Clustering Assignment

There will be some functions that start with the word "grader" ex: grader_actors(), grader_movies(), grader_cost1() etc, you should not change those function definition.

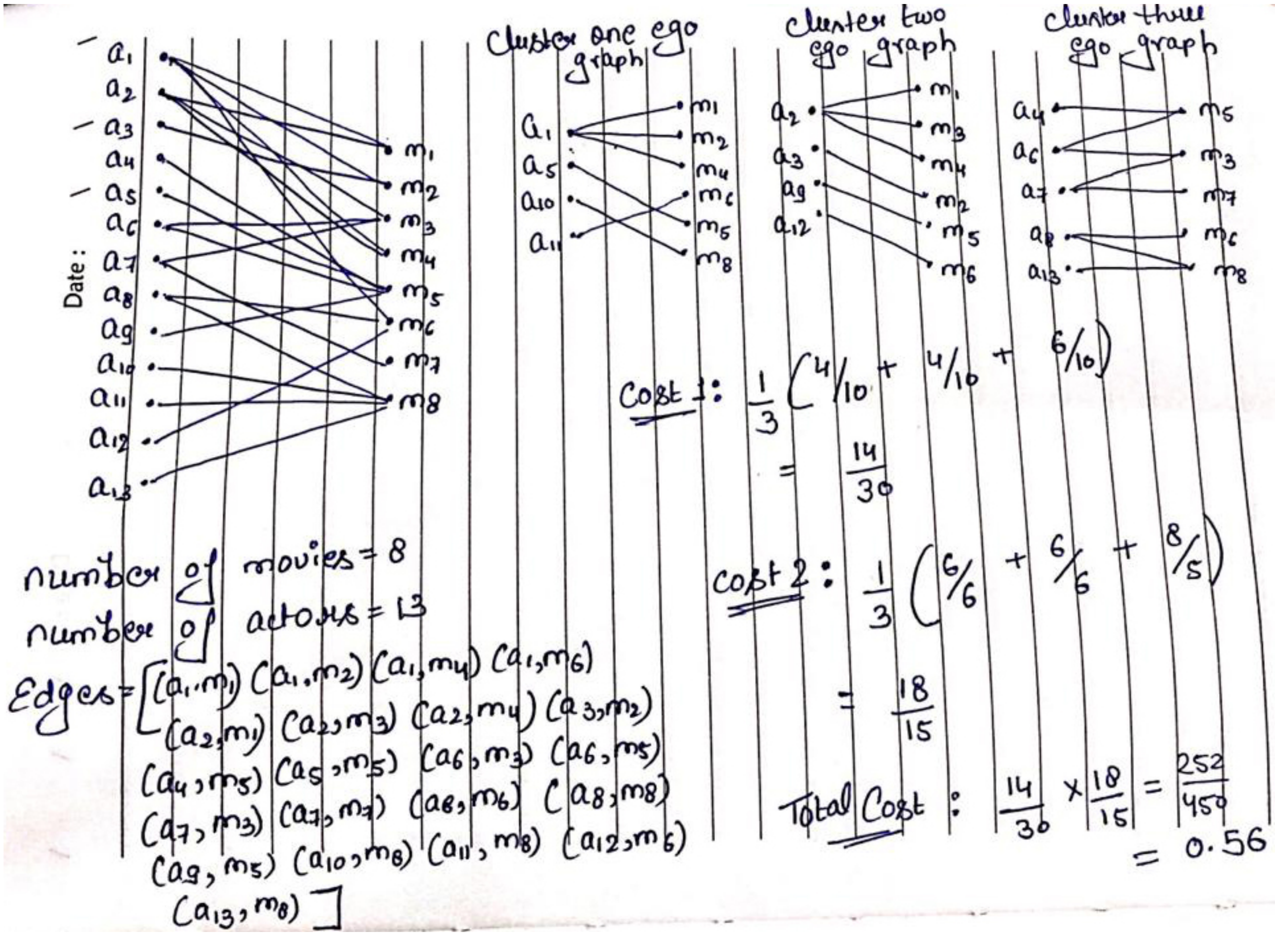
Every Grader function has to return True.

Please check clustering assignment helper functions notebook before attempting this assignment.

