Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Identificação automática de casos repetitivos no MPDFT

Daniel de Souza Costa Pedroso

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Marcelo Ladeira

Coorientador
Prof. Dr. Thiago de Paulo Faleiros

Brasília
2018

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

de Souza Costa Pedroso, Daniel
dD184i Identificação automática de casos repetitivos no MPDFT / Daniel de Souza Costa Pedroso; orientador Marcelo Ladeira; co-orientador Thiago de Paulo Faleiros. -- Brasília, 2018.
141 p.

Tese (Doutorado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2018.

1. Modelagem de Tópicos. 2. Recuperação de Documentos Jurídicos. 3. Mineração de Dados. 4. Ministério Público. I. Ladeira, Marcelo, orient. II. de Paulo Faleiros, Thiago, co orient. III. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Identificação automática de casos repetitivos no MPDFT

Daniel de Souza Costa Pedroso

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Marcelo Ladeira (Orientador)
CIC/UnB

Prof. Dr. Díbio Leandro Borges
CIC/UnB

Prof. Dr. Sandro José Rigo
Unisinos

Prof. Dr. Vinicius Ruela Pereira Borges
CIC/UnB

Prof. Dr. Aletéia Patrícia Favacho de Araújo
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 14 de Dezembro de 2018

Dedicatória

Dedico este trabalho ao Ministério Público do Distrito Federal e Territórios, a Deus e à minha família.

Agradecimentos

Agradeço a todos os colegas de trabalho que se dispuseram para prestar informações e contribuir com o trabalho de um modo geral. Agradeço a Deus e a minha família pelo apoio quase incondicional e pela confiança em mim depositada por todos os envolvidos.

Resumo

O Ministério Público do Distrito Federal e Territórios (MPDFT) aprecia um volume de casos da ordem de 200 mil novos feitos anualmente. Entre os casos apreciados é notável a ocorrência de casos semelhantes ou repetitivos. O tratamento destes casos pode ser mais célere se os casos semelhantes puderem ser encontrados rapidamente para servirem como embasamento para o caso em tratamento. Até então, o problema é abordado de modo descentralizado entre as diversas equipes de trabalho do órgão. Este trabalho tem o objetivo de avaliar o uso de técnicas de recuperação de informações para viabilizar a identificação automatizada de casos semelhantes. Como prova de conceito, as técnicas de indexação sintática (TF-IDF e BM25) e semântica (*Latent Semantic Indexing* - LSI e *Latent Dirichlet Allocation* - LDA) foram avaliadas com o uso de bases de documentos de duas áreas do MPDFT: Procuradorias de Justiça Criminal e Procuradorias de Justiça Criminal Especializada. Além disso, avaliamos o enriquecimento dos modelos obtidos com o uso dos dados cadastrais acumulados acerca dos casos, e também com as citações às normas jurídicas observadas nos documentos. Os modelos foram avaliados com o uso de bases de referência produzidas a partir de amostras extraídas das bases de documentos das Procuradorias de Justiça Criminal e Criminal Especializada. A métrica utilizada para medir a performance dos modelos foi a *Normalized Discounted Cumulated Gain* - NDCG. Ao final dos experimentos, concluímos que, no âmbito das bases de documentos analisadas, não houve diferença significativa de performance entre as técnicas de indexação semântica e sintática. Além disso, não foi verificado ganho de performance significativo com o enriquecimento dos modelos. Considerando isto, elegemos a técnica BM25 como mais adequada por ter bom equilíbrio entre performance e simplicidade.

Palavras-chave: Modelagem de Tópicos, Recuperação de Documentos Jurídicos, Mineração de Dados, Ministério Público

Abstract

The Public Ministry of the Federal District and Territories (MPDFT) appreciates a volume of 200,000 new cases annually. Among these cases, the occurrence of similar or repetitive cases is remarkable. The response for these cases may be improved if similar cases can be found quickly to serve as a basement or template for the case under treatment. Nowadays, this problem is addressed in a decentralized way among the various corporate teams, and it may be improved. This work aims to evaluate the use of information retrieval techniques to enable the automated identification of similar cases. As a proof of concept, syntactic indexing (TF-IDF and BM25) and semantic indexing (Latent Semantic Indexing - LSI and Latent Dirichlet Allocation - LDA) techniques were evaluated using document collections from two public prosecutor's offices. In addition, we evaluated model enrichment with the use of recorded data about the cases, and also with the legal norm citations observed in documents. The models were evaluated using baseline document collections sampled from full document collection from two public prosecutor's offices. The metric used to measure the performance of the models was the Normalized Discounted Cumulated Gain - NDCG. We concluded that, considering the document bases used, there was no significant performance difference between semantic and syntactic indexing techniques. In addition, we observe no significant performance gain with model enrichment. So, we have chosen the BM25 technique as more adequate because it has a good balance between performance and simplicity.

Keywords: Topic Model, Legal Information Retrieval, Data Mining, Public Ministry

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

BM25 *Best Match 25*.

BOW *Bag of Words*.

CNMP Conselho Nacional do Ministério Público.

CRISP-DM *Cross Industry Standard Process for Data Mining*.

LDA *Latent Dirichilet Allocation*.

LSI *Latent Semantic Indexing*.

MAP *Mean Average Precision*.

MPDFT Ministério Público do Distrito Federal e Territórios.

MPE Ministério Público dos Estados.

MPU Ministério Público da União.

MRI Modelo de recuperação de informações.

NDCG *Normalized Discounted Cumulated Gain*.

.

pLSI *Probabilistic Latent Semantic Indexing*.

STF Superior Tribunal de Justiça.

SVD *Singular Value Decomposition*.

TF-IDF *Term Frequency - Inverse Document Frequency*.

TJDFT Tribunal de Justiça do Distrito Federal e Territórios.

URN *Unified Resource Name*.

Capítulo 1

Introdução

Nesta seção serão introduzidos o contexto do trabalho, a definição do problema abordado, a justificativa e objetivos do projeto, as hipóteses de pesquisa e a contribuição esperada.

1.1 O Ministério Público Brasileiro

O Ministério Público tem como missão defender o interesse público, a ordem jurídica e a sociedade perante a Administração Pública e demais Poderes, inclusive o Judiciário. Ao órgão compete a instauração de inquéritos civis e policiais, a requisição de diligências investigatórias e procedimentos administrativos, o exercício do controle externo da atividade policial, a participação nos conselhos penitenciários, e a fiscalização da execução penal [1].

O Ministério Público Brasileiro é composto pelo Ministério Público da União (MPU) e pelo Ministério Público dos Estados (MPE)

O MPU é composto de 4 ramos a saber:

- MPF - Ministério Público Federal
- MPT - Ministério Público do Trabalho
- MPM - Ministério Público Militar
- MPDFT - Ministério Público do Distrito Federal e Territórios

Além dos 4 ramos existe o Conselho Nacional do Ministério Público (CNMP), que tem como finalidade executar a fiscalização administrativa, financeira e disciplinar de todo o Ministério Público Brasileiro, bem como expedir atos regulamentares.

1.2 O Ministério Público do Distrito Federal e Territórios

O Ministério Público do Distrito Federal e Territórios (MPDFT) exerce suas funções nas causas de competência do Tribunal de Justiça e dos Juízes do Distrito Federal e Territórios. A instituição atua em diversas áreas de interesse público, nas áreas cíveis, controle da atividade policial, crime organizado, saúde, deficientes, educação, eleitoral, entorpecentes, execuções penais, ordem tributária, patrimônio público, registros públicos, direitos do consumidor, dos idosos, dos menores, e outros.

No cumprimento de sua missão, o MPDFT pode propor ações judiciais ou inquéritos baseados em denúncias ou mesmo nos fatos observados durante suas atividades de fiscalização e termos circunstanciados. O órgão também atua quando for intimado pelo Tribunal de Justiça do Distrito Federal e Territórios (TJDFT) a manifestar-se em processos judiciais, agindo como defensor do interesse público e agente fiscalizador da aplicação da lei. O MPDFT tem o dever de se manifestar oficialmente acerca de todo procedimento submetido à sua apreciação. Denomina-se "Feito" os procedimentos recebidos para apreciação, e "Ato praticado" ou "Movimento" as manifestações expedidas em função destes.

Segundo o último senso [1] do IBGE, o Distrito Federal tem uma população de 2,57 milhões de habitantes, de modo que a missão de defesa do interesse público envolve a apreciação de grande volume de processos judiciais e procedimentos extrajudiciais.

Para apreciar tal volume de procedimentos, a instituição conta com 379 membros ativos, entre promotores e procuradores, além de 1.780 servidores e quase 500 estagiários [2]. O órgão dispõe de 288 promotorias e 41 procuradorias, além da Procuradoria Geral de Justiça.

1.2.1 O volume de casos apreciados pelo órgão

Nesta seção apresentamos algumas estatísticas obtidas no portal da transparência do órgão [2], com o intuito de fornecer estimativa do volume de feitos apreciados e dos atos neles praticados. A Figura 1.1 apresenta apenas a quantidade de feitos novos recebidos por ano. Os feitos novos são aqueles conhecidos somente no ano apurado. Os demais são feitos conhecidos em anos anteriores que retornaram ao órgão para novas providências.

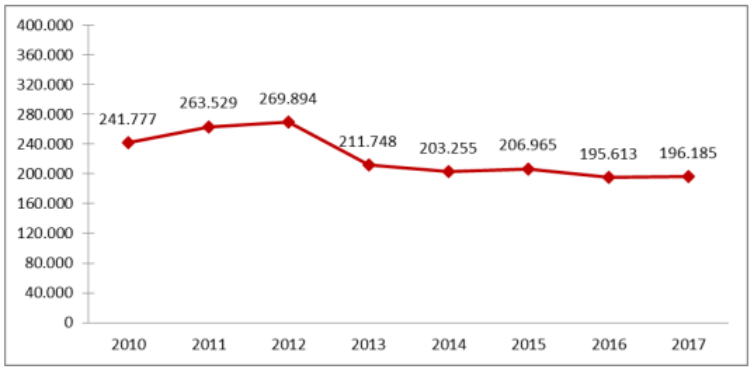
A Figura 1.2 apresenta o volume de todos os feitos recebidos em cada ano.

A Figura 1.3 apresenta o volume de atos praticados anualmente nos feitos apreciados. Observa-se uma média anual que ultrapassa a marca dos 550 mil, e que atualmente se aproxima dos 600 mil.

[1] `https://cidades.ibge.gov.br/brasil/df/brasilia/panorama`

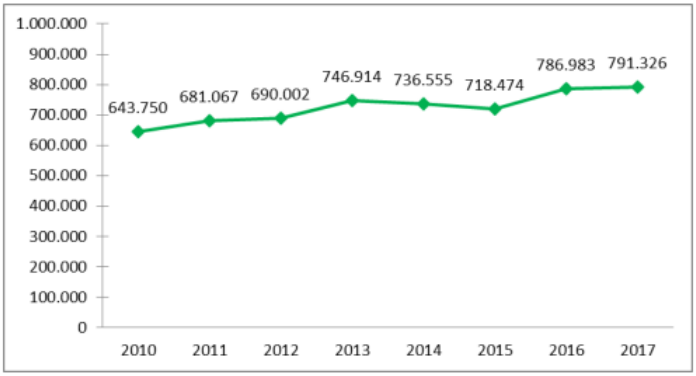
[2] `http://www.mpdft.mp.br/portal/pdf/unidades/corregedoria/AnuarioEstatistico2017.pdf`

Figura 1.1: Evolução do número de feitos judiciais novos recebidos entre 2010 e 2017



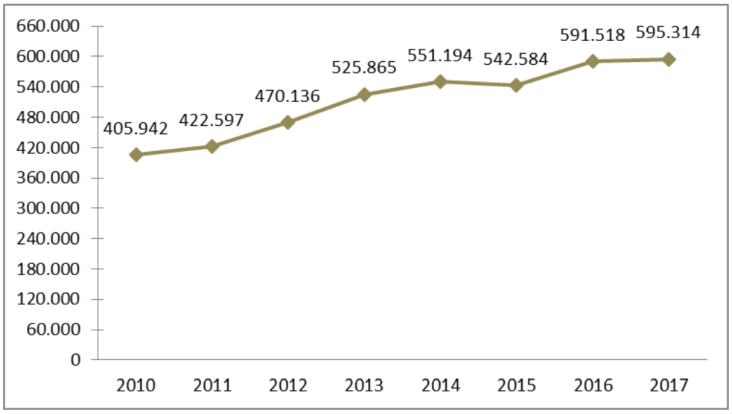
Fonte: MPDFT - Anuário estatístico (2017)

Figura 1.2: Evolução do número de feitos judiciais recebidos entre 2010 e 2017



Fonte: MPDFT - Anuário estatístico (2017)

Figura 1.3: Evolução do número de atos praticados em feitos judiciais entre 2010 e 2017



Fonte: MPDFT - Anuário estatístico (2017)

1.2.2 A variedade de casos apreciados pelo órgão

Existe uma grande variedade de tipos de feitos, atos praticados. Esta variedade pode ser verificada examinando os itens de classificação da taxonomia estabelecida pelo CNMP [3].

A taxonomia classifica os tipos de feitos, os assuntos que podem ser atribuídos aos feitos, e também os tipos de atos praticados. A taxonomia tem por objetivo facilitar o planejamento estratégico do MPU e permitir a interoperabilidade com o Poder Judiciário, que dispõe de taxonomia muito similar. A seguir descrevemos os itens de classificação estabelecidos pela taxonomia.

Classe

A classe corresponde ao tipo do feito, de modo que cada feito tem apenas uma classe. As classes são organizadas de modo hierarquizado, da mais abrangente para a mais específica. Como exemplo de classes, temos:

- *PROCESSO CRIMINAL > Recursos > Apelação Criminal*: refere-se a apreciação de apelação interposta contra uma sentença em um processo criminal;
- *PROCESSO CRIMINAL > Medidas Garantidoras > Habeas Corpus Criminal*: refere-se à apreciação de Habeas Corpus impetrado contra uma decisão de prisão preventiva.

Ao todo, a taxonomia estabelece 646 classes.

Assunto

O assunto especifica as matérias do direito relacionadas ao feito. O feito pode ter vários assuntos. Os assuntos também são organizados de forma hierárquica, do mais abrangente para o mais específico. Como exemplo de assuntos temos:

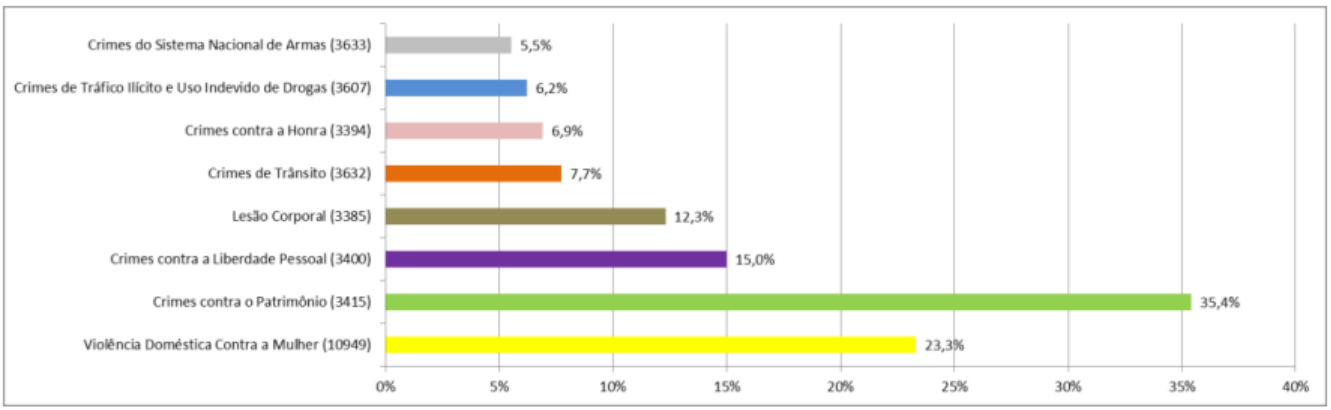
- *DIREITO PENAL > Crimes contra a vida > Homicídio Simples*;
- *DIREITO PENAL > Crimes contra o Patrimônio > Roubo.*

A Figura 1.4 apresenta a proporção de assuntos nos processos judiciais em 2017, considerados somente os assuntos mais comuns.

A taxonomia define assuntos de cunho finalístico e administrativo. Para o escopo deste trabalho consideramos somente os finalísticos, que somam 3378 assuntos.

[3] `https://sgt.cnmp.mp.br`

Figura 1.4: Proporção de Processos judiciais recebidos por assunto em 2017



Fonte: MPDFT - Anuário estatístico (2017)

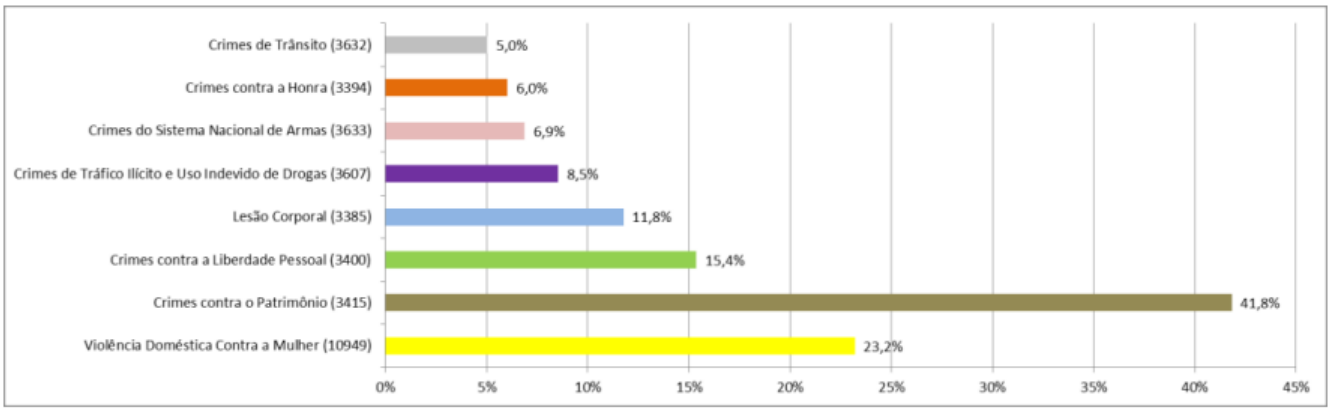
Tipo de Movimento

O tipo de movimento classifica os atos praticados no feito. A cada ato praticado é atribuído um único tipo. Trata-se também de uma classificação hierarquizada, do mais abrangente para o mais específico. Como exemplo, temos:

- *ATOS FINALÍSTICOS > Recursos > Razões > Apelação: quando o promotor ao tomar conhecimento da sentença de um processo interpõe apelação contra esta.*
- *ATOS FINALÍSTICOS > Alegações finais: manifestação proferida pelo Ministério Público após o encerramento da instrução processual.*

A Figura 1.5 apresenta a proporção de assuntos nos movimentos do tipo "Alegações Finais" em processos judiciais em 2017, considerados somente os assuntos mais comuns.

Figura 1.5: Alegações finais apresentadas em Processos judiciais em 2017 por assunto



Fonte: MPDFT - Anuário estatístico (2017)

A taxonomia define tipos de movimento de cunho finalístico e administrativo, sendo que somente os finalísticos são escopo deste trabalho. Estes somam 268 tipos.

1.2.3 O controle do acervo de casos

Os feitos são recebidos em mídia física (papel) e eletrônica. Os Feitos são classificados na ocasião do recebimento, com a classe e assuntos da forma como foram classificados pelo TJDFT.

Os feitos recebidos são registrados em um sistema informatizado que acumula todos os dados e documentos produzidos durante a apreciação dos feitos. O sistema executa a distribuição dos feitos entre as promotorias, controla os trâmites dos feitos, e apoia às atividades da corregedoria do órgão.

Desde que iniciaram os registros digitais dos feitos até agosto de 2017, acumulou-se um acervo de mais de 3.6 milhões de feitos. Os feitos anteriores a 2011 não estão classificados com a taxonomia estabelecida pelo CNMP, pois esta entrou em vigência em dezembro de 2010.

1.3 Definição do problema

Durante a apreciação de um feito, os especialistas que apoiam o promotor costumam pesquisar o acervo de atos praticados e selecionar alguns para embasar o novo ato praticado.

O sistema de informações do órgão mantem todo o acervo de atos praticados, porém não dispõe de serviço de pesquisa adequado a esta tarefa. Os serviços de pesquisas disponíveis são otimizados para a recuperação de feitos, e não de atos praticados. Criar um serviço de pesquisa capaz de recuperar atos praticados por item de classificação da taxonomia, também não é adequado, pois podem haver milhares de atos praticados dentro de um item de classificação. Por exemplo, consideremos que o especialista quer encontrar atos praticados do tipo 'contra-razões de apelação' em feitos cujo assunto é "Roubo" e classe 'Apelação'. Com estes atributos existem 1.423 atos praticados registrados. Alterando o assunto para 'furto', mantendo a mesma classe e tipo de movimento, temos 754 casos. Não é viável para o especialista examinar esta quantidade de documentos.

Por conta disso, no que se refere a pesquisa de atos praticados, os especialistas dispensam o uso do sistema de informações. Eles preferem manter repositórios particulares de documentos em arquivos digitais, e explorar este repositório utilizando o serviço de pesquisa disponível no sistema operacional de seus computadores. Promotorias que tratam feitos de mesma matéria costumam ter repositório de arquivos compartilhados.

Este serviço permite a pesquisa por palavra-chave, abrangendo tanto o nome do arquivo quanto seu inteiro teor. Além disso é possível restringir a pesquisa somente aos documentos de uma dada pasta do repositório.

Baseando-se nisto, as equipes organizam seus repositórios do modo que convier, visando facilitar a recuperação. Não há ato normativo que estabeleça regras de nomenclatura e organização dos arquivos.

Segundo Dabney [3], a eficiência deste método de recuperação de documentos está limitada pela habilidade do especialista indexar os documentos.

Além disso, o método de pesquisa utilizado neste serviço é desconhecido, cuja propriedade intelectual pertence ao fabricante, não sendo possível aprimorá-lo nem adequá-lo às necessidades do especialista. Segundo observado, o serviço recupera somente os documentos que contém as palavras-chave utilizada na pesquisa, em sua forma exata. O serviço não é capaz de recuperar documentos que contenham termos sinônimos ou o mesmo termo flexionado, nem mesmo considerar a polisemia, ou seja, uma mesma palavra-chave pode ter conceitos diferentes.

Outro problema com esta abordagem é a manutenção de repositórios alternativos ao repositório do sistema de informação do MPDFT, que além de exigir recursos de armazenamento, não estão sujeitos às mesmas restrições de acesso e segurança que o sistema de informação impõe.

1.4 O escopo do projeto

Considerando as dimensões do órgão, entendemos que não é viável realizar os experimentos abrangendo todas as promotorias e procuradorias.

Para reduzir o escopo do trabalho optamos por selecionar uma coleção de atos praticados de um único tipo de movimento, porque este é utilizado na formação de equipes de trabalho. Além disso preferimos as coleções em que ocorrem uma boa variedade de assuntos abordados, e que tenham volume suficiente para os objetivos do trabalho.

Diante destes critérios, selecionamos as coleções das Procuradorias de Justiça Criminais e Criminais Especializadas. As procuradorias atuam principalmente na emissão de pareceres em processos criminais que vão à segunda instância. O acervo destas procuradorias abrange grande parte das matérias criminais, o que é conveniente. Além disso, o acervo é volumoso. As procuradorias criminais tem um acervo de 16.244 atos praticados e as procuradorias criminais especializadas um acervo de 7.665, totalizando 23.909. Estas coleções não contém documentos com imposição de sigilo ou segredo de justiça.

1.5 Justificativa do projeto

A pesquisa por atos praticados é parte importante na atividade das promotorias e procuradorias. O aprimoramento da pesquisa ao acervo de atos praticados pode ter impacto positivo, tanto no tempo de resposta do órgão, quanto na qualidade.

Além disso, este trabalho estuda alternativas de pesquisa ao acervo que dispensam a manutenção e indexação dos documentos nos repositórios de forma manual, além de poupar recursos de armazenamento de dados e viabilizar a imposição das mesmas regras de acesso impostas pelo sistema de informações do órgão.

Entre as técnicas de recuperação de informações avaliadas neste trabalho estão as técnicas de busca semântica. Estas técnicas permitem recuperar documentos relevantes que não contém os termos de busca exatamente como constam do texto de busca. Com o uso destas técnicas espera-se que a qualidade do resultados obtidos se mantenha para uma dada necessidade de informação, mesmo que a redação do texto de pesquisa varie entre os especialistas.

É relevante pontuar que a adoção de técnicas de recuperação conhecidas nos permite medir sua qualidade, viabilizando a evolução e aprimoramento de sua implementação e aplicação de novas técnicas, de acordo com as necessidades e particularidades de cada equipe.

1.6 Objetivo geral

Aplicar técnicas de mineração de textos para auxiliar a recuperação de atos praticados similares no MPDFT.

1.7 Objetivos específicos

- Realização de prova de conceito de software para recuperação de atos praticados pelas Procuradorias de Justiça Criminal e Criminal Especializada, com a aplicação das técnicas de mineração de textos.
- Desenvolver protótipo de software para extração e análise de características dos atos praticados e suas peças processuais para a construção dos sistemas de recuperação de informações.
- Comparar a performance dos modelos induzidos.

1.8 Hipóteses de pesquisa

- O uso de técnicas de busca semântica no teor dos documentos aumenta a performance de modelos de induzidos em mineração de dados para a recuperação de atos praticados.
- Um modelo enriquecido com os dados cadastrais dos feitos é mais eficiente na localização dos casos semelhantes.
- Um modelo enriquecido com as citações às normas jurídicas presentes nos documentos é mais eficiente na recuperação de atos praticados.

1.9 Contribuições

O trabalho estabelece uma abordagem e artefatos de software para apoiar a implementação de serviço de pesquisa de atos praticados em várias promotorias ou procuradorias do órgão. Estes serviços podem ser adaptados às necessidades e particularidades de cada equipe porque as técnicas utilizadas são conhecidas e estarão sob domínio do órgão. Além disso, os serviços dispensam a tarefa de indexação manual dos documentos e a manutenção de repositórios paralelos dentro do órgão.

A metodologia proposta permite comparar outras técnicas de recuperação de informações e avaliar de modo objetivo se houve ganho real de qualidade, orientando a evolução dos serviços de pesquisa.

Como atos praticados são documentos jurídicos, a metodologia deve ser capaz de apoiar o desenvolvimento de serviços de recuperação de informações jurídicas em outras organizações.

Como inovação tecnológica, este trabalho introduz a utilização de técnicas de modelagem de tópicos e mineração de textos no MPDFT. Espera-se que o uso destas técnicas contribua para a celeridade e aprimoramento do atendimento às demandas da população por parte do MPDFT.

1.10 Estrutura da dissertação

Essa dissertação está estruturada da seguinte forma. O Capítulo 2 apresenta a fundamentação teórica e trabalhos correlatos. O Capitulo 3 apresenta a metodologia e o desenvolvimento de cada etapa do trabalho. O Capítulo 4 apresenta os resultados obtidos e o Capítulo as conclusões. O Apêndice A e B detalham os experimentos referentes às procuradorias criminais e criminal especializadas, respectivamente. O Apêndice C apresenta exemplos de pareceres utilizados nos experimentos. Todas as figuras e tabelas que

não apresentarem fonte foram produzidas pelo autor dessa dissertação durante a pesquisa realizada.

Capítulo 2

Fundamentação Teórica

2.1 Modelos de Recuperação de Informações

Os Modelo de recuperação de informações (MRI) são necessários quando o usuário precisa de uma pesquisa mais complexa do que se pode fazer utilizando um sistema de recuperação de dados estruturados.

O objetivo principal de todo MRI é recuperar o conjunto de documentos relevantes para uma dada pesquisa feita pelo usuário do modelo. Quanto menos documentos irrelevantes estiverem contidos no conjunto recuperado, melhor. Tipicamente, a pesquisa ou *Query* é expressa por um texto, mas pode conter dados.

Segundo Baeza-Yates e Ribeiro-Neto [4], um MRI pode ser definido da seguinte maneira:

$$[D, Q, F, R(q_i, d_j)] \tag{2.1}$$

Onde:

- D é o conjunto das representações dos documentos;
- Q é o conjunto das representações das *queries*;
- F é um *framework* de modelagem das representações das necessidades de informações e dos documentos, e seus relacionamentos;
- $R(q_i, d_j)$ é uma função de ranking que associa um escore para a relação entre a *query* $q_i \in D$ e o documento $d_i \in Q$. Este escore é utilizado para ordenar documentos por relevância.

Para a construção de um MRI, os documentos e as *queries* precisam ser transformados em uma representação simplificada para a recuperação de informações. Tipicamente, a

transformação remove preposições, artigos e demais palavras que sabidamente não contribuem para a recuperação de informações (*stopwords*). Além disso, é comum substituir as palavras por seu radical (*stemming*).

O *framework* é um método ou técnica que permite a comparação entre as *queries* e os documentos, e fornece um meio de utilizar a função de *ranking* para ordenar os documentos por relevância.

As abordagens mais comuns para modelagem de um *framework* são a booleana, os modelos vetoriais e os probabilísticos.

A abordagem booleana é composta de conjuntos de documentos e de operações tradicionais de conjuntos. Na abordagem de modelos vetoriais, os documentos e as *queries* são representados como vetores e de operações tradicionais da álgebra linear. A abordagem probabilística utiliza a distribuição de probabilidades de ocorrência dos termos nos documentos e nas *queries* e operações baseadas no teorema de Bayes.

Neste trabalho utilizaremos as abordagens vetoriais *Bag of Words* (BOW) com *Term Frequency - Inverse Document Frequency* (TF-IDF), e *Latent Semantic Indexing* (LSI), e as probabilísticas *Latent Dirichilet Allocation* (LDA) e *Best Match 25* (BM25).

2.2 *Bag of Words* e *TF-IDF*

O *Bag of Words* (BOW) é uma representação de um documento por um vetor cujas dimensões correspondem às frequências de cada termo do vocabulário que será considerado pelo *framework*. Esta representação despreza a ordem em que as palavras estão dispostas nos documentos. Outra característica é que os termos mais frequentes no documento serão tidos como mais informativos, o que na prática nem sempre corresponde às necessidades do especialista. Para corrigir isto, pode-se utilizar a função *Term Frequency - Inverse Document Frequency* (TF-IDF) [5], que reduz a importância dos termos que são frequentes em muitos documentos, e aumenta a importância dos termos que forem muito frequentes em um grupo pequeno de documentos.

A função TF-IDF pode ser definida como:

$$tfidf(d,t) = tf(d,t) \times log\Big(\frac{|D|}{df(t)}\Big) \tag{2.2}$$

Onde:

- $tf(d,t)$ é a frequência do termo t no documento d;
- $|D|$ é o tamanho médio dos documentos da coleção;
- $df(t)$ é a quantidade de documentos em que o termo t ocorre.

A representação BOW ponderada com TF-IDF é uma abordagem largamente utilizada em sistemas de recuperação de informações.

Existem variantes do TF-IDF, porém não as abordaremos neste trabalho.

2.3 *Latent Semantic Indexing*

A recuperação de informações baseada na comparação de vetores BOW com TF-IDF não é capaz de reagir a certos fenômenos linguísticos como a sinonimia e polisemia. Sendo assim, a relevância dos documentos recuperados depende da habilidade em se formular a *query* contendo termos em comum com os documentos desejados. Esta limitação é indesejável na prática, porque os documentos podem conter narrativas diferentes acerca do mesmo conceito.

Deerwester propôs o *Latent Semantic Indexing* (LSI)[6] para abordar esta limitação a partir da comparação de conceitos presentes na *query* e nos documentos, ao invés de comparar os termos em comum.

A técnica consiste na aplicação do *Singular Value Decomposition* (SVD) para decompor a matriz termo-documento construída a partir de todos os vetores BOW da coleção de documentos organizados como colunas, e os termos do vocabulário como linhas. A decomposição SVD é definida da seguinte forma:

$$M = K \cdot S \cdot D^T \tag{2.3}$$

Onde:

- M é a matriz termo-documento;
- K é a matriz de autovetores derivada da matriz de correlação termo-termo obtida por $C = M \cdot M^T$;
- D^T é a matriz de autovetores derivada da transposta da matriz documento-documento, dada por $M^T \cdot M$;
- S é a matriz de autovalores de dimensões $r \times r$ onde r é o rank da matriz M.

Preservam-se somente os s maiores autovalores, juntamente com suas colunas correspondentes em K e D^T, e desprezam-se os demais autovetores, obtendo a matriz $M_s = K_s \cdot S_s \cdot D_s^T$, com $s < r$. Deerwester advoga que os conceitos semânticos presentes na coleção de documentos podem ser capturados pelos autovetores. O valor de s precisa ser ajustado de modo a ser grande o suficiente para preservar a estrutura original contida nos dados, e pequeno o suficiente para eliminar a parte não relevante. Com relação às técnicas de modelagem de tópicos, os autovetores correspondem aos tópicos.

A representação vetorial da *query* é obtida juntando-se seu vetor BOW como uma linha em D. Assim teremos o vetor D_q que representa a *query*, e então podemos verificar sua similaridade com os vetores dos documentos utilizando a similaridade cosseno [6].

Outra característica interessante da abordagem é a redução da dimensionalidade de M para M_s, que favorece a performance e economia no uso de recursos computacionais.

Hofmann [7] propôs uma versão probabilística do LSI conhecida como *Probabilistic Latent Semantic Indexing* (pLSI), que dispensa a manutenção das matrizes para a recuperação de informações.

2.4 *Latent Dirichlet Allocation*

O LDA é um modelo de tópicos probabilístico generativo para documentos, proposto por Bley et all [8].

A ideia básica é que os documentos são representados como misturas aleatórias sobre tópicos latentes de K, onde cada tópico é caracterizado por uma distribuição sobre palavras. Em uma formulação simplificada do LDA, a probabilidade dos termos é parametrizada por uma matriz β com dimensões $K \times W$, onde W é o tamanho do vocabulário considerado. Um tópico k $(1 \leq k \leq \mathcal{K})$ é uma distribuição discreta sobre palavras com vetor de probabilidade β_k. Cada documento d_j $(1 \leq j \leq D)$, onde D é o número de documentos, mantém uma distribuição separada θ_j que descreve a contribuição de cada tópico. As implementações de algoritmos de inferência LDA tipicamente usam Dirichlet simétrico antes de $\Theta = \{\theta_1, \ldots, \theta_D\}$, no qual o parâmetro de concentração α é fixo. Uma distribuição de tópico de um documento d_j e uma palavra w_i é associada em uma variável de distribuição $z_{j,i}$.

Dados os parâmetros α e β, o problema computacional do LDA é inferir a distribuição conjunta de uma mistura de tópicos Θ, dada por $p(\Theta, z, w|\alpha, \beta)$. Uma vez inferida a distribuição do tópico do documento Θ, podemos interpretar cada distribuição θ_j como uma representação reduzida de um documento d_j. Da mesma forma, podemos aplicar uma consulta q ao modelo de LDA inferido e induzir uma distribuição representativa θ_q. A similaridade entre uma consulta q e um documento d_j é calculada pela similaridade cosseno entre os vetores de distribuição θ_j e θ_q.

Tanto o LSI quanto o LDA tem como parâmetro a quantidade de tópicos a utilizar. Neste trabalho experimentaremos uma faixa de valores para estimar qual o valor adequado para cada aplicação e modelo.

2.5 BM25

Na definição de MRI fornecida por Baeza-Yates e Ribeiro-Neto, LSI e LDA podem ser utilizados como *frameworks*. A técnica *Best Match 25* (BM25) [9], também conhecida como Okapi BM25, trata-se de uma função de escore probabilística. A função atribui um escore entre a *query* e os documentos, baseando-se nos termos em comum entre eles. A função pode ser descrita da seguinte maneira:

$$score(D, Q) = \sum_{i=1}^{n} IDF(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{avgdl})} \tag{2.4}$$

Onde:

- $f(q_1, D)$ é a frequência do termo q_1 da *query* no documento D;
- $|D|$ é o tamanho do documento D, em palavras;
- $avgdl$ é o tamanho médio em palavras de toda a coleção de documentos;
- b e k_1 são parâmetros livres, ajustáveis. Na ausência de um procedimento de otimização, costuma-se utilizar $k_1 \in [1.2, 2.0]$ e $b = 0.75$

A função IDF é definida como:

$$IDF(q_1) = log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \tag{2.5}$$

Onde:

- N é o total de documentos na coleção;
- $n(q_i)$ é o total de documentos que contém o termo q_i da *query*.

A função BM25 é muito utilizada em MRI, e tem muitas variações. Atualmente vem sendo utilizada como linha de base de comparação para os novos MRI propostos pela comunidade científica [4].

Métricas de avaliação

Entre as técnicas mais tradicionais para avaliação de MRIs temos o *Precision* e *Recall*. Seja R o conjunto dos documentos relevantes para uma dada *query*, e A o conjunto dos documentos recuperados pelo MRI. Então podemos definir o *Precision* e *Recall* da seguinte forma:

- *Precision* consiste na fração dos documentos recuperados que são relevantes.

$$Precision = \frac{R \cup A}{A} \tag{2.6}$$

- *Recall* consiste na fração dos documentos relevantes que foram recuperados.

$$Recall = \frac{R \cup A}{R} \tag{2.7}$$

Em muitos casos, não se espera que o usuário de um sistema de recuperação de informações examine todos os documentos recuperados, de forma que é muito comum utilizar a métrica *precision* fixando uma dada posição no *ranking*, como por exemplo 10 (neste caso, utiliza-se a notação p@10).

Estas medidas nos fornecem a performance do modelo para uma *query*. Quando se pretende comparar dois ou mais MRI, utilizamos um conjunto de *queries*, e precisamos então de um valor que resuma as medidas de todas as *queries* utilizadas na avaliação. Pode-se utilizar a *Precision* média de todas as *queries*.

Outra métrica muito popular é a *Mean Average Precision* (MAP). O MAP para uma *query* é definida da seguinte maneira:

$$MAP_i = \frac{1}{|R_i|} \sum_{k=1}^{|R_i|} P(R_i[k]) \tag{2.8}$$

Onde:

- $|R_i|$ é a quantidade de documentos relevantes para a *query* i;
- $P(R_i[k])$ é a precisão quando o documento relevante k é encontrado no ranking de documentos retornado pelo MRI.

O escore MAP para o modelo, é obtido a partir da média dos escores MAP de cada *query* utilizada na avaliação.

O MAP considera todos os documentos igualmente relevantes para uma *query*. Quando se quer diferenciar os modelos que recuperam os documentos de maior relevância nas primeiras posições do ranking, então teremos que utilizar uma noção de relevância que não seja binária, e uma métrica apropriada.

A métrica *Normalized Discounted Cumulated Gain* (NDCG) permite medir modelos com uma medida de relevância em graduações, como por exemplo, variando de 0 a 2, de modo que 0 significa "sem relevância" e 2, "muito relevante". Além disso, esta métrica aplica um desconto à medida que a posição no ranking aumenta, para penalizar modelos que posicionam documentos relevantes distantes das primeiras posições.

Definimos a NDCG da seguinte forma:

- Seja R_j o ranking retornado pela *query* j, e G_j o vetor de escores de relevância para os documentos pertencentes a R_j, na mesma ordem. Chamemos G_j de vetor de ganho para a *query* j.
- Seja DCG_j o vetor de ganho acumulado com desconto, definido da seguinte forma:

$$DCG_j[i] = \begin{cases} G_j[1], \text{ se } i = 1; \\ \frac{G_j[i]}{\log_2 i} + DCG_j[i-1], \text{ se } i > 1 \end{cases}$$

- Seja IG_j o vetor de ganho ideal, correspondente a G_j ordenado por ordem decrescente de escore, e $IDCG_j$ o vetor de ganho ideal com desconto, obtido da mesma forma que o vetor $DCG_j[i]$.

Então temos o $NDCG_j[i]$ definido como:

$$NDCG_j[i] = \frac{DCG_j[i]}{IDCG_j[i]} \tag{2.9}$$

Sendo i a posição do ranking. O escore $NDCG$ para um modelo é obtido a partir da média do escore $NDCG$ de todas as *queries* utilizadas na avaliação.

2.6 Trabalhos correlatos

Nesta seção faremos um breve resumo sobre as principais linhas de pesquisa referentes à recuperação de informações no contexto jurídico.

2.6.1 *Legal Information Retrieval*

Entre os trabalhos avaliados, destacamos o de Van Opijnem[10], que aborda as principais características que se espera de uma solução de recuperação de informações legais. Van Opjinem conceitua a relevância de documentos legais em seis dimensões:

- Relevância algorítmica: trata-se da semelhança entre o texto de pesquisa e o texto dos documentos atribuída por algoritmos, como ocorre na maioria dos sistemas de recuperação de informações.
- Relevância tópica: trata-se da correspondência entre assunto ou tópico pesquisados com e os documentos recuperados.

- Relevância Bibliográfica: trata-se do alinhamento e coerência bibliográfica dos documentos recuperados. Esta dimensão da relevância mede a pertinência dos documentos no que se refere às legislações e normas referenciadas.
- Relevância Cognitiva: trata-se da opinião ou preferência do pesquisador pelo modo como a questão legal é abordada nos documentos recuperados.
- Relevância Situacional: trata-se das características da atividade desempenhada pelo especialista para as quais busca suporte nos documentos. Em uma dada situação, o especialista busca por documentos que contenham argumentos de acusação, e em outras por argumentos de defesa;
- Relevância de domínio: refere-se à importância ou destaque atribuído ao documento pela comunidade jurídica.

Este trabalho aborda a relevância algorítmica ao aplicar um conjunto de algoritmos para estimar a similaridade entre os documentos e a *queries*. A relevância tópica é abordada com o uso da taxonomia de assuntos e classe associada a cada documento. A relevância bibliográfica é abordada verificando as citações em comum entre as *queries* e os documentos. A relevância situacional é tratada a partir do tipo de movimento associado ao documento. O tipo de movimento informa o contexto situacional da atividade que o especialista desempenha. As demais dimensões de relevância não serão abordadas neste trabalho.

2.6.2 *Legal Information Retrieval* com apoio de ontologias

Na literatura há diversas abordagens para estimar a relevância dos documentos legais. Entre as abordagens mais comuns está o uso de ontologias para mapear os conceitos presentes nos escritos jurídicos. Com um objetivo muito semelhante ao deste trabalho, Quaresma et all [11] propôs um sistema de perguntas e respostas acerca de documentos legais emanados pela Suprema Corte, Tribunais superiores e Procuradoria Geral da República Portuguesa. Quaresma utilizou um *framework* que combina a recuperação de informações tradicional com a tradução semântica do texto de pesquisa (*queries*) e com o uso de ontologias. A avaliação do *framework* consiste em verificar o percentual de documentos recuperados que contém a resposta às perguntas submetidas. Em seus experimentos, Quaresma reporta que de 200 documentos recuperados pelo *framework*, cerca de 33% continham a resposta correta.

Em nossa proposta, descartamos o uso de ontologias. Como o MPDFT aborda um grande espectro de temas jurídicos, estimamos ser inviável estabelecer taxonomias para um contexto tão amplo, considerado o curto e médio prazo. Além disso, as frequen-

tes alterações legislativas e volume jurisprudencial exigiriam constante atualização das ontologias. Por este motivo, preferimos abordagens que generalizem um pouco mais a necessidade de recuperação de informações nas várias equipes de trabalho do MPDFT.

2.6.3 *Legal Information Retrieval* com apoio de rede de citações

Outra abordagem comum consiste em explorar as citações ou referências legais encontradas nos documentos. As citações informam o relacionamento do documento com as normas legais, e são utilizadas para enriquecer as soluções de recuperação de informações.

Entre os trabalhos semelhantes a este destacamos o de Raghav et all (2015) [12]. Raghav aplicou e adaptou a técnica de clusterização *K-Means* para agrupar documentos semelhantes utilizando sua rede suas citações.

O autor parte do pressuposto que documentos que contém citações em comum são semelhantes. Além da rede de citações, o modelo é enriquecido com escore de semelhança entre cada parágrafo das sentenças. Os parágrafos são vetorizado utilizando a técnica BOW com TF-IDF.

A comparação é feita utilizando a similaridade do cosseno, de modo que se um par de sentenças contiver ao menos 3 parágrafos com escore maior que 0,5, então considera-se que existe uma relação entre as sentenças.

A base de dados utilizada foram os julgamentos expedidos pela Suprema Corte da Índia, considerando o período de 1970 até 1993, contendo 3.738 julgamentos. Os resultados foram avaliados com o apoio de especialistas, que atribuíram escore de similaridade, de 0 a 10, para 47 pares de documentos que o modelo apontou como sendo semelhantes. Escores acima de 5 foram considerados como verdadeiro-positivo e, abaixo de 5, falso-positivo.

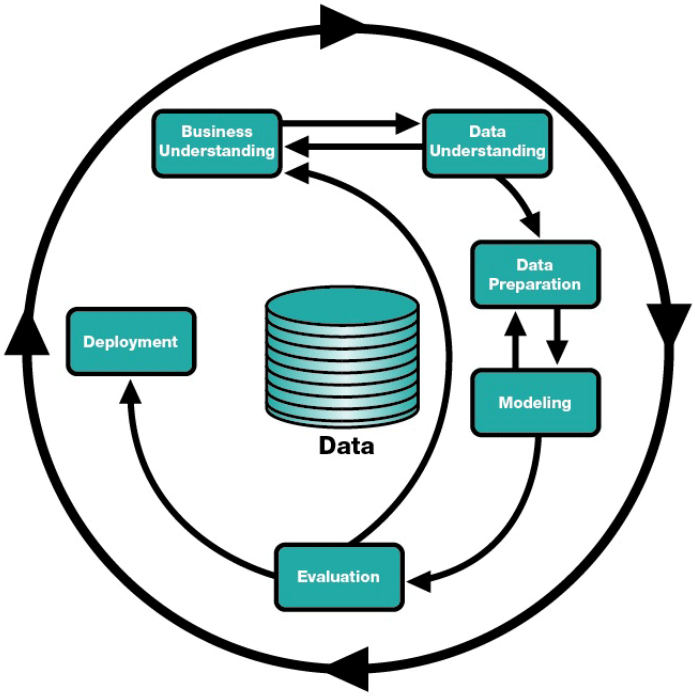
Com esta métrica, o modelo utilizando somente as citações atingiu valor médio de 86% de *precision*, 73% de recall e 79% na medida F1. Enriquecendo o modelo com as ligações derivadas dos parágrafos semelhantes os escores para estas métricas foram 81%, 75% e 80%, respectivamente. Os resultados obtidos neste trabalho reforçam a hipótese de que as citações são fonte de informação relevante para estimativa de similaridade entre os casos.

2.7 CRISP-DM

O CRISP-DM [13] é uma metodologia utilizada em projetos de mineração de dados. A metodologia propõe a divisão do processo de mineração de dados em 6 fases, que

podem ocorrer em ciclos sucessivos para o aperfeiçoamento do modelo, de acordo com o apresentado na Figura 2.1.

Figura 2.1: Diagrama do modelo CRISP-DM.



Fonte: `http://crisp-dm.eu/reference-model`

De forma resumida, as fases são as seguintes:

- fase de entendimento do negócio (*Business Understanding*). Nesta fase, o objetivo do trabalho é detalhado e refinado, as principais características do problema são analisadas, e os dados preliminares são levantados. Esta fase pode ser revisitada quando a avaliação do modelo apontar a necessidade de refinar alguns pontos do objetivo;
- fase de entendimento dos dados (*Data understanding*). Nesta fase, os dados disponíveis são estudados com o objetivo de elencar as primeiras ideias sobre como abordar o problema com os dados disponíveis. Hipóteses são elaboradas para verificação nas fases subsequentes. Nesta fase também afere-se a qualidade dos dados e a necessidade de coletar dados adicionais;
- fase de preparação dos dados (*Data preparation*). Nesta fase, os dados são pré-processados com o objetivo de produzir os indicadores e as características relevantes para o problema, que muitas vezes não estão registrados de forma direta nos dados

disponíveis. Além disso, faz-se o tratamento dos dados eventualmente faltantes requeridos para treinamento do modelo;

- fase de modelagem (*Modeling*). Nesta fase, os modelos de mineração de dados são produzidos e testados. Os parâmetros do modelo são ajustados para seleção de um pequeno grupo de modelos para a fase de avaliação;
- fase de avaliação (*Evaluation*). Nesta fase os modelos selecionados na fase de modelagem são revisados e avaliados em relação aos objetivos propostos. Algumas simulações são feitas com dados não considerados pelo modelo com o objetivo de aferir a qualidade do modelo.
- fase de implantação (*Deployment*). Nesta fase, os modelos aprovados são implantados e usados no ambiente real. O monitoramento dos modelos deve ser realizado com o objetivo de levantar dados para revisão e atualização futura do modelo.

Capítulo 3

A identificação de casos similares

Neste capítulo detalharemos a metodologia utilizada e cada uma das etapas do trabalho.

3.1 Metodologia

Primeiramente, fizemos o levantamento dos trabalhos relacionados e a revisão do estado da arte, atividade que perdurou durante todo o planejamento e execução do trabalho. As pesquisas foram orientadas pelo tema "recuperação de informações jurídicas", dentro linha de pesquisa sobre recuperação de informações. Para localizar trabalhos correlatos, utilizamos o termo "legal information retrieval" como parâmetro de pesquisa nas ferramentas de busca Web of Science[1], Microsoft Academic[2] e Google Acadêmico[3], considerando o período de 1980 até 2018. Os textos selecionados para estudo foram escolhidos de acordo com a similaridade com problema abordado neste trabalho, considerando também a quantidade de citações registradas. O captítulo 2 detalha os principais trabalhos selecionados.

Em seguida, procedemos com o entendimento do negócio e dos dados. Estudamos os dados e documentos disponíveis através de análises exploratórias e realizamos entrevistas com especialistas com o objetivo de entender em detalhes como o problema de recuperação de informações é tratado na instituição, e como este procedimento pode ser aprimorado. Como pretendemos que a solução seja aplicada em várias equipes do MPDFT, decidimos selecionar duas coleções de documentos provenientes de equipes de especialistas que desempenham a mesma atividade, porém especializadas em matérias diferentes. Com isso, poderemos avaliar a capacidade de generalização das técnicas de recuperação de informação estudadas.

[1] `http://wokinfo.com/`
[2] `https://academic.microsoft.com`
[3] `https://scholar.google.com.br/`

Selecionadas as coleções de documentos, desenvolvemos os procedimentos para a extração dos respectivos dados e documentos junto ao sistema de controle de feitos do órgão. Além da extração, foram desenvolvidas rotinas de transformação dos dados e pré-processamento dos documentos.

Em seguida, estabelecemos o método de avaliação dos modelos de recuperação de informações. A avaliação dos modelos requer um conjunto de *queries* e do conjunto de documentos relevantes para cada *query*. Como as coleções de documentos são numerosas (as duas coleções somam cerca de 23 mil documentos), consideramos inviável estabelecer a relevância de cada documento em relação às *queries*. Desta forma optamos por construir as bases de referência com documentos obtidos a partir da amostragem de cada coleção. Para cada base de referência, elaboramos uma lista de 10 *queries*, e para cada documento da amostra atribuímos seu nível de relevância em relação a cada *query*. A relevância é medida em 3 níveis: não relevante (0), relevante(1) e muito relevante(2). Além disso, para cada query temos a lista de assuntos e a lista de citações pertinentes, a serem utilizada na avaliação dos modelos enriquecidos. Além destas 10 *queires* produzimos mais 2 versões de cada *query*, modificando apenas seu texto e mantendo a lista de assuntos e citações, de modo que cada base de referência passou a contar com 30 queries. Isto foi feito com o objetivo de avaliar a robustez dos modelos em relação às diferentes formas de se descrever a mesma necessidade de informação. As bases de referência foram construídas com o apoio de especialistas de cada equipe.

