



Learning Apache Spark with Python

Release v1.0

Wenqiang Feng

September 24, 2017

CONTENTS

1	Preface	3
1.1	About	3
1.2	Motivation for this tutorial	3
1.3	Acknowledgement	4
1.4	Feedback and suggestions	4
2	Why Spark with Python ?	5
2.1	Why Spark?	5
2.2	Why Spark with Python (PySpark)?	7
3	Configure Running Platform	9
3.1	Run on Databricks Community Cloud	9
3.2	Configure Spark on Mac and Ubuntu	14
3.3	Configure Spark on Windows	17
3.4	PySpark With Text Editor or IDE	17
3.5	Set up Spark on Cloud	18
3.6	Demo Code in this Section	18
4	An Introduction to Apache Spark	21
4.1	Core Concepts	21
4.2	Spark Components	21
4.3	Architecture	24
4.4	How Spark Works?	24
5	Programming with RDDs	25
5.1	Create RDD	25
5.2	Spark Transformations	28
5.3	Spark Actions	28
6	Statistics Preliminary	29
6.1	Notations	29
6.2	Measurement Formula	29
7	Regression	31
7.1	Linear Regression	31
7.2	Generalized linear regression	37

7.3	Decision tree Regression	42
7.4	Random Forest Regression	47
7.5	Gradient-boosted tree regression	47
8	Classification	49
8.1	Logistic regression	49
8.2	Decision tree Classification	49
8.3	Random forest Classification	49
8.4	Gradient-boosted tree Classification	49
8.5	Naive Bayes Classification	49
8.6	Support Vector Machines Classification	50
9	Clustering	51
9.1	K-Means Model	51
10	Text Mining	53
10.1	Source of text	53
10.2	Text Preprocessing	56
10.3	Text Classification	59
10.4	Sentiment analysis	59
10.5	N-grams and Correlations	59
10.6	Topic Model: Latent Dirichlet Allocation	59
11	Social Network Analysis	61
11.1	Co-occurrence Network	61
11.2	Correlation Network	64
12	Neural Network	65
12.1	Feedforward Neural Network	65
13	Main Reference	69
	Bibliography	71
	Index	73



Welcome to our **Learning Apache Spark with Python** note! In these note, you will learn a wide array of concepts about **PySpark** in Data Mining, Text Mining, Machine Learning and Deep Learning.

1.1 About

1.1.1 About this note

This is a shared repository for Learning Apache Spark Notes. The first version was posted on Github in [\[Feng2017\]](#). This shared repository mainly contains the self-learning and self-teaching notes from Wenqiang during his [IMA Data Science Fellowship](#).

In this repository, I try to use the detailed demo code and examples to show how to use each main functions. If you find your work wasn't cited in this note, please feel free to let me know.

Although I am by no means an data mining programming and Big Data expert, I decided that it would be useful for me to share what I learned about PySpark programming in the form of easy tutorials with detailed example. I hope those tutorials will be a valuable tool for your studies.

The tutorials assume that the reader has a preliminary knowledge of programing and Linux. And this document is generated automatically by using [sphinx](#).

1.1.2 About the authors

- **Wenqiang Feng**
 - Data Scientist and Phd in Mathematics
 - University of Tennessee at Knoxville
 - Email: wfeng1@utk.edu

1.2 Motivation for this tutorial

I was motivated by the [IMA Data Science Fellowship](#) project to learn PySpark. After that I was impressed and attracted by the PySpark. And I foud that:

1. It is no exaggeration to say that Spark is the most powerful Bigdata tool.

2. However, I still found that learning Spark was a difficult process. I have to Google it and identify which one is true. And it was hard to find detailed examples which I can easily learned the full process in one file.
3. Good sources are expensive for a graduate student.

1.3 Acknowledgement

At here, I would like to thank Ming Chen, Jian Sun and Zhongbo Li at the University of Tennessee at Knoxville for the valuable discussion and thank the generous anonymous authors for providing the detailed solutions and source code on the internet. Without those help, this repository would not have been possible to be made. Wenqiang also would like to thank the [Institute for Mathematics and Its Applications \(IMA\)](#) at [University of Minnesota, Twin Cities](#) for support during his IMA Data Scientist Fellow visit.

1.4 Feedback and suggestions

Your comments and suggestions are highly appreciated. I am more than happy to receive corrections, suggestions or feedbacks through email (wfeng1@utk.edu) for improvements.

WHY SPARK WITH PYTHON ?

Note: Sharpening the knife longer can make it easier to hack the firewood – old Chinese proverb

I want to answer this question from the following two parts:

2.1 Why Spark?

I think the following four main reasons from [Apache Spark™](#) official website are good enough to convince you to use Spark.

1. Speed

Run programs up to 100x faster than Hadoop MapReduce in memory, or 10x faster on disk.

Apache Spark has an advanced DAG execution engine that supports acyclic data flow and in-memory computing.

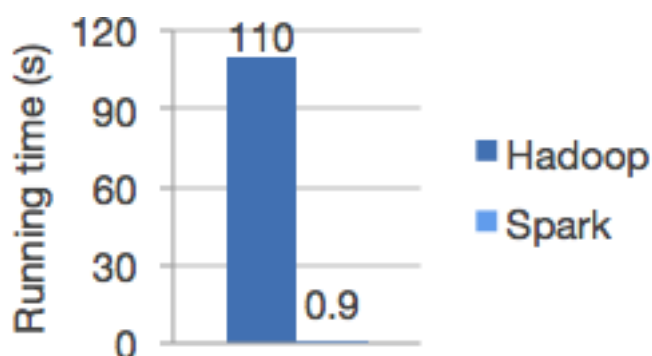


Figure 2.1: Logistic regression in Hadoop and Spark

2. Ease of Use

Write applications quickly in Java, Scala, Python, R.

Spark offers over 80 high-level operators that make it easy to build parallel apps. And you can use it interactively from the Scala, Python and R shells.

3. Generality

Combine SQL, streaming, and complex analytics.

Spark powers a stack of libraries including SQL and DataFrames, MLlib for machine learning, GraphX, and Spark Streaming. You can combine these libraries seamlessly in the same application.

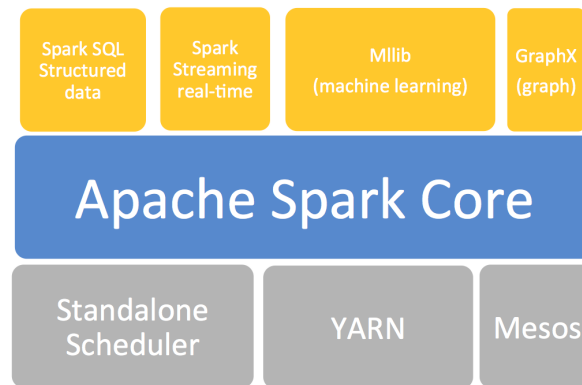


Figure 2.2: The Spark stack

4. Runs Everywhere

Spark runs on Hadoop, Mesos, standalone, or in the cloud. It can access diverse data sources including HDFS, Cassandra, HBase, and S3.



Figure 2.3: The Spark platform

2.2 Why Spark with Python (PySpark)?

No matter you like it or not, Python has been one of the most popular programming languages.

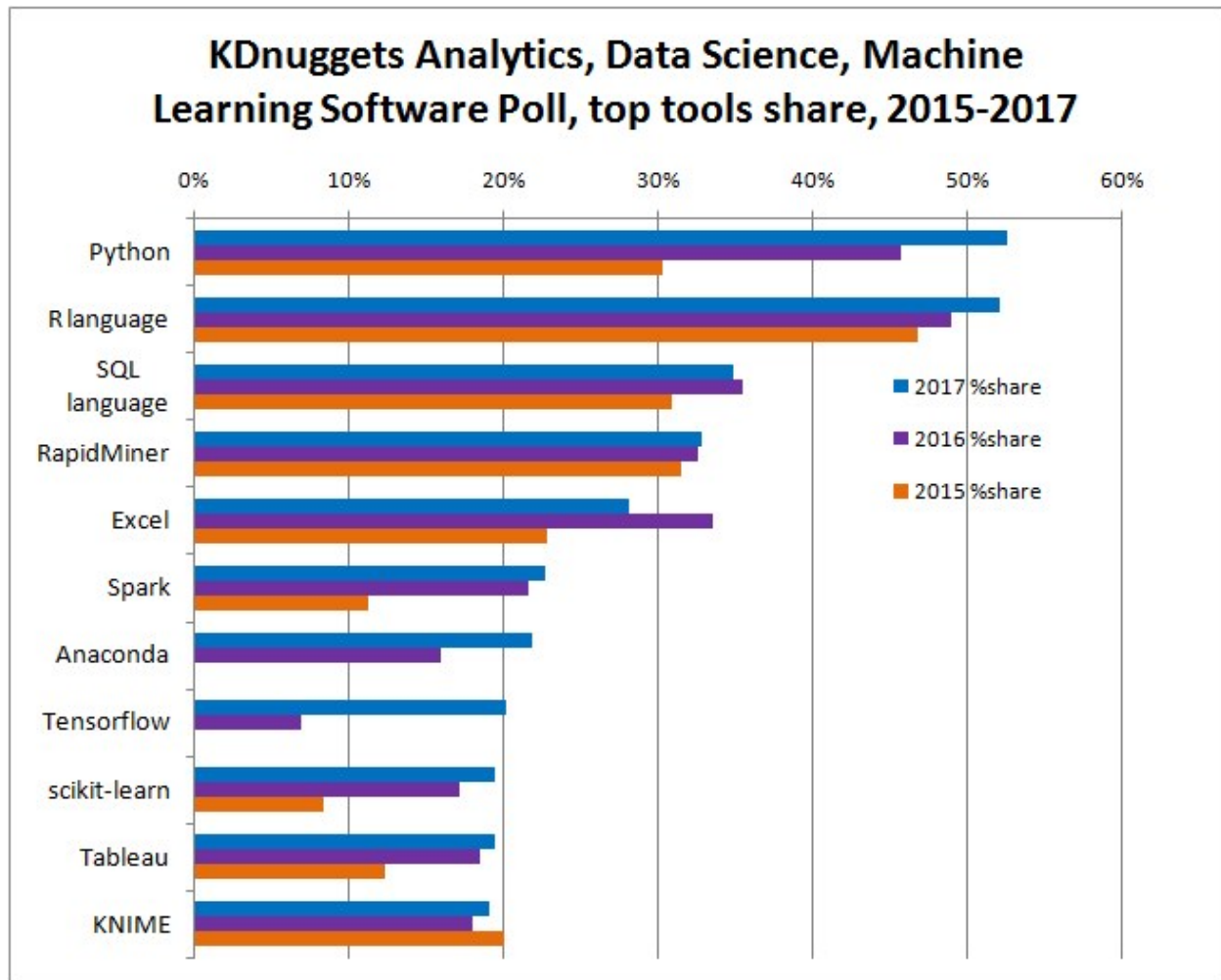


Figure 2.4: KDnuggets Analytics/Data Science 2017 Software Poll from [kdnuggets](https://www.kdnuggets.com/2017/07/data-science-software-poll.html).

CONFIGURE RUNNING PLATFORM

Note: Good tools are prerequisite to the successful execution of a job. – old Chinese proverb

A good programming platform can save you lots of troubles and time. Herein I will only present how to install my favorite programming platform and only show the easiest way which I know to set it up on Linux system. If you want to install on the other operator system, you can Google it. In this section, you may learn how to set up Pyspark on the corresponding programming platform and package.

3.1 Run on Databricks Community Cloud

If you don't have any experience with Linux or Unix operator system, I would love to recommend you to use Spark on Databricks Community Cloud. Since you do not need to setup the Spark and it's totally **free** for Community Edition. Please follow the steps listed below.

1. Sign up a account at: <https://community.cloud.databricks.com/login.html>
2. Sign in with your account, then you can creat your cluster(machine), table(dataset) and notebook(code).
3. Create your cluster where your code will run
4. Import your dataset

Note: You need to save the path which appears at Uploaded to DBFS: /File-Store/tables/05rmhuqv1489687378010/. Since we will use this path to load the dataset.

5. Creat your notebook



File Edit View History Tools People Help

Databricks

Secure | <https://community.cloud.databricks.com/?o=4622560542654492>

Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments

Upgrade ? @George

Community Edition (2.45)

Welcome to databricks™

Featured Notebooks

- Introduction to Apache Spark on Databricks
- Databricks for Data Scientists
- Introduction to Structured Streaming

New

- Notebook
- Job
- Cluster
- Table
- Library

Documentation

- Databricks Guide
- Python, R, Scala, SQL
- Importing Data

Open Recent

- Databricks for Data Scientists
- linearRegression

What's new?