- Read graph from the given movie_actor_network.csv (note that the graph is bipartite graph.)
- Using stellergaph and gensim packages, get the dense representation(128dimensional vector) of every node in the graph. [Refer Clustering_Assignment_Reference.ipynb]
- Split the dense representation into actor nodes, movies nodes.(Write you code in def data_split())

Task 1 : Apply clustering algorithm to group similar actors

1. For this task consider only the actor nodes
2. Apply any clustering algorithm of your choice
   Refer : https://scikit-learn.org/stable/modules/clustering.html
3. Choose the number of clusters for which you have maximum score of $Cost1 * Cost2$
4. Cost1 = $\frac{1}{N}\sum_{\text{each cluster i}} \frac{\text{(number of nodes in the largest connected component in the graph with the actor nodes and its movie neighbours in cluster i)}}{\text{(total number of nodes in that cluster i)}}$ where N= number of clusters
   (Write your code in def cost1())
5. Cost2 = $\frac{1}{N}\sum_{\text{each cluster i}} \frac{\text{(sum of degress of actor nodes in the graph with the actor nodes and its movie neighbours in cluster i)}}{\text{(number of unique movie nodes in the graph with the actor nodes and its movie neighbours in cluster i)}}$ where N= number of clusters
   (Write your code in def cost2())
6. Fit the clustering algorithm with the opimal number_of_clusters and get the cluster number for each node
7. Convert the d-dimensional dense vectors of nodes into 2-dimensional using dimensionality reduction techniques (preferably TSNE)
8. Plot the 2d scatter plot, with the node vectors after step e and give colors to nodes such that same cluster nodes will have same color



Task 2 : Apply clustering algorithm to group similar movies

1. For this task consider only the movie nodes
2. Apply any clustering algorithm of your choice 3.Choose the number of clusters for which you have maximum score of $Cost1 * Cost2$

   Cost1 = $\frac{1}{N}\sum_{\text{each cluster i}}\frac{\text{(number of nodes in the largest connected component in the graph with the movie nodes and its actor neighbours in cluster i)}}{\text{(total number of nodes in that cluster i)}}$ where N= number of clusters
   (Write your code in def cost1())
3. Cost2 = $\frac{1}{N}\sum_{\text{each cluster i}}\frac{\text{(sum of degress of movie nodes in the graph with the movie nodes and its actor neighbours in cluster i)}}{\text{(number of unique actor nodes in the graph with the movie nodes and its actor neighbours in cluster i)}}$ where N= number of clusters
   (Write your code in def cost2())

Algorithm for actor nodes

```
    for number_of_clusters in [3, 5, 10, 30, 50, 100, 200, 500]:
        algo = clustering_algorith(clusters=number_of_clusters)
        # you will be passing a matrix of size N*d where N number of actor nodes and d is dimension
from gensim
        algo.fit(the dense vectors of actor nodes)
        You can get the labels for corresponding actor nodes (algo.labels_)
        Create a graph for every cluster(ie., if n_clusters=3, create 3 graphs)
        (You can use ego_graph to create subgraph from the actual graph)
        compute cost1,cost2
           (if n_cluster=3, cost1=cost1(graph1)+cost1(graph2)+cost1(graph3) # here we are doing
summation
            cost2=cost2(graph1)+cost2(graph2)+cost2(graph3)
        computer the metric Cost = Cost1*Cost2
    return number_of_clusters which have maximum Cost
```

In [ ]:
```
!pip install networkx==2.3
!pip install stellargraph
import networkx as nx
from networkx.algorithms import bipartite
import matplotlib.pyplot as plt
from sklearn.cluster import KMeans
import numpy as np
import warnings
warnings.filterwarnings("ignore")
import pandas as pd
# you need to have tensorflow
from stellargraph.data import UniformRandomMetaPathWalk
from stellargraph import StellarGraph
from sklearn.manifold import TSNE
```

In [ ]:
```
data=pd.read_csv('movie_actor_network.csv', index_col=False, names=['movie','actor'])
data.head()
```

