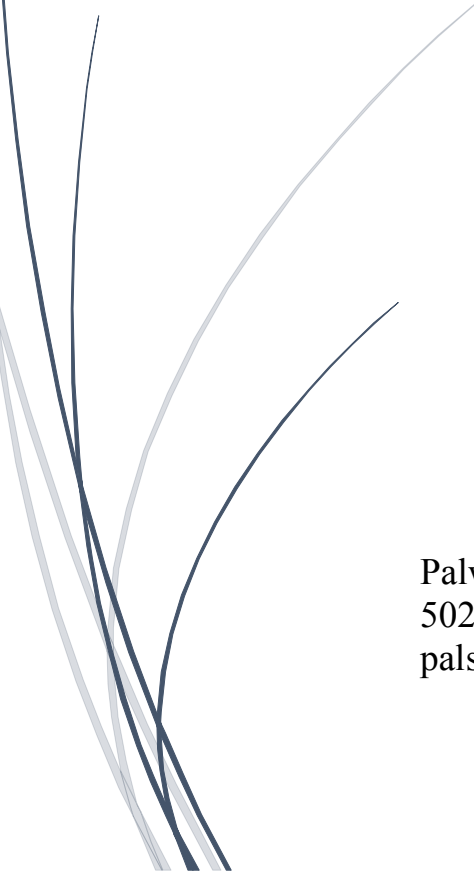


A dark blue vertical bar on the left side of the slide. A blue arrow points from this bar towards the right, containing the date.

5/11/2018

# Data Analytics Pipeline using *Apache Spark*

Several thin, curved, light blue lines that sweep upwards from the bottom left towards the center of the slide.

Palwinder Singh  
50247454  
palsingh@buffalo.edu

Gursimrat Singh  
50212333  
gsingh37@buffalo.edu

# 1. Understand Apache Spark with Titanic data analysis

Being new to spark we first understood the working and functionality of Apache Spark. We took reference from the following link and executed all the code they had.

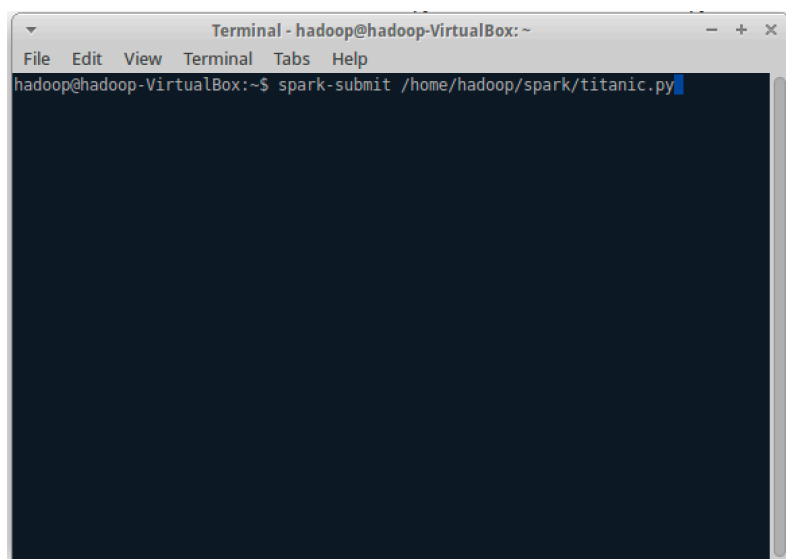
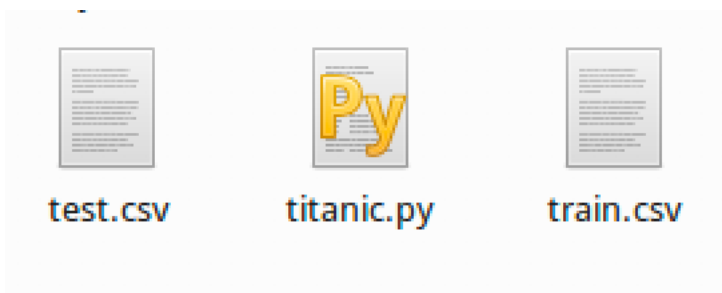
Link : [https://6chaoran.wordpress.com/2016/08/13/\\_\\_\\_trashed/](https://6chaoran.wordpress.com/2016/08/13/___trashed/)

The language used in this is: python and library used is pyspark which we used throughout the lab.

There are 4 areas we had the hands-on experience here:

- I. Data Loading and Parsing
- II. Data Manipulation
- III. Feature Engineering
- IV. Apply Spark ml/mllib models

This step helped us familiarize with all the steps we needed to accomplish our goal and create a meticulous Data analytics pipeline.



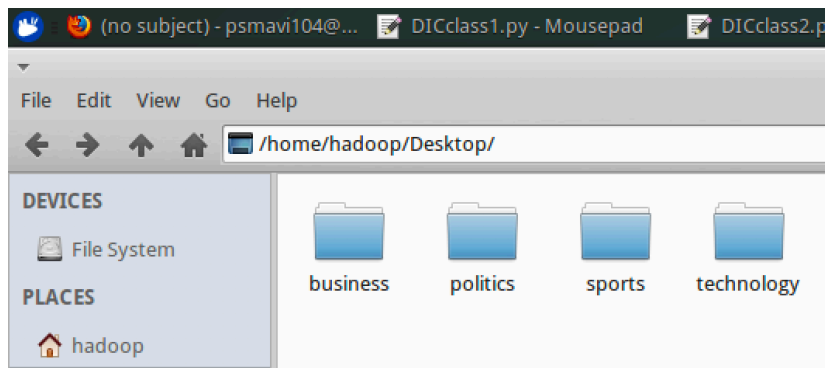
```
e.scala:45) finished in 0.001 s
18/05/11 14:50:24 INFO DAGScheduler: Job 68 finished: aggregate at AreaUnderCurv
e.scala:45, took 0.014335 s
18/05/11 14:50:24 INFO MapPartitionsRDD: Removing RDD 321 from persistence list
18/05/11 14:50:24 INFO BlockManager: Removing RDD 321
{'RandomForest': 0.8502677205507397, 'LogisticRegression': 0.8287225905150437, '
DecistionTree': 0.5850012748597654}
18/05/11 14:50:24 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 14:50:24 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 14:50:24 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEnd
point stopped
```

## 2. Collect and clean data

We used the 'beautifulsoup' library of python to crawl the urls containing articles related to our keyword. We collect data in a .txt file, and every article fetched has not been preprocessed and is stored in an individual file. For 100 articles we have 100 .txt files.

We collected data on the following 4 categories:

- I. Business
- II. Politics
- III. Sports
- IV. Technology



We have enough articles to split them as training and testing data.

## 3. Feature Engineering

Now this is the most crucial part of our project and also the determining step for our results and accuracy at the end of our data analytics pipeline.

There are two main processes in this part:

- I. Data Ingestion
- II. Feature Extraction

Now we'll take the data and engineer it. We use the functions from the library 'pyspark' which'll help us do the above meticulously.

We use the 'regular expression tokenizer' to manipulate our dataframe, and get the words from the value column. Now we filter and funnel the data by removing any special characters and stopwords from the words we have. Now we have another column in our dataframe called 'filtered', which has the words we desire and which we'll use to train our system.

Here in this step we didn't process raw count you but will compute the probability of the word frequency to the total words in the article.

## Function Description:

**StringIndexer** :Encodes a string column of labels to a column of label indices. The indices are in  $[0, \text{numLabels})$ , ordered by label frequencies, so the most frequent label gets index 0.

