ĐẠI HỌC QUỐC GIA TP.HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

Báo cáo đồ án môn học

MẠNG XÃ HỘI

Chủ Đề : Phân tích mạng xã hội trực tuyến Meetup.com

Giáo viên hướng dẫn : Thái Bảo Trân
Lớp : IS353.J21
Thực hiện : PhạmVăn Hữu -15520293

TP HCM, Ngày 12 tháng 6 năm 2019

MỤC LỤC

Contents

I. Lý do chọn đề tài

Meetup là một trang web mạng xã hội nhằm mục đích mang mọi người đến với nhau để cùng làm việc, khám phá, dạy và học về những điều mà có thể giúp cuộc sống của họ trở nên sống động có ý nghĩa hơn.

Meetup cho phép các người dùng có thể tìm và tham gia vào các nhóm (nhóm được tạo ra từ tập hợp những người có cùng chung sở thích,nghề nghiệp ,lĩnh vực ,..).Kể từ năm 2017,Meetup đã có hơn 32 triệu người dùng với 280 nghìn nhóm tại 182 quốc gia.Các thành viên xác định nhóm hay hoạt động mà họ quan tâm nhất để có thể sử dụng mạng xã hội này hiệu quả.

Để hiểu hơn về cấu trúc mạng xã hội Meetup là như thế nào?Mối quan hệ giữa các thành viên trong nhóm?Các nhóm có mối quan hệ gì với nhau?Các nhóm và các thành viên trọng mạng có ảnh hưởng với nhau như thế nào?Xong,từ những hiểu biết rút ra từ việc phân tích mạng xã hội Meetup.com chúng ta có thể xây dựng một hệ thống khuyến nghị các nhóm hay các hoạt động đến các thành viên trong mạng dựa trên sở thích hay sự quan tâm về một lĩnh vực nào đó của họ.

II. Thông tin dữ liệu

- Tên dataset : Nashville Meetup NetworkX
- Data sources:
    - member-to-group-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *member-to-group* với trọng số là số lượng các sự kiện mà thành viên đã tham dự trong mỗi nhóm.
    - group-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *group-to-group* với trọng số là số thành viên chung giữa 2 nhóm.
    - member-edges.csv : Danh sách cạnh cho việc xây dựng đồ thị *member-to-member* với trọng số là số nhóm chung giữa 2 thành viên.
    - rsvps.csv : Dữ liệu thô về việc tham dữ các sự kiện của các thành viên trong một nhóm ,được tổng hợp từ *member-to-group-edges.csv.*
    - meta-Groups.csv: Thông tin cho từng nhóm, bao gồm tên và danh mục.
    - meta-thành viên.csv: Thông tin cho từng thành viên, bao gồm tên và địa điểm

- meta-event.csv: Thông tin cho từng sự kiện, bao gồm tên và thời gian.

- Nguồn dataset: https://www.kaggle.com/stkbailey/nashville-meetup
- Thông số dataset *member-to-group-edges.csv*:
  - Số thuộc tính: 3
  - Số dòng: 45600
  - Dung lượng file: 943.3 KB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1	member_id	ID của thành viên
2	group_id	ID của nhóm
3	weight	Trọng số - số lượng các sự kiện mà thành viên đã tham dự trong mỗi nhóm

- Thông số dataset *group-edges.csv*:
  - Số thuộc tính: 4
  - Số dòng: 6693
  - Dung lượng file: 163.24 KB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1		Thứ tự
2	group1	ID Nhóm 1
3	group2	ID Nhóm 2
4	weight	Trọng số - số thành viên chung giữa 2 nhóm.

- Thông số dataset *member-edges.csv*:
  - Số thuộc tính: 4
  - Số dòng: 1180000
  - Dung lượng file: 32.87 MB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1		Thứ tự
2	member1	ID thành viên 1
3	member2	ID thành viên 2
4	weight	Trọng số - số nhóm chung giữa 2 thành viên.

- Thông số dataset *rsvps.csv*:
  - Số thuộc tính: 4
  - Số dòng: 127000
  - Dung lượng file: 4.37 MB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1		Thứ tự
2	event_id	ID sự kiện
3	member_id	ID thành viên
4	group_id	ID nhóm

- Thông số dataset *meta-groups.csv*:
  - Số thuộc tính: 7
  - Số dòng: 602

- Dung lượng file: 59.66 KB
- Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1	group_id	ID nhóm
2	group_name	Tên nhóm
3	num_members	Số lượng thành viên
4	category_id	ID Danh mục
5	category_name	Tên danh mục
6	organizer_id	ID người tổ chức
7	group_urlname	Tên url của nhóm

- Thông số dataset *meta-members.csv*:
  - Số thuộc tính: 7
  - Số dòng: 24600
  - Dung lượng file:1.22 MB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1	member_id	ID thành viên
2	name	Tên thành viên
3	hometown	Quê quán
4	city	Thành phố
5	state	Tiểu Bang
6	lat	
7	lon	

- Thông số dataset *member-to-group-edges.csv*:
  - Số thuộc tính: 4
  - Số dòng: 19300
  - Dung lượng file:1.46 MB
  - Bảng thông tin của thuộc tính

STT	Tên	Ý nghĩa
1	event_id	ID sự kiện
2	group_id	ID nhóm
3	name	Tên sự kiện
4	time	Thời gian diễn ra

III. Phân tích mạng xã hội nhóm với nhóm trên Meetup ở Nashville

1. Xây dựng và vẽ đồ thị

1.1. Nhập thư viện và đọc dữ liệu

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import seaborn as sns
import numpy as np
plt.style.use('fivethirtyeight')

