NLP Research Workshop

NLP-Anwendung in NFP73 und NFP77

PD Dr. Matthias Stürmer
Dominic Schweizer
Joel Niklaus

Forschungsstelle Digitale Nachhaltigkeit
der Universität Bern

Freitag, 14. August 2020
Universität Bern

Nachhaltige Wirtschaft
Nationales Forschungsprogramm

Digitale Transformation
Nationales Forschungsprogramm

UNIVERSITÄT BERN

NLP-Aktivitäten der FDN

Die Forschungsstelle Digitale Nachhaltigkeit beschäftigt sich im Bereich NLP aktuell mit folgenden Projekten:

- **NFP73:** Datenveredelung und Texterkennung
- **NFP77:** Topic Modelling, Named Entity Recognition etc.
- **Kanton Bern:** Chatbot-Analyse, Natural Language Understanding, Question & Answering Systems, Voicerecognition
- **SECO/ALV:** Matching zwischen offenen Stellen und Stellensuchen

NFP73 «Nachhaltigkeit bei Beschaffungen»

Università della Svizzera Italiana

University of Berne

Prof. Dr. Dr. Peter Seele
CSR and business ethics professor

Prof. Dr. Federica de Rossa Law professor and judge at the Federal Supreme Court

Sebastian Knebel
Corporate communication

Clarissa David
Lawyer

Dr. Matthias Stürmer
Computer scientist

Tobias Welz
Industrial engineer

UNIVERSITÄT BERN

SNSF NRP73 on Sustainable Procurement

- **Research question:**
  «How can public procurement become more sustainable?»
  - Public procurement accounts for CHF 41 bn spendings annually
  - Ecological and social criteria may make a big difference
  - In addition new procurement law in 2021 requires sustainability
- **Goal:** measuring sustainability criteria in public tender documents
- **Data and method:** crawling and refining of procurement data

73 NFP

Nachhaltige Wirtschaft
Nationales Forschungsprogramm

UNIVERSITÄT BERN

1. Beschaffungsstatistik.ch

- *NLP light*
- Simap.ch Publikationen als unstrukturierte Daten
- **Informationen aus Publikation extrahieren**

12.08.2020 | Projekt-ID 194015 | Meldungsnummer 1097745 | Ausschreibungen

Ausschreibung

Publikationsdatum Simap: 12.08.2020

1. Auftraggeber

1.1 Offizieller Name und Adresse des Auftraggebers
Bedarfsstelle/Vergabestelle: Post CH AG / Postmail / ganzer Konzern Post
Beschaffungsstelle/Organisator: Post CH AG /
F41 / Beschaffung, zu Hdn. von CC-WTO - Projekt «Taxi", Wankdorfallee 4, 3030 Bern, Schweiz, E-Mail: wto-taxi-dl@post.ch

1.2 Angebote sind an folgende Adresse zu schicken
Adresse gemäss Kapitel 1.1

1.3 Gewünschter Termin für schriftliche Fragen
26.08.2020
Bemerkungen: Fragen des Anbieters zur Ausschreibung und zu den Ausschreibungsunterlagen sind direkt in der Ausschreibungsplattform simap.ch bis spätestens 26.08.2020 zu stellen.
Es werden keine telefonischen oder mündlichen Auskünfte erteilt. Die Beantwortung der Fragen erfolgt direkt im Frageforum in der Ausschreibungsplattform simap.ch bis am 07.09.2020. Verspätet eintreffende Fragen werden nicht mehr beantwortet.

1.4 Frist für die Einreichung des Angebotes
Datum: 28.09.2020 **Uhrzeit:** 12:00, **Spezifische Fristen und Formvorschriften:** Das Angebot muss zur Wahrung seiner Rechtzeitigkeit zum genannten Zeitpunkt vollständig am genannten Ort eingetroffen sein; Eingaben per E-Mail oder Fax genügen nicht. Die Aufgabe eines Angebotes per Post zum genannten Zeitpunkt (Poststempel) reicht zur Wahrung der Rechtzeitigkeit nicht. Zu spät eingetroffene Angebote werden nicht berücksichtigt.

1.5 Datum der Offertöffnung:
28.09.2020, **Uhrzeit**: 12:00, **Bemerkungen**: Die Öffnung der Angebote ist nicht öffentlich.

