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: 1.0.2
The pandas version is: 2.2.3
```

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. Java Packages

1.1 Generate Node Embeddings using Fast Random Projection (Fast RP) for Java Packages

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.

```
Received notification from DBMS server: {severity: WARNING} {code: Neo.ClientNotification.Statement.UnknownRelationshipTypeWarning} {category: UNRECOGNIZED} {title: The provided relationship type is not in the database.} {description: One of the relationship types in your query is not available in the database, make sure you didn't misspell it or that the label is available when you run this statement in your application (the missing relationship type is: HAS_ROOT)} {position: line: 9, column: 44, offset: 696} for query: '// Query already calculated and written node embeddings on nodes with label in parameter $dependencies_projection_node including a communityId and centrality. Variables: dependencies_projection_node, dependencies_projection_write_property. Requires "Add_file_name and_extension.cypher".\n \n  MATCH (codeUnit)\n  WHERE $dependencies_projection_node IN LABELS(codeUnit)\n    AND codeUnit[$dependencies_projection_write_property] IS NOT NULL\n    // AND codeUnit.notExistingToForceRecalculation IS NOT NULL // uncomment this line to force recalculation\n OPTIONAL MATCH (artifact:Java:Artifact)-[:CONTAINS]->(codeUnit)\n    WITH *, artifact.name AS artifactName\n OPTIONAL MATCH (projectRoot:Directory)<-[:HAS_ROOT]-(proj:TS:Project)-[:CONTAINS]->(codeUnit)\n    WITH *, last(split(projectRoot.absoluteFileName, \'/\')) AS projectName\n  RETURN DISTINCT \n         coalesce(codeUnit.fqn, codeUnit.globalFqn, codeUnit.fileName, codeUnit.signature, codeUnit.name) AS codeUnitName\n        ,codeUnit.name AS shortCodeUnitName\n        ,coalesce(artifactName, projectName) AS projectName\n        ,coalesce(codeUnit.communityLeidenId, 0)           AS communityId\n        ,coalesce(codeUnit.centralityPageRank, 0.01)       AS centrality\n ,codeUnit[$dependencies_projection_write_property] AS embedding\n   ORDER BY communityId'
```

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	org.axonframework.disruptor.commandhandling	commandhandling	axon-disruptor-4.11.0	0	0.012156	[0.13631552457809448, 0.03722062706947327, 0.3...
1	org.axonframework.modelling.command	command	axon-modelling-4.11.0	0	0.033013	[0.12645640969276428, 0.00017376989126205444, ...
2	org.axonframework.modelling.command.inspection	inspection	axon-modelling-4.11.0	0	0.019850	[0.1149144172668457, -0.015440702438354492, 0....
3	org.axonframework.modelling.command.legacyjpa	legacyjpa	axon-modelling-4.11.0	0	0.011941	[0.04084441438317299, -0.055164188146591187, 0...
4	org.axonframework.eventsourcing	eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	[0.094574473798275, 0.02428731508553028, 0.292...

