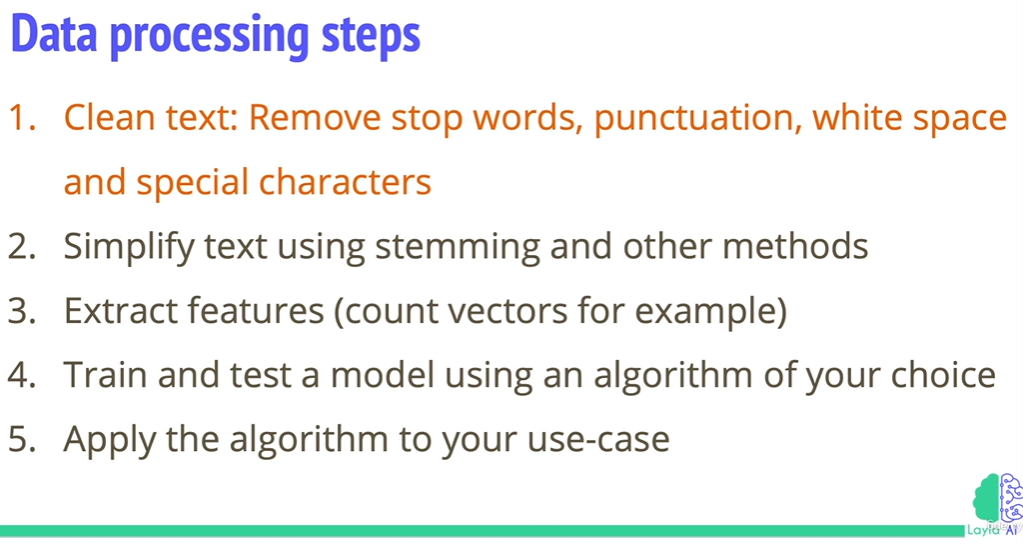
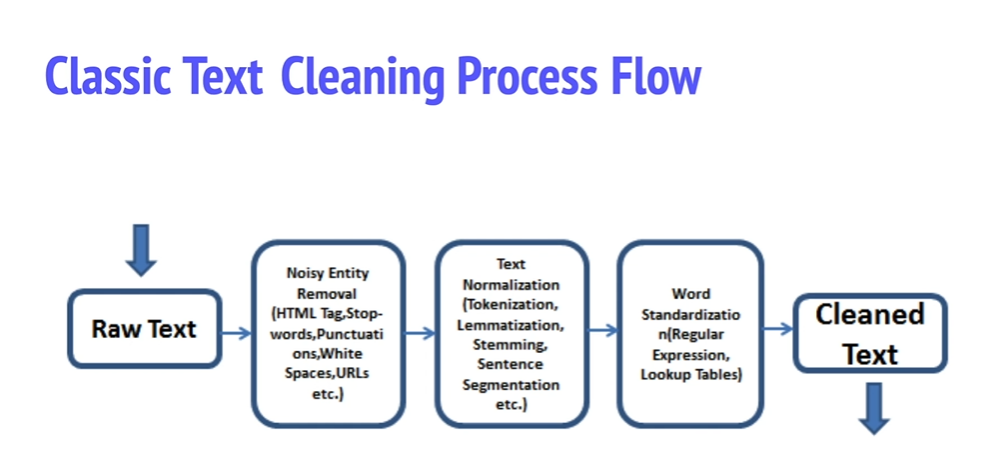


