1. Apache Spark ML

|  |  |
| --- | --- |
| **S.No** | **Transformations & Meaning** |
| 1 | **map(func)**  Returns a new distributed dataset, formed by passing each element of the source through a function **func**. |
| 2 | **filter(func)**  Returns a new dataset formed by selecting those elements of the source on which **func** returns true. |
| 3 | **flatMap(func)**  Similar to map, but each input item can be mapped to 0 or more output items (so *func* should return a Seq rather than a single item). |
| 4 | **mapPartitions(func)**  Similar to map, but runs separately on each partition (block) of the RDD, so **func** must be of type Iterator<T> ⇒ Iterator<U> when running on an RDD of type T. |
| 5 | **mapPartitionsWithIndex(func)**  Similar to map Partitions, but also provides **func** with an integer value representing the index of the partition, so **func** must be of type (Int, Iterator<T>) ⇒ Iterator<U> when running on an RDD of type T. |
| 6 | **sample(withReplacement, fraction, seed)**  Sample a **fraction** of the data, with or without replacement, using a given random number generator seed. |
| 7 | **union(otherDataset)**  Returns a new dataset that contains the union of the elements in the source dataset and the argument. |
| 8 | **intersection(otherDataset)**  Returns a new RDD that contains the intersection of elements in the source dataset and the argument. |
| 9 | **distinct([numTasks])**  Returns a new dataset that contains the distinct elements of the source dataset. |
| 10 | **groupByKey([numTasks])**  When called on a dataset of (K, V) pairs, returns a dataset of (K, Iterable<V>) pairs.  **Note** − If you are grouping in order to perform an aggregation (such as a sum or average) over each key, using reduceByKey or aggregateByKey will yield much better performance. |
| 11 | **reduceByKey(func, [numTasks])**  When called on a dataset of (K, V) pairs, returns a dataset of (K, V) pairs where the values for each key are aggregated using the given reduce function *func*, which must be of type (V, V) ⇒ V. Like in groupByKey, the number of reduce tasks is configurable through an optional second argument. |
| 12 | **aggregateByKey(zeroValue)(seqOp, combOp, [numTasks])**  When called on a dataset of (K, V) pairs, returns a dataset of (K, U) pairs where the values for each key are aggregated using the given combine functions and a neutral "zero" value. Allows an aggregated value type that is different from the input value type, while avoiding unnecessary allocations. Like in groupByKey, the number of reduce tasks is configurable through an optional second argument. |
| 13 | **sortByKey([ascending], [numTasks])**  When called on a dataset of (K, V) pairs where K implements Ordered, returns a dataset of (K, V) pairs sorted by keys in ascending or descending order, as specified in the Boolean ascending argument. |
| 14 | **join(otherDataset, [numTasks])**  When called on datasets of type (K, V) and (K, W), returns a dataset of (K, (V, W)) pairs with all pairs of elements for each key. Outer joins are supported through leftOuterJoin, rightOuterJoin, and fullOuterJoin. |
| 15 | **cogroup(otherDataset, [numTasks])**  When called on datasets of type (K, V) and (K, W), returns a dataset of (K, (Iterable<V>, Iterable<W>)) tuples. This operation is also called group With. |
| 16 | **cartesian(otherDataset)**  When called on datasets of types T and U, returns a dataset of (T, U) pairs (all pairs of elements). |
| 17 | **pipe(command, [envVars])**  Pipe each partition of the RDD through a shell command, e.g. a Perl or bash script. RDD elements are written to the process's stdin and lines output to its stdout are returned as an RDD of strings. |
| 18 | **coalesce(numPartitions)**  Decrease the number of partitions in the RDD to numPartitions. Useful for running operations more efficiently after filtering down a large dataset. |
| 19 | **repartition(numPartitions)**  Reshuffle the data in the RDD randomly to create either more or fewer partitions and balance it across them. This always shuffles all data over the network. |
| 20 | **repartitionAndSortWithinPartitions(partitioner)**  Repartition the RDD according to the given partitioner and, within each resulting partition, sort records by their keys. This is more efficient than calling repartition and then sorting within each partition because it can push the sorting down into the shuffle machinery. |

