Ciência de Dados para Todos (Data Science For All) - 2018.1 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Modelo de Relatório Final da Disciplina - Departamento de Ciência da Computação da UnB

Jorge H. C. Fernandes, Ricardo B. Sampaio, João Ribas de Moura e Jerônimo A. Filho

11/06/2018

Introdução

Este documento apresenta um modelo básico para a construção do relatório final da disciplina Tópicos Avançados em Computadores - Turma D - 2018.2, do Departamento de Ciência da Computação da Universidade de Brasília, que trata da análise da produção científica e acadêmica na Universidade de Brasília, em uma ou mais áreas de conhecimento.

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.

Fernandes, Jorge H C, Ricardo Barros Sampaio, João Ribas de Moura e Jerônimo AVelar Filho. "Ciência de Dados para Todos (Data Science For All) - 2018.1 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Modelo de Relatório Final da Disciplina - Departamento de Ciência da Computação da UnB". Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2018.1, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 13 de junho de 2018.

CRISP-DM (Corresponderia à seção de Metodologia)

Para desenvolvimento do trabalho devem ser seguidos, da forma mais simplificada e coerente possível, as fases e atividades genéricas do ciclo de vida de um projeto executado em aderência ao CRISP-DM, conforme ilustra de forma geral a figura 1. Em outras palavras, a produção do relatorio deve seguir a metodologia CRISP-DM.

Fases do Ciclo de Vida do CRISM-DM. Fonte: (Chapman et al., 2000).

Perceba que a Figura 1 sugere haver grande flexibilidade na execução das fases, de modo que se pode retornar a fases anteriores em muitos pontos.

"A widely used methodology for data mining is the CRoss-Industry Standard Process for Data Mining (CRISP-DM) which mas initiated in 1996 (. . .) with the intent of providing a process that is **reliable and repeatable** by people with little data-mining background, with a framework within which experience can be recorded, to support the replication of projects, to support planning and management, as well as to demonstrate data mining as a mature discipline (. . .)" [Sullivan, Rob. Introduction to Data Mining for the Life Sciences. Springer Science & Business Media. 2012]

O seu trabalho deve conter uma seção metodologia, onde você faz uma breve descrição da metodologia que seu grupo adotou para realização do trabalho, que pode ser baseada no texto dessa seção, desde que citado adequadamente.

Delimitações iniciais

Em aderência à estrutura do CRISP-DM, algumas delimitações de contexto para o trabalho são apresentadas a seguir:

Domínio de Aplicação do projeto

O domínio de aplicação do projeto é o da produção científica e acadêmica brasileira, mais especificamente a produção científica ou produção acadêmica de um subgrupo de pesquisadores vinculados à Universidade de Brasília. O domínio de aplicação do projeto deve ser declarado na introdução ao relatório.

Tipo de Problema abordado

O tipo de problema abordado é o da produção de análises descritivas, quantitativas e de modelagem computacional ou estatística, que permitam caracterizar como e porque ocorre a produção científica e acadêmica de um grupo de pesquisadores. Essa caracterização visa subsidiar a tomada de decisão por membros do Sistema Nacional de Pós-Graduação. O tipo de problema abordado no projeto deve ser declarado na introdução ao relatório.

Conjunto de Ferramentas e Técnicas

O conjunto de requisitos relativos a ferramentas e técnicas a serem empregadas na execução do trabalho é:

- O relatório deve ser entregue no formato R Markdown, apto à geração de saída LaTeXe PDF, composto por comandos em R entremeados por descrições textuais que auxiliem na interpretação dos resultados, bem como na compreensão do domínio de conhecimento sob análise.
- As análises descritivas devem empregar de forma criativa as funções das bibliotecas de ciência de dados em R propostas por Wickham e Grolemund (2016).
- As análises quantitativas devem lançar mão de recursos gráficos variados, que complementarão análises descritivas com *insights* sobre de que forma os processos de produção científica e acadêmica contribuem para os resultados apresentados. Por exemplo, os dados analisados possibilitam justificar o eventual crescimento ou decréscimo de índices de produção observados?
- A modelagem computacional ou estatística avançada dos dados deve usar uma das quatro técnicas prescritas:
  - Aprendizado de Máquina (Datacamp, 2018; Kuhn et al., 2018; Bruce e Bruce, 2017);
  - Aprendizado Estatístico;
  - Mineração de Texto ou;
  - Análise de Redes (Kolaczyk e Csárdi, 2014; Lusher et al., 2013; de Nooy et al., 2005).

O conjunto de requisitos relativos a ferramentas e técnicas a serem empregadas na execução do trabalhodo projeto deve ser declarado na parte de metodologia do relatório.

Modelo de Referência CRISP-DM

Miner (2012), aprofunda: "(. . .) In CRISP-DM, the complete life cycle of a data mining project is represented with **six phases**: business understanding (determining the purpose of the study), data understanding (data exploration and understanding), data preparation, modeling, evaluation, and deployment.(. . .). [Miner, Gary.

Practical Text Mining and Statistical Analysis for Non-structure Text Data Applications. Academic Press, 2012.]

Por que usar o CRISP-DM?

Imagine uma analogia entre um projeto de datamining e a preparação de uma receita de bolo para ser usada em uma fábrica. Para iniciar a produção, com base numa receita de comprovada eficácia (metodológica e científica), você tem que minerar os ingredientes (dados) em um grande supermercado (*dataset*). Com os ingredientes você precisa aplicar um método (a forma de misturá-los), colocar os ingredientes numa determinada ordem, mexer por um certo tempo, aquecer por tantos minutos até o bolo ficar pronto e ser aprovado em um ou mais testes de degustação.

Tendo por objetivo fazer com que essa receita (script de mineração de dados) possa ser executada com sucesso diversas vezes, numa fábrica, será que outro cozinheiro (cientista) que reproduzisse a receita (método) chegaria ao mesmo resultado? Se a metodologia (receita) já foi bastante testada, então é bem provável que o resultado será o mesmo e seu produto (receita de bolo) será aceito para a produção (*deployment*) de análises para consumo futuro, com base em fundamentos científicos.

Organização hierárquica de atividades em fases

Dentro de cada fase no CRISP-DM existe uma estrutura hierárquica de atividades genéricas para serem realizadas. Cada uma dessas atividades **genéricas** pode determinar a execução de atividades **específicas**.

Voltando ao exemplo do bolo, a atividade " 1. Entendimento do Bolo" poderia conter uma atividade genérica chamada "1.1. Determinar para que o bolo servirá (simples café da manhã? bolo de aniversário? bolo de casamento?)". Dentro dessa atividade genérica poderia haver atividades específicas como "1.1.1.Entrevistar o contratante para obter detalhes de onde o bolo será usado¿'; "1.1.2. Conversar com os convidados sob alguma necessidade especial (sem lactose? sem glútem?)", etc.

Seis Fases do CRISP-DM

Com base no apresentado, segue uma descrição um pouco mais detalhada das seis fases de um projeto no CRISP-DM, interpretadas no contexto do relatório que você e seu grupo deverão produzir.

Todas as fases deverão ser adequadamente relatadas no relatório, em seções que aparecem após a seçãoda metodologia

1. O propósito da fase de **Entendimento do Negócio** é o desenvolvimento dos objetivos e declaração das necessidades do projeto sob a perspectiva do negócio, para transformar isso tudo em definição de um problema de data mining.

As atividades genéricas dentro dessa fase envolvem:

- Identificar o que a organização realmente necessita alcançar. No caso específico desta disciplina, a necessidade do Sistema Nacional de Pós-Graduação do Brasil de produzir análises de alta qualidade de suas pós-graduações, com baixo custo. Como produzir um projeto de mineração de dados se você não sabe o que necessita encontrar ou resolver? Se você não entender os objetivos da organização pode levar ao erro de procurar as respostas certas para as perguntas erradas.
- Avaliação das Circunstâncias. Envolve identificar quais recursos ou dificuldades podem influenciar os objetivos da mineração ou do projeto em si. No caso específico desta disciplina, isso envolve refletir, entre vários outros aspectos, sobre as limitações de tempo do projeto, que precisa ser realizado dentro de um semestre letivo, de modo que considerável parte das atividades já foram pré-organizadas pelos docentes responsáveis pela disciplina.

- O projeto de mineração é o grande objetivo desta etapa e o relatório precisa conter uma seção sobre Metodologia, apresentando em detalhes o que se pretende fazer adiante.

2. A fase de **Entendimento dos Dados** inicia determinando quais são os dados realmente disponíveis na organização, se existe permissão para utilizá-los, se existem dados confidenciais ou cobertos pelo sigilo. Por exemplo, um *dataset* das declarações de imposto de renda da Receita Federal certamente seria protegido pelo sigilo fiscal. Dados de pacientes de hospitais podem conter restrições.

Também é necessário acessar os dados para compreendê-los melhor para ter o *insight* de como será feita a modelagem mais tarde.

Na fase de entendimento dos dados pode-se trabalhar com quatro atividades genéricas:

- Coleta inicial dos dados. Essa atividade envolve a análise das permissões de acesso e outras questões envolvendo sigilo e outros proprietários dos dados (terceiros). Por exemplo, eu poderia estar acessando uma base de dados que foi obtida de outro órgão por convênio, mas nesse convênio (contrato) não foi dada permissão para qualquer outro tipo de acesso ou exploração dos dados. Neste projeto, a coleta inicial foi feita pelos autores deste relatório. O relatório final deve conter indicações de como foi realizada a coleta inicial dos dados.

- Descrição dos dados. 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 este modelo já oferece uma descrição inicial. O relatório final deve conter descrições significativas e aprofundadas dos dados.

- Análise exploratória dos dados. A análise exploratória dos dados possibilita um entendimento mais profundo da relação estatística existente entre os dados dos *datasets* para um melhor entendimento da qualidade daqueles dados para o objetivo do projeto. Neste projeto, a análise exploratória dos dados é responsabilidade parcial dos alunos, tendo em vista que este relatório apresenta uma análise exploratória preliminar. O relatório final deve conter análises exploratórias dos dados que sejam significativas e aprofundadas.

