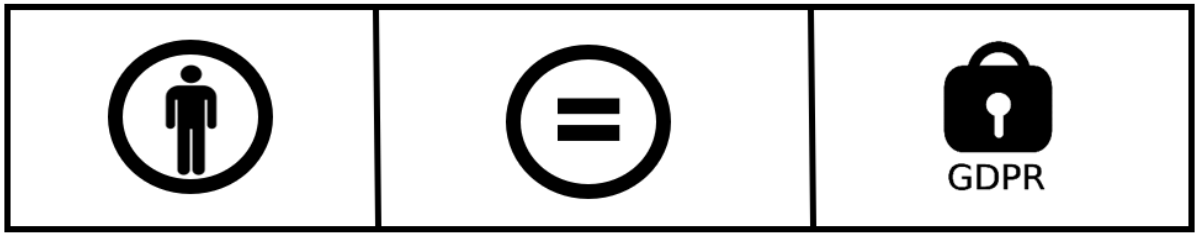
Εξόρυξη Μεγάλων Δεδομένων

2η Εργαστηριακή Άσκηση

«Διανυσματική Αναπαράσταση Γράφων και Οπτικοποίηση: Εφαρμογή σε Πρωτεϊνικά Δίκτυα»

Βίννη Παναγιώτα
Α.Μ. : 170
Εξάμηνο: 2ο
penyvinni@gmail.com

Πίνακας Περιεχομένων



Στόχος

Στην παρούσα άσκηση, θα εξερευνήσουμε την εφαρμογή του αλγορίθμου node2vec για την ανάλυση δικτύων πρωτεϊνών και την δημιουργία αναπαραστάσεων χαρακτηριστικών από αυτά τα δίκτυα. Θα χρησιμοποιήσουμε τη μέθοδο t-SNE για τη μείωση της διάστασης των embeddings και την οπτικοποίηση τους σε δισδιάστατο/τρισδιαστατο χώρο, παρέχοντας μια οπτική εικόνα των δομικών και λειτουργικών σχέσεων των πρωτεϊνών. Οι βιολογικές δομές δεδομένων που θα χρησιμοποιηθούν περιλαμβάνουν αλληλεπιδράσεις πρωτεϊνών από τη Βάση Δεδομένων Αναφοράς Ανθρώπινων Πρωτεϊνών (HPRD) και δομικές πληροφορίες από την Τράπεζα Δεδομένων Πρωτεϊνών (PDB).

- Να κατανοήσετε την εφαρμογή του node2vec στην παραγωγή αναπαραστάσεων χαρακτηριστικών από δίκτυα γράφων πρωτεϊνών.
- Να υλοποιήσετε το t-SNE για τη μείωση της διαστατικότητας των ενσωματωμάτων για λόγους οπτικοποίησης.
- Να αναλύσετε και να ερμηνεύσετε τα οπτικά μοτίβα στα δεδομένα αλληλεπίδρασης πρωτεϊνών για να διατυπώσετε υποθέσεις για βιολογικές λειτουργίες και αλληλεπιδράσεις.
- Να εξερευνήσετε πραγματικά σύνολα δεδομένων αλληλεπιδράσεων πρωτεϊνών και δομικών πληροφοριών.

Περιγραφή Άσκησης

1. **Φόρτωση και Προετοιμασία Δικτύων Πρωτεϊνών:** Αρχικά, θα φορτώσετε τα δεδομένα από τις βάσεις δεδομένων όπως η Human Protein Reference Database (HPRD) ή η Protein Data Bank (PDB) και θα προετοιμάσετε τα δίκτυα για ανάλυση, μετατρέποντας τις δομικές πληροφορίες σε Γράφους.
2. **Διανυσματοποίηση των Κόμβων με το node2vec:** Στη συνέχεια, θα χρησιμοποιήσετε το node2vec για να μετατρέψετε τους κόμβους των γράφων πρωτεϊνών σε διανύσματα. Αυτή η διαδικασία περιλαμβάνει την εκμάθηση ενσωματώσεων που αντικατοπτρίζουν τη δομική και λειτουργική πληροφορία των πρωτεϊνών.
3. **Εφαρμογή της Μεθόδου t-SNE για Μείωση Διαστάσεων:** Μετά τη διανυσματοποίηση, θα εφαρμόσετε τη μέθοδο t-SNE για να μειώσετε τις διαστάσεις των διανυσμάτων που δημιουργήθηκαν από το node2vec, με στόχο τη δημιουργία ενός δισδιάστατου χάρτη των πρωτεϊνών για ευκολότερη οπτική ανάλυση.
4. **Οπτικοποίηση των Δεδομένων:** Τέλος, θα οπτικοποιήσετε τα δεδομένα στον νέο χώρο διαστάσεων που δημιουργήθηκε από τη μέθοδο t-SNE. Αυτή η οπτικοποίηση θα σας βοηθήσει να κατανοήσετε καλύτερα τυχόν πρότυπα, συσχετίσεις ή δομικές ανωμαλίες μεταξύ των πρωτεϊνών, διευκολύνοντας την ερμηνεία βιολογικών λειτουργιών.

