Machine learning/ Natural language processing

INFO319 – Research Topics in Big data

Vimala Nunavath

vimala.nunavath@uia.no

04.10.2019

Agenda

- Machine learning
- Natural language processing
- Presentations by you.
- Practical session

Programming project

- Deadline is today!!

Machine learning

Machine learning

"Programming computer to optimize the performance using example data and past experience"

Machine learning

Field of study that gives “computers the ability to learn without being explicitly programmed”

---- Arthur Samuel, 1959

Machine learning

- Subset of Artificial Intelligence
- Design and development of algorithms
- Computers evolve behavior based on empirical data

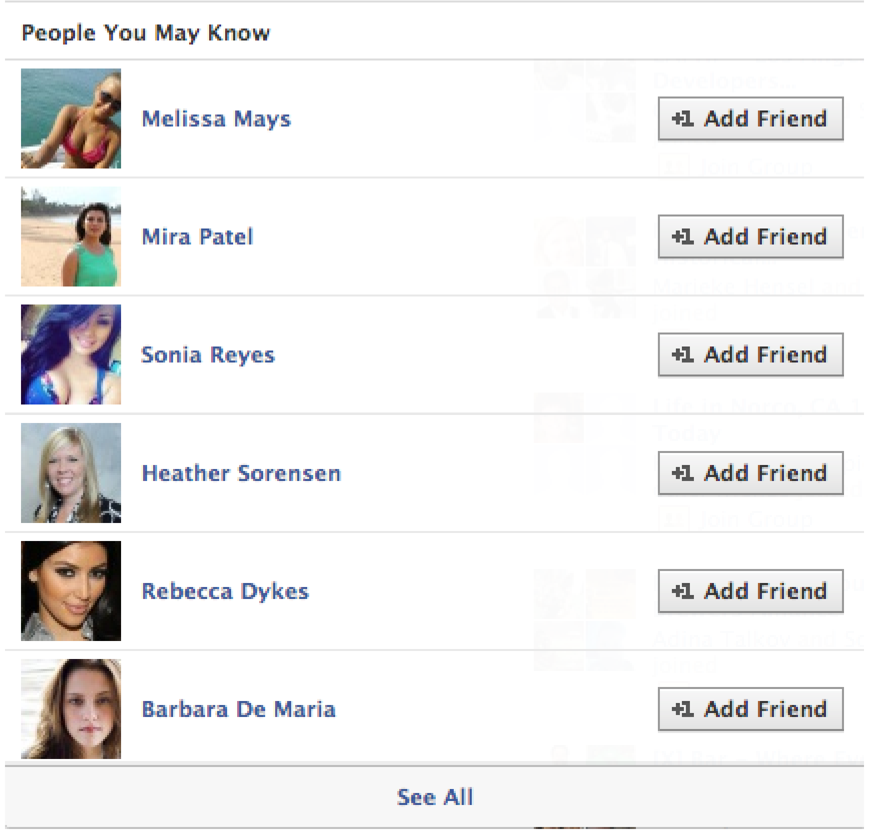
Machine learning

- The **input data** to a machine learning system can be **numerical, textual, audio, visual, or multimedia.**
- The corresponding **output data** of the system can be a **floating-point number.**
- **E.g., -** the velocity of a rocket, an integer representing a category or a class,

  - a pigeon or a sunflower from image recognition.

Machine learning - Applications

- Recommend friends, dates and products to end-users.



People You May Know

Melissa Mays — +1 Add Friend

Mira Patel — +1 Add Friend

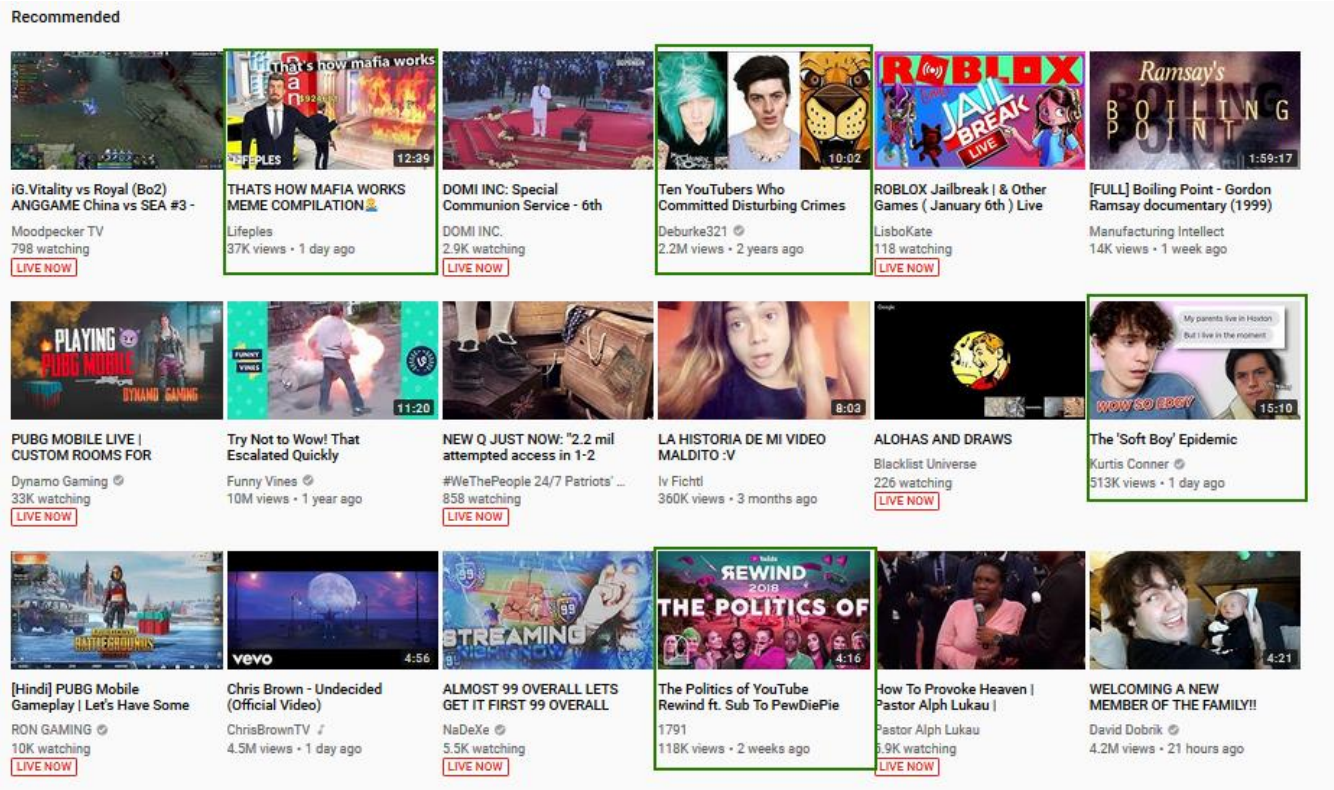
Sonia Reyes — +1 Add Friend

Heather Sorensen — +1 Add Friend

Rebecca Dykes — +1 Add Friend

Barbara De Maria — +1 Add Friend

See All



Recommended

iG.Vitality vs Royal (Bo2) ANGGAME China vs SEA #3 -
Moodpecker TV
798 watching
LIVE NOW

THATS HOW MAFIA WORKS MEME COMPILATION
Lifeples
37K views • 1 day ago
12:39

DOMI INC: Special Communion Service - 6th
DOMI INC.
2.9K watching
LIVE NOW

Ten YouTubers Who Committed Disturbing Crimes
Deburke321
2.2M views • 2 years ago
10:02

ROBLOX Jailbreak | & Other Games (January 6th) Live
LisboKate
118 watching
LIVE NOW

[FULL] Boiling Point - Gordon Ramsay documentary (1999)
Manufacturing Intellect
14K views • 1 week ago
1:59:17

PUBG MOBILE LIVE | CUSTOM ROOMS FOR
Dynamo Gaming
33K watching
LIVE NOW

Try Not to Wow! That Escalated Quickly
Funny Vines
10M views • 1 year ago
11:20

NEW Q JUST NOW: "2.2 mil attempted access in 1-2
#WeThePeople 24/7 Patriots' ...
858 watching
LIVE NOW

LA HISTORIA DE MI VIDEO MALDITO :V
lv Fichtl
360K views • 3 months ago
8:03

ALOHAS AND DRAWS
Blacklist Universe
226 watching
LIVE NOW

The 'Soft Boy' Epidemic
Kurtis Conner
513K views • 1 day ago
15:10

[Hindi] PUBG Mobile Gameplay | Let's Have Some
RON GAMING
10K watching
LIVE NOW

Chris Brown - Undecided (Official Video)
ChrisBrownTV
4.5M views • 1 day ago
4:56

ALMOST 99 OVERALL LETS GET IT FIRST 99 OVERALL
NaDeXe
5.5K watching
LIVE NOW

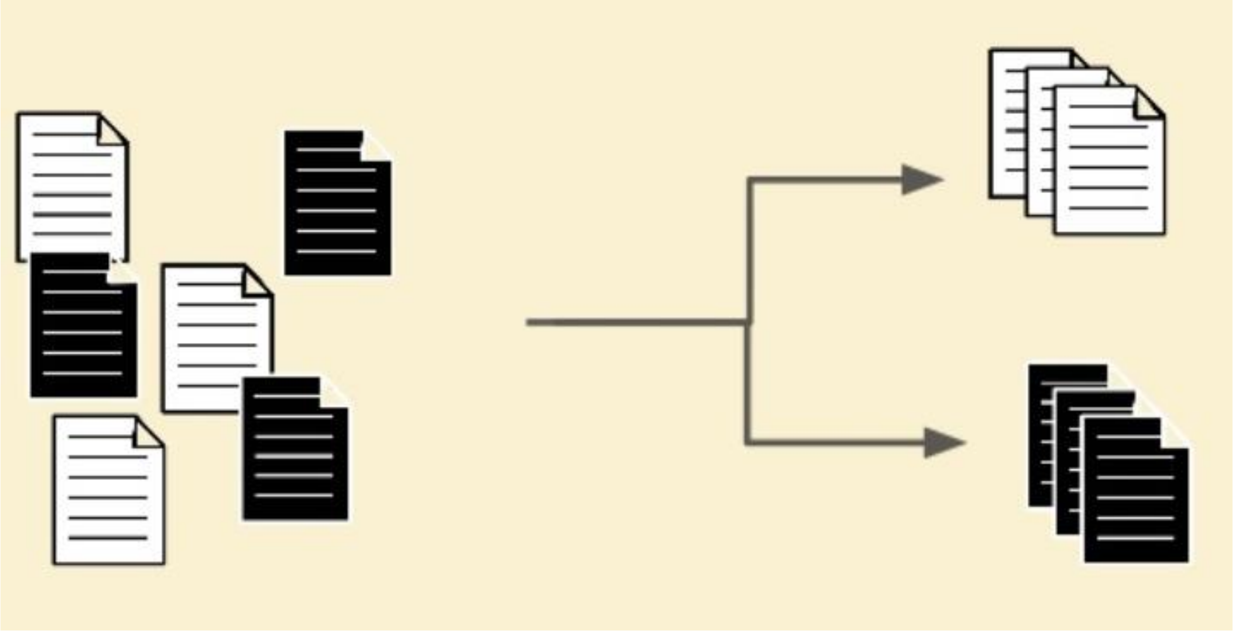
The Politics of YouTube Rewind ft. Sub To PewDiePie
1791
118K views • 2 weeks ago
4:16

How To Provoke Heaven | Pastor Alph Lukau |
Pastor Alph Lukau
5.9K watching
LIVE NOW

WELCOMING A NEW MEMBER OF THE FAMILY!!
David Dobrik
4.2M views • 21 hours ago
4:21

Machine learning - Applications

- Classify content into predefined groups



Machine learning - Applications

- Identify key topics in large collections of text

ans achieve
The term is generally u
cribe the causes for anim
nan motivation.
d in a basic need to mini
nclude specific needs suc
ing, ideal, o
wiseGEEK

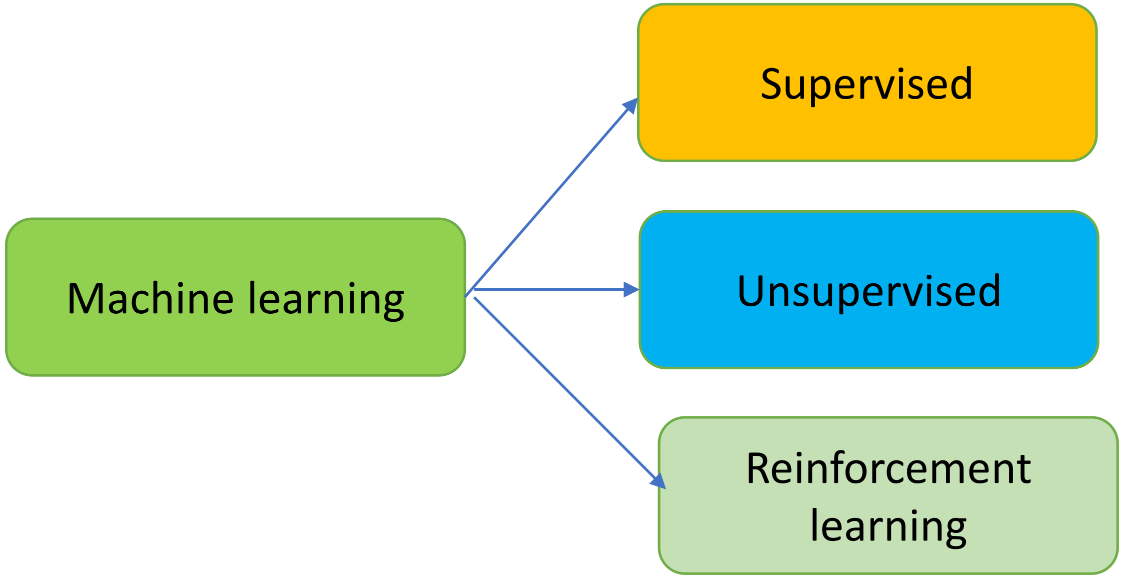
Machine learning - Applications

- Natural language processing

Machine learning - Applications

- Detect Anomalies within given data
- Ranking Search Results with user feedback learning
- Classifying DNA sequences
- Sentiment analysis/ opinion mining
- Bioinformatics
- Speech and Handwriting Recognition

