Ciência de Dados para Todos (Data Science For All) - 2018.2 - Relatório dos Programas de Pós Graduação de Educação - Departamento de Ciência da Computação da UnB

Bruno Helder Rodrigues Guedes, Gabriel Almeida Campos, Thiago Luis Rodrigues Pinho

17/11/2018

Introdução

Este documento apresenta uma análise exploratória de dados retirados da plataforma e-Lattes e da plataforma Sucupira para a construção do relatório final da disciplina Ciência de Dados para Todos (Data Science For All), do Departamento de Ciência da Computação da Universidade de Brasília.

Os dados utilizados na análise estão relacionados aos programs de pós-graduação de: Educação, Educação Física e Ciências de Reabilitação da Universidade de Brasília.

A metodologia para desenvolvimento do relatório é baseada no modelo de mineração de dados denominado CRISP-DM (Chapman et al., 2000, Mariscal et al., 2010).

Este documento deve ser referenciado do modo como aparece na seção de referências ao final do texto, abaixo reproduzida.

Bruno Helder Rodrigues Guedes, Gabriel Almeida Campos, Thiago Luis Rodrigues Pinho. "Ciência de Dados para Todos (Data Science For All) - 2018.2 - Relatório dos Programas de Pós Graduação de Educação - Departamento de Ciência da Computação da UnB". Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2018.2, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 17 de novembro de 2018.

1 - Busca dos dados

Os dados utilizados na análise contida neste relatório foram obtidas na plataforma e-Lattes e na plataforma Sucupira, sendo estes parte dos programas de pós-graduação de Educação, Educação Física e Ciências de Reabilitação.

2 - Contextualização dos Programas

3 - Descrição da Metodologia

Nesta sessão iremos descrever e analisar o processo de exploração, pesquisa e desenvolvimento utilizados para a criação deste projeto, o CRISP-DM. CRISP DM é abreviação de Cross-industry standard process for data mining que é um padrão industrial aberto de processos para abordar trabalhos de data mining e análise de dados em geral. Esta metodologia basicamente se divide em seis etapas: Business Understanding , Data Understanding, Data preparation , Modeling , Evaluation e Deployment. Iremos descrever casa uma das etapas mais detalhadamente à seguir. Business Understanding

A primeira etapa do processo é talvez a mais abstrata de todas, que é o conhecimento sobre o problema. O objetivo é que se avalie os custos e os impactos da solução a ser proposta. Nesta etapa também, são definidos os objetivos e metas do processo. Neste projeto estamos analisando os dados relativos aos Programas de Pós-Graduação da Universidade de Brasília. À partir disso desejamos realizar uma análise exploratória sobre os dados para oferecer novas perspectivas sobre estes programas.

Data Understanding

Data Understanding, que pode ser traduzido para Entendimento dos Dados, é a etapa que consiste em avaliar os dados já disponíveis. Deve-se organizar e descrever os dados obtidos, e, avaliar se é possível atingir os objetivos determinados na etapa anterior com estes dados. Esta etapa tem um teor também exploratório, pois é necessário avaliar que tipo de análise é possível realizar com os dados disponíveis e tomar decisões baseadas nesta análise. Por exemplo, é possível determinar que é necessário realizar o processo de obtenção dos dados novamente, ou retornar para a etapa de Business Understanding.

Neste projeto, temos acesso tanto aos dados disponibilizados pelo professor, em diferentes arquivos em formato JSON, com dados sobre todos os projetos de pós-graduação da UnB, e também os dados disponíveis no E-lattes. O arquivo profile.json contém informações referentes ao perfil dos docentes, identificados pelo número de matrícula e contém nome, resumo do currículo, áreas de atuação, endereço profissional, produção bibliográfica que contém os capítulos de livro, eventos, livros e artigos em periódicos publicados pelo docente. O arquivo publication.json tem informações referentes as publicações da determinada área no período entre 2010 e 2017.

Data Preparation

A etapa de Data Preparation, ou Preparação dos dados, é uma etapa de cunho técnico que tem como objetivo realizar a limpeza dos dados para o processo de modelagem. Dependendo da maneira de como foi realizada a coletagem de dados, esta etapa pode se tornar menos ou mais relevante. Nesta etapa também é realizada a construção de características derivadas, ou seja, dados que são possíveis ser inferidos dos dados disponíveis.

Tomando em conta o projeto aqui sendo desenvolvido, esta etapa é menos relevante pois já obtemos os dados semi-filtrados, que podem ser facilmente utilizados no ambiente de desenvolvimento utilizando a linguagem R, restando apenas selecionar as caracterísitcas a serem utilizadas.

Modelagem

Nesta etapa é aplicado as técnicas de Data Science e Data Analisys para construir modelos adequados para o contexto da situação e consequentemente tirar conclusões ou proposições para as soluções do problema proposto na etapa de Business Understanding.

No projeto em questão, iremos ((inserir o que vai ser feito no projeto)).

Evaluation

Fase de avaliação dos resultados obtidos na etapa de modelagem e verificando se cumpre os objetivos definidos na etapa de business understanding.

Deployment

Fase de implantação de fato dos resultados obtidos nas análises realizadas e assim concluindo o processo de desenvolvimento.

4 - Resultados Descritivos

5 - Modelos de Análise