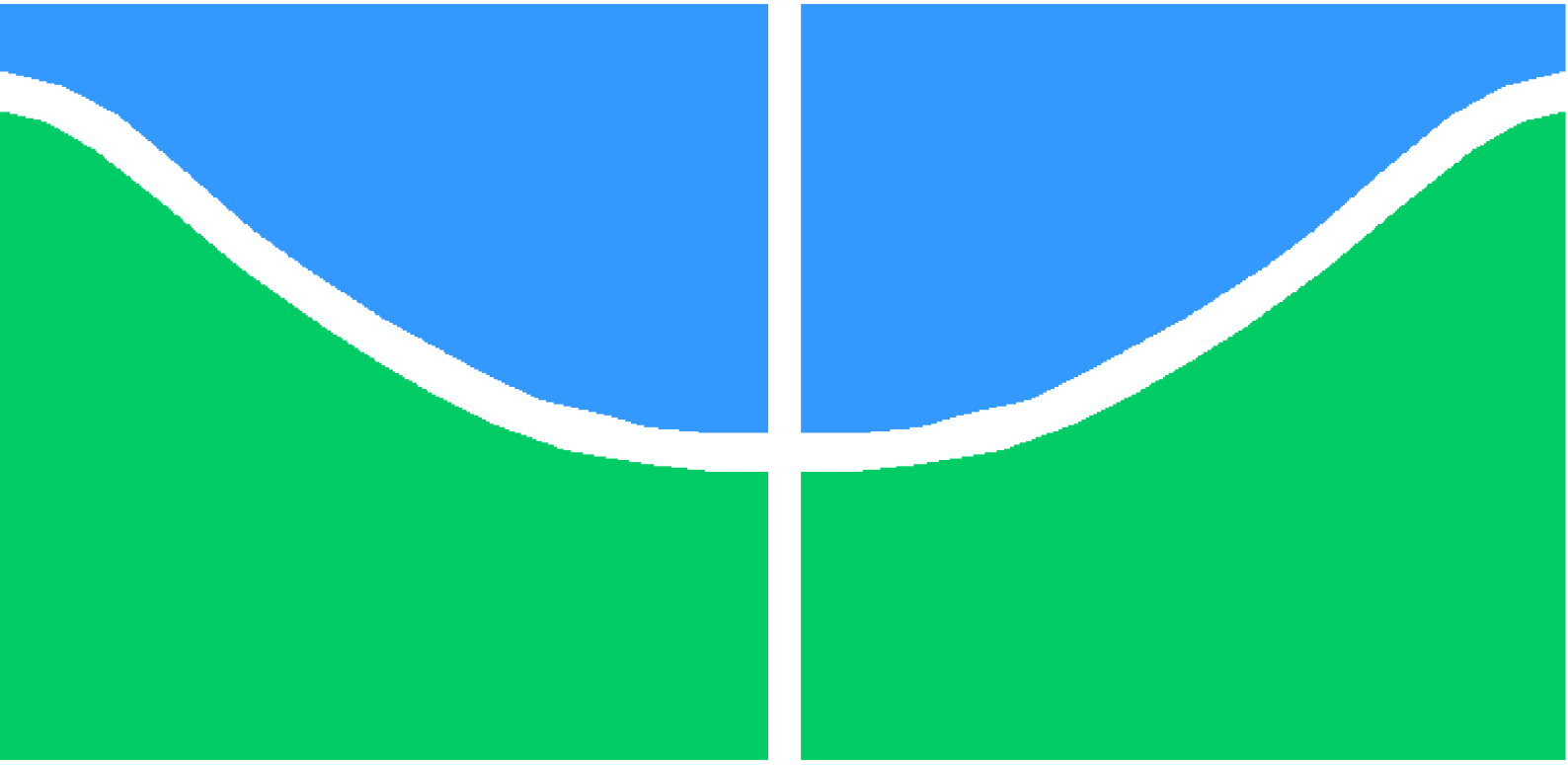


Universidade de Brasília – UnB

Data Science 4 All e Ciência de Dados Aplicada

Relatório Final: Pós Graduação - Cursos de Arquitetura e Design

Autores: Cecília Dib, Thiago C. Moreira e Vitor Bertulucci

Professores: Ricardo Barros e Jorge Henrique Cabral Fernandes

Brasília, DF

2018

DS4A & Ciêcia de Dados Aplicada

Relatório Final: Pós Graduação - Cursos de Arquitetura e Design

Cecilia França Dib de Oliveira Bessa - 14/0134425
Thiago Canabrava Moreira - 14/0163859
Vitor Bertulucci Borges - 14/0171126

1. Introdução

Este relatório se propõe a explorar, interpretar e analisar cursos de Pós Graduação da Universidade de Brasília (UnB). Mais especificamente na área de conhecimento de Arquitetura e Urbanismo e Design. Como fonte de pesquisa será utilizada a plataforma e-Lattes UnB, que disponibiliza arquivos *json* de todos os cursos da UnB. Nestes arquivos podemos explorar dados que contém informações sobre *publicações*, *perfil dos autores das publicações*, *período de atuação dos autores*, *categorias das publicações em cada ano*, dentre outros.

Além da análise dos dados, será apresentado um referencial teórico sobre as áreas de conhecimento presentes neste relatório, além da aplicação da metodologia *CRISP-DM* que é utilizada na Ciência de Dados.

1.1 Ciência de Dados

Data Science ou Ciência de Dados permite a extração de informações extremamente valiosas a partir de dados contidos em qualquer corporação ou empresa. Na era do Big Data, a ciência de dados está se tornando um campo promissor para processar grandes volumes gerados por diversas fontes e em diferentes velocidades. [5]

Em resumo é uma ciência que visa estudar as informações, seu processo de captura, transformação, geração e, posteriormente, análise de dados. [5]

A ciência de dados difere das análises estatísticas e da ciência da computação em seu método que é aplicado a dados coletados usando princípios científicos. A ciência de dados está procurando descobrir conhecimento a partir de uma quantidade grande e pesada de dados que podem ser usadas para tomar decisões e fazer previsões, e não simplesmente a interpretação de números. [5]

Para organizar, categorizar, estruturar e, depois, analisar essa quantidade exorbitante de dados, o data scientist é o profissional mais indicado. Este especialista também deve ser capaz de realizar tarefas, como extrair dados de bancos de dados MySQL e transformá-los em informações úteis e decifráveis por executivos de negócios por meio de gráficos e dashboards. [5]

No livro R for Data Science, os autores definem o fluxograma do processo de Data Science com seis etapas que englobam desde a coleta dos dados até a comunicação dos resultados com os públicos interessados, de maneira automatizada e rápida. As etapas são: importação, organização, transformação, visualização, modelagem e comunicação. Portanto, data science é a maneira de gerar conhecimento, de fazer ciência a partir dos dados. [5]

1.2 Cursos de Pós Graduação

Segundo a Capes, Programa de Pós-Graduação é o conjunto formado pelos cursos de mestrado e/ou doutorado acadêmicos ou de mestrado profissional de uma Instituição de Ensino Superior - IES atuante numa mesma área do conhecimento - sua área básica - que compartilha essencialmente o mesmo corpo docente e tem uma estrutura administrativa comum. Programas são divididos em acadêmicos ou profissionais, de acordo com a natureza do(s) curso(s).

A aplicação Coleta de Dados é um sistema informatizado da Capes, desenvolvido com o objetivo de coletar informações dos programas de pós-graduação stricto sensu do país. Com o lançamento da Plataforma Sucupira, o Coleta de Dados foi reformulado e passa a ser um dos módulos que a constituem.

A Plataforma Sucupira é uma importante ferramenta para coletar informações, realizar análises e avaliações e ser a base de referência do Sistema Nacional de Pós-Graduação - SNPG. A Plataforma deve disponibilizar em tempo real e com muito mais transparência as informações, processos e procedimentos que a Capes realiza no SNPG para toda a comunidade acadêmica. [3]

2. CRISP-DM

O desenvolvimento e apresentação de resultados, como sugerido no plano de ensino, deve seguir a metodologia CRISP-DM (**Cross Industry Standard Process for Data Mining**), que é um dos melhores e mais conhecidos modelos de análise e exploração de dados. Proposto em 1996, o modelo trata-se de um processo de data mining ou mineração de dados desvendo fases e o correto caminho a ser seguido quando estamos lidando com análise de dados e seus problemas relacionados. [6]

O processo é composto por 6 (seis) grandes etapas principais: Business Understanding; Data Understanding; Data Preparation; Modeling; Evaluation; Deployment. É importante destacar que o processo e ela é interativa e incremental e não necessariamente linear de uma fase para a outra, o que pode ser visto na figura a seguir. [6]

No relatório, o CRISP-DM será utilizado como metodologia de exploração de dados, a fim de obter uma

2.1 Compreensão de Negócios

A primeira e uma das etapas mais importantes do processo consiste em realizar e extrair informações relevantes sobre qual o contexto a ser trabalhado, qual a área de conhecimento, quais os objetivos e requisitos do projeto para poder, assim, traçar um plano claro sobre as ações que devem ser tomadas. [6]

No contexto da disciplina de Data Science 4 All (DS4A) & Ciência de Dados Aplicada, nos cursos de Arquitetura e Urbanismo e Design, o objetivo traçado foi extrair informações relevantes referentes as publicações e os autores destes cursos, tais como publicações por ano, publicações por país, palavras relevantes nas publicações, autores com mais publicações, dentre outros.

2.1.1 O Contexto

Atualmente, a sociedade vive um processo de transição para algo que nós chamamos de Quarta Revolução Industrial [7]. “Estamos a bordo de uma revolução tecnológica que transformará fundamentalmente a forma como vivemos, trabalhamos e nos relacionamos. Em sua escala, alcance e complexidade, a transformação será diferente de qualquer coisa que o ser humano tenha experimentado antes” como cita o alemão Klaus Schwab em seu livro “A Quarta Revolução Industrial”, e esta, por sua vez, é caracterizada pela era em que vivemos hoje em dia: a sociedade do conhecimento.

Uma característica importantíssima que caracteriza nossa sociedade hoje em dia é a aquisição de conhecimento e a valorização do mesmo no mercado profissional, onde o foco principal é a transformação de informação em conhecimento [8]. Com isso em mente, tem-se que a evolução da sociedade tem como objetivo aumentar o número de pessoas qualificadas nos seus diversos contextos científicos e acadêmicos.

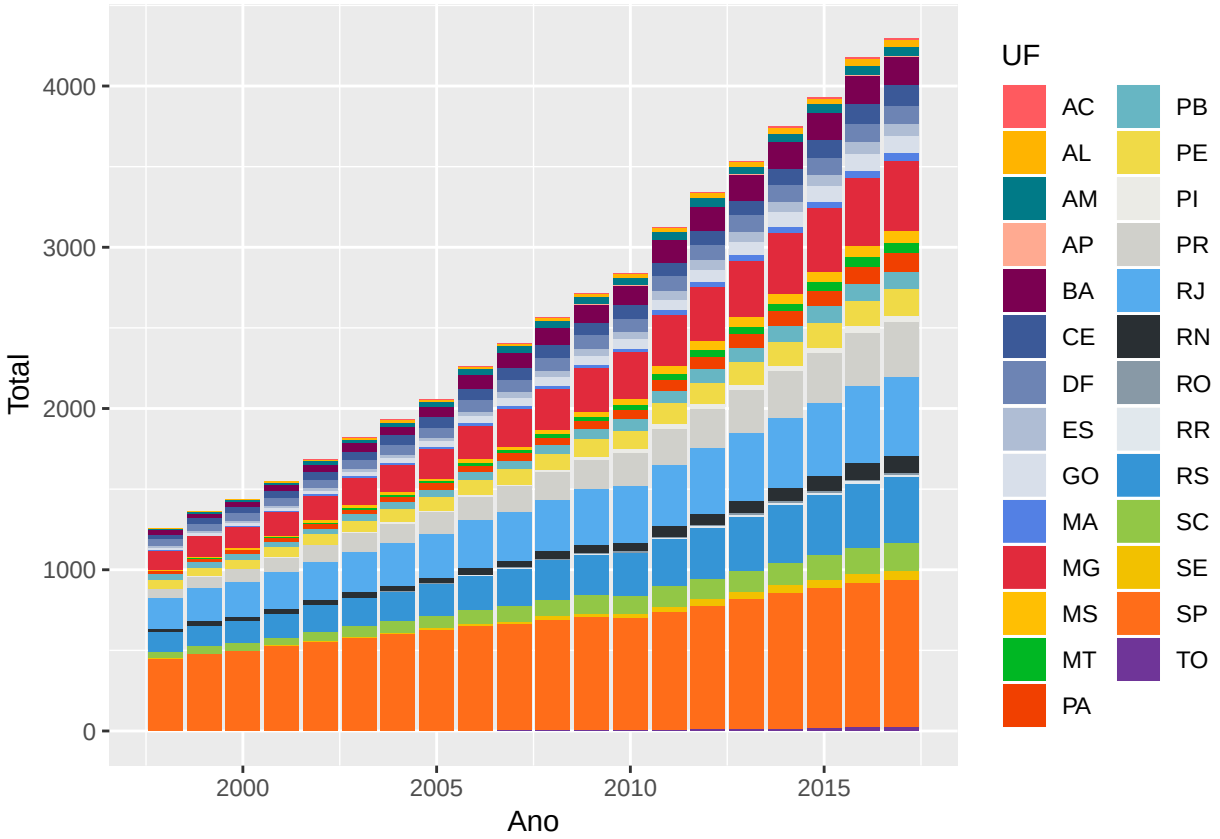
O Programa Nacional de Pós-Graduação integrado por programas de pós-graduação avaliados e reconhecidos pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) é um dos principais responsáveis pela produção de conhecimento científico no país [9].

