



ANÁLISE DAS VOTAÇÕES NA CÂMARA DOS DEPUTADOS USANDO MINERAÇÃO DE DADOS

Trabalho de Conclusão de Curso

Engenharia da Computação

Edson Galileu Oliveira Melo
Orientador: Prof. Alexandre Magno Andrade Maciel



**UNIVERSIDADE
DE PERNAMBUCO**

**Universidade de Pernambuco
Escola Politécnica de Pernambuco
Graduação em Engenharia de Computação**

EDSON GALILEU OLIVEIRA MELO

**ANÁLISE DAS VOTAÇÕES NA
CÂMARA DOS DEPUTADOS USANDO
MINERAÇÃO DE DADOS**

Monografia apresentada como requisito parcial para obtenção do diploma de Bacharel em Engenharia de Computação pela Escola Politécnica de Pernambuco – Universidade de Pernambuco.

Recife, Dezembro de 2018.

MONOGRAFIA DE FINAL DE CURSO

Avaliação Final (para o presidente da banca)*

No dia 19 de dezembro de 2018, às 14:00 horas, reuniu-se para deliberar a defesa da monografia de conclusão de curso do discente **EDSON GALILEU OLIVEIRA MELO**, orientado pelo professor **Alexandre Magno Andrade Maciel**, sob título **ANALISE DAS VOTAÇÕES NA CÂMARA DOS DEPUTADOS USANDO MINERAÇÃO DE DADOS**, a banca composta pelos professores:

Sérgio Campello Oliveira

Alexandre Magno Andrade Maciel

Após a apresentação da monografia e discussão entre os membros da Banca, a mesma foi considerada:

Aprovada Aprovada com Restrições* Reprovada

e foi-lhe atribuída nota: 8,0 (OITO)

*(Obrigatório o preenchimento do campo abaixo com comentários para o autor)

O discente terá 15 dias para entrega da versão final da monografia a contar da data deste documento.


SÉRGIO CAMPELLO OLIVEIRA


ALEXANDRE MAGNO ANDRADE MACIEL

* Este documento deverá ser encadernado juntamente com a monografia em versão final.

Agradecimentos

Sou grato a Deus por prover a mim os meios e os recursos que estão além da competência humana tais como saúde e suporte familiar.

Agradeço a minha família e aos meus amigos que me apoiaram e incentivaram durante essa jornada de estudos.

E finalmente agradeço aos professores da Escola Politécnica de Pernambuco os quais com maestria compartilharam seu conhecimento comigo.

Resumo

Tão importante quanto votar é acompanhar como votam os parlamentares. Atualmente isso é possível graças ao portal de dados abertos do Congresso Nacional que disponibiliza entre outras informações as sessões de votações das proposições. Esses dados estão disponíveis em um repositório de dados que utiliza a arquitetura RESTful, que tem por finalidade atender as demandas de softwares aplicativos, portanto não são para serem consultados por usuários humanos. Neste trabalho foram analisados como estão agrupados os partidos políticos ao votarem as proposições encaminhadas pelo Poder Executivo à Câmara dos Deputados. Existem proposições encaminhadas pelos próprios deputados e postas em votação, mas não servem para esta análise em particular porque não temos como provar quais delas é uma iniciativa de apoio ou oposição ao governo. O estudo foi feito aplicando uma funcionalidade da mineração de dados chamada análise de agrupamento que resolve problemas onde é preciso identificar quantas e quais classes existem na base de dados. Os resultados obtidos mostram os partidos políticos que apoiaram, fizeram oposição e os que se revelaram divididos em relação ao governo. Pode ser verificado também como a oposição se mostrou coesa comparada à fragmentação do partido do governo.

Abstract

Just as important as voting is to follow up how the parliamentarians vote. This is currently possible thanks to the open data portal of the National Congress that provides among other information the voting sessions of the proposals. This data is available in a data repository that uses the RESTful architecture, which is intended to meet the demands of software applications, so they are not to be queried by human users. In this work, we analyzed how political parties are grouped by voting the propositions sent by the Executive Branch to the Chamber of Deputies. There are proposals put forward by the Members themselves and put to a vote, but they do not serve this particular analysis because we can not prove which of them is an initiative of support or opposition to the government. The study was done by applying a data mining functionality called cluster analysis that solves problems where you need to identify how many and which classes exist in the database. The results show the political parties that supported, opposed and those who were divided in relation to the government. It can also be verified how the opposition has proved cohesive compared to the fragmentation of the government party.

Sumário

Capítulo 1.....	11
Introdução.....	11
Capítulo 2.....	13
Fundamentação Teórica.....	13
2.1 Mineração de Dados.....	13
2.1.1 Agrupamento.....	15
2.1.2 O Algoritmo K-means.....	16
2.1.3 Métricas de escolha do k.....	17
Capítulo 3.....	20
Análise dos Dados.....	20
3.1 Limpeza e integração dos dados.....	20
3.2 Seleção e transformação dos dados.....	20
3.3 Mineração de dados.....	22
3.4 Avaliação de padrões.....	26
Capítulo 4.....	35
Conclusão e trabalhos futuros.....	35
4.1 Conclusão.....	35
4.2 Trabalhos Futuros.....	36
4.2.1 Análise comparativa entre proposições votadas nos governos Dilma e Temer.....	36
4.2.2 Análise das votações sobre a perspectiva das frentes parlamentares ou bancadas.....	36
4.2.3 Análise comparativa com outros algoritmos de agrupamentos....	37
Bibliografia.....	38

Índice de Figuras

Figura 1: Mineração de dados como um passo no processo de descoberta do conhecimento	14
Figura 2: Diferentes formas de representar clusters	16
Figura 3: Etapas do algoritmo K-means	17
Figura 4: Curva Elbow - $SSE = f(k)$	18
Figura 5: Representação gráfica do coeficiente Silhouette	19
Figura 6: Gráfico método Elbow	23
Figura 7: Definição do k pelo coeficiente Silhouette.....	25
Figura 8: Distribuição de partidos para 8 clusters referente à proposição da primeira denúncia contra Temer.....	29
Figura 9: Distribuição dos partidos para 6 clusters referente à proposição da primeira denúncia contra Temer.....	30

Índice de Tabelas

Tabela 1: Amostra dos dados estruturados.....	21
Tabela 2: Dados tratados para análise com o K-means	22
Tabela 3: Relação entre numero de cluster e a soma dos erros quadráticos	23
Tabela 4: Instâncias sorteadas para os centroides inicial com $k = 8$	26
Tabela 5: Instâncias sorteadas para os centroides inicial com $k = 6$	26
Tabela 6: Instâncias para os centroides ao final das iterações com $k = 8$	27
Tabela 7: Instâncias para os centroides ao final das iterações com $k = 6$	27
Tabela 8: Distribuição das instâncias para 8 clusters	28
Tabela 9: Distribuição das instâncias para 6 clusters	28
Tabela 10: k-means Classes to Clusters Evaluation para $k = 8$	32
Tabela 11: k-means Classes to Clusters Evaluation para $k = 6$	32
Tabela 12: Análise k-means para $k = 8$	33
Tabela 13: Análise k-means para $k = 6$	34

Tabela de Símbolos e Siglas

ARFF – *Attribute-Relation File Format*

CSV – *Comma Separated Values*

HTML-JS – *Hyper Text Markup Language JavaScript*

JSON – *JavaScript Object Notation*

KDD – *Knowledge Discovery Databases*

MP – *Medida Provisória*

PDF – *Portable Document Format*

PEC – *Proposta de Emenda Constitucional*

RESTful – *Representational State Transfer*

SSE – *Sum of Squared Errors*

URL – *Uniform Resource Locator*

XML – *Extensible Markup Language*

Capítulo 1

Introdução

Os dados abertos disponibilizados pelos governos significam um grande avanço na transparência dos atos públicos [2]. No entanto, muitos deles se encontram em um estado bruto [8], isso significa que necessitam ser processados para que sejam obtidas informações relevantes. Em alguns casos, a plataforma de dados abertos foi projetada para que esses dados sejam consumidos por *softwares* aplicativos, ficando assim indispensável o trabalho de um profissional de *software* na mineração desses dados, possibilitando então que eles comuniquem alguma informação ao usuário final. Diversos aplicativos com esse fim [1], resultantes de trabalhos voluntários, têm sido desenvolvidos e disponibilizados a população de forma gratuita.

