Assignment 1 + Language Properties

(SNLP tutorial 2)

Vilém Zouhar, Awantee Deshpande, Julius Steuer

May 2, 2021

Organisational Issues

- Teammates
- Assignment submissions
- ▸ Naming your assignment folder: Name1_id1_Name1_id2
- ▸ Your Notebooks and files should be directly inside the main folder (no unnecessary nesting)
- ▸ Do not submit the following files

__pycache__
.ipynb_checkpoints
data folder/files
any other pdf or information file accompanying the assignment

- ▸ Only submit: Notebook + Python files! Otherwise we could deduct your marks.

Part 1: Discussion of Assignment 1

- Exercise 1: Instructions for setup
- Exercise 2: Stick breaking
- Exercise 3: Zipf's Law at word level
- Bonus: Zipf's Law at character level

Part 2: Overview of current topics

- Basics of Probability Theory
- Perplexity
- Maximum Likelihood Estimation
- Smoothing

Probability Theory for language models

Predict

$P(w_i|h_i)$ or $P(w_1, w_2...w_N)$

Compare for uniform,unigram, bigram, trigram... ngram models. Where do we assume statistical independence?

Probability Theory for language models

Predict

$P(w_i|h_i)$ or $P(w_1, w_2 ... w_N)$

Compare for uniform,unigram, bigram, trigram... ngram models. Where do we assume statistical independence?

Expectation value

$$E[f(V)] = \sum_{w_i \in V} p(w_i) f(w_i)$$

For LMs,

$$H = E[-log(p(w_i))] = -\sum_{w_i \in V} p(w_i) log(p(w_i))$$

What does this mean? What are we capturing by the entropy of the LM?

Perplexity

Formulae

$$PP = 2^{\frac{1}{n}\sum_1^n \log p(w_i|w_{i-1})}$$

$$PP = 2^{-\sum_{w,h} f(w,h)\log_2 P(w|h)}$$

How do these two formulae relate to each other?

TODO

Maximum Likelihood Estimation

TODO

- A way to estimate language model (distribution) parameters
- Trying to maximize probability of the training data

LM Smoothing

TODO

- Q: What happens if an unknown token is encountered and LM assigns it 0 probability?

Different smoothing methods will be covered in the further chapters.

Homework

- Exercise 1: Perplexity calculation by han
- Exercise 2: Plotting n-gram distributions
- Exercise 3: MLE language models, smoothing, perplexity calculation
- Bonus: Custom alternative to perplexity

Resources

1. TODO