

Entwicklung einer Toolbox zur Evaluation von LLM-basiertem semantischem Clustering

Diplomand



Lukas Derungs

Problemstellung: Das IFSAR führt regelmässig Umfragen durch, die offene Fragen enthalten. Diese ermöglichen es den Befragten, ihre Meinung frei zu äussern und neue Themen einzubringen.

Die Auswertung solcher Antworten ist jedoch aufwendig: Sie erfolgt manuell und ist zeitintensiv. Offene Fragen liefern häufig viele unterschiedliche Antworten, was eine automatisierte Auswertung erschwert.

Ziel der Arbeit: Ziel des Projekts war es daher, eine Toolbox zu entwickeln die auch Benutzern mit wenig Programmiererfahrung ermöglicht, offene Antworten effizient zu analysieren und zentrale Themen sichtbar zu machen.

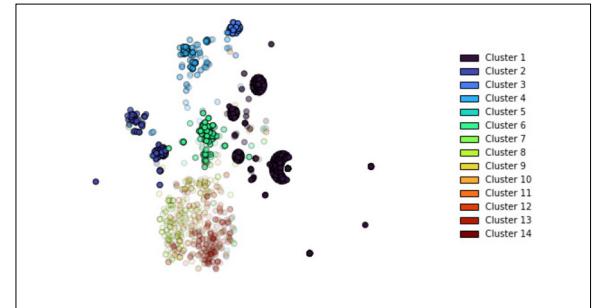
Dazu sollte die Toolbox in der Lage sein, semantisch ähnliche Antworten automatisch zu gruppieren, zentrale Themen mithilfe eines Sprachmodells zu erklären und die Ergebnisse zu visualisieren. Ein zusätzliches Ziel war die Entwicklung von Darstellungen, die zeigen, wie stabil die gefundenen Antwortgruppen bei wiederholter Analyse sind. So erhalten Benutzer eine Einschätzung darüber, wie zuverlässig die Gruppierungen sind und ob sich Überschneidungen zwischen den Gruppen ergeben.

Ergebnis: Die Toolbox wandelt offene Antworten mithilfe externer Embedding-Modelle in numerische Vektorrepräsentationen um, die anschliessend in semantisch ähnliche Gruppen eingeteilt werden. Um die Qualität und Verlässlichkeit dieser Gruppierungen besser einschätzen zu können, visualisiert die Toolbox die Stabilität der Gruppierungen anhand zweier Darstellungen:

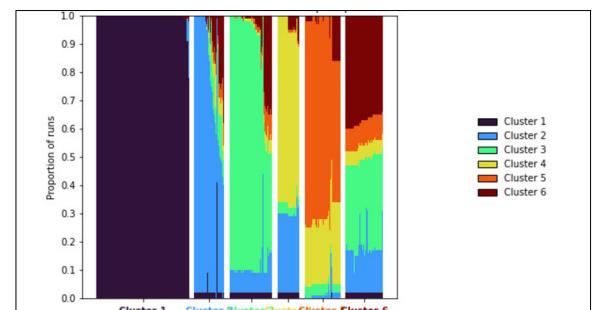
Beim Scatterplot werden die Antworten als Punkte dargestellt, farblich nach Themen gruppiert. Je transparenter der Punkt, desto öfters wechselte sich seine Zuordnung.

Der Balkendiagramm zeigt für jede Antwort, wie oft sie in den verschiedenen Gruppen eingeteilt wurde. Dabei sind auf der X-Achse die Antworten entsprechend ihrer häufigsten Gruppenzugehörigkeit sortiert, während die Y-Achse den prozentualen Anteil der Zuweisungen über alle Durchläufe hinweg darstellt

Aufteilung und Stabilität der Gruppen im Scatterplot
Eigene Darstellung



Gruppenzuordnung im Balkendiagramm
Eigene Darstellung



Referent

Dr. Beat Tödtli

Korreferent

Leo Büttiker, yonesu
GmbH, Olten, SO

Themengebiet

Data Science

Projektpartner
IFSAR (OST)



Diese Darstellungen helfen Benutzern die entstandenen Gruppierungen, deren Stabilität und mögliche Überschneidungen einzuschätzen. Die Resultate können danach in eine Excel-Datei exportiert werden um sie weiter zu analysieren.