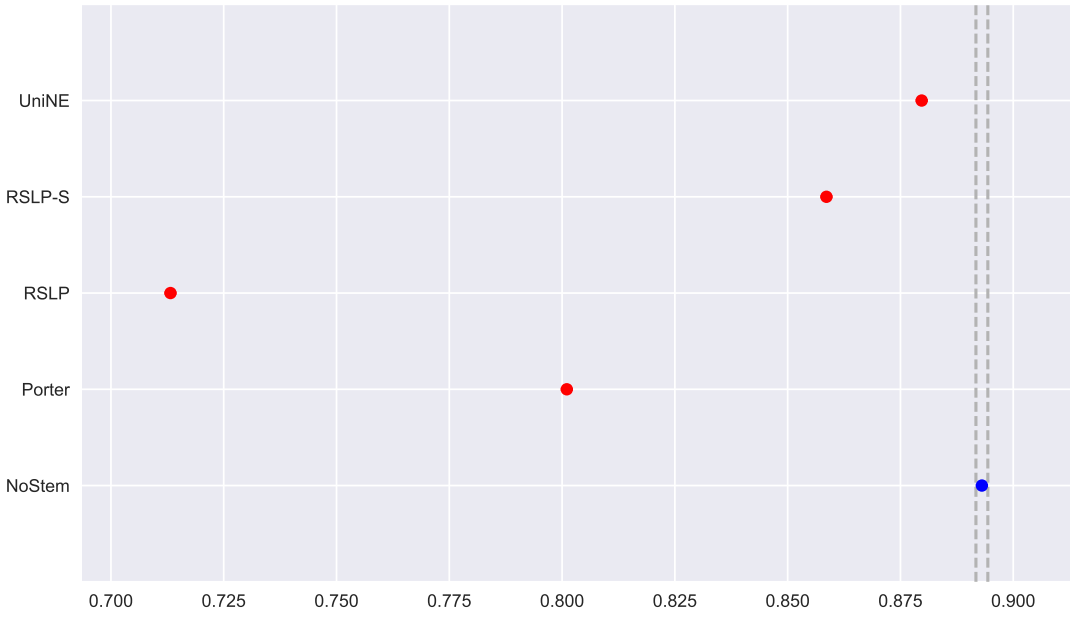

Figura 16 – Comparação da MAP nas Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

Por outro lado, a premissa de normalidade da MPC, Figura 17, foi violada pelos algo-

ritmos Porter e UniNE, com *p-values* iguais a 0,017 e 0,006, respectivamente. Além disso, confirmamos a heterocedasticidade dos dados e a significância das diferenças entre os tratamentos (Figura 18), refutando a hipótese $H0^{Pr@10}$, haja vista que os testes encontraram *p-value* inferior a 0,001.

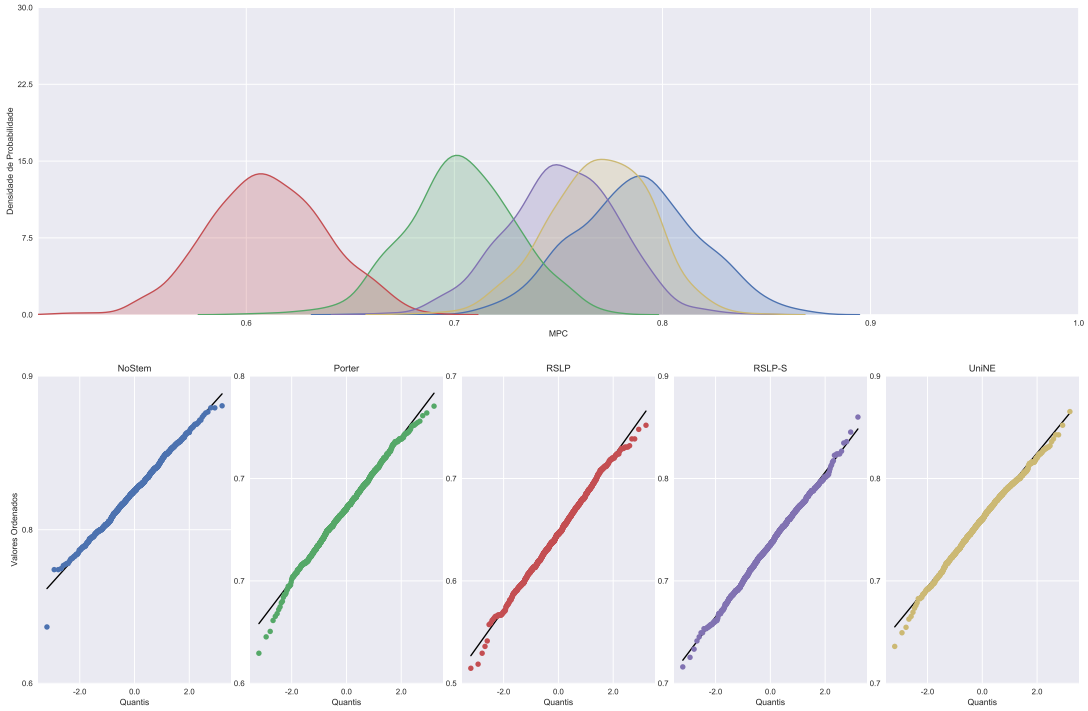

Figura 17 – Distribuição e gráficos de normalidade da métrica MPC das Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

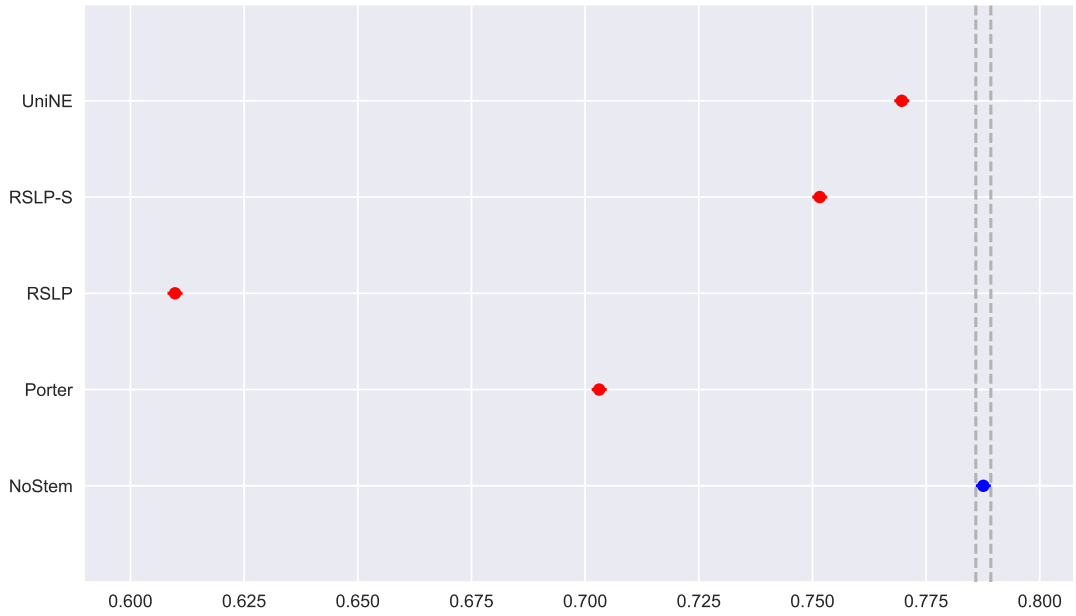

Figura 18 – Comparação da MPC nas Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

Apesar de a normalidade da MRP, Figura 19, ter sido evidenciada pelos testes de Shapiro-Wilk, com todos os tratamentos apresentando *p-value* superior ao nível de significância adotado pelo experimento, o teste de Levene refutou a hipótese de homocedasticidade dos dados. Com isso, optamos novamente pelo uso do teste não-paramétrico de Kruskal-Wallis para testar a hipótese de igualdade da MRP entre os grupos ($H0^{RP}$). Essa hipótese foi rejeitada (*p-value* < 0,001) e o teste de Mann-Whitney evidenciou a diferença entre a radicalização e o grupo de controle (Figura 20).