1.6 Art des Auftraggebers
Bund (Dezentrale Bundesverwaltung - öffentlich rechtliche Organisationen)

1.7 Verfahrensart
Offenes Verfahren

1.8 Auftragsart
Dienstleistungsauftrag

1.9 Gemäss GATT/WTO-Abkommen, resp. Staatsvertrag
Ja

2. Beschaffungsobjekt

2.1 Dienstleistungskategorie CPC:
[27] Sonstige Dienstleistungen

2.2 Projekttitel der Beschaffung
Taxi DL (Taxi Dienstleistungen)

Link: https://beschaffungsstatistik.ch

UNIVERSITÄT BERN

1. Beschaffungsstatistik.ch

- Auftraggeber identifizieren
- Struktur erkennen
- Informationen aus String extrahieren
- **Auftraggeber mit Trainingsdaten vergleichen**

12.08.2020 | Projekt-ID 194015 | Meldungsnummer 1097745 | Ausschreibungen

Ausschreibung

Publikationsdatum Simap: 12.08.2020

1. Auftraggeber

1.1 Offizieller Name und Adresse des Auftraggebers

Bedarfsstelle/Vergabestelle: Post CH AG / Postmail / ganzer Konzern Post
Beschaffungsstelle/Organisator: Post CH AG /
F41 / Beschaffung, zu Hdn. von CC-WTO - Projekt «Taxi", Wankdorfallee 4, 3030 Bern, Schweiz, E-Mail: wto-taxi-dl@post.ch

1.2 Angebote sind an folgende Adresse zu schicken

Adresse gemäss Kapitel 1.1

1.3 Gewünschter Termin für schriftliche Fragen

26.08.2020
Bemerkungen: Fragen des Anbieters zur Ausschreibung und zu den Ausschreibungsunterlagen sind direkt in der Ausschreibungsplattform simap.ch bis spätestens 26.08.2020 zu stellen.
Es werden keine telefonischen oder mündlichen Auskünfte erteilt. Die Beantwortung der Fragen erfolgt direkt im Frageforum in der Ausschreibungsplattform simap.ch bis am 07.09.2020. Verspätet eintreffende Fragen werden nicht mehr beantwortet.

1.4 Frist für die Einreichung des Angebotes

Datum: 28.09.2020 **Uhrzeit:** 12:00, **Spezifische Fristen und Formvorschriften:** Das Angebot muss zur Wahrung seiner Rechtzeitigkeit zum genannten Zeitpunkt vollständig am genannten Ort eingetroffen sein; Eingaben per E-Mail oder Fax genügen nicht. Die Aufgabe eines Angebotes per Post zum genannten Zeitpunkt (Poststempel) reicht zur Wahrung der Rechtzeitigkeit nicht. Zu spät eingetroffene Angebote werden nicht berücksichtigt.

1.5 Datum der Offertöffnung:

28.09.2020, **Uhrzeit**: 12:00, **Bemerkungen**: Die Öffnung der Angebote ist nicht öffentlich.

1.6 Art des Auftraggebers

Bund (Dezentrale Bundesverwaltung - öffentlich rechtliche Organisationen)

1.7 Verfahrensart

Offenes Verfahren

1.8 Auftragsart

Dienstleistungsauftrag

1.9 Gemäss GATT/WTO-Abkommen, resp. Staatsvertrag

Ja

2. Beschaffungsobjekt

2.1 Dienstleistungskategorie CPC:

[27] Sonstige Dienstleistungen

2.2 Projekttitel der Beschaffung

Taxi DL (Taxi Dienstleistungen)

Link: https://beschaffungsstatistik.ch

1. Beschaffungsstatistik.ch

- Auftrag der Post CH AG
- Beschaffungsstelle:
  - Post CH AG / F41 / Beschaffung, Wankdorfallee 4 3030 Bern

