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 stellergraph 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 [1]:
```python
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
```

In [2]:
```python
data=pd.read_csv('movie_actor_network.csv', index_col=False, names=['movie','actor'])
```

In [3]:
```python
edges = [tuple(x) for x in data.values.tolist()]
```

In [4]:
```python
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 [5]:
```python
A = list(nx.connected_component_subgraphs(B))[0]
```

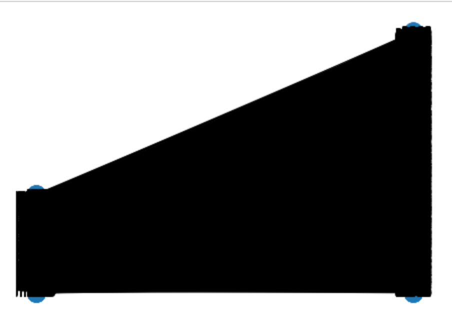
In [6]:
```python
print("number of nodes", A.number_of_nodes())
print("number of edges", A.number_of_edges())
```

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

In [7]:
```python
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 [8]:
```python
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 [8]:
```python
# 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 [9]:
```python
from gensim.models import Word2Vec
model = Word2Vec(walks, vector_size=128, window=5)
```

In [10]:
```python
model.wv.vectors.shape  # 128-dimensional vector for each node in the graph
```

Out[10]:
```
(4703, 128)
```

In [11]:
```python
# Retrieve node embeddings and corresponding subjects
node_ids = model.wv.index_to_key  # 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]
```

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

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

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

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

In [14]:
```python
def data_split(node_ids,node_targets,node_embeddings):
    '''In this function, we will split the node embeddings into actor_embeddings , movie_emb
eddings '''
    actor_nodes,movie_nodes=[],[]
    actor_embeddings,movie_embeddings=[],[]
    # split the node_embeddings into actor_embeddings,movie_embeddings based on node_ids
    # By using node_embedding and node_targets, we can extract actor_embedding and movie emb
edding
    # By using node_ids and node_targets, we can extract actor_nodes and movie nodes
    for i,x in enumerate(node_ids):
        if node_targets[i]=='actor':
            actor_nodes.append(x)

    for i,x in enumerate(node_ids):
        if node_targets[i]=='movie':
            movie_nodes.append(x)

    for i,x in enumerate(node_embeddings):
        if node_targets[i]=='actor':
            actor_embeddings.append(x)

    for i,x in enumerate(node_embeddings):
        if node_targets[i]=='movie':
            movie_embeddings.append(x)

    return actor_nodes,movie_nodes,actor_embeddings,movie_embeddings

actor_nodes,movie_nodes,actor_embeddings,movie_embeddings = data_split(node_ids,node_targets
,node_embeddings)
```

Grader function - 1

In [15]:
```python
def grader_actors(data):
    assert(len(data)==3411)
    return True
grader_actors(actor_nodes)
```

Out[15]: True

Grader function - 2

In [16]:
```python
def grader_movies(data):
    assert(len(data)==1292)
    return True
grader_movies(movie_nodes)
```

Out[16]: True

Calculating cost1

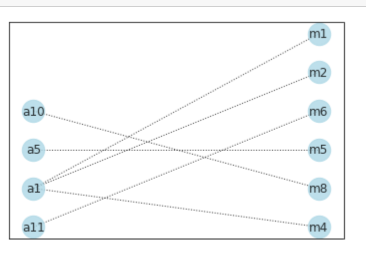
Cost1 =
$\frac{1}{N} \sum_{\text{each cluster i}} \frac{\text{(number of nodes in the largest connected component in the graph with the actor nodes and its movie neighbours in cluster i)}}{\text{(total number of nodes in that cluster i)}}$
where N= number of clusters

In [46]:
```python
def cost1(graph,number_of_clusters):
    '''In this function, we will calculate cost1'''

    largest_conn_comp = len(max(nx.connected_components(graph)))
    nodes = graph.number_of_nodes()
    cost1 = largest_conn_comp/nodes
    cost1 = cost1/number_of_clusters

    return cost1
```

In [17]:
```python
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,st
yle='dotted',node_size=500)
```



Grader function - 3

In [47]:
```python
graded_cost1=cost1(graded_graph,3)
def grader_cost1(data):
    assert(data==((1/3)*(4/10))) # 1/3 is number of clusters
    return True
grader_cost1(graded_cost1)
```