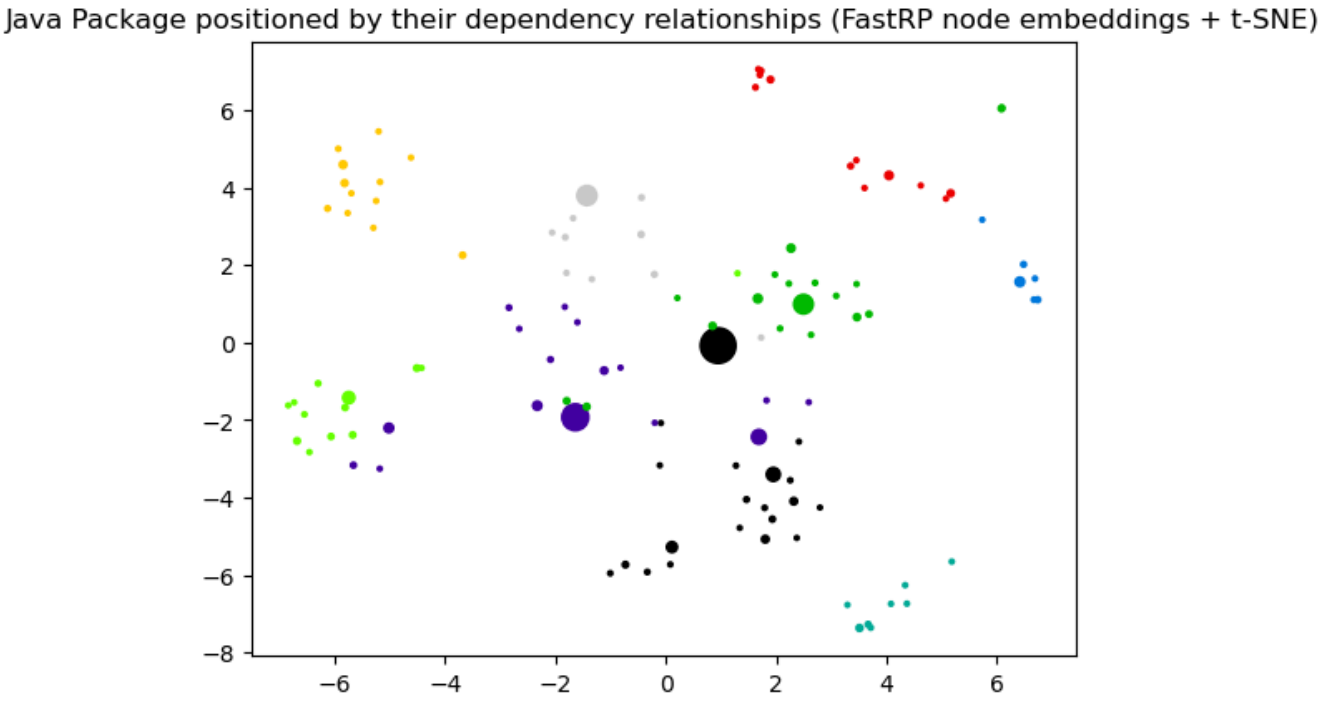
1.2 Dimensionality reduction with t-distributed stochastic neighbor embedding (t-SNE)

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".

```
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 90 nearest neighbors using exact search using euclidean distance...
   --> Time elapsed: 0.03 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=9.67 for 250 iterations...
Iteration   50, KL divergence -0.5190, 50 iterations in 0.0420 sec
Iteration  100, KL divergence 1.2405, 50 iterations in 0.0111 sec
Iteration  150, KL divergence 1.2405, 50 iterations in 0.0101 sec
Iteration  200, KL divergence 1.2405, 50 iterations in 0.0101 sec
Iteration  250, KL divergence 1.2405, 50 iterations in 0.0101 sec
   --> Time elapsed: 0.08 seconds
===> Running optimization with exaggeration=1.00, lr=116.00 for 500 iterations...
Iteration   50, KL divergence 0.1821, 50 iterations in 0.0376 sec
Iteration  100, KL divergence 0.1582, 50 iterations in 0.0497 sec
Iteration  150, KL divergence 0.1520, 50 iterations in 0.0457 sec
Iteration  200, KL divergence 0.1522, 50 iterations in 0.0457 sec
Iteration  250, KL divergence 0.1521, 50 iterations in 0.0457 sec
Iteration  300, KL divergence 0.1521, 50 iterations in 0.0528 sec
Iteration  350, KL divergence 0.1520, 50 iterations in 0.0460 sec
Iteration  400, KL divergence 0.1521, 50 iterations in 0.0461 sec
Iteration  450, KL divergence 0.1521, 50 iterations in 0.0446 sec
Iteration  500, KL divergence 0.1520, 50 iterations in 0.0447 sec
   --> Time elapsed: 0.46 seconds
(116, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	org.axonframework.disruptor.commandhandling	axon-disruptor-4.11.0	0	0.012156	1.337732	-4.781068
1	org.axonframework.modelling.command	axon-modelling-4.11.0	0	0.033013	1.800289	-5.072100
2	org.axonframework.modelling.command.inspection	axon-modelling-4.11.0	0	0.019850	1.928533	-4.555024
3	org.axonframework.modelling.command.legacyjpa	axon-modelling-4.11.0	0	0.011941	2.371203	-5.040282
4	org.axonframework.eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	1.459869	-4.049171

1.3 Visualization of the node embeddings reduced to two dimensions



