INFS3200 Advanced Database Systems

**Prac 3: Distributed Databases (5%)**

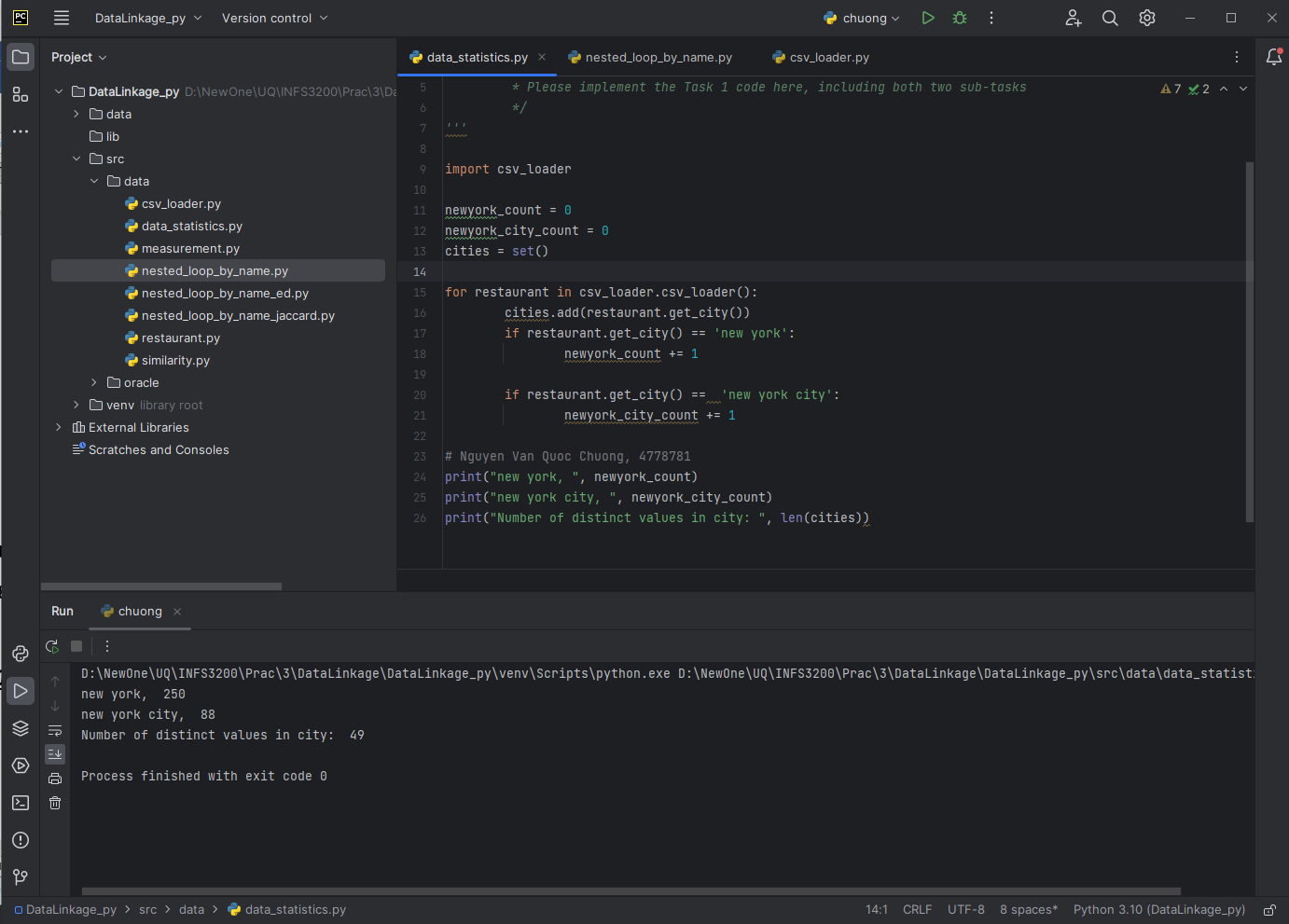
*Semester 1, 2023*

**Student’s name: Nguyen Van Quoc Chuong**

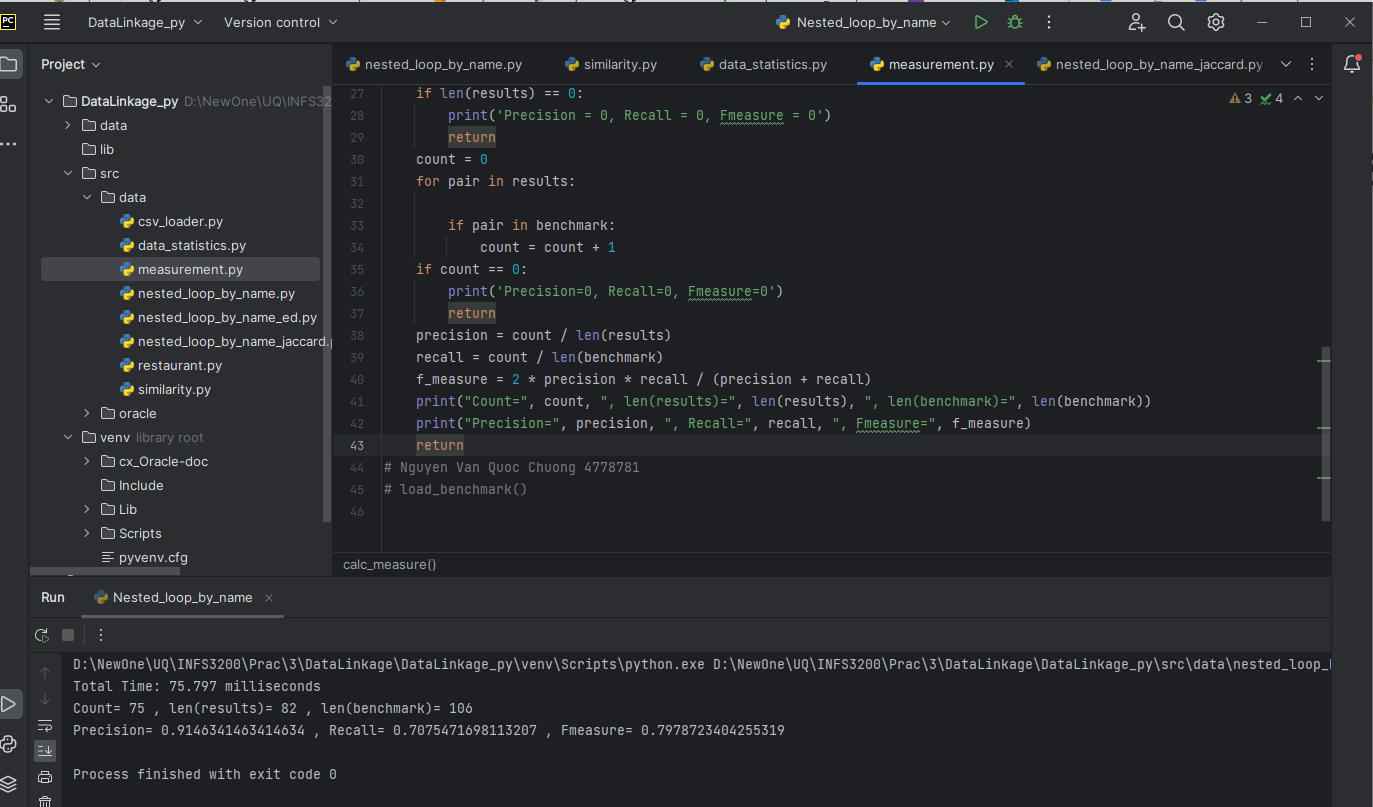
**ID: 47787810**

***Task 1:***

import csv\_loader  
  
newyork\_count = 0  
newyork\_city\_count = 0  
cities = set()  
  
for restaurant in csv\_loader.csv\_loader():  
 cities.add(restaurant.get\_city())  
 if restaurant.get\_city() == 'new york':  
 newyork\_count += 1  
  
 if restaurant.get\_city() == 'new york city':  
 newyork\_city\_count += 1  
  
# Nguyen Van Quoc Chuong, 4778781  
print("new york, ", newyork\_count)  
print("new york city, ", newyork\_city\_count)  
print("Number of distinct values in city: ", len(cities))



