Kanton Zürich
Direktion der Justiz und des Innern
Statistisches Amt

Kreative Challenge für den NLP-Hackathon an der Universität Bern 24./25. März 2021

Ausgangslage

Es ist nicht immer leicht in möglichst kurzer Zeit an benötigte statistische Informationen zu gelangen. Öffentlich zugängliche Daten und Informationen sind vertikal auf verschiedenen föderalen Ebenen, innerhalb dieser Ebenen horizontal auf verschiedenen Ämtern und innerhalb der Ämter auf verschiedenen Seiten/Kanälen verteilt. Bei diesen Gegebenheiten scheint selbst die berüchtigte Nadel im Heuhaufen leicht auffindbar.

Nur schon die rein statistischen Informationen sind mit Hilfe von Suchmaschinen nicht immer leicht auffindbar, da nicht alle vorhandenen Ressourcen indiziert werden.

Das erschwerte Auffinden von Fakten, Daten und auch methodologischen Erklärungen ist auch ein Risiko für demokratische Prozesse: Je schwerer es für die/den Durchschnittsbürger/in ist, die Informationen zu finden, desto leichter ist es falsche Fakten zu verbreiten.

Vor diesem Hintergrund möchte das Statistische Amt Kanton Zürich zusammen mit anderen Organisationen einen Schweizerischen statistischen Bot (STATBOT.CH) entwickeln, welcher über alle Organisationen hinweg Daten und statistische Auskünfte direkt und schnell liefert. Ein Projektantrag zwischen KORSTAT und dem BFS läuft, und auch im Kanton Zürich laufen Abklärungen.

Das Ziel der Challenge mit «kreativem Charakter»

Nach unserem Verständnis gibt es zwei grosse Komponenten eines solchen statistischen Bots. Einerseits das Verständnis der Frage und andererseits die Datenabfrage. In dieser Challenge möchten wir uns auf die erste Komponente fokussieren.

Primäres Ziel: Einen Weg finden, um die wichtigsten Informationen aus einer Frage zu extrahieren, um dann damit weiterarbeiten zu können.

Sekundäres Ziel: Ein brauchbares Modell für die Wissensextraktion zu trainieren.

Lösungsansätze

Die Teilnehmerinnen und Teilnehmer sind absolut frei in der Wahl ihres Lösungsansatzes. Im Folgenden stellen wir den Lösungsansatz vor, den wir momentan am geeignetsten halten Aber vielleicht seht ihr eine viel bessere Lösung? Ihr seid da völlig frei. Wir sind gespannt auf eure Rückmeldungen und Ideen!

Den jeweils aktuellsten Stand unserer folgenden Überlegungen findet ihr in der Dokumentation unseres github-repos. Dort findet ihr auch noch zwei weitere mögliche Lösungsansätze, die uns durch den Kopf gegangen sind.

Möglicher Lösungsansatz A: NER trainieren

Mittels Named Entity Recognition (NER) können in einem Text relevante Entitäten identifiziert werden. Im Englischen gibt es sehr fortgeschrittene NER, die viele für uns relevante Entitäten erkennen (cardinal, ordinal, time, money usw.).

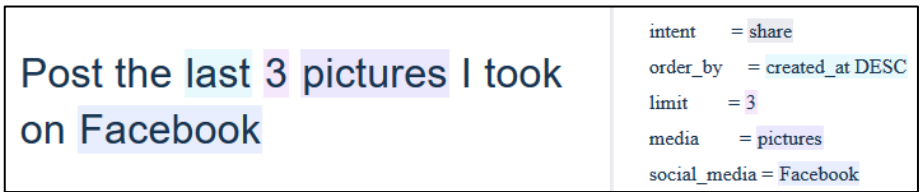

Abbildung: Beispielentitäten aus wit.ai wie order_by und limit

Allerdings gibt es bisher für die deutsche Sprache nur NER mit 4 Entitäten: Lokalitäten, Organisationen, Personen und «Misc». Dabei hat einzig die Erkennung von Lokalitäten wie beispielsweise Zürich, Wil usw. einen Nutzen für uns. Wobei einige der für uns relevanten Lokalitäten (wie z.B. gewisse Gemeinden im Kanton Zürich) nur unzureichend identifiziert werden. Vielleicht könnten auch die Part of Speech (POS) Elemente von Nutzen sein – ein erster Blick auf das Stuttgart-Tübingen-Tagset (STTS) hat uns einige interessante Elemente gezeigt, aber wir haben dies noch nicht detailliert angeschaut.