1. Auftraggeber

1.1 Offizieller Name und Adresse des Auftraggebers

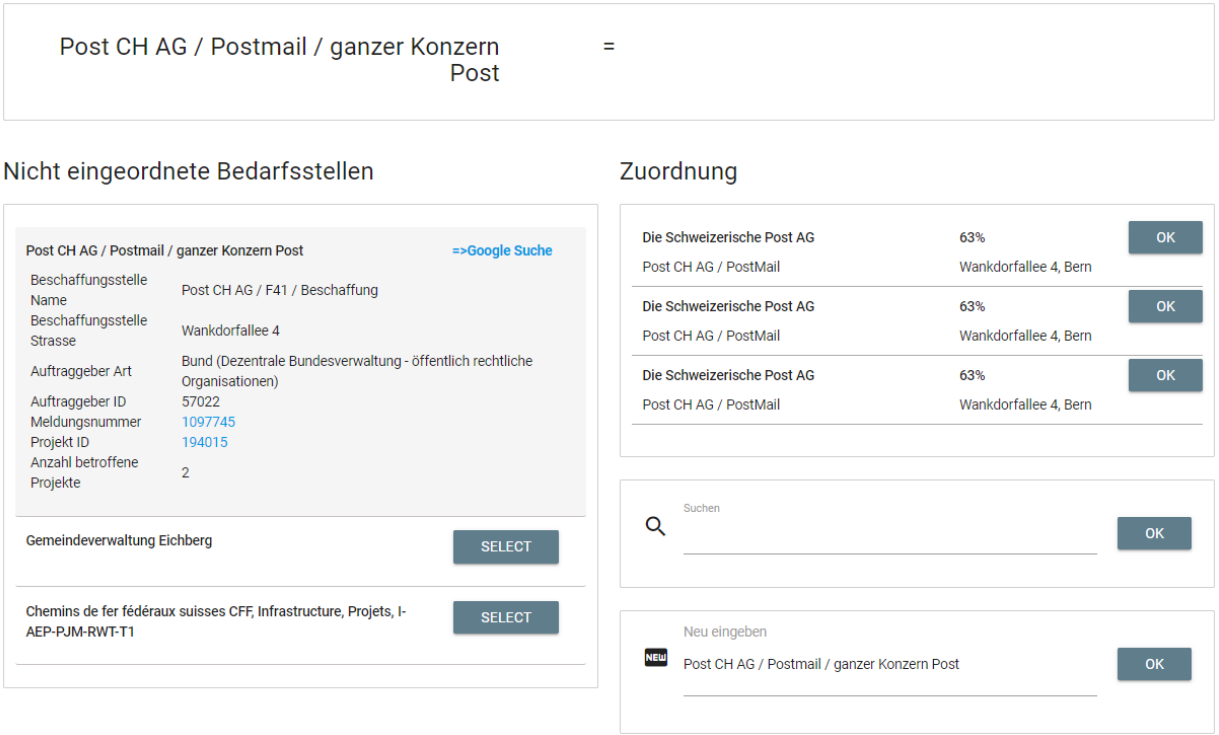
Bedarfsstelle/Vergabestelle: Post CH AG / Postmail / ganzer Konzern Post

Beschaffungsstelle/Organisator: Post CH AG /

F41 / Beschaffung, zu Hdn. von CC-WTO - Projekt «Taxi", Wankdorfallee 4, 3030, Bern, Schweiz, E-Mail: wto-taxi-dl@post.ch

UNIVERSITÄT BERN

1. Beschaffungsstatistik.ch



UNIVERSITÄT BERN

1. Beschaffungsstatistik.ch

Ausschreibung

(20 Einträge von 78'138 angezeigt)