Na Câmara dos Deputados os dados abertos estão disponíveis em um repositório de dados que utiliza a arquitetura RESTful [4], ela viabiliza o acesso aos dados formatados em arquivos paginados XML ou JSON, de onde eles podem ser acessados por meio de URLs no portal de dados abertos da Câmara dos Deputados. Normalmente os arquivos possuem muitas páginas, portanto a melhor maneira de consultar esses dados é escrevendo um programa que automatize esse trabalho. Uma sugestão do portal de dados abertos da Câmara dos Deputados é utilizar um script HTML-JS. Os arquivos que contêm as votações da Câmara dos Deputados ainda não estavam disponíveis nessa plataforma até o momento deste trabalho. Todos eles podiam ser encontrados na versão anterior do portal no formato PDF para consulta ou para serem baixados.

As sessões de votação na Câmara dos Deputados acabam por ficarem esquecidas ou sequer são percebidas pela população ao longo da legislatura. É possível que um fator colaborativo para isso esteja associado ao fato de que os resultados destas votações não estejam facilmente disponíveis, conforme já foi demonstrado no parágrafo anterior. O voto é a melhor maneira de identificar como os parlamentares e partidos se posicionam diante das proposições do governo e

também, diante da sua posição ideológica. Ter esse conhecimento ajudaria os eleitores a votarem de forma mais consciente.

O objetivo deste trabalho é reunir dez arquivos que documentam as votações das proposições encaminhadas pelo Poder Executivo à Câmara dos Deputados e empregando de técnicas de mineração de dados visualizar como estão agrupados os partidos quando votam. É importante que as proposições sejam apenas aquelas encaminhadas pelo Poder Executivo porque fica possível garantir a distinção dos partidos que apoiam dos que se opõe ao governo por meio dos votos de seus parlamentares.

Para alcançar o objetivo proposto neste trabalho, foi necessário pesquisar em matérias jornalísticas os números de cada proposição e a data de votação, porque é assim que elas se encontram indexadas no portal de dados abertos da Câmara dos Deputados. No passo seguinte foram pesquisados e obtidos, neste portal, os arquivos das votações disponíveis apenas no formato PDF. Como os dados não estão estruturados neste formato houve a conversão dos arquivos para os formatos XML e CSV, sendo esta a exigência das ferramentas empregadas na mineração de dados. Neste trabalho foi usada a técnica de mineração de dados chamada *Clustering* ou agrupamento [13], que consiste em identificar quantos e quais são os subconjuntos de dados, chamados de *clusters*, que compõe a base de dados. Estudando o perfil desses *clusters* é possível extrair conhecimento da base de dados.

Foi obtido como resultados uma classificação dos partidos quanto a um posicionamento de apoio e oposição ao governo e até mesmo a identificação de partidos que tiveram seus membros divididos. Também foi observado o nível de engajamento dos membros do partido do governo em relação ao partido que expressou maior representação na oposição.

Este trabalho está estruturado da seguinte forma: o Capítulo 2 descreve todo embasamento teórico necessário à aplicação das metodologias utilizadas. No Capítulo 3 está registrado todo o desenvolvimento deste trabalho. Finalmente no Capítulo 4 estão a conclusão e sugestões de trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Atualmente a indústria da informação e a sociedade de um modo geral tem voltado sua atenção para o tema chamado mineração de dados. Isso ocorre em um momento em que enormes quantidades de dados estão disponíveis e aliadas a iminente necessidade de transformar esses dados em informações e conhecimento úteis. Neste capítulo são abordados os conceitos fundamentais da mineração de dados, os processos envolvidos para obtenção de resultados confiáveis e as ferramentas utilizadas [7].

2.1 Mineração de Dados

Conceituar mineração de dados não é uma tarefa fácil visto que não há um consenso entre alguns que entendem como sendo sinônimo de outro termo popularizado como *Knowledge Discovery from Data* (KDD), e outros que percebem como sendo uma etapa importante no processo de descoberta do conhecimento [5]. Esse processo é mostrado na Figura 1, que obedece a sequência iterativa de passos a seguir:

1. Limpeza dos dados – remove ruídos e inconsistências de dados.
2. Integração dos dados – uma combinação de várias fontes de dados.
3. Seleção de dados – os dados relevantes para a tarefa de análise são recuperados da base de dados.
4. Transformação dos dados – os dados são transformados e consolidados em formulários apropriados para mineração executando operações de resumo ou agregação.
5. Mineração de dados – um processo essencial onde métodos inteligentes são aplicados para extrair padrões de dados.

6. Avaliação de padrões – identificar os padrões verdadeiramente interessantes que representam o conhecimento com base em medidas de interesse.
7. Apresentação de conhecimento – técnicas de visualização e representação de conhecimento são usadas para apresentar conhecimento minerado aos usuários.

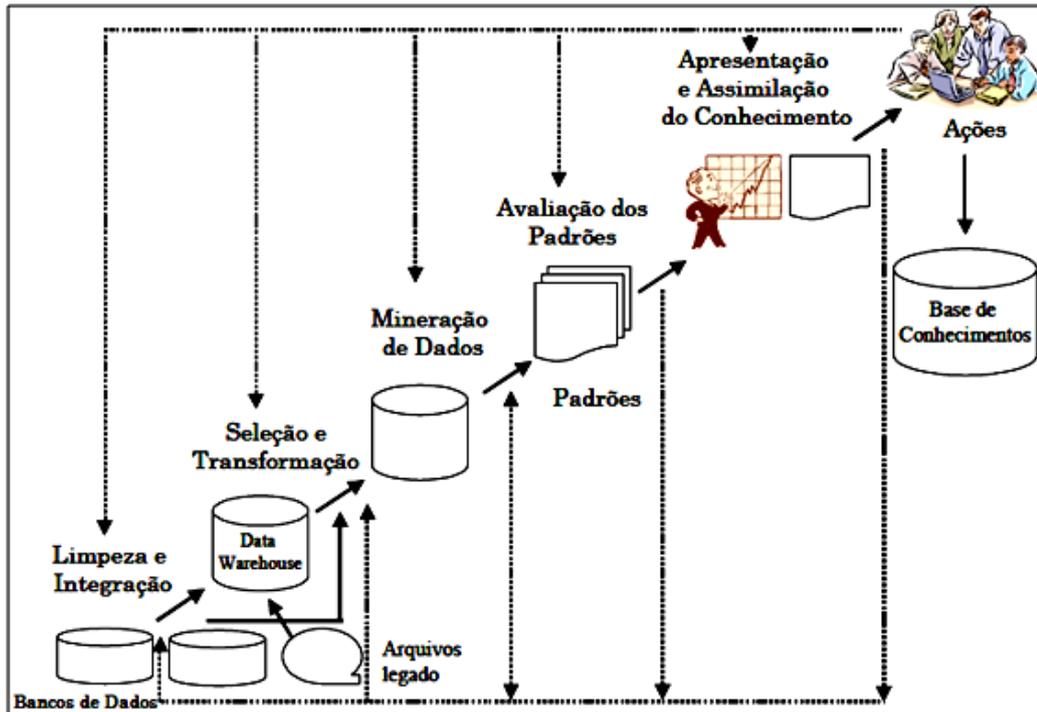


Figura 1: Mineração de dados como um passo no processo de descoberta do conhecimento

Fonte: ftp://ftp.inf.puc-rio.br/pub/docs/techreports/02_10_cortes.pdf acessado em 03-11-2018

Mineração de dados é uma área interdisciplinar que reúne um conjunto de disciplinas tais como sistemas de banco de dados, estatística, aprendizado de máquina, ciência da informação e reconhecimento de padrões. Essas disciplinas causaram forte influência no desenvolvimento de métodos de mineração de dados.

As tarefas de mineração de dados são geralmente classificadas como descritiva e preditiva. Tarefas de mineração descritivas caracterizam as propriedades gerais dos dados no banco de dados. Devem descrever padrões e tendências reveladas pelos dados. Tarefas de mineração preditivas realizam inferência sobre os dados atuais para fazer previsões. Baseados no tipo de tarefa a ser executada é que se define qual será a funcionalidade utilizada na mineração de

dados. As funcionalidades de mineração de dados incluem caracterização e discriminação, mineração de padrões frequentes, associações e correlações, classificação e regressão, análise de agrupamento e análise de *outlier* [7]. Aqui será usada a funcionalidade *Cluster Analysis* ou análise de agrupamento tendo em vista a tarefa ser descritiva e será detalhada na seção 2.1.1.

2.1.1 Agrupamento

A análise de agrupamentos é conhecida por *clustering* e está relacionada com uma técnica de mineração de dados em que se espera obter como resultado grupos de dados chamados de *clusters*. A representação da saída mostra como as instâncias de dados são associadas a cada *cluster* de modo exclusivo conforme Figura 2a. Pode acontecer de uma instância pertencer a mais de um *cluster* como na Figura 2b. Outra maneira é expressa de forma probabilística, assim cada instância está associada a um *cluster* por um grau de associação ou probabilidade de acordo com a Figura 2c. Existe ainda a estrutura hierárquica de *cluster* que lembra uma estrutura em árvore contendo poucos *clusters* no nível superior e estes se dividem em seus próprios *subclusters* no nível abaixo, e assim por diante, mostrado na Figura 2d [13].

