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:

Teammitglieder

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

* Vincenzo Timmel

Welche Tools gedenken Sie einzusetzen für die SNA?

Gephi

Pajek

R

Python

Anderes? Falls ja, welche?

# Datenquelle

Woher kriegen Sie Ihre Daten?

<https://snap.stanford.edu/data/twitch_gamers.html>

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?

Für die Datensätze konnte ich keine Angabe zu den AGBs oder Terms of use etc. finden auf der Seite. Im Forum hat jemand aber genau diese Frage gestellt: <https://groups.google.com/g/snap-datasets/c/t-jP8cTDqvs/m/EImi4ArBBgAJ>

Das Benutzen der Datasets für unsere akademische Zwecke ist also OK.

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?

Sind nicht limitiert, da es sich um ein .ZIP File handelt.

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

Einlesen mit `open` oder `pd.read\_csv`.

# 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?

Ich denke, es hat hier zwei Möglichkeiten, welche ich erforschen werde: Es wird unterscheidet zwischen im Datenset zwischen einem Zuschauer und einem Twitch-Partner. Somit lässt es sich theoreisch ein two-mode Netzwerk erstellen.

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: 168,114 Anzahl Kanten: 6,797,557

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.)

* Numeric\_id: Anonymisierte ID eines Users
  + Wohl sehr gute Datenqualität, da diese ID das ganze Netz zusammenbaut und der Datensatz gut gereinigt ist.
* Affiliate: Ob der User ein Twitch-Partner ist, oder nicht.
  + Gute Datenqualität, da ein Bool.
* Dead Account: Ob der Account vom User deaktiviert wurde, oder eine lange Zeit inaktiv ist.
  + Gute Datenqualität, da ein Bool.
* Views: Wieviele totale Views der User hat.
  + Denke gute Datenqualität, da einfach auslesbar und der Datensatz gut gereinigt ist.
* Mature: Ob der Kanal vom USer 18+ ist.
  + Denke gute Datenqualität, da einfach auslesbar und der Datensatz gut gereinigt ist. Ebenfalls ist es ein Bool.
* Language: Sprache des Channels
  + Datenqualität ist nicht klar, muss noch analysiert werden.
* Created\_at: Wann der User erstellt wurde.
  + Datenqualität ist nicht klar, muss noch analysiert werden.
* Updated\_at: Wann der User das letzte Mal gestreamed hat.
  + Datenqualität ist nicht klar, muss noch analysiert werden.
* Life\_time: Zeitabstand zwischen User-Erstellung und letztem Stream.
  + Datenqualität ist nicht klar, muss noch analysiert werden.

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?

Die Views könnte man Binnen und so Kategorien erstellen.

# 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: | Wie sehen die Communties auf Twitch aus? Gibt es User, welche das ganze Twitch verbinden, gibt es abgetrennte (keine edges zu anderen communities) communities? |
| Filterung: | Nein |
| Analyse: | Weiss ich noch nicht, dies muss ich noch lernen in der Theorie |
| Erwartung: | Ich denke, es wird User haben, welche eine grosse Allgemeinheit und User, welche nur eine sehr kleine Gruppe ansprechen. |

|  |  |
| --- | --- |
| These / Frage: | Gibt es abgetrennte Communities? |
| Filterung: | Nein |
| Analyse: | Weiss ich noch nicht, dies muss ich noch lernen in der Theorie |
| Erwartung: | Ich denke nicht, dass ich eine vollständige, abgetrennte Community finden werde. |

# Fragen und Unklarheiten?

Nicht wirklich, eher das mein Team nur aus mir besteht. Dies war mir im Vorhinein nicht bewusst.