Node Embeddings

This notebook demonstrates different methods for node embeddings and how to further reduce their dimensionality to be able to visualize them in a 2D plot.

Node embeddings are essentially an array of floating point numbers (length = embedding dimension) that can be used as "features" in machine learning. These numbers approximate the relationship and similarity information of each node and can also be seen as a way to encode the topology of the graph.

Considerations

Due to dimensionality reduction some information gets lost, especially when visualizing node embeddings in two dimensions. Nevertheless, it helps to get an intuition on what node embeddings are and how much of the similarity and neighborhood information is retained. The latter can be observed by how well nodes of the same color and therefore same community are placed together and how much bigger nodes with a high centrality score influence them.

If the visualization doesn't show a somehow clear separation between the communities (colors) here are some ideas for tuning:

- Clean the data, e.g. filter out very few nodes with extremely high degree that aren't actually that important
- Try directed vs. undirected projections
- Tune the embedding algorithm, e.g. use a higher dimensionality
- Tune t-SNE that is used to reduce the node embeddings dimension to two dimensions for visualization.

It could also be the case that the node embeddings are good enough and well suited the way they are despite their visualization for the down stream task like node classification or link prediction. In that case it makes sense to see how the whole pipeline performs before tuning the node embeddings in detail.

Note about data dependencies

PageRank centrality and Leiden community are also fetched from the Graph and need to be calculated first. This makes it easier to see if the embeddings approximate the structural information of the graph in the plot. If these properties are missing you will only see black dots all of the same size.

References

- jqassistant

- Neo4j Python Driver
- Tutorial: Applied Graph Embeddings
- Visualizing the embeddings in 2D
- scikit-learn TSNE
- AttributeError: 'list' object has no attribute 'shape'
- Fast Random Projection (neo4j)
- HashGNN (neo4j)
- node2vec (neo4j) computes a vector representation of a node based on second order random walks in the graph.
- Complete guide to understanding Node2Vec algorithm

```
The openTSNE version is: 1.0.1
The pandas version is 1.5.1.
```

Dimensionality reduction with t-distributed stochastic neighbor embedding (t-SNE)

The following function takes the original node embeddings with a higher dimensionality, e.g. 64 floating point numbers, and reduces them into a two dimensional array for visualization.

> It converts similarities between data points to joint probabilities and tries to minimize the Kullback-Leibler divergence between the joint probabilities of the low-dimensional embedding and the high-dimensional data.

(see https://opentsne.readthedocs.io)

1. Typescript Modules

1.1 Generate Node Embeddings for Typescript Modules using Fast Random Projection (Fast RP)

Fast Random Projection is used to reduce the dimensionality of the node feature space while preserving most of the distance information. Nodes with similar neighborhood result in node embedding with similar vectors.

👉 **Hint:** To skip existing node embeddings and always calculate them based on the parameters below edit `Node_Embeddings_0a_Query_Calculated` so that it won't return any results.

The results have been provided by the query filename: ../cypher/Node_Embeddings/Node_Embeddings_0a_Query_Calculated.cypher

	codeUnitName	shortCodeUnitName	projectName	communityId	centrality	embedding
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	react-router	0	0.871695	[0.1136832982301712, -0.14023642241954803, -0....
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	react-router-native	0	0.249387	[0.09540732204914093, -0.13414444029331207, -0...
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	react-router-dom	1	0.340235	[0.049455150961875916, -0.2577980160713196, -0...
3	/home/runner/work/code-graph-analysis-pipeline...	server	react-router-dom	1	0.249387	[0.13975170254707336, -0.15085220336914062, -0...