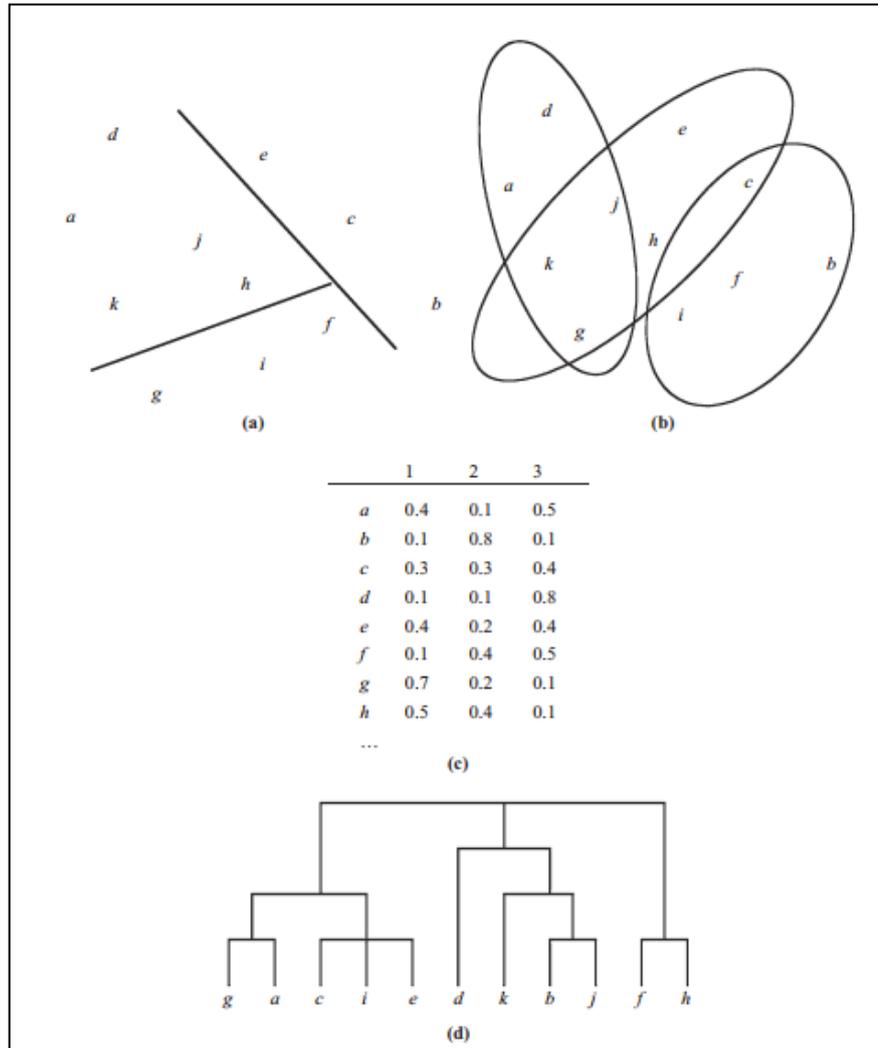


Figura 2: Diferentes formas de representar clusters

Fonte: obtida de [13]

2.1.2 O Algoritmo K-means

Uma técnica clássica usada para resolver problemas de agrupamento é o K-means. O primeiro passo a dar antes de executar este algoritmo é especificar o parâmetro *k*, que significa escolher a quantidade de *clusters* que representarão as classes existentes na base de dados. Após esta definição *k* pontos são escolhidos aleatoriamente na base dados como centro de *clusters*. Obedecendo a métrica de distância euclidiana comum cada instância é atribuída ao centro de *cluster* mais próximo. Em seguida, o centroide, ou a média, das instâncias em cada *cluster* é calculado. Assim são obtidos os centroides que representam os novos valores centrais de cada *cluster*. Esse processo é iterativo até que os mesmos pontos sejam atribuídos a cada *cluster* em rodadas consecutivas, alcançando o estágio em que os

centros do *cluster* se estabilizaram e permanecerão os mesmos para sempre conforme Figura 3.

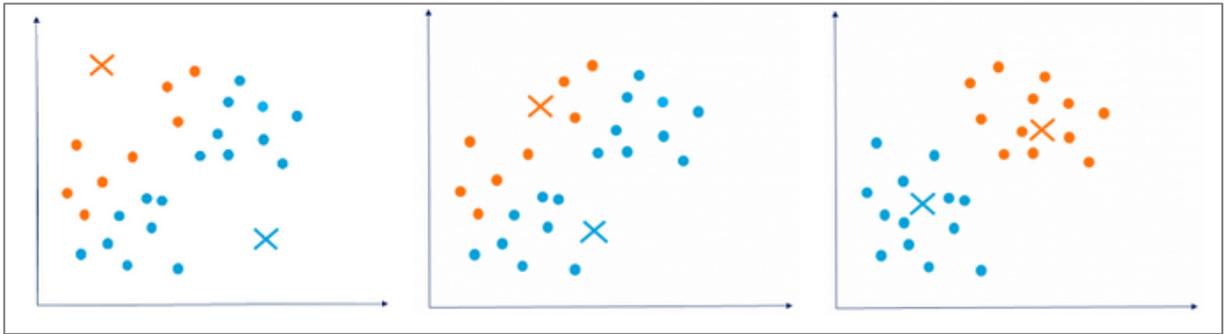


Figura 3: Etapas do algoritmo K-means

Fonte: obtido de [10]

A escolha do centro do *cluster* como o centroide minimiza a distância total ao quadrado de cada um dos pontos do *cluster* até o centro. Depois que a iteração se estabilizar, cada ponto é atribuído ao centro do *cluster* mais próximo, de modo que o efeito geral é minimizar a distância total do quadrado de todos os pontos até os centros do *cluster*. Os *clusters* finais são bastante sensíveis aos centros iniciais do *cluster*. Arranjos completamente diferentes podem surgir de pequenas mudanças na escolha aleatória inicial. De fato, isso se aplica a todas as técnicas práticas de agrupamento: é quase sempre inviável encontrar *clusters* globalmente ótimos. Uma maneira de se aumentar as chances de encontrar um mínimo global é rodar o algoritmo várias vezes com diferentes escolhas iniciais e escolher o melhor resultado final, aquele com a menor distância quadrada [13].

2.1.3 Métricas de escolha do k

Escolher o valor de k não é uma tarefa fácil quando não se sabe quantos grupos ou *clusters* existem na base de dados. Uma maneira de fazer isso é pelo método Elbow [7]. Esse método calcula a soma dos erros quadráticos SSE (*Sum of Squared Error*) em relação ao valor de k. O SSE é a soma da distância ao quadrado entre cada membro do cluster e seu centroide. Inicialmente o algoritmo K-means deve ser executado para k = 1 que gera um relatório com o valor do SSE, esses valores devem ser anotados em uma tabela. Executa novamente o K-means para k = 2 e anota na tabela junto com o SSE obtido do relatório gerado. Esse processo é repetido incrementando o número de k até que a variação do valor do SSE não seja

mais significativa. Na Figura 4 o gráfico, resultado dos dados coletados na tabela, mostra a formação de uma curva acentuada que lembra um cotovelo, daí o nome Elbow [10].

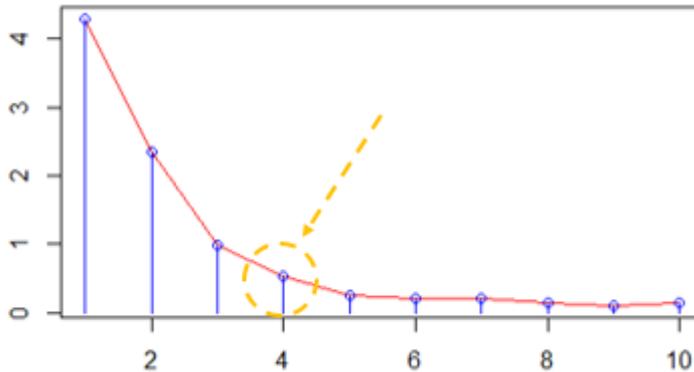


Figura 4: Curva Elbow - SSE = f(k)

Fonte: <http://www.diegonogare.net/2015/08/entendendo-como-funciona-o-algoritmo-de-cluster-k-means/> acessado em 03-11-2018

Em alguns casos a curva é suave e não é possível ter uma definição clara do valor de k , então uma alternativa é usar o método Silhouette [7]. Ele apresenta como resultado um coeficiente obtido por meio do cálculo da proximidade de cada instância de um *cluster* com as instâncias dos *clusters* vizinhos. Esse coeficiente está contido no intervalo $[-1, 1]$. Quanto mais próximo de 1, mais distante a instância estará do *cluster* vizinho e mais próximo do *cluster* a que foi designado. O oposto ocorre caso o valor se aproxime de -1. A ocorrência de 0 indica uma situação em que a instância está no limite da distância entre dois *clusters*. Portanto, a melhor configuração para um *cluster* é quando o índice se aproxima de 1. A Figura 5 mostra o gráfico como resultado da aplicação desse método em uma base de dados sendo indicando três *clusters* [7].