O gráfico abaixo mostra a evolução na quantidade de programas de Pós-graduação desde de 1998 à 2017, dados fornecidos pelo GEOCAPES.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(magrittr)
library(ggplot2)

dados.geocapes <- read.csv("dados/dados-pos-graduacao-geocapes.CSV", sep=";")
summarized <- dados.geocapes %>% group_by(Ano, UF) %>% summarise(Total=n())
color.list <- c("#FF5A5F", "#FFB400", "#007A87", "#FFAA91", "#7B0051", "#3b5998", "#6d84b4", "#afbdd4",
ggplot(title="Teste",data=summarized, aes(x=Ano, y=Total)) + geom_bar( mapping = aes(x=Ano, fill=UF),
```



De acordo com o gráfico, com o passar dos anos, o interesse e criação de cursos de pós-graduação obtiveram um aumento significativo dentro dos útimos 19 anos, onde, em São Paulo, por exemplo, o número de programas em 1998 era de 450 e hoje ultrapassa o dobro desse valor, com 915 programas. A quantidade total de programas aumentou de 1.259 para 4.296 no mesmo período. A mesma tendência de evolução seguiu tanto para o número de discentes que aumentou de 44274 para 90188 também no estado de São Paulo.

Dos principais objetivos do PNPG de 2011 à 2020 estão:

- Estímulo à criação de Programas com ciclo completo (Mestrado e Doutorad);
- Inserção do Mestrado profissional;
- Internacionalização
  - envio de estudantes e docentes ao exterior;
  - aumento da quantidade e qualidade de publicações;
  - estímulo à participação em eventos no exterior;
  - atrair discentes e docentes à estudarem no exterior.

2.1.2 Arquitetura e Urbanismo

Simultaneamente à criação da Universidade de Brasília em 1962, o Curso de Mestrado em Arquitetura foi iniciado como o primeiro do país nesta área. Em 1965 o curso foi sumariamente encerrado devido à crise

política do período, retornando apenas em 1976 no quadro do 2o Plano Nacional de Desenvolvimento com a criação do Mestrado em Planejamento Urbano.

O objetivo principal era atender as demandas por capacitação de recursos humanos no campo das políticas públicas e do desenvolvimento das cidades, contando com apoio do então Conselho Nacional de Política Urbana da Secretaria de Planejamento. [1]

Em 1986, foi criado o Mestrado em Desenho Urbano para o estudo da configuração físico-espacial das cidades. Este programa se consolidou como um dos principais centros de irradiação de conhecimentos nas áreas de Planejamento e Desenho Urbano.

Com mais de 40 anos de existência, O Programa de Pesquisa e Pós-Graduação da Faculdade de Arquitetura e Urbanismo da Universidade de Brasília (PPG-FAU/UnB) enfrentou desafios e transformações. Na última década o perfil do corpo de docentes e pesquisadores perceberam a necessidade de uma nova estrutura para abranger as mudanças no Curso de Mestrado e a criação do Curso de Doutorado em Arquitetura e Urbanismo. Em 2002 a proposta foi aprovada pela CAPES.

As áreas de concentração de pesquisa em que o Programa se fundamenta são:

Teoria, história e crítica

A área de concentração Teoria, História e Crítica realiza pesquisas em torno dos processos históricos de concepção, transformação e reflexão sobre arte, arquitetura e cidade. A formação nessa área transita desde os fundamentos epistemológicos e sociais da cultura material até a articulação crítica do pensamento sobre o patrimônio cultural, passando pelo estudo empírico e documental, com metodologias diversas e enfoque interdisciplinar. Acolhe recortes espaço-temporais diversos, com ênfase na região de Brasília em perspectiva diacrônica. [2]

Dentro desta área estão as seguintes linhas de pesquisa:

- História e Teoria da Arquitetura;
- História e Teoria da Cidade e do Urbanismo;
- Patrimônio e Preservação;
- Estética, Hermenêutica e Semiótica.

Tecnologia, ambiente e sustentabilidade

A área de Tecnologia, Ambiente e Sustentabilidade compreende estudos relativos a técnicas e processos ligados à produção da Arquitetura e do ambiente construído, com especial atenção à questão da sustentabilidade. Os sistemas estruturais são estudados no âmbito específico da Arquitetura e etapas da construção, desde os projetos e técnicas de produção até seu desempenho são objeto de pesquisas. O contexto urbano comparece com pesquisas relativas a qualidade do espaço e sua gestão, reabilitação em seus aspectos físicos e socioambientais. [2]

Dentro desta área estão as seguintes linhas de pesquisa:

- Estruturas e Arquitetura;
- Sustentabilidade, Qualidade e Eficiência do ambiente construído;
- Tecnologia de produção do ambiente construído.

Projeto e Planejamento

A área de Projeto e Planejamento abarca estudos sobre políticas, planos e gestão em escalas diversas, do edifício ao território. As pesquisas voltadas para o âmbito edilício envolvem estratégias projetuais, configuração, representação e acessibilidade. Entre os temas diversos dos estudos relacionados ao urbano, podem-se citar análises e proposições sobre planejamento urbano e territorial, legislação, reabilitação, regularização fundiária urbana e mobilidade. [2]

Dentro desta área estão as seguintes linhas de pesquisa:

- Projeto e Planejamento Edilício;
- Projeto e Planejamento Urbano e Regional.

O programa conta com uma nota 4 de acordo com a última avaliação Quadrienal dentro do valor estabelecido das notas dos cursos da CAPES, que podem variar de 1 à 7. Segue o quadro com a avaliação do programa, tanto de mestado quanto para doutorado, desde sua criação.

Nível	1996	1998	2001	2004	2007	2008	2009	2010	2011	2012	2013	2014	2017
Mestrado	-	-	4	4	3	3	3	4	4	4	4	4	4
Doutorado	C	3	4	4	3	3	3	4	4	4	4	4	4

O curso atualmente conta 187 alunos de pós-graduação, sendo 114 de Mestrado e 73 de Doutorado. Dentro da função idealizada para a UnB, o programa tem como principal objetivo a formação de recursos humanos de alto nível para o desenvolvimento de atividades de docência e pesquisa, bem como a estruturação de um corpo técnico de especialistas para atuação em órrgãos públicos. Em relação à avaliação anterior, 8 alunos permanecem como estudantes, 17 alunos trabalham fora da esfera pública ou da docência, 57 alunos trabalham no setor de educação e 39 alunos são servidores públicos. À partir disso, pode-se observar que de maneira geral os objetivos têm sido alcançados, com a maior espera de participantes do programa atuam na educação ou serviço público.

O número total de Docentes é de 42, sendo 30 permanentes, 12 colaboradores e 0 visitantes. De acordo com o quadro de Corpo Docente da avaliação, o quadro de DP ou Docentes Permanentes iniciou com 24 docentes permanentes, aumentando para 29 em 2014, com perda de três, totalizando 26 em 2016 e no final do período com os mesmos 29 de 2014. Já o número de colaboradores aumentou de 11 para 12 no período. Em relação à área de atuação do corpo estão: Arquitetura e Urbanismo, Eng. Civil e Construção Civil, Ergonomia, Historia, Sociologia e Tecnologia.

Para o programa de arquitetura e urbanismo, o programa contou com 79 disciplinas dentro das 9 linhas de pesquisas citadas, com registro de 119 Projetos de Pesquisa.

2.1.3 Design

No segundo semestre de 2013 foi criado o curso de mestrado acadêmico em Design. O Programa de Pós Graduação em Design da Universidade de Brasília - PPG Design UnB - faz parte do Instituto de Artes e do Departamento de Design da UnB.

Os objetivos do Programa são:

- Possibilitar a formação de docentes na área de design com vistas a ampliação e melhor qualificação dos docentes na região Centro-Oeste e Norte do país.
- Desenvolver pesquisadores em design, possibilitando o desenvolvimento tecnológico, cultural e econômico do país.
- Aproximar o design da sociedade por meio de pesquisas relacionadas aos processos de produção e inovação com base tecnológica.

O grupo do PPG Design UnB busca construir um programa de qualidade, integrado com as necessidades da pesquisa em design, ao mesmo tempo, ativo na produção do conhecimento e na busca por projetos cooperativos. Através de suas ações, contribui para ampliar as pesquisas em Design desenvolvidas na região, aumentando a produção bibliográfica e a geração de conhecimento e inovação. A diversidade da temática das pesquisas pode ser observada também no projetos de pesquisas dos discentes, sendo que muitos são oriundos de outros estados. Isto demonstra a vocação do PPG Design UnB em agregar pluralidade cultural e interdisciplinar que pode ser vista nos projetos de pesquisas desenvolvidos no programa. [4]

As áreas de concentração de pesquisa em que o Programa se fundamenta são *Design, Tecnologia e Sociedade*.

Segue a tabela disposta das notas do programa em sua última avaliação quadrienal

Nível	2013	2014	2017
Mestrado	3	3	3

2.2 Compreensão dos dados

Com a primeira fase já previamente estabelecida, uma outra etapa tão importante quanto, é a fase de onde somos capaz de inspecionar, organizar e descrever os dados disponíveis para a resolução do problema. É

uma etapa importante também para a familiarização com os dados de insumo para a realização das análises, possibilitando também que alguns *insights* sejam previamente levados em consideração para etapas futuras. [6]

O estudo de todas as tabelas disponíveis e realizar a medição da possibilidade de criação de uma visão única para a análise também é parte dessa fase. [6]

Visto isso, foram realizados os passos abaixo para realizar análises relevantes dos datasets disponíveis.

2.2.1 Importando as bibliotecas

Primeiramente foi importado as bibliotecas para possibilitar a leitura e manipulação dos dados.

```
library(jsonlite) # Biblioteca para lidar com arquivos do formato JSON
library(plyr) # Splitting e combinação de dados
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## ------------------------------------------------------------------------------
```

```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(dplyr) # Manipulação de dados
library(RColorBrewer) # Editar paleta de cores dos gráficos
library(stringr) # Facilita a manipulação de strings
library(ggplot2) # Criação e plot de gráficos
library(tidyr) # Manipulação de Dataframes
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
library(igraph) # Criação de grafos
```

```
##
## Attaching package: 'igraph'
```

```
## The following object is masked from 'package:tidyr':
##
##     crossing
```

```
## The following objects are masked from 'package:dplyr':
##
##     as_data_frame, groups, union
```

```
## The following objects are masked from 'package:stats':
##
##     decompose, spectrum

## The following object is masked from 'package:base':
##
##     union
```

```
Sys.setenv(JAVA_HOME='C:\\Program Files\\Java\\jre1.8.0_191') # Setando o local da instalação do Java
library(rJava) # Java para R (dependência do qdap)
library(qdap) # Manipulação de texto e linguagem natural
```

```
## Loading required package: qdapDictionaries

## Loading required package: qdapRegex

##
## Attaching package: 'qdapRegex'

## The following object is masked from 'package:jsonlite':
##
##     validate

## The following object is masked from 'package:ggplot2':
##
##     %+%

## The following object is masked from 'package:dplyr':
##
##     explain

## Loading required package: qdapTools

##
## Attaching package: 'qdapTools'

## The following object is masked from 'package:plyr':
##
##     id

## The following object is masked from 'package:dplyr':
##
##     id

##
## Attaching package: 'qdap'

## The following objects are masked from 'package:igraph':
##
##     %>%, diversity

## The following object is masked from 'package:tidyr':
##
##     %>%

## The following object is masked from 'package:stringr':
##
##     %>%
```

```
## The following object is masked from 'package:magrittr':
##
##     %>%

## The following object is masked from 'package:dplyr':
##
##     %>%

## The following object is masked from 'package:base':
##
##     Filter
```

```
library(tm) # Mineração de texto
```

```
## Loading required package: NLP

##
## Attaching package: 'NLP'

## The following object is masked from 'package:qdap':
##
##     ngrams

## The following object is masked from 'package:ggplot2':
##
##     annotate

##
## Attaching package: 'tm'

## The following objects are masked from 'package:qdap':
##
##     as.DocumentTermMatrix, as.TermDocumentMatrix
```

```
library(SnowballC) # Steaming
library(wordcloud2) # Visualização em nuvem de palavras
```

2.2.2 Importando arquivos de publicações no formato JSON

O arquivos importados abaixos representam os dados sobre publicações disponibilizados pela plataforma e-Lattes da UnB dos cursos citados acima.

