UM SISTEMA DE BUSCA PARA FÓRUNS DE DISCUSSÃO NA WEB

Por

DANIEL SANTOS PEIXOTO

Trabalho de Graduação

Universidade Federal da Bahia
dcc@ufba.br
wiki.dcc.ufba.br/DCC/

SALVADOR/2019

Monografia apresentada por **Daniel Santos Peixoto** ao programa de Bacharelado em Ciência da Computação do Departamento de Ciência da Computação da Universidade Federal da Bahia, sob o título **Um Sistema de Busca para Fóruns de Discussão na Web**, orientada pelo **Prof. Frederico Araújo Durão** e aprovada pela banca examinadora formada pelos professores:

Prof. Frederico Araújo Durão
Departamento de Ciência da Computação/UFBA

Prof. Danilo Barbosa Coimbra
Departamento de Ciência da Computação/UFBA

Prof. Roberto Freitas Parente
Departamento de Ciência da Computação/UFBA

SALVADOR
2019

Universidade Federal da Bahia
Departamento de Ciência da Computação

DANIEL SANTOS PEIXOTO

"UM SISTEMA DE BUSCA PARA FÓRUNS DE DISCUSSÃO NA WEB"

Trabalho apresentado ao Programa de do Departamento de Ciência da Computação da Universidade Federal da Bahia como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Orientador: *Frederico Araújo Durão*

SALVADOR, 2019

In God we trust, others bring data.

—W. EDWARDS DEMING

Resumo

Fóruns de discussão são meios populares para se obter informação na Web, permitindo que usuários compartilhem conhecimentos ao criarem tópicos para discutir os mais variados assuntos. Sendo assim, os sistemas de recuperação tentam filtrar os resultados mais relacionados à busca realizada pelo usuário. No entanto, os resultados encontrados nem sempre são relevantes. Por conta disso, os usuários não têm uma boa experiência em acessar o conteúdo existente e em alguns casos nunca acessam. Consequentemente, sem encontrar a informação que desejam, tópicos vão sendo criados com a mesma intenção de tópicos já existentes, o que acaba aumentando o número de tópicos duplicados. Portanto, este trabalho propõe um método que, além dos meios convencionais de busca em texto, utiliza o feedback da comunidade sobre um tópico para avaliar sua relevância visando proporcionar resultados mais precisos diante da busca realizada. Ainda assim, para a avaliação realizada, foi percebido que essa metodologia permite a obtenção de resultados mais eficazes do que os métodos convencionais disponíveis.

Palavras-chave: busca, fóruns, crowdsourcing, perguntas e repostas, pln, recuperação da informação

Abstract

Discussion forums are a popular way of getting information online. It allows users to share knowledge by creating topics to discuss about a wide range of subjects. Because of that, information retrieval systems try to filter results that are related to the user query. Although, the results found are not frequently relevant and users don't have a good experience accessing existing content and, sometimes, never have any access to useful content. Consequently, without finding the needed information, new topics are created with the same intention of existing topics and this increases the amount of duplicates. Because of that, this work propose a method that, more than conventional text retrieval methods, uses the community feedback of a topic to evaluate its relevance. In this way, evaluations prove that this methodology guarantees more precise results are found to a given and an overall increase over 100% in the effectiveness against conventional text retrieval methods.

Keywords: search, forums, crowdsourced knowledge, qa, nlp, information retrieval

Sumário

1 Introdução **19**

1.1 Motivação 20
1.2 Problema 20
1.3 Objetivos da Solução Proposta 21
1.4 Estrutura 21

2 Fóruns de Discussão na Web **23**

2.1 Introdução 23
2.2 Histórico 24
2.3 Fóruns Populares 25
2.4 Usos Mercadológicos 26
2.4.1 Vendas Direcionadas 26
2.4.2 Recrutamento 26
2.4.3 Educação à distância 27
2.5 Sumário 28

3 Recuperação da Informação **31**

3.1 Introdução 31
3.2 Recuperação de Informação versus Recuperação de Dados 31
3.3 Estratégias de Recuperação 33
3.3.1 Modelo Booleano 33
3.3.2 Modelo de Espaço Vetorial 34
3.3.3 Modelos Probabilísticos 35
3.3.3.1 Okapi BM25 35
3.4 Métodos de avaliação 37
3.5 Tecnologias e Frameworks 38
3.6 Sumário 39

4 Um Sistema de Busca para Fóruns de Discussão na Web **41**

4.1 Requisitos 41
4.2 Visão Geral da Solução 42
4.3 Modelagem dos Dados 43
4.4 Pré-processamento 43
4.4.1 Filtros de caracteres 44
4.4.2 Tokenização 44
4.4.3 Filtros de tokens 44

4.4.3.1 Uniformização . . . 44
4.4.3.2 Remoção de Stop Words . . . 45
4.4.3.3 Stemming . . . 45
4.4.4 Itens pré-processados . . . 46
4.5 Indexação e Recuperação . . . 46
4.6 Modelo de Busca . . . 47
4.7 Tecnologias Utilizadas . . . 48
4.8 Sumário . . . 50

5 Avaliação **51**
5.1 Conjunto de Dados . . . 51
5.2 Metodologia . . . 52
5.3 Métricas de Avaliação . . . 53
5.4 Resultados . . . 54
5.4.1 Comparação do uso dos campos . . . 54
5.4.2 Variações de Influência da Similaridade . . . 55
5.4.3 Comparações com outros métodos . . . 56
5.5 Discussão . . . 56
5.6 Sumário . . . 57

6 Conclusão **59**
6.1 Contribuições . . . 59
6.2 Trabalhos Futuros . . . 60
6.3 Sumário . . . 61

Referências **62**

Lista de Figuras

2.1 Fórum do AnswerPoint. 25
2.2 Anúncios nas perguntas feitas no website Quora. 27
2.3 Recrutamento usado como forma de monetização no StackExchange. 28
2.4 Exposição de dúvida em um fórum no Moodle. 29

3.1 Exemplo de consulta no modelo booleano [3]. 33
3.2 Saturação da frequência dos termos no BM25 e sua comparação com o TF/IDF [2]. 37

4.1 Diagrama de dependências da solução proposta. 43
4.2 Regras usadas na primeira fase do Porter Stemmer. 46

5.1 A pergunta foi marcada como duplicata de uma já existente. 52
5.2 Comparação entre variantes da solução que usam ou não certos campos. 54
5.3 Comparação de diferentes usos da variável i da Equação 4.1. 55
5.4 Valores de precisão e cobertura ao se comparar com outros métodos de recuperação. 57
5.5 Valores de F-measure ao se comparar com outros métodos de recuperação. . . . 58

Lista de Tabelas

3.1 Diferença entre sistemas Recuperação de Informação (RI) e Recuperação de Dados (RD)[14]. 32
3.2 Frases de exemplo. 34
3.3 Frequência de termos nos documentos. 34

4.1 Requisitos da solução proposta. 42
4.2 Formato exigido para a modelagem de um tópico. 43
4.3 Frases de exemplo. 44
4.4 Frases de exemplo após os filtros de caracteres. 44
4.5 Frases de exemplo após a tokenização. 45
4.6 Frases de exemplo após as transformações. 45
4.7 Frases de exemplo após a remoção de *stop words*. 45
4.8 Frases de exemplo após o stemming. 46
4.9 Frases de exemplo ao final do pré-processamento. 46
4.10 Tópico após o pré-processamento. 47
4.11 Frases de exemplo no índice invertido. 47
4.12 Cálculo de relevância das frases de exemplo. $i = 4$. 48

Lista de Acrônimos

APM	Application Perfomance Monitoring 49
AVA	Ambientes Virtuais de Aprendizado 27
ELK	Elasticsearch Logstash Kibana
DCG	Discounted Cumulative Gain 53
IDF	Inverse Document Frequency 35
MAP	Mean Average Precision 53
MRR	Mean Recurrent Rank 53
NLP	Natural Language Processing 49
PLN	Processamento de Linguagem Natural 43
RI	Recuperação de Informação 31
RD	Recuperação de Dados 31

Monografia apresentada por **Daniel Santos Peixoto** ao programa de Bacharelado em Ciência da Computação do Departamento de Ciência da Computação da Universidade Federal da Bahia, sob o título **Um Sistema de Busca para Fóruns de Discussão na Web**, orientada pelo **Prof. Frederico Araújo Durão** e aprovada pela banca examinadora formada pelos professores:

Prof. Frederico Araújo Durão
Departamento de Ciência da Computação/UFBA

Prof. Danilo Barbosa Coimbra
Departamento de Ciência da Computação/UFBA

Prof. Roberto Freitas Parente
Departamento de Ciência da Computação/UFBA

SALVADOR
2019

1

Introdução

A popularização da Internet permitiu que a obtenção e geração de conhecimento também se tornasse acessível à maior parte da população. Em números, há mais de 3.8 bilhões de usuários [6], mais de 2.5 petabytes gerados todos os dias [9]. As enciclopédias foram substituídas por wikis, aulas e cursos já estão migrando para serviços de *streaming* e as discussões saíram dos espaços físicos para os fóruns. Além disso, o conhecimento agora é mais democrático, mais pessoas podem produzir conteúdo, são mais de 600 edições de páginas na Wikipédia[1] por minuto [5]. Dessa maneira, atualmente qualquer indivíduo que possua acesso à internet pode consumir todo esse conteúdo, e não mais somente grandes empresas e/ou acadêmicos.

Existem diversos tipos de fóruns na Internet, alguns mais conhecidos como Yahoo Respostas[2], StackOverflow[3] e AskUbuntu[4]. Além destes, há também alguns com domínios específicos para comunidades, que podem falar sobre animes, filmes, música, empreendedorismo, etc. Os grandes fóruns normalmente contam com equipes especializadas para seu funcionamento. Entretanto, os pequenos são normalmente criados em plataformas que também os hospedam, como é o caso da Fandom[5], que permite que usuários hospedem fóruns de entretenimento gratuitamente. Diante disso, aqueles que tem perguntas, possuem meios para interagir com os que detêm conhecimento de um determinado domínio, além de compartilhar essa interação com o resto da comunidade, que está ao redor de todo o globo.

Devido aos mecanismos de interação do usuário com a informação, é possível verificar a relevância de uma dada informação. O StackOverflow[6] usa dessa estratégia para que as respostas possam ser validadas ou não pela comunidade. Por meio de votos positivos e negativos, comentários, avaliação do próprio autor do tópico, a comunidade pode expressar sua opinião sobre as respostas ofertadas.

Com um número tão grande de informação, faz-se necessário que ferramentas automatizadas auxiliem os usuários na busca por informações relevantes aos seus interesses, sendo assim,

[1] https://www.wikipedia.org
[2] https://www.answers.yahoo.com
[3] https://www.stackoverflow.com
[4] https://www.askubuntu.com
[5] https://www.fandom.com
[6] https://www.stackoverflow.com

surgiram os buscadores como o Google[7] e Bing[8]. O próprio Google, realiza mais de 3.5 bilhões de buscas todos os dias [24]. Além destes, cada vez mais surgem buscadores específicos para determinados conteúdos, como é o caso do Google Scholar[9], YouTube[10] e Buscapé[11]. Logo, para ajudar os usuários a lidarem com o grande volume de informações em fóruns, buscadores para fóruns se fazem necessários.

1.1 Motivação

A ação de buscar algo na Web é algo que pode se tornar uma tarefa repetitiva ao logo do tempo. Constantemente indivíduos buscam por informações à respeito de dúvidas relacionadas a algum contexto. Sendo assim, cerca de 80% à 84% das buscas no Google já foram feitas anteriormente [13]. Da mesma forma, tópicos em fóruns tendem a se repetir. É muito importante que um buscador de fóruns traga bons resultados. Se um usuário não possui sua dúvida sanada por meio do histórico já existente, ele deve criar um novo tópico para esse banco de dados, o que implicará, quando essa resposta já existe, em um retrabalho para a discussão nesse tópico e um aumento desnecessário do problema, já que agora temos mais um item no banco de dados. Por conta disso, durante o desenvolvimento desse trabalho foi notado que 10% das perguntas do AskUbuntu são perguntas duplicatas, ou seja, alguma pessoa já realizou essa mesma pergunta anteriormente, esse número poderia ser diminuído com ferramentas de buscas mais adequadas. Dessa forma, não somente existem mais perguntas sendo criadas, mas também a experiência do usuário é depreciada, já que ele não tem um acesso fácil à informação.

1.2 Problema

O problema que esse trabalho investiga é a precisão nos buscadores de fóruns. O seguinte estudo verifica meios para retornar bons resultados ainda que não se haja grandes equipes de desenvolvimento, ou que a solução cause grandes impactos na estrutura já existente, garantindo que a solução seja simples e eficiente.

Com milhões de informações disponíveis, precisamos usar filtros para encontrar a informação desejada. Para isso, precisamos verificar a relevância e a utilidade que esta informação possui, pois nem toda informação recebida é necessariamente verdade ou capaz de solucionar o problema do usuário. Porém, esta mesma tarefa tem que ser executada em tempo hábil para se prover uma resposta satisfatória para o usuário, o que é outro problema já que consideramos um número massivo de dados.

[7] https://www.google.com
[8] https://www.bing.com
[9] https://www.scholar.google.com
[10] https://www.youtube.com
[11] https://www.buscape.com.br

Ainda que tenhamos metadados, é preciso entender o valor que estes trazem para a busca. Informações como datas, feedbacks positivos e interação com conteúdo podem ter influências diferentes dependendo do domínio utilizado, já que alguns conteúdos podem ser desvalorizados com o passar do tempo e outros perdurarem e, assim, inviabilizar o uso de datas como parâmetro de buscas. Da mesma forma, algumas perguntas por serem mais frequentemente encontradas podem ter a tendência de ter mais interações que outras, o que pode prejudicar as perguntas que não ocorrem tão frequentemente e, consequentemente, não recebem pontuações tão altas.

