

Projektziel

Die Watchlist Internet ist eine unabhängige Informationsplattform zu Internetbetrug. Aktuell sieht das Team dringenden Handlungsbedarf bei Betrug über Investmentplattformen und bei Fake-Shops, die gefälschte Medikamente oder Nahrungsergänzungsmittel verkaufen. Mit Fraud Seeker sollen daher potenzielle Betrugsopfer im Moment des Zweifels via Suchmaschinen erreicht werden. Werden die Warnungen der Watchlist Internet vor den eigentlichen Fallen angezeigt, können erhebliche Schäden finanzieller oder gesundheitlicher Art verhindert werden. Um möglichst rasch und gezielt entsprechende Warnungen veröffentlichen zu können entwickelten wir einen WebCrawler, der nach diesen Websites und einen meta AdCrawler, der nach der dazugehörigen Werbung sucht. Dabei machten wir uns das „Copy & Paste“-Prinzip der Kriminellen zunutze. Die vielen betrügerischen Webseiten existieren nur, weil sie immer wieder kopiert werden. Die Suche nach solchen (meist textbasierten) Ähnlichkeiten wurde durch einen Crawler automatisiert und die Präventionsarbeit (in Form von Warnlisten und Warnartikeln) deutlich ausgebaut.

Zielgruppe

Insbesondere der WebCrawler unterstützt das Watchlist Internet Team bei der täglichen Arbeit. Die Redakteur:innen sollen dabei neue Textphrasen anlegen und die Ergebnisse bearbeiten können. Wobei die Bearbeitung und Einordnung der Ergebnisse über die Datenbank der Watchlist Internet¹ funktioniert, da eine Schnittstelle dafür sorgt, dass die Crawler-Ergebnisse regelmäßig an uns zur Weiterbearbeitung übermittelt werden.

Der meta AdCrawler wird aktuell ausschließlich vom „Fraud Seeker“-Team verwendet. Die Ergebnisse werden anschließend an die Watchlist Internet Redaktion kommuniziert. Da es sich beim meta-AdCrawler um einen Prototyp handelt, ist derzeit eine breite Anwendung nicht vorgesehen. Informationen zur Funktionsweise finden sich stattdessen in der Entwickler:innen-Dokumentation.

Anleitung WebCrawler

(<https://github.com/oiat/fraud-seeker-webcrawler>)

1. Erste Schritte

- Download WebCrawler (<https://github.com/oiat/fraud-seeker-crawler>)
- Download und Installation Heidi SQL (<https://www.heidisql.com/>)
- API-Anmeldedaten für die Custom Search API erstellen (<https://console.cloud.google.com/projectselector2/apis/library/customsearch.googleapis.com>)

¹ Da es sich bei der Watchlist Internet Datenbank um ein geschlossenes System handelt, wird die Anwendung dieser Datenbank nicht veröffentlicht, stattdessen wird die

2. Tabellen erstellen

- Datenbank mit Heidi SQL erstellen
- config.php im Editor öffnen, Datenbank- und API-Autorisierungsdaten ausfüllen
- In der Datenbank die Datei sqls.sql ausführen
- Tabellen zum Eingeben der Daten und zum Abrufen der Ergebnisse werden erstellt

3. Nach Textphrasen suchen

- Tabelle wi_keywords mit Textphrasen befüllen
- WebCrawler ausführen und Suchanfragen durchführen
 - Call, um nach Ergebnissen zu den angegebenen Phrasen zu suchen und die Anzahl der angegebenen Ergebnisseiten zu überwachen. Als Standardsuchmaschine wird Google verwendet. Die Suchergebnisse werden anschließend in der Tabelle wi_search_engine_result gespeichert.

```
php PhraseFinder.php crawl <number of phrases> <number of result pages>  
<search engine>
```

Beispiel Call: `php PhraseFinder.php crawl 5 2 google`

- Call, um die Ergebnisse eines Crawls innerhalb einer bestimmten Zeit zu analysieren. Neu gefundene Domains werden in der Tabelle wi_findings gespeichert.

```
php PhraseFinder.php store <start date (YYYY-MM-DD)> <end date (YYYY-MM-DD)>  
<search engine>
```

Beispiel Call: `php PhraseFinder.php crawl 5 2 google`

- Suchergebnisse können in den jeweiligen Tabellen überprüft und weiterbearbeitet werden.