- Automatic termination of clusters
- Autoscaling local storage with EBS volumes (beta)

[Latest release notes](#)

Send Feedback

File Edit View History Tools People Help

Create Cluster - Databricks

Secure | <https://community.cloud.databricks.com/?o=4622560542654492#create/cluster>

Apps Bookmarks job Foreign National The FORTRAN Pr Fortran Tutorials Using the cluster Attachments

Upgrade ? @George

Create Cluster

New Cluster

Cancel Create Cluster

0 Workers: 0.0 GB Memory, 0 Cores, 0 DBU
1 Driver: 6.0 GB Memory, 0.88 Cores, 1 DBU

Cluster Name
MLmachine

Databricks Runtime Version
Spark 2.1 (Auto-updating, Scala 2.10)

Instance

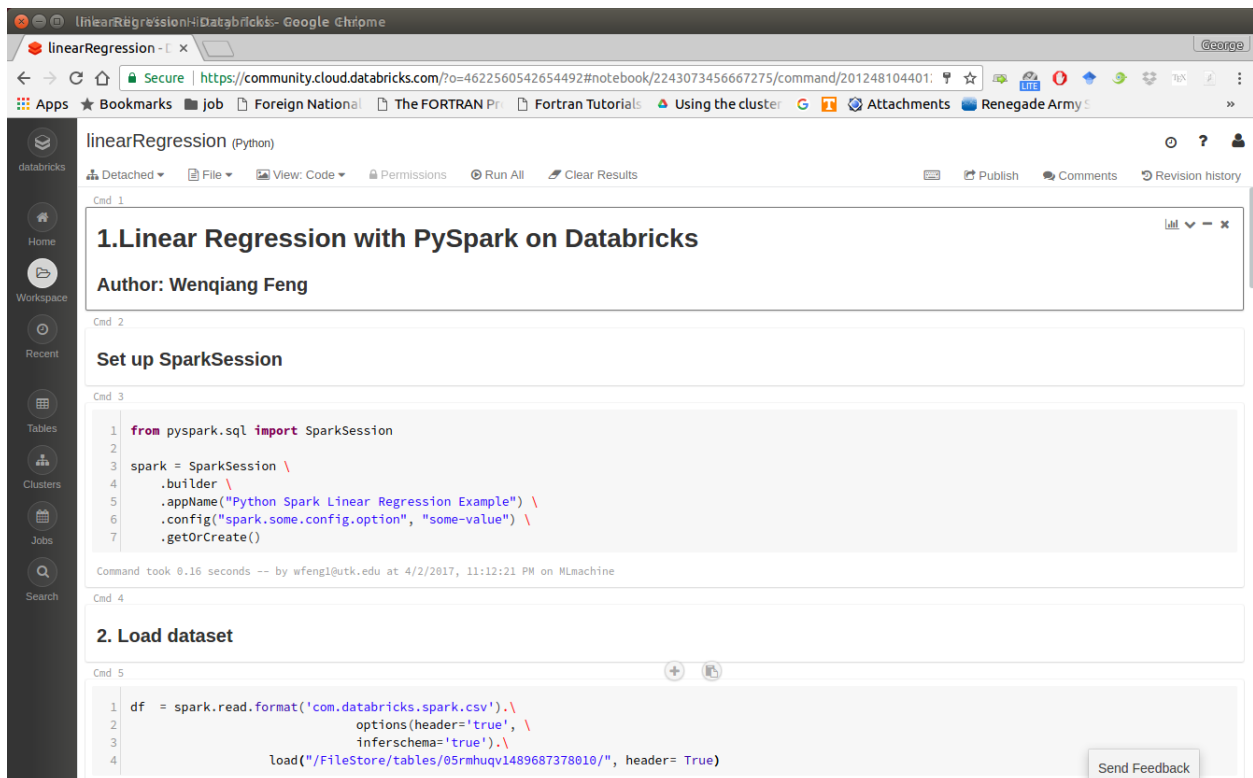
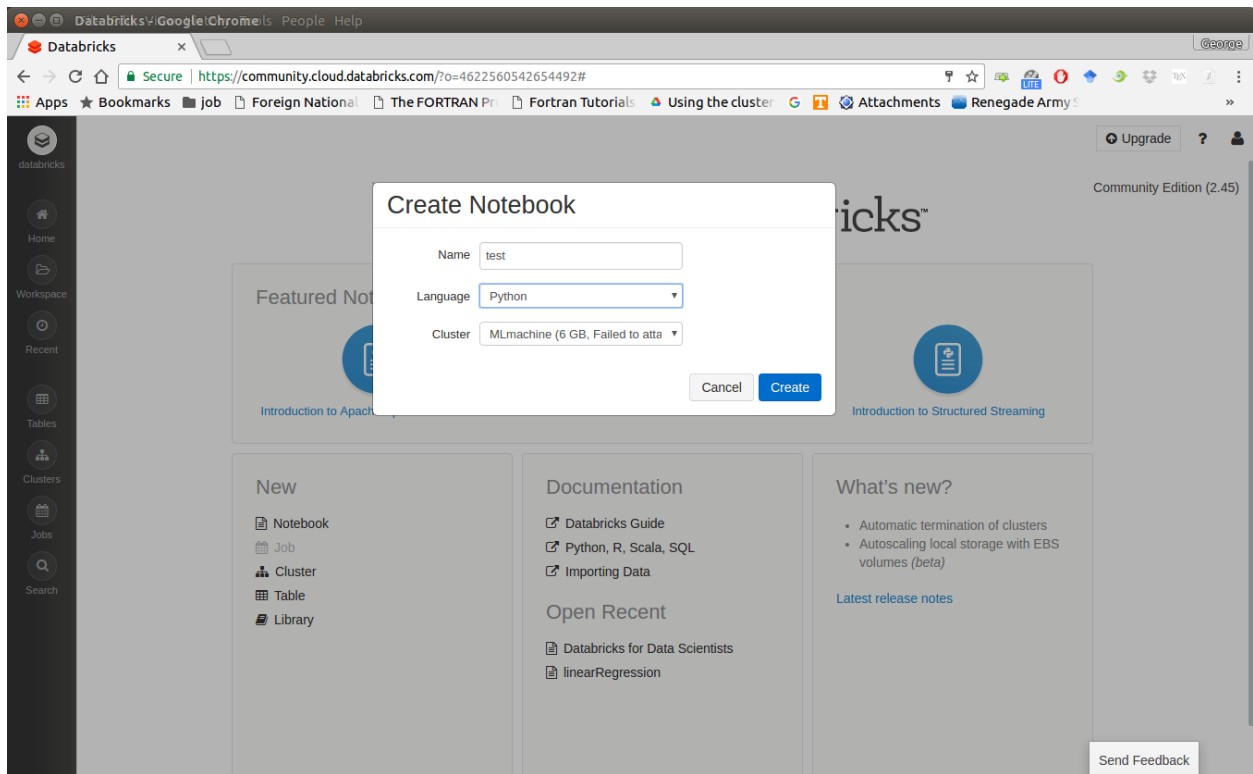
Free 6GB Memory: As a Community Edition user, your cluster will automatically terminate after an idle period of two hours. For [more configuration options](#), please [upgrade your Databricks subscription](#).

AWS Spark

Availability Zone
us-west-2c

Send Feedback





After finishing the above 5 steps, you are ready to run your Spark code on Databricks Community Cloud. I will run all the following demos on Databricks Community Cloud. Hopefully, when you run the demo code, you will get the following results:

```
+---+-----+-----+-----+-----+
|_c0|      TV|Radio|Newspaper|Sales|
+---+-----+-----+-----+-----+
|  1|230.1| 37.8|      69.2| 22.1|
|  2| 44.5| 39.3|      45.1| 10.4|
|  3| 17.2| 45.9|      69.3|  9.3|
|  4|151.5| 41.3|      58.5| 18.5|
|  5|180.8| 10.8|      58.4| 12.9|
+---+-----+-----+-----+-----+
only showing top 5 rows

root
|-- _c0: integer (nullable = true)
|-- TV: double (nullable = true)
|-- Radio: double (nullable = true)
|-- Newspaper: double (nullable = true)
|-- Sales: double (nullable = true)
```

3.2 Configure Spark on Mac and Ubuntu

3.2.1 Installing Prerequisites

I will strongly recommend you to install [Anaconda](#), since it contains most of the prerequisites and support multiple Operator Systems.

1. Install Python

Go to Ubuntu Software Center and follow the following steps:

1. Open Ubuntu Software Center
2. Search for python
3. And click Install

Or Open your terminal and using the following command:

```
sudo apt-get install build-essential checkinstall
sudo apt-get install libreadline-gplv2-dev libncursesw5-dev libssl-dev
                        libsqlite3-dev tk-dev libgdbm-dev libc6-dev libbz2-dev
sudo apt-get install python
sudo easy_install pip
sudo pip install ipython
```

3.2.2 Install Java

Java is used by many other softwares. So it is quite possible that you have already installed it. You can try using the following command in Command Prompt:

```
java -version
```

Otherwise, you can follow the steps in [How do I install Java for my Mac?](#) to install java on Mac and use the following command in Command Prompt to install on Ubuntu:

```
sudo apt-add-repository ppa:webupd8team/java
sudo apt-get update
sudo apt-get install oracle-java8-installer
```

3.2.3 Install Java SE Runtime Environment

I installed ORACLE [Java JDK](#).

Note: Installing Java and Java SE Runtime Environment steps are very important, since Spark is a domain-specific language written in Java.

You can check if your Java is available and find its version by using the following command in Command Prompt:

```
java -version
```

If your Java is installed successfully, you will get the similar results as follows:

```
java version "1.8.0_131"
Java(TM) SE Runtime Environment (build 1.8.0_131-b11)
Java HotSpot(TM) 64-Bit Server VM (build 25.131-b11, mixed mode)
```

3.2.4 Install Apache Spark

Actually, the Pre-build version doesn't need installation. You can use it when you unpack it.

1. Download: You can get the Pre-built Apache Spark™ from [Download Apache Spark™](#).
2. Unpack: Unpack the Apache Spark™ to the path where you want to install the Spark.
3. Test: Test the Prerequisites: change the direction
spark-#.##-bin-hadoop#.##/bin and run

```
./pyspark
```

```
Python 2.7.13 |Anaconda 4.4.0 (x86_64)| (default, Dec 20 2016, 23:05:08)
[GCC 4.2.1 Compatible Apple LLVM 6.0 (clang-600.0.57)] on darwin
Type "help", "copyright", "credits" or "license" for more information.
Anaconda is brought to you by Continuum Analytics.
Please check out: http://continuum.io/thanks and https://anaconda.org
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
```


3.3 Configure Spark on Windows

Installing open source software on Windows is always a nightmare for me. Thanks for Deelesh Mandloi. You can follow the detailed procedures in the blog [Getting Started with PySpark on Windows](#) to install the Apache Spark™ on your Windows Operator System.

3.4 PySpark With Text Editor or IDE

3.4.1 PySpark With Jupyter Notebook

After you finishing the above setup steps in *Configure Spark on Mac and Ubuntu*, then you should be good to use write and run your PySpark Code in Jupyter notebook.



```
In [1]: ## set up SparkSession
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

df = spark.read.format('com.databricks.spark.csv') \
    .options(header='true', \
              inferSchema='true') \
    .load("/home/feng/Spark/Code/data/Advertising.csv", header=True)

df.show(5)
df.printSchema()

+---+-----+-----+-----+
|_c0|  TV|Radio|Newspaper|Sales|
+---+-----+-----+-----+
| 1|230.1| 37.8|   69.2| 22.1|
| 2| 44.5| 39.3|   45.1| 10.4|
| 3| 17.2| 45.9|   69.3|  9.3|
| 4|151.5| 41.3|   58.5| 18.5|
| 5|180.8| 10.8|   58.4| 12.9|
+---+-----+-----+-----+
only showing top 5 rows

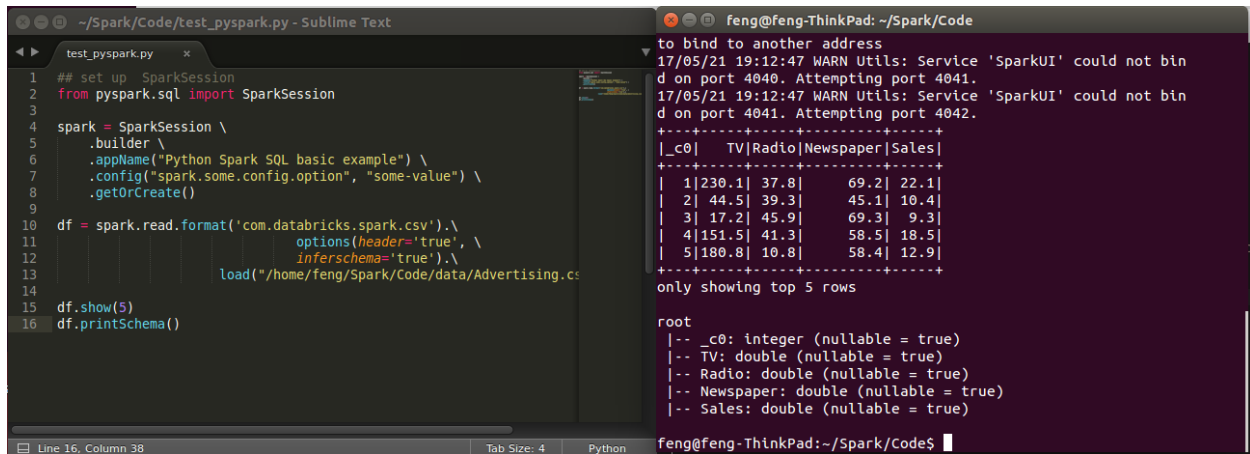
root
 |-- _c0: integer (nullable = true)
 |-- TV: double (nullable = true)
 |-- Radio: double (nullable = true)
 |-- Newspaper: double (nullable = true)
 |-- Sales: double (nullable = true)
```

3.4.2 PySpark With Sublime Text

After you finishing the above setup steps in *Configure Spark on Mac and Ubuntu*, then you should be good to use Sublime Text to write your PySpark Code and run your code as a normal python code in Terminal.

```
python test_pyspark.py
```

Then you should get the output results in your terminal.



```
test_pyspark.py
1 ## set up SparkSession
2 from pyspark.sql import SparkSession
3
4 spark = SparkSession \
5     .builder \
6     .appName("Python Spark SQL basic example") \
7     .config("spark.some.config.option", "some-value") \
8     .getOrCreate()
9
10 df = spark.read.format('com.databricks.spark.csv').\
11     options(header='true', \
12             inferSchema='true').\
13     load("/home/feng/Spark/Code/data/Advertising.csv")
14
15 df.show(5)
16 df.printSchema()
```

```
feng@feng-ThinkPad: ~/Spark/Code
to bind to another address
17/05/21 19:12:47 WARN Utils: Service 'SparkUI' could not bind
d on port 4040. Attempting port 4041.
17/05/21 19:12:47 WARN Utils: Service 'SparkUI' could not bind
d on port 4041. Attempting port 4042.
+-----+-----+-----+-----+
|_c0|  TV|Radio|Newspaper|Sales|
+-----+-----+-----+-----+
| 1|230.1| 37.8|    69.2| 22.1|
| 2| 44.5| 39.3|    45.1| 10.4|
| 3| 17.2| 45.9|    69.3|  9.3|
| 4|151.5| 41.3|    58.5| 18.5|
| 5|180.8| 10.8|    58.4| 12.9|
+-----+-----+-----+-----+
only showing top 5 rows

root
|-- _c0: integer (nullable = true)
|-- TV: double (nullable = true)
|-- Radio: double (nullable = true)
|-- Newspaper: double (nullable = true)
|-- Sales: double (nullable = true)
feng@feng-ThinkPad:~/Spark/Code$
```