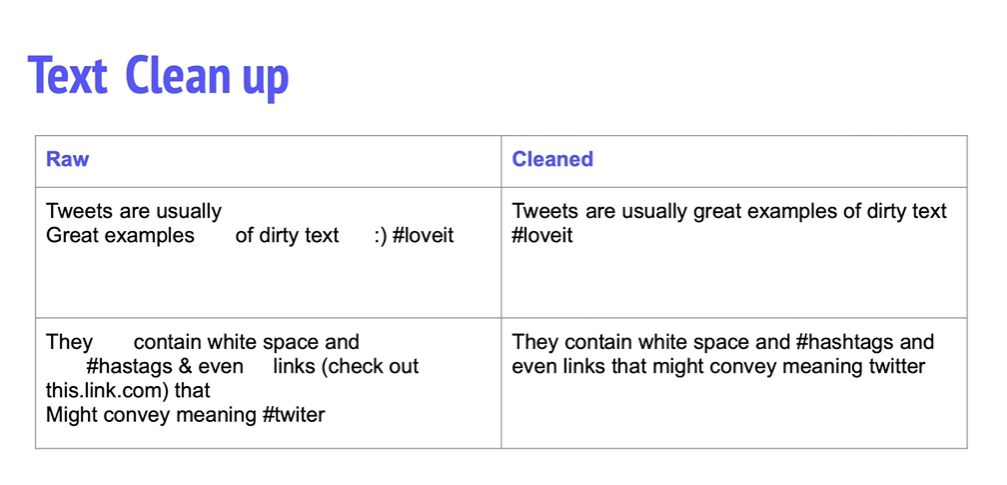


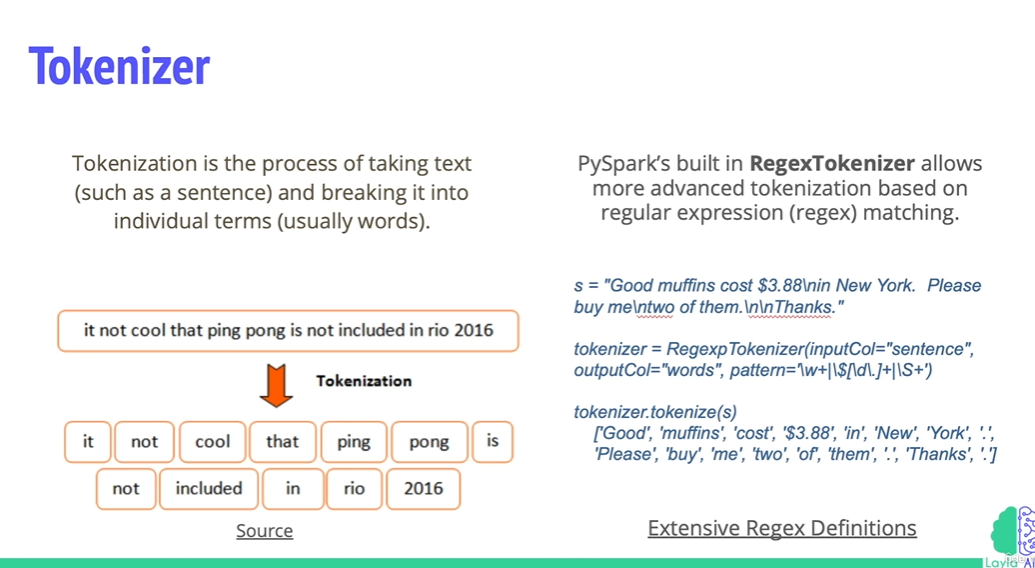
1. Clean text
2. Simplify text-stemming
3. Extract features – count vectors
4. Train and test -model