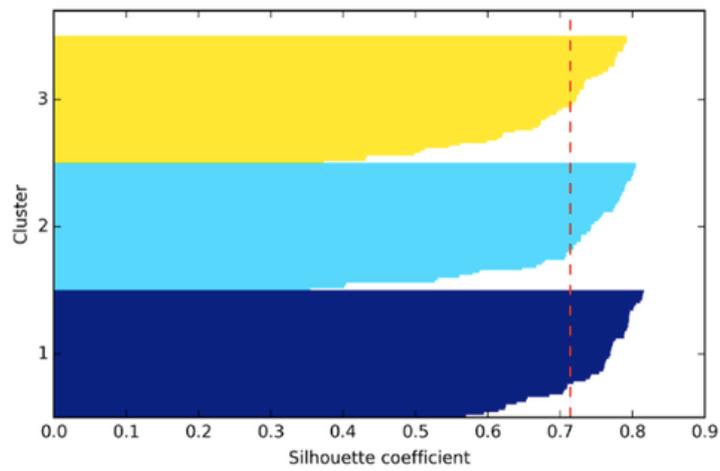


Figura 5: Representação gráfica do coeficiente Silhouette

Fonte: <https://sebastianraschka.com/faq/docs/issues-with-clustering.html> acessado em 06-11-2018

Capítulo 3

Análise dos Dados

Neste capítulo é apresentada a metodologia utilizada no desenvolvimento deste trabalho. Na sequência os passos do KDD, já detalhado no Capítulo 2, foram percorridos até a obtenção dos resultados, expresso pelo conhecimento extraído da base de dados em estudo.

3.1 Limpeza e integração dos dados

Os arquivos do repositório de dados abertos da Câmara dos Deputados referentes às votações da 55ª legislatura estão no formato PDF (*Portable Document Format*). Foram obtidos dez arquivos que contêm dez proposições encaminhadas pelo poder executivo chefiado pelo Excelentíssimo Senhor Presidente da República Michel Temer para votação no período de 29/08/2016 a 19/02/2018. Foi necessário estruturar os dados em uma planilha eletrônica e depois converter em um formato CSV (*Comma-Separated Values*) para iniciar o tratamento dos dados, que pode ser feito em um banco de dados, no Weka, que é uma ferramenta para mineração de dados, ou na própria planilha eletrônica, que foi a escolha para este projeto [7].

Existem várias maneiras de se converter um arquivo PDF para CSV. Após vários testes com aplicativos e sites que se propõe a fazer isso de forma gratuita, a melhor escolha foi converter de PDF para XLS e depois salvar como um arquivo CSV. Os testes com a conversão disponibilizada em <https://www.pdfexcel.com/pt> geraram resultados satisfatórios.

3.2 Seleção e transformação dos dados

O próximo passo foi estruturar os dez arquivos que foram convertidos para XLS em um só arquivo de modo que as proposições, que são cada um dos dez arquivos, sejam os atributos (colunas) e as votações sejam as instâncias (linhas) nesse novo arquivo.

A câmara dos deputados é composta por 513 deputados federais. Durante o período da atual 55ª legislatura, suplentes desses parlamentares ocuparam a cadeira de seus titulares por motivos diversos. Devido a isso foi registrada nesse período de votação a participação de 569 deputados. Diante dessa informação os dados devem conter 10 atributos que são as proposições votadas e 569 instâncias representando as votações dos parlamentares. Foram incluídos mais dois atributos sendo eles, deputados e partidos. Esses atributos são conhecidos por atributos de classe e terão grande importância na análise dos dados. Também é importante destacar a exclusão da instância referente ao presidente da Câmara dos Deputados, já que em todas as votações ele não votou em conformidade com o Art. 17 do regimento interno [9], ficando assim um total de 568 instâncias.

As proposições escolhidas foram: denúncia contra Temer 1, denúncia contra Temer 2, intervenção militar no Rio, lei da terceirização, licitações e contratos, MP do agronegócio, organização da presidência da república e ministérios, PEC dos gastos públicos, reforma do ensino médio e reforma trabalhista.

Uma amostra de como ficaram estruturados os dados depois de aplicadas as transformações descritas pode ser visto na Tabela 1. Feito isso resta apenas converter esse novo arquivo para o formato CSV. Agora fica possível utilizar a ferramenta para mineração de dados Weka.

Cada atributo pode ter em suas instâncias as opções SIM, NÃO, AUSENTE, ABSTENÇÃO E OBSTRUÇÃO. A instância referente ao presidente da Câmara dos Deputados, todas com valor Art. 17 foi eliminada, totalizando agora 568 instâncias.

Tabela 1: Amostra dos dados estruturados

Den_Tem	Den_Tem	Interv. M	Lei da Te	Licitaçãc	MP Agror	Org. pres	PEC dos g	Ref. Ensii	Ref. Trab	Parlame	Partido
Sim	Não	Sim	Sim	Sim	Ausente	Sim	Sim	Sim	Sim	Abel Mes	DEM
Ausente	Ausente	Sim	Sim	Sim	Ausente	Sim	Ausente	Sim	Sim	Adail Car	PP
Sim	Ausente	Ausente	Ausente	Sim	Ausente	Sim	Sim	Sim	Sim	Adalbert	PTB
Não	Não	Não	Não	Não	Obstruçã	Não	Não	Não	Não	Adelmo C	PT
Não	Não	Sim	Não	Sim	Sim	Ausente	Sim	Ausente	Não	Adelson	PR
Sim	Sim	Ausente	Não	Não	Sim	Sim	Sim	Sim	Não	Ademir C	PODE
Ausente	Ausente	Ausente	Sim	Sim	Ausente	Ausente	Ausente	Ausente	Sim	Adérmis	PSDB
Sim	Sim	Ausente	Sim	Sim	Sim	Sim	Sim	Sim	Ausente	Adilton S	PSB
Sim	Sim	Sim	Ausente	Sim	Sim	Ausente	Sim	Ausente	Sim	Aelton F	PR
Não	Não	Não	Não	Não	Obstruçã	Não	Não	Não	Não	Afonso I	PT
:	:	:	:	:	:	:	:	:	:	:	:

Fonte: Autoria própria

O que se quer obter como resultado é de que forma estão agrupados os partidos políticos que apoiam e os que se opõe ao governo expressado pelas votações de seus deputados, portanto isso envolve um problema de Agrupamento ou *Clustering*.

3.3 Mineração de dados

Para análise dos dados foi usado o K-means da ferramenta Weka, desenvolvida pela universidade de Waikato que fica em Hamilton Nova Zelândia. O Weka reconhece arquivos no formato CSV ou em um formato próprio chamado ARFF (*Attribute-Relation File Format*). A escolha aqui foi importar para o Weka o arquivo CSV e depois salvar como formato ARFF, porque os resultados da análise ficam padronizados facilitando a comparação dos dados.

Antes de executar o algoritmo foram retirados os atributos de classificação “parlamentar” e “partido”, já que o K-means é um algoritmo não supervisionado, resultando então na estrutura da Tabela 2. Em seguida foi definido o valor de k, que é a quantidade de *clusters*.

