

CURSO DE SISTEMAS DE INFORMAÇÃO – TCC ACADÊMICO	
(X) PRÉ-PROJETO () PROJETO	ANO/SEMESTRE: 2021/01

REPOSITÓRIO DE INFORMAÇÕES DE PARLAMENTARES: UM DOSSIÊ PÚBLICO ONLINE

Jean Patrick Scherer

Aurélio Faustino Hoppe - Orientador

1 INTRODUÇÃO

De acordo com Barbosa (2018), não há precedentes sobre a frequência e volume com que as pessoas publicam informações virtualmente, cujo maior atrativo é que grande parte dos usuários da internet postam suas opiniões, expectativas, elogios e frustrações, concebendo, assim, vasta base de dados primários sobre diversos produtos e serviços.

Segundo Nogueira (2014), com o advento da Internet, o volume de informações textuais vem crescendo exponencialmente. Tais informações são encontradas em maiores quantidades em diversas fontes. Segundo o autor, através das informações textuais, conhecidas também como não estruturadas, é possível extrair conhecimento útil e implícito que devido ao grande volume são impossíveis de serem processadas por um ser humano.

Por outro lado, segundo Almeida (2007), as escolhas de representantes políticos no Brasil sempre foi um trabalho árduo para o eleitor. Na maioria das vezes, não se é dada a devida atenção a esta ação que pode gerar inúmeras consequências para que a fez e para as demais que a rodeiam. Porém, nos últimos anos a política está em evidência no país, devido a troca da detenção do poder entre os extremos da democracia (direta e esquerda da política nacional) que ocasionou, como esperado, uma reviravolta no país e na percepção da população em relação ao voto.

Almeida (2007) também aponta que em toda eleição surgem novos candidatos para serem escolhidos entre a população, com o discurso quase que padrão de educação e segurança. Mas até onde isto realmente é verdade? Como posso estar em garantia que esta mesma pessoa que está me prometendo investimentos na educação, por exemplo, a alguns anos atrás não vetou algum projeto que destinaria recursos para esta mesma educação? Até onde é verdade este discurso e até onde não? Ou seja, percebe-se que há carência de informação de fácil acesso aos eleitores sobre os candidatos e isto, impulsiona escolhas equivocadas ou não baseadas em fatos comprovados por veículos de imprensa em relação aos candidatos.

Diante desta realidade, este trabalho propõem o desenvolvimento de uma aplicação que implementará técnicas de mineração de dados para apoiar a extração e o agrupamento de informações históricas de seus representantes, disponibilizando-as na forma de um dossiê político.

1.1 OBJETIVOS

O objetivo deste trabalho é disponibilizar uma aplicação para agrupar informações históricas de seus representantes, apresentando-as na forma de um dossiê político.

Os objetivos específicos são:

- a) apresentar, de forma clara e comprovada, ações políticas realizadas pelo candidato;
- b) facilitar e auxiliar na escolha de representantes políticos, conforme as preferencias do eleitor;
- c) disponibilizar, de forma centralizada, conteúdo de fácil e ampla consulta.

2 TRABALHOS CORRELATOS

Neste capítulo são apresentados três trabalhos correlatos, que possuem características semelhantes à proposta deste trabalho. A seção 2.1 detalha o SentimentALL (CAMARGO *et al.*, 2017) que extrai dados do turismo nacional com o *framework* Scrapy. Na seção 2.2 é descrito a aplicação ALUMNI tool (ALMEIDA, 2018) que recupera dados pessoais na Web em redes sociais autenticadas. Por fim, a seção 2.3, aborda o artigo ciência política na era da big data (SILVA; MEIRELES, 2016) que visa automatizar a coleta de dados políticos na internet para fins acadêmicos utilizando a linguagem de programação R.

2.1 SPIDERS DE EXTRAÇÃO DE DADOS DO TURISMO NACIONAL COM O FRAMEWORK SCRAPY

Camargo *et al.* (2017) desenvolveram um módulo de extração para coletar dados do contexto do turismo nacional. Os autores utilizaram o *framework* Scrapy para extração de dados do site TripAdvisor. O *framework* Scrapy permite definições e implementações de regras de web *crawling* e web *scraping*, denominadas *spiders*, escrito em Python possibilitando o alcance de grandes massas de dados a partir da web ou outras fontes.

Segundo Camargo *et al.* (2017), as coletas foram divididas em duas partes, sendo que para cada uma foram criadas três *spiders*, uma para cada tipo de busca, atrações, hotéis e restaurantes. Primeiramente, as *spiders* realizam a sondagem, coletando a quantidade de avaliações para cada tipo de busca para todas as cidades do Brasil. Posteriormente, foram filtradas apenas as extrações das cem cidades com a maior quantidade de avaliações.

Camargo *et al.* (2017) apontam que inicialmente deve-se informar o link da página ao será feita a navegação do *spider* até encontrar o objeto de extração (atração, hotel ou restaurante). No caso do TripAdvisor os endereços relacionados seguem um padrão de Uniform Resource Locator (URL) baseado pelo identificador ao{X} onde, X é uma progressão

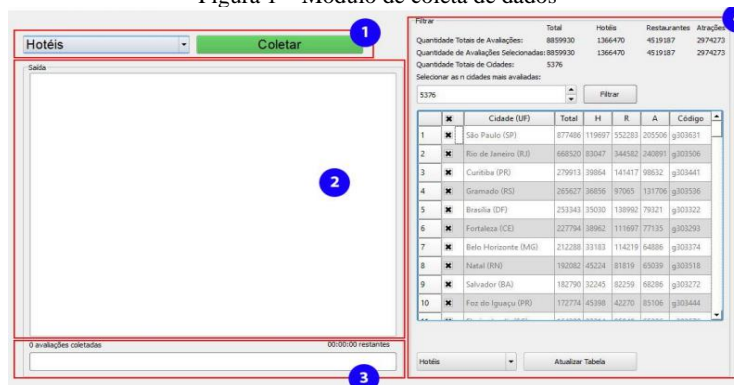
incrementada de vinte em vinte. A partir desta informação é possível fazer a extração das respectivas cidades. Caso houver paginação, o X mencionado anteriormente, deve ser incrementado de trinta em trinta.

De acordo com Camargo *et al.* (2017), o objetivo da segunda *spider* é a extração dos dados de cidade, estado, nome do objeto, quantidade de avaliações e quantidade de avaliações em português, armazenando em um arquivo Comma-Separated-Values (CSV). Esta mesma *spider* é responsável pela extração do código da cidade e código do objeto que são encontradas na URL da página.

Na terceira spider, segundo Camargo *et al.* (2017) é necessário definir os links iniciais de busca. Para isso foi criado uma função para montar os links de cada objeto com base num conjunto de cidades selecionadas pelo usuário. Os códigos de cidades e o tipo de objeto são extraídos do arquivo CSV gerado anteriormente. O padrão de link para esta etapa consiste em: `tripadvisor.com.br/{tipoobjeto}_Review{CÓDIGO CIDADE}{CÓDIGO OBJETO}`. Os dados extraídos e salvos em arquivo CSV são, Objeto, Código do Objeto, Tipo de objeto, Cidade, Código da cidade, Estado, País, Nome do autor, Cidade do autor, Nível do autor, Ano de cadastro do autor no site, Código do autor, Título, Nota, Data de postagem, Comentário e Código de avaliação.

Na Figura 1, é apresentada a interface gráfica do módulo de coleta. Através dela é possível iniciar e interromper a coleta dos dados (Figura 1 item 1), visualizar os logs gerados pelo *framework* Scrapy (Figura 1 item 2), uma barra de status do processo juntamente com o tempo estimado para finalização e quantidade de avaliações coletadas (Figura 1 item 3) e por fim, dados referentes a coleta (armazenados em CSV) são exibidos possibilidade a realização de filtro por cidade (Figura 1 item 4)

Figura 1 – Módulo de coleta de dados



Fonte: Camargo *et al.* (2017).

Camargo *et al.* (2017) apontam como possíveis melhorias a curto e médio prazo, (i) a possibilidade de se desenvolver *spiders* utilizando interface gráfica; (ii) desenvolvimento de rotinas padrões de pré-processamento para diminuir o tempo gasto no tratamento dos dados e, (iii) possibilidade de se trabalhar com mais contextos e não somente com o turismo.

