驰骋股市！手把手教你如何用Python和数据科学赚钱？

大数据文摘 2 days ago

大数据文摘出品

编译：胡笳、Aileen

金融领域或许是数据科学应用场景中最充满想象力的部分，毕竟它跟财富结合地无比紧密。

不管是否是经济达人，数据科学都是一种帮你了解一支股票的高效方式。本文作者把数据科学和机器学习技术应用到金融领域中，向你展示如何通过数据分析的方式驰骋股市，搭建自己的金融模型！

让我们先了解一些基本定义。

定义和假设

什么是交易算法？

Quantopian定义：
交易算法是一种计算机程序，它定义了一套买卖资产的规则。大多数交易算法基于研究历史数据得出的数学或统计模型来做出决策。

我们使用什么平台？

我使用Anaconda, Jupyter Notebooks, 和 PyCharm实现Python建模，使用这些工具非常容易。但是，你也可以使用Quantopian平台内置内核工具，或者甚至可以根据需要将代码修改为R或者其他语言。

我使用Mac系统，并将全程分享所用的UNIX命令。Windows用户请自行搜索答案！

我们关注哪些资产？

Apple苹果（AAPL）是一支很好的股票，因为目前为止（2018年9月）它已经是世界上价值最高的公司，不仅拥有相对稳定的股票价格，而且拥有足够多与品牌相关的体量、新闻和人气。

需要提醒：此处涵盖的原则对于较小的公司股本，或不同的行业等的适用性有所不同。

环境搭建

要在本地电脑上获取Quantopian平台，请在终端执行以下命令：

```
# create conda py35 since that's the newest version that works
conda create -n py35 python=3.5
```

```
conda install -c quantopian/label/ci -c quantopian zipline
```

为了确保Quandl正常运行，请根据账号创建说明和API文档加载金融数据。另外，请保存好你的API key，因为需要用到它来加载所有重要数据。

加载数据

让我们开始使用代码库：

```
import pandas as pd
import numpy as np
import patsy

pd.core.common.is_list_like = pd.api.types.is_list_like
from pandas_datareader import data
import quandl
quandl.ApiConfig.api_key = "###############"
```

现在让我们来拉取些Apple股票数据：

```
df = quandl.get("WIKI/" + 'AAPL', start_date="2014-01-01")
```

注意观察这些列，注意其中一个名为“分割比例”的列。这是一个非常重要的指标；它标志着股票拆分发生。在2014年，Apple决定采用7:1进行股票分割，我们可以使用Python 和pandas 来查询发生的日期：

```
len(df)
df['Split Ratio'].value_counts()
df[df['Split Ratio'] == 7.0]
```

我们从而找到了2014-06-09。让我们拉取这个日期后的股票价格信息：

```
aapl_split = quandl.get("WIKI/" + 'AAPL', start_date="2014-06-10")
aapl_split.head()
```

顺便说一句，我在GitHub上找到了所有财富500强的股票代码清单。如果你想将自己的分析扩展到股票集，可以像这样将它们加载到列表中：

```
f500 = pd.read_csv('https://raw.githubusercontent.com/datasets/s-and-p-500-companies/master/dat

tickers = f500.Symbol.tolist()
```

关键统计数据

增广迪基-福勒检验(Augmented Dickey-Fuller test),简称ADF检验。

我们需要检验单位根是否存在，可以使用ADF测试完成检验。简而言之，单位根存在则预示存在驱动AAPL的潜在趋势，从而我们可以提取模式并用于预测。

```
# run ADF to determine unit root
import statsmodels.tsa.stattools as ts
cadf = ts.adfuller(aapl_split.Close)
```

```
print('Augmented Dickey Fuller:')
print('Test Statistic =',cadf[0])
print('p-value =',cadf[1])
print('Critical Values =',cadf[4])
```

```
Augmented Dickey Fuller:
Test Statistic = -0.731194982176

p-value = 0.838503045276
Critical Values = {'1%': -3.4372231474483499, '5%': -2.8645743628401763, '10%': -2.568385665036
```

我们将上面的测试统计值与临界值进行比较；如果它低于我们选择的阈值，则拒绝存在单位根的零假设。正如你所见，p-value比较大，所以我们必须接受原假设(Ho)：即AAPL存在单位根。这个结果很好，因为我们可以利用潜在的趋势和模式进行预测。