- Verificação da qualidade dos dados. A verificação da qualidade dos dados envolve responder se os dados disponíveis estão realmente completos. As informações disponíveis são suficientes para o trabalho proposto? Neste projeto, a verificação da qualidade dos dados é responsabilidade dos alunos.

3. Na fase de **Preparação dos Dados** os *datasets* que serão utilizados em todo o trabalho são construídos a partir dos dados brutos. Aqui os dados são "filtrados" retirando-se partes que não interessam e selecionando-se os "campos" necessários para o trabalho de mineração.

São 5 as atividades genéricas nesta fase de preparação dos dados:

- Seleção dos dados. Envolve identificar quais dados, da nossa "montanha de dados", serão realmente utilizados. Quais variáveis dos dados brutos serão convertidas para o *dataset*? Não é raro cometer o erro de selecionar dados para um modelo preditivo com base em uma falsa ideia de que aqueles dados contém a resposta para o modelo que se quer construir. Surge o cuidado de se separar o sinal do ruído (Silver, Nate. The Signal and the Noise: Why so many predictions fail — but some don't. USA: The Penguin Press HC, 2012.).

- Limpeza dos dados.

- Construção dos dados. Envolve a criação de novas variáveis a partir de outras presentes nos *datasets*.

- Integração dos dados. Envolve a união (merge) de diferentes tabelas para criar um único *dataset* para ser utilizado no R, por exemplo.

- Formatação dos dados. Envolve a realização de pequenas alterações na estrutura dos dados, como a ordem das variáveis, para permitir a execução de determinado método de data mining.

4. A fase de **Modelagem** no CRISP-DM envolve a construção e avaliação do modelo, podendo ser realizada em quatro atividades genéricas:

- Seleção das técnicas de modelagem.
- Realização de testes de modelagem, onde diferentes modelos são previamente testados e avaliados. Pode-se dividir o *dataset* criado na etapa anterior para se ter uma base de treino na construção de modelos, e outra pequena parte para validar e avaliar a eficiência de cada modelo criado até se chegar ao mais "eficiente".
- Construção do modelo definitivo, com base na melhor experiência do passo anterior.
- Avaliação do modelo.

5. Na fase de **Avaliação** do CRISP-DM os resultados não são apenas avaliados, mas se verifica se existem questões relacionadas à organização que não foram suficientemente abordadas. Deve-se refletir se o uso arepetido do modelo criado pode trazer algum "efeito colateral" para a organização.

Nesta fase, pode-se trabalhar com 3 atividades genéricas:

- Avaliação dos resultados
- Revisão do processo, por meio da qual verifica-se se o modelo foi construído adequadamente. As variáveis (passadas) para construir o modelo estarão disponíveis no futuro?
- Determinação dos etapas seguintes. Pode ser necessário decidir-se por finalizar o projeto, passar à etapa de desenvolvimento, ou rever algumas fases anteriores para a melhoria do projeto.

6. Na fase de **Implantação** (*deployment*) se realiza o planejamento de implantação dos produtos desenvolvidos (scripts, no caso do executado nesta disciplina) para o ambiente operacional, para seu uso repetitivo, envolvendo atividades de monitoramento e manutenção do sistema (script) desenvolvido. A fase de implantação concluir com a produção e apresentação do relatório final com os resultados do projeto.

São atividades genéricas na fase de **implantação**:

- Planejamento da transição dos produtos;
- Planejamento do monitoramento dos produtos em utilização no ambiente operacional;
- Planejamento de manuteção a ser eventualmente efetuada no produto (scripts);
- Produção do relatório final;
- Apresentação do relatório final;
- Revisão sobre a execução do projeto, com registro de lições aprendidas etc.

No contexto do projeto realizado no âmbito desta disciplina, a responsabilidade por execução de todas essas atividades é dos alunos, com exceção da apresentação do relatório final, que não será realizada.

CRISP-DM Fase 1 - Entendimento do Negócio

O que é o Sistema Nacional de Pós-Graduação? (Contextualização)

A produção do conhecimento científico, no Brasil, é predominantemente efetuada por meio do Sistema Nacional de Pós-Graduação - SNPG, e mais fortemente relacionada com a formação de doutores nesse sistema (Pátaro e Mezzomo, 2013), por meio de cursos de pós-graduação *strictu sensu*.

Fernandes e Sampaio (2017) já indicaram que a ciência é reconhecidamente um elemento essencial para o desenvolvimento social e econômico de qualquer nação. Assim sendo, faz-se mister aprimorar o SNPG como forma de promoção desse crescimento, visando maximizar o retorno decorrente do emprego dos recursos nele aplicados. A promoção do crescimento do SNPG se dá predominantemente por meio de avaliações regulares de seus programas de pós-graduação, sob responsabilidade da CAPES, que realiza a cada quatro

anos um complexo (Leite, 2018, p. 13) e custoso processo de coleta de dados, análise e deliberação sobre as pós-graduações *strictu sensu*, em coerência com o estabelecido no Plano Nacional de Pós-Graduação (PNPG) 2012-2020 (CAPES, 2010) e nos diversos documentos que definem os critérios de organização da pós-graduação em cada área do conhecimento (CAPES, 2018). Leite (2018) faz uma apresentação geral de como se organizam e são avaliadas as pós-graduações no Brasil.

O Plano Nacional de Pós-Graduação (PNPG), por outro lado, define diretrizes estratégicas para desenvolvimento da pós-graduação brasileira, que deve abordar prioritariamente grandes temas de interesse nacional, tais como a redução das assimetrias de desenvolvimento entre as regiões do Brasil, a formação de professores para a educação básica, a formação de recursos humanos para as empresas, a resposta aos grandes desafios brasileiros sobre Água, Energia, Transporte, Controle de Fronteiras, Agronegócio, Amazônia, Amazônia Azul (Mar), Saúde, Defesa, Programa Espacial, além de Justiça, Segurança Pública, Criminologia e Desequilíbrio Regional. O PNPG também traça as diretrizes para financiamento da pós-graduação e sua internacionalização, apresentando conclusões e recomendações.

As avaliações do SNPG, ao atribuirem mensurações de desempenho às diversas pós-graduações que dele fazem parte, geram incentivos e penalidades aos programas, tendo em vista a limitada disponibilidade de recursos para investimento em bolsas, taxas de bancada etc. Embora o sistema seja altamente sofisticado ele é também altamente criticado (Azevedo et al., 2016), sobretudo porque há percalços na busca por um equilíbrio entre as diferentes concepções de finalidade da ciência. Se de um lado a promoção do conhecimento gerado predominantemente nas ditas ciências *hard* constribui para criar fluxos econômicos mais intensos, isso não significa que essa promoção possa ocorrer em detrimento da menor promoção na geração de conhecimento sobre problemas sociais, predominantemente gerado nas ditas ciências *soft*, especialmente das áreas de humanidades, sob pena de ampliação de desigualdades (Azevedo et al., 2016).

Não há solução simples, mas postula-se, nesta disciplina, que uma maior agilidade na avaliação e a utilização de critérios mais objetivos, poderá facilitar a melhoria do sistema.

Os Colégios, Grandes Áreas e Áreas da Pós-Graduação Brasileira

A partir de 2018, as diversas áreas da pós-graduação brasileira foram organizadas na forma de colégios, grandes áreas e áreas, conforme apresentam as tabelas a seguir.

Colégio de Ciências da vida

CIÊNCIAS AGRÁRIAS	CIÊNCIAS BIOLÓGICAS	CIÊNCIAS DA SAÚDE
Ciência de Alimentos	Biodiversidade	Educação Física
Ciências Agrárias I	Ciências Biológicas I	Enfermagem
Medicina Veterinária	Ciências Biológicas II	Farmácia
Zootecnia / Recursos Pesqueiros	Ciências Biológicas III	Medicina I
-	-	Medicina II
-	-	Medicina III
-	-	Nutrição
-	-	Odontologia
-	-	Saúde Coletiva

Colégio de Ciências Exatas, Tecnológicas e Multidisciplinar

CIÊNCIAS EXATAS E DA TERRA	ENGENHARIAS	MULTIDISCIPLINAR
Astronomia / Física	Engenharias I	Biotecnologia
Ciência da Computação	Engenharias II	Ciências Ambientais
Geociências	Engenharias III	Ensino

CIÊNCIAS EXATAS E DA TERRA	ENGENHARIAS	MULTIDISCIPLINAR
Matemática / Probabilidade e Estatística	Engenharias IV	Interdisciplinar
Química	-	Materiais

Colégio de Humanidades

CIÊNCIAS HUMANAS	CIÊNCIAS SOCIAIS APLICADAS	LINGUÍSTICA, LETRAS E ARTES
Antropol/Arqueol	Admin.Púb./Empr.,C.Contáb. e Tur.	Artes
Ciência Pol. e Rel. Int.	Arquit., Urban. e Design	Linguística e Literatura
Ciências da Religião e Teol.	Comunicação e Informação	-
Educação	Direito	-
Filosofia	Economia	-
Geografia	Planej. Urbano e Reg. / Demografia	-
História	Serviço Social	-
Psicologia	-	-
Sociologia	-	-

Cada um desses colégios, grandes áreas e áreas de conhecimento possuem dinâmicas próprias, e, portanto, não há um modelo universal que se aplique a todas. Existem aspectos comuns, mas também grandes peculiaridades, descritas parcialmente nos correspondentes documentos de área disponíveis em CAPES (2018).

A UnB dentro do Sistema Nacional de Pós-Graduação (Contextualização)

O que é a UnB?

Descrição da Universidade de Brasília, com foco na sua produção científica e acadêmica.

Descrição das pós-graduações da UnB

Texto a desenvolver.

Outros aspectos que caracterizam a produção científica e acadêmica da UnB

Texto a desenvolver.

O que a Organização precisa realmente alcançar?

Vários stakeholders estão envolvidos no projeto em curso, e poderíamos considerar cada um deles como distintas organizações que possuem interesses distintos e complementares. Elas são: * A Disciplina Ciência de Dados para Todos 2018.1, que quer comprovar que seus alunos dominam ferramentas e técnicas de ciência de dados, para fins de avaliação de rendimento da disciplina. * A UnB, representada pelos decanatos de pós-graduação (DPG) e de pesquisa e inovação (DPI), que querem dispor de instrumentos para realização de avaliações contínuas de suas pós-graduações. * O SNPG, que assim com o DPG e DPI, também pode se beneficiar do uso de instrumentos para realização de avaliações contínuas de suas pós-graduações. * Os interessados em melhor conhecer o que é produzido pelo Sistema Nacional de Pós-graduação, como empresas privadas, que querem desfrutar dos benefícios gerados pela ciência brasileira.

