Laboratory_Assignment_3_1_CAP_22_23

December 3, 2022

Laboratory Assignment 3.1

1 Introduction

In this laboratory, we will cover the basic element of programming using a map-reduce methodology. For that purpose, we will be using Apache Spark as a reference, but bear in mind that similar frameworks exists and principles can be extrapolated.

1.1 Some concepts and facts

- Spark is a distributed computing platform that operates on a cluster. Like MPI, we expect that nodes does NOT share a memory space but they are connected in high-speed dedicated network. Distributed filesystems that work over the network are extremely useful.
- It is considered the next generation of previous map-reduce standard Apache Hadoop. Main difference is thought to be the use of the memory instead of disk for intermediate operations, but there are many more improvements.
- It is built on Java. Despite this, it can be programmed using Java, Scala, Python or R. The complete API can only be found in JVM-based languages but the most frequent one is PySpark, since people is reluctant to use JVM-based languages in data science. Indeed, since Hadoop was only available for Java, it is likely that some Java codes of Spark are adaptations of previous Hadoop codes.
- Resilient Distributed Dataset (RDD): the basic unit that is processed in Spark. Equivalent to a numpy array but distributed.
- RDD API usually exposes the low-level operations of Apache Spark, useful for preprocessing data but useless for data analytics
- For data analysis, Dataframe and Spark SQL is used. It relies on a pandas-alike API that even accepts SQL code (which may sound crazy and useless for developers, but many *old* data scientists and statisticians are really proficient in SQL but not in Python).

1.2 How to install Spark in colab.

[58]:

```
!apt-get install openjdk-8-jdk-headless -qq > /dev/null # Install JVM v8
#!wget -q https://downloads.apache.org/spark/spark-2.4.5/spark-2.4.
↪5-bin-hadoop2.7.tgz # Download latest release. Update if necessary
!wget -q https://downloads.apache.org/spark/spark-3.3.1/spark-3.3.1-bin-hadoop3.
↪tgz
!pip install pyspark # Well, the library itself
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: pyspark in /usr/local/lib/python3.8/dist-packages
(3.3.1)
Requirement already satisfied: py4j==0.10.9.5 in /usr/local/lib/python3.8/dist-
packages (from pyspark) (0.10.9.5)
```