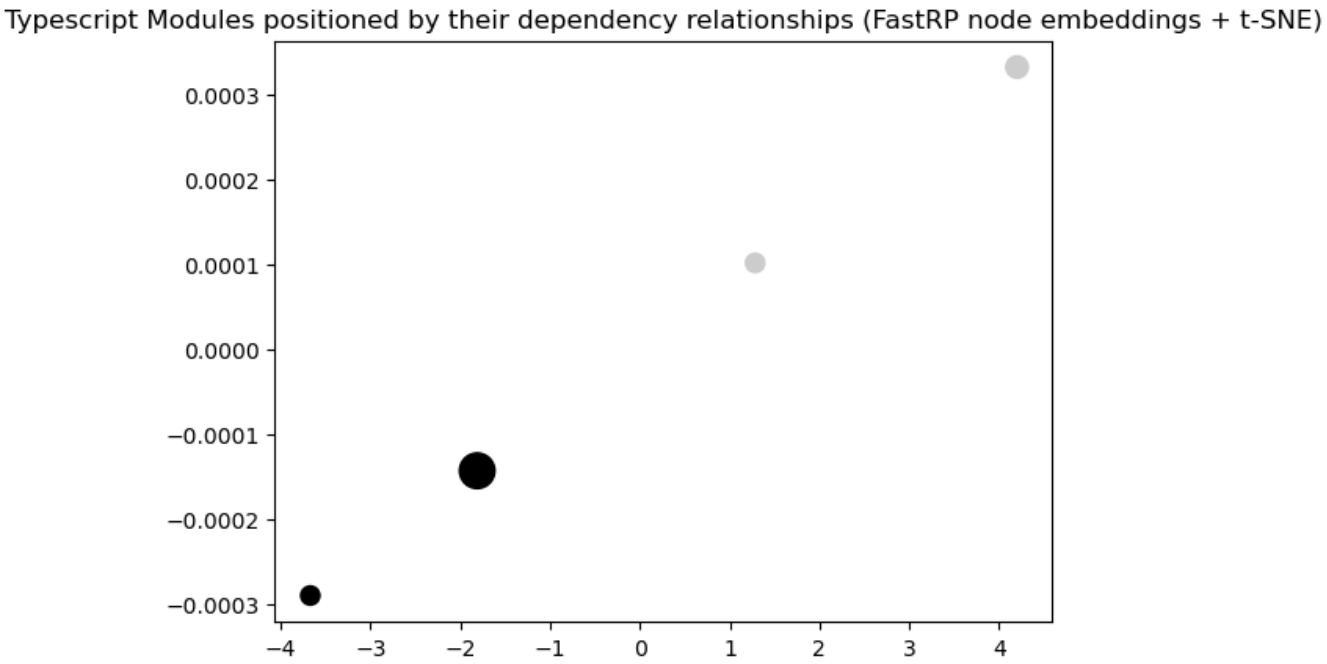
1.2 Dimensionality reduction with t-distributed stochastic neighbor embedding (t-SNE)