Tabela 2: Dados tratados para análise com o K-means

Den_Ten	Den_Ten	Interv. M	Lei da Te	Licitaçãc	MP Agror	Org. pres	PEC dos g	Ref. Ensi	Ref. Trabi
Sim	Não	Sim	Sim	Sim	Ausente	Sim	Sim	Sim	Sim
Ausente	Ausente	Sim	Sim	Sim	Ausente	Sim	Ausente	Sim	Sim
Sim	Ausente	Ausente	Ausente	Sim	Ausente	Sim	Sim	Sim	Sim
Não	Não	Não	Não	Não	Obstruçã	Não	Não	Não	Não
Não	Não	Sim	Não	Sim	Sim	Ausente	Sim	Ausente	Não
Sim	Sim	Ausente	Não	Não	Sim	Sim	Sim	Sim	Não
Ausente	Ausente	Ausente	Sim	Sim	Ausente	Ausente	Ausente	Ausente	Sim
Sim	Sim	Ausente	Sim	Sim	Sim	Sim	Sim	Sim	Ausente
Sim	Sim	Sim	Ausente	Sim	Sim	Ausente	Sim	Ausente	Sim
Não	Não	Não	Não	Não	Obstruçã	Não	Não	Não	Não
:	:	:	:	:	:	:	:	:	:

Fonte: Autoria própria

Definir o valor ideal para K não é uma tarefa fácil. Uma forma de fazer isso é pelo método Elbow que relaciona número de *clusters* com a soma dos erros quadráticos. O objetivo é identificar a partir de qual número de *clusters* o decremento da soma dos erros quadráticos passa a ser pouco significativo. Para fazer isso basta executar no Weka o algoritmo K-means. Aqui foi definido inicialmente k = 1 e incrementando o número de *clusters* em uma unidade para cada execução até k =

20. Em seguida é possível estruturar a Tabela 3 que relaciona cada *cluster* com a soma dos erros quadráticos, obtidos dos relatórios nas execuções do algoritmo K-means. A partir dessa tabela é possível desenhar o gráfico da Figura 6.

Analisando os dados obtidos é perceptível que a partir de oito clusters não há uma variação significativa da soma dos erros quadráticos sendo, portanto $k = 8$ uma boa escolha.

Tabela 3: Relação entre numero de cluster e a soma dos erros quadráticos

CLUSTERS	SSE	DECREMENTO
1	2797	
2	2323	474
3	1483	840
4	1450	33
5	1262	188
6	1247	15
7	1169	78
8	1085	84
9	1074	11
10	1052	22
11	1021	31
12	931	90
13	903	28
14	884	19
15	881	3
16	871	10
17	837	34
18	815	22
19	796	19
20	788	8

Fonte: Autoria própria

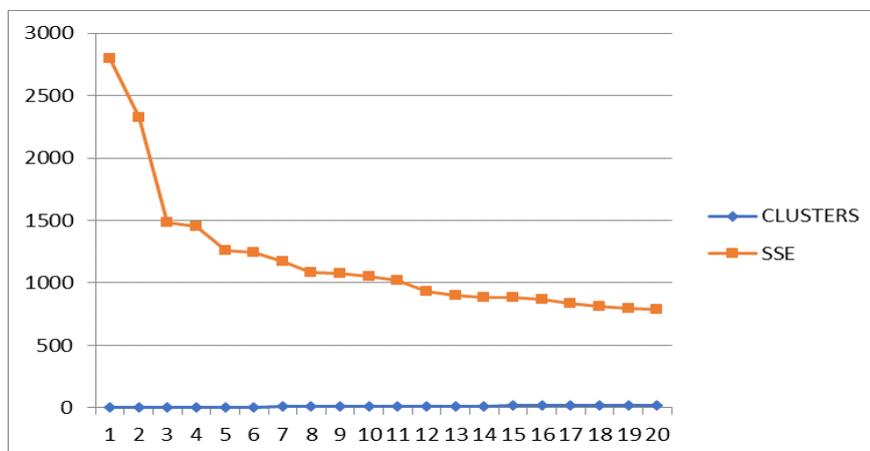


Figura 6: Gráfico método Elbow

Fonte: Autoria própria

É possível comparar escolha do k pelo método Elbow usando o coeficiente *Silhouette*. Existe uma ferramenta chamada Orange [11] que faz isso, bastando que sejam definidos os limites inferiores e superiores para o valor de k . Então se obtém o gráfico da Figura 7a que mostra a totalidade de agrupamentos ou clusters recomendados e o quanto eles estão coesos. A Figura 7b é um recorte desse gráfico ampliado em sua base para que fique visível o valor para $k = 6$ indicado por “C6” e a escala dos coeficientes *Silhouette* no eixo das ordenadas, significando uma maior coesão das instâncias em cada *cluster* à medida que se aproxima de 0,5.



Figura 7: Definição do k pelo coeficiente Silhouette

Fonte: Autoria própria

3.4 Avaliação de padrões

A partir do momento em que foram encontrados dois valores possíveis para k , o próximo passo foi identificar os padrões que representam o conhecimento extraído dessa base de dados. Foram gerados relatórios em paralelo para $k = 8$ e $k = 6$, permitindo assim também avaliar qual é a melhor opção para o valor de k .

Na Tabela 4, originada com $k = 8$, e na Tabela 5, gerada com $k = 6$, estão as definições iniciais das instâncias sorteadas para representarem os centroides de cada *cluster* antes de iniciar as iterações do algoritmo K-means. Essas tabelas podem ser comparadas com as geradas ao final das iterações verificando o quanto cada centroide mudou de posição dentro do agrupamento ou *cluster*. Com base nestas informações não é possível escolher entre os dois valores para k .

Tabela 4: Instâncias sorteadas para os centroides inicial com $k = 8$

```
kMeans
=====
Number of iterations: 5
within cluster sum of squared errors: 1085.0

Initial starting points (random):

Cluster 0: Não,Não,Não,Não,Não,Sim,Não,Não,Não,Não
Cluster 1: Não,Ausente,Sim,Ausente,Não,Ausente,Ausente,Ausente,Ausente,Não
Cluster 2: Sim,Sim,Ausente,Sim,Sim,Ausente,Sim,Sim,Sim,Sim
Cluster 3: Sim,Sim,Ausente,Ausente,Ausente,Ausente,Ausente,Sim,Ausente,Sim
Cluster 4: Sim,Sim,Sim,Sim,Sim,Sim,Sim,Ausente,Sim,Sim
Cluster 5: Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim
Cluster 6: Não,Não,Sim,Sim,Ausente,Sim,Sim,Sim,Sim,Sim
Cluster 7: Ausente,Ausente,Ausente,Ausente,Ausente,Ausente,Ausente,Sim,Ausente,Ausente
```

Fonte: Autoria própria

Tabela 5: Instâncias sorteadas para os centroides inicial com $k = 6$

```
kMeans
=====
Number of iterations: 7
within cluster sum of squared errors: 1247.0

Initial starting points (random):

Cluster 0: Não,Não,Não,Não,Não,Sim,Não,Não,Não,Não
Cluster 1: Não,Ausente,Sim,Ausente,Não,Ausente,Ausente,Ausente,Não
Cluster 2: Sim,Sim,Ausente,Sim,Sim,Ausente,Sim,Sim,Sim,Sim
Cluster 3: Sim,Sim,Ausente,Ausente,Ausente,Ausente,Ausente,Sim,Ausente,Sim
Cluster 4: Sim,Sim,Sim,Sim,Sim,Sim,Sim,Ausente,Sim,Sim
Cluster 5: Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim,Sim
```

Fonte: Autoria própria

Na sequência a Tabela 6 com $k = 8$ e a Tabela 7 para $k = 6$ mostram o resultado ao final das iterações, revelando quais instâncias passaram a conter os

centroides de seus respectivos *clusters*. Elas apresentam uma tendência dos valores contidos em cada cluster. Comparando as duas tabelas podem ser destacados na Tabela 6 o *cluster 1* como a representação de alguns partidos divididos entre governo e oposição e ainda com cinquenta por cento de ausentes e o *cluster 7* que na Tabela 7 é igual ao *cluster 1*, indicando um grupo considerado de parlamentares ausentes nas votações. Ainda não dá para concluir qual dos dois valores de k é o melhor.

