CS3723 Pgm 6 Python (40 pts)

This is a continuation of program #5. In this program, you will compare addresses to determine how close they match. This isn't just a simple comparison.

For each customer:

* Show the customer.
* In program #5, you retrieved multiple addresses for a customer. Each address should be represented by a dictionary, and you should have a list of addresses.
* For each address:
  + Show the address with a sequence number. (This was done in program #5.)
  + Break the street line (which was concatenated from multiple LINE commands in program #5) into its parts (streetNr, streetName, streetType, direction, apartmentNumber). This part of the assignment requires *thinking*.
  + Show the parts of a street line.
  + Show pairs of addresses (referencing the sequence number) and provide a score (0..100) of how closely these match.

Your business people have provided this guidance:

1. Your code should recognize equivalence of common abbreviations. It is probably best to map them to the longer value.

Street types (not an exhaustive list):

* RD, ROAD
* AVE, AVENUE
* ST, STREET

Directions (not an exhaustive list):

* SOUTH, S (when appropriate)
* SW, SOUTHWEST, SOUTH WEST (when appropriate)

2. There is usually at least one word of a street name before a street type, assuming the street type exists. (e.g., 123 E ST)

3. The direction (e.g., SOUTH) might occur after the street number or after the street type.

**Example addresses including name, street address, city, state, and zip**

BOB WIRE

1 123 DIRT RD

SAN ANTONIO, TX 78210

2 123 DIRT

SAN ANTONIO, TX 78210

3 123 DIRT LN

SAN ANTONIO, TX 78210

PENNY LOAFER

1 111 SHOE LN

SAN ANTONIO, TX 78249-1234

2 111 SHOE ST

SAN ANTONIO, TX 78249-1234

3 111 BOOT ST SOUTH APT #5A

SAN ANTONIO, TX 78230

4 111 S BOOT STREET NR 5A

SAN ANTONIO, TX 78230

FLO N WATER

1 45 S.W. VISTA RIO GRANDE RD

SAN ANTONIO, TX 78210

2 45 S WEST VISTA RIO GRANDE RD

SAN ANTONIO, TX 78210

3 45 SOUTHWEST VISTA RIO GRANDE ROAD

SAN ANTONOI, TX 78210

HOLLY WOOD

1 12 WEST AVE APT 23D

LOS ANGELES, CA 90009

2 12 WEST AVENUE #23D

LOS ANGELES, CA 90009

3 12 WEST A ST NR 23

LOS ANGELES, CA 90009

BILL MELATER

1 36A E COMMERCE ST #5

LOS ANGLES, CA 90009-2312

2 36-A COMMERCE ST EAST APT. NR. 5

LOS ANGELES, CA 90009

3 36A COMMERCE APT 5

LOS ANGELES, CA 90009

4 36A E COMMERCE ST #8

LOS ANGELES, CA 90009

5 36A EAST COMMEREC APT. 5

LOS ANGELES, CA 9009

Scoring Matches:

|  |  |  |
| --- | --- | --- |
| Seq | Information | Scoring |
| 1 | Street number | +20 both not empty and they match  -20 both have values, but they don't match  0 both empty  -20 only one empty |
| 2 | Street Type | +10 both not empty and they match  -10 both have values, but they don't match  +10 both empty  +5 only one empty |
| 3 | Direction | +5 both not empty and they match  -10 both have values, but they don't match  +5 both empty  -5 only one empty |
| 4 | AptNum | +20 both not empty and they match  -20 both have values, but they don't match  +10 both empty  -10 only one empty  5 mostly match (scale) where both have values |
| 5 | City | +20 both not empty and they match  -20 both have values, but they don't match  +10 both empty  -10 only one empty  15 mostly match (scale) where both have values |
| 6 | State | +10 both not empty and they match  -20 both have values, but they don't match  0 both empty  0 only one empty |
| 7 | Street Name | +20 both not empty and they match  -5 both have values, but they don't match  -20 both empty  -20 only one empty  +10 mostly match (scale) where both have values |
| 8 | Zip Code | +80 both not empty, both len of 10, and they match (if the ZIP+4 values match, these are most likely the same address)  +5 both not empty, both len of 5, and they match  +0 same len, but they don't match. Note that zip codes change a lot without all the addresses being updated  +5 lens not equal, the first 5 characters match  +0 lens not equal, the first 5 characters don't match. |

Notes:

1. If the computed score is less than 0, set it to 0. If it is greater than 100, set it to 100.

2. For scoring where mostly match is provided:

* Only use when the values are both not empty and they do not match.
* Use the SequenceMatcher's ratio to see if they mostly match.
* If the ratio is greater than 0.6, add the ratio \* the value in the table to the score.
* Otherwise, score using the both have values, but they do not match value.

3. Apartment numbers come in a variety of formats. They may be preceded with APT, NR, and/or #. Frequently those have periods and run into the actual apartment number (which isn't necessarily a number).

4. Some punctuation isn't valuable.

5. Initially test with the data from program #5, adding the data in the examples above. You will be provided with additional data later.