Instituto Serzedello Corrêa
Escola Superior do Tribunal de Contas da União

ALESSANDRA DE QUEIROZ REQUENA GARRIDO

DESENVOLVIMENTO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO RANQUEADO POR RELEVÂNCIA COM EXPANSÃO DE CONSULTA

Brasília

2020

ALESSANDRA DE QUEIROZ REQUENA GARRIDO

DESENVOLVIMENTO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO RANQUEADO POR RELEVÂNCIA COM EXPANSÃO DE CONSULTA

Trabalho de conclusão do curso de pós-graduação *lato sensu* em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.
Orientador: Prof. Dr. Thiago de Paulo Faleiros

Brasília

2020

REFERÊNCIA BIBLIOGRÁFICA

REQUENA, Alessandra. **Desenvolvimento de um modelo de recuperação da informação ranqueado por relevância com expansão de consulta**. 2020. Trabalho de Conclusão de Curso (Especialização em Análise de Dados para o Controle) – Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Brasília DF. 121 fl.

CESSÃO DE DIREITOS

NOME DO AUTOR: Alessandra Requena
TÍTULO: Desenvolvimento de um modelo de recuperação da informação ranqueado por relevância com expansão de consulta
GRAU/ANO: Especialista/2020

_____

Alessandra Requena
alessandra.requena@gmail.com

Ficha catalográfica

Requena, Alessandra
Desenvolvimento de um modelo de recuperação da informação ranqueado por relevância com expansão de consulta/Alessandra Requena; orientador, Thiago de Paulo Faleiros, 2020.
133 p.

Trabalho de Conclusão de Curso (especialização) - Escola Superior do Tribunal de Contas da União, Instituto Serzedello Corrêa, Especialização em Análise de Dados para o controle, Brasília, 2020.

Inclui referências.

1. Administração. 2. Recuperação da Informação. 3. Modelo Ranqueado. 4. Expansão de Consulta. I. Faleiros, Thiago de Paulo. II. Escola Superior do Tribunal de Contas da União. Especialização em Análise de Dados para o Controle. III. Título.

ALESSANDRA DE QUEIROZ REQUENA GARRIDO

DESENVOLVIMENTO DE UM MODELO DE RECUPERAÇÃO DA INFORMAÇÃO RANQUEADO POR RELEVÂNCIA COM EXPANSÃO DE CONSULTA

Trabalho de conclusão do curso de pós-graduação lato sensu em Análise de Dados para o Controle realizado pela Escola Superior do Tribunal de Contas da União como requisito para a obtenção do título de especialista.

Brasília, 23 de março de 2020

Banca Examinadora:

_____

Prof.ª Thiago de Paulo Faleiros, Dr.

Orientador

Universidade de Brasília

_____

Prof. Saul Campos Berardo, Ms.

Universidade Federal do Pará

Mãe: uma *stop word*.

RESUMO

O objetivo do trabalho foi desenvolver modelos ranqueados de recuperação da informação de complexidade crescente de forma incremental e guiados por métricas. O primeiro modelo desenvolvido foi baseado no modelo probabilístico BM25. O segundo modelo evoluiu o anterior ao implementar expansão de consultas por meio de vetores de palavras gerados automaticamente. O terceiro modelo incorporou ao segundo a extração e indexação de ideias do texto a fim de melhorar a recuperação. Por fim, o quarto modelo teve por objetivo exercitar conceitos de *re-ranking* e *learn to rank*.

Palavras-chave: Recuperação da informação. Modelo ranqueado. Modelo probabilístico. BM25. Expansão de consulta. *Word Embeddings*. Word2Vec. Solr.

ABSTRACT

The goal of the current work was to develop information retrieval ranked models of increasing complexity incrementally and guided by metrics. The first model developed was based on the BM25 probabilistic model. The second model evolved the previous one by implementing query expansion through automatically generated word vectors. The third model incorporated the extraction and indexing of ideas from the text in order to improve the recovery. Finally, the fourth model aimed to exercise concepts of re-ranking and learn to rank.

Keywords: Information retrieval. Ranked model. Probabilistic model. BM25. Query expansion. Word Embeddings. Word2Vec. Solr.

SUMÁRIO

1 INTRODUÇÃO

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

1.1 PESQUISA DE DOCUMENTOS NO TRIBUNAL DE CONTAS DA UNIÃO

O Tribunal de Contas da União é um órgão do poder legislativo federal que tem como missão apoiar o Congresso Nacional no acompanhamento da execução orçamentária e financeira do país. Ele é responsável pela fiscalização contábil, financeira, orçamentária, operacional e patrimonial dos órgãos e entidades públicas do país quanto à legalidade, legitimidade e economicidade.

Seu corpo técnico, os Auditores Federais de Controle Externo, trabalham com diversas informações durante suas jornadas de auditorias e fiscalizações. São informações corporativas que estão disponíveis em bancos de dados relacionais e base de documentos eletrônicos, além de dados obtidos por compartilhamento de informações de outros órgãos da Administração Pública, para citar as principais.

Usualmente, faz parte das jornadas de auditoria e fiscalização pesquisas em documentos, sejam eles corporativos ou não. Ser capaz de pesquisar informações é uma grande necessidade do trabalho moderno. A capacidade de localizar informações em ambiente virtual com rapidez e eficiência é fator diferencial de qualidade e produtividade nas atividades realizadas por órgãos e entidades da Administração Pública.

Diante do contexto de alto volume de informações, sistemas de recuperação de informações ganham relevância fundamental no trabalho do auditor ao fornecer a capacidade de localizar informações em grandes bases de dados de forma rápida. O TCU disponibiliza diversas bases textuais para consulta, seja para os servidores da casa como também para o cidadão. Assim, o cidadão pode, por exemplo, consultar o inteiro teor dos acórdãos do TCU ou localizar qualquer ato de pessoal (posse, aposentadoria, exoneração etc.) da Administração Pública Federal.

São atualmente dezesseis bases de documentos disponíveis para consulta em uma plataforma de recuperação da informação. Algumas bases são restritas e necessitam de perfil de

acesso e outras contém informações públicas e estão disponíveis para consulta por qualquer cidadão. Novas bases vêm sendo incorporadas ao longo do tempo. Assim, os dados contidos na tabela 1.1 é uma fotografia do que está disponível em outubro de 2019.

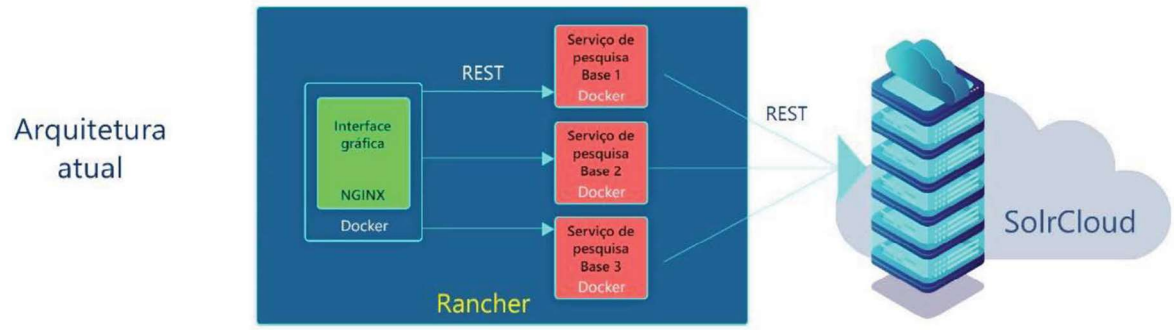
Tabela 1.1 – Bases de pesquisa na Plataforma de Pesquisa Textual do TCU (situação em 31/10/2019)

Base e Privacidade	Quantidade de documentos	Origem dos dados	Quantidade média aproximada de acessos
Acórdãos do TCU (Pública)	342.875	Dados relacionais e documentos eletrônicos	82520
Atas das sessões colegiadas (Pública)	3183	Documentos eletrônicos	3053
Atos normativos da presidência do TCU (Pública)	3165	Dados relacionais e documentos eletrônicos	7039
Atos de pessoal da Administração Pública Federal (Pública)	4.022.705	Dados relacionais	5783
BTCU (Parte pública e parte privada)	18.422	Dados relacionais e documentos eletrônicos	3077
CNPJ (Privada)	39.617.440	Dados relacionais	141
CPF (Privada)	242.302.573	Dados relacionais	179
Jurisprudência Selecionada (Pública)	14.021	Dados relacionais e documentos eletrônicos	41609
Ouvidoria (Privada)	48.350	Dados relacionais	Não disponível
Pareceres da Conjur (Privada)	23.061	Dados relacionais e documentos eletrônicos	158
Pareceres do MPTCU (Privada)	180.470	Dados relacionais e documentos eletrônicos	24
Processo (Parte pública e parte privada)	1.555.349	Dados relacionais e documentos eletrônicos	9800 (dados de acesso à parte pública)
Publicações de Pessoal, Jurisprudência e Licitações e Contratos (Pública)	5.605	Dados relacionais e documentos eletrônicos	3560

Questões de ordem (Pública)	84	Dados relacionais e documentos eletrônicos	3842
Súmulas (Pública)	289	Dados relacionais e documentos eletrônicos	3346
União (Privada)	6.450	Dados relacionais e documentos eletrônicos	Não disponível

A arquitetura da plataforma de pesquisa textual é baseada em microsserviços, sendo um microsserviço por base textual, que são implantados em diferentes containers Docker, geridos em um ambiente de nuvem privada. As bases de dados textuais estão indexadas na ferramenta Solr em modo SolrCloud (solução distribuída para a indexação e busca textual) (TRIBUNAL DE CONTAS DA UNIÃO, 2018).

Figura 1.1 – Arquitetura da Plataforma de Pesquisa Textual do TCU



Fonte: Revista Reconhe-Ser 2018 (TRIBUNAL DE CONTAS DA UNIÃO, 2018)

A plataforma de pesquisa textual é o sistema mais acessado do TCU em número de requisições, segundo dados levantados pelo Serviço de Infraestrutura de Aplicações - Sinap, o que sugere indiretamente a importância desta ferramenta para o trabalho desempenhado pelo TCU e para a sociedade. A plataforma ganhou o prêmio de trabalho de destaque no Reconhe-Ser 2018 e a pesquisa de jurisprudência ganhou o prêmio de trabalho inovador de controle externo em 2017 (TRIBUNAL DE CONTAS DA UNIÃO, 2018)

1.2 BREVE HISTÓRICO DA PESQUISA TEXTUAL NO TCU

O primeiro sistema de recuperação de informação utilizado no TCU foi o IBM *Storage and Information Retrieval System* - STAIRS. Este software provia armazenamento e busca livre em dados textuais em um momento em que todos os sistemas de informação do TCU eram baseados em grande porte. As consultas do STAIRS eram expressões booleanas dos termos desejados. Além dos operadores booleanos AND, OR e NOT, o STAIRS reconhecia ainda operadores específicos de adjacência de termos e termos em um mesmo parágrafo. Os documentos de texto poderiam ainda conter campos formatados com informações fixas para serem usados em seleções específicas como pesquisa por data ou situação. Outras funcionalidades do sistema eram salvamento e refinamento de consultas e *highlight*.

Com o *downsizing*, em 1998, foi adquirida em uma licitação uma ferramenta chamada BRS/Search para prover a pesquisa textual nas três fontes de informações do TCU de grande interesse público: Processos, Atos de Pessoal e Jurisprudência.

O BRS/Search é um banco de dados *full-text* e um sistema de recuperação da informação. Ele utiliza um sistema de índice invertido para armazenar, localizar e recuperar dados não estruturados. É uma ferramenta baseada no modelo booleano de recuperação da informação e oferece, além dos operadores booleanos (E, OU, NÃO), operadores de proximidade de texto (ADJ, PROX, MESMO, COM).

Para entender o contexto histórico do momento de implantação do BRS/Search no TCU e a motivação da aquisição desta ferramenta, é interessante citar que MANNING et al. (2008) afirmam em seu livro que sistemas implementando o modelo booleano de recuperação da informação eram a principal ou até mesmo a única opção para soluções de pesquisa providas por grandes empresas comerciais por três décadas até o início de 1990 (aproximadamente a data do início da *World Wide Web*). E de fato o produto BRS/Search era forte no mercado brasileiro e entrou como solução de pesquisa em vários órgãos públicos, a citar Banco Central do Brasil, Superior Tribunal de Justiça e Supremo Tribunal Federal. Em alguns órgãos, por exemplo o STF, é a solução de pesquisa textual utilizada até hoje.

Ainda segundo MANNING et al. (2008), esses sistemas não ofereciam somente os operadores booleanos básicos (AND, OR e NOT). Uma expressão booleana pura sobre termos

com um resultado não ordenado é limitada para muitas das necessidades de informação que as pessoas possuem e esses sistemas implementavam modelos de recuperação de informação booleanos estendidos ao incorporar operadores adicionais, como os operadores de proximidade. Um operador de proximidade é uma maneira de especificar se dois termos em uma consulta devem ocorrer perto um do outro em um documento, onde a proximidade pode ser medida pelo limite de palavras intermediárias ou por uma referência a uma unidade estrutural do texto como uma sentença ou um parágrafo.

Em meados dos anos 90, com o advento da Internet e de pesquisas acadêmicas que apontavam vantagens nos modelos de recuperação ranqueados, outras soluções de pesquisa em dados não estruturados surgiram (MANNING et al., 2008). O uso do BRS/Search pelas organizações foi diminuindo até que a parceira comercial brasileira do produto parou de oferecer suporte técnico à solução. As empresas que ainda utilizavam o BRS/Search, como o TCU, tiveram que arcar com o risco de falta de suporte e manutenção.

A substituição do BRS/Search no TCU foi motivada não somente pela falta de suporte técnico, mas também por um desejo de modernizar a solução de pesquisa do TCU para pesquisas de texto livre, também chamadas de “Google like”. Ou seja, bastaria o usuário entrar com um ou mais termos de busca ao invés de utilizar uma linguagem precisa com operadores para construir expressões e o sistema decide quais os documentos satisfazem melhor a consulta. Foi nesse contexto que surgiu o projeto de migração da pesquisa textual para o Solr em 2012. O Solr é uma plataforma de código aberto de pesquisa textual baseada no Lucene que, na época, era a ferramenta mais popular de recuperação da informação que estava sendo adotada por grandes organizações. A nova solução de pesquisa criada pela área de TI do TCU era mais moderna, oferecia pesquisas multifacetadas e resultados ranqueados por relevância. Oferecia um modelo de recuperação da informação ranqueado, não oferecendo mais os operadores de proximidade. No entanto, esta pesquisa não agradou o usuário.

A grande queixa era que localizar as informações desejadas era muito difícil. Havia um sentimento de insatisfação generalizado com a nova ferramenta e, assim, não foi possível desativar a ferramenta anterior pois muitos usuários ainda a utilizavam. Os usuários sentiam falta dos operadores de proximidade e diziam que a relevância programada na nova ferramenta não correspondia às suas necessidades. Para agravar ainda mais o problema, a nova pesquisa misturava em um único resultado documentos vindos de várias bases textuais diferentes e não

oferecia nenhum tipo de *feedback* para o usuário no resultado da pesquisa (por exemplo, fragmentos do documento recuperado), tornando a localização da informação uma tarefa ainda mais difícil.

Até o ano de 2016, a área de TI não compreendia bem porque a nova solução de recuperação da informação, mais moderna e alinhada às pesquisas acadêmicas da área, não atendia o usuário e acreditava-se que era uma questão de tempo para o usuário se adaptar, o que não ocorreu.

Segundo MANNING et al. (2008), muitos usuários, principalmente profissionais de áreas técnicas, preferem o modelo booleano de recuperação da informação. As consultas booleanas são precisas: um documento corresponde à consulta ou não. Isso oferece um grande controle e transparência em relação ao que está sendo recuperado. Além disso, os operadores de proximidade aumentam o poder nas mãos do usuário ao construir uma expressão de busca que, em certos domínios de negócio, permitem realizar buscas muito precisas em extensas bases de documentos. E, em alguns domínios, a citar a área legal, o ranqueamento por ordem cronológica inversa (do documento mais recente para o mais antigo) é muito efetivo e atende à expectativa do usuário.

Além desta questão da precisão que o modelo booleano oferece e que era muito atrativo e importante ao usuário típico da solução do TCU, os demais tribunais superiores, a citar STJ e STF, oferecem até hoje em suas pesquisas de jurisprudência um modelo baseado em operadores booleanos e de proximidade.

Assim, em 2016, com um melhor entendimento das necessidades do usuário, o TCU refez solução de pesquisa de jurisprudência. Implementou a lógica de proximidade como uma extensão da plataforma Solr, permitindo ao usuário realizar pesquisas utilizando os operadores de proximidade da mesma forma que ele utiliza ao pesquisar acórdãos no STJ ou STF. Alterou também o ordenamento padrão do resultado da consulta para ordem cronológica decrescente e modernizou a interface do usuário e das arquiteturas tecnológicas.

A nova solução de pesquisa foi bem aceita pelos interessados, o que foi ratificado pelo resultado de pesquisa de satisfação conduzida pela unidade organizacional gestora da pesquisa de jurisprudência no TCU, a Secretaria das Sessões (Seses). Conforme a pesquisa, que foi respondida por 33,7% dos servidores ativos do TCU, 73,48% dos respondentes julgam a pesquisa como “boa” ou “muito boa” (o resultado completo da pesquisa pode ser obtido com a Seses).

Esta nova solução tornou-se o padrão sobre o qual a plataforma de pesquisa textual em documentos do TCU foi construída.

Desde então, tem havido muito investimento na evolução da pesquisa textual:

1. Experiência do usuário: estudos e monitorização do uso da ferramenta de pesquisa que levam a uma construção de interface mais alinhada às necessidades e uso do usuário.
2. Modernização da arquitetura tecnológica: uso de microsserviços dentro de uma macroarquitetura tecnológica, com uso de Docker e uso de nuvem privada.
3. Integração das bases textuais: aumento o número de bases textuais em que é possível pesquisar em várias bases simultaneamente dentro de uma plataforma.

1.3 MOTIVAÇÃO

Conforme exposto, muito se avançou em relação à criação de uma nova linguagem de pesquisa (operadores de proximidade), à interface do usuário, às arquiteturas tecnológicas da pesquisa textual e à internalização de novas bases.

No entanto, ainda não houve investimentos em relação à adoção de um modelo de recuperação ranqueado. Apesar do modelo booleano estendido atender a uma parcela dos usuários da solução que sabem utilizar bem os operadores de busca, há usuários ainda insatisfeitos com a qualidade dos resultados. A pesquisa de satisfação conduzida pela Seses demonstra isso pelos comentários recebidos como *"encontro pelo Google com mais facilidade"*, *"sinceramente, é muito mais fácil pesquisar no Google sobre jurisprudência do TCU"* e *"o antigo critério relevância para pesquisar decisões era bem útil".* Além disso, aproximadamente 33% dos respondentes do levantamento julgam que os operadores é a característica que menos apreciam na pesquisa.

As estatísticas de uso da pesquisa textual do TCU mostram que grande parte das consultas não utilizam operadores, que sugere que a construção de consultas com operadores é inconveniente para a maior parte dos usuários. Comentários do tipo *"Muito confuso para estruturar uma frase com operadores. Geralmente desisto nas primeiras tentativas"*, *"Não sei utilizar os operadores de pesquisa"* e *"O uso de operadores lógicos não é tão simples quanto*

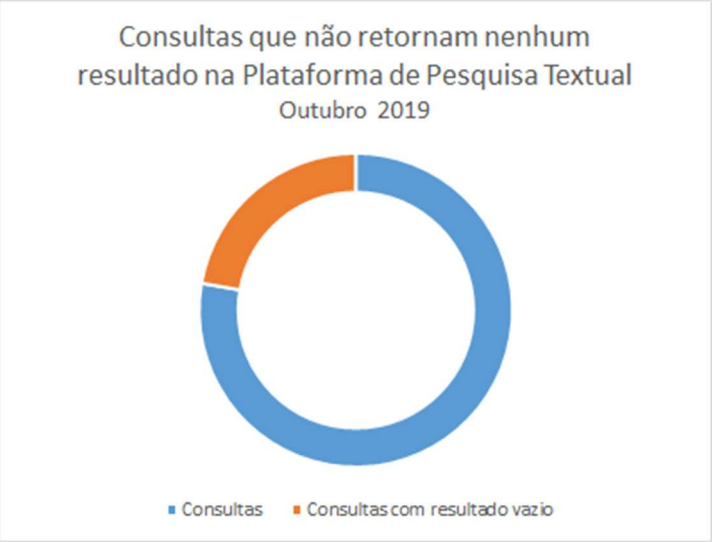
vendem. A maioria não sabe usar os recursos de forma efetiva" também foram recebidos no levantamento da Seses.

Além disso, verifica-se que, associado ao uso de operadores, está o uso de caracteres curinga, como o ?, $ e *. Estes caracteres são usados para substituir o término dos termos para tratamento das variações como conjugações verbais e plurais. É uma estratégia utilizada por usuários avançados para aumentar a revocação da pesquisa. No entanto, a maioria não utiliza *wildcards* e alguns comentários no levantamento da Seses sugerem problemas relacionados à baixa revocação, tal como *"Por vezes a busca na pesquisa de jurisprudência não acha nenhum resultado e a busca no Google com o mesmo texto leva a um acórdão do TCU"*.

Figura 1.2 – Uso dos operadores de proximidade na Plataforma de Pesquisa Textual (todas as bases) em outubro de 2019. Consultas com operadores representam aproximadamente 0,5% do total de consultas.

Fonte: Log de acessos da Plataforma de Pesquisa Textual do TCU.

Figura 1.3 – Consultas que não retornam resultado sugerem problemas relacionados à baixa revocação. Na Plataforma de Pesquisa Textual (todas as bases), aproximadamente 29% das pesquisas retornam nenhum resultado.



Fonte: Log de acessos da Plataforma de Pesquisa Textual do TCU.

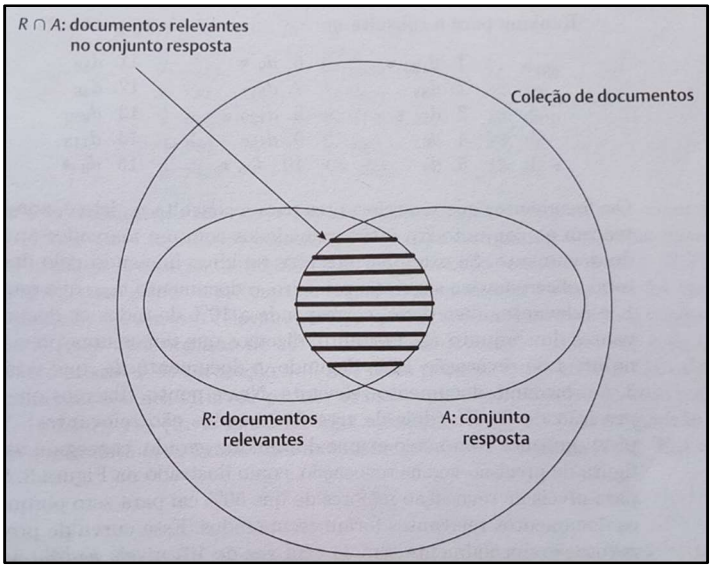
Há também um problema relacionado à ordenação dos resultados de pesquisa em ordem cronológica decrescente, que nem sempre atende a necessidade de informação do usuário. Um exemplo é a busca por "Regimento Interno do TCU", em que a necessidade de informação do usuário é clara. No entanto, o sistema de busca retorna a Resolução 155/2002 - Regimento Interno do Tribunal de Contas da União - como o último documento da lista de resultados.

Assim, as necessidades de informação atendidas pela grande precisão do modelo booleano estendido, como é o caso de pesquisa por jurisprudência em que é necessário encontrar todos os acórdãos sobre determinado assunto, são satisfeitas se o usuário especificar exaustivamente as várias possibilidades em uma complexa expressão de busca. Estas consultas são geralmente realizadas por usuários especialistas que representam somente uma parcela dos clientes da busca. Já as pesquisas chamadas livres, ou seja, sem uso de operadores, são a maioria e nem sempre retornam os melhores resultados, seja por causa da ordenação inadequada por documentos mais recentes ou por uma baixa revocação.

1.4 DEFINIÇÃO DO PROBLEMA

No modelo booleano de recuperação da informação, um documento é relevante ou não relevante. Não há satisfação parcial das condições de consulta. Esse critério binário de decisão, sem nenhuma noção de grau, pode impedir a boa qualidade na recuperação. Além disso, o modelo booleano possui um problema genérico: utilizar o operador AND tende a produzir uma grande precisão, mas uma baixa revocação, enquanto usar o OR oferece uma baixa precisão mas uma grande revocação. É difícil ou impossível achar um ponto intermediário nesta fórmula.

Figura 1.4 – Precisão e Revocação para uma dada requisição de informação



Fonte: BAEZA-YATES et al., 2011

Já o modelo de recuperação ranqueado por relevância sugere um método efetivo para ranquear ou ordenar os resultados da recuperação por meio de uma pontuação para o documento que encapsula o quão bom aquele documento representa uma dada consulta.

A calibração da relevância é um problema muito difícil, normalmente mal compreendido e difícil de detectar quando não está funcionando bem. Geralmente requer que vários exemplos ruins sejam identificados para que padrões problemáticos emerjam e é um desafio saber quais seriam os melhores resultados para a pesquisa. Infelizmente, é somente após a entrada em produção que a organização consegue identificar a grande lacuna entre a relevância

padrão da ferramenta e uma personalizada para o domínio do negócio, o que ocorreu no TCU na primeira tentativa de modernização da ferramenta de busca.

Assim, a calibração de relevância depende do domínio de negócio e cada aplicação tem expectativas de relevância diferentes. Buscas consideradas “simples e fáceis” requerem muito esforço de engenharia por trás e os usuário tem uma alta expectativa em relação aos sistemas de busca.

A relevância também está relacionada à capacidade da ferramenta de busca de identificar relações semânticas no texto. Ou seja, um documento pode ser relevante para uma consulta mesmo que não haja correspondência exata dos termos da consulta e dos termos indexados. E por isso a revocação não pode ser subestimada.

Há dois problemas relacionados à correspondência exata de termos da consulta nos documentos: polissemia e sinonímia. A polissemia é quando palavras possuem mais de um significado e a sinonímia define-se por palavras diferentes terem o mesmo significado. Ao ser capaz de identificar essas relações semânticas no texto, o motor de busca consegue melhorar a revocação. Assim, expansão dos termos por sinônimos e representações alternativas de consulta podem melhorar o conjunto resposta da busca, tornando mais provável que documentos semanticamente similares sejam localizados. Em um mundo ideal, a máquina de busca deve ir além de simples identificação de termos em documentos para tentar identificar a real necessidade de informação do usuário.

A implementação de busca semântica é uma alternativa para aumentar a revocação do conjunto resposta, o que eleva a possibilidade de localizar a informação. Um bom ranqueamento por relevância vem, por fim, ordenar os resultados conforme as expectativas do usuário, atendendo de forma ágil suas necessidades de informação ao aumentar a precisão em um conjunto resposta com alta revocação.

Assim, acredita-se que para evoluir positivamente a plataforma de pesquisa, os próximos passos deveriam ser no sentido da implantação de um modelo de recuperação ranqueado por relevância e implementação de busca semântica.

1.5 JUSTIFICATIVA DO TRABALHO

Para subsidiar o trabalho de controle externo, uma ferramenta fundamental para os auditores é a pesquisa em documentos corporativos do TCU (acórdãos, jurisprudência, atos normativos, dentre outros). A pesquisa atual realiza a busca pela presença dos termos da consulta nos documentos e o resultado é ordenado por data do documento. A ordenação por data do documento é muitas vezes inadequada pois não apresenta os documentos mais relevantes no início da lista de documentos localizados, forçando o usuário a paginar o resultado da pesquisa até localizar o documento desejado. Além disso, documentos que são relevantes para a consulta mas não apresentam os termos exatos que o usuário digitou não são recuperados.

