UNIVERSIDADE FUMEC
FACULDADE DE CIÊNCIAS EMPRESARIAIS
Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento

WENDEL VILAÇA DE ASSIS

CHAT BOT SUMÉ
WEB SCRAPING EM DADOS GOVERNAMENTAIS PARA CONSULTA DE GASTOS PÚBLICOS DOS VEREADORES DA CÂMARA MUNICIPAL DE BELO HORIZONTE

BELO HORIZONTE
2021

WENDEL VILAÇA DE ASSIS

CHAT BOT SUMÉ
WEB SCRAPING EM DADOS GOVERNAMENTAIS PARA CONSULTA DE GASTOS PÚBLICOS DOS VEREADORES DA CÂMARA MUNICIPAL DE BELO HORIZONTE

Dissertação apresentada ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento, da Universidade FUMEC, como parte dos requisitos para a obtenção do título de Mestre em Sistemas de Informação e Gestão do Conhecimento.

Área de concentração: Gestão de Sistemas de Informação e do Conhecimento.

Linha de pesquisa: Tecnologia e Sistemas de Informação.

Orientador: Prof. Dr. João Victor Boechat Gomide

BELO HORIZONTE
2021

Dados Internacionais de Catalogação na Publicação (CIP)

A848c Assis, Wendel Vilaça de, 1987-

Chat Bot Sumé: *Web Scraping* em dados governamentais para consulta de gastos públicos dos vereadores da Câmara Municipal de Belo Horizonte / Wendel Vilaça de Assis. - Belo Horizonte, 2021.

90 f. : il.

Orientador: João Victor Boechat Gomide

Dissertação (Mestrado em Sistemas de Informação e Gestão do Conhecimento), Universidade FUMEC, Faculdade de Ciências Empresariais, Belo Horizonte, 2021.

1. Inteligência artificial. 2. Transparência na administração pública. 3. Belo Horizonte (MG). Câmara Municipal. I. Título. II. Gomide, João Victor Boechat. III. Universidade FUMEC, Faculdade de Ciências Empresariais.

CDU: 681.3.72

Ficha catalográfica elaborada pela Biblioteca Universitária-FUMEC

Dissertação intitulada "**CHAT BOT SUMÉ WEB SCRAPING EM DADOS GOVERNAMENTAIS PARA CONSULTA DE GASTOS PÚBLICOS DOS VEREADORES DA CÂMARA MUNICIPAL DE BELO HORIZONTE**" de autoria de Wendel Vilaça de Assis, aprovada pela banca examinadora constituída pelos seguintes professores**:**

Prof. Dr. João Victor Boechat Gomide – Universidade FUMEC
(Orientador)

Prof. Dr. Luiz Cláudio Gomes Maia– Universidade FUMEC
(Examinador Interno)

Prof. Dr. Rodrigo Moreno Marques – UFMG
(Examinador Externo)

Prof. Dr. Fernando Silva Parreiras
Coordenador do Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC

Belo Horizonte, 04 de março de 2021.

Prof. Dr. João Victor Boechat Gomide

Luiz Maia.

Rodrigo Moreno Marques

ABARIS

REQUESTED

TITLE	**Assinatura de ata e contra-capas Universidade**
FILE NAME	**80cb1ee6-e678-4b6c-adf9-386d35bf287a.pdf**
RESQUEST ID	**signature_request_884a6dee-2f79-4f3f-bcc2-9edad**
RESQUESTED BY	**Júlio César Teixeira e Silva**
STATUS	**Completed**

Professor (jvictor@fumec.br)

SENDED	05/03/2021 22:53:37UTC±0
SIGNED	08/03/2021 12:29:17UTC±0 186.214.220.146

Professor (luiz.maia@fumec.br)

SENDED	08/03/2021 12:29:17UTC±0
SENDED	09/03/2021 19:51:20UTC±0 191.185.140.62

Professor (rodrigomorenomarques@yahoo.com.br)

SENDED	09/03/2021 19:51:20UTC±0
SENDED	09/03/2021 20:04:12UTC±0 168.195.101.145

COMPLETED

09/03/2021
20:04:12 UTC±0
The document has been completed.

Dedico esta pesquisa a minha mãe Rosângela Vilaça, pelo seu exemplo de persistência, pois por vários anos da minha infância ela trabalhou como doméstica garantindo o sustento da família. Sem ela nada disso seria possível.

AGRADECIMENTOS

Muitos desafios enfrentei para alcançar o objetivo de concluir meu mestrado, consegui pela perseverança, esta define a origem da minha força, que foi concedida por Deus.

Agradeço ao:

Alex Marques pelo incentivo;
Aline de Almeida pela relação;
Ezequiel Vilaça pela ternura;
Família Vilaça pela convivência;
Fernando Parreiras pela crítica;
Josie Menezes pelo auxílio;
João Boechat pela orientação;
José Ricardo pelo apoio;
Leonardo Prado pela trajetória;
Marcos Pereira por ser meu pai;
Marcus Pinto pelo ensinamento;
Priscila Reis pela solicitude;
Rodrigo Moreno pela inspiração;
Rony Veloso pela direção;
Rosana pela oportunidade;
Rosângela Vilaça por ser minha mãe;
Senac Minas pelo estímulo.

Cada pessoa citada acima foi fundamental, se não houvesse a participação de cada um, meu objetivo não teria sido alcançado. Muito obrigado!

“Existe apenas um bem, o saber, e apenas um mal,
a ignorância.”
Sócrates

RESUMO

A participação ativa da população na política é essencial para a democracia, portanto, quando existem mecanismos tecnológicos que facilitam interação dos cidadãos com as instituições públicas, aumentam as chances do exercício da soberania popular. Inovações tecnológicas são desenvolvidas com o intuito de disponibilizar dados para a população sobre gastos em todas as esferas públicas, sendo a transparência governamental um dos precursores nesse sentido. Entretanto, muitos dados disponibilizados pelas instituições públicas são desestruturados ou estão em formato de leitura humana, como arquivos em PDF. Ocorre que, em diversas situações, as fontes de dados são distintas e pode existir a necessidade de consulta a centenas de arquivos, o que torna inviável e moroso para a população, diminuindo, assim, o interesse da sociedade em participar dessas consultas. Após a criação da Lei de Acesso à Informação - que impõe diretivas para as instituições públicas-, ainda assim, convivemos com a escassez de dados abertos, mesmo sendo um requisito obrigatório dessa lei. Nesse contexto, diversos pesquisadores iniciaram movimentos de governo aberto e dados abertos para tratar essa lacuna. Diante dessa conjuntura em que diversas bases não são disponibilizadas pelas instituições públicas como dados abertos, a adoção de métodos de *Web Scraping* tem apresentado notáveis resultados para extração de dados na Web. Quando se trata de dados governamentais, o *Web Scraping* possibilita a extração, manipulação e armazenamento de dados que antes não estavam disponíveis para leitura de máquina, possibilitando a transformação em dados abertos. Este trabalho desenvolveu um método *Web Scraping* em *Python* para extração de dados do portal de transparência da Câmara Municipal de Belo Horizonte. O resultado desse método foi a criação de dados abertos do Custeio Parlamentar, o que permite o compartilhamento desses dados e possibilita a produção de novas soluções para a sociedade exercer controle social, como o protótipo de *Chat Bot* Sumé, desenvolvido neste trabalho.

Palavras-chave: *Chat Bot*; Dados Abertos; Dados Governamentais; Inteligência Artificial; *Web Scraping*

ABSTRACT

The active participation of the population in politics is essential for democracy, therefore, when there are technological mechanisms that facilitate interaction between citizens and public institutions, the chances of exercising popular sovereignty increase. Technological innovations are developed with the aim of making data available to the population on expenditures in all public spheres, with government transparency being one of the precursors in this regard. However, many data made available by public institutions are unstructured or in human-readable format, such as PDF files. It so happens that, in several situations, the data sources are different and there may be a need to consult hundreds of files, which makes it unfeasible and time-consuming for the population, thus reducing society's interest in participating in these consultations. After the creation of the Access to Information Law - which imposes directives on public institutions -, we still live with the scarcity of open data, even though it is a mandatory requirement of this law. In this context, several researchers have initiated open government and open data movements to address this gap. Given this situation in which several databases are not made available by public institutions as open data, the adoption of Web Scraping methods has shown remarkable results for data extraction on the Web. When it comes to government data, Web Scraping enables extraction and manipulation and storage of data that was previously not available for machine reading, enabling transformation into open data. This work developed a Web Scraping method in Python to extract data from the transparency portal of the Municipality of Belo Horizonte. The result of this method was the creation of open data from Parliamentary Costing, which allows the sharing of these data and enables the production of new solutions for society to exercise social control, such as the Chat Bot Sumé prototype, developed in this work.

Keyword: Artificial Intelligence, Chat Bot, Government Data, Open Data, Web Scraping.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

Frequentemente, diversas notícias são veiculadas na mídia envolvendo escândalos de altas cifras referentes aos gastos públicos. A corrupção nas instituições públicas e os inúmeros gastos desnecessários são informações recorrentes em nosso cotidiano e, com o advento da internet, a participação popular passou a ser mais dinâmica por meio das mídias sociais (RODRIGUES; FONTES, 2018).

A internet tem sido parte importante no que se refere a prestações de contas e a consultas de gastos públicos. Com esse aliado, a demonstração das movimentações financeiras públicas fica disponível aos cidadãos para monitoramento dos gastos públicos, entretanto, não é tão eficaz em algumas circunstâncias, como no quesito de acessibilidade e adequação às diretivas da legislação vigente (EAVES, 2009).

Várias iniciativas no Brasil foram adotadas para divulgação de dados; uma delas foi a criação do Portal de Transparência do Governo Federal, inaugurado em 2004 pelo Ministério da Transparência e Controladoria-Geral da União (PTGF, 2004).

> O Portal da Transparência do Governo Federal é um site de acesso livre, no qual o cidadão pode encontrar informações sobre como o dinheiro público é utilizado, além de se informar sobre assuntos relacionados à gestão pública do Brasil. Desde a criação, a ferramenta ganhou novos recursos, aumentou a oferta de dados ano após ano e consolidou-se como importante instrumento de controle social, com reconhecimento dentro e fora do país. (PTGF, 2004).

Melhorias foram executadas ao longo dos anos referente à transparência de dados e em “18 de novembro de 2011 foi sancionada a Lei de Acesso à Informação Pública (LAI), lei 12.527/2011 que regula o acesso a dados e informações detidas pelo governo” (BRASIL, 2011). Segundo o Portal Brasileiro de Dados Abertos, “essa lei constitui um marco para a democratização da informação pública, e preconiza, dentre outros requisitos técnicos, que a informação solicitada pelo cidadão deve seguir critérios tecnológicos alinhados com as 3 leis de dados abertos” (INDA, 2011a). Complementando, o Portal Brasileiro de Dados Abertos é uma base centralizada de busca e acesso dos dados.

Baseado na LAI, as instituições públicas devem possuir seu próprio portal de transparência, sejam elas instituições de órgãos públicos ou órgãos independentes com ressalva aos municípios com menos de 10 mil habitantes. Em Belo Horizonte temos, dentre alguns portais, o Portal da Transparência da Câmara Municipal. A Câmara Municipal de Belo Horizonte (CMBH) mantém o portal de transparência para disponibilização dos gastos dos vereadores e destaca:

> O Portal da Transparência da Câmara Municipal de Belo Horizonte – CMBH - constitui-se como um importante instrumento por meio do qual esta Casa realiza o processo de prestação de contas ao cidadão belo-horizontino, promovendo o acesso a dados e informações sobre a gestão administrativa e a execução orçamentária e financeira da CMBH (CMBH, 2020a).

O conceito de dados abertos não é novo, a *Open Knowledge Foundation* (OKF, 2020a), defini dados abertos como:

> Dados abertos são dados que podem ser usados livremente, reutilizados e redistribuídos por qualquer pessoa - sujeitos apenas, no máximo, à exigência de atribuição e compartilhamento (OKF, 2020a).

O Brasil foi membro cofundador da Parceria para Governo Aberto, ou *Open Government Partnership* (OGP). Esse Portal é um de seus compromissos que foram formalizados no primeiro Plano de ação de governo aberto, lançado na OGP e referenciado pelo Decreto sem número de 15 de setembro de 2011 (OGP, 2011).

Apesar da existência de diversos sites de transparência mantidos pelas instituições públicas, o grande volume de dados disponibilizados pode dificultar que os cidadãos acompanhem, de maneira mais direta, como o erário é utilizado nas esferas federal, estadual e municipal. Logo, uma simples análise que um cidadão tente realizar de determinada atividade de um parlamentar ou de gastos públicos, acaba sendo algo complexo devido ao grande volume de informações e fontes distintas. Dessa forma, não se torna algo popular e, consequentemente, não é um processo democrático.

Essa dificuldade em manipular grandes volumes de dados fomentou várias iniciativas para que os mesmos fossem padronizados e pudessem ser lidos por máquina, conforme abordado por (EAVES, 2009). Com os dados em formato aberto para auxiliar na sua

manipulação, consequentemente facilita-se o entendimento dos humanos para análises e tomadas de decisões.

Mesmo com tantos dados disponibilizados pelas instituições públicas, ainda faltam soluções tecnológicas que facilitem aos cidadãos realizarem suas cobranças democráticas quanto aos investimentos públicos realizados em seu bairro, cidade, estado e no país. Então, a atuação da população nesse contexto pode não estar sendo efetiva devido à falta de participação. Consequentemente, a transparência ativa acaba não sendo satisfatória, tendo em vista que existe uma descentralização, além do alto volume de dados disponibilizados pelas instituições públicas, o que pode dificultar seu entrelaçamento.

Vale ressaltar que ainda hoje temos dados disponibilizados pelas instituições públicas somente em formatos de XLS, JPEG, PDF, DOC, ou seja, são dados com licenciamento proprietário ou que não podem ser manipulados por máquina, estes são um dos motivos que enquadram esses dados por não atender aos requisitos de dados abertos como apontado pelos autores (BIZER; HEATH; BERNERS-LEE, 2009; EAVES, 2009).

Embora as instituições públicas estejam cumprindo com as obrigações de transparência disponibilizando os dados nos formatos supracitados, elas não estão cumprindo com as obrigações da Lei de Acesso à Informação (LAI) no que se refere aos dados abertos, conforme (BRASIL, 2011). Segundo (EAVES, 2009), os dados são considerados inexistentes, caso não possam ser processados por máquina.

Dessa forma, uma análise minuciosa de dados armazenados em formatos distintos pode acarretar uma complexidade devido à falta de dados abertos. Então, somente pelo fato de os dados não serem abertos, acaba gerando consequências com impactos que impossibilitam o desenvolvimento de novas soluções e a adoção de novas tecnologias, tendo em vista os grandes valores que seriam necessários.

Atualmente, há várias formas de se ter acesso aos dados governamentais no âmbito federal; inclusive existem APIs para acessar e consumir Dados Abertos. O *Comprehensive Knowledge Archive Network* (CKAN), por exemplo, é um sistema para catalogação de dados para publicação de Dados Abertos (CKAN, 2013).

Além dos catálogos de dados como o CKAN, as *Application Programming Interface* (APIs) e *Web services* também podem ser utilizadas para facilitar o acesso e manipulação dos dados. Embora boa parte das iniciativas tenha por objetivo fornecer dados na Web, ainda não há um consenso sobre qual é a melhor maneira de se realizar esse procedimento. Entretanto, os catálogos e APIs têm sido bastante utilizados em função da facilidade no consumo dos dados.

Os catálogos de dados e as APIs são utilizados de diversas formas, sendo possível desenvolver soluções como serviços ou ferramentas. Um exemplo bastante conhecido é o projeto Serenata do Amor, uma ferramenta de controle social que foi fundamentada em dados abertos. “O objetivo desse projeto era expor para as pessoas informações sobre gastos dos deputados que já são públicas, porém não tão acessíveis” (RODRIGUES; FONTES, 2018). Segundo (VILANOVA, 2017): “O ponto forte do trabalho da Operação Serenata de Amor não é fazer algumas denúncias de milhões de reais. Sua função é possibilitar milhões de denúncias de alguns reais.”

No portal da transparência da CMBH é disponibilizado o acesso aos gastos dos vereadores referentes ao Custeio Parlamentar, Serviços Postais e Gastos com Telefonia, entretanto, esses dados estão desestruturados e não são dados abertos (CMBH, 2020b). Nesse caso, a CMBH não está promovendo o acesso de dados e informações, nem cumprindo com as obrigações de dados abertos conforme a LAI (BRASIL, 2011). A LAI traz consigo conceitos de dados abertos em seu texto, art. 8º e essa forma de disponibilização da CMBH não atende à LAI, nem aos padrões do governo aberto em que se preza pela transparência do estado com a disponibilização dos dados em formato aberto (OGP, 2011).

Tendo em vista o exemplo da CMBH, há estados e municípios que não disponibilizam seus Dados em formato aberto, o que pode dificultar a acessibilidade dos dados e impedir inovações e criação de novos serviços como, por exemplo, a Serenata do Amor (RODRIGUES; FONTES, 2018).

Diante desse cenário caótico de disponibilização de dados, o *Web Scraping* tem sido uma técnica utilizada para extração de dados não estruturados na *Web*, com objetivo de transformá-los em dados estruturados no formato *Comma Separated Values* (CSV) (MATTOSINHO, 2010). De acordo com (DIOUF, et al., 2019) “O principal objetivo do *Web*

Scraping é extrair informações de um ou vários sites e processá-las em estruturas simples, como planilhas, banco de dados ou arquivo (CSV)". Conforme apontamento do (DIOUF et al., 2019) o *Web Scraping* também pode ser utilizado para extração de dados estruturados quando o objetivo é manipular dados de fontes distintas.

Algumas alternativas são adotadas para tratar problemas específicos com técnicas de *Web Scraping* para extrair dados de sites ou base de dados que não são considerados dados abertos, entretanto, essas técnicas de *Web Scraping* possibilitam o atendimento de determinadas demandas (HERNÁNDEZ, et al., 2015).

1.1. JUSTIFICATIVA

Os gastos governamentais são objeto de discussão pelas cifras elevadas que permeiam esse setor. Embora existam portais de transparência para gastos federais, estaduais e municipais, os cidadãos podem encontrar dificuldades para monitorar como seus governantes utilizam as verbas em todas essas esferas. Uma consulta de gastos públicos no próprio município dos seus representantes pode ser uma tarefa árdua, então só o fato da transparência da informação existir e ser fomentada pela LAI, não significa que os gastos e utilização do erário são acessíveis, ou seja, de fácil visualização.

Tendo em vista a LAI e seus requisitos, esse estudo visa analisar - sobre a perspectiva de dados abertos - a disponibilização que a CMBH realiza referente aos dados de gastos dos vereadores no portal de transparência.

Optou-se por estudar os dados públicos do site de transparência dos vereadores do município de Belo Horizonte em Minas Gerais. Essa definição de escopo foi baseada no não-cumprimento da CMBH referente à disponibilização de dados abertos, tendo em vista que essa disponibilização é um direito dos cidadãos, não sendo facultativo.

Essa escolha de pesquisa visa agregar valor para a academia, pesquisadores e município. Com os resultados alcançados, poderão ocorrer o fomento e o engajamento nesse tipo de trabalho com a criação de um método para extrair dados não-estruturados e transformá-los em dados abertos. O objetivo final é fomentar a transparência pública e esse tido de estudo em outros municípios que também não disponibilizam dados abertos.

1.2. PROBLEMA DE PESQUISA

A ausência de dados abertos é um tema discutido internacionalmente devido às diversas situações que suas variáveis podem gerar impacto na sociedade, economia e tecnologia (OJO, CURRY; ZELETI, 2015).

Dados abertos governamentais podem gerar diversas discussões atuais e futuras, como evidenciado por (OJO, CURRY; ZELETI, 2015) que demonstra que os dados abertos conectados são um dos pilares de infraestrutura de cidades inteligentes devido à natureza do conjunto de dados publicados que apoia a inovação.

Nesse cenário, muitas questões surgem devido à complexidade de tratamento dos dados abertos governamentais e às tendências que podem ser utilizadas por meio dessa disponibilização. Embora existam limitações dos dados abertos nos portais de transparência das instituições públicas, as técnicas de *Web Scraping* são utilizadas para tratar os dados desestruturados que não estão em formato aberto. Esse tipo de técnica pode ser utilizado para soluções que possibilitem a geração de dados abertos, mesmo com limitações de padronização e licenciamento.