Out[ ]:

	movie	actor
0	m1	a1
1	m2	a1
2	m2	a2
3	m3	a1
4	m3	a3

In [ ]:
```
edges = [tuple(x) for x in data.values.tolist()]
```

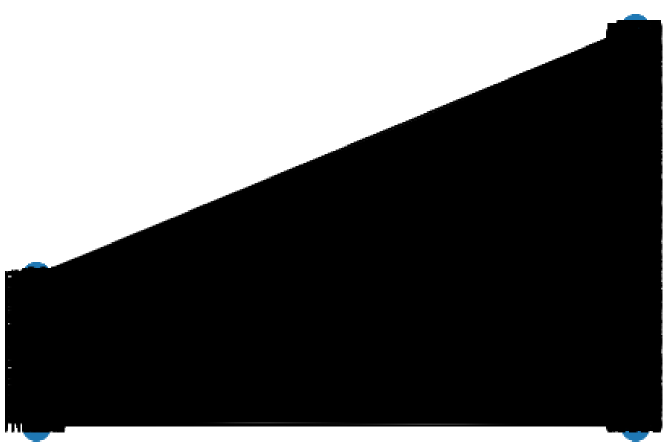
In [ ]:
```
B = nx.Graph()
B.add_nodes_from(data['movie'].unique(), bipartite=0, label='movie')
B.add_nodes_from(data['actor'].unique(), bipartite=1, label='actor')
B.add_edges_from(edges, label='acted')
```

In [ ]:
```
A = list(nx.connected_component_subgraphs(B))[0]
```

In [ ]:
```
print("number of nodes", A.number_of_nodes())
print("number of edges", A.number_of_edges())
```

```
number of nodes 4703
number of edges 9650
```

In [ ]:
```
l, r = nx.bipartite.sets(A)
pos = {}

pos.update((node, (1, index)) for index, node in enumerate(l))
pos.update((node, (2, index)) for index, node in enumerate(r))

nx.draw(A, pos=pos, with_labels=True)
plt.show()
```



In [ ]:
```
movies = []
actors = []
```

```
for i in A.nodes():
    if 'm' in i:
        movies.append(i)
    if 'a' in i:
        actors.append(i)
print('number of movies ', len(movies))
print('number of actors ', len(actors))
```

```
number of movies  1292
number of actors  3411
```

In [ ]:
```
# Create the random walker
rw = UniformRandomMetaPathWalk(StellarGraph(A))

# specify the metapath schemas as a list of lists of node types.
metapaths = [
    ["movie", "actor", "movie"],
    ["actor", "movie", "actor"]
]

walks = rw.run(nodes=list(A.nodes()), # root nodes
               length=100,  # maximum length of a random walk
               n=1,        # number of random walks per root node
               metapaths=metapaths
              )

print("Number of random walks: {}".format(len(walks)))
```

```
Number of random walks: 4703
```

In [ ]:
```
from gensim.models import Word2Vec
model = Word2Vec(walks, size=128, window=5)
```

In [ ]:
```
model.wv.vectors.shape  # 128-dimensional vector for each node in the graph
```

Out[ ]:
```
(4703, 128)
```

In [ ]:
```
# Retrieve node embeddings and corresponding subjects
node_ids = model.wv.index2word  # list of node IDs
node_embeddings = model.wv.vectors  # numpy.ndarray of size number of nodes times embeddings dimensionality
node_targets = [ A.node[node_id]['label'] for node_id in node_ids]
```

```
print(node_ids[:15], end='')
```

