Disciplina Ciência de Dados Aplicada e Ciência de Dados para Todos

Relatório 1 – Importação e Limpeza de Dados

Autor: Ricardo Barros Sampaio Data: 04/04/2018

1. Introdução

"A ciência pode ser definida como o estudo metodicamente organizado de quaisquer fenômenos que ocorrem no universo com a finalidade de explicar e prever o comportamento e a estrutura de tais fenômenos." (Fernandes e Sampaio 2017).

A Ciência da Ciência (SciSci) como uma nova área de pesquisa (Fortunato et al. 2018) oferece uma compreensão quantitativa das interações entre agentes científicos, como autores, projetos, publicações e outros, em diversas escalas com potencial para fornecer insights sobre as condições subjacentes à criatividade e à gênese da descoberta científica.

Diante dessa nova perspectiva de pesquisa e com o arcabouço tecnológico e de conhecimento disponibilizado pela Ciência de Dados, buscou-se avaliar a produção científica de professores da Universidade de Brasília por meio dos seus currículos depositados na plataforma Lattes. Além desses foram avaliadas as bases do Diretório de Grupos de Pesquisa (DGP), da Biblioteca Digital Brasileira de Teses e Dissertações (BDTD) e do repositório OÁSIS.

Os resultados apresentados são avaliações descritivas nesse primeiro momento com o objetivo de se entender o negócio, que é o desenvolvimento científico nacional, e os dados disponíveis por meio das bases acima citadas.

2. Metodologia

Para a análise dos dados foi utilizado o software RStudio e os pacotes jsonlite, readxl para importação e tratamento dos dados.

Para o entendimento e visualização das informações contidas nas bases foram utilizadas as funções summary, str para uma visão geral dos dados e as funções names, unlist e length para um melhor entendimento da estrutura dos dados que estavam em formato json. No caso da análise recursiva utilizou-se também as sunções lapply e sapply para a criação de novas estruturas e a formação de subconjuntos.

3. Resultados

Foram importados três arquivos relacionados com o currículo lattes dos professores da UnB. O primeiro unb.perfis.json com 1.592 elementos ou currículos; no segundo documento estavam as orientações contendo o período de 2012 a 2017 para as orientações de iniciação científica, mestrado, doutorado e pós-doutorado em andamento e concluídas, e o terceiro documento contendo publicações em periódicos, livros e capítulos de livros, texto em jornais e artigos cadastrados como aceitos. Para as publicações em periódicos tivemos um número de 1600, 1664, 1735, 1826, 1953 e 1028 de 2012 a 2017 respectivamente que pôde ser avaliado utilizando a função sapply(biblio$PERIODICO, length).

No que diz respeito aos dados do DGP foi importada a quarta folha da planilha com o código read_excel("UnB - Ciencia de Dados 01_2018.xls", sheet = 4). Com um total de 53.984 observações e 8 variáveis a folha da planilha continha os dados dos participantes e os seus respectivos grupos. Do total de participantes 12.615 eram do sexo feminino e 10.623 do sexo masculino calculados após ser retirado a duplicação de nomes nos grupos. A tabela abaixo demonstra o número de participantes nos grupos por tipo de participação antes e depois da deduplicação demonstrando que os pesquisadores tendem a em média se afiliar a mais de 3 grupos de pesquisa e os demais tipos a menos de 2 grupos:

COLABORADOR				
ESTRANGEIRO	ESTUDANTE	PESQUISADOR	TECNICO	
236	27031	24359	2358	
COLABORADOR				
ESTRANGEIRO	ESTUDANTE	PESQUISADOR	TECNICO	
153	15577	7294	1078	

A análise até este momento nos permitiu entender a estrutura dos dados do currículo lattes e como esta pode se relacionar com as diferentes bases de grupos de pesquisa, orientação ou demais publicações.