O portal de transparência da CMBH disponibiliza dados desestruturados referentes aos gastos Custeio Parlamentar dos vereadores; os dados são disponibilizados na página do portal em Box de seleção na página do portal. No caso dos gastos de Serviços Postais e Gastos com Telefonia, os dados estão em formato PDF. É possível afirmar que nenhum desses dados podem ser considerados dados abertos. A partir desse pressuposto, foi formulada a seguinte pergunta:

Como as técnicas de *Web Scraping* podem transformar os dados desestruturados dos portais de transparência em dados abertos?

A delimitação deste projeto de pesquisa se aplica apenas ao estudo do *Web Scraping* aplicado nos gastos de Custeio Parlamentar dos vereadores no site de transparência da Câmara Municipal de Belo Horizonte (CMBH) para transformação dos dados desestruturados, não se estendendo a outros contextos de pesquisa de *Web* Semântica, Transparência governamental, *Chat bot* ou Inteligência Artificial.

1.3. OBJETIVOS

1.3.1. Objetivo geral

O objetivo geral desse projeto será desenvolver um método de *Web Scraping* capaz de transformar os dados desestruturados do portal de transparência, da Câmara Municipal de Belo Horizonte, em dados abertos.

Quanto ao objetivo específico, traçou-se desenvolver um protótipo de *Chat Bot* baseado no resultado do método de extração dos dados no site da câmara CMBH como exemplo de uma solução fundamentada em dados abertos.

1.4. A ADERÊNCIA AO PROGRAMA DE PÓS-GRADUAÇÃO EM SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO

A pesquisa tem aderência ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento da Universidade FUMEC, tendo em vista que propõe o desenvolvimento de um método sistêmico para a extração de dados. O foco desta proposta está no campo Sistemas de Informação, em conformidade com o programa de pós-graduação da FUMEC.

1.5. ORGANIZAÇÃO DO TEXTO

Este trabalho está estruturado em 8 capítulos que foram elencados para estruturar a organização do texto.

No capítulo 1 temos a introdução, justificativa, o problema de pesquisa, o objetivo geral e o específico, além da aderência ao Programa de Pós-Graduação em Sistemas de Informação e Gestão do Conhecimento. Na introdução foi apresentada a abordagem dos temas que serão base de pesquisa para as discussões e para a fundamentação do estudo e da resolução do problema de pesquisa.

No capítulo 2 é apresentado o referencial teórico que contribuiu fundamentalmente para o embasamento dos estudos e sustentou a elaboração de propostas para atender ao objetivo geral e ao específico. Neste capítulo foram contextualizados temas propostos para

fornecer informações necessárias ao conteúdo da pesquisa. Os temas são Dados Governamentais, Transparência Governamental, Governo Aberto, Lei De Acesso À Informação, Transparência Ativa, Transparência Ativa Reversa, Transparência Passiva, Dados Abertos com suas derivações, *Web Scraping*, Inteligência Artificial, *Chat Bot*, Sumé.

O capítulo 3 demonstra os trabalhos relacionados que fomentaram o interesse pelo tema e foram fontes iniciais de estudo para adentrar nos campos dados abertos governamentais e desenvolvimento da pesquisa.

O capítulo 4 exibe a metodologia adotada para a pesquisa bibliográfica e documental com abordagem qualitativa, observando a ocorrência de Dados Governamentais e *Web Scraping* na literatura. Também é demonstrado nesse capítulo o tipo de pesquisa aplicada, ou seja, prática, pois o centro de estudo desse projeto é o de desenvolver um método de *Web Scraping* aplicável ao portal de transparência da CMBH.

O capítulo 5 traz a explanação dos dados existentes no portal de transparência da CMBH e apresenta as ferramentas necessárias para realizar a criação do método de *Web Scraping*, bem como, detalha as etapas de cada passo do experimento que foi realizado para extração, manipulação e gravação dos dados.

O capítulo 6 apresenta o protótipo do *Chat Bot* Sumé na solução de *Dialogflow* e relata a experiência da identificação de exclusão de dados do portal de transparência.

No capítulo 7 constam os resultados alcançados após a criação do novo método para extração de dados por *Web Scraping* e a avaliação do cenário de dados abertos do portal de transparência CMBH.

O capítulo 8 conclui o trabalho, com a avaliação do uso do método e a possibilidade de identificação de irregularidades referentes a dados retirados do portal de transparência, possibilitando, assim, o acesso aos gastos dos vereadores da CMBH. É explicitada a contribuição da pesquisa no meio acadêmico e social, bem como as sugestões de trabalhos futuros.

2. REFERENCIAL TEÓRICO

Neste capítulo são apresentados os conceitos principais utilizados no projeto de pesquisa para esclarecer toda a nomenclatura, com objetivo de facilitar o entendimento de todos os tópicos.

2.1. DADOS GOVERNAMENTAIS

Dados governamentais são um conjunto de elementos referente ao setor público, a Controladoria Geral da União (CGU) descreve que:

> As bases de dados governamentais são conjuntos de dados mantidos por órgãos, fundações e empresas públicos dos 3 Poderes e demais entidades controladas direta ou indiretamente pelas 3 esferas de governo: federal, estadual e municipal, enquadrando-se, portanto, no escopo da Lei de Acesso à Informação - LAI, Lei nº 12.527/2011 (CGU, 2017).

As ações públicas a respeito da abertura dos dados governamentais ainda são bastante incipientes. Prova disso é a dificuldade de se encontrar pesquisas a respeito das políticas e seus impactos (WANG; LO, 2015).

Ainda segundo (WANG; LO, 2015), “nos últimos anos, testemunhamos uma renovação de interesses em dados governamentais abertos em todo o mundo”. Nos Estados Unidos, por exemplo, um número considerável de políticas de dados abertos vem sendo desenvolvidas. Entretanto, a maioria dessas iniciativas ainda estão em fase inicial, portanto, permanece a dificuldade de acesso a pesquisas de entendimento da execução das políticas e seus impactos (WANG; LO, 2015).

Segundo (CORRÊA et al., 2017), parametrizações e artifícios técnicos para a disponibilização dos dados na internet estimularam o desenvolvimento da transparência, com a abertura de dados governamentais atrelado ao intuito de apresentar dados públicos sem restrição.

Habitualmente, os dados governamentais são publicados em formatos sem padrão ou de forma não estruturada. Além disso, as diversas bases de dados são expostas em formatos distintos com pouca ou nenhuma integração entre as mesmas. (FONSECA; AZEVEDO; ALMEIDA, 2014)

2.2. TRANSPARÊNCIA GOVERNAMENTAL

O Brasil conta com a LAI (Lei n. 12.527/2011) para reger a transparência governamental no país, além do Decreto nº 7.724/2012. Todavia, há necessidade de se criarem sistemas que minimizem os entraves ao acesso do cidadão às informações públicas de seu interesse. É crescente a demanda social por mais transparência na administração pública e democratização do acesso aos dados governamentais e científicos.

O conceito de transparência neste trabalho segue conforme exposto por (YAZIGI, 1999):

> É o acesso oportuno, suficiente e garantido do cidadão às informações relacionadas ao desempenho das funções públicas, é possível distinguir dois tipos de transparência: transparência ativa e transparência passiva (YAZIGI, 1999).

Em consonância com o ideal de transparência dos órgãos públicos, é necessário que os princípios de dados abertos (*open data*) sejam empregados. Só é possível viabilizar a transparência da administração pública por meio de um governo receptivo a plataformas tecnológicas que convertam os dados governamentais em dados abertos. Assim, o cidadão encontra não somente acesso à prestação de contas, mas, também, pode colaborar nos processos e melhorias dos serviços públicos (SANDOVAL; GARCIA, 2014).

Ressalta-se, contudo, que transparência e *open data* são expressões complementares, visto que uma concretiza os princípios da outra. A verdadeira transparência só é alcançada com a abertura dos dados, números etc.

2.2.1. Governo Aberto

Governo Aberto é a nomenclatura dada a projetos fomentadores da transparência governamental. Seus objetivos são lutar contra a corrupção, aumentar a participação social e o desenvolvimento de novas tecnologias, chegando, assim, a resultados de uma gestão pública mais consciente e otimizada.

Para abordar o conceito "Governo Aberto", o memorando "*Open Government Directive*" foi publicado em 8 de dezembro de 2009 destacando que "a transparência promove e fortalece a responsabilização dos atos governamentais, fornecendo ao público

informações sobre o que o governo está fazendo" (EAVES, 2009). Nesse formato governamental, os cidadãos podem contribuir com sugestões para as instituições públicas, levando aos órgãos públicos informações que estão amplamente disponíveis na sociedade. Isso traz eficácia ao governo e possibilidade de ampliar parcerias na Administração Pública, seja por cooperativas, ou instituições privadas, entidades não-governamentais, entre outras (OGD, 2007).

Nos Estados Unidos, em setembro de 2011 oito nações de diversos continentes instituíram a "Parceria para o Governo Aberto" (*The Open Government Partnership – OGP).* Vale ressaltar que a ação foi liderada pelos E.U.A. e pelo Brasil. Depois de dois anos de fundada, a OGP contava com 75 nações filiadas, seguindo no objetivo de assegurar compromissos concretos dos governos em promover transparência, lutar contra a corrupção e capacitar cidadãos (OGP, 2011).

2.2.2. Lei de acesso à informação

A legitimidade dos órgãos públicos está diretamente ligada à transparência das informações e ao acesso democrático da sociedade a esse banco de dados. O cidadão começa a dar mais credibilidade às atuações do poder público as quais têm acesso diretamente. Dessa forma, com mais transparência e controle social da administração pública é possível alcançar até mesmo uma redução geral no tempo demandado para monitorar os gastos públicos.

Para formalizar a garantia do direito do cidadão ao acesso democrático às informações, foi necessária a instituição da LAI, conhecida como Lei de Acesso à Informação. Ela chegou para legitimar a democracia no contexto do acesso do cidadão aos dados da gestão pública. Junto à LAI, há o Decreto n. 7.724, de 16 de maio de 2012, que a regulamentou na esfera federal, o que representou um divisor de águas para o acesso à informação pública no País (SÁ; MALIN, 2012).

A transparência de informações e a abertura dos dados passaram a ser obrigação da administração pública, englobando os órgãos públicos da administração direta e indireta dos Poderes Executivo, Legislativo e Judiciário, incluindo, também, as Cortes de Contas e o Ministério Público, em todas as esferas de acordo com art. 1º, parágrafo único. (BRASIL, 2011).

Apesar do marco da chegada da LAI no Art. 8º (BRASIL, 2011) atribuindo aos órgãos públicos a responsabilidade pela promoção de dados abertos, muitos órgãos não atendem aos requisitos de transparência ativa e dados abertos, mesmo não sendo algo facultativo. A LAI estabelece conceitos de dados abertos em seu texto, no art. 8º:

> Art. 8º É dever dos órgãos e entidades públicas promover, independentemente de requerimentos, a divulgação em local de fácil acesso, no âmbito de suas competências, de informações de interesse coletivo ou geral por eles produzidas ou custodiadas. (…) § 2º Para cumprimento do disposto no caput, os órgãos e entidades públicas deverão utilizar todos os meios e instrumentos legítimos de que dispuserem, sendo obrigatória a divulgação em sítios oficiais da rede mundial de computadores (internet). § 3º Os sítios de que trata o § 2º deverão, na forma de regulamento, atender, entre outros, aos seguintes requisitos: (…) II – possibilitar a gravação de relatórios em diversos formatos eletrônicos, inclusive abertos e não proprietários, tais como planilhas e texto, de modo a facilitar a análise das informações; III – possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina;
> A lei também define as hipóteses de sigilo e de informações pessoais, que são consideradas exceções à regra geral de que os dados devem ser abertos. (BRASIL, 2011).

Segundo o apontamento da LAI (BRASIL, 2011), para ser considerado Dado Aberto ele deve estar em formato aberto, estruturado e legível por máquina.

2.2.2.1. Transparência ativa

Para respeitar o conceito de transparência ativa, os órgãos e entidades públicas precisam garantir o acesso do cidadão às informações de interesse geral, independentemente de serem solicitadas.

Esse foi o verdadeiro avanço atingido com a regulamentação da LAI, e do Decreto n. 7.724, que regulamentou a lei na esfera federal. Esse momento representou uma evolução para o acesso à informação no país, nas iniciativas de transparência governamental (SÁ; MALIN, 2012).

Observa-se, entretanto, que nas diretrizes da LAI, há algumas exceções, como, por exemplo, é destacado no artigo “Uma análise da transparência ativa nos sites ministeriais do Poder Executivo Federal brasileiro”:

> Os princípios da transparência ativa e da transparência passiva são elementos centrais tanto da LAI quando do referido decreto federal. Ambas as modalidades de transparência são obrigatórias para todos os órgãos e entidades públicas federais, estaduais, municipais e distritais, dos poderes Executivo, Legislativo e Judiciário, além de toda a administração pública, sendo facultativo o seu cumprimento nos casos dos municípios cuja população seja inferior a dez mil habitantes (ARAÚJO; MARQUES, 2019).

Sobre a transparência ativa, ela impõe aos órgãos estaduais a obrigação de apresentarem, periodicamente, informações padronizadas que permitam o conhecimento da comunidade a respeito do desempenho das atividades administrativas. Isso significa que o cidadão precisa ter acesso às funções do governo, seus objetivos, atividades, recursos humanos, orçamentos, despesas, obras etc.

Outro viés importante é que os dados precisam ser disponibilizados de maneira estruturada, permitindo ao cidadão que deseje fazer um comparativo ao longo dos anos. De acordo com (YAZIGI, 1999), “a transparência ativa é uma maneira proativa da administração pública divulgar informações sem a necessidade de algum pedido ou solicitação da sociedade. Quando ocorre o tipo ativo de transparência, ele torna-se um instrumento central para reivindicações de direitos sociais. Já os autores Zuccolotto, Teixeira e Riccio destacam que:

> A Transparência ativa consiste na difusão periódica e sistematizada de informações sobre a gestão estatal. Resulta de ações voluntárias dos gestores públicos ou de obrigações legais impostas aos órgãos do Estado, determinando que sejam publicadas informações necessárias e suficientes para que a sociedade possa avaliar o desempenho governamental (ZUCCOLOTTO; TEIXEIRA; RICCIO, 2015, p. 148).

2.2.2.1.1. *Transparência ativa reversa*

Cunhado pelo autor da dissertação, o termo Transparência Ativa Reversa ocorre em situações que os próprios cidadãos desenvolvem métodos, técnicas, algoritmos, ferramentas ou soluções para transformar dados desestruturados e as informações públicas desordenadas em dados abertos.

O nome Transparência Ativa Reversa faz analogia a Engenharia Reversa pelo fato de criar algo baseado em alguma coisa já existente. Como exposto por (DICKEN, 96) "A Engenharia Reversa consiste em produzir novas peças, produtos ou ferramentas a partir de modelos ou componentes existentes".

Esse conceito tem harmonia com a situação exposta, tendo em vista que os dados abertos provenientes do método de *Web Scraping* se utilizam de dados desestruturados disponíveis no portal de transparência da CMBH através da Transparência Ativa.

A Transparência Ativa Reversa tem sido utilizada em dados que são disponibilizadas pelas instituições públicas em formato que não é considerado aberto como, por exemplo, XLS, JPEG, PDF, DOC. Apesar desses dados estarem disponíveis na internet nos formatos citados, os dados não são abertos e alguns não estão em formato aberto, então eles não estão acessíveis de maneira facilitada para população exercer o controle social. Dessa forma, percebe-se a diferença que existe entre disponível e acessível (RODRIGUES; FONTES, 2018).

No caso semelhante da Câmara Legislativa, portal que expõe os gastos dos Deputados, tecnicamente atenderia à Lei da Transparência, tendo em vista que disponibiliza informações referentes a gastos públicos. Todavia, essa publicação se dá de maneira não transparente devido ao formato adotado, o que praticamente inviabiliza a interpretação dos dados lá presentes (RODRIGUES; FONTES, 2018). Ainda conforme o autor o mesmo aponta:

> As informações referentes aos deputados federais, por exemplo, encontram-se disponíveis no website da Câmara Legislativa, contudo, a enorme quantidade de dados em formatos não amigáveis para seres humanos torna a fiscalização uma árdua tarefa (RODRIGUES; FONTES, 2018).

Em sua análise (RODRIGUES; FONTES, 2018) comprova a diferença entre o que é disponível e o que é acessível. Ele ressalta que os dados referentes aos deputados federais de 2017 até 2018, por exemplo, encontravam-se em um arquivo do tipo .XML com ~493Mb. (MUSSKOPF, 2016):

> Apesar de cumprir o papel exigido por lei de garantir a transparência dos pagamentos e conter várias informações úteis para qualquer cidadão verificar sua legalidade, não são acessíveis. Ferramentas como Microsoft Word e Excel não são feitas para trabalhar com arquivos de centenas de megabytes; com memória RAM suficiente, o farão com extrema lentidão, dificultando que o brasileiro comum, não especialista em análise de dados, faça pesquisas pelos nomes dos seus representantes e ajude a verificar o seu trabalho (MUSSKOPF, 2016).

A participação popular pode ser fomentada por meio da acessibilidade na utilização de recursos tecnológicos, sendo que uma parcela dessa contribuição pode ser atribuída à Transparência Ativa Reversa. Essa laboração possibilita a disponibilização de dados abertos relativos à consulta dos gastos, tendo em vista que isso demonstra a aproximação da população de seus representantes.

Os representantes populares eleitos devem zelar pelas suas demandas locais e de todo o país, não se restringindo somente a sua região em que foi eleito. Em contrapartida, a população deve cobrar, monitorar a atuação e os gastos dos seus representantes, portanto, essa aproximação entre eleito e eleitor fortalece a representatividade, o que é essencial para democracia, como já exposto por Edmund Burke:

> Você escolhe um membro, de fato; mas quando você o escolhe, ele não é membro de Bristol, mas é membro do Parlamento. Se o representante local tiver interesse ou formar uma opinião precipitada, evidentemente oposta ao verdadeiro bem do resto da comunidade, o membro desse lugar deve estar o mais longe que qualquer outro de qualquer esforço para dar efeito. (BURKE, 1774)

2.2.2.2. Transparência passiva

Quando o estado fornece aos cidadãos as informações solicitadas por eles mesmos, o autor (YAZIGI, 1999) denomina como transparência passiva. Ao contrário do que ocorre na transparência ativa, nesse formato passivo, o estado aguarda que o cidadão o acione com o

requerimento dos dados desejados. Só após esse contato, as informações são disponibilizadas ao requerente. No caso da transparência governamental passiva, o acesso aos dados de determinado órgão público é concedido pelo estado após requerimento feito ao mesmo. Admitem-se exceções aos princípios de transparência ativa e passiva, mas somente em poucos casos previamente estabelecidos em lei, justificados por razões de imparcialidade, confidencialidade ou sigilo de assuntos de segurança nacional, por exemplo.

Segundo (LOPES; ASSUMPÇÃO, 2013) “a transparência passiva possibilita que o cidadão solicite de maneira simples informações ao governo, sendo um pilar para fomento do controle social.” Entretanto, apesar do autor defender essa afirmação, não necessariamente a facilidade para o cidadão efetuar a solicitação vai implicar em resposta efetiva do governo. O mesmo poderá responder ao que foi solicitado, entretanto, isso não significa que os questionamentos serão solucionados.

Para embasar essa questão, observa-se um exemplo na pesquisa realizada no portal de transparência da CMBH. Nessa pesquisa foi identificada a ausência de dados abertos sobre os gastos dos vereadores. Diante do exposto, foi solicitado formalmente, conforme (Apêndice A), o envio de informações sobre os dados abertos e se havia algum projeto para disponibilizá-los.

O pedido foi aberto facilmente à CMBH, comprovando a afirmação do autor (LOPES; ASSUMPÇÃO, 2013), entretanto, a resposta oficial do órgão foi totalmente sem conexão com a pergunta efetuada. Diante desse exemplo percebe-se que não foi possível sequer obter resposta, muito menos exercer controle social; portanto, o contato foi totalmente ineficaz. Nessa conjuntura, esta solicitação seria registrada como atendida, mesmo sendo inexata e insuficiente.