3.4.3 PySpark With Eclipse

If you want to run PySpark code on Eclipse, you need to add the paths for the **External Libraries** for your **Current Project** as follows:

1. Open the properties of your project
2. Add the paths for the **External Libraries**

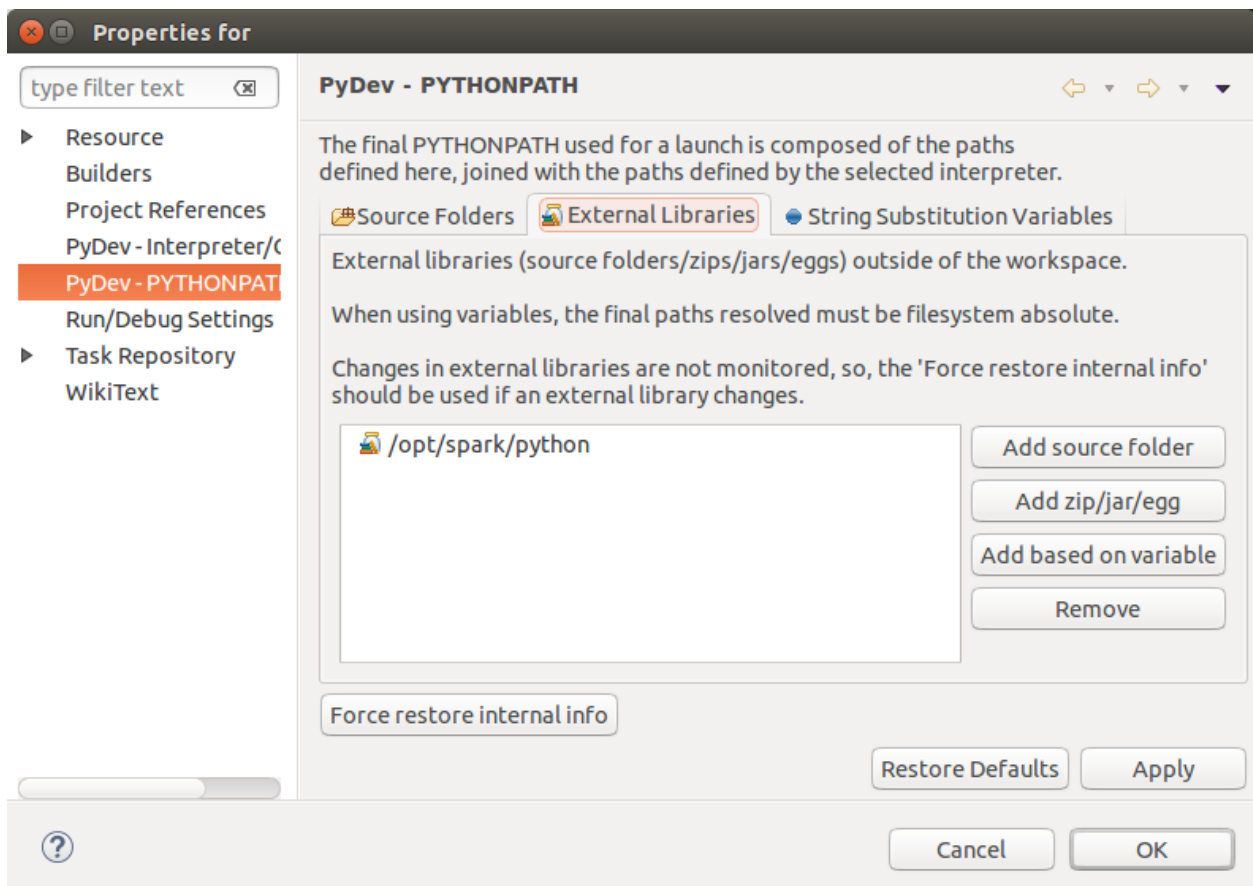
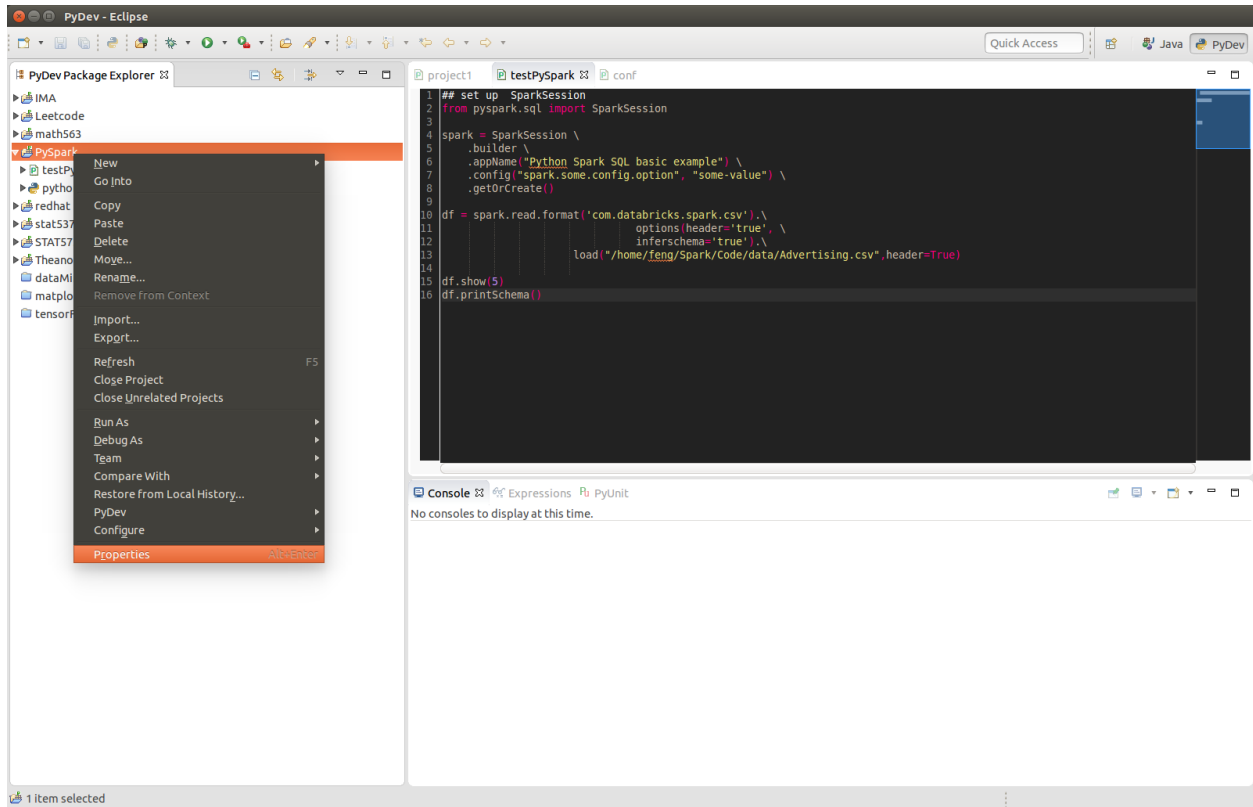
And then you should be good to run your code on Eclipse with PyDev.

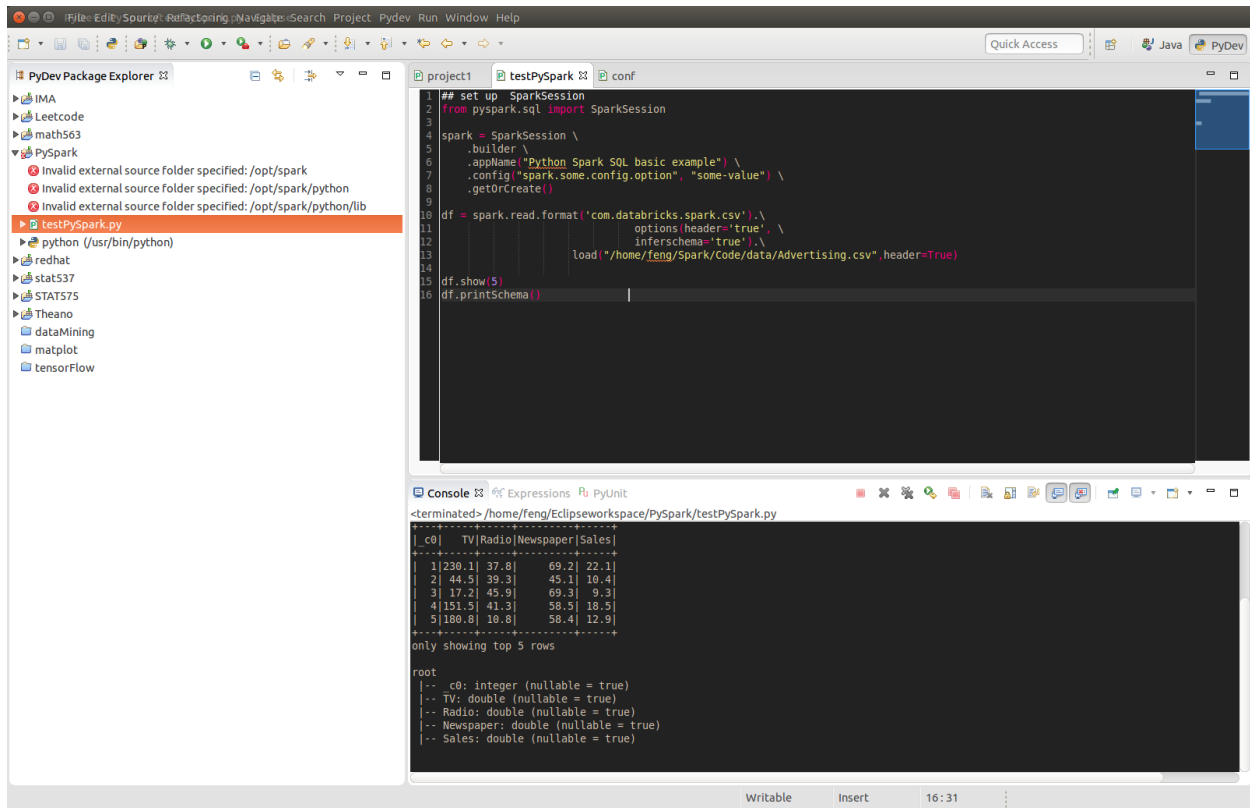
3.5 Set up Spark on Cloud

Folloing the setup steps in *Configure Spark on Mac and Ubuntu*, you can set up your own cluster on the cloud, for example AWS, Google Cloud. Actually, for those clouds, they have their own Big Data tool. You can run them directly whitout any setting just like Databricks Community Cloud. If you want more details, please feel free to contact with me.

3.6 Demo Code in this Section

The code for this section is available for download test_pyspark, and the Jupyter notebook can be download from test_pyspark_ipynb.





- Python Source code

```
## set up SparkSession
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark SQL basic example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

df = spark.read.format('com.databricks.spark.csv') \
    .options(header='true', \
              inferSchema='true') \
    .load("/home/feng/Spark/Code/data/Advertising.csv", header=True)

df.show(5)
df.printSchema()
```


AN INTRODUCTION TO APACHE SPARK

Note: **Know yourself and know your enemy, and you will never be defeated** – idiom, from Sunzi's Art of War

4.1 Core Concepts

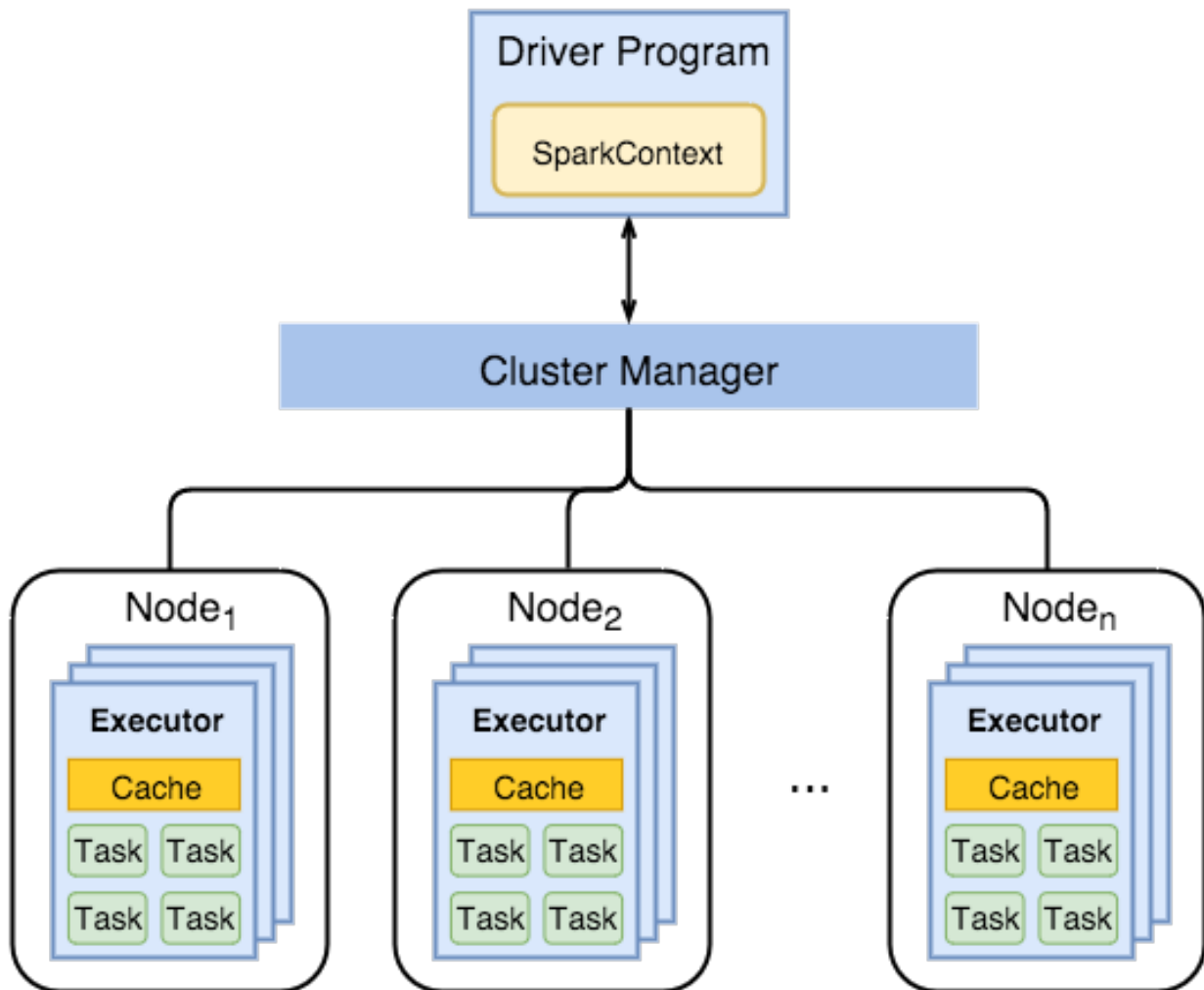
Most of the following content comes from [\[Kirillov2016\]](#). So the copyright belongs to **Anton Kirillov**. I will refer you to get more details from [Apache Spark core concepts, architecture and internals](#).

