Projeto Meantrix

Daniel Lima

15/08/2022

Contents

1 Introdução

Prospecção de empresas a partir dum arquivo csv com 21.299 empresas canadenses no Canadá seguindo os temas: *solutions on waste and water, Improve water quality and water efficiency use, water contamination, water for human consumption, water resources.*

O objetivo é digitar algum tema, e.g., *water resources*, e receber uma resposta com as informações das empresas correlacionadas a esse tema. As informações compreendem o aporte financeiro, número de empregados, descrição exata do foco da empresa, além das coordenadas de latitude e longitude e a cidade de localização da empresa.

2 Método

Os temas são escritos numa lista de Python,

```
target_list = ['solutions on waste and water','Improve water quality and water
               efficiency use', 'water contamination','water for human consumption',
               'water resources']
```

e, em seguida, é feita uma busca dos grupos de empresas que melhor se relacionem com cada tema. Nesse contexto, **a solução frontal** para essa tarefa é o algoritmo de aprendizado de máquina *Nearest Neighbor Search* (NN), porque, dado um input, o NN faz a busca pelos objetos na base de dados que mais se assemelham ao input no sentido de menor distância e/ou maior projeção num hiperplano de parâmetros. Esses parâmetros são as palavras selecionadas da base de dados cujas presenças (ausências) são quantificadas como 1 (0) nas descrições, viabilizando os algoritmos matemáticos.

Eu selecionei as 5.000 palavras com maior contagem e relevância – contagem acima de 2 e frequência não maior do que 0.95 –, assim, cada empresa terá associada a si um vetor de 5.000 dimensões relativo a contagem de palavras, o que foi cumprido através do método `TfidVectorize()` da biblioteca de processamento de linguagem natural nltk.

Finalmente, eu fitei o modelo NN, da biblioteca scikit-learn, com a opção de até 20 vizinhos e com a opção instrumental de cosseno, isto é, com a opção de projeção no hiperplano entre um objeto de pesquisa (um tema) e o modelo NN fitado.

3 Resultados

O programa exibe as 20 empresas cujas descrições mais se aproximam de cada tema. Por exemplo, os 5 primeiros resultados para o tema *water resources* é:

```
Tema de busca =  'water resources'

Distância KNN = 0.407. Neighbor idx =  14902
Nome da empresa: Tabl'eau water. Cidade: Toronto
----------------------------------------------------------------------
Distância KNN = 0.439. Neighbor idx =  387
Nome da empresa: Aqua air 247 . Cidade: Kelowna
----------------------------------------------------------------------
Distância KNN = 0.439. Neighbor idx =  11376
Nome da empresa: Aqua air 247. Cidade: Kelowna
----------------------------------------------------------------------
Distância KNN = 0.446. Neighbor idx =  5761
Nome da empresa: Clean wave products international. Cidade: Calgary
----------------------------------------------------------------------
Distância KNN = 0.481. Neighbor idx =  20120
Nome da empresa: Greenlife water. Cidade: Toronto
```

em que é fornecido o índice "Neighboor idx" que é o número da linha da base de dados csv correspondente às informações de uma empresa. Por exemplo, a quinta empresa mais distante do tema, a *Greenlife water*, possui a respectiva descrição registrada no DataFrame:

"Greenlife water is the leading provider of water filtration systems in the ontario area. we help make your water safer, healthier, and better tasting. with our point-of-entry system, you can enjoy filtered water throughout your home. we make water filtration systems become an affordable option for families of all sorts. we believe in making clean, healthy water accessible, affordable, and convenient for people across ontario. clean water should not remain a luxury. instead, people across the world should have affordable access to it. we're working hard to make that mission a reality.our company is not about a "cookie cutter" formula for water filtration. we believe that each family's unique water situation is different. that's why we offer a wide range of available filter types and services. no matter your situation, we offer a convenient and affordable solution that best serves your family. many of our customers opt for our filtration rental program. this convenient program brings all the benefits of our trusted filtration system without huge upfront expenses.we offer the following benefits: • no "installation fees" or other upfront costs. • lifetime service, filtration maintenance, and replacements for filters. • one convenient, low monthly rental payment. • full ease-of-use through integration with your enbridge gas distribution bill. • clean, pure water at the point-of-entry. this means water available for every tap in your home. • reductions from 98% to 99.9% for contaminants like chlorine, lead, arsenic, and mercury. • longer lasting appliances with less maintenance thanks to cleaner water. • outstanding water filtration without the loss of essential minerals in your water.from price to quality to environmental awareness, greenlife water is your best choice. why continue risking your family's health and safety? our filtered water is convenient, effective, and hassle-free. let a greenlife water filtration system keep you and your family healthy and happy. collapse"

