Testat 2: Dokumentation

Big Data Praktikum: Node Embedding & Graph Visualizing

Jonathan Thiemann, Oliver Perle, Moritz Huhle

15.07.2022

Einleitung

Diese Dokumentation befasst sich mit dem zweiten Testat des Big Data Praktikums zum Thema “Node Embedding & Graph Visualizing”. Dabei wurde der Programmcode für die Anwendung fertiggestellt und in Google Colab umgesetzt. Im folgenden werden die Zielsetzungen aus dem ersten Testat abgeglichen und knapp dargestellt. Daraufhin wird auf die einzelnen Funktionen der Anwendung eingegangen und diese in Bezug zur Zielsetzung gebracht. Abschließend wird dem Nutzer die Ablaufweise anhand eines Beispieldurchlaufs zum Setzen bestimmter Parameter beschrieben.

Erfüllung der Zielsetzungen

Mit der Anwendung zur Visualisierung und für den Vergleich von Node Embeddings können verschiedene Graphen eingelesen, manipuliert und mittels drei Embedding-Verfahren dargestellt werden. So kann sich der Nutzer für eine Kombination der Node-Embedding-Verfahren Node2Vec, GCN (Graph Convolutional Networks) und GraphSAGE entscheiden und die verfahrensspezifischen Parameter für Analysezwecke nach Belieben abändern.
Bei der Erstellung der Embeddings kann ein Vergleich von Embeddings auf drei Ebenen ausgeführt werden. Die erste Ebene umfasst den Vergleich zwischen einem Ausgangsgraphen G und einem manipulierten Graphen G’ innerhalb eines Embeddings. Die zweite Ebene ermöglicht den Vergleich eines Embeddings mit zwei unterschiedlichen Sätzen an verfahrensspezifischen Parametern. Auf der dritten Ebene lassen sich zwei verschiedene Embedding-Verfahren für den gleichen Graphen gegenüberstellen.

Funktionen der Anwendung

Die erstellte Anwendung ermöglicht das Erstellen und den Vergleich von Node Embeddings für den Prozess der Node Classification. Hierbei kann zu Beginn beim Import des Graphen aus drei verschiedenen “Stellargraph”-Graphen gewählt werden. Basierend auf dem gewählten Graphen werden zwei Node Embeddings erstellt, die im Anschluss verglichen werden können. Nach der Auswahl der Graphen können genauere Einstellungen für die beiden zu erstellenden Node-Embeddings vorgenommen werden. Hierbei wird die Durchführung der drei im vorangehenden Abschnitt dieser Dokumentation beschriebenen Vergleichsebenen ermöglicht. Es wurde sich für eine Zusammenführung der 3 Ebenen in ein Menü entschieden, um die Handhabung für den Nutzer übersichtlicher zu gestalten und diesem zugleich mehr Möglichkeiten und Freiheiten bei der Erstellung der Embeddings zu bieten. Nachdem die gewünschten Einstellungen im Menü getätigt wurden, werden die gewünschten Verfahren ausgeführt und der Nutzer bekommt die beiden erstellten Embeddings als grafischen Output.

Ablaufbeschreibung innerhalb der Anwendung

Beim Start der Anwendung hat der Nutzer die in Abbildung 1 dargestellte Umgebung vor sich.