1.3 Objetivos da Solução Proposta

O objetivo desse trabalho é implementar um sistema de busca que garanta acesso aos conteúdos existentes nos fóruns que podem possivelmente resolver o anseio do usuário expresso em uma busca. Tendo como objetivos específicos:

- **Objetivo 1:** Revisão da literatura existente de Recuperação da Informação;
- **Objetivo 2:** Recuperar tópicos já existentes que sejam relevantes à busca do usuário;
- **Objetivo 3:** Propor um método para avaliar a relevância do conteúdo para uma dada busca;
- **Objetivo 4:** Realizar uma avaliação experimental a fim de verificar-se a qualidade das buscas.

1.4 Estrutura

O capítulo atual foi responsável por trazer a problemática do assunto em questão e a sua importância para a sociedade. Os seguintes capítulos irão agregar o conhecimento necessário para o leitor possa analisar a solução proposta no Capítulo 4. No Capítulo 2 será explicada a trajetória dos fóruns de discussão na Web, desde o seu surgimento e popularização até os dias atuais. O Capítulo 3 trará informações essenciais para a compreensão dos desafios para criar soluções em sistemas de recuperação. No Capítulo 4 os requisitos da aplicação sejam definidos e todo o processo de implementação da solução seja facilmente compreendido. O Capítulo 5 apresenta a avaliação realizada com o objetivo de mensurar e apresentar as contribuições deste trabalho. Ao fim, será apresentada, no Capítulo 6, a conclusão deste projeto, a qual irá trazer uma visão geral do que foi feito e o que necessita de melhorias.

2

Fóruns de Discussão na Web

Ao decorrer deste capítulo, serão abordados diferentes aspectos dos fóruns na sociedade. Com o objetivo de informar o leitor sobre a necessidade deste trabalho, serão discutidos os fatores que tornaram alguns fóruns bem sucedidos e outros decadentes. Alguns dos nomes mais importantes em fóruns de discussão serão parte desse capítulo, os mesmos foram usados como referência ao se decidir criar um ambiente adequado para a aplicação deste projeto. Embora cada um deles aborde o conteúdo de forma diferente, todos são frutos de um conhecimento gerado há anos por meio das transformações dos fóruns.

2.1 Introdução

De acordo com o dicionário, fóruns são um lugar ou meio onde ideias e visões sobre um particular assunto podem ser trocadas. Quando trazemos esse conceito para a Web, essa visão vem acompanhada de um conjunto de ferramentas que irá permitir o usuário o uso de recursos multimídia e novos meios de interação interpessoal, além da vantagem de usar a aldeia global como um ambiente para a aprendizagem.

Os fóruns fazem parte de uma vertente da educação que acredita que o aprendizado é feito por meio da experiência e pelo seu compartilhar. Wittgenstein foi um famoso filósofo austríaco, sua crença é de que o conhecimento é construído pela participação do sujeito em seu meio. Essa afirmação contrasta com a visão conhecida como ingênua onde o conhecimento é visto como um conjunto de informações e a mente um espaço para guardá-las [22]. Também contrasta com a visão piagetiana que crê que o conhecimento não é formado por cópias de informações armazenadas, mas com a interpretação pessoal do mundo [22]. É perceptível que os fóruns estão em conformidade com modelo experiencialista, pois trata o conhecimento como algo dinâmico, em constante processo de mudança, e que emerge da troca de experiências de uma rede social colaborativa [16].

Embora as discussões em fóruns e espaços públicos existam há milhares de anos, como nas Ágoras, por exemplo, os fóruns na web abrem espaço para novos meios de integrar pessoas e permitir novos modelos de interação, que variam de acordo com o propósito específico de cada

fórum. Para Kenski [11]:

> *"A característica desta nova forma de ensinar é a ampliação de possibilidades de aprendizagem e o envolvimento de todos os que participam do ato de ensinar. A prática de ensino envolvida torna-se uma ação dinâmica e mista. Mesclam -se nas redes informáticas - na própria situação de produção/aquisição de conhecimentos, autores e leitores, professores e alunos. A formação de "comunidades de aprendizagem" em que se desenvolvem os princípios do ensino colaborativo, em equipe, é um dos principais pontos de alteração na dinâmica da escola. Além disso, as informações coletadas nos diversos ambientes e meios tecnológicos, em permanente transformação, devem ser analisadas e discutidas, não mais como verdades absolutas, mas compreendidas criticamente como contribuições para a construção coletiva dos conhecimentos que irão auxiliar na aprendizagem de cada um."*
>
> —KENSKI

2.2 Histórico

Um dos primeiros fóruns de perguntas e respostas a surgir foi o Answer Point, disponibilizado pelo Ask Jeeves na década de 90. Seu slogan era: "Answer Point é o lugar onde você pode fazer e responder perguntas. Tem uma dúvida? Publique-a! Sabe a resposta? Publique-a!"[19].

De acordo com Jim Lanzone, vice-presidente sênior do AskJeeves, a maior dificuldade era pra incentivar usuários à responder. Com poucas questões respondidas, Jim Lanzone defende a utilidade dos engenhos de busca, pois a maioria das buscas são únicas, ainda que a relevância não seja perfeita, é possível trazer resultados amplos [19]. Além disso, esperar por uma resposta é conflitante com o que o usuário mais precisa, velocidade.

A Ask Jeeves possuia outros produtos além do AnswerPoint, o seu principal era o Ask.com, um buscador concorrente ao do Google. Segundo Jim Lanzone, a empresa decidiu focar em outros aspectos do seu buscador e descontinou o projeto do AnswerPoint [19].

Na mesma semana do fechamento do Answer Point, foi lançado o Google Answers [26]. O Google Answers era um serviço que permitia que as pessoas submetessem perguntas e oferecessem um pagamento por sua resposta, esse pagamento podia variar de 2,50 dólares até 200 dólares. Entretanto, também era possível ver respostas para outras perguntas já respondidas sem efetuar pagamento. Seu fracasso, dentre outros motivos, foi decorrente principalmente do surgimento de serviços de perguntas e respostas gratuitos.

O Yahoo Respostas foi criado pela empresa Yahoo no ano 2005. Foi um dos primeiros a implementar o *crowdsourcing*[1] com sucesso. A empresa implementou um sistema de pontos

[1] Crowdsourcing é um modelo de produção que conta com a força do coletivo para desenvolver soluções.

Figura 2.1 Fórum do AnswerPoint.

para incentivar os usuários à responder perguntas. Com esses pontos, os usuários possuíam mais acesso à plataforma, permitindo-os publicar mais questões e respostas. A pontuação é atribuída pelo autor da pergunta e também pela comunidade que pode manifestar sua opinião [27].

Em 2004, Tim O'Reilly cunhou o termo de Web 2.0, ele se referia à uma web participativa, com conteúdo produzido pelos próprios internautas [10]. Atualmente, com a popularização dos dispositivos móveis, facilitou-se o acesso do usuário à rede e, consequentemente, sua participação. Os usuários passaram a confiar e depender cada vez mais desses meios de interação, dentre eles, os fóruns, o que tornou necessária uma web mais organizada, chamada de Web 3.0 por John Markoff. Na Web 3.0, a informação é estruturada não somente em meios inteligíveis por humanos, mas também por máquinas. Por meio das máquinas, se torna possível um acesso mais rápido à informação. É um uso mais inteligente do conteúdo já disponibilizado online.

2.3 Fóruns Populares

Com o crescimento exponencial da internet, houve espaço para novos fóruns crescerem e se desenvolverem em nichos diferentes. São estes alguns exemplos relevantes.

- **Quora**[2]: É um espaço pra perguntas e respostas de qualquer domínio. A empresa foi criada por Adam D'Angelo e Charlie Cheever e disponibilizada ao público em 2010,

[2]https://quora.com/

tornando-se uma opção mais usada pelos usuários devido ao uso mais consciente dos fóruns, onde perguntas devem ser mais bem elaboradas e bem estruturadas. Perguntas que não são adequadas são reportadas e removidas da plataforma;

Isso tornou-se necessário pois o Yahoo Respostas estava se tornando muito poluído com perguntas vagas [17], como a de estudantes para responder tarefas de casa e com respostas sem sentido, às vezes apenas com "Eu não sei", que serviam para que ainda assim ganhassem pontos na plataforma. Graças à esta nova visão, usuários da plataforma têm um acesso facilitado às perguntas de seu interesse e menos conteúdo duplicado, além de que agora a plataforma consegue melhor recomendar perguntas aos usuários que podem a responder;

- **StackOverflow**: Criado por Jeff Atwood e Joel Spolsky em 2008. Jeff era um desenvolvedor e possuia um blog chamado de Coding Horror[3] onde abordava temas relacionados à programação [1]. No mês de julho de 2008, quando foi criado, Jeff limitou o acesso para aqueles que eram assinantes do blog e no mês de setembro tornou público o acesso.

  O sucesso da plataforma foi tamanho que permitiu que crescesse para um conglomerado de fóruns de perguntas e respostas sobre vários outros tópicos, conhecido com StackExchange, onde são mantidos o próprio StackOverflow, AskUbuntu, SuperUser e mais de 150 outros. Atualmente, a plataforma possui mais de 17 milhões de perguntas.

2.4 Usos Mercadológicos

2.4.1 Vendas Direcionadas

Uma das formas mais comuns de gerar renda é através de anúncios direcionados. Um dos que usam essa abordagem é o Quora, que hoje é avaliado em mais de 1,8 bilhão de dólares [15]. Com a interação do usuário com a plataforma através de buscas, comentários e perguntas, é possível identificar possíveis necessidades que o mesmo possui e, assim, direcionar anúncios adequados para estes.

2.4.2 Recrutamento

Ao responder perguntas, os usuários criam notoriedade pra si. O StackExchange usa essa vantagem tanto para estimular usuários para responder perguntas, quanto pra identificar potenciais para determinadas vagas. Apresentadas em forma de anúncio, as vagas são ofertadas para os usuários de acordo com o seu perfil na plataforma. Da mesma forma, os recrutadores tem

[3] https://blog.codinghorror.com/

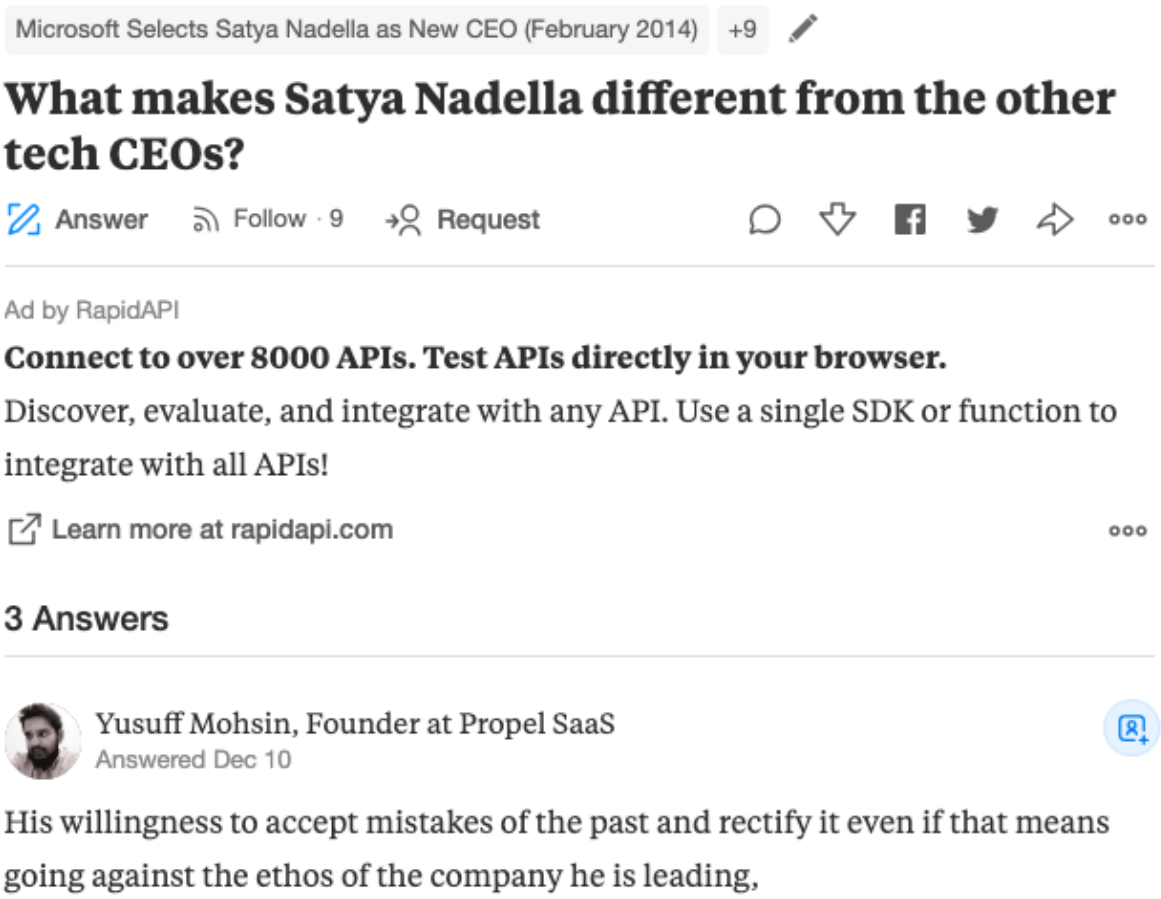

Figura 2.2 Anúncios nas perguntas feitas no website Quora.

acesso à informações mais detalhadas dos candidatos, por conta da participação dos mesmos, que os garante reconhecimento na comunidade e troféus fictícios na plataforma.

2.4.3 Educação à distância

Fóruns digitais estão cada vez mais presentes em Ambientes Virtuais de Aprendizado (AVA) e são normalmente acompanhados de outras ferramentas. Nos ambientes virtuais dúvidas de alunos podem ser compartilhadas na comunidade e respondidas por um grupo maior de pessoas, dentre elas: professores, monitores ou outros alunos. Alguns professores usam os fóruns para estimular a interação entre os alunos e as discussões de temas. Acredita-se que esta parte complementa as aulas que tendem a ser objetivas, enquanto que os fóruns trazem uma abordagem experiencialista [16]. Um dos websites que faz isso é o Moodle[4], que é utilizado por várias universidades ao redor do mundo para ajudar os estudantes a interagirem com o conteúdo das aulas.

