Disciplina Ciência de Dados Aplicada e

Ciência de Dados para Todos

Relatório 2 – Disciplina Ciência de Dados para Todos

Autor: Luis Gustavo Avelino de Lima Jacinto Data: 29/04/2018

1. Introdução

A Ciência de Dados permite a extração de informações valiosas a partir dos dados. Como estamos vivendo na era do Big Data, a Ciência de dados está se tornando um campo muito promissor para explorar e processar grandes volumes de dados gerados a partir de várias fontes e em diferentes velocidades.

Inserido nesse contexto, esse relatório tem como objetivo aplicar diversos conceitos de ciência de dados para analisar os dados referentes ao Departamento de Linguística da UnB. Além disso, também será utilizado as base do Diretório de Grupos de Pesquisa, DPG, de forma a associar os resultados obtidos das bases de dados anteriores.

Foram utilizadas as seguintes bases de dados: *"unb.Perfis.Linguistica.json", "unb.Pub.Linguistica.json", "unb.Orientacao.Linguistica.json" e "unb.DGP.xls"*

2. Metodologia

Para a análise dos foi utilizado o software RStudio, foram utilizados as seguintes bibliotecas:

library(jsonlite**)**

library(tidyr**)**

library(ggplot2**)**

library(readxl**)**

library(dplyr**)**

Para leitura dos dados foram utilizados as funções: *read_json() e read_xls()*.Para a compreensão e visualização dos dados utilizou-se as funções: *str(), summary() e*

glimpse(), e também foram utilizadas as funções *head(), names(), lenght(), unlist(), dim()*. Para melhor visualização dos dados foram utilizadas também as funções: *data.frame(), sapply(), filter(), group_by(), select(), rbind(), ggplot()*.

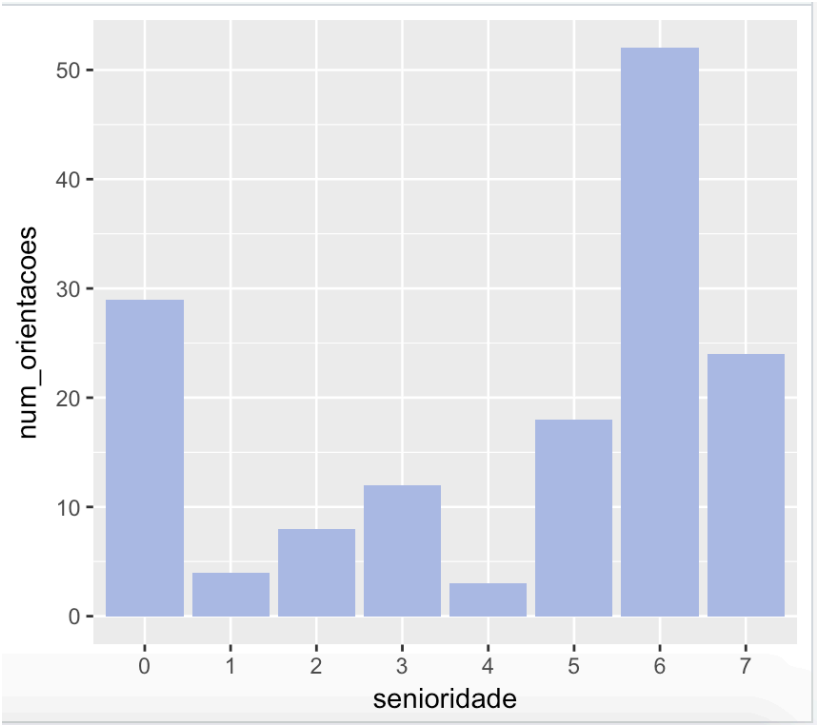
3. Resultados

Primeiramente, foi analisado o arquivo *"unb.Perfis.Linguistica.json"* que está em formato JSON e foi lido por meio da função read_json('unb.Perfis.Linguistica.json'). Esse arquivo possui perfis de 55 professores do departamento de Linguística e contém as seguintes informações sobre os professores, que pode, ser obtidas por meio da função *names()*:

```
[1] "nome"                "resumo_cv"           "areas_de_atuacao"    "endereco_profissional"
[5] "producao_bibiografica"  "orientacoes_academicas" "senioridade"
```

