Mathematical Derivation of N-gram Approach

Page

Language Models:-

The goal of probabilistic Language Modelling is to calculate the probability of sentence of sequence of words.
$P(W) = P(w_1, w_2, \ldots, w_n)$

This can be used to find the next word in the sequence.
$P(w_5 | w_1, w_2, w_3, w_4)$.

General Rule for $P(A|B) = \frac{P(A \cap B)}{P(B)}$

[Probability of A given B]

Generally the formula is:-
$P(x_1, x_2, \ldots, x_n) = P(x_1) P(x_2|x_1) \ldots P(x_n | x_1, \ldots, x_{n-1})$

$$\Rightarrow P(x_1, x_2, \ldots, x_n) = \prod_i P(x_i | x_1, \ldots, x_{i-1})$$

S = "the office is so awesome"

$P(*S) = P(\text{the}) * P(\text{office} | \text{the}) * P(\text{is} | \text{the office})$

Mathematical Derivation of N-gram Approach

Page

(*) So, a stochastic process has the Markov property if the conditional probability distribution of future states of the process depends only upon the present state, not on the sequence of events that preceded it.
A process with such property is known as Markov process.

In simpler terms the probability of next word can be estimated given only the previous K words
Let K=1,
P(awesome | the office is so) ≈ P(awesome | so)
if K=2

P(awesome | the office is so) ≈ P(awesome | is so)

∴ Basically we are estimating that what is the probability that next word is "awesome" given previous words "is so".

Mathematical Derivation of N-gram Approach

From Markov assumptions we can finally have the N-Gram Models

$$P(w_i | w_1, \ldots, w_{i-1}) \approx P(w_i | w_{i-(n-1)} \cdots w_{i-1})$$

Unigram Model ($k=1$)

$$P(w_1 w_2 \cdots w_n) \approx \prod_i P(w_i)$$

Bigram Model ($k=2$)

$$P(w_i | w_1 w_2 \cdots w_{i-1}) \approx P(w_i | w_{i-1})$$

Estimating Bigram Prob by MLE

$$P(w_i | w_{i-1}) = \frac{\text{Count}(w_{i-1}, w_i)}{\text{Count}(w_{i-1})}$$

We can extend the equations for 4, 5, 6, ... gram and so on.

We will use this approach to build a Language Model.