[59]:

```
!tar xvzf spark-3.3.1-bin-hadoop3.tgz # Unzip
```

```
spark-3.3.1-bin-hadoop3/
spark-3.3.1-bin-hadoop3/LICENSE
spark-3.3.1-bin-hadoop3/NOTICE
spark-3.3.1-bin-hadoop3/R/
spark-3.3.1-bin-hadoop3/R/lib/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/DESCRIPTION
spark-3.3.1-bin-hadoop3/R/lib/SparkR/INDEX
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/Rd.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/features.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/hsearch.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/links.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/nsInfo.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/package.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/Meta/vignette.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/NAMESPACE
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdb
spark-3.3.1-bin-hadoop3/R/lib/SparkR/R/SparkR.rdx
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/index.html
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.Rmd
spark-3.3.1-bin-hadoop3/R/lib/SparkR/doc/sparkr-vignettes.html
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/AnIndex
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdb
```

```
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/SparkR.rdx
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/aliases.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/help/paths.rds
spark-3.3.1-bin-hadoop3/R/lib/SparkR/html/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/html/00Index.html
spark-3.3.1-bin-hadoop3/R/lib/SparkR/html/R.css
spark-3.3.1-bin-hadoop3/R/lib/SparkR/profile/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/profile/general.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/profile/shell.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/tests/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/tests/testthat/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/tests/testthat/test_basic.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/worker/
spark-3.3.1-bin-hadoop3/R/lib/SparkR/worker/daemon.R
spark-3.3.1-bin-hadoop3/R/lib/SparkR/worker/worker.R
spark-3.3.1-bin-hadoop3/R/lib/sparkr.zip
spark-3.3.1-bin-hadoop3/README.md
spark-3.3.1-bin-hadoop3/RELEASE
spark-3.3.1-bin-hadoop3/bin/
spark-3.3.1-bin-hadoop3/bin/beeline
spark-3.3.1-bin-hadoop3/bin/beeline.cmd
spark-3.3.1-bin-hadoop3/bin/docker-image-tool.sh
spark-3.3.1-bin-hadoop3/bin/find-spark-home
spark-3.3.1-bin-hadoop3/bin/find-spark-home.cmd
spark-3.3.1-bin-hadoop3/bin/load-spark-env.cmd
spark-3.3.1-bin-hadoop3/bin/load-spark-env.sh
spark-3.3.1-bin-hadoop3/bin/pyspark
spark-3.3.1-bin-hadoop3/bin/pyspark.cmd
spark-3.3.1-bin-hadoop3/bin/pyspark2.cmd
spark-3.3.1-bin-hadoop3/bin/run-example
spark-3.3.1-bin-hadoop3/bin/run-example.cmd
spark-3.3.1-bin-hadoop3/bin/spark-class
spark-3.3.1-bin-hadoop3/bin/spark-class.cmd
spark-3.3.1-bin-hadoop3/bin/spark-class2.cmd
spark-3.3.1-bin-hadoop3/bin/spark-shell
spark-3.3.1-bin-hadoop3/bin/spark-shell.cmd
spark-3.3.1-bin-hadoop3/bin/spark-shell2.cmd
spark-3.3.1-bin-hadoop3/bin/spark-sql
spark-3.3.1-bin-hadoop3/bin/spark-sql.cmd
spark-3.3.1-bin-hadoop3/bin/spark-sql2.cmd
spark-3.3.1-bin-hadoop3/bin/spark-submit
spark-3.3.1-bin-hadoop3/bin/spark-submit.cmd
spark-3.3.1-bin-hadoop3/bin/spark-submit2.cmd
spark-3.3.1-bin-hadoop3/bin/sparkR
spark-3.3.1-bin-hadoop3/bin/sparkR.cmd
spark-3.3.1-bin-hadoop3/bin/sparkR2.cmd
spark-3.3.1-bin-hadoop3/conf/
spark-3.3.1-bin-hadoop3/conf/fairscheduler.xml.template
```

```
spark-3.3.1-bin-hadoop3/conf/log4j2.properties.template
spark-3.3.1-bin-hadoop3/conf/metrics.properties.template
spark-3.3.1-bin-hadoop3/conf/spark-defaults.conf.template
spark-3.3.1-bin-hadoop3/conf/spark-env.sh.template
spark-3.3.1-bin-hadoop3/conf/workers.template
spark-3.3.1-bin-hadoop3/data/
spark-3.3.1-bin-hadoop3/data/graphx/
spark-3.3.1-bin-hadoop3/data/graphx/followers.txt
spark-3.3.1-bin-hadoop3/data/graphx/users.txt
spark-3.3.1-bin-hadoop3/data/mllib/
spark-3.3.1-bin-hadoop3/data/mllib/als/
spark-3.3.1-bin-hadoop3/data/mllib/als/sample_movielens_ratings.txt
spark-3.3.1-bin-hadoop3/data/mllib/als/test.data
spark-3.3.1-bin-hadoop3/data/mllib/gmm_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/images/
spark-3.3.1-bin-hadoop3/data/mllib/images/license.txt
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/29.5.a_b_EGDP022204.jpg
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/54893.jpg
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/DP153539.jpg
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/DP802813.jpg
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/kittens/not-image.txt
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/license.txt
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/multi-channel/
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA.png
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/multi-channel/BGRA_alpha_60.png
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/multi-channel/chr30.4.184.jpg
spark-3.3.1-bin-hadoop3/data/mllib/images/origin/multi-channel/grayscale.jpg
spark-3.3.1-bin-hadoop3/data/mllib/kmeans_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/pagerank_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/pic_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/ridge-data/
spark-3.3.1-bin-hadoop3/data/mllib/ridge-data/lpsa.data
spark-3.3.1-bin-hadoop3/data/mllib/sample_binary_classification_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_fpgrowth.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_isotonic_regression_libsvm_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_kmeans_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_lda_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_lda_libsvm_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_libsvm_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_linear_regression_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_movielens_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_multiclass_classification_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/sample_svm_data.txt
spark-3.3.1-bin-hadoop3/data/mllib/streaming_kmeans_data_test.txt
spark-3.3.1-bin-hadoop3/data/streaming/
spark-3.3.1-bin-hadoop3/data/streaming/AFINN-111.txt
```

```
spark-3.3.1-bin-hadoop3/examples/
spark-3.3.1-bin-hadoop3/examples/jars/
spark-3.3.1-bin-hadoop3/examples/jars/scopt_2.12-3.7.1.jar
spark-3.3.1-bin-hadoop3/examples/jars/spark-examples_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/examples/src/
spark-3.3.1-bin-hadoop3/examples/src/main/
spark-3.3.1-bin-hadoop3/examples/src/main/java/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaHdfsLR.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaLogQuery.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaPageRank.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaSparkPi.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/JavaSta
tusTrackerDemo.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaTC.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/JavaWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
AFTSurvivalRegressionExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaALSExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
BinarizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
BisectingKMeansExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
BucketedRandomProjectionLSHExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
BucketizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ChiSqSelectorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ChiSquareTestExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
CorrelationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
CountVectorizerExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDCTExample.java
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
DecisionTreeClassificationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
DecisionTreeRegressionExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaDocument.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ElementwiseProductExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
EstimatorTransformerParamExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
FMClassifierExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
FMRegressorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
FPGrowthExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
FeatureHasherExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
GaussianMixtureExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
GeneralizedLinearRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
GradientBoostedTreeClassifierExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
GradientBoostedTreeRegressorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ImputerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
IndexToStringExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
InteractionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
IsotonicRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
KMeansExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaLDAExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
LabeledDocument.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
LinearRegressionWithElasticNetExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
LinearSVCExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
LogisticRegressionSummaryExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
LogisticRegressionWithElasticNetExample.java
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
MaxAbsScalerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
MinHashLSHExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
MinMaxScalerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ModelSelectionViaCrossValidationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
ModelSelectionViaTrainValidationSplitExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
MulticlassLogisticRegressionWithElasticNetExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
MultilayerPerceptronClassifierExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
NGramExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
NaiveBayesExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
NormalizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
OneHotEncoderExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
OneVsRestExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/ml/JavaPCAExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
PipelineExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
PolynomialExpansionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
PowerIterationClusteringExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
PrefixSpanExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
QuantileDiscretizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
RFormulaExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
RandomForestClassifierExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
RandomForestRegressorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
RobustScalerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
SQLTransformerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
StandardScalerExample.java
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
StopWordsRemoverExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
StringIndexerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
SummarizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
TfIdfExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
TokenizerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
UnivariateFeatureSelectorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
VarianceThresholdSelectorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
VectorAssemblerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
VectorIndexerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
VectorSizeHintExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
VectorSlicerExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/ml/Java
Word2VecExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/JavaALS.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaAssociationRulesExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaBinaryClassificationMetricsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaBisectingKMeansExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaChiSqSelectorExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaCorrelationsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaDecisionTreeClassificationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaDecisionTreeRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaElementwiseProductExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaGaussianMixtureExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaGradientBoostingClassificationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
```

```
avaGradientBoostingRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaHypothesisTestingExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaHypothesisTestingKolmogorovSmirnovTestExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaIsotonicRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaKMeansExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaKernelDensityEstimationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaLBFGSExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaLatentDirichletAllocationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaLogisticRegressionWithLBFGSExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaMultiLabelClassificationMetricsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaMulticlassClassificationMetricsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaNaiveBayesExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaPCAExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaPowerIterationClusteringExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaPrefixSpanExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaRandomForestClassificationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaRandomForestRegressionExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaRankingMetricsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaRecommendationExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaSVDExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaSVMWithSGDExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaSimpleFPGrowth.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaStratifiedSamplingExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
avaStreamingTestExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/mllib/J
```

```
avaSummaryStatisticsExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/Jav
aSQLDataSourceExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/Jav
aSparkSQLExample.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/Jav
aUserDefinedScalar.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/Jav
aUserDefinedTypedAggregation.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/Jav
aUserDefinedUntypedAggregation.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/sql/hive/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/hiv
e/JavaSparkHiveExample.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/sql/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredComplexSessionization.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredKafkaWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredKerberizedKafkaWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredNetworkWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredNetworkWordCountWindowed.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/sql/str
eaming/JavaStructuredSessionization.java
spark-3.3.1-bin-
hadoop3/examples/src/main/java/org/apache/spark/examples/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaCustomReceiver.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaDirectKafkaWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaDirectKerberizedKafkaWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaNetworkWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaQueueStream.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaRecord.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaRecoverableNetworkWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaSqlNetworkWordCount.java
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/java/org/apache/spark/examples/streami
ng/JavaStatefulNetworkWordCount.java
spark-3.3.1-bin-hadoop3/examples/src/main/python/
spark-3.3.1-bin-hadoop3/examples/src/main/python/__init__.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/als.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/avro_inputformat.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/kmeans.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/logistic_regression.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/__init__,py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/aft_survival_regression.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/als_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/binarizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/bisecting_k_means_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/bucketed_random_projection_lsh_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/bucketizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/chi_square_test_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/chisq_selector_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/correlation_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/count_vectorizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/cross_validator.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/dataframe_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/dct_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/decision_tree_classification_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/decision_tree_regression_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/elementwise_product_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/estimator_transformer_param_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/feature_hasher_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/fm_classifier_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/fm_regressor_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/fpgrowth_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/gaussian_mixture_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/generalized_linear_regression_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/gradient_boosted_tree_classifier_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/gradient_boosted_tree_regressor_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/imputer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/index_to_string_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/interaction_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/isotonic_regression_example.py
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/kmeans_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/lda_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/linear_regression_with_elastic_net.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/linearsvc.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/logistic_regression_summary_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/logistic_regression_with_elastic_net.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/max_abs_scaler_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/min_hash_lsh_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/min_max_scaler_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/multiclass_logistic_regressi
on_with_elastic_net.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/multilayer_perceptron_classification.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/n_gram_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/naive_bayes_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/normalizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/one_vs_rest_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/onehot_encoder_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/pca_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/pipeline_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/polynomial_expansion_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/power_iteration_clustering_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/prefixspan_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/quantile_discretizer_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/random_forest_classifier_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/random_forest_regressor_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/rformula_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/robust_scaler_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/sql_transformer.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/standard_scaler_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/stopwords_remover_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/string_indexer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/summarizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/tf_idf_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/tokenizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/train_validation_split.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/univariate_feature_selector_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/ml/variance_threshold_selector_example.py
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/vector_assembler_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/vector_indexer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/vector_size_hint_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/vector_slicer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/ml/word2vec_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/__init__.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/binary_classification_metrics_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/bisecting_k_means_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/correlations.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/correlations_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/decision_tree_classification_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/decision_tree_regression_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/elementwise_product_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/fpgrowth_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/gaussian_mixture_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/gaussian_mixture_model.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/gradient_boosting_classif
ication_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/gradient_boosting_regression_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/hypothesis_testing_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/hypothesis_testing_kolmog
orov_smirnov_test_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/isotonic_regression_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/k_means_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/kernel_density_estimation_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/kmeans.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/latent_dirichlet_allocation_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/linear_regression_with_sgd_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/logistic_regression.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/logistic_regression_with_lbfgs_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/multi_class_metrics_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/multi_label_metrics_example.py
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/naive_bayes_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/normalizer_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/pca_rowmatrix_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/power_iteration_clustering_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/random_forest_classification_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/random_forest_regression_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/random_rdd_generation.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/ranking_metrics_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/recommendation_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/regression_metrics_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/sampled_rdds.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/standard_scaler_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/stratified_sampling_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/streaming_k_means_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/streaming_linear_regression_example.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/mllib/summary_statistics_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/svd_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/svm_with_sgd_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/tf_idf_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/word2vec.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/mllib/word2vec_example.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/pagerank.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/parquet_inputformat.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/pi.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sort.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/__init__.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/arrow.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/basic.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/datasource.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/hive.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/streaming/__init__,py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/sql/streaming/structured_kafka_wordcount.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/sql/streaming/structured_network_wordcount.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/sql/streaming/structured_networ
```

```
k_wordcount_windowed.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/sql/streaming/structured_sessionization.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/status_api_demo.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/python/streaming/__init__.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/streaming/hdfs_wordcount.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/streaming/network_wordcount.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/streaming/network_wordjoinsentiments.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/streaming/queue_stream.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/streaming/recoverable_network_wordcount.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/streaming/sql_network_wordcount.py
spark-3.3.1-bin-
hadoop3/examples/src/main/python/streaming/stateful_network_wordcount.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/transitive_closure.py
spark-3.3.1-bin-hadoop3/examples/src/main/python/wordcount.py
spark-3.3.1-bin-hadoop3/examples/src/main/r/
spark-3.3.1-bin-hadoop3/examples/src/main/r/RSparkSQLExample.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/data-manipulation.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/dataframe.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/als.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/bisectingKmeans.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/decisionTree.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/fmClassifier.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/fmRegressor.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/fpm.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/gaussianMixture.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/gbt.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/glm.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/isoreg.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/kmeans.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/kstest.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/lda.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/lm_with_elastic_net.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/logit.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/ml.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/mlp.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/naiveBayes.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/powerIterationClustering.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/prefixSpan.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/randomForest.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/survreg.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/ml/svmLinear.R
spark-3.3.1-bin-hadoop3/examples/src/main/r/streaming/
```

```
spark-3.3.1-bin-
hadoop3/examples/src/main/r/streaming/structured_network_wordcount.R
spark-3.3.1-bin-hadoop3/examples/src/main/resources/
spark-3.3.1-bin-hadoop3/examples/src/main/resources/META-INF/
spark-3.3.1-bin-hadoop3/examples/src/main/resources/META-INF/services/
spark-3.3.1-bin-hadoop3/examples/src/main/resources/META-
INF/services/org.apache.spark.sql.SparkSessionExtensionsProvider
spark-3.3.1-bin-hadoop3/examples/src/main/resources/META-
INF/services/org.apache.spark.sql.jdbc.JdbcConnectionProvider
spark-3.3.1-bin-hadoop3/examples/src/main/resources/dir1/
spark-3.3.1-bin-hadoop3/examples/src/main/resources/dir1/dir2/
spark-3.3.1-bin-hadoop3/examples/src/main/resources/dir1/dir2/file2.parquet
spark-3.3.1-bin-hadoop3/examples/src/main/resources/dir1/file1.parquet
spark-3.3.1-bin-hadoop3/examples/src/main/resources/dir1/file3.json
spark-3.3.1-bin-hadoop3/examples/src/main/resources/employees.json
spark-3.3.1-bin-hadoop3/examples/src/main/resources/full_user.avsc
spark-3.3.1-bin-hadoop3/examples/src/main/resources/kv1.txt
spark-3.3.1-bin-hadoop3/examples/src/main/resources/people.csv
spark-3.3.1-bin-hadoop3/examples/src/main/resources/people.json
spark-3.3.1-bin-hadoop3/examples/src/main/resources/people.txt
spark-3.3.1-bin-hadoop3/examples/src/main/resources/user.avsc
spark-3.3.1-bin-hadoop3/examples/src/main/resources/users.avro
spark-3.3.1-bin-hadoop3/examples/src/main/resources/users.orc
spark-3.3.1-bin-hadoop3/examples/src/main/resources/users.parquet
spark-3.3.1-bin-hadoop3/examples/src/main/scala/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Accumu
latorMetricsTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/BroadcastTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/DFSReadWriteTest.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Driver
SubmissionTest.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Except
ionHandlingTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/GroupByTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/HdfsTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalALS.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalFileLR.scala
spark-3.3.1-bin-
```

```
hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalKMeans.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalLR.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/LocalPi.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/LogQuery.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/MiniRe
adWriteTest.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/MultiB
roadcastTest.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Simple
SkewedGroupByTest.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/Skewed
GroupByTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkALS.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkHdfsLR.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkKMeans.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkLR.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkPageRank.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkPi.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkR
emoteFileTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/SparkTC.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/extensions/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extens
ions/AgeExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extens
ions/SessionExtensionsWithLoader.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extens
ions/SessionExtensionsWithoutLoader.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/extens