```
# Imoportando dados de Arquitetura
advise_arquitetura <- fromJSON("ArquiteturaPos/235.advise.json")
graph_arquitetura <- fromJSON("ArquiteturaPos/235.graph.json")
list_arquitetura <- fromJSON("ArquiteturaPos/235.list.json")
perfil_arquitetura <- fromJSON("ArquiteturaPos/235.profile.json")
publicacoes_arquitetura <- fromJSON("ArquiteturaPos/235.publication.json")

# Importando dados de Design
advise_design <- fromJSON("DesignPos/267.advise.json")
graph_design <- fromJSON("DesignPos/267.graph.json")
list_design <- fromJSON("DesignPos/267.list.json")
perfil_design <- fromJSON("DesignPos/267.profile.json")
publicacoes_design <- fromJSON("DesignPos/267.publication.json")
```

pais_do_evento <chr>	cidade_do_evento <chr>	doi <chr>	classificacao <chr>	paginas <chr>	autores <list>	autores-endogeno <list>
Brasil	Campinas		NACIONAL	4231 - 4242	<chr [3]>	<chr [1]>
Equador	Quito		INTERNACIONAL	611 - 618	<chr [3]>	<chr [1]>
Brasil	Quito		INTERNACIONAL	841 - 851	<chr [3]>	<chr [1]>
Brasil	Porto		INTERNACIONAL	748 - 755	<chr [2]>	<chr [1]>
Brasil	Vila Real	10.14684/SHEWC.17.2017.173-176	INTERNACIONAL	173 - 176	<chr [2]>	<chr [1]>
Brasil	Florianópolis		INTERNACIONAL	-	<chr [3]>	<chr [1]>
Brasil	Joinville		NACIONAL	-	<chr [10]>	<chr [1]>
Brasil	Lisboa		INTERNACIONAL	-	<chr [3]>	<chr [1]>
Brasil	Lisboa		INTERNACIONAL	-	<chr [9]>	<chr [1]>
Brasil	João Pessoa		NACIONAL	-	<chr [9]>	<chr [1]>

Figure 1: Registros com valores faltantes

2.2.3 Descrição dos Dados

Foi utilizado a função `head()` para ter uma prévia das colunas e alguns registros presentes nos arquivos *json* importados.

```
head(publicacoes_arquitetura)
head(publicacoes_design)
```

Com isso, foi possível observar a presença das seguintes colunas no *json* de publicações.

- *natureza*: Status atual da publicação;
- *titulo*: Título da publicação;
- *nome_do_evento*: Evento no qual foi plubicado;
- *ano_do_trabalho*: Ano que a publicação foi feita;
- *pais_do_evento*: País sede do evento da publicação;
- *cidade_do_evento*: Cidade do evento da publicação;
- *doi*: Código universal da publicação (DOI);
- *classificação*: Internacionalização da publicação;
- *paginas*: Páginas onde se encontra a publicação;
- *autores*: Lista dos criadores da publicação;
- *autores-endogeno*: Autores endógenos da publicação.

2.2.3 Contagem de registros nulos

Na fase de compreensão, também verificamos a qualidade dos dados disponíveis. Ao observar os arquivos importados e suas colunas, nota-se que algumas colunas existem muitos valores nulos em seus registros.

```
# library(installr)
# sum(is.empty(publicacoes_arquitetura$doi))
```

Nota-se que as colunas *doi* e *paginas* contém o maior número de dados faltantes ou inelegíveis. Para fins exploratórios, essas colunas não são relevantes para as análises, então não foi um problema. Outra coluna que tinha vários valores faltantes é a

2.3 Preparação dos Dados

Esta fase abrange todas as atividades para construir o conjunto final de dados, realizando o tratamento, seleção, limpeza e estruturação dos dados disponíveis. O output desta preparação de dados são os dados finais que serão utilizados para o desenvolvimento do modelo. [6]

2.3.1 Transformando em *Dataframes*

Para a obtenção de melhores recursos para manipulação dos dados, foi transformado todos os dados em *dataframes*. Nota-se a presença de novas colunas referentes aos periódicos que foram obtidas através dos comandos baixo:

```
# Arquitetura
df_pub_arquitetura <- sapply(publicacoes_arquitetura, ldply, data.frame) %>% bind_rows() %>% as_tibble()
tbl_arq_publ <- tbl_df(df_pub_arquitetura)

# Design
df_pub_design <- sapply(publicacoes_design, ldply, data.frame) %>% bind_rows() %>% as_tibble()
tbl_design_publ <- tbl_df(df_pub_design)
```

2.3.2 Descrição dos *Dataframes*

Com a adição das novas colunas sobre os periódicos de cada registro, o *dataset* passou de 11 para 27 colunas.

```
# Arquitetura
head(df_pub_arquitetura)

# Design
head(df_pub_design)
```

2.3.3 Separação de Publicações por autor

No atual *dataset*, cada registro de publicação possui um *array* de autores que participaram na confecção do mesmo. Apesar disso, não estava sendo considerado o autor endógeno nesta contagem. Isso poderia projetar informações errôneas sobre os autores.

Então, para contornar este problema, foi utilizado os comandos abaixo para comparar a lista de autores e autores endogenos e analisar o que é distinto entre eles. Por fim foi criado mais uma coluna com os autores reais de cada publicação.

```
# Arquitetura
pub_arquitetura_autores.endo <- df_pub_arquitetura %>% unnest(autores.endogeno) %>% distinct()

# Design
pub_design_autores.endo <- df_pub_design %>% unnest(autores.endogeno) %>% distinct()
```

2.3.4 Separação de Periódicos por Autor

Nesta etapa foi separado os periódicos por cada autor, deixando as informações prontas para serem projetadas em um gráfico para realizarmos a análise.

```
# Arquitetura
pub_arquitetura_periodicos <- pub_arquitetura_autores.endo %>% filter(is.element(pub_arquitetura_autores

# Design
pub_design_periodicos <- pub_design_autores.endo %>% filter(is.element(pub_design_autores.endo$periodic
```

2.3.5 Separação de Periódicos por País

Foi feita também uma manipulação de dados para separar as publicações por país, de modo que também ficasse pronto para uma projeção no `ggplot2`.

```
# Arquitetura
pub_arquitetura_group <- select(pub_arquitetura_periodicos, c("periodico", "pais_de_publicacao"))
arquitetura_qtde_pub_pais <- pub_arquitetura_group %>% add_count(pais_de_publicacao)
arquitetura_qtde_pub_pais_distinct <- arquitetura_qtde_pub_pais %>% distinct(pais_de_publicacao, n)
```

```
# Design
pub_design_group <- select(pub_design_periodicos, c("periodico", "pais_de_publicacao"))
design_qtde_pub_pais <- pub_design_group %>% add_count(pais_de_publicacao)
design_qtde_pub_pais_distinct <- design_qtde_pub_pais %>% distinct(pais_de_publicacao, n)
```

2.4 Modelagem

Essa fase é uma etapa mais voltada para Pesquisa e Desenvolvimento, onde são selecionadas diferentes técnicas de modelagem, para realizar diferentes parâmetros, calibragens e etc. Em um problema de mineração de dados, muitas vezes é possível realizar a análise dos dados utilizando diferentes abordagens, considerando também o tipo e estrutura de dado que essa etapa recebe de input da fase de preparação de dados. [6]

A comunicação com a fase de preparação de dados nessa fase é muito importante, para a experimentação e modelagem utilizando diferentes técnicas e abordagens. [6]

Foi escolhido o *bag of words* como modelo de mineração de texto. A ideia é encontrar as palavras mais constantes no título dos periódicos, a fim de ver a tendência de palavras chaves que os autores mais utilizam em suas publicações.

2.4.1 Construção do *Bag of Words*

Como dito anteriormente, o *Bag of Words* foi o modelo de mineração de texto escolhido para o relatório, que é muito utilizado em algoritmos de processamento de linguagem natural e análise de sentimentos. Seu objetivo é fazer uma representação das palavras mais recorrentes em um texto, desconsiderando alguns fatores como a ordem em que as palavras se encontram e a gramática. No fim o *Bag of Words* tem o *tracking* da quantidade dessas palavras, e podemos tirar conclusões sobre o resultado apresentado.

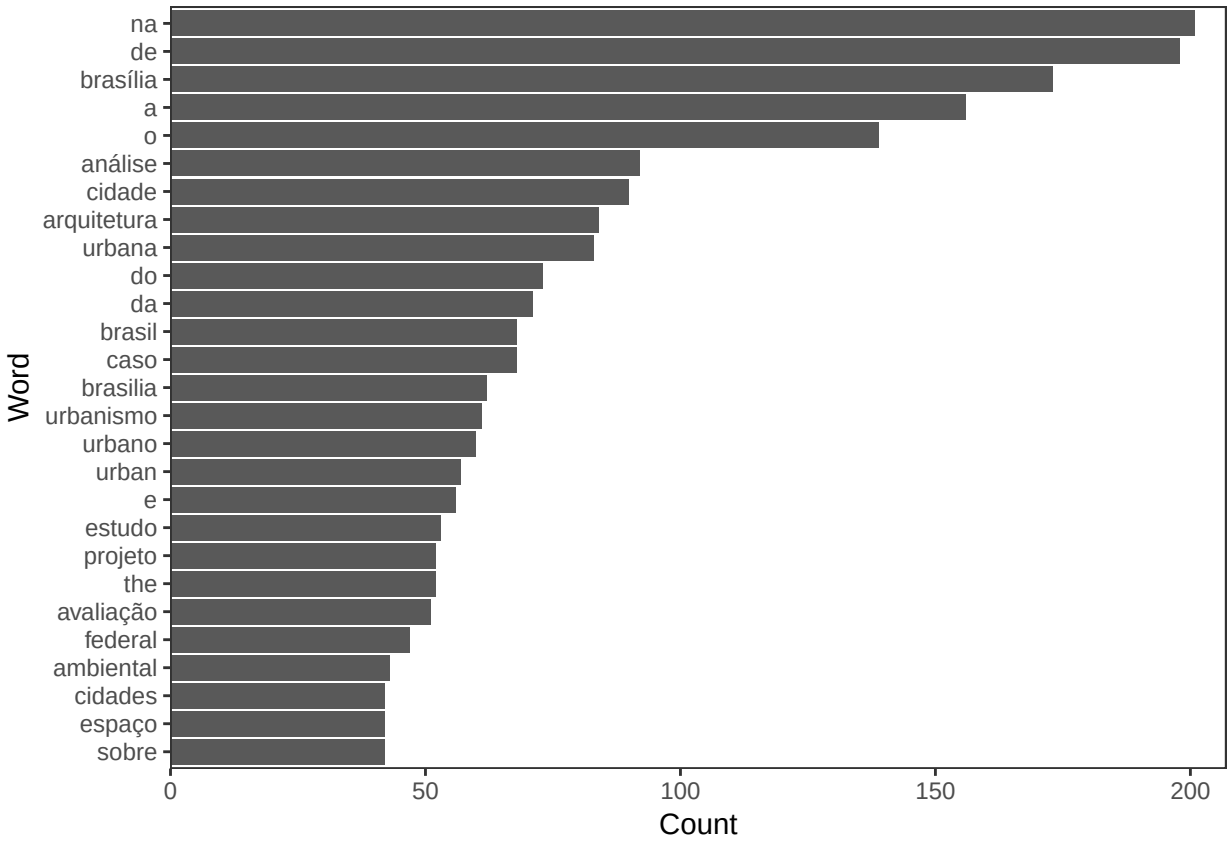
```
# Arquitetura
titulos <- paste(tbl_arq_publ$titulo, collapse = ' ')
titulos <- removeWords(titulos, c(stopwords('pt'), stopwords('en')))
term_count <- freq_terms(titulos, 25)