2.3. AS 5 ESTRELAS DOS DADOS ABERTOS

O primeiro autor a mencionar os termos Dados Ligados ou Dados Conectados foi um dos inventores da *Web*: Tim Berners-Lee (BIZER; HEATH; BERNERS-LEE, 2009). Após pensar sobre esses dados, ele propôs um formato de implementação de 5 estrelas para Dados Abertos, sistema conhecido por definir uma classificação para os dados que são

disponibilizados na *Web*. Com isso passa a ser viável a identificação da maturidade dos dados, avaliando se são considerados abertos ou não (BERNERS-LEE, 2009).

A classificação das estrelas se dá de maneira sucessiva, ou seja, a subsequente recebe os requisitos da anterior, sendo que cada classificação de estrela de ordem superior respeita as restrições de uma classificação de estrela de ordem inferior. A figura a seguir demonstra a classificação:

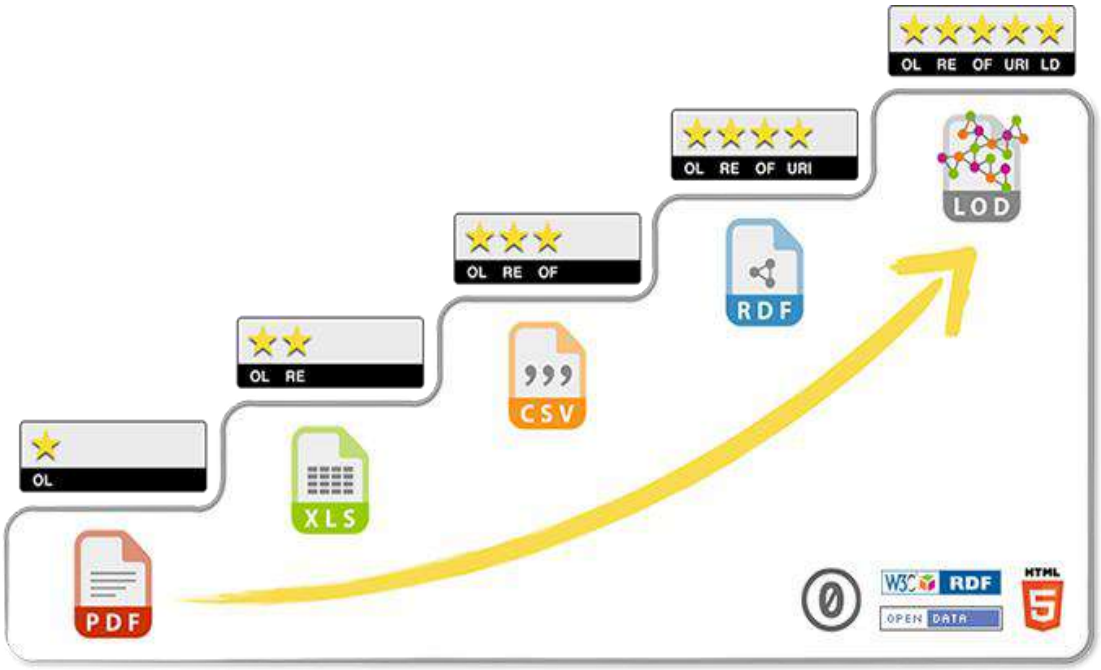


Figura 1 - Escala de desenvolvimento e enriquecimento dos dados abertos (5Stars, 2012)

Abaixo, segue os 5 exemplos descritos de classificação das estrelas elaborado por Tim Berners-Lee (BERNERS-LEE, 2009). para cada classificação de estrelas é retratado sua definição:

> 1 Estrela: O dado está disponível na web, em qualquer formato (pdf, png, jpeg); 2 Estrelas: O dado está disponível como sendo legível por máquina e estruturado (uma planilha do Excel); 3 Estrelas: O dado está disponível num formato não-proprietário (por exemplo, planilha CSV em vez de Excel). 4 Estrelas: O dado é publicado usando os padrões de dados abertos da *World Wide Web Consortium* onde é possível usar URIs para denotação. 5 Estrelas: Todos os itens acima se aplicam, além de links para dados de fontes diferentes para fornecer contexto (BERNERS-LEE, 2009).

O fato de haver publicação de dados das instituições públicas irrestrita nos portais de transparência não significa que os dados são abertos. Como abordado acima na classificação, um dado é considerado aberto quando ele pode ser manipulado por máquina e é livre de licenças, ou seja, tem que possuir no mínimo 3 estrelas. Embora existam essas definições, diversos sites governamentais não atendem a esses requisitos e não disponibilizam dados abertos.

2.3.1. Dados Abertos

Um dado aberto é considerado por Tim Berners-Lee, quando está disponível no mínimo em formato de arquivo CSV, ou seja, se trata de um dado estruturado no formato não-proprietário como por exemplo, planilha CSV em vez de XLS do Excel (BERNERS-LEE, 2009). Um dado não é considerado aberto, quando o formato do arquivo disponível possui licenciamento proprietário, como por exemplo uma planilha do Excel ou arquivos nos formatos PNG, JPEG (BERNERS-LEE, 2009).

Essa dissertação adotará a definição de Dados Abertos da LAI (BRASIL, 2011) em que sua interpretação permite o entendimento de que os dados públicos só podem ser considerados Dados Abertos se possuírem no mínimo três estrelas na formulação do esquema de estrelas proposto por Berners-Lee (2009). Entretanto serão apresentados outros autores que contextualizam esse tema para ampliar o debate no que se refere a dados abertos.

Para Davenport (2014), a definição de dado é algo fundamental, podendo ser um caractere, um símbolo ou alguma representação sem contexto passível de interpretação definida. “Os Dados podem ser representados por um conjunto de acontecimentos discretos e objetivos sobre determinados eventos, sendo capaz de ser entendidos numa organização como registros estruturados de transações.” (DAVENPORT; PRUSAK, 1998).

A *Open Knowledge Foundation* (OKF), organização sem fins lucrativos, institucionalizada na Inglaterra e País de Gales, define Dados Abertos como aqueles que podem ser acessados por qualquer pessoa livremente (OKF, 2020b). E, além do acesso democrático, para que sejam considerados abertos é necessário que os consulentes consigam, também, utilizar esses dados, fazer modificações e compartilhá-los para qualquer finalidade.

As únicas exigências apontadas pela OKF são que a proveniência e abertura dos dados se mantenham preservadas.

A primeira versão do documento de Dados Abertos foi publicada em 2005 pela OKF e até hoje está vigente. Atualmente, é mantida por um conselho consultivo de especialistas, estando a definição na versão 2.1, e tendo sido atualizada em 2015. "A definição da palavra Aberta torna preciso o significado de 'aberto' no que diz respeito ao conhecimento, promovendo um bem comum robusto, no qual qualquer pessoa pode participar e a interoperabilidade é maximizada" (OKF, 2020c). De acordo com a OKF, no que diz respeito ao software, a definição de "aberto", ou ainda, "Código Aberto" é sinônimo de "livre" ou "*libre*", como na definição de Software Livre e de Trabalhos Culturais Livres.

Já o termo *Open Definition* (OD) traz o conceito de dados abertos como sendo aqueles passíveis de serem utilizados livremente, além de reutilizados e redistribuídos por qualquer pessoa. Ressalta-se que, no máximo, haverá exigência de atribuição à fonte original das informações e, também, ao compartilhamento pelas mesmas licenças nas quais os dados foram apresentados (OKF, 2020c). Isto posto percebe-se que um dos objetivos da abertura de dados é minimizar os mecanismos de controle ou monopólio, visto que os mesmos podem acarretar restrições sobre o que foi publicado, impedindo, assim, que pessoas físicas e jurídicas explorem livremente as informações.

De acordo com essa perspectiva, o conceito de dados abertos carrega três normas fundamentais OKF:

> Disponibilidade e acesso: os dados devem estar disponíveis como um todo e sob custo não maior que um custo razoável de reprodução, e preferencialmente devem ser possíveis de ser baixados pela Internet. Os dados devem também estar disponíveis de uma forma conveniente e modificável. Reúso e redistribuição: os dados devem ser fornecidos sob termos que permitam a reutilização e a redistribuição, inclusive a combinação com outros conjuntos de dados. Participação universal: todos devem ser capazes de usar, reutilizar e redistribuir – não deve haver discriminação contra áreas de atuação ou contra pessoas ou grupos. Por exemplo, restrições de uso "não comercial" que impediriam o uso "comercial", ou restrições de uso para certos fins (ex.: somente educativos) excluem determinados dados do conceito de "abertos" (OKF, 2020b).

2.3.2. Dados Abertos Governamentais

A classificação de Dados Abertos Governamentais se dá de acordo com as mesmas diretivas e definições aplicadas aos Dados Abertos, conforme pode-se observar na publicação da *The Association of Computing Machinery's*, com a seguinte recomendação ao Governo Aberto:

> Os dados publicados pelo governo devem ser em formatos e abordagens que promovam a análise e reutilização desses dados. (OGD, 2007).

Algumas condições foram estabelecidas para que um determinado dado governamental pudesse ser considerado como aberto. Há três leis referentes ao conceito Dados Abertos Governamentais, conforme abaixo (EAVES, 2009).

> As três leis dos dados governamentais abertos:
> 1. Se o dado não pode ser encontrado e indexado na Web, ele não existe; 2. Se não estiver aberto e disponível em formato compreensível por máquina, ele não pode ser reaproveitado; e 3. Se algum dispositivo legal não permitir sua replicação, ele não é útil (EAVES, 2009).

As três leis passaram a ser um norteador referente aos Dados Abertos e, para além delas, em meados de 2007, um grupo de trabalho de 30 pessoas resolveu dar um passo ainda maior para definir os princípios dos Dados Abertos Governamentais.

O grupo reuniu-se na Califórnia, Estados Unidos da América e chegou a um consenso referente a oito princípios fundamentais do tema. Esses princípios estão divulgados em publicações feitas por diversas instituições, como a *SunLight Foundation*, *Association for Computing Machinery*, ***The White House***, entre outras (OGD, 2007).

Os Dados Abertos são compartilhados na Internet em formato aberto, ou seja, formato CSV, assim pode ser consumido por máquina e qualquer pessoa conforme (BERNERS-LEE, 2009). Essa ação possibilita a convergência de dados de diferentes bases, promovendo acesso livre pela sociedade.

Com a criação do Data.gov no ano de 2009, a visão de dados abertos governamentais ganhou força. À época o objetivo foi o de tornar acessível a qualquer cidadão a consulta dos

dados do governo norte-americano. Com o Data.gov, foi disponibilizado um portal de dados abertos na internet, constando dados governamentais (DG, 2009).

Já no Brasil houve também a criação do portal dados.gov.br, que disponibilizou os dados públicos seguindo os princípios de dados abertos. Vale ressaltar que o Brasil foi, ainda, um dos fundadores da *Open Government Partnership,* em 2011. A organização atualmente conta com a participação de 75 países (OGP, 2011).

Ainda em 2011 foram instituídos pelo governo federal, por meio da política de acesso à informação, tanto a INDA quanto os dados abertos governamentais, ambos criados em 2011 (INDA, 2011b). A necessidade de disponibilizar dados governamentais em formato aberto consta do artigo 8º da Lei de Acesso à Informação, comprovando o reconhecimento da importância desse tema por parte da Administração Pública Federal.

Outra política importante a ser mencionada é a Instrução Normativa da Infraestrutura Nacional de Dados Abertos (INDA, 2011c). Essa política do governo brasileiro regulamentada também em 2011 traz um conjunto de procedimentos, tecnologias e processos de controle. O objetivo é atender aos requisitos de disseminação e compartilhamento de dados e informações públicas por meio do modelo de Dados Abertos.

As 75 nações que compõem a *Open Government Partnership* (OGP) vem estruturando e publicando catálogos de dados governamentais, sendo os mais conhecidos e divulgados os do Brasil, dos Estados Unidos e do Reino Unido. Em relação à União Europeia é válido dar ênfase à criação do catálogo PublicData.eu, que apresenta à sociedade uma ferramenta importante. Ela consome informações de outros 29 catálogos de dados de nações da União Europeia, possibilitando a concentração dos dados num único ponto de acesso.

Para facilitar a visualização, a figura abaixo representa a aliança das nações da OGP. Por meio do esquema de cores é possível observar o nível de maturidade das nações mundiais junto à aliança. As que estão na cor marrom encontram-se no primeiro estágio; cor verde, segundo estágio e na cor amarela estão o estágio terceiro. Neste terceiro e mais avançado estágio localizam-se o Brasil e os Estados Unidos (OGP, 2011).

Figura 2 - Mapa Mundi ilustrativo com as em prol do Governo Aberto (OGP, 2011).

Sobre o conceito de *crowdsourcing*, podemos afirmar ser um caminho a ser adotado para a relação entre governo e sociedade. Trata-se de um modelo de produção baseado na inteligência coletada na internet. Os conhecimentos coletivos disponibilizados por voluntários na rede são utilizados para resolver problemas, criar conteúdos novos e propor soluções e o desenvolvimento de novas tecnologias. Considerado um modelo inovador, o *crowdsourcing* representa um modelo de criação entre o Governo e a sociedade, no qual estão presentes características essenciais ao desenvolvimento, como engajamento da população, estímulo à criatividade e participação ativa em torno da ação governamental.

Iniciativas desse contexto podem ter correlação profunda com a relevância das ações de dados abertos. No caso do escritório de big data Centro de Operações Rio (COR), observa-se uma inovação que enquadrada na análise feita acima. A agência, além de ser uma inovação urbana na forma de mecanismos de dados abertos, é vinculada a um projeto de cidade inteligente. Ou seja, é um exemplo que demonstra e exemplifica o conceito de dados abertos como uma das possíveis iniciativas de cidade inteligente – conforme sugerem (OJO, CURRY; ZELETI, 2015).

2.3.3. Dados Conectados

Dados Conectados (do inglês, *Linked Data*) podem ser definidos como um conjunto de boas práticas para publicar e conectar os diversos dados estruturados na Web. O intuito final é o de criar uma rede de dados denominada “Web de Dados” (BIZER; HEATH; BERNERS-LEE, 2006). Ressalta-se que tais práticas possuem fundamentos tecnológicos específicos da web, tais como: HTTP (*Hypertext Transfer Protocol*) e URI (*Uniform*

Resource Identifier) (BIZER; HEATH; BERNERS-LEE, 2001). São agentes de software que permitem a leitura automática dos dados conectados, de forma automática.

A padronização para a publicação de dados na web consta de um conjunto de regras estabelecido no ano de 2006. À época houve uma definição padronizando formatos em que todos e quaisquer dados publicados pudessem tornar-se parte de um espaço único de dados global, o que viabilizava, assim, a integração na rede (BIZER; HEATH; BERNERS-LEE, 2009).

Os dados conectados possibilitam uma convergência nunca possível pela interconexão dos dados, conforme destacam os autores abaixo:

> Os padrões de Dados Conectados permitem que qualquer pessoa publique os dados de uma maneira que possam ser lidos por pessoas e processados por máquinas. Isso é possível porque os dados que antes estavam "escondidos" na Web de Documentos estão agora acessíveis graças à utilização dos padrões supracitados para a conexão de dados (ISOTANI; BITTENCOURT, 2015).

Os Dados Conectados propiciam soluções do cotidiano conforme (ISOTANI; BITTENCOURT, 2015) "Essa conexão de dados permite que todos (homens e máquinas) possam trabalhar conjuntamente de forma mais eficiente (como no desenvolvimento de aplicações para os cidadãos com o objetivo de melhorar o transporte público)".

Para facilitar o entendimento do conceito de dados abertos um exemplo ilustrativo pode ser representado da seguinte forma. Visualize uma engrenagem automática, na qual, fosse possível buscar dados de fontes diversas, em formatos bem estruturados e conectados, sendo todos legíveis por máquina. Dessa forma, qualquer desenvolvedor web têm acesso a esses dados simultaneamente, podendo, inclusive, combiná-los em tempo real. Essa coleta de dados totalmente automatizada é possível por meio da utilização de Dados Conectados (WOOD, et al., 2013).

2.3.4. Dados Abertos Conectados

A publicação e o consumo de dados pode ser considerado como uma tendência, visto que diversas soluções do nosso cotidiano utilizam dados abertos e conectados disponibilizados pelo governo como o próprio Serenata do Amor. É notório o crescimento

do número de órgãos governamentais que estão disponibilizando seus dados na internet, permitindo que esses dados sejam reutilizados por empresas, cidadãos e outros órgãos governamentais, incentivando assim a publicação de dados abertos.

Há três conceitos destacados ao longo de toda a análise realizada neste trabalho. São eles: Governo Aberto, Dados Abertos e Dados Conectados. Ressalta-se, contudo, que um deles unifica os demais. São os Dados Abertos, conforme mostra o autor (BIZER; HEATH; BERNERS-LEE, 2009): “Os Dados abertos conectados se tornaram um conceito que unificou outros três, o 1° é “Governo Aberto” (*Open Government),* 2° “Dados Abertos” (*Open Data*) e o 3° “Dados Conectados” (*Linked Data*).” O objetivo é viabilizar a colaboração na utilização dos dados e preservar a transparência em órgãos, entidades públicas ou industriais. Isso é viável com a criação de uma infraestrutura de dados para cidades inteligentes. (BIZER; HEATH; BERNERS-LEE, 2009).

Conforme explicitado acima, um exemplo de utilização de dados abertos ocorre nas cidades inteligentes ou *Smart Cities*. Nessas circunstâncias os dados abertos podem auxiliar as cidades, tendo em vista que conceitualmente eles são como uma iniciativa de cidade inteligente. Em uma perspectiva emergente, de duas maneiras: quando iniciativas de dados abertos apoiam ou estão alinhadas aos objetivos de uma cidade inteligente ou quando o contexto de cidades inteligentes dá forma às iniciativas de dados abertos (OJO, CURRY; ZELETI, 2015).

A política de "Governo Aberto” preserva o acesso livre aos dados e garante que as pessoas possam usá-los e reutilizá-los livremente, segundo (OGP, 2011). É uma política que engloba os dados e informações produzidas pelas instituições públicas, e está isenta em relação a dados ou informações pessoais. Já o termo “*Open Data”* inclui, além da entidade governamental, outras entidades das áreas de negócio, industriais, organizações sem fins lucrativos. Neste contexto, para cumprir as determinações dos oito princípios dos Dados Abertos Governamentais era essencial manter transparência e padrões. “Com o uso de *Linked Data* e RDF (*Resource Description Framework*) é possível manter a interoperabilidade e estabelecer um padrão de representação dos dados, surgindo então, o conceito *Linked Open Data* (LOD), também chamado de Dados Abertos Conectados.” (BIZER; HEATH; BERNERS-LEE, 2009a).

Representando as boas práticas adotadas para publicação e conexão dos dados na web, o termo *Linked Data* utiliza tecnologias padrões na representação de dados estruturados, cujo objetivo é garantir o consumo dos dados por humanos, além do processamento por máquinas (BIZER; HEATH; BERNERS-LEE, 2009a).

Pelo volume de dados gerados por ano e analisando o contexto atual em que vive a sociedade, se fez necessário conseguir meios para compreender os dados, gerar informações, conhecimentos e agregar valor. Com isso, os Dados Abertos ganham notoriedade na gestão pública, tendo em vista que os dados governamentais podem ser reutilizados por empresas, cidadãos e outros órgãos. Acarreta, assim, em uma sinergia na gestão pública e no incentivo a ampliar, cada vez mais, a publicação de dados abertos

Com a LAI foi implementado através das instituições públicas diversos portais de transparência para a publicação de dados relativos à gestão pública e à utilização do erário. Além disso, várias entidades seguiram o mesmo caminho e passaram a publicar os dados abertos também. Com isso, aplicações diversas foram criadas num curto espaço de tempo, tendo em vista que o pré-processamento desses dados não era mais necessário.

As APIs foram utilizadas para a publicação de dados abertos e, também, viabilizaram a utilização das máquinas de maneira simplificada. A questão é que os dados abertos, ao atingirem a classificação máxima de 5 estrelas, podem ser conectados, o que possibilita maior poder de processamento para aplicações desenvolvidas (BERNERS-LEE, 2009). O resultado é a geração de novos empreendimentos, trazendo mais soluções e inovações para a sociedade, além de melhor qualidade de vida aos cidadãos, tudo isso por meio de uma melhor gestão pública de dados.

Sobre evoluções tecnológicas e a origem dos Dados Abertos Conectados *(Linked Open Data*), podemos afirmar que:

> O conceito de Dados Abertos Conectados, do inglês *"Linked Open Data*", foi criado por Tim Berners-Lee pela necessidade de padronizar a conexão entre dados na Web. Compreende-se que o uso dos padrões criados pelos Grupos de Trabalho do W3C e o trabalho da comunidade de desenvolvedores, de gestores governamentais e da sociedade interessada no desenvolvimento Web são essenciais

> para que se alcance efetivamente dados abertos e conectados (ISOTANI; BITTENCOURT, 2015, p.13).

