Analytics Vidhya
Learn everything about analytics

DATAHACK SUMMIT 2018

Transition from Pandas to Spark Dataframe using Scala

Joydeep Bhattacharjee

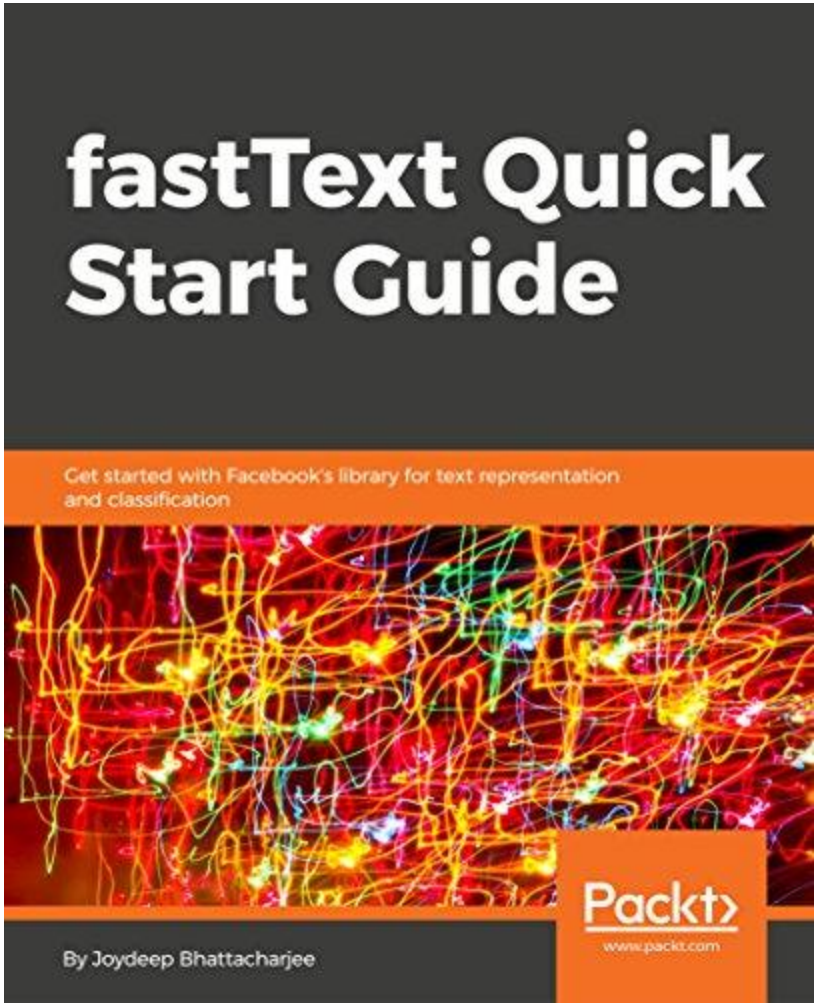
22 – 25 November, 2018

NIMHANS Convention Center, Bengaluru

Joydeep Bhattacharjee

- Machine Learning Engineer at Nineleaps
- https://twitter.com/alt227Joydeep
- https://www.linkedin.com/in/joydeep-bhattacharjee-934a1157/
- https://infinite-joy.github.io/

Book


fastText Quick Start Guide
Get started with Facebook's library for text representation and classification
By Joydeep Bhattacharjee
Packt
www.packt.com

Transition from Pandas to Spark Dataframe using Scala

Encourage usage of common paradigms

What are Dataframes

- Popular now in the data science world.
- Easy SQL like tabular interface for the user.
- Easy mental transitioning between matrix and the features.

Pandas ...

```
house_prices_df = pd.read_csv('../data/house-prices/train.csv')
```

```
melb_data = pd.read_csv('../data/melbourne_housing_snapshot/melb_data.csv')
```

Taking a look at the dataframe.

