Diego Novaes Batista

MeusDireitosConsumidor: Uma Plataforma Web para Recuperação de Artigos Legais do Código de Defesa do Consumidor a partir de Queixas-texto

Salvador, BA

2018

Diego Novaes Batista

MeusDireitosConsumidor: Uma Plataforma Web para Recuperação de Artigos Legais do Código de Defesa do Consumidor a partir de Queixas-texto

Monografia apresentada ao curso de Ciência da Computação, Departamento de Ciência da Computação, Instituto de Matemática e Estatística, Universidade Federal da Bahia, como requisito parcial para obtenção do grau de Bacharel em Ciência da Computação.

Universidade Federal da Bahia

Instituto de Matemática e Estatística

Departamento de Ciência da Computação

Orientador: Vinicius Tavares Petrucci

Salvador, BA

2018

Agradecimentos

- À Universidade Federal da Bahia, como instituição, pela oportunidade e financiamento aos meus estudos durante a graduação.
- À minha família pela educação, amor e suporte financeiro.
- À Empresa Júnior de Informática da UFBA, pelos desafios e primeiras experiências com programação e web.
- Aos professores do departamento de Ciência da Computação pela competência e apoio.
- Ao meu orientador, Vinicius Petrucci pela motivação ao tema.

"The greatest obstacle to discovery is not ignorance
- it is the illusion of knowledge."
(Daniel J. Boorstin)

Resumo

Há um grande volume de informação do campo jurídico que encontra-se inacessível por grande parte da população. Técnicas de Recuperação de Informação (RI) e Processamento de Linguagem Natural (PLN) surgem como uma abordagem poderosa na busca por respostas a partir de textos não estruturados. A partir de queixas escritas pelo usuário e utilizando essas técnicas, este trabalho apresenta uma plataforma para busca de informação legal, apoiando-se no Código de Defesa do Consumidor (CDC). O propósito deste sistema é contribuir com a democratização da informação jurídica de maneira mais simples e intuitiva a todos. A partir de uma queixa do usuário, a plataforma presta assistência informando artigos legais relevantes a sua queixa, aplicando perguntas diretas (chatbot) e técnicas de similaridade de textos. Com acesso a um navegador web, o usuário digita sua queixa e a aplicação recupera artigos relacionados à queixa. Os artigos encontrados permitem esclarecer sobre seus direitos acerca do seu problema e endossar o cidadão na solução da sua queixa. A plataforma é um passo importante na construção de um assistente jurídico.

Palavras-chave: chatbot, PLN, processamento de linguagem natural, CDC, código de defesa do consumidor, similaridade de textos, recuperação de informação, RI.

Abstract

There is a big data from the legal field that is inaccessible by the population. Natural language processing (NLP) and information retrieval (IR) techniques appear as a clear approach in the search for information on unstructured data. Based on these techniques, this work has a platform that retrieves legal information contained in the Consumer Defense Code (CDC) using complaints wrote by the user. The platform serves as an assistant application for solving general and specific problems in the area of Consumer Law. Through a user complaint, the platform provides assistance informing a legal article relevant to your complaint. With the access of a web browser, the user types their complaint and the application, which uses factoid questions and similarity of text, has the function of retrieving relevant articles. In that case, the article found could endorse the citizen in his solution or inform legal data about his rights over his problem.

Keywords: chatbot, NLP, natural language processing, consumer protection code, text similarity, information retrieval, IR.

Lista de ilustrações

Lista de abreviaturas e siglas

PLN	Processamento de Linguagem Natural
RI	Recuperação da Informação
JS	Javascript
ES	Motor de Busca Elastic Search
CDC	Código de Defesa do Consumidor
SINDEC	Sistema Nacional de Informações de Defesa do Consumidor
PROCON	Programa de Proteção e Defesa ao Consumidor
SAC	Serviço de Atendimento ao Consumidor

Sumário

1 Introdução

De acordo com as pesquisas realizadas pelo Sistema Nacional de Informações de Defesa do Consumidor (SINDEC) entre o período de 2016 a 2018, cerca de 2 milhões de reclamações do cidadão consumidor são atendidas pelo PROCON anualmente. Dentre essas reclamações as que mais destacam-se são sobre serviços de telefonia celular e fixa, cartão de crédito, bancos, energia, água e TV por assinatura. Entre 2016 e 2018, verifica-se que um grupo de reclamações repete-se em diversos estados do brasil e o número de atendimentos pelo PROCON tem diminuído com passar dos anos (SINDEC, 2017).

Apesar do número expressivo de atendimentos realizados pelo PROCON, existem ainda muitas queixas sendo registradas pelos cidadãos em sites como *ReclameAqui*[1] ou pelo próprio site do governo federal, *Consumidor.gov.br*[2]. Em 2016, de acordo com o portal de notícias do governo federal mais de 2.7 milhões de queixas foram registradas no *Consumidor.gov.br* (BRASIL, 2017). Esse valor representa um número maior de reclamações de cidadãos em relação ao publicado pelo boletim do SINDEC de 2016. No *ReclameAqui* também pode-se identificar outro número expressivo de queixas cadastradas pelos usuários do site. As três diferentes fontes indicam que o cidadão ainda tem dificuldade em realizar negociações com o consumidor de forma justa e igualitária.

No site do *ReclameAqui*, há diversas queixas mal formuladas onde os usuários mostram seu descontentamento sem ao menos saber se seu caso encontra-se suportado pela lei. Desses usuários, diversos mostram-se indignados, dizem ter sido lesados durante o uso de um serviço e não sabem o que fazer para se defenderem. Muitas das empresas no período de Fevereiro/2018 da plataforma do *ReclameAqui* conforme mostrado na Figura 1, são de telefonia e não respondem às queixas dos cidadãos.

Alguns tipos de reclamações do cidadão consumidor tendem a repetir-se anualmente, principalmente em épocas do ano que ocorrem promoções ou em feriados nacionais. De acordo com o *ReclameAqui*, as queixas repetem-se no período de ofertas de varejo como o Black Friday. Em 2017, foi registrado um aumento de mais de mil queixas referente ao Black Friday, registradas no site do *ReclameAqui* (ESTADÃO, 2017). A maioria das reclamações são sobre propaganda enganosa e divergência de valor da oferta e o produto adquirido. É comum identificar ofertas falsas, vendas casadas, empresas "fantasma", cadastros anônimos de compras não autorizadas (estelionato) ou ainda contratos abusivos de consórcios.

Devido aos dados apresentados percebe-se uma dificuldade dos consumidores em resolver conflitos gerados pela relação entre o consumidor e fornecedor. O usuário

[1] https://www.reclameaqui.com.br/
[2] https://www.consumidor.gov.br/

Figura 1 – Ranking de reclamações registradas no dia 06/02/2018 no ReclameAqui

desinformado fica vulnerável às falcatruas e abusos cometidos pelos fornecedores na venda de produtos e na prestação de serviços. É mais difícil ainda para o usuários lidar com tais abusos sem saber dos seus direitos de consumidor. Obter informação sobre os textos legais por si só apresenta uma dificuldade de acesso a toda população, seja pelo vocabulário típico da área do direito do consumidor ou pelo grande volume de informação no domínio jurídico. Por isso há uma necessidade de uma ferramenta que torne mais acessível a informação contextualizada para o usuário comum, isto é, uma informação mais explícita ligada ao problema que o usuário está enfrentando.

1.1 Definição do Problema

De acordo com uma pesquisa realizada pelo Instituto de Defesa do Consumidor (IDEC), menos da metade dos consumidores insatisfeitos com serviços ou compra de produtos buscaram um órgão de defesa do consumidor para reivindicar seus direitos. Os consumidores que reivindicaram já haviam consultado o CDC pelo menos uma vez. Essa estatística expressa um grande percentual de desconhecedores da lei (IDEC, 2016). Isto é, deveriam haver mais cidadãos que tenham tido contato com a lei e mais cidadãos

reivindicando pelos seus direitos. A participação do consumidor no combate à falta de ética do empresariado é essencial para construir um melhor relacionamento entre consumidor e fornecedor, tornando o mercado mais justo e democrático.

Os empecilhos presenciados pelos cidadãos no acesso aos seus direitos na lei são diversos: falta de escolaridade, alto custo de consulta a um advogado ou alta espera por atendimento nas delegacias do consumidor. Com o intuito de combater a desinformação, a plataforma online *MeusDireitoConsumidor.com.br* apresentada por estudo tem por fim facilitar o acesso a informação legal para a população. Para atingir esse objetivo, o estudo traz uma abordagem utilizando Recuperação da Informação, Processamento de Linguagem Natural e um conjunto de perguntas factoides para encontrar informação legal relevante a queixa digitada pelo usuário. Problemas como acesso a informação não estruturada, busca e filtro de informação relativa à queixa do usuário são assuntos discutidos na abordagem apresentada por estudo.

Este trabalho propõe uma plataforma para auxiliar o cidadão no entendimento de seus direitos como consumidor. O auxílio que a plataforma oferece está na identificação do problema do usuário com base no Código do Defesa do Consumidor (CDC). Acredita-se que o usuário bem informado poderá melhor combater injustiças que ocorrem no seu dia-dia. A aplicação tem como objetivo contribuir com a democratização e a propagação da informação jurídica de maneira mais simples e rápida a todos.

1.2 Contribuições

Com intuito de resolver o problema de acesso a informação jurídica, a plataforma irá apontar o artigo mais relevante a sua queixa sem que o usuário tenha que procurar o artigo nos textos da lei. Será usado uma motor de busca para pesquisar artigos legais de forma rápida. A tarefa de seleção do artigo mais relevante dentro dos prováveis artigos encontrados ocorre através de um chatbot que realiza perguntas relacionadas ao artigo para tentar identificar o resultado correto. Caso haja interesse, o usuário pode consultar queixas semelhantes e os artigos contidos nos resultados encontrados por outros usuários.

O *MeusDireitoConsumidor.com.br* dá passos adiante no combate a desinformação, burocracia e na qualidade de vida do consumidor brasileiro. Através desse estudo foi possível implementar um sistema capaz de:

- Extrair artigos legais baseados nas queixas digitadas pelo cidadão consumidor respondendo suas dúvidas ou defendendo sua posição no problema descrito baseado na lei.

- Apresentar queixas de outros usuários que tiveram seus problemas endossados pelos

artigos informados pelo sistema

- Classificar queixas com base no título da seção do artigo encontrado pelo sistema. Por exemplo, o sistema ao encontrar o artigo 18 como resultado relacionado à queixa, julga a queixa digitada como problema relevante ao tópico que trata sobre 'Responsabilidade por Vício do Produto e do Serviço"

- Alimentar o banco de dados do sistema com queixas validadas pelo usuário. No futuro, esses dados podem ser usados para recuperação automática de artigos baseado nas queixas registradas no sistema. Essa tipo de recuperação pode ser feita utilizando abordagem de detecções de padrões textuais

- Propagar informação legal na área do direito do consumidor usando o Código de Defesa do Consumidor (Lei 8078) para enfrentar a desinformação legal e fortalecer a cultura do respeito aos direitos do consumidor.

Assim como disposto no Art. 43 do Código de Defesa do Consumidor: "O consumidor, sem prejuízo do disposto no art. 86, terá acesso às informações existentes em cadastros, fichas, registros e dados pessoais e de consumo arquivados sobre ele, bem como sobre as suas respectivas fontes.". Assim, todo registro feito pelo usuário pode ser consultado através do sistema.

1.3 Sumário

Este capítulo introduziu contexto motivacional e o problema envolvido com a criação da plataforma online *MeusDireitoConsumidor.com.br*. No Capítulo 2 serão apresentados conceitos, tecnologias web e outros referenciais teóricos utilizados na criação da plataforma. No Capítulo 3, serão apresentados trabalhos relacionados com a aplicação deste estudo. Serão discutidos seus propósitos e similaridades com o sistema deste estudo. O Capítulo 4 apresenta a plataforma MeusDireitoConsumidor.com.br, suas funcionalidades e interface. Em seguida, o Capítulo 5 mostra o experimento realizado para validar a aplicação e analisa os resultados encontrados. Por fim o Capítulo 6 dispõe as conclusões e considerações a respeito deste trabalho.

2 Fundamentação Teórica

Com a vinda da Web 2.0 [1], qualquer usuário com acesso à internet teria o poder de criação de conteúdo. Seja em redes sociais como Orkut, Facebook, Youtube, Twitter, fóruns ou sites oficias do governo, o usuário através do navegador web poderia passar feedback dos serviços públicos prestados, publicar suas opiniões políticas, experiências no dia dia ou ainda torna-se um figura pública formadora de opinião (BIRD; KLEIN; LOPER, 2009).

A qualidade de vida do ser humano melhorou. Os usuários estão mais conectados à internet por fazer parte da rotina dos seus afazeres e do trabalho. As aplicações que utilizam internet estão em constante evolução. Através dessa grande rede, os usuários podem optar por realizar ligações com a internet, solicitar reuniões de trabalho por vídeo e voz e deslocar geograficamente de um lugar a outro sem ao menos saber seu percurso. Basta apenas ter conexão com internet, um aplicativo no celular e todas essas funcionalidades estarão disponíveis para o usuário.

Devido a tais mudanças, o usuário da internet tornou-se mais exigente. Já não basta mais o usuário digitar textos e o buscador retornar resultados dito relevantes. O usuário, na maioria das vezes, diz não saber do que se trata os textos retornados pelos motores de busca ou tem dificuldade em procurar por todas as páginas do resultado informado pelo motor de busca. O mesmo se aplica a lei 8078, conhecida como Código do Direito do Consumidor (CDC). O usuário comum, estressado e desapontado por ser vítima de uma negociação injusta, busca saber seus direitos como cidadão e acabam encontrando textos com vocabulário ilegível para ele.

A abordagem sugerida pela Recuperação da Informação (RI) vem suprir a necessidade que o usuário requisita. Este estudo mostra que é possível recuperar os artigos legais do Código de Defesa do Consumidor com base nas queixas texto apresentadas pelo cidadão. A seguir, serão apresentadas as tecnologias e conceitos utilizados por este estudo que tornaram possível a implantação da plataforma *MeusDireitoConsumidor.com.br*.

2.1 Recuperação da Informação

O processo de Recuperação de Informação (RI) consiste em determinar quais das palavras informadas pelo texto do usuário estão presentes em uma coleção de texto de um repositório. Em termos técnicos, o processo conta com uma função de busca, que

[1] Termo refere-se a websites da internet que tem ênfase em conteúdos criados por usuários, usabilidade e interoperabilidade com outros sistemas, produtos e dispositivos (O'REILLY, 2005).

compara as representações dos documentos com a expressão de busca gerada pelos termos dos textos do usuário. O processo, ao finalizar a busca, recupera os documentos que supostamente fornecem a informação que o usuário procura (FERNEDA, 2003). Esse processo é geralmente a uma grande quantidade de informação não estruturada.

É possível utilizar-se de outras técnicas em conjunto da recuperação da informação para extrair informações ligadas diretamente ao texto digitado pelo usuário. Por exemplo, a abordagem de RI pode ser utilizada em conjunto de um indexador e um ordenador de resultados. O indexador cria uma estrutura de consulta em um servidor e o ordenador ordena os resultados baseado em um valor (chamado de score ranking) associado a similaridade de texto da busca.

Uma característica importante de sistemas que utilizam a recuperação da informação é que os usuários desses sistemas buscam informações que não seguem uma representação lógica ou bem estruturada como a de um sistema gerenciador de banco de dados (SGBD). As aplicações que processam informações não estruturadas e aplicam o processo de recuperação de informação são chamados de Sistemas de Recuperação da Informação.

2.1.1 Sistemas de Recuperação da Informação

Com propagação do acesso à Web, os usuários podem utilizar de buscadores para procurar informações do seu interesse. A busca por evolução e por tecnologia, fizeram com que cientistas e tecnólogos buscassem construir sistemas que obtivessem informação especializada para usuário. O advindo da abordagem da Recuperação da Informação (RI) e a evolução dos motores de busca permitiram o surgimento e propagação dos Sistemas de Recuperação da Informação (SRI) que entregam informação filtrada e diretamente relacionada a busca do usuário. Esses sistemas desempenham tarefas como: aquisição e armazenamento de documentos, organização e controle de dados e ainda disseminação de informação (BAEZA-YATES, 2012).

Os Sistemas de Recuperação da Informação (SRI) podem utilizar de diversas técnicas para obter informação de um conjunto de dados. Algumas dessas técnicas envolvem: indexação de documentos, ordenação textual, modelagem de dados, processamento de linguagem natural e inteligência artificial. Por exemplo, através do uso de técnicas de Processamento de Linguagem Natural, identificação de expressões comuns da linguagem e a medida estatística TF-IDF (Seção 2.4.3), os sistemas de recuperação de informação são capazes de encontrar significados relevantes dentro dos documentos de um repositório.

Os Sistemas de Recuperação da Informação (SRI), para executar seu processo, contam com uma interface para entrada de dados do usuário, um servidor da aplicação responsável pelo controle e armazenamento de informação entre usuário e o sistema, um

motor de busca que entrega resultados suficientemente rápidas aos usuários e uma coleção de documentos devidamente indexada. A escolha de um correto modelo de informação permite ordenar resultados de acordo a ocorrências dos termos encontrados (MANNING, 2009). A Figura 2 apresenta uma visão esquemática dos SRI de acordo com (HIEMSTRA, 2009).

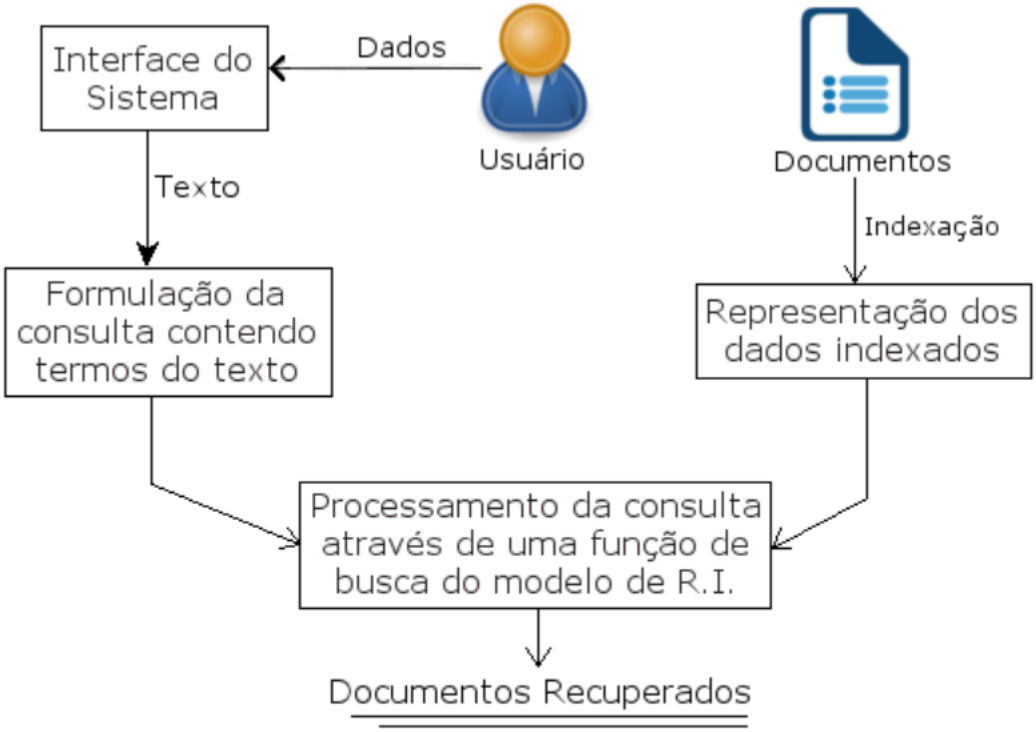

Figura 2 – Visão conceitual de sistemas de recuperação da informação

2.1.2 Modelos de Recuperação da Informação

Sabe-se que os Sistemas de Recuperação da Informação (SRI) são capazes de encontrar informação especifica de interesse do usuário. Para atingir essa tarefa, é necessário que o desenvolvedor tenha conhecimento da estrutura e do tipo de informação que será indexada. Pois sabendo do que se trata a coleção de dados a ser indexada, será possível escolher o melhor modelo de recuperação da informação que suporte os SRI a encontrar informações com maior eficiência.

O modelo de recuperação de informação ideal para a aplicação irá beneficiar os resultados encontrados pelo motor de busca. Com isso a escolha de um modelo torna essencial para que as consultas do motor não sejam tão genéricas e entreguem resultados mais relevantes. Duas medidas que estão sempre observadas nos SRI para verificar efetividade, são precisão e revocação[2]. Precisão representa a razão entre número de dados relevantes

[2] do inglês, “recall rate”

recuperados e dados totais recuperados. Revocação é a razão entre número de dados relevantes recuperados e dados relevantes totais. Em geral, é dito que existe uma relação inversa entre as duas medidas. Ao melhorar a precisão, pode ser que o sistema traga menos informações recuperadas durante a busca. Logo é necessário analisar os resultados ao tentar abranger maior número de resultados pela busca (MANNING, 2009).

Os modelos de Recuperação de Informação foram criados para obter diversos tipos de informações. Cada um desses modelos abrangem diferentes tipos de informação: informações textuais, logs de arquivos, imagens computadorizadas, elementos de conjuntos lógicos. Consequentemente, existem diferentes modelos de recuperação de informação. Podendo estes serem usados em conjunto, de forma híbrida, ou separadamente.

Nas seções a seguir serão vistos os modelos utilizados para recuperar informação na plataforma apresentada neste estudo. Nos modelos de recuperação de informação a serem apresentados são representados por um conjunto de termos de indexação. Estes termos contém a ideia de um conceito ou significado de um documento (SVENONIUS, 2009).

2.1.3 Modelo Booleano

Modelo baseado na lógica matemática concebida por George Boole e conhecida por Álgebra de Boole. A idéia é utilizar os operadores da lógica booleana em conjunto dos termos indexados para encontrar e formar novos conjuntos dos termos de documentos. Atraves das operações lógicas é possível analisar os termos como símbolos e manipulá-las de forma similar a álgebra aplicada a números (BAEZA-YATES, 1999).

As operações lógicas contidas na Álgebra de Boole é utilizada na função de busca ao procurar por documentos relevantes. Os termos das operações lógicas são os termos do texto a serem buscados nos documentos indexados do repositório. Assim ao aplicar a operação de AND, OR, NOT por exemplo, os sistemas de recuperação de informação buscam conjunto de termos ou documentos que atendam a expressão lógica da função de busca (HIEMSTRA, 2009).

Na Figura 3, é utilizado operador lógico AND na busca por documentos que contém os termos PRODUTO e OFERTA. De forma similar, pode-se buscar por todos os documentos que contenham os termos PRODUTO e OFERTA mas que não tenha o termo CONTRATO com os operadores lógicos OR e NOT (Figura 4) (HIEMSTRA, 2009).

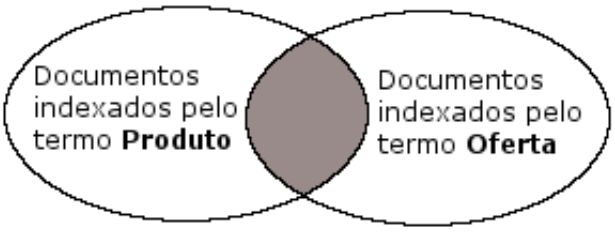

Figura 3 – Resultado da expressão de busca Produto AND Oferta

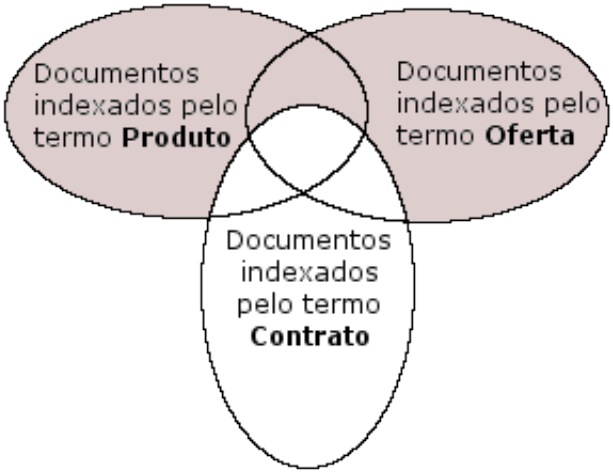

Figura 4 – Resultado da expressão de busca (Produto OR Oferta) OR NOT Contrato

É importante ressaltar que utilizando esse modelo para busca, não é possível ordenar os documentos resultantes. Essa é uma conclusão lógica oriunda deste modelo já que os termos indexados possuem mesma relevância. Isto é, não se pode afirmar que um termo é mais importante que outro neste modelo. Outra forma de chegar a mesma conclusão é observar a pontuação dada a cada resultado deste modelo. A pontuação de cada resultado é realizada através de valores binários. Se o resultado contém os termos da busca então sua pontuação é 1, caso contrário, 0.

Todos os documentos tem a mesma chance de pertencer ou não ao conjunto lógico ditado pela função de busca. Na função de busca, o resultado trata-se de pertencer ou não à representação lógica, não existe meio termo. Se não for aplicado algum processamento nas palavras antes de enviar para função de busca, os termos de alta frequência, conhecidos como *stopwords*, poderiam ter mesma relevância que palavras chave dentro de um texto de acordo com esse modelo (LASHKARI FERESHTEH MAHDAVI, 2009).

Apesar da simplicidade de suas funções de buscas, uma vantagem dessa abordagem é que é possível visualizar com clareza porque um documento foi recuperado dado uma

certa consulta. O modelo lógico serve de base para construções de consultas mais complexas, geralmente utilizado em conjunto com outro modelo de recuperação de informação. Na seção 2.5.1, será apresentado um motor de busca que utiliza modelo de recuperação de informação híbrida que utiliza também o modelo lógico.

2.1.4 Modelo Vetorial

A abordagem seguida neste modelo recupera informações através da extração parcial de informação dos documentos. Isto significa que o documento pode ser representado por parte das informações contidas nele. O modelo também chamado de Modelo Espaço Vetorial (do termo inglês, Vector Space Model), apresenta cada documento pela sua representação de termos indexados. Cada termo recebe um valor normalizado que represente a indexação do documento.

A representação vetorial dos termos de indexação e da consulta podem ser visualizada em um espaço Euclidiano com várias dimensões, onde cada dimensão representa um termo do vetor. Para criar a representação vetorial de um texto são levados em conta somente suas palavras-chave e um peso associado a essas palavras. O peso escolhido para o termo representa sua importância para o documento dentro de alguma estratégia. Seja essa estratégia escolhida pela frequência do termo contido no documento ou por outro motivo (HIEMSTRA, 2009).

Levando em conta o exemplo de queixa abaixo representada por Q1:

"Efetuei a compra de um produto e o mesmo não foi feita a entrega" (Q1)

Seja q, termos da queixa Q1 do usuário com seus devidos pesos. q contém os termos de um texto utilizado para uma busca hipotética neste modelo.

q = [produto, compra, entrega]

q = [0.9, 0.3, 0.5]

Sejam *D1, D2*, vetores que contém representação vetorial dos documentos indexados considerando os termos "produto, compra, entrega" com seus devidos pesos.

D1 = [0.6, 0.2, 0.0]

D2 = [0.0, 0.7, 0.3]

Uma estatística que é bastante utilizada para calculo dos pesos é a da frequência dos termos de um documento baseado na inversão da frequência do termos contido nos documentos da coleção. Essa medida é popularmente conhecida como TF-IDF (Seção 2.4.3). Como os vetores seguem a mesma estrutura, é possível calcular a similaridade dos textos considerando suas representações vetoriais (MANNING, 2009).

Os documentos *D1, D2* apresentados anteriormente possuem três termos de inde-

xação que os representam. Para calcular a similaridade dos documentos D1 e D2 à queixa, utiliza-se a seguinte função de similaridade da Figura 5, também chamada de Similaridade Cosseno (SALTON, 1988).

$$sim(d_j, q) = \frac{\sum_{i=1}^{N}(w_{i,j} \times w_{i,q})}{\sqrt{\sum_{i=1}^{N} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{N} w^2_{i,q}}}$$

onde $w_{i,j}$ é o peso do i-ésimo termo do documento d_j e $w_{i,q}$ é o peso do i-ésimo termo da expressão de busca q.

Figura 5 – Função de Similaridade para recuperar informações relevantes a queixa

A função de similaridade é utilizada na recuperação de informação representando a função de busca. Diferentemente do modelo booleano, os resultados númericos da função de similaridade variam de 0 a 1 sendo possível realizar um ordenamento dos resultados encontrados. Quanto mais próximo o valor estiver de 1, melhor a similaridade entre os dois vetores.

O cálculo de similaridade entre a queixa q e os documentos *D1* e *D2* é feito da seguinte maneira:

$$sim(D1,\ q) = \frac{(0.6 \times 0.9) + (0.2 \times 0.3) + (0.0 \times 0.5)}{\sqrt{(0.6)^2 + (0.2)^2 + (0.0)^2} \times (\sqrt{(0.9)^2 + (0.3)^2 + (0.5)^2})} = 0,885.$$

$$sim(D2,\ q) = \frac{(0.0 \times 0.9) + (0.7 \times 0.3) + (0.3 \times 0.5)}{\sqrt{(0.0)^2 + (0.7)^2 + (0.3)^2} \times (\sqrt{(0.9)^2 + (0.3)^2 + (0.5)^2})} = 0,4408.$$

Conclui-se que caso os documentos fossem ordenados de acordo com sua relevância, o documento *D1* seria mais relevante a queixa q pois o calculo de similaridade entre q e *D2* apresentou um valor menor.

2.2 Indexação

É o processo conhecido por construir índices ou ainda, atribuir termos ou códigos de indexação a um registro ou documento. Índice é uma estrutura contendo geralmente tabelas que apontam termos para o local de pastas, arquivos ou registros. O objetivo desse mapeamento entre termos e registros é agilizar consultas de textos utilizando uma estrutura de dados e seus termos de indexação.

A indexação faz parte de sistemas que procuram facilitar a busca em coleções de texto não estruturados como páginas web, noticias, publicações acadêmicas, bibliografias, relatórios, ofícios, registro histórico e e-mails (MANNING, 2009).

Toda informação contida no repositório é extraída e traduzida em uma linguagem ou estrutura de indexação. Essa linguagem identifica e representa o dado extraído definindo seus pontos de acesso para buscas (FERNEDA, 2003). Esses pontos de acesso podem ser utilizados para restringir o processo de consulta a documentos que contenham pelo menos uma correspondência dos termos da consulta. A estrutura de índice mais conhecida em abordagens de Sistema de Recuperação é chamada de Índice Invertido ou Arquivo Invertido (ZOBEL; MOFFAT, 2006).

2.3 Arquivos Invertidos

Arquivos Invertido ou Índice invertido é o mecanismo utilizado em indexação para realizar buscas de textos relevantes dentro de uma coleção de documentos. A busca de textos utilizando os arquivos invertidos, percorrem uma lista ordenada de termos chamado de vocabulário onde verificam a correspondência e a frequência de cada termo do vocabulário dentro de cada documento da coleção. Por não precisar percorrer todos os termos do vocabulário ou todos os documentos da coleção, a busca apresenta um grande ganho de tempo de processamento para encontrar os resultados.

Pode-se dividir a estrutura dos arquivos invertidos em duas: a primeira chama-se Vocabulário e contém todos os termos de indexação da coleção de documentos, a segunda é uma tabela ou listas ligadas a cada termo de indexação. Nessa estrutura estão contidas o número do documento da ocorrência do termo e a contagem da frequência em que o termo aparece no documento da coleção. Na Figura 6 é apresentada uma estrutura em tabela contendo os termos de uma busca representados por linhas horizontais e os números dos documentos representados por colunas. Cada valor da matriz corresponde a presença dos termos dentro dos documentos (ZOBEL; MOFFAT, 2006).

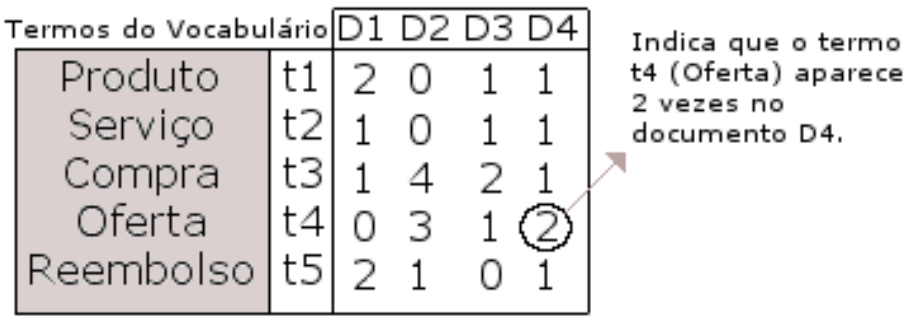

Figura 6 – Exemplo ilustrando a relação entre os termos do vocabulário com os documentos de uma coleção de texto

A busca de texto em arquivos invertidos inicia-se através da busca dos termos presentes no vocabulário. O algoritmo acessa a lista de termos utilizando tabelas *hash* ou

uma estrutura de arvores. Em seguida, recupera-se a informação da ocorrência dos termos nos documentos. Após isto, inicia-se a fase de processamento das ocorrências.

Na fase de processamento da ocorrência, diferentes abordagens podem ser tomadas dependendo do algoritmo utilizado no sistema. É comum a utilização da estatística TF-IDF para dar peso aos termos e recuperar informação dos textos. A descrição da estatística TF-IDF é encontrada na seção 2.4.3 (ZOBEL; MOFFAT, 2006).

2.4 Estatísticas Numéricas

As estatística numéricas discutidas nessa seção estão presentes no motor de busca Lucene e consequentemente, são utilizadas pelo Elastic Search no cálculo de ranqueamento dos resultados de suas buscas.

2.4.1 Frequência de Termos

É um número que indica a frequência de um determinando termo presente em um texto. Essa estatística pode ser utilizada para se obter a frequência de um termo dentro de um texto de uma consulta ou de um documento. A estatística é popularmente conhecida pelo termo em inglês *Term Frequency* (TF) e pode ser encontrada em sistemas de recuperação de informação. A medida é utilizada por sistemas que buscam identificar resultados com maior ocorrência dos termos (MANNING, 2009).

No motor de busca Lucene, a estatística é utilizada para contabilizar a frequência de um termo (q) da consulta nos documentos indexados(d). Seu valor é computado de diferentes maneiras dependendo do objetivo do sistema. A contagem pode ser feita através de valores binários, contagem pura ou pela normalização de seu valor com outras medidas utilizadas pelo sistema. No Lucene, o calculo é feito através da raiz da frequência do termo presente no documento (APACHE, 2010):

$$tf(t,d) = \sqrt{frequênciaDoTermo}$$

2.4.2 Inverso da Frequência de Termos nos Documentos

Diferente da estatística de frequência de termos da seção anterior, o Inverso da Frequência de termos nos Documentos (IDF) leva em consideração a frequência do termo em outros documentos da coleção. Seu valor expressa o quão raro é um termo de uma consulta em relação à um grupo de documentos. A estatística é utilizada em sistemas de recuperação de informação no ranqueamento de resultados e consegue diminuir a importância da frequência de termos comuns dentro de uma coleção de documentos (MANNING, 2009).

A sigla IDF vem do termo em inglês *Inverse Document Frequency* e faz parte do algoritmo de busca dos motores de busca Elastic Search, Lucene e Solr. De acordo com essa medida, a estatística de um termo é calculada através do logaritmo da fração entre o número de documentos de uma coleção e o número de documentos com presença do termo t (APACHE, 2010).

$$idf(t) = 1 + \log(\frac{numeroDocumentos}{frequenciaDocumentos + 1})$$

2.4.3 Pesagem TF-IDF

A estatística TF-IDF, do termo inglês *Term Frequency - Inverse Documento Frequency*, é formada através das medidas citadas nas duas seções anteriores. Essa estatística é utilizada por sistemas que buscam resultados levando em consideração a frequência dos termos da busca e raridade desses termos nos documentos indexados da coleção (MANNING, 2009).

O valor desta medida é utilizado na pesagem dos termos no modelo vetorial (seção 2.1.4) de sistemas de recuperação para expressar a relevância de um termo na coleção de documentos. Essa pontuação indica a raridade e a ocorrência de um termo da busca dentro de uma coleção de documentos. Seu valor é encontrado a partir da multiplicação das medidas *tf* e *idf* discutidas na seção anterior.

$$tfidf = tf(t,\ d) \times idf(t)$$

A partir da estatística TF-IDF, motores de buscas como Lucene e Elastic Search conseguem encontrar resultados ordenados de acordo com as ocorrências dos termos da busca considerando também termos não comuns dos documentos indexados (APACHE, 2010).

2.5 Motores de Busca

Motores de busca são ferramentas robustas utilizadas para encontrar textos não estruturados em documentos indexados. Geralmente, motores de busca estão hospedados em servidores que prestam serviços a aplicações. A tarefa de um motor de busca é encontrar correspondências de texto o mais rápido possível provendo confidencialidade e segurança para as aplicações que o utilizam. Em geral, os motores de busca contam com um pré-processamento dos textos da busca envolvendo processamento de linguagem natural. Os seguintes componentes que fazem parte dos motores de buscas são: uma eficiente indexação de coleção de documentos, um modelo de recuperação de informação bem definido e uma configuração de software/hardware para escalar seu serviço com buscas concorrentes e com a alta quantidade de acessos dos usuários (CROFT DONALD METZLER, 2015).

Pode-se dizer que motores de busca possuem uma certa similaridade aos sistemas de gerenciamento de banco de dados. Seus textos são salvos em um repositório e o índice para busca destes documentos são preservados. Porém sabe-se que há diferenças. Enquanto sistemas de banco de dados possuem estruturas lógicas completas de consultas, os motores de busca utilizam em sua grande maioria consultas contendo lista de termos e frases.

No sistema de banco de dados, a ocorrência dos termos de uma busca é um dado que atende a um condição lógica. No motor de busca, uma correspondência de palavras é um documento que está de acordo com uma heurística de busca seguindo um modelo de Recuperação de Informação (RI). Os modelos de RI podem ser utilizados separadamente ou de forma conjunta, chamados de modelos híbridos (ZOBEL; MOFFAT, 2006).

2.5.1 Elastic Search

Elastic Search (ES) é um motor de busca robusto criada em cima do Apache Lucene para armazenamento de dados, indexação de documentos, estruturação de informações e principalmente para realizar buscas eficientes em grandes coleções de texto. Graças a sua estrutura básica de busca, os índices Lucene, o motor de busca é capaz de encontrar ocorrências de palavras chaves diretamente conectadas aos documentos em que estão presentes. Em comparação com os modelos tradicionais de dados relacionais[3], o Elastic Search é capaz de recuperar informações específicas em grandes coleções de dados em, aproximadamente, tempo real (ELASTIC, 2017).

O Elastic Search utiliza a arquitetura REST que permite interoperabilidade entre sistemas web e a internet. Devido ao esse conjunto de regras bem definidas, o ES fornece um serviço de fácil instalação e manuseio que utiliza de objetos JSON[4] para comunicar-se com sistemas que utilizam seu serviço. Seu serviço pode ser facilmente acessado através da sua Interface de Programação da Aplicação (API) que está disponível em seu site oficial para download em diversas linguagens como Javascript, Java, PHP, Ruby e Python (ELASTIC, 2018).

Uma característica de sistemas distribuídos presente no Elastic Search é a possibilidade de escalar dados e recuperação de falhas de leitura e escrita. O indice contém grande quantidade de informações indexadas que pode exceder o limite de dados especificado pelo hardware de um de servidores da aplicação. Para isso o Elastic Search dispõe de estruturas chamadas *Shards*. Índices podem ser subdivididos em multiplas *shards*. O motor de busca é capaz de automaticamente administrar o uso das *shards* e lidar com falhas. Em caso de falhas, o motor conta com replicas, cópias das shards que podem substituir suas originais em caso de perdas (ELASTIC, 2017).

[3] baseia-se no princípio de que todos os dados estão armazenados em tabelas

[4] É uma estrutura e formato de arquivo popularmente utilizado pela facilidade de leitura e escrita de informações por computadores e humanos.

2.5.2 Apache Lucene

Apache Lucene é uma biblioteca de pesquisa completa com código aberto que contém algoritmos utilizados para busca, indexação e ranqueamento de texto. Assim como o Elastic Search, Lucene foi construída utilizando a linguagem de programação Java, com suporte a API por outras linguagens como Python, C++, Ruby e PHP. O modelo de recuperação de informação utilzado por essa ferramenta é híbrido. Utilizam-se os modelos vetorial e booleano para encontrar resultados da busca (FOUNDATION, 2016).

É devido ao Apache Lucene que motores de busca como Elastic Search e Solr conseguem indexar milhões de documentos e acessá-los em questões de segundos. Na indexação de documentos e com a utilização dos indices lucene, a biblioteca indexa coleções de documentos ocupando apenas 20% do valor de armazenamento desses documentos em disco. O propósito de indexar arquivos, é aprimorar o tempo de busca de termos em documentos e diminuir o uso de energia do hardware (FOUNDATION, 2016). Assim torna-se possível executar buscas em conjunto de outros algoritmos após a busca como classificação, ordenamento de resultados e prover serviços de busca a sistemas de recuperação de informação.

Lucene é utilizado popularmente por grandes empresas de tecnologia em seus sistemas de recuperação de informação fornecendo confiabilidade, eficiência e disponibilidade através de seu serviço de busca. A biblioteca de pesquisa pode ser utilizada para análise de informação, identificação geográfica, classificação de textos e organização de grandes quantidade de dados. A exemplo, Lucene pode identificar em uma coleção de páginas web o melhor conteúdo relevante a um determinado anúncio, encontrar um termo presente em milhões de projetos do Github, identificar geograficamente localidade de entidades presente nos documentos[5] ou ainda, analisar estatísticas sobre a experiência do usuário em um jogo[6]. Apache Lucene cria oportunidades para diversos nichos de negócio envolvendo empresas de pequeno a grande porte, fornecendo ferramentas de busca para qualquer tipo de informação não estruturada (LUCENEWIKI, 2015).

2.6 Processamento de Linguagem Natural

O Processamento de Linguagem Natural (PLN) é uma área explora as maneiras como computadores podem ser usados para entender e manipular a fala ou textos em linguagem natural (CHOWDHURY, 2003).

O PLN consiste no desenvolvimento de modelos computacionais para a realização de tarefas que dependem de informações expressas em alguma língua natural, como

[5] <https://github.com/chrismattmann/lucene-geo-gazetteer>
[6] <https://goo.gl/7Be5SW>

tradução, interpretação de textos e busca de informações em documentos (RUSSELL, 1995).

Assim, qualquer estudo que utiliza técnicas de leituras, estruturação e entendimento das palavras pelo computador está relacionado com processamento de linguagem natural. Alguns exemplos que utilizam PLN são:

- Extração de Informação: encontrar semântica, relações de texto (sujeito, verbo, objeto por exemplo) ou extrair informações em uma base de conhecimento classificada.

- Reconhecimento de Entidade de Texto: Identificar se a palavra é uma pessoa, um objeto, um animal, uma instituição ou outra entidade.