Estabelecidas as bases de referência, procedemos com a indução dos modelos de recuperação de informações a partir da base de treinamento extraída de cada coleção. A base de treinamento é composta pelos documentos que não fazem parte da base de referência. Para cada base de treinamento induzimos modelos baseados nas técnicas TF-IDF, BM25, LSI e LDA. As técnicas LSI e LDA requerem como parâmetro a quantidade de tópicos. Para este parâmetros adotamos 8 faixas de valores (50, 100, 150, 200, 250, 300, 350 e 400), resultando em 8 modelos LSI e 8 modelos LDA. Desta forma temos 18 modelos por coleção de documentos.

Concluída a indução dos modelos executamos sua avaliação. Medimos a performance dos modelos usando a métrica NDCG. Primeiramente medimos os modelos sem enriquecimento e os comparamos com o teste de Nemenyi, que revela os modelos cuja performance se destacaram dos demais. Em seguida, para cada modelo avaliamos se houve alteração significativa da performance entre suas versões sem enriquecimento, enriquecida com assunto, enriquecida com citações, e enriquecida com assuntos e citações. Para isto, usamos novamente o teste de Nemenyi para verificar se alguma versão do modelo se destaca em relação as demais.

A partir dos resultados da avaliação dos modelos verificamos as hipóteses do trabalho

e selecionamos um modelo para o desenvolvimento do protótipo para uso dos especialistas. Com o protótipo, os especialistas podem proceder com a recuperação de informações abrangendo toda a coleção de documentos de sua equipe, e não somente a base de referência. O protótipo registra as *queries* submetidas bem como a avaliação dos resultados fornecida pelos especialistas. Estes registros permitirão o aprimoramento das bases de referência e da avaliação dos modelos.

3.2 Entendimento do Negócio

A prova de conceito produzida abrange as equipes das procuradorias criminais e criminais especializadas. Nesta seção apresentaremos a organização e as atividades das procuradorias do órgão pertinentes a este trabalho.

3.2.1 As procuradorias de justiça criminal e criminal especializada

O MPDFT tem 14 procuradorias criminais, e 8 procuradorias criminais especializadas, cada uma representada pelo procurador titular. As procuradorias oficiam junto às câmaras e turmas criminais do TJDFT. A atribuição de cada procuradoria é definida por resolução do Conselho Superior do MPDFT [4].

As procuradorias de justiça criminal

As 14 procuradorias criminais se revezam no ofício nas sessões das 3 turmas criminais do TJDFT. As procuradorias criminais emitem pareceres sobre as apelações interpostas nos processos criminais, provenientes das varas criminais e nas varas de entorpecentes e contravenções penais.

O parecer apresenta o entendimento do procurador de justiça acerca das questões suscitadas pelas partes contra a decisões proferidas pelos magistrados em 1o grau. O parecer é juntado ao processo e apreciado pela turma ou câmara criminal responsável para o julgamento dos recursos.

As procuradorias de justiça criminal especializada

Das 8 procuradorias criminais especializadas, as 4 primeiras se revezam no ofício nas sessões das câmaras criminais do TJDFT, e as 4 últimas se revezam no ofício das 3

[4] `http://www.mpdft.mp.br/portal/pdf/unidades/conselho_superior/resolucoes_vigor/Res_64_alterada_pela_Res_240.pdf`

turmas criminais do TJDFT. As procuradorias criminais especializadas emitem pareceres sobre os seguintes recursos:

- *Habeas corpus*;
- Contrarrazões em recurso constitucionais;
- Agravos de instrumento;
- Apelações.

As procuradorias criminais especializadas atuam em processos provenientes do Tribunal do Juri, da Vara de Delitos de Trânsito, da Auditoria Militar, e nos processos referentes às leis 8.078/90 (direitos do consumidor) e 6.766/79 (parcelamento do solo urbano).

Os pareceres

Os pareceres analisam os recursos interpostos em processos em que o MPDFT é parte. Os recursos contestam decisão de um magistrado de 1a instância. O recurso contém as razões pelas quais a parte pede que a decisão do magistrado seja reformada, e podem ser interpostos por qualquer das partes do processo. Às partes requeridas cabe o direito de oferecer contrarrazões, que trata-se de defesa contra as razões do recurso interposto em seu desfavor. No parecer, o procurador analisa se os pré-requisitos para conhecimento do recurso foram preenchidos, e analisa o mérito das razões de todos os recursos e suas respectivas contrarrazões. O parecer é juntado ao processo e devolvido ao TJDFT para julgamento pela turma ou câmara criminal responsável.

3.3 Entendimento dos Dados

Nesta seção apresentaremos os detalhes das coleções de documentos utilizadas. Decidimos utilizar pelo menos duas coleções para verificar se a performance dos modelos alteram quando utilizados em conjunto de dados distintos.

Os dados colhidos referem-se aos pareceres produzidos pelas procuradorias criminais e criminais especializadas compreendidos nos período de 2012 até 2017. Atos praticados anteriores a 2012 tem dados cadastrais com taxonomia diferente, própria do MPDFT, haja vista que a taxonomia do CNMP, porque esta se tornou vigente a partir de 2012. Além disso, não puderam ser estudados os pareceres provenientes de feitos sigilosos ou com segredo de justiça.

3.3.1 Os pareceres das procuradorias criminais

Os pareceres das procuradorias criminais concentram-se em grande parte na apreciação de apelações, conforme se vê na Figura 3.1.

Figura 3.1: PC - Distribuição de pareceres por classes.

Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

No Apêndice C há um exemplo de parecer de apelação.

Os assuntos relativos a crimes contra o patrimônio são muito frequentes, como furto, roubo, receptação e estelionato. Outro grupo que pode se observar são os casos relativos posse e comércio de entorpecentes, os de posse ilegal de armas de fogo, e de violência doméstica.

Figura 3.2: PC - Distribuição de pareceres por assunto

Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

Os pareceres de apelações normalmente seguem o seguinte roteiro:

1. Resumo do caso: esta seção relaciona os crimes pelos quais os réus foram condenados, e quais são os seus pedidos.

2. Análise dos pré-requisitos: nesta seção são verificados o atendimento dos pré-requisitos necessários para o conhecimento do recurso, como prazos, legitimidade das partes, e demais critérios.

3. Análise do mérito: Apresenta-se posicionamento do procurador sobre caso, se concorda ou discorda das alegações e sua fundamentação. Na fundamentação são utilizados recortes dos depoimentos e de documentos da fase inquisitória, recortes da sentença e transcrições de jurisprudências que reforçam o entendimento dos magistrados em casos similares anteriores.

4. Considerações finais: Resumo do parecer e posição final sobre o deferimento das reclamações.

Na ocasião da redação da minuta do parecer, o especialista procura por casos anteriores em que as partes alegaram as mesmas coisas. Por exemplo, no caso de crime de receptação, é muito comum o réu alegar que não sabia da origem ilícita dos bens adquiridos. Se o especialista averiguar que a alegação não procede, o ele precisará não só demonstrar a partir dos autos do processo que o réu não provou o que alega, e apresentar jurisprudências que afirmam que o ônus de provar o desconhecimento da má procedência dos bens é do réu.

As partes específicas do caso, como detalhes dos fatos, pessoas, locais, diálogos não costumam ser alvo de pesquisas. As pesquisas parecem ser orientadas em função das alegações de cada parte, onde a argumentação pode ser aproveitada para as mesmas alegações em outros casos.

Examinamos também os tamanhos dos documentos em quantidade de palavras, considerando somente as palavras do dicionário do vocabulário controlado para a coleção. Vide Figura 3.3. A média de palavras por documento é de 467,42, com desvio padrão de 411,45.

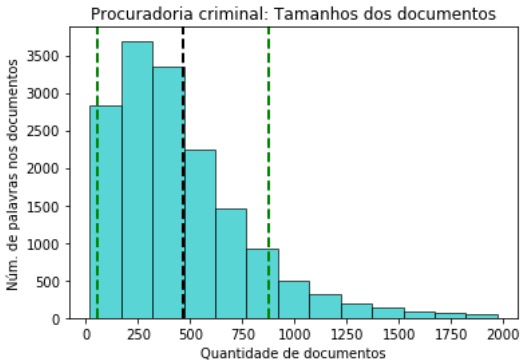
3.3.2 As procuradorias criminais especializadas

Os pareceres das procuradorias criminais concentram-se majoritariamente na análise dos recursos de *Habeas Corpus*, Apelações, Recurso ordinário e Recurso em sentido estrito, conforme mostra o gráfico da Figura 3.4.

Pareceres sobre *Habeas Corpus*

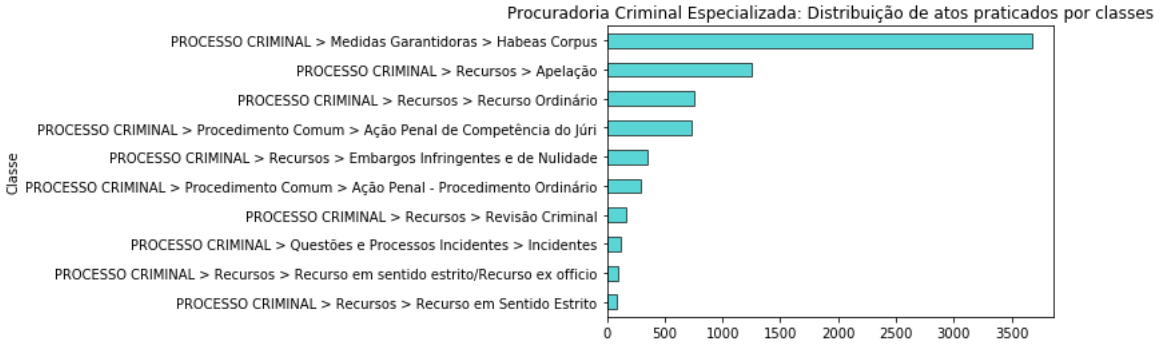
Os pareceres em *Habeas Corpus* analisam a procedência dos pedidos do paciente recluso acerca da concessão de sua liberdade, face a decisão de prisão preventiva do magistrado. No Apêndice C há um exemplo de parecer de *Habeas Corpus*.

Figura 3.3: PC - Tamanho dos documentos em palavras.



Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

Figura 3.4: PCE - Distribuição de atos praticados por classes (as 10 mais recorrentes).



Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

Pareceres sobre Apelações

As apelações apreciadas nas procuradorias criminais especializadas concentram-se casos provenientes do Tribunal do Juri e de Delitos de trânsito (2o. e 3o. itens na Figura 3.4, respectivamente) .

Os pareceres sobre apelações sobre delitos de trânsito não se diferenciam em sua estrutura em relação aos demais pareceres de apelação.

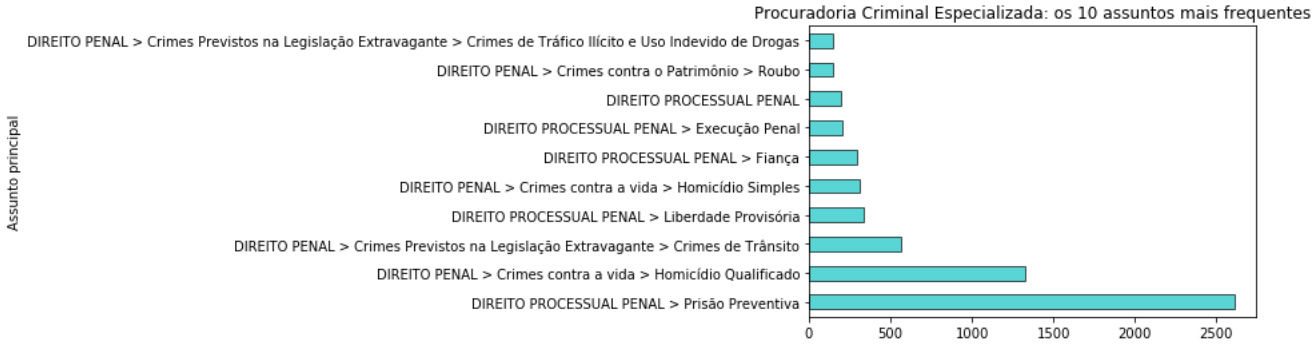
Os pareceres de apelações em casos do Tribunal do Juri diferem dos demais na análise do mérito do recurso, porque a lei estabelece no artigo 593, inciso III do Código Penal, que os pontos em que se pode contestar a sentença são os descritos nas alíneas “a”, “b”, “c” e “d”. Desta forma, a seção de análise do mérito é estruturada em função da avaliação de cada uma destas alíneas.

Pareceres sobre Recurso em sentido estrito

Os recursos em sentido estrito predominam as discussões sobre a sentença de pronúncia proferida pelo magistrado em relação ao réu recorrente. Na sentença de pronúncia, o magistrado entende que há fortes indícios de autoria e materialidade do delito e determina que o réu deverá ser julgado pelo Tribunal do Juri.

Acerca da distribuição de assuntos, observa-se na Figura 3.5 que a distribuição dos pareceres por assuntos corresponde à distribuição por classes. Diferentemente do que ocorre nas procuradorias criminais, em que as apelações abordam vários assuntos, aqui em cada classe são abordados poucos assuntos.

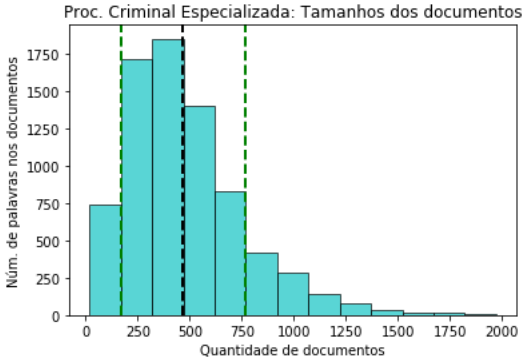
Figura 3.5: PCE - Distribuição de atos praticados por assunto



Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

Com relação ao tamanho dos documentos, a média da quantidade de palavras dos documentos é muito próxima da média das promotorias criminais (466,92 palavras por documento), porém com desvio padrão menor (298,75).

Figura 3.6: PCE - Tamanho dos documentos em palavras.



Fonte: Sistema de Controle de Feitos - MPDFT (2012-2018)

Tabela 3.1: Aproveitamento dos dados cadastrais dos feitos

Dado	Propósito
Classe	amostragem para base de treinamento
Assuntos	enriquecimento dos modelos
Nomes das partes	pré-processamento dos textos
Nomes dos membros	pré-processamento dos textos
Órgão de origem do processo	pré-processamento dos textos

3.4 Preparação dos Dados

Os dados sobre os atos praticados foram extraídos do sistema de controle de feitos de acordo com o tipo da procuradoria responsável. Todos os atos praticados são do tipo denominado "Manifestação em 2a. Instância", que corresponde aos pareceres.

De cada ato praticado, extraímos os seguintes dados do sistema de controle de feitos:

Extração do teor dos documentos

Os documentos correspondentes aos atos praticados são armazenados no sistema de controle de feitos no formato PDF, tornando necessária a extração do texto com auxílio da biblioteca Apache Tika [14].

Obtenção da lista de termos do *Thesaurus* do STF

O STF dispõe de um *Thesaurus* de termos jurídicos [5], que decidimos aproveitar para o vocabulário utilizado nos modelos. Estes termos foram utilizados para evitar separar em duas ou mais palavras os termos compostos já conhecidos.

Separação dos dados para indução e avaliação dos modelos

Para ambas as coleções, retiramos parte dos documentos para construir a base de referência a ser utilizada no processo de avaliação dos modelos. A parte restante foi utilizada para indução dos modelos. As bases de referência foram obtidas por amostragem estratificada, baseando-se na matéria principal do assunto principal dos documentos, com confiança de 95% e margem de erro de 5%. Desta forma, as coleções ficaram configuradas da seguinte maneira:

[5] `http://www.stf.jus.br/portal/jurisprudencia/pesquisarVocabularioJuridico.asp`

Tabela 3.2: Quantidade de documentos por coleção

Coleção	Base de treinamento	Base de referência	Total
Procuradorias criminais	15.718	526	16.244
Procuradorias crim. especializadas	7.161	504	7.665

Construção das bases de referência

As bases de referência são essenciais na avaliação dos modelos. A base de referência precisa conter uma lista de perguntas (*queries*), e para cada pergunta é necessário conhecer quais documentos são relevantes. Com estes dados é possível aplicar qualquer uma das medidas de performance de recuperação de informações que descrevemos na fundamentação teórica.

Com o apoio de especialistas de procuradorias criminal e criminal especializadas, foram elaboradas inicialmente 10 *queries* para cada uma das bases de referência. Em seguida, examinamos todos os seus documentos, e decidimos qual o nível de relevância cada documento tem em relação a cada uma das 10 *queries*. Os níveis de relevância são os seguintes:

- 0, se o documento for irrelevante para a *query*
- 1, se o documento for relevante para a *query*
- 2, se o documento for muito relevante para a *query*

Em seguida, produzimos outras duas versões das mesmas *queries*, porém formuladas de modo diferente, com termos diferentes, sem modificar seu sentido nem alterar os documentos relevantes correspondentes na base de referência. Isto permite avaliar resposta dos modelos face as diferentes formas de se descrever um caso. Transcrevemos aqui algumas *queries* a título de exemplo.

Tabela 3.3: Exemplo de versões de uma *query*

Query original	Versão 1	Versão 2
- "absolvição por insuficiência de provas para crime de tráfico inviável quando o comércio de drogas for atestado por policiais e pelas circunstâncias"	- "absolvição por insuficiência probatória não pode ser alegada quando a mercância for atestada por policiais e pelas circunstâncias"	- "se o comércio de entorpecentes houver sido relatado por agentes policiais a absolvição por insuficiência probatória não é cabível"
- "tentativa de furto com concurso de agentes não admite a aplicação do princípio da insignificância"	- "princípio da bagatela não pode ser alegado mediante furto com concurso de agentes"	- "princípio da bagatela incompatível com furto cometido com apoio de duas ou mais pessoas"
- "réu alega culpa da vítima porém conduziu o veículo de modo imprudente e provocando o acidente"	- "réu dirigia de forma imprudente pede compensação afirmando que a vítima provocou o acidente"	- "inviável a compensação de pena alegando que a o acidente foi provocado pela vítima"

Então, no total, cada base de referência contém 30 *queries* e seus respectivos documentos relevantes. Todas as *queries*, suas versões, bem como os quantitativos de documentos relevantes estão documentados nos Apêndices A e B.

3.4.1 O pré-processamento dos textos

O pré-processamento dos textos compreende as seguintes etapas:

- Remoção do nome de partes e dos procuradores e promotores envolvidos;
- Remoção de palavras irrelevantes (*stopwords*). Foram consideradas as "stopwords" do idioma português, obtidas na biblioteca NLTK [6].
- Identificação de termos constantes no *Thesaurus* do STF e substituição de termos alternativos pelos preferenciais, conforme indicado nas instruções de uso do Thesaurus;
- Remoção de nomes de arquivos: vários documentos continham seus endereço e nome de arquivo que persistiu no documento após produzida sua versão em PDF. Estes nomes e endereços geram termos sem sentido após o processo de remoção de caracteres especiais e, portanto, precisaram ser removidos;
- Remoção de caracteres especiais e numerais;
- Redução das palavras ao seu radical (*stemming*);

[6] `http://www.nltk.org`

- Extração das citações às normas jurídicas;

A extração das citações às normas jurídicas

Os redatores dos documentos descrevem suas referencias de forma livre, sem a preocupação de favorecer a detecção automatizada. As referências são observadas no documento das mais variadas formas e grafias. Os principais obstáculos para identificar as citações no texto são os seguintes:

- Abreviações: A referência não é redigida por extenso, mas de modo abreviado. Exemplo: "Art. 157 §2 II do CP" ou "Art. 157 §2 inc. II do CP", para referenciar "Artigo 157, parágrafo 2o, inciso 2 do Código Penal";
- Variação na ordem das partes da referência. Exemplo: "CP art 157 §2, inc. II";
- Aglutinação: Duas ou mais referências em um único texto. Exemplo: "'Art 180 §1 e 157 §2, ambos do CP"'
- Separação "(...) art. 157 (...) inciso II (...) do Código Penal"
- Sinônimos: Diplomas referenciados pelo número ou pelo nome popular. Exemplo: "Lei 10.826/03" e "Estatuto do desarmamento"; "Código Penal" e "Estatuto repressivo"; "CF", "Carta Magna" e "Carta da República";

Abordamos estes problema com o uso de diversas expressões regulares. As expressões regulares foram utilizadas inicialmente para localizar termos que fazem parte de uma referência, tais como "artigo", "inciso", "§", e também os diplomas considerando os sinônimos número e abreviações observados nos textos.

As referências são o endereço dos verbetes da legislação, e cada uma de suas partes define têm um nível de especificidade, sendo o diploma o mais abrangente, o artigo o termo intermediário, e incisos, alíneas e parágrafos como partes mais específicas. Assim, primeiramente apuramos a posição do "artigo", e então reunirmos as outras partes da referência de acordo com a proximidade da sua posição no texto em relação a ele. Derivamos a referência em função dos artigos e dos demais termos relevantes que estejam mais próximos. Se o artigo não for localizado para um determinado diploma, descartamos a referência, por ser abrangente demais para a finalidade do trabalho.

Existe uma iniciativa da parte do Senado Federal para a padronização das referências, chamada LEXML [7]. A proposta é definir URNs para as referências, de modo que estas possam ser identificadas de modo automatizado e inequívoco. Entretanto, não se trata de uma norma, mas de uma recomendação. Utilizamos a representação do LEXML para registrar as referências detectadas nos documentos.

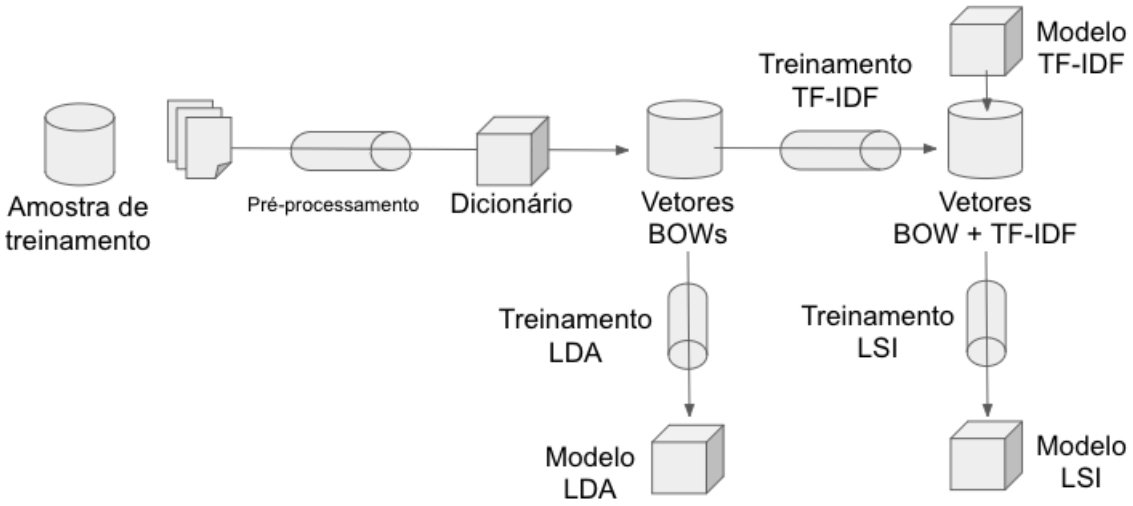
[7] `https://www.lexml.gov.br`

Evidentemente, este procedimento não é exato, e algumas referências não são corretamente detectadas ou não existem na lei. Contudo, isso não representa um problema porque as referências incorretas jamais serão correspondidas por aquelas presentes na *query* do especialista, não tendo pouco efeito prático para a aplicação que fizemos.

3.5 Modelagem

dos modelos.

Figura 3.7: Indução dos modelos.



3.5.1 O dicionário de palavras

Todas as técnicas selecionadas neste trabalho precindem de um vocabulário controlado, que aqui referenciamos como "dicionário". Para cada coleção estabelecemos seu dicionário, baseando-se nas palavras que ocorrem nos documentos pré-processados pertencentes à respectiva amostra de treinamento. Em seguida faz-se um corte dos termos pouco frequentes, e também dos termos muito frequentes. Em ambas as coleções, os parâmetros utilizados para o corte são os seguintes:

- Limite inferior: termos que ocorrem em mais de 5 documentos.
- Limite superior: termos que ocorrem em mais da metade dos documentos.

3.5.2 Modelo BOW com TF-IDF

Para cada documento da amostra de treinamento construímos o vetor BOW correspondente. O vetor BOW tem uma dimensão para cada termo do dicionário, onde se registram

a frequência de cada palavra no documento. Armazenamos os vetores BOW para uso na indução dos demais modelos. O modelo TF-IDF é treinado utilizando os vetores BOW produzidos. Com o modelo treinado construímos novos vetores ponderados com TF-IDF e os armazenamos em arquivo a parte.

3.5.3 Modelos de tópicos

A indução de modelos LSI e LDA exigem como parâmetro a quantidade de tópicos latentes. A literatura não indica critérios para a seleção do número ideal de tópicos. Como o propósito do trabalho é prover solução de recuperação de informações, decidimos experimentar 8 faixas de valores e avaliar a performance dos modelos na atividade de recuperação de informações.

- 8 modelos baseados em LSI, parametrizados com 50, 100, 150, 200, 250, 300, 350 e 400 tópicos respectivamente;
- 8 modelos baseados em LDA, parametrizado com 50, 100, 150, 200, 250, 300, 350 e 400 tópicos respectivamente;

Os modelos LSI podem ser induzidos com vetores BOW, mas optamos por usar os vetores BOW + TF-IDF.

Com relação aos modelos LDA, usamos os vetores BOW, visto que os vetores com TF-IDF são incompatíveis com a abordagem de derivação dos tópicos utilizada neste tipo de modelo.

As implementações destes modelos no Gensim são *online*, em alinhamento com a estratégia da ferramenta de permitir a indução de modelos com recursos computacionais de memória constantes, e com isso viabilizar o uso de bases de documento de tamanho virtualmente ilimitado, e viabilizar a integração de novos documentos aos modelos, a medida que eles são conhecidos.

A indução de modelos do tipo *online* tipicamente é feita com mais de uma passagem pelos documentos. Com relação aos modelos LSI, optamos pelo valor padrão do Gensim. No caso dos modelos LDA, efetuamos a indução com 15 passagens para ambas as coleções.

3.5.4 O modelo baseado em BM25

O modelo de recuperação de informações baseado em BM25 não seguiu a mesma abordagem de indução e avaliação dos demais modelos. Neste modelo, a recuperação de informações considera apenas a base de referência. O modelo estabelece um escore entre a *query* e os documentos. Se examinarmos a equação 2.5, veremos que este escore

depende do fator $n(q_i)$, que contabiliza quantos documentos da coleção contém cada um dos termos da *query*. O único fator que poderia representar algum aprendizado acerca da coleção de treinamento é o *agdl*, que contabiliza a quantidade média de palavras dos documentos. Assumimos que este valor pode ser aproximado a partir da amostra de avaliação. A otimização possível para os parâmetros b e k dependem também da base de referência. Portanto, entendemos como irrelevante a utilização da base de treinamento para o modelo BM25.

3.6 O enriquecimento dos modelos

Foram apresentados ao especialista vários dados cadastrados a respeito do feito e do ato praticado. O especialista julgou relevante para a pesquisa apenas o assunto e as citações.

Cogitamos enriquecer os vetores BOW com os assuntos e citações, mas logo nos primeiros ensaios percebemos que esta abordagem não era razoável, porque os assuntos e citações têm maior importância do que uma simples palavra no documento.

Desta forma, utilizamos os assuntos e citações da seguinte maneira:

- Além da descrição textual do caso, a *query* acompanha os assuntos e as citações;
- O modelo ordena os documentos da coleção de referência de em ordem decrescente de similaridade a *query*;
- Divide-se os documentos em dois grupos:
  - a) os que contém o dado pretendido (assunto ou citação);
  - b) os que não contém o dado pretendido (assunto ou citação).

  O ato praticado pode ter diversos assuntos e citações. Basta ter um deles em comum com a query. Em cada grupo, a ordem estabelecida pelo modelo é preservada.
- o grupo a), ocupará parte superior do ranking, e o grupo b) a inferior.

Exemplificando, considere para a uma *query* qualquer, com enriquecimento somente com assunto, que ranking fornecido pelo modelo seja o seguinte:

Após aplicar o enriquecimento com assunto, o ranking será o seguinte:

A modalidades de enriquecimento são as seguintes:

- Sem enriquecimento;

Tabela 3.4: Exemplo de ranking de documentos sem enriquecimento

Ranking	Documento	Contém o assunto?
1	1	não
2	2	sim
3	3	não
4	4	sim

Tabela 3.5: Exemplo de ranking de documentos com enriquecimento

Ranking	Documento	Contém assunto?
1	2	sim
2	4	sim
3	1	não
4	3	não

- Enriquecido com assunto;
- Enriquecido com citações;
- Enriquecido com assuntos, e em seguida com citações.

3.7 Avaliação

Nesta seção relatamos como foi feita a avaliação dos modelos e como os comparamos. Relatamos também como avaliamos o efeito do enriquecimento dos modelos com os dados dos feitos e citações. Os procedimentos de avaliação foram executados nas coleção de atos praticados das procuradorias criminais, e na coleção das procuradorias criminais especializadas.

3.7.1 Representação dos documentos

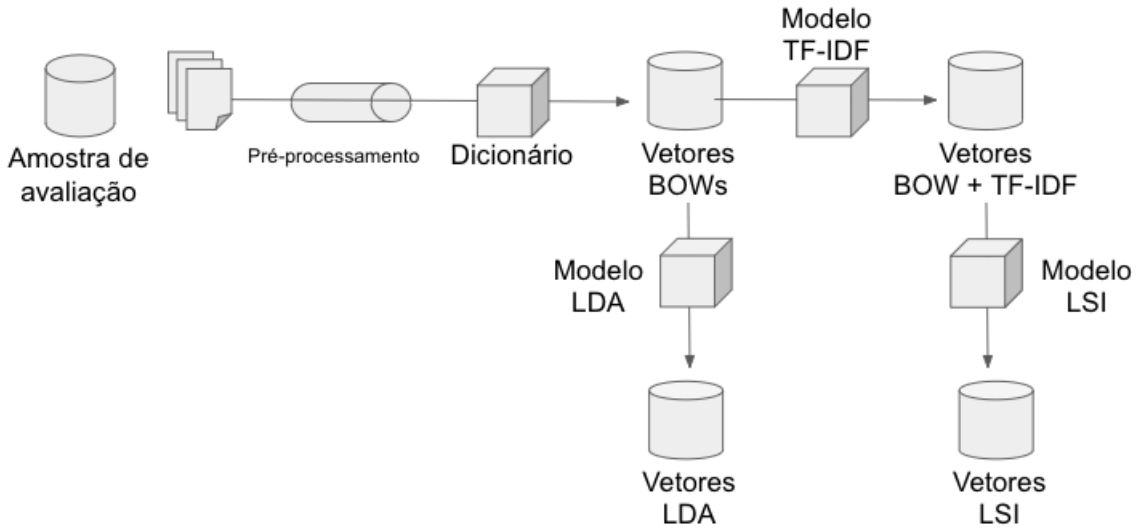
Utilizamos os modelos treinados com a coleção de treinamento para derivar os vetores da base de referência, conforme esquema apresentado na Figura 3.8.

3.7.2 Avaliação dos modelos sem enriquecimento

Para avaliar os modelos sem enriquecimento, seguimos os seguintes passos:

1. Apuração da performance dos modelos por *query*: Para cada um dos 18 modelos de coleções, submetemos as 30 *queries* da base de referência e registramos os *rankings*

Figura 3.8: Obtenção dos vetores da base de referência para cada modelo.



de documentos retornados. Em seguida, medimos a performance dos modelos por *query*, utilizando a métrica NDCG na posições 1, 5, 10, 15 e 20 do rankings.

2. Apuração da performance dos modelos para todas as *queries*: Aferimos o NDCG médio das queries para cada modelo, considerando as posições 1, 5, 10, 15 e 20 do ranking.

3. Comparação dos modelos: Aplicamos o teste Nemenyi [16] para verificar se houve diferença estatisticamente significante entre os modelos.

3.7.3 Avaliação dos modelos enriquecidos

O enriquecimento dos modelos tem as seguintes modalidades:

- Enriquecimento com assunto;
- Enriquecimento com citações;
- Enriquecimento com assunto e com citações.

Para avaliar os modelos com cada uma das modalidades de enriquecimento, seguimos os seguintes passos:

1. Apuração da performance dos modelos por *query*: Para cada um dos 18 modelos enriquecidos, submetemos as 30 *queries* da base de referência e registramos os *rankings* de documentos retornados. Em seguida, medimos a performance dos modelos por *query*, utilizando a métrica NDCG na posições 1, 5, 10, 15 e 20 do rankings.

2. Apuração da performance dos modelos enriquecidos para todas as *queries*: Aferimos o NDCG médio das *queries* para cada modelo enriquecido, considerando as posições 1, 5, 10, 15 e 20 do ranking.

3. Comparação dos modelos: Aplicamos o teste Nemenyi para verificar se houve diferença estatisticamente significante entre cada modelo, em sua versão sem enriquecimento e enriquecida com assuntos, com citações, e com assuntos e citações.

3.8 Implementação

Construímos um protótipo para que o especialista utilize o modelo selecionado para apoiar seu traballho, e coletar novas *queries* juntamente com os documentos considerados relevantes. Para cada documento apresentado, o especialista pode:

- Avançar para o próximo documento (ou retroceder): registra-se o documento como "não relevante"'
- Marcar o documento como relevante: registra-se o documento como "relevante"
- Copiar o teor do documento: registra-se o documento como "muito relevante"

O protótipo apresenta uma ementa da peça produzida com os termos do *Thesaurus* do STF, e apresenta também a lista de citações extraídas automaticamente. No futuro, pretende-se permitir que o próprio especialista modifique esta lista como achar conveniente, acrescentando, alterando ou removendo as citações.

Outro ponto a se destacar é que o protótipo reconhece as citações que o especialista escrever no texto de pesquisa. As citações serão reconhecidas automaticamente, da mesma forma que são reconhecidas no texto dos documentos.

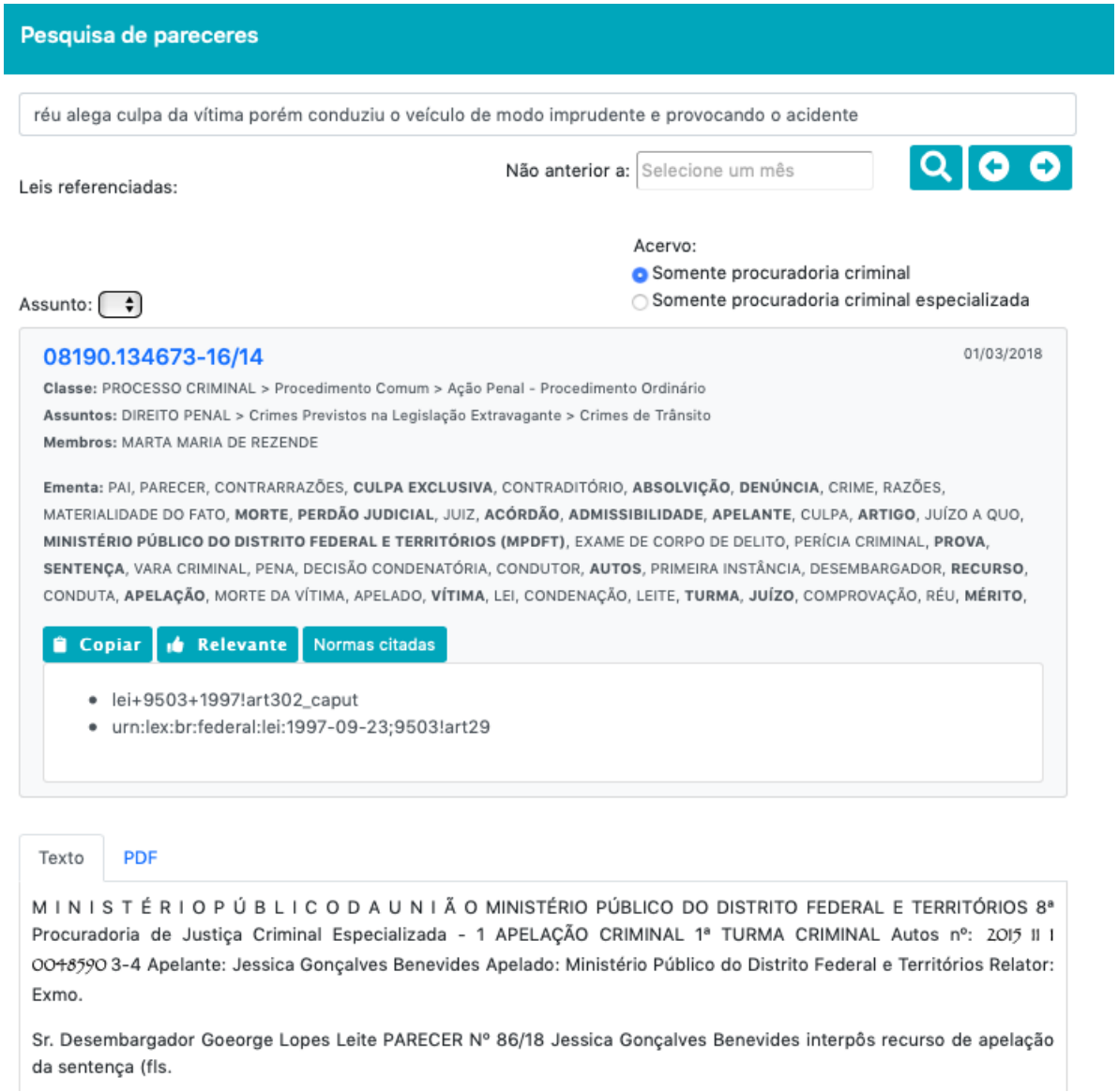
Além do texto de pesquisa, o especialista pode restrigir a busca por assunto e por data de antiguidade do documento. Não são apresentados documentos mais antigos do que a data que o especialista fornecer (mês e ano).

Participarão do uso do protótipo somente as procuradorias criminais e criminais especializadas.

O propósito inicial do protótipo era apoiar a construção das bases de referência. Contudo, o protótipo não se mostrou eficiente porque este disponibiliza apenas o texto puro dos documentos, sem a formatação e diagramação originais. Isto ocorre porque os documentos registrados no sistema de controle de feitos estão em formato PDF. Quando se extrai o seu teor para indexação, perde-se a formatação. O especialista alega que reproduzir a mesma formatação a partir do texto puro é trabalhoso e improdutivo. Com isto, a adoção do protótipo está dependente de se estabeleça uma solução para preservar a

formatação do texto. A solução considerada é a aquisição de biblioteca de *software* capaz de converter arquivos em formato PDF para o formato do editor de textos, preservando a formatação.

Figura 3.9: Protótipo disponibilizado para os especialistas



Capítulo 4

Resultados

Nesta seção, relataremos os detalhes da avaliação dos modelos.

Analisaremos primeiramente os resultados da coleção de documentos das Procuradorias Criminais, e em seguida os resultados das Procuradorias Criminais Especializadas. Para cada coleção de documentos, examinaremos se há diferenças significativas entre modelos sintáticos e semânticos, e em seguida examinaremos se há diferenças entre modelos sem enriquecimento e com enriquecimento.

4.1 Procuradoria Criminal

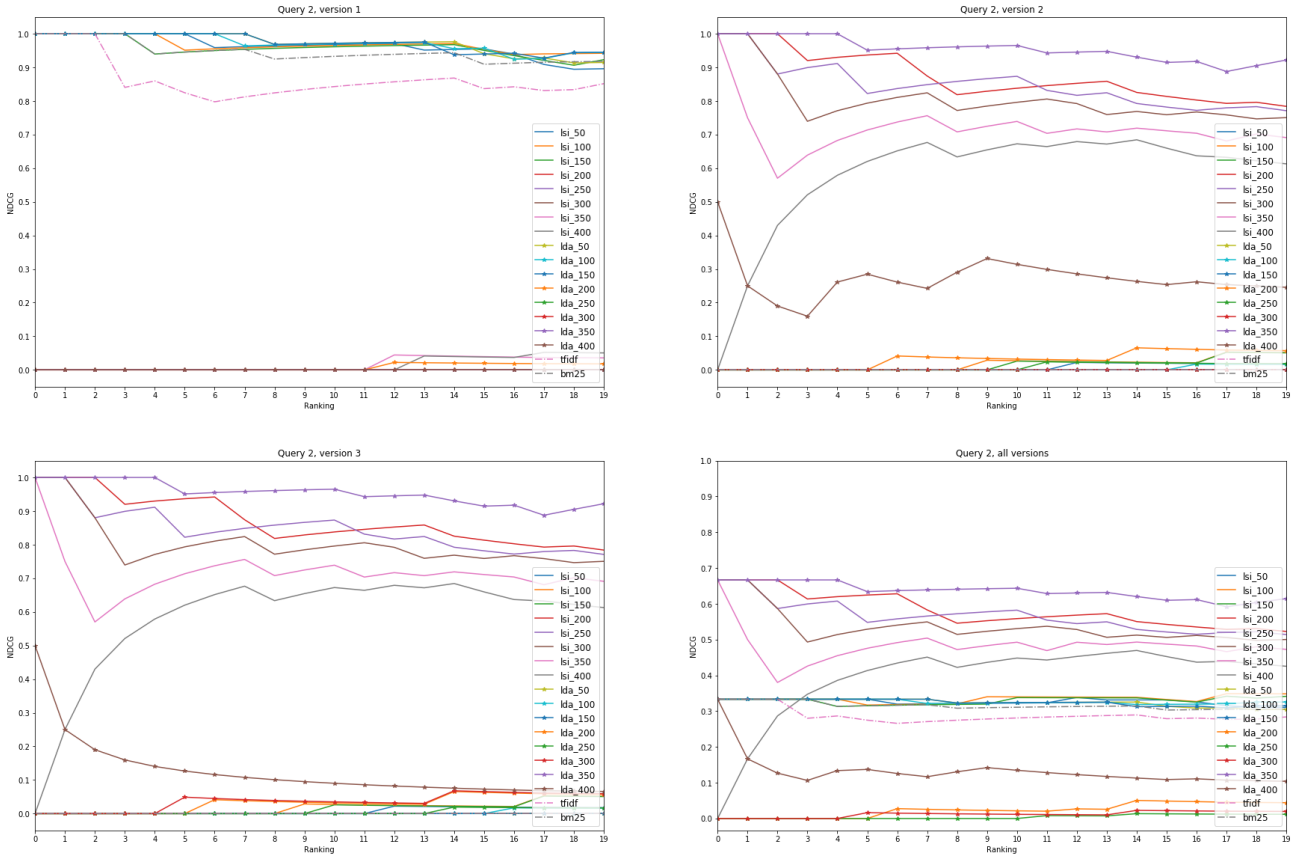
Seguem o resultados dos experimentos realizados referentes à base de referência proveniente da coleção de documentos das Procuradorias Criminais.

4.1.1 Comparação dos modelos sintáticos e semânticos

Para avaliar o desempenho dos modelos, inicialmente verificamos o comportamento para cada *query* e suas versões. O objetivo inicial é verificar anomalias e ter um panorama do comportamento dos modelos quanto as versões das *queries*.

Todos os gráficos relatando a performance de cada *query* em todas as suas versões e conjuntamente estão relacionados no Apêndice A Como são muitas *queries*, apresentamos na Figura 4.1 os resultados para a *query* 2 apenas com propósito ilustrativo. O gráfico mostra a evolução do escore NDCG de acordo com a posição no ranking.

Figura 4.1: PC - Performance dos modelos para a *Query* 2, todas as versões



Examinando os resultados em cada uma das *queries* e suas versões, percebe-se que:

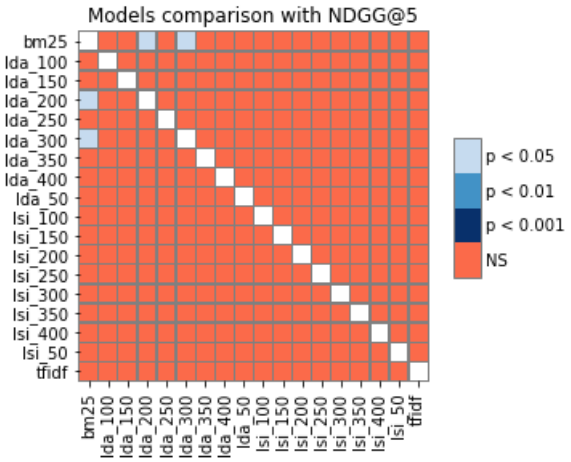
1. À exceção dos modelos LDA, ambas as modalidades sintática e semântica obtiveram resultados próximos.

2. A performance dos modelos LDA não tem qualquer padrão para esta coleção. Não há uma configuração de número de tópicos cujos resultados sejam melhores que os demais. Os resultados são imprevisíveis se comparados tanto entre as versões das queries, quanto entre uma query e outra.

Para avaliar se houve diferença estatística entre os modelos, utilizamos o teste de Nemenyi, medindo os modelos em 5 posições do ranking: 1, 5, 10, 15 e 20.

O teste mostrou diferença significativa apenas na posição 5 do ranking, conforme mostra a Figura 4.2. O mapa de calor mostra diferença significativa entre o modelo BM25 e os modelos LDA (200 e 300 tópicos). Vide Apêndice A.2.

Figura 4.2: PC - Performance dos modelos com todas as *queries* na posição 5 do ranking



A diferença entre os modelos é apresentada na Figura 4.3. Percebe-se que o modelo BM25 tem performance superior.

Figura 4.3: PC - Detalhamento das diferenças apresentadas pelo teste Nemenyi

Conclusão

Não podemos afirmar que há diferença significativa de performance entre modelos sintáticos e semânticos para a coleção de documentos das Procuradorias Criminais.

4.1.2 Comparação dos modelos sem enriquecimento e com enriquecimento

Avaliamos todos os modelos nas modalidades:

- Sem enriquecimento

- Enriquecido com assuntos
- Enriquecido com citações
- Enriquecidos com assuntos e citações

Utilizamos o teste de Nemenyi para avaliar em cada modelo, se havia diferença significante entre cada modalidade de enriquecimento. Os testes não indicaram diferença nas modalidades de enriquecimento em qualquer dos modelos. Vide apêndice A.3.

Conclusão

Não podemos afirmar que o enriquecimento com assuntos e citação melhora a performance dos modelos na coleção de documentos das Procuradorias Criminais.

4.2 Procuradoria Criminal Especializada

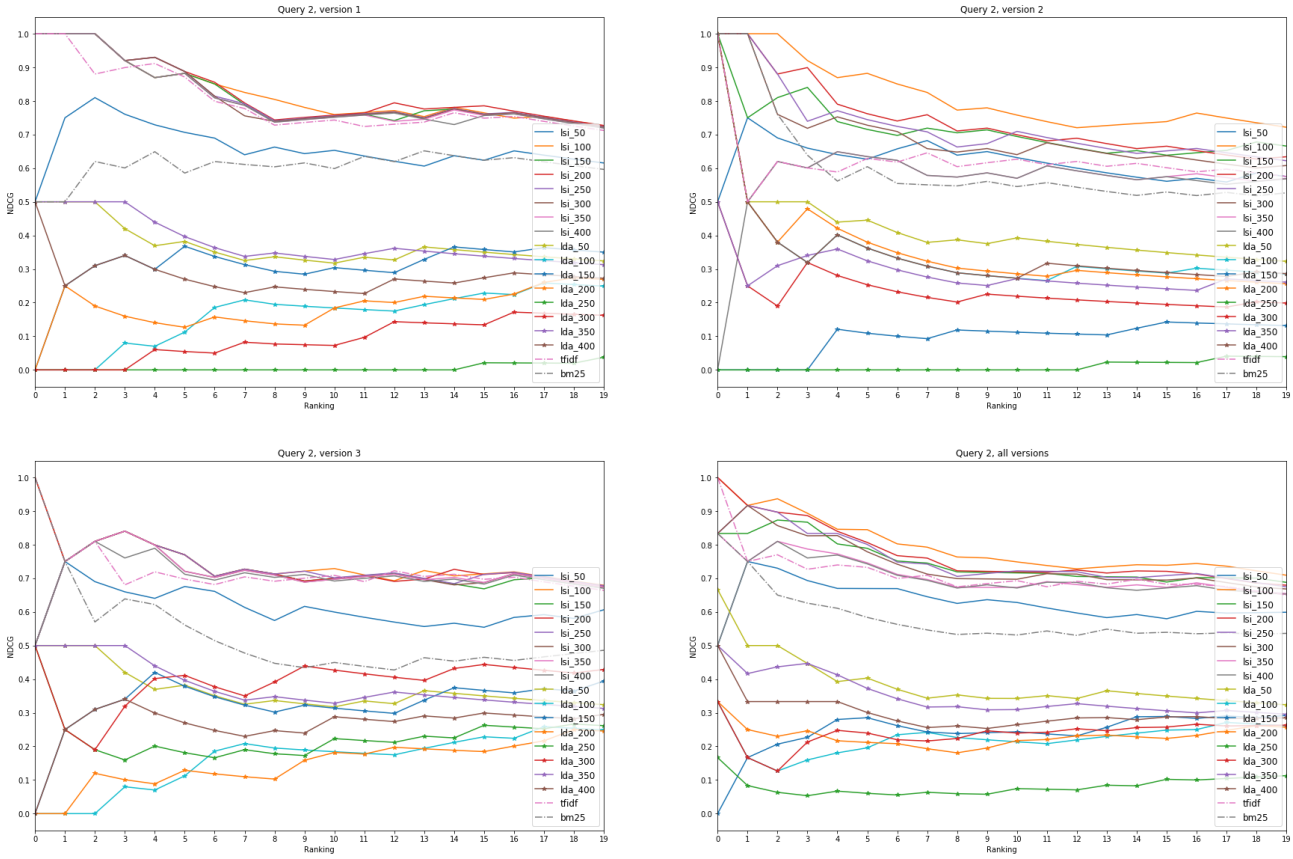
Seguem o resultados dos experimentos realizados referentes à base de referência proveniente da coleção de documentos das Procuradorias Criminais Especializadas.

4.2.1 Comparação dos modelos sintáticos e semânticos

Para avaliar o desempenho dos modelos, inicialmente verificamos o comportamento para cada *query* e suas versões. O objetivo inicial é verificar anomalias e ter um panorama do comportamento dos modelos quanto as versões das *queries*.

Todos os gráficos relatando a performance de cada *query* em todas as suas versões e conjuntamente estão relacionados no apendice B. Como são muitas *querys*, apresentamos na Figura 4.4 os resultados para a *query* 2 apenas com propósito ilustrativo. O gráfico mostra a evolução do escore NDCG de acordo com a posição no ranking.

Figura 4.4: PCE - Performance dos modelos para a *Query* 2, todas as versões



Examinando os resultados em cada uma das *queries* e suas versões, percebe-se que:

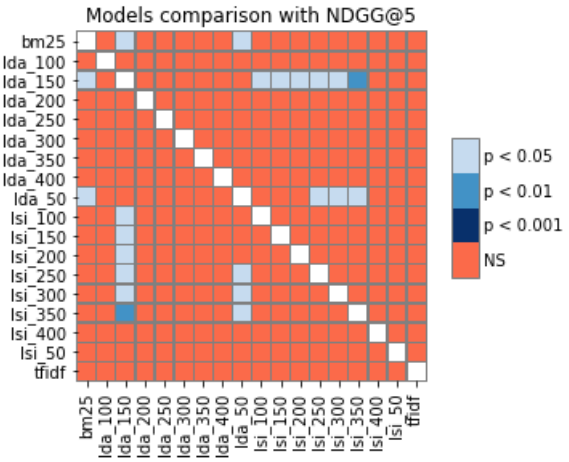
1. À exceção dos modelos LDA, ambas as modalidades sintática e semântica obtiveram resultados próximos.

