电影评论情绪分析

小组成员：
陈畅科 2006500024
张俊宇 2006500029
张达灿 2006500030

研究背景：

互联网的迅速发展以及通信工具的兴起，导致网络用户的信息交互渠道大量增加。网络用户通过各种方式来表达自己对热点事件的观点，这使得在互联网上充斥着大批的由网民所参与的，对于事物、事件等有重大研究性的评论。但是这些观点以及评论信息大多数都基于个人的主观意见，因此，情感分析的主要目的就是研究如何可以提取与情感相关的信息。伴随着生活水平的逐渐提高以及群众对自己身心的放松，极大多数的人会选择在闲暇时间去观看一场自己喜欢的电影。然而，面对逐渐扩大的电影市场以及众多但质量参差不齐的电影消费者们通常难以抉择，他们对影片的期望值越大往往失望值也越大，花钱看“烂片”的现象不在少数。因此在选择电影之前，消费者们通常会关注已经看过该影片观众的评论，这些评论主要涉及到评论者对电影本身表达的情感信息，以及评论者对电影中的人物态度观点等。但是由于每个人的喜好不同，过度的自我观点会对其他消费者造成潜移默化的影响，极大地提高了对有价值信息的获取难度。所以快速并且有效地获取、处理这些电影的评论是极其重要的

科学问题：

影评情感分析是从海量影评数据中发掘潜在价值数据的重要方法。鉴于传统的基于词典和机器学习的情感分析技术难以对海量的文字进行有效的处理,比如对离散词向量表示忽略词与词之间的语义信息、存在维度灾难、影评情感分析不准确等问题，我们将利用深度学习技术对影评进行深入的研究分类，使其分类效果更优，准确率更高。

Kaggle 链接：Movie Review Sentiment Analysis (Kernels Only) | Kaggle

数据集：

该数据集由制表符分隔的文件组成，包含烂番茄数据集中的短语。为了进行基准测试，保留了训练/测试的拆分，但句子已从其原始顺序中打乱。每个句子都被斯坦福解析器解析成许多短语。每个短语都有一个短语 ID。每个句子都有一个句子 ID。重复的短语（如短词/常用词）仅包含在数据中一次。

文件有 train.tsv、test.tsv

train.tsv 包含短语及其关联的情绪标签。

test.tsv 只包含短语。必须为每个短语分配情绪标签。

情绪标签为：

0 - 负
1 - 有点负
2 - 中性
3 - 有点正

4 - 正

数据集大小为 2.44 MB

实现方法：

使用 BiLSTM 网络。LSTM 的全称是 Long Short-Term Memory。LSTM 由于其设计的特点，非常适合用于对时序数据的建模，如文本数据。BiLSTM 是 Bi-directional Long Short-Term Memory 的缩写，是由前向 LSTM 与后向 LSTM 组合而成。两者在自然语言处理任务中都常被用来建模上下文信息。

为什么选择 Bi-LSTM 呢？将词的表示组合成句子的表示，可以采用相加的方法，即将所有词的表示进行加和，或者取平均等方法，但是这些方法没有考虑到词语在句子中前后顺序。如句子“我不觉得他好”。“不”字是对后面“好”的否定，即该句子的情感极性是贬义。使用 LSTM 模型可以更好的捕捉到较长距离的依赖关系。因为 LSTM 通过训练过程可以学到记忆哪些信息和遗忘哪些信息。但是利用 LSTM 对句子进行建模还存在一个问题：无法编码从后到前的信息。在更细粒度的分类时，如对于强程度的褒义、弱程度的褒义、中性、弱程度的贬义、强程度的贬义的五分类任务需要注意情感词、程度词、否定词之间的交互。举一个例子，“这个餐厅脏得不行，没有隔壁好”，这里的“不行”是对“脏”的程度的一种修饰，通过 BiLSTM 可以更好的捕捉双向的语义依赖。

预期目标：

对不同短语贴上正确的情绪标签，画出损失函数和精度曲线，准确率达到 70%