Hide menu

Lecture: Autocomplete

- Video: Week Introduction 1 min
- Video: N-Grams: Overview 3 min
- Reading: N-Grams Overview 5 min
- Video: N-grams and Probabilities 7 min
- Reading: N-grams and Probabilities 10 min
- Video: Sequence Probabilities 5 min
- Reading: Sequence Probabilities 6 min
- Video: Starting and Ending Sentences 8 min
- Reading: Starting and Ending Sentences 6 min
- Lab: Lecture notebook: Corpus preprocessing for N-grams 1h
- Video: The N-gram Language Model 6 min
- Reading: The N-gram Language Model 10 min
- Video: Language Model Evaluation 6 min
- Lab: Lecture notebook: Building the language model 1h
- Reading: Language Model Evaluation 10 min
- Video: Out of Vocabulary Words 4 min
- Reading: Out of Vocabulary Words 10 min
- Video: Smoothing 6 min
- Reading: Smoothing 10 min
- Lab: Lecture notebook: Language model generalization 1h
- Video: Week Summary 1 min
- Reading: Week Summary 2 min
- Video: Week Conclusion 46 sec

Lecture Notes (Optional)

Language Model Evaluation

Splitting the Data

We will now discuss the train/val/test splits and perplexity.

Train/Val/Test splits

Smaller Corpora:

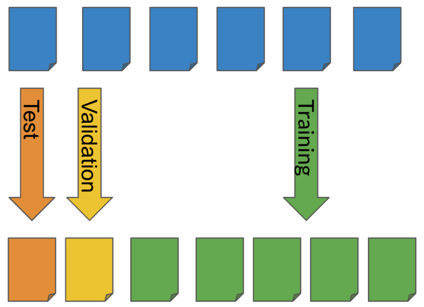
- 80% train
- 10% val
- 10% test

Larger Corpora:

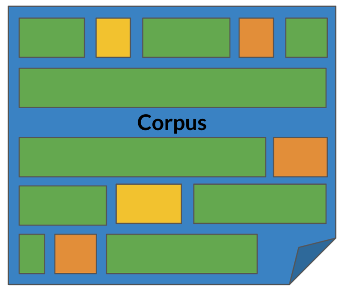
- 98% train
- 1% val
- 1% test

There are two main methods for splitting the data:

- Continuous text



- Random short sequences



Perplexity

Perplexity is used to tell us whether a set of sentences look like they were written by humans rather than by a simple program choosing words at random. A text that is written by humans is more likely to have lower perplexity, where a text generated by random word choice would have a higher perplexity.

Concretely, here are the formulas to calculate perplexity.

$$PP(W) = P(s1, s2, \ldots, sm)^{-\frac{1}{m}}$$

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \prod_{j=1}^{|si|} \frac{1}{P(wj^{(i)} \mid wj-1^{(i)})}}$$

$wj^{(i)} \rightarrow$ j corresponds to the jth word in the ith sentence. If you were to concatenate all the sentences then wi is the ith word in the test set. To compute the log perplexity, you go from

$$PP(W) = \sqrt[m]{\prod_{i=1}^{m} \frac{1}{P(wi \mid wi-1)}}$$

To

$$\log PP(W) = -\frac{1}{m} \sum_{i=1}^{m} \log 2\,(P(wi \mid wi-1))$$