- Manipulação básica de string: extração de subpalavras, comparações de textos, identificação de radicais (stemming), identificação de sinônimos, aplicação de filtro de palavras, identificação de palavras chave

- *Part of Speech Tagging*: Identificar valor morfológico e sintático das palavras através de um corpo treinado.

- Inferência de Resultados: baseado em um modelo de dados, sistemas ou algoritmos são capazes de inferir um valor ou idéia baseado em análise de palavras contidas em um corpo de texto.

- Perguntas e Respostas: dado uma pergunta em linguagem humana, o sistema é capaz de analisar a pergunta e respondê-la usando um corpo treinado. Geralmente são perguntas diretas e simples como "Qual é a capital do Canada?" ou "Qual artigo da lei 8078 que enfatiza os direitos básicos do consumidor?". A análise da pergunta aumenta a problemática cada vez que identifica perguntas mais abertas, isto é, com cunho interpretativo ou pessoal como "Qual é o sentido da vida?" ou quando uma pergunta pode apresentar várias respostas. "Quais são os direitos de restituição que o consumidor tem em caso de fraude?" Como a pergunta tem cunho muito geral, a pergunta acaba gerando outros tipos de questionamentos como: Qual tipo de fraude o consumidor sofreu? Foi através de serviço ou produto? A quanto tempo isso foi ocorrido?

Nas próximas seções são abordados com maior especificidade quais técnicas foram utilizadas no processamento do texto antes de envia-lo para o motor de busca executar consultas.

2.6.1 Stopwords

São chamadas de *Stopwords* as palavras que apresentam frequência comum nos textos ou que apresentam baixo valor semântico em relação a outras palavras do documento ou de coleções de documento. Essas palavras são geralmente removidas dos textos antes de aplicar um processamento textual (MANNING, 2009).

Antes de aplicar outras técnicas de processamento de linguagem natural, geralmente é preparado uma lista de palavras (stopwords) para que o sistema as ignore. Essa lista de palavras é também chamada de *stoplist*.

As *stopwords* consideradas para remoção neste estudo incluem artigos, pronomes, advérbios, verbos de conexão, interjeições, preposições e conjunções. Foram também consideradas algumas expressões como *stopwords*. Exemplo dessas expressões são: "Peço que", "Requisito", "Gostaria de", "Ficaria feliz em"

No campo de busca do sistema considera-se que a queixa digitada será um texto que irá ser utilizado para busca de artigos similares. As palavras ou expressões consideradas *stopwords* só dificultam a tarefa do sistema para encontrar artigos relevantes. Apesar de o motor de busca encontrar expressões similares a essas palavras, os resultados encontrados não indicam valor algum já que essas palavras estão presentes em quase todos os documentos.

Segue um exemplo de queixa abaixo com *stopwords* em negrito.

"Efetuei **o** pagamento **de uma** camisa **na** LojaProject. **Estou** aguardando **há 2** semanas **e ainda** sim **o site não reconhece meu** pagamento **apesar de ter sido** descontado **no meu** cartão **de** crédito. Entrei **em** contato **e eles** não respondem **há 10** dias. Queria receber **meu** dinheiro **de** volta."

Nesse exemplo são mostrados diferentes *stopwords* presente nas queixas do consumidor. Algumas aparecem mais de uma vez e todas não expressam a ideia da queixa do usuário. Pode-se identificar 23 stopwords no exemplo de queixa anterior:

3 - e, 4 - de, 1 - uma, 1 - na, 1- estou, 1 - o, 3 - meu, 1- ter, 1 - sido, 1 - no, 1 - em, 1 - eles, 2 - há, 1 - apesar, 1 - ainda

As palavras chaves são extraídas após a remoção de *stopwords*. O algoritmo do motor de busca utiliza as palavras chave para similaridade e comparação de texto pois essas palavras, geralmente substantivos, expressam algum valor semântico do texto. A queixa após extração das *stopwords* citadas, fica da seguinte maneira:

"Comprei efetuei pagamento camisa LojaProject. Aguardando semanas ainda sim site não reconhece pagamento apesar descontado cartão crédito. Entrei contato não respondem dias. Queria receber dinheiro volta."

Após remoção das *stopwords*, espera-se que o sistema encontre menos ruídos, isto é, que encontre menos resultados contendo palavras semelhantes sem valor semântico. Por exemplo, percebe-se que essa queixa trata-se de um caso de restituição de dinheiro. O sistema pode-se chegar a mesma conclusão graças a remoção de stopwords e da extração de palavras chave contidas nesse exemplo.

Quando as *stopwords* não são removidas, ao enviar esse conjunto de palavras ao Elastic Search, o sistema retorna como resultado vários outros artigos não relacionados a restituição como pode ser visto através da figura 7. Isto deve-se ao fato do motor de busca ter encontrado várias ocorrências de *stopwords* nesses artigos. Logo percebe-se que o uso das *stopwords* na queixa diminuiria a eficácia do motor de busca na recuperação de artigos legais relevantes.

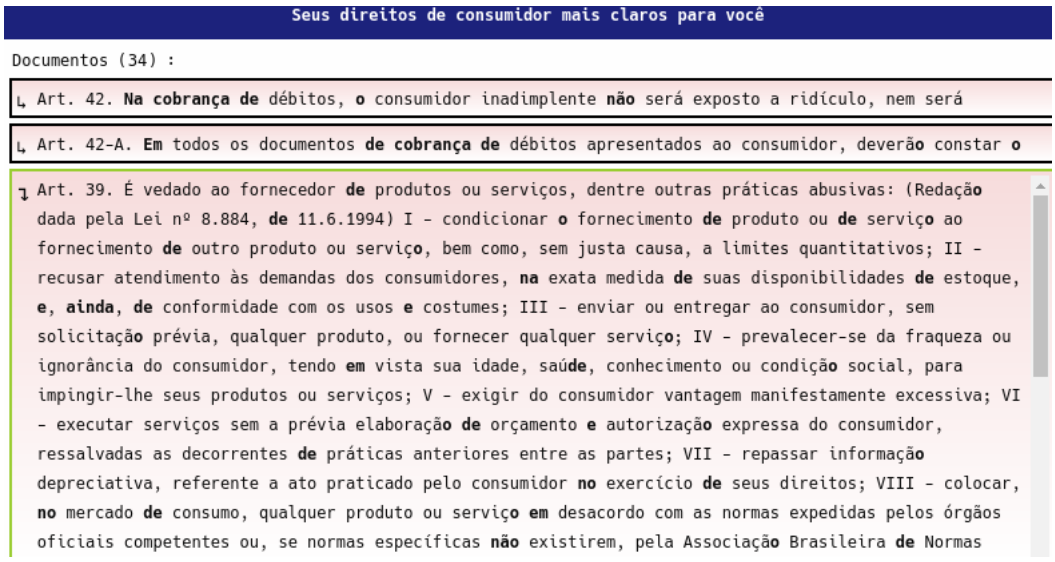

Figura 7 – Resultado de artigos encontrados pelo motor de busca sem remoção de stopwords da queixa

De maneira similar na figura 8, são apresentados os artigos legais recuperados pelo motor de busca sem as *stopwords*. Em negrito contém as ocorrências de similaridade entre as palavras chave da queixa e do texto legal.

Ao processar algumas queixas do usuário, é possível encontrar pontuações próximas as palavras fazendo com o que o computador reconheça como uma palavra nova por causa da pontuação "aglutinada" a outra palavra. Com a utilização de expressões regulares (Seção 2.6.2), é possível identificar símbolos ou caracteres especiais e removê-los da queixa do usuário. Essa remoção torna-se necessária, pois esses caracteres especiais também podem aparecer frequentemente nos textos indexados.

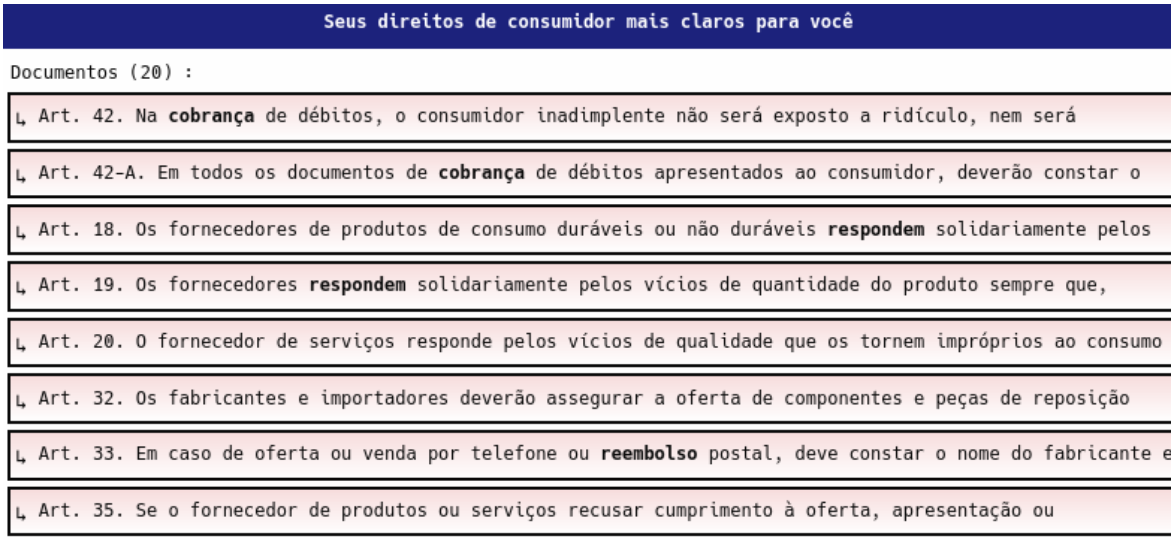

Figura 8 – Resultado de artigos encontrados pelo motor de busca com remoção de stopwords da queixa

2.6.2 Expressões Regulares

A forma mais direta de aplicar a remoção de stopwords e outras palavras indesejáveis é através das *Expressões Regulares*. As Expressões Regulares conhecidas popularmente pelo termo *Regex*, são utilizadas no reconhecimento de padrões textuais. As expressões regulares são formadas através de expressões especiais que representam a identificação de determinadas palavras, símbolos, números, caracteres ou um grupo composto por um ou mais destes elementos.

Para processar o texto, as expressões regulares podem ser utilizadas para retornar informações como presença de palavras, um novo texto contendo ocorrências de palavras removidas pelas expressões regulares, um novo texto com palavras ou caracteres substituídos e por fim, um subconjunto de palavras ou textos (MOZILLA, 2018a).

O conjunto de stopwords de um linguagem trata-se de um lista de palavras. Para melhor percorrer essa lista de palavras, as stopwords são inseridas em um vetor chamado de *stoplist*. A expressão regular responsável pela identificação dessas stopwords, percorre o texto comparando as palavras contidas nele com as palavras da lista de stopwords. A expressão regular nesse caso, são as próprias palavras da stoplist contidas entre dois caracteres *backslash* como ilustra a figura 9.

2.6.3 Bag-of-Words

No modelo “Bag of Words”, seu exato ordenamento dos termos em um documento é ignorado mas o número das ocorrências do termo é contabilizado. Consequentemente, o seguinte texto “Maria é mais rápido que João” é, neste ponto de vista, idêntico ao docu-

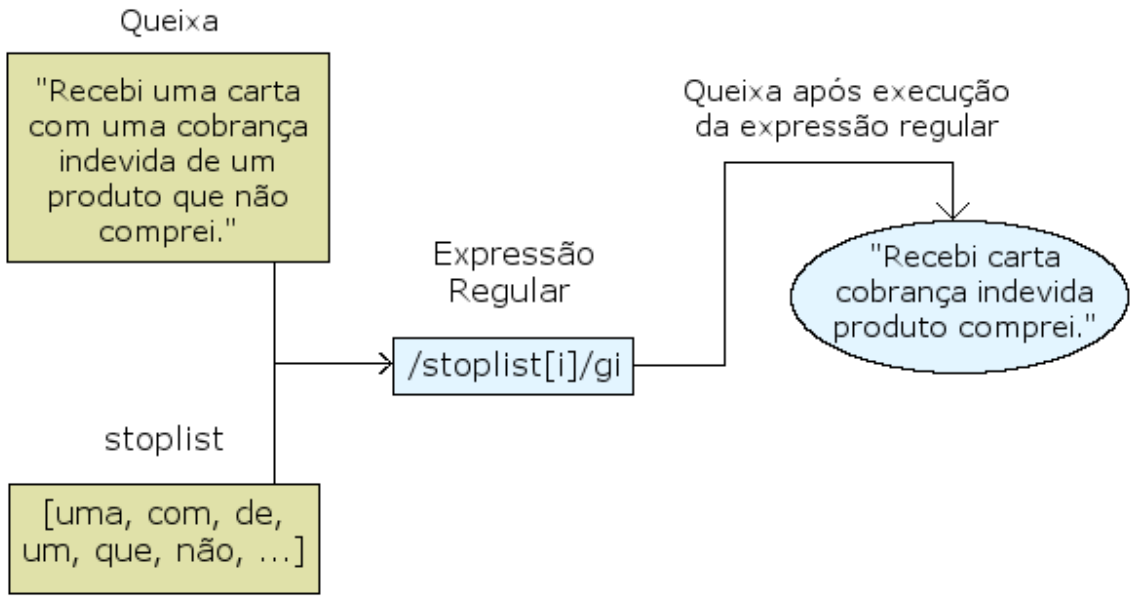

Figura 9 – Exemplo ilustrando a remoção de stopwords de um texto

mento “João é mais rápido que Maria”. Através desse exemplo observa-se a intuitividade e propósito do modelo em mostrar que dois documentos com similar representações de conjunto de palavras são similares em conteúdo (MANNING, 2009).

A ideia básica desse modelo é que um bloco de texto é visto como um conjunto de palavras e são representados em números ou especificamente, por um vetor de números. Não importa a ordem das palavras, classe gramatical, função sintática ou outro valor semântico atrelado a cada palavra do texto. A classificação de um documento de acordo com essa técnica é feita através da frequência de suas palavras num texto, logo cada documento ou texto é representado por um vetor diferente.

Abaixo será analisado um exemplo de queixa dos testes realizados no sistema. A queixa sem a utilização da remoção de *stopwords* fica da seguinte maneira:

“Eu gostaria de ter meu dinheiro de volta relacionado ao produto que eu comprei 25 dias atrás. Usei um pouco mas não estou satisfeito com o produto. Não cumpriu com o que prometia.” (Texto1)

As *stopwords* não são consideradas para contagem de frequência. Após remoção das remoção das *stopwords*, o texto fica representado da seguinte forma:

“Ter dinheiro volta relacionado produto comprei dias atrás Usei não satisfeito produto cumpriu prometia.” (Texto1 processado)

Suponha que houvesse outro exemplo de texto processado. Por fins de representação, chama-se esse texto de ‘Texto2 processado’.

"Indignado empresa enganou dinheiro prometeu desconto pagar não cumpriu prometia devolução" (Texto2 processado)

Ao compará-los entre si a fim de saber o quão similar eles são, observa-se que nos dois conjuntos de palavras, possuem palavras semelhantes. Para que o sistema chegue a esta conclusão, aplica-se a seguinte abordagem. Primeiramente, divide-se cada conjunto de palavras pelo sua representação de palavras chave. Dessa maneira, cada conjunto de palavras é representado pelas palavras que apareceram pelo menos uma vez. As palavras deste conjunto de palavras não deve conter *stopwords*. A representação do vetor de palavras contendo palavras chave dos dois conjuntos é chamado de vocabulário e pode ser visualizado deste modo:

[ter, dinheiro, volta, relacionado, produto, comprar, dias, usar, não, satisfazer, cumprir, prometer, indignado, empresa, enganar, desconto, pagar, devolução]

Para contabilizar a presença da palavra chave em um dos textos será utilizado um vetor binário em que a posição do vetor representa a palavra presente no vetor acima. Para cada texto então, tem-se a seguinte representação:

Texto1: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0]
Texto2: [0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1]

Nesse exemplo não são levados em consideração a frequência das palavras nesse texto. Outra observação é que algumas palavras são tão semelhantes ou sinônimos que ambas podem ser contabilizadas uma única vez. A exemplo:

palavras semelhantes do Texto1: pagar, comprar, efetuar, adquirir.
palavras semelhantes do Texto2: restituir, devolução, de volta.

Comparando os dois textos através dos valores de cada vetor, é encontrado o valor de similaridade de 4/18. Isto é, apenas 4 palavras do total de termos estão presentes nos dois textos. Este valor indica uma similaridade muito baixa. Porém, sabe-se que algumas palavras possuem valores semânticos relevantes entre si. Logo, se levado em conta que essas palavras são sinônimas ou apresentam uma relevância elevada com as outras isso aumentaria o valor de similaridade entre dois textos trazendo um resultado mais realístico.

Palavras como 'pagar' e 'comprar', 'devolução' e 'volta', 'indignado' e 'satisfeito', 'enganar e prometer' tem seus significados relacionados. Ao considerar essas palavras como equivalentes ou semelhantes, a contagem das palavras é considerada somente uma vez. Essa ação facilita o sistema a identificar palavras com termos associados ou equivalentes.

Diferente do modelo booleano de recuperação de informação visto na seção 2.1.3, essa é uma forma de levar em consideração que as palavras de um vocabulário não são inter-dependentes. O vetor binário de palavras de cada texto após identificação de palavras equivalentes, fica na seguinte representação:

Texto1: [1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1]
Texto2: [0, 1, 1, 0, 0, 1, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

A similaridade entre os dois vetores aumentou seu percentual para 10/18, aproximadamente: 55%. Este valor representa uma boa indicação de que os textos são similares ou pertence a uma mesma classificação. A abordagem de *Bag of Words* descrita nesta seção foi usada como inspiração para o algoritmo de similaridade de queixas. A diferença entre o algoritmo usado na plataforma *MeusDireitoConsumidor.com.br* e esta abordagem é que utiliza-se um vetor de palavras ao invés de vetor de números. O algoritmo de similaridade é descrito na seção 4.7.

2.7 Tecnologias Web

2.7.1 HTML

HTML é a linguagem que descreve a estrutura das páginas web. Pode ser usado para marcação de página fornecendo organização e estrutura para o documento. A página web pode também conter informações estruturadas sobre sua semântica ser encontrada na internet mais facilmente pelos usuários através dos motores de busca (COMUNITIES, 2011). HTML fornece aos autores de conteúdo as seguintes funcionalidades:

- Publicar documentos online com cabeçalhos, textos, tabelas, listas, imagens, etc.
- Extrair informações online através de links hipertextos.
- Construir formulários com transações para serviços remoto para uso em pesquisa de informação, pedidos de reserva, requisições de produtos e outros serviços.
- Incluem planilhas, videoclipes, clipes de sons e outras aplicações em seus documentos.

Com HTML, os autores descrevem estrutura das páginas utilizando linguagem de MARKUP. Os elementos da linguagem podem conter pedaços de elementos como parágrafos, listas, seções, links, tabelas e outros (Figura 10).

```
<!DOCTYPE HTML>
<html lang="pt-BR">
<head>
<meta charset="UTF-8">
<title>Digitar Título Aqui</title>
<meta name="description" content="Essa é a descrição da página">
<meta name="keywords" content="chatbot, CDC, Consumidor, PLN">
<meta name="author" content="Autor da página">
<meta name="viewport" content="width=device-width, initial-scale=0.85"">
<body>
<header>
    <nav>
        <ul>
            <li><a href="/">Início</a><li>
            <li><a href="/features/">Funcionalidades</a><li>
            <li><a href="/login/">Login</a><li>
        </ul>
    <nav>
</header>
<section>
    <h1> Título desta seção</h1>
    <div> conteúdo inicial do site</div>
</section>
<footer>Rodapé do site</footer>
<script src="js/javascript.js" type="text/javascript"></script>
</body>
</html>
```

Figura 10 – Exemplo de sintaxe HTML

2.7.2 Cascading Style Sheets (CSS)

CSS é uma linguagem criada para melhorar a aparência dos sites construídos com as linguagens de marcação como HTML ou XML. Atualmente, é fundamental que todo site seja bem formatado com essa linguagem. Sem a utilização das folhas de estilo (CSS), o site acaba tornado-se muito simples podendo impactar na experiência de um usuário no manuseamento e leitura de textos ou de execução de serviços online. Diversas empresas utilizam cada vez mais um design dinâmico e responsivo para atrair clientes para seus serviços (POUNCEY; YORK, 2011).

A linguagem permite também adaptar apresentações de páginas em diferentes tipos dispositivos como telas grandes geralmente em computadores ou pequenas como a de dispositivos móveis. No entanto CSS é independente do HTML e pode ser usado em qualquer documento baseados em XML. A separação também permite facilidade para manutenção de sites, compartilhamento de folhas-estilo e construção de páginas em diferentes ambientes.

A linguagem de estilização pode ser utilizada nas páginas web através de arquivos separados reconhecendo seus elementos pelos nomes de classe, id ou posicionamento dos elementos na linguagem de marcação da página (Figura 11).

```css
#results-search .title{
  width: 100%;
  display: inline-block;
  padding: 0 0.5em;
}

#div-chatbot{
  height: 100%;
}

#div-chatbot span.b{
  display: inline;
}
```

Figura 11 – Exemplo sintaxe CSS

2.7.3 MySQL

MySQL é código aberto, multithread, sistema de banco de dados relacional criado por Michael "Monty" Widenius em 1995. Atualmente, estimam 6 milhões de instalações do MySQL no mundo afora, e são reportados uma média de 50 mil downloads por dia do instalador MySQL for seu site oficial e de sites espelhos. Sua utilização cumpre importante função no armazenamento e organização de informação. Sistemas que utilizam acesso constante de informações utilizam MySQL devido a confiabilidade que a ferramenta apresenta e simplicidade das consultas utilizadas para buscar informação (DYER, 2008).

