

25 YEARS ANNIVERSARY  
SOICT

HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY



HA NOI UNIVERSITY OF SCIENCE AND TECHNOLOGY  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

# Chapter 9

# Big data analytics

Spark ML

# Machine Learning Library (MLlib)

- 2 packages
  - spark.mllib
  - spark.ml
- ML algorithms
  - Common learning algorithms such as classification, regression, clustering and collaborative filtering
- Featurization
  - Feature extraction, transformations, dimensionality reduction and selection
- Utilities
  - Linear algebra, statistics, data handling, ...

# ML: Transformer

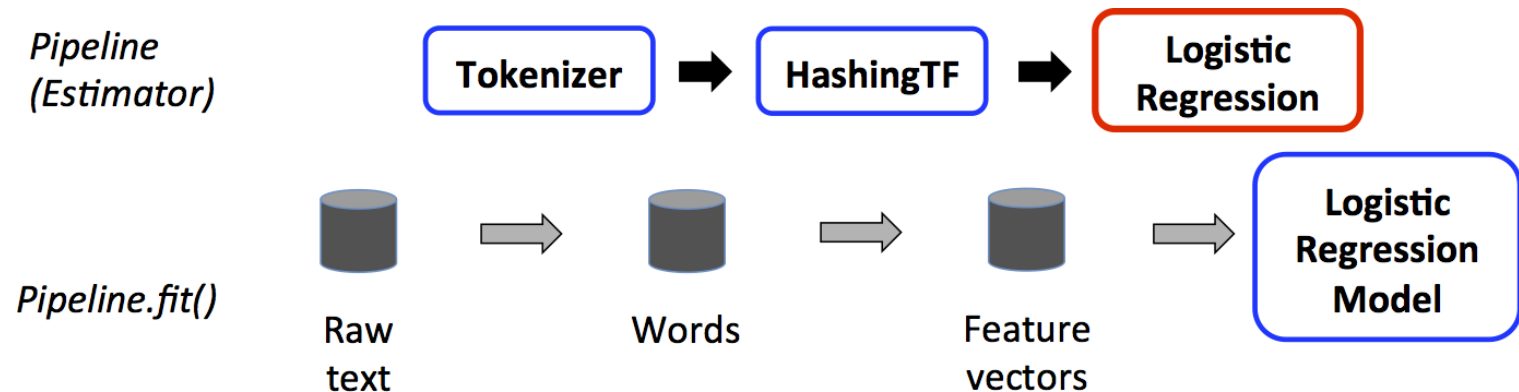
- A Transformer is a class which can transform a DataFrame into another DataFrame
- A Transformer implements transform()
- Examples
  - HashisngTF
  - LogisticRegressionModel
  - Binarizer

# ML: Estimator

- An Estimator is a class which can take a DataFrame and return a Transformer
- An Estimator implements fit()
- Examples
  - LogisticRegression
  - StandardScaler
  - Pipeline

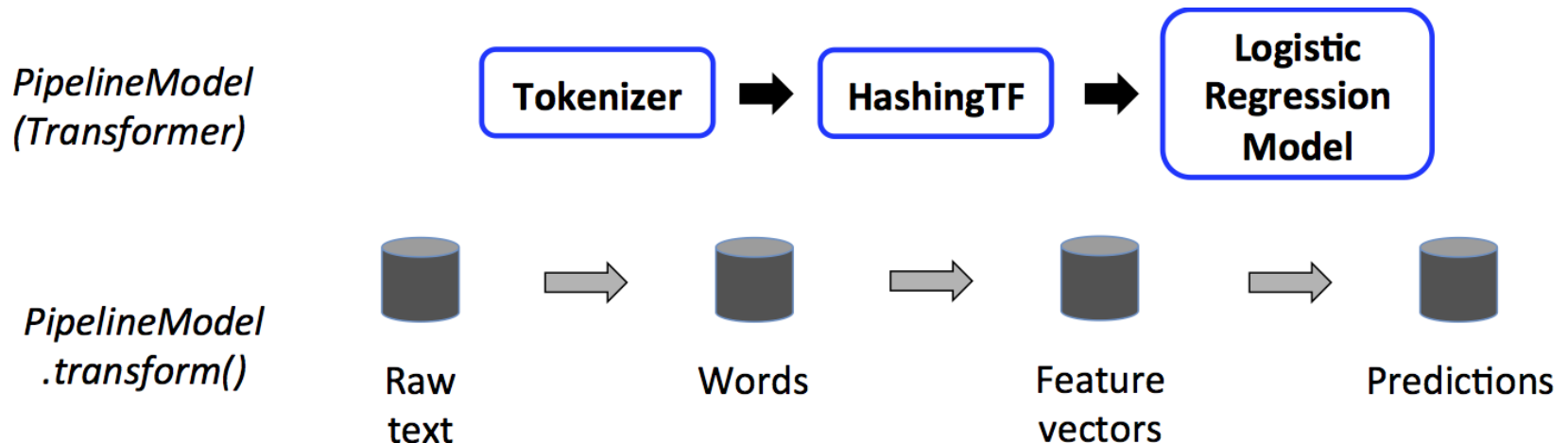
# ML: Pipeline

- A Pipeline is an estimator that specified as a sequence of stages and each stage can be either estimators or transformers.



# ML: PipelineModel

- After a `Pipeline.fit()` runs, it produces a `PipelineModel`. This `PipelineModel` is used at test time.



# Demo



# References

- Meng, Xiangrui, et al. "Mllib: Machine learning in apache spark." *The Journal of Machine Learning Research* 17.1 (2016): 1235-1241.
- Pentreath, Nick. *Machine learning with spark*. Packt Publishing Ltd, 2015.



25  
SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG  
SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Thank you  
for your  
attention!!!



[soict.hust.edu.vn/](http://soict.hust.edu.vn/)



[fb.com/groups/soict](https://fb.com/groups/soict)