2. A performance dos modelos LDA não tem qualquer padrão para esta coleção. Não há uma configuração de número de tópicos cujos resultados sejam melhores que os demais. Os resultados são imprevisíveis se comparados tanto entre as versões das queries, quanto entre uma query e outra.

Para avaliar se houve diferença estatística entre os modelos, utilizamos o teste de Nemenyi, medindo os modelos em 5 posições do ranking: 1, 5, 10, 15 e 20.

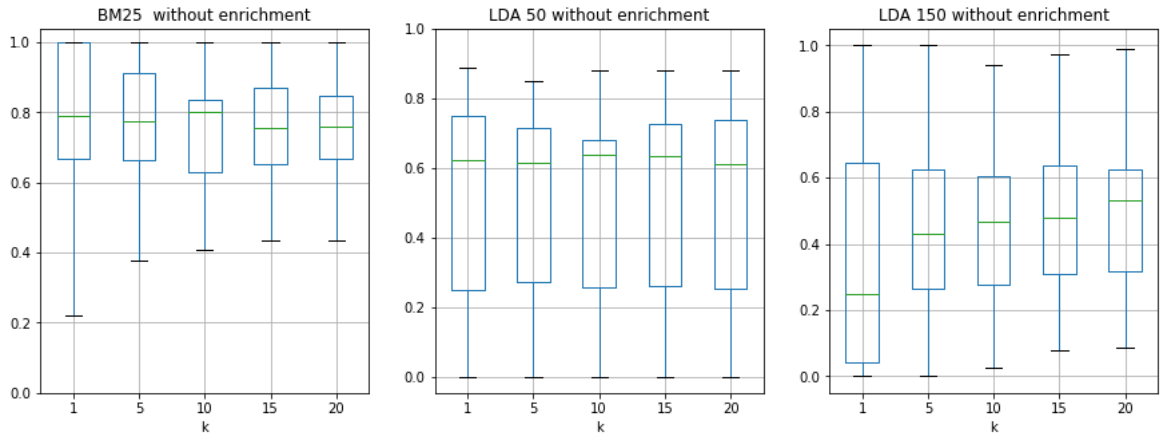
O teste mostrou diferença significativa nas posições 1, 5, 10, 15 e 20 do ranking. Com propósito ilustrativo, apresentamos a comparação no nível 5 do ranking, conforme mostra a Figura 4.5. O mapa de calor mostra diferença significativa entre os modelos LDA(50, 150, 200 e 350 tópicos) e os modelos LSI (100, 150, 200, 250, 300 e 350 tópicos) e BM25.

Figura 4.5: PCE - Performance do modelo com todas as *queries* na posição 5 do ranking



Apresentamos as diferenças entre o modelo BM25 e os modelos LDA 50 e 150 na Figura 4.6. As todas as diferenças podem ser examinadas no apêndice B.2. Percebe-se que o modelo BM25 tem performance superior aos modelos LDA 50 e 150.

Figura 4.6: PCE - Detalhamento das diferenças apresentadas pelo teste Nemenyi



Se examinarmos as diferenças em cada nível de ranking, percebe-se que somente modelos do tipo LDA se destacam com performance inferior aos demais.

Conclusão

Não podemos afirmar que há diferença significativa de performance entre modelos sintáticos e semânticos para a coleção de documentos das Procuradorias Criminais Especializadas.

4.2.2 Comparação dos modelos sem enriquecimento e com enriquecimento

Avaliamos todos os modelos nas modalidades:

- Sem enriquecimento
- Enriquecido com assuntos
- Enriquecido com citações
- Enriquecidos com assuntos e citações

Utilizamos o teste de Nemenyi para avaliar em cada modelo, se havia diferença significante entre cada modalidade de enriquecimento. Os testes não indicaram diferença nas modalidades de enriquecimento em qualquer dos modelos. Vide B.3.

Conclusão

Não podemos afirmar que o enriquecimento com assuntos e citação melhora a performance dos modelos para a coleção de documentos das Procuradorias Criminais Especializadas.

Capítulo 5

Conclusões e trabalhos futuros

Este capítulo descreve o que se pode concluir a partir dos resultados dos experimentos e os trabalhos pretendidos para o futuro.

5.1 Conclusões

Passamos então às conclusões sobre cada hipótese do trabalho.

1. **O uso de técnicas de busca semântica no teor dos documentos aumenta a performance de modelos de induzidos em mineração de dados para a recuperação de atos praticados.**

   Refutamos a hipótese, no contexto das bases de dados analisadas e da aplicação para elas pretendida. Examinamos que alguns dos modelos semânticos tiveram performance significativamente inferior, e que não houve modelo semântico significativamente superior aos modelos sintáticos, conforme relatado em A.2 e B.2.

2. **Um modelo enriquecido com os dados cadastrais dos feitos é mais eficiente na localização dos casos semelhantes.**

   Refutamos a hipótese, no contexto das bases de dados analisadas e da aplicação para elas pretendida. Examinamos que para ambas as bases e para todos os modelos, não houve diferença significativa entre os modelos sem enriquecimento e enriquecido com dados cadastrais, conforme relatado em A.3 e B.3.

3. **Um modelo enriquecido com as citações às normas jurídicas presentes nos documentos é mais eficiente na recuperação de atos praticados.**

   Refutamos a hipótese, no contexto das bases de dados analisadas e da aplicação para elas pretendida. Examinamos que para ambas as bases e para todos os modelos,

não houve diferença significativa entre os modelos sem enriquecimento e enriquecido com citações. Conforme relatado em A.3 e B.3.

5.1.1 O modelo selecionado para implantação

Considerando:

1. A performance do modelo
2. A simplicidade do modelo

Decidimos utilizar preliminarmente a técnica BM25. Acerca dos critérios estabelecidos e baseando-se nos experimentos, temos que:

- Performance: Não houve diferença significativa de performance entre ele e os melhores modelos;
- Simplicidade: Das técnicas avaliadas está entre as mais simples e com baixo custo de implementação.

5.2 Trabalhos Futuros

Neste trabalho, os modelos foram induzidos utilizando as configurações sugeridas pelas bibliotecas utilizadas. Pretendemos utilizar o método e os procedimentos de avaliação desenvolvido para otimizar os modelos e verificar se as conclusões acerca das hipóteses se mantém.

Pretendemos também explorar novas formas de enriquecimento com citações, examinando se a organização das citações com o uso grafos é capaz de aumentar a performance dos modelos.

Além disso, pretendemos examinar se a identificação automática das seções dos documentos é possível, e se a remoção das seções de menor interesse por parte do especialista pode contribuir para a performance dos modelos. Os pareceres têm seções para endereçamento e qualificação da peça, que não tem valor para a produção de novos pareceres. Ainda, a seção de análise de pré-requisitos do parecer é muito individual do caso e poderia ser suprimida no documento antes da indução dos modelos.

Referências

[1] Nacional, Congresso: *LEI COMPLEMENTAR Nº 75, DE 20 DE MAIO DE 1993*, 1993. http://www.planalto.gov.br/ccivil_03/leis/lcp/lcp75.htm. 1

[2] *MPDFT - Lei de acesso a Informação*. http://www.mpdft.mp.br/transparencia/, acesso em 2017-06-06TZ. 2

[3] Dabney, Daniel P.: *The Curse of Thamus: An Analysis of Full-Text Legal Document Retrieval*. Law Library Journal, 78:5, 1986. http://heinonline.org/HOL/Page?handle=hein.journals/llj78&id=15&div=&collection=. 7

[4] Baeza-Yates, Ricardo e Berthier Ribeiro-Neto: *Modern information retrieval: the concepts and technology behind search*. Addison Wesley, New York, second edition edição, 2011, ISBN 978-0-321-41691-9. 11, 15

[5] Salton, Gerard: *Automatic text processing: the transformation, analysis, and retrieval of information by computer*. Addison-Wesley series in computer science. Addison-Wesley, Reading, Mass, 1988, ISBN 978-0-201-12227-5. 12

[6] Deerwester, Scott, Susan T. Dumais, George W. Furnas, Thomas K. Landauer e Richard Harshman: *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41:391–407, 1990. 13, 14

[7] Hofmann, Thomas: *Probabilistic latent semantic indexing*. Em *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, páginas 50–57. ACM, 1999. http://dl.acm.org/citation.cfm?id=312649, acesso em 2017-09-03TZ. 14

[8] Blei, David M., Andrew Y. Ng e Michael I. Jordan: *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3(Jan):993–1022, 2003, ISSN ISSN 1533-7928. http://www.jmlr.org/papers/v3/blei03a.html, acesso em 2017-07-29TZ. 14

[9] Robertson, Stephen e Hugo Zaragoza: *The Probabilistic Relevance Framework: BM25 and Beyond*. Found. Trends Inf. Retr., 3(4):333–389, abril 2009, ISSN 1554-0669. http://dx.doi.org/10.1561/1500000019, acesso em 2018-12-01TZ. 15

[10] Opijnen, Marc van e Cristiana Santos: *On the concept of relevance in legal information retrieval*. Artificial Intelligence and Law, 25(1):65–87, março 2017, ISSN 0924-8463, 1572-8382. http://link.springer.com/10.1007/s10506-017-9195-8, acesso em 2017-05-17TZ. 17

[11] Quaresma, Paulo e Irene Rodrigues: *A Question-answering System for Portuguese Juridical Documents*. Em *Proceedings of the 10th International Conference on Artificial Intelligence and Law*, ICAIL '05, páginas 256–257, New York, NY, USA, 2005. ACM, ISBN 978-1-59593-081-1. http://doi.acm.org/10.1145/1165485.1165536, acesso em 2017-05-18TZ. 18

[12] Raghav, K., Pailla Balakrishna Reddy, V. Balakista Reddy e Polepalli Krishna Reddy: *Text and Citations Based Cluster Analysis of Legal Judgments*. Em *SpringerLink*, páginas 449–459. Springer, Cham, dezembro 2015. https://link-springer-com.ez54.periodicos.capes.gov.br/chapter/10.1007/978-3-319-26832-3_42, acesso em 2017-05-18TZ. 19

[13] Wirth, Rüdiger e Jochen Hipp: *CRISP-DM: Towards a Standard Process Model for Data Mining*. Em *4th international conference on the practical applications of knowledge discovery and data mining*, 2000. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.198.5133&rep=rep1&type=pdf. 19

[14] Zitting, Jukka: *Text and Metadata Extraction with Apache Tika*. Apache Lucane, Eurocone, páginas 3–12, 2010. 30

[15] Rehurek, Radim e Petr Sojka: *Software Framework for Topic Modelling with Large Corpora*. Em *In Proceedings of the Lrec 2010 Workshop on New Challenges for Nlp Frameworks*, páginas 45–50, 2010.

[16] Demsar, Janez: *Statistical Comparisons of Classifiers over Multiple Data Sets*. página 30. 38

Apêndice A

Performance dos modelos: Procuradorias Criminais

A.1 Avaliação dos modelos sem enriquecimento por query

A.1.1 *Query 0*

A *Query 0* pretende localizar pareceres sobre casos em que o réu foi condenado por tráfico de drogas e recorre sob argumento de insuficiência probatória, mesmo havendo depoimento policial favorável à tese de acusação. Vários julgados reafirmam que o depoimento dos agentes policiais tem relevância equiparável a de material probatório. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "absolvição por insuficiência de provas para crime de tráfico inviável quando o comércio de drogas for atestado por policiais e pelas circunstâncias"

- versão 2:

  > "absolvição por insuficiência probatória não pode ser alegada quando a mercância for atestada por policiais e pelas circunstâncias"

- versão 3:

  > "se o comércio de entorpecentes houver sido relatado por agentes policiais a absolvição por insuficiência probatória não é cabível"

Quantidade de documentos relevantes: 9

Quantidade de documentos muito relevantes: 17

Figura A.1: PC - performance da *Query* 0, versão 1

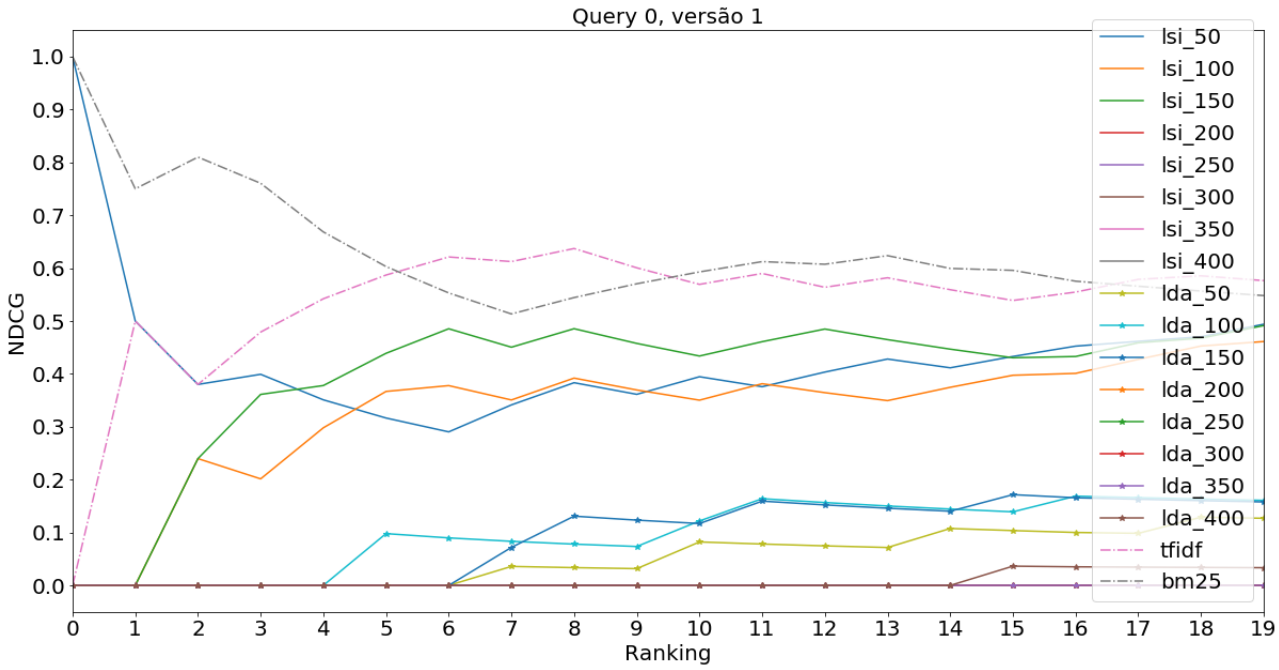


Figura A.2: PC - performance da *Query* 0, versão 2

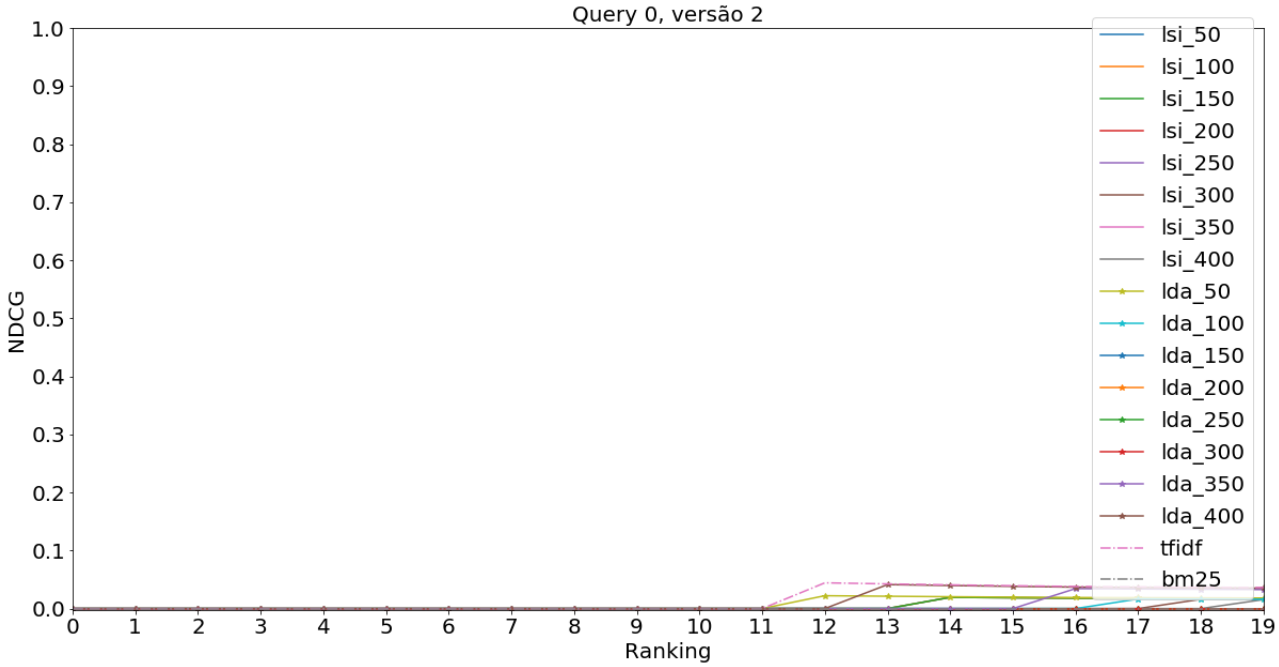


Figura A.3: PC - performance da *Query* 0, versão 3

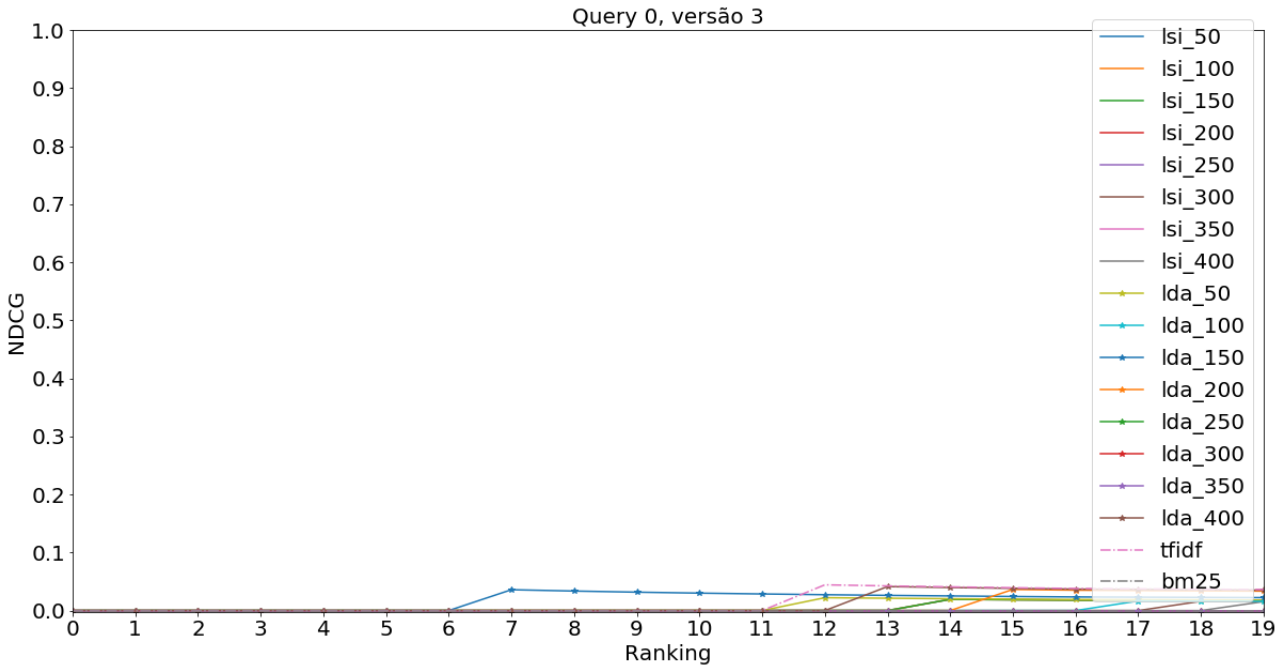
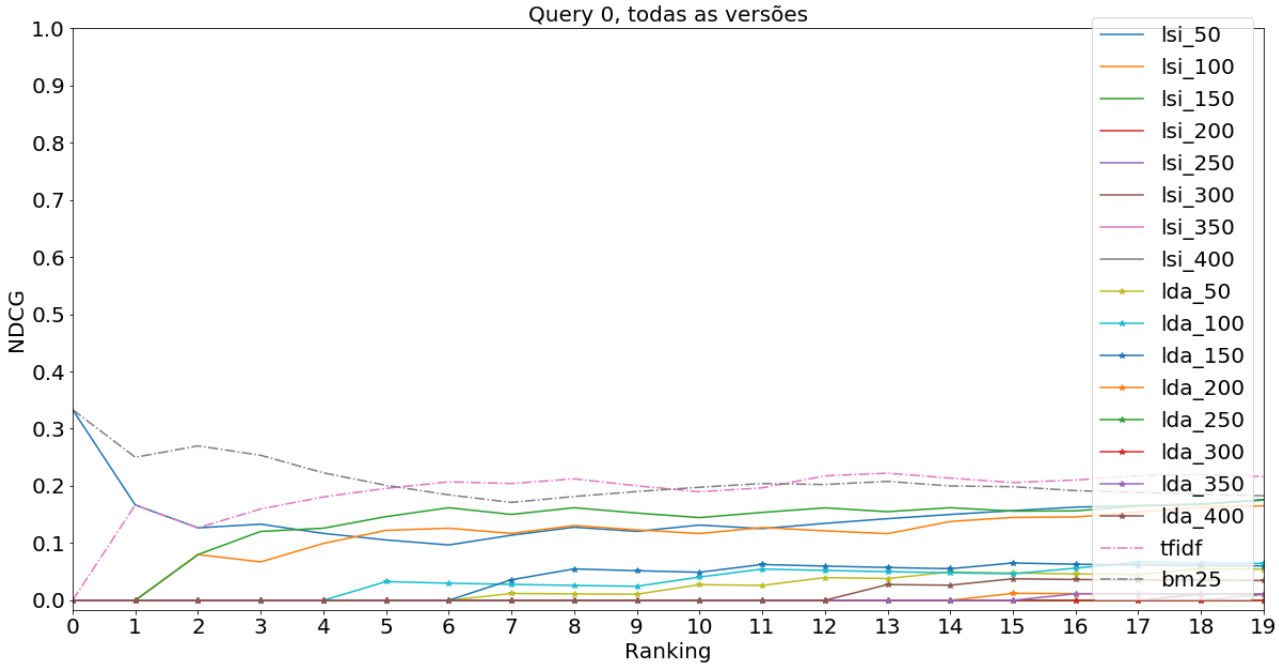


Figura A.4: PC - performance da *Query* 0, todas as versões



A.1.2 *Query 1*

A *Query 1* pretende localizar pareceres sobre casos em que o réu condenado pede reconhecimento do atenuante de confissão espontânea, mesmo verificado que se trata de réu reincidente no mesmo crime. Há julgados no sentido de aceitar ou rejeitar o acolhimento do atenuante nestes casos, porém isto não foi levado em consideração para atribuir escore de relevância aos documentos examinados. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  "A reincidência prevalece sobre a confissão espontânea"

- versão 2:

  "preponderância entre reincidência confissão espontânea"

- versão 3:

  "réu reincidente pede reconhecimento de sua confissão"

Quantidade de documentos relevantes: 10
Quantidade de documentos muito relevantes: 21

Figura A.5: PC - performance da *Query* 1, versão 1

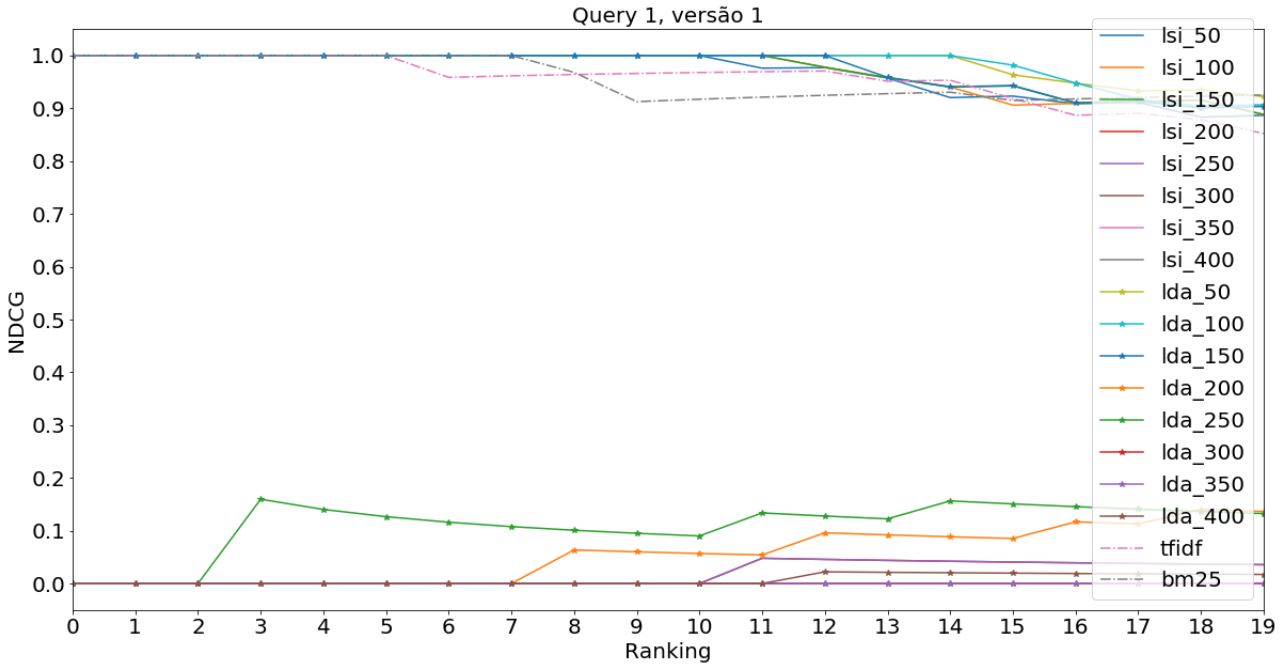


Figura A.6: PC - performance da *Query* 1, versão 2

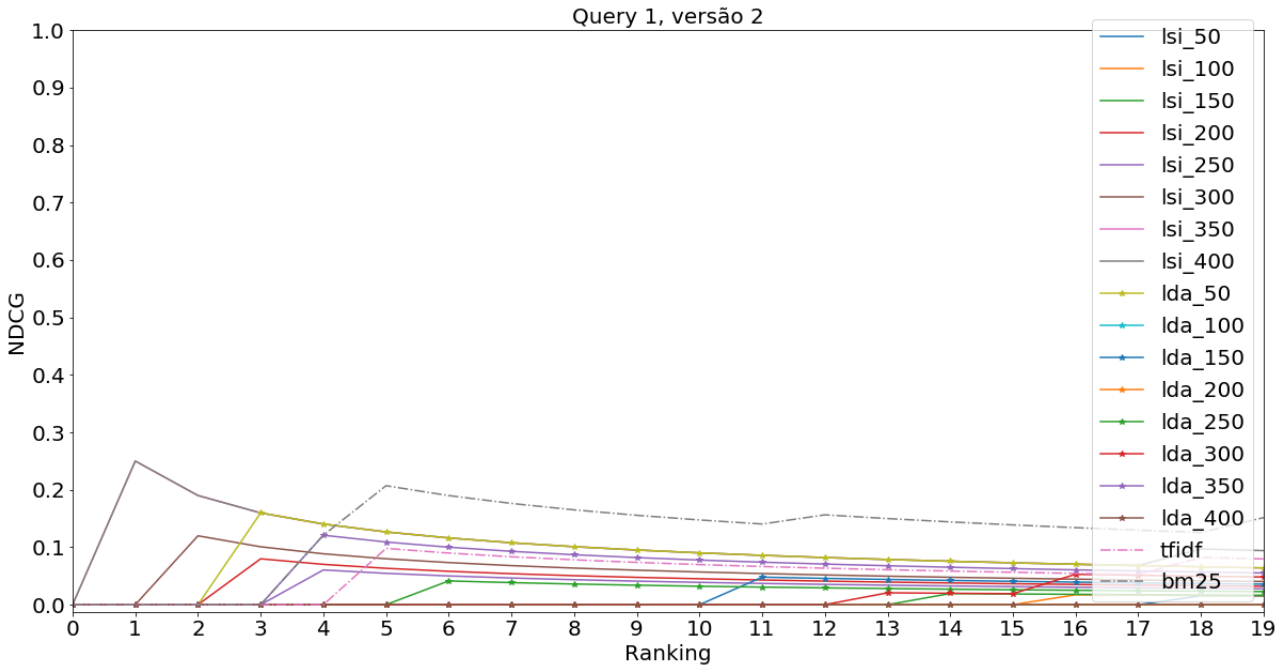


Figura A.7: PC - performance da *Query* 1, versão 3

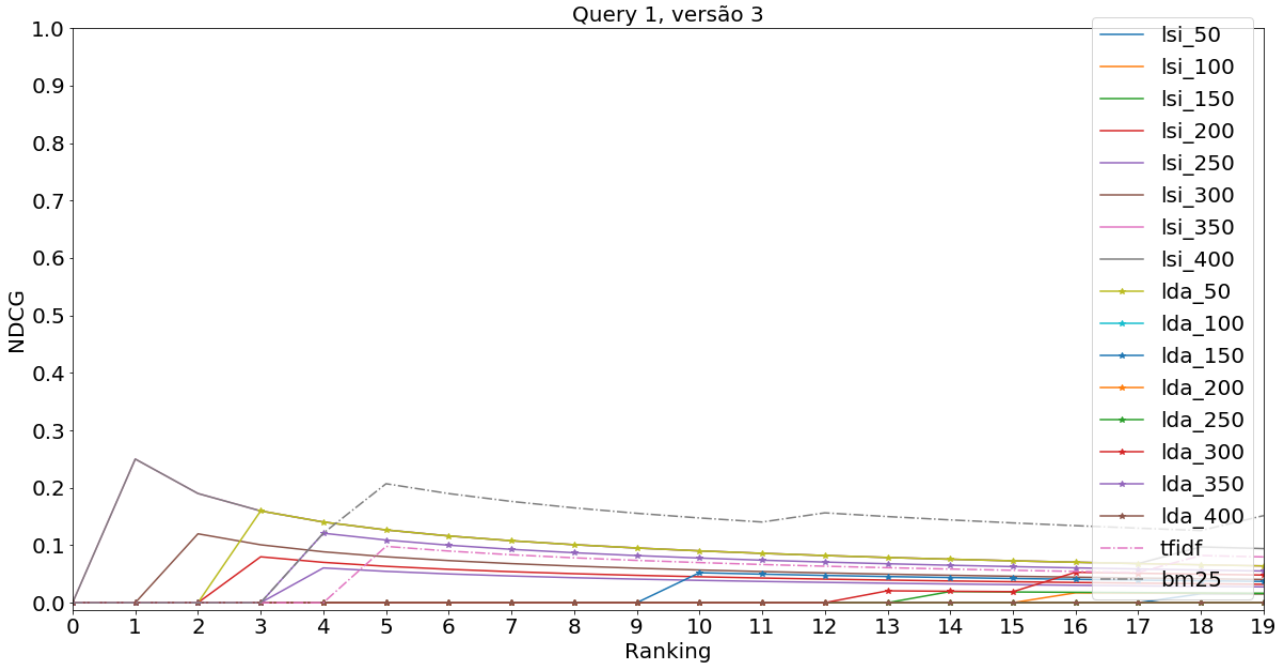
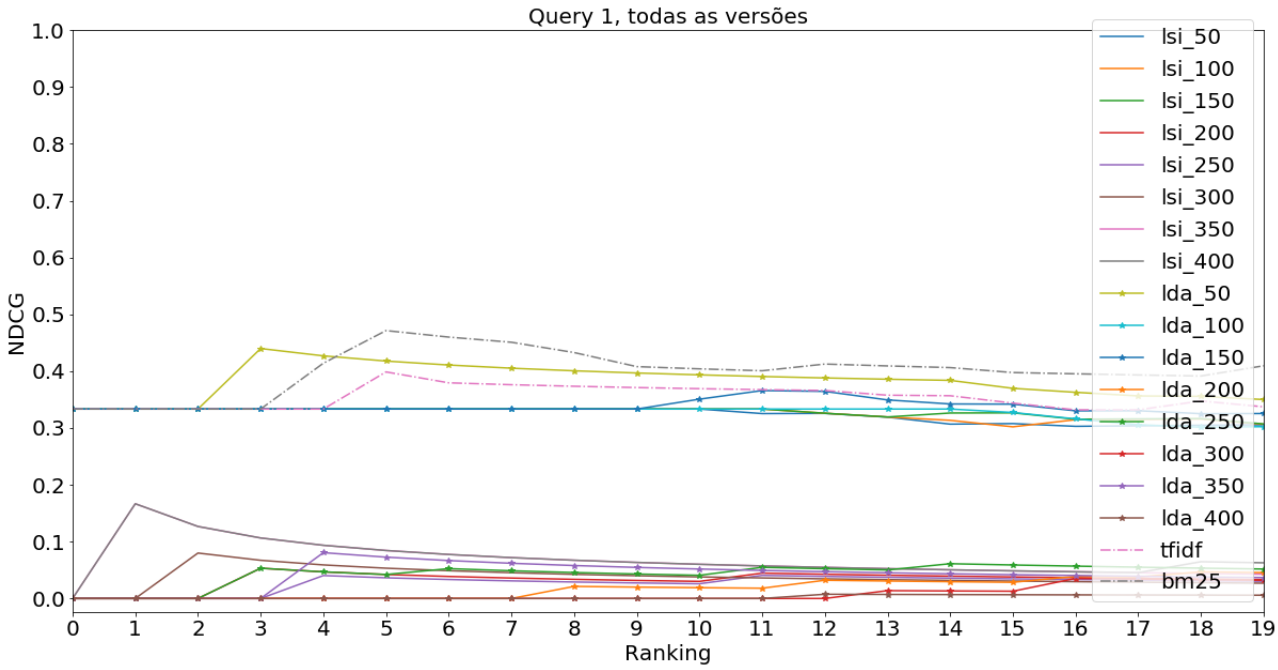


Figura A.8: PC - performance da *Query* 1, todas as versões



A.1.3 *Query 2*

A *Query 2* pretende localizar pareceres sobre casos em que o réu condenado por crime contra o patrimônio alega que o bem subtraído era de sua propriedade por tê-lo adquirido de forma legal, e que portanto sua condenação não procede. Há vários julgados que decidem que quando o bem subtraído é encontrado em poder do réu, a este se atribui o ônus de provar a propriedade e a procedência do bem se quiser ser absolvido. Para recuperar com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "apreensão do bem em posse do réu inverte o ônus da prova"

- versão 2:

  > "rés furtiva encontrada com o réu transfere a este o ônus de provar que o bem lhe pertence"

- versão 3:

  > "réu deve provar que o bem lhe pertence caso o objeto seja encontrado em sua posse"

Quantidade de documentos relevantes: 8

Quantidade de documentos muito relevantes: 18

Figura A.9: PC - performance da *Query* 2, versão 1

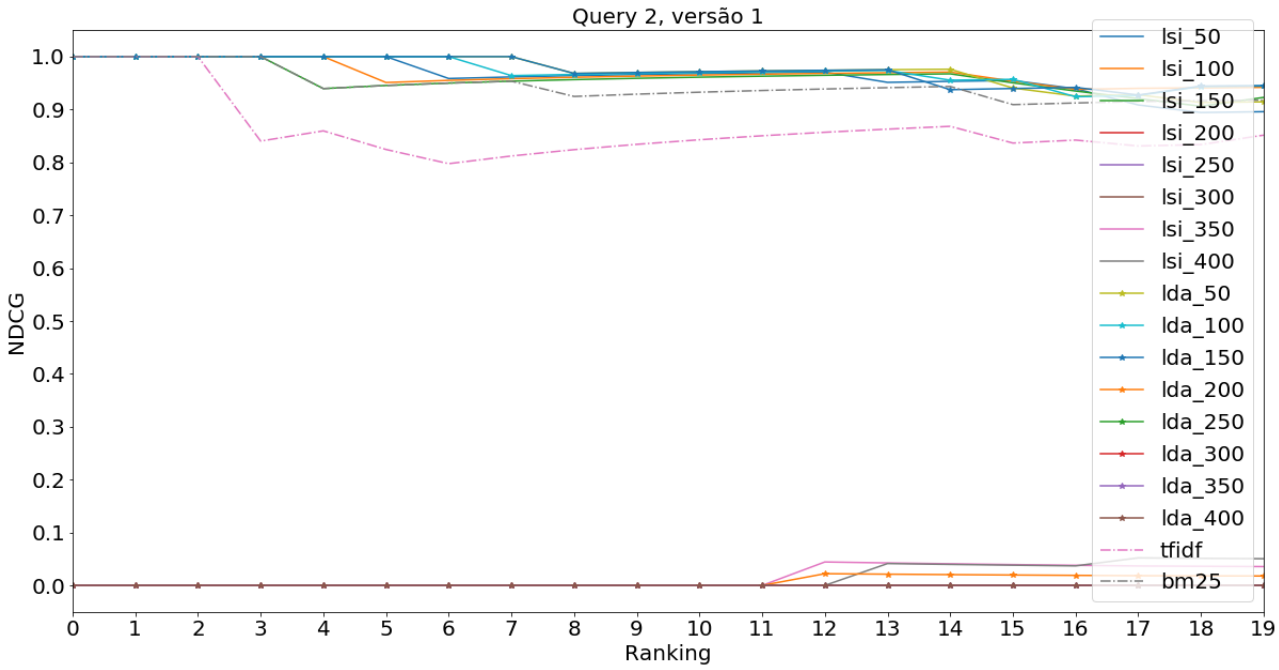


Figura A.10: PC - performance da *Query* 2, versão 2

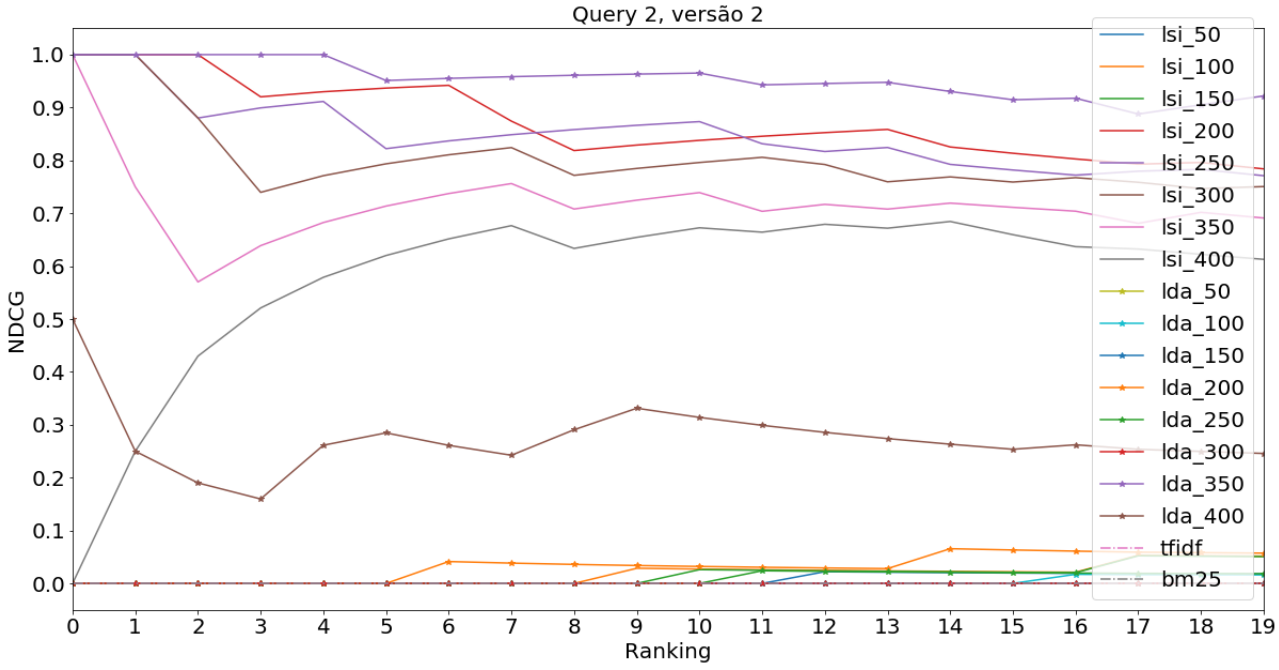


Figura A.11: PC - performance da *Query* 2, versão 3

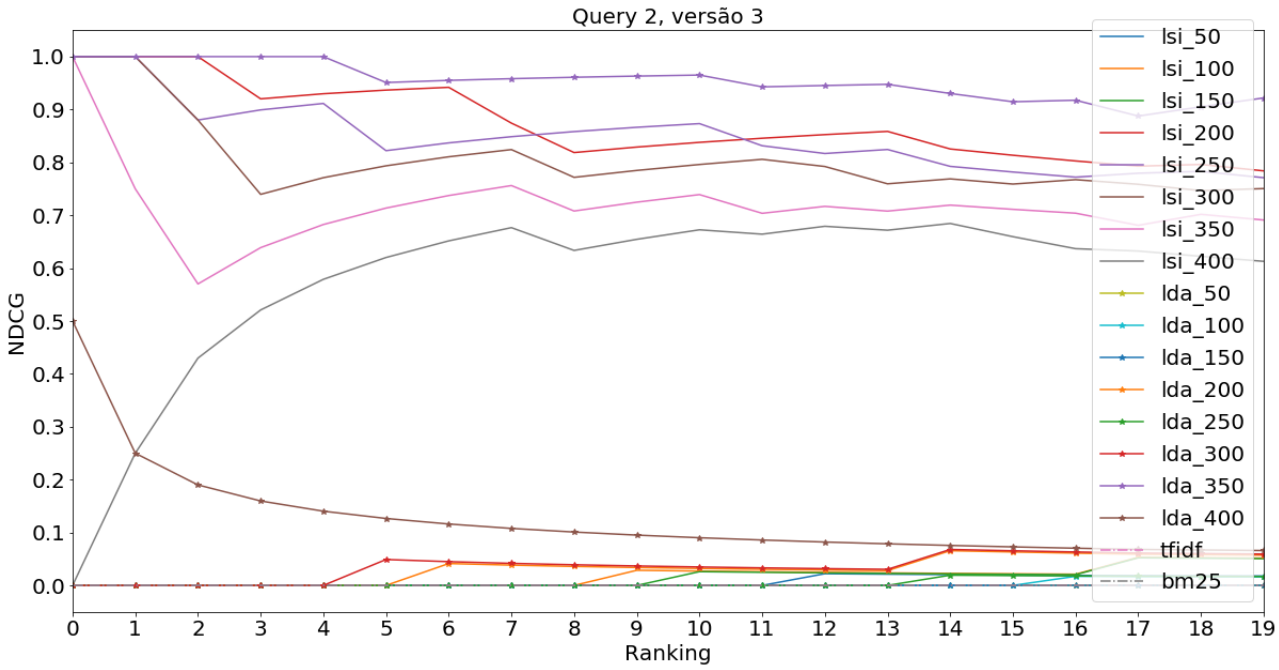
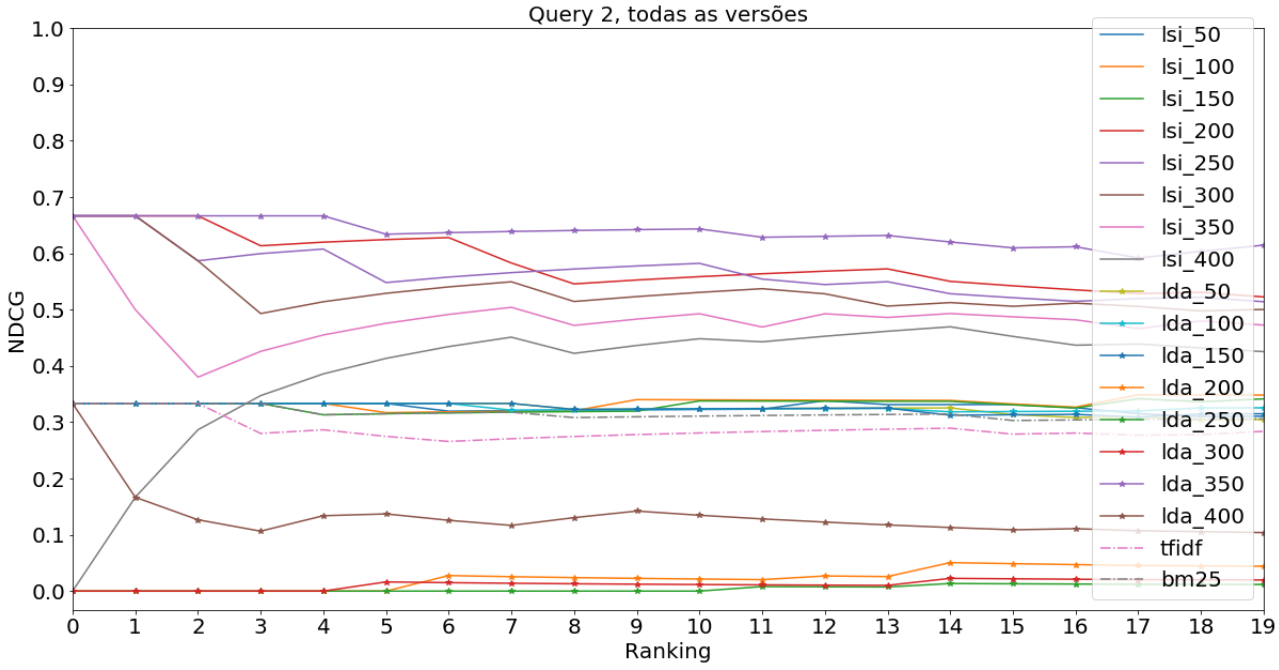


Figura A.12: PC - performance da *Query* 2, todas as versões



A.1.4 *Query 3*

A *Query 3* pretende localizar pareceres sobre casos em que o réu condenado por crime de furto cometido com concurso de agentes (com colaboração de uma pessoa) pede absolvição baseando-se no princípio da insignificância (ou bagatela), alegando que os bens furtados não tem valor significativo. Há julgados que determinam que este crime nestas circunstâncias não podem ser enquadrado no princípio da insignificância. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "tentativa de furto com concurso de agentes não admite a aplicação do princípio da insignificância"

- versão 2:

  > "princípio da bagatela não pode ser alegado mediante furto com concurso de agentes"

- versão 3:

  > "princípio da bagatela incompatível com furto cometido com apoio de duas ou mais pessoas"

