INFS 3200 : Practice Three

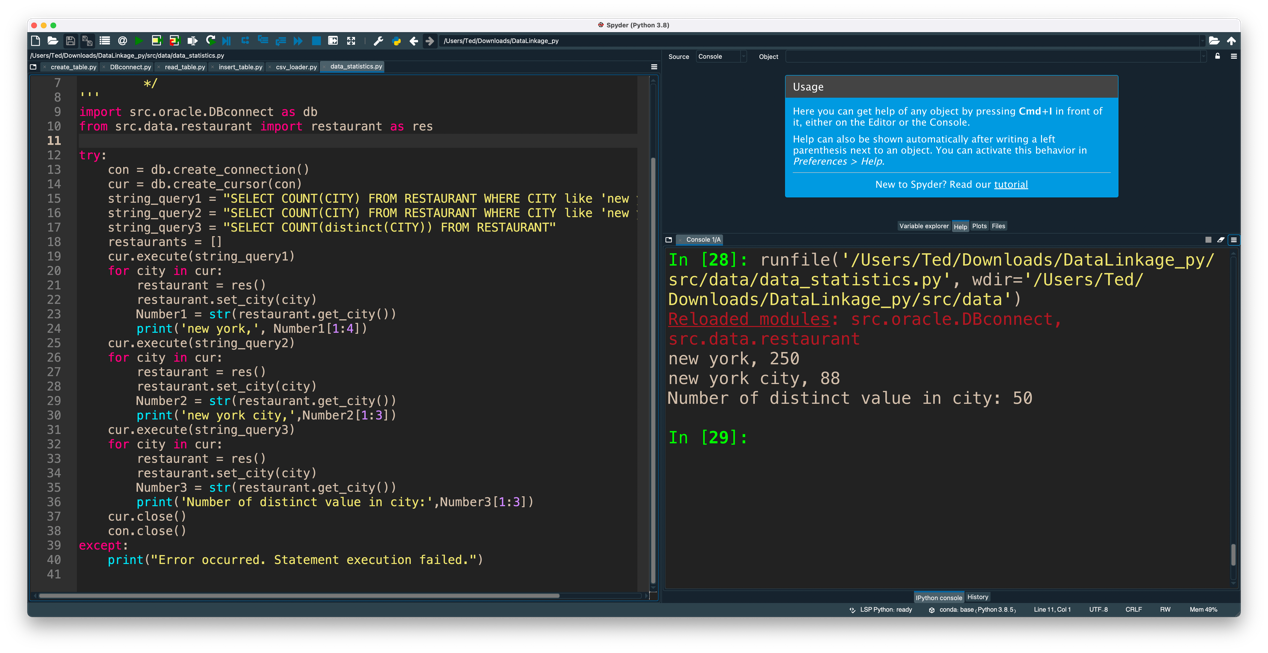
Name : Peng Yu

Student ID : 46635884

Task 1

Answer:

|  |  |
| --- | --- |
| City | Count |
| “new york” | 250 |
| “new york city” | 88 |
| Number of distinct value in city | 50 |



Task 2

Answer (1):

(1) If an instance is a positive class and the connection result is a positive class, it is a true class (True Positive TP)

(2) If an instance is a positive class, but the connection result is a negative class, it is a false negative type (False Negative FN)

(3) If an instance is a negative class, but the connection result is a positive class, it is a false positive class (False Positive FP)

(4) If an instance is a negative class, but the connection result is a negative class, it is a true negative class (True Negative TN)

TP: Number of correct connection results

FN: false negatives, the link result did not find the number of correct matches

FP: False positive, the result of the link is incorrect/mismatch

TN: Number of non--links correctly rejected

Answer (2):

Precision: The formula of precision is P = TP / (TP + FP), which calculates that all "correctly retrieved items (TP)" account for all "actually retrieved items (TP + FP)" The ratio. In this case, it represents the ratio of the correct and misjudged restaurants that are correctly retrieved. Simply put, precision means how correct the link is matched. which can be known from the code, and the calculation method is precision = count / len(results)-- > Total length of matching results.

Recall: The formula for the recall is R = TP / (TP + FN), which calculates that all "correctly retrieved items (TP)" account for all "items that should be retrieved (TP+FN)" proportion. Simply put, Recall means how comprehensive the matching link is. In this case, it represents the proportion of restaurants that want to be retrieved correctly in all restaurants that are retrieved correctly.