2.4. WEB SCRAPING

Como fonte primordial de acesso à informação, a internet ganha protagonismo ao armazenar inúmeros dados no contexto Web que, por fim, geram informação e conhecimento. Entretanto, embora os dados disponíveis na Web possam ser utilizados na maioria das vezes, eles também aparecem em formatos desestruturados em algumas ocasiões.

Sem dúvida, os dados obtidos pela Internet são úteis a diversos profissionais e setores; é uma gama de informações que são geradas para a sociedade por meio da correlação de dados. Em algumas ocasiões pode ser necessário lançar mão dos métodos de *Web Scraping* para facilitar esse serviço, visto que pode ocorrer morosidade em certas circunstâncias nas quais os dados na Web são desestruturados, o que torna o acesso deles algo muito mais complexo.

De acordo com o autor (HERNÁNDEZ et al., 2015) "*Web Scraping* ou extração de dados da Web é o processo de rastreamento e download de sites de informações e extração de dados não estruturados para um formato estruturado.".

O processo de extração por *Web Scraping* pode ser feito de forma manual, por exploração humana da *World Wide Web*. O processo se dá quando uma pessoa acessa a internet e coleta dado a dado manualmente ou um conjunto de dados. O *Web Scraping* também pode ser por implementação de scripts, sendo este um recurso de baixo nível do protocolo de transferência de hipertexto, ou via incorporação de certos navegadores Web (HERNÁNDEZ, et al., 2015).

Especialistas de outras áreas, que não tenham experiência em programação ou ferramentas de *Web Scraping*, certamente terão dificuldade quando precisam manipular dados na Web. Esse é um fator impeditivo do livre e fácil acesso de diversos especialistas que necessitam efetuar a extração de informações na Web para análise. Um exemplo desse tipo de profissional são os jornalistas. Para (DIOUF et al., 2019) a imprensa é o órgão composto por profissionais que mais necessitavam de *Web Scraping*, entretanto, era o que possuía menos ferramentas especializadas.

"O *Web Scraping* também é identificado como extração, coleta de dados da web, coleta da web ou telaraspagem. *Web Scraping* é uma forma de mineração de dados." (MATTOSINHO, 2010)

O que chamamos de processo de raspagem da web é uma espécie de mineração de dados realizadas nos sites pela Internet. Esse processo objetiva buscar dados de sites diferentes e não estruturados e transformá-los em uma estrutura compreensível. Os dados e as informações transformam-se em planilhas, banco de dados ou arquivos de valores separados por vírgula (CSV) (MATTOSINHO, 2010). Com isso, a técnica *Web Scraping* possibilita que os dados finais sejam dados abertos, que podem ser manipulados por máquinas. Segundo (MATTOSINHO, 2010), alguns exemplos do que pode ser obtido por meio desta técnica são dados de preços de itens, preços de ações, preços de mercado, além de relatórios distintos. São dados muito relevantes, visto que se tornam conteúdos que agregam valor ao negócio, auxiliando para a tomada de decisões diversas.

Figura 3 - Arquitetura Básica de *Web Scraping* (MATTOSINHO, 2010).

"*Web Scraping* é uma técnica usada para cortar informações de páginas da web com base em rotinas de script." (MATTOSINHO, 2010)

Como são compostas páginas da Web? Inicialmente, os documentos eram escritos em *Hypertext Markup Language* (HTML). Depois, passaram para XHTML, linguagem que é baseada em XML. Na realidade, os documentos Web estão estruturados na conhecida árvore DOM (*Document Object Model)* e o HTML tem como objetivo especificar o formato do texto exibido pelos navegadores Web. Veja figura abaixo:

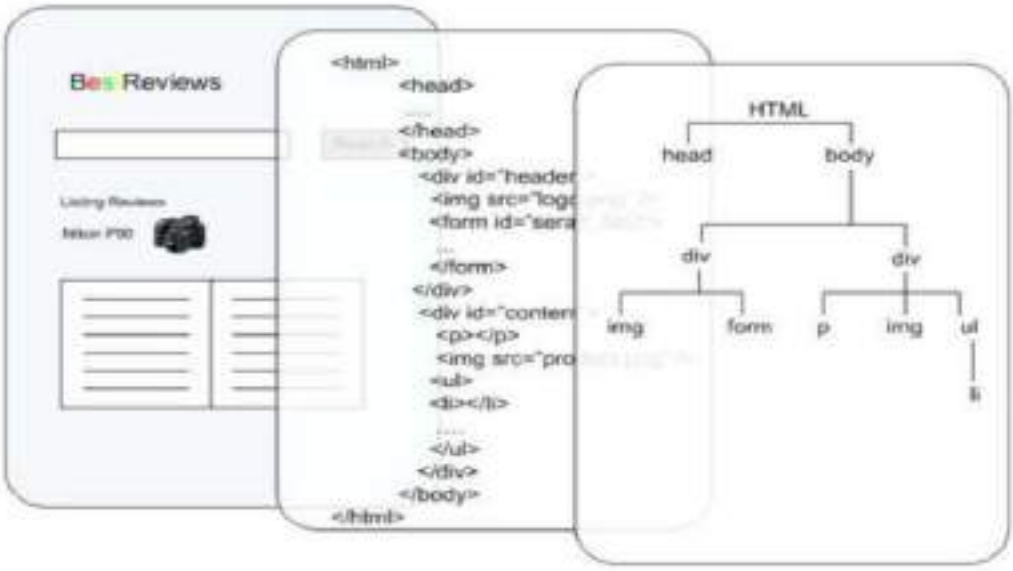


Figura 4 - Três perspectivas de documento da web (HERNÁNDEZ, et al., 2015).

Segundo (HERNÁNDEZ, et al., 2015) um programa conhecido como orquestrador é o que orquestra, organiza e executa as solicitações para o navegador. Assim se dá o processo de raspagem de dados.

2.5. INTELIGÊNCIA ARTIFICIAL

A primeira vez que surgiu o termo em título de evento foi em 1956. Isso ocorreu no *Dartmouth Summer Research Project on Artificial Intelligence*, realizado no *Dartmouth College* em *Hanover*, EUA, 1956, com o apoio da Fundação *Rockefeller*. Este evento foi liderado por Claude Shannon, Nathaniel Rochester, Marvin Minsky e John McCarthy (RUSSELL; NORVIG, 2009).

De acordo com (RICH, 1994), os estudos relacionados à melhoria na realização de tarefas realizadas por máquinas - que antes eram executadas por humanos - referem-se a Inteligência Artificial.

Segundo (RUSSELL; NORVIG, 2013) Inteligência Artificial pode ser classificada em quatro segmentos: 1) Pensando como um humano; 2) Pensando racionalmente; 3) Agindo como seres humanos; 4) Agindo racionalmente conforme tabela abaixo:

Pensando como um humano	Pensando racionalmente
"O novo e interessante esforço para fazer os computadores pensarem (...) *máquinas com mentes*, no sentido total e literal." (Haugeland, 1985) "[Automatização de] atividades que associamos ao pensamento humano, atividades como a tomada de decisões, a resolução de problemas, o aprendizado..." (Bellman, 1978)	"O estudo das faculdades mentais pelo uso de modelos computacionais." (Charniak e McDermott, 1985) "O estudo das computações que tornam possível perceber, raciocinar e agir." (Winston, 1992)
Agindo como seres humanos	**Agindo racionalmente**
"A arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas." (Kurzweil, 1990) "O estudo de como os computadores podem fazer tarefas que hoje são melhor desempenhadas pelas pessoas." (Rich and Knight, 1991)	"Inteligência Computacional é o estudo do projeto de agentes inteligentes." (Poole *et al*., 1998) "AI... está relacionada a um desempenho inteligente de artefatos." (Nilsson, 1998)

Figura 5 - Classificação de inteligência Artificial (RUSSELL; NORVIG, 2013)

O principal objetivo da IA é desenvolver inteligência similar à humana em máquinas. Esse feito pode ser realizado por meio de algoritmos de aprendizagem que tentam imitar como o cérebro humano aprende (DAS et al., 2015)

De acordo com (RUSSELL; NORVIG, 2013), as classificações 1 e 2 se relacionam a processos de pensamento e raciocínio, as de 3 e 4 são referentes a comportamento. Os conceitos 1 e 3 avaliam o êxito relacionado à fidelidade ao desempenho humano. As opções 2 e 4 estimam um conceito ideal de inteligência racional. Um sistema é racional se "faz a coisa certa", dado o que ele possui como conhecimento.

2.6. CHAT BOT

O *Chat Bot* é definido como um serviço alimentado por regras e, também, por inteligência artificial, que possibilita ao usuário humano utilizar a interface de chat para interagir (SCHLICHT, 2016).

É denominado *Chat Bot* esse sistema automatizado de conversa. Ele viabiliza as interações e a comunicação flui por linguagem natural, mesmo que no fluxo da conversa estejam sendo utilizados recursos de Inteligência Artificial e *Machine Learning.* É dessa

forma que o *Chat Bot* absorve novos conceitos a cada interação nova do usuário. O processo é automatizado e aprimorado a cada interação (DIALOGFLOW, 2020a).

Atualmente o usuário está tão familiarizado às ferramentas e softwares automatizados que chega a reconhecer a interação existente entre homem e máquina em várias ferramentas que utiliza no dia a dia. O *Chat Bot*, por exemplo, é um dos programas usuais para atendimento e resolução de problemas dos usuários. "Um *Chat Bot* é um sistema de computador cuja interface é fornecida por interação por meio de um diálogo em linguagem natural, simulando uma conversa humano-humano." (PRISCO et al,. 2019)

Vale lembrar, todavia, que os *Chat Bots* não são recursos novos. Em realidade, essas ferramentas vêm sendo utilizadas há décadas para diversas aplicações. O sistema ELIZA, de Joseph Weizenbaum, foi uma das primeiras iniciativas para construir sistemas de computador cuja interface que simulava um diálogo. Nessa aplicação foram adotadas técnicas de Inteligência Artificial para simularem um psiquiatra (ABUSHAWAR, ATWELL. 2015).

Outro exemplo é o chat de bate-papo na Internet, ou redes sociais, que ganhou espaço no dia a dia para interação entre humanos. O sucesso da ferramenta também se dá devido ao seu baixo consumo de largura de banda e consumo de dados. Para (NISHIHARA et al,. 2017), os *Bots* de bate-papo são programas de computador que funcionam online. Exemplos deles são os serviços de comunicação, ferramentas de chat, sistema de *microblog* e serviços de rede social.

A capilaridade e o poder de interação dos *Chat Bots* na sociedade e nas redes sociais atingiu, em março de 2017, um marco. Entre 9% e 15% dos usuários ativos do Twitter são *Bots* (29-49 milhões de contas de 328 milhões, de acordo com (VAROL et al,. 2017). É um número considerável e nos dá uma dimensão da quantidade de *Bots* que têm atuado nessa rede social específica.

Tendo em vista essa interação entre humanos e máquinas por meio dos *Chat Bots*, um chat que possibilite interação dos cidadãos para participação popular no que se refere aos gastos públicos dos vereadores da sua cidade pode aproximar os cidadãos da atuação de seus representantes e, possivelmente, pode fomentar a atuação e o controle social.

3. TRABALHOS RELACIONADOS

Neste capítulo, alguns exemplos de trabalhos de áreas afins foram selecionados por serem correlacionados ao tema e ao problema de pesquisa em si. São trabalhos que revelam experiências anteriores e os resultados aplicados para tratar estas questões.

3.1. CHATTERBOT CRIOULO: PROPOSTA DE UM CONVERSADOR QUILOMBOLA DAS TERRAS DE PRETO DO TERRITÓRIO LITORAL SUL – BA

A proposta do *Chatterbot Crioulo* teve como objetivo principal desenvolver um agente computacional de conversação para auxiliar na divulgação das comunidades quilombolas do território sul baiano. O objetivo secundário foi relacionar o patrimônio histórico-cultural das comunidades quilombolas do território litoral sul – BA, conforme (MENEZES, 2016).

Utilizaram nessa pesquisa as seguintes metodologias: de participante, computacional, qualitativa, aplicada e tecnológica. Delimitaram o estudo no litoral sul Bahia, tendo em vista que esse tipo de abordagem na população constrói uma identidade com base no sentimento de pertencimento ao lugar.

Foi utilizada a linguagem AIML na construção do *Chat Bot* para interagir com usuários sobre as informações peculiares e atualizadas das comunidades quilombolas. Ele permite a criação de diálogos entre usuários e máquinas por meio de escrita, porém a lógica da linguagem AIML permite que o *chatterbot* receba informações em forma de texto, imagem, áudio, dentre outras.

O estudo foi classificado como prática ou aplicada, tendo em vista que é baseado em solução de problemas em que ocorrem “a transformação de resultados em bens ou serviços”.

A pesquisa contribuiu para desenvolvimento de diversas situações que geraram mobilizações e atividades de intervenção social nas comunidades quilombolas, como mapeamento de novas comunidades quilombolas e de comunidades que ainda não são reconhecidas pelos órgãos governamentais, mas que possuem um histórico de comunidades com formações quilombolas. Tentaram, também, dar amplitude ao patrimônio histórico-cultural dessas comunidades tradicionais por meio do trabalho.

Como futuros projetos foram sugeridos trabalhos que poderão ser desenvolvidos, a fim de aprimorar o *chatterbot*. Isto deixa espaço para os programadores modificarem o código fonte do motor de busca utilizado, a fim de construir um *chatterbot* totalmente funcional e compatível com as necessidades de cada tipo de serviço.

3.2. ESTUDO DE CASO "OPERAÇÃO SERENATA DE AMOR": A ANÁLISE DE BIG DATA NO COMBATE À FESTA DOS GASTOS PÚBLICOS.

O estudo de caso da Serenata do Amor é uma investigação das dificuldades relacionadas à identificação de gastos dos Deputados Federais e o advento do desenvolvimento de um algoritmo para mineração e interpretação dos dados (RODRIGUES; FONTES, 2018). O projeto denominado Operação Serenata de Amor oferece serviço de consulta de gastos dos deputados para acompanhamento da população. Idealizado pelo Cientista de dados Irio Irineu Musskopf, o nome do projeto foi inspirado na vice-primeira ministra Mona Sahlin da Suécia, em 1990, quando ela renunciou após pressão popular ao descobrir-se que ela havia comprado uma barra de chocolate Toblerone com o cartão de crédito corporativo/governamental (RODRIGUES; FONTES, 2018).

Esse estudo de caso relata como grandes quantidades de dados são disponibilizados em formatos não amigáveis para humanos, como por exemplo, os gastos de deputados federais da Câmara Legislativa. Foi analisada a diferença entre algo disponível e acessível no segmento da informação, lembrando que a afirmação de um dado disponível na internet não o classifica como acessível, conforme já exposto pelo (EAVES, 2009) nas "3 leis de dados abertos".

Este trabalho também demonstra como a fiscalização é uma árdua tarefa, tendo em vista o grande volume de dados que são disponibilizados conforme menção de (MUSSKOPF, 2016), que aborda o exemplo dos gastos dos deputados federais disponibilizados no website nos formatos XML, JSON, CSV e XLSX, divididos por ano, de 2009 a 2017 (parcial), disponíveis para download. Os arquivos do formato .XML têm o tamanho de armazenamento em aproximadamente 315,6Mb a 462,2Mb; os do tipo .JSON de 175,4Mb a 255,4Mb; os do tipo .CSV de 52,3Mb a 76,3Mb e, os do tipo XLSX5 , de 33,5Mb a 47,8Mb. Vale ressaltar que informações anteriores a 2009 estão disponíveis em um arquivo .XML único com 6,2Gb.

Em referência ao Método do Projeto “Serenata do Amor” nos é apresentado como o projeto foi elaborado no que se refere à extração dos dados para manipulação e agrupamento dos dados para formar os sete *Datasets*. Nessa etapa são apresentadas as principais dificuldades que envolvem *reimbursements*, que é de fato o *Dataset* de recibos escaneados e fornecidos pela Câmara dos Deputados. Para realizar este procedimento, foi utilizado o OCR (*Optical Character Recognition* – Reconhecimento Ótico de Caracteres), com consultas via API Cloud Vision da Google.

Foram elencados quatro tipos de desafios encontrados nesse tipo de extração. 1° As dificuldades que envolvem a análise de recibos com escrita manual. 2° Alguns recibos são digitalizados sem cuidado, o que dificulta a visualização do conteúdo. 3° Algumas digitalizações têm qualidade muito baixa e / ou o recibo fica desbotado por ser papel térmico e isso dificulta o computador identificar o texto nele escrito. 4° A plataforma Cloud Vision do Google possui limitação de 1.000 acessos gratuitos em sua API a cada mês; desta forma, o processo é executado em múltiplas instâncias, em paralelo, tornando-se bastante lento.

Nos é apresentado, também, o Jarbas, uma interface gráfica que permite a busca e filtro dos resultados; essa visualização dos dados foi resultado da equipe do projeto que desenvolveu uma API pela comunicação entre diversos códigos.

Segundo (RODRIGUES; FONTES, 2018), outra criação foi a “robô” Rosie. Ela foi desenvolvida em *Scikit-learn*, biblioteca de aprendizado de máquina em *Python*. Nesse projeto utilizaram-se modelos de aprendizado supervisionado de classificação (ABRAHAM *et al*, 2014), que aprende a partir de exemplos rotulados das saídas de um teste e que devem ser produzidos para uma entrada (HACKELING, 2014). E, por último, utilizaram o *k-means* de aprendizado não supervisionado.

Como conclusão, é apresentado como eficaz o uso de big data no combate à corrupção com o projeto Serenata do Amor, pois foi possível identificar um volume considerável de irregularidades, mesmo havendo espaço para a evolução das técnicas de big data, aumento do conjunto de *datasets*, criação de novos classificadores ou novos modelos de *machine learning* para aumentar o controle social. O modelo *open source* já reúne diversos colaboradores e é um convite para que todos aqueles que desejam a construção de um país

mais honesto e eficiente possam contribuir na evolução da Rosie e Jarbas (RODRIGUES; FONTES, 2018).

4. METODOLOGIA

O trajeto dessa pesquisa percorrerá o caminho metodológico exploratório de estudo do *Web Scraping* e dados governamentais. Esse estudo busca os métodos existentes na extração de dados governamentais na web, para tornar os dados desestruturados de portais de transparência em dados abertos.

4.1. PESQUISA BIBLIOGRÁFICA

O tema de *Web Scraping* é o ponto central da metodologia adotada nesta dissertação que se propõe a conhecer trabalhos de *Web Scraping*, visando criar um método baseado em estudos realizados. A metodologia está pautada em vasta pesquisa bibliográfica e documental.

De acordo com (GIL, 2007, p. 44), os exemplos de pesquisa bibliográfica são sobre investigações, sobre ideologias, ou aqueles que se propõem a analisar a um determinado problema sobre diferentes prismas.

Por isso, a pesquisa documental lança mão de fontes das mais diversas sem nenhum tratamento analítico. O trabalho é feito por meio da pesquisa de materiais que ainda podem ser reelaborados de acordo com os objetos da pesquisa. (GIL, 2008, p. 45)

Esse tipo de pesquisa não busca generalizar os resultados que podem ser encontrados nas execuções de métodos de *Web Scraping* com dados governamentais, sendo que pode haver peculiaridade de acordo com a origem dos dados de cada município. Também pode depender bastante do conhecimento empírico dos autores em relação ao tema, já que se pode utilizar métodos, ferramentas e *frameworks* distintos para chegar aos resultados.

4.2. ANÁLISE QUALITATIVA

A análise qualitativa aplicada foi realizada por meio de levantamento bibliográfico e documental. De caráter exploratório, teve como objetivo levantar informações sobre os métodos de *Web Scraping* aplicados a dados governamentais.

"A abordagem qualitativa pode ser definida como aquela que se fundamenta principalmente em análises qualitativas, sendo assim caracterizada de maneira geral pela não utilização de instrumental estatístico na análise dos dados" (BARDIN, 2011).