# Design
titulos2 <- paste(tbl_design_publ$titulo, collapse = ' ')
titulos2 <- removeWords(titulos2, c(stopwords('pt'), stopwords('en')))
term_count2 <- freq_terms(titulos2, 25)
```

O foco do modelo criado foi observar a tendência que os discentes têm ao criar suas publicações, tendo como base o título das publicações. Para tal, a primeira manipulação de dados feita foi agrupar todos os títulos das publicações em um texto só, para poder aplicar o modelo. Isso foi feito tanto para o curso de Arquitetura quanto para o de Design.

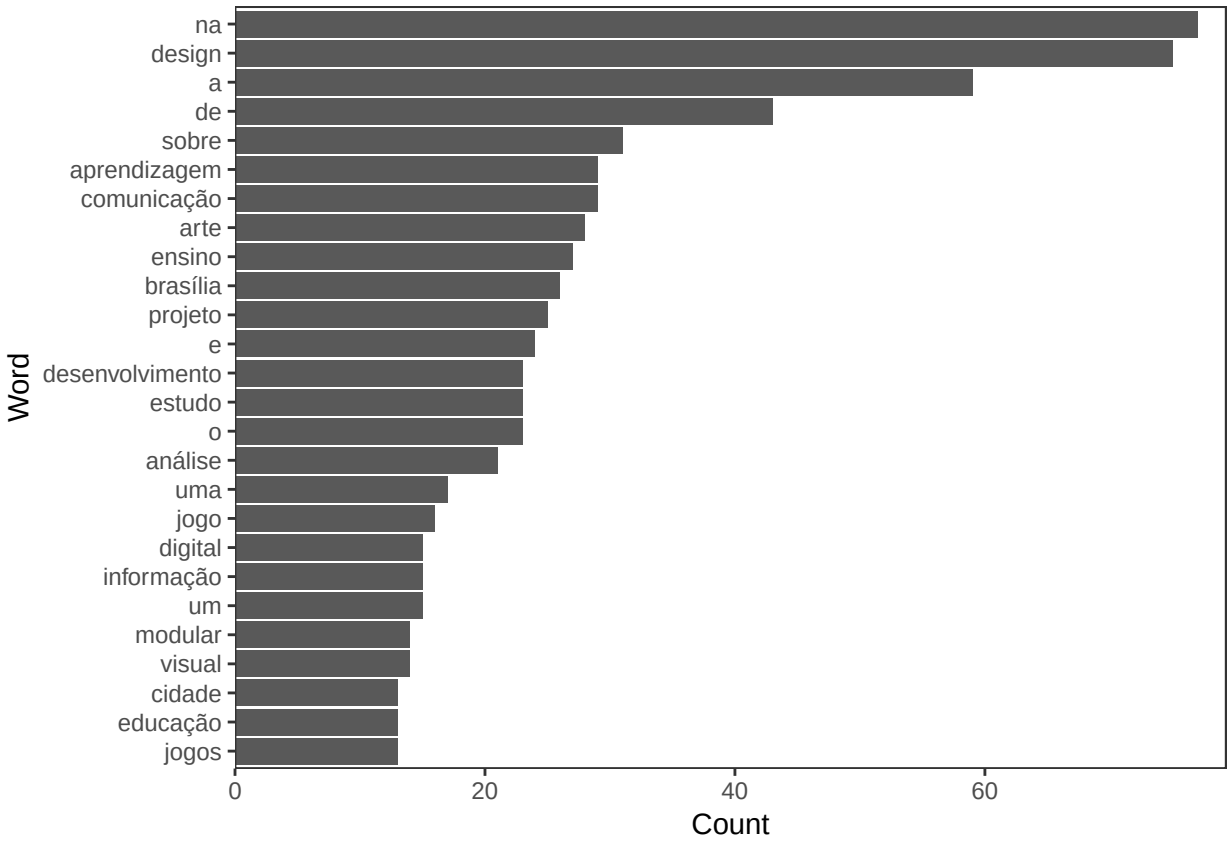
Em seguida, foi retirado dos títulos agrupados as palavras de parada, que são as palavras irrelevantes dentro dos títulos. Foi utilizado o `stopwords` do pacote `qdap`, que contém uma lista de palavras de parada de acordo com o idioma informado. Como há prevalência tanto de publicações em português quanto em inglês, foram utilizadas listas de palavras de parada nestes dois idiomas. Por fim esse agrupamento de listas de palavras foram removidos do agrupamento de títulos.

Após isso, com os dados tratados, a contagem de termos foi feita, onde foi passado como parâmetro o número 25 que diz o número de palavras mais constantes, em ordem decrescente de aparições, para observação inicial da contagem.

```
plot(term_count)
```




```
plot(term_count2)
```

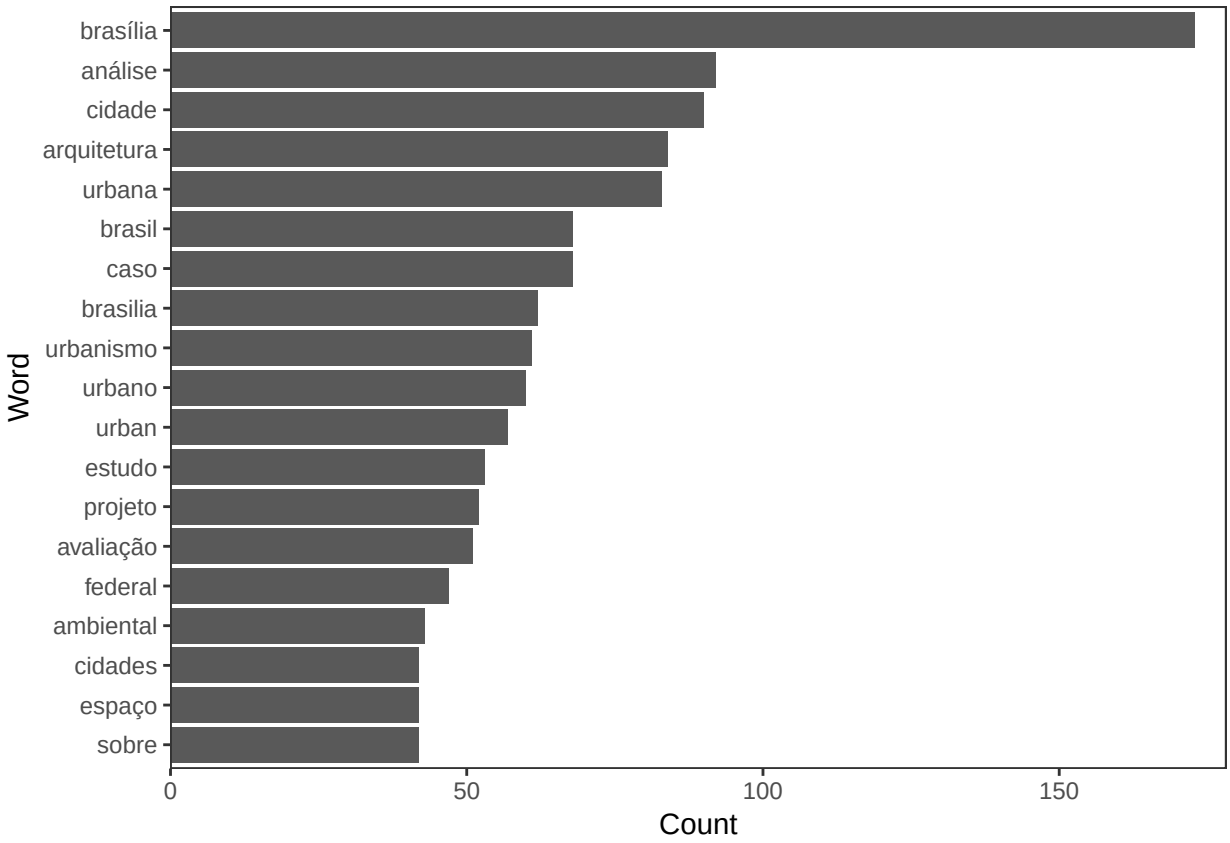


Foi observado que o uso das palavras de parada de cada idioma não foi tão eficaz quanto esperado, onde algumas palavras passaram pelo filtro criado. Para contornar este problema, foi feita outra modificação nos dados, onde foram retiradas as palavras que tinham 3 caracteres ou menos, para se obter um melhor resultado.

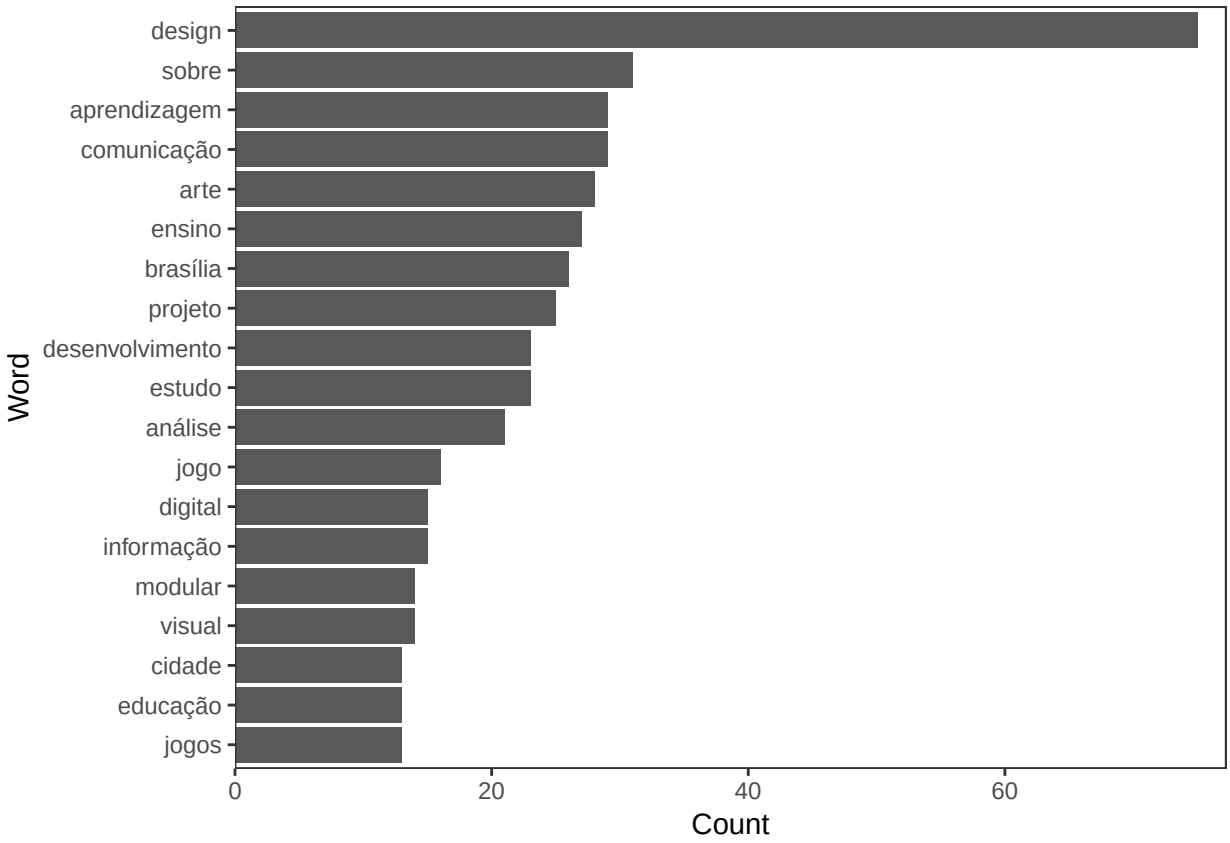
```
# Arquitetura
term_count <- term_count[!(nchar(term_count$WORD) <= 3),]

# Design
term_count2 <- term_count2[!(nchar(term_count2$WORD) <= 3),]
```

Com isso, foi possível plotar o gráfico com a nova contagem de palavras e ver os resultados.

```
plot(term_count)
```




```
plot(term_count2)
```



Após essa manipulação foi possível observar que o modelo com a contagem de palavras chave foi mais eficaz em busca do seu objetivo, pois as palavras irrelevantes foram retiradas.

Com os dados normalizados, foi possível criar um *Bag of Words* mais robusto para posteriormente criarmos uma visualização em nuvem, onde as palavras mais presentes tem tamanho maior na nuvem.