Devido a utilização de consultas SQL[7], um sistema web é capaz de acessar informações contidas em tabelas no banco de dados e gerar conteúdo dinâmico nas páginas do sistema. Ao criar uma base de dados para alimentação das informações deste sistema foi construído um modelo de entidade-relações de tabelas ilustrado abaixo (Figura 12).

[7] linguagem utilizada pelo MySQL para consulta de informações

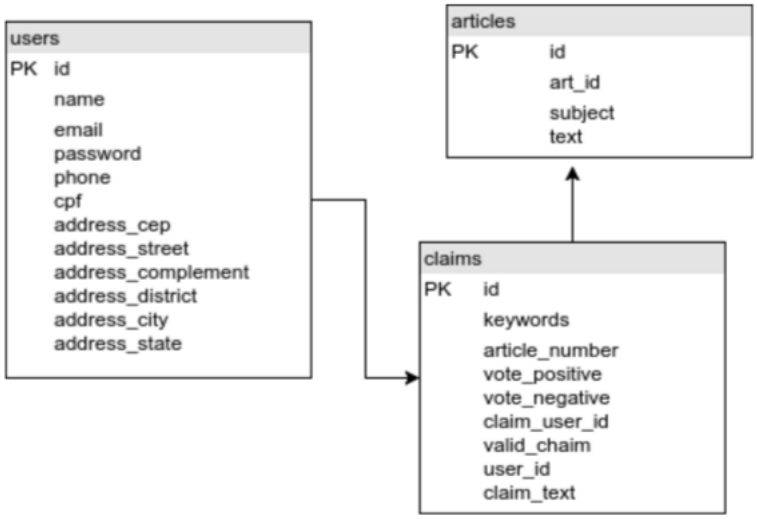

Figura 12 – Modelo entidade-relacionamento de tabelas criadas para o projeto

Através deste modelo, a aplicação deste estudo permite o cadastro e consultas de usuários do sistema, consulta de artigos do código de defesa do consumidor e registro de queixas dos cidadãos.

2.7.4 Javascript

Javascript é uma linguagem de script ou programação que permite implementar tarefas complexas em páginas da web exibindo atualizações de conteúdo oportunas, mapas interativos, Gráficos 3D e reproduções de streaming em tempo real (MOZILLA, 2018b).

Javascript é basicamente uma linguagem complementar, o que significa que é incomum para um aplicativo inteiro ser escrito exclusivamente em JavaScript sem o auxílio de outras linguagens como HTML e CSS.

O Javascript pode executar muitas tarefas no lado cliente da aplicação. A linguagem permite adicionar a interatividade necessária para um site criando menus suspensos, transformando o texto em uma página, filtrando informações digitadas pelo usuário em formulários ou ainda processar requisições assíncronas entre serviços web através de cliques do usuário (SUEHRING, 2013).

2.7.5 Vue.js

Vue[8] (pronunciado , como 'view') é uma estrutura para construir interfaces com o usuário. o Vue foi projetado desde o início para ser utilizado em paralelo com a evolução de um sistema. A biblioteca central é focada apenas na camada de visualização e é fácil

[8] <https://github.com/vuejs/vue>

de ser coletada e integrada a outras bibliotecas ou projetos existentes. o Vue também é perfeitamente capaz de alimentar aplicativos sofisticados de uma única página como é feito com a plataforma MeusDireitoConsumidor.com.br.

Vue.js suporta todos os navegadores que implementam *ECMAScript5*(ES5). Internet Explorer 8 e outros que não possuem ES5, não são suportados pelo Vue.js. Outros frameworks similares ao Vue.js bem conhecidos no mercado são o React, Ember e Angular (YOU, 2018).

Vue.js é utilizado na aplicação do MeusDireitoConsumidor.com.br para armazenamento e comunicação de informações entre a interface e os serviços da aplicação (motor de busca, banco de dados e servidor da aplicação). Além disso lida com a organização dos dados para visualização no chatbot e com a organização dos casos processados de queixas similares na interface do sistema.

2.7.6 Heroku

Heroku[9] é uma plataforma na nuvem que permite companhias construir, entregar, monitorar e escalar aplicações. A Heroku também conhecida por fornecer um tipo de serviço específico conhecido como PaaS, Plataform-as-a-Service. Permite usuário a construir aplicações customizadas e implanta-las no servidor na nuvem sem modificações no código da aplicação. Heroku irá gerenciar toda complexidade relacionada à hospedagem e processamento da aplicação (HEROKU, 2018).

Para os desenvolvedores de aplicações, Heroku tem-se mostrado uma ótima alternativa para registrar suas aplicações em servidores em nuvem, não precisam se preocupar com infraestrutura, escalamento da aplicação ou lidar com configurações do servidor. Além disso o Heroku disponibiliza planos gratuitos para aplicações em diversas linguagens como Java, Php e Javascript (HEROKU, 2018).

Serviço de hospedagem fornecido pelo Heroku foi utilizado para hospedar as aplicações: Elastic Search e MeusDireitoConsumidor.com.br por conter os requisitos mínimos necessários para rodar as duas aplicações e devido a disponibilidade do serviço no plano gratuito.

[9] <https://www.heroku.com/>

3 Trabalhos Relacionados

Neste capítulo serão apresentados alguns trabalhos existentes que atuam na mesma esfera do direito do consumidor. Serão descritos seus conceitos, propósitos e suas limitações. Todos os sistemas deste capítulo possuem semelhanças com o sistema criado através deste trabalho.

3.1 Reclame Aqui

Com a pretensão de expor sua opinião, O cidadão consumidor insatisfeito com aquisição de um produto ou serviço pode registrar uma reclamação no website *ReclameAqui* na tentativa de encurtar o contato com o fornecedor. As empresas, que tentam manter a boa imagem de sua marca, entram em contato através do site para tentar resolver o problema descrito pelo consumidor (RECLAMEAQUI, 2018).

Na plataforma é possível também verificar a estatística de queixas solucionadas, ranking de empresas do mesmo negócio e índice de satisfação dos consumidores ou ainda, sugestões de empresas para efetuar negócios novamente. Como sugerido pelo PROCON e pelo Ministério da Justiça é aconselhado que todo conflito entre consumidor e fornecedor deve ser, inicialmente, resolvida entre os envolvidos. A plataforma, com objetivo de induzir o contato entre os envolvidos, tenta aproximar o consumidor e o fornecedor para iniciar um diálogo e buscarem um solução entre si.

A plataforma traz como grande vantagem um canal a mais de comunicação entre o consumidor e o fornecedor. Além disso, o usuário pode acompanhar casos de outros usuários, verificar transparência da empresa e consultar número de casos resolvidos. Nesse sentido a plataforma torna-se uma boa opção para usuários que queiram consultar reputações das empresas antes de fechar negócio com elas.

Devido a grande quantidade de informações fica difícil o usuário encontrar casos válidos de infração na área do direito do consumidor. Isto é, existem casos registrados que não se tratam de relação entre consumidor e fornecedor e sim de uma problemática na área do direito civil. Ao descrever suas queixas, os cidadãos fazem confusão de direitos digitando queixas sem fundamento ou com informações insuficientes do ocorrido. Queixas essas que podem favorecer o fornecedor em sua defesa.

3.2 Consumidor.gov.br

Apesar da similaridade com a plataforma anterior, a plataforma *Consumidor.gov.br* é mais reconhecida pelas empresas por ser fornecida pelo governo público. Logo exerce maior pressão às empresas para que respondam as reclamações dos consumidores.

A plataforma foi criada em parceria com os criadores do *ReclameAqui*. É monitorada pela Secretaria Nacional do Consumidor - SENACON - do Ministério da Justiça, Procons, Defensorias, Ministérios Públicos e por toda a sociedade. No seu website, podem ser consultadas a imagem das empresas perante as experiências dos consumidores. Assim como o *ReclameAqui* propósito da ferramenta é permitir a aproximação entre o consumidor e fornecedor na resolução de conflitos.

A limitação desta aplicação é que não possui nenhum indicativo ou sugestão para o cidadão solucionar seu problema em caso de discordância entre os envolvidos. O consumidor lesionado em geral não sabe o que fazer em segunda instância. o contato entre os envolvidos podem durar meses e ambos não chegarem a um consenso e o consumidor desacreditado prejudicar-se ainda mais durante a busca por uma solução com o fornecedor.

Apesar de ser um sistema do governo público, a plataforma não substitui o atendimento ao PROCON. Como encontrado nas queixas da plataforma, o consumidor lesionado não deve aguardando por longos períodos em situações graves devido a indisposição das empresas quanto a solução dos seus problemas. Para resolução de conflitos, o consumidor deve conscientizar-se dos seus direitos e abrir processos nas delegacias de defesa do consumidor. Essas são as melhores atitudes a serem tomadas diante de abusos provocados pelas empresas e pela construção de um mercado mais democrático.

3.3 Mooba

Também conhecido como “Clube do Consumidor do ReclameAqui”, os usuários cadastrados no sistema tem a vantagem de realizar compras através de um cadastro de itens selecionados pelos desenvolvedores reconhecidos como “de qualidade” ou altamente confiável. Após selecionar a compra o cidadão pode receber descontos pelo site do Mooba. O objetivo principal é oferecer serviço de satisfação garantida incluindo o *CashBack*, dinheiro de volta. Caso o usuário não esteja satisfeito com o produto, a aplicação fica responsável pelo contato com a empresa para retorno do valor pago ao cliente (MOOBA, 2018).

A aplicação se mostra útil e eficiente em compras evitando possíveis estresses e transtornos causados por problemas pós-compra. No entanto, o cadastro dos itens disponibilizados para compra não cobre grande variedade de produtos e também não informa ao consumidor como resolver discordâncias entre os envolvidos.

3.4 DoNotPay

A plataforma web foi criada por um estudante de segundo ano da Universidade de Standford, Joshua Browder. Com objetivo de ajudar usuários a contestarem bilhetes de estacionamento na região de Nova York e Londres de forma intuitiva e ágil através de um chatbot (DONOTPAY, 2018).

Durante o chatbot em seu site, uma série de perguntas são feitas para confirmar se é possível invalidar a emissão do bilhete ou retorno de débito através da queixa do usuário. Por exemplo, caso o local onde houve emissão do bilhete não haja sinais claramente visíveis sobre a não permissão de estacionamento, os usuários são guiados por outras sequência de perguntas para emissão do documento final e dar entrada no processo de apelação.

Alguns dos casos são simples e comuns mas por falta de informação ou devido a burocracia dos órgãos públicos o cidadão acaba por não recorrer ao prejuízo. A maioria dos cidadãos escolhem por "evitar" o transtorno e tempo investido na solução do problema e acabam por pagar o valor do bilhete.

Na ultima vez que foi acessado o sistema (Fevereiro/2018), o site informava que é capaz de lidar também com outros casos: problemas com atraso e compensações por mudanças de data de vôo, documentos relacionados a termos de serviço e ainda alguns casos relacionados a documentação com imigração.

Este sistema assemelha-se de forma mais direta com a plataforma deste estudo. Um dos motivos principais disto é que o consumidor é informado dos seus direitos dentro do seu caso descrito. De acordo com seu site, caso o cidadão queira iniciar uma apelação, ele é informado como resolver seu problema desde que esteja dentro de Londres ou Nova York.

No momento da escrita deste estudo, as limitações desta aplicação são devidas as abrangência geográfica do publico alvo da ferramenta. É possível encontrar alguns casos não identificados pelo sistema. Porém por ser uma aplicação nova, novos casos devem ser aprendidos com o aumento de usuários por se tratar de uma aplicação que utiliza aprendizado de máquina (JOHNMANNES, 2017).

3.5 JusBrasil

A plataform online do *JusBrasil*[1] fornece conteúdo jurídico relevante aos dados digitados pelo usuário no campo de busca. Atualmente a plataforma JusBrasil pode encontrar informações como notícias, legislação, diários oficiais, modelos e peças, artigos jurídicos e jurisprudências.

[1] https://www.jusbrasil.com.br/

A JusBrasil iniciou seus trabalhos fornecendo serviço de busca de jurisprudências de forma gratuita para qualquer usuário da internet. Hoje é considerada referencia no que diz respeito a base de dados legal e principalmente, à jurisprudência no Brasil. A aplicação também utiliza o motor de busca Elastic Search e consegue realizar buscas de dados para o usuário de forma rápida e transparente para o usuário final (MATOS, 2016).

A plataforma oferece também acesso a um simples chatbot que permite encaminhar o usuário a um advogado para possível orientação acerca de seu problema. No entanto, a maioria das informações contidas nos resultados de pesquisa da plataforma contém informações especializadas aos textos jurídicos e fica difícil ao usuário leigo identificar jurisprudência ou artigos legais referentes a seu problema. Essa característica da plataforma faz com que o usuário advogado, juiz ou outro relacionado a área legal, seja o principal leitor e usuário do sistema. Isso pode ser percebido também pela seção de comentários do site em que um grande percentual dos usuários é identificado como advogado ou estudante de Direito.

4 Meus Diretos Consumidor

O sistema proposto neste estudo provê ao usuário a oportunidade de encontrar informação legal do direito do consumidor relacionado a sua queixa. Os artigos encontrados pelo sistema são os prováveis artigos que endossarão a queixa do usuário.

Como em todo planejamento de criação de sistemas, é importante listar as funcionalidades básicas que a aplicação oferece ao usuário. As principais funcionalidades da aplicação são as seguintes:

- Ao entrar no sistema, usuário deve perceber de forma intuitiva o que fazer e como usar a plataforma para enviar sua queixa; isto é, o sistema deve fornecer adequada usabilidade.
- Sistema deve permitir o usuário registrado cadastrar suas queixas e poder consultá-las quando desejar.
- O sistema deve ser capaz de identificar o tipo da queixa. A classificação da queixa ocorre após o sistema encontrar o artigo relevante. Isso acontece após a interação entre o cidadão e o chatbot do sistema.
- O sistema deve suportar um chatbot capaz de sugerir perguntas baseada no texto escrito pelo usuário com objetivo de encontrar o artigo que melhor relaciona-se com sua queixa.
- Sistema deve permitir ao usuário imprimir um documento contendo os resultados encontrados pelo sistema assim como suas: informações pessoais, queixa e a informação legal sugerida pelo sistema.

Com o desenvolvimento da aplicação surgiram novas idéias que também foram incorporadas ao sistema e documentadas durante o período deste trabalho. As novas funcionalidades que não estavam contidas nas funcionalidades básicas do sistema foram a participação do usuário na validação dos casos registrados no sistema e visualização de casos similares relacionado ao artigo encontrado pelo sistema.

Na visão técnica do sistema, as funcionalidades descritas exigem processamento da queixa, recuperação de informação e comparações entre os artigos legais. Os desafios encontrados ao implementar tais funcionalidades são os seguintes:

- A resposta dada pelo sistema precisa ser suficientemente rápida a ponto do usuário não perceber que a aplicação dependa da resposta de outra aplicação (ElasticSearch).

Uma resposta rápida do sistema indica agilidade e confiabilidade da aplicação para o usuário. Logo a transparência do serviço de busca é necessário. A falta de transparência desses serviços ou de uma resposta lenta não atrairia o cidadão a querer buscar suas informações devido a alta de espera de contato com o chatbot como acontece nos pontos de atendimento dos PROCON.

- Implementação de um banco de dados capaz de armazenar os artigos das queixas, os resultados encontrados, palavras chave das queixas registradas para busca de similaridade e dados de identificação do usuário (login)
- Registrar e consultar queixas e informações de login do usuário do sistema
- Utilizar uma biblioteca ou criar um algoritmo capaz de realizar comparações de queixas indicando similaridade entre as queixas.
- Visualizar através de uma lista, os casos mais similares ao resultado encontrado pelo sistema permitindo que o usuário os acesse.

4.1 Requisitos do Sistema

A listagem de requisitos do sistema permite que os desenvolvedores e outros participantes da produção do software possam compreender o funcionamento do sistema e suas restrições, identificar suas funcionalidades e por fim planejar de forma organizada a implementação dos requisitos através de tarefas (SOMMERVILLE, 2010).

Ao projetar o sistema foram criados e documentados 13 requisitos funcionais. A seguir possui uma lista com os items contidos nesses requisitos:

- **Abrir Queixa**: O sistema deve permitir que o usuário cadastrado abra queixas através de um texto relatando um problema encontrado nas relações de consumidor e fornecedor seja este um serviço ou produto.
- **Iniciar Contato com o chatbot:** Após realizar queixa, o usuário deve ser capaz de receber um contato do chatbot do sistema. O contato deve ser iniciado pelo sistema logo após o mesmo coletar informações referentes às queixas do consumidor.
- **Manter histórico das queixas registradas:** O sistema deve manter em registro para consulta todas as queixas registradas no sistema. Usuário poderá consultá-las futuramente, assim como, o sistema poderá utiliza-las para criar links com outras queixas. As queixas cadastradas poderão ser encontradas na área de login. Os dados que serão armazenados são a queixa do usuário, palavras chave da queixa, user_id e o número do artigo identificado como relevante.

- **Processamento da Queixa:** Ao coletar as informações da queixa, o sistema extrai as palavras chave e as envia para motor de busca encontrar um cruzamento de dados entre os artigos do código de defesa do consumidor e a queixa do usuário.

- **Relatório da queixa processada**: Ao final do processamento da queixa, caso o usuário esteja cadastrado e logado no sistema, o mesmo poderá visualizar a opção de gerar um relatório. Ao gerar o relatório, esse documento conterá a queixa registrada acompanhadas do artigo legal bem como os dados pessoais como nome, e-mail, telefone e cpf. O relatório deverá ser gerado em formato de arquivo PDF.

- **Votação da queixa analisada:** Ao final do processo, isto é, após gerar a tela de resultado, o usuário poderá dar um feedback sobre o resultado encontrado pelo sistema. O usuário realizará a votação validando o resultado encontrado através dos ícones positivo ou negativo. A resposta indicará para o sistema que a queixa e a resposta entregue ao usuário é válida.

- **Visualização dos Votos das Queixas:** O sistema deve ser capaz de mostrar para os usuários os votos realizados na queixa selecionada pelo usuários. A votação ajudar a identificar se o artigo legal gerado relaciona-se com a queixa em questão.

- **Visualização de Queixas Similares:** Ao gerar o resultado após o diálogo entre o usuário e o chatbot, o sistema mostra na tela uma lista de queixas similares ao resultado encontrado. Pode-se encontrar também a classificação dessas queixas e seu grau de similaridade com a queixa visualizada pelo usuário. As queixas similares poderão ser visualizadas ao serem clicadas.

- **Manter Usuário:** O sistema deve permitir cadastrar e atualizar informações de login ou pessoais quando solicitado pelo usuário.

- **Efetuar Login**: Ao acessar o sistema, o usuário poderá realizar login no sistema, permitindo-o ter acesso a área de login do sistema.

- **Acessar Área de Login**: O sistema dispõe de um conjunto de ações para usuários cadastrados. O usuário logado irá visualizar um botão (àrea de login) onde ao clicar será mostrado uma lista de ações que o usuário poderá tomar. A exemplo: consultar queixas de usuário cadastradas no sistema, consultar queixas realizadas pelo usuário logado, atualizar informações pessoais e gerar resultado em formato de arquivo em PDF.

- **Efetuar logout:** O sistema deve permitir que o usuário efetue logout no sistema ao clicar em um botão ou de forma automática: após 5 minutos sem realizar alguma ação.

- **Visualizar queixas registradas no sistema:** Todas as queixas poderão ser visualizadas contendo as seguintes informações: a queixa, o artigo relacionado com a queixa e links para queixas similares.

4.2 Indexando o Código de Defesa do Consumidor

O texto escolhido para indexar foi coletado no site do governo federal onde contém informação completa da Lei 8.078/1990[1]. O texto foi indexado ao motor de busca separadamente, por artigos. Isto é, cada artigo do texto legal era considerado um documento dentro da coleção de documentos indexado no motor de busca. O motor de busca (Elasticsearch) verifica ocorrências de palavras da busca nestes documentos.

A indexação dos documentos no motor de busca é feita enviando documentos JSON usando a aplicação do Elastic Search. Como a linguagem da aplicação servidor é Javascript, foi escolhida a biblioteca *elasticsearch.js* para facilitar a interação entre a aplicação web e a aplicação do motor de busca, também disponibilizada gratuitamente pelo site oficial da aplicação [2].

Foram inseridas informações juntamente com os textos para indexação no motor de buscas. Essas informações foram o título, capítulo, seção do documento legal além do conteúdo da seção composta por artigos. A pasta que contém os documentos que foram indexados no Elastic Search pode ser encontrada no repositório da aplicação do Elastic Search criado durante o desenvolvimento deste trabalho.[3]

Para atingir o objetivo desse estudo não foi necessário indexar todos os artigos da Lei 8.078/1990. Foram identificados e indexados cerca de 50 artigos que referem-se aos problemas recorrentes dos consumidores. Logo os artigos escolhidos começam a partir do artigo número 6 e estende-se até o artigo de número 54. Após os artigo 54, os artigos discorrem sobre infrações penais e por isso não contém informação relevante as queixas do usuário.

