IME INSTITUTO DE MATEMÁTICA E ESTATÍSTICA UNIVERSIDADE DE SÃO PAULO

CARIME BUMARUF

QUAIS FATORES DEVEM SER CONSIDERADOS NA CONSTRUÇÃO DE UM CORPUS DE NOTÍCIAS EM PORTUGUÊS?

Trabalho de conclusão do curso de Bacharelado em Ciência da Computação do Instituto de Matemática e Estatística da Universidade de São Paulo, como pré-requisito para a obtenção do grau de Bacharela em Ciência da Computação.

Orientador: Prof. Dr. Denis D. Mauá

São Paulo – SP

2021

Resumo

BUMARUF, Carime. **Quais fatores devem ser considerados na construção de um corpus de notícias em português?** 2021. 43 fls. Trabalho de conclusão de curso (Bacharelado em Ciência da Computação) – Instituto de Matemática e Estatística da Universidade de São Paulo, São Paulo, 2021.

A busca por dados de alta qualidade que alimentem modelos de Inteligência Artificial está começando a ganhar mais adeptos, tanto na academia quanto nas empresas. Há uma campanha pertinente pregando que, para que os modelos produzam resultados mais precisos, o primeiro passo a ser dado é na direção da maior qualidade dos dados. Tendo a internet como fonte de dados, esse problema torna-se mais evidente por conta da alta quantidade de informações não-estruturadas, diferentes interfaces e falta de indexação. Muitas vezes, é difícil encontrar o equilíbrio entre uma coleta de dados volumosa e, ao mesmo tempo, frutífera. Com esse contexto em mente, o presente trabalho buscou listar, através de pesquisa bibliográfica, quais fatores deveriam ser levados em consideração na prática de *web scraping* para a construção de um corpus de notícias em português. Ao mesmo tempo, deu-se atenção, também, ao fato de que esse corpus poderá ser utilizado como entrada para um modelo de NLP de português e, portanto, quais seriam os fatores que deveriam ser considerados por conta disso.

Palavras-chave: corpus em português; NLP de português; coleta de notícias na internet.

Abstract

The search for high quality data to feed models of Artificial Intelligence is starting to gain more followers, both in academia and in companies. There is a pertinent campaign advocating that, in order for models to produce more accurate results, the first step to be taken is towards greater data quality. With the internet as a data source, this problem becomes more evident due to the high amount of unstructured information, different interfaces and lack of indexing. It is often difficult to strike the balance between voluminous and, at the same time, fruitful data collection. With this context in mind, the present work sought to list, through bibliographical research, which factors should be taken into account in the practice of web scraping for the construction of a news corpus in Portuguese. At the same time, attention was also paid to the fact that this corpus could be used as an input to a NLP of Portuguese model and, therefore, what factors should be considered adding this fact.

Keywords: corpus in Portuguese; NLP of Portuguese; web scraping.

LISTA DE FIGURAS

SUMÁRIO

1. INTRODUÇÃO

1.1 CONSIDERAÇÕES INICIAIS

O contexto acadêmico e empresarial no qual estão inseridas grandes pesquisas e desenvolvimento de modelos de inteligência artificial, passa atualmente por uma mudança de paradigma. Essa mudança, liderada pelo proeminente cientista da computação Andrew Ng, especializado em Inteligência Artificial (IA), muda o foco das pesquisas na área, deixando de serem centradas nos modelos (em inglês, usa-se o termo *model-centric AI*) para serem centradas nos dados que alimentam esses algoritmos (*data-centric AI*). Ou seja, agora que os modelos funcionam bem até certo ponto, os pesquisadores têm que se concentrar em fazer os dados funcionarem também. Em uma palestra[1] esse ano, Ng afiirmou que os modelos atuais são normalmente treinados por 10.000 ou menos exemplos, em vez de milhões de exemplos, como utilizavam os modelos antigos. Esse é um bom motivo para prestar mais atenção à qualidade dos dados.

Nesse cenário, surge a preocupação por uma coleta de dados focada na qualidade da informação e livre de ruídos, principalmente quando se trata de uma coleta de dados na internet. Aqui, será investigada e apresentada, através de uma pesquisa qualitativa bibliográfica, possíveis respostas para algumas dessas questões de qualidade, no caso voltadas para a construção de um corpus[2] de notícias em português. Por exemplo, segundo os autores pesquisados, quais seriam os fatores que deveriam ser levados em consideração na coleta dessas notícias? Mais a fundo, tendo em vista agora que o corpus alimentará um modelo de IA focado em processamento de linguagem natural[3] (em inglês, *Natural Processing Language* - NLP), quais características os textos devem conter para colaborarem com um resultado satisfatório?

---

[1] Disponível em: https://www.youtube.com/watch?v=06-AZXmwHjo. Acesso em 09/10/2021.

[2] Segundo o dicionário Priberam de Língua Portuguesa, corpus significa "coletânea acerca de um mesmo assunto; conjunto de documentos que servem de base para a descrição ou o estudo de um fenômeno".

[3] Segundo Bird *et al.* (2019), "linguagem natural" significa uma linguagem que é usada para a comunicação diária pelos humanos, ou seja, os idiomas como inglês, hindi ou português. Em contraste com as linguagens artificiais,

1.2 OBJETIVO

Pretende-se com o desenvolvimento do presente trabalho científico mostrar quais fatores devem ser considerados na construção de um corpus de notícias em português, tendo em mente não só a confiabilidade das fontes consultadas e a qualidade do conteúdo apresentado como também sua utilização futura em modelos de NLP em português. Os resultados esperados devem identificar esses fatores, segundo pesquisa bibliográfica realizada em técnicas de *web scraping*[4] e NLP em português.

1.3 JUSTIFICATIVA E RELEVÂNCIA

Por mais que seja maçante repetir o que já foi dito centenas de vezes, e talvez até possa parecer uma armadilha para o lugar comum, não podemos deixar de afirmar que o mundo hoje é movido por dados. Segundo estimativas do Fórum Econômico Mundial[5], em 2020 foram gerados 44 ZB (zetabytes) de informação digital. Parece impressionante, mas até descobrirmos que em 2025, a mesma organização estima que o mundo produzirá 463 EB (exabytes) de dados diariamente! A título de recordação, 1 exabyte equivale a 10^6 TB e 1 zetabyte equivale a 10^9 TB.

Uma das causas para esse enorme volume gerado é a necessidade que as empresas têm por dados que alimentem seus modelos de IA para, por exemplo, melhorarem seus produtos, detectar novas demandas e saírem na frente quando se trata de inovação. Em uma pesquisa

como as linguagens de programação e notações matemáticas, as linguagens naturais evoluem à medida que passam de geração em geração e são difíceis de definir com regras explícitas.

[4] Segundo Mitchell (2015), *Web scraping* é o processo de extração de dados de sites específicos na internet para posterior recuperação e análise. Embora ele possa ser feito manualmente por um usuário, o termo normalmente refere-se a processos automatizados executados por um robô ou *web crawler*.

[5] As estimativas podem ser acessada em: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/. Acesso em 09/10/2021.

global sobre a adoção empresarial de IA em 2020[6], a consultoria McKinsey descobriu que, com a adoção de ML (do inglês, *Machine Learning*, uma das áreas de estudo de IA), 66% das empresas aumentaram suas receitas e 40% diminuíram os custos. E qualquer modelo de IA ou ML é totalmente dependente de dados subjacentes. Ou melhor, de bons dados subjacentes.

Por isso, e para se manterem competitivas, as empresas devem priorizar a qualidade dos dados que coletam, e não apenas bons modelos de IA. Isso significa mudar o foco de um desenvolvimento apenas centrado em IA para um desenvolvimento centrado também na qualidade dos dados que alimentam esses modelos. Andrew Ng, em uma palestra já citada anteriormente, compilou estudos[7] que mostram como uma melhoria nos dados melhora significativamente a precisão do modelo, enquanto o aprimoramento do algoritmo do modelo teve pouco ou nenhum efeito, evidenciando, assim, os benefícios dos dados de alta qualidade. Com isso, processos de coleta de dados, como o *web scraping*, ganham suma importância e atenção.

A montagem de um corpus para processamento de linguagem natural deve refletir não só as regras do idioma a ser analisado como também apresentar um conteúdo claro, coeso e coerente. O corpus do presente trabalho focará em notícias em português sobre o bioma costeiro e marinho brasileiro, cujos temas são biodiversidade, exploração econômica e conservação ambiental. Essa escolha foi feita tendo em vista a Década do Oceano – uma iniciativa da Organização das Nações Unidas (ONU) para conscientizar a população mundial sobre a importância dos oceanos e mobilizar atores públicos, privados e da sociedade civil organizada em ações que favoreçam a saúde e a sustentabilidade dos mares. A iniciativa terá uma série de eventos, no mundo todo, entre 2021 e 2030, para incentivar a reflexão sobre as ações urgentes e necessárias para o uso e proteção do espaço costeiro e marinho.

---

[6] A pesquisa está disponível em: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020. Acesso em 09/10/2021.

[7] A palestra em que Ng menciona seu estudo está disponível em: https://www.youtube.com/watch?v=06-AZXmwHjo. Acesso em 09/10/2021.

1.4 METODOLOGIA E ESTRUTURA DO TRABALHO

A fundamentação teórica aqui exposta foi baseada em pesquisa exploratória aplicada e revisão bibliográfica de artigos científicos, livros e periódicos publicados na internet, iniciada em março de 2021 e concluída em dezembro de 2021. O paradigma dessa pesquisa será qualitativo. Todas as fontes consultadas e/ou citadas encontram-se na seção de Referências Bibliográficas.

