Orange - Testes vocacionais baseados em Text Mining

Giulio Machado Falcão Silva, Anderson Silva do Nascimento

Escola de Ciência e Tecnologia – Universidade do Grande Rio Professor José de Souza Herdy (Unigranrio)
Duque de Caxias – Rio de Janeiro - Brasil

gfalcao@unigranrio.br, anderson.nascimento@unigranrio.edu.br

Abstract. *Nowadays many young people who have just graduated from high school have found it a great challenge to decide which path to follow in order to find an ideal career. This is because they do not know what the tasks or characteristics should be acquired during graduation. The mining of texts together with the most diverse algorithms aims to help the process of knowledge discovery in textual databases. In this work will be presented the use of text mining allied to the use of the logistic regression algorithm in order to provide a statistical and accurate knowledge about what are the similarities between the personal profile and the profile required for a given degree course.*

Resumo. *Hoje em dia diversos jovens, recém formados no ensino médio tem encontrado um grande desafio de decidir qual o caminho a ser seguido para encontrar a carreira ideal. Isso se dá pelo fato deles não saberem quais as tarefas ou características devem ser adquiridos durante a graduação. A mineração de textos em conjunto com os mais diversos algoritmos possui como objetivo auxiliar o processo de descoberta de conhecimento em bases de dados textuais. Neste trabalho será apresentado o uso da mineração de textos aliada ao uso do algoritmo de regressão logística com o objetivo de prover um conhecimento estatístico e preciso sobre quais são as semelhanças entre o perfil pessoal e o perfil exigido para um determinado curso de graduação.*

1. Introdução

A técnica de mineração de texto, ou text mining, é um processo de extração de conhecimento em bases de dados textuais, possuindo como objetivo analisar e explorar textos, buscando padrões, e até realizar o agrupamento desses dados através de suas características, com foco em previsões para a tomada de decisões.

De acordo com Beppler et al, KDT engloba técnicas e ferramentas inteligentes e automáticas que auxiliam na análise de grandes volumes de dados com o intuito de "garimpar" conhecimento útil, beneficiando não somente usuários de documentos eletrônicos da Internet, mas qualquer domínio que utiliza textos não estruturados.

A evasão escolar no ensino superior brasileiro é um fenômeno grave que acontece tanto nas instituições públicas quanto nas privadas e requer medidas eficazes de combate. Segundo Lobo, a taxa de abandono diminui para 7,4% quando ocorre a negociação informal entre universidade-aluno, o que estabelece relação direta entre a evasão e a possibilidade de financiamento indireto do valor.

Sendo assim, para esse trabalho serão utilizadas as técnicas de mineração de texto junto ao algoritmo de regressão logística para podermos identificar em quais perfis e características acadêmicas são encaixados os atributos pessoais dos alunos, tendo em vista que serão utilizados características e atributos pessoais. E assim resolver o seguinte problema: como auxiliar os alunos de forma precisa na escolha de um curso de graduação, para que seja possível promover a diminuição da taxa de evasão de alunos no ensino superior das universidades brasileiras?

2. Orange Canvas

A ferramenta Orange Canvas é uma ferramenta open source desenvolvida no laboratório de bioinformática na Faculdade de Ciência da Computação e Tecnologias da Universidade de Ljubljana, na Eslovénia. Trata-se de um conjunto de software compreensível e baseado em componentes para machine learning e data mining.

O Orange conta com diversas *widgets*, divididos em diferentes grupos: Dados, Visualização, Classificação, Regressão, Avaliação, Não Supervisionado e também conta com grupos integrados: Associação, Bioinformática, Rede, Mineração de texto e etc.

O Orange foi escolhido devido a dois fatores, sendo o primeiro deles a menor curva de aprendizado e a agilidade para obter resultados.

3. Arquitetura do Projeto e o Algoritmo de Regressão Logística

A arquitetura do projeto consiste em duas fases: Treinamento do modelo e classificação baseado nos resultados obtidos.

Os documentos escolhidos para o treinamento, foram feitos baseados em descrições apresentadas em sites de Universidades com nota máxima no MEC, os conteúdos descrevem os cursos de: Sistemas de Informação, Medicina Veterinária e Direito.

Os documentos foram estruturados de forma a conter informações sobre o curso de graduação e as funções desenvolvidas nas respectivas carreiras.

A primeira etapa do projeto, consistiu na criação das classes que foram utilizadas na classificação, nesta fase foram utilizados os widgets: Importar Documento, Seleção de Coluna, Concatenar, Colorir, Pré-Processamento, Bag of Word.

O algoritmo de Regressão Logística é uma técnica de modelagem preditiva utilizada para a descrição e classificação em situações em que se avalia resultados binários (0 e 1), sendo um recurso que permite buscar ou estimar uma probabilidade associada as quantidades de ocorrência de determinado evento.

O widget de Teste e Pontuação, neste cenário irá "testar" o algoritmo definido para a aprendizagem. Para informar o resultado, o mesmo conta com uma tabela com a precisão, o mesmo também conta os resultados de avaliação, que podem ser usados por outros widgets para analisar o desempenho, tais como: AUC e CA.

O método de AUC (Area Under ROC), realiza o cálculo baseado na área sob a curva ROC, dado um conjunto de resultados experimentais. Já o método CA (Classification Accuracy), realiza o cálculo percentual de correspondências entre as classes reais e as previstas.

Por fim, o widget de Predição realiza classificação e estatísticas de dados. Sendo um widget que possui várias entradas de dados sendo os principais: dados, textos e elementos customizados.

A segunda etapa do projeto, consistiu na criação da estrutura para importação dos experimentos e obtenção dos resultados. Ao final da criação da estrutura teremos o modelo contido na figura 1.