2.2 RECUPERAÇÃO DE DADOS PESSOAIS NA WEB EM REDES SOCIAIS AUTENTICADAS

Almeida (2018) desenvolveu a aplicação ALUMNI tool, tendo como objetivo a recuperação de dados pessoais na Web em redes sociais autenticadas. O autor implementou um web *scraping*, combinando um robô de busca em PHP e um programa para testes funcionais para coletar informações publicadas por alumni (plural de ex-aluno em latim) em redes sociais.

Segundo Almeida (2018), a grande maioria dos *crawlers* foram escritos em Java pois não se encontrou nenhum método automatizado de extração de informações de usuários de redes sociais diretamente da web, principalmente, aquelas que possuem autenticação.

Almeida (2018) subdividiu a aplicação em duas partes, sendo elas: (i) o robô, conjunto de programas em PHP vinculado com o software Selenium, que navega pelas URL's do site LinkedIn para recuperar **perfis** de ex-alunos e seus respectivos ID's que ficarão disponíveis via requisição HTTP em porta 4444 para a aplicação web. E, (ii) a aplicação web, que irá realizar a análise e processamento dos dados coletados, resumizando e disponibilizando-os de forma visual, no formato HTML, gráficos e tabelas. Na aplicação web, o autor também utilizou a linguagem de programação PHP, porém, utilizando o padrão de projeto MVC.

Na etapa de autenticação da plataforma LinkedIn, Almeida (2018) utilizou suas próprias informações de acesso, de usuário e senha. Relatando que os desafios encontrados foram a manutenção da sessão ativa no navegador, a geração de conteúdo HTML via AJAX e eventuais barreiras que o próprio LinkedIn adota contra robôs de busca. Para resolver a situação de manter a sessão ativa, o autor utilizou a ferramenta Selenium WebDriver, sendo responsável por carregar a janela do navegador e de forma autônoma fornecer os dados de login e senha válido da plataforma e, posteriormente, extraindo e armazenando o conteúdo HTML em um banco de dados. Com estes dados, o robô executa um laço de iteração para percorrer todas as URL's recuperadas utilizando a biblioteca DOMDocument da linguagem PHP combinada com utilização de expressões regulares, realizando a mineração de dados. Os dados coletados foram armazenados em banco de dados MySQL, totalizando 9 tabelas, que a aplicação final web terá acesso para realizar consultas em T-SQL.

A validação e limpeza dos dados foi feita através de regras T-SQL combinadas com scripts em PHP. Exemplos de validação e limpeza seriam dados duplicados de perfis e históricos profissionais sem vínculo direto com perfil de usuário na plataforma. Neste processo de validação, foram encontrados alguns pontos de dificuldade, sendo eles: *encoding* de dados, utilização de espaçamento único entre os nomes, utilização de letras maiúsculas e minúsculas, não utilização de acentos.

Segundo Almeida (2018) foram coletados no total, 388.868 registros, sendo eles aproximadamente, 118.000 registros de edus (histórico acadêmico), 200.000 jobs (histórico profissional), 11.000 empresas, 580 escolas e 300 cursos de graduação. Almeida (2018) também aponta que a ferramenta foi muito eficaz, atingindo 88% de hashes (perfis) e 66% de alumni (ex-alunos) com um tempo de processamento de aproximadamente 3 meses de duração.

Almeida (2018) sugere como trabalhos futuros, (i) a verificação da veracidade das informações inseridas nos perfis das plataformas, (ii) a mineração em mais modelos de web sites para acrescer a quantidade da amostragem e, (iii) a utilização de informações complementares disponibilizadas no perfil dos usuários, tal como a informação de *skills* que necessitam ser endossadas por outros perfis, aumentando assim, a integridade das informações contidas.

2.3 CIÊNCIA POLÍTICA NA ERA DO BIG DATA: AUTOMAÇÃO NA COLETA DE DADOS DIGITAIS

Silva e Meireles (2016) utilizaram a linguagem de programação R para a coleta de informações da Assembleia Legislativa de Minas Gerais, além de mais sete estados, sendo coletadas informações de **PL's PLC's e PEC's** realizadas pelas instituições. Os autores separaram as coletas em dois conjuntos de dados: (i) número, do ano, do tipo e link das proposições de interesse e, (ii) algumas outras informações oriundas do link como, tipo, sigla, ano, número, autor, partido, ementa, local fim e situação.

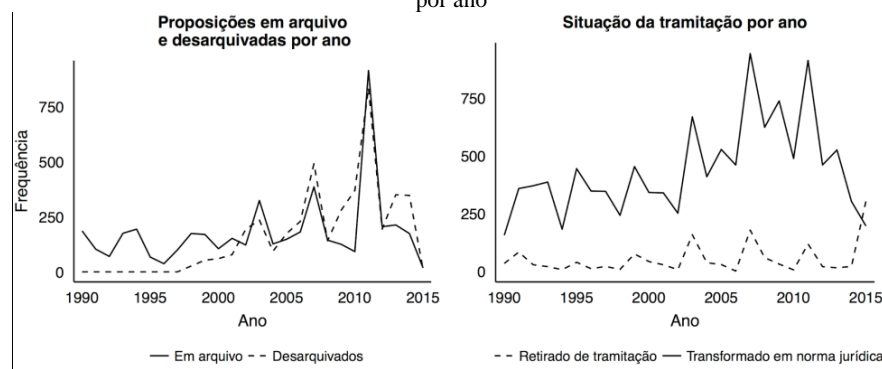
Segundo Silva e Meireles (2016), o primeiro conjunto de dados foi coletado diretamente da página web da Assembleia de Minas Gerais, ou seja, mineraram os conteúdos diretamente da página HTML. Já para o segundo, utilizaram a API disponibilizada pela própria Assembleia, tendo acesso a arquivos no formato JSON e XML. Os autores utilizaram como ferramenta de manipulação o pacote "RSelenium".

Silva e Meireles (2016) trabalharam bastante na parte de mineração do HTML. No qual, a partir do link de acesso, o conteúdo HTML era carregado para processamento no R e posteriormente, realizando o processo de seleção de conteúdo com métodos de eliminação e

contagem de caracteres. Ainda segundo os autores, o processo de coleta de informações foi facilitado, uma vez que, sabendo do modelo de retorno das requisições das API's, basta extrair o valor do campo pelo nome.

De acordo com Silva e Meireles (2016), no total, foram capturadas 25.418 proposições em aproximadamente 2h. Dados gerados entre 1990 e 2015. Após a coleta, os dados foram sumarizados em gráficos para análises, conforme mostra a Figura 2.

Figura 2 – Distribuição da proposição em arquivo ou desarquivadas e situação de tramitação, por ano



Fonte: Silva e Meireles (2016).

Segundo Silva e Meireles (2016), o número de proposições em arquivo é maior do que as desarquivadas até 2000. Depois desse ponto, a distribuição é semelhante nos dois tipos. O único ano em que houve maior quantidade de retiradas de tramitação foi em 2015. Nos demais, houve quantidade relevante maior de tramitações que se tornaram normas nas proposições legislativas.

Como principal argumentação para o trabalho, Silva e Meireles (2016) ponderam as características de “eliminação de erros de imputação, digitação ou até mesmo do copiar e colar”, além do tempo economizado na etapa de pesquisa empírica. Eles também ressaltam que, grande parte do volume da dados não estão disponíveis para download em formatos aceitos por softwares estatísticos, mesmo nas plataformas que disponibilizam uma API. Por fim, Silva e Meireles (2016) sugerem melhorias na visualização e filtros dos resultados coletados e, também quanto ao recebimento de informações de outras assembleias do território nacional.

3 PROPOSTA

A seguir é apresentada a justificativa para o desenvolvimento desse trabalho, os principais requisitos e a metodologia de desenvolvimento que será utilizada. Também são relacionados os assuntos e as fontes bibliográficas que irão fundamentar o estudo proposto.

3.1 JUSTIFICATIVA

No Quadro 1, é feita comparação entre três trabalhos correlatos voltados a estas técnicas computacionais, listando diferenças pontuais em suas técnicas, finalidades e assuntos envolvidos na mineração. As linhas representam as características e as colunas os trabalhos.

