Hochschule RheinMain
Prof. Dr. Adrian Ulges

28.10.2021

Anwendungen der KI (WS 21/22)

Aufgabenblatt 2

zu bearbeiten bis: 04.11.2021

Aufgabe 2.1 (Fragen annotieren)
Für Aufgabe 2.3 benötigen wir Ihre Fragen. Finalisieren Sie daher Ihre 40 Fragen pro Team. Senden Sie sie **bis Freitag den 29.10.2021 23:59 Uhr** in einer ";"-separierten csv-Datei an `johannes.villmow@hs-rm.de`.

Aufgabe 2.2 (Suchmaschine)
Wir wollen in diesem Übungsblatt mit ES eine größere Dokumentsammlung indexieren: Die **Wikibase**, die als Basis für unser Question-Answering-Projekt dienen soll kennen Sie bereits aus der ersten Übung. Die Wikibase besteht aus $266,341$ Wikipedia-Artikeln.

a) Lesen Sie die Dokumentensammlung zeilenweise in Python ein. Laden Sie dabei jede Zeile mit Hilfe des Python json-Moduls `import json`. Jedes Dokument enthält drei Feldern `doc_id` (die Wikipedia-ID des Artikels), `title` (der Titel des Artikels) und `text` (der Text des Artikels).

b) Fügen Sie die geladenen Artikel in Ihre ElasticSearch-Suchmaschine ein. *Hinweis: Mit bulk-Requests können Sie die Dokumente deutlich schneller indexieren! (siehe elasticsearch.helper()).*

c) Schreiben Sie ein einfaches textuelles Shell-Interface, so dass man Queries (z.B. einfache Fragen) in die Shell eingeben kann und die Titel der 20 Top-Treffer-Dokumente erhält.

d) Testen Sie mit ein paar Beispiel-Fragen (z.B. "Who murdered Abraham Lincoln?"), ob Sie vielversprechende Dokumente für die Beantwortung der Frage finden. Notieren Sie sich ein paar Beispiel-Fragen und Ihre Beobachtungen.

Aufgabe 2.3 (Messung)
Messen Sie wie gut Ihre Retrieval-Ergebnisse sind:

a) Sie finden ab Sonntag den 31.10.2021 im Read.MI unter `Trainingsfragen` die csv-Datei `train_all.csv`. Diese enthält neben Ihren noch weitere Beispiel-Fragen, die Ihnen später für das Projekt als Trainingsmenge dienen. Spalte 2 enthält jeweils die Frage, Spalte 5 die ID des Wikipedia-Artikels mit der korrekten Antwort.

b) Schreiben Sie ein Beispiel-Skript, das nacheinander alle Fragen als Queries für den ES-Index benutzt. Prüfen Sie jeweils, ob Sie das korrekte **Dokument** gefunden wurde: Messen Sie PREC@K für K=1,5,10,20,50,100, sowie die Average Precision.

c) Versuchen Sie mit *Field Boosting* in ES den Titel des Dokuments stärker zu gewichten. Gelingt es Ihnen die Güte der Ergebnisse messbar zu verbessern?