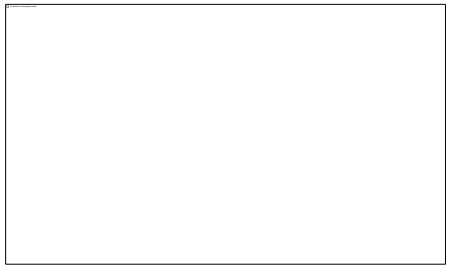
Volltextsuche in Tabelle Ausschreibung

Datum	Projekt	Meldung	Projekttitel	IT Projekt	CT projekt	Verfahrensart	Auftraggeber	Auftraggeber-Art
2020-08-12	194015	1097745	Taxi DL (Taxi Dienstleistungen)	Nein	Nein	Offenes Verfahren	Die Schweizerische Post AG	Bund (Dezentrale Bundesverwaltung - öffentlich rechtliche Organisationen)
2020-08-12	204584	1135925	KHK St. Gallen, Erneuerung der Rauchgasreinigung	Nein	Nein	Offenes Verfahren	Stadt St. Gallen	Gemeinde/Stadt
2020-08-12	206995	1145087	IWB, Fachplaner GP Bau Retrofit	Nein	Nein	Offenes Verfahren	IWB - Industrielle Werke Basel	Andere Träger kantonaler Aufgaben
2020-08-12	207883	1148399	Umbau und Sanierung Wirtschaftsgymnasium, BKP 283.2 Deckenverkleidung aus Gips	Nein	Nein	Offenes Verfahren	Kanton Basel-Stadt	Kanton
2020-08-12	207934	1148679	Appel d'offres pour des tunnels de lavage, désinfection à la vapeur et séchage de lits hospitaliers	Nein	Nein	Offenes Verfahren	Universitätsspital Lausanne CHUV	Kanton

UNIVERSITÄT BERN

2. Intelliprocure.ch

- Grösserer Datensatz
  - 490‘991 indexierte Ausschreibungsdokumente
  - 171‘297 Simap Publikationen
- Durchsuchen der Unterlagen
  - Volltextsuche



UNIVERSITÄT BERN

2. Intelliprocure

Nachhaltigkeit

SUCHEN

Total Resultate: 6738
Zeitdauer der Suche:0.184 Sekunden

Anzahl indexierte Projekte: 19'495
Total indexierte Ordner: 137'679
Total indexierte Dokumente: 490'991
Grösse der Daten: 1'665.132 GB
Neuste Ausschreibung: 13.08.2020

0-10/6738 ZURÜCK WEITER

Erneuerung und Erweiterung Campus Horw Das Verfahren besteht aus einer öffentlich ausgeschriebenen Präqualifikation (Phase 1) und einem nachfolgenden, einstufigen und nicht anonymen Studienauftrag (Phase 2) für 4 bis 5 Generalplanerteams. *.pdf*

Projekt ID: 188232 | Datum: 01.06.2019
Auftraggeber: Kanton Luzern (Simap:Kanton Luzern / Finanzdepartement)

... ja / nein Bauherr Adresse, PLZ Ort Kurzbeschrieb Adresse, PLZ Ort Name Adresse, PLZ Ort Funktion Beschrieb in Stichworten: - Projektbeschrieb - Innovationen bezüglich Nachhaltigkeit, Raumgestaltung, Gemeinschaft und Unterricht Referenz 2 Vorname, Name im Betrieb tätig seit Funktion Referenz 1 Adresse, PLZ Ort Name Adresse, PLZ Ort Funktion Beschrieb in Stichworten:...

Datei Score: 62.045002 | Dateiname: | Grösse: 0.25 MB
Ausschreibung Anzahl Dokumente: 4 | Anzahl Ordner: 0 | Grösse: 2.959 MB

In Beschaffungsstatistik ansehen
Auf Simap.ch ansehen
Dokument ansehen
Alle Dokumente zu diesem Projekt ansehen
Projekt als Zip herunterladen

UNIVERSITÄT BERN

2. Intelliprocure.ch: Ausblick

- Identifizierung von EK & ZK
  - Topic Modelling mit Random Forest & Active Learning
- Erstellung eines Kataloges für EK & ZK

Rang	Kombiniertes Modell		Einfaches Modell	
	Wort	Quelle	Wort	Quelle
1	zuschlagskriterium	Überschrift	zuschlagskriterium	Überschrift
2	zk	Überschrift	preis	Text
3	gewichtung	Text	zuschlagskriterium	Text
4	zuschlagskriterium	Text	zk	Text
5	angebot	Text	punkt	Text
6	punkt	Text	gewichtung	Text
7	zuschlagskriterium	Eltern-Überschrift	angebot	Text
8	schlüsselperson	Text	nr	Text
9	qualität	Text	angabe	Text
10	preis	Text	table	Text

UNIVERSITÄT BERN

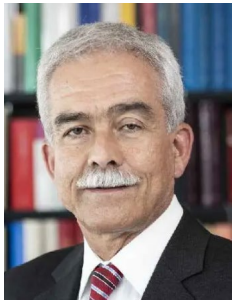
NFP77 «Open Justice vs. Privacy»

kpm kompetenzzentrum für public management universität bern

Institut für Wirtschaftsinformatik

Forschungsstelle Digitale Nachhaltigkeit am Institut für Informatik

Prof. Dr. Andreas Lienhard
Rechtsprofessor und Dekan Rechtswissenschaftliche Fakultät



