1. INTRODUÇÃO

Neste artigo, avaliamos a eficácia de algoritmos de aprendizagem de máquina na classificação de perguntas de acordo com seus assuntos.

A importância disso surge quando um usuário tenta recuperar informações em documentos, pois, dada a potencial grandiosidade da tarefa e o tédio de ter de analisar inúmeros possíveis documentos, delegar essa atividade a um algoritmo de inteligência artificial acaba sendo a melhor alternativa. Além disso, esse tipo de algoritmo também é de grande valor na criação de chatbots, para que possam identificar sobre o que está sendo perguntado por um usuário.

No TREC (Text Retrieval Conference), um número de perguntas e respostas tentam responder uma lista de perguntas predefinidas a partir do uso de um conjunto de documentos predeterminado.

1. MATERIAIS

Utilizaremos a base de dados organizada por Xin Li e Dan Roth, que foi utilizada em seu artigo Learning Question Classifiers (2002).

Usaremos três tipos de pré-processamento para cada tipo de algoritmo escolhido. Serão eles o BOW (Bag of Words), o TF (Term Frequency) e o TF-IDF (Term Frequency – Inverse Document Frequency).

1. MÉTODOS

Os algoritmos utilizados serão o KNN (K-Nearest Neighbors – modelo do vizinho mais próximo) e o Naive Bayes.

3.1. KNN

Para classificarmos conforme o KNN, primeiro encontramos *f*(*k*, *xq*) e tomamos o voto da maioria dos vizinhos (que é o voto majoritário, em caso de classificação binária). Para evitar empates, *k* é sempre escolhido como número ímpar. Para fazer regressão, podemos tirar a média ou mediana de *k* vizinhos, ou podemos resolver um problema de regressão linear sobre os vizinhos (Russel e Norvig, 2004).

A função *f* é definida assim:

*f*(*xq*) = arg max *i*))

*v* *V*

3.2. Naive Bayes

1. EXPERIMENTOS E RESULTADOS

4.1. KNN

|  |  |  |
| --- | --- | --- |
| **Método de Pré-Processamento** | **k** | **Acurácia** |
| BOW | 4 | 0,59 |
| Tf | 4 | 0,702 |
| Tfidf | 12 | 0,73 |

4.2. Naive Bayes

1. CONCLUSÃO