[4]https://www.moodle.ufba.br/

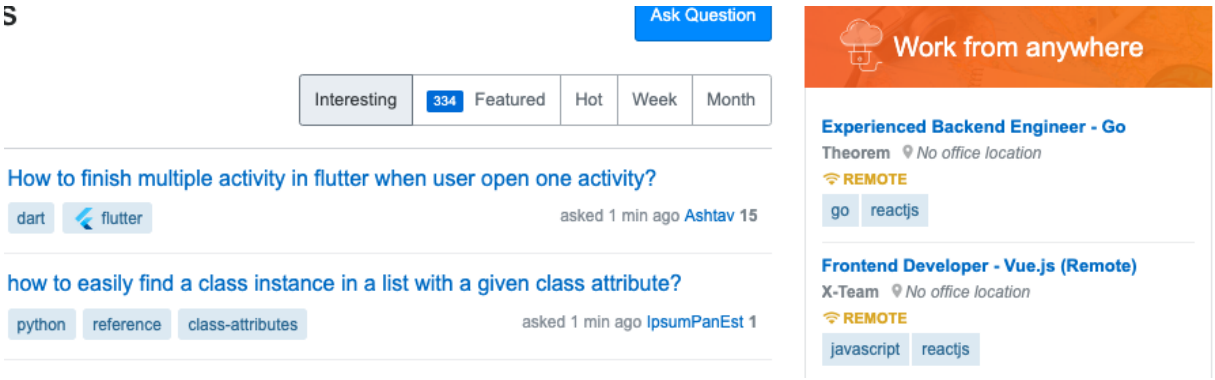

Figura 2.3 Recrutamento usado como forma de monetização no StackExchange.

2.5 Sumário

O capítulo trouxe uma instrução ao leitor sobre a presença e o valor dos fóruns no contexto rotineiro das pessoas. A história foi abordada para apresentar a sequência de eventos que trouxe as transformações que tornaram o modelo de fórum de hoje real. As opiniões de especialistas em educação foram usadas para fundamentar a afirmação de que fóruns têm uma capacidade de ir além das salas de aula e trazer uma forma mais experiencial e mais cheia de opiniões distintas entre aqueles que compartilham essa interação. Ao fim, usos mercadológicos foram adicionados para ilustrar o potencial dessa rede e mostrar o porquê deve-se continuar investindo em soluções relacionadas à mesma.

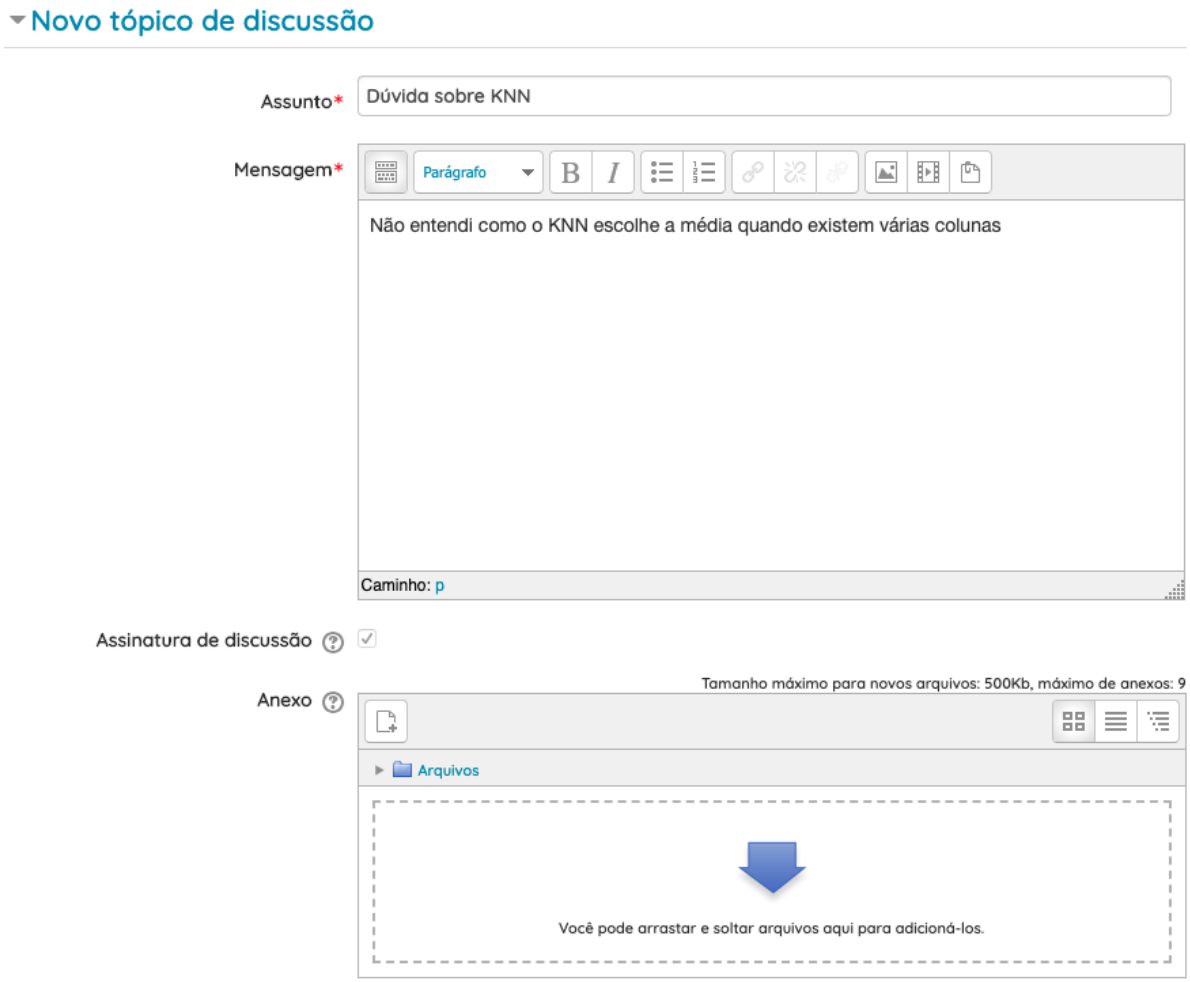

Figura 2.4 Exposição de dúvida em um fórum no Moodle.

3

Recuperação da Informação

Conforme foi descrito no Capítulo 1, na WEB 3.0, o volume de informações cada vez maior tornou necessária a existência de meios para encontrar a informação. O conteúdo, atualmente, não é feito só pra a compreensão humana, mas também para a compreensão das máquinas. Os usuários exigem que as informações desejadas sejam disponibilizadas quase que imediatamente e com precisão. Neste capítulo será abordado tópico de Recuperação de Informação (RI), onde serão vistos as diferenças entre este e Recuperação de Dados (RD), além das técnicas mais comumente utilizadas para realizar suas tarefas. Ao fim, serão abordadas as formas que se avaliam os diferentes metódos de recuperação.

3.1 Introdução

Recuperar informações é um processo que pode ser bastante impreciso, é exigido de uma máquina uma capacidade quase humana, a da interpretação. Por conta disso, muitas vezes é esperado do usuário a capacidade de saber explicar corretamente para uma máquina os seus anseios, já não existe nenhuma tecnologia capaz de interpretar exatamente o pensamento humano nem de compreender por completo a relevância dos resultados que podem ser trazidos.

Além disso, há um requisito de tempo, todo o processo de interpretação das buscas e recuperação dos resultados deve ser feito quase que instantaneamente. Tendo isso em mente, a indexação é uma das principais técnicas utilizadas para facilitar a recuperação e o cálculo de relevância de vários algoritmos [7]. Não somente isso, mas técnicas para divisão da carga de trabalho tornaram-se mais comuns pra lidar com tamanha celeridade dos dados.

3.2 Recuperação de Informação versus Recuperação de Dados

Recuperação de Informação pode ser facilmente confundida com outros meios de acesso à informação. Ao contrário de sistemas baseados em conhecimento, sistemas de recuperação de informação não realizam conclusões a partir dos documentos acessados. Os sistemas baseados

em conhecimento dependem de uma visão de mundo pré-definida para poder realizar inferências e gerar informação, muitas vezes o limitando. Já o propósito geral da recuperação de informação é conseguir entregar ao usuário documentos que possam satisfazer sua necessidade de informação [21], permitindo uma visão mais ampla. As diferenças desses tipos de recuperação podem ser visualizadas na Tabela 3.1.

Tabela 3.1 Diferença entre sistemas RI e RD[14].

Características / Métodos	**RI**	**RD**
Combinação exata		x
Alta sensibilidade a erros		x
Tratamento semântico	x	
Busca em dados não estruturados	x	
Inferência Dedutiva		x
Consulta de sintaxe controlada		x
Consulta com linguagem Natural	x	
Mais utilizado em meios acadêmicos	x	
Comum em Produtos Comerciais		x

A recuperação da informação permite que existam diferenças entre o que foi requisitado pelo usuário e o que foi apresentado como resultado, desde que haja uma similaridade. A Recuperação de Dados (RD), por outro lado, tenta garantir que a consulta irá trazer apenas o que foi especificado na consulta. Por conta disso, embora a recuperação da informação esteja sujeita à imprecisão e possíveis erros, a mesma é capaz de abranger um conteúdo maior e trazer resultados relacionados que não dependam tanto da capacidade do usuário de gerar uma consulta adequada, enquanto que a recuperação de dados, exige que o usuário tenha a capacidade de expressar perfeitamente sua necessidade e tem uma falha total quando não o consegue. Este mérito deve-se a capacidade de lidar com textos semi-estruturados de linguagem natural, enquanto que a recuperação de dados, acessa um ambiente, que embora seja estruturado, é limitado, como é o caso de um banco de dados relacional, por exemplo [7].

Seus casos de uso diferem, para este projeto, onde a maior parte do conteúdo é textual e desestruturado, é exatamente onde se encontra a informação desejada. As pessoas que usam sistemas de recuperação de informação procuram muito mais além do que apenas aquela exata combinação de palavras [7], inclusive, muito frequentemente o autor da consulta não detém informação suficiente sobre o domínio da sua busca para realizar buscas mais exatas. Sua consulta serve para encontrar um espaço mais amplo de respostas onde suas palavras-chave estão inseridas, ou talvez ainda, sinônimos ou referências similares, ainda que não sejam idênticos ao que foi especificado. Entretanto, mais comumente, são usados sistemas com recuperação de dados, onde informações relacionadas não são capazes de resolver o problema de um usuário.

Aplicações com funcionalidades de cadastro, acesso à bancos, operações administrativas e outros, são muito comuns e não fariam sentido ao recuperar informações senão as exatamente definidas pelo usuário.

3.3 Estratégias de Recuperação

Um dos principais objetivos na pesquisa de recuperação de informação é a capacidade de entender e formalizar os processos que ocorrem quando uma pessoa toma a decisão de que um pedaço de texto é relevante para sua necessidade de informação [3]. Entretanto, a tecnologia atual ainda é limitada nesse aspecto e tem se optado por modelos matemáticos que, então, são comparados com ações humanas para verificar sua eficácia. Nesses modelos, são atribuídas medidas de similaridade de consulta com documentos, onde uma maior similaridade indica uma maior relevância [8]. Todavia, o cálculo da similaridade envolve novos problemas, deve-se levar em conta sinônimos, entender sarcasmos, erros de ortografia, grau de importância de um termo para um documento e etc. Para endereçar esses desafios, algumas estratégias foram criadas.

3.3.1 Modelo Booleano

O modelo booleano foi um dos primeiros usados em engenhos de busca. Ele é limitado no aspecto de ranking e na recuperação de documentos que não são completamente adequados ao que se foi especificado na busca. Sua consulta é feita utilizando operadores booleanos (AND, OR, NOT), que determinam o que pode e o que não pode ser recuperado. Todos os documentos que cumprem as especificações, são recuperados e considerados de mesma relevância. Os documentos que não cumprem as especificações são descartados, ainda que cumpram parcialmente as especificações. Ele pode ser muito valioso quando usado em ambientes que auxiliem o usuário a criar uma consulta, pois a qualidade da consulta será ainda mais determinante para a qualidade dos resultados.

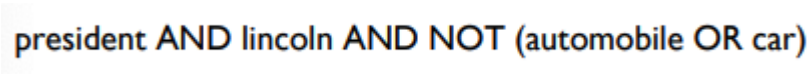

Figura 3.1 Exemplo de consulta no modelo booleano [3].

Na Figura 3.1 é exemplificada uma consulta neste modelo. Nela é determinado que documentos com as palavras *president* e *lincoln* são considerados relevantes desde que não tenham as palavras *automobile* ou *car*. No caso, procuram-se informações sobre o presidente Lincoln, mas o autor da consulta percebeu que Lincoln pode ser associado à um modelo de carro, por isso adicionou essa restrição. São definições como estas que definem as consultas no modelo booleano.

3.3.2 Modelo de Espaço Vetorial

Este modelo é considerado um dos mais simples e nele que se encaixa um dos métodos mais conhecidos de cálculo de similaridade, a similaridade do cosseno. Nele os documentos e termos são dispostos em uma matriz que indica o número de ocorrências de cada termo em cada documento. Na Tabela 3.3 pode-se ver o resultado dessa operação sobre as frases descritas na Tabela 3.2.

Tabela 3.2 Frases de exemplo.

1	Bola de brinquedo
2	Mar azul e céu azul
3	Brinquedo azul
4	Azul do mar

Com estes dados, cada documento pode ser representado como vetor ou ponto no espaço, onde cálculos de similaridade podem variar com diferentes medidas de distâncias como, por exemplo a euclideana ou a similaridade do cosseno. Algumas outras atribuições também podem ser adicionadas, por exemplo, acredita-se que termos muito recorrentes não devem ter tanta influência sobre o que se trata um determinado documento. Por exemplo, a palavra *azul* na Tabela 3.3 repete-se muito, nesses casos, pode-se usar uma ponderação inversamente proporcional para que, ao multiplicar esse valor com a frequência do termo, tenha seu valor de influência diminuído. Logo, de acordo com a Tabela 3.3, o documento 4 vai ter muito mais similaridade com documentos que possuam o termo *mar* do que com os que possuem o termo *azul*.

Tabela 3.3 Frequência de termos nos documentos.

Documento	brinquedo	bola	azul	mar	céu
1	1	1	0	0	0
2	0	0	2	1	1
3	1	0	1	0	0
4	0	0	1	1	0

Além disso, apenas a frequência de um termo em um documento, não garante a relevância do mesmo. Se um termo t aparece duas vezes mais no documento a do que no documento b não significa que o a seja duas vezes mais relevante [12]. Deve-se ao fato de que relevância não é linearmente proporcional à frequência de termos. Por conta disso, normalmente a ponderação da influência da frequência um termo é logarítmica.

3.3.3 Modelos Probabilísticos

Um dos principais usado no mercado, este modelo usa probabilidade para realizar seus rankings. Dadas algumas suposições, como a de que a relevância de um documento é independente de outros documentos, pode-se ver que ao enumerar os resultados em ordem decrescente de probabilidade de relevância ao usuário, onde as probabilidade são estimadas, a efetividade do sistema será a melhor para o que é possível obter com os dados existentes [20]. Entretanto, não é determinado como se pode obter essas probabilidades. Existem vários modelos probabilísticos e cada um deles propõe uma maneira diferente de alcançar esses números.

Nestes modelos, a recuperação da informação é dada como um problema de classificação, onde um documento D é dito relevante se e somente se a probabilidade de D ser relevante é maior do que ele não ser relevante. Sistemas que classificam documentos assim são chamados de classificadores Bayesianos. Logo, de acordo com a Regra de Bayes[1], a probabilidade de D ser relevante pode ser expressa na Equação 3.1. Onde *P(R)* é a probabilidade de qualquer documento pertencer ao conjunto dos relevantes e *P(D)* é usado para normalizar a equação. A maior parte do problema se dá em calcular *P(D|R)*, para realizar essa tarefa, o conjunto R vai precisar ser definido pelos termos usados na consulta, já que é a única informação que se tem sobre os termos relevantes, assim *P(D|R)* pode ser definido como a soma dos pesos de todos os termos em comum entre o documento e a consulta. Os pesos de cada termo são atribuídos de diferentes formas pelos modelos, um exemplo é o Inverse Document Frequency (IDF) que usa uma escala logarítmica inversa à frequência do termo no corpus para calcular seu peso.

$$P(R|D) = \frac{P(D|R)P(R)}{P(D)} \tag{3.1}$$

Por conta da dificuldade de obter as probabilidades corretas, a deficiência da informação de termos relevantes pode prejudicar muito a qualidade da recuperação. Melhores rankings podem ser obtidos com uso de *relevance feedback*, onde usuários dão sua opinião sobre os documentos recuperados quanto à sua relevância e, assim, permite um cálculo mais preciso da probabilidade de um termo pertencer ao conjunto relevante.

3.3.3.1 Okapi BM25

Esta fórmula é um exemplo de modelo probabilístico e foi o algoritmo escolhido para realizar os cálculos de similaridade nesse projeto. Existem algumas variações na sua equação, mas a que iremos usar será a definida na Equação 3.2. Na equação define-se o cálculo da similaridade de um documento D e uma consulta q.

$$sim(D,q) = \sum_{n=1}^{n} IDF(q_i) \frac{f(q_i(k+1)}{f(q_i,D) + k(1-b+b\frac{fieldLen}{avgFieldLen})} \tag{3.2}$$

[1] Thomas Bayes foi um matemático que criou os Teoremas de Bayes

$$IDF(q) = ln\left(1 + \frac{docCount - f(q_i) + 0.5}{f(q_i) + 0.5}\right) \tag{3.3}$$

- $\boldsymbol{q_i}$: É o i-ésimo termo da consulta;
- $\boldsymbol{IDF(q_i)}$: Função de cálculo do peso do termo definida na Equação 3.3;
- $\boldsymbol{fieldLen}$: Número de termos que existem no documento D;
- $\boldsymbol{avgFieldLen}$: Número médio de termos nos documentos do corpus;
- $\boldsymbol{b}$: Define o quanto de normalização sobre o tamanho do documento será utilizada. Varia de 0 até 1;
- $\boldsymbol{f(q_i, D)}$: A frequência do termo q_i no documento D;
- $\boldsymbol{k}$; Determina a saturação da frequência dos termos, ou seja, o quanto a frequência de um termo pode influenciar na pontuação final;
- $\boldsymbol{docCount}$: Número de documentos no corpus;
- $\boldsymbol{f(q_i)}$: Número de documentos no corpus que possuem o termo q_i.

A ideia desse modelo se dá primariamente sobre as variáveis b e k. A variável b é responsável por regular os efeitos da normalização dos documentos, sendo que documentos menores tendem a receber pontuações maiores e documentos maiores, pontuações menores. Valores maiores de b serão usados quando se deseja dar maior importância à cada termo em documentos menores e menos importância à cada termo quando o documento for mais longo. Valores menores de b serão usados quando o tamanho do texto não deve impactar tanto a influência de cada termo sobre o cálculo da similaridade.

Enquanto isso, a variável k regula a curva de influência da frequência de termos de uma consulta. Essa variável define a saturação da repetição de um termo, ou seja, quantas repetições são suficientes para que ele já não tenha muito mais influência sobre o domínio daquele tópico. A intuição disso é que certas palavras podem ser muito comuns em um texto e não serem o principal foco, o que em valores muito altos de k esses termos teriam muito mais influência. Já em valores menores de k, a repetição dos termos pouco importaria para o ranking e com $k = 0$ só se seria checado a presença ou ausência de um termo, independente de sua frequência. Na Figura 3.2 pode se ver como a curva desse algoritmo se comporta ao ser comparado com o TF/IDF, outro método utilizado para cálculo de similaridade.

Assim como outros modelos probabilísticos, esse algoritmo pode ser modificado para ter uma melhor visão dos documentos relevantes, a partir dessa visão é possível obter melhores resultados ao melhor valorar os pesos e probabilidades dos termos.

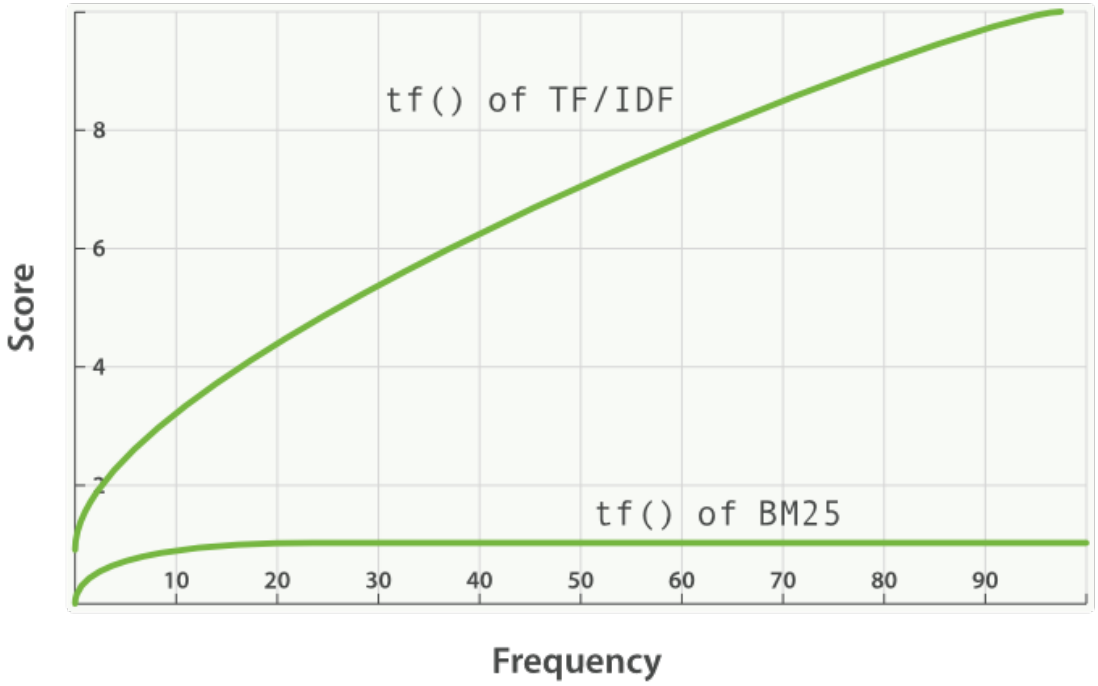

Figura 3.2 Saturação da frequência dos termos no BM25 e sua comparação com o TF/IDF [2].

3.4 Métodos de avaliação

A partir das opções disponibilizadas para construir sistemas de recuperação da informação, resta buscar meios para os avaliar. Considerando que não existe método perfeito para todos os problemas é necessário se conduzir uma pesquisa para averiguar-se qual a melhor solução para determinado problema. Para fazer isso, determinam-se métricas que irão indicar se há conformidade da solução com o problema.

Uma das formas de se executar um experimento para avaliar o sistema é utilizar um conjunto de buscas e seus resultados esperados. A partir dessas buscas, avaliam-se as condições com que foram apresentados os resultados, levando em consideração desde o tempo que custou para achar essas informações até a quantidade de esforço necessária do usuário para encontrar uma informação relevante dentro do conjunto de resultados. São estas algumas das métricas mais utilizadas:

- **Cobertura**: Representa o quanto de informações relevantes foram apresentadas como resultados. Considerando que para uma busca X existam 10 resultados relevantes, se apenas 3 foram apresentados para o usuário, a cobertura é 30%;
- **Ranking**: Métricas que utilizam o rank como parâmetro, avaliam o sistema de acordo com a posição dos itens nos resultados da busca. Um exemplo deste é o Média do Rank Recíproco, onde para cada busca, a posição do primeiro documento

relevante chamada K é usada pra determinar o que é $\frac{1}{K}$, em seguida, deve-se continuar executando mais N buscas onde encontraremos a média desses valores a partir de $\frac{(\frac{1}{K_1}+\frac{1}{K_2}+...+\frac{1}{K_n})}{N}$ para encontrar o Média do Rank Recíproco;

- **Precisão**: Informa o quanto de conteúdo relevante existe dentre os documentos apresentados. Por exemplo, se dentre 10 resultados retornados, apenas 2 resultados forem relevantes, temos uma precisão de 20%;

- **Tempo de resposta**: É o intervalo médio entre o momento da consulta e a apresentação dos resultados;

- **Esforço do usuário**: Representa o esforço despendido pelo usuário para obter resultados em sua busca. Isso pode incluir diversos fatores que variam de acordo com a especificação do problema, sendo exemplos deles: número de linhas lida, número de ações necessárias para executar a busca, número de documentos abertos até encontrar o resultado desejado e outros.

3.5 Tecnologias e Frameworks

Para melhor lidar com os desafios de Recuperação da Informação, foram criadas novas tecnologias e *frameworks*. Essas tecnologias implementam os conceitos aqui utilizados e permitem modificações para que se adaptem melhor ao problema em questão. Os seguintes *frameworks* são os mais populares.

- **Lucene**: O Apache Lucene é um motor de busca escrito em Java com ferramentas para busca em texto. É uma API open source disponível para download gratuito. Possui uma linguagem própria que permite fazer buscas parametrizadas com expressões regulares;

- **Solr**: Solr é uma aplicação web construída ao redor do Lucene que adiciona funcionalidades como: busca geospacial, replicação, cacheamento e interfaces de administração;

- **Elasticsearch**: Assim como Solr, o Elasticsearch é uma aplicação construída com o uso do Lucene. O Elasticsearch é distribuído pela empresa Elastic e possui diversos plugins gratuitos e pagos para adicionar-se ferramentas de administração. Diferencia-se do Solr por focar mais em aspectos da administração do banco de dados e na escalabilidade.

3.6 Sumário

Durante esse capítulo, foram enunciados conceitos de RI e comparados com a área de RD, que, embora semelhante, difere em alguns aspectos e propósitos. Algumas das estratégias de recuperação mais utilizadas foram abordadas, tendo um aprofundamento maior no modelo probabilístico, o qual será usado na proposta de solução deste trabalho. Em seguida, foram trazidas métricas capazes de avaliar sistemas de recuperação da informação, desde aspectos de usabilidade à aspectos relacionados à qualidade dos resultados. Ao fim, algumas ferramentas famosas do mercado foram listadas e explicadas quanto ao seu propósito. No capítulo seguinte, esses conhecimentos serão utilizados para construir a proposta.

4

Um Sistema de Busca para Fóruns de Discussão na Web

Com o objetivo de facilitar o acesso de usuários à fóruns, este trabalho traz como proposta um buscador que além de tratar a similaridade semântica, também leva em consideração a opinião da comunidade. Durante esse capítulo serão analisadas e explicadas as decisões tomadas para construir esta proposta. Inicialmente será definida uma visão geral da solução, onde serão indicadas as formas de aproximar um sistema pré-existente ao modelo proposto. Em seguida, será indicado o modelo que é necessário que o fórum use para poder fazer uso da solução e suas possíveis adaptações. Após, será explicada a solução em ação, passando pelo pré-processamento de novos itens, a indexação desses mesmos itens para a sua recuperação que será realizada no futuro e, por fim, o modelo de busca utilizado para julgar a relevância dos itens que devem ser recuperados. Após a conclusão da explicação da solução, as tecnologias utilizados que tornaram este trabalho possível serão indicadas e explanadas quanto à sua função na solução.

4.1 Requisitos

Os requisitos são definidos como as funções que a solução deve ser capaz de exercer e suas restrições [23]. Os requisitos são documentados com o objetivo de explicitar as necessidades que devem ser cumpridas para que o sistema esteja considerado pronto e permitir consulta posterior. A linguagem utilizada deve ser clara e objetiva para garantir que a implementação esteja sujeita ao mínimo de ambiguidades possíveis. Os requisitos se dividem entre funcionais e não funcionais. Os requisitos funcionais falam à respeito do comportamento do sistema e como os usuários podem interagir com o mesmo, enquanto que os requisitos não funcionais são usados para garantir métricas de qualidade e segurança como, por exemplo, o tempo de resposta da aplicação. Na Tabela 4.1 são enunciados os requisitos funcionais e também os não funcionais.

