Meilenstein SNA-Projekt

In diesem Meilenstein formulieren Sie ihre Projektidee. Sie sind keinesfalls darauf beschränkt, dass Sie schlussendlich dann nur genau die hier beschriebenen Analysen durchführen dürfen oder alle hier beschriebenen Analysen durchführen müssen. Der Meilenstein dient zum Austausch zwischen den Studierenden und dem Dozenten in der Anfangsphase, um Ihnen frühzeitig Feedback zur Projektidee geben zu können.

Die von Ihnen auszufüllenden Teile sind jeweils gelb hinterlegt.

Organisatorisches

Die Fragen in diesem Abschnitt betreffen die rein organisatorischen Aspekte des Projekts.

Projekttitel / Projekt Kurzbeschrieb:

Marvel Universe Social Network

Teammitglieder (min. 2, max. 3)

- Florian Baumgartner
- Janis Fröhlich
- Alexander Schilling

Datenquelle

Woher kriegen Sie Ihre Daten?

Marvel Universe Social Network

- edges.csv
- nodes.csv
- hero-edge.csv

Dürfen Sie die Daten einsammeln und verwenden. Welche Dokumente (AGBs, Terms of use, robots.txt usw.) wurden berücksichtigt, um diese Frage zu beantworten?

Creative Commons Lizensierung

Ist der Zugang zu den Daten limitiert? (beispielsweise haben APIs häufig Zugriffs-Limitierungen wie beispielsweise maximal 100 Anfragen pro Tag). Falls ja, inwiefern schränkt Sie dies ein? Wie gehen Sie damit um, damit dies nicht zu einem Problem wird?

Unlimitiert. Die Daten können von Kaggle einfach heruntergeladen werden.

Datenmodellierung

Was bildet in Ihrem Netzwerk die Knoten? Welche Bedeutung(en) haben die Kanten? Handelt es sich um ein One-Mode oder Two-Mode Netzwerk? Planen Sie verschiedene Modellierungen?

Die Knoten entsprechen die Namen der Charaktere und deren Typ (Hero oder Comic) und die Kanten stellen ihre Beziehungen durch gemeinsame Auftritte in Comics dar.

Ob es sich um ein One- oder Two-Mode Netzwerk handelt, kommt auf das Ziel der Exploration an.

- **One-Mode**: Alle Knoten könnten Superhelden sein und die Kanten repräsentieren gemeinsame Auftritte in Comics.
- **Two-Mode**: Es gibt eine Art von Knoten für Superhelden und eine andere Art für Comics. Die Kanten repräsentieren dann die Auftritte der Charaktere in den Comics.

Die Idee besteht also darin, beide Modellierungen zu analysieren und unterschiedliche Gewichtungen für die Kanten zu generieren, basierend auf der Anzahl der gemeinsamen Auftritte, der Wichtigkeit der Charaktere im Marvel-Universum oder anderen relevanten Faktoren.

Mit welcher Netzwerk-Grösse rechnen Sie? (Brechen Sie die Abschätzung auf den Typ herunter, falls sie ein Two-Mode Netzwerk verwenden):

One-Mode (Helden zu Helden, basierend auf gemeinsamen Comics):

- Anzahl Knoten: 6440
- Anzahl Kanten: 171644

Two-Mode (Helden zu Comics):

- Anzahl Knoten: 19091
- Anzahl Kanten: 96104

Welche Attribute haben Sie auf den Knoten und Kanten? Geben Sie für jedes Attribut, welches Sie in ihren Analysen verwenden, eine Prognose an, was für eine Datenqualität / Probleme Sie nach Ihren ersten Untersuchungen erwarten. (Wie vollständig sind die Daten, wie korrekt sind die Daten, gibt es unterschiedliche Schreibweisen für dasselbe Konzept usw.)

Knoten-Attribute:

- node: Dies ist der eindeutige Identifikator für jeden Knoten, sei es ein Held oder ein Comic.
- type: Dies gibt den Typ des Knotens an, entweder "hero" oder "comic".

Kanten-Attribute:

- hero: Der Held, der in einem bestimmten Comic erscheint.
- comic: Der Comic, in dem ein bestimmter Held erscheint.

Prognose der Datenqualität

node Attribut:

- Vollständigkeit: Es scheinen alle Knoten einen eindeutigen Identifikator zu haben. Es wäre jedoch sinnvoll, zu überprüfen, ob es Duplikate gibt.

