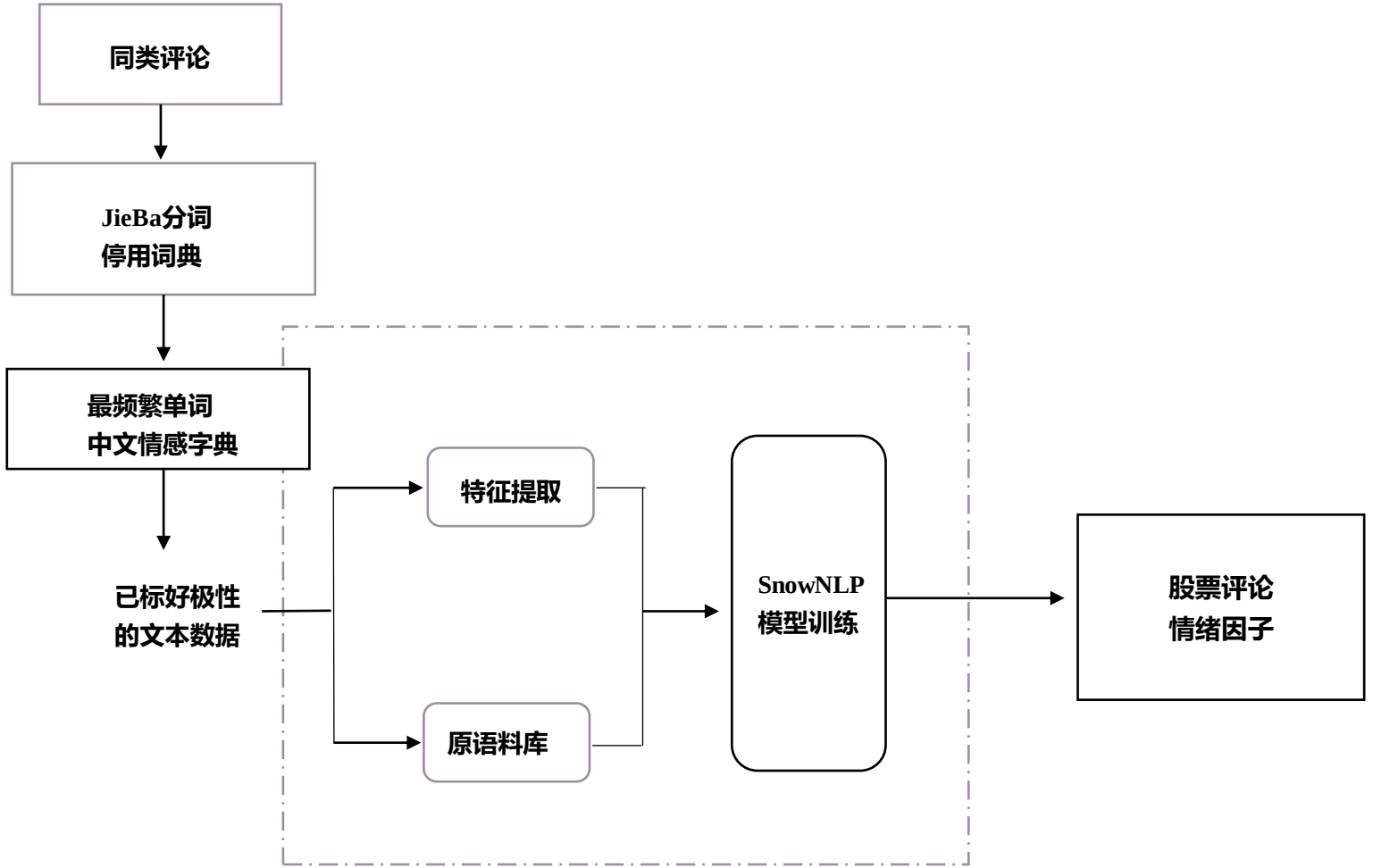
特征工程搭建

KK-Kris



对数据进行情感词典分类

用JieBa库使用已经准备好的停止词库，获取同类股票评论在去除停止词后的前1000个频繁单词。
后对最频繁单词手动标注词性结合中文金融情感词典对同类的评论进行分类。

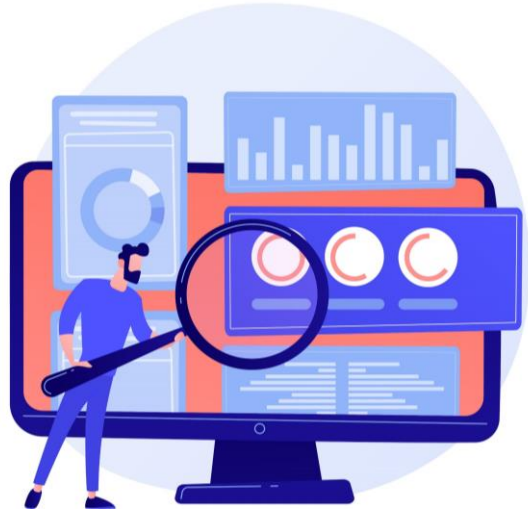
训练SnowNLP库中情感模型

将分类好的同类评论数据存入语料库中，同时为了减弱因词典错误分类造成的影响，与SnowNLP原有语料库进行结合，从而训练出新的情感模型。

分类评论，获取情绪因子

使用新训练的情感模型，对研究的个股评论数据进行模型分类，从而获取出此股票评论的情绪指数，作为预测的一大重要指标。

总结思路



- **本项目基于投资者情绪对股票价格进行了预测，发现在添加情绪指数后股价预测模型是有较为明显的效果提升**

思路如下：

1 数据获取

本文使用东方财富Chioce金融终端获取了君实生物60分钟线，获取了2021年7月5日至2022年8月12日共271个交易日，每天10:30，11:30，14:00，15:00四个时间段共1084个时间点的股票数据：开盘价，最高价，最低价，收盘价，成交量。

后使用Java语言爬取了这271个交易日在东方财富股吧的评论数据。以及同类股票（如同在科创板或同为生物股或医药股）的股吧评论数据共60000条以便进行情绪分析。

2 情绪分析

对同类股票评论使用JieBa分词，获取最频繁1000个单词，手动标明词性，结合中文金融情感词典对同类股票评论数据进行分类，后与SnowNLP库中原有评论数据结合训练模型。

对君实生物这只股票评论进行模型分类，并给出处于[0，1]之间情绪指数。

3 特征工程

每个时间点数据采用其前两个小时之间评论的指数评论值，以判断此时间点的股票的情绪。并将情绪指数调整到[-0.5，0.5]以便更好区分情绪性质。

情绪指数与其他技术性指标进行相关性分析

与原技术性指标相结合搭建新的股票特征工程。

4 模型预测

对构建出的特征工程采用已经搭建好的模型进行预测。

将带有情绪指标的股票数据，与纯技术性指标的数据进行对比预测，以探究出股票的情绪指数对股价预测的影响作用。

对此项目进行总结

KK-Kris