Algoritmo build

- Entrada:
  - conjunto O de vetores de dados;
  - valor K.
- Saída:
  - conjunto S, contendo K amostras selecionadas como medoids;
  - conjunto U, contendo as amostras restantes.

1. Inicialize o conjunto S adicionando a ele uma amostra cuja soma das diferenças a todas as outras amostras é mínima.

(**OBS**: Este passo é exatamente o mesmo que o realizado na Tarefa 4 do Projeto 2). Porém, aconselhamos fortemente que seja usada uma matriz que armazene as distâncias entre todos os pares de amostras, pois isso agiliza significantemente o restante do algoritmo.

Como temos $|O|$ amostras, os resultados dos cálculos de distância devem ser armazenados em uma matrix $D_{|O|\times|O|}$ onde cada elemento na linha i e coluna j é uma distância d(i, j).

A distância a ser usada aqui é a mesma que a do Projeto 2, i.e., ℓ_1:

$$d_1(\vec{p}, \vec{q}) = ||\vec{p} - \vec{q}||_1 = \sum_{m=1}^{M} |p_m - q_m|,$$

onde M é a dimensionalidade dos vetores (no nosso caso, M=26). Para simplificar o restante desse texto, os vetores $\vec{p}$ e $\vec{q}$ são representados por identificadores de amostras, tais como i, j e h.

Dada a matriz de distâncias, gere um vetor que contém a soma de todos os elementos de cada linha p:

$$D_p = \sum_{h}^{|O|} D_{p,h}$$

e localize a linha j tal que essa soma é a mínima, i.e.:

$$j = \arg\min_p D_p$$

Para k=2⋯K:

a. Para cada amostra i do conjunto U, considere i como uma candidata para inclusão no conjunto S de amostras selecionadas.

   i. Para cada amostra j do conjunto U−{i},

      i. Compute D_j, definida pela distância entre j e a amostra mais próxima do conjunto S.

         Para tal, recomendamos o uso da matriz $D_{|O|\times|O|}$ sugerida para o passo anterior, examinando o índice da linha j e as colunas $i \in S$.

ii. Se $D_j > d(i,j)$, a amostra j vai contribuir positivamente para a decisão de selecionar a amostra i, pois nesse caso, j faria parte do agrupamento que tem i como medoid.

a. Compute o valor dessa contribuição da seguinte forma:

$$C_{i,j} = D_j d(i,j)$$

Senão, $C_{i,j}=0$

i.Calcule g_i, o ganho total de se adicionar i a S como sendo a soma de todas as contribuições relacionadas a i, ou seja

$$g_i = \sum_{j \in U} C_{i,j}$$

Escolha a amostra i que maximize g_i,

a.Remova i do conjunto U e adicione-a ao conjunto S.