COMP 150 NLP

Fall 2016

Problem Set 2

Language Modeling

Mingzhe Li

This program strives at evaluating different language models and smoothing methods.

It utilizes the defaultdict in python to count the unigrams and bigrams. A bigram dictionary was implemented for the sole purpose of check distribution. Other than that, I pretty much followed the provided design.

From the final output, it is obvious that smoothing is critical in language modeling. Without smoothing, the bigram ("party", "!") appeared in the test set but never in the training set caused the perplexity to be infinity. With smoothing, the perplexity is significantly reduced.

Program outputs

```
<S> The Fulton County Grand Jury said Friday an investigation of
Atlanta's recent primary election produced `` no evidence '' that any
irregularities took place . </S>

<S> Several were firing into the barn when Billy <UNK> arrived . </S>

<S> `` I can't leave the party ! ! </S>

Distribution Check:  Valid distribution!

Perplexity without smoothing:  inf

Perplexity with Laplace smoothing:  945.403396886

Perplexity with linear interpolation:  176.721565305

lambda 1 =  0.354212631617

lambda 2 =  0.645787368383

Perplexity with deleted interpolation:  168.263365104


Process finished with exit code 0
```