A forma de desenvolvimento do trabalho está dividida conforme a seguinte estrutura:

Capítulo 1 – Introdução; onde apresentam-se as justificativas e objetivos deste trabalho;

Capítulo 2 – Desenvolvimento do Tema; onde serão descritos, de forma sumarizada, como funciona o *web scraping*, seus benefícios, usos, técnicas mais comuns e introdução ao NLP e NLP em português;

Capítulo 3 – Os Fatores Observados na Prática de *Web Scraping* de Notícias em Português; onde são discutidos a relevância e utilidade dos fatores observados nessa prática e onde obter as notícias relevantes ao objetivo;

Capítulo 4 – Conclusão; onde serão respondidos os objetivos propostos pelo trabalho, relacionando-os com os pontos fundamentais e mais importantes da pesquisa bibliográfica;

Capítulo 5 – Possíveis Desdobramentos, onde apresentam-se conceitos e interpretações de tópicos atuais da tecnologia para possíveis aplicações do corpus em português;

E por fim, as Referências Bibliográficas.

2. DESENVOLVIMENTO DO TEMA

Neste capítulo, segue-se uma discussão teórica inicial sobre a construção de um corpus de notícias usando-se *web scraping*, para em seguida ser apresentada a discussão teórica sobre os fatores que devem ser considerados na construção de um corpus de notícias em português que, futuramente, alimentará um modelo NLP de português.

2.1 COMO FUNCIONA O *WEB SCRAPING*

Segundo Mitchell (2015), *Web scraping* é uma técnica empregada para extrair grandes quantidades de dados de sites e repositórios da internet e são salvos localmente para uso instantâneo ou análise que será realizada posteriormente. Essa técnica também é conhecida com outros nomes, como *screen scraping, web data extraction* e *web harvesting*. Os dados são salvos em um sistema de arquivos local ou tabelas de banco de dados, de acordo com a estrutura dos dados extraídos.

A maioria dos sites na internet permite-nos apenas ver o conteúdo apresentado na tela do computador, e geralmente não permite a instalação de uma cópia ou download do conteúdo. A solução mais óbvia seria fazer um script automatizado que possa extrair dados das páginas escolhidas e salvá-los em um formato estruturado. Mas também existem softwares de *web scraping* que carregam automaticamente várias páginas da web, uma por uma, e extraem os dados, de acordo com os requisitos. Isso pode ser feito sob medida, para um site específico, ou pode ser configurado, com base em um conjunto de parâmetros, para funcionar com qualquer site. É muito comum, atualmente, que *bots* inteligentes façam *web scraping* usando APIs ou interfaces web. Ao contrário do *screen scraping*, a ser discutido adiante, onde apenas se copia os pixels exibidos na tela, esses *bots* extraem o código HTML subjacente, bem como os dados armazenados em um banco de dados em segundo plano.

Para explicar como funciona um *web scraper* de forma simplificada, é necessário antes falar sobre *webcrawlers*, ou *spider*. Segundo Williams (2021), um *webcrawler* é um programa de computador que pesquisa automaticamente documentos na web com o objetivo principal de construir um índice. Há também os que pesquisam diferentes tipos de informações, como *feeds* RSS e endereços de e-mail. O *webcrawler* mais conhecido é o Googlebot.

Listaremos a seguir, o funcionamento básico e a configuração de um *webcrawler*, segundo Willians (2021):

a) A semente

É um procedimento semelhante a uma travessia de árvore, em que o rastreador primeiro passa pela “URL semente”, também chamada de URL base e, em seguida, procura a próxima URL nos próprios dados da URL semente e assim por diante. A URL semente

seria informada no início pelo usuário. Por exemplo, para extrair todos os dados das diferentes páginas de um site, a URL semente serviria como uma base incondicional;

b) Definir direções

Uma vez que os dados da URL semente tenham sido extraídos e armazenados na memória temporária, as URLs presentes ali serão fornecidas ao ponteiro e então o sistema deve se concentrar em extrair os dados dessas URLs;

c) Fila

O *crawler* precisa extrair e armazenar todas as páginas que analisa enquanto percorre um único repositório, de arquivos HTML. A etapa final de extração e limpeza de dados acontece, de fato, neste repositório local;

d) Extração de dados

Agora, todos os dados necessários estão no repositório. Mas os dados não podem ser usados. Portanto, é importante ensinar o *crawler* a identificar nós, *tags* de dados e extrair apenas os dados necessários;

e) Desduplicação e limpeza

Somente os dados sem ruídos devem ser extraídos e as entradas duplicadas devem ser excluídas automaticamente. Esses tipos de atividades devem ser incorporadas à inteligência do *scraper* para torná-lo mais prático e os dados provenientes como saída, mais utilizáveis;

f) Estruturação

Seria possível criar um pipeline de utilização direta desses dados somente se o *scraper* for capaz de estruturar os dados não estruturados, para alimentar, por exemplo, um site diretamente com o resultado do mecanismo de extração.

2.2 BENEFÍCIOS E USOS

O primeiro benefício a ser mencionado, segundo Williams (2021), é o mais popular quando se faz uma busca online sobre serviços de *web scraping*, que é obter detalhes e preços de produtos

em sites de vendas online. Muitas empresas rastreiam sites de comércio eletrônico em busca de preços, descrições de produtos e imagens, para obter a partir deles possíveis meios de impulsionar a modelagem analítica e preditiva do serviço oferecido. A comparação de preços nos últimos anos tornou-se muito importante para lojas online saberem mais sobre seus concorrentes. Um exemplo, são os sites de viagens que extraem preços de sites de companhias aéreas e grandes redes hoteleiras para manterem seu negócio. Desta forma, eles coletam esses dados para criarem seu próprio *data warehouse*, para uso presente e futuro.

Outro benefício bem popular, segundo Williams (2021), é o investigativo, seja a título profissional ou mercadológico. No primeiro caso, um exemplo seria o rastreamento de antecedentes que empresas fazem em seu quadro de empregados. Os dados coletados são posteriormente usados para checar a veracidade de um currículo ou verificar antecedentes criminais. No segundo âmbito, há empresas especializadas em fazer análise e curadoria personalizadas a novos sites ou canais de atendimento, onde os dados coletados podem ajudar a entender a demanda e o comportamento do público-alvo. Isso impulsiona novas empresas, logo no início de suas atividades, e a venda de produtos baseados em descobertas de padrões que ganharão mais visitas orgânicas. Dessa forma, gasta-se menos com anúncios ao descobrir, através de dados, qual anúncio seria mais adequado para determinados usuários do site. Isso catalisa a economia de receita de marketing ao mesmo tempo que atrai mais consumidores.

Ainda no âmbito de pesquisa mercadológica, a reputação online é outro fator muito importante e levada muito a sério por empresas expostas às mídias sociais e que dependem, quase que inteiramente, da publicidade “boca a boca” e da recomendação entre conhecidos, para ajudá-las a crescer. Nesse ponto, é crucial retirar informações das mídias sociais para compreender a opinião pública e os sentimentos expostos nos comentários. Formadores e influenciadores de opinião, tópicos de tendência e fatores demográficos podem ser proeminentes por meio da coleta de dados e, então, podem ser usados para garantir que a empresa consiga reparar sua imagem ou ter uma maior pontuação de satisfação do público online (WILLIAMS, 2021).

Por último, e também no âmbito comercial, um dos benefícios do *web scraping* é detectar e analisar avaliações online fraudulentas. Essas avaliações são extremanente importantes ao comércio online, uma vez que pode estimular ou não a decisão sobre a compra de um produto. Já falamos aqui o quanto é valioso a propaganda “boca a boca” de um produto e, certamente,

as avaliações online é uma de suas principais manifestações. Por serem tão importantes, criou-se uma modalidade de fraude, o *spamming* de opinião, totalmente ilegal, que consiste, por exemplo, em escrever comentários falsos nos portais de compra, prejudicando assim o anunciante em questão. Também é chamado de *shilling* - uma atividade que visa unicamente enganar os usuários do site. O *scraping* das avaliações do site pode ajudar a rastrear as avaliações fraudulentas, detectá-las, bloqueá-las ou verificá-las, porque essas avaliações geralmente são as que se destacam na multidão.

2.3 TÉCNICAS DE *WEB SCRAPING*

A coleta de dados na internet pode se tornar muito torturosa, principalmente quando se trata de coletar dados não estruturados e, no universo da internet, os dados não estruturados são os mais abundantes. Em contrapartida aos dados estruturados, os dados não estruturados não permitem serem gerenciados ativamente em um sistema transacional; por exemplo, são dados que não residem em um sistema de gerenciamento de banco de dados relacional (RDBMS) e, portanto, não podem ser recuperados com uma simples consulta em SQL. Os dados não estruturados têm uma estrutura interna, mas não são predefinidos por meio de modelos de dados. O motivo deles serem tão prolíficos é simples: eles podem ser qualquer coisa, como imagens, áudios, dados de sensores, dados de texto gerados por discurso, entre outros, podendo ser, portanto, gerados tanto por humanos como por máquinas.

Vale ressaltar, antes de continuarmos a discussão, que coletar dados na internet é considerado legal se for feito com permissão e não violar a privacidade das entidades envolvidas, como explicitado no caso HiQ Labs vs. LinkedIn[8]. Entretanto, para evitar quaisquer desentendimentos, sempre consulte o arquivo robots.txt, disponibilizado por todo site na internet, para verificar se ele permite ou não ser rastreado por mecanismos automáticos de coleta de dados.

[8] Mais informações podem ser encontradas em https://en.wikipedia.org/wiki/HiQ_Labs_v._LinkedIn. Acesso em: 16/10/2021.