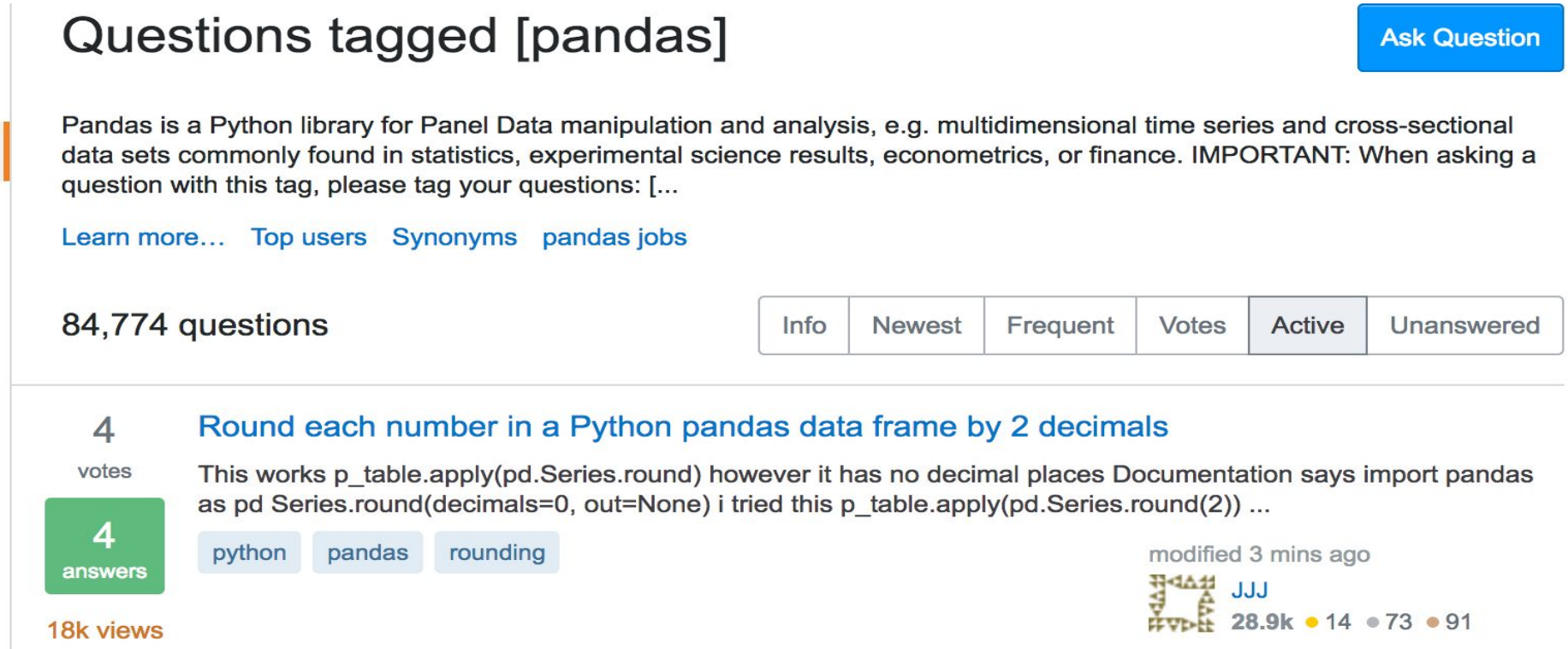
```
house_prices_df.head()
```

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape	LandContour	Utilities	...	PoolArea	PoolQC
0	1	60	RL	65.0	8450	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN
1	2	20	RL	80.0	9600	Pave	NaN	Reg	Lvl	AllPub	...	0	NaN
2	3	60	RL	68.0	11250	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
3	4	70	RL	60.0	9550	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN
4	5	60	RL	84.0	14260	Pave	NaN	IR1	Lvl	AllPub	...	0	NaN

- Good for data gathering and transformations.
- Flexible reshaping and pivoting datasets.
- Easy merges and joins.

Result

It is very easy to write code in pandas.