Επεξήγηση του κώδικα

1. **Φόρτωση δεδομένων:**

   - Φορτώνει τις απαραίτητες βιβλιοθήκες: Pandas, NetworkX, Node2Vec, TSNE, NumPy, Matplotlib.
   - Χρησιμοποιεί το Google Colab για να συνδεθεί με το Google Drive.

2. **Επιλογή και φόρτωση δεδομένων:**

   - Ζητάει από τον χρήστη να επιλέξει ανάμεσα σε δύο αρχεία: BRCA ή LEUK.
   - Φορτώνει τον γράφο από ένα αρχείο TSV χρησιμοποιώντας τις βιβλιοθήκες Pandas και NetworkX.

3. **Εκπαίδευση μοντέλου Node2Vec:**

   - Ορίζει ένα αντικείμενο Node2Vec με συγκεκριμένες παραμέτρους.
   - Εκπαιδεύει το μοντέλο χρησιμοποιώντας τους γράφους που φορτώθηκαν προηγουμένως.

4. **Ενσωματώσεις κόμβων:**

   - Αποθηκεύει τις ενσωματώσεις κόμβων σε ένα λεξικό.

5. **Μείωση διαστάσεων με t-SNE:**

   - Μετατρέπει τις ενσωματώσεις κόμβων σε πίνακα NumPy.
   - Εφαρμόζει τον αλγόριθμο t-SNE για να μειώσει τις διαστάσεις σε 2, προετοιμάζοντάς τις για οπτικοποίηση.

6. **Οπτικοποίηση:**

   - Δημιουργεί ένα scatter plot για να οπτικοποιήσει τις μειωμένες διαστάσεις των ενσωματώσεων.
   - Προαιρετικά, προσθέτει ετικέτες στους κόμβους του γράφου στο scatter plot.

Ο Αλγόριθμος Node2Vec

Ο αλγόριθμος **Node2Vec** είναι ένας αλγόριθμος ενσωμάτωσης κόμβων γράφου που αποτελεί μια επέκταση του πιο γνωστού αλγορίθμου **Word2Vec** για ενσωμάτωση λέξεων. Μερικά σημαντικά στοιχεία σχετικά με τον αλγόριθμο Node2Vec είναι τα εξής:

1. **Σκοπός**
   - Ο σκοπός του Node2Vec είναι να μετατρέψει κάθε κόμβο ενός γράφου σε ένα διάνυσμα χαμηλής διάστασης (ενσωμάτωση) έτσι ώστε να αντιπροσωπεύει τις σχέσεις του με άλλους κόμβους στο γράφο.

2. **Ιδέα**
   - Ο αλγόριθμος βασίζεται στην ιδέα της περιήγησης σε γράφους με σκοπό την εξαγωγή συναφών δεδομένων. Χρησιμοποιείται για να παράγει ενσωματώσεις κόμβων που διατηρούν τις δομικές και λειτουργικές συσχετίσεις των κόμβων σε ένα γράφο.

3. **Μέθοδος**
   - Ο αλγόριθμος Node2Vec χρησιμοποιεί μια προσαρμοσμένη διαδικασία περιήγησης στον γράφο, η οποία επιτρέπει στους κόμβους να περιηγούνται μεταξύ των γειτονικών κόμβων με διάφορους τρόπους.
   - Χρησιμοποιεί έναν παραμετροποιημένο κατανεμημένο τύπο περιήγησης που επιτρέπει στους κόμβους να επιλέγουν μονοπάτια με βάση την πιθανότητα. Αυτός ο παραμετροποιημένος τύπος περιήγησης ονομάζεται biased random walk.
   - Οι περιπάτοι που προκύπτουν από αυτήν την προσέγγιση καταγράφονται και χρησιμοποιούνται για την εκπαίδευση του μοντέλου.