Continuando, as técnicas de coleta de dados não estruturados na internet incluem desde o simples copiar e colar até os mais avançados softwares pagos, como ParseHub (parsehub.com) e Octoparse (octoparse.com), geralmente adotados por empresas para obter informações sobre seus competitidores online. Falaremos mais em detalhes sobre cada uma delas nas seções a seguir.

2.3.1 COPIAR E COLAR MANUALMENTE

No *web scraping* manual, o que se faz é simplesmente copiar e colar o conteúdo do site desejado e transformá-lo em dado estruturado. Isso, no entanto, é demorado, repetitivo e sujeito a erros, mas muito eficaz porque as defesas que um site pode oferecer ao *web scraping* automático, não são oferecidas à coleta manual. Entretanto, esse tipo de atividade tomaria muito tempo e dinheiro e correria o risco de não chegar ao objetivo final, dado que o volume de informação a ser coletado pode ser enorme. Daremos mais atenção, então, às técnicas automatizadas discutidas adiante.

2.3.2 *SCREEN SCRAPING*

Se no *web scraping* manual só é necessário selecionar o texto e colar, nessa técnica só é preciso clicar na área do site onde encontra-se a informação desejada e esperar o software especializado estruturar os dados em uma planilha ou arquivo CSV. Não é necessário nenhum código de programação.

Os softwares disponíveis atualmente, como VisualPing (visualping.io) e Octoparse (octoparse.com), só pra citar alguns, oferecem uma interface amigável ao usuário, semelhantes a navegar em um site, além de outras facilidades como serviços em nuvem, login automático e manejo de *scrolling* infinito.

Porém, essa técnica é extremamente dependende do design do site em questão, como por exemplo, a disposição das informações importantes, display de imagens e estética envolvida. Se, por algum motivo, o administrador do site decidir mudar seu layout ou design, todo o trabalho de seleção das áreas interessantes deverá ser refeito. Outro fator que deve ser

considerado é se o custo da licença do software valerá a pena quando comparado ao retorno esperado.

2.3.3 ANÁLISE DE HTML E REQUISIÇÕES HTTP

Todo site na internet é um documento estruturado escrito em HTML. Porém, nem sempre é possível extrair dados das páginas e, ao mesmo tempo, manter sua estrutura, seja por conta da maneira como a informação é apresentada ou porque nem todo site fornece seus dados em formatos estruturados, como CSV ou JSON.

A análise de HTML é feita em conjunto com requisições HTTP, usando-se JavaScript ou bibliotecas em Python, como lxml, Requests e BeautifulSoup. Tem como alvo páginas HTML lineares ou aninhadas. É um método rápido e robusto usado para extração de texto, captura de tela, como o *point-and-click*, extração de recursos, entre outros (CHAPMAN, 2020).

Como o objetivo aqui é apenas citar algumas das técnicas mais usadas, não entraremos em detalhes sobre cada biblioteca e seus diversos usos, o que pode ser facilmente encontrado em suas devidas documentações.

2.3.4 ANÁLISE DE DOM

DOM é a abreviação em inglês de *Document Object Model* e define a estrutura de estilo e o conteúdo dos arquivos XML que determinado site emprega. A partir de sua análise, obtém-se uma visão detalhada de como é feita a estruturação de dados desse site. Assim, é possível extrair os nós que contêm as informações desejadas e, em seguida, por exemplo, usar a linguagem de consulta Xpath, ou XML Path, para copiá-las e extraí-las (CHAPMAN, 2020). Essa linguagem de consulta é usada para navegar em documentos XML por causa de sua estrutura em árvore, selecionando os nós com base em parâmetros diferentes. Ainda, navegadores como o Firefox podem ser incorporados na análise com o objetivo de extrair a página inteira ou apenas partes dela, mesmo se o conteúdo gerado for de natureza dinâmica.

2.3.5 AGREGAÇÃO VERTICAL

Segundo Arguello *et al*. (2015), plataformas verticais são mecanismos de busca para verticais, ou seja assuntos específicos, demandando alto poder computacional, muitas vezes em nuvem. Essas plataformas criam e monitoram uma infinidade de robôs voltados para verticais específicas, sem envolvimento humano direto durante o processo. A preparação envolve estabelecer, primeiramente, uma base de conhecimento[9] para toda a vertical e, em seguida, a plataforma cria os robôs automaticamente para buscar as informações desejadas, não em um site especificamente mas em toda a base de conhecimento recém criada.

A robustez da plataforma é medida pela qualidade das informações que ela recupera da base de conhecimento e sua escalabilidade, ou seja, a rapidez com que pode escalar para centenas ou milhares de sites. Essa escalabilidade é usada principalmente para atingir sites que os agregadores comuns consideram complicados ou muito trabalhosos para coletar conteúdo (SIRISURIYA, 2015).

2.3.6 RECONHECIMENTO DE NOTAÇÃO SEMÂNTICA

As páginas que estão sendo consideradas para extração de dados podem incluir metadados ou marcações semânticas e anotações, que podem ser usadas para localizar fragmentos de dados específicos. Se as anotações forem incorporadas às páginas, como feito pelo *Microformat*, essa técnica pode ser vista como um caso especial de análise DOM. Em outro caso, as anotações, que são organizadas em uma camada semântica, são armazenadas e gerenciadas separadamente das páginas da web para que os scripts possam recuperar o esquema de dados e as instruções dessa camada antes de extrair os dados das páginas. Também pode ser adicionada a essa técnica a correspondência de padrão de texto (*text pattern matching*). Essa técnica envolve o uso do

[9] Segundo Ceta (2018), base de conhecimento é um repositório ou biblioteca de autoatendimento usado para armazenar informações facilmente recuperáveis sobre um produto, serviço ou assunto. Serão discutidas mais pra frente, no capítulo 5 - Possíveis Desdobramentos.

comando *grep* do UNIX e é usada com linguagens de programação populares como Perl ou Python.

2.3.7 FERRAMENTAS E SERVIÇOS ONLINE

As técnicas de *web scraping* também envolvem o uso de ferramentas e serviços que podem ser facilmente acessados online. Essas ferramentas e serviços automatizados incluem *limeproxies*, cURL, Wget, HTTrack, Import.io, Node.js e uma lista de outros. Para fins de extração de dados, geralmente são usados navegadores *headless* como Phantom.js, Slimmer.js e Casper.js.

Citaremos brevemente uma das ferramentas mais empregadas, segundo Chapman (2020), chamada Selenium. Ela é um projeto *open source* que abrange uma variedade de ferramentas e bibliotecas que permitem e dão suporte à automação de navegadores da web. Segundo sua documentação (2021), ela fornece extensões para emular a interação do usuário com os navegadores, um servidor de distribuição para dimensionar a alocação do navegador e a infraestrutura para implementações da especificação W3C WebDriver, o que permite escrever código intercambiável para todos os principais navegadores da web. O Selenium ajuda muito a descobrir como os sites funcionam, seja simulando uma visita humana normal a uma página usando um navegador da web comum ou emulando chamadas *ajax* em *web scraping*. Por ser uma ferramenta de automação poderosa, também possibilita testar sites e automatizar qualquer ação demorada na web.

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL (NLP)

Segundo Borcan (2020), Processamento de Linguagem Natural é um subcampo de pesquisa da grande área de Inteligência Artificial que se concentra no desenvolvimento de modelos e pontos de interação entre humanos e computadores baseados na linguagem natural. Isso inclui texto e sistemas baseados na fala.

De acordo com o relatório de pesquisa de 2021 "*AI in Healthcare*"[10], o NLP se tornou um dos componentes de tecnologia mais importantes para as tendências em IA e saúde. Outras verticais de negócios, como serviços financeiros, têm um longo histórico de alavancagem de mineração de texto e impulsionaram a adoção da NLP em todos os setores de negócios. Hoje, essa área desempenha um papel vital em muitas tarefas que envolvem documentos e comunicações de texto. Outras aplicações incluem análise de patentes, artigos científicos para pesquisa farmacêutica, otimização da cadeia de suprimentos global, recomendações de notícias entre outras, como veremos adiante.

2.4.1 COMO FUNCIONA O NLP

Uma verdade geralmente aceita na ciência da computação é que todo problema complexo se torna mais fácil se o dividirmos em pedaços menores. Isso é especialmente verdadeiro no campo da Inteligência Artificial. Para um determinado ambiente, constroem-se vários componentes pequenos e altamente especializados, que são bons para resolver um e apenas uma parte do ambiente. Em seguida, alinham-se todos esses componentes, processa-se a entrada por cada um deles e obtém-se uma saída no final do processamento. Isso é chamado de *pipeline*.