Questions tagged [pandas]

Ask Question

Pandas is a Python library for Panel Data manipulation and analysis, e.g. multidimensional time series and cross-sectional data sets commonly found in statistics, experimental science results, econometrics, or finance. IMPORTANT: When asking a question with this tag, please tag your questions: [...

Learn more… Top users Synonyms pandas jobs

84,774 questions

Info | Newest | Frequent | Votes | Active | Unanswered

4 votes

4 answers

18k views

Round each number in a Python pandas data frame by 2 decimals

This works p_table.apply(pd.Series.round) however it has no decimal places Documentation says import pandas as pd Series.round(decimals=0, out=None) i tried this p_table.apply(pd.Series.round(2)) ...

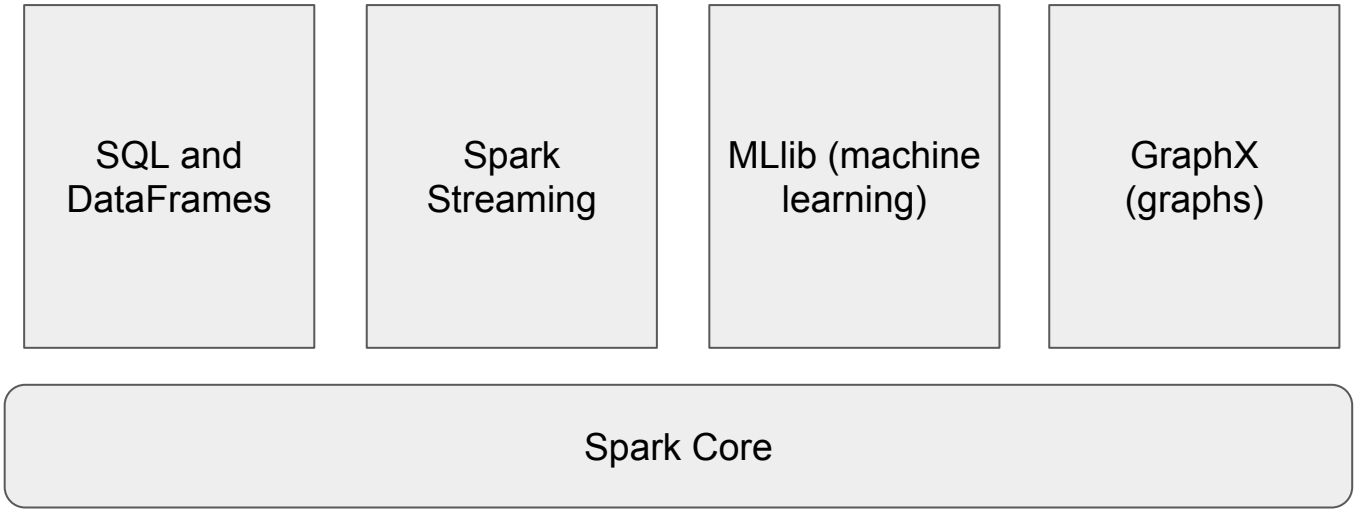
python pandas rounding

modified 3 mins ago

JJJ

28.9k ● 14 ● 73 ● 91

Spark Components



Quick words on some new concepts

- **Spark context**: This is basically an instance of the spark environment and enables your application to be “run”.
- **Sql context**: This is a “sub-environment” for running the Spark SQL functionalities such as running dataframe transformations.
- **scala/python**: You can use either of these two languages to create your spark applications
- **SBT**: Simple Build Tool for your Scala apps.
- **Spark-submit**: submit your code to the execution environment.
- **spark-shell/pyspark**: try out some code here.
- **Transformations and Actions**: spark transformations are generally lazy and dont run until an “action” is called.

Demo and transformations

Head over to the notebooks for -

- Creating dataframes.
- Get the dataframe shape and columns.
- Changing the column names of the dataframes.
- Orderby and groupby.
- Filtering data.
- Membership in dataframe.
- Missing value imputation.
- Discretization and Binning.
- Getting a particular data.
- Reshaping and Pivoting.
- Merges and Joins.
- Function application, transformations and mapping.

Thank You