Before diving deep into how Apache Spark works, let's understand the jargon of Apache Spark

- **Job:** A piece of code which reads some input from HDFS or local, performs some computation on the data and writes some output data.
- **Stages:** Jobs are divided into stages. Stages are classified as a Map or reduce stages (It's easier to understand if you have worked on Hadoop and want to correlate). Stages are divided based on computational boundaries, all computations (operators) cannot be updated in a single stage. It happens over many stages.
- **Tasks:** Each stage has some tasks, one task per partition. One task is executed on one partition of data on one executor (machine).
- **DAG:** DAG stands for Directed Acyclic Graph, in the present context it's a DAG of operators.
- **Executor:** The process responsible for executing a task.
- **Master:** The machine on which the Driver program runs
- **Slave:** The machine on which the Executor program runs

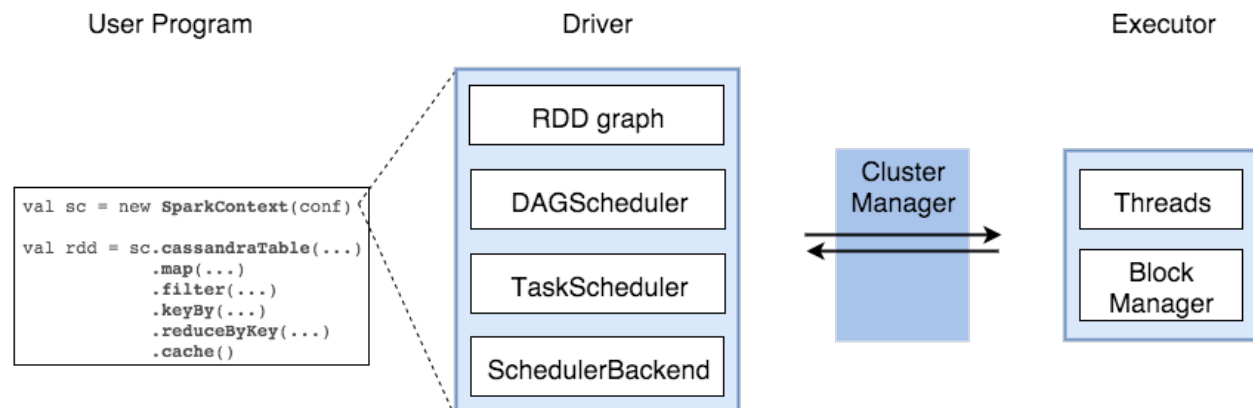
4.2 Spark Components

1. Spark Driver
 - separate process to execute user applications



- creates SparkContext to schedule jobs execution and negotiate with cluster manager
2. Executors
 - run tasks scheduled by driver
 - store computation results in memory, on disk or off-heap
 - interact with storage systems
 3. Cluster Manager
 - Mesos
 - YARN
 - Spark Standalone

Spark Driver contains more components responsible for translation of user code into actual jobs executed on cluster:



- SparkContext
 - represents the connection to a Spark cluster, and can be used to create RDDs, accumulators and broadcast variables on that cluster
- DAGScheduler
 - computes a DAG of stages for each job and submits them to TaskScheduler determines preferred locations for tasks (based on cache status or shuffle files locations) and finds minimum schedule to run the jobs
- TaskScheduler
 - responsible for sending tasks to the cluster, running them, retrying if there are failures, and mitigating stragglers
- SchedulerBackend
 - backend interface for scheduling systems that allows plugging in different implementations (Mesos, YARN, Standalone, local)

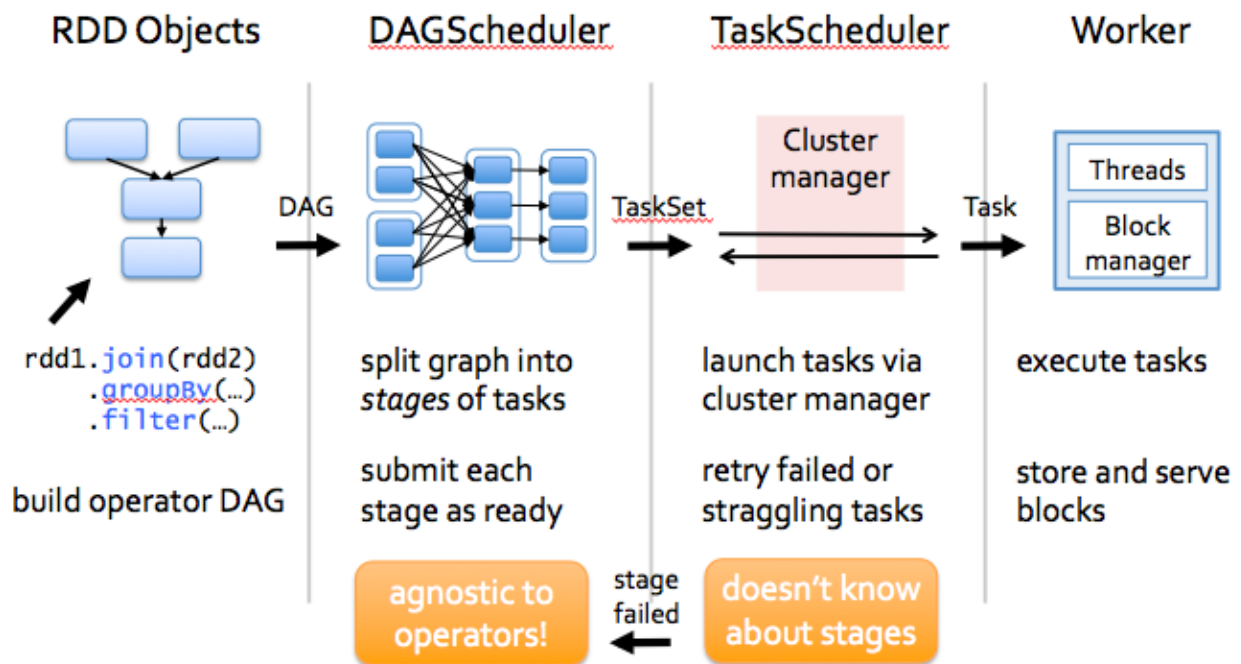
- BlockManager
 - provides interfaces for putting and retrieving blocks both locally and remotely into various stores (memory, disk, and off-heap)

4.3 Architecture

4.4 How Spark Works?

Spark has a small code base and the system is divided in various layers. Each layer has some responsibilities. The layers are independent of each other.

The first layer is the interpreter, Spark uses a Scala interpreter, with some modifications. As you enter your code in spark console (creating RDD's and applying operators), Spark creates an operator graph. When the user runs an action (like collect), the Graph is submitted to a DAG Scheduler. The DAG scheduler divides operator graph into (map and reduce) stages. A stage is comprised of tasks based on partitions of the input data. The DAG scheduler pipelines operators together to optimize the graph. For e.g. Many map operators can be scheduled in a single stage. This optimization is key to Spark's performance. The final result of a DAG scheduler is a set of stages. The stages are passed on to the Task Scheduler. The task scheduler launches tasks via cluster manager. (Spark Standalone/Yarn/Mesos). The task scheduler doesn't know about dependencies among stages.



PROGRAMMING WITH RDDS

Note: If you only know yourself, but not your opponent, you may win or may lose. If you know neither yourself nor your enemy, you will always endanger yourself – idiom, from Sunzi’s Art of War

RDD represents **Resilient Distributed Dataset**. An RDD in Spark is simply an immutable distributed collection of objects sets. Each RDD is split into multiple partitions (similar pattern with smaller sets), which may be computed on different nodes of the cluster.

5.1 Create RDD

Usually, there are two popular way to create the RDDs: loading an external dataset, or distributing a set of collection of objects. The following examples show some simplest ways to create RDDs by using `parallelize()` function which takes an already existing collection in your program and pass the same to the Spark Context.

1. By using `parallelize()` function

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark create RDD example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

df = spark.sparkContext.parallelize([(1, 2, 3, 'a b c'),
                                     (4, 5, 6, 'd e f'),
                                     (7, 8, 9, 'g h i')]).toDF(['col1', 'col2', 'col3', 'col4'])
```

Then you will get the RDD data:

```
df.show()

+---+---+---+---+
|col1|col2|col3| col4|
+---+---+---+---+
|  1|  2|  3|a b c|
|  4|  5|  6|d e f|
```

```
| 7| 8| 9|g h i|
+---+---+---+---+
```

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession \
    .builder \
    .appName("Python Spark create RDD example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

```
myData = spark.sparkContext.parallelize([(1,2), (3,4), (5,6), (7,8), (9,10)])
```

Then you will get the RDD data:

```
myData.collect()
```

```
[(1, 2), (3, 4), (5, 6), (7, 8), (9, 10)]
```

2. By using createDataFrame() function

```
from pyspark.sql import SparkSession
```

```
spark = SparkSession \
    .builder \
    .appName("Python Spark create RDD example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

```
Employee = spark.createDataFrame([
    ('1', 'Joe', '70000', '1'),
    ('2', 'Henry', '80000', '2'),
    ('3', 'Sam', '60000', '2'),
    ('4', 'Max', '90000', '1')],
    ['Id', 'Name', 'Sallary', 'DepartmentId']
)
```

Then you will get the RDD data:

```
+---+---+---+---+
| Id| Name|Sallary|DepartmentId|
+---+---+---+---+
| 1| Joe| 70000| 1|
| 2| Henry| 80000| 2|
| 3| Sam| 60000| 2|
| 4| Max| 90000| 1|
+---+---+---+---+
```

3. By using read and load functions

1. Read dataset from .csv file

```
## set up SparkSession
from pyspark.sql import SparkSession
```

```

spark = SparkSession \
    .builder \
    .appName("Python Spark create RDD example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

df = spark.read.format('com.databricks.spark.csv').\
    options(header='true', \
             inferSchema='true').\
    load("/home/feng/Spark/Code/data/Advertising.csv", header=True)

df.show(5)
df.printSchema()

```

Then you will get the RDD data:

```

+---+-----+-----+-----+-----+
|_c0|    TV|Radio|Newspaper|Sales|
+---+-----+-----+-----+
|  1|230.1| 37.8|    69.2| 22.1|
|  2| 44.5| 39.3|    45.1| 10.4|
|  3| 17.2| 45.9|    69.3|  9.3|
|  4|151.5| 41.3|    58.5| 18.5|
|  5|180.8| 10.8|    58.4| 12.9|
+---+-----+-----+-----+

```

only showing top 5 rows

```

root
 |-- _c0: integer (nullable = true)
 |-- TV: double (nullable = true)
 |-- Radio: double (nullable = true)
 |-- Newspaper: double (nullable = true)
 |-- Sales: double (nullable = true)

```

Once created, RDDs offer two types of operations: transformations and actions.

2. Read dataset from DataBase

```

## set up SparkSession
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark create RDD example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()

## User information
user = 'your_username'
pw   = 'your_password'

## Database information
table_name = 'table_name'
url = 'jdbc:postgresql://##.##.##.##:5432/dataset?user='+user+'&password='+pw

```

```
properties = {'driver': 'org.postgresql.Driver', 'password': pw, 'user': user}

df = spark.read.jdbc(url=url, table=table_name, properties=properties)

df.show(5)
df.printSchema()
```

Then you will get the RDD data:

```
+---+-----+-----+-----+-----+
|_c0|      TV|Radio|Newspaper|Sales|
+---+-----+-----+-----+-----+
|  1|230.1| 37.8|      69.2| 22.1|
|  2| 44.5| 39.3|      45.1| 10.4|
|  3| 17.2| 45.9|      69.3|  9.3|
|  4|151.5| 41.3|      58.5| 18.5|
|  5|180.8| 10.8|      58.4| 12.9|
+---+-----+-----+-----+-----+
```

only showing top 5 rows

```
root
 |-- _c0: integer (nullable = true)
 |-- TV: double (nullable = true)
 |-- Radio: double (nullable = true)
 |-- Newspaper: double (nullable = true)
 |-- Sales: double (nullable = true)
```

Note: Reading tables from Database needs the proper drive for the corresponding Database. For example, the above demo needs `org.postgresql.Driver` and **you need to download it and put it in “jars” folder of your spark installation path**. I download `postgresql-42.1.1.jar` from the official web-site and put it in `jars` folder.

