Question Duplicate Detection

This repository contains a comprehensive solution for detecting duplicate questions using machine learning techniques. The script preprocesses, engineers features, and builds models to classify pairs of questions as duplicates or not.

Table of Contents

Overview

The main objective of this project is to identify whether pairs of questions are duplicates. The script connects to a MySQL database to fetch data, preprocesses the questions, engineers various features, and builds machine learning models to make predictions.

Requirements

- Python 3.x
- MySQL database
- Python libraries:
  - numpy
  - pandas
  - matplotlib
  - seaborn
  - BeautifulSoup
  - mysql-connector-python
  - scikit-learn
  - xgboost
  - nltk
  - fuzzywuzzy
  - plotly

Installation

1. Clone the repository:

   ```
   git clone https://github.com/yourusername/question-duplicate-detection.git
   ```

2. Navigate to the project directory:

```
cd question-duplicate-detection
```

3. Install the required libraries:

```
pip install -r requirements.txt
```

Usage

1. Configure the MySQL database connection in the script:

```
python
Copy code
host = '127.0.0.1'
user = 'root'
password = 'yourpassword'
database = 'train'
table_name = 'questions'
```

2. Run the script:

```
python duplicate_detection.py
```

Feature Engineering

The script creates several features to enhance model performance:

- **Token Features**: Counts of common words and tokens, and their ratios.
- **Length Features**: Length of each question and the difference in lengths.
- **Fuzzy Features**: Fuzzy matching scores using fuzzywuzzy library.

Models

The script builds and evaluates two models:

- **Random Forest Classifier**
- **XGBoost Classifier**

Both models are trained on the engineered features and evaluated using accuracy and confusion matrices.

Results

The performance of the models is printed in terms of accuracy and confusion matrices. This allows for comparison and selection of the best performing model.

Contributing

Contributions are welcome! Please open an issue or submit a pull request for any improvements or bug fixes.