DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
UNIVERSITY AT BUFFALO

# CSE 587 LAB 3

## DATA ANALYTICS PIPELINE USING APACHE SPARK

Shivam Sahu
(#50247673)

# 1. Apache Spark with Titanic data analysis

```
root
 |-- PassengerId: string (nullable = true)
 |-- Survived: double (nullable = true)
 |-- Pclass: string (nullable = true)
 |-- FirstName: string (nullable = true)
 |-- Name: string (nullable = true)
 |-- Sex: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- SibSp: double (nullable = true)
 |-- Parch: double (nullable = true)
 |-- Ticket: string (nullable = true)
 |-- Fare: double (nullable = true)
 |-- Cabin: string (nullable = true)
 |-- Embarked: string (nullable = true)
 |-- Mark: string (nullable = false)

Train Data Number of Row: 637

Validate Data Number of Row: 254

Test Data Number of Row: 418

{'RandomForest': 0.8498852626211115,
'LogisticRegression': 0.8287225905150437,
'DecistionTree': 0.5850012748597654}
```

# Logistic Regression

### Test Data:

Accuracy : -   **77.7366 %**

Confusion Matrix: -

| | Sports | Poltics | Business | Education |
|---|---|---|---|---|
| Sports | 10 | 1 | 0 | 2 |
| Poltics | 0 | 9 | 0 | 3 |
| Business | 0 | 2 | 7 | 2 |
| Education | 0 | 0 | 0 | 6 |

### Unknown Set Data:

Accuracy : -    **77.6252 %**

Confusion Matrix: -

| | Sports | Poltics | Business | Education |
|---|---|---|---|---|
| Sports | 8 | 2 | 0 | 0 |
| Poltics | 0 | 9 | 0 | 1 |
| Business | 0 | 4 | 5 | 1 |
| Education | 0 | 1 | 0 | 9 |

# Naive Bayes

**Test Data**:

Accuracy : -  **71.006 %**

Confusion Matrix: -

| | Sports | Poltics | Business | Education |
|---|---|---|---|---|
| Sports | 7 | 1 | 0 | 5 |
| Poltics | 0 | 8 | 1 | 3 |
| Business | 0 | 1 | 9 | 1 |
| Education | 1 | 0 | 0 | 5 |

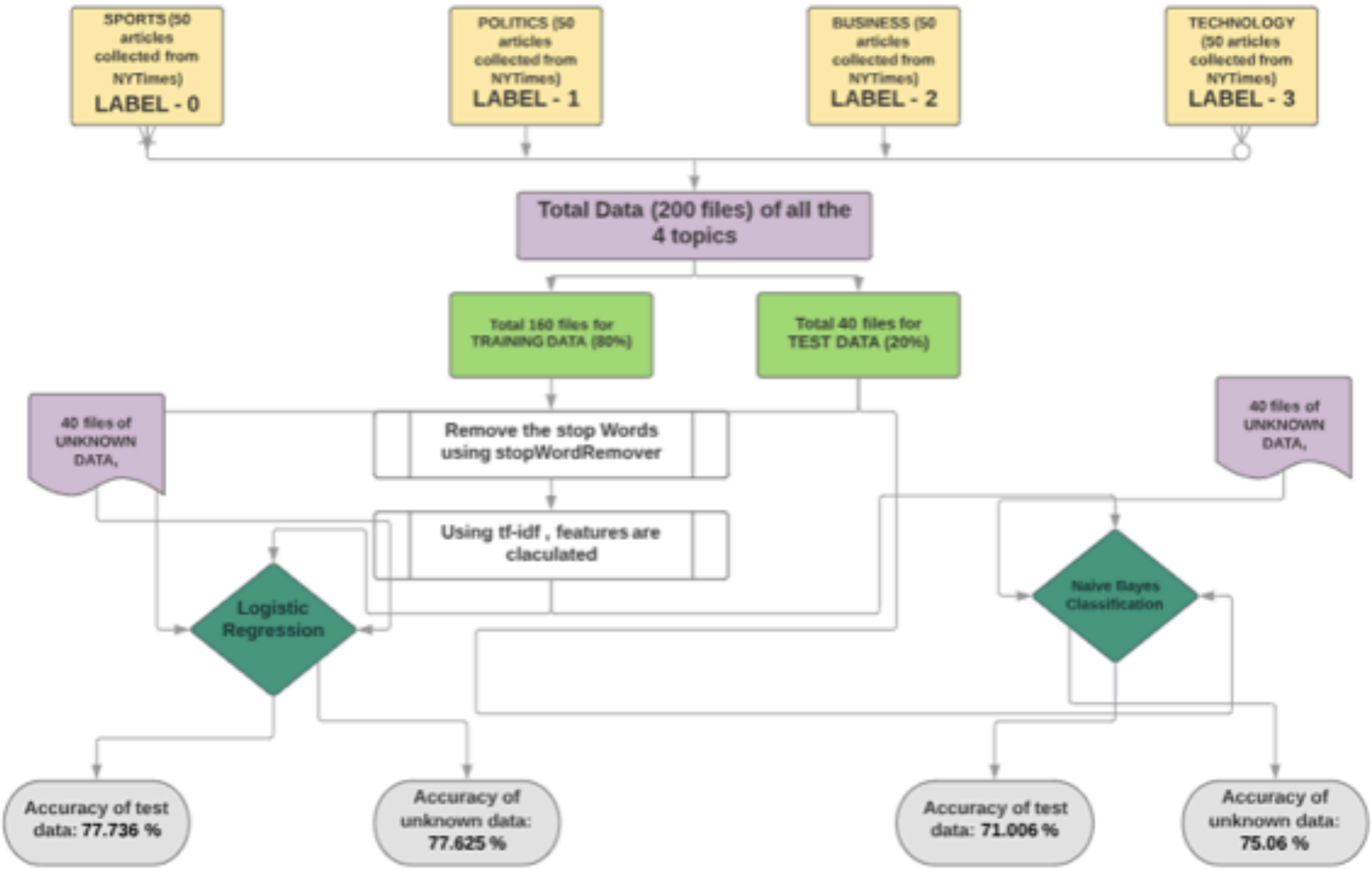
**Unknown Set Data:**

Accuracy : -  **75.0652 %**

Confusion Matrix: -

| | Sports | Poltics | Business | Education |
|---|---|---|---|---|
| Sports | 6 | 0 | 1 | 3 |
| Poltics | 0 | 6 | 2 | 2 |
| Business | 0 | 0 | 9 | 1 |
| Education | 0 | 0 | 1 | 9 |

# Flow Diagram –


SPORTS (50 articles collected from NYTimes)
LABEL - 0
POLITICS (50 articles collected from NYTimes)
LABEL - 1
BUSINESS (50 articles collected from NYTimes)
LABEL - 2
TECHNOLOGY (50 articles collected from NYTimes)
LABEL - 3
Total Data (200 files) of all the 4 topics
Total 160 files for TRAINING DATA (80%)
Total 40 files for TEST DATA (20%)
40 files of UNKNOWN DATA,
Remove the stop Words using stopWordRemover
Using tf-idf , features are claculated
40 files of UNKNOWN DATA,
Logistic Regression
Naive Bayes Classification
Accuracy of test data: 77.736 %
Accuracy of unknown data: 77.625 %
Accuracy of test data: 71.006 %
Accuracy of unknown data: 75.06 %