Assim, investir na evolução do motor de busca para identificar relações semânticas no texto e oferecer resultados ranqueados ao usuário é dar um salto de qualidade dos resultados que a plataforma de pesquisa textual oferece. Com a forte tendência de internalização de novas bases, esta evolução torna-se fundamental para agregar ainda mais valor para o usuário e pode representar um aumento de produtividade ao reduzir o tempo para localizar a informação. Com os resultados positivos deste projeto para uma base de documentos, os conceitos poderão ser aplicados em todas as demais bases de pesquisa.

1.6 OBJETIVO GERAL

Desenvolver um modelo de recuperação da informação ranqueado por relevância com identificação de relações semânticas de texto em uma coleção de documentos específica do Tribunal de Contas da União que aumente a precisão dos resultados da pesquisa.

1.7 OBJETIVOS ESPECÍFICOS

- Remodelar a base textual de uma coleção de documentos para atender as necessidades de recuperação de informação dos usuários.
- Implementar um modelo de recuperação ranqueado para uma coleção de documentos específica.

- Implementar a expansão semântica dos documentos e consultas por meio de geração automática de relações semânticas de texto.

1.8 CONTRIBUIÇÕES

Todo o aprendizado proporcionado por este trabalho sobre como modelar bases textuais para fins de pesquisa, ranqueamento por relevância e expansão semântica poderá ser aplicado nas dezesseis bases textuais atualmente disponíveis na plataforma de pesquisa textual.

Além disso, alguns artefatos de software, como a análise exploratória de logs de pesquisa, o processamento de *word embeddings* e a avaliação dos resultados por meio de métricas, serão produtos de softwares a serem utilizados pelo time de TI do TCU para a evolução das pesquisas textuais.

Este trabalho pode ainda ser usado para formatar um treinamento sobre o tema a ser ministrado internamente no TCU. Há pouco conhecimento do assunto na área de TI, apesar de haver uma tendência de se oferecer buscas embutidas nos sistemas para melhorar a experiência do usuário. Pode-se também expandir a oferta do treinamento para outros órgãos da Administração Pública, uma vez que a unidade responsável pela plataforma de pesquisa textual do TCU - o Serviço de Integração e Métricas de Sistemas (Seint) - vem apoiando alguns desses órgãos na implantação de seus sistemas de busca e nota-se que há uma lacuna de conhecimento do assunto nessas organizações.

2 FUNDAMENTAÇÃO TEÓRICA

2.1 RECUPERAÇÃO DE INFORMAÇÃO

O campo de estudo acadêmico chamado de Recuperação da Informação (RI) trata da **representação**, **armazenamento, organização** e **acesso** a itens de informação, como documentos, páginas Web, catálogos online, registros estruturados e semiestruturados, objetos multimídia etc. A representação e a organização dos itens de informação devem fornecer aos usuários facilidade de acesso às informações de seu interesse (BAEZA-YATES et al. 2011). A RI vem se tornando rapidamente a forma dominante de acesso à informação, ultrapassando as buscas tradicionais em bancos de dados, o que confere a este tema grande importância.

O objetivo principal de um sistema de RI é recuperar todos os documentos que são relevantes à necessidade de informação do usuário e ao mesmo tempo recuperar o menor número possível de documentos irrelevantes. Este é o problema fundamental de RI: a dificuldade é saber não somente como extrair a informação dos documentos, mas também saber como utilizá-la para decidir quanto à sua relevância. Isto é, a noção de **relevância** tem um papel central em RI.

No entanto, a relevância é um julgamento pessoal que depende da tarefa a ser resolvida e o seu contexto. Por exemplo, a relevância pode mudar com o tempo, com o local ou até mesmo com o dispositivo. Nesse sentido, nenhum sistema de RI pode fornecer resposta perfeita a todos os usuários (BAEZA-YATES et al. 2011).

Os usuários de sistemas modernos de RI têm necessidades de informação de diferentes níveis de complexidade. A descrição completa da necessidade do usuário não necessariamente fornece a melhor formulação de consulta para o sistema de RI. Em vez disso, normalmente o usuário traduz essa necessidade de informação em uma consulta ou em uma sequência de consultas a serem submetidas ao sistema. Em sua forma mais comum, essa tradução gera uma série de palavras-chave, ou termos de indexação, que sumarizam a necessidade de informação do usuário. Dada a consulta do usuário, o objetivo maior do sistema de RI é recuperar informações que sejam úteis ou relevantes para o usuário.

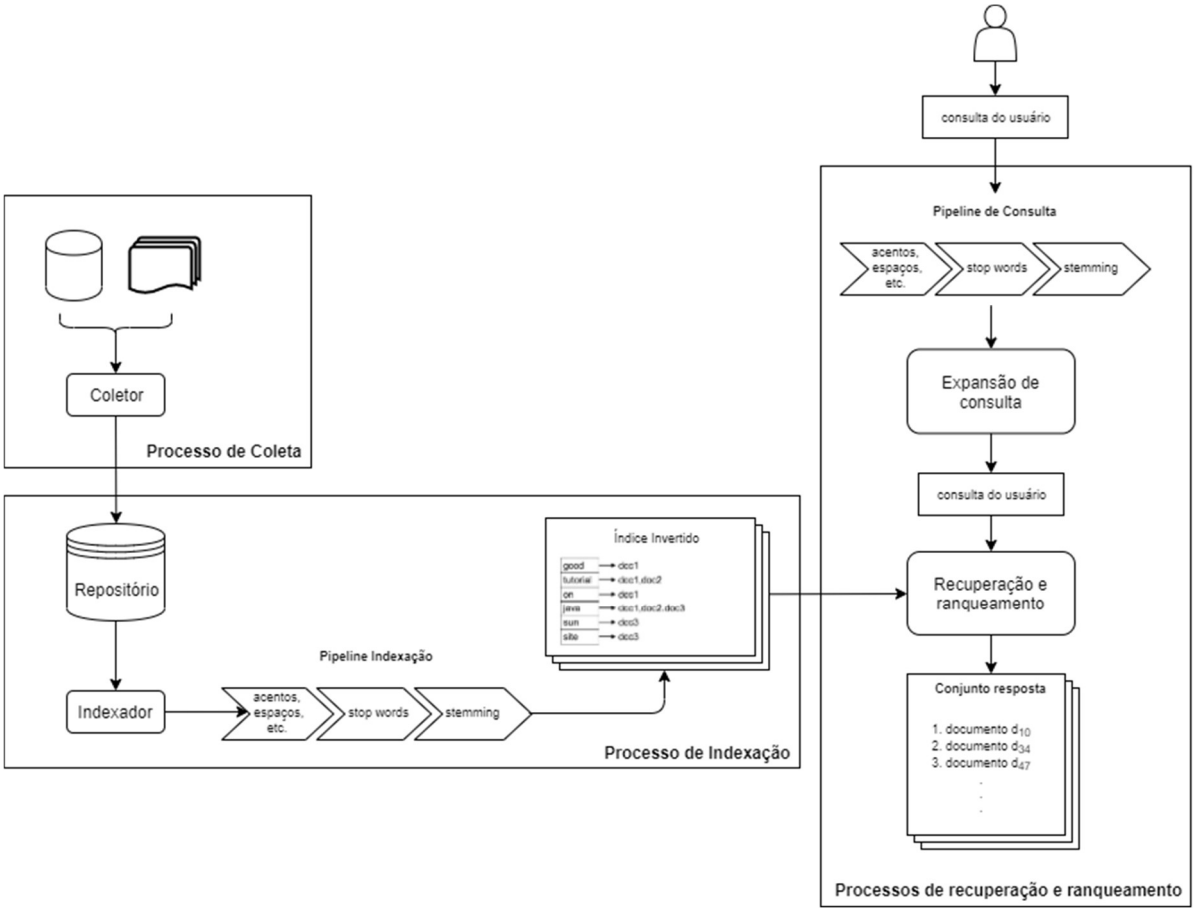
A fim de ser efetivo em sua tentativa de satisfazer a necessidade de informação do usuário, o sistema de RI deve de alguma forma interpretar o conteúdo dos itens de informação,

isto é, dos documentos de uma coleção, e classificá-los de acordo com o grau de relevância à consulta do usuário. Essa interpretação do conteúdo de um documento envolve a extração de informações sintáticas e semânticas do texto do documento e sua utilização para satisfazer a necessidade de informação do usuário.

2.2 ARQUITETURA DE UM SISTEMA DE RI

Uma visão de alto nível da arquitetura de um sistema de RI pode ser vista na figura 2.1 com o intuito de entender como funciona um sistema genérico de recuperação da informação.

Figura 2.1 – Arquitetura de alto nível do software de um sistema de RI



Fonte: BAEZA-YATES et al. 2011 (Adaptada)

Um sistema de RI começa com o processo de coleta da coleção de documentos, que pode ser particular (corporativa) ou de informações públicas (dados disponíveis na Web). A partir da coleta, a indexação processa a coleção de documentos com o objetivo de fornecer estruturas de dados adequadas para que o processo de recuperação e ranqueamento consiga recuperar as informações com performance satisfatória.

Para a indexação, a estrutura de dados mais utilizada é o índice invertido, que, de forma simplificada, é composto por todas as palavras distintas da coleção e para cada palavra a lista de documentos que a contém.

O processo de indexação envolve a aplicação de operações textuais em sequência, tais como *tokenização*, redução do texto para letras minúsculas, eliminação de *stopwords*, radicalização (*stemming*) e a seleção de um subconjunto de termos para serem utilizados como termos de indexação. Os termos de indexação são utilizados para compor a representação do documento, que pode ser menor ou maior do que o documento original.

Uma vez que os documentos estão indexados, o processo de recuperação da informação pode acontecer. Ele consiste em recuperar os documentos que satisfaçam uma consulta do usuário. O usuário informa a consulta e ela é então analisada e modificada por operações semelhantes às aplicadas durante a indexação. A seguir, a consulta pode ser expandida e modificada. A expansão de consultas pode ocorrer de formas diferentes e será detalhada na seção 2.11. Após a transformação da consulta, ela é utilizada para a recuperação dos documentos.

Em seguida, os documentos recuperados são ranqueados segundo alguma função/modelo de ranqueamento e aqueles que estão no topo do ranking são retornados para o usuário.

O propósito do ranqueamento é identificar os documentos que têm maior probabilidade de serem considerados relevantes para o usuário. Esta é a parte mais crítica de um sistema de RI e para qual há diversos modelos de RI.

2.3 MODELOS DE RI

Para recuperar respostas para uma consulta, qualquer sistema de RI tem de lidar com um problema central: prever quais documentos os usuários irão considerar relevantes e quais irão considerar irrelevantes. O problema é essencialmente difícil. Além disso, existe um grau de incertezas ou de imprecisão devido ao fato de que dois usuários podem discordar sobre o que é e o que não é relevante. A esse respeito, o sistema implementa um algoritmo preditivo que almeja aproximar-se da opinião de uma grande fração dos usuários quanto à relevância dos resultados de uma grande fração de consultas. Esse algoritmo preditivo é essencialmente a função de ranqueamento utilizada para estabelecer um ordenamento dos documentos recuperados.

Assim, a modelagem em RI é um processo complexo que tem o objetivo de produzir uma função de ranqueamento, ou seja, uma função que atribui escores a documentos em relação a uma consulta.

Esse processo pode ser dividido em duas tarefas principais:

1. a concepção de um arcabouço lógico para representar documentos e consultas;
2. a definição de uma função de ranqueamento que computa o grau de similaridade de cada documento em relação à consulta dada.

O arcabouço lógico é normalmente baseado em conjuntos, vetores ou em distribuição de probabilidades. Ele afeta diretamente a computação dos graus de similaridade entre documento e consulta que são utilizados para ordenar os documentos recuperados em resposta à consulta dada.

Diferentes conjuntos de premissas (a respeito da relevância de um documento) produzem modelos de RI diferentes. O modelo de RI adotado determina as predições do que é e do que não é relevante.

Assim, as premissas fundamentais que formam a base de um algoritmo de ranqueamento determinam o modelo de RI. Segundo BAEZA-YATES et al. (2011), um modelo de RI é caracterizado por

$$[D, Q, F, R(q_i, d_j)]$$

Onde:

- D é o conjunto das representações dos documentos;
- Q é o conjunto das representações das consultas (necessidades de informação);
- F é um arcabouço para modelar as representações dos documentos e das necessidades de informação e seus relacionamentos, como conjuntos e relações Booleanas, vetores e operações de álgebra linear, espaços amostrais e distribuições de probabilidade;
- $R(q_i, d_j)$ é uma função de ranqueamento que associa um escore para a relação entre a query $q_i \in Q$ e o documento $d_j \in D$. Este escore é utilizado para ordenar documentos por relevância.

Segundo BAEZA-YATES et al. (2011), os modelos clássicos de RI são **booleano**, **vetorial** e **probabilístico**. Para cada um desses modelos, partindo das representações lógicas de documentos e consultas, um arcabouço lógico é definido, que fornece a intuição para construção de uma função de ranqueamento.

Para o modelo booleano clássico, o arcabouço é baseado na teoria de conjuntos. O modelo vetorial clássico utiliza um espaço vetorial com n dimensões, representações dos documentos e das consultas como vetores de operações de álgebra linear. Já o modelo probabilístico clássico, as distribuições de probabilidades dos termos nos documentos e nas consultas e o teorema de Bayes apoiam a definição do modelo.

No modelo booleano, os termos de indexação não têm peso algum associado; são simplesmente elementos de um conjunto. Nos modelos vetorial e probabilístico, os termos de indexação possuem pesos associados (ponderação de termos) com o objetivo de melhorar a ordenação dos documentos.

Considerando que o propósito principal de um modelo de RI é produzir um conjunto de resultados que provavelmente seja relevante para o usuário, implementações modernas de sistemas de RI incluem características de vários modelos de RI e não de apenas um.

2.4 PONDERAÇÃO DE TERMOS TF-IDF

Para aferir o grau de importância de um termo em relação a um documento, uma abordagem lógica é medir algumas propriedades de ocorrência do termo no documento: a frequência do termo (TF, *Term Frequency*) e a frequência inversa de documento (IDF, *Inverse Document Frequency)*. Essas propriedades são facilmente mensuráveis e formam a base do esquema de ponderação mais popular em RI, chamado TF-IDF.

O TF reflete o quão frequente um termo ocorre em um documento. Assim, o TF atribui a um termo *t* o número de ocorrências do termo no documento *d*.

O IDF informa o quão raro (e consequentemente valioso) é um termo. Sendo DF o número de documentos em que o termo ocorre, o IDF é o inverso dessa medida.

Portanto, a métrica TF-IDF pode ser descrita como:

$$\text{TF-IDF}_{t,d} = \text{TF}_{t,d} \times \text{IDF}_t$$

Em outras palavras, TF-IDF atribui ao termo *t* um peso no documento *d* que:

- é máximo quando *t* ocorre muitas vezes em um pequeno número de documentos (assim conferindo grande poder discriminatório para aqueles documentos);
- é pequeno quando o termo ocorre poucas vezes em um documento ou ocorre em muitos documentos (provendo assim um sinal de relevância menor);
- mínimo quando o termo ocorre virtualmente em todos os documentos.

Esta ponderação é crucial para fazer pontuação e ranqueamento dos documentos em um sistema de RI.

2.5 NORMALIZAÇÃO PELO TAMANHO DOS DOCUMENTOS

Apesar da ponderação TF-IDF ser muito intuitiva e corresponder à intuição de grande parte dos usuários que entendem que termos raros são mais específicos que termos comuns, experimentação na área de IR mostra que o TF-IDF puro não corresponde à intuição de relevância do usuário. Se um termo consultado ocorrer 10 vezes mais em um conteúdo, isso não

torna o documento 10 vezes mais relevante. Mais menções ao termo correlaciona com a relevância, mas a relação não é linear. (BAEZA-YATES et al., 2011) (TURNBULL e BERRYMAN, 2016)

Em uma grande coleção, o tamanho dos documentos pode variar bastante. Isso é problemático, porque documentos mais longos têm mais chances de serem recuperados pelas consultas simplesmente porque eles contêm mais palavras. A fim de compensar esse efeito indesejado, pode-se dividir o número de ordem de cada documento pelo seu tamanho - um procedimento comumente chamado de normalização pelo tamanho do documento. Como essa normalização geralmente leva a um melhor ranqueamento (isto é, um ranqueamento que melhor corresponde à percepção de relevância de um usuário), a normalização pelo tamanho dos documentos é amplamente adotada pelos modelos de RI.

Para reduzir o impacto dos tamanhos dos documentos na ordenação, os melhores sistemas de RI adotam alguma forma de normalização dos documentos em sua função de ranqueamento. Isso é particularmente benéfico em coleções que possuem grandes variações no tamanho dos documentos. (BAEZA-YATES et al., 2011)

2.6 O MODELO VETORIAL

Em recuperação da informação, o texto pode ser representado como vetores em um espaço vetorial comum, sendo as dimensões representadas pelas palavras do texto. Este modelo é conhecido como o modelo vetorial (*vector space model*) e é fundamental para as operações de RI como ranqueamento de documentos em relação a uma consulta, classificação e *clusterização* de documentos. (MANNING et al., 2008)

Sendo cada palavra do texto uma dimensão do vetor, faz-se necessária a atribuição de pesos não-binários para medir a força ou magnitude de cada dimensão (ou *feature*). Os pesos dos termos são utilizados para computar o grau de similaridade entre cada documento armazenado no sistema e a consulta do usuário.

Se a consulta também for vista como um vetor no mesmo espaço vetorial, o modelo vetorial calcula o grau de similaridade do documento *d* em relação à consulta *q* sob forma de

correlação entre os vetores $\overrightarrow{d}$ e $\vec{q}$. Essa correlação pode ser quantificada, por exemplo, pelo produto escalar dos vetores ou o cosseno do ângulo entre esses dois vetores.

Assim, o modelo vetorial ordena os documentos de acordo com o grau de similaridade em relação à consulta. Um documento pode ser recuperado mesmo que satisfaça a consulta apenas parcialmente. Por exemplo, podemos estabelecer um limiar para o grau de similaridade e retornar somente os documentos acima do limiar.

Os pesos no modelo vetorial são basicamente os pesos TF-IDF. Apesar de sua simplicidade, o modelo vetorial consegue bons resultados com coleções genéricas devido ao esquema de ponderação de termos e da adoção da normalização pelo tamanho dos documentos. (BAEZA-YATES et al., 2011)

Segundo BAEZA-YATES et al. (2011), as principais vantagens do modelo vetorial são:

- seu esquema de ponderação de termos melhora a qualidade da recuperação;
- sua estratégia de casamento parcial permite a recuperação de documentos que aproximam as condições da consulta;
- a fórmula do cosseno ordena os documentos de acordo com o seu grau de similaridade em relação à consulta;
- a normalização pelo tamanho do documento está naturalmente embutida no modelo.

Teoricamente, o modelo vetorial tem a desvantagem de que os termos de indexação são considerados mutuamente independentes. Contudo, na prática, considerar as dependências entre termos é desafiador e pode levar a resultados ruins se não for feito de maneira adequada. (BAEZA-YATES et al., 2011)

Apesar de sua simplicidade, o modelo vetorial é uma boa estratégia de ranqueamento para coleções genéricas. Ele fornece resultados ranqueados que dificilmente podem ser melhorados sem o uso de expansão de consultas e realimentação de relevância. Vários métodos alternativos de ranqueamento foram comparados ao modelo vetorial e o consenso parece ser que, com coleções genéricas, o modelo vetorial é um bom e sólido método básico de ranqueamento. Além disso, é simples e rápido. Por essas razões, o modelo vetorial continua sendo um modelo

de recuperação popular que é constantemente utilizado para fins de comparação na avaliação de fórmulas alternativas de ranqueamento e novos modelos de RI. (BAEZA-YATES et al., 2011)

2.7 MODELO PROBABILÍSTICO E A FÓRMULA BM25

O modelo probabilístico propõe uma solução ao problema de RI com base em um arcabouço probabilístico. O escore de relevância deve refletir a probabilidade de um documento ser relevante para uma necessidade de informação, conforme julgado pelo usuário.

Os métodos probabilísticos são um dos mais antigos modelos em RI, sendo propostos inicialmente em 1976 (BAEZA-YATES et al., 2011). No entanto, foi somente nos anos 90 que, com os bons resultados apresentados pela fórmula de ranqueamento BM25, que os modelos probabilísticos começaram a ser adotados como um esquema de ponderação de termos. (MANNING et al., 2008)

A diferença entre os modelos de RI de "espaço vetorial" e "probabilístico" é muito sutil. No modelo probabilístico, ao invés de calcular o escore das consultas por meio da similaridade do cosseno e TF-IDF, o cálculo baseia-se em uma fórmula um pouco diferente motivada pela teoria probabilística (MANNING et al., 2008).

O esquema de ponderação BM25 significa *Best Match* 25 e aperfeiçoa a ponderação TF-IDF. Ele foi lançado em 1994 e as experimentações que culminaram neste modelo foram motivadas pela observação de que uma boa ponderação de termos é baseada em três princípios:

1. a frequência inversa de documentos;
2. a frequência de termos; e
3. normalização pelo tamanho dos documentos.

Assim, sua fórmula baseia-se em parâmetros mais robustos (BAEZA-YATES et al., 2011):

$$sim_{BM25}(d_j, q) \sim \sum_{t_i[q,d_j]} \frac{(k+1)f_{i,j}}{k\left[(1-b)+b\,\frac{len(d_j)}{avdl}\right]+f_{i,j}} \times log\left(\frac{N-n_i+0{,}5}{n_i+0{,}5}\right)$$

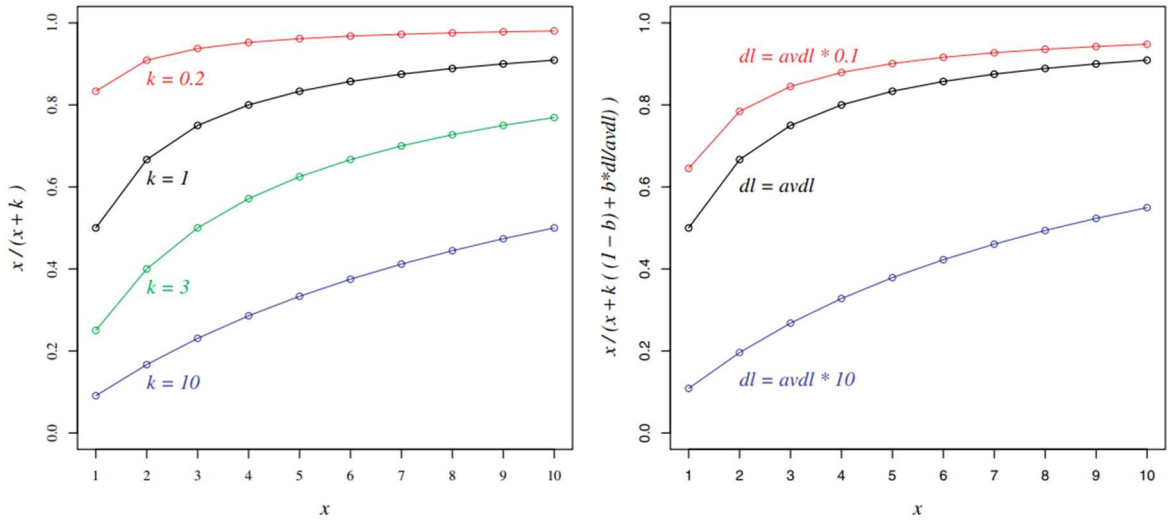
Onde:

- q é a consulta;
- d_j é um documento relevante para q;
- $t_i[q,d_j]$ é uma notação condensada para $t_i \in q \wedge t_i \in d_j$, onde t_i é um termo na consulta q;
- k é uma constante empírica, comumente recebe o valor 1.2. Alterar o valor de k modifica o impacto do TF. k maiores fazem o TF levar mais tempo para alcançar saturação;
- $f_{i,j}$ é a frequência do termo t_i no documento d_j;
- b é uma constante empírica que pode assumir valores no intervalo [0,1];
- $len(d_j)$ é o tamanho do documento d_j;
- avdl é o tamanho médio dos documentos na coleção;
- N é o número de documentos da coleção;
- n_i é o número de documentos que contém o termo t_i.

A partir desta fórmula, observa-se que:

1. o impacto da frequência de termos atinge um ponto de saturação, cuja intensidade é controlada por k;
2. o impacto do tamanho do documento (normas) é relativo à média dos tamanhos dos documentos - se o tamanho do documento for maior que o tamanho médio, a saturação ocorre menos intensamente, caso contrário, maior saturação. O parâmetro b controla o quanto as normas influenciarão o modelo;
3. o IDF é computado como na similaridade TF-IDF clássica.

Figura 2.2 – Saturação da frequência dos termos à esquerda e frequências normalizadas à direita. Maior saturação é obtida para valores pequenos de k (esquerda) e para documentos menores (direita). O gráfico à direita refere-se à k = 1 e b = 0.5.



Fonte: ROBERTSON e ZARAGOZA, 2009

O uso do BM25 tem-se tornado prevalente na comunidade de recuperação da informação (TURNBULL e BERRYMAN, 2016). Existe um crescente consenso que, uma vez que tenha sido finamente ajustado, ele fornece resultados melhores do que o modelo vetorial para coleções genéricas. Assim, ele tem sido usado como comparativo para a avaliação de novos métodos de ranqueamento, substituindo o modelo vetorial. (BAEZA-YATES et al., 2011) (MANNING et al., 2008)

2.8 SOLR/LUCENE E SEU MODELO DE RELEVÂNCIA

O Solr é uma plataforma de pesquisa de código aberto que facilita a construção de aplicações de busca sofisticadas e de alta performance. Ele é construído sobre outra tecnologia de código aberto, o Lucene, uma biblioteca Java que provê recursos de indexação e busca, com funcionalidades como *spellchecking*, *hit highlighting* e recursos avançados de análise e *tokenização*. O desenvolvimento do Solr e do Lucene caminham juntos e por isso eles possuem os mesmos *releases*. É a principal plataforma de busca utilizada pelo TCU desde 2012.

O modelo de relevância do Lucene é derivado do campo de estudo teórico de recuperação da informação. Utiliza heurísticas inspiradas na teoria e experiência aplicada em relação ao que funciona bem.

No Lucene, os objetos para os quais calcula-se a relevância são os documentos. Um documento é uma coleção de campos e cada campo possui uma semântica de como ele é criado e armazenado. O ranqueamento do Lucene é calculado nos campos e depois são combinados para retornar para os documentos. Isso é muito importante pois dois documentos com o mesmo conteúdo, mas com campos distintos, podem retornar escores diferentes por causa da normalização de tamanho dos campos e da combinação dos escores dos campos.

Em geral, o Lucene primeiramente encontra os documentos que precisam ser ranqueados baseado em três tipos de cláusulas booleanas: mandatória (+), opcional e proibida (-). Por padrão, todas as palavras ou frases especificadas na consulta são tratadas como opcionais caso não sejam precedidas por "+" ou "-". As cláusulas opcionais são recuperadas conforme a especificação do parâmetro *mm - Minimum Should Match*, que especifica qual o número mínimo de cláusulas que devem ser recuperadas. Após a recuperação dos documentos que satisfazem a condição booleana, é feito o ranqueamento deste subconjunto de documentos a partir do modelo de recuperação desejado. O Lucene suporta diferentes tipos de modelos de recuperação da informação, como o modelo vetorial, probabilístico e modelos de linguagem. Esses modelos são plugados ao Lucene por meio da *Similarity API*. Esta API fornece alguns modelos *built-in* e a possibilidade de se prover uma nova implementação de similaridade. Os modelos *built-in* podem ser modificados por meio de ajuste de parâmetros e por sobrescrita de seu comportamento.

Este conceito de similaridade no Lucene é, portanto, a implementação da ponderação dos termos e atua no nível do campo do documento. A similaridade utiliza estatísticas do índice em relação aos termos encontrados para a computação de um peso numérico para o termo em um documento. A maior parte dos modelos de similaridade são baseados na fórmula TF-IDF, mas normalmente adaptações à esta métrica são necessárias para resultados ótimos. A similaridade a ser aplicada pode ser configurada por campo da base textual. Atualmente, a similaridade padrão do Lucene é o BM25.

