INTELIGÊNCIA ARTIFICIAL CLUSTERIZAÇÃO: PROBLEMA DOS AEROPORTOS EM BELÉM

Kalil Saldanha Kaliffe[1],**Guilherme Farias da Silva Leite**[1],**Marcos Eduardo Nascimento Lima**[1]

[1]Instituto de Ciências Exatas e Naturais - Faculdade de Computação
Universidade Federal do Pará (UFPA) - Belém, PA - Brasil

{kalil.kaliffe, guilherme.leite, marcos.lima}@icen.ufpa.br

Abstract. *This article aims to implement a reasonable solution to the problem introduced by the professor Reginaldo Cordeiro dos Santos Filho, called: "problem of airports in Belém"from the various concepts and learning acquired in the classroom such as clustering, then illustrate and discuss the solution, showing the methods, functions, libraries and particularities applied in the development of the implementation.*

Resumo. *Este artigo tem como objetivo implementar uma solução satisfatória para o problema apresentado pelo professor Reginaldo Cordeiro dos Santos Filho, denominado: "problema dos aeroportos em Belém" a partir dos diversos conceitos e aprendizados adquiridos em sala de aula como a clusterização, e em seguida ilustrar e discorrer sobre a solução, mostrando os métodos, funções, bibliotecas e particularidades aplicados no desenvolvimento da implementação.*

1. Introdução

O agrupamento de dados ou "clustering" é uma técnica focada em no primeiro momento particionar um determinado conjunto de dados de acordo com suas características em comum, ou seja, necessariamente trabalhando os pontos de similaridade ou dissimilaridade desses dados através de funções. E dentro desse universo de "clusterização" existem diversos tipos e variações dessa técnica, podendo ser classificadas como hierárquicas ou não.

Neste relatório técnico escolhemos o método de caráter não-hierarquico conhecido como "K-means" para poder exemplificar e solucionar o problema dos aeroportos sobre agrupamento de dados. As motivações que levam a escolha do "K-means" para implementar a solução além de que se faz uso de dados do tipo real, partem do princípio de que o método "K-means" trabalha atualizando o valor dos centróides durante a execução do programa, e posteriormente compara a relação de distância para agrupa-los com base nesses centróides, ou seja, uma análise será efetuada em cima das informações coletadas e os elementos similares serão distintos a partir de uma função para poder definir a relação entre os dados individualmente, e posteriormente as relações entre os grupos identificados, partindo do uso das distâncias Euclidianas, da soma e do máximo para delimitar as fronteiras dos agrupamentos encontrados.

Dito isso, o relatório será especificado de maneira qual o tópico seguinte "2" irá descrever a base de dados utilizada na construção da implementação; "3" vai descorrer sobre como o trabalho foi desenvolvido em si, quais técnicas usadas na implementação e realização do relatório; "4" os dados de saída da implementação; e "5" discorre sobre a conclusão assimilada ao fim do relatório.

2. Descrição da base de dados

A base de dados utilizada para a implementação do algoritmo ”K-means”consiste nas coordenadas de cada bairro da cidade de Belém delimitados pelo enunciado do trabalho, onde foram verificados um número de 38 bairros compondo a tabela da Fig 1, dividida em 2 subgrupos: nome do bairro e coordenadas, sendo coordenadas divididas em duas colunas: latitude e longitude. Utilizando esses dados, a implementação apresentará o devido resultado para o problema dos aeroportos, a fonte das coordenadas dos bairros foi o site ’https://mapacep.com.br/index.php’.

Bairro,	Latitude,	Longitude
Val-de-Cães	-1.3820615,	-48.4775227
Aurá	-1.4043379159816836,	-48.3973207817254
Águas Lindas	-1.3917029,	-48.3888384
Barreiro	-1.4137485,	-48.484351
Batista Campos	-1.4611392,	-48.4898427
Bengui,	-1.3752594,	-48.4552816
Cabanagem,	-1.366319368356491,	-48.43215060129148
Campina,	-1.4523804,	-48.4992935
Canudos,	-1.453612,	-48.4612362
Castanheira,	-1.40629215405936,	-48.43145136846568
Cidade Velha,	-1.4619499,	-48.5062207
Condor,	-1.4723764,	-48.4812248
Coqueiro,	-1.3344158441322356,	-48.44344734808546
Cremação,	-1.4612731,	-48.476227
Curió-Utinga,	-1.430418,	-48.4464734
Fátima,	-1.440747232085611,	-48.47234190438533
Guamá,	-1.461980750110683,	-48.46339585838604
Guanabara,	-1.3981178,	-48.4202024
Jurunas,	-1.4708282,	-48.4937213
Mangueirão,	-1.3871,	-48.4458
Maracangalha,	-1.3991768,	-48.4812248
Marambaia,	-1.4040646,	-48.4537424
Marco,	-1.4337274026202258,	-48.46161739778115
Miramar,	-1.40670643,	-48.4912218
Nazaré,	-1.452286257609553,	-48.48125227059778
Parque Verde,	-1.367325,	-48.4437522
Pedreira,	-1.4233092090503592,	-48.47113678113785
Pratinha,	-1.3606557,	-48.4737282
Reduto,	-1.4460735,	-48.4924716
Sacramenta,	-1.4136943,	-48.476227
Souza,	-1.41255440721719,	-48.45077831798063
São Brás,	-1.4515674,	-48.4707716
São Clemente,	-1.3658691,	-48.4612362
Tapanã,	-1.3387568,	-48.4737282
Telégrafo,	-1.4248878,	-48.4862231
Terra Firme,	-1.457048,	-48.4512446
Umarizal,	-1.4413019,	-48.4837239
Una,	-1.370855676060548,	-48.4252926583561
Universitário,	-1.4585135004434184,	-48.44110298384286

