题目 3. 中文金融新闻的聚类

问题背景

在建立量化交易策略时，除了时间序列本身的信号，金融新闻的也是非常重要的分析依据。在本场景中，我们需要分析金融新闻中的高频词，预测哪些金融信息是相近的，去除这些近似重复的信息，然后将金融新闻聚为合适的几类。金融分析师可以通过这样的数据处理流程，减少阅读量，并且从宏观上了解信息的趋势。如何定义重复的信息？如何在聚类中选择合适的类数？请对此展开研究。

数据集和资料

金融新闻数据集 `q_3_news.csv`

根据 “CC-BY 4.0 转载需署名” 从 HeyWhale `https://www.heywhale.com/mw/dataset/5eb69242366f4d002d77d2b7/content` 获取。

最简单的自然语言处理算法是词频法，中文分词函数库 `https://github.com/fxsjy/jieba` 可以将文段分解成词语，去除停用词（包括连词、数词、语气词、人称代词等虚词），这样就可以统计文段中的词频。除了词频法以外，还有一类深度学习模型可以将词语转换成词向量。每个词对应一个词向量，词向量是高维空间中的一个矢量，两个词语的矢量距离越近，代表词义越相近。FinBert 模型 `https://github.com/valuesimplex/FinBERT` 不需要自己训练，可以直接下载其他人已经训练好的模型，调用推理功能将文段转化成词向量。

近义词需要使用到最近邻算法。词向量的维度很高，如果文本量比较大，最近邻算法非常耗时。我们可以将词向量矩阵存储为 KDTree 结构 `https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KDTree.html` 加速寻找最近邻的方法。

有些函数库集成了最近邻的功能，例如 `https://github.com/chatopera/Synonyms` 给定两个句子，可以直接得到相似度。

在用 DBSCAN 聚类时，可以使用 Elbow Method 选定最合适的邻域半径和最少样本点。

研究任务

问题 3-1. 将金融文本拆分为 80% 的训练集，20% 的测试集，不需要使用交叉验证。拆分数据集时使用 `sklearn.metrics.train_test_split(..., random_state=71193832)` 构建为高频词表，阐述选择词频的依据，在测试集上绘制词频的相关性矩阵。训练 PCA 算法（该算法虽然是无监督的，但实际上要根据数据计算变换矩阵，仍应该在训练集上 `fit_transform`, 然后在测试集上只使用 `tramsform`），将测试集投影到向量空间中。

问题 3-2. 根据上述结果计算文本的余弦距离，选择合适的阈值，去除近似重复的文段，报告保留的样本比例。训练 DBSCAN 聚类算法，将剩余样本聚为几类，计算 `silhoutte` 指标。

问题 3-3. 使用 FinBert 模型计算词向量，重复问题 3-2 并比较结果；用 Synonyms 函数库直接计算句子之间的相似度，使用 `sklearn.cluster.DBSCAN(..., metric='precomputed')` 直接完成聚类，与问题 3-2 的结果作比较。