

Pontifícia Universidade Católica de Minas Gerais

Pós-graduação em Ciência de Dados e Big Data

Web Scraping no apoio à Análise Linguística: Construção de um Corpus Textual

Aluno: Marcelo Honório de Oliveira

Orientador: Cristiano Rodrigues de Carvalho

Belo Horizonte

2018

SUMÁRIO

1	RESUMO EXECUTIVO.....	6
2	CARACTERIZAÇÃO DO PROBLEMA, MOTIVAÇÕES, OBJETIVO.....	7
3	WEB SCRAPING E A FERRAMENTA KNIME.....	8
3.1	Técnicas de <i>Web Scraping</i>	8
3.2	Ferramenta KNIME Analytics Platform.....	13
4	O FLUXO DE DADOS DO KNIME.....	15
4.1	Passo 1: Planejamento.....	16
4.2	Passo 2: Definição do método de extração.....	17
4.3	Passo 3: Instalação e configuração da ferramenta KNIME.....	18
4.4	Passo 4: Criação e configuração do fluxo de dados.....	18
4.4.1	Scraping no Site da SciELO.....	19
4.4.2	Scraping no Site da Oxford University.....	19
4.5	Passo 5: Execução do fluxo e armazenamento dos dados.....	20
4.6	Passo 6: Validação dos dados extraídos.....	22
5	APLICAÇÃO PRÁTICA NA ANÁLISE LINGUÍSTICA.....	24
6	CONCLUSÕES.....	26
7	TRABALHOS FUTUROS.....	27
8	REFERÊNCIAS BIBLIOGRÁFICAS.....	28

LISTA DE ILUSTRAÇÕES

Figura 1: Protocolo de Exclusão de Robôs - SciELO	9
Figura 2: Protocolo de Exclusão de Robôs - Oxford University Press	9
Figura 3: Meta Tag Robot – SciELO (Encontrada com valor “all”).....	11
Figura 4: Meta Tag Robot – SciELO (Não encontrada).....	11
Figura 5: Extensão Scraper - Área de <i>download</i>	12
Figura 6: Exemplo de uso do Scraper.....	13
Figura 7: Tela inicial do KNIME	14
Figura 8: Ciclo de Projeto de Mineração de Dados (Barbieri, 2011)	15
Figura 9: Caminho de navegação SciELO	16
Figura 10: Caminho de navegação Oxford.....	17
Figura 11: Mineração Web (Kosala & Blockeel, 2000).....	17
Figura 12: KNIME - Document View – SciELO	20
Figura 13: KNIME - Document View – Oxford University Press	21
Figura 14: Saída Arquivos de Texto - SciELO - Português	21
Figura 15: Saída Arquivos de Texto - SciELO - Inglês	22
Figura 16: Saída Arquivos de Texto - Oxford - Inglês.....	22
Figura 18: COCA - Corpus of Contemporary American English.....	24
Figura 17: Repositório GitHub - marcelohonoliveira - TCC	25

LISTA DE TABELAS

Tabela 1: Páginas raspadas – Português	23
Tabela 2: Páginas raspadas – Inglês	23

AGRADECIMENTOS

Agradeço ao Me. Estêvão Carvalho Batista pela mentoria dispensada desde a concepção da ideia à concretização deste estudo.

Ao professor orientador Cristiano Rodrigues de Carvalho pelas revisões atentas e criativas.

E a todos que, de várias formas, contribuíram para a realização deste trabalho.

1 RESUMO EXECUTIVO

O trabalho mostra os passos realizados no desenvolvimento de um fluxo de dados para coleta e organização de textos de páginas da *web* utilizados para uma análise linguística por meio da construção de um *corpus*¹. Os problemas que podem ocorrer e as respectivas soluções são apresentadas de forma detalhada.

As atividades realizadas no caso apresentado podem servir de referência para novas coletas onde as técnicas de *Web Scraping* se aplicam e concernentes às limitações próprias de uma coleta automática na *web*.

¹ Um *corpus* linguístico é uma coleção de textos que foram selecionados e reunidos para que a linguagem possa ser estudada no computador. (Wynne, 2005)

2 CARACTERIZAÇÃO DO PROBLEMA, MOTIVAÇÕES, OBJETIVO

Um das dificuldades apresentadas por linguistas é a aquisição de material para a realização de estudos que abrangem o uso de termos e expressões em um idioma e a escassez dos recursos computacionais na pesquisa linguística. Uma das razões é o a falta de conhecimento dos instrumentos disponíveis. Diversos trabalhos na área da linguística utilizam poucos textos devido à carência de material formatado e organizado especialmente quando este é proveniente da *web* ou de redes sociais (Sardinha, 1999).

Ao se deparar com as limitações dos especialistas em idioma, viu-se a necessidade de prover uma forma de facilitar a extração, organização e armazenamento de um número razoavelmente grande de texto para a análise linguística. A partir desta motivação, o desafio foi desenvolver uma rotina que, por meio de automatização do processo de coleta, fosse capaz também de realizar a organização dos textos em documentos classificados e disponibilizar aos interessados o *corpus* em arquivos de texto para aplicação da análise linguística em si.

O produto final deste trabalho foi, portanto, a construção de um *corpus* textual que possibilite ao linguista analisar o uso de verbos modais do inglês (*will, should, would, may, might, must*) nas normas de submissão de artigos de periódicos disponíveis na plataforma SciELO - Scientific Electronic Library (SciELO, 2018) e a tradução de tais normas do português para o inglês, bem como as escolhas tradutórias e o uso das mesmas estruturas por nativos conforme as instruções da Oxford University Press (Oxford University, 2018).

A motivação inicial se deu a partir da demanda particular por uma coleta de dados para uma futura pesquisa de pós-doutoramento do Mestre em Estudos Literários Estêvão Carvalho Batista que pretende analisar traços da escrita acadêmica por falantes do inglês e português brasileiro. Logo, o objetivo deste relatório foi demonstrar como se deu o desenvolvimento da ferramenta que extraiu da *web* textos específicos de fontes pré-determinadas e que, cujo resultado, proporcionasse, a quem interessar, a possibilidade de analisar o material coletado e observar o uso de termos específicos de um determinado idioma.

3 WEB SCRAPING E A FERRAMENTA KNIME

3.1 Técnicas de *Web Scraping*

Scraping ou Raspagem é a forma de se extrair dados da *web* e movê-los para um formato mais simples com o objetivo de facilitar a sua análise e cruzá-los com outras fontes com mais flexibilidade. Para realizar, portanto, essa extração, se faz necessária a utilização de ferramentas computacionais que aplicam tais técnicas. (Andriolo, 2012)

Web Scraping é uma técnica de software de computador utilizada para extrair informações de sites e consiste, principalmente, na transformação de dados não estruturados (HTML - *HyperText Markup Language*) na *web* em dados estruturados - banco de dados ou planilha eletrônica por exemplo (Ray, 2015).

Após a definição da fonte da coleta, é importante verificar se o dado ali disponibilizado está aberto para ser acessado por meio de ferramentas de automatização de leitura (robôs). Uma das formas que as fontes de dados têm para sinalizar que é autorizado aplicar *scraping* é o uso do Protocolo de Exclusão de Robôs - método empregado pelos administradores de sistemas para informar aos robôs visitantes quais partes de um site não devem ser raspados por eles (Wikipédia, 2017).

A coleta deste trabalho se deu em dois sites:

- SciELO - Scientific Electronic Library (SciELO, 2018)
- Oxford University Press (Oxford University, 2018)

Em ambas fontes, a aplicação do *scraping* foi permitida, pois o arquivo Robots.txt não existe no site do SciELO e no Oxford University Press não há restrição para os diretórios raspados neste trabalho².

² O arquivo “Robots.txt” para a Oxford University Press restringiu o acesso a todo o conteúdo do site apenas ao agente “008” que é utilizado pelo provedor de servidos de rastreamento da *web* 80legs (Datafiniti, LLC, 2018). O 80legs permite que seus usuários criem e executem rastreamentos da *web* personalizados (BotReports, 2014).

SciELO: <http://www.scielo.org/Robots.txt>



Figura 1: Protocolo de Exclusão de Robôs - SciELO

Oxford University Press: <https://academic.oup.com/Robots.txt>



Figura 2: Protocolo de Exclusão de Robôs - Oxford University Press

Além dos arquivos tipo “Robots.txt”, foram verificadas as *Meta Tags* incorporadas aos sites estudados. *Meta Tags* são estruturas de dados sobre os próprios dados, uma breve descrição do

conteúdo da página, seu autor, data de criação, linguagem e outras informações relevantes (Gazola, 2016).

A *Meta Tag* relacionada ao comportamento que o robô deve assumir acerca da permissão ou não para a coleta é a definida pela especificação “*robots*” conforme exemplo abaixo:

```
<meta name="robots" content="all" />
```

Por meio do navegador *web* Google Chrome (Alphabet Inc., 2018), foram verificadas tais *Meta Tags*:

- *view-source:http://www.scielo.org/php/index.php*
- *view-source:https://academic.oup.com/journals*

Em ambas fontes, a aplicação do *scraping* foi permitida, pois a *Meta Tag* não existe no site da Oxford University Press e no do SciELO não há restrição para a raspagem: a *Meta Tag* exibe o valor “*all*”³.

