# 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 Umfang für die Lösungen sollte jeweils eine bis maximal zwei A4-Seiten (in 10p-Schrift) 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 1/3: Ausgabe am 12.6.2020, Abgabe bis 26.6.2020

## 1. (Theoretisch): Relationen zwischen Wörtern: Für Liste von Wörtern (davon einige mehrdeutig)

Gegeben ist die folgende Liste von Wörtern: *Anhänger, LKW, Partei, CDU, SPD, Mitglieder, gewählt, langsam, Motor, Fahrzeug, Kupplung, fahren*
Teilaufgaben:

1. Geben Sie die bestehenden Relationen zwischen diesen Wörtern an (Ober- / Unterbegriff, Kohyponym, Meronym (Bestandteil, Element), typisches Objekt für, typische Eigenschaft von).
2. Fügen Sie der Liste oben fünf selbst gewählte Wörter hinzu, die untereinander oder zu den Wörtern oben in verschiedenen Relationen stehen und ordnen Sie diese ein.
3. Beschreiben Sie drei verschiedene Verfahren, die Kandidaten für die entsprechenden Relationen liefern (darunter mindestens jeweils eines *mit* bzw. *ohne* Verwendung von Trainingsdaten). Welche zusätzlichen Daten sind nützlich bzw. erforderlich?
4. Wie werden mehrdeutige Wörter durch diese Verfahren eingeordnet?
5. Beschreiben Sie mögliche Fehler, die von diesen Verfahren gemacht werden können.

## 2. (Praktisch): Cognates mittels Levenshtein-Ähnlichkeit in Wikipedia-Artikeln finden

Daten: Gegeben sind Sätze aus jeweils drei einander entsprechenden Wikipedia-Artikeln in den drei Sprachen Deutsch, Englisch und Französisch, ohne und mit POS-Tags[1].
Ermitteln Sie Cognates (incl. völlig übereinstimmender Wörter) in den gegebenen Texten entsprechend selbstgewählter Parameter zu den unten aufgeführten Kriterien und geben Sie diese an.
Definieren Sie genau den von Ihnen benutzten Levenshtein-Abstand!
Erklären Sie Ihr Vorgehen und diskutieren Sie Ihr Ergebnis.
Welche Kriterien für Cognates erscheinen Ihnen wichtig:

1. Umgang mit Groß- / Kleinschreibung und mit diakritischen Zeichen.
2. Maximaler Levenshtein-Abstand
3. Mindestwortlänge
4. Relative Übereinstimmung als Kombination von Levenshtein-Abstand und Wortlänge
5. Zueinander passende POS-Tags

Wie erklären Sie unterschiedliche Größe der Ergebnismengen für die verschiedenen Sprachpaare?
Lassen sich Qualitätsmaße wie Precision und Recall definieren und bestimmen?

[1] Für das POS-Tagging wurde der TreeTagger verwendet. Informationen über die Tagsets gibt es hier:
Deutsch: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/STTS-Tagset.pdf
Französisch: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html
Englisch: https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf