Social Network Analysis on StackOverflow

# Einleitung

StackOverflow ist Teil des StackExchange Netzwerks und bietet eine Plattform für Fragestellungen im Bereich Softwarentwicklung. Im Rahmen des Moduls «Social Network Analysis» der FHNW Technik führen wir eine Analyse von Fragen, Antworten und Tags von StackOverflow aus.

Wir entwickeln eine Anwendung zur automatischen Abfrage, Aufbereitung und Speicherung dieser Daten. Die Software bietet Exportfunktionen an, die es erlauben, die gesammelten Daten in geeigneter Form in SNA-Tools wie Gephi weiter zu analysieren.

Die zentralen Fragestellungen dieser Arbeit sind folgende:

* Was sind die populärsten Technologien?
* In welcher Community hat man die besten Antwortchancen (relativ zur Community Grösse)?
* Welche Technologien werden zu Arbeits- und welche zu Hobby-Zwecken verwendet?
* Welche Technologien sind am stärksten miteinander verbunden?
* Welche Technologien stehen eher allein?
* Was haben schnell beantwortete Fragen gemeinsam?

# Datenbeschaffung

Die benötigten Daten sind alle frei auf StackOverflow.com verfügbar. Wir können den StackExchange-weiten REST Endpoint verwenden, um die Daten im json Format zu scrapen – wir sind hier bezüglich der Datenmenge zwar auf 10’000 API Requests (300, wenn nicht registriert) pro Tag eingeschränkt, dies ist aber mehr als ausreichend für dieses Projekt. Pro Request bekommen wir bis zu 100 Datensätze.

## API Einschränkungen

Eine Einschränkung ist, dass wir die Throttle Violation beachten. Wir dürfen nicht mehr als 300 (10'000 wenn registriert) tägliche und 30 sekündliche API Requests absenden. Zusätzlich muss der «backoff» Parameter respektiert werden, welcher bei einer Response mitgeliefert werden kann. All dies wurde bei der Implementierung unseres Scrapers beachtet und stellt keinerlei Probleme dar. Wir erhalten pro Request bis zu 100 Datensätze, können also täglich bis zu einer Million Datensätze pro IP scrapen. Wir aggregieren die Datensätze zusätzlich in einer MongoDB und werden so definitiv genug Daten zur Verfügung haben.

Zusätzliche Infos hier: <https://api.stackexchange.com/docs/throttle>

## Softwarearchitektur & Datenspeicherung

Die verwendeten Technologien beschränken sich auf Kotlin und Spring Boot. Mit einem eigenen Websraper fragen wir die Daten im json Format vom REST Endpoint ab, mappen sie zu Mongo Documents und speichern sie in einer MongoDB zwischen. Von dort werden die Daten dann transformiert und zu einem Graph in CSV Format umgewandelt, der dann in Gephi (oder auch weiteren Tools) importiert werden kann.

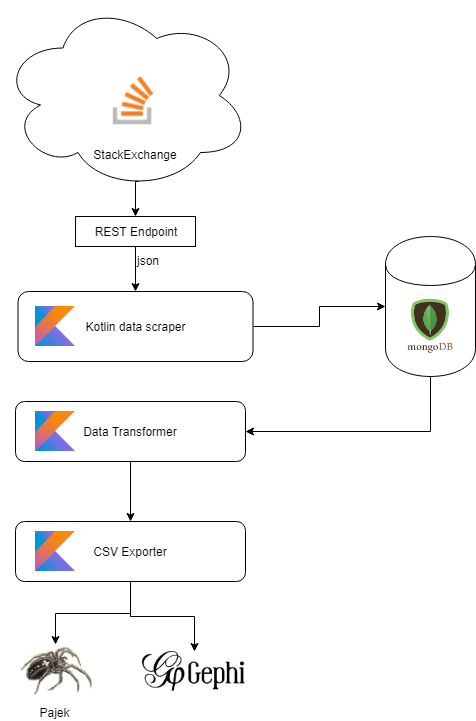


Abbildung : Software Architektur

## How to & Setup

Unsere Software ist auf Github verfügbar: <https://github.com/patricsteiner/sna_soscraper>