Um NER nutzen zu können, müssen wir eigene Entitäten erstellen. Beispiele für solche Entitäten sind Datensatznamen, Variablennahmen, die Granularität (z.B. Bevölkerung auf Bundes, Kantons, Bezirks oder Gemeindeebene?), und selbst wichtige Informationen im Text wie «grösser als», «pro Monat», «für alle Gemeinden im Kanton Zürich» oder «von 2005 bis 2008» (wie man es z.B. auch oben sieht bei der Abbildung von wit.ai). Auf Github befinden sich die Datengrundlagen und der Code für unsere ersten Versuche, die als Inspiration verwendet werden dürfen (siehe Bild).

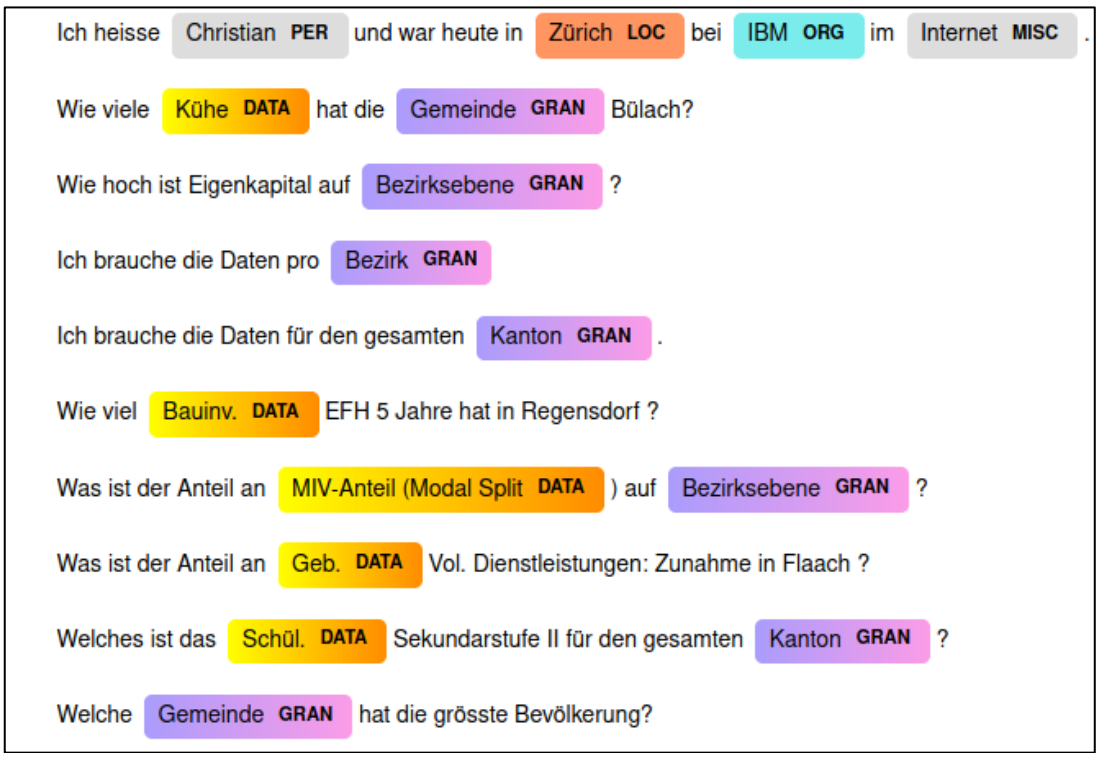

Abbildung: Beispiel von zwei zusätzlich trainierten NER zu Granularität und Datensets

Der Code um eigene NERs zu trainieren stammt grösstenteils von hier:
https://deepnote.com/publish/2cc2d19c-c3ac-4321-8853-0bcf2ef565b3

Andere Quellen mit vortrainierten NER auf Deutsch findet man z.B. hier:
https://huggingface.co/transformers/v2.2.0/examples.html#named-entity-recognition
https://sites.google.com/site/germeval2014ner/data

Daten

Alle öffentlich verfügbaren Daten und Metadaten vom Statistischen Amt des Kantons Zürichs sind hier als JSON abrufbar: https://www.web.statistik.zh.ch/data/zhweb.json

Aber die Daten können natürlich auch erweitert werden: Alle Daten unter opendata.swiss sind relevant (primär Daten von Statistikämtern).

Das github-repo befindet sich hier:
https://github.com/statistikZH/statbot

Dort befindet sich im Unterverzeichnis Dokumentation auch die aktuellste Version unserer Überlegungen und der Script-Beschreibungen.

Teilnehmende seitens Statistisches Amt Kanton Zürich

Christian Ruiz
christian.ruiz@statistik.ji.zh.ch
076 / 448 75 00

Corinna Grobe

Manuela Paganini

Michelle Donzallaz

Thomas Lo Russo