Projeto Orientado em Computação II
Classificador de votos em proposições da Câmara dos Deputados

Adler Melgaço Ferreira
adlermf@dcc.ufmg.br
Orientador: Marcos Augusto dos Santos
marcos@dcc.ufmg.br

Universidade Federal de Minas Gerais

2 de dezembro de 2019

Resumo

Previsões sempre fizeram parte do imaginário humano, e a área da política é umas das áreas que mais tenta se aproveitar disso. Assim, o projeto consiste na criação de uma ferramenta para prever votos de deputados baseando-se no texto da ementa das proposições votadas em plenário. Além disso, é feita também uma busca pelas palavras mais frequentes nas proposições, agrupando essas palavras por partido.

Palavras-chaves: classificação, regressão, política, ementa, Câmara dos Deputados.

1 Introdução

O ser humano sempre se interessou pelo futuro. Seja para predizer quando será a próxima chuva, o próximo ataque inimigo ou se um determinado produto será um sucesso de vendas ou não, a espécie humana utilizou-se de mecanismos e misticismos para realizar previsões, na tentativa de facilitar o desenvolvimento da civilização, seja para o interesse coletivo ou para um determinado grupo.

Na Grécia Antiga, a figura do Oráculo de Delfos possuía um papel tanto espiritual quanto político, sendo consultado antes de decisões bélicas, com suas respostas sempre sob a justificativa de estarem de acordo com a orientação recebida pelos deuses.

Passados pouco mais de dois milênios, ainda há aqueles que se baseiam na espiritualidade para realizar suas previsões, mas a sociedade criou meios mais concretos que pudessem formular palpites cada vez mais certeiros, com conhecimentos matemáticos oriundos da estatística e ferramentas desenvolvidas pela computação.

No mundo atual, um dos setores mais interessados em fazer uso dessas abordagens é o da política, cujo proveito, por parte dos candidatos, está em saber qual o perfil do indivíduo mais disposto a votar em sua candidatura[1]. Há também a atividade do *lobby*, cuja definição mais clássica se refere à atividade, por parte de um grupo, de influenciar o poder público a tomar decisões que sejam de interesse desse grupo. Dessa forma, quando é feita uma votação no plenário, há uma busca por parte de políticos em identificar quais são os votos que já estão garantidos e os que ainda são duvidosos, de forma a garantir a aprovação ou não da proposição em questão.

Com isso em mente, o projeto teve como foco analisar as atividades legislativas da Câmara dos Deputados, visando traçar o comportamento de determinados deputados e prever qual será o seu voto em uma dada proposição, verificando quais são os assuntos com os quais ele mais se importa, a influência de seu partido em suas decisões, entre outros fatores. Além disso, uma análise sobre os próprios partidos também foi feita, com relação às proposições elaboradas por seus membros, na tentativa de identificar quais são os temas mais recorrentes em suas propostas, e se estes estão de acordo com o que é esperado do partido em questão.

2 Referencial Teórico

Como dito antes, a política é uma das áreas nas quais a possibilidade de se realizar uma previsão possui uma relevância considerável. Assim, existem trabalhos de classificação que focam em diversos aspectos, como a orientação política no discurso[1], diferenciação de notícias de cunho político[2], entre outros.

Um trabalho de caráter mais semelhante foi publicado[3] em 2011, por pesquisadores da Universidade de Princeton, nos Estados Unidos, com um modelo que previa o comportamento de legisladores e o padrão de suas votações.

No Brasil, foi publicado um artigo em 2015[4], analisando a semelhança entre os diversos partidos que existem no país, e levantando a discussão sobre quais deles são realmente diferentes uns dos outros. Além disso, há o projeto "Operação Serenata de Amor"[2], que tem como objetivo fiscalizar os gastos públicos e tornar mais fácil o acesso a esse tipo de informação.

3 Desenvolvimento do projeto

O projeto foi desenvolvido utilizando-se a linguagem Python no ambiente do Google Colab, para realizar todo o tratamento inicial dos dados, a formação da base de dados de interesse e a a análise sobre as palavras mais relevantes para cada partido, enquanto o ambiente de protipagem MATLAB foi utilizado para gerar os resultados do classificador. Os detalhes do desenvolvimento serão explanados nas subseções a seguir.

