**任务序号：1（没有与组交互，但是也是一种集成，而且构建了知识图谱）**

**组员分工：**

曹润泽：框架搭建，前端设计，echart构图

王 阳：数据清洗，数据提取

赵 航：前后端交互逻辑处理

黄恺晟：数据清洗，数据合并

项目地址：

**数据源：**

数据源来自第二阶段我们四个人分别爬取的数据，即中国证券网、金融界、东方财富网和Tushare的数据。

**集成过程与方法：**

即手工将这四个数据源中其中的三张表进行合并，相互补充、相互约束，来提取关键信息。

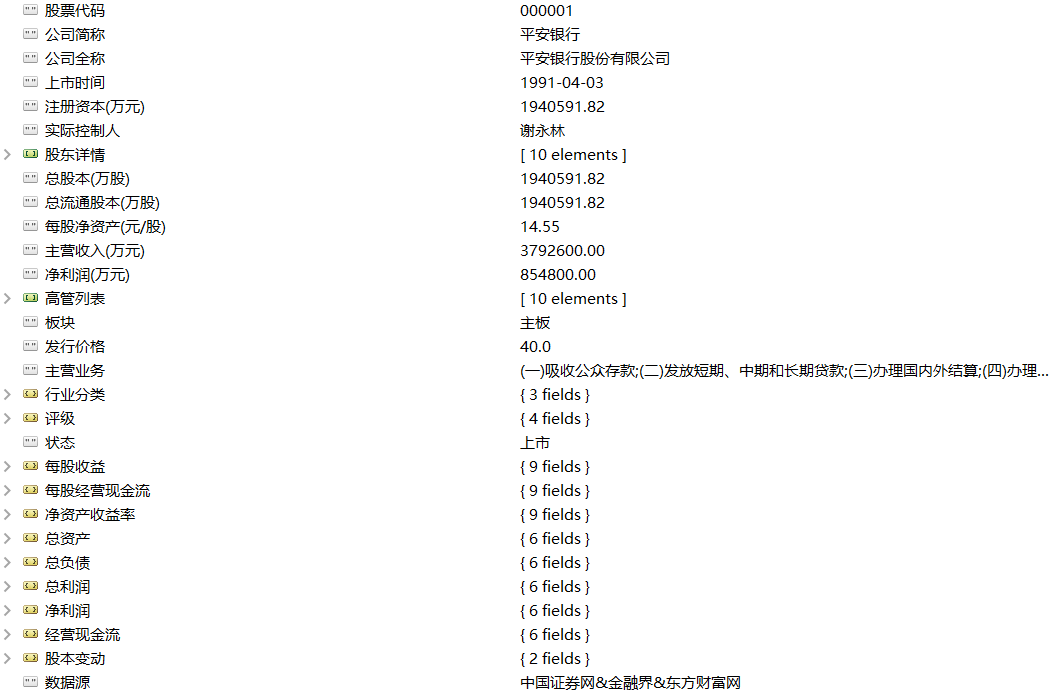
通过多重字典数据分割提取来获得需要的信息，相关代码在connect.py中。

详情如下：



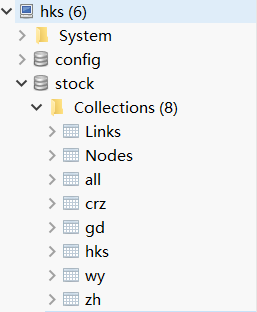
代码较为笨拙，简言之，就是将不同数据源的部分字段进行结合添加。

之前的三个数据源的字段已经给出，集成之后的总字段为：



这就是我们最终得到的单条数据信息。

**结果数据库地址与描述：**

与之前的四个数据源的数据放到一块，地址同样是：

地址为 47.100.220.26，端口为 27017，账号密码都是 mongoadmin。

解释：

这里的all为最终集成的数据，也就是上面给出的例子。

Crz、wy、zh为三人爬取的三个网站的数据，集成来源。

Hks是黄恺晟从tushare接口上获取的实时数据。

Gd是数据清洗过程中的产物，最终被抛弃。

Links和Nodes分别是知识图谱需要的边和点数据。（最终也被抛弃，因为数据量和mongodb读取等问题，最终这两个数据被txt文件替代。）