5、Doc2Vec模型

Doc2Vec方法是一种无监督算法，能从变长的文本（例如：句子、段落或文档）中学习得到固定长度的特征表示。Doc2Vec也可以叫做 Paragraph Vector、Sentence Embeddings，它可以获得句子、段落和文档的向量表达，是Word2Vec的拓展，可以不用固定句子长度，接受不同长度的句子做训练样本。

Doc2Vec模型是受到了Word2Vec模型的启发。Word2Vec预测词向量时，预测出来的词是含有词义的，Doc2Vec中也是构建了相同的结构，所以Doc2Vec克服了词袋模型中没有语义的缺点。假设现在存在训练样本，每个句子是训练样本，和Word2Vec一样，Doc2Vec也有两种训练方式，一种是分布记忆的段落向量（Distributed Memory Model of Paragraph Vectors , PV-DM）类似于Word2Vec中的CBOW模型，另一种是分布词袋版本的段落向量（Distributed Bag of Words version of Paragraph Vector，PV-DBOW）类似于Word2Vec中的Skip-gram模型。

Doc2vec相对于Word2vec不同之处在于，在输入层增添了一个新的句子向量Paragraph vector，Paragraph vector可以被看作是另一个词向量，它扮演了一个记忆角色。Average Word Vectors中，使用Word2Vec训练词向量，因为每次训练只会截取句子中一小部分词训练，而忽略了除了本次训练词以外该句子中的其他词，这样仅仅训练出来每个词的向量表达，句子只是每个词的向量累加在一起取平均的一种表达。正如上面所说的Average Word Vectors的缺点，忽略了文本的词序问题。而Doc2vec中的Paragraph vector则弥补了这方面的不足，它每次训练也是滑动截取句子中一小部分词来训练，Paragraph Vector在同一个句子的若干次训练中是共享的，所以同一句话会有多次训练，每次训练中输入都包含Paragraph vector。它可以被看作是句子的主旨，有了它，该句子的主旨每次都会被作为输入的一部分来训练。这样每次训练过程中，不光是训练了词，得到了词向量。同时随着一句话每次滑动取若干词训练的过程中，作为每次训练的输入层一部分的共享Paragraph vector，该向量表达的主旨会越来越准确。

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，并将文本数据转换为gensim的TaggedDocument格式，这是训练Doc2Vec模型所需的格式。接着训练Doc2Vec模型，设置向量大小为100，窗口大小为5，最小词频为1，并使用4个工作线程。在模型训练之后，定义一个函数compute\_document\_vector，用于计算文档的向量表示，计算文档向量并将其存储在DataFrame的doc\_vector列中，再准备特征矩阵X和目标变量y，移除NaN值以确保数据的一致性。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。

具体真实值和预测值的对比图如下：