4.3 Pre-processamento da queixa

Antes de enviar a queixa para o motor de busca, o sistema aplica um processamento no texto escrito pelo usuário. Esse processamento ocorre no servidor da aplicação e nele são removidas *stopwords*, caracteres especiais. Após isto realiza-se uma verificação de sinônimos, aplicação de *lowerCase* para facilitar a comparação de palavras e por fim tem-se a extração de palavras chave. Ao final desse pré-processamento, a queixa é representada

[1] <http://www.planalto.gov.br/ccivil_03/Leis/L8078.htm>
[2] <https://www.elastic.co/guide/en/elasticsearch/client/javascript-api/current/index.html>
[3] <https://github.com/dnovaes/ConsumerRightsSearcher/tree/master/source/texts/cdc_pt>

por um vetor contendo suas palavras chave. A partir daí a queixa está pronta para ser enviada ao motor de busca (Elastic Search). Nesta seção é descrito o processo de adição de sinônimos e apresentados os filtros adicionais utilizados para remoção de palavras ou caracteres especiais.

4.3.1 Filtros Adicionais

Além das *stopwords* presentes nos textos de linguagem natural, existem alguns outros caracteres ou palavras que não possuem valor semântico e por isso não são importantes enviá-las para o Elastic Search buscar por textos similares. Abaixo tem-se uma lista de grupos de caracteres que podem ser removidos juntos com *stopwords* também utilizando expressões regulares:

- Artigos de vogais acentuadas "àaeiíouéèêëçàâäîïùûüôóö". Removidas separadamente das stopwords por um quesito de organização.
- Pontuações, marcas ou símbolos contidos na queixa: “. , ? ! / \ $ & * #%”. Exemplo: ‘problema?’ é diferente de ‘problema’
- Números romanos: “I II III IV V VI VII VIII IX X XI XII XIII XIV L C” Os textos do corpus contém vários números romanos. Com objetivo de não induzir a pesquisa a algum artigo específico esses números romanos foram adicionados ao filtro para remoção na queixa.
- Espaçamento branco duplicado ou nulo criado após remoção de palavras: Em expressão regular pode ser removidos através da seguinte expressão: “\\s+” .

Todos esses filtros podem ser encontrados na pasta pública de funções em javascript no repositório online do sistema[4].

4.3.2 Sinônimos

O fato dos Sistemas de Recuperação de Informação identificar palavras de mesmo significado constitui-se uma vantagem para reconhecimentos de termos dentro texto. Essa vantagem permite ao sistema que palavras semelhantes em seu significado aponte para os mesmos documentos indexados no motor de busca. Para consulta e possível acréscimos de sinônimos ao sistema, escolheu-se a API *http://thesaurus.altervista.org/*. O serviço *thesaurus* informa sinônimos de palavras requisitas pelo usuário ou por outro sistema.

A API funciona da seguinte forma, o sistema envia uma palavra e o serviço disponibiliza alguns sinônimos relacionados a esta palavra. Os possíveis sinônimos encontrados

[4] <https://github.com/dnovaes/chatbot-cdc/blob/master/public/js/ext_functions.js>

pela API poderiam ser adicionados a requisição de busca no ElasticSearch com objetivo de encontrar mais similaridade de palavras entre o código de defesa do consumidor e a queixa do usuário.

Nem todos os sinônimos são encontrados no serviço disponibilizado. Para lidar com esse problema, foi criado um vetor local contendo os sinônimos comuns as palavras mais buscadas ou importantes de cada artigo. Caso fossem encontrados mais sinônimos, estes poderiam ser adicionados a esse vetor de sinônimos da aplicação.

Por exemplo ao enviar a palavra 'devolver', a API retorna a palavra '*remunerar*' como resultado. Através dos testes de busca que foram realizados pela plataforma *MeusDireitoConsumidor.com.br*, as palavras 'restituir', 'remunerar' ou 'devolver' presentes nas queixas de teste indicavam que estavam relacionados aos artigos referentes a restituição. Logo a palavra 'devolver' confirma-se uma ótima candidata para sinônimo de 'remunerar' e, por isso, foi adicionado ao vetor de sinônimos da aplicação.

Ao adicionar esses sinônimos ao vetor local de sinônimos da plataforma, se um usuário usar palavra 'devolver' em sua queixa, a plataforma *MeusDireitoConsumidor.com.br* retorna o artigo correto. Isto deve-se ao reconhecimento das palavras 'devolver' e 'restituição' como sinônimas pela plataforma.

O grande benefício com uso do sinônimos é o aumento do alcance de resultados. O sistema é capaz de enviar mais palavras relacionadas com a queixa e por isso encontrar maior número de resultados. No entanto, isso não acontece todas as vezes pois pode haver que o documento indexado não contenha a nova palavra adicionada.

4.4 Realizando consultas utilizando o Motor de Busca

Ao realizar busca de queixas-texto, o código da API do ElasticSearch é similar ao usado para indexar o corpus jurídico (CDC). Isto é, utiliza-se também a biblioteca fornecida pelos site oficial do ElasticSearch, o elasticsearch.js.

O usuário digita e envia uma queixa através da plataforma web, o sistema processa a requisição no servidor da aplicação, extraindo as palavras chaves e em seguida as envia ao ElasticSearch. É sugerido ao usuário que escreva sua queixa com maior número de palavras possível, assim o motor de busca tem maior chances de encontrar ocorrências de suas palavras chave e por isso maior chances de encontrar o artigo correto.

O ElasticSearch ao receber as palavras chave, consulta os artigos que contenham as palavras chaves ordenados pela sua relevância. Caso o ElasticSearch encontre algum resultado, envia-se um dado em JSON contendo título e conteúdo dos textos encontrados para o servidor da aplicação. Caso não encontre textos semelhantes, envia-se uma mensagem de falha também em formato de JSON para o servidor da aplicação.

A Figura 13 apresenta um exemplo de resultado encontrado e retornado pelo motor de busca em objeto JSON. Neste objeto contém como parâmetro o nome do índice do corpus jurídico indexado e uma variável contendo o conteúdo do documento encontrado pelo motor de busca.

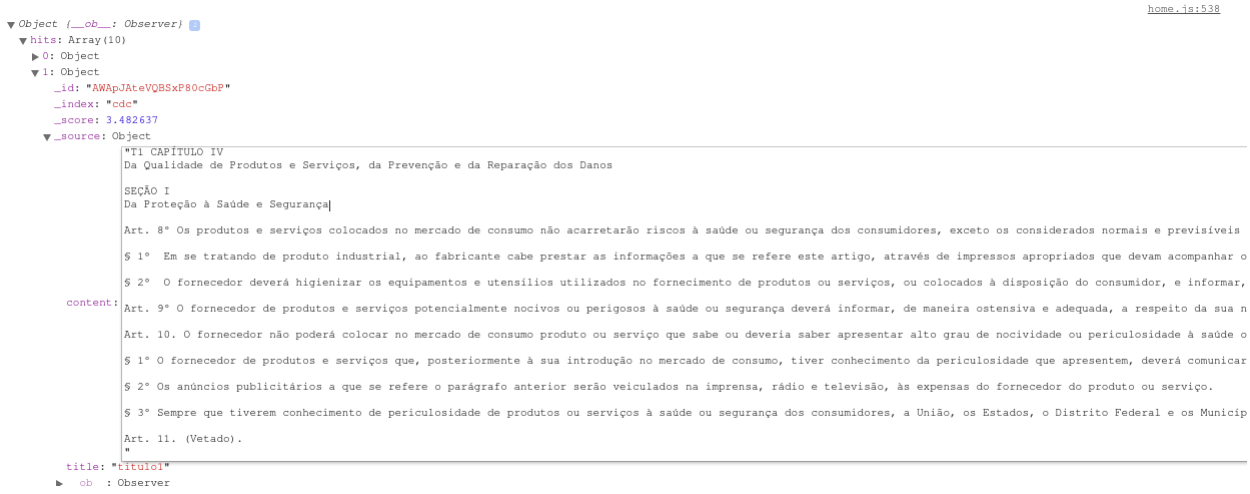

Figura 13 – Screenshot tirada do console do navegador mostrando o conteúdo do arquivo JSON retornado pela aplicação web

4.5 Interface Web

Ao acessar a plataforma web[5] a primeira informação que o cidadão visualiza é o campo de queixas. É através deste campo que o usuário digita sua queixa e o sistema inicia o processo para encontrar um artigo do CDC relacionado a queixa. É possível visualizar também em destaque as funcionalidades que o sistema oferece (Figura 14).

A plataforma web possui acesso aberto ao público cadastrado ou não. O usuário cadastrado no sistema pode acessar a área de login, visualizar registro de todas queixas realizadas, artigos legais relevantes, histórico de queixas realizadas pelo usuário e ainda dados pessoais do usuário cadastrado (Figura 15).

5 <https://wwww.meusdireitosconsumidor.com.br>

Figura 14 – Tela inicial da plataforma web

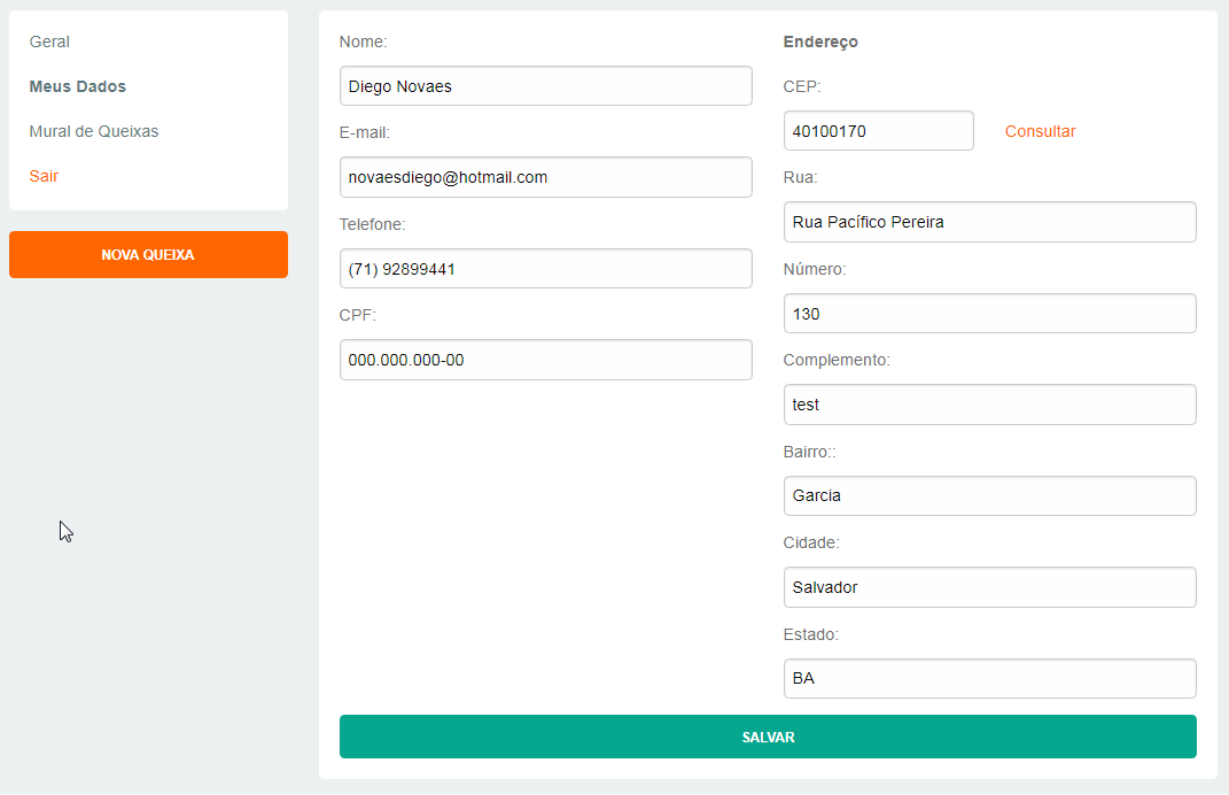

Figura 15 – Visualização e edição de informações pessoais

O usuário ao digitar sua queixa é redirecionado para a parte inferior do website onde o chatbot é iniciado (Figura 16). Caso o chatbot não seja iniciado, significa que o sistema não pôde encontrar artigos relevantes. Um dos motivos pode ser porque a queixa contém um número pequeno de palavras. o usuário precisa descrever melhor seu problema para que o sistema encontre mais artigos relevantes. Na parte inferior do website é a seção onde o chatbot interage com o usuário através de perguntas relacionadas a sua queixa. Com a interação de perguntas e respostas simples (afirmação ou negação), o chatbot elimina possíveis resultados de artigos não relacionados ao texto do usuário.

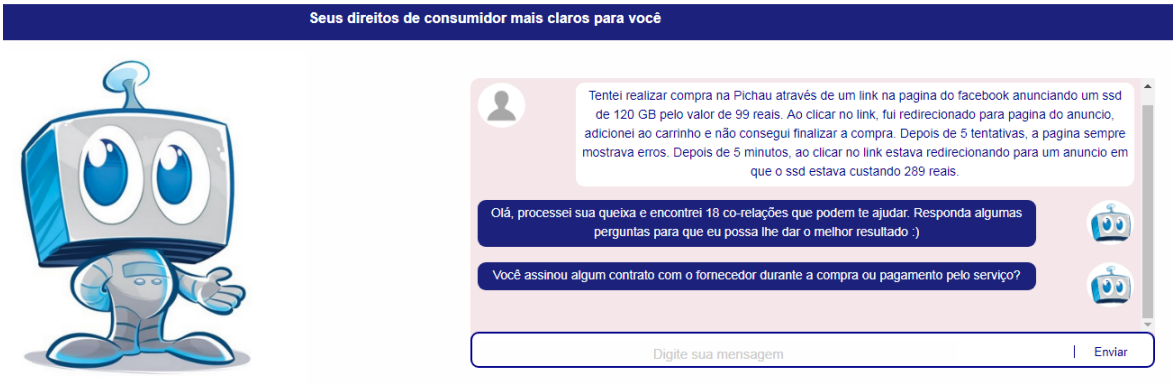

Figura 16 – Exemplo interface do chatbot

O chatbot ao receber uma resposta positiva do usuário de uma de suas perguntas, o redireciona para uma nova tela contendo o artigo que pode endossar a queixa digitada. Na mesma tela também é possível visualizar as queixas similares cadastradas no sistema caso assim existam. O processo de recuperação de artigos do CDC é descrito na seção 4.6 e o algoritmo de similaridade de queixas descrito na seção 4.7. Apesar da lista de queixas similares ser gerada na tela de resultado do sistema de forma rápida, o calculo de similaridade de queixas é feita no servidor da aplicação em contato com seu banco de dados. Na figura 17 mostra a tela de resultado do sistema ao encontrar um artigo legal que endossa a queixa do usuário.

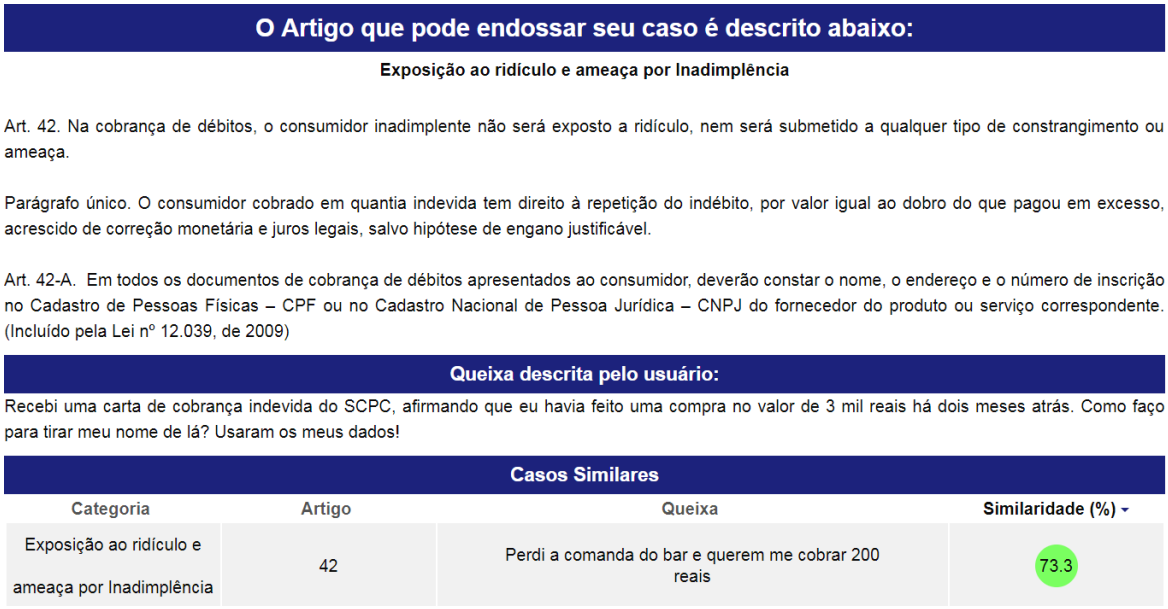

Figura 17 – Tela de resultado gerada para uma das queixas cadastradas do sistema

4.6 Recuperação de Artigos do CDC

O processo de recuperação de artigos do CDC inicia quando o usuário fornece um texto contendo sua queixa. O sistema ao receber a queixa, envia para o servidor da aplicação para iniciar um pré-processamento: remoção de *stopwords*, aplicação de *lowercase*, remoção de símbolos e caracteres especiais e por fim adição de sinônimos a busca. O servidor finaliza o pré-processamento da queixa e envia um conjunto de palavras chave para o motor de busca que está hospedado em outro servidor. O motor de busca então procura por artigos legais similares às palavras chave contidas no CDC.

A aplicação ao receber do motor de busca os prováveis textos encontrados no CDC, separa os documentos encontrados em artigos. Normalmente são encontrados mais de 5 artigos de acordo com os testes realizados. Após recuperar os prováveis artigos que endossam as queixas do usuário, inicia-se uma nova etapa com objetivo de responder a seguinte pergunta: Como identificar qual dos artigos encontrados é o mais relevante à queixa do usuário ?

4.6.1 Chatbot com perguntas diretas

Como o sistema não possui informação suficiente para identificar quais palavras são mais relevantes a um assunto do artigo do que outro, fica difícil o sistema encontrar o artigo que mais se assemelha com a queixa. Para resolver este problema surgiu-se a idéia de prolongar o contato com usuário de tal forma que através do diálogo seja possível filtrar os artigos encontrados pelo motor de busca através de perguntas diretas.

O sistema realiza perguntas diretas em que espera-se respostas simples, positivas ou negativas, como: “sim”, “com certeza”, “não”, “de acordo”, “nunca”, “claro” ou expressões similares. Mesmo que o usuário não saiba no momento a resposta das perguntas, ele poderá pulá-las ou ainda repetir o teste caso assim escolha. O usuário após responder as perguntas selecionadas pela aplicação poderá visualizar o suposto melhor artigo escolhido pelo sistema.

4.6.2 Seleção das Perguntas do Chatbot

A decisão de criar um chatbot simples com perguntas diretas para cada artigo vem da prospecção de que cada artigo contém um tema acerca dele. Ao separar os documentos para indexação no motor de busca, percebeu-se que em cada seção da lei 8078 [6] está associado a um assunto a respeito das relações entre consumidor e fornecedor. Para os artigos da lei, isso ocorre de maneira mais específica. Segue um exemplo abaixo mostrando o Artigo 18 para verificar a afirmação anterior:

> Art. 18. Os fornecedores de produtos de consumo duráveis ou não duráveis respondem solidariamente pelos vícios de qualidade ou quantidade que os tornem impróprios ou inadequados ao consumo a que se destinam ou lhes diminuam o valor, assim como por aqueles decorrentes da disparidade, com a indicações constantes do recipiente, da embalagem, rotulagem ou mensagem publicitária, respeitadas as variações decorrentes de sua natureza, podendo o consumidor exigir a substituição das partes viciadas.
>
> § 1º Não sendo o vício sanado no prazo máximo de trinta dias, pode o consumidor exigir, alternativamente e à sua escolha:
>
> I - a substituição do produto por outro da mesma espécie, em perfeitas condições de uso;
>
> II - a restituição imediata da quantia paga, monetariamente atualizada, sem prejuízo de eventuais perdas e danos;
>
> III - o abatimento proporcional do preço.
>
> § 2º Poderão as partes convencionar a redução ou ampliação do prazo previsto no parágrafo anterior, não podendo ser inferior a sete nem superior a cento e oitenta dias. Nos contratos de adesão, a cláusula de prazo deverá ser convencionada em separado, por meio de manifestação expressa do consumidor.
>
> § 3º O consumidor poderá fazer uso imediato das alternativas do § 1º deste artigo sempre que, em razão da extensão do vício, a substituição das partes viciadas puder comprometer a qualidade ou características do produto, diminuir-lhe o valor ou se tratar de produto essencial.
>
> § 4º Tendo o consumidor optado pela alternativa do inciso I do § 1º deste artigo, e não sendo possível a substituição do bem, poderá haver substituição por outro de espécie, marca ou modelo diversos, mediante complementação ou restituição de eventual diferença de preço, sem prejuízo do disposto nos incisos II e III do § 1º deste artigo.
>
> § 5º No caso de fornecimento de produtos in natura, será responsável perante o consumidor o fornecedor imediato, exceto quando identificado claramente seu produtor.

[6] Lei 8078 - CDC <http://www.planalto.gov.br/ccivil_03/Leis/l8078.htm>

> § 6º São impróprios ao uso e consumo:
>
> I - os produtos cujos prazos de validade estejam vencidos;
>
> II - os produtos deteriorados, alterados, adulterados, avariados, falsificados, corrompidos, fraudados, nocivos à vida ou à saúde, perigosos ou, ainda, aqueles em desacordo com as normas regulamentares de fabricação, distribuição ou apresentação;
>
> III - os produtos que, por qualquer motivo, se revelem inadequados ao fim a que se destinam."(BRASIL, 1990)

O artigo 18 descreve que "Os fornecedores de produto de consumo respondem pelos vícios de qualidade e quantidade que os tornem impróprios ou inadequados ao consumo". Em seguida os parágrafos contidos no artigo descrevem ainda mais, especificando cada problema inerente ao tema do artigo. Em cada artigo do código do consumidor, contém informações legais que protegem o consumidor acerca de um tema. Seguindo este raciocínio, o usuário leigo precisaria ler todos os artigos do CDC para entender do que se trata cada um. Leva-se em consideração também que uma queixa pode estar relacionada a mais de um artigo do CDC.

