

National University of Singapore
School of Computing
CS3245: Information Retrieval
Tutorial 1

Ngram Language Models

Readings: IIR, Sections 12.1 – 12.2

1. **‘80s Pop songs.** Let’s practice the math in ngram language models again, but this time on song titles.

- (a) Generate a probabilistic unigram language model of word tokens for the following two pop song artists. Apply add-one smoothing over the entire vocabulary.

Artist	Song Title	Artist	Song Title
Phil Collins	I Don’t Want To Go	Air Supply	All Out Of Love
Phil Collins	A Groovy Kind Of Love	Air Supply	Here I Am
Phil Collins	You Can’t Hurry Love	Air Supply	I Remember Love
Phil Collins	This Must Be Love	Air Supply	Love Is All
Phil Collins	Take Me With You	Air Supply	Don’t Tell Me

- (b) Do the same as the above but for a bigram language model of tokens. Be sure to use special <START> and <END> tokens to delimit tokens that begin and end at the document boundaries. Apply add-one smoothing as above.
 - (c) Predict which artist would most likely be responsible for songs entitled 1) “I Remember You” and 2) “Don’t Want You To Love Me” and 3) “Goodbye”. Show your math.
2. **Homework 1 revisited.** In our Homework 1, we asked you to create a 4-gram model of strings in different local languages. However, in the math to calculate the probability in the model, we treated these 4-grams as entirely independent tokens, in effect treating them as a unigram model, where context plays no role.
 - (a) In lecture we stated that ngram models grow exponentially in complexity as n gets larger, and we also stated that in practice, many researchers restrict their ngram models to $n = 3$ or lower. However in our assignment we set $n = 4$. Why isn’t this a problem?
 - (b) Describe how a true 4-gram model would calculate its probability. Remember, a ngram model with $n = k$ uses $k - 1$ previous elements to predict the k^{th} element.
 - (c) If we chose to use a true 4-gram model, how would this affect the probability computation? Give a concrete example in which the computation would change.
 - (d) Give your opinion about which of the two models (treating 4-gram letter sequences as unigrams or as 4-grams) would be better to follow.
 3. **Smoothing.** We’ve covered the idea of *smoothing* in class, where we assign probabilities to unseen items by effectively subtracting some probability mass from items that we did observe. It’s also known as *discounting*, as some actual probability mass is discounted from actual observed events.

We mentioned that add-one smoothing is a simple introduction to smoothing but isn’t used in practice. In particular, why should it be add one instead of some other constant k ?

- (a) Describe what happens when add-one smoothing is applied on an LM constructed from a small number of observations, but which has a large vocabulary space. Is add-one smoothing appropriate in this case?
- (b) Describe what happens when add-one smoothing is applied on an LM constructed from a large number of observations, and which has a small vocabulary space. Is add-one smoothing appropriate in this case?
- (c) What does add-one or add- k smoothing represent in terms of our belief in the distribution of occurrences of events?