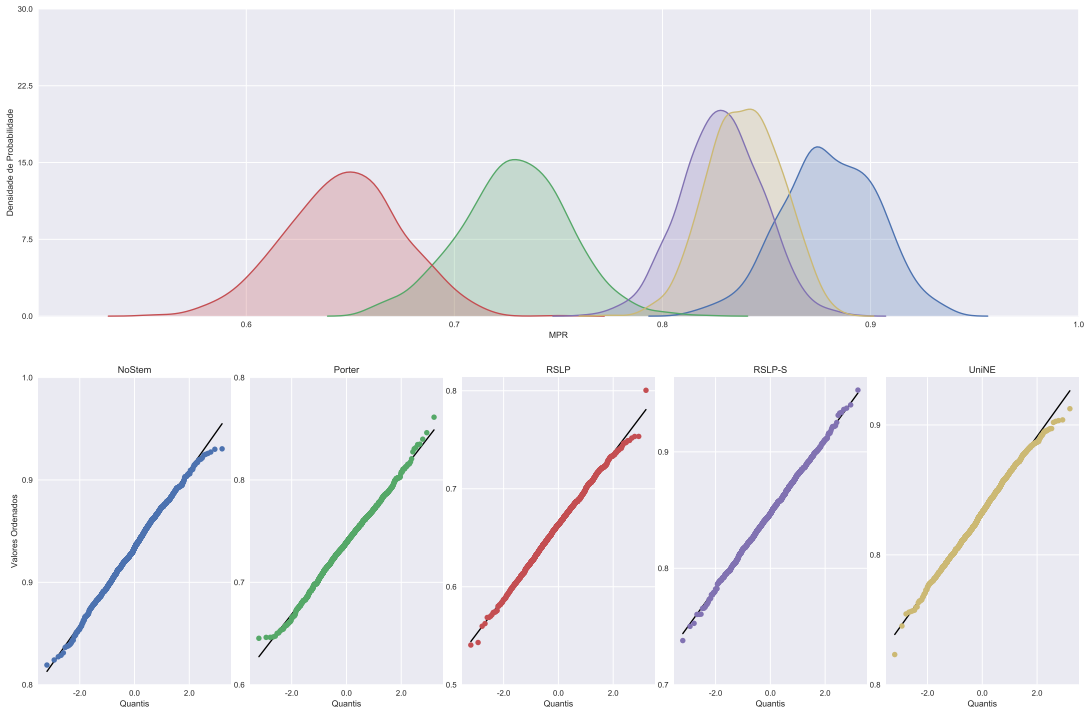

Figura 19 – Distribuição e gráficos de normalidade da métrica MRP das Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

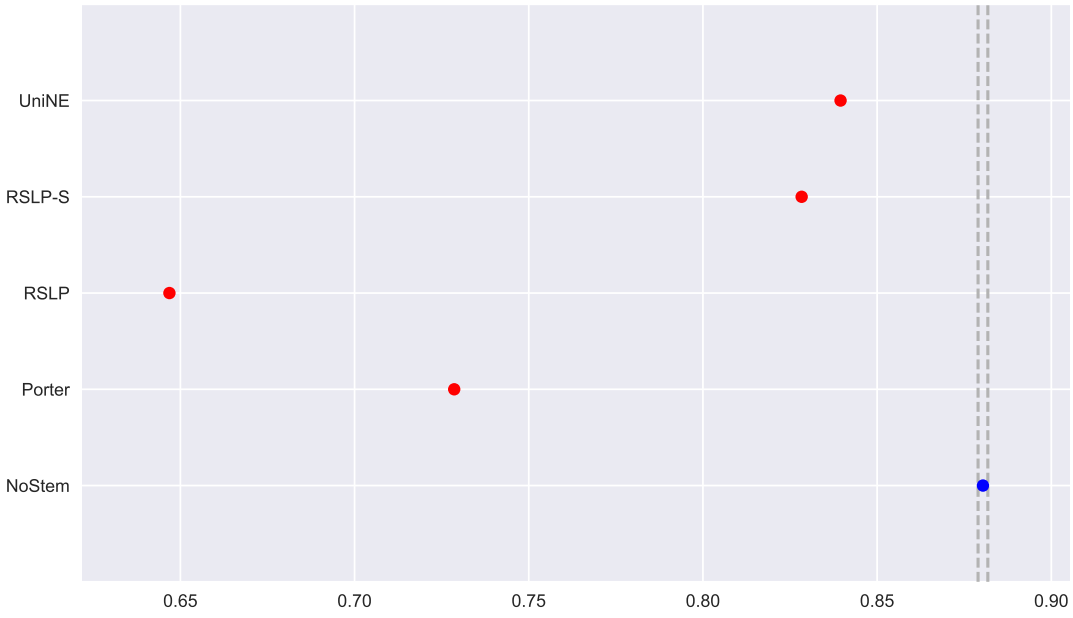

Figura 20 – Comparação da MRP nas Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

Por último, podemos visualizar as múltiplas variáveis envolvidas no experimento por

meio da Figura 21.

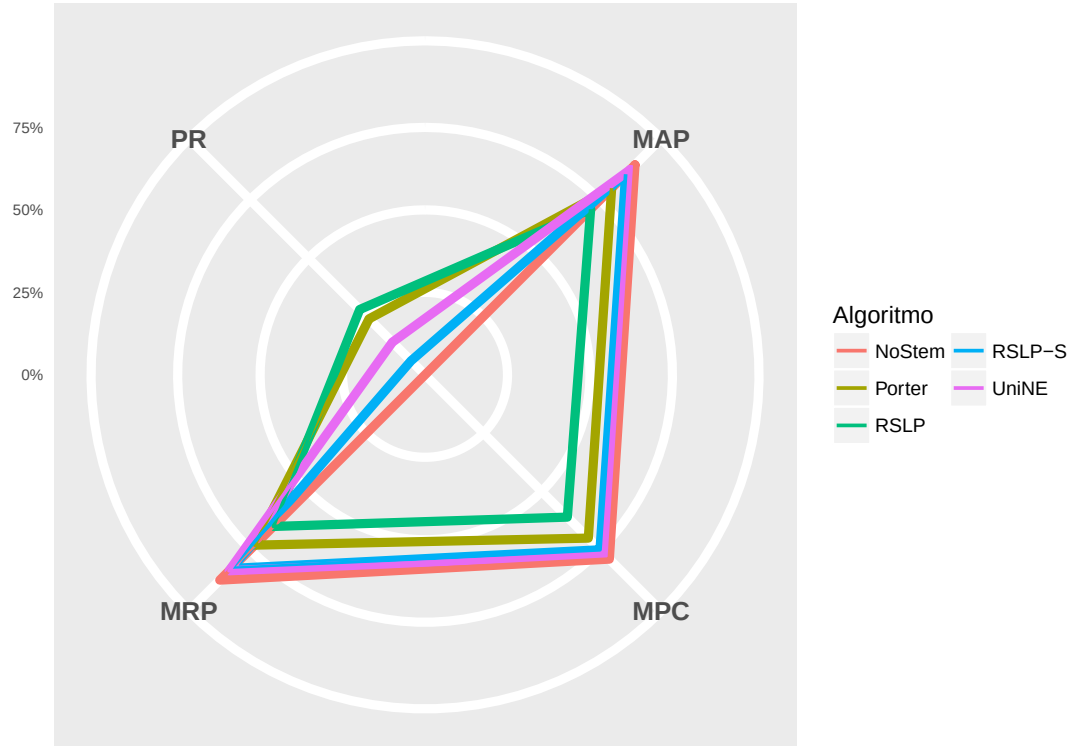

Figura 21 – Percentual de Redução (PR), MAP, MPC e MRP das Decisões Monocráticas do Segundo Grau.

Fonte: Elaborada pelo autor.

4.3.1.3 Acórdãos da Turma Recursal

Segundo a Tabela 11, esta coleção foi a única dentre as quatro estudadas, na qual a radicalização causou um aumento das três métricas. Sendo assim, analisaremos se essa diferença com relação ao grupo de controle foi estatisticamente significativa.

O grupo de controle NoStem não apresentou normalidade dos dados (*p-value* = 0,045) com respeito à MAP (Figura 22) e a hipótese de homocedasticidade foi rejeitada (*p-value* < 0,001). Seguindo o processo, a condução do teste de Kruskal-Wallis mostrou haver diferença entre os grupos estudados (*p-value* < 0,001), ou seja, a hipótese $H0^{AP}$ foi rejeitada. A análise *post hoc* evidenciou que os algoritmos RSLP-S e UniNE (Figura 23) apresentaram uma melhoria significativa da MAP.

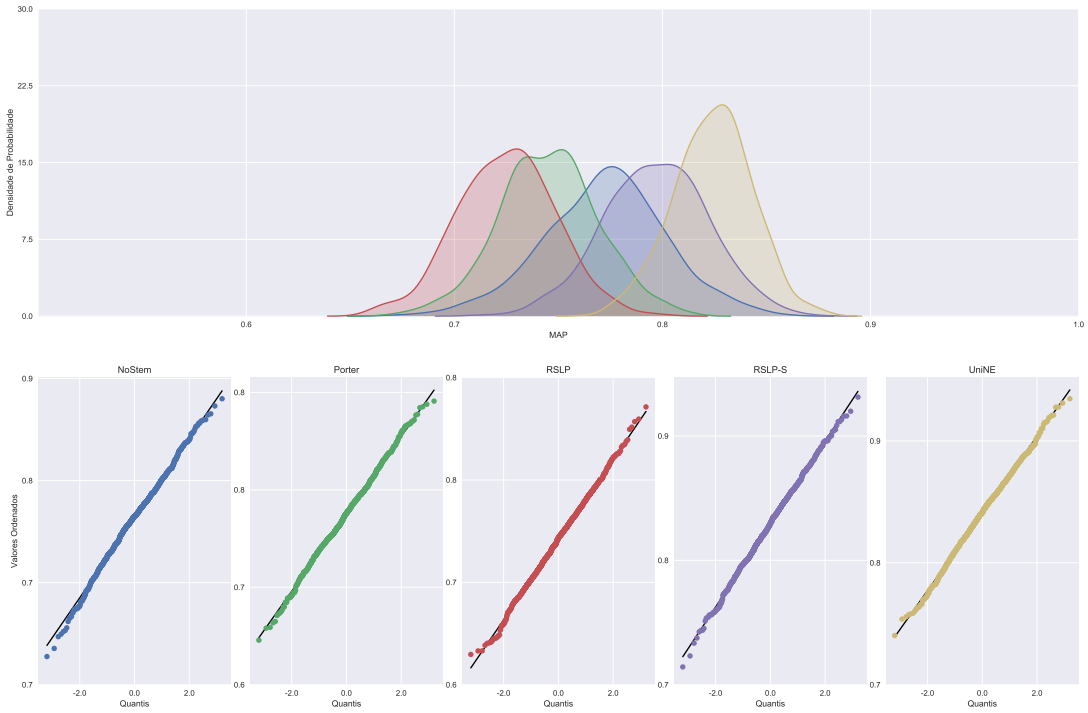

Figura 22 – Distribuição e gráficos de normalidade da métrica MAP dos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

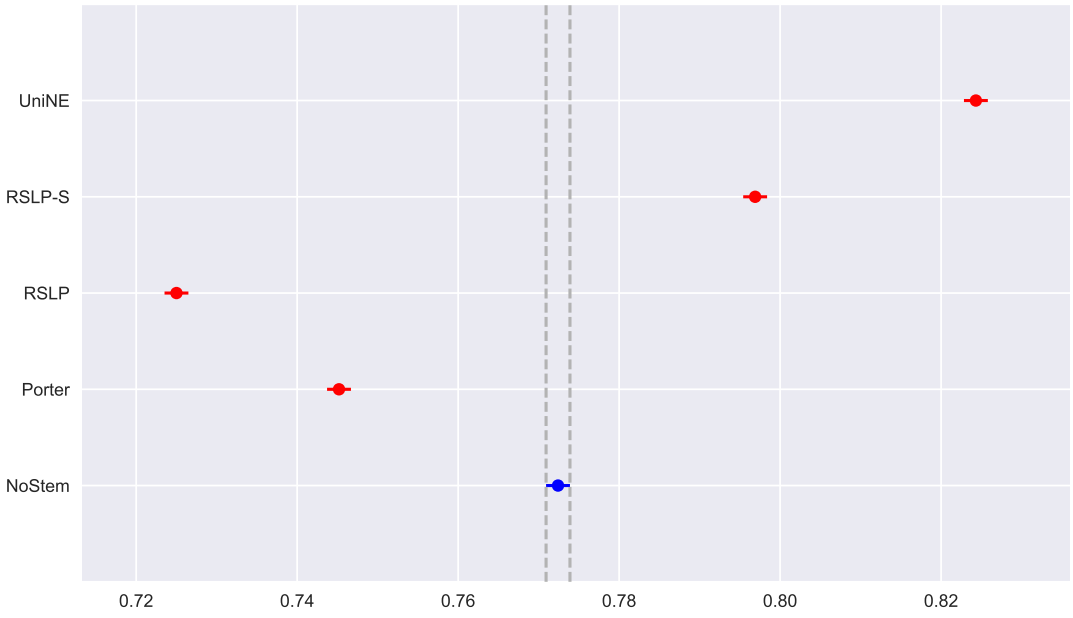

Figura 23 – Comparação da MAP nos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

Além de uma melhoria da MAP, esses dois algoritmos também causaram uma melhora

na MPC dos documentos nessa coleção. Para tal análise, conduzimos o teste de Shapiro-Wilk sobre a distribuição (Figura 24) e testamos a homocedasticidade dos dados. Apesar de todos os grupos terem apresentado normalidade, houve heterocedasticidade dos tratamentos. Assim, rejeitamos a hipótese $H0^{Pr@10}$, igualdade da MPC entre os grupos, com o teste de Kruskal-Wallis (*p-value* < 0,001) e validamos a significância desta diferença, ilustrada pela Figura 25 com Mann-Whitney.

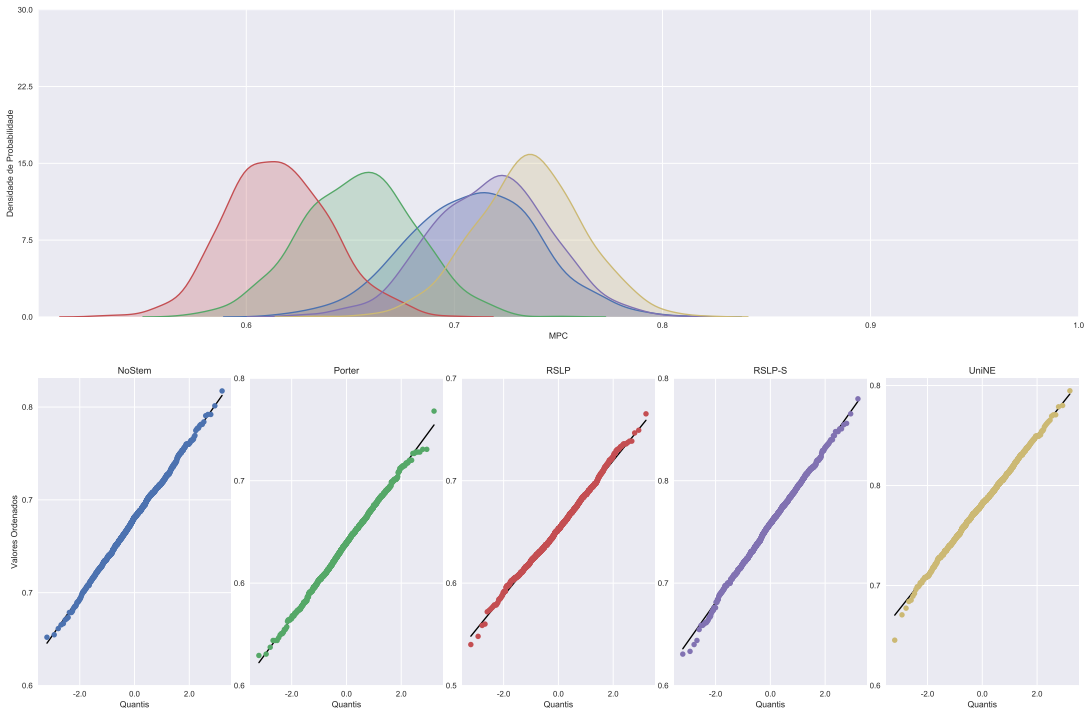

Figura 24 – Distribuição e gráficos de normalidade da métrica MPC dos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

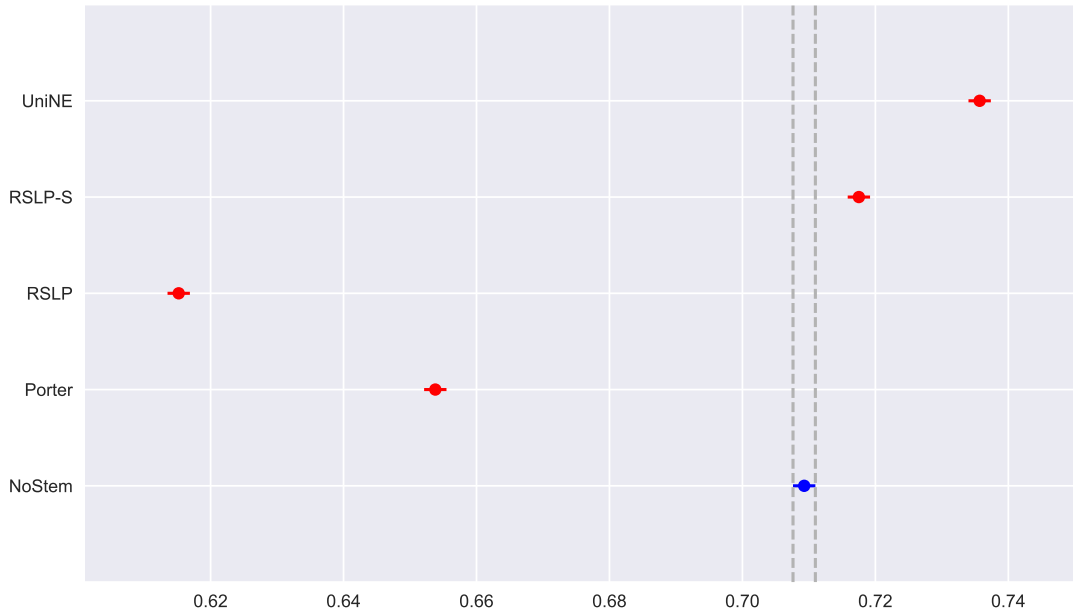

Figura 25 – Comparação da MPC nos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

De modo análogo à métrica anterior, houve normalidade dos tratamentos e heterocedasticidade dos dados relativos à MRP (Figura 26). Contudo, conforme análise *post hoc*, somente o algoritmo UniNE causou uma melhoria da métrica (Figura 27), rejeitando, desse modo, a hipótese de igualdade entre a MRP dos tratamentos ($H0^{RP}$).