4. **Χρήση**
   - Χρησιμοποιείται σε πολλές εφαρμογές, όπως η πρόβλεψη συνδέσεων σε κοινωνικά δίκτυα, η συσταδοποίηση και η πρόταση περιεχομένου σε κοινωνικά δίκτυα και η ανάλυση βιολογικών δικτύων.

Με τον αλγόριθμο **Node2Vec**, ο γράφος μπορεί να μετατραπεί σε έναν χώρο χαμηλής διάστασης, όπου οι σχέσεις μεταξύ των κόμβων μπορούν να αξιολογηθούν και να αναλυθούν ευκολότερα. Αυτό επιτρέπει στους αλγορίθμους μηχανικής μάθησης να εργαστούν πιο αποτελεσματικά πάνω σε δομικά και λειτουργικά χαρακτηριστικά των γράφων.

Ο Κώδικας Υλοποίησης

```
import pandas as pd
import networkx as nx
from node2vec import Node2Vec
from sklearn.manifold import TSNE
import numpy as np
import matplotlib.pyplot as plt
from google.colab import drive

# Attach Google Drive
drive.mount('/content/drive')

# Paths to files within Google Drive
brca_path = '/content/drive/MyDrive/Colab Notebooks/brca.tsv'
leuk_path = '/content/drive/MyDrive/Colab Notebooks/leuk.tsv'

# Question about file selection
file_choice = input("Select the file you want to run ('brca' or 'leuk'):
").strip().lower()

if file_choice == 'brca':
    file_path = brca_path
elif file_choice == 'leuk':
    file_path = leuk_path
else:
    raise ValueError("Incorrect file selection. Please select 'brca' or
'leuk'.")

def load_graph_from_tsv(file_path):
    G_df = pd.read_csv(file_path, sep='\t')
    G = nx.from_pandas_edgelist(G_df, source=G_df.columns[0],
target=G_df.columns[1])
    return G

# Example of use
protein_graph = load_graph_from_tsv(file_path)

# Create Node2Vec object
node2vec = Node2Vec(protein_graph, dimensions=64, walk_length=30,
num_walks=200, workers=4)

# Training the model
model = node2vec.fit(window=10, min_count=1, batch_words=4)

# Storage of embodiments
```

```
embeddings = {str(node): model.wv[str(node)] for node in
protein_graph.nodes()}

# Print some of the vectors
print("Some of the vectors generated by node2vec:")
for i, (node, vector) in enumerate(embeddings.items()):
    print(f"Node {node}: {vector}")
    if i == 4:  # Print only the first 5 vectors
        break

# Conversion of integrations into a table
embedding_values = np.array(list(embeddings.values()))

# Applying t-SNE
tsne = TSNE(n_components=2, random_state=42)
embeddings_2d = tsne.fit_transform(embedding_values)

# Create visualisation
plt.figure(figsize=(10, 10))
plt.scatter(embeddings_2d[:, 0], embeddings_2d[:, 1], s=5)

# Add labels (optional)
for i, node in enumerate(protein_graph.nodes()):
    plt.annotate(str(node), (embeddings_2d[i, 0], embeddings_2d[i, 1]),
fontsize=8)

plt.show()
```

Ο κώδικας βρίσκεται στο παρακάτω αρχείο στο Google Collab:
https://colab.research.google.com/drive/1dgWfrQ7PPkwft4VVCFFlSc5W-Syj_fwH#scrollTo=HTvw8ay7rOyZ

Οπτικοποίηση των Αποτελεσμάτων

Η οπτικοποίηση που πραγματοποιήθηκε στον κώδικα που παρουσιάστηκε παραπάνω χρησιμοποίησε τη μείωση των διαστάσεων με τον αλγόριθμο t-distributed Stochastic Neighbor Embedding (t-SNE). Ο αλγόριθμος t-SNE μετατρέπει πολυδιάστατα δεδομένα σε δύο διαστάσεις έτσι ώστε να διατηρείται η δομή των δεδομένων όσο το δυνατόν περισσότερο. Συγκεκριμένα, η διαδικασία οπτικοποίησης περιλαμβάνει τα ακόλουθα βήματα:

1. **Μετατροπή των ενσωματώσεων κόμβων σε πίνακα:** Οι ενσωματώσεις κόμβων που παράχθηκαν από τον αλγόριθμο Node2Vec αποθηκεύονται σε έναν πίνακα NumPy.

