Graph Data Construction Pipeline

Soojung Hong

August 8. 2019

raw data : chicago_newyork_place.csv

```
1    8401r4ve-ceb55322b3dd0f54c584b2758ba9831f | 2010 Parking | 700-7400-0000 | 9990 | Westchester Ave | USA | 10462 | en | 40.83363 | -73.85706
2    8401r4ve-1a02bbb9a72406e35a823430aa2493b2 | Kaleidoscope Concepts | 700-7400-0141 | 9500 | 5th Ave | USA | 10016 | en | 40.74666 | -73.98582
3    840dr5ru-85f703aebc1448ffa5ede3f5893e4838 | Theracare Preschool Services | 700-7400-0000 | 700-7400-0141 | W 32nd St | USA | 10001 | en | 40.74879 | -73.98944
4    840dr72w-c7b8f21d6b4a4092bf7a12942603fdf6 | Roger Singer Consulting | 700-7200-0264 | 541618 | Paul Ave | USA | 10468 | en | 40.87749 | -73.88886
5    8403fv6k-b21f31bbc1a9032bb302977cf028c862 | Lee Tai Enterprises USA | 700-7400-0145 |   | Madison Ave | USA | 10017 | en | 40.75696 | -73.97643
6    8401r4ve-7d11607267a10cd355bdaaf2775e531a | Aapt | 700-7400-0000 | 9990 | W 14th St | USA | 10011 | en | 40.73815 | -73.9988
7    840dr5r7-00112b6b5baf44029e878fe809c1eb8c | L'Epicerie | 100-1000-0000 | 7221 | Ferris St | USA | 11231 | en | 40.67786 | -74.01739
8    8401r4ve-8c50977e99080cd64daa12b85417b94e | Lidmark Home Renovation | 700-7400-0000 | 9990 | Gary Pl | USA | 10314 | en | 40.59805 | -74.16505
9    8403fv6k-0bb48db1d0bf0e5ef4506b1e0f5c91db | Riverview II | 700-7400-0000 | 100-1000-0000 | W 135th St | USA | 10031 | en | 40.82065 | -73.95677
10   8403fv6k-b40be9097f5e0dd87ced4c6eb006cb4e | Kessler Investment Advisors | 700-7200-0271 | 52391 | E 76th St | USA | 10021 | en | 40.77469 | -73.96403
11   8403fv6k-38b037606f470b473f9b26fa0779f7c4 | MPDL Design and Architert | 700-7400-0000 | 700-7200-0256 | Little West 12th St | USA | 10014 | en | 40.74002 | -74.00702
12   840dr5qb-1b718460f7ec43f788e7cb44a54391db | Jus That Graphics | 700-7400-0141 | 9500 | DINSMORE ST | USA | 10314 | en | 40.60419 | -74.17006
13   840dr5q8-1090a8b935f04b91a5e995b48c9fd831 | 4215 Good Taste Chinese Food | 100-1000-0000 | 201-000 | HYLAN BLVD | USA | 10308 | en | 40.53733 | -74.15047
14   840dr72j-d98b339328f84efba751bcaa40342b21 | Compucare Team | 700-7200-0261 |   | E 92nd St | USA | 10128 | en | 40.78296 | -73.95173
15   840dr72h-0ff90328e6fa46d0b0f3fe98cebf5469 | Tom McWilliam Photography | 700-7400-0144 | 9507 | W 99th St | USA | 10025 | en | 40.79759 | -73.97282
```