No contexto do NLP, um problema básico seria que, para um determinado parágrafo, o computador entende exatamente o significado dele e então, possivelmente, age de acordo ao esperado. Para que isso funcione, precisamos seguir algumas etapas. São elas, resumidamente, segundo Borcan (2020):

a) Segmentação do limite de frase (*Sentence boundary segmentation*)

Para um determinado texto, é necessário identificar corretamente todos os parágrafos para que cada frase resultante dele tenha seu significado extraído nas próximas etapas. Com isso, espera-se que seja mais preciso, e fácil, extrair o significado de cada frase de um texto e, em seguida, juntá-lo do que tentar identificar o significado de todo o texto de uma só vez. Afinal, quando falamos (ou escrevemos) não queremos dizer,

[10] O relatório completo está disponível em https://www.nlpsummit.org/industry-survey-analysis-ai-in-healthcare-2021/. Acesso em: 02/11/2021.

necessariamente, apenas uma coisa. Frequentemente, tendemos a transmitir mais ideias em uma só e a beleza da linguagem natural (e a maldição do NLP) é que geralmente textos e discurssos fazem isso. Uma abordagem inicial, e ingênua, para tal seria pesquisar apenas alguns pontos finais em um trecho de texto e defini-los como o final de uma frase. O problema é que os pontos também podem ser usados para outros fins (por exemplo, abreviações), entretanto, na prática, os modelos de aprendizado de máquina foram definidos para identificar corretamente os sinais de pontuação usados para encerrar frases;

b) Tokenização de palavras (*Word tokenization*)

Esta parte envolve utilizar uma frase da etapa anterior e dividi-la em uma lista contendo todas as palavras (e sinais de pontuação) que ela contém. Isso será usado nas próximas etapas para realizar uma análise de cada palavra;

c) Parte de marcação de fala (*Part of speech tagging - PoS*)

Essa etapa envolve classificar cada palavra da etapa anterior de acordo com a classe gramatical que ela representa. Esse é um passo essencial para identificar o significado por trás de um texto. Identificar os substantivos nos permite descobrir de quem ou do que se trata o texto fornecido. Então, os verbos e adjetivos nos permitem entender o que as entidades fazem, ou como são descritas, ou qualquer outro significado que possamos obter de um texto. Segundo Borcan (2020), a marcação de PoS é um problema difícil, mas já foi resolvido e implementações para isso podem ser encontradas na maioria das bibliotecas e ferramentas de aprendizado de máquina modernas;

d) Reconhecimento de entidade nomeada (*Named Entity Recognition – NER*)

Essa tarefa inclui identificar os nomes em uma frase e classificá-la corretamente em uma lista de categorias predefinidas. Essas categorias podem envolver Pessoas, Organizações, Locais, Tempo, Quantidades e assim por diante. Listas de categorias podem ser personalizadas para um caso de uso específico, mas, em geral, quase todos precisam de pelo menos essas categorias para serem corretamente identificadas. Ainda segundo Borcan (2020), existem muitas implementações para este tipo de problema e os modelos construídos recentemente alcançaram um desempenho quase humano, como o

GPT-3[11]. Basicamente, esta etapa também é dividida em duas subtarefas: identificar corretamente os nomes em uma frase e, em seguida, classificar cada nome de acordo com sua lista de categorias.

É claro que há muitas outras tarefas que o NLP hoje em dia é capaz de resolver para serem usadas em aplicativos do mundo real, mas a maioria delas são adaptadas a cada caso de uso e a cada necessidade de negócio. Tendo tudo isso dito, as etapas apresentadas anteriormente são as etapas básicas em quase todos os casos de uso que podemos pensar.

2.4.2 ALGUMAS APLICAÇÕES DE NLP

Atualmente, a Inteligência Artificial é largamente empregada em uma variedade de aplicativos usados no dia-a-dia por usuários que não fazem ideia de que a estão usando de fato. Vários casos de uso foram identificados como solucionáveis por meio da implantação de modelos de NLP. Alguns desses exemplos são:

a) Assistentes virtuais

O Processamento de Linguagem Natural permite que assistentes virtuais, ou *chatbots*, entendam a linguagem como nós, humanos, a falamos. Isso significa que o assistente virtual não apenas entende as palavras, mas pode entender, também, a intenção da pergunta de um consumidor e o contexto da conversa. Isso permite que a interação flua como uma conversa em vez de uma sessão de perguntas e respostas robótica. Segundo Krayewski (2021), isso é muito mais difícil do que se parece e foi tornado possível graças aos recentes avanços nos chips mobile dedicados à IA, embutidos na grande maioria dos celulares comercializados recentemente.

Um *chatbot* funciona com as seguintes chaves: enunciados (maneiras como o usuário se refere a uma intenção específica), intenção (o significado por trás das palavras que um usuário fala ou digita), entidade (detalhes que são importantes para a intenção, como datas e locais), contexto (que ajuda a salvar e compartilhar parâmetros em uma sessão)

[11] Mais informações em https://openai.com/blog/gpt-3-apps/. Acesso em: 02/11/2021.

e sessão (uma conversa do início ao fim, mesmo se interrompida) (KRAYEWSKI, 2021);

b) Tradução automática

A tradução automática é executada por softwares que traduzem texto em um idioma para outro sem a contribuição humana. Em seu nível fundamental, ela realiza uma substituição direta de palavras atômicas em um determinado idioma por palavras em outro. Segundo Kumari (2020), usando-se métodos de corpus, traduções mais complicadas podem ser conduzidas levando em consideração um melhor tratamento de contrastes em tipologia fonética, reconhecimento expresso e traduções de expressões idiomáticas, assim como a reclusão de particularidades no idioma em questão. Algumas técnicas de NLP empregadas em tradução automática são *Statistical Machine Translation* (SMT), *Rule-based Machine Translation* (RBMT) e *Neural Machine Translation* (NMT).

c) Reconhecimento de fala

São raros os tablets, notebooks ou smartphones top de linha que não possuem função de reconhecimento de fala. Entretanto, apesar de tamanha popularidade, o que poucos sabem é que os programas empregados nessa função usam diversas técnicas de NLP para entender o que é dito. Essa função analisa os dados na forma de palavras, escritas ou faladas, ao usar o áudio como uma fonte primária de dados e, em seguida, ajuda a treinar, também, outros modelos por meio de aprendizado profundo, como texto-para-voz e voz-para-texto (IFTEKHAR, 2021).

2.4.3 NLP EM PORTUGUÊS

De acordo com a pesquisa *Ethnologue*[12] de 2021, da SIL International[13], atualmente o inglês é a língua mais falada no mundo com, aproximadamente, 1.3 bilhões de falantes. Isso, no entanto, não pode ser um limitante de conhecimento, pois sabemos que o mundo é multilíngue; por exemplo, a mesma pesquisa contabilizou 34 línguas com mais de 45 milhões de falantes cada. O português encontra-se em nono lugar no ranking de línguas mais faladas, sendo 232.4 milhões de falantes como primeira língua e 25 milhões como segunda língua, totalizando 258 milhões pessoas.

Nesse contexto, trabalhar cientificamente em uma tarefa de NLP em qualquer outro idioma diferente do inglês é extensivamente trabalhoso, talvez até impossível ainda. Nos últimos anos, progrediu-se muito em abordagens baseadas em aprendizagem profunda para tarefas de processamento de linguagem natural, e há muito para se entusiasmar. No entanto, esses avanços podem demorar para passar do inglês para outros idiomas. No passado, a maioria dos acadêmicos mostrou pouco interesse em publicar pesquisas ou construir conjuntos de dados que vão além da língua inglesa, embora as aplicações da indústria precisem desesperadamente de técnicas independentes de idioma.

Segundo Ruder e Eisenschlos (2019), as abordagens existentes para NLP multilíngue dependem de:

a) Dados paralelos entre idiomas, ou seja, um corpus de documentos com exatamente o mesmo conteúdo, mas escritos em idiomas diferentes. Isso é muito difícil de adquirir em um ambiente geral;
b) Um vocabulário compartilhado, ou seja, um vocabulário comum a vários idiomas. No entanto, essa abordagem sobrerepresenta as linguagens com muitos dados e subrepresentaria as com poucos dados. Um exemplo de ferramenta que usa essa

[12] A pesquisa completa está disponível em https://www.ethnologue.com/ethnoblog/gary-simons/welcome-24th-edition. Acesso em: 06/11/2021.

[13] Mais informações em https://www.sil.org/. Acesso em: 06/11/2021.

abordagem é o BERT multilíngue[14], que consome muitos recursos para ser treinada mas, mesmo assim, pode ter dificuldades quando os idiomas são diferentes.

O principal apelo dos modelos multilíngues, como o BERT multilíngue, são suas capacidades de transferência *zero-shot*: com apenas rótulos em um idioma de muitos recursos, como o inglês, ele pode transferí-los para outro idioma sem nenhum dado de treinamento no idioma destino. Além disso, quando o idioma destino é muito diferente do idioma de origem (geralmente inglês), a transferência *zero-shot* pode ter um desempenho ruim ou falhar completamente.

Segundo Vecchietti (2017), modelos baseados em dados são capazes de alcançar resultados de ponta em algoritmos de Processamento de Linguagem Natural apresentados em um sistema de conversão de texto em fala do português brasileiro. Além disso, os modelos baseados em dados propostos no estudo de 2017 superaram os modelos baseados em regras para a conversão G2P e problemas de silabificação.

Avanços recentes na representação de linguagem usando redes neurais tornaram viável a transferência dos estados internos aprendidos de grandes modelos de linguagem pré-treinados (LMs) para tarefas *downstream* de NLP. Essa abordagem de transferência de aprendizagem melhora o desempenho geral em muitas tarefas e é altamente benéfica quando os dados rotulados são escassos, tornando os LMs pré-treinados recursos valiosos, especialmente para idiomas com poucos exemplos de treinamento anotados.

Em 2017, Vaswani *et al.* apresentou o modelo *Transformer* em seu estudo seminal *Attention Is All You Need*[15]. Os *transformers* eram originalmente uma resposta a algumas das limitações dos modelos de sequência a sequência (*seq2seq*) então dominantes - RNNs e LSTMs. Esses modelos sofrem pelo fato de serem verdadeiramente sequenciais. Eles levavam muito tempo para treinar, principalmente por processarem os tokens em uma sequência, um de cada vez. Isso significa que quanto mais longa a sequência, maior o tempo de treinamento (HENRIQUEZ, 2020). Outra limitação era a "memória". À medida que cada token é processado, um estado

[14] A ferramenta é *open source* e está disponível em https://github.com/google-research/bert/blob/master/multilingual.md. Acesso em: 06/11/2021.

