Relatório Grupo 12 - Educação, Profissional Educação, Educação Física e Ciências de Reabilitação

Bruno Helder, Gabriel Almeida, Thiago Luis

November 17, 2018

Introdução

Este documento apresenta uma análise exploratória de dados retirados da plataforma e-Lattes e da plataforma Sucupira para a construção do relatório final da disciplina Ciência de Dados para Todos (Data Science For All), do Departamento de Ciência da Computação da Universidade de Brasília.

Os dados utilizados na análise estão relacionados aos programas de pós-graduação de: Educação, Educação Física e Ciências de Reabilitação da Universidade de Brasília.

A metodologia para desenvolvimento do relatório é baseada no modelo de mineração de dados denominado CRISP-DM (Chapman et al., 2000, Mariscal et al., 2010).

Este documento deve ser referenciado do modo como aparece na seção de referências ao final do texto, abaixo reproduzida.

Bruno Helder Rodrigues Guedes, Gabriel Almeida Campos, Thiago Luis Rodrigues Pinho. "Ciência de Dados para Todos (Data Science For All) - 2018.2 - Relatório dos Programas de Pós Graduação de Educação - Departamento de Ciência da Computação da UnB". Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2018.2, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 17 de novembro de 2018.

Contextualização

Nesta sessão iremos discorrer sobre os programas de pós-graduação avaliados para a execução deste projeto e também como a Universidade de Brasília se inclui neste contexto.

O que é a CAPES e o Plano Nacional de Pós-Graduação - PNPG

A CAPES, ou Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, é uma instituição vinculada ao Ministério da Educação, que atua na organização, expansão e avaliação de programas de pós-graduação (mestrado e doutorado) em todo o país. Sua missão institucional é definida como *"Avaliação da pós-graduação stricto sensu, acesso e divulgação da produção científica, investimentos na formação de especialistas de alto nível e promoção da cooperação científica internacional."*, e então, possui diversos programas para assegurar que os objetivos de sua missão sejam cumpridos.

O Plano Nacional de Pós-Graduação é um programa institucional que visa definir as estratégias e propostas de políticas públicas para a área de pós-graduação e pesquisa no Brasil.

Como forma de consolidar o crescimento científico e a qualidade da esfera acadêmica, a CAPES também realiza a Avaliação Quadrienal dos programas de pós-graduação. O programa realiza a avaliação utilizando os seguintes quesitos:

- Proposta do Programa
- Corpo Docente
- Corpo Discente, Teses e Dissertações

- Produção Intelectual
- Inserção Social

A partir de notas dadas entre 1 e 7 para cada um destes quesitos (tendo em vista que os quesitos possuem diferentes pesos para a avaliação), é no final concedido ao programa uma nota relativo ao todo do programa de pós-graduação. Assim, a CAPES possue a capacidade de determinar o descredenciamento dos cursos que apresentem nota baixa.

Programas de Pós-Graduação

Neste projeto iremos analisar e explorar quatro diferentes programas de Pós-Graduação da Universidade de Brasília, Ciências de reabilitação (53001010107P2), Educação (53001010001P0) , Profissional Educação(53001010087P1) e Educação Física(53001010066P4).

Programa de Pós-Graduação de Educação Física

"O Programa de Pós-Graduação em Educação Física (PPGEF) é diretamente vinculado ao Decanato de Pesquisa e Pós-Graduação da Universidade de Brasília (UnB), à sua Câmara de Pesquisa e Pós-Graduação e ao Conselho de Ensino, Pesquisa e Extensão". (Retrieved from http://www.ppgef.unb.br/o-programa)

O programa tem como objetivo consolidar e expandir a pesquisa e o desenvolvimento científico na área de Educação Física, assim como promover a troca de conhecimento acadêmico à sociedade, de forma a atender as demandas sociais.

O programa, na última Avaliação Quadrienal da CAPES recebeu nota 4, que segundo a legenda dos conceitos significa *"bom"*.

Descrição da Metodologia

Nesta seção iremos descrever e analisar o processo de exploração, pesquisa e desenvolvimento utilizados para a criação deste projeto, o CRISP-DM. CRISP DM é abreviação de Cross-industry standard process for data mining que é um padrão industrial aberto de processos para abordar trabalhos de data mining e análise de dados em geral. Esta metodologia basicamente se divide em seis etapas: Business Understanding , Data Understanding, Data preparation , Modeling , Evaluation e Deployment. Iremos descrever cada uma das etapas mais detalhadamente a seguir. Business Understanding

A primeira etapa do processo é talvez a mais abstrata de todas, que é o conhecimento sobre o problema. O objetivo é que se avalie os custos e os impactos da solução a ser proposta. Nesta etapa também, são definidos os objetivos e metas do processo. Neste projeto estamos analisando os dados relativos aos Programas de Pós-Graduação da Universidade de Brasília. À partir disso desejamos realizar uma análise exploratória sobre os dados para oferecer novas perspectivas sobre estes programas.