ions/SparkSessionExtensionsTest.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/AggregateMessagesExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx/Analytics.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
```

```
/ComprehensiveExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/ConnectedComponentsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/LiveJournalPageRank.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/PageRankExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/SSSPExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/SynthBenchmark.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/graphx
/TriangleCountingExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/AFT
SurvivalRegressionExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ALSExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Bin
arizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Bis
ectingKMeansExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Buc
ketedRandomProjectionLSHExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Buc
ketizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Chi
SqSelectorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Chi
SquareTestExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Cor
relationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Cou
ntVectorizerExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/DCTExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Dat
aFrameExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Dec
isionTreeClassificationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Dec
isionTreeExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Dec
isionTreeRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Dev
eloperApiExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Ele
mentwiseProductExample.scala
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/EstimatorTransformerParamExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FMClassifierExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FMRegressorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FPGrowthExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/FeatureHasherExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GBTExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GaussianMixtureExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GeneralizedLinearRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeClassifierExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/GradientBoostedTreeRegressorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/ImputerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/IndexToStringExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/InteractionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/IsotonicRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/KMeansExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LDAExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearRegressionWithElasticNetExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LinearSVCExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionSummaryExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/LogisticRegressionWithElasticNetExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MaxAbsScalerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/MinHashLSHExample.scala
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Min
MaxScalerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Mod
elSelectionViaCrossValidationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Mod
elSelectionViaTrainValidationSplitExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Mul
ticlassLogisticRegressionWithElasticNetExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Mul
tilayerPerceptronClassifierExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/NGramExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Nai
veBayesExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Nor
malizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/One
HotEncoderExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/One
VsRestExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/PCAExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Pip
elineExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Pol
ynomialExpansionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Pow
erIterationClusteringExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Pre
fixSpanExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Qua
ntileDiscretizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/RFo
rmulaExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Ran
domForestClassifierExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Ran
domForestExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Ran
domForestRegressorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Rob
ustScalerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/SQL
TransformerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Sta
ndardScalerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Sto
pWordsRemoverExample.scala
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Str
ingIndexerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Sum
marizerExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/TfIdfExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Tok
enizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Una
ryTransformerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Uni
variateFeatureSelectorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Var
ianceThresholdSelectorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Vec
torAssemblerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Vec
torIndexerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Vec
torSizeHintExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Vec
torSlicerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/ml/Wor
d2VecExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
AbstractParams.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
AssociationRulesExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
BinaryClassification.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
BinaryClassificationMetricsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
BisectingKMeansExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
ChiSqSelectorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
Correlations.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
CorrelationsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
CosineSimilarity.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
DecisionTreeClassificationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
DecisionTreeRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
```

```
DecisionTreeRunner.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
DenseKMeans.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
ElementwiseProductExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
FPGrowthExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
GaussianMixtureExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
GradientBoostedTreesRunner.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
GradientBoostingClassificationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
GradientBoostingRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
HypothesisTestingExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
HypothesisTestingKolmogorovSmirnovTestExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
IsotonicRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
KMeansExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
KernelDensityEstimationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
LBFGSExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/LDAExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
LatentDirichletAllocationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
LogisticRegressionWithLBFGSExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
MovieLensALS.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
MultiLabelMetricsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
MulticlassMetricsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
MultivariateSummarizer.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
NaiveBayesExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
NormalizerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
PCAOnRowMatrixExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
```

```
PCAOnSourceVectorExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
PMMLModelExportExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
PowerIterationClusteringExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
PrefixSpanExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
RandomForestClassificationExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
RandomForestRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
RandomRDDGeneration.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
RankingMetricsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
RecommendationExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/SVDExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
SVMWithSGDExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
SampledRDDs.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
SimpleFPGrowth.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
SparseNaiveBayes.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StandardScalerExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StratifiedSamplingExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StreamingKMeansExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StreamingLinearRegressionExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StreamingLogisticRegression.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
StreamingTestExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
SummaryStatisticsExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
TFIDFExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
TallSkinnyPCA.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
TallSkinnySVD.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/mllib/
```

```
Word2VecExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/pythonconverters/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/python
converters/AvroConverters.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/RDDRelation.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/SQ
LDataSourceExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/Si
mpleTypedAggregator.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/Sp
arkSQLExample.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/Us
erDefinedScalar.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/Us
erDefinedTypedAggregation.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/Us
erDefinedUntypedAggregation.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/hive/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/hi
ve/SparkHiveExample.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/jdbc/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/jd
bc/ExampleJdbcConnectionProvider.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredComplexSessionization.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredKafkaWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredKerberizedKafkaWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredNetworkWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredNetworkWordCountWindowed.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/sql/st
reaming/StructuredSessionization.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/CustomReceiver.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/DirectKafkaWordCount.scala
```

```
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/DirectKerberizedKafkaWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/HdfsWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/NetworkWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/QueueStream.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/RawNetworkGrep.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/RecoverableNetworkWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/SqlNetworkWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/StatefulNetworkWordCount.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/StreamingExamples.scala
spark-3.3.1-bin-
hadoop3/examples/src/main/scala/org/apache/spark/examples/streaming/clickstream/
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/clickstream/PageViewGenerator.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scala/org/apache/spark/examples/stream
ing/clickstream/PageViewStream.scala
spark-3.3.1-bin-hadoop3/examples/src/main/scripts/
spark-3.3.1-bin-hadoop3/examples/src/main/scripts/getGpusResources.sh
spark-3.3.1-bin-hadoop3/jars/
spark-3.3.1-bin-hadoop3/jars/HikariCP-2.5.1.jar
spark-3.3.1-bin-hadoop3/jars/JLargeArrays-1.5.jar
spark-3.3.1-bin-hadoop3/jars/JTransforms-3.1.jar
spark-3.3.1-bin-hadoop3/jars/RoaringBitmap-0.9.25.jar
spark-3.3.1-bin-hadoop3/jars/ST4-4.0.4.jar
spark-3.3.1-bin-hadoop3/jars/activation-1.1.1.jar
spark-3.3.1-bin-hadoop3/jars/aircompressor-0.21.jar
spark-3.3.1-bin-hadoop3/jars/algebra_2.12-2.0.1.jar
spark-3.3.1-bin-hadoop3/jars/annotations-17.0.0.jar
spark-3.3.1-bin-hadoop3/jars/antlr-runtime-3.5.2.jar
spark-3.3.1-bin-hadoop3/jars/antlr4-runtime-4.8.jar
spark-3.3.1-bin-hadoop3/jars/aopalliance-repackaged-2.6.1.jar
spark-3.3.1-bin-hadoop3/jars/arpack-2.2.1.jar
spark-3.3.1-bin-hadoop3/jars/arpack_combined_all-0.1.jar
spark-3.3.1-bin-hadoop3/jars/arrow-format-7.0.0.jar
spark-3.3.1-bin-hadoop3/jars/arrow-memory-core-7.0.0.jar
spark-3.3.1-bin-hadoop3/jars/arrow-memory-netty-7.0.0.jar
spark-3.3.1-bin-hadoop3/jars/arrow-vector-7.0.0.jar
spark-3.3.1-bin-hadoop3/jars/audience-annotations-0.5.0.jar
spark-3.3.1-bin-hadoop3/jars/automaton-1.11-8.jar
spark-3.3.1-bin-hadoop3/jars/avro-1.11.0.jar
```

```
spark-3.3.1-bin-hadoop3/jars/avro-ipc-1.11.0.jar
spark-3.3.1-bin-hadoop3/jars/avro-mapred-1.11.0.jar
spark-3.3.1-bin-hadoop3/jars/blas-2.2.1.jar
spark-3.3.1-bin-hadoop3/jars/bonecp-0.8.0.RELEASE.jar
spark-3.3.1-bin-hadoop3/jars/breeze-macros_2.12-1.2.jar
spark-3.3.1-bin-hadoop3/jars/breeze_2.12-1.2.jar
spark-3.3.1-bin-hadoop3/jars/cats-kernel_2.12-2.1.1.jar
spark-3.3.1-bin-hadoop3/jars/chill-java-0.10.0.jar
spark-3.3.1-bin-hadoop3/jars/chill_2.12-0.10.0.jar
spark-3.3.1-bin-hadoop3/jars/commons-cli-1.5.0.jar
spark-3.3.1-bin-hadoop3/jars/commons-codec-1.15.jar
spark-3.3.1-bin-hadoop3/jars/commons-collections-3.2.2.jar
spark-3.3.1-bin-hadoop3/jars/commons-collections4-4.4.jar
spark-3.3.1-bin-hadoop3/jars/commons-compiler-3.0.16.jar
spark-3.3.1-bin-hadoop3/jars/commons-compress-1.21.jar
spark-3.3.1-bin-hadoop3/jars/commons-crypto-1.1.0.jar
spark-3.3.1-bin-hadoop3/jars/commons-dbcp-1.4.jar
spark-3.3.1-bin-hadoop3/jars/commons-io-2.11.0.jar
spark-3.3.1-bin-hadoop3/jars/commons-lang-2.6.jar
spark-3.3.1-bin-hadoop3/jars/commons-lang3-3.12.0.jar
spark-3.3.1-bin-hadoop3/jars/commons-logging-1.1.3.jar
spark-3.3.1-bin-hadoop3/jars/commons-math3-3.6.1.jar
spark-3.3.1-bin-hadoop3/jars/commons-pool-1.5.4.jar
spark-3.3.1-bin-hadoop3/jars/commons-text-1.9.jar
spark-3.3.1-bin-hadoop3/jars/compress-lzf-1.1.jar
spark-3.3.1-bin-hadoop3/jars/core-1.1.2.jar
spark-3.3.1-bin-hadoop3/jars/curator-client-2.13.0.jar
spark-3.3.1-bin-hadoop3/jars/curator-framework-2.13.0.jar
spark-3.3.1-bin-hadoop3/jars/curator-recipes-2.13.0.jar
spark-3.3.1-bin-hadoop3/jars/datanucleus-api-jdo-4.2.4.jar
spark-3.3.1-bin-hadoop3/jars/datanucleus-core-4.1.17.jar
spark-3.3.1-bin-hadoop3/jars/datanucleus-rdbms-4.1.19.jar
spark-3.3.1-bin-hadoop3/jars/derby-10.14.2.0.jar
spark-3.3.1-bin-hadoop3/jars/dropwizard-metrics-hadoop-
metrics2-reporter-0.1.2.jar
spark-3.3.1-bin-hadoop3/jars/flatbuffers-java-1.12.0.jar
spark-3.3.1-bin-hadoop3/jars/generex-1.0.2.jar
spark-3.3.1-bin-hadoop3/jars/gson-2.2.4.jar
spark-3.3.1-bin-hadoop3/jars/guava-14.0.1.jar
spark-3.3.1-bin-hadoop3/jars/hadoop-client-api-3.3.2.jar
spark-3.3.1-bin-hadoop3/jars/hadoop-client-runtime-3.3.2.jar
spark-3.3.1-bin-hadoop3/jars/hadoop-shaded-guava-1.1.1.jar
spark-3.3.1-bin-hadoop3/jars/hadoop-yarn-server-web-proxy-3.3.2.jar
spark-3.3.1-bin-hadoop3/jars/hive-beeline-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-cli-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-common-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-exec-2.3.9-core.jar
spark-3.3.1-bin-hadoop3/jars/hive-jdbc-2.3.9.jar
```

```
spark-3.3.1-bin-hadoop3/jars/hive-llap-common-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-metastore-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-serde-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-service-rpc-3.1.2.jar
spark-3.3.1-bin-hadoop3/jars/hive-shims-0.23-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-shims-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-shims-common-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-shims-scheduler-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hive-storage-api-2.7.2.jar
spark-3.3.1-bin-hadoop3/jars/hive-vector-code-gen-2.3.9.jar
spark-3.3.1-bin-hadoop3/jars/hk2-api-2.6.1.jar
spark-3.3.1-bin-hadoop3/jars/hk2-locator-2.6.1.jar
spark-3.3.1-bin-hadoop3/jars/hk2-utils-2.6.1.jar
spark-3.3.1-bin-hadoop3/jars/httpclient-4.5.13.jar
spark-3.3.1-bin-hadoop3/jars/httpcore-4.4.14.jar
spark-3.3.1-bin-hadoop3/jars/istack-commons-runtime-3.0.8.jar
spark-3.3.1-bin-hadoop3/jars/ivy-2.5.0.jar
spark-3.3.1-bin-hadoop3/jars/jackson-annotations-2.13.4.jar
spark-3.3.1-bin-hadoop3/jars/jackson-core-2.13.4.jar
spark-3.3.1-bin-hadoop3/jars/jackson-core-asl-1.9.13.jar
spark-3.3.1-bin-hadoop3/jars/jackson-databind-2.13.4.1.jar
spark-3.3.1-bin-hadoop3/jars/jackson-dataformat-yaml-2.13.4.jar
spark-3.3.1-bin-hadoop3/jars/jackson-datatype-jsr310-2.13.4.jar
spark-3.3.1-bin-hadoop3/jars/jackson-mapper-asl-1.9.13.jar
spark-3.3.1-bin-hadoop3/jars/jackson-module-scala_2.12-2.13.4.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.annotation-api-1.3.5.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.inject-2.6.1.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.servlet-api-4.0.3.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.validation-api-2.0.2.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.ws.rs-api-2.1.6.jar
spark-3.3.1-bin-hadoop3/jars/jakarta.xml.bind-api-2.3.2.jar
spark-3.3.1-bin-hadoop3/jars/janino-3.0.16.jar
spark-3.3.1-bin-hadoop3/jars/javassist-3.25.0-GA.jar
spark-3.3.1-bin-hadoop3/jars/javax.jdo-3.2.0-m3.jar
spark-3.3.1-bin-hadoop3/jars/javolution-5.5.1.jar
spark-3.3.1-bin-hadoop3/jars/jaxb-runtime-2.3.2.jar
spark-3.3.1-bin-hadoop3/jars/jcl-over-slf4j-1.7.32.jar
spark-3.3.1-bin-hadoop3/jars/jdo-api-3.0.1.jar
spark-3.3.1-bin-hadoop3/jars/jersey-client-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jersey-common-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jersey-container-servlet-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jersey-container-servlet-core-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jersey-hk2-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jersey-server-2.36.jar
spark-3.3.1-bin-hadoop3/jars/jline-2.14.6.jar
spark-3.3.1-bin-hadoop3/jars/joda-time-2.10.13.jar
spark-3.3.1-bin-hadoop3/jars/jodd-core-3.5.2.jar
spark-3.3.1-bin-hadoop3/jars/jpam-1.1.jar
```

```
spark-3.3.1-bin-hadoop3/jars/json-1.8.jar
spark-3.3.1-bin-hadoop3/jars/json4s-ast_2.12-3.7.0-M11.jar
spark-3.3.1-bin-hadoop3/jars/json4s-core_2.12-3.7.0-M11.jar
spark-3.3.1-bin-hadoop3/jars/json4s-jackson_2.12-3.7.0-M11.jar
spark-3.3.1-bin-hadoop3/jars/json4s-scalap_2.12-3.7.0-M11.jar
spark-3.3.1-bin-hadoop3/jars/jsr305-3.0.0.jar
spark-3.3.1-bin-hadoop3/jars/jta-1.1.jar
spark-3.3.1-bin-hadoop3/jars/jul-to-slf4j-1.7.32.jar
spark-3.3.1-bin-hadoop3/jars/kryo-shaded-4.0.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-client-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-admissionregistration-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-apiextensions-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-apps-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-autoscaling-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-batch-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-certificates-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-common-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-coordination-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-core-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-discovery-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-events-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-extensions-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-flowcontrol-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-metrics-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-networking-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-node-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-policy-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-rbac-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-scheduling-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/kubernetes-model-storageclass-5.12.2.jar
spark-3.3.1-bin-hadoop3/jars/lapack-2.2.1.jar
spark-3.3.1-bin-hadoop3/jars/leveldbjni-all-1.8.jar
spark-3.3.1-bin-hadoop3/jars/libfb303-0.9.3.jar
spark-3.3.1-bin-hadoop3/jars/libthrift-0.12.0.jar
spark-3.3.1-bin-hadoop3/jars/log4j-1.2-api-2.17.2.jar
spark-3.3.1-bin-hadoop3/jars/log4j-api-2.17.2.jar
spark-3.3.1-bin-hadoop3/jars/log4j-core-2.17.2.jar
spark-3.3.1-bin-hadoop3/jars/log4j-slf4j-impl-2.17.2.jar
spark-3.3.1-bin-hadoop3/jars/logging-interceptor-3.12.12.jar
spark-3.3.1-bin-hadoop3/jars/lz4-java-1.8.0.jar
spark-3.3.1-bin-hadoop3/jars/mesos-1.4.3-shaded-protobuf.jar
spark-3.3.1-bin-hadoop3/jars/metrics-core-4.2.7.jar
spark-3.3.1-bin-hadoop3/jars/metrics-graphite-4.2.7.jar
spark-3.3.1-bin-hadoop3/jars/metrics-jmx-4.2.7.jar
spark-3.3.1-bin-hadoop3/jars/metrics-json-4.2.7.jar
spark-3.3.1-bin-hadoop3/jars/metrics-jvm-4.2.7.jar
spark-3.3.1-bin-hadoop3/jars/minlog-1.3.0.jar
spark-3.3.1-bin-hadoop3/jars/netty-all-4.1.74.Final.jar
```

```
spark-3.3.1-bin-hadoop3/jars/netty-buffer-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-codec-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-common-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-handler-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-resolver-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-tcnative-classes-2.0.48.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-classes-epoll-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-classes-kqueue-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-native-epoll-4.1.74.Final-linux-
aarch_64.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-native-epoll-4.1.74.Final-
linux-x86_64.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-native-kqueue-4.1.74.Final-osx-
aarch_64.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-native-kqueue-4.1.74.Final-
osx-x86_64.jar
spark-3.3.1-bin-hadoop3/jars/netty-transport-native-unix-common-4.1.74.Final.jar
spark-3.3.1-bin-hadoop3/jars/objenesis-3.2.jar
spark-3.3.1-bin-hadoop3/jars/okhttp-3.12.12.jar
spark-3.3.1-bin-hadoop3/jars/okio-1.14.0.jar
spark-3.3.1-bin-hadoop3/jars/opencsv-2.3.jar
spark-3.3.1-bin-hadoop3/jars/orc-core-1.7.6.jar
spark-3.3.1-bin-hadoop3/jars/orc-mapreduce-1.7.6.jar
spark-3.3.1-bin-hadoop3/jars/orc-shims-1.7.6.jar
spark-3.3.1-bin-hadoop3/jars/oro-2.0.8.jar
spark-3.3.1-bin-hadoop3/jars/osgi-resource-locator-1.0.3.jar
spark-3.3.1-bin-hadoop3/jars/paranamer-2.8.jar
spark-3.3.1-bin-hadoop3/jars/parquet-column-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/parquet-common-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/parquet-encoding-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/parquet-format-structures-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/parquet-hadoop-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/parquet-jackson-1.12.2.jar
spark-3.3.1-bin-hadoop3/jars/pickle-1.2.jar
spark-3.3.1-bin-hadoop3/jars/protobuf-java-2.5.0.jar
spark-3.3.1-bin-hadoop3/jars/py4j-0.10.9.5.jar
spark-3.3.1-bin-hadoop3/jars/rocksdbjni-6.20.3.jar
spark-3.3.1-bin-hadoop3/jars/scala-collection-compat_2.12-2.1.1.jar
spark-3.3.1-bin-hadoop3/jars/scala-compiler-2.12.15.jar
spark-3.3.1-bin-hadoop3/jars/scala-library-2.12.15.jar
spark-3.3.1-bin-hadoop3/jars/scala-parser-combinators_2.12-1.1.2.jar
spark-3.3.1-bin-hadoop3/jars/scala-reflect-2.12.15.jar
spark-3.3.1-bin-hadoop3/jars/scala-xml_2.12-1.2.0.jar
spark-3.3.1-bin-hadoop3/jars/shapeless_2.12-2.3.7.jar
spark-3.3.1-bin-hadoop3/jars/shims-0.9.25.jar
spark-3.3.1-bin-hadoop3/jars/slf4j-api-1.7.32.jar
spark-3.3.1-bin-hadoop3/jars/snakeyaml-1.31.jar
```

```
spark-3.3.1-bin-hadoop3/jars/snappy-java-1.1.8.4.jar
spark-3.3.1-bin-hadoop3/jars/spark-catalyst_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-core_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-graphx_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-hive-thriftserver_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-hive_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-kubernetes_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-kvstore_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-launcher_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-mesos_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-mllib-local_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-mllib_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-network-common_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-network-shuffle_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-repl_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-sketch_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-sql_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-streaming_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-tags_2.12-3.3.1-tests.jar
spark-3.3.1-bin-hadoop3/jars/spark-tags_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-unsafe_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spark-yarn_2.12-3.3.1.jar
spark-3.3.1-bin-hadoop3/jars/spire-macros_2.12-0.17.0.jar
spark-3.3.1-bin-hadoop3/jars/spire-platform_2.12-0.17.0.jar
spark-3.3.1-bin-hadoop3/jars/spire-util_2.12-0.17.0.jar
spark-3.3.1-bin-hadoop3/jars/spire_2.12-0.17.0.jar
spark-3.3.1-bin-hadoop3/jars/stax-api-1.0.1.jar
spark-3.3.1-bin-hadoop3/jars/stream-2.9.6.jar
spark-3.3.1-bin-hadoop3/jars/super-csv-2.2.0.jar
spark-3.3.1-bin-hadoop3/jars/threeten-extra-1.5.0.jar
spark-3.3.1-bin-hadoop3/jars/tink-1.6.1.jar
spark-3.3.1-bin-hadoop3/jars/transaction-api-1.1.jar
spark-3.3.1-bin-hadoop3/jars/univocity-parsers-2.9.1.jar
spark-3.3.1-bin-hadoop3/jars/velocity-1.5.jar
spark-3.3.1-bin-hadoop3/jars/xbean-asm9-shaded-4.20.jar
spark-3.3.1-bin-hadoop3/jars/xz-1.8.jar
spark-3.3.1-bin-hadoop3/jars/zjsonpatch-0.3.0.jar
spark-3.3.1-bin-hadoop3/jars/zookeeper-3.6.2.jar
spark-3.3.1-bin-hadoop3/jars/zookeeper-jute-3.6.2.jar
spark-3.3.1-bin-hadoop3/jars/zstd-jni-1.5.2-1.jar
spark-3.3.1-bin-hadoop3/kubernetes/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/Dockerfile
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/Dockerfile.java17
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/R/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/R/Dockerfile
```

```
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/python/
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/bindings/python/Dockerfile
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/decom.sh
spark-3.3.1-bin-hadoop3/kubernetes/dockerfiles/spark/entrypoint.sh
spark-3.3.1-bin-hadoop3/kubernetes/tests/
spark-3.3.1-bin-hadoop3/kubernetes/tests/autoscale.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/decommissioning.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/decommissioning_cleanup.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/py_container_checks.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/pyfiles.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/python_executable_check.py
spark-3.3.1-bin-hadoop3/kubernetes/tests/worker_memory_check.py
spark-3.3.1-bin-hadoop3/licenses/
spark-3.3.1-bin-hadoop3/licenses/LICENSE-AnchorJS.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-CC0.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-JLargeArrays.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-JTransforms.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-antlr.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-arpack.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-automaton.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-blas.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-bootstrap.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-cloudpickle.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-d3.min.js.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-dagre-d3.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-datatables.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-dnsjava.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-f2j.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-graphlib-dot.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-istack-commons-runtime.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jakarta-annotation-api
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jakarta-ws-rs-api
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jakarta.activation-api.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jakarta.xml.bind-api.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-janino.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-javassist.html
spark-3.3.1-bin-hadoop3/licenses/LICENSE-javax-transaction-transaction-api.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-javolution.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jaxb-runtime.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jline.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jodd.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-join.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jquery.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-json-formatter.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-jsp-api.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-kryo.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-leveldbjni.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-machinist.txt
```

```
spark-3.3.1-bin-hadoop3/licenses/LICENSE-matchMedia-polyfill.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-minlog.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-modernizr.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-mustache.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-netlib.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-paranamer.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-pmml-model.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-protobuf.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-py4j.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-pyrolite.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-re2j.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-reflectasm.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-respond.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-sbt-launch-lib.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-scala.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-scopt.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-slf4j.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-sorttable.js.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-spire.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-vis-timeline.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-xmlenc.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-zstd-jni.txt
spark-3.3.1-bin-hadoop3/licenses/LICENSE-zstd.txt
spark-3.3.1-bin-hadoop3/python/
spark-3.3.1-bin-hadoop3/python/.coveragerc
spark-3.3.1-bin-hadoop3/python/.gitignore
spark-3.3.1-bin-hadoop3/python/MANIFEST.in
spark-3.3.1-bin-hadoop3/python/README.md
spark-3.3.1-bin-hadoop3/python/dist/
spark-3.3.1-bin-hadoop3/python/docs/
spark-3.3.1-bin-hadoop3/python/docs/Makefile
spark-3.3.1-bin-hadoop3/python/docs/make.bat
spark-3.3.1-bin-hadoop3/python/docs/make2.bat
spark-3.3.1-bin-hadoop3/python/docs/source/
spark-3.3.1-bin-hadoop3/python/docs/source/_static/
spark-3.3.1-bin-hadoop3/python/docs/source/_static/copybutton.js
spark-3.3.1-bin-hadoop3/python/docs/source/_static/css/
spark-3.3.1-bin-hadoop3/python/docs/source/_static/css/pyspark.css
spark-3.3.1-bin-hadoop3/python/docs/source/_templates/
