HW1

Jackson Crook

February 2016

(1.1)
true
false
true

(1.2)
document = [
'.', 'the', 'dog', 'bit', 'the', 'man', '.',
'.', 'the', 'dog', 'ate', 'the', 'cheese', '.',
'.', 'the', 'mouse', 'bit', 'the', 'cheese', '.',
'.', 'the', 'mouse', 'drank', 'coffee', '.',
'.', 'the', 'man', 'drank', 'tea', '.'
]

Unigram FreqDist:

- cheese 2
- coffee 1
- ate 1
- tea 1
- dog 2
- . 10 (I'm using a period as the pad symbol)
- drank 2
- the 8
- mouse 2
- bit 2
- man 2

$\mid W \mid = 11$

$n_o(d) = 33$

$\tilde{\theta} = \frac{n_w(d)+1}{n_o(d)+1(|W|)}$

$\tilde{\theta}_{the} = \frac{8+1}{11+33} = 9/44$

$\tilde{\theta}_{mouse} = \frac{2+1}{44} = 3/44$

$\tilde{\theta}_{ate} = \frac{1+1}{44} = 2/44$

$\tilde{\theta}_{.} = \frac{10+1}{44} = 11/44$

$\tilde{\theta}_{*U*} = \frac{0+1}{44} = 1/44$

$$n_{.\ the}(d) = 5$$
$$n_{the\ mouse}(d) = 2$$
$$n_{mouse\ ate}(d) = 0$$
$$n_{ate\ the}(d) = 1$$
$$n_{the\ potato}(d) = 0$$
$$n_{potato\ .}(d) = 0$$

$$\tilde{\Theta}_{.\ the} = \frac{5 + 10\frac{9}{44}}{9+10}$$
$$\tilde{\Theta}_{the\ mouse} = \frac{2 + 10\frac{3}{44}}{8+10}$$
$$\tilde{\Theta}_{mouse\ ate} = \frac{0 + 10\frac{2}{44}}{2+10}$$
$$\tilde{\Theta}_{ate\ the} = \frac{1 + 10\frac{9}{44}}{1+10}$$
$$\tilde{\Theta}_{the\ potato} = \frac{0 + 10\frac{1}{44}}{8+10}$$
$$\tilde{\Theta}_{potato\ .} = \frac{0 + 10\frac{11}{44}}{0+10}$$

Multiply all those together, take the negative logarithm, and you get 13.21

Now for KN:

$k_o(d) = 19$

w	k$_w(d)$	K$_{w'}$
mouse	3	$\frac{3+1}{19+11}$
ate	1	$\frac{1+1}{30}$
the	1	$\frac{1+1}{30}$
.	5	$\frac{5+1}{30}$
$*U*$	0	$\frac{0+1}{30}$

So:

$$\bar{\Theta}_{.\ the} = \frac{5 + 10\frac{4}{30}}{9 + 10}$$

$$\bar{\Theta}_{the\ mouse} = \frac{2 + 10\frac{2}{30}}{8 + 10}$$

$$\bar{\Theta}_{mouse\ ate} = \frac{0 + 10\frac{2}{30}}{2 + 10}$$

$$\bar{\Theta}_{ate\ the} = \frac{1 + 10\frac{4}{30}}{1 + 10}$$

$$\bar{\Theta}_{the\ potato} = \frac{0 + 10\frac{1}{30}}{8 + 10}$$

$$\bar{\Theta}_{potato\ .} = \frac{0 + 10\frac{6}{30}}{0 + 10}$$

multiply all those together, take the negative log, and you get 13.05

(1.3)
Aside from the fact that there will be one less bi-gram in the corpus, contextual dependencies are weaker between words separated by another word than between adjacent words.

(1.4)
Additive smoothing:

$$\tilde{\Theta}_{ww'w''} = \frac{n_{ww'w''}(d) + \beta\tilde{\theta}_{w''}}{n_{ww'o}(d) + \beta}$$

KN smoothing:

$$\bar{\Theta}_{ww'w''} = \frac{n_{ww'w''}(d) + \beta K_{w''}}{n_{ww'o}(d) + \beta}$$

where $K_w = \frac{k_w(d)}{k_o(d)}$ and $k_w(d)$ is the number of unique bi-grams that precede w in d.

(1.5)
Part1
$\sum_{w'} \frac{n_{ww'}(d)}{n_{wo}(d)}$ equals 1 by definition, because $n_{wo} = \sum_{w'} n_{ww'}$

$\sum_{w'} \tilde{\theta}_{w'}$ also equals 1 by definition, because $\tilde{\theta}$ is a probability distribution. Therefore, using the distributive property of summation,

$$\sum_{w'} \lambda \frac{n_{ww'}(d)}{n_{wo}(d)} + (1-\lambda)\tilde{\theta}_{w'} = \lambda + (1-\lambda) = 1$$

Part2
Setting lambda equal to zero yields an unsmoothed unigram model, while setting lambda to one yields an unsmoothed bigram model. So higher lambdas mean more reliance on the bigram distribution. Because there are nearly always going to be more possible bigrams from a set of possible words than there will be unigrams (because of the many combinations of those unigrams), much more data is required to get a sample size large enough to inform an accurate distribution over bigrams. So, with more training data, the bigram-based compononent of this model will improve, and the optimal lambda will increase.