³ All: Valor *default*, significa vazio, o robô de busca não recebe nenhuma informação. Não há restrições para a indexação ou a veiculação. Essa diretiva é o valor padrão e não terá efeito se for listada explicitamente. (Google Inc, 2018)

```

1 <!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
2 "http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
3 <html>
4   <head>
5     <title>
6       SciELO - Scientific Electronic Library Online     </title>
7     <meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1"/>
8     <meta http-equiv="Expires" content="-1"/>
9     <meta http-equiv="pragma" content="no-cache"/>
10
11    <meta name="author" content="BIREME (http://www.bireme.br/)"><meta name="keywords" content="in
12    <meta name="description" content="Biblioteca Virtual em Saúde"/>
13    <meta name="robots" content="all" />
14    <meta name="MSSmartTagsPreventParsing" content="true" />
15    <meta name="generator" content="BVS-Site 5.2.1" />
16
17    <script type="text/javascript">var lang = 'pt';</script>
18    <script type="text/javascript" src="/js/functions.js"></script>
19    <script type="text/javascript" src="/js/showHide.js"></script>
20    <script type="text/javascript" src="/js/metasearch.js"></script>
21    <script type="text/javascript" src="/applications/scielo-org/js/metaSearch.js"></script>
22    <script type="text/javascript" src="/js/jquery.js"></script>

```

Figura 3: Meta Tag Robot – SciELO (Encontrada com valor “all”)

```

1 <!DOCTYPE html>
2 <html lang="en">
3
4   <head><title>
5     Journals | Oxford Academic
6   </title>
7   <!-- Responsive View Port -->
8   <meta name="viewport" content="width=device-width, initial-scale=1, maximum-scale=1" />
9
10  <!-- Meta -->
11  <meta charset="utf-8" /><meta http-equiv="Content-Type" content="text/html; charset=utf-8" /><meta http-equiv="X-UA-Compatible" content="IE=Edge" /><script type="text/javascript">window.N
12  ({});NREUM.info = { "beacon": "bam.nr-data.net", "errorBeacon": "bam.nr-
13  data.net", "licenseKey": "8d76caef26", "applicationID": "29565620", "transactionName": "ZQBRMKVUCtRVkIXLxKcJvNg9WVFBOS1B8FUs=", "queueTime": 0, "applicationTime": 206, "agent": "", "atts": "" }</script><
14  type="text/javascript">(window.NREUM||(NREUM={})).loader.config={xpid:"XAYBV15MGwEJUVdXQMH"};window.NREUM||(NREUM={}).__nr_require=function(t,e,n){function r(n){if(!e[n]){var o=e[n]={exports:
15  {}},call(o,exports,function(e){var o=t[n][1][e];return r(o[1],e)},o,o,exports)}return e[n].exports}if("function"==typeof __nr_require)return __nr_require;for(var o=0;o<n.length;o++)r(n[o]);retu
16  r(n.stack);c.dev&&(n("HR AGENT IN DEVELOPMENT MODE"),r({flags: "a(c,function(t,e){return t}).join(", ");}));2:[function(t,e,n){function r(t,e,n,r,c){try{h2h=1;o(c)||new
17  UncaughtException(t,e,n,0)}catch(f){try{i("ierr",[f.s.now(),0])}return"function"==typeof u&&u.apply(this,a(arguments))}function UncaughtException(t,e,n){this.message||"Uncaught
18  additional information",this.sourceURL,e,this.line}function o(t,e){var n=e?null:s.now();i("err",[t,n])}var
19  i="handle",a=21,c="ee",s="loader",f="gos",u=window.onerror,d=1,p="nr@seenError",h=0,s.features.err=0,t(1),window.onerror=try(throw new Error("stack in 1&&
20  (t(13),t(12),addEventListener in window&&t(6),s.xhrWrappable&&t(14),d=10)}c.on("fn-start",function(t,e,n){d&&(h=1)}),c.on("fn-err",function(t,e,n){d&&in[p]&&(f(n,p,function()
21  {return 0}),this.throw=1,o(n)}),c.on("fn-end",function(){d&&this.throw&&h&&(h=1)}),c.on("internal-error",function(t){i("ierr",[t.s.now(),0])}),3:[function(t,e,n){t("loader").fea
22  t(1),4:[function(t,e,n){function r(){(M+=,s=s.hash,this[u]=b.now())function o(){(M--,y.hash+=5&&(h=10),var t=b.now(),this[i]=this[i]+t-this[u],this[d]=t}function i(t,e){t.emit("newURL",[t,y
23  a(t,e){t.on(e,function(){this[e]=b.now()})}var c="start",s="end",f="

```

Figura 4: Meta Tag Robot – SciELO (Não encontrada)

Para facilitar o mapeamento do conteúdo das páginas de interesse, foi utilizada uma extensão do navegador *web* Google Chrome (Alphabet Inc., 2018) chamada Scraper na versão 1.7 atualizada em 20 de abril de 2015. O Scraper é uma extensão de Mineração de Dados⁴ muito simples (mas limitada) para facilitar a pesquisa on-line quando é necessário obter dados

⁴ A Mineração de Dados é o processo de descoberta de informações acionáveis em grandes conjuntos de dados. A mineração de dados usa análise matemática para derivar padrões e tendências que existem nos dados. Normalmente, esses padrões não podem ser descobertos com a exploração de dados tradicional pelo fato de as relações serem muito complexas ou por haver muitos dados. (Microsoft Corporation, 2017)

rapidamente em formato tabular. Destina-se como uma ferramenta fácil de usar para usuários intermediários a avançados que se sentem confortáveis com o XPath⁵ (Google, 2015).

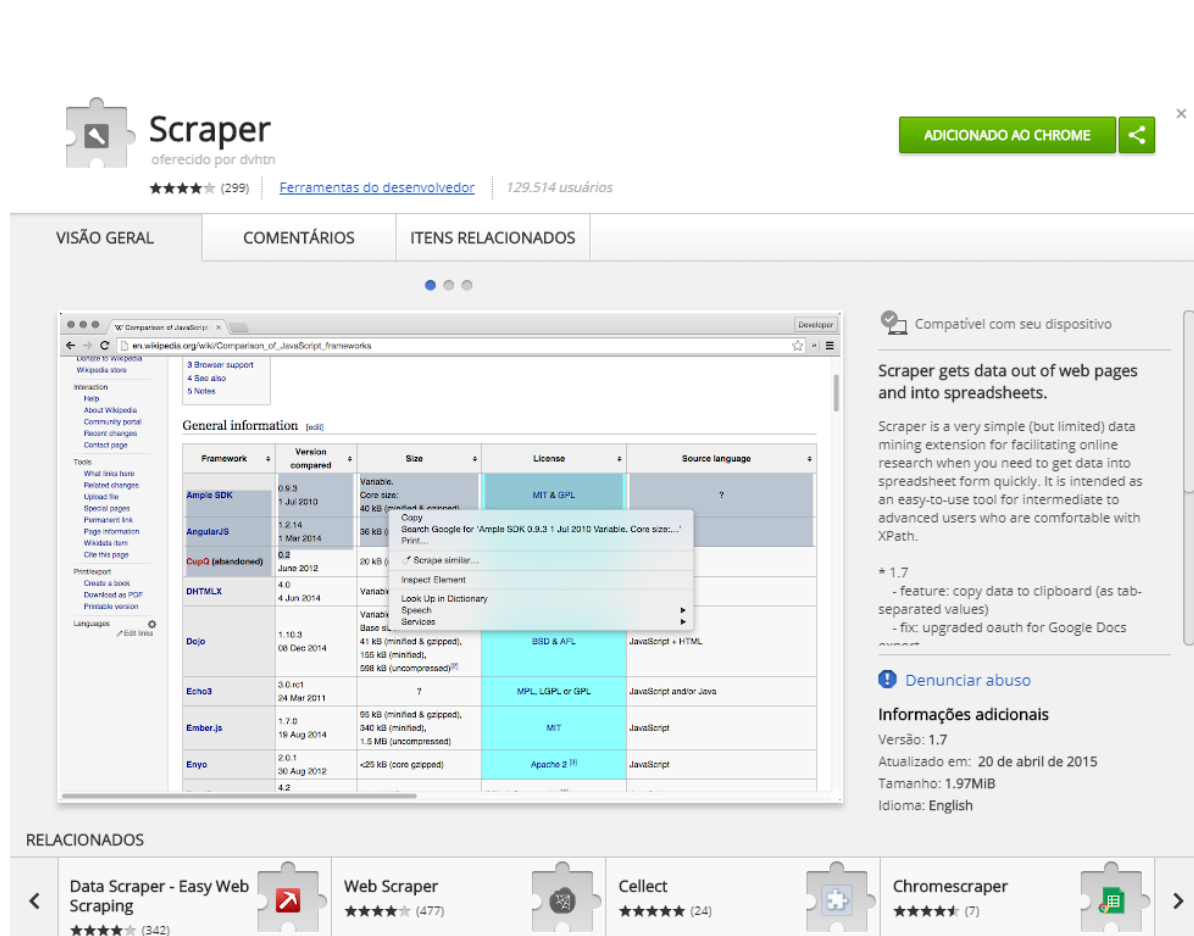


Figura 5: Extensão Scraper - Área de *download*

A extensão Scraper foi utilizada, especialmente, para listar as partes do site que endereçassem os conteúdos de interesse. No exemplo abaixo, observa-se a ferramenta fornecendo a referência XPath para os itens de menu dos Assuntos do site SciELO.