***Task 2:***

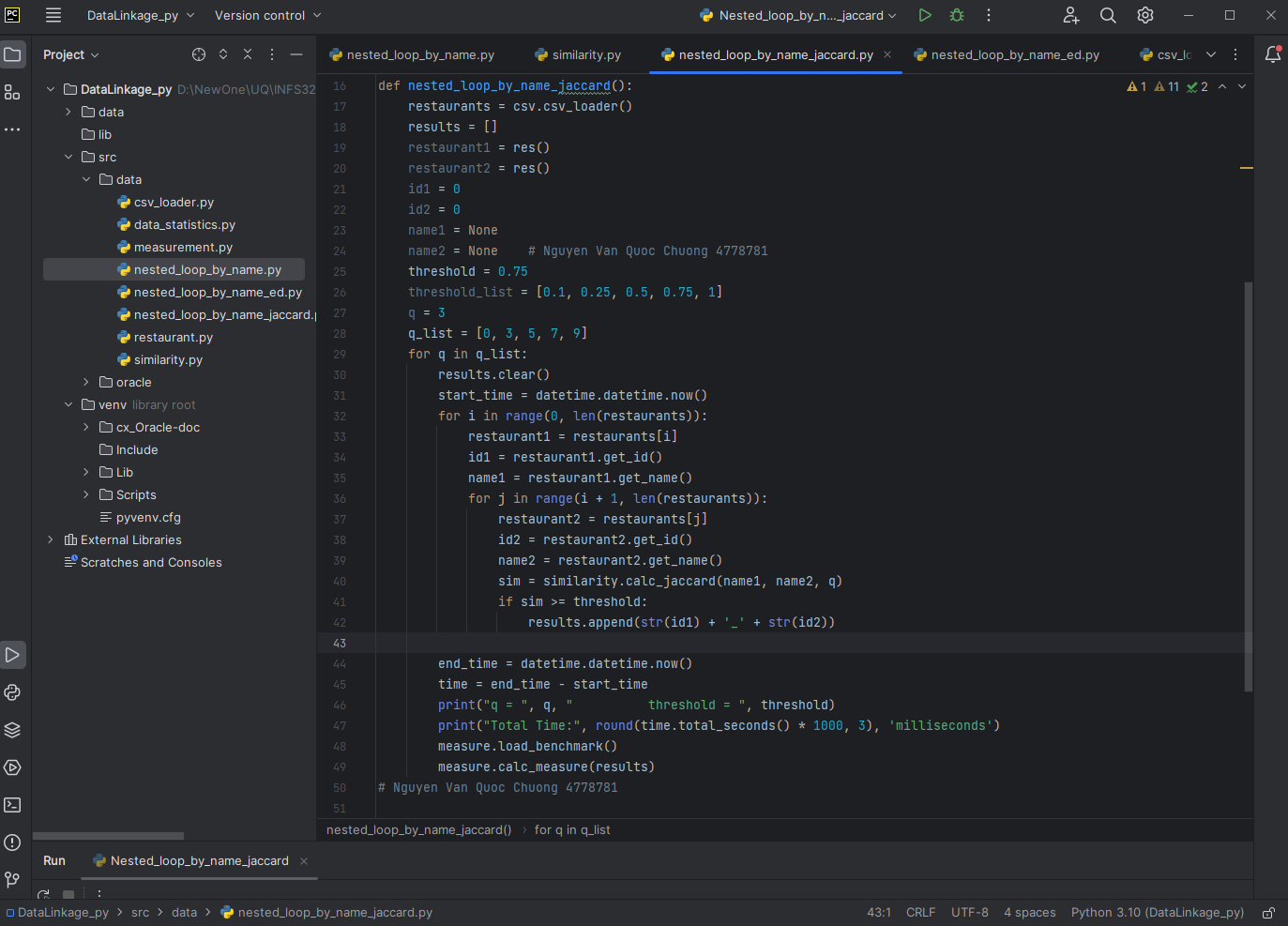


Calc\_measure(results) functions:

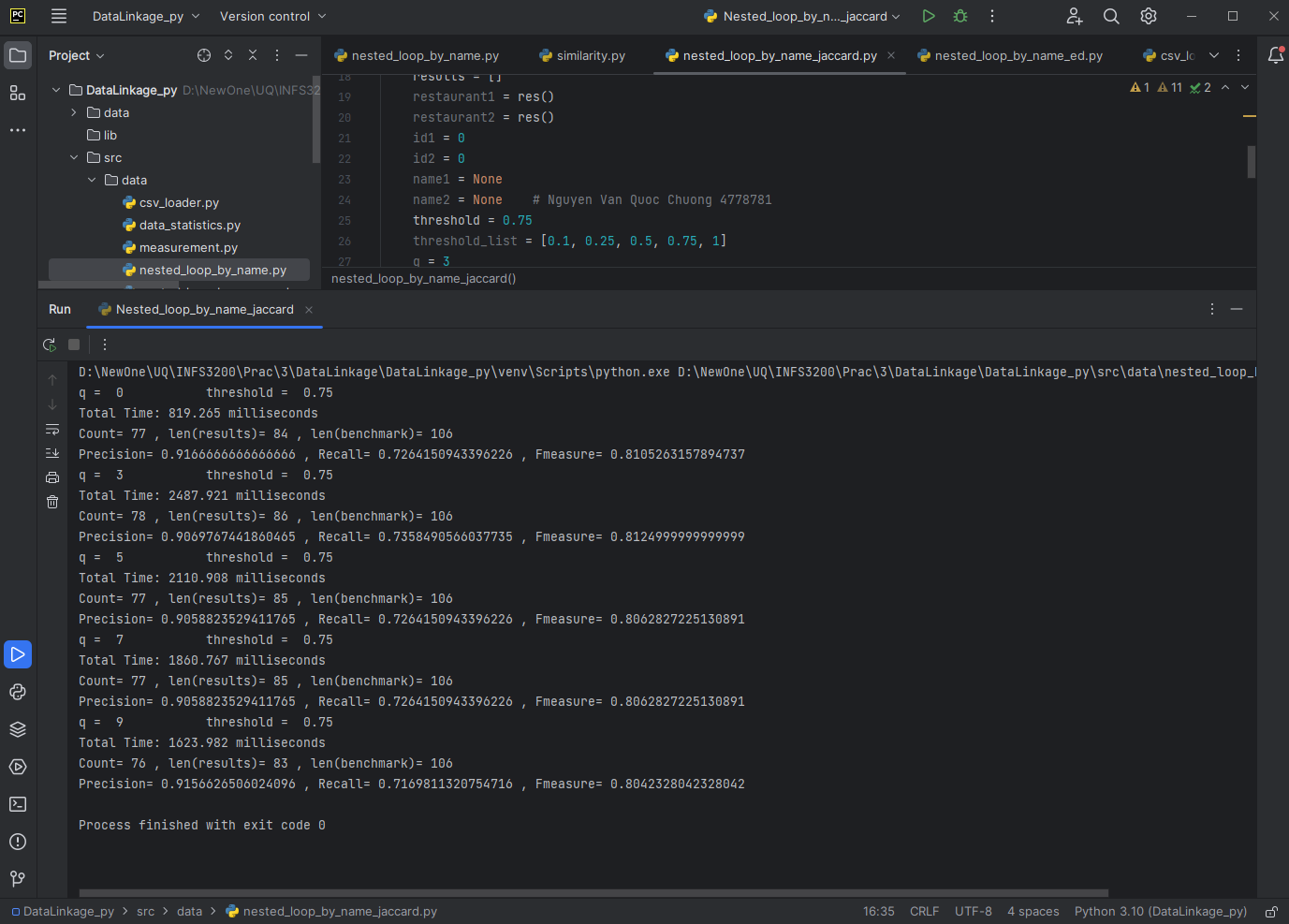
* The variable count is initialized as 0 and is incremented for each pair in results that matches with the pairs in the benchmark list. (True positive TP)
* Len(result): refers to the number of pairs that are linked by the algorithm.
* Len(benchmark): represents the number of pairs that should be linked based on the ground truth benchmark.
* f\_measure: refers to a statistical measure that combines both precision and recall to provide a single numerical value as the harmonic mean of precision and recall that represents the overall performance of the record linkage algorithm.
* True Positive (TP): The number of record pairs that are correctly linked by the algorithm and are also present in the benchmark dataset.
  + TP = count = 75
* True Negative (TN): In the context of data linkage, TN is not applicable as there is no definitive set of non-matches. This is not calculated in the code.
  + TN = 1 – TP – FN - FP
* False Positive (FP): The number of record pairs that are linked by the algorithm but are not present in the benchmark dataset. This is calculated by the difference of len(result) and count
  + FP = len(result) – count = 82 – 75 = 7
* False Negative (FN): The number of record pairs that are not linked by the algorithm but are present in the benchmark dataset. This is computed by subtracting count from len(benmark)
  + F = len(benchmark) – count = 106 – 75 = 31
* Precision: The fraction of correctly linked record pairs out of the total number of record pairs linked by the algorithm.
  + TP / (TP + FP) = count / len(result) = 75/82
* Recall: The fraction of correctly linked record pairs out of the total number of record pairs that should be linked according to the ground truth benchmark.
  + TP / (TP + FN) = count / len(benchmark) = 75 / 106