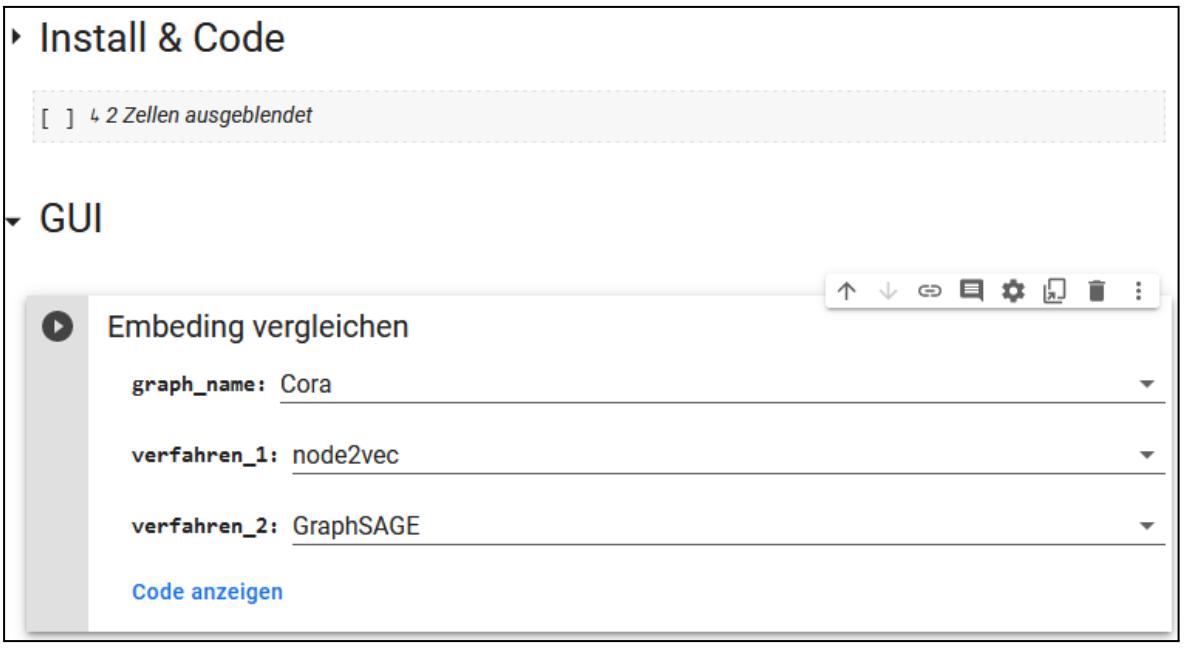

Abbildung 1: Startseite

Zu Beginn ist es nötig, auf den kleinen **“Play-Button”** im Abschnitt **“Install & Code”** zu klicken. Daraufhin kann nach dem Klick auf den weiteren **“Play-Button”** unterhalb des Blocks **“GUI”** die Interaktion mit der dargestellten Benutzeroberfläche starten. Es wird ein Graph ausgewählt, auf dessen Basis die Embeddings erstellt werden sollen. Anschließend werden die gewünschten Embedding-Verfahren für die beiden Embeddings angegeben. Hierbei ist es möglich, für beide Embeddings das gleiche Verfahren zu nutzen oder zwei unterschiedliche Verfahren zu wählen. Für den Graphen sowie für die Verfahren sind default Werte hinterlegt. Demnach ist auch eine Erstellung der Embeddings ohne eigene Eingabe möglich. Je nach Auswahl der Verfahren erscheint im Anschluss eine Übersicht der verfahrensspezifischen Parameter, welche vom Nutzer angepasst werden können. Diese sind in der folgenden Übersicht aufgelistet und kurz erläutert:

Node2Vec:

- *Walke length:* Maximale Länge eines random walks
- *Number of walks:* Anzahl an walks vom jeweiligen Knoten aus
- *P:* Wahrscheinlichkeit von 1/p, um zum vorherigen Knoten zurückzukehren
- *Q:* Wahrscheinlichkeit von 1/q, um zu einem bisher noch nicht erforschten Knoten zu gelangen

GraphSAGE:

- *Node features at each layer:* Anzahl der Dimensionen der Knotenmerkmale pro Schicht
- *Nodes observed at each level:* Anzahl der Knoten, die auf der ersten Ebene des Modells untersucht werden (2-Ebenen Modell: zweite Ebene wird automatisch aus diesem Wert berechnet, eine Betrachtung von noch mehr Ebenen würde die Laufzeit unverhältnismäßig in die Höhe treiben)
- *Epochs:* Anzahl der Durchläufe der Trainingsdaten durch den Algorithmus

GCN:

- *Layer size:* Größe der Verborgenen Einheiten
- *Aktivierungsfunktion:* Funktion, die die Eingaben umwandelt
- *Dropout:* Dropout kann optional auf die Eingabeknotenmerkmale angewendet werden, bevor sie transformiert werden
- *Predict Aktivierungsfunktion:* Funktion, die die Eingaben für die Kreuzvalidierung umwandelt
- *Epochs:* Anzahl der Durchläufe der Trainingsdaten durch den Algorithmus

- *Train size:* Größe der Trainingsmenge
- *Val Train size:* Größe der Validierungsmenge

Die Eingabemaske der Parameter sieht für den Nutzer wie in den folgenden Abbildungen 2 und 3 dargestellt aus. Hierbei kann der Nutzer mit Schiebereglern oder Eingabefenstern interagieren.