⁵ XPath: forma pela qual se pode referir partes de um documento XML (W3C, 2018).

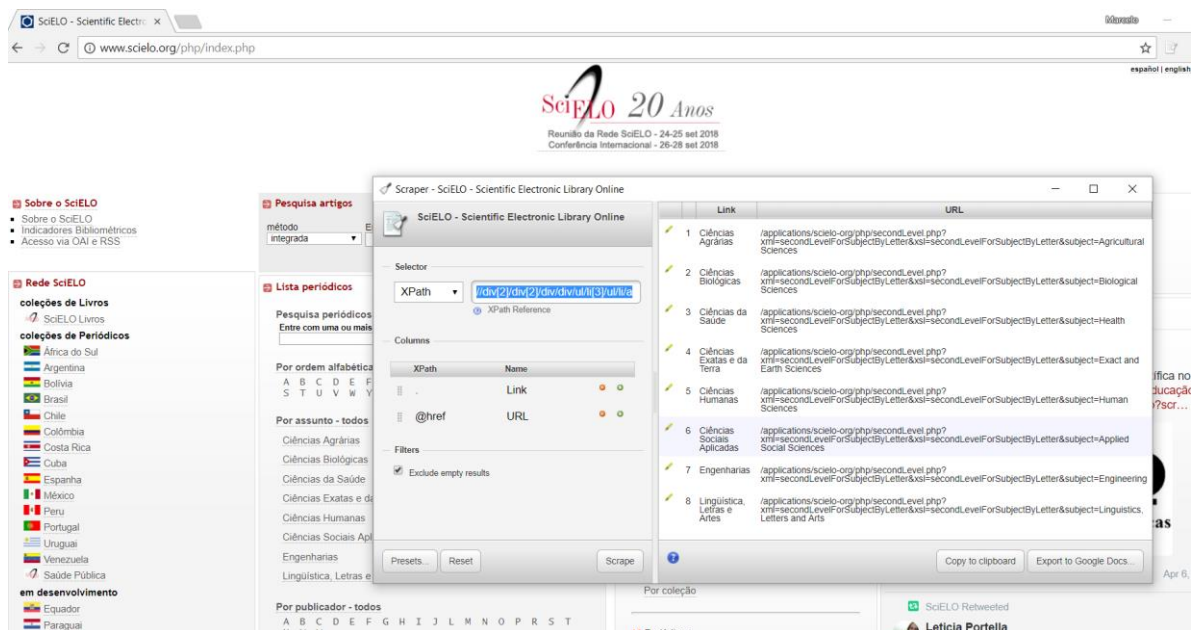


Figura 6: Exemplo de uso do Scraper

3.2 Ferramenta KNIME Analytics Platform

De acordo com o *site* da empresa, a plataforma analítica KNIME é uma solução aberta à inovação baseada em dados e projetada para descobrir o potencial da Mineração de Dados para novos *insights* ou previsões (KNIME, 2017).

A ferramenta integra vários componentes para Aprendizagem de Máquina e Mineração de Dados por meio de um conceito modular de pipeline de dados. A interface gráfica do usuário permite a montagem rápida e fácil de nós (*nodes*) para o pré-processamento de dados (ETL: *Extract, Transform and Load* - Extração, Transformação e Carregamento), para modelagem, análise e visualização de dados (Russell & Cohn, 2012).

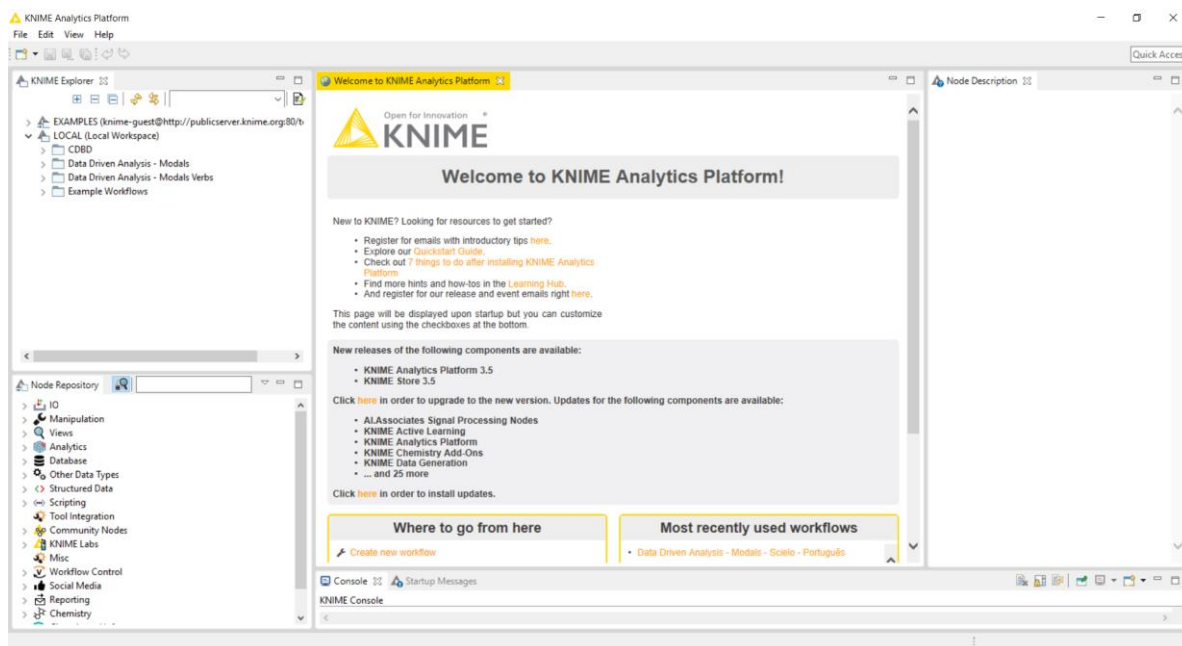


Figura 7: Tela inicial do KNIME

Neste trabalho, foi utilizado o KNIME na versão 3.4.0 e licenciada sob a Licença Pública Geral (GNU) – Versão 3 com todos as extensões então disponíveis.

4 O FLUXO DE DADOS DO KNIME

Este capítulo é dedicado à descrição do procedimento completo em alto nível para desenvolvimento do fluxo de dados na ferramenta KNIME.

O procedimento foi testado completamente e cada um de seus passos será descrito detalhadamente. Assim, a execução das etapas descritas poderá ser de grande valia para obtenção de sucesso em uma nova extração de outras fontes da *web*.

Todo o trabalho foi baseado no ciclo de desenvolvimento de um projeto de Mineração de Dados conforme diagrama abaixo cuja proposta original é composta por seis fases básicas:

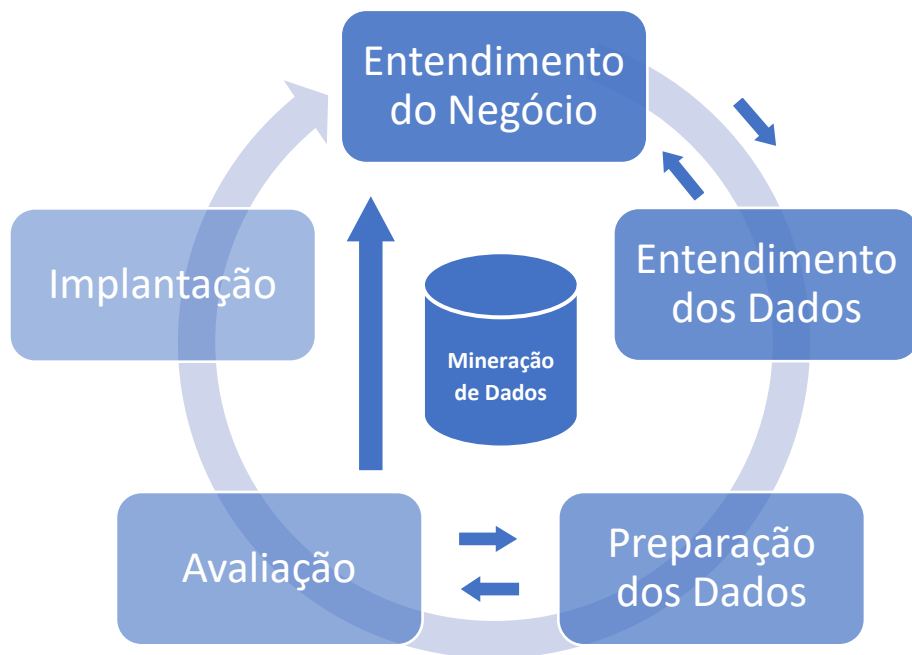


Figura 8: Ciclo de Projeto de Mineração de Dados (Barbieri, 2011)

O trabalho de desenvolvimento do fluxo foi composto por uma série de passos, citados e descritos a seguir:

- Planejamento
- Definição do método de extração
- Instalação e configuração da ferramenta KNIME

- Criação e configuração do fluxo de dados
- Execução do fluxo e armazenamento dos dados
- Validação dos dados extraídos

4.1 Passo 1: Planejamento

A demanda inicial foi extrair as páginas de instruções aos autores que desejassem submeter artigos de periódicos nas plataformas SciELO - Scientific Electronic Library e Oxford University Press.