A fim de analisar a relação entre senioridade e orientações acadêmicas, foi gerado um gráfico de barra dessas duas variáveis. Primeiramente separando a senioridade e o número de orientações em variáveis diferentes, por meio da função *sapply()*, após isso criando um data frame com essas duas variáveis e por fim utilizando a função *ggplot()* para criar o gráfico:

```
senioridade <- sapply(perfisLinguistica, function(x) x$senioridade)
num_orientacoes <- sapply(perfisLinguistica, function(x) length(x$orientacoes_academicas))
data <- data.frame(num_orientacoes, senioridade)
ggplot(data, aes(x=senioridade, y=num_orientacoes)) + geom_bar(stat="identity", fill="#adbce6")
```



Para a base de dados *unb.Orientacao.Linguistica.json* inicialmente o foco foi contemplar o conteúdo dessa base, isso deu-se por meio do código abaixo:

```
sapply(orientacoesLinguistica, names)
```

Através desse código pode-se observar que o conteúdo dessa base de dados trata-se de orientações acadêmicas de **"Andamento em Pós Doutorado", "Andamento em Mestrado", "Andamento em Iniciação Científica", "Andamento Doutorado", "Andamento Graduação", "Concluída Doutorado", "Concluída Mestrado", "Concluída Pós Doutorado" e "Outras Orientações Concluídas"** e para cada tipo de orientação há dados dos anos de 2010 até 2017.

O enfoque será dado nas orientações concluídas de doutorado. Para verificar quantas orientações concluídas de haviam sido realizadas foi utilizado o seguinte comando:

```
> length(orientacoesLinguistica$ORIENTACAO_CONCLUIDA_DOUTORADO)
[1] 8
```

Transformando os dados para o tipo data frame, foi é possível selecionar apenas as variáveis desejadas para a análise, através da função *select()*, como observa-se abaixo:

```
> select(orientacoesLinguistica.df, titulo)
                                                                                titulo
1 Práticas identitárias do professor de língua materna: dialogismo, alteridade e identidade
```

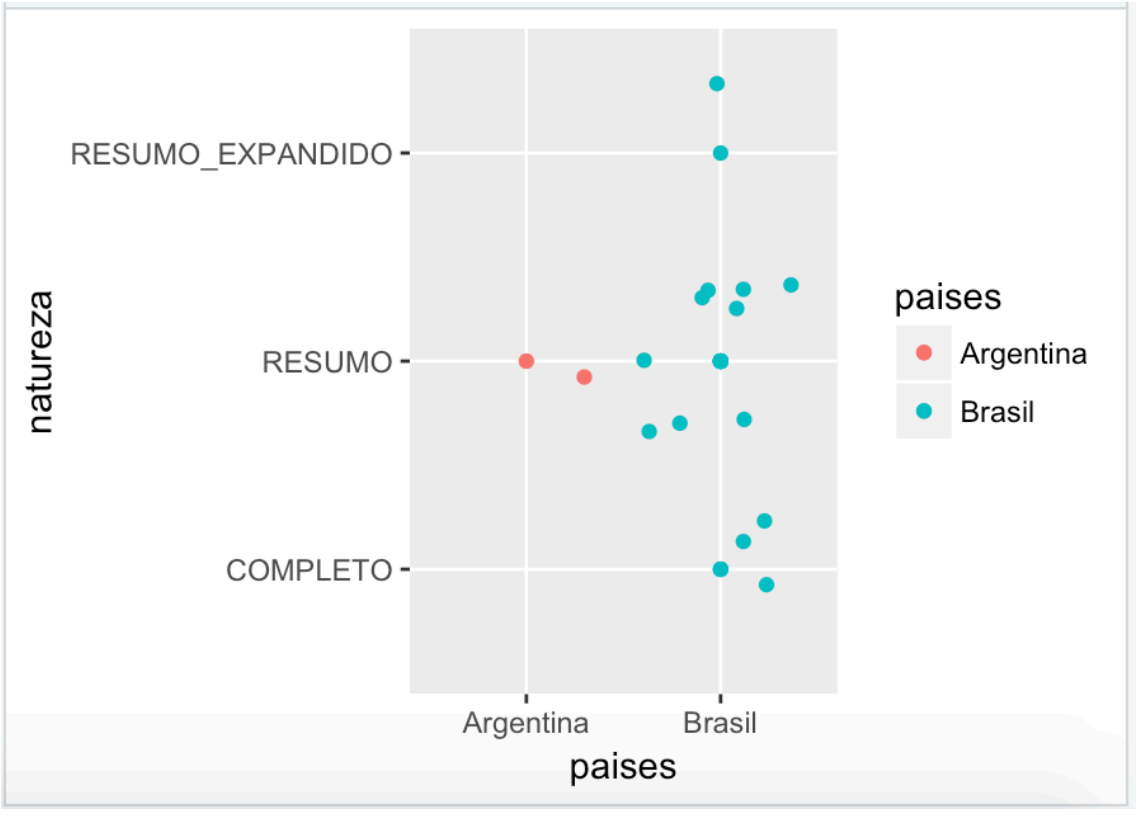
Para a terceira base de dados *"unb.Pub.Linguistica.json"*, buscou-se o entendimento de seus dados através dos códigos abaixo:

```
> sapply(pub, names)
     PERIODICO LIVRO  CAPITULO_DE_LIVRO TEXTO_EM_JORNAIS EVENTO ARTIGO_ACEITO DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
[1,] "2010"    "2010" "2010"           "2010"           "2010" "2010"        "2010"
[2,] "2011"    "2011" "2011"           "2011"           "2011" "2011"        "2011"
[3,] "2012"    "2012" "2012"           "2012"           "2012" "2012"        "2012"
[4,] "2013"    "2013" "2013"           "2013"           "2013" "2013"        "2013"
[5,] "2014"    "2014" "2014"           "2014"           "2014" "2014"        "2014"
[6,] "2015"    "2015" "2015"           "2015"           "2015" "2015"        "2015"
[7,] "2016"    "2016" "2016"           "2016"           "2016" "2016"        "2016"
[8,] "2017"    "2017" "2017"           "2017"           "2017" "2017"        "2017"
```