Quadro 1 – Comparativo dos trabalhos correlatos

Trabalhos Características	Camargo <i>et al.</i> (2017)	Almeida (2018)	Silva e Meireles (2016)
Realizam a extração de dados	Sim	Sim	Sim
Utilizam técnicas de mineração de dados	Sim	Sim	Sim
Sumarização os dados coletados	Não	Sim	Sim
Dados vinculados a política brasileira	Não	Não	Sim
Api utilizadas	Scrapy	RSelenium	RSelenium

Fonte: elaborado pelo autor.

A partir do Quadro 1, pode-se observar que todas as aplicações extraem dados de sites e utilizam de técnicas de mineração para alcançarem seus resultados. Além disso, também percebe-se que apenas o trabalho de Silva e Meireles (2016) realizam a manipulação de dados da política brasileira. Camargo *et al.* (2017) optaram que extração e sumarização de informações turísticas. Já Almeida (2018) pela coleta informações das redes sociais.

O trabalho proposto é relevante pois agrupará informações históricas de seus representantes, apresentando-as na forma de um dossiê político, sendo exposto de forma categorizada (negativamente ou positivamente) e de fácil leitura, as ações realizadas ou ações vinculadas a determinado parlamentar, estando ou não em processo eleitoral. Também serão vinculados *links* de matérias oriundas de portais oficiais de notícias aos nomes dos parlamentares envolvidos diretamente ou indiretamente no conteúdo da matéria disponível no *link* informado. Dessa forma, espera-se que os eleitores possam ter acesso facilitado a informações sobre seus candidatos e discursos apresentados, agregando na qualidade das argumentações em discussões construtivas, debates e nas escolhas feitas pelos eleitores nas urnas. Acredita-se que através de um *web crawling* e técnicas de mineração de dados será possível disponibilizar na forma de um repositório centralizado, informações agrupadas, simplificadas e de fácil acesso, visando beneficiar de forma geral os eleitores do país.

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

A aplicação de mineração de dados deverá:

- permitir o cadastro de sites ao qual serão extraídas as informações (Requisito Funcional – RF);

- b) extrair as informações dos sites através de um web scraping (RF);
- c) implementar o processo de mineração e as suas etapas, sendo elas: seleção, pré-processamento, limpeza, transformação, mineração de dados e interpretação dos dados (RF);
- d) implementar a técnica de agrupamento para realizar a mineração dos dados (RF);
- e) disponibilizar as informações na forma de um repositório/dossiê com link para notícias dos mandatários (RF);
- f) permitir ao usuário consultar as informações a partir do nome completo do mandatário (RF);
- g) ser implementada utilizando o ambiente de desenvolvimento Microsoft Visual Studio 2010 e na plataforma ASP.NET (Requisito Não-Funcional - RNF);
- h) ser implementada utilizando Kubernetes ou Docker Swarm (RNF).

Comentado [ARV1]: Va implementar ou usar algo pronto (bot)?

Comentado [ARV2R1]: Considerando o que está na metodologia será desenvolvido, então sugiro detalhar os requisitos dele (web scraping)

Comentado [ARV3]: Isso aqui constitui-se de vários RFs, ou não?

Comentado [ARV4]: Considerando quais premissas?

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

- a) levantamento bibliográfico: realizar o levantamento bibliográfico sobre mineração de dados, web scraping e trabalhos correlatos;
- b) elicitação de requisitos: baseando-se no levantamento bibliográfico, refinar os requisitos propostos para a mineração de dados proposto;
- c) especificação da aplicação: especificar a aplicação com análise orientada a objetos utilizando a Unified Modeling Language (UML). Utilizar a ferramenta Enterprise Architect 7.5 (EA) para o desenvolvimento dos diagramas de casos de uso e classes;
- d) definição dos sites de busca: analisar e escolher quais sites disponibilizam informações são relevantes sobre os mandatários;
- e) obtenção das informações dos sites: desenvolvimento de um web scraping na linguagem ASP.NET, que busque as informações a partir da etapa (d);
- f) definição da técnica de mineração: pesquisar, analisar a técnica de agrupamento que irá ser utilizada para compor a aplicação de mineração dos dados, levando em consideração seu desempenho e sua eficácia;
- g) implementação o processo de extração e agrupamento das informações: a partir das etapas (e) e (f) implementar o processo de mineração e agrupamento das informações dos mandatários utilizando a plataforma ASP.NET com o ambiente Microsoft Visual Studio 2010;
- h) testes: realizar testes com a técnica implementada na etapa (g) verificando se a

mesma é eficiência e possui baixo custo computacional para realizar a extração das informações. Além disso também será verificado a assertividade dos agrupamentos de notícias associadas aos mandatários.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2 – Cronograma de atividades a serem realizadas

etapas / quinzenas	2021							
	ago.		set.		out.		nov.	
	1	2	1	2	1	2	1	2
levantamento bibliográfico								
elicitação de requisitos								
especificação da aplicação								
definição dos sites de busca								
obtenção das informações dos sites								
definição da técnica de mineração								
implementação o processo extração e agrupamento das informações								
testes								

Fonte: elaborado pelo autor.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo explorar os principais assuntos para realização deste trabalho. A seção 4.1 aborda mineração de dados. E, por fim, a seção 4.2 discorre sobre *web scraping*.

4.1 MINERAÇÃO DE DADOS

A mineração de dados é o processo de descoberta de padrões independentemente da quantidade de dados a ser analisados e do que se referem aos dados. Este processo utiliza-se de técnicas para reconhecimento de padrões, técnicas de estatística e matemática (LAROSE, 2005, p. 2).

Segundo Castanheira (2008, p. 18), a mineração de dados caracteriza-se pela utilização de algoritmos/técnicas para diante de um propósito extrair informações relevantes de uma base de dados, sendo que muitas das informações encontradas podem ser desconhecidas. Para o autor, a mineração de dados é a extração de informações úteis e desconhecidas de grandes bancos de dados, que conforme o propósito pode descobrir comportamentos que seriam dificilmente identificados por especialistas.

De acordo com Ferrari e Nunes (2017, p. 28), a mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases*, ou KDD). Embora muitos usem mineração de dados como

sinônimo de KDD, na primeira conferência internacional sobre KDD, realizada na cidade de Montreal, Canadá, em 1995, foi proposto que a terminologia descoberta de conhecimentos em bases de dados se referisse a todo o processo de extração de conhecimentos a partir de dados. Foi proposto também que a terminologia mineração de dados fosse empregada exclusivamente para a etapa de descoberta do processo de KDD, que inclui a seleção e integração das bases de dados, a limpeza da base, a seleção e transformação dos dados, a mineração e a avaliação dos dados.

Ferrari e Nunes (2017) apontam quatro etapas fortemente relacionadas e interdependentes, tanto que a inter-relações entre cada uma delas está ligada diretamente no resultado final da mineração, sendo elas: (i) base de dados, (ii) pré-processamento, (iii) mineração e (iv) validação. A base de dados possui valores quantitativos ou qualitativos de um conjunto de itens que permite uma recuperação eficiente dos dados. No pré-processamento existem etapas que consistem em limpeza, integração, seleção/redução e transformação dos dados oriundos da base de dados. A mineração é o processo responsável pela aplicação de algoritmos de extração de conhecimento a partir dos dados pré-processados. E, por fim, a etapa de validação consiste em identificar conhecimentos úteis e não triviais da mineração.

4.2 WEB SCRAPING

Segundo Mitchell (2019) a coleta automatizada de dados da internet é quase tão antiga quanto a própria internet. Embora *web scraping* não seja um termo novo, no passado, a prática era mais conhecida como *screen scraping*, *data mining*, *web harvesting* ou variações similares.

Barbosa e Cavalcanti (2020) definem *Web Scraping* como uma técnica ‘raspagem’ de dados diretamente da web, onde extraímos informações relevantes de sites através de *bots* [...]. Essa técnica é importante pois a análise auxilia a encontrar padrões e a tomar decisões com maior probabilidade de acerto. Para Borges e Ganimi (2018, p. 24) *Web scraping*, ou raspagem da Web é o processo de aquisição automatizada de dados não estruturados de site da internet, seguido do armazenamento estruturado.

