Fuzzy String Matching im Kontext der EU-Sanktionsliste

Integrationsseminar WWI20DSA

Joshua Brenzinger
Luis Steinert
Pascal Breucker

10.02.2023

Identifizierung von Sanktionierten Personen

Motivation und Inhaltliche Abgrenzung des Kontextes

Gesetzliche Vorschriften: Geld- und Haftstrafen bei Nichteinhaltung

Ruf: negative Publicity und Geschäftseinbußen

Risiko: finanzielle und rechtliche Risiken

Nationale Sicherheit: Schutz finanzieller und rechtlicher Interessen von Unternehmen

Content overview

1. Datengrundlage

Art und Form der für die Aufgabe zugrundeliegenden Informationen und Daten

Grundlage:

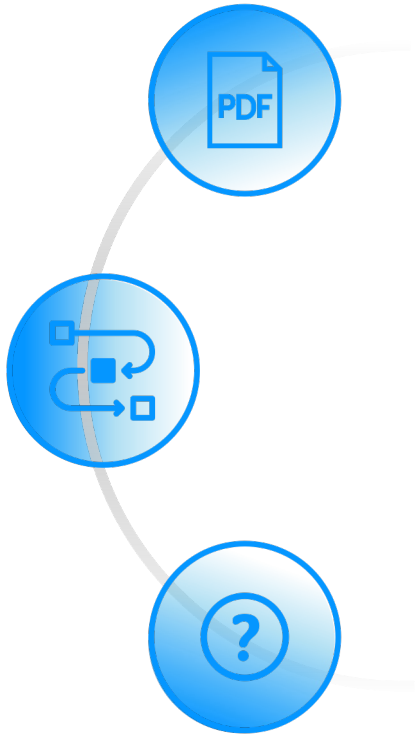
Consolidated list of persons, groups and entities subject to EU financial sanctions

Herangehensweise:

Datenaufnahmen → Datenaufbereitung → Datenbereinigung

Herausforderungen:

Uneinheitliche Einträge (Identity Information, Citizenship Information, Contact Information)



1. Datengrundlage

Beispieleintrag in der PDF

EU reference number: EU.2301.53
Legal basis: 2017/404 (OJ L63)
Programme: AFG - Afghanistan
Identity information:

- **Name/Alias:** Agha Jan Alizai
- **Name/Alias:** Haji Agha JanAlizai
- **Name/Alias:** Abdul Habib
- **Name/Alias:** Abdul Habib Alizai **Title:** Haji **Function:** Has managed a drug trafficking network in Helmand Province, Afghanistan
- **Name/Alias:** Hajji Agha Jan
- **Name/Alias:** Agha Jan Alazai
- **Name/Alias:** Haji Loi Lala
- **Name/Alias:** Loi Agha

Birth information:

- **Birth date:** 1967 **Birth place:** Unknown country
- **Birth date:** 1957 **Birth place:** Unknown country
- **Birth date:** Circa **Birth place:** Afghanistan, Musa Qala District, Helmand Province, Yatimchai village **Remark:** Yatimchai village, Musa Qala District, Helmand Province, Afghanistan
- **Birth date:** 14/02/1973 **Birth place:** Unknown country
- **Birth date:** 15/10/1963 **Birth place:** Unknown country
- **Birth place:** Afghanistan, Kandahar

