N-gram similarity and distance

Abstract: In many applications, it is necessary to algorithmically quantify the similarity exhibited by two strings composed of symbols from a finite alphabet. Numerous string similarity measures have been proposed. A particularly well-known method is edit distance, which is defined as the minimum number of elementary edit operations needed to transform one string into another. Another, closely related approach relies on finding the length of the longest common subsequence (LCS) of the two strings.

We develop a notion of n-gram similarity and distance. We show that edit distance and the measure based on the length of the LCS are special cases of n-gram distance and similarity, respectively. We provide formal, recursive definitions of n-gram similarity and distance. Experiments suggest that measures based on bigram/trigram distance and similarity outperform their unigram equivalents.