Figura 1. Representação da base de dados por meio de uma tabela

3. Metodologia do trabalho

O trabalho foi dividido em algumas etapas, a coleta dos dados dos bairros, construção da base de dados dos bairros, implementação do processamento de dados no código e avaliação dos resultados no artigo.

Foi utilizada uma base de dados contendo 38 bairros da zona metropolitana de Belém, divididas em distritos onde as fronteiras foram delimitadas pelo enunciado do problema. Essa base de dados foi adquirida através da ferramenta mapacep, um site de pesquisa de localidades que retorna dados dos locais pesquisados. Cada bairro foi pesquisado, um de cada vez, e incrementado na tabela .csv da base de dados sua respectiva latitude e longitude.

A linguagem de programação utilizada para a implementação do algoritmo foi python, onde pode-se ter acesso a bibliotecas como numpy, que foi utilizada para calcular a média das distâncias dos bairros para assim encontrar os centroids e a distância euclidiana de um bairro para um centroid. Matplotlib, biblioteca que foi utilizada para fazer a plotagem dos gráficos, tanto da representação visual da região metropolitana de Belém e seus respectivos bairros quanto do resultado em si, mostrando os centroids como resultado. Pandas, biblioteca utilizada para a manipulação da base de dados, basicamente convertendo a base .csv para um 'DataFrame' que pode ser utilizado no algoritmo.

A partir disso, para cumprir o objetivo de encontrar as melhores longitudes e latitudes dos K centroids para a construção dos aeroportos, primeiro, o algoritmo recebe o número 4 como valor de K e o 'DataFrame' dos bairros, calcula o tamanho do DF como o número total de amostras, inicializa um dicionario de centroids que deve guardar os um par de coordenadas como item de um key que representa o número de um centroid, sendo o primeiro a ser adicionado ou seja a key 0 vai ser a amostra de Val-de-Cans, a posição 0 de Val-de-Cans permite sua identificação para o bloqueio de sua atualização de posição nos proximos passos do algoritmo.

Então, em adição ao centroid de Val-de-Cans o algoritmo irá lançar mais três centroids no plano de forma aleatória.

Assim, as amostras dos bairros são submetidas a uma analise para determinar sua classificaçaõ em relação aos centroids, elas são classificadas de acordo com suas menores Distâncias Euclidianas dos centroids. Logo, cada amostra vai ser um item da key do seu centroid mais perto no dict de classes.

Portanto, a partir da classificação um cluster é formado e os centroids são atualizados (exceto o centroid 0 de Val-de-Cans) com base na média das distâncias das amostras de sua classe.

Esse processo de atualização dos centroids tem como critério de parada o número de amostras. Quando o número de atualizações for igual ao número de amostras então o algoritmo retorna o dict das classes de amostras, do centroids e distância total calculada a partir soma das distâncias euclidianas entre os centroids e seus bairros.

Dessa forma, a métrica de qualidade da saída vai ser a distância total, a qual vai ser comparada com outras execuções para tentar minimizar ela e encontrar o melhor resultado, ou seja, a menor distância total, assim como encontrar o pior resultado, a maior distância total.

Além disso, os dicionários de centroids e de classes, vão ser usados para plotar as coordenadas dos bairros e centroids no mapa.

4. Resultados

Após exposta a representação do melhor e do pior cenário encontrados quando aplicadas 20 execuções do algoritmo desenvolvido (Figura 2 e 3), e adotando-se o método K-means, percebe-se que a implementação foi feliz ao suprir uma solução satisfatoria para o problema dos aeroportos sugerido no relatório, já que a melhor solução quando posta em comparação com a pior, demonstra que as distâncias entre os centróides foram minimizadas levando em consideração o ponto de partida aleatório dos centroídes e em todas as iterações o centróide fixo de Val-de-Cans não foi deslocado.

