

**Aufgabe 1)** Manuelles POS Tagging

**Punkte: 6**

1. Taggen Sie folgenden Textauszug mithilfe des Penn Treebank Tagsets<sup>1</sup>. Berücksichtigen Sie auch die erforderliche Tokenisierung. (2 Punkte)

My mama always said 'Life was like a box of chocolates; you never know what you're gonna get.'  
aus: *Forrest Gump*

2. Taggen Sie folgenden Textauszug mithilfe des Stuttgart-Tübingen-Tagsets<sup>2</sup>. Berücksichtigen Sie auch die erforderliche Tokenisierung. (2 Punkte)

Die Hummel hat eine Flügelfläche von 0,7 Quadratzentimeter, bei 1,2 Gramm Gewicht. Nach den bekannten Gesetzen der Aerodynamik ist es unmöglich, bei diesen Verhältnissen zu fliegen.  
*Arthur Lassen, Motivationstrainer*

3. Nennen Sie zwei Unterschiede im Tagging von Eigennamen zwischen STTS und Penn Treebank Tagset (2 Punkte).

---

<sup>1</sup><https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf>

<sup>2</sup><http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf>

**Aufgabe 2) HMM Tagging Simulation****Punkte: 14**

Gegeben ist das folgende deutsche Korpus. Groß- und Kleinschreibung haben wir ignoriert, sonst wäre es zu einfach. Getaggt wurde es mit dem Universal Tagset. Das Tagging enthält Punktuation. Nehmen Sie dennoch auch Satzanfangszustände und Satzendzustände an.

```
regenwuermer/NOUN regen/VERB sich/PRON im/ADP regen/NOUN ./PUNCT
im/ADP regen/ADJ Verkehr/NOUN regen/VERB sich/PRON menschen/NOUN auf/PART ./PUNCT
im/ADP november/NOUN folgt/VERB auf/ADP regen/NOUN regen/NOUN ?/PUNCT
```

**a) Trainieren des HMMs***Punkte: 4*

Berechnen Sie für das Korpus die Übergangs- und Emissionswahrscheinlichkeiten eines Hidden Markov Modells, welches die POS-Tag-Folge als unbeobachtete Zustände annimmt. Vergessen Sie bei den Übergangswahrscheinlichkeiten nicht die Satzanfänge und Satzenden. Smoothing ist nicht nötig.

**b) Tagging***Punkte: 10*

Gegeben ist nun folgender Testsatz:

im regen regen regen sich regenwuermer.

1. Ermitteln Sie die wahrscheinlichste POS-Tag Sequenz für den Testsatz durch Anwendung Ihres trainierten HMM-Modells. Benutzen Sie hierfür den Viterbi-Algorithmus und zeigen Sie Ihre Trellis. (Es gibt sehr viele Nullen, was das ganze stark vereinfacht). (7 Punkte)
2. Welche Wahrscheinlichkeit und welche Taggingsequenz weist Ihr Modell dem Satz zu? Ist die zugewiesene Sequenz richtig? (1 Punkt)
3. Wie würde ein greedy-Vorgehen den Satz taggen? Welche Entscheidungen sind dann richtig, welche falsch? (2 Punkte)

**Aufgabe 3)** Bonus: HMM Fragen

**Punkte: 4**

Ein weiteres geläufiges Beispiel, um HMMs zu erklären, besteht aus dem Zusammenhang von Wetter und Tätigkeiten. Dabei sind Wetterfeatures (**warm**, **regnerisch**, **kalt** ...) die (*hidden*) *states*, und Tätigkeiten (*kochen*, *schwimmen*, ...) die *observations/Beobachtungen*. Aus einer Beobachtungssequenz mit je einer Beobachtung pro Tag wie *kochen schwimmen schwimmen Fußball lesen Lego* soll man dann das Wetter an den (im Beispiel 6) Tagen dekodieren.

1. Welche Wahrscheinlichkeiten braucht man, um solch ein Modell zu trainieren? (2 Punkte)
2. In wiefern werden sich die Übergangswahrscheinlichkeiten in diesem Modell von denen für POS Tags unterscheiden? Begründen Sie Ihre Antwort linguistisch? (2 Punkte)