A partir do artigo de número 54, os artigos discorrem sobre sanções admnistrativas e infrações penais utilizadas geralmente por juízes e advogados. Portanto para este estudo, foram escolhidos os artigos de número 1 ao de número 54 para criar as perguntas e respostas utilizadas pelo chatbot. No total foram contabilizados 54 artigos da Lei 8078 Por isso 54 perguntas, no mínimo, foram criadas para implementar o chatbot. A exemplo do artigo 18, a pergunta escolhida para o chatbot foi a seguinte:

"Você não está satisfeito com a inadequação do produto mostrado na propaganda, com as informações contidas no recipiente do produto ou o produto encontra-se inadequado para uso e deseja substituir, ressarcir ou trocar-lo? (prazo máximo de 30 dias)"

Essa pergunta sumariza o artigo 18. Logo se o usuário responder sim para esta pergunta, o artigo 18 será escolhido como artigo que pode endossar a queixa do usuário ou que responderia sua dúvida em relação a um tema do direito do consumidor. Todas as perguntas criadas para o chatbot podem ser visualizadas pela planilha criada no *Google Drive* para melhor gerenciamento. [7].

4.7 Similaridade de Queixas

Ao criar essa funcionalidade, o propósito foi de mostrar ao usuário queixas similares de outras pessoas encontradas como resultado do sistema e seus devidos artigos relacionados.

[7] Detalhamento Artigos e Perguntas do chatbot - Lei 8078: <https://goo.gl/BYwE4r>

Ao finalizar o contato com o chatbot, o usuário irá visualizar: a queixa, o artigo que o endossa, uma seção para realizar votação do resultado encontrado e uma lista contendo queixas registradas no sistema que tiveram palavras chave semelhantes. Nessa lista também encontra-se um valor de 0 a 100 que representa o valor de relevância entre as queixas. Este é o grau de similaridade que a queixa da lista possui em relação a queixa digitada pelo usuário. O usuário poderá clicar nas queixas similares e uma nova tela aparecerá contendo as mesmas informações citadas e uma nova lista de queixas similares.

4.7.1 O Algoritmo de Similaridade

Neste trabalho, as queixas são ditas similares quando as palavras chave entre as queixas são iguais, sinônimas ou de alguma forma tenham valor semântico relacionados.

Inspirado na abordagem de *Bag-of-Words* descrito na subseção 2.6.3, cada queixa é representada por um conjunto de palavras armazenado no banco de dados. No momento que o usuário visualiza a tela de resultado, tela onde contém informações relacionada a queixa e o artigo selecionado pelo sistema, é executado uma requisição para o servidor web com objetivo de encontrar queixas similares cadastradas no sistema.

A busca por queixas similares é executada no servidor da aplicação em conexão com o banco de dados (MySQL). São selecionadas todas as queixas registradas no sistema e armazenada em um objeto de vetores. Cada elemento do objeto representa uma queixa registrada no sistema. Cada elemento também contém as palavras chave da sua queixa, não sendo preciso extrair suas palavras chave novamente. Em seguida, reproduz-se, uma comparação de palavras entre as queixas registradas no sistema com a queixa visualizada pelo usuário. Ao término da comparação entre as queixas, cada queixa recebe um valor em porcentagem representando a comparação entre as palavras chaves das queixa visualizada no sistema pelo usuário e as queixas selecionadas no banco de dados.

O valor encontrado pela comparação da queixa do usuário com cada queixa registrada no banco de dados é representado em porcentagem. Esse valor é chamado de "Grau de Similaridade" no sistema. Para uma queixa do banco de dados aparecer na lista de queixas similares, deve-se ter no mínimo 30% de similaridade com a queixa encontrada pelo usuário. Esse valor foi escolhido ao realizar os testes com as queixas do ReclameAqui. A partir desse valor, a interface não fica poluída com excesso de queixas na listagem de queixas similares e diminui a chance de mostrar queixas não relacionadas com a queixa do usuário. Os graus de similaridade entre queixas pode ser visualizado pelos usuário do sistema conforme mostra a figura 18.

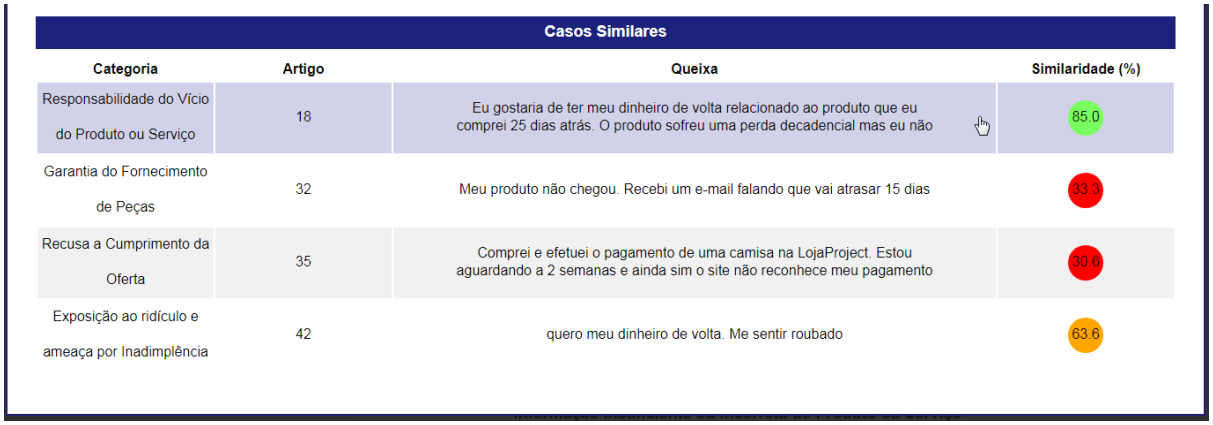

Figura 18 – Exemplo de visualização das queixas similares

4.7.2 Registrando queixas únicas

Com objetivo de controlar a quantidade de informações cadastradas no banco, surgiu a ideia de remover queixas repetidas através de um filtro. As queixas que passarem por esse filtro são reconhecidas como queixas únicas. Considera-se que uma queixa é única quando não há outra queixa registrada no banco com grau de similaridade maior que 90%. No caso, o filtro é o grau de similaridade maior que 90% entre as queixas. Quando esse valor é atingido, considera-se nesse sistema que as queixas são idênticas ou muito semelhante.

Quando o usuário envia uma queixa e o sistema reconhece que alguém já registrou uma queixa muito semelhante a ela, o sistema não inicia o processo de recuperação de artigos. Ao invés disto, a aplicação carrega imediatamente as informações do artigo que endossou a queixa considerada idêntica anteriormente. A suposta nova queixa então não é registrada no sistema.

Se o registro de queixas iguais fosse permitido, o banco teria muito mais queixas idênticas para consultar e calcular o grau de similaridade desnecessariamente por se tratarem de queixas repetidas. Além disso o surgimento de queixas repetidas na lista de queixas similares não agregaria valor algum para o consumidor.

O usuário poderá ver na lista de queixas similares outras queixas que tiveram palavras semelhantes a queixa digitada, desde que essas queixas tenham grau de similaridade abaixo de 90%. As queixas da lista, consideradas queixas únicas, também podem ajudar o cidadão na identificação de outro artigo relacionado ao seu problema.

4.7.3 Validação de Queixas por Votação

Apesar do sistema poder identificar queixas únicas, nem sempre uma queixa única pode ser considerada uma queixa válida. Para que o sistema reconheça a queixa como

válida adotou-se a seguinte medida:

Se o resultado da queixa foi acessada por outros usuários, tem número de votos maior que 3 e número de votos positivos maior que o número de votos negativos, então queixa é válida.

No resultado mostrado pela ferramenta, encontra-se uma seção em que o usuário pode votar indicando ao sistema se a queixa foi útil a ele (Figura 19). Entende-se dessa forma que o resultado encontrado ajudou algum cidadão a solucionar seu problema ou informou o artigo correto para a queixa relacionada. A verificação da quantidade de votos por queixa é feita e registrada no banco de dados do sistema assim que uma nova votação é realizada pelo usuário. No momento em que o usuário visualiza a queixa e clica no botão de votar (positivo ou negativo), o voto é contabilizado. Nesse instante, recalcula-se a contagem total de votos e atualiza o campo de queixa válida da queixa no banco de dados com a informação de queixa válida ou inválida.

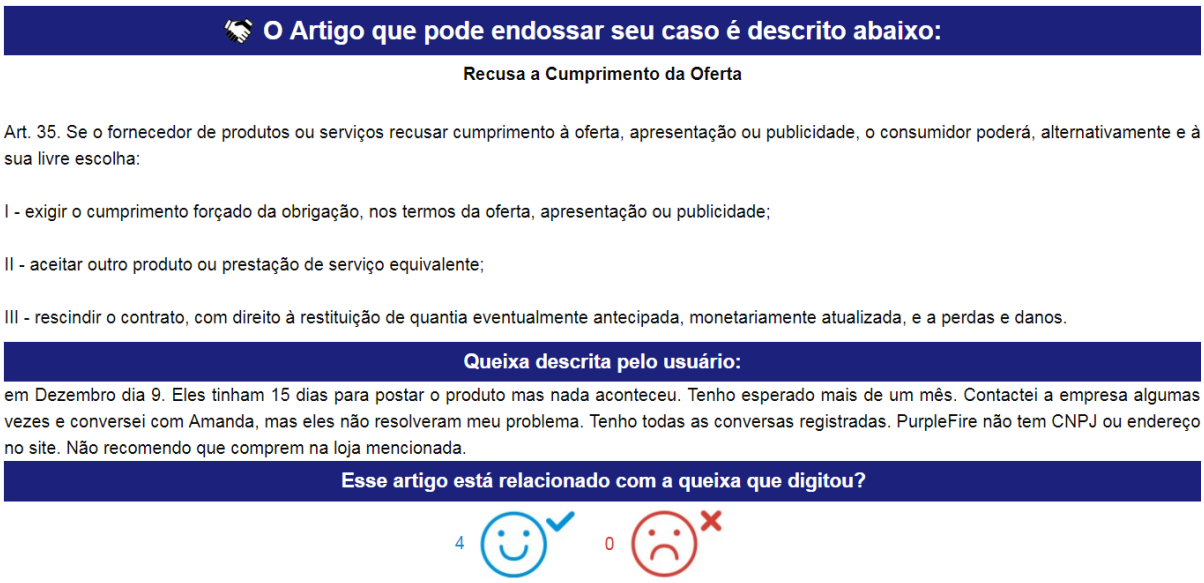

Figura 19 – Tela de resultado do artigo encontrado com opção para o usuário realizar votação

É através da lista de queixas similares ou no mural de queixas (Figura 20) que o usuário poderá selecionar a queixa, visualizá-la e votar. Ao votar, a queixa pode torna-se válida ou inválida dependendo do saldo de votos. O título de queixa válida é um valor que indica o quão importante foi o resultado da queixa para os usuários do sistema.

Figura 20 – Tela de consulta de queixas na área de login do sistema

5 Avaliação

Neste capítulo é discutido o experimento realizado com o objetivo de validar a plataforma, analisando os resultados obtidos pelos usuários no uso do sistema. Visa também identificar possíveis limitações na heurística escolhida para extrair os artigos das queixas dos usuários.

5.1 Metodologia

Só através da experimentação que é possível analisar e testar soluções. A experimentação em software pode ser feita através de testes com um grupo fechado de pessoas ou ainda utilizando somente algoritmo de testes. Porém ao considerar a experimentação aberta com pessoas em sistemas de software significa levar em consideração situações não planejadas. A criatividade e a inocência do usuário podem revelar erros ou resultados inesperados dentro de um sistema. Por isso a experimentação é importante para compreensão do problema e para criar soluções (KOZIOLEK, 2005).

Nos primeiros testes criados foram utilizados exemplo de queixas retiradas do *ReclameAqui* e do PROCON. Vários artigos foram encontrados utilizando similaridade de texto e a ajuda do chatbot. Para confirmar os resultados encontrados no primeiro teste e validar a efetividade da ferramenta foi realizado um experimento com cerca de 21 usuários de diferentes perfis.

Através do experimento online, foram colhidas informações sobre o perfil do usuário e suas conclusões sobre o uso da plataforma. Ao iniciar o experimento através do site da plataforma, o usuário recebe instruções de como será aplicado o experimento porém não é explicado como utilizar o sistema, propositalmente. Assim é possível identificar desde o início dificuldades que o usuário teve ao utilizar a plataforma.

Dividi-se o experimento em duas etapas. Na primeira, o usuário utiliza a ferramenta enviando uma queixa voltada ao mercado consumidor e verifica se o artigo legal encontrado tem haver com o problema descrito. Essa verificação é contabilizada através da pesquisa na segunda etapa do experimento ou através da votação do usuário sobre resultado encontrado dentro do sistema. Na segunda etapa, o usuário recebe um formulário criado com o *Google Forms* contendo questões que buscam compreender a experiência do usuário com a plataforma e seu perfil relacionado ao Código de Direito do Consumidor. Por fim, também foi possível colher sugestões e críticas do usuários sobre o sistema.

Um grupo de pessoas foram convidadas a participar do experimento e incentivadas a convidar outras pessoas através do link do experimento. Foram coletados cerca de 40

e-mails para participarem do experimento. O experimento esteve aberto durante 30 dias e foi realizado com 22 pessoas. O experimento poderia ser realizado através de qualquer computador desde que tenha acesso a internet e consiga acessar o site da ferramenta *MeusDireitoConsumidor.com.br*.

5.2 Resultados

Nesta seção são apresentados os dados coletados pela pesquisa respondida durante o experimento. Durante essa seção, serão feitas breves análises sobre as decisões tomadas pelos usuários e o que estas podem significar para o sistema em termos de mudanças. Ao fim desta seção será apresentado um sumário do resultado do experimento.

5.2.1 Pesquisa de perfil do consumidor

Pode-se visualizar o formulário em duas etapas. Na primeira etapa, são perguntas de caracterização do usuário, são questões sobre seu perfil e nível de conhecimento relacionado ao Código de Defesa do Consumidor (CDC). Na segunda parte, foram feitas perguntas sobre a plataforma deste estudo. Abaixo, estão listadas todas as perguntas realizadas. As respostas do fomulário podem ser visualizadas através do link direto à planilha do *Google Drive*[1].

1) Qual sua formação acadêmica?

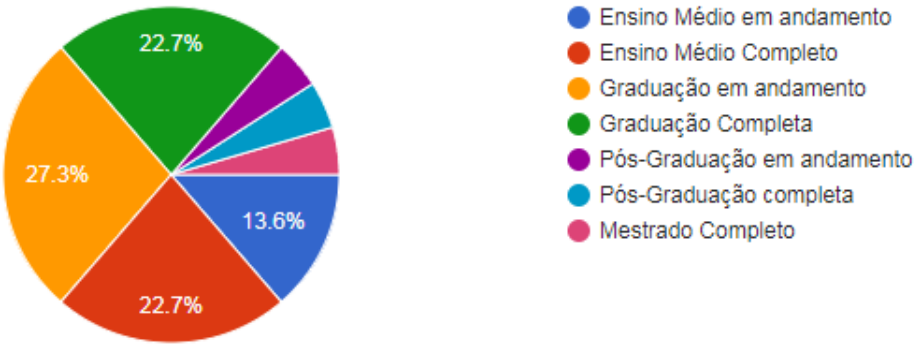

Figura 21 – Resultado da pesquisa - Questão 1

Ao iniciar a primeira etapa do questionário, questiona-se sobre a escolaridade do usuário do experimento. Foi obtido como resposta que a maioria dos entrevistados responderam ter graduação completa ou em andamento. Percentual correspondente a 50% do total de participantes (Figura 21). Com essa informação, esperava-se um bom discernimento do usuário no momento de interação com o chatbot, ler os artigos encontrados

[1] https://goo.gl/CnGc65

e comparar se a resposta dada pelo sistema confere com a digitada na queixa. Todos usuários que participaram do experimento tinham pelo menos 18 anos de idade.

2) Na constituição brasileira, existe uma lei responsável por estabelecer normas de proteção e defesa ao consumidor também conhecida como Código de Defesa do Consumidor (Lei 8078). Já ouviu falar dela?

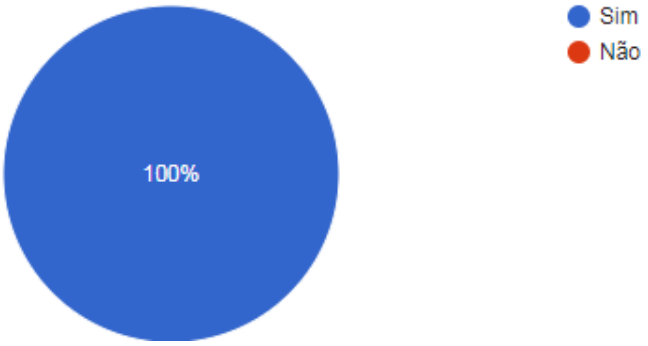

Figura 22 – Resultado da pesquisa - Questão 2

Na Figura 22, mostrou um resultado bem otimista. Todos os entrevistados já ouviram falar do Código de Defesa do Consumidor. Este resultado indica que a divulgação dos direitos do consumidor pela mídia ou qualquer meio de divulgação tem tido bom alcance. Em seguida (Figure 23), relacionado a pergunta anterior, as estatísticas mostraram que desses 100% apenas um pouco mais da metade tem conhecimento da lei 8078. Isto é, apesar dos cidadãos saberem que seus direitos existem, pouco mais da metade declarou estar informado sobre o que se tratam esses direitos. Esse resultado fortifica utilidade e propósito desse sistema: trazer as informações para o cidadão de maneira filtrada e relacionada com a queixa, sem precisar que o usuário consulte diretamente o Código do Consumidor. O acesso a informação legal deve atingir maiores números de cidadãos.

Dos participantes que responderam “Não” para a questão 3 foi questionado o porquê da decisão. Majoritariamente (77,8%) os entrevistados responderam que encontraram dificuldades seja por causa da leitura rebuscada ou pela grande quantidade de informações contidas na lei 8078 (Figura 24). Somente 2 de 22 pessoas, votaram nas opções restantes.

Na questão 5 (Figura 25), mostra que em sua maioria os entrevistados tiveram contato pela primeira vez à lei através de propagandas de TV ou de site de terceiros pela internet. Somente 25% tiveram acesso direto ao conteúdo seja pela escola, faculdade ou pelo site do governo federal onde está documentada a lei 8078. Um resultado positivo dessa estatística no entanto é que 25% dos entrevistados foram informados da lei através das lojas (físicas ou online). Isto mostra o interesse dos empresários em mostrar os direitos ao consumidor e evitar possíveis transtornos com a lei.

Na figura 26, apresenta uma visão otimista dos entrevistados. Um pouco mais da

3) Você já procurou se informar sobre essa lei para saber em que situações possa lhe proteger em casos de desigualdade ou descumprimento de responsabilidade nos contrato de serviços ou compra de produtos?

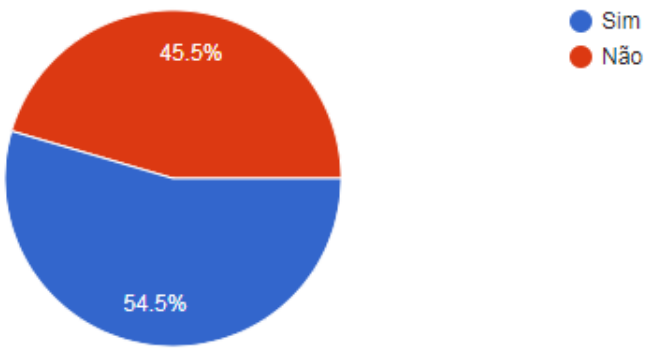

Figura 23 – Resultado da pesquisa - Questão 3

4) Qual motivo te levou a tomar essa decisão?

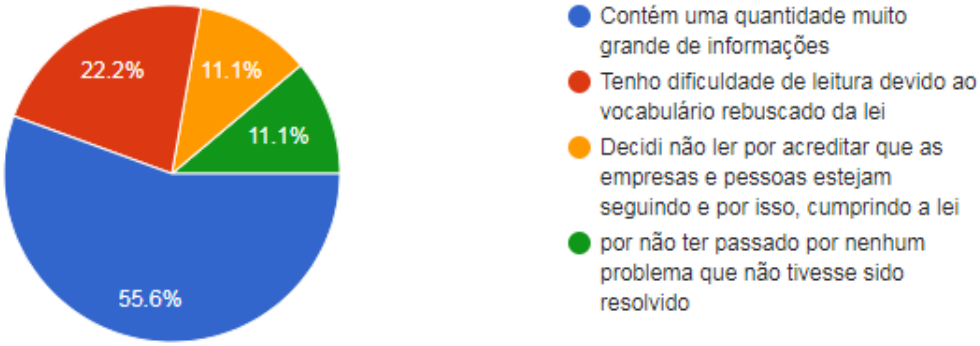

Figura 24 – Resultado da pesquisa - Questão 4 (Condicional)

metade dos entrevistados declararam estar satisfeitos com as negociações realizadas, 8 pessoas declararam neutras (Razoáveis) e apenas 2 entrevistados disseram estar insatisfeitos com as experiências no mercado de consumo. Nenhum entrevistado votou que as negociações realizadas são de excelência ou péssimas, respectivamente, representadas por melhor e pior qualificação das negociações realizadas pelo consumidor.

