Gebresilassie Takele
NetId : GTT210000

Biruk Mamo
NETID: BGM210001

A.

Contiguous sequences of N items, which can be words, characters, or even phonemes, make up an N-gram in a text. N-grams are frequently used in natural language processing (NLP) to create language models, which are statistical models that forecast the likelihood of a word given a series of preceding words. For instance, a bigram language model determines the likelihood of the subsequent word in a text given the preceding word. These probabilities are calculated using the frequency of bigrams, which are groups of words that follow one another. In other words, the model divides the frequency of the previous word by the frequency of the bigram to determine the likelihood of a word given its prior word. From unigrams (single words) to trigrams, quadrigrams, and so forth, n-grams can be used to construct language models of various sizes. The work at hand and the volume of data at hand will determine which N to choose. More information is needed to estimate larger N-grams properly since they represent more intricate word associations. NLP applications, including machine translation, speech recognition, text categorization, and information retrieval all make use of language models created with n-grams. They can also be utilized for activities like text production, grammar checking, and text completion.

B.

For a variety of purposes, natural language processing (NLP) makes extensive use of n-grams. Machine translation, text classification, and language modeling for speech recognition are a few of them. They are also utilized in named entity recognition, sentiment analysis, text completion, information retrieval, spell checking, and correction. Using bilingual N-gram models to align and translate words or phrases between languages and estimate translation probabilities, N-grams can also be used for machine translation. N-grams are an effective method for processing and analyzing language, enabling precise language modeling and the discovery of patterns and relationships in textual data.

C.

Using unigrams and bigrams, the process of developing statistical language models in natural language processing (NLP) determines the likelihood of a word given a series of preceding words. The probability of a word w given its previous word v is calculated as the frequency of that bigram (v, w) divided by the frequency of the previous word v. The

probability of a word w given its first word v is calculated as the frequency of that word in the text divided by the total number of words in the text. When a word or bigram is absent from the training data, zero probability arises. To address this issue, these probabilities are frequently smoothed using methods like Laplace smoothing. Smoothing increases language correctness.

D.

When building a language model, the source text used for training is critical to its accuracy and generalizability. The corpus must have a broad vocabulary coverage, naturalness, and diversity to handle unseen words or phrases that occur in real-world data. Additionally, the size and quality of the corpus should be considered to avoid noise or errors that can negatively impact the model's performance. Finally, the source text should be adaptable to different tasks and domains to improve its accuracy and generalization capabilities. Choosing the right corpus is essential for building an accurate and robust language model.

E.

To prevent zero probabilities for words or sequences that don't appear in the training data, smoothing is a language modeling technique. Laplace smoothing, also known as add-one smoothing, is a widely used technique that involves adding a fixed value of 1 to the count of each word or bigram in the training data to guarantee that the probabilities are not zero. Adding the constant value to the count and dividing by the overall word count or vocabulary size yields the smoothed probability. Laplace smoothing is easy to use and efficient, however there are different smoothing methods that can be applied depending on the task and the data.

F.

The next word in a series can be predicted using language models, which can then be used to generate text. Several sampling techniques, including random sampling, greedy decoding, and beam search, can be applied during the creation process. Unfortunately, generated text might not be imaginative or coherent, and it might repeat prejudices and stereotypes found in the training data. Moreover, language models are restricted to particular domains or datasets, and their performance may suffer when used for various tasks. When utilizing language models for text production in real applications, certain restrictions must be taken into consideration.

G.

Depending on the task at hand and the nature of the data, a variety of measures can be employed to gauge how well language models work. Perplexity, a metric that assesses how effectively a language model predicts test data, is one of the most frequently used metrics for language models. Accuracy, F1-score, BLEU score, and human evaluation are additional metrics that are applied to certain activities or applications. These measures aid in evaluating how well language models produce content that is accurate and coherent.

H.

The Google Books corpus' word and phrase usage frequency can be examined using the potent n-gram viewer, which is a feature of Google. The corpus has a sizable collection of books that have been scanned over the course of many centuries, allowing users to research cultural patterns, language development, and the birth of new ideas and concepts in society. Users can input particular words or phrases and a time frame to produce a graph that displays the frequency of the word or phrase over time. The phrase "artificial intelligence" serves as an illustration of how the n-gram viewer may be used to research the historical evolution of linguistic usage and cultural fads, showing that the frequency of the term increased noticeably in the 1950s and continued to climb substantially in the 1980s.