This step takes the original node embeddings with a higher dimensionality, e.g. 64 floating point numbers, and reduces them into a two dimensional array for visualization. For more details look up the function declaration for "prepare_node_embeddings_for_2d_visualization".

```
Perplexity value 30 is too high. Using perplexity 1.00 instead
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 3 nearest neighbors using exact search using euclidean distance...
   --> Time elapsed: 0.06 seconds
===> Calculating affinity matrix...
   --> Time elapsed: 0.00 seconds
===> Calculating PCA-based initialization...
   --> Time elapsed: 0.00 seconds
===> Running optimization with exaggeration=12.00, lr=0.33 for 250 iterations...
Iteration   50, KL divergence 0.5239, 50 iterations in 0.0070 sec
Iteration  100, KL divergence 0.9939, 50 iterations in 0.0068 sec
Iteration  150, KL divergence 0.9939, 50 iterations in 0.0063 sec
Iteration  200, KL divergence 0.9939, 50 iterations in 0.0064 sec
Iteration  250, KL divergence 0.9939, 50 iterations in 0.0063 sec
   --> Time elapsed: 0.03 seconds
===> Running optimization with exaggeration=1.00, lr=4.00 for 500 iterations...
Iteration   50, KL divergence 0.1135, 50 iterations in 0.0065 sec
Iteration  100, KL divergence 0.1151, 50 iterations in 0.0065 sec
Iteration  150, KL divergence 0.1287, 50 iterations in 0.0066 sec
Iteration  200, KL divergence 0.1236, 50 iterations in 0.0064 sec
Iteration  250, KL divergence 0.1332, 50 iterations in 0.0067 sec
Iteration  300, KL divergence 0.1343, 50 iterations in 0.0066 sec
Iteration  350, KL divergence 0.1326, 50 iterations in 0.0067 sec
Iteration  400, KL divergence 0.1335, 50 iterations in 0.0066 sec
Iteration  450, KL divergence 0.1314, 50 iterations in 0.0066 sec
Iteration  500, KL divergence 0.1310, 50 iterations in 0.0064 sec
   --> Time elapsed: 0.07 seconds
(4, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	0	0.871695	-1.813226	-0.000143
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	0	0.249387	-3.672437	-0.000290
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.340235	4.201760	0.000332
3	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.249387	1.283902	0.000102

1.3 Plot the node embeddings reduced to two dimensions for Typescript



1.4 Node Embeddings for Typescript Modules using HashGNN

HashGNN resembles Graph Neural Networks (GNN) but does not include a model or require training. It combines ideas of GNNs and fast randomized algorithms. For more details see HashGNN. Here, the latter 3 steps are combined into one for HashGNN.

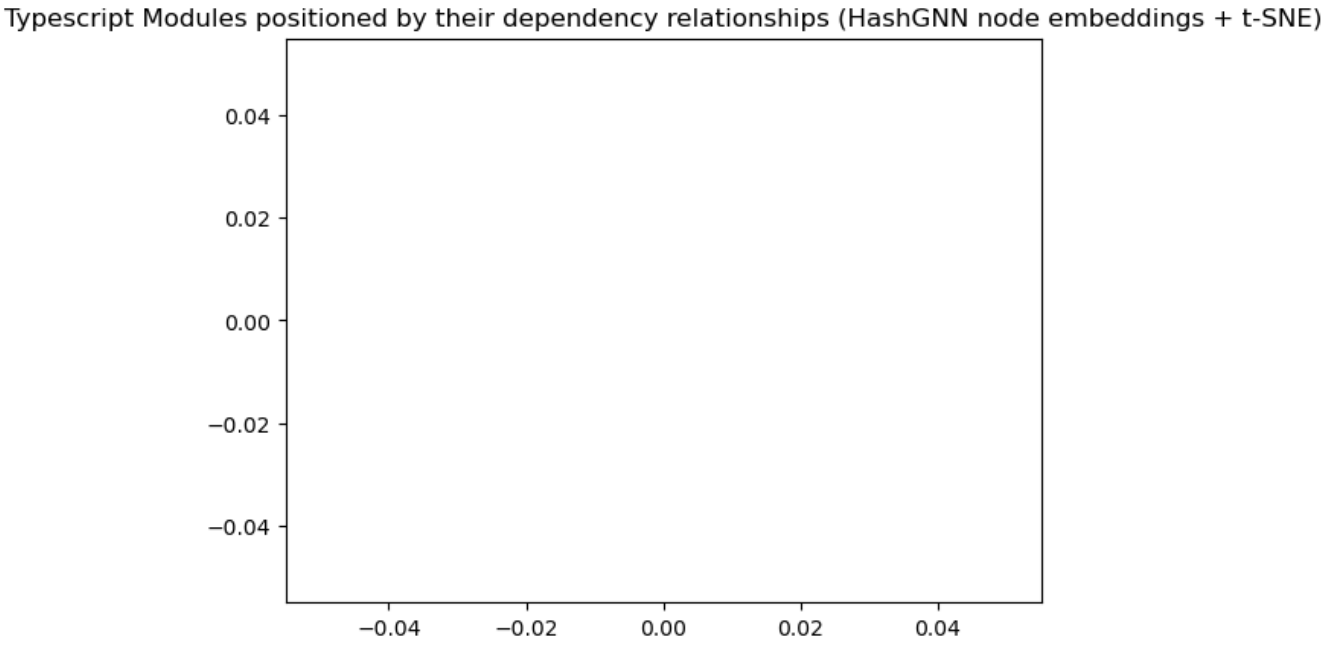
```
The results have been provided by the query filename: ../cypher/Node_Embeddings/Node_
Embeddings_0a_Query_Calculated.cypher
```

	codeUnitName	shortCodeUnitName	projectName	communityId	centrality	embedding
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	react-router	0	0.871695	[0.3061862289905548, 0.6123724579811096, -1.22...
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	react-router-native	0	0.249387	[0.3061862289905548, 0.6123724579811096, -1.22...
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	react-router-dom	1	0.340235	[0.3061862289905548, 0.6123724579811096, -1.22...
3	/home/runner/work/code-graph-analysis-pipeline...	server	react-router-dom	1	0.249387	[0.3061862289905548, 0.6123724579811096, -1.22...