Data Understanding

Data Understanding, que pode ser traduzido para Entendimento dos Dados, é a etapa que consiste em avaliar os dados já disponíveis. Deve-se organizar e descrever os dados obtidos, e, avaliar se é possível atingir os objetivos determinados na etapa anterior com estes dados. Esta etapa tem um teor também exploratório, pois é necessário avaliar que tipo de análise é possível realizar com os dados disponíveis e tomar decisões baseadas nesta análise. Por exemplo, é possível determinar que é necessário realizar o processo de obtenção dos dados novamente, ou retornar para a etapa de Business Understanding.

Neste projeto, temos acesso tanto aos dados disponibilizados pelo professor, em diferentes arquivos em formato JSON, com dados sobre todos os projetos de pós-graduação da UnB, e também os dados disponíveis

no E-lattes. O arquivo profile.json contém informações referentes ao perfil dos docentes, identificados pelo número de matrícula e contém nome, resumo do currículo, áreas de atuação, endereço profissional, produção bibliográfica que contém os capítulos de livro, eventos, livros e artigos em periódicos publicados pelo docente. O arquivo publication.json tem informações referentes as publicações da determinada área no período entre 2010 e 2017.

Data Preparation

A etapa de Data Preparation, ou Preparação dos dados, é uma etapa de cunho técnico que tem como objetivo realizar a limpeza dos dados para o processo de modelagem. Dependendo da maneira de como foi realizada a coletagem de dados, esta etapa pode se tornar menos ou mais relevante. Nesta etapa também é realizada a construção de características derivadas, ou seja, dados que são possíveis ser inferidos dos dados disponíveis.

Tomando em conta o projeto aqui sendo desenvolvido, esta etapa é menos relevante pois já obtemos os dados semi-filtrados, que podem ser facilmente utilizados no ambiente de desenvolvimento utilizando a linguagem R, restando apenas selecionar as características a serem utilizadas.

Modelagem

Nesta etapa é aplicado as técnicas de Data Science e Data Analisys para construir modelos adequados para o contexto da situação e consequentemente tirar conclusões ou proposições para as soluções do problema proposto na etapa de Business Understanding.

No projeto em questão, iremos ((inserir o que vai ser feito no projeto)).

Evaluation

Fase de avaliação dos resultados obtidos na etapa de modelagem e verificando se cumpre os objetivos definidos na etapa de business understanding.

Deployment

Fase de implantação de fato dos resultados obtidos nas análises realizadas e assim concluindo o processo de desenvolvimento.

CRISP-DM (Corresponderia à seção de Metodologia)

Essa parte eu acho que devemos manter bastante coisa, não tem muito o que acrescentar sobre o CRIPS, no máximo que além de tudo o que foi dito, nós somos estimulados a usar esse método por questões didáticas.

CRISP-DM Fase 1 - Entendimento do Negócio

O que é a UnB? O que são as grandes áreas? Como são avaliadas as POS?(Usando aquela avaliação quadrienal da CAPES) Quais são as POS que falaremos? E o que elas tem em relação às grandes áreas e a avaliação quadrienal?

CRISP-DM Fase 2 - Entendimento dos Dados

Doravante, a fim de facilitar aos alunos seguirem a metodologia CRISP-DM, os nomes das seções e subseções de texto serão prefixadas com o número e nome da fase e atividade genérica do CRISP-DM. Fica facultado aos grupos seguir ou não a sequência prevista, tendo em vista que se pode retornar às fases anteriores, bem como podem haver atividades que não foram adequadas às características do problema específico sob análise.

CRISP-DM Fase.Atividade 2.1 - Coleta inicial dos dados

Todos os arquivos com dados iniciais a seguir apresentados foram fornecidos pelos professores responsáveis pela disciplina. Os dados foram gerados no mês de maio de 2018, e compilam informações entre os anos de 2010 e 2017. Os arquivos estão no formato JSON, e seus atributos iniciais e conteúdos são apresentados a seguir.

