# Bachelorarbeit von Viet Hoa Nguyen bei der Immowelt AG

Betreuer: Maxim Fridental, Axel Schwanke, Immowelt AG

## Titel der Arbeit

Entwicklung eines verteilten Cloud-basierten Systems zum Crawling von öffentlichen immobilien-relevanten Daten für ein großes deutsches Immobilienportal

## Aufgabenstellung

Bauvorhaben, Bebauungspläne, Umwidmungen von Flurstücken etc. können drastische Änderungen im Wert der betroffenen Immobilien und Flurstücken bedeuten. Die Veränderungen können im Voraus in Erfahrung gebracht werden, z.B. durch öffentliche Protokolle der Bauausschüsse oder Amtsblätter. Diese werden aber nicht von allen Eigentümern gelesen. Die entsprechenden Webseiten aller Städte, Gemeinden, anderer Behörden und Ämter sollen daher regelmäßig gecrawlt und neue/aktualisierte Immobilienwert-relevante Dokumente gespeichert werden. Wichtig ist der effiziente Einsatz von Crawlern, d.h. Webseiten möglichst nur dann zu crawlen, wenn auch neue/aktualisierte Dokumente zu erwarten sind.

Aufgaben sind die

* Konzeption einer verteilten Architektur auf Basis von Amazon Simple Queue Service und Amazon RDS zur Koordination und Steuerung der Web-Crawler-Worker.
* Recherche von Machine-Learning Modellen zur Steuerung und Optimierung der Web-Crawler Worker auf Basis historischer Daten bzgl. der Webseiten/Dokumenten Änderungen.

Aufstellen eines Kriterien-Katalogs; Bewertung der Modelle; Auswahl eines Modells.

* Implementierung des ausgewählten Modells zur Steuerung der Web-Crawler-Worker, inkl.

automatisiertem Deployment und Skalierung in der Amazon Cloud.

* Konzeption und Implementierung einer einfachen GUI zur Überwachung der verteilten Web-Crawler-Worker.
* Prototypischer Einsatz; Test der Funktionalität und Usability; Analyse und Auswertung des Tests, z.B. Abhängigkeit der Crawling-Frequenz von der Einwohnerzahl der jeweiligen Stadt/Gemeinde, Zeitdifferenz zwischen Dokument-Erstellung/Aktualisierung und Crawling-Zeitpunkt.

Die Implementierung der Funktionen und Services soll in C# oder Python erfolgen.

Ziel der Arbeit ist einerseits die Entwicklung eines ersten Prototyps eines verteilten Cloud-basierten Systems zum Crawlen von öffentlichen immobilien-relevanten Daten. Andererseits sollen die Grundlagen für ein Framework gelegt werden, dass sich auch zum effizienten verteilten crawlen anderer für den Immobilienbereich interessanter Information eignet.

<https://medium.com/@tonywangcn/how-to-build-docker-cluster-with-celery-and-rabbitmq-in-10-minutes-13fc74d21730>

https://benbernardblog.com/the-tale-of-creating-a-distributed-web-crawler/