***Task 3:***

Change the q from 0 to 9 with the fixed threshold = 0.75

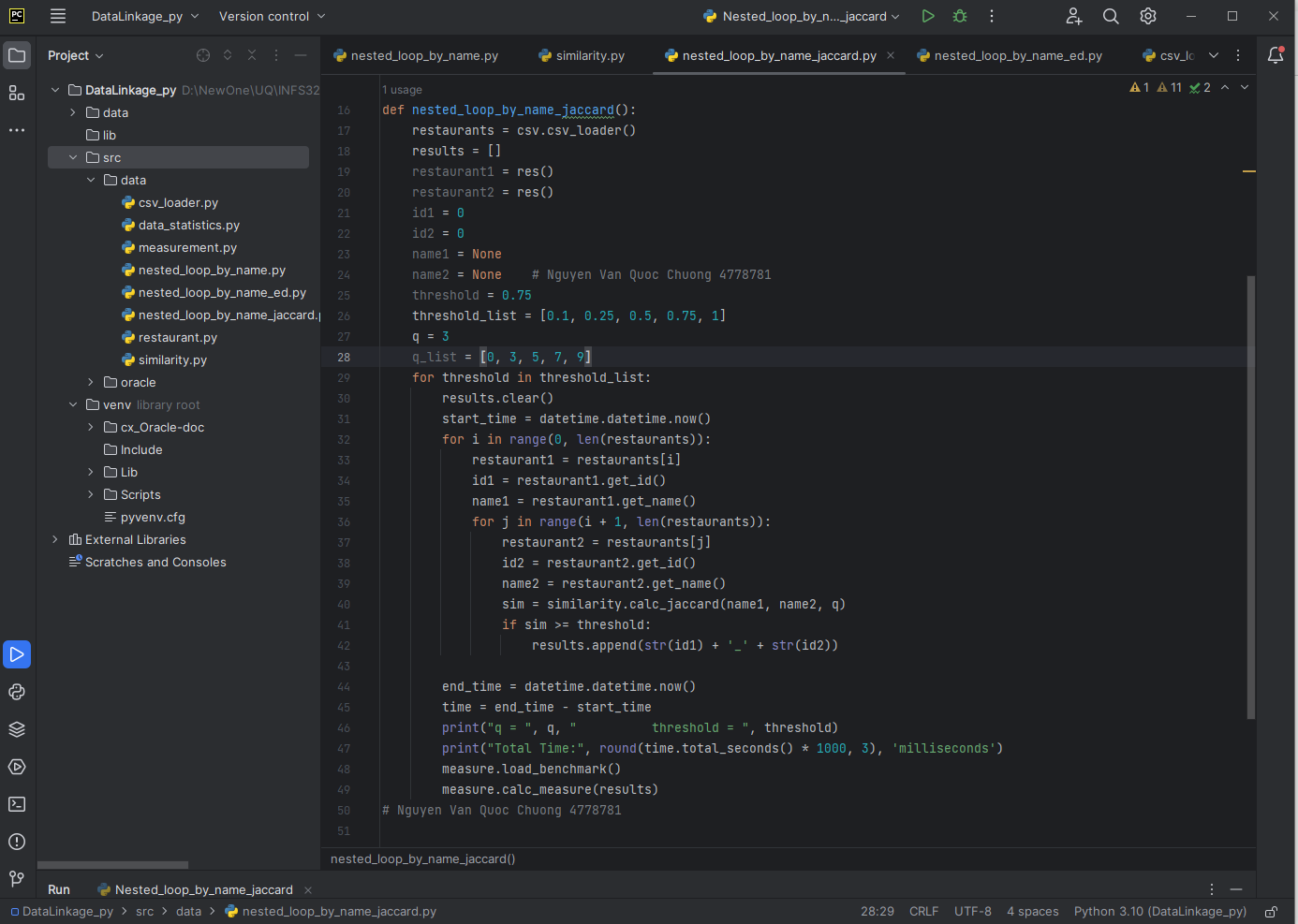


Output:



Explaination: we can see that increasing the value of q from 0 to 3 increases the number of pairs of restaurants in the results list, but does not significantly affect the precision, recall, or F1 score of the similarity measure. However, increasing the value of q further to 5, 7, or 9 does not increase the number of pairs significantly but decreases the precision, recall, and F1 score of the similarity measure. The threshold of 0.75 seems to be a reasonable value for this dataset, as increasing it would decrease the number of pairs in the results list but increasing it further would not significantly affect the precision, recall, or F1 score. Overall, the results suggest that a value of q = 3 and threshold = 0.75 provide a good balance between performance and accuracy for this dataset. Moreover, the execution time will descrease due to the increasing in value of q.

Change the threshold from 0.1 to 1 with a constant q = 3:



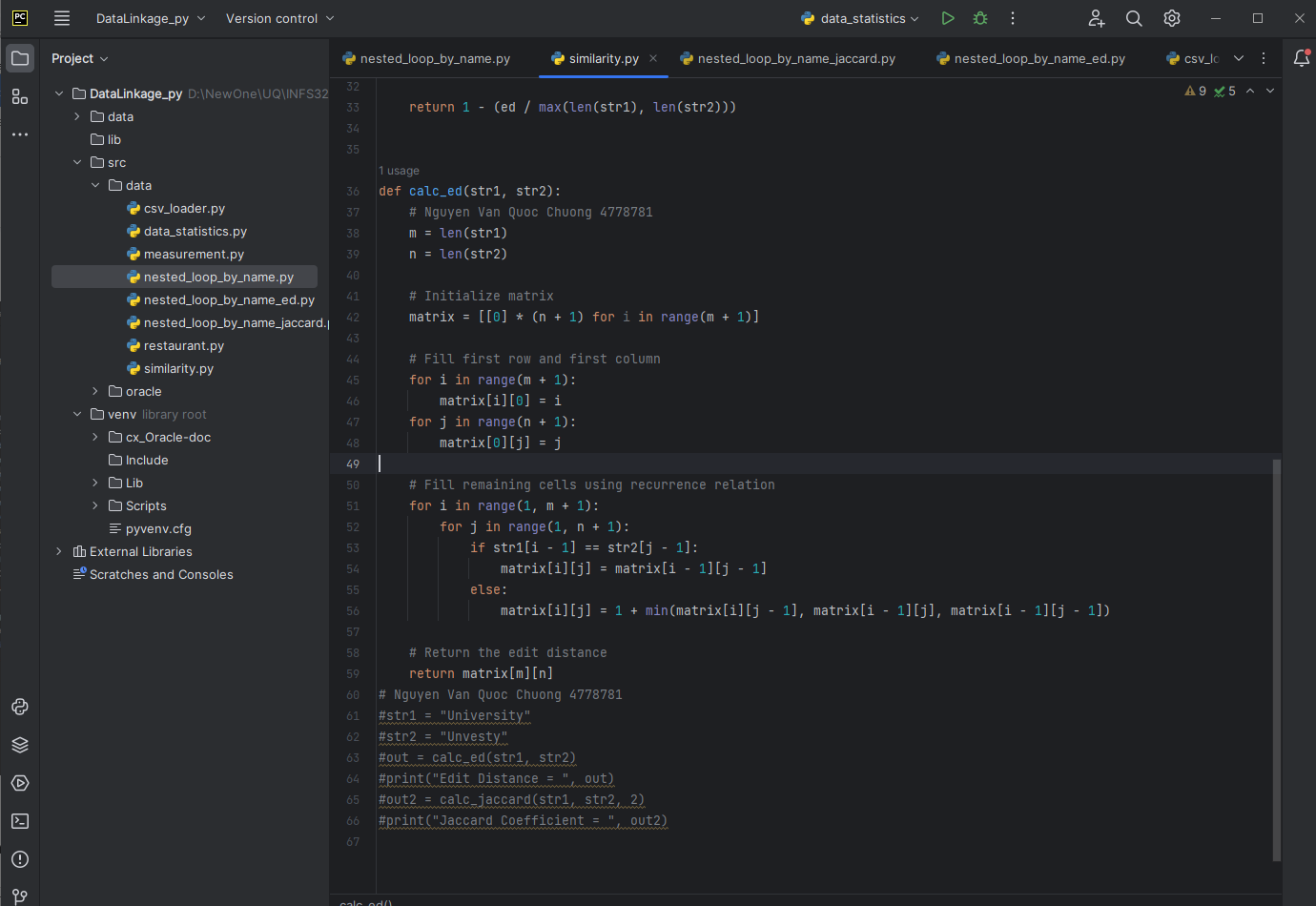
Output:

A screenshot of a computer program

Description automatically generated with medium confidence

Explaination: The number of results (len(results)) for each iteration decreases as the threshold increases, which is expected since a higher threshold means a stricter requirement for similarity between restaurant names. The recall value decreases as the threshold increases, meaning that more relevant results are being filtered out. On the other hand, precision generally increases as the threshold increases, indicating that the results are becoming more accurate. The F-measure reaches its peak at a threshold of 0.75

