Indexação Automática de Conteúdo: uma análise comparativa das estratégias de indexação para gestão de informação

Thainara Barbosa da Silva[1], Carlos Eduardo Pagani[1]
[1]Instituto Federal de Educação, Ciência e Tecnologia de São Paulo (IFSP) – Campus Hortolândia

t.barbosa@aluno.ifsp.edu.br, pagani@ifsp.edu.br

Abstract. *This work is a case study that aims to identify which strategy is best applied, based on qualitative requirements, to manage and retrieve information in extensive databases, focused on building data products in a Web 4.0 and Big Data scenario. Thus, two techniques of Automatic Content Indexing were compared and analyzed from the construction of workflows and application of the techniques in the FAPESP database using the Orange Data Mining tool.*

Resumo. *Esse trabalho trata-se de um estudo de caso que tem como propósito identificar qual estratégia melhor se aplica, a partir de critérios qualitativos, para gerir e recuperar informações em bases de dados extensas, com foco na construção de produtos de dados em um cenário de Web 4.0 e Big Data. Dessa forma, foram comparadas e analisadas duas técnicas de Indexação Automática de Conteúdo a partir da construção de fluxos de trabalhos e aplicação das técnicas em uma base de dados da FAPESP na ferramenta Orange Data Mining.*

1. Introdução

O crescente volume, variedade e velocidade que os dados são gerados na atualidade, *Big Data* (Gartner Group, 2021), é consequência do comportamento dos usuários na Web, os quais, segundo Content (2019), deixam de ser passivos para serem ativos, compartilhando informações e gerando dados a cada instante (Domo, 2021). Este cenário representa a era conhecida como Web 4.0, que para as organizações gerenciarem esses ativos e promoverem a recuperação de informações para a construção de um produto, pode ser uma tarefa demorada. O gerenciamento eficiente das informações para transformá-las em conhecimento, está diretamente relacionada com a forma que os dados são capturados, armazenados e transformados em informações, ou seja, quanto maior a quantidade de informações, maior a necessidade de gerenciamento (Bianque, 2015).

Para Bianque (2015) e Jacobino Et. Al. (2017) a informação é um componente indispensável em uma organização para tomada de decisões estratégicas e para que seja útil e relevante precisa ser acessível e disponível. Já para Almeida e Alves (2020) a informação é qualquer coisa que é de importância na resposta a uma questão.

Assim, pode-se dizer que, a informação é um componente indispensável para as organizações, as quais não são capazes de perceber a importância deste ativo e de suas tecnologias, não compreendendo os processos pelos quais a informação se transforma em conhecimento (Jacobino Et. Al., 2017). Por esta razão, as informações precisam ser recuperadas de forma rápida e corretamente quando requeridas pelo usuário.

Dessa forma, pensando em melhorar a maneira como as informações são recuperadas em bases de dados e transformadas em produtos, o presente trabalho tem como objetivo analisar, e identificar, através de um estudo de caso neste trabalho e em artigos e tutoriais técnicos, a melhor estratégia de Indexação Automática de Conteúdo para proporcionar às organizações uma forma de recuperar informações em ambiente Web 4.0 (Content, 2019).

Este trabalho trata-se de um estudo de caso e uma revisão de artigos com a finalidade de identificar a viabilidade da aplicação de uma estratégia de recuperação de informação em base de dados organizacionais. Dessa forma comparar as técnicas de indexação automática de conteúdo por extração e atribuição, para identificar a que melhor se aplica a esse estudo de caso. Visando garantir a capacidade de pesquisar e recuperar as informações contidas nas bases de dados. Assim, melhorar a eficiência da organização, permitindo que seus funcionários pesquisem informações sem a necessidade de explorar bases de dados. A comparação das estratégias será feita através dos fluxos de trabalho construídos na ferramenta Orange Data Mining. Os critérios qualitativos utilizados na comparação das técnicas serão: a viabilidade na aplicação, cenário tecnológico e performance.

As próximas seções serão apresentadas da seguinte forma: Seção 2 Fundamentação teórica; Seção 3 Materiais e métodos; Seção 4 Desenvolvimento; Seção 5 Discussão e resultados; Seção 6 Considerações finais; Referências e Apêndices.

2. Fundamentação teórica

2.1. Big Data

Big Data é um termo que significa grande volume de dados, essa definição foi originada nos anos 2000 por um analista do Gartner Group (Machado, 2018). Ainda de acordo com Gartner Group (2021), *Big Data* se refere aos dados gerados em grande volume, variedade e velocidade que necessitam de maneiras inovadoras e econômicas de processamento de informações para que seja possível ter uma visão aprimorada dos ativos de informação, tomada de decisões orientadas a dados e automação de processos.

O *Big Data* está relacionado ao grande volume de dados que ganha relevância conforme a sociedade se depara com um aumento sem precedentes no número de informações geradas a cada dia (IBM, 2014).

Em concordância com as definições citadas, é possível dizer que o *Big Data* é o crescimento exponencial dos dados, sua utilização e armazenamento em grande volume, variedade que desafiam o método convencional de gerenciamento e análise de dados e a forma que é transformada em conhecimento perante as organizações (Rêgo, 2013, p. 20).

2.2. A evolução da Web 1.0 para a Web 4.0

Há quinhentos anos, algumas descobertas contribuíram para o salto de desenvolvimento intelectual da humanidade, partindo do desenvolvimento da tipografia de Guttemberg no ano de 1436, a invenção da máquina a vapor em 1784, a eletricidade por Thomas Edison em 1879 e, outras descobertas e invenções que melhoraram a vida de milhares de pessoas no mundo inteiro. Porém, a descoberta do século XX a *"World Wide Web"* (www) proporcionou a

conectividade entre as pessoas (Barros e Caiado, 2017).

Segundo Barros e Caiado (2017) o termo Web muitas vezes é relacionado à internet, mas não é a mesma coisa. A Web é a ferramenta que proporciona que as pessoas acessem a internet por meio de navegadores (*browsers*) criada por Tim Berners-Lee. Essa ferramenta evoluiu com o passar dos anos, sendo aprimorada e está em constante transformação. Para Aghaei, Nematbakhsh, Farsani (2012, p. 1) e Klein, Adolfo (2020, p. 4) houve muito progresso da Web 1.0 para a Web 4.0 nas últimas duas décadas decorrente da evolução das tecnologias e para entender sobre o que é a Web 4.0, é preciso entender o progresso da Web 1.0 para a Web 4.0.

A Web 1.0, primeira geração, considerada como a Web da cognição em que era caracterizada pela utilização da internet para leitura cujos usuários tinham uma interação limitada com a internet, assumindo uma postura passiva. Já a Web 2.0, segunda geração, foi considerada como a Web da comunicação, marcada pela revolução na indústria dos computadores e usuários que passam a ser um pouco mais expressivos na internet, passando a utilizar o meio para escrita, criando e atualizando conteúdos através de *blogs*.

Após a Web 1.0 e 2.0, a terceira geração, Web 3.0, permitiu à tecnologia vincular, integrar e analisar dados para obter novos fluxos de informações, sendo assim, considerada como a Web da cooperação. Segundo Content (2018) os avanços tecnológicos os usuários da internet deixam de ser passivos para serem mais ativos na Web, compartilhando conteúdos e gerando dados a cada minuto (Domo, 2020).

Esse perfil de usuário representa a quarta geração, ou Web 4.0, a Web da integração, caracterizada pela geração que avança no uso da Inteligência Artificial, Big Data para interpretar grande volume de dados (Almeida, 2017).

2.3. Dado, informação e conhecimento

Dado, informação e conhecimento possuem diferentes definições e o momento de transição entre eles é fundamental para a construção de uma inteligência coletiva. Em um ambiente organizacional, dados são registros estruturados e não estruturados de transações. Estes tornam-se informações quando os seus registros são processados e aplicados a um contexto (Zeferino, 2020).

Dessa forma, é possível exemplificar a transição dos dados para informação dizendo que os dados são observações documentadas, ou resultado de medições que acontecem no contexto, por exemplo, de uma pessoa almoçando em um restaurante, essa medição isolada não representa muita coisa, mas se avaliarmos dez pessoas no mesmo restaurante, estas dez coletam dados e ao agrupá-los, é possível avaliar se existe padrões como, o prato preferido, o valor gasto ou, até o mesmo, o tempo ocupado no restaurante, esses dados trabalhados são as informações.

Entretanto, a informação por si só não é o conhecimento, pois este está relacionado com a forma que a informação é captada, assimilação, associação e a construção, desconstrução e reconstrução de conceitos pela mente humana (Borges e Rhaddour, 2017).

Borges e Rhaddour (2017) acrescentam que o significado da palavra “conhecimento” se

refere ao conhecimento acumulado ao longo do tempo e da sua socialização e compartilhamento, facilitados pelo contexto atual da internet.

2.4. A Inteligência Coletiva e Gestão do Conhecimento

A inteligência coletiva trata-se do compartilhamento de informações entre um grupo de pessoas, o qual é capaz de resolver mais problemas do que apenas um indivíduo (Padilha e Graeml, 2019). Lévy (2015) acrescenta que a inteligência está distribuída e, quando valorizada, coordenada em tempo real, resulta em uma consolidação efetiva de habilidades, sendo a base da inteligência coletiva o enriquecimento mútuo das pessoas.