## Network
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pylab as plt
from itertools import count
from operator import itemgetter
from networkx.drawing.nx_agraph import graphviz_layout
import pylab

df=pd.read_csv('../test/input/group-edges.csv')
df.head()
```

Kết quả:

	Unnamed: 0	group1	group2	weight
0	0	19292162	535553	2
1	1	19292162	19194894	1
2	2	19292162	19728145	1
3	3	19292162	18850080	2
4	4	19292162	1728035	1

1.2. Xem thông tin Dataset

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6692 entries, 0 to 6691
Data columns (total 4 columns):
Unnamed: 0    6692 non-null int64
group1        6692 non-null int64
group2        6692 non-null int64
weight        6692 non-null int64
dtypes: int64(4)
memory usage: 209.2 KB
```

1.3. Xem các thống kê cơ bản của dữ liệu

```
df.describe()
```

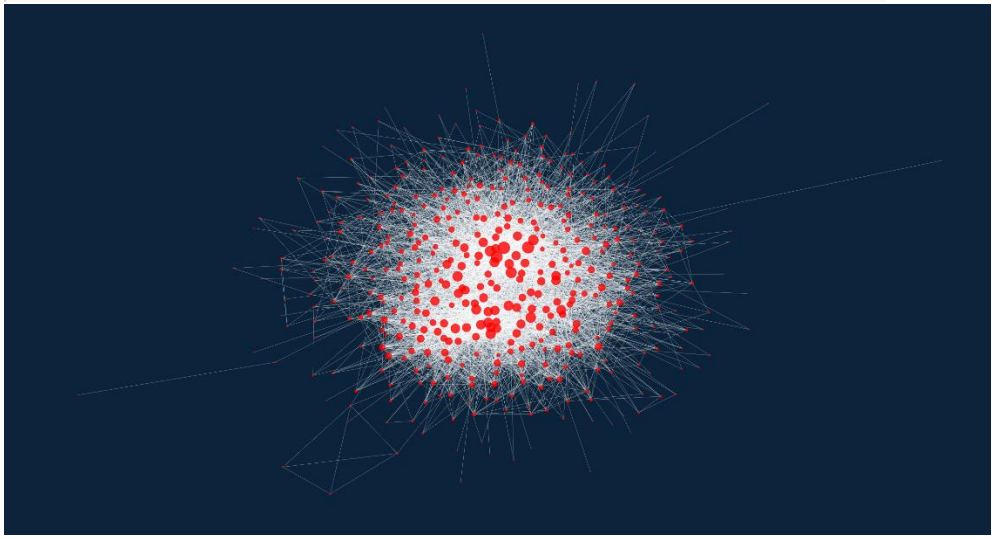
	Unnamed: 0	group1	group2	weight
count	6692.000000000	6.6920000000e+03	6.6920000000e+03	6692.0000000000
mean	3345.500000000	1.2535057028e+07	1.2962960003e+07	2.3018529588
std	1931.958332884	8.2965403778e+06	8.1157206251e+06	4.0899733233
min	0.000000000	1.6801400000e+05	4.7094000000e+04	1.0000000000
25%	1672.750000000	2.1504910000e+06	3.3760420000e+06	1.0000000000
50%	3345.500000000	1.6487812000e+07	1.6487812000e+07	1.0000000000
75%	5018.250000000	1.9266390000e+07	1.9416348000e+07	2.0000000000
max	6691.000000000	2.6091301000e+07	2.6091301000e+07	91.0000000000

1.4. Vẽ đồ thị mạng với các loại layouts khác nhau

Đối với dữ liệu các nhóm trên Meetup.com ở khu vực Nashville thì mỗi đỉnh sẽ là các nhóm độc lập trên Meetup và các liên kết giữa hai nhóm sẽ hình thành từ số thành viên chung của hai nhóm đó.Trọng số là số lượng thành viên chung của hai nhóm.

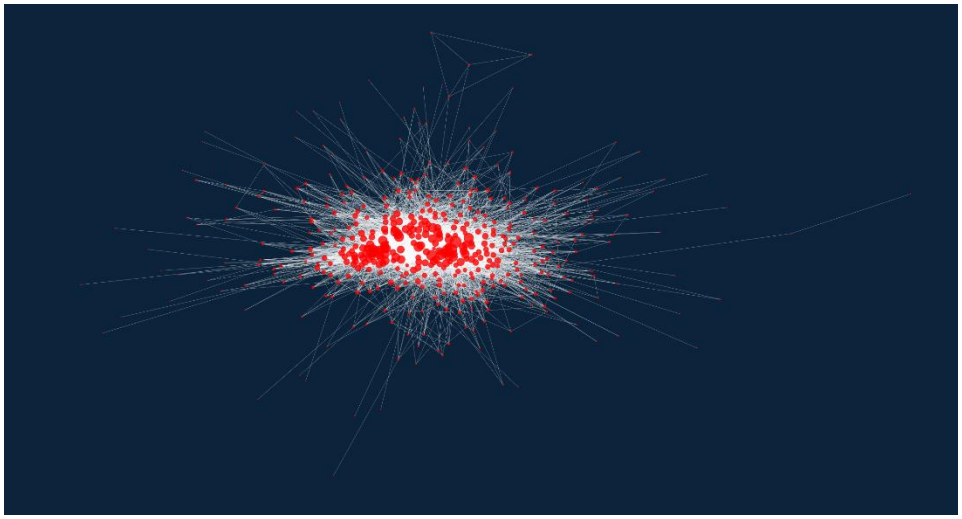
1.4.1. Kamada_kawai_layout

```
pd.set_option('precision',10)
G = nx.from_pandas_edgelist(df, source='group1', target='group2',
                            edge_attr='weight',
                            create_using = nx.Graph())

nodes = G.nodes()
degree = G.degree()
size = [(degree[n]) for n in nodes]

pos = nx.kamada_kawai_layout(G)
#pos = nx.spring_layout(G, k = 0.2)
#pos = nx.circular_layout(G)
#pos = nx.random_layout(G)
cmap = plt.cm.viridis_r
cmap = plt.cm.Purples

fig = plt.figure(figsize = (15,8), dpi=150)

nx.draw(G,pos,alpha = 0.8, nodelist = nodes, node_color = 'r',
        node_size = size, with_labels= False,font_size = 6,
        width =0.2 , edge_color ='w')
fig.set_facecolor('#0B243B')

plt.show()
```



