N-grams are contiguous sequences of n items (words, letters, or symbols) within a text. N-grams are often used in natural language processing (NLP) and machine learning applications, such as language modeling, text classification, and sentiment analysis. They can provide important insights into the structure and meaning of a text, as well as help identify patterns and relationships between words.

N-grams are used to build language models, which are statistical models that predict the probability of a sequence of words or phrases occurring in each language. It is used to estimate the probability of a word given the previous n-1 words, which can be used to generate coherent sentences and improve the accuracy of speech recognition and machine translation systems. It is also used to represent text data as vectors, which can be used to train machine learning models for text classification tasks, such as sentiment analysis, topic classification, and spam detection.

The probability of a unigram (single word) is calculated as P(w) = count(w) / N where count(w) is the number of occurrences of w in C, and N is the total number of words in C. The probability of a bigram (sequence of two words) w1w2 in a text corpus C is calculated as: P (w2 | w1) = count(w1w2) / count(w1) where count(w1w2) is the number of occurrences of the bigram w1w2 in C, and count(w1) is the number of occurrences of the word w1 in C.

The source text is extremely important in building a language model because it determines the accuracy and effectiveness of the model. A language model is a statistical model that predicts the probability of a sequence of words in a language. To build a language model, a large corpus of text is required, and the quality and diversity of the corpus have a significant impact on the quality of the language model.

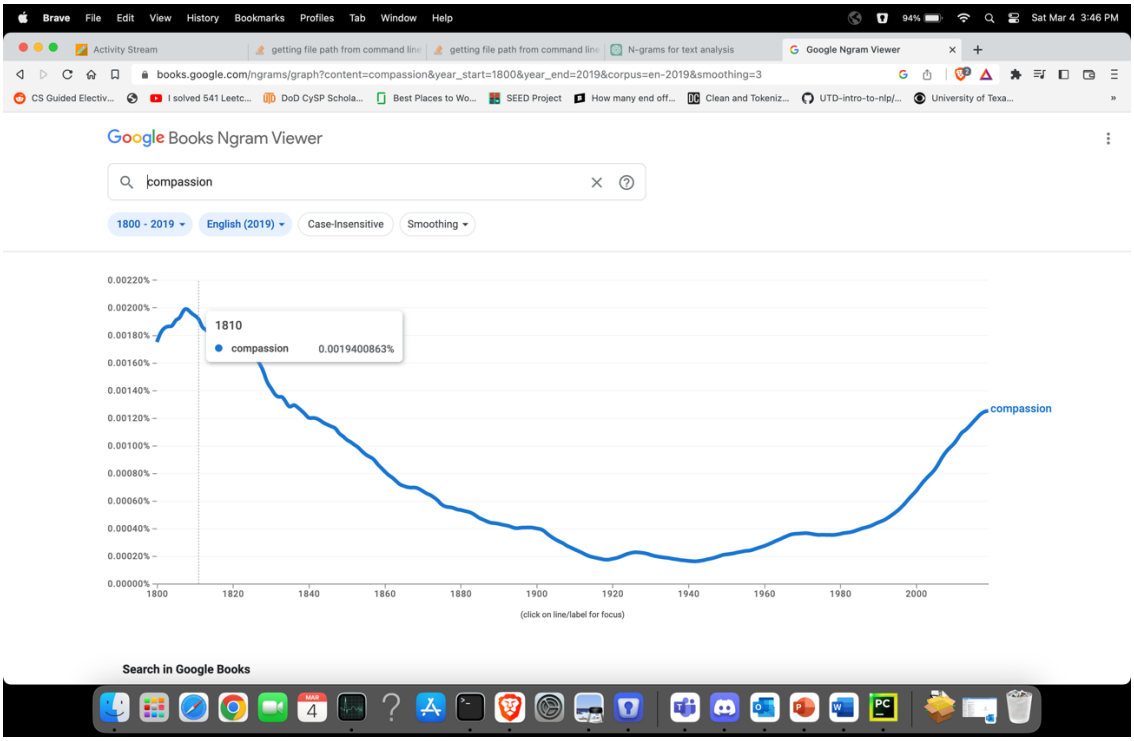
Smoothing is an important technique in natural language processing (NLP) and machine learning that helps to address the problem of zero probabilities for unseen n-grams in a language

model. Unseen n-grams are those that are not present in the training corpus but may appear in the test or evaluation data. Without smoothing, the probability of unseen n-grams would be zero, which can lead to poor performance and inaccurate predictions. The basic idea of smoothing is to redistribute some probability mass from seen n-grams to unseen n-grams, which helps to make the language model more robust and accurate. There are several smoothing techniques used in NLP, including Laplace smoothing, Good-Turing smoothing, and Jelinek-Mercer smoothing.

Language models can be used for text generation by predicting the probability of the next word in a sentence given the previous words. This can be done by training a language model on a large corpus of text and then using the model to generate new text by sampling from the predicted probabilities.

To evaluate a language model, a test corpus is typically used to measure the model's performance on unseen data. The test corpus should be representative of the target domain and should include a range of sentence structures and vocabulary. In summary, language models can be evaluated using various metrics to measure their accuracy and performance on a test corpus. The choice of evaluation metrics depends on the application and the goals of the language model.

Google's Ngram Viewer is a web-based tool that allows users to search and analyze the frequency of words or phrases in a large corpus of text. The tool uses Google's vast digital book collection and other text sources to generate n-gram graphs, which display the frequency of a given word or phrase over a specific time. Below is an example of Google’s N-gram:


Brave File Edit View History Bookmarks Profiles Tab Window Help
Sat Mar 4 3:46 PM
Google Ngram Viewer
books.google.com/ngrams/graph?content=compassion&year_start=1800&year_end=2019&corpus=en-2019&smoothing=3
Google Books Ngram Viewer
compassion
1800 - 2019
English (2019)
Case-Insensitive
Smoothing
0.00220%
0.00200%
0.00180%
0.00160%
0.00140%
0.00120%
0.00100%
0.00080%
0.00060%
0.00040%
0.00020%
0.00000%
1800
1820
1840
1860
1880
1900
1920
1940
1960
1980
2000
1810
compassion 0.0019400863%
compassion
(click on line/label for focus)
Search in Google Books