Prof. Dr. Andreas Hotho,
M.Sc. Janna Omeliyanenko
Lecture Chair X for Data Science, Universität Würzburg

9. Exercise for “Sprachverarbeitung und Text Mining”

21.01.2022

1 Knowledge Questions

1. Describe the idea of an N-gram language model in your own words.
2. Give 3 possible applications of a language model in the field of NLP.
3. The Brown corpus has long been used as a representative corpus for the English language. The Google Crawl presented in the lecture is much larger, and contains about 13,500,000 word forms, as opposed to about 300,000 in the Brown corpus. Thus, it also contains many more word forms than English dictionaries. Explain how this phenomenon can occur.
4. In the lecture, Shannon’s method for generating text was introduced. Describe how resulting sentences change when using more and more complex language models (i.e. first unigrams, then bigrams, etc...).
   Are more complex language models always better? What role does the training corpus play in this?
5. Explain the concepts of extrinsic evaluation and intrinsic evaluation based on the example of language models. How do both concepts differ?
6. Justify why it is unlikely that there will ever be a corpus large enough to contain reliable counts of rare N-grams. How can language models deal with very rarely occurring or even unknown words? What possible solutions were shown in the lecture?

2 Smoothing

Given are the following sentences:

```
beer is a difficult lecture
to brew beer is difficult
i brew beer in the lecture
in the lecture i drink beer
the lecture is difficult
```

Note: In addition to the words in the sentence, consider the sentence-start token `<s>`. A sentence ending token should not be used in this task.

a) First, create an overview table with the counts of all bigrams that occur.

b) Using the created counts, calculate the probability of the following sentence each with unigrams and bigrams (calculate unigrams without sentence-start tokens): `lecture is a difficult beer`

c) Calculate the perplexity for the sentence given above, for bigrams and unigrams.

d) Smooth the bigram counts with Laplace-smoothing and calculate the probability and perplexity of the above sentence again. It is sufficient to smooth only the bigrams that are used in the sentence.

e) Smooth the bigrams with Good-Turing-smoothing and calculate the probability and perplexity of the sentence again. It is sufficient to smooth only the bigrams that are used in the sentence.

3 Shannons Method for Text Generation

In the lecture, Shannon's method for text generation was presented using Shakespeare as an example.
Download the file
`Lewis_Carroll_Alices_Adventures_in_Wonderland.txt`
from WueCampus and determine the unigram, bigram and trigram counts of this novel.
Generate one sentence each for unigrams, bigrams, and trigrams using Shannon's method.
Begin with:

- <s>
- She
- She had

Always choose the N-gram with the highest probability.
Note: Treat punctuation marks as separate tokens.

`http://guidetodatamining.com/ngramAnalyzer/`