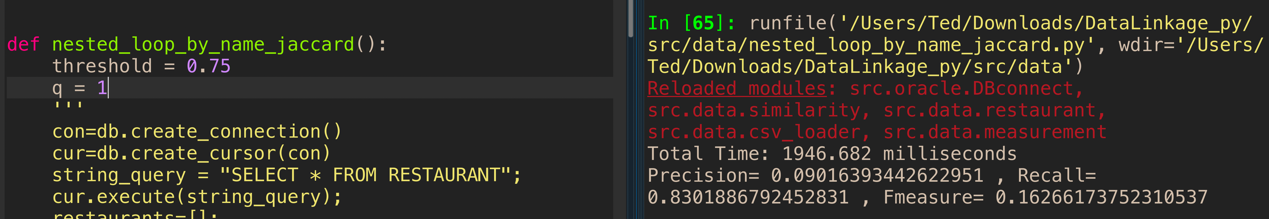
Task 3

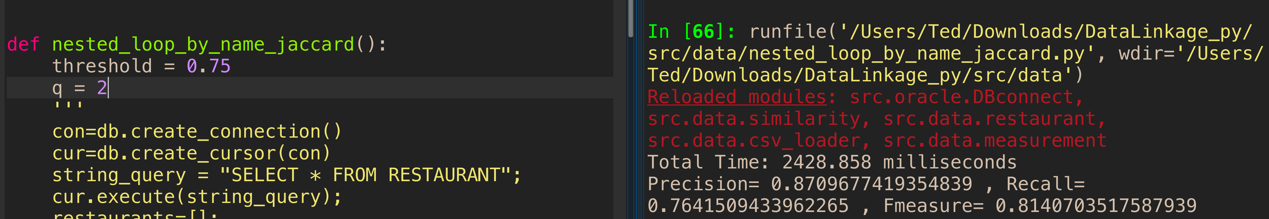
The default value of q and threshold is 3 and 0.75.

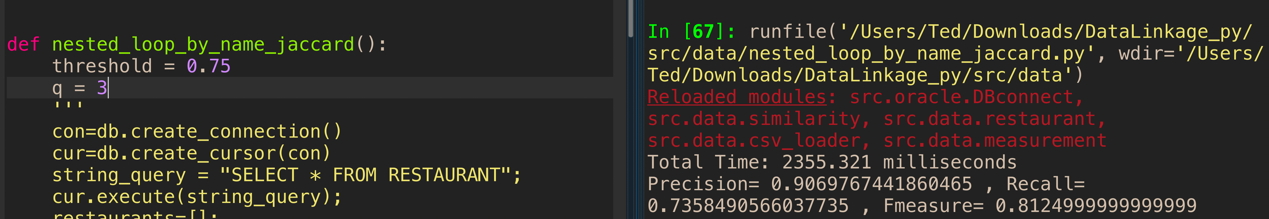
First I choice 5 values of q which is 0 1 2 3 4, keep threshold as 0.75 and the result show below:

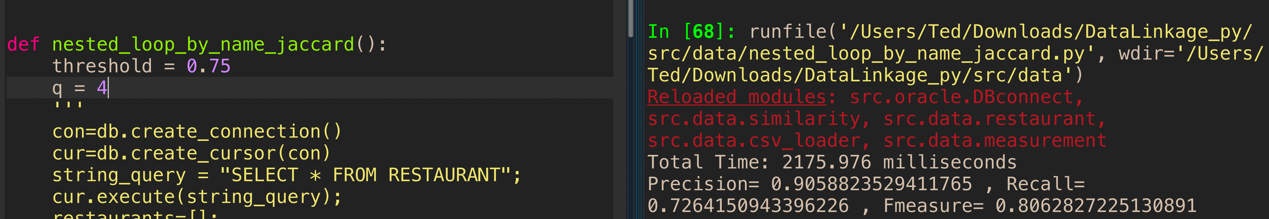
|  |  |  |  |
| --- | --- | --- | --- |
| q | Precision | Recall | Fmeasure |
| 1 | 0.0901639344262295 | 0.8301886792452831 | 0.1622661737523105 |
| 2 | 0.8709677419354839 | 0.7641509433962265 | 0.8140703517587939 |
| 3 | 0.9069767441860465 | 0.7358490566037735 | 0.8124999999999999 |
| 4 | 0.9058823529411765 | 0.7264150943396226 | 0.8062827225130891 |
| 5 | 0.9058823529411765 | 0.7264150943396226 | 0.8062827225130891 |

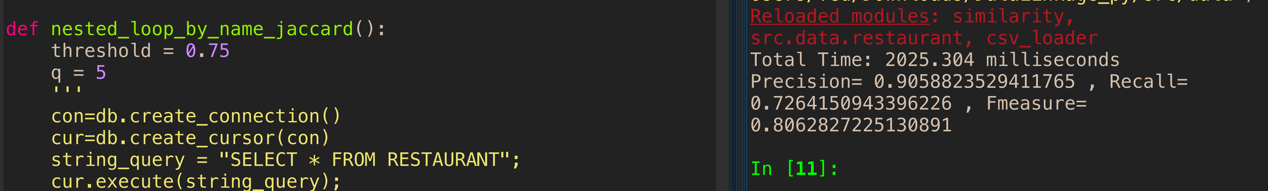
As q increases, the Precision gradually increases, reaches a maximum when q=3, and then gradually decreases as q increases again. When q increases, the Recall decreases. When q = 2, Fmeasure has the best result.