## **Actions**

The following table gives a list of Actions, which return values.

|  |  |
| --- | --- |
| **S.No** | **Action & Meaning** |
| 1 | **reduce(func)**  Aggregate the elements of the dataset using a function **func** (which takes two arguments and returns one). The function should be commutative and associative so that it can be computed correctly in parallel. |
| 2 | **collect()**  Returns all the elements of the dataset as an array at the driver program. This is usually useful after a filter or other operation that returns a sufficiently small subset of the data. |
| 3 | **count()**  Returns the number of elements in the dataset. |
| 4 | **first()**  Returns the first element of the dataset (similar to take (1)). |
| 5 | **take(n)**  Returns an array with the first **n** elements of the dataset. |
| 6 | **takeSample (withReplacement,num, [seed])**  Returns an array with a random sample of **num** elements of the dataset, with or without replacement, optionally pre-specifying a random number generator seed. |
| 7 | **takeOrdered(n, [ordering])**  Returns the first **n** elements of the RDD using either their natural order or a custom comparator. |
| 8 | **saveAsTextFile(path)**  Writes the elements of the dataset as a text file (or set of text files) in a given directory in the local filesystem, HDFS or any other Hadoop-supported file system. Spark calls toString on each element to convert it to a line of text in the file. |
| 9 | **saveAsSequenceFile(path) (Java and Scala)**  Writes the elements of the dataset as a Hadoop SequenceFile in a given path in the local filesystem, HDFS or any other Hadoop-supported file system. This is available on RDDs of key-value pairs that implement Hadoop's Writable interface. In Scala, it is also available on types that are implicitly convertible to Writable (Spark includes conversions for basic types like Int, Double, String, etc). |
| 10 | **saveAsObjectFile(path) (Java and Scala)**  Writes the elements of the dataset in a simple format using Java serialization, which can then be loaded using SparkContext.objectFile(). |
| 11 | **countByKey()**  Only available on RDDs of type (K, V). Returns a hashmap of (K, Int) pairs with the count of each key. |
| 12 | **foreach(func)**  Runs a function **func** on each element of the dataset. This is usually, done for side effects such as updating an Accumulator or interacting with external storage systems.  **Note** − modifying variables other than Accumulators outside of the foreach() may result in undefined behavior. See Understanding closures for more details. |