3.1 Base de dados

No ano de 2006, a Câmara dos Deputados lançou um serviço de dados abertos com o nome de "SIT Câmara", o qual exigia o cadastro do cidadão que quisesse trabalhar com esses dados, porém, em 2011, com a promulgação da Lei de Acesso à Informação, esse cadastro

[1] <https://www.economist.com/graphic-detail/2018/11/03/how-to-forecast-an-americans-vote>
[2] <https://serenata.ai>

deixou de existir, e o nome do serviço passou a ser "Dados Abertos". Então, em 2017, a API do Dados Abertos[3] foi lançada, sendo atualizada e renovada. O portal disponibiliza diversos tipos de dados públicos a respeito da Câmara dos Deputados, existindo informações sobre proposições, discursos de deputados, comissões formadas, etc.

Devido a natureza do classificador sendo desenvolvido, era necessário que a API possuísse meios de relacionar as proposições com os votos dos deputados, entretanto, como se trata de uma ferramenta relativamente nova, essa API ainda não possui esse tipo de informação e, por causa disso, o serviço antigo foi utilizado.[4]

3.2 Pré-processamento

Pela API antiga, através de chamadas GET no protocolo HTTP é possível extrair os dados que eram necessários. Primeiramente, foram extraídas todas as proposições que foram votadas em plenário do ano 1999 até o ano de 2017, sendo estas do tipo PEC(Proposta de Emenda à Constituição) e PL(Projeto de Lei), obtendo assim o número, tipo e ano dessas proposições.

A partir desses três valores, é possível então obter tanto a ementa da proposição, ou seja, o texto que explica resumidamente a que ela se refere, quanto também informações sobre a sua votação, como a orientação de cada partido ou bancada sobre ela, o voto de cada um dos deputados, etc.

Para cada proposição, cada uma das palavras de sua ementa eram separadas, removendo ainda palavras consideradas irrelevantes, que apareceriam demais devido a sua natureza na língua portuguesa, como preposições(de, da, do, pois...), artigos(o, a, um, uma...), entre outras, realizando em seguida a extração dos radicais nas palavras que sobraram.

3.3 Radicais

Apenas para clarificar, radical é o elemento básico e significativo de uma palavra, ou seja é o local onde se encontra o seu significado principal. Essa métrica foi usada justamente por isso, já que assim seria possível preservar e centralizar a importância de cada palavra nas proposições. Segue um exemplo, onde os trechos em negrito são os radicais:

Perder - **Perd**ido - **Perd**edor - Im**perd**ível

3.4 A Matriz

Depois que esses passos são feitos, a matriz então é formada, sendo que cada um dos radicais encontrados representa um atributo, enquanto as linhas representam as instâncias. Caso a proposição possua aquele radical, então é marcado com o número 1, caso contrário, o número 0 é que fica marcado. A última coluna representa o voto do deputado, sendo 1 para "Sim", e 0 para "Não".

Devido à forma a qual os dados foram organizados, não há uma diferenciação entre as diferentes votações que podem haver sobre a mesma proposição. Por exemplo, a votação sobre emendas substitutivas, alterações e a proposição em si têm a mesma identificação,

[3] <https://dadosabertos.camara.leg.br/index.html>

[4] <https://www2.camara.leg.br/transparencia/dados-abertos/dados-abertos-legislativo/dados-abertos-legislativo>

possuindo assim o mesmo texto da ementa. Dessa forma, o mesmo deputado pode ter votado de maneiras diferentes com relação a esses tipos distintos, mas como não é possível detectar essas distinções, apenas uma votação sobre a proposição foi mantida, reduzindo assim o número de entidades possíveis.

No fim, a matriz final possui dimensões de 282 linhas por 1050 colunas. Segue um exemplo ilustrativo de como se dispõe a matriz:

ID	alt	disposi	art	feder	...	Voto
PEC 33/1999	1	1	0	0	...	1
PEC 82/1995	0	1	0	1	...	1
PEC 85/1999	0	0	1	0	...	0
...	...	...	...	...	...	...
PEC 169/1993	1	0	0	1	...	0

3.5 Modelo

No classificador desenvolvido, o modelo é feito por deputado, não sendo então um modelo genérico para realizar uma previsão para qualquer um. Dito isso, o modelo foi desenvolvido com base no deputado Pedro Chaves, pois este foi identificado como o que possuía o maior volume de dados, já que ele exerceu mandatos do ano de 1999 até o ano de 2017, permanecendo no partido MDB(antigo PMDB) durante todos esses anos.

Para a obtenção dos resultados, a técnica de estatística conhecida como Regressão Logística foi utilizada, todavia, em virtude da natureza da matriz, que possui um número de atributos consideravelmente maior do que o de instâncias, um modelo de regressão comum não poderia funcionar corretamente, já que o posto da matriz não permite que o algoritmo funcione de maneira esperada.