Quantidade de documentos relevantes: 3

Quantidade de documentos muito relevantes: 2

Figura A.13: PC - performance da *Query* 3, versão 1

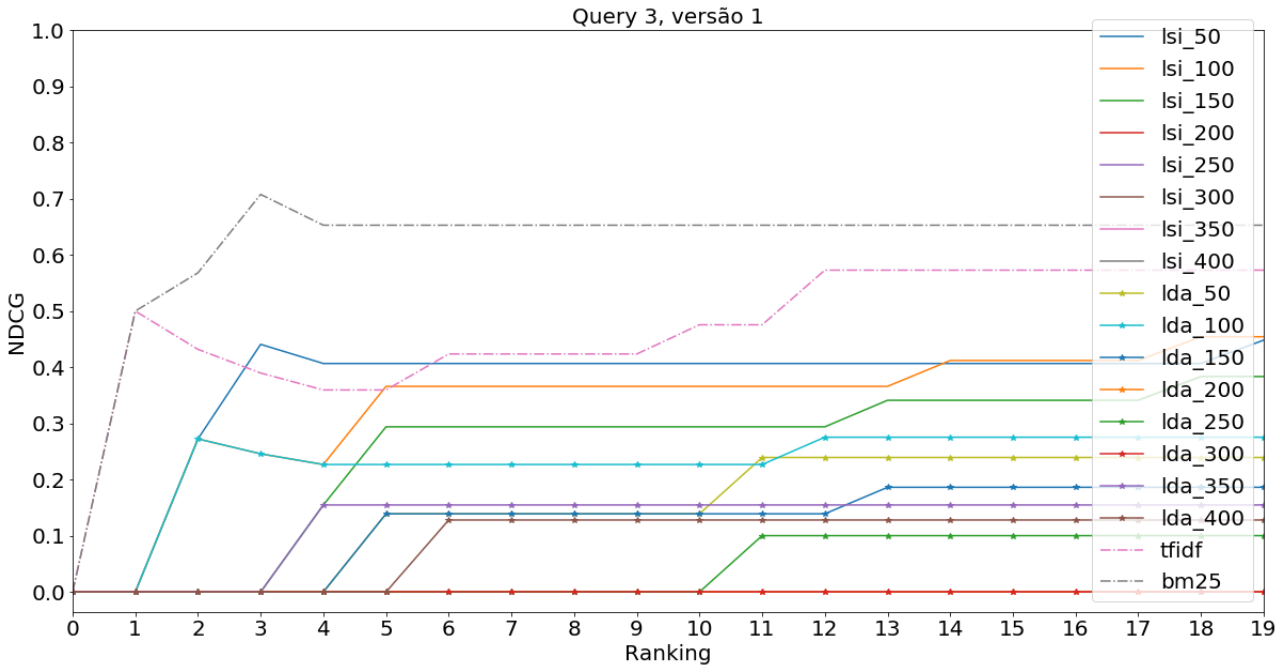


Figura A.14: PC - performance da *Query* 3, versão 2

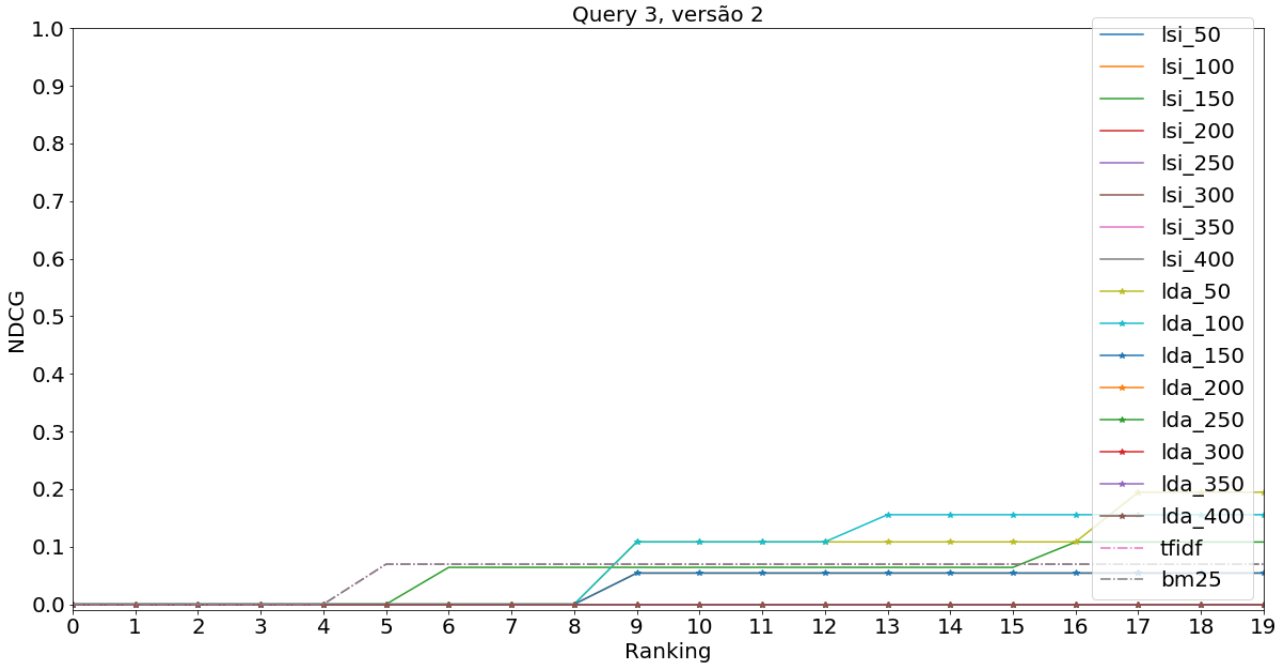


Figura A.15: PC - performance da *Query* 3, versão 3

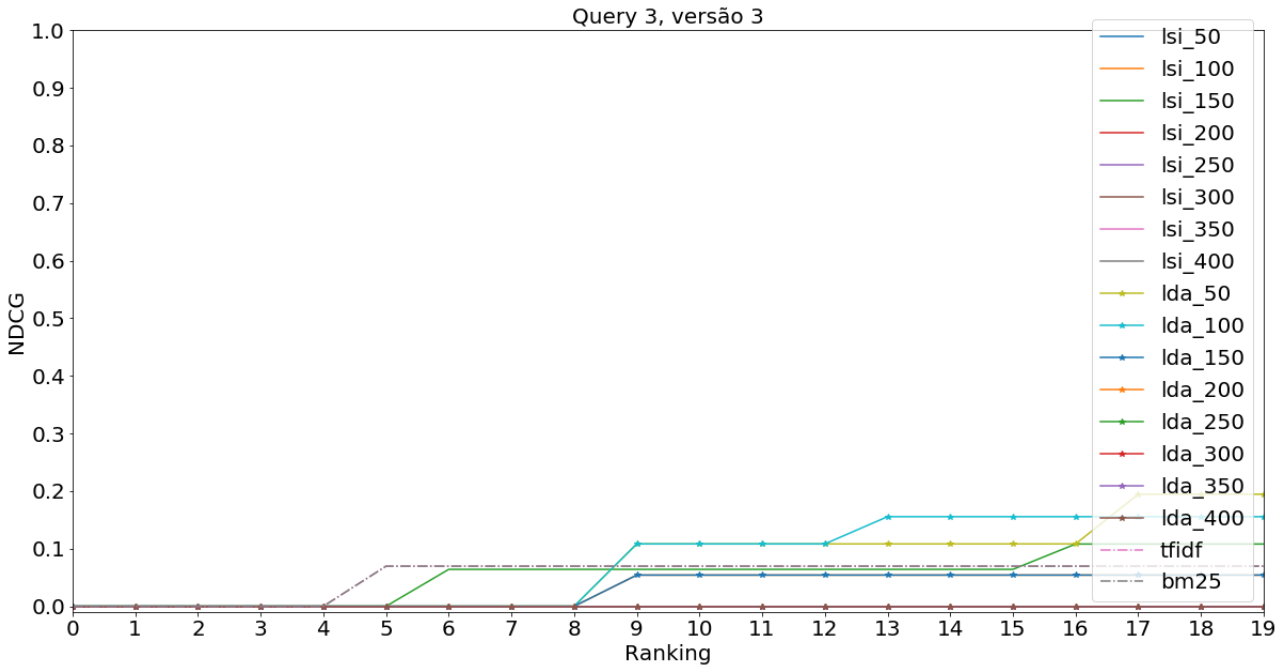
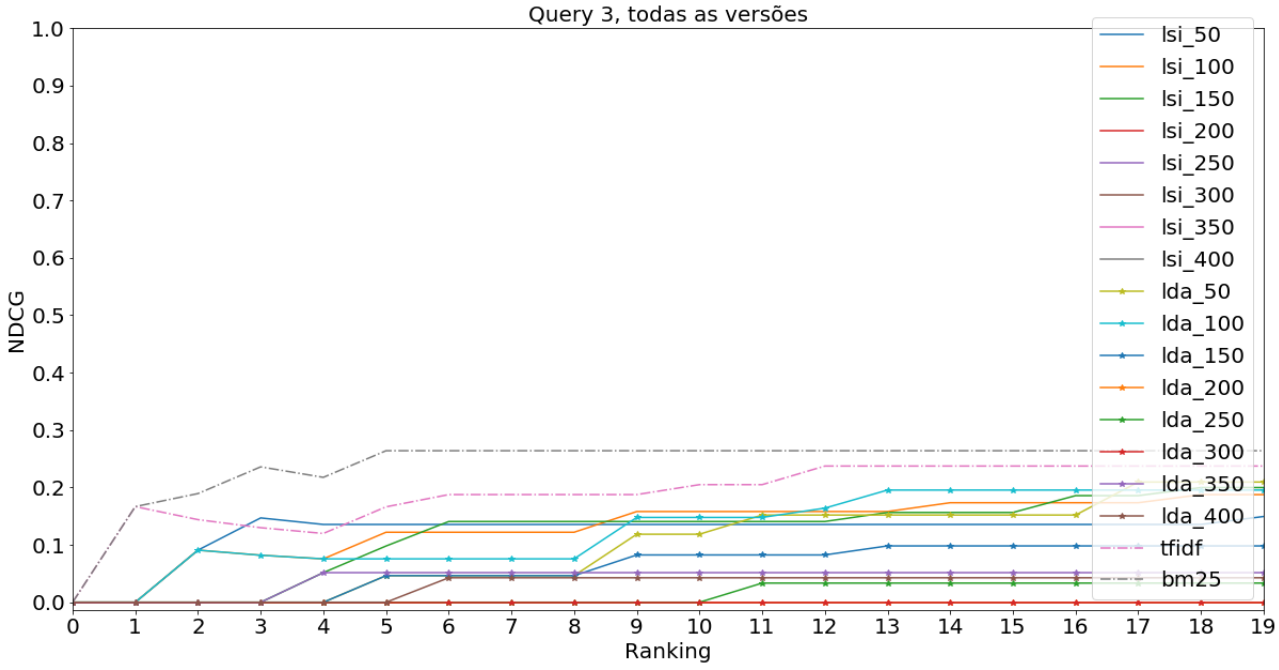


Figura A.16: PC - performance da *Query* 3, todas as versões



A.1.5 *Query 4*

A *Query 4* pretende localizar pareceres sobre casos em que o réu condenado por crime de roubo circunstanciado (com emprego de arma) alega que a arma utilizada não tinha potencial lesivo, e que portanto a majorante de ameaça com arma deve ser removida. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "Uso de arma de fogo desmuniciada impede o enquadramento como roubo circunstanciado"

- versão 2:

  > "Não se pode considerar roubo circunstanciado se a arma de fogo empregada pelo réu não tiver potencial lesivo"

- versão 3:

  > "desclassificação de roubo circunstanciado considerando artefato sem potencial lesivo ou simulacro"

Quantidade de documentos relevantes: 2

Quantidade de documentos muito relevantes: 4

Figura A.17: PC - performance da *Query* 4, versão 1

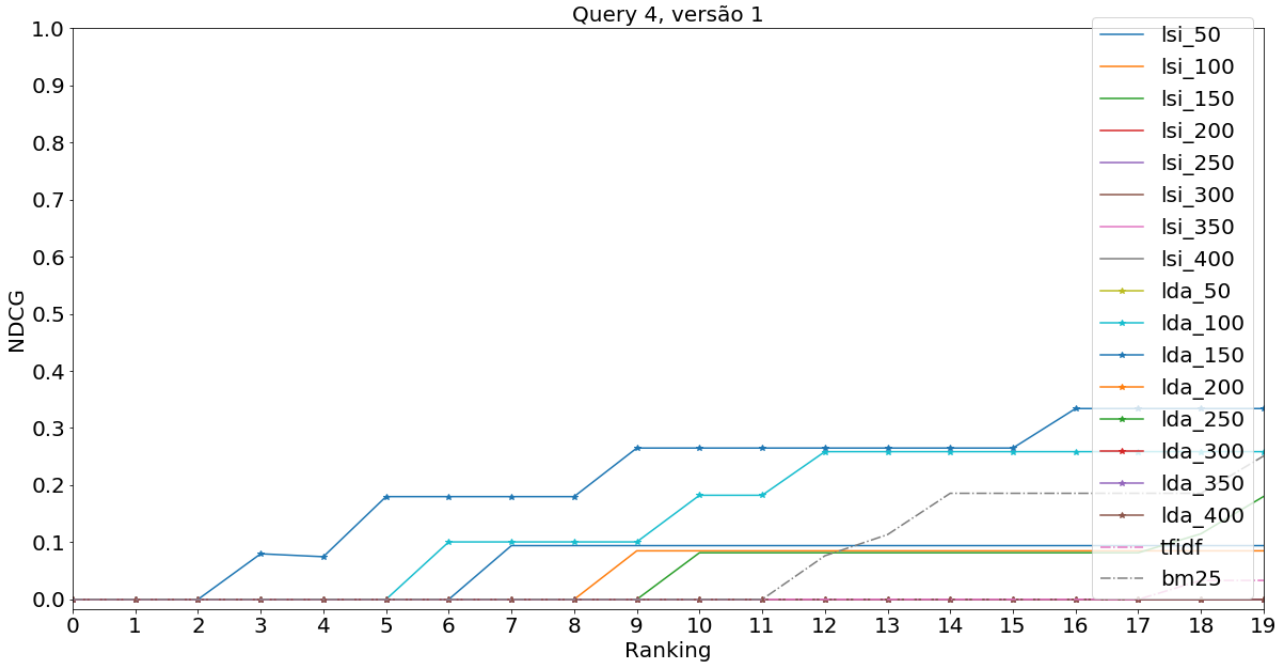


Figura A.18: PC - performance da *Query* 4, versão 2

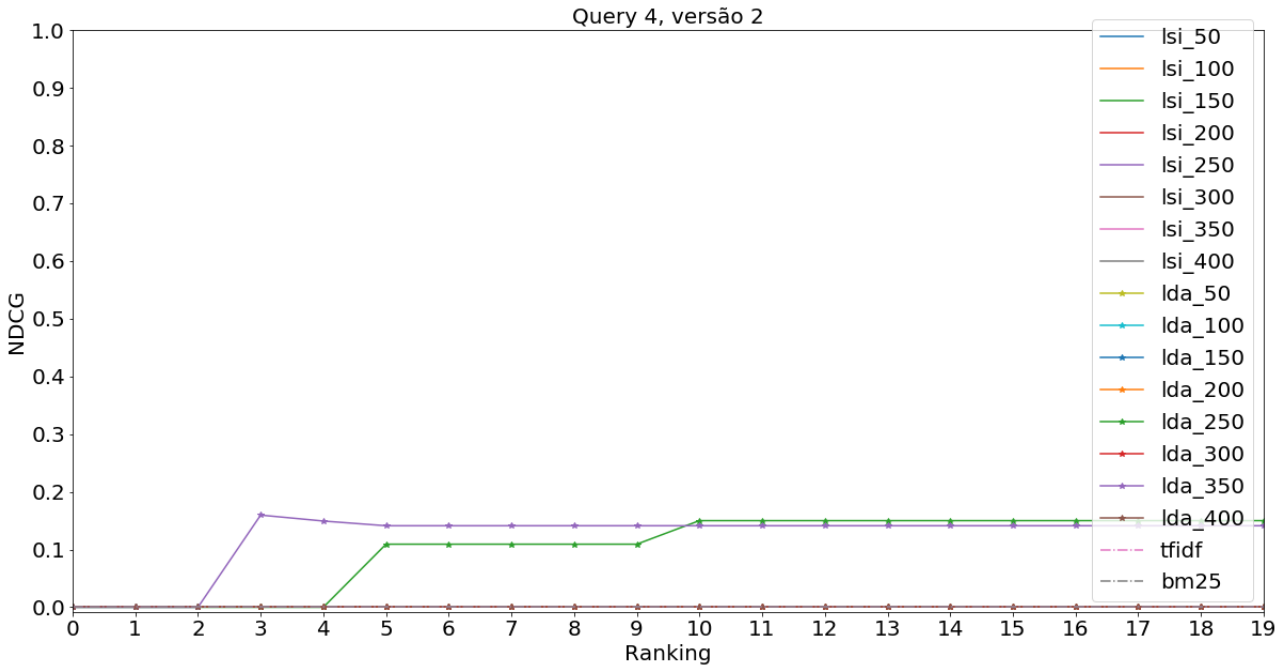


Figura A.19: PC - performance da *Query* 4, versão 3

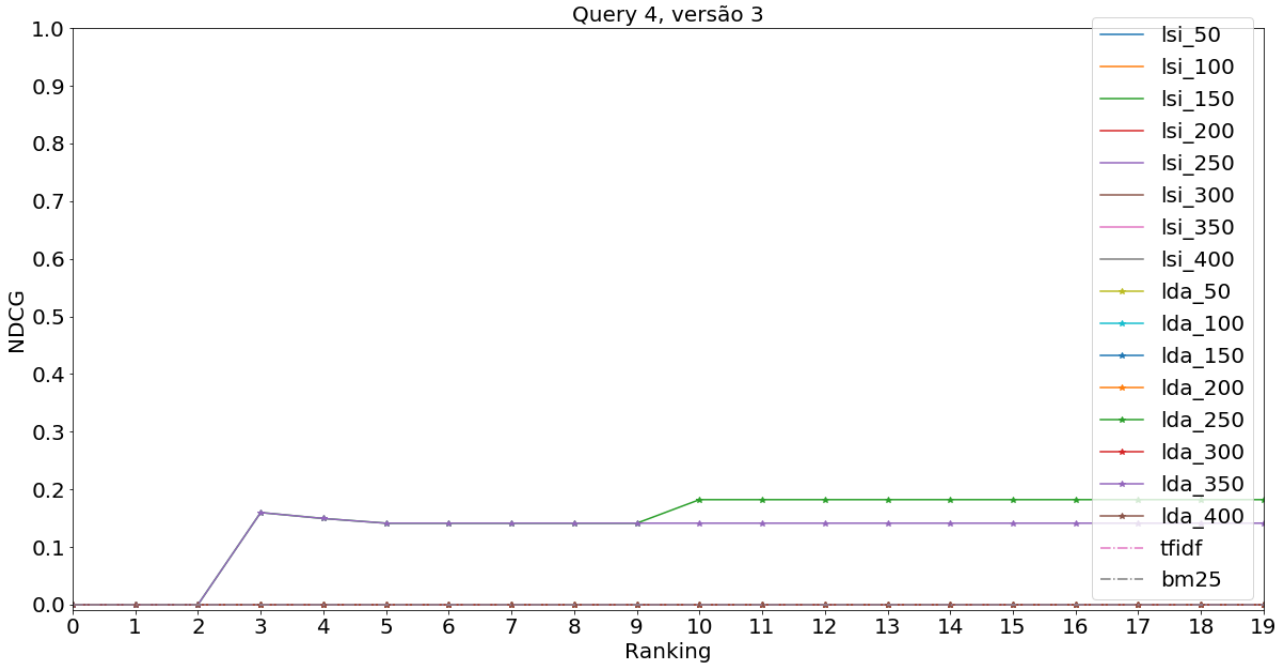
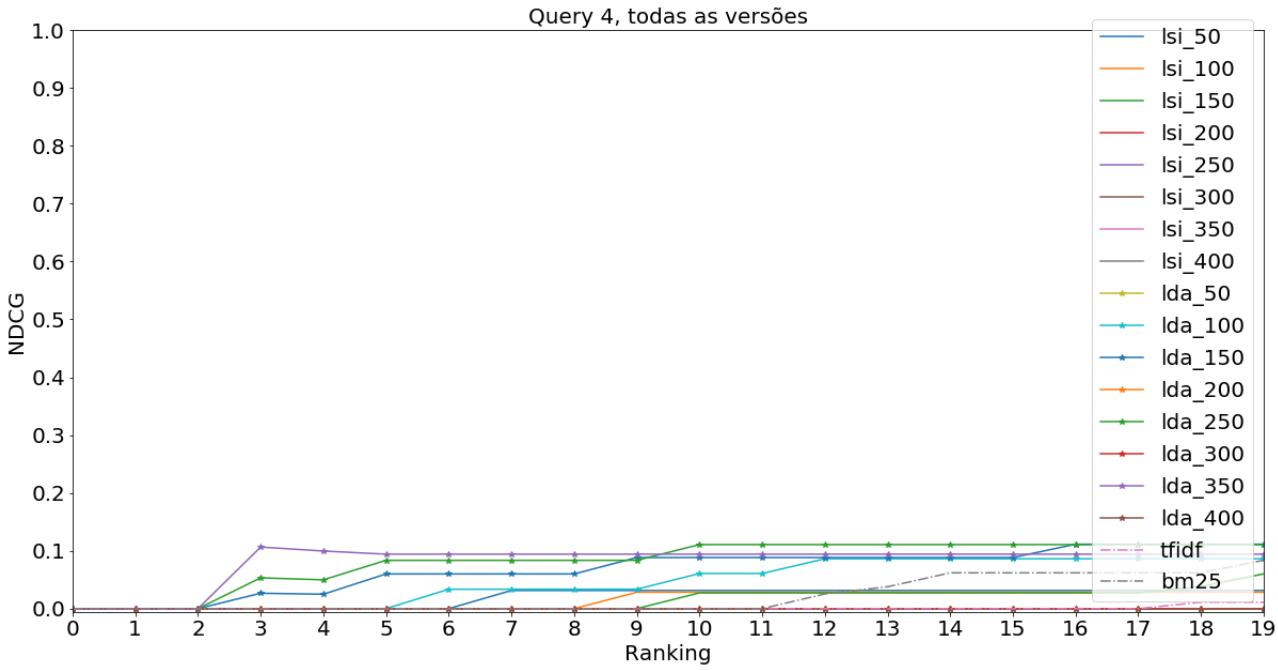


Figura A.20: PC - performance da *Query* 4, todas as versões



A.1.6 *Query 5*

A *Query 5* pretende localizar pareceres sobre casos em que o réu pede indulto, porém fora condenado por crime ediondo ou equiparado. Há vários julgados negando o benefício ao réu. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  "impossibilidade de concessão de indulto para crimes hediondos ou similares"

- versão 2:

  "indulto inviável para réu condenado por crime hediondo ou assemelhado"

- versão 3:

  "incabível a concessão de indulto para casos de crime hediondo ou tráfico de drogas"

Quantidade de documentos relevantes: 0

Quantidade de documentos muito relevantes: 6

Figura A.21: PC - performance da *Query* 5, versão 1

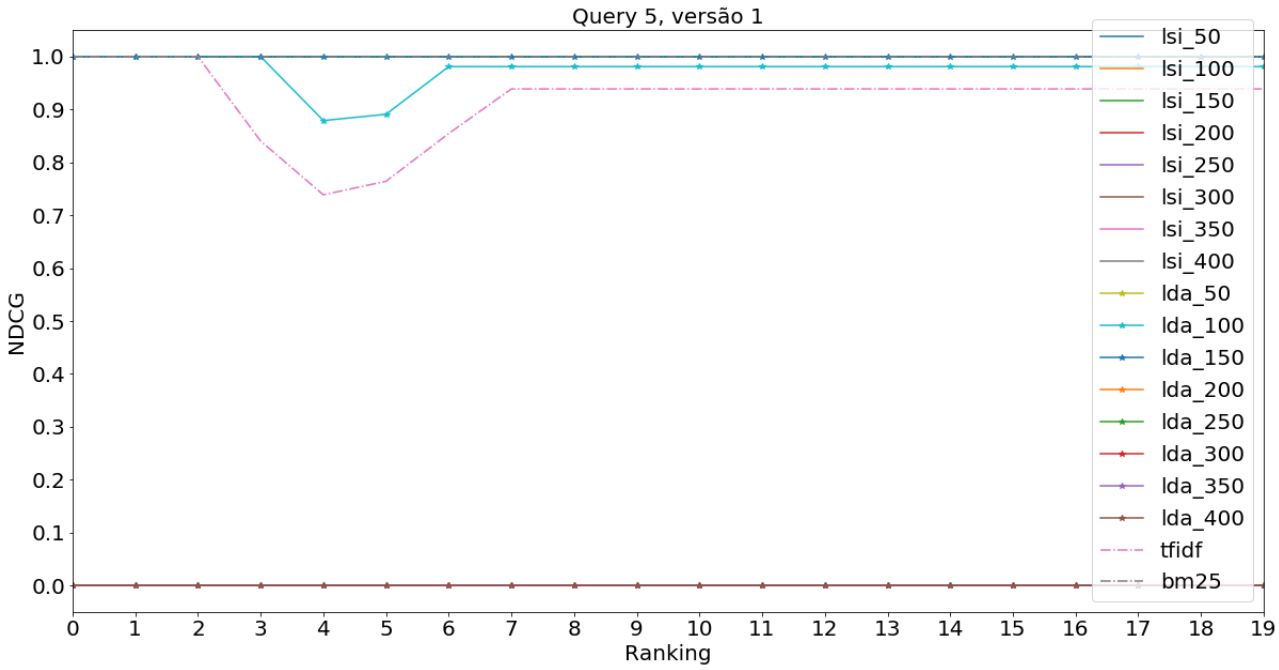


Figura A.22: PC - performance da *Query* 5, versão 2

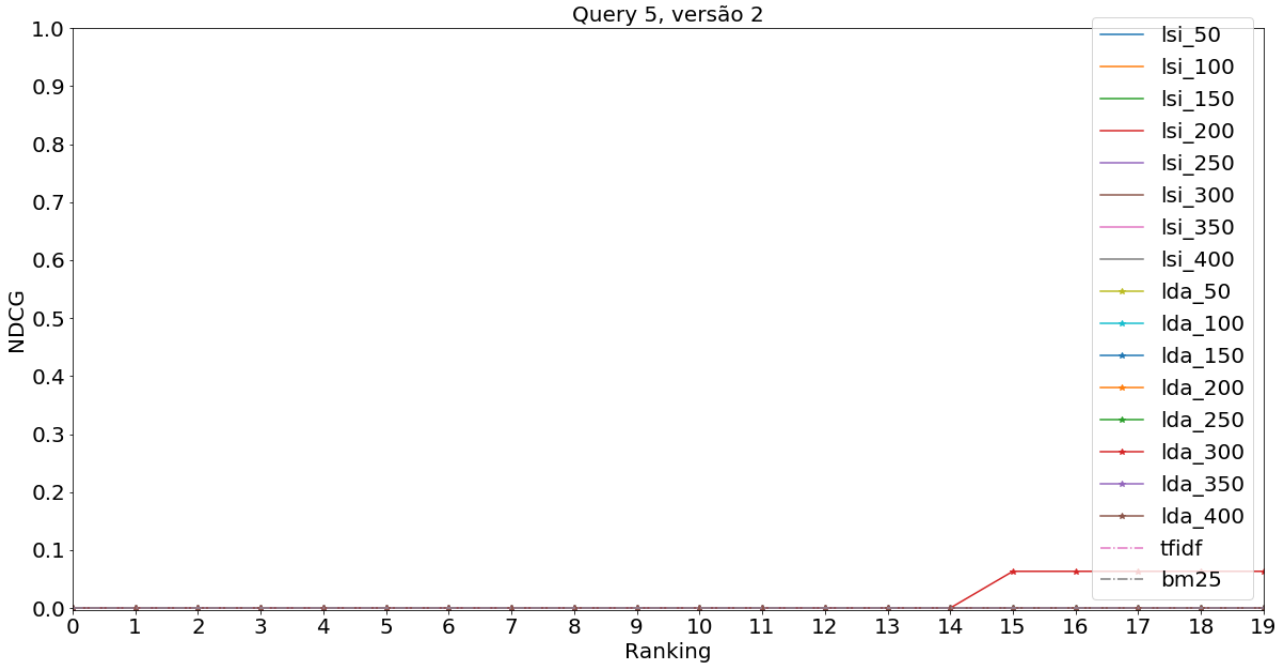


Figura A.23: PC - performance da *Query* 5, versão 3

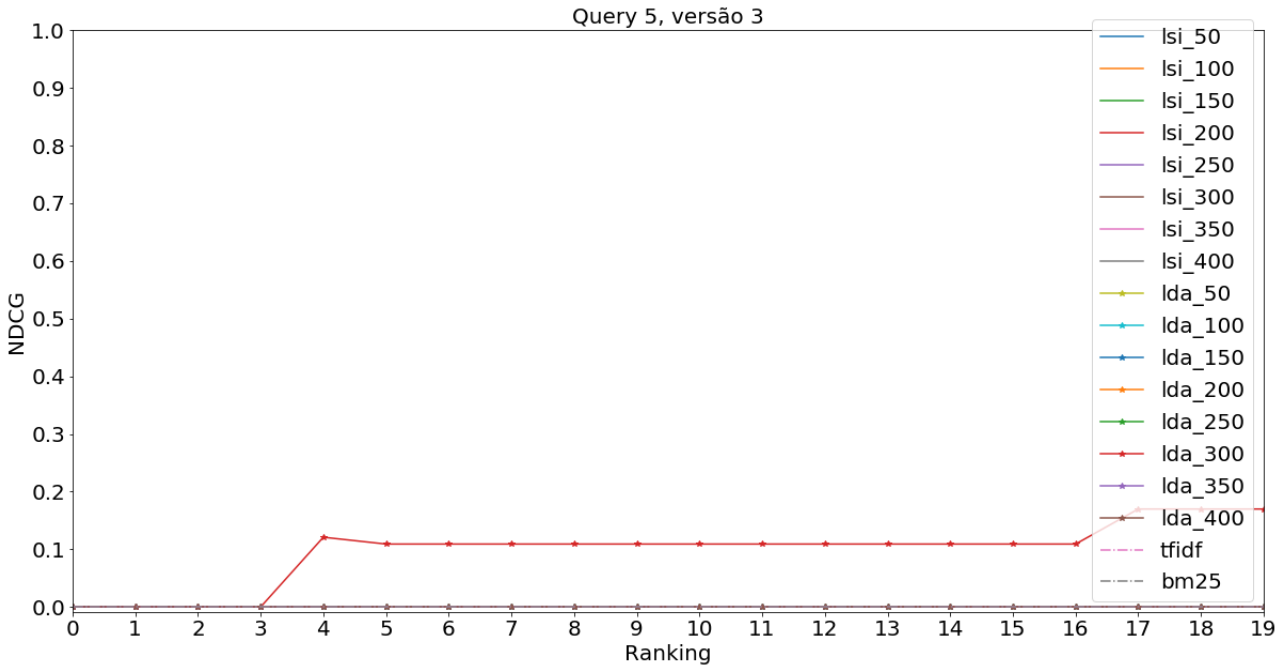
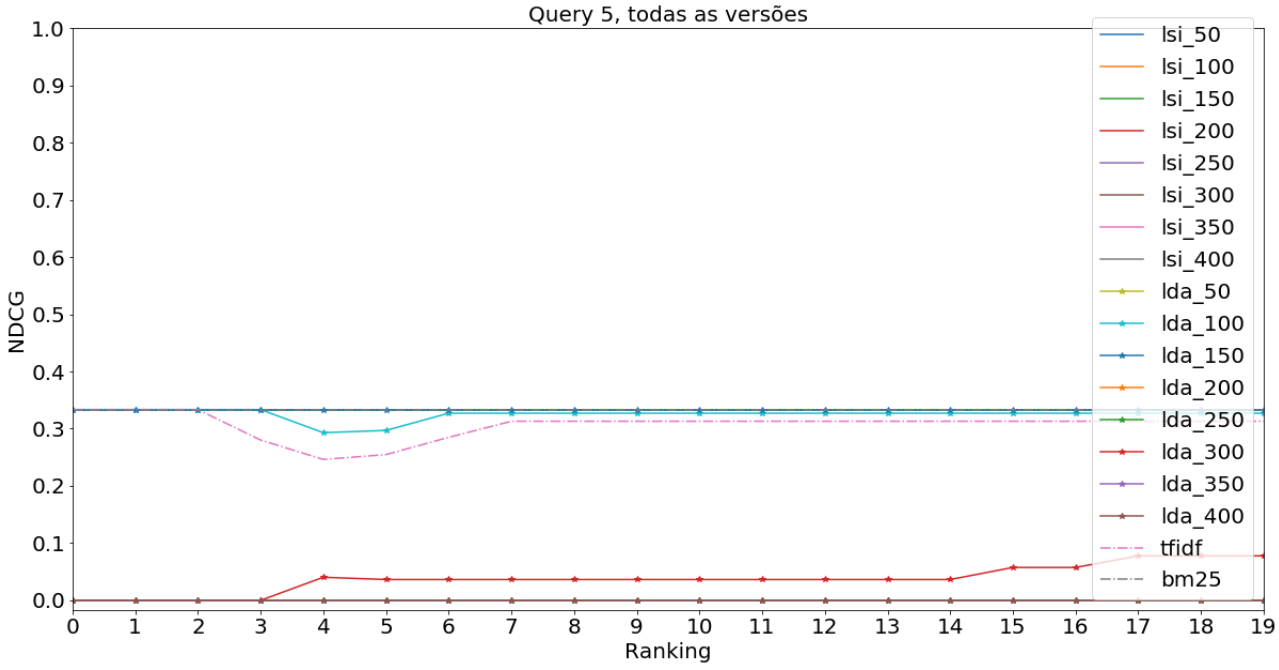


Figura A.24: PC - performance da *Query* 5, todas as versões



A.1.7 *Query 6*

A *Query 6* pretende localizar pareceres sobre casos em que o réu pede absolvição do crime em que foi condenado alegando que estava embriagado. O código penal prevê absolvição somente para os casos de embriaguês ou entorpecimento involuntário, desde que comprovado. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "Alegação de absolvição por embriaguês sem comprovar que esta ocorreu de forma acidental"

- versão 2:

  > "inviável absolver por embriaguês voluntária"

- versão 3:

  > "absolvição possível somente mediante embriaguês involuntária e comprovada"

Quantidade de documentos relevantes: 4
Quantidade de documentos muito relevantes: 3

Figura A.25: PC - performance da *Query* 6, versão 1

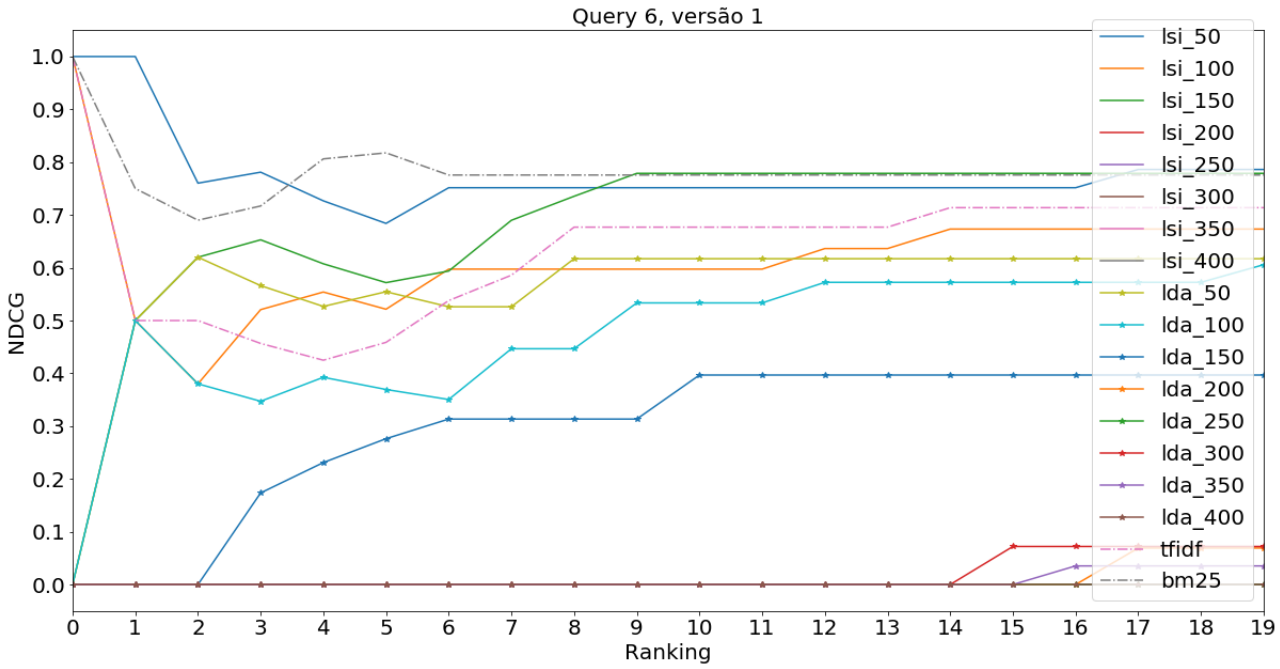


Figura A.26: PC - performance da *Query* 6, versão 2

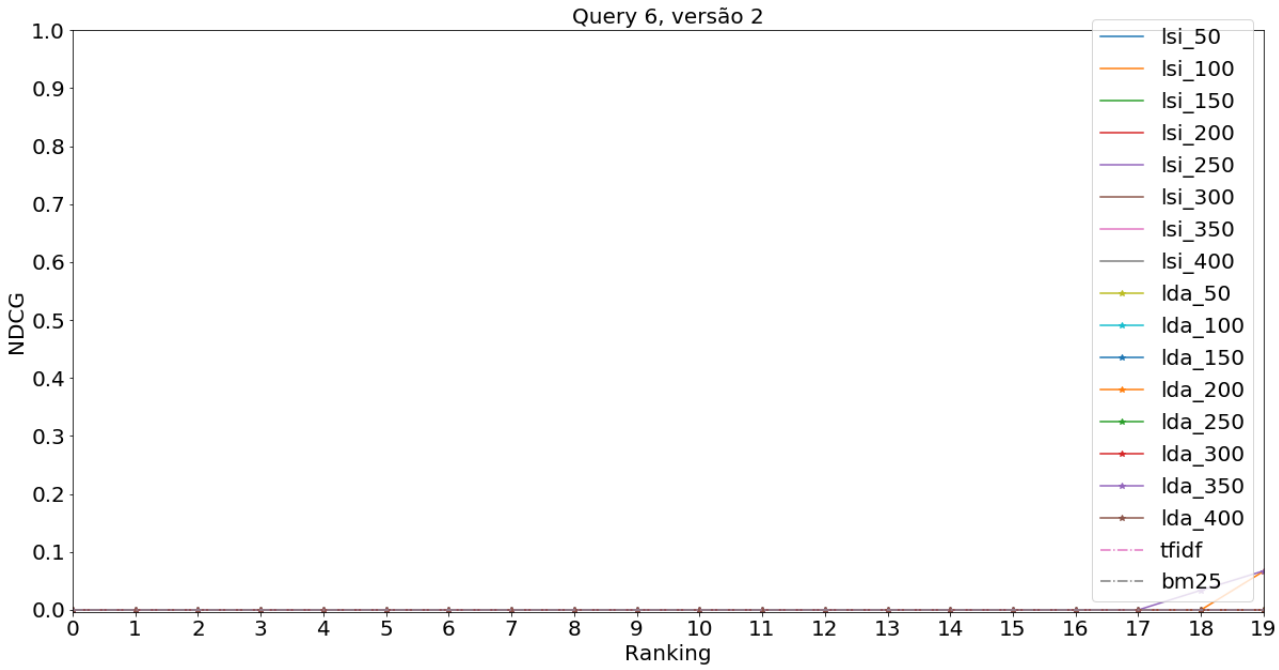


Figura A.27: PC - performance da *Query* 6, versão 3

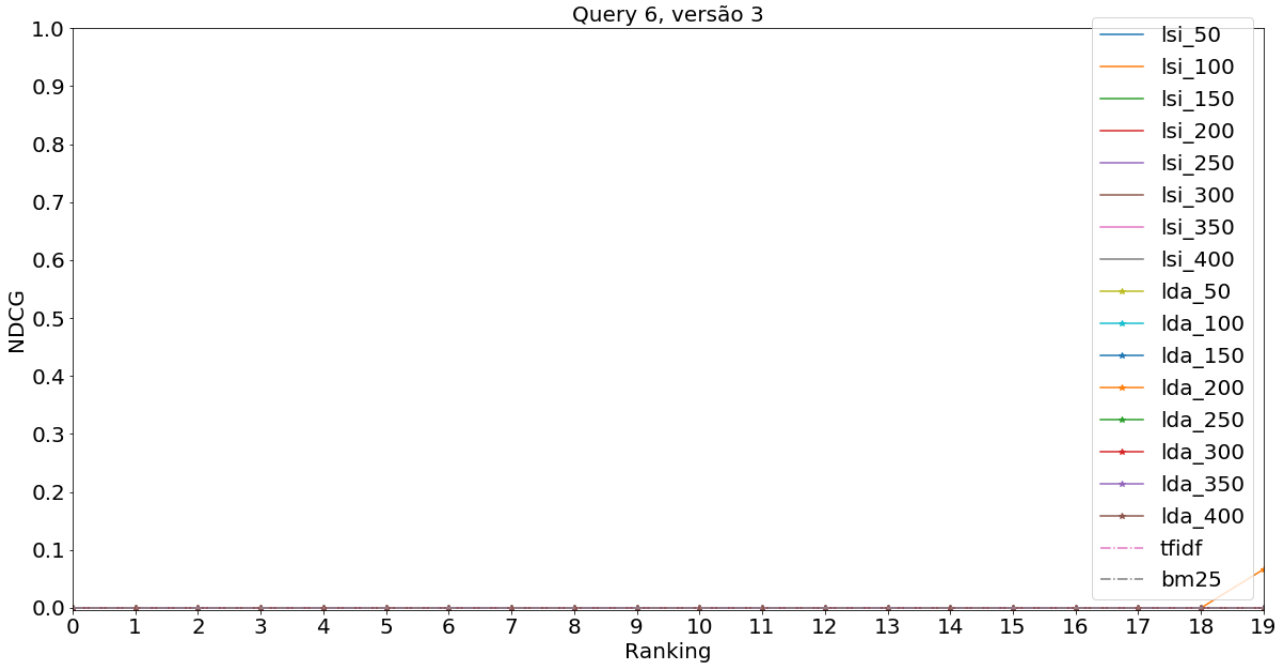
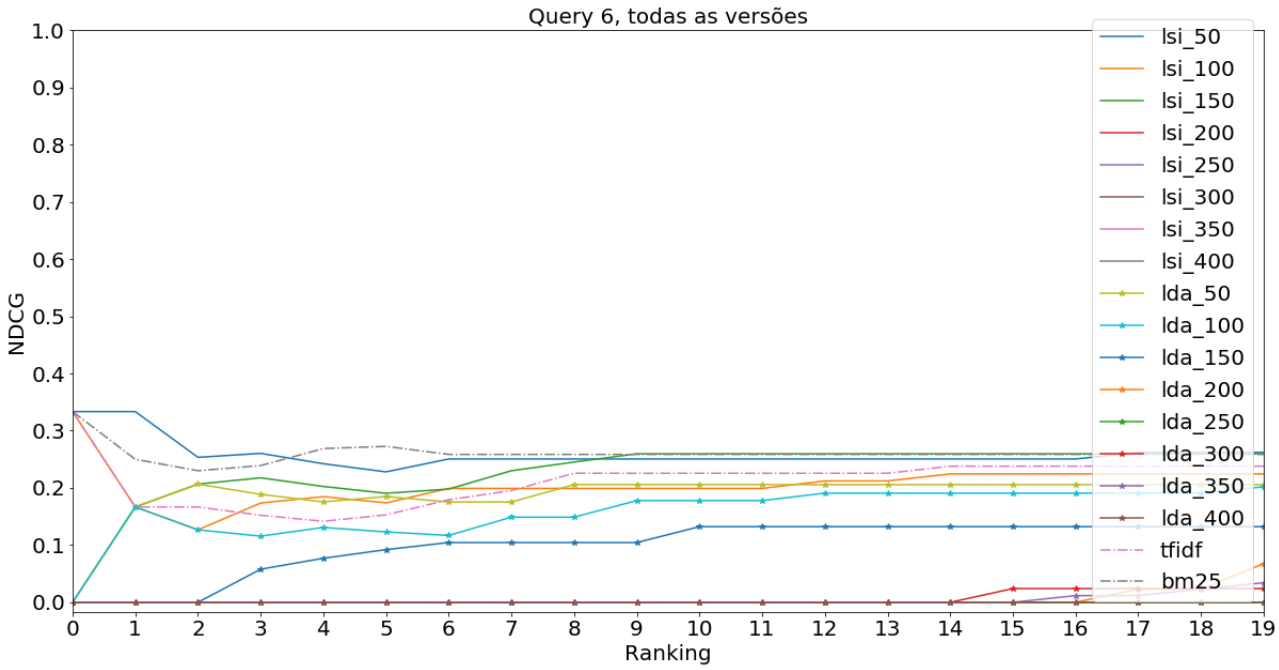


Figura A.28: PC - performance da *Query* 6, todas as versões



A.1.8 *Query 7*

A *Query 7* pretende localizar pareceres em casos em que o réu condenado por crime de violência doméstica pede absolvição baseando-se no princípio da insignificância. Há julgados que determinam que o pedido não pode ser atendido, dada a gravidade do crime de violência doméstica. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "princípio da insignificância não pode ser considerado em casos de violência no âmbito das relações domésticas"

- versão 2:

  > "crime de violência doméstica não admite aplicação do princípio da insignificância"

- versão 3:

  > "princípio da insignificância incabível quando a violência ou ameaça ocorre no âmbito das relações domésticas"

Quantidade de documentos relevantes: 0

Quantidade de documentos muito relevantes: 5

Figura A.29: PC - performance da *Query* 7, versão 1

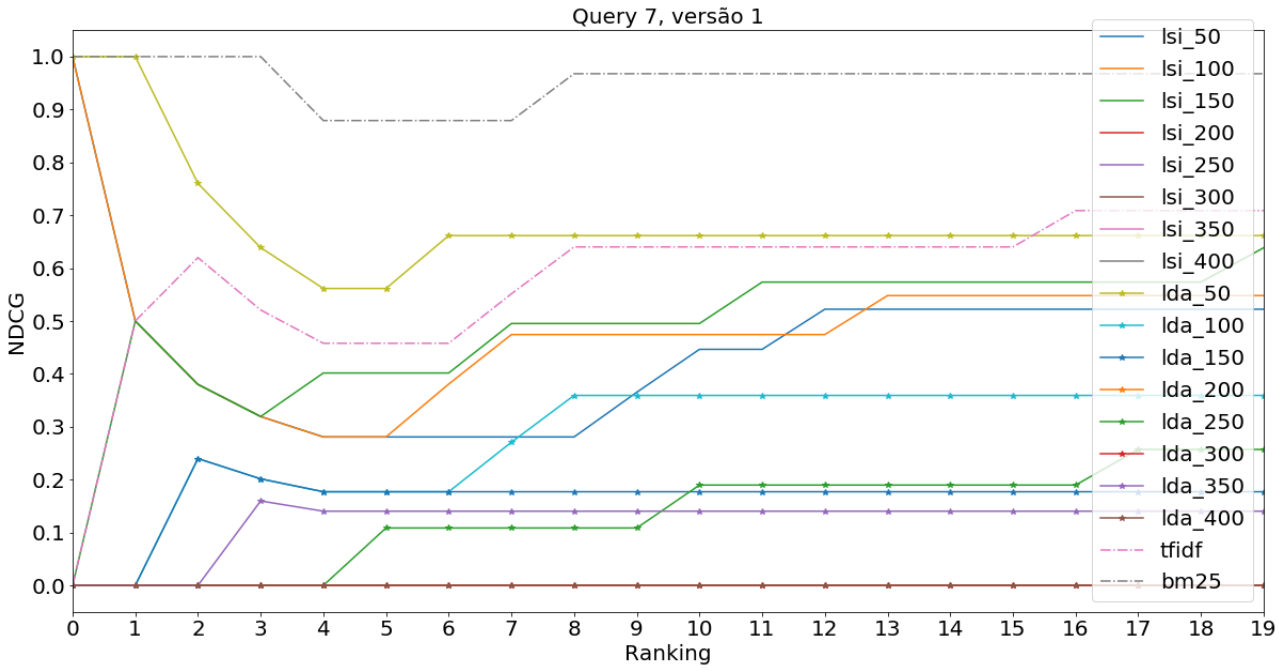


Figura A.30: PC - performance da *Query* 7, versão 2

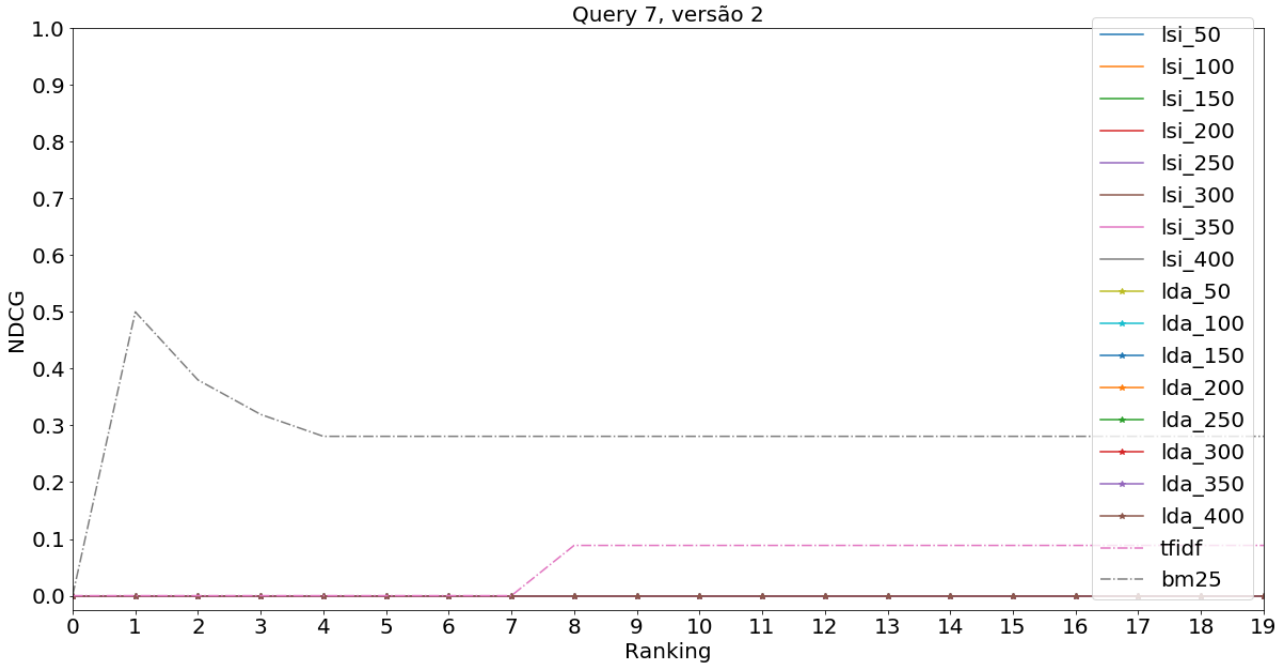


Figura A.31: PC - performance da *Query* 7, versão 3

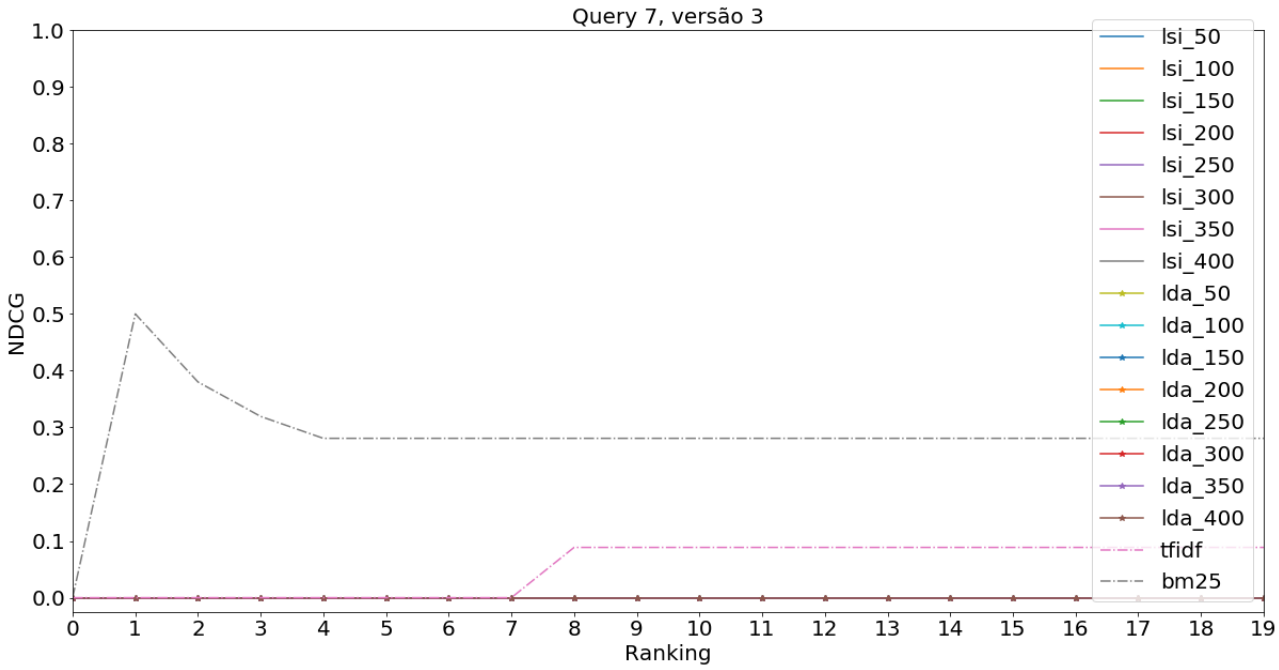
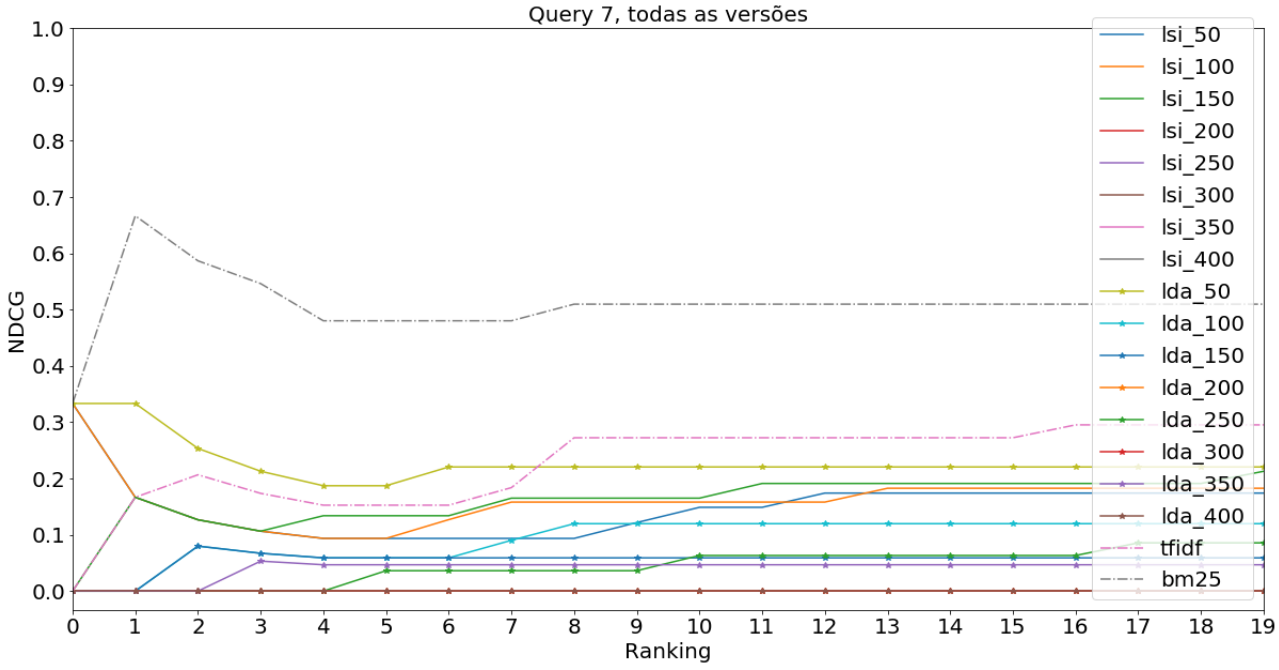