Abaixo, pode-se observar a quantidade de publicação de cada tipo:

```
> sapply(pub, function(x) sapply(x, function(y) length(y)))
     PERIODICO LIVRO CAPITULO_DE_LIVRO TEXTO_EM_JORNAIS EVENTO ARTIGO_ACEITO DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
2010        31    18               33                3     58             2                                      3
2011        30    13               15               12     71             1                                      1
2012        25     9               21               22     57             2                                      3
2013        36    19               34               35     50             3                                      4
2014        27    12               31               12     51             4                                      4
2015        32    17               23                6     67             2                                      1
2016        40    10               39               20     60             1                                      2
2017        12    10               26                0     14             4                                      3
```

Analisando os eventos do ano de 2017 buscou-se encontrar uma relação entre os países referentes ao evento com a natureza da publicação. Pode-se observar essa relação a partir dos seguinte gráfico:



Esse gráfico foi gerado a partir do seguinte código:

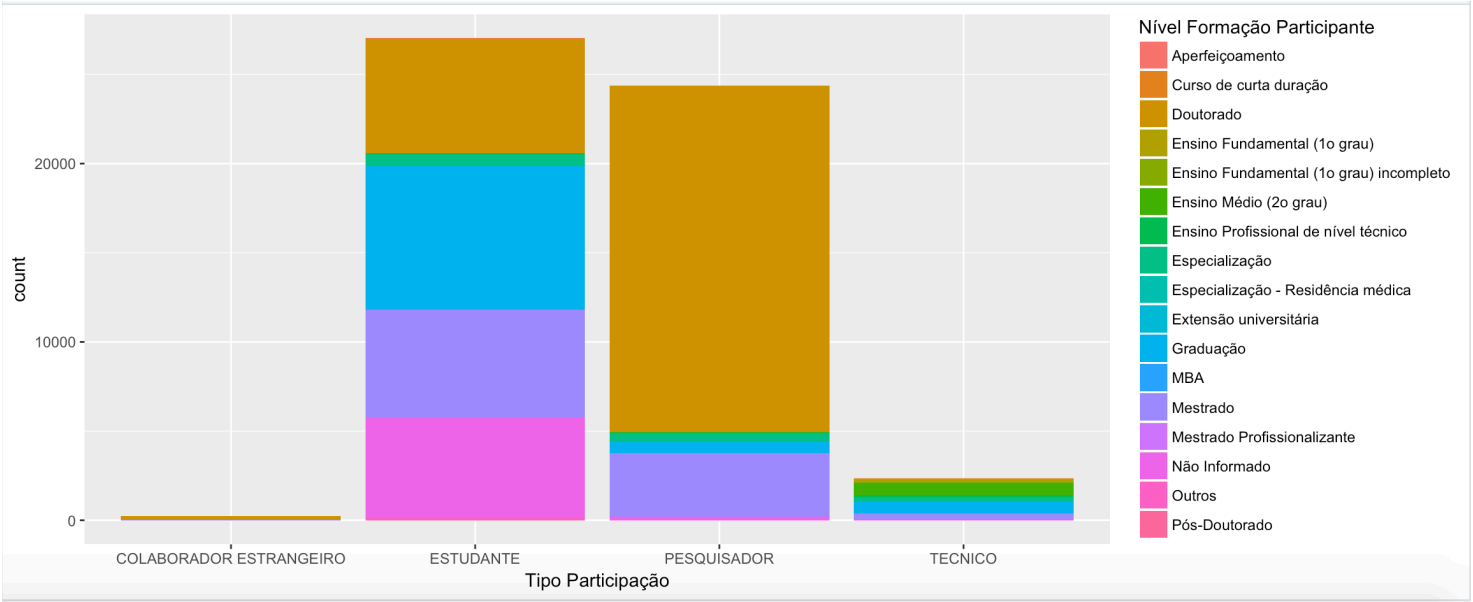
```
eventos2017 <- pub$EVENTO$`2017`
paises <- sapply(eventos2017, function(x) x$pais_do_evento)
natureza <- sapply(eventos2017, function(x) x$natureza)
data <- data.frame(paises, natureza)
ggplot(data, aes(x=paises, y=natureza, col=paises)) + geom_point() + geom_jitter()
```

Por fim, para o arquivo DGP foi transformado de .xls para .xlsx e foi importado apenas a página 5, através do código abaixo. Também foi utilizado a função *glimpse()* para melhor entendimento dos dados.

