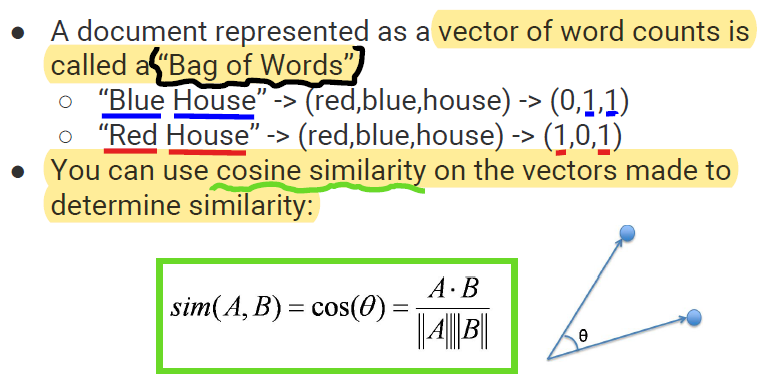
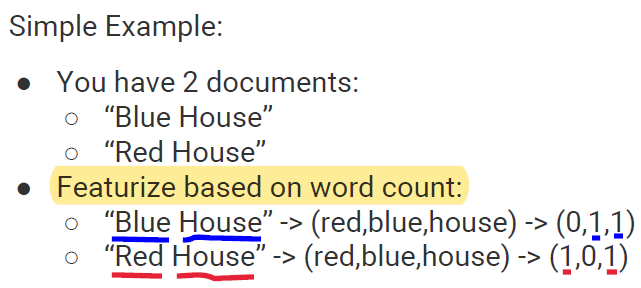
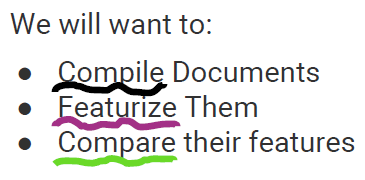
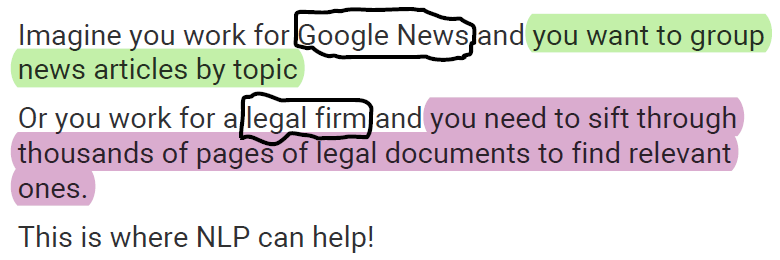
**Natural Language Processing (NLP)**



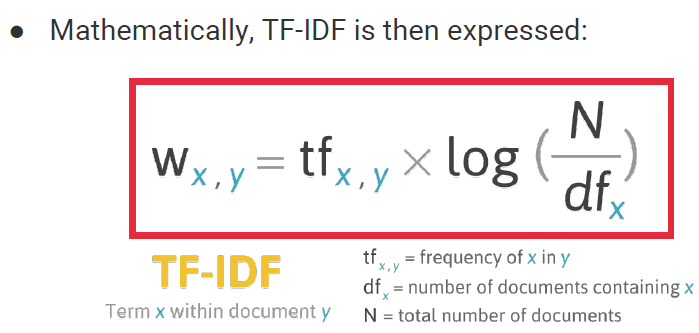
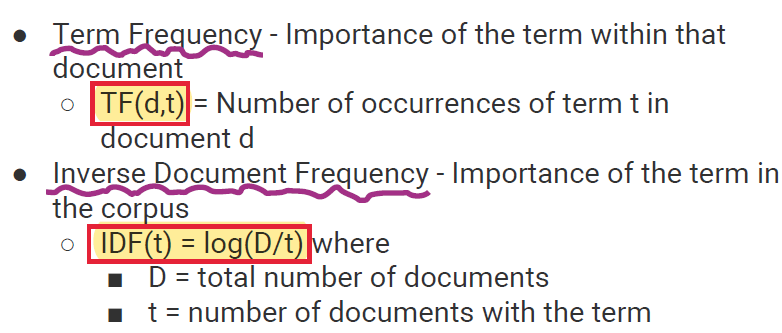
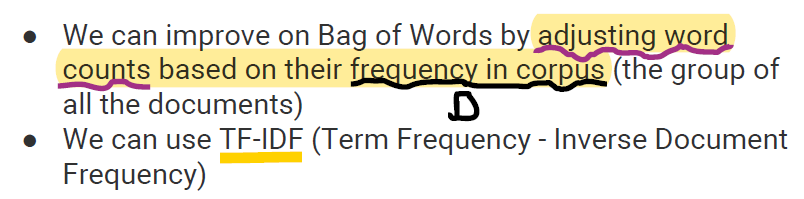
**Featurize**