**regexTokenizer**: Tokenization (with Regular Expression)

The method we use in the lab for the above is: Logistic Regression (using TF, IDF)

Also, in the next part we will use another classifier: Naïve Bayes

value Category	words	filtered	rawFeatures	features label
Sections SEARCH Skip to ...	[business sections, search, skip, to...	[content, site, index, subs...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0
...	[business monies, needed, to, fund, ...	[monies, needed, fund, musi...	(120,[4,19,22,24,44,47,48,5...	(120,[4,19,22,24,44,47,48,5... 0.0
...	[business philanthropist, civic, aff...	[philanthropist, civic, aff...	(120,[4,10,12,30,33,34,37,3...	(120,[4,10,12,30,33,34,37,3... 0.0
...	[business his, career, as, a, partne...	[career, partner, lehman, b...	(120,[22,47,52,53,61,62,65...	(120,[22,47,52,53,61,62,65... 0.0
...	[business it, the, fortunate, opport...	[fortunate, opportunity, ex...	(120,[12,21,24,25,30,31,40...	(120,[12,21,24,25,30,31,40... 0.0
...	[business aip, which, became, a, res...	[aip, became, respected, gr...	(120,[14,21,24,25,42,45,46...	(120,[14,21,24,25,42,45,46... 0.0
...	[business wife, wendy, wasson, bingh...	[wife, wendy, wasson, bingh...	(120,[3,4,8,19,21,24,32,33...	(120,[3,4,8,19,21,24,32,33... 0.0
...	[business with, this, line, taken, f...	[line, taken, us, navy, day...	(120,[14,21,22,44,49,65,68...	(120,[14,21,22,44,49,65,68... 0.0
...	[business of, geneva, and, paris, th...	[geneva, paris, throughout...	(120,[1,20,24,25,38,44,50,5...	(120,[1,20,24,25,38,44,50,5... 0.0
...	[business town, school, for, boys, c...	[town, school, boys, class...	(120,[6,24,31,35,48,49,54,5...	(120,[6,24,31,35,48,49,54,5... 0.0

only showing top 10 rows

value	Category	words	filtered	rawFeatures	features label
Sections SEARCH Skip to ...	business	[sections, search, skip, to...	[content, site, index, subs...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0
Sections SEARCH Skip to ...	business	[sections, search, skip, to...	[content, site, index, econ...	(120,[0,1,2,3,4,5,6,7,8,10...	(120,[0,1,2,3,4,5,6,7,8,10... 0.0
NYTimes.com no...	business	[nytimes, com, no, longer, ...	[nytimes, com, longer, supp...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0
NYTimes.com no...	business	[nytimes, com, no, longer, ...	[nytimes, com, longer, supp...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 1.0
Sections SEARCH skip to ...	business	[sections, search, skip, to...	[content, site, index, busi...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0
Sections SEARCH skip to ...	politics	[sections, search, skip, to...	[content, site, index, styl...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 1.0
Sections SEARCH skip to ...	technology	[sections, search, skip, to...	[content, site, index, subs...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 2.0
Sections SEARCH skip to ...	business	[sections, search, skip, to...	[content, site, index, new...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0
NYTimes.com no...	business	[nytimes, com, no, longer, ...	[nytimes, com, longer, supp...	(120,[0,1,2,3,4,5,6,7,8,9,1...	(120,[0,1,2,3,4,5,6,7,8,9,1... 0.0

only showing top 10 rows

```
18/05/11 12:24:42 INFO CodeGenerator:
+-----+
|      value|Category|
+-----+
| Sections SEARC...|business|
| NYTi...|business|
| NYTi...|business|
| NYTi...|business|
| NYTi...|business|
| Sections SEARC...|business|
| NYTi...|business|
| NYTi...|business|
| Sections SEARC...|business|
| NYTi...|business|
| Sections SEARC...|business|
| NYTi...|business|
| NYTi...|business|
| Sections SEARC...|business|
| Section...|business|
| NYTi...|business|
| Sections SEARC...|business|
| NYTi...|business|
+-----+
only showing top 20 rows
```

## 4. Multi-class Classification

We selected two classifiers for the lab:

- I. Logistic Regression (using TF and IDF)
- II. Naïve Bayes

Now we have a lot of data for all of the 4 categories, so we use the training to the testing set in the ratio .8 : 0.2 .Next we model our pipeline .

