CSE4/587 Data-intensive Computing Spring 2017
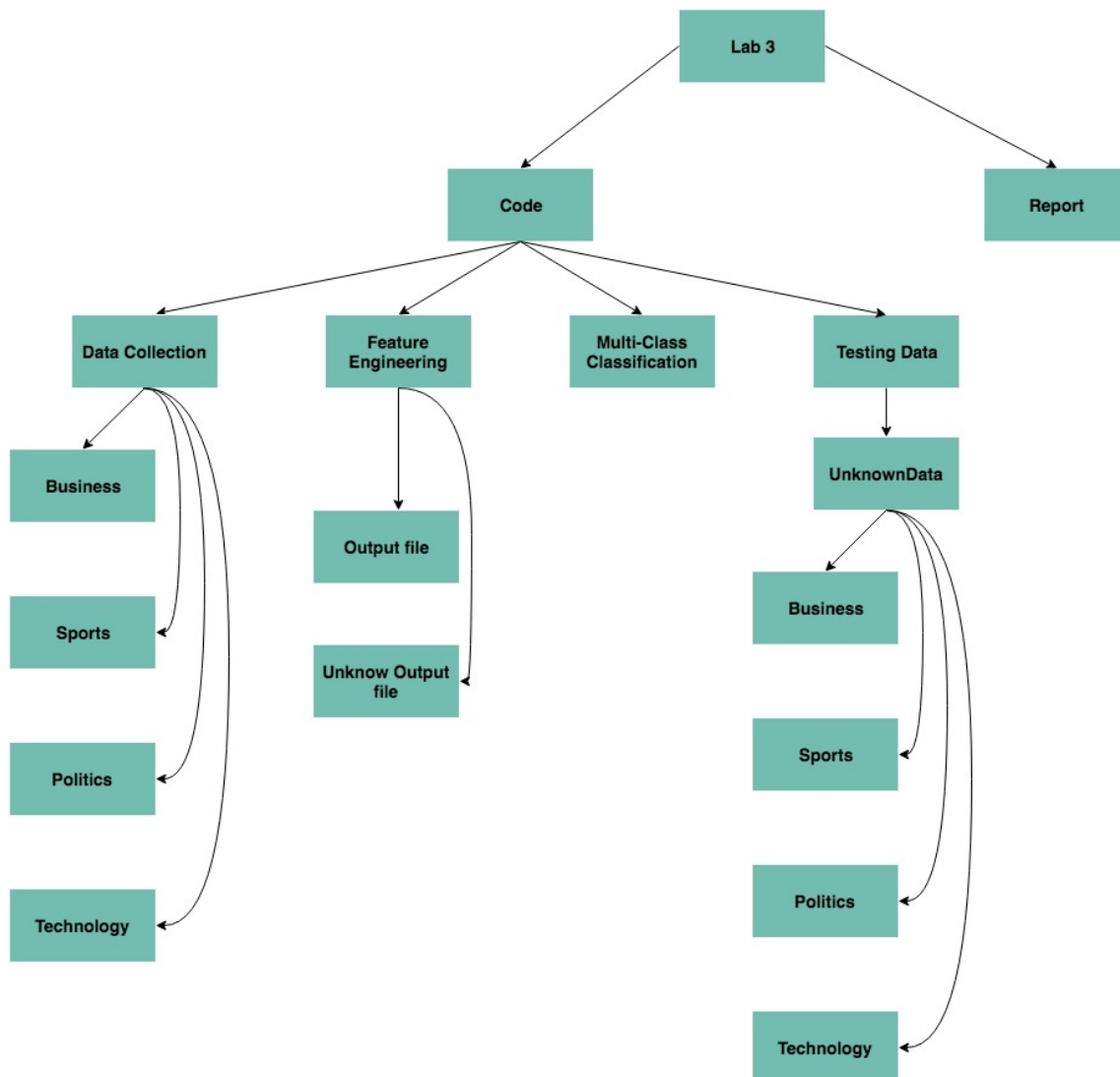