Machine learning types

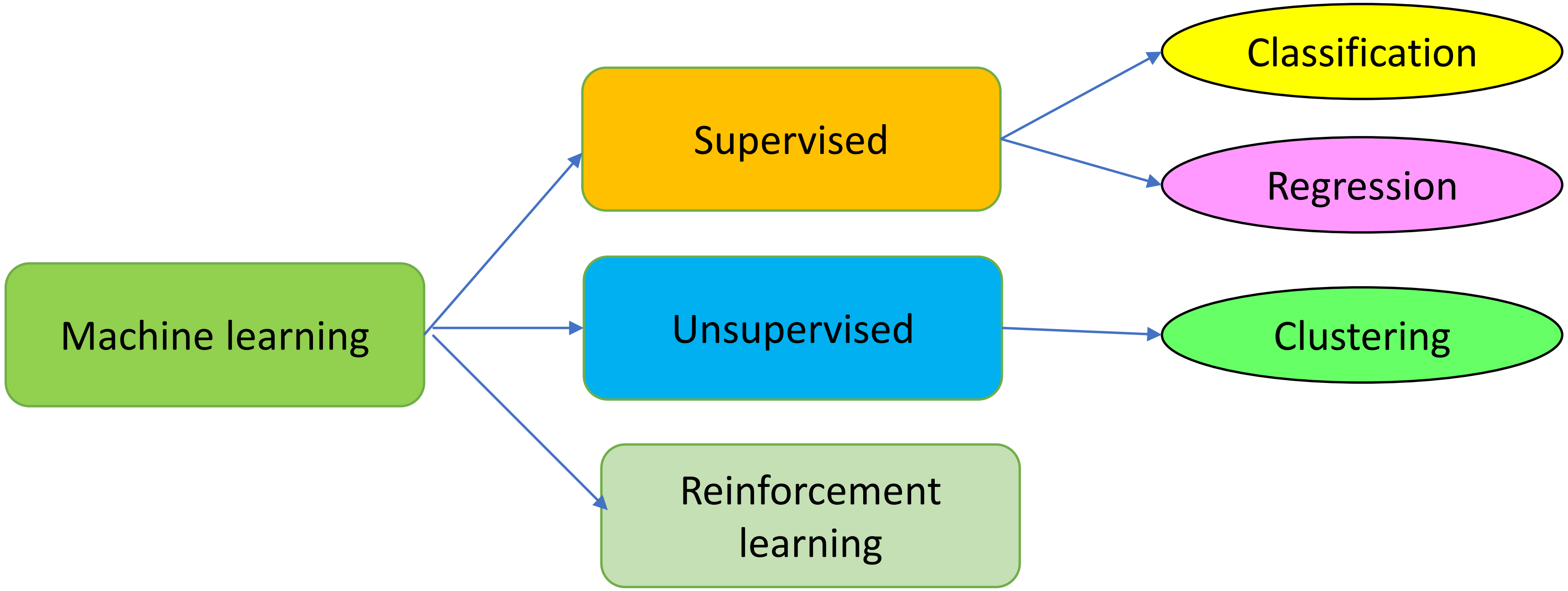


Deals with labeled training data, learn to map inputs to outputs

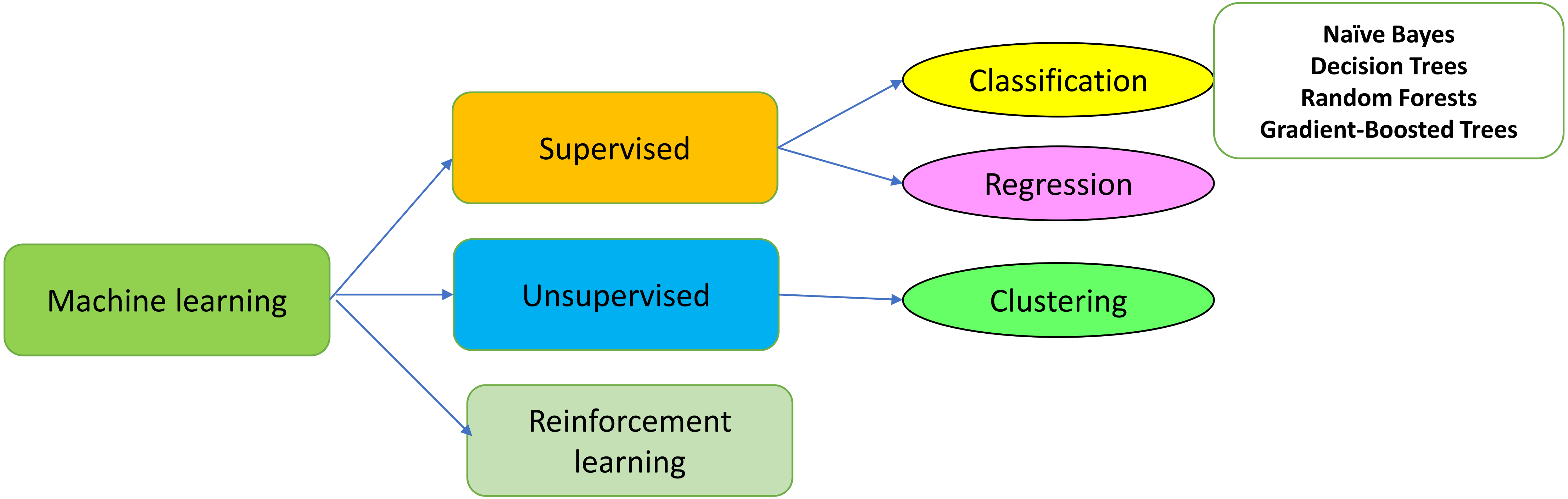
Deals with unlabeled data, observe structure in the data and find patterns.

Dynamic environment, perform a certain goal

Machine learning types



Machine learning - Classification



Machine learning

Supervised

Unsupervised

Reinforcement learning

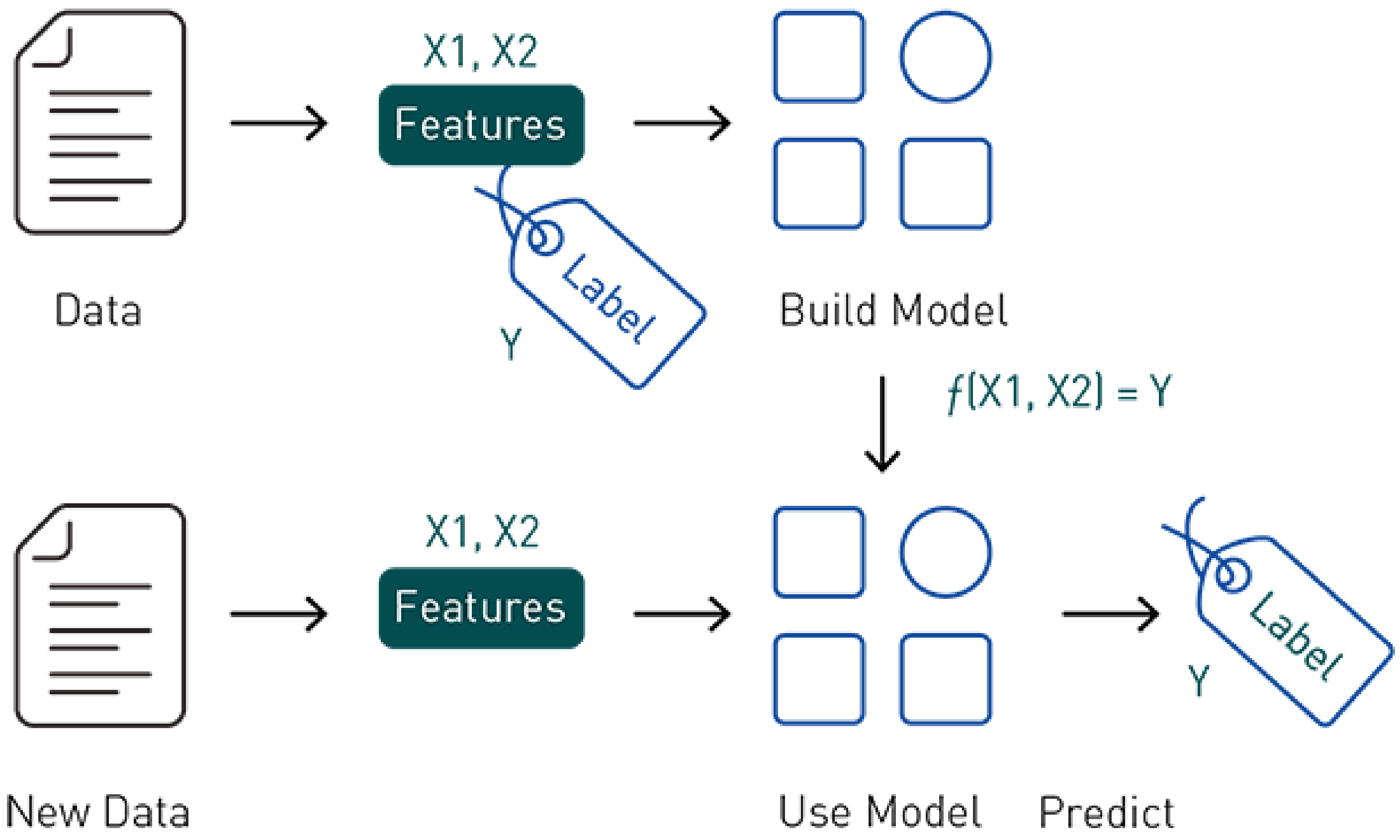
Classification

Regression

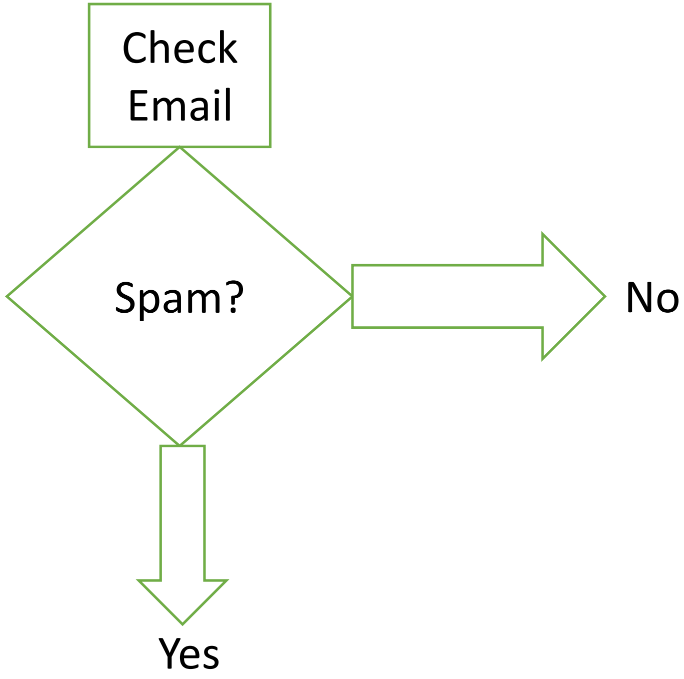
Clustering

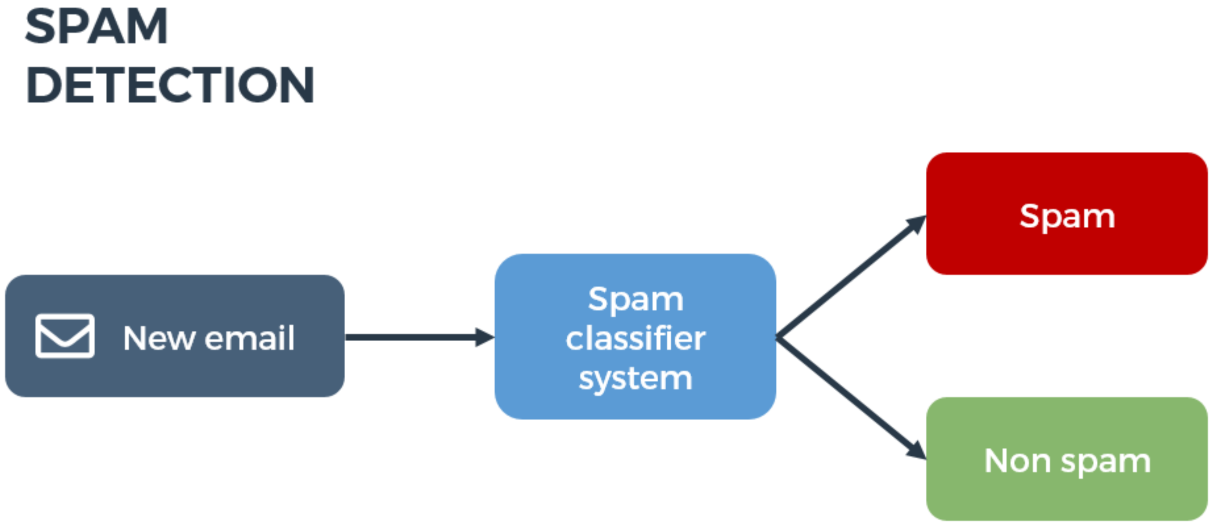
Naïve Bayes
Decision Trees
Random Forests
Gradient-Boosted Trees