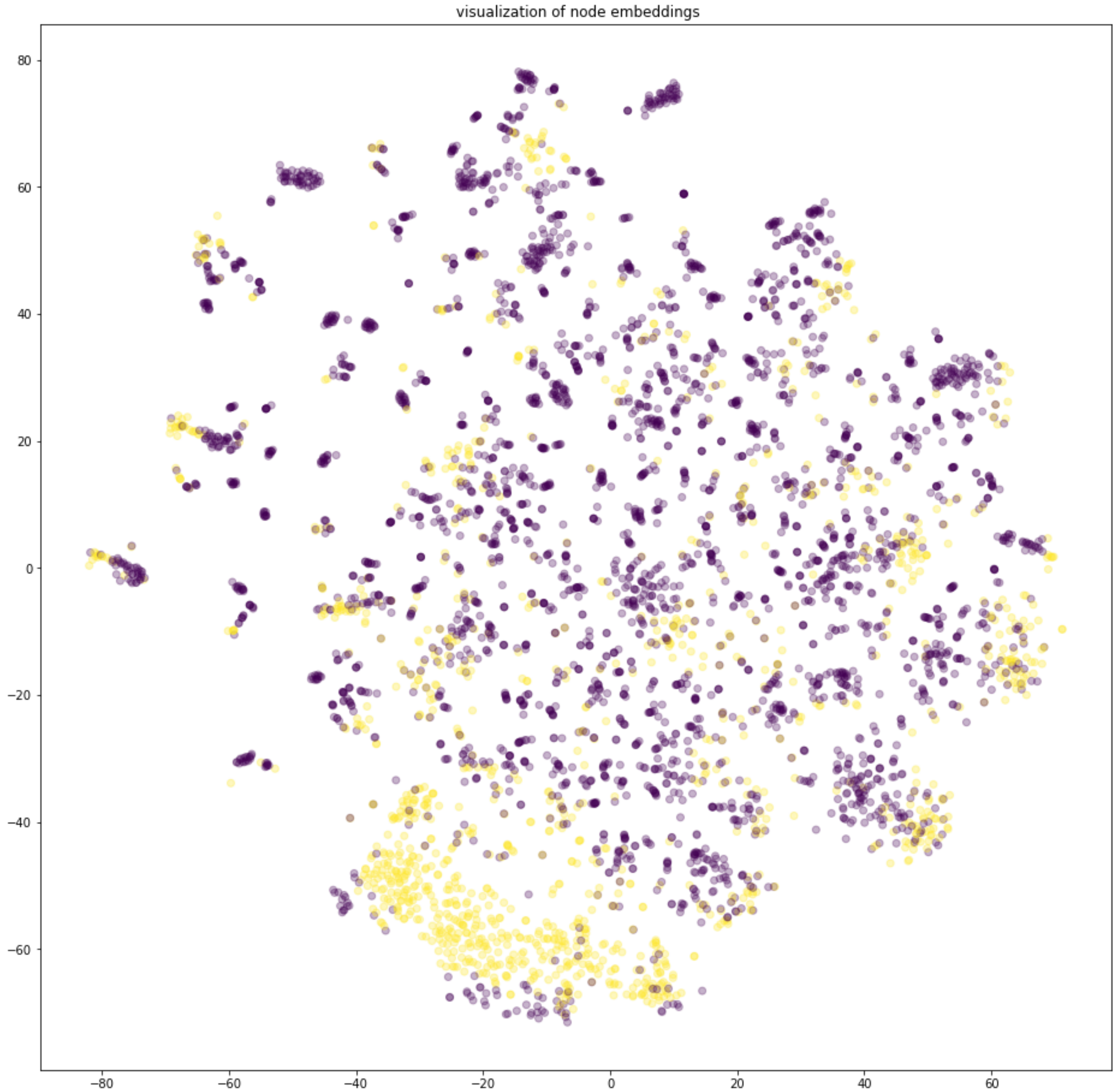
```
['a973', 'a967', 'a964', 'a1731', 'a969', 'a970', 'a1028', 'a1057', 'a965', 'a1003', 'm1094', 'a966', 'm67', 'a988', 'm1111']
```

```
print(node_targets[:15],end='')
```

```
['actor', 'actor', 'actor', 'actor', 'actor', 'actor', 'actor', 'actor', 'actor', 'actor', 'movie', 'actor', 'movie', 'actor', 'movie']
```

In [ ]:
```
from sklearn.manifold import TSNE
transform = TSNE #PCA

trans = transform(n_components=2)
node_embeddings_2d = trans.fit_transform(node_embeddings)
```

In [ ]:
```
import numpy as np
# draw the points

label_map = { l: i for i, l in enumerate(np.unique(node_targets))}
node_colours = [ label_map[target] for target in node_targets]

plt.figure(figsize=(20,16))
plt.axes().set(aspect="equal")
plt.scatter(node_embeddings_2d[:,0],
            node_embeddings_2d[:,1],
            c=node_colours, alpha=0.3)
plt.title(' visualization of node embeddings')

plt.show()
```



In [ ]:
```
def data_split(node_ids,node_targets,node_embeddings):
    '''In this function, we will split the node embeddings into actor_embeddings , movie_embeddings '''
    actor_nodes,movie_nodes=[],[]
    actor_embeddings,movie_embeddings=[],[]
    for i in range(len(node_ids)):
      if node_targets[i] == 'actor':
        actor_nodes.append(node_ids[i])
        actor_embeddings.append(node_embeddings[i])
      else:
        movie_nodes.append(node_ids[i])
        movie_embeddings.append(node_embeddings[i])

    return actor_nodes,movie_nodes,actor_embeddings,movie_embeddings
```