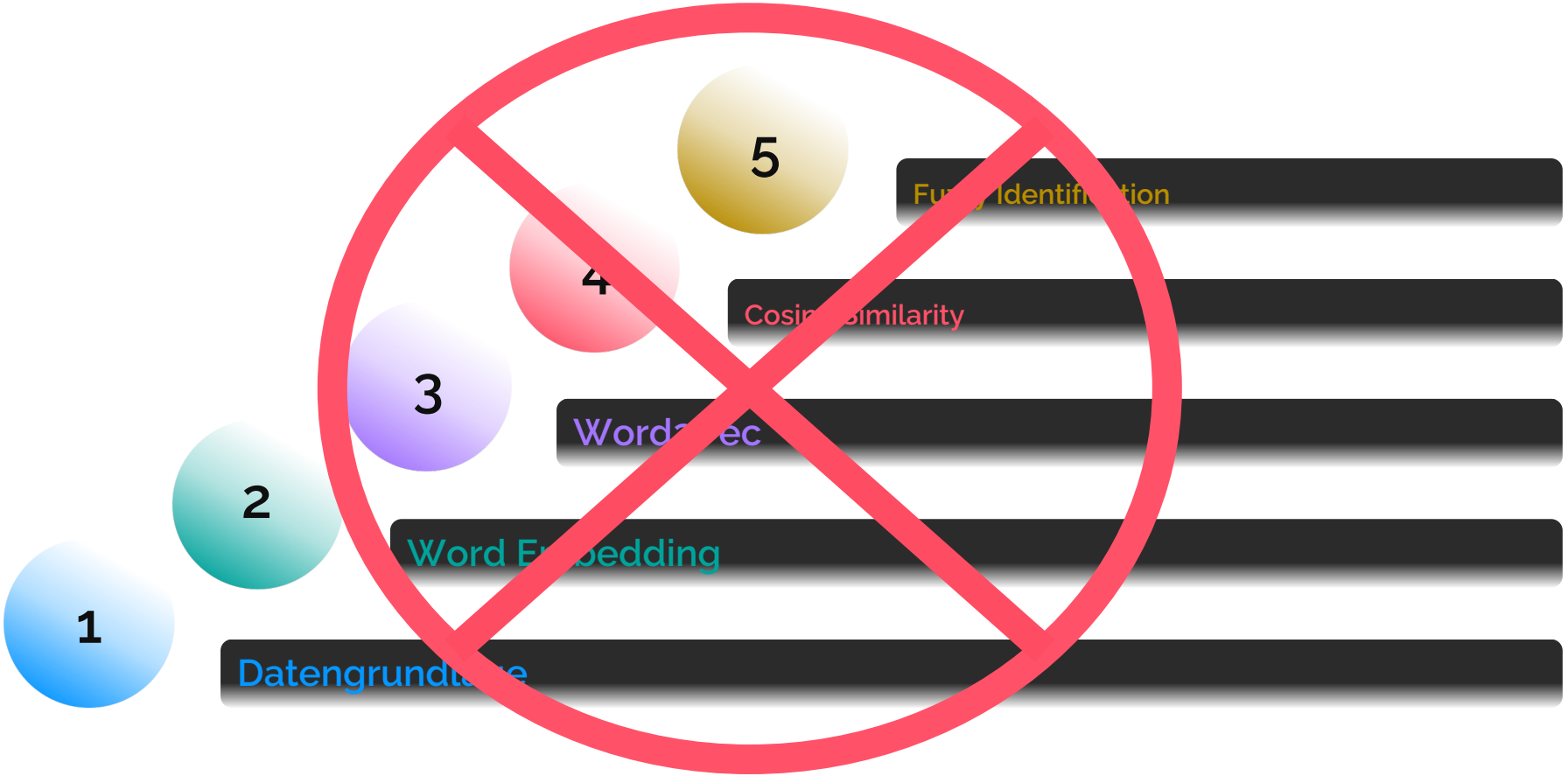
Citizenship information:

- **Citizenship:** Afghanistan

Remark: INTERPOL-UN Security Council Special Notice web link: https://www.interpol.int/en/notice/ search/un/1684147

2. Word Embedding

Versuchter Ansatz für Fuzzy Identifikation



3. Named Entity Recognition

Vortrainiertes BERT Modell nutzen um Namen + Orte zu filtern

Base_bert_NER

Entitätserkennung

- LOC → Location
- PER → Person
- ORG → Organisation
- MISC → Miscellaneous

Datengrundlage für Finetuning

- CoNLL-2003 NER
- Problem → Herkunft der Namen

Performance

Funktionalität

- bei Orten → GUT
- bei Namen → nicht unbedingt gut

Programme: A ORG FG - Afghanistan LOC Identity information: • Name/Alias: Abdul Baqi Basir Awal Shah PER Title: (a) Ma PER ulavi, (b) Mu PER llah Function: (a) Governor of Khost LOC and Paktika LOC provinces under the Taliban MISC regime, (b) Vice-Minister of Information and Culture ORG under the Taliban ORG regime, (c) Consulate Department, Ministry of Foreign Affairs ORG under the Taliban ORG regime. • Name/Alias: Abdul Baqi PER Birth information: • Birth date: Circa LOC from 1960 to 1962 Birth place: Afghanistan LOC , Shinwar District LOC , Nangarhar Province LOC • Birth date: from 1960 to 1962 Birth place: Afghanistan LOC , Jalalabad City LOC , Nangarhar Province LOC Citizenship information: • Citizenship: Afghanistan LOC

4. Fuzzy Matching
Abgrenzung der Begrifflichkeit

Probleme / Hintergrund

Konzept:
Unscharfe Suche für Schlüsselbegriffe

Probleme:
- Berücksichtigung ähnlicher Begriffe
- Kontext
- Verbesserte Suchergebnisse

Konzept

Package theFuzz:
Berechnung der Unterschiede zwischen den Sequenzen und Ausgabe in Prozent