Conforme descrito na seção anterior, o BM25 é um modelo baseado na fórmula TF-IDF que utiliza parâmetros para o cálculo do escore. Alguns deles podem ser ajustados no Lucene para customização do BM25:

- k: constante empírica que influencia o impacto da frequência de termos (TF);
- b: constante empírica que influencia o efeito da normalização pelo tamanho do campo.

A partir dos cálculos do escore por termo e por campo, os documentos recuperados para a consulta são combinados de diferentes formas para se chegar ao resultado final. A combinação pode ocorrer de diferentes maneiras: elas podem ser o máximo entre dois valores ou uma soma de valores. O que determina como será feita a combinação da pontuação dos campos e documentos recuperados é o *query parser* utilizado.

Query Parser é um analisador léxico que interpreta uma consulta (*string*) e gera uma *query* Lucene. Há diferentes *query parsers* providos pelo Solr e os mais utilizados são o **DisMax** e sua versão estendida **eDisMax**. DisMax é um acrônimo para *Maximum Disjunction* e foi projetado para processar sentenças simples (sem sintaxe complexa) e buscar por termos individuais em vários campos utilizando diferentes pesos (*boost*) de acordo com a significância de cada campo. Os campos que serão utilizados para o cálculo do valor final do escore é o peso máximo dos campos pesquisados para o termo. Assim, é um *parser* centrado no termo pois busca por termos individuais em vários campos (mais na seção 2.10).

É possível também construir um novo *query parser* para o Solr, como o que foi construído pelo TCU em 2016 para o processamento de consultas com operadores de proximidade (adj, prox, com, mesmo).

É importante mencionar que há alguns parâmetros que influenciam o cálculo do escore final que são utilizados pelos *query parsers*, a citar:

- *boost* (^): impulsiona o escore de um campo específico.
- *boost query* (bq): consultas secundárias que modificam a principal para alterar a função de ranqueamento final.
- *coordination factor* (coord): introduz um viés no cálculo a favor dos documentos em que são localizados todos os termos da consulta. Este parâmetro pode ser desabilitado.

- *tie breaker* (tie): se especificado, controla o quanto o escore final da consulta será influenciado pelos escores dos campos que pontuaram menos em relação ao campo de maior pontuação.

Percebe-se, por fim, que o modelo de ranqueamento do Solr/Lucene é altamente configurável e customizável, o que corresponde com a noção de que relevância para o usuário depende do contexto em que se está pesquisando. Assim, é necessário avaliar o motor de busca a fim de verificar se o ranqueamento está produzindo resultados que correspondem à intuição do usuário. Os usuários não pensam em termos matemáticos conforme foram expostos e é nesse sentido que a arte e o desafio se envolvem na sintonização do modelo de relevância de uma máquina de busca.

Figura 2.3 – Decomposição do cálculo de relevância no Solr de uma pesquisa por **dispensa de licitação** na base de atos normativos do TCU utilizando o eDisMax. A SchemaSimilarity apontada é o nome usado pelo Lucene para a similaridade BM25.

```
19.229858 = sum of:
  10.381365 = sum of:
    4.2301774 = max of:
      4.2301774 = weight(TEXTODOCUMENTO:dispensa in 1372) [SchemaSimilar-
ity], result of:
        4.2301774 = score(doc=1372,freq=1.0 = termFreq=1.0), product of:
          3.071761 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq
+ 0.5)) from:
            228.0 = docFreq
            4930.0 = docCount
          1.377118 = tfNorm, computed as (freq * (k1 + 1)) / (freq + k1 * (1 - b + b
* fieldLength / avgFieldLength)) from:
            1.0 = termFreq=1.0
            1.2 = parameter k1
            0.75 = parameter b
            1011.39777 = avgFieldLength
            334.36734 = fieldLength
    6.1511874 = max of:
      6.1511874 = weight(TEXTODOCUMENTO:licitacao in 1372) [SchemaSimilarity], result
of:
        6.1511874 = score(doc=1372,freq=3.0 = termFreq=3.0), product of:
          3.3528998 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq
+ 0.5)) from:
            172.0 = docFreq
            4930.0 = docCount
          1.8345873 = tfNorm, computed as (freq * (k1 + 1)) / (freq + k1 * (1 - b +
b * fieldLength / avgFieldLength)) from:
            3.0 = termFreq=3.0
            1.2 = parameter k1
            0.75 = parameter b
            1011.39777 = avgFieldLength
            334.36734 = fieldLength
  8.848494 = max of:
    8.848494 = weight(TEXTODOCUMENTO:\"dispensa de licitacao\" in 1372) [SchemaSimi-
larity], result of:
      8.848494 = score(doc=1372,freq=1.0 = phraseFreq=1.0), product of:
        6.4253707 = idf(), sum of:
          3.071761 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq
+ 0.5)) from:
            228.0 = docFreq
            4930.0 = docCount
          7.100472E-4 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq
+ 0.5)) from:
            4927.0 = docFreq
            4930.0 = docCount
          3.3528998 = idf, computed as log(1 + (docCount - docFreq + 0.5) / (docFreq
+ 0.5)) from:
            172.0 = docFreq
            4930.0 = docCount
        1.377118 = tfNorm, computed as (freq * (k1 + 1)) / (freq + k1 * (1 - b + b *
fieldLength / avgFieldLength)) from:
          1.0 = phraseFreq=1.0
          1.2 = parameter k1
          0.75 = parameter b
          1011.39777 = avgFieldLength
          334.36734 = fieldLength
```

2.9 MODELAGEM DE *FEATURES*

A fundação de uma boa recuperação da informação é a geração dos *tokens* que serão indexados ou pesquisados (TURNBULL e BERRYMAN, 2016). Este processo é comumente chamado de análise e ele controla o processo de recuperação (*match*) dos documentos. Os tokens extraídos dos documentos são *features* que descrevem os documentos e pelas quais o documento pode ser classificado. Há *features* que são melhores para a classificação da informação e é pela análise que as *features* adequadas são extraídas. Quando configurada de forma apropriada, a análise gera *features* ricas que podem melhorar a relevância da pesquisa.

Assim, a modelagem de *features* pode ser usada para enriquecer a experiência de pesquisa ao capturar o significado, ou seja, ideias presentes no documento, e também a intenção do usuário. Idealmente, a análise não deve mapear palavras em tokens, mas sim significado e intenção em tokens. E, para isso, pode fazer uso não somente de técnicas de processamento do texto, como remoção de acentuação, lematização, radicalização, como também enriquecimento da base a partir de fontes externas como um tesauro.

A análise impacta diretamente os níveis de precisão e revocação e por isso as consequências das operações aplicadas no pré-processamento do texto devem ser avaliadas. Por exemplo, a radicalização é uma técnica de modelagem de *feature* que sacrifica a precisão para aumentar a revocação. Sem radicalização, há uma diminuição importante da revocação e a precisão aumenta, mas em um nível que pode não ser útil para o usuário típico. Assim, ao modificar o *pipeline* de análise, é possível criar tokens que equilibram o *trade-off* entre precisão e revocação.

A análise também manipula o TF x IDF para refletir de forma mais acurada a força de uma *feature* do texto. Ela pode melhorar a computação do TF x IDF ao normalizar representações diferentes de uma mesma ideia. Isso melhora a acurácia do ranqueamento, alinhando-o às noções de relevância do usuário.

Ferramentas de pesquisa são feitas para encontrar a informação que o usuário não tem por completo. Assim, se a consulta não traz os documentos óbvios que o usuário está procurando, há falha de modelagem de *feature*. Portanto, o devido investimento deve ser feito no processo de análise do texto por prover a fundação do sistema de recuperação da informação.

2.10 MODELAGEM DE *SIGNALS*

O escore final de um documento no Solr/Lucene é uma combinação dos escores dos campos do documento que são pesquisáveis. Dois documentos que contenham exatamente o mesmo conteúdo mas projetado com diferentes campos provavelmente terão pontuações diferentes para uma mesma consulta. Uma vez que o escore final é uma combinação variável dos escores dos campos e que o cálculo da similaridade entre a consulta e o conteúdo do campo leva em consideração estatísticas do campo tal como o tamanho e a frequência do termo, a modelagem da base textual em campos influencia diretamente na qualidade do modelo de ranqueamento.

Assim, cada campo influencia o comportamento do ranqueamento. A combinação dos escores é uma das maneiras de balancear os critérios que são críticos para o negócio e para os usuários e para atender esses critérios é necessário construir campos explicitamente para prover a informação certa para a solução de busca. Desta maneira, é possível programar a função de ranqueamento para levar em conta critérios importantes para o negócio ou usuários.

Chama-se *signal* o componente do cálculo da relevância que mede precisamente uma informação do negócio considerada importante para a relevância. Modelagem de *signal* é a modelagem de dados para relevância, ou seja, uma forma de construir campos para responder perguntas dos usuários. Portanto, a modelagem para relevância é bem diferente da modelagem de uma base de dados para armazenamento de conteúdo.

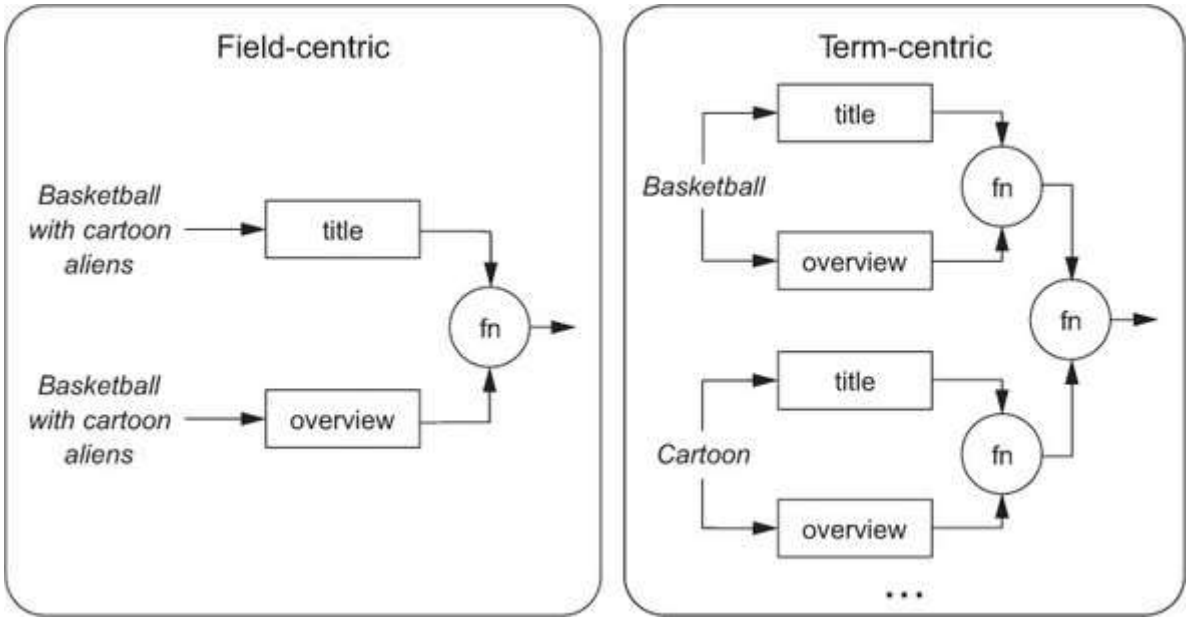
2.11 QUERY PARSERS

A combinação dos escores de similaridade calculados para cada campo do documento textual pode ser realizada de várias maneiras, sendo que é importante entender como é a combinação centrada no campo e a centrada no termo.

A consulta centrada no campo (*field-centric*) busca todos os termos da consulta em cada campo isoladamente. Os escores de cada campo são posteriormente combinados.

A pesquisa centrada no termo (*term-centric*) busca cada termo da consulta em todos os campos. O resultado é uma pontuação por termo que combina a influência de cada campo para determinado termo.

Figura 2.4 - Diferença entre uma pesquisa centrada no campo (*field-centric*) e centrada no termo (*term-centric*) para os campos *title* e *overview*. Pesquisas centradas no campo buscam todos os termos em um campo de forma isolada; já as pesquisas centradas no termo buscam em cada campo termo a termo



Fonte: TURNBULL e BERRYMAN, 2016

Os dois métodos apresentam problemas e por isso usualmente são combinados de tal forma a conseguir equilibrar o que está sendo medido e as expectativas do usuário. Este é um dos grandes trabalhos de refinamento de relevância.

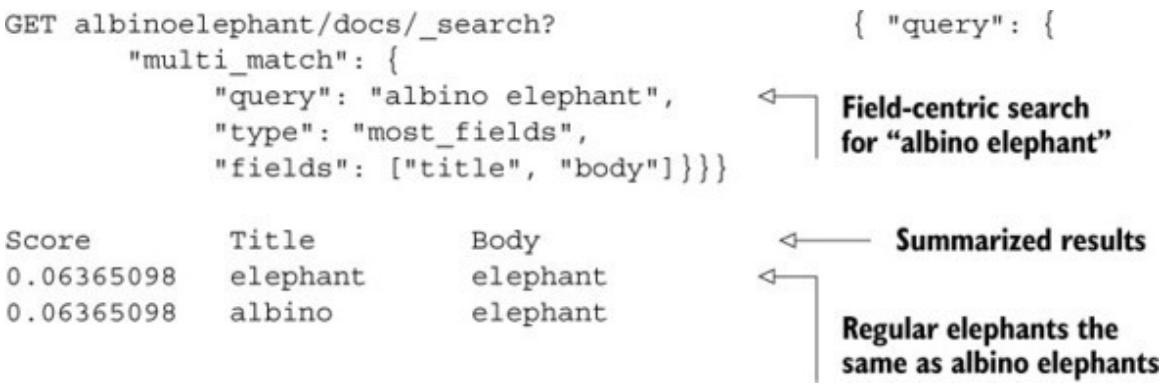
A pesquisa centrada no campo possui os seguintes problemas (TURNBULL e BERRYMAN, 2016):

1. Problema do elefante albino: este método falha em dar escore mais alto para documentos que possuem todos os termos de pesquisa.

   Em uma base que contém dois documentos que possuem os campos título e corpo, no método centrado no campo para uma pesquisa por “elefante albino”, o documento que possui o termo “elefante” no título e “elefante” no corpo pode

ser visto como na mesma situação que um documento que possui “albino” no título e “elefante” no corpo. Isso acontece porque este método de pesquisa não leva em consideração quando um termo de pesquisa ocorre em um campo e o outro ocorre em outro, conforme pode ser visto na figura 2.5. Ao ignorar termos da consulta do usuário, a ferramenta de busca pode parecer pouco inteligente para os usuários.

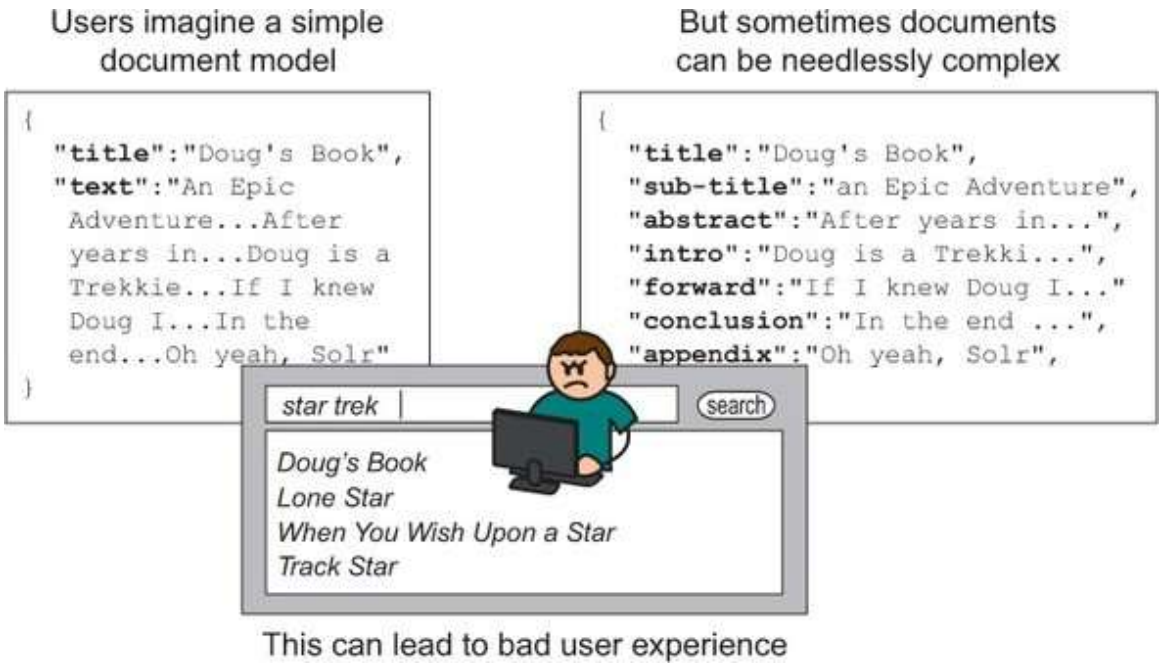
Figura 2.5 - O problema do elefante albino, conforme descrito por Chuck Williams em *2004*



Fonte: TURNBULL e BERRYMAN, 2016

2. Discordância de *signals*: o escore é baseado em partes do documento (campos) que muitas vezes não é intuitiva para o usuário ao invés de considerar o documento como um todo. Este problema é resultante da modelagem de *signals*, mas é amplificado nas pesquisas centradas no campo pois a pontuação é calculada para o campo isoladamente. Isso faz com que os escores sejam enviesados em direções específicas.

Figura 2.6 - Discordância de signals



Fonte: TURNBULL e BERRYMAN, 2016

Já as pesquisas centradas no termo possuem o problema da sincronicidade de campos: um *parser* centrado no termo requer que termos idênticos sejam pesquisados em cada campo, ou seja, cada campo deve ser pesquisado da mesma maneira (devem conter o mesmo *pipeline* de análise). Isso gera limitações e muitas vezes dificulta utilizar o campo para o objetivo que ele foi modelado.

Assim, é necessário balancear as forças de ambas as técnicas para a criação de soluções de relevância que satisfazem o usuário. É uma tarefa complexa que requer um trabalho contínuo de ajuste para se obter o balanceamento correto.

2.12 EXPANSÃO DE CONSULTAS

A expansão de consultas, também chamada por BAEZA-YATES et al. (2011) de realimentação implícita, é uma abordagem para obter informações relacionadas com a intenção por trás da consulta e utilizá-las para reformulação da consulta inicial. Os autores a chamam de

realimentação implícita pois trata-se de um método que reformula a consulta por meio de informações derivadas implicitamente pelo sistema, em contraste com a realimentação explícita cujas informações derivadas são fornecidas pelos usuários.

Existem duas abordagens básicas para a coleta de informações implícitas de realimentação, conforme descrito por BAEZA-YATES et al (2011) e MANNING et al. (2008): métodos globais e métodos locais.

Os métodos locais obtêm a informação de realimentação dos documentos do topo do ranking no conjunto de resultados e estão fora do escopo deste trabalho.

Os métodos globais obtêm a informação de realimentação a partir de fontes externas, como um tesauro ou das relações de termos extraídas da coleção de documentos. Métodos globais incluem:

- expansão/reformulação de consulta por meio de um tesauro:

  Uso de um vocabulário controlado mantido por editores humanos que especifica os sinônimos e termos relacionados.

- expansão de consulta por meio da geração automática de um tesauro:

  Um tesauro pode ser derivado automaticamente por meio da análise da coleção de documentos. Há algumas abordagens possíveis.

  Uma primeira abordagem é explorar a coocorrência de palavras e construir um tesauro estatístico global.

  Outra forma é a construção de um tesauro de similaridade que se baseia nos relacionamentos entre termos em vez de em uma matriz de coocorrência.

A expansão de consultas é efetiva para aumentar a revocação e é amplamente usada em muitas áreas. No entanto, a expansão de consultas pode diminuir significativamente a precisão, principalmente quando a query possui termos ambíguos. (MANNING et al., 2008)

2.13 WORD2VEC

Word2Vec são técnicas que podem ser usadas para o aprendizado de vetores de palavras de alta qualidade a partir de grandes conjuntos de dados. (MIKOLOV et a., 2013)

Os modelos propostos pelo Word2Vec são arquiteturas de redes neurais *feed-forward*, cuja entrada são trechos de texto e a saída um conjunto de vetores, uma para cada palavra do texto. Quando os vetores de saída são plotados em um gráfico de duas dimensões, vetores cujas palavras são similares em termos semânticos estão muito próximos uns dos outros. Distâncias como a do cosseno podem ser usadas para encontrar as palavras mais próximas. Esses vetores gerados pelo Word2Vec são muitas vezes referenciados como *word embeddings*. (TEOFILI e MATTMANN, 2019)

O conceito geral por trás do Word2Vec é que a rede neural recebe como entrada um trecho de texto, que é dividido em fragmentos de tamanho específicos (*window size*). Cada fragmento é submetido à rede como um par consistindo de uma palavra e o contexto associado.

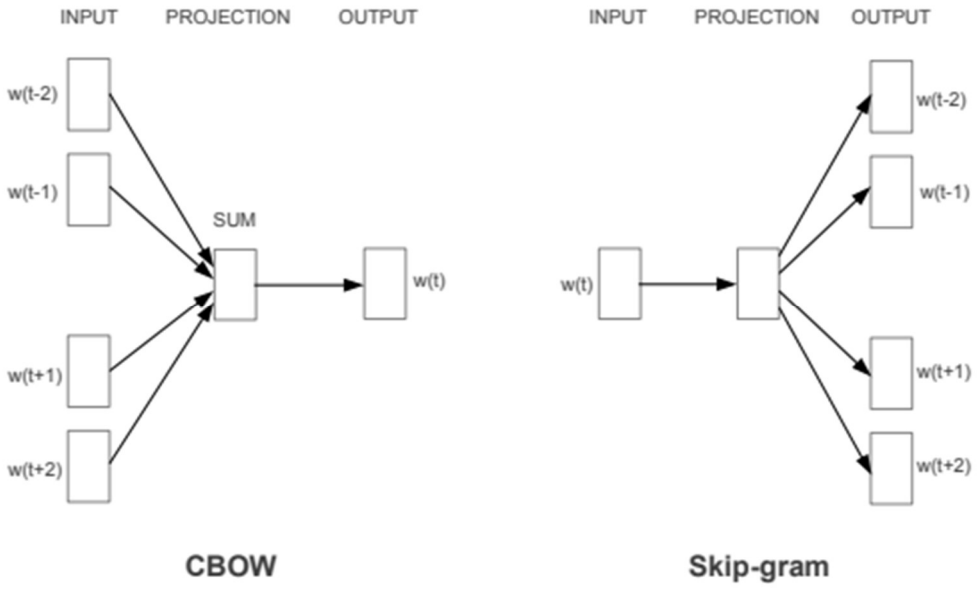
A camada intermediária da rede contém um conjunto de pesos para cada palavra. Esses vetores serão usados como a representação da palavra quando o aprendizado da rede terminar. É importante ressaltar que a saída da rede neural para o Word2Vec não é relevante, pois é o estado interno da camada *hidden* que será utilizado como a representação de cada palavra.

O Word2Vec propõe dois modelos de arquiteturas de redes neurais para a aprendizagem de representações distribuídas de palavras: *continuous bag of words* (CBOW) e *continuous skip-gram*.

No modelo CBOW, a palavra alvo é utilizada como a saída da rede, e as palavras remanescentes do fragmento de texto (contexto) são usadas como entrada. Esta arquitetura é chamada de modelo *bag-of-words* pois a ordem das palavras anteriores não influencia a projeção.

Já no modelo *skip-gram*, a palavra alvo é utilizada como entrada e as palavras do contexto são a saída.

Figura 2.7 - O modelo CBOW prediz a palavra corrente baseado no contexto, e o *Skip-gram* prediz palavras circundantes dada uma palavra



Fonte: MIKOLOV et al., 2013

2.14 POINTWISE MUTUAL INFORMATION

Representações de palavras em vetores são limitadas pela inabilidade de representar expressões idiomáticas que não são simplesmente uma composição das palavras isoladas. Por exemplo, “Tribunal de Contas da União” é um órgão da administração pública brasileira e não uma combinação dos significados das palavras individuais.

Os modelos que geram *word embeddings* podem ser adaptados para serem modelos baseados em frases ao invés de baseados em palavras. Esta extensão é relativamente simples. Primeiramente, são identificadas as expressões frequentes no texto que são transformadas em termos únicos. Com o texto adaptado, o modelo Word2Vec pode ser treinado. (MIKOLOV et al., 2013)

Uma técnica que pode ser utilizada para extração dessas expressões é a *Pointwise Mutual Information* - PMI. O PMI é uma medida de associação que identifica colocações, que são

strings formadas de duas ou mais palavras que ocorrem juntas frequentemente. Dada duas palavras, a e b, o PMI identifica quanto da probabilidade delas juntas difere da probabilidade individual esperada, assumindo que elas são independentes. Isso pode ser expresso como:

$$\operatorname{pmi}(x;y) \equiv \log \frac{p(x,y)}{p(x)p(y)} = \log \frac{p(x|y)}{p(x)} = \log \frac{p(y|x)}{p(y)}.$$

Bons pares de palavras possuem PMI alto pois a probabilidade da coocorrência é só um pouco inferior que as probabilidades da ocorrência de cada palavra individualmente. Por outro lado, um par de palavras cujas probabilidades de ocorrência são mais altas que a probabilidade de coocorrência obtém um escore baixo de PMI.

2.15 LEARN TO RANK

Tipicamente, modelos de ranqueamento contém parâmetros que precisam ser ajustados para resultados ótimos. Por exemplo, o BM25 contém os parâmetros b e k que precisam ser moldados para fornecer boa performance para um conjunto de dados de validação. No entanto, o ajuste de parâmetros não é uma tarefa trivial: um modelo perfeitamente ajustado para um conjunto de validação pode apresentar baixa performance em consultas nunca vistas (*overfiitting*).

A aprendizagem de máquina tem se mostrado efetiva em ajustar automaticamente parâmetros, combinando múltiplas evidências e evitando o *overfitting*. Por isso, é vista como promissora para resolver problemas de ranqueamento.

Assim, o *Learn to Rank* em recuperação da informação é a tarefa de construir modelos de ranqueamento automáticos utilizando aprendizagem de máquina, de tal forma que o modelo possa ordenar objetos de acordo com o grau de relevância, preferência ou importância. (LIU, 2011)

O *Learn to Rank* na plataforma Solr é utilizado no contexto de *re-ranking*, ou seja, dado um ranqueamento inicial, um segundo nível de ranqueamento por meio do *Learn to Rank* pode ser aplicado para reordenar os primeiros n documentos.

3 DESENVOLVIMENTO

3.1 ENTENDIMENTO DOS DADOS

Cada aplicação de busca possui expectativas de relevância muito distintas. Apesar de muitas vezes ser vista como um problema único, aplicações de busca diferem muito umas das outras. É fundamental entender os objetivos dos usuários, o domínio de negócio e o contexto para medir o que importa quando uma busca é realizada. É por meio da seleção das *features* corretas e da implementação dos *signals* que mapeiam para as necessidades dos usuários e do negócio que o sistema de recuperação da informação atende às necessidades de informação dos usuários.

Assim, esta seção tem por objetivo explorar o contexto de negócio relacionado à aplicação de busca. Será apresentada a justificativa da escolha da base de atos normativos da Presidência do TCU, o entendimento sob o ponto de vista do negócio dos atos normativos, a compreensão de como a pesquisa de atos normativos é utilizada, quais são as expectativas do usuário quando realizam busca por atos normativos, quais são os dados relacionados a atos normativos do TCU, quais são os temas mais buscados e os julgamentos de relevância para algumas pesquisas.