A inteligência coletiva passou a ser reconhecida na década de 1990 quando a gestão do conhecimento começou a ser utilizada como complemento para melhorar a gestão da informação, partindo da compreensão da diferença entre o significado de dado e informação, os quais passaram a serem compreendidos como base para o conhecimento, fundamentando-se na organização dos dados e informações para a interpretação de eventos (Padilha e Graeml, 2019).

Padilha e Graeml (2019) ainda afirmam que a inteligência coletiva também está atrelada a forma que um grupo de pessoas evoluem a partir da resolução de problemas e integração colaborativa, com base no "reconhecimento e enriquecimento mútuo" dos envolvidos em um grupo em que cada integrante "contribui com o que sabe para a construção do conhecimento coletivo".

Em ambientes digitais a inteligência coletiva pode ser utilizada como uma forma de melhorar processos simples, porém, difíceis de automatizar, bem como na coleta de informação sobre dados (Svobodová; Koudelková, 2011). Maries e Scarlat (2011) complementam que à medida que a sociedade evolui com as tecnologias da informação, a comunicação possibilita a "coordenação de pensamentos e ações entre um número cada vez maior de indivíduos", assim, possibilitando a gestão do conhecimento coletivo.

2.5. Recuperação de Informação por Indexação Automática de Conteúdo

A recuperação de informação tem como finalidade encontrar uma informação específica dentre um grande volume de dados que podem ser dos mais variados tipos, estruturados, semiestruturados ou, até mesmo, não estruturados (Muller Et. Al., 2015). Neste contexto existem algumas técnicas de recuperação de informação que facilitam e otimizam o tempo de busca por uma informação específica, sendo uma dessas técnicas a Indexação Automática de Conteúdo.

A técnica de indexação automática utiliza a estratégia de representar um conteúdo por meio da sintetização de texto, ou seja, realiza o resumo de um texto ressaltando os termos que são mais importantes, servindo como ponto de acesso a localização de informações em um sistema (Silva e Corrêa, 2015).

Silva e Correa (2020) corrobora afirmando que:

> "A organização e recuperação da informação se materializam pela indexação, que por sua vez é realizada com a finalidade de determinar, por meio do

> conteúdo dos documentos, um conjunto de palavras-chave ou assuntos, facilitando sua armazenagem em bases de dados e atendendo deste modo, as necessidades de recuperação da informação (Fujita; Gil-Leiva, 2010).”

Ainda de acordo com Silva e Correa (2020) essa técnica de recuperação de informação se dá através da seleção automática de palavras por meio de um software ou sistema, que leva em consideração estatísticas e ocorrência das palavras da base.

Segundo Silva e Correa (2020), existem dois tipos de indexação automática de conteúdo: por extração e por atribuição. A técnica que realiza a extração, a seleção dos termos fica restrito ao contexto do próprio documento, o indexador utiliza critérios institucionais e pessoais, selecionando no texto palavras que serão utilizadas para representar o documento. Já a técnica por atribuição, utiliza-se de um elemento externo ao documento, um conjunto de termos previamente definidos e normalizados (léxico), cuja complexidade pode variar desde uma lista de cabeçalhos de assunto até uma ontologia. Após a leitura do texto, o indexador escolhe os termos mais adequados para representar o conteúdo do documento.

2.6. Sumarização de textos

A sumarização de textos é uma técnica que consiste na realização de um resumo de um texto, sintetizando o conteúdo, diferente de uma narração que é detalhada. Também pode ser definida como a tarefa de realizar uma síntese concisa e fluente, conservando as informações e significados gerais, conforme o texto original (Dubey, 2018).

Esta técnica pode ocorrer de duas formas, computacionalmente explicando, sumarização por abstração e a sumarização por extração. Na primeira, o texto é analisado com base em sua semântica e relação entre as sentenças (frases), gerando um sumário reformulado do texto. Já na segunda técnica, o sumário gerado é baseado em frases selecionadas por meio de métodos estatísticos, eliminando termos que não agregam informações, atribuindo pesos às sentenças (Goularte, Bif Et. Al., 2014).

3. Materiais e métodos

Este estudo trata-se de um estudo de caso e uma revisão de artigos com a finalidade de identificar a viabilidade da aplicação de uma estratégia de recuperação de informação em base de dados organizacionais. Dessa forma comparar as técnicas e identificar a estratégia que melhor se aplica a esse estudo. Para isso, o estudo foi dividido nas seguintes etapas: estudo literário, sumarização de conteúdo, criação de base de dados, aplicação na *Orange Data Mining*, construção dos fluxos de Indexação Automática de Conteúdo e análise e interpretação das técnicas.

3.1 Materiais

A linguagem de programação *Python* foi utilizada na implementação do algoritmo de *Textrank*, juntamente com a biblioteca NLTK, NumPY e NetworkX. O recurso NLTK é uma biblioteca que estabelece infraestrutura necessária para trabalhar com linguagem natural para a construção de programas na linguagem Python como classificação textual, análise sintática e classificação gramatical. A biblioteca NumPY permite a utilização de matrizes multidimensionais e álgebra linear para tarefas de probabilidade. Já a biblioteca NetworkX é

utilizada para armazenar e manipular estruturas de redes, permitindo assim, a visualização de redes semânticas.

O algoritmo *Textrank* foi disponibilizado pelo autor Praveen Dubey, autor do artigo "Understand Text Summarization and create your own summarizer in python" (Dubey, 2019). O qual disponibilizou um algoritmo em uma versão para a implementação através das execuções de linha de comando.

Além do algoritmo de sumarização de textos, serviram de base para este estudo artigos científicos da FAPESP, bibliografias e a ferramenta *Orange Data Mining* para análise comparativa de dados e técnicas de indexação automática de conteúdo. O *Orange Data Mining* é uma ferramenta de código aberto que possibilita criar fluxos de trabalho de um projeto sem a necessidade de codificar através de sua interface gráfica que permite inserir, manipular, classificar e extrair dados.

3.2 Métodos

Primeiramente foi realizada uma revisão de artigos para as técnicas de indexação automática de conteúdo por extração e atribuição, Web 4.0, *Big Data*, Inteligência Coletiva e Gestão da Informação. Posteriormente, foi utilizado o algoritmo *Textrank* para sumarizar textos, ou seja, resumir textos com base na sua semântica e citações de outros documentos. Os artigos científicos utilizados para criar a base de dados foram retirados da plataforma da revista científica FAPESP. A partir disso, a base de dados foi utilizada com os textos sumarizados em que o resumo é definido como conciso e fluente, preservando as principais informações e o significado geral (Dubey, 2019).

Na sequência foi aplicada a base sumarizada na ferramenta *Orange Data Mining* para uma comparação entre as técnicas de indexação automática de conteúdo, as quais foram aplicadas por cada algoritmo por meio da construção de fluxos comparativos de resultados e dados sobre cada técnica, com o objetivo de visualizar as principais diferenças entre elas.

Para este estudo foi escolhida a ferramenta *Orange Data Mining* que possibilita criar o fluxo de trabalho de um projeto. Este *software* é próprio para aprendizado de máquina, visualização e análise de dados, além de ser gratuito e sem a necessidade de precisar programar (Batista, 2019).

Por fim, foram analisadas as duas técnicas de indexação automática de conteúdo observando dois fluxos de dados. O primeiro aplicando a técnica de indexação automática de conteúdo por extração, extraindo palavras que não agregam ao conteúdo como pronomes, artigos, preposições com o objetivo de apresentar informações.

O processo de Indexação Automática de Conteúdo para este estudo contém cinco etapas:

1. Sumarização de textos;
2. Formação da base de dados;
3. Pré-processamento dos textos;
4. Indexação de conteúdo;
5. Análise dos dados.

4. Desenvolvimento

Através de estudo de caso neste trabalho e em artigos e tutoriais técnicos, foram analisadas as características descritivas e a aplicação prática das técnicas aplicadas no software *Orange Data Mining* e, assim, identificar a melhor estratégia para gerenciar e recuperar informações de bases de dados dos dados da organização.

Com esse objetivo, o estudo apresenta a hipótese de que a indexação automática de conteúdo é uma estratégia que traz rapidez e eficiência para recuperar informações em bases de dados, abaixo será apresentada a comparação das duas técnicas: indexação automática de conteúdo por extração e indexação automática por atribuição e o processo de sumarização de textos.

Este processo tem a finalidade de resumir os artigos e gerar uma base de dados, preservando as informações originais e o significado geral do texto para assim aplicar as técnicas de indexação automática, foi utilizado o processo de sumarização de textos (figura 1).

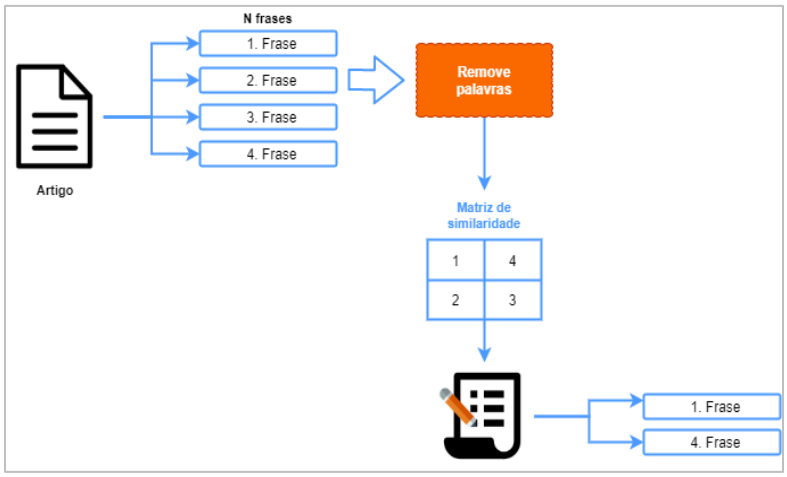

Figura 1. Processo de sumarização de textos

Fonte: Adaptação Towards Data Science, 2019.