spark-3.3.1-bin-hadoop3/python/docs/source/_templates/autosummary/
spark-3.3.1-bin-hadoop3/python/docs/source/_templates/autosummary/class.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/_templates/autosummary/class_with_docs.rst
spark-3.3.1-bin-hadoop3/python/docs/source/conf.py
spark-3.3.1-bin-hadoop3/python/docs/source/development/
spark-3.3.1-bin-hadoop3/python/docs/source/development/contributing.rst
spark-3.3.1-bin-hadoop3/python/docs/source/development/debugging.rst
spark-3.3.1-bin-hadoop3/python/docs/source/development/index.rst
```

```
spark-3.3.1-bin-hadoop3/python/docs/source/development/setting_ide.rst
spark-3.3.1-bin-hadoop3/python/docs/source/development/testing.rst
spark-3.3.1-bin-hadoop3/python/docs/source/getting_started/
spark-3.3.1-bin-hadoop3/python/docs/source/getting_started/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/getting_started/install.rst
spark-3.3.1-bin-hadoop3/python/docs/source/getting_started/quickstart_df.ipynb
spark-3.3.1-bin-hadoop3/python/docs/source/getting_started/quickstart_ps.ipynb
spark-3.3.1-bin-hadoop3/python/docs/source/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/koalas_to_pyspark.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_1.0_1.2_to_1.3.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_1.4_to_1.5.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_2.2_to_2.3.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_2.3.0_to_2.3.1_above.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_2.3_to_2.4.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_2.4_to_3.0.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_3.1_to_3.2.rst
spark-3.3.1-bin-hadoop3/python/docs/source/migration_guide/pyspark_3.2_to_3.3.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/
spark-3.3.1-bin-hadoop3/python/docs/source/reference/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.ml.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.mllib.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/extensions.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/frame.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/general_functions.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/groupby.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/indexing.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/io.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/ml.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/series.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.pandas/window.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.resource.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/avro.rst
```

```
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/catalog.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/column.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/reference/pyspark.sql/configuration.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/reference/pyspark.sql/core_classes.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/data_types.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/dataframe.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/functions.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/grouping.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/io.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/observation.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/row.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/reference/pyspark.sql/spark_session.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.sql/window.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/core_classes.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.ss/io.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/reference/pyspark.ss/query_management.rst
spark-3.3.1-bin-hadoop3/python/docs/source/reference/pyspark.streaming.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/arrow_pandas.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/index.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/pandas_on_spark/
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/best_practices.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/pandas_on_spark/faq.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/from_to_dbms.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/pandas_on_spark/index.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/options.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/pandas_pyspark.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/supported_pandas_api.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/transform_apply.rst
spark-3.3.1-bin-
hadoop3/python/docs/source/user_guide/pandas_on_spark/typehints.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/pandas_on_spark/types.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/python_packaging.rst
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/sql/
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/sql/arrow_pandas.rst
```

```
spark-3.3.1-bin-hadoop3/python/docs/source/user_guide/sql/index.rst
spark-3.3.1-bin-hadoop3/python/lib/
spark-3.3.1-bin-hadoop3/python/lib/PY4J_LICENSE.txt
spark-3.3.1-bin-hadoop3/python/lib/py4j-0.10.9.5-src.zip
spark-3.3.1-bin-hadoop3/python/lib/pyspark.zip
spark-3.3.1-bin-hadoop3/python/mypy.ini
spark-3.3.1-bin-hadoop3/python/pyspark/
spark-3.3.1-bin-hadoop3/python/pyspark/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/__pycache__/
spark-3.3.1-bin-hadoop3/python/pyspark/__pycache__/install.cpython-38.pyc
spark-3.3.1-bin-hadoop3/python/pyspark/_globals.py
spark-3.3.1-bin-hadoop3/python/pyspark/_typing.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/accumulators.py
spark-3.3.1-bin-hadoop3/python/pyspark/broadcast.py
spark-3.3.1-bin-hadoop3/python/pyspark/cloudpickle/
spark-3.3.1-bin-hadoop3/python/pyspark/cloudpickle/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/cloudpickle/cloudpickle.py
spark-3.3.1-bin-hadoop3/python/pyspark/cloudpickle/cloudpickle_fast.py
spark-3.3.1-bin-hadoop3/python/pyspark/cloudpickle/compat.py
spark-3.3.1-bin-hadoop3/python/pyspark/conf.py
spark-3.3.1-bin-hadoop3/python/pyspark/context.py
spark-3.3.1-bin-hadoop3/python/pyspark/daemon.py
spark-3.3.1-bin-hadoop3/python/pyspark/files.py
spark-3.3.1-bin-hadoop3/python/pyspark/find_spark_home.py
spark-3.3.1-bin-hadoop3/python/pyspark/install.py
spark-3.3.1-bin-hadoop3/python/pyspark/instrumentation_utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/java_gateway.py
spark-3.3.1-bin-hadoop3/python/pyspark/join.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/
spark-3.3.1-bin-hadoop3/python/pyspark/ml/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/_typing.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/ml/base.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/classification.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/clustering.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/common.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/evaluation.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/feature.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/fpm.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/image.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/linalg/
spark-3.3.1-bin-hadoop3/python/pyspark/ml/linalg/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/param/
spark-3.3.1-bin-hadoop3/python/pyspark/ml/param/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/param/_shared_params_code_gen.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/param/shared.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/pipeline.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/recommendation.py
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/ml/regression.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/stat.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_algorithms.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_base.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_evaluation.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_feature.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_image.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_linalg.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_param.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_persistence.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_pipeline.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_stat.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_training_summary.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_tuning.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_util.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/test_wrapper.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_classification.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_clustering.yaml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_evaluation.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_feature.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_param.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_readable.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tests/typing/test_regression.yml
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tree.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/tuning.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/util.py
spark-3.3.1-bin-hadoop3/python/pyspark/ml/wrapper.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/_typing.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/classification.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/clustering.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/common.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/evaluation.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/feature.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/fpm.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/linalg/
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/linalg/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/linalg/distributed.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/random.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/random.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/recommendation.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/recommendation.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/regression.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/KernelDensity.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/_statistics.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/distribution.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/stat/test.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_algorithms.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_feature.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_linalg.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_stat.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_streaming_algorithms.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tests/test_util.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/tree.py
spark-3.3.1-bin-hadoop3/python/pyspark/mllib/util.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/_typing.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/accessors.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/base.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/categorical.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/config.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/base.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/binary_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/boolean_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/categorical_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/complex_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/date_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/datetime_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/null_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/num_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/string_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/timedelta_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/data_type_ops/udt_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/datetimes.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/exceptions.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/extensions.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/frame.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/generic.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/groupby.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/base.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/category.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/datetimes.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/multi.py
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/numeric.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexes/timedelta.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/indexing.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/internal.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/common.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/frame.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/general_functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/groupby.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/indexes.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/series.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/missing/window.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/ml.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/mlflow.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/namespace.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/numpy_compat.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/plot/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/plot/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/plot/core.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/plot/matplotlib.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/plot/plotly.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/series.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/spark/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/spark/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/spark/accessors.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/spark/functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/spark/utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/sql_formatter.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/sql_processor.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/strings.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_base.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_binary_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_boolean_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_categorical_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_complex_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_date_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/data_type_ops/test_datetime_ops.py
```

```
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/test_null_ops.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/test_num_ops.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/test_string_ops.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/test_timedelta_ops.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/test_udt_ops.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/data_type_ops/testing_utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test_base.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test_category.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test_datetime.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/indexes/test_timedelta.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/plot/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/plot/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test_frame_plot.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/plot/test_frame_plot_matplotlib.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/plot/test_frame_plot_plotly.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/plot/test_series_plot.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/plot/test_series_plot_matplotlib.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/plot/test_series_plot_plotly.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_categorical.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_config.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_csv.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_dataframe.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_dataframe_conversion.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_dataframe_spark_io.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_default_index.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_expanding.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_extension.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_frame_spark.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_groupby.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_indexing.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_indexops_spark.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_internal.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_namespace.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_numpy_compat.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_ops_on_diff_frames.py
spark-3.3.1-bin-
```

```
hadoop3/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby_expanding.py
spark-3.3.1-bin-
hadoop3/python/pyspark/pandas/tests/test_ops_on_diff_frames_groupby_rolling.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_repr.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_reshape.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_rolling.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_series.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_series_conversion.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_series_datetime.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_series_string.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_spark_functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_sql.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_stats.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_typedef.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/tests/test_window.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/typedef/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/typedef/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/typedef/typehints.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/usage_logging/
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/usage_logging/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/usage_logging/usage_logger.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/pandas/window.py
spark-3.3.1-bin-hadoop3/python/pyspark/profiler.py
spark-3.3.1-bin-hadoop3/python/pyspark/py.typed
spark-3.3.1-bin-hadoop3/python/pyspark/python/
spark-3.3.1-bin-hadoop3/python/pyspark/python/pyspark/
spark-3.3.1-bin-hadoop3/python/pyspark/python/pyspark/shell.py
spark-3.3.1-bin-hadoop3/python/pyspark/rdd.py
spark-3.3.1-bin-hadoop3/python/pyspark/rddsampler.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/
spark-3.3.1-bin-hadoop3/python/pyspark/resource/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/information.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/profile.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/requests.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/resource/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/resource/tests/test_resources.py
spark-3.3.1-bin-hadoop3/python/pyspark/resultiterable.py
spark-3.3.1-bin-hadoop3/python/pyspark/serializers.py
spark-3.3.1-bin-hadoop3/python/pyspark/shell.py
spark-3.3.1-bin-hadoop3/python/pyspark/shuffle.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/_typing.pyi
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/sql/avro/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/avro/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/avro/functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/catalog.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/column.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/conf.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/context.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/dataframe.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/group.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/observation.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/__init__.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/protocols/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/protocols/__init__.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/protocols/frame.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/_typing/protocols/series.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/conversion.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/functions.pyi
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/group_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/map_ops.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/serializers.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/typehints.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/types.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/pandas/utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/readwriter.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/session.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/sql_formatter.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/streaming.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_arrow.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_arrow_map.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_catalog.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_column.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_conf.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_context.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_dataframe.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_datasources.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_functions.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_group.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_cogrouped_map.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_grouped_map.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_map.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf.py
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf_grouped_agg.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf_scalar.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf_typehints.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf_typehints_with_
future_annotations.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_pandas_udf_window.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_readwriter.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_serde.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_session.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_streaming.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_types.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_udf.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_udf_profiler.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/test_utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_column.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_dataframe.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_functions.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_readwriter.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_session.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/tests/typing/test_udf.yml
spark-3.3.1-bin-hadoop3/python/pyspark/sql/types.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/udf.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/sql/window.py
spark-3.3.1-bin-hadoop3/python/pyspark/statcounter.py
spark-3.3.1-bin-hadoop3/python/pyspark/status.py
spark-3.3.1-bin-hadoop3/python/pyspark/storagelevel.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/context.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/dstream.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/kinesis.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/listener.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/test_context.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/test_dstream.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/test_kinesis.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/tests/test_listener.py
spark-3.3.1-bin-hadoop3/python/pyspark/streaming/util.py
spark-3.3.1-bin-hadoop3/python/pyspark/taskcontext.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/
spark-3.3.1-bin-hadoop3/python/pyspark/testing/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/mllibutils.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/mlutils.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/pandasutils.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/sqlutils.py
```

```
spark-3.3.1-bin-hadoop3/python/pyspark/testing/streamingutils.py
spark-3.3.1-bin-hadoop3/python/pyspark/testing/utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/
spark-3.3.1-bin-hadoop3/python/pyspark/tests/__init__.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_appsubmit.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_broadcast.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_conf.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_context.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_daemon.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_install_spark.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_join.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_pin_thread.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_profiler.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_rdd.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_rddbarrier.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_readwrite.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_serializers.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_shuffle.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_statcounter.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_taskcontext.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_util.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/test_worker.py
spark-3.3.1-bin-hadoop3/python/pyspark/tests/typing/
spark-3.3.1-bin-hadoop3/python/pyspark/tests/typing/test_context.yml
spark-3.3.1-bin-hadoop3/python/pyspark/tests/typing/test_core.yml
spark-3.3.1-bin-hadoop3/python/pyspark/tests/typing/test_rdd.yml
spark-3.3.1-bin-hadoop3/python/pyspark/tests/typing/test_resultiterable.yml
spark-3.3.1-bin-hadoop3/python/pyspark/traceback_utils.py
spark-3.3.1-bin-hadoop3/python/pyspark/util.py
spark-3.3.1-bin-hadoop3/python/pyspark/version.py
spark-3.3.1-bin-hadoop3/python/pyspark/worker.py
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/PKG-INFO
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/SOURCES.txt
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/dependency_links.txt
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/requires.txt
spark-3.3.1-bin-hadoop3/python/pyspark.egg-info/top_level.txt
spark-3.3.1-bin-hadoop3/python/run-tests
spark-3.3.1-bin-hadoop3/python/run-tests-with-coverage
spark-3.3.1-bin-hadoop3/python/run-tests.py
spark-3.3.1-bin-hadoop3/python/setup.cfg
spark-3.3.1-bin-hadoop3/python/setup.py
spark-3.3.1-bin-hadoop3/python/test_coverage/
spark-3.3.1-bin-hadoop3/python/test_coverage/conf/
spark-3.3.1-bin-hadoop3/python/test_coverage/conf/spark-defaults.conf
spark-3.3.1-bin-hadoop3/python/test_coverage/coverage_daemon.py
spark-3.3.1-bin-hadoop3/python/test_coverage/sitecustomize.py
spark-3.3.1-bin-hadoop3/python/test_support/
```

```
spark-3.3.1-bin-hadoop3/python/test_support/SimpleHTTPServer.py
spark-3.3.1-bin-hadoop3/python/test_support/hello/
spark-3.3.1-bin-hadoop3/python/test_support/hello/hello.txt
spark-3.3.1-bin-hadoop3/python/test_support/hello/sub_hello/
spark-3.3.1-bin-hadoop3/python/test_support/hello/sub_hello/sub_hello.txt
spark-3.3.1-bin-hadoop3/python/test_support/sql/
spark-3.3.1-bin-hadoop3/python/test_support/sql/ages.csv
spark-3.3.1-bin-hadoop3/python/test_support/sql/ages_newlines.csv
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/_SUCCESS
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/c=0/
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/c=0/.part-r-
00000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=0/c=0/part-r-0
0000-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/.part-r-
00000-829af031-b970-49d6-ad39-30460a0be2c8.orc.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/orc_partitioned/b=1/c=1/part-r-0
0000-829af031-b970-49d6-ad39-30460a0be2c8.orc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/_SUCCESS
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/_common_metadata
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/_metadata
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2014/month=9/
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2014/month=9/day=1/
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mo
nth=9/day=1/.part-r-00008.gz.parquet.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2014/mo
nth=9/day=1/part-r-00008.gz.parquet
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2015/month=10/
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=25/
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=10/day=25/.part-r-00002.gz.parquet.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=10/day=25/.part-r-00004.gz.parquet.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=10/day=25/part-r-00002.gz.parquet
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
```

```
nth=10/day=25/part-r-00004.gz.parquet
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2015/month=10/day=26/
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=10/day=26/.part-r-00005.gz.parquet.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=10/day=26/part-r-00005.gz.parquet
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2015/month=9/
spark-3.3.1-bin-
hadoop3/python/test_support/sql/parquet_partitioned/year=2015/month=9/day=1/
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=9/day=1/.part-r-00007.gz.parquet.crc
spark-3.3.1-bin-hadoop3/python/test_support/sql/parquet_partitioned/year=2015/mo
nth=9/day=1/part-r-00007.gz.parquet
spark-3.3.1-bin-hadoop3/python/test_support/sql/people.json
spark-3.3.1-bin-hadoop3/python/test_support/sql/people1.json
spark-3.3.1-bin-hadoop3/python/test_support/sql/people_array.json
spark-3.3.1-bin-hadoop3/python/test_support/sql/people_array_utf16le.json
spark-3.3.1-bin-hadoop3/python/test_support/sql/streaming/
spark-3.3.1-bin-hadoop3/python/test_support/sql/streaming/text-test.txt
spark-3.3.1-bin-hadoop3/python/test_support/sql/text-test.txt
spark-3.3.1-bin-hadoop3/python/test_support/userlib-0.1.zip
spark-3.3.1-bin-hadoop3/python/test_support/userlibrary.py
spark-3.3.1-bin-hadoop3/sbin/
spark-3.3.1-bin-hadoop3/sbin/decommission-slave.sh
spark-3.3.1-bin-hadoop3/sbin/decommission-worker.sh
spark-3.3.1-bin-hadoop3/sbin/slaves.sh
spark-3.3.1-bin-hadoop3/sbin/spark-config.sh
spark-3.3.1-bin-hadoop3/sbin/spark-daemon.sh
spark-3.3.1-bin-hadoop3/sbin/spark-daemons.sh
spark-3.3.1-bin-hadoop3/sbin/start-all.sh
spark-3.3.1-bin-hadoop3/sbin/start-history-server.sh
spark-3.3.1-bin-hadoop3/sbin/start-master.sh
spark-3.3.1-bin-hadoop3/sbin/start-mesos-dispatcher.sh
spark-3.3.1-bin-hadoop3/sbin/start-mesos-shuffle-service.sh
spark-3.3.1-bin-hadoop3/sbin/start-slave.sh
spark-3.3.1-bin-hadoop3/sbin/start-slaves.sh
spark-3.3.1-bin-hadoop3/sbin/start-thriftserver.sh
spark-3.3.1-bin-hadoop3/sbin/start-worker.sh
spark-3.3.1-bin-hadoop3/sbin/start-workers.sh
spark-3.3.1-bin-hadoop3/sbin/stop-all.sh
spark-3.3.1-bin-hadoop3/sbin/stop-history-server.sh
spark-3.3.1-bin-hadoop3/sbin/stop-master.sh
spark-3.3.1-bin-hadoop3/sbin/stop-mesos-dispatcher.sh
spark-3.3.1-bin-hadoop3/sbin/stop-mesos-shuffle-service.sh
spark-3.3.1-bin-hadoop3/sbin/stop-slave.sh
spark-3.3.1-bin-hadoop3/sbin/stop-slaves.sh
```

```
spark-3.3.1-bin-hadoop3/sbin/stop-thriftserver.sh
spark-3.3.1-bin-hadoop3/sbin/stop-worker.sh
spark-3.3.1-bin-hadoop3/sbin/stop-workers.sh
spark-3.3.1-bin-hadoop3/sbin/workers.sh
spark-3.3.1-bin-hadoop3/yarn/
spark-3.3.1-bin-hadoop3/yarn/spark-3.3.1-yarn-shuffle.jar
```