3.1.1 Escolha da base de atos normativos

O primeiro passo dado para escolher a base textual para realizar este trabalho foi realizar o levantamento de quais bases textuais são mais pesquisadas. A partir do *log* que registra todas as pesquisas feitas na plataforma de busca integrada, verificou-se que, em primeiro lugar, a base de jurisprudência é a mais consultada, seguida da base de processos e, em terceiro lugar, atos normativos do TCU.

Uma escolha natural seria a pesquisa de jurisprudência. No entanto, apesar de ser a base textual mais acessada do TCU, esta pesquisa apresenta complexidades que dificultariam ou até mesmo inviabilizaria este trabalho.

A base de acórdãos do TCU possui aproximadamente 300 mil documentos, sendo que a maioria deles contém dezenas ou centenas de páginas. Ou seja, são documentos longos e em grande número. Verificou-se, ao longo desses anos de desenvolvimento da pesquisa de jurisprudência, que um usuário típico que busca jurisprudência do TCU não objetiva localizar um acórdão sobre determinado assunto, mas entender como o TCU vem decidindo sobre um tema, o que reflete uma necessidade de alta revocacão nos resultados da busca.

A primeira tentativa de implantar um modelo ranqueado de busca de jurisprudência foi malsucedida e muito criticada. Não houve ajustes no modelo de ranqueamento de acordo com as necessidades de negócio e foi utilizado o cálculo de relevância padrão da ferramenta em uma base textual que continha mais de 60 campos modelados, o que justifica o insucesso. Assim, esta circunstância, somada ao fato de a pesquisa de acórdãos ser o sistema corporativo do TCU mais utilizado - o que o torna um sistema bem crítico, impõem riscos para uma provável entrada em produção dos resultados deste trabalho. Não seria, portanto, uma boa estratégia utilizar esta base textual.

A pesquisa de processos, a segunda colocada, poderia se beneficiar de uma evolução para um modelo ranqueado. No entanto, a base atual de processos é muito restrita, pois somente contém os metadados dos processos, não alcançando suas peças. É uma pesquisa muito simples e ofereceria poucos desafios para o presente trabalho.

Assim, optou-se por utilizar a base de atos normativos do TCU não somente por ela ser a terceira mais acessada, mas também porque um usuário típico desta pesquisa busca satisfazer uma necessidade de informação que normalmente está contida em um único documento, ou seja, não é necessário fazer uma análise de vários documentos para atender o seu objetivo.

3.1.2 Os Atos Normativos da Presidência do TCU

O artigo 67 do Regimento Interno do TCU estabelece os tipos de deliberações do Plenário do TCU:

Art. 67. As deliberações do Plenário e, no que couber, das câmaras, terão a forma de:

I – instrução normativa, quando se tratar de disciplinamento de matéria que envolva pessoa física, órgão ou entidade sujeita à jurisdição do Tribunal;

II – resolução, quando se tratar de:

a) aprovação do Regimento Interno, de ato definidor da estrutura, atribuições e funcionamento do Tribunal, das unidades de sua Secretaria e demais serviços auxiliares;

b) outras matérias de natureza administrativa interna que, a critério do Tribunal, devam revestir-se dessa forma;

III – decisão normativa, quando se tratar de fixação de critério ou orientação, e não se justificar a expedição de instrução normativa ou resolução;

IV – parecer, quando se tratar de:

a) Contas do Presidente da República;

b) outros casos em que, por lei, deva o Tribunal assim se manifestar;

V – acórdão, quando se tratar de deliberação em matéria da competência do Tribunal de Contas da União, não enquadrada nos incisos anteriores.

Parágrafo único. As deliberações previstas neste artigo serão formalizadas nos termos estabelecidos em ato normativo.

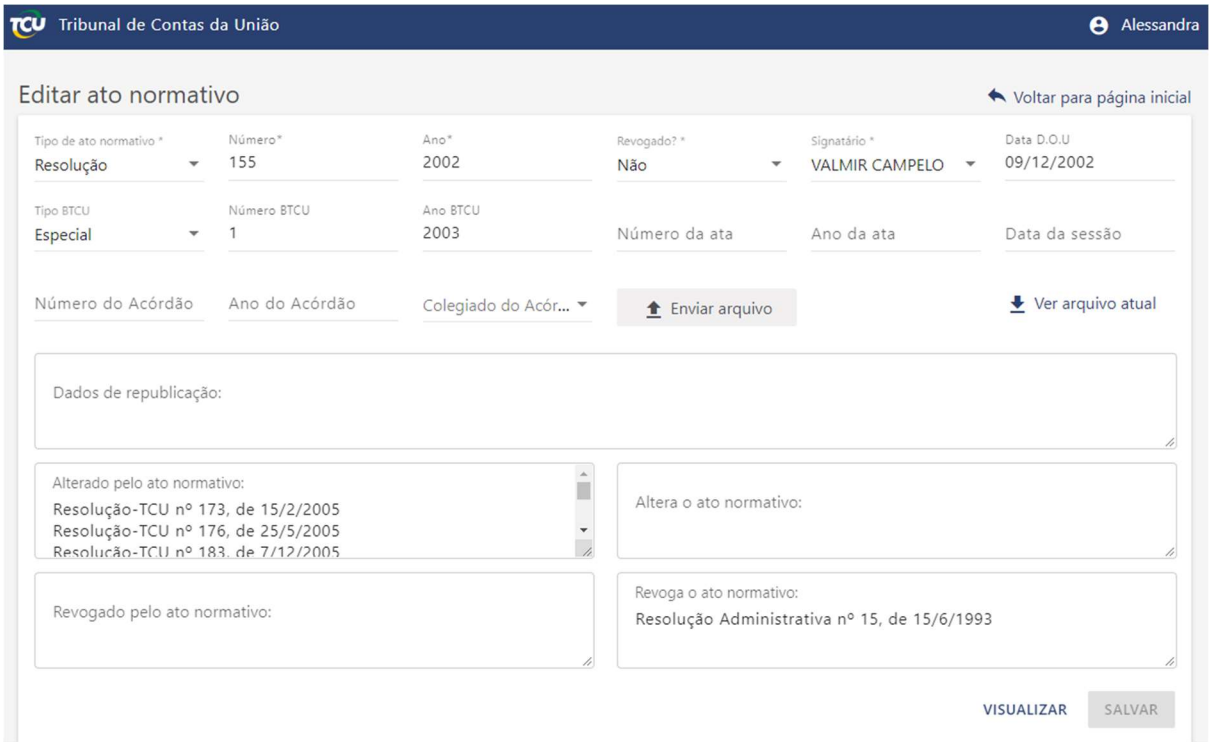
Assim, os tipos de atos normativos contemplados pela base de atos normativos do TCU são:

- Instruções normativas e Decisões normativas: são atos normativos que impactam o público externo ao TCU.
- Resoluções: são atos normativos sobre matérias internas ao TCU.
- Portarias: são atos normativos de interesse geral

A Secretaria das Sessões (Seses) possui a atribuição regimental de manter os atos normativos do TCU em base de dados atualizada, por isso esta unidade é a gestora da solução de pesquisa de atos normativos.

Após a aprovação do ato normativo por meio de um acórdão (instruções normativas, decisões normativas e resoluções) ou após a publicação de uma portaria da Presidência no Boletim do TCU - BTCU, a Seses realiza o cadastramento do ato normativo em um sistema de gerenciamento de atos normativos. É registrado neste sistema os metadados do ato normativo com referência para o documento que representa o ato.

Figura 3.1 – Tela de cadastramento de um ato normativo da Presidência do TCU



Após o cadastro ou atualização, o ato normativo fica disponível para indexação, que é um processo *batch* que executa duas vezes por dia - às 0h e às 13h. Após a indexação, o ato normativo fica disponível na plataforma de pesquisa.

Há alguns problemas potenciais nesse processo, relacionados ao erro humano durante o cadastro ou atualização:

- Pode ocorrer erro na atualização do relacionamento entre os atos, por exemplo, o indicativo de que um ato altera ou revoga outro.
- Quando um ato altera outro, há a replicação do conteúdo do ato que altera no ato alterado, quando podem ocorrer erros.
- Os atos normativos que somente alteram outros acabam representando um ruído, pois seu conteúdo é sempre replicado no ato alterado.

3.1.3 A base textual de Atos Normativos

A coleção de atos normativos da Presidência do TCU possui atualmente 3178 documentos (posição em dezembro/2019). Um documento desta coleção contém os seguintes campos:

Tabela 3.1 – Campos da base textual de Atos Normativos da Presidência do TCU

Nº	Nome	Descrição	Tipo	Pesquisável?
1	ALTERADOS	Atos normativos que foram alterados por este ato.	text_pt multivalorado	Sim
2	ALTERADOSPOR	Atos normativos que alteraram este ato.	text_pt multivalorado	Sim
3	ANEXOS	Texto contido nos anexos do ato normativo.	text_pt	Sim
4	ANOACORDAO	Ano do Acórdão que aprovou o ato, quando for o caso.	text_pt	Não
5	CARGOSIGNATARIO	Cargo do signatário do ato. Como esta base contém somente atos da Presidência do TCU, possui o mesmo conte-	text_pt	Sim

		údo para todos os atos: Presidente.		
6	COLEGIADO	Colegiado que aprovou o ato. Como esta base contém somente atos da Presidência do TCU, possui o mesmo conteúdo para todos os atos: Plenário.	text_pt	Sim
7	COLEGIADOACORDAO	Colegiado do Acórdão que aprovou o ato, quando for o caso.	text_pt	Não
8	DADOSREPUBLICACAO	Republicação oficial	text_pt	Sim
9	DATADOU	Data da publicação do ato normativo no Diário Oficial da União. Formato dd/mm/aaaa.	text_pt	Sim
10	DATADOUORDENACAO	Data da publicação do ato normativo no Diário Oficial da União usada para fins de ordenação do resultado da pesquisa. Formato aaaammdd.	text_pt	Sim
11	DATAEXPEDICAO	Campo existente, mas sem sentido para o negócio e não utilizado.	tdate	Não
12	DATAEXPEDICAOORDENACAO	Campo existente, mas sem sentido para o negócio e não utilizado.	string	Não
13	DATASESSAO	Data da sessão em que o ato foi aprovado. Formato dd/mm/aaaa	text_pt	Sim

14	DATASESSAOORDENACAO	Data da sessão em que o ato foi aprovado, mas utilizada para fins de ordenação dos resultados da pesquisa. Formato aaaammdd.	text_pt	Sim
15	DTATUALIZACAO	Data a última indexação do ato normativo. Formato aaaammdd.	string	Não
16	DTRELEVANCIA	Data de expedição do ato normativo para fins de ordenação. Formato aaaammdd.	string	Não
17	EMENTA	Ementa do ato normativo. A ementa é o assunto do ato normativo.	text_pt	Sim
18	FILENAME	URL para acesso ao documento no GED	string	Não
19	KEY	Chave do ato normativo, utilizada para identificar unicamente o documento na base textual. Não tem significado de negócio. Formato: ATO-NORMATIVO-NNNNNN	string	Não
20	NOMESIGNATARIO	Nome do signatário do ato. Como esta base contém somente atos da Presidência do TCU, é o nome do presidente da época.	text_pt	Sim
21	NOTACOMPILACAO	Nota de compilação	text_pt	Sim
22	NUMANOATA	Ano da ata da sessão em que o ato normativo foi aprovado.	string	Não

23	NUMANOATO	Ano do ato normativo	string	Não
24	NUMATO	Número do ato normativo	text_pt	Sim, com boost de 1000000000
25	NUMATOINT	Número do ato normativo em formato inteiro.	int	Não
26	NUMD	Identificador único do ato normativo de uso interno do sistema. Não tem sentido para o negócio. Formato: Sigla-Ano-Numero. Ex:PRT2019-378.	string	Não
27	NUMEROACORDAO	Número do Acórdão que aprovou o ato, quando for o caso.	text_pt	Não
28	NUMEROATA	Número da ata da sessão em que o ato normativo foi aprovado.	text_pt	Sim
29	NUMEROBOLETIMTCU	Número completo do identificador do BTCU em que o ato normativo foi publicado. Está no formato número/ano: NN/AAAA.	text_pt	Sim
30	REDACAOANTERIOR	Redação anterior do ato normativo, antes de alterações.	text_pt	Sim
31	REVOGADOS	Atos normativos revogados por este ato.	text_pt multivalorado	Sim
32	REVOGADOSPOR	Atos normativos que revogaram este ato.	text_pt multivalorado	Sim

33	SITUACAO/COPIASITUA-CAO	Situação do ato normativo: - Não consta revogação expressa - Revogada	text_pt/ string	Sim/ Não
34	TEXTODOCUMENTO	Texto contido no corpo do ato normativo.	text_pt	Sim
35	TIPO/COPIATIPO	Tipo do ato normativo: - Portaria - Resolução - Decisão Normativa - Resolução Administrativa - Instrução normativa	string/string	Não/Não
36	TIPOBTCU	Tipo do BTCU em que o ato normativo foi publicado.	text_pt	Sim
37	TITULO/COPIATITULO	Título do ato normativo que é apresentado na tela de resultado da pesquisa. Seu conteúdo é o número do ato e o ano, no formato NNN/AAAA.	text_pt/string	Sim, com boost de 1000000000/ string

Nesta base, são utilizados quatro tipos de campos:

- text_pt: campo texto que passa por análise.

  Os *pipelines* de análise para indexação e consulta são:

Os dois *pipelines* são parecidos, mas há algumas diferenças. Na análise da consulta, há remoção de *stop words*.

As *stop words* são palavras que possuem pouco valor para a seleção de documentos para uma dada consulta. No entanto, há situações em que elas fazem a diferença na recuperação (MANNING et al., 2008). Por exemplo, em consultas frasais (entre aspas), *stop words* impactam diretamente a precisão da recuperação: a consulta "vôo para Londres" perde muita precisão caso a *stop word* seja removida. Outro exemplo é a pesquisa por "A Cabana", cuja intenção é a busca por um filme. Nesse caso, se houver remoção da *stop word*, o motor de busca retornará documentos não relacionados à intenção do usuário. MANNING et al (2008) afirmam que a tendência atual em recuperação da informação é o uso de listas de *stop words* muito pequenas ou simplesmente não as remover.

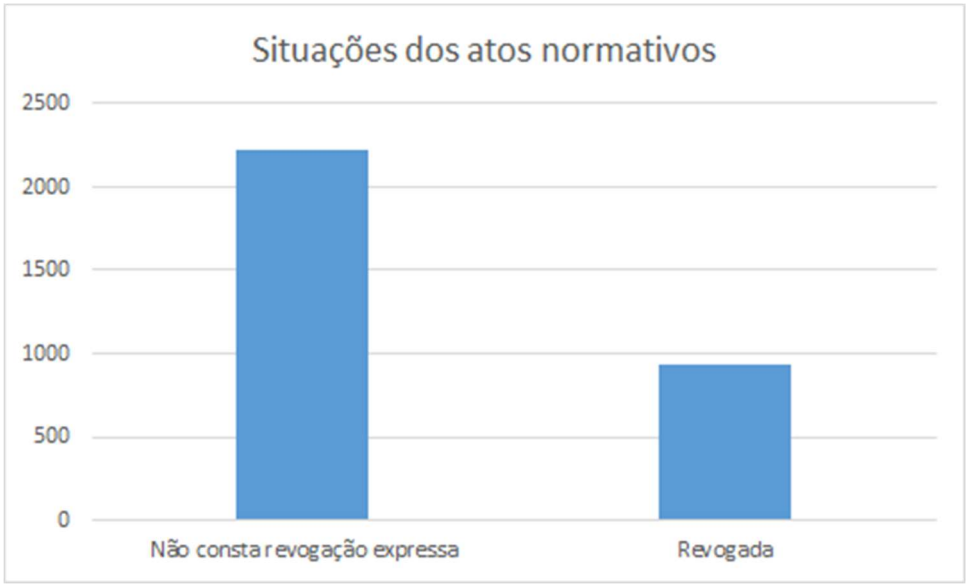
Apesar da plataforma de pesquisa textual ainda não considerar as *stop words* na recuperação da informação ao removê-las no *pipeline* de consulta, tomou-se a decisão de mantê-las no índice. Isso aumenta o custo de armazenamento, o que não é significativo, mas abre a possibilidade de se implementar a consideração de *stop words* na recuperação sem necessidade de reindexação da base.

- string: é um campo texto não *tokenizado*, ou seja, o conteúdo total do campo é interpretado como um token. Normalmente, este tipo de campo é usado para gerar facetas.
- tdate: campo do tipo data
- int: campo inteiro

A partir de uma análise exploratória desta base, observou-se:

1. 42% dos atos normativos da coleção estão revogados. Este número impressiona, pois, a situação não é levada em consideração para o retorno dos resultados da pesquisa. É claro que pode haver a necessidade de se encontrar atos revogados e isso será confirmado com o negócio. No entanto, acredita-se que, no uso mais comum desta pesquisa, os atos revogados representam ruído.

Figura 3.2 – Quantidade de atos normativos por situação



2. Ao examinar o conteúdo dos atos normativos vigentes, ou seja, aqueles em situação “Não consta revogação expressa”, observa-se que eles se subdividem em:
    - Normas que possuem nova disposição
    - Normas que alteram e que possuem nova disposição
    - Normas somente que alteram outras

- Normas que somente revogam outras

As normas das categorias c e d acabam representando ruído dentro da coleção de documentos. Uma norma que somente altera outra traz o novo texto em seu corpo, mas o novo texto é replicado na norma alterada. Assim, seu conteúdo pode ser localizado na norma alterada e a norma que altera representa somente o meio formal. Usualmente, o usuário busca o texto original que contém todas as alterações e não as normas que fizeram a alteração. De qualquer forma, a norma que fez a alteração está relacionada à norma alterada, não somente por referência no texto mas também em metadado do ato normativo.

A norma que somente revoga outra segue a mesma linha de raciocínio descrita anteriormente; a revogação será refletida na norma revogada por meio da alteração de situação e o ato revogador é um metadado do ato revogado.

Assim, é interessante avaliar essa condição para fins de cálculo de relevância.

3. Verificou-se que há 39 atos normativos que foram revogados por outros, mas que não estão com a situação atualizada. A verificação foi feita por meio da consulta por atos que possuem o campo REVOGADOPOR preenchido cuja situação é "Não consta revogação expressa". Esta inconsistência foi validada manualmente por meio de análise do texto.
4. Os pares de campos (REVOGADOS, REVOGADOSPOR) e (ALTERADOS, ALTERADOSPOR) são cadastrados manualmente e há o risco de inconsistência no cadastro.
5. Os campos TITULO e NUMEROBOLETIMTCU possuem exatamente o mesmo formato (nnn/aaaa) e ambos são pesquisáveis. São *signals* que podem impactar negativamente o cálculo da relevância, pois a busca por número/ano é muito comum e o *score* de um documento pode subir indevidamente pois houve um *match* indevido no número ou ano no campo NUMEROBOLETIMTCU.
6. O fato de o campo TIPO não ser pesquisável e estar definido como do tipo *string* chama atenção pois, apesar da interface do sistema oferecer filtros pelo tipo do ato, este não está sendo pesquisado com os termos da pesquisa livre. O campo

COPIATIPO, que é uma cópia deste campo, também está definido como *string*, o que representa uma redundância desnecessária.

7. A pesquisa por atos normativos ordena os resultados por data decrescente. No entanto, quando um ato normativo é muito antigo e vigente e outro é muito recente e revogado, a pesquisa retornará o ato revogado antes do ato vigente, o que pode não fazer sentido. A classificação por data decrescente neste contexto portanto parece não ser muito importante como na base de jurisprudência, pois o tipo de necessidade de informação dos usuários é diferente.

8. Há vários campos de datas que são pesquisáveis, o que chama atenção. Na tela de filtros da pesquisa, há opção para filtrar a pesquisa somente por data de expedição. É interessante, portanto, revisar a necessidade de se ter vários campos de data sendo pesquisados.

9. Há dois campos de data cujos nomes contêm a palavra "ORDENACAO", mas que estão sendo pesquisados. Não faz sentido um campo de data no formato aaaammdd ser pesquisável.

10. Os campos pesquisáveis para *queries* e *phrase queries* (quando os termos estão entre aspas) estão configurados diferentemente na base de atos normativos. Isso é contra intuitivo para o usuário uma vez que ele pode localizar documentos totalmente diferentes somente pelo fato de informar sua consulta entre aspas. O uso das aspas nas expressões de busca aumenta a precisão e diminuem a revocação e normalmente é essa a intenção do usuário quando utiliza aspas em sua expressão de busca.

11. Há campos pesquisáveis que não fazem sentido estarem configurados como tal. Por exemplo, número e tipo do boletim TCU. Percebe-se, portanto, que há uma confusão em relação a campo pesquisável e campo a ser utilizado somente para armazenamento para fins de apresentação da informação para o usuário. Configurar um campo como pesquisável quando ele não faz sentido para o usuário pode gerar vieses indesejados no cálculo da pontuação do documento em relação à consulta.

3.1.4 O uso da pesquisa de Atos Normativos

Para entender como a pesquisa de atos normativas é utilizada pelos usuários, foi realizada uma análise exploratória do log de consultas e foram conduzidas entrevistas com usuários selecionados.

Uma forma indireta de entender as necessidades de informação dos usuários é analisando o log de pesquisa. Assim, foi construído um notebook Python para fazer a análise exploratória dos logs de pesquisa da base de atos normativos do período de 15/2/2019 a 31/07/2019.

O resumo da análise exploratória pode ser visto a seguir.

Figura 3.3 – Resumo da análise exploratória dos logs de pesquisa em atos normativos

```
Análise pesquisas por * ----------------------------------------------------------
Percentual de pesquisas por *: 84.14%
    Percentual de pesquisas por * com filtro: 11.78%
    Os filtros são:
     ('^TIPO:.*$', 63.45792661958045)
     ('^TIPO:(.*) NUMATO:.*$', 46.69206147660239)
     ('^TIPO:(.*) NUMATO:.* NUMANOATO:.*$', 36.235219427029605)
     ('^TIPO:(.*) NUMANOATO:.*$', 41.687167685104356)
     ('^TIPO:(.*) NUMANOATO:.* DATAEXPEDICAOORDENACAO:.*$', 0.5660926381504113)
     ('^TIPO:(.*) DATAEXPEDICAOORDENACAO:.*$', 1.6771155728381344)
     ('^NUMATO:.*$', 26.185752453508982)
     ('^NUMATO:.* NUMANOATO:.*$', 22.76274369759014)
     ('^NUMANOATO:.*$', 2.833108483453694)
     ('^DATAEXPEDICAOORDENACAO:.*$', 1.304129301907256)
     ('^COPIASITUACAO:"Não consta revogação expressa" TIPO:(.*) NUMATO:.* NUMANOATO:.*$', 1.3623257413432797)
    Percentual de pesquisas por * sem filtro - externo: 84.20%
    Percentual de pesquisas por * sem filtro - publico: 91.79%

Análise pesquisas com 0 resultados-----------------------------------------------
Percentual de pesquisas com 0 resultados: 3.50%
    Percentual de pesquisas com 0 resultado com filtro: 21.78%
```

Nessa análise, chama atenção o número impressionante e inesperado de 84% das consultas serem por “*”. E, destes 84%, um filtro é especificado somente em 12%. É importante tentar entender com o negócio porque isso ocorre.

Além disso, o levantamento de quais são os filtros mais usados mostra algumas necessidades de informação importantes para o usuário. Assim, a pesquisa por tipo do ato, número e ano se destacam. A data de expedição é usada em somente 3% das pesquisas com filtro.

Outra informação importante é que 3,5% das pesquisas não retornam resultados. É muito frustrante para o usuário realizar uma pesquisa e não encontrar nenhum resultado. Ao avaliar as consultas com zero resultados, constatou-se:

- há casos em que o usuário informa corretamente número e ano do ato, mas informa o tipo do ato incorretamente. Por exemplo, na pesquisa "INSTRUÇÃO NORMATIVA 156", provavelmente o usuário buscava a decisão normativa 156.

Há estratégias para lidar com esta questão, como fazer correção ortográfica dos termos da pesquisa ou expansão de consultas. Este ponto precisa ser considerado no trabalho de melhoria de relevância.

A partir do log de pesquisas, levantou-se as pesquisas mais realizadas na base de atos normativos e quais são os documentos mais acessados:

Tabela 3.2 – Pesquisas mais realizadas na base de atos normativos

Posição	Termos de pesquisa
1	Licença capacitação
2	Feriados
3	Teletrabalho
4	Regimento interno
5	Recesso
6	Monitoramento
7	Tomada de contas especial
8	Avaliação de desempenho
9	Diárias e passagens
10	Marinha
11	concessionário MESMO \"contratação direta\
12	licitação
13	remoção
14	relatório de gestão

15	estrutura do tcu
16	assistência pré-escolar

Tabela 3.3 – Documentos mais acessados da base de atos normativos do TCU

Posição	Número do Ato	Ementa
1	259/2014 (ATO-NORMATIVO-120268)	Estabelece procedimentos para constituição, organização e tramitação de processos e documentos relativos à área de controle externo.
2	305/2018 (ATO-NORMATIVO-1041)	Define a estrutura, as competências e a distribuição das funções de confiança das unidades da Secretaria do Tribunal de Contas da União.
3	71/2012 (ATO-NORMATIVO-111822)	Dispõe sobre a instauração, a organização e o encaminhamento ao Tribunal de Contas da União dos processos de tomada de contas especial.
4	175/2005 (ATO-NORMATIVO-50587)	Dispõe sobre normas atinentes à distribuição de processos a ministros e auditores no âmbito do Tribunal de Contas da União.
5	101/2019 (ATO-NORMATIVO-1342)	Dispõe sobre a realização de teletrabalho por servidores ocupantes de cargos efetivos do Quadro de Pessoal da Secretaria do Tribunal de Contas da União.
6	31/2019 (ATO-NORMATIVO-1081)	Divulga os feriados nacionais e define os dias de ponto facultativo em 2019 no âmbito do Tribunal de Contas da União.
7	444/2018 (ATO-NORMATIVO-1028)	Dispõe sobre o processo de contratação de serviços, no âmbito da Secretaria do Tribunal de Contas da União (TCU).
8	170/2018 (ATO-NORMATIVO-641)	Dispõe acerca das unidades cujos dirigentes máximos devem prestar contas de suas gestões ocorridas no exercício de 2018, especificando a forma, os conteúdos e os prazos de apresentação, nos termos do art. 3º da Instrução Normativa TCU 63, de 1º de setembro de 2010.
9	63/2010 (ATO-NORMATIVO-86212)	Estabelece normas de organização e de apresentação dos relatórios de gestão e das peças complementares que constituirão os processos

		de contas da administração pública federal, para julgamento do Tribunal de Contas da União, nos termos do art. 7º da Lei nº 8.443, de 1992.
10	443/2018 (ATO-NORMATIVO-1021)	Disciplina, no âmbito do Tribunal de Contas da União, a emissão de passagens, a concessão de diárias e as demais indenizações relativas a viagens a serviço.
11	1/2019 (ATO-NORMATIVO-1062)	Delega competência ao Secretário-Geral de Administração para os fins que especifica
12	294/2018 (ATO-NORMATIVO-42)	Dispõe sobre a classificação da informação quanto à confidencialidade no âmbito do Tribunal de Contas da União.
13	181/2019 (ATO-NORMATIVO-1601)	Aprova o Plano de Gestão do Tribunal de Contas da União para o período de abril de 2019 a março de 2021 e a distribuição, nos períodos avaliativos, dos valores das metas que compõem o resultado institucional
14	265/2014 (ATO-NORMATIVO-122857)	Dispõe sobre a expedição e o monitoramento de deliberações que tratam de determinações, recomendações e de ciência a unidades jurisdicionadas, no âmbito do Tribunal de Contas da União.
15	308/2018 (ATO-NORMATIVO-782)	Dispõe sobre o funcionamento das unidades da Secretaria do Tribunal de Contas da União durante o período de recesso relativo a 2018-2019.
16	212/2008 (ATO-NORMATIVO-71739)	Dispõe sobre o desenvolvimento de ações de educação no âmbito do Tribunal de Contas da União.

