1 Language Model

We begin with our language model, $P(w)$
Using the "Naive assumption" that our variables are independent then, $P(A, B, C) = P(A)P(B)P(C)$
. Given a document D,

$$P(D) = \prod_i P(w_i) = \prod_w P(w)^{N_w}$$

Here $P(w)^{N_w}$ is the count of tokens, where w is the word in the vocabulary of the document, and N_w is the frequency of a word w in a document D.

2 Bayes Theorem

However we must remember Bayes theorem when using probabilities. Bayes theorem states the following

$$P(C_i|D) = \frac{P(C_i)P(D|C_i)}{P(D)}$$

When calculating probabilities your output can become very small. Taking products often leads to underflow. The maximum number of words before underflow is around 30

3 Log Probabilities

To combat this underflow we take the log of the product of probabilities.

$$\log P(D) = \log \prod_w P(w)^{N_w} = \sum_w \log N_w P(w)$$

However, we must still remember to offset our data in case the probability of a word goes to zero, becauselog(0) is undefined.

4 Additional Notes

The biggest sin is testing on your training data. Figure out why they are getting misclassified? what are the words in romeo and julie that make it look more like a comedy?