Projektarbeit – Soziale Netzwerkanalyse

Einführung

Sie erlernten in diesem Modul während des Unterrichts hauptsächlich die theoretischen Grundlagen der Sozialen Netzwerkanalyse sowie den Umgang mit Tools. In der Projektarbeit wird dieses erworbene Wissen vertieft und praktisch angewendet.

Sie werden als kleine Gruppe ein Thema frei wählen, welches schlussendlich genauer untersuchen werden soll. Achten Sie bei Ihrer Themenwal darauf, dass Daten für Ihre Analysen öffentlich verfügbar sind.

Administratives

Gruppengrösse & Gruppenzuordnung:

Arbeiten Sie in Gruppen à 2-3 Personen. Sie sehen im DS Space (Bsc Data Science) oder bei der Liste der Team-Mitglieder in MS Teams (Bsc Informatik / iCompetence), welche anderen Personen ebenfalls da Modul besuchen.

Tragen Sie Ihre Gruppen in die nachfolgende Excel-Tabelle ein:

Bachelor Informatik / iCompetence (Gruppennummer 1 – 20): https://fhnw365-my.sharepoint.com/:x:/g/personal/michael_henninger_fhnw_ch/ESpms_L-9pxNji7Ncwt8hbcB7hTHHyic_5hFKK_73ajdvg?e=9lL8wh

Bachelor Data Science (ab Gruppennummer 21): https://fhnw365-my.sharepoint.com/:x:/g/personal/michael_henninger_fhnw_ch/EXNE-v5oODxPp0EnyFDiRccBq_n86lRqN0rdbCq1gbovuA?e=f28Paf

Kommunikation:

Bevorzugt wird die Kommunikation über einen Chat in Microsoft Teams. Sie können jedoch auch ein E-Mail schreiben. Wichtig: Geben Sie immer Ihre Gruppen-Nummer an: Den Chat bitte umbenennen zu „SAN-Projekt – Gruppe xx“, wo dann alle Teammitglieder und ich eingeladen werden. (Bitte verwenden Sie für die interne Kommunikation dann einen anderen Chat, ansonsten wird es bei mir sehr unübersichtlich).

Themenfindung

Suchen Sie in der Gruppe ein Thema, welches Sie genauer untersuchen möchten. Berücksichtigen Sie unbedingt, welche Daten Sie dafür verwenden möchten. Behalten Sie im Hinterkopf, dass mit der Menge der verfügbaren Informationen auch die Menge der Analysemöglichkeiten steigt. Grundsätzlich gilt:

- Gerichtete Kanten sind informativer als ungerichtete Kanten
- Umso mehr Attribute auf Knoten und Kanten existieren, desto umfangreicher sind die Analyse- und Filtermöglichkeiten
- Multirelationale Netzwerke ermöglichen Analysen aus verschiedenen Perspektiven, können jedoch nicht mit allen Tools entsprechend abgebildet werden. Gephi unterstützt noch keine Multirelationalen Netzwerke. Erstellen Sie einfach ein einzelnes Netzwerk für jede Relation.
- Verschiedene Netzwerke ermöglichen einen Vergleich der Netzwerk-Metriken (können auch verschiedene Sub-Netzwerke sein)

Nehmen Sie sich für die Voruntersuchung der Datenquelle genug Zeit! Sie können auch verschiedene Netzwerke kombinieren. In Gephi kann beim Öffnen eines Graphen angegeben werden, ob ein neuer Graph erzeugt werden soll, oder ob die Graph-Daten zum bestehenden Graphen (*Append Graph*) hinzugefügt werden sollen (Knoten-ID muss jedoch übereinstimmend sein für eine Vereinigung). So können Sie die beiden Graphen separat, jedoch auch vereint untersuchen und schauen, ob es Überlappungen gibt.

Projektablauf

Das Projekt kann als zwei Teile (Datensammlung / explorative Analyse und Netzwerkanalyse) betrachtet werden. Falls Sie während dem Sammeln oder der explorativen Analyse der Daten bemerken, dass der Datensatz für die weiterführende Analyse ungeeignet erscheint, melden Sie sich beim Dozenten, so dass dafür eine Lösung gefunden werden kann (anderer Datensatz oder sogar komplett anders Thema).

Datensammlung

Sie können entweder einen bereits aufbereiteten Datensatz (z.B. von kaggle) verwenden oder selbst Daten einsammeln, wobei zweiteres Bonuspunkte gibt. Melden Sie sich beim Dozenten, falls Sie nicht frei zugängliche Daten für die Projektarbeit verwenden möchten. Beachten Sie die Privatsphäre von Drittpersonen sowie Nutzungsbestimmungen.