4.3. PESQUISA APLICADA

A natureza dessa pesquisa será a aplicação prática, visto que o objetivo é desenvolver um método com resultado aplicável para a sociedade. Partindo inicialmente da pesquisa bibliográfica e documental, chega-se ao desenvolvimento de um método de *Web Scraping* com resultado aplicável para transformar os dados do portal de transparência, da Câmara Municipal de Belo Horizonte (CMBH), em dados abertos. De acordo com o autor (SILVEIRA; CÓRDOVA, 2009), a pesquisa aplicada visa investigar, gerar conhecimentos para aplicação prática e comprovar ou rejeitar hipóteses.

A pesquisa se limitada a descoberta de dados a respeito dos gastos dos vereadores no site de transparência da Câmara municipal de Belo Horizonte

5. CHAT BOT SUMÉ

O nome Sumé foi atribuído ao *Chat Bot* como forma de tributo aos índios brasileiros por serem meus antepassados e uma parte da base cultural deste país. A escolha deste nome Sumé também teve alinhamento com a representação que esse personagem possui na cultura Tupi-Guarani com a organização social, tendo em vista que essa narrativa social vem ao encontro da proposta desta dissertação.

"Sumé remete a crença atribuída a cultura Tupi-Guarani em que se cultuavam esse ser considerado um herói civilizador a quem os Tupis da Costa e outros grupos atribuíam, em especial, o conhecimento da agricultura e sua organização social" (CLASTRES, 1978).

Baseado nas afirmações de (NÓBREGA, 1988):

> Sim, a doutrina cristã foi transmitida aos índios na Antiguidade pelo apóstolo São Tomé, o que também foi afirmado pelo Padre Vieira e foi acolhido nos séculos XVI e XVII. No Brasil, o que se fez foi interpretar o mito de Sumé como uma narrativa da vinda do apóstolo São Tomé para a América, principalmente pelo conhecimento que eles tinham da agricultura e de sua organização social (NÓBREGA, 1988).

Em relação ao trabalho relacionado "Operação Serenata do Amor", o ponto principal em comparação com o *Chat Bot* Sumé está na geração dos *Dataset* de gastos públicos. O projeto desta dissertação tem a proposta de realizar um método para extração e criação de *Dataset* em formato aberto .CSV com os dados dos gastos públicos no site de transparência da câmara municipal de Belo Horizonte. Estes dados por sua vez estão desestruturados e não estão em formato aberto. No caso dos dados *Dataset* da Serenata do Amor, os mesmos já estão disponíveis para *download* e se encontram estruturados em formato aberto CSV e em outros formatos como XML, JSON e XLSX.

O *Chat Bot* Sumé é uma solução fundamentada em dados abertos, tendo em vista que sua base de dados será o resultado do método de Web Scraping aplicado aos dados de custeio parlamentar no portal de transparência da CMBH.

5.1. DADOS DE CUSTEIO PARLAMENTAR DA CMBH

Paralelamente aos mecanismos de controle tradicionais, a Câmara Municipal de Belo Horizonte (CMBH) promove a ampliação das ações de divulgação dos gastos empreendidos pelo próprio órgão. Todavia, como exposto na introdução, os dados referentes aos gastos não são disponibilizados em formato aberto, conforme exigido pela LAI.

Dessa forma, a CMBH não está em conformidade com a LAI, fato que acontece em proporção maior em municípios, como neste caso analisado, em Belo Horizonte. Já os órgãos da união estão mais avançados nesse sentido, disponibilizando diversas bases de dados abertos.

Podemos apontar que há três tipos de gastos disponibilizados pela CMBH atualmente. Eles são referentes aos gastos de gabinete dos Vereadores, ao Custeio Parlamentar - em formato de menu do elemento select HTML -, e aos Serviços Postais e Gastos com Telefonia

- em formato PDF. A figura abaixo ilustra como acessar o menu de dados dos gastos de Custeio Parlamentar.

Figura 6 - Custeio Parlamentar

O foco central deste trabalho de pesquisa é desenvolver um método de *Web Scraping* para a extração dos dados de custeio parlamentar. Após manipular os dados, o intuito é gerá-los como dados em formato aberto, ou seja, no formato .CSV disponível para acesso na WEB. A linguagem utilizada para a extração será a *Python* através do *Web Scraping*, tendo como suporte o auxílio de algumas bibliotecas, como por exemplo, a *Beautiful Soup* (BROUCKE; BAESENS, 2018).

5.2. PYTHON

"Python é uma linguagem de programação interpretada, interativa e orientada a objetos. Ela incorpora módulos, exceções, tipagem dinâmica, tipos de dados dinâmicos de nível muito alto e classes." (PSF, 2001a).

A origem do nome dessa linguagem foi baseada em uma série de comédia da BBC nos anos 1970. Guido van Rossum desejava um nome objetivo, curto, único e misterioso,

com isso se inspirou em roteiros publicados desta série "*Monty Python's Flying Circus*", denominando a linguagem como Python (PSF, 2001a). Python é uma linguagem bastante manejável como apresentado na documentação da linguagem:

> Ele oferece suporte a vários paradigmas de programação além da programação orientada a objetos, como a programação procedural e funcional. Python combina poder notável com sintaxe muito clara. Ele tem interfaces para muitas chamadas de sistema e bibliotecas, bem como para vários sistemas de janela, e é extensível em C ou C ++. Também pode ser usado como uma linguagem de extensão para aplicativos que precisam de uma interface programável. Finalmente, o Python é portátil: ele roda em muitas variantes do Unix, incluindo Linux e macOS, e no Windows (PSF, 2001a).

Além de Python, há várias linguagens de programação que poderiam ser usadas nativamente para *web scraping*, incluindo Java e Ruby. Numerosas ferramentas, plug-ins de navegador e APIs estão disponíveis para facilitar o *Web Scraping* (HADI; Al-ZEWAIRI, 2017). Optou-se por escolher Python na versão 3.6 para este projeto, visto que é uma linguagem comumente utilizada em pesquisas científicas por possuir licenciamento de código aberto. Além disso, o autor da dissertação é especialista na linguagem.

Para alcançar a extração, manipulação e formatação dos dados da maneira desejada, foram utilizadas soluções para desenvolvimento do método, dentre elas, se encontram a IDE Pycharm e bibliotecas de Python.

5.2.1. PyCharm

Neste trabalho foi utilizada a versão PyCharm 2020.2.3 (Community Edition), programa de computador para desenvolvedores profissionais. A *Integrated Development Environment (IDE*) ou Ambiente de Desenvolvimento em *Python* reúne características capazes de agilizar o desenvolvimento de um software. Há inúmeras funções como: autocompletar código inteligente; efetuar inspeções de código; alcançar realce de erros em tempo real, além de efetuar correções rapidamente junto a refatorações de códigos automatizadas e recursos de navegação avançados (JETBRAINS, 2020).

5.2.2. Selenium Webdriver

O *API Selenium Python WebDriver* foi o sistema adotado para criar o método de *Web Scraping*. Por meio dele foi possível interagir com o navegador e extrair as informações necessárias.

"*Selenium* é um projeto abrangente para uma variedade de ferramentas e bibliotecas que permitem e dão suporte à automação de navegadores do mercado através do uso de *WebDriver."* (SELENIUM, 2020)

WebDriver é uma API e protocolo para definição de interface para controlar os navegadores Web. O processo se dá da seguinte forma: cada navegador tem como suporte o *driver*, que é uma implementação *WebDriver* específica. Esse *driver* é o componente responsável por controlar a comunicação entre o Selenium e o navegador. (SELENIUM, 2020)

Quanto ao *Selenium*, trata-se de ferramenta utilizada para escrever testes automatizados de websites. Sua função é imitar o comportamento de um usuário real e, portanto, faz a interação com o HTML da aplicação. (SELENIUM, 2020)

5.2.3. Beautiful Soup

Para extrair dados de arquivos HTML e XML, *a Beautiful Soup* é a biblioteca *Python* utilizada. É uma ferramenta que otimiza o dia a dia, economizando horas de trabalho dos programadores ou, até mesmo, dias. A *Beautiful Soup* funciona como um método de análise inteligente que identifica maneiras idiomáticas de navegar, pesquisar e modificar a árvore de análise. (CRUMMY, 2004).

Um exemplo de construção pode ser observado a seguir como se pode passar uma *string* ou um identificador de arquivo aberto (CRUMMY, 2004):

```
1  from bs4 import BeautifulSoup
2  with open("index.html") as fp:
3      soup = BeautifulSoup(fp, 'html.parser')
4  soup = BeautifulSoup("<html>a web page</html>", 'html.parser')
```

Figura 7 – Exemplo de código *Python Beautiful Soup* (CRMMY, 2004).

Foi utilizado *Beautiful Soup* em conjunto de outras ferramentas para captar os dados referente aos gastos de Custeio Parlamentar.

5.2.4. Time

O módulo *Python Time* é aquele que fornece várias funções relativas ao tempo. É um módulo que está sempre disponível, entretanto, muitas funções dele não ficam disponíveis em todas as plataformas (PSF, 2001b).

Para este método será necessário utilizar o módulo *Time*, visto que em alguns momentos o código deverá aguardar alguns segundos para execução. Assim, a função tem o papel de disponibilizar tempo hábil para que o robô *Selenium* possa navegar nos conteúdos selecionados sem a ocorrência de conflitos entre as instruções e a coleta de dados.

Calculado por segundos, o módulo em questão chama-se *time.sleep.* Nele é possível suspender - por determinado número de segundos - a execução do *thread* de chamada. “O argumento pode ser um número de ponto flutuante para indicar um tempo de sono mais preciso.” (PSF, 2001b)

Segundo (PSF, 2001b) o tempo de suspensão real pode ser menor do que o solicitado. Essa ocorrência se dá porque qualquer sinal capturado encerrará a *sleep()* execução seguinte da rotina de captura do sinal.

5.2.5. Pandas

A biblioteca Pandas foi utilizada para manipular os dados em CSV dos resultados de *Web Scraping* realizados pelo *Selenium Web Driver* tendo em vista que os dados são extraídos da fonte e precisam ser manipulados para gerar informação e, posteriormente, conhecimento.

A biblioteca Pandas, segundo (MCKINNEY, 2012), disponibiliza estruturas de dados de alto nível. São várias funções cujo intuito é tornar ágil o trabalho com dados estruturados ou tabulares. Além da agilidade, com a Pandas, o processo fica mais fácil e expressivo,

possibilitando criação de *DataFrame*. Esta nada mais é do que uma estrutura de dados tabular, com rótulos de linha e coluna, *Series* e objeto de matriz rotulado unidimensional.

Utilizou-se a Pandas neste projeto visando a construção de uma base de conjunto de dados. É um *Dataset* como fonte de dados unificados do Custeio Parlamentar e, consequentemente, a base de dados do *Chat Bot Sumé*. Para a construção dos *Datasets*, as técnicas aplicadas são definidas como pré-processamento dos dados (HAN & KAMBER, 2006), conforme figura abaixo:

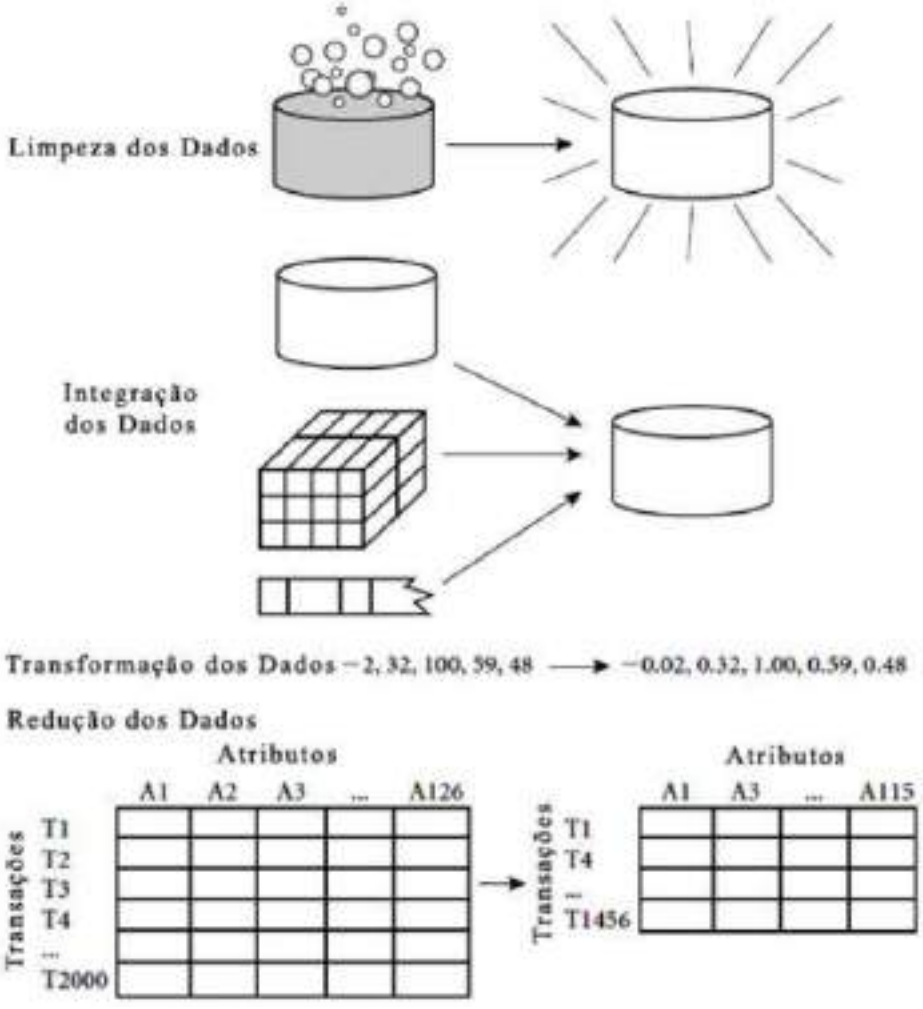


Figura 8 - Atividades do pré-processamento de datasets (Han & Kamber, 2006)

5.3. EXPERIMENTO DO MÉTODO WEB SCRAPING

Como a técnica Web Scraping é baseada em rotinas de *script* (MATTOSINHO, 2010), é necessário que, antes da construção do método, se realize a extração. Portanto,

inicialmente analisa-se o conteúdo Web e identificam-se quais dados precisam ser extraídos. Somente após a conclusão desses processos, partimos para a elaboração do *script*, lembrando que essa etapa de análise é uma boa prática para a extração de dados.

O experimento do método de *Web Scraping* foi dividido em cinco etapas que envolvem toda a extração e a manipulação de dados que foram necessárias.

5.3.1. Scraping Vereador

Nessa primeira etapa deverá ser realizado o acesso ao portal da CMBH, com o objetivo de analisar e identificar os vereadores que compõem a câmara.

Para analisarmos os dados e o conteúdo de uma página na *Web*, os desenvolvedores podem utilizar no navegador a opção de "ferramentas desenvolvedor" para Google Chrome ou "*Web Developer - Inspetor*" para o *Firefox*. Após selecionar esta opção conseguimos verificar todos elementos de HTML que compões a página Web.

No portal da CMBH na aba de "Vereadores", ao fazermos a pesquisa de vereador e partido ao qual ele pertence, precisamos identificar através da opção de "ferramentas desenvolvedor" quais os dados representam o nome do vereador e o partido, ou seja, qual "id" eles assumem no HTML. No caso, o "id" que contempla o conteúdo com nome e partido do vereador é "view-content" que está dentro de uma div class HTML, conforme a figura abaixo.

Figura 9 - Div “view-content” nome do partido e vereador.

Após realizar a identificação da ”view-content”, é possível elaborar um código automatizado por um robô do *Selenium Web Driver* para capturar esses dados que estão dentro do id ”view-content”. Observe como está escrito no código a seguir:

```
import time
from bs4 import BeautifulSoup
from selenium import webdriver

driver = webdriver.Firefox(executable_path="C:/Users/Wandel/PycharmProjects/pythonProject/Driver/geckodriver.exe")
driver.get("https://www.cmbh.mg.gov.br/vereadores")
time.sleep(3)
dados_0 = driver.find_element_by_class_name("view-content")
html_0 = dados_0.get_attribute("innerHTML")
soup = BeautifulSoup(html_0, 'html.parser')
resultado = (soup.get_text())
print(soup.get_text())
with open("Vereador.csv", "w", encoding="utf-8") as f:
    s = "".join(resultado)
    f.write(s + "\n")
time.sleep(2)
driver.close()
```

Figura 10 – Código *Scraping* Vereador

O resultado da extração após interação do robô é um arquivo nomeado "Vereador.csv" e sua apresentação será conforme figura abaixo:

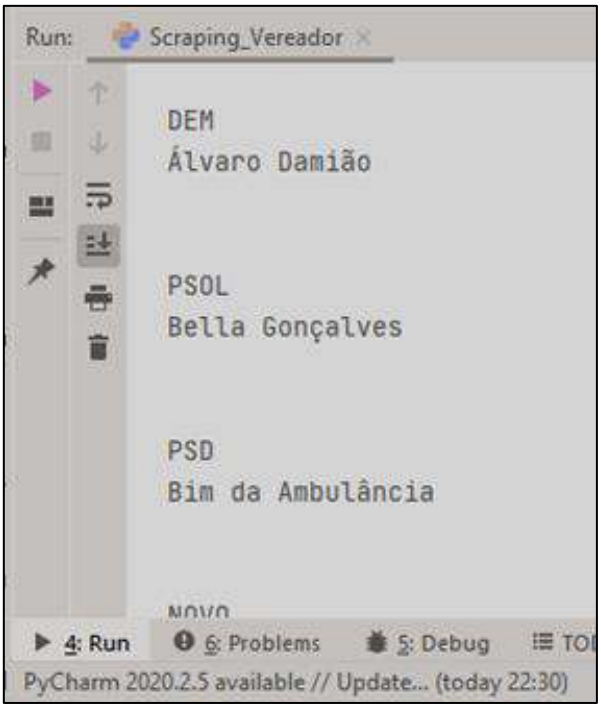


Figura 11 - Resultado da extração div view-content.

Esse método de *Web Scraping* foi utilizado para extrair somente o nome do vereador e seu partido, pois, posteriormente essas informações serão usadas para agrupar aos dados do

Custeio Parlamentar, tendo em vista que nesse último não temos informação do partido do vereador.

5.3.2. Scraping Custeio Parlamentar

Nesse segundo momento precisará ser acessado a página de Custeio Parlamentar para realizar esta análise (https://www.cmbh.mg.gov.br/transparencia/vereadores/custeio-parlamentar).

Ao fazermos a pesquisa de gasto mensal do Custeio Parlamentar é necessária interação com a página, por se tratar de um elemento select HTML dentro de uma div. Dessa forma, é preciso identificar através da opção de “ferramentas desenvolvedor” qual “id” é referente ao mês para registrá-lo. Somente assim o robô do *Selenium WebDriver* conseguirá navegar pelas marcações HTML do conteúdo Web da página e selecionará os dados registrados a serem coletados. O “id” encontrado referente à seleção do mês está nomeado como ”data”, id=”data”, como apresentado abaixo:

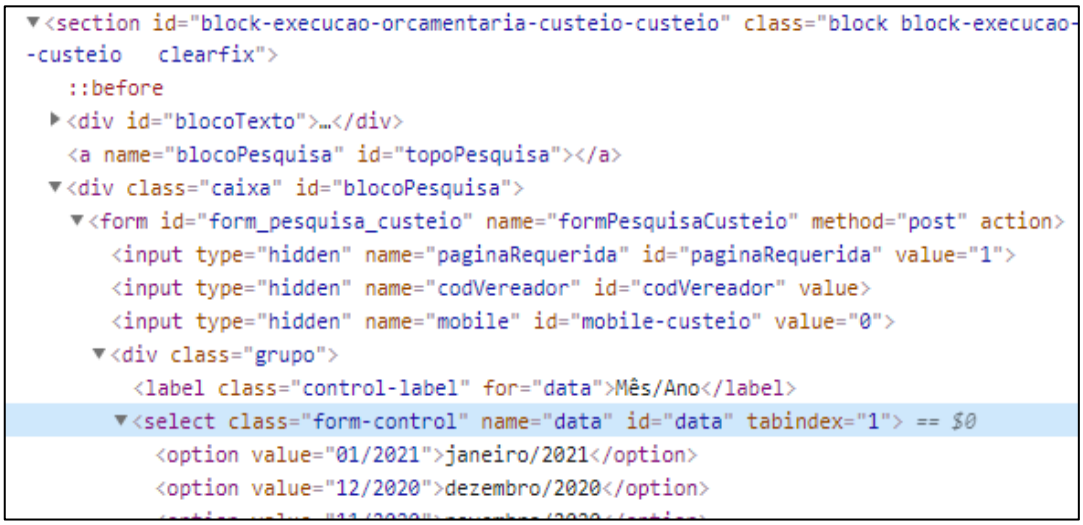


Figura 12 - Id “data” Custeio Parlamentar.