Um trecho do código utilizado na sumarização de texto (apêndice A) exemplificado pelo artigo do *Towards Data Science*, traduzida as etapas e detalhado na Tabela 1.

Tabela 1: Etapas da sumarização de conteúdo

Etapa	Descrição	Detalhe
1	Ler o texto e realizar a tokenização	Nesta etapa é utilizado o método de tokenização, que separa o texto em componentes menores como palavras e sentenças .

2	Gerar matrizes de similaridades	Nesta etapa é gerada uma matriz similaridade, uma matriz de pontuação que apresenta a similaridade entre os artigos. A matriz de similaridade mede as semelhanças entre os pares de artigos, quanto maior a similaridade, maior o valor da medida.
3	Classificar os *tokens* em matrizes de similaridades	Nesta etapa são classificados os *tokens* em uma matriz de similaridade para comparar as semelhanças dos artigos.
4	Ordenar os *tokens* e seleção dos melhores	Nesta etapa os *tokens* são ordenados em um *ranking* e selecionados os melhores, constituindo um texto sumarizado.
5	Imprimir o resultado na tela	Nesta etapa é exibido na tela o texto sumarizado com base nos *tokens* formados, ordenados e selecionados.

Em resumo, a sumarização acontece a partir da criação de uma base de dados (*Corpus*), que se trata de uma coleção de textos que representam uma linguagem ou um conjunto de linguagens naturais (Matos Et. Al, 2019). Após a sumarização de textos e a criação da base é realizado o pré-processamento dos textos (tokenização e remoção de *stopwords*) em que obtém uma representação estruturada dos artigos.

4.1 Fluxos de indexação automática de conteúdo

A indexação automática de conteúdo tem como objetivo possibilitar o acesso rápido à informação, por meio da busca por palavras e, assim, recuperar informações. Sobre o resultado das etapas anteriores foram aplicadas as técnicas de indexação automática de conteúdo. Por fim, a análise, leitura e a interpretação dos dados.

4.1.1 Indexação automática por extração

A indexação automática de conteúdo por extração (figura 3) representa a recuperação de conteúdo por meio da técnica de extração aplicada a um vocabulário controlado, onde são extraídos os termos que aparecem com frequência.

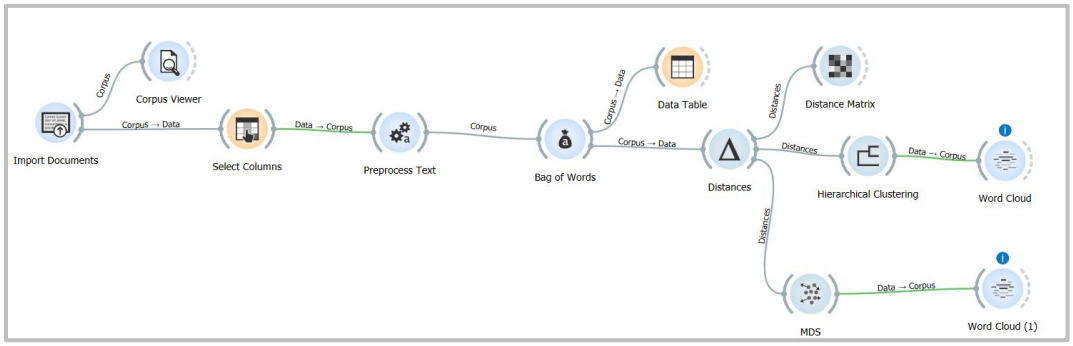

Figura 3. Indexação automática por extração

4.1.2 Indexação automática por atribuição

A indexação automática de conteúdo por atribuição (figura 4) tem o mesmo objetivo da primeira técnica, o diferencial é que visa representar o conteúdo do documento através de termos autorizados de um vocabulário controlado, ou seja, a partir da atribuição de termos, busca na base de dados os termos relacionados, apresentando as informações.

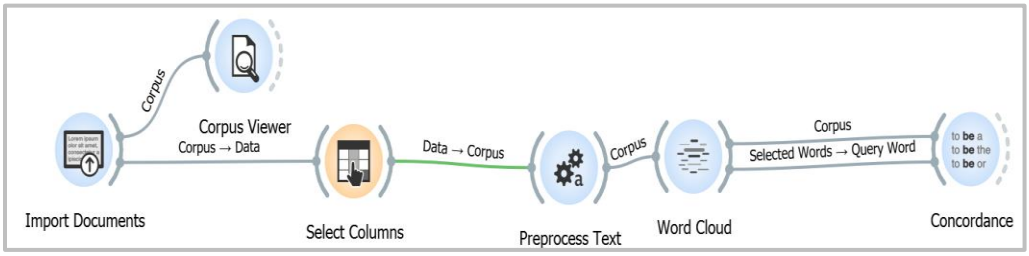

Figura 4. Indexação automática por atribuição

5. Discussão e resultados

A partir do estudo de caso, análise e comparação das técnicas de indexação automática de conteúdo, aplicado a base de dados de artigos científicos da FAPESP e na construção de fluxos de trabalho na ferramenta *Orange Data Mining*, foi possível identificar as diferenças entre as estratégias (tabela 2).

Tabela 2: Comparação das técnicas de Indexação Automática de Conteúdo

Indexação Automática por Extração	**Indexação Automática por Atribuição**
Utiliza um índice com termos isolados não contemplando palavras compostas;	Utiliza uma linguagem artificial (codificada);
Os termos são extraídos e traduzidos em "termos autorizados de um vocabulário controlado";	São extraídas as palavras pertinentes à representação do conteúdo;
Termos que aparecem com frequência no texto ou na base de dados;	Atribuição de termos ao documento a partir de um outro documento;
Utiliza a linguagem natural em que o vocabulário é extraído de forma automática;	Os termos podem ser originários da cabeça do indexador;
	Utiliza a linguagem artificial em que o vocabulário é indexado;
	Classificação de termos e alocação de um perfil;

Além da identificação das diferenças entre as técnicas é possível enfatizar que para

realizar a recuperação e a gestão da informação, é necessário entender sobre os dados da organização e o cenário da tecnologia. As organizações também precisam disseminar o conhecimento sobre os dados para gerar a inteligência coletiva.

De acordo com Fujita (2010) a organização da informação compreende as atividades e operações de tratamento da informação que abrange conhecimento teórico e metodológico, tanto para o tratamento descritivo quanto para o tratamento temático de conteúdo das informações. Assim, a recuperação consiste na localização de documentos inseridos em bases de dados e requer técnicas e métodos da organização da informação a fim de satisfazer as necessidades de busca.

Para a realizar a recuperação e a gestão da informação, é preciso entender sobre os dados da organização e o cenário da tecnologia, além disso as organizações precisam disseminar o conhecimento sobre os dados para gerar inteligência coletiva. Devido a isso, a técnica de indexação automática de conteúdo por extração (figura 5), é considerada a mais adequada para aplicação em bases de dados extensas, pois a extração e categorização da informação é automática, de acordo com o estudo de caso, artigos e tutoriais técnicos.

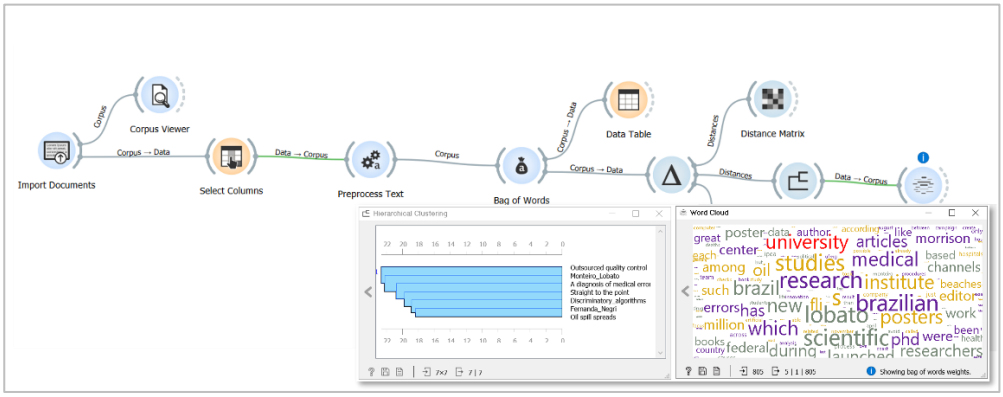

Figura 5. Resultado da técnica de indexação automática de conteúdo por extração.

Já a técnica de indexação automática de conteúdo por atribuição (figura 6), apesar de ser a mais prática na construção, se torna mais demorada, pois, é preciso indexar termos e classificá-los antes de retornar informações.