No gráfico de colunas representado pela Figura 27, os resultados da questão 7 mostraram também um dado otimista desses entrevistados. Maioria está satisfeita com atendimento prestado pela empresas ao tentaram solucionar um problema que tiveram. E representando uma parcela pequena, 7 pessoas insatisfeitas com o atendimento. Por fim, 1 pessoa não conseguiu resolver seu problema contactando o fornecedor.

Apesar de saber que o comercio brasileiro esteja atuando de forma justa e legal, esperava-se do grupo entrevistado bastantes respostas negativas em relação a questão

5) Onde você encontrou ou ouviu falar dessa lei ?

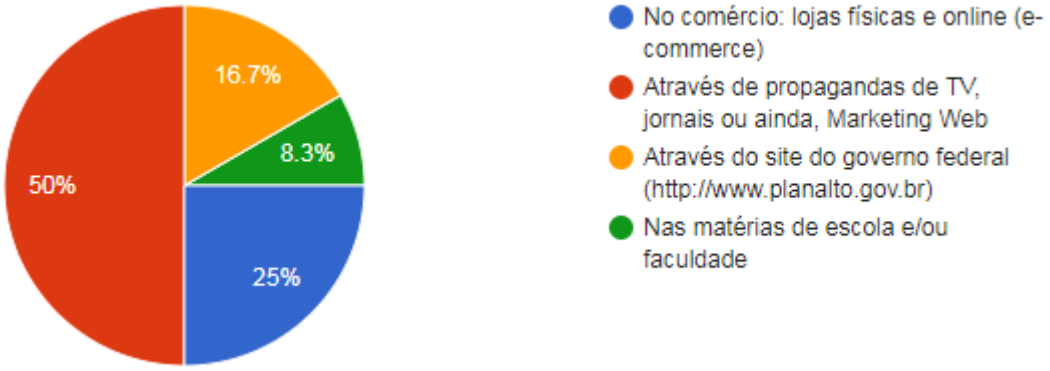

Figura 25 – Resultado da pesquisa - Questão 5

6) De todas as compras de produtos ou serviços, de forma geral, como você qualifica as negociações realizadas? (compras, serviços e contratos)

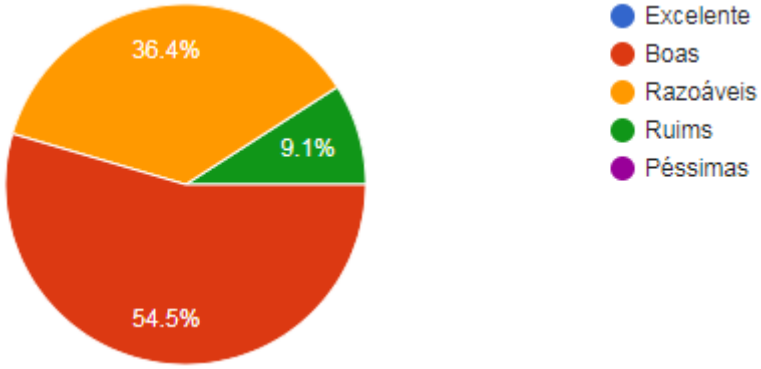

Figura 26 – Resultado da pesquisa - Questão 6

7. A exemplo, em 2017 no site ReclameAqui circulou-se bastantes notícias sobre venda de serviços e produto com disparidade de informações na oferta. No mesmo site , as operadoras de celular deixaram de atender uma quantidade significativa de clientes com problemas em seus serviços[2].

[2] Cidadãos Insatisfeitos com o serviços prestados pelas operadores de telefone <https://noticias.reclameaqui.com.br/noticias/operadoras-de-celular-deixaram-de-responder-223-mil-reclamac_3101/>

7) Caso você tenha sido desfavorecido na compra de um produto ou serviço, de forma geral, como você qualifica o atendimento prestado pelos fornecedores na tentativa de solucionar o seu problema?)

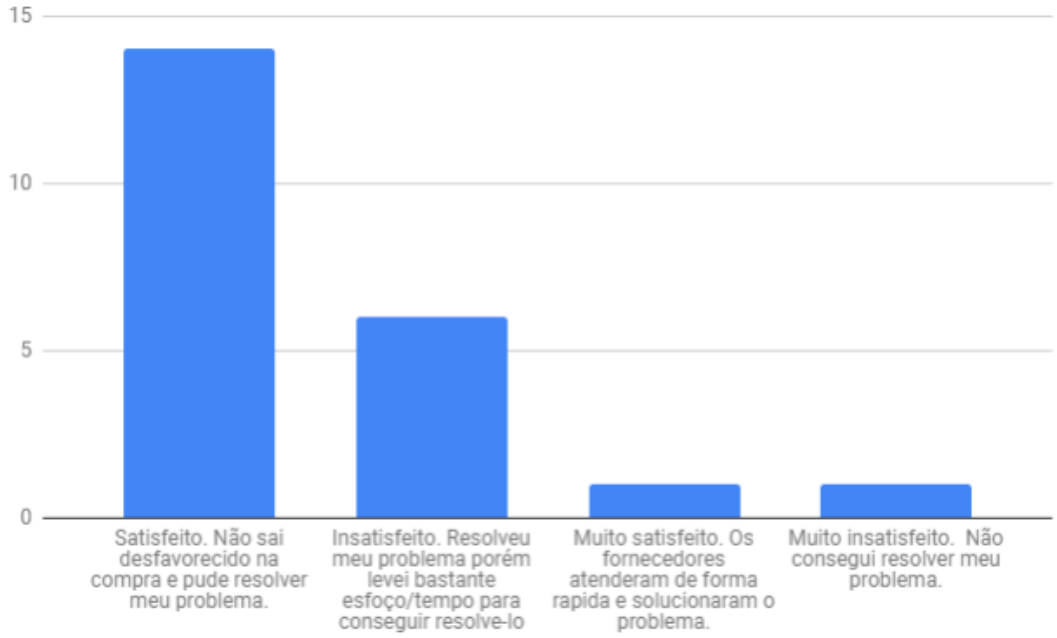

Figura 27 – Resultado da pesquisa - Questão 7

Na questão 8, foi dado espaço para o participante escolher a opção que mais se encaixava a sua situação, incluindo escrever uma nova alternativa como resposta. Como foram encontrado varias respostas, algumas com mesma semântica e palavras diferentes, foram agrupadas em uma única opção como visto na Figura 28.

Apesar do PROCON atender em todo território nacional, normalmente estão presentes nas capitais deixando alguns municípios bem afastados e de difícil acesso como dito por um dos participantes. Uma parcela muito pequena (9%) declarou ter tentado o acesso ao atendimento mas não obteve sucesso devido a alta burocracia ou ao tempo duradouro de espera por atendimento. A resposta encontrada na questão 8 já era esperada. como visto nos comentários e sugestões da aplicação do *Google Maps* os usuários reclamam da falta de um atendimento de qualidade pelo PROCON[3].

Maior parte dos entrevistados (45.4%) declararam desconhecer o atendimento ou não saber como ter acesso ao serviço prestado pelo meio público confirmando a desinformação dos usuários quanto ao benefício. 6 pessoas (27.3%) declararam que o atendimento prestado é eficiente e funciona.

[3] Cidadãos reclamam do mal atendimento prestado pelo PROCON <https://goo.gl/6iZkJG>

8) Quando o contato direto com o fornecedor não resolve a reclamação do consumidor, o cidadão tem o direito a um advogado público disponíveis nas capitais do país nas devidas Delegacias de Defesa ao Consumidor (PROCON). De acordo com as opções abaixo, qual a que você mais se identifica?

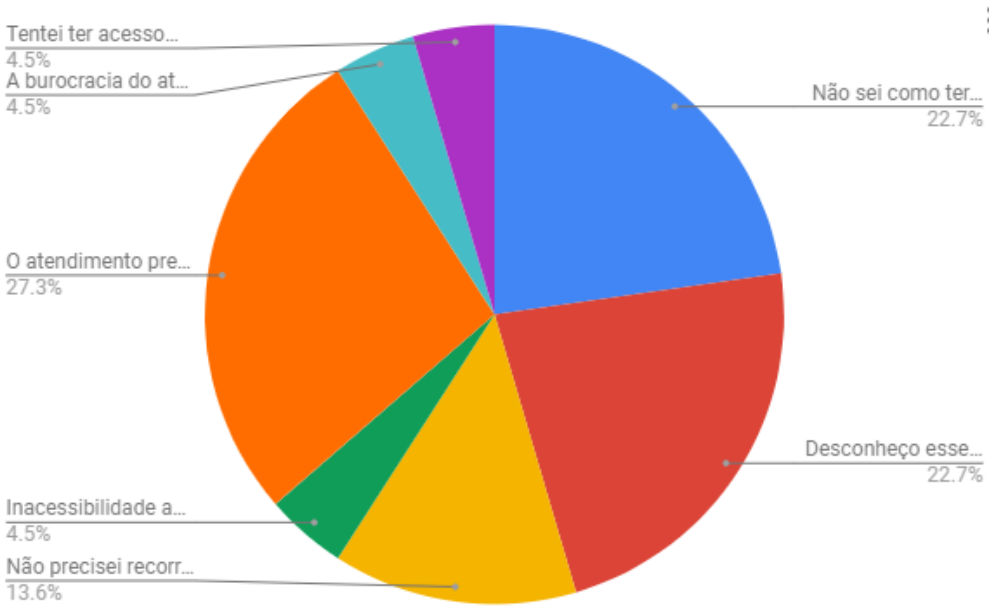

Figura 28 – Resultado da pesquisa - Questão 8

5.2.2 Pesquisa sobre a plataforma

Objetivo dessa pesquisa foi verificar se a plataforma estaria atingindo o objetivo pelo qual ela foi proposta. Para testar também se a interface do sistema estaria acessível e intuitiva, os participantes não tiveram nenhum guia sobre como usar a plataforma. Através dos entrevistados, foi possível identificar falhas e encontrar melhorias para o sistema.

9) Você conseguiu enviar sua queixa para o sistema identificar um artigo legal relacionado?

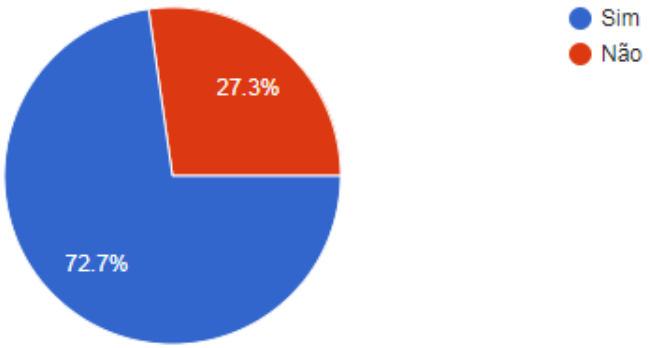

Figura 29 – Resultado da pesquisa - Questão 9

Na questão 9, foi verificado se uma tarefa básica do sistema estava sendo atendida:

usuário ser capaz de enviar sua queixa para a plataforma analisá-la. Através da Figura 29, pode-se verificar que nem todos os usuários conseguiram enviar sua queixa. 6 pessoas (27.3%) declararam não ter conseguido enviar sua queixa. Ao verificar no banco de dados se elas realmente não haviam enviado a queixa, verificou-se que dessas declarações, 2 dessas pessoas na verdade conseguiram enviar a queixa para o sistema apesar de terem informado o contrário.

Ao questionar as pessoas que responderam "Não", elas disseram que achavam que a pergunta referia-se ao sistema do PROCON já que as perguntas anteriores se tratavam do PROCON. 4 das 6 pessoas responderam que o chatbot não conseguiu encontrar o artigo relacionado e por isso finalizou o experimento sem enviar a queixa. Logo, somente 4 das 22 pessoas (18.2%) não conseguiram enviar a queixa. Esse resultado corresponde a um resultado positivo para o sistema em estudo, pois a maioria dos entrevistados conseguiu utilizar a ferramenta.

10) O artigo informado pelo sistema estava relacionado ao seu problema?

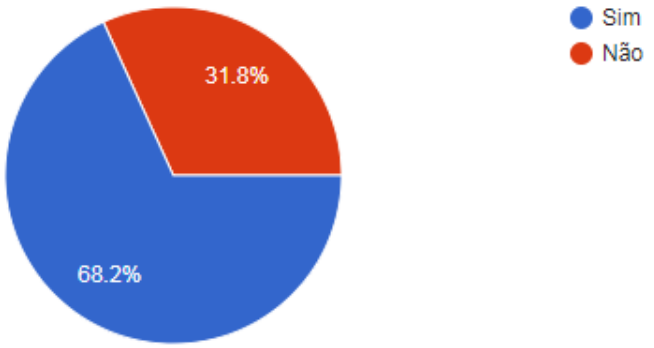

Figura 30 – Resultado da pesquisa - Questão 10

Na questão 10, 15 das 22 pessoas (68.2%) declararam que o artigo encontrado estava relacionado com a queixa. 7 pessoas responderam que artigo informado não estava relacionado (Figura 30). Dos que reportaram que os artigos não estavam relacionados, 2 disseram que o artigo continha informação relevante mas também informações não relacionadas a queixa. A exemplo, um usuário reportou que enviaram a ele uma cobrança acima do preço estipulado. O sistema encontrou o artigo 42 como o artigo mais relacionado com sua queixa que também cita sobre exposição ao ridículo. O usuário informou que o fornecedor não o expôs ao ridículo mas que o artigo estava relacionado com sua queixa por tratar-se de cobranças inadequadas.

Parágrafo Único da Seção V, artigo 42 [4]:

> Parágrafo único. O consumidor cobrado em quantia indevida tem direito à repetição do indébito, por valor igual ao dobro do que pagou em

[4] Código de Defesa do Consumidor <http://www.planalto.gov.br/ccivil_03/Leis/l8078.htm>

excesso, acrescido de correção monetária e juros legais, salvo hipótese de engano justificável.

Em seguida foi verificado se as perguntas expostas pelo chatbot tinham relevância com a queixa digitada (Figura 31). As respostas encontradas da questão 11 foram positivas: 16 participantes disseram que as questões expostas ali tinham relação com a queixa, valor acima da média. E 6 pessoas responderam que as perguntas encontradas ali não tinham haver com a queixa digitada.

11) Das perguntas informadas pelo chatbot, você encontrou uma relacionada com sua queixa?

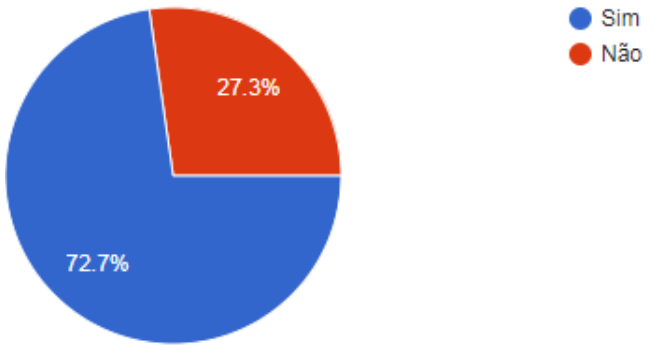

Figura 31 – Resultado da pesquisa - Questão 11

Uma importante funcionalidade que foi preciso validar é da listagem de queixas similares onde usava o algoritmo de comparação de palavras chave da queixa com as queixas registradas no sistema. Nessa questão, a grande maioria dos participantes responderam que todas ou pelo menos uma das queixas listas ali tinha relação com a queixa que digitou correspondendo a 77.3% do percentual de participantes. Apenas 5 pessoas não conseguiram encontrar alguma informação na lista de queixas similares. Os motivos foram: lista de queixas similares vazia, não conseguiu visualizar a lista ou ainda as queixas cadastradas não tinham relevância com a queixa digitada (Figura 32).

12) Com relação a lista de queixas similares disponibilizadas após encontrar o artigo, escolha uma opção:

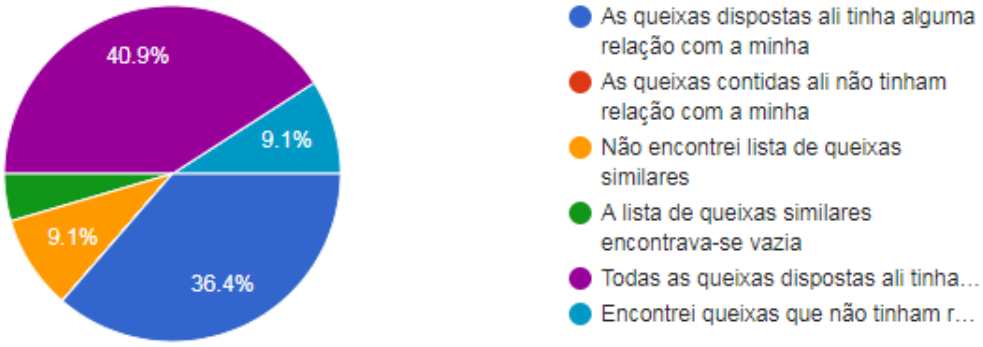

Figura 32 – Resultado da pesquisa - Questão 12

Nas pergunta 13 e 14, que estão intimamente relacionadas, foi perguntado se o usuário estava satisfeito com o sistema utilizado. O resultado encontrado também foi muito bom. Mostrando que o usuário acredita no potencial da ferramenta. Os participantes afirmaram majoritariamente que recomendariam o uso do site *MeusDireitoConsumidor.com.br* para um amigo. No entanto 3 pessoas correspondendo 13.6% não conseguiram encontrar a uma informação legal relacionada a queixa e por isso não recomendam o sistema (Figura 33).

13) Você recomendaria a um amigo que utilizasse esse sistema para identificar quais são seus direitos de consumidor baseado em suas queixas?

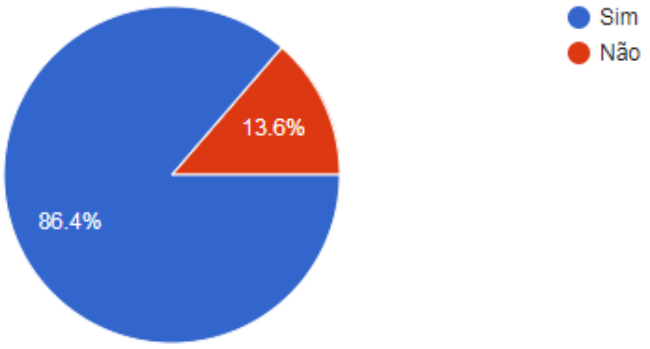

Figura 33 – Resultado da pesquisa - Questão 13

Na Figura 34 confirmando o resultado anterior encontrado, 19 pessoas deram nota mediana ou acima da média com 10 pessoas nota máxima acerca do chatbot utilizado. Sabe-se que foi encontrado perguntas do chatbot que puderiam ser melhor filtradas para atingir um melhor percentual de acerto. No entanto, esse resultado constitue um valor importante para ferramenta por ter atingido o percentual acima de 80% de satisfação dos usuários.

14) Com uma nota de 1 a 5, como você avalia o chatbot utilizado?

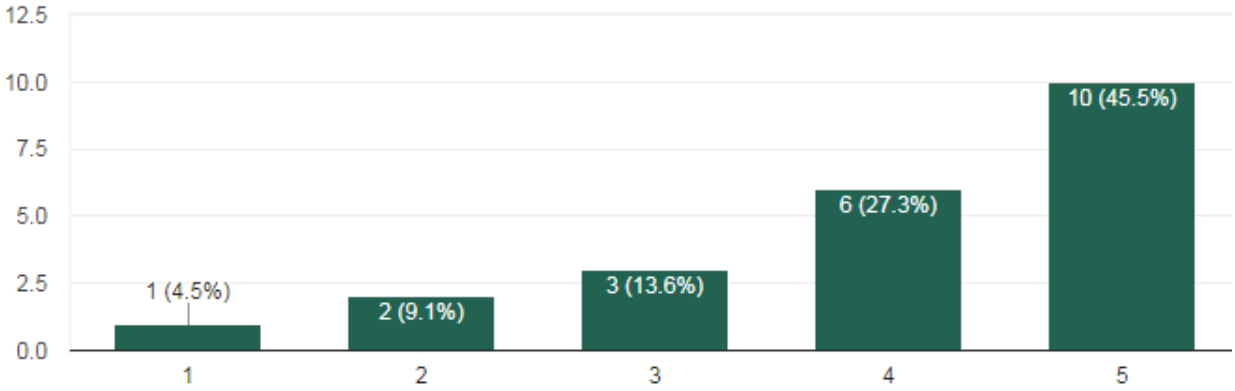

Figura 34 – Resultado da pesquisa - Questão 14

5.3 Discussão

Se comparar os resultados deste experimento com os resultados da pesquisa realizada pelo Idec em 2016 (IDEC, 2016), pode-se afirmar que o perfil dos participantes das duas pesquisas são muito similares. A maioria dos usuários nas duas pesquisas declararam ter ouvido falar do CDC porém um percentual grande declarou não ter consultados seus direitos na lei. O experimento mostrou que, ainda que mais de 50% dos entrevistados possuem graduação completa ou em andamento, muitos tem dificuldade para encontrar seus direitos na lei seja devido a grande quantidade de texto do CDC ou pela sua dificuldade de seu vocabulário.

Quanto às questões referente ao uso da plataforma criada, foi visto que há ainda trabalho a ser feito. Uma pequena parcela não conseguiu enviar a queixa para a plataforma buscar pelo artigo relacionado. Foi pedido por uma pequena parcela de usuários do experimento para o sistema especificar exatamente qual artigo estaria relacionado com a queixa já que no resultado encontrado pelo sistema podem aparecer mais de um artigo.