与其他股票的相关性

Apple被认为是一个巨头技术品牌。假如我们能够计算与其他股票的强相关性会怎么样？

请注意相关性并不意味着因果关系，并且可能存在着哪个股票是先行者的问题，但是模式和关系对于提高模型性能总是一件好事。

我建议你查看三支股票，以及AAPL如何与它们关联：

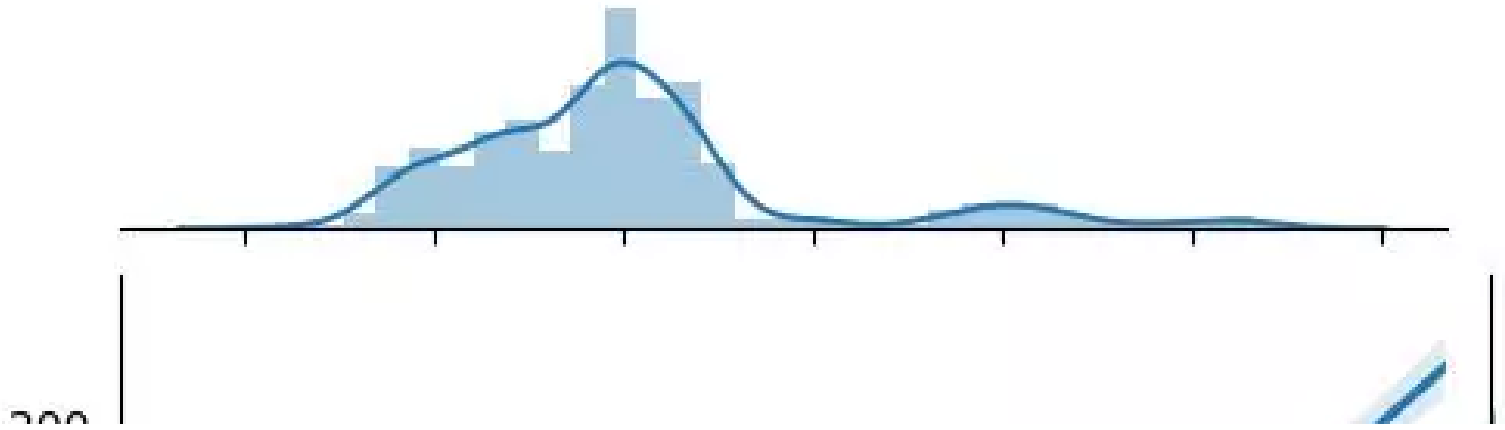
- 微软Microsoft (MSFT)
- 因特尔Intel (INTC)
- 蒂芙尼Tiffany & Co. (TIF)

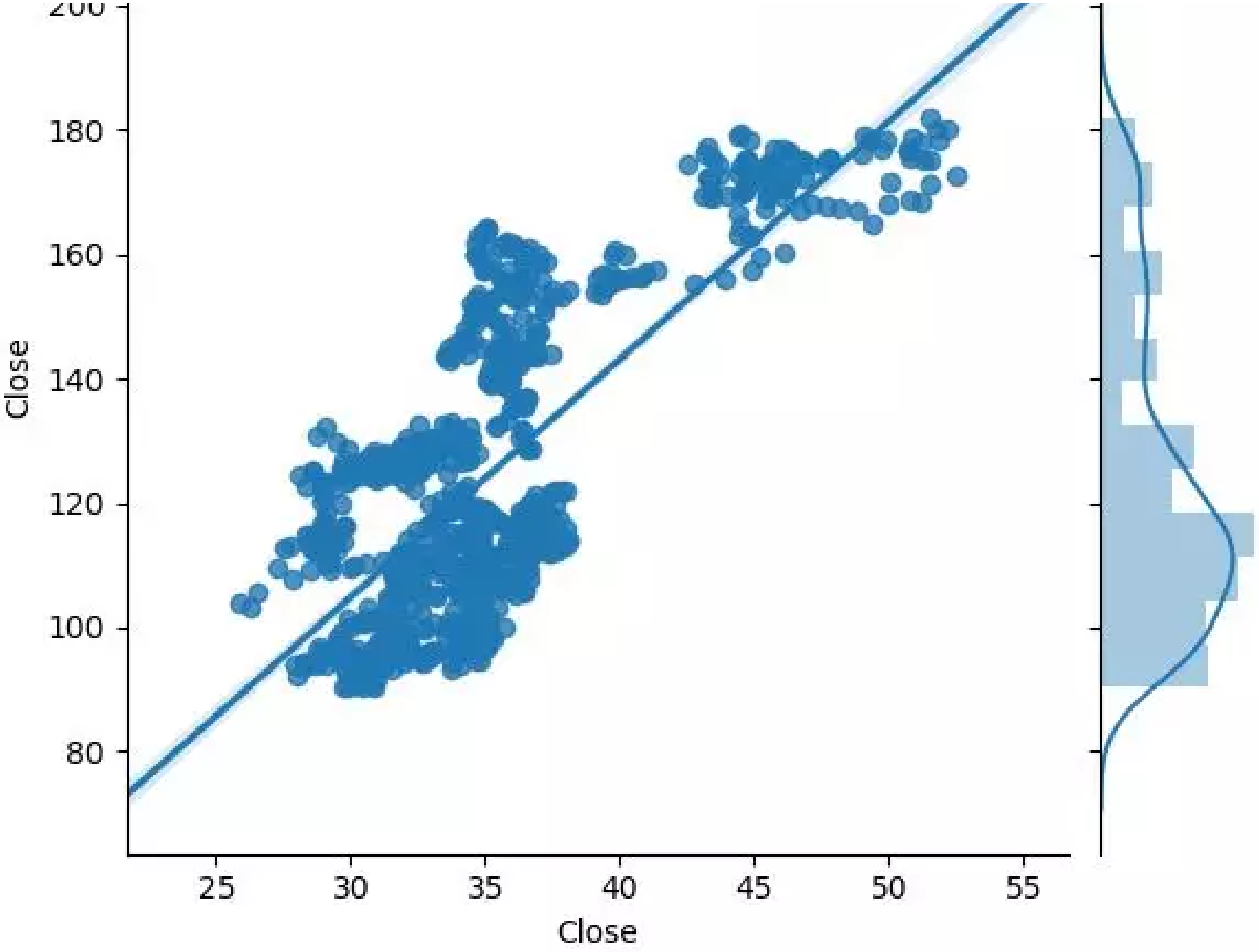
```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline

MSFT = quandl.get("WIKI/" + 'MSFT', start_date="2014-06-10")
INTC = quandl.get("WIKI/" + 'INTC', start_date="2014-06-10")
TIF = quandl.get("WIKI/" + 'TIF', start_date="2014-06-10")
```