Figura A.32: PC - performance da *Query* 7, todas as versões



A.1.9 *Query 8*

A *Query 8* pretende localizar pareceres em casos em que o réu condenado por crime de violência doméstica pede absolvição alegando que a contenda que originou o crime fora resolvida. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  "reconciliação não absolve agressor de crime de ameaça contra a vítima"

- versão 2:

  "absolvição de crime de ameaça alegando que o casal reatou"

- versão 3:

  "absolvição da ameaça é incabível ainda que o réu tenha retornado a convivência do lar"

Quantidade de documentos relevantes: 0
Quantidade de documentos muito relevantes: 3

Figura A.33: PC - performance da *Query* 8, versão 1

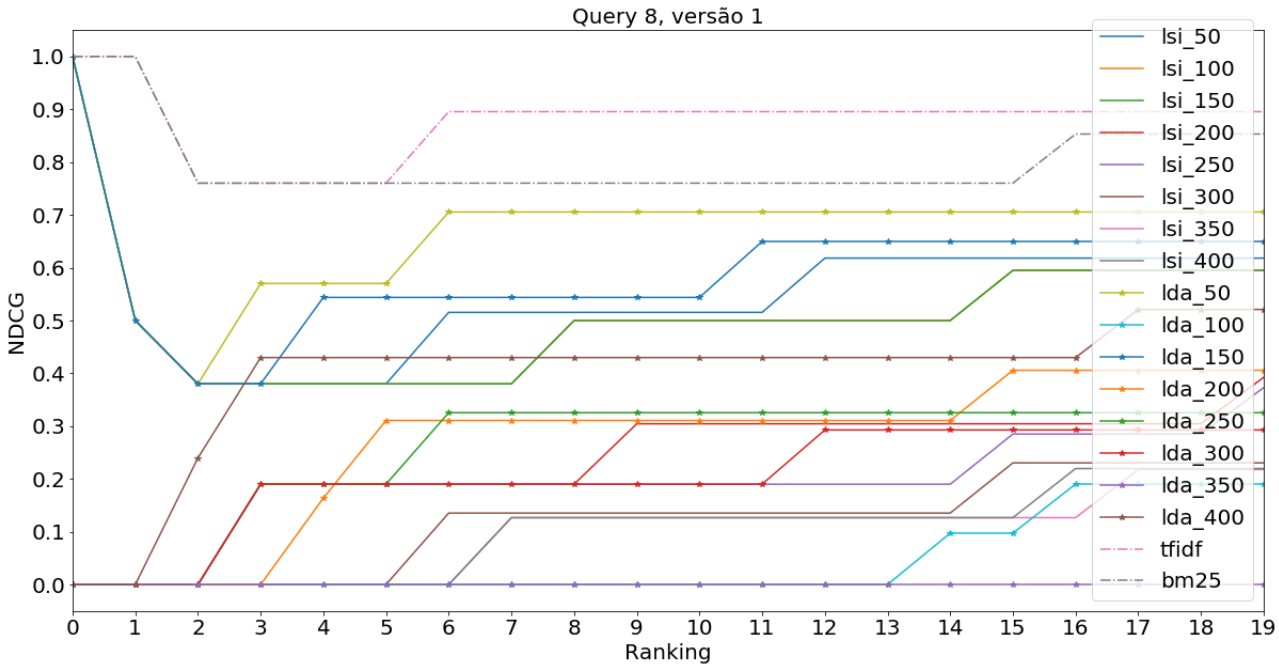


Figura A.34: PC - performance da *Query* 8, versão 2

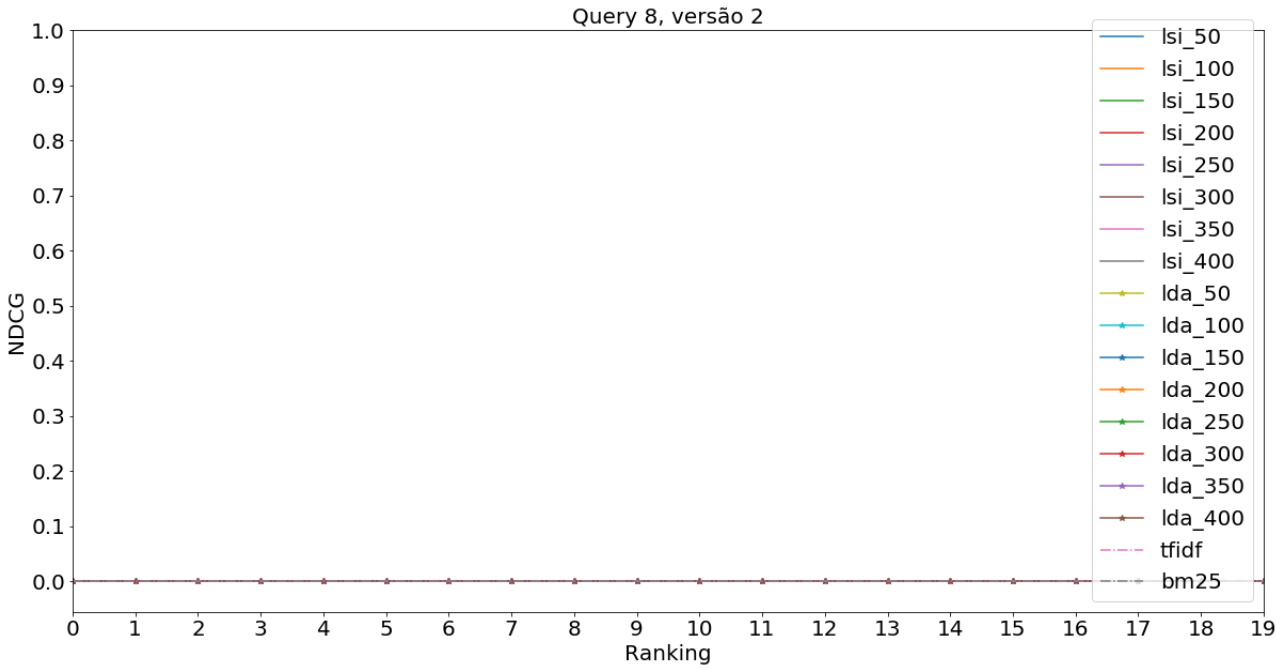


Figura A.35: PC - performance da *Query* 8, versão 3

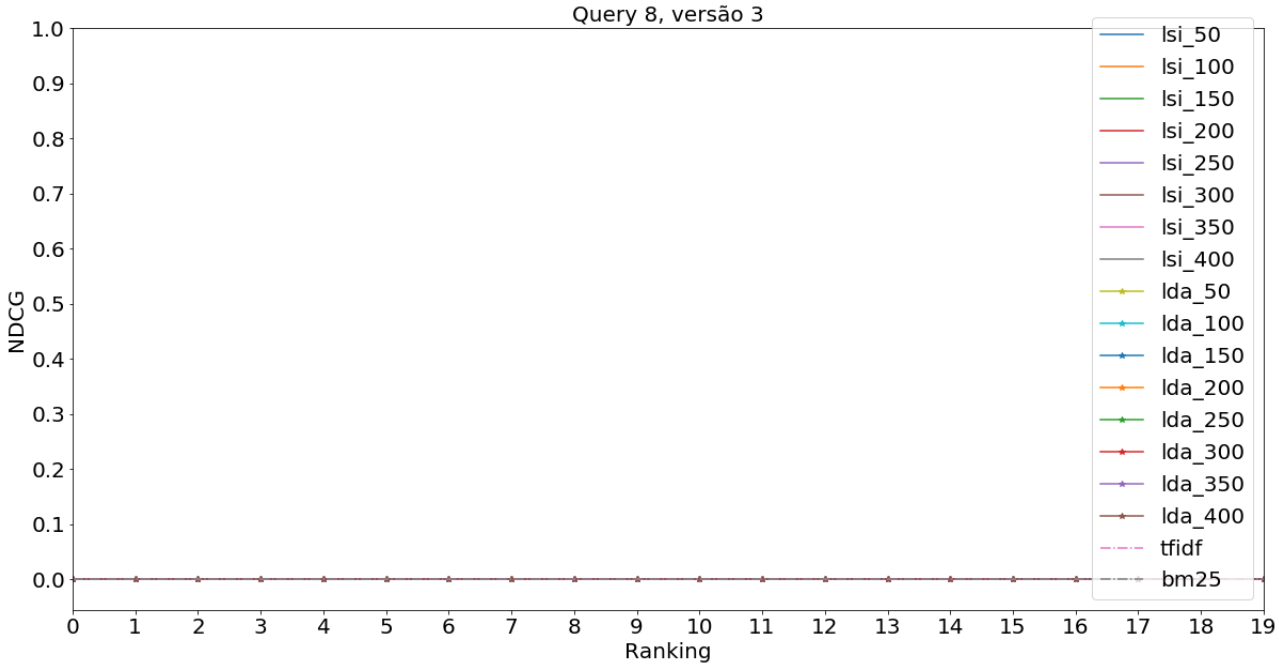
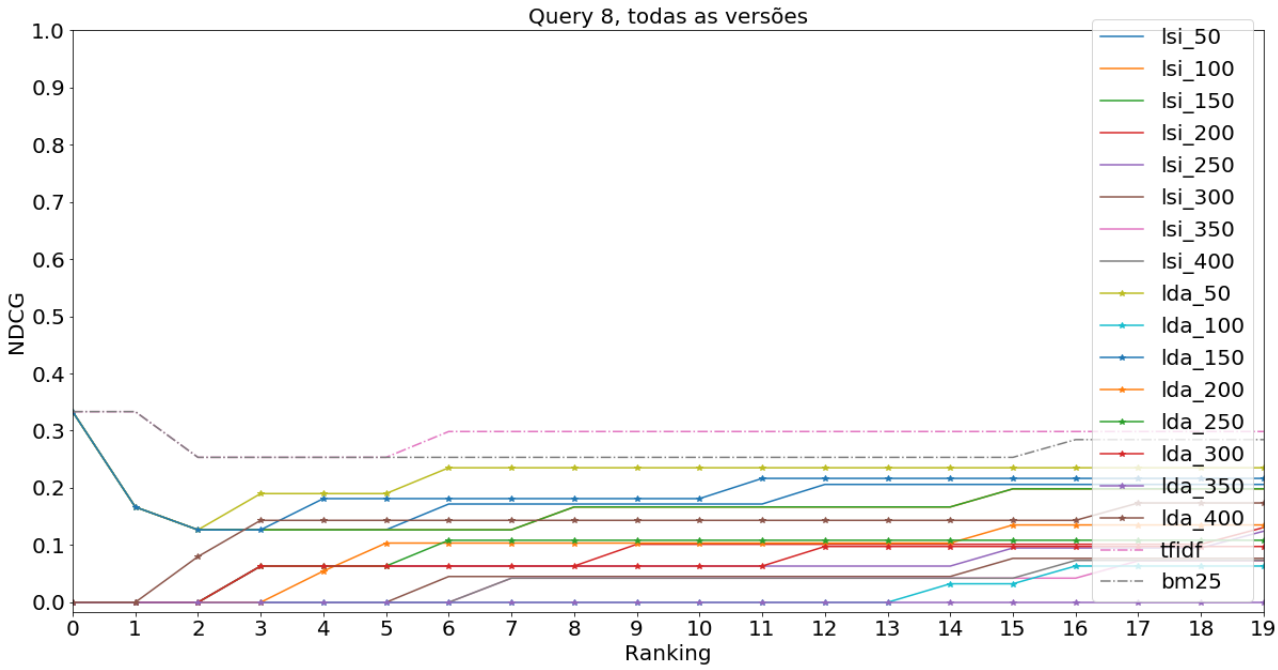


Figura A.36: PC - performance da *Query* 8, todas as versões



A.1.10 *Query 9*

A *Query 9* pretende localizar pareceres em casos em que o réu condenado por crime de violência doméstica descumpre as medidas protetivas a ele impostas. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "descumprimento de medidas protetivas por parte do agressor enseja em crime de desobediência"

- versão 2:

  > "agressor incorre em crime de desobediência ou prevaricação quando desrespeita as medidas protetivas"

- versão 3:

  > "réu desconsiderou as medidas protetivas impostas ensejando em crime de desobediência"

Quantidade de documentos relevantes: 1

Quantidade de documentos muito relevantes: 1

Figura A.37: PC - performance da *Query* 9, versão 1

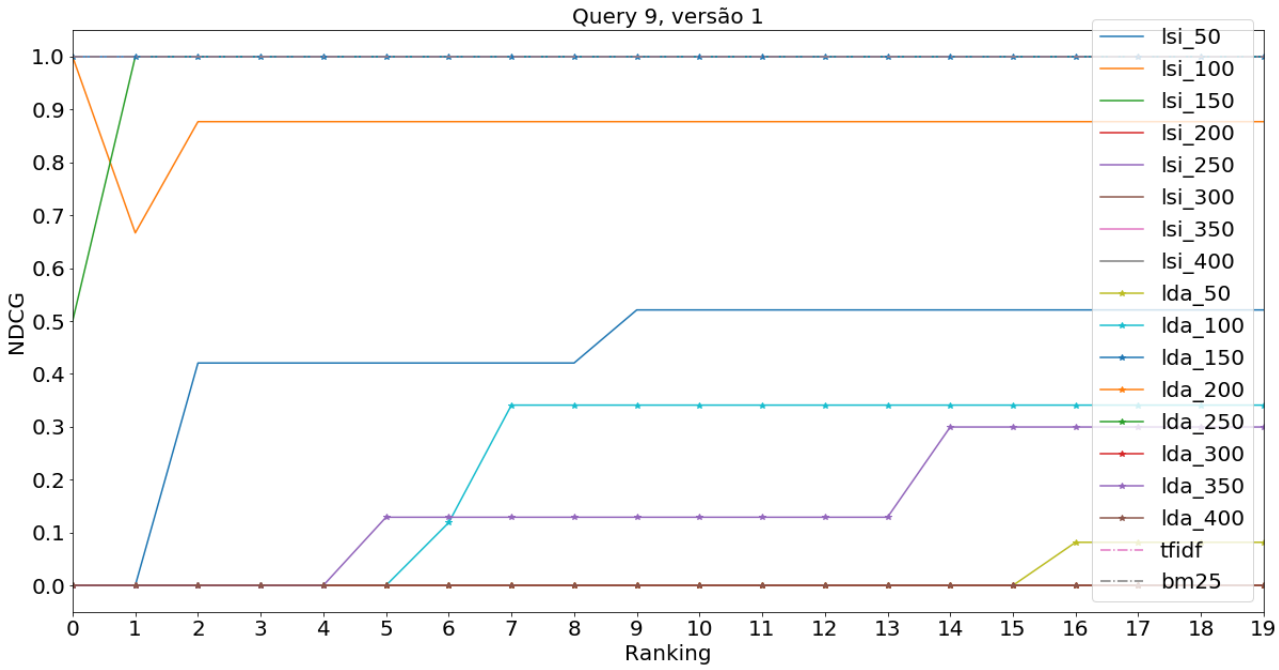


Figura A.38: PC - performance da *Query* 9, versão 2

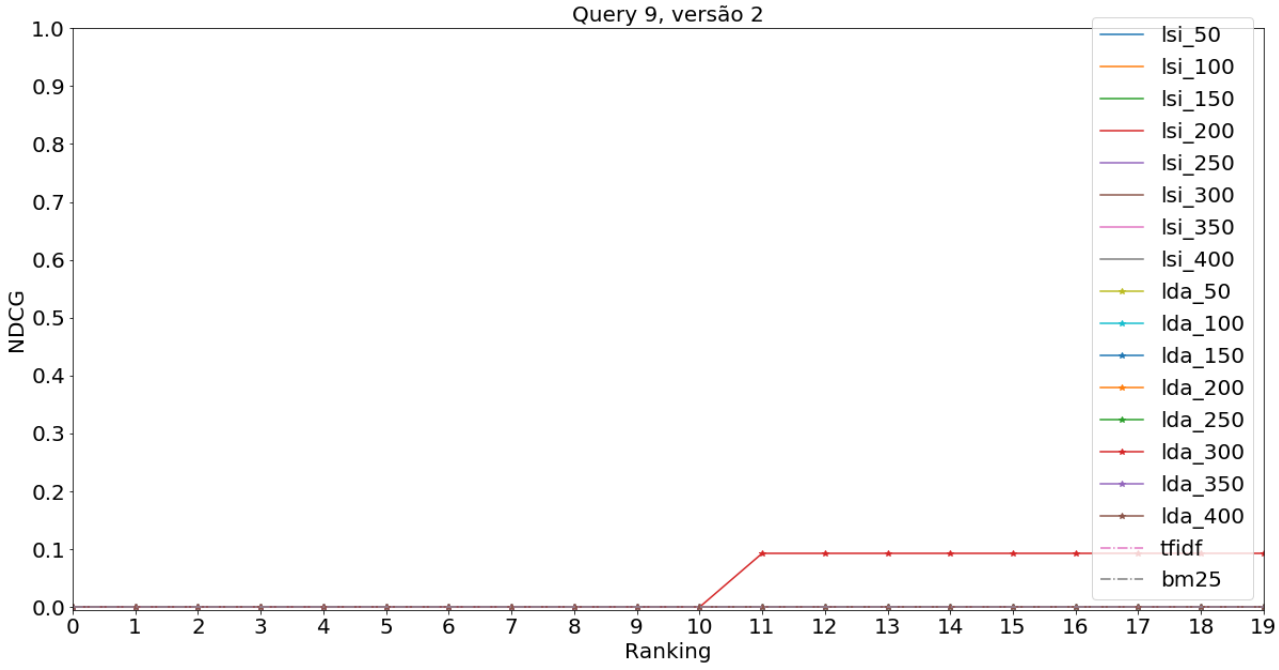


Figura A.39: PC - performance da *Query* 9, versão 3

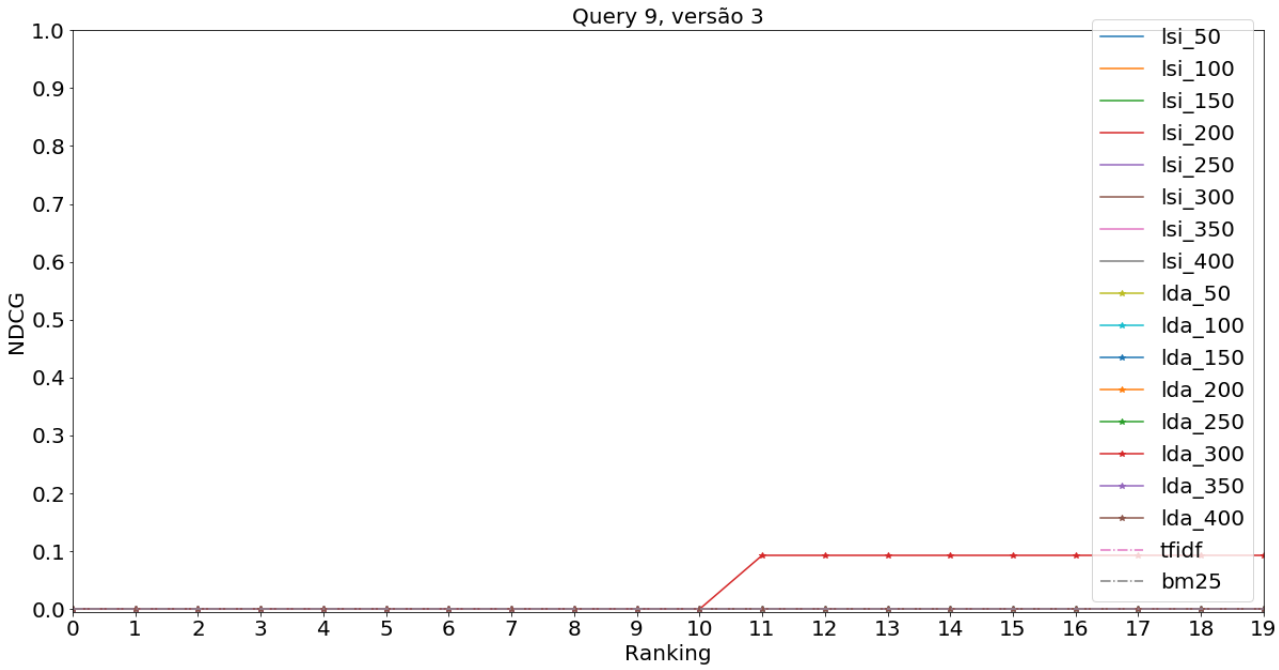
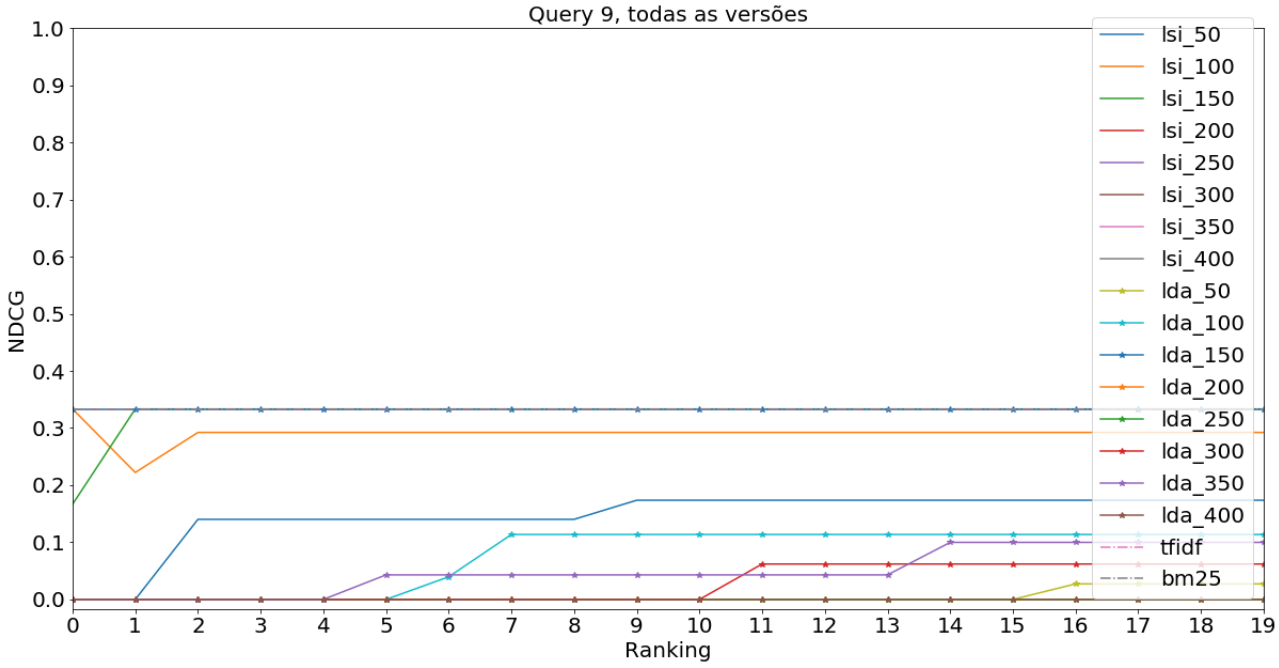


Figura A.40: PC - performance da *Query* 9, todas as versões



A.2 Comparação dos modelos sem enriquecimento

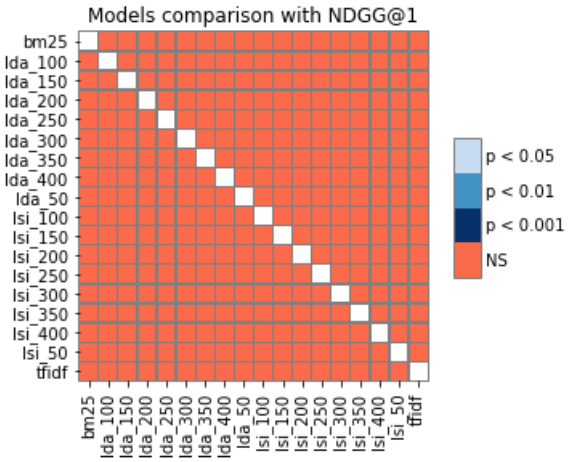


Figura A.41: Comparação entre todos os modelos na posição 1 do ranking

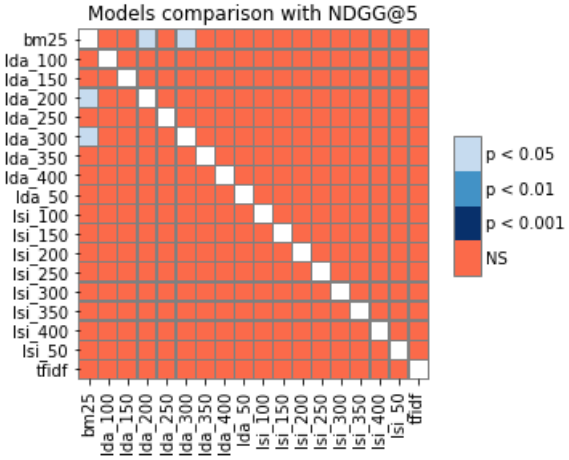


Figura A.42: Comparação entre todos os modelos na posição 5 do ranking

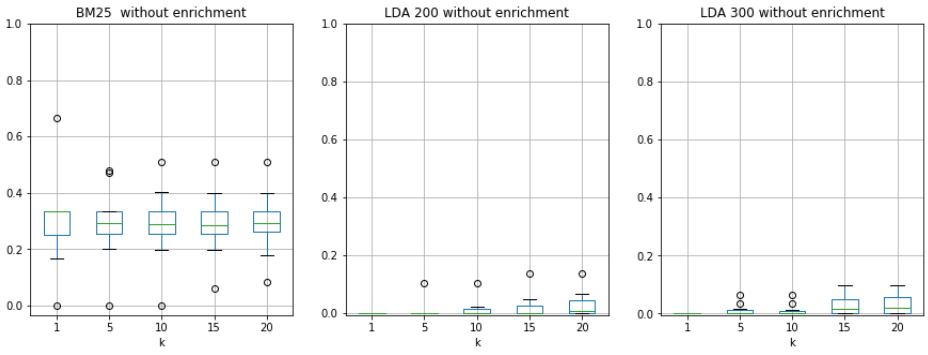


Figura A.43: Diferenças entre os modelos na posição 5 do ranking

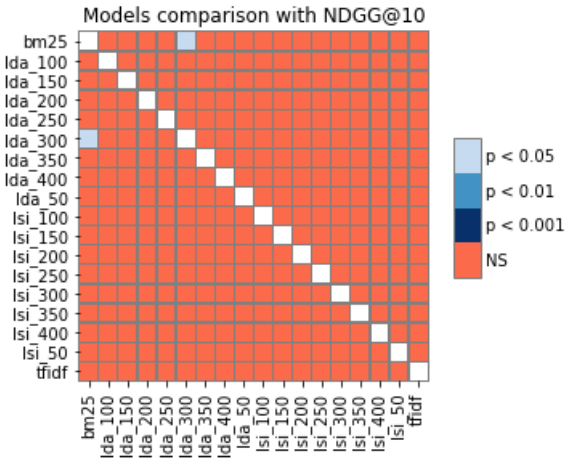


Figura A.44: Comparação entre todos os modelos na posição 10 do ranking

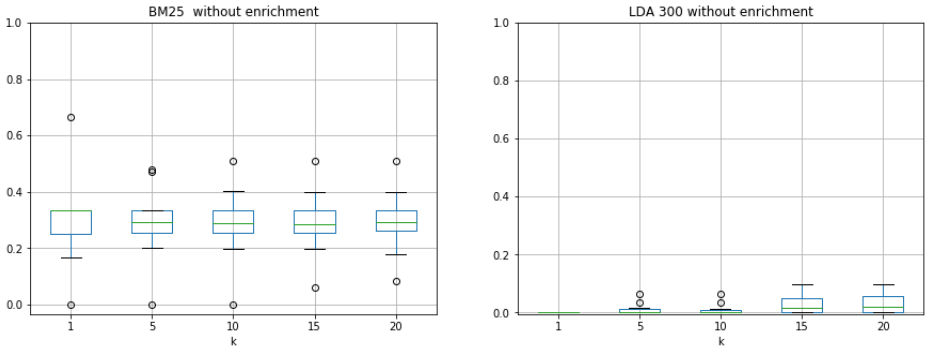


Figura A.45: Diferenças entre os modelos na posição 10 do ranking

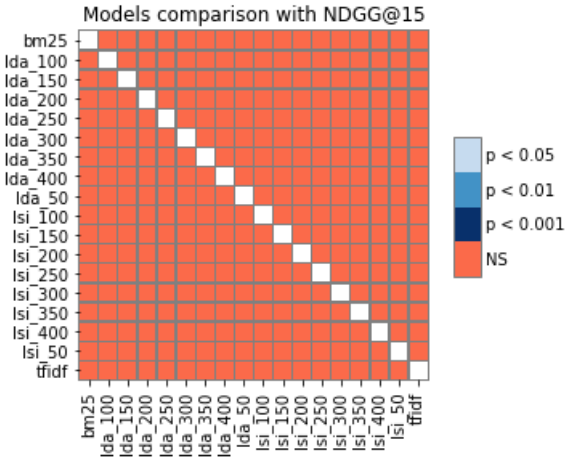


Figura A.46: Comparação entre todos os modelos na posição 15 do ranking

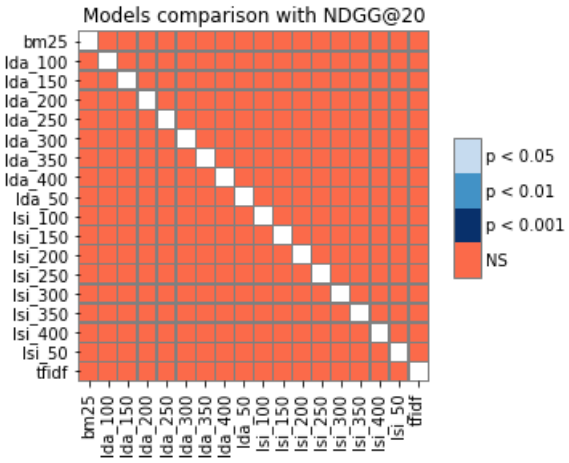


Figura A.47: Comparação entre todos os modelos na posição 20 do ranking

A.3 Comparação dos modelos com enriquecimento

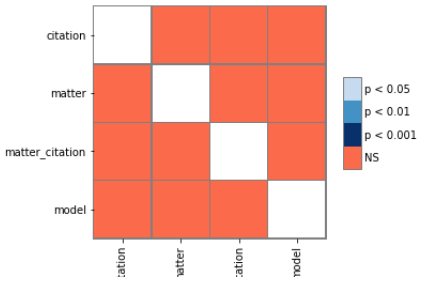


Figura A.48: Comparação de modelos TF-IDF com enriquecimento

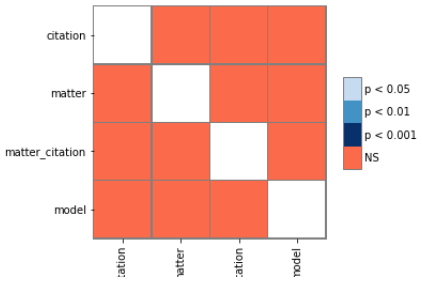


Figura A.49: Comparação de modelos BM25 com enriquecimento

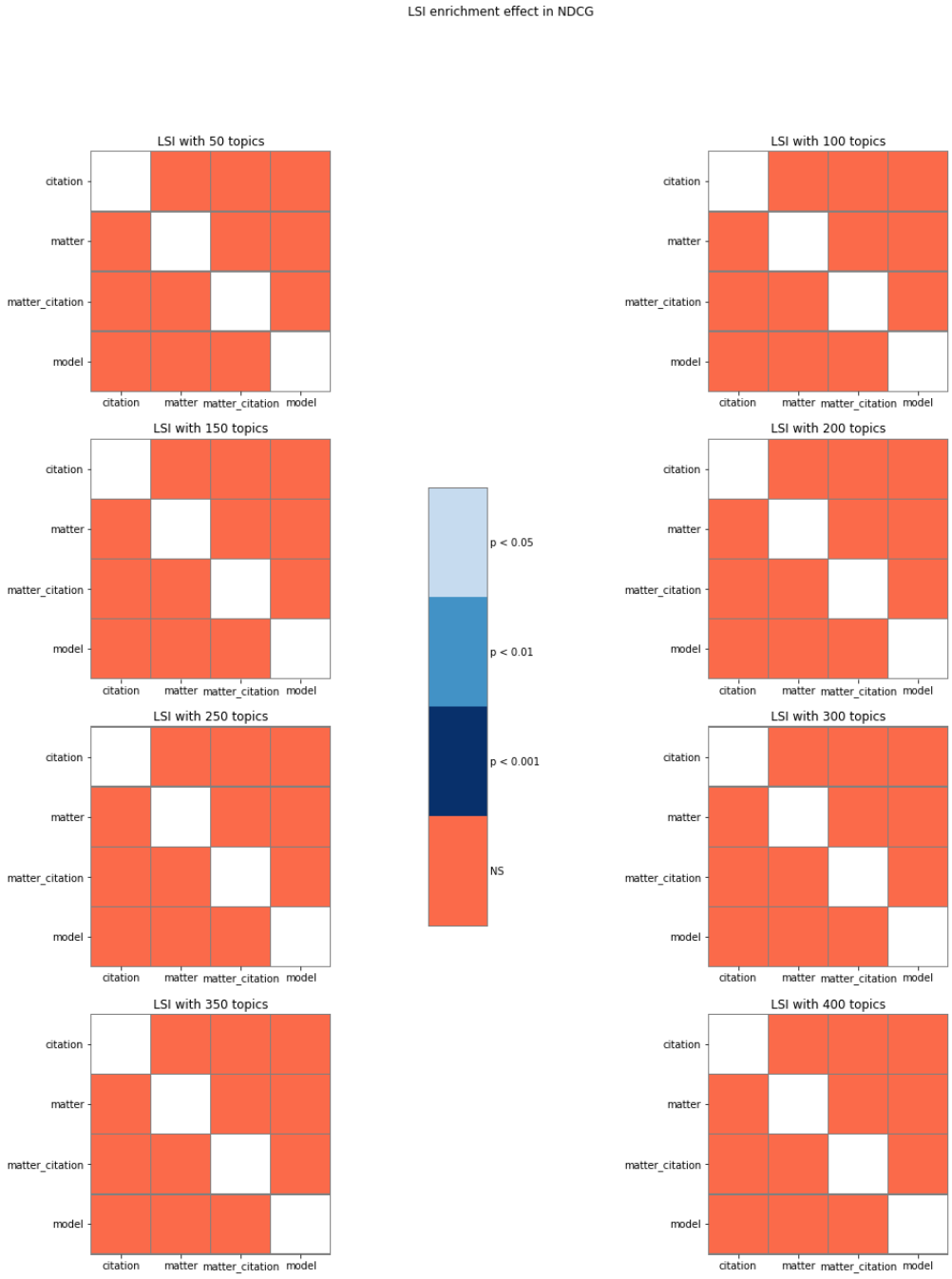


Figura A.50: Comparação de modelos LSI com enriquecimento

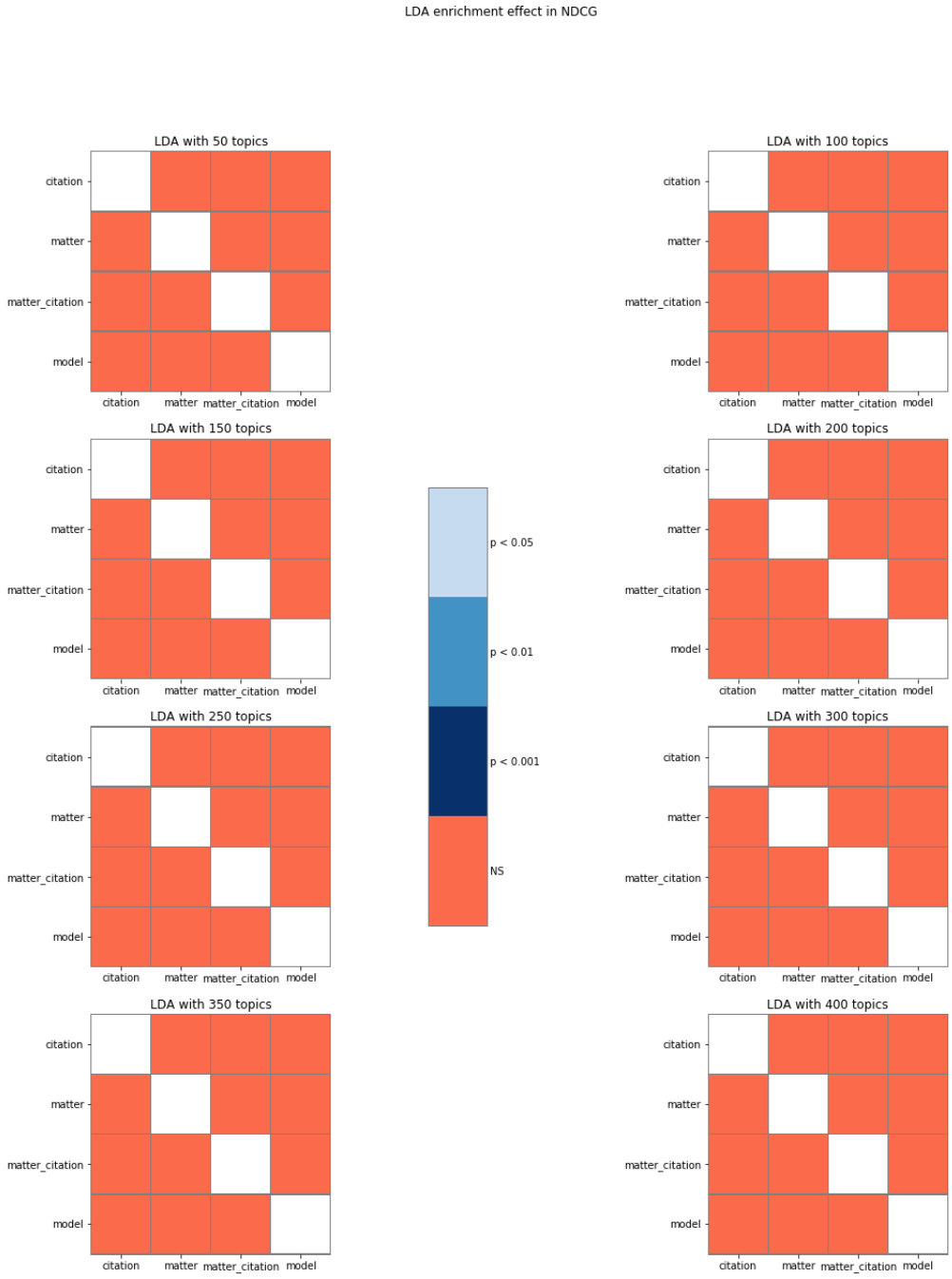


Figura A.51: Comparação de modelos LDA com enriquecimento

Apêndice B

Performance dos modelos: Procuradorias Criminais Especializadas

B.1 Avaliação dos modelos sem enriquecimento por query

B.1.1 *Query 0*

A *Query 0* pretende localizar pareceres sobre casos em que o réu condenado por crime de trânsito alega que o acidente de trânsito em que era condutor do veículo envolvido ocorreu por culpa da vítima, e que deve haver absolvição ou redução das reprimendas. Tipicamente, o processo reúne perícia ou testemunho de agentes públicos relatando a imprudência do condutor. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa: Texto de pesquisa:

- versão 1:

  > "réu alega culpa da vítima porém conduziu o veículo de modo imprudente e provocando o acidente"

- versão 2:

  > "réu dirigia de forma imprudente pede compensação afirmando que a vítima provocou o acidente"

- versão 3:

  > "inviável a compensação de pena alegando que a o acidente foi provocado pela vítima"

Quantidade de documentos relevantes: 0

Quantidade de documentos muito relevantes: 10

Figura B.1: PCE - performance da *Query* 0, versão 1

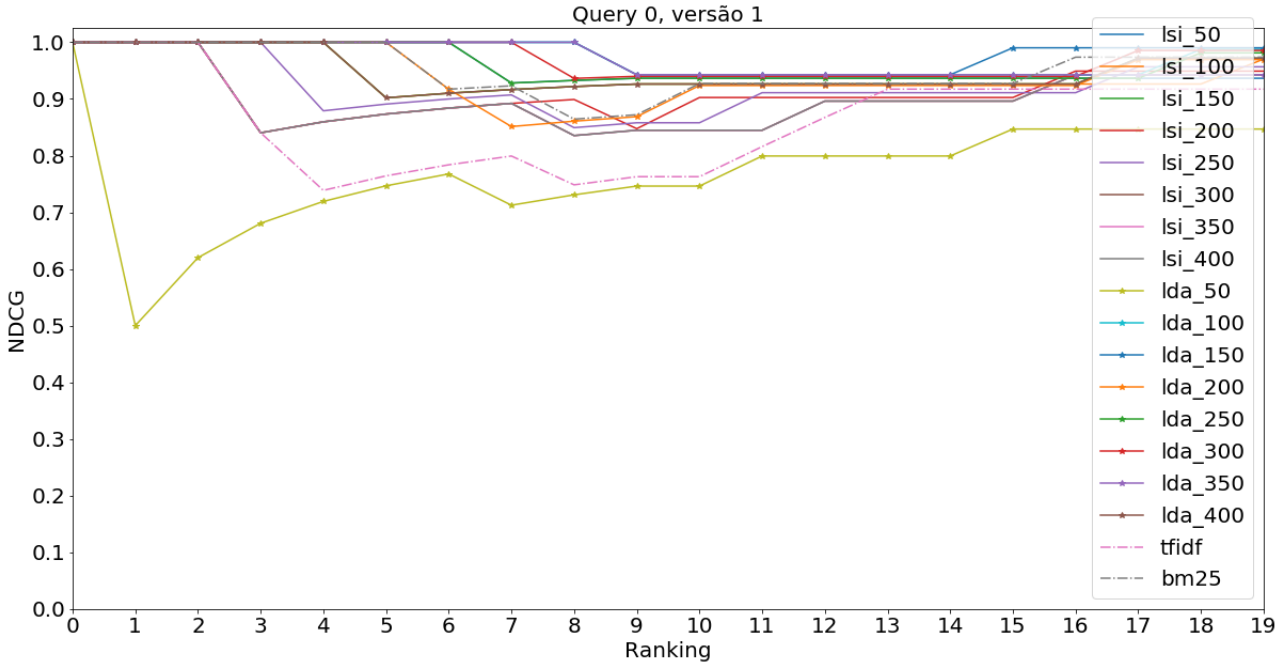


Figura B.2: PCE - performance da *Query* 0, versão 2

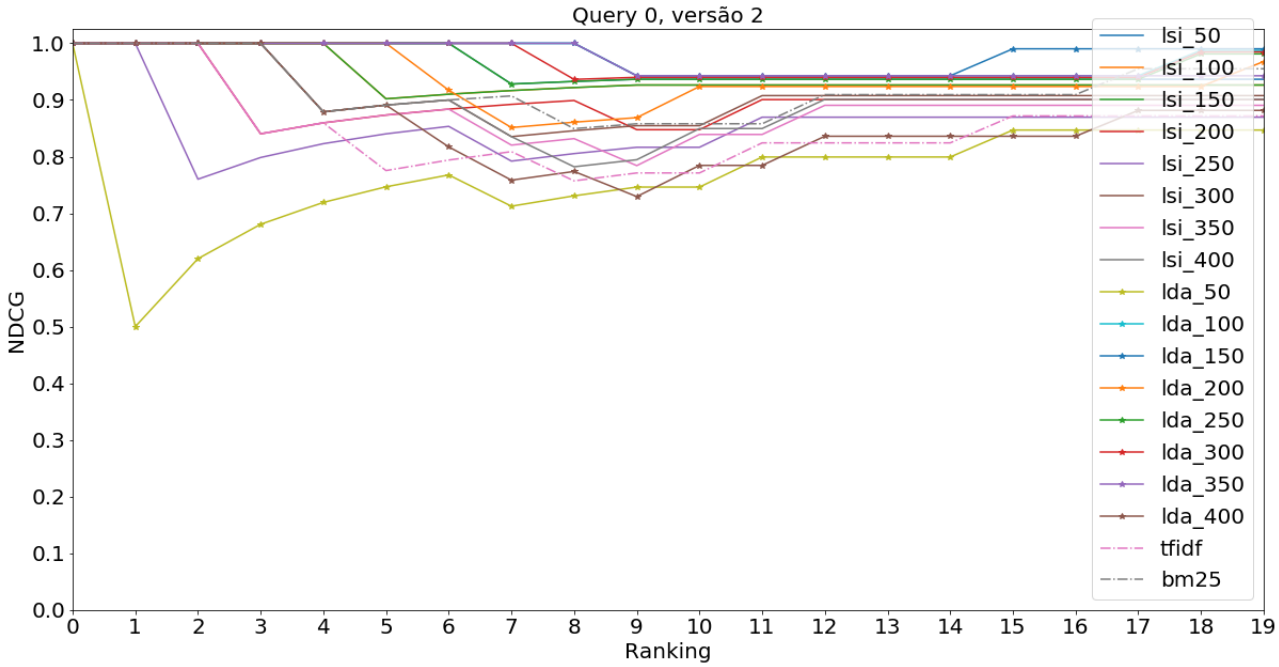


Figura B.3: PCE - performance da *Query* 0, versão 3

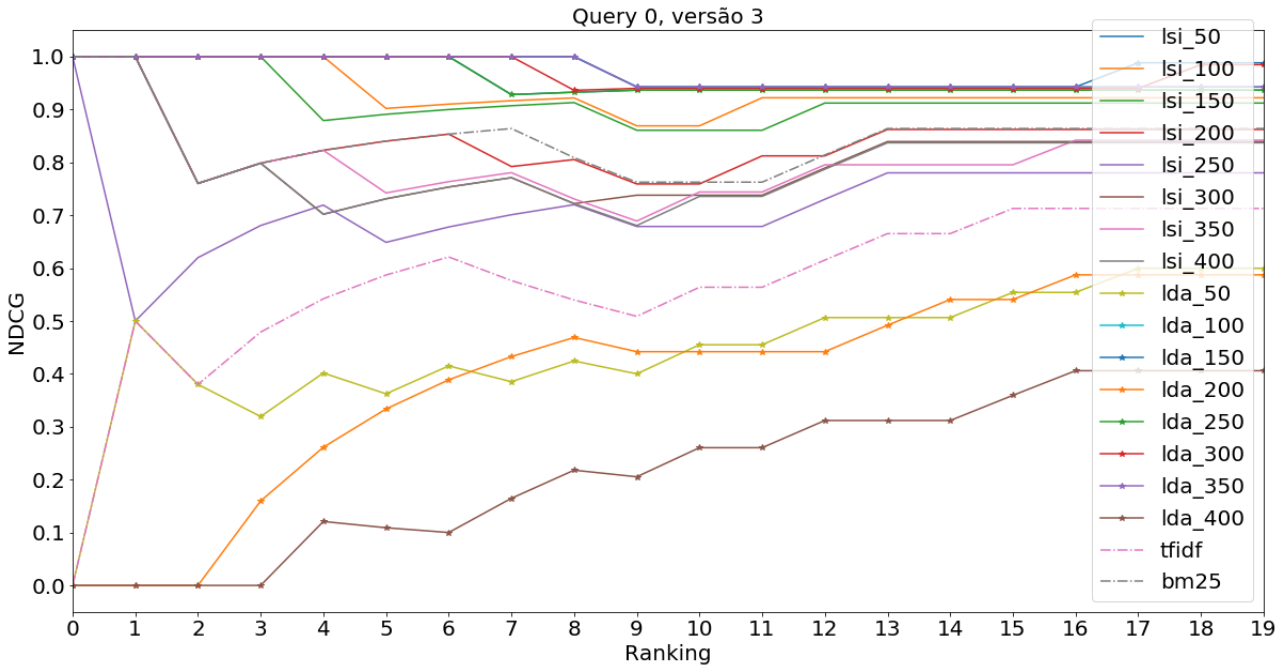
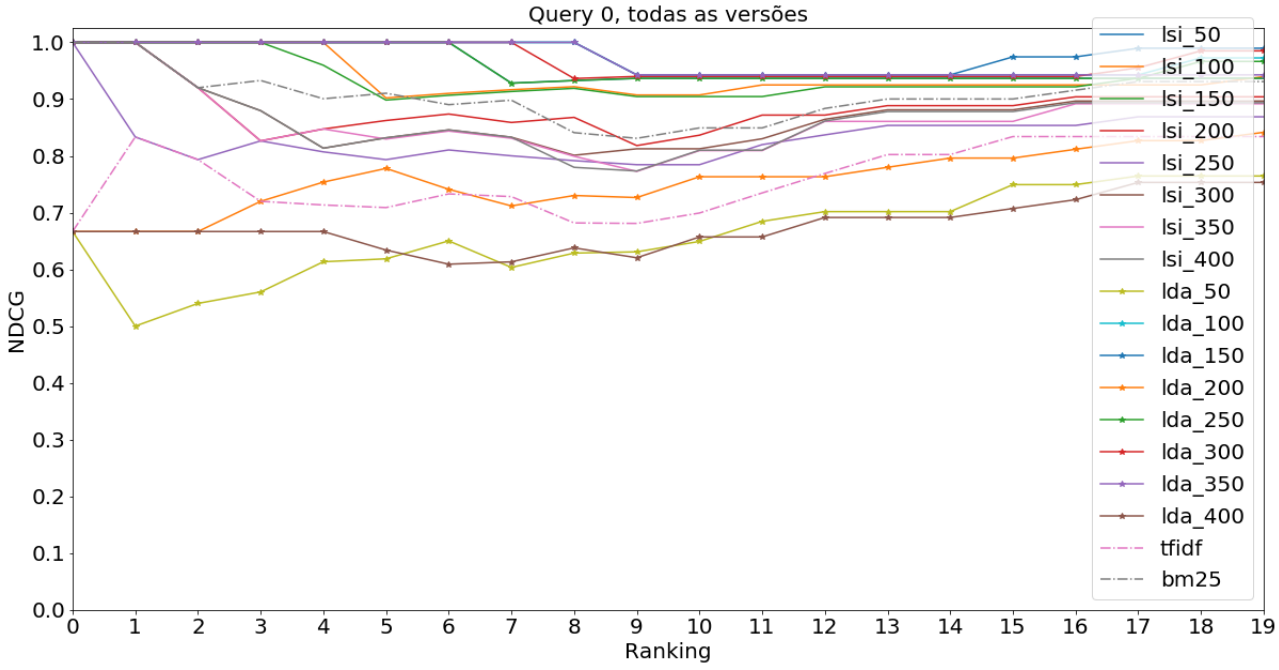


Figura B.4: PCE - performance da *Query* 0, todas as versões



B.1.2 *Query 1*

A *Query 1* pretende localizar pareceres sobre casos em que o paciente em *Habeas Corpus* alega que a prisão preventiva é desnecessária em função de que este tem bons antecedentes criminais e não há risco com relação à sua liberdade. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "Paciente alega identidade civil comprovada estar empregado e bons antecedentes porém é acusado de crime de roubo qualificado representando risco a ordem pública"

