CS159 - Assignment 2a: Solution

1. Unigram probabilities:

	MLE prob
$p(a)$	6/12
$p(b)$	5/12
$p(c)$	1/12

2. The bigram probabilities:

	MLE prob
$p(a\|a)$	1/5
$p(b\|a)$	3/5
$p(c\|a)$	1/5
$p(a\|b)$	2/3
$p(b\|b)$	1/3
$p(c\|b)$	0
$p(a\|c)$	1/1
$p(b\|c)$	0
$p(c\|c)$	0

3. Interpolated bigram probabilities with $\lambda = 1$:

	MLE prob
$p(a\|a)$	2/8
$p(b\|a)$	4/8
$p(c\|a)$	2/8
$p(a\|b)$	3/6
$p(b\|b)$	2/6
$p(c\|b)$	1/6
$p(a\|c)$	2/4
$p(b\|c)$	1/4
$p(c\|c)$	1/4

4. Adding in `<UNK>`

```
<UNK> a <UNK> b
a <UNK> a b
b a b a
```

5. Updated unigram and bigram probabilities

	MLE prob
$p(a)$	5/12
$p(b)$	4/12
$p(\texttt{<UNK>})$	3/12

The bigram probabilities:

	MLE prob
$p(a\|a)$	0
$p(b\|a)$	2/4
$p(\texttt{<UNK>}\|a)$	2/4
$p(a\|b)$	2/2
$p(b\|b)$	0
$p(\texttt{<UNK>}\|b)$	0
$p(a\|\texttt{<UNK>})$	2/3
$p(b\|\texttt{<UNK>})$	1/3
$p(\texttt{<UNK>}\|\texttt{<UNK>})$	0

6. Backoff model with $D = 0.5$:

   - a
     $reserved\_mass(a) = (2 * 0.5)/4 = 1/4$

     Since there is only one unseen bigram, it will get all of the reserved mass (i.e. 1/4).
     If you wanted to do the entire calculation (e.g. for practice):
     denominator $= 1 - \sum_{x:count(ax)>0} p(x) = 1 - (p(b) + p(\texttt{<UNK>}) = 1 - (4/12 + 3/12) = 5/12$
     $\alpha(a) = \frac{1/4}{5/12} = 12/20 = 3/5$
   - b
     $reserved\_mass(b) = (1 * 0.5)/2 = 1/4$

     denominator $= 1 - \sum_{x:count(bx)>0} p(x) = 1 - p(a) = 1 - 5/12 = 7/12$
     (Note this is the same as $p(b) + p(\texttt{<UNK>})$, but for efficiency, since the number of words that do occur following a particular word is usually much, much smaller than the number that don't, programmatically we calculate it using the 1 minus formulation).
     $\alpha(b) = \frac{1/4}{7/12} = 12/28 = 3/7$

   - $\texttt{<UNK>}$
     $reserved\_mass(\texttt{<UNK>}) = (2 * 0.5)/3 = 1/3$

     Since there is only one unseen bigram, it will get all of the reserved mass.
     If you wanted to do the entire calculation (e.g. for practice):
     denominator $= 1 - \sum_{x:count(<UNK>x)>0} p(x) = 1 - (p(a) + p(b)) = 1 - (5/12 + 4/12) = 3/12 = 1/4$
     $\alpha(\texttt{<UNK>}) = \frac{1/3}{1/4} = 4/3$

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
$p(a\|a)$	3/5 * 5/12	2/8
$p(b\|a)$	(2-0.5)/4	3/8
$p($`<UNK>`$\|a)$	(2-0.5)/4	3/8
$p(a\|b)$	(2-0.5)/2	3/4
$p(b\|b)$	3/7 * 4/12	4/28
$p($`<UNK>`$\|b)$	3/7 * 3/12	3/28
$p(a\|$`<UNK>`$)$	(2-0.5)/3	3/6
$p(b\|$`<UNK>`$)$	(1-0.5)/3	1/6
$p($`<UNK>`$\|$`<UNK>`$)$	4/3 * 3/12	2/6