pid	place name	category code	...	lat	lon

raw data : PDS-categories.csv, Level_1_Categories.csv

```
1   PDS Category Code, PDS Category Name, description
2   100-1000-0000,Restaurant,"An establishment that sells refreshments and prepared meals. This category should be used for all Places that do not fit other categories defined for Restaurant (100-1000-xxxx)."
3   100-1000-0001,Casual Dining,"A restaurant serving moderately-priced food in a casual atmosphere that usually includes table service."
4   100-1000-0002,Fine Dining,"A full-service restaurant that serves full-course meals in a formal setting. Places in this category usually have high quality décor, highly-trained chefs, wait staff in formal attire and vi
5   100-1000-0003,Take Out and Delivery Only,"A restaurant that offers take-out service, delivery service or both."
6   100-1000-0004,Food Market-Stall,"A restaurant that serves specialty foods at a food court, marketplace or outdoor setting including hawker centers (common in Southeast Asia)."
7   100-1000-0005,Taqueria,"A street vendor stand or small restaurant that serves traditional Mexican food, such as tacos or burritos."
8   100-1000-0006,Deli,"A restaurant that sells ready-to-serve delicatessens including cold cut meats, cheeses and salads. This type of restaurant can be found as a stand-alone establishment or within a grocery store."
9   100-1000-0007,Cafeteria,"A restaurant that provides food service with little or no wait staff. This type of establishment is common in schools, large office buildings, hospitals and other public establishments. "
10  100-1000-0008,Bistro,"A restaurant that serves moderately priced meals in a European-styled casual setting."
11  100-1000-0009,Fast Food,"A restaurant that serves food that is prepared and served quickly. Common examples of fast food restaurants include McDonald's, Taco Bell and KFC. "
12  100-1000-0230,Unspecified Food Category,"Not on the PDS list."
13  100-1100-0000,Coffee-Tea,"An establishment that sells drinks, such as coffee and tea, as well as refreshments. This category should be used for all Places that do not fit other categories defined for Coffee-Tea (100-1
14  100-1100-0010,Coffee Shop,"An establishment that primarily sells coffee, but may also serve light foods, such as pastries or other snacks."
15  100-1100-0331,Tea House,"An establishment that primarily sells tea and other related products."
16  200-2000-0000,Nightlife-Entertainment,"An establishment that provides evening entertainment and usually serves alcoholic beverages. This base-level category should be used for all Places that do not fit other categori
17  200-2000-0011,Bar or Pub,"An establishment that sells alcoholic beverages. Places in this category may also serve food or light meals. Common examples include public houses, sports bars and grills. "
18  200-2000-0012,Night Club,"An establishment that provides evening entertainment generally until the very early morning. Common examples include dance clubs, lounges, wine bars.  "
19  200-2000-0013,Dancing,"An establishment that provides dancing facilities. Common examples include dance halls, dance studios and other locations with dance floors. "
20  200-2000-0014,Karaoke,"An establishment that provides karaoke entertainment."
```

All Categories

```
1   Level_1_Code,Level_1_Name
2   100,Eat and Drink
3   200,Going Out-Entertainment
4   300,Sights and Museums
5   350,Natural and Geographical
6   400,Transport
7   500,Accommodation
8   550,Leisure and Outdoor
9   600,Shopping
10  700,Business and Services
11  800,Facilities
12  900,Areas and Buildings
```

Level 1 Categories

Merged place data with category
(source code : merge_raw_place_with_category_and_split_per_city.py)

chicago_newyork_place.csv ➕ PDS-categories.csv & Level_1_Categories.csv

Merged Data Frame (merged_data_frame.csv)


```
merged_data_df['pid_int'] = pd.factorize(merged_data_df.pid)[0]
```

Id_in_merged	lat	lon	pid

chicago.csv   nyc.csv

Split merged_data_frame.csv to each city's dataframe
(source code : merge_raw_place_with_category_and_split_per_city.py)

Id_in_merged	lat	lon	pid

```
merged_data_df['pid_int'] =
pd.factorize(merged_data_df.pid)[0]
```
using 'pid', make 'pid_int' by factorize

```
bbox_nyc = {'left_lon':-74.2591, 'bottom_lat':40.4774, 'right_lon':-73.7002, 'top_lat':40.9162}
bbox_chicago = {'left_lon':-88.133, 'bottom_lat':41.6062, 'right_lon':-87.4656, 'top_lat':42.1603}
```

```
1   id_in_merged,lat,lon,pid
2   48,41.8807,-87.63414,8401r4ve-b0a1a89079ed0ab0710901ff2137f386
3   49,41.88837,-87.63535999999999,840dp3wm-864ebee4834a4100bffab290ba7f2263
4   50,41.74507,-87.71377,8401r4ve-dcee5b4aabf60abb29fef88c18ac3564
5   51,41.8004,-87.60359,840dp3ty-2b534805468b46e2b168efe0fbd11d7b
6   52,41.75124,-87.58309,840dp3tw-125d1dfd49df49c59733705c39bf90f3
7   53,41.80066,-87.60289,840dp3ty-2df955b9847c485c9a08e965354a81ab
8   54,41.90788,-87.65101,8403fv6k-de2a37520ee001781720048d9e436e82
9   55,41.8663,-87.71826,8401r4ve-8377e192b622028ccfecd8697a7c8f66
10  56,41.87573,-87.63172,8401r4ve-96b51ea086790aae6a28218c918e79ee
11  57,41.89647,-87.64286,840dp3wm-56b5c2be2f7e4aa8995ac6f79dfc6a6f
12  58,41.853491399999996,-87.6341677,840dp3wj-b0f7a7e090f841b7983df6465d66e442
13  59,41.71945,-87.57523,840dp3tr-3632b6247bac4bd1b5f355151f1ea761
14  60,41.99184,-87.66186,840dp3wv-d22f43ac405c4840956e56f2e894d7fc
15  61,41.874423,-87.6243649,840dp3wn-46ffbc071f0841c6aa8382b54a5d6af2
16  62,41.7805,-87.59510999999999,8401r4ve-788f2e4255820ffed58c51e5c739fdfe
17  63,41.89485,-87.76432,840dp3w6-ab58056c12264e6990d67e3b2136f004
18  64,41.78632,-87.79415,840dp3td-c1cb813ded7e49e5949fb07e6ce68e5f
```