A fim de dar maior fidelidade e homogeneidade ao exercício realizado na disciplina, focaremos em atendimento aos interesses comuns das organizações DPI, DPG e CAPES, que desejam dispor de instrumentos ágeis para avaliação contínua da pós-graduação brasileira.

Com base no exposto, o objetivo do trabalho final a ser alcançado pelos produtos d emineração de dados desenvolvido pelos alunos da disciplina Ciência de Dados para Todos é produzir, tomando por base inicial os dados fornecidos pelos professores responsáveis pela disciplina, ferramentas para análise e avaliação contínuas e de baixo custo, do desempenho de um conjunto de pós-graduações que estão vinculadas a uma mesma subárea ou grupo de conhecimento. Cada área de pós-graduação apresenta suas características peculiares, assim como cada um dos programas vinculados a essas áreas. Como já informado, características peculiares de cada programa podem ser obtidas a partir de visita ao sítio da CAPES (2018).

Avaliação das Circunstâncias

Este documento serve como base para a realização dos trabalhos dos alunos. apresenta limitações no tocante à quantidade pequena de dados que serão empregados para análises e avaliações, tendo em vista sua finalidade maior que é a didática, de permitir aos alunos demonstrarem a capacidade de aplicação das técnicas e ferramentas apreendidas durante o semestre.

Avaliação preliminar das pós-graduações na UnB

Texto a desenvolver.

Avaliação preliminar da produção científica e acadêmica da UnB

Texto a desenvolver.

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

```
json.perfil <- "data/UnBPosGeral/profile.json"
file.info(json.perfil)
```

```
## Warning in as.POSIXlt.POSIXct(x, tz): unknown timezone 'zone/tz/2018e.1.0/
## zoneinfo/America/Sao_Paulo'
```

```
##                                    size isdir mode               mtime
## data/UnBPosGeral/profile.json 75162725 FALSE  644 2018-06-22 08:22:26
##                                           ctime               atime uid
## data/UnBPosGeral/profile.json 2018-06-28 01:14:24 2018-09-15 02:32:31 502
##                               gid  uname grname
## data/UnBPosGeral/profile.json  20 Barros  staff
```

O arquivo data/UnBPosGeral/profile.json apresenta dados sobre o perfil de todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017. Esse arquivo foi fornecido pelos docentes responsáveis pela disciplina.

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

```
json.advise <- "data/UnBPosGeral/advise.json"
file.info(json.advise)
```

```
##                                  size isdir mode               mtime
## data/UnBPosGeral/advise.json 29828920 FALSE  644 2018-06-22 08:22:26
##                                          ctime               atime uid
## data/UnBPosGeral/advise.json 2018-06-28 01:14:11 2018-09-15 02:32:31 502
##                              gid  uname grname
## data/UnBPosGeral/advise.json  20 Barros  staff
```

O arquivo data/UnBPosGeral/advise.json apresenta dados sobre o orientações de mestrado e doutorado feitas por todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017. Esse arquivo foi fornecido pelos docentes responsáveis pela disciplina.

Produção bibliográfica gerada pelos docentes vinculados às pós-graduações

```
json.producao.bibliografica <- "data/UnBPosGeral/publication.json"
file.info(json.producao.bibliografica)
```

```
##                                       size isdir mode               mtime
## data/UnBPosGeral/publication.json 33546293 FALSE  644 2018-06-22 08:22:26
##                                              ctime               atime
## data/UnBPosGeral/publication.json 2018-06-28 01:14:31 2018-09-15 02:32:31
##                                   uid gid  uname grname
## data/UnBPosGeral/publication.json 502  20 Barros  staff
```

O arquivo data/UnBPosGeral/publication.json apresenta dados sobre a produção bibliográfica gerada por todos os docentes vinculados a programas de pós-graduação da UnB, entre 2010 e 2017.

Agrupamento dos docentes conforme áreas de atuação

```
json.researchers_by_area <- "data/UnBPosGeral/researchers_by_area.json"
file.info(json.researchers_by_area)
```

```
##                                             size isdir mode
## data/UnBPosGeral/researchers_by_area.json 64366 FALSE  644
##                                                      mtime
## data/UnBPosGeral/researchers_by_area.json 2018-06-22 08:22:26
##                                                      ctime
## data/UnBPosGeral/researchers_by_area.json 2018-06-28 01:14:36
```

```
##                                                       atime uid gid
## data/UnBPosGeral/researchers_by_area.json 2018-09-15 02:32:31 502  20
##                                            uname grname
## data/UnBPosGeral/researchers_by_area.json Barros  staff
```

O arquivo data/UnBPosGeral/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.

```
file.info('data/UnBPosGeral/graph.json')
```

```
##                               size isdir mode               mtime
## data/UnBPosGeral/graph.json 503798 FALSE  644 2018-06-22 08:22:26
##                                            ctime               atime uid
## data/UnBPosGeral/graph.json 2018-06-28 01:14:18 2018-09-15 01:54:20 502
##                              gid  uname grname
## data/UnBPosGeral/graph.json  20 Barros  staff
```

Redes de colaboração entre docentes

O arquivo data/UnBPosGeral/graph.json apresenta 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.

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(readtext)
```

```
##
## This data.table install has not detected OpenMP support. It will work but slower in single threaded r
```

```
library(ggplot2)
library(tidyverse)
```

```
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## filter(): dplyr, stats
## lag():    dplyr, stats
```

```
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

O arquivo unb.perfis.json, que contém dados que caracterizam o perfil profissional de todos os docentes do grupo sob análise, podem ser lido por meio do comando seguinte.

```
unb.prof <- fromJSON("data/UnBPosGeral/profile.json")
```

A quantidade de docentes sob análise é apresentada a seguir.

```
length(unb.prof)
```