Mögliche Quellen zum Einsammeln von Daten sind beispielsweise:

- **APIs**: Verschiedene Internet-Dienste bieten APIs an (Schnittstellen für Programmierer), über welche auf deren Daten zugegriffen werden können. (z.B. Wikipedia API, Twitter API usw.)
- **Crawler**: Implementation eines eigenen Webcrawlers

Die Implementation eines eigenen Crawlers, welcher auch Attribute und Beziehungen liefert, ist sicher aufwändiger als die Anbindung einer API. Beim Einsammeln der Daten empfiehlt es sich, die Rohdaten zu persistieren. So brauchen bei nachträglichen Anpassungen von Transformationen nicht nochmals die Daten einzusammeln.

Explorative Datenanalyse

Nachdem die Daten beschaffen wurden, sollen einfache Datenanalysen (Häufigkeitsanalyse, Statistiken, Daten-Visualisierungen…) helfen, die Daten besser zu verstehen und Unstimmigkeiten wie auch fehlende Werte in den Daten frühzeitig zu erkennen. Passen Sie wenn nötig ihre Programme zur Datensammlung und Daten-Transformationen an, wenn Sie Probleme mit der Datenqualität erkennen. Prüfen Sie auch (wenn möglich), ob die eingesammelten Daten plausibel sind und mit der Datenquelle / Realität (z.B: anhand einer Stichprobe) übereinstimmen. Mögliche Probleme:

Dieser Schritt ist wichtig, ist aber schlussendlich im Bewertungsraster im Vergleich zu anderen Kriterien nur schwach gewichtet. Ich empfehle die Nutzung einer Python-Library wie z.B. *Pandas Profiling* oder *Sweetviz*, welche das Leben für die explorative Datenanalyse erheblich vereinfacht.

Soziale Netzwerkanalyse

In der Analysephase das das eingesammelte Netzwerk / die eingesammelten Netzwerke analysiert.

Im Meilenstein haben Sie sich überlegen, welche Analysen Sie auf dem Netzwerk ausführen möchten. Es sollten verschiedene Analysen aus dem SAN-Unterricht sein. Sie dürfen aber gerne auch nicht im Unterricht behandelte Netzwerk-Analysen durchführen. Es ist auch gut möglich, dass Sie plötzlich im zweiten Teil noch weitere Ideen für Analysen haben, die Sie nicht im Meilenstein erwähnt haben. Diese dürfen Sie gerne umsetzen.

Zeitplan & Termine

1. Individuell: Abgabe des Projekt-Meilensteins (siehe Unterabschnitt „Meilenstein“ im Kapitel «Abgaben»). Empfohlener Zeitpunkt: Vor der Umsetzung des Projekts. Das Feedback zum Meilenstein erfolgt schriftlich.
2. 31.12.2023, 23:59:59: Abgabe des Projektes (Inhalt im nachfolgenden Abschnitt «Abgaben» festgelegt). Das Projekt wird anhand des zur Verfügung gestellten Bewertungsrasters bewertet.
3. Individuell (ca. 2 Wochen nach Abgabe der Projektarbeit): Projekt Schlussgespräch (ca. 15 Minuten).

Abgaben

Ort und Ordnerstruktur

Erstellen Sie in Microsoft Teams im Kanal „Allgemein“ einen Unterordner im Ordner „Gruppenarbeiten“ mit ihrer Gruppennummer und einem Titel (z.B. «02_E-Mail-Analyse») und platzieren Sie da alle dort alle Abgaben. Es wird folgende Unterordner-Struktur erwartet:

- **Meilenstein:** Ausgefüllter Meilenstein
- **Präsentation:** Aufgezeichnete Präsentation (Zeitlimit: Maximal 20 Minuten (nach 20 Minuten wird das Video gestoppt und alles, was nachher kommt (z.B. Ausblick) nicht mehr gewertet resp. als «nicht vorhanden» gewertet.)). Der erwartete Inhalt wird weiter unten erläutert.
- **Source:** Source-Code des Fetchers / Datenverarbeitungs-Scripts / Jupyter Notebooks.
- **Rohdaten:** Rohdaten-Ausgaben
- **Graphen:** Ihre Graphen (z.B. im CSV-Format / SQL Dump / Gephi File, NetworkX Export …)

Falls Sie die Daten nicht über diesen öffentlichen Kanal abgeben können (z.B., weil die Daten nicht öffentlich verfügbar sind), melden Sie sich. Sollten die Abgaben ohne vorherige Absprache nicht via MS Teams (Dateien im Kanal «Allgemein» -> «Gruppenarbeiten») erfolgen, hat dies ein Abzug in der Schlussnote zur Folge.