Após essa etapa precisamos identificar o “id” que contém o resultado da pesquisa, este em questão foi identificado e está nomeado como id="resultadoPesquisa_custeio" conforme imagem abaixo:

Figura 13 - Resultado da extração Id “data”.

Munido desses dados, partimos para a elaboração do código que irá manipular os dados - através da interação com o robô do *Selenium WebDriver* - e realizará a extração. O robô simulará um humano clicando nos meses representados em “Mês/Ano”, depois, em “Pesquisar” e, assim, captará os dados de resultado da pesquisa. A simulação do resultado dessa etapa está conforme imagem abaixo:

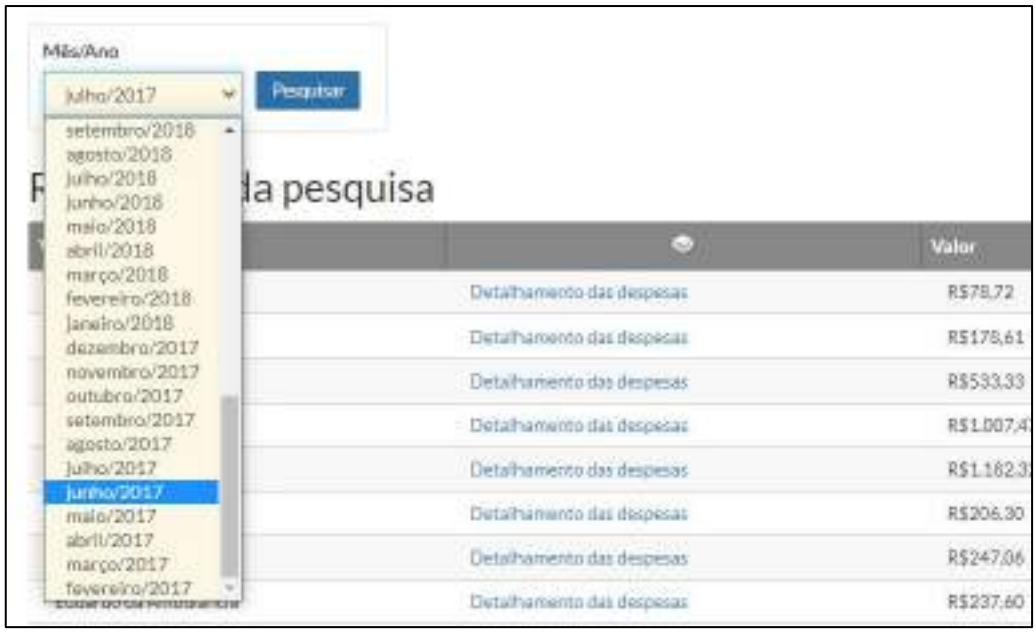


Figura 14 - Menu de seleção dos gastos mensais.

Após o robô captar o resultado da pesquisa, será armazenado na memória através da estrutura de repetição *for*, assim a *div*, que possui a tabela dentro da variável *dados*, *assim.*

Os dados são armazenados até a próxima repetição *for*, conforme código abaixo do Custeio Parlamentar.

```
import time
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver.support.ui import Select

driver = webdriver.Firefox(executable_path="C:/Users/Wendel/PycharmProjects/pythonProject/Driver/geckodriver.exe")
driver.get("https://www.cmbh.mg.gov.br/transparencia/vereadores/custeio-parlamentar")
time.sleep(3)
clicar0 = driver.find_element_by_id("data").click()
time.sleep(1)
mes = ["05/2017","06/2017","07/2017","08/2017","09/2017","10/2017","11/2017","12/2017",
       "04/2018","05/2018","06/2018","07/2018","08/2018","09/2018","10/2018","11/2018","12/2018",
       "02/2019","03/2019","04/2019","05/2019","06/2019","07/2019","08/2019","09/2019","10/2019","11/2019","12/2019",
       "01/2020","02/2020","03/2020","04/2020","05/2020","06/2020","07/2020","08/2020","09/2020","10/2020"]
for a in mes:
      select = Select(driver.find_element_by_id("data"))
    select.select_by_value(a)
    filtrar = driver.find_element_by_id("pesquisar-custeio")
    filtrar.click()
    time.sleep(2)
    dados = driver.find_element_by_id("resultadoPesquisa_custeio")
    html = dados.get_attribute("innerHTML")
    soup = BeautifulSoup(html, "html.parser")
    table = soup.select_one("table")
    headers = [header.text+";" for header in table.select("tr.success td")]
    print(headers)
    with open("Custeio_Parlamentar.csv", "a") as f:
        s = "".join(headers)
        f.write(s + "\n")
time.sleep(2)
driver.close()
```

Figura 15 – Código *Scraping* Custeio Parlamentar.

Neste código acima temos a coleta de todo o conteúdo HTML que está dentro da *div* "resultadoPesquisa_custeio".

O conteúdo HTML está dentro da variável “html”e, nesta etapa, iremos utilizar *Beautiful Soup* para fazer o parser desse HTML.

Dentro da variável “soup” temos o conteúdo da variável “html” retornado pelo robô do Selenium já com os dados convertidos pelo *Beautiful Soup*.

Foi utilizado o método select_one para buscar o elemento table dentro desse resultado da variável “soup”.

No conteúdo das tabelas temos várias vírgulas e espaços. Como será necessário converter esses dados para o formato CSV, é preciso definir um delimitador, neste caso foi definido o caractere ponto e virgula (;). Dando continuidade, é realizada a busca de todos os resultados dos elementos “tr”, que possuem a classe "success", e os elementos filhos cuja tag é "td" para armazená-los em linhas e colunas distintas.

A última etapa é a gravação dos dados que já foram parseados e serão gravados em .CSV, o formato aberto que atende aos requisitos da LAI. Essa gravação é realizada através do comento “with”.

Após a finalização da seleção, a gravação do arquivo é concluída e o robô *Selenium* encerra a seção do navegador. Abaixo, um exemplo do resultado da extração:

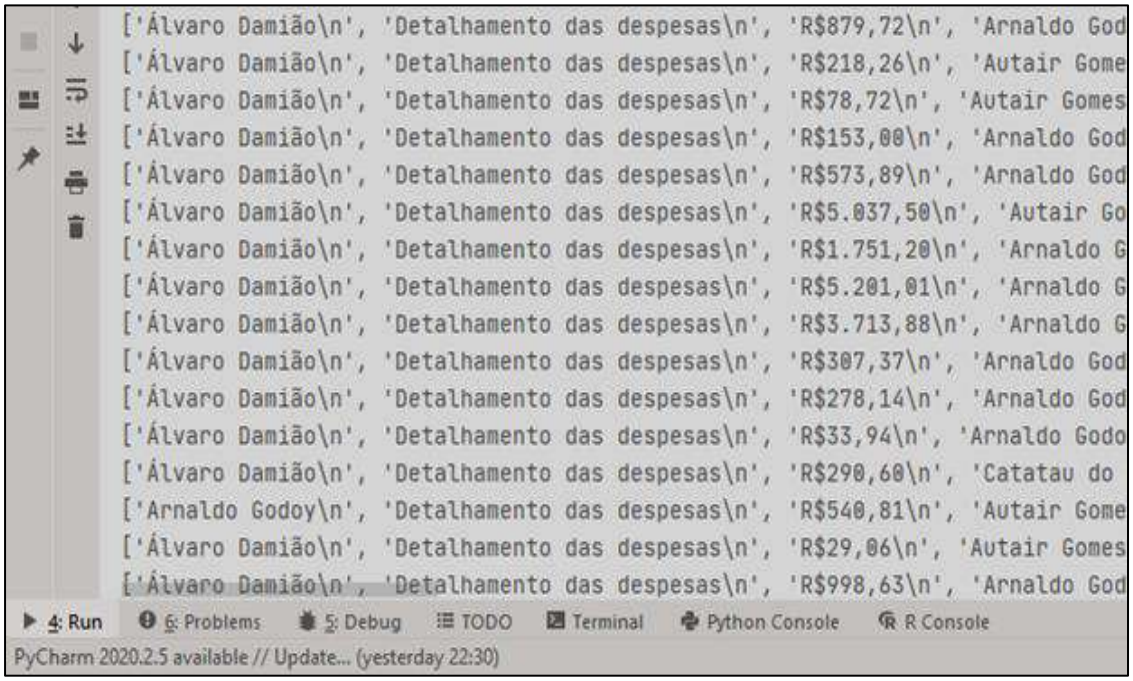


Figura 16 - Resultado da extração Custeio Parlamentar.

5.3.3. Manipulação de dados no Pandas

O método da biblioteca Pandas tem o *DataFrame* como grande destaque apontado pelo próprio site (PANDAS, 2020). É ele que torna rápido e eficiente o processo para

manipulação de dados com indexação integrada. Além disso, o *DataFrame* dispõe de ferramentas para ler e gravar dados, sendo estruturas de dados na memória e os diferentes formatos de arquivos, como o formato rápido HDF5, os CSV, arquivos de texto, Microsoft Excel e bancos de dados SQL.

Ao realizarmos manipulação utilizando Pandas conseguimos otimizar tempo e recursos devido a sua estrutura de indexação que facilita as interações:

> Um DataFrame representa uma tabela retangular de dados e contém uma coleção ordenada de colunas, cada uma das quais pode ser um tipo de valor diferente (numérico, string, booleano, etc.). O DataFrame possui um índice de linha e coluna; pode ser pensado como um dicionário de series, todos compartilhando o mesmo índice. Sob o capô, os dados são armazenados como um ou mais blocos bidimensionais em vez de uma lista, dicionário ou alguma outra coleção de matrizes unidimensionais (MCKINNEY, 2012).

5.3.3.1. DataFrame partido e vereador

A biblioteca Pandas foi utilizada para criar três *DataFrames:* "Partido_Vereador.csv", "Vereador_Custeio.csv" e o "dataset-bot.csv", *DataSet* do *Bot* Sumé. Cada *DataFrame* foi responsável por uma tarefa distinta para agrupamento de informações dos vereadores e seus gastos.

Foi necessário realizar essas etapas para que se pudesse agrupar os dados em somente um arquivo. Assim, eles serviram como fonte de dados para utilização no *Chat Bot*.

Tendo em vista o resultado da extração do arquivo "Vereador.csv ", será necessário manipular o conteúdo para que seja possível agrupar o nome do Vereador e seu Partido em colunas distintas. Para que fosse possível foi utilizada estrutura "while" com a propriedade "df.shape" para concatenar os dados e os agrupar em colunas separadas, tendo em vista que todos os dados se encontravam em uma única coluna.

```
import pandas as pd

df = pd.read_csv('Vereador.csv', header=None)
ii=0
df_saida = pd.DataFrame()
while ii < df.shape[0]:
    df_aux = pd.DataFrame([[df.values[ii][0].strip(),df.values[ii+1][0].strip()]])
    df_saida = pd.concat([df_aux,df_saida])
    ii+=2

df_saida = df_saida.rename(columns={0:'Partido',1:'Vereador'})

print(df_saida)

df_saida.to_csv('Partido_Vereador.csv', index=False, sep=';',encoding='cp1252')
```

Figura 17 – Código *DataFrame* partido e vereador

O resultado do código gerou um arquivo no formato CSV “Partido_Vereador.csv”, abaixo segue um exemplo do resultado após manipulação no Pandas.

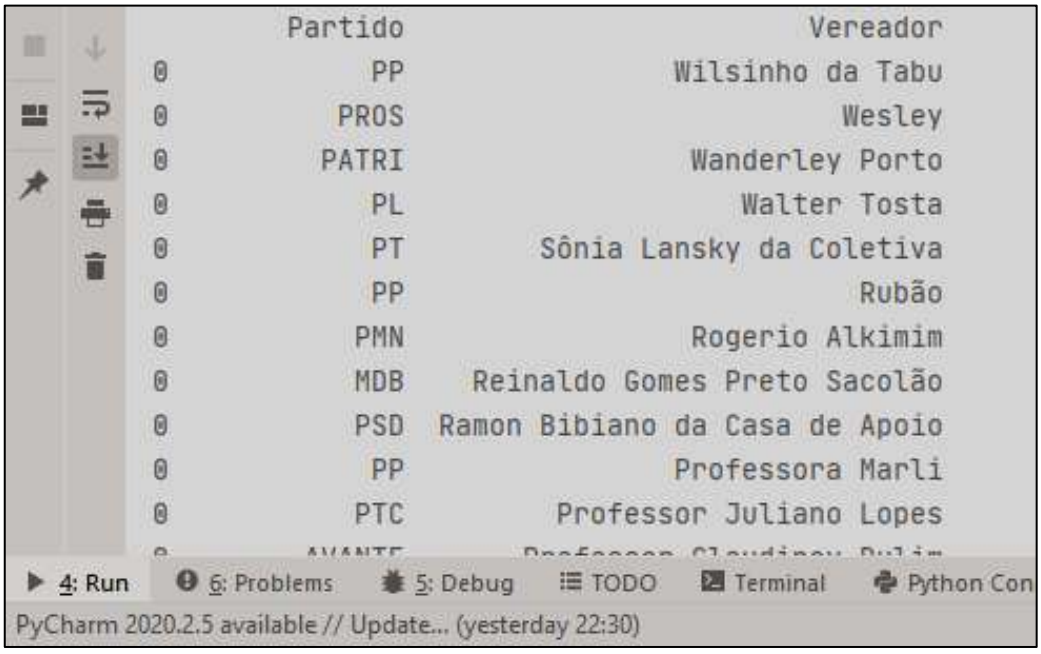


Figura 18 - Resultado *DataFrame* Partido e Vereador.

5.3.3.2. DataFrame vereador e custeio parlamentar

O segundo *DataFrame* teve como propósito manipular os dados “Custeio_Parlamentar.csv” para agrupá-los em colunas distintas: o Vereador e o Gasto do Custeio Parlamentar.

Foi utilizada a lógica semelhante ao do *DataFrame* Partido e Vereador, como pode ser visto no código abaixo:

```
import pandas as pd

df = pd.read_csv('Custeio_Parlamentar.csv', header=None, sep=';')
ii=0
df_saida = pd.DataFrame()
while ii < df.shape[0]:
    df_aux = pd.DataFrame([[df.values[ii][0].strip(),df.values[ii+2][0].strip()]])
    df_saida = pd.concat([df_aux,df_saida])
    ii+=3

df_saida = df_saida.rename(columns={0:'Vereador',1:'Gasto'})

print(df_saida)

df_saida.to_csv('Vereador_Custeio.csv', index=False, sep=';',encoding='cp1252')
```

Figura 19 – Código *DataFrame* vereador e custeio parlamentar.

O resultado do código gerou um arquivo no formato CSV “Vereador_Custeio.csv”, abaixo segue um exemplo do resultado:

Figura 20 - Resultado *DataFrame* Vereador e Custeio Parlamentar.

5.3.3.3. DataFrame Chat Bot Sumé

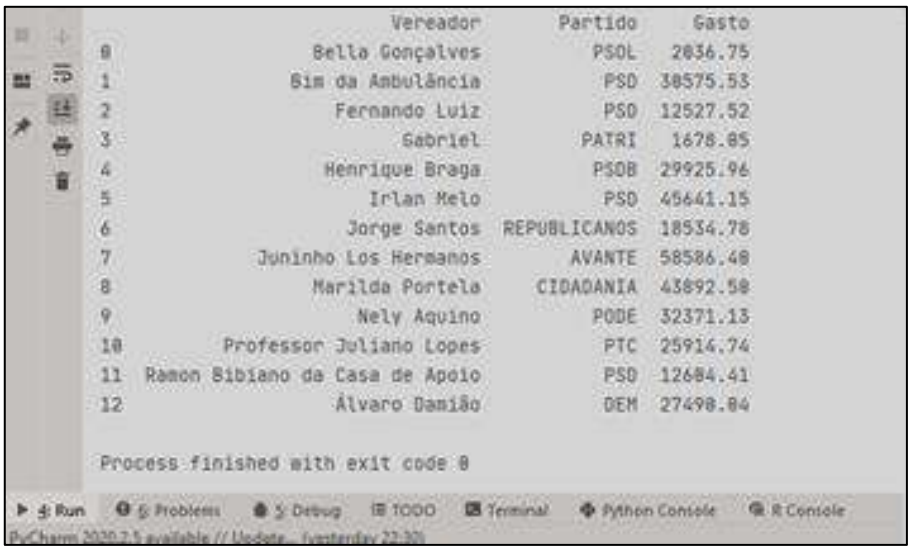
O terceiro DataFrame e a quinta etapa que concluem o *Web Scraping*, tiveram como propósito manipular os resultados dos *DataFrames* "Partido_Vereador.csv" e "Vereador_Custeio.csv", gerados após manipulação feita via biblioteca PANDAS.

Este código concatena os dados dos *DataFrames* em questão, através do id Vereador. Assim, os resultados de ambos os arquivos são agrupados e, depois, é realizada a soma dos valores gatos de cada Vereador, possibilitando o agrupamento em uma única *tupla* referente à soma dos gastos do Custeio. Em outra coluna é alocado o Vereador e na outra o Partido ao qual ele pertence.

```
import pandas as pd

df_vereador = pd.read_csv("Partido_Vereador.csv", sep=';', encoding='cp1252')
df_gasto = pd.read_csv("Vereador_Custeio.csv", sep=";', encoding='cp1252')

df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('R','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('$','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace('.','')
df_gasto["Gasto"]=df_gasto["Gasto"].str.replace(',','.').astype(float)

df_gasto_agregado=df_gasto.groupby(by="Vereador").Gasto.sum().reset_index()

df_gasto_agregado = round(df_gasto_agregado, 2)
df_vereador_partido_gasto=df_gasto_agregado.merge(df_vereador,on='Vereador')

df_vereador_partido_gasto=df_vereador_partido_gasto[['Vereador','Partido','Gasto']]

print(df_vereador_partido_gasto)

df_vereador_partido_gasto.to_csv('dataset-bot.csv',index=False,sep=';',encoding='cp1252')
```

Figura 21 – Código *DataFrame Chat Bot* Sumé

Segue o exemplo do resultado gerado após execução do código:



	Vereador	Partido	Gasto
0	Bella Gonçalves	PSOL	2036.75
1	Bim da Ambulância	PSD	38575.53
2	Fernando Luiz	PSD	12527.52
3	Gabriel	PATRI	1678.85
4	Henrique Braga	PSDB	29925.96
5	Irlan Melo	PSD	45641.15
6	Jorge Santos	REPUBLICANOS	18534.78
7	Juninho Los Hermanos	AVANTE	58586.48
8	Marilda Portela	CIDADANIA	43892.58
9	Nely Aquino	PODE	32371.13
10	Professor Juliano Lopes	PTC	25914.74
11	Ramon Bibiano da Casa de Apoio	PSD	12684.41
12	Álvaro Damião	DEM	27498.84

Figura 22 - Resultado *DataFrame Chat Bot* Sumé.

Nesta última etapa foi criado o arquivo “dataset-bot.csv” que contém todos os dados de gastos dos vereadores e partidos referentes ao Custeio Parlamentar no mandato de 2017 até 2020.

O valor total dos gastos com o Custeio Parlamentar da Câmara foi de R$ 1.222.690,08 (Um milhão e duzentos e vinte e dois mil e seiscentos e noventa reais e oito centavos), segundo os dados do portal de transparência da CMBH. Por meio deste método também foi possível elaborar um ranking dos vereadores que mais gastaram em seu mandato as verbas públicas relativas a Custeio Parlamentar. Abaixo, temos o ranking dos 5 vereadores e dos 5 partidos que mais gastaram.

Tabela 1 - *Ranking* dos 5 vereadores que mais gastaram.

Posição	Vereador	Partido	Gasto
1 °	Catatau do Povo	PSD	R$ 61.071,08
2 °	Juninho Los Hermanos	AVANTE	R$ 58.540,10
3 °	Gilson Reis	PCdoB	R$ 51.409,24
4 °	Dr. Nilton	PSD	R$ 45.804,42
5 °	Irlan Melo	PSD	R$ 45.641,15

Tabela 2 - *Ranking* dos 5 partidos que mais gastaram.