Borges e Ganimi (2018) também apontam que há duas principais maneiras de utilização de web scraping sendo, (i) uma quando o próprio autor analisa a estrutura do site e implementa as técnicas de raspagem em seus alvos, e a outra (ii) quando são utilizados softwares terceiros e padronizados de raspagem. A vantagem da primeira opção é que a raspagem pode ser personalizada e adaptada a qualquer página web, porém demanda mais tempo e conhecimento em programação. Já na segunda, você ganha em tempo de implementação/utilização e perde espaço para customizações e adaptações aos seus sites alvo.

Após a extração, há diversas formas de armazenamento dos resultados, Borges e Ganimi (2018) mencionam duas, sendo elas: (i) armazenamento em arquivos do sistema operacional e (ii) armazenamento por meio de um Sistema de Gerenciamento de Banco de Dados (SGBD). Considerando que em arquivos do sistema operacional a manipulação por usuário é facilitada, os autores também apontaram que o armazenamento via SGBD dificulta a inserção ou manipulação de dados de forma errônea.

REFERÊNCIAS

- ALMEIDA, Leânia. **Cláusula de barreira: comportamento eleitoral e desempenho partidário nas eleições de 2002 e perspectivas para 2006**. Monografia (especialização) – Centro de Formação, Treinamento e Aperfeiçoamento (Cefor), da Câmara dos Deputados, 2007.
- ALMEIDA, Luis Gustavo. **Recuperação de dados pessoais na Web em redes sociais autenticadas**. 2018. Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2018.
- BARBOSA, Ana Beatriz Gomes; CAVALCANTI, Alessandro Bezerra. **Web Scraping e Análise de dados**, Anais do V CONAPESC... Campina Grande: Realize Editora, 2020.
- BARBOSA, Júlio César. **Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor**. 2018. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento.
- BORGES, Thiago da Cunha; GANIMI, Zeus Olenchuk. **Extração de dados com web scraping para análise da variação de preço de veículos automotores**. 2018
- CAMARGO, Pedro Henrique Gomes *et al.* **Spiders de extração de dados do turismo nacional com o framework Scrapy**. XIX Encontro – Congresso de Computação e Tecnologias da Informação, Canoas – RS. 2017.
- CASTANHEIRA, Luciana G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões**. 2008. 91 f. Dissertação (Mestrado em Engenharia Elétrica) - Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte
- FERRARI, Daniel Gomes; SILVA, Leandro Nunes de Castro. **Introdução a mineração de dados**. Saraiva Educação SA, 2017.
- LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining**. New Jersey: John Wiley & Sons, 2005.
- MITCHELL, Ryan. **Web Scraping com Python: Coletando mais dados da web moderna**. Novatec Editora, 2019.
- NOGUEIRA, Thaís Caldoncelli. **Mineração de texto para descoberta de conhecimento em bulas de medicamentos**.
- SILVA, Denisson; MEIRELES, Fernando. Ciência Política na era do Big Data: automação na coleta de dados digitais. **Revista Política Hoje**, v. 24, n. 2, p. 87-102, 2016.

ASSINATURAS

(Atenção: todas as folhas devem estar rubricadas)

Assinatura do(a) Aluno(a): _____

Assinatura do(a) Orientador(a): _____

Assinatura do(a) Coorientador(a) (se houver): _____

Observações do orientador em relação a itens não atendidos do pré-projeto (se houver):

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Acadêmico: Jean Patrick Scherer

Avaliador(a): _____

ASPECTOS AVALIADOS ¹		atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado? O problema está claramente formulado?			
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado? Os objetivos específicos são coerentes com o objetivo principal?			
	3. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?			
	4. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada? São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta? São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	5. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?			
	6. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC? Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?			
	7. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC? As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?			
	8. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica? A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			
	9. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?			
	10. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?			
ASPECTOS METODOLÓGICOS	11. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT? As citações obedecem às normas da ABNT? Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?			

PARECER – PROFESSOR DE TCC I OU COORDENADOR DE TCC (PREENCHER APENAS NO PROJETO):

O projeto de TCC será reprovado se:

- qualquer um dos itens tiver resposta **NÃO ATENDE**;
- pelo menos **4 (quatro)** itens dos **ASPECTOS TÉCNICOS** tiverem resposta **ATENDE PARCIALMENTE**; ou
- pelo menos **4 (quatro)** itens dos **ASPECTOS METODOLÓGICOS** tiverem resposta **ATENDE PARCIALMENTE**.

PARECER: () APROVADO () REPROVADO

Assinatura: _____ Data: _____

¹ Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR AVALIADOR

Acadêmico: Jean Patrick Scherer _____

Avaliador(a): Alexander Roberto Valdameri

ASPECTOS AVALIADOS ¹		atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	X		
	O problema está claramente formulado?		X	
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	X		
	Os objetivos específicos são coerentes com o objetivo principal?	X		
	3. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?	X		
	4. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?	X		
	São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?	x		
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?	X		
	5. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?			X
	6. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?	X		
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?	X		
	7. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?		X	
ASPECTOS METODOLÓGICOS	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?	x		
	8. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?		X	
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?		x	

PARECER – PROFESSOR AVALIADOR: (PREENCHER APENAS NO PROJETO)

O projeto de TCC deverá ser revisado, isto é, necessita de complementação, se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos 5 (cinco) tiverem resposta ATENDE PARCIALMENTE.

PARECER: () APROVADO (X) REPROVADO

Assinatura:



Data: 10/05/2021

¹ Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

CURSO DE SISTEMAS DE INFORMAÇÃO – TCC ACADÊMICO	
(X) PRÉ-PROJETO () PROJETO	ANO/SEMESTRE: 2021/01

REPOSITÓRIO DE INFORMAÇÕES DE PARLAMENTARES: UM DOSSIÊ PÚBLICO ONLINE

Jean Patrick Scherer

Aurélio Faustino Hoppe - Orientador

1 INTRODUÇÃO

De acordo com Barbosa (2018), não há precedentes sobre a frequência e volume com que as pessoas publicam informações virtualmente, cujo maior atrativo é que grande parte dos usuários da internet postam suas opiniões, expectativas, elogios e frustrações, concebendo, assim, vasta base de dados primários sobre diversos produtos e serviços.

Segundo Nogueira (2014), com o advento da Internet, o volume de informações textuais vem crescendo exponencialmente. Tais informações são encontradas em maiores quantidades em diversas fontes. Segundo o autor, através das informações textuais, conhecidas também como não estruturadas, é possível extrair conhecimento útil e implícito que, devido ao grande volume são impossíveis de serem processadas por um ser humano.

Por outro lado, segundo Almeida (2007), as escolhas de representantes políticos no Brasil sempre foi um trabalho árduo para o eleitor. Na maioria das vezes, não se é dada a devida atenção a esta ação que pode gerar inúmeras consequências para que a fez e para as demais que a rodeiam. Porém, nos últimos anos a política está em evidência no país, devido a troca da detenção do poder entre os extremos da democracia (direta e esquerda da política nacional) que ocasionou, como esperado, uma reviravolta no país e na percepção da população em relação ao voto.

Almeida (2007) também aponta que em toda eleição surgem novos candidatos para serem escolhidos entre a população, com o discurso quase que padrão de educação e segurança. Mas até onde isto realmente é verdade? Como posso estar em garantia que esta mesma pessoa que está me prometendo investimentos na educação, por exemplo, a alguns anos atrás não vetou algum projeto que destinaria recursos para esta mesma educação? Até onde é verdade este discurso e até onde não? Ou seja, percebe-se que há carência de informação de fácil acesso aos eleitores sobre os candidatos e ~~isto~~ isso, impulsiona escolhas equivocadas ou não baseadas em fatos comprovados por veículos de imprensa em relação aos candidatos.

Diante desta realidade, este trabalho propõem o desenvolvimento de uma aplicação que implementará técnicas de mineração de dados para apoiar a extração e o agrupamento de informações históricas de seus representantes, disponibilizando-as na forma de um dossiê político.

Comentado [AS1]: Frase longa. Rever a redação. Não se faz parágrafo com uma única frase.

Comentado [AS2]: Como?

Comentado [AS3]: Rever a redação.

Comentado [AS4]: Normalmente em documento científico não se faz perguntas diretas.

1.1 OBJETIVOS

O objetivo deste trabalho é disponibilizar uma aplicação para agrupar informações históricas de seus representantes, apresentando-as na forma de um dossiê político.