Portanto, o primeiro passo foi entender como e onde se localizam tais instruções e o caminho de navegação até elas. O diagrama a seguir demonstra como isso se deu:

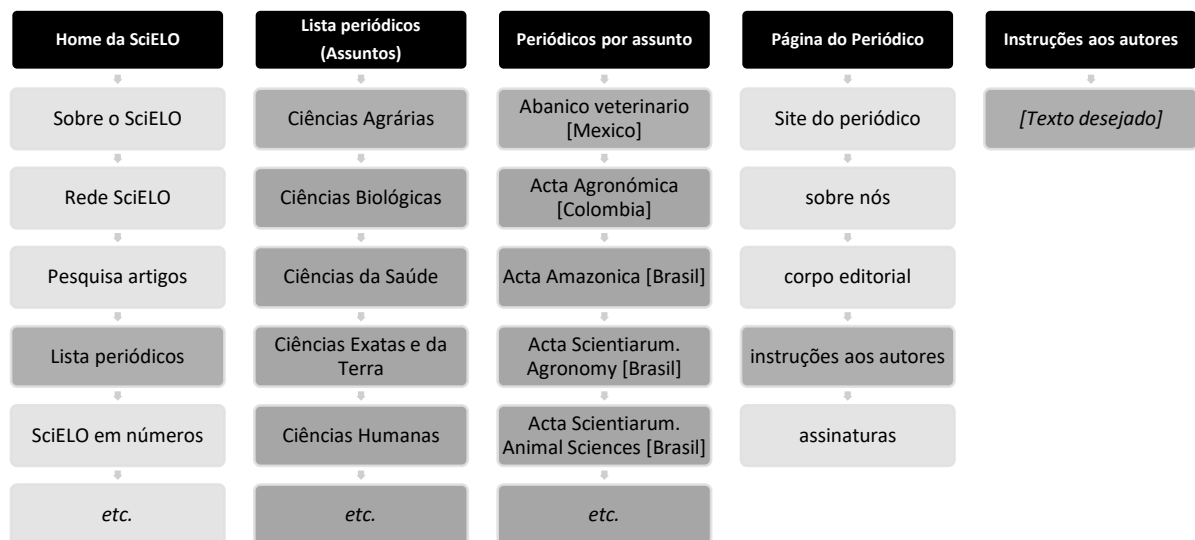


Figura 9: Caminho de navegação SciELO

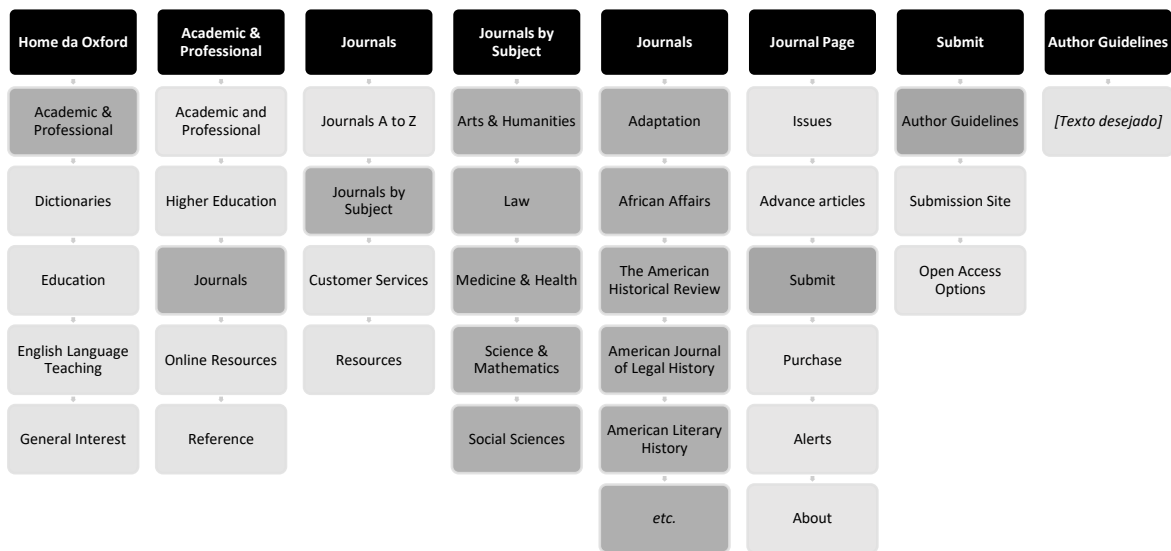


Figura 10: Caminho de navegação Oxford

4.2 Passo 2: Definição do método de extração

De acordo com (Kosala & Blockeel, 2000), mineração *web* pode ser dividida em três subáreas: Mineração de Estrutura (*Web Structure Mining*), Mineração de Uso (*Web Usage Mining*) e Mineração de Conteúdo (*Web Content Mining*), como observado na figura baixo:

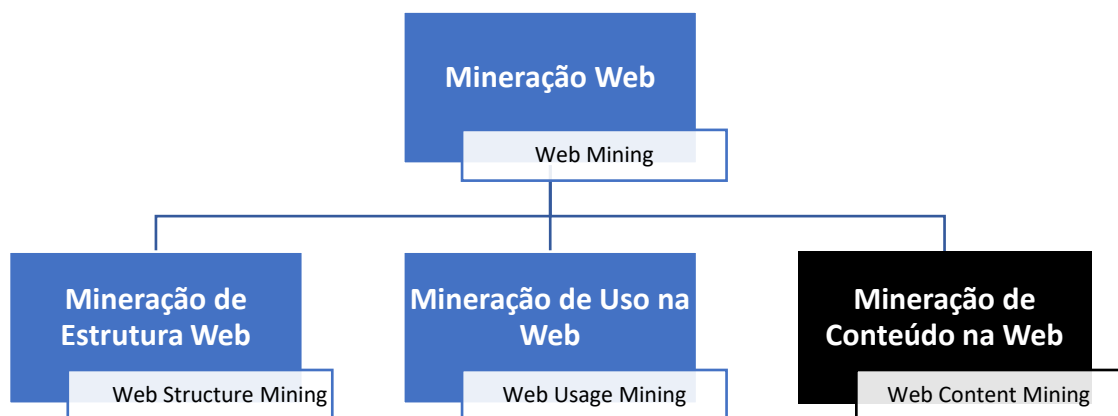


Figura 11: Mineração Web (Kosala & Blockeel, 2000)

A Mineração de Estrutura Web é a linha de pesquisa inspirada no estudo das redes sociais e de análise de citações e está interessada na estrutura dos hiperlinks dentro da *Web*. Já a Mineração de Uso na *Web* se concentra em técnicas que podem prever o comportamento do usuário enquanto o usuário interage com a *Web* (Kosala & Blockeel, 2000).

Apesar da aplicação também das técnicas das duas linhas de mineração anteriormente citadas, o trabalho se concentrou na Mineração de Conteúdo na *web* já que a busca foi realizada com o objetivo de extrair o texto das páginas (conteúdo da *Web*). Ou seja, textos no formato com pouca ou sem estrutura: HTML. A técnica de Mineração de Conteúdo na *web* procura descobrir informações úteis de conteúdo, dados e documentos ali, por meio de busca programática.

4.3 Passo 3: Instalação e configuração da ferramenta KNIME

Após realizado o download do Instalador do KNIME Analytics Platform versão 3.5.2 para Windows 64 Bits a partir do *site* oficial da aplicação (KNIME, 2017), foi realizada a instalação padrão incluindo todas as extensões. Não foi necessária nenhuma configuração extra ou programação adicional.

4.4 Passo 4: Criação e configuração do fluxo de dados

A coleta dos dados exigiu a divisão do trabalho em três fluxos de trabalho:

- Site da SciELO – Versão em Português
- Site da SciELO – Versão em Inglês
- Site da Oxford University – Versão Única em Inglês

Independentemente do site raspado, a construção do fluxo seguiu um modelo único onde iniciava-se pela página inicial e seguiu com uma navegação pelos menus e submenus até atingir a página de instruções aos autores. Cada página visitada fornecia o próximo passo da busca e, por fim, a extração em si. Nas próximas seções, serão detalhados os fluxos de trabalhos criados.

4.4.1 Scraping no Site da SciELO

A raspagem se inicia a partir da definição do site (Node 1 – *Table Creator*) inserindo o URL⁶ do site escolhido:

- Português: <http://www.scielo.org/php/index.php?lang=pt>.
- Inglês: <http://www.scielo.org/php/index.php?lang=en>

Como os site da SciELO permite a navegação em Português e em Inglês, a raspagem foi realizada em dois fluxos independentes e respectivo ao idioma. A definição do idioma é realizada por meio de *cookie*⁷ e essa foi a diferença básica entre os fluxos para o SciELO.