Meilenstein

Im Ordner „Meilenstein“ finden Sie den Projekt-Meilenstein. Dieser dient als roten Faden für das Projekt. Initial definieren Sie, was von welcher Datenquelle in welcher Team-Zusammensetzung genauer untersucht werden sollte. Sie beschäftigen sich mit der Datenquelle sowie auch möglichen Hindernissen. Der Meilenstein ist **NICHT** dazu gedacht, dass Sie seitenweise Text schreiben müssen / sollen. Beschreiben Sie kurz und klar.

Präsentation (max. 20 min)

Wichtig: Konsultieren Sie unbedingt die Bewertungskriterien weiter und und prüfen Sie, dass alle darin erwarteten Punkte in der Präsentation enthalten sind. Erwähnen Sie jeweils, in welchen Dateien / Notebooks die Implementation dazu gefunden werden können. Der Fokus sollte klar auf der Sozialen Netzwerkanalyse liegen. Der vorgeschlagene Ablauf ist:

1.) **Datenbeschaffung**: Beschreiben Sie kurz, wie und woher Sie die Daten einsammeln und ein paar zentrale Analyse-Ideen, damit ein grober Einblick in Ihr Vorhaben greifbar wird. Erklären Sie auch, wie sie überprüft haben, ob die Datenquelle geeignete (vollständige und korrekte) Daten liefert und ob allfällige alternativen oder weitere Datenquellen verwendet wurden.
2.) **Explorative Datenanalyse:** Entsprechen die Rohdaten den Erwartungen / der Realität? Was wurde in der explorativen Datenanalyse wie analysiert? Welche Erkenntnisse und Massnahmen leiten Sie für die weiteren Analysen ab? Musste das Einsammeln oder Transformieren

der Daten nochmals überarbeitet werden? Es wird erwartet, dass Sie sich mit den Daten auseinandersetzen, Probleme erkennen und entscheiden, wie damit umgegangen wird

3.) **Modellierung:** Beschreiben Sie, welche/s Netzwerk/e Sie aus den eingesammelten Netzwerken wie modelliert haben (was wurde al Knoten / Kanten abgebildet, Ist es ein One- oder Two-Mode Netzwerk, welche Transformation wurde verwendet, welche (wichtigen) Attribute gibt es auf den Knoten / Kanten usw.
4.) **Soziale Netzwerkanalyse:** Hier werden die Netzwerkanalysen vorgestellt. Erwähnen Sie gewisse Kennzahlen zum Netzwerk (z.B. Anzahl Knoten und Kanten) und führen Sie dann Ihre Analysen aus. Wenn Sie auf gefilterten Netzwerken arbeiten, dann erwähnen Sie die Filter-Kriterien. Wenn Sie mehrere Netzwerke modelliert haben, ist es wichtig jeweils zu erwähnen, welches Netzwerk nun verwendet wird. Wichtig ist, dass Sie beschreiben, was Sie wie analysieren und ihr Resultat auf Ihren Anwendungsfall bezogen interpretieren. Zahlen und Grafiken ohne Interpretation in Bezug auf den vorliegenden Anwendungsfall werden bei der Bewertung keine Beachtung geschenkt und kann sich auch negativ auswirken, wenn der Eindruck entsteht, dass Quantität vor Qualität gestellt wurde.
5.) **Ausblick**: Was wäre Ihrer Meinung nach noch interessant mit Sozialer Netzwerkanalyse zu analysieren, wenn das Projekt weitergeführt würde? Gäbe es noch weitere interessante Fragestellungen / weitere interessante Attribute, die eingesammelt oder sauber verarbeitet werden müssten?
6.) **Lessons Learned**: Was lief gut? Was würden Sie in Zukunft anders machen?