2. **Εφαρμογή του αλγορίθμου t-SNE:** Ο αλγόριθμος t-SNE εφαρμόζεται στον πίνακα των ενσωματώσεων κόμβων για να μειώσει τις διαστάσεις από τον αρχικό χώρο υψηλής διάστασης σε έναν χώρο δύο διαστάσεων.

3. **Οπτικοποίηση με Scatter Plot:** Τα δεδομένα που μειώθηκαν σε δύο διαστάσεις απεικονίζονται σε ένα scatter plot, όπου κάθε σημείο αντιπροσωπεύει έναν κόμβο.

4. **Προσθήκη Ετικετών:** Προαιρετικά, μπορούν να προστεθούν ετικέτες σε κάθε σημείο του scatter plot για να δείξουν την ταυτότητα του αντίστοιχου κόμβου.

Καρκίνος του Μαστού (BRCA):

- Η οπτικοποίηση των δεδομένων για τον καρκίνο του μαστού μπορεί να αποκαλύψει τις δομικές και λειτουργικές σχέσεις μεταξύ των πρωτεϊνών που σχετίζονται με την πάθηση.

- Οι ενσωματώσεις κόμβων μπορούν να απεικονίσουν τις αλληλεπιδράσεις και τις συνδέσεις μεταξύ των γονιδίων και των βιολογικών παραμέτρων που σχετίζονται με τον καρκίνο του μαστού.

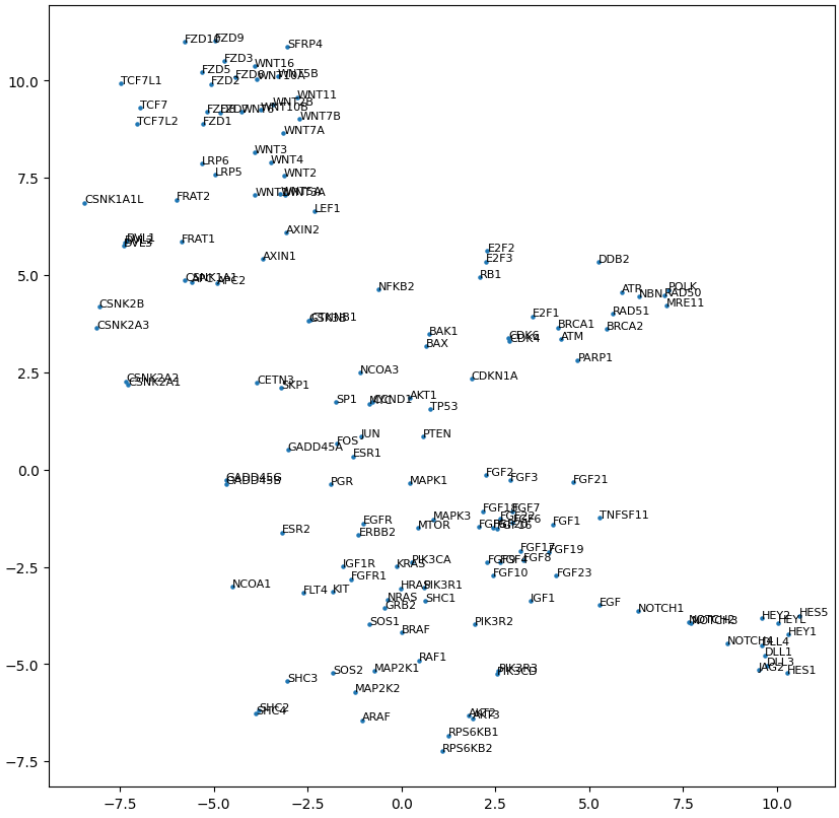
- Μια οπτικοποίηση των ενσωματώσεων κόμβων μπορεί να αποκαλύψει τη δομή του δικτύου πρωτεϊνών και τη συνδεσιμότητα μεταξύ των βιολογικών στοιχείων που μπορεί να επηρεάζουν τον καρκίνο του μαστού.