A utilização do *cookie* se fez necessária para que todo o fluxo se dê no respectivo idioma.

Em seguida, foram se incluindo sequencialmente nodes que liam a página inicial, mapeavam, elencavam e armazenavam os links para as páginas subsequentes, até a página final de interesse a ser coletada: A página de instrução aos autores.

O processo se encerra com a geração dos arquivos de texto em um diretório do computador (Node 40 – *StringCell to File*).

4.4.2 Scraping no Site da Oxford University

A raspagem se inicia a partir da definição do site (Node 1 – *Table Creator*) inserindo o URL do site escolhido: <https://academic.oup.com/journals/>.

Como também descrito na seção anterior para o *site* da SciELO, para o da Oxford, foram se incluindo sequencialmente nodes que liam a página inicial, mapeavam, elencavam e armazenavam os links para as páginas subsequentes, até a página final de interesse a ser coletada: A página de instrução aos autores.

O processo se encerra com a geração dos arquivos de texto em um diretório do computador (Node 43 – *StringCell to File*).

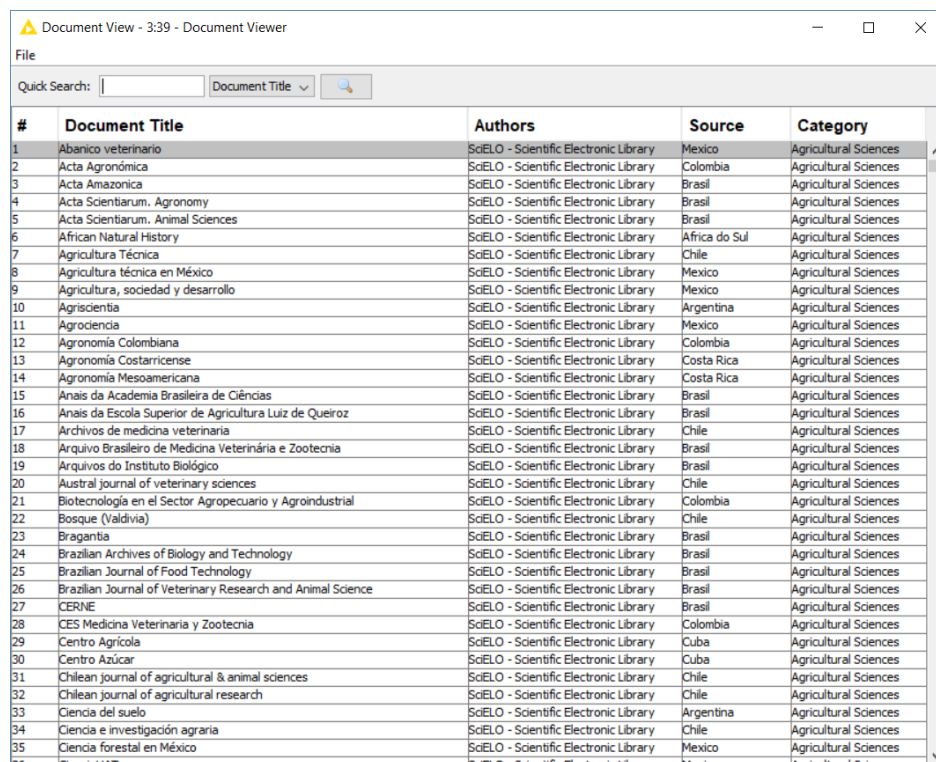
⁶ URL (*Uniform Resource Locator* - Localizador Padrão de Recursos) é o formato de atribuição universal para designar um recurso na Internet (CCM Benchmark Group, 2018).

⁷ *Cookie* é um pedaço de texto que um servidor Web pode armazenar no disco rígido do usuário. São utilizados pelos sites principalmente para identificar e armazenar informações sobre os visitantes (Martinez, 2018).

4.5 Passo 5: Execução do fluxo e armazenamento dos dados

Após a configuração dos fluxos, a rotina de raspagem foi iniciada lendo, portanto, todas as páginas listas com respostas breves e sem erros em tempo de execução.

Os dados foram armazenados nos fluxos de trabalho do KNIME em formato de documentos conforme imagens abaixo:



The screenshot shows the 'Document View' window in KNIME, titled 'Document View - 3:39 - Document Viewer'. It contains a table with the following columns: #, Document Title, Authors, Source, and Category. The table lists 36 documents, all from 'SciELO - Scientific Electronic Library', categorized under 'Agricultural Sciences'. The documents include titles in various languages, such as Spanish, Portuguese, and English, covering topics like veterinary medicine, agriculture, and food technology.

#	Document Title	Authors	Source	Category
1	Abanico veterinario	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences
2	Acta Agronómica	SciELO - Scientific Electronic Library	Colombia	Agricultural Sciences
3	Acta Amazonica	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
4	Acta Scientiarum. Agronomy	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
5	Acta Scientiarum. Animal Sciences	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
6	African Natural History	SciELO - Scientific Electronic Library	África do Sul	Agricultural Sciences
7	Agricultura Técnica	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
8	Agricultura técnica en México	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences
9	Agricultura, sociedad y desarrollo	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences
10	Agriscientia	SciELO - Scientific Electronic Library	Argentina	Agricultural Sciences
11	Agrociencia	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences
12	Agronomía Colombiana	SciELO - Scientific Electronic Library	Colombia	Agricultural Sciences
13	Agronomía Costarricense	SciELO - Scientific Electronic Library	Costa Rica	Agricultural Sciences
14	Agronomía Mesoamericana	SciELO - Scientific Electronic Library	Costa Rica	Agricultural Sciences
15	Anais da Academia Brasileira de Ciências	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
16	Anais da Escola Superior de Agricultura Luiz de Queiroz	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
17	Archivos de medicina veterinaria	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
18	Arquivo Brasileiro de Medicina Veterinária e Zootecnia	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
19	Arquivos do Instituto Biológico	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
20	Austral journal of veterinary sciences	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
21	Biotechnología en el Sector Agropecuario y Agroindustrial	SciELO - Scientific Electronic Library	Colombia	Agricultural Sciences
22	Bosque (Valdivia)	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
23	Bragantia	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
24	Brazilian Archives of Biology and Technology	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
25	Brazilian Journal of Food Technology	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
26	Brazilian Journal of Veterinary Research and Animal Science	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
27	CERNE	SciELO - Scientific Electronic Library	Brasil	Agricultural Sciences
28	CES Medicina Veterinaria y Zootecnia	SciELO - Scientific Electronic Library	Colombia	Agricultural Sciences
29	Centro Agrícola	SciELO - Scientific Electronic Library	Cuba	Agricultural Sciences
30	Centro Azúcar	SciELO - Scientific Electronic Library	Cuba	Agricultural Sciences
31	Chilean journal of agricultural & animal sciences	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
32	Chilean journal of agricultural research	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
33	Ciencia del suelo	SciELO - Scientific Electronic Library	Argentina	Agricultural Sciences
34	Ciencia e investigación agraria	SciELO - Scientific Electronic Library	Chile	Agricultural Sciences
35	Ciencia forestal en México	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences
36	Ciencia y Tecnología	SciELO - Scientific Electronic Library	Mexico	Agricultural Sciences

Figura 12: KNIME - Document View – SciELO

Os textos exibidos na tabela acima que figuram outros idiomas como, por exemplo, “*Agricultura, sociedad y desarrollo*” se referem, entre outros motivos, a apenas o nome do Periódico em si que, no caso, tem origem Mexicana, e não tem vínculo com o idioma do texto coletado. O idioma dos textos coletados tem relação, unicamente, à versão de idioma do site.