[60]:
```python
import os
os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
os.environ["SPARK_HOME"] = "/content/spark-3.3.1-bin-hadoop3"
```

1.3 How to initialize Spark

[61]:
```python
from pyspark.sql import SparkSession

APP_NAME = "CAP-lab3"
SPARK_URL = "local[*]"
spark = SparkSession.builder.appName(APP_NAME).master(SPARK_URL).getOrCreate()
sc = spark.sparkContext
```

2 First Part: RDDs

2.1 Basic operations

2.1.1 Parallelize & collect

It creates a RDD out of a list or array. Second argument indicates the number of pieces of the RDD

[62]:
```python
array = sc.parallelize([1,2,3,4,5,6,7,8,9,10,1], 2)
array
```

[62]:
```
ParallelCollectionRDD[138] at readRDDFromFile at PythonRDD.scala:274
```

[63]:
```python
import numpy as np
randomSamples = sc.parallelize(np.random.randn(100))
randomSamples
```

[63]:
```
ParallelCollectionRDD[139] at readRDDFromFile at PythonRDD.scala:274
```

Cool, RDDs can not be printed…

Of course, RDDs can not be printed unless they are reduced

[64]:
```
print(array.collect())
print(randomSamples.collect())
```

```
[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 1]
[-0.3027062236855682, 0.8497831254788817, 0.09673139429774069,
1.4486033134932121, 0.5828954795420782, -1.0478346281975415, 0.261767932896776l,
-1.1595333716013794, -0.6336456219904765, 1.16259652682523, -1.985481680973185,
-0.2640986191401976, -0.22105764165938327, -0.739614522636875,
0.8020039518750812, 1.458928223172202, 0.26495412202300656, 0.384838311982052l,
1.731270038748271, 2.0094155841840897, -0.5622799007667083, 1.4864952988142057,
0.6924581758483126, -0.642485394024644, 0.31144282696463993,
-0.9711242313676898, -0.7205011376271563, -0.797219445776185,
0.3542311532168802, 0.15241286590188843, -1.58206065516638, -1.2072504042571452,
0.44334327976335985, 0.1698368540835224, -0.12292140323970031,
-1.3737921177116919, -0.6768578772415462, 0.2514626961817422,
-1.0475308238105037, -1.0312168483462714, -1.5160953472412126,
-0.42173094620116175, 1.2084701490775913, -2.0378565736640883,
-0.19655563212740776, -0.7974080817585756, -0.9139066890377837,
1.1776448531532107, -1.175099221187751, -0.6446836478857331,
-0.04907370667287337, 0.2862424962482331, -0.570033389347673,
0.7836576730434606, 1.1919672651028799, -0.4694175158743068,
-1.0329906023202093, 0.553730711979216, -0.36861966041134325,
0.6260035859464578, -0.6935673064001514, 0.3220043517292291, 0.6907565281692103,
0.5238444338757086, 0.826805437054279, 2.3248540907997057, 0.9956580219023234,
-1.0555479569224053, -0.6137908716176694, -1.503939825742022,
-1.1524257128834112, -1.586849001047177, -0.694139528395303,
-0.39569744836536397, -0.8552936803715557, -0.3882170513098532,
0.8542936798832164, 0.4707589832451836, -2.4131272136750956,
0.08894705113928336, 0.2454181347212917, 0.9494809379274589,
-0.5346888028558663, 0.935471690399495, -1.2880996064253882, 0.9453701894488286,
-0.20443886213480375, 0.5194134426628358, -0.5551991063393502,
-0.16111478703708795, -1.0913309487860883, -1.6331561968333206,
-0.507974896548668, 0.961838488475515, 1.03305419750189, -0.7603360409278795,
0.8493996714616835, -0.6372880046593861, 0.8282624878923938,
-0.7927802450360271]
```

Spark uses lazy operations for everything, this means that nothing is evaluated until an action, a reduce operation normally, is performed. The basic reduce operation is collect, which returns the whole RDD (i.e. no reduction is performed).

2.1.2 Other ways of loading data

[65]:
```
import requests

request = requests.get("https://gist.githubusercontent.com/jsdario/
 ↪6d6c69398cb0c73111e49f1218960f79/raw/
 ↪8d4fc4548d437e2a7203a5aeeace5477f598827d/el_quijote.txt")
```

```
with open("elquijote.txt", "wb") as f:
  f.write(request.content)
```

[66]:
```
quijote = sc.textFile("elquijote.txt")
quijote.take(10)
```

[66]:
```
['DON QUIJOTE DE LA MANCHA',
 'Miguel de Cervantes Saavedra',
 '',
 'PRIMERA PARTE',
 'CAPÍTULO 1: Que trata de la condición y ejercicio del famoso hidalgo D.
Quijote de la Mancha',
 'En un lugar de la Mancha, de cuyo nombre no quiero acordarme, no ha mucho
tiempo que vivía un hidalgo de los de lanza en astillero, adarga antigua,
rocín flaco y galgo corredor. Una olla de algo más vaca que carnero, salpicón
las más noches, duelos y quebrantos los sábados, lentejas los viernes, algún
palomino de añadidura los domingos, consumían las tres partes de su hacienda.
El resto della concluían sayo de velarte, calzas de velludo para las fiestas
con sus pantuflos de lo mismo, los días de entre semana se honraba con su
vellori de lo más fino. Tenía en su casa una ama que pasaba de los cuarenta, y
una sobrina que no llegaba a los veinte, y un mozo de campo y plaza, que así
ensillaba el rocín como tomaba la podadera. Frisaba la edad de nuestro hidalgo
con los cincuenta años, era de complexión recia, seco de carnes, enjuto de
rostro; gran madrugador y amigo de la caza. Quieren decir que tenía el
sobrenombre de Quijada o Quesada (que en esto hay alguna diferencia en los
autores que deste caso escriben), aunque por conjeturas verosímiles se deja
entender que se llama Quijana; pero esto importa poco a nuestro cuento; basta
que en la narración dél no se salga un punto de la verdad. Es, pues, de saber,
que este sobredicho hidalgo, los ratos que estaba ocioso (que eran los más del
año) se daba a leer libros de caballerías con tanta afición y gusto, que
olvidó casi de todo punto el ejercicio de la caza, y aun la administración de
su hacienda; y llegó a tanto su curiosidad y desatino en esto, que vendió
muchas hanegas de tierra de sembradura, para comprar libros de caballerías en
que leer; y así llevó a su casa todos cuantos pudo haber dellos; y de todos
ningunos le parecían tan bien como los que compuso el famoso Feliciano de
Silva: porque la claridad de su prosa, y aquellas intrincadas razones suyas, le
parecían de perlas; y más cuando llegaba a leer aquellos requiebros y cartas
de desafío, donde en muchas partes hallaba escrito: la razón de la sinrazón
que a mi razón se hace, de tal manera mi razón enflaquece, que con razón me
quejo de la vuestra fermosura, y también cuando leía: los altos cielos que de
vuestra divinidad divinamente con las estrellas se fortifican, y os hacen
merecedora del merecimiento que merece la vuestra grandeza. Con estas y
semejantes razones perdía el pobre caballero el juicio, y desvelábase por
entenderlas, y desentrañarles el sentido, que no se lo sacara, ni las
entendiera el mismo Aristóteles, si resucitara para sólo ello. No estaba muy
bien con las heridas que don Belianis daba y recibía, porque se imaginaba que
por grandes maestros que le hubiesen curado, no dejaría de tener el rostro y
todo el cuerpo lleno de cicatrices y señales; pero con todo alababa en su autor
aquel acabar su libro con la promesa de aquella inacabable aventura, y muchas
veces le vino deseo de tomar la pluma, y darle fin al pie de la letra como allí
se promete; y sin duda alguna lo hiciera, y aun saliera con ello, si otros
mayores y continuos pensamientos no se lo estorbaran.',
 'Tuvo muchas veces competencia con el cura de su lugar (que era hombre docto
graduado en Sigüenza), sobre cuál había sido mejor caballero, Palmerín de
Inglaterra o Amadís de Gaula; mas maese Nicolás, barbero del mismo pueblo,
decía que ninguno llegaba al caballero del Febo, y que si alguno se le podía
comparar, era don Galaor, hermano de Amadís de Gaula, porque tenía muy
acomodada condición para todo; que no era caballero melindroso, ni tan llorón
como su hermano, y que en lo de la valentía no le iba en zaga.',
 'En resolución, él se enfrascó tanto en su lectura, que se le pasaban las
noches leyendo de claro en claro, y los días de turbio en turbio, y así, del
poco dormir y del mucho leer, se le secó el cerebro, de manera que vino a
perder el juicio. Llenósele la fantasía de todo aquello que leía en los
libros, así de encantamientos, como de pendencias, batallas, desafíos,
heridas, requiebros, amores, tormentas y disparates imposibles, y asentósele de
tal modo en la imaginación que era verdad toda aquella máquina de aquellas
soñadas invenciones que leía, que para él no había otra',
 'historia más cierta en el mundo.',
 'Decía él, que el Cid Ruy Díaz había sido muy buen caballero; pero que no
tenía que ver con el caballero de la ardiente espada, que de sólo un revés
había partido por medio dos fieros y descomunales gigantes. Mejor estaba con
Bernardo del Carpio, porque en Roncesvalle había muerto a Roldán el encantado,
valiéndose de la industria de Hércules, cuando ahogó a Anteo, el hijo de la
Tierra, entre los brazos. Decía mucho bien del gigante Morgante, porque con ser
de aquella generación gigantesca, que todos son soberbios y descomedidos, él
solo era afable y bien criado; pero sobre todos estaba bien con Reinaldos de
Montalbán, y más cuando le veía salir de su castillo y robar cuantos topaba,
y cuando en Allende robó aquel ídolo de Mahoma, que era todo de oro, según
dice su historia. Diera él, por dar una mano de coces al traidor de Galalón,
al ama que tenía y aun a su sobrina de añadidura.']
```

Here, you can see both a method to load a text file line per line and a another reduction operation.

[67]:

```python
quijote.take?
```

2.1.3 Transformations

Let's review all the transformation that can be performed to data.

[68]:

```python
charsPerLine = quijote.map(lambda s: len(s))
allWords = quijote.flatMap(lambda s: s.split())
allWordsNoArticles = allWords.filter(lambda a: a.lower() not in ["el", "la"])
allWordsUnique = allWords.map(lambda s: s.lower()).distinct()
sampleWords = allWords.sample(withReplacement=True, fraction=0.2, seed=666)
```

```
weirdSampling = sampleWords.union(allWordsNoArticles.sample(False, fraction=0.
 ↪3))
```

```
[69]: weirdSampling.take(5)
```

```
[69]: ['DON', 'Que', 'ejercicio', 'del', 'D.']
```

Assignment question

Explain the use and purpose of each operation above.

Comment also on the size of the resulting RDD in terms of the size of the original RDD, e.g. if original RDD is of size N, then rdd.filter() is of size $K \leq N$

Answer: - map -> Transforma cada item del array, retornando un solo elemento. $N => N$ - flatMap => Transforma cada item del array, retornando uno o varios elementos transformados. $N => kN$ - filter -> Filtra los elementos del array, usando una función de condición. $N => N - k$ - distinct -> Elimina los items repetidos en el array. $N => N/k$ - sample -> Devuelve items aleatorios del array. $N => k$ - union -> Une los items de dos arrays evitando la repetición. $N, M => N + M$

Actions

```
[70]: numLines = quijote.count()
      numChars = charsPerLine.reduce(lambda a,b: a+b) # also charsPerLine.sum()
      sortedWordsByLength = allWordsNoArticles.takeOrdered(10, key=lambda x: -len(x))
      numLines, numChars, sortedWordsByLength
```

```
[70]: (2186,
       1036211,
       ['procuremos.Levántate,',
        'extraordinariamente,',
        'estrechísimamente,',
        'convirtiéndoseles',
        'entretenimientos,',
        'inadvertidamente.',
        'cortesísimamente',
        'Agredeciéronselo',
        'Pintiquiniestra,',
        'entretenimiento,'])
```

Assignment question

Explain the use and purpose of each action above.

Implement the count operation using reduce as the unique option. You can use transformations. Is it possible to achieve a solution without any transformation? Does it make sense?

Answer: - count: Cuenta el número de entradas que tiene el RDD. - reduce: Ejecuta una función dada, por pares de entradas hasta reducirlo a un solo dato. - takeOrdered: Ordena en base a la condición dada y devuelve la cantidad de entradas pedidas.

```
[71]: numLines = quijote.map(lambda line: 1).reduce(lambda a,b: a+b)
      numLines
```

```
[71]: 2186
```

No es posible conseguir el mismo efecto sin transformaciones, pues la función *reduce* simplemente agrupa los valores que tiene el RDD. No tiene sentido si no transformamos primero cada línea a un 1, pues *reduce* no podrá ”contar” el número de líneas, sino que sumará (concatenará) las strings de cada entrada.

2.2 Key-Value RDDs

```
[72]: import re
      allWords = allWords.flatMap(lambda w: re.sub(""";|:|\.|,|-|-|"|'|\s"""," ", w.
       ↪lower()).split(" ")).filter(lambda a: len(a)>0)
      allWords2 = sc.parallelize(requests.get("https://gist.githubusercontent.com/
       ↪jsdario/9d871ed773c81bf217f57d1db2d2503f/raw/
       ↪585de69b0631c805dabc6280506717943b82ba4a/el_quijote_ii.txt").iter_lines())
      allWords2 = allWords2.flatMap(lambda w: re.sub(""";|:|\.|,|-|-|"|'|\s"""," ", w.
       ↪decode("utf8").lower()).split(" ")).filter(lambda a: len(a)>0)
```

```
[73]: allWords.take(10), allWords2.take(10)
```

```
[73]: (['don',
        'quijote',
        'de',
        'la',
        'mancha',
        'miguel',
        'de',
        'cervantes',
        'saavedra',
        'primera'],
       ['don',
        'quijote',
        'de',
        'la',
        'mancha',
        'miguel',
        'de',
        'cervantes',
        'saavedra',
        'segunda'])
```

Next, we move to more interesting operations that involve key-value RDDs. Key-value RDDs are a special kind of RDDs where each is element is a tuple (K,V) where K is the key and V the value.

```
[74]: words = allWords.map(lambda e: (e,1))
      words2 = allWords2.map(lambda e: (e,1))

      words.take(10), words2.take(10)
```

```
[74]: ([('don', 1),
        ('quijote', 1),
        ('de', 1),
        ('la', 1),
        ('mancha', 1),
        ('miguel', 1),
        ('de', 1),
        ('cervantes', 1),
        ('saavedra', 1),
        ('primera', 1)],
       [('don', 1),
        ('quijote', 1),
        ('de', 1),
        ('la', 1),
        ('mancha', 1),
        ('miguel', 1),
        ('de', 1),
        ('cervantes', 1),
        ('saavedra', 1),
        ('segunda', 1)])
```

2.2.1 How to manipulate K-V RDDs

```
[75]: frequencies = words.reduceByKey(lambda a,b: a+b)
      frequencies2 = words2.reduceByKey(lambda a,b: a+b)
      frequencies.takeOrdered(10, key=lambda a: -a[1])
```

```
[75]: [('que', 10705),
       ('de', 9033),
       ('y', 8668),
       ('la', 5015),
       ('a', 4815),
       ('en', 4046),
       ('el', 3857),
       ('no', 3083),
       ('se', 2382),
       ('los', 2148)]
```

[76]:

```
res = words.groupByKey().takeOrdered(10, key=lambda a: -len(a)) # -len(a[0])␣
 ↪longitud palabra y -len(a[1]) por frecuencia
res # To see the content, res[i][1].data
for i in range(10):
  print(res[i][0])
```

```
don
mancha
saavedra
primera
parte
1
que
condición
y
del
```

[77]:

```
joinFreq = frequencies.join(frequencies2)
joinFreq.take(10)
```

[77]:

```
[('don', (1072, 1606)),
 ('mancha', (50, 101)),
 ('saavedra', (2, 1)),
 ('primera', (39, 55)),
 ('parte', (178, 158)),
 ('1', (1, 1)),
 ('que', (10705, 10040)),
 ('condición', (33, 39)),
 ('y', (8668, 9650)),
 ('del', (1128, 1344))]
```

[78]:

```
joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1]))).
 ↪takeOrdered(10, lambda v: -v[1]), joinFreq.map(lambda e: (e[0], (e[1][0] -␣
 ↪e[1][1])/(e[1][0] + e[1][1]))).takeOrdered(10, lambda v: +v[1])
```

[78]:

```
([('bacía', 0.9393939393939394),
  ('venia', 0.9230769230769231),
  ('hermandad', 0.9),
  ('andrés', 0.8823529411764706),
  ('peña', 0.8823529411764706),
  ('micomicona', 0.8823529411764706),
  ('barca', 0.875),
  ('novela', 0.875),
  ('yerme', 0.875),
  ('acertó', 0.8666666666666667)],
 [('teresa', -0.9767441860465116),
  ('roque', -0.96),
  ('refranes', -0.9375),
  ('condesa', -0.9333333333333333),
  ('leones', -0.9333333333333333),
  ('gobernadores', -0.9166666666666666),
  ('lacayo', -0.9166666666666666),
  ('visorrey', -0.9130434782608695),
  ('antonio', -0.9076923076923077),
  ('zaragoza', -0.9047619047619048)])
```

Assignment question

Explain the use and purpose of each action above.

Implement the frequency with groupByKey and transformations.

Which of the two following cells is more efficient?

Answer: - reduceByKey: Ejecuta por cada clave una función dada, sumando/concatenando pares de entradas hasta reducirlo a un solo dato. - groupByKey: Agrupa/Concatena los valores según la clave. - join: Genera tuplas con los valores por clave de dos RDDs.

