# Uma abordagem de mineração descritiva aplicada a dados abertos governamentais empregando a ferramenta R

**Lydia C. C. Braga[1], Isabela Neves Drummond[1]**

[1]Instituto de Matemática e Computação – Universidade Federal de Itajubá (UNIFEI)
Caixa Postal 15.064 – 91.501-970 – Itajubá – MG – Brazil

{lydia.braga,isadrummond}@unifei.edu.br

***Abstract.*** *Open government data is a source of information on the administration and daily life of a country, so extracting this knowledge is extremely advantageous. This article presents the use of the R language as mining tool of this type of data, in order to verify its potential in this context. As an application, the data set Enem 2013, provided by the National Institute of Studies and Research (INEP), was used in order to demonstrate the advantages of this type of process in the population daily.*

***Resumo.*** *Dados abertos governamentais são fonte de informação sobre a administração e cotidiano de um país, de forma que extrair este conhecimento é extremamente vantajoso. Este artigo aborda o uso da linguagem R como ferramenta de mineração deste tipo de dados, buscando verificar seu potencial neste contexto. Como aplicação foi empregado o conjunto de dados Enem 2013, provido pelo Instituto Nacional de Estudos e Pesquisa (INEP), de forma a demonstrar as vantagens deste tipo de processo no cotidiano populacional.*

## 1. Introdução

As áreas governamentais coletam grandes quantidades de dados todos os dias, mas apenas uma pequena parte é utilizada para gerar conhecimento relevante para a população. A partir do paradigma de dados abertos governamentais, estes dados passam a estar disponíveis em formatos abertos, de forma que a sociedade tem a possibilidade de produzir cruzamentos, novas interpretações e aplicações [Dutra e Lopes 2013].

Este processo é realizado através da Mineração de dados, área caracterizada pela extração de conhecimento relevante de grandes bases de dados, descobrindo relacionamentos inesperados e resumindo-os em uma nova forma de fácil entendimento e grande utilidade para o proprietário dos dados [Hand, Mannila e Smyth 2001]. As técnicas de mineração de dados têm fundamentos embasados na estatística e no aprendizado de máquina, utilizando algoritmos complexos no processo de análise de dados. Dessa forma, ferramentas eficientes são essenciais neste tipo de trabalho.

Neste contexto, este artigo propõe um modelo de mineração de dados capaz de descrever dados abertos governamentais, utilizando a ferramenta R como instrumento de mineração. O emprego desta ferramenta por estatísticos e analistas de dados tem crescido nos últimos anos, de forma que ela é um instrumento promissor na área de mineração, tornando-se uma opção para a mineração de dados abertos governamentais.

Este artigo está organizado da seguinte maneira: a Seção 2 visa definir o processo de mineração de dados e exibir suas aplicações em dados abertos governamentais. A Seção 3 define a linguagem R, apresentando alguns pacotes utilizados na aplicação descrita neste artigo. A Seção 4 descreve a metodologia de mineração empregada, especificando as técnicas e os pacotes da linguagem R. A seção 5 apresenta o uso do modelo desenvolvido em um conjunto de dados abertos governamentais, como forma de validação. Por fim, a Seção 6 apresenta a análise geral do processo realizado, e possíveis trabalhos futuros.

## 2. Mineração de dados e Dados abertos governamentais

O processo de mineração de dados visa extrair informações úteis e compreensíveis de grandes conjuntos de dados, permitindo a aplicação deste conhecimento para gerar valor ao proprietário. Essas informações são expressas em equações, regras, agrupamentos, estruturas de árvores entre outras formas de relações de dados, gerados por meio de métodos estatísticos e de aprendizado de máquina.

