Übung 3

Aufgabe 3.1

1) Erstellen Sie ein Bigram-Modell analog zu Tabelle 3.2 aus [Jurafsky 2020] für den folgenden Korpus:
   <s> I am Sam </s>
   <s> Sam I am </s>
   <s> Sam I like </s>
   <s> Sam I do like </s>
   <s> do I like Sam </s>
2) Was ist die wahrscheinlichste Fortsetzung der folgenden Wortsequenzen?
   a. <s> Sam …
   b. <s> do I like …
   c. <s> Sam I am Sam …
   d. <s> Sam I do …
3) Welcher der folgenden Sätze ist "besser" (hat die größte Wahrscheinlichkeit gemäß dem Bigram-Modell)?
   a. <s> I am Sam </s>
   b. <s> Sam I do like </s>
   c. <s> Sam do I like </s>
   d. <s> I do like Sam I am </s>
4) Welches grundsätzliche Problem ergibt sich, wenn man die Wahrscheinlichkeiten für unterschiedlich lange Sätze berechnet?

Aufgabe 3.2

1) Berechnen Sie (bezogen auf den Korpus und das Bigram-Modell aus Aufgabe 3.1) die Perplexitäten folgender Sätze:
   a. <s> I am Sam </s>
   b. <s> Sam I do like </s>
   c. <s> Sam do I like </s>
   d. <s> I do like Sam I am </s>
   e. <s> I do like Sam I do like </s>
2) Welches Problem ergibt sich, wenn Sie die Perplexität von <s> like I do </s> berechnen wollen?

Übung 3

Aufgabe 3.3

Berechnen Sie für den Korpus aus Aufgabe 3.1 die Bigram-Wahrscheinlichkeiten, diesmal aber mit Laplace-Smoothing.

Berechnen Sie mit diesen Wahrscheinlichkeiten die Perplexitäten aus Aufgabe 3.2.

Hinweise:

- Da </s> als Folgetoken in Frage kommt, muss das Endymbol </s> in Formel 3.23 bei der Berechnung der Vokabulargröße V mit eingerechnet werden.
- Die Symbole <s> und </s> können nur am Anfang bzw. am Ende eines Biwords vorkommen, d.h. auch bei Laplace-Smoothing gilt $P_{Laplace}(<s>|x) = P_{Laplace}(x|</s>) = 0$.

Aufgabe 3.4

Nehmen Sie an, das Endsymbol </s> würde nicht verwendet werden. Trainieren Sie ein Bigram-Modell (ohne Smoothing) auf folgendem Korpus:

<s> a a
<s> a b
<s> b b
<s> b a

Zeigen Sie, dass dadurch keine gültige Wahrscheinlichkeitsverteilung für alle Sequenzen über dem Vokabular {a, b} definiert ist.

Tipp: Was muss für die Summe der Wahrscheinlichkeiten aller Sequenzen der Form <s>{a,b}* gelten? Zeigen Sie, dass die Summe der Wahrscheinlichkeiten aller 2-Wort-Sequenzen 1 ist, ebenso wie die Summe der Wahrscheinlichkeiten aller 3-Wort-Sequenzen.