[15] O estudo está disponível em https://arxiv.org/abs/1706.03762. Acesso em: 06/11/2021.

oculto anterior é usado em combinação com o token atual para calcular um vetor de contexto. O vetor captura os relacionamentos entre o token atual e os contextos anteriores dos tokens anteriores. A expectativa é que, quando chegar ao último token, o vetor de contexto terá preservado todos esses relacionamentos. Entretanto, quanto mais longa a sequência, maior a possibilidade de que algumas das relações entre tokens distantes podem ser diminuídas ou perdidas.

Muitos são os diferenciais dos *transformers*, segundo Henriquez (2021), para resolver esses entraves. O primeiro é que eles recebem sua entrada em paralelo. Eles ingerem toda a sequência de uma vez, obtendo assim a capacidade de processar todos os tokens na sequência ao mesmo tempo, aproveitando-se do poder de processamento paralelo das GPUs. Isso acontece tanto no codificador quanto no decodificador, o que significa que o decodificador não precisa esperar que o codificador envie seu vetor de contexto.

Os *transformers* também são capazes de resolver o problema de memória com “auto-atenção”. Como toda a sequência é processada em paralelo, cada token tem acesso a todos os outros tokens ao calcular a "atenção". Isso significa que cada token pode levar consigo sua relação com todos os outros tokens. Em uma linguagem tão complexa gramaticalmente como a portuguesa, certamente isso é um avanço e por esses e outros fatores que os modelos de *transformers* estão ganhando cada vez mais espaço no campo de aprendizado profundo e modelos recorrentes (HENRIQUEZ, 2021).

O modelo que se beneficia dos *transformers* mais proeminente em NLP multilíngue, atualmente, é o BERT[16], acrônimo em inglês para *Bidirectional Encoder Representations from Transformers*. Publicado por pesquisadores do Google AI Language, o artigo causou um rebuliço na comunidade de aprendizado de máquina ao apresentar resultados de última geração em uma ampla variedade de tarefas de NLP, incluindo resposta a perguntas (SQuAD v1.1), inferência de linguagem natural (MNLI) e outras.

A principal inovação técnica do BERT é a aplicação do treinamento bidirecional do *Transformer*, o modelo popular citado anteriormente, à modelagem de linguagem. Isso está em

[16] O artigo completo que divulga essa ferramenta está disponível em https://arxiv.org/pdf/1810.04805.pdf. Acesso em: 06/11/2021.

contraste com os esforços anteriores que olhavam para uma sequência de texto da esquerda para a direita ou treinamento combinado da esquerda para a direita e da direita para a esquerda. Os resultados do artigo mostram que um modelo de linguagem que é treinado bidirecionalmente pode ter um senso mais profundo de contexto e fluxo de linguagem do que modelos de linguagem de direção única. No artigo, os pesquisadores detalham uma nova técnica chamada *Masked LM* (MLM), que permite o treinamento bidirecional em modelos nos quais anteriormente era impossível.

O primeiro trabalho a empregar modelos BERT para a NER (Named Entity Recognition) em português, e que encontra-se em evidência atualmente, é o BERTimbau[17], treinado pela startup brasileira de inteligência artificial NeuralMind. No trabalho, foram avaliadas diversas arquiteturas neurais usando-se modelos BERT para a NER em português e comparadas estratégias de treinamento baseadas em recursos e em ajuste fino. Os modelos obtidos foram avaliados em três tarefas posteriores ao NLP: similaridade textual de sentenças, reconhecimento de vinculação textual e reconhecimento de entidades nomeadas. Eles aprimoram o estado da arte em todas essas tarefas, superando o BERT multilíngue e confirmando a eficácia de grandes LMs pré-treinados para o português.

2.5. MACHINE LEARNING OPERATIONS (MLOps)

O objetivo deste tópico não é entrar em detalhes no vasto campo de MLOps, mas sim listar algumas das necessidades que um sistema especializado requer e, assim, reiterar que para desenvolver e operar sistemas complexos como os de NLP é necessário um saldo técnico muito grande e multidisciplinar.

Assim como existe um termo para designar a combinação entre desenvolvimento contínuo de software (“Dev”) com operações de TI (“Ops”), ou DevOps como é popularmente conhecido, criou-se também o termo equivalente para as operações de TI envolvidas no processo de desenvolvimento de aprendizado de máquina, ou seja, MLOps.

[17] O artigo e o modelo estão disponíveis em https://github.com/neuralmind-ai/portuguese-bert. Acesso em: 06/11/2021.

Segundo a empresa Google Cloud[18], MLOps é um conjunto de práticas que visa implantar e manter modelos de aprendizado de máquina em produção de maneira confiável e eficiente. As operações envolvem a automação e o monitoramento em todas as etapas da construção de um sistema de ML, incluindo integração, teste, liberação, implantação e gerenciamento de infraestrutura. A adoção dessas práticas tornam-se mandatórias a medida que os modelos de IA ficam mais robustos e capazes de resolver problemas cada vez mais complexos, já que todo esse avanço demanda alto poder computacional e expertise em lidar com tudo o que isso implica (migração à nuvem, gerenciamento de recursos na nuvem, os custos envolvimentos em cada recurso entre outros).

Assim, segundo Sculley *et al.* (2015), o verdadeiro desafio atualmente não é mais construir um modelo de ML e codificá-lo, mas sim construir um sistema integrado de ML e operá-lo continuamente em produção. Conforme mostrado na figura 1 a seguir, apenas uma pequena fração de um sistema de ML no mundo real é composta por código de ML em si. Os elementos envolventes necessários são vastos e complexos.

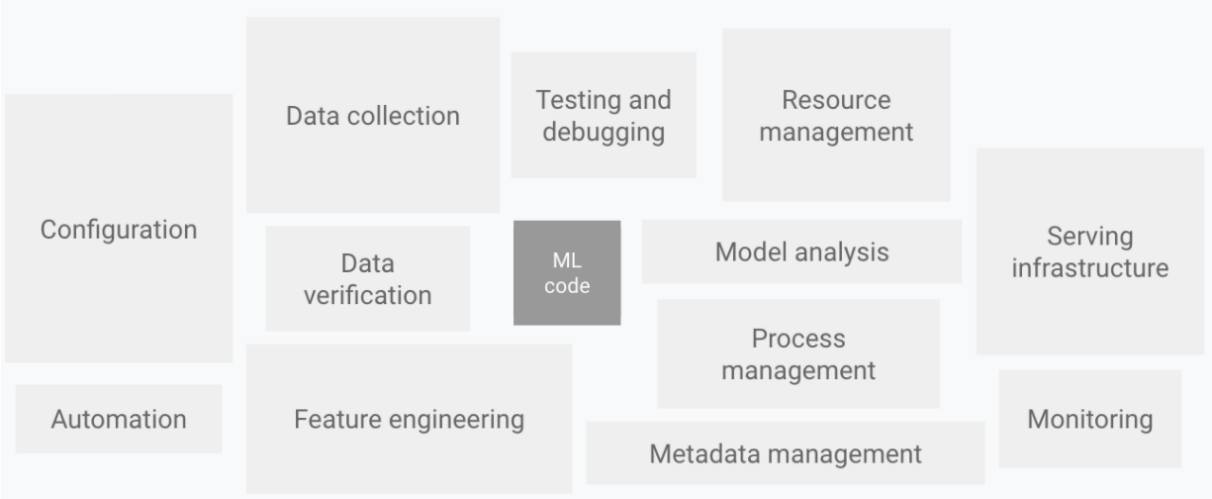

Figura 1: Elementos de um sistema de ML.

Fonte: adaptado de SCULLY, D. *et al*., 2015, p.4.

[18] Mais informações estão disponíveis em https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning. Acesso em: 12/11/2021.

Nesse diagrama, ainda há o restante do sistema, que não está mostrado, que é composto de configuração, automação, coleta de dados, verificação de dados, teste e depuração, gerenciamento de recursos, análise de modelo, gerenciamento de processo e metadados, infraestrutura de serviço e monitoramento.

Outro fato muito pertinente é o custo dessas operações. O fato de que o custo marginal do software se aproxima de zero tem sido a base do modelo de negócios da indústria de software desde os anos 1980, diz Talby (2019). Por exemplo, antigamente, ao codificar um aplicativo de calculadora, o desenvolvedor tinha a certeza que ele continuaria funcionando corretamente por um mês, um ano ou 10 anos depois, com aproximadamente custo zero envolvido no processo.

Esse não é mais o caso quando você está implementando modelos de ML. Segundo o autor, essa suposição, errada, se mantém dentre pesquisadores focados apenas na teoria dos modelos que se esquecem do custo operacional, não só os referentes aos recursos computacionais, mas também do expertise exigido durante todo o processo.

Ainda segundo Talby, para mitigar esse desastre, existem algumas medidas remediadoras:

a) Medição online de precisão do modelo

   Assim como é preciso saber a latência de um site, é preciso saber a precisão dos modelos que estão em produção. Quantas previsões realmente se concretizaram? Isso requer a coleta e registro de resultados de uso real. Isso é um requisito elementar (TAULBY, 2019). Rastrear o desempenho de muitos modelos ao mesmo tempo é bastante difícil, e é provável que ocorram erros se não houver uma abordagem coordenada para monitorá-los. Segundo Lawrence (2020), MLOps ajuda a fazer o controle de versão de diferentes modelos e garantir que o desempenho deles esteja melhorando em vez de piorando;

b) Cuidado com a lacuna entre os dados

   Deve-se atentar para não haver diferenças entre os conjuntos de dados de treinamento e os conjuntos de dados online, ou seja, os que estão chegando “fresquinhos”. Esta é uma heurística simples de medir e que pode revelar uma variedade de problemas. Se os dados de treinamento têm 50% de determinado tipo, mas na produção está sendo previsto apenas 30% desse tipo, então provavelmente é hora de retreinar o modelo;

c) Alertas de qualidade dos dados online

> Se o número ou proporção dos dados de entrada mudar de forma inesperada, um alerta deve ir para a equipe de operações. Se o modelo não foi treinado para esses tipos de dados novos, e por isso houve um alerta, os resultados podem estar fazendo previsões erradas e um novo treinado deve ser iniciado, já considerando essas mudanças abruptas.

