Mineração de textos - Agrupamento de datasets públicos

Henrique Gomes Zanin - 10441321

[1]Relatório desenvolvido para a
Disciplina de Mineração de Dados não Estruturados - ICMC-USP

Resumo. *Os dados abertos brasileiros possui um conjunto de metadados associados aos datasets. Esses metadados foram utilizados no ambito deste projeto para agrupar dados semelhantes de acordo com o título e a descrição dos dados. Foram aplicados os métodos TF-IDF e BERT para geração de vetores de representação para posterior agrupamento utilizando os algoritmos Kmeans e DBScan. Por meio dos experimentos realizados conclui-se que o título do dataset não é um bom candidato para geração de agrupamentos, sendo a descrição um atributo capaz de gerar melhores medidas de silhueta em todos os casos. Outro ponto a ser considerado é o fato do algoritmo DBScan superar a qualidade dos agrupamentos do Kmeans.*

1. Identificação do Problema

O Plano de Dados Abertos instituído pelo governo federal possui como principal objetivo garantir a acessibilidade de dados abertos originários de organizações públicas brasileiras. No Portal Brasileiro de Dados Abertos, estão presentes 265 organizações públicas de diversos setores econômicos e sociais.

Apesar de possuir um mecanismo de busca unificado, capaz de percorrer os dados das 265 organizações, não há nenhum caminho que permita encontrar dados similares intrainstituições e interinstituições. Dessa forma, fica atribuído ao consumidor o papel de especialista para encontrar dados relacionados entre si.

O objetivo desse projeto é agrupar dados similares entre si para facilitar ao consumidor a exploração de dados relacionados ao que o mesmo procura. Em verdade, o portal apresenta uma forma de agrupamento baseada em etiquetas, sendo essa, porem, limitada e demandante de ajustes manuais de atribuições de dados e rótulos.

A abordagem de agrupamento proposta nesse trabalho compreende a aplicação de técnicas de aprendizado não supervisionado para determinação de grupos de dados similares. Para atingir esse objetivo, avaliamos eficácia da utilização do título do dado em comparação com a sua descrição.

Como estudo de caso foi escolhida a *api* do Instituto de Pesquisa Econômica Aplicada (IPEA) para extração dos metadados de todas as series temporais do Instituto. Todos os metadados foram extraídos e organizados em um CSV contendo o título e a descrição de um dataset. O CSV com os metadadados acompanha a entrega deste relatório. O *endpoint* da *api* utilizado para a extração é o que se segue: `http://www.ipeadata.gov.br/api/odata4/Metadados`

A Figura 1 a seguir exibe um exemplo das variáveis utilizadas (titulo, descrição) para a atividade de agrupamento:

Para avaliar o resultado obtido pelo agrupamento utilizou-se critérios quantitativos como a comparação de silhueta dos métodos de agrupamento Kmeans e DBSCAN

	name	description
0	Balanço de pagamentos - conta capital - despesa	O balanço de pagamentos é uma peça contábil qu...
1	Base Monetária restrita - M0 - reservas bancár...	A base monetária corresponde ao passivo monetá...
2	Produção industrial - bens de consumo não durá...	A produção industrial de bens de consumo não d...
3	Exportações - produtos de madeira - preços - í...	Este índice busca captar o efeito dos preços s...
4	Taxa de câmbio real bilateral - IPA-DI - Brasi...	A taxa de câmbio real bilateral é definida pel...
...	...	...
4405	Indicador IPEA de FBCF - índice real (média 19...	A Formação Bruta de Capital Fixo (FBCF) da eco...
4406	Taxa de juros prefixada - estrutura a termo - ...	A taxa de juros é o coeficiente que determina ...
4407	Índice de custo da tecnologia da informação (I...	O Índice de Custos da Tecnologia da Informação...
4408	IPA-M - 1º decêndio	O Índice de Preços ao Produtor Amplo (IPA) é u...
4409	INCC-10 - geral - índice (ago. 1994 = 100)	O Índice Nacional de Custo da Construção (INCC...

Figura 1. Conjunto de dados

aplicados nos títulos e nas descrições dos dados separadamente. Outra forma de avaliar a qualidade do agrupamento foi a verificação da distância dos centroides de cada cluster por meio de uma matriz de distâncias, essa visualização permite comparar qualitativamente se os grupos próximos possuem significado semântico.

