Edit Distance in NLP

By Prof Nirali Arora

Minimum Edit Distance

Minimum edit distance is another similarity measure that can be used to find strings that are close to a given string.

A popular use in NLP is in autocorrection systems, where it is used to suggest corrections for misspelled words. Autocorrect is a feature that is built into many devices and applications, such as smartphones, email clients, and text processors.


Your mom and I are going to divorce next month
what??? why! call me please?
I wrote Disney and this phone changed it. We are going to Disney.

Edit distance, also known as Levenshtein distance, is a measure of the similarity between two strings by calculating the minimum number of single-character edits required to change one string into the other. It provides a quantitative measure of how *different* or *similar* two strings are.

Operations for Edit distance

The operations that can be performed are:

- Insertion
- Deletion
- Replacement

For the minimum edit distance, we want to find the minimum number of operations that are needed to transform one string into another string.

Operations on Edit Distance

Consider the words `kitten` 🐈 and `sitting` 🧘. The edit distance between them is 3 because the following three operations can transform one into the other:

- Replace `k` with `s`
- Replace `e` with `i`
- Insert `g` at the end

Applications on Edit distance

- Autocorrection / Spell checking: Minimum Edit Distance is often used in spell-checking algorithms to suggest corrections for misspelled words. By calculating the minimum edit distance between a misspelled word and candidate words in a dictionary, a spell checker can identify potential corrections.
- Information Retrieval: Minimum Edit Distance can be used in information retrieval systems to find similar words or phrases in a database. This is particularly useful in search engines when dealing with queries that may contain typos or slight variations.
- Plagiarism Detection: Minimum Edit Distance can be applied to compare and analyze text documents for plagiarism detection. By measuring the similarity between documents in terms of edit operations, it becomes possible to identify instances of copied or closely paraphrased content.
- OCR (Optical Character Recognition): OCR systems may use Minimum Edit

Collocations in NLP

In natural language processing (NLP), collocations are combinations of words that frequently appear together more often than would be expected by chance. These combinations can include idiomatic expressions, fixed phrases, or simply pairs of words that commonly occur together in a particular order, such as "strong tea," "heavy rain," or "make a decision." Understanding collocations is crucial for applications like machine translation, information retrieval, and language generation because they reflect the way language is naturally used.

Collocations in NLP

Collocations can be classified into several types based on the grammatical relationship between the words:

1. **Adjective + Noun**: e.g., *strong tea*, *heavy rain*.
2. **Verb + Noun**: e.g., *make a decision*, *take a look*.
3. **Noun + Noun**: e.g., *data analysis*, *traffic light*.
4. **Adverb + Adjective**: e.g., *deeply concerned*, *highly unlikely*.
5. **Verb + Adverb**: e.g., *run quickly*, *whisper softly*.

Importance of Collocations

Natural Language Understanding: Recognizing collocations helps in understanding the meaning and context of text more accurately.

Machine Translation: Properly identifying collocations can improve the quality of translations by ensuring that idiomatic expressions are translated correctly.

Text Generation: Using collocations allows for more fluent and natural-sounding text generation.

Search and Information Retrieval: Enhancing search algorithms to recognize collocations can improve the relevance and accuracy of search results.

Methods to identify Collocations

Frequency-Based Methods

These methods identify collocations based on the frequency of word pairs appearing together.

- **Bigram Frequency: Counts the occurrences of two words appearing consecutively.**

Statistical Methods:(Mutual associations,Chi square test and T score)

Using NLP Libraries