Classification

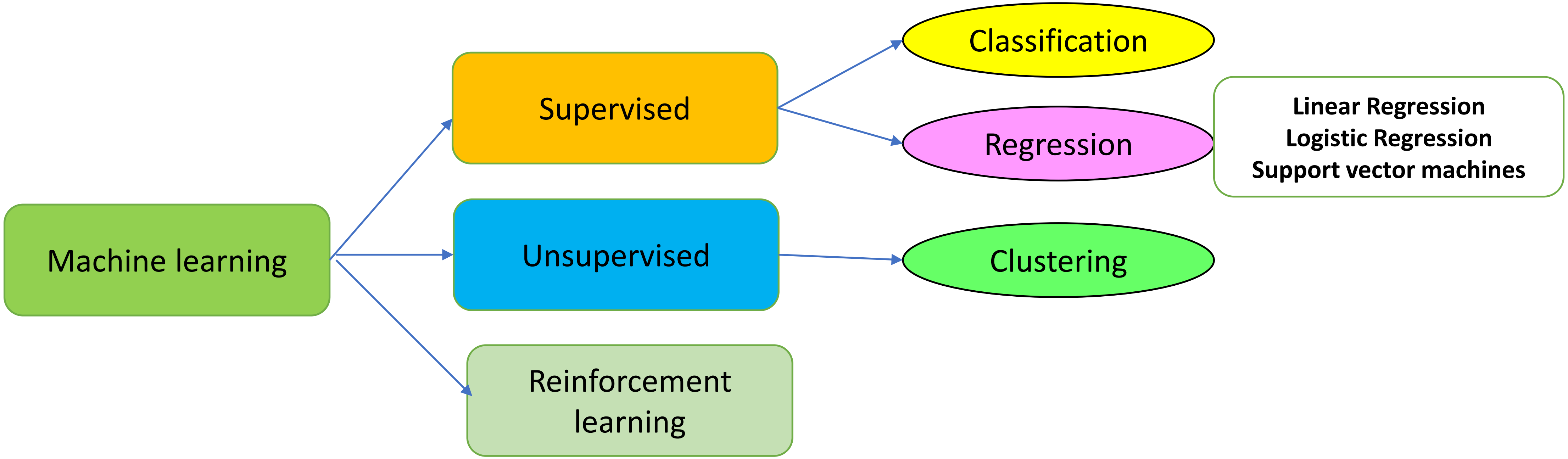


Machine learning - Classification





Machine learning - Regression



Machine learning

Supervised

Unsupervised

Reinforcement learning

Classification

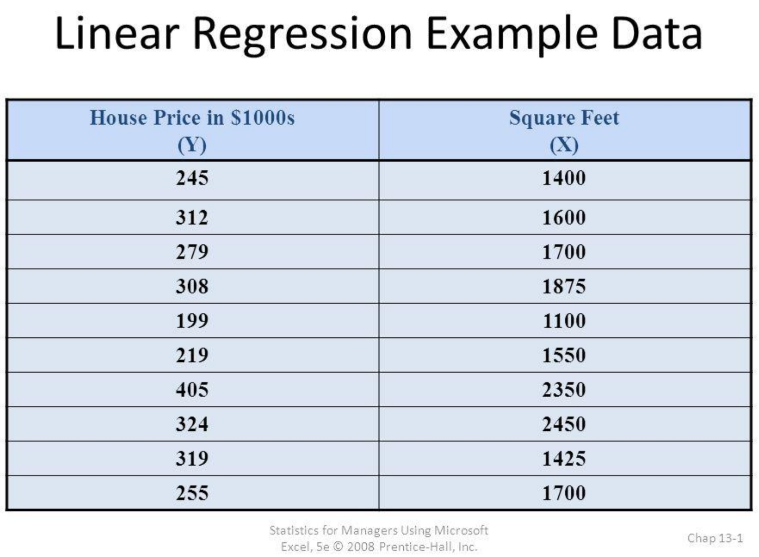
Regression

Clustering

Linear Regression
Logistic Regression
Support vector machines

Machine learning - Regression

- Regression models are used to predict a continuous value.
  - E.g., predicting prices of a house given the features of house like size, price etc
- In liner regression, we draw all possible lines going through the points such that it is closest to all.

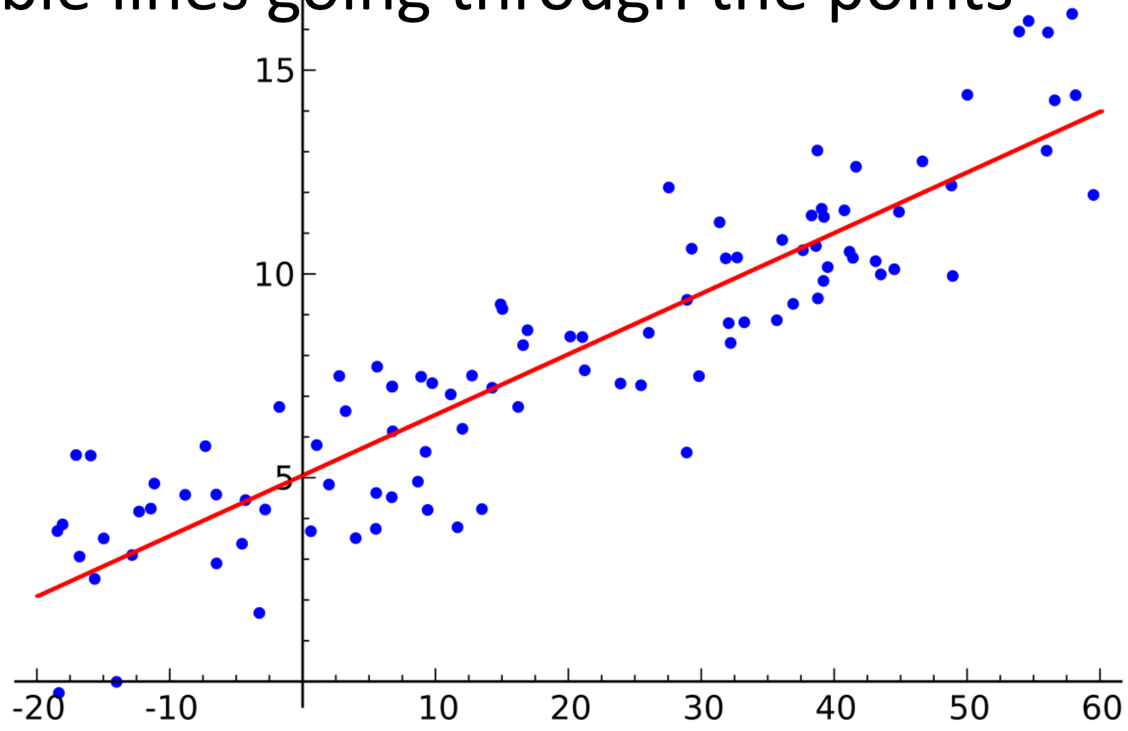


Linear Regression Example Data

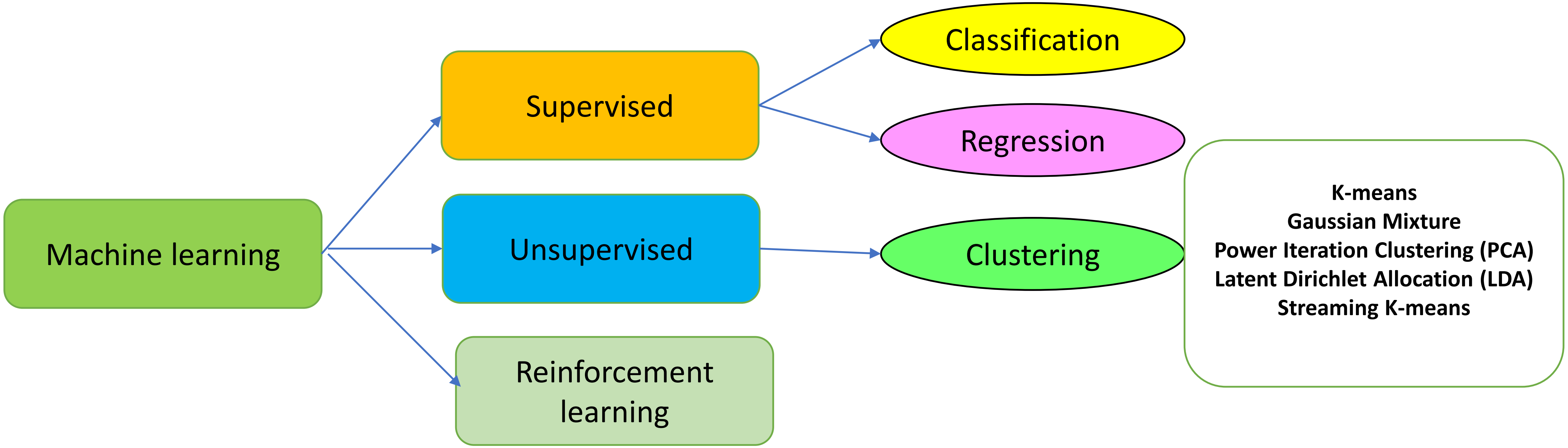
House Price in $1000s (Y)	Square Feet (X)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Statistics for Managers Using Microsoft Excel, 5e © 2008 Prentice-Hall, Inc.

Chap 13-1



Machine learning types



Machine learning

Supervised

Unsupervised

Reinforcement learning

Classification

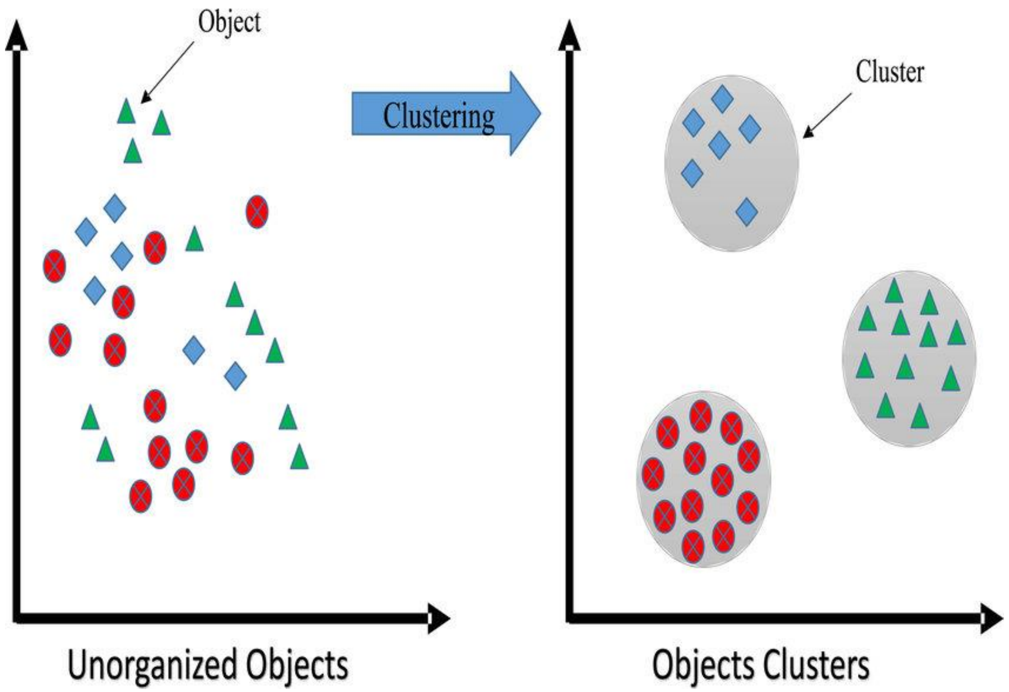
Regression

Clustering

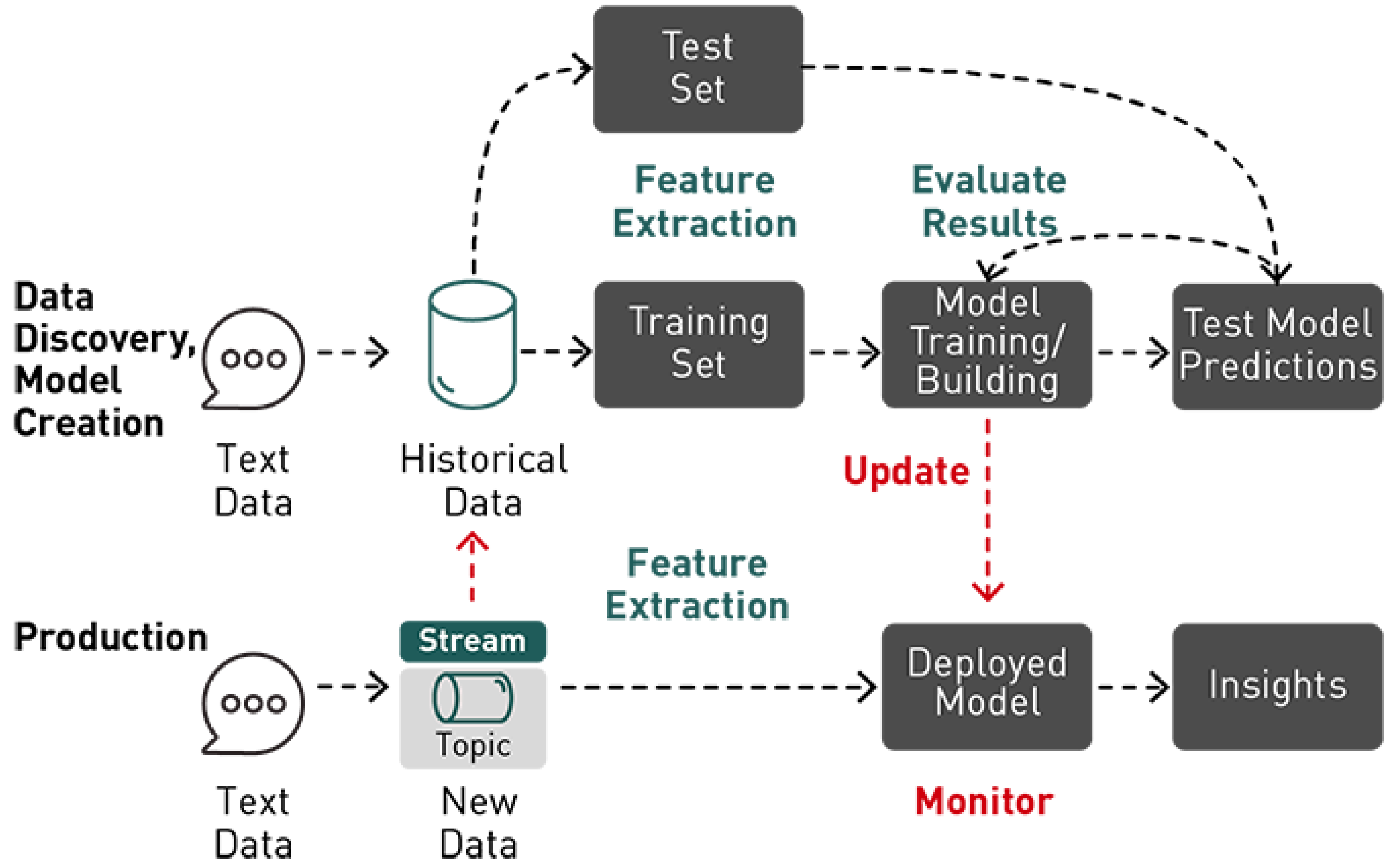
K-means
Gaussian Mixture
Power Iteration Clustering (PCA)
Latent Dirichlet Allocation (LDA)
Streaming K-means

Machine learning - Clustering

- **Clustering** is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups.
- It is basically a collection of objects on the basis of similarity and dissimilarity between them.