5.2 Spark Transformations

Transformations construct a new RDD from a previous one. For example, one common transformation is filtering data that matches a predicate.

5.3 Spark Actions

Actions, on the other hand, compute a result based on an RDD, and either return it to the driver program or save it to an external storage system (e.g., HDFS).

STATISTICS PRELIMINARY

Note: If you only know yourself, but not your opponent, you may win or may lose. If you know neither yourself nor your enemy, you will always endanger yourself – idiom, from Sunzi’s Art of War

6.1 Notations

- m : the number of the samples
- n : the number of the features
- y_i : i -th label
- \bar{y} : the mean of y .

6.2 Measurement Formula

- Mean squared error

In statistics, the **MSE** (**Mean Squared Error**) of an estimator (of a procedure for estimating an unobserved quantity) measures the average of the squares of the errors or deviations—that is, the difference between the estimator and what is estimated.

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2$$

- Root Mean squared error

$$\text{RMSE} = \sqrt{\text{MSE}} = \sqrt{\frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2}$$

- Total sum of squares

In statistical data analysis the **TSS** (Total Sum of Squares) is a quantity that appears as part of a standard way of presenting results of such analyses. It is defined as being the sum, over all observations, of the squared differences of each observation from the overall mean.

$$\text{TSS} = \sum_{i=1}^m (y_i - \bar{y})^2$$

REGRESSION

Note: A journey of a thousand miles begins with a single step – old Chinese proverb

In statistical modeling, regression analysis focuses on investigating the relationship between a dependent variable and one or more independent variables. [Wikipedia Regression analysis](#)

In data mining, Regression is a model to represent the relationship between the value of label (or target, it is numerical variable) and on one or more features (or predictors they can be numerical and categorical variables).

7.1 Linear Regression

7.1.1 Introduction

Given that a data set $\{x_{i1}, \dots, x_{in}, y_i\}_{i=1}^m$ which contains n features (variables) and m samples (data points), in simple linear regression model for modeling m data points with one independent variable: x_{i1} , the formula is given by:

$$y_i = \beta_0 + \beta_1 x_{i1}, \text{ where, } i = 1, \dots, m.$$

In matrix notation, the data set is written as $\mathbf{X} = [\mathbf{X}_1, \dots, \mathbf{X}_n]$ with $\mathbf{X}_i = \{x_{i1}\}_{i=1}^m$, $\mathbf{y} = \{y_i\}_{i=1}^m$ and $\boldsymbol{\beta}^\top = \{\beta_i\}_{i=1}^m$. Then the normal equations are written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}.$$

7.1.2 How to solve it?

1. Direct Methods
2. Iterative Methods

7.1.3 Demo

- The Jupyter notebook can be download from Linear Regression which was implemented without using Pipeline.
- The Jupyter notebook can be download from Linear Regression with Pipeline which was implemented with using Pipeline.
- I will only present the code with pipeline style in the following.
- For more details about the parameters, please visit [Linear Regression API](#).

1. Set up spark context and SparkSession

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark regression example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

2. Load dataset

```
df = spark.read.format('com.databricks.spark.csv').\
    options(header='true', \
             inferSchema='true').\
    load("../data/Advertising.csv", header=True);
```

check the data set

```
df.show(5, True)
df.printSchema()
```

Then you will get

```
+-----+-----+-----+-----+
|   TV|Radio|Newspaper|Sales|
+-----+-----+-----+-----+
|230.1| 37.8|      69.2| 22.1|
| 44.5| 39.3|      45.1| 10.4|
| 17.2| 45.9|      69.3|  9.3|
|151.5| 41.3|      58.5| 18.5|
|180.8| 10.8|      58.4| 12.9|
+-----+-----+-----+-----+
only showing top 5 rows
```

```
root
 |-- TV: double (nullable = true)
 |-- Radio: double (nullable = true)
 |-- Newspaper: double (nullable = true)
 |-- Sales: double (nullable = true)
```

You can also get the Statistical results from the data frame (Unfortunately, it only works for numerical).

```
df.describe().show()
```

Then you will get

summary	TV	Radio	Newspaper	Sales
count	200	200	200	200
mean	147.0425	23.264000000000024	30.553999999999995	14.022500000000003
stddev	85.85423631490805	14.846809176168728	21.77862083852283	5.217456565710477
min	0.7	0.0	0.3	1.6
max	296.4	49.6	114.0	27.0

3. Convert the data to dense vector (features and label)

```
from pyspark.sql import Row
from pyspark.ml.linalg import Vectors

# I provide two ways to build the features and labels

# method 1 (good for small feature):
#def transData(row):
#    return Row(label=row["Sales"],
#               features=Vectors.dense([row["TV"],
#                                       row["Radio"],
#                                       row["Newspaper"]]))

# Method 2 (good for large features):
def transData(data):
    return data.rdd.map(lambda r: [Vectors.dense(r[:-1]), r[-1]]).toDF(['features', 'label'])
```

4. Transform the dataset to DataFrame

```
transformed= transData(df)
transformed.show(5)
```

```
+-----+-----+
|      features|label|
+-----+-----+
|[230.1, 37.8, 69.2]| 22.1|
|[44.5, 39.3, 45.1]| 10.4|
|[17.2, 45.9, 69.3]|  9.3|
|[151.5, 41.3, 58.5]| 18.5|
|[180.8, 10.8, 58.4]| 12.9|
+-----+-----+
only showing top 5 rows
```

Note: You will find out that all of the machine learning algorithms in Spark are based on the **features** and **label**. That is to say, you can play with all of the machine learning algorithms in Spark when you get ready the **features** and **label**.

5. Deal With Categorical Variables

7.1. Linear Regression

```
from pyspark.ml import Pipeline
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.evaluation import RegressionEvaluator

# Automatically identify categorical features, and index them.
# We specify maxCategories so features with > 4 distinct values are treated as continuous.

featureIndexer = VectorIndexer(inputCol="features", \
                                outputCol="indexedFeatures", \
                                maxCategories=4).fit(transformed)

data = featureIndexer.transform(transformed)
```

Now you check your dataset with

```
data.show(5, True)
```

you will get

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
|[230.1, 37.8, 69.2]| 22.1|[230.1, 37.8, 69.2]|
|[44.5, 39.3, 45.1]| 10.4|[44.5, 39.3, 45.1]|
|[17.2, 45.9, 69.3]|  9.3|[17.2, 45.9, 69.3]|
|[151.5, 41.3, 58.5]| 18.5|[151.5, 41.3, 58.5]|
|[180.8, 10.8, 58.4]| 12.9|[180.8, 10.8, 58.4]|
+-----+-----+-----+
only showing top 5 rows
```

6. Split the data into training and test sets (40% held out for testing)

```
# Split the data into training and test sets (40% held out for testing)
(trainingData, testData) = transformed.randomSplit([0.6, 0.4])
```

You can check your train and test data as follows (In my opinion, it is always to good to keep tracking your data during prototype pahse):

```
trainingData.show(5)
testData.show(5)
```

Then you will get

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
|[4.1, 11.6, 5.7]|  3.2|[4.1, 11.6, 5.7]|
|[5.4, 29.9, 9.4]|  5.3|[5.4, 29.9, 9.4]|
|[7.3, 28.1, 41.4]|  5.5|[7.3, 28.1, 41.4]|
|[7.8, 38.9, 50.6]|  6.6|[7.8, 38.9, 50.6]|
|[8.6, 2.1, 1.0]|  4.8|[8.6, 2.1, 1.0]|
+-----+-----+-----+
only showing top 5 rows
```

```

+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [0.7,39.6,8.7]| 1.6| [0.7,39.6,8.7]|
| [8.4,27.2,2.1]| 5.7| [8.4,27.2,2.1]|
| [11.7,36.9,45.2]| 7.3| [11.7,36.9,45.2]|
| [13.2,15.9,49.6]| 5.6| [13.2,15.9,49.6]|
| [16.9,43.7,89.4]| 8.7| [16.9,43.7,89.4]|
+-----+-----+-----+
only showing top 5 rows

```

7. Fit Ordinary Least Square Regression Model

For more details about the parameters, please visit [Linear Regression API](#).

```

# Import LinearRegression class
from pyspark.ml.regression import LinearRegression

# Define LinearRegression algorithm
lr = LinearRegression()

```

8. Pipeline Architecture

```

# Chain indexer and tree in a Pipeline
pipeline = Pipeline(stages=[featureIndexer, lr])

model = pipeline.fit(trainingData)

```

9. Summary of the Model

Spark has a poor summary function for data and model. I wrote a summary function which has similar format as **R** output for the linear regression in PySpark.

```

def modelsummary(model):
    import numpy as np
    print ("Note: the last rows are the information for Intercept")
    print ("##", "-----")
    print ("##", " Estimate      | Std.Error | t Values  | P-value")
    coef = np.append(list(model.coefficients),model.intercept)
    Summary=model.summary

    for i in range(len(Summary.pValues)):
        print ("##", '{:10.6f}'.format(coef[i]),\
              '{:10.6f}'.format(Summary.coefficientStandardErrors[i]),\
              '{:8.3f}'.format(Summary.tValues[i]),\
              '{:10.6f}'.format(Summary.pValues[i]))

    print ("##", '---')
    print ("##", "Mean squared error: % .6f" \
          % Summary.meanSquaredError, ", RMSE: % .6f" \
          % Summary.rootMeanSquaredError )
    print ("##", "Multiple R-squared: %f" % Summary.r2, ", \
          Total iterations: %i" % Summary.totalIterations)

```

```
modelsummary(model.stages[-1])
```

You will get the following summary results:

Note: the last rows are the information **for** Intercept

```
('##', '-----')
('##', ' Estimate | Std.Error | t Values | P-value')
('##', ' 0.044186', ' 0.001663', ' 26.573', ' 0.000000')
('##', ' 0.206311', ' 0.010846', ' 19.022', ' 0.000000')
('##', ' 0.001963', ' 0.007467', ' 0.263', ' 0.793113')
('##', ' 2.596154', ' 0.379550', ' 6.840', ' 0.000000')
('##', '----')
('##', 'Mean squared error: 2.588230', ' RMSE: 1.608798')
('##', 'Multiple R-squared: 0.911869', ' Total iterations: 1')
```

10. Make predictions

```
# Make predictions.
predictions = model.transform(testData)

# Select example rows to display.
predictions.select("features", "label", "predictedLabel").show(5)
```

```
+-----+-----+-----+
|      features|label|      prediction|
+-----+-----+-----+
| [0.7,39.6,8.7]| 1.6| 10.81405928637388|
| [8.4,27.2,2.1]| 5.7| 8.583086404079918|
| [11.7,36.9,45.2]| 7.3|10.814712818232422|
| [13.2,15.9,49.6]| 5.6| 6.557106943899219|
| [16.9,43.7,89.4]| 8.7|12.534151375058645|
+-----+-----+-----+
```

only showing top 5 rows

9. Evaluation

```
from pyspark.ml.evaluation import RegressionEvaluator
# Select (prediction, true label) and compute test error
evaluator = RegressionEvaluator(labelCol="label",
                                predictionCol="prediction",
                                metricName="rmse")

rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)
```

The final Root Mean Squared Error (RMSE) is as follows:

```
Root Mean Squared Error (RMSE) on test data = 1.63114
```

You can also check the R^2 value for the test data:

```
y_true = predictions.select("label").toPandas()
y_pred = predictions.select("prediction").toPandas()
```



```
import sklearn.metrics
r2_score = sklearn.metrics.r2_score(y_true, y_pred)
print('r2_score: {0}'.format(r2_score))
```

Then you will get

```
r2_score: 0.854486655585
```

Note: You should know most softwares are using different formula to calculate the R^2 value when no intercept is included in the model. You can get more information from the [discussion at StackExchange](#).

7.2 Generalized linear regression

7.2.1 Introduction

7.2.2 How to solve it?

7.2.3 Demo

- The Jupyter notebook can be download from [Generalized Linear Regression](#).
- For more details about the parameters, please visit [Generalized Linear Regression API](#).

1. Set up spark context and SparkSession

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark regression example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

2. Load dataset

```
df = spark.read.format('com.databricks.spark.csv').\
    options(header='true', \
             inferSchema='true').\
    load("../data/Advertising.csv", header=True);
```

check the data set

```
df.show(5, True)
df.printSchema()
```

Then you will get

```
+-----+-----+-----+-----+
|   TV | Radio | Newspaper | Sales |
+-----+-----+-----+-----+
```

```
|230.1| 37.8|      69.2| 22.1|
| 44.5| 39.3|      45.1| 10.4|
| 17.2| 45.9|      69.3|  9.3|
|151.5| 41.3|      58.5| 18.5|
|180.8| 10.8|      58.4| 12.9|
+-----+-----+-----+-----+
```

only showing top 5 rows

```
root
|-- TV: double (nullable = true)
|-- Radio: double (nullable = true)
|-- Newspaper: double (nullable = true)
|-- Sales: double (nullable = true)
```

You can also get the Statistical results from the data frame (Unfortunately, it only works for numerical).

```
df.describe().show()
```

Then you will get

```
+-----+-----+-----+-----+-----+
|summary|          TV|          Radio|          Newspaper|          Sales|
+-----+-----+-----+-----+-----+
|  count|          200|          200|          200|          200|
|   mean|    147.0425|23.2640000000000024|30.553999999999995|14.022500000000003|
| stddev|85.85423631490805|14.846809176168728| 21.77862083852283| 5.217456565710477|
|   min|           0.7|           0.0|           0.3|           1.6|
|   max|          296.4|          49.6|          114.0|          27.0|
+-----+-----+-----+-----+-----+
```

3. Convert the data to dense vector (features and label)

```
from pyspark.sql import Row
from pyspark.ml.linalg import Vectors

# I provide two ways to build the features and labels

# method 1 (good for small feature):
#def transData(row):
#    return Row(label=row["Sales"],
#               features=Vectors.dense([row["TV"],
#                                       row["Radio"],
#                                       row["Newspaper"]]))

# Method 2 (good for large features):
def transData(data):
    return data.rdd.map(lambda r: [Vectors.dense(r[:-1]),r[-1]].toDF(['features','label']))

transformed= transData(df)
transformed.show(5)

+-----+-----+
|      features|label|
+-----+-----+
```

```
| [230.1, 37.8, 69.2] | 22.1 |
| [44.5, 39.3, 45.1] | 10.4 |
| [17.2, 45.9, 69.3] | 9.3 |
| [151.5, 41.3, 58.5] | 18.5 |
| [180.8, 10.8, 58.4] | 12.9 |
```

```
+-----+-----+
```

only showing top 5 rows

Note: You will find out that all of the machine learning algorithms in Spark are based on the **features** and **label**. That is to say, you can play with all of the machine learning algorithms in Spark when you get ready the **features** and **label**.