Os objetivos específicos são:

- apresentar, de forma clara e comprovada, ações políticas realizadas pelo candidato;
- facilitar e auxiliar na escolha de representantes políticos, conforme as preferências do eleitor;
- disponibilizar, de forma centralizada, conteúdo de fácil e ampla consulta.

Comentado [AS5]: Não é objetivo. Os objetivos específicos devem ser verificáveis no fim do trabalho.

Comentado [AS6]: Isso é um resultado e não objetivo

2 TRABALHOS CORRELATOS

Neste capítulo são apresentados três trabalhos correlatos, que possuem características semelhantes à proposta deste trabalho. A seção 2.1 detalha o SentimentALL (CAMARGO *et al.*, 2017) que extrai dados do turismo nacional com o *framework* Scrapy. Na seção 2.2 é descrito a aplicação ALUMNI tool (ALMEIDA, 2018) que recupera dados pessoais na Web em redes sociais autenticadas. Por fim, a seção 2.3, aborda o artigo ciência política na era da big data (SILVA; MEIRELES, 2016) que visa automatizar a coleta de dados políticos na internet para fins acadêmicos utilizando a linguagem de programação R.

2.1 SPIDERS DE EXTRAÇÃO DE DADOS DO TURISMO NACIONAL COM O FRAMEWORK SCRAPY

Camargo *et al.* (2017) desenvolveram um módulo de extração para coletar dados do contexto do turismo nacional. Os autores utilizaram o *framework* Scrapy para extração de dados do site TripAdvisor. O *framework* Scrapy permite definições e implementações de regras de web *crawling* e web *scraping*, denominadas *spiders*, escrito em Python possibilitando o alcance de grandes massas de dados a partir da web ou outras fontes.

Comentado [AS7]: Do que?

Segundo Camargo *et al.* (2017), as coletas foram divididas em duas partes, sendo que para cada uma foram criadas três *spiders*, uma para cada tipo de busca, atrações, hotéis e restaurantes. Primeiramente, as *spiders* realizam a sondagem, coletando a quantidade de avaliações para cada tipo de busca para todas as cidades do Brasil. Posteriormente, foram filtradas apenas as extrações das cem cidades com a maior quantidade de avaliações.

Camargo *et al.* (2017) apontam que inicialmente deve-se informar o link da página ao ser feita a navegação do *spider* até encontrar o objeto de extração (atração, hotel ou restaurante). No caso do TripAdvisor, os endereços relacionados seguem um padrão de Uniform Resource Locator (URL) baseado pelo identificador ao{X} onde, X é uma progressão

Comentado [AS8]: Nomes de pacotes, classes, entidades, atributos, métodos ou diálogos de interface devem ser escritos em fonte courier (tem estilo).

Comentado [AS9]: Coloca em fonte courier (tem estilo)

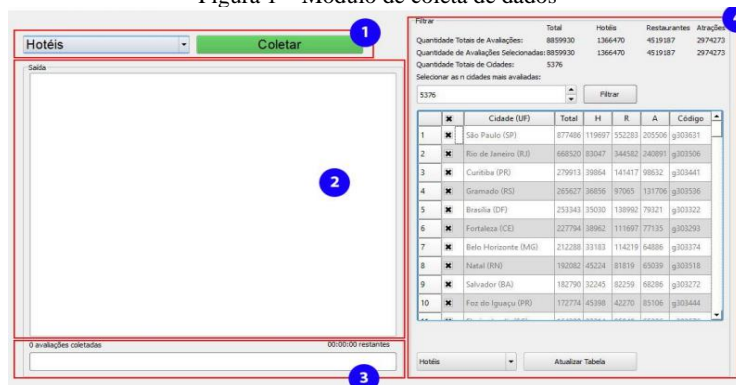
incrementada de vinte em vinte. A partir desta informação é possível fazer a extração das respectivas cidades. Caso houver paginação, o X mencionado anteriormente, deve ser incrementado de trinta em trinta.

De acordo com Camargo *et al.* (2017), o objetivo da segunda **spider** é a extração dos dados de cidade, estado, nome do objeto, quantidade de avaliações e quantidade de avaliações em português, armazenando em um arquivo Comma-Separated-Values (CSV). Esta mesma **spider** é responsável pela extração do código da cidade e código do objeto que são encontradas na URL da página.

Na terceira **spider**, segundo Camargo *et al.* (2017) é necessário definir os links iniciais de busca. Para isso foi criado uma função para montar os links de cada objeto com base num conjunto de cidades selecionadas pelo usuário. Os códigos de cidades e o tipo de objeto são extraídos do arquivo CSV gerado anteriormente. O padrão de link para esta etapa consiste em: `tripadvisor.com.br/{tipoobjeto}_Review{CÓDIGOCIDADE}{CÓDIGOOBJETO}`. Os dados extraídos e salvos em arquivo CSV são: Objeto, Código do Objeto, Tipo de objeto, Cidade, Código da cidade, Estado, País, Nome do autor, Cidade do autor, Nível do autor, Ano de cadastro do autor no site, Código do autor, Título, Nota, Data de postagem, Comentário e Código de avaliação.

Na Figura 1, é apresentada a interface gráfica do módulo de coleta. Através dela é possível iniciar e interromper a coleta dos dados (Figura 1 item 1), visualizar os logs gerados pelo *framework* Scrapy (Figura 1 item 2), uma barra de status do processo juntamente com o tempo estimado para finalização e quantidade de avaliações coletadas (Figura 1 item 3) e por fim, dados referentes a coleta (armazenados em CSV) são exibidos possibilidade a realização de filtro por cidade (Figura 1 item 4)

Figura 1 – Módulo de coleta de dados



Fonte: Camargo *et al.* (2017).

Formatado: Realce

Comentado [AS10]: Rever a redação

Formatado: Realce

Formatado: Realce

Comentado [AS11]: Coloca em fonte courier (tem estilo)

Comentado [AS12]: Esta informação é realmente importante?

Camargo *et al.* (2017) apontam como possíveis melhorias a curto e médio prazo, (i) a possibilidade de se desenvolver **spiders** utilizando interface gráfica; (ii) desenvolvimento de rotinas padrões de pré-processamento para diminuir o tempo gasto no tratamento dos dados e, (iii) possibilidade de se trabalhar com mais contextos e não somente com o turismo.

Formatado: Realce

2.2 RECUPERAÇÃO DE DADOS PESSOAIS NA WEB EM REDES SOCIAIS AUTENTICADAS

Almeida (2018) desenvolveu a aplicação ALUMNI tool, tendo como objetivo a recuperação de dados pessoais na Web em redes sociais autenticadas. O autor implementou um *web scraping*, combinando um robô de busca em PHP e um programa para testes funcionais para coletar informações publicadas por alumni (plural de ex-aluno em latim) em redes sociais.

Segundo Almeida (2018), a grande maioria dos *crawlers* foram escritos em Java pois não se encontrou nenhum método automatizado de extração de informações de usuários de redes sociais diretamente da web, principalmente, aquelas que possuem autenticação.

Almeida (2018) subdividiu a aplicação em duas partes, sendo elas: (i) o robô, conjunto de programas em PHP vinculado com o software Selenium, que navega pelas URL's do site LinkedIn para recuperar os perfis de ex-alunos e seus respectivos ID's, que ficarão disponíveis via requisição HTTP em porta 4444 para a aplicação web ~~Fe~~, (ii) a aplicação web, que irá realizar a análise e processamento dos dados coletados, resumizando e disponibilizando-os de forma visual, no formato HTML, gráficos e tabelas. Na aplicação web, o autor também utilizou a linguagem de programação PHP, porém, utilizando o padrão de projeto MVC.

Na etapa de autenticação da plataforma LinkedIn, Almeida (2018) utilizou suas próprias informações de acesso, de usuário e senha. Relatando que os desafios encontrados foram a manutenção da sessão ativa no navegador, a geração de conteúdo HTML via AJAX e eventuais barreiras que o próprio LinkedIn adota contra robôs de busca. Para resolver a situação de manter a sessão ativa, o autor utilizou a ferramenta Selenium WebDriver, sendo responsável por carregar a janela do navegador e de forma autônoma fornecer os dados de login e senha válido da plataforma e, posteriormente, extrair e armazenando o conteúdo HTML em um banco de dados. Com estes dados, o robô executa um laço de iteração para percorrer todas as URL's recuperadas utilizando a biblioteca DOMDocument da linguagem PHP combinada com utilização de expressões regulares, realizando a mineração de dados. Os dados coletados foram armazenados em banco de dados MySQL, totalizando 9 tabelas, que a aplicação final web terá acesso para realizar consultas em T-SQL.