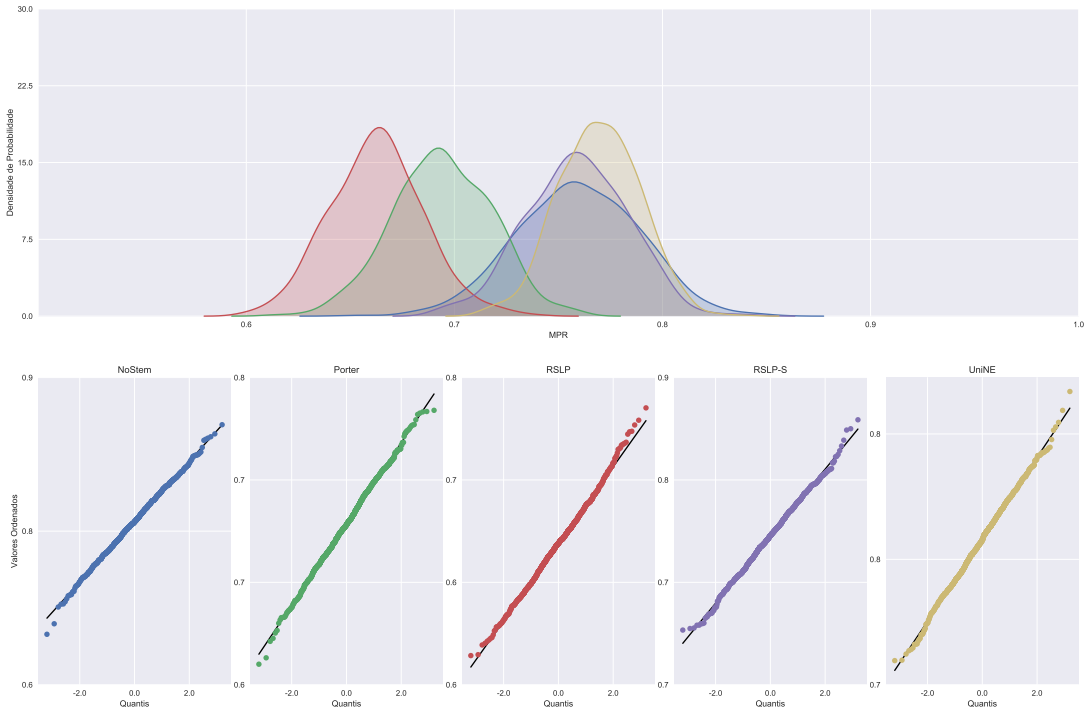

Figura 26 – Distribuição e gráficos de normalidade da métrica MRP dos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

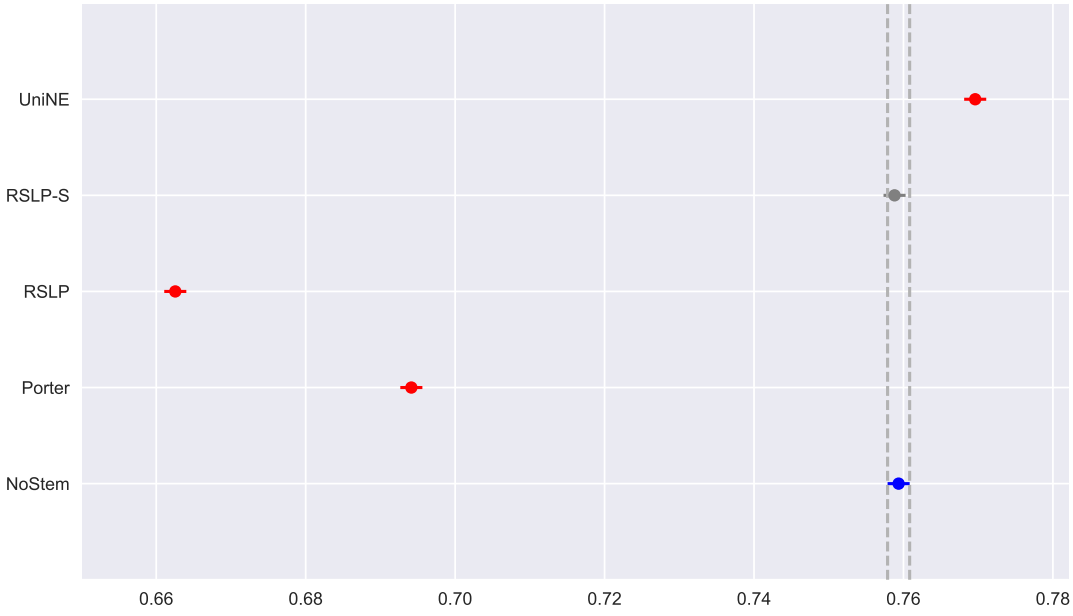

Figura 27 – Comparação da MRP nos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

No entanto, como podemos notar pela Figura 28, o algoritmo RSLP-S apresentou uma redução de dimensionalidade superior à do grupo de controle, assim, mesmo que possuam a MRP

idênticas, torna-se uma opção mais vantajosa em virtude da maior eficiência no armazenamento.

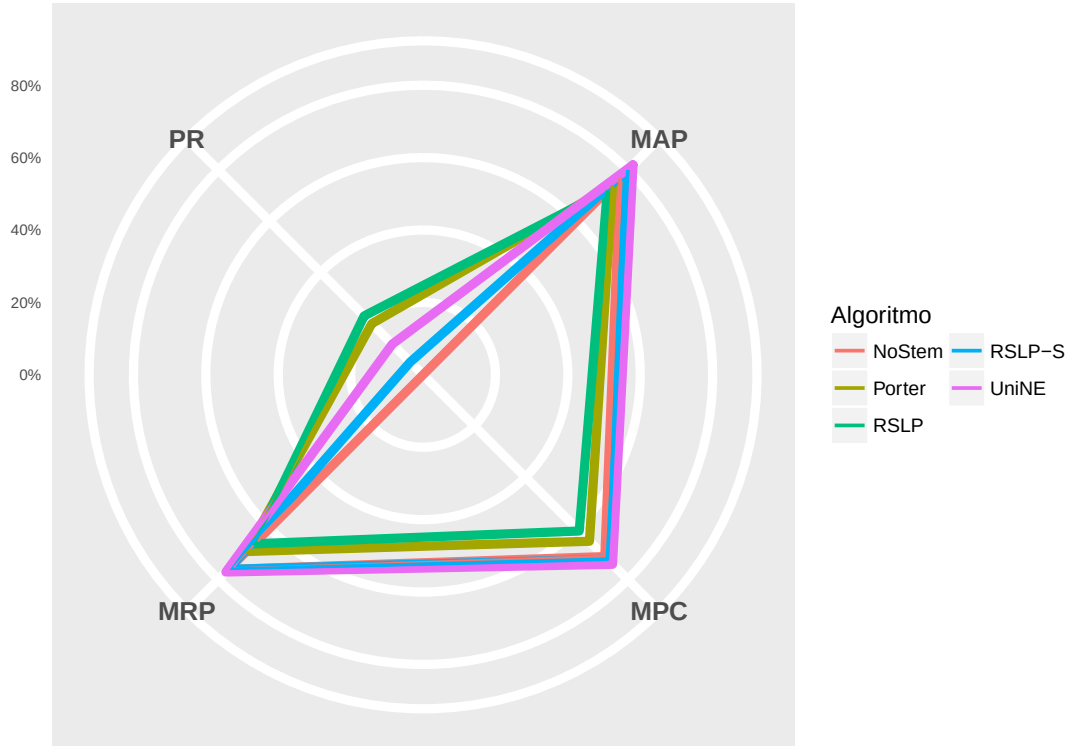

Figura 28 – Percentual de Redução (PR), MAP, MPC e MRP dos Acórdãos da Turma Recursal.