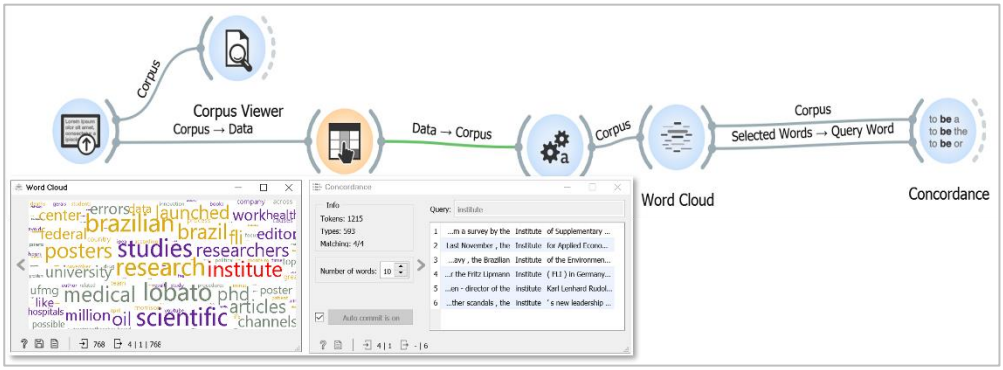

Figura 6. Resultado da técnica de indexação automática de conteúdo por atribuição.

Dessa forma, com base no estudo de caso, a indexação automática por extração apresentou a melhor performance na aplicação na base de dados construída a partir dos artigos científicos da FAPESP, demonstrando rapidez e eficiência para recuperar informações com base nos termos frequentes e classificação em grupos (*clusters*) que correlacionam termos e artigos.

6. Considerações finais e trabalho futuro

Com o advento tecnológico, os usuários da Web tornaram-se mais ativos, compartilhando informações e gerando dados a cada instante. Como consequência, as organizações passaram a lidar com o grande fluxo de informações em suas bases de dados e a dificuldade de gerir e extrair informações necessárias para a construção de produtos de dados.

A partir desse cenário, este trabalho realizou um estudo de caso e uma revisão de artigos referente a duas estratégias de recuperação de informações. Dessa forma, foi comprovado que a técnicas de indexação automática de conteúdo são eficientes para aplicar em bases de dados extensas e recuperar informações com qualidade. E, a melhor técnica para aplicar nesses tipos de bases, conforme os critérios qualitativos apresentados no estudo de caso, é a de indexação automática por extração, a qual não necessita de um indexador para recuperar informações, sendo uma extração automática.

De acordo os exemplos apresentados, é possível criar uma ferramenta que sugere a leitura de artigos e retorna insumos correlatos para a elaboração de estudos e desenvolvimento de produtos de dados.

Como trabalho futuro, serão necessários: I) a definição de métricas qualitativas para mensurar a eficiência de cada fluxo de trabalho elaborado (extração e atribuição); II) a inclusão de critério de qualidade para cada estratégia de recuperação de informação; III) a inclusão de possibilidades de produtos de dados no caso de uso; IV) o desenvolvido uma aplicação Web para conectar com a(s) técnica(s) de Indexação Automática de Conteúdo e promover o resultado prático.

Referências

Aghaei, S.; Nematbakhsh, M. A. e F.; H. Khosravi. Evolution of the World Wide Web: from Web 1.0 to Web 4.0. International Journal of Web & Semantic Technology (IJWesT), v. 3, n. 1, jan. 2012. DOI: 10.5121/ijwest.2012.3101. Disponível em: <http://airccse.org/journal/ijwest/papers/3112ijwest01.pdf>. Acesso em: 13, jun. de 2021.

Almeida, L. F. Concept and Dimensions of Web 4.0. International Journal of Computers & Technology, Punjab, v. 16, n. 7, nov. 2017, p. 7040-7046. Disponível em: <https://doi.org/10.24297/ijct.v16i7.6446>. Acesso em: 13, jun. de 2021.

Anderson, J. D.; Pérez-Carballo, J. (2001). The Nature of Indexing: how humans and machines analyze messages and texts for retrieval - Part I: Research, and The Nature of Human Indexing. Information Processing and Management, v.37, n.2.

Andrade, I. A.; Junior, D. W. B.; Tomaél, M. I.; Corgosinho, R. J. M. (2011). Inteligência Coletiva e Ferramentas WEB 2.0: a busca da gestão da informação e do conhecimento em organizações. Perspectivas em Gestão do Conhecimento, João Pessoa, v. 1, Número Especial, p. 27-43, out.

Araújo J., R. H. (2007). Precisão no Processo de Busca e Recuperação da Informação. Brasília: Thesaurus.

Barros, F. A. do R.; Caiado, R. V. R. Língua Portuguesa na Web 3.0: relações complexas de ensino por meio dos Recursos Educacionais Abertos (REAs). Entremeios (Revista de Estudos do Discurso, Seção Temática (Linguagem e Tecnologia), Programa de Pós-Graduação em Ciências da Linguagem (PPGCL), Universidade do Vale do Sapucaí (UNIVÁS), Pouso Alegre (MG), vol. 15, p. 247-266, jul. - dez. 2017. DOI: <http://dx.doi.org/10.20337/ISSN21793514revistaENTREMEIOSvol15pagina247a266

Batista, B. Machine Learning Sem Código. Usando Orange Data Mining para Criar um Modelo Preditivo sem Usar uma Linha de Código! Ensina.AI. (s. l.), 9 jul. 2019. Disponível em: Acesso em: 28, fev. 2021.

Bianque, A. G. (2015). Um Estudo de Caso sobre a Indexação Automática de Documentos Oficiais da UENP Baseado em Layouts. Disponível em: <http://200.201.11.152/bitstream/handle/123456789/37/TCC%20Vers%C3%A3o%20Final%20-%20Corre%C3%A7%C3%B5es.pdf?sequence=1&isAllowed=y>. Acesso em: 15, mai. de 2021.

Borges e Rhaddour. A Arquitetura da Informação em Plataformas Colaborativas como Suporte para a Gestão da Inteligência Coletiva nas Organizações. Disponível em: <http://www.scielo.org.pe/pdf/biblios/n69/a04n69.pdf>. Acesso em: 30, jun. de 2021.

Content, R. (2019). Internet das Coisas, Integração de Serviços e Interação Social: O Que Esperar da Web 4.0. Disponível em: <https://rockcontent.com/blog/Web- 4-0/>. Acesso em: 17 ago. Choo, C. W. (2003). A Organização do Conhecimento: Como as Organizações Usam a Informação para Criar Significado, Construir Conhecimento e Tomar Decisões. São Paulo: Editora SENAC.

Domo, 2021. Data Never Sleeps 8.0. Disponível em: <https://www.domo.com/learn/infographic/data-never-sleeps-8>. Acesso em: 20, set de 2021.

Dubey, P. (2019). Understand Text Summarization and Create Your own Summarizer in Python. Disponível em: <https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70>. Acesso em: 15, set. de 2021.

Fujita, M. S. L.; Gil-Leiva, I. As Linguagens de Indexação em Bibliotecas Nacionais, Arquivos Nacionais e Sistemas de Informação na América Latina. Rio de Janeiro: Universidade Federal do Rio de Janeiro, 2010.

Gartner Group. Big Data. Disponível em: <https://www.gartner.com/en/information-technology/glossary/big-data>. Acesso em: 13, jun. de 2021.

Goularte, F. Bif et al. Métricas de Sumarização Automática de Texto em Tarefas de um Ambiente Virtual de Aprendizagem. Brazilian Symposium on Computers in Education (Simpósio Brasileiro de Informática na Educação - SBIE), (S.l.), p. 752, nov. 2014. ISSN 2316-6533. Disponível em: <http://br-ie.org/pub/index.php/sbie/article/view/3007>. Acesso em: 17, jul. de 2021. doi: <http://dx.doi.org/10.5753/cbie.sbie.2014.752>.

Lévy, P. A Inteligência Coletiva: por uma antropologia do ciberespaço. 10. ed. São Paulo: Edições Loyola, 2015.

Machado, F. N. R. Disponível em: <https://www.researchgate.net/profile/Rafaela-Silva-6/publication/348691368/HIPERTEXTUALIDADES-E-TECNOLOGIAS-DE-AUTOMACAO.pdf#page=175>. Acesso em: 15, mai. de 2021.

Matos, F. F.; Souza, R. R.; Reis, Z. S. N. Análise de Dados de Saúde: mineração de texto com a utilização do Orange Canvas para exploração da informação. Encontro Nacional de Pesquisa em Ciência da Informação, n. XX ENANCIB, 2019. Disponível em: <http://hdl.handle.net/20.500.11959/brapci/122468>. Acesso em: 13, out. de 2021.

Muller, E.; Granatyr, j.; Lessing, O. R. Comparativo Entre o Algoritmo de Luhn e o Algoritmo GistSumm para Sumarização de Documentos. 2015.

Nascimento, G. F. C. de L. (2008). Folksonomia como Estratégia de Indexação dos Bibliotecários no Delicious / Geysa Flávia Câmara de Lima Nascimento. João Pessoa: PPGCI. Nunes, M.G.V.; Dias da Silva, B.C.; Rino, L. H. M.; Oliveira J., O. N.; Martins, R.T.; Montilha, G. (1999). Introdução ao Processamento das Línguas Naturais. Notas Didáticas do ICMC, N. 38. São Carlos/SP, Junho, 91p.

Padilha, M, A, O; Graeml, A, R. Perspectivas em Gestão & Conhecimento. João Pessoa, v. 9, n. 2, p. 153-173. Disponível em: <I: http://dx.doi.org/10.21714/2236-417X2019v9n2p153>. Acesso em: 03, jul de 2021.

