

Hide menu

Lecture: Autocomplete

- Video: Week Introduction1 min
- Video: N-Grams: Overview3 min
- Reading: N-Grams Overview5 min
- Video: N-grams and Probabilities7 min

Reading: N-grams and Probabilities10 min

Video: Sequence Probabilities5 min

Reading: Sequence Probabilities6 min

Video: Starting and Ending Sentences8 min

Reading: Starting and Ending Sentences6 min

Lab: Lecture notebook: Corpus preprocessing for N-grams1h

Video: The N-gram Language Model6 min

Reading: The N-gram Language Model10 min

Video: Language Model Evaluation6 min

Lab: Lecture notebook: Building the language model1h

Reading: Language Model Evaluation10 min

Video: Out of Vocabulary Words4 min

Reading: Out of Vocabulary Words10 min

Video: Smoothing6 min

Reading: Smoothing10 min

Lab: Lecture notebook: Language model generalization1h

Video: Week Summary1 min

Reading: Week Summary2 min

Video: Week Conclusion46 sec

Lecture Notes (Optional)

> Week 3 > N-grams and Probabilities

< PreviousNext >

An N-gram is a sequence of N words

Corpus: I am happy because I am learning

Unigrams: { I , am , happy , because , learning }

Bigrams: { I am , am happy , happy because ... } I happy

Trigrams: { I am happy , am happy because, ... }

Now given the those definitions, we can label a sentence as follows:

Corpus: This is great ... teacher drinks tea. m = 500
w1 w2 w3 w498 w499 w500

In other notation you can write:

- w1^m = w1w2w3...wm
- w1^3 = w1w2w3
- wm-2^m = wm-2wm-1wm

Given the following corpus: I am happy because I am learning.

- Size of corpus m = 7.
- P(I) = 2/7
- P(happy) = 1/7

To generalize, the probability of a unigram is P(w) = C(w)/m

Bigram Probability:

Corpus: I am happy because I am learning

$$P(am|I) = \frac{C(I\ am)}{C(I)} = \frac{2}{2} = 1$$
$$P(happy|I) = \frac{C(I\ happy)}{C(I)} = \frac{0}{2} = 0$$
 I happy

$$P(learning|am) = \frac{C(am\ learning)}{C(am)} = \frac{1}{2}$$

Probability of a bigram:
$$P(y|x) = \frac{C(x\ y)}{\sum_w C(x\ w)} = \frac{C(x\ y)}{C(x)}$$

Trigram Probability:

To compute the probability of a trigram:

- $P(w_3 | w_1^2) = \frac{C(w_1^2 w_3)}{C(w_1^2)}$
- $C(w_1^2 w_3) = C(w_1 w_2 w_3) = C(w_1^3)$

N-gram Probability:

$$C(w_1^{N-1} w_N)$$