2. Pré-processamento

A etapa de pré-processamento pode ser dividida em dois grupos, o pré processamento para o método TF-IDF e para o BERT. Ambos os métodos consistem na construção de vetores numéricos que representam os textos usados como base para a aplicação dos métodos de agrupamento. Fazem parte do conjunto de dados 4406 datasets públicos

O pré-processamento realizado para o cálculo do TF-IDF apresenta as seguintes etapas em ordem de execução:

1. Remoção de stopwords: Consiste em remover palavras não relevantes para o agrupamento. Para essa tarefa utilizou-se o pacote nltk contendo as stopwords da lingua portuguesa.
2. Stemming de palavras: Reduz uma palavra flexionada à sua raiz. Dessa forma remove-se a derivação de uma palavra para que o radical possa ser interpretado igualmente. Utilizou-se a função word_tokenizer do pacote nltk para a lingua portuguesa.
3. Aplicação do TF-IDF: Construção do vetor numérico que representa o conjunto de palavras presentes no texto. O TF-IDF aplica uma ponderação em palavras que ocorrem raramente, aumentando a sua relevância em relação às palavras mais frequentes.

O resultado final para o TF-IDF quanto ao número de atributos deu-se da seguinte forma:

- Aplicado no título: **364** atributos no vetor de representação
- Aplicado na descrição: **543** atributos no vetor de representação

A segunda forma de representação vetorial consistiu na aplicação do Bert tanto no título como na descrição. Para construção do vetor de representação textual utilizou-se o modelo ”distiluse-base-multilingual-cased-v2”, resultando em 512 atributos para ambos os casos (título, descrição).

3. Extração de Padrões

Para o agrupamento de dados foram aplicados dois algoritmos, o Kmeans e o DBSCAN. Os parâmetros de cada algoritmo estão representados a seguir:

- **Kmeans:** Numero de clusters(n_clusters), estado inicial(random_state), numero de centroides iniciais(n_init)
- **DBScan:** EPS(Raio máximo entre dois pontos para serem considerados do mesmo cluster), min_samples(número mínimo de pontos para garantir uma densidade desejada), metrica de distância

As otimizações realizadas no âmbito do trabalho concentraram-se no número de clusteres do algoritmo Kmeans e no EPS do DBScan. A medida de silhueta foi utilizada como métrica principal de avaliação da qualidade do agrupamento, os melhores resultados obtidos para o título e a descrição dos datasets são os seguintes:

- **Kmeans:**
  - TF-IDF Título do dataset: k = 43, silhueta = 0.1386
  - TF-IDF Descrição do dataset: k = 49, silhueta = 0.4223
  - BERT Título do dataset: k = 48, silhueta = 0.1588
  - BERT Descrição do dataset: k = 50, silhueta = 0.4638
- **DBScan:**
  - TF-IDF Título do dataset: eps = 0.49, silhueta = 0.2
  - TF-IDF Descrição do dataset: eps = 0.25, silhueta = 0.6219
  - BERT Título do dataset: eps = 0.55, silhueta = 0.1041
  - BERT Descrição do dataset: eps = 0.15, silhueta = 0.5955

Os experimentos realizados com o TF-IDF utilizaram os ngramas (2,2) e (3,3), porém não houve variação significativa quando o parâmetro ngram_range foi (3,3). Os testes envolvendo o algoritmo DBScan foram executados até a falha para os valores do parâmetro *eps*.

4. Pós-processamento

As subseções a seguir discutem os resultados para cada um dos dos métodos de geração dos vetores de representação associados aos algoritmos de agrupamento. Fica visível a partir das visualizações que a utilização do título como fator de geração dos vetores é insuficiente para a construção de bons agrupamentos.

4.1. Título do dataset

O título de um dataset não apresenta um bom agrupamento. No melhor caso a silhueta não ultrapassa o valor de 0.2, o que demonstra a ineficácia de sua utilização como fator de agrupamento.