4. Convert the data to dense vector

```
# convert the data to dense vector
def transData(data):
    return data.rdd.map(lambda r: [r[-1], Vectors.dense(r[:-1])]).\
        toDF(['label', 'features'])
```

```
from pyspark.sql import Row
from pyspark.ml.linalg import Vectors
```

```
data= transData(df)
data.show()
```

5. Deal with the Categorical variables

```
from pyspark.ml import Pipeline
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.evaluation import RegressionEvaluator
```

```
# Automatically identify categorical features, and index them.
# We specify maxCategories so features with > 4
# distinct values are treated as continuous.
```

```
featureIndexer = VectorIndexer(inputCol="features", \
                                outputCol="indexedFeatures", \
                                maxCategories=4).fit(transformed)
```

```
data = featureIndexer.transform(transformed)
```

When you check you data at this point, you will get

```
+-----+-----+-----+
|          features|label| indexedFeatures|
+-----+-----+-----+
| [230.1, 37.8, 69.2] | 22.1 | [230.1, 37.8, 69.2] |
| [44.5, 39.3, 45.1] | 10.4 | [44.5, 39.3, 45.1] |
| [17.2, 45.9, 69.3] | 9.3 | [17.2, 45.9, 69.3] |
| [151.5, 41.3, 58.5] | 18.5 | [151.5, 41.3, 58.5] |
| [180.8, 10.8, 58.4] | 12.9 | [180.8, 10.8, 58.4] |
+-----+-----+-----+
```

only showing top 5 rows

6. Split the data into training and test sets (40% held out for testing)

```
# Split the data into training and test sets (40% held out for testing)
(trainingData, testData) = transformed.randomSplit([0.6, 0.4])
```

You can check your train and test data as follows (In my opinion, it is always to good to keep tracking your data during prototype pahse):

```
trainingData.show(5)
testData.show(5)
```

Then you will get

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [5.4,29.9,9.4]| 5.3| [5.4,29.9,9.4]|
| [7.8,38.9,50.6]| 6.6| [7.8,38.9,50.6]|
| [8.4,27.2,2.1]| 5.7| [8.4,27.2,2.1]|
| [8.7,48.9,75.0]| 7.2| [8.7,48.9,75.0]|
| [11.7,36.9,45.2]| 7.3| [11.7,36.9,45.2]|
+-----+-----+-----+
only showing top 5 rows
```

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [0.7,39.6,8.7]| 1.6| [0.7,39.6,8.7]|
| [4.1,11.6,5.7]| 3.2| [4.1,11.6,5.7]|
| [7.3,28.1,41.4]| 5.5| [7.3,28.1,41.4]|
| [8.6,2.1,1.0]| 4.8| [8.6,2.1,1.0]|
| [17.2,4.1,31.6]| 5.9| [17.2,4.1,31.6]|
+-----+-----+-----+
only showing top 5 rows
```

7. Fit Generalized Linear Regression Model

```
# Import LinearRegression class
from pyspark.ml.regression import GeneralizedLinearRegression

# Define LinearRegression algorithm
glr = GeneralizedLinearRegression(family="gaussian", link="identity",\
                                  maxIter=10, regParam=0.3)
```

8. Pipeline Architecture

```
# Chain indexer and tree in a Pipeline
pipeline = Pipeline(stages=[featureIndexer, glr])

model = pipeline.fit(trainingData)
```

9. Summary of the Model

Spark has a poor summary function for data and model. I wrote a summary function which has similar format as **R** output for the linear regression in PySpark.

```
def modelsummary(model):
    import numpy as np
    print ("Note: the last rows are the information for Intercept")
    print ("##", "-----")
    print ("##", " Estimate | Std.Error | t Values | P-value")
    coef = np.append(list(model.coefficients), model.intercept)
    Summary=model.summary

    for i in range(len(Summary.pValues)):
        print ("##", '{:10.6f}'.format(coef[i]), \
              '{:10.6f}'.format(Summary.coefficientStandardErrors[i]), \
              '{:8.3f}'.format(Summary.tValues[i]), \
              '{:10.6f}'.format(Summary.pValues[i]))

    print ("##", '---')
    # print ("##", "Mean squared error: % .6f" \
    #       % Summary.meanSquaredError, ", RMSE: % .6f" \
    #       % Summary.rootMeanSquaredError )
    # print ("##", "Multiple R-squared: %f" % Summary.r2, ", \
    #       Total iterations: %i" % Summary.totalIterations)

modelsummary(model.stages[-1])
```

You will get the following summary results:

Note: the last rows are the information **for** Intercept

```
('##', '-----')
('##', ' Estimate | Std.Error | t Values | P-value')
('##', ' 0.042857', ' 0.001668', ' 25.692', ' 0.000000')
('##', ' 0.199922', ' 0.009881', ' 20.232', ' 0.000000')
('##', ' -0.001957', ' 0.006917', ' -0.283', ' 0.777757')
('##', ' 3.007515', ' 0.406389', ' 7.401', ' 0.000000')
('##', '---')
```

10. Make predictions

```
# Make predictions.
predictions = model.transform(testData)

# Select example rows to display.
predictions.select("features", "label", "predictedLabel").show(5)
```

```
+-----+-----+-----+
| features|label| prediction|
+-----+-----+-----+
| [0.7,39.6,8.7]| 1.6|10.937383732327625|
| [4.1,11.6,5.7]| 3.2| 5.491166258750164|
| [7.3,28.1,41.4]| 5.5| 8.8571603947873|
| [8.6,2.1,1.0]| 4.8| 3.793966281660073|
| [17.2,4.1,31.6]| 5.9| 4.502507124763654|
+-----+-----+-----+
```

only showing top 5 rows

11. Evaluation

```
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.evaluation import RegressionEvaluator
# Select (prediction, true label) and compute test error
evaluator = RegressionEvaluator(labelCol="label",
                                predictionCol="prediction",
                                metricName="rmse")

rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)
```

The final Root Mean Squared Error (RMSE) is as follows:

Root Mean Squared Error (RMSE) on test data = 1.89857

```
y_true = predictions.select("label").toPandas()
y_pred = predictions.select("prediction").toPandas()

import sklearn.metrics
r2_score = sklearn.metrics.r2_score(y_true, y_pred)
print('r2_score: {0}'.format(r2_score))
```

Then you will get the R^2 value:

r2_score: 0.87707391843

7.3 Decision tree Regression

7.3.1 Introduction

7.3.2 How to solve it?

7.3.3 Demo

- The Jupyter notebook can be download from [Decision Tree Regression](#).
- For more details about the parameters, please visit [Decision Tree Regressor API](#).

1. Set up spark context and SparkSession

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark regression example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

2. Load dataset

```
df = spark.read.format('com.databricks.spark.csv').\
    options(header='true', \
             inferSchema='true').\
    load("../data/Advertising.csv", header=True);
```

check the data set

```
df.show(5, True)
df.printSchema()
```

Then you will get

```
+-----+-----+-----+-----+
|   TV|Radio|Newspaper|Sales|
+-----+-----+-----+-----+
|230.1| 37.8|      69.2| 22.1|
| 44.5| 39.3|      45.1| 10.4|
| 17.2| 45.9|      69.3|  9.3|
|151.5| 41.3|      58.5| 18.5|
|180.8| 10.8|      58.4| 12.9|
+-----+-----+-----+-----+
```

only showing top 5 rows

```
root
 |-- TV: double (nullable = true)
 |-- Radio: double (nullable = true)
 |-- Newspaper: double (nullable = true)
 |-- Sales: double (nullable = true)
```

You can also get the Statistical results from the data frame (Unfortunately, it only works for numerical).

```
df.describe().show()
```

Then you will get

```
+-----+-----+-----+-----+
|summary|          TV|          Radio|          Newspaper|          Sales|
+-----+-----+-----+-----+
| count|          200|          200|          200|          200|
| mean|    147.0425|23.264000000000024|30.553999999999995|14.022500000000003|
| stddev|85.85423631490805|14.846809176168728|21.77862083852283|5.217456565710477|
|  min|           0.7|           0.0|           0.3|           1.6|
|  max|          296.4|          49.6|          114.0|          27.0|
+-----+-----+-----+-----+
```

3. Convert the data to dense vector (features and label)

```
from pyspark.sql import Row
from pyspark.ml.linalg import Vectors

# I provide two ways to build the features and labels

# method 1 (good for small feature):
```

```
#def transData(row):
#     return Row(label=row["Sales"],
#                 features=Vectors.dense([row["TV"],
#                                         row["Radio"],
#                                         row["Newspaper"]]))

# Method 2 (good for large features):
def transData(data):
    return data.rdd.map(lambda r: [Vectors.dense(r[:-1]),r[-1]]).toDF(['features','label'])

transformed= transData(df)
transformed.show(5)

+-----+-----+
|          features|label|
+-----+-----+
| [230.1, 37.8, 69.2]| 22.1|
| [44.5, 39.3, 45.1]| 10.4|
| [17.2, 45.9, 69.3]|  9.3|
| [151.5, 41.3, 58.5]| 18.5|
| [180.8, 10.8, 58.4]| 12.9|
+-----+-----+
only showing top 5 rows
```

Note: You will find out that all of the machine learning algorithms in Spark are based on the **features** and **label**. That is to say, you can play with all of the machine learning algorithms in Spark when you get ready the **features** and **label**.

4. Convert the data to dense vector

```
# convert the data to dense vector
def transData(data):
    return data.rdd.map(lambda r: [r[-1], Vectors.dense(r[:-1])]).\
        toDF(['label','features'])

transformed = transData(df)
transformed.show(5)
```

5. Deal with the Categorical variables

```
from pyspark.ml import Pipeline
from pyspark.ml.regression import LinearRegression
from pyspark.ml.feature import VectorIndexer
from pyspark.ml.evaluation import RegressionEvaluator

# Automatically identify categorical features, and index them.
# We specify maxCategories so features with > 4
# distinct values are treated as continuous.

featureIndexer = VectorIndexer(inputCol="features", \
                               outputCol="indexedFeatures",\
                               maxCategories=4).fit(transformed)
```



```
data = featureIndexer.transform(transformed)
```

When you check you data at this point, you will get

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [230.1, 37.8, 69.2] | 22.1 | [230.1, 37.8, 69.2] |
| [44.5, 39.3, 45.1] | 10.4 | [44.5, 39.3, 45.1] |
| [17.2, 45.9, 69.3] | 9.3 | [17.2, 45.9, 69.3] |
| [151.5, 41.3, 58.5] | 18.5 | [151.5, 41.3, 58.5] |
| [180.8, 10.8, 58.4] | 12.9 | [180.8, 10.8, 58.4] |
+-----+-----+-----+
only showing top 5 rows
```

6. Split the data into training and test sets (40% held out for testing)

```
# Split the data into training and test sets (40% held out for testing)
(trainingData, testData) = transformed.randomSplit([0.6, 0.4])
```

You can check your train and test data as follows (In my opinion, it is always to good to keep tracking your data during prototype pahse):

```
trainingData.show(5)
testData.show(5)
```

Then you will get

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [4.1, 11.6, 5.7] | 3.2 | [4.1, 11.6, 5.7] |
| [7.3, 28.1, 41.4] | 5.5 | [7.3, 28.1, 41.4] |
| [8.4, 27.2, 2.1] | 5.7 | [8.4, 27.2, 2.1] |
| [8.6, 2.1, 1.0] | 4.8 | [8.6, 2.1, 1.0] |
| [8.7, 48.9, 75.0] | 7.2 | [8.7, 48.9, 75.0] |
+-----+-----+-----+
only showing top 5 rows
```

```
+-----+-----+-----+
|      features|label| indexedFeatures|
+-----+-----+-----+
| [0.7, 39.6, 8.7] | 1.6 | [0.7, 39.6, 8.7] |
| [5.4, 29.9, 9.4] | 5.3 | [5.4, 29.9, 9.4] |
| [7.8, 38.9, 50.6] | 6.6 | [7.8, 38.9, 50.6] |
| [17.2, 45.9, 69.3] | 9.3 | [17.2, 45.9, 69.3] |
| [18.7, 12.1, 23.4] | 6.7 | [18.7, 12.1, 23.4] |
+-----+-----+-----+
only showing top 5 rows
```

7. Fit Decision Tree Regression Model

```
from pyspark.ml.regression import DecisionTreeRegressor
```

```
# Train a DecisionTree model.
dt = DecisionTreeRegressor(featuresCol="indexedFeatures")
```

8. Pipeline Architecture

```
# Chain indexer and tree in a Pipeline
pipeline = Pipeline(stages=[featureIndexer, dt])

model = pipeline.fit(trainingData)
```

9. Make predictions

```
# Make predictions.
predictions = model.transform(testData)

# Select example rows to display.
predictions.select("features", "label", "predictedLabel").show(5)
```

```
+-----+-----+-----+
|prediction|label|      features|
+-----+-----+-----+
|      7.2|  1.6| [0.7,39.6,8.7]|
|      7.3|  5.3| [5.4,29.9,9.4]|
|      7.2|  6.6| [7.8,38.9,50.6]|
|     8.64|  9.3| [17.2,45.9,69.3]|
|     6.45|  6.7| [18.7,12.1,23.4]|
+-----+-----+-----+
only showing top 5 rows
```

10. Evaluation

```
from pyspark.ml.evaluation import RegressionEvaluator
from pyspark.ml.evaluation import RegressionEvaluator
# Select (prediction, true label) and compute test error
evaluator = RegressionEvaluator(labelCol="label",
                                predictionCol="prediction",
                                metricName="rmse")

rmse = evaluator.evaluate(predictions)
print("Root Mean Squared Error (RMSE) on test data = %g" % rmse)
```

The final Root Mean Squared Error (RMSE) is as follows:

```
Root Mean Squared Error (RMSE) on test data = 1.50999
```

```
y_true = predictions.select("label").toPandas()
y_pred = predictions.select("prediction").toPandas()
```

```
import sklearn.metrics
r2_score = sklearn.metrics.r2_score(y_true, y_pred)
print('r2_score: {0}'.format(r2_score))
```

Then you will get the R^2 value:

```
r2_score: 0.911024318967
```

You may also check the importance of the features:

```
model.stages[1].featureImportances
```

The you will get the weight for each features

```
SparseVector(3, {0: 0.6811, 1: 0.3187, 2: 0.0002})
```

7.4 Random Forest Regression

7.4.1 Introduction

7.4.2 How to solve it?

7.4.3 Demo

- The Jupyter notebook can be download from [Random Forest Regression](#).
- For more details about the parameters, please visit [Random Forest Regressor API](#).

