SWISS INSIGHTS

Swiss Data Insights Association

Swiss Insights News

#12

Projekt statbot.swiss: Mehr als nur ein statistischer Bot

Projekt statbot.swiss: Mehr als nur ein statistischer Bot

Dr. Christian Ruiz
Projektleiter Machine Learning
Statistisches Amt Kanton Zürich

In welcher Schweizer Gemeinde ist der Anteil der Elektroautos am höchsten? Und welche Gemeinde hat die meisten Dieselfahrzeuge? Das Projekt statbot.swiss hat zum Ziel, die Machbarkeit eines Bots zu prüfen, der statistische Fragen eigenständig mittels KI beantworten kann.

Eine Bürgerin sucht nach Daten der öffentlichen Statistik in der Schweiz: Die Wahrscheinlichkeit ist hoch, dass damit eine digitale Odyssee beginnt, statt, dass die Suche nach einer kurzen Suchmaschinenanfrage endet. Denn selbst für erfahrene Recherchierende ist es nicht immer leicht, Daten zu finden: Befinden sich die gesuchten Daten vertikal auf Ebene Bund, Kantone oder Gemeinden? Und welche Behörde oder welches Amt stellt innerhalb einer Ebene die Daten bereit? Die Suchmaschinen haben Mühe, die gesuchten Informationen zu indizieren, da sie sich in verschiedenen Applikationen oder sogar Dokumenten befinden. So klicken sich die Bürger durch, bis sie die gewünschten Informationen erhalten – oder bis sie die zuständige Stelle gefunden haben, bei der sie ihre Anfrage platzieren können.

Machine learning to the rescue?

Können wir nicht einen Algorithmus trainieren, uns die richtigen Antworten auf unsere Fragen zu liefern? Ist es nicht genau das, was Google und Co. machen und nicht immer schaffen? Unsere Schlussfolgerung ist: Wenn die Daten nicht «richtig aufbereitet» sind, dann fehlen die Grundlagen für die Anwendung eines fortgeschrittenen Algorithmus.

Die benötigten Daten der Schweizer Behörden stehen qualitativ hochwertig zur Verfügung. Trotzdem ist es nahezu unmöglich, einen Algorithmus im Umgang mit diesen Daten so zu trainieren, dass er die Frage «was ist der Anteil der registrierten Elektroautos im Kanton Zürich» korrekt beantwortet.

Das seit längerer Zeit bekannte Problem von maschinell nicht lesbaren Datenformaten wie Excel oder PDF spielt dabei nur eine kleine Rolle. Vor allem fehlt es in der Schweiz an einheitlichen Ansätzen für Datenstrukturen; nicht nur zwischen den verschiedenen Datenproduzenten, sondern sogar innerhalb deren jeweiligen Datenportalen. Die Probleme sind:

- Qualität (wahr und nicht suggestiv)
- Ohne menschliches Zutun sind zeitliche Elemente, räumliche Elemente, Dimensionen und Attribute nicht identifizierbar.
- Die Daten sind uneinheitlich sowohl im «Long-Format» wie auch im «Wide-Format» – zusätzlich findet man manchmal sogar Kreuztabellen oder gar Listen.
- Es fehlt eine Standardisierung der verwendeten Codes und Labels.
- Die Definitionen der relevanten Daten sind nicht in jedem Fall harmonisiert – beispielsweise errechnen Bund, Kantone und Gemeinden die Bevölkerungszahlen der Stadt Zürich leicht unterschiedlich.
- Spalten gleicher Datenkategorien sind unterschiedlich benannt

Als Konsequenz dieser unterschiedlichen Datenstrukturen ist der Bau jeglicher datensatzübergreifenden Applikation mit beträchtlichem Aufwand verbunden – unabhängig davon, ob es sich um eine High End Machine-Learning-Lösung oder um ein simples Indikatorenportal handelt. Selbst wenn nur die Daten aus einem einzigen Datenportal eingelesen werden, kann

die Menge der eingelesenen Datasets nicht beliebig vergrössert werden, da man für jedes Dataset einen separaten Code schreiben muss.

