Smoothing

What do we do with words that are in our vocabulary but appear in a test set in an unseen context (for example they appear after a word they never appeared after in training) ?

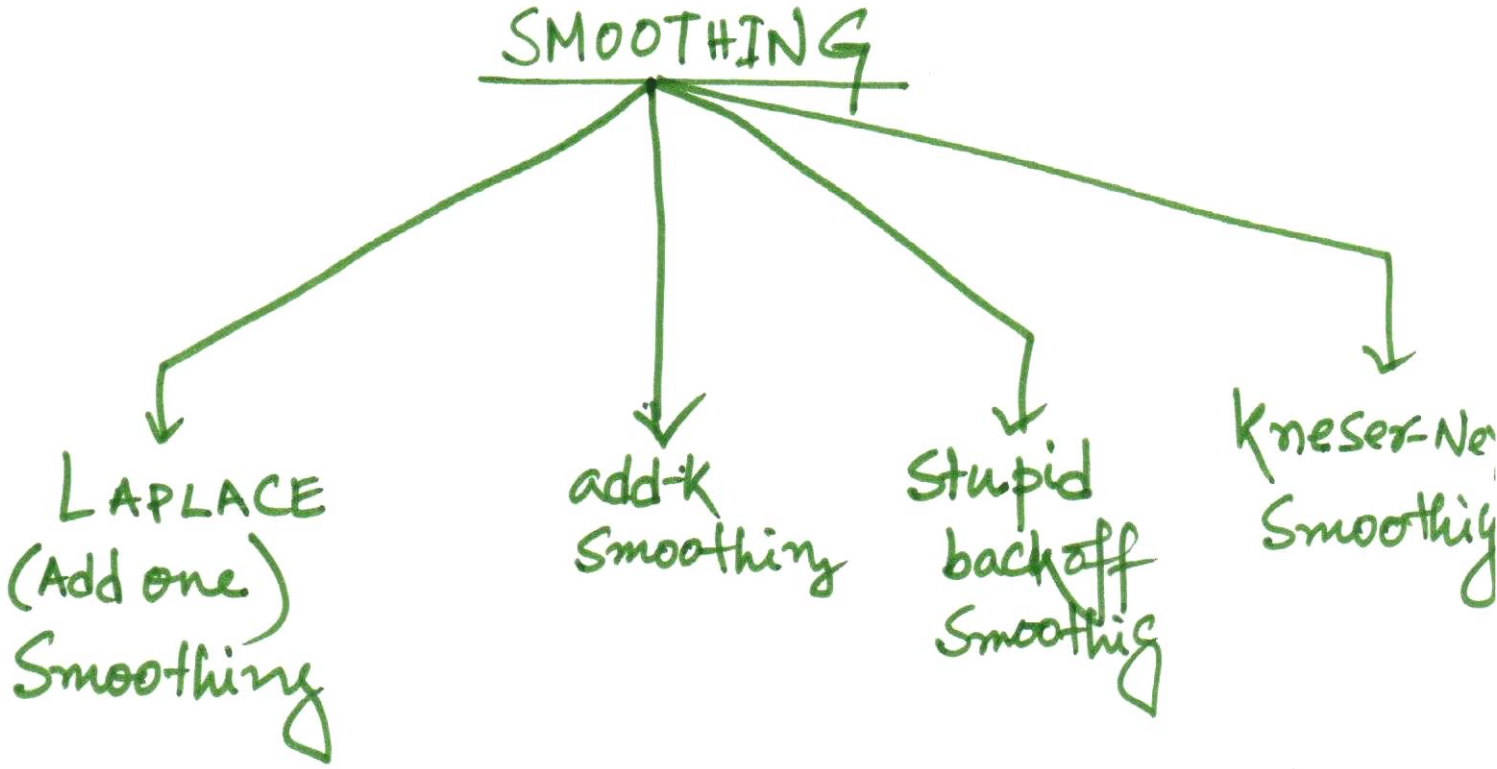
— In this case Zero probility to these unseen event

for example

Given Corpus:

<s> I am Henry </s>
<s> I like College </s>
<s> Do Henry Like College </s>
<s> Henry I am </s>
<s> Do I like Henry </s>
<s> Do I like College </s>
<s> Do I like Henry </s>

Calculate (Big
<s> I like a Colleg
</s>


SMOOTHING
LAPLACE (Add one) Smoothing
add-k Smoothing
Stupid backoff Smoothig
Kneser-Ne Smoothig

<s> I like a College </s>

$$\equiv P(\langle s\rangle) \times P(I \mid \langle s\rangle) \times P(a \mid like) \times P(college \mid a) \times P(\langle /s\rangle \mid college)$$

$$\equiv \quad ?$$

LAPLACE SMOTHING

- The simplest way to do smoothing is to add one to all the n-gram counts, before we normalize them into probalilities
- All the counts that used to be zero will now have a count of 1, the counts of 1 will be 2 and so on. This algorithm is called Laplace Smoothing

$$P_{Laplace}(w_i) = \frac{C_i + 1}{N + V}$$

V = words in Vocabulary