```
# Arquitetura
titulos_bag <- paste(tbl_arq_publ$titulo, collapse = ' ')
titulos_bag <- removeWords(titulos, c(stopwords('pt'), stopwords('en')))
term_count_bag <- freq_terms(titulos, 2500)
term_count_bag <- term_count_bag[!(nchar(term_count_bag$WORD) <= 3),]

# Design
titulos2_bag <- paste(tbl_design_publ$titulo, collapse = ' ')
titulos2_bag <- removeWords(titulos2, c(stopwords('pt'), stopwords('en')))
term_count2_bag <- freq_terms(titulos2, 2500)
term_count2_bag <- term_count2_bag[!(nchar(term_count2_bag$WORD) <= 3),]
```

2.5 Avaliação do Modelo

Com o modelo em mãos é possível realizar a avaliação dos seus resultados de maneira detalhada e revisar as etapas executadas para que os mesmo estejam dentro dos previstos e cumpram os requisitos identificados na fase de entendimento do negócio. [6]

De forma geral, os objetivos traçados na Compreensão de Negócios que envolvia extrair informações relevantes sobre as publicações e sobre seus autores foi alcançado. Foi possível obter várias informações interessantes, onde a aplicação prática da mineração de texto e manipulação de dados será explicada no próximo tópico.

2.6 Desenvolvimento

Com todas as etapas tendo sido percorridas, mesmo não sendo o fim do processo - como citado inicialmente, o processo é iterativo e incremental -, é preciso demonstrar os resultados e para que o modelo preparado seja utilizado. No caso da disciplina, a parte de implantação consiste basicamente em elaborar um relatório final para demonstrar os dados e as mais diversas informações que eles fornecem sobre os projetos de Pós-Graduação de Arquitetura e Urbanismo. Em diferentes contextos, poderia ser utilizado o modelo para entregar informações em tempo real para o usuário final, por exemplo. [6]

2.6.1 Publicações por Ano

Para cada programa disposto para a realização do trabalho, é possível realizar uma contagem de publicações por ano. Dentro da avaliação quadrienal dos programas, um ponto muito importante e que conta para a nota total do programa (40%) é referente à Produção Intelectual realizada no período de avaliação. Não somente a quantidade das produções mas também a qualidade das mesmas são levadas em consideração na hora da avaliação e pontuação dos programas.

Arquitetura

Para o programa de Arquitetura e Urbanismo foram aplicados alguns procedimentos para extrair informações relevantes sobre as publicações.

Quantidade total de publicações

```
length(pub_arquitetura_periodicos$.id)
```

```
## [1] 1178
```

Quantidade de publicações que possuem alguma natureza descrita

```
pub_arquitetura_periodicos %>% group_by(complete.cases(natureza)) %>%
  dplyr::summarize(Quantidade=n())
```

```
## # A tibble: 2 x 2
##   `complete.cases(natureza)` Quantidade
##   <lgl>                           <int>
## 1 FALSE                             186
## 2 TRUE                              992
```

Pode-se observar que os valores onde o natureza = TRUE foram os que possuem alguma natureza, ou seja, 992 das 1178. Os outros 186 são publicações as quais a natureza não foram informadas.

Desses 186, podemos observar os tipos de publicações:

```
pub_arquitetura_periodicos %>% filter(!complete.cases(natureza)) %>% distinct(tipo)
```

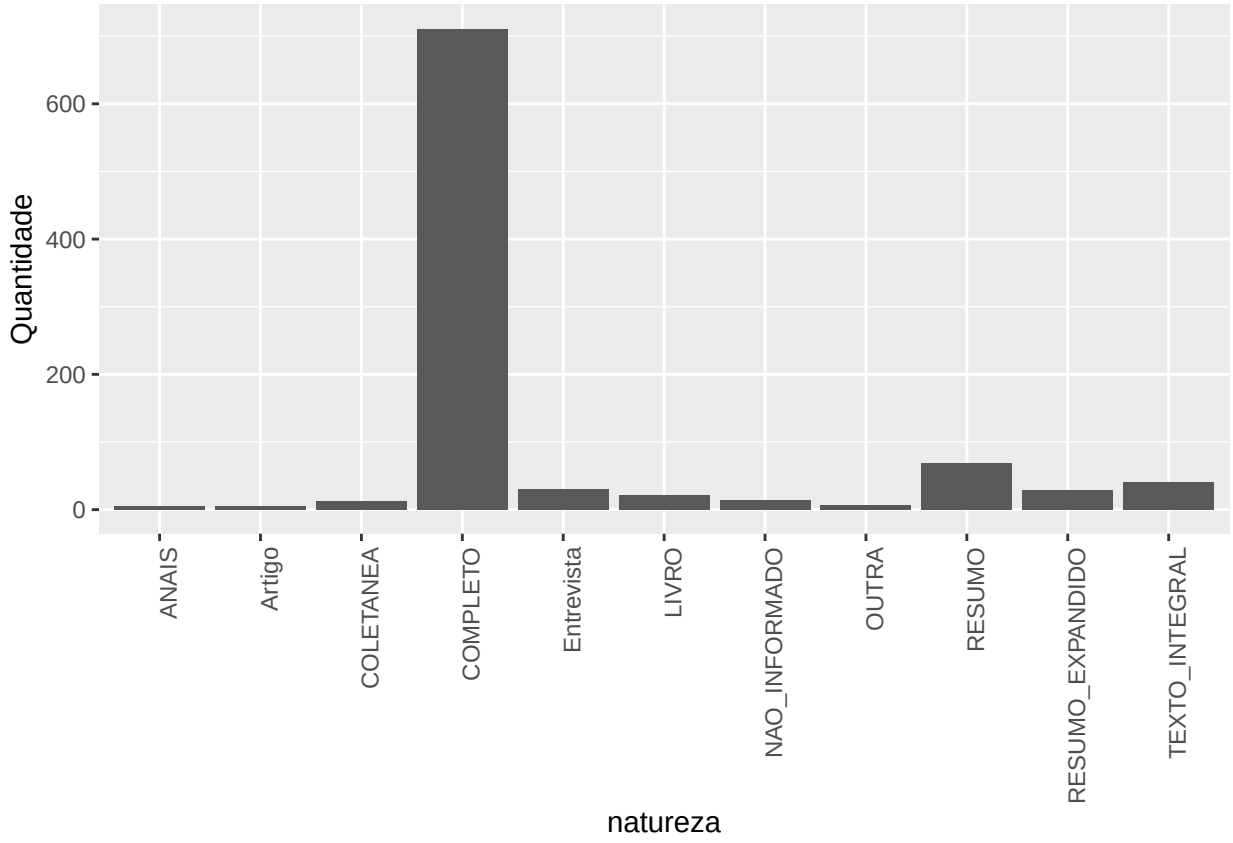
```
## # A tibble: 1 x 1
##   tipo
##   <chr>
## 1 Capítulo de livro publicado
```

Para os tipos de natureza da publicação, as que aparecem mais vezes - apenas as 10 primeiras - foram:

```
pub_grouped_by_length <- pub_arquitetura_periodicos %>%
  filter(complete.cases(natureza)) %>%
  group_by(natureza) %>%
  dplyr::summarize(Quantidade=n()) %>%
  arrange(Quantidade) %>% top_n(10)
```

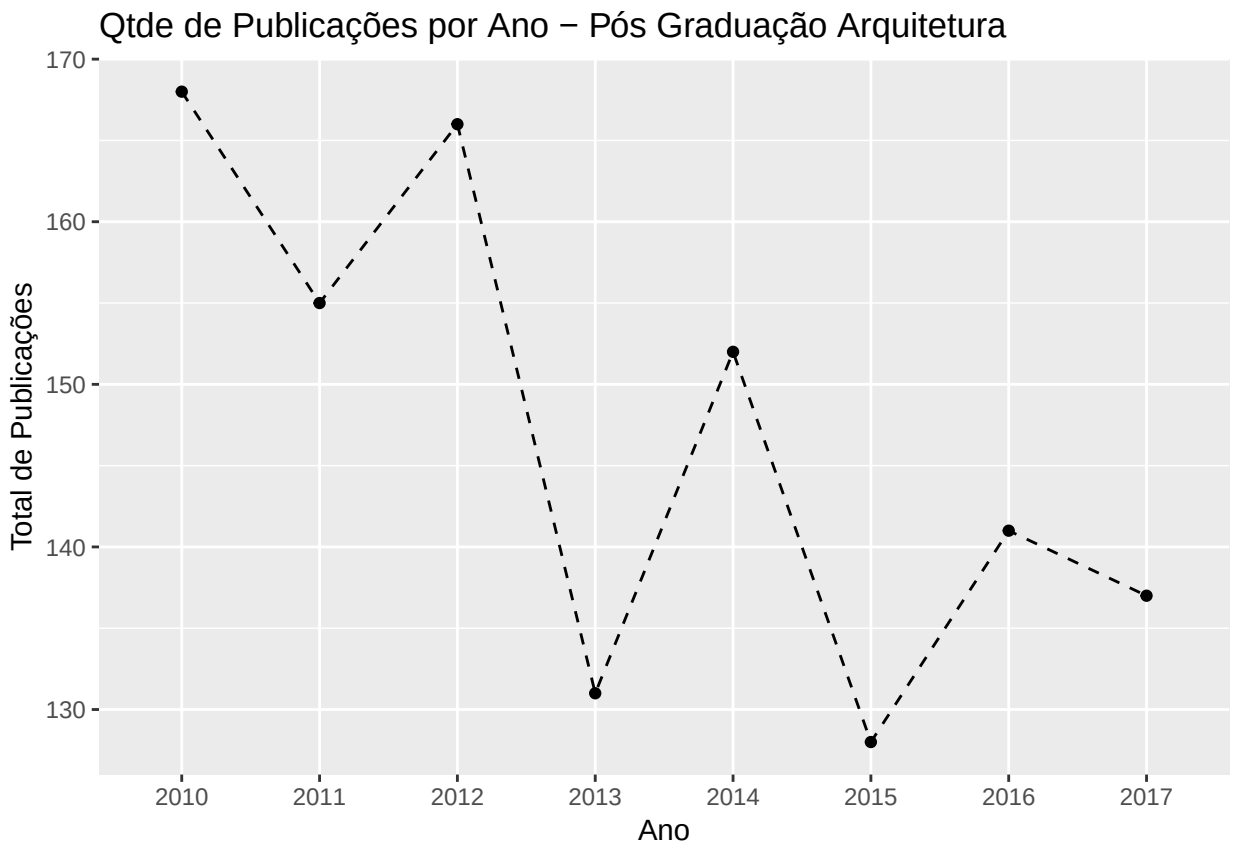
```
## Selecting by Quantidade
```

```
ggplot(data=pub_grouped_by_length, aes(x=natureza, y=Quantidade)) +
  geom_bar( mapping = aes(x=natureza),  stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Com os dados e o acesso ao ano referente à cada publicação, podemos visualizar também a quantidade de publicações por ano.

```
# Arquitetura
periodicos_por_ano <- pub_arquitetura_periodicos %>% group_by(.id) %>% dplyr::summarize(Quantidade=n())
ggplot(periodicos_por_ano, aes(x = .id, y=Quantidade, group = 1 )) +
ggtitle("Qtde de Publicações por Ano - Pós Graduação Arquitetura") +
labs(y = "Total de Publicações", x = "Ano") +
geom_line(linetype = "dashed") +
geom_point()
```




```
mean(periodicos_por_ano$Quantidade)
```

```
## [1] 147.25
```

Neste gráfico podemos ver o que aconteceu com a quantidade de publicações por ano para o programa de Arquitetura: desde a primeira avaliação do programa, do ano de 2010 até o ano de 2017, a quantidade de publicações por ano caiu de 168 (2010) para 137 (2017), a qual terminou com um valor cerca de 7% abaixo da média dos 8 anos (147.25 publicações em média).

Design

O mesmo aplicado para o programa de Arquitetura pode ser aplicado para o programa de Design.

Quantidade total de publicações

```
length(pub_design_periodicos$.id)
```

```
## [1] 422
```

Quantidade de publicações que possuem alguma natureza descrita

```
pub_design_periodicos%>% group_by(complete.cases(natureza)) %>%
  dplyr::summarize(Quantidade=n())
```

```
## # A tibble: 2 x 2
##   `complete.cases(natureza)` Quantidade
```

```
##   <lgl>                          <int>
## 1 FALSE                             73
## 2 TRUE                             349
```

Pode-se observar que os valores onde o natureza = TRUE foram os que possuem alguma natureza, ou seja, 349 das 422. Os outros 73 são publicações as quais a natureza não foram informadas.