We have the following columns in our dataframe at this point:

"value", "Category", "words", "filtered", "rawFeatures", "features", "label"

We'll order this on the basis of the probability which is indeed the core of our machine learning approach.

Next we use the function from the pyspark library called 'MulticlassClassificationEvaluator' to get the accuracy value .

Classifier	Split Ratio	Accuracy achieved
Logistic Regression (using TF IDF)	0.8 training : 0.2 testing	60.47%
Naïve Bayes	0.8 training : 0.2 testing	66.8%

```
18/05/11 12:01:48 INFO Executor: Finished task 8.0 in stage 75.0 (TID 914). 1826 bytes result sent to driver
18/05/11 12:01:48 INFO TaskSetManager: Finished task 8.0 in stage 75.0 (TID 914) in 7 ms on localhost (executor driver) (10/10)
18/05/11 12:01:48 INFO TaskSchedulerImpl: Removed TaskSet 75.0, whose tasks have all completed, from pool
18/05/11 12:01:48 INFO DAGScheduler: ResultStage 75 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.047 s
18/05/11 12:01:48 INFO DAGScheduler: Job 42 finished: collectAsMap at MulticlassMetrics.scala:53, took 3.238468 s
('-----The accuracy is : ', 0.6047141264532568, '-----')
18/05/11 12:01:48 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 12:01:48 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 12:01:48 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
```

### Logistic Regression

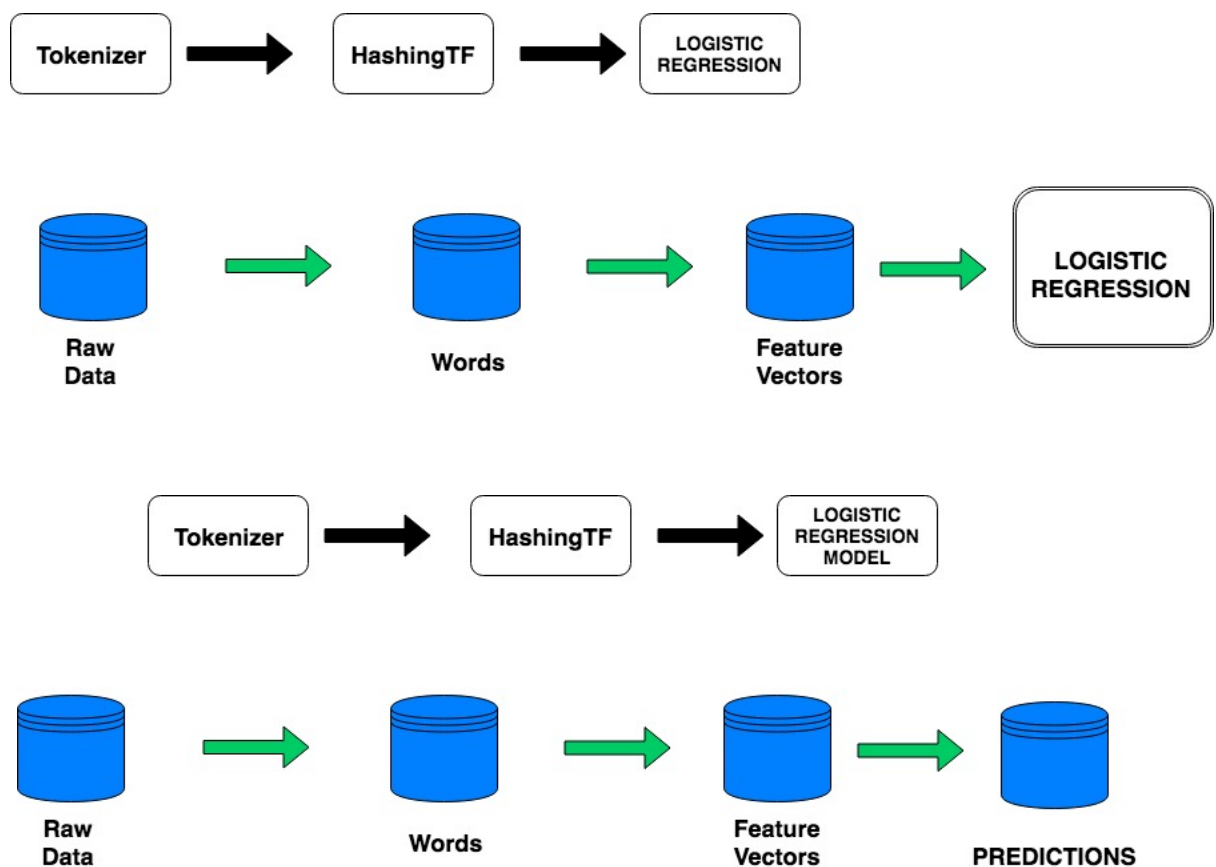
```
18/05/11 12:02:57 INFO TaskSchedulerImpl: Removed TaskSet 28.0, whose tasks have all completed, from pool
18/05/11 12:02:57 INFO DAGScheduler: ResultStage 28 (collectAsMap at MulticlassMetrics.scala:53) finished in 0.097 s
18/05/11 12:02:57 INFO DAGScheduler: Job 19 finished: collectAsMap at MulticlassMetrics.scala:53, took 3.368098 s
('-----The accuracy is : ', 0.6681864235055724, '-----')
18/05/11 12:02:57 INFO SparkContext: Invoking stop() from shutdown hook
18/05/11 12:02:57 INFO SparkUI: Stopped Spark web UI at http://10.0.2.15:4040
18/05/11 12:02:57 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
18/05/11 12:02:57 INFO MemoryStore: MemoryStore cleared
```