Há um outro aspecto em relação à pesquisa de atos normativos que não será tratado no contexto deste trabalho que é em relação ao escopo desta base. Há somente os atos normativos da presidência do TCU. Atos normativos expedidos pelas unidades e subunidades do TCU não constam desta base. No entanto, verificou-se no log de pesquisas que há usuários que buscam informações que são normatizadas por outras unidades, o que mostra que pode ser intuitivo para o usuário que todos os atos normativos, independentemente da autoria, estarem em uma base

única. Além disso, há uma grande dificuldade em encontrar os atos normativos das unidades do TCU pois eles são armazenados de forma dispersa e frequentemente não há possibilidade de se pesquisar em seu conteúdo. Seria interessante tratar este problema para aumentar a eficiência interna do TCU.

Em relação às entrevistas com usuários, foram entrevistados 9 colaboradores do TCU de diferentes áreas. Os dados obtidos nas entrevistas estão no Anexo II.

De forma sumarizada, as principais constatações são:

- 75% dos entrevistados entendem que se a informação for localizada na ementa do ato normativo, isto torna o ato mais relevante para a consulta.
- 100% dos entrevistados afirmaram que os tipos de atos normativos não determinam importância diferenciada ou relações de hierarquia e que nem sempre sabem qual o tipo do ato que buscam.
- 75% dos entrevistados concordam que a data de expedição nem sempre é o melhor para a ordenação do resultado da pesquisa.
- 75% dos entrevistados entendem que os atos revogados são menos importantes que os atos vigentes.
- 87,5% dos entrevistados citaram dificuldades em relação à pesquisa por termos, pois não encontram o que necessitam seja por causa que buscam por expressões sinônimas ou relacionadas ou por variações nos termos (plurais, conjugações verbais etc.).
- 62,5% dos entrevistados citaram que a descentralização dos atos normativos do TCU em geral é um grande dificultador de seu trabalho.
- 50% dos entrevistados sugeriram que os atos normativos poderiam ser classificados por tema.

3.1.5 Julgamentos de relevância de pesquisa de atos normativos

A partir da análise dos logs de pesquisa e das entrevistas com os usuários, elaborou-se uma lista de julgamentos de relevância de pesquisa de atos normativos que será utilizada para aferição da qualidade dos modelos de recuperação da informação propostos. A lista de julgamentos de pesquisa encontra-se na tabela 3.4.

Ressalta-se que foram considerados relevantes somente os atos normativos principais, ou seja, os atos normativos que alteram o principal não foram considerados como resultado esperado na lista de julgamentos de relevância. Esta decisão foi tomada para simplificar a avaliação, caso contrário seria necessário utilizar alguma métrica em que o nível de importância do documento fosse adotado.

Tabela 3.4 – Julgamentos de pesquisa de atos normativos

	Termo de pesquisa	Normas esperadas (Número/ano e chave do ato normativo)
1	teletrabalho	101/2019 (ATO-NORMATIVO-1342)
2	trabalho fora das dependências	
3	trabalho remoto	
4	feriados 2019	31/2019 (ATO-NORMATIVO-1081)
5	feriados 2017	115/2017 (ATO-NORMATIVO-134102)
6	feriados	10/2020 (ATO-NORMATIVO-2283)
7	regimento interno	155/2002 (ATO-NORMATIVO-46063)
8	regimento interno do tcu	
9	jornada de trabalho	100/2019 (ATO-NORMATIVO-1341) 383/2018 (ATO-NORMATIVO-1022) 138/2008 (ATO-NORMATIVO-71107)
10	organização de processos de controle externo	259/2014 (ATO-NORMATIVO-120268) 292/2018 (ATO-NORMATIVO-134201)

11	parecer de controle interno	180/2019 (ATO-NORMATIVO-2241) 63/2010 (ATO-NORMATIVO-8621)
12	suprimento de fundos	193/2018 (ATO-NORMATIVO-361) 117/2016 (ATO-NORMATIVO-131829)
13	adiantamento de numerário	
14	compra de pequeno valor	
15	atribuição de cargos do tcu	154/2002 (ATO-NORMATIVO-46060)
16	atribuições de cargos do tcu	
17	estrutura do tcu	305/2018 (ATO-NORMATIVO-1041) 2/2019 (ATO-NORMATIVO-1061)
18	competências das unidades	
19	licença capacitação	316/2019 (ATO-NORMATIVO-2022) 212/2008 (ATO-NORMATIVO-71739)
20	recesso	315/2019 (ATO-NORMATIVO-2021)
21	recesso de fim de ano	
22	monitoramento de deliberações	265/2014 (ATO-NORMATIVO-124279)
23	monitoramento de decisões	
24	acompanhamento decisões	
25	tomada de contas especial	71/2012 (ATO-NORMATIVO-111822) 155/2016 (ATO-NORMATIVO-134058)
26	avaliação de desempenho	307/2019 (ATO-NORMATIVO-2002)
27	diárias e passagens	443/2018 (ATO-NORMATIVO-1021)
28	remoção	69/2017 (ATO-NORMATIVO-134094) 286/2017 (ATO-NORMATIVO-134088)
29	distribuição de processos	175/2005 (ATO-NORMATIVO-50587) 304/2018 (ATO-NORMATIVO-742)
30	assistência pré-escolar	642/1996 (ATO-NORMATIVO-45636) 363/2019 (ATO-NORMATIVO-2181)

31	contratação de serviços	444/2018 (ATO-NORMATIVO-1028)
32	horário de funcionamento do tcu	396/2019 (ATO-NORMATIVO-2322) 11/2020 (ATO-NORMATIVO-2321) 138/2008 (ATO-NORMATIVO-71107)
33	horário de funcionamento do tribunal de contas	
34	prestação de contas dirigente máximo	180/2019 (ATO-NORMATIVO-2241)
35	prestar contas dirigente máximo	
36	relatório de gestão	378/2019 (ATO-NORMATIVO-2221) 308/2019 (ATO-NORMATIVO-2003) 182/2019 (ATO-NORMATIVO-1603) 180/2019 (ATO-NORMATIVO-2241) 62/2019 (ATO-NORMATIVO-1201) 178/2019 (ATO-NORMATIVO-2101) 63/2010 (ATO-NORMATIVO-86212)
37	140/2014	140/2014 (ATO-NORMATIVO-122680)
38	63/2010	63/2010 (ATO-NORMATIVO-86212)
39	155/2002	155/2002 (ATO-NORMATIVO-46063)

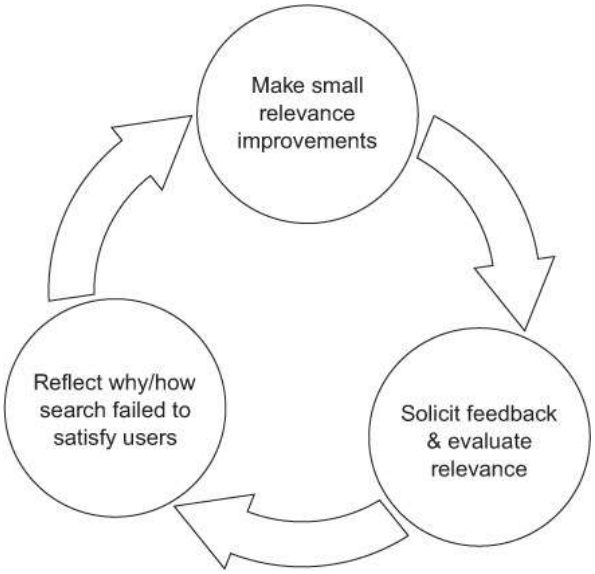
3.2 PREPARAÇÃO DOS DADOS

Após o estudo teórico dos tópicos relacionados à recuperação da informação e o entendimento dos dados (base de atos normativos), realizou-se um conjunto de implementações no sistema de busca de atos normativos com o objetivo de gerar uma pesquisa ranqueada por relevância com identificação de relações semânticas de texto.

Relevância é um conceito subjetivo e bastante dependente das necessidades/expectativas do usuário e do negócio. Por isso, o trabalho de melhoria de relevância em pesquisa tem uma natureza iterativa e que requer falha rápida (*fail fast*). É iterativa no sentido de que o foco é entregar rapidamente uma solução em produção para avaliação. Requer falha rápida para antecipar desvios pequenos e imediatos como o melhor mecanismo para ajustar o curso de ação e evitar falhas mais graves. TURNBULL e BERRYMAN (2016) sugerem um ciclo iterativo para

melhoria de relevância: pequenos incrementos são entregues para avaliação, *feedback* é coletado e o curso de ação é ajustado repetidamente.

Figura 3.4 – Iteração ideal para melhoria de relevância de pesquisa: implementar ajustes simples, solicitar feedback e falhar rapidamente para melhorar continuamente



Fonte: TURNBULL e BERRYMAN (2016)

No entanto, para este trabalho, não será possível seguir este processo com envolvimento do usuário por restrições de tempo. Por isso, serão realizados ciclos de implementação, avaliação objetiva por meio de métricas e ajustes. As implementações serão descritas juntamente com sua fundamentação e motivação.

Assim, a partir do que foi estudado, sabe-se que as seguintes estratégias podem ser usadas e combinadas com o objetivo de melhorar a relevância de pesquisa:

- Otimização da análise dos tokens (modelagem de *features*)
- Modelagem de *signals*
- Manipulação da função de ranqueamento
- Expansão de consultas

Todos os códigos desenvolvidos neste trabalho estão disponíveis no GitHub (REQUENA, 2020).

3.2.1 Primeira Iteração

A intenção, na primeira iteração, foi implementar um modelo ranqueado básico, já que a solução da pesquisa de atos normativos que está em produção não oferece classificação por relevância. Assim, o foco neste primeiro ciclo foi:

1. Melhoria do processo de análise de tokens (modelagem de *features*)
2. Nova modelagem da base conforme necessidades de negócio identificadas (modelagem de *signals*)
3. Manipulação da função de ranqueamento

3.2.1.1 Modelagem de *features*

Conforme comentado na seção 2.8, uma análise adequada do texto é a fundação de uma pesquisa de qualidade. Uma vez que a análise controla a correspondência dos termos buscados nos documentos, ela precisa ser realizada com qualidade para que os usuários encontrem a informação que buscam. Caso contrário, enfrentarão muitas dificuldades.

A partir da análise dos logs de pesquisa e das entrevistas com usuários, percebe-se que há problemas com o *pipeline* de análise atual da base de atos normativos. O fato de que 84% das consultas serem por "*" e pelo alto índice do uso de filtros sugere dificuldades na recuperação de documentos por termos. Além disso, 87,5% dos usuários entrevistados afirmaram ter dificuldade em localizar um ato normativo por termos.

Avaliando o *pipeline* de análise que está sendo utilizado, verifica-se que as variações das palavras (conjugações verbais e plurais) não estão sendo tratadas. Uma alternativa comum para resolver este problema é de *stemming* para português, o que já é oferecido pelo Solr para português do Brasil. No entanto, há a desvantagem de diminuir a precisão da recuperação (MANNING et al., 2008).

O uso da lematização poderia ser uma alternativa ao *stemming*, para transformar as conjugações verbais em verbos no infinitivo e tratar os plurais de substantivos e conectores do Português do Brasil. No entanto, não há uma solução de lematização para português do Brasil integrada ao Solr. Após buscas por alternativas na Internet, identificou-se uma biblioteca chamada CoGroo (Corretor Gramatical acoplável ao LibreOffice), que é o corretor gramatical utilizado pelo *Libre Office* e *Open Office*. É um software livre distribuído sob a licença GPL-3.0 que foi inicialmente desenvolvido pelo IME/USP e que hoje possui uma comunidade de desenvolvedores que colaboram para sua evolução e manutenção.

O Cogroo está desenvolvido em Java e oferece a funcionalidade de lematização de texto. O resultado gerado é de boa qualidade, conforme exemplo abaixo.

Texto:

Dispõe sobre os horários de funcionamento e de atendimento do Tribunal de Contas da União e sobre a gestão da frequência dos servidores

Texto após análise do Cogroo:

dispor sobre o horário de funcionar e de atender de o Tribunal de Contas da União e sobre o gestão de a frequência de o servir

Dada a boa qualidade da análise, tentou-se, portanto, implementar um *plugin* para integração entre Cogroo e Solr, por meio da criação de um filtro a ser usado no *pipeline* de análise do texto. No entanto, a implementação não foi bem-sucedida.

Foram duas as principais dificuldades enfrentadas. A primeira delas é que, para melhor resultado, o Cogroo precisa analisar toda a sentença e não somente um *token*, pois ele realiza *part-of-speech tagging*. A lógica padrão do processo de análise do Solr ocorre com uma análise *token* a *token*. Mas é possível sim analisar toda a sentença no Solr e esta dificuldade pôde ser superada por meio da criação de *cache* do *stream* de *tokens* de tal forma a fornecer para o Cogroo o texto completo.

Apesar do *cache* resolver o problema da análise da sentença completa, o Cogroo pode gerar resultados que aumentam o número de *tokens* da sentença analisada. Por exemplo, para a

sentença “As bases de dados **dos** órgãos de controle estaduais são compartilhadas.”, o Cogroo analisa em “o base de dado **de o** órgão de controle estadual ser compartilhar”. Ao aumentar o número de *tokens*, cria-se uma dessincronia com o texto original. Essa falta de sincronia impacta o *highlight* do texto. Para fazer o *highlight* dos termos encontrados no texto, é necessário conhecer as posições dos termos no documento. Após realizar uma pesquisa, o Solr obtém as posições dos termos localizados no documento e inclui marcações no texto do documento retornado que indicam quais foram as correspondências. Quando a posição do termo indexado está incorreta, o *highlight* gera resultados incorretos. Tentou-se, portanto, corrigir os *offsets* dos termos após a análise do Cogroo de tal forma que as posições corretas fossem armazenadas no índice. No entanto, após um bom período de tentativa, não houve sucesso e essa alternativa de solução precisou ser abortada por restrição de tempo. Acredita-se que é possível realizar este ajuste dos *offsets* dos termos após a análise do Cogroo, por isso este será um trabalho indicado como futuro.

Uma outra solução para uso da lematização seria analisar o texto antes de enviar para a indexação do Solr. No entanto, essa alternativa também prejudicaria o *highlight* pois haveria dessincronia com o documento original.

Há um cuidado em manter o correto *highlight* do texto pois é uma funcionalidade muito apreciada pelos usuários da plataforma de pesquisa textual. Poder caminhar pelo documento todo identificando os termos que foram localizados é algo que o usuário da plataforma valoriza muito, apesar de ser uma funcionalidade cara em termos de armazenamento e processamento.

Após estas tentativas de se incorporar o Cogroo ao pipeline de análise do Solr, percebeu-se que há um risco relacionado à lematização de todo o texto do documento que é uma possível falta de sincronicidade entre a lematização do documento e a da consulta. Essa falta de sincronicidade não é desejada, pode gerar resultados que inviabilizam a localização do documento. Uma alternativa, portanto, seria a lematização *token* a *token*. No entanto, a qualidade do resultado teria que ser avaliada.

Assim, por causa das dificuldades de integração com o Solr e da qualidade duvidosa do resultado da lematização *token* a *token*, optou-se pelo uso da radicalização do texto utilizando o *BrazilianStemFilter* do Solr.

Após a análise dos textos dos atos normativos, verificou-se que o *BrazilianStemFilter* possui um problema ao tratar plurais de palavras terminadas em “ão”. Por exemplo, atribuição é reduzida para “atribuica” e atribuições para “atribuico”.

Por isso, foi necessário incorporar ao *pipeline* de análise um passo que resolve essas inconsistências por meio de um filtro de sinônimos que mapeia um radical a outro. Por exemplo, ao encontrar o token “atribuico”, ele converte para “atribuica” (atribuico => atribuica).

O pipeline de análise final é o descrito abaixo.

```
<charFilter class="solr.HTMLStripCharFilterFactory"/>

<tokenizer class="solr.WhitespaceTokenizerFactory"/>

<filter class="solr.KeywordMarkerFilterFactory" protected="stemprotected.txt"/>

<filter  class="solr.WordDelimiterGraphFilterFactory"  types="delimeter_rules.txt"  gener-
ateNumberParts="0" splitOnCaseChange="0" generateWordParts="1" splitOnNumerics="0" preserve-
Original="1" catenateWords="1"/>

<filter class="solr.FlattenGraphFilterFactory"/>

<filter class="solr.BrazilianStemFilterFactory"/>

<filter class="solr.LowerCaseFilterFactory"/>

<filter class="solr.ASCIIFoldingFilterFactory"/>

<filter class="solr.SynonymGraphFilterFactory" synonyms="tratamento_plurais.txt"/>

<filter class="solr.FlattenGraphFilterFactory"/>

<filter  class="solr.PatternReplaceFilterFactory"  pattern="(^[^a-zA-Z0-9]*|[^a-zA-Z0-9]*$)"
replacement=""/>

<filter class="solr.PatternReplaceFilterFactory" pattern="([\d]+)\." replacement="$1"/>

<filter  class="solr.PatternReplaceFilterFactory"  pattern="^([0]+)([1234567890]{1,3})([/-
])([\d]+)" replacement="$2$3$4"/>

<filter class="solr.StopFilterFactory" words="stopwords.txt" ignoreCase="true"/>

<filter class="solr.PatternReplaceFilterFactory" pattern="^([0]+)([\d]+)" replacement="$2"/>
```

Após a remoção de *tags* HTML (*solr.HTMLStripCharFilterFactory*), o documento é quebrado em termos pelo *solr.WhitespaceTokenizerFactory*. Os termos listados no arquivo *stemprotected.txt* são protegidos do processo de *stemming* (*solr.KeywordMarkerFilterFactory*) e as palavras com delimitadores (hífen, barra, etc.) são tratadas (*solr.WordDelimiterGraphFil-*

terFactory). Os termos resultantes são radicalizados (*solr.BrazilianStemFilterFactory*), reduzidos para letras minúsculas (*solr.LowerCaseFilterFactory*) e sem acentuação (*solr.ASCIIFoldingFilterFactory*). O filtro de sinônimo trata os plurais incorretamente radicalizados pelo *stemming* (*solr.SynonymGraphFilterFactory*). Os caracteres especiais no início e fim do termo são removidos e a máscara de números como CPF e CNPJ são removidas. Estas operações utilizam o filtro *solr.PatternReplaceFilterFactory* que aplica expressões regulares ao termo. As *stop words* são removidas (*solr.StopFilterFactory*) e os zeros à esquerda de números também são removidos (*solr.PatternReplaceFilterFactory*).

Para ilustração, segue um exemplo de análise utilizando este *pipeline.*

Texto:

Dispõe acerca das prestações de contas anuais da Administração Pública Federal referentes ao exercício de 2019, que devem ser apresentadas em 2020, especificando a forma, os elementos de conteúdo, as unidades que devem prestar contas e os prazos de apresentação, nos termos do art. 3º da Instrução Normativa-TCU 63, de 1º de setembro de 2010.

Resultado da análise:

disp acerc prestaca cont anu administr public federal referent exercici 2019 dev apresent 2020 especific form element conteud unidad dev prest cont praz apresent term art 3 instruca normativa-tcu normativatcu normat tcu 63 1 setembr 2010

Em comparação com a solução de pesquisa atualmente em produção, esta análise inova somente em relação à redução das variações das palavras da língua portuguesa por meio do *stemming*.

3.2.1.2 Modelagem de *signals*

A base de atos normativos em produção tem 23 campos pesquisáveis. A modelagem desta base textual está muito próxima à da base relacional que atende o sistema de cadastramento dos atos normativos do TCU. Isso é problemático, pois como a relevância é um cálculo

com medidas extraídas dos campos, campos pesquisáveis desnecessários podem enviesar o cálculo indesejadamente.

Assim, a partir do entendimento das expectativas de relevância do usuário, ficou claro que um ato normativo, do ponto de vista do usuário, é composto por número, ano, ementa e texto. Para ele, se a informação for localizada na ementa, isso torna o ato normativo mais relevante. O conteúdo do ato (texto) é percebido como algo único e o número e ano são importantes pois muitos o sabem de cor.

Assim, a proposta é ter três campos pesquisáveis - **Título**, **Ementa** e **Texto**. O Título é um *signal*, importante para capturar pesquisas por número e ano, muito comuns conforme detectado no log de pesquisas. A Ementa também representa um *signal*, pois mede informação importante para o usuário que é o assunto do ato. Será um campo pesquisável à parte, pois corresponde com a noção de relevância do usuário, de que se a informação estiver presente na Ementa, isso torna o ato normativo mais relevante. O Texto é todo o conteúdo do ato e será um campo que contém todas as informações do ato concatenadas (tipo, número, ano, título, situação, texto do documento e anexos).

Será criado um novo campo que indica se o ato é um novo ato normativo, ou seja, se ele traz conteúdo novo. Na base de atos normativos, temos os atos que trazem conteúdo novo e atos que alteram outros atos. O conteúdo do ato que altera é sempre replicado no texto do ato original, ou seja, seu conteúdo também está presente no ato alterado. Assim, a informação pode ser localizada no ato original, o que faz com que os atos que alteram representem ruído no resultado de pesquisa. Assim, identificar que o ato é novo ou não é importante para que esse fato possa ser utilizado no cálculo de relevância.

Para esta identificação, foi feito um processamento simples que marca o ato como novo se a ementa dele não começa com “altera” ou “convalida”. Adotou-se esta heurística a partir da observação de várias ementas de atos que alteram. É claro que ela é falha e que seria interessante o uso de técnicas com maior acurácia, como, por exemplo, um classificador de atos normativos. No entanto, o custo para se gerar tal classificador é alto, uma vez que não há dados anotados. Fica como sugestão de trabalho futuro desenvolver uma técnica para melhor detectar atos que inovam em matéria.

O novo campo gerado - **Novo Ato** - será utilizado no modelo de ranqueamento fornecendo um peso extra caso o ato seja novo.

Além disso, os campos de **data de publicação** do ato e a **situação** serão utilizados pois fornecem informações importantes do negócio que influenciam a expectativa de relevância do usuário.

Em suma, o modelo utilizará os seguintes campos:

- Título
- Texto
- Ementa
- Situação
- Novo Ato
- Data de Publicação

3.2.1.3 Manipulação da função de ranqueamento

A similaridade que será utilizada é o BM25. Ela será manipulada por meio de pesos (*boosts*) nos campos.

Os valores adotados para os parâmetros k e b do BM25 foram os valores que o Solr utiliza por padrão: k = 1.2 e b = 0.75. ROBERTSON e ZARAGOZA (2009) afirmam que um número significativo de experimentos foram realizados e que sugerem que os valores de b entre 0.5 e 0.8 e de k entre 1.2 e 2 são suficientemente bons para a grande maioria dos cenários. Portanto, optou-se por, nesta iteração, manter estes valores padrão.

De forma experimental, o Título recebeu um *boost* de 10, o Texto de 2 e Ementa de 3. Estes números refletem a importância para a relevância de quando a informação é encontrada em cada um dos campos. Assim, se o usuário entra com número/ano na pesquisa, essa informação será localizada no título e por isso terá o seu escore multiplicado por 10 para que este documento seja priorizado em relação a outros que possuem o número/ano no texto do ato. No caso de consultas frasais, ou seja, entre aspas, o peso será o dobro: Título recebe *boost* de 20, Texto de 4 e Ementa de 6.

Os atos normativos em situação diferente de revogado terão seu escore multiplicado por 2, pois grande parte dos usuários entrevistados disseram que os atos revogados são menos importantes.

Um *boost* por data também foi aplicado. No entanto, a pesquisa de atos normativos apresenta uma particularidade: nem sempre os documentos mais recentes serão mais relevantes que os mais antigos. Por exemplo, uma busca por "regimento interno do tcu" possui uma clara intenção que é localizar o ato normativo 155/2002 - Regimento Interno do TCU. No entanto, este ato é antigo (2002) e há vários outros atos recentes que alteram o regimento interno. Neste caso, a data de publicação não é tão importante. Por outro lado, há atos que dispõem sobre determinada matéria que foram substituídos por atos novos que revogam os anteriores tacitamente. E, para este cenário, a data é importante.

Assim, para lidar com esta necessidade, foi adotada a função **recip** do Solr e experimentou-se um conjunto diferente de parâmetros para a função. A função recip é uma função recíproca do tipo a/(m*x+b), onde m, a, b são constantes e x é uma função complexa arbitrária.

Após algumas experimentações de parâmetros, adotou-se os valores a=b=12 e m=3.16e-11. Ressalta-se que valores maiores de a e b suavizam a influência da data no *boost* (ver figura 3.5).

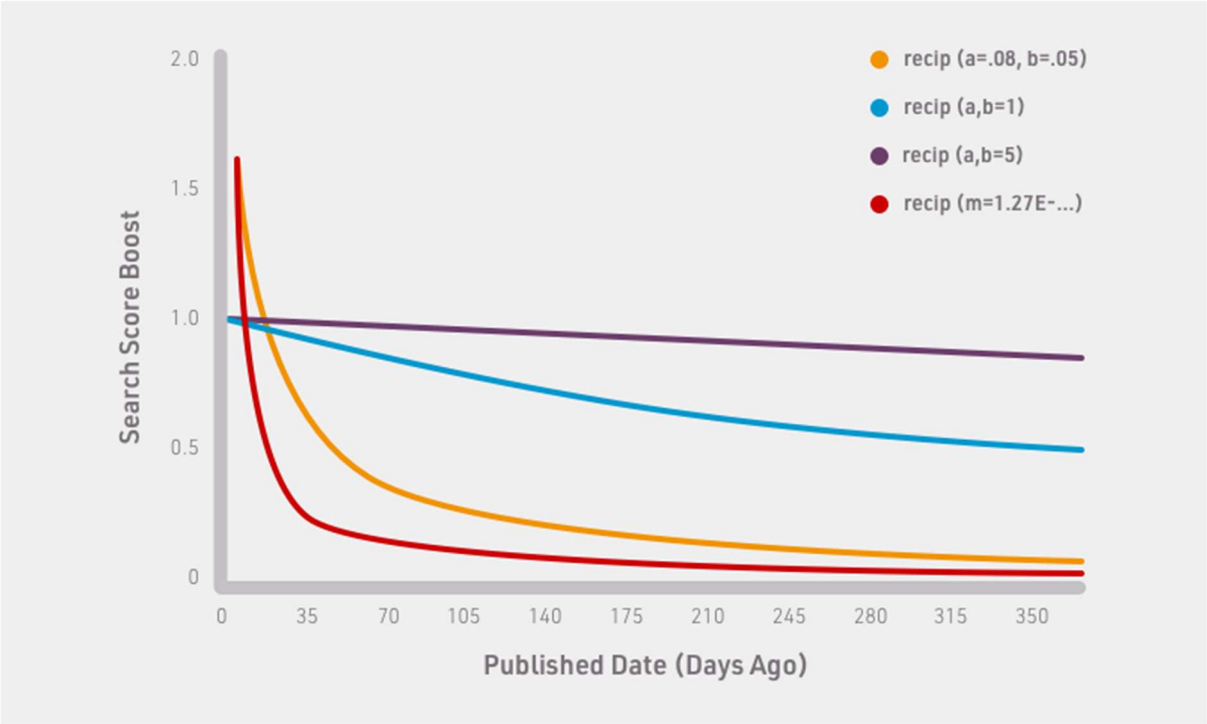
Em suma, a função de ranqueamento será influenciada da seguinte maneira:

- *boost* por campo:
    a. TITULO^10
    b. TEXTO^2
    c. EMENTA^3
- *boost* por campo em consultas frasais:
    a. TITULO^20
    b. TEXTO^4
    c. EMENTA^6
- *boost* multiplicativo:
    a. if(termfreq(SITUACAO,'Revogada'),1,2)

b. if(termfreq(NOVOATO,'N'),1,2)

c. recip(ms(NOW,DATAEXPEDICAO),3.16e-11,12,12)

Figura 3.5 – Gráfico da função recip do Solr para conjuntos diferentes de parâmetros



Fonte: TOUPIN et al. (2018)

3.2.2 Segunda Iteração

Na segunda iteração, o foco foi implementar a expansão de consultas com o objetivo de identificar similaridade semântica entre termos. O objetivo é melhorar a revocação da pesquisa. O modelo base utilizado foi o gerado pela primeira iteração.