Dados Abertos Governamentais consistem em dados coletados pelos poderes públicos disponibilizados à sociedade na sua forma bruta e em formato aberto, permitindo que a população produza suas próprias interpretações e aplicações [Dutra e Lopes, 2013]. Objetivando basicamente a transparência, liberação de valor social e comercial e a participação governamental, qualquer pessoa interessada tem a possibilidade de minerar e extrair conhecimento útil destes dados, tornando os cidadãos mais informados sobre as ações e decisões governamentais e dando-lhes a oportunidade de contribuir com o Governo [Open Govenment Data, 2014].

Exemplos deste tipo de aplicação podem ser encontrados em países como Estados Unidos, Canadá e Alemanha, onde são empregadas técnicas de mineração em dados oficiais, trazendo benefícios ao cotidiano populacional. Dentre estes projetos, destaca-se o *Sunlight Foundation*, uma organização americana apartidária e sem fins lucrativos que visa aumentar a transparência e a responsabilidade no Congresso Nacional. Suas iniciativas incluem a *Sunlight Labs*, uma comunidade *open source* que coleta e organiza dados públicos. No site oficial[1] dessa fundação, encontra-se uma série de ferramentas baseadas em mineração de dados que permitem o acesso a informações governamentais e a possível descoberta de novos conhecimentos. No Brasil estes dados vêm sendo utilizados em iniciativas independentes, como o Prestação de Contas da Câmara Municipal de São Paulo[2] e o projeto Alagamentos[3], ambos desenvolvidos por Maurício Maia [W3C Brasil, 2009]. Entretanto, faltam projetos capazes de atingir um grande número de pessoas em suas diversas necessidades cotidianas.

## 3. Linguagem R

A linguagem R é um conjunto integrado de ferramentas para estatística computacional e visualização gráfica, sendo uma das linguagens que mais cresce no mundo devido a seu caráter *open-source* e grande comunidade ativa [Torgo, 2010]. Seus pacotes trazem

[1] Disponível em: http://sunlightfoundation.com/

[2] Disponível em: http://cmsp.topical.com.br/

[3] Disponível em: http://alagamentos.topical.com.br/

funcionalidades específicas, que permitem a mineração de dados de forma simples, mas customizada de acordo com as necessidades do analista.

Devido à grande comunidade de desenvolvedores, diversos algoritmos de mineração de dados são desenvolvidos e publicados todos os anos. O Projeto *R and Data Mining*[4] (RDM) tem o objetivo de agrupar os diversos pacotes desenvolvidos e facilitar o compartilhamento de códigos, funções e algoritmos desenvolvidos para essa finalidade [Zhao, 2013].

As tarefas de mineração costumam possuir pacotes específicos contendo funções que auxiliam as aplicações destas técnicas nos conjuntos de dados trabalhados. Algoritmos populares como árvores de decisão e o agrupamento k-médias podem ser encontrados em diversos pacotes com diferentes implementações. Também estão disponíveis algoritmos complexos, como a rede neural de Kohonen, trazendo facilidades para o treinamento e visualização de resultados. Respectivamente, estes algoritmos podem ser encontrados, por exemplo, nos pacotes party [Horthon *et al*, 2015], NbClust [Charrad *et al*, 2015] e Kohonen [Wehrens, 2014]. A linguagem R também disponibiliza ferramentas para mineração de texto, permitindo que se extraia conhecimento oculto em documentos textuais, como atas, pautas de reuniões, documentos oficiais, entre outros, muito comuns dentro do domínio governamental. Um dos pacotes disponíveis para esse processo é o tm [Feinerer, 2015], que traz facilidades para importação de dados textuais, pré-processamento, manipulação de texto e gerenciamento de metadado.

## 4. Metodologia

A metodologia empregada neste trabalho visou à aplicação de técnicas de mineração de dados a dados abertos governamentais por meio dos pacotes e facilidades da ferramenta R. Neste contexto, definiu-se um processo capaz de verificar o relacionamento entre as variáveis presentes em um conjunto de dados, focando na extração de informações ocultas no conjunto trabalhado e embasando-se nas ferramentas disponibilizadas nesta linguagem, tornando-se um guia para aplicações reais.

