Seminararbeit AI-assisted programming and data analysis

Data Profiling with LLM's - A Case Study on HR Data

Vincent-Konstantin Kapp Prof. Dr. Martin Spindler

31. October 2024

Abstract

Datengetriebene Prozesse können nicht nur Unternehmen beim Optimieren der Prozesse sondern auch bei der entwicklung neuer Strategien helfen. Aber uum festzustellen welche Daten in den Datenbanken vorliegen und wie diese zu einernander stehen braucht es Mitarbeiter*innen die einzelnd überprüfen welche Daten es gibt und in welchen zusammenhängen, ausgehend von Commen Sense, es gibt. Data Profiling stellt Unternehmen mit historisch gewachsene Datenbanksysteme vor Herausforderungen für die Dategetriebene Transformation da. Können die aktuellen entwicklunge von Large Language Modellen (LLM) dazu beitragen, autamtisiert die zusammenhängen von Spalten zu erkennen? Um diese Frage zu beantworten, widmet sich diese Seminararbeit aufbauen von der Arbeit von Trummer 2024, wie gut LLM, anhand einer Case Study, die Korrelation von Daten zu erkennen.

1 Einleitung

2 Hintergrund

2.1 Data Profiling

2.2 Korrelationserkennung in Datenbanken

(Trummer 2023).

2.3 Sprachmodelle und deren Anwendung in der Datenanalyse

(Arora et al. 2023).

3 Methodik der Fallstudie

3.1 Datenbasis und Vorbereitung

3.2 Vorgehensweise zur Korrelationsanalyse mit LLM

3.3 Benchmark und Bewertungsmetriken

4 Durchführung der Fallstudie: Gender-Pay-Gap-Analyse

(Bach, Chernozhukov, and Spindler 2024).

References

Arora, Simran, Brandon Yang, Sabri Eyuboglu, Avanika Narayan, Andrew Hojel, Immanuel Trummer, and Christopher Ré. 2023. "Language Models Enable Simple Systems for Generating Structured Views of Heterogeneous Data Lakes." Journal Article. *arXiv Preprint arXiv:2304.09433*.

Bach, Philipp, Victor Chernozhukov, and Martin Spindler. 2024. "Heterogeneity in the US Gender Wage Gap." *Journal of the Royal Statistical Society Series A: Statistics in Society* 187 (1): 209–30.

Trummer, Immanuel. 2023. "Can Large Language Models Predict Data Correlations from Column Names?" Journal Article. *Proceedings of the VLDB Endowment* 16 (13): 4310–23.