Machine Learning Workflow



Natural language processing

What is Natural language processing?

- Shortened as NLP,
- **Field in ML** that **deals with** the **interaction between computers** and **humans** using the **natural language**.
- The ultimate objective of NLP is to **read**, **decipher**, **understand**, and **make sense of the human languages** in a manner that is valuable.
- Most NLP techniques rely on machine learning to derive meaning from human languages.

The goal of NLP

- The goal of NLP is to “accomplish human-like language processing”.

- A full NLP System would be able to:
    1. Paraphrase an input text
    2. Translate the text into another language
    3. Answer questions about the contents of the text
    4. Draw inferences from the text

Why NLP?

- Aids communication between two humans
  - Machine translation
  - Speech-to-speech translation
  - Speech-to-text & text-to-speech
  - Editorial aids (spelling & grammar checkers)
  - Aids communication between human and machine
  - Personal assistants
  - Interactive Voice Response systems
  - Aids communication between two machines

Why NLP for Social Media?

- Social Media generates BIG UNSTRUCTURED NATURAL LANGUAGE DATA
    - Volume: 1.3 Billion monthly active FB users
    - Velocity: 5700 Tweets/sec. 2500 FB-msg/sec
    - Variety: scripts, languages, style, topic, …
- Today's world resides in social media
- It is impossible to process (consume, understand or summarize) this information manually.

Why NLP?

- Trending Topic Detection
- Information Retrieval & Extraction
- Information Summarization
- **Sentiment Detection**
- Rumor Detection

Challenges of NLP

- Traditionally, NLP systems are designed to handle input that is
  - Grammatically correct
  - No spelling errors
  - Single language
  - The right script

Sentiment Analysis

- Sentiment analysis refers to the class of computational and natural language processing-based techniques **used to identify, extract or characterize subjective information**, such as opinions, expressed in a given piece of text.

- The main purpose of sentiment analysis is to classify a writer’s attitude towards various topics into positive, negative or neutral categories.

Sentiment Analysis

- Sentiment analysis has many applications in different domains e.g., business intelligence, politics, sociology, social networking websites, microblogs, wikis and Web applications

- Data such as web-postings, Tweets, videos, etc., all express opinions on various topics and events, offer immense opportunities to study and analyze human opinions and sentiment.

Latent Dirichlet Allocation (LDA)

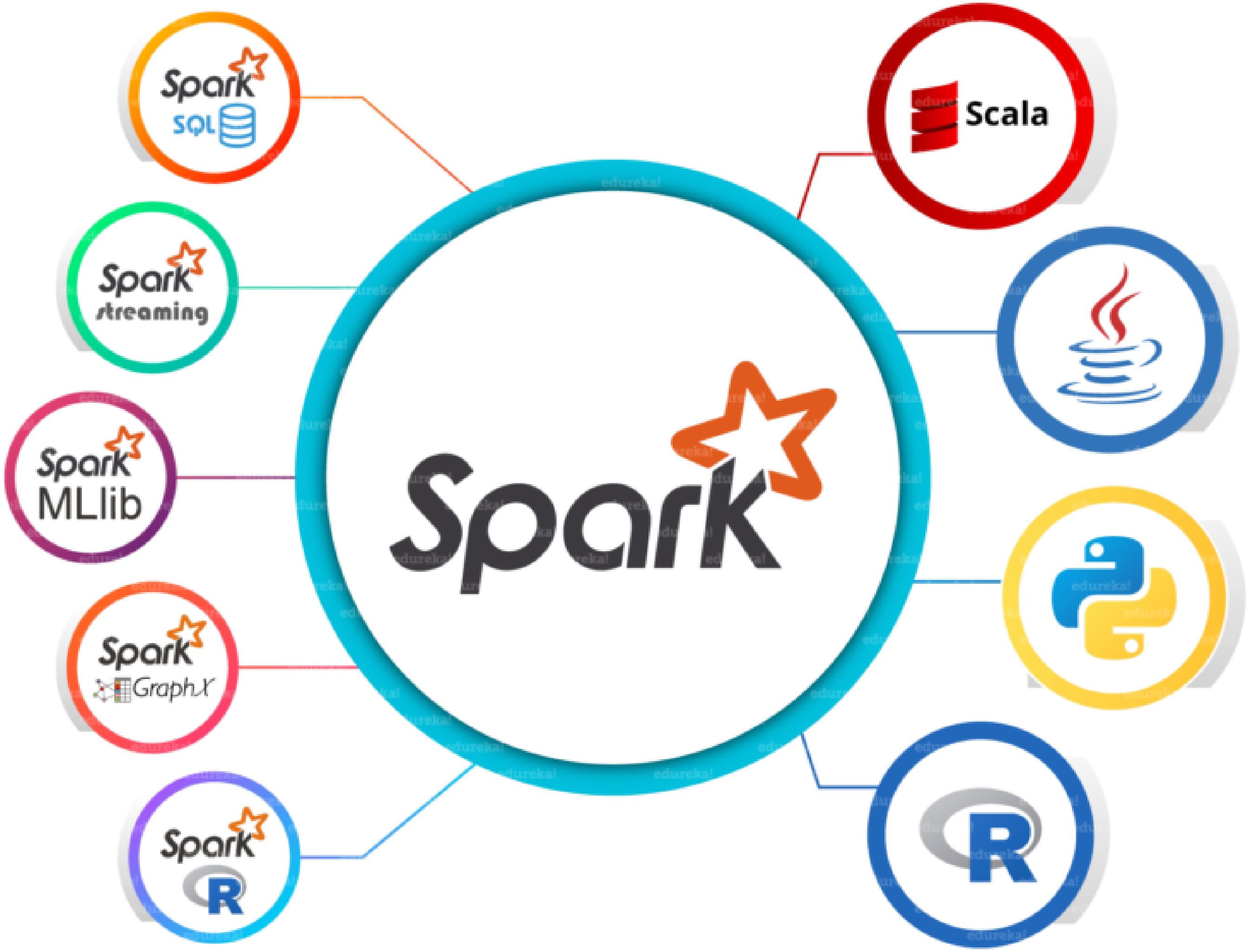
- It is a “generative probabilistic model” of a collection of composites made up of parts.
- Here composites are documents and the parts are words and/or phrases.
- Allows a word to simultaneously belong to several clusters with varying degree.

Latent Dirichlet Allocation (LDA)

- **Latent**: refers to everything that we don’t know a priori and are hidden in the data.
  - E.g., the themes or topics that document consists of are unknown, but they are believed to be present as the text is generated based on those topics.
- **Dirichlet:** It is a ‘distribution of distributions.
  - E.g.,in the context of topic modeling, the Dirichlet is the distribution of topics in documents and distribution of words in the topic.
- **Allocation**: This means that once we have Dirichlet, we will allocate topics to the documents and words of the document to topics.

Spark MLib

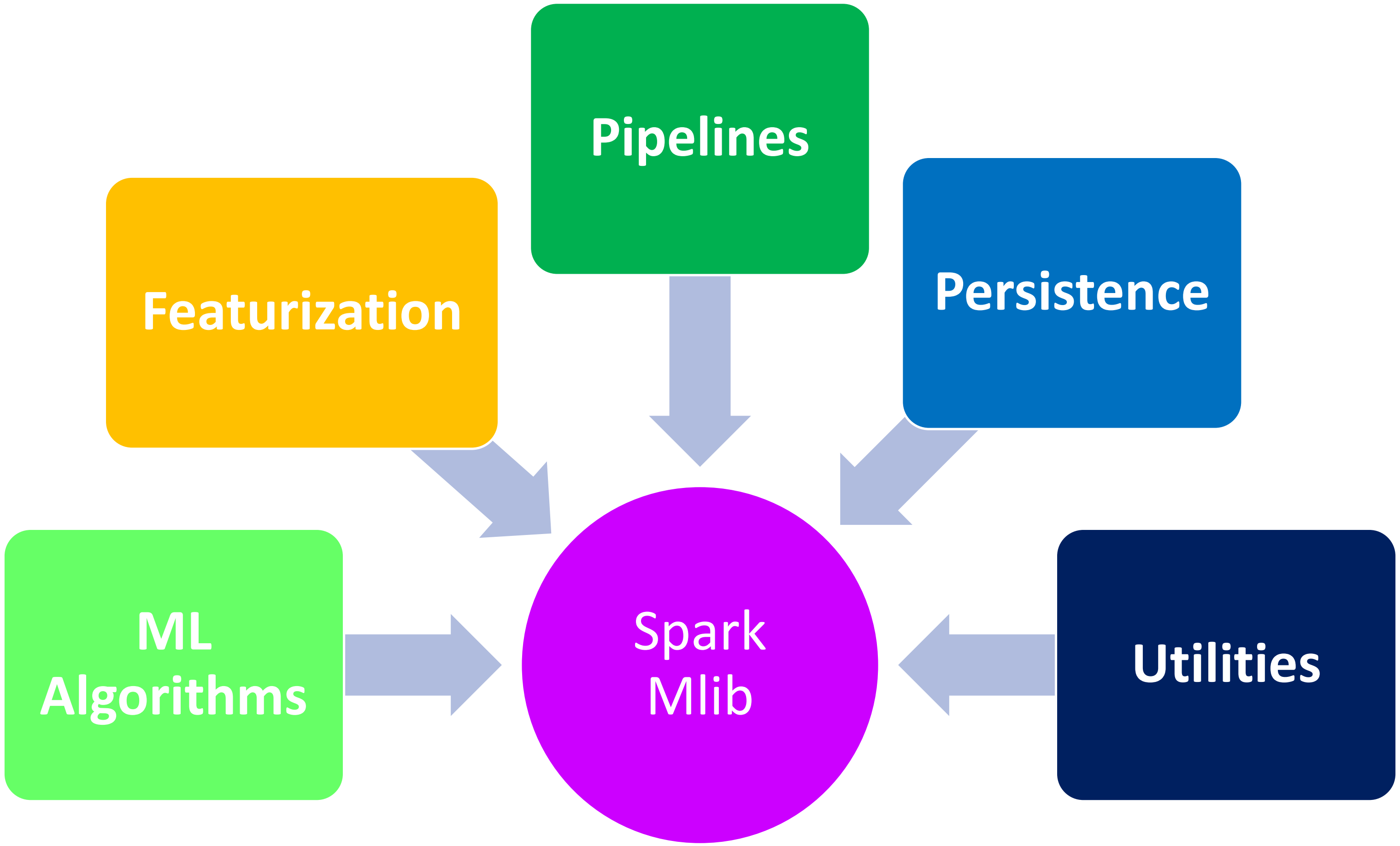
Spark ecosystem



MLib

- It is Spark’s library of machine learning functions designed to run in parallel on clusters.
- Contains variety of learning algorithms
- MLib invokes various algorithms on RDDs
- Soma basic ML algorithms are not included with Spark Mlib as they are not designed for parallel.

Spark MLib


Pipelines
Featurization
Persistence
ML Algorithms
Spark Mlib
Utilities

Spark MLlib Overview

- Divided into two packages:

  - *spark.mllib* contains the original API built on top of RDDs.

  - *spark.ml* provides higher level API built on top of DataFrames.

- Using spark.ml is recommended because with DataFrames, the API is more versatile and flexible.