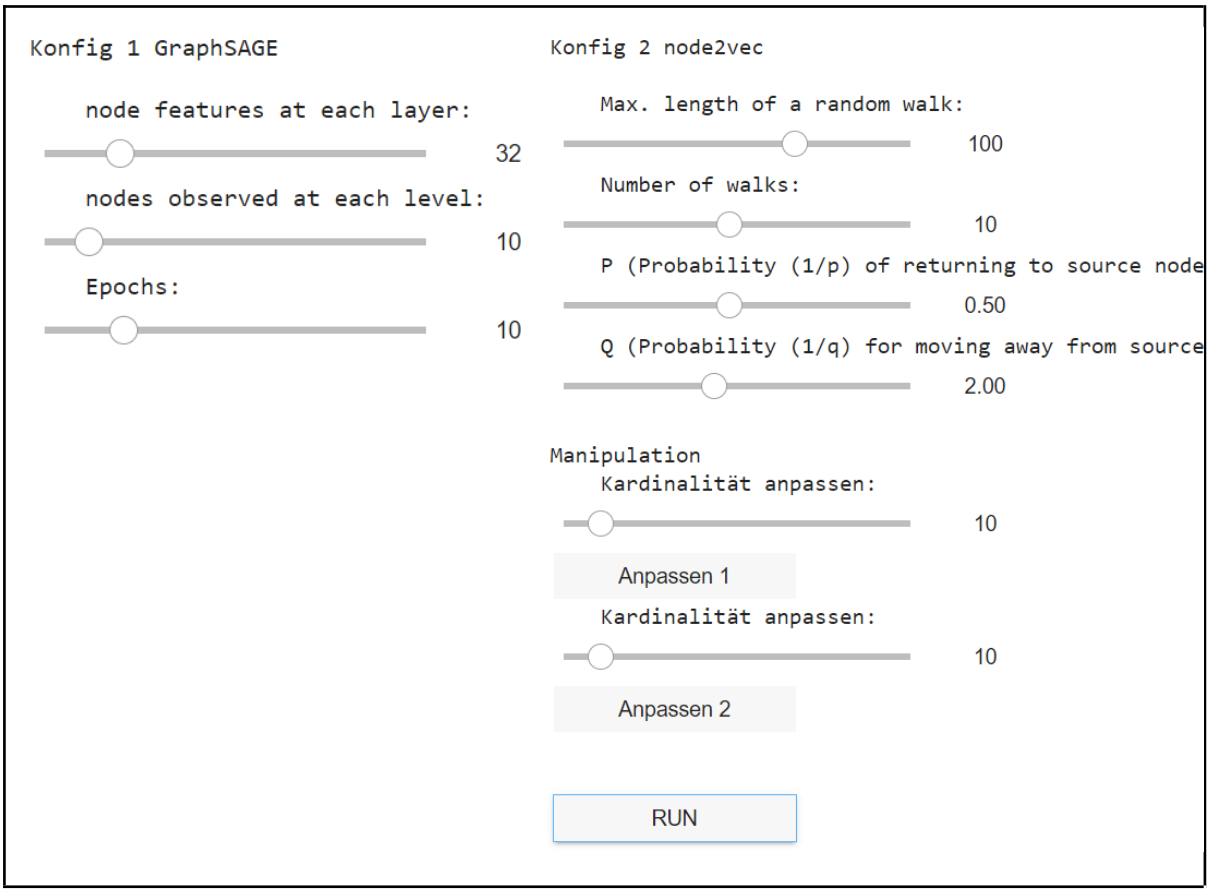

Abbildung 2: Parametereinstellungen für Konfiguration 1

Abbildung 3: Parametereinstellungen für Konfiguration 2 + Manipulation

Bei den Parametereinstellungen sind ebenfalls default Werte hinterlegt, welche die Ausführung des Programms ohne spezifische Eingaben ermöglichen. Neben den verfahrensspezifischen Parametern ist ebenfalls eine Manipulation des Ausgangsgraphen möglich. Somit kann beispielsweise das gleiche Embedding-Verfahren für einen Ausgangsgraphen und eine leicht veränderte Version dieses Graphen mit anschließendem Vergleich der Änderungen in den Embeddings durchgeführt werden. Die Manipulation erfolgt hierbei so, dass der Nutzer einen Wert für die Kardinalität wählen kann, beispielsweise 10%. In diesem Falle werden die 10% der Knoten mit den höchsten Kardinalitäten gelöscht. Dementsprechend sind die meist frequentierten Knoten anschließend nicht mehr vorhanden,

wodurch eine Veränderung bei den erstellten Embeddings zu erwarten ist. Sofern eine Manipulation erwünscht ist, kann nach Anpassung des Schiebereglers der entsprechende “**Anpassen 1**”, bzw. der ”**Anpassen 2**” Button geklickt werden. Die Manipulation wird anschließend durchgeführt und angezeigt.