5. String Matchin Methods

Longest Common Substring

Länge der längsten zusammenhängenden Zeichenkett

- Ganzzahliger Wert
- Inputabhängig

contra → Aussagekraft

Thomas	Thomas
Tobias	Tobias

Levenshtein Distance

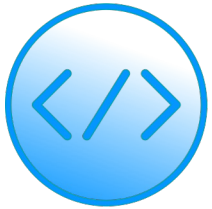
Anzahl Bearbeitungsschritte

1. Zeichen Hinzufügen
2. Zeichen löschen
3. Zeichen ersetzen

contra → Berechnungsintensiv

Thomas	Tomas	Tobas	Tobias
Tobias	Tobias	Tobias	Tobias

5. String Matchin Methods



Jaro-Winkler Similarity

- basiert auf Jaro Similarity

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

- Präfix Skala
- bestimmt Stringlänge
- Wert zwischen 0 – 1

$$sim_w = sim_j + \ell p(1 - sim_j)$$

Fuzzy Similarity

- Verschiedene Vergleichsmethoden
- Implementierung von Levenshtein
- Unterschied der Strings als Output

6. Livedemo

Fuzzy Search in Streamlit

https://fuzzy-sanctions.streamlit.app

Fazit

Kritische Auseinandersetzung

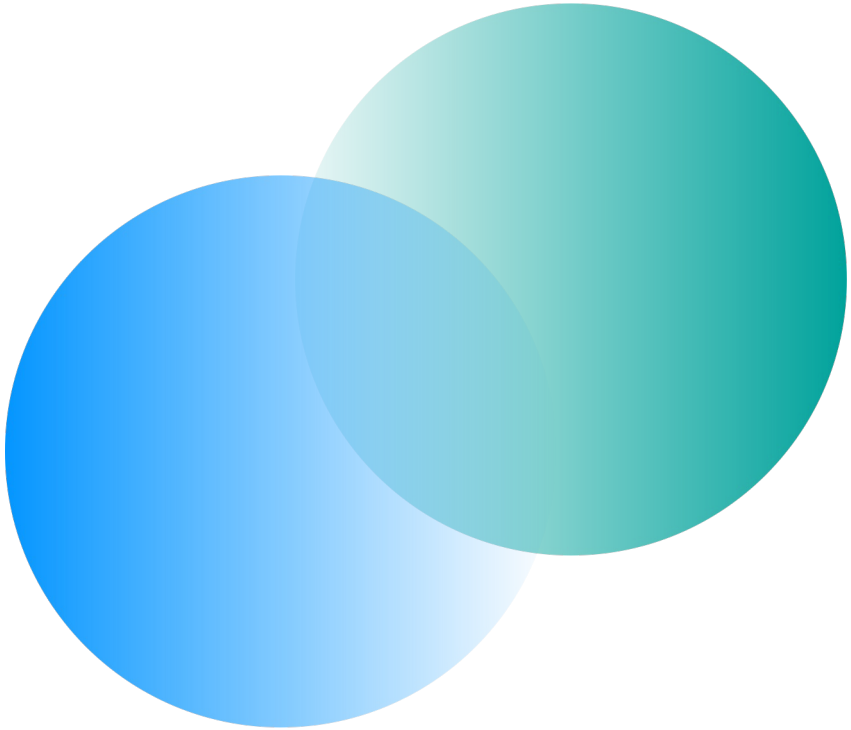
Garbage in → Garbage out

- Preprocessing

- fuzzy score bei langen texten weil in prozent betrachtet wird.