Perfil profissional dos docentes vinculados às pós-graduações

Os arquivos utilizados para os dados sobre o perfil dos docentes vinculados aos programas de pós-graduação da UnB, entre 2010 e 2017, em específico são:
>Todas - "dataset/pos_unb_todas/unbpos.profile.json"
Ciências de reabilitação(53001010107P2) - "dataset/pos_ciencias_de_reabilitacao/250.profile.json"
Educação(53001010001P0) - "dataset/pos_educacao/273.profile.json"
Profissional Educação(53001010087P1) - "dataset/pos_profissional_educacao/274.profile.json"
Educação Física(53001010066P4) - "dataset/pos_educacao_fisica/277.profile.json"

Orientações de mestrado e doutorado realizadas pelos docentes vinculados às pós-graduações

Os arquivos com os dados sobre as vinculações de todos os docentes que declararam atuar em cada uma das áreas de pós-graduação do Sistema Nacional de Pós-Graduação da CAPES, conforme apresenta-se registrada essa informação no currículo Lattes de cada um, em data recente são:
>Todas - "dataset/pos_unb_todas/unbpos.advise.json"
Ciências de reabilitação(53001010107P2) - "dataset/pos_ciencias_de_reabilitacao/250.advise.json"
Educação(53001010001P0) - "dataset/pos_educacao/273.advise.json"
Profissional Educação(53001010087P1) - "dataset/pos_profissional_educacao/274.advise.json"
Educação Física(53001010066P4) - "dataset/pos_educacao_fisica/277.advise.json"

Produção bibliográfica gerada pelos docentes vinculados às pós-graduações

Os arquivos usados sobre os dados da produção bibliográfica gerada por todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017, foram:
>Todas - "dataset/pos_unb_todas/unbpos.publication.json"
Ciências de reabilitação(53001010107P2) - "dataset/pos_ciencias_de_reabilitacao/250.publication.json"
Educação(53001010001P0) - "dataset/pos_educacao/273.publication.json"
Profissional Educação(53001010087P1) - "dataset/pos_profissional_educacao/274.publication.json"
Educação Física(53001010066P4) - "dataset/pos_educacao_fisica/277.publication.json"

Agrupamento dos docentes conforme áreas de atuação

O arquivo dataset/pos_unb_todas/unbpos.researchers_by_area.json apresenta as vinculações de todos os docentes que declararam atuar em cada uma das áreas de pós-graduação do Sistema Nacional de Pós-

Graduação da CAPES, conforme apresenta-se registrada essa informação no currículo Lattes de cada um, em data recente.

Redes de colaboração entre docentes

Os arquivo que apresentam redes de colaboração na co-autoria de artigos cientpificos, feitas entre os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017, são: >Todas - "dataset/pos_unb_todas/unbpos.graph.json" Ciências de reabilitação(53001010107P2) - "dataset/pos_ciencias_de_reabilitacao/250.graph.json" Educação(53001010001P0) - "dataset/pos_educacao/273.graph.json" Profissional Educação(53001010087P1) - "dataset/pos_profissional_educacao/274.graph.json" Educação Física(53001010066P4) - "dataset/pos_educacao_fisica/277.graph.json"

CRISP-DM Fase.Atividade 2.2 - Descrição dos Dados

Para ler e manipular inicialmente esses dados, serão usadas primordialmente as bibliotecas seguintes

```
library(jsonlite)
library(listviewer)
library(readxl)
library(readr)
library(ggplot2)
library(tidyverse)
library(stringr)
```

Como já informado, a descrição dos dados verifica se os dados sendo acessados terão potencial para responder às questões de *data mining*. Além disso, deve-se avaliar qual o volume de dados, a estrutura dos dados (tipos), codificações usadas, etc. Neste projeto, a descrição dos dados é responsabilidade parcial dos alunos, tendo em vista que esta seção já oferece uma descrição inicial simplificada. O relatório final deve conter descrições significativas e aprofundadas dos dados.

