Description

In this question, you are required to do fuzzy matching and link two sets of names: the acquirers in bank M&A activities in 2015, and all bank holding companies in 2015. It is common that two datasets contain the same set of identities, such as firm names, but they are not exactly matched with each other. Examples like Chemical Financial in one dataset and Chemical Financial Corporation in the other make data merging difficult. Fuzzy matching gives any pair of strings a similarity score. Therefore, for any acquirer string, we will select a number of bank holding companies with highest similarity scores.

There are mainly 4 kinds of similarity scores: fuzz.ratio, fuzz.partial\_ratio, fuzz.partial\_token\_sort\_raio, and partial\_token\_set\_ratio. In this question, we adopt a naive metric, that is the sum of the 4 ratios, as the similarity score. Since fuzzy matching is not perfect, we select 10 candidate bank holding companies with highest scores for each acquirer and append them along with the acquirers. Usually we have to manually select the best match among these candidates as post-processing, but this is not the main concern of this question and you don't have to do it.

Recommended Python Packages

fuzzywuzzy (process)

Useful Links

<https://github.com/seatgeek/fuzzywuzzy>

Input

acquirers.xlsx

bank\_names.csv

Output

output.csv

Sample Input

acquirers bank\_names

Chemical Financial INC. HARTLAND FINANCIAL, INC.

CHEMICAL FINANCIAL CORPORATION

CHOICEONE FINANCIAL SERVICES

Sample Output

acquirers 0 1 2 ...

Chemical Financial CHEMICAL FINANCIAL CORPORATION CHOICEONE... ...