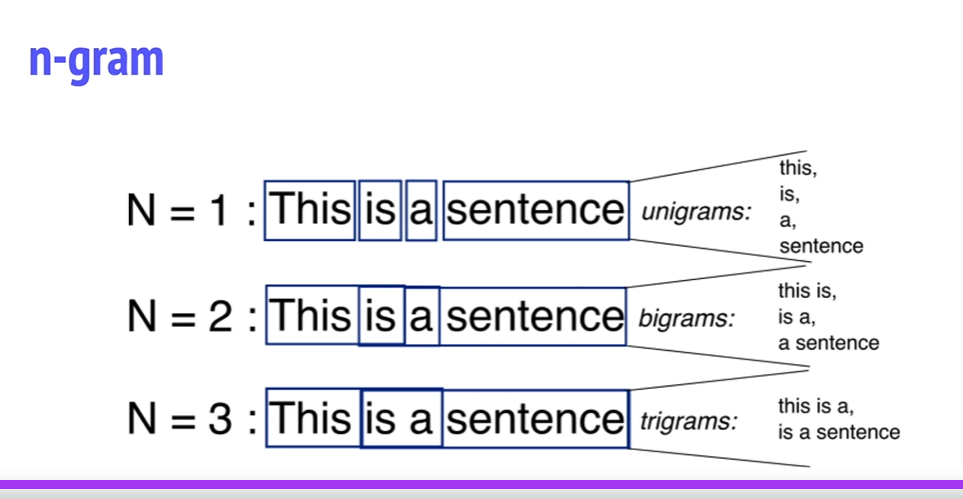










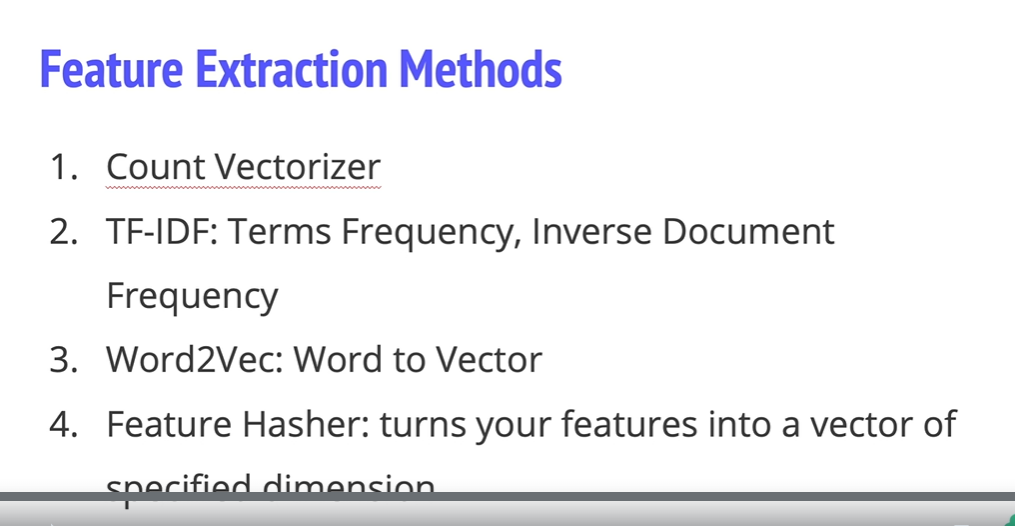


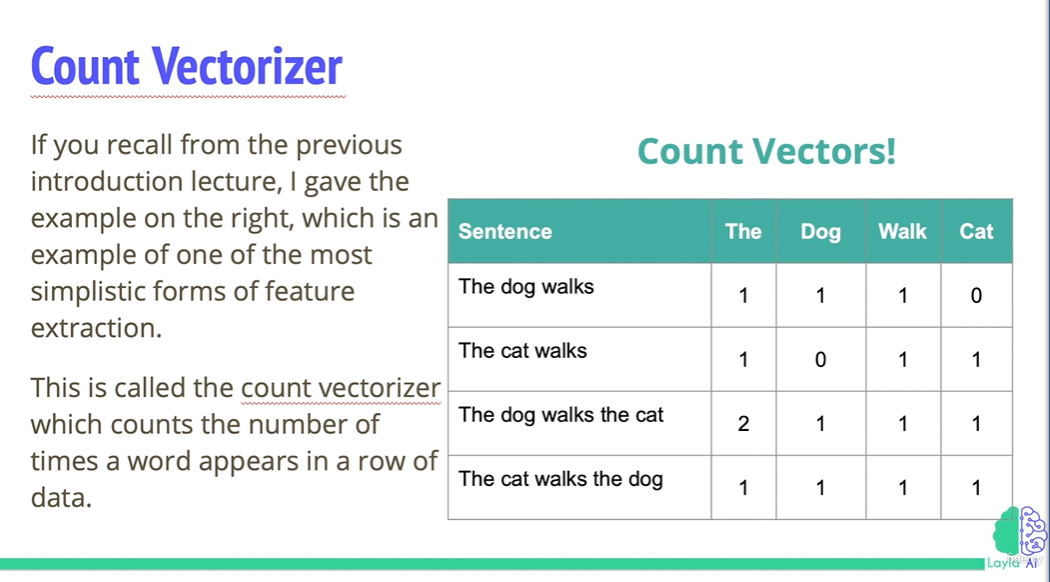
Notes

[Extracting, transforming and selecting features - Spark 3.3.1 Documentation (apache.org)](https://spark.apache.org/docs/latest/ml-features.html#feature-transformers)

[JohnSnowLabs/spark-nlp: State of the Art Natural Language Processing (github.com)](https://github.com/JohnSnowLabs/spark-nlp)

[Ultimate Regex Cheat Sheet - KeyCDN Support](https://www.keycdn.com/support/regex-cheat-sheet)