Out[47]: 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 [48]:
```python
def cost2(graph,number_of_clusters):
    '''In this function, we will calculate cost1'''

    degrees = graph.number_of_edges()
    nodes=0
    for i in graph.nodes():
        if 'm' in i:
            nodes+=1

    cost2 = (degrees/nodes)/(number_of_clusters)

    return cost2
```

Grader function - 4

In [49]:
```python
graded_cost2=cost2(graded_graph,3)
def grader_cost2(data):
    assert(data==((1/3)*(6/6))) # 1/3 is number of clusters
    return True
grader_cost2(graded_cost2)
```

Out[49]: True

Grouping similar actors

In [27]:
```python
from sklearn.cluster import KMeans
```

In [50]:
```python
number_of_clusters = [3, 5, 10, 30, 50, 100, 200, 500]
actor_embeddings = np.array(actor_embeddings)
cost_score = dict()                                                      # Dict to store cost1*co
st2

#Applying Kmeans algorithm
for cl in number_of_clusters:
    algo = KMeans(n_clusters=cl)
    algo.fit(actor_embeddings)
    list_of_clusters = []
    for i in range(cl):
        clusters=[]
        for labels,nodes in zip(algo.labels_,actor_nodes):          #algo.labels will give t
he cluster number for each nodes
            if labels==i:
                clusters.append(nodes)
        list_of_clusters.append(clusters)

    sum_cost1=0
    sum_cost2=0
    product=1

    for j in list_of_clusters:
        cost_1=0
        cost_2=0
        G=nx.Graph()                                         # Create Graph for each cluster of n
odes
        for node in j:                                       # for each node of the cluster numbe
r
            sub_graph = nx.ego_graph(B,node)                 # create subgraph and add it to the
 main graph for the cluster group
            G.add_nodes_from(sub_graph.nodes())
            G.add_edges_from(sub_graph.edges())
            cost_1 = cost1(G,cl)
            cost_2 = cost2(G,cl)
        sum_cost1 += cost_1
        sum_cost2 += cost_2
    product = sum_cost1*sum_cost2
    cost_score[cl]=product
```

In [51]:
```python
cost_score
```

Out[51]:
```
{3: 3.710905635443258,
 5: 2.4951654033728676,
 10: 2.374834147352468,
 30: 1.7613078045058368,
 50: 1.7981179986542903,
 100: 1.5984403337600437,
 200: 1.6510307569080513,
 500: 1.8223793994277457}
```

In [52]:
```python
algo=KMeans(n_clusters=3)
algo.fit(actor_embeddings)
```

Out[52]: KMeans(n_clusters=3)

In [53]:
```python
algo.labels_
```

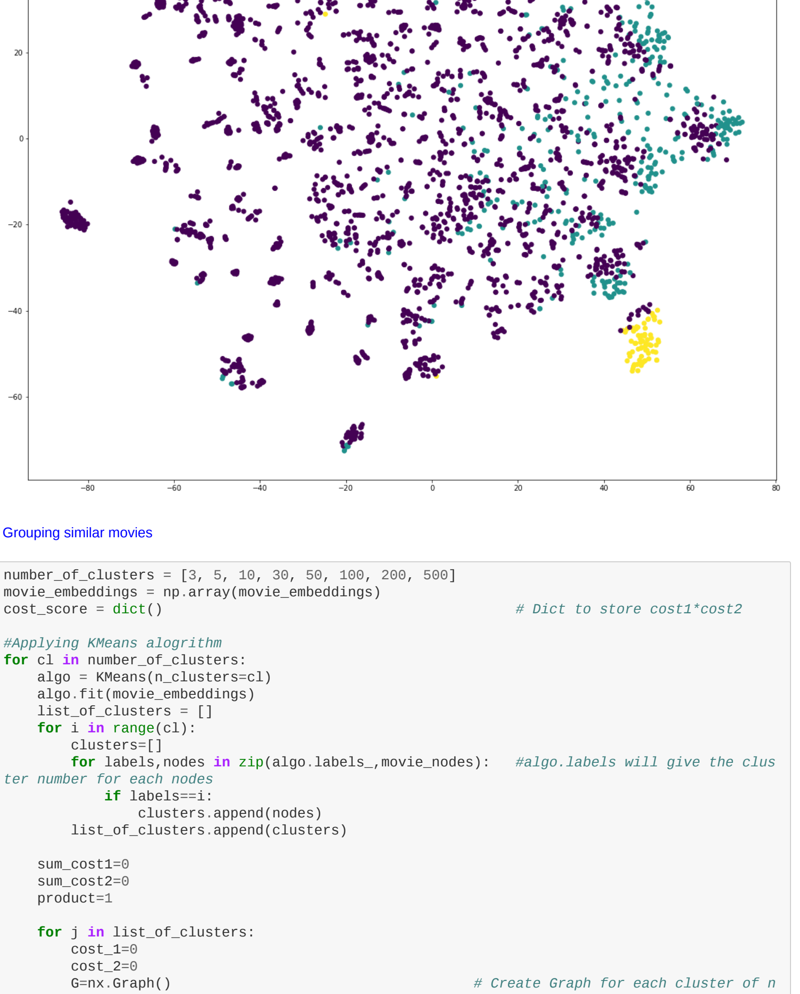
Out[53]: array([2, 2, 2, ..., 0, 0, 0])

