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 3: Ausgabe am 10.7.2020, Abgabe bis 24.7.2020

5. (Praktisch) Quasidubletten in vorgegebener größerer Satzmenge finden.

DATEN: Satzliste aus einem deutschen Korpus, ca. 1M Sätze.

Beschreiben Sie einen für Sie realisierbaren, aber möglichst allgemeingültigen Ansatz, um aus der gegebenen Liste von rund 1 Millionen Sätzen Quasidubletten zu ermitteln, also Mengen von solchen Sätzen, die sich nur minimal voneinander unterscheiden.

5.1. Finden Sie für die zwei angegebenen Kriterien jeweils Mengen von mindestens 5 Quasidubletten. Die Sätze sollen mindestens 30 Zeichen lang sein. Beschreiben Sie, wie Sie dabei vorgegangen sind (Quellcode ist *nicht* einzureichen).

- Je zwei Sätze unterscheiden sich ausschließlich durch Zahlen.
- Je zwei Sätze unterscheiden sich je ein oder zwei Wörter, nicht durch Zahlen.

5.2 Ermitteln Sie aus der gegebenen Liste von Sätzen eine *möglichst große Anzahl* an *Paaren* von Quasidubletten für jedes der folgenden zwei Kriterien. Die Sätze sollen mindestens 30 Zeichen lang sein.

- Je zwei Sätze unterscheiden sich nur durch ihre Satzzeichen
- Ein Satz ist die verlängerte Version eines anderen (wobei die abschließenden Satzzeichen des kürzeren Satzes sich dabei ändern können: "*So ist das.*" - "*So ist das, sagte sie.*") .

Beschreiben Sie Ihre dafür verwendeten Ansätze (Quellcode ist *nicht* einzureichen). Geben Sie die Listen aller so gefundenen Paare für jedes Kriterium in zwei separaten Textdateien ab (Typ *.txt*).

6. (Praktisch) Sprachstatistik

DATEN: Wie bei 5., zusätzlich Wortliste mit Häufigkeit.

6.1. Stellen Sie die Verteilung der Satzlängen (*Satzlänge in Zeichen* vs. *Häufigkeit dieser Satzlänge*) graphisch dar. Können Sie eine Abweichung von der erwarteten Verteilung feststellen? Falls ja: Können Sie diese erklären?

6.2. Überprüfen Sie das Zipfsche Gesetz mit der gegebenen Wortliste, indem Sie den Graphen für *Rang* vs. *Häufigkeit* doppelt-logarithmisch darstellen. Berechnen Sie den Anstieg einer im mittleren Bereich der Kurve angenäherten Gerade. Beschreiben Sie Ihr Vorgehen, speziell:

- Welchen Bereich wählen Sie für die Berechnung des Anstieges? Warum?
- Wie groß ist der von Ihnen ermittelte Anstieg?

6.3. In der Wortliste befinden sich auch die Jahreszahlen. Welche Informationen liefert Ihnen die Verteilung der Jahreszahlen? Geben Sie die ermittelte Verteilung für den interessanten Bereich an.

6.4. Ermitteln Sie jeweils die 5 Sätze mit höchster bzw. niedrigster durchschnittlicher Wortlänge und mindestens 5 Wörtern pro Satz. Dabei ist es ausreichend, nur Leerzeichen als Wortgrenzen zu berücksichtigen. Haben diese Sätze irgendwelche ungewöhnlichen Eigenschaften?