Standards and data harmonization to the rescue!

Hier setzen wir mit dem Projekt statbot.swiss an. Finanziert wird es durch eGov Schweiz, geleitet vom Bundesamt für Statistik und der Konferenz der regionalen statistischen Ämter der Schweiz. Technologische Partner sind die Zürcher Hochschule für angewandte Wissenschaften (ZHAW) sowie das Swiss Data Science Center (SDSC).

Das Projekt besteht aus zwei Teilen: Im chronologisch zweiten Teil experimentieren Forschende der Zürcher Hochschule für angewandte Wissenschaften an einem Machine-Learning-Algorithmus, der natürliche Sprache in Datenbankabfragen übersetzen kann.[1] Dieser soll die Antworten in Form von Datenwerten, Tabellen oder Grafiken anzeigen. Das aktuelle Projekt setzt dabei den Fokus absichtlich und bescheiden auf die Abklärung der Machbarkeit und der technischen Limitierungen.

Im bereits abgeschlossenen ersten Teil des Projekts haben wir ein Open Source Data Warehouse[2] gebaut, das Daten aus unterschiedlichen Quellen an einem Ort integriert und harmonisiert. Es lädt die Daten aus unterschiedlichen Schnittstellen, führt die gesamte Transformation («ETL») aus und erstellt alle Daten in einer einheitlichen, definierten Datenstruktur mit dazugehörenden Mapping-Tabellen für Dimensionen. Ziel war auch, die Datenstandardisierung und -harmonisierung voranzubringen.

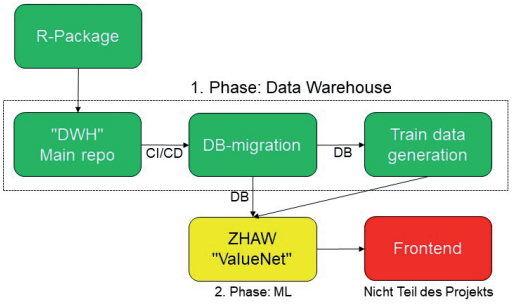

Abbildung 1: Übersicht über die verschiedenen open source Repositories. Die Machine-Learning-Lösung der ZHAW wird dann sowohl auf die Datenbank, wie auch auf die generierten Trainingsdaten zugreifen. Ein sehr einfaches Frontend soll erstellt werden.

Das Hinzufügen der Daten ist dabei skalierbar: wir konnten Stand heute über 240 «Statbot-Datasets» mit Inputdaten sowohl aus dem Bundesamt für Statistik wie auch aus sechs weiteren regionalen Statistikämtern hinzufügen. Wir haben dabei einen Prozess geschaffen, der keine Programmierung durch Mitarbeitende von teilnehmenden Statistikämtern benötigt. Stattdessen brauchen diese nur das sprichwörtliche Rezept anzugeben, wie die Daten verarbeitet werden sollen: Für einen Milchshake müssen die Früchte gewaschen, geschält, dann in den Mixer gegeben und am Ende mit Milch gemischt werden. Die Reihenfolge dieser Schritte gibt der Mitarbeitende ein einziges Mal ein, dann kann die Maschine regelmässig die aktuellsten Inputdaten holen und die notwendigen Transformationen durchführen.

Wichtig ist dabei auch, dass dieser Endzustand aller Datasets gleich ist, wobei wir hier eine Lösung gefunden haben, die sich an bisherigen Ansätzen orientiert[3]. Als zusätzlicher Schritt müssen daher Dimensionen «harmonisiert» werden. Jede Dimension wie Geschlecht oder Staatsangehörigkeit hat dabei gewisse Codes. Wenn die Standardcodes (üblicherweise aus dem Bundesamt für Statistik) verwendet werden, läuft das Matching ohne grösseren Zusatzaufwand. Ansonsten können neue Mappings hinzugefügt werden, um eine flexible Zuordnung herzustellen.

Das Resultat sind relationale Daten in einer definierten Struktur mit einheitlichen Spaltenbezeichnungen, Zuordnung von Raum, Zeit, Dimensionen sowie Attributen und klar bestimmten Kodierungen[4].

Diese «Statbot-Datasets» sind somit einheitlich strukturiert. Damit ist es leichter, daraus datensatzübergreifende Applikationen zu bauen, beispielsweise eine Machine-Learning-Lösung wie im zweiten Teil des Projekts oder ganz andere Applikationen, die nun auf solche bereinigten Daten zugreifen könnten.

1 Die ZHAW leitet das Projekt «INODE – Intelligent Open Data Exploration», das vom EU-Forschungsprogramm Horizon 2020 mit knapp 6 Millionen Euro gefördert wird.).

2 https://renkulab.io/gitlab/christian.ruiz/statbot-swiss

3 Es orientiert sich sowohl an der Linked Data Lösung von Statistik Stadt Zürich wie auch am SDMX-Format, welches u.a. von Eurostat verwendet wird. Die Überlegungen gehen dabei noch weiter, ob sogar diese Formate generiert werden könnten.

4 Momentan fehlt eine Vorgabe für eine «definierte Struktur» von Statistikdaten in Schweizer Statistikämtern. Diese funktionale Vorgangsweise könnte auch als Bottom-up-Lösung dienen, um einer Harmonisierung schrittweise näher zu kommen.

spatialunit_ontology	spatialunit_hist_id	time_value	period_value	value	gender_of_child	citizenship_category_of_child	age_class_of_mother
CH	0	1985-12-31	NA	26939	-111	1	2
CH	0	1985-12-31	NA	17809	-111	1	3
CH	0	1985-12-31	NA	5424	-111	1	4
CH	0	1985-12-31	NA	846	-111	1	5
CH	0	1985-12-31	NA	11675	-111	2	-111
CH	0	1985-12-31	NA	3768	-111	2	1
CH	0	1985-12-31	NA	4065	-111	2	2
CH	0	1985-12-31	NA	2614	-111	2	3
CH	0	1985-12-31	NA	1036	-111	2	4
CH	0	1985-12-31	NA	192	-111	2	5
CH	0	1985-12-31	NA	37965	1	-111	-111
CH	0	1985-12-31	NA	7918	1	-111	1
CH	0	1985-12-31	NA	15908	1	-111	2
CH	0	1985-12-31	NA	10315	1	-111	3
CH	0	1985-12-31	NA	3317	1	-111	4
CH	0	1985-12-31	NA	507	1	-111	5
CH	0	1985-12-31	NA	32023	1	1	-111

Abbildung 2: Ein Beispiel der definierten Datenstruktur eines «Statbot-Datasets». Die ersten beiden Spalten definieren eindeutig die Raumeinheiten. Die nächsten beiden Spalten definieren Zeitelemente (Stichtage sowie Perioden). Danach kommt der Beobachtungswert des Datenpunkts. Alles was rechts der eingezeichneten Linie ist, sind Dimensionen, die von Dataset zu Dataset unterschiedlich sind aber die in Dimensionstabellen definiert sind. Im Bild nicht sichtbar sind angedachte Attributsspalten: Quellenangaben, Definitionsangaben und «Flags» mit weiteren Informationen pro Datenpunkt.

Weiteres Potenzial

Das Projekt lehrt uns viel über Machine Learning. In Machine-Learning-Projekten geht es um weit mehr als den Gebrauch der Algorithmen. Viel Arbeit muss in Grundlagenarbeit gesteckt werden, um die Daten für Algorithmen überhaupt brauchbar zu machen. Dazu gehört auch die Bereitstellung von «Trainingsdaten», also Paare von Fragen in natürlicher Sprache und passenden Antworten in Form von Datenbankabfragen. Die werden dann verwendet, um den Algorithmus zu trainieren. Hier haben wir mit zusätzlichen Sprachinformationen gearbeitet, um sinnvollere Sätze zu generieren, die aber dennoch etwas hölzern erscheinen.