Tabela 6: Instâncias para os centroides ao final das iterações com k = 8

Final cluster centroids:										
Attribute	Full Data (568.0)	Cluster# 0 (120.0)	1 (41.0)	2 (87.0)	3 (89.0)	4 (13.0)	5 (93.0)	6 (63.0)	7 (62.0)	
Den_Temer_1	Sim	Não	Não	Sim	Sim	Sim	Sim	Não	Ausente	
Den_Temer_2	Sim	Não	Não	Sim	Sim	Sim	Sim	Não	Ausente	
Interv. Mil. Rio	Sim	Não	Sim	Ausente	Sim	Sim	Sim	Sim	Ausente	
Lei da Terceir.	Sim	Não	Ausente	Sim	Sim	Sim	Sim	Sim	Ausente	
Licitacão e contratos	Sim	Não	Sim	Sim	Sim	Sim	Sim	Sim	Ausente	
MP Agroneg.	Sim	Sim	Ausente	Ausente	Sim	Sim	Sim	Sim	Ausente	
Org. presid. Rep. E ministérios	Sim	Não	Ausente	Sim	Ausente	Sim	Sim	Sim	Ausente	
PEC dos gastos Públicos	Sim	Não	Ausente	Sim	Sim	Ausente	Sim	Sim	Ausente	
Ref. Ensino Médio	Sim	Não	Ausente	Sim	Ausente	Sim	Sim	Sim	Ausente	
Ref. Trabalhista	Sim	Não	Não	Sim	Sim	Sim	Sim	Sim	Ausente	

Fonte: Autoria própria

Tabela 7: Instâncias para os centroides ao final das iterações com k = 6

Final cluster centroids:								
Attribute	Full Data (568.0)	Cluster# 0 (138.0)	1 (81.0)	2 (97.0)	3 (103.0)	4 (16.0)	5 (133.0)	
Den_Temer_1	Sim	Não	Ausente	Sim	Sim	Sim	Sim	
Den_Temer_2	Sim	Não	Ausente	Sim	Sim	Sim	Sim	
Interv. Mil. Rio	Sim	Não	Ausente	Ausente	Sim	Sim	Sim	
Lei da Terceir.	Sim	Não	Ausente	Sim	Sim	Sim	Sim	
Licitacão e contratos	Sim	Não	Ausente	Sim	Sim	Sim	Sim	
MP Agroneg.	Sim	Sim	Ausente	Ausente	Sim	Sim	Sim	
Org. presid. Rep. E ministérios	Sim	Não	Ausente	Sim	Ausente	Sim	Sim	
PEC dos gastos Públicos	Sim	Não	Ausente	Sim	Sim	Ausente	Sim	
Ref. Ensino Médio	Sim	Não	Ausente	Sim	Ausente	Sim	Sim	
Ref. Trabalhista	Sim	Não	Ausente	Sim	Sim	Sim	Sim	

Fonte: Autoria própria

A distribuição das instâncias por *clusters* pode ser observada na Tabela 8, que usa k = 8, e na Tabela 9, produzida com k = 6. Esses dados mostram o quanto é significativo ou não um *cluster* em relação à base de dados. É notável o nível de coesão da oposição que com 120 instancias possui o maior *cluster*. Para este trabalho o que se deseja observar são as representações com uma maioria de votos SIM ou votos NÃO, logo o único cluster que não vai ajudar na análise é o *cluster 1* da Tabela 7. Ainda assim não é visível qual dos dois valores de k é mais adequado.

Tabela 8: Distribuição das instâncias para 8 clusters

```

=== Model and evaluation on training set ===
Clustered Instances
0      120 ( 21%)
1       41 (  7%)
2       87 ( 15%)
3       89 ( 16%)
4       13 (  2%)
5       93 ( 16%)
6       63 ( 11%)
7       62 ( 11%)
    
```

Fonte: Autoria própria

Tabela 9: Distribuição das instâncias para 6 clusters

```

=== Model and evaluation on training set ===
Clustered Instances
0      138 ( 24%)
1       81 ( 14%)
2       97 ( 17%)
3      103 ( 18%)
4       16 (  3%)
5      133 ( 23%)
    
```

Fonte: Autoria própria

Outra forma de visualização desta análise é por meio de gráficos. O Orange [6], ferramenta para mineração de dados, explora bem essa funcionalidade. Na Figura 8, desenhada usando $k = 8$, e na Figura 9, gerada para $k = 6$, estão representadas as distribuições dos partidos por *clusters*. É possível identificar os agrupamentos com seus respectivos votos, nesse caso em particular é referente à primeira denúncia contra Temer. As numerações dos *clusters* não vão coincidir com as dos relatórios do Weka, mas dá para identificar, por exemplo, no gráfico da Figura 9 no *cluster* 1 uma concentração de ausentes e no cluster 5 uma concentração de votos NÃO semelhantes aos dados das tabelas do Weka.

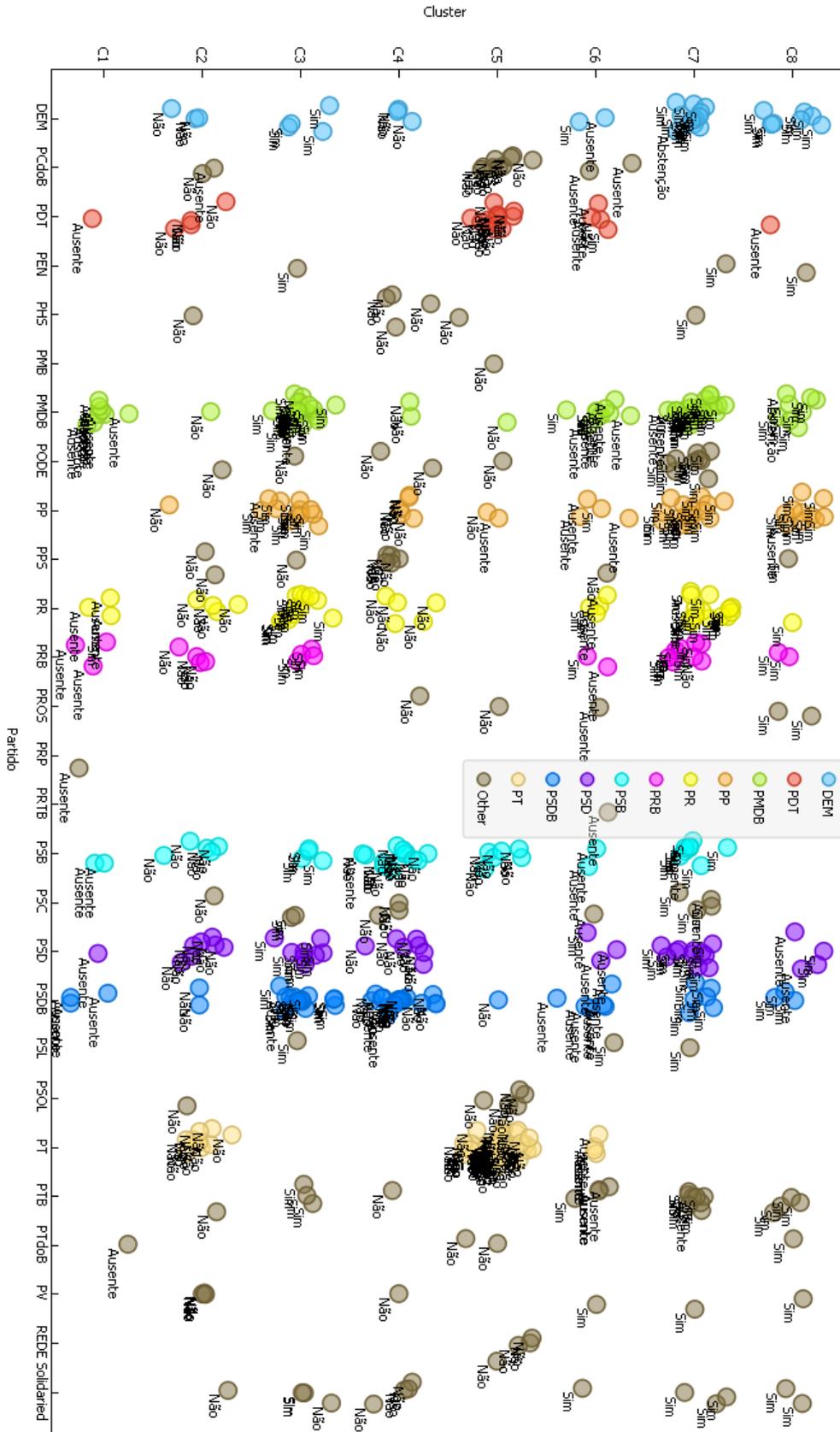


Figura 8: Distribuição de partidos para 8 clusters referente à proposição da primeira denúncia contra Temer