Desses 73, podemos observar os tipos de publicações:

```
pub_design_periodicos %>% filter(!complete.cases(natureza)) %>% distinct(tipo)
```

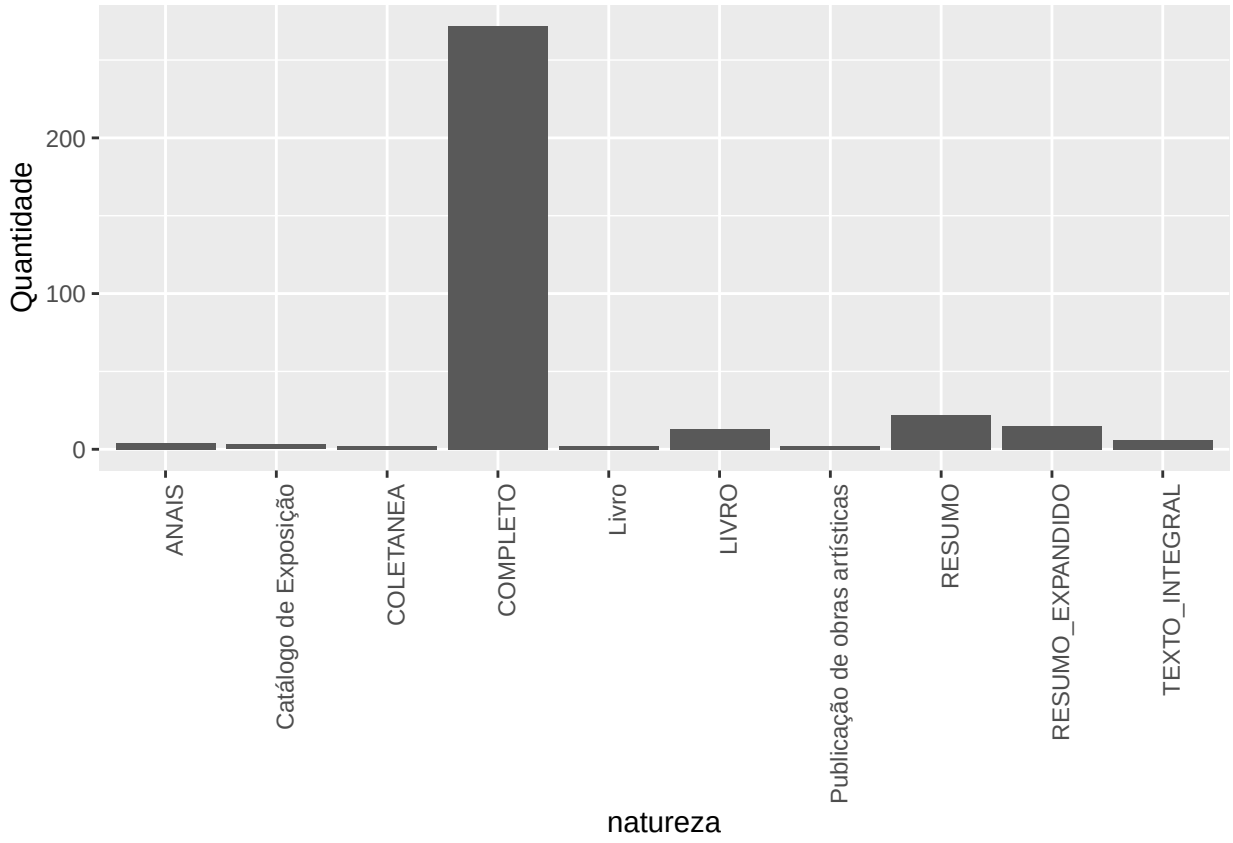
```
## # A tibble: 1 x 1
##   tipo
##   <chr>
## 1 Capítulo de livro publicado
```

Para os tipos de natureza da publicação, as que aparecem mais vezes - apenas as 10 primeiras - foram:

```
pub_design_grouped_by_length <- pub_design_periodicos %>%
  filter(complete.cases(natureza)) %>%
  group_by(natureza) %>%
  dplyr::summarize(Quantidade=n()) %>%
  arrange(Quantidade) %>% top_n(10)
```

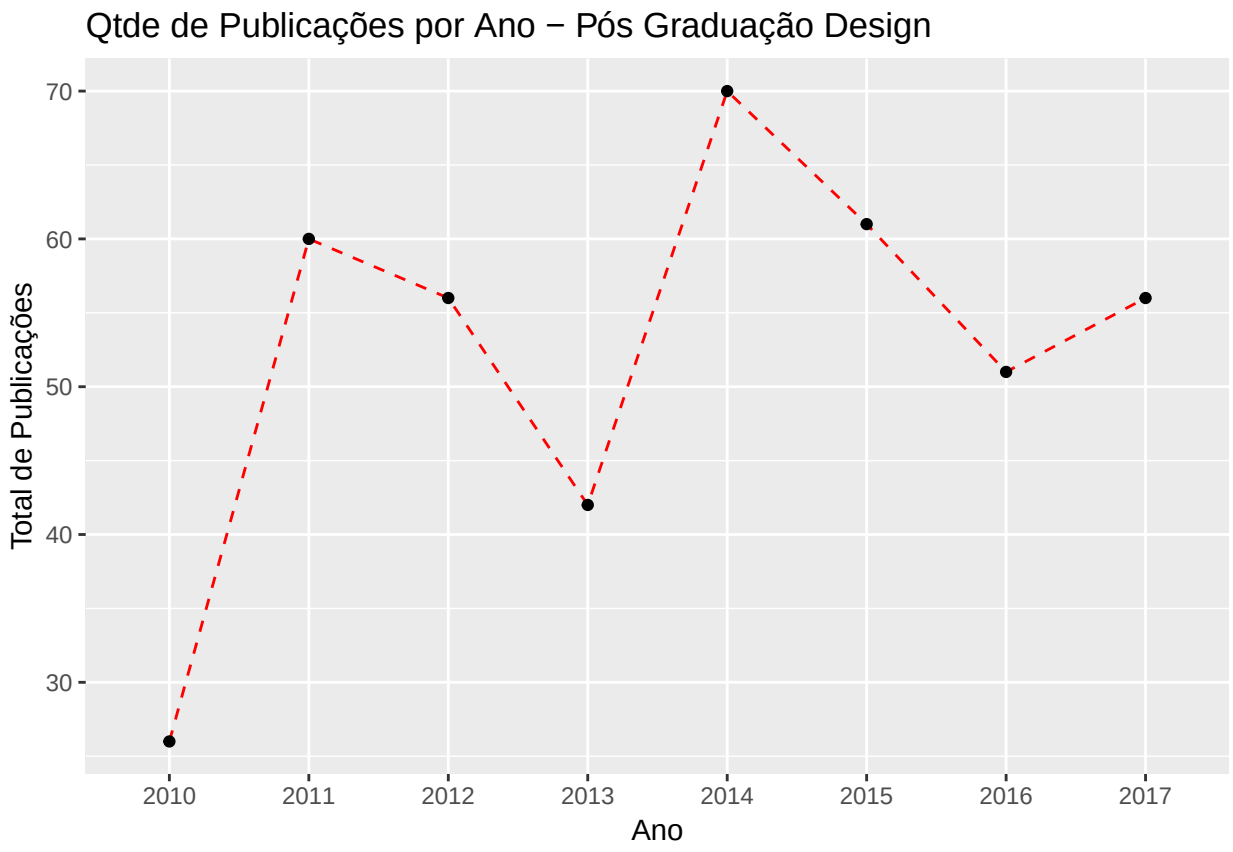
```
## Selecting by Quantidade
```

```
ggplot(data=pub_design_grouped_by_length , aes(x=natureza, y=Quantidade)) +
  geom_bar( mapping = aes(x=natureza),  stat = "identity") +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Com os dados e o acesso ao ano referente à cada publicação, podemos visualizar também a quantidade de publicações por ano.

```
# Design
periodicos_por_ano_design <- pub_design_periodicos %>% group_by(.id) %>% dplyr::summarize(Quantidade=n()
ggplot(periodicos_por_ano_design , aes(x = .id, y=Quantidade, group = 1 )) +
ggtitle("Qtde de Publicações por Ano - Pós Graduação Design") +
labs(y = "Total de Publicações", x = "Ano") +
geom_line(color="red", linetype = "dashed") +
geom_point()
```




```
mean(periodicos_por_ano_design$Quantidade)
```

```
## [1] 52.75
```

Já em relação ao programa de Arquitetura e Urbanismo, o programa de Design teve uma melhora significativa no que se diz respeito à quantidade de publicações. Do ano de 2010 e 2017, a evolução nas publicações foi de 30 publicações, e no ano de 2017 a quantidade ultrapassou a média de 52.75 publicações anuais em aproximadamente 6%.

2.6.2 Publicações por País

É um objetivo generalizado entre os programas de pós graduação internacionalizar suas publicações e ter maior relevância em um âmbito global.

Para mostrar indícios desta internacionalização, foram feitas análises da proporção de publicações por país, mostrando que os cursos de Arquitetura e Design ao longos dos anos vêm cumprido, mesmo que aos poucos, estes objetivos de globalização.

Para visualização destes dados, foi escolhido o gráfico de pizza, que é um diagrama circular que mostra proporcionalmente os valores de acordo com suas frequências. No gráfico de pizza os ângulos representam o percentual de cada valor específico, em vista do valor total. [11]

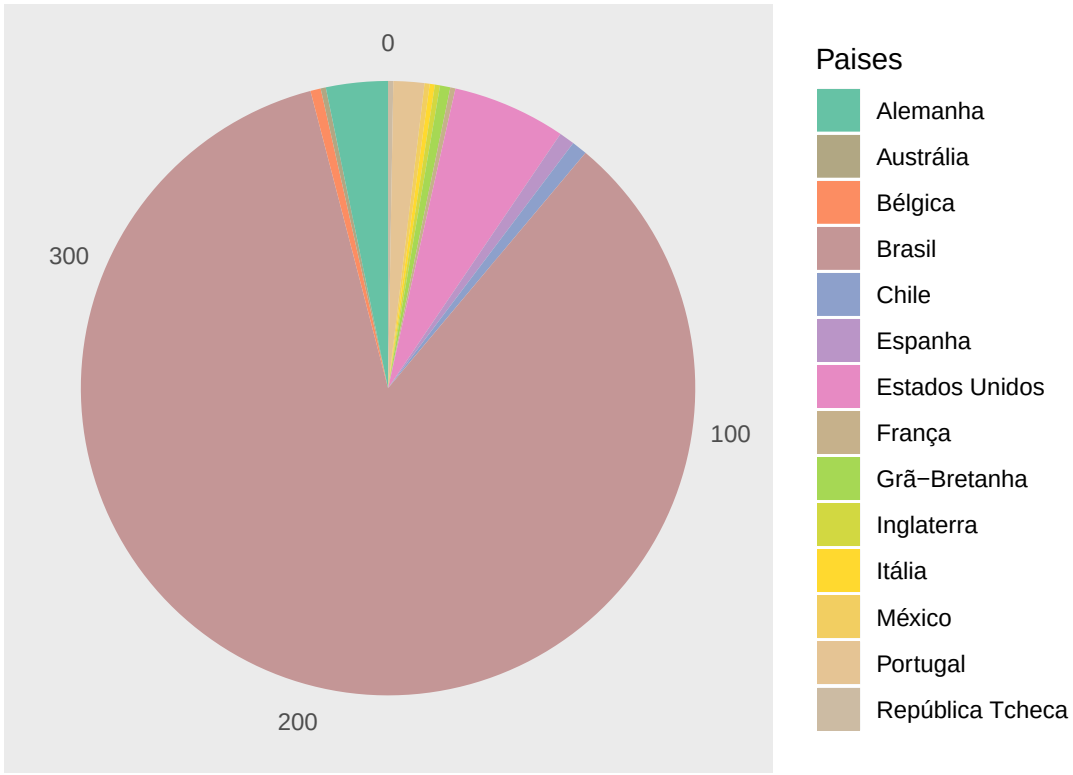
2.6.2.1 Publicações por País - Arquitetura e Urbanismo

Foi plotado o gráfico de pizza das publicações por país do programa. Devido o programa de Arquitetura ter vários países com publicações vinculadas, foi criada uma paleta de cor para este gráfico. Além disso foi mostrado a quantidade de cada publicação por país.