O processo foi dividido em quatro fases principais: *(i)* Pré-processamento, onde os dados escolhidos como material de trabalho foram explorados e corrigidos, *(ii)* Agrupamento, onde estes dados foram relacionados em grupos por meio de técnicas de aprendizado de máquina não supervisionadas, como a rede neural de Kohonen e o algoritmo k-médias, *(iii)* Extração de Regras, com o objetivo de identificar as regras de agrupamento por meio de uma árvore de decisão e *(iv)* Mineração de Texto, para encontrar novas informações relevantes em campos textuais presentes no conjunto trabalhado. Como todo procedimento de mineração, após a análise dos resultados, este processo pode ser reiniciado várias vezes até que o objetivo seja atingido.

### 4.1. Pré-Processamento

Devido aos tamanhos tipicamente grandes e origens heterogêneas, as bases de dados atuais muitas vezes possuem dados faltantes e inconsistentes. Dados abertos governamentais estão incluídos neste problema, de forma que é necessário melhorar a

[4] Disponível em: http://www.rdatamining.com/

qualidade destes dados. A ferramenta R disponibiliza estratégias para analisar estatisticamente conjuntos de dados, possibilitando verificar problemas de completude, credibilidade e interpretação, que podem ser detectados através de funções chamadas de descritivas, com a visualização de médias, medianas, variâncias e contagens para a fase de exploração de dados [Zumel e Mount 2014].

Tipicamente, o comando *summary* é utilizado para realizar a primeira verificação dos dados, permitindo a identificação de campos vazios nos dados e em quais colunas eles ocorrem. Informações como faixa de valores em campos numéricos e frequência de valores assumidos em campos categóricos e lógicos permitem que o analista encontre incoerências. Zumel e Mount (2014) alertam para a importância desta fase no processo de mineração de dados. Não reservar um tempo para examinar os dados antes de começar a fase de modelagem pode causar retrabalho ou, até mesmo, gerar predições incorretas.

### 4.2. Agrupamento

A fase de agrupamento consiste no coração do modelo proposto, visto que implementa técnicas de mineração que verificam a relação entre as variáveis de conjuntos de dados abertos governamentais, através de técnicas de aprendizado não supervisionado.

O método de agrupamento escolhido utiliza o mapa de Kohonen. Esta rede consiste em um mapeamento organizado de cada entrada de um conjunto de dados para uma camada de neurônios, formando agrupamentos de acordo com as características presentes nestas entradas [Kohonen, 2012]. A partir dele, as instâncias são classificadas por meio do algoritmo k-médias, que resolve o problema de agrupamentos a partir de um número de grupos conhecidos. A ferramenta R possui pacotes específicos para cada uma destas técnicas, como o pacote Kohonen e o pacote NbClust, que determina o número ótimo de grupos a partir de índices de validade de agrupamento e gera partições por meio de métodos como o algoritmo k-médias.

A função básica *som()* está disponível no pacote Kohonen e gera a forma usual deste tipo de mapa. A partir de um conjunto de parâmetros, a rede é adaptada ao tipo de dado utilizado, gerando como resultado um objeto do tipo Kohonen. Os vetores de características, componentes deste objeto, são essenciais para se classificar as instâncias em grupos. Cada um destes neurônios gera um vetor de características que o define, ou seja, representa todas as instâncias relacionadas a ele.

Através destes vetores, o algoritmo k-médias é capaz de classificar cada instância do conjunto de dados em grupos. Para tanto é necessário definir o número de agrupamentos ótimo representados no mapa de Kohonen. A literatura propõe uma ampla variedade de índices que combinam informações sobre a compactação e isolamento dos agrupamentos, assim como propriedades de geometria ou estatística e medidas de dissimilaridade ou similaridade entre os dados para verificar o número ótimo de grupos, sendo extremamente úteis em processos reais de agrupamento. Para realizar essa tarefa, a ferramenta R disponibiliza o pacote NbClust que provê trinta índices que visam determinar o número de agrupamentos em um conjunto de dados. Além disso, também é gerada a partição dos grupos. O método de agrupamento escolhido permite que a função *NbClust()* agrupe os dados de entrada utilizando o número ideal de grupos definido pela maioria dos índices.