7.5 Gradient-boosted tree regression

7.5.1 Introduction

7.5.2 How to solve it?

7.5.3 Demo

- The Jupyter notebook can be download from [Gradient-boosted tree regression](#).
- For more details about the parameters, please visit [Gradient boosted tree API](#).

CLASSIFICATION

Note: *Birds of a feather flock together.* – old Chinese proverb

8.1 Logistic regression

8.1.1 Binomial logistic regression

8.1.2 Multinomial logistic regression

8.2 Decision tree Classification

- The Jupyter notebook can be download from Decision Tree Classification.
- For more details, please visit [DecisionTreeClassifier API](#) .

8.3 Random forest Classification

- The Jupyter notebook can be download from Random forest Classification.
- For more details, please visit [RandomForestClassifier API](#) .

8.4 Gradient-boosted tree Classification

- The Jupyter notebook can be download from Gradient boosted tree Classification.
- For more details, please visit [GBClassifier API](#) .

8.5 Naive Bayes Classification

- The Jupyter notebook can be download from Naive Bayes Classification.

- For more details, please visit [NaiveBayes API](#).

8.6 Support Vector Machines Classification

CLUSTERING

Note: Sharpening the knife longer can make it easier to hack the firewood – old Chinese proverb

9.1 K-Means Model

TEXT MINING

Note: Sharpening the knife longer can make it easier to hack the firewood – old Chinese proverb

I want to answer this question in two folders:

10.1 Source of text

10.1.1 Image to text

```
def img2txt(img_dir):
    """
    convert images to text
    """
    import os, PythonMagick
    from datetime import datetime
    import PyPDF2

    from PIL import Image
    import pytesseract

    f = open('doc4img.txt', 'wa')
    for img in [img_file for img_file in os.listdir(img_dir)
                 if (img_file.endswith(".png") or
                     img_file.endswith(".jpg") or
                     img_file.endswith(".jpeg"))]:

        start_time = datetime.now()

        input_img = img_dir + "/" + img

        print('-----')
        print(img)
        print('Converting ' + img + '.....')
        print('-----')

        # extract the text information from images
        text = pytesseract.image_to_string(Image.open(input_img))
        print(text)
```

```
# ouput text file
f.write( img + "\n")
f.write(text.encode('utf-8'))

print "CPU Time for converting" + img + ":" + str(datetime.now() - start_time) + "\n"
f.write( "\n-----\n")

f.close()
```

10.1.2 Image Enhnaced to text

```
def pdf2txt_enhance(img_dir, scaler):
    """
    convert images files to text
    """

    import numpy as np
    import os, PythonMagick
    from datetime import datetime
    import PyPDF2

    from PIL import Image, ImageEnhance, ImageFilter
    import pytesseract

    f = open('doc4img.txt', 'wa')
    for img in [img_file for img_file in os.listdir(img_dir)
                 if (img_file.endswith(".png") or
                    img_file.endswith(".jpg") or
                    img_file.endswith(".jpeg"))]:

        start_time = datetime.now()

        input_img = img_dir + "/" + img
        enhanced_img = img_dir + "/" + "Enhanced" + "/" + img

        im = Image.open(input_img) # the second one
        im = im.filter(ImageFilter.MedianFilter())
        enhancer = ImageEnhance.Contrast(im)
        im = enhancer.enhance(1)
        im = im.convert('1')
        im.save(enhanced_img)

        for scale in np.ones(scaler):
            im = Image.open(enhanced_img) # the second one
            im = im.filter(ImageFilter.MedianFilter())
            enhancer = ImageEnhance.Contrast(im)
            im = enhancer.enhance(scale)
            im = im.convert('1')
            im.save(enhanced_img)
```

```

print('-----')
print(img)
print('Converting ' + img + '.....')
print('-----')

# extract the text information from images
text = pytesseract.image_to_string(Image.open(enhanced_img))
print(text)

# ouput text file
f.write( img + "\n")
f.write(text.encode('utf-8'))

print "CPU Time for converting" + img + ":" + str(datetime.now() - start_time) + "\n"
f.write( "\n-----\n")

f.close()

```

10.1.3 PDF to text

```

def pdf2txt(pdf_dir, image_dir):
    """
    convert PDF to text
    """

    import os, PythonMagick
    from datetime import datetime
    import PyPDF2

    from PIL import Image
    import pytesseract

    f = open('doc.txt', 'wa')
    for pdf in [pdf_file for pdf_file in os.listdir(pdf_dir) if pdf_file.endswith(".pdf")]:

        start_time = datetime.now()

        input_pdf = pdf_dir + "/" + pdf

        pdf_im = PyPDF2.PdfFileReader(file(input_pdf, "rb"))
        npage = pdf_im.getNumPages()

        print('-----')
        print(pdf)
        print('Converting %d pages.' % npage)
        print('-----')

        f.write( "\n-----\n")

```

```
for p in range(npage):

    pdf_file = input_pdf + '[' + str(p) + ']'
    image_file = image_dir + "/" + pdf + '_' + str(p) + '.png'

    # convert PDF files to Images
    im = PythonMagick.Image()
    im.density('300')
    im.read(pdf_file)
    im.write(image_file)

    # extract the text information from images
    text = pytesseract.image_to_string(Image.open(image_file))

    #print(text)

    # ouput text file
    f.write(pdf + "\n")
    f.write(text.encode('utf-8'))

print "CPU Time for converting" + pdf + ":" + str(datetime.now() - start_time) + "\n"

f.close()
```

10.2 Text Preprocessing

- check to see if a row only contains whitespace

```
def check_blanks(data_str):
    is_blank = str(data_str.isspace())
    return is_blank
```

- Determine whether the language of the text content is english or not: Use langid module to classify the language to make sure we are applying the correct cleanup actions for English langid

```
def check_lang(data_str):
    predict_lang = langid.classify(data_str)
    if predict_lang[1] >= .9:
        language = predict_lang[0]
    else:
        language = 'NA'
    return language
```

- Remove features

```
def remove_features(data_str):
    # compile regex
    url_re = re.compile('https?:/(www.)?\w+\.\w+(/w+)*/?')
    punc_re = re.compile('[%s]' % re.escape(string.punctuation))
    num_re = re.compile('(\\d+)')
```

```

mention_re = re.compile('@(\w+)')
alpha_num_re = re.compile("[a-z0-9_]+$")
# convert to lowercase
data_str = data_str.lower()
# remove hyperlinks
data_str = url_re.sub(' ', data_str)
# remove @mentions
data_str = mention_re.sub(' ', data_str)
# remove punctuation
data_str = punc_re.sub(' ', data_str)
# remove numeric 'words'
data_str = num_re.sub(' ', data_str)
# remove non a-z 0-9 characters and words shorter than 3 characters
list_pos = 0
cleaned_str = ''
for word in data_str.split():
    if list_pos == 0:
        if alpha_num_re.match(word) and len(word) > 2:
            cleaned_str = word
        else:
            cleaned_str = ' '
    else:
        if alpha_num_re.match(word) and len(word) > 2:
            cleaned_str = cleaned_str + ' ' + word
        else:
            cleaned_str += ' '
    list_pos += 1
return cleaned_str

```

- removes stop words

```

def remove_stops(data_str):
    # expects a string
    stops = set(stopwords.words("english"))
    list_pos = 0
    cleaned_str = ''
    text = data_str.split()
    for word in text:
        if word not in stops:
            # rebuild cleaned_str
            if list_pos == 0:
                cleaned_str = word
            else:
                cleaned_str = cleaned_str + ' ' + word
            list_pos += 1
    return cleaned_str

```

- tagging text

```

def tag_and_remove(data_str):
    cleaned_str = ''
    # noun tags
    nn_tags = ['NN', 'NNP', 'NNP', 'NNPS', 'NNS']
    # adjectives

```

```
jj_tags = ['JJ', 'JJR', 'JJS']
# verbs
vb_tags = ['VB', 'VBD', 'VBG', 'VBN', 'VBP', 'VBZ']
nltk_tags = nn_tags + jj_tags + vb_tags

# break string into 'words'
text = data_str.split()

# tag the text and keep only those with the right tags
tagged_text = pos_tag(text)
for tagged_word in tagged_text:
    if tagged_word[1] in nltk_tags:
        cleaned_str += tagged_word[0] + ' '

return cleaned_str
```

- lemmatization

```
def lemmatize(data_str):
    # expects a string
    list_pos = 0
    cleaned_str = ''
    lmtzr = WordNetLemmatizer()
    text = data_str.split()
    tagged_words = pos_tag(text)
    for word in tagged_words:
        if 'v' in word[1].lower():
            lemma = lmtzr.lemmatize(word[0], pos='v')
        else:
            lemma = lmtzr.lemmatize(word[0], pos='n')
        if list_pos == 0:
            cleaned_str = lemma
        else:
            cleaned_str = cleaned_str + ' ' + lemma
        list_pos += 1
    return cleaned_str
```

define the preprocessing function in PySpark

```
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType
import preproc as pp

check_lang_udf = udf(pp.check_lang, StringType())
remove_stops_udf = udf(pp.remove_stops, StringType())
remove_features_udf = udf(pp.remove_features, StringType())
tag_and_remove_udf = udf(pp.tag_and_remove, StringType())
lemmatize_udf = udf(pp.lemmatize, StringType())
check_blanks_udf = udf(pp.check_blanks, StringType())
```

10.3 Text Classification

```
from nltk.stem.wordnet import WordNetLemmatizer
from nltk.corpus import stopwords
from nltk import pos_tag
import string
import re
import langid
```

10.4 Sentiment analysis

10.5 N-grams and Correlations

10.6 Topic Model: Latent Dirichlet Allocation

SOCIAL NETWORK ANALYSIS

Note: Sharpening the knife longer can make it easier to hack the firewood – old Chinese proverb