Praveen, D. (2018). Entenda o Resumo de Texto e Crie seu Próprio Resumidor em Python. Disponível em: <https://towardsdatascience.com/understand-text-summarization-and-create-your-own-summarizer-in-python-b26a9f09fc70>. Acesso em: 17, jun. de 2021.

Silva, B. C. D.; Montilha, G.; Rino, L. H. M.; Specia, L.; Nunes, M. G. V.; Junior, O. N. O.; Martins, R. T.; Pardo, T. A. S. (2007). Introdução ao Processamento das Línguas Naturais e Algumas Aplicações. NILC.

Silva, S. R. de B. Corrêa, R. F. (2020). Sistemas de Indexação Automática por Atribuição: uma análise comparativa. Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação, 25, 01-25. Disponível em: <https://doi.org/10.5007/1518-2924.2020.e70740>. Acesso em: 15, mai. de 2022.

Silva, T. J.; Corrêa, R. F. Ferramentas para Indexação Automática: uma análise comparativa entre o OGMA, Parser PALAVRAS, LX-Parser e a extração manual de sintagmas nominais. In: Encontro Nacional de Pesquisa em Ciência da Informação, 16, 2015, Anais. João Pessoa: UFPB, 2015.

Svobodová, A.; Koudelková, P. Collective Intelligence and Knowledge Management as a Tool for Innovations. Economics and Management, v. 2011, n. 16, p. 942-946, 2011.

World Information System For Science and Technology. Princípios de Indexação. R. Esc. Bibliotecon. UFMG. 1981. v. 10, n. 1, p. 83-94 SILVA, T. E.; TOMAÉL, M. I. (2019). A Gestão da Informação nas Organizações. Inf. Disponível em: <http://www.uel.br/revistas/uel/index.php/informacao/article/view/1806/1540>. Acesso em: 15, abr. de 2021.

Zeferino, D. Dados, Informação e Conhecimento: qual a diferença dos conceitos?. Disponível em: <https://www.certifiquei.com.br/dados-informacao-conhecimento/>. Acesso em: 30, jun. de 2021.

Apêndice A

Tutorial *Textrant*: Exemplo de código *Python* utilizado na sumarização com artigos da FAPESP.

1. Instalar bibliotecas (CMD -> pip install biblioteca)
    a. panda
    b. nltk
    c. stopwords
    d. networkx
    e. numpy
2. Adicionar arquivo txt na mesma pasta do código
3. Rodar o comando abaixo primeiro

```
In [5]: import nltk
        nltk.download('stopwords')

        [nltk_data] Downloading package stopwords to
        [nltk_data]     C:\Users\a1620941\AppData\Roaming\nltk_data...
        [nltk_data]   Unzipping corpora\stopwords.zip.

Out[5]: True
```

4. Executar as etapas

```
def generate_summary(file_name, top_n=5):
    stop_words = stopwords.words('english')
    summarize_text = []

# Step 1 - Read text and tokenize
    sentences =  read_article(file_name)]

# Step 2 - Generate Similary Martix across sentences
    sentence_similarity_martix = build_similarity_matrix(sentences,
stop_words)

# Step 3 - Rank sentences in similarity martix
    sentence_similarity_graph =
nx.from_numpy_array(sentence_similarity_martix)
    scores = nx.pagerank(sentence_similarity_graph)

# Step 4 - Sort the rank and pick top sentences
    ranked_sentence = sorted(((scores[i],s) for i,s in
enumerate(sentences)), reverse=True)
    print("Indexes of top ranked_sentence order by ",
ranked_sentence)for i in range(top_n):
      summarize_text.append(" ".join(ranked_sentence[i][1]))

# Step 5 - Off course, output the summarize texr
    print("Summarize Text: \n", ". ".join(summarize_text))
```

5. Apresentará texto sumarizado

In an attempt to build an AI-ready workforce, Microsoft announced Intelligent Cloud Hub which has been launched to empower the next generation of students with AI-ready skills

Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services

As part of the program, the Redmond giant which wants to expand its reach and is planning to build a strong developer ecosystem in India with the program will set up the core AI infrastructure and IoT Hub for the selected campuses

The company will provide AI development tools and Azure AI services such as Microsoft Cognitive Services, Bot Services and Azure Machine Learning.According to Manish Prakash, Country General Manager-PS, Health and Education, Microsoft India, said, "With AI being the defining technology of our time, it is transforming lives and industry and the jobs of tomorrow will require a different skillset

This will require more collaborations and training and working with AI

Thatâ€™s why it has become more critical than ever for educational institutions to integrate new cloud and AI technologies

The program is an attempt to ramp up the institutional set-up and build capabilities among the educators to educate the workforce of tomorrow." The program aims to build up the cognitive skills and in-depth understanding of developing intelligent cloud connected solutions for applications across industry

Earlier in April this year, the company announced Microsoft Professional Program In AI as a learning track open to the public

The program was developed to provide job ready skills to programmers who wanted to hone their skills in AI and data science with a series of online courses which featured hands-on labs and expert instructors as well

This program also included developer-focused AI school that provided a bunch of assets to help build AI skills.

Indexes of top ranked_sentence order are [(0.15083257041122708, ['Envisioned', 'as', 'a', 'three-year', 'collaborative', 'program,', 'Intelligent', 'Cloud', 'Hub', 'will', 'support', 'around', '100', 'institutions', 'with', 'AI', 'infrastructure,', 'course', 'content', 'and', 'curriculum,', 'developer', 'support,', 'development', 'tools', 'and', 'give', 'students', 'access', 'to', 'cloud', 'and', 'AI', 'services']), (0.13161201335715553, ['The', 'company', 'will', 'provide', 'AI', 'development', 'tools', 'and', 'Azure', 'AI', 'services', 'such', 'as', 'Microsoft', 'Cognitive', 'Services,', 'Bot', 'Services', 'and', 'Azure', 'Machine', 'Learning.According', 'to', 'Manish', 'Prakash,', 'Country', 'General', 'Manager-PS,', 'Health', 'and', 'Education,', 'Microsoft', 'India,', 'said,', '"With', 'AI', 'being', 'the', 'defining', 'technology', 'of', 'our', 'time,', 'it', 'is', 'transforming', 'lives', 'and', 'industry', 'and', 'the', 'jobs', 'of', 'tomorrow', 'will', 'require', 'a', 'different', 'skillset']), (0.11403047674961146, ['Earlier', 'in', 'April', 'this', 'year,', 'the', 'company', 'announced', 'Microsoft', 'Professional', 'Program', 'In', 'AI', 'as', 'a', 'learning', 'track', 'open', 'to', 'the', 'public']), (0.10721749759953528, ['In', 'an', 'attempt', 'to', 'build', 'an', 'AI-ready', 'workforce,', 'Microsoft', 'announced', 'Intelligent', 'Cloud', 'Hub', 'which', 'has', 'been', 'launched', 'to', 'empower', 'the', 'next', 'generation', 'of', 'students', 'with', 'AI-ready', 'skills']), (0.10404298514456578, ['As', 'part', 'of', 'the', 'program,', 'the', 'Redmond', 'giant', 'which', 'wants', 'to', 'expand', 'its', 'reach', 'and', 'is', 'planning', 'to', 'build', 'a', 'strong', 'developer', 'ecosystem', 'in', 'India', 'with', 'the', 'program', 'will', 'set', 'up', 'the', 'core', 'AI', 'infrastructure', 'and', 'IoT', 'Hub', 'for', 'the', 'selected', 'campuses']), (0.10031366655994461, ['Thatâ€™s', 'why', 'it', 'has', 'become', 'more', 'critical', 'than', 'ever', 'for', 'educational', 'institutions', 'to', 'integrate', 'new', 'cloud', 'and', 'AI', 'technologies']),

(0.10001137283486655, ['The', 'program', 'is', 'an', 'attempt', 'to', 'ramp', 'up', 'the', 'institutional', 'set-up', 'and', 'build', 'capabilities', 'among', 'the', 'educators', 'to', 'educate', 'the', 'workforce', 'of', 'tomorrow.'", 'The', 'program', 'aims', 'to', 'build', 'up', 'the', 'cognitive', 'skills', 'and', 'in-depth', 'understanding', 'of', 'developing', 'intelligent', 'cloud', 'connected', 'solutions', 'for', 'applications', 'across', 'industry']), (0.09916750119894317, ['This', 'will', 'require', 'more', 'collaborations', 'and', 'training', 'and', 'working', 'with', 'AI']), (0.09277191614415067, ['The', 'program', 'was', 'developed', 'to', 'provide', 'job', 'ready', 'skills', 'to', 'programmers', 'who', 'wanted', 'to', 'hone', 'their', 'skills', 'in', 'AI', 'and', 'data', 'science', 'with', 'a', 'series', 'of', 'online', 'courses', 'which', 'featured', 'hands-on', 'labs', 'and', 'expert', 'instructors', 'as', 'well'])]

Summarize Text:

Envisioned as a three-year collaborative program, Intelligent Cloud Hub will support around 100 institutions with AI infrastructure, course content and curriculum, developer support, development tools and give students access to cloud and AI services. The company will provide AI development tools and Azure AI services such as Microsoft Cognitive Services, Bot Services and Azure Machine Learning.According to Manish Prakash, Country General Manager-PS, Health and Education, Microsoft India, said, "With AI being the defining technology of our time, it is transforming lives and industry and the jobs of tomorrow will require a different skillset

ARQUIVO TXT

No arquivo txt não pode pular linha, os parágrafos precisam seguir na mesma linha, além disso existe um limite de quantidade.