Dessa forma, o seguinte cálculo foi feito:

$$\vec{\alpha} = (I + A^T * A) \backslash A^T * \vec{b}$$

Fazendo o cálculo dessa maneira, isso permite que o problema com a disparidade entre instâncias e atributos seja tratado. Nessa equação, A é a matriz formada por votações e radicais, I é a matriz identidade de A, e b é um vetor com valores logaritmos correspondentes. Já o vetor alfa representa a relevância de cada um dos radicais para a classificação, quanto maior o seu valor, mais relevância ele possui.

A partir da variável alfa, é possível então obter a probabilidade p do deputado votar "Sim" em uma dada proposição, através da fórmula a seguir:

$$p = \frac{e^{(q^T * \alpha)}}{1 + e^{(q^T * \alpha)}}$$

Na qual a variável q representa a instância que está sendo analisada, ou seja, a proposição que se deseja obter a previsão.

3.6 Dados de saída

O gráfico gerado a partir dos dados de alfa é o seguinte:

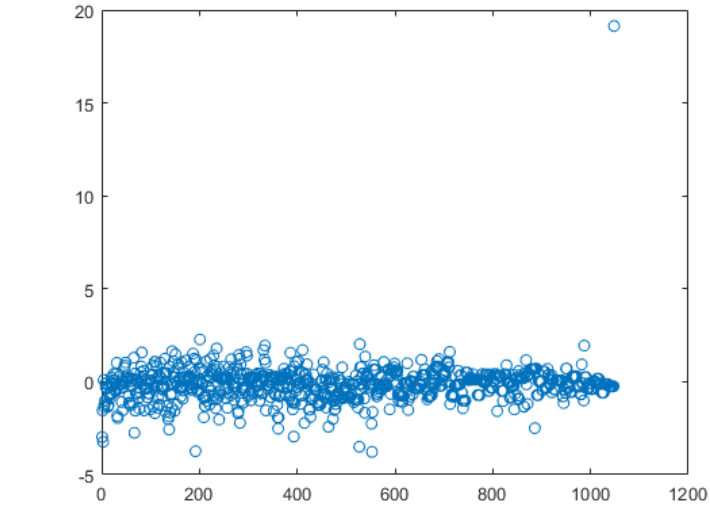
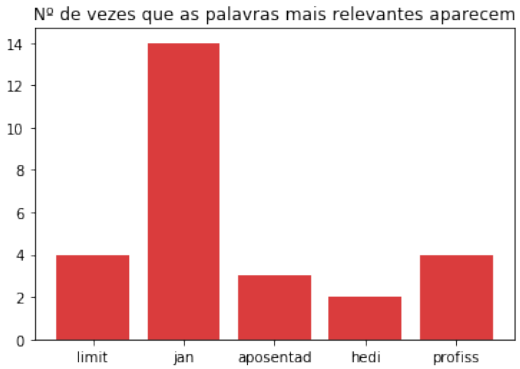
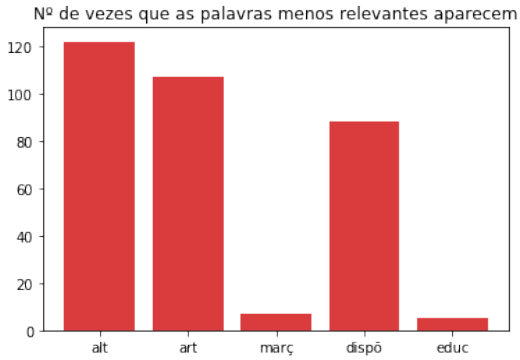


Figura 1 – Valores do vetor alfa.

É possível então verificar quais são os radicais correspondentes aos maiores e e menores valores de alfa, examinando também quantas vezes esses radicais aparecem ao longo da base de dados. Os gráficos estão ordenados de maneira descrescente em relação aos seus valores alfa.



Radicais com os maiores valores de alfa



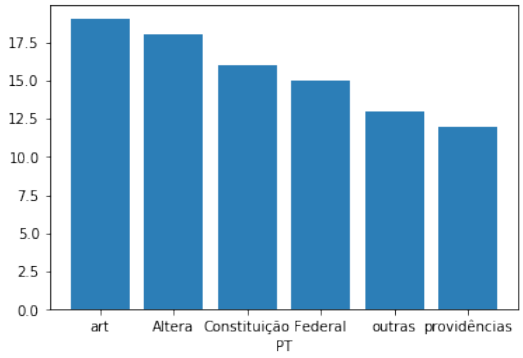
Radicais com os menores valores de alfa

É interessante notar que o número de vezes que o radical aparece não significa simplesmente que ele será ou não relevante para a classificação do voto, embora o atributo "jan" apareça mais vezes que "limit", este é mais relevante do que aquele. E embora os radicais "alt" e "art" apareçam bastante vezes ao longo da base, eles não são tão relevantes. Na verdade, no caso desses dois indivíduos, é justamente por aparecerem tantas vezes é que eles não possuem tanta importância.