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

Para gerar uma apresentação inicial dos dados que estão contido nos dados de perfil dos docentes, pode-se usar a função glimpse, da biblioteca dplyr, como ilustra o código seguinte, que apresenta os atributos típicos que podem ser obtidos relativamente a um pesquisador específico, o mais antigo docente ainda em exercício na UnB a ter criado seu registro na plataforma Lattes.

```
glimpse(unb.prof[[1]], width = 30)
```

```
## List of 7
##  $ nome                  : chr "Norai Romeu Rocco"
##  $ resumo_cv             : chr "Possui gradua\u00e7\u00e3o em Matem\u00e1tica (licenciatura plena) p
##  $ areas_de_atuacao      :'data.frame':  5 obs. of  4 variables:
##   ..$ grande_area  : chr [1:5] "CIENCIAS_EXATAS_E_DA_TERRA" "CIENCIAS_EXATAS_E_DA_TERRA" "CIENCIAS_E
##   ..$ area         : chr [1:5] "Matem\u00e1tica" "Matem\u00e1tica" "Matem\u00e1tica" "Ci\u00eancia da
##   ..$ sub_area     : chr [1:5] "" "\u00c1lgebra" "\u00c1lgebra" "Matem\u00e1tica da Computa\u00e7\u0
##   ..$ especialidade: chr [1:5] "" "" "Grupos de \u00c1lgebra N\u00e3o-Comutaviva" "Matem\u00e1tica S
##  $ endereco_profissional :List of 8
##   ..$ instituicao: chr "Universidade de Bras\u00edlia"
##   ..$ orgao      : chr "Instituto de Ci\u00eancias Exatas"
##   ..$ unidade    : chr "Departamento de Matem\u00e1tica"
##   ..$ DDD        : chr "061"
##   ..$ telefone   : chr "31076442"
##   ..$ bairro     : chr "Asa Norte"
##   ..$ cep        : chr "70910900"
##   ..$ cidade     : chr "Bras\u00edlia"
##  $ producao_bibiografica :List of 4
##   ..$ ARTIGO_ACEITO                         :'data.frame':   1 obs. of  10 variables:
##   .. ..$ natureza        : chr "NAO_INFORMADO"
##   .. ..$ titulo          : chr "Finiteness conditions for the non-abelian tensor product of groups"
##   .. ..$ periodico       : chr "MONATSHEFTE FUR MATHEMATIK"
##   .. ..$ ano             : chr "2017"
##   .. ..$ volume          : chr ""
##   .. ..$ issn            : chr "00269255"
##   .. ..$ paginas         : chr " - "
##   .. ..$ doi             : chr "10.1007/s00605-017-1143-x"
##   .. ..$ autores         :List of 1
##   .. ..$ autores-endogeno:List of 1
##   ..$ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA:'data.frame':   7 obs. of  9 variables:
##   .. ..$ natureza          : chr [1:7] "DIVULGA\u00c7\u00c3O DE RESULTADOS DE PESQUISA" "DIVULGA\u00
##   .. ..$ titulo            : chr [1:7] "NON-ABELIAN TENSOR SQUARE OF FINITE-BY-NILPOTENT GROUPS" "The
##   .. ..$ ano               : chr [1:7] "2015" "2016" "2016" "2016" ...
##   .. ..$ pais_de_publicacao: chr [1:7] "Estados Unidos" "Estados Unidos" "Estados Unidos" "Estados Un
##   .. ..$ editora           : chr [1:7] "" "ArXiv.com - Cornell University Library" "ArXiv.com - Corne
##   .. ..$ doi               : chr [1:7] "" "" "" "" ...
##   .. ..$ numero_de_paginas : chr [1:7] "8" "12" "12" "11" ...
```

```
##   .. ..$ autores           :List of 7
##   .. ..$ autores-endogeno  :List of 7
##   ..$ EVENTO                                :'data.frame':   1 obs. of  11 variables:
##   .. ..$ natureza        : chr "RESUMO"
##   .. ..$ titulo          : chr "On Semidirect Products and non-abelian Tensor Products of Groups"
##   .. ..$ nome_do_evento  : chr "XIX Col\u00f3quio Latinoamericano de \u00c1lgebra"
##   .. ..$ ano_do_trabalho : chr "2012"
##   .. ..$ pais_do_evento  : chr "Chile"
##   .. ..$ cidade_do_evento: chr "Puc\u00f3n - Chile"
##   .. ..$ doi             : chr ""
##   .. ..$ classificacao   : chr "INTERNACIONAL"
##   .. ..$ paginas         : chr " - "
##   .. ..$ autores         :List of 1
##   .. ..$ autores-endogeno:List of 1
##   ..$ PERIODICO                             :'data.frame':   6 obs. of  10 variables:
##   .. ..$ natureza        : chr [1:6] "COMPLETO" "COMPLETO" "COMPLETO" "COMPLETO" ...
##   .. ..$ titulo          : chr [1:6] "On the q-tensor square of a group" "A survey of non-abelian ten
##   .. ..$ periodico       : chr [1:6] "Journal of Group Theory" "Boletim da Sociedade Paranaense de Ma
##   .. ..$ ano             : chr [1:6] "2011" "2012" "2016" "2016" ...
##   .. ..$ volume          : chr [1:6] "14" "30" "16" "107" ...
##   .. ..$ issn            : chr [1:6] "14335883" "21751188" "02194988" "0003889X" ...
##   .. ..$ paginas         : chr [1:6] "785 - 805" "77 - 89" "1750211 - " "127 - 133" ...
##   .. ..$ doi             : chr [1:6] "10.1515/JGT.2010.084" "10.5269/bspm.v30i1.13350" "10.1142/S021
##   .. ..$ autores         :List of 6
##   .. ..$ autores-endogeno:List of 6
##  $ orientacoes_academicas:List of 3
##   ..$ ORIENTACAO_CONCLUIDA_DOUTORADO   :'data.frame':    3 obs. of  13 variables:
##   .. ..$ natureza                   : chr [1:3] "Tese de doutorado" "Tese de doutorado" "Tese de dou
##   .. ..$ titulo                     : chr [1:3] "Cotas superiores para o expoente e o n\u00famero m\u
##   .. ..$ ano                        : chr [1:3] "2011" "2011" "2017"
##   .. ..$ id_lattes_aluno            : chr [1:3] "9037151037918091" "8664599889120339" "0723203301483
##   .. ..$ nome_aluno                 : chr [1:3] "Eunice C\u00e2ndida Pereira Rodrigues" "Ivonildes R
##   .. ..$ instituicao                : chr [1:3] "Universidade de Bras\u00edlia" "Universidade de Bras
##   .. ..$ curso                      : chr [1:3] "Matem\u00e1tica" "Matem\u00e1tica" "Matem\u00e1tica
##   .. ..$ codigo_do_curso            : chr [1:3] "51500035" "51500035" "51500035"
##   .. ..$ bolsa                      : chr [1:3] "SIM" "SIM" "NAO"
##   .. ..$ agencia_financiadora       : chr [1:3] "Funda\u00e7\u00e3o de Amparo \u00e0 Pesquisa do Esta
##   .. ..$ codigo_agencia_financiadora: chr [1:3] "035600000004" "002200000000" ""
##   .. ..$ nome_orientadores          :List of 3
##   .. ..$ id_lattes_orientadores     :List of 3
##   ..$ ORIENTACAO_CONCLUIDA_MESTRADO    :'data.frame':    3 obs. of  13 variables:
##   .. ..$ natureza                   : chr [1:3] "Disserta\u00e7\u00e3o de mestrado" "Disserta\u00e7\u
##   .. ..$ titulo                     : chr [1:3] "Algumas Cotas S\u00faperiores para aordem do Quadrad
##   .. ..$ ano                        : chr [1:3] "2010" "2011" "2012"
##   .. ..$ id_lattes_aluno            : chr [1:3] "5367744818899315" "" "0121355793029434"
##   .. ..$ nome_aluno                 : chr [1:3] "Bruno Cesar Rodrigues Lima" "M\u00f4nica Aparecida C
##   .. ..$ instituicao                : chr [1:3] "Universidade de Bras\u00edlia" "Universidade de Bras
##   .. ..$ curso                      : chr [1:3] "Matem\u00e1tica" "Matem\u00e1tica" "Matem\u00e1tica
##   .. ..$ codigo_do_curso            : chr [1:3] "51500035" "51500035" "51500035"
##   .. ..$ bolsa                      : chr [1:3] "NAO" "SIM" "SIM"
##   .. ..$ agencia_financiadora       : chr [1:3] "" "Coordena\u00e7\u00e3o de Aperfei\u00e7oamento de
##   .. ..$ codigo_agencia_financiadora: chr [1:3] "" "045000000000" "045000000000"
##   .. ..$ nome_orientadores          :List of 3
##   .. ..$ id_lattes_orientadores     :List of 3
```

```
##   ..$ ORIENTACAO_EM_ANDAMENTO_DOUTORADO:'data.frame':    1 obs. of  13 variables:
##   .. ..$ natureza                   : chr "Tese de doutorado"
##   .. ..$ titulo                     : chr "Quadrado Tensorial n\u00e3o Abeliano de certas classes de
##   .. ..$ ano                        : chr "2014"
##   .. ..$ id_lattes_aluno            : chr "1933036212945705"
##   .. ..$ nome_aluno                 : chr "Juliana Silva Canella"
##   .. ..$ instituicao                : chr "Universidade de Bras\u00edlia"
##   .. ..$ curso                      : chr "Matem\u00e1tica"
##   .. ..$ codigo_do_curso            : chr "51500035"
##   .. ..$ bolsa                      : chr "SIM"
##   .. ..$ agencia_financiadora       : chr "Coordena\u00e7\u00e3o de Aperfei\u00e7oamento de Pessoal
##   .. ..$ codigo_agencia_financiadora: chr "045000000000"
##   .. ..$ nome_orientadores          :List of 1
##   .. ..$ id_lattes_orientadores     :List of 1
##  $ senioridade           : chr "8"
```

Uma breve inspeção visual dos atributos anteriormente apresentados permite inferir que o pesquisador sob análise:

- Atua predominantemente na área de matemática.
- Trabalha no Instituto de Ciências Exatas da UnB.
- Possui três artigos recentes publicados, além de um aceito para publicação.
- Possui uma orientação de doutorado em andamento, iniciada em 2014.
- Foi classificado com senioridade 5.

Potencial de utilização dos dados do perfil dos docentes

Esses dados terão potencial para responder às questões de *data mining*? O que é possível gerar a partir desses dados, para o conjunto dos 1592 docentes da UnB? A fim de compreender a relevância dos dados para a avaliação da produção acadêmica nas pós-graduações brasileiras pode-se recorrer a trabalhos como os seguintes:

- Leite (2018) apresenta, em suas "Considerações básicas sobre a Avaliação do Sistema Nacional de Pós-Graduação", o conjunto dos itens que são tópicos de avaliação das pós-graduações pela CAPES, e que envolvem, entre outros:
  - Avaliação do corpo docente, com 20% a 30% de peso na avaliação total do programa, a depender do seu tipo. Analisando-se de forma mais detalhada os critérios de avaliação do corpo docente, indicados por Leite, o que é possível gerar com base nos dados disponíveis em unb.prof? Há dados que permitam identificar o perfil do docente, como proposto pela CAPES, inclusive no documento de área específica na qual atua o pesquisador? Que outros aspectos relevantes para a CAPES podem ser levantados com base nos dados dessa fonte?
  - Avaliação do corpo discente, Teses e dissertações, com 30% a 20% de peso na avaliação total do programa, a depender de seu tipo. Os dados sobre orientação permitem fazer quais tipos de avaliações do corpo discente?
  - Avaliação da produção intelectual, com 40% de peso na avaliação total. Qual a relevância dos dados em unb.prof para essa avaliação? Que outros arquivos podem melhor subsidiar essa avaliação?
- Em busca de considerar outros fatores relevantes para a avaliação da pós-graduação, não considerados no modelo da CAPES, pode-se recorrer ao trabalho de Kalpazidou Schmidt e Graversen (2018), que apresentam um conjunto de fatores persistentes que facilitam a existência de ambientes de pesquisa inovadores e dinâmicos, dentre os quais se destaca:
  - Atividade em pesquisas com elevado grau de impacto social;
  - Promoção de elevado grau de autonomia individual, tanto do ponto de vista teórico quanto metodológico;
  - Possuem um bom clima de trabalho, baseado no trabalho em times;
  - São internacioinalmente bem conhecidas etc.

Estariam esses fatores contemplados, de alguma forma, memso que parcialmente, nos dados presentes em unb.prof? Ou em qualquer outros dos arquivos? Cabe explorar.

