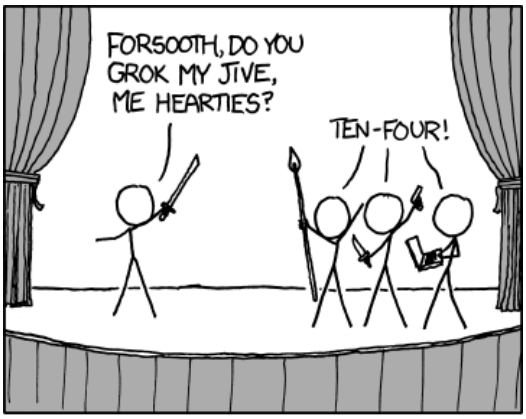
CS159 - Assignment 2a

Due: Thursday, 9/12 @ 11:59pm



`http://xkcd.com/771/`

For the first part of this assignment, we're going to estimate probabilities for some of the different language models by hand. I know this can be a bit tedious, but it serves two important goals. First, it makes sure that you actually understand how the probabilities are calculated. It's easy to think that you do, but then get bogged down once you start to try and implement it. Second, your answers here will provide a test bank for you to compare your answers from your program with.[1] I'll post solutions to these after 24 hours (to accommodate for late submissions) so that you can check to make sure your answers and understanding are correct.

On this assignment, you may work with a partner. If you do, whenever either of you are working on the assignment, you should both be there.

When you're done with this assignment, *DO* start working on part 2b of the assignment. They are meant to be done concurrently.

Given the following corpus of three "sentences", where we only have one letter words:

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

[1] Since we're not including the start and end tags in our calculations below, you may have to tweak your program a bit to make a comparison, but it shouldn't be too hard.

Calculate and the results for the following:

1. The unigram probabilities *without* `<s>` and `</s>` and without the `<UNK>` symbol.

2. The bigram probabilities ($p(Y|X)$) without any smoothing and *without* `<s>` and `</s>` and without the `<UNK>` symbol. Since we're not using start and end tokens, be careful about your counts. In particular,

$$P(Y|X) = \frac{\text{number of times X Y occured}}{\text{number of times X occurred } \textit{with a word following it}}$$

   Notice that the denominator is **not** simply the number of times Y occurred. For example, $p(a|a) = 1/5$, since there are only 5 a's that are followed by letters.

   This problem won't happen in your implementation since we will include start and end tags.

3. Same as in problem 2, but using add-lambda smoothing with $\lambda = 1$ (i.e. add-one smoothing).

4. The corpus after `<UNK>` symbol processing. Specifically, replace the first occurrence of each word with the `<UNK>` symbol.

5. The unigram and bigram probabilities *on this new corpus*, still without start and end tokens (i.e. redo problems 1 and 2 on the corpus with unknown word handling).

6. Calculate the bigram probabilities on the new corpus from problem 4 using absolute discounting with $D = 0.5$, again, *without* `<s>` and `</s>`.

   Make sure to show your intermediary work, specifically the reserved mass and α for each "word".