CS3245

Information Retrieval

AY2022/23 Semester 2

Notes by Jonathan Tay

Last updated on January 15, 2023

Contents

1 Language Models

A **language model** is a grammarless, computational model created from collections of text.

They are used to assign scores (e.g. probabilities) to a sequence of words.

unigram model

Create a **frequency table** of all tokens (words) that appear in the collection.

Unigram models have insufficient context to model the order of words in a sentence.

***n*-gram model**

By remembering sequences of n tokens we can predict the n-th token given only the previous $n - 1$ tokens as context (**Markov assumption**).

A unigram model is a 1-gram model, bigram model is a 2-gram model, etc.

However, n-gram models require exponentially more space as n increases.

The **count** of an input is the *sum* of the counts of all tokens in the input, while the **probability** of an input is the *product* of the probabilities of all tokens in the input.

However, if a token does not appear in the collection, its probability is 0, resulting in a probability of 0 for the entire input, which is undesirable.

1-smoothing is a technique to avoid this problem. It adds a count of 1 to every token in the collection, even if it does not appear in the input.