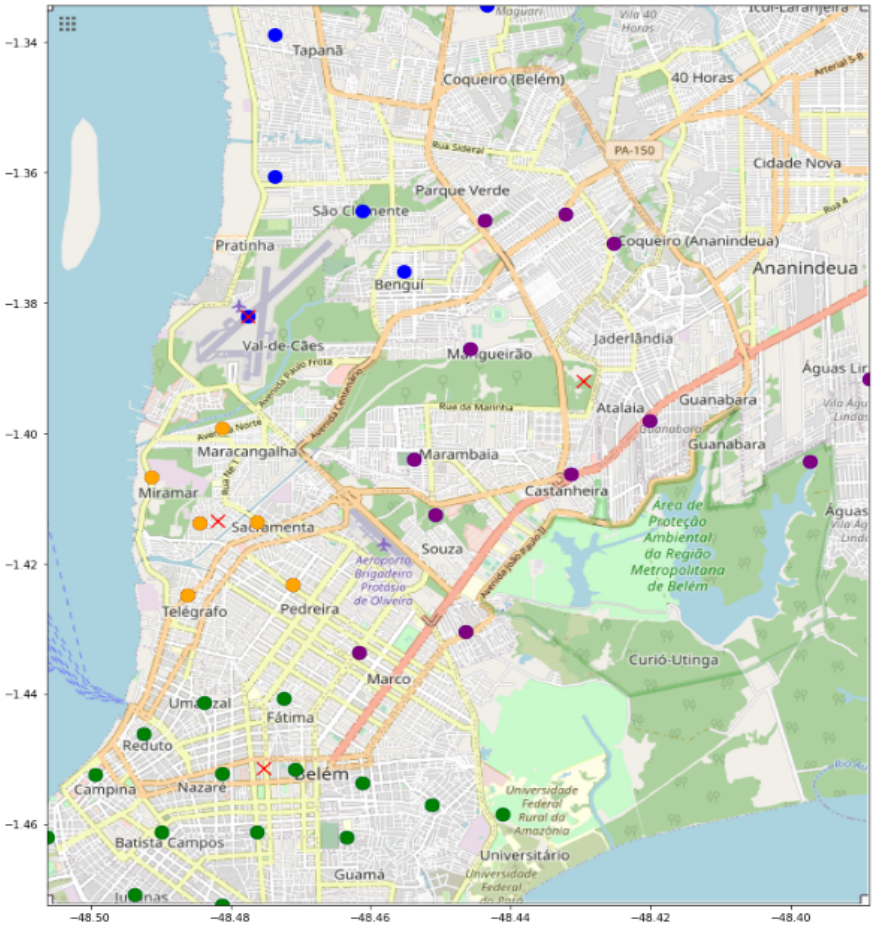

Figura 2. Plotagem das amostras e centroides do melhor resultado no mapa de Belém

As Figuras 2 e 3, são a saída da 'Scatter' plot feita com a ajuda do 'MatPlotLib', que recebe a imagem do mapa de Belém que é sincronizada junto as coordenadas dos bairros e dos centroides seguindo o agrupamento de classes realizado pelo algoritmo.

Nota-se que nas Figuras 2 e 3 a descrição de cada cluster e de cada um dos bairros que fazem parte dele é feita usando a própria posição no mapa das coordenadas no mapa identificado, assim como as cores que representam um agrupamento por cor de acordo com sua classe, ou seja, o cluster que o bairro pertence.