### 4.3. Extração de Regras

Extrair regras de um conjunto de dados é um tipo de aprendizado de máquina supervisionado que permite identificar as influências de cada um dos atributos que compõem o conjunto e aprofundar os resultados da mineração descritiva. A ferramenta R possui pacotes que criam esse tipo de modelo, como o party e o partykit que utilizam a inferência condicional como método de partição em subconjuntos de forma binária e recursiva. Esta metodologia é aplicável a todas as classes de problemas de regressão, incluindo os nominais, ordinais, numéricos e até mesmo variáveis de resposta multivariada e escalas de medidas arbitrárias. O funcionamento destes pacotes tem como base uma estrutura unificada que incorpora particionamento binário recursivo com a teoria de permutação desenvolvida por Strasser e Weber (1999). A base para a seleção imparcial entre as co-variáveis é a distribuição condicional de estatísticas entre as respostas [Horthon, Hornik e Zeileis 2015].

A função básica é a *ctree()*. Seus principais parâmetros são o conjunto de dados de treinamento e uma "fórmula", que consiste na especificação da variável alvo e do relacionamento das demais variáveis. Por exemplo, a fórmula *classe* ~ $variável_1$ + $variável_2$ + $variável_3$ define que o campo classe dos dados é o alvo para a árvore de decisão, ou seja, as folhas da árvore, e as variáveis 1, 2 e 3 são independentes [Zhao 2013]. Uma funcionalidade importante e disponível apenas no pacote partykit é o método *list.rules.party()*, que permite a extração de todas as regras geradas pela árvore de decisão, automatizando a análise.

### 4.4. Mineração de Texto

Os modelos aplicados a dados textuais diferem dos utilizados em dados numéricos ou categóricos, uma vez que a informação encontra-se implícita, ou seja, não existem campos que identificam atributos e simplificam as informações representadas. Em dados abertos governamentais, este tipo de técnica é essencial na análise de documentos.

O processo definido visa extrair os termos mais frequentes nos atributos textuais presentes no conjunto de dados, de forma a construir uma nuvem de palavras. Esse tipo permite a identificação de assuntos e pontos mais presentes em textos, como contratos e reportagens relacionados aos dados. Para isso é preciso: *(i)* transformar o texto para caixa baixa facilitando os processos de comparação, *(ii)* remover pontuação, número, símbolos e *stop words*, *(iii)* realizar expansão semântica, *(iv)* verificar a frequência de cada palavra e, finalmente, *(v)* gerar a nuvem.

A ferramenta R disponibiliza o pacote tm para realizar mineração de texto, permitindo especificar a língua trabalhada. As três primeiras etapas podem ser realizadas através de um único método disponibilizado no pacote, o *tm_map()*. Este método aplica uma função, definida a partir de uma função de mapeamento englobada no conceito de transformação, em todos os elementos de um corpo de texto. Dessa forma, transformar textos em caixa baixa, remover padrões, números e pontuação encontram-se neste contexto. No caso da transformação relacionada às *stop words*, o ponto chave está na utilização de um tipo de dicionário disponibilizado no pacote, que permite a definição da língua. Para realizar a expansão semântica, o R disponibiliza a transformação *stemDocument*. Esta transformação está contida no pacote SnowballC e remove o sufixo das palavras. O cálculo das frequências dos termos contidos em um

corpo de texto pode ser feito com a função *findFreqTerms()*, que é realizada sobre um objeto do tipo *Document Term Matrix*, um tipo de matriz onde as linhas representam os documentos, as colunas os termos e cada uma das células define a frequência destes termos em cada documento [Williams, 2014]. Por fim, através do pacote wordcloud, a nuvem de palavras é gerada. O analista pode definir um número máximo de palavras na visualização e também a frequência mínima de cada uma delas.

