

Textklassifikation

Beschreibung

In dieser Aufgabe wirst du eine Textklassifikation auf einem offenen Datensatz von Reuters durchführen. Die Aufgabenstellung ist bewusst vage und offen formuliert, sodass du eigene Entscheidungen treffen kannst und dich auf Aspekte konzentrierst, die du besonders interessant findest.

Anweisungen

- Lade den [Reuters-Datensatz](#) herunter.
- Deine Aufgabe ist es, die Dokumente den in **TOPICS** enthaltenen Kategorien zuzuordnen.
- Halte dich an die in der XML-Datei unter **LEWISSPLIT** angegebene Aufteilung in Trainings- und Testdaten.
- Falls du Zeit hast, bewerte und dokumentiere die Ergebnisse.
- Reiche ein lauffähiges Code-Stück ein, dokumentiert nach deinen eigenen Präferenzen (z. B. direkt im Code, in separaten Dokumenten oder durch Tests – Hauptsache, der Code ist verständlich).
- Wir sind an deinen Ideen sowie an deinem Programmierstil und deinen Fähigkeiten interessiert. Mach dir keine Sorgen, falls die Klassifikationsergebnisse nicht optimal sind.
- Falls du Ideen hast, wie die Lösung weiter verbessert werden kann, schreibe diese bitte ebenfalls auf.
- Plane etwa **3 Stunden** für diese Aufgabe ein. Kein Problem, wenn du es nicht komplett schaffst. Schreib einfach kurz auf, was du noch machen würdest.

Diskussionspunkt für das Meeting

Falls du bei Reuters arbeiten würdest und ein solches Modell entwickelt hättest:

- Wie könnte es genutzt werden, um Mehrwert für das Unternehmen und/oder die Nutzer zu schaffen?
- Welche Einschränkungen hat das Modell?
- Würdest du etwas anders machen?