2. Azure ML toolkit:

| **Module** | **Description** |
| --- | --- |
| [Add Columns](https://msdn.microsoft.com/en-us/library/azure/dn906007.aspx) | Adds a set of columns from one dataset to another |
| [Add Rows](https://msdn.microsoft.com/en-us/library/azure/dn905871.aspx) | Appends a set of rows from an input dataset to the end of another dataset |
| [Apply Filter](https://msdn.microsoft.com/en-us/library/azure/dn905876.aspx) | Applies a filter to specified columns of a dataset |
| [Apply Math Operation](https://msdn.microsoft.com/en-us/library/azure/dn905975.aspx) | Applies a mathematical operation to column values |
| [Apply SQL Transformation](https://msdn.microsoft.com/en-us/library/azure/dn905914.aspx) | Runs a SQLite query on input datasets to transform the data |
| [Apply Transformation](https://msdn.microsoft.com/en-us/library/azure/dn913055.aspx) | Applies a well-specified data transformation to a dataset |
| [Assign to Clusters (deprecated)](https://msdn.microsoft.com/en-us/library/azure/dn905957.aspx) | Assigns data to clusters using an existing trained clustering model  This module has been deprecated, but is available for use with existing experiments. For new experiments, use [Assign Data to Clusters](https://msdn.microsoft.com/en-us/library/azure/mt484313.aspx). |
| [Assign Data to Clusters](https://msdn.microsoft.com/en-us/library/azure/mt484313.aspx) | Assigns data to clusters using an existing trained clustering model |
| [Bayesian Linear Regression](https://msdn.microsoft.com/en-us/library/azure/dn906022.aspx) | Creates a Bayesian linear regression model |
| [Boosted Decision Tree Regression](https://msdn.microsoft.com/en-us/library/azure/dn905801.aspx) | Creates a regression model using the Boosted Decision Tree algorithm |
| [Build Count Table (deprecated)](https://msdn.microsoft.com/en-us/library/azure/dn913606.aspx) | Creates counts to use in building features |
| [Build Counting Transform](https://msdn.microsoft.com/en-us/library/azure/mt243845.aspx) | Creates counts to use in building features |
| [Clean Missing Data](https://msdn.microsoft.com/en-us/library/azure/dn906028.aspx) | Specifies how to handle the values missing from a dataset |
| [Clip Values](https://msdn.microsoft.com/en-us/library/azure/dn905918.aspx) | Detects outliers and then clips or replaces their values |
| [Compute Elementary Statistics](https://msdn.microsoft.com/en-us/library/azure/dn905936.aspx) | Calculates specified summary statistics for selected dataset columns |
| [Detect Languages](https://msdn.microsoft.com/en-us/library/azure/mt762913.aspx) | Detects the language of each line in the input file |
| [Compute Linear Correlation](https://msdn.microsoft.com/en-us/library/azure/dn905819.aspx) | Calculates the linear correlation between column values in a dataset |
| [Convert To Arff](https://msdn.microsoft.com/en-us/library/azure/dn905864.aspx) | Converts data input to the attribute relation file format used by the Weka toolset |
| [Convert to CSV](https://msdn.microsoft.com/en-us/library/azure/dn906002.aspx) | Converts data input to a comma-separated values format |
| [Convert to Dataset](https://msdn.microsoft.com/en-us/library/azure/dn905850.aspx) | Converts data input to the internal Dataset format used by Microsoft Azure Machine Learning |
| [Convert to Indicator Values](https://msdn.microsoft.com/en-us/library/azure/dn905831.aspx) | Converts categorical values in columns to indicator values |
| [Convert to SVMLight](https://msdn.microsoft.com/en-us/library/azure/dn905860.aspx) | Converts data input to the format used by the SVM-Light framework |
| [Convert to TSV](https://msdn.microsoft.com/en-us/library/azure/dn905866.aspx) | Converts data input to the tab-delimited format |
| [Count Featurizer (deprecated)](https://msdn.microsoft.com/en-us/library/azure/dn913059.aspx) | Generates a set of compact features based on a count table |
| [Create R Model](https://msdn.microsoft.com/en-us/library/azure/dn955435.aspx) | Creates an R model using custom resources |
| [Cross-Validate Model](https://msdn.microsoft.com/en-us/library/azure/dn905852.aspx) | Cross-validates parameter estimates for classification or regression models by partitioned the data |
| [Decision Forest Regression](https://msdn.microsoft.com/en-us/library/azure/dn905862.aspx) | Creates a regression model using the decision forest algorithm |
| [Detect Languages](https://msdn.microsoft.com/en-us/library/azure/mt762913.aspx) | Detects the language of each line in the input file |
