Disciplina Ciência de Dados Aplicada e Ciência de Dados para Todos

Relatório 1 – Importação e Limpeza de Dados

Autor: Pedro Henrique Luz de Araujo Data: 06/04/2018

1. Introdução

Os dados podem ser encarados como observações ou medições de eventos no mundo. A partir de dados organizados e estruturados podemos extrair compreensões e entendimentos sobre a realidade descrita pelos dados ou até mesmo construir modelos de explicação dessa realidade.

Diante desse potencial de construção do conhecimento, a análise dos dados relacionados à produção acadêmica dos professores da Universidade de Brasília podem informar e fornecer *insights* que auxiliem a elaboração de políticas internas da Universidade que promovam o desenvolvimento científico da UnB. Serão avaliados os currículos depositados na plataforma Lattes, composto de três bases de dado, além de dados sobre grupos de pesquisa da UnB, a partir de base do Diretório de Grupos de Pesquisa (DGP).

Inicialmente, busca-se fazer uma avaliação descritiva das bases mencionadas com o objetivo de melhor entendê-las.

2. Metodologia

Foi utilizada a plataforma RStudio para a utilização da linguagem R, juntamente com o pacote jsonlite para a importação de dados armazenados em arquivo json. Para importação e limpeza dos dados salvos como tabela do Excel, usou-se o pacote readxl.

As funções summary e str foram usadas para se obter uma descrição das variáveis e respectivas classes das bases analisadas. Por outro lado, as funções names e length foram utilizadas para analisar os dados estruturados em formato json. Por fim, foi utilizada a função sapply para a criação recursiva de vetores de dados.

3. Resultados

O primeiro arquivo, unb.perfis.json, foi lido a partir da função fromJSON e é composto de 1592 currículos, que são lidos em formato de lista. A função names retorna o nome de cada currículo, os quais são strings de caracteres numéricos que identificam cada professor de maneira única. Cada currículo contém como variáveis nome do professor, resumo do currículo, áreas de atuação, especialidade, endereços, entre outras.

O segundo arquivo, unb.relatorioProducaoBibliografica.json, também importado com fromJSON, contém publicações de periódicos, livros, capítulos de livros, textos em jornais e artigos aceitos, no período de 2012 a 2017. Tal informação foi obtida pela chamada de função sapply(prod, names), em que prod é a lista retornada pela função fromJSON.

O arquivo unb.relatorioOrientacao.json foi igualmente importado usando fromJSON. Contém orientações em andamento a nível de graduação, iniciação científica, graduação, mestrado, doutorado e pós-doutorado; além de orientações concluídas a nível de mestrado, doutorado e pós-doutorado. O período de análise também está compreendido entre os anos 2012 e 2017. A partir da chamada sapply(orientacoes$ORIENTACAO_CONCLUIDA_MESTRADO, function(x) length(x$titulo)), foi possível obter a quantidade de orientações de mestrado concluídas em cada ano, como demonstra a figura 1.

Finalmente, usou-se o pacote readxl para importar a quinta folha da planilha com dados do DGP. O conjunto de dados consiste em 2949 observações de 10 variáveis, podendo ser extraídas informações relativas à quantidade de publicações realizadas por cada grupo de pesquisa nos anos 2000, 2002, 2004, 2006, 2008, 2010, 2014 e 2016, informação obtida pelo código sort(unique(dpg$”Ano Censo”)). Os tipos de publicação compreendem artigos completos de circulação internacional, artigos completos de circulação nacional, capítulos de livro, dissertações, livros, produção técnica, teses e trabalhos completos publicados em anais.

Tais análises contribuiram para o entendimento a alto nível dos dados da plataforma lattes, quanto aos currículos, publicações e orientações, e dos grupos de pesquisa, principalmente em relação à forma como foram estruturados.

Figura 1