11º Congresso de Inovação, Ciência e Tecnologia do IFSP - 2020

LEVANTAMENTO E CARACTERIZAÇÃO DE COLEÇÕES DE DOCUMENTOS PARA MINERAÇÃO DE TEXTOS EM PORTUGUÊS

MARCOS R. V. SIQUEIRA[1], ROBERTA A. SINOARA[2]

[1]Graduando em Tecnologia de Análise e Desenvolvimento de Sistemas, Bolsista PIBIFSP, IFSP, Câmpus Boituva, marcos.vicente@ifsp.edu.br

[2]Professora EBTT, IFSP, Câmpus Boituva, roberta.sinoara@ifsp.edu.br

Área de conhecimento (Tabela CNPq): 1.03.03.03-0 Banco de Dados

RESUMO: Com o aumento da quantidade e variedade de textos em formato digital, as técnicas de Mineração de Textos tornam-se essenciais no apoio à extração de conhecimento e têm sido foco de muitos trabalhos de pesquisa. Uma necessidade comum em tais pesquisas é a necessidade de coleções de textos rotulados para avaliação dos métodos desenvolvidos. Essa necessidade surge tanto no treinamento supervisionado de modelos quanto na avaliação de modelos gerados de maneira não supervisionada. Apesar da grande relevância, considerando-se o idioma português, não há um inventário consolidado de coleções de textos rotulados disponíveis para pesquisa. Este projeto visa tratar essa lacuna, sendo desenvolvido com o objetivo de contribuir com avanços das pesquisas de Mineração de Textos no idioma português por meio da disponibilização de informações consolidadas sobre coleções de documentos rotulados e disponíveis para pesquisas da área. Para realizar o levantamento das coleções, foram consultados artigos recentes das principais conferências da área, resultando na identificação de 52 coleções de documentos (ou *datasets*) utilizadas em pesquisas da área de Mineração de Textos.

PALAVRAS-CHAVE: mineração de textos; classificação de textos; agrupamento de textos; datasets rotulados

GATHERING AND CHARACTERIZATION OF PORTUGUESE TEXT COLLECTIONS FOR TEXT MINING

ABSTRACT: Considering the increase in quantity and variety of digital texts, Text Mining methods become a crucial support for knowledge discovery and are the focus of several research projects. The evaluation of the developed methods often needs collections of labeled text, which are required for training supervised models as well as for evaluating unsupervised models. Although labeled text collections are very important for Text Mining researches, there is not a consolidated inventory of labeled text collection written in Portuguese and available for research. In this context, the objective of this project is to contribute to the advances of Portuguese Text Mining researches through the gathering and characterization of Portuguese text collections that are available for the community. The gathering of the text collections was based on recent publications of the main conferences in the research field. As a result, we identified 52 text collections (datasets) used in Text Mining researches.
KEYWORDS: text mining; text classification; text clustering; labeled datasets

INTRODUÇÃO

Os avanços e a disseminação do uso de tecnologias de informação e comunicação têm causado um aumento expressivo na geração e armazenamento de dados em formato digital. Projeções realizadas em 2017 pela *International Data Corporation* indicavam que em 2025 seriam gerados 165 *zettabytes* (REINSEL; GANTZ; RYDNING, 2017). Já no estudo realizado em 2018, essa projeção para 2025 passou para 175 *zettabytes* (REINSEL; GANTZ; RYDNING, 2018). Uma parte desse universo é composta por dados não estruturados, como os documentos textuais gerados internamente nas empresas, as revisões e comentários sobre produtos e serviços em páginas da web, e os *posts* em redes sociais.

Tendo em vista esse cenário, as técnicas de Mineração de Textos (AGGARWAL; ZHAI, 2012) tornam-se essenciais para apoio à extração de conhecimento de dados textuais. Uma necessidade comum em pesquisas na área de Mineração de Textos é a disponibilidade de coleções de textos rotulados. Essa necessidade surge tanto no treinamento supervisionado de modelos quanto na avaliação de modelos gerados de maneira não supervisionada. Apesar da grande relevância, considerando-se o idioma português, não há um inventário consolidado de coleções de textos rotulados disponíveis para pesquisa.

