自然语言处理开题报告

小组成员：
郭宛怡 2006300022 杨君 2006300026 崔楷雄 2006300067

课题名称：
词袋遇见影评
https://www.kaggle.com/competitions/word2vec-nlp-tutorial/overview/part-%201-for-beginners-bag-of-words

题目背景:
情感分析是机器学习中一个具有挑战性的主题。人们用语言表达自己的情感，这些语言经常被讽刺、模棱两可和文字游戏所掩盖，所有这些都可能对人类和计算机产生很大的误导。还有另一个用于电影评论情绪分析的 Kaggle 竞赛
• 词袋模型的经典方法：one-hot(独热编码) TF 编码(词频) TF-IDF 表示法(词频-逆文档频率)

• one-hot 其方法是使用 N 位状态寄存器来对 N 个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。

• TF 核心思想：考虑单个文档中词频的重要性，忽略词序，词义，语境

• TF-IDF 核心思想：如果某个单次在一篇文章中出现的频率明显高于其他文章，则认为该词具有很好的类别区分能力，适合用于分类用以评估词语在一个文档中的重要程度。字词的重要性随着它在单个文档中出现的次数增加，随着它在所有文档中出现的频率下降

选题原因：
词袋模型是自然语言处理中的一个基本模型，它为自然语言处理和文本分析提供了基础和支持。但是词袋模型存在无法反映语义问题，故提出用其他方法完成情感分析的项目，同时希望能够在电影行业的消费者和生产者之间搭建起一座有效沟通的桥梁，促进相关产业的发展

研究内容：
利用谷歌的 Word2Vec 算法进行自然语言处理以及情感分析，对 IMDB 电影评论的情绪进行评分

数据集:
• 一共有三个数据文件

• 标记的数据集由 50,000 条 IMDB 电影评论组成，这些评论是专门为进行情感分析选择的

• 评论的情绪是二元的，这意味着 IMDB 评分< 5 会导致情绪得分为 0，评分>=7 的情绪得分为 1

• 没有一部电影的评论超过 30 条。标记为训练集的 25,000 条评论不包含与 25,000 条评论测试集相同的任何电影。此外，还有另外 50,000 条没有任何评级标签的 IMDB 评论

• labeledTrainData-标记的训练集。该文件以制表符分隔，并具有一个标题行，后跟 25,000 行，其中包含每个评论的 ID、情绪和文本

• testData-测试集。制表符分隔的文件有一个标题行，后跟 25,000 行，其中包含每个审阅的 ID 和文本。我们的任务是预测每个人的情绪

• unlabeledTrainData-没有标签的额外训练集。制表符分隔的文件有一个标题行，后跟 50,000 行，其中包含每个审阅的 ID 和文本

计划使用算法：
• Google 的 World2Vec 深度学习启发算法

• Word2Vec 是一群用来产生词向量的相关模型。这些模型为浅而双层的神经 网络，在 Word2Vec 中词袋模型假设下，词的顺序是不重要的，将次映射为 一个向量，向量可以表示词与词之间的关系

• 它专注于单词的含义，并试图理解单词之间的含义和语义关系，工作方式类似于深度学习方法，但计算效率更高

预计达到目标：
• 对 25,000 条无标记的数据进行二元情感分析预测，1 代表正面评论，0 代表负面评论

• 预测模型的 AUC > = 0.6 (AUC, area under the ROC curve)

科学意义：
1. 词袋模型是自然语言处理中的一个基本模型，它为自然语言处理和文本分析提供了基础和支持。本次课题使用的词嵌入模型将解决词袋模型存在的无法反映语义的问题。词嵌入模型将词汇进行向量化，从而定量的分析和挖掘词汇之间的联系。目前 Word2Vec 已经成为自然语言处理中一个重要的技术，并且在多个领域得到了广泛应用，例如文本分类、语音识别、机器翻译等

2. 通过词嵌入模型完成本次课题影评情感分析，训练出能够对影评情感做出的大致判断，在消费者和生产者之间促成有效的联系，促使影视行业进一步的发展

词袋模型具有一些明显的缺点：

•词与词之间是相互独立的，而现实生活中是相互影响，上下文有联系的

•向量间的距离无法反映语义差异

•特征是离散稀疏的，会造成维度灾难

• 词嵌入模型把一个维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中，每个单词或词组被映射为实数域上的向量，语义相近的词语，通过相同算法提取出来的特征词向量，在向量空间上不能说一定重合，但至少方向上是一致的。词嵌入模型可以将文本通过一个低维向量来表达，语意相似的词在向量空间上也会比较相近，从而解决词袋模型无法反映语义的问题