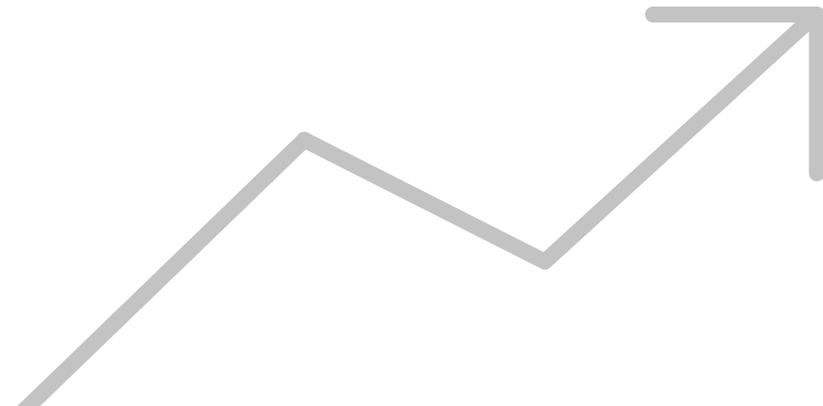


Textklassifikation

Maschinelles Lernen zur
Klassifikation von freien Texteingaben
in der amtlichen Statistik



Vom Text zur Klasse

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet.

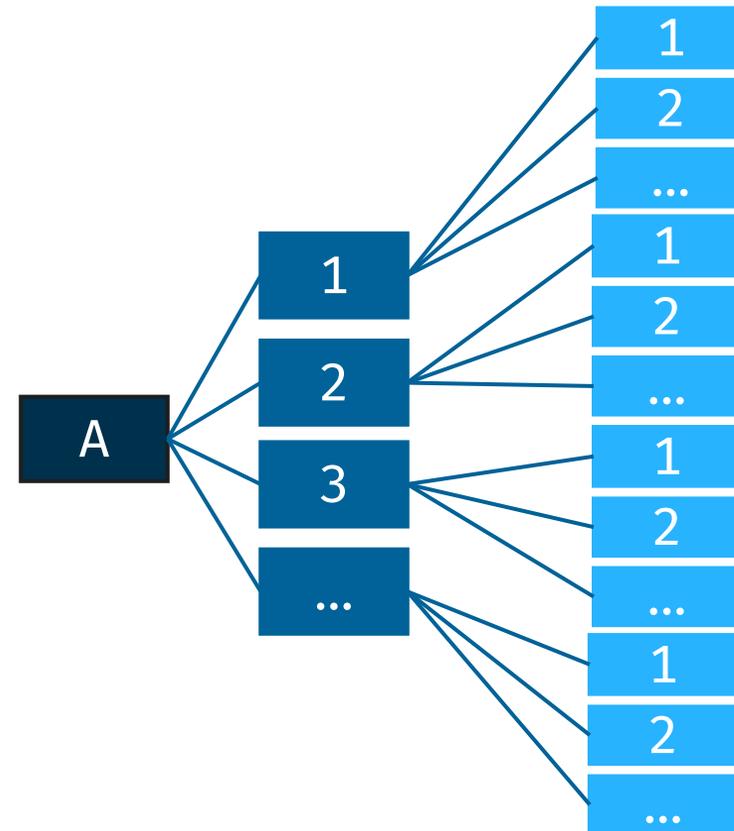
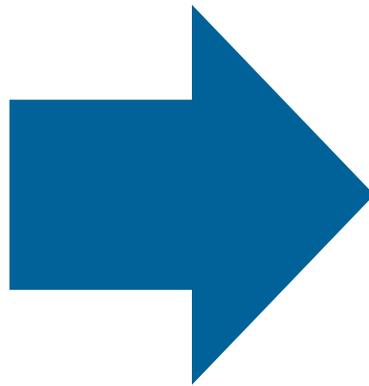
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugiat nulla facilisis. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wis enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et justo odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugiat nulla facilisis.

Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wis enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, At accusam aliquyam diam diam dolore dolores duo eirmod eos erat, et nonummy sed tempor et et invidunt justo labore Stet clita ea et gubergren, kasd magna no rebum. sanctus sea sed takimata ut vero voluptua. est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat.

Consetetur sadipscing elitr, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus.