Uma técnica comum é a expansão por sinônimos por meio do uso de um filtro de sinônimos no pipeline de análise. Este filtro pode utilizar um tesauro mantido por editores humanos ou um tesauro gerado automaticamente.

O TCU possui um tesauro que é chamado de Vocabulário de Controle Externo - VCE. Este tesauro é atualmente utilizado pelas pesquisas de Jurisprudência Selecionada e Súmulas para expansão. A expansão de consultas pelo VCE é útil e oferece expansões de alta qualidade,

mas é limitada ao se restringir ao sentido denotativo da linguagem. Ou seja, a conotação, que é definida pelo contexto em que a palavra aparece, não é levada em consideração em dicionários. Por isso, neste trabalho, deseja-se implementar a expansão de consultas por meio da geração automática de um tesauro que capture a similaridade das palavras considerando o seu contexto.

Assim, optou-se por utilizar o algoritmo Word2Vec para aprender as representações de palavras, por ser uma boa alternativa para aprender relações que são sensíveis ao contexto (TEOFILI e MATTMANN, 2019).

O primeiro passo foi treinar o Word2Vec com os dados dos atos normativos. Optou-se inicialmente por usar um modelo gerado com os dados da própria base e não utilizar *embeddings* previamente treinados. Segundo DIAZ et al. (2016), a expansão de consulta por meio de *embeddings* locais melhoram a performance, enquanto utilizar *embeddings* treinados com dados que possuem linguagem muito distinta dos documentos a serem recuperados pode prejudicar a recuperação da informação. Os autores ressaltam também que pode ser benéfico aumentar os dados de treinamento dos *embeddings* com outros dados que possuem linguagem semelhante.

Para o treinamento, os atos normativos foram pré-processados somente para conversão para letras minúsculas e remoção de acentuação. O algoritmo foi treinado com 100 dimensões na versão CBOW.

Conforme pontuado por TEOFILI e MATTMANN (2019), é preciso considerar alguns aspectos antes de usar os resultados do Word2Vec em expansão de consulta:

1. Se os *embeddings* forem usados durante o processo de indexação (para cada termo do índice, obter as palavras próximas), o índice pode crescer muito e isso pode não ser aceitável do ponto de vista de performance, tanto em relação a tempo de execução quanto armazenamento. Estratégias como a expansão para somente alguns papéis semânticos, a partir de análises de PoS (*Part of Speech*) *Tagger*, ou expansão que considere o peso dos termos.
2. É interessante utilizar os resultados do Word2Vec que tiverem um bom escore de similaridade. Por exemplo, se a similaridade do cosseno for utilizada para obter os vizinhos mais próximos, o resultado pode ser uma baixa similaridade, mesmo que sejam os mais próximos. Por isso, é interessante ter um limite mínimo.

3. Sinônimos *versus* antônimos: palavras antônimas podem ser identificadas pelo algoritmo como similares. Em alguns contextos, isso não é relevante. No entanto, há contextos em que isso é indesejado. A similaridade entre vetores de palavras pode ajudar a excluir aqueles casos em que não há "similaridade suficiente". Uma estratégia é ter um valor mínimo fixo de similaridade (por exemplo, valores acima 0,5) e utilizar as palavras que possuem uma similaridade de até 0.1 menor que a de maior similaridade.

Assim, após o treinamento do modelo, a primeira tentativa para incorporar os resultados do Word2Vec ao motor de busca foi gerar um arquivo com o mapeamento dos sinônimos e utilizar este arquivo em um filtro de sinônimos do Solr. Foram considerados como sinônimos somente aqueles termos similares com valor mínimo de 0,5 e que a diferença com o termo de maior similaridade fosse até 0,1. Segue um trecho do arquivo gerado (o número após o | é a similaridade do cosseno):

procuradoria=>agu|0.727438 mpf|0.708749 advocacia|0.667032 controladoria|0.642001

telegrafos=>ect|0.923806 correios|0.915287

economicidade=>eficacia|0.916901 eficiencia|0.859981 legitimidade|0.827532

cargos=>ocupantes|0.981139 funcoes|0.975228 efetivos|0.974050 comissionadas|0.971335 vagos|0.970461 confianca|0.970018 alocadas|0.951903 carreira|0.947097 cargo|0.946785 ocupados|0.940241

ritcu=>8.906|0.997460 295|0.997410 10.522|0.997273 supressiva|0.996968 fundamentadas|0.996917 regido|0.996745 260|0.996191 fiscalizara|0.995916 analogo|0.995886 ocasionara|0.995600

suprimentos=>operados|0.988290 juntos|0.985135 tornarao|0.985101 economicas|0.983808 enquadram|0.982255 saber|0.981127 honra|0.980962 universidades|0.980337 estabelecimentos|0.979438 esportivos|0.979211

naval=>navais|0.976564 exercito|0.970636 comando|0.964574 marinha|0.960176 aeronautica|0.959147 consolidando|0.957583 militares|0.953937 fuzileiros|0.952422 centro|0.948566 mercante|0.946001

devedores=>duvidosos|0.994026 cruzeiros|0.992409 cz$|0.989929 descontado|0.989395 somado|0.988911 despendido|0.986909 1%|0.986600 retidos|0.986591 etaria|0.986204 realizavel|0.985230

prestar=>inspecoes|0.954812 gestores|0.949151 instancias|0.946025 acompanhamentos|0.937110 concernentes|0.936525 competentes|0.930358 previsao|0.930340 inclusive|0.927417 auditorias|0.925145 relativas|0.924336

Apesar dos resultados serem interessantes, há alguns problemas. As variações das palavras, em alguns casos, foram capturadas pelo modelo, a exemplo de naval => navais e cargos => cargo, mas para outros não (prestar e prestação). Este problema poderia ser contornado com o uso de *stemming*, por exemplo.

No entanto, outro problema são expressões formadas por mais de uma palavra, como "Tribunal de Contas da União". Se os termos forem expandidos isoladamente, não alcançaremos o sinônimo "TCU".

Assim, por ser uma solução de expansão estática, com processamento termo a termo, ou seja, que não considera todos os termos da consulta para identificar similaridade, os resultados não foram bons. A expansão muitas vezes ocorria em direções não desejadas, agregando termos sem similaridade com o contexto da consulta. Inclusive há um estudo feito por ZUCCON et al. (2015) que afirma que é altamente desejável levar em consideração o contexto da consulta para a recuperação.

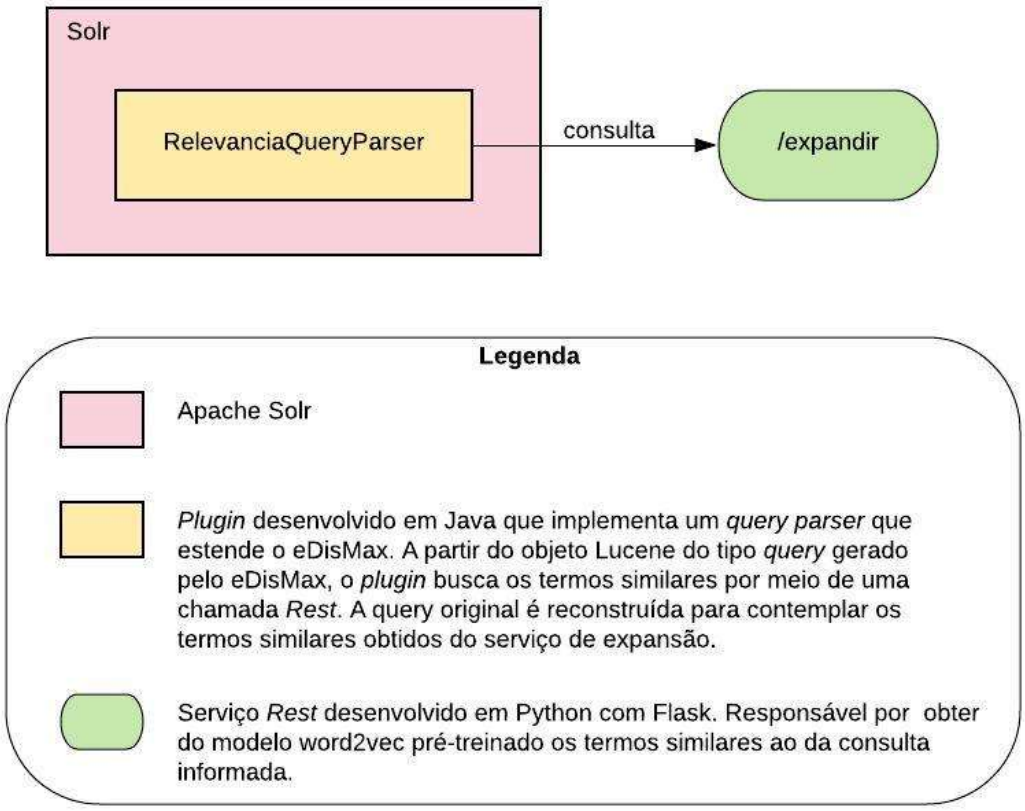
Uma evolução desta alternativa seria expandir a consulta com as palavras similares a todos os termos da consulta. Desta forma, conseguiria-se um resultado mais alinhado ao contexto da consulta. Isso requer que o modelo Word2Vec seja consultado em tempo de consulta ao invés de uma expansão estática. Uma solução para esta operação foi documentada por TEOFILI e MATTMANN (2019), que integra o Word2Vec com o Solr por meio do DeepLearning4J. A sugestão dos autores é a construção de um filtro customizado para o Lucene que é invocado no *pipeline* de análise da consulta. Este filtro utiliza o modelo Word2Vec pré-treinado para realizar a expansão. A escolha dos termos similares utiliza a mesma estratégia descrita anteriormente - termos cuja similaridade seja até 0,1 menor que a maior similaridade e um limite mínimo de 0,5.

O *plugin* foi construído e integrado ao Solr. Porém, a performance do motor de busca foi muito prejudicada. Quando o filtro que usa o DeepLearning4J era acionado, o Solr parava de responder. Tentou-se aumentar a quantidade de RAM de 8GB para 16GB, mas sem sucesso. Por isso, essa alternativa foi abortada.

A terceira alternativa foi utilizar o Python para treinar e carregar o modelo Word2Vec e a integração com o Solr ocorrer por meio da chamada a serviços REST. Para isso, foi construído um novo *query parser* para o Solr, chamado de RelevanciaQueryParser. Este parser é

uma extensão do eDisMax e ele modifica a consulta gerada pelo eDisMax ao adicionar novas cláusulas com o resultado da similaridade do Word2Vec.

Figura 3.6 – Arquitetura do plugin RelevanciaQueryParser



Foi implementada também melhorias no treinamento do modelo Word2Vec. O pré-processamento dos dados de treinamento foi alterado para contemplar a radicalização (*stemming*) a fim de reduzir as variações de linguagem, seguindo o mesmo *pipeline* de análise do Solr (3.2.1.1). Além disso, os dados de treinamento foram processados para a identificação dos conceitos contidos no texto, por meio da identificação de bigramas e trigramas utilizando *Pointwise Mutual Information* (PMI). O modelo Word2Vec foi treinado com os bigramas e trigramas, também com 100 dimensões e na versão CBOW.

Abaixo segue exemplo da expansão da consulta por **regimento interno do tcu** realizada pelo RelevanciaQueryParser:

Query gerada pelo eDisMax antes da expansão:

```
FunctionScoreQuery(+
        (((TITULO:regiment)^10.0 | (TEXTO:regiment)^2.0 | (EMENTA:regiment)^3.0)
        ((TITULO:intern)^10.0 | (TEXTO:intern)^2.0 | (EMENTA:intern)^3.0)
        ((TITULO:tcu)^10.0 | (TEXTO:tcu)^2.0 | (EMENTA:tcu)^3.0))
        ((TITULO:"regiment intern ? tcu")^20.0 | (TEXTO:"regiment intern ? tcu")^4.0
        | (EMENTA:"regiment intern ? tcu")^6.0),
        scored by boost(product(if(termfreq(COPIASITUACAO,Revogada),
        const(1),const(2)),if(termfreq(NOVOATO,N),const(1),const(2)),12.0/(3.16E-
        11*float(ms(const(1582372740130),date(DATAEXPEDICAO)))+12.0))))
```

Query expandida pelo RelevanciaQueryParser:

```
FunctionScoreQuery(+
        (((TITULO:regiment)^10.0 | (TEXTO:regiment)^2.0 | (EMENTA:regiment)^3.0)
        ((TITULO:intern)^10.0 | (TEXTO:intern)^2.0 | (EMENTA:intern)^3.0)      ((TI-
        TULO:tcu)^10.0 | (TEXTO:tcu)^2.0 | (EMENTA:tcu)^3.0))
        ((TITULO:"regiment intern ? tcu")^20.0 | (TEXTO:"regiment intern ? tcu")^4.0
        | (EMENTA:"regiment intern ? tcu")^6.0)
        EMENTA:tribunal TEXTO:tribunal EMENTA:dest TEXTO:dest EMENTA:term TEXTO:term
        EMENTA:"preench requisit admissibil" TEXTO:"preench requisit admissibil"
        EMENTA:"tribunal cont unia" TEXTO:"tribunal cont unia",
        scored by boost(product(if(termfreq(COPIASITUACAO,Revogada),
        const(1),const(2)),if(termfreq(NOVOATO,N),const(1),const(2)),12.0/(3.16E-
        11*float(ms(const(1582372740130),date(DATAEXPEDICAO)))+12.0))))
```

Observa-se que, ao treinar o modelo Word2Vec com bigramas e trigramas, é possível que a expansão da consulta gere um resultado interessante como o da consulta exemplo acima: a expansão inclui uma consulta frasal por “tribunal cont unia”, que foi um trigrama identificado que se relaciona com tcu.

3.2.3 Terceira Iteração

Na terceira iteração, o foco foi investir na análise de *tokens* para a extração de ideias ao invés de simples extração de termos. Assim, este ciclo investiu em:

1. Modelagem de *features*, para a extração de ideias do texto e da consulta
2. Modelagem de *signals*: os documentos que contém as ideias relacionadas à consulta podem ser priorizados
3. Manipulação da função de relevância para considerar as ideias identificadas

Para a extração das ideias do texto, utilizou-se o PMI no processo de indexação. O modelo utilizado foi o mesmo do treinamento do Word2Vec da segunda iteração. Os conceitos (bigramas e trigramas) foram extraídos do texto e foram indexados em um campo à parte. A análise da EMENTA gerou os conceitos que foram indexados no campo EMENTA_CONCEITOS e a análise do TEXTO gerou conceitos que foram salvos no campo TEXTO_CONCEITOS. O objetivo de gerar novos campos contendo os conceitos é reforçar, no cálculo de relevância, os documentos que possuem os conceitos associados à consulta, sejam conceitos explícitos na consulta ou conceitos similares identificados pelo Word2Vec.

A função de ranqueamento foi manipulada para conter as seguintes regras:

- os conceitos extraídos da consulta do usuário serão buscados nos campos EMENTA_CONCEITOS e TEXTO_CONCEITOS com um *boost* de 4 e 2 respectivamente.
- os conceitos similares identificados pelo Word2Vec serão buscados nos campos EMENTA_CONCEITOS e TEXTO_CONCEITOS sem *boost*.

Para a integração com o Solr, foi construído um novo *query parser* chamado ConceptQueryParser, que também estende o eDisMax modificando a consulta por ele gerada para incluir as expansões descritas. A arquitetura segue o mesmo padrão descrito na figura 3.6.

Abaixo segue exemplo da expansão da consulta por **tomada de contas especial** realizada pelo ConceptQueryParser:

Query gerada pelo eDisMax antes da expansão:

```
FunctionScoreQuery(+
        (((TITULO:tom)^10.0 | (TEXTO:tom)^2.0 | (EMENTA:tom)^3.0)
        ((TITULO:cont)^10.0 | (TEXTO:cont)^2.0 | (EMENTA:cont)^3.0)
        ((TITULO:especial)^10.0 | (TEXTO:especial)^2.0 | (EMENTA:especial)^3.0))
        ((TITULO:"tom ? cont especial")^20.0 | (TEXTO:"tom ? cont especial")^4.0 |
        (EMENTA:"tom ? cont especial")^6.0),
        scored by boost(product(if(termfreq(COPIASITUACAO,Revogada),const(1),
        const(2)),if(termfreq(NOVOATO,N),const(1),const(2)),12.0/(3.16E-11*float(
        ms(const(1582375018841),date(DATAEXPEDICAO)))+12.0))))
```

Query expandida pelo ConceptQueryParser:

```
FunctionScoreQuery(+
        (((TITULO:tom)^10.0 | (TEXTO:tom)^2.0 | (EMENTA:tom)^3.0)
        ((TITULO:cont)^10.0 | (TEXTO:cont)^2.0 | (EMENTA:cont)^3.0)
        ((TITULO:especial)^10.0 | (TEXTO:especial)^2.0 | (EMENTA:especial)^3.0))

        ((TITULO:"tom ? cont especial")^20.0 | (TEXTO:"tom ? cont especial")^4.0 |
        (EMENTA:"tom ? cont especial")^6.0)

        EMENTA:"especi cuj" TEXTO:"especi cuj" EMENTA_CONCEITOS:especi_cuj TEXTO_CON-
        CEITOS:especi_cuj

        EMENTA:"fat origin dat" TEXTO:"fat origin dat"
        EMENTA_CONCEITOS:fat_origin_dat TEXTO_CONCEITOS:fat_origin_dat

        EMENTA:"especi instaur" TEXTO:"especi instaur"
        EMENTA_CONCEITOS:especi_instaur TEXTO_CONCEITOS:especi_instaur

        EMENTA:"nrinstruca normat dou" TEXTO:"nrinstruca normat dou"
        EMENTA_CONCEITOS:nrinstruca_normat_dou TEXTO_CONCEITOS:nrinstruca_normat_dou

        EMENTA:"pel onus decorrent" TEXTO:"pel onus decorrent"
        EMENTA_CONCEITOS:pel_onus_decorrent TEXTO_CONCEITOS:pel_onus_decorrent

        EMENTA:"sobr instaur correspondent" TEXTO:"sobr instaur correspondent"
        EMENTA_CONCEITOS:sobr_instaur_correspondent TEXTO_CONCEITOS:sobr_instaur_cor-
        respondent

        (EMENTA_CONCEITOS:tom_cont)^4.0 (TEXTO_CONCEITOS:tom_cont)^2.0,

        scored by boost(product(if(termfreq(COPIASITUACAO,Revogada),
        const(1),const(2)),if(termfreq(NOVOATO,N),const(1),const(2)),12.0/(3.16E-
        11*float(ms(const(1582375018841),date(DATAEXPEDICAO)))+12.0))))
```

Observa-se uma expansão interessante neste caso, que é o conceito identificado de tomada de contas (tom_cont) e ônus decorrente (pel_onus_decorrent). Nota-se, ainda, espaço para melhoria, pois algumas palavras (pelo, sobre) não estão sendo consideradas como *stop word,* o que pode reduzir a revocação em alguns casos.

3.2.4 Quarta Iteração

Há trabalhos que apontam que a expansão de consultas por meio de *word embeddings* aumentam a revocação mas diminuem a precisão. Esta conclusão é intuitiva, uma vez que há um claro *trade-off* entre as métricas de precisão e revocação (MANNING et al., 2008).

A expansão de consulta para captura de similaridades semânticas do texto é importante justamente para aumentar a revocação. Documentos que não seriam recuperados passam a ser e novas soluções para a melhoria da precisão devem entrar em ação.

Uma alternativa é o *re-ranking* em que uma quantidade parametrizada de documentos recuperados, ranqueados seguindo algum modelo de ranqueamento, serão reordenados. Esta

reordenação pode acontecer por meio de qualquer tipo de operação. Por exemplo, pode utilizar a ponderação de diferentes métricas de similaridade (BM25 ou similaridade do cosseno) como também algum modelo de aprendizagem de máquina.

Assim, o foco da quarta iteração foi utilizar o *Learn to Rank* no Solr com o objetivo de melhorar a precisão da recuperação. A ideia foi experimentar o uso da técnica no contexto de *re-ranking*, utilizando um modelo simples que não utiliza aprendizagem de máquina.

As *features* utilizadas pelo modelo foi o vetor médio do conjunto de palavras contidas em cada parâmetro - ementa e consulta. Este vetor médio foi obtido a partir do modelo Word2Vec. No caso da ementa, o vetor médio foi indexado como um campo do documento (EMENTA_MEAN_VECTOR) e o da consulta é gerado quando ela é realizada. O modelo adotado utilizou a seguinte fórmula:

(*originalScore* / 100) + (similaridade_cosseno_consulta_ementa * 2,3) + (similaridade_cosseno_consulta_texto * 1,3)

onde *originalScore* é o valor gerado pelo BM25, a similaridade_cosseno_consulta_ementa é a similaridade do cosseno entre os vetores médios da consulta e da ementa do documento e a similaridade_cosseno_consulta_texto é a similaridade do cosseno entre os vetores médios da consulta e do texto do documento.

A ideia dessa fórmula é atender à intuição de relevância do usuário de que se os termos da consulta forem localizados na ementa do documento, isso torna o documento mais relevante.

3.3 AVALIAÇÃO

Há várias métricas para a avaliação da qualidade da recuperação de um sistema de RI. As mais básicas e frequentes são precisão e revocação, combinadas de diferentes maneiras (BAEZA-YATES et al, 2011). A escolha da métrica depende das necessidades de uso do sistema de recuperação. Em alguns cenários, por exemplo a Web, a precisão é mais importante, pois o objetivo é satisfazer rapidamente a necessidade de informação. Já a busca por documentos da área legal visa normalmente uma ampla recuperação para o estudo do assunto, o que privilegia a revocação.

Conforme explicitado anteriormente, uma boa parte dos usuários da pesquisa de atos normativos busca encontrar a norma referente a determinado assunto. Ou seja, neste cenário, uma boa precisão faz toda a diferença. Por este motivo, será utilizada uma métrica de precisão para a avaliação dos resultados. Ademais, deseja-se que essa métrica compare a qualidade para consultas isoladas, pois o uso de métricas médias de várias consultas pode encobrir anomalias importantes nos algoritmos de recuperação. A análise de performance para consultas isoladas permite também identificar grupos de consultas problemáticos.

Assim, será adotado um único valor de precisão em diferentes níveis de revocação (P@n). Esta métrica visa explicitar que uma maior concentração de documentos relevantes no topo do ranking sugere uma impressão mais positiva do usuário.

Serão utilizados os níveis de revocação 3, 5 e 10 (P@3, P@5 e P@10 respectivamente). Como uma página de resultado da Plataforma de Pesquisa Textual possui por padrão 10 documentos, essas métricas foram adotadas com o objetivo alvo de que o sistema de recuperação apresente o documento desejado na primeira página de resultado, livrando o usuário da necessidade de mudar de página.

Histogramas de precisão serão também utilizados para uma rápida comparação entre os diferentes modelos de recuperação por meio de inspeção visual.

A avaliação será conduzida a partir da dos resultados ranqueados obtidos dos diversos modelos para as consultas especificadas em 3.1.5. São um total de 39 consultas e 6 diferentes sistemas de recuperação (4 gerados por este trabalho e duas versões do sistema atualmente em produção).

Cada consulta será submetida a cada base do Solr. Os resultados serão comparados com os resultados esperados (julgamentos de pesquisa) e a precisão para cada um dos níveis de revocação é calculada seguindo a fórmula:

$$nº\ documentos\ relevantes\ no\ resultado$$
$$/\ mínimo\ (nível\ de\ revocação\ ou\ qtde\ de\ documentos\ relevantes\ nos\ julgamentos)$$

Será calculada também a média das precisões para cada nível de revocação, para ter uma avaliação sobre qual algoritmo seria preferível do ponto de vista do usuário.

4 RESULTADOS

Para a avaliação dos resultados, cada uma das iterações descritas em 3.2 - Preparação de dados - serão chamadas de alternativa A1, A2, A3 e A4 respectivamente.

Além dos modelos gerados por este trabalho, optou-se por incorporar nos resultados os modelos que hoje estão em produção na Plataforma de Pesquisa Textual para atos normativos. Estes modelos serão chamados de alternativa A0.1 e A0.2.

A alternativa A0.1 corresponde aos resultados que o usuário hoje recebe na pesquisa, ou seja, é a consulta que utiliza o *query parser* Swan com ordenação por data decrescente. O Swan é um *query parser* que foi desenvolvido pelo TCU capaz de gerar consultas para o Solr a partir de uma sintaxe de pesquisa que utiliza operadores de proximidade (ADJ e PROX), além dos operadores booleanos clássicos (OR e AND). Foi inspirado na sintaxe e funcionamento das pesquisas da antiga ferramenta BRS.

A alternativa A0.2 corresponde à consulta na base de produção utilizando o *query parser* eDisMax com ordenação por escore.

A comparação dos resultados obtidos por este trabalho com os resultados oferecidos hoje em produção tem por objetivo somente **evidenciar que há margem para melhoria nos resultados atualmente oferecidos** para o usuário. Não se tem por intuito comparar a performance com os modelos em produção, pois eles nunca foram customizados para oferecer uma pesquisa classificada por relevância.

Assim, como não há um modelo ranqueado de referência para a comparação de performance, optou-se por utilizar o modelo mais simples (A1) como referência para os demais modelos.

As métricas apresentadas nas figuras 4.1, 4.2, 4.4, 4.5, 4.7 e 4.8 mostram os valores do modelo referência (A1) em laranja. O objetivo é destacar a métrica utilizada como referência na comparação com os valores das demais alternativas.

Quando há melhoria na métrica, a célula é destacada em verde. Um resultado inferior ao modelo de referência é apresentado em vermelho. E quando não há mudança, a célula permanece com fundo branco.

Em suma, os modelos que estão sendo comparados nos resultados a seguir são:

- Alternativa A0.1:

  Modelo de produção, que considera os resultados que o usuário vê (handler /selectSwan e ordenação por data de relevância)

- Alternativa A0.2:

  Modelo de produção, utilizando query parser eDisMax com ordenação por escore (handler /select). Ressalta-se que esta configuração é da relevância *default* do Solr, sem otimizações.

- Alternativa A1:

  Modelo ranqueado por meio da similaridade BM25, análise com uso de *stemming*. *Boost* por campo TITULO^10 TEXTO^2 EMENTA^3, boost por campo em consultas frasais de TITULO^20 TEXTO^4 EMENTA^6 e *boosts* multiplicativos:

  - if(termfreq(SITUACAO,'Revogada'),1,2)
  - if(termfreq(NOVOATO,'N'),1,2)
  - recip(ms(NOW,DATAEXPEDICAO),3.16e-11,12,12)

- Alternativa A2:

  Modelo A1 com expansão de consulta por meio de *word embeddings*

- Alternativa A3:

  Modelo A2 com expansão de conceitos extraídos por meio de PMI

- Alternativa A4

  Modelo A3 com *Learn to Rank* para os primeiros 10 resultados

4.1 AVALIAÇÃO P@3

As precisões médias para o nível 3 de revocação podem ser vistas na figura 4.1.

Figura 4.1 – P@3 média. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

	A01 P@3	A02 P@3	A1 P@3	A2 P@3	A3 P@3	A4 P@3
0	0.187692	0.384359	0.683333	0.644872	0.764615	0.734615

A melhoria da recuperação da informação é evidente ao comparar a precisão média do modelo de referência (A1) com os valores dos sistemas de produção (A.01 e A.02). No entanto, observa-se que o modelo com expansão de consulta A2 perde um pouco de precisão em relação ao modelo básico adotado como referência. Já o A3 e A4, que também tem expansão de consulta, oferecem resultados um pouco melhores.

Na figura 4.2, é possível ter uma visão por consulta e alguns resultados chamam atenção.

As consultas por “trabalho remoto” e “relatório de gestão” perdem precisão em A4 pois estes termos não estão presentes na ementa do ato normativo, o que sugere que o modelo *learn to rank* deve incluir outras *features*.