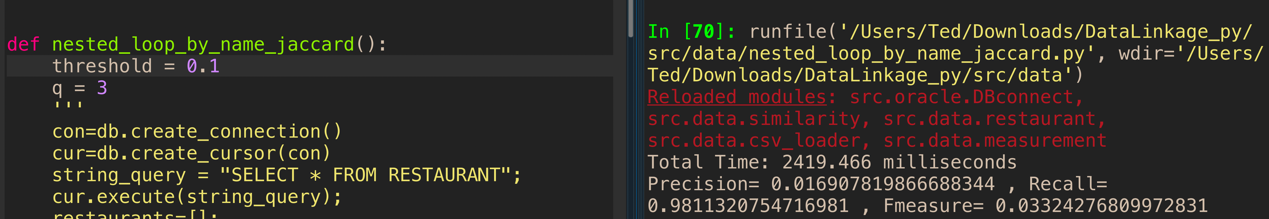


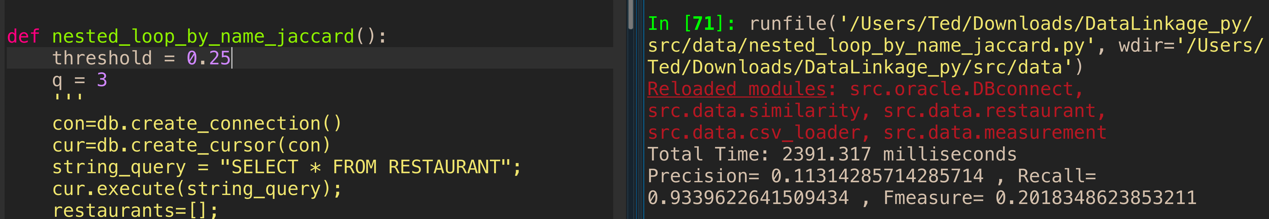
Second I keep q value as 3, and change five threshold value which is : 0.10, 0.25, 0.50, 0.70, 0.90 and the result show below:

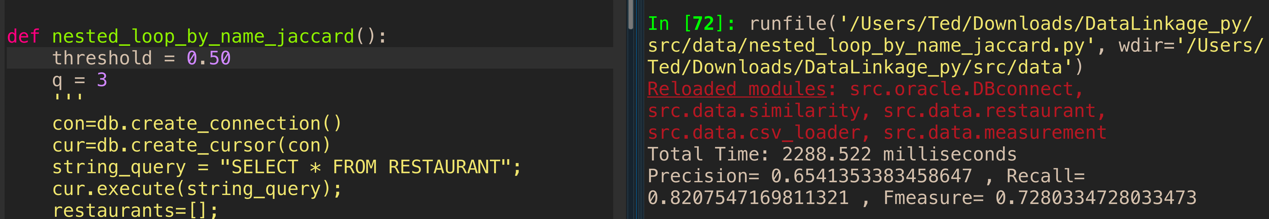
|  |  |  |  |
| --- | --- | --- | --- |
| threshold | Precision | Recall | Fmeasure |
| 0.10 | 0.01690781986668834 | 0.981132075471698 | 0.0332427680997283 |
| 0.25 | 0.11314285714285714 | 0.933962264150943 | 0.2018348623853211 |
| 0.50 | 0.6541353383458647 | 0.820754716981132 | 0.7280334728033473 |
| 0.70 | 0.8764044943820225 | 0.735849056603773 | 0.8 |
| 0.90 | 0.9156626506024096 | 0.716981132075471 | 0.8042328042328042 |

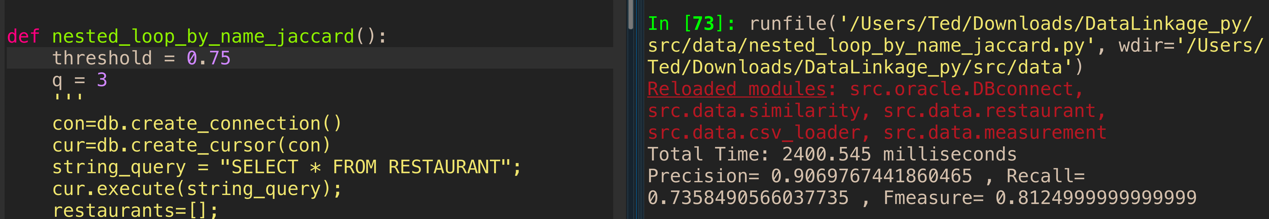
As the threshold increases, Precision increases, Recall decreases. When threshold = 0.90, Fmeasure has the best result.

