PROJEKTAUSSCHREIBUNG
2020/2021
Prof. Dr. Martin Przewloka

Vergleich von open source web crawlern am Anwendungsfall xyz für das data mining

Problemstellung/Ausgangssituation:

Das Internet stellt eines der essenziellsten Datenquellen unserer Zeit dar. Die Informationszunahme steigt exponentiell und die manuelle Generierung, Wertung und Weiterverarbeitung von Informationen ist schon längst an ihre Grenzen gestoßen. In diesem Zusammenhang spielen maschinelle Werkzeuge zur automatisierten Datengenerierung basierend auf öffentlich, im Internet zugänglichen Informationen, eine zentrale Rolle. Die sogenannten web crawler oder scraper sind in der Lage, HTML-Seiten automatisiert aufzurufen und die gewünschte(n) Information(en) herauszulesen. Die so gewonnenen Daten können mit entsprechenden Anwendungen weiter prozessiert, wie bspw. statistisch ausgewertet werden. Eine besondere Bedeutung kommt hierbei den open source – Plattformen zu, die in den letzten Jahren erheblich zum Fortschritt dieser Domäne beigetragen haben.

Masterarbeit, ggf. auch Bachelorarbeit:

Im Rahmen einer Masterarbeit soll ein Vergleich von mindestens 3 führenden open source web crawler Systemen durchgeführt werden, indem ein konkreter Anwendungsfall umzusetzen ist. Der Fall kann direkt aus dem Unternehmenskontext des Studierenden stammen. Alternativ kann er auch in Absprache mit der Institutsleitung vor der Umsetzung des Projekts identifiziert werden. Der Fall erfordert zwingend eine nur maschinell umsetzbare Datengenerierung (bspw. aufgrund der zu generierenden Datenmenge) und eine weitere Aufbereitung der Daten (bspw. eine statistische Analyse). Optional können auch Werkzeuge der maschinellen Sprachverarbeitung (bspw. eine Sentimentanalyse) zum Einsatz kommen.

Ziel ist es, mit den ausgewählten Systemen den identischen Anwendungsfall umzusetzen und anhand der Erkenntnisse wie auch dem erzielten Ergebnis einen Vergleich durchzuführen. Die hierfür notwendigen Kriterien und ein damit einhergehendes Bewertungsverfahren sind ebenfalls zu entwickeln.

Zum Abschluss der Arbeit sind Handlungsempfehlungen für den Praktiker und Entscheider zu geben, die unmittelbar aus den Erfahrungen der Projektumsetzung heraus resultieren. Optional kann die Arbeit zudem mit einem qualitativen Vergleich zu alternativen Verfahren der Datengenerierung ergänzt werden.

Zielgruppe:

Die ausgeschriebene Arbeit adressiert:

- Technisch mittel aber auch stark versierte Studierende
- Betriebswirtschaftlich orientierte Studierende, die Interesse an einer modernen Datengenerierung haben, und über Grund-Programmierkenntnisse verfügen
- Betriebswirtschaftlich und/oder technisch Interessierte aus dem Beratungs- und Projektmanagementumfeld

Erweitertes Ziel ist es, die Ergebnisse der Arbeit zu veröffentlichen, sowie ggf. die gewonnenen Kenntnisse im Berufsumfeld anzuwenden.