Daniel Kettiger
Rechtsanwalt

Nathalie Schwager
Doktorandin Rechtswissenschaften

Prof. Dr. Thomas Myrach
Professor für Wirtschaftsinformatik

PD Dr. Matthias Stürmer
Leiter Forschungsstelle Digitale Nachhaltigkeit

Joel Niklaus
Doktorand Informatik

UNIVERSITÄT BERN

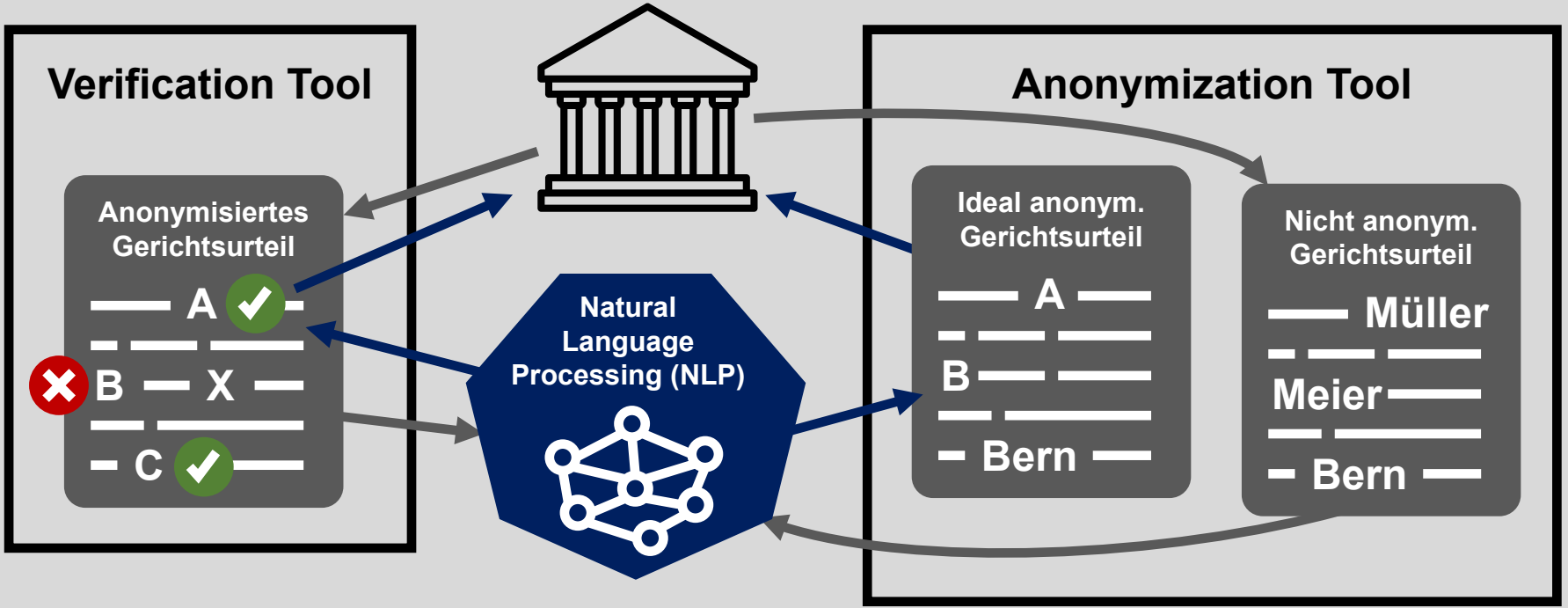
SNSF NRP77 on Open Justice vs. Privacy

- **Research question: «How can court decisions be released as open data without compromising privacy?»**
  - Transparency in justice expects court decisions to be published openly
  - Usually plaintiffs in a court decision are pseudonymized (e.g. «A vs. B»)
  - But big data and artificial intelligence may deanonymize court decisions
- **Goal:** conduct ethical deanonymization for improving anonymization
- **Data and method:** crawling and NLP analysis of Swiss court data