- versão 2:

  > "Emprego e residência fixos e família não são suficientes para afastar a prisão preventiva considerando o crime de roubo qualificado"

- versão 3:

  > "Liberdade do paciente indiciado por crime de roubo qualificado representa risco à órdem pública mesmo que ostente condições pessoais favoráveis"

Quantidade de documentos relevantes: 37

Quantidade de documentos muito relevantes: 46

Figura B.5: PCE - performance da *Query* 1, versão 1

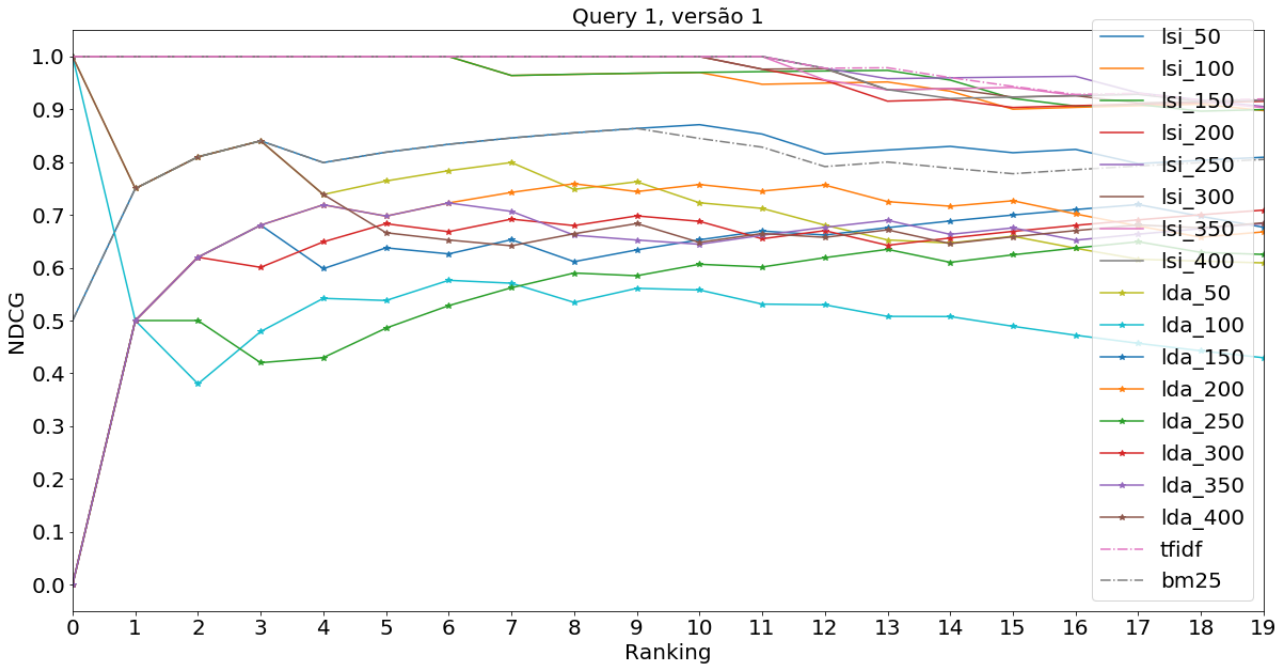


Figura B.6: PCE - performance da *Query* 1, versão 2

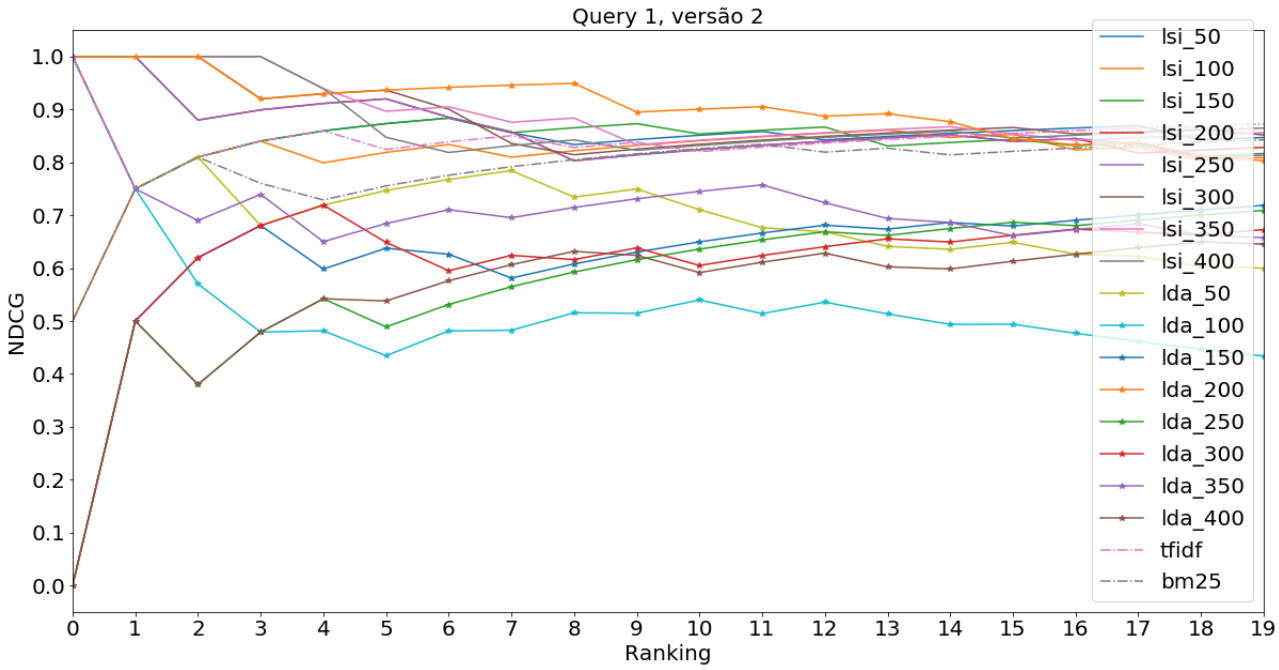


Figura B.7: PCE - performance da *Query* 1, versão 3

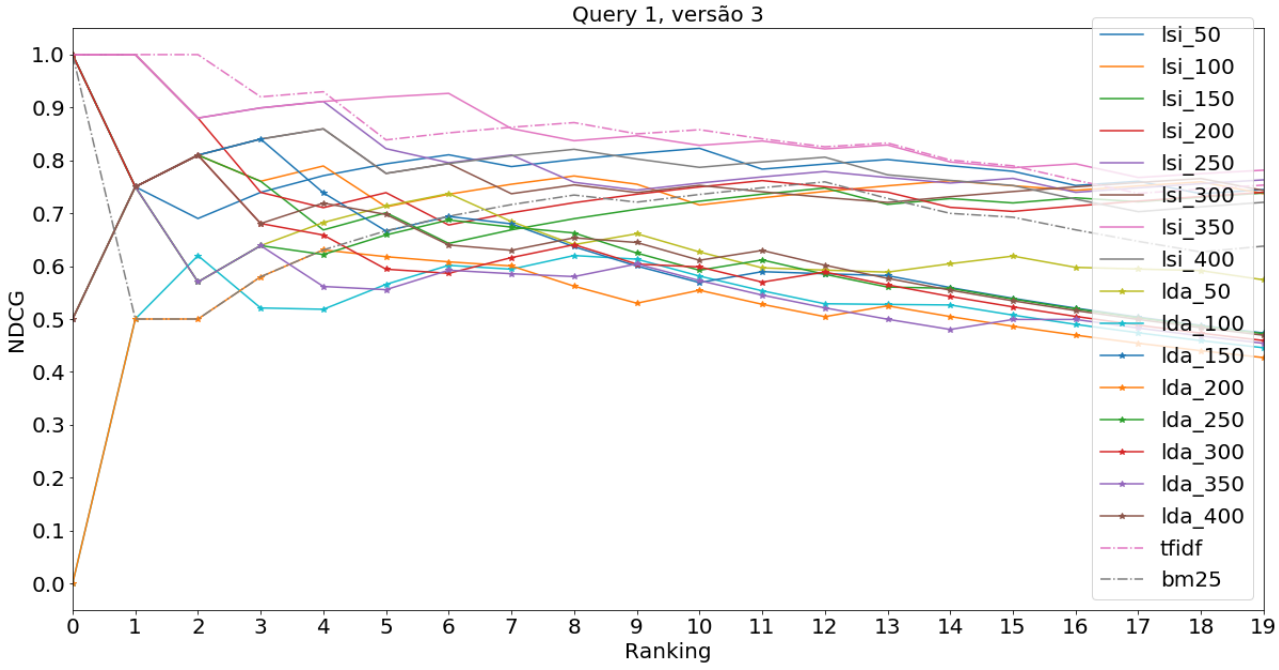
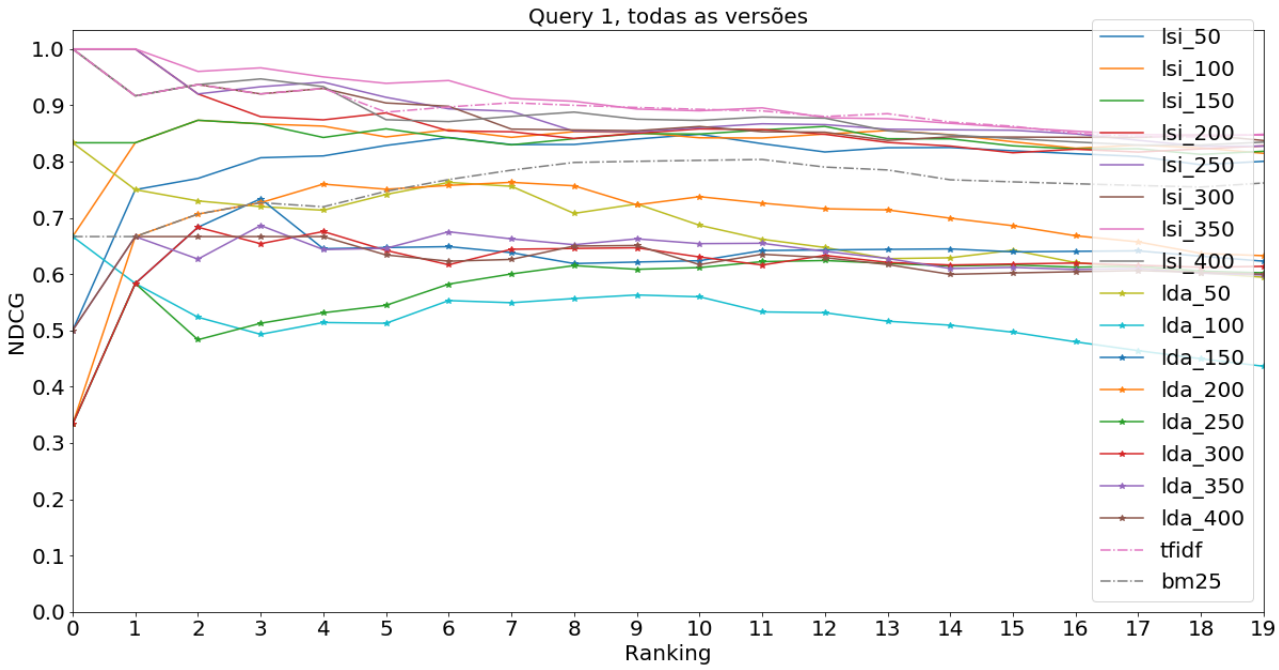


Figura B.8: PCE - performance da *Query* 1, todas as versões



B.1.3 *Query 2*

A *Query 2* pretende localizar pareceres sobre casos em que o paciente em *Habeas Corpus* alega que a prisão preventiva desconsidera a presunção de sua inocência, hava vista que o caso ainda não foi julgado. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  "Paciente alega que a medida fere presunção de sua inocência"

- versão 2:

  "Comprovados o periculum libertatis e o fumus commissi delicti não há que se falar em violação da presunção de inocência"

- versão 3:

  "Segregação cautelar em nada ofende o princípio da presunção da inocência havendo forte indícios de materialidade e autoria "

Quantidade de documentos relevantes: 17

Quantidade de documentos muito relevantes: 9

Figura B.9: PCE - performance da *Query* 2, versão 1

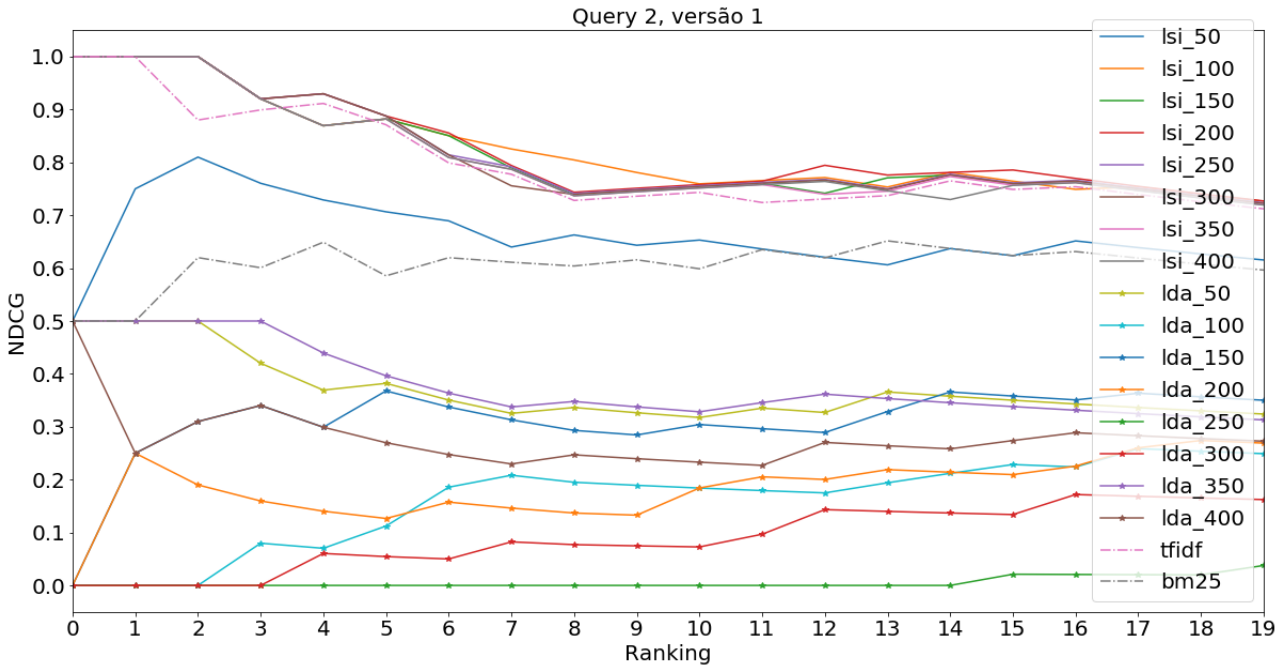


Figura B.10: PCE - performance da *Query* 2, versão 2

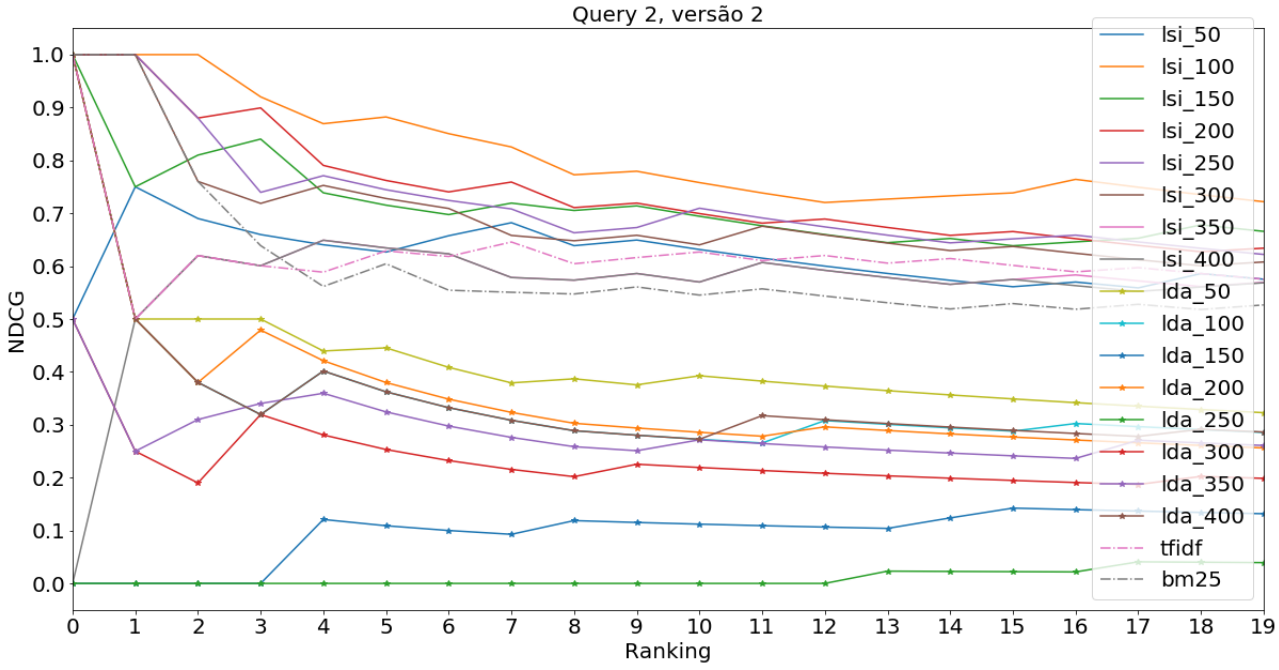


Figura B.11: PCE - performance da *Query* 2, versão 3

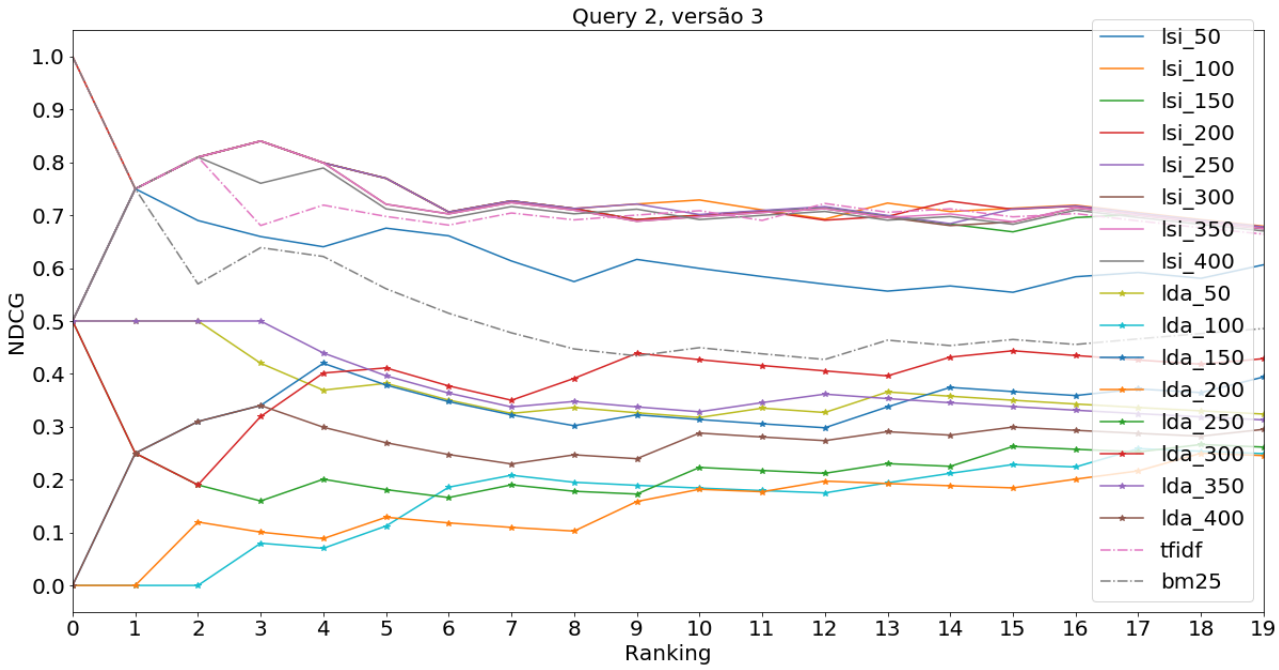
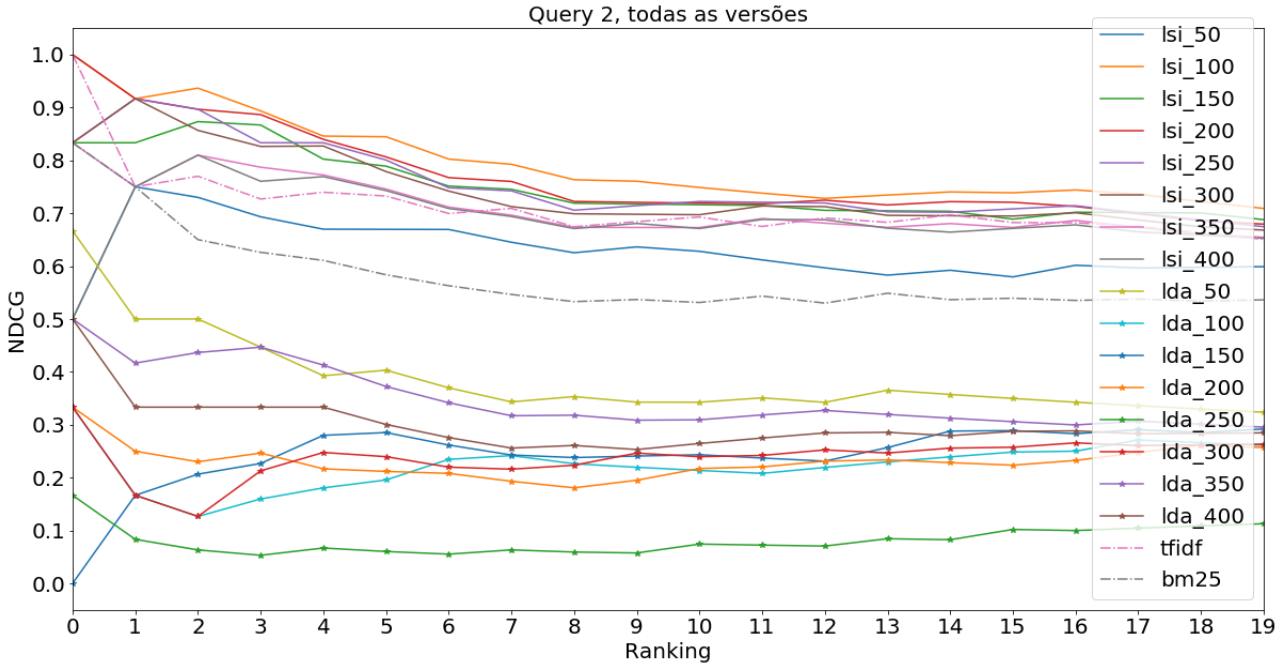


Figura B.12: PCE - performance da *Query* 2, todas as versões



B.1.4 *Query 3*

A *Query 3* pretende localizar pareceres sobre casos em que o paciente em *Habeas Corpus* acusado de homicídio (ou tentativa) alega que a prisão preventiva é desnecessária em função de que não há risco com relação à sua liberdade, mas o paciente evadiu do local do fato. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "homicídio tentado e evasão do autor demostra a presença do periculum libertatis"

- versão 2:

  > "Paciente representa risco à órdem pública por ser acusado de tentativa de homicídio e ainda ter fugido do local do fato"

- versão 3:

  > "necessária a prisão cautelar considerando a gravidade do delito de homicídio com posterior evasão do acusado"

Quantidade de documentos relevantes: 5

Quantidade de documentos muito relevantes: 4

Figura B.13: PCE - performance da *Query* 3, versão 1

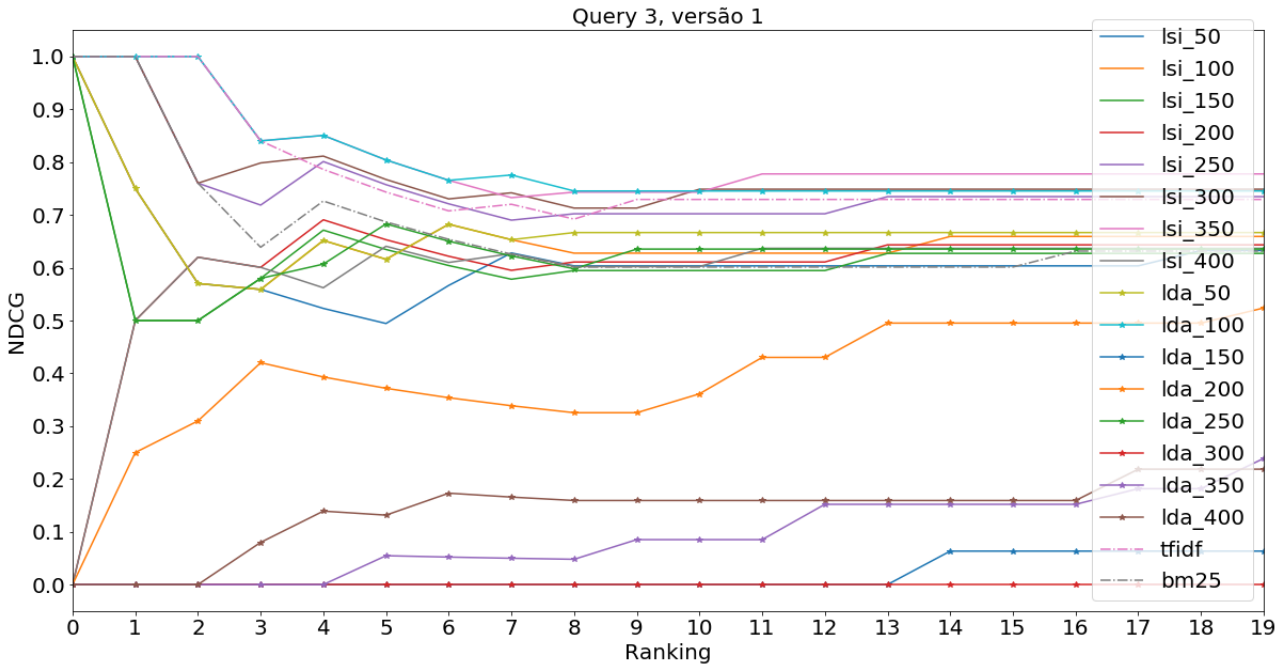


Figura B.14: PCE - performance da *Query* 3, versão 2

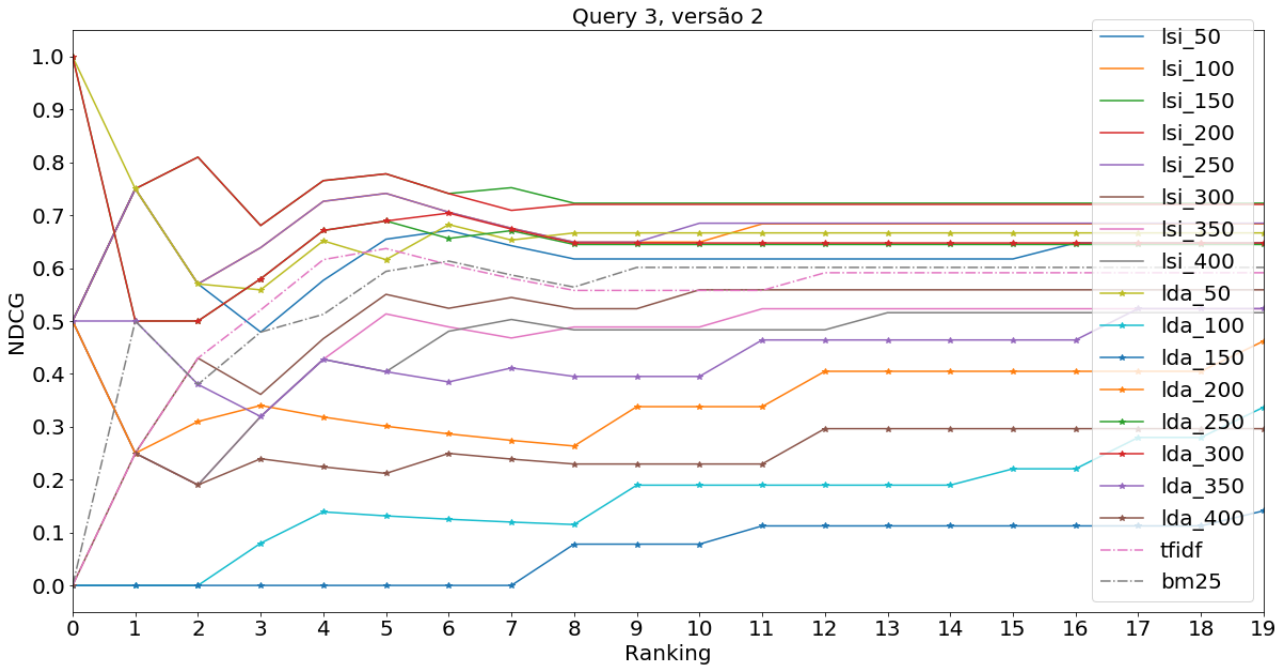


Figura B.15: PCE - performance da *Query* 3, versão 3

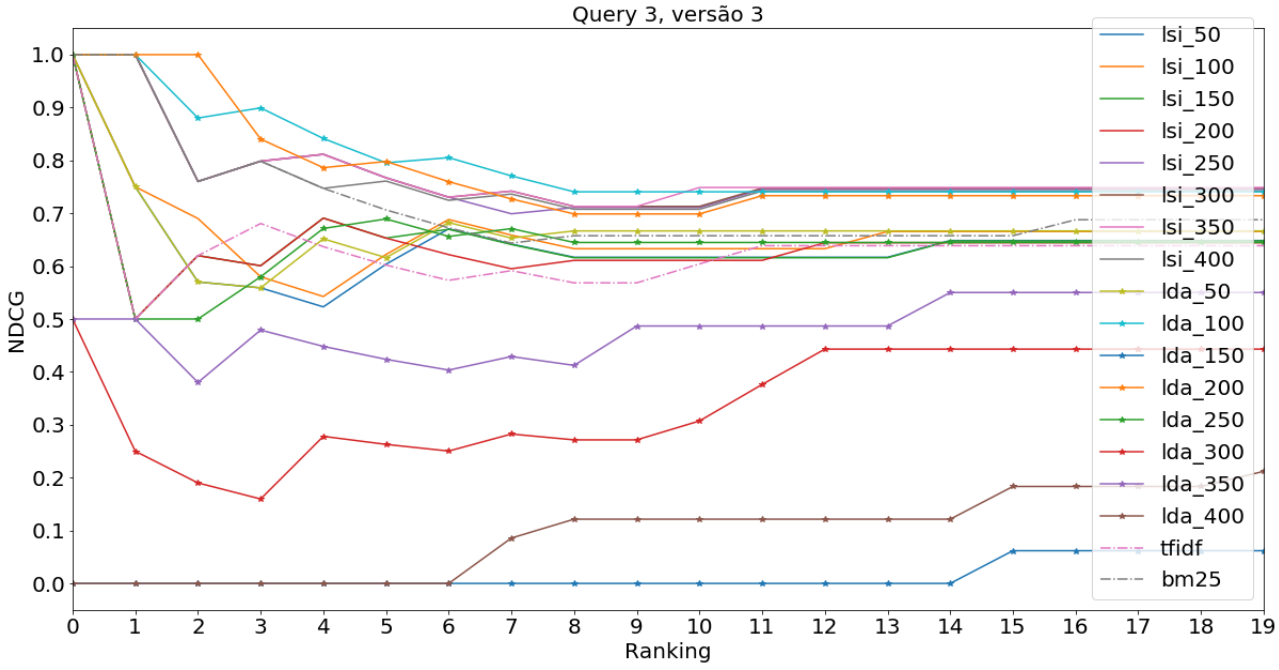
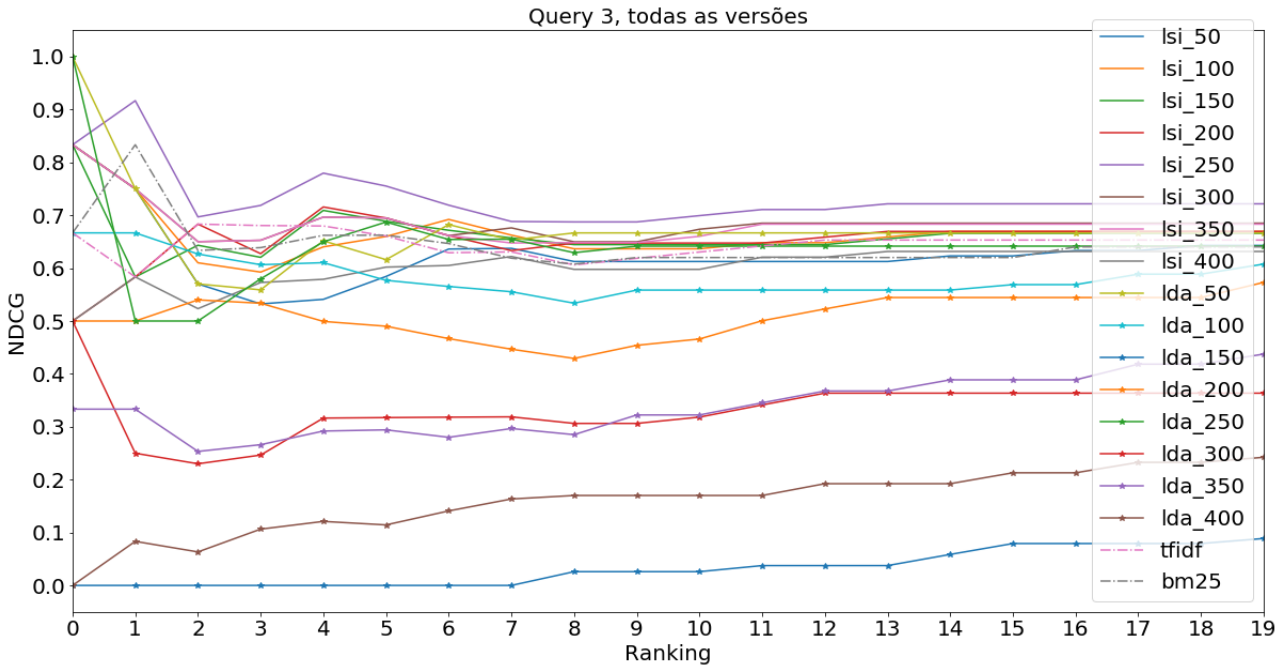


Figura B.16: PCE - performance da *Query* 3, todas as versões



B.1.5 *Query 4*

A *Query 4* pretende localizar pareceres sobre casos em que o réu condenado deseja afastar a majorante de corrupção de menores alegando que desconhecia a menoridade do colaborador no crime, e que não foram apresentados documentos que provasse esta condição. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "menoridade pode ser comprovada por outros documentos que não o de identidade"

- versão 2:

  > "réu alega que não sabia a idade do menor que concorreu para o crime"

- versão 3:

  > "ônus de provar desconhecimento da condição de menoridade é da defesa"

Quantidade de documentos relevantes: 0
Quantidade de documentos muito relevantes: 2

Figura B.17: PCE - performance da *Query* 4, versão 1

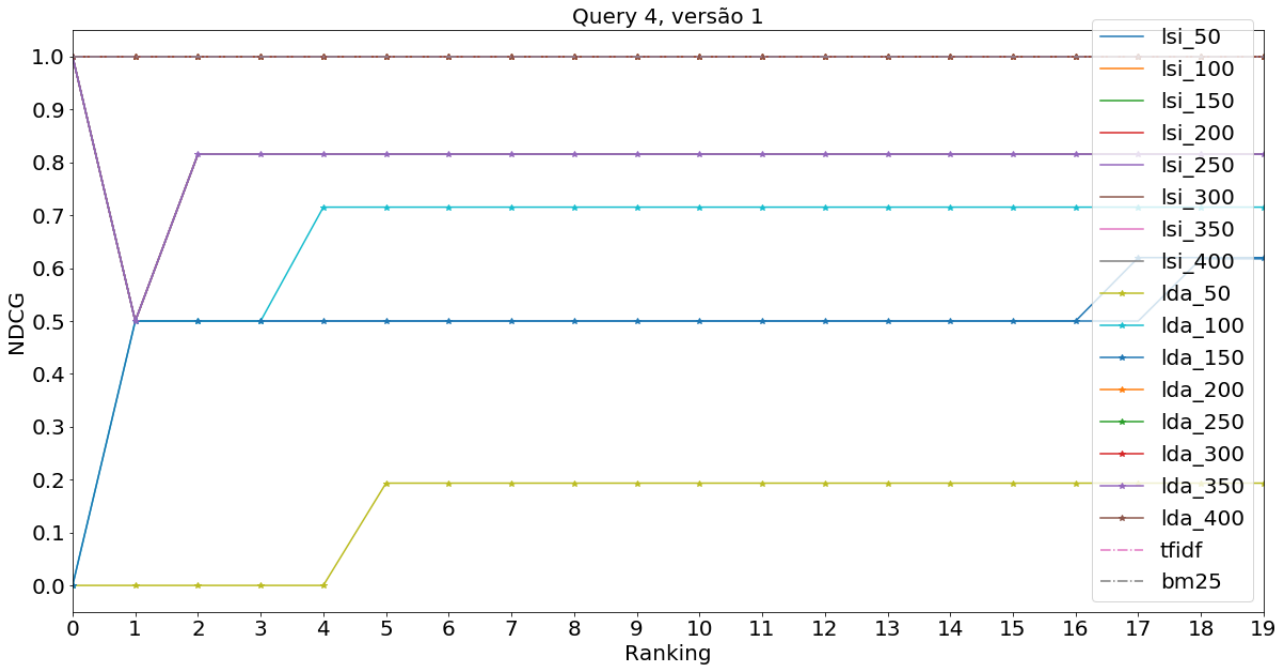


Figura B.18: PCE - performance da *Query* 4, versão 2

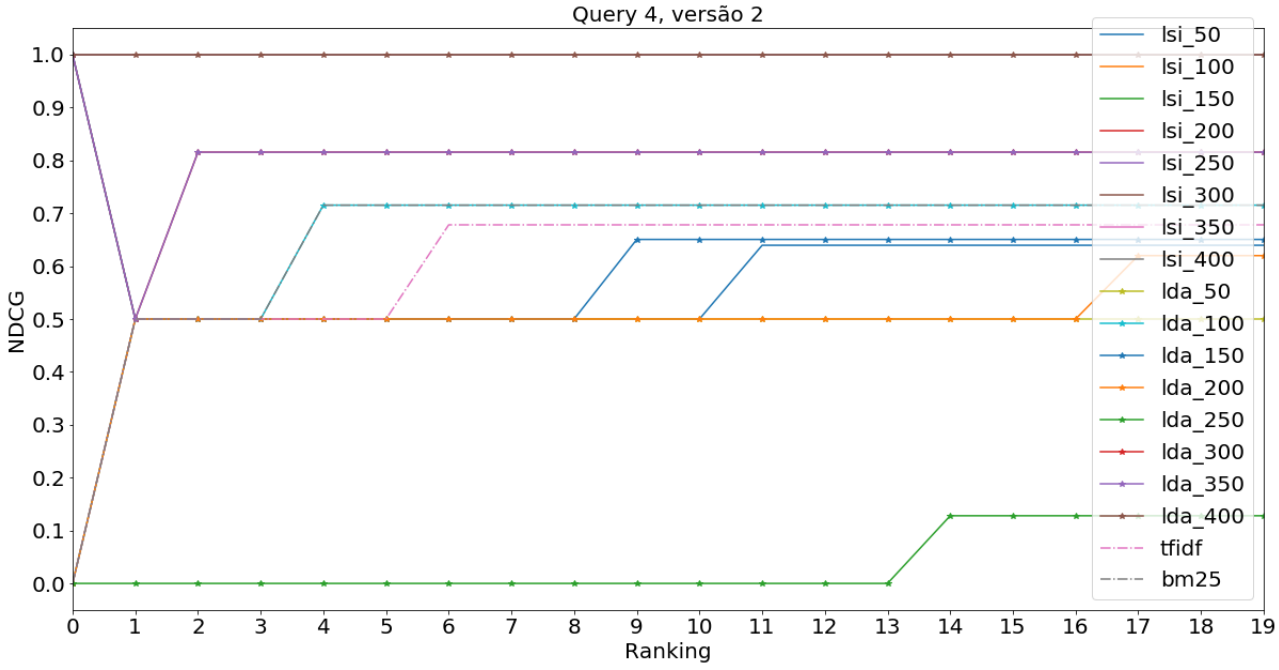


Figura B.19: PCE - performance da *Query* 4, versão 3

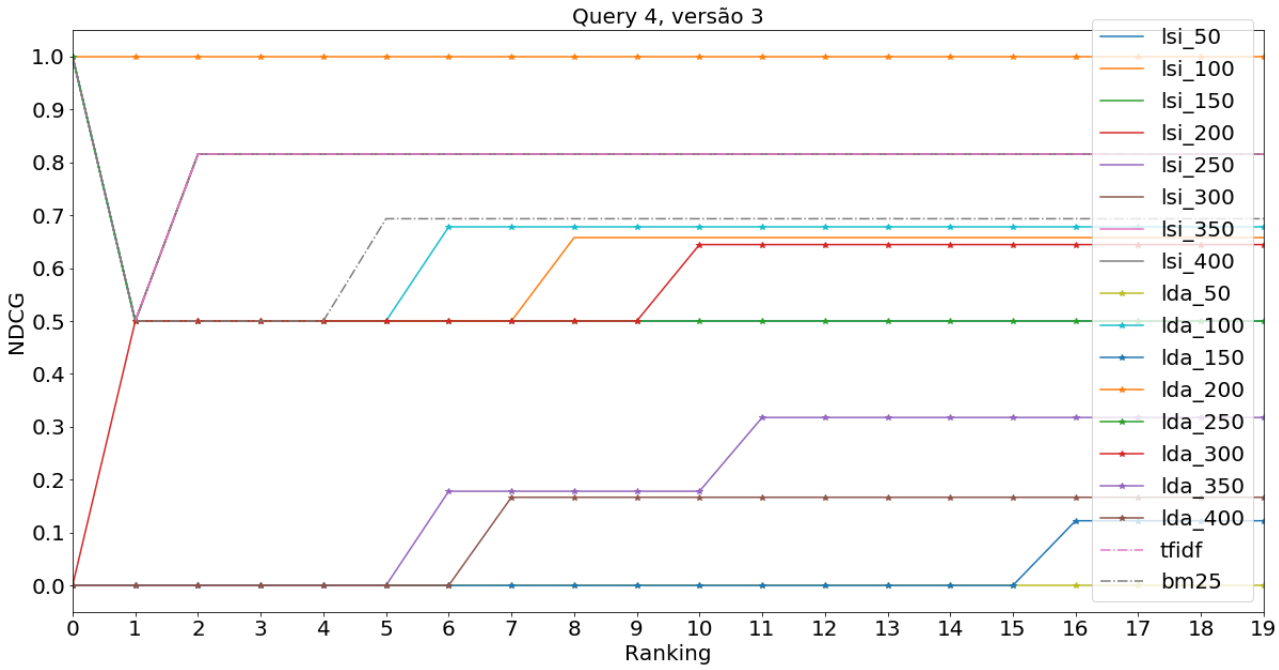
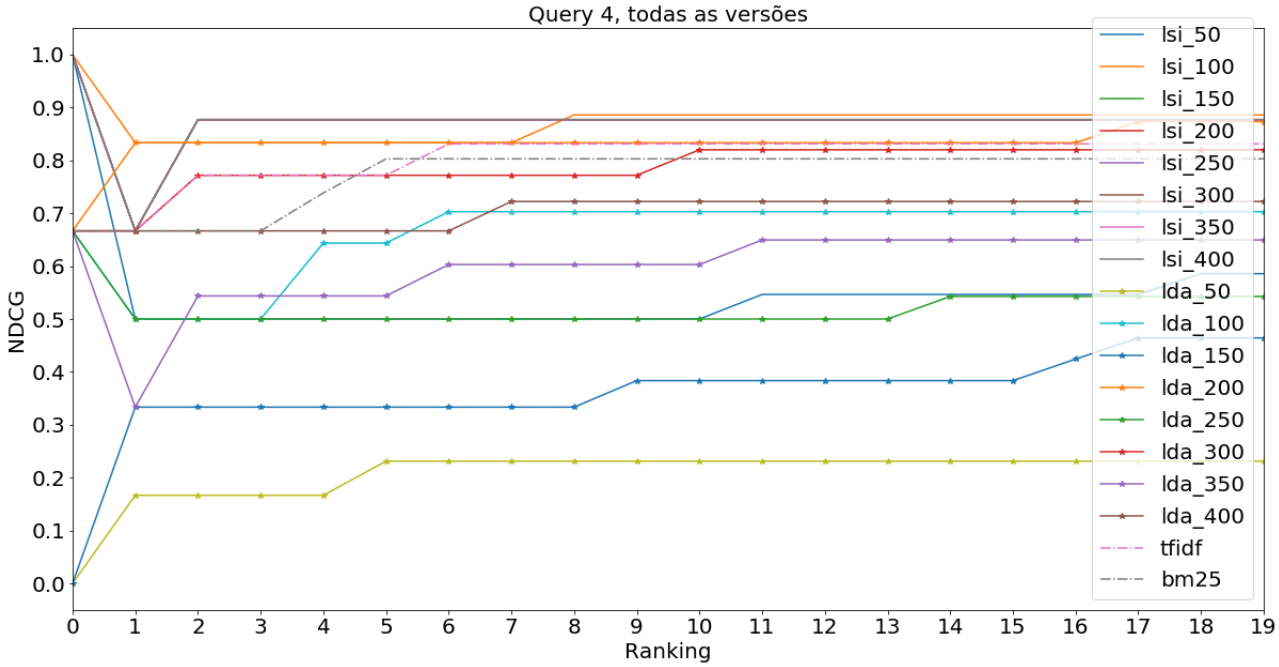


Figura B.20: PCE - performance da *Query* 4, todas as versões



B.1.6 *Query 5*

A *Query 5* pretende localizar pareceres sobre casos em que o paciente em *Habeas Corpus* alega que a decorreu muito tempo desde sua prisão preventiva. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "paciente alega coação ilegal por excesso de prazo para conclusão da instrução criminal"

- versão 2:

  > "coação e contrangimento ilegal por demora demasiada da prisão cautelar"

- versão 3:

  > "delonga injustificável da segregação cautelar"

Quantidade de documentos relevantes: 5
Quantidade de documentos muito relevantes: 17