Exemplo em português

2020/04505-3
Pesquisador Responsável: Marcelo Urbano Ferreira
Instituição sede
Instituto de Ciências Biomédicas / USP

O vÃ-rus SARS-CoV-2 disseminou-se globalmente e representa agora um importante desafio para os paÃ-ses de baixa e mÃ©dia renda, onde a infraestrutura de saÃºde pode rapidamente tornar-se sobrecarregada

Esta proposta parte de nossas pesquisas de campo em andamento, financiadas pela FAPESP, para investigar a epidemiologia e o controle da infecÃ§Ã£o por SARS-CoV-2 em MÃ¢ncio Lima, uma pequena cidade amazÃ´nica

Tem-se como objetivo geral traduzir as informaÃ§Ãµes geradas pelo estudo de campo em evidÃªncias para orientar o controle de COVID-19 em uma das regiÃµes mais pobres do Brasil

Partimos da hipÃ³tese de que muitas infecÃ§Ãµes por SARS-CoV-2 permanecem despercebidas e os portadores assintomÃ¡ticos de infecÃ§Ã£o podem continuar disseminando o patÃ³geno em suas interaÃ§Ãµes sociais cotidianas atÃ© sua eliminaÃ§Ã£o espontÃ¢nea, tornando-se imunes a reinfecÃ§Ãµes ou, pelo menos, Ã doenÃ§a grave

Os meios propostos para testar esta hipÃ³tese sÃ£o: (a) utilizar ensaios sorolÃ³gicos seriados para detectar retrospectivamente eventos de soroconversÃ£o, estimar o tamanho do surto de SARS-CoV-2 e identificar fatores de risco associados Ã soroconversÃ£o na comunidade; (b) identificar as interaÃ§Ãµes sociais e os espaÃ§os compartilhados, como o domicÃ-lio, o local de trabalho, as escolas e as igrejas, que possam ter contribuÃ-do para a transmissÃ£o local de SARS-Cov-2; (c) calcular a proporÃ§Ã£o de infecÃ§Ãµes por SARS-CoV-2 diagnosticadas retrospectivamente que permaneceram assintomÃ¡ticas ou cursaram com sintomas leves, geralmente sem diagnÃ³stico prÃ©vio, e aquelas associadas Ã doenÃ§a (COVID-19), resultando em visitas a serviÃ§os de saÃºde e atÃ© mesmo em hospitalizaÃ§Ã£o, e (d) determinar a proporÃ§Ã£o de indivÃ-duos que, ao se tornarem soropositivos durante o surto, permanecem com anticorpos contra SARS-CoV-2 ao longo dos 12 meses eguintes.

Indexes of top ranked_sentence order are [(0.4134928080574074, ['Esta', 'proposta', 'parte', 'de', 'nossas', 'pesquisas', 'de', 'campo', 'em', 'andamento,', 'financiadas', 'pela', 'FAPESP,', 'para', 'investigar', 'a', 'epidemiologia', 'e', 'o', 'controle', 'da', 'infecÃ§Ã£o', 'por', 'SARS-CoV-2', 'em', 'MÃ¢ncio', 'Lima,', 'uma', 'pequena', 'cidade', 'amazÃ´nica']), (0.21858918123102267, ['Partimos', 'da', 'hipÃ³tese', 'de', 'que', 'muitas', 'infecÃ§Ãµes', 'por', 'SARS-CoV-2', 'permanecem', 'despercebidas', 'e', 'os', 'portadores', 'assintomÃ¡ticos', 'de', 'infecÃ§Ã£o', 'podem', 'continuar', 'disseminando', 'o', 'patÃ³geno', 'em', 'suas', 'interaÃ§Ãµes', 'sociais', 'cotidianas', 'atÃ©', 'sua', 'eliminaÃ§Ã£o', 'espontÃ¢nea,', 'tornando-se', 'imunes', 'a', 'reinfecÃ§Ãµes', 'ou,', 'pelo', 'menos,', 'Ã\xa0', 'doenÃ§a', 'grave']), (0.199007522131989, ['Tem-se', 'como', 'objetivo', 'geral', 'traduzir', 'as', 'informaÃ§Ãµes', 'geradas', 'pelo', 'estudo', 'de', 'campo', 'em', 'evidÃªncias', 'para', 'orientar', 'o', 'controle', 'de', 'COVID-19', 'em', 'uma', 'das', 'regiÃµes', 'mais', 'pobres', 'do', 'Brasil']), (0.16891048857958096, ['O', 'vÃ\xadrus', 'SARS-CoV-2', 'disseminou-se', 'globalmente', 'e', 'representa', 'agora', 'um', 'importante', 'desafio', 'para', 'os', 'paÃ\xadses', 'de', 'baixa', 'e', 'mÃ©dia', 'renda,', 'onde', 'a', 'infraestrutura', 'de', 'saÃºde', 'pode', 'rapidamente', 'tornar-se', 'sobrecarregada'])]

Summarize Text:

Esta proposta parte de nossas pesquisas de campo em andamento, financiadas pela FAPESP, para investigar a epidemiologia e o controle da infecÃ§Ã£o por SARS-CoV-2 em MÃ¢ncio Lima, uma pequena cidade amazÃ´nica. Partimos da hipÃ³tese de que muitas infecÃ§Ãµes por SARS-CoV-2 permanecem despercebidas e os portadores assintomÃ¡ticos de infecÃ§Ã£o podem continuar disseminando o patÃ³geno em suas interaÃ§Ãµes sociais cotidianas atÃ© sua eliminaÃ§Ã£o espontÃ¢nea, tornando-se imunes a reinfecÃ§Ãµes ou, pelo menos, Ã doenÃ§a grave

Apêndice B

Guia prático da Ferramenta *Orange Data Mining*: Aplicação das Bases de Dados na Ferramenta e Construção dos Fluxos de Trabalho

Nesta etapa acontece a coleta de dados através de textos, criando base de dados a ser utilizada em todo o processo. Os documentos textuais utilizados para formar a base foram artigos da FAPESP sumarizados

Indexação por Extração

Por meio dessa técnica de indexação de conteúdo é extraído informações da base de dados de forma automática a partir de um termo frequente.

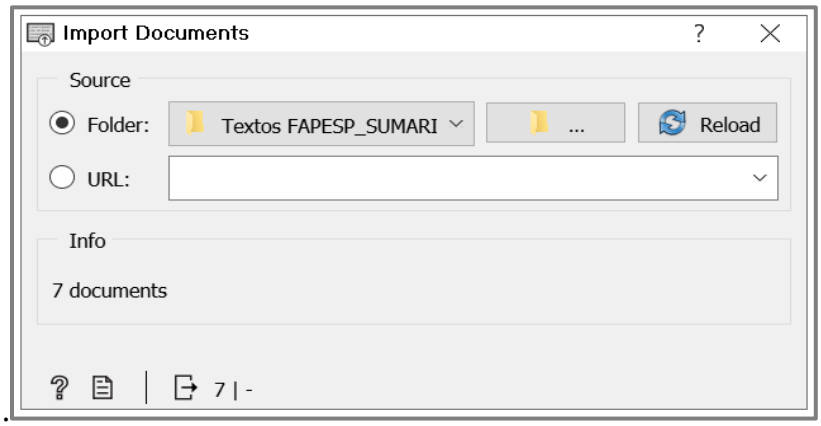

Figura 1. *Import Documents* (Importa documentos)

A figura 2 apresenta uma coleção de artigos da base de dados que serão analisados.

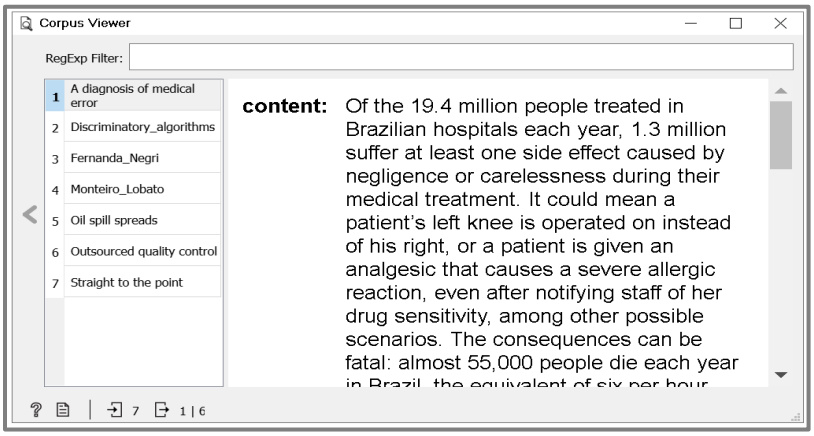

Figura 2. *Corpus Viewer* (visualizador de conteúdo).

A figura 3 seleciona as colunas que serão trabalhadas (*name* e *content*), ignorando apenas o caminho do arquivo.