Document View - 2:39 - Document Viewer

File

Quick Search: Document Title

#	Document Title	Authors	Source	Category
1	Adaptation	Oxford University Press	Reino Unido	Arts & Humanities
2	African Affairs	Oxford University Press	Reino Unido	Arts & Humanities
3	The American Historical Review	Oxford University Press	Reino Unido	Arts & Humanities
4	American Journal of Legal History	Oxford University Press	Reino Unido	Arts & Humanities
5	American Literary History	Oxford University Press	Reino Unido	Arts & Humanities
6	Analysis	Oxford University Press	Reino Unido	Arts & Humanities
7	Applied Linguistics	Oxford University Press	Reino Unido	Arts & Humanities
8	Aristotelian Society Supplementary Volume	Oxford University Press	Reino Unido	Arts & Humanities
9	The British Journal of Aesthetics	Oxford University Press	Reino Unido	Arts & Humanities
10	The British Journal for the Philosophy of Science	Oxford University Press	Reino Unido	Arts & Humanities
11	The Cambridge Quarterly	Oxford University Press	Reino Unido	Arts & Humanities
12	Christian bioethics: Non-Ecumenical Studies in Medical Morality	Oxford University Press	Reino Unido	Arts & Humanities
13	Classical Receptions Journal	Oxford University Press	Reino Unido	Arts & Humanities
14	Contemporary Women's Writing	Oxford University Press	Reino Unido	Arts & Humanities
15	Digital Scholarship in the Humanities	Oxford University Press	Reino Unido	Arts & Humanities
16	Diplomatic History	Oxford University Press	Reino Unido	Arts & Humanities
17	Early Music	Oxford University Press	Reino Unido	Arts & Humanities
18	ELT Journal	Oxford University Press	Reino Unido	Arts & Humanities
19	English: Journal of the English Association	Oxford University Press	Reino Unido	Arts & Humanities
20	The English Historical Review	Oxford University Press	Reino Unido	Arts & Humanities
21	Environmental History	Oxford University Press	Reino Unido	Arts & Humanities
22	Essays in Criticism	Oxford University Press	Reino Unido	Arts & Humanities
23	European Review of Economic History	Oxford University Press	Reino Unido	Arts & Humanities
24	Forum for Modern Language Studies	Oxford University Press	Reino Unido	Arts & Humanities
25	French History	Oxford University Press	Reino Unido	Arts & Humanities
26	French Studies	Oxford University Press	Reino Unido	Arts & Humanities
27	French Studies Bulletin	Oxford University Press	Reino Unido	Arts & Humanities
28	German History	Oxford University Press	Reino Unido	Arts & Humanities
29	History Workshop Journal	Oxford University Press	Reino Unido	Arts & Humanities
30	Holocaust and Genocide Studies	Oxford University Press	Reino Unido	Arts & Humanities
31	Industrial and Corporate Change	Oxford University Press	Reino Unido	Arts & Humanities
32	ISLE: Interdisciplinary Studies in Literature and Environment	Oxford University Press	Reino Unido	Arts & Humanities
33	International Journal of Lexicography	Oxford University Press	Reino Unido	Arts & Humanities
34	Journal of the American Academy of Religion	Oxford University Press	Reino Unido	Arts & Humanities
35	Journal of American History	Oxford University Press	Reino Unido	Arts & Humanities
36	Journal of Church and State	Oxford University Press	Reino Unido	Arts & Humanities

Figura 13: KNIME - Document View – Oxford University Press

Os documentos foram exportados em arquivo texto para disponibilização aos interessados:

03 Saída - SciELO - Português

03 Saída - SciELO - Português

Arquivo Início Compartilhar Exibir

← → ↑ ↓ Este Computador > DATA (D:) > Marcelo > Google Drive > 01 Marcelo > 06 Projetos > 07 Data Driven Analysis - Modais > 03 Desenvolvimento > 01 Arquivos > 03 Saída - SciELO - Português

	Nome	Data de modificaç...	Tipo	Tamanho
Acesso rápido	Abanico veterinario.txt	17/09/2017 14:12	Documento de Te...	1 KB
OneDrive	ABCD, Arquivos Brasileiros de Cirurgia DL...	17/09/2017 14:12	Documento de Te...	8 KB
Este Computador	Acción Psicológica.txt	17/09/2017 14:12	Documento de Te...	1 KB
	ACIMED.txt	17/09/2017 14:12	Documento de Te...	3 KB
	Acta Agronómica.txt	17/09/2017 14:12	Documento de Te...	11 KB
	Acta Amazonica.txt	17/09/2017 14:12	Documento de Te...	4 KB
	Acta bioethica.txt	17/09/2017 14:12	Documento de Te...	13 KB
	Acta Biológica Colombiana.txt	17/09/2017 14:12	Documento de Te...	7 KB
	Acta bioquímica clínica latinoamericana...	17/09/2017 14:12	Documento de Te...	26 KB
	Acta Botanica Brasileira.txt	17/09/2017 14:12	Documento de Te...	1 KB

Figura 14: Saída Arquivos de Texto - SciELO - Português

\02 Saída - SciELO - Inglês

Nome	Data de modificação	Tipo	Tamanho
Abanico veterinario.txt	17/09/2017 13:16	Documento de Te...	1 KB
ABCD. Arquivos Brasileiros de Cirurgia Di...	17/09/2017 13:16	Documento de Te...	8 KB
Acción Psicológica.txt	17/09/2017 13:16	Documento de Te...	5 KB
ACIMED.txt	17/09/2017 13:16	Documento de Te...	3 KB
Acta Agronómica.txt	17/09/2017 13:16	Documento de Te...	11 KB
Acta Amazonica.txt	17/09/2017 13:16	Documento de Te...	4 KB
Acta bioethica.txt	17/09/2017 13:16	Documento de Te...	11 KB
Acta Biológica Colombiana.txt	17/09/2017 13:16	Documento de Te...	33 KB
Acta bioquímica clínica latinoamericana...	17/09/2017 13:16	Documento de Te...	24 KB
Acta Botanica Brasílica.txt	17/09/2017 13:16	Documento de Te...	3 KB

Figura 15: Saída Arquivos de Texto - SciELO - Inglês

\01 Saída - Oxford - Inglês

Nome	Data de modificação	Tipo	Tamanho
Acta Biochimica et Biophysica Sinica.txt	17/09/2017 13:59	Documento de Te...	25 KB
Adaptation.txt	17/09/2017 13:59	Documento de Te...	10 KB
Aesthetic Surgery Journal.txt	17/09/2017 13:59	Documento de Te...	48 KB
African Affairs.txt	17/09/2017 13:59	Documento de Te...	16 KB
Age and Ageing.txt	17/09/2017 13:59	Documento de Te...	23 KB
Alcohol and Alcoholism.txt	17/09/2017 13:59	Documento de Te...	32 KB
American Entomologist.txt	17/09/2017 13:59	Documento de Te...	1 KB
American Journal of Agricultural Econom...	17/09/2017 13:59	Documento de Te...	1 KB
American Journal of Clinical Pathology.txt	17/09/2017 13:59	Documento de Te...	20 KB
American Journal of Epidemiology.txt	17/09/2017 13:59	Documento de Te...	1 KB

Figura 16: Saída Arquivos de Texto - Oxford - Inglês

4.6 Passo 6: Validação dos dados extraídos

Para verificação do conteúdo extraído, foi gerada uma planilha eletrônica listando todos as 1.317 páginas de periódicos raspadas organizadas conforme tabelas abaixo:

Idioma Português

País	Site	Assunto	Periódico	Quantidade	Percentual
Brasil				407	100,00%
	SciELO - Scientific Electronic Library			407	100,00%
		Ciências Agrárias		47	11,55%
		Ciências Biológicas		42	10,32%
		Ciências da Saúde		115	28,26%
		Ciências Exatas e da Terra		21	5,16%
		Ciências Humanas		96	23,59%
		Ciências Sociais Aplicadas		44	10,81%
		Engenharias		26	6,39%
		Linguística, Letras e Artes		16	3,93%
Total Geral				407	100,00%

Tabela 1: Páginas raspadas – Português

Idioma Inglês

País	Site	Assunto	Periódico	Quantidade	Percentual
Brasil				408	44,84%
	SciELO - Scientific Electronic Library			408	100,00%
		Agricultural Sciences		47	11,52%
		Applied Social Sciences		44	10,78%
		Biological Sciences		42	10,29%
		Engineering		26	6,37%
		Exact and Earth Sciences		21	5,15%
		Health Sciences		115	28,19%
		Human Sciences		97	23,77%
		Literature and Arts		16	3,92%
Reino Unido				502	55,16%
	Oxford University Press			502	100,00%
		Arts & Humanities		84	16,73%
		Law		56	11,16%
		Medicine & Health		118	23,51%
		Science & Mathematics		147	29,28%
		Social Sciences		97	19,32%
Total Geral				910	100,00%

Tabela 2: Páginas raspadas – Inglês

5 APLICAÇÃO PRÁTICA NA ANÁLISE LINGUÍSTICA

Os *corpora* produzidos podem ser utilizados para diversos fins. O objetivo principal do trabalho foi disponibilizar os meios para que linguistas possam coletar e estruturar dados de grande valia para análises textuais, principalmente investigar o uso de verbos modais do inglês (*will, should, would, may, might, must*) nas normas de submissão de artigos de periódicos disponíveis na *Web*.

Uma vez que este trabalho mostra como produzir corpora em larga escala, os dados estruturados podem também ser utilizados por cientistas de dados no treinamento de modelos de classificação textual, processamento de linguagem natural (NLP), mineração automática de padrões, entre outras diversas análises.