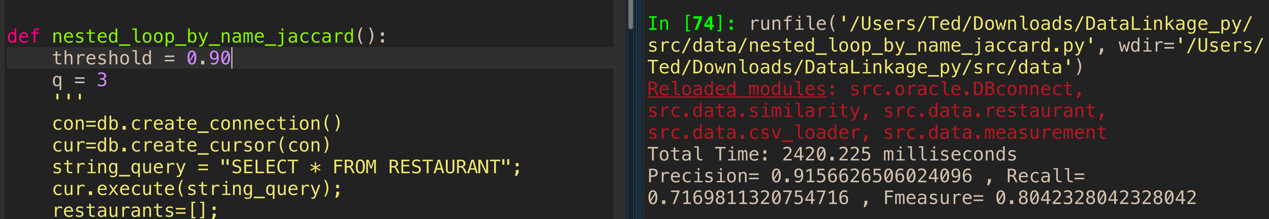
In summary, when threshold = 0.90 and q = 2, the predicted result is the best.





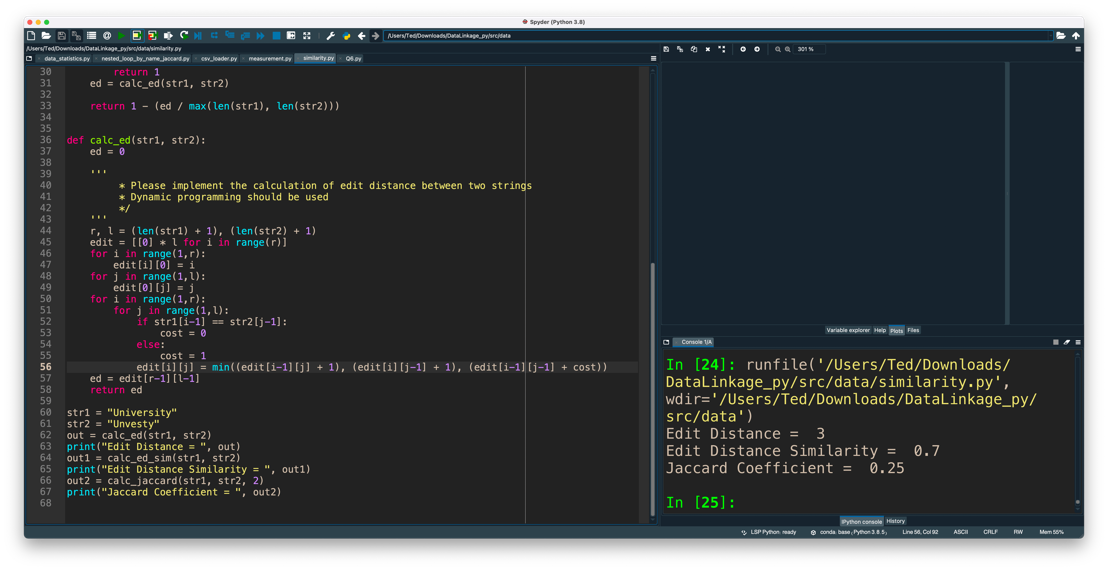






Task 4

|  |  |
| --- | --- |
| Edit Distance | 3 |
| Edit Distance Similarity | 0.7 |
| Jaccard Coefficient | 0.25 |



I change five threshold value which is : 0.05, 0.20, 0.45, 0.75, 0.90 and the result show below:

|  |  |  |  |
| --- | --- | --- | --- |
| threshold | Precision | Recall | Fmeasure |
| 0.05 | 0.00031302139471 | 1.0 | 0.00062584688596 |
| 0.20 | 0.00112923738104 | 0.9716981132075 | 0.00225585317243 |
| 0.45 | 0.04759441282979 | 0.86792452830188 | 0.09024031387935 |
| 0.75 | 0.72477064220183 | 0.74528301886792 | 0.73488372093023 |
| 0.90 | 0.89534883720930 | 0.72641509433962 | 0.80208333333333 |

As the threshold increases, the Precision increases, Recall decreases. When threshold = 0.90, Fmeasure has the best result.