## 5. Estudo de caso: uma aplicação em Dados Abertos Governamentais

A apresentação dos resultados obtidos durante a mineração de dados tem como objetivo permitir a visão concreta dos elementos que compõem o processo definido na Seção 4, avaliando as técnicas e algoritmos encontrados na linguagem R e utilizados para minerar o conjunto de dados selecionado.

### 5.1. Dados empregados

O conjunto de dados ENEM 2013 refere-se aos resultados por Escola do Exame Nacional do Ensino Médio aplicado no ano de 2013, podendo ser encontrado no portal do Instituto Nacional de Estudos e Pesquisa[5] (INEP), juntamente com as provas aplicadas e seus gabaritos. Também são disponibilizados documentos descritivos sobre os dados. O conjunto de dados utilizado neste trabalho consiste na *Planilha ENEM por Escola*, um arquivo que divide o resultado por escola de acordo com as áreas avaliadas na prova do ENEM: Linguagens e Códigos, Redação, Matemática, Ciências Humanas e Ciências da Natureza. Este conjunto possui 14714 instâncias que representam as escolas participantes do exame. Os atributos utilizados na análise referem-se à unidade da federação, dependência administrativa, localização e porte da escola, incluindo ainda o nível social médio dos alunos, o índice da formação de adequação da formação docente e as médias referentes às provas do ENEM.

### 5.2. Resultados

Os resultados referem-se à aplicação da abordagem de mineração proposta neste artigo ao conjunto de dados ENEM 2013.

*Pré-Processamento*

Através do comando *summary*, avaliou-se o conjunto de dados trabalhado. Os campos referentes à unidade federativa, dependência administrativa, localização, porte da escola e nível social foram corretamente definidos como campos categóricos e apresentam a frequência em que cada um destes valores aparece no conjunto de dados. Os campos que representam a média das escolas em cada uma das áreas do conhecimento avaliadas estão definidos como campos numéricos e apresentam valores que tendem a 1000, de forma que estão dentro do intervalo esperado como nota da prova. Já a partir desta análise pode-se perceber que a prova de Ciências Naturais possui a menor média e a prova de redação, a maior. Por último, o campo de formação de docente assume valores dentro da faixa [0, 100], estando consistente em relação ao seu contexto. Apesar disso, é possível perceber que existem três instâncias do conjunto de dados que estão com este campo vazio, como pode ser visto através do valor NA. Para

[5] Disponível em: http://portal.inep.gov.br/

resolver essa inconsistência foi seguida a abordagem de remoção, eliminando as instâncias que possuíam campos vazios. O resultado pode ser visto na Figura 1.

```
       UF          DEP_ADM       LOCALIZACAO                 PORTE_ESCOLA          NIVEL_SOCIAL
SP     :3006   Estadual :7914   Rural :  627   De 1 a 30 alunos    :4488   Alto       :3215
MG     :1688   Federal  : 284   Urbana:14088   De 31 a 60 alunos   :3603   Baixo      : 840
RJ     :1312   Municipal: 115                  De 61 a 90 alunos   :2100   Medio      :3376
RS     :1071   Privada  :6402                  Maior que 90 alunos:4524   Medio Alto :3826
PR     : 925                                                              Medio Baixo:1916
CE     : 812                                                              Muito Alto :1471
(Other):5901                                                              Muito Baixo:  71

 FORMACAO_DOCENTE     NATURAIS         HUMANAS         LINGUAGENS       MATEMATICA        REDACAO
Min.    :  0.0   Min.   :382.4   Min.   :384.7   Min.   :365.8   Min.   :382.6   Min.   :196.7
1st Qu.: 48.6   1st Qu.:450.8   1st Qu.:491.0   1st Qu.:467.3   1st Qu.:484.3   1st Qu.:482.7
Median : 60.9   Median :474.8   Median :517.7   Median :496.8   Median :520.7   Median :528.0
Mean   : 59.4   Mean   :487.8   Mean   :529.3   Mean   :501.6   Mean   :534.7   Mean   :536.9
3rd Qu.: 71.9   3rd Qu.:518.0   3rd Qu.:564.4   3rd Qu.:535.3   3rd Qu.:577.7   3rd Qu.:586.2
Max.   :100.0   Max.   :734.0   Max.   :738.8   Max.   :658.3   Max.   :868.3   Max.   :869.0
NA's   :3
```