Descrição dos dados do perfil

Os arquivos que contém dados que caracterizam o perfil profissional de todos os docentes dos grupos sob análise, podem ser lidos por meio do comando seguinte.

```
unb.perfil.geral <- fromJSON(json.perfil)
unb.perfil.ciencias_de_reabilitacao <- fromJSON(json.perfil.ciencias_de_reabilitacao)
unb.perfil.educacao <- fromJSON(json.perfil.educacao)
unb.perfil.profissional_educacao <- fromJSON(json.perfil.profissional_educacao)
unb.perfil.educacao_fisica <- fromJSON(json.perfil.educacao_fisica)
```

A quantidade de docentes sob análise é apresentada a seguir.

```
length(unb.perfil.geral)
```

```
## [1] 1764
```

Sendo que desses, por pós-graduação, temos:

```
length(unb.perfil.ciencias_de_reabilitacao)
```

```
## [1] 18
```

```
length(unb.perfil.educacao)
```

```
## [1] 50
```

```
length(unb.perfil.profissional_educacao)
```

```
## [1] 16
```

```
length(unb.perfil.educacao_fisica)
```

```
## [1] 28
```

Descrição dos dados de orientações

Os arquivos que contém os dados referentes às orientações de todos os docentes dos grupos sob análise podem ser lidos utilizando os seguintes comandos:

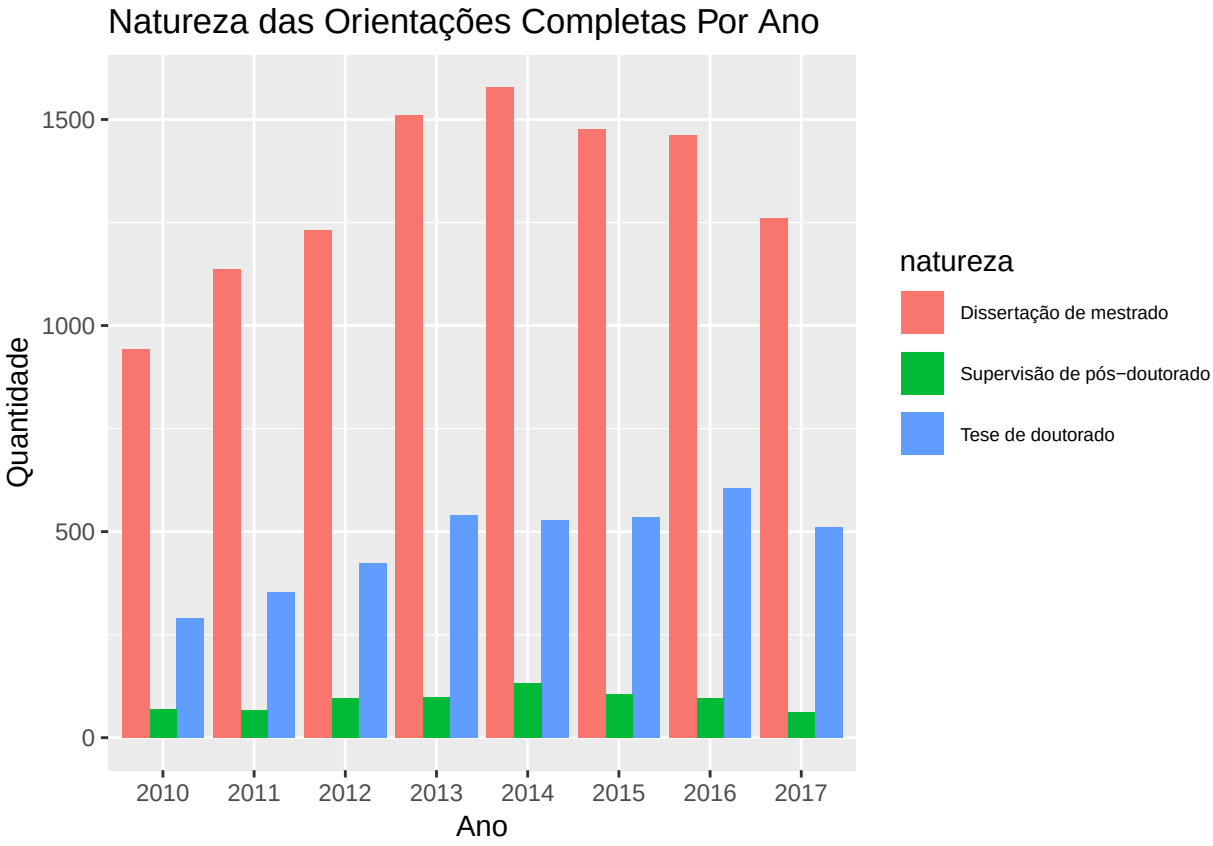
```
unb.orientacao.geral <- fromJSON(json.advise)
unb.orientacao.ciencias_de_reabilitacao <- fromJSON(json.advise.educacao)
unb.orientacao.educacao <- fromJSON(json.advise.educacao)
unb.orientacao.profissional_educacao <- fromJSON(json.advise.profissional_educacao)
unb.orientacao.educacao_fisica <- fromJSON(json.advise.educacao_fisica)
```