Die klare Trennung von Inhalt und Bezeichnung und die einheitliche Strukturierung und Harmonisierung der Bezeichnungen im Data Warehouse werden es darüber hinaus erlauben, automatisiert hochwertige «Linked Open Data» zu generieren. Linked Open Data sind Daten, die schnittstellenartig per HTTP abgerufen werden können, und die auf weitere Ressourcen verweisen können.

Dies würde vernetzte Datenlösungen ermöglichen, welche Daten aus den verschiedensten Quellen kombinieren könnten und es würde ein semantischer Datenraum geschaffen werden. Das wäre ein zusätzlicher Mehrwert für Nutzende, Firmen sowie das gesamte Statistiksystem der Schweiz, da es weitere maschinelle Anwendungen ermöglichen würde.

Ein Projekt-Teilnehmer hat kürzlich Folgendes treffend formuliert: «Auch nach dem offiziellen Ende des Projekts statbot.swiss glaube ich, dass das Data Warehouse bestehen bleiben wird… Es ist womöglich das erste Mal in der Schweiz, dass öffentliche Daten von verschiedenen Datenproduzenten an einem Ort in einer gemeinsamen, harmonisierten Struktur vorhanden sind.». Die Datenbedürfnisse der potentiellen Machine-Learning-Anwendung haben somit als wesentliche Anforderung dazu geführt, Daten zu strukturieren und zu harmonisieren. Dieses Resultat – das Data Warehouse – könnte viel Potenzial für andere Anwendungen haben.

Ich erinnere an dieser Stelle zunächst an die Bürgerin, die hoffentlich in der Zwischenzeit die gesuchten Daten gefunden hat. Jede Entwicklung, egal ob ein Data Warehouse, ein Bot oder eine andere Anwendung, muss direkt oder indirekt auf die Bedürfnisse der Bürger ausgerichtet sein. Öffentliche Behördendaten sollten einfacher auffindbar, vernetzter und leichter in andere Applikationen integrierbar werden.

Hier möchten wir mit der Community durch persönliche Gespräche, Events und Online stärker ins Gespräch kommen. Einerseits möchten wir die Datenbedürfnisse der Nutzenden besser kennen, um darauf bedarfsorientiert entwickeln zu können. Andererseits können wir damit leichter Lösungen wie den Bot in einer frühen Phase testen und Feedback integrieren. Wenn Sie Interesse haben, können Sie mir gerne formlos eine E-Mail mit Betreff statbot.swiss schreiben.

Der Autor

Dr. Christian Ruiz leitet aktuell das Machine Learning im Statistischen Amt Kanton Zürich. Er hat 13 Jahre Erfahrung in Data Science und 6 Jahre Erfahrung in der praktischen Entwicklung von Machine-Learning-Anwendungen. Er ist Mitgründer von zwei Startups im Bereich Deep Learning, hat drei weitere Machine-Learning-Projekte umgesetzt und andere Projekte begleitet.

Kontakt

Christian Ruiz, +41 43 259 75 12
christian.ruiz@statistik.ji.zh.ch

Swiss Insights

Swiss Insights ist der Verband und die Interessensvertretung aller Unternehmen, die Daten und prädiktive Modelle im Rahmen von Marketing, Innovationsprozessen, Kundenservice, Angebotsgestaltung, Kommunikation und Zielgruppendefinitionen erheben, analysieren, einsetzen und daraus Handlungsempfehlungen ableiten.

Swiss Insights pflegt einen aktiven Dialog mit politisch und gesellschaftlich wichtigen Akteuren und fördert den Austausch mit anderen nationalen und internationalen Fachorganisationen.

