WIDaT 2022

WORKSHOP DE INFORMAÇÃO, DADOS E TECNOLOGIA

CLASSIFICAÇÃO AUTOMÁTICA DE ARTIGOS PUBLICADOS NO ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - 2021

Liliane Cristina Soares Sousa - Doutoranda do PPGCI – UEL

Dr. José Eduardo Santarém Segundo – Universidade de São Paulo - USP

Dr. Fábio Parra Furlanete – Universidade Estadual de Londrina - UEL

INTRODUÇÃO

Destaca-se a importância de estudos de Data Science e algoritmos de Machine Learning, para aplicação na Ciência da Informação, como ferramenta de processamento e análise de corpus textual.

Em que medida algoritmos de Machine Learning são capazes de processar métricas de validação em classificação de dados textuais?

OBJETIVO:

- Diagnosticar se algoritmos de Machine Learning, aplicados em dados não-estruturados, são capazes de classificar documentos textuais, pelo olhar das técnicas de classificação dos bibliotecários.

PROBLEMATIZAÇAO:

- Em que medida algoritmos de Machine Learning são capazes de processar métricas de validação em classificação de dados textuais?

METODOLOGIA:

- A investigação fundamentou-se na Mineração de Texto, uma vez que, estabelece procedimentos de extração de padrões em significativas quantidades de textos, tendo como atributo a linguagem natural.

DATA SCIENCE

Os estudos de Data Science envolvem as informações, o processo de seleção, preparação, transformação, desenvolvimento, processamento e análise de dados.

PROCEDIMENTOS METODOLÓGICOS

Fonte: Os autores (2022)

PROCEDIMENTOS METODOLÓGICOS

"A Mineração de Texto, está permeada por quatro macro etapas: "coleta, pré-processamento, indexação e análise da informação."

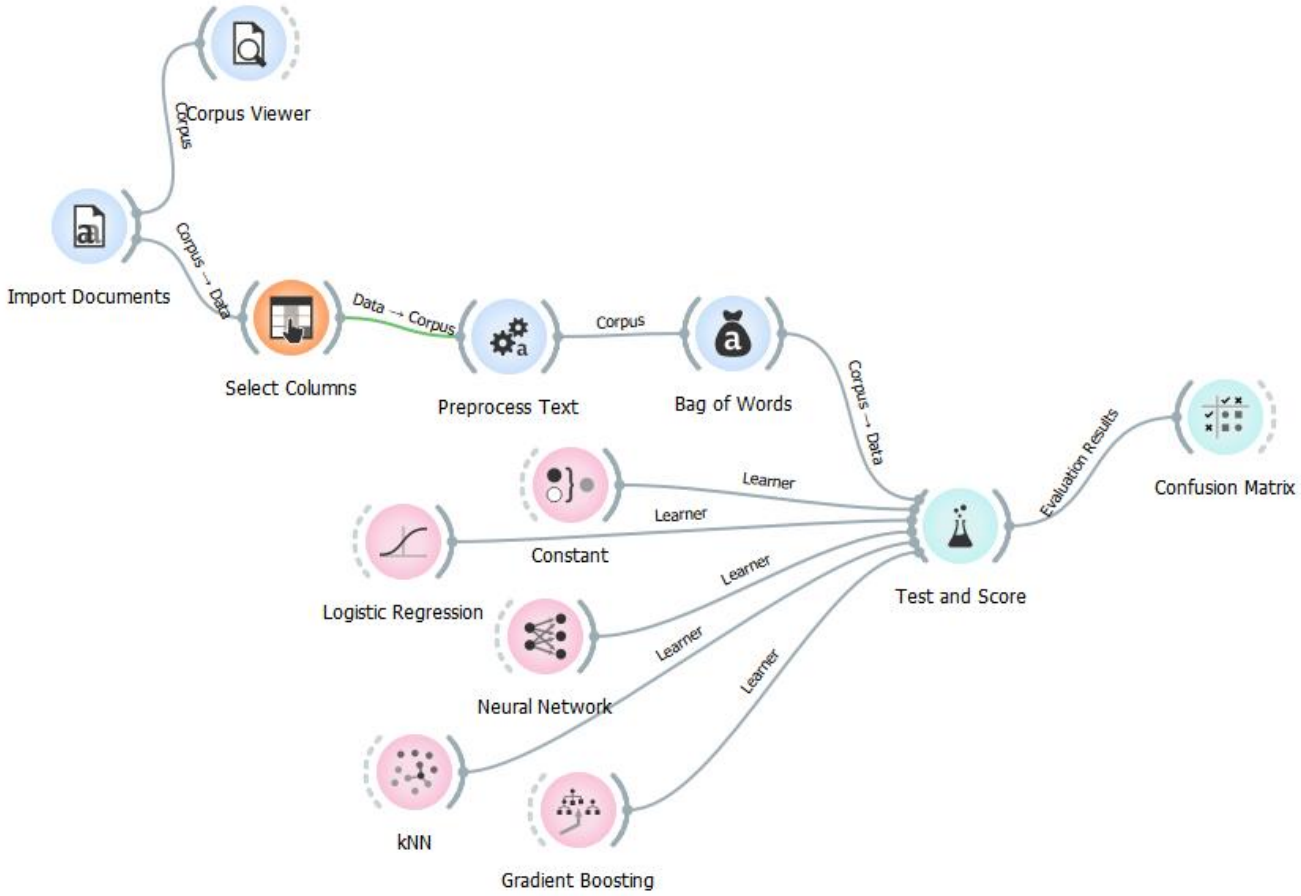
Etapa 1 – Definição do corpus da pesquisa, nesta pesquisa, foram os trabalhos publicados no Anais do ENANCIB 2021.