[79]:

```
words.groupByKey().map(lambda s: (s[0], len(s[1]))).take(10) #, frequencies.
↪take(10)
```

[79]:

```
[('don', 1072),
 ('mancha', 50),
 ('saavedra', 2),
 ('primera', 39),
 ('parte', 178),
 ('1', 1),
 ('que', 10705),
 ('condición', 33),
 ('y', 8668),
 ('del', 1128)]
```

La segunda celda es la más eficiente puesto que no tiene que usar la función *map* dos veces; sino que lo hace una vez y luego ordena y coge los valores que necesita.

---

[80]:

```
joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1]))).takeOrdered(10, lambda v: -v[1]), joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1]))).takeOrdered(10, lambda v: +v[1])
```

[80]:

```
([('bacía', 0.9393939393939394),
  ('venia', 0.9230769230769231),
  ('hermandad', 0.9),
  ('andrés', 0.8823529411764706),
  ('peña', 0.8823529411764706),
  ('micomicona', 0.8823529411764706),
  ('barca', 0.875),
  ('novela', 0.875),
  ('yerme', 0.875),
  ('acertó', 0.8666666666666667)],
 [('teresa', -0.9767441860465116),
  ('roque', -0.96),
  ('refranes', -0.9375),
  ('condesa', -0.9333333333333333),
  ('leones', -0.9333333333333333),
  ('gobernadores', -0.9166666666666666),
  ('lacayo', -0.9166666666666666),
  ('visorrey', -0.9130434782608695),
  ('antonio', -0.9076923076923077),
  ('zaragoza', -0.9047619047619048)])
```

[81]:

```
result = joinFreq.map(lambda e: (e[0], (e[1][0] - e[1][1])/(e[1][0] + e[1][1])))
result.takeOrdered(10, lambda v: -v[1]), result.takeOrdered(10, lambda v: +v[1])
```

[81]:

```
([('bacía', 0.9393939393939394),
  ('venia', 0.9230769230769231),
  ('hermandad', 0.9),
  ('andrés', 0.8823529411764706),
  ('peña', 0.8823529411764706),
  ('micomicona', 0.8823529411764706),
  ('barca', 0.875),
  ('novela', 0.875),
  ('yerme', 0.875),
  ('acertó', 0.8666666666666667)],
 [('teresa', -0.9767441860465116),
  ('roque', -0.96),
  ('refranes', -0.9375),
  ('condesa', -0.9333333333333333),
  ('leones', -0.9333333333333333),
  ('gobernadores', -0.9166666666666666),
  ('lacayo', -0.9166666666666666),
  ('visorrey', -0.9130434782608695),
  ('antonio', -0.9076923076923077),
  ('zaragoza', -0.9047619047619048)])
```

2.3 Optimizations and final notes

2.3.1 Optimizing the data movement around the cluster

One of the main issues could be that if data after an operation is not balanced, we may not be using the cluster properly. For that purpose, we have two operations

[82]:

```
result.coalesce(numPartitions=2) # Avoids the data movement, so it tries to␣
 ↪balance inside each machine
# los accesos a memoria son consecutivos
result.repartition(numPartitions=2) # We don't care about data movement, this␣
 ↪balance the whole thing to ensure all machines are used
```

[82]:

```
MapPartitionsRDD[195] at coalesce at NativeMethodAccessorImpl.java:0
```

2.3.2 Persistance for intermediate operations

In contrast to Hadoop, intermediate RDDs are not preserved, each time we use an action/reduction, the whole data pipeline is executed from the datasources. To avoid this:

[83]:

```
result.take(10)
allWords.cache() # allWords RDD must stay in memory after computation, we made␣
 ↪a checkpoint (well, it's a best effort, so must might be too strong)
result.take(10)
```

[83]:

```
[('don', -0.19940253920836445),
 ('mancha', -0.337748344370860093),
 ('saavedra', 0.3333333333333333),
 ('primera', -0.1702127659574468),
 ('parte', 0.05952380952380952),
 ('1', 0.0),
 ('que', 0.03205591708845505),
 ('condición', -0.08333333333333333),
 ('y', -0.05360847254067038),
 ('del', -0.08737864077669903)]
```

[84]:

```
from pyspark import StorageLevel
# https://spark.apache.org/docs/2.2.0/rdd-programming-guide.html#rdd-persistence
allWords2.persist(StorageLevel.MEMORY_AND_DISK) # Now it will be preserved on␣
 ↪disk also
```

[84]:

```
PythonRDD[199] at RDD at PythonRDD.scala:53
```

[85]:

```
!rm -rf palabras_parte2
allWords2.saveAsTextFile("palabras_parte2")
```

[86]:

```
!ls palabras_parte2
```

```
part-00000  part-00001  _SUCCESS
```

Assignment question

Before saving with saveAsTextFile, use coalesce with different values. What's the difference in the previous ls?

Answer: Con numPartitions = 2 el resultado del ls es el siguiente: part-00000 part-00001 _SUCCESS Probando con con numPartitions entre 2 y 10 el resultado sobrenido del ls es el siguiente: part-00000 part-00001 _SUCCESS el resultado del ls es el mismo

---

2.3.3 Global variables

There are two kind of global variables, read-only and write-only.

[87]:

```
articles = sc.broadcast(["el", "la"])
articles.value
```

[87]:

```
['el', 'la']
```

Broadcast variables are read-only. They help us to avoid local variables of the closures (the functions we use inside map, reduce, ...) to be transferred in every single Spark operation. In that way, they are only transferred only once.

[88]:

```
acc = sc.accumulator(0)
def incrementar(x):
  global acc
  acc += x

allWords.map(lambda l:1).foreach(incrementar)
acc
```

[88]:

```
Accumulator<id=1, value=187045>
```

Write-only variables can be also declared and initialized, but they can not be read since reading will force a complete synchronization of the cluster.

3 Second part: Spark SQL

Next, we will do a short review of the high-level API of Spark

[89]:

```
import pandas as pd

size = int(1e6)
def loadRedditToPandas(subreddit=None, size=size):
  if subreddit is not None:
    redditData = requests.get(f"https://api.pushshift.io/reddit/search/
 ↪submission/?subreddit={subreddit}&sort=desc&sort_type=created_utc&size={size:
 ↪d}").json()
  else:
    redditData = requests.get(f"https://api.pushshift.io/reddit/search/
 ↪submission/?sort=desc&sort_type=created_utc&size={size:d}").json()
  return pd.DataFrame(redditData["data"])
```

[90]:

```
pdf = loadRedditToPandas()
pdf.head(10)
```

[90]:

```
  all_awardings  allow_live_comments          author author_flair_css_class  \
0            []                False         Smoke01                   None
1            []                False  freiburger1884                   None
2            []                 True   Amirali_toxic                   None
3            []                False   AutoModerator                   None
4            []                False   AutoModerator                   None
5            []                False   AutoModerator                   None
6            []                False        DlngoLex                   None
7            []                False   AutoModerator                   None
8            []                False        SiiZzL4c                   None
9            []                False    ReveErotique                   None

  author_flair_richtext author_flair_text author_flair_type author_fullname  \
0                    []              None              text     t2_uccs05ie
1                    []              None              text     t2_6zy5hbut
2                    []              None              text     t2_t1k1kp5n
3                    []              None              text        t2_6l4z3
4                    []              None              text        t2_6l4z3
5                    []              None              text        t2_6l4z3
6                    []              None              text        t2_zg6ay
7                    []              None              text        t2_6l4z3
8                    []              None              text     t2_bn3o7f1x
9                    []              None              text     t2_j8i6y045

   author_is_blocked  author_patreon_flair  ...  pwls whitelist_status  wls  \
0              False                 False  ...   NaN              NaN  NaN
1              False                 False  ...   NaN              NaN  NaN
```

```
2              False                 False  …   NaN              NaN  NaN
3              False                 False  …   6.0          all_ads  6.0
4              False                 False  …   0.0           no_ads  0.0
5              False                 False  …   0.0           no_ads  0.0
6              False                 False  …   NaN              NaN  NaN
7              False                 False  …   0.0           no_ads  0.0
8              False                 False  …   6.0          all_ads  6.0
9              False                 False  …   NaN              NaN  NaN

   distinguished  media_metadata author_flair_background_color  \
0            NaN             NaN                           NaN
1            NaN             NaN                           NaN
2            NaN             NaN                           NaN
3            NaN             NaN                           NaN
4            NaN             NaN                           NaN
5            NaN             NaN                           NaN
6            NaN             NaN                           NaN
7            NaN             NaN                           NaN
8            NaN             NaN                           NaN
9            NaN             NaN                           NaN

  author_flair_template_id author_flair_text_color gallery_data  is_gallery
0                      NaN                    NaN          NaN         NaN
1                      NaN                    NaN          NaN         NaN
2                      NaN                    NaN          NaN         NaN
3                      NaN                    NaN          NaN         NaN
4                      NaN                    NaN          NaN         NaN
5                      NaN                    NaN          NaN         NaN
6                      NaN                    NaN          NaN         NaN
7                      NaN                    NaN          NaN         NaN
8                      NaN                    NaN          NaN         NaN
9                      NaN                    NaN          NaN         NaN

[10 rows x 83 columns]
```

[91]:

```
pdf.selftext = pdf.selftext.apply(lambda e: str(e))
```

[92]:

```
attrs = ["author", "created_utc", "title", "subreddit", "selftext", "over_18"]
df = spark.createDataFrame(pdf[attrs])
```

3.1 Basic operations

[93]:

```
df.show()
```

```
+----------------+-----------+--------------------+-----------------+------------
--------+-------+
|         author|created_utc|               title|        subreddit|
selftext|over_18|
+---------------+-----------+--------------------+-----------------+------------
--------+-------+
|        Smoke01| 1669971610|Ariana Grande Fan…|            55Tra|
|  false|
| freiburger1884| 1669971610|I wish i had anot…| u_freiburger1884|
|   true|
|  Amirali_toxic| 1669971610|           iran news|     irannews2022|Iran  \nHave
peop…|   true|
|  AutoModerator| 1669971610|Icon Ecosystem We…|        helloicon|Greetings
and wel…|  false|
|  AutoModerator| 1669971610|Get Free ISK and …|      evereferral|
|  false|
|  AutoModerator| 1669971610|Daily Chat - Dece…|         TTC_PCOS|
|  false|
|       DlngoLex| 1669971610|             Shenise|            sheer|
|   true|
|  AutoModerator| 1669971610|   New Members Intro|    ProVSNatVSFun|If you’re
new to …|  false|
|       SiiZzL4c| 1669971610| Do you use discogs?|            vinyl|
|  false|
|   ReveErotique| 1669971610|      Amelia Braymut|       Bellissima|
|   true|
|  AutoModerator| 1669971610|Weekly FPL Analyt…|     fplAnalytics|This thread
is fo…|  false|
|  AutoModerator| 1669971610|Free Talk Friday:…|         ultimate|Use this
thread f…|  false|
|Unusual-Ad-7339| 1669971610|NEW BNB MINER PRO…| solarfarmfinance|Solar Farm
[Solar…|  false|
|  AutoModerator| 1669971610| The Interior I…|      SDRUntucked|*
**Welcome to…|  false|
|  AutoModerator| 1669971610|Daily Discussion …|      covidstocks|
|  false|
|  AutoModerator| 1669971610|* DAILY CHAT THRE…|           ekinde|Good
Morning! \n\…|  false|
|       blink045| 1669971610|     Today is Friday|whichdayoftheweek|
|  false|
|    chopichopfr| 1669971610|Desinfectant Auto…|        INSOLITES|
|  false|
|Unfair-Law-8858| 1669971610|she been bully me…|Bullyporncaptions|
|   true|
| SmartCryptoBot| 1669971610| Ask Anything Thread| smart_crypto_bot|Use this
thread t…|  false|
+---------------+-----------+--------------------+-----------------+------------
--------+-------+
only showing top 20 rows
```

3.1.1 Filtering

[94]:
```
df.filter(~df.over_18).show()
```

```
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
|         author|created_utc|               title|         subreddit|
selftext|over_18|
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
|        Smoke01| 1669971610|Ariana Grande Fan…|             55Tra|
|  false|
|  AutoModerator| 1669971610|Icon Ecosystem We…|         helloicon|Greetings
and wel…|  false|
|  AutoModerator| 1669971610|Get Free ISK and …|       evereferral|
|  false|
|  AutoModerator| 1669971610|Daily Chat - Dece…|          TTC_PCOS|
|  false|
|  AutoModerator| 1669971610|   New Members Intro|     ProVSNatVSFun|If you’re
new to …|  false|
|       SiiZzL4c| 1669971610| Do you use discogs?|             vinyl|
|  false|
|  AutoModerator| 1669971610|Weekly FPL Analyt…|      fplAnalytics|This thread
is fo…|  false|
|  AutoModerator| 1669971610|Free Talk Friday:…|          ultimate|Use this
thread f…|  false|
|Unusual-Ad-7339| 1669971610|NEW BNB MINER PRO…|  solarfarmfinance|Solar Farm
[Solar…|  false|
|  AutoModerator| 1669971610| The Interior I…|       SDRUntucked|*
**Welcome to…|  false|
|  AutoModerator| 1669971610|Daily Discussion …|       covidstocks|
|  false|
|  AutoModerator| 1669971610|* DAILY CHAT THRE…|            ekinde|Good
Morning! \n\…|  false|
|       blink045| 1669971610|     Today is Friday| whichdayoftheweek|
|  false|
|    chopichopfr| 1669971610|Desinfectant Auto…|         INSOLITES|
|  false|
| SmartCryptoBot| 1669971610| Ask Anything Thread|  smart_crypto_bot|Use this
thread t…|  false|
|        Eferver| 1669971610|Is anyone going t…|MotorsportsReplays|
|  false|
|    Fonnster420| 1669971610|G,day auzzies no …|   Auzziemegamemes|no
affending abor…|  false|
|    chopichopfr| 1669971610|Scie adaptable po…|         INSOLITES|
|  false|
|  AutoModerator| 1669971610|Daily Discussion …|              SSBM|Yahoooo!
```

```
Welcome …|  false|
|  AutoModerator| 1669971610|Daily Discussion …|             BBBY|Shop  [Bed
Bath &…|  false|
+---------------+-----------+--------------------+-----------------+-----------
---------+-------+
only showing top 20 rows
```

[95]:
```
df.where(~df.over_18).show()
```

```
+---------------+-----------+--------------------+-----------------+-----------
---------+-------+
|         author|created_utc|               title|        subreddit|
selftext|over_18|
+---------------+-----------+--------------------+-----------------+-----------
---------+-------+
|        Smoke01| 1669971610|Ariana Grande Fan…|            55Tra|
|  false|
|  AutoModerator| 1669971610|Icon Ecosystem We…|        helloicon|Greetings
and wel…|  false|
|  AutoModerator| 1669971610|Get Free ISK and …|      evereferral|
|  false|
|  AutoModerator| 1669971610|Daily Chat - Dece…|         TTC_PCOS|
|  false|
|  AutoModerator| 1669971610|   New Members Intro|    ProVSNatVSFun|If you’re
new to …|  false|
|       SiiZzL4c| 1669971610| Do you use discogs?|            vinyl|
|  false|
|  AutoModerator| 1669971610|Weekly FPL Analyt…|     fplAnalytics|This thread
is fo…|  false|
|  AutoModerator| 1669971610|Free Talk Friday:…|         ultimate|Use this
thread f…|  false|
|Unusual-Ad-7339| 1669971610|NEW BNB MINER PRO…| solarfarmfinance|Solar Farm
[Solar…|  false|
|  AutoModerator| 1669971610|  The Interior I…|      SDRUntucked|*
**Welcome to…|  false|
|  AutoModerator| 1669971610|Daily Discussion …|      covidstocks|
|  false|
|  AutoModerator| 1669971610|* DAILY CHAT THRE…|           ekinde|Good
Morning! \n\…|  false|
|       blink045| 1669971610|     Today is Friday|whichdayoftheweek|
|  false|
|    chopichopfr| 1669971610|Desinfectant Auto…|        INSOLITES|
|  false|
| SmartCryptoBot| 1669971610| Ask Anything Thread| smart_crypto_bot|Use this
thread t…|  false|
|        Eferver| 1669971610|Is anyone going t…|MotorsportsReplays|
|  false|
|    Fonnster420| 1669971610|G,day auzzies no …|   Auzziemegamemes|no
affending abor…|  false|
|    chopichopfr| 1669971610|Scie adaptable po…|         INSOLITES|
|  false|
|  AutoModerator| 1669971610|Daily Discussion …|              SSBM|Yahoooo!
Welcome …|  false|
|  AutoModerator| 1669971610|Daily Discussion …|              BBBY|Shop  [Bed
Bath &…|  false|
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
only showing top 20 rows
```

