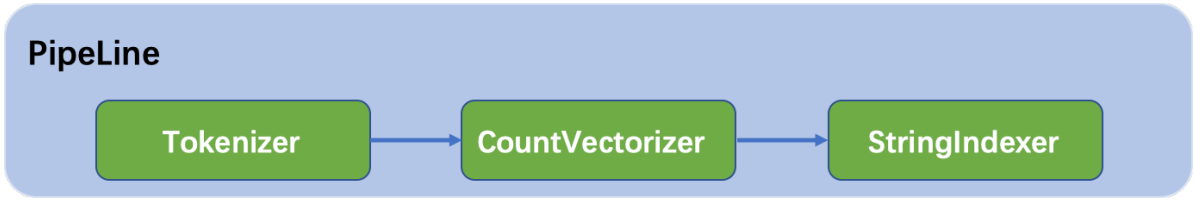
Yuan Gao z5239220 Report

1. Evaluation of stacking model on the test data

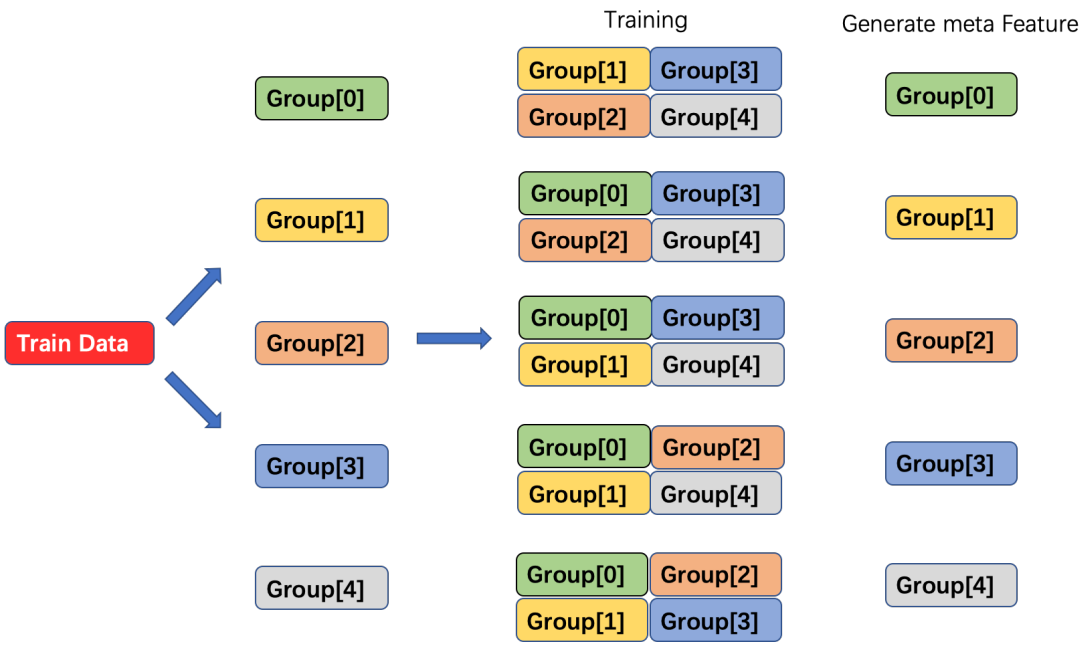
1.1 Build a Preprocessing Pipeline

- Process `train_data`, use `pipline` increase `Tokenizer`, `CountVectorizer` and `StringIndexer` columns

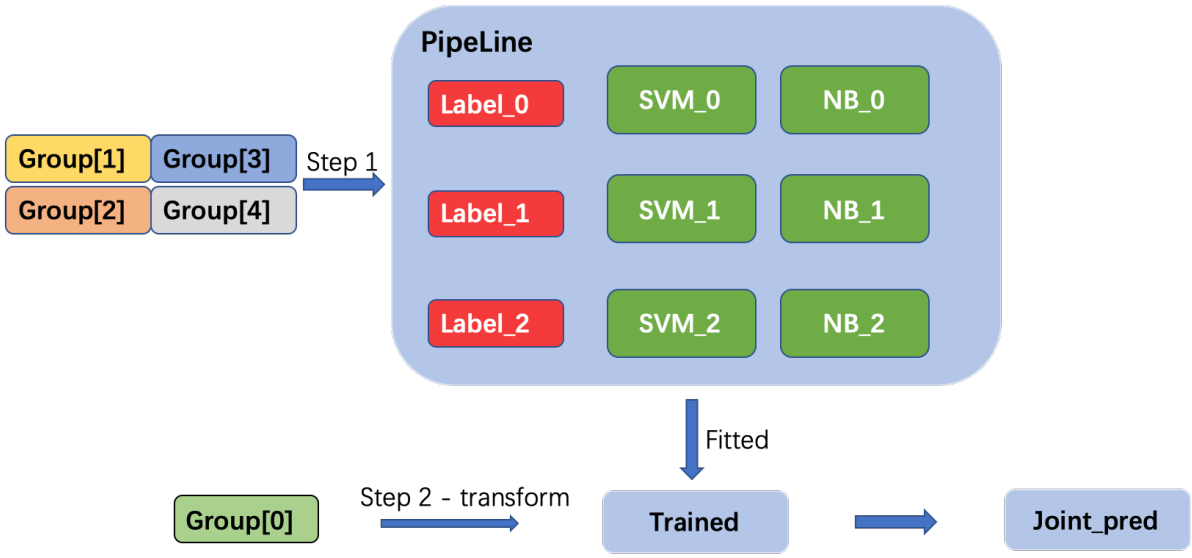


1.2 Generate Meta Features for Training

- Generate Meta Features:



- For every Feature:



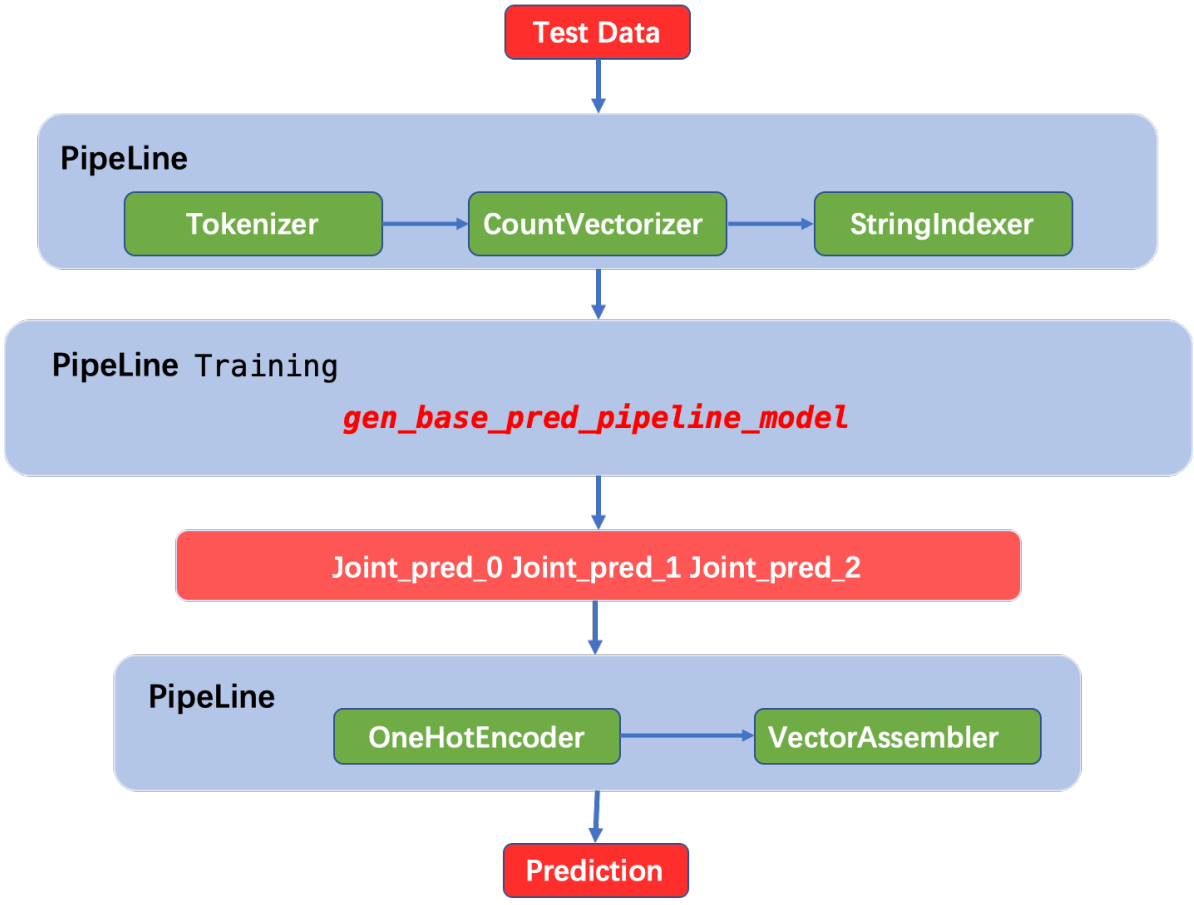
- Combine 9 Column:

```
+---------+----------+ +---------+----------+ +---------+----------+ +------------+ +------------+ +------------+
|nb_pred_0|svm_pred_0| |nb_pred_1|svm_pred_1| |nb_pred_2|svm_pred_2| |joint_pred_0| |joint_pred_1| |joint_pred_2|
+---------+----------+ +---------+----------+ +---------+----------+ +------------+ +------------+ +------------+
|      0.0|       0.0| |      0.0|       0.0| |      0.0|       1.0| |         0.0| |         0.0| |         1.0|
|      0.0|       0.0| |      1.0|       1.0| |      0.0|       0.0| |         0.0| |         3.0| |         0.0|
|      0.0|       1.0| |      0.0|       0.0| |      0.0|       0.0| |         1.0| |         0.0| |         0.0|
|      0.0|       0.0| |      1.0|       1.0| |      0.0|       0.0| |         0.0| |         3.0| |         0.0|
|      1.0|       1.0| |      0.0|       0.0| |      0.0|       0.0| |         3.0| |         0.0| |         0.0|
+---------+----------+ +---------+----------+ +---------+----------+ +------------+ +------------+ +------------+
```

1.3 Obtain the prediction for the test data

- Prediction

> Using meta Features to predict



2. How to improve the performance

In `pyspark.ml.feature`, there are `RegexTokenizer` function can address punctuation. We can use this function to reduce the effect of nonsense character. Moreover, we can add Decision Tree model geting `dt_pred_0, dt_pred_1, dt_pred_2` and compose with `NB` and `SVM`. Nine columns geting meta feature can improve the performance of the stacking model.