1. **As principais cidades (polos de desenvolvimentos)**: inserindo os temas duma só vez da `target_list` fornecida e escolhendo as 20 cidades para cada tema, totalizando 100 cidades para a `target_list` traduzidas formatadas dentro dum DataFrama de Pandas de 100 respostas, após usar o método `value_counts()` na coluna de cidades do DataFrame, o resultado é o seguinte:

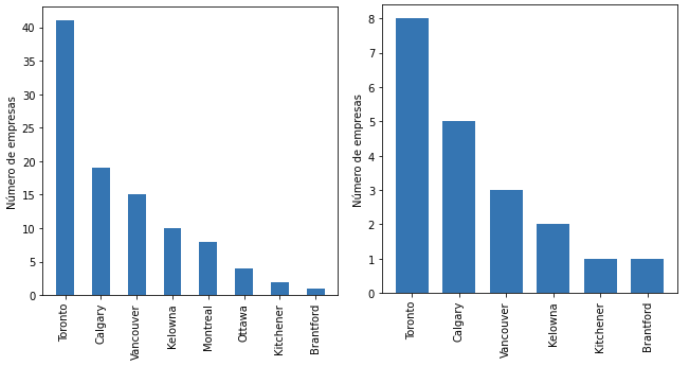

Figure 1: À esquerda, gráfico de barras do número de empresas em cada cidade resultante da pesquisa do tema *solutions on waste and water*, especificamente. À direita, o resultado de todos os temas em `target_list`.

Portanto, as cidades de **Toronto e Calgary despontam como polos**. As coordenadas de latitude e longitude discriminam somente a cidade das empresas, não a localização exata delas. Dessa maneira, não é possível depreender informações mais detalhadas sobre como as empresas estão distribuídas em bairros nas cidades e, então, um mapa das empresas utilizando a biblioteca `folium` não é revelador pois somente aponta as cidades.

2. **Distribuição do aporte financeiro das empresas**: infelizmente não existem dados o sufciente para uma análise acurada, porque a maior parte da vezes essa informação não é fornecida. Das 100 respostas, 84 são `Nan`: `df_respostas["aporte"].isna().sum() = 84` e 8 não são preenchidas. As 8 entradas restantes compreendem 4 empresas. A empresa com maior aporte de 5.034.285 de unidades monetárias (pode ser USD ou CAD) é a *Acuva technologies*, baseada em Vancouver, vindo logo em seguida a *Island water technologies*, baseada em Toronto, com um aporte de 690.742 unidades monetárias.

3. **Distribuição do número de funcionários por empresa**

A maior parte das empresas relacionadas a esses temas compreendem empresas com somente 1 funcionário.

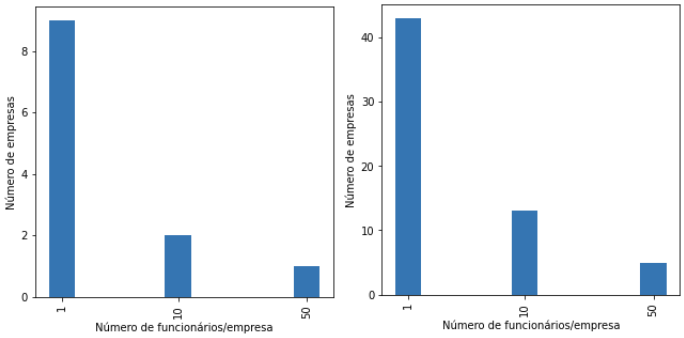

Figure 2: À esquerda, gráfico de barras do número de empregados por empresa da busca pelo tema *solutions on waste and water*, especificamente. À direita, gráfico de barras do resultado de todos os temas em `target_list`.