chicago.csv

```
1   ,lat,lon,pid
2   0,40.74666,-73.98581999999999,8401r4ve-1a02bbb9a72406e35a823430aa2493b2
3   1,40.60419,-74.17005999999999,840dr5qb-1b718460f7ec43f788e7cb44a54391db
4   2,40.63257,-73.99473,8401r4ve-c2094740dbe103b5f47aeba00925ad35
5   3,40.720659999999995,-74.01012,8401r4ve-ab33b1d342f50712735b75b022bb7a0d
6   4,40.81612,-73.94521999999999,840dr72j-dfc41fce42f841e6a5ba7b3d6a762635
7   5,40.720490000000005,-74.00627,840dr5re-1f27ebba4d4f401bba42ece3de1f4557
8   6,40.63675,-74.02461,8401r4ve-b27fe1e200740c7a692b0c6309b1fcff
9   7,40.69767,-73.93472,840dr5rt-146d25f01bcb4be09c2763e1c6c6bcdd3
10  8,40.7102871,-73.9651084,840dr5rs-4d9e119ddfe14ca5822c284edc3dd3f8
11  9,40.81002470000001,-73.9280416,8403fv6k-f937cd60a486004ebe14b23010dc9ad3
12  10,40.70998,-73.9277,8403fv6k-1ab673e0c9940a004e6ee9b0367b66ef
13  11,40.79977,-73.96969,840dr72h-59775cb3f0a34093829d1bd3b0abd3e5
14  12,40.633415,-74.0114569,840dr5r5-525a44fb92d54505af7ceb62ebd0edbd
15  13,40.646390000000004,-73.96158,840dr5rh-9ac2a6e1eb564a7b95beeea55d896a0f
16  14,40.74853,-73.99243,8401r4ve-71d1470639c4053a469193137aeac54b
17  15,40.690059999999995,-73.92498,840dr5rm-386564310d4648eda3625eef8fd33dba
18  16,40.7638,-73.97518000000001,8401r4ve-b4d7f3be35bd0648b9b93637324c3a25
19  17,40.5814373,-73.9677997,840dr5qu-59a8352b61c345f3aa999031375d6609
20  18,40.63867,-73.95363,840dr5rj-ce141ed188c24c74a0f4d62e6e341047
21  19,40.73494,-74.00265,8401r4ve-568cd28971eb054e1e03a414a0d3553b
22  20,40.72822,-73.98671999999999,840dr5rs-9f0f4aa010ea48dfbc16d9691add6a57
23  21,40.77693,-73.97981999999999,8401r4ve-cf08a76663d308bc19f99a9501e27619
24  22,40.73656,-73.85772,840dr5rx-51e8e9a8eea94afca3c2bda731a82fd2
25  23,40.67236,-73.81111999999999,840dr5x2-69feac63e46f424da5f682b8c285bd28
```
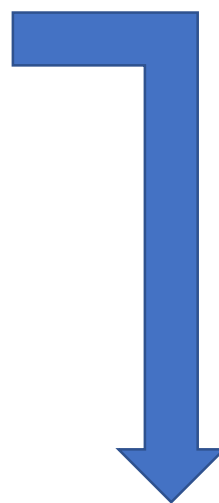
nyc.csv