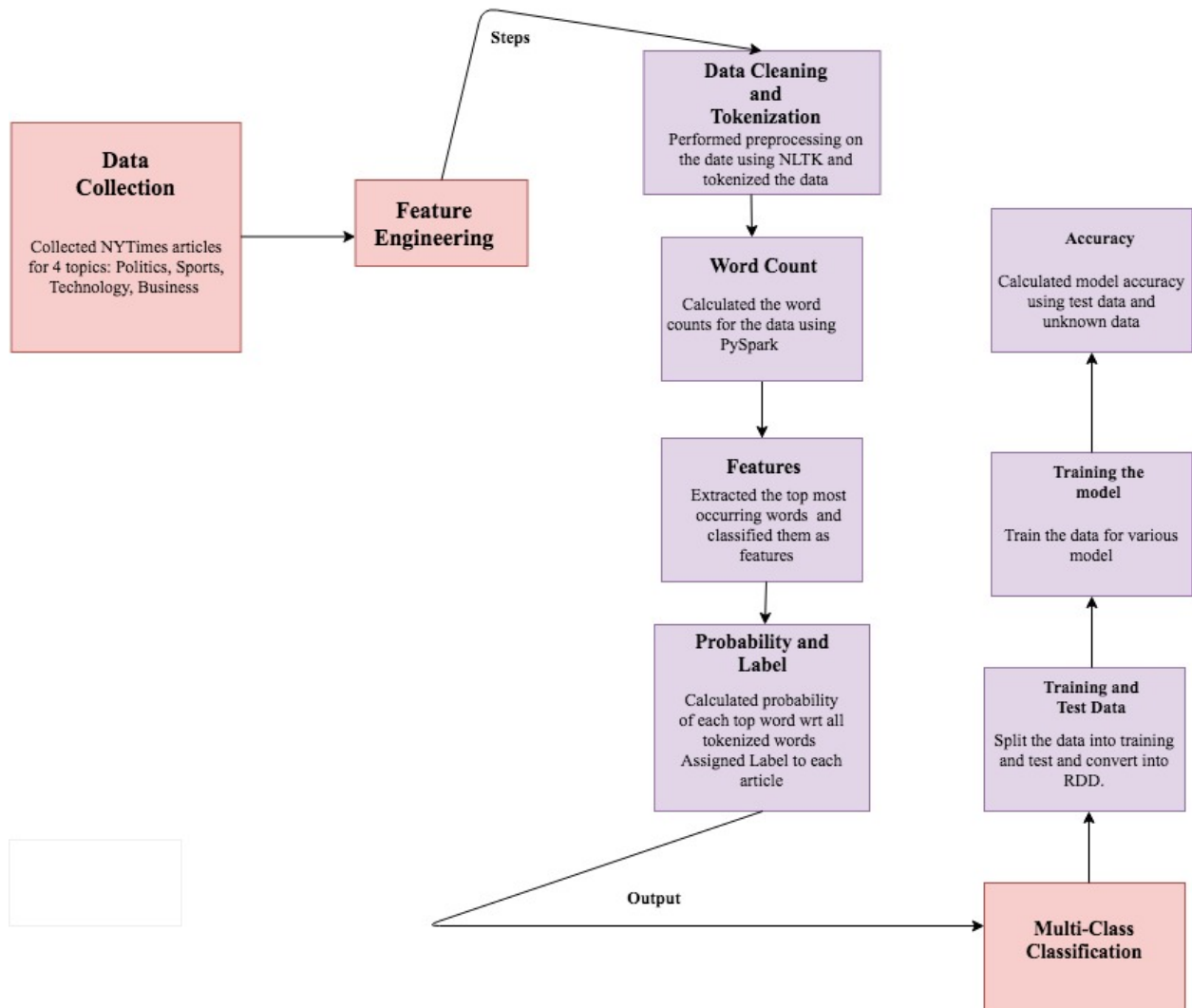## LAB 3: DATA ANALYTICS PIPELINE USING APACHE SPARK: B. RAMAMURTHY