Descrição dos dados de orientações

```
unb.adv <- fromJSON("data/UnBPosGeral/advise.json")
names(unb.adv)
```

```
## [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"
```

```
names(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO)
```

```
## [1] "2010" "2011" "2012" "2013" "2014" "2015" "2016" "2017"
```

```
length(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO$`2016`$natureza)
```

```
## [1] 606
```

```
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO$`2017`$curso), decreasing = TRUE), 10)
```

```
##
##                  Ci<U+00EA>ncias da Sa<U+00FA>de
##                                               17
##                                        Geografia
##                                               15
##                            Educa<U+00E7><U+00E3>o
##                                               14
##              Psicologia Cl<U+00ED>nica e Cultura
##                                               14
## Processos de Desenvolvimento Humano e Sa<U+00FA>de
##                                               13
##                                         Economia
##                                               12
##                                        Geotecnia
##                                               11
##                                   Biologia Animal
##                                               10
##      Ci<U+00EA>ncias da Informa<U+00E7><U+00E3>o
##                                               10
##                                         Geologia
##                                               10
```

```
head(sort(table(unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO$`2017`$curso), decreasing = TRUE), 10)
```

```
##
##                                         Economia
##                                               43
##                                          Direito
```

```
##                                                 34
##                  Ci<U+00EA>ncias da Sa<U+00FA>de
##                                                 28
##    Ci<U+00EA>ncias e Tecnologias em Sa<U+00FA>de
##                                                 26
##      Estruturas e Constru<U+00E7><U+00E3>o Civil
##                                                 25
##                Estudos de Tradu<U+00E7><U+00E3>o
##                                                 25
##                                       Literatura
##                                                 21
## Processos de Desenvolvimento Humano e Sa<U+00FA>de
##                                                 20
##                                         Geologia
##                                                 19
##                 Ci<U+00EA>ncias Mec<U+00E2>nicas
##                                                 18
```

Descrição dos dados de produção bibliográfica

```
unb.pub <- fromJSON("data/UnBPosGeral/publication.json")
names(unb.pub)
```

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

```
names(unb.pub$PERIODICO$`2012`)
```

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

```
head(sort(table(unb.pub$PERIODICO$`2017`$periodico), decreasing = TRUE), 10)
```

```
##
## REVISTA DE SA<U+00DA>DE P<U+00DA>BLICA (ONLINE)
##                                             23
##                                       PLoS One
##                                             21
##                             ESPACIOS (CARACAS)
##                                             16
##                             Scientific Reports
##                                             16
##                       Ciencia & Saude Coletiva
##                                             15
##                GENETICS AND MOLECULAR RESEARCH
##                                             15
##           CADERNOS DE PROSPEC<U+00C7><U+00C3>O
```

```
##                                           14
##        JOURNAL OF SOUTH AMERICAN EARTH SCIENCES
##                                           14
##           Journal of Molecular Modeling (Print)
##                                           13
## RBC. REVISTA BRASILEIRA DE CARTOGRAFIA (ONLINE)
##                                           13
```

```
head(sort(table(unb.pub$LIVRO$`2015`$nome_da_editora), decreasing = TRUE), 10)
```

```
##
##                                    ANPOF
##                                       23
##
##                                       13
## Novas Edi<U+00E7><U+00F5>es Acad<U+00EA>micas
##                                       12
##                     Laccademia Publishing
##                                        8
## Editora Universidade de Bras<U+00ED>lia
##                                        7
##                           Pontes Editores
##                                        7
##                               Lumen Juris
##                                        6
##                          Fino Tra<U+00E7>o
##                                        5
##                                  INEP/MEC
##                                        5
##                                     LTr
##                                        5
```

Descrição dos dados de agregação de docentes por área

```
unb.area <- fromJSON("data/UnBPosGeral/researchers_by_area.json")
unb.area.df <- cbind(names(unb.area$`Areas dos pesquisadores`),
          (sapply(unb.area$`Areas dos pesquisadores`, function(x) length(x))))
rownames(unb.area.df) <- c(1:nrow(unb.area.df)); colnames(unb.area.df) <- c("Area", "Professores")
glimpse(unb.area.df)
```

```
##  chr [1:85, 1:2] "Administra\u00e7\u00e3o" "Agronomia" ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:85] "1" "2" "3" "4" ...
##   ..$ : chr [1:2] "Area" "Professores"
```

```
head(unb.area.df[])
```

```
##   Area                        Professores
## 1 "Administra<U+00E7><U+00E3>o" "101"
## 2 "Agronomia"                 "65"
## 3 "Antropologia"              "52"
## 4 "Arqueologia"               "2"
## 5 "Arquitetura e Urbanismo"   "42"
## 6 "Artes"                     "85"
```

Descrição dos dados de redes de colaboração

CRISP-DM Fase.Atividade 2.3 - Análise exploratória dos dados

Como já informado, a análise exploratória dos dados possibilita um entendimento mais profundo da relação estatística existente entre os dados dos *datasets* para um melhor entendimento da qualidade daqueles dados para os objetivos do projeto.

Como já informado, a análise exploratória dos dados é responsabilidade parcial dos alunos, tendo em vista que este relatório apresenta uma análise exploratória preliminar. O relatório final deve conter análises exploratórias dos dados que sejam significativas e aprofundadas.

Arquivo Profile

```
# jsonedit(unb.prof)
# Número de áreas de atuação cumulativo
sum(sapply(unb.prof, function(x) nrow(x$areas_de_atuacao)))
```

```
## [1] 7119
```

```
# Número de áreas de atuação por pessoa
table(unlist(sapply(unb.prof, function(x) nrow(x$areas_de_atuacao))))
```

```
##
##   1   2   3   4   5   6  10
## 119 205 350 331 342 416   1
```

```
# Número de pessoas por grande area
table(unlist(sapply(unb.prof, function(x) (x$areas_de_atuacao$grande_area))))
```

```
##
##                                  CIENCIAS_AGRARIAS
##                         27                       427
##        CIENCIAS_BIOLOGICAS          CIENCIAS_DA_SAUDE
##                        780                       601
## CIENCIAS_EXATAS_E_DA_TERRA           CIENCIAS_HUMANAS
##                       1075                      1641
## CIENCIAS_SOCIAIS_APLICADAS                ENGENHARIAS
##                       1172                       697
## LINGUISTICA_LETRAS_E_ARTES                    OUTROS
##                        634                        65
```

```
# Número de pessoas que produziram os específicos tipos de produção
table(unlist(sapply(unb.prof, function(x) names(x$producao_bibiografica))))
```

```
##
##                          ARTIGO_ACEITO
##                                    349
##                      CAPITULO_DE_LIVRO
##                                   1346
## DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA
##                                    532
##                                 EVENTO
##                                   1504
##                                  LIVRO
##                                    867
```

```
##                              PERIODICO
##                                   1716
##                         TEXTO_EM_JORNAIS
##                                    492
```

```
# Número de publicações por tipo
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano)))
```

```
## [1] 563
```

```
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano)))
```

```
## [1] 8816
```

```
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$LIVRO$ano)))
```

```
## [1] 2932
```

```
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$PERIODICO$ano)))
```

```
## [1] 30352
```

```
sum(sapply(unb.prof, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano)))
```

```
## [1] 3042
```

```
# Número de pessoas por quantitativo de produções por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
##    0    1    2    3    4    5    6    7    9   10   11   15
## 1415  242   66   21    8    2    3    3    1    1    1    1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
## 418 258 204 139 130  99  68  64  52  43  32  36  31  33  19  14  10  11
##  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  37
##  13   9  11   6   7   5   6   4   2   7   4   1   3   4   2   3   3   1
##  38  39  40  44  47  52  56  69
##   3   1   2   1   1   1   2   1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$LIVRO$ano))))
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  14  15  16  18  19
## 897 322 189 100  67  50  29  29  19  14  12   7   8   2   2   3   2   3
##  20  21  26  28  31  32  39  49
##   1   2   1   1   1   1   1   1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$PERIODICO$ano))))
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
##  48  55  63  74  74 101  83  86  66  75  69  61  60  53  46  47  48  27
##  18  19  20  21  22  23  24  25  26  27  28  29  30  31  32  33  34  35
##  34  44  32  38  31  32  20  25  19  26  22  17  18  13  13  11  21  10
##  36  37  38  39  40  41  42  43  44  45  46  47  48  49  50  51  52  53
##  10  11  13   6   9   8   7   6   7  14   5   8   5   4   1   6   9   2
##  54  55  56  57  58  59  60  61  62  63  64  66  67  68  69  70  71  73
```

```
##   1   3   2   3   3   3   2   4   2   6   2   3   2   3   1   1   4   2
##  74  75  76  77  78  83  86  88  89  90 103 104 126 146 222 233
##   1   1   1   4   3   1   2   1   2   2   1   1   1   1   1   1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   14   15
## 1272  219   92   46   25   20    9    9   10    4    9    4    5    5    3
##   17   18   19   21   22   23   25   27   29   30   31   34   35   38   39
##    1    5    1    1    1    1    1    1    3    2    1    1    1    1    1
##   49   51   67   86  109  146  148  176  178  181
##    1    1    1    1    1    1    1    1    1    1
```

```
# Número de produções por ano
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$ARTIGO_ACEITO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   21   17   29   44   48   60   93  251
```

```
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$CAPITULO_DE_LIVRO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 1042 1052 1325  891 1135 1185 1162 1024
```

```
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$LIVRO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##  353  301  373  383  388  426  371  337
```

```
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$PERIODICO$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 3097 3193 3633 3859 3943 4154 4322 4151
```

```
table(unlist(sapply(unb.prof, function(x) (x$producao_bibiografica$TEXTO_EM_JORNAIS$ano))))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##  459  440  424  374  384  310  360  291
```

```
# Número de pessoas que realizaram diferentes tipos de orientações
length(unlist(sapply(unb.prof, function(x) names(x$orientacoes_academicas))))
```

```
## [1] 7317
```

```
# Número de pessoas por tipo de orientação
table(unlist(sapply(unb.prof, function(x) names(x$orientacoes_academicas))))
```

```
##
##               ORIENTACAO_CONCLUIDA_DOUTORADO
##                                          935
##                ORIENTACAO_CONCLUIDA_MESTRADO
##                                         1546
##           ORIENTACAO_CONCLUIDA_POS_DOUTORADO
##                                          267
```

```
##            ORIENTACAO_EM_ANDAMENTO_DOUTORADO
##                                         1061
##            ORIENTACAO_EM_ANDAMENTO_GRADUACAO
##                                          195
## ORIENTACAO_EM_ANDAMENTO_INICIACAO_CIENTIFICA
##                                          591
##             ORIENTACAO_EM_ANDAMENTO_MESTRADO
##                                         1168
##               OUTRAS_ORIENTACOES_CONCLUIDAS
##                                         1554
```

```
#Número de orientações concluidas
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano)))
```

```
## [1] 10875
```

```
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)))
```

```
## [1] 3899
```

```
sum(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$ano)
```

```
## [1] 740
```

```
# Número de pessoas por quantitativo de orientações por pessoa 0 = 1; 1 = 2...
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17
## 218 142 126 137 153 139 134 105 113  92  95  76  45  37  42  25  23   6
##  18  19  20  21  22  23  24  25  26  27  28  30  31  34  38  41  45
##  15  12   4   2   4   4   2   2   1   3   1   1   1   1   1   1   1
```

```
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10  11  12  13  14  16  17  19
## 829 203 158 119 116  84  68  55  55  19  13  14  13   4   4   2   3   1
##  20  21  22
##   1   1   2
```

```
table(unlist(sapply(unb.prof, function(x) length(x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTO
```

```
##
##    0    1    2    3    4    5    6    7    8    9   10   14   16   17   21
## 1497  109   66   28   23   10   11    7    3    2    2    3    1    1    1
```

```
# Número de orientações por ano
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_MESTRADO$ano))
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##  962 1167 1288 1541 1621 1513 1502 1281
```

```
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_DOUTORADO$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##  304  360  433  557  550  554  617  524
```

```r
table(unlist(sapply(unb.prof, function(x) (x$orientacoes_academicas$ORIENTACAO_CONCLUIDA_POS_DOUTORADO$a
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
##   75   66   95  103  134  106   98   63
```

