Universität Regensburg

Prof. Dr. U. Kruschwitz 11.2.2020

Klausur "Einführung in die Informationslinguistik 1" WS 2019/2020

Nachname, Vorname	
Abschluss (BA, MA, FKN etc.)	
Matrikelnummer, Semester	
Versuch (1/2/3)	

Bitte füllen Sie zuerst den Kopf des Angabenblattes aus!

Studieren Sie nach der ***neuen*** **Prüfungsordnung,** so dauert die Klausur 90 Minuten. Beantworten Sie bitte alle Fragen, also Fragen 1,2 und 3. In der Klausur sind insgesamt 75 Punkte zu erreichen.

Studieren Sie nach der ***alten*** **Prüfungsordnung,** so dauert die Klausur 60 Minuten. Beantworten Sie bitte Fragen 1 und 2. In der Klausur sind insgesamt 50 Punkte zu erreichen.

Die Klausur besteht aus 10 Seiten.

Bitte beantworten Sie alle Fragen im freien Platz direkt nach der jeweiligen Teilfrage. Falls der Platz nicht reichen sollte, so nutzen Sie gegebenenfalls das leere Blatt im Anhang und kennzeichnen dies entsprechend. Sie können das Blatt auch als Schmierpapier benutzen. Eigene Schmierblätter sind nicht erlaubt.

Viel Erfolg!

Question 1 / Aufgabe 1: (30 marks / 30 Punkte)

Basics / *Grundlagen*.

Question 1.1 / Aufgabe 1.1: (5 marks / 5 Punkte)

A major problem in Natural Language Engineering (NLE) is ambiguity. List three examples to illustrate this.
Mehrdeutigkeit ist ein wesentliches Problem der automatischen Sprachverarbeitung. Nennen Sie drei Beispiele, um diese Aussage zu belegen.

Question 1.2 / Aufgabe 1.2: (5 marks / 5 Punkte)

This is a simple regular expression (RE) using Perl syntax:

```
/^([\d]+)\.([0-9]+)\.([\d]+)\.([0-9]+)$/
```

Write down one string that matches this RE and one that does not. Justify your examples.

Oben sehen Sie einen regulären Ausdruck in Perl-Notation. Geben je ein Beispiel für Texteingaben, die von diesem Ausdruck erkannt bzw. nicht erkannt werden und erläutern Sie Ihre Antwort kurz.

Question 1.3 / Aufgabe 1.3: (5 marks / 5 Punkte)

Briefly explain the motivation for using the factor *tf* in the weighting formula *tf.idf*.
Geben Sie eine kurze Motivation für den Faktor tf in der Formel tf.idf.

Question 1.4 / Aufgabe 1.4: (5 marks / 5 Punkte)

Look at the following short text:

```
The exam will be in lecture theatre H4. That's great!
```

Assign part-of-speech tags to each of the words in the text.
Oben sehen Sie einen kurzen Text. Ermitteln Sie die syntaktische Wortklasse für jedes einzelne Wort.

Question 1.5 / Aufgabe 1.5: (5 marks / 5 Punkte)

Briefly discuss the problems a tokenizer might encounter when processing texts which contain the period character ('.').
Erklären Sie kurz, welche Probleme ein 'Tokenizer' bei Texten haben kann, die einen Punkt enthalten.

Question 1.6 / Aufgabe 1.6: (5 marks / 5 Punkte)

Provide a single reason as to why sparse word embeddings might have become so popular.
Benennen Sie einen Grund dafür, warum 'sparse word embeddings' so populär geworden sind.

Question 2 / Aufgabe 2: (20 marks / 20 Punkte)

Grammars and Parsing / *Grammatiken und Syntaxanalyse*.

Question 2.1 / Aufgabe 2.1: (10 marks / 10 Punkte)

Imagine you want to communicate with your smart home via natural language input. Develop a very simple context-free grammar (CFG) that accepts these grammatically correct user inputs:

```
"start the heating and close the kitchen window"
"make me a White Russian"
```

Ensure that ungrammatical inputs such as the following should be rejected by your grammar:

```
* "start the heating and close window"
* "make me"
```

Stellen Sie sich vor, Sie wollen mit Ihrer intelligenten Wohnung über natürliche Sprachverarbeitung kommunizieren. Entwickeln Sie eine sehr einfache kontextfreie Grammatik, die die oben genannten korrekten Eingaben akzeptiert (die ersten zwei Bespiele), inkorrekte Eingaben (wie die anderen zwei) jedoch nicht erkennt.

