Aprendizado de Máquina em Textos

Identificando Idiomas com Machine Learning

Jorge Ashkar Ferreira Simondi — 8517081,
Leonardo de Almeida Lima Zanguetin — 8531866,
Victor Luiz da Silva Mariano Pereira — 8602444

Resumo

Reconhecer em qual idioma um texto está escrito, hoje em dia, é algo de suma importância. Apesar de ter várias ferramentas prontas na internet, nesse trabalho, mostraremos uma forma de como aplicar inteligência artificial para fazer essa classificação. Uma visão desde a análise dos dados até o treinamento para o aprendizado de máquina.

Introdução

A análise de textos em línguas diferentes pode ser feita de diversas maneiras, uma delas é em relação as palavras, a qual temos que verificar em qual dicionário está. Porém, uma outra forma de analisar é pela frequência de cada caractere, e essa é a abordagem que usaremos.

```
 Perceived end knowledge
certainly day sweetness why
cordially. Ask quick six seven
offer see among. Handsome
met debating sir dwelling age
material. As style lived he worse
dried. Offered related so visitor
we private removed. Moderate do
subjects to distance.
Of friendship on inhabiting
diminution discovered as. Did
friendly eat breeding building
few nor. Object he barton no
effect played valley afford.
Period so to oppose we little
seeing or branch. Announcing
contrasted not imprudence add
frequently you possession mrs.
Period saw his houses square and
misery.
```

Figura 1: Texto do Random Text Generator [1]

```
Ainda assim, existem dúvidas
a respeito de como a execução
dos pontos do programa facilita
a criação das direções
preferenciais no sentido do
progresso. Por outro lado, o
novo modelo estrutural aqui
preconizado cumpre um papel
essencial na formulação das novas
proposições.
O empenho em analisar o
consenso sobre a necessidade de
qualificação exige a precisão e
a definição dos relacionamentos
verticais entre as hierarquias.
Todas estas questões, devidamente
ponderadas, levantam dúvidas
sobre se a revolução dos costumes
deve passar por modificações a
longo prazo.
```

Figura 2: Texto do Lero Lero [2]

Nesses dois textos acima, só de olhar o computador não sabe em qual idioma é, mas se analisarmos os histogramas da quantidade de cada letra para esses dois textos, podemos ver que eles são diferentes. Isto acontece porque cada lingua tem letras que utilizam mais frequentemente do que outras. Nosso objetivo é descobrir como diferenciar os dois textos usando inteligência artificial, mais especificamente, *machine learning*.

Dados do problema

Os dados que utilizaremos são os caracteres do texto, tirando toda a pontuação e espaçamento. Apesar de parecer fazer parte de ruído, os caracteres especiais ajudam na predição de qual tipo de texto é, se é inglês ou em português.

A quantidade de cada caractere é uma informação relevante, mas não podendo ser utilizada crua, tendo que ser utilizada a porcentagem de ocorrência cada letra no texto.

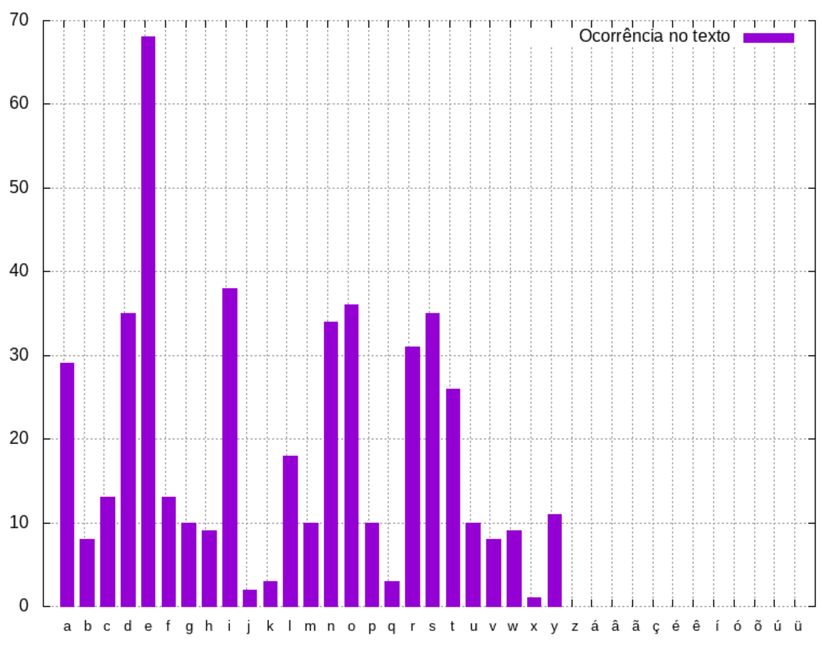

Figura 3: Histograma do texto em inglês

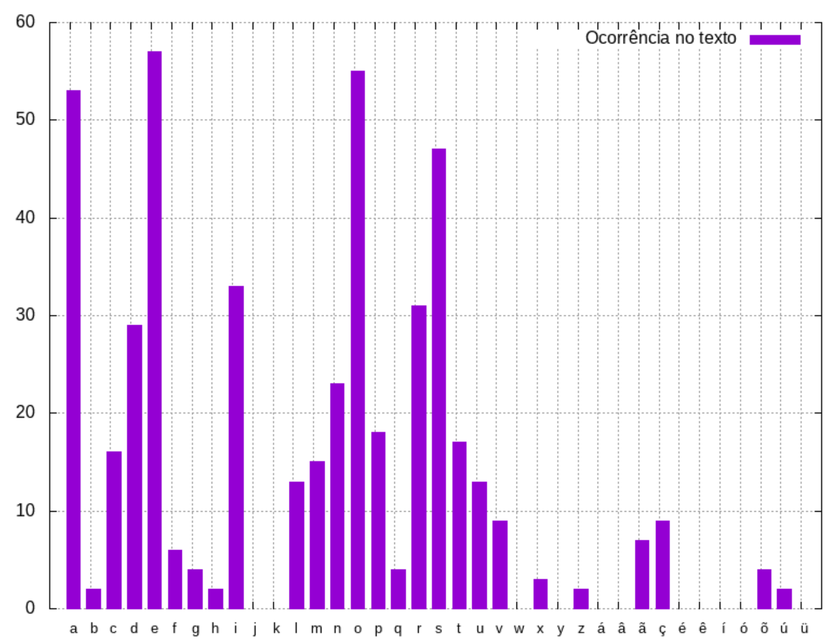

Figura 4: Histograma do texto em português

Método de classificação

Pela análise do nosso problema, consideramos utilizar **modelos preditivos** para poder classificar o texto com base em outros textos de entrada previamente classificados.

Modelo baseado em distância

A técnica que decidimos utilizar é baseado na **distância euclidiana** 1 entre o texto sem classificação e os textos da base de conhecimento. Como a distância envolvendo a quantidade de ocorrência poderia dar uma falsa informação, usamos a porcentagem que cada letra é encontrada no texto.

$$d(x_i, x_j) = \sqrt{\sum_{l=1}^{d} (x_i^l - x_j^l)^2} \tag{1}$$

Como a distância para o mais próximo pode não ser tão precisa, a utilização do algoritmo *k-NN*, do inglês *k-Nearest Neighbour* se fez necessária, assim calculamos a distância para cada texto e fazemos uma média dos n primeros.

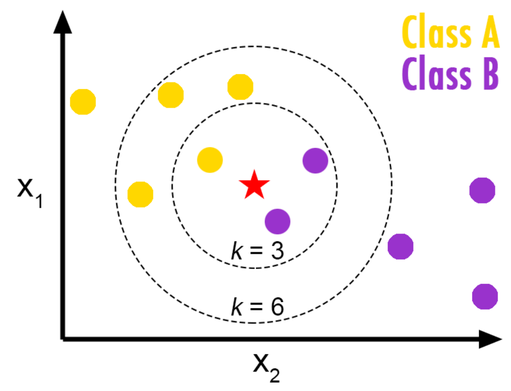

Figura 5: Representação do *k-NN*

Conclusão

Sabemos que o método que escolhemos para a identificação pode ser um pouco custoso para uma base de dados muito grande, mas a predição que fazemos é importante para casos em que se utiliza redes neurais ou outras teorias mais avançadas de Inteligência Artificial. Assim, temos uma idéia da potencialização que algoritmos baseados em *Machine Learning* fornecem para um projeto, maximizando a eficiência e a velocidade em que diversos algoritmos são executados.

Referências

[1] Vangelis Bibakis. Random text generator. `http://www.randomtextgenerator.com/`.

[2] Miguel Borges. O fabuloso gerador de lero lero v3. `http://lerolero.miguelborges.com/`.

[3] Katti Faceli, Ana Carolina Lorena, João Gama, and André Carlos Ponce de Leon Ferreira Carvalho. *Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina*. LTC, 1 edition, 2011.