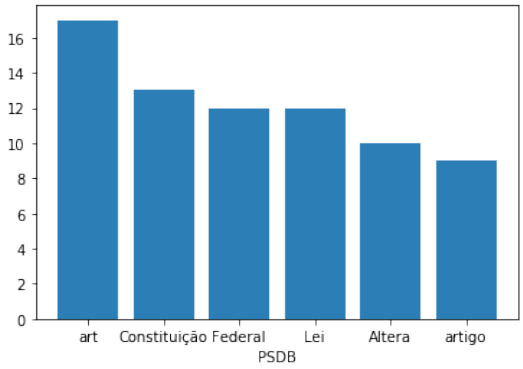
3.7 Análise dos partidos

O outro foco do projeto foi realizar uma análise das palavras mais frequentes nas proposições de acordo com cada partido, em busca de encontrar algum padrão, e investigar se os temas que mais aparecem estão em sintonia com o que o partido prega.

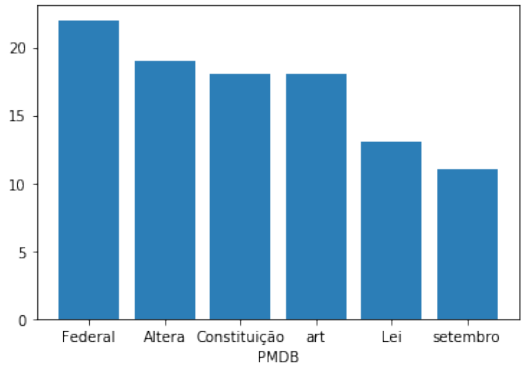
Seguem alguns gráficos com as palavras mais frequentes:



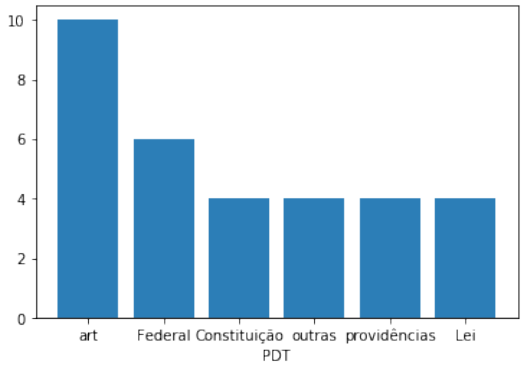
Palavras mais frequentes do PT



Palavras mais frequentes do PSDB



Palavras mais frequentes do PMDB



Palavras mais frequentes do PDT

Como é possível perceber, os gráficos estão bem parecidos, e isto se repete para grande parte dos partidos, indicando que, futuramente, é necessário realizar uma remoção das palavras mais relacionadas à linguagem legislativa da Câmara para obter resultados mais interessantes.

4 Conclusões

Embora até o momento o classificador possua um modelo apenas para o deputado Pedro Chaves, o desenvolvimento do projeto se mostrou bastante promissor, levantando questões sobre quanto o conteúdo textual de uma proposição realmente infuencia na decisão do deputado. No futuro, modelos para mais deputados podem ser desenvolvidos, possibilitando também que um modelo mais genérico, que consiga obter respostas para qualquer indivíduo, seja implementado. Além disso, levar em conta outros fatores, como a orientação do partido a respeito daquela votação, diferentes votações sobre a mesma proposição, decisões distintas do deputado, como "Abstenção"e "Obstrução", são imprescidíveis para obter resultados mais robustos e fiéis.

Sobre a análise das palavras mais frequentes dos partidos, foi importante compreender que há palavras no vocabulário legislativo que surgem repetidas vezes, e por causa disso, um tratamento mais adequado sobre essas palavras é necessário para que se realize um estudo mais aprofundado.

5 Referências Bibliográficas

[1] Yan, H., Lavoie, A., Das, S. (2017). The Perils of Classifying Political Orientation From Text.

[2] Budak, C., Goel, S., Rao, J.M.: Fair and balanced? Quantifying media bias through crowdsourced content analysis. Public Opin. Quarterly 80(S1), 250–271 (2016)

[3] Gerrish, S., Blei, D.M.: Predicting legislative roll calls from text. In: Proc. ICML. pp. 489– 496 (2011)

[4] Vaz de Melo POS (2015) How Many Political Parties Should Brazil Have? A Data-Driven Method to Assess and Reduce Fragmentation in Multi-Party Political Systems. PLoS ONE 10(10): e0140217