WDB Mini-Challenge 2

Einführung

Wir haben einen Selenium-Scraper für die Webseite CASH, spezifisch ihre News (unter https://www.cash.ch/news/alle verfügbar) erstellt.

CASH ist die grösste Schweizer Finanzplattform mit News, Börsenkursen und Online-Trading. Monatlich besuchen über 300'000 User das Portal. Dabei ist die Seite sehr aktuell: Im Minutentakt werden neue Artikel publiziert. Bei diesen handelt es sich um Analysen mit Wirtschafts- und Finanzinformationen.

Mit unserem Scraper werden diese Artikel automatisch ausgelesen und lokal in einer csv- oder Feather-Datei abgespeichert. Dabei speichern wir die URL, das Veröffentlichungsdatum, den Titel, die Kategorie und den Teaser-Text.

Motivation

Das Ziel dieser Datenbeschaffung ist, dass sie anschliessend gemeinsam mit Börsenkursen abgeglichen werden kann. Unsere These ist, dass Kursschwankungen mit den News korrelieren. Unsere anschliessende Überlegung ist, dass es einen gewissen Delay zwischen den Nachrichten und Kursschwankungen geben muss, da der Aktienmarkt auch von Privatpersonen beeinflusst wird, welche diese Nachrichten erst sehen und verarbeiten müssen, bevor sie eine Handlung davon ableiten. Das Minimieren dieses Delays könnte einen entscheidenden Vorteil im Aktienmarkt bringen. Bei erfolgreichem Beweisen der These könnte also anschliessend ein Modell trainiert werden, um Kursschwankungen auf eine kurze Zeit vorherzusagen.

Eine Schwierigkeit dabei ist, dass es sich bei unseren gescrapten Daten um textuelle Daten handelt. Diese müssen zuerst noch verarbeitet werden, bevor sie in einem Modell verwendet werden können. Dabei kann in einer einfachen Variation der Name des betroffenen Unternehmens im Text gesucht werden und dann anhand des Datums in den Aktienkursen geschaut werden, ob es nach dieser Nennung einen grösseren Effekt gegeben hat.

Mit Natural Language Processing oder ähnlichem könnte man noch weiter gehen und die genau Bedeutung der Neuigkeit erkennen: Genauer gesagt, ob es sich um eine positive, negative oder neutrale Nachricht über das Unternehmen handelt. So könnten noch genauere Analysen gemacht werden.

Wie das Skript die Nachrichten ausliest

Wir finden die benötigten Elemente mittels HTML Klassen-Tags und lesen ihren textuellen Inhalt aus. Dabei werden die Elemente in einem Pandas Dataframe gespeichert.

Datenanalyse

Unsere Datenanalyse ist in data-analysis.ipynb zu sehen.