Houston Holman

3/2/23

N-grams Project Narrative

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

- N-grams are continuous sequences of tokens extracted from a given text. They are important tools in building a language model because they can be used to predict the likelihood of the next word in a sentence given the previous words.

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

- Some applications that can utilize n-grams include:
    - Language modeling: By predicting the probability of the next word in a sentence, language models can be created.
    - Spell check: Spelling errors can be identified by comparing the n-grams in a word against a dictionary of valid n-grams.
    - Speech recognition: N-grams can model the likelihood of phonemes given the preceding phonemes, allowing for improved speech recognition.

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

- To calculate the probability of a unigram in a corpus, first you must count all the occurrences of the target unigram in the corpus and divide that number by the total number of tokens in the corpus. To calculate the probability of bigrams in a corpus, first count all the occurrences of the target bigram. Then you must divide that number by the count of the total number of occurrences of the first unigram in the target bigram.

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

- The source text is of great importance when building a language model, as it is the foundation that the model is constructed on. A large and varied source provides the model with a greater understanding of the language, including more nuances and complexities. The selection of the source text also depends heavily on the purpose of the model. For example, a language model trained on legal documents may have a difficult time understanding social media posts.

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

- Smoothing addresses the issue of sparse data, which occurs when some words in a dataset have very few occurrences. By assigning a small probability to unseen words, smoothing prevents the language model from making unrealistic predictions and sticking too closely to the training data. A simple approach to smoothing is LaPlace smoothing which adds a small constant value to the count of each word in the dataset, giving each word a nonzero probability.

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

- Language models can be used to generate text by predicting the most likely next word based on previous words. One limitation to this approach is that the generated text is often repetitive or lacks creativity. The model is limited by the patterns it observed in the training data. Another limitation is that the generated text is often not contextually appropriate or relevant. This is due to the model only generating text based on statistical patterns, rather than understanding the meaning of the text.

g. describe how language models can be evaluated

- Language models can be evaluated both extrinsically and intrinsically. Extrinsically, models can be evaluated by comparing their output with human annotators. Intrinsically, a metric like perplexity can be used, which is the inverse probability of seeing the words we observe, normalized by the number of words.

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

- Google's n-gram viewer allows a user to search and visualize the frequency of n-grams in Google's massive collection of digitized books which span centuries. For example, entering the phrase "artificial intelligence" shows that there was a large spike in the late 1980s followed by a sharp decline and rampant growth throughout the 2010s.
- https://books.google.com/ngrams/graph?content=artificial+intelligence&year_start=1900&year_end=2019&corpus=en-2019&smoothing=3