

Θεωρία Γραφημάτων και Εφαρμογές

Applied Social Network Analysis in Python

1ο Εργαστήριο

Εργασία 1:

Οκτώ υπάλληλοι σε μία μικρή εταιρεία ερωτήθηκαν να διαλέξουν τρεις ταινίες που θα ήθελαν να δουν στην ερχόμενη βραδιά ταινίας. Οι επιλογές τους υπάρχουν στο αρχείο “Employee_Movie_Choices.txt”.

Ένα δεύτερο αρχείο, “Employee_Relationships.txt”, περιέχει δεδομένα για τις σχέσεις μεταξύ των υπαλλήλων. Ο συντελεστής σχέσης έχει εύρος από -100(εχθροί) έως +100(κολλητοί). Τιμή συντελεστή 0 σημαίνει πως οι δύο υπάλληλοι δεν έχουν αλληλεπιδράσει ή έχουν ουδέτερη άποψη ο ένας για τον άλλον.

Περιεχόμενα των αρχείων:

Employee_Movie_Choices.txt:

```
#Employee      Movie
Andy   Anaconda
Andy   Mean Girls
Andy   The Matrix
Claude Anaconda
Claude Monty Python and the Holy Grail
Claude Snakes on a Plane
Frida   The Matrix
Frida   The Shawshank Redemption
Frida   The Social Network
Georgia Anaconda
Georgia Monty Python and the Holy Grail
Georgia Snakes on a Plane
Joan    Forrest Gump
Joan    Kung Fu Panda
Joan    Mean Girls
Lee     Forrest Gump
Lee     Kung Fu Panda
Lee     Mean Girls
Pablo   The Dark Knight
Pablo   The Matrix
Pablo   The Shawshank Redemption
Vincent The Godfather
Vincent The Shawshank Redemption
Vincent The Social Network
```

Employee_Relationships.txt:

```
Andy   Claude 0
Andy   Frida   20
Andy   Georgia -10
Andy   Joan    30
Andy   Lee     -10
Andy   Pablo   -10
Andy   Vincent 20
Claude Frida   0
Claude Georgia 90
Claude Joan    0
Claude Lee     0
Claude Pablo   10
Claude Vincent 0
Frida   Georgia 0
Frida   Joan    0
Frida   Lee     0
Frida   Pablo   50
Frida   Vincent 60
Georgia Joan    0
Georgia Lee     10
Georgia Pablo   0
Georgia Vincent 0
Joan    Lee     70
Joan    Pablo   0
Joan    Vincent 10
Lee     Pablo   0
Lee     Vincent 0
Pablo   Vincent -20
```

Κώδικας για την εισαγωγή των βιβλιοθηκών που θα χρειαστούμε, και για την σχεδίαση των γράφων:

```
import networkx as nx
import pandas as pd
from networkx.algorithms import bipartite

employees = set([
    'Pablo',
    'Lee',
    'Georgia',
    'Vincent',
    'Andy',
    'Frida',
    'Joan',
    'Claude'
])

movies = set([
    'The Shawshank Redemption',
    'Forrest Gump',
    'The Matrix',
    'Anaconda',
    'The Social Network',
    'The Godfather',
    'Monty Python and the Holy Grail',
    'Snakes on a Plane',
    'Kung Fu Panda',
    'The Dark Knight',
    'Mean Girls'
])

def plot_graph(G, weight_name=None):
    import matplotlib.pyplot as plt

    plt.figure()
    pos = nx.spring_layout(G)

    if weight_name:
        labels = nx.get_edge_attributes(G, weight_name)
        nx.draw_networkx_edge_labels(G, pos, edge_labels=labels)
        nx.draw_networkx(G, pos)
    else:
        nx.draw_networkx(G, pos)
```