Figura B.21: PCE - performance da *Query* 5, versão 1

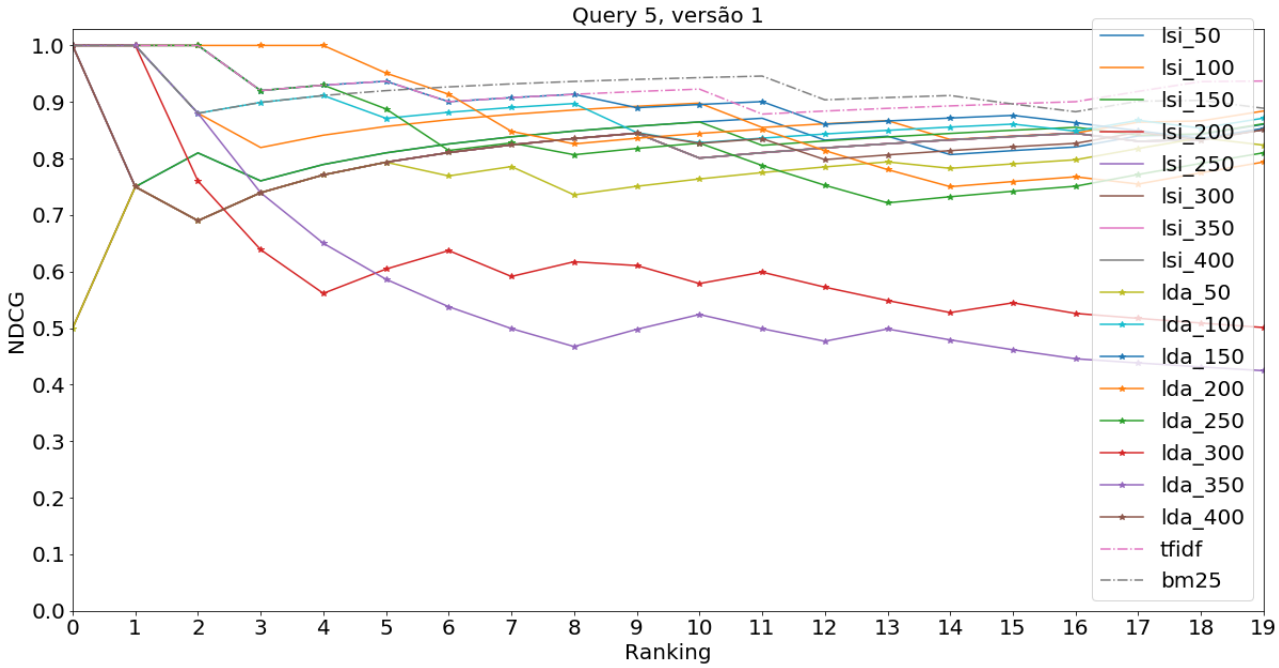


Figura B.22: PCE - performance da *Query* 5, versão 2

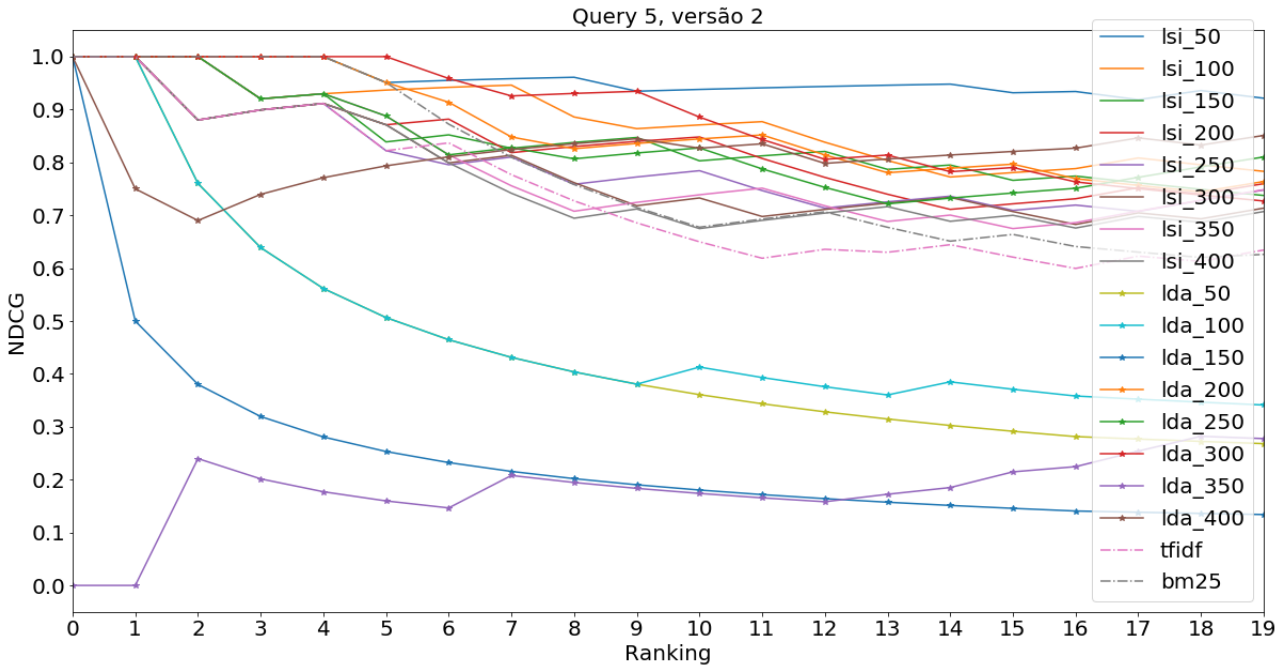


Figura B.23: PCE - performance da *Query* 5, versão 3

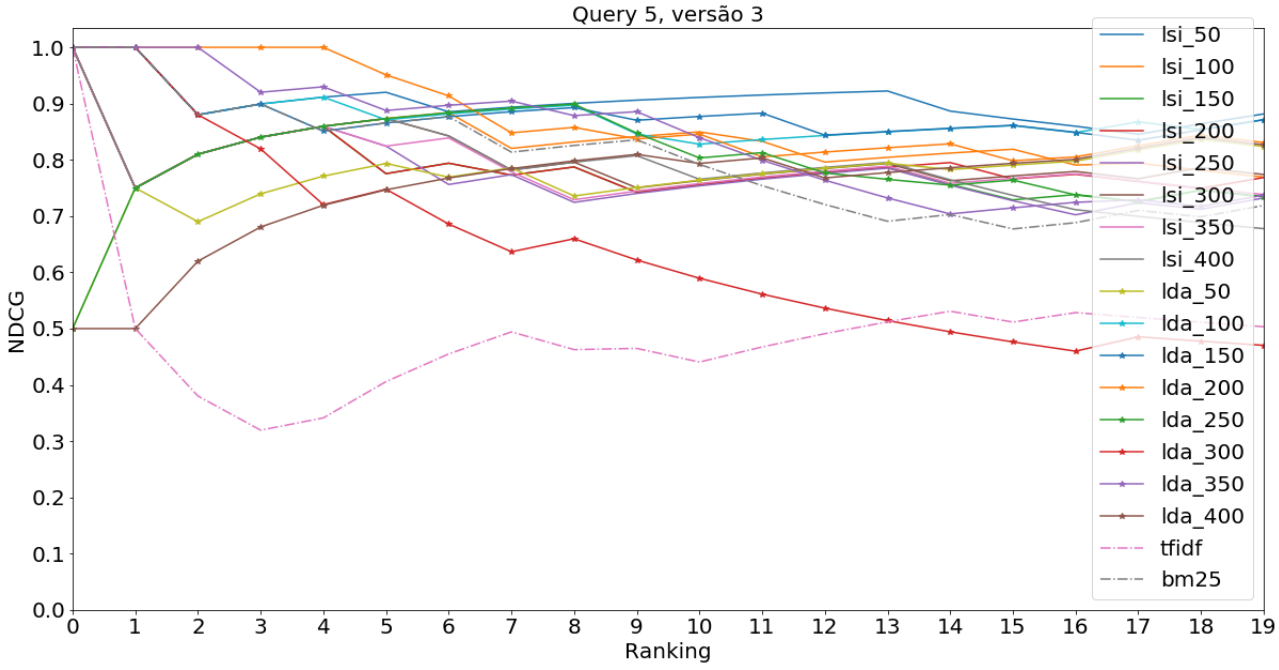
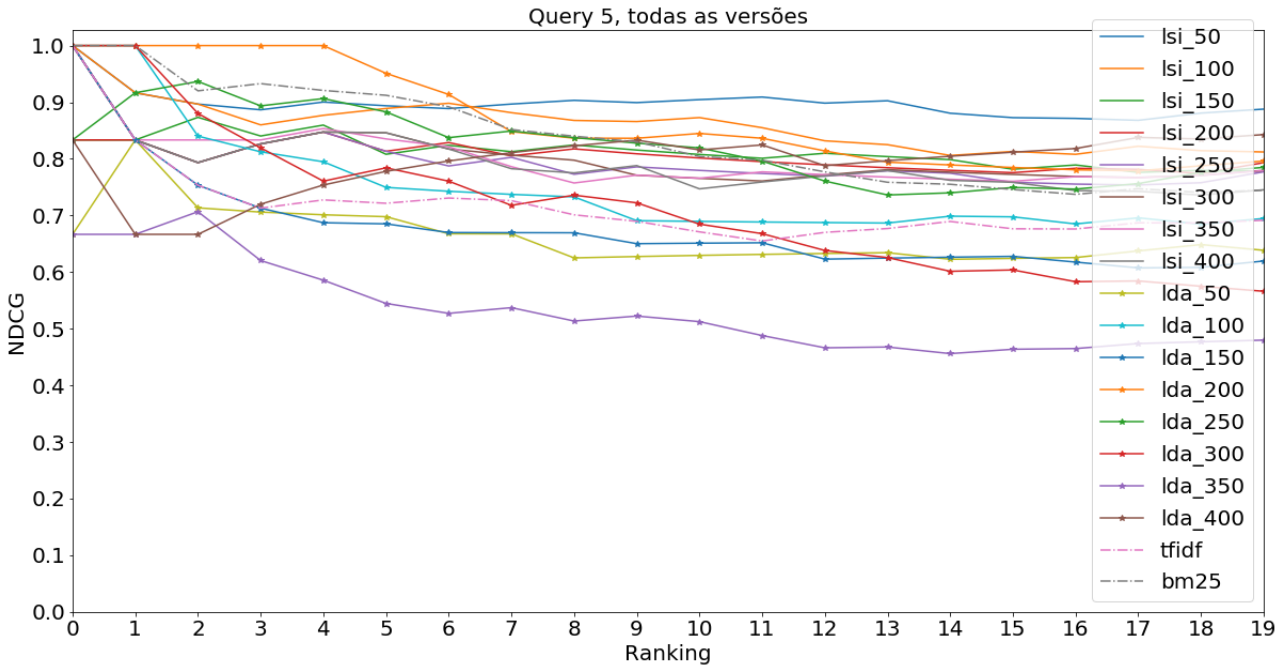


Figura B.24: PCE - performance da *Query* 5, todas as versões



B.1.7 *Query 6*

A *Query 6* pretende localizar pareceres sobre casos em que o réu condenado por dirigir embriagado alega ausência de provas, sendo que houve depoimento de pessoas ou agente público atestando a condição de embriaguês. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "testemunhas atestam que o réu tinha sinais claros de alcoolemia, com andar cambaleante, olhos vermelhos, andar trôpego e hálito com odor etílico"

- versão 2:

  > "embriaguês atestada por testemunhas e policiais que observaram fala embolada e halitose"

- versão 3:

  > "réu recusou-se a realizar teste do bafômetro porém havia sinais claros de que estava embrigado"

Quantidade de documentos relevantes: 5
Quantidade de documentos muito relevantes: 14

Figura B.25: PCE - performance da *Query* 6, versão 1

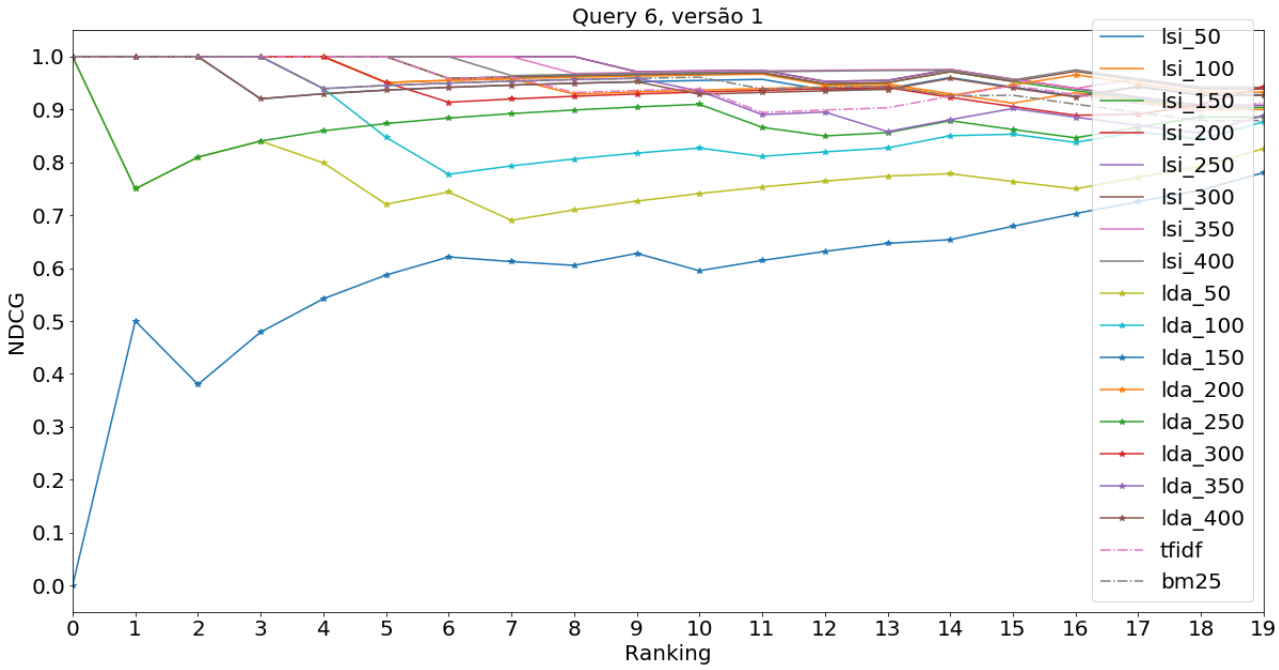


Figura B.26: PCE - performance da *Query* 6, versão 2

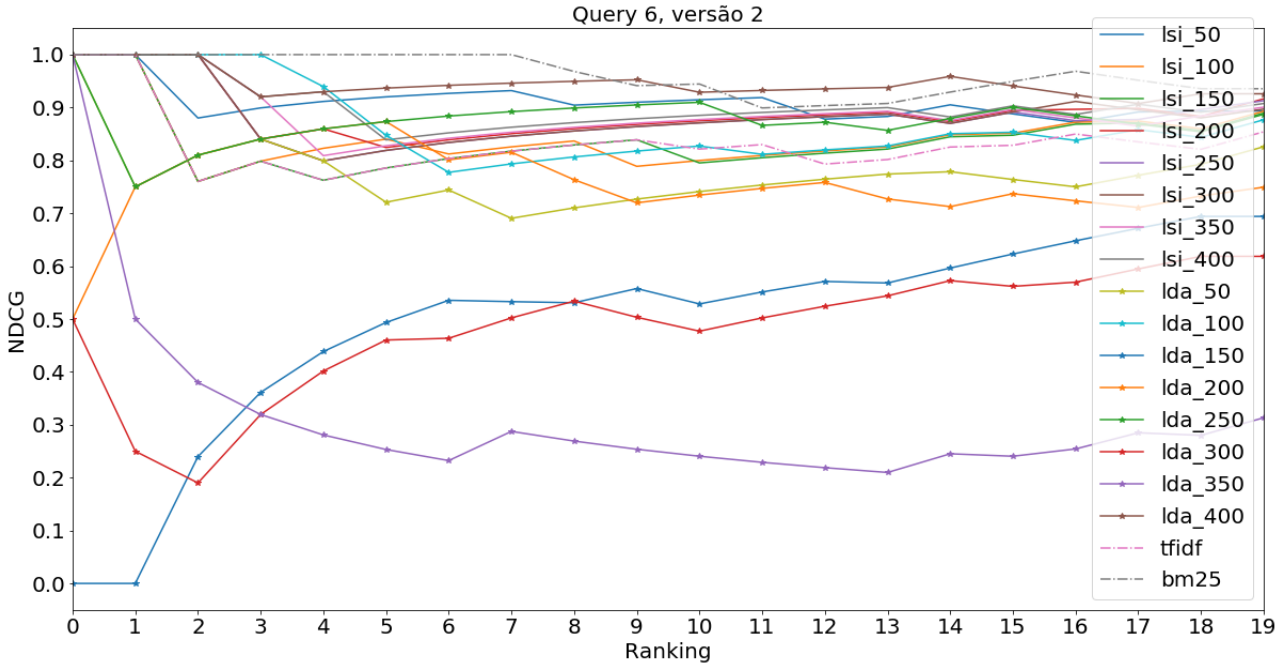


Figura B.27: PCE - performance da *Query* 6, versão 3

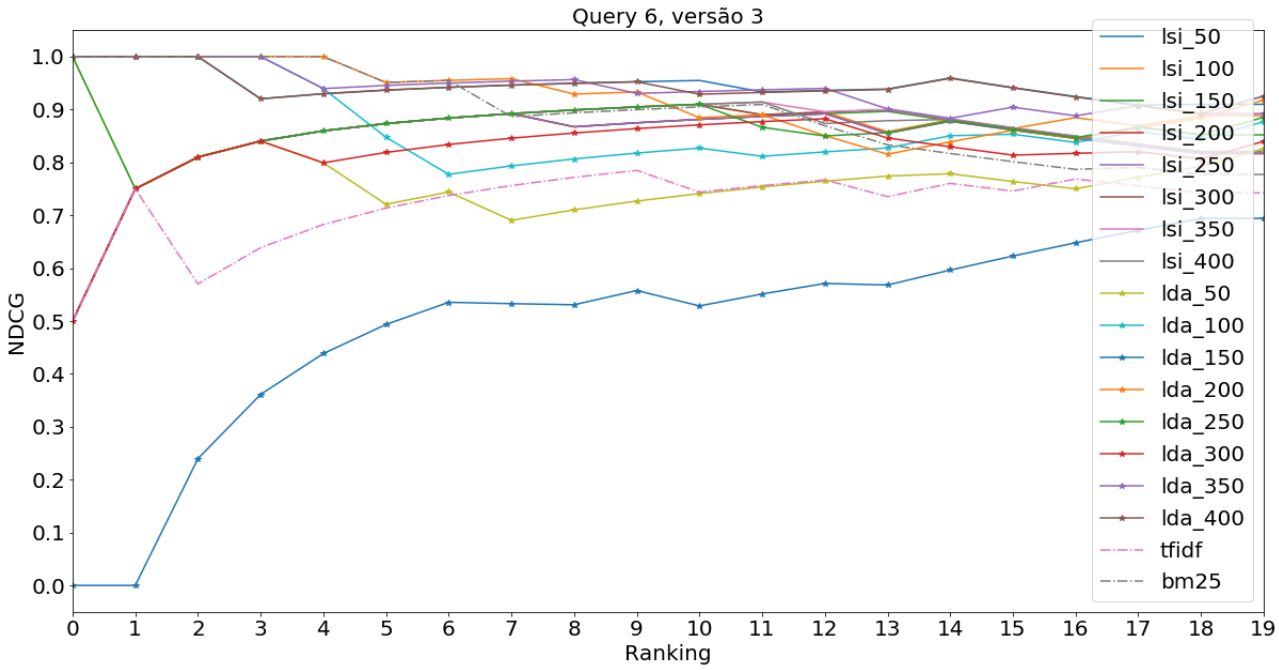
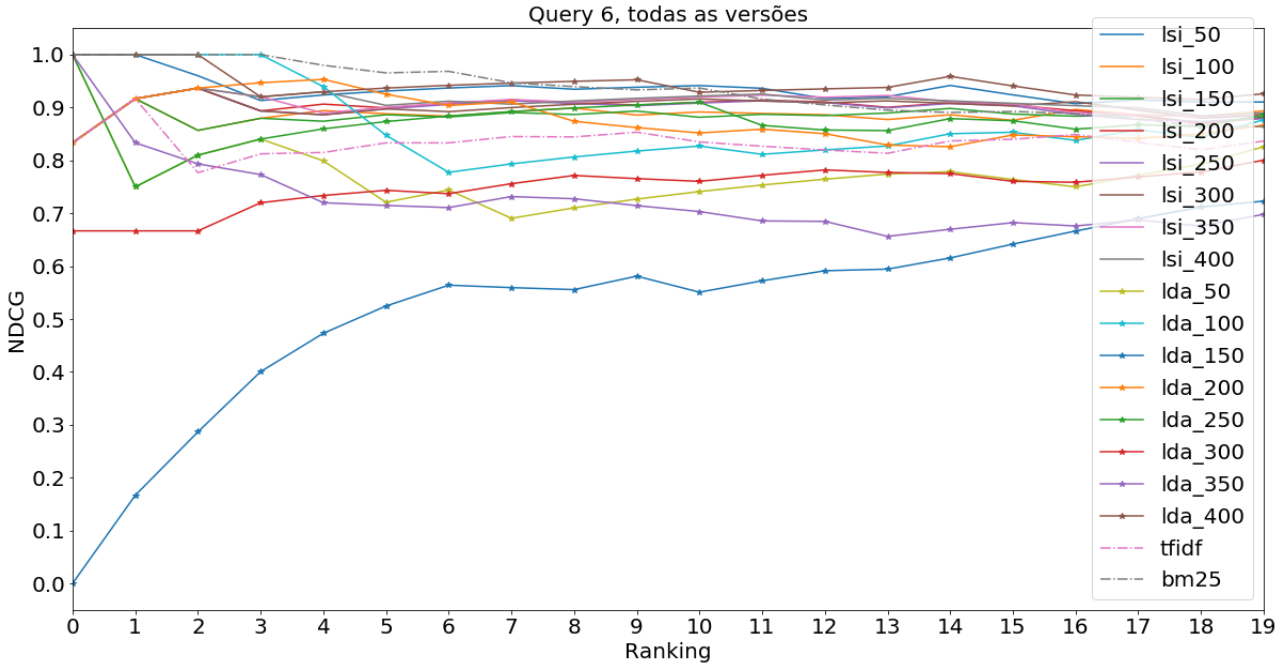


Figura B.28: PCE - performance da *Query* 6, todas as versões



B.1.8 *Query 7*

A *Query 7* pretende localizar pareceres sobre casos submetidos ao tribunal do juri em que o promotor apelante entende que a decisão dos jurados está em nítida contradição com as provas dos autos do processo. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "decisão dos jurados contrariam as provas porque reconhecidas a autoria e materialidade a absolvição só poderia ter ocorrido em função do excludente de ilicitude e isso não foi parte da tese defensiva"

- versão 2:

  > "legítima defesa não foi aventada pela defesa de modo que há contradição nos quesitos"

- versão 3:

  > "réu absolvido mesmo tendo os jurados o reconhecido autor do fato"

Quantidade de documentos relevantes: 2
Quantidade de documentos muito relevantes: 1

Figura B.29: PCE - performance da *Query* 7, versão 1

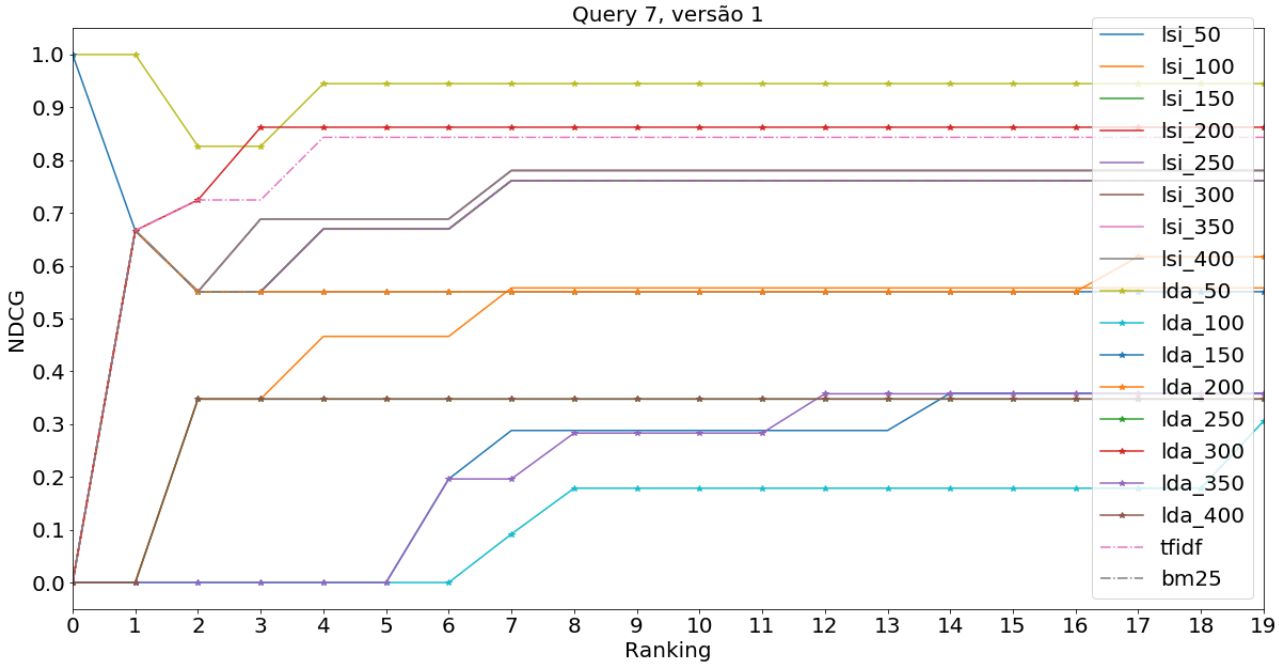


Figura B.30: PCE - performance da *Query* 7, versão 2

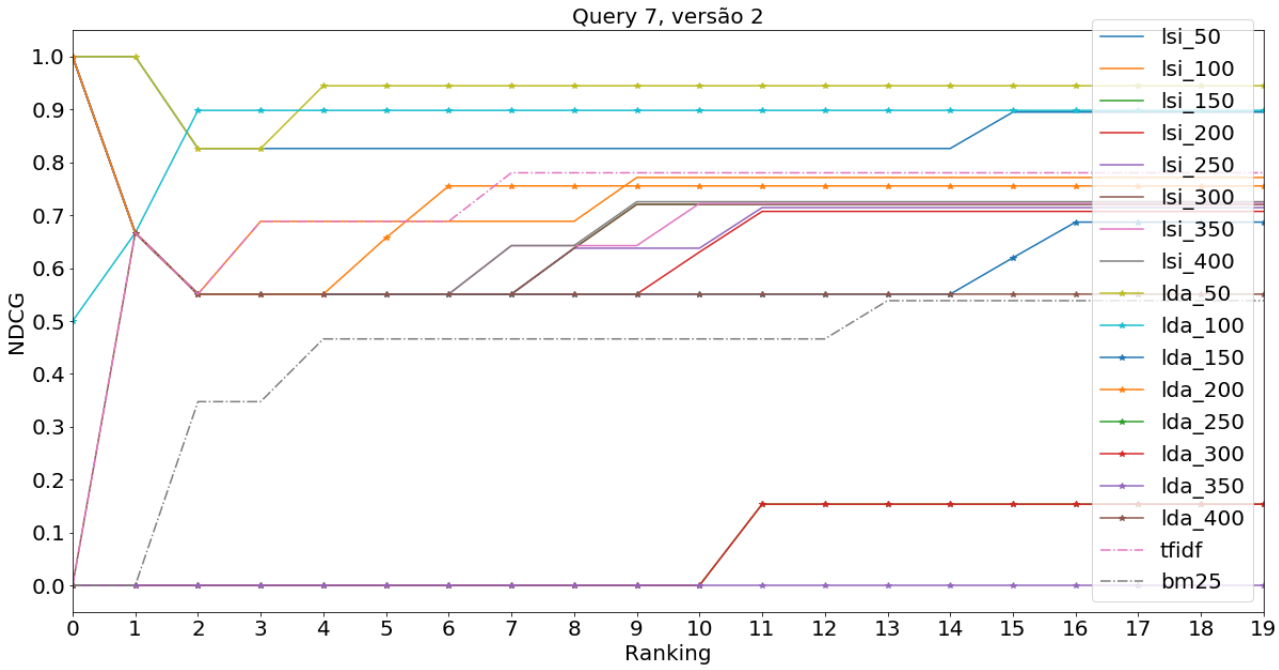


Figura B.31: PCE - performance da *Query* 7, versão 3

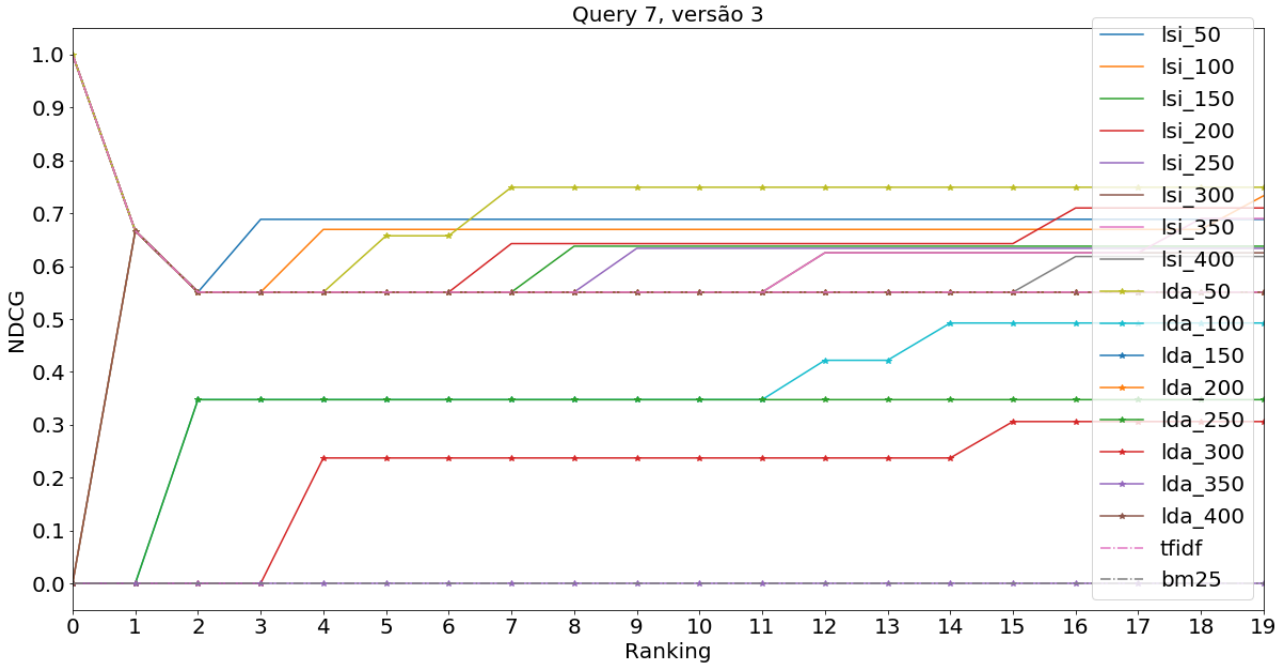
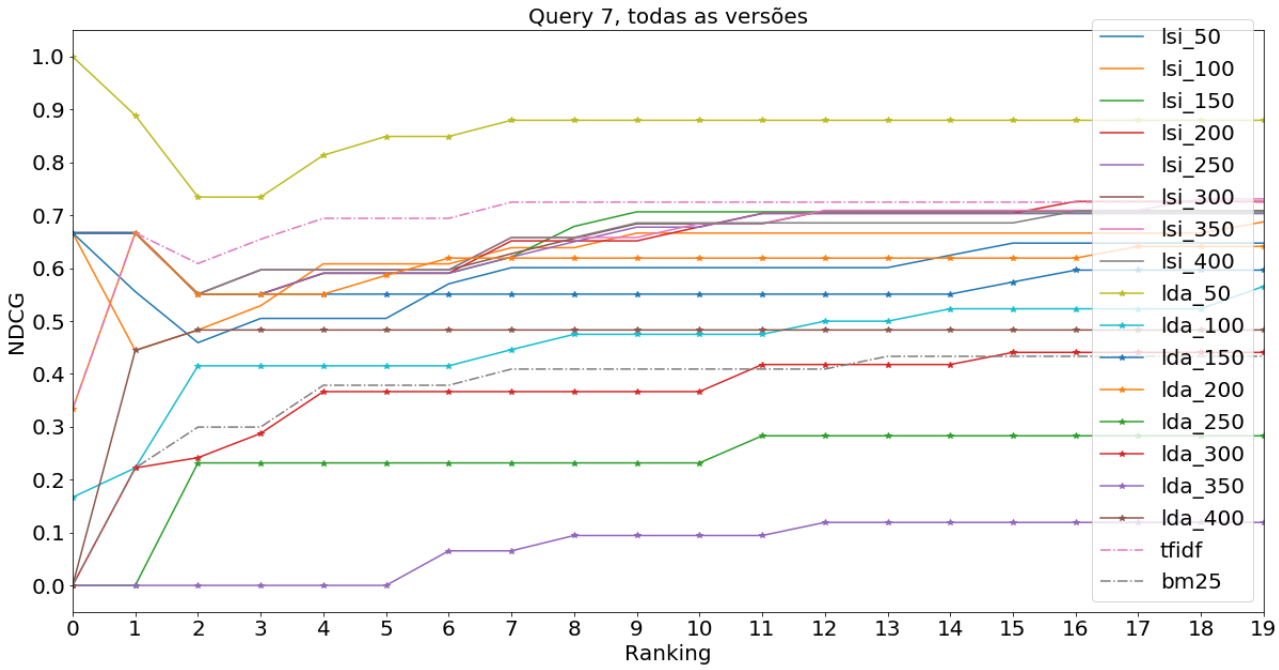


Figura B.32: PCE - performance da *Query* 7, todas as versões



B.1.9 *Query 8*

A *Query 8* pretende localizar pareceres sobre casos submetidos ao tribunal do juri em que o apelante entende que há contradição nas decisões dos jurados acerca dos quesitos estabelecidos para o caso. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "Não houve esclarecimento das contradições das decisões dos quesitos incorrendo em nulidade do julgamento"

- versão 2:

  > "nulidade do julgamento deve ser admitida considerando que o juiz deveria ter se pronunciado acerca das decisões contraditórias nos quesitos anteriores"

- versão 3:

  > "decisão dos jurados nos quesitos não são harmônicas entre si havendo necessidade de esclarecimento por parte do juiz"

Quantidade de documentos relevantes: 0

Quantidade de documentos muito relevantes: 1

Figura B.33: PCE - performance da *Query* 8, versão 1

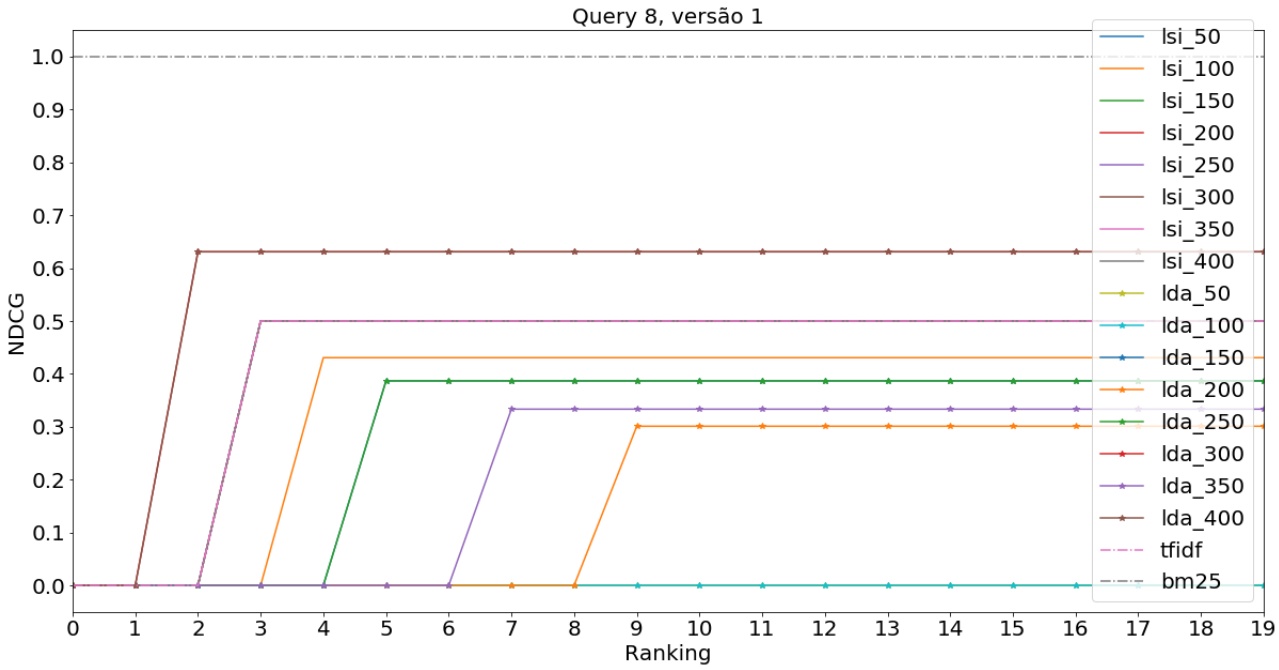


Figura B.34: PCE - performance da *Query* 8, versão 2

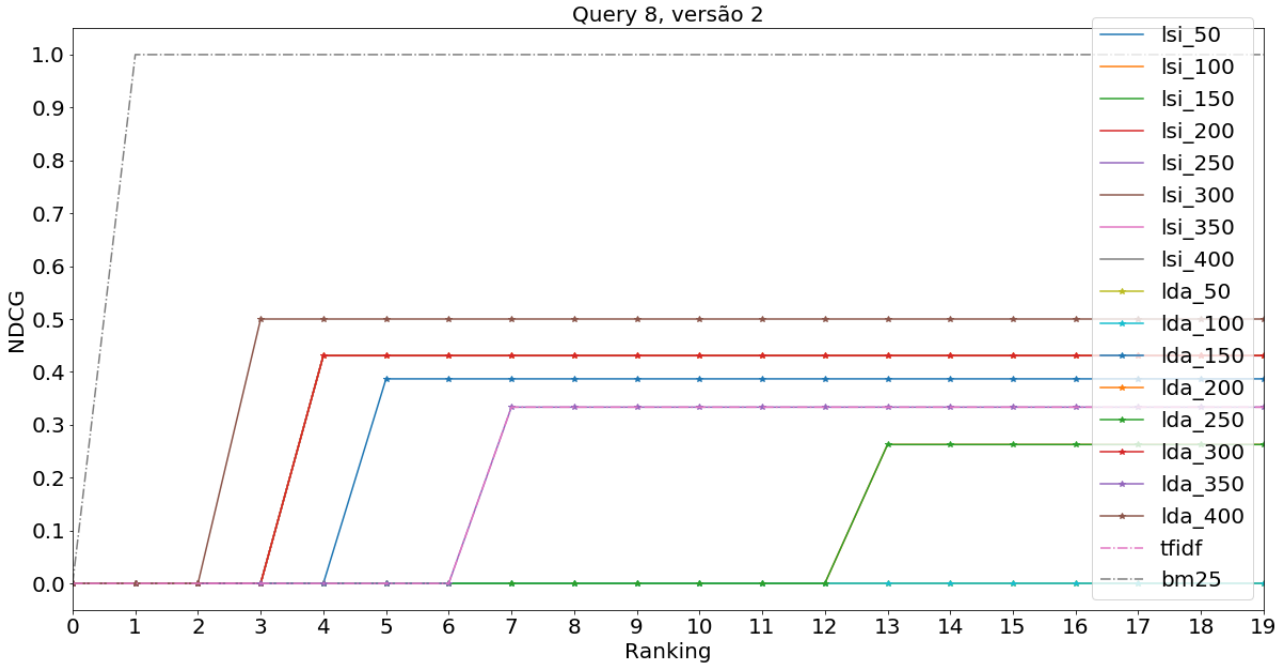


Figura B.35: PCE - performance da *Query* 8, versão 3

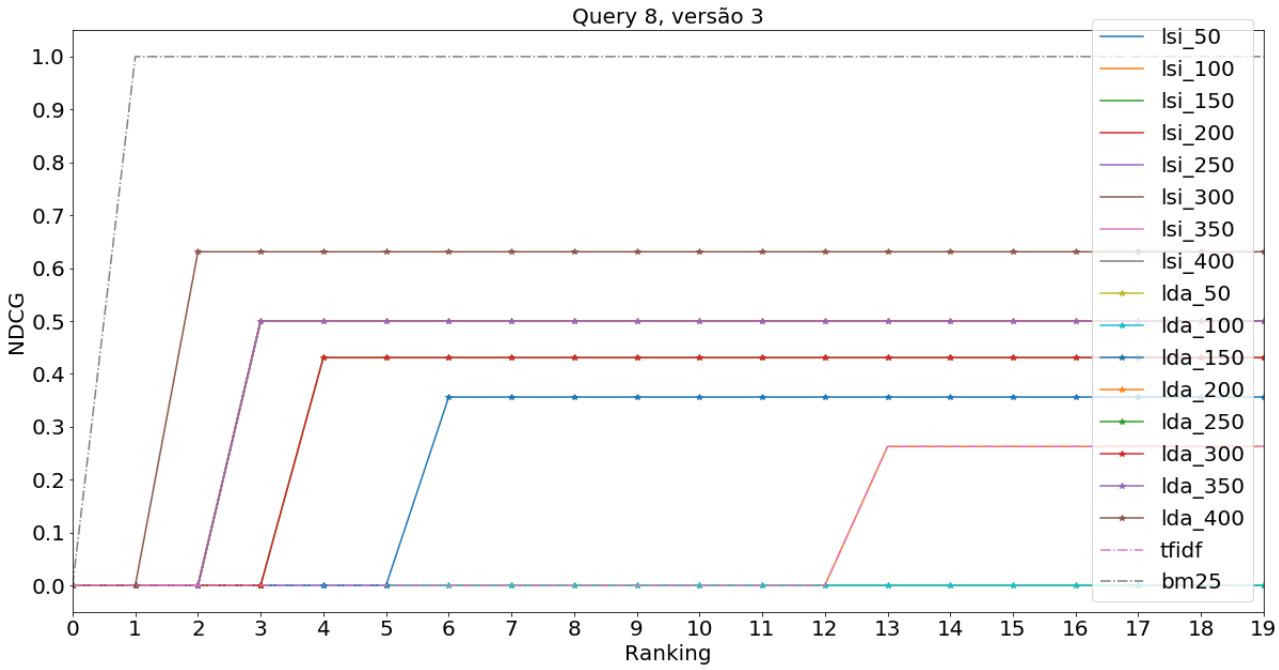
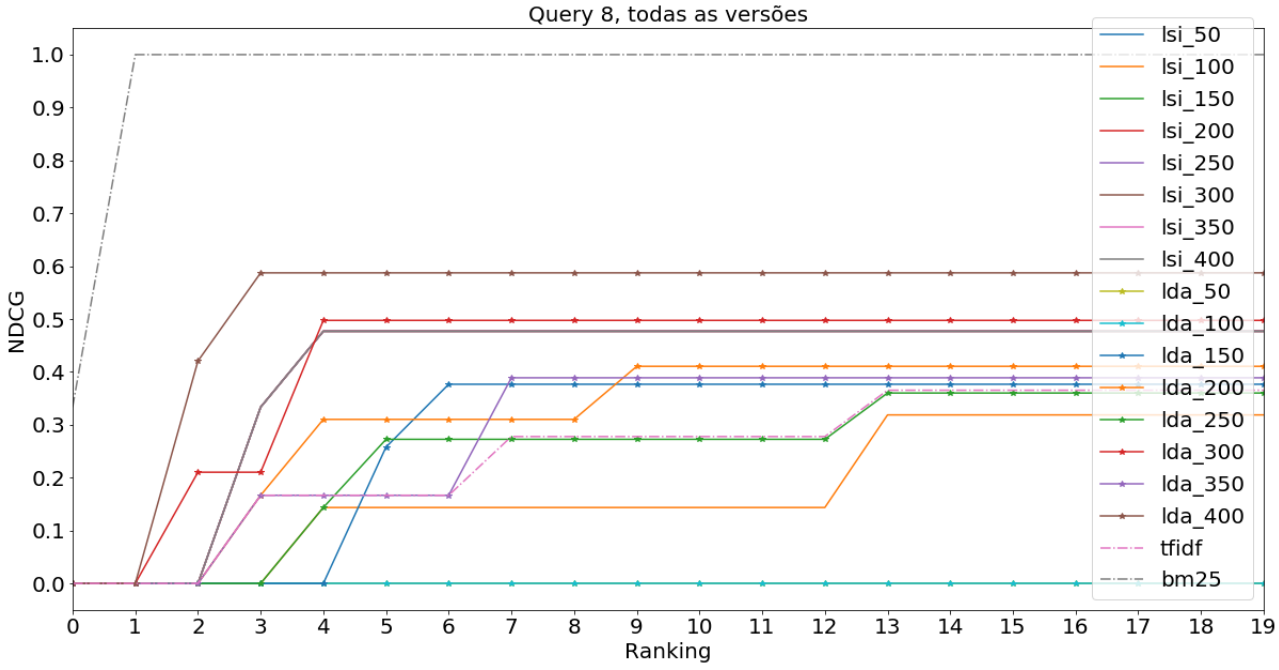


Figura B.36: PCE - performance da *Query* 8, todas as versões



B.1.10 *Query 9*

A *Query 8* pretende localizar pareceres sobre casos submetidos ao tribunal do juri em que o apelante entende que o magistrado que proferiu a sentença não foi o mesmo que conduziu a instrução processual. Para recuperar documentos com esta argumentação, temos 3 versões de textos de pesquisa:

- versão 1:

  > "réu alega nulidade da sentença proferida por juiz diferente daquele que conduziu a instrução"

- versão 2:

  > "juiz que conduziu a instrução deve ser o mesmo a proferir a sentença"

- versão 3:

  > "sentença proferida por magistrado incompente"

Quantidade de documentos relevantes: 0
Quantidade de documentos muito relevantes: 1

Figura B.37: PCE - performance da *Query* 9, versão 1

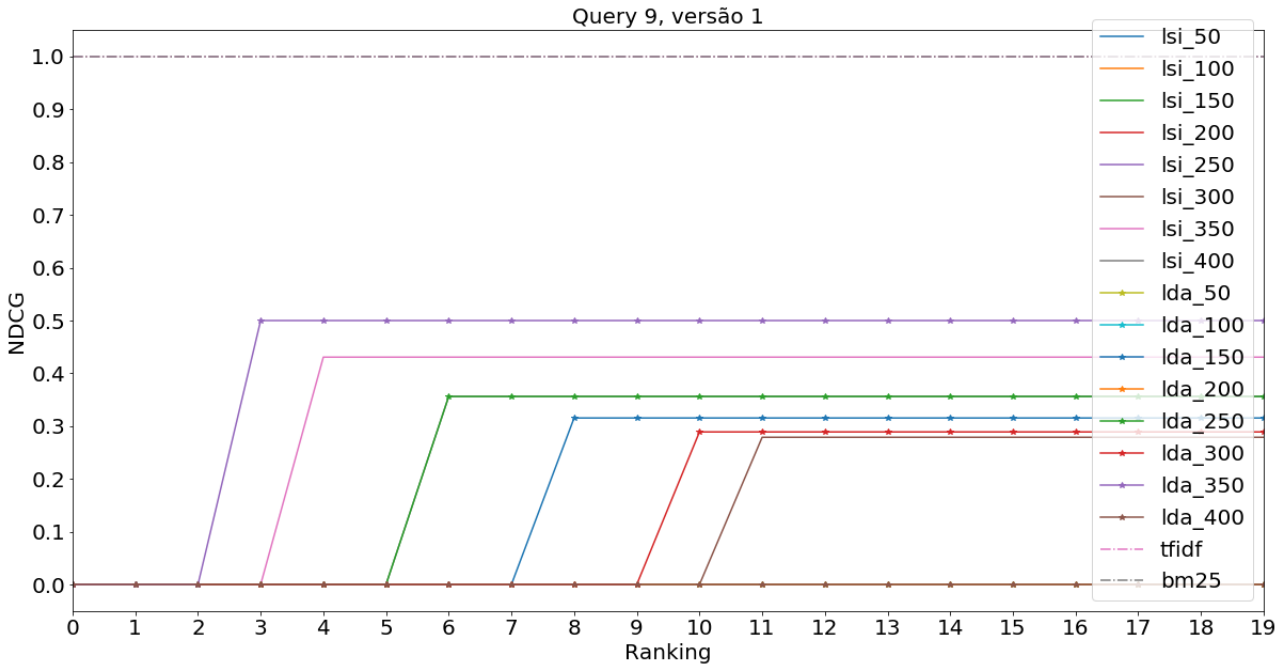


Figura B.38: PCE - performance da *Query* 9, versão 2

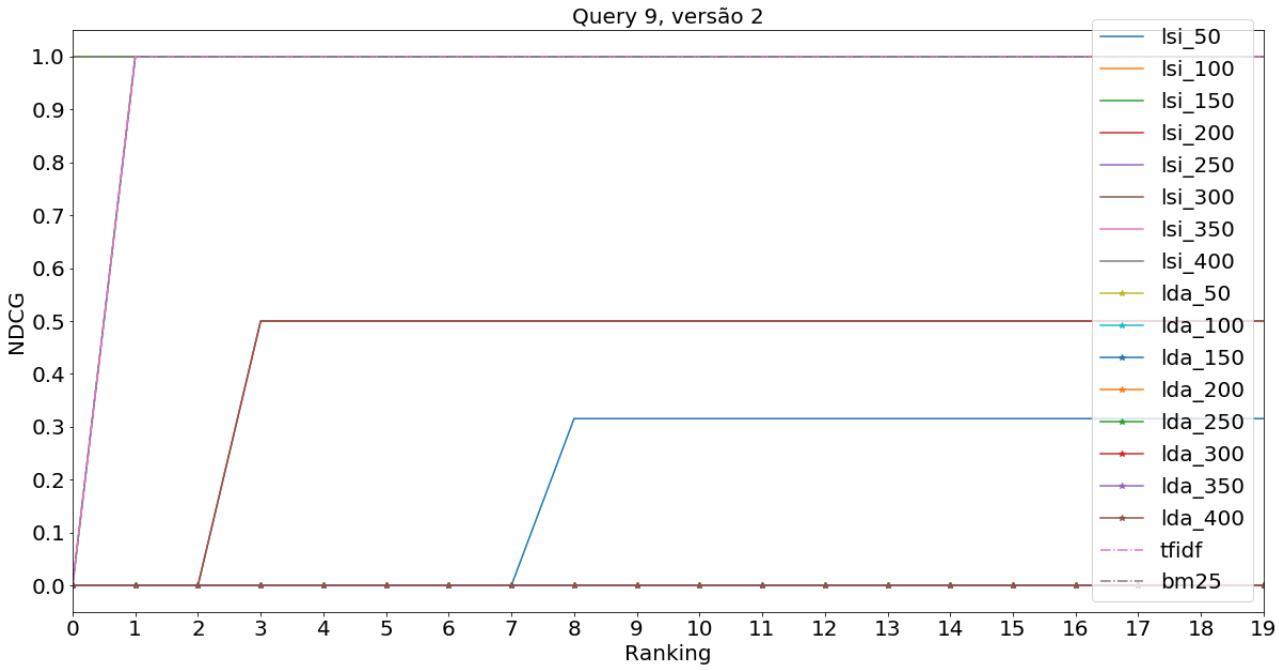


Figura B.39: PCE - performance da *Query* 9, versão 3

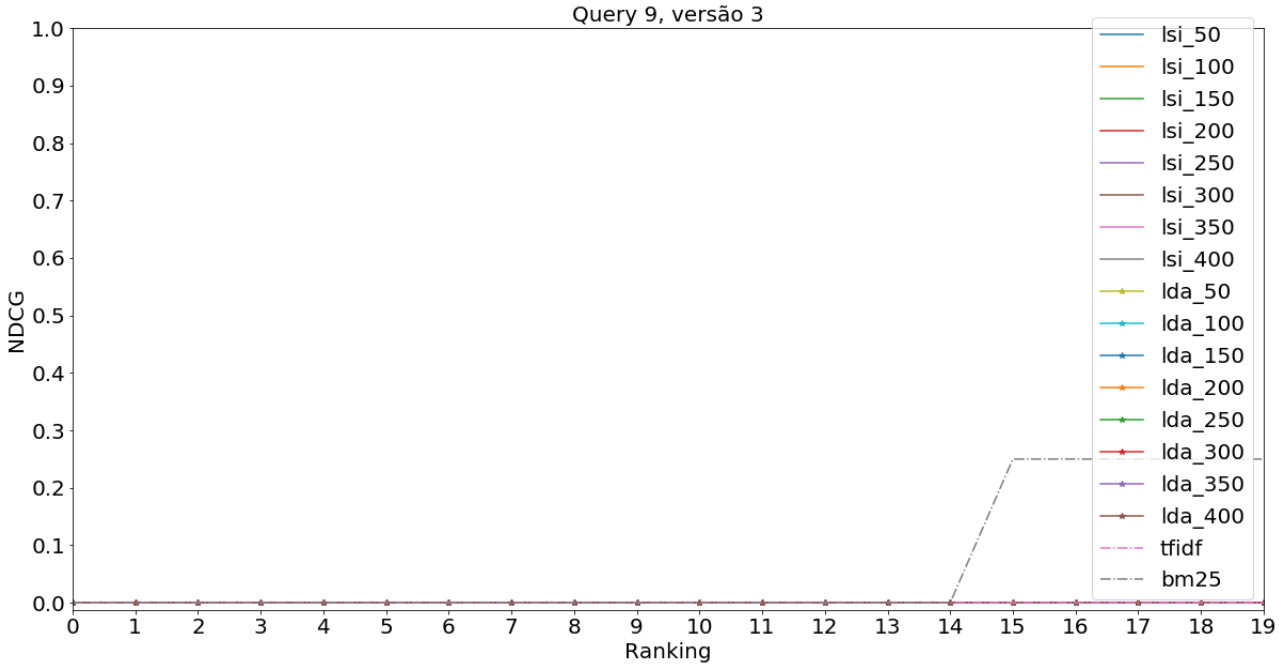
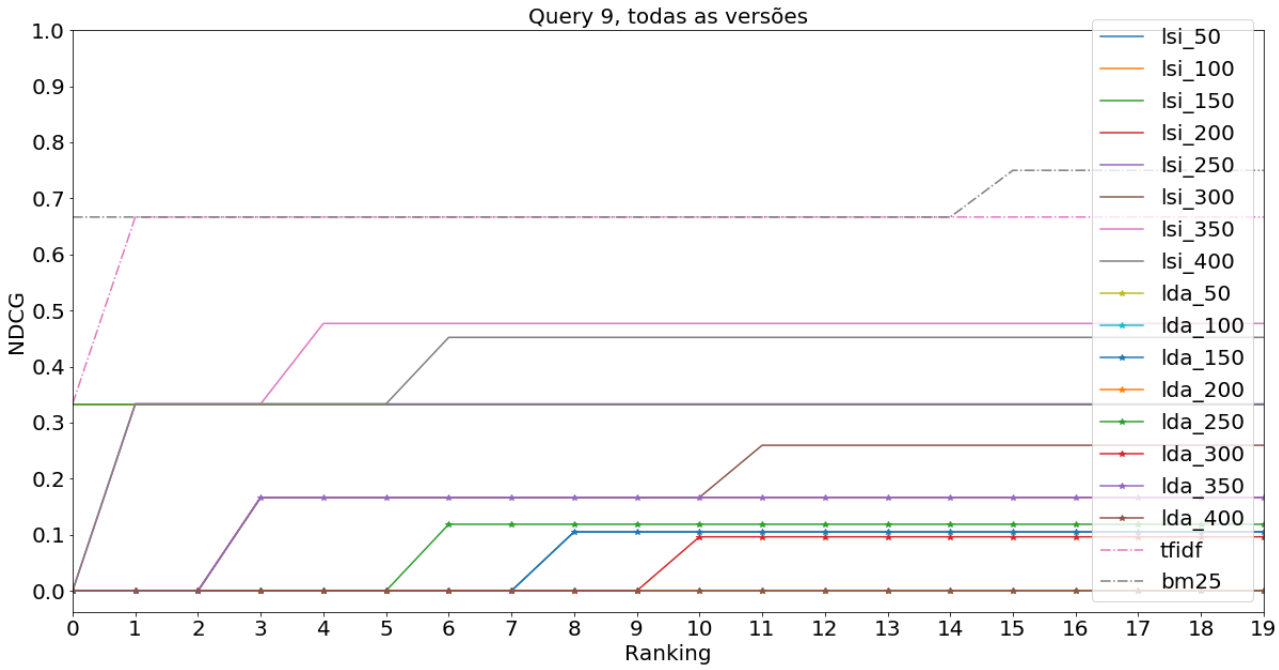