***Task 4:***

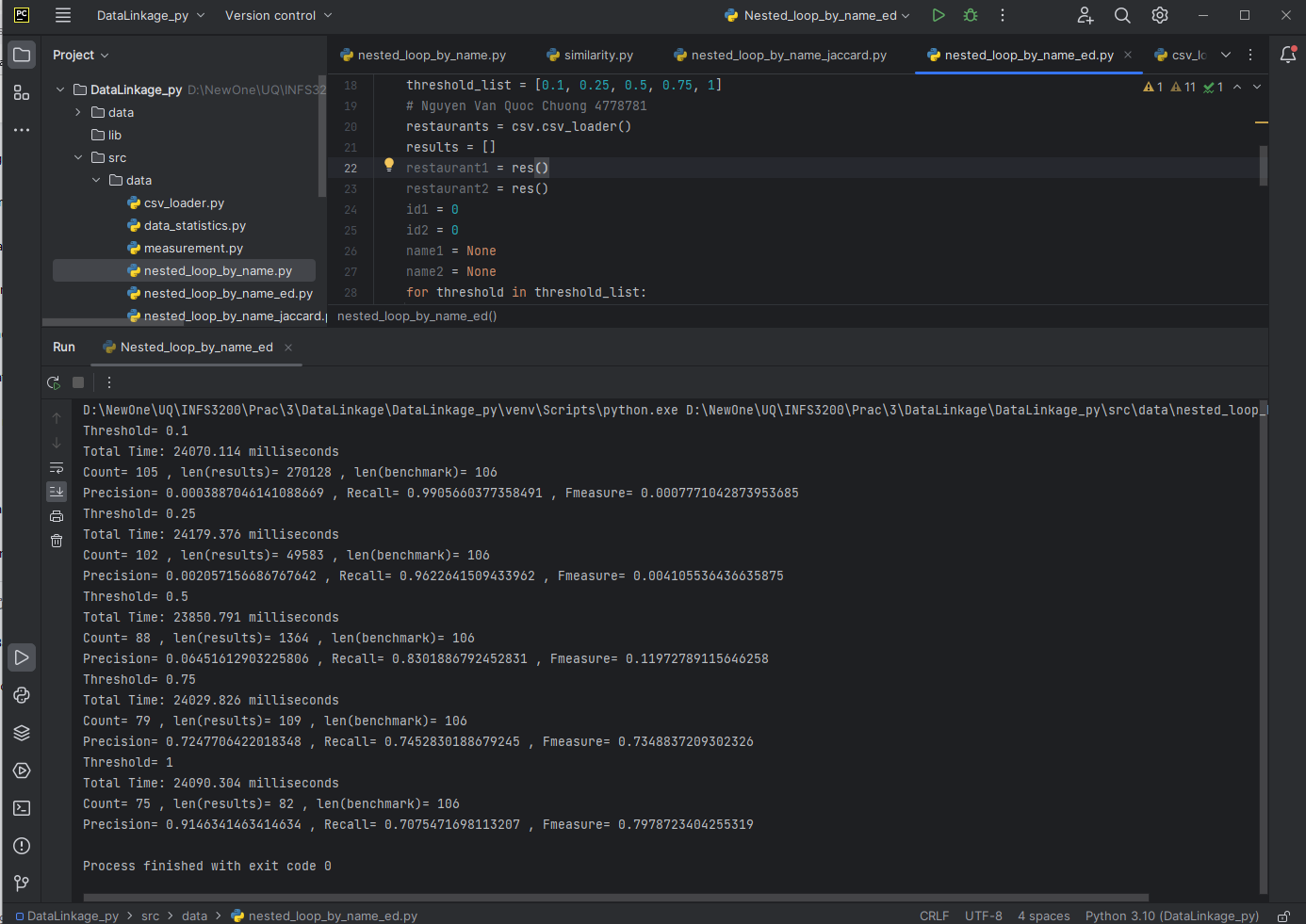


I change the values of threshold from 0.1 to 1:

A screenshot of a computer program

Description automatically generated with medium confidence

Output:



Explaination: As the threshold value increases from 0.1 to 1, the precision increases while the recall decreases. This is because as the threshold becomes more strict, fewer pairs are included in the results, leading to a higher precision but a lower recall. The F-measure also shows a similar trend, reaching its peak at a threshold of 0.75. Overall, my ouput shows that a threshold of 0.75 provides a good balance between precision and recall. Moreover, the running time of each threshold does not seem to be changed.

Top of Form

Bottom of Form