Disciplina Ciência de Dados Aplicada e Ciência de Dados para Todos - Relatório 2 – Importação e Limpeza de Dados

Gabriel Martins de Miranda

April 29, 2018

1. Introdução

O presente relatório realiza análises quantitativas e qualitativas nos dados disponíveis na base e-lattes da unb, em especial para o curso de Engenharia elétrica, sendo **unb_ePERFIS**, que contém informações dos perfis dos professores do departamento na base lattes, **unb_eORIENT**, que contém dados de orientações de pós-doutorado, doutorado, mestrado e graduação e **unb_ePUB**, que contém as publicações em livros, periódicos, revistas e outros. Além disso foi usada a base do DGB, **unb_DGP_df**, que contém dados de grupos de pesquisa realizados em todo o Brasil. Os dados do e-lattes foram todos coletados desde 2010 a 2017. Já o censo DGP, na folha 4 que foi utilizada, tem dados que vão de 2000 a 2016.

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 a análise dos dados foi usado o pacote tidyverse, que inclui os pacotes dplyr, ggplot2, purr. Para o relatório, foi utilizado um documento R Markdown, cuja extensão é Rmd e gerado um arquivo em pdf.

3. Resultados

3.1 Importações

A importação do DGP foi a mais simples, já que foi possível importar diretamente como dataframe. Foi decidido a importação de apenas a página 4. Com **glimpse** é possível ter uma ideia geral dos dados, já que engloba informações como **names**, **dim** e **str**. A importação e informações dos dados do DGP são mostradas a seguir.

```
# unb_DGP
unb_DGP_df <- read_xls("unb.DGP.xls", sheet = 4)
glimpse(unb_DGP_df)
```

```
## Observations: 53,984
## Variables: 8
## $ `Token Grupo Pesquisa`            <chr> "0025913153775147", "0025913...
## $ `Ano Censo`                       <chr> "2014", "2016", "2014", "201...
## $ `Nome Participante`               <chr> "Anna Francisca Salles Marqu...
## $ `Ano Nascimento Participante`     <dbl> 1994, 1994, 1995, 1995, 1988...
## $ `Tipo Participação`               <chr> "ESTUDANTE", "ESTUDANTE", "E...
## $ `Nível Formação Participante`     <chr> "Graduação", "Graduação", "G...
## $ `País de Nascimento Participante` <chr> "Brasil", "Brasil", "Brasil"...
## $ `Género Participante`             <chr> "Feminino", "Feminino", "Fem...
```

Para a importação dos dados de publicações dos professores, foi decidido que seriam usadas apenas as publicações ocorridas em eventos. Abaixo esta a importação, assim como sua estrutura.

```
# unb_ePUB
unb_ePUB <- fromJSON("unb.Pub.EngEletrica.json")
unb_ePUB_df <- data.frame()
for (i in 1:length(unb_ePUB$EVENTO))
  unb_ePUB_df <- rbind(unb_ePUB_df, unb_ePUB$EVENTO[[i]])
glimpse(unb_ePUB_df)
```

```
## Observations: 966
## Variables: 11
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "Stable Sampling Period Adaptation for Ener...
## $ nome_do_evento     <chr> "XVIII Congresso Brasileiro de Automática, ...
## $ ano_do_trabalho    <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ pais_do_evento     <chr> "Brasil", "Brasil", "Portugal", "Portugal",...
## $ cidade_do_evento   <chr> "Bonito", "São Paulo", "Lisbon, Portugal", ...
## $ doi                <chr> "", "", "", "", "", "", "", "", "", "", "",...
## $ classificacao      <chr> "NACIONAL", "INTERNACIONAL", "INTERNACIONAL...
## $ paginas            <chr> " - ", "1 - 5", " - ", " - ", "1 - 5", "214...
## $ autores            <list> [<"Steffen, T.", "Ishihara, J.Y.", "Bauchs...
## $ `autores-endogeno` <list> ["0301021863146083", "0363748060813357", "...
```