Figura B.40: PCE - performance da *Query* 9, todas as versões



B.2 Comparação dos modelos sem enriquecimento

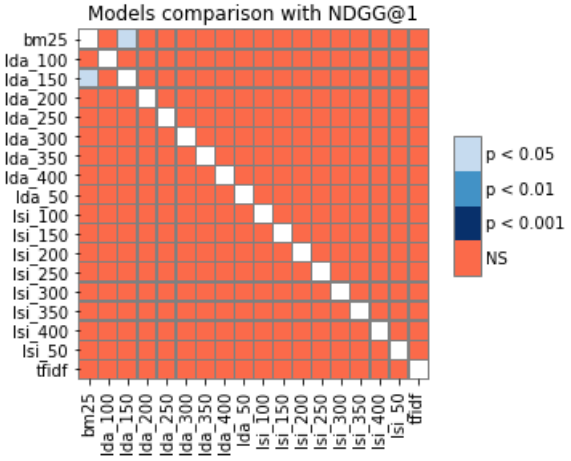


Figura B.41: Comparação entre todos os modelos na posição 1 do ranking

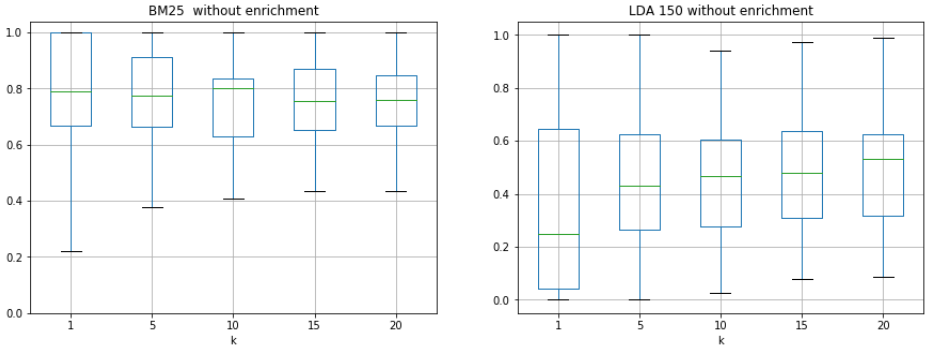


Figura B.42: Diferenças entre os modelos na posição 1 do ranking

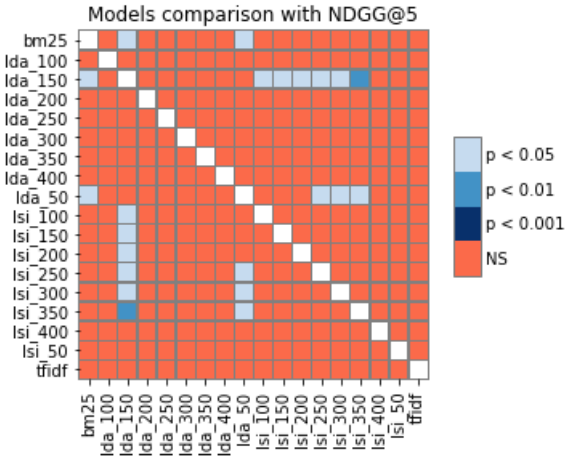


Figura B.43: Comparação entre todos os modelos na posição 5 do ranking

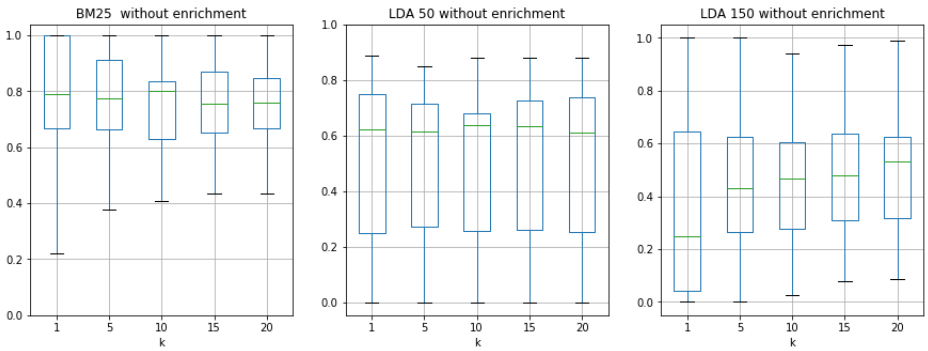


Figura B.44: Diferenças entre os modelos na posição 5 do ranking

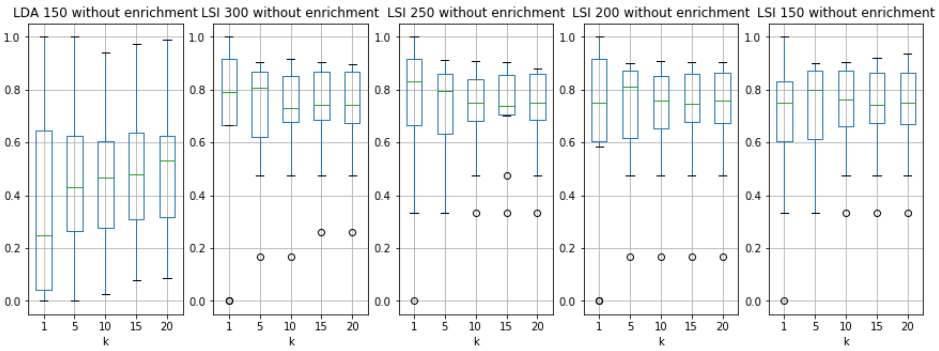


Figura B.45: Diferenças entre os modelos na posição 5 do ranking

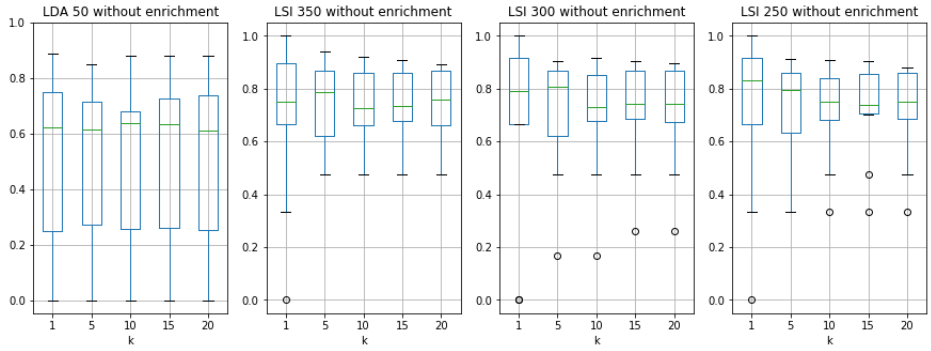


Figura B.46: Diferenças entre os modelos na posição 5 do ranking

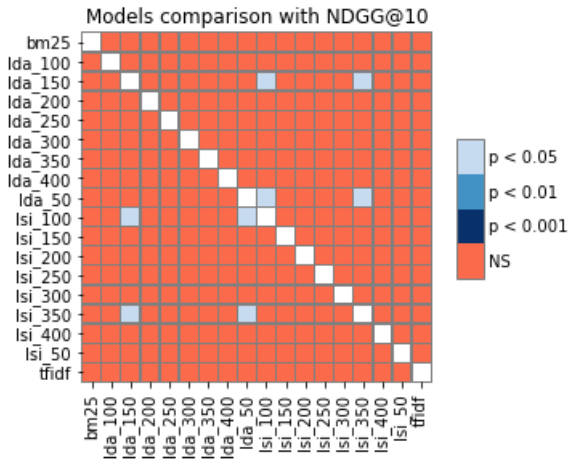


Figura B.47: Comparação entre todos os modelos na posição 10 do ranking

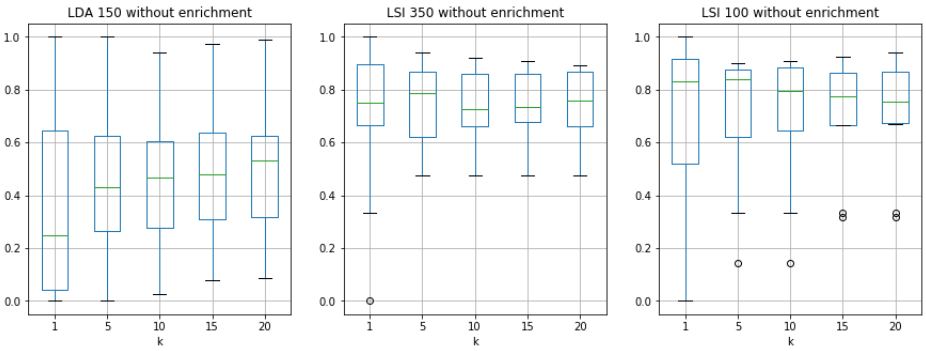


Figura B.48: Diferenças entre os modelos na posição 10 do ranking

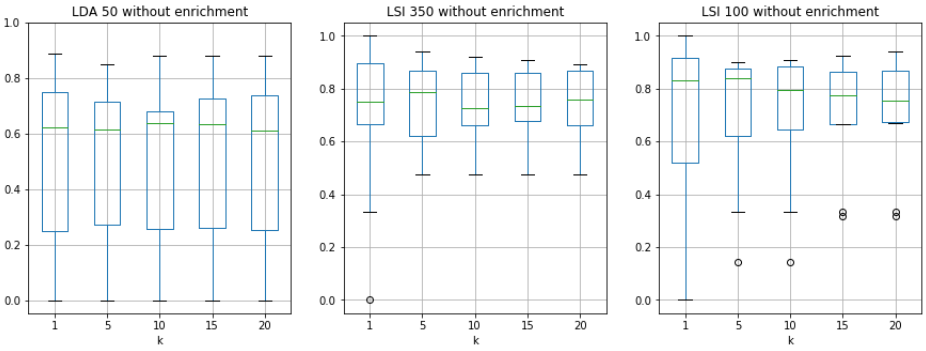


Figura B.49: Diferenças entre os modelos na posição 10 do ranking

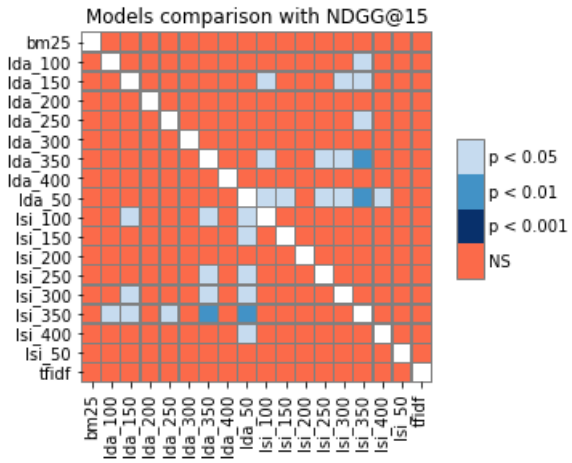


Figura B.50: Comparação entre todos os modelos na posição 15 do ranking

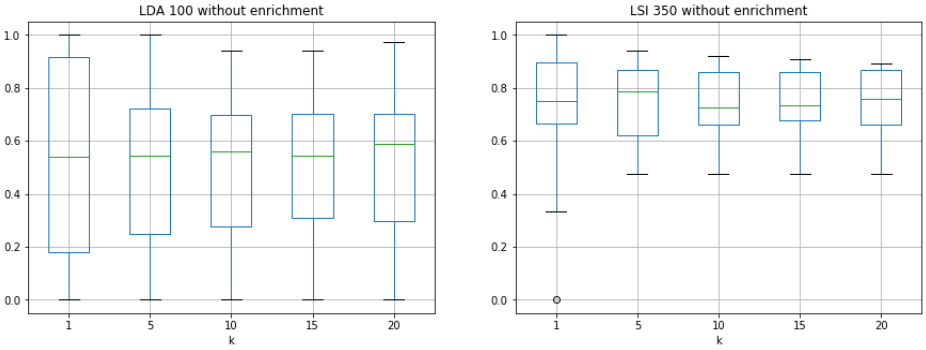


Figura B.51: Diferenças entre os modelos na posição 15 do ranking

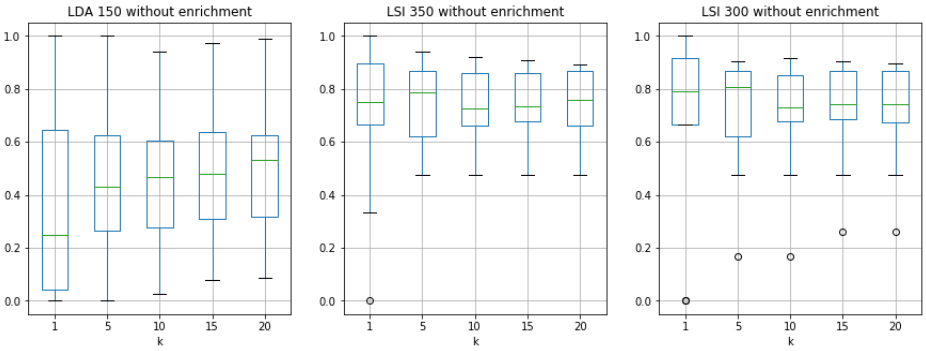


Figura B.52: Diferenças entre os modelos na posição 15 do ranking

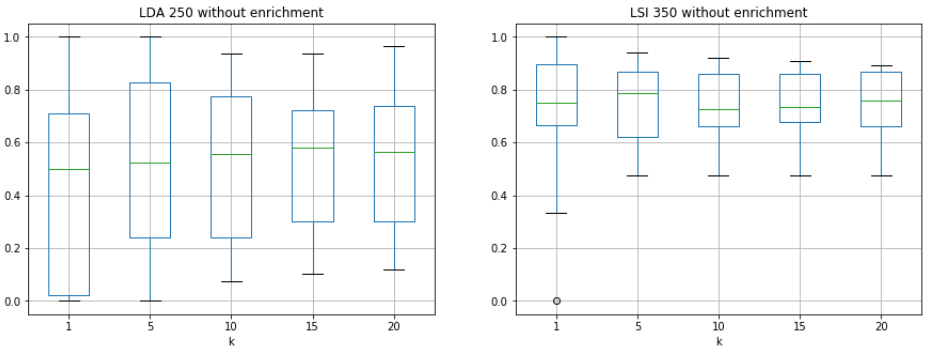


Figura B.53: Diferenças entre os modelos na posição 5 do ranking

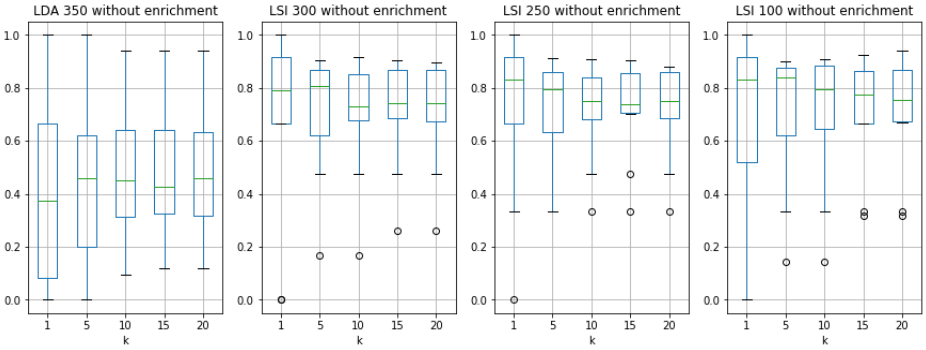


Figura B.54: Diferenças entre os modelos na posição 5 do ranking

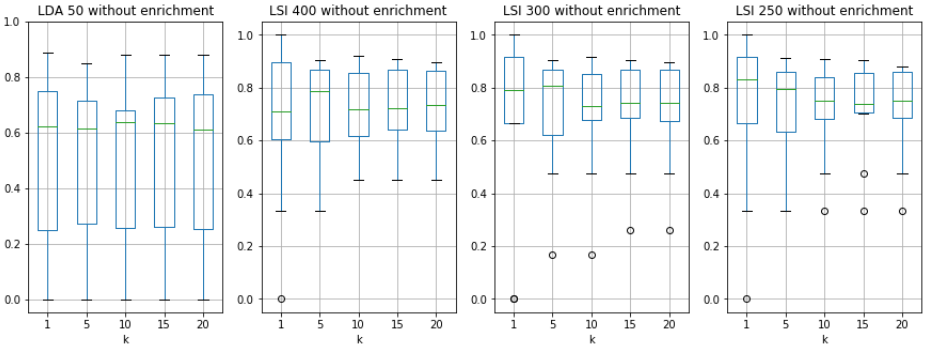


Figura B.55: Diferenças entre os modelos na posição 5 do ranking

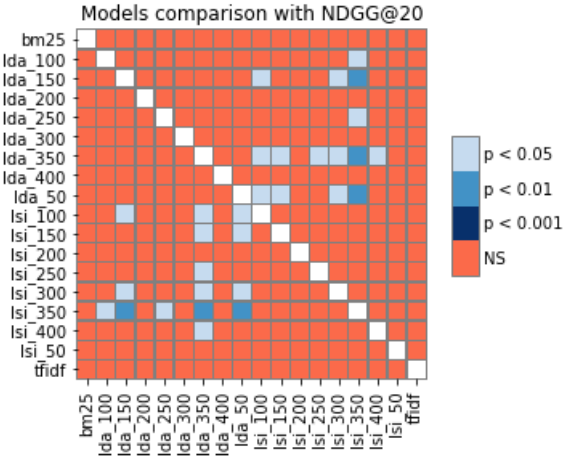


Figura B.56: Comparação entre todos os modelos na posição 20 do ranking

B.3 Comparação dos modelos com enriquecimento

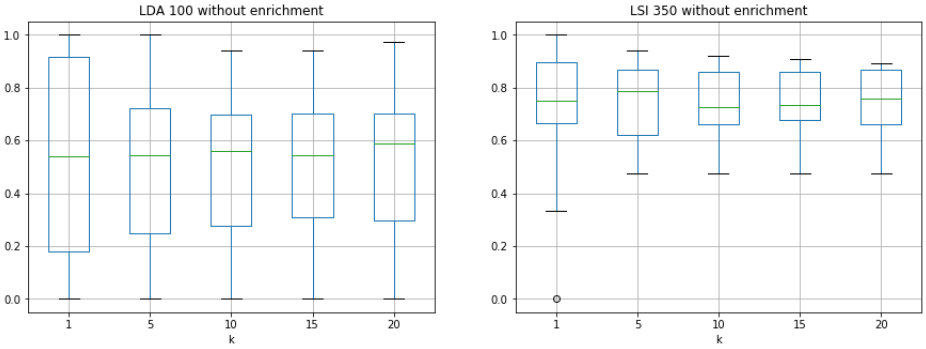


Figura B.57: Diferenças entre os modelos na posição 20 do ranking

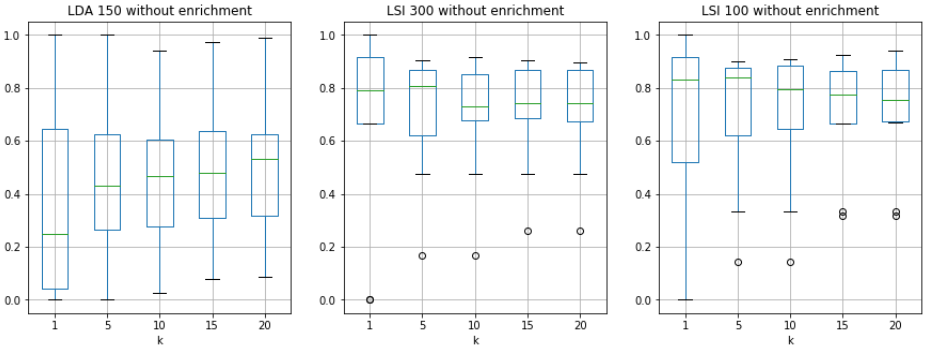


Figura B.58: Diferenças entre os modelos na posição 20 do ranking

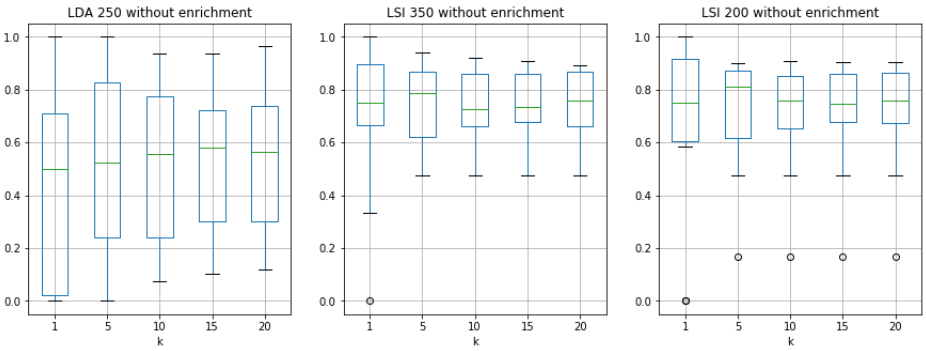


Figura B.59: Diferenças entre os modelos na posição 20 do ranking

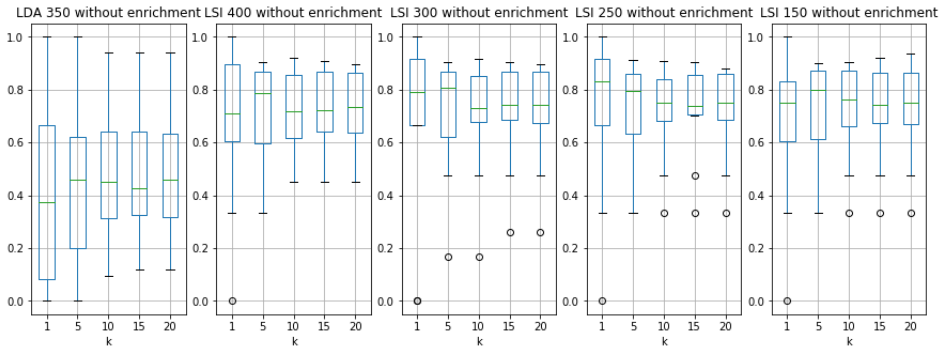


Figura B.60: Diferenças entre os modelos na posição 20 do ranking

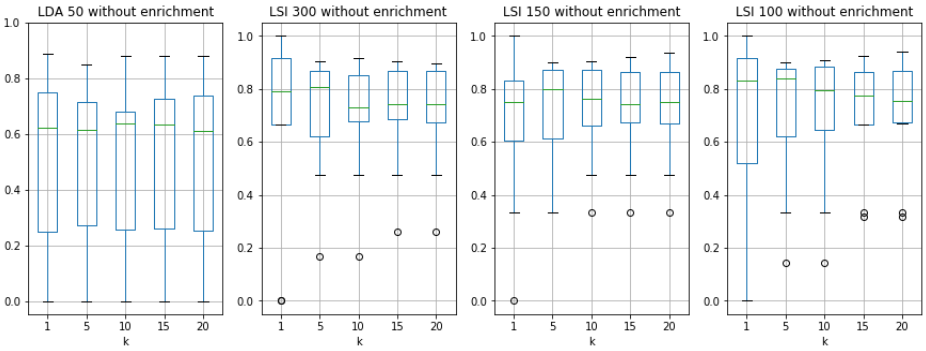


Figura B.61: Diferenças entre os modelos na posição 20 do ranking

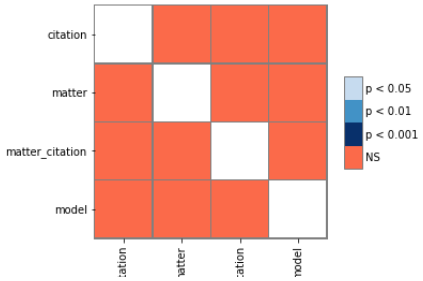


Figura B.62: Comparação de modelos TF-IDF com enriquecimento

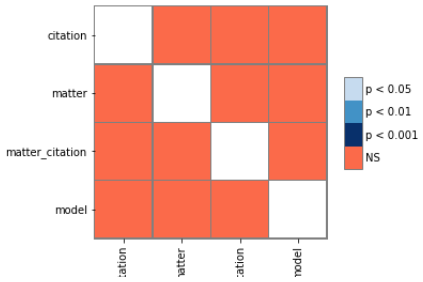


Figura B.63: Comparação de modelos BM25 com enriquecimento

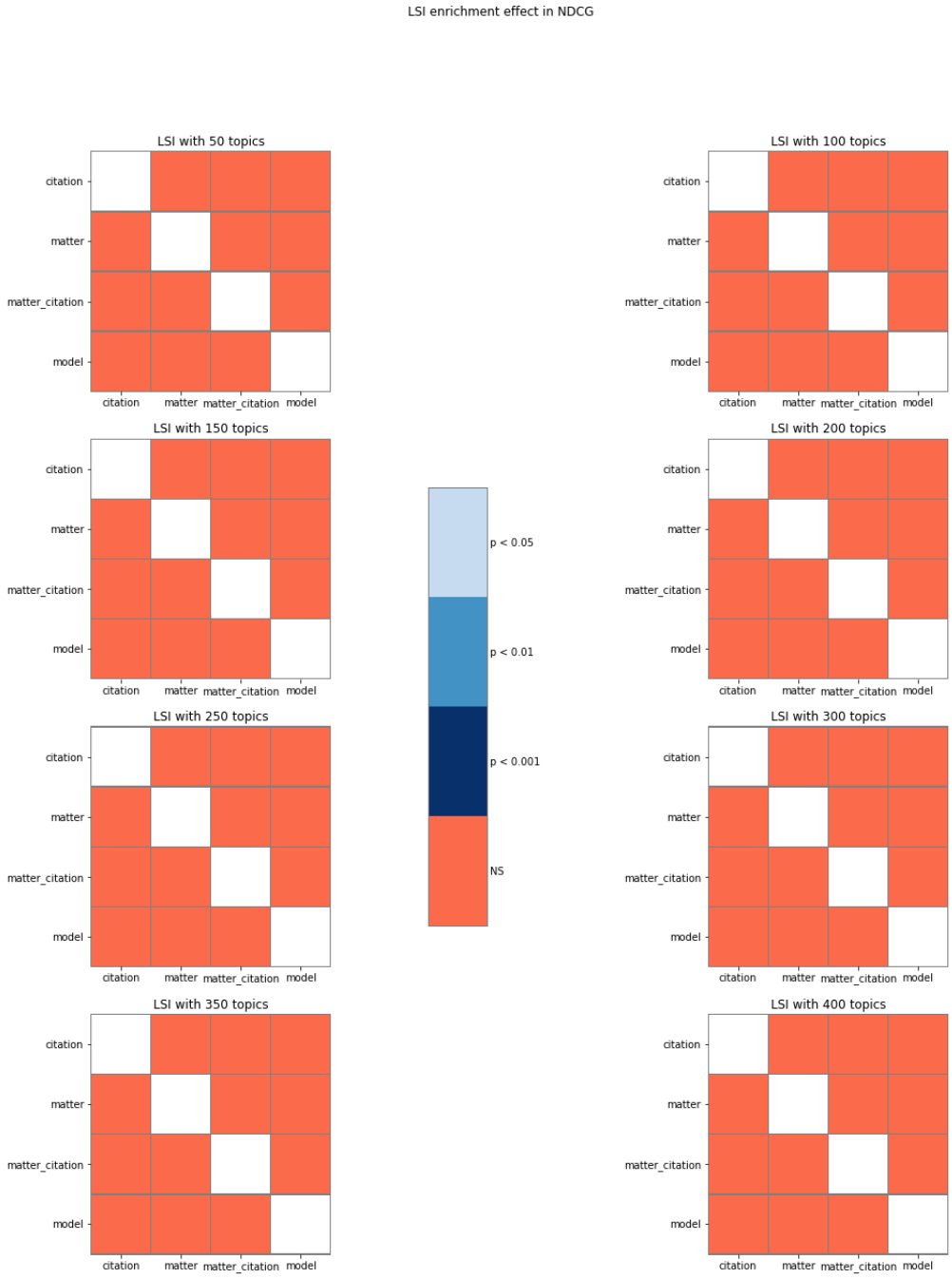


Figura B.64: Comparação de modelos LSI com enriquecimento

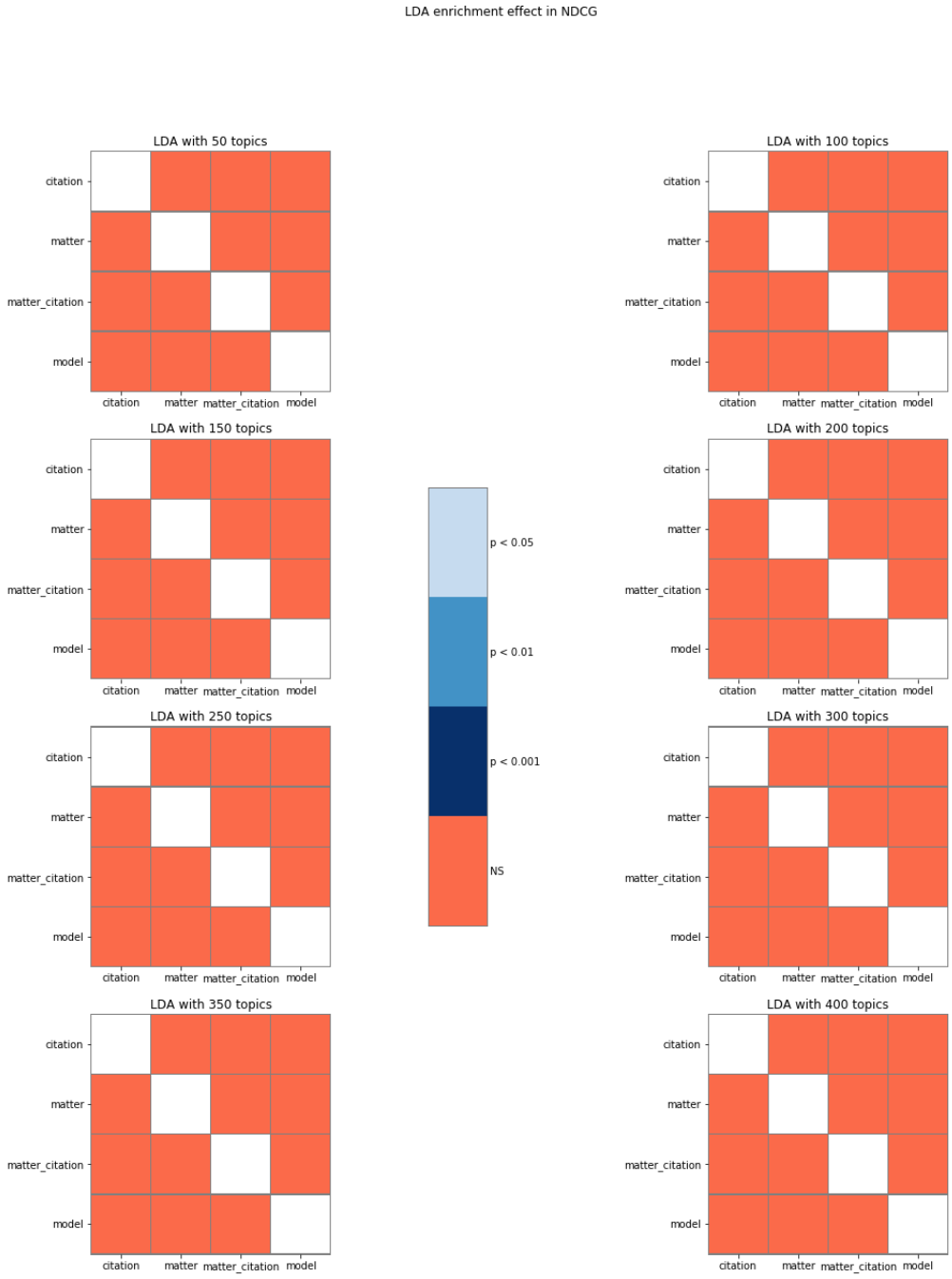


Figura B.65: Comparação de modelos LDA com enriquecimento

Apêndice C

Exemplo de pareceres

Neste apêndice transcrevemos o texto extraído de dois pareceres, um referente a uma apelação criminal, emitido pela 1a Procuradoria de Justiça criminal, e um parecer em *Habeas Corpus* da 3a Procuradoria de Justiça Criminal Especializada. Os pareceres aqui transcritos tiveram omitidos os nomes dos envolvidos, número do processo e do parecer.

C.1 Exemplo de parecer em apelação criminal

APELAÇÃO CRIMINAL Nº (OMITIDO)

SEGUNDA TURMA CRIMINAL

APELANTE: (OMITIDO)

APELADO: MINISTÉRIO PÚBLICO DO DISTRITO FEDERAL E TERRITÓRIOS

PARECER nº (OMITIDO)

I – Relatório:

Trata-se de apelação criminal interposta por (OMITIDO) em face da sentença de fls. 131/136, proferida pelo MM. Juiz de Direito da Vara Criminal e do Tribunal do Júri de Brazlândia/DF, que julgou procedente a pretensão punitiva estatal, para condená-lo como incurso no artigo 155, §1º e § 4º, Inciso II, e art. 147, caput, ambos do Código Penal, à pena de 01 (um) ano e 09 (nove) meses de reclusão, bem como ao pagamento de 10 (dez) dias-multa, calculados à razão mínima legal, ao primeiro crime, e ao segundo, 01 (um) mês de detenção. Ambos os delitos devem ser cumpridos no regime inicial aberto. As penas corporais foram substituídas por duas restritivas de direitos.

O sentenciado, em suas razões de apelação (fls. 148-v/160), pugna pela absolvição dos crimes que lhes foram imputados, nos termos do art. 386, Incisos III, V ou VIII, do Código de Processo Penal. Requer, ainda, seja desclassificada a modalidade qualificada do crime de furto para sua forma simples, além da absolvição por atipicidade da conduta no que concerne ao crime de ameaça.

Assim, analisados os autos, esta Procuradoria de Justiça oficia.

II – Apreciação:

Presentes os requisitos de admissibilidade do recurso necessário ao processamento, passa-se ao exame do mérito.

Não assiste razão ao recorrente (OMITIDO), ante a juntada aos autos de provas robustas a fim de sustentarem sua condenação criminal.

Consta da peça acusatória (fls. 02/02-a):

"DO CRIME DE FURTO QUALIFICADO DURANTE O REPOUSO NOTURNO

No dia 28 de novembro de 2015, por volta de 01h10min, durante o período do repouso noturno, na Quadra 19, Casa 03, Setor Tradicional de Brazlândia/DF, o denunciado, agindo com vontade livre e consciente, mediante escalada, subtraiu, em proveito próprio: 01 (uma) caixa sifonada, marca Tigre, cor branca; 01 (um) ralo marca Tigre, cor branca; 02 (dois) pares de luva multiuso de látex, cor amarela, marca Kalipso; 03 (três) discos de corte, diamantado, marca Irwin; 01 (um) disco de corte diamantado, marca Nak Fast; 01 (um) disco de corte diamantado, marca Vonder; 14 (quinze) abraçadeiras copo 3/4, marca Thlemar; 03 (três) conexões de cano, marca Tigre, 50x40; 10 (dez) conexões de cano NBR 5648, marca Tigre; 04 (quatro) curvas 90o, 25x1/2, marca Tigre, 05 (cinco) Ts marca Tigre, NBR 5648; 04 (quatro) curvas de conexão, marca Tigre, TP242, bens esses de propriedade de (OMITIDO). Apurou-se que no lote onde ocorreram os fatos havia uma edificação que era residência onde a vítima residia com sua família e outra ainda em construção. No dia, hora e local mencionados o denunciado, mediante escalada, pulou o muro do lote, adentrou a edificação em construção, separou os citados objetos, colocou-os em uma sacola de plástico azul, e saiu da edificação em construção. Ocorre que a vítima havia escutado barulho de cães e percebido uma movimentação, tendo saído de sua residência com seu filho, logrando abordar o denunciado. O denunciado, então, largou a sacola com os objetos subtraídos e tentou fugir, utilizando um cavalete para pular o muro do lote, mas foi contido ate a chegada da Polícia Militar.

DO CRIME DE AMEAÇA

Nas mesmas circunstâncias de tempo e lugar, o denunciado, agindo com vontade livre e consciente, ameaçou (OMITIDO) por palavras de causar-lhe mal injusto e grave. Apurou-se que após ser detido por ocasião do crime de furto que cometeu, o denunciado disse à vítima acima citada que iria matá-lo quando saísse da prisão, usando inclusive uma peixeira. As ameaças foram feitas inclusive na presença dos policiais militares que foram ao local."

A materialidade e a autoria dos crimes de furto qualificado e de ameaça restaram sobejamente comprovadas nos autos, mormente pela juntada do Auto de Prisão em Flagrante nº 630/2015 (fls. 02-d e seguintes); Auto de Apresentação e Apreensão nº 699/2015 (fls.

17/18); Registro de Ocorrência Policial nº 11.683/2015-0 (fls. 25/30); Laudo de Avaliação Econômica Indireta (fls. 39/40); Laudo de Exame de Local fls. (94/101), bem como pelas provas orais produzidas no curso da instrução processual.

A vítima (OMITIDO), em juízo (mídia de fl. 109), ratificou a narrativa trazida por ocasião da fase inquisitiva (fls. 02-d/03). Afirmou que no lote onde se deu a empreitada criminosa existe a edificação onde reside e outra ainda em construção. Informou que no dia dos fatos, por volta de 0h30, ouviu o barulho dos cães e por isso ficou atento para o que estava acontecendo do lado de fora. Esclareceu que viu o recorrente no interior do lote, momento em que segurava uma sacola. Alegou que diante de tais circunstâncias se municiou com dois tijolos, a fim de detê-lo, contudo, (OMITIDO) tentou evadir-se do local, pulando o muro que separa o lote da rua, e que para tanto, utilizou-se de um cavalete que estava sobre um monte de terra na área em construção, contudo, com a ajuda de seu filho conseguiu conter a fuga do recorrente até a chegada da Policia Militar. Afirmou que o viu saindo com a referida sacola, porém, no momento da abordagem ela foi deixada dentro do lote. Confirmou que quando os policiais chegaram ao local, o que foi encontrado na sacola se limita aos objetos descritos na exordial acusatória. Por fim, no que diz respeito às ameaças, a vítima concluiu que elas foram perpetradas contra si no interior da Delegacia de Polícia, com os seguintes dizeres: "eu vou te matar, seu desgraçado! Você vai ver".

A testemunha ocular, (OMITIDO), filho da vítima, confirmou em juízo as declarações prestadas na Delegacia de Polícia (fls. 04/05). Disse que, ao contrário do alegado pelo recorrente, ele e seu pai não haviam ingerido bebida alcoólica. Contou que em virtude do latido dos cães, a vítima, (OMITIDO), foi observar o que estava acontecendo na área de construção, momento em que o recorrente foi visto com uma sacola cheia de objetos. Destacou que viu o apelante dispensá-la no momento da abordagem, e que o recorrente utilizou-se do cavalete para subir no muro e empreender sua fuga. Atestou que (OMITIDO) certamente escalou o muro para ingressar no lote, pois, não existe outro meio de acesso, e que provavelmente o degrau que existe no muro vizinho foi utilizado para tanto. Afirmou, ainda, que as ameaças começaram no momento em que o recorrente foi recuado contra a parede pela testemunha e pela vítima e que persistiram desde a chegada da policia até o interior da Delegacia – mídia de fl. 109.

Com efeito, é matéria pacífica, na doutrina e na jurisprudência, que nos crimes contra o patrimônio a palavra da vítima, quando apresentada de maneira firme e coerente, reveste-se de importante força probatória, mormente quando corroborada pelos demais meios de provas, restando, pois, apta a embasar decreto condenatório.

O apelante (OMITIDO), por sua vez, em juízo (mídia de fl. 109), ratificou a negativa de autoria dos crimes declarada na Delegacia (fl. 08/09). Alegou que de fato pulou o muro do lote para ingressar e para sair do local, contudo, adentrou o lugar para realizar

suas necessidades fisiológicas e não para furtar quaisquer bens. Disse que após defecar na residência em construção, foi surpreendido pelo vizinho, qual seja, a pessoa de (OMITIDO) e seu filho, (OMITIDO). Disse que a sacola em questão se tratava de um pedaço de saco de cimento. Quanto às ameaças, asseverou que elas não ocorreram, entretanto, proferiu palavras de baixo calão contra a vítima quando na delegacia.

Como se vê, tal versão exculpatória encontra-se isolada nos autos, é fantasiosa e não se sustenta ante o arcabouço probatório constante do caderno processual, o qual demonstra, indubitavelmente, em especial pelas circunstâncias fáticas apuradas na prova oral, que o réu agiu com animus furandi.

Nesse sentido, são incensuráveis os fundamentos da sentença condenatória, quando concluiu que "a versão apresentada pelo acusado não possui qualquer credibilidade. Não é plausível que alguém adentre uma casa em construção de pessoa supostamente conhecida, durante a madrugada, com a finalidade de fazer suas necessidades fisiológicas. Inclusive, para adentrar o local, foi necessário pular um muro alto, conforme fotografias costadas às fls. 97/101. Entendo que o local e o horário confirmam que a versão dos fatos distoa da realidade" - sentença de fls. 131/136.

Ainda, quanto ao crime de ameaça, este também restou comprovado, pois, as vítimas foram firmes em afirmar que o recorrente as ameaçou, situação em que foi presenciada por policiais militares (fl. 26), logo, não há que se falar em absolvição.

Ressalta-se, que a palavra da vítima em crimes que não deixam vestígios, como no caso da ameaça tem especial relevância. No caso em tela, todo o alegado pelos ofendidos encontra-se respaldo no arcabouço probatório coligido aos autos, além de não terem sido eivadas ou desqualificadas pelas demais informações prestadas em juízo.

Noutro giro, o sentenciado requer que, alternativamente, seja desclassificado o crime de furto qualificado para a sua modalidade simples, além da aplicação do princípio da insignificância.

A configuração da qualificadora do abuso da escalada encontra-se escorreita nos autos. Além dos depoimentos prestados pelo ofendido e pela testemunha, o Laudo de Perícia Criminal (fls. 94/101), evidenciou que "havia, na extremidade anterior esquerda, marcas de fricção de sujidades, similar a um solado, na parte interna e mediana do portão, assim como a presença de um cavalete próximo ao muro e ao portão, na parte interna do lote, indicando que pelo menos uma pessoa tenha se apoiado no portão, no muro e no cavalete para escalar para fora do lote e/ou para dentro do lote, nas adjacências da mureta retrodescrita do referido muro (fotografias nº 06 a 10)." [grifou-se].

No mais, não há que se falar em aplicação do princípio da insignificância, haja vista não estão presentes requisitos objetivos e subjetivos exigidos para sua aplicação. E, por mais que não se faça necessário a aferição do montante subtraído, sabe-se que o quantum

subtraído não é de pequeno valor (R$ 609,32 – seiscentos e nove reais e trinta e dois centavos) – fls. 39/40.

Assim, resta demonstrada a ausência de qualquer equívoco por parte do magistrado de primeiro grau, pois, se utilizou de fundamentação precisa, advinda da apreciação dos fatos ocorridos, da legislação e da jurisprudência aplicada ao deslinde do caso motivo pelo qual deve ser mantida a r. sentença condenatória pelos seus próprios fundamentos.

III – Conclusão:

Pelo exposto, esta Procuradoria de Justiça manifesta-se pelo conhecimento e desprovimento da apelação defensiva.

Brasília, XX de XXXXXX de 2017.

(OMITIDO)

Procurador de Justiça

C.2 Exemplo de parecer em *Habeas Corpus*

HABEAS CORPUS Nº (OMITIDO)

1ª TURMA CRIMINAL

IMPETRANTE: (OMITIDO)

PACIENTE: OMITIDO

RELATOR: Exm(a). Des(a). Ana Maria Duarte Amarante Brito

PARECER Nº (OMITIDO)

3ª PROCURADORIA DE JUSTIÇA CRIMINAL ESPECIALIZADA

I - Relatório

Trata-se de habeas corpus, com pedido liminar, impetrado em favor de (OMITIDO), contra decisão do Juízo da Segunda Vara Criminal da Circunscrição Judiciária de Taguatinga-DF, que indeferiu o pedido de revogação da prisão preventiva, formulado em favor de paciente.

O impetrante sustenta, em síntese, que a decisão em ataque carece de fundamentação legal, pois não demonstra que, na espécie, estejam presentes quaisquer dos requisitos constantes do artigo 312 do Código de Processo Penal.

Assevera que não há nos autos quaisquer indícios de circunstâncias que autorizem a decretação da prisão preventiva.

Discorre sobre a tipicidade dos fatos, alegando a atipicidade do delito, vez que não houve a subtração de valores consideráveis e, portanto, o crime imputado ao paciente sequer seria roubo.

Afirma, ainda, violação ao princípio constitucional da individualização da pena. Ressalta que o paciente se encontra preso a mais de 90 (noventa) dias, sendo que nem ele nem seus defensores deram motivos para que a data de seu julgamento fosse retardada.

Por fim, aduz que o ora paciente ostenta todas as condições pessoais para responder ao processo em liberdade, eis que possui residência fixa, trabalho e ainda se compromete a comparecer perante o juízo, quando solicitado, e a não mudar de residência sem prévia permissão da autoridade judiciária.

Requer, assim, a concessão de liminar, para que o paciente seja imediatamente posto em liberdade; no mérito, requer a confirmação da medida, com a concessão da ordem em definitivo. A inicial está instruída com os documentos de fls. 14/35. O pedido liminar foi indeferido às fls. 39/42; Informações à fl. 45.

Os autos vieram ao Ministério Público para parecer.

É, em síntese, o breve relato dos fatos.

PARECER

II - Conhecimento

Os requisitos exigidos para o processamento do presente habeas corpus estão presentes, assim, esta Procuradoria de Justiça manifesta-se pelo seu conhecimento.

III - Mérito

Consta dos autos que o Paciente foi preso em flagrante pela suposta prática dos crimes previstos no art. 157, § 1º e § 2º, inciso II, do Código Penal e art. 306, da Lei 9503/97.

O impetrante se insurge contra a r. decisão, sustentando que o paciente reúne todos os requisitos para responder ao processo em liberdade, uma vez que inexistem motivos legais para a manutenção da segregação cautelar.

Na espécie, e ao contrário do que sustenta o impetrante, não há constrangimento ilegal a ser sanada pela via do habeas corpus, tanto porque há indícios suficientes de autoria e prova certa da materialidade, como porque a r. decisão está fundamentada na persistência do motivo que ensejou a prisão preventiva, qual seja, a necessidade de se garantir a ordem pública, atestada pelas circunstâncias concretas em que a conduta delituosa foi praticada, aliada à inexistência de fato novo capaz de desconstituir a necessidade da segregação.

A necessidade da custódia cautelar ficou devidamente fundamentada na decisão de fls. 22/23. Vejamos:

(. . .)

Por meio da análise das peças que instruem a comunicação da prisão em flagrante, constata-se a materialidade do delito, bem como a existência de indícios de que o indiciado seja, em tese, o autor das condutas a ele imputadas, conforme declarações do condutor, da testemunha e da vítima. O modus operandi adotado na execução do delito retrata, in concreto, a periculosidade dos autuados. Segundo consta, os autores do fato abordaram as

vítimas no posto, em concurso de agentes, com emprego de graves ameaças e de violência por empurrões. O fato é grave e a prisão se mostra necessária.

(...)

Sobre a outra tese do Impetrante, referente ao suposto excesso de prazo, imperioso destacar o seguinte entendimento o E. TJDFT,

HABEAS CORPUS. PENAL E PROCESSUAL PENAL. ROUBO CIRCUNSTANCIADO. FUNDAMENTOS DA PRISÃO PREVENTIVA. EXCESSO DE PRAZO NA FORMAÇÃO DA CULPA. NÃO OCORRÊNCIA. PRINCÍPIO DA RAZOABILIDADE. CONSTRANGIMENTO ILEGAL INEXISTENTE.

1. Segundo a orientação jurisprudencial dominante, o prazo para a formação da culpa não deve ser observado a partir de regra aritmética rígida, devendo ter como norte o princípio da razoabilidade.

2. No caso, tendo em vista que a tramitação do feito segue curso condizente com os limites estabelecidos pela Instrução Normativa nº 01/2011, da Corregedoria de Justiça do Distrito Federal, e com suas peculiaridades, não há se falar em excesso injustificado ou desarrazoado na tramitação.

3. Ordem denegada.

(Acórdão n.990416, 20160020490827HBC, Relator: JESUINO RISSATO 3ª TURMA CRIMINAL, Data de Julgamento: 26/01/2017, Publicado no DJE: 01/02/2017. Pág.: 330/350) (grifou-se)

Verifica-se que, in casu, não há nenhum elemento que aponte para uma demora injustificada do processo, eis que ele se desenvolve em sua perfeita normalidade e no prazo condizente com as suas peculiaridades.

Quanto às alegações sobre a atipicidade da conduta delitiva, é de se ponderar que esta é tese de mérito, cuja análise em sede de habeas corpus mostra-se inviável por não haver, nos autos, elementos suficientes para se proceder a tal apuração.

De outra parte, vale lembrar que as condições pessoais favoráveis não garantem ao paciente o direito líquido e certo à liberdade provisória, quando presentes quaisquer dos requisitos previstos nos artigos 312 e 313 do Código de Processo Penal. In verbis:

HABEAS CORPUS. PRISÃO EM FLAGRANTE. ROUBO QUALIFICADO. CONCURSO DE PESSOAS. USO ARMA DE FOGO. PRESENÇA DOS REQUISITOS DA PREVENTIVA. BONS ANTECEDENTES. GRAVIDADE CONCRETA. GARANTIA DA ORDEM PÚBLICA. ORDEM DENEGADA.

I - Preenchidos os requisitos elencados nos artigos 312 e 313 do Código de Processo Penal, a prisão preventiva é medida que se impõe, principalmente quando a liberdade do paciente representa periculosidade para a garantia da ordem pública dadas a violência e a gravidade em concreto de sua conduta.

II - As condições pessoais favoráveis do paciente, como bons antecedentes e domicílio certo, não bastam para afastar a custódia cautelar quando evidenciada a gravidade concreta da conduta a ele imputada, demandando medida efetiva para garantia da ordem pública.

III. As circunstâncias em que praticados os delitos demonstram que as medidas cautelares do art. 319 do CPP são inadequadas.

IV - Ordem denegada. (Acórdão n.941426, 20160020105312HBC, Relator: WALDIR LEÔNCIO LOPES JÚNIOR, 3ª Turma Criminal, Data de Julgamento: 12/05/2016, Publicado no DJE: 18/05/2016. Pág.: 177) (grifou-se)

Ressalte-se, por fim, que não há falar em qualquer violação a princípios constitucionais, em especial ao da individualização da pena, porquanto a prisão preventiva não é pena em si, mas apenas medida que visa garantir a ordem pública e o escorreito processamento do Feito.

Assim, sendo a prisão do paciente cabível e necessária, impossível se falar em coação ilegal sanável por Habeas Corpus.

IV - Conclusão

Ante o exposto, esta Procuradoria de Justiça manifesta-se pelo conhecimento e denegação da ordem.

Brasília, XX de XXXXXXX de 2017.

(OMITIDO)

PROCURADORA DE JUSTIÇA