Assignment 5,6 + Smoothing 2

(SNLP Tutorial 6)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

1st, 2nd June 2021

Assignment 5

- Exercise 1: OOV Words
- Exercise 2: Additive smoothing
- Exercise 3: Perplexity, infinite smoothing, interpolation
- Bonus: Other language models

Good-Turing

Data: 🍏 🍏 🍏 🍆 🍏 🍌 🍌 🍒 🍏 🍆 🍌 🍌 🍒 🍆 🍇 🌿

Good-Turing

Data: 🍏 🍏 🍏 🍆 🍏 🍌 🍌 🍒 🍏 🍆 🍌 🍌 🍒 🍆 🍇 🌿

- $N_4 = \{$🍌$\}$
- $N_3 = \{$🍏, 🍆$\}$
- $N_2 = \{$🍒$\}$
- $N_1 = \{$🍇, 🌿$\}$
- $N_0 = \{$🍨$\}$

Good-Turing

Data: 🍏 🍏 🍏 🍆 🍏 🍌 🍌 🍒 🍏 🍆 🍌 🍌 🍒 🍆 🍇 🌿

- $N_4 = \{$🍌$\}$
- $N_3 = \{$🍏, 🍆$\}$
- $N_2 = \{$🍒$\}$
- $N_1 = \{$🍇, 🌿$\}$
- $N_0 = \{$🍨$\}$

$$p_r = \frac{(r+1)N_{r+1}}{N_r} \cdot \frac{1}{N}$$

Good-Turing

Data: 🍏 🍏 🍏 🍆 🍏 🍌 🍌 🍒 🍏 🍆 🍌 🍌 🍒 🍆 🍇 🌿

- $N_4 = \{$🍌$\}$
- $N_3 = \{$🍏, 🍆$\}$
- $N_2 = \{$🍒$\}$
- $N_1 = \{$🍇, 🌿$\}$
- $N_0 = \{$🍨$\}$

$$p_r = \frac{(r+1)N_{r+1}}{N_r} \cdot \frac{1}{N}$$

- Nominator: expected total number of occurences of words that occur $r+1$ times
- Denominator-left: previous bucket size
- Fraction-left: expected number of occurences of a single word from that bucket
- Denominator-right: divide by total occurences

Good-Turing - Questions

- Let k be the maximum occurence of a word. What's the issue?

Good-Turing - Questions

- Let k be the maximum occurence of a word. What's the issue?
- A similar issue related to the one above?

Good-Turing - Questions

- Let k be the maximum occurence of a word. What's the issue?
- A similar issue related to the one above?
- Do the probabilities sum up to 1?

Good-Turing - Questions

- Let k be the maximum occurence of a word. What's the issue?
- A similar issue related to the one above?
- Do the probabilities sum up to 1?
- How to make it work for anything above unigrams?

Assignment 6

- TODO

Resources

1. UdS SNLP Class: https://teaching.lsv.uni-saarland.de/snlp/
2. Twitter emojis