| [Edit Metadata](https://msdn.microsoft.com/en-us/library/azure/dn905986.aspx) | Edits metadata associated with columns in a dataset |
| [Enter Data Manually](https://msdn.microsoft.com/en-us/library/azure/dn905948.aspx) | Enables entering and editing small datasets by typing values |
| [Evaluate Model](https://msdn.microsoft.com/en-us/library/azure/dn905915.aspx) | Evaluates a scored classification or regression model with standard metrics |
| [Evaluate Probability Function](https://msdn.microsoft.com/en-us/library/azure/dn905813.aspx) | Fits a specified probability distribution function to a dataset |
| [Evaluate Recommender](https://msdn.microsoft.com/en-us/library/azure/dn905954.aspx) | Evaluates the accuracy of recommender model predictions |
| [Execute Python Script](https://msdn.microsoft.com/en-us/library/azure/dn955437.aspx) | Executes a Python script from an Azure Machine Learning experiment |
| [Execute R Script](https://msdn.microsoft.com/en-us/library/azure/dn905952.aspx) | Executes an R script from an Azure Machine Learning experiment |
| [Export Count Table](https://msdn.microsoft.com/en-us/library/azure/mt243849.aspx) | Exports counts from a count transform |
| [Export Data](https://msdn.microsoft.com/en-us/library/azure/dn905984.aspx) | Writes a dataset to web URLs or to various forms of cloud-based storage in Azure, such as tables, blobs, and Azure SQL databases  This module was formerly named **Writer**. |
| [Extract Key Phrases from Text](https://msdn.microsoft.com/en-us/library/azure/mt786708.aspx) | Extracts key words and phrases from a text column |
| [Extract N-Gram Features from Text](https://msdn.microsoft.com/en-us/library/azure/mt762916.aspx) | Creates N-Gram dictionary features and does feature selection on them |
| [Fast Forest Quantile Regression](https://msdn.microsoft.com/en-us/library/azure/dn913093.aspx) | Creates a quantile regression model |
| [Feature Hashing](https://msdn.microsoft.com/en-us/library/azure/dn906018.aspx) | Converts text data to integer-encoded features using the Vowpal Wabbit library |
| [Filter Based Feature Selection](https://msdn.microsoft.com/en-us/library/azure/dn913071.aspx) | Identifies the features in a dataset with the greatest predictive power |
| [FIR Filter](https://msdn.microsoft.com/en-us/library/azure/dn906035.aspx) | Creates a finite impulse response filter for signal processing |
| [Fisher Linear Discriminant Analysis](https://msdn.microsoft.com/en-us/library/azure/dn913100.aspx) | Identifies the linear combination of feature variables that can best group data into separate classes |
| [Group Categorical Values](https://msdn.microsoft.com/en-us/library/azure/dn906014.aspx) | Groups data from multiple categories into a new category |
| [Group Data into Bins](https://msdn.microsoft.com/en-us/library/azure/dn913065.aspx) | Puts numerical data into bins |
| [IIR Filter](https://msdn.microsoft.com/en-us/library/azure/dn905814.aspx) | Creates an infinite impulse response filter for signal processing |
| [Import Count Table](https://msdn.microsoft.com/en-us/library/azure/mt243848.aspx) | Imports counts from an existing count table |
| [Import Data](https://msdn.microsoft.com/en-us/library/azure/dn905997.aspx) | Loads data from external sources on the web; from various forms of cloud-based storage in Azure such as tables, blobs, and SQL databases; and from on-premises SQL Server databases  This module was formerly named **Reader**. |
| [Import Images](https://msdn.microsoft.com/en-us/library/azure/dn905877.aspx) | Loads images from Azure BLOB storage into a dataset |
| [Join Data](https://msdn.microsoft.com/en-us/library/azure/dn905836.aspx) | Joins two datasets |
| [K-Means Clustering](https://msdn.microsoft.com/en-us/library/azure/dn905944.aspx) | Configures and initializes a K-means clustering model |
| [Latent Dirichlet Allocation](https://msdn.microsoft.com/en-us/library/azure/mt762914.aspx) | Performs topic modeling using the Vowpal Wabbit library for LDA |
| [Linear Regression](https://msdn.microsoft.com/en-us/library/azure/dn905978.aspx) | Creates a linear regression model |
| [Load Trained Model](https://msdn.microsoft.com/en-us/library/azure/mt809133.aspx) | Gets a trained model so that you can use it in an experiment for scoring |
| [Median Filter](https://msdn.microsoft.com/en-us/library/azure/dn905833.aspx) | Creates a median filter used to smooth data for trend analysis |
| [Merge Count Transform](https://msdn.microsoft.com/en-us/library/azure/mt243847.aspx) | Merges two sets of count tables |