**Figura 1. Pré-processamento do conjunto de dados ENEM 2013.**

*Agrupamento*

A rede neural de Kohonen foi aplicada utilizando uma grade de 38x39 neurônios, com vizinhança circular e taxa de aprendizado no intervalo [0.06, 0.01], gerando o mapa de Kohonen. A Figura 2 apresenta os resultados obtidos através desta rede. À esquerda encontra-se o mapa auto-organizável gerado com os agrupamentos e, ao seu lado, são expostos os mapas de calor de cada uma das variáveis do conjunto. No mapa referente à localização, por exemplo, é possível observar que existem dois grupos, um notavelmente maior que o outro. A comprovação desta característica pode ser conferida no pré-processamento (Figura 1), que mostra que existem 627 escolas na zona rural e 14088 na zona urbana. A mesma análise pode ser feita no mapa de dependência administrativa, onde existem dois grandes grupos que representam as categorias estadual e privada, e dois grupos menores referentes aos valores federal e municipal. Também é interessante observar que valores altos para as provas estão localizados na mesma região do mapa, coincidindo também com a região onde o nível social tende a ser maior. Posteriormente, foi verificado através do pacote NbClust que o número ótimo de grupos consiste em 7, gerando as partições referentes com o algoritmo K-médias.

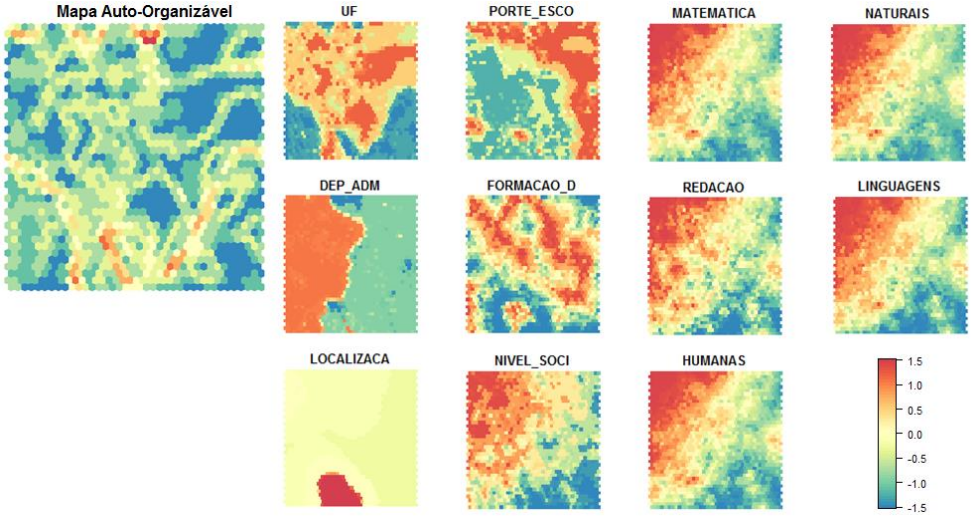


**Figura 2. Resultado da Rede de Kohonen.**

*Regras*

Para gerar as regras, foi utilizada a árvore de decisão provida pelos pacotes party e partykit. A fórmula utilizada como base do processo definiu como variável alvo as classes geradas pelo algoritmo K-médias, mantendo os demais atributos como independentes. Após o processo de poda foi gerada uma árvore com 191 folhas, revelando a dificuldade de generalização do algoritmo para este conjunto de dados. A acurácia do modelo atingiu 55,6%, evidenciando um resultado considerado razoável, mas que pode indicar que as regras geradas não representam seu comportamento em totalidade. A Figura 3 apresenta a porcentagem de regras que indicam escolas abaixo da média geral dos inscritos do ano de 2013, separando-as por área de conhecimento. Com base neste gráfico, fica evidente que 54% das regras apresentaram notas abaixo da média, sendo que 36% relacionavam-se coma prova de ciências da natureza.

