Machine Learning
Exercises: language models (n-grams)

Laura Kallmeyer

Summer 2016, Heinrich-Heine-Universität Düsseldorf

Exercise 1 *Consider the following toy example (similar to the one from Jurafsky & Martin (2015)):*

Training data:

```
<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>
```

Assume that we use a bigram language model based on the above training data.

1. *What is the most probable next word predicted by the model for the following word sequences?*

   (1) `<s>` *Sam ...*

   (2) `<s>` *Sam I do ...*

   (3) `<s>` *Sam I am Sam ...*

   (4) `<s>` *do I like ...*

2. *Which of the following sentences is better, i.e., gets a higher probability with this model?*

   (5) `<s>` *Sam I do I like* `</s>`

   (6) `<s>` *Sam I am* `</s>`

   (7) `<s>` *I do like Sam I am* `</s>`

Exercise 2 *Consider again the same training data and the same bigram model. Compute the perplexity of*

<s> *I do like Sam*

Exercise 3 *Take again the same training data. This time, we use a bigram LM with Laplace smoothing.*

1. *Give the following bigram probabilities estimated by this model:*

   P(do|<s>) P(do|Sam) P(Sam|<s>) P(Sam|do)
   P(I|Sam) P(I|do) P(like|I)

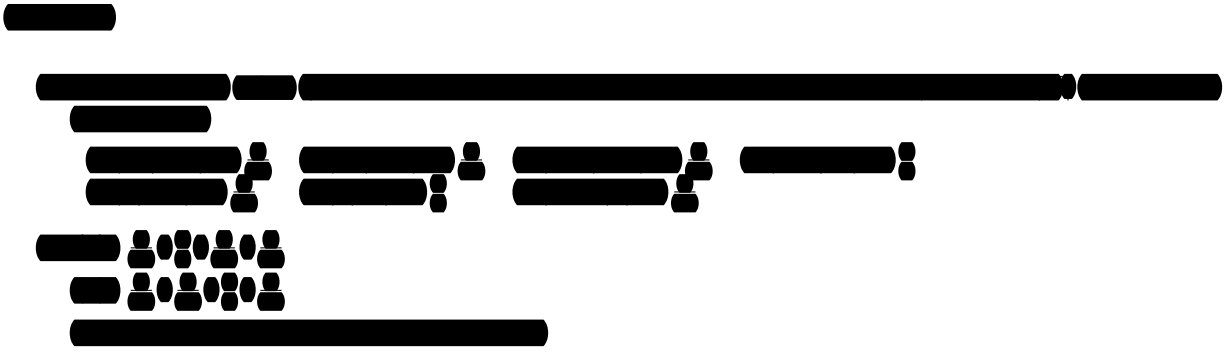
   Note that for each word w_{n-1}, we count an additional bigram for each possible continuation w_n. Consequently, we have to take the words into consideration and also the symbol </s>*.*

2. *Calculate the probabilities of the following sequences according to this model:*

   (8) <s> *do Sam I like*

   (9) <s> *Sam do I like*

   Which of the two sequences is more probable according to our LM?



References

Jurafsky, Daniel & James H. Martin. 2015. Speech and language processing. an introduction to natural language processing, computational linguistics, and speech recognition. Draft of the 3rd edition.