| [Missing Values Scrubber (deprecated)](https://msdn.microsoft.com/en-us/library/azure/dn906010.aspx) | Specifies how to handle values that are missing from a dataset |
| [Modify Count Table Parameters](https://msdn.microsoft.com/en-us/library/azure/mt243846.aspx) | Builds a compact set of count-based features from count tables |
| [Moving Average Filter](https://msdn.microsoft.com/en-us/library/azure/dn905894.aspx) | Creates a moving average filter that smooths data for trend analysis |
| [Multiclass Decision Forest](https://msdn.microsoft.com/en-us/library/azure/dn906015.aspx) | Creates a multiclass classification model using the decision forest algorithm |
| [Multiclass Decision Jungle](https://msdn.microsoft.com/en-us/library/azure/dn905963.aspx) | Creates a multiclass classification model using the decision jungle algorithm |
| [Multiclass Logistic Regression](https://msdn.microsoft.com/en-us/library/azure/dn905853.aspx) | Creates a multiclass logistic regression classification model |
| [Multiclass Neural Network](https://msdn.microsoft.com/en-us/library/azure/dn906030.aspx) | Creates a multiclass classification model using a neural network algorithm |
| [Named Entity Recognition](https://msdn.microsoft.com/en-us/library/azure/dn905955.aspx) | Recognizes named entities in a text column |
| [Neural Network Regression](https://msdn.microsoft.com/en-us/library/azure/dn905924.aspx) | Creates a regression model using a neural network algorithm |
| [Normalize Data](https://msdn.microsoft.com/en-us/library/azure/dn905838.aspx) | Rescales numeric data to constrain dataset values to a standard range |
| [One-Class Support Vector Machine](https://msdn.microsoft.com/en-us/library/azure/dn913103.aspx) | Creates a one-class Support Vector Machine model for anomaly detection |
| [One-vs-All Multiclass](https://msdn.microsoft.com/en-us/library/azure/dn905887.aspx) | Creates a multiclass classification model from an ensemble of binary classification models |
| [Ordinal Regression](https://msdn.microsoft.com/en-us/library/azure/dn906029.aspx) | Creates an ordinal regression model |
| [Partition and Sample](https://msdn.microsoft.com/en-us/library/azure/dn905960.aspx) | Creates multiple partitions of a dataset based on sampling |
| [Permutation Feature Importance](https://msdn.microsoft.com/en-us/library/azure/dn997262.aspx) | Computes the permutation feature importance scores of feature variables given a trained model and a test dataset |
| [PCA-Based Anomaly Detection](https://msdn.microsoft.com/en-us/library/azure/dn913102.aspx) | Creates an anomaly detection model using Principal Component Analysis |
| [Poisson Regression](https://msdn.microsoft.com/en-us/library/azure/dn905988.aspx) | Creates a regression model that assumes data has a Poisson distribution |
| [Preprocess Text](https://msdn.microsoft.com/en-us/library/azure/mt762915.aspx) | Performs cleaning operations on text |
| [Pretrained Cascade Image Classification](https://msdn.microsoft.com/en-us/library/azure/dn913079.aspx) | Creates a pretrained image classification model for frontal faces using the OpenCV Library |
| [Principal Component Analysis](https://msdn.microsoft.com/en-us/library/azure/dn913053.aspx) | Computes a set of features with reduced dimensionality for more efficient learning |
| [Remove Duplicate Rows](https://msdn.microsoft.com/en-us/library/azure/dn905805.aspx) | Removes the duplicate rows from a dataset |
| [Replace Discrete Values](https://msdn.microsoft.com/en-us/library/azure/dn905815.aspx) | Replaces discrete values from one column with numeric values based on another column |
| [Score Matchbox Recommender](https://msdn.microsoft.com/en-us/library/azure/dn905970.aspx) | Scores predictions for a dataset using the Matchbox recommender |
| [Score Model](https://msdn.microsoft.com/en-us/library/azure/dn905995.aspx) | Scores predictions for a trained classification or regression model |
| [Score Vowpal Wabbit 7-4 Model](https://msdn.microsoft.com/en-us/library/azure/dn905869.aspx) | Scores data using the Vowpal Wabbit machine learning system  Requires a trained model built using VW versions 7-4 and 7-6 |
| [Score Vowpal Wabbit 7-10 Model](https://msdn.microsoft.com/en-us/library/azure/mt674684.aspx) | Scores data using the Vowpal Wabbit machine learning system  Requires a trained model built using VW version 7-10 |
| [Score Vowpal Wabbit 8 Model](https://msdn.microsoft.com/en-us/library/azure/mt712720.aspx) | Scores data using the Vowpal Wabbit machine learning system from the command line interface  Requires a trained model built using VW version 8 |