Para estudarmos as orientações produzidas primeiro, estudaremos a estrutura com que foram armazendadas:

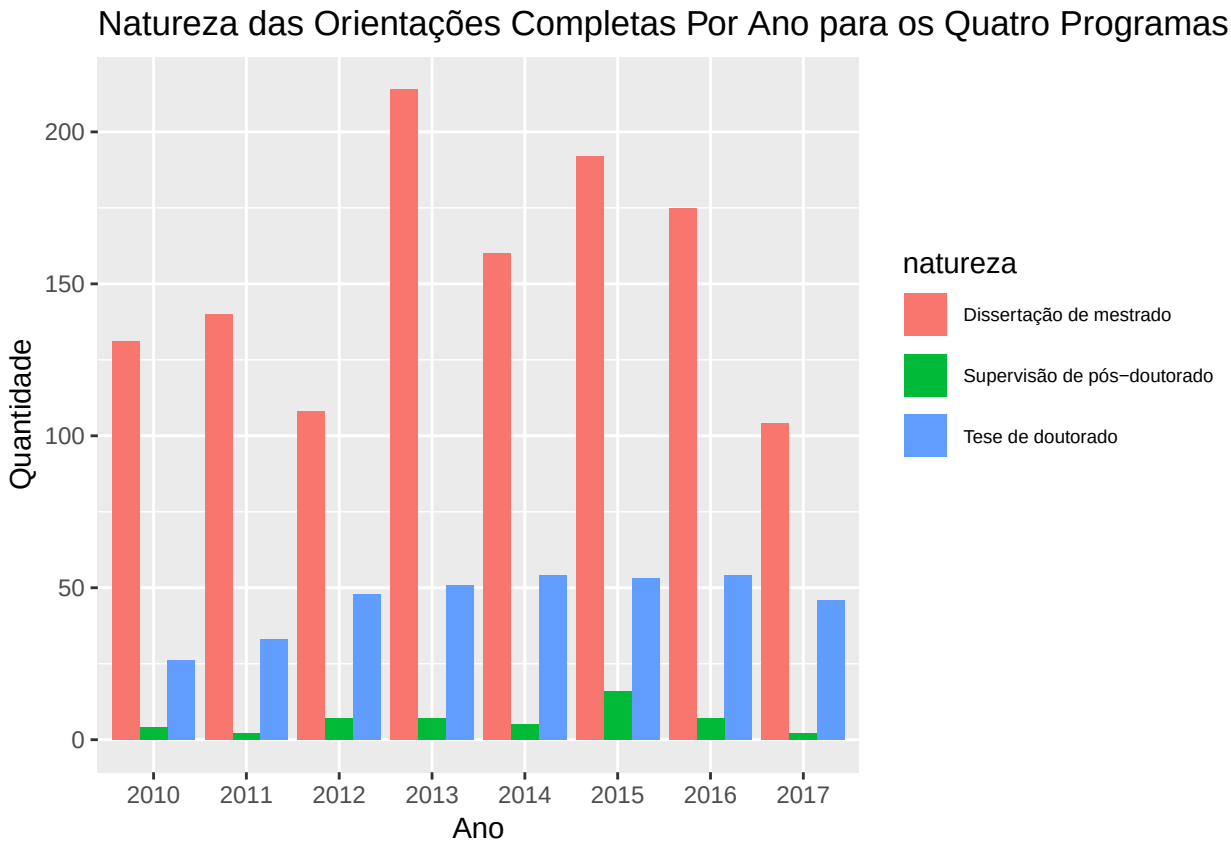
```
## [1] "ORIENTACAO_EM_ANDAMENTO_DE_POS_DOUTORADO"
## [2] "ORIENTACAO_EM_ANDAMENTO_DOUTORADO"
## [3] "ORIENTACAO_EM_ANDAMENTO_MESTRADO"
## [4] "ORIENTACAO_EM_ANDAMENTO_GRADUACAO"
## [5] "ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA"
## [6] "ORIENTACAO_CONCLUIDA_POS_DOUTORADO"
## [7] "ORIENTACAO_CONCLUIDA_DOUTORADO"
## [8] "ORIENTACAO_CONCLUIDA_MESTRADO"
## [9] "OUTRAS_ORIENTACOES_CONCLUIDAS"
```

```
##  [1] "natureza"                   "titulo"
##  [3] "ano"                        "id_lattes_aluno"
##  [5] "nome_aluno"                 "instituicao"
##  [7] "curso"                      "codigo_do_curso"
##  [9] "bolsa"                      "agencia_financiadora"
## [11] "codigo_agencia_financiadora" "nome_orientadores"
## [13] "id_lattes_orientadores"
```

