Auxiliary AI GmbH

Classification with spaCy

- Block 04-

Auxiliary AI GmbH

Exkurs: Bürgerbeteiligung

- Informelle Bürgerbeteiligungen werden in vielen Städten durchgeführt, um die Meinungen der Bürgerinnen und Bürgern anonym, digital, zeit- und ortunabhängig zu erheben. Dies erfolgt für z. B. Bauvorhaben, kulturelle, soziale und weiteren Projekte und Initiativen der öffentlichen Verwaltung, Vereinen, Verbänden und Unternehmen.
- Beteiligungen werden z. B. mit digitalen Plattformen wie DIPAS durchgeführt. Die Daten werden nach dem Ende der Beteiligung ausgewertet, indem Inhalte **induktiv** und **deduktiv kategorisiert** und nach dem **Sentiment** (Stimmung) analysiert werden.
- Die Ergebnisse fließen in die Entscheidungsfindung ein und sollen so den realen Mehrwert der und die Akzeptanz für die Projekte erhöhen.

Auxiliary AI GmbH

Datensatz Bürgerbeteiligung

- Ca. 19.000 gelabelte Texte mit einer Länge von einem Satz bis zwei Sätzen.

- 0 <-> Kultur
- 1 <-> Sport
- 2 <-> Umwelt & Grün
- 3 <-> Mobilität
- 4 <-> Öffentliche Dienste & Sicherheit
- 5 <-> Soziales & Netzwerk
- 6 <-> Wohnen
- 7 <-> Wirtschaft
- 8 <-> Bildung
- 9 <-> Undefiniert
- 10 <-> Sauberkeit
- 11 <-> Lautstärke & Emissionen
- 12 <-> Erholung

Auxiliary AI GmbH

Training eines Machine Learning Klassifzierer

- **Aufgabe 1:** Ladet den Datensatz und analysiert diesen manuell und indem ihr euch die Klassenverteilung genauer anschaut. Was fällt auf und wie aussagekräftig ist der Datensatz dieser?
- **Aufgabe 2:** Vektorisiert den Datensatz mithilfe von spaCy. Überprüft, ob ihr Wortvektoren mit Länge X für jeden Satz erhalten habt.
- **Aufgabe 3:** Trainiert einen *Decision Tree* und eine *SVM* (sowie nach Bedarf weitere Modelle) mit 80% der Daten.

Auxiliary AI GmbH

Training eines Machine Learning Klassifzierer

- **Aufgabe 4:** Evaluiert die Zuverlässigkeit der Vorhersagen mit den verbliebenden 20% der Daten und erstellt die *Confusion Matrix*.
- **Aufgabe 5:** Fügt die Cross-Validation ein und überprüft, ob es Abweichungen zwischen den Modellen gibt.
- **Aufgabe 6:** Überlegt, wie die Zuverlässigkeit mithilfe von *Punktuation (entfernen), PoS, um einzelne Typen auszuschließen, Stop-Words und Lemmatization (ggf. weiteren)* erhöht werden. Erhöht die Zuverlässigkeit der Vorhersage, indem ihr unterschiedliche Möglichkeiten anwendet.

Auxiliary AI GmbH

Training eines Machine Learning Klassifzierer

- **Aufgabe 7:** Sind die Klassen gleichverteilt? Nein! Wie können wir das Problem lösen? Benennt die zwei Möglichkeiten und löst das Problem ohne händische Arbeit.

  Downscaling: Reduktion von Klassen und Anzahl der Inhalte von Einträgen auf eine feste Anzahl. Dadurch werden teils viele Daten entfernt.

  Upscaling: Ergänzung der Daten durch weitere gelabelte Beispiele. Teils herausfordernd und mit viel Aufwand verbunden.

- **Aufgabe 8:** Trainiert weitere Klassifikatoren auf den erweiterten Datensatz (einen erhaltet ihr von uns) und vergleicht die Zuverlässigkeit der Ergebnisse. Variiert dabei auch die Preprocessing Pipeline. Welchen F1-Score etc. erreicht ihr und wer hat das beste Modell?

Auxiliary AI GmbH

Auxiliary AI GmbH

Geschäftsführer	**Marten Borchers & Benjamin Klinkigt**
Anschrift	Am Ziegelteich 74 22525 Hamburg Deutschland
Handelsregister	HR B 185519
Registergericht	Amtsgericht Hamburg
Umsatzsteuer-ID	DE 366 814 276
Kontakt	info@auxiliary-ai.de
Webseite	https://auxiliary-ai.de/