Uma ferramenta já disponível no mercado que faz um trabalho similar é o COCA (*Corpus of Contemporary American English*) – um corpus do inglês americano de uso gratuito que possui um módulo que permite analisar um termo específico em diversos contextos. (Davies, 2017). Essa ferramenta utiliza um *corpus* próprio. A imagem abaixo mostra como o COCA possibilita a verificação dos termos que aparecem antes e depois de uma palavra escolhida. Por exemplo, o verbo *can*:

[https://corpus.byu.edu/coca/](#)

Corpus of Contemporary American English

SEARCH
FREQUENCY
CONTEXT
HELP

FIND SAMPLE: [100](#) [200](#) [500](#) [1000](#)
 PAGE: << < 1 / 1000 > >

CLICK FOR MORE CONTEXT				SHOW DUPLICATES
1	2017	ACAD	Vanderbilt Law Review	A B C and her conduct-rather than on the occurrence of harm outside of the offender's control-they can not provide adequate justification for the pro
2	2017	ACAD	Vanderbilt Law Review	A B C a victim who desires to show " mercy " to the offender, victim-facing theories can not justify differential punishment, rendering the practice cate
3	2017	ACAD	Vanderbilt Law Review	A B C influential 1974 article Harm and Punishment, is that differentiating punishment based on its results can not be justified as a matter of practice
4	2017	ACAD	Vanderbilt Law Review	A B C differential punishment in many circumstances. # Therefore, to the extent that differential punishment can be justified at all, it can only be justi
5	2017	ACAD	Vanderbilt Law Review	A B C Therefore, to the extent that differential punishment can be justified at all, it can only be justified in reference to these victim-facing justification
6	2017	ACAD	Vanderbilt Law Review	A B C cases. Because previous authors have focused solely on the question of whether differential punishment can be justified writ large, they have fa
7	2017	ACAD	Vanderbilt Law Review	A B C discussion in the academic literature of why many of the traditional theories of criminal punishment can not justify the practice of differential p
8	2017	ACAD	Vanderbilt Law Review	A B C territory, identifying three distinct categories of criminal offenses to which victim-facing justifications for punishment can not apply, even in pri
9	2017	ACAD	Vanderbilt Law Review	A B C caused any statutory harm. As we will argue in this Part, offender-facing justifications can not justify differential punishment, precisely because
10	2017	ACAD	Vanderbilt Law Review	A B C are not the first to suggest that many of the classic theories of criminal punishment can not justify the role that the results of an offender's conc
11	2017	ACAD	Vanderbilt Law Review	A B C any objections and to affirmatively make the case that offender-facing justifications do not, and can not, justify differential punishment. # A.Det
12	2017	ACAD	Vanderbilt Law Review	A B C We submit that it does not. # By punishing criminal offenders, the state can simultaneously accomplish two forms of deterrence.16 First, punishi
13	2017	ACAD	Vanderbilt Law Review	A B C generally has a neutral or even positive effect on his likelihood of **25;0:TOOLONG, we can conclude at the outset that it would be categorically, pr
14	2017	ACAD	Vanderbilt Law Review	A B C attempt. At most, then, this theory suggests that the criminal justice system can substantially deter intentional crimes without punishing attorn
15	2017	ACAD	Vanderbilt Law Review	A B C

Figura 17: COCA - Corpus of Contemporay American English

Com o *corpus* aqui produzido ou realizando a extração de outros *corpora*, o linguista poderá buscar padrões de uso dos diversos vocábulos e expressões ali presentes e, até mesmo, a não ocorrência de termos em contexto pré-definidos.

O Fluxo de Trabalho do KNIME e demais artefatos gerados estão disponíveis para *download* no endereço abaixo:

<https://github.com/marcelohonoliveira/Ciencia-de-Dados-e-Big-Data/tree/master/TCC>

README.md

TCC - Trabalho de Conclusão de Curso

Web Scraping no apoio à Análise Linguística: Construção de um Corpus Textual

RESUMO EXECUTIVO

O trabalho mostra os passos realizados no desenvolvimento de um fluxo de dados para coleta e organização de textos de páginas da web utilizados para uma análise linguística por meio da construção de um corpus . Os problemas que podem ocorrer e as respectivas soluções são apresentadas de forma detalhada.

As atividades realizadas no caso apresentado podem servir de referência para novas coletas onde as técnicas de Web Scraping se aplicam e concernentes às limitações próprias de uma coleta automática na web.

1. [Texto do TCC](#) - Relatório Técnico - Versão Final Aprovada
2. [Artefatos Gerados](#) - Fluxo de Trabalho do KNIME

Pontifícia Universidade Católica de Minas Gerais
Pós-graduação em Ciência de Dados e Big Data

Web Scraping no apoio à Análise Linguística: Construção de um Corpus Textual

Aluno: Marcelo Honório de Oliveira
Orientador: Cristiano Rodrigues de Carvalho

Selo Horezom
2018

Figura 18: Repositório GitHub - marcelohonoliveira - TCC

6 CONCLUSÕES

O desenvolvimento do presente trabalho possibilitou a prática de técnicas de *Web Scraping* envolvendo os assuntos relativos à mineração de texto da *web* para extração, organização e disponibilização de conteúdo antes disponíveis apenas nos *sites* de origem.

Os objetivos foram alcançados com êxito confirmando a aderência da ferramenta escolhida para extração com as necessidades da mineração realizada. O *corpus* construído está disponível a quem interessar em especial linguistas que desejam consumir expressões idiomáticas do inglês e respectivas tradução para o português brasileiro.

Por fim, o trabalho permitiu ao acadêmico experimentar umas das atividades inerentes à Mineração de Dados no que toca os processos de extração, transformação e carga de dados, atividades essas de extrema importância para aqueles se expõem aos desafios da Ciência de Dados.

7 TRABALHOS FUTUROS

A utilização dos *corpora* pode ser feita diretamente pelo interessado, especialmente linguistas, numa análise humana sem a utilização de ferramentas específicas. Contudo, isso é viável quando o *corpus* é pequeno e a análise é mais qualitativa do que quantitativa.

Devido à dificuldade em examinar, manualmente, dados massivos, como trabalhos futuros, pretende-se desenvolver ferramentas que ajudem a minerar padrões nos dados coletados e estruturados para a análise linguística textual. Um exemplo seria minerar sequências probabilísticas de termos onde os verbos modais foram usados nesses textos.

8 REFERÊNCIAS BIBLIOGRÁFICAS

- Alphabet Inc. (1 de Janeiro de 2018). *Chrome - Navegador*. (Google LLC) Acesso em 15 de Fevereiro de 2018, disponível em Use um navegador da Web gratuito e mais rápido: <https://www.google.com.br/chrome/>
- Andriolo, E. (09 de Abril de 2012). *Desvendando 'Data Scraping': Entenda como raspar dados pode facilitar o trabalho jornalístico*. (Texas University) Acesso em 17 de Fevereiro de 2018, disponível em Knight Center for Journalism in the Americas: <https://knightcenter.utexas.edu/pt-br/blog/00-9586-desvendando-o-data-scraping-entenda-como-raspar-dados-pode-facilitar-o-trabalho-jornali>
- Barbieri, C. (2011). *BI2 - Business Intelligence: Modelagem & Qualidade*. Rio de Janeiro: Elsevier.
- BotReports. (22 de Fevereiro de 2014). *BotReports - Updates on the latest spiders, crawlers, scrapers*. (BotReports) Acesso em 8 de Abril de 2018, disponível em <http://www.botreports.com/user-agent/008.shtml>
- CCM Benchmark Group. (8 de Abril de 2018). *O que é um URL*. Fonte: CCM Brasil: <https://br.ccm.net/contents/288-o-que-e-um-url>
- Datafiniti, LLC. (1 de Janeiro de 2018). *80legs - Easy Web Scraping Tools and Cloud-Based Web Crawling*. (Datafiniti, LLC) Acesso em 8 de Abril de 2018, disponível em <http://80legs.com/>
- Davies, M. (1 de Dezembro de 2017). *COCA*. (M. Davies, Produtor, & Brigham Young University) Acesso em 22 de Abril de 2018, disponível em Corpus of Contemporary American English: <https://corpus.byu.edu/coca/>
- Gazola, A. (19 de Janeiro de 2016). *Utilizando meta tags*. (Mozilla) Acesso em 28 de Fevereiro de 2018, disponível em https://developer.mozilla.org/pt-PT/docs/Utilizando_meta_tags
- Google. (20 de Abril de 2015). *Scraper*. (Google) Acesso em 25 de Fevereiro de 2018, disponível em <https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecacpepngjd>