Levenshtein sinnvoll bei kurzen texten und wörtern

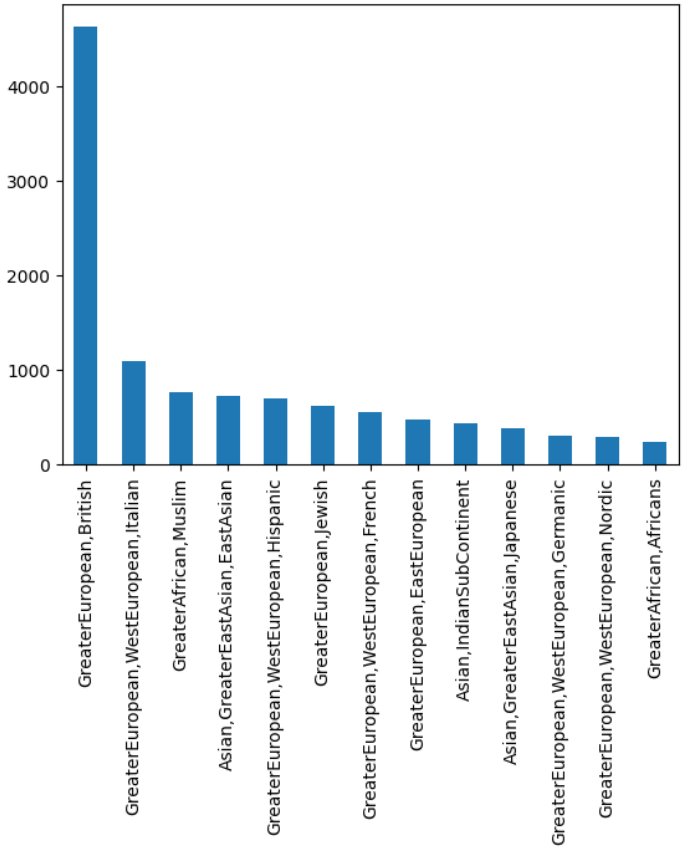
Backup Slides



Dieses Projekt wurde im Rahmen des Integrationsseminars des Studienganges WWI20DSA an der DHBW Mannheim durchgeführt. Ziel war es das erlernte in den Modulen der Wirtschaftsinformatik und Data Science zu kombinieren und im Rahmen des Seminars ein Projekt zu entwickeln und dieses zu dokumentieren und Präsentieren

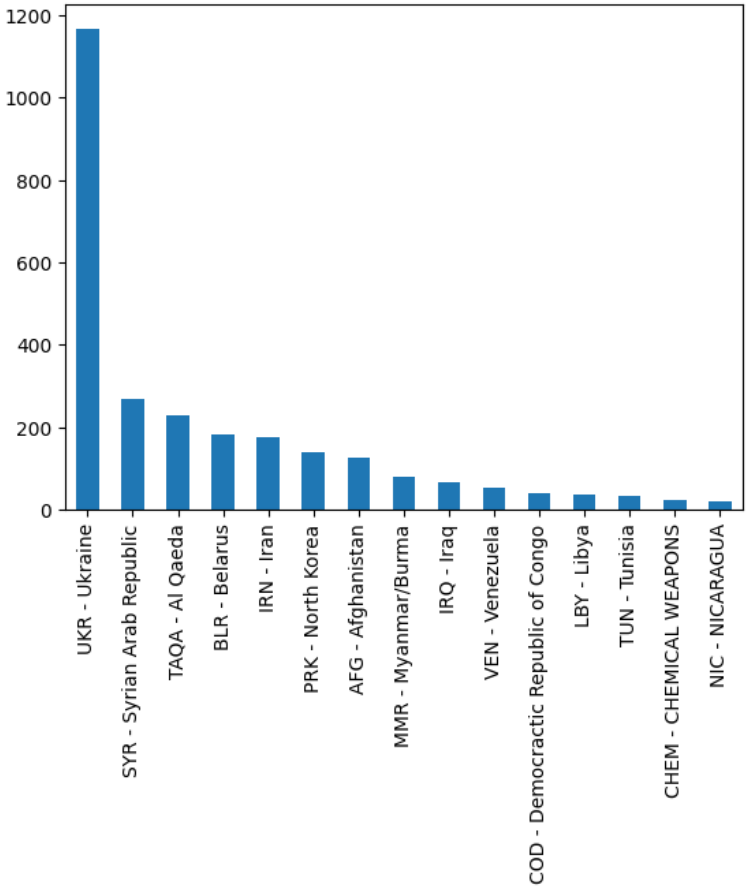
NER

Daten von Bert kommen aus folgenden Regionen



NER

Häufigkeit der Programme



PDF

Zweites Beispiel aus PDF

EU reference number: EU.312.72
Legal basis: 2017/404 (OJ L63)
Programme: AFG - Afghanistan
Identity information:

- **Name/Alias:** Abdul Baqi Basir Awal Shah **Title:** (a) Maulavi, (b) Mullah **Function:** (a) Governor of Khost and Paktika provinces under the Taliban regime, (b) Vice-Minister of Information and Culture under the Taliban regime, (c) Consulate Department, Ministry of Foreign Affairs under the Taliban regime.
- **Name/Alias:** Abdul Baqi

Birth information:

- **Birth date:** Circa from 1960 to 1962 **Birth place:** Afghanistan, Shinwar District, Nangarhar Province
- **Birth date:** from 1960 to 1962 **Birth place:** Afghanistan, Jalalabad City, Nangarhar Province

Citizenship information:

- **Citizenship:** Afghanistan

Remark: Believed to be in Afghanistan/Pakistan border area. Review pursuant to Security Council Resolution 1822 (2008) was concluded on 1 Jun. 2010. INTERPOL-UN Security Council Special Notice web link: https://www.interpol.int/en/notice/search/ un/1493921