Os resultados para a consulta por acompanhamento de decisões são bem interessantes. O termo acompanhamento não é o utilizado pelas normas. As normas utilizam o termo monitoramento. A consulta por monitoramento de decisões apresenta P@3 igual a 1 para todos os quatro modelos desenvolvidos - A1, A2, A3 e A4. No entanto, a consulta por acompanhamento de decisões somente obtém os documentos relevantes quando há expansão de consulta. Avaliando também os outros níveis de revocação (5 e 10), isto continua ocorrendo e os modelos de produção e de referência não conseguem recuperar os documentos relevantes até o nível 10 de revocação.

Figura 4.2 – Valores de P@3 para cada consulta. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

	A01 P@3	A02 P@3	A1 P@3	A2 P@3	A3 P@3	A4 P@3
teletrabalho	0	1	1	1	1	1
trabalho fora das dependências	0	0	1	1	1	1
trabalho remoto	0	0	1	1	1	0
feriados 2019	1	1	1	1	1	1
feriados 2017	0	1	0	0	0	0
feriados	1	0	1	1	1	1
regimento interno	0	1	0	0	1	1
regimento interno do tcu	0	0	0	0	0	0
jornada de trabalho	0	0	1	1	1	1
organização de processos de controle externo	0	0	1	1	1	1
parecer de controle interno	0.5	0	0	0	0.5	0.5
suprimento de fundos	0	0.5	0.5	0.5	0.5	0.5
adiantamento de numerário	0	0	0.5	0.5	0.5	0.5
compra de pequeno valor	0.5	0.5	1	0.5	0.5	0.5
atribuição de cargos do tcu	0	0	1	0	1	1
atribuições de cargos do tcu	0	1	1	0	1	1
estrutura do tcu	0	0	0.5	0.5	0.5	0.5
competências das unidades	0	0	0.5	0.5	0.5	0.5
licença capacitação	0.5	0	0.5	0.5	0.5	0.5
recesso	0	0	1	1	1	1
recesso de fim de ano	0	0	1	1	1	1
monitoramento de deliberações	0	1	1	1	1	1
monitoramento de decisões	0	0	1	1	1	1
acompanhamento decisões	0	0	0	1	1	1
tomada de contas especial	0	0.5	0.5	0.5	0.5	0.5
avaliação de desempenho	0	0	1	1	1	1
diárias e passagens	0	1	1	1	1	1
remocao	0	0	1	1	1	1
distribuição de processos	0	0.5	1	1	1	1
assistência pré-escolar	0.5	1	0	0	0.5	1
contratação de serviços	0	1	1	1	1	1
horário de funcionamento do tcu	0.66	0.66	0.66	0.66	1	0.66
horário de funcionamento do tribunal de contas da união	0.66	0.33	0.33	0.33	0.66	0.66
prestação contas dirigente máximo	1	0	0	0	0	0
prestar contas dirigente máximo	0	0	0	0	0	0
relatório de gestão	0	0	0.66	0.66	0.66	0.33
140/2014	1	1	1	1	1	1
63/2010	0	1	1	1	1	1
155/2002	0	1	1	1	1	1

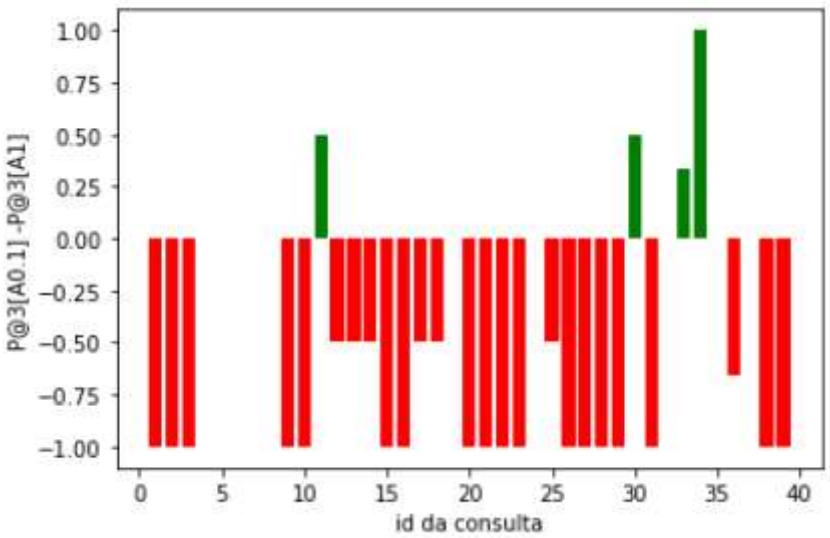
O resultado para a consulta compra de pequeno valor também chama atenção. A expansão de consulta prejudica a precisão, quando comparado com o modelo de referência.

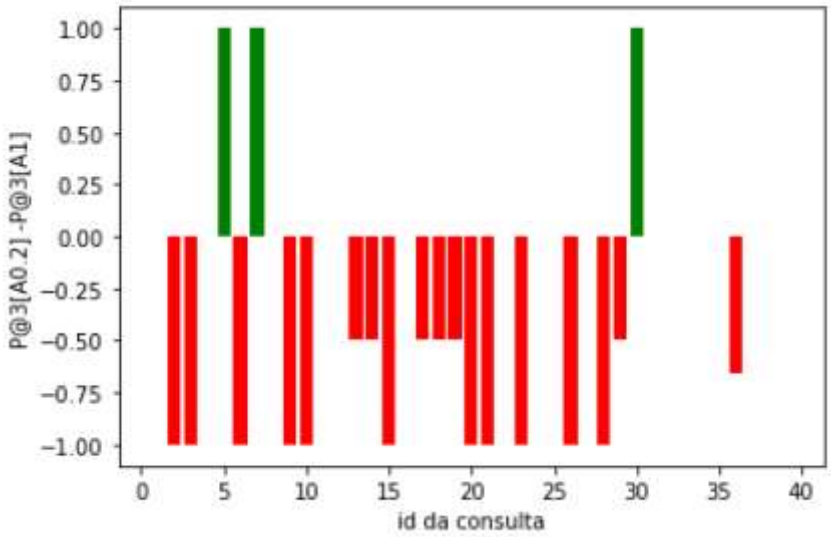
Outro aspecto interessante são as pesquisas por número/ano do ato. Essas pesquisas são extremamente comuns nesta base e todos os modelos propostos (A1, A2, A3 e A4) oferecem P@3 máxima.

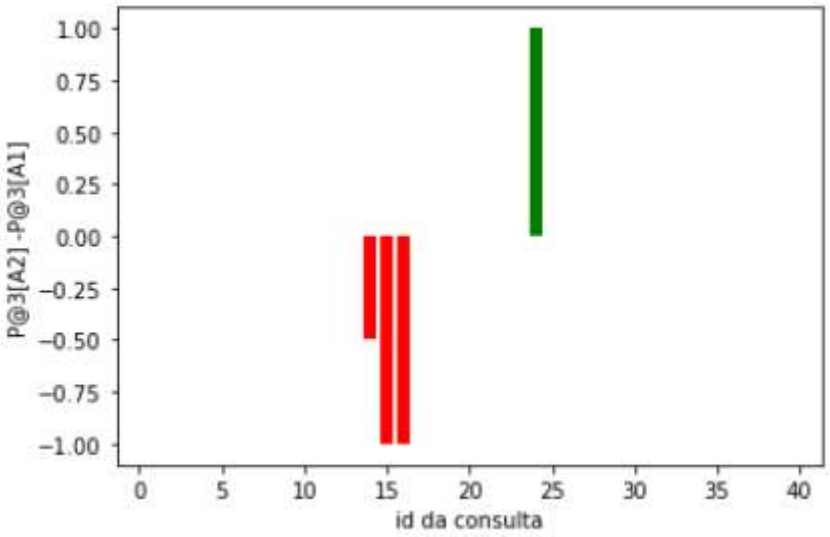
A consulta por regimento interno do tcu é um desafio para se obter uma excelente precisão na recuperação, pois o regimento interno é citado com frequência em outras normas. É preciso pensar em outras *features* que possam ser usadas para maior precisão neste caso.

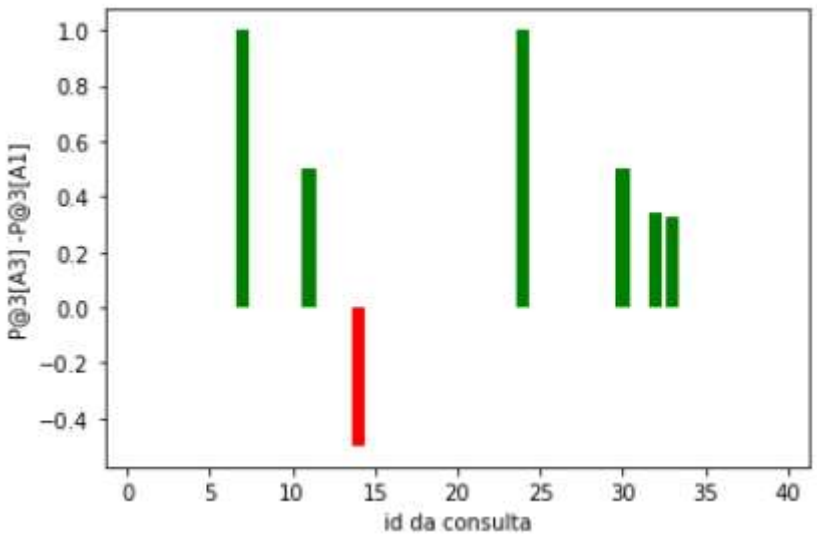
Os histogramas de precisão facilitam a comparação dos modelos por meio da identificação visual (figura 4.3). As barras verdes indicam melhor performance e as vermelhas, performance inferior.

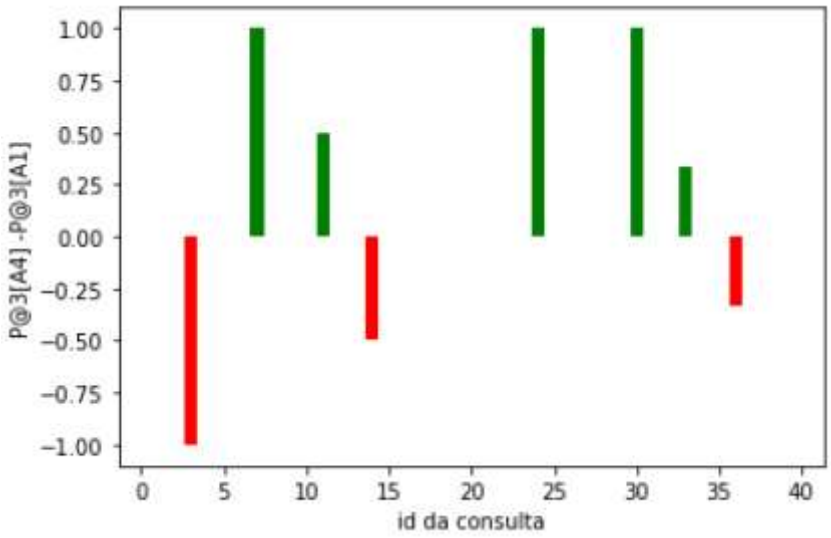
Figura 4.3 – Histogramas de precisão (P@3) calculados como a subtração entre a precisão do modelo comparado e a precisão do modelo referência, para cada consulta avaliada. As barras verdes indicam melhor performance e as vermelhas, performance inferior.











4.2 AVALIAÇÃO P@5

As precisões médias para o nível 5 de revocação podem ser vistas na figura 4.4.

Figura 4.4 – P@5 média. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

	A01 P@5	A02 P@5	A1 P@5	A2 P@5	A3 P@5	A4 P@5
0	0.346667	0.474103	0.758718	0.81	0.832308	0.804103

Da mesma forma que o observado em P@3, a melhoria da recuperação da informação é evidente em relação à produção. Mas, para este nível de revocação, todos os outros três modelos - A2, A3 e A4 - apresentam leve melhora nos níveis de precisão.

Na figura 4.5, é possível ter uma visão detalhada por consulta.

Apesar dos modelos A2 e A4 apresentarem resultados de precisão um pouco maiores que o modelo de referência, eles prejudicaram a precisão da recuperação de 4 consultas. O modelo A2 não conseguiu recuperar os atos de atribuição de cargos do tcu neste nível de revocação, provavelmente por aumentar a revocação com a expansão de consulta com perda de precisão. Já o modelo A4 (*learn to rank)* prejudicou a precisão da recuperação das consultas "trabalho remoto" e "relatório de gestão", da mesma forma que ocorreu em P@3.

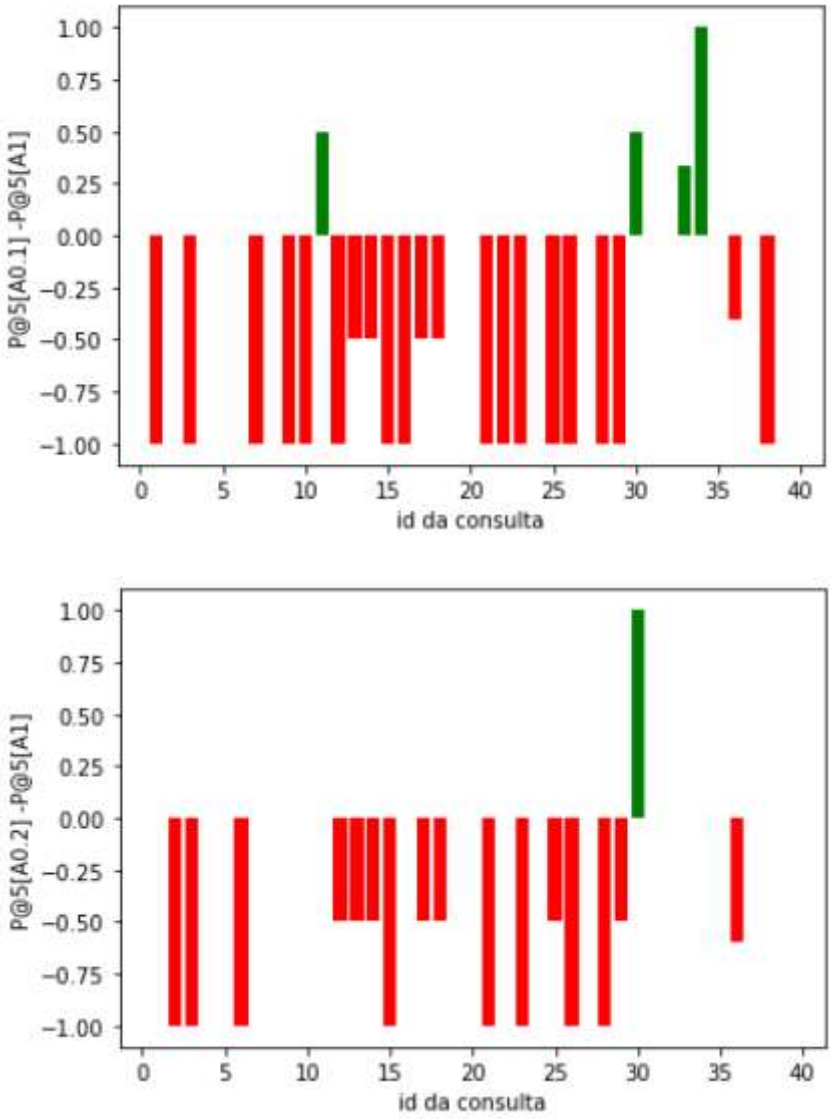
Figura 4.5 – Valores para P@5 para cada consulta. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

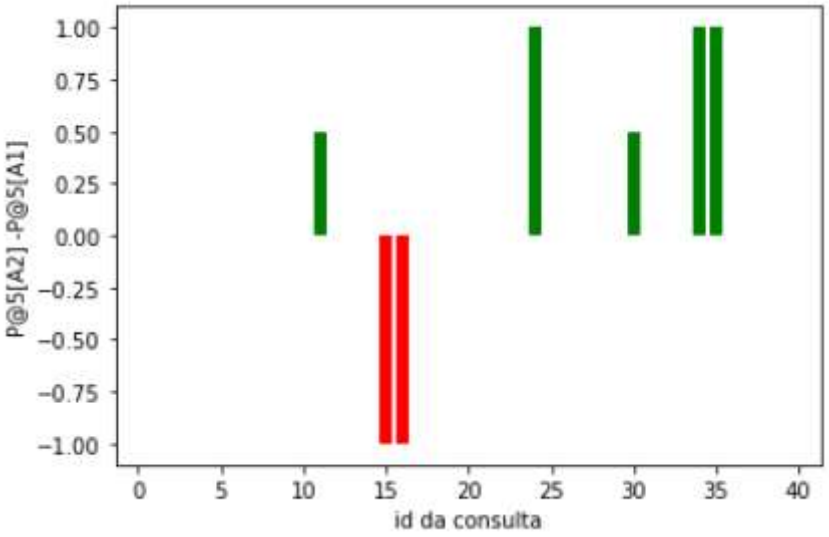
	A01 P@5	A02 P@5	A1 P@5	A2 P@5	A3 P@5	A4 P@5
teletrabalho	0	1	1	1	1	1
trabalho fora das dependências	1	0	1	1	1	1
trabalho remoto	0	0	1	1	1	0
feriados 2019	1	1	1	1	1	1
feriados 2017	1	1	1	1	1	1
feriados	1	0	1	1	1	1
regimento interno	0	1	1	1	1	1
regimento interno do tcu	0	0	0	0	0	0
jornada de trabalho	0	1	1	1	1	1
organização de processos de controle externo	0	1	1	1	1	1
parecer de controle interno	0.5	0	0	0.5	0.5	0.5
suprimento de fundos	0	0.5	1	1	1	1
adiantamento de numerário	0	0	0.5	0.5	0.5	0.5
compra de pequeno valor	0.5	0.5	1	1	1	1
atribuição de cargos do tcu	0	0	1	0	1	1
atribuições de cargos do tcu	0	1	1	0	1	1
estrutura do tcu	0	0	0.5	0.5	0.5	0.5
competências das unidades	0	0	0.5	0.5	0.5	0.5
licença capacitação	0.5	0.5	0.5	0.5	0.5	0.5
recesso	1	1	1	1	1	1
recesso de fim de ano	0	0	1	1	1	1
monitoramento de deliberações	0	1	1	1	1	1
monitoramento de decisões	0	0	1	1	1	1
acompanhamento decisões	0	0	0	1	1	1
tomada de contas especial	0	0.5	1	1	1	1
avaliação de desempenho	0	0	1	1	1	1
diárias e passagens	1	1	1	1	1	1
remocao	0	0	1	1	1	1
distribuição de processos	0	0.5	1	1	1	1
assistência pré-escolar	0.5	1	0	0.5	0.5	1
contratação de serviços	1	1	1	1	1	1
horário de funcionamento do tcu	0.66	0.66	0.66	0.66	1	1
horário de funcionamento do tribunal de contas da união	0.66	0.33	0.33	0.33	0.66	0.66
prestação contas dirigente máximo	1	0	0	1	0	0
prestar contas dirigente máximo	0	0	0	1	0	0
relatório de gestão	0.2	0	0.6	0.6	0.8	0.2
140/2014	1	1	1	1	1	1
63/2010	0	1	1	1	1	1
155/2002	1	1	1	1	1	1

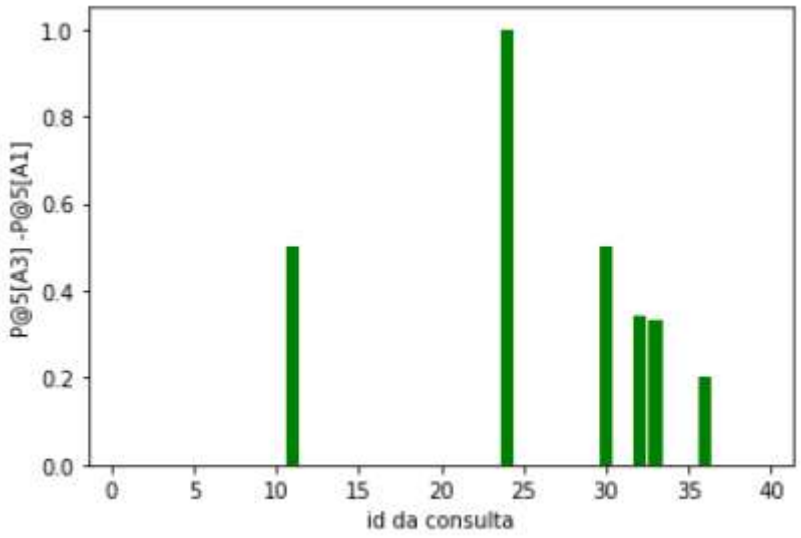
O modelo A3 ofereceu bons resultados no nível 5 de revocação pois não prejudicou a precisão de nenhuma consulta, somente melhorando os níveis de precisão para o conjunto de consultas avaliadas.

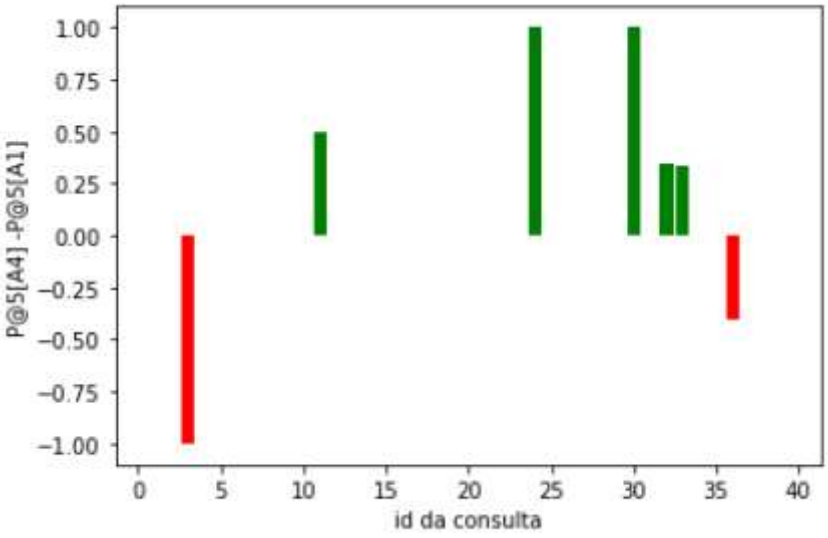
Os histogramas de precisão podem ser visualizados na figura 4.6 para identificação visual rápida da comparação entre os modelos.

Figura 4.6 – Histogramas de precisão (P@5) calculados como a subtração entre a precisão do modelo comparado e a precisão do modelo referência, para cada consulta avaliada. As barras verdes indicam melhor performance e as vermelhas, performance inferior.









4.3AVALIAÇÃO P@10

As precisões médias para o nível 10 de revocação podem ser vistas na figura 4.7.

Figura 4.7 - P@10 média. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

	A01 P@10	A02 P@10	A1 P@10	A2 P@10	A3 P@10	A4 P@10
0	0.449487	0.589487	0.792051	0.907436	0.928974	0.933846

O resultado médio de P@10, comparado aos valores médios de P@3 e P@5, confirmam a intuição de que a expansão de consultas aumenta a revocação. Todos os modelos com expansão (A2, A3, A4) possuem uma precisão maior que o modelo referência para este nível de revocação. O fato do modelo A4 apresentar um valor levemente superior que o A3 pode ser justificado pelo *re-ranking* ocorrer para os 12 primeiros documentos, o que pode priorizar documentos que estavam na segunda página de resultado.

Ter uma precisão média próxima de 90% na primeira página de resultados da pesquisa é um avanço muito interessante para o usuário da pesquisa. Além disso, a expansão de consulta e a extração de conceitos conseguiu melhorar em quase 15% a precisão média em relação ao modelo de referência.

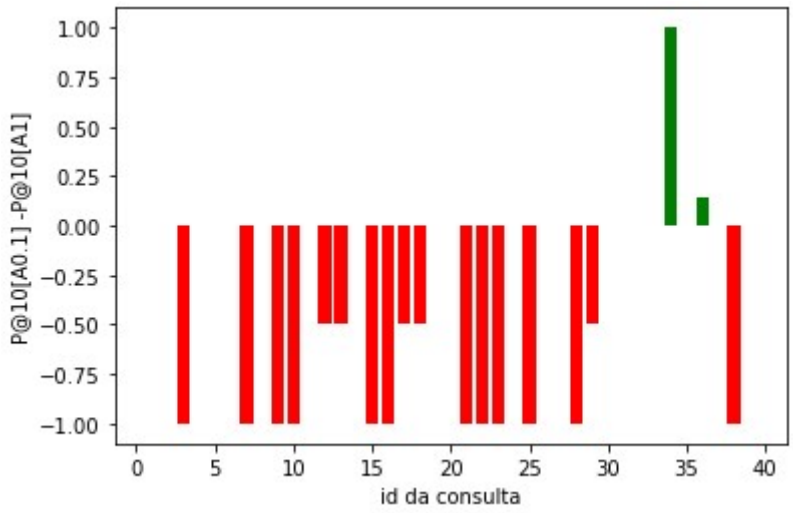
Um olhar mais detalhado nos resultados das consultas específicas evidencia que somente a pesquisa por **relatório de gestão** piorou com o modelo *learn to rank* (A4). Isto já estava claro em P@5 e continuou em P@10. As demais consultas tiveram seus resultados resgatados no nível 10 de revocação.

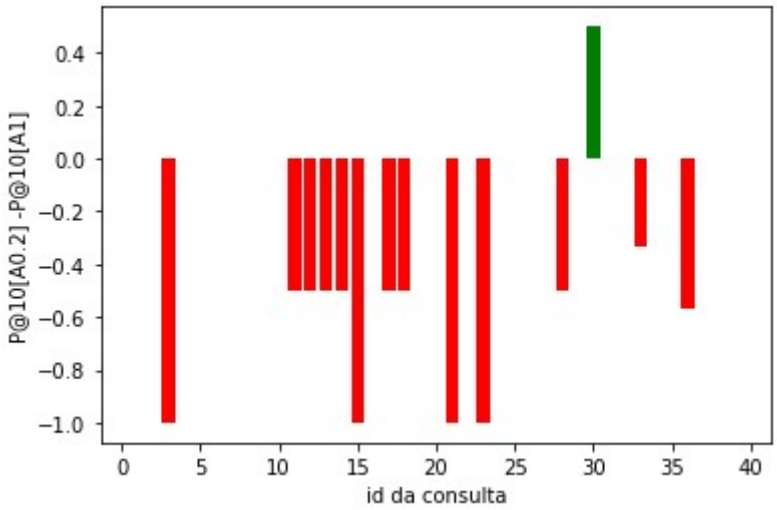
Uma visão detalhada dos resultados para cada consulta pode ser visualizada na figura 4.8 e os histogramas de precisão estão na figura 4.9.

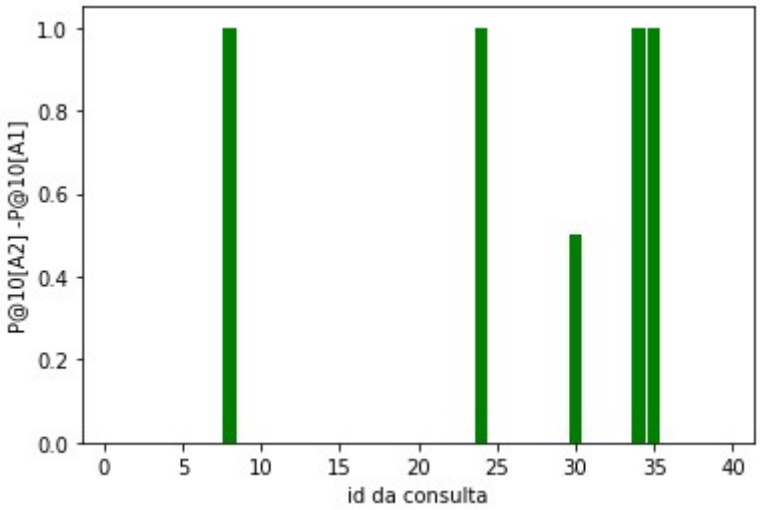
Figura 4.8 – Valores para P@10 para cada consulta. A célula em laranja é o modelo de referência (A1). Células em vermelho indicam resultado inferior ao modelo de referência e as em verde resultado superior.