MLlib Algorithms

- The popular algorithms and utilities in Spark MLlib are:
  - Basic Statistics
  - Regression
  - Classification
  - Clustering
  - Recommendation System
  - Dimensionality Reduction
  - Feature Extraction

MLlib Algorithms - Basic Statistics

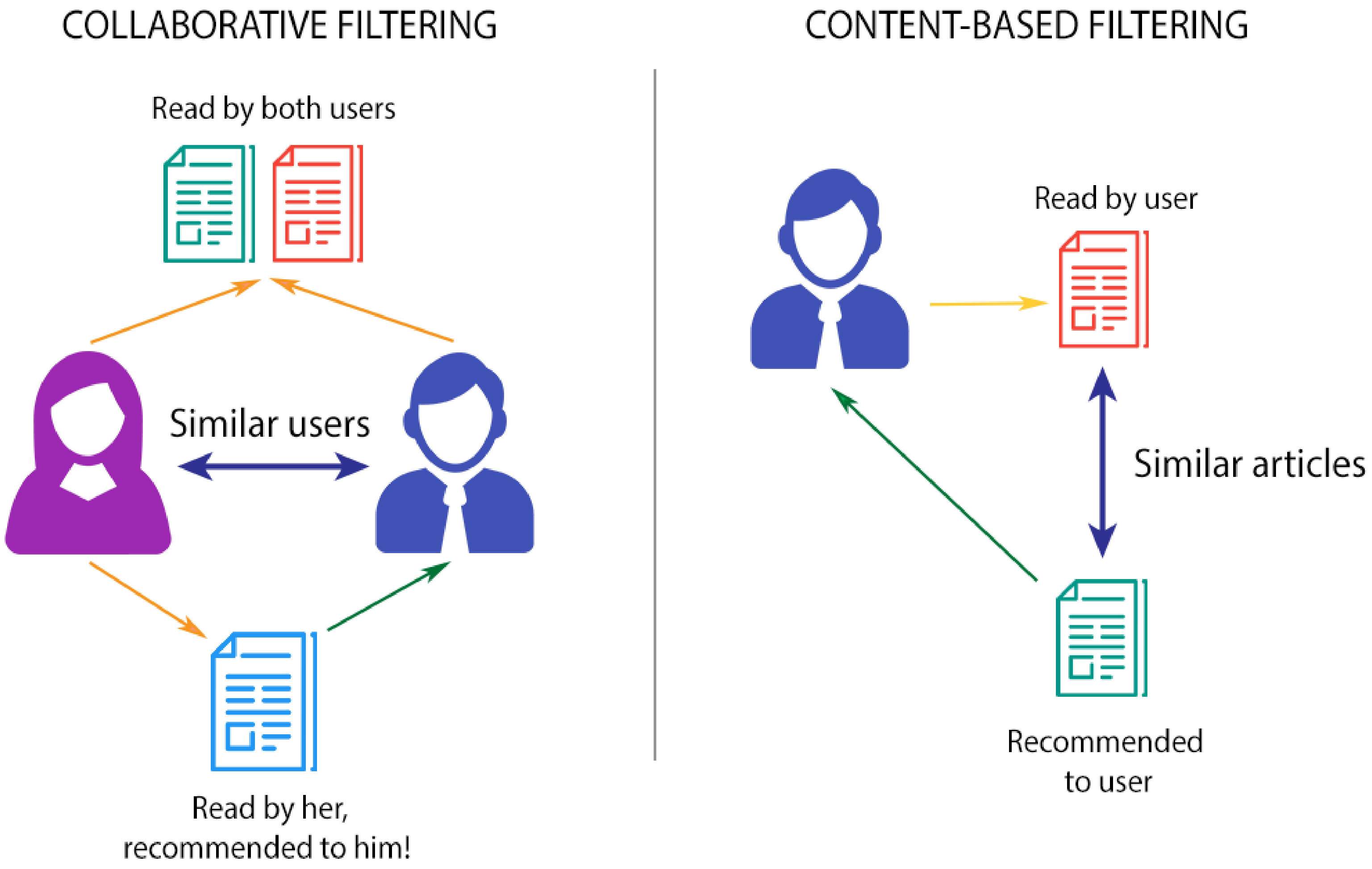
- Basic Statistics includes the most basic machine learning techniques.

  - **Summary Statistics**: provides column summary statistics for RDD[Vector]
    - e.g., include mean, variance, count, max, min and numNonZeros.
  - **Correlations**: pairwise correlations, Spearman and Pearson are some ways to find correlation.
  - **Stratified Sampling**: two methods: sampleBykey and sampleByKeyExact.
  - **Hypothesis Testing**: Pearson's chi-squared test is an example of hypothesis testing.
  - **Random Data Generation**: RandomRDDs, Normal and Poisson are used to generate random data.

MLlib Algorithms - Recommendation System

- A recommendation system provides suggestions to the users through a filtering process that is based on user preferences and browsing history.
- A recommendation system is a subclass of **information filtering system** that seeks to predict the “rating” or “preference” that a user would give to an item.
- Recommender systems have become increasingly popular in recent years and are utilized in a variety of areas.
  - e.g., movies, music, news, books, research articles, search queries, social tags, and products in general.

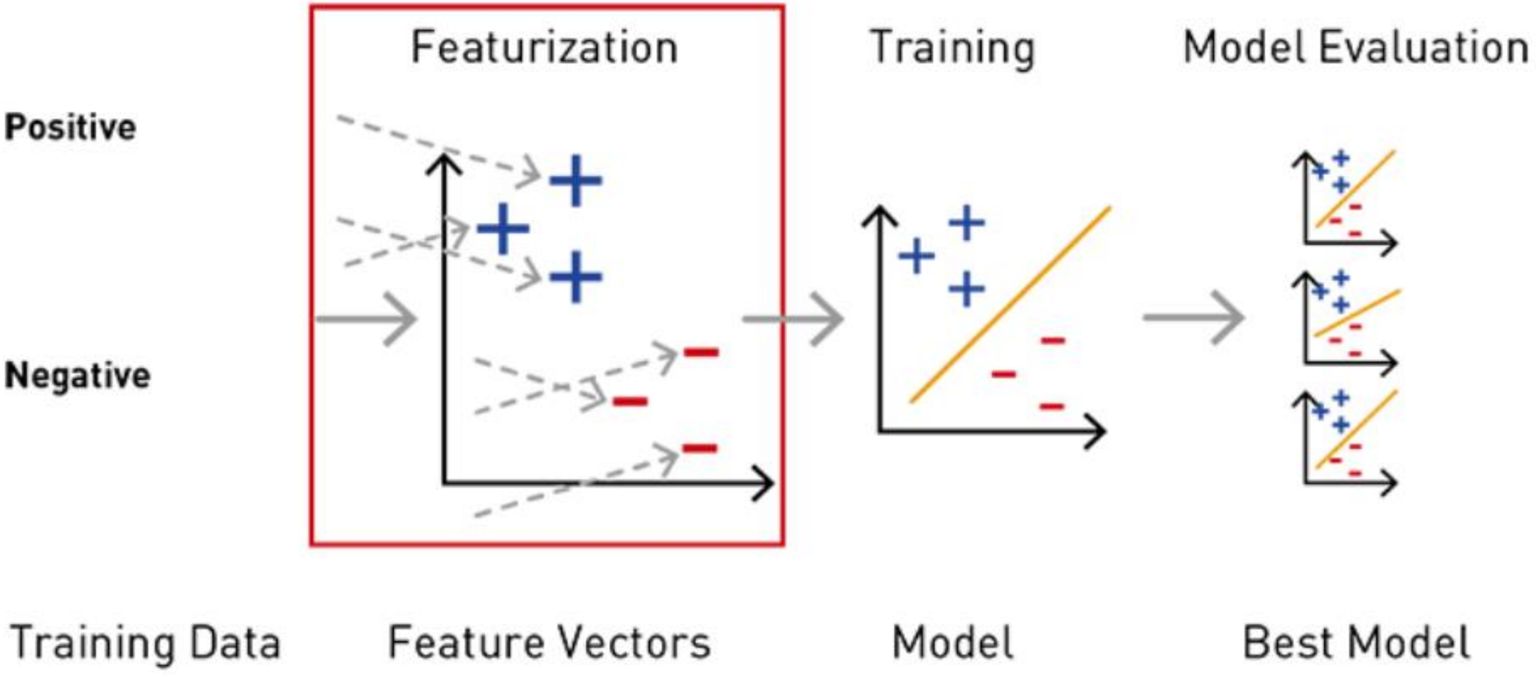
MLlib Algorithms - Recommendation System



MLlib Algorithms - Dimensionality Reduction

- Dimensionality Reduction is the process of reducing the number of random variables under consideration, via obtaining a set of principal variables. It can be divided into **feature selection and feature extraction.**
  - **Feature Selection:** Feature selection finds a subset of the original variables (also called features or attributes).
  - **Feature Extraction:** transforms the data in the high-dimensional space to a space of fewer dimensions.

MLlib Algorithms - Feature Extraction



(Image reference O'Reilly Learning Spark)

MLlib Algorithms - Feature Extraction

- Feature Extraction starts from **an initial set of measured data** and builds derived values (features) intended to be informative and non-redundant, facilitating the **subsequent learning** and generalization steps, and in some cases leading to better human interpretations.
- Algorithms for working with features, roughly divided into these groups:
  1. Extraction: Extracting features from “raw” data
  2. Transformation: Scaling, converting, or modifying features
  3. Selection: Selecting a subset from a larger set of features
  4. Locality Sensitive Hashing (LSH): This class of algorithms combines aspects of feature transformation with other algorithms.

Feature Extraction and transformation

- It contains 5 techniques:

  - Term frequency (TF) and inverse document frequency (IDF)

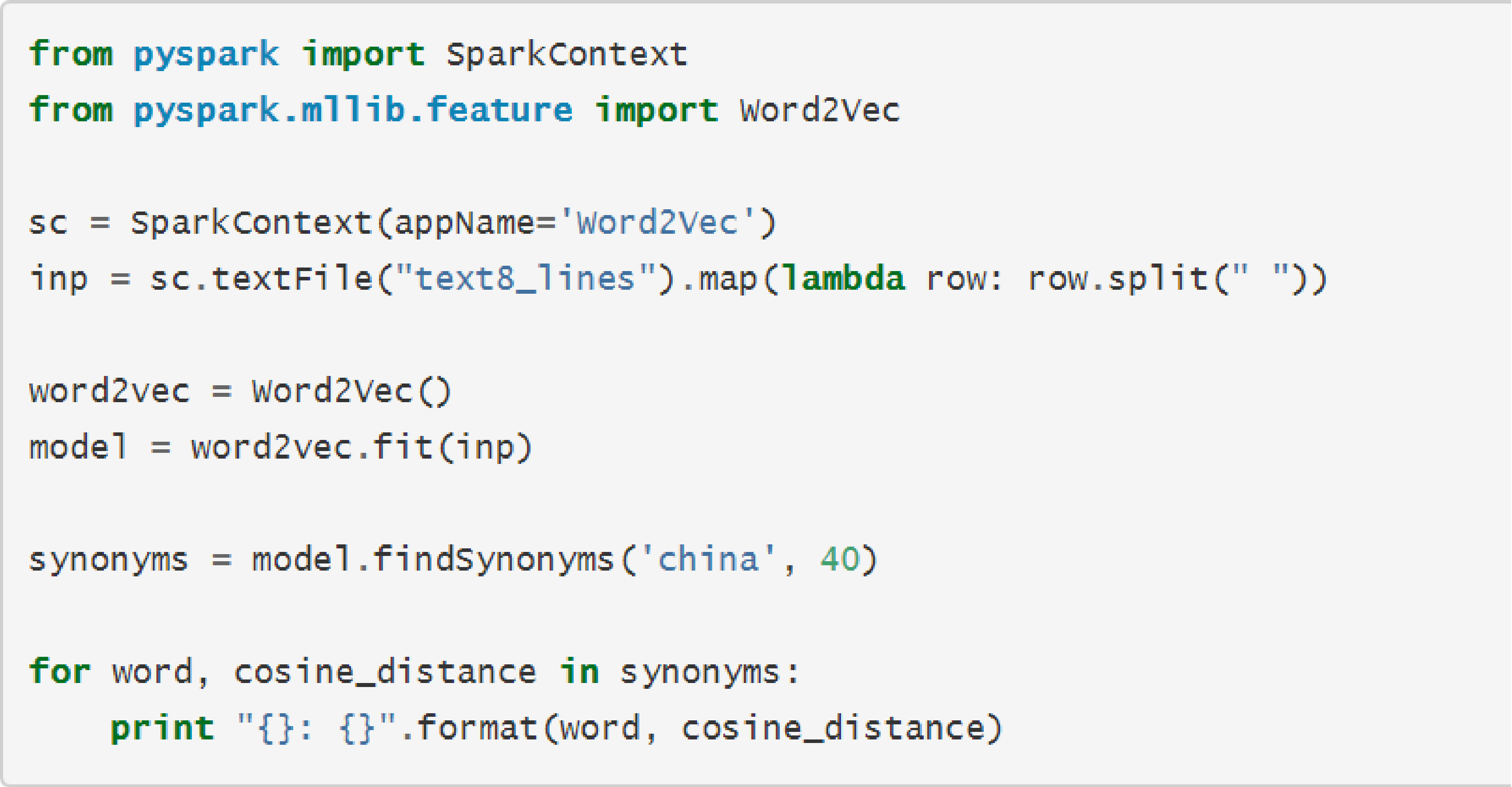
  - Word2Vec (skip-gram and continuous bag of words**)**

  - Standard scaler

  - Normalizer

  - Chi-square Selector

Word2Vec- display the top 40 synonyms of the specified word.



```
from pyspark import SparkContext
from pyspark.mllib.feature import Word2Vec

sc = SparkContext(appName='Word2Vec')
inp = sc.textFile("text8_lines").map(lambda row: row.split(" "))

word2vec = Word2Vec()
model = word2vec.fit(inp)

synonyms = model.findSynonyms('china', 40)

for word, cosine_distance in synonyms:
    print "{}: {}".format(word, cosine_distance)
```

Sentiment Analysis in Disaster Relief

- Sentiment analysis of disaster related posts in social media is one of the techniques that could gear up **detecting posts for situational awareness.**
- Useful to better **understand** the dynamics of the network including **users' feelings, panics and concerns** as it is **used to** identify polarity **of sentiments** expressed by users during disaster events to improve decision making.
- It **helps** authorities to **find answers** to their questions and make better decisions regarding the event assistance without paying the cost as the traditional public surveys.