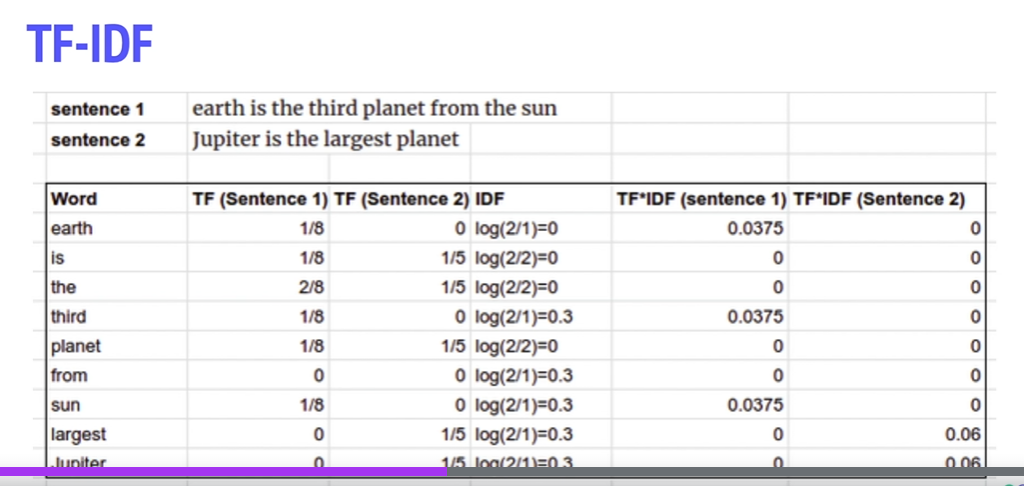




Measures the imp of key word by comparing freq of term by larger set of documents

