VECTOR

Fallstudie AI

Aufgabe

Das Gebiet des Natural Language Processings (NLP) ermöglicht es, automatisiert natürliche Sprache zu analysieren und zu verstehen. Blog-Einträge bieten eine einfache Möglichkeit, Meinungen und Informationen mit anderen Menschen zu teilen. Durch diese öffentlichen Beiträge können mittels NLP automatisierte Rückschlüsse auf den Verfasser gezogen werden.

Im Rahmen dieser Aufgabe soll eine Pipeline für eine automatisierte Analyse und Darstellung eines Datensatzes mit Blog-Einträgen aufgebaut werden. Der Datensatz wurde durch Crawling (blog.fefe.de) und entfernen von HTML-Tags generiert. Die Pipeline sollte mindestens aus den folgenden vier Bestandteile bestehen:

- Data Preprocessing: Vorbereitung und Säuberung des Datensatzes. Dies kann beispielsweise das Klassifizieren von Zitaten, Tokenisierung, Generelle Bereinigung und/oder Vektorisierung beinhalten.
- Classification: Beispielsweise die Einordnung von Texten zu Sentiments und übergeordneten Themen.
- Clustering: Einzelne Texte können beispielsweise in Themen aufgrund von Ähnlichkeiten eingeordnet werden. Das gleiche kann ebenso mit Themen durchgeführt werden. Ein üblicher Algorithmus für das Clustering ist z.B. k-Means-Clustering.
- Result presentation: Darstellung der Ergebnisse aus den Abschnitten 'Classification' und 'Clustering'. Dies kann über unterschiedliche Visualisierungsarten erfolgen wie beispielsweise Tag-Clouds.

Die beschriebenen Schritte dienen als Orientierungspunkte und sollen nach blieben angepasst werden. Ebenso ist der mitgelieferte Datensatz eine Empfehlung und kann bei Wunsch auch durch Alternativen ersetzt werden.

Vorstellung

- Bitte stelle deine Lösung in maximal 20 Minuten vor.
- Dabei kannst du ein Medium deiner Wahl nutzen.