Sentiment Analysis in Disaster Relief

- Sentiment information could also be **used to project** the information regarding **the devastation and recovery situation** and **donation requests to the crowd** in better ways.
- Using the results obtained from sentiment analysis, **authorities can figure out** where they should look for particular information regarding the disaster such as the most affected areas, types of emergency needs.

Use case – Sentiment analysis from Twitter data

Sentiment analysis

- Sentiment refers to the emotion behind a social media mention online.
- Sentiment analysis is categorizing the tweets related to particular topic and performing data mining using sentiment automation analytics tools.
- We will be performing Twitter Sentiment analysis as our use case for Spark streaming.
- Sentiment analysis helps in disaster management.

Problem statement

- To design a Twitter Sentiment analysis system where we populate real time sentiments for disaster management.
- Sentiment analysis is used to:

  - Identify, extract or characterize subjective information, such as opinions, expressed in a given piece of text.

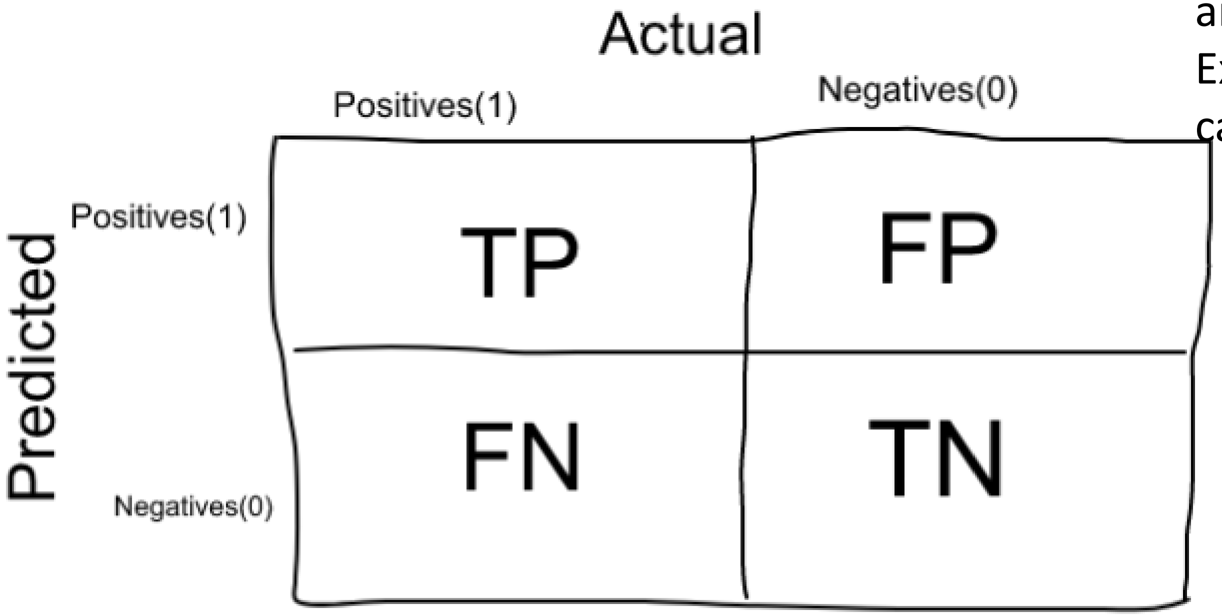
  - Classify a writer's attitude towards various topics into positive, negative or neutral categories.

  - To study and analyze human opinions and sentiment.

Performance metrics for classification

- Confusion Matrix
- Accuracy
- Precision
- Recall or Sensitivity
- Specificity

Confusion matrix



		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

Confusion Matrix

False Positives (FP): False positives are the cases when the actual class of the data point was 0(False) and the predicted is 1(True). False is because the model has predicted incorrectly and positive because the class predicted was a positive one. Ex: A person NOT having cancer and the model classifying his case as cancer comes under False Positives.

True Positives (TP): True positives are the cases when the actual class of the data point was 1(True) and the predicted is also 1(True)
Ex: The case where a person is actually having cancer(1) and the model classifying his case as cancer(1) comes under True positive

True Negatives (TN): True negatives are the cases when the actual class of the data point was 0(False) and the predicted is also 0(False).
Ex: The case where a person NOT having cancer and the model classifying his case as Not cancer comes under True Negatives.

False Negatives (FN): False negatives are the cases when the actual class of the data point was 1(True) and the predicted is 0(False). False is because the model has predicted incorrectly and negative because the class predicted was a negative one. (0)
Ex: A person having cancer and the model classifying his case as No-cancer comes under False Negatives.

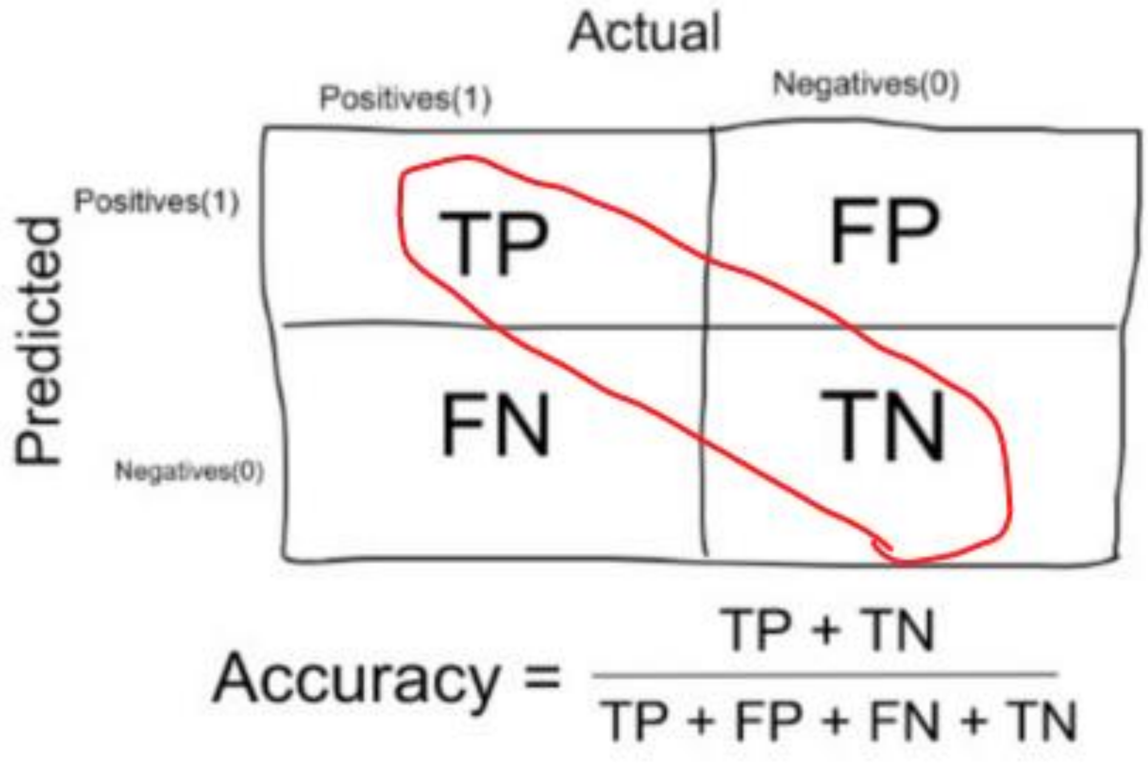
Accuracy

- **When to use Accuracy:**

Accuracy is a good measure when the target variable classes in the data are nearly balanced.
Ex:60% classes in our fruits' images data are apple and 40% are oranges.
A model which predicts whether a new image is Apple or an Orange, 97% of times correctly is a very good measure in this example.

- **When NOT to use Accuracy:**

Accuracy should NEVER be used as a measure when the target variable classes in the data are a majority of one class.



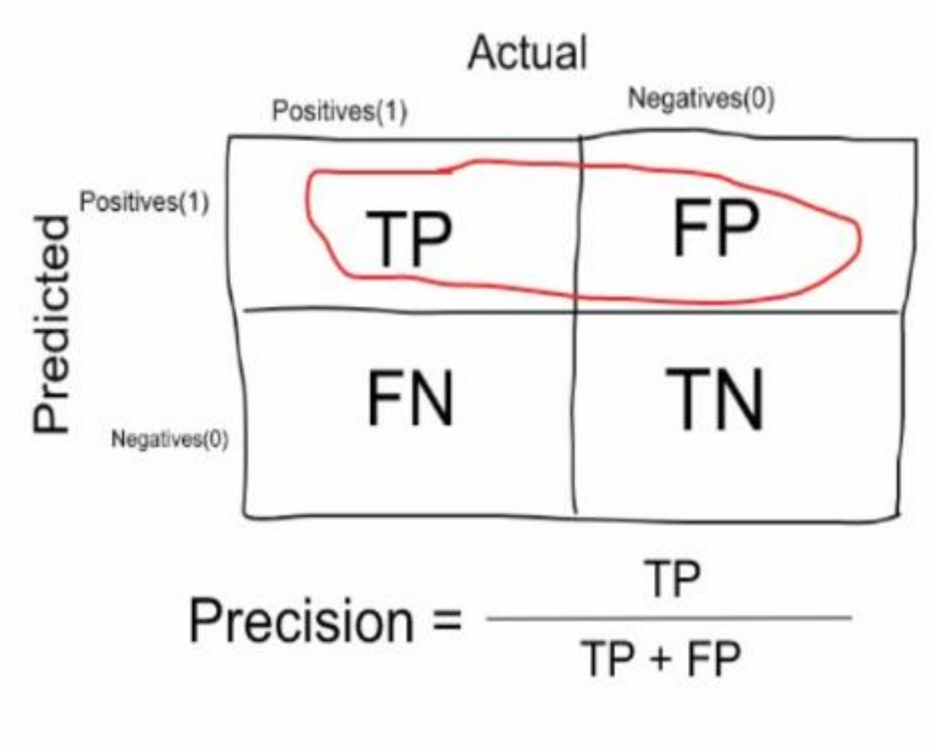
		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

Accuracy

Precision

Precision is a measure that tells us what proportion of patients that we diagnosed as having cancer, actually had cancer. The predicted positives (People predicted as cancerous are TP and FP) and the people actually having a cancer are TP.



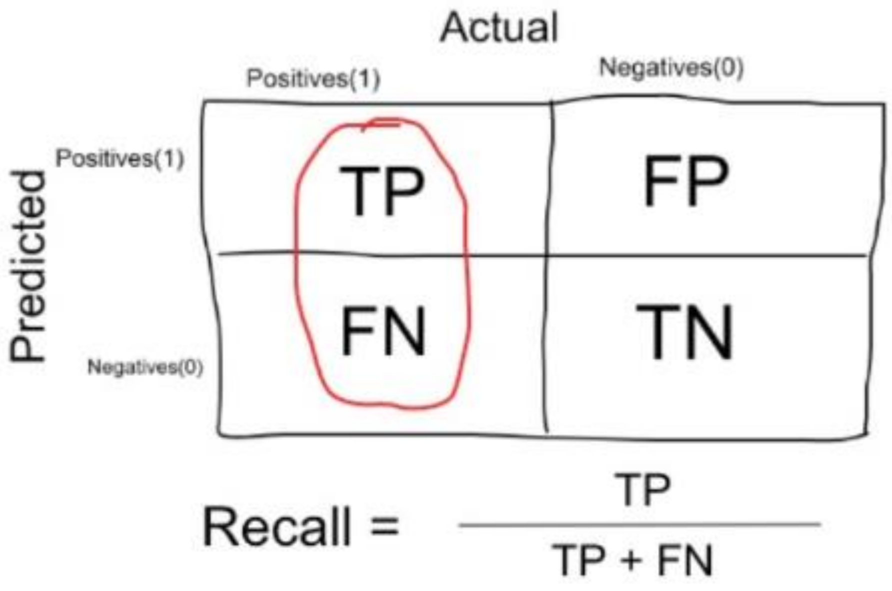
		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

$$\text{Precision} = \frac{TP}{TP + FP}$$

Precision

Recall or Sensitivity

Recall is a measure that tells us what proportion of patients that actually had cancer was diagnosed by the algorithm as having cancer. The actual positives (People having cancer are TP and FN) and the people diagnosed by the model having a cancer are TP. (Note: FN is included because the Person actually had a cancer even though the model predicted otherwise).