```
cols <- colorRampPalette(brewer.pal(12, "Set2"))
```

```
## Warning in brewer.pal(12, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

```
myPal <- cols(length(unique(arquitetura_qtde_pub_pais_distinct$pais_de_publicacao)))

arquitetura_qtde_pub_pais_distinct <- arquitetura_qtde_pub_pais_distinct[-c(10), ]

ggplot(arquitetura_qtde_pub_pais_distinct, aes(x= "", y= n, fill = arquitetura_qtde_pub_pais_distinct$pa
ggtitle("Qtde de Publicações por País - Pós Graduação Arquitetura") +
geom_bar(width = 1, stat ="identity") +
coord_polar("y") +
labs(fill='Paises') +
theme (
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank()
) +
# geom_text(aes(y = n/2 + c(0, cumsum(n)[-length(n)]), label = percent(n/100)), data = arquitetura_qtde_
scale_fill_manual(values = myPal)
```




```
head(arquitetura_qtde_pub_pais_distinct, n = 14L)
```

```
## # A tibble: 14 x 2
##    pais_de_publicacao     n
##    <chr>              <int>
##  1 Brasil               313
##  2 Estados Unidos        22
##  3 Portugal               6
##  4 Espanha                3
##  5 Bélgica                2
##  6 Chile                  3
##  7 Alemanha              12
##  8 Grã-Bretanha           2
##  9 República Tcheca       1
## 10 França                 1
## 11 México                 1
## 12 Itália                 1
## 13 Inglaterra             1
## 14 Austrália              1
```

De acordo com o gráfico, nota-se que ainda há uma grande prevalência de publicações em português, o que já era um resultado esperado.

Um ponto a se destacar é a presença de publicações em mais de 13 países; o que pode ser reflexo da atuação internacional dos Grupos de Pesquisa do programa de Arquitetura na organização de congressos e seminários internacionais, que resultam em publicações em outros países.

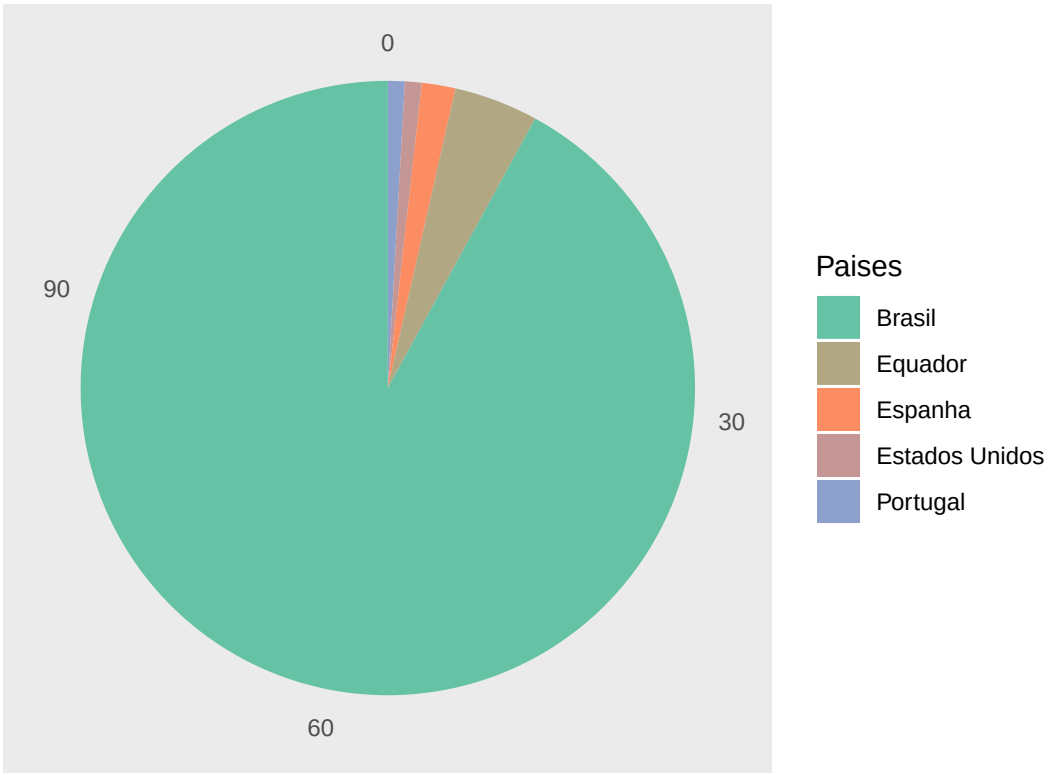
Um dos objetivos descritos nas Propostas do Programa de Arquitetura é ampliar a rede de cooperação internacional e intercâmbio de alunos e pesquisadores. O programa PPG-FAU, que tem participação de

discentes tanto brasileiros quanto estrangeiros, tem resultado em publicações conjuntas entre alunos brasileiros e estrangeiros, além dos professores. Este programa é outro fator que reflete no gráfico das publicações, onde pouco mais de 17% tem participação estrangeira.

2.6.2.2 Publicações por País - Design

Em seguida foi plotado o gráfico de pizza das publicações por país do programa de Design e um quadro com seus dados.

```
design_qtde_pub_pais_distinct <- design_qtde_pub_pais_distinct[-c(6), ]

ggplot(design_qtde_pub_pais_distinct, aes(x= "", y= n, fill = design_qtde_pub_pais_distinct$pais_de_publ
ggtitle("Qtde de Publicações por País - Pós Graduação Design") +
geom_bar(width = 1, stat ="identity") +
coord_polar("y") +
labs(fill='Paises') +
theme (
  axis.title.x = element_blank(),
  axis.title.y = element_blank(),
  panel.border = element_blank(),
  panel.grid=element_blank(),
  axis.ticks = element_blank()
) +
scale_fill_manual(values = myPal, na.value = "#FFa1e2")
```



```r
head(design_qtde_pub_pais_distinct, n = 6L)
```

```
## # A tibble: 5 x 2
##   pais_de_publicacao     n
##   <chr>              <int>
## 1 Brasil               104
## 2 Equador                5
## 3 Espanha                2
## 4 Portugal               1
## 5 Estados Unidos         1
```

Um ponto forte do programa é sua perspectiva de internacionalização constante, independente de indicadores (poucas publicações). Isso ocorre pelo fato do curso ser novo, mas é importante ressaltar os diversos eventos acadêmicos realizados pelo programa sozinho e/ou em associação de outros órgãos/sociedades durante os quatro primeiros anos de sua existência, principalmente o I Seminário Internacional de Pesquisa em Design (2017) que trouxe convidados nacionais de renomadas universidades do país e palestrantes internacionais.

É interessante destacar a parceria com o Programa de Pós-Graduação de Arte da UnB, que desde o início tem dado bons frutos ampliando o contato no sentido de internacionalização com projetos vinculados a mais de 5 países, como mostra o gráfico.

As atividades do programa em relação à internacionalização são eficientes, pois 3 professoras têm Pós-doutorado concluídos com bolsa de pesquisa integral e internacionais, mesmo com pouco tempo de existência. Com isso são esperadas novas publicações internacionais em breve.

2.6.3 Nuvem de palavras

Para visualização dos dados gerados pelo *Bag of Words* de Arquitetura e Design criados, uma das melhores formas para visualização é a Nuvem de Palavras (*Word Cloud*). Ela mostra de forma visual a frequência da ocorrência das palavras, onde, quanto maior o tamanho a visualização de determinada palavra na nuvem, maior sua ocorrência. [10]

O *Bag of Words* dos dois cursos contém as 2500 palavras mais frequêntes nos títulos. Por isso, uma visualização através de histogramas, neste caso, não seriam interessantes.

As visualizações das nuvens de palavras dos cursos de Arquitetura e Design e sua interpretação podem ser visualizados abaixo:

Com a Nuvem de Palavras do curso de Arquitetura e Urbanismo é possível notar que o programa tem buscado capacitar através de publicação envolvendo recursos humanos, políticas públicas e desenvolvimento das cidades. Além disso, com a criação do "Mestrado em Desenho Urbano", também é perceptível o enfoque nos estudos da configuração físico-espacial, acentuando a ênfase nos estudos de urbanismo e das questões correlatas à cidade, consolidando-se como um pólo nas áreas de Planejamento e Desenho Urbano. Todas estas áreas fazem parte da Proposta de Programa do curso, obtido através da Plataforma Sucupira.

Também é importante destacar a linha de Sustentabilidade aplicada à Arquitetura do programa, justificada pela relevância das palavras "Eficiência" (19 vezes), "Água" (19 vezes) e "Natural" (19 vezes).

Palavras como "Cidade" (90 vezes), "Análise" (92 vezes), "Urbana" (60 vezes) e "Urban" (56 vezes) evidenciam o lado voltado para o desenvolvimento das cidades.

Palavras como "Espacial" (25 vezes), "Cidade" (90 vezes), "Social" (28 vezes), "Distrito" (34 vezes), "Distric" (9 vezes) e "Public" (15 vezes) evidenciam a ênfase da configuração físico-espacial.

Ao observar a Nuvem de Palavras do curso de Design, foi possível perceber que seus discentes seguem uma linha ligada aos objetivos gerais e específicos do programa. Um de seus objetivos é possibilitar o desenvolvimento tecnológico, cultural e econômico do país através de pesquisas em Design, capacitando bons pesquisadores na área.

Palavras como "Ensino" (27 vezes), "Estudo" (23 vezes) e "Pesquisa" (10 vezes) evidenciam esse lado voltado para pesquisa dos objetivos do programa.

Palavras como "Digital" (15 vezes) e "Digitais" (12 vezes), "Jogos" (12), "Robô" (6) evidenciam o lado tecnológico do objetivo.

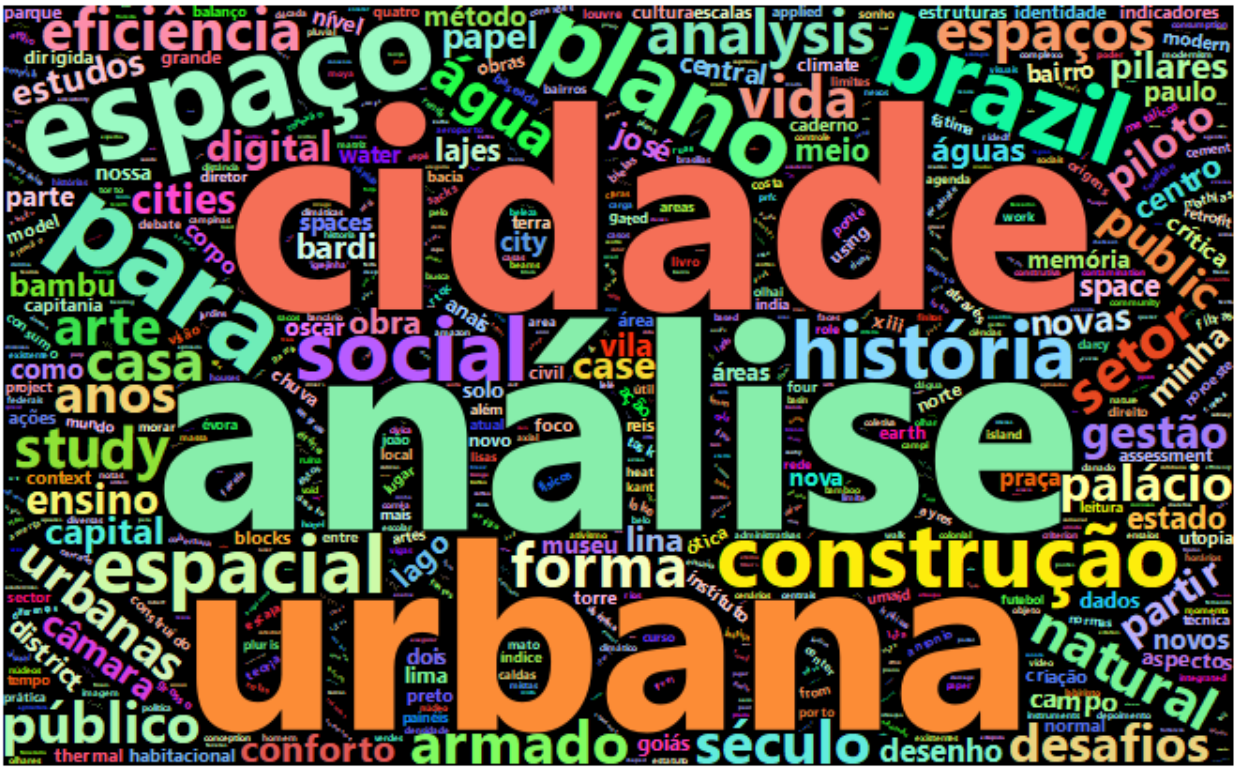


Figure 2: Departamento de Arquitetura - Nuvem de palavras chave

Figure 3: Departamento de Design - Nuvem de palavras chave

Outras palavras como “Arte” (28 vezes) e “Cultural/Cultura” (11 vezes) evidenciam o lado cultural do objetivo.

