Disciplina Ciência de Dados Aplicada e Ciência de Dados para Todos - Relatório 1 – Importação e Limpeza de Dados

Gabriel Martins de Miranda

April 07, 2018

1. Introdução

Com cada vez mais dados sendo gerados pelo mundo a fora, com seus diversos significados, torna-se difícil gerenciar e obter informação útil a partir deles. Diante da percepção da importância competitiva que estes dados podem vir a oferecer, grande demanda vêm sendo gerada por empresas em busca de pessoas capacitadas em extrair informação de grande volume de dados, sendos as áreas de Data Science e Big Data focadas neste objetivo.
Diante desta perspectiva, o presente trabalho buscou analisar dados e extrair informações quantitativas e qualitativas úteis em bases de dados referentes ao escopo da UnB. Foram analisadas quatro bases, sendo a de orientação dos professores, publicações dos professores, do diretório de grupos de pesquisa e do repositório de publicações OASIS.

2. Metodologia

A plataforma utilizada foi o RStudio e sua linguagem de programação R. Os pacotes envolvidos foram o jsonlite para o tratamento dos dados em json e readxl para o excell. Para o relatório, foi utilizado um documento R Markdown, cuja extensão é Rmd e gerado um arquivo em pdf.

3. Resultados

A primeira base de dados, **orient_prof_unb_json**, importada com *fromJSON("orientacoes_professores_unb.json")*, representa as orientações realizadas pelos professores da UnB concluídas ou em andamento nos cursos de graduação, mestrado, doutorado, pós-doutorado e iniciação científica, *summary(orient_prof_unb_json)*. Com *sapply(orient_prof_unb_json[["ORIENTACAO_CONCLUIDA_POS_DOUTORADO"]], function (x) unlist(x$"ano"))* foi possível observar que os dados de doutorado concluídos foram para os anos de 2012 a 2017, informação que pode ser replicada para as outras orientações. A seguir é possível observar os seis primeiros alunos de graduação com orientação em andamento de 2017 em ordem alfabética.

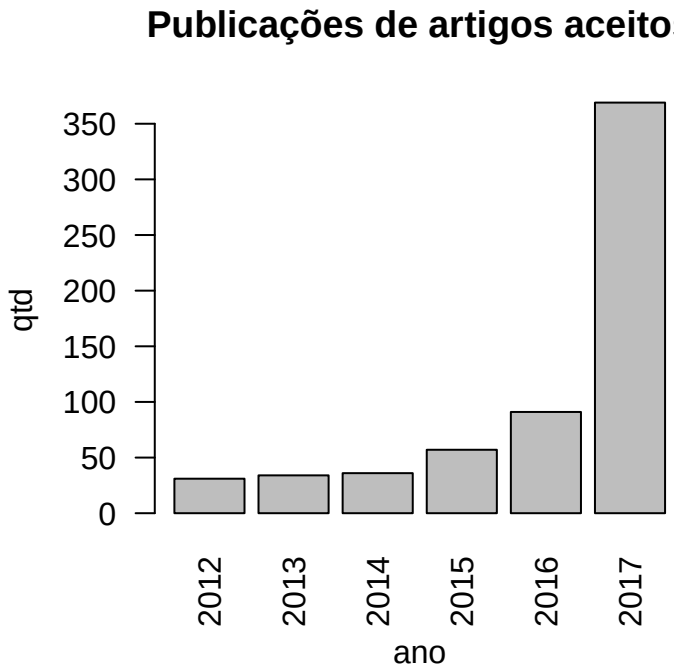
```
head(sapply(orient_prof_unb_json[["ORIENTACAO_EM_ANDAMENTO_GRADUACAO"]][["2017"]],
            function (x) {x$"nome_aluno"}) %>% sort())
```

```
## [1] "ADONAI PADILHA"
## [2] "Adrielly Nascimento"
## [3] "Afonso Henrique Dutra Manso"
## [4] "Alana Luysla Silva Lima"
## [5] "Alessandra de Souza Carvalho"
## [6] "Alexandre Correia Mesquita Oliveira"
```

A segunda base de dados, **pub_prof_unb_json**, importada com *fromJSON("publicacoes_professores_unb.json")*, representa as publicações dos professores da UnB em periódicos, livros, capítulos de livros, textos em jornais e artigos, como visto por meio de *summary*. Novamente com *sapply* na variável ano foi possível observar que

os anos em questão são de 2012 a 2017. A seguir são mostrados o número de artigos aceitos por ano da base em questão.

```
sapply(pub_prof_unb_json[["ARTIGO_ACEITO"]], function (x) length(x)) %>%
  barplot(main="Publicações de artigos aceitos", xlab="ano", ylab="qtd", las=2)
```



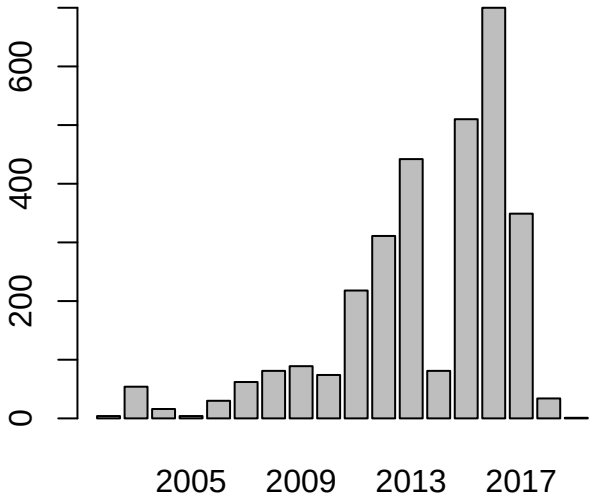
Para a terceira base, **dir_gru_pesq_unb_xls**, a primeira folha da planilha foi importada com *read_excel("diretorio_grupos_pesquisa_unb.xls", sheet = 1, skip = 2)*, sendo que possui as colunas de token do grupo de pesquisa, ano do censo, nome do grupo, nome da grande área, nome da área e ano de formação, *names(dir_gru_pesq_unb_xls)*. Com *summary(dir_gru_pesq_unb_xls)*, foi descoberto que o dataset possui 2955 observações e que o ano de formação mínimo foi 1970 e o máximo 2016. A seguir, é possível ver as áreas com o maior número de grupos de pesquisas já criados.

```
dir_gru_pesq_unb_xls %>% select(`Nome Área`) %>% group_by(`Nome Área`) %>%
  mutate(Número = n()) %>% unique() %>% arrange(desc(`Número`)) %>% head()
```

```
## # A tibble: 6 x 2
## # Groups:   Nome Área [6]
##   `Nome Área`           Número
##   <chr>                  <int>
## 1 Educação                 169
## 2 Psicologia               122
## 3 Química                  111
## 4 Administração            105
## 5 Letras                   102
## 6 Ciência da Informação     97
```

A quarta e última base, **pub_rep_oasis**, importada com *read.table("publicacoes_repositorio_oasis.txt", sep = "", fill = T, header = TRUE)*, representa o repositório de publicações OASIS e é composto por 18 variáveis, sendo id, titulo, autor, tópico, descrição, data de publicação, lingua e outros, *names(pub_rep_oasis)*, sendo 3060 observações, *nrow(pub_rep_oasis)*. A seguir é possível observar o número de publicações por ano.

```
plot(pub_rep_oasis$publishDate)
```



A presente análise realizada neste trabalho permitiu abstrair informações de valor de variadas bases de dados acadêmicas, principalmente do escopo da Universidade de Brasília.