Com tudo isso, o investimento de tempo na configuração de MLOps pode ser considerável, mas os benefícios também são consideráveis, principalmente para as organizações que levam a sério o aprendizado de máquina reprodutível em produção e sua sustentabilidade a longo prazo. Os pipelines de ML automatizados permitem treinar e reimplantar modelos, bem como uma integração contínua e a garantia de que os processos não vão quebrar e os modelos vão continuar a melhorar.

Conforme a inevitável empregabilidade do aprendizado de máquina avança, tanto no campo empresarial como acadêmico, novos sistemas e infraestrutura capazes de lidar com ela também devem avançar, para torná-la disponível, confiável e catalisadora do desenvolvimento sustentável. Em pouco tempo, espera Lawrence (2020), a área de MLOps provavelmente será considerada tão comun e importantes quanto a de DevOps.

3. FATORES OBSERVADOS NA PRÁTICA DE *WEB SCRAPING* DE NOTÍCIAS EM PORTUGUÊS

Como não há fatores ou regras acordados formalmente por nenhum órgão normativo, como a Associação Brasileira de Normas Técnicas (ABNT), ou associações computacionais, como a Sociedade Brasileira de Computação (SBC), a seguir serão apresentadas as boas práticas e fatores reunidos pela autora após pesquisa bibliográfica e prática em construção de corpus de notícias em língua portuguesa.

A título de clareza e organização, os fatores estão divididos em duas seções. Na primeira seção estão aqueles que dizem respeito à seleção das fontes, e na segunda seção os que dizem respeito aos usos dessas fontes.

Conforme dito antes, o tema das notícias é o bioma costeiro e marinho brasileiro e o foco buscado, dentro desse tema, é biodiversidade, conservação e aspectos econômicos do referido

bioma. As notícias são copiadas e armazenadas em um arquivo JSON para posterior recuperação, a ser discutido na seção 3.2. O corpus foi construído tendo em mente sua futura utilização em modelos de NLP em português. Para tal, priorizou-se os principais divulgadores científicos em português, visando suas explicações simples com linguagem de alto alcance, diferentemente das encontradas em artigos científicos – linguagem voltada a outros cientistas, com referência constante a figuras e tabelas ao longo do texto – que dificulta a interpretação pelos modelos de NLP.

3.1 AS FONTES – ONDE BUSCAR OS DADOS

O processo de pesquisa das fontes não foi o mais óbvio, como pensado logo de início, de, por exemplo, simplesmente digitar o tema das notícias em algum mecanismo de busca e, então, explorar os links listados dentre as dezenas de páginas de resultados. A pesquisa iniciou-se, de fato, em mecanismos de buscas, porém a decisão quanto ao termo a ser buscado não foi tão simples. Inicialmente, o termo buscado era "Amazônia Azul"[19], o que trazia uma série de links de sites de compra online devido a semelhança com a loja Amazon, mas pouquíssimos resultados cabiam no objetivo. A busca não afunilava para o tema desejado nem com uma série de palavras-chaves e quantificadores propostos pelo site de busca. Depois das sucessivas buscas sem sucesso, foi percebido que esse termo ainda não é largamente empregado pela imprensa, livros didáticos e academia para se referir ao bioma marinho e costeiro brasileiro, sendo de uso corriqueiro apenas nos sites da Marinha do Brasil.

Então, outro termo passou a ser explorado, no caso, "Década do Oceano", com a intenção de deixar explícito o tema da busca. Dessa vez, os resultados alinhavam com a intenção, porém ou redirecionavam para páginas em inglês – língua largamente empregada nas comunicações da ONU, mas que não é o foco desse trabalho – ou eram notícias pontuais, sobre fatos relevantes

[19] Segundo a Marinha do Brasil, o termo "Amazônia Azul" foi primeiramente utilizado pelo Almirante-de-Esquadra Roberto de Guimarães Carvalho em 2004 no período em que era o Comandante da Marinha. Foi cunhado através da comparação das propriedades desse território marítimo com as do território amazônico, ambos abundantes em recursos naturais de importância estratégica para o Brasil. Mais informações estão disponíveis em: https://www.marinha.mil.br/spp/amaz%C3%B4nia-azul. Acesso em 08/11/2021.

ocorridos naquele dia ou semana, tais como acordos assinados visando ampliar o alcance das ações da Década do Oceano, a filiação de órgãos educacionais para maior divulgação das ciências oceânicas etc. No entanto, se em determinado momento do texto algum aspecto científico sobre o bioma era citado, então a notícia poderia ser considerada a entrar no corpus. Caso contrário, ela seria descartada, pois conforme já citado, o tema, dentro do foco, é biodiversidade, conservação e aspectos econômicos do bioma costeiro e marinho brasileiro.

Assim, após sucessivos insucessos, foi identificada a necessidade de se acrescentar nos termos de busca algumas palavras-chaves referentes aos focos citados, caso contrário o objetivo não seria alcançado. Alguns exemplos são “biodiversidade marinha no Brasil”, “conservação das praias brasileiras” e “exploração mineral marítmica no Brasil”. Os resultados da busca foram muito mais animadores, tanto no conteúdo das notícias quanto na descoberta de fontes.

As fontes, então, foram selecionadas e classificadas de acordo com sua natureza: podem ser projetos temáticos, universidades e seus laboratórios especializados, portal de notícias, fundações de amparo à pesquisa, entidades governamentais e empresas de impacto ambiental. Cada caso será detalhado a seguir.

Inicialmente, deu-se prioridade aos links de projetos que visam o estudo e preservação dos habitats e de animais marinhos, tanto da iniciativa privada quanto pública como, por exemplo, o Projeto Tamar, Coral Vivo e Baleia Jubarte. Em suas seções de notícias e publicações, eles tratam do tema de forma científica, simples e didática, raramente fazendo menções a tabelas e imagens, portanto, aumentando as chances de resultados frutíferos aos modelos NLP.

Posteriormente, ampliou-se a exploração de notícias para os links divulgados nos sites desses projetos. Isso, estrategicamente, possibilitou “pular” a etapa exaustiva de procurar resultados úteis nas páginas dos buscadores online, e ao mesmo tempo manter o tema e foco em questão. Um aspecto positivo, é que a maioria deles levavam aos sites de universidades públicas e seus laboratórios de pesquisa marinhas – que também indicavam outros links úteis dentro do tema, o que aumentou muito a descoberta de fontes com o perfil citado anteriormente. Com isso, conforme o número de sites visitados aumentava, também aumentava o número de fontes a serem exploradas. Alguns exemplos: Centro de Biologia Marinha da USP (CEBImar), Ciências do Mar da FURG e o Instituto de Estudos Avançados do Mar da UNESP (IEAMar).

Observou-se, durante o processo de *web scraping* nos sites dos projetos e universidades, que algumas das notícias eram replicadas de portais de notícias e jornais, como Folha de São Paulo, UOL e Globo.com. Alguns desses, no entanto, têm restrições quanto a replicação de suas notícias, e, portanto, foram desconsiderados para exploração. Outros, os que permitem a exploração por robôs, foram considerados e acrescentados à lista, como Jornal da USP, Revista FAPESP e Mongabay.

Além disso, muitas das notícias exploradas, independentemente da natureza da fonte, fazem menção à entidades governamentais, como IBAMA, ICMBio e Ministério do Meio Ambiente. Mais uma vez, fez-se o trabalho de seguir esses links para descobrir se o conteúdo era de interesse. Entretanto, se houvesse ali publicações de nenhuma ou baixa qualidade científica, os sites eram excluídos da lista de fontes. Além disso, sites preocupados em defender alguma ideologia foram descartados para se evitar, ao máximo, repassar esse viés aos modelos de NLP.

Essa seleção foi repetida para todos os links encontrados e, assim, foram classificados e salvos para exploração um total de 278 sites, com 7 naturezas diferentes.

