Ngrams Narrative

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

- N-grams are sliding windows over text, where the sliding window has n words. For example, unigrams have a sliding window of size 1 word, bigrams have a sliding window of 2 words, and so on. The unigrams for the sentence, “How are you?” are [‘How’, ‘are’, ‘you’, ‘?’], and the bigrams are [(‘How’, ‘are’), (‘are’, you’), (‘you’, ‘?’)]. By counting how often sequences of words occur in a text, n-grams can be used to build a language model.

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

- N-grams could be used in spell and grammar check software, language translation software, speech recognition software, and autocomplete software.

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

- For unigrams, the probability of a certain unigram occurring in a text is calculated by dividing the number of occurrences of that unigram by the total number of unigrams in the text.
- For bigrams, the probability of a certain 2-word sequence occurring in a text can be calculated by multiplying the probability of the first word occurring * the probability of the second word occurring after the first word. For example, the probability of ‘are you’ occurring given the very small corpus of, ‘how are you. i am good. are you good?’ is P(are, you) = P(are) * P(you | are).

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

- The source text is very important in order to create a good language model. The source text serves as a training set for the language model, so a large training set/source text is

necessary to create an effective model. The source text should also be preprocessed to remove noise.

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

- Smoothing is important in probabilistic computation because it removes zero counts. When a zero count occurs during calculation, it will completely zero out the probability of the current feature (sparsity problem). A simple approach to smoothing is 1-laplace smoothing, which adds 1 to all counts (to remove the zero counts). This is balanced out by adding the total vocabulary count to the denominator

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

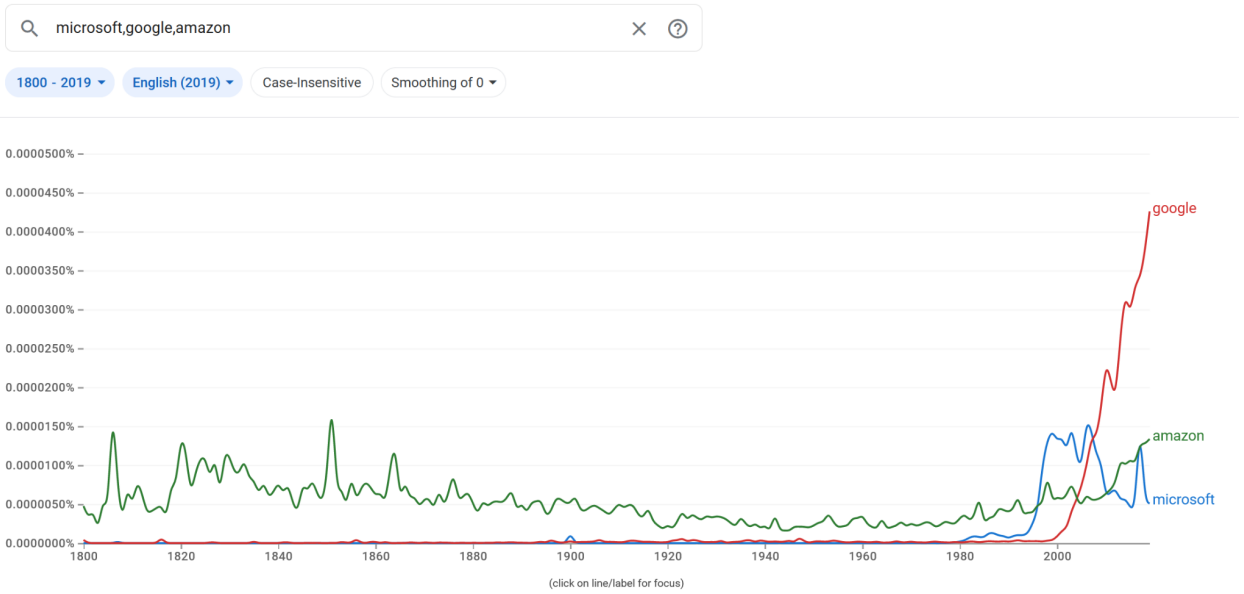
- Language models can be used for text generation by calculating probabilities of bigrams and unigrams and continually taking the bigram with the highest probability given the start word's position. The limitations of this approach are its small size (higher n-grams yield better results) and its simplicity.

g. Describe how language models can be evaluated

- There are two ways language models can be evaluated. Extrinsic evaluation uses human annotators to determine how good a language model is and intrinsic evaluation uses an internal metric like complexity to evaluate the language model.

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

- Google's n-gram viewer is an online search engine that takes n-grams as search queries and plots their frequencies on a graph by year. This data is derived from a corpus of books dated by year. Here is an example of the n gram viewer with the query "microsoft,google,amazon".


microsoft,google,amazon
1800 - 2019
English (2019)
Case-Insensitive
Smoothing of 0
0.0000500%
0.0000450%
0.0000400%
0.0000350%
0.0000300%
0.0000250%
0.0000200%
0.0000150%
0.0000100%
0.0000050%
0.0000000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
google
amazon
microsoft
(click on line/label for focus)