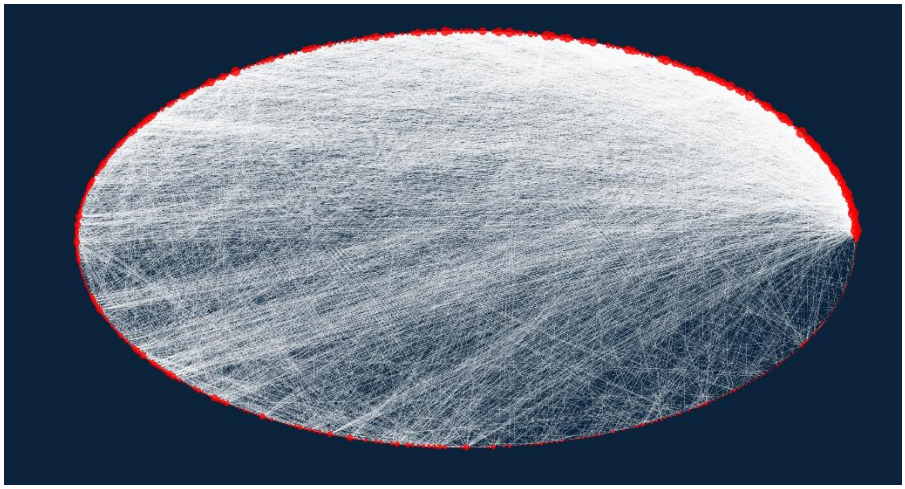
1.4.2. Spring_layout

```
pd.set_option('precision',10)
G = nx.from_pandas_edgelist(df, source='group1', target='group2',
                            edge_attr='weight',
                            create_using = nx.Graph())

nodes = G.nodes()
degree = G.degree()
size = [(degree[n]) for n in nodes]

#pos = nx.kamada_kawai_layout(G)
pos = nx.spring_layout(G, k = 0.2)
#pos = nx.circular_layout(G)
#pos = nx.random_layout(G)
cmap = plt.cm.viridis_r
cmap = plt.cm.Purples

fig = plt.figure(figsize = (15,8), dpi=150)

nx.draw(G,pos,alpha = 0.8, nodelist = nodes, node_color = 'r',
        node_size = size, with_labels= False,font_size = 6,
        width =0.2 , edge_color ='w')
fig.set_facecolor('#0B243B')

plt.show()
```



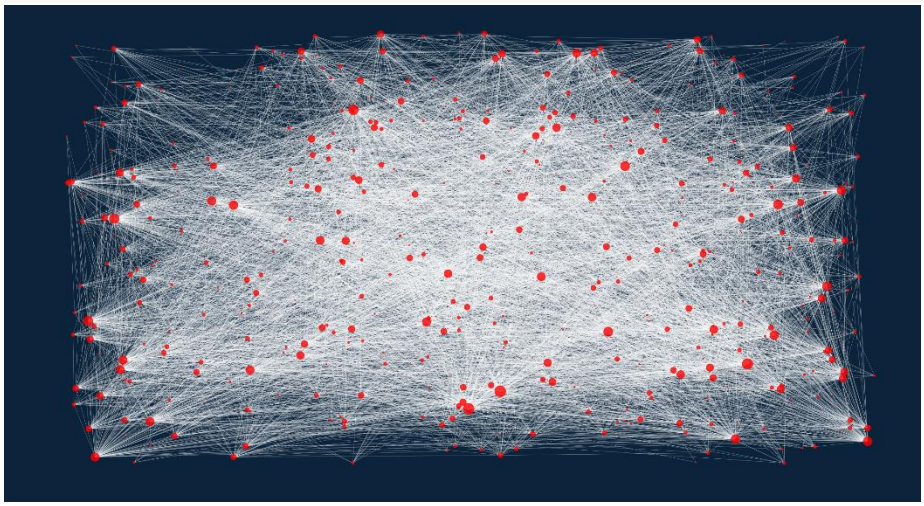
1.4.3. Circular_layout

```
pd.set_option('precision',10)
G = nx.from_pandas_edgelist(df, source='group1', target='group2',
                            edge_attr='weight',
                            create_using = nx.Graph())

nodes = G.nodes()
degree = G.degree()
size = [(degree[n]) for n in nodes]

#pos = nx.kamada_kawai_layout(G)
#pos = nx.spring_layout(G, k = 0.2)
pos = nx.circular_layout(G)
#pos = nx.random_layout(G)
cmap = plt.cm.viridis_r
cmap = plt.cm.Purples

fig = plt.figure(figsize = (15,8), dpi=150)

nx.draw(G,pos,alpha = 0.8, nodelist = nodes, node_color = 'r',
        node_size = size, with_labels= False,font_size = 6,
        width =0.2 , edge_color ='w')
fig.set_facecolor('#0B243B')

plt.show()
```



1.4.4. Random_layout

```
pd.set_option('precision',10)
G = nx.from_pandas_edgelist(df, source='group1', target='group2',
                            edge_attr='weight',
                            create_using = nx.Graph())

nodes = G.nodes()
degree = G.degree()
size = [(degree[n]) for n in nodes]

#pos = nx.kamada_kawai_layout(G)
#pos = nx.spring_layout(G, k = 0.2)
#pos = nx.circular_layout(G)
pos = nx.random_layout(G)
cmap = plt.cm.viridis_r
cmap = plt.cm.Purples

fig = plt.figure(figsize = (15,8), dpi=150)

nx.draw(G,pos,alpha = 0.8, nodelist = nodes, node_color = 'r',
        node_size = size, with_labels= False,font_size = 6,
        width =0.2 , edge_color ='w')
fig.set_facecolor('#0B243B')

plt.show()
```