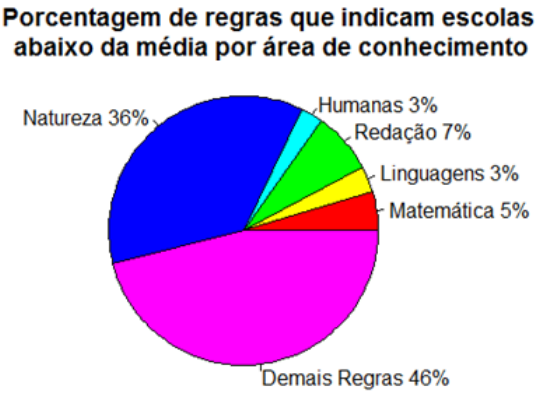


**Figura 3. Gráfico da porcentagem de Regras com escolas abaixo da média por área de conhecimento**

*Mineração de Texto*

A etapa de mineração de texto foi aplicada sobre as provas do exame do ano de 2013, trazendo resultados voltados ao conteúdo escolhido como foco das áreas de conhecimento do ENEM, como as matérias mais cobradas. Esse tipo de informação relaciona-se com as notas apresentadas nas regras obtidas, visto que específica os conteúdos onde os alunos apresentaram maiores dificuldades. A Figura 4 demonstra as nuvens geradas. Em relação à prova de matemática, pode-se inferir que seu foco foi em questões relacionadas a conversões de tempo, como indica as palavras “tempo”, “horas”, “dia” e “anos”, e geometria plana, espacial e analítica através das palavras “raio”, “centro”, “sistema”, “vértices”, entre outras. Sobre a prova de redação, vê-se com clareza o tema do ano de 2013: a lei seca. A prova de português destaca questões sobre gêneros textuais, como demonstra os termos “carta” e “crítica”, E também podem ser observados temas voltados à literatura, com base nas palavras “arte”, “poema”, “obra” e “fragmento”. Em relação a Ciências Humanas, percebe-se que as palavras características deste contexto são naturalmente mais difíceis de enquadrar em conteúdos específicos do ensino médio, podendo-se inferir com base nas palavras “brasil”, “rei”, “poder, “guerra” e “filosofia” que os conteúdos cobrados giravam em torna de história do brasil, monarquias, conflitos históricos e conteúdos filosóficos. Por fim, em relação a prova de Ciências da Natureza, vê-se a presença de termos de química como “carbono” e “moléculas”. Também percebe-se termos típicos de física mecânica e eletrodinâmica,

como “movimento”, “força”, “elétrico” e “circuito”. Os termos “dna”, “pai” e “mãe” podem ser enquadrados em questões de genética e as palavras “água”, “árvores” e “aquecimento” indicam matérias ligadas a ecologia.


**Figura 4. Nuvens de Palavra para as provas do Enem 2013.**

Através desse resultado, é possível estabelecer uma relação entre as regras geradas e as matérias cobradas na prova, concluindo-se que no ano de 2013 os inscritos na prova do Enem tiveram mais dificuldade em matérias relacionadas a química, física e biologia, evidenciando uma deficiência no ensino destas matérias.

