Hochschule RheinMain
Prof. Dr. Adrian Ulges

23.04.2018

Anwendungen der KI (SS 2018)

Aufgabenblatt 2

Aufgabe 2.1 (Suchmaschine)
Wir wollen in diesem Übungsblatt mit ES eine größere Dokumentsammlung indexieren: Die **Wikibase**, die als Basis für unser Question-Answering-Projekt dienen soll. Die Wikibase besteht aus $266,341$ Wikipedia-Artikeln. Sie befindet sich in der PostgreSQL-Datenbank `ulges_anwki18` auf `db.intern.mi.hs-rm.de`. Sie besitzen Lesezugriff.

a) Verbinden Sie sich mit Hilfe des Python-Moduls `psycopg2` mit der Datenbank. Die Tabelle `DOCS` enthält die Dokumente, jedes mit den drei Feldern `pageid` (die Wikipedia-ID des Artikels), `pagetitle` (der Titel des Artikels) und `pagetext` (der Text des Artikels). Laden Sie die Artikel in einer Schleife batch-weise mit `fetchmany()` herunter.

b) Fügen Sie die heruntergeladenen 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 2.2 (Messung)
Messen Sie wie gut Ihre Retrieval-Ergebnisse sind:

a) Sie finden auf der Homepage unter `Trainingsfragen` die csv-Datei `train_all.csv`. Diese enthält 631 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 631 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?