Am einfachsten klont man das Projekt und importiert es in IntelliJ. Von dort wird gradle automatisch alle nötigen Schritte vornehmen, um die Software zum Laufen zu bringen. Im Readme findet man zusätzliche Informationen: <https://github.com/patricsteiner/sna_soscraper/blob/master/README.md>

Wir haben während mehreren Tagen Daten gesammelt und diese in einer MongoDB (gratis hosting auf mlab.com) gespeichert. In der Version, die wir zu diesem Dokument mitliefern, sind die nötigen Credentials bereits konfiguriert, die Applikation kann also as-is gestartet und verwendet werden. Man kann aber auch eine eigene MongoDB verwenden, um Daten zu speichern. Dazu muss einfach die Datei appliaction.properties entsprechen angepasst werden.

Wenn man nur eine kleinere Menge an Daten abfragen und gleich weiterverwenden möchte ist es einfacher, die Daten einfach temporär im Memory zu halten, anstatt zu persistieren. Wir bieten dazu entsprechende Repository Klassen an (InMemoryQuestionRepository und InMemoryTagRepository). Um diese zu verwenden, muss einfach die @Repository Annotation dort gesetzt und beim MongoRepository entfernt werden.

# Datenanalyse

Im Rahmen dieser Arbeit verwenden wir insgesamt 77587 Datensätze, die wir während mehreren Tagen gescraped haben. Im Schnitt hat jede Frage 2.92 Tags.

## Rohdaten

Wie die Rohdaten im Detail aussehen, kann der Klasse Question entnommen werden: <https://github.com/patricsteiner/sna_soscraper/blob/master/src/main/kotlin/ch/fhnw/sna/soscraper/domain/Question.kt>

Die für uns wichtigsten Attribute zu jeder Frage sind nachfolgend aufgelistet:

* Id
* Label
* Titel
* Erstellungsdatum
* Anzahl Antworten
* Hat akzeptierte Antwort
* Score
* Anzahl Views
* Tags

Die Daten kommen garantiert immer genau gleich an und es sind somit keine Bereinigungsschritte zwingend.

## Analyse der Rohdaten

Eine erste Analyse der relevanten Attribute der Rohdaten liefert folgende Ergebnisse:

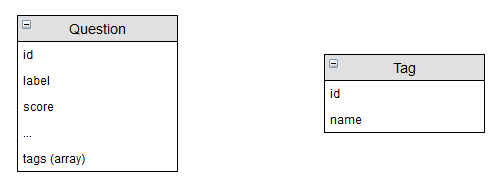
*Tabelle 1: Analyse der Rohdaten*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribut** | **Min** | **Max** | **Average** | **Kommentar** |
| CreationDate | 2018-10-09 06:18:44 | 2018-10-21 15:45:30 | - | Täglich wurden also rund 6500 Fragen gestellt. |
| Views | 2 | 11’233 | 39.87 | Einige wenige Ausreisser nach oben («viral Questions») |
| Answers | 0 | 12 | 0.9 | - |
| Comments | 0 | 31 | 2.0 | - |
| Favorites | 0 | 12 | 0.08 | Interessant zu untersuchen, welche Fragen oft favorisiert werden. |
| Score | -11 | 77 | 0.13 | Das einzige Attribut, das negativ sein kann. |
| Tags | 1 | 5 | 2.92 | Es muss mindestens ein und höchstens 5 Tags gesetzt werden. |

## Transformation

Aus den gesammelten Daten kreieren wir zwei verschiedene Sammlungen von Entitäten. Einerseits «Question» und andererseits «Tag», wobei eine n-n Verknüpfung von Question zu Tag besteht. Es ist nötig, dass wir Tags (die einfach als String-Array einer Question geliefert werden) zu einer Entität mit eigener ID umwandeln, da das Tool Gephi eine id voraussetzt.

Wir erhalten also zwei Repositories mit folgenden Entitäten:



Aus diesen zwei Repositories können wir nun ein 2-Mode Netzwerk generieren. Jede Question und jedes Tag werden jeweils zu einem Knoten mit dem dazugehörigen Typ. Für jedes Tag einer Question wird eine Kante zwischen den entsprechenden Knoten generiert.