4.1.1. Kmeans

O método BERT apresentou valores levemente superiores de silhueta quando comparado com o TF-IDF. Porém, o melhor valor de silhueta entre os testes foi de 0.1588 para 48

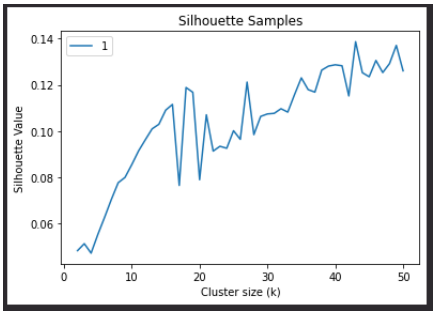

Figura 2. TF-IDF - Silhueta para os valores de K

grupos. A Figura 2 exibe os valores das silhuetas usando o TF-IDF, a partir do k = 18 não há ganhos expressivos no aumento do número de agrupamentos.

Ao utilizar o BERT é possível observar que há certa estabilidade quando o numero de clusters atinge 29 unidades. A Figura 3 exibe a trajetória da silhueta para os valores de k testados.

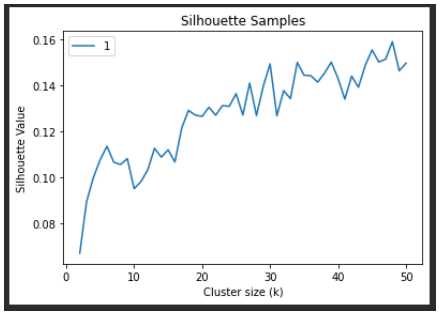

Figura 3. BERT - Silhueta para os valores de K

4.1.2. DBScan

O algoritmo DBScan utiliza uma estratégia de calcular a densidade de uma região, sendo o raio máximo entre dois pontos o atributo mais importante para determinar a melhor silhueta entre os grupos. Mesmo substituindo o kmeans pelo DBScan não houve melhora significativa da silhueta. A Figura 4 exibe os valores do EPS para o TF-IDF.

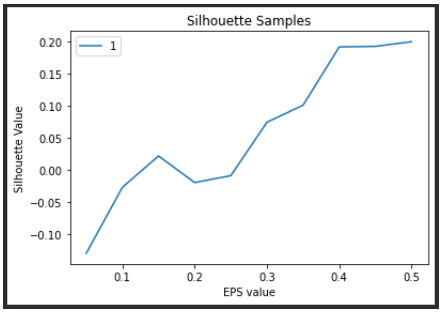

Figura 4. TF-IDF - Silhueta para os valores de EPS

O BERT apresentou um resultado consideravelmente inferior ao TF-IDF quando aplicado o DBScan. No melhor caso o valor foi a metade do melhor valor do TF-IDF.

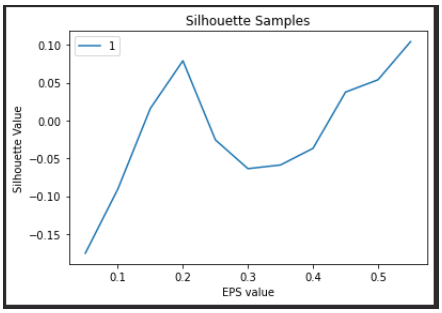

Figura 5. BERT - Silhueta para os valores de EPS

4.2. Descrição do dataset

A descrição do dataset contém uma quantidade consideravelmente maior de texto, fator que colabora para um agrupamento melhor de acordo com a medida de silhueta dos testes. Nesse caso houve uma melhora sigfificativa na utilização do DBScan quando comparado com o Kmeans.

4.2.1. Kmeans

É perceptível que acréscimos nos valores de k melhora a métrica de silhueta consideravelmente. A Figura 6 exibe a trajetória da silhueta para o TF-IDF na descrição e a Figura 7 para o BERT

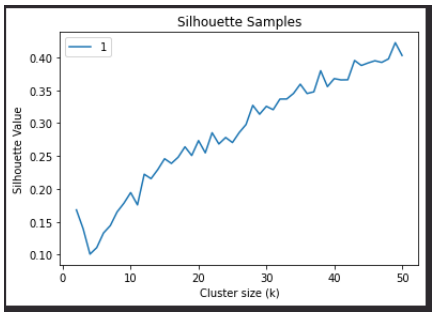

Figura 6. TF-IDF - Silhueta para os valores de K

4.2.2. DBScan

Quando aplicado à descrição o DBScan apresentou os melhores resultados tanto no TF-IDF como no BERT. Há uma leve mudança quanto ao parâmetro eps entre as duas aplicações. O melhor valor para o TF-IDF foi 0.25, enquanto para o BERT foi 0.15. A variação na silhueta para os melhores valores do eps não são expressivas quando comparados os métodos TF-IDF e BERT.