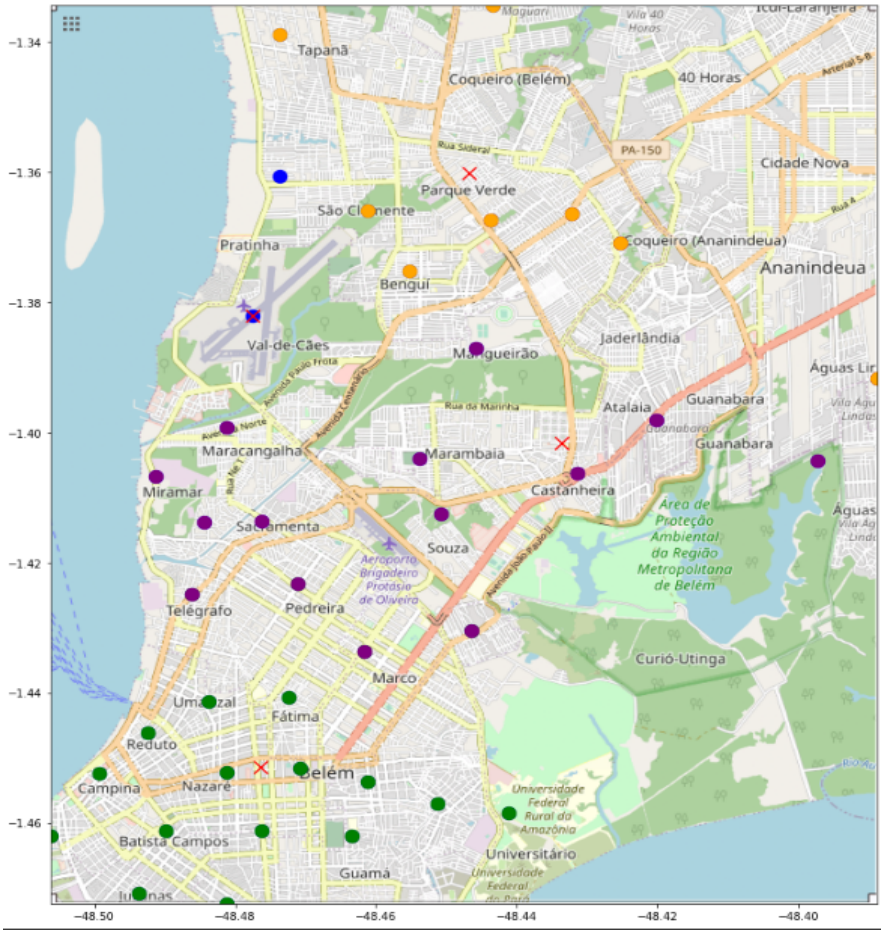

Figura 3. Plotagem das amostras e centroides do pior resultado no mapa de Belém

O melhor resultado representado na Figura 2 foi selecionado a partir da melhor distância, ou seja, a menor distância total entre todas as distâncias totais retornadas pelo K-means apresentadas na tabela da Figura 4, na qual a melhor clusterização foi a primeira (0) com a distância total de 8.036815546704123 unidades a partir soma das Distâncias Euclidianas.

O pior resultado representado na Figura 3 foi selecionado a partir da pior distância, ou seja, a maior distância total entre todas as distâncias totais retornadas pelo K-means apresentadas na tabela da Figura 4, na qual a pior clusterização foi execução de index 13 com a distância total de 8.728730351613885 unidades a partir soma das Distâncias Euclidianas.

A variação nas distâncias totais na Figura 4 exemplificam a natureza estocástica da clusterização feita pelo ”K-means”, uma vez que a posição inicial dos centroids (exceto o centroid de Val-de-Cans) é aleatória. Além disso, essas execuções desmonstram que o número de iterações igual ao número de amostras foi o suficiente para manter um equilíbrio entre a velocidade do algoritmo e a minimização da distância total.

0	Distancia Total,
0	8.036816
1	8.037930
2	8.541567
3	8.631628
4	8.624280
5	8.413000
6	8.373842
7	8.175760
8	8.358733
9	8.312362
10	8.158571
11	8.705563
12	8.221403
13	8.728730
14	8.393379
15	8.217563
16	8.324244
17	8.698079
18	8.172134
19	8.096716

Figura 4. Tabela com as distâncias totais por execução do algoritmo

5. Conclusão

O algoritmo de clusterização "K-means" foi bem executado de acordo com os padrões definidos na atividade em questão. E o teórico problema dos aeroportos em belém foi suprido através das diversas iterações da implementação afim de aprimorar os cenários até ser encontrado o melhor possível. O trabalho foi desenvolvido pelos integrantes fazendo uso das seguintes ferramentas de edição cooperativa em tempo real: Overleaf, para organizar um template SBC e atualizarmos os tópicos, um documento denominado "IAK-means02.ipynb" do Google Colaborative para as implementações, fonte do banco de dados e todos os testes antecessores a versão final do código-fonte em anexo junto a esse relatório técnico, e por fim a base de dados que foi atribuida através do link ao documento.csv no GitHub.

6. Anexos

Código-Fonte:`https://colab.research.google.com/drive/1yemLmNDQJTqbEgEUnvUL8rXpSntGAwgd#scrollTo=JQQanGKiwNmA`.

Fonte Base de Dados: `https://mapacep.com.br/index.php`.

.

Base de Dados Csv:https://gist.githubusercontent.com/hawkilol/9a5e64f50372346a35bb5098c59c7380/raw/7cdb42466b91704e474b9e7737ff22a7fd46e833/bairrosBelemNovo.csv

Referências

https://www.sbc.org.br/documentos-da-sbc/category/169-templates-para-artigos-e-capitulos-de-livros

https://www.semantix.com.br/escolhendo-o-algoritmo-de-clusterizacao-apropriado-para-seus-dados/

https://medium.com/programadores-ajudando-programadores/