2. Phân tích đồ thị

2.1. Tính các đặc trưng của mạng

Đọc dữ liệu file *meta-groups.csv* để lấy ra thông tin những nhóm tồn tại trong đồ thị.Từ đó ta tính các đặc trưng như Degree, Clustering,Path length và thêm các giá trị đó vào DataFrame

```
groups=pd.read_csv('../test/input/meta-groups.csv',index_col='group_id')
len(groups)
```

```
602
```

```
groups=groups.loc[[x for x in G.nodes]]
len(groups)
```

```
456
```

Tính các đặc trưng và thêm vào DataFrame:

```
groups['degree'] = pd.Series(dict(nx.degree(G)))
groups['clustering'] = pd.Series(nx.clustering(G))
avg_length_dict = {}
for node, path_lengths in nx.shortest_path_length(G):
    path_lengths = [x for x in path_lengths.values()]
    avg_length_dict[node] = np.mean(path_lengths)
groups['path_length'] = pd.Series(avg_length_dict)
```

Kết quả:

```
groups.head()
```

group_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname	degree	clustering	path_length
19292162	Nashville CocoaHeads	237	34	Tech	145632652	Nashville-CocoaHeads	37	0.6921921922	2.1447368421
535553	Nash.rb	881	34	Tech	14344641	nashrb	65	0.5596153846	2.0285087719
19194894	Nashville Christian Technologists and Entrepre...	613	34	Tech	193181718	Nashville-Christian-Technologists-and-Entrepre...	55	0.5616161616	2.1337719298
19728145	Stepping Out Social Dance Meetup	1778	5	Dancing	118484462	steppingoutsocialdance	182	0.1716349948	1.6425438596
18850080	NashReact	438	34	Tech	10083866	NashReact-Meetup	55	0.6020202020	2.1644736842

2.1.1. Top 10 nhóm có giá trị Degree cao nhất trong mạng

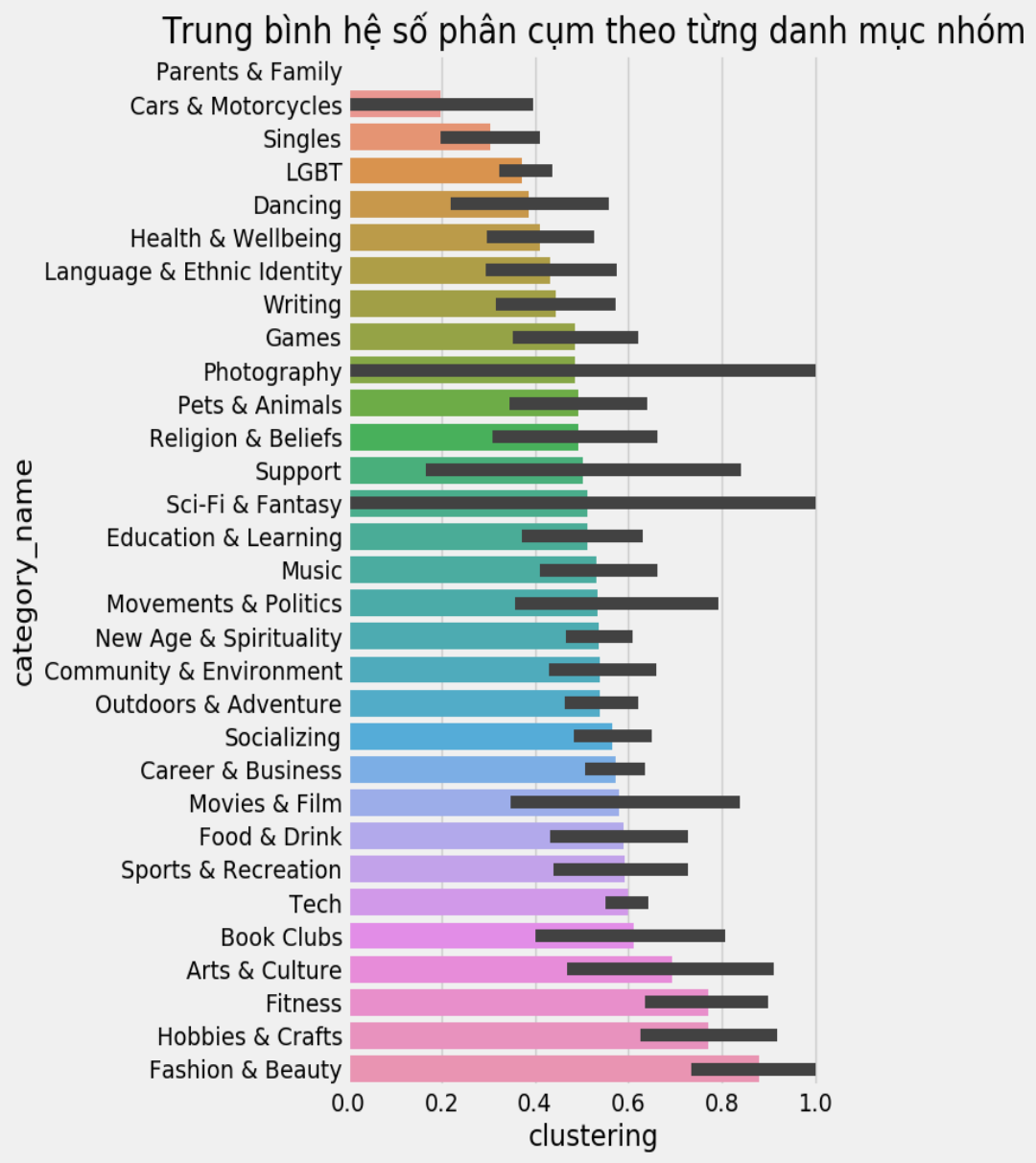
```
groups.sort_values(by='degree', ascending=False)[['group_name','degree']].head(10)
```

group_id	group_name	degree
19728145	Stepping Out Social Dance Meetup	182
18955830	Eat Love Nash	176
1187715	What the Pho!	168
18506072	20s in Nashville	145
339011	Nashville Hiking Meetup	136
4126912	Nashville Online Entrepreneurs	129
18243826	Middle TN 40+ singles	129
1776274	Nashville SEO & Internet Marketing, Over 1,600...	125
11077852	Sunday Assembly Nashville	125
16487812	Code for Nashville	117