Etapa 2 – Pré-processamento dos dados, caracterizada por ser responsável pela construção de uma estrutura representativa dos documentos textuais. Inclui-se nesta etapa de pré-processamento a indexação dos dados, objetiva-se uma acessibilidade rápida e eficiente das informações.

Etapa 3 – Construção de uma estrutura de dados composta pelas etapas anteriores, no qual, aplica-se algoritmos de mineração de dados, com o objetivo de extrair os conhecimentos.

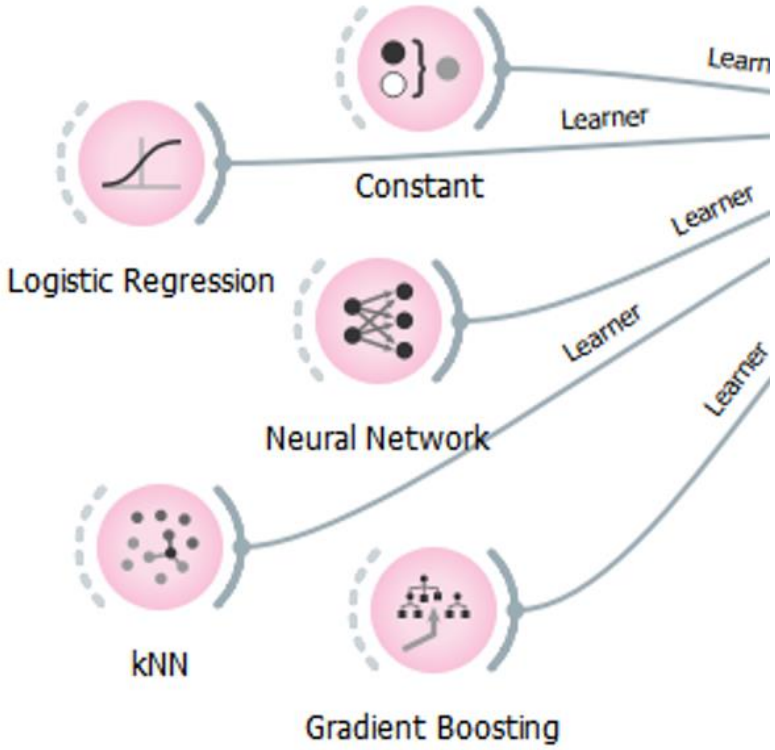
Etapa 4 – Análise da aplicação dos algoritmos, e posterior, o processo de leitura dos dados gerados.

