Disciplina Ciência de Dados Aplicada e

Ciência de Dados para Todos

Relatório 1 – Importação e Limpeza de Dados

Autor: Luis Gustavo Avelino de Lima Jacinto Data: 07/04/2018

1. Introdução

A ciência de dados é uma área interdisciplinar voltada para o estudo e a análise de dados, estruturados ou não, que visa a extração de conhecimento para possíveis tomadas de decisão.

Diante disso e de acordo com os dados disponibilizados, buscou-se avaliar a produção científica de professores, publicações, orientações e teses e dissertações da UnB. Os resultados são dados descritivos com objetivo de compreender os dados disponíveis nessas bases de dados.

2. Metodologia

Para realizar a análise dos dados foi utilizado o programa RStudio e o pacote jsonlite para importação e limpeza dos dados.

Para melhor entendimento e visualização dos dados foram utilizados as funções *str() e summary()*, que proporcionaram uma visão geral dos dados, e também as funções l*ength(), class() e names()* para entender melhor a estrutura dos dados que estavam em formato json. Por fim, foram utilizadas as funções *lapply() e sapply()* para a criação de novas estruturas e a formação de subconjuntos

3. Resultados

A importação de todas as bases de dados foram por meio do método *read_json()*.

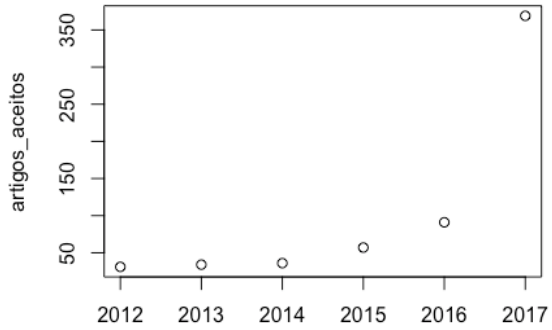
3.1. Perfils Professores UnB

O primeiro arquivo, *unb_perfis_json*, contém 1592 observações, que pode ser visto através do método: *length(unb_perfis_json)*. Por meio da função *names(unb_perfis_json),* pode-se perceber que cada observação tem uma string composta de 16 caracteres numéricos que identifica a observação. E, por meio da função: *summary()*, pode-se observar que cada observação é divida em 7 atributos principais: **nome, resumo_cv, areas_de_atuacao, endereco_profissional, producao_bibliografica, orientacoes_academicas e senioridade**.

3.2. Publicações Professores UnB

No segundo arquivo, *unb_relatorio_plublicao_json*, temos os dados referentes a publicações de professores em: **periódicos, livros e capítulos de livros, texto em jornais e artigos aceitos**, isso pode ser visto utilizando a função *names()*. Para os artigos aceitos pode-se perceber que o número vem crescendo desde 2012, como pode ser visto por meio da função *plot(anos, artigos_aceitos),* em que *anos*

<- names(artigos_aceitos) e *artigos_aceitos <- lapply(unb_relatorio_plublicao_json$ARTIGO_ACEITO, length)*.



3.3. Orientações Professores UnB

O terceiro arquivo, *unb_relatorio_orientacao_json*, pode-ser perceber, por meio das funções **length(unb_relatorio_orientacao_json) e names(unb_relatorio_orientacao_json)** , que os dados são dividios em 8 observações quer seriam orientação em andamento de pós doutorado, doutorado, mestrado, graduação e iniciação científica e orientação concluída de pós doutorado, doutorado e mestrado. Para as orientações concluídas de mestrado tem os números 760, 933, 877, 817, 777, 463 para os anos de 2012 a 2017, respectivamente. Esse dado pode ser avaliado por meio da função *sapply(unb_relatorio_orientacao_json$ORIENTACAO_CONCLUIDA_MESTRADO, length)*

3.4. Teses e Dissertações da UnB

O quarto arquivo, *unb_teses_dissertacoes_json*, tem os dados referentes a teses e dissertações na unb. O arquivo contém 18462 observações, que pode ser vista através da função *length(unb_teses_dissertacoes_json$response$docs)*. As observações contém informações como descrição, linguagem, data de publicação, título, autor, entre outras, referentes a tese ou dissertação.