时间缘故，我们在这里只关注Intel数据；让我们绘制AAPL与INTC的收盘价：

```
sns.jointplot(INTC.Close, aapl_split.Close, kind="reg");
```





Intel vs. Apple

我们还可以看看相关值数据(correlation value)：

```
np.corrcoef(INTC.Close, aapl_split.Close)
```

我们注意到r-value为0.7434；就预测来说不错，但我们需要记住一个重要的事实：如果我们知道INTC的收盘价，我们也可以看查AAPL的收盘价。所以让我们查看下INTC七天前的收盘价的相关性，来获得更可行的指标：

```
# seven day lead
np.corrcoef(INTC.Close[:-7], aapl_split.Close[7:])
```

这次我们得到r-value为0.7332；还是很不错的！

谷歌趋势（Google Trends）

我们可以比较Twitter和其他社交网络人气数据如何影响股价。现在让我们看看Google Trends是否与AAPL相关。请确保为指定时间范围，或使用此链接（https://trends.google.com/trends/explore?date=2014-06-10%202018-04-02&q=%2Fm%2F0k8z）来进行准确搜索（注意我在四月多添加了几天来处理半周问题），然后将CSV加载到Python中：

```
# Google Trends
```

```
# Google Trends

aapl_trends = pd.read_csv('/Users/jb/Desktop/multiTimeline.csv', header=1)

aapl_trends.tail()
```

注意每周的数据格式，因此我们需要使用 pandas.resample() 转换我们的股票价格数据集：

API链接：
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.resample.html

```
aapl_split_week = aapl_split.resample('W', convention='end').last()
```

现在让我们检查相关性并绘制给定周的Google搜索请求数据总和图表，以及该周最后一个工作日的收盘价格：

```
# trend and price corr
np.corrcoef(aapl_trends['Apple: (Worldwide)'], aapl_split_week.Close)
```

哎呀！我们得到了一个微不足道的0.0454，这个数据有些道理，我们可以想一下：AAPL相关的新闻/活动/闲谈并不是影响股票价格的积极因素。像人气这种对其有重要影响应该能提供更强的信号，但我们在下次再做讨论。

结语

我们只是浅显的讨论了下部分EDA（ Exploratory Data Analysis）可以做的金融分析，但是在下一篇文章中，我们将过渡到建立预测模型并通过高级软件包来为我们实现繁重的工作。