Tabela 4.1 Requisitos da solução proposta.

Código	Nome	Descrição
RF01	Buscar documentos	Um usuário pode usar linguagem natural para enunciar seu problema e receber resultados relacionados à sua busca.
RF02	Comparar relevância	A solução deve ser capaz de entregar resultados ordenados pela maior relevância para o usuário até a menor.
RF03	Permitir adaptações	A forma que a relevância é calculada deve ser modificável para adaptar-se à diferentes fóruns.
RF04	Navegar sobre os resultados	O usuário pode escolher dentre os resultados encontrados.
RNF01	Acoplável	Administradores devem ser capaz de interagir com a solução para criar sua própria interface ou acoplar à funcionalidade ao site já criado.
RNF02	Migração contínua	Novos dados inseridos na base original do fórum devem ser incorporados
RNF03	Independente	Erros não devem propagados para o fórum e o processo de migração não pode deixar o fórum indisponível
RNF05	Fácil configuração	O processo para acoplarem a solução ao seu sistema pré-existente deve ser simples.

4.2 Visão Geral da Solução

Neste projeto será considerada uma arquitetura plugável à uma base pré-existente. A visão geral pode ser analisada na Figura 4.1. Conforme descrito na própria figura, os fóruns já possuem uma base de dados própria (Componente 2) e uma interface (Componente 1) sendo utilizadas. Esta parte se manterá isolada, sem necessidades de modificações, com o objetivo de garantir a integridade do sistema original e, como dito anteriormente, funcionar como um plugin. A solução tem acesso às informações da base através de uma ferramenta de migração (Componente 3) que irá buscar as informações necessárias na base de dados do fórum para popular o engenho de busca com as informações no formato especificado na Tabela 4.2 (Componente 4), que, por sua vez, irá pré-processar os tópicos recebidos e os salvar em um banco de dados próprio (Componente 5), para que, enfim, o usuário tenha acesso à funcionalidade de busca através de uma interface (Componente 6) que deve ser criada pelos administradores do fórum, sendo que a mesma pode ser unida à interface original (Componente 1) para evitar a necessidade de interfaces distintas.

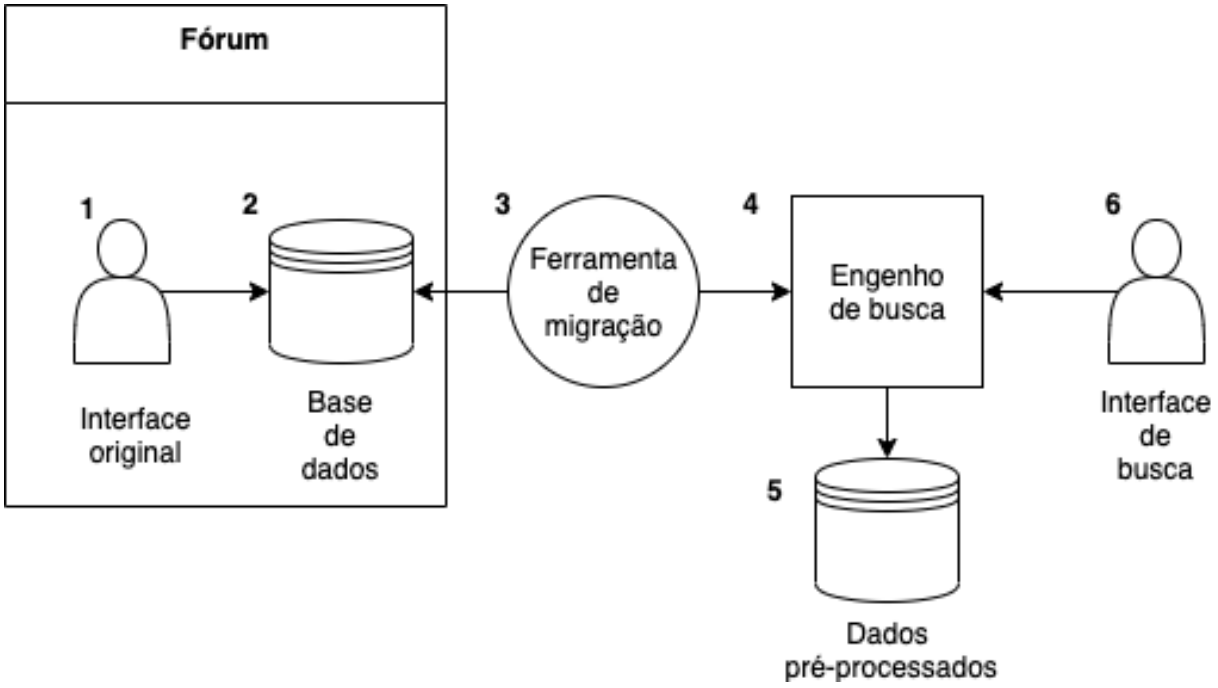

Figura 4.1 Diagrama de dependências da solução proposta.

4.3 Modelagem dos Dados

Nos fóruns observados percebe-se que os tópicos podem ter tanto título e corpo. Entende-se como corpo, o conteúdo textual descritivo da primeira postagem. Ao permitir que usuários votem indicado a utilidade ou qualidade de um documento, é adicionado um conceito de pontuação. Fóruns que não implementam recursos de avaliação podem usar outros parâmetros para definir a pontuação de um documento, como número de visualizações ou alguma outra métrica de engajamento, por exemplo. Logo, para usar a solução, os administrados dos fóruns devem configurar a migração para seguir o formato descrito na Tabela 4.2.

Tabela 4.2 Formato exigido para a modelagem de um tópico.

Propriedade	**Tipo**	**Exemplo**
Título	texto	Como instalar o Linux?
Corpo	texto	Quero saber como faço pra instalar o Ubuntu 17.04.
Pontuação	numérico	23

4.4 Pré-processamento

O pré-processamento é responsável por transformar a entrada, neste caso um tópico, em uma estrutura mais adequada para a análise e recuperação desses mesmos itens mais à frente. Por meio de tecnologias de Processamento de Linguagem Natural (PLN), é possível extrair mais

sentido de textos construídos por humanos e os estruturar de maneiras mais adequadas para uma máquina também ser capaz de o compreender.

Neste problema, cada tópico passará por transformações antes de serem armazenados. Considere os títulos descritos na Tabela 4.3 no seu estado inicial. Estes irão ilustrar o pré-processamento. Vale ressaltar que o mesmo processo que está sendo descrito com os títulos dos tópicos também ocorre com os corpos dos tópicos.

Tabela 4.3 Frases de exemplo.

1	I will organize this room
2	All rooms are organized and clean
3	Cleaners are very effective
4	I will open this window

4.4.1 Filtros de caracteres

Na web é comum encontrar documentos com tags HTML. Para este projeto estas tags não terão importância, portanto serão removidas e apenas será deixado apenas o texto. É possível verificar o resultado exemplificado na Tabela 4.4.

Tabela 4.4 Frases de exemplo após os filtros de caracteres.

	Antes	**Depois**
1	I will organize this room	I will organize this room
2	All rooms are organized and clean	All rooms are organized and clean
3	Cleaners are very effective	Cleaners are very effective
4	I will open this window	I will open this window

4.4.2 Tokenização

O processo de tokenização é responsável por dividir um texto em unidades menores. Essas unidades podem ser palavras, caracteres, cadeia de palavras ou cadeia de caracteres.

Para este projeto foi usado a tokenização dividindo o conteúdo do texto em palavras usando o algoritmo Unicode Text Segmentation [4].

4.4.3 Filtros de tokens

4.4.3.1 Uniformização

Foi usada nesse projeto uma transformação responsável por deixar todos os caracteres minúsculos.

Tabela 4.5 Frases de exemplo após a tokenização.

	Antes	**Depois**
1	I will organize this room	[I, will, organize, this, room]
2	All rooms are organized and clean	[All, rooms, are, organized, and, clean]
3	Cleaners are very effective	[Cleaners, are, very, effective]
4	I will open this window	[I, will, open, this, window]

Tabela 4.6 Frases de exemplo após as transformações.

	Antes	**Depois**
1	[I, will, organize, this, room]	[i, will, organize, this, room]
2	[All, rooms, are, organized, and, clean]	[all, rooms, are, organized, and, clean]
3	[Cleaners, are, very, effective]	[cleaners, are, very, effective]
4	[I, will, open, this, window]	[i, will, open, this, window]

4.4.3.2 Remoção de Stop Words

Existem palavras que adicionam pouco valor semântico ao texto, são conhecidas como *stop words*. *Stop words* são palavras como: isso, um, a, o, que. Estas também serão removidas. Existem diversas formas de as detectar em um corpus, as mais comuns são com uso de listas com *stop words* predefinidas ou definindo um limiar máximo de ocorrências que uma palavra pode ter. Assume-se que palavras que ocorrem frequentemente em vários documentos não agregam muito conteúdo ao texto.

Tabela 4.7 Frases de exemplo após a remoção de *stop words*.

	Antes	**Depois**
1	[i, will, organize, this, room]	[organize, room]
2	[all, rooms, are, organized, and, clean]	[rooms, organized, clean]
3	[cleaners, are, very, effective]	[cleaners, effective]
4	[i, will, open, this, window]	[open, window]

4.4.3.3 Stemming

O propósito do *stemming* é reduzir a variação morfológica das palavras [25]. Documentos podem usar formas diferentes de uma palavra, por exemplo, um deles pode usar organizar, outro pode usar organizando e outro organizado. Embora as palavras não sejam idênticas, elas trazem consigo um sentido similar, o que pode ser útil em um sistema de recuperação da informação onde se deseja conteúdo relacionado, ainda que não idêntico.

O algoritmo mais comum de Stemming é o Porter [18]. Este algoritmo usa fases com conjuntos definidos de regras para definir as transformações que uma dada palavra irá passar. A Figura 4.2 mostra algumas das regras executadas na primeira fase do algoritmo.

(F)

Rule			Example		
SSES	→	SS	caresses	→	caress
IES	→	I	ponies	→	poni
SS	→	SS	caress	→	caress
S	→		cats	→	cat

Figura 4.2 Regras usadas na primeira fase do Porter Stemmer.

Pode-se ver que agora, que as frases descritas na Tabela 4.8, possuem mais palavras em comum do que anteriormente, isso garante uma busca mais abrangente, permitindo o usuário ter resultados com palavras diferentes da sua, mas que ainda acessem o mesmo domínio. Entretanto o resultado não é perfeito, palavras relacionadas podem acabar tendo finais diferentes e palavras diferentes podem possuir o mesmo resultado.

Tabela 4.8 Frases de exemplo após o stemming.

	Antes	**Depois**
1	[organize, room]	[organ, room]
2	[rooms, organized, clean]	[room, organ, clean]
3	[cleaners, effective]	[cleaner, effect]
4	[open, window]	[open, window]

4.4.4 Itens pré-processados

Percebe-se na Tabela 4.9 o quanto cada uma das frases mudou e agora é possível verificar alguns padrões se repetindo onde não seria tão fácil se identificar previamente.

Tabela 4.9 Frases de exemplo ao final do pré-processamento.

	Inicial	**Final**
1	I will organize this room	[organ, room]
2	All rooms are organized and clean	[room, organ, clean]
3	Cleaners are very effective	[cleaner, effect]
4	I will open this window	[open, window]

Na Tabela 4.10 percebe-se o mesmo processo aplicado aos campos título e corpo de um tópico.

4.5 Indexação e Recuperação

Concluído o pré-processamento, deve-se armazenar estes documentos, de forma que seja possível os recuperar em outro momento. Entretanto, métodos comuns encontrados na maioria

Tabela 4.10 Tópico após o pré-processamento.

	Inicial	**Final**
Título	O brinquedo é azul?	[brinq, azul]
Descrição	Meu brinquedo é da cor do mar?	[brinq, cor, mar]
Pontuação	7	7

dos bancos de dados são inadequados para lidar com texto já que os mesmos dependem de estratégias de RD e não RI.

Índices invertidos foram criados para serem uma forma rápida para lidar com dados massivos de texto. Para cada termo, é salvo o número de ocorrências e em que documentos eles ocorrem, conforme descrito na Tabela 4.11. Para cada campo de um tópico, existe um índice, ou seja, além do índice dos títulos na Tabela 4.11, ao realizar o pré-processamento do corpo, também haverá um índice para os corpos.

Tabela 4.11 Frases de exemplo no índice invertido.

Termo	**Frequência**	**Documentos**
organ	2	1, 2
room	1	2
clean	1	2
cleaner	1	3
effect	1	3
open	1	4
window	1	4

Ao executar uma busca, o texto informado pelo usuário passa pelo mesmo processo de preprocessamento. Dessa forma, os tokens gerados ao final do pré-processamento serão usados para encontrar resultados na base que possuam um ou mais tokens em comum. Estes índices serão usados para também otimizar o processo de ranking, que o usará para avaliar a relevância de cada um dos termos.

4.6 Modelo de Busca

Na seção anterior foi explicado como encontrar documentos relacionados. Nesta seção será explicado como definir a prioridade entre os selecionados, ou seja, determinar quais devem aparecer primeiro.

Será usado o algoritmo Okapi BM25, uma função de ranking baseada nos modelos probabilísticos conforme descrito no Capítulo 3. Com o uso somente do cálculo de similaridade,

já é possível ordenar os resultados. Entretanto, o Okapi BM25 não leva em consideração a pontuação adquirida pelo tópico.