PRIYA MURTHY:50248887
SHEFALI SHARMA: 50247677

Directory Structure:

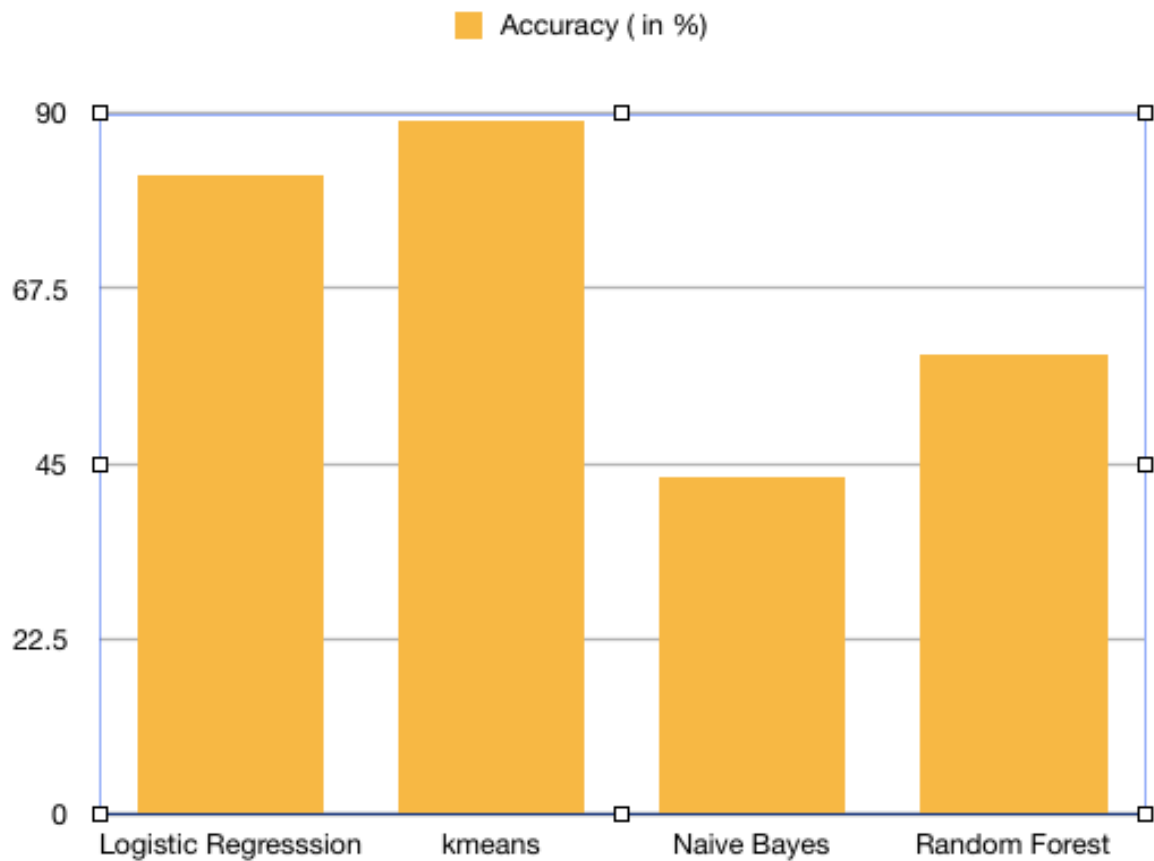**DATA PIPELINE:**

Data Collection

Collected NYTimes articles for 4 topics: Politics, Sports, Technology, Business

Feature Engineering

Steps

Data Cleaning and Tokenization

Performed preprocessing on the date using NLTK and tokenized the data

Word Count

Calculated the word counts for the data using PySpark

Features

Extracted the top most occurring words and classified them as features

Probability and Label

Calculated probability of each top word wrt all tokenized words Assigned Label to each article

Accuracy

Calculated model accuracy using test data and unknown data

Training the model

Train the data for various model

Training and Test Data

Split the data into training and test and convert into RDD.

Output

Multi-Class Classification

Model Accuracy for Test data:

Accuracy for Test Data

Model	Accuracy (in %)
Logistic Regresssion	82
kmeans	89
Naive Bayes	43.21
Random Forest	59

■ Accuracy (in %)

Model Accuracy for Unknown data:

Accuracy for Unknown Data

Model	Accuracy (in %)
Logistic Regresssion	94.65
kmeans	89.62
Naive Bayes	40.9
Random Forest	31.1

Bar chart with legend "Accuracy (in %)". Y-axis scaled 0 to 100. Logistic Regresssion ≈ 94.65, kmeans ≈ 89.62, Naive Bayes ≈ 40.9, Random Forest ≈ 31.1.