Fonte: Autoria própria

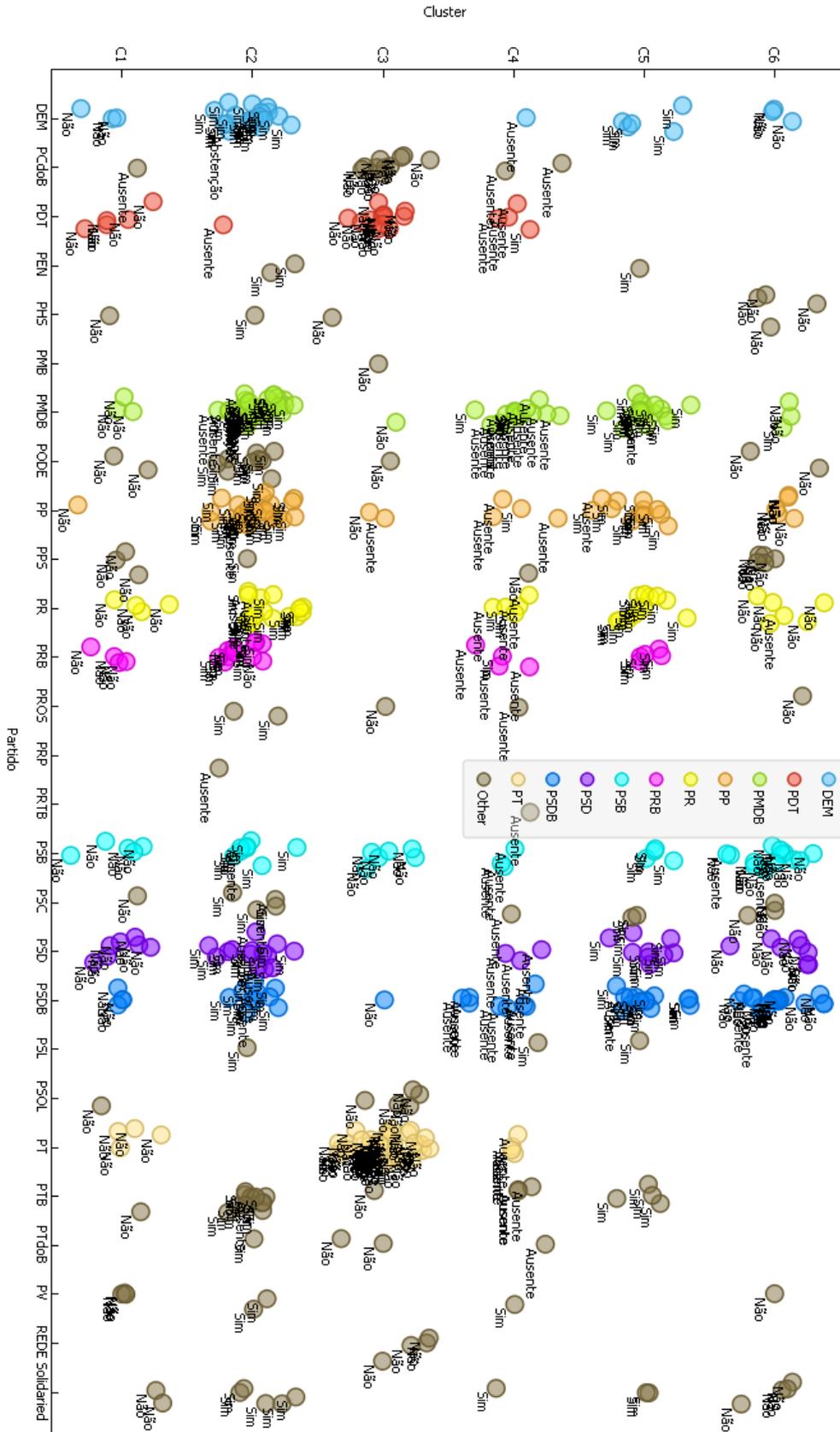


Figura 9: Distribuição dos partidos para 6 clusters referente à proposição da primeira denúncia contra Temer

Fonte: Autoria própria

Existe ainda outra funcionalidade do Weka usando o algoritmo K-means chamada *Classes to Clusters Evaluation*, que gera um relatório das participações de cada partido nos clusters. A Tabela 10 e a Tabela 11 com $k = 8$ e $k = 6$ respectivamente, contêm o resultado dessa análise. Cada coluna representa um *cluster* e nas linhas está a quantidade de votos por partido distribuídos pelos clusters. Esses resultados foram analisados em conjunto com a Tabela 6 e a Tabela 7 objetivando gerar um relatório que contenha informações que relacionem os partidos com os tipos de votos. O resultado disso pode ser observado na Tabela 12 que usa $k = 8$ e Tabela 13 originada a partir de $k = 6$. Na coluna instâncias estão totalizados os votos de cada partido. Os partidos em que mais de cinquenta por cento da sua bancada votaram a favor do governo estão na coluna governo. Seguindo a mesma regra, estão listados na coluna oposição os partidos que fizeram oposição ao governo. Há ainda partidos que não demonstraram um posicionamento efetivo tendo seus votos fracionados entre governo e oposição, e estão na coluna nomeada como dividido. O método de criação dessas duas últimas tabelas consiste em, por exemplo, analisar a Tabela 6 e identificar em quais clusters estão os votos favoráveis e em quais estão os votos de oposição ao governo. Transportando essas informações para Tabela 10 fica possível gerar a Tabela 12 que totaliza para cada partido o percentual de votos favoráveis ou não ao governo. A partir dessas informações dá para perceber quem apoia e quem faz oposição ao governo.

Finalmente comparando a Tabela 12 com a Tabela 13 fica visível de que usando o $k = 8$ ou $k = 6$ será obtido o mesmo resultado.

Tabela 10: k-means Classes to Clusters Evaluation para k = 8

```

Class attribute: Partido
Classes to Clusters:
 0  1  2  3  4  5  6  7  <-- assigned to cluster
 0  1  8  6  1 10  3  2  | DEM
 2  2 13 11  2 11  5  4  | PP
 2  2  7  2  0  5  0  3  | PTB
56  1  0  0  0  0  0  4  | PT
 2  3  3 11  0 14  5  5  | PR
 2  2  5  0  1  3  1  0  | PODE
 1  2  5 16  0  5 18  8  | PSDB
10  5  2  3  3  6  8  2  | PSB
15  2  1  0  0  0  0  4  | PDT
 1  2 16 19  1 21  2 15  | PMDB
 4  0  0  0  0  0  0  0  | REDE
 2  3  1  1  0  0  3  0  | PPS
 0  0  1  1  0  0  0  1  | PSL
10  1  0  0  0  0  0  2  | PCdoB
 1  7  8  6  3  6  6  4  | PSD
 0  1  1  2  1  2  3  1  | PSC
 1  3  4  6  0  7  0  3  | PRB
 0  1  2  2  0  0  1  1  | PV
 0  2  2  3  0  3  3  1  | Solidaried
 1  0  2  0  0  0  1  1  | PROS
 2  0  2  0  0  0  0  0  | PTdoB
 1  1  0  0  1  0  4  0  | PHS
 6  0  0  0  0  0  0  0  | PSOL
 0  0  3  0  0  0  0  0  | PEN
 0  0  1  0  0  0  0  0  | PRP
 0  0  0  0  0  0  0  1  | PRTB
 1  0  0  0  0  0  0  0  | PMB
    
```

Fonte: Autoria própria

Tabela 11: k-means Classes to Clusters Evaluation para k = 6

```

Class attribute: Partido
Classes to Clusters:
 0  1  2  3  4  5  <-- assigned to cluster
 0  3  9  6  1 12  | DEM
 4  5 14 12  2 13  | PP
 2  5  7  2  0  5  | PTB
57  4  0  0  0  0  | PT
 2  7  2 14  1 17  | PR
 4  0  6  0  1  3  | PODE
 1  9  8 17  0 20  | PSDB
15  5  2  3  3 11  | PSB
15  6  1  0  0  0  | PDT
 2 16 17 19  1 22  | PMDB
 4  0  0  0  0  0  | REDE
 3  2  1  2  1  1  | PPS
 0  1  1  1  0  0  | PSL
11  2  0  0  0  0  | PCdoB
 1  6  9 11  3 11  | PSD
 1  1  2  2  2  3  | PSC
 2  5  4  6  0  7  | PRB
 0  1  2  3  0  1  | PV
 2  1  2  4  0  5  | Solidaried
 1  1  3  0  0  0  | PROS
 2  0  2  0  0  0  | PTdoB
 2  0  1  1  1  2  | PHS
 6  0  0  0  0  0  | PSOL
 0  0  3  0  0  0  | PEN
 0  0  1  0  0  0  | PRP
 0  1  0  0  0  0  | PRTB
 1  0  0  0  0  0  | PMB
    
```

Fonte: Autoria própria

Tabela 12: Análise k-means para k = 8

INSTÂNCIAS	PARTIDOS	SIM	NÃO	AUSENTE	GOVERNO	OPOSIÇÃO	DIVIDIDO
31	DEM	0,77			DEM		
50	PP	0,70			PP		
21	PTB	0,57		0,14	PTB		
61	PT		0,92	0,07		PT	
43	PR	0,71		0,12	PR		
14	PODE	0,57	0,14		PODE		
55	PSDB	0,62		0,15	PSDB		
39	PSB	0,36	0,26				PSB
22	PDT		0,68	0,18		PDT	
77	PMDB	0,73			PMDB		
4	REDE		1,00			REDE	
10	PPS	0,30	0,20				PPS
3	PSL	0,66		0,33	PSL		
13	PCdoB		0,77	0,15		PCdoB	
41	PSD	0,65			PSD		
11	PSC	0,63			PSC		
24	PRB	0,71			PRB		
7	PV	0,72		0,14	PV		
14	Solidariedade	0,63			Solidariedade		
5	PROS	0,60	0,20	0,20	PROS		
4	PTdoB	0,50	0,50				PTdoB
7	PHS	0,71	0,14		PHS		
6	PSOL		1,00			PSOL	
3	PEN	1,00			PEN		
1	PRP	1,00			PRP		
1	PRTB			1,00			PRTB
1	PMB		1,00			PMB	