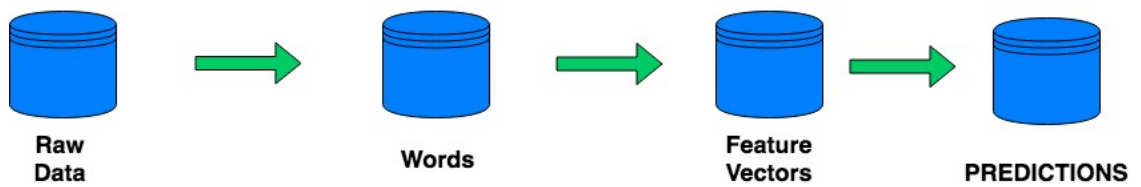
### Naïve Bayes

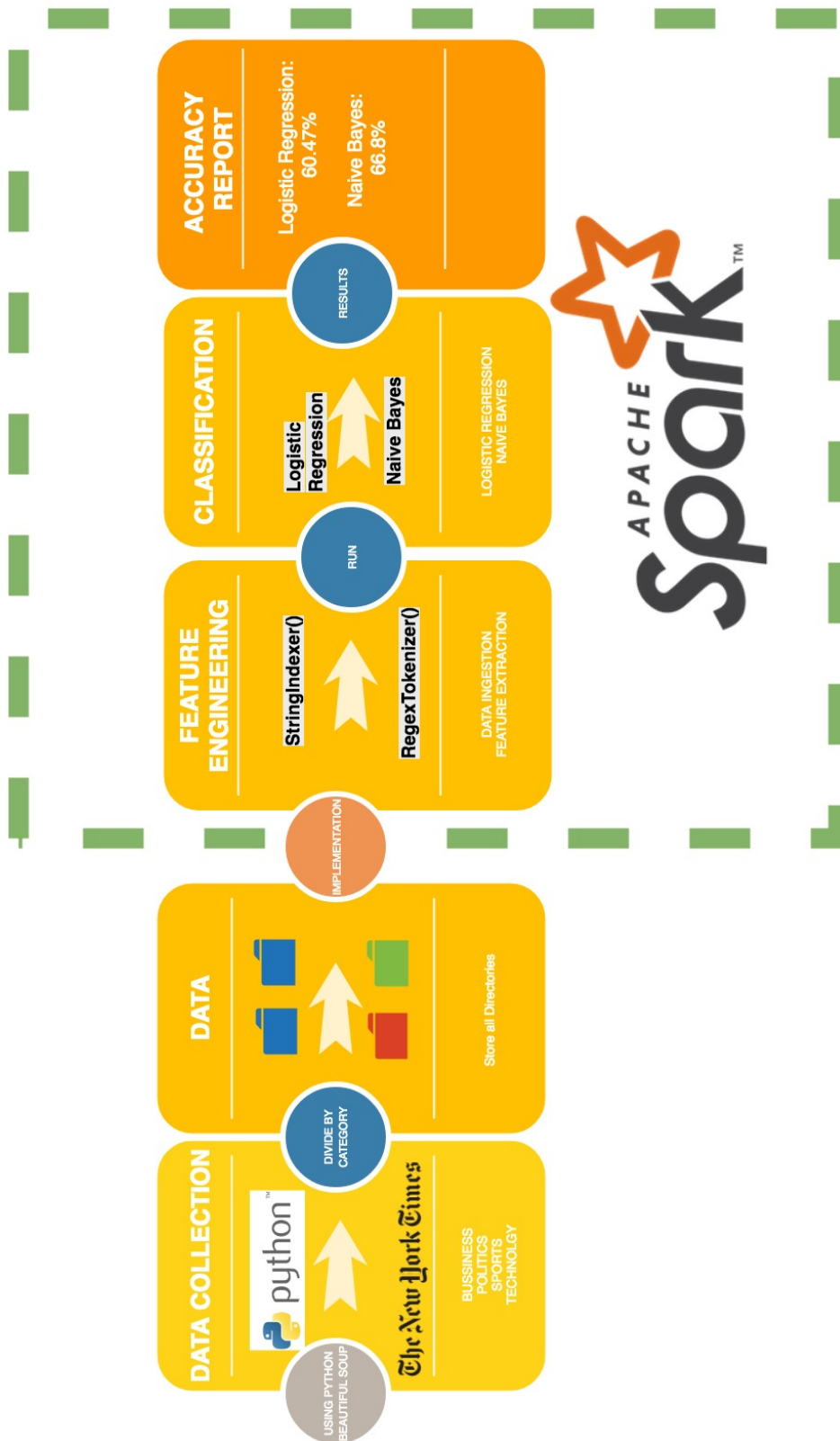
## 5. Testing:

In this step we have collected data again with keyword 'Google' which falls under technology. We take training data as the one collected before and pass the new data as testing data . This will categorize the new data for us,i.e. technology . Since we were getting higher accuracy in Naïve Bayes we used this classifier for random data.

## 6. Documentation/Block Diagram







Data Analytics Pipeline Block Diagram



**Video of running code:** <https://drive.google.com/file/d/1jTbhv-g1BSJayhPU2Gugrfa0HHeY3iy6/view?usp=sharing>

#### References:

- [https://6chaoran.wordpress.com/2016/08/13/\\_trashed/](https://6chaoran.wordpress.com/2016/08/13/_trashed/)
- Professor Bina Ramamurthy's lectures and notes.(Also the class demos).
- <https://spark.apache.org/docs/0.9.0/python-programming-guide.html>
- <https://stackoverflow.com/questions/36217090/how-do-i-get-python-libraries-in-pyspark>
- <https://towardsdatascience.com/multi-class-text-classification-with-pyspark>
- <https://spark.apache.org/docs/2.3.0/ml-pipeline.html>
- Charts are made on draw.io