| [Select Columns in Dataset](https://msdn.microsoft.com/en-us/library/azure/dn905883.aspx) | Selects columns to include or exclude from a dataset in an operation |
| [SMOTE](https://msdn.microsoft.com/en-us/library/azure/dn913076.aspx) | Increases the number of low incidence examples in a dataset using synthetic minority oversampling |
| [Split Data](https://msdn.microsoft.com/en-us/library/azure/dn905969.aspx) | Partitions the rows of a dataset into two distinct sets |
| [Summarize Data](https://msdn.microsoft.com/en-us/library/azure/dn905933.aspx) | Generates a basic descriptive statistics report for the columns in a dataset |
| [Sweep Clustering](https://msdn.microsoft.com/en-us/library/azure/mt484327.aspx) | Performs a parameter sweep on a clustering model to determine the optimum parameter settings |
| [Test Hypothesis Using t-Test](https://msdn.microsoft.com/en-us/library/azure/dn905917.aspx) | Compares means from two datasets using a t-test |
| [Threshold Filter](https://msdn.microsoft.com/en-us/library/azure/dn905843.aspx) | Creates a threshold filter that constrains values |
| [Time Series Anomaly Detection](https://msdn.microsoft.com/en-us/library/azure/mt775197.aspx) | Learns a trend in time series data and uses it to detect anomalies |
| [Train Anomaly Detection Model](https://msdn.microsoft.com/en-us/library/azure/dn913097.aspx) | Trains an anomaly detector model and labels data from the training set |
| [Train Clustering Model](https://msdn.microsoft.com/en-us/library/azure/dn905873.aspx) | Trains a clustering model and assigns data from the training set to clusters |
| [Train Matchbox Recommender](https://msdn.microsoft.com/en-us/library/azure/dn905987.aspx) | Trains a Bayesian recommender using the Matchbox algorithm |
| [Train Model](https://msdn.microsoft.com/en-us/library/azure/dn906044.aspx) | Trains a classification or regression model in a supervised manner |
| [Train Vowpal Wabbit 7-4 Model](https://msdn.microsoft.com/en-us/library/azure/dn905861.aspx) | Trains a model from the Vowpal Wabbit machine learning system  This module is for compatibility with VW versions 7-4 and 7-6 |
| [Train Vowpal Wabbit 7-10 Model](https://msdn.microsoft.com/en-us/library/azure/mt674683.aspx) | Trains a model from the Vowpal Wabbit machine learning system  This module is for the current VW version, 7-10 |
| [Train Vowpal Wabbit 8 Model](https://msdn.microsoft.com/en-us/library/azure/mt712721.aspx) | Trains a model using version 8 of the Vowpal Wabbit machine learning system  This module is for the VW version 8 |
| [Tune Model Hyperparameters](https://msdn.microsoft.com/en-us/library/azure/dn905810.aspx) | Performs a parameter sweep on a regression or classification model to determine the optimum parameter settings |
| [Two-Class Averaged Perceptron](https://msdn.microsoft.com/en-us/library/azure/dn906036.aspx) | Creates an averaged perceptron binary classification model |
| [Two-Class Bayes Point Machine](https://msdn.microsoft.com/en-us/library/azure/dn905930.aspx) | Creates a Bayes point machine binary classification model |
| [Two-Class Boosted Decision Tree](https://msdn.microsoft.com/en-us/library/azure/dn906025.aspx) | Creates a binary classifier using a boosted decision tree algorithm |
| [Two-Class Decision Forest](https://msdn.microsoft.com/en-us/library/azure/dn906008.aspx) | Creates a two-class classification model using the decision forest algorithm |
| [Two-Class Decision Jungle](https://msdn.microsoft.com/en-us/library/azure/dn905976.aspx) | Creates a two-class classification model using the decision jungle algorithm |
| [Two-Class Locally Deep Support Vector Machine](https://msdn.microsoft.com/en-us/library/azure/dn913070.aspx) | Creates a binary classification model using the locally deep Support Vector Machine algorithm |
| [Two-Class Logistic Regression](https://msdn.microsoft.com/en-us/library/azure/dn905994.aspx) | Creates a two-class logistic regression model |
| [Two-Class Neural Network](https://msdn.microsoft.com/en-us/library/azure/dn905947.aspx) | Creates a binary classifier using a neural network algorithm |
| [Two-Class Support Vector Machine](https://msdn.microsoft.com/en-us/library/azure/dn905835.aspx) | Creates a binary classification model using the Support Vector Machine algorithm |
| [Unpack Zipped Datasets](https://msdn.microsoft.com/en-us/library/azure/mt683839.aspx) | Unpacks datasets from a zip package in user storage |
| [User-Defined Filter](https://msdn.microsoft.com/en-us/library/azure/dn905965.aspx) | Creates a custom finite or infinite impulse response filter |