A relevância *Rel* de um tópico x para uma busca b é definida pela média dos valores de similaridades obtidos pela função *sim* da busca com título t e da busca com corpo c multiplicados com a pontuação p obtida pelo tópico. Além disso, também é definido um parâmetro i para regular a influência da similaridade no cálculo. Este cálculo é formalmente apresentado na Equação 4.1 e na Tabela 4.12 é exemplificado o processo de ranking com dados fictícios.

$$Rel(b,x) = \left(\frac{sim(t_x,b) + sim(c_x,b)}{2} \right)^i p_x \tag{4.1}$$

Tabela 4.12 Cálculo de relevância das frases de exemplo. $i = 4$.

Tópico	Sim. Título	Sim. Corpo	Pontuação	**Relevância**
1	3	11	7	102487
2	0	1	1600	100
3	0	0	327	0
4	4	4	80	20480

4.7 Tecnologias Utilizadas

Para tornar essa aplicação real foi usado um conjunto de ferramentas que auxiliam os processos descritos anteriormente nesse capítulo.

- **Elasticsearch**: Este engenho de busca já mencionado no capítulo 3 foi configurado para trabalhar com fóruns. O Elasticsearch[1] foi criado pela empresa Elastic[2] e é uma ferramenta open source e gratuita, ainda que pertença à uma empresa privada. Ele pode ser configurado com diferentes plugins que interagem desde o processo de indexação até a sua administração. É o engenho de busca mais popular do mercado devido suas ferramentas empresariais, frequentemente usado para hospedar os logs dos sistemas. Já funciona como serviço na AWS[3] e no GCP[4];

- **Logstash**: Criado inicialmente com o objetivo de tratar os logs das aplicações, o Logstash[5] é uma ferramenta produzida também pela Elastic. O mesmo é capaz de extrair diversas informações de fontes de dados, atualmente não só mais de logs, mas também de bancos de dados, arquivos, páginas web e outros, tratar esses dados com

[1] https://www.elastic.co/products/elasticsearch
[2] https://www.elastic.co/
[3] https://aws.amazon.com/
[4] https://cloud.google.com/
[5] https://www.elastic.co/products/logstash

operações desde simples expressões regulares até complexos joins em sistemas de banco de dados para, enfim, os salvar em uma outra localidade, que pode ser um arquivo ou, mais comumente utilizado, no Elasticsearch. Foi usado para migrar os dados do banco de dados para o Elasticsearch;

- **Kibana**: Para finalizar, o Kibana[6] é o último participante do grupo ELK. Esta é uma ferramenta também produzida pela empresa Elastic que cuida de visualização de dados com suporte a linguagem Lucene[7]. Esta também é capaz de monitorar o Elasticsearch através do seu Application Perfomance Monitoring (APM) e gerar visualizações a partir disso. Conta com um painel de gerenciamento do Elasticsearch para realizar operações críticas e não-críticas e também com um espaço para realizar benchmarks e testes de operações específicas;

- **MySQL**: O banco de dados MySQL[8] é um dos mais populares bancos relacionais existentes. Neste projeto, ele foi responsável por manter a base de dados organizada, e também por representar um fórum já existente com uma base própria. Suas funcionalidades permitiram uma rápida iteração no projeto, graças à facilidade de extrair novas informações ao mudar os requisitos. Este banco de dados também conta com estruturas de índices, que garantiu que as migrações fossem feitas rapidamente. Sua robustez e tempo de existência fazem diferença, possui estruturas e ferramentas que usam a performance das máquinas para distribuir as cargas de trabalho de forma mais eficiente;

- **Docker**: Docker[9] é um ambiente de virtualização que faz uso de contêineres. Foi criado com o objetivo de criar ambientes prontos para a produção, onde a configuração de um sistema é definida em um arquivo chamado Dockerfile e, ao ser executado, configura um contâiner com as definições impostas. É uma opção que tem sido considerada a sucessora ao de uso de máquinas virtuais, pois usa uma tecnologia que garante o isolamento com o resto do sistema sem comprometer tanto a performance. Neste projeto, foi usado para rapidamente ter acesso às outras tecnologias descritas aqui e poder construir e reconstruir ambientes de forma simples;

- **Python**: Esta linguagem de programação é comumente usada para aplicações de RI e de Natural Language Processing (NLP), graças à um grande número de bibliotecas, como NLTK[10], Scikit-learn[11], Pandas[12], NumPy[13] e outras, além de uma comunidade

[6] https://www.elastic.co/products/kibana
[7] http://lucene.apache.org/
[8] https://www.mysql.com/
[9] https://docker.com/
[10] https://www.nltk.org/
[11] https://scikit-learn.org/
[12] https://pandas.pydata.org/
[13] http://www.numpy.org/

participativa. Python[14] foi criada em 1991 e tem sido usada em projetos de diferentes tipos, é extremamente popular e considerada uma linguagem fácil para estudantes aprenderem como primeira linguagem de programação. A linguagem foi utilizada para a realização de buscas, testes e análises nesta pesquisa.

4.8 Sumário

Neste capítulo apresentou-se uma visão geral sobre o processo de desenvolvimento da solução, métodos e tecnologias utilizadas. Os requisitos funcionais e não-funcionais foram listados para definir as tarefas que a proposta deve ser capaz de cumprir. A arquitetura permitiu visualizar a forma como o buscador poderia se acoplar à sistemas pré-existentes sem causar grandes problemas. A partir disso, a modelagem do problema e cada um dos passos para a sua execução são explicados desde o pré-processamento, até a indexação e recuperação.

[14] https://www.python.org/

5 Avaliação

Neste capítulo são apresentadas as formas de avaliação utilizadas para julgar o cumprimento do propósito deste trabalho. Espera-se que, com a solução proposta, seja possível entregar ao usuário documentos mais relevantes e em melhores posições do que com as soluções mais simples. Nas próximas seções são explicadas: o conjunto de dados escolhido e como ele se adapta ao problema, a metodologia utilizada para conduzir o experimento, métricas de avaliação com o objetivo de julgar objetivamente os resultados, para que, ao fim, haja uma discussão sobre a utilidade do trabalho e possíveis melhorias.

5.1 Conjunto de Dados

O conjunto de Dados escolhido corresponde aos arquivos do AskUbuntu, que foram acessíveis graças à empresa StackExchange que disponibiliza as postagens dos seus sites gratuitamente. Os dados disponíveis no StackExchange foram escolhidos devido à grande diversidade de conjuntos de dados disponíveis, existindo mais de 150 sites disponíveis, variando de tópicos sobre jogos de tabuleiro, língua inglesa até os mais famosos que tratam problemas de programação. A complexidade é percebida quando o site Freelancing[1] possui pouco mais de 1500 (mil e quinhentas) perguntas, enquanto o StackOverflow[2] já alcançou mais de 17 (dezessete) milhões de perguntas.

Durante esse trabalho, haverá um enfoque maior no AskUbuntu, pois o mesmo possui um número aproximado de 300 (trezentas) mil perguntas, o que garantiu um conjunto de teste considerável e, ao mesmo tempo, não tão grande à ponto de inviabilizar os experimentos. Este site, possui perguntas e respostas relacionadas ao sistema operacional Ubuntu. Os usuários utilizam o site com o propósito principal de sanar problemas encontrados com o uso deste sistema operacional e para aprender o funcionamento de determinadas ferramentas. São tópicos frequentes àqueles relacionados à instalação do sistema operacional e outros programas, resolução de mensagens de erros emitidas por programas e configuração de arquivos e interfaces.

[1] https://freelancing.stackexchange.com
[2] hhtps://stackoverflow.com

Um dos maiores problemas no StackExchange é a existência de perguntas duplicatas, ou seja, quando um usuário faz uma pergunta que já existe no sistema. No conjunto de dados utilizados, existem marcações que indicam essa relação de duplicidade. Logo se uma pergunta P_1 é criada com o objetivo de resolver um problema X, caso já existe uma outra pergunta qualquer P_2 que resolva o problema X, P_1 é chamada de pergunta duplicata de P_2. Essa marcação é feita pela própria comunidade que identifica que ambas as perguntas tem a intenção de resolver o mesmo problema. No conjunto de dados utilizados, existe mais de 33 (trinta e três) mil duplicatas, correspondendo à mais de 10% do total.

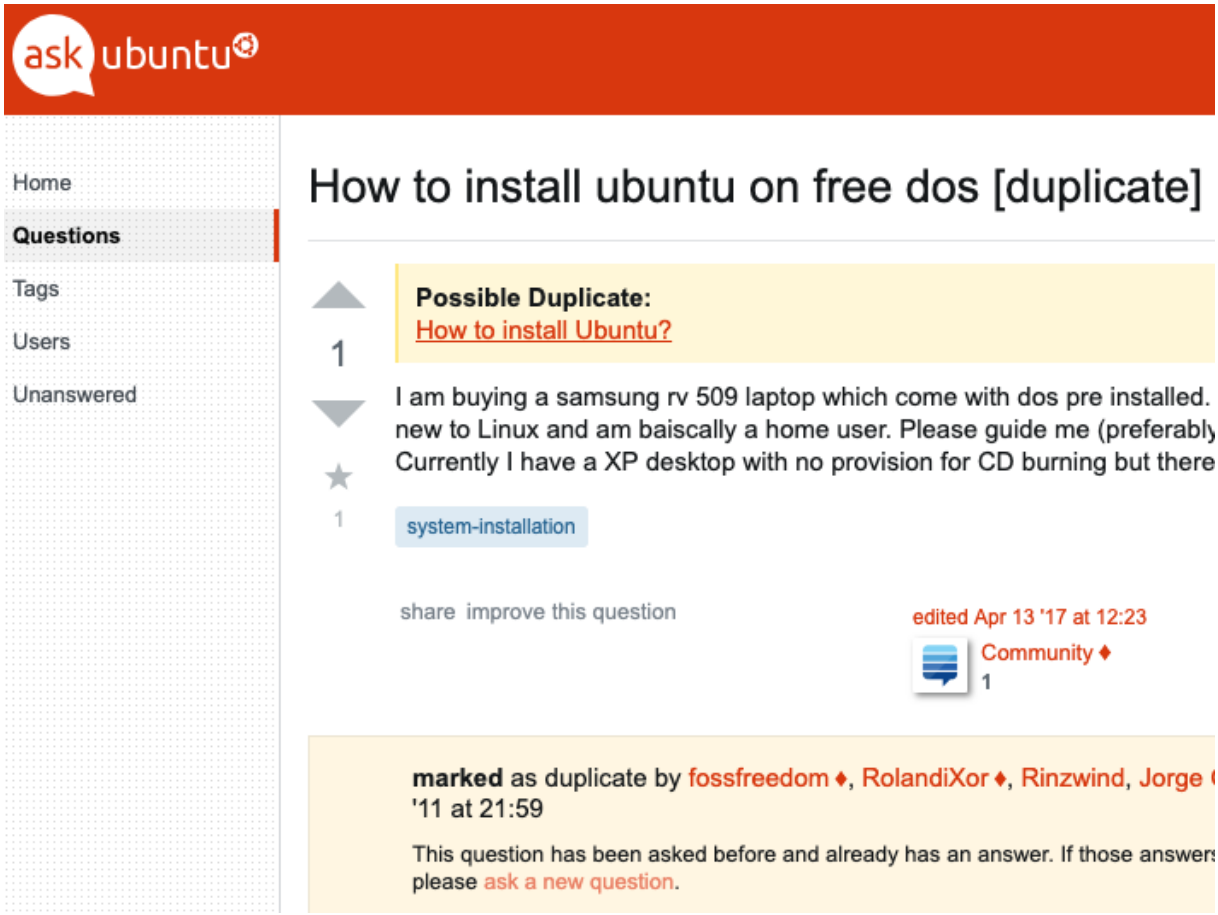

Figura 5.1 A pergunta foi marcada como duplicata de uma já existente.

5.2 Metodologia

Com uma ferramenta de busca eficiente, a intenção é mostrar que com a solução apresentada, os autores das duplicatas conseguem encontrar a pergunta já existente que resolve o seu problema, desfazendo a necessidade de criar uma nova pergunta. Para cada duplicata, usa-se o título da pergunta para encontrar possíveis outras perguntas relevantes e verifica-se se a pergunta que os usuários identificaram como solução para o mesmo problema também foi recuperada e sua posição no ranking da busca.

- **Experimento 1**: Este experimento verificará a importância dos campos textuais para a recuperação da informação. São testadas variações que usam somente o título, somente o corpo e ambos os campos;

- **Experimento 2**: Durante esse experimento será avaliado o melhor valor da variável i descrita na Equação 4.1 para o conjunto de dados utilizado;

- **Experimento 3**: Será verificado à importância desse trabalho contra às abordagens convencionais. As etapas descritas no Capítulo 4 são postas à prova ao se comparar a solução proposta e outras soluções que não usam ou usam parcialmente as etapas descritas.

5.3 Métricas de Avaliação

As métricas escolhidas são as mais comumente utilizadas para avaliar sistemas de recuperação da informação. As métricas Discounted Cumulative Gain (DCG), Mean Average Precision (MAP) e Mean Recurrent Rank (MRR), embora também muito úteis em avaliações de buscadores, não são usadas pois em mais de 95% das perguntas duplicatas, as mesmas são duplicatas em relação à apenas uma questão, o que torna as métricas citadas um resultado visualmente igual ao cálculo da precisão.

Em todos as métricas são avaliados conjuntos de resultados de diferentes tamanhos, ou seja, para uma busca B, para cada métrica M, existirá um grupo de tópicos de tamanho K que corresponde aos resultados encontrados pela busca B. Logo $M@K$ corresponde à métrica M quando o resultado da busca contém apenas os K melhores resultados escolhidos pela solução utilizada.

- **Precisão**: Corresponde à porcentagem de documentos encontrados que são relevantes para a busca. Então, para uma busca B, se metade dos documentos encontrados forem relevantes, sua precisão é de 50% ou 0.5. Seu cálculo pode ser visualizado na Equação 5.1.

$$P = \frac{DocumentosRelevantesRecuperados}{NmerodeDocumentosRecuperados} \tag{5.1}$$

- **cobertura**: É útil para se saber quantos dos documentos relevantes foram encontrados. Assim, existe um grupo G de documentos relevantes para cada busca B. A cobertura é o número de documentos pertencentes à G que foram encontrados sobre o número total de documentos relevantes, que é $|G|$. Sua fórmula é descrita na Equação 5.2.

$$R = \frac{DocumentosRelevantesRecuperados}{DocumentosRelevantes} \tag{5.2}$$

- **F-Measure**: É uma métrica derivada da cobertura e da precisão, seu cálculo é através

da média harmônica de ambos. Sua forma pode ser vista na Equação 5.3.

$$F_1 = \frac{2PR}{P+R} \tag{5.3}$$

5.4 Resultados

Durante essa seção são apresentados encontrados a partir dos experimentos descritos anteriormente.

