组员分工：

曹润泽：爬取**中国证券网**股票数据
王　阳：爬取**金融界**股票数据
赵　航：爬取**东方财富网**股票数据
黄恺晟：爬取 **tushare** 股票数据

数据源：

本次数据集成集成了 4 个数据源，分别来自**中国证券网、金融界、东方财富网和 tushare 接口。**主要就是让各个数据源的数据相互补充，有些在一个数据源上得不到的数据可以通过另一个数据源来获得，同时，有些相互冲突的数据可以通过集成更多的数据进行约束整理。

获取方式：

其中，中国证券网、金融界、东方财富网我们采用传统的代码爬取方法，即 bs4+request 框架从网页中爬取。而 tushare 是 python 自带的一个 api 接口，我们可以非常方便的获取股票信息，更多的是包含动态信息。

由于股票信息不只是停留在一个网页，因此中间可能包含多个子网页，所以常规的爬取可能比较慢，效率低下。（中国证券网、金融界）

同时有些实时数据在网页中无法获得，比如今日开盘这些信息等等。常规操作只能获得静态数据，比如公司名称、注册资本等等。

关键代码：

因此要突破限制，对于实时数据要找到其对应的 api 接口，这样就可以直接进行数据对接，而不需要解析 html 文本，从而实现更快速的数据爬取。（东方财富网）

```
# 主要指标：http://f10.eastmoney.com/NewFinanceAnalysis/MainTargetAjax?type=0&code=SZ000001
# 资产负债表：http://f10.eastmoney.com/NewFinanceAnalysis/zcfzbAjax?companyType=4&reportDateType=0&reportType=1&endDate=&code=SZ000001
# 利润表：http://f10.eastmoney.com/NewFinanceAnalysis/lrbAjax?companyType=4&reportDateType=0&reportType=1&endDate=&code=SZ000001
# 现金流表：http://f10.eastmoney.com/NewFinanceAnalysis/xjllbAjax?companyType=4&reportDateType=0&reportType=1&endDate=&code=SZ000001
# 股权变动：http://dcfm.eastmoney.com/em_mutisvcexpandinterface/api/js/get?type=GBJG&token=70f12f2f4f091e459a279469fe49eca5&ps=3&st=CHANG
```

同时，这里提供了爬取网页的另一种方法，中国证券网和金融界都是通过标签子元素来获取数据，东方财富网是通过正则表达式进行的：

```
result1 = [str(item)[15:] for item in re.compile(r'REPORTDATE\\":\\"[^\\]*').findall(r_zcfzb_text)]
result2 = [str(item)[13:] for item in re.compile(r'SUMASSET\\":\\"\d*').findall(r_zcfzb_text)]
result3 = [str(item)[12:] for item in re.compile(r'SUMLIAB\\":\\"\d*').findall(r_zcfzb_text)]
```