77 NFP

Digitale Transformation
Nationales Forschungsprogramm

Verification and Anonymization Tool



UNIVERSITÄT BERN

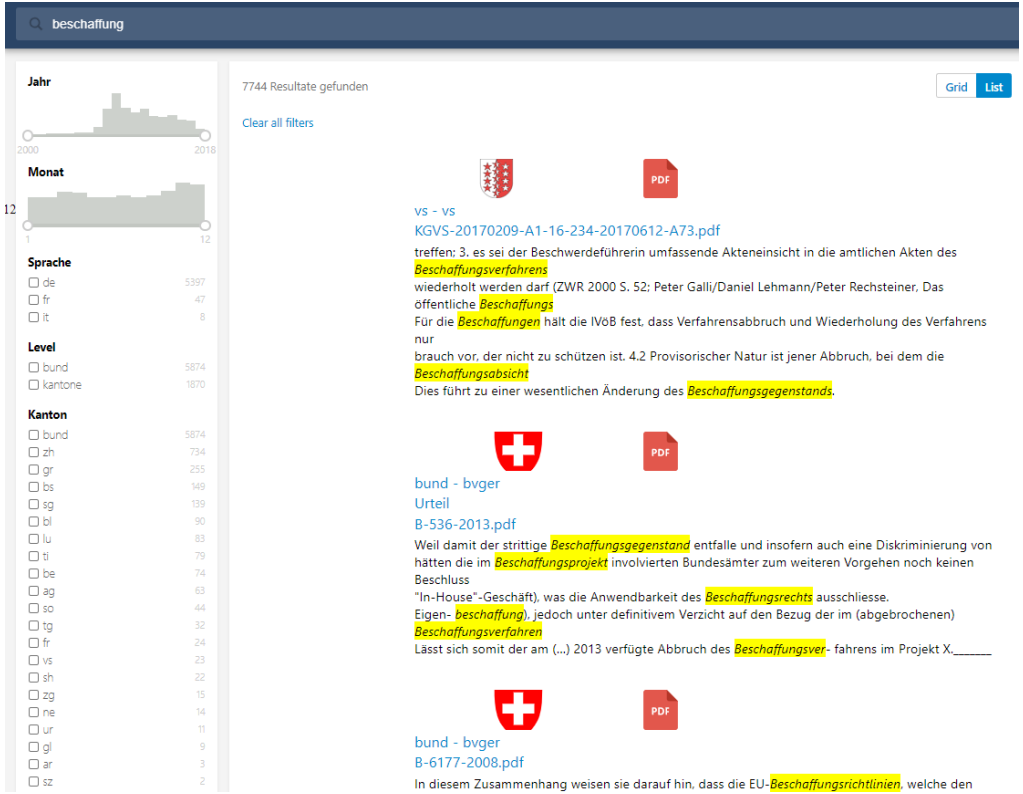
Plan

1. Datenset vorbereiten mit Entscheidsuche.ch
2. Topic Modeling für Gerichtsurteile (wichtig für Linkage)
3. Externe Daten für Linkage crawlen (bspw Zeitungen, Social Media)
4. Ein anonymisiertes Gerichtsurteil auswählen und re-identifizieren
5. Anonymisierungssystem (mit NER) für ein Gerichtsurteil bauen
6. Die beiden Systeme für eine Kategorie (topic) ausbauen
7. Die beiden Systeme für alle Kategorien ausbauen