Para os dados de orientações, todas as orientações foram importadas, desde pós-doutorado a graduação. Abaixo está o código da importação.

```
# unb_eORIENT
unb_eORIENT <- fromJSON("unb.Orientacao.EngEletrica.json")
unb_eORIENT_df <- data.frame()
for (i in 1:length(unb_eORIENT))
  for (j in 1:length(unb_eORIENT[[1]]))
    unb_eORIENT_df <- rbind(unb_eORIENT_df, unb_eORIENT[[i]][[j]])
glimpse(unb_eORIENT_df)
```

```
## Observations: 1,064
## Variables: 13
## $ natureza                    <chr> "Supervisão de pós-doutorado", "Te...
## $ titulo                      <chr> "", "SISTEMAS DE MODULAÇÃO EM TECN...
## $ ano                         <chr> "2017", "2011", "2011", "2011", "2...
## $ id_lattes_aluno             <chr> "", "8125419373679302", "", "36679...
## $ nome_aluno                  <chr> "RICARDO KEHRLE MIRANDA", "GENIVAL...
## $ instituicao                 <chr> "Universidade de Brasília", "Unive...
## $ curso                       <chr> "", "ENGENHARIA DE SISTEMAS ELETRÔ...
## $ codigo_do_curso             <chr> "", "60057840", "60057840", "60057...
## $ bolsa                       <chr> "SIM", "NAO", "SIM", "NAO", "NAO",...
## $ agencia_financiadora        <chr> "Coordenação de Aperfeiçoamento de...
## $ codigo_agencia_financiadora <chr> "045000000000", "", "045000000000"...
## $ nome_orientadores           <list> ["Joao Paulo Carvalho Lustosa da ...
## $ id_lattes_orientadores      <list> ["1786889674911887", "11132346902...
```

A importação mais complexa foi a dos perfis dos professores, já que muitos dos dados internos são do formato lista pura. Para todos os perfis de professores, foram importados dados de nome, resumo do currículo, áreas de atuação e senioridade.

```
# unb_ePERFIS
unb_ePERFIS <- fromJSON("unb.Perfis.EngEletrica.json")
```

```
unb_ePERFIS_df <- data.frame()
for (i in 1:length(unb_ePERFIS))
{
  temp <- data.frame(nome = unb_ePERFIS[[i]]$nome,
                     resumo_cv = unb_ePERFIS[[i]]$resumo_cv,
                     unb_ePERFIS[[i]]$areas_de_atuacao,
                     senioridade = unb_ePERFIS[[i]]$senioridade)
  unb_ePERFIS_df <- rbind(unb_ePERFIS_df, temp)
}
glimpse(unb_ePERFIS_df)
```

```
## Observations: 220
## Variables: 7
## $ nome          <fct> Ronaldo Sergio Chacon Camargos, Ronaldo Sergio C...
## $ resumo_cv     <fct> Possui graduação em Engenharia Elétrica (2013) e...
## $ grande_area   <chr> "ENGENHARIAS", "ENGENHARIAS", "ENGENHARIAS", "EN...
## $ area          <chr> "Engenharia Elétrica", "Engenharia Elétrica", "E...
## $ sub_area      <chr> "Sistemas Elétricos de Potência", "Energia Solar...
## $ especialidade <chr> "", "", "Controle de Processos Eletrônicos, Retr...
## $ senioridade   <fct> 0, 0, 2, 2, 2, 2, 2, 2, 5, 5, 7, 2, 2, 2, 2, 2, ...
```

3.2 Análises

Para o grupo de pesquisa DGB, é possível observar que existem muitas duplicações dos dados dos participantes. O número distinto de participantes é dado pelo comando a seguir. Das 53984 observações do dataset, caímos para 21888.

