Ngrams (word & character)

Ngrams: What are they?

In the fields of computational linguistics and probability, an n-gram is a contiguous sequence of n items from a given sample of text or speech (a corpus). The items can be phonemes, syllables, letters, words or base pairs according to the application. The n-grams typically are collected from a text or speech corpus. When the items are words, n-grams may also be called shingles.
https://en.wikipedia.org/wiki/N-gram

Using Latin numerical prefixes, an n-gram of size 1 is referred to as a "unigram"; size 2 is a "bigram" (or, less commonly, a "digram"); size 3 is a "trigram". English cardinal numbers are sometimes used, e.g., "four-gram", "five-gram", and so on.

Word Ngrams: Examples

Consider the sentence: “The train arrived late at destination because of a hailstorm along the way.”
“The,” “train,” “arrived,” “late,” are word unigrams. “The train,” “train arrived,” “arrived late,” … are bigrams. “The train arrived,” “arrived late,” … are bigrams. “The train,” “train arrived,” “arrived late,” … are bigrams. “The train arrived,” “train arrived late,” … are trigrams. Although in principle one could compute Ngrams of any size, typically bigrams or, at most, trigrams yield the most significant information.

Character Ngrams: Examples

Consider the same sentence: “The train arrived late at destination because of a hailstorm along the way.”
“T,” “h,” “e,” “t,” “r,” … are character unigrams. “Th,” “he,” “et,” “tr,” … are character bigrams. “The,” “het,” “etr,” “tra,” … are character trigrams.

What do Ngrams tell you?

Word Ngrams have been used in the field of culturomics to study changing temporal shifts in culture. But they have also been used as markers of authorial style (e.g., some authors using certain combinations of words and characters more than other authors; see the TIPS TIPS_NLP_Style analysis.pdf). But they can also be used to improve online spelling checker and automatic rewriting or to predict the next word while writing an online text.

Word and character Ngrams by Sentence Index: What are they and what do they tell you?

The NLP Suite algorithm not only computes Ngrams but it computes them by sentence index, reporting in a csv file each Ngram by specific sentence numbers (1, 2, 3, …). Are certain Ngrams of interest used at the beginning, middle, or end of a document or used across a document? This information may give us clues about an author's style (see the TIPS TIPS_NLP_Style analysis.pdf).

More TIPS

TIPS_NLP_Ngram and Word Co-Occurrence Viewer.pdf
TIPS_NLP_Ngram Google Ngram Viewer.pdf
TIPS_NLP_Style analysis.pdf