Assignment 3 - Spell Check

Sharath Chandra Bagur Suryanarayanaprasad
800974802
sbagursu@uncc.edu

Accuracies with the given default Holbrook training corpus-

```
Uniform Language Model:
correct: 31 total: 471 accuracy: 0.065817
Laplace Unigram Language Model:
correct: 52 total: 471 accuracy: 0.110403
Laplace Bigram Language Model:
correct: 64 total: 471 accuracy: 0.135881
Custom Language Model:
correct: 43 total: 471 accuracy: 0.091295
```

Accuracies with the updated Holbrook training corpus with more training data (Additional data fetched from Santa Barbara Corpus of Spoken American English)-

```
Uniform Language Model:
correct: 36 total: 471 accuracy: 0.076433
Laplace Unigram Language Model:
correct: 63 total: 471 accuracy: 0.133758
Laplace Bigram Language Model:
correct: 84 total: 471 accuracy: 0.178344
Custom Language Model:
correct: 52 total: 471 accuracy: 0.110403
```

Discussion-

Using the default Holbrook Corpus handed out, Laplace Unigram Model is implemented based on the formula to calculate Unigrams. Based on Unigrams we develop Laplace Bigram Model. Laplace Bigram Model with smoothing-

score(wi-1, wi) = (count(wi-1,wi) + 1)/(count of bigrams + unigram count(wi-1))

The above scores are computed as logprobs in logarithm space. The accuracies obtained are documented above.

Based on the bigram model, in the custom model, we extend the implementation to Laplace Trigram Model, as followsscore(wi-2,wi-1,wi) = (count(wi-2,wi-1,wi)+1)/ (count of trigrams + bigram count(wi-2,wi-1)

The accuracies for this model is noted in the Custom Language Model.

Difficulties faced-

The main difficulties faced during the course of implementation were that of handling unseen data. Though smoothing is implemented the accuracy seen due to trigram(Custom model) is due to large number of unseen bigrams. This hampers the overall performance of the model because smoothing treats all unseen data with equal probability. However we do observe a significant increase in accuracies when the training data corpus of Holbrook is extended. When more data is provided to the training set, we see an increase in the accuracies. The reason behind this is that the language model is now able to recognize more unseen bigrams or unigrams and incorporate the occurrences of these in the formula to calculate trigrams. However the context of the corpus plays a significant role. The additional data extended may not be the same context as that of the provided HolBrook corpus. Through the execution of models on dev set we observe that Laplace Bigram and Laplace Unigram accuracies are significantly high with respect to each other solving the problem provided in part 2 of the assignment.

Execution Steps-

There are 2 folders, assignment3_problem1 and assignment3_problem2. assignment3_problem1 contains the entire library with the default holbrook corpus. assignment3_problem2 contains the entire library with holbrook corpus with additional training data.

The paths for training files are relative paths. No hardcoded paths are used. Run the SpellCheck.py file in order to obtain the accuracy results.