1.4 Node Embeddings for Java Packages 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.

```
Received notification from DBMS server: {severity: WARNING} {code: Neo.ClientNotifica
tion.Statement.UnknownRelationshipTypeWarning} {category: UNRECOGNIZED} {title: The p
rovided relationship type is not in the database.} {description: One of the relations
hip types in your query is not available in the database, make sure you didn't misspe
ll it or that the label is available when you run this statement in your application
(the missing relationship type is: HAS_ROOT)} {position: line: 9, column: 44, offset:
696} for query: '// Query already calculated and written node embeddings on nodes wit
h label in parameter $dependencies_projection_node including a communityId and centra
lity. Variables: dependencies_projection_node, dependencies_projection_write_propert
y. Requires "Add_file_name and_extension.cypher".\n \n  MATCH (codeUnit)\n  WHERE $de
pendencies_projection_node IN LABELS(codeUnit)\n    AND codeUnit[$dependencies_projec
tion_write_property] IS NOT NULL\n    // AND codeUnit.notExistingToForceRecalculation
IS NOT NULL // uncomment this line to force recalculation\n OPTIONAL MATCH (artifact:
Java:Artifact)-[:CONTAINS]->(codeUnit)\n    WITH *, artifact.name AS artifactName\n O
PTIONAL MATCH (projectRoot:Directory)<-[:HAS_ROOT]-(proj:TS:Project)-[:CONTAINS]->(co
deUnit)\n    WITH *, last(split(projectRoot.absoluteFileName, \'/\')) AS projectName
\n  RETURN DISTINCT \n         coalesce(codeUnit.fqn, codeUnit.globalFqn, codeUnit.fi
leName, codeUnit.signature, codeUnit.name) AS codeUnitName\n        ,codeUnit.name
AS shortCodeUnitName\n        ,coalesce(artifactName, projectName)
AS projectName\n        ,coalesce(codeUnit.communityLeidenId, 0)           AS communi
tyId\n        ,coalesce(codeUnit.centralityPageRank, 0.01)       AS centrality\n
,codeUnit[$dependencies_projection_write_property] AS embedding\n   ORDER BY communit
yId'
```

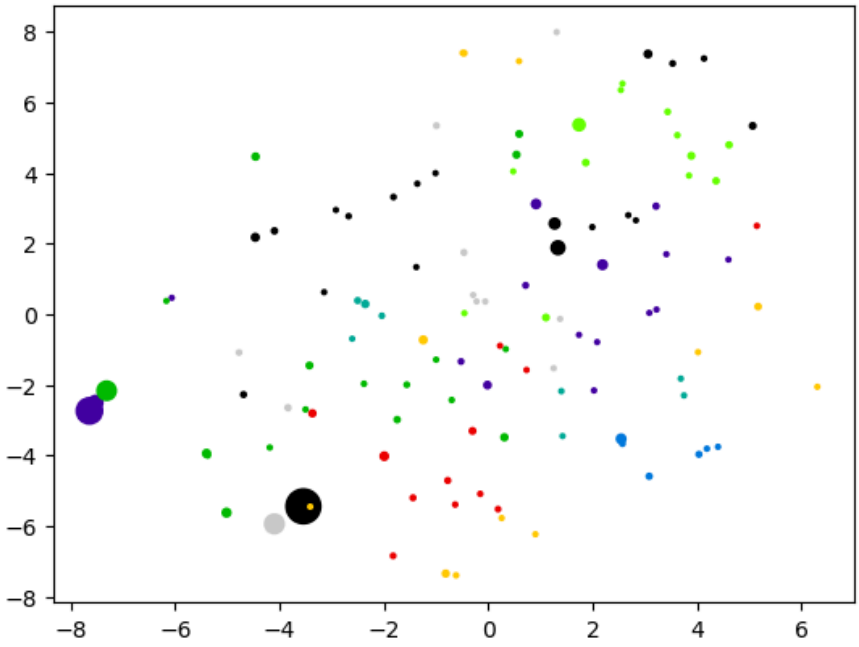
```
The results have been provided by the query filename: ../cypher/Node_Embeddings/Node_
Embeddings_0a_Query_Calculated.cypher
```

	codeUnitName	shortCodeUnitName	projectName	communityId	centrality	embedding
0	org.axonframework.disruptor.commandhandling	commandhandling	axon-disruptor-4.11.0	0	0.012156	[0.0, -1.2990380823612213, -1.0825317353010178...