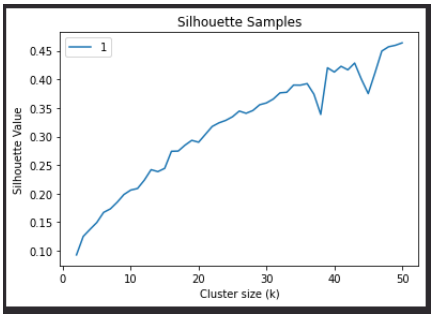

Figura 7. BERT - Silhueta para os valores de K

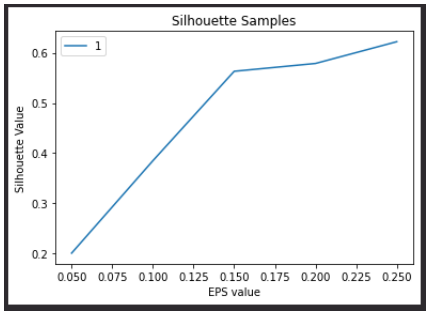

Figura 8. TF-IDF - Silhueta para os valores de EPS

5. Uso do Conhecimento

A partir dos experimentos realizados fica patente que os métodos de agrupamentos devem ser aplicados na descrição do dataset. Todas as abordagens de construção dos vetores de representação e os algoritmos de agrupamento apresentaram resultados superiores de separação de grupos. As Figuras 10 e 11 representam a diferença entre a utilização do título e da descrição quanto a distância entre os agrupamentos no Kmeans. A primeira representa a distância entre os grupos no título e a segunda figura na descrição, ambos utilizando o BERT. Quanto menor o valor há mais proximidade entre grupos. A distância entre os grupos permite identificar grupos semelhantes textualmente, o que pode facilitar a recomendação de dados relacionados.

O DBScan não permite verificar a distância entre os grupos pois não há um centroide no agrupamento, o algoritmo utiliza um cálculo de densidade que inviabiliza a definição de um centro. Uma consideração importante acerca do DBScan é a quanti-

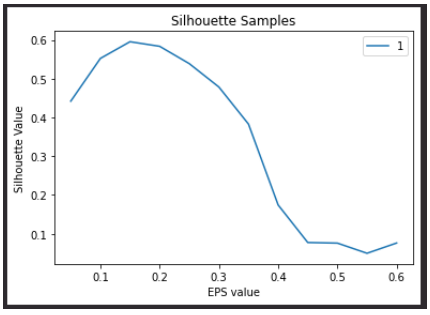

Figura 9. BERT - Silhueta para os valores de EPS

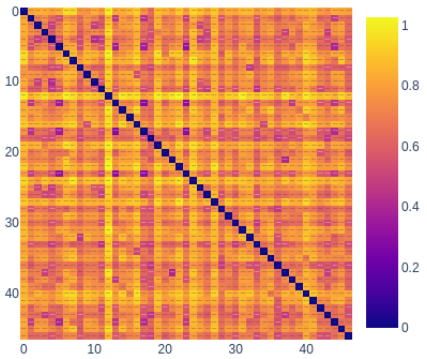

Figura 10. Matriz de distância entre os agrupamentos - BERT - Título

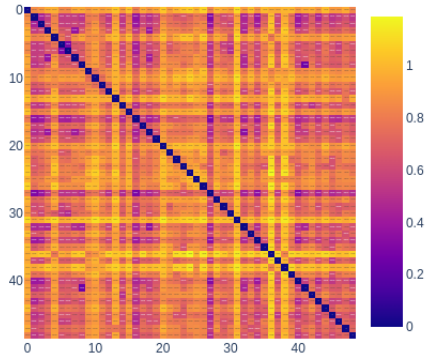

Figura 11. Matriz de distância entre os agrupamentos - BERT - Descrição

dade de grupos gerada, no melhor valor de silhueta são gerados 133 agrupamentos para a descrição usando o BERT, essa quantidade de grupos pode prejudicar a interpretabilidade por parte de agentes humanos, sendo difícil compreender a semântica por trás de cada grupo.