2.1.2. Tính trung bình hệ số phân cụm theo từng danh mục nhóm

```
fig, ax = plt.subplots(1,1, figsize=(5,10), dpi=100)

sns.barplot(data=groups, x='clustering', y='category_name',
            order=groups.groupby('category_name').clustering.mean().sort_values().index)
ax.set_title('Trung bình hệ số phân cụm theo từng danh mục nhóm')

plt.show()
```

Kết quả :



Ta có thể thấy được các nhóm như Tech, Book Clubs, Art & Culture, Fitness, Hoobies & Crafts và Fashion & Beauty là những danh mục rất phổ biến.Như đã đề cập khi xây dựng đồ thị thì những liên kết được tạo ra từ số thành viên chung giữa hai nhóm nên những thành viên thuộc những nhóm của các danh mục phổ biến này cũng sẽ có xu hướng tham gia vào các nhóm trong các danh mục phổ biến kể trên.

2.2. Tìm nhóm trung tâm trong mạng

2.2.1. Tính Betweenness Centrality

Tính giá trị betweenness centrality trong mạng và thêm vào Dataframe

```
groups['betweenness_centrality'] = pd.Series(dict(nx.betweenness_centrality(G)))
groups.head()
```

ıp_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname	degree	clustering	path_length	betweenness_centrality
2162	Nashville CocoaHeads	237	34	Tech	145632652	Nashville-CocoaHeads	37	0.6921921922	2.1447368421	0.0002551306
5553	Nash.rb	881	34	Tech	14344641	nashrb	65	0.5596153846	2.0285087719	0.0026079135
4894	Nashville Christian Technologists and Entrepre...	613	34	Tech	193181718	Nashville-Christian-Technologists-and-Entrepre...	55	0.5616161616	2.1337719298	0.0013327107
8145	Stepping Out Social Dance Meetup	1778	5	Dancing	118484462	steppingoutsocialdance	182	0.1716349948	1.6425438596	0.0731614127
0080	NashReact	438	34	Tech	10083866	NashReact-Meetup	55	0.6020202020	2.1644736842	0.0009975528

2.2.2. Top 10 nhóm trung tâm trong mạng

```
groups.sort_values(by='betweenness_centrality', ascending=False)[['group_name','betweenness_centrality']].head(10)
```

group_id	group_name	betweenness_centrality
19728145	Stepping Out Social Dance Meetup	0.0731614127
18955830	Eat Love Nash	0.0531121856
1187715	What the Pho!	0.0525033206
18243826	Middle TN 40+ singles	0.0394879706
339011	Nashville Hiking Meetup	0.0362513964
18506072	20s in Nashville	0.0357014185
1776274	Nashville SEO & Internet Marketing, Over 1,600...	0.0301803387
4126912	Nashville Online Entrepreneurs	0.0298541449
11077852	Sunday Assembly Nashville	0.0273215992
4705492	WOMEN "Word of Mouth Entrepreneurial Networkers"	0.0270924489

IV. Phân tích các nhóm thuộc danh mục 'Tech' trên Meetup ở Nashville

1. Xây dựng và vẽ đồ thị

1.1. Nhập thư viện và đọc dữ liệu

```
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import seaborn as sns
import numpy as np
plt.style.use('fivethirtyeight')

## Network
import networkx as nx
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import pylab as plt
from itertools import count
from operator import itemgetter
from networkx.drawing.nx_agraph import graphviz_layout
import pylab

df=pd.read_csv('../test/input/group-edges.csv')
df.head()
```

Kết quả :

	Unnamed: 0	group1	group2	weight
0	0	19292162	535553	2
1	1	19292162	19194894	1
2	2	19292162	19728145	1
3	3	19292162	18850080	2
4	4	19292162	1728035	1

1.2. Tạo đồ thị các nhóm thuộc danh mục 'Tech' trên Meetup ở Nashville

- Tạo đồ thị các nhóm trên Meetup

```
g0 = nx.from_pandas_edgelist(df, source='group1', target='group2', edge_attr='weight')
```

- Đọc dữ liệu thông tin các nhóm

```
groups=pd.read_csv('../test/input/meta-groups.csv',index_col='group_id')
```

- Lấy ra các nhóm thuộc danh mục 'Tech'