Comentado [AS13]: Rever a redação.

Comentado [AS14]: Rever a redação.

A validação e limpeza dos dados foi feita através de regras T-SQL combinadas com scripts em PHP. Exemplos de validação e limpeza seriam dados duplicados de perfis e históricos profissionais sem vínculo direto com perfil de usuário na plataforma. Neste processo de validação, foram encontrados alguns pontos de dificuldade, sendo eles: *encoding* de dados, utilização de espaçamento único entre os nomes, utilização de letras maiúsculas e minúsculas, não utilização de acentos.

Segundo Almeida (2018) foram coletados no total, 388.868 registros, sendo eles aproximadamente, 118.000 registros de **edus** (histórico acadêmico), 200.000 **jobs** (histórico profissional), 11.000 empresas, 580 escolas e 300 cursos de graduação. Almeida (2018) também aponta que a ferramenta foi muito eficaz, atingindo 88% de **hashes** (perfis) e 66% de **alumni** (ex-alunos) com um tempo de processamento de aproximadamente 3 meses de duração.

Almeida (2018) sugere como trabalhos futuros: (i) a verificação da veracidade das informações inseridas nos perfis das plataformas; (ii) a mineração em mais modelos de web sites para acrescer a quantidade da amostragem; e, (iii) a utilização de informações complementares disponibilizadas no perfil dos usuários, tal como a informação de **skills** que necessitam ser endossadas por outros perfis, aumentando assim, a integridade das informações contidas.

2.3 CIÊNCIA POLÍTICA NA ERA DO BIG DATA: AUTOMAÇÃO NA COLETA DE DADOS DIGITAIS

Silva e Meireles (2016) utilizaram a linguagem de programação R para a coleta de informações da Assembleia Legislativa de Minas Gerais, além de mais sete estados, sendo coletadas informações de **PL's PLC's e PEC's** realizadas pelas instituições. Os autores separaram as coletas em dois conjuntos de dados: (i) número, do ano, e do tipo, e link das proposições de interesse e, (ii) algumas outras informações oriundas do link como, tipo, sigla, ano, número, autor, partido, ementa, local fim e situação.

Segundo Silva e Meireles (2016), o primeiro conjunto de dados foi coletado diretamente da página web da Assembleia de Minas Gerais, ou seja, mineraram os conteúdos diretamente da página HTML. Já para o segundo, utilizaram a API disponibilizada pela própria Assembleia, tendo acesso a arquivos no formato JSON e XML. Os autores utilizaram como ferramenta de manipulação o pacote "RSelenium".

Silva e Meireles (2016) trabalharam bastante na parte de mineração do HTML. No qual, a partir do link de acesso, o conteúdo HTML **era-foi** carregado para processamento no R e posteriormente, **foi** realizado o processo de seleção de conteúdo com métodos de eliminação

Comentado [AS15]: itálico

Comentado [AS16]: itálico

Comentado [AS17]: itálico

Comentado [AS18]: coloque a palavra em português

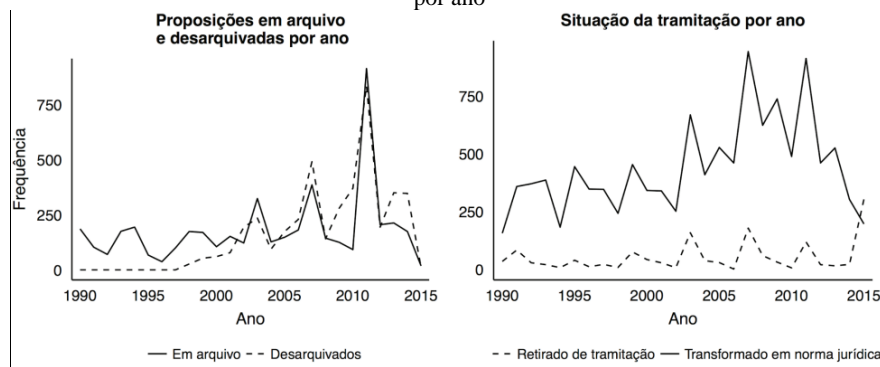
Comentado [AS19]: o que estes termos significam?

Comentado [AS20]: Esta informação é relevante? Se sim, deve ser escrita de modo formal.

e contagem de caracteres. Ainda segundo os autores, o processo de coleta de informações foi facilitado, uma vez que, sabendo do modelo de retorno das requisições das API's, basta é possível extrair o valor do campo pelo nome.

De acordo com Silva e Meireles (2016), no total, foram capturadas 25.418 proposições em aproximadamente 2h. Dados gerados entre 1990 e 2015. Após a coleta, os dados foram sumarizados em gráficos para análises, conforme mostra a Figura 2.

Figura 2 – Distribuição da proposição em arquivo ou desarquivadas e situação de tramitação, por ano



Segundo Silva e Meireles (2016), o número de proposições em arquivo é maior do que as desarquivadas até 2000. Depois desse ponto, a distribuição é semelhante nos dois tipos. O único ano em que houve maior quantidade de retiradas de tramitação foi em 2015. Nos demais, houve quantidade relevante maior de tramitações que se tornaram normas nas proposições legislativas.

Como principal argumentação para o trabalho, Silva e Meireles (2016) ponderam as características de “eliminação de erros de imputação, digitação ou até mesmo do copiar e colar”, além do tempo economizado na etapa de pesquisa empírica. Eles Os autores também ressaltam que, grande parte do volume ~~de~~ dados não estão disponíveis para download em formatos aceitos por softwares estatísticos, mesmo nas plataformas que disponibilizam uma API. Por fim, Silva e Meireles (2016) sugerem melhorias na visualização e filtros dos resultados coletados e, também quanto ao recebimento de informações de outras assembleias do território nacional.

Comentado [AS21]: Rever a redação.

Comentado [AS22]: Rever a redação.

Comentado [AS23]: Citação direta tem que ter o número da página

3 PROPOSTA

A seguir é apresentada a justificativa para o desenvolvimento desse trabalho, os principais requisitos e a metodologia de desenvolvimento que será utilizada. Também são relacionados os assuntos e as fontes bibliográficas que irão fundamentar o estudo proposto.

3.1 JUSTIFICATIVA

No Quadro 1, é feita comparação entre três trabalhos correlatos voltados a estas técnicas computacionais, listando diferenças pontuais em suas técnicas, finalidades e assuntos envolvidos na mineração. As linhas representam as características e as colunas os trabalhos.

Quadro 1 – Comparativo dos trabalhos correlatos

Características \ Trabalhos	Camargo <i>et al.</i> (2017)	Almeida (2018)	Silva e Meireles (2016)
Realizam a extração de dados	Sim	Sim	Sim
Utilizam técnicas de mineração de dados	Sim	Sim	Sim
Sumarização os dados coletados	Não	Sim	Sim
Dados vinculados a política brasileira	Não	Não	Sim
Api utilizadas	Scrapy	RSelenium	RSelenium

Fonte: elaborado pelo autor.

A partir do Quadro 1, pode-se observar que todas as aplicações extraem dados de sites e utilizam de técnicas de mineração para alcançarem seus resultados. Além disso, também percebe-se que apenas o trabalho de Silva e Meireles (2016) realizam a manipulação de dados da política brasileira. Camargo *et al.* (2017) optaram que extração e sumarização de informações turísticas. Já Almeida (2018) pela coleta informações das redes sociais.