3.2 OS USOS – COMO COLETAR E ARMAZENAR OS DADOS

Conforme dito antes, a seleção de fontes foi minuciosa, mas para um punhado de sites selecionados, houve outro punhado de sites descartados por impossibilitarem o processo de coleta. Antes de aplicar o processo de coleta, e durante a seleção das fontes, considerou-se os seguintes fatores:

a) Atualização dos dados

   Se o site em questão não apresentasse matérias e notícias atuais, digamos, com no máximo um ano da data da última publicação, era reconsiderada sua posição na lista de fontes. Era verificado, no entanto, se poderia ser vantajoso coletar o conteúdo antigo, divulgado até a data de visita. Mas futuramente, a inclusão do site na rotina do robô poderia consumir recursos à toa. O mesmo vale para sites cuja codificação é altamente dinâmica, como os de órgãos governamentais – eles mudam de "cara" a cada novo mandado de governo e isso pode quebrar a rotina do processo de coleta;

b) Links quebrados

Se muitos links do site estivessem quebrados, seja por falta de manutenção ou por se referirem a sites muito dinâmicos, a lista de novos links não era atualizada, senão impossibilitaria que o *crawler* execute sua função;

c) Frequência de visitas (*hits*) ao servidor

Não é boa prática, em nenhuma atividade online, solicitar dados aos servidores dos sites com uma frequência muito alta. Esse comportamento adiciona uma carga desnecessária aos servidores e se ela exceder um certo ponto, o serviço pode ficar lento ou travar, destruindo a boa navegação e experiência dos outros usuários. Além disso, o IP utilizado pelo *scraper* pode ser banido, o que impossibilita a rotina do processo de coleta. Recomenda-se usar no código do script funções que funcionem como um "intervalo" entre uma requisição e outra, como a *time.sleep()* em Python;

d) Pico de tráfego

Historicamente, a navegação na internet é melhor fora do horário de pico de alguns sites, como em lançamentos de eventos locais importantes. Assim, é possível evitar durante o rastreamento por dados que a rotina fique presa em filas de acesso por conta do tempo de inatividade de determinado servidor;

e) Uso responsável das informações

Conforme já citado anteriormente, as políticas de uso das informações de cada site devem ser respeitadas, tendo em mente que a publicação de dados protegidos por direitos autorais podem acarretar graves repercussões indesejadas. Isso pode ser evitado ao consultar o documento robots.txt, que todo site tem divulgado publicamente. Ele possui o conjunto de regras que definem como os robôs podem interagir com o site, e caso o *scraping* for feito de forma contrária a essas regras, pode levar a ações judiciais e multas.

Uma vez passada por todos os critérios anteriormente descritos, uma notícia era coletada através de um simples script em Python, usando-se o módulo *requests* e a biblioteca BeautifulSoup[20], que basicamente solicita o código-fonte de uma URL específica ao servidor, baixa e salva o conteúdo que é retornado, identifica os elementos da página que fazem parte do trecho desejado

[20] Documentação disponível em: https://www.crummy.com/software/BeautifulSoup/bs4/doc/. Acesso em 12/11/2021.

e extrai e reformata esses elementos em um conjunto de dados que podemos analisar ou usar da maneira preferida. No caso, transformar em formato de texto puro, sem código HTML, e salvá-lo num arquivo JSON.

A titulo informativo, a linguagem em que as páginas web são escritas, o HTML, tem muitas funções semelhantes às encontradas em um processador de texto como o Microsoft Word, por exemplo – como deixar o texto em negrito, criar parágrafos e assim por diante. O código final em HTML consiste em elementos chamados tags. A tag mais básica é a tag <html>. Essa tag informa ao navegador web que tudo dentro dela, até a tag </html>, é HTML. Em seguida, o conteúdo principal da página vai para a tag <body>, sendo os parágrafos separados pela tag <a> e . As tag mais importantes <class> e <id> têm propriedades especiais que fornecem nomes de elementos HTML e os tornam mais fáceis de interagir quando é feito o *scraping*. Um elemento pode ter várias classes e uma classe pode ser compartilhada entre os elementos. Cada elemento pode ter apenas um id, e um id só pode ser usado uma vez em uma página. Classes e ids são opcionais e nem todos os elementos os terão, mas eles são usados por CSS para determinar a quais elementos HTML aplicar certos estilos, e portanto, também podem ser usados para especificar os elementos visados. Saber como manipulá-los usando a biblioteca BeautifulSoup é essencial para automatizar o processo e economizar tempo retirando as tags do código fonte e transformando-o em texto puro.

Ao ser acrescentada ao corpus, uma notícia tinha não só o texto selecionado, mas também as datas de publicação e acesso, o link, a fonte e o autor do texto. Uma última classificação era acrescentada, a qual foi pensada em dar o modelo de NLP alguma pista sobre como interpretar o texto. Por exemplo, sabemos que ao ler uma entrevista, teremos uma experiência de interpretação de texto diferente de quando lemos um artigo, ou uma coluna de jornal que retrata um ponto de vista subjetivo. Então, pensando em facilitar o trabalho do modelo futuramente, foram acrescentadas as classificações de reportagem, entrevista, artigo e notícia.

A reportagem trata-se de uma matéria especialmente feita em torno de um assunto, por exemplo explicar um pouco sobre as espécies tubarões vistas no estado da Paraíba. A entrevista foi armazenada de forma que a pergunta ficasse isolada em uma linha e a resposta na linha seguinte. O artigo não tem sub-seções, é um texto direto e reflete um ponto de vista pessoal, que pode ou não ser aproveitado pelo modelo. E a notícia trata de um evento isolado, onde narra-se uma

história sobre ele citando pessoas, idades, profissões e datas, além de conteúdo científico, que não necessariamente seriam usados em um capítulo de livro, mas que podem ajudar o modelo a interpretar outras notícias, reportagens etc. Por exemplo, o derramamento de óleo ocorrido no Nordeste em 2019 foi muito noticiado pela imprensa brasileira e, nas principais reportagens, citava-se não só conteúdo científico como também pontual, eventual, de uma determinada semana ou mês. Isoladamente, essa notícia não seria de grande utilidade para o modelo montar o conhecimento do evento como um todo, mas ela pode ajudá-lo a organizar os fatos temporais, os acontecimentos e as informações de interesse ao resultado final.

A maneira mais prática e rápida encontrada para manter o corpus atualizado, ou seja, com notícias que ainda serão publicadas sobre o tema, foi usar um agregador de notícias (RSS *reader*, em inglês). Um *feed* RSS é um arquivo que contém um resumo das atualizações de um site, geralmente na forma de uma lista de artigos com links atualizados sobre novos conteúdos dos sites de interesse. Fundamentalmente, o RSS é simplesmente um arquivo de texto XML, criado pelo editor do site e contém uma lista contínua dos conteúdos publicados, com a entrada mais recente sempre no topo da lista. Cada entrada contém detalhes como o título do artigo, descrição e link para o conteúdo. Os *feeds* RSS são publicados e atualizados em tempo real, portanto, os assinantes sempre têm acesso ao conteúdo publicado mais recentemente e assim, se o corpus se manteria atualizado após rodar o script em cima dessas notícias.

Ao total, após dois meses de atualizações, o corpus armazenava um total de 705 notícias.

4. CONCLUSÕES

Muitos dos autores consultados para esse trabalho, como Chapman, Ilyas e Lorica, reforçam que “IA começa com dados bons”. Essa é uma declaração que recebe amplo acordo de cientistas de dados e acadêmicos. E, tanto na iniciativa privada como na acadêmica, houve um aumento significativo na capacidade de construir modelos complexos de IA para previsões, classificações e várias tarefas de análise. Já está amplamente disponível uma abundância de ferramentas que permitem aos profissionais do setor treinar modelos complexos em apenas alguns dias. À medida que a construção do modelo se torna mais fácil, o problema dos dados de alta qualidade torna-se mais evidente do que nunca.

Foi com esse cenário em mente que elaborou-se esse trabalho. O objetivo dele foi identificar quais fatores devem ser considerados na construção de um corpus de notícias em português, visando sua utilização futura em modelos de NLP em português, segundo os autores pesquisados e aplicação prática da autora. Os fatores identificados foram expostos no capítulo 3, item por item. Quando aplicados na prática, em conjunto com scripts em Python e agregadores de notícias, possibilitaram a construção de um corpus pela autora que pode ser acessado sob consulta.

O processo de descobrimento e seleção de fontes demandou muito mais tempo do que se imaginava inicialmente. Apesar dos buscadores online serem bem-sucedidos em buscas no dia-a-dia, em uma busca temática não se mostraram muito animadores. Foram encontradas muito mais notícias do que reportagens, ou seja, há uma preocupação maior em informar sobre fatos corriqueiros do que escrever matérias mais elaboradas sobre assuntos científicos. Por exemplo, muitas são as notícias sobre encalhamento de baleias, pinguins encontrados pelas praias do Sul e derramamento de óleo do Nordeste. Mas há poucas reportagens sobre a biodiversidade de espécies nas restingas e interação ecológica portuária.

Além disso, encontrar conteúdo de alta qualidade científica em Língua Portuguesa, sobre o tema e foco desejados, mostrou-se desafiador também. Infelizmente, são poucos os divulgadores científicos e os sites das universidades com cursos de ciências oceânicas e laboratórios dedicados a pesquisa e, mesmo dentre esses, os sites não são atualizados com tanta frequência. Alguns são claramente abandonados após o fim de um projeto, sem compromisso com a continuidade de divulgação dos resultados. Muitas vezes, na parte de publicações, são divulgados apenas links para artigos científicos do grupo, sem nenhuma preocupação em construir uma seção de “tradução” deles para um material que transmita o conhecimento em linguagem mais acessível, dedicada aos leitores leigos.

Com a Década do Oceano à frente, está sendo mostrado mais interesse, tanto acadêmico como político, em contruir essa literatura de alto alcance popular. Mas o processo está lento e poucos divulgadores têm se mostrado engajados de fato. Espera-se que com a pressão por medidas urgentes em favor do clima também se crie uma agenda de educação oceânica popular, para que todos possam participar juntos das mudanças necessárias. Isso favoreceria não só a

educação das próximas gerações como também facilitaria a alimentação de modelos de NLP de português.

Por fim, e mais importante, deve-se enfatizar que sem os recursos trazidos pela computação em nuvem dificilmente um projeto de IA será bem-sucedido. Só a nuvem oferece uma infraestrutura completa e capaz de treinar um modelo e posteriormente alimentá-lo com dados atualizados.

5. POSSÍVEIS DESDOBRAMENTOS

Durante a pesquisa, foram detectados dois tópicos bem interessantes que derivam de alguns dos assuntos aqui discutidos, mas, por não se encaixarem perfeitamente no tema do trabalho, não foram descritos detalhadamente com a atenção que merecem.