1	org.axonframework.modelling.command	command	axon-modelling-4.11.0	0	0.033013	[1.2990380823612213, -1.948557123541832, -1.08...
2	org.axonframework.modelling.command.inspection	inspection	axon-modelling-4.11.0	0	0.019850	[1.7320507764816284, -1.7320507764816284, -0.4...
3	org.axonframework.modelling.command.legacyjpa	legacyjpa	axon-modelling-4.11.0	0	0.011941	[-0.4330126941204071, -2.5980761647224426, 0.6...
4	org.axonframework.eventsourcing	eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	[0.0, -1.7320507764816284, -0.2165063470602035...

```
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 90 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=9.67 for 250 iterations...
Iteration   50, KL divergence -1.0528, 50 iterations in 0.0725 sec
Iteration  100, KL divergence 1.2151, 50 iterations in 0.0176 sec
Iteration  150, KL divergence 1.2151, 50 iterations in 0.0147 sec
Iteration  200, KL divergence 1.2151, 50 iterations in 0.0115 sec
Iteration  250, KL divergence 1.2151, 50 iterations in 0.0100 sec
   --> Time elapsed: 0.13 seconds
===> Running optimization with exaggeration=1.00, lr=116.00 for 500 iterations...
Iteration   50, KL divergence 0.6200, 50 iterations in 0.0375 sec
Iteration  100, KL divergence 0.6033, 50 iterations in 0.0512 sec
Iteration  150, KL divergence 0.5972, 50 iterations in 0.0517 sec
Iteration  200, KL divergence 0.5969, 50 iterations in 0.0533 sec
Iteration  250, KL divergence 0.5969, 50 iterations in 0.0531 sec
Iteration  300, KL divergence 0.5970, 50 iterations in 0.0550 sec
Iteration  350, KL divergence 0.5970, 50 iterations in 0.0542 sec
Iteration  400, KL divergence 0.5970, 50 iterations in 0.0535 sec
Iteration  450, KL divergence 0.5971, 50 iterations in 0.0538 sec
Iteration  500, KL divergence 0.5971, 50 iterations in 0.0547 sec
   --> Time elapsed: 0.52 seconds
(116, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	org.axonframework.disruptor.commandhandling	axon-disruptor-4.11.0	0	0.012156	-1.012244	3.993793
1	org.axonframework.modelling.command	axon-modelling-4.11.0	0	0.033013	-4.466856	2.176367
2	org.axonframework.modelling.command.inspection	axon-modelling-4.11.0	0	0.019850	-4.099579	2.358433
3	org.axonframework.modelling.command.legacyjpa	axon-modelling-4.11.0	0	0.011941	-2.921124	2.949980
4	org.axonframework.eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	-4.692043	-2.277287

Java Package positioned by their dependency relationships (HashGNN node embeddings + t-SNE)



2.5 Node Embeddings for Java Packages using node2vec

```
Received notification from DBMS server: {severity: WARNING} {code: Neo.ClientNotifica
tion.Statement.UnknownRelationshipTypeWarning} {category: UNRECOGNIZED} {title: The p
rovided relationship type is not in the database.} {description: One of the relations
hip types in your query is not available in the database, make sure you didn't misspe
ll it or that the label is available when you run this statement in your application
(the missing relationship type is: HAS_ROOT)} {position: line: 9, column: 44, offset:
696} for query: '// Query already calculated and written node embeddings on nodes wit
h label in parameter $dependencies_projection_node including a communityId and centra
lity. Variables: dependencies_projection_node, dependencies_projection_write_propert
y. Requires "Add_file_name and_extension.cypher".\n \n  MATCH (codeUnit)\n  WHERE $de
pendencies_projection_node IN LABELS(codeUnit)\n    AND codeUnit[$dependencies_projec
tion_write_property] IS NOT NULL\n    // AND codeUnit.notExistingToForceRecalculation
IS NOT NULL // uncomment this line to force recalculation\n OPTIONAL MATCH (artifact:
Java:Artifact)-[:CONTAINS]->(codeUnit)\n    WITH *, artifact.name AS artifactName\n O
PTIONAL MATCH (projectRoot:Directory)<-[:HAS_ROOT]-(proj:TS:Project)-[:CONTAINS]->(co
deUnit)\n    WITH *, last(split(projectRoot.absoluteFileName, \'/\')) AS projectName
\n  RETURN DISTINCT \n         coalesce(codeUnit.fqn, codeUnit.globalFqn, codeUnit.fi
leName, codeUnit.signature, codeUnit.name) AS codeUnitName\n        ,codeUnit.name
AS shortCodeUnitName\n        ,coalesce(artifactName, projectName)
AS projectName\n        ,coalesce(codeUnit.communityLeidenId, 0)           AS communi
tyId\n        ,coalesce(codeUnit.centralityPageRank, 0.01)       AS centrality\n
,codeUnit[$dependencies_projection_write_property] AS embedding\n   ORDER BY communit
yId'
```

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	org.axonframework.disruptor.commandhandling	commandhandling	axon-disruptor-4.11.0	0	0.012156	[0.24945801496505737, -0.13298842310905457, -0...