```
Perplexity value 30 is too high. Using perplexity 1.00 instead
```

```
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 3 nearest neighbors using exact search using euclidean distance...
   --> Time elapsed: 0.01 seconds
===> Calculating affinity matrix...
   --> Time elapsed: 0.00 seconds
===> Calculating PCA-based initialization...
   --> Time elapsed: 0.00 seconds
===> Running optimization with exaggeration=12.00, lr=0.33 for 250 iterations...
Iteration   50, KL divergence    nan, 50 iterations in 0.0064 sec
Iteration  100, KL divergence    nan, 50 iterations in 0.0062 sec
Iteration  150, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  200, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  250, KL divergence    nan, 50 iterations in 0.0062 sec
   --> Time elapsed: 0.03 seconds
===> Running optimization with exaggeration=1.00, lr=4.00 for 500 iterations...
Iteration   50, KL divergence    nan, 50 iterations in 0.0062 sec
Iteration  100, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  150, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  200, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  250, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  300, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  350, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  400, KL divergence    nan, 50 iterations in 0.0061 sec
Iteration  450, KL divergence    nan, 50 iterations in 0.0060 sec
Iteration  500, KL divergence    nan, 50 iterations in 0.0061 sec
   --> Time elapsed: 0.06 seconds
/home/runner/miniconda3/envs/codegraph/lib/python3.11/site-packages/sklearn/decomposi
tion/_pca.py:527: RuntimeWarning: invalid value encountered in divide
  explained_variance_ratio_ = explained_variance_ / total_var
/home/runner/miniconda3/envs/codegraph/lib/python3.11/site-packages/openTSNE/initiali
zation.py:27: RuntimeWarning: invalid value encountered in divide
  x /= np.std(x[:, 0]) / target_std
(4, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	0	0.871695	NaN	NaN
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	0	0.249387	NaN	NaN
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.340235	NaN	NaN
3	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.249387	NaN	NaN



1.5 Node Embeddings for Typescript Modules using node2vec

node2vec computes a vector representation of a node based on second order random walks in the graph. The node2vec algorithm is a transductive node embedding algorithm, meaning that it needs the whole graph to be available to learn the node embeddings.

The results have been provided by the query filename: ../cypher/Node_Embeddings/Node_Embeddings_0a_Query_Calculated.cypher

	codeUnitName	shortCodeUnitName	projectName	communityId	centrality	embedding
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	react-router	0	0.871695	[0.18593083322048187, -0.5021991729736328, -0....
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	react-router-native	0	0.249387	[0.18710149824619293, -0.5070731043815613, -0....
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	react-router-dom	1	0.340235	[0.18546485900878906, -0.4823518395423889, -0....
3	/home/runner/work/code-graph-analysis-pipeline...	server	react-router-dom	1	0.249387	[0.19117136299610138, -0.491073876619339, -0.2...

```
Perplexity value 30 is too high. Using perplexity 1.00 instead
```

```
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 3 nearest neighbors using exact search using euclidean distance...
   --> Time elapsed: 0.00 seconds
===> Calculating affinity matrix...
   --> Time elapsed: 0.00 seconds
===> Calculating PCA-based initialization...
   --> Time elapsed: 0.00 seconds
===> Running optimization with exaggeration=12.00, lr=0.33 for 250 iterations...
Iteration   50, KL divergence 0.0820, 50 iterations in 0.0072 sec
Iteration  100, KL divergence 0.0588, 50 iterations in 0.0075 sec
Iteration  150, KL divergence 0.0479, 50 iterations in 0.0075 sec
Iteration  200, KL divergence 0.0414, 50 iterations in 0.0074 sec
Iteration  250, KL divergence 0.0370, 50 iterations in 0.0073 sec
   --> Time elapsed: 0.04 seconds
===> Running optimization with exaggeration=1.00, lr=4.00 for 500 iterations...
Iteration   50, KL divergence 0.0117, 50 iterations in 0.0071 sec
Iteration  100, KL divergence 0.0064, 50 iterations in 0.0072 sec
Iteration  150, KL divergence 0.0044, 50 iterations in 0.0072 sec
Iteration  200, KL divergence 0.0034, 50 iterations in 0.0073 sec
Iteration  250, KL divergence 0.0027, 50 iterations in 0.0074 sec
Iteration  300, KL divergence 0.0023, 50 iterations in 0.0074 sec
Iteration  350, KL divergence 0.0020, 50 iterations in 0.0075 sec
Iteration  400, KL divergence 0.0017, 50 iterations in 0.0075 sec
Iteration  450, KL divergence 0.0015, 50 iterations in 0.0076 sec
Iteration  500, KL divergence 0.0014, 50 iterations in 0.0075 sec
   --> Time elapsed: 0.07 seconds
(4, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	/home/runner/work/code-graph-analysis-pipeline...	react-router	0	0.871695	-19.001085	0.247846
1	/home/runner/work/code-graph-analysis-pipeline...	react-router-native	0	0.249387	-19.001167	0.247845
2	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.340235	19.001078	-0.247846
3	/home/runner/work/code-graph-analysis-pipeline...	react-router-dom	1	0.249387	19.001174	-0.247846