Für unsere Arbeit ist es jedoch nutzbringender, dieses 2-Mode Netzwerk (Question zu Tag) in ein 1-Mode Netzwerk (Tag zu Tag) zu transformieren. So können wir bei der Transformation noch nachfolgend beschriebene zusätzliche Informationen gewinnen. Diese Attribute ermöglichen uns erweiterte Analysen und werden uns helfen, die Kernfragestellungen dieser Arbeit zu beantworten.

* **Occurence**: Anzahl Vorkommen in allen Fragen
* **Views**: Summe der Views aller Fragen, die dieses Tag beinhalten
* **Answered**: Summe der beantworteten Fragen, die dieses Tag beinhalten
* **Unanswered**: Summe der unbeantworteten Fragen, die dieses Tag beinhalten
* **Bounty**: Summe der Bounties aller Fragen, die dieses Tag beinhalten
* **AnsweredRatio**: Verhältnis von beantworteten zu unbeantworteten Fragen, die dieses Tag beinhalten

Bei Ursprungsnetzwerk handelt es sich um einen gerichteten Graphen (Question zu Tag). Bei der Transformation behalten wir die Kantenrichtungen bei, operieren also weiterhin auf einem gerichteten Tag zu Tag Graphen. Man könnte die Kantenrichtungen auch ignorieren, würde dadurch aber potentiell wertvolle Informationen verlieren. Darum tun wir das nur für gewisse Visualisierungen, bei denen die Kantenrichtung irrelevant ist.

# Soziale Netzwerkanalyse

Zur Analyse wurde der generierte gerichtete Graph in Gephi importiert. Alle Mehrfachkanten wurden zusammengefügt und bekamen die Summe der Anzahl zusammengeführter Kanten als Gewicht.

## Kennzahlen zum Netzwerk

Die nachfolgenden Werte wurden direkt mit Gephi berechnet.

*Tabelle 2: Kennzahlen zum Netzwerk*

|  |  |  |
| --- | --- | --- |
| **Kennzahl** | **Wert** | **Kommentar** |
| Anzahl Knoten | 16’777 | Erstaunlich, dass es so viele Tags gibt! |
| Anzahl Kanten | 239’908 | Waren ursprünglich viel mehr, doppelte Kanten wurden zusammengefügt. |
| Average Degree (Kantenrichtung ignoriert) | 3.484 | Jedes Tag ist im Schnitt mit so vielen anderen Tags direkt verbunden. |
| Average Weighted Degree (Kantenrichtung ignoriert) | 37.299 | Diese Zahl sagt nicht viel aus, da die Streuung enorm gross ist (s. Abbildung 1) |
| Graph Density | 0 | Die Dichte ist sicher nicht genau 0, jedoch so klein, dass sie von Gephi als 0 ausgegeben wird. |

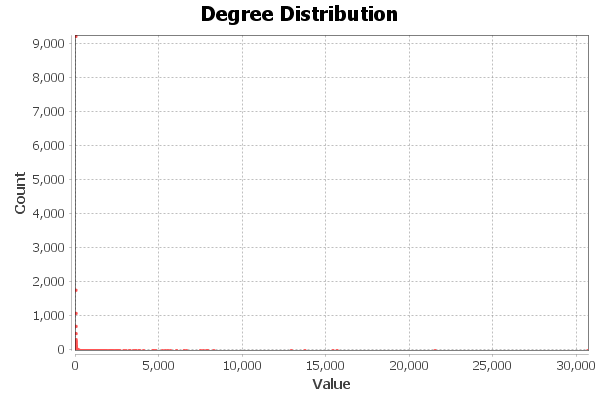


Abbildung 2: Verteilung der gewichteten Knotengrade

Weitere Metriken können in diesem riesigen Netzwerk nur mit enorm hohen Rechenaufwand berechnet werden, darum werden wir in den weiteren Analysen nur bestimmte Teilnetzwerke genauer untersuchen.