Fonte: Autoria própria

Tabela 13: Análise k-means para k = 6

INSTÂNCIAS	PARTIDOS	SIM	NÃO	AUSENTE	GOVERNO	OPOSIÇÃO	DIVIDIDO
31	DEM	0,87			DEM		
50	PP	0,78			PP		
21	PTB	0,57		0,24	PTB		
61	PT		0,93	0,07		PT	
43	PR	0,73		0,16	PR		
14	PODE	0,64	0,29		PODE		
55	PSDB	0,67		0,16	PSDB		
39	PSB	0,28	0,38	0,13			PSB
22	PDT	0,05	0,68	0,27		PDT	
77	PMDB	0,76			PMDB		
4	REDE		1,00			REDE	
10	PPS	0,20	0,30	0,20			PPS
3	PSL	0,66		0,33	PSL		
13	PCdoB		0,85	0,15		PCdoB	
41	PSD	0,76			PSD		
11	PSC	0,81			PSC		
24	PRB	0,54		0,21	PRB		
7	PV	0,86		0,14	PV		
14	Solidariedade	0,79	0,14		Solidariedade		
5	PROS	0,60	0,20	0,20	PROS		
4	PTdoB	0,50	0,50				PTdoB
7	PHS	0,71	0,29		PHS		
6	PSOL		1,00			PSOL	
3	PEN	1,00			PEN		
1	PRP	1,00			PRP		
1	PRTB			1,00			PRTB
1	PMB		1,00			PMB	

Fonte: Autoria própria

Capítulo 4

Conclusão e trabalhos futuros

4.1 Conclusão

Neste trabalho foi realizada uma análise utilizando as ferramentas Weka e Orange aplicando o K-means no processamento dos dados disponíveis no portal da Câmara dos Deputados. Os dados selecionados foram dez votações de proposições encaminhadas pelo Poder Executivo à Câmara dos Deputados.

Uma dificuldade típica neste tipo de análise é a definição do parâmetro k . Para realizar essa definição foram usados os métodos Elbow e o coeficiente de Silhouette. Os valores encontrados foram diferentes para cada método e a partir daí foi desenvolvida uma análise para cada possibilidade de valor de k .

A análise comparativa com os dois valores do parâmetro k mostrou que, para este trabalho, não fez diferença nos resultados obtidos.

Ficou claro que é possível classificar os partidos políticos a partir das proposições encaminhadas pelo Poder Executivo para votação na Câmara dos Deputados. Pôde ser destacada a existência de partidos divididos na hora de se posicionarem e também que um dos partidos de oposição, o PT obteve mais engajamento de seus membros do que o PMDB, partido do governo.

A utilização desse tipo de análise poderá ser feita de modo interativo com o portal de dados abertos da Câmara dos Deputados a partir do momento em que as votações estiverem disponíveis na arquitetura RESTful, bastando apenas o desenvolvimento de um aplicativo que faça essa análise. Neste caso a população teria uma ferramenta com dados atualizados, dando uma visão do comportamento dos partidos políticos frente ao voto.

Utilizando uma base de dados de votação dos congressistas semelhante a que foi empregada neste trabalho é possível também fazer outras análises, como por exemplo, aplicando técnicas de mineração de dados preditivas, possibilita inferir

previsões de como votariam os partidos políticos para determinadas questões polêmicas da sociedade atual tais como, o desarmamento, aborto, descriminalização de drogas entre e outras.

4.2 Trabalhos Futuros

4.2.1 Análise comparativa entre proposições votadas nos governos Dilma e Temer

Houve dois presidentes da república em uma mesma legislatura e nessa transição também houve mudanças no poder legislativo. Estudar a dinâmica de como os partidos políticos, expressados pelos votos, se articularam passando de governo para oposição e vice-versa, pode dar transparência quanto às motivações de tais decisões.

4.2.2 Análise das votações sobre a perspectiva das frentes parlamentares ou bancadas

É sabido da existência de diversas bancadas no congresso nacional que tem uma expressividade que transcende as diferenças político-partidárias. Essa verdade é revelada por serem compostas por parlamentares de partidos diversos, alguns até sem nenhuma afinidade ideológica. A representação dessas bancadas é tão significativa que o governo sucessor tem apostado sua capacidade de governabilidade no apoio dessas bancadas, tendo em vista que o modelo tradicional de negociação com as lideranças partidárias vem se traduzindo em comportamentos viciados, dificultando que mudanças ocorram.

As proposições postas em votação nas casas legislativas são apoiadas ou rejeitadas mais veementemente pelas bancadas, que realmente expressam alguma ideologia, do que pelos partidos políticos, que se tornaram mais útil como uma plataforma de lançamento de candidatos a uma eleição. Estudar as votações sob a perspectiva das bancadas pode revelar um modelo de representação que não é transparente para os eleitores, que só tem percebido os partidos políticos sendo apresentado como bandeira ideológica de seus candidatos.

4.2.3 Análise comparativa com outros algoritmos de agrupamentos

Neste trabalho foi utilizado o algoritmo K-means, mas também é possível empregar outros métodos de agrupamento, como por exemplo, o Hierárquico, que consiste em decompor o conjunto de dados em vários níveis. Comparar os resultados destes dois métodos permitiria uma visão dos dados sob outra perspectiva.

Bibliografia

- [1] *Aplicativos, visualizações e infográficos produzidos com dados abertos*. (s.d.). Acesso em 03 de 11 de 2018, disponível em Portal Brasileiro de Dados Abertos: <http://dados.gov.br/aplicativos>
- [2] *Benefits of Open Data*. (s.d.). Acesso em 08 de 11 de 2018, disponível em European Data Portal: <https://www.europeandataportal.eu/en/using-data/benefits-of-open-data>
- [3] *Business Case for Open Data*. (s.d.). Acesso em 08 de 11 de 2018, disponível em Project Open Data: <https://project-open-data.cio.gov/business-case/>
- [4] *Dados Abertos*. (s.d.). Acesso em 08 de 11 de 2018, disponível em Câmara dos Deputados: <https://dadosabertos.camara.leg.br/swagger/api.html#api>
- [5] de Castro, L. N., & Ferrari, D. G. (2016). *Introdução à mineração de dados: conceitos básicos, algoritmos e aplicações* (1 ed.). São Paulo: Saraiva.
- [6] *Documentation*. (s.d.). Acesso em 10 de 11 de 2018, disponível em Orange: <https://orange.biolab.si/>
- [7] Han, J., Kamber, M., & Pei, J. (2012). *Data Mine: Concept and Techniques* (3 ed.). Waltham, MA, USA: Elsevier.
- [8] *Mineração de dados: do estado bruto à sabedoria*. (07 de 11 de 2018). Acesso em 08 de 11 de 2018, disponível em Tribunal de Contas do Estado da Bahia: <https://www.tce.ba.gov.br/noticias/desafios-para-grandes-volumes-de-dados-e-encontrar-as-relacoes-entre-dados-e-entender-as-relacoes-encontradas>
- [9] *Regimento Interno da CD*. (s.d.). Acesso em 18 de 11 de 2018, disponível em Câmara dos Deputados: http://www2.camara.leg.br/atividade-legislativa/legislacao/Constituicoes_Brasileiras/regimento-interno-da-camara-dos-deputados
- [10] Santana, F. (12 de 12 de 2017). *Entenda o Algoritmo K-means e Saiba como Aplicar essa Técnica*. Acesso em 17 de 11 de 2018, disponível em Minerando

Dados: <http://minerandodados.com.br/index.php/2017/12/12/entenda-o-algoritmo-k-means/>

[11] *Silhouette Plot*. (s.d.). Acesso em 04 de 11 de 2018, disponível em Orange Visual Programming: <https://orange-visual-programming.readthedocs.io/widgets/visualize/silhouetteplot.html>

[12] *Vantagens da publicação de dados abertos*. (s.d.). Acesso em 08 de 11 de 2018, disponível em Kit para dados abertos: <http://kit.dados.gov.br/vantagens-dados-abertos/>

[13] Witten, I. H., Eibe, F., & Hall, M. A. (2011). *Data mining : practical machine learning tools and techniques* (3 ed.). Burlington, MA, USA: Elsevier.