Displaying similar actor clusters

In [54]:
```python
#Applying TSNE
from sklearn.manifold import TSNE
transform = TSNE #PCA

trans = transform(n_components=2)
actor_embeddings_2d = trans.fit_transform(actor_embeddings)
```

In [62]:
```python
actor_targets=[x for x in node_targets if x=='actor']
actor_colours = [ label_map[target] for target in actor_targets]
plt.figure(figsize=(20,16))
plt.axes().set(aspect="equal")
plt.scatter(actor_embeddings_2d[:,0],actor_embeddings_2d[:,1],c=algo.predict(actor_embedding
s))
plt.title('{} visualization of actor embeddings'.format(transform.__name__))
plt.show()
```



Grouping similar movies

In [66]:
```python
number_of_clusters = [3, 5, 10, 30, 50, 100, 200, 500]
movie_embeddings = np.array(movie_embeddings)
cost_score = dict()                                              # Dict to store cost1*cost2

#Applying KMeans alogrithm
for cl in number_of_clusters:
    algo = KMeans(n_clusters=cl)
    algo.fit(movie_embeddings)
    list_of_clusters = []
    for i in range(cl):
        clusters=[]
        for labels,nodes in zip(algo.labels_,movie_nodes):   #algo.labels will give the clus
ter number for each nodes
            if labels==i:
                clusters.append(nodes)
        list_of_clusters.append(clusters)

    sum_cost1=0
    sum_cost2=0
    product=1

    for j in list_of_clusters:
        cost_1=0
        cost_2=0
        G=nx.Graph()                                         # Create Graph for each cluster of n
odes
        for node in j:                                       # for each node of the cluster numbe
r
            sub_graph = nx.ego_graph(B,node)                 # create subgraph and add it to the
 main graph for the cluster group
            G.add_nodes_from(sub_graph.nodes())
            G.add_edges_from(sub_graph.edges())
            cost_1 = cost1(G,cl)
            cost_2 = cost2(G,cl)
        sum_cost1 += cost_1
        sum_cost2 += cost_2
    product = sum_cost1*sum_cost2
    cost_score[cl]=product
```

In [67]:
```python
cost_score
```

Out[67]:
```
{3: 8.419237566746045,
 5: 9.15481206397498,
 10: 8.924174794131906,
 30: 12.33258600943409,
 50: 13.459487140715378,
 100: 13.903572407151659,
 200: 12.927853830295758,
 500: 10.355596931930783}
```

In [68]:
```python
algo=KMeans(n_clusters=100)
algo.fit(movie_embeddings)
```

Out[68]: KMeans(n_clusters=100)

In [69]:
```python
algo.labels_
```

Out[69]: array([22, 26, 38, ..., 31, 31, 31])

Displaying similar movie clusters

In [72]:
```python
#Applying TSNE
from sklearn.manifold import TSNE
transform = TSNE #PCA

trans = transform(n_components=2)
movie_embeddings_2d = trans.fit_transform(movie_embeddings)
```

In [73]:
```python
movie_targets=[x for x in node_targets if x=='movie']
actor_colours = [ label_map[target] for target in movie_targets]
plt.figure(figsize=(20,16))
plt.axes().set(aspect="equal")
plt.scatter(movie_embeddings_2d[:,0],movie_embeddings_2d[:,1],c=algo.predict(movie_embedding
s))
plt.title('{} visualization of actor embeddings'.format(transform.__name__))
plt.show()
```