

Промышленное машинное обучение на Spark



Содержание курса

- 01 Оборачиваем модель в сервис. Первая часть: Docker. Flask
- 02 Оборачиваем модель в сервис. Вторая часть: Requests. REST API
- 03 Распределенные вычисления. HDFS. MapReduce. Spark DataFrame
- 04 Погружение в среду Spark. RDD, SQL, Pandas API.
- 05 Генерация признаков. Spark feature engineering
- 06 Распределенное обучение моделей. Spark ML
- 07 Обработка и хранение текстовых данных и картинок.
Spark image processing. Spark NLP.
- 08 Обработка потоковых данных. Spark Streaming.

Что позволяет Spark.ML

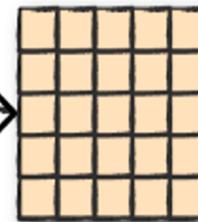
<http://spark.apache.org/docs/latest/ml-features.html>

Raw Data



Preprocessing
cleaning &
feature
engineering

Clean &
Structured



Modeling &/or
analytical
techniques

Tuning



Evaluation



Success!

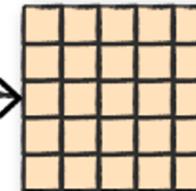
Structured
APIs
Raw Data



Preprocessing
cleaning &
feature
engineering

Transformers
& Estimators

Clean &
Structured



Estimators
& Models

Modeling &/or
analytical
techniques

Pipelines
Cross
Validation



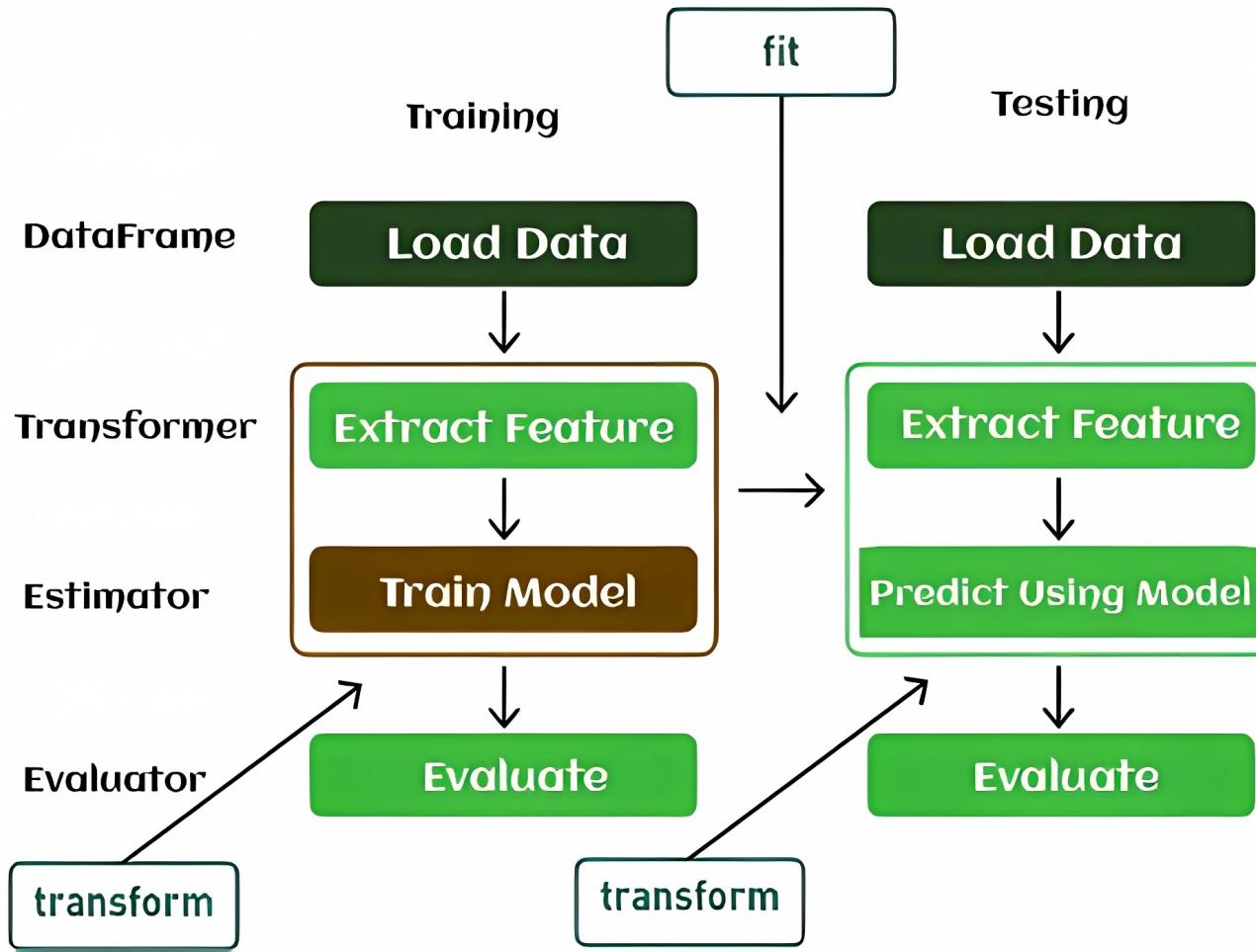
Tuning

Evaluation

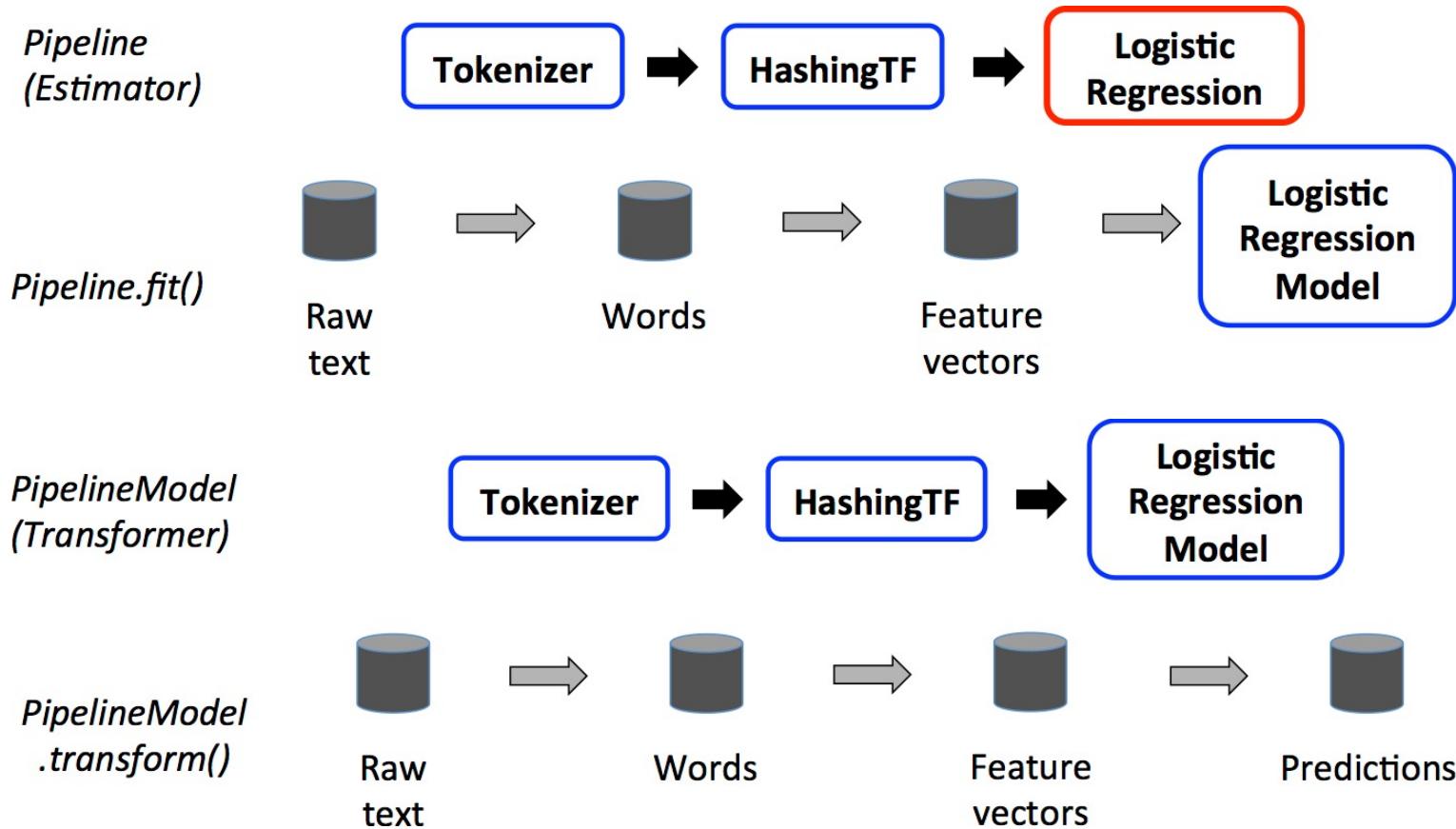
Success!

