NPFL124 Introduction to Natural Language Processing

Practicals: Probability in NLP

Pavel Pecina & others

📅 Feubruary 25, 2026

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Probability in NLP

Why?

- Clean formalization of models of a natural language

What?

- Probability of a word (unigram)
- Probability of a word sequence (bigram, trigram, …)
- Bigram probability: joint / conditional

How?

- Maximum likelihood estimation: $p(w) = c(w)/T$
- The best estimation …

Task 1: My first corpus

Create a text corpus from data available on the Web.

Format

- Plain text in UTF-8

Size

- Minimum 500K words

Possible sources

- https://en.wikipedia.org/
- https://www.gutenberg.org/

Save it as one file corpus.txt, e.g.:

```
wget http://www.gutenberg.org/files/135/135-0.txt -O corpus.txt
```

Task 2: Unigram probabilities

1. Guess 5 most frequent words (unigrams) in English
2. Obtain frequency list of all unigrams from your corpus
   - Show 5 most frequent unigrams
   - Show 5 least frequent unigrams
   - Sample code: `http://ufal.mff.cuni.cz/~pecina/NLP/count1.py`
3. Estimate unigram probabilities (i.e. relative frequencies)
   - Show 5 most probable unigrams
   - Show 5 least probable unigrams
   - Sample code: `http://ufal.mff.cuni.cz/~pecina/NLP/prob1.py`

Task 3: Bigram probabilities

1. Guess 5 most frequent word pairs (bigrams) in English
2. Obtain frequency list of all bigrams from your corpus
   - Show 5 most/least frequent bigrams
   - Sample code: `http://ufal.mff.cuni.cz/~pecina/NLP/count2.py`
3. Estimate (joint) bigram probabilities (i.e. relative frequencies)
   - Show 5 most/least probable bigrams
   - Sample code: `http://ufal.mff.cuni.cz/~pecina/NLP/prob2.py`
4. Estimate conditional bigram probabilities
   - Show 5 most/least (cond.) probable bigrams
   - Sample code: `http://ufal.mff.cuni.cz/~pecina/NLP/probcond.py`