# Entity Matching System Design with Code Explanation

## 1. Data Loading and Setup

This part imports necessary libraries and loads the dataset from an Excel file. The dataset should contain columns like Name1, Name2, and Label. Here's the code:

import pandas as pd  
from sklearn.metrics import precision\_score, recall\_score  
from sklearn.metrics.pairwise import cosine\_similarity  
from sklearn.feature\_extraction.text import CountVectorizer  
from fuzzywuzzy import fuzz  
from metaphone import doublemetaphone  
  
file\_path = 'Mathwizzathon\_Entity\_Matching\_Dataset.xlsx' # Replace with your Excel file path  
data = pd.read\_excel(file\_path)

Explanation:  
This code segment loads necessary libraries for processing, similarity computation, and phonetic encoding. The dataset is loaded from an Excel file, which will be used for the matching process.

## 2. Preprocessing Functions

Normalization and Title Removal: Convert names to lowercase, remove special characters, and strip common honorifics. This standardization step improves matching accuracy.

def normalize\_name(name):  
 return str(name).lower().replace(".", "").replace("-", "").replace(",", "").strip()  
  
def remove\_titles(name):  
 titles = {'dr', 'mr', 'mrs', 'ms', 'ph.d.', 'professor', 'sir', 'lady', 'captain', 'major'}  
 parts = normalize\_name(name).split()  
 return ' '.join(part for part in parts if part not in titles)

Explanation:  
The normalize\_name function removes punctuation and converts the name to lowercase. The remove\_titles function removes common honorifics like 'Dr.' or 'Mr.' to help standardize the names and improve the accuracy of matching.

## 3. Similarity Functions

Initial Matching: Compares the initials of each name part to check for exact matches.

def extract\_initials(name):  
 parts = name.split()  
 return [part[0] for part in parts]  
  
def initials\_match(name1, name2):  
 initials1 = set(extract\_initials(name1))  
 initials2 = set(extract\_initials(name2))  
 return initials1 == initials2

Explanation:  
The initials\_match function extracts the first letters of each name part and checks for exact matches. This is useful for identifying initials-based matches, such as 'John D.' matching 'J. Doe'.