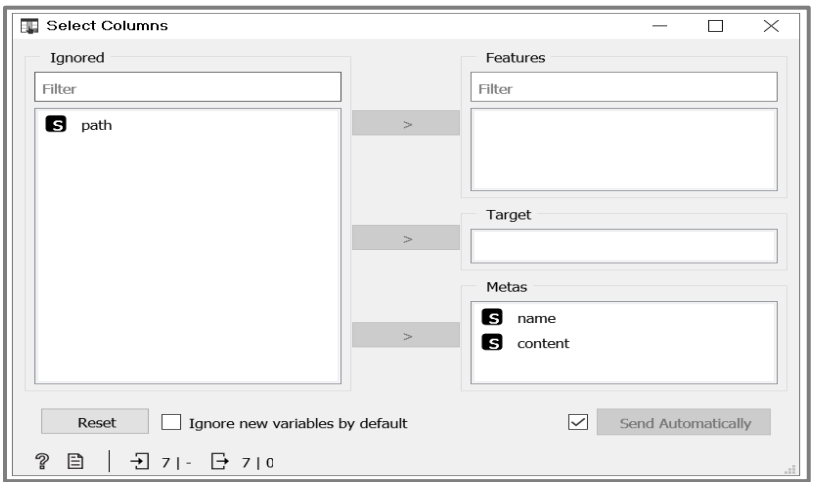

Figura 3. Select *Columns* (seleciona colunas)

É realizado um pré-processamento de dados, limpando os dados irrelevantes a partir de: padronização, filtros, remoção de acentos, alterar as palavras para minúsculo, remover URLs, *stopword*, expressão regulares (números e sinais).

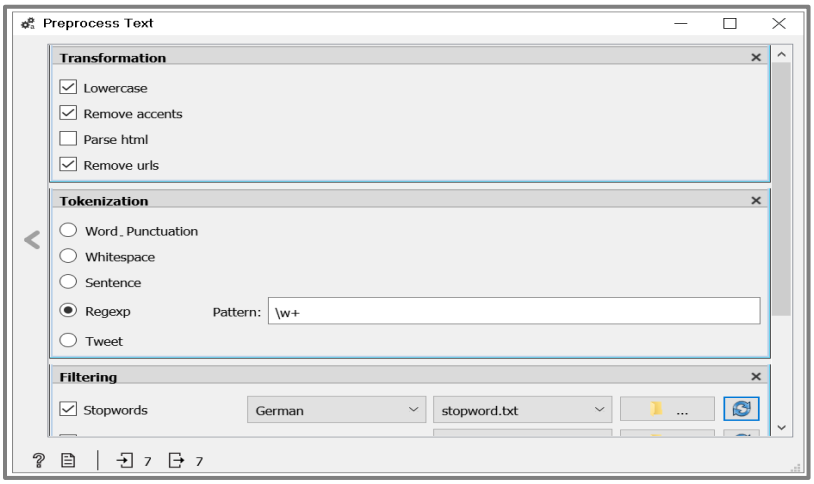

Figura 4. *Preprocess Text* (pré-processamento de texto)

Uma lista de palavras externas criada e utilizada para remover palavras que são irrelevantes para a análise como: preposições, artigos, gírias, sinais, números, entre outras.

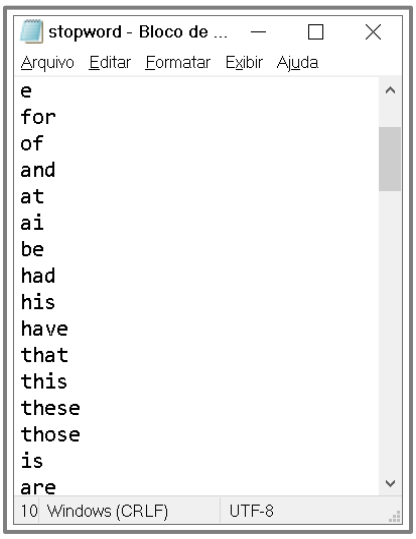

Figura 5. *Stopword*

Após o pré-processamento de texto é criado a *Bags of Words* (bolsa de palavras). Nesta etapa os textos são representados como multiconjuntos de palavras, no qual é desconsiderando a estrutura gramatical, mas é mantido as palavras que aparecem com mais frequência, assim, criando uma métrica para as palavras mais frequentes.

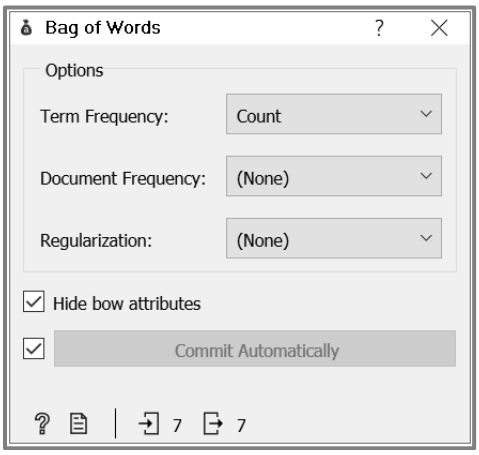

Figura 6. *Bag of Words* (bolsa de palavras)

No *Data Table* são exibidos os textos e as *Bags of Words* geradas.

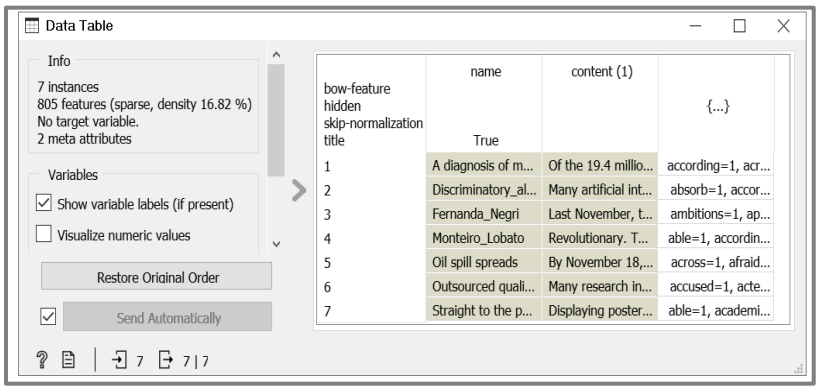

Figura 7. *Data Table* (tabela de dados)

6Nesta etapa é ligada a *Bag of Word* ao *Distances* para calcular a distância entre os artigos, baseando-se na distância euclidiana ($d=\sqrt{\sum_{i=1}^{n}(a_i-b_i)^2}$).

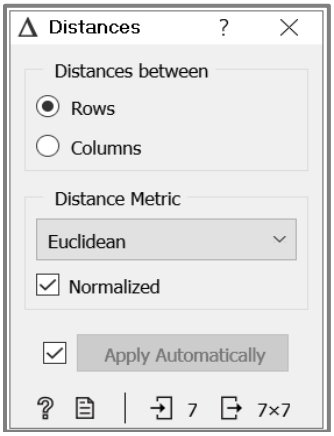

Figura 8. *Distances* (distâncias)

Para visualizar a distância calculada entre os artigos no item anterior, foi inserido a *Distance Matrix*.

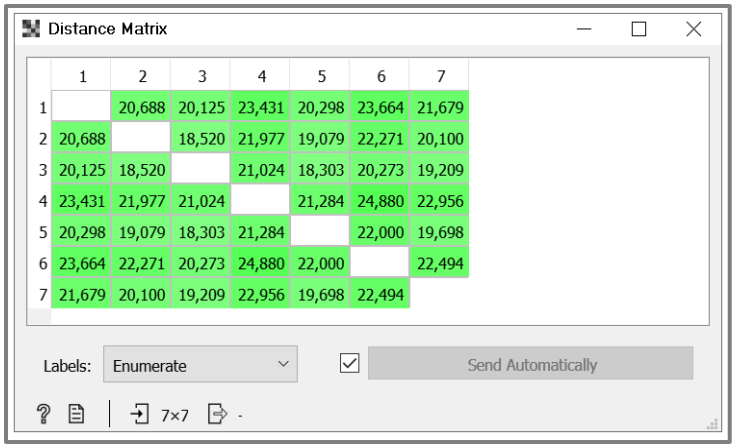

Figura 9. *Distance Matrix* (matriz de distância)

A partir do cálculo de distância é possível fazer um agrupamento hierárquico dos

textos (clusterização), possibilitando assim entender a relação entre os artigos.

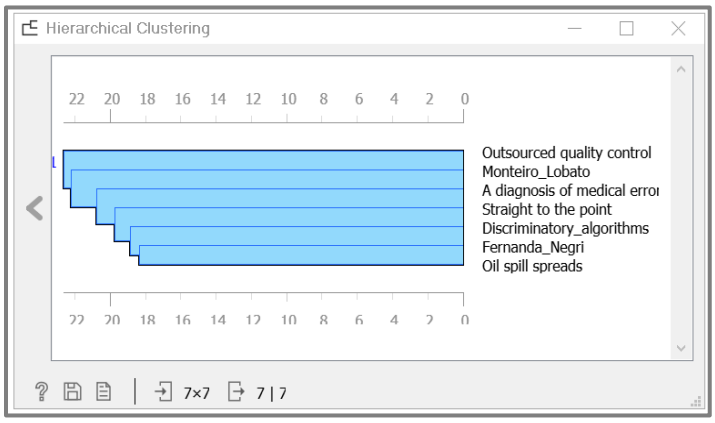

Figura 10. *Hierarchical Clustering* (agrupamento hierárquico)

O item anterior é ligado a uma *Word Cloud* (nuvem de palavras) para visualizar a hierarquia e relacionamento entre os artigos.