3. Google Refine ( now known as Open Refine) :

* insert into a database or merging data with other data files.
* Open Refine as a powerful data-cleaning tool.
* Encourage dataset exploration; look at the data with the visualization tools in Open Refine.
* Introduce concept of facets
* Show split columns by defined separator
* Show power of include / exclude, sort by name / count
* Show the power of clustering algorithms to reveal data patterns, data snafus
* Allow undo / redo.
* scripting
* regular expressions
* APIs

4. DataScience ToolKit

Coordinates to Statistics

Returns characteristics like population density, elevation, climate, ethnic makeup, and other statistics for points all around the world at a 1km-squared or finer resolution.

#### Text to Sentiment

Tries to guess whether the text represents a roughly positive or negative comment, and returns a score between -5 and 5

#### Text to Times

Searches the input for strings that represent dates or times, and parses them into the standard form of a Ruby date/time string

#### Text to People

Extracts any sequences of words that look like people's names, and tries to guess their gender, likely age, and ethnicity from the statistical properties of any names found.

#### HTML to Story

This call takes an HTML document, and extracts the sections of text that appear to be the main body of a news story, or more generally the long, descriptive passages in any page. This is especially useful when you want to run an analysis only on the unique content of each page, ignoring all the repeated navigation elements.

#### HTML to Text

This call takes an HTML document, and analyzes it to determine what text would be displayed when it is rendered.

#### Text to Sentences

This call takes some text, and returns any fragments of it that look like proper sentences organized into paragraphs.

#### File to Text

If you pass in an image, this API will run an optical character recognition algorithm to extract any words or sentences it can from the picture. If you upload a PDF file, Word document, Excel spreadsheet or HTML file, it will return a plain text version of the content.