**Note**: In NLP, the tokens (individual words) are used as the features.

Each row (sample) is usually a sentence or document, while each column (feature) is a token.

Usually, the most frequently occurring tokens are used as features.

The entry i, j in such a matrix captures the frequency (or weight) of the j'th term of the *vocabulary* in document i.

**Tokenization**: Separating words using space as delimiter (Words Tokenization). Besides words, we can also tokenize sentences based on individual characters (Character Tokenization).

**Embedding**: Converting each word into a vector representation. The numbers of the vector are adjusted and “learned” during the training process such that words that have similar meaning will have similar vector representations.

**Hashing Vectorizer**: A function that turns text into a matrix representation of rows x columns, where rows are the individual sentences or documents and the columns are the tokens.

**Maximum Sequence Length**: The maximum length of the phrase being processed. Phrases shorter will be zero-padded to match the maximum sequence length, while phrases longer will be truncated.

**Stopwords**: Words like “the”, “at”, ”when”, etc. that do not contribute to the meaning of a sentence and therefore can be removed during the preprocessing stage.

**Lemmatization/Word Stemming**: A process which converts words to their root forms (word stems). This is done to group words with similar meanings. This reduces the number of words from the text. E.g. the words “running”, “outrun”, and “runner” all have the same base word “run”.