UNIVERSITÄT BERN

1. Aktueller Stand der Datensammlung

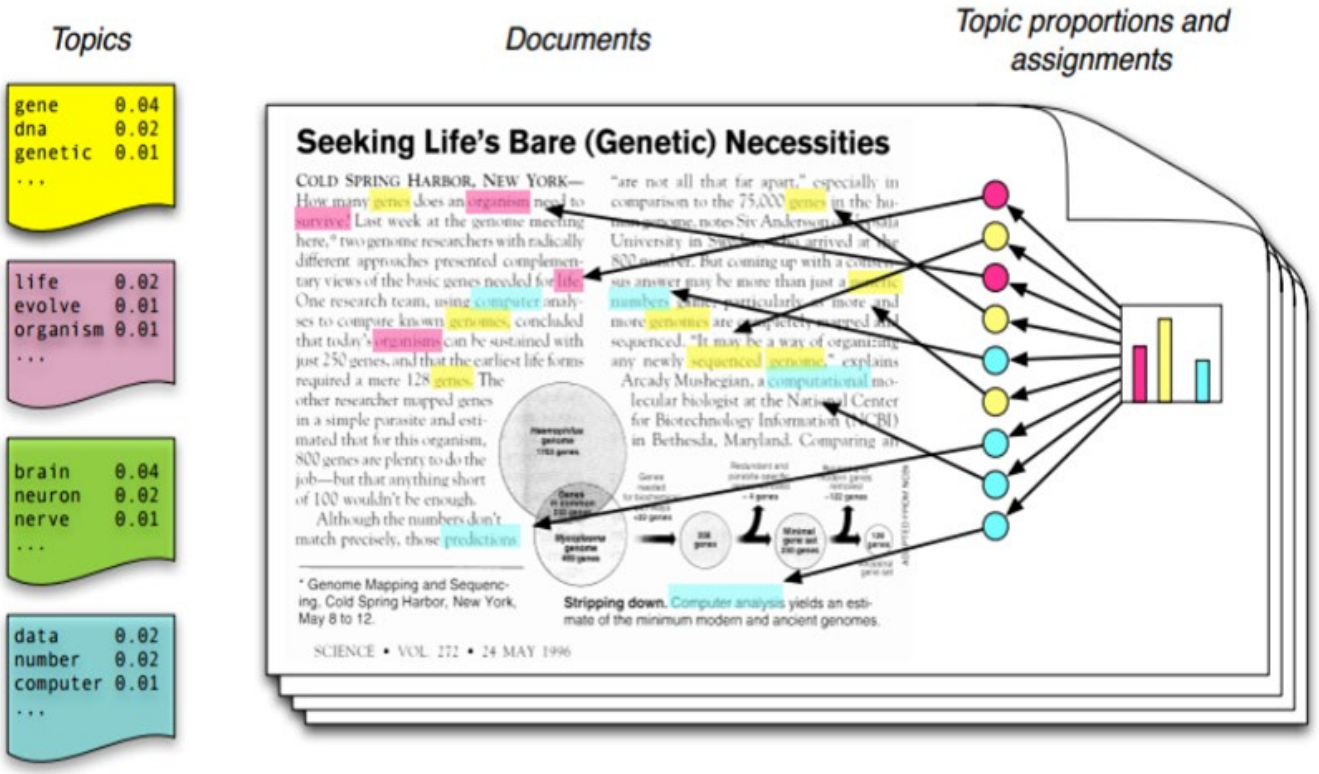
- Im Gespräch für Kooperation mit Verein **entscheidsuche.ch**
- Hat 2018 rund **500'000 Gerichtsurteile** aus rund 80 Gerichts-datenbanken gecrawlt
- **Wiederkehrendes Crawling** durch neue Programmierung der Scraper



Link: https://entscheidsuche.ch

UNIVERSITÄT BERN

2. Topic Modeling



UNIVERSITÄT BERN

2. Topic Modeling

- Unsupervised oder Supervised (Topic Classification)
- Methoden:
  - Latent Dirichlet Allocation (LDA)
  - Latent Semantic Analysis (LSA)