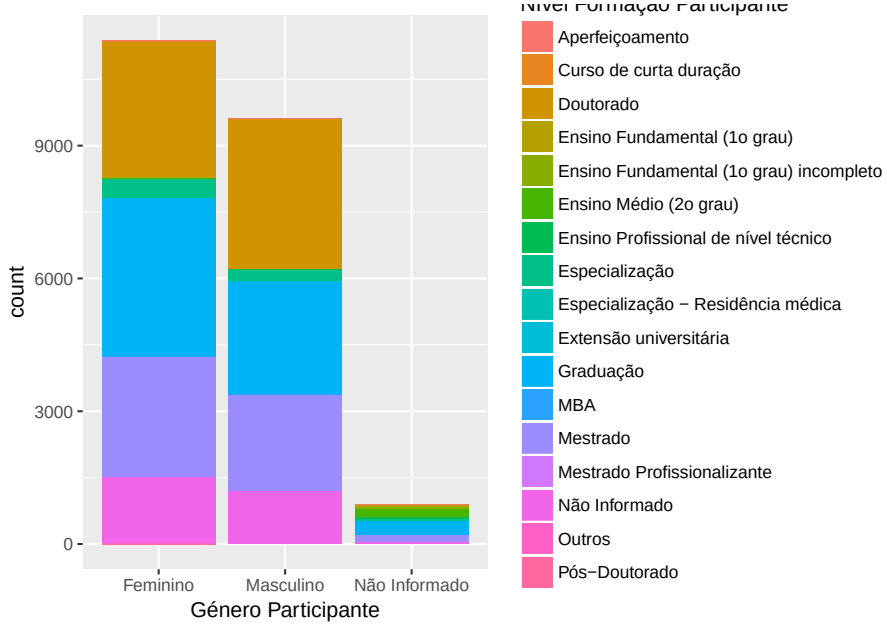
```
  unb_DGP_df %>% select(`Nome Participante`) %>% distinct %>% nrow
```

```
## [1] 21888
```

Destes participantes, podemos diferenciar por gênero e pelo nível de formação e plotar a seguir. Podemos ver que grande parte dos participantes é de doutorado, mestrado ou graduação.

```
  df <- unb_DGP_df %>%
  select(`Nome Participante`, `Género Participante`, `Nível Formação Participante`) %>%
  distinct(`Nome Participante`, .keep_all = TRUE)

  ggplot(data = df) +
  geom_bar(mapping = aes(x = `Género Participante`,
                         fill = `Nível Formação Participante` ))
```



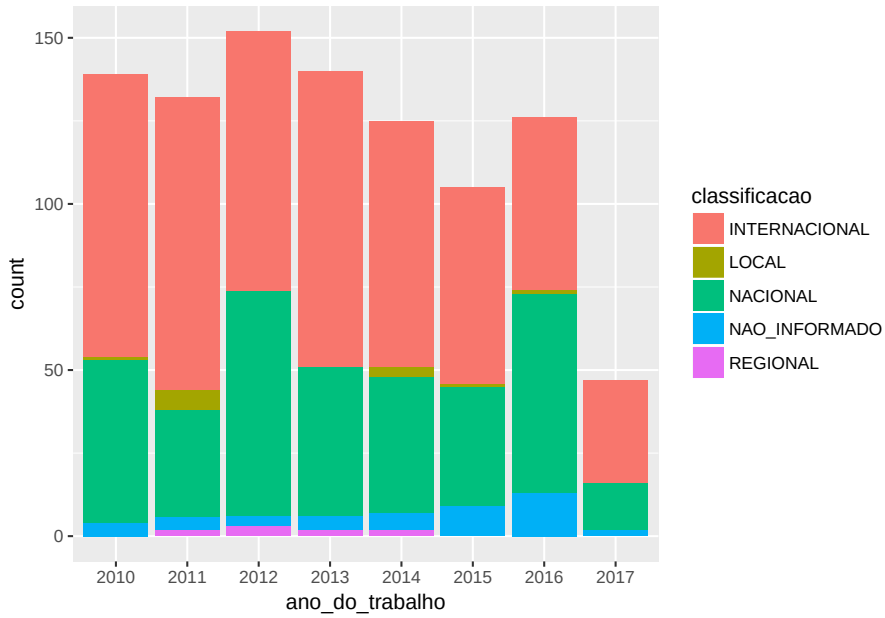
Para as publicações dos professores de engenharia elétrica em eventos, podemos determinar em quais deles os professores da UnB estão mais engajados e publicam mais com o comando a seguir. Como resultado, observamos que o XXIII CBEB foi o mais publicado dentro os professores, em Engenharia Biomédica.

```
  df <- group_by(unb_ePUB_df, nome_do_evento) %>%
  summarise(count = n())
  head(arrange(df, desc(count)), 10)
```

```
## # A tibble: 10 x 2
##    nome_do_evento                                              count
##    <chr>                                                       <int>
##  1 XXIII Congresso Brasileiro de Engenharia Biomédica             18
##  2 Congresso Brasileiro de Automática                             12
##  3 Simpósio Brasileiro de Sistemas Elétricos                       8
##  4 Simpósio Brasileiro de Sistemas Elétricos - SBSE                8
##  5 XXIV Congresso Brasileiro de Engenharia Biomédica               8
##  6 Conferência Brasileira sobre Qualidade de Energia Elétrica      7
##  7 Simpósio Brasileiro de Telecomunicações (SBrT)                  7
##  8 IS&amp;amp;T/SPIE Electronic Imaging                            6
##  9 Simpósio Brasileiro de Automação Inteligente                    6
## 10 Simpósio Brasileiro de Automação Inteligente (SBAI)             6
```

