Webspidering mit KI-Unterstützung

Studiengang: MAS Data Science

Künstliche Intelligenz ist in unserem Alltag bereits weit verbreitet und ist in der Lage Autos autonom fahren zu lassen, faszinierende Roboter und Drohnen zu steuern, Fotos aufzubereiten oder sonstige nützlichen Alltagsaufgaben diskret und zuverlässig zu erledigen. Es liegt auf der Hand, dass sich damit vermutlich auch ein Webbrowser als Recherchewerkzeug automatisch bedienen lassen kann.

Problemstellung

Wir verwenden in einigen Projekten Daten von externen, öffentlichen Webseiten welche mittels automatisierten Webcrawlern (Bots) extrahiert und für die Weiterverwendung aufbereitet werden. Dieser Prozess ist sehr aufwandintensiv in Bezug auf Entwicklung und Wartung, da jede Webseite spezifisch konfiguriert werden muss. Jede planmässige oder unvorhersehbare Änderung an einer externen Webseite erfordert eine zeitnahe Anpassung an der entsprechenden Konfiguration auf unserer Seite. Dieser Lösungsansatz ist nicht skalierbar für eine grössere Anzahl Webseiten. Im Rahmen dieser Master Thesis soll ein moderner, generischer Lösungsansatz aufgebaut und evaluiert werden, mit Unterstützung von künstlicher Intelligenz (KI).

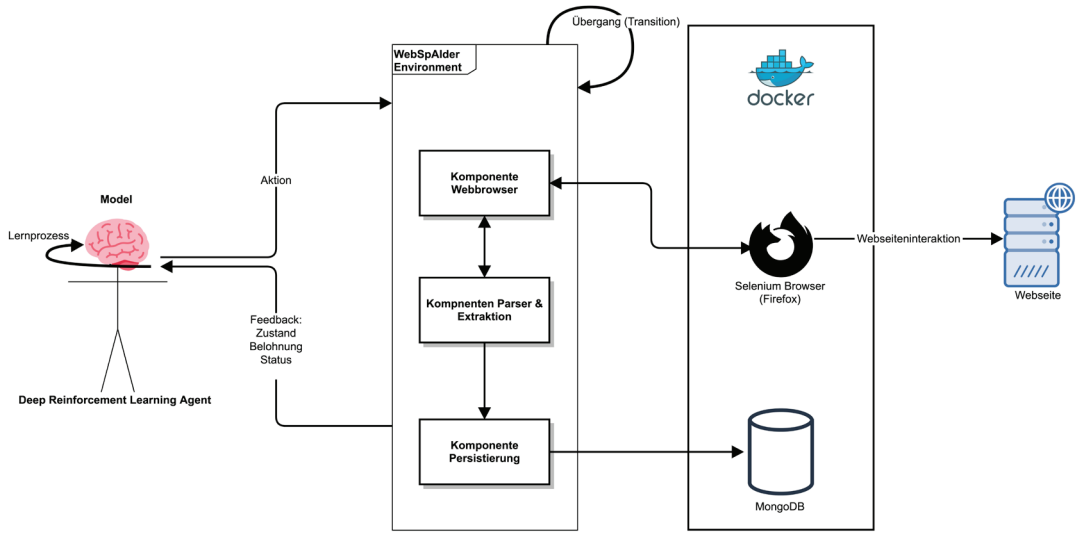
Lösungsaufbau

Es wurde eine Deep Reinforceement Learning Umgebung aufgebaut, welche von einem generischen Agenten angesteuert und angelernt werden kann. Die Umgebung übernimmt die Bedienung eines Selenium Webbrowsers und simuliert auf der entsprechenden Webseite einen normalen Benutzer. Jede Aktion wird ausgewertet und an den Agenten zurückgemeldet. In der Initialversion übernimmt ein DQN Agent die Ansteuerung der Umgebung und wird darauf trainiert, die gesuchten Fachbegriffe möglichst schnell auf der Webseite zu finden. Die gefundenen Begriffe und dazugehörige Informationen werden mit einem generischen Prozess aufbereitet und zur Weiterverwendung bereitgestellt.

Resultate

Es konnte eine Lösung aufgebaut werden, welche selbstständig die Bedienung einer nahezu beliebigen Webseite erlernt und nach fachspezifischen Inhalten suchen kann. Jede neue Webseite muss mindestens einmalig automatisch angelernt werden, damit die Daten dann periodisch möglichst effizient wieder neu extrahiert werden können. Die gesuchten Informationen werden sauber strukturiert in einer MongoDB bereitgestellt.

Michael Bieri
079 699 23 82
michael@bieri.club



Aufbau Deep Reinforcement Learning Umgebung