Arquivo Publicação

```r
# Visualizar a estrutura do arquivo de Publicacao
#jsonedit(unb.pub)
#Criando um data-frame com todos os anos
unb.pub.df <- data.frame()
for (i in 1:length(unb.pub[[1]]))
  unb.pub.df <- rbind(unb.pub.df, unb.pub$PERIODICO[[i]])
glimpse(unb.pub.df)
```

```
## Observations: 24,456
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "An unusual presentation of pediatric Cushi...
## $ periodico          <chr> "Journal of Pediatric Endocrinology & Metab...
## $ ano                <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume             <chr> "23", "5", "78", "32", "13", "7", "259", "2...
## $ issn               <chr> "0334018X", "17446651", "00099163", "180611...
## $ paginas            <chr> "607 - 612", "697 - 709", "457 - 463", "1 -...
## $ doi                <chr> "", "10.1586/eem.10.47", "10.1111/j.1399-00...
## $ autores            <list> [<"AZEVEDO, M. F.;Azevedo, M;AZEVEDO, M F;...
## $ `autores-endogeno` <list> ["0017467628165816", "0017467628165816", "...
```

```r
# Limpando o data-frame de listas
unb.pub.df$autores <- gsub("\",\"|\", \"", "; ", unb.pub.df$autores)
unb.pub.df$autores <- gsub("\"|c\\(|\\)", "", unb.pub.df$autores)
unb.pub.df$`autores-endogeno` <- gsub(",", ";", unb.pub.df$`autores-endogeno`)
unb.pub.df$`autores-endogeno` <- gsub("\"|c\\(|\\)", "", unb.pub.df$`autores-endogeno`)
glimpse(unb.pub.df)
```

```
## Observations: 24,456
## Variables: 10
## $ natureza           <chr> "COMPLETO", "COMPLETO", "COMPLETO", "COMPLE...
## $ titulo             <chr> "An unusual presentation of pediatric Cushi...
## $ periodico          <chr> "Journal of Pediatric Endocrinology & Metab...
## $ ano                <chr> "2010", "2010", "2010", "2010", "2010", "20...
## $ volume             <chr> "23", "5", "78", "32", "13", "7", "259", "2...
## $ issn               <chr> "0334018X", "17446651", "00099163", "180611...
## $ paginas            <chr> "607 - 612", "697 - 709", "457 - 463", "1 -...
## $ doi                <chr> "", "10.1586/eem.10.47", "10.1111/j.1399-00...
## $ autores            <chr> "AZEVEDO, M. F.;Azevedo, M;AZEVEDO, M F;AZE...
## $ `autores-endogeno` <chr> "0017467628165816", "0017467628165816", "00...
```

Arquivo Orientação

```
#Orientação
#Visualizar a estrutura do json no painel Viewer
#jsonedit(unb.adv)
#Reunir todos os anos e orientações concluidas em um mesmo data-frame
unb.adv.tipo.df <- data.frame(); unb.adv.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_POS_DOUTORADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df); unb.adv.tipo.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_DOUTORADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df); unb.adv.tipo.df <- data.frame()
for (i in 1:length(unb.adv[[1]]))
  unb.adv.tipo.df <- rbind(unb.adv.tipo.df, unb.adv$ORIENTACAO_CONCLUIDA_MESTRADO[[i]])
unb.adv.df <- rbind(unb.adv.df, unb.adv.tipo.df)
glimpse(unb.adv.df)
```

```
## Observations: 15,109
## Variables: 13
## $ natureza                    <chr> "Supervis\u00e3o de p\u00f3s-douto...
## $ titulo                      <chr> "A Interculturalidade na sala de a...
## $ ano                         <chr> "2010", "2010", "2010", "2010", "2...
## $ id_lattes_aluno             <chr> "", "", "", "", "", "", "", "", ""...
## $ nome_aluno                  <chr> "Lucielena Mendon\u00e7a de Lima",...
## $ instituicao                 <chr> "Universidad de Bras\u00edlia", "U...
## $ curso                       <chr> "", "", "", "", "", "", "", "", ""...
## $ codigo_do_curso             <chr> "", "", "", "", "", "", "", "", ""...
## $ bolsa                       <chr> "NAO", "SIM", "SIM", "SIM", "SIM",...
## $ agencia_financiadora        <chr> "", "Funda\u00e7\u00e3o de Ci\u00e...
## $ codigo_agencia_financiadora <chr> "", "005100000992", "000700000992"...
## $ nome_orientadores           <list> ["Maria Luisa Ort\u00edz Alvarez"...
## $ id_lattes_orientadores      <list> ["0562632464695581", "05626324646...
```

```
#Transformar as colunas de listas em caracteres eliminando c("")
unb.adv.df$nome_orientadores <- gsub("\"|c\\(|\\)", "", unb.adv.df$nome_orientadores)
unb.adv.df$id_lattes_orientadores <- gsub("\"|c\\(|\\)", "", unb.adv.df$id_lattes_orientadores)
#Separar as colunas com dois orientadores
unb.adv.df <- separate(unb.adv.df, nome_orientadores, into = c("ori1", "ori2"), sep = ",")
```

```
## Warning: Too many values at 6 locations: 34, 35, 36, 2771, 5282, 5283
```

```
## Warning: Too few values at 14710 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, ...
```

```
unb.adv.df <- separate(unb.adv.df, id_lattes_orientadores, into = c("idLattes1", "idLattes2"), sep = ",
```

```
## Warning: Too many values at 6 locations: 34, 35, 36, 2771, 5282, 5283
```

```
## Warning: Too few values at 14710 locations: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10,
## 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, ...
```

```
#Numero de orientacoes por ano
table(unb.adv.df$ano)
```

```
##
## 2010 2011 2012 2013 2014 2015 2016 2017
## 1301 1556 1749 2150 2237 2117 2166 1833
```

```r
#Tabela com nome de professor e numero de orientacoes
head(sort(table(rbind(unb.adv.df$ori1, unb.adv.df$ori2)), decreasing = TRUE), 20)
```

```
##
##                         Octavio Luiz Franco
##                                          76
##       C<U+00E9>lia Maria de Almeida Soares
##                                          61
##                   Maria Fatima Grossi de Sa
##                                          59
##                      Jorge Madeira Nogueira
##                                          53
##         Concepta Margaret McManus Pimentel
##                                          51
##             Juliana de F<U+00E1>tima Sales
##                                          41
##                 Ana Maria Resende Junqueira
##                                          40
##                                Debora Diniz
##                                          39
##     Ana Suelly Arruda C<U+00E2>mara Cabral
##                                          36
##                           Gabriele Cornelli
##                                          36
##                          Helena Eri Shimizu
##                                          36
##                          Nivaldo dos Santos
##                                          36
##                   Lucio Fran<U+00E7>a Teles
##                                          35
##                       Edson Silva de Farias
##                                          34
##                Ileno Iz<U+00ED>dio da Costa
##                                          34
##                   Ricardo Bentes de Azevedo
##                                          34
## Stella Maris Bortoni de Figueiredo Ricardo
##                                          34
##                   Anderson de Rezende Rocha
##                                          33
##                    Aparecido Divino da Cruz
##                                          33
##               Andr<U+00E9> Pacheco de Assis
##                                          32
```

CRISP-DM Fase.Atividade 2.4 - Verificação da qualidade dos dados.

Como já informado, a verificação da qualidade dos dados envolve responder se os dados disponíveis estão realmente completos.

As informações disponíveis são suficientes para o trabalho proposto?

Neste projeto, a verificação da qualidade dos dados é responsabilidade dos alunos.

CRISP-DM Fase 3 - Preparação dos Dados

Como já informado, na fase de **Preparação dos Dados** os *datasets* que serão utilizados em todo o trabalho são construídos a partir dos dados brutos. Aqui os dados são "filtrados" retirando-se partes que não interessam e selecionando-se os "campos" necessários para o trabalho de mineração.

São 5 as atividades genéricas nesta fase de preparação dos dados, a seguir divididas em subseções

CRISP-DM Fase.Atividade 3.1 - Seleção dos dados.

Como já informado, a seleção dos dados envolve identificar quais dados, da nossa "montanha de dados", serão realmente utilizados.

Quais variáveis dos dados brutos serão convertidas para o *dataset*?

Não é raro cometer o erro de selecionar dados para um modelo preditivo com base em uma falsa ideia de que aqueles dados contém a resposta para o modelo que se quer construir. Surge o cuidado de se separar o sinal do ruído (Silver, Nate. The Signal and the Noise: Why so many predictions fail — but some don't. USA: The Penguin Press HC, 2012.).