Adding place name vector & category vector

(source code : create_dataframe_with_place_and_category_embedding_vector.py)

```
1   id_in_merged,lat,lon,pid
2   48,41.8807,-87.63414,8401r4ve-b0a1a89079ed0ab0710901ff2137f386
3   49,41.88837,-87.63535999999999,840dp3wm-864ebee4834a4100bffab290ba7f2263
4   50,41.74507,-87.71377,8401r4ve-dcee5b4aabf60abb29fef88c18ac3564
5   51,41.8004,-87.60359,840dp3ty-2b534805468b46e2b168efe0fbd11d7b
6   52,41.75124,-87.58309,840dp3tw-125d1dfd49df49c59733705c39bf90f3
7   53,41.80066,-87.60289,840dp3ty-2df955b9847c485c9a08e965354a81ab
8   54,41.90788,-87.65101,8403fv6k-de2a37520ee001781720048d9e436e82
9   55,41.8663,-87.71826,8401r4ve-8377e192b622028ccfecd8697a7c8f66
10  56,41.87573,-87.63172,8401r4ve-96b51ea086790aae6a28218c918e79ee
11  57,41.89647,-87.64286,840dp3wm-56b5c2be2f7e4aa8995ac6f79dfc6a6f
12  58,41.853491399999996,-87.6341677,840dp3wj-b0f7a7e090f841b7983df6465d66e442
13  59,41.71945,-87.57523,840dp3tr-3632b6247bac4bd1b5f355151f1ea761
14  60,41.99184,-87.66186,840dp3wv-d22f43ac405c4840956e56f2e894d7fc
15  61,41.874423,-87.6243649,840dp3wn-46ffbc071f0841c6aa8382b54a5d6af2
16  62,41.7805,-87.59510999999999,8401r4ve-788f2e4255820ffed58c51e5c739fdfe
17  63,41.89485,-87.76432,840dp3w6-ab58056c12264e6990d67e3b2136f004
18  64,41.78632,-87.79415,840dp3td-c1cb813ded7e49e5949fb07e6ce68e5f
```

chicago.csv

+ Place name to BERT embedding vector
+ Label embedding vector

chicago_dataset_prep.csv

```
1    ,pid,place_name,pid_int,Level_1_Code,category_id,category_name,place_embedding,label_encoding
2    0,840dp3sy-20d1775fb30c4a02b7e6537727814c09 , St. Mary of the Assumption ,0,700,700-7200-0269,Organizations and Societies,"[-1.74715489e-01  4.08351809e-01  3.96962285e-01  2.23932058e-01
3       5.26237726e-01 -2.35716030e-01 -3.87203068e-01 -5.26686966e-01
4       3.90139930e-02 -5.88226438e-01  6.77395880e-01  3.01174492e-01
5       1.40942007e-01 -8.65686983e-02 -5.39555430e-01 -3.50973547e-01
6      -3.66449147e-01  1.68439955e-01  4.95820522e-01 -2.77181685e-01
7       3.62269402e-01  3.86260539e-01 -9.61957455e-01 -2.38257378e-01
8      -4.48122501e-01 -1.97863579e-01  3.71831954e-02 -6.33582950e-01
9      -2.48921350e-01 -1.65049404e-01  9.23781097e-01  3.96046042e-03
10      6.47879779e-01  3.40692937e-01 -2.13416554e-02  7.78934360e-01
11     -5.19545913e-01 -1.87932640e-01 -2.05483532e+00 -2.61593200e-02
12      4.16051924e-01  1.15724802e-01 -5.18061101e-01 -2.79989183e-01
187    -1.19119681e-01 -3.97050560e-01  2.03351349e-01  4.32683378e-01
188     1.56413361e-01 -4.97340500e-01 -6.52063405e-03  9.47097123e-01
189    -7.33256042e-01  7.45610952e-01  5.91144264e-01 -3.48486543e-01
190     4.56334978e-01 -1.03427398e+00  7.59260476e-01  2.37306401e-01
191     1.69161588e-01  1.04476643e+00  2.56296657e-02 -8.53758335e-01
192    -8.55177522e-01 -2.29032516e-01  2.04653174e-01  3.76439244e-01
193     1.47533882e-02  6.80688024e-01  9.53419805e-02  1.64584279e-01]",[0 0 1 0 0 0 0 0 0 0 0]
```