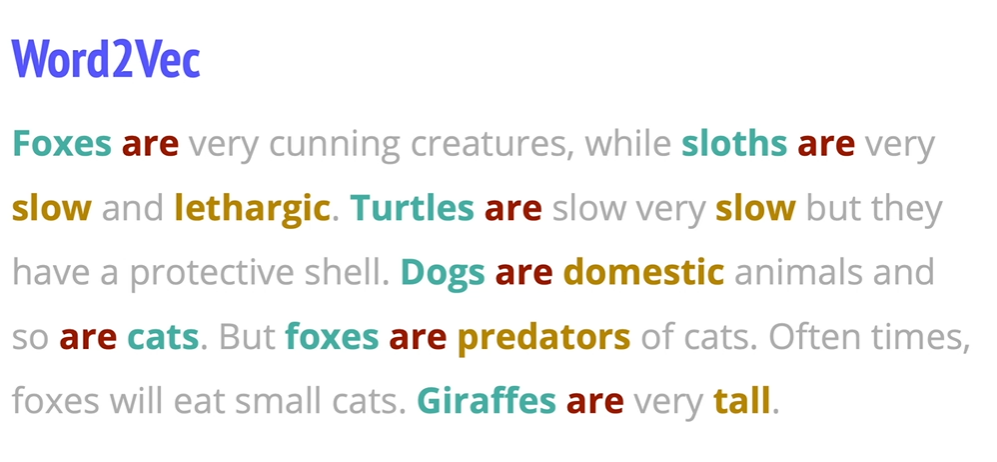
Basket ball : mentioned every document

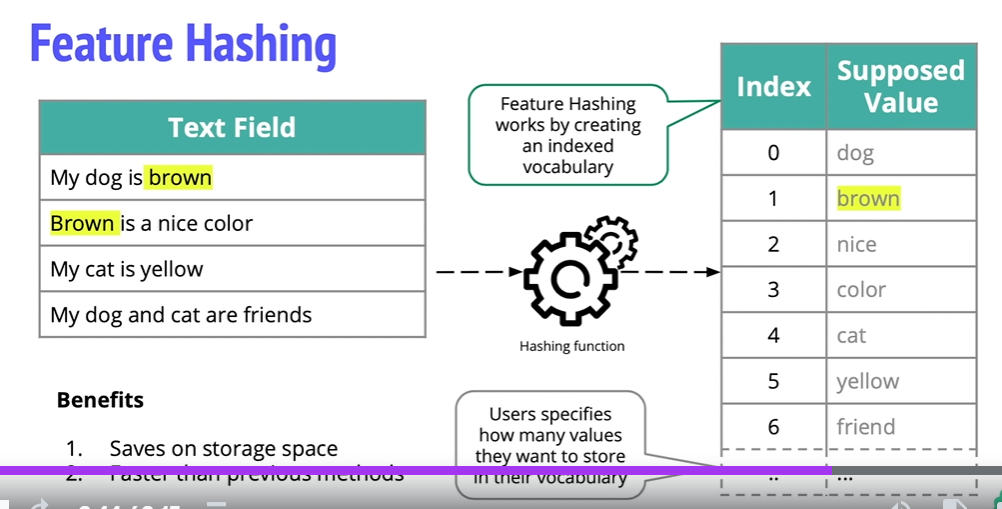
Less freq term: higher tfidf



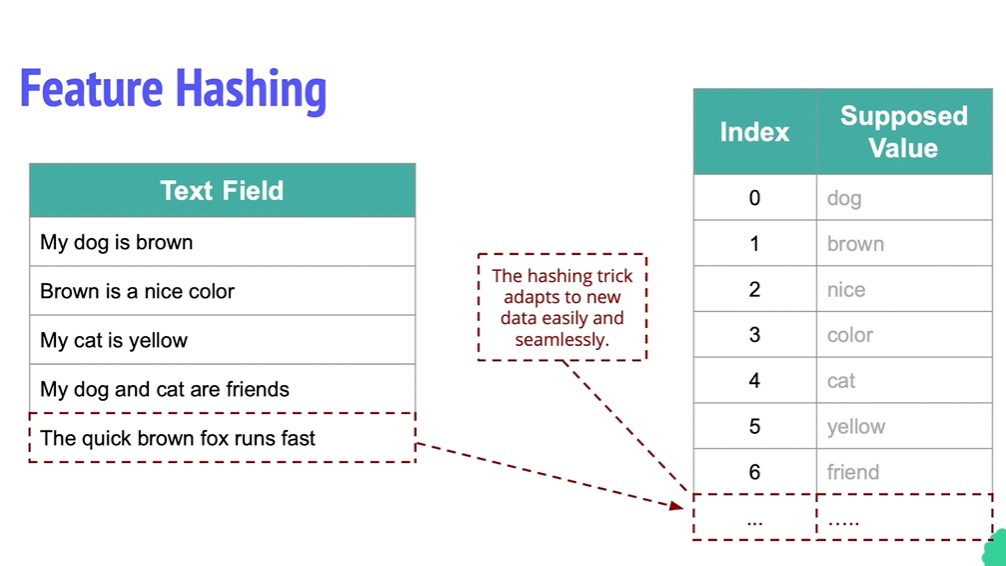








Turns dataframe of featue vector into small dimensions

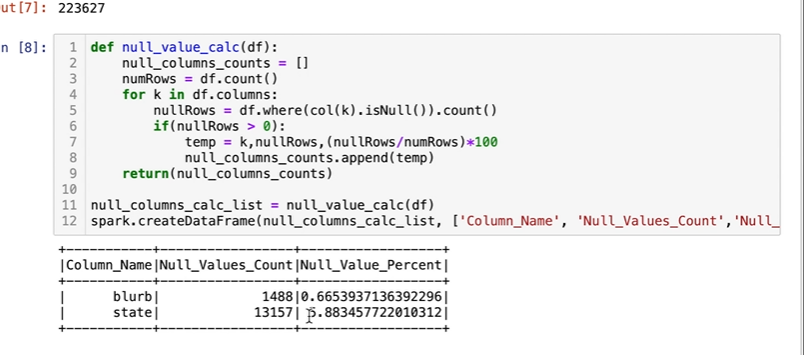


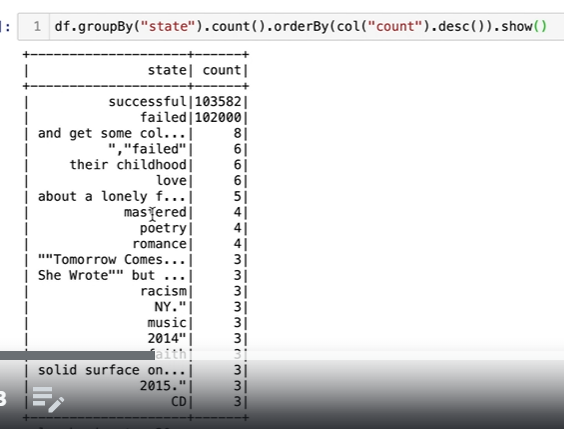
[Extracting, transforming and selecting features - Spark 3.3.1 Documentation (apache.org)](https://spark.apache.org/docs/latest/ml-features.html#feature-extractors)