CRISP-DM Fase.Atividade 3.2 - Limpeza dos dados

CRISP-DM Fase.Atividade 3.3 - Construção dos dados

Como já informado, a construção dos dados envolve a criação de novas variáveis a partir de outras presentes nos *datasets*.

```
# Funcoes

# converte as colunas de um dataframe tipo lista em tipo character
cv_tplista2tpchar <- function( df  ) {
  for( variavel in names(df)) {
    if (class(df[[variavel]]) == "list" ) {
      df[[variavel]] <- lapply(df[[variavel]] ,   function(x)   lista2texto( x  ) )
      df[[variavel]] <- as.character( df[[variavel]] )
    }
  }
  return(df)
}
###


# converte o conteudo de lista em array de characters
lista2texto <- function( lista  ) {
  if(is.null(lista)) {
    return ( NULL )
  }
  saida <- ""
  for( j in 1:length(lista)) {
    for( i in 1:length(lista[[j]]) ) {
      elemento <- lista[[j]][i]
      if( !is.null(elemento)) {
        if( i == length(lista[[j]]) & j == length(lista)  ) {
          # se for o ultimo elemento nao coloque o ponto e virgula no final
```

```
          saida <- paste0( saida , elemento  )
        } else {
          # enquanto nao for o ultimo coloque ; separando os elementos concatenados
          saida <- paste0( saida , elemento , sep = " ; ")
        }
      }
    }
  }
  return( saida )
}

# Converte producao elattes separada por anos em um unico dataframe
converte_producao2dataframe<- function( lista_producao ) {
  df_saida <- NULL

  for( ano in names(lista_producao)) {
    df_saida <- rbind(df_saida , lista_producao[[ano]])
  }

  # converte tipo lista em array de character
  df_saida <- cv_tplista2tpchar(df_saida)
  return(df_saida)


}

#concatena dois dataframes com  colunas diferentes
concatenadf <- function( df1, df2) {
  #cria colunas de df1 que faltam em df2
  for( coluna in names(df1 ) ) {
    if( !is.element(coluna, names(df2) )) {
      df2[coluna] <- NA
    }
  }

  #cria colunas de df2 que faltam em df1
  for( coluna in names(df2 ) ) {

    if( !is.element(coluna, names(df1) )) {
      df1[coluna] <- NA
    }
  }


  #faz o rbind dos dois dataframes
  df_final <- rbind(df1 , df2)
  return(df_final)

}

# Extracao dos perfis dos professores

extrai_1perfil <- function( professor ) {
```

```
  idLattes <- names(professor)
  nome <- professor[[idLattes]]$nome
  resumo_cv <- professor[[idLattes]]$resumo_cv
  endereco_profissional <- professor[[idLattes]]$endereco_profissional #list
  instituicao <- endereco_profissional$instituicao
  orgao <- endereco_profissional$orgao
  unidade <- endereco_profissional$unidade
  DDD <- endereco_profissional$DDD
  telefone <- endereco_profissional$telefone
  bairro <- endereco_profissional$bairro
  cep <- endereco_profissional$cep
  cidade <- endereco_profissional$cidade
  senioridade <- professor[[idLattes]]$senioridade
  df_1perfil <- data.frame( idLattes , nome, resumo_cv ,instituicao ,
                            orgao, unidade , DDD, telefone, bairro,cep,cidade , senioridade,
                            stringsAsFactors = FALSE)

  return(df_1perfil)
}

extrai_perfis <- function(jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_professor <- extrai_1perfil(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- rbind(df_saida , df_professor)
    } else {
      df_saida <- df_professor
    }
  }

  return(df_saida)
}

# Extracao da producao bibliografica dos professores

extrai_1producao <- function(professor) {
  idLattes <- names(professor)
  df_1producao <<- NULL
  producao_bibliografica <- professor[[idLattes]]$producao_bibiografica  #list
  for( tipo_producao in names(producao_bibliografica)) {
    df_temporario <- cv_tplista2tpchar ( producao_bibliografica[[tipo_producao]])
    df_temporario$tipo_producao <-  tipo_producao
    df_temporario$idLattes <-  idLattes
    df_1producao <- concatenadf( df_1producao , df_temporario  )
  }
  return(df_1producao)
}

extrai_producoes <- function( jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
```

```
    jsonProfessor <- jsonProfessores[i]
    df_producao <- extrai_1producao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- concatenadf(df_saida , df_producao)
    } else {
      df_saida <- df_producao
    }
  }
  df_saida <- df_saida %>% filter( !is.na(tipo_producao))
  return(df_saida)
}

# Extracao das orientacoes dos professores

extrai_1orientacao <- function(professor) {
  idLattes <- names(professor)
  df_1orientacao <- NULL
  orientacoes_academicas  <- professor[[idLattes]]$orientacoes_academicas  #list
  for( orientacao in names(orientacoes_academicas )) {
    df_temporario <- cv_tplista2tpchar ( orientacoes_academicas[[orientacao]])
    df_temporario$orientacao <-  orientacao
    df_temporario$idLattes <-  idLattes
    df_1orientacao <- concatenadf( df_1orientacao , df_temporario  )
  }
  return(df_1orientacao)
}

extrai_orientacoes <- function(jsonProfessores) {
  df_saida <- data.frame()
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_orientacao <- extrai_1orientacao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- concatenadf(df_saida , df_orientacao)
    } else {
      df_saida <- df_orientacao
    }
  }
  df_saida <- df_saida %>% filter(!is.na(idLattes))
  return(df_saida)
}

# Extracao das areas de atuacao dos professores

extrai_1area_de_atuacao <- function(professor){
  idLattes <- names(professor)
  df_1area <-  professor[[idLattes]]$areas_de_atuacao
  df_1area$idLattes <- idLattes
  return(df_1area)
}

extrai_areas_atuacao <- function(jsonProfessores){
  df_saida <- data.frame()
```

```
  for( i in 1:length(jsonProfessores)) {
    jsonProfessor <- jsonProfessores[i]
    df_area_atuacao <- extrai_1area_de_atuacao(jsonProfessor)
    if( nrow(df_saida) > 0 ) {
      df_saida <- concatenadf(df_saida , df_area_atuacao)
    } else {
      df_saida <- df_area_atuacao
    }
  }
  df_saida <- df_saida %>% filter( !is.na(idLattes))
  return(df_saida)
}
############################ Inicio

# colocar o diretorio onde está o arquivo json de perfis a serem lidos
unb.prof.json <- read_file("data/UnBPosGeral/profile.json")
unb.prof.df.capes <- read.csv("data/PesqPosCapes.csv",
                              sep = ";", header = TRUE, colClasses = "character")
unb.prof <- fromJSON(unb.prof.json)
length(unb.prof)
```

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

```
# extrai perfis dos professores
unb.prof.df.professores <- extrai_perfis(unb.prof)

# extrai producao bibliografica de todos os professores
unb.prof.df.publicacoes <- extrai_producoes(unb.prof)

#extrai orientacoes
unb.prof.df.orientacoes <- extrai_orientacoes(unb.prof)

#extrai areas de atuacao
unb.prof.df.areas.de.atuacao <- extrai_areas_atuacao(unb.prof)

#salva os daframes
save(unb.prof.df.professores, unb.prof.df.publicacoes,
     unb.prof.df.orientacoes, unb.prof.df.areas.de.atuacao, file = "dataframes.Rda")

#cria arquivo para análise
unb.prof.df <- data.frame()
unb.prof.df <- unb.prof.df.professores %>%
  select(idLattes, nome, resumo_cv, senioridade) %>%
  left_join(
    unb.prof.df.orientacoes %>%
      select(orientacao, idLattes) %>%
      filter(!grepl("EM_ANDAMENTO", orientacao)) %>%
      group_by(idLattes) %>%
      count(orientacao) %>%
      spread(key = orientacao, value = n),
    by = "idLattes") %>%
  left_join(
    unb.prof.df.publicacoes %>%
      select(tipo_producao, idLattes) %>%
```

```
      filter(!grepl("ARTIGO_ACEITO", tipo_producao)) %>%
      group_by(idLattes) %>%
      count(tipo_producao) %>%
      spread(key = tipo_producao, value = n),
    by = "idLattes") %>%
  left_join(
    unb.prof.df.areas.de.atuacao %>%
      select(area, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(area)),
    by = "idLattes") %>%
  left_join(
    unb.prof.df.capes %>%
      select(AreaPos, idLattes) %>%
      group_by(idLattes) %>%
      summarise(n_distinct(AreaPos)),
    by = "idLattes")

glimpse(unb.prof.df)
```

```
## Observations: 1,764
## Variables: 16
## $ idLattes                                <chr> "0000507838194708", "00...
## $ nome                                    <chr> "Norai Romeu Rocco", "A...
## $ resumo_cv                               <chr> "Possui gradua\u00e7\u0...
## $ senioridade                             <chr> "8", "9", "7", "8", "9"...
## $ ORIENTACAO_CONCLUIDA_DOUTORADO          <int> 3, 3, NA, NA, NA, NA, N...
## $ ORIENTACAO_CONCLUIDA_MESTRADO           <int> 3, 14, 1, 5, 2, 3, NA, ...
## $ ORIENTACAO_CONCLUIDA_POS_DOUTORADO      <int> NA, NA, NA, NA, NA, NA,...
## $ OUTRAS_ORIENTACOES_CONCLUIDAS           <int> NA, 6, 11, 7, 5, 14, 10...
## $ CAPITULO_DE_LIVRO                       <int> NA, 3, 1, 5, 1, NA, 3, ...
## $ DEMAIS_TIPOS_DE_PRODUCAO_BIBLIOGRAFICA <int> 7, 10, NA, NA, NA, NA, ...
## $ EVENTO                                  <int> 1, 8, 25, 17, 9, 1, 26,...
## $ LIVRO                                   <int> NA, 2, NA, 2, NA, NA, 1...
## $ PERIODICO                               <int> 6, 27, 3, 6, 27, 2, 14,...
## $ TEXTO_EM_JORNAIS                        <int> NA, NA, NA, 1, NA, NA, ...
## $ `n_distinct(area)`                      <int> 2, 1, 2, 2, 1, 1, 1, 4,...
## $ `n_distinct(AreaPos)`                   <int> 1, 1, 1, 2, 1, 1, 1, 2,...
```

CRISP-DM Fase.Atividade 3.4 - Integração dos dados

Como já informado, a integração dos dados envolve a união (merge) de diferentes tabelas para criar um único *dataset* para ser utilizado no R, por exemplo.

CRISP-DM Fase.Atividade 3.5 - Formatação dos dados

Como já informado, a formatação de dados envolve a realização de pequenas alterações na estrutura dos dados, como a ordem das variáveis, para permitir a execução de determinado método de data mining.

CRISP-DM Fase 4 - Modelagem

Como já informado, na fase de **Modelagem** no CRISP-DM ocorre a construção e avaliação de modelos estatísticos ou computacionais, podendo ser realizada em quatro atividades genéricas, a seguir organizadas na forma de seções

CRISP-DM Fase.Atividade 4.1 - Seleção das técnicas de modelagem

CRISP-DM Fase.Atividade 4.2 - Realização de testes de modelagem

Como já informado, na realização de testes de modelagem diferentes modelos estatísticos ou computacionais são previamente testados e avaliados. Pode-se dividir o *dataset* criado na etapa anterior para se ter uma base de treino na construção de modelos, e outra pequena parte para validar e avaliar a eficiência de cada modelo criado até se chegar ao mais "eficiente".

