Prüfungsrelevante Übungsaufgaben zur Vorlesung Textdatenbanken 2020

Prof. Uwe Quasthoff

Auf Grund der besonderen Situation in diesem Semester gibt es auch für die Prüfungen besondere Rahmenbedingungen.
Für die Vorlesung Textdatenbanken wird es in diesem Semester keine schriftliche Prüfungsklausur geben, sondern wir verfahren wie unten beschrieben.
Wir stellen drei Serien von prüfungsrelevanten Übungsaufgaben, und zwar an den folgenden Tagen: 12.6.2020, 26.6.2020 und 10.7.2020
Für die Lösung stehen jeweils 14 Tage zur Verfügung, Abgabe per E-Mail als PDF.
Pro Serie sind maximal 20 Punkte erreichbar. Aus der Gesamtpunktzahl wird eine Prüfungsnote ermittelt.
Es steht frei, an dieser Prüfung teilzunehmen. Note 5 wird wie Nichtteilnahme gewertet. Alle anderen Noten sind endgültig.
Für Nicht-Teilnehmer wird es zu einem späteren Zeitpunkt (frühestens im Oktober 2020) eine Prüfung als Klausur geben.

Es gelten die folgenden Regeln:

- Sie versichern durch Ihre Teilnahme, dass Sie die Aufgaben selbständig und ohne fremde Hilfe gelöst haben.
- Alle Aussagen sind zu begründen.
- Hilfsmittel sind erlaubt. Ggf. verwendete fremde Programme sind anzugeben. Bei der Verwendung eigener Programme ist deren Funktionsweise zu skizzieren, so dass die Ergebnisse nachvollziehbar sind.
- Soweit nötig, werden Daten als Download zu den Aufgaben zur Verfügung gestellt.
- Der Text der Lösungen sollte jeweils eine bis maximal zwei A4-Seiten (in 10p-Schrift, gemessen ohne Abbildungen) pro Aufgabe betragen und keinen Quellcode enthalten.
- Pro Aufgabe werden 10 Punkte vergeben.
- Abgabe als PDF an Felix Helfer *helfer@informatik.uni-leipzig.de*.
- Bitte Namen und Matrikel-Nr. angeben.

Serie 2/3: Ausgabe am 26.6.2020, Abgabe bis 10.7.2020

3. Speicherung in Baumstrukturen

(Die Abbildungen für die Darstellung der Bäume werden bei der Platzbeschränkungn nicht berücksichtigt. Die Beschränkung betrifft maximal zwei Seiten *Text* pro Lösung.)

3.1. Konstruieren Sie (von Hand) einen **Trie** aus der folgenden Wortmenge:
Antwort, Bahn, Anlage, Beta, Ananas, Bett, Banane, Anlauf, Bahnen
3.2. Konstruieren Sie (von Hand) einen **Patricia-Trie** aus der Wortmenge der vorigen Teilaufgabe.
3.3. Konstruieren Sie (von Hand) einen PAT-Tree aus dem folgenden Text, bis (inklusive) zum siebten Substring: *0010100110*
3.4. Beschreiben Sie, wie im Baum der letzten Teilaufgabe dessen häufigstes Ziffern-Bigramm identifiziert werden kann und geben Sie dieses an.
3.5. Beschreiben Sie allgemein, wie eine *Präfix*-Suche in einem solchen **PAT-Tree** verlaufen würde. Wie könnte alternativ eine *Suffix*-Suche für einen gegebenen Text ermöglicht werden?

4. Korpusabfragen mit CQL auf der NoSketch-Engine

Zur Bearbeitung der Aufgabe benutzen Sie bitte die NoSketchEngine mit dem Korpus deu_newscrawl-public_2018 unter http://cql.corpora.uni-leipzig.de/bonito/run.cgi/first?corpname=deu_newscrawl-public_2018

Verwendete CQL-Queries sind anzugeben.

4.1. Suchen Sie nach möglichst vielen Formen der Redewendung *einen Streit vom Zaun brechen*. Was statt *Streit* wird im Korpus noch *vom Zaun gebrochen* (gerankte Liste, Mindestanzahl 2)? Welche anderen Formen (flektiert, umgestellt) der Redewendung finden Sie (gerankte Liste, Mindestanzahl 2)?
4.2. Gesucht ist eine Liste von Namen von Hauptstädten, und zwar sowohl bestehend aus einem Wort (*Paris*) wie auch aus mehreren Wörtern (*Ulan Bator*). Geben Sie für jedes untersuchte Muster die 20 häufigsten so gefundenen Namen an.
4.3. Extrahieren Sie eine möglichst lange Liste von Vornamen aus dem Korpus mit Hilfe einer einzigen oder mit wenigen Abfragen. Wenn Sie diese nach Häufigkeit sortieren: Was stellen Sie fest? Geben Sie die häufigsten 30 Vornamen an.
4.4. Qualität der Ergebnisse: Welche fehlerhaften Ergebnisse liefern Ihre Abfragen? Was kann man dagegen tun? Fehlen wichtige Einträge?