- Konsistenz: Es könnte unterschiedliche Schreibweisen oder Variationen für denselben Helden oder Comic geben, z. B. "Spider-Man" und "Spiderman". Dies sollte überprüft werden.

type Attribut:

- Vollständigkeit: Jeder Knoten sollte einen Typ haben. Es wäre gut zu überprüfen, ob es Knoten ohne Typ gibt.
- Korrektheit: Es sollte nur zwei Typen geben: "hero" und "comic". Jegliche Abweichungen wären ein Zeichen für inkorrekte Daten.
- Konsistenz: Solange die Daten nur die beiden erwarteten Typen enthalten, sollte dieses Attribut konsistent sein.

hero und comic Attribute (Kanten):

- Vollständigkeit: Jede Kante sollte einen Helden und einen Comic haben. Es wäre gut zu überprüfen, ob es Kanten ohne diese Attribute gibt.
- Konsistenz: Es sollte überprüft werden, ob es Kanten gibt, die zwei Helden oder zwei Comics verbinden, da dies im Kontext dieses Netzwerks nicht zulässig ist.

Leiten Sie aus gesammelten Daten neue Attribute ab (z.B. Kategorisierung verschiedener Werte, Extraktion von Alter anhand der Jahreszahl, usw.)? Falls ja, welches sind diese neuen Attribute und wie sieht Ihre Strategie aus, diese abzuleiten? Welche Datenqualität erwarten Sie?

Grad des Helden:

- Dies wäre die Anzahl der Comics, in denen ein Held erscheint. Dies könnte verwendet werden, um die Popularität eines Helden zu bestimmen.
- Strategie: Grad-Funktion von Gephi
- Datenqualität: Da dies direkt aus den Kanten des Netzwerks abgeleitet wird, erwarten wir eine hohe Datenqualität.

Grad des Comics:

- Dies wäre die Anzahl der Helden, die in einem bestimmten Comic erscheinen. Dies könnte Hinweise auf Crossover-Comics geben.
- Strategie: Grad-Funktion von Gephi
- Datenqualität: Da dies direkt aus den Kanten des Netzwerks abgeleitet wird, erwarten wir eine hohe Datenqualität.

Jahr des Comics (falls möglich):

- Wenn die Comic-Namen Jahreszahlen enthalten, könnten wir versuchen, das Jahr zu extrahieren. Dies würde uns eventuell ermöglichen, Trends zu analysieren.
- Strategie: Ein regulärer Ausdruck könnte verwendet werden, um das Jahr aus dem Comic-Namen zu extrahieren.
- Datenqualität: Hängt von der Konsistenz der Comic Namen ab. Wir erwarten nicht, dass alle Comic Namen eine Jahreszahl beinhaltet. Dies könnte also zu ungenauen Ergebnissen führen.

Zentralitätsmasse:

- Es gibt verschiedene Zentralitätsmasse wie Betweenness oder Closeness für jeden Helden zu berechnen.

- Strategie: Funktionen von Gephi
- Datenqualität: Hängt von der Struktur des Netzwerks ab. Dies scheint aber recht stabil zu sein, darum erwarten wir hier genaue Ergebnisse.

Analysen

Beschreiben Sie in diesem Abschnitt, was sie wie analysieren möchten. Verwenden Sie für jede Analyse die dafür vorgegebene Tabelle. Jede Analyse soll in einer eigenen Tabelle beschrieben werden.

These / Frage:	Welche Helden sind die zentralsten im Netzwerk?
Filterung:	Keine. Die Analyse wird auf dem kompletten Netzwerk durchgeführt.
Analyse:	Verwendung von Zentralitätsmassen wie Degree-Centrality, Betweenness-Centrality und Closeness-Centrality.
Erwartung:	Die populärsten oder bekanntesten Helden könnten die höchsten Zentralitätswerte haben, da sie in vielen Comics erscheinen und somit viele Verbindungen im Netzwerk haben.

These / Frage:	Wie ist die Verteilung der Anzahl der Comics pro Held?
Filterung:	Keine. Die Analyse wird auf dem kompletten Netzwerk durchgeführt.
Analyse:	Berechnung der Degree-Distribution für Helden-Knoten.
Erwartung:	Es könnte eine schräge Verteilung geben, bei der nur wenige Helden in vielen Comics erscheinen, während die Mehrheit der Helden nur in wenigen Comics erscheint.

These / Frage:	Gibt es Gruppen oder Gemeinschaften von Helden, die häufig zusammen in Comics erscheinen?
Filterung:	Keine. Die Analyse wird auf dem kompletten Netzwerk durchgeführt.
Analyse:	Verwendung von Community-Detection-Algorithmen, z. B. dem Louvain-Algorithmus.
Erwartung:	Es könnten klar definierte Gemeinschaften oder Gruppen von Helden identifiziert werden, die aufgrund gemeinsamer Geschichten oder Crossovers häufig zusammen erscheinen.

These / Frage:	Gibt es Helden, die nur in einem einzigen Comic erscheinen?
Filterung:	Ja, Filterung nach Helden mit einem Grad von 1.

Analyse:	Berechnung des Grades für jeden Helden und Identifizierung von Helden mit einem Grad von 1.
Erwartung:	Es könnten einige Helden identifiziert werden, die nur in einem Comic erscheinen, was auf Nebencharaktere oder weniger bekannte Helden hinweisen könnte.

Fragen und Unklarheiten?

-