ĐẠI HỌC
BÁCH KHOA

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

Chương 9

Phân tích dữ liệu lớn

Tia lửa ML

Thư viện học máy (MLlib)

- 2 gói •
  - spark.mllib •
  - spark.ml •

Thuật toán ML • Các
  - thuật toán học tập phổ biến như phân loại, hồi quy, phân cụm và lọc cộng tác

- Đặc điểm hóa • Trích
  - xuất đặc điểm, chuyển đổi, giảm chiều và lựa chọn

- Tiện ích
  - Đại số tuyến tính, thống kê, xử lý dữ liệu, .

ML: Máy biến áp

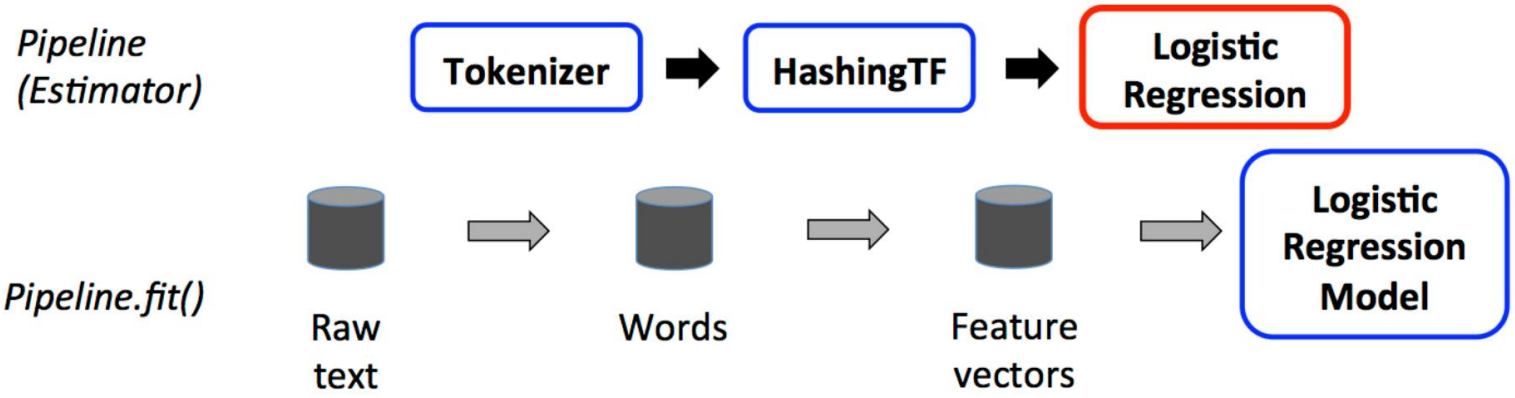
- Một Transformer là một lớp có thể biến đổi một DataFrame vào một DataFrame khác
- Một Transformer thực hiện transform() • Ví dụ • HashisngTF •
  - LogisticRegressionModel •
  - Binarizer

ML: Ước tính

- Một Estimator là một lớp có thể lấy một DataFrame và trả về một Transformer
- Một trình ước tính thực hiện fit()
- Ví dụ
  - Hồi quy logistic
  - ChuẩnScaler
  - Đường ống

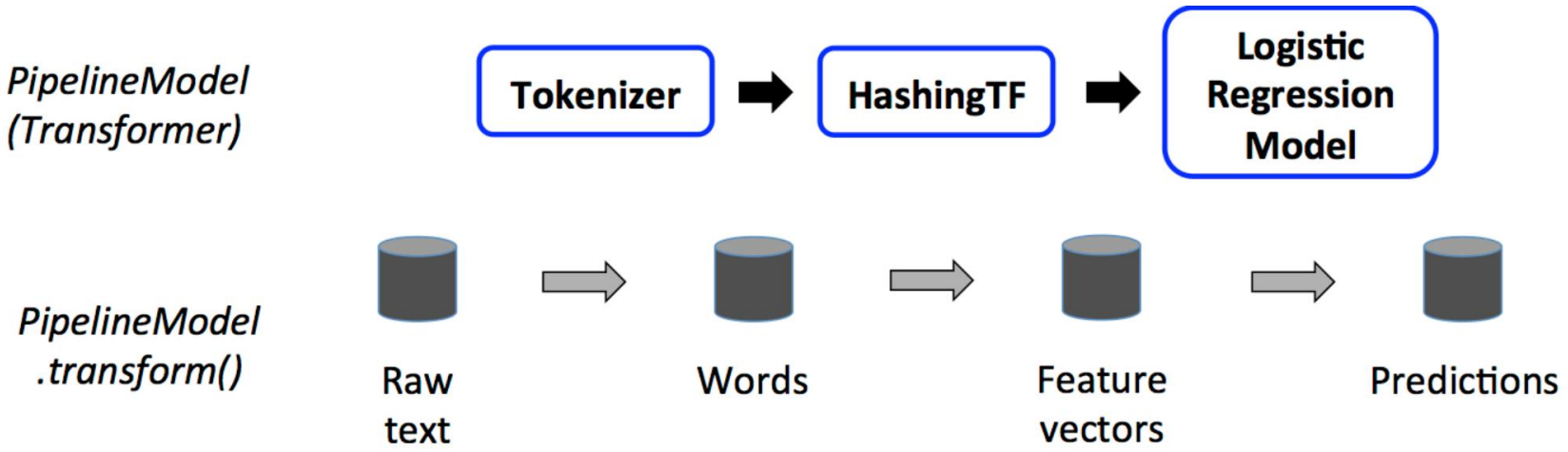
ML: Đường ống

- Pipeline là một bộ ước tính được chỉ định như một chuỗi các giai đoạn và mỗi giai đoạn có thể là bộ ước tính hoặc bộ chuyển đổi.



ML: Mô hình đường ống

- Sau khi Pipeline.fit() chạy, nó sẽ tạo ra PipelineModel.
  PipelineModel này được sử dụng tại thời điểm thử nghiệm.



Thử nghiệm

Tài liệu tham khảo

- Meng, Xiangrui, et al. "Mllib: Học máy trong apache tia lửa." Tạp chí nghiên cứu máy học 17.1 (2016): 1235-1241.

- Pentreath, Nick. Học máy với spark. Packt Publishing Công ty TNHH, 2015.

ĐẠI HỌC BÁCH KHOA

25 YEARS ANNIVERSARY SOICT

VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGY

Cảm ơn sự chú ý của bạn!!!