**Compile**

**Compare**

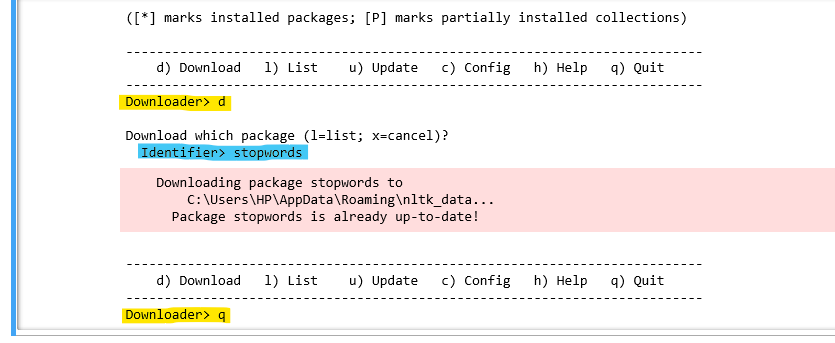
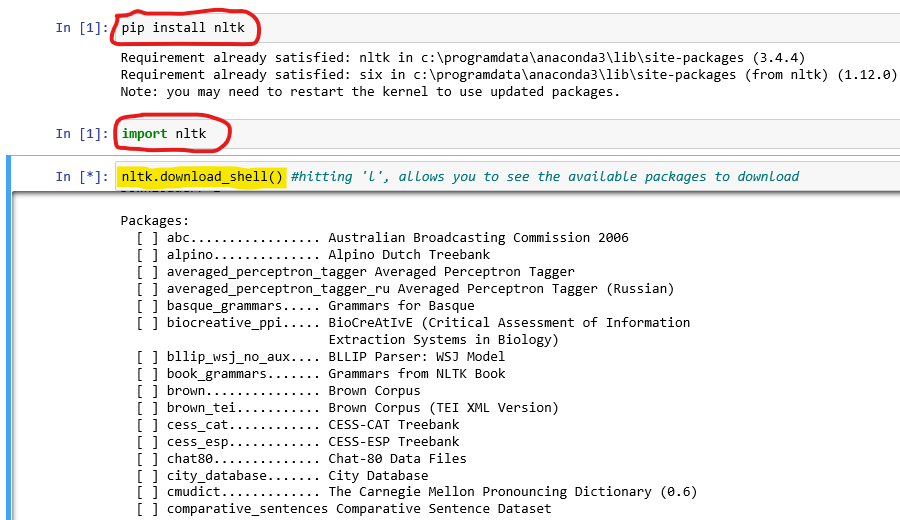
Vector of word counts = **Bag of Words**