CRISP-DM Fase.Atividade 4.3 - Construção do modelo definitivo

Como já informado, a construçao do modelo definitivo é realizada com base na melhor experiência do passo anterior.

CRISP-DM Fase.Atividade 4.4 - Avaliação do modelo

CRISP-DM Fase 5 - Avaliação

Como já informado, na fase de **Avaliação** do CRISP-DM os resultados não são apenas avaliados, mas se verifica se existem questões relacionadas à organização que não foram suficientemente abordadas. Deve-se refletir se o uso arepetido do modelo criado pode trazer algum "efeito colateral" para a organização.

Como já informado, nesta fase, pode-se trabalhar com 3 atividades genéricas, a seguir distribuídas em seções.

CRISP-DM Fase.Atividade 5.1 - Avaliação dos resultados

CRISP-DM Fase.Atividade 5.2 - Revisão do processo

Como já informado, durante a revisão do processo verifica-se se o modelo foi construído adequadamente. As variáveis (passadas) para construir o modelo estarão disponíveis no futuro?

CRISP-DM Fase.Atividade 5.3 - Determinação dos etapas seguintes

Como já informado, pode ser necessário decidir-se por finalizar o projeto, passar à etapa de desenvolvimento, ou rever algumas fases anteriores para a melhoria do projeto.

CRISP-DM Fase 6 - Implantação (*deployment*)

Como já informado, na fase de **Implantação** (*deployment*) se realiza o planejamento de implantação dos produtos desenvolvidos (scripts, no caso do executado nesta disciplina) para o ambiente operacional, para seu

uso repetitivo, envolvendo atividades de monitoramento e manutenção do sistema (script) desenvolvido. A fase de implantação concluir com a produção e apresentação do relatório final com os resultados do projeto.

Como já informado, são seis as atividades genéricas na fase de **implantação**, a seguir apresentadas na forma de seções.

CRISP-DM Fase.Atividade 6.1 - Planejamento da transição

De que forma os produtos desenvolvidos pelo grupo poderiam ser colocados em uso prático regular, na organização cliente?

CRISP-DM Fase.Atividade 6.2 - Planejamento do monitoramento dos produtos

De que forma seria possível realizar o monitoramento do funcionamento dos produtos em utilização no ambiente operacional?

CRISP-DM Fase.Atividade 6.3 - Planejamento de manuteção

que manutenções, ajustes, mudanças, poderia ter que ser eventualmente realizadas no produto (scripts), quando em uso no ambiente operacional do cliente?

CRISP-DM Fase.Atividade 6.4 - Produção do relatório final

A entrega do relatório do grupo, tomando como base este aqui, reflete a execução desta etapa.

CRISP-DM Fase.Atividade 6.5 - Apresentação do relatório final

Como já informado, não será feita apresentação do relatório, mas esperamos que publicações científicas possam ser geradas com pelo seu grupo, com o apoio dos professores da disciplina.

CRISP-DM Fase.Atividade 6.6 - Revisão sobre a execução do projeto

Deve-se fazer aqui o registro de lições aprendidas, bem como traçadas perspectivas futuras de aprimoramento deste trabalho, da disciplina de Ciência de Dados para Todos etc.

Referências

- Azevedo, Mário Luiz Neves de, João Ferreira de Oliveira, e Afrânio Mendes Catani. "O Sistema Nacional de Pós-Graduação (SNPG) e o Plano Nacional de Educação (PNE 2014-2024): regulação, avaliação e financiamento". Revista Brasileira de Política e Administração da Educação 32, nº 3 (2016). http://dx.doi.org/10.21573/vol32n32016.68576.
- Can, Fazli, Tansel Özyer, e Faruk Polat, orgs. State of the Art Applications of Social Network Analysis. Lecture Notes in Social Networks. Switzerland: Springer International Publishing, 2014.
- CAPES. "Documentos de Área". CAPES.gov.br. Acessado 12 de junho de 2018. http://avaliacaoquadrienal.capes.gov.br/documentos-de-area.
- ———. "Plano Nacional de Pós-Graduação - PNPG 2011/2020 Vol. 1". Brasília - DF, dezembro de 2010. http://www.capes.gov.br/images/stories/download/Livros-PNPG-Volume-I-Mont.pdf.

- ———. “Plano Nacional de Pós-Graduação - PNPG 2011/2020 Vol. 2”. Brasília - DF, dezembro de 2010. http://www.capes.gov.br/images/stories/download/PNPG_Miolo_V2.pdf.
- ———. “Sucupira: coleta de dados, docentes de pós-graduação stricto sensu no Brasil 2015”. CAPES - Banco de Metadados, 16 de março de 2016. http://metadados.capes.gov.br/index.php/catalog/63.
- Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, e Rüdiger Wirth. “CRISP-DM 1.0: Step-by-Step Data Mining Guide”. USA: CRISP-DM Consortium, 2000. https://www.the-modeling-agency.com/crisp-dm.pdf.
- Datacamp. “Machine Learning with R (Skill Track)”. Datacamp, 2018. https://www.datacamp.com/tracks/machine-learning.
- Fernandes, Jorge H C, e Ricardo Barros Sampaio. “DataScienceForAll”. Zotero, 13 de junho de 2018. https://www.zotero.org/groups/2197167/datascienceforall.
- ———. “Especificação do Trabalho Final da Disciplina de Ciência de Dados para Todos 2017.2: Estudo sobre a visibilidade internacional da produção científica das pós-graduações vinculadas às áreas de conhecimento da CAPES, na Universidade de Brasília (Comunicação Interna)”. Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2017.2, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 28 de novembro de 2017. https://aprender.ead.unb.br/pluginfile.php/474549/mod_resource/content/1/Estudo%20da%20Cie%CC%82ncia.pdf.
- Fernandes, Jorge H C, Ricardo Barros Sampaio, e João Ribas de Moura. “Ciência de Dados para Todos (Data Science For All) - 2018.1 - Análise da Produção Científica e Acadêmica da Universidade de Brasília - Modelo de Relatório Final da Disciplina - Departamento de Ciência da Computação da UnB”. Disciplina 116297 - Tópicos Avançados em Computadores, turma D, do semestre 2018.1, do Departamento de Ciência da Computação do Instituto de Ciências Exatas da Universidade de Brasília, 13 de junho de 2018.
- Frickel, Scott, e Kelly Moore. The New Political Sociology of Science: Institutions, Networks, and Power. Science and technology in society. USA: The University of Wisconsin Press, 2006.
- Graduate Prospects Ltd. “Job profile: Higher education lecturer”, 2018. https://www.prospects.ac.uk/job-profiles/higher-education-lecturer.
- Kalpazidou Schmidt, Evanthia, e Ebbe Krogh Graversen. “Persistent factors facilitating excellence in research environments”. Higher Education 75, nº 2 (1º de fevereiro de 2018): 341–63. https://doi.org/10.1007/s10734-017-0142-0.
- Kilduff, Martin, e Wenpin Tsai. Social Networks and Organizations. UK: Sage Publications, 2003.
- Kolaczyk, Eric D., e Gábor Csárdi. Statistical Analysis of Network Data with R. USA: Springer, 2014.
- Kuhn, Max, Jed Wing, Steve Weston, Andre Williams, Chris Keefer, Allan Engelhardt, Tony Cooper, et al. “Package ‘Caret’ - Classification and Regression Training”, 27 de maio de 2018. https://cran.r-project.org/web/packages/caret/caret.pdf.
- Leite, Fernando César Lima. “Considerações básicas sobre a Avaliação do Sistema Nacional de Pós-Graduação”. Comunicação Pessoal (slides). Universidade de Brasília, abril de 2018. https://aprender.ead.unb.br/pluginfile.php/502250/mod_resource/content/1/Considera%C3%A7%C3%B5es%20b%C3%A1sicas%20sobre%20a%20Avalia%C3%A7%C3%A3o%20do%20Sistema%20Nacional.pdf.
- Lusher, Dean, Johan Koskinen, e Garry Robins, orgs. Exponential Random Graph Models for Social Networks: Theory, methods, and applications. Structural Analysis in the Social Sciences. USA: Cambridge University Press, 2013.
- Mariscal, Gonzalo, Óscar Marbán, e Covadonga Fernández. “A survey of data mining and knowledge discovery process models and methodologies”. The Knowledge Engineering Review 25, nº 2 (2010): 137–66. https://doi.org/10.1017/S0269888910000032.
- Nery, Guilherme, Ana Paula Bragaglia, Flávia Clemente, e Suzana Barbosa. “Nem tudo parece o que é: Entenda o que é plágio”. Instituto de Arte e Comunicação Social da UFF, 2009. http://www.noticias.uff.br/arquivos/cartilha-sobre-plagio-academico.pdf.
- Nooy, Wouter de, Andrej Mrvar, e Vladimir Batagelj. Exploratory Social Network Analysis with Pajek. Structural Analysis in the Social Sciences. USA: Routldge, 2005.
- Pátaro, Cristina Saitê de Oliveira, e Frank Antonio Mezzomo. “Sistema Nacional de Pós-Graduação no Brasil: estrutura, resultados e desafios para política de Estado - Lívio Amaral”. Revista Educação e Linguagens 2, nº 3 (julho de 2013): 11–17.

- Schwartzman, Simon. “A Ciência da Ciência”. Ciência Hoje 2, nº 11 (março de 1984): 54–59.
- Silver, Nate. The Signal and the Noise: Why so many predictions fail — but some don’t. USA: The Penguin Press HC, 2012.
- Vicari, Donatella, Akinori Okada, Giancarlo Ragozini, e Claus Wiehs. Analysis and Modeling of Complex Data in Behavioral and Social Sciences. Studies in Classifi cation, Data Analysis, and Knowledge Organization. Switzerland: Springer, 2014.
- Wickham, Hadley, e Garrett Grolemund. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. USA: O’Reilly, 2016.