Essa grande variação de palavras com contextos distantes, condiz bem com o perfil dos discentes da pós graduação de Design, que, de acordo com a proposta de programa disponibilizado pela CAPES, o curso realiza pesquisas e atrai várias áreas, como Engenharia, Educação, Comunicação e Artes.

2.6.4 Análise de perfil e relacionamento de docentes

Um outro modelo de análise de dados é realizar, à partir do arquivo `graph.json` criar uma espécie de “rede de relacionamentos” entre os autores das diversas publicações dos dois programas dispostos. Isso é possível utilizando a biblioteca `igraph` e permite visualizar, de maneira gráfica por meio de um grafo, qual docente publica juntamente com outros docentes, como se da essa relação, etc.

Arquitetura

Para o programa de arquitetura temos os seguintes dados:

```
# Quantidade de docentes
length(perfil_arquitetura)
```

```
## [1] 42
```

```
# Diferentes áreas de atuação de grande área
unique(unlist(sapply(
  sapply(
    perfil_arquitetura, function(x) x$areas_de_atuacao$grande_area
  ), function(x) x)
))
```

```
## [1] "CIENCIAS_HUMANAS"           "CIENCIAS_SOCIAIS_APLICADAS"
## [3] "LINGUISTICA_LETRAS_E_ARTES" "ENGENHARIAS"
## [5] ""                           "CIENCIAS_EXATAS_E_DA_TERRA"
```

```
# Diferentes áreas de atuação de área
unique(unlist(sapply(
  sapply(
    perfil_arquitetura, function(x) x$areas_de_atuacao$area
  ), function(x) x)
))
```

```
##  [1] "Filosofia"                    "Arquitetura e Urbanismo"
##  [3] "Artes"                        "Planejamento Urbano e Regional"
##  [5] "Engenharia Civil"             "Comunicação"
##  [7] "História"                     "Letras"
##  [9] "Educação"                     "Sociologia"
## [11] ""                             "Desenho Industrial"
## [13] "Engenharia de Transportes"    "Engenharia de Produção"
## [15] "Engenharia Sanitária"         "Geociências"
```

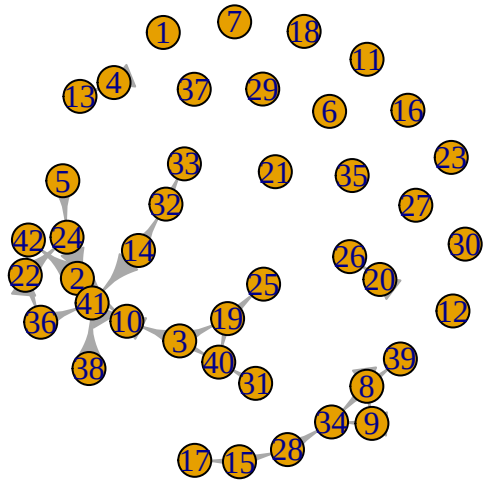
Os docentes se concentram em 5 grandes áreas que se dividem em outras 15 áreas.

Para a relação de publicação desses professores, ou seja, qual professor publica com qual temos:

```
nodes <- graph_arquitetura$nodes
nodes$name <- nodes$properties$name
nodes$properties <- NULL
nodes
```

```
##                   id label                                name
## 1  0186938954124031     1             Miguel Gally de Andrade
## 2  0287848411905739     2         Marta Adriana Bustos Romero
## 3  0414142132580629     3          Cláudia Naves David Amorim
## 4  0505514768226348     4             Rodrigo Santos de Faria
## 5  0718568790650154     5                 Monica Fiuza Gondim
## 6  1218180830143253     6          Eduardo Pierrotti Rossetti
## 7  1740856393308920     7               Cláudia Estrela Porto
## 8  1766107518538219     8   Ana Elisabete de Almeida Medeiros
## 9  1796841203235489     9               Elane Ribeiro Peixoto
## 10 2593051036451822    10              Caio Frederico e Silva
## 11 2861675901591901    11         Claudia da Conceição Garcia
## 12 2969670428458905    12                  Flavio René Kothe
## 13 2978568603389237    13                  Benny Schvarsberg
## 14 3178676800830416    14         Carlos Eduardo Luna de Melo
## 15 3228132238224370    15                      Sylvia Ficher
## 16 3384102431683538    16                  Sergio Rizo Dutra
## 17 3458904726233428    17                   Ricardo Trevisan
## 18 3486886457253995    18         Virginia de Siqueira Barros
## 19 3699291030017215    19         José Manoel Morales Sánchez
## 20 3886525216818735    20               Neander Furtado Silva
## 21 3904543030005792    21          Jaime Gonçalves de Almeida
## 22 4095512091254454    22    Frederico Rosa Borges de Holanda
## 23 4241493690450737    23      Paulo Roberto Carvalho Tavares
## 24 4671263508814146    24  Valério Augusto Soares de Medeiros
## 25 5507673131571546    25 Carolina Pescatori Candido da Silva
## 26 5623972959038895    26              Francisco Leite Aviani
## 27 5654879697444080    27               Maria Fernanda Derntl
## 28 5767592881382885    28      Pedro Paulo Palazzo de Almeida
## 29 5942041330805370    29     Marcos Thadeu Queiroz Magalhães
## 30 6031008181034694    30                  Julia Issy Abrahao
## 31 6313415782219480    31            Márcio Albuquerque Buson
## 32 6339433870219875    32           Márcio Augusto Roma Buzar
## 33 6879105340639188    33                Joao da Costa Pantoja
## 34 7477058101351856    34         Luciana Saboia Fonseca Cruz
## 35 7899321988947015    35   Carlos Henrique Magalhães de Lima
## 36 8302669735998016    36           Gabriela de Souza Tenorio
## 37 8485430007155664    37      Maria do Carmo de Lima Bezerra
## 38 8626162886111282    38             Daniel Richard Sant'Ana
## 39 8653037550466873    39        Ivan Manoel Rezende do Valle
## 40 9313626132229535    40           Vanda Alice Garcia Zanoni
## 41 9661028896672932    41         Liza Maria Souza de Andrade
## 42 9959872053260157    42        Rômulo José da Costa Ribeiro
```

```
relations <- graph_arquitetura$links

g <- graph_from_data_frame(relations, directed=TRUE, vertices=nodes)
plot(g)
```



Onde podemos observer que a docente Liza Maria Souza de Andrade é o que mais aparece em publicações relacionado com outros docentes.

Design

Para o programa de design temos os seguintes dados:

```
# Quantidade de docentes
length(perfil_design)
```

```
## [1] 13
```

```
# Diferentes áreas de atuação de grande área
unique(unlist(sapply(
  sapply(
    perfil_design, function(x) x$areas_de_atuacao$grande_area
  ), function(x) x)
))
```

```
## [1] "CIENCIAS_SOCIAIS_APLICADAS" "CIENCIAS_EXATAS_E_DA_TERRA"
## [3] "LINGUISTICA_LETRAS_E_ARTES" "CIENCIAS_HUMANAS"
## [5] "OUTROS"                     "ENGENHARIAS"
```

```
# Diferentes áreas de atuação de área
unique(unlist(sapply(
  sapply(
    perfil_design, function(x) x$areas_de_atuacao$area
```

```
  ), function(x) x)
))
```

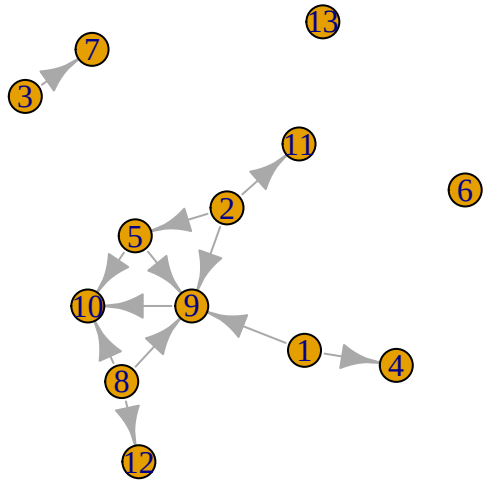
```
##  [1] "Desenho Industrial"
##  [2] "Psicologia"
##  [3] "Ciência da Informação"
##  [4] "Ciência da Computação"
##  [5] "Artes"
##  [6] "Comunicação"
##  [7] "Educação"
##  [8] "Ciências Ambientais"
##  [9] "Engenharia Mecânica"
## [10] "Robótica, Mecatrônica e Automação"
## [11] "Engenharia Civil"
```

Os docentes se concentram em 6 grandes áreas que se dividem em outras 11 áreas.
Para a relação de publicação desses professores, ou seja, qual professor publica com qual temos:

```
nodes <- graph_design$nodes
nodes$name <- nodes$properties$name
nodes$properties <- NULL
nodes
```

```
##                  id label                                    name
## 1  0727171902745709     1            Tiago Barros Pontes e Silva
## 2  1113356615802141     2              Virgínia Tiradentes Souto
## 3  1114384517661428     3                      Marisa Cobbe Maass
## 4  1158798915256133     4            Christus Menezes da Nobrega
## 5  1414472867163536     5            Fátima Aparecida dos Santos
## 6  2146841563850436     6 Silvio Romero Botelho Barreto Campello
## 7  2977168844333740     7          Célia Kinuko Matsunaga Higawa
## 8  3462848897175909     8                 Dianne Magalhães Viana
## 9  4684085149051278     9                    Rogerio José Camara
## 10 4887266557816573    10              Daniela Fávaro Garrossini
## 11 6119310102978688    11                 Ricardo Ramos Fragelli
## 12 7249895826115757    12                  Shirley Gomes Queiroz
## 13 7406666262139231    13                  Ana Claudia Maynardes
```

```
relations <- graph_design$links

g <- graph_from_data_frame(relations, directed=TRUE, vertices=nodes)
plot(g)
```



Onde podemos observer que o docente Rogerio José Camara é o que mais aparece em publicações relacionado com outros docentes.

3. Conclusão

Os programas de pós graduação de Arquitetura e Urbanismo e Design da Universidade de Brasília (UnB) contém grandes fontes de dados e é uma rica fonte de informações para análise, mas como qualquer outro Big Data, seu estudo necessita ser aprofundado. Para isso, a Ciência de Dados é o recurso do conhecimento capaz de trazer informações relevantes dessas grandes massas de dados, e assim fornecer um meio para tomadas de decisão e predições.

Para trabalhar de forma correta e produtiva com Big Data, é recomendado a implementação de metodologias de mineração de dados, e o CRISP-DM é uma boa alternativa, pois tem um processo bem desenhado e atividades bem definidas para tal.

Em relação aos programas, nota-se que eles estão em constante crescimento, seja em número de docentes, discentes ou publicações, e os cursos de Arquitetura e Design não são diferentes. Conclui-se que programa de Arquitetura, através das análises dos dados, tem buscado ao longo dos anos cumprir os objetivos gerais e específicos da Proposta do Programa e tem obtido êxito. O programa da UnB tem um âmbito global com docentes e várias publicação internacionais, em 13 países diferentes. O conteúdo de suas publicações remetem fielmente aos assuntos citados no Histórico e Contextualização do Programa, envolvendo sustentabilidade, desenvolvimento de cidades e outros assuntos como urbanismo, contrução e espaço.

Conclui-se também que o programa de Design, apesar de não ser bem pontuado pela avaliação quadrienal, tem objetivos bem definidos e busca alcança-los. Seus objetivos gerais e específicos incluem capacitar pesquisadores na área de Design tendo foco em auxiliar o crescimento do país tecnologia, cultura e economia, e seus estudos e publicações seguem fielmente essa linha de pensamento. Apesar de ser um programa recente, a Pós-Graduação em Design da UnB tem grande potencial para virar referencia tanto nacionalmente, quanto internaciolnamente, visto que suas publicações fora do Brasil tem crescido regularmente.

Referências

1. Histórico da PGG - FAU.
2. Áreas de Concentração - FAU.
3. COLETA DE DADOS Conceitos e orientações - Manual de preenchimento da Plataforma Sucupira.
4. PGG - Departamento de Design.
5. Introdução à Ciência de Dados 2.0 - Data Science Academy.
6. Afinal, o que é Data Science? - ABG Consultoria Estatística.
7. O que é a 4ª revolução industrial - e como ela deve afetar nossas vidas
8. A sociedade do conhecimento
9. Formação docente e o campo educacional: políticas, regulações e processos. In: OLIVEIRA, Dalila Andrade; VIEIRA, Lívia Fraga (Orgs.). Trabalho na Educação Básica: a condição docente em sete estados brasileiros. Belo Horizonte: Fino Traço, 2012. p. 131-151.
10. Crie a sua própria nuvem de palavras - ONBIZ
11. Gráfico de pizza - Portal Action