Data Engineering Applied to the Swiss Job Market

Studiengang: BSc in Informatik | Vertiefung: Data Engineering
Betreuer: Prof. Dr. Erik Graf
Experte: Doktor Federico Flueckiger (Eidgenössisches Finanzdepartement EFD)

Online Jobportale bieten sehr viele Informationen über den aktuellen Arbeitsmarkt der Schweiz. Informationen wie Popularität des Berufes, sowie auch gewünschte Fähigkeiten, befinden sich in den Beschreibungen vieler Jobangebote. Das Ziel dieser Thesis war es diese Informationen aus dem Internet zu extrahieren, zu verarbeiten, auszuwerten, und zu visualisieren.

Einleitung

Das Ziel ist es Jobangebote aus dem Internet zu scrapen, daraus einen Datensatz zu erstellen. Diesen Datensatz mit zusätzlichen Informationen zu annotieren, damit es für Analysen und Visualisierungen verwendet werden kann.

Datensammlung

Die Datensammlung wurde mit Beautiful Soup4 bewerkstelligt. Mit einem Script wurden jeden Tag die neusten Berufe von www.indeed.ch aus allen Kantonen extrahiert und in eine Datenbank eingelesen. Beim Extrahieren der Daten wurde grosser Wert auf die Instandhaltung der Informationen gelegt und darauf das Meta-Daten für spätere Überprüfungen auch enthalten sind.

Datenverarbeitung

Bei der Datenverarbeitung wurde der Text von den HTML-Seiten extrahiert. Der Datensatz wurde mit der Sprache erweitert, indem ein Language Classifier die Beschreibung klassifiziert und annotiert hat. Manuelles Stemming wurde für Vereinheitlichung der geschlechtsspezifischen Berufe verwendet (Fachfrau / Fachmann ergibt Fachperson) und Stopword-Listen um Wörter mit wenig Aussagekraft zu entfernen.

Analyse

Die Analyse wurden mit normalen Regex Pattern-Searches und Spacy Rule-Based Matching bewerkstelligt. Mit diesen beiden Methoden wurden Informationen, wie der Schulabschluss, Qualifikationen und Sprachkenntnisse entnommen. Aus diesen gewonnenen Informationen wurden dann Erkenntnisse gezogen, wie der meistgesuchte EFZ Abschluss.

Visualisation

Die Informationen, welche aus der Analyse entstanden sind wurden mit der Library Matplotlib, Pyplot, Seaborn und Wordcloud dargestellt.

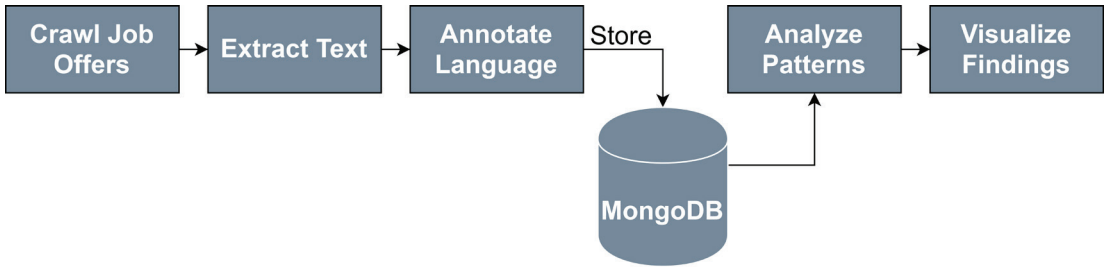
Resultat

Insgesamt wurden ~55'000 Stellenausschreibungen gesammelt, davon sind ~75% auf Deutsch, ~15% auf Französisch, ~8% auf Englisch, ~1% auf Italienisch und ~1% in anderen Sprachen ausgeschrieben. Alle Stellenausschreibungen sind in einem Datensatz verfügbar und können weitergeführt werden oder für eigene Analysen verwendet werden.

Ausblick

Da sich die Bachelorthesis nur in einem Zeitraum von einem Semester abspielt, konnten keine Erkenntnisse gezogen werden, welche die Veränderung über Zeit wieder spiegeln. Eine weitere Option die Arbeit fortzuführen, wäre es andere Jobportale zu scrapen um möglicherweise ein breiteres Spektrum abzufangen. Zusätzlich wurden längst nicht alle Untersuchungen bewerkstelligt und besonders geschlechtsspezifische Untersuchungen könnten sich als interessant herausstellen.

Dennis Gjokaj



Data Engineering Workflow