In [ ]:
```
actor_nodes,movie_nodes,actor_embeddings,movie_embeddings = data_split(node_ids,node_targets,node_embeddings)
```

Grader function - 1

In [ ]:
```
def grader_actors(data):
    assert(len(data)==3411)
    return True
grader_actors(actor_nodes)
```

Out[ ]: True

Grader function - 2

In [ ]:
```
def grader_movies(data):
    assert(len(data)==1292)
    return True
grader_movies(movie_nodes)
```

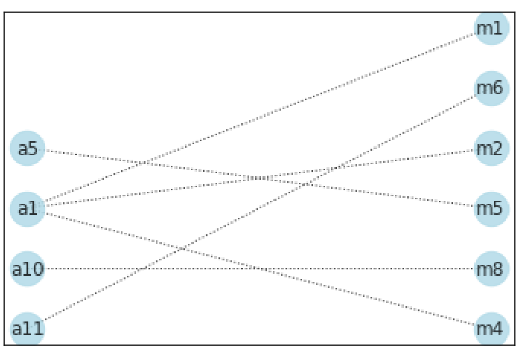
Out[ ]: True

Calculating cost1

Cost1 = $\frac{1}{N}\sum_{\text{each cluster i}}\frac{(\text{number of nodes in the largest connected component in the graph with the actor nodes and its movie neighbours in cluster i})}{(\text{total number of nodes in that cluster i})}$ where N= number of clusters

In [ ]:
```
def cost1(graph,number_of_clusters,for_actor = 0):
  lists = [ list(),list()]
  for i in graph.nodes():
    if 'm' in i:

      lists[1].append(i)
    if 'a' in i:

      lists[0].append(i)

  node_collection = []
  for j in lists[for_actor]:
    node_collection.append(graph.degree(j)+1)
  max_node = max(node_collection)

  cost = max_node/graph.number_of_nodes()

  cost1 = cost/number_of_clusters
  return cost1
```

In [ ]:
```
import networkx as nx
from networkx.algorithms import bipartite
graded_graph= nx.Graph()
graded_graph.add_nodes_from(['a1','a5','a10','a11'], bipartite=0) # Add the node attribute "bipartite"
graded_graph.add_nodes_from(['m1','m2','m4','m6','m5','m8'], bipartite=1)
graded_graph.add_edges_from([('a1','m1'),('a1','m2'),('a1','m4'),('a11','m6'),('a5','m5'),('a10','m8')])
l={'a1','a5','a10','a11'};r={'m1','m2','m4','m6','m5','m8'}
pos = {}
pos.update((node, (1, index)) for index, node in enumerate(l))
pos.update((node, (2, index)) for index, node in enumerate(r))
nx.draw_networkx(graded_graph, pos=pos, with_labels=True,node_color='lightblue',alpha=0.8,style='dotted',node_size=500)
```



Grader function - 3

In [ ]:
```
graded_cost1=cost1(graded_graph,3,0)
print(graded_cost1)
def grader_cost1(data):
    assert(data==((1/3)*(4/10))) # 1/3 is number of clusters
    return True
grader_cost1(graded_cost1)
```

```
0.13333333333333333
```

Out[ ]: True

Calculating cost2

Cost2 = $\frac{1}{N}\sum_{\text{each cluster i}}\frac{(\text{sum of degress of actor nodes in the graph with the actor nodes and its movie neighbours in cluster i})}{(\text{number of unique movie nodes in the graph with the actor nodes and its movie neighbours in cluster i})}$ where N= number of clusters

In [ ]:
```
def cost2(graph,number_of_clusters,for_actor = 1):
  lists = [ list(),list()]
  for i in graph.nodes():
    if 'm' in i:
      lists[1].append(i)
    if 'a' in i:
      lists[0].append(i)

  cost = graph.number_of_edges()/len(lists[for_actor])

  cost2 = cost/number_of_clusters
  return cost2
```

Grader function - 4

In [ ]:
```
graded_cost2=cost2(graded_graph,3,1)
print(graded_cost2)
def grader_cost2(data):
    assert(data==((1/3)*(6/6))) # 1/3 is number of clusters
    return True
grader_cost2(graded_cost2)
```

```
0.3333333333333333
```