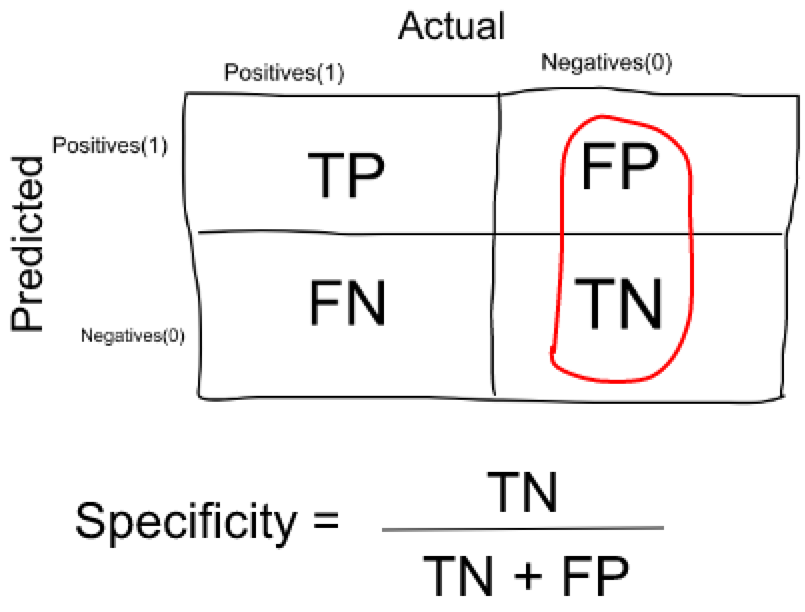


$$\text{Recall} = \frac{TP}{TP + FN}$$

Recall or Sensitivity

Specificity

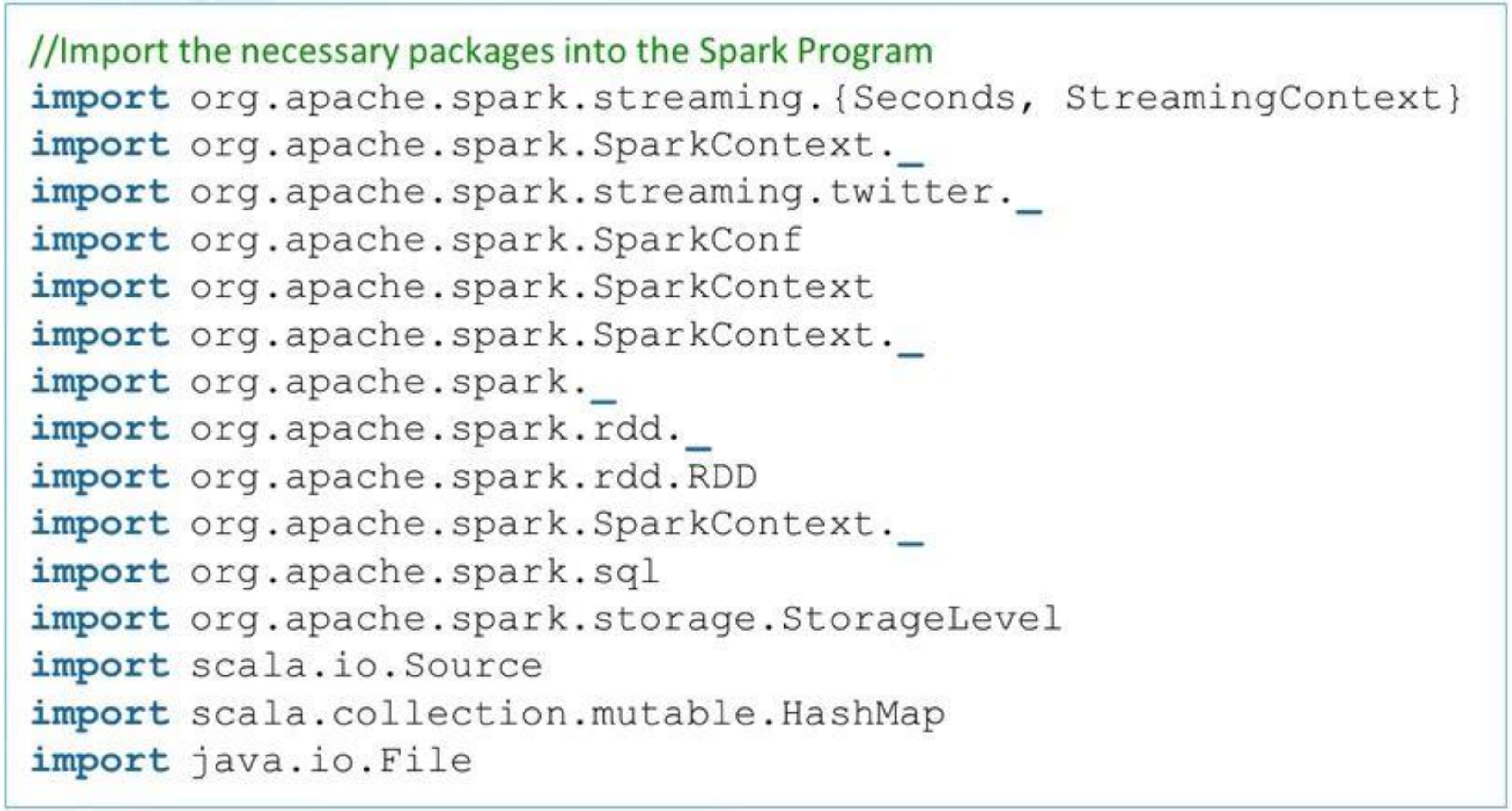
Specificity is a measure that tells us what proportion of patients that did NOT have cancer, were predicted by the model as non-cancerous. The actual negatives (People actually NOT having cancer are FP and TN) and the people diagnosed by us not having cancer are TN. (Note: FP is included because the Person did NOT actually have cancer even though the model predicted otherwise).



		Actual	
		Positives(1)	Negatives(0)
Predicted	Positives(1)	TP	FP
	Negatives(0)	FN	TN

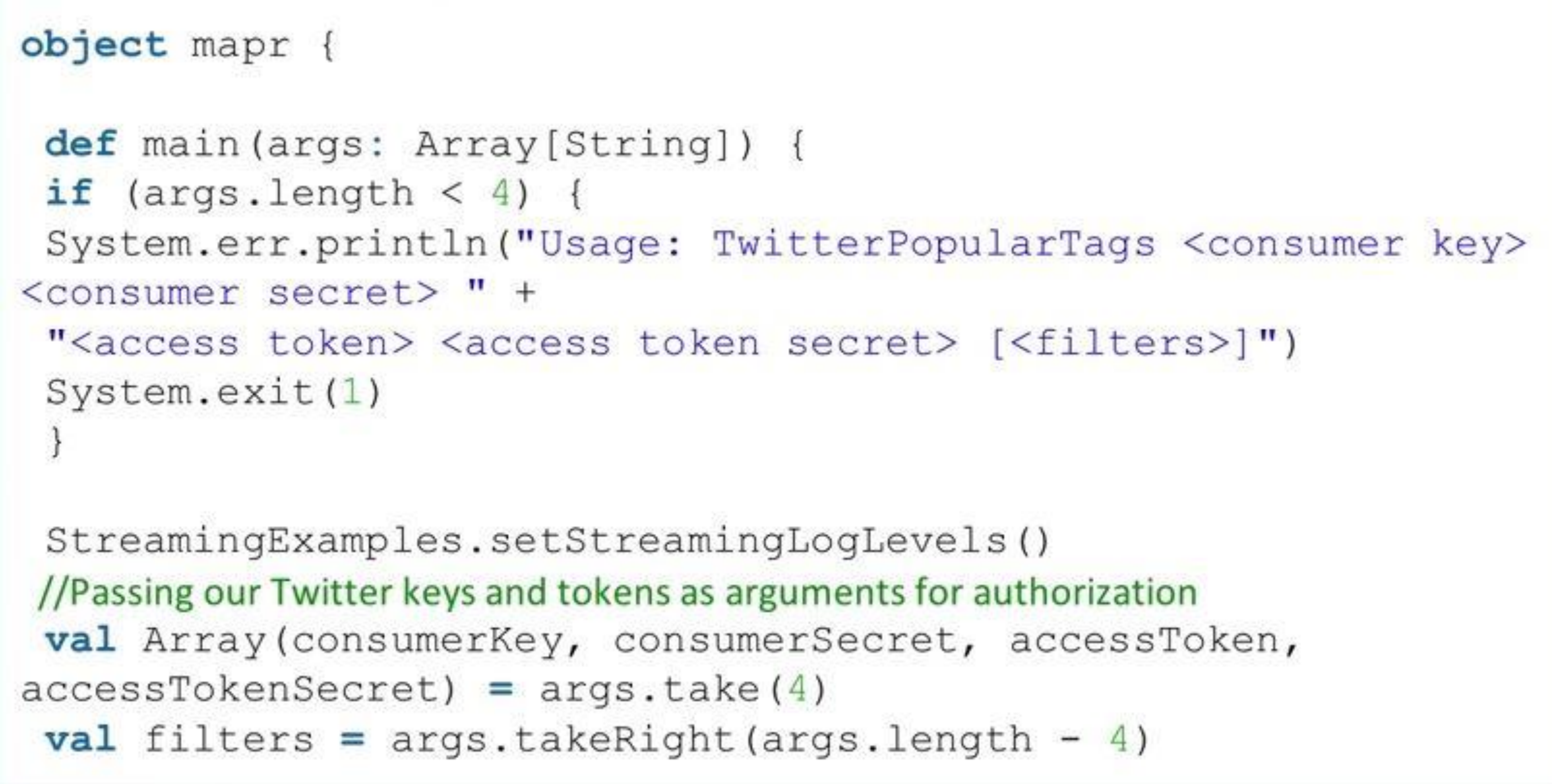
$$\text{Specificity} = \frac{TN}{TN + FP}$$

Use Case – Importing packages



```
//Import the necessary packages into the Spark Program
import org.apache.spark.streaming.{Seconds, StreamingContext}
import org.apache.spark.SparkContext._
import org.apache.spark.streaming.twitter._
import org.apache.spark.SparkConf
import org.apache.spark.SparkContext
import org.apache.spark.SparkContext._
import org.apache.spark._
import org.apache.spark.rdd._
import org.apache.spark.rdd.RDD
import org.apache.spark.SparkContext._
import org.apache.spark.sql
import org.apache.spark.storage.StorageLevel
import scala.io.Source
import scala.collection.mutable.HashMap
import java.io.File
```

Use case- Twitter token authorization

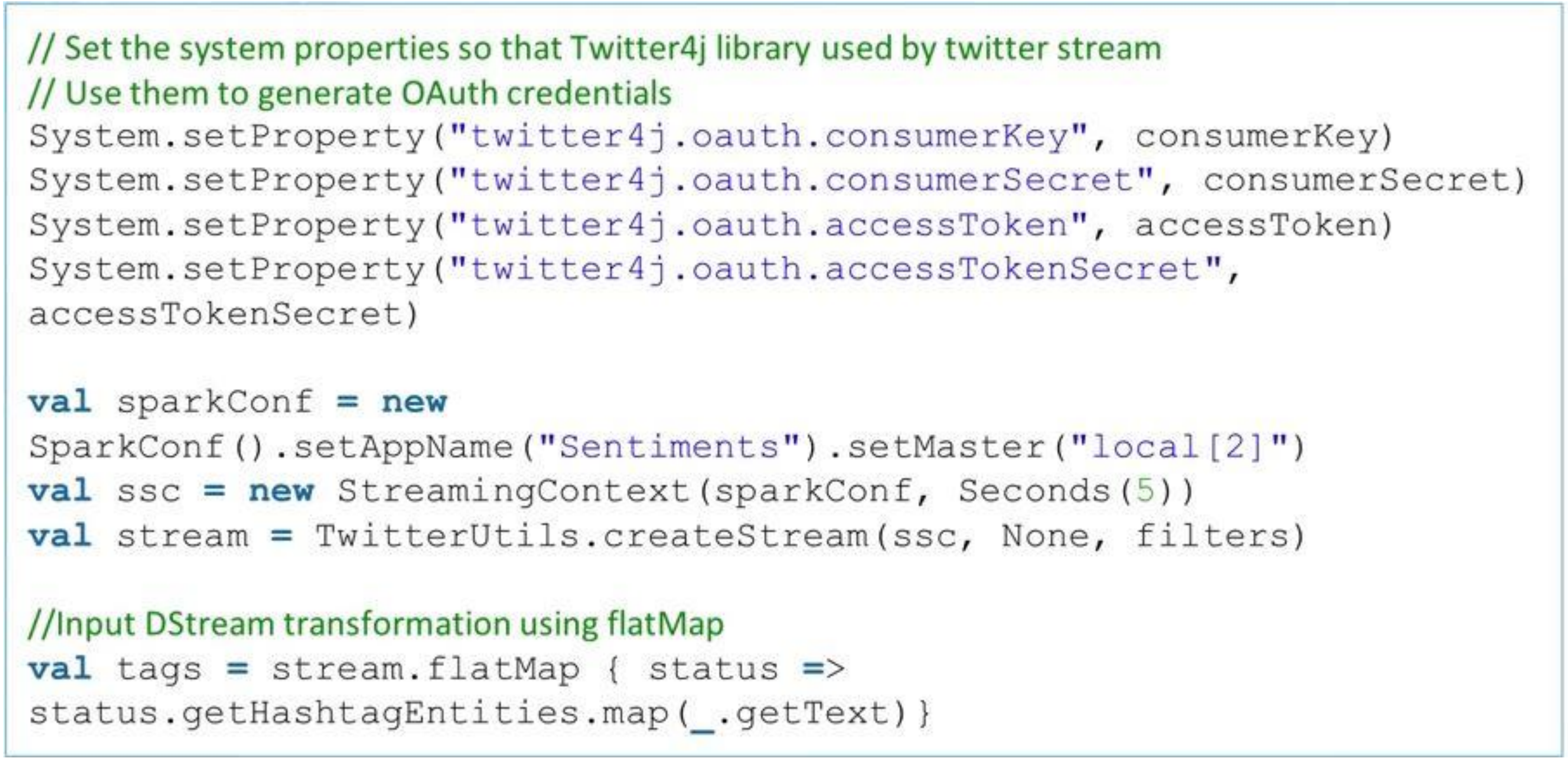


```
object mapr {

 def main(args: Array[String]) {
 if (args.length < 4) {
 System.err.println("Usage: TwitterPopularTags <consumer key>
<consumer secret> " +
 "<access token> <access token secret> [<filters>]")
 System.exit(1)
 }

 StreamingExamples.setStreamingLogLevels()
 //Passing our Twitter keys and tokens as arguments for authorization
 val Array(consumerKey, consumerSecret, accessToken,
accessTokenSecret) = args.take(4)
 val filters = args.takeRight(args.length - 4)
```