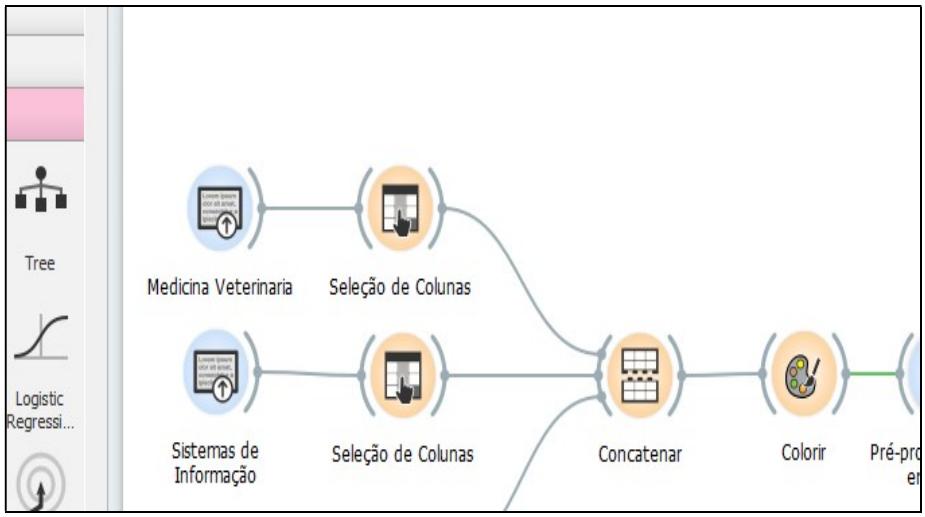

Figura 1. Modelo Completo

4. Classificação e Resultados

Na primeira etapa, os documentos foram importados para o repositório do orange, mantendo a estrutura original. Em seguida, foi inserido o widget de selecionar colunas, este widget fez a remoção de elementos originais do texto, mantendo somente o texto.

Ao inserir o widget de concatenar, foi possível criar uma identificação para os arquivos, ou seja, o primeiro arquivo é indexado com o valor "untitled 0", deste modo, tornando fácil a identificação dos documentos. Logo em seguida, utilizando o widget colorir, foram inseridos os nomes das classes. O widget de pré-processamento atuou realizando a limpeza e retirada de caracteres especiais e números.

O widget de Bag of Word, foi adicionado, devido a necessidade de se contar a quantidade de palavras contidas nos textos, distribuindo-as para cada bolsa de palavras.

Este widget atendeu a necessidade devido ao seu método de análise, contando a frequência em que as palavras se repetem.

Ao realizar todo o processo de importação, foi feito o teste da aplicação do algoritmo, utilizando o widget de teste e pontuação, tendo como principais indicadores AUC e CA.

Ao final da classificação dos experimentos obtemos os resultados expostos na figura 2:

	Logistic Regression	content
1	0.26 : 0.41 : 0.34 → Sistemas de Informação	Classificação 1
2	0.31 : 0.42 : 0.26 → Sistemas de Informação	Classificação 2
3	0.34 : 0.27 : 0.39 → Direito	Classificação 3

Figura 2. Classificação utilizando o algoritmo Regressão Logística.

5. Conclusão e Trabalhos Futuros

Este trabalho apresentou a técnica de mineração de texto junto ao algoritmo de regressão logística, este conhecido como algoritmo de aprendizado de máquina. Foram analisados dois indicadores para a predição, sendo eles a Curva ROC e Acurácia, apresentando ao fim do processo o percentual de proximidade entre as classes usadas para o treinamento e os experimentos. Este trabalho teve como principal objetivo apresentar estatísticas

reais e previsões cada vez mais confiável, visando apresentar um novo método de se decidir qual curso de graduação o aluno pode escolher.

Foram criadas duas estruturas, onde ambas consistiram em apresentar um modelo de treinamento, sendo uma para o treinamento do modelo e a outra permitindo importar os documentos para obtenção das classificações desejadas, obtendo bons resultados em relação a acurácia e a área da curva roc, deste modo classificando os documentos de maneira correta. Com estes resultados tivemos certeza de que o algoritmo de regressão logística tenha sido o melhor para realizar este trabalho.

Como trabalhos futuros, foi pensado na ideia da implementação do algoritmo de redes neurais, tornando assim os resultados cada vez mais eficientes. Outra sugestão de trabalhos futuros poderia ser aumentar a quantidade de classes, tornando um modelo cada vez mais robusto e dispondo assim de diversos outros caminhos para a classificação dada pelo algoritmo, fazendo com que o estudo fique ainda mais real, com um nível maior de confiança.

Referências

Américo, Pedro Amaral. (2017) 'Comparação entre Regressão Logística e Redes Neurais em Aplicação de Reconhecimento De Dígitos Manuscritos', documento de graduação apresentada a Instituto Federal de Educação, Ciência e Tecnologia Fluminense.

BARROS, P. Aprendizagem de Maquina: Supervisionada ou Não Supervisionada? Medium, 2016. Disponivel em: https://medium.com/opensanca/aprendizagem-de-maquina-supervisionada-ou-n%C3%A3o-supervisionada-7d01f78cd80a.

BEPPLER, M; FERNANDES, A. Aplicação de text mining para a extração de conhecimento jurisprudencial. In: PRIMEIRO CONGRESSO SUL CATARINENSE DE EDUCAÇÃO, 2005.

LOBO, RICARDO (2018). A Evasão no Ensino Superior Brasileiro – Novos Dados. Disponível em: https://educacao.estadao.com.br/blogs/roberto-lobo/497-2/

ROMAO, L, Análise do uso de técnicas de pré-processamento de dados em algoritmos para classificação de proteínas General Terms. Universidade da Região de Joinville (Univelle). Joinville, p. 5. 2016.