Documentation of TrialMatcher

# Overview

This documentation provides a detailed overview of the TrialMatcher system, starting from all aspects from data preprocessing to patient-to-trial matching, and output generation. It explains the approach, implementation details, testing, and how to run the system step-by-step.

# Data Preprocessing

The data preprocessing step is essential for ensuring that patient and trial data are clean and in the correct format. The script "data\_preprocessing.py" performs the following key tasks:

1. Loading patient and trial data from CSV files.

2. Handling missing values and formatting columns such as "AGE", "DESCRIPTION\_x" (medical conditions), and "DESCRIPTION\_y" (medications) in patient data.

3. Ensuring that trial data columns like "minimum\_age", "maximum\_age", and "exclusion\_criteria" are properly formatted.

4. Converting the date of birth of the patients in years with respect to the current calendar.

5. Saving the cleaned data to new CSV files for use in the matching process.

# Patient Matching

The core functionality of the system is the patient matching process, implemented in "patient\_matching.py". This script performs the following steps:

1. Prompts the user for patient ID as input.

2. Matches the patient to eligible clinical trials by checking two main criteria:

- Age-based inclusion criteria: The patient’s age must fall within the trial’s defined minimum and maximum age range.

- Exclusion criteria: The patient’s medical conditions and medications are compared to the trial’s exclusion criteria. If there is any match between the patient’s conditions/medications and the exclusion criteria, the patient is excluded from the trial.

3. Summarizes the patient’s history (conditions and medications).

4. Generates an explanation for why the patient is eligible for each matched trial.

5. Outputs the results in JSON format, including patient details, trial matches, and explanations.

# Output Example

An example of the output generated by the system is a JSON file that contains:

1. The patient’s ID and a summary of their medical history.

2. A list of eligible trials along with the criteria they met.

3. A detailed explanation for why the patient is eligible for each trial.

# Testing and Validation

Unit tests and integration tests are included in the system to validate the individual components and the entire workflow. These tests ensure that both patient-to-trial matching and data preprocessing are functioning correctly.

Key components tested include:

- Extracting keywords from patient conditions and medications for exclusion criteria matching.

- Age range validation to ensure patients fall within the correct age range for trials.

- Summarizing patient history and generating match explanations.

Integration testing is also performed using sample patient and trial data.

# Conclusion

The TrialMatcher System provides an automated solution to match patients to relevant clinical trials. The system uses both age and exclusion criteria to ensure that patients are only matched to trials for which they are eligible. The system also generates detailed explanations and summaries of patient history, making the output both useful and informative. This system can be further extended with additional features, such as including more criteria or integration with additional data sources.