Para melhor descrever esses dados, aqui está a relação de orientações por natureza de todos os programas por ano:



Os mesmos gráficos especificamente para os programas que esse relatório foca:



Descrição dos dados de produção bibliográfica

Os arquivos que contém os dados referentes às produções bibliográficas desse mesmo grupo são:

```
unb.publicacao.geral <- fromJSON(json.publication)
unb.publicacao.ciencias_de_reabilitacao <- fromJSON(json.publication.educacao)
unb.publicacao.educacao <- fromJSON(json.publication.educacao)
unb.publicacao.profissional_educacao <- fromJSON(json.publication.profissional_educacao)
unb.publicacao.educacao_fisica <- fromJSON(json.publication.educacao_fisica)
```

E suas respectivas estruturas são:

```
## [1] "PERIODICO"
## [2] "LIVRO"
## [3] "CAPITULO_DE_LIVRO"
## [4] "TEXTO_EM_JORNAIS"
## [5] "EVENTO"
## [6] "ARTIGO_ACEITO"
## [7] "DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA"
```

```
##  [1] "natureza"         "titulo"           "periodico"
##  [4] "ano"              "volume"           "issn"
##  [7] "paginas"          "doi"              "autores"
## [10] "autores-endogeno"
```

```
##
##               REVISTA DE SAÚDE PÚBLICA (ONLINE)
##                                              23
```

```
##                                     PLoS One
##                                           21
##                             ESPACIOS (CARACAS)
##                                           16
##                             Scientific Reports
##                                           16
##                       Ciencia & Saude Coletiva
##                                           15
##                GENETICS AND MOLECULAR RESEARCH
##                                           15
##                          CADERNOS DE PROSPECÇÃO
##                                           14
##        JOURNAL OF SOUTH AMERICAN EARTH SCIENCES
##                                           14
##           Journal of Molecular Modeling (Print)
##                                           13
## RBC. REVISTA BRASILEIRA DE CARTOGRAFIA (ONLINE)
##                                           13
```

```
##
##                            ANPOF
##                               23                                13
##         Novas Edições Acadêmicas            Laccademia Publishing
##                               12                                 8
## Editora Universidade de Brasília                  Pontes Editores
##                                7                                 7
##                      Lumen Juris                       Fino Traço
##                                6                                 5
##                         INEP/MEC                              LTr
##                                5                                 5
```

Descrição dos dados de agregação de docentes por área

```
##   Area                       Professores
## 1 "Administração"            "101"
## 2 "Agronomia"                "65"
## 3 "Antropologia"             "52"
## 4 "Arqueologia"              "2"
## 5 "Arquitetura e Urbanismo" "42"
## 6 "Artes"                    "85"
```

Se analisarmos por cada docente dos programas que este relatório estuda:

```
## [1] "Ciências de Reabilitação"
```

```
##
##        CIENCIAS_BIOLOGICAS          CIENCIAS_DA_SAUDE
##                          6                         55
##           CIENCIAS_HUMANAS CIENCIAS_SOCIAIS_APLICADAS
##                          4                          1
##                ENGENHARIAS                     OUTROS
##                          4                          1
```

```
## [1] "Educação"
```

```
##
```

```
##          CIENCIAS_DA_SAUDE CIENCIAS_EXATAS_E_DA_TERRA
##                          3                          6
##           CIENCIAS_HUMANAS CIENCIAS_SOCIAIS_APLICADAS
##                        221                          8
## LINGUISTICA_LETRAS_E_ARTES
##                          8
```

```
## [1] "Profissional Educação"
```

```
##
##                                  CIENCIAS_DA_SAUDE
##                          1                          1
##           CIENCIAS_HUMANAS LINGUISTICA_LETRAS_E_ARTES
##                         63                          7
```

```
## [1] "Educação Fìsica"
```

```
##
##        CIENCIAS_BIOLOGICAS          CIENCIAS_DA_SAUDE
##                          8                         66
##           CIENCIAS_HUMANAS CIENCIAS_SOCIAIS_APLICADAS
##                         19                          3
##               ENGENHARIAS
##                          3
```