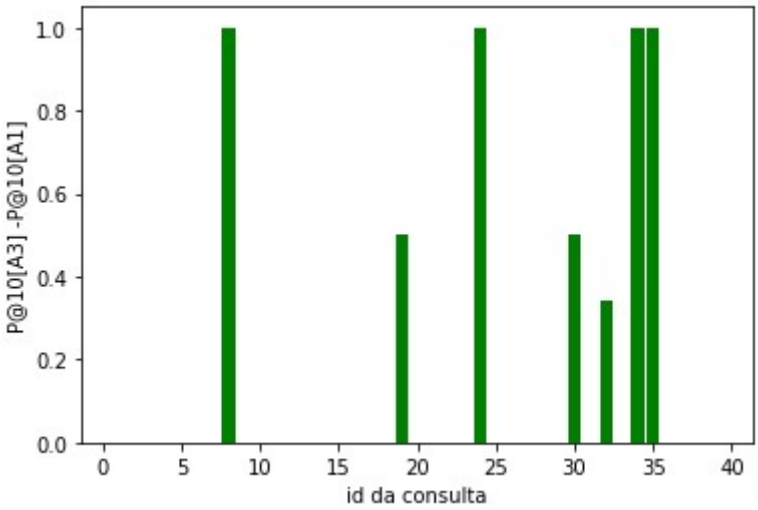
	A01 P@10	A02 P@10	A1 P@10	A2 P@10	A3 P@10	A4 P@10
teletrabalho	1	1	1	1	1	1
trabalho fora das dependências	1	1	1	1	1	1
trabalho remoto	0	0	1	1	1	1
feriados 2019	1	1	1	1	1	1
feriados 2017	1	1	1	1	1	1
feriados	1	1	1	1	1	1
regimento interno	0	1	1	1	1	1
regimento interno do tcu	0	0	0	1	1	1
jornada de trabalho	0	1	1	1	1	1
organização de processos de controle externo	0	1	1	1	1	1
parecer de controle interno	0.5	0	0.5	0.5	0.5	0.5
suprimento de fundos	0.5	0.5	1	1	1	1
adiantamento de numerário	0	0	0.5	0.5	0.5	0.5
compra de pequeno valor	1	0.5	1	1	1	1
atribuição de cargos do tcu	0	0	1	1	1	1
atribuições de cargos do tcu	0	1	1	1	1	1
estrutura do tcu	0	0	0.5	0.5	0.5	0.5
competências das unidades	0	0	0.5	0.5	0.5	0.5
licença capacitação	0.5	0.5	0.5	0.5	1	1
recesso	1	1	1	1	1	1
recesso de fim de ano	0	0	1	1	1	1
monitoramento de deliberações	0	1	1	1	1	1
monitoramento de decisões	0	0	1	1	1	1
acompanhamento decisões	0	0	0	1	1	1
tomada de contas especial	0	1	1	1	1	1
avaliação de desempenho	1	1	1	1	1	1
diárias e passagens	1	1	1	1	1	1
remocao	0	0.5	1	1	1	1
distribuição de processos	0.5	1	1	1	1	1
assistência pré-escolar	0.5	1	0.5	1	1	1
contratação de serviços	1	1	1	1	1	1
horário de funcionamento do tcu	0.66	0.66	0.66	0.66	1	1
horário de funcionamento do tribunal de contas da união	0.66	0.33	0.66	0.66	0.66	1
prestação contas dirigente máximo	1	0	0	1	1	1
prestar contas dirigente máximo	0	0	0	1	1	1
relatório de gestão	0.71	0	0.57	0.57	0.57	0.42
140/2014	1	1	1	1	1	1
63/2010	0	1	1	1	1	1
155/2002	1	1	1	1	1	1

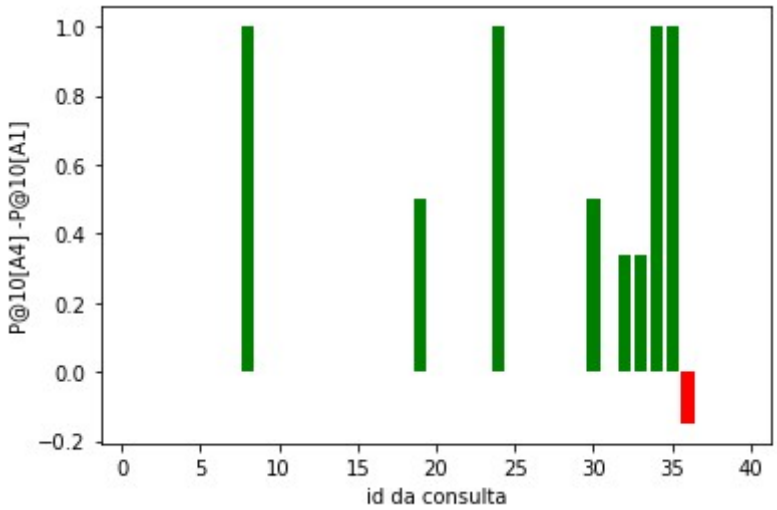
Figura 4.9 – Histogramas de precisão (P@10) calculados como a subtração entre a precisão do modelo comparado e a precisão do modelo referência, para cada consulta avaliada. As barras verdes indicam melhor performance e as vermelhas, performance inferior.











5 CONCLUSÃO

5.1 CONCLUSÕES

Os objetivos propostos para o trabalho foram alcançados.

A fundação para o alcance dos objetivos foi o estudo em relação ao tema de recuperação de informação, relevância em pesquisa e algumas técnicas para identificação de representações textuais, além do entendimento do contexto de negócio relacionado à pesquisa de atos normativos.

Esta base teórica possibilitou a criação incremental de modelos de recuperação de informação ranqueados com o objetivo de aumentar a precisão dos resultados da pesquisa. De uma maneira iterativa, novas capacidades foram agregadas para solucionar problemas específicos de complexidade crescente, com foco no objetivo principal de melhoria da precisão.

Assim, primeiramente foi desenvolvido um modelo ranqueado simples que contemplou a remodelagem da base textual e a manipulação da função de ranqueamento para contemplar o entendimento de relevância do ponto de vista do usuário do sistema.

Este modelo foi incrementado com a expansão semântica da consulta por meio de geração automática de relações entre termos. Ao observar que a expansão semântica de consultas levou a um nível maior de revocação mas com uma piora no nível de precisão, optou-se por implementar um segundo nível de ranqueamento (*learn to rank)* com o foco no objetivo principal do trabalho de elevar os níveis de precisão.

Demonstrou-se, na avaliação dos resultados, que os modelos propostos atingiram níveis crescentes de precisão no nível de revocação 10, ou seja, estes modelos são capazes de recuperar a informação na primeira página de resultado em 90% das pesquisas testadas, contra 45% de precisão mensurados dos resultados atuais de produção. Por outro lado, em níveis de revocação baixos (3), a precisão dos modelos com expansão de consultas piorou em relação ao modelo de referência e, em alguns poucos casos, em relação à produção.

Ressalta-se, no entanto, que os modelos ranqueados propostos sugerem boa performance para o conjunto de validação definido, o que não significa que estes modelos extrapolem bem para casos ainda não vistos.

Por fim, entende-se que os resultados obtidos podem ajudar a implementação futura de modelos ranqueados na Plataforma de Pesquisa Textual, servindo como um ponto de partida já avançado para o time de pesquisa textual da TI do TCU.

5.2 TRABALHOS FUTUROS

A análise dos resultados evidencia o que não está funcionando bem e, a partir disso, pode-se sugerir caminhos de ação para melhoria dos resultados do trabalho:

- Aprimoramento da análise dos termos

  O processo de análise é a fundação de um sistema de recuperação da informação. A análise dos resultados do trabalho apresenta indícios de problemas relacionados ao processamento dos termos, principalmente em relação às palavras com delimitadores, aos números e às formas flexionadas das palavras.

  Pretende-se fazer uma pesquisa sobre a contribuição da lematização para performance de sistemas de recuperação em português.

- Classificação dos atos normativos

  A identificação dos atos normativos que trazem novos assuntos pode ser aprimorada por meio da construção de um classificador que utilize uma maior quantidade de *features*.

- Utilização de mais dados para treinamento do modelo de *word embeddings*

  Pretende-se utilizar o tesauro do TCU (VCE) para enriquecer os dados de treinamento do modelo Word2Vec com o objetivo de capturar mais relações semânticas entre termos. O tesauro do TCU é uma fonte rica e atualizada de relações entre termos, especialmente relacionados ao negócio do TCU.

  Além disso, pretende-se utilizar também outros dados que seguem o mesmo tipo de linguagem dos atos normativos para treinar o modelo, como, por exemplo, utilizar o Vade Mecum do Direito.

- Evolução do modelo utilizado pelo *Learn to Rank*

Pretende-se utilizar aprendizagem de máquina para aprimorar o processo de *re-ranking*, com o objetivo de gerar um modelo de ranqueamento que consiga generalizar bem a performance para todo tipo de consulta. *Features* derivadas dos dados de interação com o sistema de pesquisa (cliques) podem ser utilizadas.

- Classificação dos atos normativos por tema

  50% dos usuários entrevistados sugeriram a classificação do ato normativo por tema. Existe uma espécie de classificação pela separação dos atos normativos do TCU por tema no Portal. Internalizar os temas dos atos normativos na pesquisa pode apoiar bastante o desenvolvimento de certos trabalhos no TCU, como contratação ou gestão de pessoas. Assim, pretende-se estudar este assunto com o objetivo de futura implantação de classificação de atos por tema.

REFERÊNCIAS

BAEZA-YATES, R.; RIBEIRO BERTHIER DE ARAÚJO NETO. **Modern information retrieval: the concepts and technology behind search**. Traducao . [s.l.] Addison Wesley, 2011.

Corretor Gramatical acoplável ao LibreOffice. Disponível em: <http://cogroo.sourceforge.net/>

DIAZ, F.; MITRA, B.; CRASWELL, N. Query Expansion with Locally-Trained Word Embeddings. **Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, 2016.

KUMAR, A.; PAUL, A. **Mastering text mining with R: master text-taming techniques and build effective text-processing applications with R**. Traducao . [s.l.] Packt Publishing Limited, 2016.

LIU, T.-Y. Learning to Rank for Information Retrieval. 2011.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **Introduction to information retrieval**. Traducao . [s.l.] Cambridge University Press, 2018.

REQUENA, Alessandra. **Código Fonte**. Disponível em: https://github.com/alessandrarequena/relevancia-pesquisa. Acesso em: 23 mar. 2020

ROBERTSON, S.; ZARAGOZA, H. **The Probabilistic Relevance Framework: BM25 and Beyond. Vol. 4**. Traducao . [s.l.] Now Publishers, 2009.

TEOFILI, T.; MATTMANN, C. **Deep learning for search**. Traducao . [s.l.] Manning Publications Co., 2019.

TOUPIN, E. et al. **Drupal 8 & Apache Solr: Boost Search Term Relevance by Publish**

Date. Disponível em: <https://atendesigngroup.com/blog/drupal-8-apache-solr-boost-search-term-relevance-publish-date>

TRIBUNAL DE CONTAS DA UNIÃO. **Regimentos internos: Portal TCU**. Disponível em: <https://portal.tcu.gov.br/normativos/regimentos-internos/>

TRIBUNAL DE CONTAS DA UNIÃO. **Revista Reconhe-ser: Portal TCU**. Disponível em: <https://portal.tcu.gov.br/biblioteca-digital/revista-reconhe-ser-2017.htm>

TURNBULL, D.; BERRYMAN, J. **Relevant search with applications for Solr and Elasticsearch**. Traducao . [s.l.] Manning, 2016.

ZUCCON, G. et al. Integrating and Evaluating Neural Word Embeddings in Information Retrieval. **Proceedings of the 20th Australasian Document Computing Symposium on ZZZ - ADCS '15**, 2015.

ANEXO I

Este anexo apresenta as informações obtidas nas entrevistas com servidores do TCU sobre o uso da pesquisa de atos normativos.

Wemerson Soares de Araújo e Renata Nunes de Almeida

Designer de Experiência do Usuário (UX) e Desenvolvedora de Interface Gráfica (UI) na Secretaria de Soluções de TI - STI

Os dois profissionais trabalham no projeto da plataforma de pesquisa textual integrada há dois anos e foram entrevistados por causa de sua grande experiência interagindo com os usuários da pesquisa textual. As seguintes observações foram coletadas:

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	■ **trabalho fora das dependências** O termo que está na base de pesquisa é "trabalho fora de suas dependências", o que dificulta a localização do documento. ■ **teletrabalho** ■ **feriados** ■ **regimento interno** Alguns usuários buscam pela expressão "aprova regimento interno" por causa da baixa precisão dos resultados quando a pesquisa é somente por regimento interno. ■ **reconhe-ser** ■ **jornada de trabalho**

Quais as principais dificuldades em relação à pesquisa de atos normativos?	O resultado das consultas frasais (entre aspas) é diferente da pesquisa livre, o que confunde o usuário. O refinamento da consulta às vezes atrapalha, acontece que quanto mais ele refina tentando achar o que quer, pior fica o resultado da pesquisa.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Não. Os usuários em geral percebem o ato com um todo, não diferenciam campos. Eles buscam um tema ou termos. Há uma intuição que se os termos estiverem mais próximos, maior a relevância.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não. O usuário não distingue o tipo de documento, seja em relação a todas as bases, seja em relação aos tipos do ato normativo.
Como você percebe a influência da data de expedição no resultado da pesquisa?	A ordenação por data decrescente não reflete a intuição do usuário. Como uma alternativa à baixa precisão dos resultados, o usuário utiliza bastante o filtro de data pois às vezes sabe o período em que foi publicado o ato normativo.
Os atos revogados são menos importantes que os atos vigentes?	Sim. Resta uma dúvida se o usuário entra no termo de pesquisa a palavra “revogado”

Vilmar Agapito Teixeira

Assessor do Ministro-Substituto André Luís de Carvalho

Vilmar foi identificado pelos logs da consulta como um usuário frequente e, por trabalhar como assessor de ministro-substituto, pode fornecer uma visão do trabalho em gabinete. As seguintes observações foram coletadas:

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	■ **número, ano ou data** Normalmente, busca por esses metadados. Usa pouco termos na base de atos normativos porque dá trabalho para localizar nos resultados da busca. O problema é que não se sabe exatamente qual o termo que está no documento. Há muita tentativa e erro e demora-se para acertar. Vem muita coisa que não interessa. O gabinete normalmente usa atos normativos que têm normatização para fora do TCU - IN e DN. Exemplo: contas, contas especiais, resoluções Na área de pessoal, é mais comum utilizar atos normativo. Facilitaria se tivesse áreas temáticas para os atos, assim como acontece na jurisprudência selecionada. Ele hoje busca os atos normativos nas páginas temáticas do portal quando precisa, porque lá já está listado. Exemplo: Pessoal, Prestação de Contas. A classificação por temas facilitaria para para público interno e externo.

Quais as principais dificuldades em relação à pesquisa de atos normativos?	A segregação dos atos normativos (os que são da Presidência e os de expedição das unidades) é muito ruim, porque o usuário não enxerga essa diferença. Para o usuário, tudo é ato normativo. O TCU é modelo/espelho para outros órgãos. Então, servidores da Administração Pública vem muito ao portal para saber como o TCU disciplina certos assuntos. Muitas vezes, os órgãos não acham a informação e ligam para a Segedam para pedir a informação.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. Há uma prioridade em relação às partes do ato normativo. Se estiver na ementa, é mais importante. Se estiver no corpo, segundo lugar. E anexo, terceiro lugar. Ex: 64/2003 É interessante poder pesquisar somente por parte do ato normativo, ou seja, só ementa, corpo ou anexos. A ementa é o assunto, o corpo é a regulamentação do assunto e os anexos é o resto.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não. O tipo de ato normativo não tem hierarquia em relação à importância. O tipo tem a ver com a natureza do ato. IN e DN são para o público externo, os demais para interno.
Como você percebe a influência da data de expedição no resultado da pesquisa?	Nem sempre a data é o critério mais importante para ordenação.

Os atos revogados são menos importantes que os atos vigentes?	Sim. Os atos revogados não têm tanta importância pois eles já são referenciados no novo ato e o ato alterado tem o texto novo e o texto anterior. Seria interessante ter hyperlink para os atos que alteram ou revogam

Angerico Alves Barroso Filhos

Secretaria de Controle Externo da Administração - Secex Administração

Angerico foi indicado pelo time de UX como um usuário que tem muito colaborado com o desenvolvimento da plataforma de pesquisa textual. Por pertencer à Secex Administração, ele pode fornecer uma visão de servidor que trabalha diretamente com instrução de processos de controle externo que utiliza a pesquisa de atos normativos para dar apoio a este trabalho.

Pergunta	Resposta
Que tipo de informação você mais busca na base de atos normativos? Descreva suas principais necessidades.	Normalmente busca normas de processos de controle (resoluções): portarias da Segecex e DN de Contas. Há dificuldade para achar o ato normativo por causa do termo específico. Eventualmente, vai no Google para encontrar. ■ **relatório de gestão** A expectativa com essa busca é localizar a decisão normativa sobre como as contas devem ser apresentadas. Deve retornar 178/2019 e o que se espera já vem na primeira posição na pesquisa atual.

	■ **organização de processos de controle externo** O resultado da pesquisa atual não atende. A resposta é a Resolução 259/2014. ■ **relatório de auditoria ou parecer de controle interno** Busca-se a DN 172/2018 . Vai sair uma DN nova agora em dezembro sobre este assunto.
Quais as principais dificuldades em relação à pesquisa de atos normativos?	Os atos normativos do TCU estão espalhados, não há uma base centralizada de todos eles. Por exemplo, a portaria da Segecex que fala de distribuição de competências não é um ato da Presidência, por isso não está na base de atos normativos. Essa portaria está em servidor de arquivos e não é pesquisável: _sarq_prod\Segecex\Publico\Portarias\Portarias2019
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. A questão central está na ementa, por isso é mais importante. Não vê diferença entre corpo e anexos.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.

Como você percebe a influência da data de expedição no resultado da pesquisa?	Nem sempre a data é a melhor para ordenação dos resultados da pesquisa.
Os atos revogados são menos importantes que os atos vigentes?	Sim. Eventualmente, é necessário conhecer como era antes, qual foi o caminho/histórico até chegar na situação atual. Mas é exceção.

George Atsushi Murakami

Assessor da Secretaria de Infraestrutura de TI – Setic

George também foi indicado pelo time de UX como um usuário que colabora com o desenvolvimento da plataforma de pesquisa textual e tem bastante experiência no uso da pesquisa de atos normativos do TCU para desempenhar seu trabalho na assessoria da Setic.

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	Busca muito por contratação em geral (planejamento e execução de contratos). Mas há dificuldades. Por exemplo, a busca pelo termo terceirização não acha nada, porque este não é o termo usado na norma. ■ **licença capacitação** ■ **projeto de especialista** ■ **planejamento do Tribunal** O problema da variação dos termos é importante. Por exemplo, buscar tecnologia da informação e TI.

Quais as principais dificuldades em relação à pesquisa de atos normativos?	A dificuldade em relação a atos normativos é achar/localizar todos os atos de um tema em questão. Esse é o único normativo ou tem outros relacionados? Este é um problema genérico de legislações em geral. Outro ponto é o relacionamento dos atos que não é claro. Há dificuldade em saber tudo sobre determinado tema. A resposta é exaustiva? Um grande desafio seria a organização por temas. Algumas pessoas do TCU fazem um trabalho de curadoria de atos normativos em relação a grandes temas. Por exemplo, a Francis da Segep é uma espécie de curadora de atos normativos do tema de contratação. Ele entende que se deve avaliar todas as normas para saber sobre o assunto.Há temas amplos, que são mais complicados. Estes precisam de um estudo mais aprofundado da base e é necessário avaliar tudo que tem sobre o tema. Há temas mais específicos, que são resolvidos com o acesso a um ato e pronto. Assim, temos os dois tipos de necessidades de informação nesta base, uma que é um estudo aprofundado de um tema e outro que é localizar uma informação e pronto.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Não. A ementa não importa. O importante é o tema sendo tratado no corpo do ato normativo.

Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.
Como você percebe a influência da data de expedição no resultado da pesquisa?	Nem sempre a data é a melhor para ordenação dos resultados.
Os atos revogados são menos importantes que os atos vigentes?	Sim. As revogadas não deveriam aparecer na pesquisa explicitamente.

Francismary Souza Pimenta Maciel

Assessora de Secretário-Geral da Secretaria Geral de Administração – Segedam

A Francis foi indicada pela Secretaria das Sessões como uma importante usuária da base de atos normativos do TCU. Ela tem longa experiência de trabalho na assessoria da Secretaria de Administração do TCU.

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	Há grande dificuldade em se pesquisar por assunto. Não se sabe ao certo os termos que foram usados nos atos e não acha a norma. Então, o que se faz é ligar para alguém para descobrir o número e consultar por número. Normalmente, se busca pelo número e tem um registro paralelo dos números de atos que são muito usados. ■ **suprimento de fundos**

	A pesquisa encontra o ato normativo por esse termo, mas não acha por **adiantamento de numerário** ou **compra de pequeno valor**. Poderia haver um jeito de pesquisar por assunto de forma mais eficiente. Há uma crença entre alguns usuários que consomem atos normativos de que "aqui no TCU você tem que saber o número, senão já era". ■ **atribuição dos cargos do TCU** Esta norma ela não conseguiu localizar pelo assunto. É uma resolução, ligou para a Segep e perguntou pelo número.
Quais as principais dificuldades em relação à pesquisa de atos normativos?	Primeiro grande problema: os atos normativos estão dispersos. Não há unificação, é muito difícil encontrar os atos das unidades. Um outro problema importante é em relação às portarias de delegação de competência (portaria conjunta segedam-segecex). Muito difícil achar as subdelegações. Poderia haver um link entre as portarias, a delegação e as subdelegações. Enfim, a parte de normativos de competências é um caos. Poderia haver um repositório específico de portarias de competências.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. Quando o assunto aparece na ementa, é porque é o tema da norma. Entende que este ato é específico daquele tema e por isso ela dá prioridade.

Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.
Como você percebe a influência da data de expedição no resultado da pesquisa?	Nem sempre é a melhor ordenação.
Os atos revogados são menos importantes que os atos vigentes?	Não. As revogadas são importantes pois, a depender do trabalho, o histórico pode ser relevante. Além disso, às vezes instrui um processo e precisa da norma da época. Por isso, não vê problema nas revogadas aparecerem no resultado, desde que esteja claro que é uma norma revogada e qual revogou.

Maria Vanda Lima Pinto

Assessoria Secretaria de Gestão de Pessoas - Segep

A Vanda foi indicada pelo time de UX como uma servidora que colabora com o desenvolvimento da plataforma de pesquisa textual e tem experiência no uso de atos normativos do TCU em seu trabalho na assessoria da Secretaria de Gestão de Pessoas.

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	Temas que normalmente busca são os temas da área de pessoal. ■ **assistência médica** ■ **criação do Pro-TCU**

	Tem várias portarias relacionadas a essas, nem sempre é fácil alcançar todas elas. ■ **licença capacitação** ■ **regime especial** ■ **jornada de trabalho** ■ **recesso** Tem o recesso previsto no RI. Tem o recesso do servidor. Desde quando o recesso existe? É uma pergunta que foi respondida por busca na base de atos normativos Como o TCU funcionava em determinada época? Para responder esta questão, ela busca a que está em vigor e vai traçando o histórico para chegar na situação do passado. A Lei Orgânica deveria estar lá na base de atos normativos. Tem uma página no portal que lista os normativos importantes que regem o TCU, nem todos estão na pesquisa. Ver a página do Portal que tem as normas. Tem dificuldade de buscar por tema. Apesar de que a busca está bem melhor que antes. Mas ainda há dificuldade. Exemplo: coloca uma palavra e vem um monte de coisa que não tem nada a ver com o tema. Faz muita busca por ano e número. Mas também usa a busca por assunto, pois na sua percepção melhorou. A busca por palavras: saúde, médica, jornada retorna coisas que não tem nada a ver.

Quais as principais dificuldades em relação à pesquisa de atos normativos?	O principal problema é que as normas não estão unificadas. A pesquisa das normas TCU é boa, mas a pesquisa do Sisnormas é péssima. O Sisnormas é o sistema que contém as normas administrativas, da segedam e segepres. As normas da Segecex não estão em lugar nenhum. A base das normas do TCU está melhor que a anterior.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. Se a informação está na ementa, aumenta a relevância do ato.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.
Como você percebe a influência da data de expedição no resultado da pesquisa?	Nem sempre é a melhor ordenação.
Os atos revogados são menos importantes que os atos vigentes?	Sim. As normas revogadas deveriam ter menos importância. Mas elas também são importantes. Às vezes não se sabe se o ato foi revogado, descobre-se pesquisando. Às vezes vem mais revogados do que o que interessa. A impressão que dá é que o revogado é mais importante que o em vigor.

Luciana de Freitas Mourão

Chefe do Serviço de Administração do Gabinete da Presidência

Todos os atos normativos da Presidência passam pela Luciana e por isso ela foi indicada pela Secretaria das Sessões para apoiar este trabalho.

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	Busca pelas normas de rotinas de expedição e encaminhamento de assuntos da presidência. Faz pesquisas para a alteração de portarias, busca as portarias antigas. Busca por normas que disciplinam a área de controle externo. Às vezes, pesquisa por número e às vezes por assunto. Muitas vezes já sabe o número do ato. A busca às vezes traz muitos documentos e há uma dificuldade de triagem.
Quais as principais dificuldades em relação à pesquisa de atos normativos?	Não tem dificuldades com a pesquisa.
Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. A ementa traz o assunto do ato e com certeza faz o ato ser mais relevante.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.

Como você percebe a influência da data de expedição no resultado da pesquisa?	A ordenação pela cronologia é mais adequada.
Os atos revogados são menos importantes que os atos vigentes?	Sim. Seria interessante reduzir a importância do ato revogado. Mas ela acha que o retorno da pesquisa é satisfatório.

Barnabé Tomas Pereira

Assessor da Secretaria de Fiscalização de Pessoal – Sefip

O Barnabé tem longa experiência na Assessoria da Sefip e usa a pesquisa textual para a execução do seu trabalho diário. Colabora com o desenvolvimento da plataforma de pesquisa textual e foi indicado pelo time de UX para apoiar este trabalho.

Pergunta	Resposta
Que tipo de informação você mais busca na base de datos normativos? Descreva suas principais necessidades.	Busca por assunto sem problemas Normalmente busca por nome/assunto, não sabe o número. Temas relacionados à área de pessoal: pensão, aposentadoria, quintos, VPNI.
Quais as principais dificuldades em relação à pesquisa de atos normativos?	Não tem dificuldade em localizar acórdãos. A jurisprudência antiga é um problema para recuperar, pois os processos foram digitalizados e não tem qualidade. Sobre os atos normativos, não tem um local para acesso direto ao ato. Vê isso como uma desvantagem. Seria interessante ter um local com link para os atos.

Se a informação for localizada na ementa do ato, isso é um indicativo de que aquele ato é mais importante para a dada consulta?	Sim. A ementa é sim superior em relação aos outros campos, ele olha a ementa e se a informação estiver presente na ementa, usa aquele ato normativo.
Em relação aos tipos de ato normativos, você sabe distinguir os tipos ou percebe algum tipo de hierarquia entre os tipos de atos normativos?	Não.
Como você percebe a influência da data de expedição no resultado da pesquisa?	Sem opinião.
Os atos revogados são menos importantes que os atos vigentes?	Não. Os atos revogados são importantes para o trabalho da Sefip pois eles têm muitos processos em que se deve aplicar legislação da época. Assim, é importante para a natureza do trabalho da Sefip.