[96]:

```
df.where("not over_18").show() # SQL syntax
```

```
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
|         author|created_utc|               title|         subreddit|
selftext|over_18|
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
|        Smoke01| 1669971610|Ariana Grande Fan…|             55Tra|
|  false|
|  AutoModerator| 1669971610|Icon Ecosystem We…|         helloicon|Greetings
and wel…|  false|
|  AutoModerator| 1669971610|Get Free ISK and …|       evereferral|
|  false|
|  AutoModerator| 1669971610|Daily Chat - Dece…|          TTC_PCOS|
|  false|
|  AutoModerator| 1669971610|   New Members Intro|     ProVSNatVSFun|If you’re
new to …|  false|
|       SiiZzL4c| 1669971610| Do you use discogs?|             vinyl|
|  false|
|  AutoModerator| 1669971610|Weekly FPL Analyt…|      fplAnalytics|This thread
is fo…|  false|
|  AutoModerator| 1669971610|Free Talk Friday:…|          ultimate|Use this
thread f…|  false|
|Unusual-Ad-7339| 1669971610|NEW BNB MINER PRO…|  solarfarmfinance|Solar Farm
[Solar…|  false|
|  AutoModerator| 1669971610|  The Interior I…|       SDRUntucked|*
**Welcome to…|  false|
|  AutoModerator| 1669971610|Daily Discussion …|       covidstocks|
|  false|
|  AutoModerator| 1669971610|* DAILY CHAT THRE…|            ekinde|Good
Morning! \n\…|  false|
|       blink045| 1669971610|     Today is Friday| whichdayoftheweek|
|  false|
|    chopichopfr| 1669971610|Desinfectant Auto…|         INSOLITES|
|  false|
| SmartCryptoBot| 1669971610| Ask Anything Thread|  smart_crypto_bot|Use this
thread t…|  false|
|        Eferver| 1669971610|Is anyone going t…|MotorsportsReplays|
|  false|
|    Fonnster420| 1669971610|G,day auzzies no …|   Auzziemegamemes|no
affending abor…|  false|
|    chopichopfr| 1669971610|Scie adaptable po…|          INSOLITES|
|  false|
|  AutoModerator| 1669971610|Daily Discussion …|               SSBM|Yahoooo!
Welcome …|  false|
|  AutoModerator| 1669971610|Daily Discussion …|               BBBY|Shop  [Bed
Bath &…|  false|
+---------------+-----------+--------------------+------------------+-----------
---------+-------+
only showing top 20 rows
```

3.1.2 Operations

[97]:

```
df.select(df.created_utc * 2).show()
```

```
+-----------------+
|(created_utc * 2)|
+-----------------+
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
|       3339943220|
+-----------------+
only showing top 20 rows
```

[98]:

```
from pyspark.sql.functions import log
df.select(log(df.created_utc * 2)).show()
```

```
+---------------------+
|ln((created_utc * 2))|
+---------------------+
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
|   21.929219643790518|
+---------------------+
only showing top 20 rows
```

3.1.3 Aggregations

[99]:

```
df.where("not over_18").groupby(["author", df.subreddit]).count().toPandas()
```

[99]:

```
                author             subreddit  count
0             royjones           StopDipping      1
1              feelz09             u_feelz09      1
2           IWNDWYTmod               IWNDWYT      1
3    NachrichtenDE_Bot         NachrichtenDE      7
4              Eferver    MotorsportsReplays      1
..                 …                   …    …
172   milkywayteapasta      WondrousDiscount      1
173    Sam_75_thoughts               answers      1
174   Vast-Piccolo9032    u_Vast-Piccolo9032      2
```

```
175      APD-Supernova  AudioProductionDeals      1
176   Vast-Piccolo9032             MediaFilm      1

[177 rows x 3 columns]
```

3.1.4 Custom functions

[100]:

```
from pyspark.sql.functions import length

df = df.withColumn("length", length(df.selftext)) # This adds a column

df.where("length > 1000").toPandas()
```

[100]:

```
                  author  created_utc   \
0        Unusual-Ad-7339   1669971610
1          AutoModerator   1669971610
2          AutoModerator   1669971610
3          AutoModerator   1669971610
4        AmazingCulinary   1669971610
5            pumpkinduel   1669971609
6               dochavoc   1669971609
7          AutoModerator   1669971609
8          Blogsnark_mod   1669971609
9            Porygon-Bot   1669971609
10         AutoModerator   1669971609
11         AutoModerator   1669971609
12         AutoModerator   1669971609
13                pleck0   1669971607
14  The_Dwemer_Automaton   1669971606
15  The_Dwemer_Automaton   1669971603
16          zhichengping   1669971601
17               Arnadus   1669971600

                                                title           subreddit  \
0   NEW BNB MINER PROJECT - 14,2 % APR daily - Sol…   solarfarmfinance
1   Daily Discussion Thread Dec 02, 2022 - Upcomin…               SSBM
2         Daily Discussion Thread | December 02, 2022                BBBY
3             Fresh Vegetable and Shrimp Spring Rolls     Amazingculinary
4           Panzanella Salad Recipe is for 4 servings     Amazingculinary
5            [F4A] I suppose sex work pays the bills…        dirtypenpals
6           CARV Daily Discussion - December 02, 2022           carvstock
7   [XBOX One] Daily Sales Thread: December 02, AM…       hutcoinsales
8                Influencer Discussion, Friday Dec 02           blogsnark
9   Supported Older Games - Weekly Casual Trade Th…      pokemontrades
10   [PS4] Daily Sales Thread: December 02, AM Thread        hutcoinsales
11  Boston Daily Discussion Thread, Friday Decembe…             boston
```

```
12                            General Chat December 02      TryingForABaby
13  [SELL] New Balance, Nike SB dunk, Norse Projec…   MaleFashionMarket
14  Guild Fair Friday - Advertise your guild, Find…  elderscrollsonline
15                  [Daily] Set Discussion: Morkuldin  elderscrollsonline
16                                  chuban120       u_zhichengping
17                 Most trending cryptos today so far  cryptopricesalerts

                                             selftext  over_18  length
0   Solar Farm [SolarFarmMinerOffical](https://app…    False    1923
1   Yahoooo! Welcome to the Daily Discussion Threa…    False    3240
2   Shop  [Bed Bath &amp; Beyond.com](https://www…    False    3733
3   Kitchen Recipe \nFresh Vegetable and Shrimp Sp…    False    2892
4   Kitchen Recipe \nPanzanella Salad\nRecipe is f…    False    1761
5   Lin smoked a cigarette as she noticed a woman …     True    2150
6    [ ](https://i.redd.it/9oohuwgcjdh71.jpg) **Bu…    False    4445
7   **XBOX One** users: Are you looking to buy or …    False    1120
8   Here's your daily place to snark on the antics…    False    1087
9   # Welcome to the /r/pokemontrades Weekly Casua…    False    1610
10  **PS4** users: Are you looking to buy or sell …    False    1115
11  Hey r/Boston\n\nThis thread is for chatting ab…    False    1350
12  Anything, within the [rules](https://www.reddi…    False    2429
13  PRICES LOWERED. OPEN TO OFFERS. BE REASONABLE\…    False    2855
14  Hey folks,\n\n Welcome to our new recurring po…    False    1113
15  **[Morkuldin](https://eso-hub.com/en/sets/mork…    False    2064
16          \r  \n\r  \n            …    False    1914
17  |Crypto|Number of users mentioning it|\n:--|:-…    False    5049
```

```
[101]: from pyspark.sql.functions import udf

       def splitWords(e):
         return e.split(" ")

       splitWords = udf(splitWords)
       df.select(splitWords(df.selftext)).show()
```

```
+--------------------+
|splitWords(selftext)|
+--------------------+
|                  []|
|                  []|
|[Iran, , \nHave, …|
|[Greetings, and, …|
|                  []|
|                  []|
|                  []|
|[If, you’re, new,…|
|                  []|
|                  []|
|[This, thread, is…|
|[Use, this, threa…|
|[Solar, Farm, [So…|
|[*,  , **Welcome…|
|                  []|
|[Good, Morning!, …|
|                  []|
|                  []|
|                  []|
|[Use, this, threa…|
+--------------------+
only showing top 20 rows
```

[102]:

```
#df.where("author = 'TheStartupChime'").toPandas()
df.groupby(["author"]).count().where("count > 1000").toPandas()
```

[102]:

```
Empty DataFrame
Columns: [author, count]
Index: []
```

Assignment question

Obtain the users who have posted in reddit more than 1k posts in any subreddit

---

Answer: agrupas por author sin importar el subreddit, cuentas las apiriciones y filtras para obtener aquellos con un número de apariciones de 1000 o más df.groupby([”author”]).count().where(”count > 1000”).toPandas()

3.2 SQL operations

3.2.1 How to declare a view from a Dataframe

[103]:

```
df.createOrReplaceTempView("reddit")
```

[104]:

```
spark.sql("select * from reddit limit 10").show()
```

```
+--------------+-----------+--------------------+----------------+--------------
------+-------+------+
|        author|created_utc|               title|       subreddit|
selftext|over_18|length|
+--------------+-----------+--------------------+----------------+--------------
```

```
------+-------+------+
|       Smoke01| 1669971610|Ariana Grande Fan…|           55Tra|
|  false|     0|
|freiburger1884| 1669971610|I wish i had anot…|u_freiburger1884|
|   true|     0|
| Amirali_toxic| 1669971610|           iran news|    irannews2022|Iran  \nHave
peop…|   true|   297|
| AutoModerator| 1669971610|Icon Ecosystem We…|       helloicon|Greetings and
wel…|  false|   646|
| AutoModerator| 1669971610|Get Free ISK and …|     evereferral|
|  false|     0|
| AutoModerator| 1669971610|Daily Chat - Dece…|         TTC_PCOS|
|  false|     0|
|      DlngoLex| 1669971610|             Shenise|           sheer|
|   true|     0|
| AutoModerator| 1669971610|   New Members Intro|   ProVSNatVSFun|If you're new
to …|  false|    51|
|      SiiZzL4c| 1669971610| Do you use discogs?|           vinyl|
|  false|     0|
|  ReveErotique| 1669971610|      Amelia Braymut|      Bellissima|
|   true|     0|
+--------------+-----------+--------------------+----------------+--------------
------+-------+------+
```

Assignment question

Obtain the users who have posted in reddit more than 1k characters in any subreddit with SQL (without using any column named length)

Answer: respuesta en la celda siguiente

```
[105]: spark.sql("select author, SUM(CHAR_LENGTH(selftext)) as s from reddit  group by author having s > 1000").show()
```

```
+--------------------+-----+
|              author|    s|
+--------------------+-----+
|            dochavoc| 4445|
|         Porygon-Bot| 1610|
|       AutoModerator|24508|
|         pumpkinduel| 2150|
|       Blogsnark_mod| 1087|
|     Unusual-Ad-7339| 1923|
|     AmazingCulinary| 1761|
|        zhichengping| 1914|
|             Arnadus| 5049|
|The_Dwemer_Automaton| 3177|
|              pleck0| 2855|
+--------------------+-----+
```

3.3 Other libraries

Beyond dataframes, we can find other libraries that also rely on Spark...

[106]:
```
!pip install koalas
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: koalas in /usr/local/lib/python3.8/dist-packages
(1.8.2)
Requirement already satisfied: pandas>=0.23.2 in /usr/local/lib/python3.8/dist-
packages (from koalas) (1.3.5)
Requirement already satisfied: pyarrow>=0.10 in /usr/local/lib/python3.8/dist-
packages (from koalas) (9.0.0)
Requirement already satisfied: numpy>=1.14 in /usr/local/lib/python3.8/dist-
packages (from koalas) (1.21.6)
Requirement already satisfied: python-dateutil>=2.7.3 in
/usr/local/lib/python3.8/dist-packages (from pandas>=0.23.2->koalas) (2.8.2)
Requirement already satisfied: pytz>=2017.3 in /usr/local/lib/python3.8/dist-
packages (from pandas>=0.23.2->koalas) (2022.6)
Requirement already satisfied: six>=1.5 in /usr/local/lib/python3.8/dist-
packages (from python-dateutil>=2.7.3->pandas>=0.23.2->koalas) (1.15.0)
```

[107]:
```python
import databricks.koalas as ks
import pandas as pd

# Create a Koalas DataFrame from pandas DataFrame
kdf = ks.from_pandas(pdf[attrs])

kdf.head()
```

[107]:
```
           author  created_utc
title          subreddit
selftext  over_18
0         Smoke01   1669971610   Ariana Grande Fans React To Paige Niemann's
OnlyFans Account             55Tra
False
1  freiburger1884   1669971610      I wish i had another cock in my mouth! Who
wants to join?  u_freiburger1884
True
```

```
2   Amirali_toxic   1669971610
iran news      irannews2022
Iran   \nHave people's protests in Iran reduced the chances of revitalizing the
JCPOA?  \nAfter 20 days of popular protests all over Iran, how have the
positions of Iran and the West changed regarding the negotiations to revive the
JCPOA? Are the parties still looking for nuclear talks?  \n09/10/2022     True
3   AutoModerator   1669971610    Icon Ecosystem Weekly Discussion Thread -
December 02, 2022          helloicon  Greetings and welcome to the Icon ecosystem
weekly discussion thread!   \n\n\nHere everyone is encouraged to ask questions,
learn, teach, and discuss Icon ecosystem projects. Price discussion and off
topic discussions are also welcome. Please show respect to your fellow user and
follow our subreddit rules.   \n\n\nIn need of assistance? Please check out -
[Icon FAQ and handy links](https://www.reddit.com/r/helloicon/comments/xs0v89/ic
on_ecosystem_faq_frequently_asked_questions_and/)  \n\n\n*Mods/Icon staff will
never message you directly, please be wary of scammers contacting you. Do not
share your private information or private keys with anyone.*    False
4   AutoModerator   1669971610  Get Free ISK and Skill Points in Eve Online -
evereferral.com       evereferral
False
```

[108]:

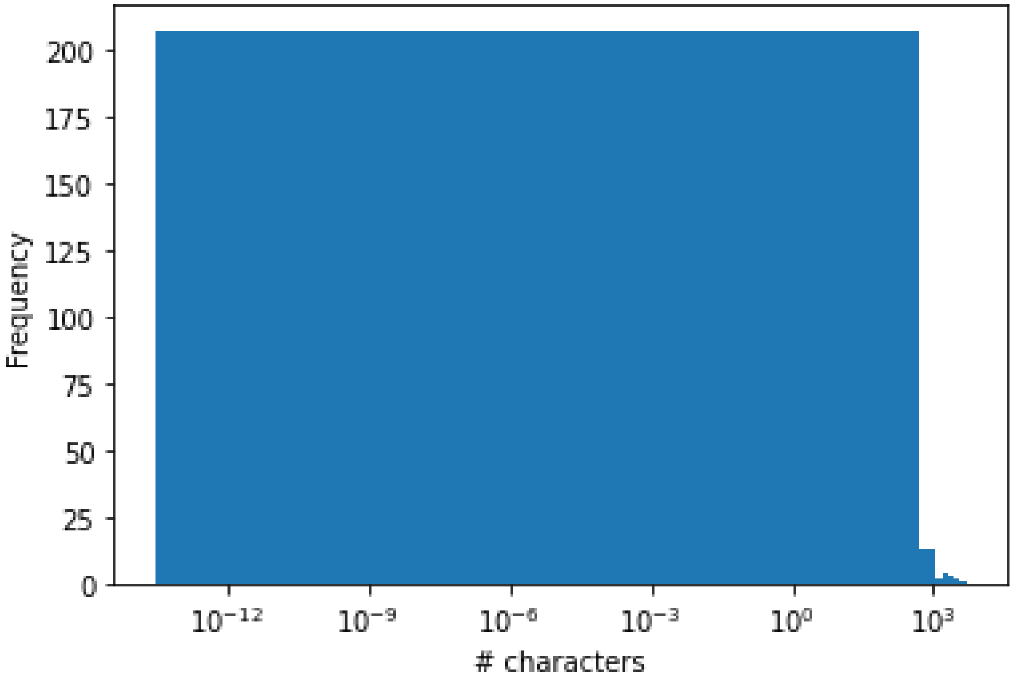
```
kdf["sumChars"] = kdf.selftext.str.len()
res = kdf.groupby(["author", "subreddit"]).sum()
res[res.sumChars > 1000]
```