Figura 11. *Word Cloud* (nuvem de palavras)

Este componente representa de forma visual as distâncias entre os artigos.

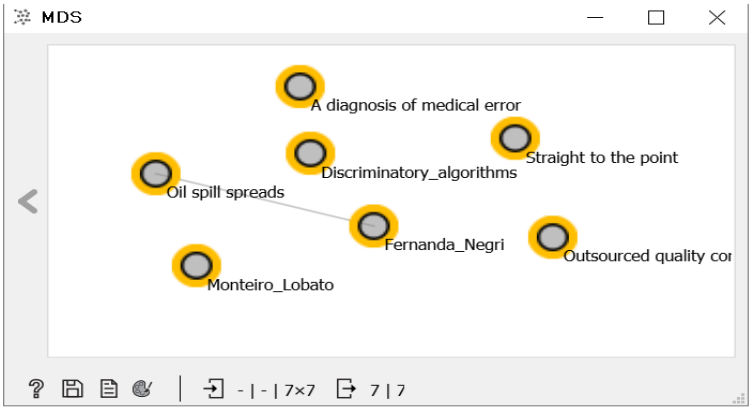

Figura 12. *Multidimensional Scaling* – MDS (dimensionamento multidimensional)

O item anterior é ligado a uma *Word Cloud* para visualizar e comparar as distâncias entre os clusters.

Figura 13*. Word Cloud* (nuvem de palavras)

Indexação por Atribuição

Por meio dessa técnica de indexação de conteúdo é atribuído um termo a base de dados e a extração acontece de forma manual a partir das informações correlatas que são classificadas por perfil e apresentadas.

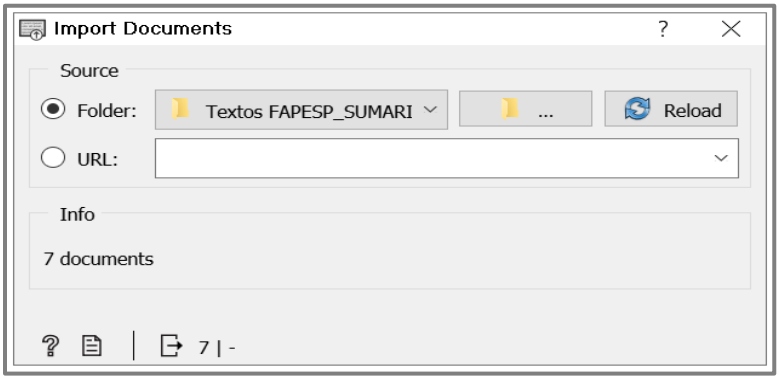

Figura 14. *Import Documents* (importa documentos)

Apresenta uma coleção de artigos da base de dados que serão analisados.

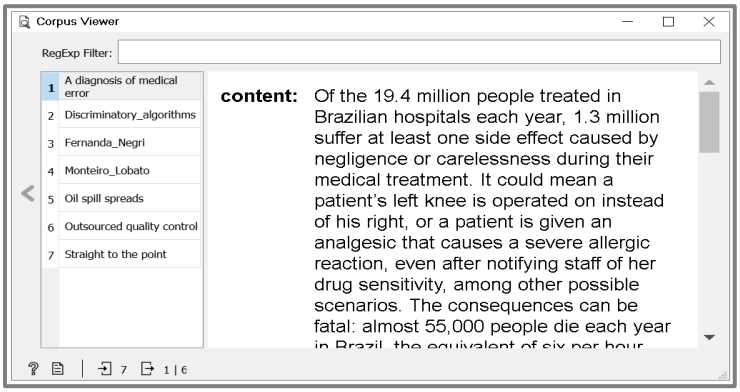

Figura 15. *Corpus Viewer* (visualizador de conteúdos)

Seleciona as colunas que serão trabalhadas (*name e content*), ignorando apenas o caminho do arquivo.

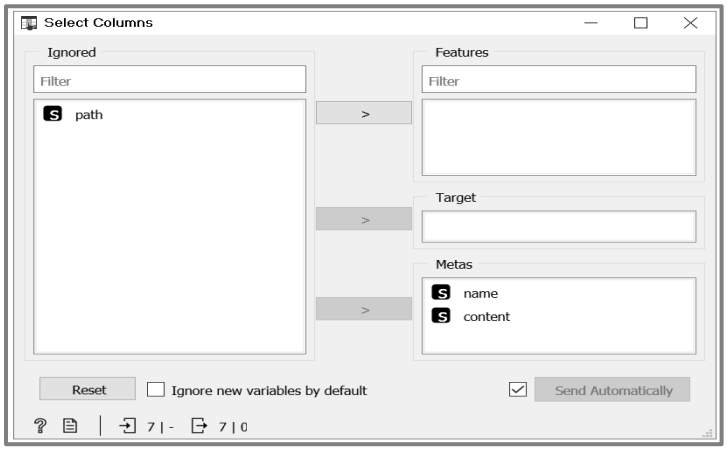

Figura 16. *Select Columns* (seleciona colunas)

É realizado um pré-processamento de dados, limpando os dados irrelevantes a partir de: padronização, filtros, remoção de acentos, alterar as palavras para minúsculo, remover URLs, adicionar uma *stopword*, remover expressões regulares (números e sinais).

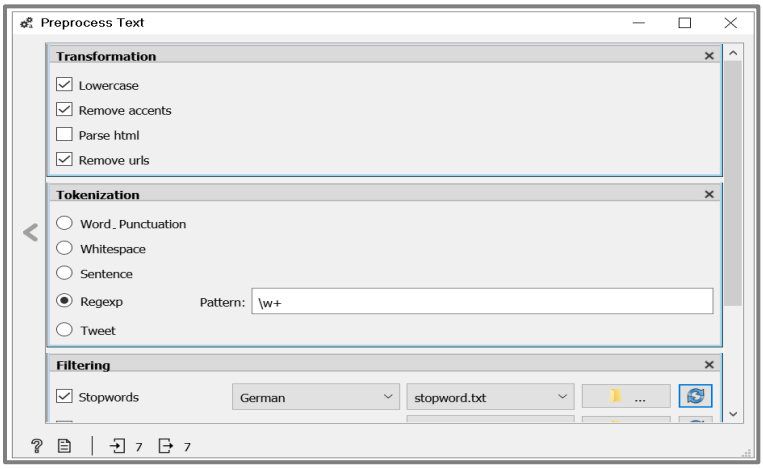

Figura 17. *Preprocess Text* (pré-processamento de texto)

Uma *Stopword* é uma lista de palavras externas criada e utilizada para remover palavras que são irrelevantes para a análise como: preposições, artigos, gírias, sinais, números, entre outras.

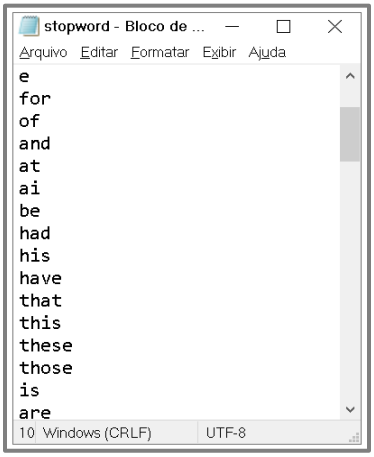

Figura 18. *Stopword*

Nesta etapa a *Word Cloud* é utilizada para atribuir (selecionar) as palavras para a análise da concordância na sequência.

Figura 19. *Word Cloud* (nuvem de palavras).

Nesta etapa são apresentadas as palavras selecionadas no *Word Cloud* dentro de um ou mais contextos.

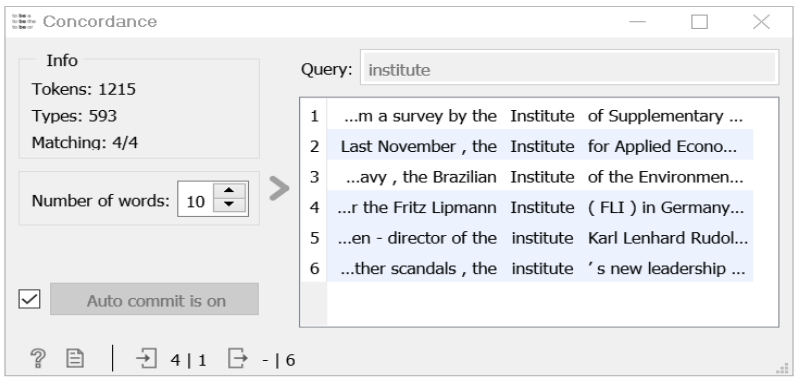

Figura 20. *Concordance* (concordância)

Documento Digitalizado Público

Artigo TCC - Thainara Barbosa da Silva

Assunto: Artigo TCC - Thainara Barbosa da Silva
Assinado por: Carlos Pagani
Tipo do Documento: Estudo
Situação: Finalizado
Nível de Acesso: Público
Tipo do Conferência: Documento Digital

Documento assinado eletronicamente por:

- **Carlos Eduardo Pagani**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 05/10/2022 11:49:11.

Este documento foi armazenado no SUAP em 05/10/2022. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse https://suap.ifsp.edu.br/verificar-documento-externo/ e forneça os dados abaixo:

Código Verificador: 1120166
Código de Autenticação: 9c6c7bce12