Question 2.2 / Aufgabe 2.2: (5 marks / 5 Punkte)

Outline how you would extend the coverage of your grammar so that it can deal with simple questions such as:

"do I have new email messages?"

"could you make me a cocktail?"

Skizzieren Sie, wie Sie Ihre soeben entwickelte Grammatik erweitern könnten, um einfache Fragen wie die obigen zwei erkennen zu können.

Question 2.3 / Aufgabe 2.3: (5 marks / 5 Punkte)

Briefly discuss whether a finite state automaton (FSA) can be developed that defines the same language that your grammar accepts.

Diskutieren Sie kurz, ob die von Ihrer Grammatik definierte Sprache auch mit Hilfe eines endlichen Automaten beschrieben werden kann.

Question 3 / Aufgabe 3: (25 marks / 25 Punkte)

Applications / ***Anwendungen*****.**

Question 3.1 / Aufgabe 3.1: (10 marks / 10 Punkte)

Imagine you want to build a word processor that is able to predict the next word you will type, based on what you have typed so far. Such a system could be trained on the user's specific writing style. Outline how you could train such a system on n-gram probabilities using maximum likelihood estimation. Use the mini-corpus below (starting with "The iSchools ...") as an example to illustrate what the system would suggest to follow after the last two tokens ("*in iSchools*") using different types of n-grams.

Stellen Sie sich vor, Sie sollen eine Anwendung programmieren, die bei einer Texteingabe in der Lage ist, auf Basis der bereits eingebenen Wörter das nächste Wort vorherzusagen. Man kann sich vorstellen, dass dieses System auf den jeweiligen Schreibstil eines einzelnen Autors trainiert wird. Skizzieren Sie, wie Sie unter Anwendung von N-Gram-Wahrscheinlichkeiten und Maximum-Likelihood-Schätzungen solch ein System trainieren würden. Zeigen Sie am folgenden Minitext, welches Wort so ein System am Ende des Textes (nach "in iSchools") vorhersagen würde für unterschiedliche Werte von N.

> "The iSchools organization has admitted three new member-schools to its consortium of leading Information Schools. Our newest members: the University of Chinese Academy of Sciences, Department of Library, Information and Archives Management, which has joined at the prestigious iCaucus level; the University of Regensburg, Institute for Information and Media, Language and Culture; and the University of São Paulo, School of Communication and Arts. The iSchools movement originated from a small group of schools in the United States; the iSchools organization how has 110 members worldwide. Researchers in iSchools"

Question 3.2 / Aufgabe 3.2: (5 marks / 5 Punkte)

What is the motivation for *smoothing* in statistical language processing? Outline the main advantage in comparison to a statistical approach that does not include smoothing. Use the text in Question 3.1 to illustrate your answer.

Was ist der Grund für die Anwendung des 'Smoothing' in statistischer Sprachverarbeitung? Was ist der entscheidende Vorteil gegenüber statistischer Sprachverarbeitung ohne 'Smoothing'? Beziehen Sie sich in Ihrer Antwort auf den Beispieltext in Aufgabe 3.1.

Question 3.3 / Aufgabe 3.3: (5 marks / 5 Punkte)

Discuss the implications of Zipf's law on the distribution of words in a document collection for applications such as that outlined in Question 3.1. Briefly discuss how increasing the corpus size might or might not address these implications.
Diskutieren Sie die Auswirkungen des Zipfschen Gesetzes auf die Verteilung von Wörtern in Dokumentensammlungen, wie sie bei Anwendungen wie der in Aufgabe 3.1 vorkommen. Diskutieren Sie kurz, ob die Vergrößerung der Dokumentensammlung die Implikationen der Zipfschen Verteilung zu lösen hilft.

Question 3.4 / Aufgabe 3.4: (5 marks / 5 Punkte)

Outline how you might go about evaluating the word processor described in Question 3.1.
Skizzieren Sie, wie Sie das in Aufgabe 3.1 beschriebene System evaluieren könnten.