Fonte: Elaborada pelo autor.

4.3.1.4 Decisões Monocráticas da Turma Recursal

Nesta coleção, a hipótese de normalidade dos dados relacionados à MAP (Figura 29) foi rejeitada para os algoritmos NoStem (*p-value* < 0,001), Porter (*p-value* = 0,001), RSLP-S (*p-value* = 0,002) e UniNE (*p-value* < 0,001).

Após rejeitarmos a homocedasticidade dos dados, testamos a hipótese de igualdade da MAP entre os tratamentos ($H0^{AP}$). Tendo sido refutada pelo teste de Kruskal-Wallis, conduzimos uma análise *post-hoc* com o teste de Mann-Whitney. A diferença entre os tratamentos e o grupo de controle, exibida pela Figura 30, foi estatisticamente significativa, porém, o algoritmo RSLP-S apresentou *p-value* igual a 0,02, muito próximo do nível de significância adotado pelo experimento.

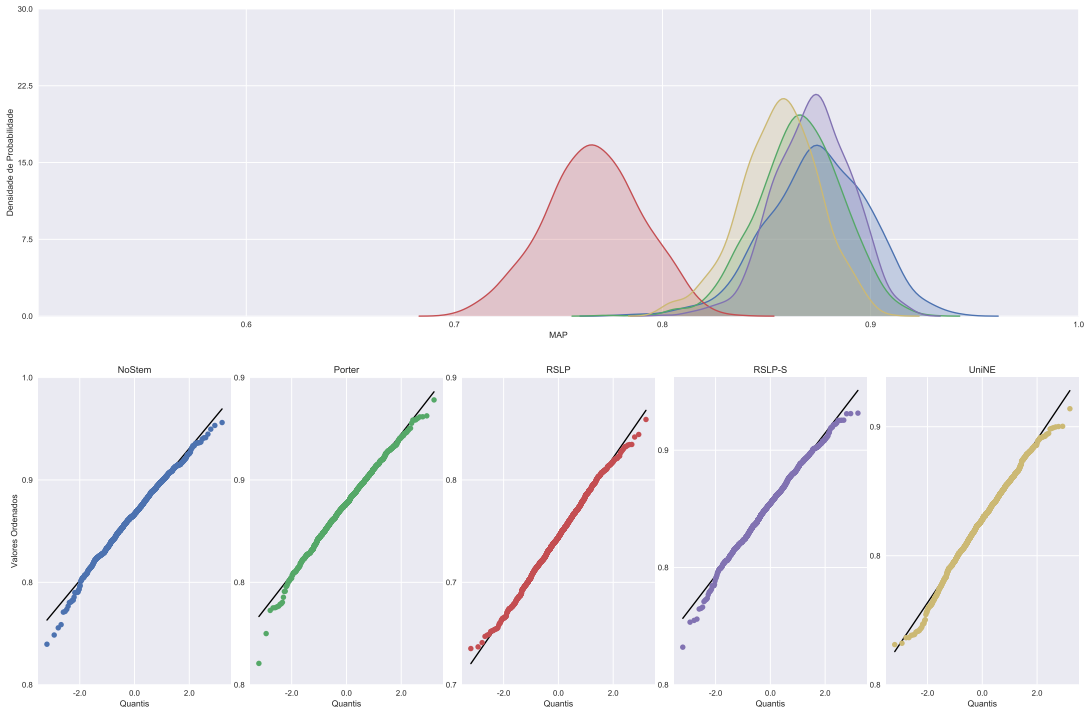

Figura 29 – Distribuição e gráficos de normalidade da métrica MAP das Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

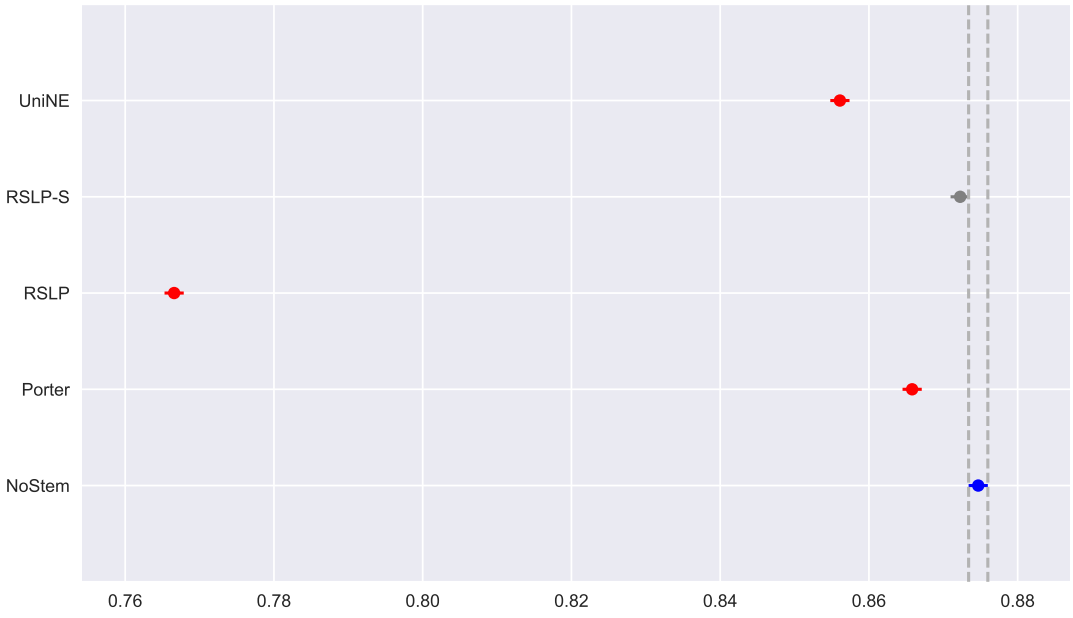

Figura 30 – Comparação da MAP nas Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

Em seguida, a premissa de normalidade da MPC, Figura 31, foi violada pelos algoritmos

RSLP-S e UniNE, com *p-values* iguais a 0,021 e 0,016, respectivamente. Além disso, confirmamos a heterocedasticidade dos dados e a significância das diferenças entre os tratamentos (Figura 32), refutando a hipótese $H0^{Pr@10}$, haja vista que os testes encontraram *p-value* inferior a 0,001.

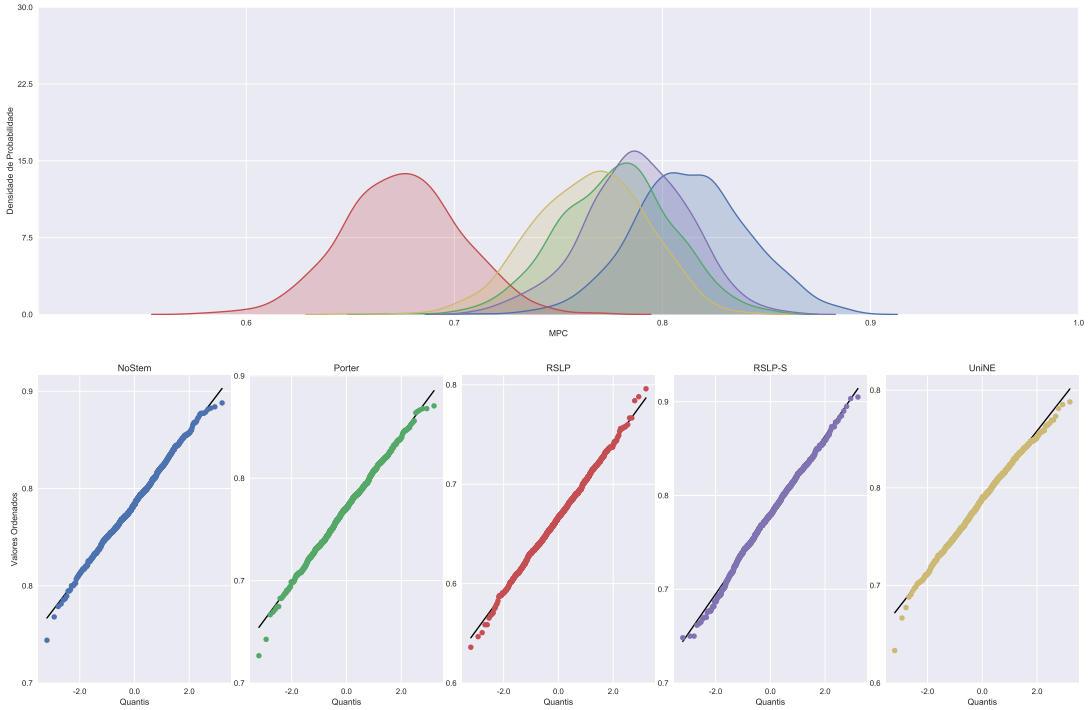

Figura 31 – Distribuição e gráficos de normalidade da métrica MPC das Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

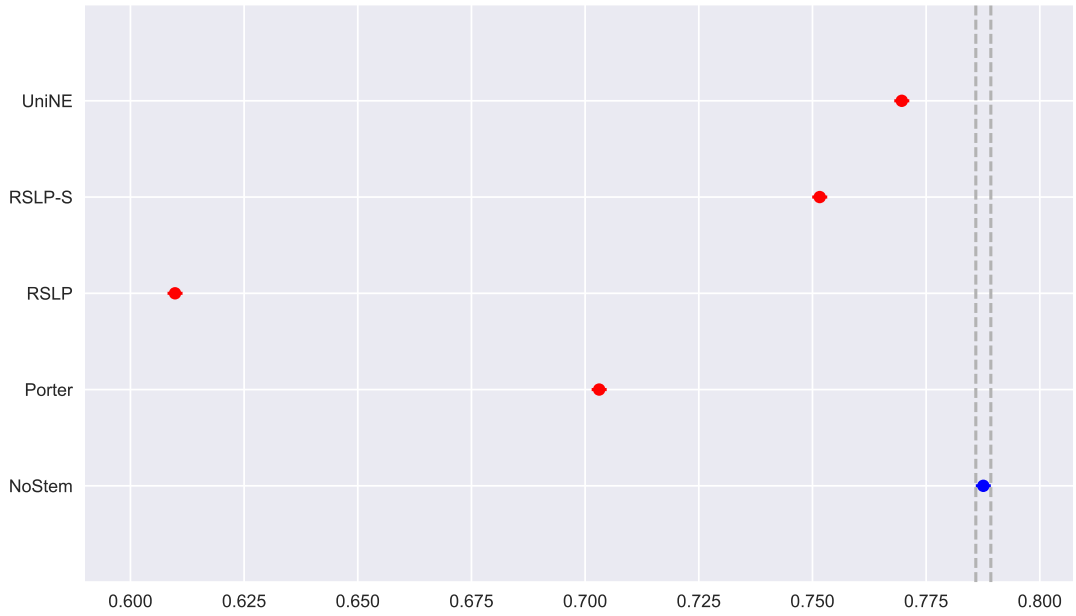

Figura 32 – Comparação da MPC nas Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

Por último, a análise dos dados da métrica MRP (Figura 33) mostrou que os algoritmos NoStem (*p-value* = 0,01), RSLP-S (*p-value* < 0,001) e UniNE (*p-value* = 0,001) não aderem à distribuição normal. Assim como nas demais métricas, o teste de Levene evidenciou a heterocedasticidade entre os grupos. Nesse cenário, ao executarmos o teste de Kruskal-Wallis, evidenciou-se a não igualdade da MRP entre os tratamentos e os testes de Mann-Whitney mostraram uma diferença estatisticamente significativa, com p-value inferior a 0,001, entre todos os tratamentos e o grupo de controle.

Figura 33 – Distribuição e gráficos de normalidade da métrica MRP das Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

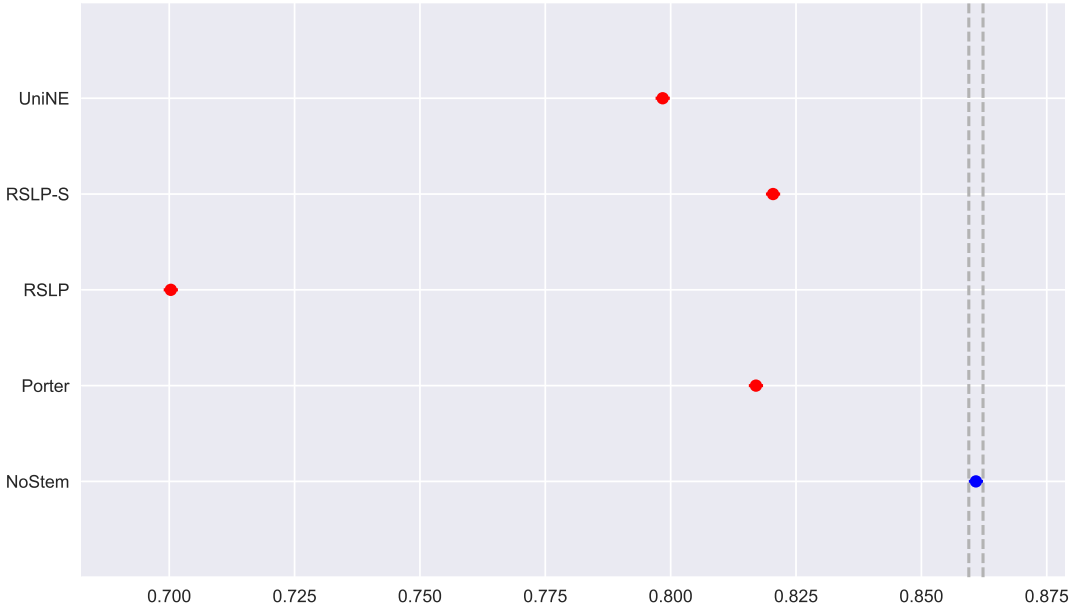

Figura 34 – Comparação da MPR nas Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

Por meio da Figura 35, podemos perceber que os algoritmos RSLP-S e UniNE apresentam MAP, MPC e MRP muito próximas das do grupo de controle. Sendo assim, o analista de

dados pode fazer a opção por um desses algoritmos e beneficiar-se da redução de dimensionalidade.

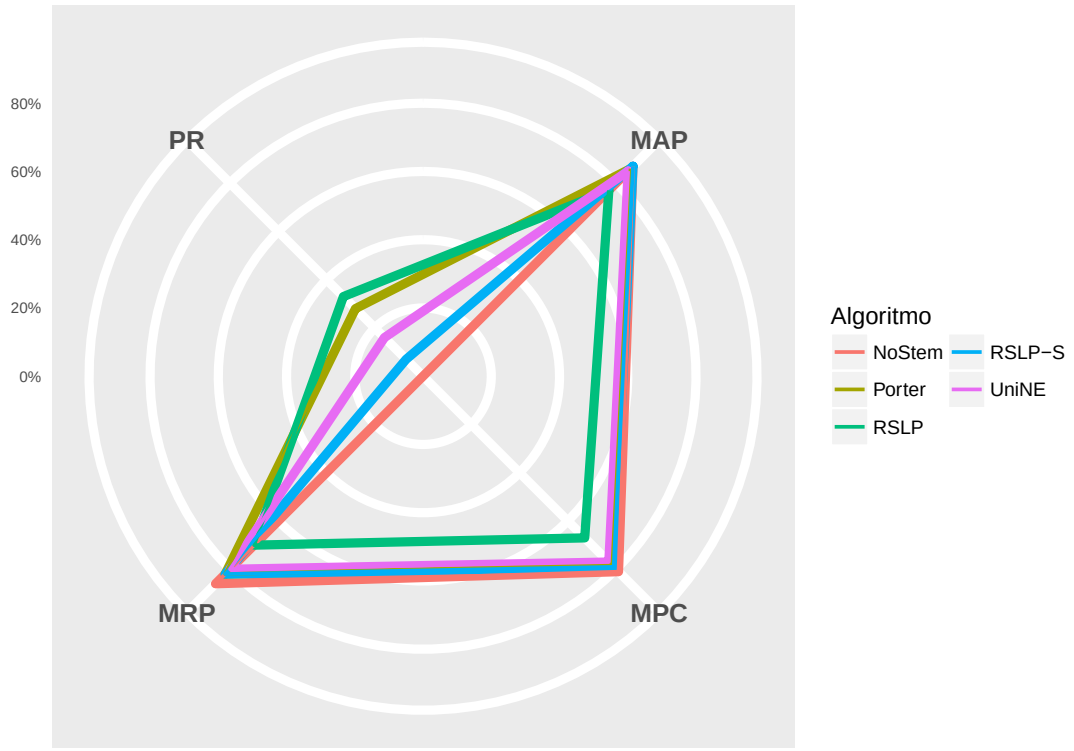

Figura 35 – Percentual de Redução (PR), MAP, MPC e MRP das Decisões Monocráticas da Turma Recursal.

Fonte: Elaborada pelo autor.

4.3.2 Ameaças à Validade

Ainda que os resultados obtidos por meio do experimento tenham sido satisfatórios, o mesmo apresenta ameaças a sua validade que precisam ser consideradas:

Ameaças à validade interna. Considerando que os dados foram coletados e analisados pelos autores, há uma forte ameaça à validade interna. No entanto, não há motivos para privilegiar um determinado algoritmo, uma vez que não existe conflito de interesses. Para mitigar qualquer viés possível, as consultas utilizadas durante o experimento foram escolhidas aleatoriamente, nos moldes do *design* experimental RCBD.