Use case- Dstream transformation

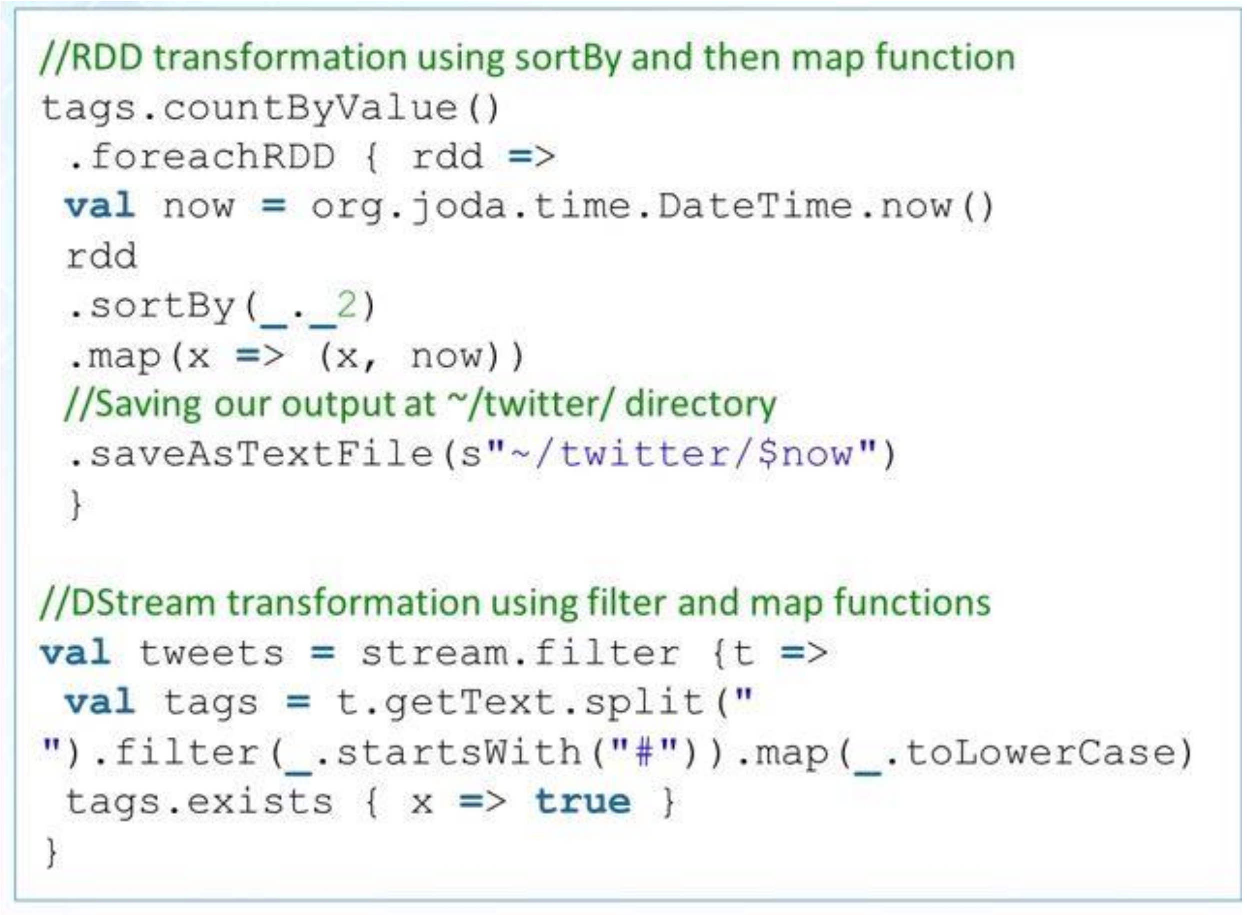


```
// Set the system properties so that Twitter4j library used by twitter stream
// Use them to generate OAuth credentials
System.setProperty("twitter4j.oauth.consumerKey", consumerKey)
System.setProperty("twitter4j.oauth.consumerSecret", consumerSecret)
System.setProperty("twitter4j.oauth.accessToken", accessToken)
System.setProperty("twitter4j.oauth.accessTokenSecret",
accessTokenSecret)

val sparkConf = new
SparkConf().setAppName("Sentiments").setMaster("local[2]")
val ssc = new StreamingContext(sparkConf, Seconds(5))
val stream = TwitterUtils.createStream(ssc, None, filters)

//Input DStream transformation using flatMap
val tags = stream.flatMap { status =>
status.getHashtagEntities.map(_.getText)}
```

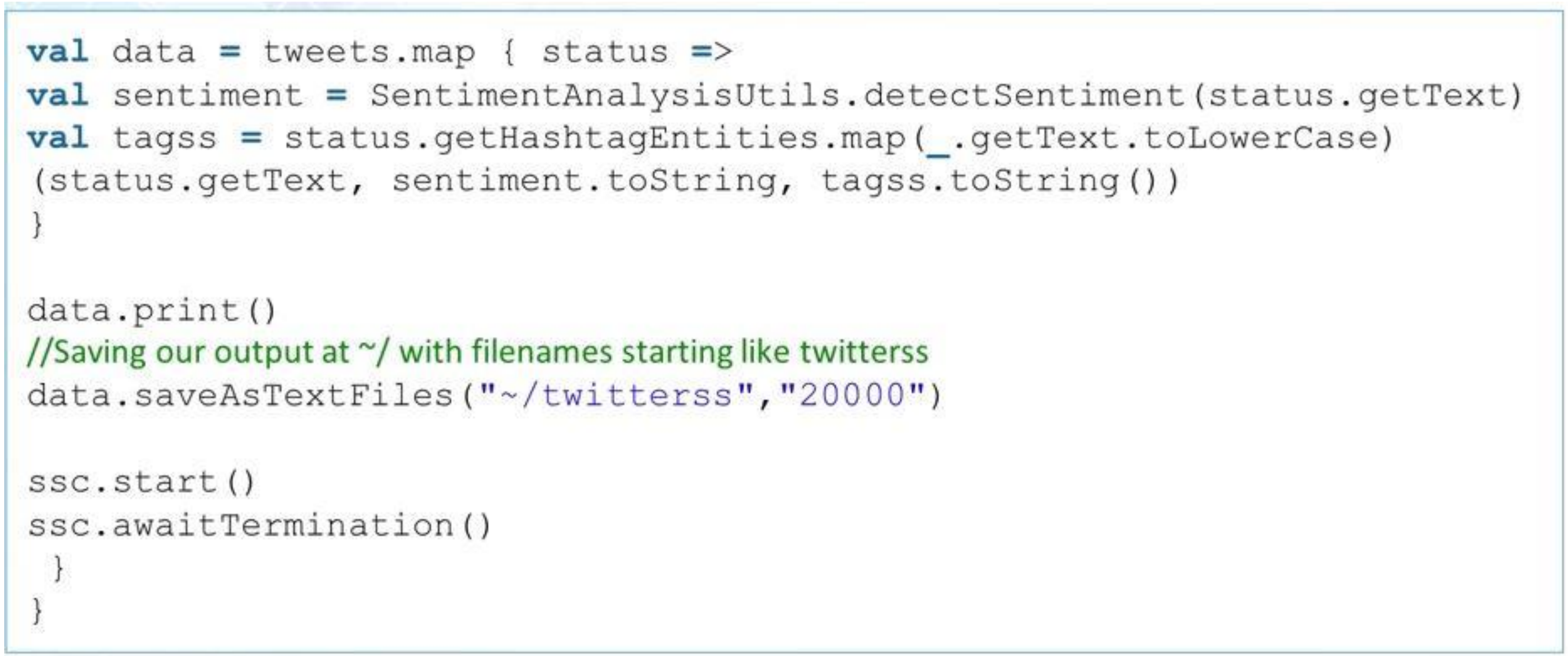
Use case – Generating tweet data



```
//RDD transformation using sortBy and then map function
tags.countByValue()
 .foreachRDD { rdd =>
 val now = org.joda.time.DateTime.now()
 rdd
 .sortBy(_._2)
 .map(x => (x, now))
 //Saving our output at ~/twitter/ directory
 .saveAsTextFile(s"~/twitter/$now")
 }

//DStream transformation using filter and map functions
val tweets = stream.filter {t =>
 val tags = t.getText.split("
").filter(_.startsWith("#")).map(_.toLowerCase)
 tags.exists { x => true }
}
```

Use case – Extracting Sentiments

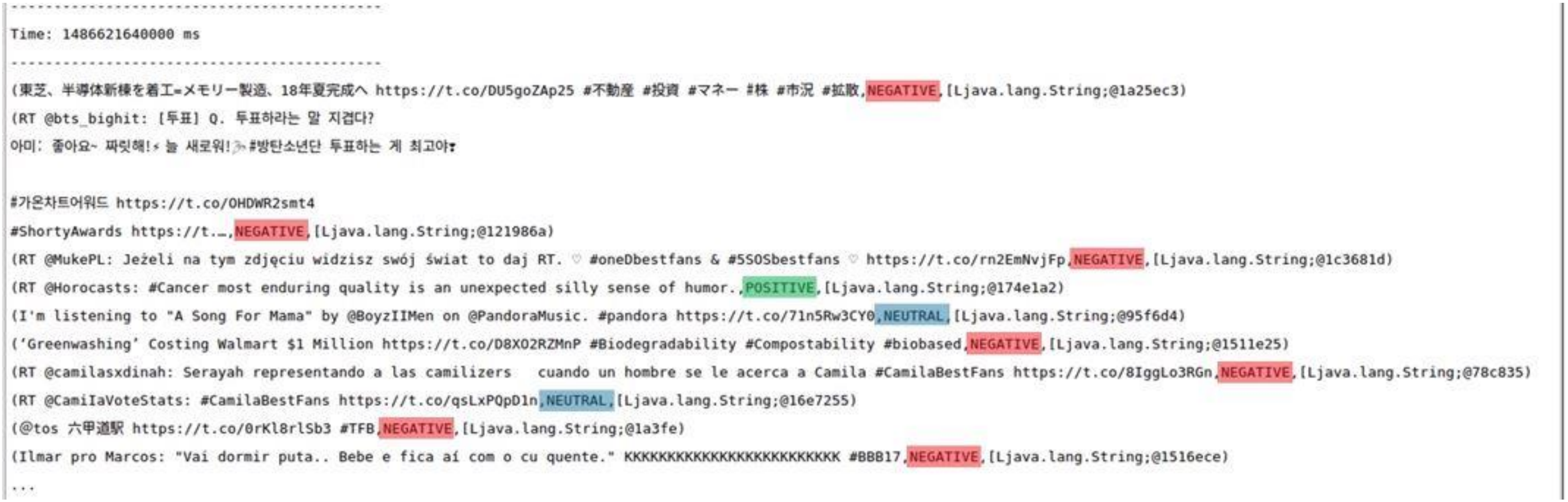


```
val data = tweets.map { status =>
val sentiment = SentimentAnalysisUtils.detectSentiment(status.getText)
val tagss = status.getHashtagEntities.map(_.getText.toLowerCase)
(status.getText, sentiment.toString, tagss.toString())
}

data.print()
//Saving our output at ~/ with filenames starting like twitterss
data.saveAsTextFiles("~/twitterss","20000")

ssc.start()
ssc.awaitTermination()
 }
}
```

Use case - Results



```
-------------------------------------------
Time: 1486621640000 ms
-------------------------------------------
(東芝、半導体新棟を着工=メモリー製造、18年夏完成へ https://t.co/DU5goZAp25 #不動産 #投資 #マネー #株 #市況 #拡散,NEGATIVE,[Ljava.lang.String;@1a25ec3)
(RT @bts_bighit: [투표] Q. 투표하라는 말 지겹다?
아미: 좋아요~ 짜릿해!⚡ 늘 새로워!🎺#방탄소년단 투표하는 게 최고야❣

#가온차트어워드 https://t.co/OHDWR2smt4
#ShortyAwards https://t.…,NEGATIVE,[Ljava.lang.String;@121986a)
(RT @MukePL: Jeżeli na tym zdjęciu widzisz swój świat to daj RT. ♡ #oneDbestfans & #5SOSbestfans ♡ https://t.co/rn2EmNvjFp,NEGATIVE,[Ljava.lang.String;@1c3681d)
(RT @Horocasts: #Cancer most enduring quality is an unexpected silly sense of humor.,POSITIVE,[Ljava.lang.String;@174e1a2)
(I'm listening to "A Song For Mama" by @BoyzIIMen on @PandoraMusic. #pandora https://t.co/71n5Rw3CY0,NEUTRAL,[Ljava.lang.String;@95f6d4)
('Greenwashing' Costing Walmart $1 Million https://t.co/D8XO2RZMnP #Biodegradability #Compostability #biobased,NEGATIVE,[Ljava.lang.String;@1511e25)
(RT @camilasxdinah: Serayah representando a las camilizers   cuando un hombre se le acerca a Camila #CamilaBestFans https://t.co/8IggLo3RGn,NEGATIVE,[Ljava.lang.String;@78c835)
(RT @CamiIaVoteStats: #CamilaBestFans https://t.co/qsLxPQpD1n,NEUTRAL,[Ljava.lang.String;@16e7255)
(@tos 六甲道駅 https://t.co/0rKl8rlSb3 #TFB,NEGATIVE,[Ljava.lang.String;@1a3fe)
(Ilmar pro Marcos: "Vai dormir puta.. Bebe e fica aí com o cu quente." KKKKKKKKKKKKKKKKKKKKKKKKKK #BBB17,NEGATIVE,[Ljava.lang.String;@1516ece)
...
```

🟩 Positive

🟦 Neutral

🟥 Negative

Figure: Output file containing tweet and its sentiment

Conclusions

- Machine learning
- Natural language processing
- Spark MLib
- Usecase