Dos usuários que enviaram a queixa, o resultado mostrado pelos gráficos pizza e pelos comentários do sistema mostram uma aprovação da plataforma de maneira majoritária. No entanto, dos usuários que não conseguiram enviar a queixa, o sistema não conseguiu recuperar artigos relevantes por não encontrar palavras semelhantes entre a queixa do usuário e os artigos legais.

Esse é um problema esperado pela heurística utilizada na recuperação de artigos. Uma abordagem para tentar mitigar o problema seria através do registro da queixa em questão no banco de dados para posterior análise de quais palavras chave não foram relacionadas ao artigo correto do CDC. Essas palavras poderão ser adicionadas ao vetor de sinônimos da aplicação caso sejam constatadas que são de fato palavras chaves sinônimas

de artigos do CDC.

6 Conclusão

6.1 Considerações Finais

Observou-se que através das queixas cadastradas, o sistema pôde encontrar em sua maioria artigos satisfatórios. Com os exemplos extraídos do *ReclameAqui* foi possível testar inicialmente o sistema e realizar melhorias na identificação de palavras chave induzindo o motor de busca a entregar resultados mais relevantes. Após ter testado o sistema com as queixas do *ReclameAqui*, decidiu-se realizar um experimento aberto para validar o funcionamento do sistema. Havia uma necessidade de confirmar que outras pessoas chegassem a mesma conclusão sobre o uso da ferramenta. Com o experimento realizado, foi verificado um percentual de satisfação acima de 80% dos entrevistados.

Através da plataforma online *MeusDireitoConsumidor.com.br*, mais de 70 queixas únicas foram cadastradas contendo exemplos retirados da plataforma *ReclameAqui* e dos participantes do experimento. Com isso mais de 70 artigos foram recuperados e registrados com as queixas dos usuários. Todas as queixas podem ser visualizadas através da área de login do sistema online[1].

Foram também encontrados durante os testes, queixas duplicadas ou muitos similares porém estas não foram adicionadas ao banco de dados do sistema devido a seguinte funcionalidade implementada: “Identificação de queixas similares”. Através dessa funcionalidade foi possível evitar o registro de queixas consideradas idênticas.

O sistema confirmou sua utilidade na identificação de artigos legais do código de defesa do consumidor relevantes a queixa do usuário. No entanto, devido a falta de informações a respeito de sinônimos ou expressões similares contido no sistema, a taxa de acerto do artigo legal relacionado pode diminuir.

Com o uso da plataforma e o aumento de queixas cadastradas foi possível analisar os resultados gerados pelo sistema. Permitiu-se identificar as palavras chaves contidas nas queixas que relacionavam com os artigos corretamente. Nos artigos que não foram encontrados corretamente, a causa identificada foi que haviam palavras contidas na queixa descrita do usuários não reconhecidas pelo sistema. Isto é, essas palavras não possuíam ocorrência nos artigos da lei. No entanto, pôde verificar-se que ao reconhecer as novas palavras chave da queixa como sinônimo das palavras presentes no artigo legal, a resposta poderia ser encontrada pelo motor de busca e entregue ao usuário corretamente.

O experimento realizado permitiu identificar as falhas do sistema e comemorar

[1] *MeusDireitosConsumidor.com.br*

acertos, mostrando que o sistema fornece um bom suporte na conscientização do consumidor e na divulgação de seus direitos. Dito isso, o trabalho atingiu seu objetivo criando uma plataforma gratuita e transparente para os consumidores terem acesso a informação legal baseado no seu problema escrito. Ao fim do processo de recuperação do artigo, o usuário tem acesso a um folheto em formato PDF com sugestões para ajuizamento da queixa no PROCON da sua cidade ou no SAC mais próximo do seu município. (Apêndice A)

Com a satisfação dos participantes deste experimento, este estudo confirma que a abordagem escolhida tem grande potencial para evolução. Para isso é necessário aumentar o número de acertos na busca de artigos relevantes e aprimorar a interface melhorando experiência do usuário na plataforma.

6.2 Trabalhos Futuros

6.2.1 Adição de Sinônimos

Apesar de a API *thesaurus.altervista* ser um ótimo candidato para busca de sinônimos, ela não mostra casos de palavras ou expressões que tenham significado distinto mas que possuam valor semântico relevante. A Exemplo tem-se as palavras 'restituição' e 'devolução'. As duas palavras são sinônimos encontrados pelo *thesaurus.altervista.org/* porém as expressões 'retorno' e 'de volta' não são encontradas pelo serviço e fazem parte geralmente dos textos escritos pelos usuários. Essas palavras podem sugerir a mesma ideia de reembolso. Somente através dos testes cadastrados no sistema foi possível verificar essa perspectiva já que seria impossível ter acesso a todos os sinônimos possíveis de uma expressão ou palavra ao criar a plataforma.

Após verificar que "retorno" e "volta" são palavras que estavam presentes nos artigos referentes a restituição. Foi necessário adicionar essas palavras a um vetor local de sinônimos. Caso o usuário digite uma da queixa contendo palavras desse vetor, o sistema deve considerar todas as palavras sinônimas para busca no Elastic Search. Porém mapear todos os possíveis sinônimos é uma tarefa que irá ser gradualmente preenchida com o uso da plataforma pelos usuarios.

A medida que a quantidade de dados for aumentando e análise dos dados forem feitas gradualmente, o número de sinônimos cresce e com isso melhor a identificação de palavras chave das queixas. Uma outra melhoria para identificação de palavras-chave, é fazer com que o sistema reconheça as palavras baseado em seu radical, isto é aplicação do 'stemming'. Seria suficiente identificar a palavras através do radical (stem) sem se preocupar com sua conjugação variada pelo tempo, pessoa ou número.

6.2.2 Erros de Digitação

Um dos testes que consolidou-se um desafio no início construção da aplicação foi o tratamento de caracteres especiais com posicionamento indevido. Por exemplo, foi encontrado queixas do ReclameAqui que haviam pontuações ligando as palavras:

"... cobraram de mim uma taxa que eu não solicitei.isso é um ABSURDO!!!quero meu dinheiro.R$15,05 de volta"

Para lidar com problemas assim foram criados expressões regulares que identificassem os caracteres especiais e não comprometesse o conteúdo da frase. No caso acima era comum encontrar erros após remoção das pontuações gerando palavras muito próximas como 'absurdoquero' ou então expressões como "quero meu dinheiro.R$15,05" tornar-se em "quero, dinheiro R". Ao processar as queixa era removido pontuações e cifrões mas sobrava a letra R que não se aplicava regra alguma da expressão regular. Foi necessário incluir filtros adicionais que removem vogais e consoantes de tamanho um para evitar as situações listadas.

Outro erro comum que não foi levado em conta durante este estudo é erro de digitação quanto a grafia. Levou-se em conta que os usuário iria digitar queixas com grafia coerente e correta. E por isso não foi aplicado nenhum algoritmo de correção automática ou identificação de erros de grafia como o de *Levenstein* ou *Edit Distance*, que podem ser adicionados no futuro para completude do sistema. Outra alternativa é a utilização de um biblioteca que sugira correção ortográfica das queixas do usuário durante a escrita.

6.2.3 Distinção entre Produto e Serviço

Todas as perguntas geradas para o usuário através do chatbot são baseadas nos artigos resultantes da consulta do motor de busca. O resultado contém diversos artigos identificados apenas pelo seu número. As ordens das perguntas geradas para o usuário também segue a ordem dos artigos encontrados pelo motor de busca. Esses artigos levam em consideração apenas a presença ou não das palavras chaves da queixa nos artigos indexados. Logo podem aparecer artigos relacionados a contratos, consórcios, produtos ou serviços.

A limitação encontrada é que se o usuário escreve uma queixa sobre um produto que ele adquiriu e o sistema, no entanto, pode mostrar perguntas no chatbot relacionados a serviço por exemplo. Isto é, perguntas não relacionadas com a queixa do usuário podem aparecer devido a similaridade de algumas palavras chave da queixa no artigo legal. O sistema deveria informar artigos focados apenas em assuntos relacionados a queixa. Isso acontece porque o motor de busca encontrou nas palavras da queixa semelhanças com outros artigos envolvendo contrato ou serviço.

Para futuras implementações, uma forma sútil de lidar com esse problema é pedir ao usuário que informe se sua queixa é sobre produto, serviço ou contrato. Outra alternativa, é separar algumas palavras ou expressões que seriam bons indicadores para o sistema classificar a queixa (Reconhecimento de Padrões Textuais).

6.2.4 Aprimorar o Chatbot

Atualmente, o chatbot cumpre um função fundamental na recuperação do artigo legal relacionado à queixa do usuário. Através dele, em conjunto com o usuário, o sistema é capaz de filtrar e encontrar o artigo ideal que pode endossar a queixa do usuário na solução de seu problema. Isto é, após receber a lista dos prováveis artigos relevantes à queixa, o sistema consegue selecionar um artigo dessa lista de acordo com a resposta dada pelo usuário.

Durante o experimento, foi sugerido a ideia de utilizar o chatbot como um agente virtual. Dessa maneira o chatbot realizaria uma sequência de perguntas gerais ou específicas, baseando-se nas respostas do usuário até chegar ao problema do cidadão e informar o artigo legal. O usuário dialogaria com o sistema contribuindo com mais informações e não apenas participando com respostas diretas e simples. O chatbot seria capaz também de responder perguntas que não se encontram nos textos legais do Código de Defesa do Consumidor mas que estão relacionadas ao tema da área.

A exemplo, segue algumas perguntas que seriam de grande ajuda para o usuário na solução de seu problema caso fossem respondidas: “Qual o site oficial do PROCON de Salvador-Ba?”, “Poderia indicar algum advogado online para ajudar na solução do meu problema? Estou ciente dos meus direitos e gostaria de prosseguir com a causa.”

6.2.5 Sumário

- Implementar a técnica de *Stemming* no processamento da queixa do usuário para facilitar na identificação de palavras chave pelo motor de busca.
- Implementar sugestão de correção ortográfica do texto da queixa do usuário.
- Expandir o sistema para área de aplicativos móveis com registro de queixas por voz.
- Reconhecimento e classificação de queixas. A plataforma online poderia identificar se a queixa esta relacionada com produto ou serviço. Isso faria com que o chatbot filtrasse perguntas não relevantes baseado na identificação da queixa.
- Aprimorar o contato entre o chatbot e usuário da plataforma. Permitir que o sistema possa extrair mais informações além da queixa do usuário para recuperar artigos legais do CDC.

- Aprimorar interface da ferramenta a fim de fornecer melhor suporte ao usuário. Além disso, adicionar meios para prosseguir na solução de seu problema como indicação de um advogado especializado na área de direito do consumidor ou indicação de um contato do PROCON mais próximo da cidade do usuário.

Referências

APACHE. *TFIDFSimilarity*. 2010. Disponível em: <https://lucene.apache.org/core/5_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html>. Acesso em: 25 julho 2018. Citado 2 vezes nas páginas 29 e 30.

BAEZA-YATES, B. R.-N. R. *Modern Information Retrieval*. 2nd. ed. [S.l.]: Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA, 1999. ISBN 020139829X. Citado na página 24.

BAEZA-YATES, B. R.-N. R. *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd. ed. [S.l.]: Addison-Wesley Professional, 2012. ISBN 978-0321416919, 0321416910. Citado na página 22.

BIRD, S.; KLEIN, E.; LOPER, E. *Natural Language Processing with Python*. 1st. ed. O'Reilly Media, Inc., 2009. ISBN 0596516495, 9780596516499. Disponível em: <http://www.nltk.org/book/ch00.html>. Citado na página 21.

BRASIL. *LEI Nº 8.078, DE 11 DE SETEMBRO DE 1990*: Código de defesa do consumidor. 1990. Disponível em: <http://www.planalto.gov.br/ccivil_03/leis/l8078.htm>. Acesso em: 23 de março de 2018. Citado na página 60.

BRASIL, P. *Reclamações de consumidores chegam a 2,7 milhões em 2016*. 2017. Disponível em: <https://goo.gl/ymHPmL>. Acesso em: 22 de julho de 2018. Citado na página 17.

CHOWDHURY, G. Natural language processing. *Annual Review of Information Science and Technology*, v. 37, p. 51–89, 2003. Citado na página 32.

COMUNITIES, W. *What is HTML*. 2011. Disponível em: <https://www.w3.org/wiki/Html/Training/What_is_HTML>. Acesso em: 25 julho 2018. Citado na página 39.

CROFT DONALD METZLER, T. S. W. B. *Search Engines, Information Retrieval in Practice*. Pearson Education, Inc., 2015. Disponível em: <http://ciir.cs.umass.edu/downloads/SEIRiP.pdf>. Citado na página 30.

DONOTPAY. *Donotpay*. 2018. Disponível em: <https://www.donotpay.com/>. Acesso em: 12 de julho de 2018. Citado na página 47.

DYER, R. J. *MySQL in a Nutshell*. 2nd. ed. [S.l.]: O'Reilly, 2008. Citado na página 41.

ELASTIC. *Basic Concepts*. 2017. Disponível em: <https://www.elastic.co/guide/en/elasticsearch/reference/6.2/_basic_concepts.html>. Acesso em: 24 de abril de 2018. Citado na página 31.

ELASTIC. *Elasticsearch Clients*. 2018. Disponível em: <https://www.elastic.co/guide/en/elasticsearch/client/index.html>. Acesso em: 18 de julho de 2018. Citado na página 31.

ESTADÃO. *Reclamações aumentam na Black Friday*. 2017. Disponível em: <https://economia.estadao.com.br/noticias/geral,black-friday-2017-ja-tem-mais-reclamacoes-em-relacao-ao-ano-passado,70002095995>. Acesso em: 30 de julho de 2018. Citado na página 17.

FERNEDA, E. *Recuperação de Informação: Análise sobre a contribuição da Ciência da Computação para a Ciência da Informação*. Tese (Doutorado), São Paulo, Brasil, 2003. Disponível em: <www.teses.usp.br/teses/disponiveis/27/27143/tde-15032004-130230/publico/Tese.pdf>. Acesso em: 13 de julho de 2017. Citado 2 vezes nas páginas 22 e 28.

FOUNDATION, A. *"Apache Lucene Core"*. 2016. Disponível em: <https://lucene.apache.org/core/>. Acesso em: 18 de julho de 2018. Citado na página 32.

HEROKU. *What is Heroku?* 2018. Disponível em: <https://www.heroku.com/what>. Acesso em: 09 de julho de 2018. Citado na página 43.

HIEMSTRA, D. Information retrieval models. 2009. Disponível em: <http://wwwhome.cs.utwente.nl/~hiemstra/papers/IRModelsTutorial-draft.pdf>. Citado 3 vezes nas páginas 23, 24 e 26.

IDEC. *Brasileiros conhecem seus direitos de consumidor, mas não reclamam de forma efetiva*. 2016. Disponível em: <https://goo.gl/hgEVTm>. Acesso em: 09 de julho de 2018. Citado 2 vezes nas páginas 18 e 75.

JOHNMANNES. *DoNotPay launches 1,000 new bots to help you with your legal problems*. 2017. Disponível em: <https://techcrunch.com/2017/07/12/donotpay-launches-1000-new-bots-to-help-you-with-your-legal-problems/>. Acesso em: 23 de março de 2018. Citado na página 47.

KOZIOLEK, H. The role of experimentation in software engineering. *Seminar "Research Methods", Summer Term 2005*, 2005. Disponível em: <https://pdfs.semanticscholar.org/7bf6/75bca6641dcec2353e6c1e6cea6e4f9d5ad8.pdf>. Citado na página 65.

LASHKARI FERESHTEH MAHDAVI, V. G. A. H. A boolean model in information retrieval for search engines. *International Conference on Information Management and Engineering*, IEEE, 2009. Citado na página 25.

LUCENEWIKI. *"PoweredBy - Lucene-java"*. 2015. Disponível em: <https://wiki.apache.org/lucene-java/PoweredBy>. Acesso em: 18 de julho de 2018. Citado na página 32.

MANNING, P. R. . H. S. C. D. *Introduction to Information Retrieval*. 1st. ed. Cambridge University Press, 2009. Disponível em: <https://nlp.stanford.edu/IR-book/>. Citado 8 vezes nas páginas 23, 24, 26, 27, 29, 30, 34 e 37.

MATOS, O. *Como o Jusbrasil funciona*. 2016. Disponível em: <https://medium.com/@tupydabahia/como-o-jusbrasil-funciona-4303f2b1d356>. Acesso em: 02 de Agosto de 2018. Citado na página 48.

MOOBA. *Cupons de Desconto e Dinheiro de Volta | Mooba*. 2018. Disponível em: <https://www.mooba.com.br/>. Acesso em: 12 de julho de 2018. Citado na página 46.

MOZILLA. *"Regular Expressions - Javascript"*. 2018. Disponível em: <https://developer.mozilla.org/en-US/docs/Web/JavaScript/Guide/Regular_Expressions>. Acesso em: 22 de julho de 2018. Citado na página 36.

MOZILLA. *What is JavaScript?* 2018. Disponível em: <https://developer.mozilla.org/en-US/docs/Learn/JavaScript/First_steps/What_is_JavaScript>. Acesso em: 8 de julho de 2018. Citado na página 42.

O'REILLY, T. *What Is Web 2.0.* 2005. Disponível em: <http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>. Acesso em: 15 de julho de 2018. Citado na página 21.

POUNCEY, I.; YORK, R. *Beginning CSS: Cascading Style Sheets for Web Design.* 3rd. ed. Wrox Press Ltd., 2011. Disponível em: <http://www.nltk.org/book/ch00.html>. Citado na página 40.

RECLAMEAQUI. *Reclame Aqui - Consumidores exponham suas reclamações.* 2018. Disponível em: <https://www.reclameaqui.com.br/>. Acesso em: 02 de fevereiro de 2018. Citado na página 45.

RUSSELL, P. N. S. *Artificial Intelligence - a modern approach.* [S.l.]: In Prentice-Hall., 1995. Citado na página 33.

SALTON, C. B. G. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, Pergamon Press, v. 24, n. 5, p. 513–523, 1988. Citado na página 27.

SINDEC. *Boletim Sindec.* 2017. Disponível em: <http://www.justica.gov.br/news/crescem-reclamacoes-contra-bancos-cartoes-de-credito-energia-e-saneamento/boletim_sindec_2017.pdf>. Acesso em: 01 agosto 2018. Citado na página 17.

SOMMERVILLE, I. *Software Engineering.* 9. ed. Harlow, England: Addison-Wesley, 2010. ISBN 978-0-13-703515-1. Citado na página 50.

SUEHRING, S. *Javascript Step by step.* 3rd. ed. [S.l.]: Microsoft Press, 2013. ISBN 978-0735665934, 0735665931. Citado na página 42.

SVENONIUS, E. *The Intellectual Foundation of Information Organization.* 1st. ed. [S.l.]: Digital Libraries and Electronic Publishing MIT Press, 2009. Citado na página 24.

YOU, E. *Introduction - What is Vue.js?* 2018. Disponível em: <https://vuejs.org/v2/guide/index.html>. Acesso em: 08 de julho de 2018. Citado na página 43.

ZOBEL, J.; MOFFAT, A. Inverted files for text search engines. *ACM Computing Surveys, Article 6*, v. 38, n. 2, 2006. Citado 3 vezes nas páginas 28, 29 e 31.

Apêndices

APÊNDICE A – Exemplo de Relatório em formato PDF gerado pelo sistema

MeusDireitosConsumidor.com.br			
Nome:	Diego Novaes	**CPF:**	000.000.000-00
Email:	novaes@hotmail.com	**Telefone:**	(00) 000000000
CEP:	00000000	**Logradouro:**	Rua Pacífico Nogreira
Numero:	100	**Complemento:**	
Bairro:	Centro	**Cidade-UF:**	Salvador-BA

Queixa descrita pelo cidadão:

Não recebo franquia dados de internet e bônus de meu plano há 2 semanas. Desde que fiz adesão ao plano, tenho mantido o saldo suficiente para renovação semanal do plano, porém , a empresa TIM não tem cumprido com o que é previsto no plano o qual tenho contrato junto a empresa, tendo tido à minha disposição apenas da franquia de voz, e esse tempo todo apenas utilizando a internet através do WIFI. Acho extremamente desreipeitoso e outras coisas mais por parte da empresa. Como usuário, de forma direta, desde o início da adesão ao plano, estou usufruindo apenas de transtornos.

O Artigo que pode endossar seu caso é descrito abaixo:

Recusa a Cumprimento da Oferta

Art. 35. Se o fornecedor de produtos ou serviços recusar cumprimento à oferta, apresentação ou publicidade, o consumidor poderá, alternativamente e à sua livre escolha:

I - exigir o cumprimento forçado da obrigação, nos termos da oferta, apresentação ou publicidade;

II - aceitar outro produto ou prestação de serviço equivalente;

III - rescindir o contrato, com direito à restituição de quantia eventualmente antecipada, monetariamente atualizada, e a perdas e danos.

Informações Adicionais:

Esse sistema te auxilia a encontrar informações legais baseado na queixa digitada porém não tem caráter decisório.

Para ajuizar a queixa ou iniciar um processo para solucionar seu problema entre em contato com o Procom responsável do seu município através do link abaixo:

http://www.portaldoconsumidor.gov.br/procon.asp?acao=buscar

Ou ainda entre em contato com o Serviço de Atendimento ao Consumidor (SAC) do seu munícipio.

No exemplo abaixo temos o link para agendamento de queixas em Salvador-Ba:

http://www.sac.ba.gov.br/index.php/Servicos-de-parceiros/Servicos-do-TJ-BA-%E2%80%93-Tribunal-de-Justica-da-Bahia/SAJ-%E2%80%93-Servico-de-Atendimento-Judiciario.html