5.4.1 Comparação do uso dos campos

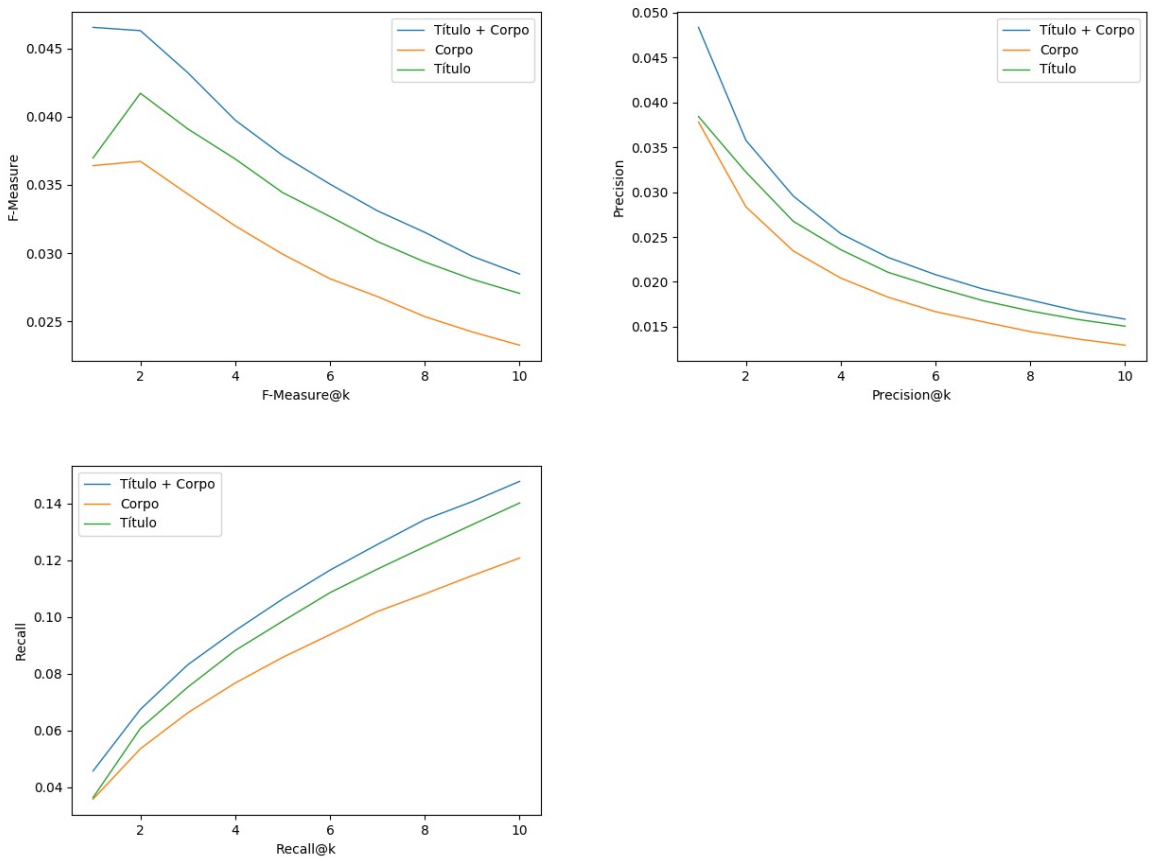

Figura 5.2 Comparação entre variantes da solução que usam ou não certos campos.

O objetivo desse experimento é mensurar o quanto os campos título e corpo são relevantes para a busca, buscando comparar o quanto o uso individual de cada um desses campos influencia nos resultados. A Figura 5.2 ilustra os resultados obtidos nos experimentos. Fica claro que a opção à ser utilizada é a com uso dos dois campos. Entretanto, percebe-se que o título tem uma capacidade maior de encontrar similaridades em outros documentos, no gráfico que ilustra o F-Measure, sua curva acentuada quando $K = 2$ indica que os resultados, quando encontrados,

ocupam mais as primeiras posições do ranking, o que é uma boa indicação. Conclui-se que, embora o corpo possua mais texto e uma melhor contextualização do tópico, o título indica o domínio do tópico de uma forma mais concisa, sendo mais informativo do que corpo.

5.4.2 Variações de Influência da Similaridade

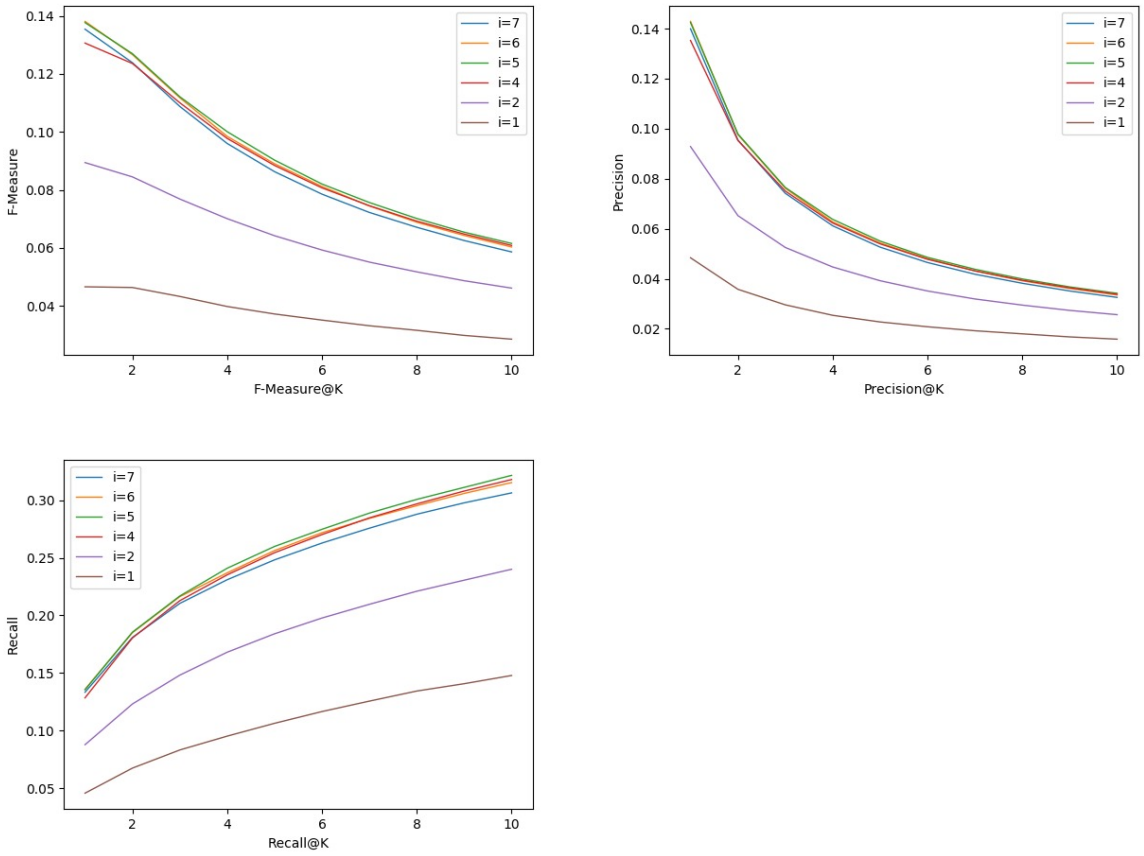

Figura 5.3 Comparação de diferentes usos da variável i da Equação 4.1.

Neste experimento, será encontrado o valor de i descrito na Equação 4.1 que indica o quanto uma busca usa a similaridade entre a busca e os campos título e corpo para decidir os textos relevantes e o quanto da pontuação é usado para tomar essa decisão. Para maiores valores de i, a similaridade terá maior influência e para menores, a pontuação terá mais influência na equação. Nos resultados encontrados, demonstrados na Figura 5.3, percebe-se que a diferença que esse aspecto traz para a busca é de grande magnitude, podendo melhorar os resultados em até três vezes, tornando essa decisão muito mais importante do que a escolha de uso de determinados campos ou não.

Nos gráficos apresentados, os valores de i que apresentam os melhores resultados estão conglomerados na faixa $4 <= i <= 7$, a partir de $i = 4$, o ganho obtido passa a ser muito pouco e de $i = 6$ em diante, os resultados começam a ter leves prejuízos, sendo $i = 7$ a opção com os

piores resultados dentro dessa faixa. Com isto em vista, para este conjunto de dados foi escolhido $i = 5$ para maximizar os acertos.

5.4.3 Comparações com outros métodos

O seguinte experimento avalia o potencial de usar essa solução em sistemas reais. O presente trabalho é comparado com outros métodos comumente utilizados pelo mercado para resolver os problemas de busca. Como pode se perceber nas Figuras 5.4 e 5.5, a solução proposta supera todas as outras avaliadas com muita vantagem, obtendo resultados de duas e até três vezes melhores em alguns casos. É importante notar, que em sua versão sem a configuração de parâmetros, com $i = 1$, ela já supera os outros métodos, embora seja uma posição confortável, a decisão pela configuração desses parâmetros faz com que os resultados se tornem muito melhores.

Enquanto isso, os modelos concorrentes se beneficiam do uso do pré-processamento descrito no Capítulo 4, mas não o suficiente para se tornarem uma opção considerável. Os mesmos foram apresentados no Capítulo 3 e fazem parte de diferentes modos de lidar com buscas. Em especial, o modelo booleano, que apenas checa as palavras em comum, teve os piores resultados, isso atesta o fato de que no problema em questão é muito mais importante saber ponderar os valores das palavras no texto, como é feito no BM25 e no TF/IDF, que são as alternativas mais usadas nos motores de busca.

5.5 Discussão

A partir dos resultados encontrados, conclui-se que o uso de um parâmetro adicional correspondente à pontuação atribuída pela comunidade à um tópico impacta fortemente nos resultados obtidos. Além disso, o pré-processamento, embora não seja o fator mais determinante, tem um papel crucial para alcançar ainda melhores resultados. Vale a pena notar que o processo mais importante a ser tomado, depois da adição de um campo de pontuação, é a configuração dos parâmetros para garantir o melhor funcionamento. Ainda tendo resultados bem mais altos, a implementação não deixou de ser simples, por conta disso a aplicabilidade em fóruns existentes é também simples. Todavia, para maximizar os resultados é necessária a correta configuração dos parâmetros o que vai exigir que cada fórum execute experimentos na própria base para encontrar os parâmetros adequados para sua base, ou seja, é necessário que exista um conjunto de buscas e seus respectivos conjunto de resultados esperados para que novas avaliações e comparações sejam feitas.

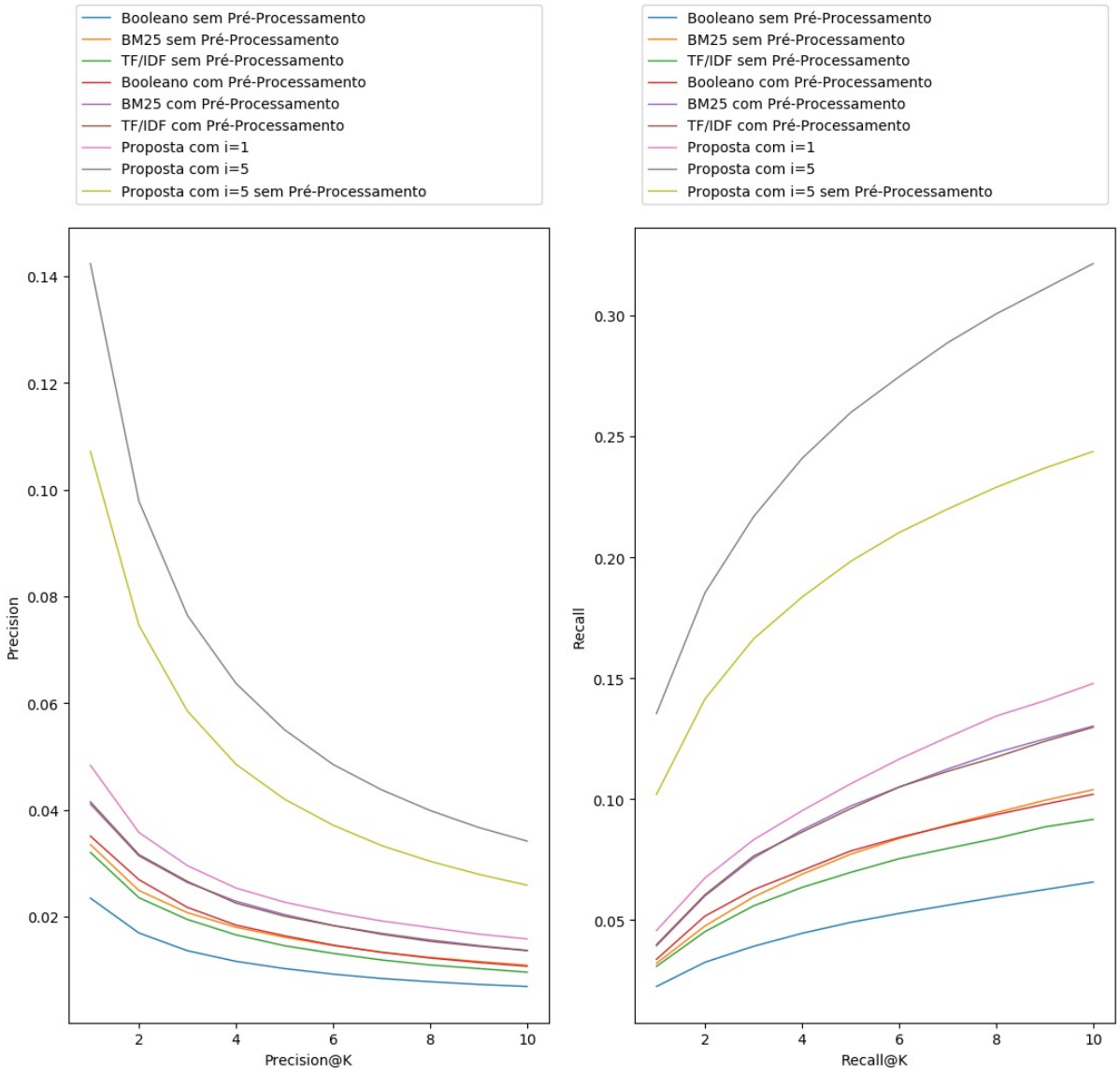

Figura 5.4 Valores de precisão e cobertura ao se comparar com outros métodos de recuperação.