11.1 Co-occurrence Network

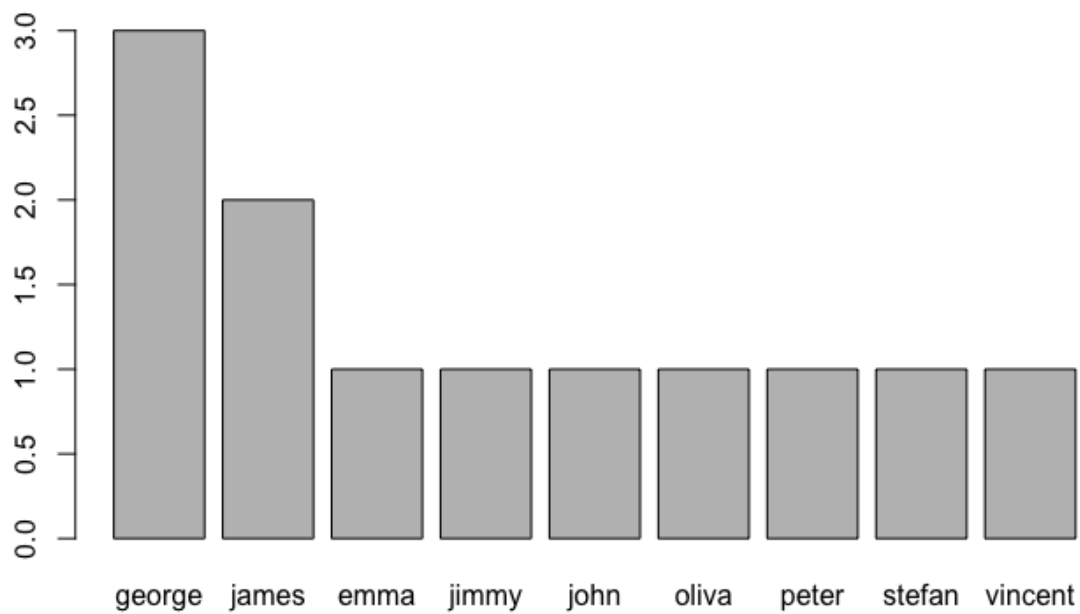


Figure 11.1: Name frequency

Then you will get Figure *Co-occurrence network*

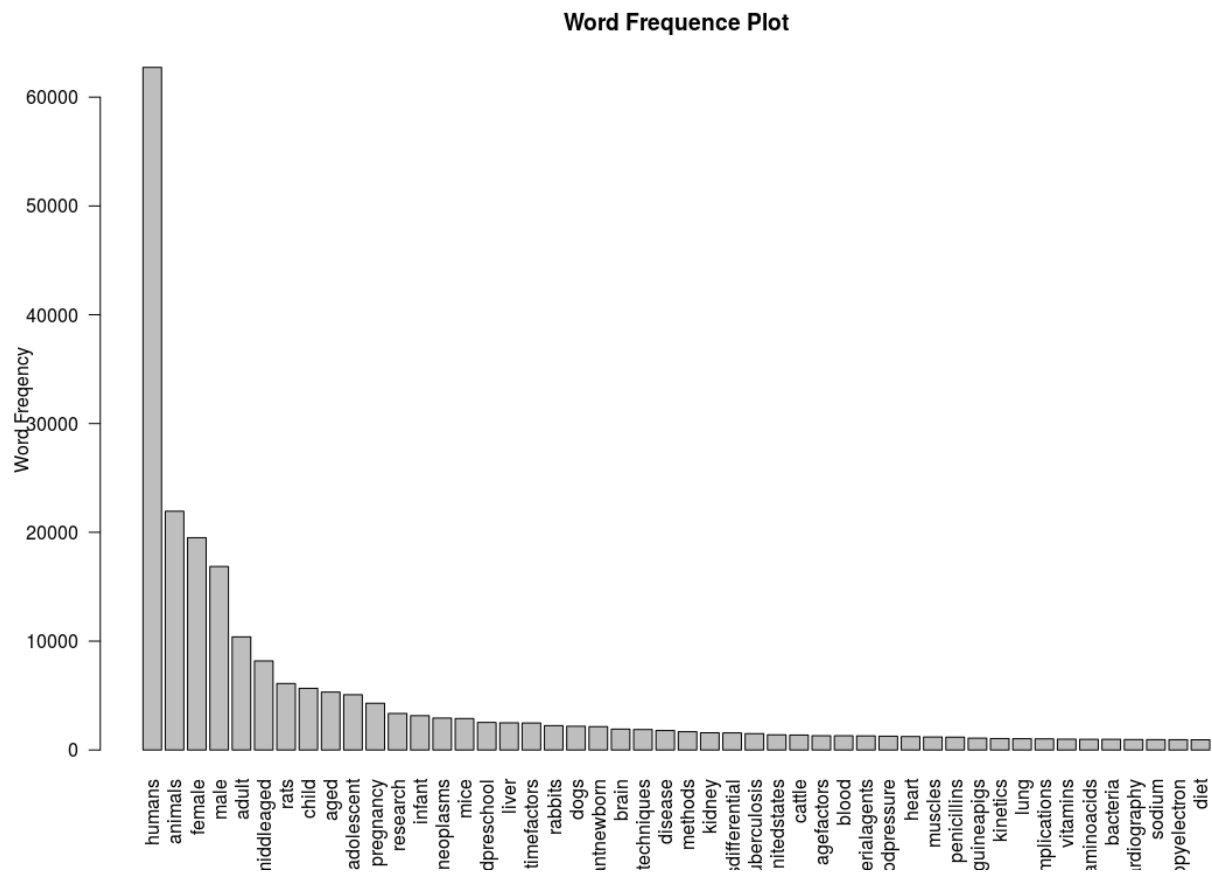


Figure 11.2: Word frequency

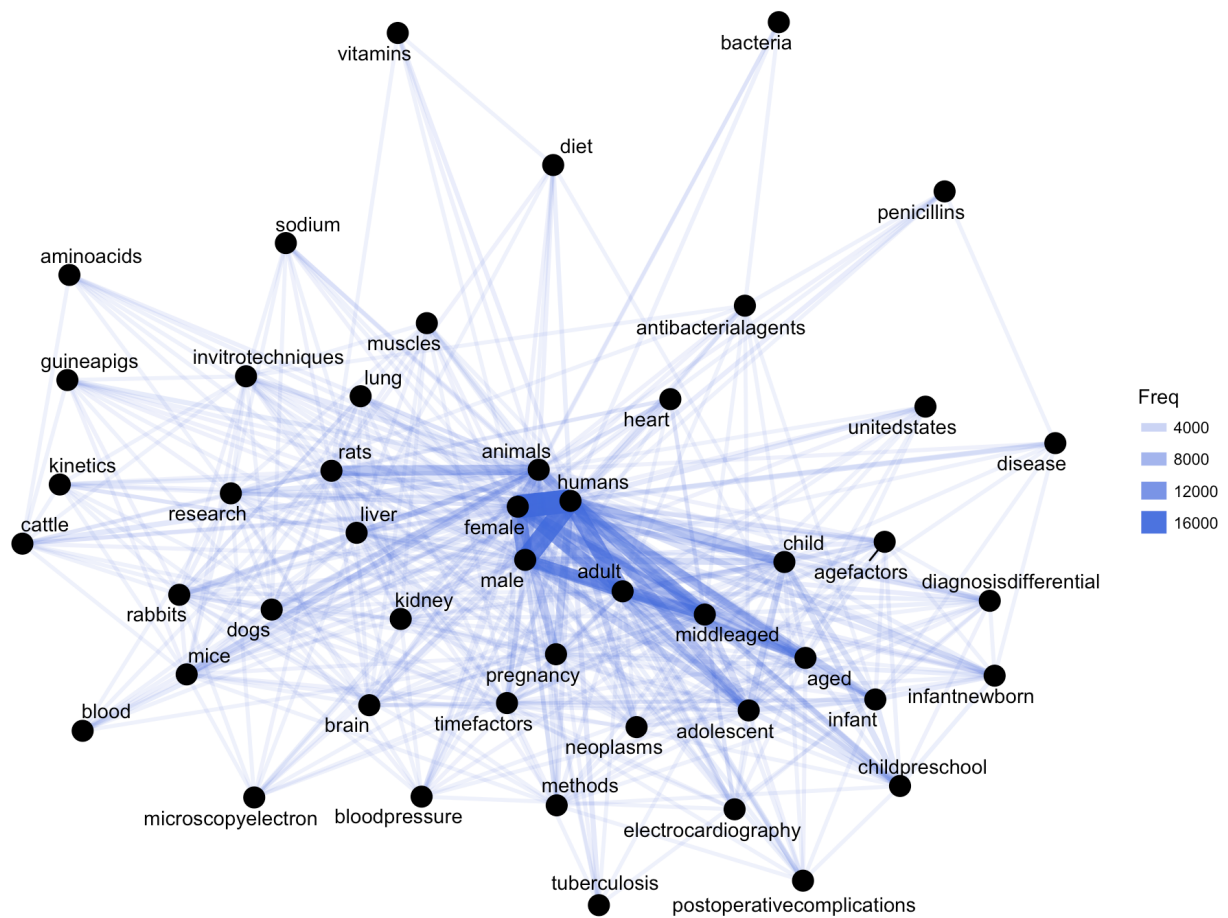


Figure 11.3: Co-occurrence network

11.2 Correlation Network

NEURAL NETWORK

Note: Sharpening the knife longer can make it easier to hack the firewood – old Chinese proverb

12.1 Feedforward Neural Network

12.1.1 Introduction

A feedforward neural network is an artificial neural network wherein connections between the units do not form a cycle. As such, it is different from recurrent neural networks.

The feedforward neural network was the first and simplest type of artificial neural network devised. In this network, the information moves in only one direction, forward (see Fig. *MultiLayer Neural Network*), from the input nodes, through the hidden nodes (if any) and to the output nodes. There are no cycles or loops in the network.

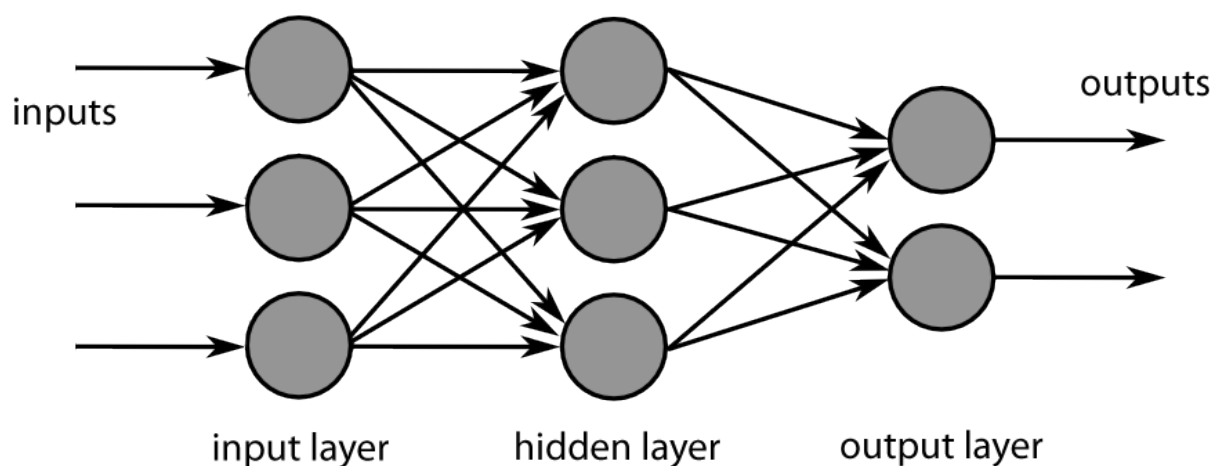


Figure 12.1: MultiLayer Neural Network

12.1.2 Demo

1. Set up spark context and SparkSession

```
from pyspark.sql import SparkSession

spark = SparkSession \
    .builder \
    .appName("Python Spark Feedforward neural network example") \
    .config("spark.some.config.option", "some-value") \
    .getOrCreate()
```

2. Load dataset

```
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
|fixed|volatile|citric|sugar|chlorides|free|total|density|pH|sulphates|alcohol|quality|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| 7.4| 0.7| 0.0| 1.9| 0.076|11.0| 34.0| 0.9978|3.51| 0.56| 9.4| 5|
| 7.8| 0.88| 0.0| 2.6| 0.098|25.0| 67.0| 0.9968| 3.2| 0.68| 9.8| 5|
| 7.8| 0.76| 0.04| 2.3| 0.092|15.0| 54.0| 0.997|3.26| 0.65| 9.8| 5|
| 11.2| 0.28| 0.56| 1.9| 0.075|17.0| 60.0| 0.998|3.16| 0.58| 9.8| 6|
| 7.4| 0.7| 0.0| 1.9| 0.076|11.0| 34.0| 0.9978|3.51| 0.56| 9.4| 5|
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
only showing top 5 rows
```

3. change categorical variable size

```
# Convert to float format
def string_to_float(x):
    return float(x)

#
def condition(r):
    if (0<= r <= 4):
        label = "low"
    elif(4< r <= 6):
        label = "medium"
    else:
        label = "high"
    return label

from pyspark.sql.functions import udf
from pyspark.sql.types import StringType, DoubleType
string_to_float_udf = udf(string_to_float, DoubleType())
quality_udf = udf(lambda x: condition(x), StringType())
df= df.withColumn("quality", quality_udf("quality"))
```

4. Convert the data to dense vector

```
# convert the data to dense vector
def transData(data):
    return data.rdd.map(lambda r: [r[-1], Vectors.dense(r[:-1])]).\
        toDF(['label', 'features'])
```

```
from pyspark.sql import Row
from pyspark.ml.linalg import Vectors
```

```
data= transData(df)
data.show()
```

5. Split the data into training and test sets (40% held out for testing)

```
# Split the data into train and test
(trainingData, testData) = data.randomSplit([0.6, 0.4])
```

6. Train neural network

```
# specify layers for the neural network:
# input layer of size 11 (features), two intermediate of size 5 and 4
# and output of size 7 (classes)
layers = [11, 5, 4, 4, 3, 7]

# create the trainer and set its parameters
FNN = MultilayerPerceptronClassifier(labelCol="indexedLabel", \
                                     featuresCol="indexedFeatures", \
                                     maxIter=100, layers=layers, \
                                     blockSize=128, seed=1234)

# Convert indexed labels back to original labels.
labelConverter = IndexToString(inputCol="prediction", outputCol="predictedLabel",
                              labels=labelIndexer.labels)

# Chain indexers and forest in a Pipeline
from pyspark.ml import Pipeline
pipeline = Pipeline(stages=[labelIndexer, featureIndexer, FNN, labelConverter])
# train the model
# Train model. This also runs the indexers.
model = pipeline.fit(trainingData)
```

7. Make predictions

```
# Make predictions.
predictions = model.transform(testData)
# Select example rows to display.
predictions.select("features", "label", "predictedLabel").show(5)
```

8. Evaluation

```
# Select (prediction, true label) and compute test error
evaluator = MulticlassClassificationEvaluator(
    labelCol="indexedLabel", predictionCol="prediction", metricName="accuracy")
accuracy = evaluator.evaluate(predictions)
print("Predictions accuracy = %g, Test Error = %g" % (accuracy, (1.0 - accuracy)))
```

**CHAPTER
THIRTEEN**

MAIN REFERENCE

BIBLIOGRAPHY

- [Bird2009] 19. Bird, E. Klein, and E. Loper. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.
- [Feng2017] 23. Feng and M. Chen. [Learning Apache Spark](#), Github 2017.
- [Karau2015] 8. Karau, A. Konwinski, P. Wendell and M. Zaharia. Learning Spark: Lightning-Fast Big Data Analysis. O'Reilly Media, Inc., 2015
- [Kirillov2016] Anton Kirillov. Apache Spark: core concepts, architecture and internals. <http://datastrophic.io/core-concepts-architecture-and-internals-of-apache-spark/>

C

Configure Spark on Mac and Ubuntu, [14](#)

R

Run on Databricks Community Cloud, [9](#)

S

Set up Spark on Cloud, [18](#)