Constructing graph in adjacency_matrix

(source code : construct_adjacency_matrix.py)

chicago_dataset_prep.csv


```
1    ,pid,place_name,pid_int,Level_1_Code,category_id,category_name,place_embedding,label_encoding
2    0,840dp3sy-20d1775fb30c4a02b7e6537727814c09 , St. Mary of the Assumption ,0,700,700-7200-0269,Organizations and Societies,"[-1.74715489e-01  4.08351809e-01  3.96962285e-01  2.23932058e-01
3     5.26237726e-01 -2.35716030e-01 -3.87203068e-01 -5.26686966e-01
4     3.90139930e-02 -5.88226438e-01  6.77395880e-01  3.01174492e-01
5     1.40942007e-01 -8.65686983e-02 -5.39555430e-01 -3.50973547e-01
6    -3.66449147e-01  1.68439955e-01  4.95820522e-01 -2.77181685e-01
7     3.62269402e-01  3.86260539e-01 -9.61957455e-01 -2.38257378e-01
8    -4.48122501e-01 -1.97863579e-01  3.71831954e-02 -6.33582950e-01
9    -2.48921350e-01 -1.65049404e-01  9.23781097e-01  3.96046002e-03
10    6.47879779e-01  3.40692937e-01 -2.13416554e-01  7.78934360e-01
11   -5.19545913e-01 -1.87932640e-01 -2.05483532e+00 -2.61593200e-02
12    4.16051924e-01  1.15724802e-01 -5.18061101e-01 -2.79989183e-01
187  -1.19119681e-01 -3.97050560e-01  2.03351349e-01  4.32683378e-01
188   1.56413361e-01 -4.97340500e-01 -6.52063405e-03  9.47097123e-01
189  -7.33256042e-01  7.45610952e-01  5.91144264e-01 -3.48486543e-01
190   4.56334978e-01 -1.03427398e+00  7.59260476e-01  2.37306401e-01
191   1.69161588e-01  1.04476643e+00  2.56296657e-02 -8.53758335e-01
192  -8.55177522e-01 -2.29032516e-01  2.04653174e-01  3.76439244e-01
193   1.47533882e-02  6.80688024e-01  9.53419805e-02  1.64584279e-01]",[0 0 1 0 0 0 0 0 0 0]
```



```
1    ,pid_int,neighbor_pid_int
2    0,0,3
3    1,0,4
4    2,0,5
5    3,0,6
6    4,0,7
7    5,3,4
8    6,3,5
9    7,3,6
10   8,3,7
11   9,4,5
12   10,4,6
13   11,4,7
14   12,5,6
15   13,5,7
16   14,6,7
17   15,8,11
18   16,12,13
19   17,14,15
20   18,14,16
21   19,14,17
22   20,14,18
23   21,14,19
24   22,14,20
25   23,14,21
26   24,14,22
27   25,14,23
28   26,15,16
29   27,15,17
30   28,15,18
31   29,15,19
32   30,15,20
```

<start_node, end_node>

Chicago_adj_matrix.csv


```
1    ,pid_int,neighbor_pid_int,distance
2    0,3,4,0.0
3    1,3,5,0.0
4    2,3,6,0.0
5    3,3,7,0.0
6    4,4,5,0.0
7    5,4,6,0.0
8    6,4,7,0.0
9    7,5,6,0.0
10   8,5,7,0.0
11   9,6,7,0.0
12   10,8,11,0.031492871094327295
13   11,12,13,0.0
14   12,14,15,0.012001048284246177
15   13,14,16,0.012001048284246177
16   14,14,17,0.012001048284246177
17   15,14,18,0.015915523676496936
18   16,14,19,0.015915523676496936
19   17,14,20,0.015915523676496936
20   18,14,21,0.015915523676496936
21   19,14,22,0.015915523676496936
22   20,14,23,0.025636815204265428
23   21,15,16,0.0
24   22,15,17,0.0
25   23,15,18,0.008466233775933672
26   24,15,19,0.008466233775933672
27   25,15,20,0.008466233775933672
28   26,15,21,0.008466233775933672
29   27,15,22,0.008466233775933672
30   28,15,23,0.02220278318828604
```

<start_node, end_node, distance>

Chicago_adj_matrix_with_distance.csv

Run GCN using graph saved as pt

(source code : train_test_graph_convolutional_model.py)

```
# load Chicago tensors - x, y and edges
edge_index = torch.load('edge_index_tensor.pt')
x = torch.load('x_tensor.pt')
y = torch.load('y_tensor.pt')
dataset = Data(x=x, edge_index=edge_index, y=y)
```

Model define
Model train
Model evaluation