4、Word2Vec模型

Word2Vec是一种将词汇映射到低维向量空间的神经网络模型，最初由Google在2013年提出。其核心思想是将每个词表示为一个固定大小的向量（即词嵌入），这些向量能够捕捉词汇之间的语义关系。在向量空间中，语义相似的词通常会有相似的向量表示。Word2Vec有两种主要的训练方法：CBOW（Continuous Bag of Words）和Skip-Gram。CBOW模型：它根据上下文词预测目标词。例如，给定一个句子，CBOW模型会尝试预测其中的每个词的上下文。这种方法在文本数据量较小时通常表现较好，因为它更关注全局的语义信息。Skip-Gram模型：与CBOW相反，Skip-Gram模型根据目标词预测其上下文词。也就是说，给定一个词，模型会尝试预测在其前后可能出现的词。Skip-Gram通常在大规模数据集上表现更佳，因为它更关注细粒度的词对关系。Word2Vec模型的目标是通过优化使相似词的向量在向量空间中距离更近。例如，“king”和“queen”的词向量可能在某个维度上非常接近，而与“apple”的距离较远。这种语义信息使得Word2Vec能够被广泛应用于文本分类、信息检索、问答系统等任务。训练Word2Vec模型的过程实际上是通过调整模型的权重，使得词与上下文之间的关系符合训练数据中的实际情况。通过反向传播和梯度下降，模型逐步调整词向量，使它们能够捕捉到上下文中的语义关系。训练完成后，模型可以通过多种方式使用，例如获取词的向量表示、计算词之间的相似度、查找与给定词最相似的词等。

在代码具体实现过程中，首先加载了股票收益率数据，并读取每份报告的文本内容。通过将报告文件从PDF转换为文本文件并存储在DataFrame中，程序为后续的文本处理打下了基础。接下来，使用gensim库中的Word2Vec类对报告文本进行训练。训练时，指定词向量维度为100，窗口大小为5，最小词频为1，以确保所有词语都参与模型训练。训练完Word2Vec模型后，代码计算了每个报告的平均词向量。具体来说，通过将每个词的词向量取平均值，生成一个固定维度的向量来表示整篇报告。这些向量随后被用作线性回归模型的特征输入，目标变量则是股票的下一年或第二年的收益率。在此基础上，将数据按照4：1的比例划分为训练集和测试集，使用训练集对线性回归模型进行拟合，并在测试集上进行预测。最后，计算并输出均方误差（MSE），以评估模型在样本外的预测性能。此外，代码还通过绘制真实值和预测值的对比图，直观展示了模型的表现。

使用Word2Vec模型对股票下一年收益率预测的MSE约为0.219，，而对第二年股票收益率预测的MSE约为0.089，下面给出了真实值和预测值的对比图，可以看到对第二年股票收益率进行预测的效果要明显更好一些。