Este projeto visa tratar essa lacuna e será desenvolvido com o objetivo de contribuir com avanços das pesquisas de Mineração de Textos no idioma português por meio da disponibilização de informações consolidadas sobre coleções de documentos rotulados e disponíveis para pesquisas da área. Assim, os principais objetivos específicos deste projeto são: levantamento de coleções de textos rotulados e escritos em português que estão disponíveis para pesquisa, e execução de uma avaliação experimental considerando as tarefas de classificação e agrupamento de textos. Este projeto encontra-se em andamento, sendo que neste artigo são apresentados os resultados referentes ao primeiro objetivo citado.

MATERIAIS E MÉTODOS

Inicialmente foi realizada uma revisão da literatura e estudo de conceitos fundamentais das áreas de Mineração de Textos, Aprendizados de Máquina e Processamento de Linguagem Natural. Considerando os temas utilizados no projeto, foi estudado material teórico por meio de livros e textos científicos (REZENDE, 2003; SINOARA, 2018). A revisão teve foco em conceitos de classificação e agrupamento de textos, além da avaliação de algoritmos de Aprendizado de Máquina.

O desenvolvimento deste projeto foi dividido em duas etapas: (i) levantamento e caracterização das coleções de textos em português disponíveis para pesquisa; e (ii) avaliação de coleções em tarefas de classificação e agrupamento. Este projeto está em andamento e, neste artigo, são apresentados os resultados obtidos principalmente na primeira etapa.

Na primeira etapa, a identificação de coleções de textos rotulados disponíveis para pesquisas em Mineração de Textos foi realizada por meio de duas atividades principais: identificação dos principais grupos de pesquisa que trabalham com textos em português e posterior identificação das coleções de textos utilizadas nessas pesquisas. Essas atividades foram realizadas por meio de consultas ao Currículo Lattes de pesquisadores e a publicações das principais conferências nacionais da área (BRACIS[1], ENIAC[2], WebMedia[3], PROPOR[4] e STIL[5]).

As coleções de textos identificadas foram obtidas, estudadas e caracterizadas. Para os casos em

[1]BRACIS - Brazilian Conference on Intelligent Systems
[2]ENIAC - Encontro Nacional de Inteligência Artificial e Computacional
[3]WebMedia - Simpósio Brasileiro de Sistemas Multimídia e Web
[4]PROPOR - Conferência Internacional de Processamento Computacional da Língua Portuguesa
[5]STIL - Brazilian Symposium in Information and Human Language Technology

que as coleções não estavam disponíveis para *download*, os autores foram contactados. A partir das coleções obtidas, foi construída uma descrição padronizada de todas as coleções e *datasets* derivados. A descrição padronizada irá contribuir para pesquisas em Mineração de Textos visto que resultados da aplicação de algoritmos de AM poderão ser analisados com base nas características dos dados de entrada. As principais propriedades das coleções de textos e *datasets* consideradas baseiam-se nas propriedades levantadas para coleções de textos em inglês por (ROSSI; MARCACINI; REZENDE, 2013) e (SINOARA, 2018). Vale ressaltar que algumas propriedades se referem especificamente à representação de textos adotada e serão levantadas para cada representação gerada a partir dos textos pré-processados, na segunda etapa deste projeto.

A segunda etapa da pesquisa, que está em andamento, seguirá o processo de Mineração de Textos apresentado na Figura 1 para realização de avaliação experimental de um conjunto das coleções de textos obtidas na primeira etapa. Assim, serão realizadas as etapas de pré-processamento, extração de padrões e pós-processamento para cada coleção selecionada. Na extração de padrões serão considerados algoritmos de classificação e de agrupamento.

