SeemsDancy

Aktueller Zustand:

Aktuell ist unser Prototyp in der Lage über einen Button auf der Webseite den internen CodeceptJS-Test/ Workflow zu starten. CodeceptJS seitig kamen hier in letzter Zeit allerdings einige Veränderungen. Während bis zu unserer Präsentation vor einigen Wochen noch die Anmeldung und Erreichung von Webseiten auf Facebook ohne größere Hindernisse möglich war, haben sich seit der Sperrung unseres ursprünglichen Facebook-Bot-Accounts hier einige Probleme aufgetan.

Größte Veränderung hierbei ist das einige Buttons wie „Interessiert“ oder „Suche nach“ die vorher ohne Probleme vom Bot zu betätigen waren, nun entweder nicht mehr vom Bot gefunden werden oder einen „Out of Memory“-Fehler hervorrufen (ohne das ein Out of Memory-Fehler hier logisch wäre.). Infolgedessen bricht der Workflow aufgrund dieser Probleme häufig bereits früh ab.

Des Weiteren sind einzelne Elemente auf der Webseite, um Ihnen beispielsweise Informationen zu entziehen, nur schwer bis gar nicht zu erreichen. Daher funktioniert unser Workflow aktuell (auch aufgrund der verlorenen Zeit durch die bereits genannten Probleme) noch mit einer fest vorgegebenen URL zu einem Bild.

Die Übertragung findet anschließend über Flask an den NLP-Prozess statt, der die Informationen aus den Plakaten und Bildern ausliest und im Terminal ausgibt.

Gesammelte Erkenntnisse:

Das Harvesting von Informationen im Internet ist deutlich komplizierter als gedacht. Anbieter wie Facebook versuchen Ihr bestes den Zugang zu Informationen zu erschweren und gehen auch teilweise aktiv dagegen vor, wie man an der Sperrung unseres Accounts, sowie an den Veränderungen auf der Facebook-Webseite sehen kann.

Hierzu wurden auch im Rahmen der Präsentation einige Erkenntnisse zu rechtlichen Themen in Bezug auf das Web-Harvesting gesammelt.

Flask eignet sich gut als Schnittstelle um die Verbindung zwischen Webseite, JS und Python zu ermöglichen. Vor allem auch um Informationen zu übertragen.

Weitere Verbesserungsmöglichkeiten/ Ausblick:

Zur weiteren Verbesserung unseres Prototypen sollen in der Zukunft die Informationen direkt von mehreren Seiten entnommen werden. Damit kann man die Problematik einzelner nicht funktionierender Zugriffe abdämpfen. Außerdem sollen die Ausgaben übersichtlicher und auf der Webseite erfolgen. Die Filterung und Anpassung der Suchbegriffe ist ebenfalls angedacht.