Out[ ]: True

Grouping similar actors

In [ ]:
```
from sklearn.cluster import KMeans

# only considering actor row for similarity
index = []
for i,j in enumerate(node_ids):
  if 'a' in j:
    index.append(i)
new_id = [node_ids[i] for i in index]

metric_Cost = []
K = [3, 5, 10, 30, 50, 100, 200, 500]
for k in K:
    km_bow = KMeans(n_clusters=k, random_state=0)
    km_bow = km_bow.fit(actor_embeddings)
    labels = km_bow.labels_

    #creating dataframe with actor column and its corresponding labels
    data1 = {'unique':new_id,'labels':labels}
    df = pd.DataFrame(data1)

    Cost1 , Cost2 = 0,0
    #we are creating loop for calculating cost1 andd cost2 for each labels
    for i in range(k):
      li = []
      for j, row in df.iterrows():
        if 'a' in row['unique'] and row['labels'] == i:
          li.append(j)

      a = df.iloc[li]


      l = []
      for m , row in data.iterrows():
        if row['actor'] in a['unique'].values:
          l.append(m)
      new_data = data.iloc[l]

      edges = [tuple(x) for x in new_data.values.tolist()]
      C = nx.Graph()
      C.add_nodes_from(new_data['movie'].unique(), bipartite=0, label='movie')
      C.add_nodes_from(new_data['actor'].unique(), bipartite=1, label='actor')
      C.add_edges_from(edges, label='acted')

      Cost1 += cost1(C,k+1,for_actor = 0)

      Cost2 += cost2(C,k+1,for_actor = 1)

    metric_Cost.append(Cost1*Cost2)
```

```
metric_Cost
```

In [ ]:

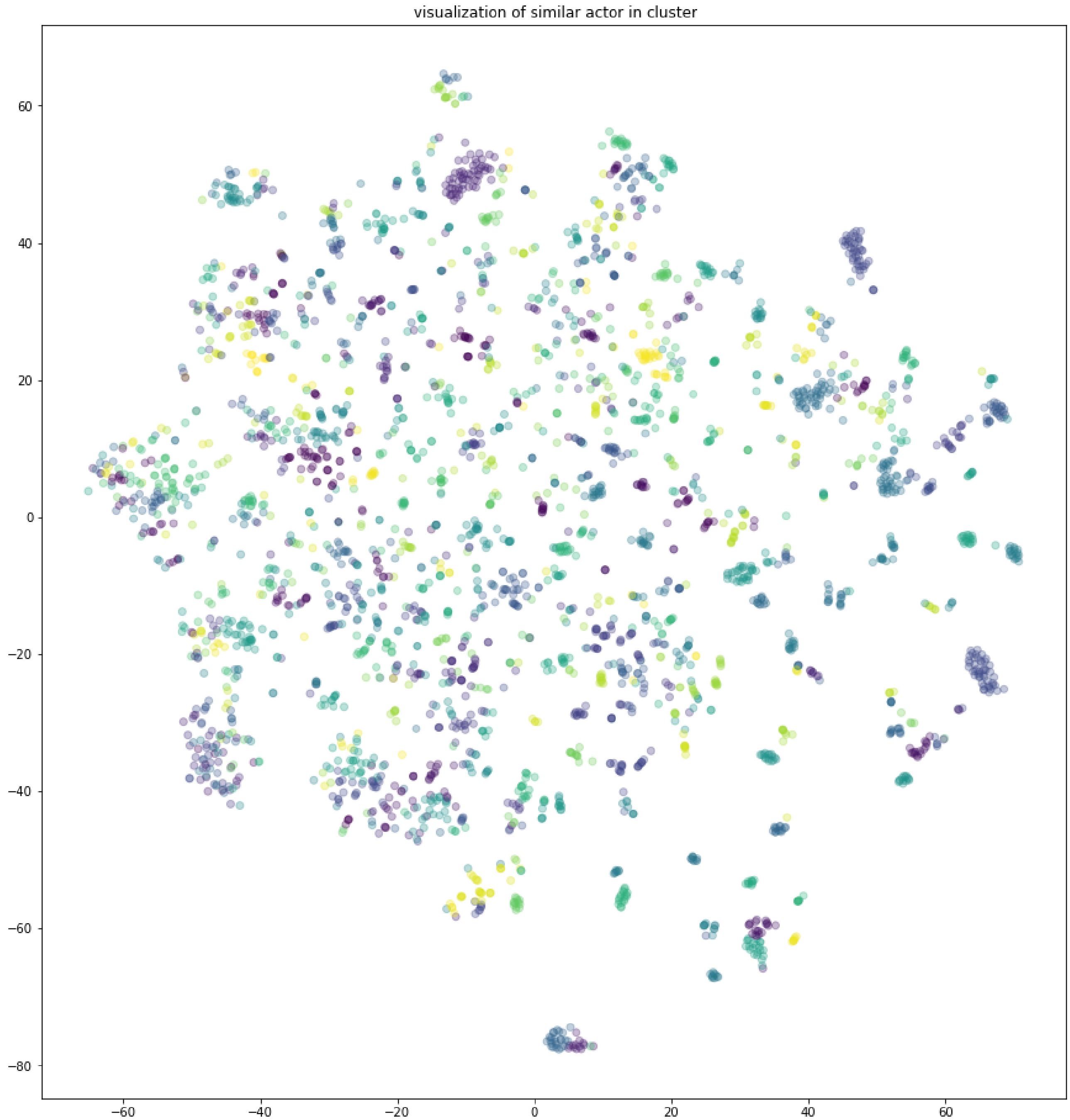
Out[ ]:
```
[0.21138872873798542,
 0.3258505595462498,
 0.33810155188220586,
 0.6074837571389242,
 0.6952940008665235,
 0.9204395458500396,
 1.0640756451275462,
 1.3164037008078986]
```

