Programming Problem - Bigram Language Model

In this assignment, you will be implementing a Bigram Language Model with smoothing, as is detailed in the first chapter of the course text. You will also be evaluating the strength of your language model on a set of test data, and reporting your results.

You will be using a subset of the Penn Treebank Corpus - a standard dataset used to benchmark language models. Specifically, you will be training your language models on the file `penn-tree-bank-train.txt`, using the heldout data `penn-tree-bank-heldout.txt` to find the optimal β for your bigram model, and using the test data `penn-tree-bank-test.txt` to evaluate your models. All of these files can be found in the directory `/course/cs1460/asgn/langmod/data/`.

Note: These files are preprocessed with UNK and STOP symbols for you - you just need to read in the files as is, treating the entire corpus as one continuous stream of tokens.

The evaluation metric you will be using in this assignment is **perplexity**. As a brief refresher from class, the formula for unigram perplexity is as follows:

$$\text{Perplexity} = \exp(-\frac{1}{N}\sum_{i=1}^{N} \ln \theta_i)$$

N is the number of unigrams (words) in the test corpus, and θ_i is the unigram probability computed via your model. Bigram perplexity is similar except you will be replacing the unigram probability with the appropriate bigram probability (Θ_{w_1,w_2}).

We strongly suggest you walk through the following steps to do this:

1. Create a unigram language model, trained on the training data. Note that because the Penn Treebank data is preprocessed and UNK'd for you, there is no need for smoothing (hint: you will want to use equation 1.3 from the course textbook to learn the unigram parameters). Compute the Unigram model perplexity on the test data.
2. Construct a smoothed bigram model, trained on the training data, with the parameter $\beta = 1$. Compute the Bigram model perplexity on the test data.
3. Now set the bigram parameter β to optimize the likelihood of the heldout data, as described in the course textbook. What values of β optimize the heldout data? What is the new perplexity of the test data?

The template script is `/course/cs1460/asgn/langmod/langmod`. *Copy this file and fill in the specified line with the command that runs your code and* *include this with your handin.* **Make sure to adhere to the output guidelines detailed in the template script - any submissions that fail to do so will have points deducted from the total score**. To hand in, run `/course/cs1460/bin/cs146_handin langmod` from the directory that contains your code.