Pipeline

[ML Pipelines - Spark 3.3.1 Documentation (apache.org)](https://spark.apache.org/docs/latest/ml-pipeline.html)

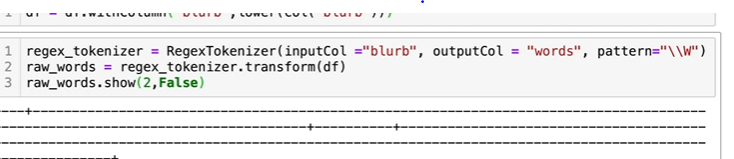
Null values :

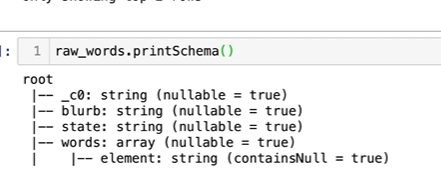


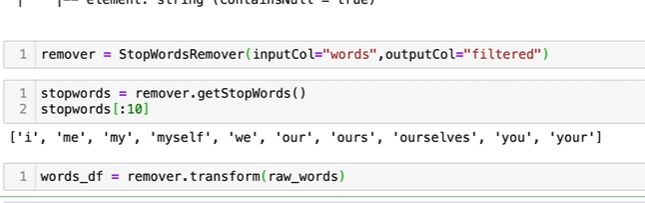


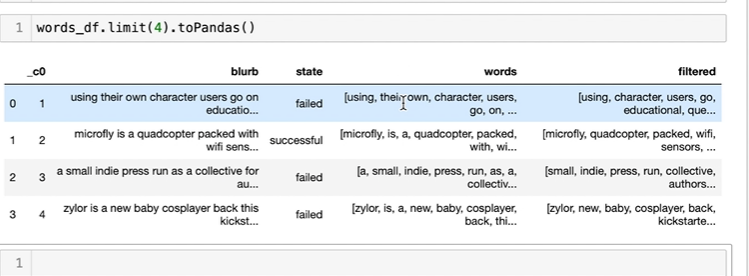


Tokenizer

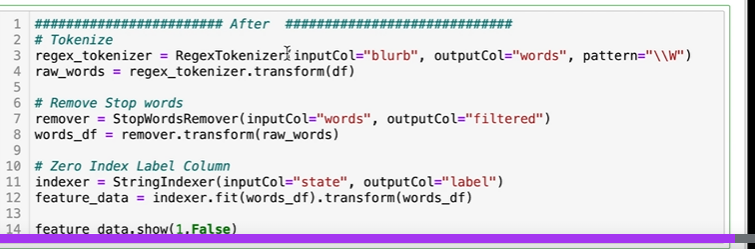


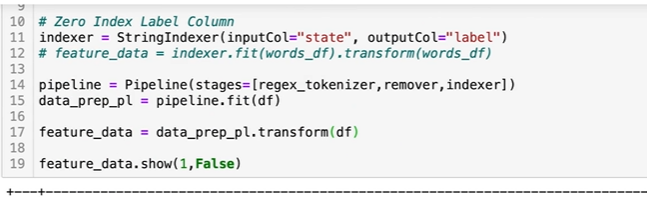