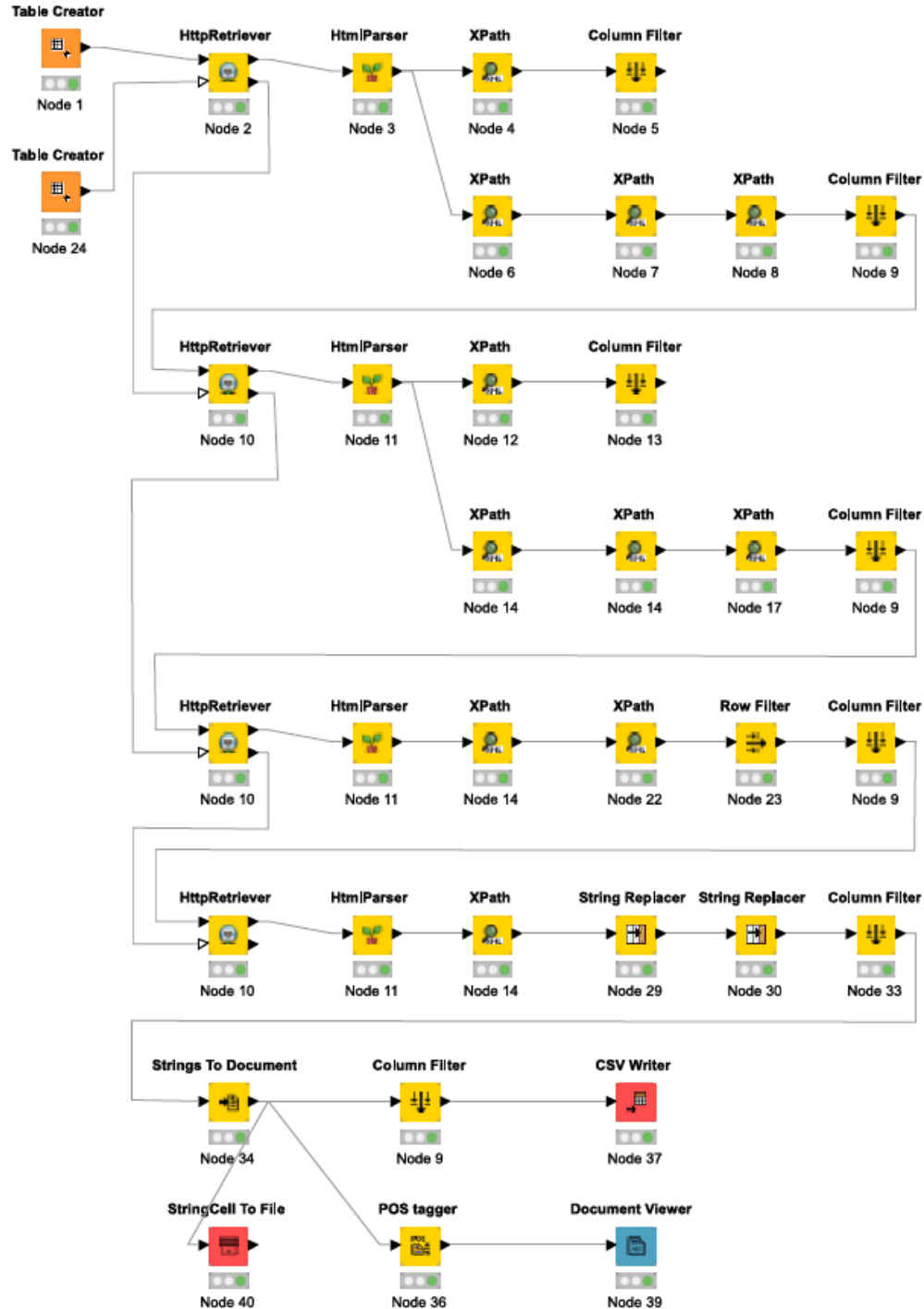
- Google Inc. (8 de Abril de 2018). *Especificações para metatags robots e cabeçalhos HTTP X-Robots-Tag*. Fonte: Google - Developers: https://developers.google.com/search/reference/robots_meta_tag?hl=pt-br
- KNIME. (1 de Janeiro de 2017). *About KNIME*. (KNIME - Open for Innovation) Acesso em 2 de Março de 2018, disponível em <https://www.knime.com/about>
- Kosala, R., & Blockeel, H. (Julho de 2000). Web Mining Research: A Survey. *ACM SigKDD Exploration*, II(1), 1-15. doi:10.1145/360402.360406
- Martinez, M. (5 de Abril de 2018). *Cookies*. Fonte: InfoEscola: <https://www.infoescola.com/informatica/cookies/>
- Microsoft Corporation. (14 de Março de 2017). *Conceitos de Mineração de Dados*. Acesso em 2018 de Abril de 2018, disponível em Microsoft Docs: <https://docs.microsoft.com/pt-br/sql/analysis-services/data-mining/data-mining-concepts>
- Oxford University. (1 de Fevereiro de 2018). *Oxford University Press - Journals*. (Oxford University Press) Acesso em 1 de Fevereiro de 2018, disponível em Oxford University Press: <https://academic.oup.com>
- Ray, S. (22 de Outubro de 2015). *Beginner's guide to Web Scraping in Python (using BeautifulSoup)*. (Analytics Vidhya) Acesso em 19 de Fevereiro de 2018, disponível em <https://www.analyticsvidhya.com/blog/2015/10/beginner-guide-web-scraping-beautiful-soup-python/>
- Russell, J., & Cohn, R. (2012). KNIME. Em *KNIME* (p. 128). Stoughton: Book on Demand Ltd.
- Sardinha, T. B. (1999). Usando WordSmith Tools na investigação da linguagem. 20. (P. d.-G. Linguagem, Ed.) São Paulo, SP, Brasil: Pontifícia Universidade Católica de São Paulo. Acesso em 31 de Março de 2018, disponível em <http://www2.lael.pucsp.br/direct/DirectPapers40.pdf>
- SciELO. (1 de Fevereiro de 2018). *SciELO - Scientific Electronic Library*. Acesso em 1 de Fevereiro de 2018, disponível em SciELO - Scientific Electronic Library: <http://www.scielo.org>

W3C. (18 de Abril de 2018). *W3C Standards*. Fonte: W3C: <https://www.w3.org/TR/xpath/>

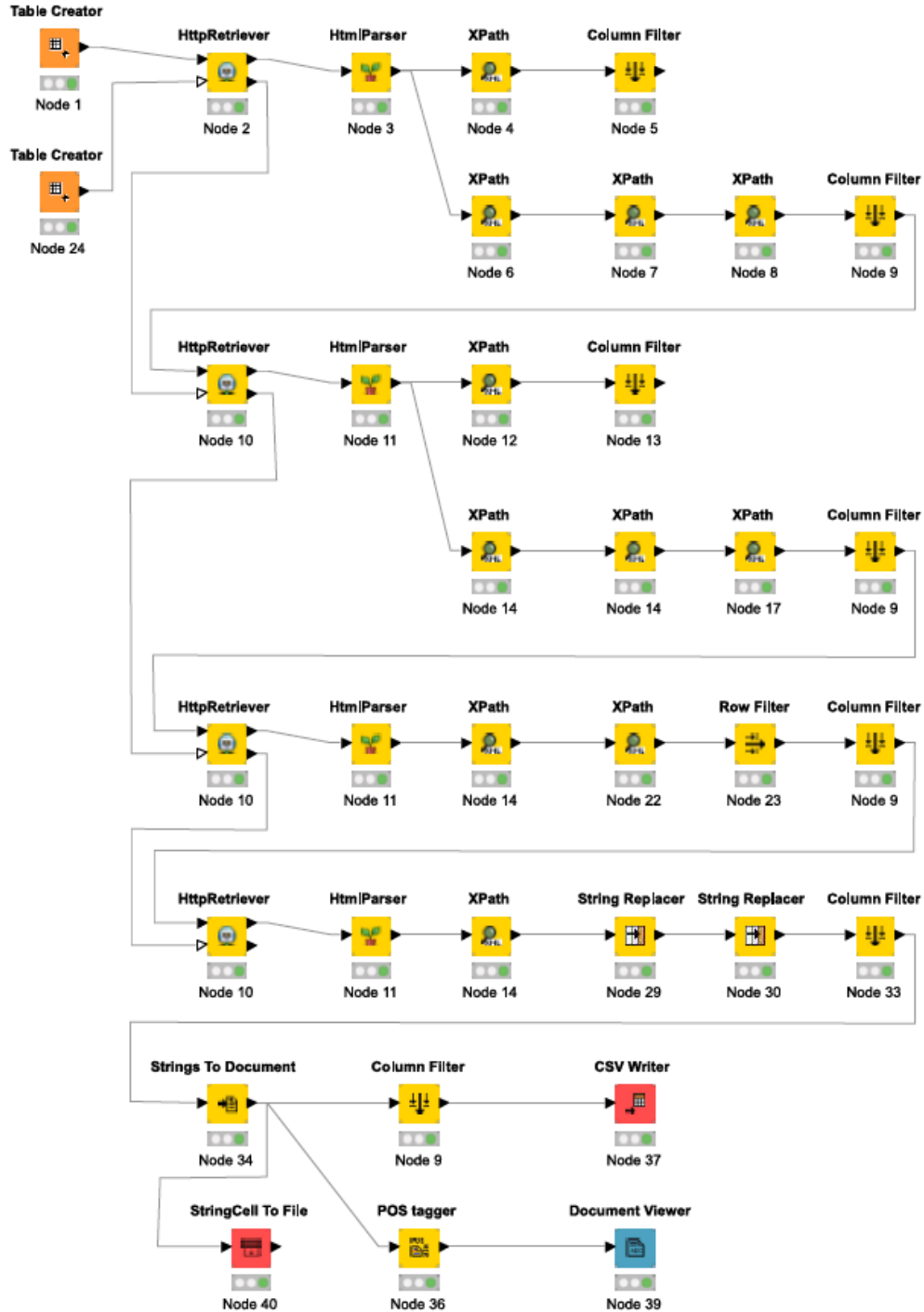
Wikipédia. (5 de Agosto de 2017). *Protocolo de Exclusão de Robôs*. (Wikipédia) Acesso em 1 de Março de 2018, disponível em Wikipédia - A enciclopédia livre: https://pt.wikipedia.org/wiki/Protocolo_de_Exclus%C3%A3o_de_Rob%C3%B4s

Wynne, M. (2005). *Developing linguistic corpora: A guide to good practice*. Oxford: AHDS literature, languages and linguistics.

ANEXO I – Data Driven Analysis – Modals – SciELO – Português



ANEXO II – Data Driven Analysis – Modals – SciELO – Inglês



ANEXO III – Data Driven Analysis – Modals – Oxford – Inglês