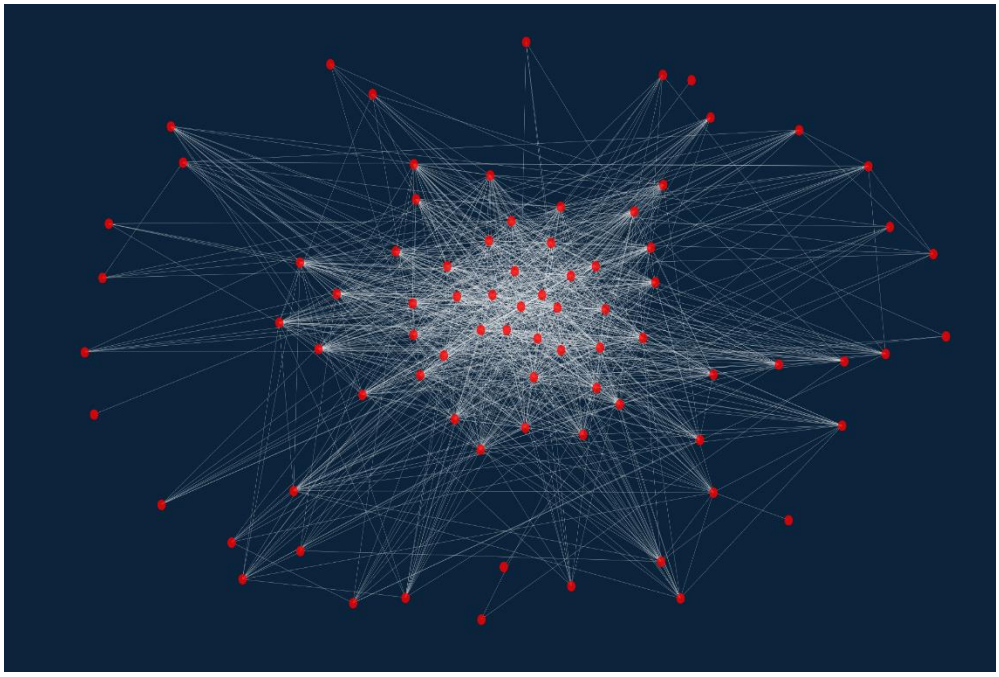
```
tech=groups.loc[groups.category_name=='Tech
```

- Tạo đồ thị các nhóm thuộc danh mục 'Tech' từ đồ thị các nhóm trên Meetup

```
gt = g0.subgraph(tech.index) ### Tạo đồ thị con của g0
g = [gt.subgraph(c) for c in nx.connected_components(gt)][0]
tech = tech.loc[(n for n in g.nodes)]
```

1.3. Vẽ đồ thị các nhóm thuộc danh mục 'Tech'

```
pos = nx.spring_layout(g, k = 2)

fig = plt.figure(figsize = (15,8), dpi=150)

nx.draw(g,pos,alpha = 0.8, node_color = 'r',
        node_size = 80, with_labels= False,font_size = 6,
        width =0.2 , edge_color ='w')
fig.set_facecolor('#0B243B')

plt.show()
```



2. Phân tích đồ thị

2.1. Tính các đặc trưng của mạng

Tính các đặc trưng như Degree, Clustering,Path length và thêm các giá trị đó vào DataFrame

```
tech['degree'] = pd.Series(dict(nx.degree(g)))
tech['clustering'] = pd.Series(nx.clustering(g))
avg_length_dict = {}
for node, path_lengths in nx.shortest_path_length(g):
    path_lengths = [x for x in path_lengths.values()]
    avg_length_dict[node] = np.mean(path_lengths)
tech['path_length'] = pd.Series(avg_length_dict)
```

```
tech.head()
```

Kết quả:

group_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname	degree	clustering	path_length
535553	Nash.rb	881	34	Tech	14344641	nashrb	46	0.7420289855	1.4320987654
19292162	Nashville CocoaHeads	237	34	Tech	145632652	Nashville-CocoaHeads	29	0.8497536946	1.6419753086
19194894	Nashville Christian Technologists and Entrepre...	613	34	Tech	193181718	Nashville-Christian-Technologists-and-Entrepre...	33	0.8106060606	1.6049382716
19275797	Mindful Cyborgs Meditation	256	34	Tech	112028452	mindfulcyborgs	16	0.8750000000	1.8641975309
21533726	State & Local Government Developers Network	123	34	Tech	4611657	SLGDN_Nashville	36	0.7936507937	1.5679012346

2.2. Tìm nhóm trung tâm trong mạng

2.2.1. Tính Betweenness Centrality

```
tech['betweenness_centrality'] = pd.Series(dict(nx.betweenness_centrality(g)))
tech.head()
```

Kết quả :

group_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname	degree	clustering	path_length	betweenness_centrality
535553	Nash.rb	881	34	Tech	14344641	nashrb	46	0.7420289855	1.4320987654	0.0105087552
19292162	Nashville CocoaHeads	237	34	Tech	145632652	Nashville-CocoaHeads	29	0.8497536946	1.6419753086	0.0010423881
19194894	Nashville Christian Technologists and Entrepre...	613	34	Tech	193181718	Nashville-Christian-Technologists-and-Entrepre...	33	0.8106060606	1.6049382716	0.0031670944
19275797	Mindful Cyborgs Meditation	256	34	Tech	112028452	mindfulcyborgs	16	0.8750000000	1.8641975309	0.0004173492
21533726	State & Local Government Developers Network	123	34	Tech	4611657	SLGDN_Nashville	36	0.7936507937	1.5679012346	0.0035208503

2.2.2. Top 10 nhóm trung tâm của mạng các nhóm 'Tech'

```
tech.sort_values(by='betweenness_centrality', ascending=False)[['group_name','betweenness_centrality']].head(10)
```

group_id	group_name	betweenness_centrality
10016242	NashJS	0.0651802515
1728035	WordPress Nashville	0.0457699182
20947040	Nashville Blockchain Meetup	0.0416132935
6707902	Data Science Nashville	0.0385245350
13560402	Nashville Modern Excel & Power BI User Group	0.0330723538
16487812	Code for Nashville	0.0315800431
18494105	The Iron Yard - Nashville	0.0285798775
23353167	Nashville Docker Meetup	0.0276636149
10016162	Nashville PowerShell User Group (NashPUG)	0.0266104008
18616278	Franklin Developer Lunch & Learn	0.0264267442

V. Ứng dụng-Xây dựng một hệ tư vấn đơn giản (Recommender System)

1. Đặt vấn đề

Giả sử khi bạn mới bắt đầu sử dụng mạng xã hội Meetup và muốn tham gia vào một nhóm nào đó.Ví dụ cụ thể ở đây là bạn muốn tìm một nhóm thuộc danh mục 'Tech' và muốn tham gia vào nhóm 'NashJS' (nhóm trung tâm của danh mục).Và sau khi tham gia vào nhóm này rồi thì bạn cũng có nhu cầu tham gia vào những nhóm khác .Vậy bạn sẽ tham gia vào nhóm nào trong hàng trăm nhóm? Nhóm nào có tính tương đồng như nhóm mình đã tham gia?Hệ tư vấn sẽ giúp bạn làm điều đó.