Bewertungskriterien:

Die Bewertung wird entsprechend dem beiliegenden Bewertungskriterien durchgeführt. Ausschlaggebend sind die Informationen aus der Präsentation und der Schlussbesprechung. Es wird jedoch auch ein Blick in die praktische Umsetzung geworfen, weshalb alle in der Präsentation vorgestellten Analysen und Rohdaten ebenfalls abzugeben sind. Bewertet werden folgende Kriterien:

Leicht gewichtete Kriterien:

- **Explorative Datenanalyse / Validierung**
  - Explorative Datenanalyse zu den eingesammelten und in der Netzwerkanalyse verwendeten Attribute (Verteilungen der Werte (sind diese plausibel?), Vergleich mit der «Realität», Untersuchen wie viele Daten fehlen (missing Values), ob es ungewöhnlich hohe oder tiefe Werte (Ausreisser) gibt, …. Es soll schlussendlich anhand dieser Analyse entschieden werden können, ob und welche Daten qualitativ genug hergeben, um diese für weitere Analysen zu verwenden und welche Attribute verworfen resp. nochmals neu eingesammelt oder transformiert werden müssen. Ein Vergleich mit der «Realität» (z.B. Datenquelle oder alternative Datenquellen) hilft zu erkennen, ob beim Sammeln der Daten Probleme aufgetreten sind.
  - Welches sind die Erkenntnisse und Schlussfolgerungen aus der Explorativen Datenanalyse?
- **Anschaulichkeit der Resultate**
  - Sind die Netzwerke visuell ansprechend dargestellt? Unterstützt die Visualisierung die Kernaussage? Werden verschiedene Möglichkeiten zur Hervorhebung verwendet?
- **Ausblick**
  - Ausblick über weitere Analysen aus dem Bereich der Netzwerkanalyse. Welche fortführenden Analysen könnten noch gemacht werden, wenn mehr Zeit zur Verfügung

steht? Welche weiteren Daten und Datenquellen wären spannend hinzuzuziehen, um dann weitere (und welche) Soziale Netzwerkanalysen zu ermöglichen?

Stark gewichtete Kriterien:

- **Vielfalt der Analysen:** Unterschiedliche Netzwerkanalyse-Methoden, Themengebiete oder Netzwerke/Subnetzwerke werden berücksichtigt.
  - Seien sie sparsam mit Degree Centrality Analysen. Auch wenn diese Metrik wohl die intuitivste ist, können Degree-Centrality Analysen grundsätzlich auch ohne die Netzwerkstruktur gemacht werden (z.B. mit einem simplen *«Select count(*) .. group by..»* Statement).
- **Korrekte Anwendung der Masse und Metriken der Netzwerk-Analyse:**
  - Angewendete Netzwerkanalyse-Methoden sind für den Anwendungsfall geeignet und korrekt angewendet.
- **Nachvollziehbarkeit der Analyse & Plausibilität der Schlussfolgerungen:**
  - Klare, präzise und nachvollziehbare Erklärungen zu den Analysen und Schlussfolgerungen. Interpretation auf den Anwendungsfall und Datensatz bezogen (Keine allgemeine Aussagen wie «Die Degree-Zentralität sagt aus, wie viele direkte Verbindungen ein Knoten hat», sondern was dies konkret in Ihrem Anwendungsfall bedeutet (z.B. Die Degree-Zentralität sagt aus, wie viele Freunde eine Person auf Facebook zum Zeitpunkt X der Datensammlung gehabt hat»)). Vermeiden von Fehlinterpretationen oder übermäßige Verallgemeinerungen und Berücksichtigung möglicher Einflussfaktoren.

Zusatzpunkte gibt es für:

- Sehr aufwändige Datensammlung & Aufbereitung
- Besonders kreative Netzwerkanalysen
- Weiterführende Netzwerk-Analysen, die über den behandelten Stoff des Moduls hinausgehen.

Abzug gibt es für:

- Dateien ohne vorherige Absprache nicht im Gruppenordner in MS Teams abgegeben
- Verspätete Abgabe.

Der Notenmassstab soll sich an demjenigen der Bachelor-Arbeit orientieren. Darin ist folgendes vermerkt: «*Grundsatz: Die Note 5.0 ist zu erteilen, wenn für das jeweilige Kriterium die Leistung in vollem Umfang die Anforderungen an eine(n) in der Industrie tätige(n) Ingenieur(in) erfüllt.*»

Hinweise

- Sie dürfen das Thema auch wechseln, wenn es zu dem von ihnen ausgewählten Thema zu wenig Daten gibt oder diese nicht verwertbar sind.
- **Jegliche Form von Plagiat** (nicht deklarierte Kopien aus dem Internet oder anderer Projektarbeiten) führt automatisch zur Note 1 für die komplette Gruppe.
- Sollte ein Gruppenmitglied die Zusammenarbeit verweigern, muss dies frühzeitig gemeldet werden, damit entsprechende Massnahmen vollzogen werden können.
- Gephi kann auch Daten über die Zeit dynamisch visualisieren.
- Gephi kann mit dem GeoLayout Knoten geografisch anordnen, wenn Latitude und Longitude Informationen vorhanden sind.