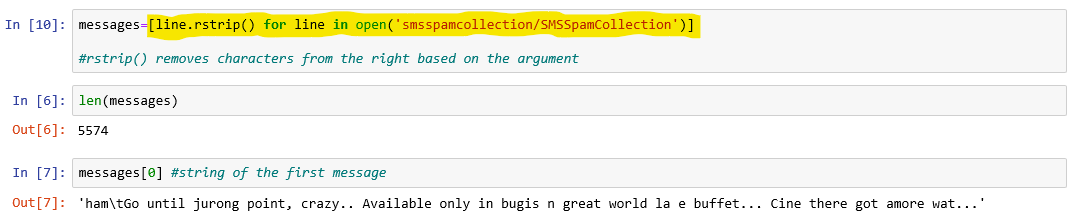
**Bag of Words**

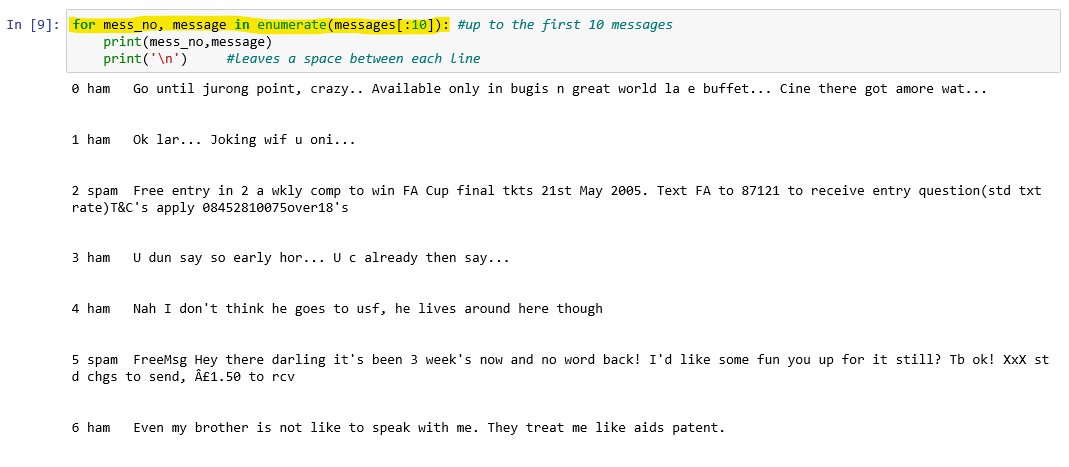


Term Frequency - Inverse Document

Frequency



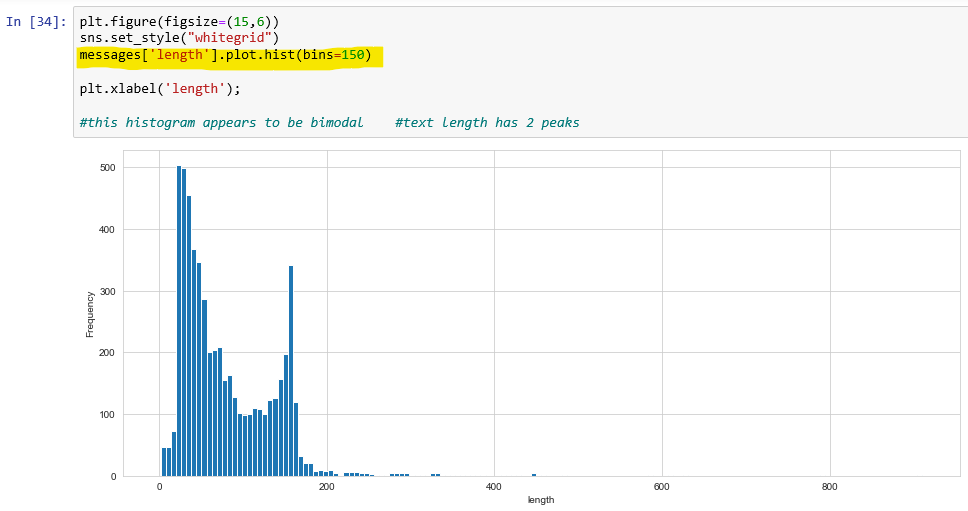
****



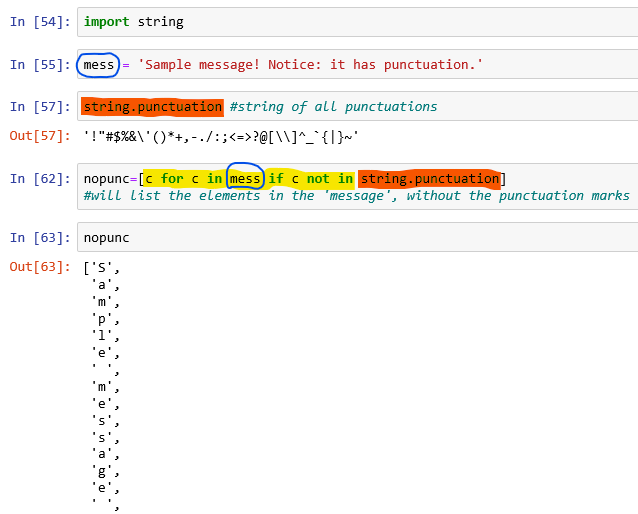
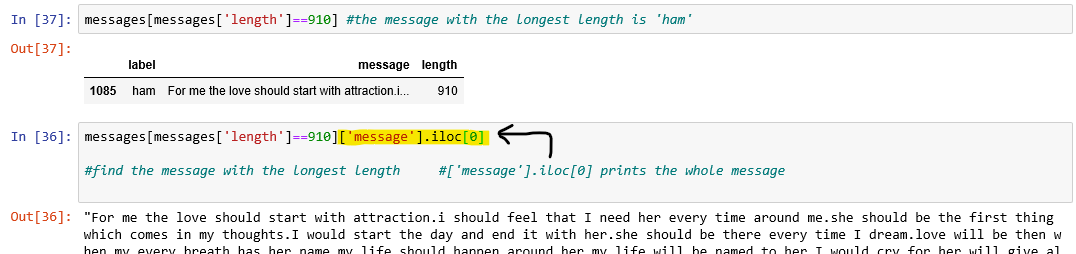
**… … … … … … … …**

