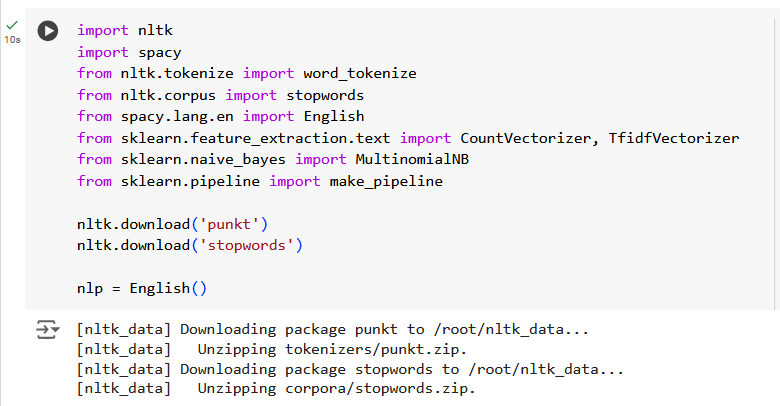
We understood the differences between different feature extraction functions available in the Python machine learning package *scikit-learn*. There are multiple vectorizers available, two of which are CountVectorizer and TfidfVectorizer.

# Preparing the Packages and Dataset

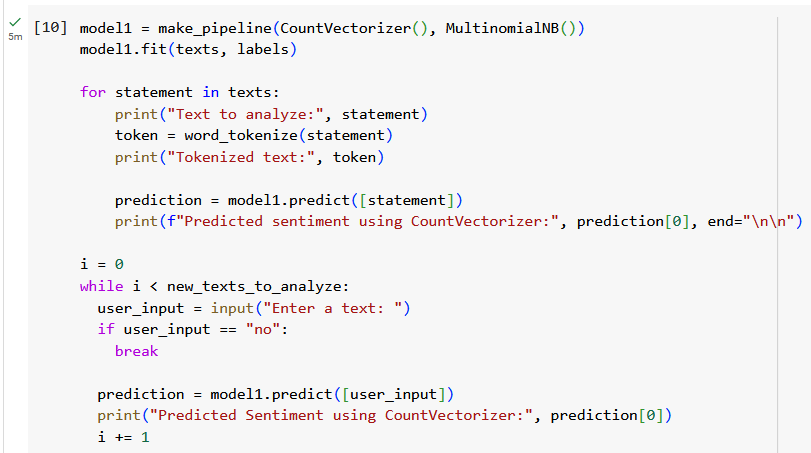






# Initiating the Vectorizer

## With CountVectorizer



## With TfidfVectorizer



# Outputs

|  |  |
| --- | --- |
|  |  |

# Insights on the Two Functions

CountVectorizer simply converts the given texts into a matrix of token counts. This means that if a token is repeated multiple times, it will be prioritized as important parts of a label. Similarly, whatever input is inserted into the function is the only source of prioritization for CountVectorizer.

Since we have not provided the module with a dictionary categorizing every word in the English language, text containing new words are of an unknown entity, and are simply set aside. This also means that very common words, such as articles (i.e. “the”, “a” and “an”) and conjunctions (“and” and “or”), are rated highly, which is contrary to the inner workings of any language.

TfidfTransformer transforms this basic count matrix into a normalized representation based on TF-IDF (meaning, term-frequency times inverse document-frequency).

Instead of using the raw frequencies of occurrence in a given set of texts, the goal is to scale down the tokens that occur too frequently in favor of ones that occur in a smaller scale in the training corpus.