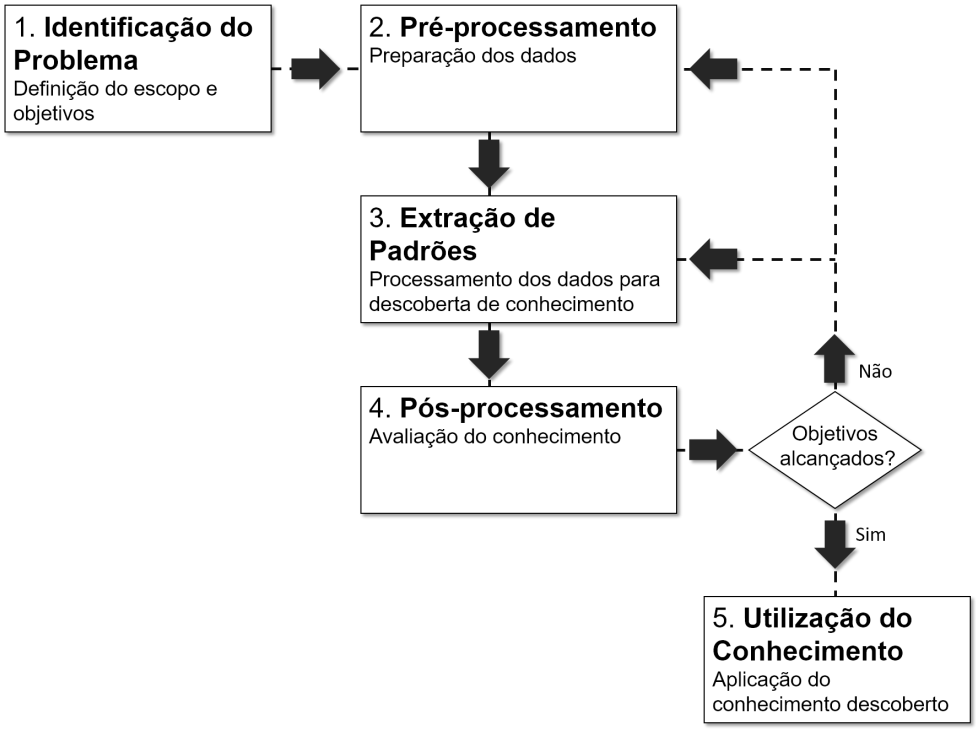

Figura 1: Processo de Mineração de Textos. Adaptada de (SINOARA; ANTUNES; REZENDE, 2017)

Para realizar o pré-processamento das coleções foi desenvolvida uma ferramenta de apoio, com base em uma ferramenta desenvolvida por (SCHEICHER et al., 2019). Essa ferramenta de pré-processamento foi escrita na linguagem *Python*, utilizando bibliotecas como *NLTK*, *Scikit Learn* e *Pandas*. O desenvolvimento teve como foco a criação de uma ferramenta genérica, ou seja, que pode ser utilizada em outros projetos. Nessa atividade foram consultados livros da linguagem *Python*, ciência de dados e expressões regulares (BEAZLEY; JONES, 2013; GRUS, 2016; BORGES, 2014; JARGAS, 2016). Como resultado final obteve-se representações estruturadas das coleções, em formato ARFF (WITTEN; FRANK, 2005), que serão utilizadas para conclusão da avaliação experimental das coleções selecionadas.

RESULTADOS E DISCUSSÃO

Para o levantamento das coleções de textos foram consultados os artigos publicados nos últimos anos (2018 e 2019) das conferências BRACIS, PROPOR, STIL, ENIAC e WebMedia, além de consultas diretas a pesquisadores e grupos de pesquisa da área. No total, foram considerados 350 artigos, destes 42 foram selecionados por serem relacionados a processamento de língua natural e classificação de textos em português. Na Tabela 1 são apresentados os números de artigos selecionados por conferência. Os artigos contabillizados em "Outras fontes" foram identificados por meio de referências em artigos das conferências consultadas ou diretamente através de grupos de pesquisa da área.