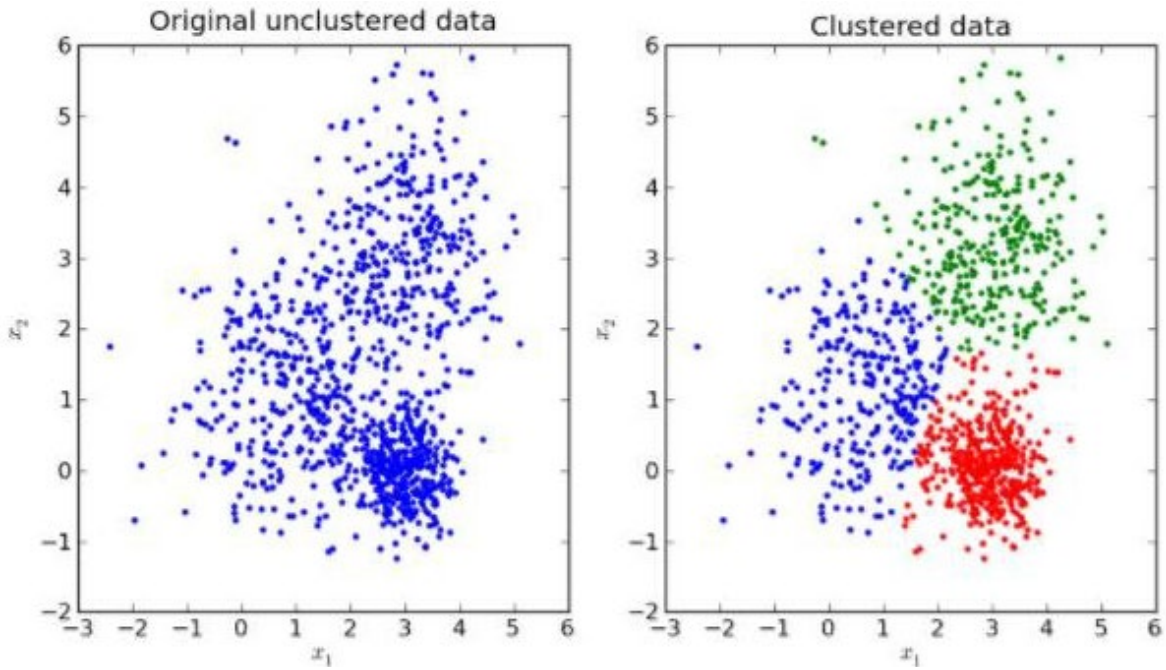


 Bild: https://www.researchgate.net/publication/333868156_Moderne_Data_Mining_Methoden

UNIVERSITÄT BERN

2. Topic Modeling

 Bild: https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

UNIVERSITÄT BERN

3. Externe Daten Crawlen

Bild: https://hackernoon.com/how-to-build-a-web-crawler-from-scratch-wps32ia

UNIVERSITÄT BERN

4. Re-Identifizierung

- Linkage mit externen Informationsquellen (Knowledge Base) mit
  - Identifikatoren
  - Quasi-Identifikatoren (nur zusammen mit anderen Informationen aussagekräftig)

  → Entity Linking (NER + NED)

…

 Bilder: https://www.britannica.com/topic/newspaper, https://www.johnlocke.org/update/court-decision-could-clear-the-way-for-final-payments-to-eugenics-victims/

5. Anonymisierung: Named Entity Recognition

In fact, the Chinese NORP market has the three CARDINAL most influential names of the retail and tech space – Alibaba GPE , Baidu ORG , and Tencent PERSON (collectively touted as BAT ORG), and is betting big in the global AI GPE in retail industry space . The three CARDINAL giants which are claimed to have a cut-throat competition with the U.S. GPE (in terms of resources and capital) are positioning themselves to become the 'future AI PERSON platforms'. The trio is also expanding in other Asian NORP countries and investing heavily in the U.S. GPE based AI GPE startups to leverage the power of AI GPE . Backed by such powerful initiatives and presence of these conglomerates, the market in APAC AI is forecast to be the fastest-growing one CARDINAL , with an anticipated CAGR PERSON of 45% PERCENT over 2018 - 2024 DATE .

To further elaborate on the geographical trends, North America LOC has procured more than 50% PERCENT of the global share in 2017 DATE and has been leading the regional landscape of AI GPE in the retail market. The U.S. GPE has a significant credit in the regional trends with over 65% PERCENT of investments (including M&As, private equity, and venture capital) in

UNIVERSITÄT BERN

5. Anonymisierung

- Vier Methoden:
  - Suppression (Peter → XXX)
  - Tagging (Peter → Vorname_1)
  - Random substitution (Peter → Hans)
  - Generalization (Peter → Name)

24 F. Dias, N. Mamede, and J. Baptista, “Automated Anonymization of Text Documents,” 2016.

Bild: https://www.solarwindsmsp.com/blog/data-anonymization-overview

Diskussion

Anregende Fragen:

- Bei welchen NLP Techniken ist der state-of-the-art noch nicht gut genug?
- Welche anderen NLP Techniken könnten relevant werden?

Bild: https://www.global-english.com/blog/tips-for-tefl-discussion-classes/