Evaluators
Metrics

Spark ML Workflow



Pipeline



Transformers

- Tokenizer
- StopWordsRemover
- n-gram
- Binarizer
- PCA
- PolynomialExpansion
- Discrete Cosine Transform (DCT)
- StringIndexer
- IndexToString
- OneHotEncoder
- VectorIndexer
- Interaction

- Normalizer
- StandardScaler
- RobustScaler
- MinMaxScaler
- MaxAbsScaler
- Bucketizer
- ElementwiseProduct
- SQLTransformer
- VectorAssembler
- VectorSizeHint
- QuantileDiscretizer
- Imputer

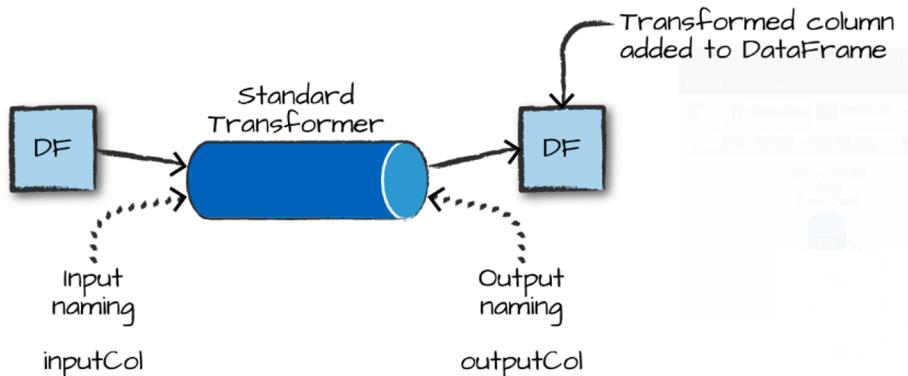


Figure 24-3. A standard transformer

Estimators

- LogisticRegression
- DecisionTreeClassifier
- RandomForestClassifier
- NaiveBayes
- OneVsRest
- MultilayerPerceptronClassifier
- KMeans
- LDA
- GaussianMixture
- ALS
- DecisionTreeRegressor
- LinearRegression
- RandomForestRegressor

Evaluators

- `BinaryClassificationEvaluator`
- `RegressionEvaluator`
- `MulticlassClassificationEvaluator`
- `MultilabelClassificationEvaluator`
- `ClusteringEvaluator`