**mess\_no**

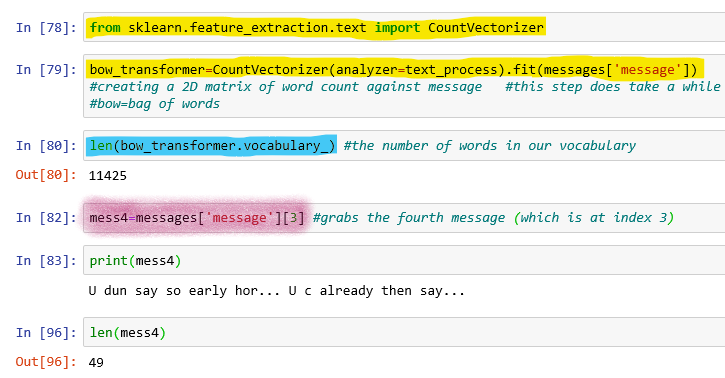
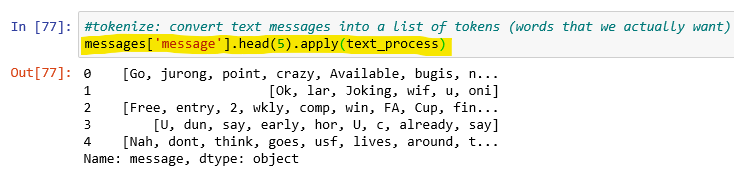
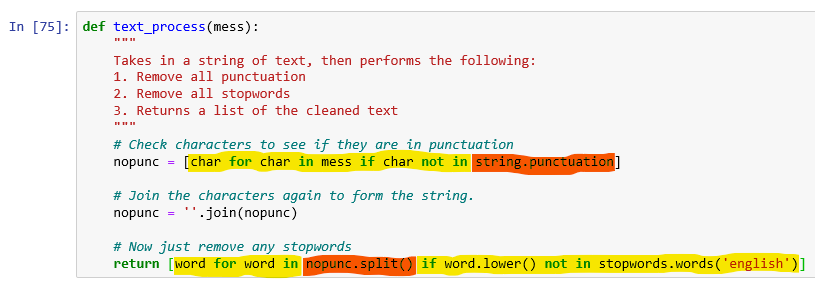
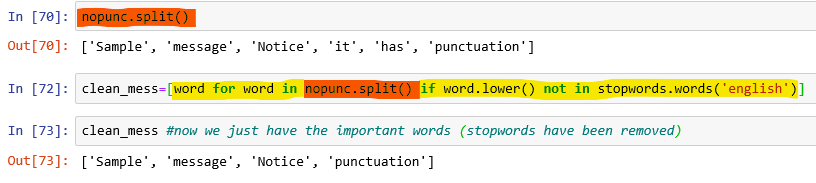
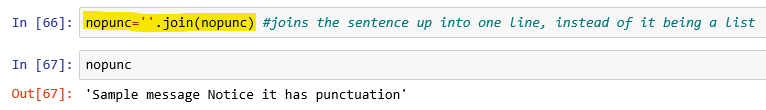
**message**



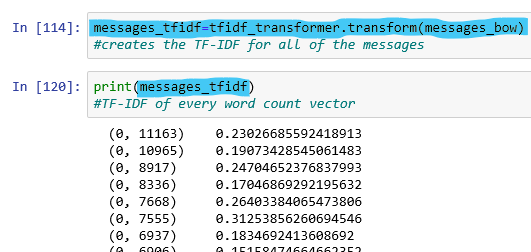
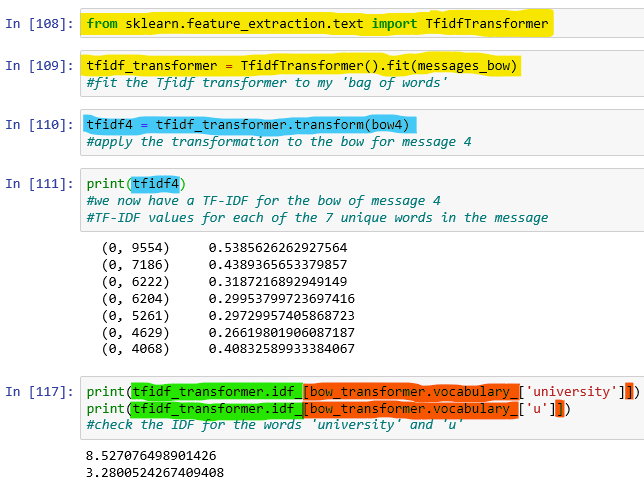
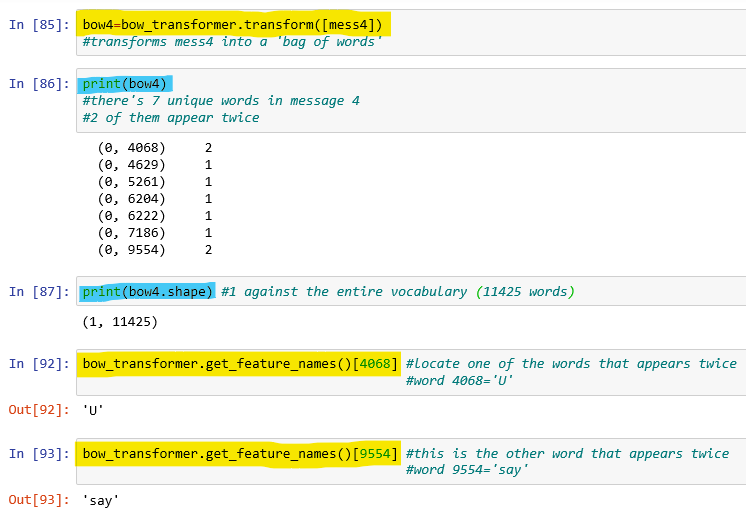
Histogram of the character lengths of the messages



