

《用 Python 玩转数据》之利用免费财经数据接口 TuShare

获取和分析数据

by Dazhuang@NJU

1. 安装

在 Anaconda Prompt 窗口中输入如下命令安装：

```
> pip install tushare
```

2. 介绍

"TuShare 是一个免费、开源的 python 财经数据接口包。主要实现对股票等金融数据从数据采集、清洗加工 到 数据存储的过程，能够为金融分析人员提供快速、整洁、和多样的便于分析的数据，为他们在数据获取方面极大地减轻工作量，使他们更加专注于策略和模型的研究与实现上。考虑到 Python pandas 包在金融量化分析中体现出的优势，TuShare 返回的绝大部分的数据格式都是 pandas DataFrame 类型，非常便于用 pandas/NumPy/Matplotlib 进行数据分析和可视化。当然，如果您习惯了用 Excel 或者关系型数据库做分析，您也可以通过 TuShare 的数据存储功能，将数据全部保存到本地后进行分析。"这是 TuShare 官网(<http://tushare.org/index.html>)上对于 TuShare 的描述，它提供了便捷的各类财经数据和新闻等的接口。

3. 简单示例

例如要想获取股票代码是 600848 的股票在 2018 年 3 月 1 日至 3 月 10 日间的基本历史数据，只要使用如下代码即可：

```
>>> import tushare as ts
>>> ts.get_hist_data('600848',start='2018-03-01',end='2018-03-08')
```

	open	high	close	low	volume
date					
2018-03-08	23.82	24.10	23.99	23.71	34416.47
2018-03-07	24.06	24.28	23.90	23.89	43007.21
2018-03-06	24.20	24.28	24.03	23.82	48066.02
2018-03-05	24.48	24.48	24.13	23.91	37519.75
2018-03-02	23.90	24.39	24.33	23.63	61194.21
2018-03-01	23.69	24.48	24.18	23.56	46819.00

...

get_hist_data()函数可以获取三年内 A 股历史行情，其他 Tushare 中功能相似的函数还有 get_h_data()和 get_k_data()。

提示：如果要做正式发表的研究，数据尽量要与权威的财经网站比对核对。

小项目任务：

利用 Tushare 包中的接口函数获取招商银行（股票代码 600036）2018 年下半年的股票数据并完成如下数据处理和分析任务：

1. 数据只保留 date、open、high、close、low 和 volume 这几个属性，并按时间先后顺序对数据进行排序；
2. 输出这半年内成交量最低和最高那两天的日期和分别的成交量；
3. 列出成交量在 1000000 以上的记录；

4. 计算这半年中收盘价 (close) 高于开盘价 (open) 的天数;
5. 计算前后两天开盘价的涨跌情况, 用两种方式表示, 第一种输出每两天之间的差值 (后一天减去前一天), 第二种输出一个开盘价涨跌列表, 涨用 1 表示, 跌用-1 表示;
6. 计算每月收盘价的平均值。

【参考程序】

```
import tushare as ts
import numpy as np
```

```
# 1
```

```
df = ts.get_hist_data('600036', start = '2018-07-01', end = '2018-12-31')
```

```
df = df.iloc[:, :5] # 获取前 5 列
```

```
df.sort_index(inplace = True) # 按 date 列进行排序
```

```
# 2
```

```
min_day = df.sort_values('volume').iloc[0,]
```

```
min_volume = min_day.volume
```

```
min_volume_date = min_day.name
```

```
print("the min volume of {} is at {}".format(min_volume, min_volume_date))
```

```
max_day = df.sort_values('volume').iloc[-1,]
```

```
max_volume = max_day.volume
```

```
max_volume_date = max_day.name
```

```
print("the max volume of {} is at {}".format(max_volume, max_volume_date))
```

```
# 3
```

```
print(df[df.volume >= 1000000])
```

```
# 4
```

```
print(len(df[df.close > df.open]))
```

```
# 5
```

```
print(df.open.diff())
```

```
print(np.sign(np.diff(df.open)))
```

```
# 6
```

```
month = [item[5:7] for item in df.index]
```

```
print(df.close.groupby(month).apply(np.mean))
```