Eine der Hauptaufgaben des Verbands ist die Förderung der Markt-, Meinungs- und Sozialforschung im Allgemeinen und der Wissenschaftlichkeit im Besonderen. Er entwickelt, definiert und unterhält strenge Leitlinien zur Qualitätssicherung und grenzt sich im Bereich der Markt- und Sozialforschung klar von Werbung und Direktmarketing ab. Hierzu führt der Verband das Qualitätslabel «Market & Social Research by Swiss Insights».

Darüber hinaus engagiert sich Swiss Insights dafür, dass die Nutzung von Daten und die Anwendung von datengetriebenen Modellen transparent, nachvollziehbar und in diesem Sinne fair gestaltet wird. Hierfür wurde das Label «Data Fairness by Swiss Insights» geschaffen.

Market and Social Research by SWISS INSIGHTS

DATA FAIRNESS BY SWISS INSIGHTS

Herausgeber und Kontakt
Swiss Insights, Swiss Data Insights Association, Gruebengasse 10, 6055 Alpnach, Switzerland
+41 44 3501960, info@swiss-insights.ch, www.swiss-insights.ch

Unsere Mitglieder

SWISS INSIGHTS Institute Member

Im Verband sind alle relevanten **Markt- und Sozialforschungsinstitute** unter einem Dach organisiert. Alle Mitgliedsinstitute unterliegen einem strengen Regelwerk von schweizerischen und internationalen Normen.

Mitgliedsinstitute dürfen das Label **Market & Social Research by SWISS INSIGHTS** und je nach Tätigkeitsgebiet das Label **Data Fairness by SWISS INSIGHTS** tragen.

Die Mitgliederliste finden Sie auf der nachfolgenden Seite.

SWISS INSIGHTS Corporate Member

Die Corporate Mitglieder sind **Unternehmen**, die sich für den fairen Umgang mit Auskunftspersonen und Auftraggeber sowie für den Schutz der Privatsphäre engagieren.

Corporate Member, die das Label **Data Fairness by SWISS INSIGHTS** tragen, stehen für den wissenschaftlichen, seriösen und respektvollen Umgang mit Daten ein.

Die Mitgliederliste finden Sie auf der nachfolgenden Seite.

SWISS INSIGHTS

Institute Member

amPuls Market Research
Hirschengraben 49, 6000 Luzern 7
+41 41 612 14 14 / info@ampuls.ch
www.ampuls.ch

amrein+heller MarktforschungsTreuhand AG
Südweid 7, 6274 Eschenbach
+41 748 63 70 / contact@ah-feedback.ch
www.ah-feedback.ch

Bilendi Schweiz AG
Reinhardstrasse 19, 8008 Zürich
+41 79 801 88 80 / contact.ch@bilendi.com
www.bilendi.ch

Boomerang Ideas AG
Sihlquai 131, 8005 Zürich
+41 44 500 88 60 / raphael@boomerangideas.com
www.boomerangideas.com

Constant Dialog
Alte Steinhauserstrasse 33, 6330 Cham
+41 41 310 05 40 / info@constant-dialog.ch
www.constant-dialog.ch

DemoSCOPE Data + Research
Klusenstrasse 17, 6043 Adligenswil
+41 41 375 40 00 / demoscope@demoscope.ch
www.demoscope.ch

dr-ouwerkerk ag – just-medical!
Blegistrasse 5, 6340 Baar
+41 41 766 11 55 / info@just-medical.com
www.pharmaagentur.ch

Gallup AG
Reinhardstrasse 19, 8008 Zürich,
+41 78 891 31 15 / office@gallup.swiss
www.gallup.swiss

gff Swiss Research Services
Baarerstrasse 25, 6300 Zug
+41 41 560 01 60 / gut@gff.ag,
www.gff.ag