Ameaças à validade externa. De acordo com os trabalhos relacionados, a redução de dimensionalidade deveria melhorar a recuperação de documentos, no entanto, a melhoria ocorreu somente em uma das coleções analisadas. Assim, o algoritmo *nruns* pode não ter sido adequado para julgar a relevância de documentos jurisprudenciais. Como forma de mitigar esta ameaça, os julgamentos efetuados pelo algoritmo foram preservados, possibilitando que trabalhos futuros façam uso de especialistas do Direito para validarem a consistência dos mesmos.

4.3.3 Considerações Finais

Neste capítulo, realizamos o experimento responsável por analisar o impacto que a radicalização gerou sobre a recuperação de documentos judiciais. Por meio do estudo das métricas MAP, MPC e MPR, podemos observar que cada coleção apresentou características próprias, reforçando a necessidade de estudá-las individualmente.

De posse da análise individual das bases jurisprudenciais, os analistas de dados poderão elaborar estratégias para priorizar métricas que sejam mais relevantes para determinados cenários, seja para aumentar a eficácia da recuperação dos documentos, seja para trabalhar sobre a redução de dimensionalidade dos dados.

5
Conclusão

Dia após dia, os magistrados, no exercício de suas funções, produzem um vasto número de documentos. Assim, é preciso avaliar técnicas que auxiliem no armazenamento e busca de tais informações. No entanto, como fazer avaliações quantitativas de forma sistemática?

Não é raro observarmos nos veículos de comunicação, notícias que mostram o desperdício de recursos por parte do poder público. Tecnologias, abordagens e métodos são simplesmente trocados, sem evidências empíricas, baseando-se no *feeling* do gestor, carecendo de critérios objetivos para nortear as decisões.

Nesse sentido, a abordagem experimental proposta por Wohlin et al. (2012) foi o alicerce sobre o qual conduzimos as análises da radicalização e da recuperação de documentos. Dessa forma, os dados quantitativos descritos neste trabalho foram obtidos por meio de um processo sistemático, mitigando possíveis vieses atribuíveis aos autores, fornecendo subsídio para tomadas de decisão mais objetivas.

5.1 Contribuições

Assim, dando continuidade ao tópico anterior, salientamos que a principal contribuição deste estudo consiste na condução de um processo experimental para analisar a redução de dimensionalidade obtida radicalizando-se as quatro coleções de documentos jurisprudenciais do Tribunal de Justiça do Estado de Sergipe. Além disso, analisou-se o impacto desta redução sobre a recuperação dos documentos.

Dentre as principais contribuições deste trabalho, destacam-se:

- Geração de uma massa de teste com documentos jurisprudenciais obtidos a partir de um ambiente real em produção;

- Experimento que analisa o impacto da radicalização sobre a redução de dimensionalidade das bases. Ressalte-se que os resultados foram publicados em conferências internacionais: na ITNG 2017, sob o título *Summary Report of Experimental Analysis of Stemming Algorithms Applied to Judicial Jurisprudence*, e na ICEIS 2017, com o título *Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence - An Experimental Analysis*.

- Avaliação experimental da recuperação dos documentos radicalizados.

Como consequência destas contribuições, conseguimos responder às questões de pesquisa elaboradas no início do trabalho:

Q1: No contexto jurisprudencial, a aplicação de algoritmos de radicalização reduz de forma significativa a quantidade de termos únicos por documento? Sim, a radicalização reduziu a quantidade de termos únicos em todas as coleções. A maior redução ocorreu com o uso do algoritmo RSLP sobre os Acórdãos do Segundo Grau, chegando a 52%.

Q2: A eficácia dos algoritmos de radicalização é a mesma em todas as coleções judiciais? Não, os algoritmos de radicalização variaram sua efetividade a depender da coleção sobre a qual foram aplicados.

Q3: A radicalização tem efeito sobre os resultados obtidos mediante as buscas jurisprudenciais? Sim, a radicalização alterou a recuperação dos documentos em todas as coleções. No entanto, teve efeito positivo somente nos documentos referentes aos Acórdãos da Turma Recursal. Nas demais coleções, houve uma degradação na recuperação dos documentos jurisprudenciais.

Dessa maneira, outros órgãos do Poder Judiciário poderão utilizar os dados levantados nesta pesquisa para escolher algoritmos de radicalização mais adequados, ou, até mesmo, para replicar o experimento e comparar os resultados. Registre-se, por oportuno, que a Divisão de Banco de Dados do TJSE, atualmente gerida pelo autor deste trabalho, tem desenvolvido ações visando promover o uso da abordagem experimental como subsídio para a tomada de decisões.

5.2 Perspectivas

Em função dos custos envolvidos na criação de testes que utilizam julgamentos feitos por humanos, optamos pelo método *nruns*, proposto por Sakai e Lin (2010), considerando ter apresentado resultados satisfatórios quando aplicado a coleções de testes disponibilizadas por conferências especializadas.

No entanto, sabendo que o algoritmo não foi testado sobre bases jurídicas e que os experimentos aqui realizados, em contraposição aos trabalhos relacionados, mostraram que a radicalização não gerou uma melhora na recuperação dos documentos em todas as coleções, uma

possível evolução desta pesquisa seria coletar uma amostra dos julgamentos efetuados para ser avaliada por especialistas do Direito. Ao final, teríamos como checar a eficácia do *nruns* no contexto jurisprudencial.

Além dessa sugestão, outros possíveis desdobramentos desta pesquisa são:

- Criar um algoritmo de radicalização específico para o domínio jurídico e comparar sua eficácia contra os algoritmos genéricos;
- Combinar outras técnicas de recuperação de informação, avaliando variações da eficácia quando comparada à aplicação somente da radicalização;
- Analisar o impacto da radicalização sobre a classificação de documentos jurisprudenciais;
- Mensurar os efeitos da redução de dimensionalidade sobre a clusterização dos documentos jurídicos.

Destaque-se que os resultados integrados desta dissertação serão submetidos ao periódico *Artificial Intelligence and Law*, ISSN 0924-8463, em razão de seu escopo estar diretamente relacionado ao presente trabalho.

5.3 Considerações Finais

Este trabalho apresentou os resultados da análise do processo de radicalização e seu impacto sobre a recuperação de documentos jurisprudenciais por meio de um processo experimental. Além disso, relaram-se as contribuições obtidas, bem como as perspectivas de pesquisas futuras, tendo por intuito uma evolução do presente estudo.

Referências

AGARWAL, Nidhi; DEEP, Prakhar. 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**, p. 742–747, 2014. DOI: 10.1109/CONFLUENCE.2014.6949233.

AHAD, Nor Aishah et al. Sensitivity of normality tests to non-normal data. **Sains Malaysiana**, v. 40, n. 6, p. 637–641, 2011. ISSN 01266039. DOI: http://www.ukm.my/jsm/pdf_files/SM-PDF-40-6-2011/15%20NorAishah.pdf.

ALVARES, Reinaldo Viana; GARCIA, Ana Cristina Bicharra; FERRAZ, Inhaúma. 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, p. 693–701, 2005. ISSN 03029743. DOI: 10.1007/11595014_67.

ASF. **Lucene**. [S.l.: s.n.], 2011. Disponível em: <https://lucene.apache.org/core/>.

BASILI, Victor R; CALDIERA, Gianluigi; ROMBACH, H Dieter. The goal question metric approach. **Encyclopedia of Software Engineering**, v. 2, p. 528–532, 1994. ISSN <null>. DOI: 10.1.1.104.8626. Disponível em: <http://maisqual.squoring.com/wiki/index.php/The%20Goal%20Question%20Metric%20Approach>.

BASILI, V. et al. **Aligning Organizations Through Measurement - The GQM+Strategies Approach**. [S.l.: s.n.], 2014. p. 205. ISBN 978-3319050461. Disponível em: <http://www.springer.com/computer/swe/book/978-3-319-05046-1%7B%5C%%7D5Cnhttp://www.amazon.com/Aligning-Organizations-through-Measurement-Engineering/dp/331905046X?tag=donations09-20>.

CÂMARA JÚNIOR, Auto Tavares. **Indexação automática de acórdãos por meio de processamento de linguagem natural**. 2007. f. 141. Tese (Doutorado) – Universidade de Brasília.

CARTERETTE, B.; VOORHEES, E. M. Overview of information retrieval evaluation. **In Current challenges in patent information retrieval.**, p. 69–85, 2011.

CLEVERDON, Cyril. **The Cranfield tests on index language devices**. v. 19. [S.l.: s.n.], 1967. p. 173–194. ISBN 0-89791-448-1. DOI: 10.1108/eb050097.

