NLP Assignment 2

Total Number of Words: 1857
Total Number of Sentences: 107
Number of Unique Words: 770

For our data scrubbing, we ran the following code on the corpus, which replaced many of the things we didn't want to cause issues with our tokenizing steps. In addition to the following replacements, we replaced every formatted quotation and apostrophe marks with the corresponding ASCII characters. The changes that we made were to make parsing and tokenization easier. We removed or properly spaced periods to improve our sentence tokenizing ability. We also removed contractions and instead used the expanded version of the words. This makes tokenizing words easier, as the parser would automatically separate pre- and post- apostrophe text. We did ensure that the possessive would be tokenized as we wanted, which is to say as a single word. So the word "Blake's" tokenizes as ["Blake's"] and not as ["Blake", "'", "s"], which is what it would've been without our specification.

".I " => ". I "
".It " => ". It "
".That " => ". That "
"Mr. " => "Mr "
"Dr. " => "Dr "
"F.B.I." => "FBI"
"Catherine E." =>
"Catherine E"
"ä" => "a"
"'re" => " are"
"'ve" => " have"
"n't" => " not"
"I'm" => "I am"
"It's" => "It is"
"it's" => "it is"
"What's" => "What is"
"what's" => "what is"
"There's" => "There is"
"there's" => "there is"
"that's" => "that is"
"That's" => "That is"
"He's" => "He is"
"he's" => "he is"
"Let's" => "Let us"
"\"" => ""

Perplexity values for our models given different splits

	70/30	**80/20**	**90/10**
Unigram	18.962	18.983	13.401
Bigram	10.186	10.126	9.605
Trigram	9.694	9.739	9.341

The values were as expected in that the 90/10 split always did the best of any ngram model we tried. This makes sense as the model has more training data, which means that it has been exposed to more words and word combinations, thus allowing it to better recognize any ngram. It also had a smaller test data set, which would mean a lower likelihood of running into something it doesn't understand and/or recognize. It also functions as expected because typically, the n+1 gram does better than the n gram. Unexpectedly, there would be a slight increase from the .7 split to the .8 split. This could be various reasons due to where the split occurs, context of the split, and other factors.

Our model couldn't be generalized very well, as it is based upon a very small corpus (containing only 770 unique words, with only 1857 total). This is not nearly big enough to be a representative corpus (with a good size for a corpus being around 1 million words). It is also a very targeted passage, so is not generalizable to multiple contexts, this is better shown below with the sentence MLE scores.

MLE values for sentences using our best models

	Unigram	**Bigram**	**Trigram**
Please return the television to Mr. Lynch on Sunday.	9.314×10^{-57}	9.999×10^{-55}	9.999×10^{-49}
Odysseus was 17 when he first had novocain.	9.314×10^{-51}	9.999×10^{-49}	9.999×10^{-43}
Johnny Cash was drowning in imagery of dusty roads at night.	9.314×10^{-69}	2.929×10^{-53}	9.999×10^{-61}
While wearing blue shoes, I took a detour to find Laura Frost, the Wizard of Python Empire.	7.476×10^{-104}	6.417×10^{-111}	9.999×10^{-109}
I built a box for the rabbits.	5.781×10^{-40}	2.468×10^{-39}	9.999×10^{-37}
She was wearing a leather jacket.	5.781×10^{-34}	2.436×10^{-31}	9.999×10^{-31}
I was born in 1976 and graduated high school in 1990.	9.314×10^{-69}	1.202×10^{-58}	9.999×10^{-61}
That explanation is irrelevant to surrealists, existentialists and film snobs for purposes of interpretation, narrative	1.369×10^{-128}	1.202×10^{-130}	9.999×10^{-130}

from, humorous asides and soap.			
It is a truth universally acknowledged that a zombie in possession of brains must be in want of more brains.	3.587×10^{-113}	9.872×10^{-118}	9.999×10^{-115}
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair.	2.963×10^{-380}	7.422×10^{-354}	4.936×10^{-406}

These MLE values are much smaller than we were expecting. Especially for sentences that contain words that are very likely to appear in our corpus with some frequency, the MLE never gets even close to 1%. Interestingly, they appear to be proportional to each other in a way that we might expect. For example, our last sentence (which contains basically no meaningful words that appear in our corpus), the MLE is significantly lower than for every sentence. This holds for what we would reasonably expect to see for a sentence that has little to no overlap with the initial corpus. The sentences with more frequently-appearing words also have higher MLEs, which also makes sense. Since our best models are all based on 90/10 splits, it would also be reasonable to assume that our MLEs would on average be lower if we were comparing these sentences to a smaller testing set.