Posição	Partido	Gasto
1 °	PSD	R$ 345.802,13
2 °	AVANTE	R$ 99.219,33
3 °	CIDADANIA	R$ 95.465,69
4 °	PSB	R$ 81.799,73
5 °	PSC	R$ 80.262,43

Ressalta-se que no portal de transparência da CMBH constam os dados originais. Contudo, eles não estão acessíveis em plataforma de dados abertos, fato abordado a todo momento nesta dissertação (RODRIGUES; FONTES, 2018). Segue abaixo o link da plataforma em que estão disponíveis os dados gerados através do método de *Web Scraping*:

Git Hub - https://github.com/wendelvilaca/chat_bot_sume

5.4. DIALOGFLOW

A plataforma api.ai foi adquirida no ano de 2016 pelo Google e veio a se transformar posteriormente em *Dialogflow*. A ferramenta é focada no desenvolvimento de agentes virtuais (*chatbots*) baseados em Inteligência Artificial, *Machine Learning* e processamento de linguagem natural. Dessa forma, a *Dialogflow* utiliza NLU (*Natural Language*

Understanding). É uma ferramenta gratuita, salvo algumas restrições com limites de mensagens. (BRANDES, 2017).

"O *Dialogflow* é uma plataforma de processamento de linguagem natural que facilita o design e a integração de uma interface do usuário conversacional com apps para dispositivos móveis, aplicativos da Web, dispositivos, *bots*, sistemas interativos de resposta de voz e etc". (DIALOGFLOW, 2020b)

A plataforma *Dialogflow* foi escolhida como agente virtual para consulta dos gastos de Custeio Parlamentar dos Vereadores de Belo Horizonte. Foram criadas *Intentes* com as respostas baseadas nos resultados obtidos pelo método Sumé de *Web Scraping* e armazenadas no arquivo "dataset-bot.csv".

Nas Intentes criadas temos os gastos de Custeio Parlamentar realizados no mandato entre 2017 e 2020, segmentados por ranking contendo o Vereador que mais gastou, os cinco Vereadores que mais gastaram, os cinco partidos que mais gastaram e o gasto total. Abaixo, segue o exemplo em formato JSON da *Intent* contendo *"responses*" e *"messages"* com as respostas dos cinco Vereadores que mais gastaram:

```
"id": "848462f5-c614-48a4-80f4-9f3222d09bfc", "name": "2_Top_5_Vereadores",
 "auto": true,
 "contexts": [],
 "responses": [
   {
     "resetContexts": false,
     "action": "",
     "affectedContexts": [],
     "parameters": [
       {
         "id": "86fdfca4-87c0-4347-b76a-7d52e0b23e4c",
         "name": "number",
         "required": false,
         "dataType": "@sys.number",
         "value": "$number",
         "defaultValue": "",
         "isList": false,
         "prompts": [],
         "promptMessages": [],
         "noMatchPromptMessages": [],
         "noInputPromptMessages": [],
         "outputDialogContexts": []
       },
       {
         "id": "5db986b4-09f3-4815-b59e-fcc1d2dc9516",
         "name": "location",
         "required": false,
         "dataType": "@sys.location",
         "value": "$location",
         "defaultValue": "",
         "isList": false,
         "prompts": [],
         "promptMessages": [],
         "noMatchPromptMessages": [],
         "noInputPromptMessages": [],
```


Figura 23 - *Intent* em formato JSON "responses"

```
      "messages": [
        {
          "type": "0",
          "title": "",
          "textToSpeech": "",
          "lang": "pt-br",
          "speech": [
            "Top 5 vereadores que mais gastaram:\n\n1 - Catatau do Povo, PSD, R$ 61.071,08
            \n2 - Juninho Los Hermanos, AVANTE, R$ 58.548,10\n3 - Gilson Reis, PCdoB, R$ 51.489,24
            \n4 - Dr. Wilton, PSD, R$ 45.884,42\n5 - Irlan Melo, PSD, R$ 45.641,15\n
            \nEsses gastos são referente a verba de custeio parlamentar da Câmara Municipal de Belo Horizonte."
          ],
          "condition": ""
        }
      ],
      "speech": []
    }
  ],
  "priority": 500000,
  "webhookUsed": false,
  "webhookForSlotFilling": false,
  "fallbackIntent": false,
  "events": [],
  "conditionalResponses": [],
  "condition": "",
  "conditionalFollowupEvents": []
```


Figura 24 - *Intent* em formato JSON *"messages"*

Foi utilizado o recurso *Integrations* do *Dialog Flow* para utilizar a *Web Demo* e nos possibilitar essa interação com *Chat Bot* Sumé pela internet. O link a seguir é passível de interação com o Sumé (https://bot.dialogflow.com/sume)

Este protótipo *Chat Bot* Sumé baseado em *Web Scraping* e dados abertos é pioneiro no Brasil para as interações entre os cidadãos e os gastos dos vereadores de câmaras municipais. É esperado que o projeto *Chat Bot* Sumé seja uma inspiração para fomentar soluções semelhantes e a Transparência Ativa Reversa. Assim, espera-se que os cidadãos possam utilizar de maneira facilitada os serviços e possam exercer seu controle social.

6. DADOS EXCLUÍDOS DO PORTAL DE TRANSPARÊNCIA

Após a conclusão do método de *Web Scraping* e criação dos Data Set, foi observado que alguns dados de gastos do Custeio Parlamentar dos vereadores tinham sido excluídos do portal de transparência sem qualquer aviso ou informação a respeito. Isto só foi possível, visto que os dados extraídos mensalmente eram armazenados localmente e na nuvem, seguindo o exemplo de boas práticas de backup.

Após ser realizada a comparação de dados do *Data Set* - que foram extraídos em outubro de 2020 e em janeiro de 2021 -, foi possível identificar inconsistência de alguns dados e a ausência de alguns vereadores. Ao todo, foi identificada a ausência de valores aproximados de R$ 194.489,24 (Cento e noventa e quatro mil e quatrocentos e oitenta e nove reais e vinte e quatro centavos). Esses gastos foram extintos do portal de transparência sem nenhum aviso e no momento não estão disponíveis. Segue abaixo discriminado o nome dos vereadores, bem como os gastos que foram excluídos do portal de transparência.

Tabela 3 - Dados excluídos do portal CMBH

Vereador	Partido	Gasto
Professor Wendel Mesquita	PSB	R$ 41.586,95
Jair Di Gregório	PP	R$ 35.676,16
Cláudio Duarte	PSL	R$ 31.501,47
Rafael Martins	PSD	R$ 22.726,21
Osvaldo Lopes	PHS	R$ 16.383,87
Cesar Gordin	PROS	R$ 16.278,51
Áurea Carolina	PSOL	R$ 12.913,67
Doorgal Andrada	PSD	R$ 8.706,73
Wellington Magalhães	PTN	R$ 6.146,92
Ronaldo Batista	PCC	R$ 2.128,97
Mateus Simões	NOVO	R$ 384,88
Ricardo da Farmácia	PMN	R$ 54,90
	TOTAL	**R$ 194.489,24**

Este é um dos exemplos que demonstra a eficácia do *Chat Bot* Sumé com procedimentos e ferramentas automatizados. Caso não existisse esse método no qual desde o ano passado foram coletados os dados, estas informações nunca estariam disponíveis para serem questionadas pelos cidadãos e não poderiam ser exigidas, tendo em vista que os dados não foram disponibilizados. Diante do exposto, fora esses dados de gastos excluídos, quais outros poderiam estar inacessíveis também?

Essa situação relembra algumas passagens do romance distópico 1984, de George Orwell. Nessa obra literária nos é apresentado um estado totalitário em que em que fora criado o Ministério da Verdade, entretanto esse Ministério seria responsável pela falsificação de documentos e literatura que referenciava o passado, assim o estado conseguia sempre dominar com sua retórica a tornando verdadeira. Como apontado no romance:

> “Smith trabalhava no Ministério da Verdade e tinha como função “corrigir” ou eliminar notícias e demais documentos do passado que não estavam de acordo com o presente. Assim, por exemplo, se algum artigo publicado tivesse uma previsão não consolidada, alterava-se este dado para que o “Partido” alcançasse os objetivos postos (ORWELL, 2009).

Nasce neste contexto a frase “quem controla o passado controla o presente e quem controla o presente controla o futuro”, toda informação era controlado pelo “Partido”, ou seja, eles conseguiam se perpetuar no poder através do controle dos dados e impondo narrativas (ORWELL, 2009, p. 47).

Parece que estarmos vivendo algo semelhante, guardadas, obviamente, as devidas proporções, em contexto bem específico, logo fazer uma analogia dessa situação com a distopia de 1984 é algo plausível. Toda essa conjuntura é bem passível de questionamentos referente aos portais de transparência do governo, tendo em vista que são geridos por eles próprios, na medida que eles detêm os dados primários e os inserem no portal.

Essa reflexão e preocupação é plausível, em virtude de que se não tivesse sido criado esse método de *Web Scraping* do *Chat Bot* Sumé, nunca saberíamos dessa inconsistência e que um dia esses dados estavam faltando. A ausência de participação popular pode facilitar essa anomalia e potencializar as inconsistências de dados existentes nos portais de transparência divulgados pelo estado, sendo que não se tem uma base externa para se comparar.

O cenário ideal ou uma boa prática seria a população exercer seu controle social e ser responsável também pelo controle da transparência, ao ponto de criar mecanismos como fontes externas das instituições públicas para que pudesse ter uma contrapartida da análise dos dados. Um bom exemplo é como tem sido feito em alguns trabalhos de dados abertos, em que as bases são armazenadas na plataforma CKAN.

Somente com acessibilidade aos dados podemos aumentar a participação da população. Dessa forma, se faz importante que existam meios mais fáceis na interação dos cidadãos com os gastos públicos, para, assim, exercerem pleno controle social.

7. ANÁLISE DOS RESULTADOS

Após analisar os dados referentes aos gastos dos vereadores, disponibilizados no portal de transparência da Câmara Municipal de Belo Horizonte, foi possível mapear as fontes de gastos dos parlamentares. Entretanto, nenhum dado está em formato aberto e não existe nenhuma API da CMBH que disponibilize esses dados. No momento, os dados disponibilizados pela CMBH não atendem aos requisitos da LAI, Art. 8°, § 3º (...) III que deixa clara a exigência do fornecimento dos dados em formato aberto estruturados e legíveis por máquina (BRASIL, 2011).

A CMBH possui maturidade mínima no que se refere a dados abertos dos gastos dos Vereadores, pois disponibiliza os dados em formato DOM, PDF e, algumas vezes, em XLS. Baseado na qualificação de cinco estrelas (BERNERS-LEE, 2009) a CMBH atinge o valor mínimo de uma estrela para os gastos do Custeio Parlamentar.

O método de *Web Scraping* possibilitou a geração de dados abertos e a evolução desses dados para 3 estrelas. Depois da transformação dos dados de gastos dos vereadores em dados abertos, será possível utilizá-los para manipulação de máquina, conforme abordado por (EAVES, 2009).

Após disponibilização dos dados nas plataformas *CKAN* e *Git Hub* eles se tornam acessíveis através do método de *Resource Description Framework* (RDF) alcançando, assim, o patamar de 4 estrelas. Os mais importantes itens de dados possuem uma *Uniform Resource Identifier* (URI) e podem ser compartilhados na Web (BERNERS-LEE, 2009).

Após consultas e solicitações junto à CMBH, não foi obtido o resultado no que tange à disponibilização dos dados abertos, nem previsão de quando será implementado algum projeto relacionado. A solicitação formal ocorreu em 30 de setembro de 2020, através da Solicitação 64668 conforme (Apêndice A).

O retorno oficial da CMBH foi totalmente desconexo e não respondeu o que foi questionado e solicitado; sequer mencionam Dados Abertos ou algo relativo a isso. Por este retorno fica evidente que o conceito Dados Abertos não é algo discutido na Câmara e provavelmente qualquer solicitação semelhante eles respondem por mensagem padrão. A resposta completa está disponível no (Apêndice A) deste projeto.

Um dos dados referentes aos gastos que mais chamaram atenção foi o de Serviço Postal, este recurso é um dos ofertados pela CMBH, como a própria câmara evidencia:

> Além do gabinete, a Câmara oferece serviços e materiais complementares a cada um dos vereadores, mediante processos de aquisição definidos nos termos da legislação federal de licitações:
>
> VII - Serviços Postais - Serviço fornecido pela Empresa Brasileira de Correios e Telégrafos, disponibilizado aos gabinetes dos vereadores, por meio do Contrato nº 9912468166/2019. O contrato está em vigor desde setembro de 2019 (CMBH, 2020c).

O controle social é importante para exercício da fiscalização dos gastos públicos, em virtude de que os parlamentares podem utilizar recursos que em algumas circunstâncias não são essenciais ou que não tem prioridade, como pode ser o caso do serviço postal.

Mesmo hoje com diversos serviços e recursos tecnológicos que facilitam a comunicação e divulgação, a CMBH ainda gasta valores elevados com o serviço postal. Somente no mandato de 2017 a 2020 os vereadores gastaram com esse serviço no mínimo impressionantes R$ 701.845,29 (setecentos e um mil oitocentos e quarenta e cinco reais e vinte e nove centavos). Vale ressaltar que no período de pandemia entre fevereiro e maio de 2020 os vereadores continuaram utilizando esse serviço chegando ao gasto extraordinário de R$ 99.875,77 (noventa e nove mil oitocentos e setenta e cinco reais e setenta e sete centavos).

Por esses dados podemos avaliar essa situação e perceber que não se teve bom senso com os gastos públicos nesse período de pandemia que tão delicado, então isso nos leva a questionar se em plena pandemia esse serviço não poderia ter sido evitado para gerar mais economia ao estado, possibilitando acumular recursos para adquirir materiais de saúde, como por exemplo equipamentos respiradores ou cilindros de oxigênio.

Os vereadores e a CMBH poderiam explicar por qual motivo esse serviço foi utilizado e qual foi a prioridade, assim estaríamos exercendo nosso controle social para avaliar se esse recurso foi utilizado corretamente como algo essencial ou prioritário, também poderá colocar em pauta se ainda precisa ser utilizado pela CMBH o Serviço Postal. Caso este serviço esteja sendo utilizado somente para divulgações de parabenizações ou mala direta como ocorre constantemente, é algo totalmente questionável. Hoje temos e-mails, redes sociais e diversos

recursos mais acessíveis, sem contar que em alguns casos alguns serviços são gratuitos para divulgação.

Durante o período de estudo deste trabalho ocorreu outra situação que foi a mais chamou atenção. Após o desenvolvimento do método de *Web Scraping* foi possível identificar a exclusão de dados de gastos do Custeio Parlamentar dos vereadores no portal de transparência. Juntando todos os valores que foram extintos do portal chegamos a valores de aproximadamente R$ 194.489,24 (Cento e noventa e quatro mil e quatrocentos e oitenta e nove reais e vinte e quatro centavos). São dados que não estão mais disponíveis, ou seja, a população sequer sabe da existência desses gastos e só é possível ter essa informação em função dos dados extraídos do portal em setembro de 2020, antes de realizarem a retirada. Isto demonstra a eficiência do método de *Web Scraping* para controle de gastos.

O *Chat Bot* Sumé está disponível para que a população possa interagir e utilizar essa tecnologia, possibilitando visualizar os gastos do mandato dos vereadores e, consequentemente, exercer o controle social através desta solução.

8. CONCLUSÃO E TRABALHOS FUTUROS

A eficácia do *Web Scraping* como método adotado para a extração de dados desestruturados, seguida de manipulação para transformá-los em Dados Abertos, foi comprovada neste trabalho. Com o *Chat Bot* Sumé, foi possível identificar irregularidades referentes aos dados analisados na Câmara Municipal de Belo Horizonte (CMBH). Além disso, o projeto viabilizou o desejado acesso aos gastos dos Vereadores da CMBH.

As técnicas de *Web Scraping* estão em constante evolução, assim como a maioria das ferramentas na área da Tecnologia. Entretanto, ainda há espaço para inovações relativas a dados abertos referentes a outros gastos dos vereadores, como gastos com serviços postais e gastos com telefonia. A inteligência artificial é o caminho apontado para identificar padrões de gastos elevados.

Ressalta-se que a pesquisa foi elaborada a partir de dúvidas relacionadas às discussões de dados abertos e a transparência das instituições públicas com a sociedade e os gastos públicos, no que se refere aos constructos de *Web Scraping* e Dados Governamentais. O

modelo conceitual proposto - via *Web Scraping* – foi o que permitiu disponibilizar dados desestruturados e dados abertos baseando-se no referencial teórico para dar suporte à pesquisa. O método foi desenvolvido e testado no portal de transparência da CMBH, realizando com eficácia a extração dos dados de Custeio Parlamentar e os agrupando para consulta na WEB.

Sobre o objetivo geral de desenvolver um método de *Web Scraping*, partindo de pesquisa bibliográfica e de ferramentas, o método mostrou-se capaz de tornar os dados desestruturados do portal de transparência, da Câmara Municipal de Belo Horizonte, em dados abertos. Ou seja, a pesquisa conseguiu atingi-lo.

A Contribuição da pesquisa de cunho acadêmico apresentou na literatura o relacionamento dos constructos que foram alvo desta pesquisa: *Web Scraping* e Dados Governamentais. Sob o aspecto de contribuição ao conhecimento, a pesquisa criou um método para extração de dados do portal de transparência. Sob o aspecto social, a maior contribuição é o *Chat Bot Sumé* que está disponível para consulta dos gastos dos Vereadores da CMBH para que os cidadãos possam exercer seu controle social. Também no campo social esperasse que esse trabalho gere debates e discussões acerca dos gastos com Serviços Postais, fomentando a verificação da utilização e prioridade desse serviço, se possível que ocorra sua extinção, caso não atenda às necessidades da população.

As limitações da pesquisa são os temas diversos que foram englobados teoricamente e superficialmente por não serem objetos principais de estudo da pesquisa.

Como proposta de projetos futuros relativos ao método adotado, recomendamos o aprimoramento do código do método de *Web Scraping*, com a ampliação do escopo para extração de gastos com serviços postais e gastos com telefonia, aumentando a integração entre as aplicações. Também é perfeitamente recomendável reaplicar o método em outros municípios no país, repetindo a realização do estudo do impacto do *Chat Bot Sumé* feito na cidade de Belo Horizonte relacionando a sua utilização pelos cidadãos. É recomendável elaborar estudo que analise e correlacione *Web Scraping* e Privacidade.

Nas considerações finais, enfatiza-se que a pesquisa teve importância tanto acadêmica quanto social, ao apresentar um método de extração de dados eficaz via *Chat Bot Sumé* e que

propôs o empoderamento da população com livre acesso aos gastos dos vereadores da sua cidade. Academicamente, cabe aos cientistas a concretização dos projetos sugeridos ou, ao menos, a realização de mais pesquisas a respeito dos temas estudados. Visando incentivar a melhoria contínua no campo da ciência, essa contribuição foi realizada para que a sociedade possa utilizar, da melhor forma possível, as informações aqui analisadas, estudadas e registradas.

REFERÊNCIAS

5 STARS OPEN DATA. **5 Stars Open Data**. 2012. Disponível em:<https://5stardata.info/en/>. Acesso em: 15 setembro 2020.

ABRAHAM, A. et al. (2014). Machine learning for neuroimaging with scikit-learn. **Frontiers in neuroinformatics.** Disponível em:<https://doi.org/10.3389/fninf.2014.00014.> Acesso em: 28 setembro 2020.

ARAÚJO, Luís & MARQUES, Rodrigo. (2019). Uma análise da transparência ativa nos sites ministeriais do Poder Executivo Federal brasileiro. **Revista Ibero-Americana de Ciência da Informação**. 12. 419-439. 10.26512/rici.v12.n2.2019.9236.

ABUSHAWAR, B. e ATWELL, E "Alice chatbot: Trials and outputs,"**Computacion y Sistemas**, vol. 19, pp. 625-632, 12 2015.

BARDIN, L. (2011) **Análise de conteúdo**. São Paulo: Edições 70.

BERNERS-LEE, T. Linked Data. 2009. Disponível em: <https://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 18 agosto 2020.

BIZER, C.; HEATH, T.; BERNERS-LEE. The semantic Web. **Scientific American**, v. 284, n. 5, p. 34-43, 2001.

BIZER, C.; HEATH, T.; BERNERS-LEE, T**.** Linked data – the story so far. **International journal on semantic web and information systems**, v. 5, n. 3, p. 1-22, 2009.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data – The story so far. [ed.] Tim Heath, M. Hepp and Christian Bizer. **International Journal on Semantic Web and Information System**, Special Issue on Linked Data, 2006.