GfK Switzerland AG
Suurstoffi 18A, 6343 Rotkreuz
+41 41 632 91 11 / info.ch@gfk.com
www.gfk.ch / www.gfk.com

gfs.bern. Menschen. Meinungen. Märkte.
Effingerstrasse 14, Postfach, 3001 Bern
+41 31 311 08 06 / info@gfsbern.ch
www.gfsbern.ch

gfs-zürich, Markt- & Sozialforschung
Riedtlistrasse 9, 8006 Zürich
+41 44 360 40 20 / gfs@gfs-zh.ch
www.gfs-zh.ch

gfs-befragungsdienst
Schaffhauserstrasse 491, 8052 Zürich
+41 44 360 26 40 / info@gfs-bd.ch
www.gfs-bd.ch

GIM Suisse AG
General-Wille-Strasse 10, 8002 Zürich
+41 44 283 18 18 / info@g-i-m.ch
www.g-i-m.ch

Happy Thinking People AG
Staufacherstrasse 101, 8048 Zürich
+41 44 204 16 26 / contact-zurich@happythinkingpeople.com
www.happythinkingpeople.com

INNOFACT (Schweiz) AG Research & Consulting
Flurstrasse 50, 8048 Zürich
+41 43 931 77 82, Info@innofact.ch
www.innofact.ch

Instight Institute AG
Bergstrasse 138, 8032 Zürich
+41 44 387 90 90 / info@insightinstitute.ch
www.insightinstitute.ch

intervista
Optingenstrasse 5, 3013 Bern
+41 31 511 39 00 / anfragen@intervista.ch
www.intervista.ch

IPSOS Suisse SA
11, Chemin du Château-Bloch, 1219 Le Lignon
+41 22 591 06 00 / Contact_Switzerland@ipsos.com
www.ipsos.com/de-ch

Kantar Media Switzerland AG
Bahnhofstrasse 4, 3073 Gümligen
+41 31 537 79 00 / ch.panel@kantarmedia.com
www.kantarmedia.com

SWISS INSIGHTS
Institute Member

LINK
Baslerstrasse 60, 8048 Zürich
+41 41 367 73 73 / zurich@link.ch
www.link.ch

Marketagent.com Schweiz AG
Seefeldstrasse 19, 8008 Zürich
+41 43 555 06 50, schweiz@marketagent.com
www.marketagent.com

M.I.S. Trend SA
Pont Bessières 3, 1005 Lausanne
+41 21 320 95 03 / info@mistrend.ch
www.mistrend.ch

onlineumfragen.com GmbH
Kernserstrasse 15, 6056 Kägiswil
+41 44 500 50 54 / info@onlineumfragen.com
www.onlineumfragen.com

POLYQUEST AG
Flurstrasse 26, 3014 Bern
+41 31 335 64 00 / info@polyquest.ch
www.polyquest.ch

Publicom AG
Alte Landstrasse 55, 8802 Kilchberg
+41 44 716 55 11 / publicom@publicom.ch
www.publicom.ch

Qualitest AG, Institut für Marketing- und Sozialforschung
Rosenberghöhe 3, 6004 Luzernr
+41 41 712 12 21 / qualitest@qualitestag.ch
www.qualitestag.ch

SensoPLUS
Industriestrasse 16, 6300 Zug
+41 41 710 71 61 / info@sensoplus.ch
www.sensoplus.ch

NielsenIQ (Switzerland) GmbH
Park 4, 6039 Root D4
+41 41 445 64 64 / nielsen-ch@nielsen.com
www.nielsen.com

TransferPlus AG Market Research
Haldenstrasse 11, 6006 Luzern
+41 41 618 33 11 / transfer@transferplus.ch
www.transferplus.ch

SWISS INSIGHTS
Corporate Member

BSI Business Systems Integration AG
Täfernweg 1, 5405 Baden
+41 58 255 90 00, info@bsi-software.com
www.bsi-software.com