Mit einem Klick auf den **“RUN”** Button wird bestätigt, dass die Eingaben bezüglich der Konfigurationen der Embeddings soweit abgeschlossen sind. Die Konfigurationen werden noch einmal ausgegeben und die Erstellung der Embeddings gestartet. Zur Laufzeit werden darüber hinaus verfahrensspezifische Ausgaben angezeigt. Je nach Auswahl der Parameter dauert die Erstellung der Embeddings unterschiedlich lange bis sehr lange. Hierbei wurden durch die Vorgabe von min und max Werten für die Parameter eine zu hohe Laufzeit des Programms verhindert. Nach dem Durchlauf der Verfahren werden die beiden erstellten Embeddings ausgegeben und der Nutzer kann diese vergleichen. Eine Beispielausgabe ist in der folgenden Abbildung 4 dargestellt.

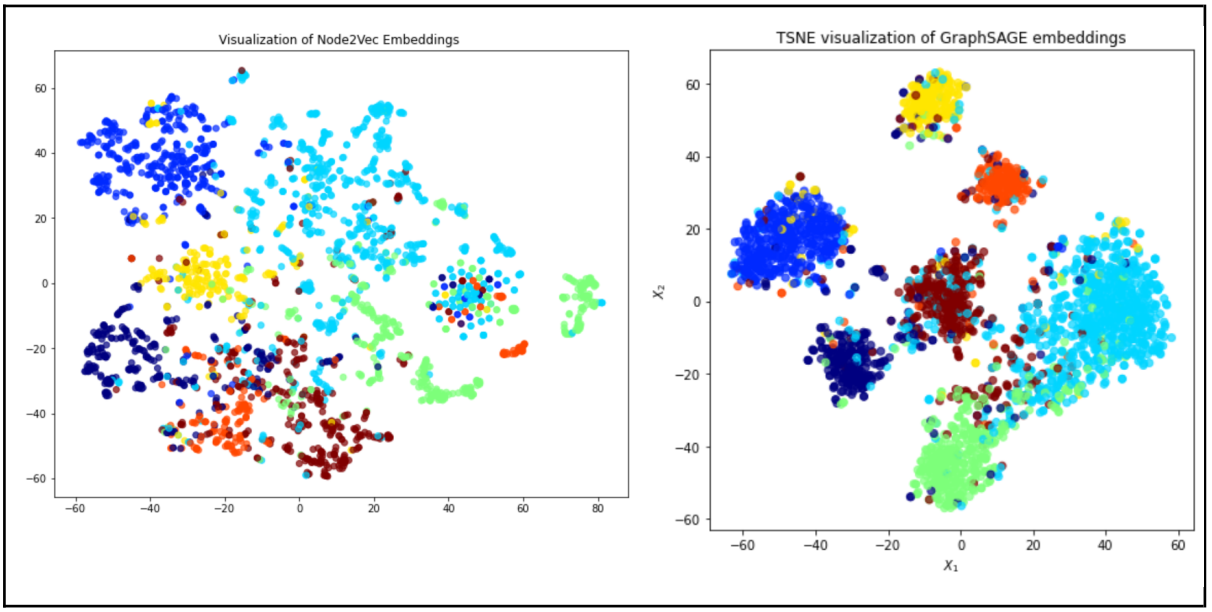

Abbildung 4: Ausgabe der Embeddings für die in Abbildung 2 und 3 genutzten default Werte

Um eine Ausführung des Programms mit neuen Parametern zu starten, kann der Nutzer die Konfigurationen wieder anpassen und erneut auf **“RUN”** klicken. Ist jedoch eine andere Manipulation des Graphen gewünscht, muss die Zelle über den **“Play-Button”** erneut ausgeführt werden.