Λευχαιμία (LEUK):

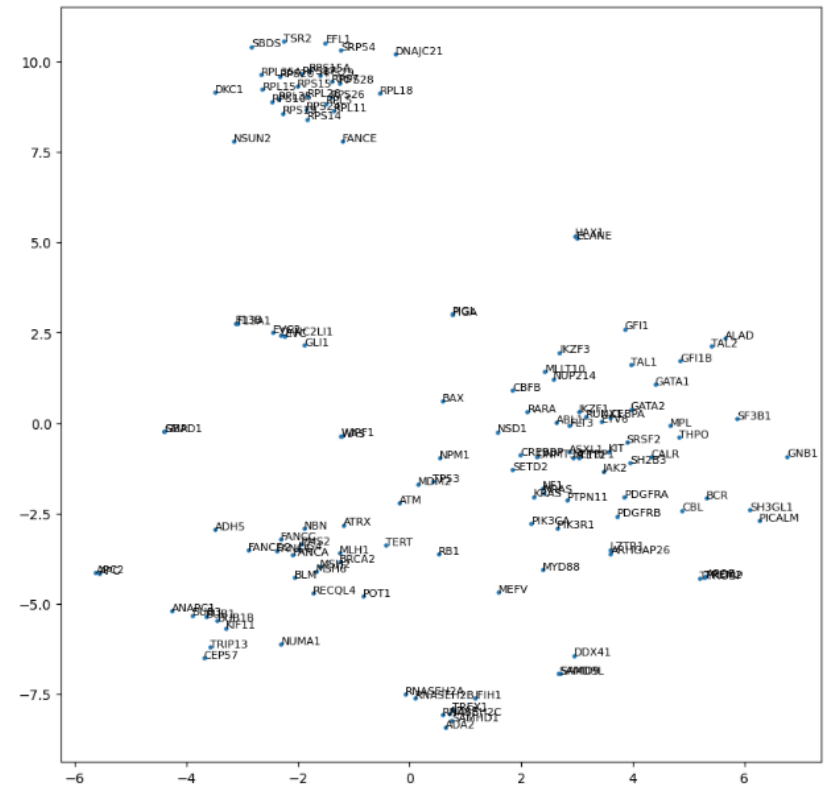
- Η οπτικοποίηση των δεδομένων για τη λευχαιμία μπορεί να βοηθήσει στην κατανόηση των μηχανισμών που βρίσκονται πίσω από τη νόσο και την αλληλεπίδραση μεταξύ των γονιδίων και των πρωτεϊνών που σχετίζονται με αυτήν.

- Οι ενσωματώσεις κόμβων μπορούν να αποτυπώσουν τις πολύπλοκες διασυνδέσεις μεταξύ των γονιδίων και των βιολογικών διαδρόμων που σχετίζονται με τη λευχαιμία.

- Μια οπτικοποίηση των ενσωματώσεων κόμβων μπορεί να αναδείξει την δομή του δικτύου πρωτεϊνών και να αποτυπώσει τις πιθανές πορείες που οδηγούν στην ανάπτυξη της λευχαιμίας.



Εικόνα 1: Οπτικοποίηση των δεδομένων καρκίνου του μαστού



Εικόνα 2: Οπτικοποίηση των δεδομένων λευχαιμίας

Συμπεράσματα

- Ο κώδικας προσφέρει έναν απλό τρόπο για την ανάλυση γράφων χρησιμοποιώντας τον αλγόριθμο Node2Vec. Αυτός ο αλγόριθμος επιτρέπει τη μετατροπή των κόμβων ενός γράφου σε ενσωματώσεις διανυσμάτων, που μπορούν να χρησιμοποιηθούν για ανάλυση και οπτικοποίηση.

- Οι ενσωματώσεις κόμβων εκπαιδεύτηκαν χρησιμοποιώντας τον αλγόριθμο Node2Vec σε ένα δεδομένο γράφο που περιέχει πληροφορίες για τον καρκίνο του μαστού (BRCA) ή τη λευχαιμία (LEUK).

- Τα ενσωματώσεις κόμβων μειώθηκαν σε δύο διαστάσεις χρησιμοποιώντας τον αλγόριθμο t-SNE, προκειμένου να επιτρέψουν την οπτικοποίηση σε ένα scatter plot.

- Η οπτικοποίηση των δεδομένων παρέχει μια εικόνα των σχέσεων μεταξύ των κόμβων στο γράφο. Αυτό μπορεί να βοηθήσει στην ανάλυση των δομικών και λειτουργικών συσχετίσεων στον γράφο και στην αναγνώριση πιθανών προτύπων ή δομών.

- Ο κώδικας μπορεί να προσαρμοστεί εύκολα για την ανάλυση και οπτικοποίηση διαφορετικών γράφων, ανάλογα με τις ανάγκες και τα δεδομένα του χρήστη.