BRANDES, B. Dialogflow (api.ai). **Breve introdução da plataforma**. Bots Brasil, 2017. Disponível em:<https://medium.com/botsbrasil/api-ai-breve-introdu%C3%A7%C3%A3o-da-plataforma-ecb2d77107a2> Acesso em: 02 setembro 2020.

BRASIL. Lei nº 12.527, de 18 de novembro de 2011. **Presidência da República**. Disponível em:<http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2011/lei/l12527.htm> Acesso em: 05 setembro 2020.

BROUCKE S. Vanden, BAESENS B. (2018) From Web Scraping to Web Crawling. In: Practical Web Scraping for Data Science. **Apress, Berkeley, CA**. Disponível em:<https://doi.org/10.1007/978-1-4842-3582-9_6> Acesso em: 17 setembro 2020.

BURKE, Edmund. 1774 Speech to the Electors of Bristol, on His Being Declared by the Sheriffs Duly Elected One of the Representatives In Parliament for That City, on Thursday, the 3rd of November, 1774" (in: NIMMO, J. C. (ed.). 1887. **The Works of the Right**

Honourable Edmund Burke, in Twelve Volumes. V. II. London: s/n. Disponível em:<http://socserv.mcmaster.ca/econ/ugcm/3ll3/burke/Works02.pdf> Acesso em: 12 Outubro 2020.

CÂMARA MUNICIPAL DE BELO HORIZONTE - CMBH. **Transparência.** Belo Horizonte, 07 de julho de 2020. Disponível em:<https://www.cmbh.mg.gov.br/transparencia-principal>. Acesso em: 20 de junho de 2020

CÂMARA MUNICIPAL DE BELO HORIZONTE - CMBH. **Custeio Parlamentar**, 2020. Belo Horizonte, 07 de jul. de 2020. Disponível em: <https://www.cmbh.mg.gov.br/transparencia/vereadores/custeio-parlamentar#resultadoPesquisa_custeio >. Acesso em: 20 de junho de 2020

CÂMARA MUNICIPAL DE BELO HORIZONTE - CMBH. **Como são custeados os mandatos parlamentares?** Belo Horizonte, 07 de jul. de 2020. Disponível em: <https://www.cmbh.mg.gov.br/A-C%C3%A2mara/entenda-a-camara >. Acesso em: 20 de junho de 2020.

CÂMARA MUNICIPAL DE BELO HORIZONTE - CMBH. **Custeio Parlamentar - Serviços Postais.** Belo Horizonte, 07 de jul. de 2020. Disponível em: < https://www.cmbh.mg.gov.br/transparencia/custeio-parlamentar/servicos-postais >. Acesso em: 20 de junho de 2020.

COMPREHENSIVE KNOWLEDGE ARCHIVE NETWORK – CKAN**. Association. API guide**, 2013. Disponível em:<https://docs.ckan.org/en/2.9/user-guide.html#what-is-ckan> Acesso em: 10 de setembro de 2020.

CLASTRES, H., **Terra sem Mal, O Profetismo Tupi-Guarani**. São Paulo, Editora Brasiliense, 1978.

CONTROLADORIA GERAL DA UNIÃO - CGU. **Verba Indenizatória**. Belo Horizonte, 11 de nov. de 2017. Disponível em: < https://www.gov.br/cgu/pt-br/governo-aberto/noticias/2017/consulta-sobre-dados-governamentais-de-interesse-da-comunidade-cientifica>. Acesso em: 13 de agosto de 2020

CORRÊA, A. et al. (2017). Transparency and open government data: a wide national assessment of data openness in Brazilian local governments. Transforming Government**: People, Process and Policy,** Vol. 11, Issue: 1, pp.58-78. Disponível em:<https://doi.org/10.1108/TG-12-2015-0052> Acesso em: 13 de agosto de 2020

CRUMMY. **Documentation Beautiful Soup.** 2004. Disponível em:<https://www.data.gov/about> Acesso em: 15 setembro 2020

DAS et al., 2015; Das, S., Dey, A., Pal, A., and Roy, N. (2015). Article: Applications of artificial intelligence in machine learning: Review and prospect. **International Journal of Computer Applications,** 115(9):31–41.

DATA.GOV - DG. **About Data.gov**. 2009. Disponível em: <https://www.data.gov/about> Acesso em: 15 setembro 2020

DAVENPORT, T. H. **Big Data at Work: Dispelling the Mhyts, Uncovering the Opportunities**. Boston, Massachusetts: Harvard Business School Publishing Corporation, 2014.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento empresarial: como as organizações gerenciam o seu capital intelectual**. 4. ed. Tradução de Lenke Peres. Rio de Janeiro: Campus, 1998. 237 p.

DEMOGRAPHIA WORLD URBAN AREAS. Built-Up Urban Areas or Urban Agglomerations, **11th Annual Edition: January**, 2015. Disponível em:<http://www.demographia.com/dbworldua.pdf> Acesso em: 5 outubro 2020

DIALOGFLOW. DialogFlow. **Documentation -Integrations.**2020. Disponível em:<https://cloud.google.com/dialogflow/docs> Acesso em: 3 setembro 2020.

DIALOGFLOW. **DialogFlow** 2020. Disponível em:<https://cloud.google.com/dialogflow> Acesso em: 3 setembro 2020.

DICKEN, Peter. Reverse engineering regains popularity, in **IEE Review**, vol. 42, no. 5, pp. 213-S2, 19 Sept. 1996, doi: 10.1049/ir:19960513.

DIOUF, E. N. Sarr, O. Sall, B. Birregah, M. Bousso and S. N. Mbaye, "Web Scraping: State-of-the-Art and Areas of Application," **2019 IEEE International Conference on Big Data (Big Data)**, Los Angeles, CA, USA, 2019, pp. 6040-6042, doi: 10.1109/BigData47090.2019.9005594.

EAVES, D. 2009. **The Three Laws of Open Government Data**. Acesso em 04 de outubro de 2014. Disponível em<http://eaves.ca/2009/09/30/three-law-of-open-government-data/> Acesso em: 3 dezembro 2020.

FONSECA, Lucas & AZEVEDO, C.L.B. & ALMEIDA, João. (2014). **Mapeando Dados Governamentais com uma Ontologia de Organizações**.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2007.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa.** 4. ed. São Paulo: Atlas, 2008.

HACKELING, G. (2014). **Mastering Machine Learning with scikit-learn.** Packt Publishing Ltd.

HADI, Ali & AL-ZEWAIRI, Malek. (2017). **Using IPython for Teaching Web Scraping**. 10.1007/978-3-319-55354-2_5.

HAN, J., Pei, J., & KAMBER, M. (2011). **Data mining: concepts and techniques**: Elsevier.

HERNÁNDEZ, A. et al. (2015). Metodologías para análisis político utilizando Web Scraping. **Research in Computing Science.** 95. 113-121. 10.13053/rcs-95-1-9.

INFRAESTRUTURA NACIONAL DE DADOS ABERTOS - INDA. **O que são dados abertos?**, 2011. Disponível em: <https://dados.gov.br/pagina/dados-abertos> Acessado em: 22 de setembro de 2020.

INFRAESTRUTURA NACIONAL DE DADOS ABERTOS - INDA. **Perguntas mais frequentes**, 2011. Disponível em: <http://dados.gov.br/pagina/faq> Acessado em: 22 de setembro de 2020.

INFRAESTRUTURA NACIONAL DE DADOS ABERTOS - INDA. **Sobre o dados.gov.br**, 2011. Disponível em: <https://dados.gov.br/pagina/sobre> Acessado em: 22 de setembro de 2020.

ISOTANI, Seiji; BITTENCOURT, Ig Ibert. **Dados abertos conectados**. 1 Ed. São Paulo: Editora Novatec, 2015.

JETBRAINS. PyCharm. Praga: **JetBrains**. Disponível em:<https://www.jetbrains.com/pycharm>. Acesso em: 18 setembro de 2020.

LOPES, Karen M. G.; ASSUMPÇÃO, Rita C. Processos e solução tecnológica para implementação da lei de acesso à informação (LAI). In: **CONGRESSO CONSAD DE GESTÃO PÚBLICA**, 2013, Brasília. Anais... Brasília, 2013

MATTOSINHO, F. J. A. P. **Thesis on Mining Product Opinions and Reviews on the Web.** Technische Universitat Dresden ,2010.

MCKINNEY, W. 2012. **Python for Data Analysis: Data Wranglingwith Pandas, NumPy, and IPython**. Sebastopol, California:O'Reilly Media Sebastopol, CA.

MENEZES, Romeu Araújo. **Chatterbot crioulo: proposta de um conversador quilombola das terras de preto do Território Litoral Sul - BA.** 109 f. Dissertação (Mestrado) - Mestrado Profissional em Gestão e Tecnologia da Educação, Departamento de Educação, Universidade do Estado da Bahia, 2016.

MUSSKOPF, I. (2016) **Disponível é diferente de acessível.** Data Science Brigade. Disponível em: <https://medium.com/serenata/dispon%C3%ADvel-%C3%A9-diferente-de-acess%C3%ADvel-56e1f76188c1>. Acesso em: 15 novembro 2020.

NISHIHARA, Y. et al. A Generation Method of Back-Channel Response to Let a Chatting Bot Be a Member of Discussions in a Text-Based Chat, 2**017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)**, Hamamatsu, 2017, pp. 324-329

NÓBREGA, M.- **Cartas do Brasil, 1549-1560**. Belo Horizonte, Ed. Itatiaia; São Paulo, EDUSP, 1988.

OJO, Adegboyega; CURRY, Edward; ZELETI, Fatemeh Ahmadi. A Tale of Open Data Innovations in Five Smart Cities. In: System Sciences (HICSS), **2015 48th Hawaii International Conference on. IEEE**, 2015. p. 2326-2335.

OPEN GOV DATA - OGD. **The Annotated 8 Principles of Open Government Data**. 2007. Disponível em: http://www.opengovdata.org/. Acesso em: 2 de outubro 2020

OPEN GOVERNMENT PARTNERSHIP - OGP. **Declaração de governo aberto**. set. 2011. Disponível em: <www.opengovpartnership.org/open-government-declaration>. Acesso em: 10 setembro 2020.

OPEN KNOWLEDGE FOUNDATION - OKF. **OPEN DATA HANDBOOK - What is Open Data?**. 2020 Londres: Open Knowledge Foundation. Disponível em: < http://opendatahandbook.org/guide/en/what-is-open-data//>. Acesso em: 18 setembro 2020.

OPEN KNOWLEDGE FOUNDATION - OKF. **What is open?** 2020.Londres: Open Knowledge Foundation. Disponível em: < https://okfn.org/opendata/ >. Acesso em: 18 agosto. 2020.

OPEN KNOWLEDGE FOUNDATION - OKF. **Announcement – Open Definition 2.1.** 2020.Londres: Open Knowledge Foundation. Disponível em: <https://blog.okfn.org/2015/11/10/announcement-open-definition-2-1/>. Acesso em: 18 setembro 2020.

ORWELL, G. **1984**. Trad. Alexandre Hubner e Heloisa Jahn. São Paulo: Companhia das Letras, 2009.

PANDAS. **About Pandas**. Disponível em: <https://pandas.pydata.org/about/> Acesso em: 8 de junho 2020

PORTAL DA TRANSPARÊNCIA DO GOVERNO FEDERAL PTGF. **O que é e como funciona**, 2004. Disponível em: http://www.portaltransparencia.gov.br/sobre/o-que-e-e-como-funciona. Acesso em: 28 agosto 2020.

PRISCO V. (2019). A Facebook chat bot as recommendation system for programming problems**. 2019 IEEE Frontiers in Education Conference (FIE)** 1-5. 10.1109/FIE43999.2019.9028655.

PYTHON SOFTWARE FOUNDATION - PSF. **What is python?** 2001.Disponível em: < https://docs.python.org/3/faq/general.html#what-is-python >. Acesso em: 18 agosto 2020.

PYTHON SOFTWARE FOUNDATION - PSF. **Time access and conversions**. 2001. Disponível em: < https://docs.python.org/3/library/time.html >. Acesso em: 18 agosto 2020.

RICH, E.; KNIGHT, K. **Inteligência Artificial**. 2ª ed. Rio de Janeiro. Editora McGraw-Hill, 1994. 722 p.

RODRIGUES, Jc & FONTES, Carlos. (2018). Estudo de Caso "Operação Serenata de Amor": a análise de Big Data no combate à festa dos gastos públicos. **Conference: XIV Congreso de la Asociación Latinoamericana de Investigadores de la Comunicación (ALAIC)**. Disponível em: <https://www.researchgate.net/publication/323585318_Estudo_de_Caso_Operacao_Serenata_de_Amor_a_analise_de_Big_Data_no_combate_a_festa_dos_gastos_publicos> Acesso em: 18 outubro 2020.

RUSSELL, S.J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. New Jersey: Prentice Hall,2009 (3º Ed.).

RUSSELL, S.; NORVIG, P. (2013) **Inteligência Artificial**. 3. ed. Rio de Janeiro: Elsevier.

SÁ, Maria Irene da Fonseca; MALIN, Ana Maria Barcelos. Lei de Acesso à Informação: Um Estudo Comparativo com Outros Países. In: **XIII Encontro Nacional de Pesquisa em Ciência da Informação (XIII ENANCIB)**, 2012, Rio de Janeiro. Disponível em: <https://periodicos.ufmg.br/index.php/revistaagora/article/view/2624>

SANDOVAL-Almazan R., GARCIA Gil J.R. (2014) Towards an Evaluation Model for Open Government: A Preliminary Proposal. In: Janssen M., Scholl H.J., Wimmer M.A., Bannister F. (eds) Electronic Government. EGOV 2014**. Lecture Notes in Computer Science, vol 8653. Springer**, Berlin, Heidelberg. https://doi.org/10.1007/978-3-662-44426-9_4

SCHLICHT, Matt. **The Complete Beginner's Guide To Chatbots: Everything you need to know**. 2016. Disponível em: <https://chatbotsmagazine.com/the-complete-beginner-sguide-to-chatbots-8280b7b906ca>. Acesso em: 20 setembro 2020.

SELENIUM**. Projeto de automação do navegador Selenium**. Disponível em: <https://www.selenium.dev/documentation/en/>. Acesso em: 15 novembro 2020.

SILVEIRA, D.T.; CÓRDOVA, F.P. A pesquisa científica. In: GERHARDT, T.E.; SILVEIRA, D.T.(Org.). **Métodos de pesquisa**. Porto Alegre: UFRGS, 2009.

VAROL O. et al. **Online Human-Bot Interactions: Detection, Estimation, and Characterization**, arXiv: 1703.03107v1, 2017. Disponível em: <https://arxiv.org/pdf/1703.03107.pdf> Acesso em: 28 setembro 2020.

VILANOVA, P. (2017). **Convertendo suspeitas em dinheiro devolvido: Como são as denúncias da Operação Serenata de Amor**. Disponível em: <https://medium.com/data-sciencebrigade/convertendo-suspeitas-em-dinheiro-devolvido-como-são-as-denúncias-da-operaçãoserenata-de-amor-34fe8425631e> Acesso em: 20 setembro 2020.

WANG, H.-J., & LO, J. (2015). **Adoption of open government data among government agencies. Government Information Quarterly**. Disponível em: <http://dx.doi.org/10.1016/j.giq.2015.11.004> Acesso em: 20 setembro 2020.

YAZIGI, A. F. (1999). Dinero, política y transparencia: El imperativo democrático de combatir la corrupción. **9th International Anti-Corruption Conference** (IACC), 10–15.

WOOD, D. et al. Linked Data: Structured data on the Web. New York: Manning Publications Co., 2013.

APÊNDICE A – PROTOCOLO DE ATENDIMENTO - SOLICITAÇÃO 64668

1 – Solicitação do cidadão:

Link da solicitação: https://www.cmbh.mg.gov.br/participe/lai/protocolo/64668

Data de solicitação: 26/09/2020 - 02:42

Tipo de Solicitação: Solicitações Solicitação: Wendel Vilaça de Assis

Boa noite! Não está disponível no site de transparência da câmara dos vereadores de Belo Horizonte e em nenhuma outra plataforma os Dados Abertos referente aos gastos de Custeio Parlamentar, Gastos com serviços postais, Gastos com telefonia e Verba Indenizatória dos Vereadores de BH. Formalizo minha solicitação desses dados abertos, que eles sejam disponibilizados baseado na Lei de Acesso à Informação (LAI), Art. 8°, § 3º (...) III – possibilitar o acesso automatizado por sistemas externos em formatos abertos, estruturados e legíveis por máquina. De acordo com a Lei de Acesso à Informação (Lei Federal nº 12.527/2011) esses dados devem estar disponíveis. Diante do exposto, qual a data provável de disponibilização desses dados? Obrigado!

2 - Resposta Final da CMBH:

1) Quais são as verbas a que cada vereador tem direito? Todas elas estão englobadas em "Custeio Parlamentar" e "Verba Indenizatória", disponíveis no link https://www.cmbh.mg.gov.br/transparencia/vereadores? Os vereadores fizeram uso de verba indenizatória até meados de 2017, quando os suprimentos passaram a ser realizados, em sua totalidade, por meio de licitações, como regulamenta a Deliberação nº 18/2016. As licitações estão disponíveis no Portal da Transparência da CMBH >Transparência>Licitações.

2) Qual o valor mensal do "Custeio Parlamentar" disponível para cada parlamentar? Com exceção do Impresso de divulgação do mandato parlamentar, dos serviços postais e do serviço de telefonia, os contratos são feitos por Sistema de Registro de Preços, cujos valores podem sofrer alterações a cada ano, conforme a data em que as licitações e contratos ficam prontos. Com o modelo de fornecimento por licitação, para viabilizar o planejamento, é feita uma definição de cota anual por item e não um cômputo de valor

mensal. A totalidade dos itens pode ser solicitada do primeiro mês de vigência do contrato até o mês de dezembro ou podem ser feitas solicitações parciais, conforme demanda. Dada alteração do calendário eleitoral, para o ano de 2020 não houve licitação para o item “Impresso de divulgação do Mandato Parlamentar”. Os gastos com serviços postais e de telefonia estão disponíveis no Portal da Transparência da CMBH > Transparência > Vereadores. E as licitações estão disponíveis no Portal da Transparência da CMBH >Transparência > Licitações.

3) O "Custeio Parlamentar" dos vereadores não está disponível para os meses Janeiro, Fevereiro, Março e Abril/2017; para Janeiro, Fevereiro, Março/2018; para janeiro/2019. Por qual motivo? O Custeio Parlamentar iniciou a partir de maio de 2017. Em relação aos meses dos anos de 2018 e 2019, os gastos não aparecem no portal pelo fato de os procedimentos administrativos licitatórios não terem abrangido tais períodos, por terem se findado recentemente ou ainda estarem findando. Os gastos dependem, portanto, do término dos procedimentos licitatórios e/ou datas de celebração dos contratos.

4) A partir de Maio/2017, os valores estão disponíveis, mas não para todos os vereadores. Por que? Quando o valor para determinado vereador não está disponível é porque ele não teve gastos com custeio? Sim. A cota das categorias de despesa é isonômica a todos os vereadores. Se o gasto com custeio de determinado parlamentar não foi disponibilizado, certamente não realizou gastos dessas categorias que tenham sido custeados pela Câmara Municipal.

5) Com relação à Verba Indenizatória os valores não estão disponíveis para o mês de Maio/2017 e a partir de Agosto do mesmo ano. Por qual motivo? Como relatado na questão 03, as contratações se iniciaram em maio/2017. Todavia, porque a empresa que forneceria o Impresso de Divulgação do Mandato Parlamentar não cumpriu com as exigências contratuais, foi necessário contratar outra licitante. E neste ínterim (meses de junho e julho), somente para essa categoria, foi disponibilizado o uso de verba. Os demais serviços/materiais foram fornecidos mediante contratos celebrados a partir de licitações.

6) Onde encontro a informação sobre o número de assessores de cada parlamentar e o total gasto com os mesmos? Todas as informações solicitadas quanto aos assessores parlamentares estão disponíveis no Portal da Transparência da CMBH, que pode ser

acessado por meio do link: https://www.cmbh.mg.gov.br/transparencia/pessoal/consulta-a-remuneracao A relação completa de servidores, com as respectivas lotações, encontra-se disponível no site. A consolidação não compete à Câmara, nos termos do art. 9º, III, da Deliberação nº 5/2013. Cordialmente.