a. what are n-grams and how are they used to build a language model

Ngrams are useful as we can use them to "train" a model as we did in our code. We created a dictionary of ngrams weighted by their frequency in the training data which we can then use to calculate the probability of n grams in other new texts.

b. list a few applications where n-grams could be used

An obvious application is language detections such as we did in our code, in which we could predict the language in the test data with a high degree of success. Along this line, I can imagine a grammar checker using n grams in which the likelihood of an n-gram could be calculated simply by training the data on grammatically correct data.

c. a description of how probabilities are calculated for unigrams and bigrams

Probability is calculated by splitting the test data into bigrams, calculating the laplace smoothing probability of each bigram, and then multiplying the probabilities of all the bigrams. The formula for laplace smoothing is ((hits of the bigrams in the bigram dict) +1) / ((hits of the first element of the bigram in the unigram dictionary) + the number of vocabulary across training data).

d. the importance of the source text in building a language model

The source text obviously must be accurate and large enough to sufficiently train the model. If the English source text was all German, then the model would fail to categorize english. If the source text was not large enough to have sufficient entries, then it would also fail. Say the German source text happened to have words that overlapped with English, it is possible that an English test text could be miscategorized if it were too short.

e. the importance of smoothing, and describe a simple approach to smoothing

Smoothing refers to improving a model to be able to handle n-grams which were not present in the training data. A simple approach would be to simply split test data into n-grams as we did in the code, and multiply the probabilities of each of these n-grams.

f. describe how language models can be used for text generation, and the limitations of this approach

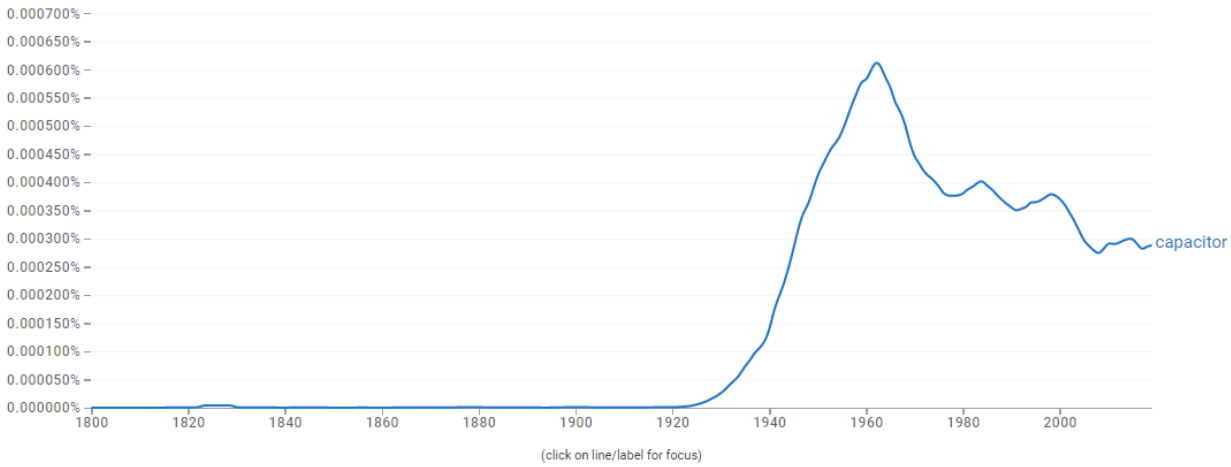
Language models can be used to complete text or fill in the blank. Comprehension is another issue which limits the length of coherent generations.

g. describe how language models can be evaluated

A model can be evaluated given the accuracy of its predictions, and by analyzing the locations of its failures.

h. give a quick introduction to Google's n-gram viewer and show an example

Simply enter an nGram and google Ngram viewer displays the frequency in books on google books. An example is included as a chart below:


0.000700%
0.000650%
0.000600%
0.000550%
0.000500%
0.000450%
0.000400%
0.000350%
0.000300%
0.000250%
0.000200%
0.000150%
0.000100%
0.000050%
0.000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
capacitor
(click on line/label for focus)