2. Ý tưởng

Mạng xã hội Meetup được tạo ra nhằm mục đích có thể tạo ra một mạng lưới các nhóm để những người dùng có thể tham gia vào các nhóm thuộc sở thích,lĩnh vực học tập,nghiên cứu,…và để tham gia vào các sự kiện trong nhóm để gặp những người có cùng sở thích,kiến thức,kinh nghiệm trong nhóm để học tập,chia sẽ,giao lưu,..Như vậy thì ta có thể sử dụng thuộc tính số lần tham gia sự kiện của các thành viên trong nhóm và trung bình số lần tham dự sự kiện trên một thành viên của nhóm để tìm ra sự tương đồng này.Hiểu một cách đơn giản là bạn sẽ tìm ra những nhóm có nhiều thành viên,nhiều sự kiện được tổ chức và số lần tham gia sự kiện của các thành trong nhóm có cao hay không?Điều này giúp bạn có thể thỏa mãn được nhu cầu tìm ra được nhóm có nhiều sự kiện được tổ chức,nhóm hoạt động

sôi nổi,trong nhóm có nhiều thành viên tích cực thông qua số lần tham gia sự kiện của thanh viên đó,…

3. Thực hiện ý tưởng

3.1. Import thư viên và lấy những thông tin cần thiết từ dataset

```
import pandas as pd
import numpy as np
```

- Lấy số lần tham gia sự kiện của các thanh viên trong nhóm

```
###Lấy thông tin số lần tham gia sự kiện của các thành viên trong nhóm
df1=pd.read_csv('../test/input/member-to-group-edges.csv ')
df1.head()
```

	member_id	group_id	weight
0	2069	19277993	3
1	625050	19277993	2
2	1939496	19277993	1
3	2606806	19277993	4
4	3438546	19277993	1

- Lấy thông tin của nhóm

```
groups=pd.read_csv('../test/input/meta-groups.csv',index_col='group_id') ###Lấy thông tin của các nhóm
groups.head()
```

group_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname
339011	Nashville Hiking Meetup	15838	23	Outdoors & Adventure	4353803	nashville-hiking
19728145	Stepping Out Social Dance Meetup	1778	5	Dancing	118484462	steppingoutsocialdance
6335372	Nashville soccer	2869	32	Sports & Recreation	108448302	Nashville-soccer
10016242	NashJS	1975	34	Tech	8111102	nashjs
21174496	20's & 30's Women looking for girlfriends	2782	31	Socializing	184580248	new-friends-in-Nashville

❖ Lọc các nhóm thuộc danh mục 'Tech'

```
tech=groups.loc[groups.category_name=='Tech']###Lọc các nhóm thuộc danh mục 'Tech'
tech.head()
```

group_id	group_name	num_members	category_id	category_name	organizer_id	group_urlname
10016242	NashJS	1975	34	Tech	8111102	nashjs
11625832	PyNash	1442	34	Tech	215201845	PyNash
19218850	Greater Nashville Healthcare Analytics	764	34	Tech	12825115	Greater-Nashville-Healthcare-Analytics
18589616	Agile Nashville User Group	862	34	Tech	126249582	Agile-Nashville-User-Group
19277993	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps

3.2. Gộp dữ liệu hai bảng trên

```
df = pd.merge(df1,tech,on='group_id') ### Gộp dữ liệu 2 dataframe trên
df.head()
```

	member_id	group_id	weight	group_name	num_members	category_id	category_name	organizer_id	group_urlname
0	2069	19277993	3	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps
1	625050	19277993	2	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps
2	1939496	19277993	1	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps
3	2606806	19277993	4	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps
4	3438546	19277993	1	Nashville DevOps Meetup	502	34	Tech	183378188	NashDevOps

3.3. Lấy giá trị trung bình số lần tham gia sự kiện

Lấy giá trị trung bình số lần tham gia sự kiện của các thành viên theo tên nhóm và tạo một dataframe **attendings**:

```
### Tạo 1 dataframe attendings
attendings = pd.DataFrame(df.groupby('group_name')['weight'].mean())
attendings.head()
```

group_name	weight
Agile Nashville User Group	2.885117
All Things Angular	1.954023
BIG FUN GAME nite	2.000000
Blockchain Technology Disrupting Healthcare	1.000000
Brentwood Artificial Intelligence Meetup	1.257143

Giá trị '*weight*' bây giờ là giá trị trung bình số lần tham gia sự kiện của các thành viên trong nhóm.