Tabela 1: Número de artigos consultados para identificação de coleções de textos rotulados escritos em português.

Conferência	Ano	Número de artigos
BRACIS	2019	5
ENIAC	2018	2
	2019	9
PROPOR	2018	4
STIL	2018	1
	2019	7
WebMedia	2018	1
	2019	3
Outras fontes		10

Os 42 artigos selecionados foram lidos em busca de coleções de textos rotulados e escritos em português. A partir destes artigos foram encontradas 52 coleções de textos (*datasets*). Na próxima etapa deste projeto, um subconjunto desses *datasets* será selecionado para execução de avaliações experimentais considerando-se as tarefas de classificação e agrupamento de textos. Também serão extraídas características das representações das coleções, obtidas com a utilização da ferramenta de apoio desenvolvida neste projeto. Todas as informações das coleções de textos serão disponibilizadas para a comunidade de pesquisa.

CONCLUSÕES

Visando tratar a lacuna referente à disponibilidade de um inventário consolidado de coleções de textos escritos em português para pesquisas em Mineração de Textos, nesta primeira etapa do projeto, foi realizado um extenso levantamento. Foram consultados diversos artigos, publicados em importantes conferências da área. Com este trabalho, constatou-se que a comunidade de pesquisa tem desenvolvido um número considerável de coleções de textos rotulados no idioma português. No entanto, essas coleções se encontravam dispersas, dificultando a sua identificação e posterior uso por outros grupos de pesquisa.

Como trabalho futuro, dando continuidade ao projeto, este inventário será publicado e avaliações experimentais em classificação e agrupamento de textos serão realizadas com uma seleção dos *datasets* obtidos. Com isso, será disponibilizado à comunidade de pesquisa em Mineração de Textos uma importante ferramenta para avaliação de métodos desenvolvidos para o idioma português.

AGRADECIMENTOS

Os autores agradecem ao auxílio fornecido pelo Programa Institucional de Bolsas de Iniciação Científica e Tecnológica do IFSP (PIBIFSP).

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. (Ed.). *Mining Text Data*. [S.l.]: Springer, 2012.

BEAZLEY, D.; JONES, B. K. *Python Cookbook: Receitas para dominar o Python 3*. [S.l.]: Novatec, 2013.

BORGES, L. E. *Python para desenvolvedores: aborda Python 3.3*. [S.l.]: Novatec Editora, 2014.

GRUS, J. *Data Science do zero: Primeiras regras com o Python*. [S.l.]: Alta Books, 2016.

JARGAS, A. M. *Expressões regulares: uma abordagem divertida*. [S.l.]: Novatec Editora, 2016.

REINSEL, D.; GANTZ, J.; RYDNING, J. *Data Age 2025: The Evolution of Data to Life-Critical*. 2017. IDC White Paper.

REINSEL, D.; GANTZ, J.; RYDNING, J. *Data Age 2025: The Digitization of the World - From Edge to Core*. 2018. IDC White Paper.

REZENDE, S. O. (Ed.). *Sistemas Inteligentes: Fundamentos e Aplicações*. [S.l.]: Editora Manole, 2003.

ROSSI, R. G.; MARCACINI, R. M.; REZENDE, S. O. *Benchmarking Text Collections for Classification and Clustering Tasks*. [S.l.], 2013.

SCHEICHER, R. B. et al. Sentiment classification improvement using semantically enriched information. In: *Proceedings of the ACM Symposium on Document Engineering 2019*. [S.l.: s.n.], 2019. p. 1–4.

SINOARA, R. A. *Aspectos semânticos na representação de textos para classificação automática*. Tese (Doutorado) — Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2018. Programa de Pós-Graduação em Ciências de Computação e Matemática Computacional.

SINOARA, R. A.; ANTUNES, J.; REZENDE, S. O. Text mining and semantics: a systematic mapping study. *Journal of the Brazilian Computer Society*, v. 23, n. 9, p. 1–20, 2017.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd. ed. [S.l.]: Morgan Kaufmann, 2005.