CS159 - Absolute Discount Smoothing Handout

David Kauchak - Fall 2020

To help understand the absolute discounting computation, below is a walkthrough of the probability calculations on as *very* small corpus.

Given the following corpus (where we only have one letter words):

```
a a a b
a b b a
c a a a
```

We would like to calculate an absolute discounted model with D = 0.5. We'll ignore the begin and end sentence tokens, assume that our vocabulary is all three "words" and not worry about handling out of vocabulary words (i.e. `<UNK>`).

We first calculate the unigram MLE probabilities as:

	MLE prob
$p(a)$	8/12
$p(b)$	3/12
$p(c)$	1/12

and the bigram MLE probabilities as:

	MLE prob
$p(a\|a)$	4/6
$p(b\|a)$	2/6
$p(c\|a)$	0
$p(a\|b)$	1/2
$p(b\|b)$	1/2
$p(c\|b)$	0
$p(a\|c)$	1
$p(b\|c)$	0
$p(c\|c)$	0

Using this information, we can now calculate the reserved mass and the αs for each of the words in our vocabulary (i.e. things that we're conditioning on):

- a

  The numerator for our α is the reserved mass:

$$reserved\_mass(a) = (2 * 0.5)/6 = 1/6$$

and the denominator is:

$$1 - \sum_{x:count(ax)>0} p(x) = 1 - (p(a) + p(b)) = 1 - (8/12 + 3/12) = 1/12$$

giving us an α of:

$$\alpha(a) = \frac{1/6}{1/12} = 2$$

- b

  $reserved\_mass(b) = (2 * 0.5)/2 = 1/2$

  $1 - \sum_{x:count(bx)>0} p(x) = 1 - (p(a) + p(b)) = 1 - (8/12 + 3/12) = 1/12$

  $\alpha(b) = \frac{1/2}{1/12} = 6$

- c

  $reserved\_mass(c) = (1 * 0.5)/1 = 1/2$

  $1 - \sum_{x:count(cx)>0} p(x) = 1 - p(a) = 1 - (8/12) = 4/12 = 1/3$

  $\alpha(c) = \frac{1/2}{1/3} = 3/2$

Finally, now that we have the αs, we can calculate the smoothed bigram probabilities. For those that occurred, we simply discount the count. For those that did not occur, we calculate the probability as *alpha* times the unigram probability of the word.

	eqn	prob
$p(a\|a)$	(4 - 0.5)/6	3.5/6
$p(b\|a)$	(2 - 0.5)/6	1.5/6
$p(c\|a)$	2 * 1/12	1/6
$p(a\|b)$	(1-.05)/2	1/4
$p(b\|b)$	(1-0.5)/2	1/4
$p(c\|b)$	6 * 1/12	1/2
$p(a\|c)$	(1-0.5)/1	1/2
$p(b\|c)$	3/2 * 3/12	3/8
$p(c\|c)$	3/2 * 1/12	1/8

Notice that after the smoothing, the three distributions all still sum to 1. In this case discounting by 0.5 may be a bit aggressive.

Note, for the first two distributions ($p(\cdot|a)$ and $p(\cdot|b)$), we actually didn't need to go through the exercise of calculating α since there was only one "word" we were backing off to and it therefore would get all of the reserved mass. However, I included the computation here so you can see that the math still works out.