**基于文本挖掘的投资者情绪对股票收益率的影响研究**

**文本挖掘** 是指从大量 文本数据 中抽取事先未知的、可理解的、最终可用的知识的过程，同时运用这些知识更好地组织信息以便将来参考。 文本挖掘的主要用途是从原本未经处理的文本中提取出未知的知识，但是文本挖掘也是一项非常困难的工作,因为它必须处理那些本来就模糊而且非结构化的文本数据,所以它是一个多学科混杂的领域,涵盖了 信息技术 、文本分析、 模式识别 、 统计学 、 数据可视化 、 数据库技术 、 机器学习 以及数据挖掘等 技术 。文本挖掘是从数据挖掘发展而来，因此其 定义 与我们熟知的数据挖掘 定义相类似。

**投资者情绪** 是指投资者对未来预期的系统性偏差称作投资者情绪。 投资者情绪是个难以度量的概念，它反映了市场参与者的投资意愿或者预期。 投资者能感觉到它的客观存在，但是要问它到底有多高、近期发生了何种变化，每个个体的投资者有因为有持仓、风格、财富、地位等因素的不同，给出不同的答案。

**词典法** 看这个词就能大概知道这个方法的逻辑。 但是再深入想一想就可以知道其实没那么简单。 个人觉得，以下几个方面就会对该算法造成很大的困难： 分词. 举个例子就很好理解了。 比如“好坏”这两个字。 如果把它理解成两个形容词，‘好’，‘坏’，那么这就是一个 正面的词+一个负面的词；如果把它理解成一个修饰形容词的副词和另一个被修饰的形容词。 那么这就是 一个 负面的词。 那么‘好坏’到底是很坏的意思还是一个中心词呢？ 语义转折. 大部分语义转折出现于 讽刺。

举个例子“我简直太喜欢这个手机了，用了两天就坏了，好开心。 ” 这种语义转折的情况，如果用词典法，会发现这全是“喜欢”，‘开心”这种正面的词。 那么词典法具体是怎么样的呢？ 基础版本真的是简单到不能再简单了。

**中文分词(**Chinese Word Segmentation) 指的是将一个汉字序列切分成一个一个单独的词。分词就是将连续的字序列按照一定的规范重新组合成词序列的过程。

基于词典的方法、基于统计的方法、基于规则的方法

基于词典的方法（[字符串](https://so.csdn.net/so/search?q=%E5%AD%97%E7%AC%A6%E4%B8%B2&spm=1001.2101.3001.7020" \t "https://blog.csdn.net/u011402896/article/details/_blank)匹配，机械分词方法）   
定义:按照一定策略将待分析的汉字串与一个“**大机器词典**”中的词条进行匹配，若在词典中找到某个字符串，则匹配成功。

例子

长春市\_长春\_药店\_

长春市\_长\_春药\_

长春\_市长\_春药\_

长春\_市\_长春\_

长\_春\_市长\_

**LSTM**

长短期记忆（LSTM）单位是递归神经网络（RNN）的单位。由LSTM单元组成的RNN通常称为LSTM网络（或仅称为LSTM）。公共LSTM单元由单元，输入门，输出门和忘记门组成。该单元记住任意时间间隔内的值，并且三个门控制进出单元的信息流。

LSTM网络非常适合基于时间序列数据进行分类，处理和预测，因为在时间序列中的重要事件之间可能存在未知持续时间的滞后。开发LSTM是为了处理在训练传统RNN时可能遇到的爆炸和消失的梯度问题。对于间隙长度的相对不敏感性是LSTM相对于RNN，隐马尔可夫模型和其他序列学习方法在许多应用中的优势。

**噪声交易理论**（Noise Deal），在 有效市场假说 中，证券价格与价值之间存在着一个偏差，这个偏差就是噪声，由于 证券市场 中的交易者依靠信息对交易品种的价值作出判断，尽管价值是先于价格的产生而客观存在的，但在实际交易过程中对价值的判断表现为一种“共识”，“价值发现”就是对价值的“共识”达成一致的行为，共识的达成受到众多的投资者行为的影响，所以这种一致的认识即价值判断的标准无法实现静态的均衡，同时无法先验地判断哪些信息与价值有关，哪些信息与价值无关。 于是价格与价值之间就会出现偏差，噪声由此产生。

**日内效应** 主要指：由于金融市场的交易机制限制（如交易时间、竞价方式的限制导致无法交易）和市场信息的消化、吸收和释放（如中午休市期间进行信息吸收而下午开始信息释放）而导致的日内波动率和收益率呈现出非平稳趋势，日内的某一或某几个交易时段的平均收益或平均波动水平出现明显不同于其余交易时段的现象。

**周内效应** 是指一周中某一个交易日金融市场的平均收益或平均波动水平明显不同于其余交易日，并且这种效应持续存在的现象。

**牛市** 的定义是市场在一段时间内大幅上涨。随着市场开始上涨，股票市场变得越来越贪婪。你看到越来越多的人在想，“哦，是的，让我们把钱投入市场，因为它正在上升。”

**熊市**

牛市的特点是普遍的乐观情绪和积极的增长，这往往会催化贪婪。熊市与普遍的下跌感有关，这往往会在股东的心中灌输恐惧。作为规则第一的投资者，我们采取与投资公众相反的方式 – 当涉及到牛市与熊市时 – 并通过在熊市期间以低价找到优质股票并在牛市中重新获得其价值时出售来充分利用其情绪。