SNA-Projekt: Meilenstein Projektklärung

Die Projektklärung definiert die Rahmenbedingungen Ihres Projektes. Sie sollten, bevor Sie richtig durchstarten, in Ruhe einmal verschiedene Überlegungen zum Projekt machen, erste Voruntersuchungen zu tätigen und das Ganze zusammengefasst niederzuschreiben. Dies hilft schon zu Beginn einen roten Faden verfolgen zu können und mögliche Problematiken frühzeitig zu erkennen.

Zu diesem Meilenstein gehören unter anderem eine kurze Beschreibung Ihrer Gruppe, der geplanten Infrastruktur, der Datenmodellierung, erwartete Datenmenge und –qualität sowie auch die beabsichtigte Analysen Es handelt sich einmal um eine initiale Definition, Sie dürfen aber schlussendlich im zweiten Teil gerne weitere Analysen durchführen oder (begründet) gewisse der hier definierten Analysen weglassen. Möglicherweise klappt nicht alles, wie sich das zu Beginn vorstellen. Doch dies bringt dann vielleicht einen guten Lerngewinn.

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:

Filmdatenanalyse

Teammitglieder

In meinem Team befinden sich die folgenden Team-Mitglieder (min. 2, max. 3):

- Firat Saritas
- Kajenthini Kobivasan

Welche Tools gedenken Sie einzusetzen für die SNA?

☒ Gephi

☐ Pajek

☐ R

☒ Python

☐ Anderes? Falls ja, welche?

Datenquelle

Woher kriegen Sie Ihre Daten?

Für unsere praktische Arbeit werden wir die öffentliche Filmdatenbank der ProCinema, Schweizer Verband für Kino und Filmverleih nutzen. Die Filmdatenbank umfasst alle Filme, die seit 1995 in den Schweizer Kinos aufgeführt wurden. Zur Zeit verfügt die Datenbank Informationen über mehr als 12'000 Titel.

Wie können Sie auf die Daten dieser Datenquelle zugreifen?

☐ Sparql Endpoint

☐ API

☒ Eigener Webcrawler

☐ Anderes? Falls ja, was?

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

Die Daten stammen aus einer staatlichen Organisation. Auf der Webseite stehen keine AGBs oder Terms of use ähnliche Informationen. Da wir diese Daten nur für schulische Zwecke verwenden werden, nehmen wir an, dass keine Probleme entstehen werden.

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

Der Zugang zu den Daten ist nicht limitiert.

Wie sieht ihre geplante System-Umgebung aus? Zeichen Sie ein Datenfluss-Diagramm mit den einzelnen Komponenten. (Siehe Dokument ETL Prozess für eine Beispiel-Umgebung)

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?

In unserem Netzwerk bilden folgende Informationen die Knoten:

- Titel
- Kinostart
- Besucherzahlen
- Regie
- Produzent
- Drehbuch
- Musik
- Schauspieler
- Produktionsland

Es handelt sich um eine Two-Mode Netzwerk, welches für die weitere Analyse in eine One-Mode transformiert werden muss.

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

Anzahl Knoten: ca. 700 (Schauspielöer/in) Anzahl Kanten: 12'000 (Filme)

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: Geschlecht

Kanten: Genre, Länge, Sprache, Erscheinungsjahr, Erscheinungsland, Regie, Produzent

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?

Erscheinungsjahr: 1995-2019 (25 Jahre)

Filmdauer: >1h, 1-2h, <2h (wurde nicht umgesetzt)

Erscheinungsland: Kontinent (wurde nicht umgesetzt)

Analysen

Beschreiben Sie in diesem Abschnitt, was sie wie analysieren möchten. Verwenden Sie die dafür vorgegebene Tabelle. Jede Analyse soll in einer eigenen Tabelle beschrieben werden. Erwähnen Sie, welche berechneten Messwerte Sie verwenden (sofern das Thema bereits im Unterricht behandelt wurde).

These / Frage:	Mit wie vielen Schauspielern hat, ein Schauspieler durchschnittlich gespielt?
Filterung:	Für diese Analyse brauchen wir nur die Angaben der Schauspieler und die von ihnen gespielte Filmtitel.
Analyse:	Um diesen Durchschnitt zu berechnen, werden wir Python nutzen und den Mittelwert ausrechnen lassen.
Erwartung:	Es wird erwartet, dass ein/e Schauspieler/in mindestens mit zwei weiteren Personen gespielt haben wird. Als einen Maximalwert setzten wir bei 4 Personen an, weil meistens nur die Hauptfiguren aufgelistet werden und nie alle Schauspieler.

These / Frage:	Welcher Regisseur hat bei den meisten Filmen Regie geführt?
Filterung:	Für diese Analyse nehmen wir die Daten zur Regie und Filmtitel.
Analyse:	Wir werden in Python und in Gephi die Analyse erstellen und auflisten welche Regisseure bei den meisten Filmen Regie geführt haben. Hierfür werden wir rechnen, wie viele Filme pro Regisseur zugeteilt werden.
Erwartung:	Wir haben keine Erwartung für diese Frage und werden uns von dem Resultat überraschen lassen.

These / Frage:	Welche Filme wurden am meisten besucht und wessen Schauspieler waren diese?
Filterung:	Wir werden diese Analyse gleich in Python durchführen und dabei nur die Daten zu den Besucherzahlen genauer beobachten. Anschliessend werden wir die Informationen zu den Schauspielern hinzufügen.
Analyse:	Durch eine Auflistung mit den meistbesuchten Filmen, können wir identifizieren, welche Filme am beliebtesten waren. Anhand dieser Liste können wir dann überprüfen, ob es gewisse Star-Schauspieler/innen in den letzten 25 Jahren gab, die viele Zuschauen gelockt haben.

Erwartung:	Wir denken, dass es sicher Starschauspieler/innen geben wir, die einen gewissen Fan-Base besitzen. In solchen Fällen kann eine erhöhte Besucherzahl erwartet werden.

These / Frage:	Wer ist die produktivsten Schauspieler und in welchen Genres haben sie am meisten gespielt?
Filterung:	Für diese Analyse werden für die Schauspielern mit dem Genre vergleichen. Bevor wir diese Analyse machen können, müssen wir die Schauspieler filtrieren. Aus diesem Grund haben wir die produktiven Schauspieler mit den meisten Filmen filtriert. Mit diesen Schauspielern werden wir analysieren in welchen Genres sie am meisten gespielt haben.
Analyse:	Wir wollen sehen, wer die meisten Filme dreht und was dabei die Genres dieser Filme ist.
Erwartung:	Bei den ersten Analysen hat sich ergeben, dass Drama, Comedy und Action einer das beliebteste Genre sind. Aus diesem Grund gehen wir davon aus, dass auch dieses Genre von den produktivsten Schauspielern gespielt wird.

These / Frage:	Welche Filmgenres sind die beliebtesten?
Filterung:	Hier wird der Durchschnitt der Besucherzahlen pro Genre analysiert.
Analyse:	Diese Analyse geht vor allem nur durch Python, weil wir hier nummerische Spalten haben und dadurch das besser rechnerisch angehen statt grafisch.
Erwartung:	Es wird erwartet, dass Actionfilme die meisten Besucher anlocken, weil es auch nach unseren Vorstellungen die höchsten Kosten verursacht.

Fragen und Unklarheiten?