## 6. Conclusão

Os estudos realizados indicam que a área de mineração de dados é extremamente útil na análise de dados abertos governamentais. O processo de mineração desenvolvido demonstrou ser adequado para a realização de mineração desse tipo de dados, já que dados abertos governamentais costumam se apresentar sem classificação prévia, sendo necessária a elaboração de um procedimento que, através do aprendizado não supervisionado, obtenha as partições presentes nos dados para posterior uso de algoritmos de extração de regras. A etapa de mineração de texto também mostrou potencial para descobrir novas informações sobre este tipo de dado, já que muita informação relevante encontra-se camuflada em textos. Com base na utilização das ferramentas da linguagem, percebe-se que ela facilita o processo de mineração desde a leitura do dado até a aplicação de técnicas de aprendizado de máquina complexas. A comunidade ativa em volta desta ferramenta desenvolve diversos pacotes voltados para essa tarefa, sendo que todos eles são muito bem documentados e mantidos atualizados. A própria estrutura da linguagem facilita a manipulação de dados e permite que eles sejam armazenados em objetos que simulam tabelas e facilitam a visualização das transformações realizadas. Ela também está preparada para realizar operações com matrizes e vetores de forma ágil e simplificada, facilitando o desenvolvimento de funções próprias necessárias ao processo desenvolvido. Com base nestas conclusões, trabalhos futuros podem estar relacionados com o desenvolvimento de plataformas

voltadas a dados abertos governamentais com base na linguagem R, de forma a facilitar o acesso da população.

## Referências


Charrad, M. *et al* (2015). "Package 'NbClust'". CRAN-R Project. Disponível em: <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf>. Acesso em 07 de ago. 2015.

Dutra, C. e Lopes, K. M. G. (2013) "Dados Abertos: Uma forma Inovadora de Transparência". In: VI Congresso de Gestão Pública. Brasília.

Feinerer, I. *et al* (2015). "Package 'tm'". CRAN-R Project. Disponível em: <https://cran.r-project.org/web/packages/tm/tm.pdf>. Acesso em 10 de ago. 2015.

Hand, D., Mannila, H. e Smyth, P. (2001) "Principles of Data Mining". The MIT Press. Cambridge.

Horthon, T. *et al* (2015). "Package 'party'". CRAN-R Project. Disponível em: <https://cran.r-project.org/web/packages/party/party.pdf>. Acesso em 19 de jun. 2015.

Horthon, T.; Hornik, K. e Zeileis, A. (2015) "party: A Laboratory for Recursiva Partytioning". https://cran.r-project.org/web/packages/party/vignettes/party.pdf, Setembro.

Kohonen, T. (2012) "Essentials of the self-organizing map". In: Neural Networks, v.37, p. 52-65.

Open Government Data (2014). "Open Government Data". Versão 3.0. Disponível em: <http://opengovernmentdata.org/>. Acesso em 02 de mai. 2015.

Strasser H., Weber C. (1999). "On the asymptotic theory of permutation statistics." In: Mathematical Methods of Statistics, v. 8, p. 220–250.

Torgo, L. (2010). "Data Mining with R: Learning with Case Studies". Chapman & Hall/CRC.

W3C Brasil (2009). "Manual dos dados abertos: governo". 1. Ed. Traduzido e adaptado de opendatamanual.org. 92p. Comitê Gestor da Internet no Brasil. São Paulo.

Wehrens, R. (2014). "Package 'Kohonen'". CRAN-R project. Disponível em: <https://cran.r-project.org/web/packages/kohonen/kohonen.pdf>. Acesso 04 de mai. 2015.

Williams, G. (2014) "Hands-On Data Science with R: Text Mining". http://handsondatascience.com/TextMiningO.pdf, Julho.

Zhao, Y (2013). "R and Data Mining: Examples and Case Studies". 1. Ed. Academic Press, Elsevier Inc. São Diego.

Zumel, N. e Mount, J. (2014) "Practical Data Science with R". Manning Publications Co, 1ª Edição. Shelter Island, NY.