NTMI 2016: Project Exercises
Part A (steps 3)

This project has four steps. The steps are gradual and allow you to build the project in blocks that you can re-use in subsequent steps.

1 STEP3: Smoothing ngram statistics (Due date: 16th Feb 2016)

In this exercise, we will again use the Austen Train corpus of the previous exercises. For training we will use the same corpus as before (`http://www-nlp.stanford.edu/fsnlp/statest/austen.txt`), but we will test the model on additional data (`http://www-nlp.stanford.edu/fsnlp/statest/ja-pers-clean.txt`).

1. Construct a bigram language model based on the training corpus. For this, use the program of the previous exercise.
2. Implement Add-1 smoothing and apply it to the bigram model of the first part.
3. Implement Good-Turing smoothing and apply it to the bigram model of the first part.
   To avoid trouble, use smoothing only for bigrams which appeared $k \leq 5$ times in the training corpus. Use the standard Good-Turing formula (4.27)[1] for $r = 0$ *and distribute mass uniformly over unseen (zero occuring* $r = 0$*) bigrams*, and the following formula (4.31 in the book) for $1 \leq r \leq k$:
   $$r^* = \frac{(r+1)\frac{n_{r+1}}{n_r} - r\frac{(k+1)n_{k+1}}{n_1}}{1 - \frac{(k+1)n_{k+1}}{n_1}}$$
   Use no smoothing for $r > 5$ (keep $r^* = r$ for $k > 5$).

[1] I am referring to formulas in Jurafsky and Martin Edition 2, 2009.

Remarks:

1. Assume that $|V|$ (the number of words in the vocabulary) is equal to the number of *word types* in the full AUSTIN corpus.
2. To calculate the smoothed probability $\hat{P}(w_i|w_{i-1})$, you need to use the formula $\hat{P}(w_i|w_{i-1}) = \frac{Count^*(w_{i-1}w_i)}{\sum_{w \in V \cup \{STOP\}} Count^*(w_{i-1},w)}$ (where $Count^*$ is the frequency after smoothing).[2] Make sure you calculate the sum $\sum_{w \in V \cup \{STOP\}} Count^*(w_{i-1}, w)$ in an efficient way.
3. Test the model without smoothing and with both forms of smoothing (with $k = 5$ in the second smoothing model) on the test corpus, as follows:
   (a) Report the percentage of sentences which are assigned probability zero by each of the models.

   (b) Report the first 5 sentences in the test corpus which are assigned a probability of zero by each of the models.

   (c) Explain the differences between the models.
4. Please submit the following:
   - The program with clear instructions as to how to compile and run it.
   - The output for part 3.

The requested information should print when run with the following parameters:

```
./a1-step3 -train-corpus [path] -test-corpus [path] -n [value] -smoothing
[no|add1|gt]
```

[2] Why is it no longer possible to use the formula $P(w_i|w_{i-1}) = \frac{Count(w_{i-1}w_i)}{Count(w_{i-1})}$?