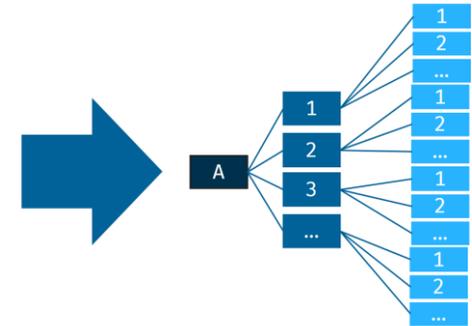
Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea takimata sanctus est Lorem ipsum dolor sit amet. Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed diam nonummy eirmod tempor invidunt ut labore et dolore magna aliquyam erat, sed diam voluptua. At vero eos et accusam et justo duo dolores et ea rebum. Stet clita kasd gubergren, no sea.



Vom Text zur Klasse



Quellenhinweise: Statistisches Bundesamt https://www.destatis.de/DE/Ueber-uns/Geschichte/_inhalt.html
Fotorechte: © Statistisches Bundesamt



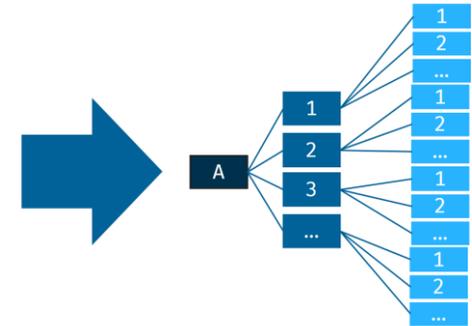
zeitintensiv

personalintensiv

Vom Text zur Klasse

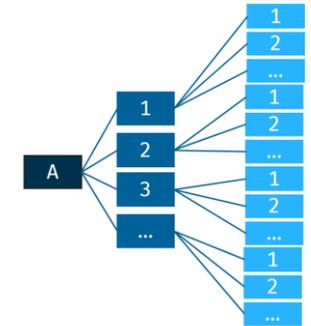
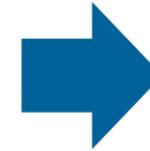


Fotorechte: Wim Klerkx Photography <https://www.wimklerkx.nl>



Verfügbares
Textvolumen steigt

Projekte und Projektideen in der amtlichen Statistik



Geburtsstaatsklassifikation in den Wanderungsstatistiken

Trainingsdaten

rund 6.000.000 Wanderungsfälle jährlich,
teils manuell plausibilisiert

Merkmale Herzugland, Nationalität,
Geburtsdatum, Religion und *Geburtsort* liegen vor.

Klassifikation

Geburtsstaat eines Wanderungsfalles



COICOP-Klassifikation mit Scannerdaten

Trainingsdaten

Textbeschreibungen von Produkten + GTIN
regelmäßige Lieferungen großer Datenmengen

Klassifikation

- » Zielgröße: COICOP 10-Steller (> 200 Klassen)
- » Mehr als 85% Accuracy bei großen Unterschieden in den Klassen
- » Weitere Handelsketten aufnehmen



COICOP: Klassifikation der Verwendungszwecke des Individualverbrauchs (Classification of Individual Consumption by Purpose)
Foto: commons.wikimedia.org

Klassifikation der Wirtschaftszweige in den Gewerbeanzeigen

Trainingsdaten

Freitextangaben von Gewerbeanmeldungen

Klassifikation

- » Zielgröße WZ-4-Steller (ca. 600 Klassen)
- » Knapp 60% Accuracy bei großen Unterschieden in den Klassen
- » Weitere Hilfsmerkmale könnten die Ergebnisse verbessern



Quelle: Deutsche Fotothek

Freitexte in den freiwilligen Haushaltserhebungen

Trainingsdaten

Freitextangaben zu Ausgaben und Aktivitäten

Klassifikation

- » Systematik der Ein- und Ausgaben (ca. 250 Klassen)
- » ca. 89% Accuracy bei großen Unterschieden in den Klassen
- » Besonderheiten beim Meldeverhalten könnten ins Modell integriert werden



Klassifikation der Handelsregister- und Insolvenzbekanntmachungen

Projektidee

Trainingsdaten

frei verfügbare, unstrukturierte und heterogene Bekanntmachungen im Internet

Ziel

- » Identifikation bestimmter thematischer Sachverhalte für Fachstatistiken
- » Automatische Zuordnung zu Einheiten im Unternehmensregister



Kontakt

Statistisches Bundesamt
Gustav-Stresemann-Ring 11
65189 Wiesbaden

www.destatis.de

www.destatis.de/kontakt

Ansprechpartner
Joerg Feuerhake
joerg.feuerhake@destatis.de
Telefon +49 611 75-4116

