题目：

使用谷歌的 Word2Vec 来分析电影评论

Bag of Words Meets Bags of Popcorn

Use Google's Word2Vec for movie reviews

成员：

2006500037 罗方彬

2006500039 王 立

题目链接：

https://www.kaggle.com/competitions/word2vec-nlp-tutorial

主要内容：

谷歌的 Word2Vec 是一种深度学习启发的方法，它关注单词的含义。Word2Vec 试图理解单词之间的含义和语义关系。它的工作方式类似于循环神经网络或深度神经网络等深度方法，但计算效率更高。这个教程侧重于用 Word2Vec 进行情感分析。

情感分析在机器学习中是一个具有挑战性的课题。人们用语言表达情感，这种语言往往被讽刺、模糊和文字游戏所掩盖，这些都可能对人和计算机产生误导。这里还有另一个针对电影评论情感分析的 Kaggle 竞赛。在这个教程中，我们探讨了如何将 Word2Vec 应用于类似的问题。

本次题目选择了电影评论情感分析作为示例问题。需要使用 Word2Vec 对电影评论进行情感分析，以确定评论者表达的情感状态是正面、中性还是负面。这个问题的解决对电影产业有重要的意义，能够为观众提供更好的观影体验和建议，同时也能够帮助制片方更好地了解观众反馈和需求。

数据集：

IMDb 电影评论数据集：

由互联网电影数据库（IMDb）提供。该数据集包含 50,000 个来自 IMDb 的电影评论，其中 25,000 个评论用作训练数据集，另外 25,000 个评论用作测试数据集。每个评论都被标记为积极或消极。

这个数据集被广泛用于情感分析的研究中，也是该领域最常用的数据集之一。由于该数据集的标签被认为是准确和可靠的，因此该数据集被用于评估和比较不同情感分析算法和模型的性能。此外，该数据集还包括其他有用的元数据，例如电影名称和发布日期，可以用于进一步的研究和分析。

链接：

http://ai.stanford.edu/~amaas/data/sentiment/

方法：

首先对电影评论数据进行预处理，剔除非文本内容并将文本划分成单词，为每个单词生成唯一的整数编码。接着使用 Word2Vec 模型为每个单词生成对应的词向量，将序列化后的文本数据转化为神经网络可以识别的形式。然后定义深度神经网络模型，包括 LSTM 或 GRU 层在内，并选择适当的损失函数和优化算法，在预先划分好的训练集上进行模型训练，并在测试集上进行模型评估。最后使用模型对新的电影评论进行情感分析预测。

预期目标：

使用 Word2Vec 算法对情感分析建立深度神经网络模型，即通过分析一段文本来确定其中的情感倾向，例如正面、负面或中性情感。