```
> dgp <- read_xlsx("dgp.xlsx", sheet=4, col_names = TRUE, col_types = NULL, skip=0)
> glimpse(dgp)
Observations: 53,984
Variables: 8
$ `Token Grupo Pesquisa`            <chr> "0025913153775147", "0025913153775147", "0025913153775147",...
$ `Ano Censo`                       <chr> "2014", "2016", "2014", "2016", "2014", "2016", "2014", "20...
$ `Nome Participante`               <chr> "Anna Francisca Salles Marques da Silva", "Anna Francisca S...
$ `Ano Nascimento Participante`     <dbl> 1994, 1994, 1995, 1995, 1988, 1988, 1986, 1986, 1974, 1974,...
$ `Tipo Participação`               <chr> "ESTUDANTE", "ESTUDANTE", "ESTUDANTE", "ESTUDANTE", "ESTUDA...
$ `Nível Formação Participante`     <chr> "Graduação", "Graduação", "Graduação", "Graduação", "Gradua...
$ `País de Nascimento Participante` <chr> "Brasil", "Brasil", "Brasil", "Brasil", "Brasil", "Brasil",...
$ `Género Participante`             <chr> "Feminino", "Feminino", "Feminino", "Feminino", "Masculino"...
> DGP_producao <- read_xls("unb.dgp.xls", sheet=5, col_names = TRUE, col_types = NULL, skip=0)
```

Pode-se diferenciar os participantes por nível de formação e o tipo de participação, gerando o gráfico a seguir:

```
ggplot(data = dgp) + geom_bar(mapping = aes(x = `Tipo Participação`, fill = `Nível Formação Participante` ))
```



Percebe-se que para os estudantes tem um certo equilíbrio e entre os diferentes níveis de formação, já para os pesquisadores, em maioria absoluta são de doutorados.

4. Conclusão

Este trabalho teve grande utilidade para a análise de dados de interesse nos datasets e-lattes e DGP. Com ele foi possível colaborar para o maior entendimento do funcionamento da área de Ciências de Dados, assim como as plataformas utilizadas e métodos de maior uso. Com esse trabalho também foi possível observar a grande participação do doutores na área acadêmica da UnB.