Mapeamento do processo de execução da Mineração de Texto



Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

ALGORITMOS UTILIZADOS PARA O PROCESSAMENTO - TESTE



Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

RESULTADOS – ANÁLISE DOS ALGORITMOS

Model	AUC	CA	F1	Precision	Recall
kNN	0.932	0.952	0.733	0.733	0.733
Neural Network	0.970	0.945	0.710	0.688	0.733
Logistic Regression	0.960	0.952	0.714	0.769	0.667
Gradient Boosting	0.927	0.939	0.643	0.692	0.600
Constant	0.500	0.909	0.000	0.000	0.000

Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

ANÁLISE COM ALGORITMO LOGISTIC REGRESSION

	GT8	GT10	GT4	GT1	GT5	GT3	GT11	GT2	GT9	GT7	GT6	∑
GT8	7	0	0	1	0	0	1	1	1	3	1	15
GT10	0	10	0	0	1	1	0	0	2	1	0	15
GT4	1	0	11	0	1	0	0	0	0	0	2	15
GT1	1	0	0	10	1	1	0	0	0	0	2	15
GT5	1	1	0	1	9	1	1	0	0	1	0	15
GT3	2	0	3	0	0	7	1	0	0	0	2	15
GT11	1	0	1	0	1	0	9	0	0	2	1	15
GT2	2	0	0	2	0	0	0	11	0	0	0	15
GT9	2	2	0	0	0	0	0	0	11	0	0	15
GT7	0	0	0	0	0	0	1	0	0	13	1	15
GT6	0	0	0	2	0	1	0	0	0	1	11	15
∑	17	13	15	16	13	11	13	12	14	21	20	165

Fonte: Elaborado pelos autores usando o software Orange Canvas (2022)

ANÁLISE – LEITURA DOS DADOS

GT 1 – 66,66% artigos pertencem a sua classificação prévia;

GT 2 – 73,33% artigos pertencem a sua classificação prévia;

GT 3 – 46,66% artigos pertencem a sua classificação prévia;

GT 4 – 73,33% artigos pertencem a sua classificação prévia;

GT 5 – 60% artigos pertencem a sua classificação prévia;

GT 6 – 73,33% artigos pertencem a sua classificação prévia;

GT 7 – 86,66% artigos pertencem a sua classificação prévia;

GT 8 – 46,66% artigos pertencem a sua classificação prévia;

GT 9 – 73,33% artigos pertencem a sua classificação prévia;

GT 10 – 66,66% artigos pertencem a sua classificação prévia;

GT 11 – 60% artigos pertencem a sua classificação prévia.

Recorte de 15 artigos por GT

RESULTADOS

- ✓ Analisar informações em documentos não estruturados significa inúmeros desafios;
- ✓ A importância do pré-processamentos dos dados textuais, para qualificar análise e leitura dos algoritmos;
- ✓ Observamos que ao aumentar o recorte e a quantidade de dados testados, há a probabilidade de identificar que determinados GTs (ENANCIB) podem ter conteúdos mais aderentes entre eles, ou ainda, GTs que tenham termos mais significativos, que os colocam em situação de exclusividade de tema em relação a outros GTs;
- ✓ A Mineração de Texto, por meio de Data Science, permite a construção de ferramentas de extração e uso de dados, que têm a fortalecer as pesquisas na Ciência da Informação;
- ✓ Verificamos viabilidade de aprofundamento neste modelo de investigação.

REFERÊNCIAS

- ✓ ARANHA, Christian; PASSOS, Emmanuel. A Tecnologia de Mineração de Textos. Revista Eletrônica de Sistemas de Informação, [S.l.], v. 5, n. 2, ago. 2006. ISSN 1677-3071. Disponível em: http://periodicosibepes.org.br/index.php/reinfo /article/view/171>. Acesso em: 12 set. 2022. doi:https://doi.org/10.21529/RESI.2006.0502 001.
- ✓ ARANHA, Christian N. Uma Abordagem de Pré-Processamento Automático para Mineração de Textos em Português: Sob o Enfoque da Inteligência Computacional. 2007. 144f. Tese (Doutorado) - Programa de Pós-graduação em Engenharia Elétrica, Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, 2007. Disponível em: https://www.maxwell.vrac.pucrio.br/10081/10081_4.PDF.Acesso em 03 jun. 2022.
- ✓ BARION, E. C.N. Mineração de textos. 2008. Disponível em: https://revista.pgsskroton.com.br/index.php/rc ext/article/view/2372/2276. Acesso em: 24 nov. 2018.
- ✓ INGERSOLL, Grant S.; MORTON, Thomas S.; FARRIS, Andrew L. 2013. Taming Text: How to find, organize and manipulate it. Shelter Island, NY (USA): Manning Publications Co., 2013. 298p.

OBRIGADA!

lilianeli.sousa@uel.br

santarem@usp.br