# Classifiação de livros com base em sua descrição

Gustavo Saad - RA: 10332747

Neste artigo, vamos explorar o desenvolvimento de um projeto de classificação de livros com base nas suas descrições, com o objetivo de identificar automaticamente o gênero literário de uma obra. O projeto utiliza técnicas de aprendizado de máquina, especificamente a **classificação de texto**. Através desse processo, procuramos construir um modelo de Inteligência Artificial (IA) capaz de categorizar livros em gêneros como **Aventura**, **Ficção Científica**, **Mistério**, **Romance**, **Thriller**, entre outros.

## Objetivo do Projeto

O principal objetivo deste projeto é criar um sistema capaz de classificar livros em diferentes categorias literárias, com base em suas descrições. Usando um modelo de IA, o sistema deverá ser capaz de prever corretamente o gênero literário de um livro a partir da análise do seu resumo ou descrição. Com isso, é possível automatizar a organização de grandes catálogos de livros, facilitando a pesquisa e organização de bibliotecas digitais ou plataformas de leitura.

## Construção da Base de Dados

Para treinar o modelo de IA, é necessário uma base de dados robusta, que contenha exemplos suficientes de livros com suas respectivas descrições e gêneros. A base de dados foi criada utilizando informações de livros públicos disponíveis na **Google Books API**. Através dessa API, foi possível coletar uma variedade de dados, incluindo o título, categoria (gênero literário) e a descrição do livro.

Inicialmente, a coleta está focada em livros de **Ficção**, para cobrir uma gama de gêneros populares, como **Aventura**, **Ficção Científica**, **Mistério**, **Romance**, **Thriller**, **Horror**, e outros. Com esses dados, o objetivo é treinar um modelo que consiga, dado o resumo ou a descrição de um livro, prever com precisão qual gênero ele pertence.

## Processo de Coleta de Dados

A coleta de dados foi realizada a partir de uma integração com a **Google Books API**. O primeiro passo foi acessar a API, realizar uma busca pelos gêneros literários definidos e armazenar informações detalhadas sobre os livros.

Um ponto interessante durante a coleta foi a remoção de livros com descrições vazias ou `None`, que não poderiam ser usados para treinar o modelo. Além disso, foi realizada uma limpeza nos

dados para garantir que o texto estivesse no formato adequado, sem caracteres indesejados ou erros de codificação que poderiam prejudicar a análise.

## Método de Inteligência Artificial Utilizado

O modelo de IA utilizado para realizar a classificação de gêneros literários é baseado em **Aprendizado Supervisionado**, um dos principais métodos de aprendizado de máquina. Neste tipo de aprendizado, o modelo é treinado usando um conjunto de dados rotulado, ou seja, um conjunto em que as descrições dos livros já são associadas aos seus gêneros literários.

Para transformar as descrições de texto em dados que possam ser processados pela IA, utilizamos a técnica de **TF-IDF (Term Frequency-Inverse Document Frequency)**. O TF-IDF é uma técnica de vetorização de texto que converte palavras em números, atribuindo maior peso às palavras mais importantes de cada descrição, ou seja, palavras que são frequentes em um único documento (livro), mas raras em outros. Isso ajuda o modelo a identificar as palavras-chave que caracterizam cada gênero literário.

Após a vetorização, utilizamos um algoritmo de **Regressão Logística**, que é eficiente para tarefas de classificação de texto. A regressão logística foi escolhida devido à sua simplicidade e eficácia em problemas de classificação binária ou multi-classe, como é o caso deste projeto, onde temos várias classes de gêneros literários.

## Treinamento e Avaliação do Modelo

Após preparar os dados e configurar o modelo de IA, dividimos o conjunto de dados em **conjunto de treino** e **conjunto de teste**. O conjunto de treino é utilizado para ensinar o modelo a identificar padrões nas descrições dos livros e associá-los aos gêneros. Já o conjunto de teste é usado para avaliar a capacidade do modelo de generalizar, ou seja, prever corretamente o gênero de livros que ele nunca viu antes.

A avaliação do modelo foi realizada utilizando métricas como **precisão**, **recall** e **F1-score**, que nos ajudam a medir a qualidade da classificação em termos de quantas vezes o modelo acertou o gênero correto, quantos gêneros ele conseguiu identificar corretamente, e o equilíbrio entre esses acertos.

## Desafios e Melhorias Futuras

Durante o desenvolvimento do projeto, enfrentamos alguns desafios. A base de dados inicialmente coletada estava desequilibrada em relação aos gêneros, com uma maior concentração de livros em determinados gêneros literários. Esse desequilíbrio pode afetar a precisão do modelo em categorias menos representadas. Para resolver isso, estratégias como **balanceamento de classes** e **aumento de dados** podem ser aplicadas, o que implica gerar mais exemplos para as classes minoritárias.

Além disso, o modelo pode ser aprimorado utilizando técnicas mais avançadas de aprendizado de máquina, como **redes neurais** e **modelos baseados em BERT** (Bidirectional Encoder Representations from Transformers), que são conhecidos por sua eficácia em tarefas de processamento de linguagem natural.

## Conclusão

Este projeto demonstra como a Inteligência Artificial pode ser aplicada para automatizar a classificação de livros com base nas suas descrições. Utilizando técnicas de aprendizado de máquina, foi possível treinar um modelo capaz de categorizar livros em diferentes gêneros literários com um nível aceitavel de precisão. Com o uso de bases de dados como a Google Books API e a aplicação de técnicas de vetorização como o TF-IDF, conseguimos transformar dados de texto em informações valiosas para o treinamento do modelo.

Futuras melhorias podem incluir a expansão da base de dados, a adição de novos gêneros literários, o melhor tratamento dos dados e o aprimoramento do modelo com técnicas mais avançadas (tecnicas alternativas de ML). O objetivo final é criar uma ferramenta eficiente para classificar automaticamente grandes coleções de livros, facilitando a organização de bibliotecas digitais e plataformas de leitura.

## Referencias

https://pt.wikipedia.org/wiki/Tf%E2%80%93idf
https://huggingface.co/tasks/text-classification
https://scikit-learn.org/1.5/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html