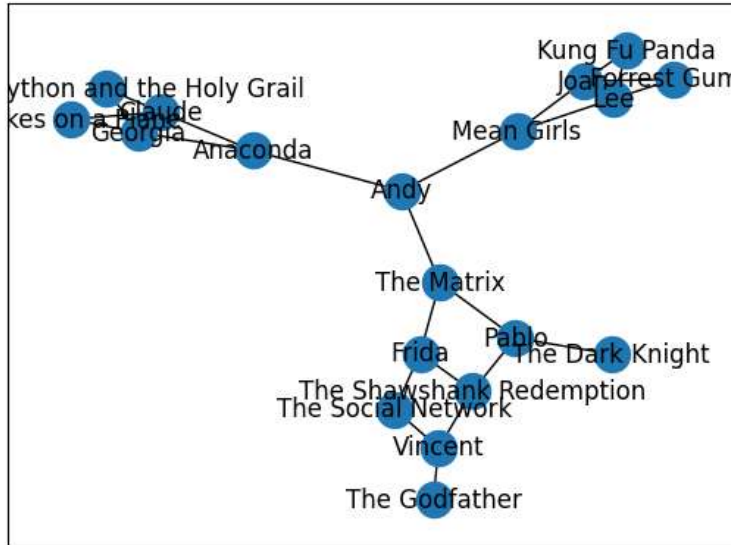
Ερώτημα 1ο:

Χρησιμοποιώντας την βιβλιοθήκη NetworkX, φορτώνουμε τον διμερή γράφο από το αρχείο Employee_Movie_Choices.txt

Κώδικας για την επίλυση του ερωτήματος:

```
g1_df = pd.read_csv('Employee_Movie_Choices.txt', sep='\t',  
                    header=None, skiprows=1, names=['Employees', 'Movies'])  
g1 = nx.from_pandas_edgelist(g1_df, 'Employees', 'Movies')  
plot_graph(g1)
```

Ο Γράφος που προκύπτει:



Πίνακας γειτνίασης του γράφου:

Andy	Anaconda
Andy	Mean Girls
Andy	The Matrix
Claude	Anaconda
Claude	Monty Python and the Holy Grail
Claude	Snakes on a Plane
Frida	The Matrix
Frida	The Shawshank Redemption
Frida	The Social Network
Georgia	Anaconda
Georgia	Monty Python and the Holy Grail
Georgia	Snakes on a Plane
Joan	Forrest Gump
Joan	Kung Fu Panda
Joan	Mean Girls
Lee	Forrest Gump
Lee	Kung Fu Panda
Lee	Mean Girls
Pablo	The Dark Knight
Pablo	The Matrix
Pablo	The Shawshank Redemption
Vincent	The Godfather
Vincent	The Shawshank Redemption
Vincent	The Social Network

Ερώτημα 2ο:

Χρησιμοποιώντας τον προηγούμενο γράφο προσθέτουμε στους κόμβους των γράφων την ιδιότητα 'type' η οποία παίρνει τιμές 'movie' ή 'value' αναλόγως με το αν ο κόμβος αντιστοιχεί σε ταινία ή υπάλληλο.

Κώδικας για την επίλυση του ερωτήματος:

```
g2 = g1
g2.add_nodes_from(employees, bipartite=0, type='employee')
g2.add_nodes_from(movies, bipartite=1, type='movie')
plot_graph(g2)
```

Καθώς δεν έγινε κάποια αλλαγή στην σχέση των κόμβων ο γράφος που προκύπτει είναι ο ίδιος με αυτόν του ερωτήματος 1.

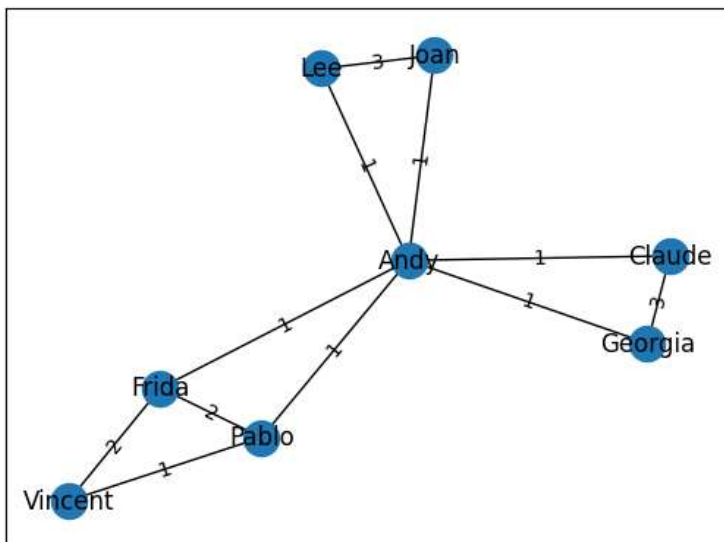
Ερώτημα 3ο:

Με βάση τον γράφο του ερωτήματος 2 βρίσκουμε τον γράφο με βάρη στις ακμές του που μας δείχνει ποιο υπάλληλοι έχουν επιλέξει κοινές ταινίες και πόσες κοινές έχει κάθε ζευγάρι υπαλλήλων. Οι υπάλληλοι με μηδέν κοινές ταινίες παραλείπονται.

Κώδικας για την επίλυση του ερωτήματος:

```
g3 = g2
weighted_projection = bipartite.weighted_projected_graph(g3, employees)
plot_graph(weighted_projection, 'weight')
```

Ο Γράφος που προκύπτει:



Πίνακας γειτνίασης και βαρών του γράφου:

Claude	Georgia	3
Claude	Andy	1
Lee	Joan	3
Lee	Andy	1
Vincent	Pablo	1
Vincent	Frida	2
Pablo	Frida	2
Pablo	Andy	1
Frida	Andy	1
Joan	Andy	1
Georgia	Andy	1

Ερώτημα 4ο:

Ας υποθέσουμε ότι θέλουμε να βρούμε εάν τα άτομα που έχουν μεγάλο συντελεστή φιλίας έχουν και κοινά γούστα σε ταινίες. Βρίσκουμε τον συντελεστή Pearson correlation (με χρήση της συνάρτησης DataFrame.corr()) μεταξύ των υπαλλήλων και τις κοινές ταινίες που επέλεξαν. Άμα δύο υπάλληλοι δεν έχουν κοινές ταινίες τότε δεν παραλείπουμε αυτή την σχέση από τον υπολογισμό μας αλλά θεωρούμε την τιμή εώς 0.

Κώδικας για την επίλυση του ερωτήματος:

```
g4 = pd.DataFrame(weighted_projection.edges(data=True),
                  columns=['Emp1', 'Emp2', 'movies_in_common'])
g4['movies_in_common'] = g4['movies_in_common'].map(lambda x: x['weight'])
g4_duplicate = g4.copy()
g4_duplicate.rename(columns={'Emp1': 'Emp2', 'Emp2': 'Emp1'}, inplace=True)
g4_common_movies = pd.concat([g4, g4_duplicate], ignore_index=True)
g4_employee_rel = pd.read_csv('Employee_Relationships.txt', sep='\t', header=None,
                             names=['Emp1', 'Emp2', 'Relationship'])
g4_merged = pd.merge(g4_employee_rel, g4_common_movies, how='left')
g4_merged['movies_in_common'].fillna(value=0, inplace=True)
correlation_score = g4_merged['movies_in_common'].corr(g4_merged['Relationship'])

print(correlation_score)
```

Η τιμή που προκύπτει είναι: 0.7883962221733474, άρα η τιμή του συντελεστή φιλίας με τις κοινές ταινίες που έχουν υπάλληλοι μεταξύ τους συσχετίζονται περίπου στο 79%.

Πίνακας με τις τιμές των συντελεστών φιλίας και με τον αριθμό των κοινών ταινιών που έχουν:

Andy	Claude	0	1
Andy	Frida	20	1
Andy	Georgia	-10	1
Andy	Joan	30	1
Andy	Lee	-10	1
Andy	Pablo	-10	1
Andy	Vincent	20	0
Claude	Frida	0	0
Claude	Georgia	90	3
Claude	Joan	0	0
Claude	Lee	0	0
Claude	Pablo	10	0
Claude	Vincent	0	0
Frida	Georgia	0	0
Frida	Joan	0	0
Frida	Lee	0	0
Frida	Pablo	50	2
Frida	Vincent	60	2
Georgia	Joan	0	0
Georgia	Lee	10	0
Georgia	Pablo	0	0
Georgia	Vincent	0	0
Joan	Lee	70	3
Joan	Pablo	0	0
Joan	Vincent	10	0
Lee	Pablo	0	0
Lee	Vincent	0	0
Pablo	Vincent	-20	1