O primeiro deles é como seria o processo de construção de um corpus de artigos científicos e capítulos de livros especializados em português. Esses tipos de material trazem mais informações ao leitor, como descrições mais ricas em detalhes, gráficos e tabelas que certamente beneficiariam os modelos de NLP. Ao fornecerem um contexto mais amplo e, portanto, mais conhecimento para o modelo, damos mais subsídios para que ele calcule os parâmetros envolvidos ao longo do processo mais precisamente.

Pensando nessa possibilidade, e já aproveitando que a coleta de notícias estava em andamento, também foram salvos alguns links de livros e artigos sobre o bioma marinho e costeiro brasileiro. Assim, pode-se descartar a dolorosa etapa de busca online em torno do tema. Porém, uma observação deve ser feita nesse caso: a vasta maioria dos artigos observados estão em inglês, pouquíssimo em português. Então antes de treinar um modelo deve-se ter isso em mente: a literatura acadêmica é em inglês.

O outro tópico a ser desdobrado seria o uso desse corpus para alimentar programas de perguntas e respostas. Esse tipo de software, chamado de Q&A, pode ser encontrado nas seções de atendimento ao consumidor ou de perguntas mais frequentes (FAQ) em vários websites. Eles são empregados, por exemplo, em chats que simulam interação com humanos mas que, na

verdade, são processados por robôs. Para construir uma resposta à determinada pergunta, são feitas consultas a uma, ou algumas, base de conhecimento.

Segundo Patterson (2021), uma base de conhecimento é uma coleção de documentos cujo conteúdo fornece respostas às perguntas frequentes, guias de procedimentos e instruções para solução de problemas. Deve-se ter em mente que isso é um projeto de longo prazo – além de requerer constante atualizações à medida que novas notícias são lançadas, também deve-se considerar "novas versões" científicas sobre determinado assunto. Por exemplo, em novembro de 2021 foi divulgada pela Revista Fapesp[21] que as baleias de grande porte consomem três vezes mais alimento por ano do que se estimava anteriormente. Com isso, não deve ser negligenciado o fato de que todas as notícias já inclusas no corpus que citam os hábitos das baleias devem ser, de alguma forma, atualizadas com essa nova informação ao serem recuperadas pelo usuário ou modelo de NLP.

[21] A notícia completa está disponível em: https://revistapesquisa.fapesp.br/as-grandes-adubadoras-dos-oceanos/. Acesso em 02/12/2021.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ARGUELLO, Jaime; DIAZ, Fernando; CALLAN, Jamie. **Learning to Aggregate Vertical Results into Web Search Results**. Disponível em: https://ils.unc.edu/~jarguell/ArguelloCIKM11.pdf. Acesso em: 25/10/2021.

BIRD, Steven; KLEIN, Ewan; LOPER, Edward. **Natural Language Processing with Python**. Disponível em: http://nltk.org. Acesso em: 10/10/2021.

BORCAN, Marius. **What Is Natural Language Processing? A Gentle Introduction to NLP**. Disponível em: https://towardsdatascience.com/what-is-natural-language-processing-a-gentle-introduction-to-nlp-4ed219a768ad. Acesso em: 02/11/2021.

CETA, Noel. **What Is a Knowledge Base and Why Do I Need One?** Disponível em: https://document360.com/blog/what-is-a-knowledge-base/. Acesso em: 26/10/2021.

CHAPMAN, Rachel. **Top 10 Web Scraping Techniques**. Disponível em: https://limeproxies.netlify.app/blog/top-10-web-scraping-techniques. Acesso em: 26/10/2021.

CORPUS. In.: **Dicionário Priberam da Língua Portuguesa** [em linha], 2021. Disponível em: https://dicionario.priberam.org/corpus. Acesso em: 10/10/2021.

CRAWLER. In.: **RyteWiki**, 2021. Disponível em: https://en.ryte.com/wiki/Crawler. Acesso em: 31/10/2021.

DALE, Kyran. **Data Visualization with Python and JavaScript**. 1ª Ed – Sebastopol, CA: O'Reilly Media Inc., 2016.

DESJARDINS, Jeff. **How much data is generated each day?** Disponível em: https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/. Acesso em: 10/10/2021.

DOCUMENTAÇÂO Selenium. **The Selenium Browser Automation Project**. Disponível em: https://www.selenium.dev/documentation/. Acesso em: 26/10/2021.

GUILLOU, Pierre. **NLP – Datasets em português**. Disponível em: https://medium.com/@pierre_guillou/nlp-datasets-em-portugu%C3%AAs-7e1790a44d42. Acesso em: 06/11/2021.

HENRIQUEZ, Alvaro. **Age of the Transformers**. Disponível em: https://medium.com/swlh/age-of-the-transformers-31b208cdbed4. Acesso em: 06/11/2021.

IFTEKHAR, Muhammad Hanan. **Natural Language Processing (NLP) speech to text (Technical)**. Disponível em: https://www.finsliqblog.com/ai-and-machine-learning/natural-language-processing-nlp-speech-to-text-technical/. Acesso em: 06/11/2021.

ILYAS, Ihab; LORICA, Ben. **The quest for high-quality data**. Disponível em: https://www.oreilly.com/radar/the-quest-for-high-quality-data/. Acesso em: 31/10/2021.

JOHNSON, Dave. **A guide to using RSS feeds, the files that contain real-time updates from websites**. Disponível em: https://www.businessinsider.com/what-is-rss-feed. Acesso em: 15/11/2021.

KAZIL, Jacqueline; JARMUL, Katharine. **Data Wrangling with Python**. 1ª Ed – Sebastopol, CA: O'Reilly Media Inc., 2016.

KRAYEWSKI, Kalia. **How NLP Text-Based Chatbots Work**. Disponível em: https://www.ultimate.ai/blog/ai-automation/how-nlp-text-based-chatbots-work. Acesso em 03/11/2021.

KUMARI, Riya. **4 Types of Machine Translation in NLP**. Disponível em: https://www.analyticssteps.com/blogs/4-types-machine-translation-nlp. Acesso em: 03/11/2021.

LAWRENCE, Trey. **MLOps: What It Means and Why It Matters**. Disponível em: https://spell.ml/blog/mlops-what-it-means-and-why-it-matters-Xw8uYhEAACAAtDnG. Acesso em: 28/11/2021.

McKINSEY Consulting. **Global Survey: The state of AI in 2020**. Disponível em: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/global-survey-the-state-of-ai-in-2020. Acesso em: 10/10/2021.

MITCHELL, Ryan. **Web scraping with Python**. 1ª Ed – Sebastopol, CA: O'Reilly Media Inc., 2015.

PATTERSON, Mathew. **The Ultimate Guide to Using a Knowledge Base for Self-Service Support**. Disponível em: https://www.helpscout.com/playlists/knowledge-base/. Acesso em: 02/12/2021.

PILGRIM, Mark. **HTTP Web Services**. Disponível em: https://diveintopython3.net/http-web-services.html. Acesso em: 25/10/2021.

PRESS, Gil. **Andrew Ng Launches A Campaign For Data-Centric AI**. Disponível em: https://www.forbes.com/sites/gilpress/2021/06/16/andrew-ng-launches-a-campaign-for-data-centric-ai/. Acesso em: 10/10/2021.

REINSEL, David; GANTZ, John; RYDNING, John. **The Digitization of the World: From Edge to Core**. Disponível em: https://www.seagate.com/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf. Acesso em: 10/10/2021.

RUDER, Sebastian; EISENSCHLOS, Julian. **Efficient multi-lingual language model fine-tuning**. Disponível em: https://nlp.fast.ai/. Acesso em: 06/11/2021.

SAHA, Debanjan. **How The World Became Data-Driven, And What's Next**. Disponível em: https://www.forbes.com/sites/googlecloud/2020/05/20/how-the-world-became-data-driven-and-whats-next/. Acesso em: 10/10/2021.

SCULLEY, D.; HOLT, Gary; GOLOVIN, Daniel; DAVYDOV, Eugene; PHILLIPS, Todd *et al*. **Hidden Technical Debt in Machine Learning Systems**. Disponível em: https://proceedings.neurips.cc/paper/2015/file/86df7dcfd896fcaf2674f757a2463eba-Paper.pdf. Acesso em: 12/11/2021.

SO, Kenn; LORICA, Ben. **Data Quality Unpacked**. Disponível em: https://gradientflow.com/data-quality-unpacked/. Acesso em: 10/10/2021.

SIRISURIYA, SCM de S. **A Comparative Study on Web Scraping**. Disponível em: http://ir.kdu.ac.lk/bitstream/handle/345/1051/com-059.pdf?sequence=1&isAllowed=y. Acesso em: 25/10/2021.

TALBY, David. **Why Machine Learning Models Crash And Burn In Production**. Disponível em: https://www.forbes.com/sites/forbestechcouncil/2019/04/03/why-machine-learning-models-crash-and-burn-in-production/. Acesso em: 12/11/2021.

VECCHIETTI, Luiz Felipe Santos. **Comparison between rule-based and data-driven natural language processing algorithms for Brazilian Portuguese speech synthesis**. 2017. 72 fls. Dissertação de mestrado – Programa de Pós-graduação em Engenharia Elétrica, UFRJ/COPPE, Rio de Janeiro, 2017.

WEB SCRAPING. Disponível em: https://en.wikipedia.org/wiki/Web_scraping. Acesso em: 12/10/2021.

WILLIAMS, Janet. **What is web scraping?** Disponível em: https://www.promptcloud.com/blog/what-is-web-scraping/. Acesso em: 30/10/2021.