ABSTRACT

In the age of big data the field of Data Science gets more relevant. A lot of data has to be collected, cleaned and analyzed. But the role of a Data Scientist is broad. The responsibilities reaches from data collecting und cleaning, analyzing, visualizing to Machine Learning and Artificial Intelligence. But how is the work split in Data Science projects? Are specific roles assigned or does everyone work every role? Especially for recruiting this topic is interesting. The result could give insight on recruiting more specifically. Even the job descriptions and titles could get more specific to the employees wishes.
Within the scope of this paper Data Science projects will be analyzed to find role distribution. For this task publicly available projects hosted on GitHub are used as the database. Specifically only IPython files will be analyzed, since it's assumed it's more likely to find Data Science projects there compared to other formats. Based on the commit changes the written source code will be assigned to its author. Using abstract syntax trees used functions are extracted and matched with their module.
The collected data will be grouped by authors. Based on the functions used a topic modeling will be done to find a possible role distribution. A distribution of used functions can be observed, although it isn't possible to deduce subject areas.

ZUSAMMENFASSUNG

Im Zeitalter der großen Datenmengen wird der Bereich Data Science immer relevanter. Viele Daten müssen erhoben, bereinigt und analysiert werden. Doch die Rollen des Data Scientist sind weitreichend. Die Aufgabenbereiche reichen von Datenerhebung, -bereinigung, Analyse und Visualisierung bis hin zu Machine Learning und künstliche Intelligenz. Doch wie erfolgt die Aufgabenverteilung innerhalb eines Data Science Projektes? Werden spezifische Rollen angenommen oder bearbeitet jeder jeden Aufgabenbereich? Gerade fürs Recruiting ist diese Fragestellung interessant, da mit dieser Erkenntnis gezielter nach potenziellen Arbeitnehmern gesucht werden kann. Auch die Stellenbeschreibungen selbst können dadurch spezifischer auf die Arbeitnehmer angepasst werden.
Im Rahmen dieser Arbeit werden Data Science Projekte auf Rollenverteilung untersucht. Hierbei werden öffentliche Repositories von GitHub als Datenbasis verwendet. Es werden ausschließlich IPython Dateien untersucht, da angenommen wird, dass in dieser Umgebung am wahrscheinlichsten Data Science Projekte geschrieben sind. Anhand der Commit Änderungen wird geschriebener Source Code den jeweiligen Autoren zugewiesen. Mit Hilfe von Abstract Syntax Trees (AST) werden im Source Code aufgerufene Funktionen extrahiert und den entsprechenden Modulen zugewiesen.
Die gesammelten Daten werden nach Autoren gruppiert. Anhand der verwendeten Funktionen wird ein Topic Modeling durchgeführt um mögliche Rollen abzuleiten. Es kann eine Verteilung von aufgerufenen Funktionen gebildet werden, diese ist aber nicht eindeutig in feste Themenbereiche unterteilbar.