Podemos descobrir também, para cada ano, se houveram mais publicações nacionais ou internacionais.

```
ggplot(data = unb_ePUB_df) +
  geom_bar(mapping = aes(x = ano_do_trabalho,
                         fill = classificacao ))
```



Para as orientações dos professores, podemos filtrar os trabalhos relacionados a robótica com o comando a seguir.

```
df <- unb_eORIENT_df %>%
      select(titulo) %>%
      filter(grepl('robô|robo|robotic|robotics', titulo))
head(df, 10)
```

```
##                                                                                titulo
## 1                                                          Controle híbrido de robô aéreo
## 2                  Sistema de navegação por ultrassom para uso em cirurgia auxiliada por robô
## 3                                                                       SLAM em robô aéreo
## 4   Kinematic control based on dual quaternion algebra and its application to robot manipulators
## 5              Controle de robô para auxílio em cirurgia laparoscópica usando quatérnios duais
## 6                                                  Estabilização de marcha para um robô humanóide
## 7  Implementação de módulos de fusão sensorial e de controle para cooperação de dois robôs móveis
## 8          Implementação de técnicas de visão computacional para cooperação de dois robôs móveis
## 9                                                  Controle de equilíbrio de um robô humanoide
## 10           Algoritmos de processamento sensorial para estimação de estado de um robô humanoide
```

Podemos descobrir também quantas teses de doutorado receberam suporte de bolsa nestes anos. Podemos observar que depende do ano, sendo que em 2016 o número dos que receberam bolsa foi maior do que os que não receberam.

```
df <- unb_eORIENT_df %>%
select(titulo, natureza, ano, bolsa) %>%
filter(natureza == "Tese de doutorado") %>%
distinct(titulo, .keep_all = TRUE)

ggplot(data = df) +
geom_bar(mapping = aes(x = ano,
                       fill = bolsa ))
```

Por fim, sobre os perfis dos professores, podemos obter qual a senioridade dos professores do departamento de Engenharia elétrica. Podemos ver que a maior parte tem senioridade 0 seguida de 7.

```
df <- unb_ePERFIS_df %>%
  select(nome, senioridade) %>%
  distinct(nome, .keep_all = TRUE) %>%
  group_by(senioridade) %>%
  summarise(count = n())
  head(arrange(df, desc(count)), 10)
```

```
## # A tibble: 8 x 2
##   senioridade count
##   <fct>       <int>
## 1 0             17
## 2 7             13
## 3 5              8
## 4 2              7
## 5 6              7
## 6 4              6
## 7 3              3
## 8 1              2
```

4. Conclusão

O presente trabalho teve grande utilidade para a análise de dados de interesse nos datasets e-lattes e DGP para o autor, sendo que contribuiu para o melhor entendimento do funcionamento da área de Ciências de Dados, assim como as plataformas utilizadas e métodos de maior uso.