Themenvorschlag BA-Arbeit

Ziel

Textklassifikation: Machine-Learning-Ansatz zur Bewertung von Beschreibungstexten aus dem Gegenstand einer Firma im Handelsregister anhand der Klassifikation der Wirtschaftszweige (WZ2008)

Hintergrund

Im Handelsregister gibt es zu jeder Firma eine Beschreibung deren Tätigkeitsbereichs oder -bereiche, den sogenannten Gegenstand. Dabei handelt es sich um einen kurzen Beschreibungstext[1], der je nach Firma sehr kurz oder etwas detaillierter sein kann.

Bei BDS verwenden wir die Klassifikation der Wirtschaftszweige des statistischen Bundesamtes (WZ)[2], um Branchen systematisch über verschiedene Quellen zu vereinheitlichen. Das Ziel ist also, einen HR-Gegenstand-Text auf einen WZ-Code abzubilden.

Aufgabe

Für diese Einteilung bieten sich Machine-Learning-Ansätze aus dem Bereich der Textklassifikation an. Herausforderungen können dabei der geringe Umfang sowie die recht variable Länge der Texte sein. Auch handelt es sich eher um „Behördensprache“ mit ggf. spezieller Wortwahl.

Als Datengrundlage können Firmen verwendet werden, für die ein HR-Gegenstand vorliegt, sowie eine WZ-Zuordnung aus einer anderen Quelle als Zielvariable. BDS kann eine höhere sechsstellige Menge an Datensätzen bereitstellen. Die Zuordnung von WZ-Codes zu Firmen muss nicht eindeutig sein, eine Firma kann also mehrere WZ-Codes erhalten.

[Hier wären uU Erweiterungen möglich: Herausfiltern weiterer möglicher Unternehmensmerkmale aus dem Text]

Mögliche Aspekte

- Gewinnung und Modellierung von Features aus den Gegenstands-Texten
- Erprobung verschiedener Klassifikations-Modelle, gerne auch in Richtung Wortvektoren/Neuronale Netze
- Auswertung der Klassifikationsergebnisse
- Resultierender Klassifikationsprozess in Python (oder Java) für BDS nutzbar und wiederverwertbar

---

[1] Beispiele siehe Anhang

[2] https://www.destatis.de/DE/Methoden/Klassifikationen/Gueter-Wirtschaftsklassifikationen/klassifikation-wz-2008.html

Beispiele für HR Gegenstand Beschreibungen

- Die Produktion individueller Holzmöbel.
- Der Betrieb eines Frisörgeschäftes.
- Gegenstand des Unternehmens sind die Erbringung von Dienstleistungen aller Art im Zusammenhang mit mobilen Zahlungssystemen, u.a. bei der Parkraumbewirtschaftung in Kommunen. Gegenstand sind ferner die Entwicklung, der Betrieb und die Vermarktung von Software in diesem Bereich sowie die Entwicklung und Durchführung von Zertifizierungsprozessen. Gegenstand ist ferner die Vermarktung von Werberechten.
- Der Kfz-Ersatzteilhandel, der Import und Export von neuen, gebrauchten und verbrauchten Kfz-Ersatzteilen, die Kfz-Vermietung, der Handel mit Kosmetik- und Drogerieartikeln sowie der Import und Export von Kosmetik- und Drogerieartikeln sowie Transporte bis zu 3,5 Tonnen.
- Gerüstbau
- die Vorbereitung und Einbringung von Verwaltungs- und Vertriebsarbeiten, von Reinigungsarbeiten außer Gebäudereinigung, von Service-Leistungen und sonstigen Dienstleistungen für Restaurant- und Catering-Unternehmen, soweit diese nicht genehmigungspflichtig sind.