COHEN, Jacob. A power primer. **Psychological Bulletin**, v. 112, n. 1, p. 155–159, 1992. ISSN 0033-2909. DOI: 10.1037/0033-2909.112.1.155. arXiv: arXiv:1011.1669v3. Disponível em: <http://doi.apa.org/getdoi.cfm?doi=10.1037/0033-2909.112.1.155>.

DUNN, Olive Jean. Multiple Comparisons Among Means. **Journal of the American Statistical Association**, v. 56, n. 293, p. 52–64, 1961. ISSN 0162-1459. DOI: 10.2307/2282330. arXiv: arXiv:1011.1669v3.

ELLIS, Paul D. **The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results**. [S.l.: s.n.], 2010. v. 53, p. 170. ISBN 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

FANG, Hui; TAO, Tao; ZHAI, Chengxiang. Diagnostic Evaluation of Information Retrieval Models. **ACMTransactions on Information Systems**, v. 29, n. 7, 7:2–7:41, 2011. ISSN 10468188. DOI: 10.1145/1961209.1961210. Disponível em: <http://portal.acm.org/citation.cfm?id=1961210>.

FLORES, Felipe N.; MOREIRA, Viviane P. Assessing the impact of Stemming Accuracy on Information Retrieval – A multilingual perspective. **Information Processing & Management**, Elsevier Ltd, v. 0, p. 1–15, 2016. ISSN 03064573. DOI: 10.1016/j.ipm.2016.03.004. Disponível em: <http://linkinghub.elsevier.com/retrieve/pii/S0306457316300358>.

HOLM, Sture. A Simple Sequentially Rejective Multiple Test Procedure. **Scandinavian Journal of Statistics**, v. 6, n. 2, p. 65–70, 1979. ISSN 03036898. DOI: 10.2307/4615733. arXiv: arXiv:1011.1669v3. Disponível em: <http://www.jstor.org/stable/10.2307/4615733>.

JAMES, Gareth et al. **An Introduction to Statistical Learning**. New York: Springer, 2013. ISBN 9780387781884. DOI: 10.1016/j.peva.2007.06.006.

JONES, Eric et al. **SciPy: Open source scientific tools for Python**. [S.l.: s.n.], 2001. Disponível em: <http://www.scipy.org/>.

KITCHENHAM, Barbara. Procedures for performing systematic reviews. **Keele, UK, Keele University**, v. 33, TR/SE-0401, p. 28, 2004. ISSN 13537776. DOI: 10.1.1.122.3308.

KITCHENHAM, Barbara et al. Robust Statistical Methods for Empirical Software Engineering. **Empirical Software Engineering**, Empirical Software Engineering, p. 1–52, 2016. ISSN 15737616. DOI: 10.1007/s10664-016-9437-5.

KONTOSTATHIS, April; KULP, Scott. The Effect of Normalization when Recall Really Matters. v. 101, 2007.

LEVENE, Howard et al. Robust tests for equality of variances. **Contributions to probability and statistics**, v. 1, p. 278–292, 1960.

LOUREIRO, Luís; GAMEIRO, Manuel. Interpretação crítica dos resultados estatísticos: para lá da significância estatística. **Revista de Enfermagem Referência**, III Série, nº 3, p. 151–162, 2011. ISSN 08740283. DOI: 10.12707/RIII1009. Disponível em: <http://www.esenfc.pt/rr/index.php?module=rr%7B%5C&%7Dtarget=publicationDetails%7B%5C&%7Dpesquisa=%7B%5C&%7Did%7B%5C_%7Dartigo=2224%7B%5C&%7Did%7B%5C_%7Drevista=9%7B%5C&%7Did%7B%5C_%7Dedicao=35>.

MAGALHÃES, Cristiane Costa. **Minerjur : Uma Ferramenta Para Mineração De Bases De Jurisprudência**. 2008. Tese (Doutorado) – 2008. Dissertação (Mestrado em Sistemas e Computação) – Universidade Salvador, Salvador.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **An Introduction to Information Retrieval**. [S.l.]: Cambridge University Press, 2009. ISBN 0521865719. DOI: 10.1109/LPT.2009.2020494. arXiv: 0521865719978052186571 5. Disponível em: <http://nlp.stanford.edu/IR-book/>.

MAXIMILIANO, C. **Hermenêutica e Aplicação do Direito**. 20. ed. Rio de Janeiro: Forense, 2011. p. 1–352. ISBN 9788530934477.

MESSIAS, André et al. New standardized texts in Brazilian Portuguese to assess reading speed: comparison with four European languages. **Arquivos brasileiros de oftalmologia**, SciELO Brasil, v. 71, n. 4, p. 553–558, 2008.

NIST. **TREC Eval**. [S.l.]: GitHub, 2016. https://github.com/usnistgov/trec_eval.

OLIVEIRA, Robert A. N. de; COLAÇO JÚNIOR, Methanias. Assessing the Impact of Stemming Algorithms Applied to Judicial Jurisprudence - An Experimental Analysis. ScitePress, p. 99–105, 2017. DOI: 10.5220/0006317100990105.

_____. Summary Report of Experimental Analysis of Stemming Algorithms Applied to Judicial Jurisprudence. **ITNG Proceedings**, 2017.

ORENGO, Viviane Moreira; BURIOL, Luciana S.; COELHO, Alexandre Ramos. A Study on the Use of Stemming for Monolingual Ad-Hoc Portuguese Information Retrieval. **Evaluation of Multilingual and Multi-modal Information Retrieval**, p. 91–98, 2007. DOI: 10.1007/978-3-540-74999-8_12. Disponível em: <http://link.springer.com/10.1007/978-3-540-74999-8_12>.

ORENGO, Viviane Moreira; HUYCK, Christian. A stemming algorithm for the portuguese language. In: IEEE. STRING Processing and Information Retrieval, 2001. SPIRE 2001. Proceedings. Eighth International Symposium on. [S.l.: s.n.], 2001. p. 186–193.

PORTER, M. F. An algorithm for suffix stripping. **Program**, p. 130–137, 1980. DOI: https://doi.org/10.1108/eb046814.

RAZALI, Nornadiah Mohd; WAH, Yap Bee. Power comparisons of Shapiro-Wilk , Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests. **Journal of Statistical Modeling and Analytics**, v. 2, n. 1, p. 21–33, 2011. ISSN 9789673631575. DOI: doi:10.1515/bile-2015-0008.

ROBERTSON, Stephen; ZARAGOVA, Hugo. **The Probabilistic Relevance Framework: BM25 and Beyond**. [S.l.: s.n.], 2009. v. 3, p. 333–389. ISBN 1500000019. DOI: 10.1561/1500000019.

ROITBLAT, Herbert L.; KERSHAW, Anne; OOT, Patrick. Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review. **International Review of Research in Open and Distance Learning**, v. 14, n. 4, p. 90–103, 2009. ISSN 14923831. DOI: 10.1002/asi. arXiv: 0803.1716.

SAKAI, Tetsuya; KANDO, Noriko et al. Overview of the {NTCIR}-7 {ACLIA} {IR4QA} Task. **Proceedings of the Seventh NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access**, p. 77–114, 2008.

SAKAI, Tetsuya; LIN, Chin-Yew. Ranking Retrieval Systems without Relevance Assessments — Revisited. **EVIA 2010Proceedings of the 3rd International Workshop on Evaluating Information Access (EVIA)**, p. 25–33, 2010.

SANTOS, Washington dos. **Dicionário Jurídico Brasileiro**. [S.l.]: Livraria Del Rey Editora LTDA, 2001. ISBN 9788578110796. DOI: 10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

SPSS, IBM. Statistical Package for Social Science. **USA: International Business Machines Corporation SPSS Statistics**, 2012.

TEAM, R Development Core. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2008. Disponível em: <http://www.r-project.org/>.

THEODORSSON-NORHEIM, Elvar. 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**, Elsevier, v. 17, n. 2, p. 85–99, 1987.

VOORHEES, Ellen M; HARMAN, Donna K. **TREC: experiment and evaluation in information retrieval**. [S.l.]: MIT press Cambridge, 2005. ISBN 0262220733.

WEBBER, William. Re-examining the Effectiveness of Manual Review. **Sire '11**, 2011. Disponível em: <http://www.umiacs.umd.edu/%7B~%7Doard/sire11/papers/webber.pdf>.

WEISS, S. M. et al. **Text mining: predictive methods for analyzing unstructured information**. [S.l.]: Springer Science & Business Media, 2010.

WOHLIN, Claes et al. **Experimentation in Software Engineering**. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012. ISBN 978-3-642-29043-5. DOI: 10.1007/978-3-642-29044-2. arXiv: arXiv:1011.1669v3. Disponível em: <http://link.springer.com/10.1007/978-3-642-29044-2>.

ZHAI, Chengxiang; MASSUNG, Sean. **Text Data Management and Analysis**. [S.l.: s.n.], 2016. ISBN 9781970001198.