O trabalho proposto é relevante pois agrupará informações históricas de seus representantes, apresentando-as na forma de um dossiê político, sendo exposto de forma categorizada (negativamente ou positivamente) e de fácil leitura, as ações realizadas ou ações vinculadas a determinado parlamentar, estando ou não em processo eleitoral. Também serão vinculados *links* de matérias oriundas de portais oficiais de notícias aos nomes dos parlamentares envolvidos diretamente ou indiretamente no conteúdo da matéria disponível no *link* informado. Dessa forma, espera-se que os eleitores possam ter acesso facilitado a informações sobre seus candidatos e discursos apresentados, agregando na qualidade das argumentações em discussões construtivas, debates e nas escolhas feitas pelos eleitores nas urnas. Acredita-se que através de um *web crawling* e técnicas de mineração de dados será possível disponibilizar na forma de um repositório centralizado, informações agrupadas, simplificadas e de fácil acesso, visando beneficiar de forma geral os eleitores do país.

Comentado [AS24]: Aqui no quadro poderias especificar as técnicas utilizadas ao invés de colocar sim e não

3.2 REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO

A aplicação de mineração de dados deverá:

- a) permitir o cadastro de sites ao qual serão extraídas as informações (Requisito Funcional – RF);
- b) extrair as informações dos sites através de um web scraping (RF);
- c) implementar o processo de mineração e as suas etapas, sendo elas: seleção, pré-processamento, limpeza, transformação, mineração de dados e interpretação dos dados (RF);
- d) implementar a técnica de agrupamento para realizar a mineração dos dados (RF);
- e) disponibilizar as informações na forma de um repositório/dossiê com link para notícias dos mandatários (RF);
- f) permitir ao usuário consultar as informações a partir do nome completo do mandatário (RF);
- g) ser implementada utilizando o ambiente de desenvolvimento Microsoft Visual Studio 2010 e na plataforma ASP.NET (Requisito Não-Funcional - RNF);
- h) ser implementada utilizando Kubernetes ou Docker Swarm (RNF).

3.3 METODOLOGIA

O trabalho será desenvolvido observando as seguintes etapas:

- a) levantamento bibliográfico: realizar o levantamento bibliográfico sobre mineração de dados, web scraping e trabalhos correlatos;
- b) elicitação de requisitos: baseando-se no levantamento bibliográfico, refinar os requisitos propostos para a mineração de dados proposto;
- c) especificação da aplicação: especificar a aplicação com análise orientada a objetos utilizando a Unified Modeling Language (UML). Utilizar a ferramenta Enterprise Architect 7.5 (EA) para o desenvolvimento dos diagramas de casos de uso e classes;
- d) definição dos sites de busca: analisar e escolher quais sites disponibilizam informações são relevantes sobre os mandatários;
- e) obtenção das informações dos sites: desenvolvimento de um web scraping na linguagem ASP.NET, que busque as informações a partir da etapa (d);
- f) definição da técnica de mineração: pesquisar, analisar a técnica de agrupamento que irá ser utilizada para compor a aplicação de mineração dos dados, levando em consideração seu desempenho e sua eficácia;
- g) implementação o processo de extração e agrupamento das informações: a partir das

etapas (e) e (f) implementar o processo de mineração e agrupamento das informações dos mandatários utilizando a plataforma ASP.NET com o ambiente Microsoft Visual Studio 2010;

- h) testes: realizar testes com a técnica implementada na etapa (g) verificando se a mesma é eficiente e possui baixo custo computacional para realizar a extração das informações. Além disso também será verificado a assertividade dos agrupamentos de notícias associadas aos mandatários.

As etapas serão realizadas nos períodos relacionados no Quadro 2.

Quadro 2 – Cronograma de atividades a serem realizadas

etapas / quinzenas	2021							
	ago.		set.		out.		nov.	
	1	2	1	2	1	2	1	2
levantamento bibliográfico								
elicitação de requisitos								
especificação da aplicação								
definição dos sites de busca								
obtenção das informações dos sites								
definição da técnica de mineração								
implementação o processo extração e agrupamento das informações								
testes								

Fonte: elaborado pelo autor.

Comentado [AS25]: Rever a redação

Comentado [AS26]: Podes usar a 2ª quinzena de novembro.

4 REVISÃO BIBLIOGRÁFICA

Este capítulo tem como objetivo explorar os principais assuntos para realização deste trabalho. A seção 4.1 aborda mineração de dados. E, por fim, a seção 4.2 discorre sobre *web scraping*.

4.1 MINERAÇÃO DE DADOS

A mineração de dados é o processo de descoberta de padrões independentemente da quantidade de dados a ser analisados e do que se referem aos dados. Este processo utiliza-se de técnicas para reconhecimento de padrões, técnicas de estatística e matemática (LAROSE, 2005, p. 2).

Segundo Castanheira (2008, p. 18), a mineração de dados caracteriza-se pela utilização de algoritmos/técnicas para diante de um propósito extrair informações relevantes de uma base de dados, sendo que muitas das informações encontradas podem ser desconhecidas. Para o autor, a mineração de dados é a extração de informações úteis e desconhecidas de grandes

Comentado [AS27]: No pré-projeto devem ser descritos brevemente os assuntos que fundamentarão o estudo a ser realizado, relacionando a(s) principal(is) referência(s) bibliográfica(s), a(s) qual(is) deve(m) constar nas REFERÊNCIAS. Cada assunto abordado deve ser descrito em um parágrafo.

bancos de dados, que conforme o propósito pode descobrir comportamentos que seriam dificilmente identificados por especialistas.

De acordo com Ferrari e Nunes (2017, p. 28), a mineração de dados é parte integrante de um processo mais amplo, conhecido como descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases*, ou KDD). Embora muitos usem mineração de dados como sinônimo de KDD, na primeira conferência internacional sobre KDD, realizada na cidade de Montreal, Canadá, em 1995, foi proposto que a terminologia descoberta de conhecimentos em bases de dados se referisse a todo o processo de extração de conhecimentos a partir de dados. Foi proposto também que a terminologia mineração de dados fosse empregada exclusivamente para a etapa de descoberta do processo de KDD, que inclui a seleção e integração das bases de dados, a limpeza da base, a seleção e transformação dos dados, a mineração e a avaliação dos dados.

Ferrari e Nunes (2017) apontam quatro etapas fortemente relacionadas e interdependentes, tanto que a inter-relações entre cada uma delas está ligada diretamente no resultado final da mineração, sendo elas: (i) base de dados, (ii) pré-processamento, (iii) mineração e (iv) validação. A base de dados possui valores quantitativos ou qualitativos de um conjunto de itens que permite uma recuperação eficiente dos dados. No pré-processamento existem etapas que consistem em limpeza, integração, seleção/redução e transformação dos dados oriundos da base de dados. A mineração é o processo responsável pela aplicação de algoritmos de extração de conhecimento a partir dos dados pré-processados. E, por fim, a etapa de validação consiste em identificar conhecimentos úteis e não triviais da mineração.

4.2 WEB SCRAPING

Segundo Mitchell (2019) a coleta automatizada de dados da internet é quase tão antiga quanto a própria internet. Embora *web scraping* não seja um termo novo, no passado, a prática era mais conhecida como *screen scraping*, *data mining*, *web harvesting* ou variações similares.

Barbosa e Cavalcanti (2020) definem *Web Scraping* como uma técnica ‘raspagem’ de dados diretamente da web, onde extraímos informações relevantes de sites através de *bots* [...]. Essa técnica é importante pois a análise auxilia a encontrar padrões e a tomar decisões com maior probabilidade de acerto. Para Borges e Ganimi (2018, p. 24) *Web scraping*, ou raspagem da Web é o processo de aquisição automatizada de dados não estruturados de site da internet, seguido do armazenamento estruturado.

Borges e Ganimi (2018) também apontam que há duas principais maneiras de utilização de *web scraping* sendo, (i) uma quando o próprio autor analisa a estrutura do site e implementa

Comentado [AS28]: Não se coloca o número de página na citação indireta. A referência do autor não está de acordo. Ver comentário na seção Referências.

Comentado [AS29]: É realmente relevante esta informação?

Comentado [AS30]: É uma citação direta? Se sim, não está de acordo coma norma

Comentado [AS31]: No texto todo você coloca muitos desses itens. Deve-se evitar. Neste caso aqui por exemplo é desnecessário.

Formatado: Fonte: Itálico

as técnicas de raspagem em seus alvos, e a outra (ii) quando são utilizados softwares terceiros e padronizados de raspagem. A vantagem da primeira opção é que a raspagem pode ser personalizada e adaptada a qualquer página web, porém demanda mais tempo e conhecimento em programação. Já na segunda, você ganha em tempo de implementação/utilização e perde espaço para customizações e adaptações aos seus sites alvo.