我们希望这篇文章对你有帮助，并且很乐意在评论中听到你的意见：

运行代码是否遇到任何问题？有时候环境和版本会搞砸一切……

- 你使用什么包和技术？
- 那些可视化工具有助于了解股票价格的变动？
- 你认为哪些因素会最大化模型预测效果？

最后，如果你恰好知道一种持续赚大钱的建模技术，请直接告诉我们！如果你们喜欢这个系列，请持续关注大数据文摘的后续文章。

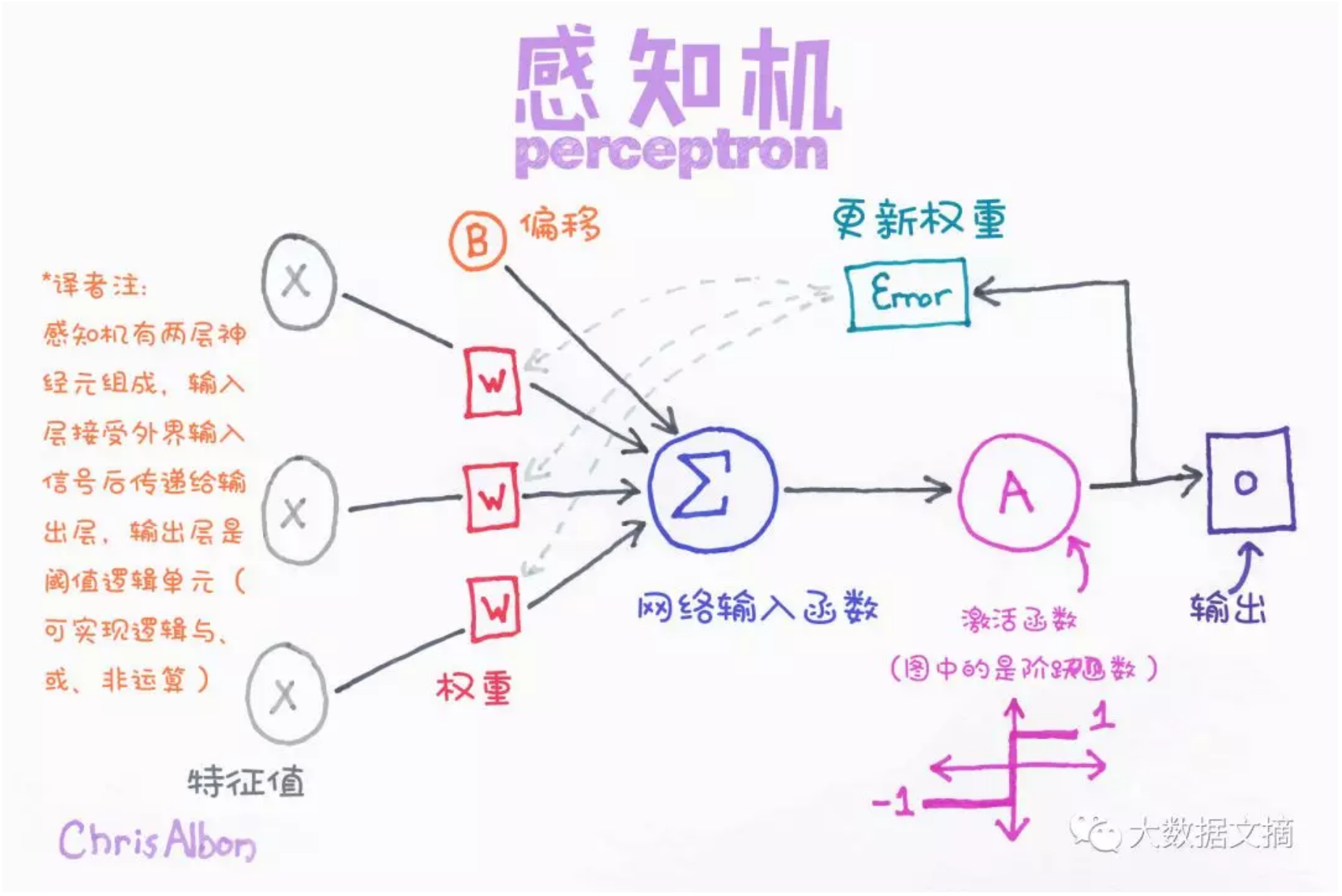
最后的最后，本文提供的信息和随附材料仅供参考。本文不应被当做法律或财务建议。你应咨询律师或其他专业人士来确定什么最适合你的个人需求。

祝大家赚钱开心！

相关报道：
https://towardsdatascience.com/on-making-money-with-python-and-data-science-1-setup-and-statistics-1d69f1a68661?from=singlemessage&isappinstalled=0

【今日机器学习概念】

Have a Great Definition



志愿者介绍

回复“志愿者”加入我们

胡笳

IT攻城狮一枚。好奇心浓郁，爱喝饮料，喜欢音乐以及给别人推荐音乐。希望能多做点有意思的事儿。自称为笳yi大王。开心就好，其他没什么大不了。