3.4. Lấy giá trị tổng số lần tham gia sự kiện

Đếm tổng số lần tham gia sự kiện của các thành viên theo tên nhóm và thêm giá trị này vào bảng **attendings** ở trên:

```
attendings['num of attending'] = pd.DataFrame(df.groupby('group_name')['weight'].count())
attendings.head()
```

group_name	mean_attending	num of attending
Agile Nashville User Group	2.885117	383
All Things Angular	1.954023	174
BIG FUN GAME nite	2.000000	25
Blockchain Technology Disrupting Healthcare	1.000000	22
Brentwood Artificial Intelligence Meetup	1.257143	35

3.5. Thực hiện đề xuất nhóm tương tự

Trước tiên chúng ta cần tạo một ma trận *group_matrix* trong đó *member_id* sẽ nằm trên một trục và trục còn lại là thông tin tên nhóm.Mỗi ô sẽ là số lần mà thành viên đó tham dự sự kiện trong nhóm trên.

```
group_matrix = df.pivot_table(index='member_id',columns='group_name',values='weight')
group_matrix.head()
```

group_name	Agile Nashville User Group	All Things Angular	BIG FUN GAME nite	Blockchain Technology Disrupting Healthcare	Brentwood Artificial Intelligence Meetup	Code for Nashville	Cryptocurrency And Ecommerce	Data Science Nashville	Design Thinking Nashville	Developer Launchpad Nashville	...	Tennessee Red Hat User Group (RHUG)	The Data Warehouse Institute (TDWI) Nashville Meetup	The Y Nash
member_id														
2069	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
8386	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
17903	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	
42506	NaN	NaN	NaN	NaN	NaN	NaN	NaN	1.0	NaN	NaN	...	NaN	NaN	
76671	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	

Ta có thể thấy có rất nhiều giá trị NaN là bởi vì không phải tất cả các người dùng đều thuộc các thành viên của các nhóm trên.

- Những nhóm có số lần tham gia sự kiện của các thành viên cao nhất:

```
### Những nhóm có số lần tham gia sự kiện của các thành viên cao nhất
attendings.sort_values('num of attending',ascending=False).head(10)
```

group_name	mean_attending	num of attending
NashJS	2.475000	760
The Iron Yard - Nashville	1.882448	621
Nashville .NET User Group	2.683241	543
PyNash	2.631179	526
Nashville UX	2.335878	524
Code for Nashville	2.014737	475
Data Science Nashville	1.897619	420
Agile Nashville User Group	2.885117	383
Nashville Mobile Developers	1.967391	368
Nashville Women Programmers	2.487500	320

- Chọn nhóm 'NashJS' và lấy thông tin tham dự sự kiện của các thành viên:

```
NashJS_member_attendings = group_matrix['NashJS']
NashJS_member_attendings.head(10)
```

```
member_id
2069     NaN
8386     NaN
17903    NaN
42506    NaN
76671    NaN
128185   NaN
199844   NaN
201249   NaN
208295   NaN
247029   NaN
Name: NashJS, dtype: float64
```

- Sử dụng phương thức *corrwith()* để lấy được những tương quan với nhóm 'NashJS'

```
similar_to_NashJS = group_matrix.corrwith(NashJS_member_attendings)
```

- Xóa các giá trị NaN và sử dụng DataFrame để hiển thị sự tương quan:

```
corr_NashJS = pd.DataFrame(similar_to_NashJS,columns=['Correlation'])
corr_NashJS.dropna(inplace=True)
corr_NashJS.head()
```

group_name	Correlation
Agile Nashville User Group	0.166691
All Things Angular	0.179256
Code for Nashville	0.199901
Data Science Nashville	-0.049477
Design Thinking Nashville	-0.181353

- Sắp xếp các giá trị tương đồng:

```
corr_NashJS.sort_values('Correlation',ascending=False).head(10)
```

group_name	Correlation
NashJS	1.000000
Mindful Cyborgs Meditation	0.942809
Nashville Lisp Sync	0.837241
Murfreesboro Web Development Meetup	0.703526
The Nashville Web Design Meetup Group	0.579741
Nashville ColdFusion User Group	0.500000
NashReact	0.410800
freeCodeCamp Nashville	0.391251
Nashville .NET User Group	0.378197
PyNash	0.327531

Ta có thể thấy bên trên là 9 nhóm tương đồng với 'NashJS' được đề xuất với tỉ lệ tương đồng từ 32% đến 94%.Tương tự ta có thể làm điều này với các nhóm còn lại.

VI. Kết luận

1. Ưu điểm

Việc phân tích mạng xã hội mang đến những kiến thức mới,cũng như được hiểu thể về một mạng xã hội mới giúp ích cho mọi người.Việc thực hiện những phân tích mạng xã hội Meetup giúp chúng ta có thể hiểu kết cấu của mạng xã hội này và có thể rèn luyện và thực hành các kỹ năng phân tích một mạng xã hội với NetworkX.

Dataset **Nashville Meetup NetworkX** là một dataset cơ bản và được xây dựng sẵn các danh sách các liên kết trong mạng vì thể rất tiện lợi trong việc phân tích.

2. Nhược điểm

Những kết quả đạt được chỉ nằm ở mức cơ bản nên cũng chưa có những đánh giá xác thực và chính xác nhất về mạng xã hội Meetup.

3. Hướng phát triển đồ án

Do đồ án tập trung vào phân tích đối tượng *Group* trên Meetup nên chưa có cái nhìn tổng quát về hai đối tượng *Member* và *Event* nên việc cần làm:

- Tiếp tục phân tích hai đối tượng Member và Event trong dataset.
- Sử dụng các thuật toán mới để khám phá cộng đồng
- Dự đoán những liên kết mới trong mạng.

VII. Bảng phân công công việc

	Tìm dataset	Mô tả dữ liệu	Phân tích mạng giữa các nhóm trong dataset	Phân tích mạng giữa các nhóm trong danh mục 'Tech'	Ứng dụng-xây dựng hệ tư vấn đơn giản	Kết luận	Viết báo cáo,slide, hướng dẫn demo	Thuyết trình
Hữu	**X**	**X**	**X**	**X**	**X**	**X**	**X**	**X**

VIII. Tài liệu tham khảo

- https://towardsdatascience.com/build-your-own-recommender-system-within-5-minutes-30dd40388fbf
- https://networkx.github.io/documentation/networkx-1.9.1/index.html
- https://www.kaggle.com/stkbailey/nashville-meetup
- https://www.kaggle.com/sirpunch/meetups-data-from-meetupcom