Após a extração, há diversas formas de armazenamento dos resultados. Borges e Ganini (2018) mencionam duas, sendo elas: (i) armazenamento em arquivos do sistema operacional e (ii) armazenamento por meio de um Sistema de Gerenciamento de Banco de Dados (SGBD). Considerando que em arquivos do sistema operacional a manipulação por usuário é facilitada, os autores também apontaram que o armazenamento via SGBD dificulta a inserção ou manipulação de dados de forma errônea.

REFERÊNCIAS

ALMEIDA, Leânia. **Cláusula de barreira: comportamento eleitoral e desempenho partidário nas eleições de 2002 e perspectivas para 2006**. Monografia (especialização) – Centro de Formação, Treinamento e Aperfeiçoamento (Cefor), da Câmara dos Deputados, 2007.

ALMEIDA, Luis Gustavo. **Recuperação de dados pessoais na Web em redes sociais autenticadas**. 2018. Dissertação (Mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2018.

BARBOSA, Ana Beatriz Gomes; CAVALCANTI, Alessandro Bezerra. **Web Scraping e Análise de dados**, Anais do V CONAPESC... Campina Grande: Realize Editora, 2020.

BARBOSA, Júlio César. **Mineração de texto: uso de técnicas de processamento de linguagem natural para suporte à geração de projeções baseadas em opiniões do consumidor**. 2018. Tese de Doutorado. Mestrado em Sistemas de Informação e Gestão do Conhecimento.

BORGES, Thiago da Cunha; GANIMI, Zeus Olenchuk. **Extração de dados com web scraping para análise da variação de preço de veículos automotores**. 2018

CAMARGO, Pedro Henrique Gomes *et al.* **Spiders de extração de dados do turismo nacional com o framework Scrapy**. XIX Encontro – Congresso de Computação e Tecnologias da Informação, Canoas – RS. 2017.

CASTANHEIRA, Luciana G. **Aplicação de técnicas de mineração de dados em problemas de classificação de padrões**. 2008. 91 f. Dissertação (Mestrado em Engenharia Elétrica) - Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal de Minas Gerais, Belo Horizonte

FERRARI, Daniel Gomes; SILVA, Leandro Nunes de Castro. **Introdução a mineração de dados**. Saraiva Educação SA, 2017.

LAROSE, Daniel T. **Discovering knowledge in data: an introduction to data mining**. New Jersey: John Wiley & Sons, 2005.

MITCHELL, Ryan. **Web Scraping com Python: Coletando mais dados da web moderna**. Novatec Editora, 2019.

Comentado [AS32]: Texto deve ser escrito no impessoal.

Comentado [AS33]: Ver meu comentário anterior

Formatado: Fonte: Negrito

Comentado [AS34]: Essa referência está com a cor da fonte errada.

A forma como você referencia o autor no texto está incorreta.

Comentado [AS35]: Os prenomes e demais sobrenomes dos autores podem ser ou não abreviados, desde que seja seguido o mesmo padrão de abreviação.

Comentado [AS36]: cor

NOGUEIRA, Thaís Caldoncelli. **Mineração de texto para descoberta de conhecimento em bulas de medicamentos.**

SILVA, Denisson; MEIRELES, Fernando. Ciência Política na era do Big Data: automação na coleta de dados digitais. **Revista Política Hoje**, v. 24, n. 2, p. 87-102, 2016.

Comentado [AS37]: cor

ASSINATURAS

(Atenção: todas as folhas devem estar rubricadas)

Assinatura do(a) Aluno(a): _____

Assinatura do(a) Orientador(a): _____

Assinatura do(a) Coorientador(a) (se houver): _____

Observações do orientador em relação a itens não atendidos do pré-projeto (se houver):

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR TCC I

Acadêmico: Jean Patrick Scherer _____

Avaliador(a): Andreza Sartori _____

ASPECTOS AVALIADOS ¹		atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?	X		
	O problema está claramente formulado?	X		
	2. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?	X		
	Os objetivos específicos são coerentes com o objetivo principal?			X
	3. JUSTIFICATIVA São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?	X		
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?	X		
ASPECTOS METODOLÓGICOS	4. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?	X		
	Os métodos, recursos e o cronograma estão devidamente apresentados?	X		
	5. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?	X		
	6. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?		X	
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?		X	
	7. ORGANIZAÇÃO E APRESENTAÇÃO GRÁFICA DO TEXTO A organização e apresentação dos capítulos, seções, subseções e parágrafos estão de acordo com o modelo estabelecido?	X		
	8. ILUSTRAÇÕES (figuras, quadros, tabelas) As ilustrações são legíveis e obedecem às normas da ABNT?	X		
	9. REFERÊNCIAS E CITAÇÕES As referências obedecem às normas da ABNT?		X	
	As citações obedecem às normas da ABNT?		X	
Todos os documentos citados foram referenciados e vice-versa, isto é, as citações e referências são consistentes?			X	

PARECER – PROFESSOR DE TCC I OU COORDENADOR DE TCC (PREENCHER APENAS NO PROJETO):

O projeto de TCC será reprovado se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos 4 (quatro) itens dos **ASPECTOS TÉCNICOS** tiverem resposta ATENDE PARCIALMENTE; ou
- pelo menos 4 (quatro) itens dos **ASPECTOS METODOLÓGICOS** tiverem resposta ATENDE PARCIALMENTE.

PARECER: () APROVADO () REPROVADO

Assinatura: _____ Data: 07/05/2021 _____

¹ Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.

FORMULÁRIO DE AVALIAÇÃO – PROFESSOR AVALIADOR

Acadêmico: Jean Patrick Scherer

Avaliador(a): _____

ASPECTOS AVALIADOS ¹		atende	atende parcialmente	não atende
ASPECTOS TÉCNICOS	1. INTRODUÇÃO O tema de pesquisa está devidamente contextualizado/delimitado?			
	O problema está claramente formulado?			
	1. OBJETIVOS O objetivo principal está claramente definido e é passível de ser alcançado?			
	Os objetivos específicos são coerentes com o objetivo principal?			
	2. TRABALHOS CORRELATOS São apresentados trabalhos correlatos, bem como descritas as principais funcionalidades e os pontos fortes e fracos?			
	3. JUSTIFICATIVA Foi apresentado e discutido um quadro relacionando os trabalhos correlatos e suas principais funcionalidades com a proposta apresentada?			
	São apresentados argumentos científicos, técnicos ou metodológicos que justificam a proposta?			
	São apresentadas as contribuições teóricas, práticas ou sociais que justificam a proposta?			
	4. REQUISITOS PRINCIPAIS DO PROBLEMA A SER TRABALHADO Os requisitos funcionais e não funcionais foram claramente descritos?			
	5. METODOLOGIA Foram relacionadas todas as etapas necessárias para o desenvolvimento do TCC?			
	Os métodos, recursos e o cronograma estão devidamente apresentados e são compatíveis com a metodologia proposta?			
	6. REVISÃO BIBLIOGRÁFICA (atenção para a diferença de conteúdo entre projeto e pré-projeto) Os assuntos apresentados são suficientes e têm relação com o tema do TCC?			
ASPECTOS METODOLÓGICOS	As referências contemplam adequadamente os assuntos abordados (são indicadas obras atualizadas e as mais importantes da área)?			
	7. LINGUAGEM USADA (redação) O texto completo é coerente e redigido corretamente em língua portuguesa, usando linguagem formal/científica?			
	A exposição do assunto é ordenada (as ideias estão bem encadeadas e a linguagem utilizada é clara)?			

PARECER – PROFESSOR AVALIADOR: (PREENCHER APENAS NO PROJETO)

O projeto de TCC ser deverá ser revisado, isto é, necessita de complementação, se:

- qualquer um dos itens tiver resposta NÃO ATENDE;
- pelo menos 5 (cinco) tiverem resposta ATENDE PARCIALMENTE.

PARECER: () APROVADO () REPROVADO

Assinatura: _____ Data: _____

¹ Quando o avaliador marcar algum item como atende parcialmente ou não atende, deve obrigatoriamente indicar os motivos no texto, para que o aluno saiba o porquê da avaliação.