## Kennzahlen zu Knotenattributen

Diese Kennzahlen werden innerhalb unserer Webanwendung berechnet. Die Berechnung dauert mehrere Sekunden, da sehr viele Daten ins RAM geladen werden müssen und kann unter dem Rest Endpoint <http://localhost:8080/stats> abgerufen werden.

Tabelle : Kennzahlen zu Knotenattributen

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Attribut** | **Min** | **Max** | **Average** | **Kommentar** |
| Occurence | 1 | 8471 | 13.52 | - |
| Views | 3 | 345844 | 554.97 | - |
| Answered | 0 | 4290 | 5.84 | - |
| Bounty | 0 | 500 | 0.3 | Wird nur bei sehr wenigen Fragen gesetzt. Siehe <https://stackoverflow.com/help/bounty> für weitere Infos. |

## Die populärsten Tags

Die erste und interessanteste Frage ist sicher: Was sind die beliebtesten/häufigsten Tags? Gefiltert nach den 10 meistvorkommenden Tags erhält man folgendes Bild:

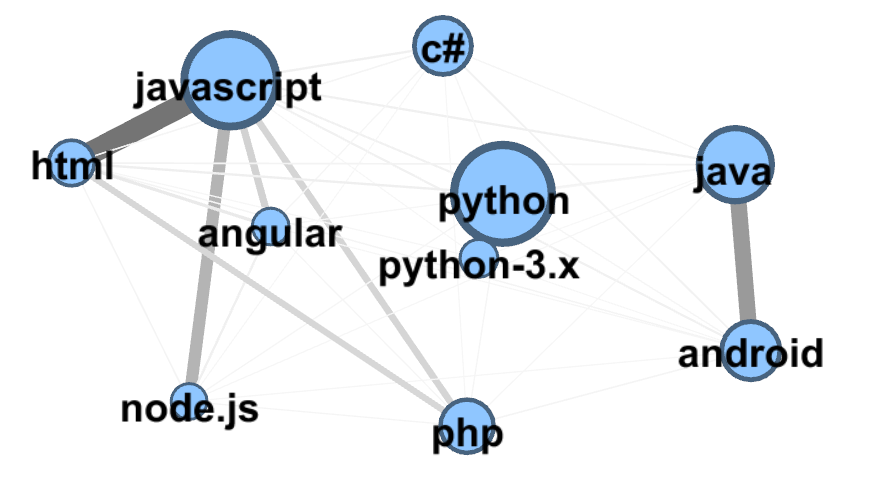


Abbildung : Top 10 Tags

Wenn man dieses Resultat mit der Tag-Statistik von der gesamten StackOverflow Database (siehe <https://stackoverflow.com/tags>, Stand 22.12.2018) vergleicht, stellt man schnell fest, welche Technologien heutzutage wovon abgelöst wurden:

Tabelle : All-time top 10 Tags auf StackOverflow

|  |  |  |
| --- | --- | --- |
| **Rang** | **Tag** | **Popularität** |
| 1 | JavaScript | 1494422 |
| 2 | Java | 1494422 |
| 3 | C# | 1270228 |
| 4 | PHP | 1249791 |
| 5 | Android | 1159674 |
| 6 | Python | 1080161 |
| 7 | jQuery | 938323 |
| 8 | HTML | 793431 |
| 9 | C++ | 598303 |
| 10 | iOS | 585958 |

StackOverflow gibt es seit 10 Jahren, wobei sich die Popularität der Tags natürlich im Verlaufe dieser Zeit stark verändert hat. Wir analysieren nur die Daten der aktuellsten 2 Wochen und ziehen daraus folgende Schlüsse:

* **JavaScript** ist nach wie vor der Leader
* **Python** scheint heutzutage viel beliebter als auch schon zu sein, was wohl mit dem Data Science Hype zu tun hat
* Mit dem aufkommen von Single Page Applications hat **Angular** weitgehend **jQuery** abgelöst und auch **node.js** in den Vordergrund gerückt
* **PHP** ist nach wie vor relevant, aber definitiv am nicht mehr so sehr wie auch schon

Betrachtet man zusätzlich noch die 40 nächst-meistvorkommenden Tags, ist gut erkennbar, wie sich die heute prominenten Libraries und Frameworks um die dazugehörigen Technologien anordnen:

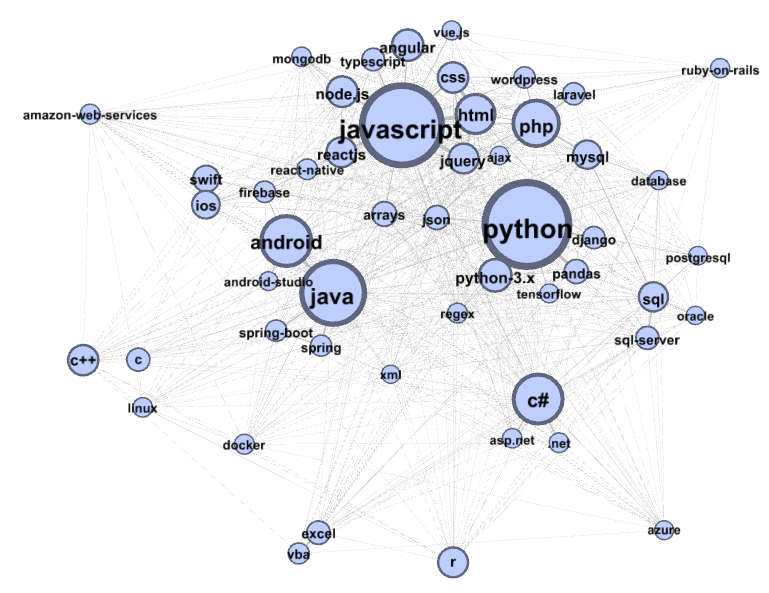


Abbildung : Top 50 Tags

## Korrelationen zwischen Popularität und Knotenattributen

In diesem Abschnitt wird untersucht, wie die Popularität mit verschiedenen Knotenattributen korreliert. Um die Visualisierungen zu vereinfachen verwenden wir hier die 3000 meist vorkommenden Tags. Als erstes wollen wir herausfinden, wie stark die Anzahl views mit der Vorkommnis eines Tags korreliert. Es bestätigt sich hier die Erwartung eines sehr startken Zusammenhangs dieser beiden Attribute, wie nachfolgendem Chart entnommen werden kann. Der Pearson Korrelationskoeffizient ist 0.99.

Viel interessanter ist aber die Frage, ob die Chance, dass eine Frage beantwortet, positiv korreliert mit der Popularität eines Tags. Wenn wir dazu wieder die 3000 häufigsten Tags verwenden, ist das zuerst überhaupt nicht ersichtlich:

Wenn die 200 höchsten Ausreisser entfernt werden, ist das Resultat immer noch nicht sehr klar. Es sieht leicht danach aus, als ob der AnsweredRatio mit erhöhter Occurence sogar sinkt. Der Person Korrelationskoeffizient von 0.05 ist sehr gering und bestätigt diese Vermutung nur teilweise. Wir erachten dies als keine relevante Korrelation, da es plausibel ist, dass mit höherer Popularität eines Tags auch mehr Unbeantwortete Fragen existieren, zum Beispiel auf Grund von Verlinkung auf Duplikate.

### Bounties

Die Idee von Bounties ist, dass man einen Teil seiner StackOverflow Reputation als Prämie (Bounty) offeriert für den, der die Frage beantwortet. Reputation ist keine leicht verdiente Sache auf StackOverflow, und eine Bounty der Grössenordnung 100 ist sehr viel mehr Wert als zum Beispiel eine Akzeptierte Antwort auf eine Frage. Wir wollen herausfinden, ob sich die ausgesetzte Prämie tatsächlich für den Fragesteller lohnt und vergleichen den durchschnittlichen answeredRatio von Fragen mit Bounties mit dem von Fragen ohne Bounties.

# Ausblick

Was wäre Ihrer Meinung nach noch interessant zu analysieren, wenn das Projekt weitergeführt würde? Gäbe es noch weitere interessante Fragestellungen / weitere interes-sante Attribute, die eingesammelt oder sauber verarbeitet werden müssten? War die Daten-quelle passend oder sollte eine andere in Betracht gezogen werden? Falls ja, haben Sie eine gefunden?