1	org.axonframework.modelling.command	command	axon-modelling-4.11.0	0	0.033013	[0.3946908712387085, -0.17252349853515625, -0....
2	org.axonframework.modelling.command.inspection	inspection	axon-modelling-4.11.0	0	0.019850	[0.3345467746257782, -0.10661914199590683, -0....
3	org.axonframework.modelling.command.legacyjpa	legacyjpa	axon-modelling-4.11.0	0	0.011941	[0.240640327334404, -0.24923254549503326, -0.3...
4	org.axonframework.eventsourcing	eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	[0.43496444821357727, -0.022817663848400116, -...

```
--------------------------------------------------------------------------------
TSNE(early_exaggeration=12, random_state=47, verbose=1)
--------------------------------------------------------------------------------
===> Finding 90 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=9.67 for 250 iterations...
Iteration   50, KL divergence -0.6131, 50 iterations in 0.0491 sec
Iteration  100, KL divergence 1.1723, 50 iterations in 0.0131 sec
Iteration  150, KL divergence 1.1723, 50 iterations in 0.0103 sec
Iteration  200, KL divergence 1.1723, 50 iterations in 0.0103 sec
Iteration  250, KL divergence 1.1723, 50 iterations in 0.0101 sec
   --> Time elapsed: 0.09 seconds
===> Running optimization with exaggeration=1.00, lr=116.00 for 500 iterations...
Iteration   50, KL divergence 0.3555, 50 iterations in 0.0418 sec
Iteration  100, KL divergence 0.3360, 50 iterations in 0.0509 sec
Iteration  150, KL divergence 0.3280, 50 iterations in 0.0507 sec
Iteration  200, KL divergence 0.3282, 50 iterations in 0.0489 sec
Iteration  250, KL divergence 0.3282, 50 iterations in 0.0486 sec
Iteration  300, KL divergence 0.3281, 50 iterations in 0.0492 sec
Iteration  350, KL divergence 0.3278, 50 iterations in 0.0488 sec
Iteration  400, KL divergence 0.3280, 50 iterations in 0.0495 sec
Iteration  450, KL divergence 0.3281, 50 iterations in 0.0488 sec
Iteration  500, KL divergence 0.3281, 50 iterations in 0.0474 sec
   --> Time elapsed: 0.48 seconds
(116, 2)
```

	codeUnit	artifact	communityId	centrality	x	y
0	org.axonframework.disruptor.commandhandling	axon-disruptor-4.11.0	0	0.012156	-2.323706	-1.860799
1	org.axonframework.modelling.command	axon-modelling-4.11.0	0	0.033013	-3.248882	-2.484244
2	org.axonframework.modelling.command.inspection	axon-modelling-4.11.0	0	0.019850	-3.046569	-2.972749
3	org.axonframework.modelling.command.legacyjpa	axon-modelling-4.11.0	0	0.011941	-3.853874	-2.592007
4	org.axonframework.eventsourcing	axon-eventsourcing-4.11.0	0	0.017131	-1.090431	-1.478530

Java Package positioned by their dependency relationships (node2vec node embeddings + t-SNE)