**…**



**Vectorization**



**… … …………………**

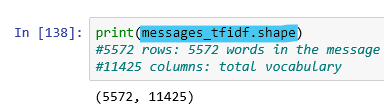
**TF-IDF**

**TF-IDF:** term frequency-inverse document frequency

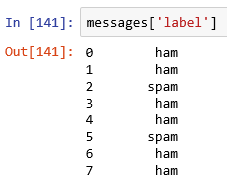
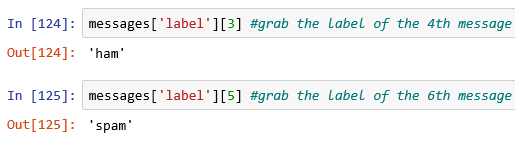
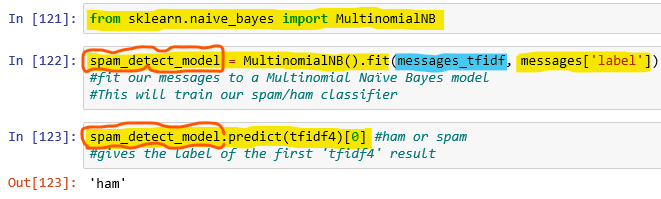
* TF-IDF weight is often used in information retrieval and text mining.
* This weight is a **statistical measure** used to evaluate how important a word is to a document in a collection or corpus.
* The importance increases proportionally to the number of times a word appears in the document. but is offset by the frequency of the word in the corpus.
* Variations of the tf-idf weighting scheme are often used by search engines as a central tool in **scoring** and **ranking** a document's relevance, given a user query.

**TF:** Term Frequency

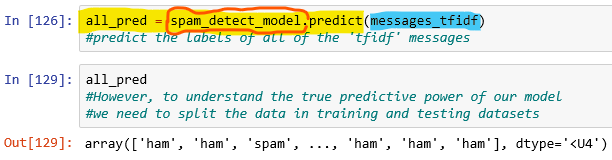
**IDF:** Inverse Document Frequency



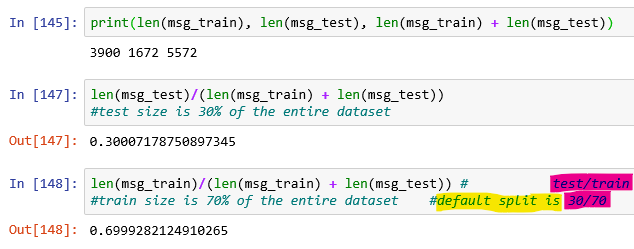
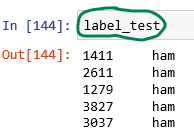
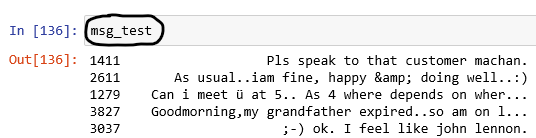
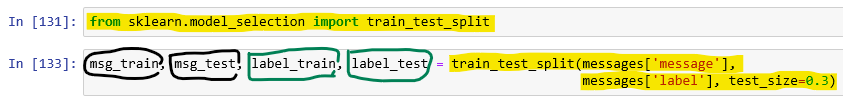
**Training Model**



**. …**



**Train-Test Split**



**Creating a data pipeline**