5.6 Sumário

O capítulo de avaliação, por meio de métricas e experimentos, validou a proposta deste trabalho de forma objetiva. O conjunto de dados foi detalhado e sua escolha justificada para as avaliações subsequentes. A metodologia explicou como a avaliação sera conduzida e as correspondentes métricas foram listadas. Os experimentos iniciais foram executados para se obter os melhores parâmetros da solução proposta e, ao fim, comparou-se o resultado com outros métodos propostos para resolução deste problema.

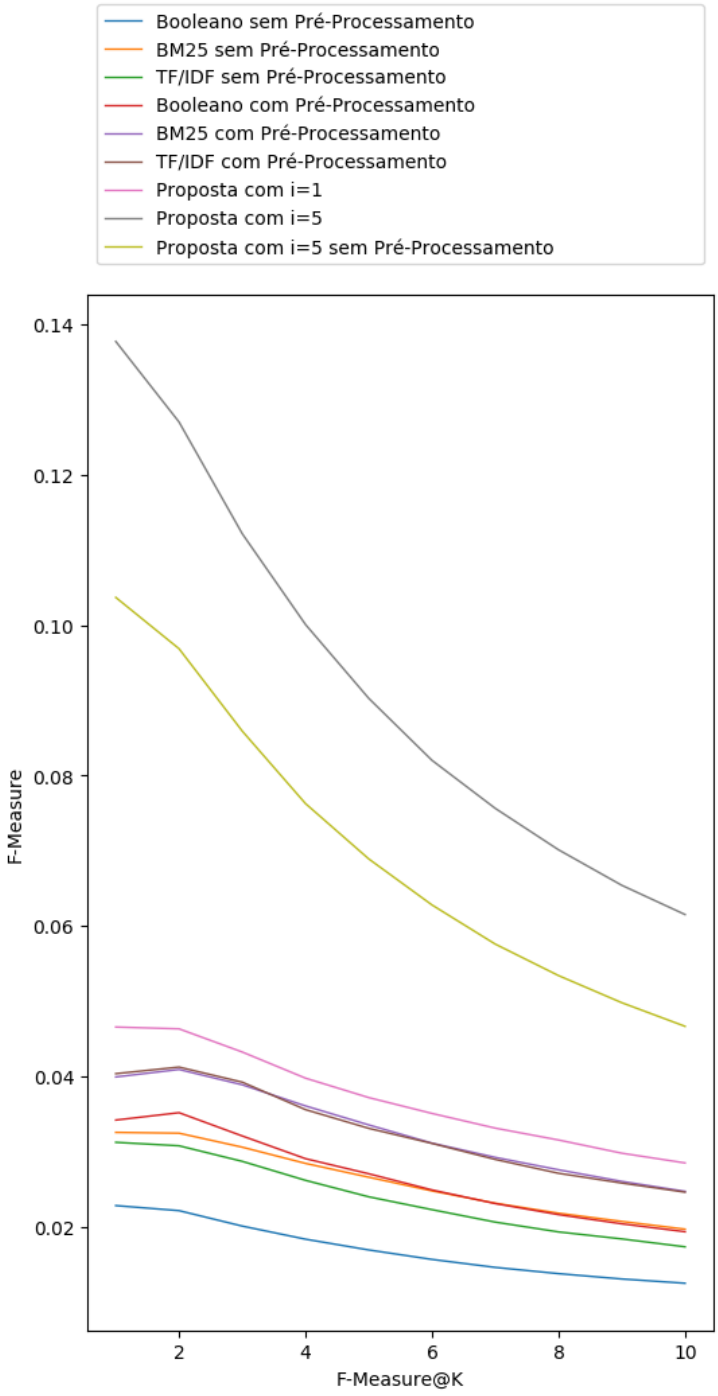

Figura 5.5 Valores de F-measure ao se comparar com outros métodos de recuperação.

6

Conclusão

O presente trabalho apresentou um sistema de recuperação da informação cujo objetivo maior é beneficiar fóruns através do uso da pontuação presente nos tópicos e por meio de um uso inteligente de técnicas de pré-processamento. Na introdução, foi abordada a relevância do problema que esse projeto trata e como isso pode prejudicar os fóruns se não for tratado. Logo, foi indicado um conjunto de objetivos à serem cumpridos nesse trabalho para obter a solução.

O Capítulo 2 contextualizou o conceito de fóruns, trazendo opiniões de especialistas e a trajetória deles no mercado e na sociedade. Ao abordar sistemas de recuperação da informação no Capítulo 3 foi explicado os conceitos necessários para a compreensão do domínio do problema e formas de verificar a qualidade dos métodos apresentados.

Ao chegar na proposta, apresentada no Capítulo 4, a solução é apresentada desde os seus requisitos até sua implementação com explicações sobre as técnicas utilizadas para armazenar, estruturar e recuperar os tópicos relevantes e garantir a resolução do desafio apresentado.

Na avaliação, a metodologia usada para conduzir os experimentos foi apresentada e a forma de comparar resultados foi apresentada nas métricas, a fim de que o trabalho possa ser avaliado objetivamente e que seja verificada a potencialidade de se ter uma solução para o problema, conforme foi demonstrado no Capítulo 5.

6.1 Contribuições

Após serem analisados os aspectos que esse trabalho abordou, suas principais contribuições são listadas à seguir.

- **Revisão de métodos de Recuperação da Informação**: Foi realizada uma revisão nos trabalhos relativos à sistemas de recuperação da informação e, por meio da opinião de especialistas na área, como estes podem contribuir para uma melhor experiência nos fóruns. O estudo também explicou como outros métodos podem ser mais apropriados para outros casos de uso para melhores resultados na recuperação da informação;

- **Modelo para a recuperação de tópicos em Fóruns**: O objetivo principal do trabalho foi propor um modelo para a recuperação de tópicos em fóruns de discussão na web para obter resultados que possam minimizar o número de perguntas duplicatas e facilitar o acesso do usuário à informação;

- **Avaliação Experimental**: O trabalho trouxe comparações objetivas com os métodos mais famosos, trazendo assim, uma avaliação crítica de como o uso da proposta seria útil para fóruns. Os experimentos mostraram como cada uma das etapas descritas no Capítulo 4 podem contribuir para melhores resultados.

6.2 Trabalhos Futuros

Este projeto ainda pode ser melhorado de algumas maneiras, algumas limitações de recursos e escopo não permitiram que fossem feitas no presente trabalho. Dentre as melhorias possíveis, pode-se destacar as mais importantes:

- **Avaliações com diferentes conjuntos de dados**: Outros conjuntos de dados também estão disponíveis para teste na Web. Seria interessante a avaliação sobre o comportamento dos algoritmos quando aplicados sobre bases de diferentes tamanhos e diferentes domínios, além de ilustrar as possíveis adaptações para os critérios de pontuação, por exemplo;

- **Avaliações com os próprios fóruns**: A comparação com os resultados dos próprios fóruns feita por humanos traria resultados ainda mais interessantes, com uma visão mais subjetiva pode-se avaliar não somente as perguntas duplicadas, mas se a busca realmente traz conteúdo relevante de modo geral;

- **Uso de técnicas mais avançadas de NLP**: NLP é uma área crescente nos últimos anos e muitas ferramentas tem sido propostas, inclusive o próprio Google, maior referência em sistemas de recuperação, disponibiliza um algoritmo chamado Word2Vec [1] que entraria pra a lista de possíveis melhorias, este algoritmo é capaz de transformar as palavras em uma representação vetorial, o que permite melhores comparações entre busca e documentos buscáveis;

- **Adição de novos metadados**: O campo da pontuação foi o único campo não textual utilizado para este trabalho. Embora durante o desenvolvimento tenha-se testado com outros campos, nenhum deles apresentou resultados significativos nos testes iniciais. Entretanto, é possível que hajam adaptações para que estes campos, ou outros campos exclusivos para um determinado domínio, possam ser usados para obter resultados melhores.

[1] https://code.google.com/archive/p/word2vec/

6.3 Sumário

O capítulo final deste trabalho apresentou uma visão geral de tudo que foi discutido e as formas que existem para que ainda seja melhorado. Os objetivos citados no Capítulo 1 foram cumpridos e explicados na subseção de contribuições deste capítulo. Outras técnicas e tarefas que poderiam ser feitas são apresentadas como as potenciais melhorias para este trabalho e/ou trabalhos futuros.

Bibliografia

[1] Jeff Attwood. *The Gamification*. 2011. URL: https://blog.codinghorror.com/the-gamification/ (acedido em 16/03/2019).

[2] Shane Connelly. *Practical BM25 - Part 2: The BM25 Algorithm and its Variables*. 2018. URL: https://www.elastic.co/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables (acedido em 17/03/2019).

[3] Bruce Croft, Donald Metzler e Trevor Strohman. *Search Engines: Information Retrieval in Practice*. 1st. USA: Addison-Wesley Publishing Company, 2009. ISBN: 0136072240, 9780136072249.

[4] Mark Davis. *Unicode Text Segmentation*. 2019. URL: http://unicode.org/reports/tr29/ (acedido em 17/03/2019).

[5] DOMO. *Data Never Sleeps 5.0*. 2017. URL: https://www.domo.com/learn/data-never-sleeps-5 (acedido em 15/03/2019).

[6] DOMO. *Data Never Sleeps 6.0*. 2018. URL: https://www.domo.com/learn/data-never-sleeps-6 (acedido em 15/03/2019).

[7] Frederico A. Durão. "Applying a Semantic Layer in a Source Code Retrieval Tool". Tese de mestrado. Universidade Federal de Pernambuco, 2008.

[8] David A. Grossman e Ophir Frieder. *Information Retrieval*. 2nd. The Information Retrieval Series. Springer Netherlands, 2004.

[9] IBM. *10 Key Marketing Trends for 2017*. 2017. URL: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=WRL12345USEN (acedido em 15/03/2019).

[10] Joice. *Web 1.0, Web 2.0 e Web 3.0... Enfim, o que é isso?* 2013. URL: http://ex2.com.br/blog/web-1-0-web-2-0-e-web-3-0-enfim-o-que-e-isso/ (acedido em 15/03/2019).

[11] V. M. Kenski. "As novas tecnlogias de comunicação e informação e as mudanças necessárias nas instituições educacionais". Em: *Educação e Linguagem* 3 (2000).

[12] Christopher D. Manning, Prabhakar Raghavan e Hinrich Schütze. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN: 0521865719, 9780521865715.

[13] Jon Mitchell. *How Google Search Really Works*. 2012. URL: http://readwrite.com/2012/02/29/interview_changing_engines_mid-flight_qa_with_goog (acedido em 16/03/2019).

[14] Bruno Beltrão Moiteiro. “Aplicação de uma Camada de Feedback de Relevância em um Sistema de Recuperação de Informação”. Em: (2018).

[15] Aashish Pahwa. *Quora Revenue Model | How does Quora make money?* 2017. URL: https://www.feedough.com/quora-business-model/ (acedido em 16/03/2019).

[16] VMO Paiva. “Ambientes virtuais de aprendizagem: implicações epistemológicas”. Em: *Educ* 26.3 (2010), pp. 353–70.

[17] Koustubh Pareek. *What is the difference between Quora and Yahoo Answers, etc.?* 2017. URL: https://www.quora.com/What-is-the-difference-between-Quora-and-Yahoo-Answers-etc (acedido em 15/03/2019).

[18] Martin F. Porter. “An algorithm for suffix stripping”. Em: *Program* 14.3 (1980), pp. 130–137.

[19] Gary Price. *Web Search History: Before Google Answers and Yahoo Answers There Was Answer Point From Ask Jeeves*. 2005. URL: https://searchenginewatch.com/sew/news/2060212/web-search-history-before-google-answers-yahoo-answers-there-was-answer-point-from-ask-jeeves (acedido em 20/03/2019).

[20] Stephen Robertson. “The Probability Ranking Principle in IR”. Em: *Journal of Documentation* 33 (dez. de 1977), pp. 294–304. DOI: 10.1108/eb026647.

[21] Ian Ruthven e Mounia Lalmas. “A Survey on the Use of Relevance Feedback for Information Access Systems”. Em: *Knowl. Eng. Rev.* 18.2 (jun. de 2003), pp. 95–145. ISSN: 0269-8889. DOI: 10.1017/S0269888903000638. URL: http://dx.doi.org/10.1017/S0269888903000638.

[22] D. Johnson S. D. Harlow. “An epistemology of technology”. Em: *Educational Technology Review* 9 (1998), pp. 15–20.

[23] I. Sommervile. *Software Engineering*. 9nd. Harlow, England: Addison-Wesley, 2010.

[24] Internet Live Stats. *Google Search Statistics*. 2019. URL: internetlivestats.com/google-search-statistics (acedido em 15/03/2019).

[25] Cambridge University. *Stemming and lemmatization*. 2009. URL: https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html (acedido em 17/03/2019).

[26] Wikipedia. *Google Answers*. URL: https://pt.wikipedia.org/wiki/Google_Answers (acedido em 22/03/2019).

[27] Wikipedia. *Yahoo Respostas*. 2018. URL: https://pt.wikipedia.org/wiki/Yahoo!_Respostas (acedido em 22/03/2019).