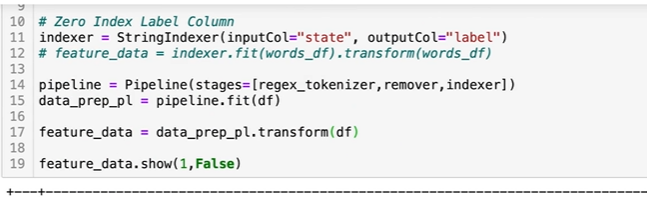


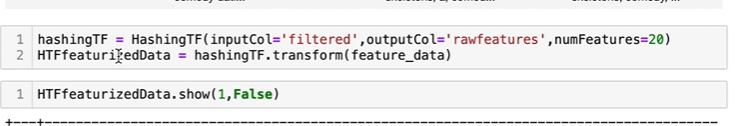


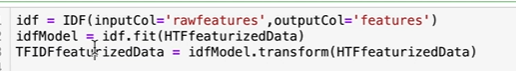


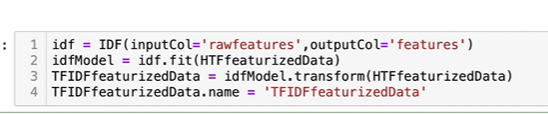


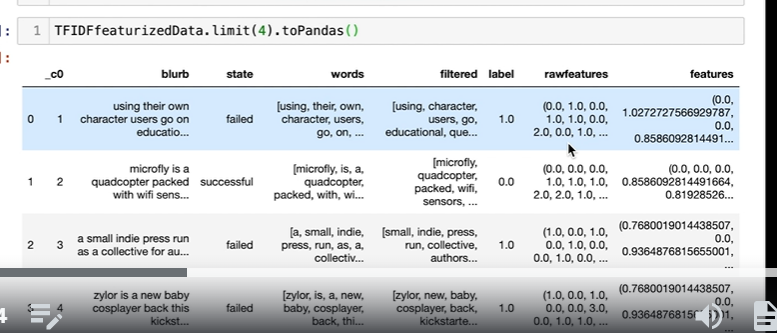


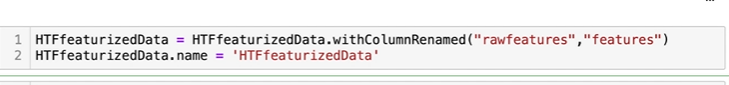


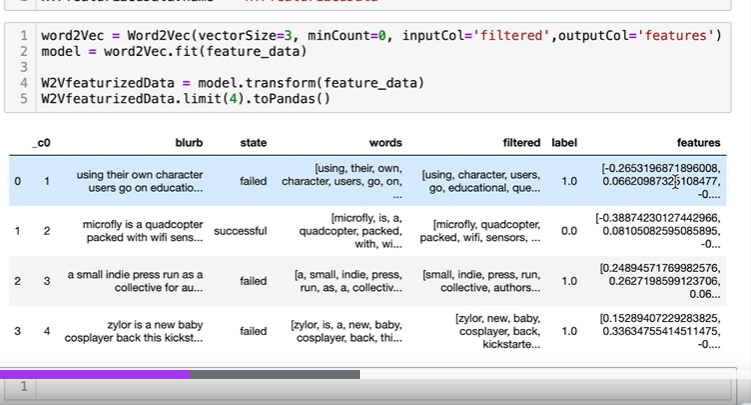


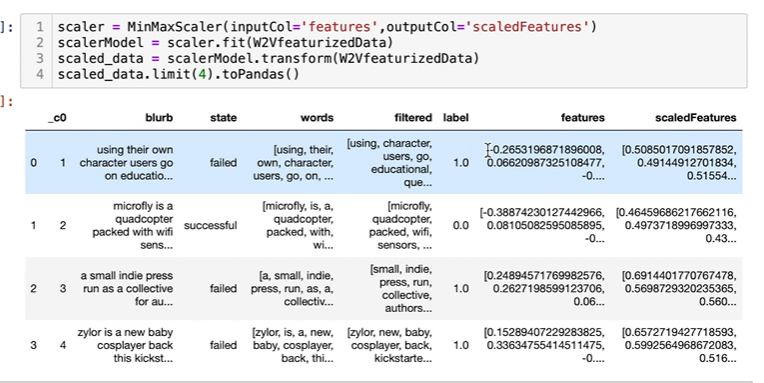




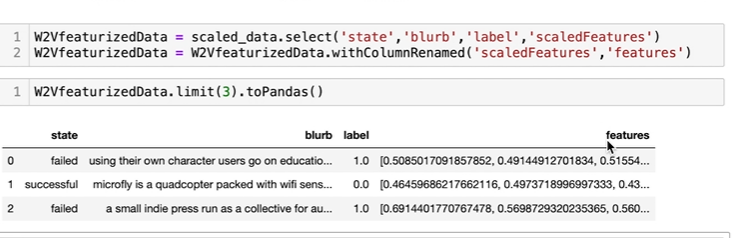












Now do xlassiffication