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:

Soziale Himmelskrieger

Teammitglieder (min. 2, max. 3)

- Patrick Schürmann
- Si Ben Tran
- Flurina Riner

Datenquelle

Woher kriegen Sie Ihre Daten?

https://www.kaggle.com/datasets/ruchi798/star-wars

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

Wie auf der Website auch sichtbar ist, ist der Datensatz lizenzfrei.

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?

Wir haben keine Limitierung, da wir nicht mit einer API arbeiten. Es wäre zwar grundsätzlich möglich, aber der Zugriff funktioniert nur für einen privaten Account. Weil das ganze aber für alle hier beteiligten nutzbar sein sollte, und sich der Datensatz nicht mehr verändern wird gemäss Website, haben wir den Datensatz heruntergeladen.

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?

1. **Knoten im Netzwerk**: Die Knoten in diesem Netzwerk sind die Charaktere aus der Star-Wars-Reihe. Jeder Charakter wird durch einen Knoten dargestellt.

2. **Bedeutung der Kanten**: Die Kanten repräsentieren die Interaktionen und Erwähnungen zwischen den Charakteren. Eine Kante zwischen zwei Charakteren bedeutet, dass diese Charaktere entweder in derselben Szene gesprochen haben oder in derselben Szene erwähnt wurden.
3. **One-Mode oder Two-Mode Netzwerk**: Es handelt sich um ein One-Mode-Netzwerk.
4. **Verschiedene Modellierungen**: Im Moment planen wir mit dem bestehenden Netzwerk zu arbeiten. Je nach Entwicklung haben wir uns überlegt, aus den verschiedenen Filmepisoden ein Two-Mode-Netzwerk zu erstellen. Dies würde uns weitere Auswertungen ermöglichen.

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

Anzahl Knoten: 110 Anzahl Kanten: 443

Diese Angaben gelten über alle Filme. Wir haben auch die Daten für jede einzelne Episode, welche wir zuerst verwenden werden. Diese werden kleiner sein, in der ersten Episode hat es 37 Knoten und 134 Kanten.

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.)

- **Knotenattribute**:
  - Name des Charakters
  - Anzahl der Szenen, in denen der Charakter auftritt
  - Farbe in der Visualisierung (wir können dieses Attribut nicht nachvollziehen und werden es vorerst ignorieren)
- **Kantenattribute**:
  - Index des Start- und Zielcharakters
  - Anzahl der Szenen, in denen beide Charaktere gemeinsam auftreten
- **Datenqualität und -probleme**: Wir erwarten noch keine Unregelmässigkeiten oder Probleme bei den Daten. Denn es handelt sich um einen viel genutzten Datensatz von Kaggle, für den auch schon Notebooks/Anwendungen veröffentlicht wurden. Wir erwarten, in der explorativen Datenanalyse mögliche Fehler abfangen zu können.

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?

Wie erwähnt überlegen wir uns, für weitere Untersuchungen basierend auf der Filmepisode ein Two-Mode-Netzwerk zu erstellen. Neue Attribute wären zwar interessant, wir müssten sie aber von einem

fremden Datensatz joinen. Dies würden wir im Moment noch nicht machen, zum Beispiel weil wir einen Datensatz finden müssten, wo die gleichen Charakteren erfasst sind und es einen gemeinsamen Nenner gibt. Alternativ würden wir jede Film schauen und manuell erfassen ☺

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:	Wie entwickeln sich die Wichtigkeiten einzelner Charakteren/Gruppen im Verlauf der Filme (inkl. Prestige)? Lassen sich Cluster/Communities erkennen oder gar erstellen? Gibt es Cliquen? Wie entwickeln sich die Zentralisierung der Netzwerk? Wie unterscheiden sich die Netzwerke? Falls neue Attribute: Mögliche Hypothesentests auf den Knotenattribute Welche Charakteren sollten sich kennenlernen? Wie schnell würde sich die Geburt Yodas im Netzwerk verbreiten? (Dies sind viele Fragen, die wir gesammelt haben. Wenn es sich als zu aufwändig herausstellt, werden wir nur ein Teil davon beantworten)
Filterung:	Mangels Attributen können wir kaum Filterungen durchführen. Einzig die Häufigkeit der Erscheinung der Charakteren könnten gefiltert werden.
Analyse:	Gemäss Script und den tatsächlich beantworteten Fragen.
Erwartung:	Wir erwarten beim Vergleich unterschiedlicher Episoden sichtbare Differenzen, da sich die Filme zum Teil stark unterscheiden. Bei der Wichtigkeit von Charakteren erwarten wir besonders bei den berühmten höhere Werte.

Fragen und Unklarheiten?

Nennen Sie Fragen und Unklarheiten hier.