股票数据及股票新闻数据抓取及分析预测

**由于6号有考试，还有部分文档代码没上传，6号上午考完后再整理上传**。

一数据获取

先在东方财务网获取当前上市公司所有名称及股市代码。再分别进入每家上市公司对应的股吧获取关注人数及每日文章发表量阅读量及评论量。用爬出来的数据算出每支股票的热度及热度时间段等进而预测股票走势，并通过api获取股票上市以来的所有数据并进行验证。

二所用技术

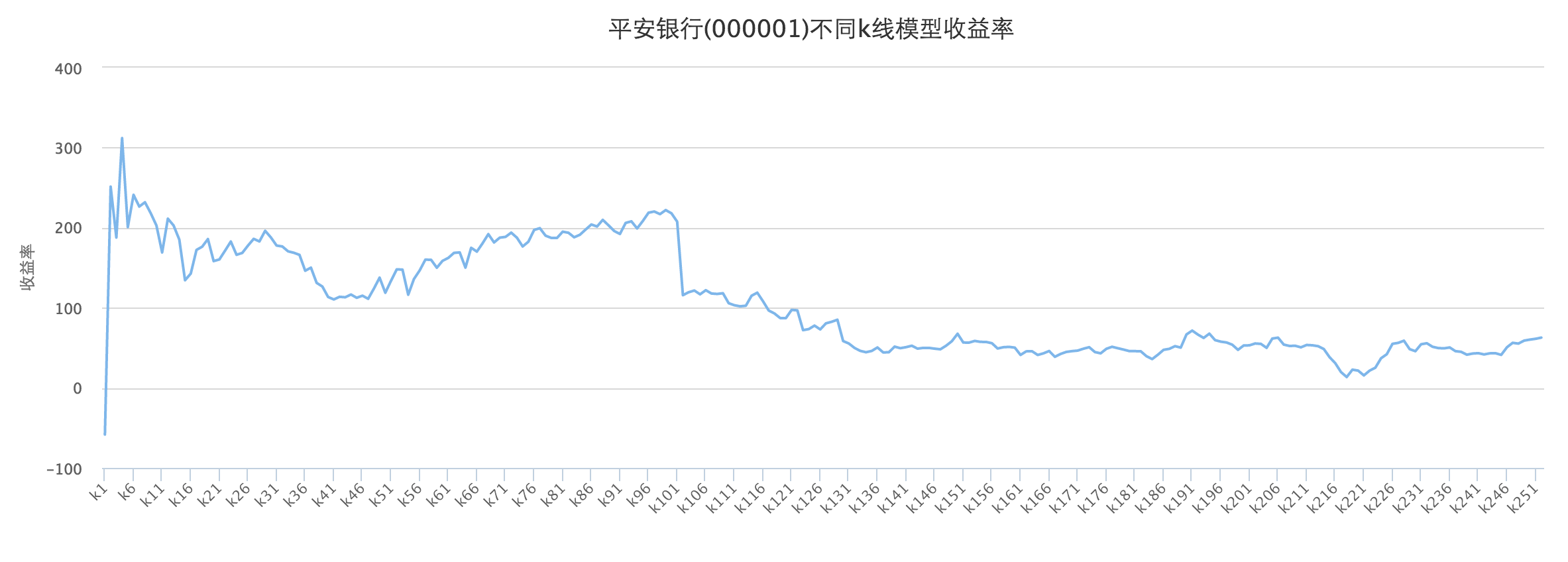
爬出采用webmagic框架，进行数据爬取。对爬出来的数据用hadoop平台进行数据统计排序处理。最后抽查某些热门股票进行统计分析。

三 数据分析

建立k线收益模型，即若当日收盘价大于k线值时则买入(若已入手则继续持有),若当日收盘价低于k线值时则卖出(若已卖出则继续等待机会).用历史数据即可跑出不同k线模型的收益率为多少。

四 实验分析

本实验以平安银行(000001)为例进行252(一年有252个交易日)个不同k线模型数据分析，结果如下。



五 结论

即采用k＝4时，收益率最高能达到300%。最差为k＝1，收益率为负。

故进行投资时建议采用k＝4时操作模型。