At the time of this recording Spark’s latest release - version 2.2.1 - supports two machine learning libraries spark.ml and spark.mllib. Both libraries are part of Spark's Machine Learning Library known as MLlib.

Spark.mllib is an RDD based library and it has been in maintenance mode since version 2.0. According to the current plans spark.ml, the Dataframe based API, will be feature complete by version 2.3 and then the older spark.mllib will be removed in Spark 3.0. So currently we might need to use a mixture of these two libraries but as time goes on you should focus on spark.ml as it’s becoming Spark’s standard machine learning library.

The term "Spark ML" is sometimes used to refer to the Spark Machine Learning library, which is officially called "MLlib". For further details see the [MLlib documentation](https://spark.apache.org/docs/latest/ml-guide.html" \t "_blank). In the following tutorials we'll use the DataFrame-based API.

# Feature Extraction

You can find more information about the different scaler and indexer options through the links below.

Scalers

* [Normalizer](https://spark.apache.org/docs/latest/ml-features.html#normalizer)
* [StandardScaler](https://spark.apache.org/docs/latest/ml-features.html#standardscaler)
* [MinMaxScaler](https://spark.apache.org/docs/latest/ml-features.html#minmaxscaler)
* [MaxAbsScale](https://spark.apache.org/docs/latest/ml-features.html#maxabsscaler)

Indexers

* [StringIndexer](https://spark.apache.org/docs/latest/ml-features.html#stringindexer)
* [IndexToString](https://spark.apache.org/docs/latest/ml-features.html#indextostring)
* [VectorIndexer](https://spark.apache.org/docs/latest/ml-features.html#vectorindexer)

To learn more about text processing you can read further in the [documentation](https://spark.apache.org/docs/latest/ml-features.html#feature-extractors).