In [ ]:
```
print(f'Optimal number of cluster {K[metric_Cost.index(max(metric_Cost))]} ')
```

```
Optimal number of cluster 500
```

Displaying similar actor clusters

In [ ]:
```
km_bow = KMeans(n_clusters=500, random_state=0)
km_bow = km_bow.fit(actor_embeddings)
labels = km_bow.labels_


from sklearn.manifold import TSNE
transform = TSNE #PCA

trans = transform(n_components=2)
node_embeddings_2d = trans.fit_transform(actor_embeddings)

# draw the points

plt.figure(figsize=(20,16))
plt.axes().set(aspect="equal")
plt.scatter(node_embeddings_2d[:,0],
            node_embeddings_2d[:,1],
            c=labels, alpha=0.3)
plt.title(' visualization of similar actor in cluster')

plt.show()
```



Grouping similar movies

In [ ]:
```
# only considering movies row for similarity
index = []
for i,j in enumerate(node_ids):
  if 'm' in j:
    index.append(i)
new_id = [node_ids[i] for i in index]


metric_Cost = []
K = [3, 5, 10, 30, 50, 100, 200, 500]
for k in K:
    km_bow = KMeans(n_clusters=k, random_state=0)
    km_bow = km_bow.fit(movie_embeddings)
    labels = km_bow.labels_

    #creating dataframe with movie column and its corresponding labels
    data1 = {'unique':new_id,'labels':labels}
    df = pd.DataFrame(data1)

    Cost1 , Cost2 = 0,0
    #we are creating loop for calculating cost1 and cost2 for each labels
```

```
    for i in range(k):
      li = []
      for j, row in df.iterrows():
        if 'm' in row['unique'] and row['labels'] == i:
          li.append(j)

      a = df.iloc[li]


      l = []
      for m , row in data.iterrows():
        if row['movie'] in a['unique'].values:
          l.append(m)
      new_data = data.iloc[l]

      edges = [tuple(x) for x in new_data.values.tolist()]
      C = nx.Graph()
      C.add_nodes_from(new_data['movie'].unique(), bipartite=0, label='movie')
      C.add_nodes_from(new_data['actor'].unique(), bipartite=1, label='actor')
      C.add_edges_from(edges, label='acted')

      Cost1 += cost1(C,k+1,for_actor = 1)

      Cost2 += cost2(C,k+1,for_actor = 0)

    metric_Cost.append(Cost1*Cost2)
```

In [ ]:

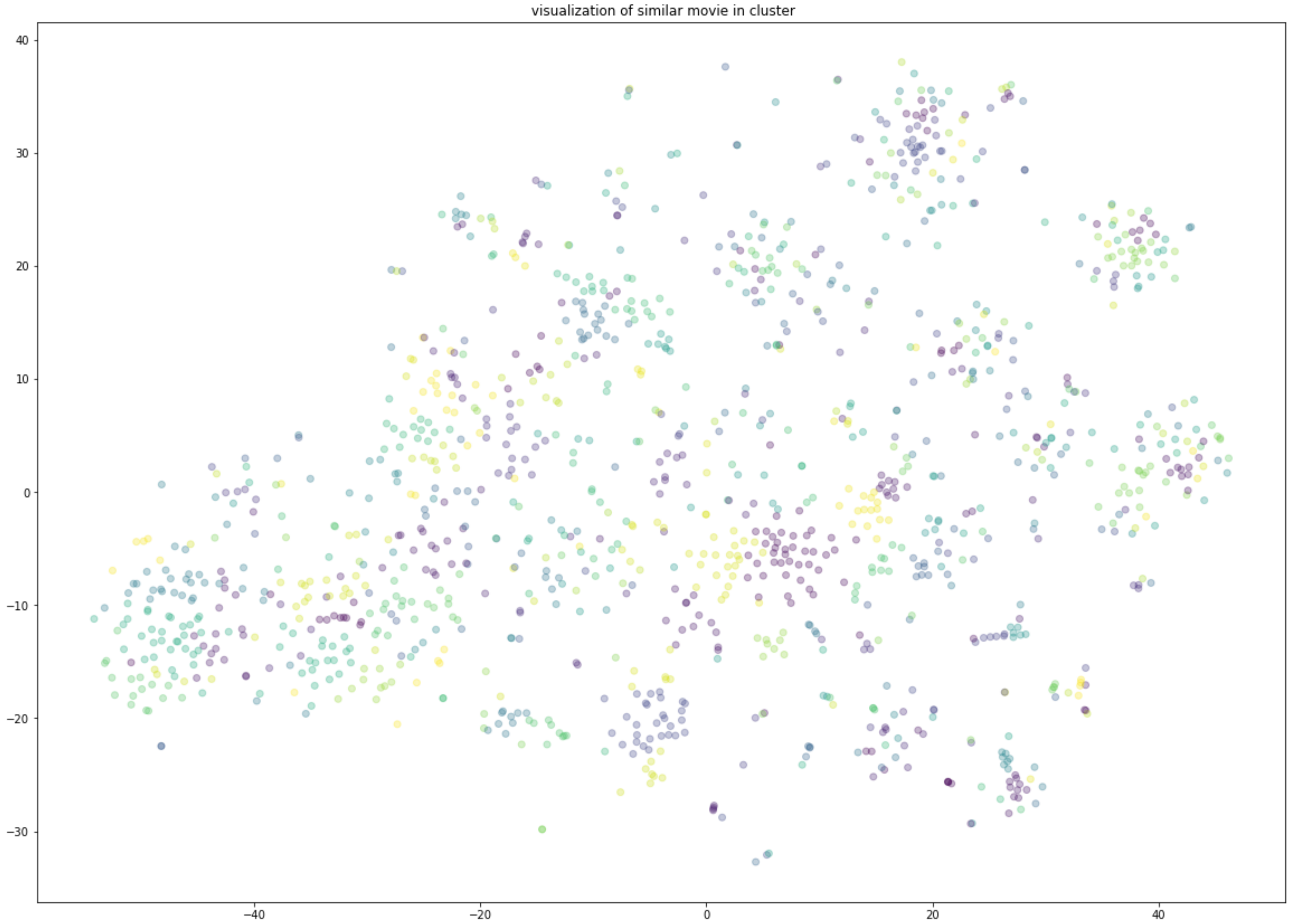
```
print(f'Optimal number of cluster {K[metric_Cost.index(max(metric_Cost))]} ')
```

```
Optimal number of cluster 500
```

Displaying similar movie clusters

In [ ]:

```
km_bow = KMeans(n_clusters=500, random_state=0)
km_bow = km_bow.fit(movie_embeddings)
labels = km_bow.labels_



transform = TSNE #PCA

trans = transform(n_components=2)
node_embeddings_2d = trans.fit_transform(movie_embeddings)

# draw the points

plt.figure(figsize=(20,16))
plt.axes().set(aspect="equal")
plt.scatter(node_embeddings_2d[:,0],
            node_embeddings_2d[:,1],
            c=labels, alpha=0.3)
plt.title(' visualization of similar movie in cluster')

plt.show()
```


visualization of similar movie in cluster
40
30
20
10
0
−10
−20
−30
−40
−20
0
20
40