[108]:

```
                                         created_utc  sumChars
author               subreddit
AutoModerator        hutcoinsales         3339943218      2235
                     TryingForABaby       1669971609      2429
Blogsnark_mod        blogsnark            1669971609      1087
Porygon-Bot          pokemontrades        1669971609      1610
pumpkinduel          dirtypenpals         1669971609      2150
AutoModerator        boston               1669971609      1350
Unusual-Ad-7339      solarfarmfinance     1669971610      1923
AutoModerator        Amazingculinary      1669971610      2892
                     BBBY                 1669971610      3733
dochavoc             carvstock            1669971609      4445
AutoModerator        SSBM                 1669971610      3240
AmazingCulinary      Amazingculinary      1669971610      1761
pleck0               MaleFashionMarket    1669971607      2855
zhichengping         u_zhichengping       1669971601      1914
Arnadus              cryptopricesalerts   1669971600      5049
The_Dwemer_Automaton elderscrollsonline   3339943209      3177
```

[109]:

```
import matplotlib.pyplot as plt
plt.hist(res.sumChars.to_numpy())
plt.xlabel("# characters")
plt.ylabel("Frequency")
plt.xscale("log")
```



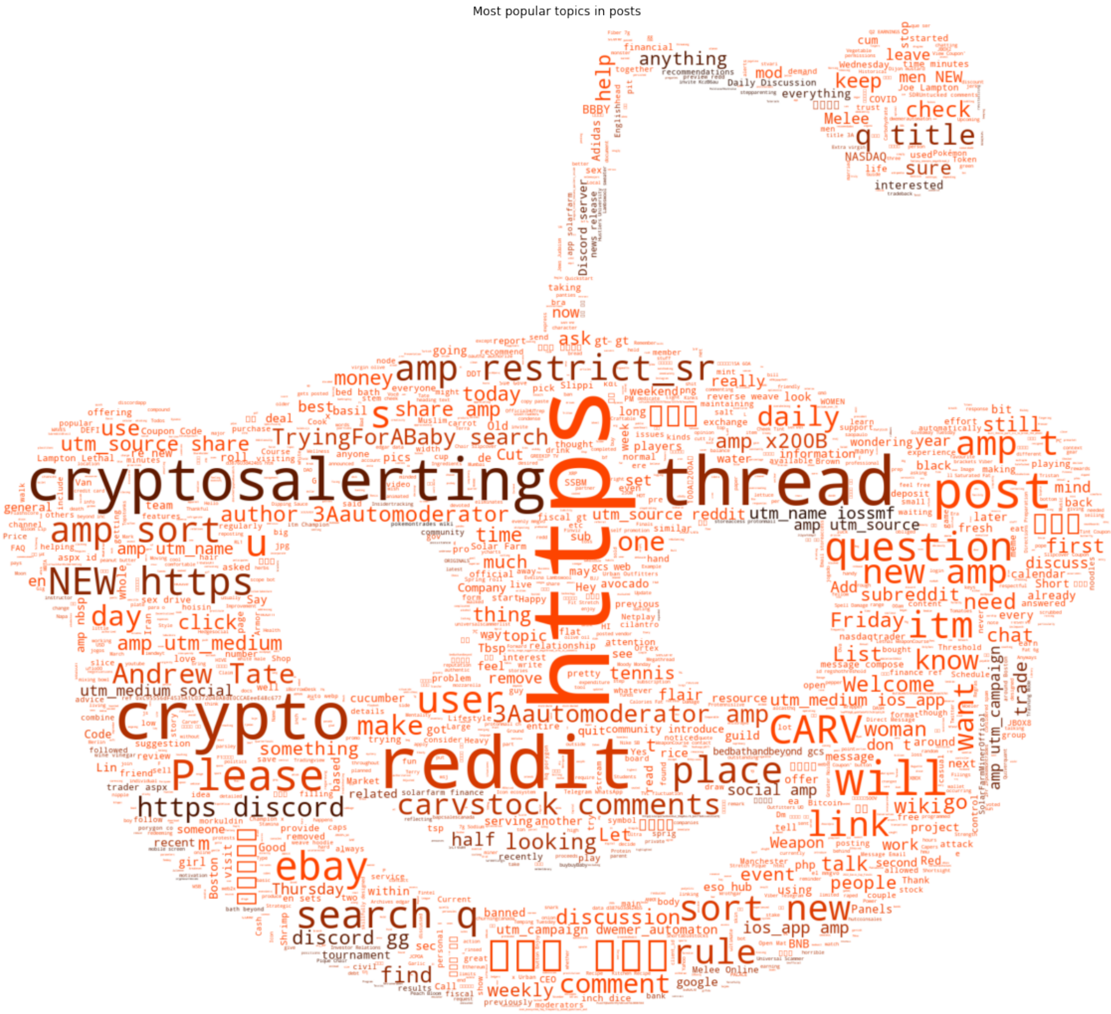
[110]:

```
!curl https://2.bp.blogspot.com/-eGskF3n8_Ag/XE7F3P_de2I/AAAAAAAAHU8/
 ↪WJw0un2nHqMGA8cFVtv_yFfpBVQJSYyVACK4BGAYYCw/s1600/Icon-Reddit.png > reddit.
 ↪png
from wordcloud import WordCloud, ImageColorGenerator
from PIL import Image

mask = np.array(Image.open("reddit.png"))
text = " ".join([i for i in kdf.selftext.to_numpy() if len(i) > 0 and i !=␣
 ↪"[removed]" and i!="[deleted]"])
```

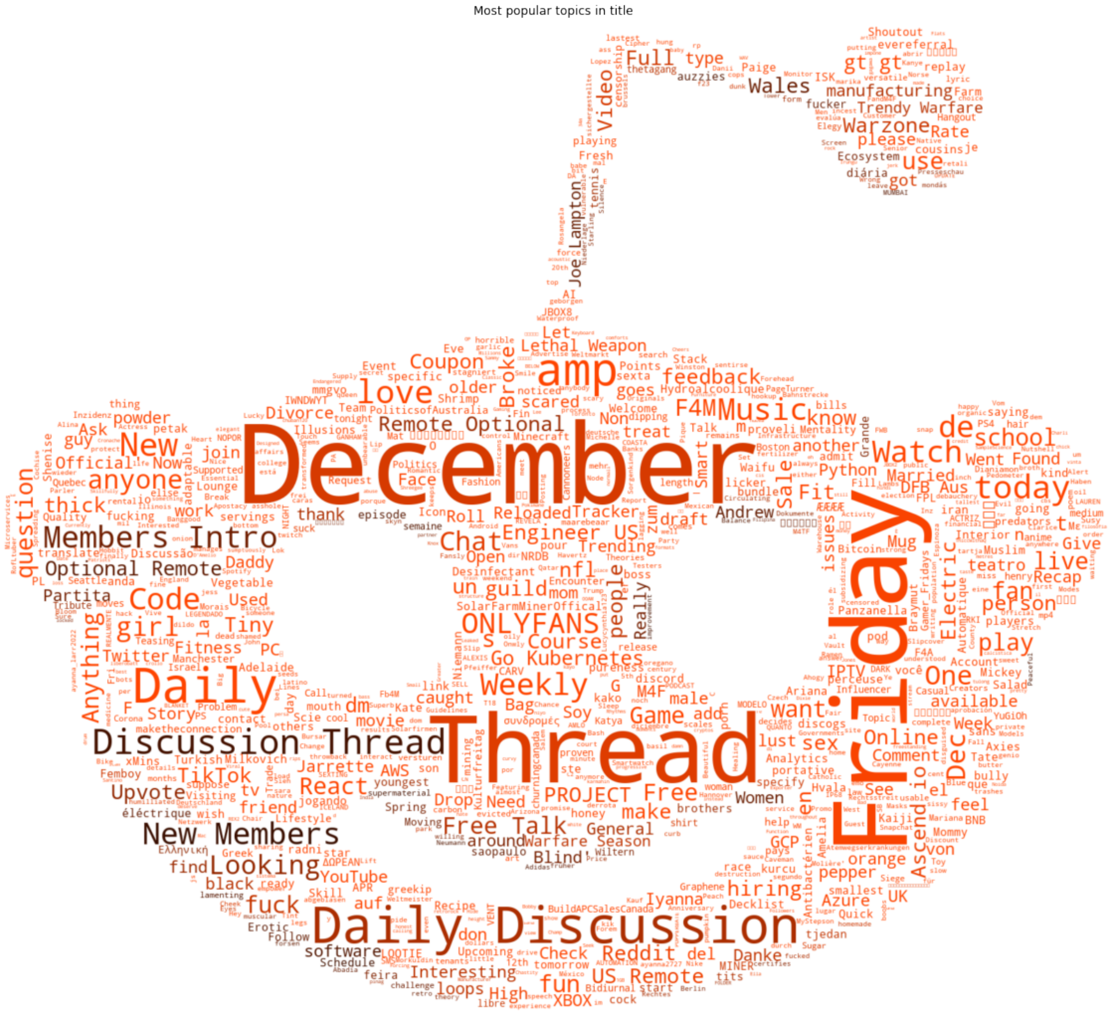
```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 76154  100 76154    0     0  3098k      0 --:--:-- --:--:-- --:--:-- 3098k
```

[111]:

```
text = " ".join([i for i in kdf.selftext.to_numpy() if len(i) > 0 and i !=␣
 ↪"[removed]" and i!="[deleted]"])
wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0],␣
 ↪background_color="white", mode="RGBA").generate(text)
# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
plt.title("Most popular topics in posts")
plt.axis("off");
```



```
[55]: text = " ".join([i for i in kdf.title.to_numpy() if len(i) > 0 and i !=␣
       ↪"[removed]" and i!="[deleted]"])
      wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0],␣
       ↪background_color="white", mode="RGBA").generate(text)
      # create coloring from image
      image_colors = ImageColorGenerator(mask)
      plt.figure(figsize=(20,20))
      plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
      plt.title("Most popular topics in title")
      plt.axis("off");
```



```
[ ]: text = " ".join([i for i in kdf.subreddit.to_numpy() if len(i) > 0 and i !=␣
     ↪"[removed]" and i!="[deleted]"])
     wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0],␣
     ↪background_color="white", mode="RGBA").generate(text)
     # create coloring from image
     image_colors = ImageColorGenerator(mask)
     plt.figure(figsize=(20,20))
     plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
     plt.title("Most popular subreddits")
     plt.axis("off");
```

Assignment question

Choose a subreddit you like and build a worldcloud using Koalas. Feel free to change the mask or the colors...

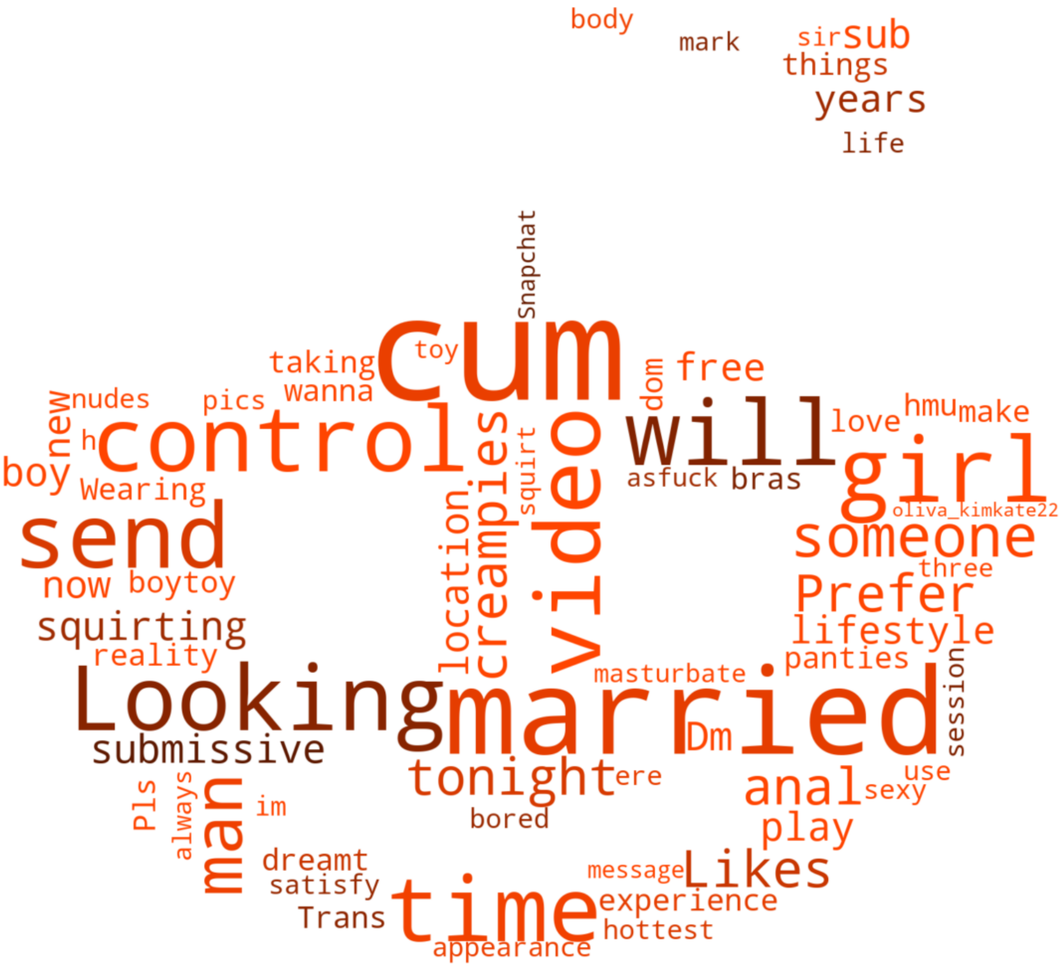
Answer:

[57]:

```
from wordcloud import WordCloud, ImageColorGenerator
from PIL import Image

!curl https://2.bp.blogspot.com/-eGskF3n8_Ag/XE7F3P_de2I/AAAAAAAAHU8/
 ↪WJwOun2nHqMGA8cFVtv_yFfpBVQJSYyVACK4BGAYYCw/s1600/Icon-Reddit.png > reddit.
 ↪png
from wordcloud import WordCloud, ImageColorGenerator
from PIL import Image

mask = np.array(Image.open("reddit.png"))



# obtenemos el subreddit (me sale uno sin self.text)
kdf = ks.from_pandas(pdf[attrs])
kdf_aux = kdf.filter(items=["subreddit"])
subreddit = kdf.at[0,"subreddit"]

#vamos a recorrerlos para ver el subreddit con más texto
dic = {}
for i, j in zip(kdf.subreddit.to_numpy(), kdf.selftext.to_numpy()):
  if(len(i)>0 and len(j)>0 and i != "[removed]" and i!="[deleted]" and j!=␣
 ↪"[removed]" and j!="[deleted]"):
      if i in dic:
        dic[i].append(j)
      else:
        dic[i]=[j]

max_value=max(dic, key=lambda k: len(dic[k]))
#print(max_value, dic[max_value])
#cojo los textos de los validos
#text = " ".join([j for i, j in zip(kdf.subreddit.to_numpy(), kdf.selftext.
 ↪to_numpy()) if i==max_value])
#print(text)
text = " ".join(dic[max_value])
wordcloud = WordCloud(max_words=5000, mask=~mask[:,:,0],␣
 ↪background_color="white", mode="RGBA").generate(text)
# create coloring from image
image_colors = ImageColorGenerator(mask)
plt.figure(figsize=(20,20))
plt.imshow(wordcloud.recolor(color_func=image_colors), interpolation="bilinear")
plt.title("Most topics in subreddit " + max_value)
plt.axis("off");
```

```
  % Total    % Received % Xferd  Average Speed   Time    Time     Time  Current
                                 Dload  Upload   Total   Spent    Left  Speed
100 76154  100 76154    0     0  1906k      0 --:--:-- --:--:-- --:--:-- 1906k
```

Most topics in subreddit dirtyr4r


body
mark
sir
sub
things
years
life
Snapchat
taking
toy
cum
free
wanna
dom
nudes
pics
hmu
make
new
love
h
control
will
girl
boy
squirt
creampies
video
asfuck
bras
Wearing
oliva_kimkate22
send
location
someone
three
now
boytoy
Prefer
squirting
lifestyle
reality
masturbate
panties
Looking
married
session
Dm
submissive
tonight
ere
anal
use
sexy
Pls
always
man
im
bored
play
dreamt
message
Likes
satisfy
experience
Trans
time
hottest
appearance