**Final Project**

**(Due on 2024/09/01)**

1. 下载沪深股市上市公司的年度报告，要求至少500份报告，可以是同一个公司不同年份的报告，或者不同公司同一年份的报告，也可以是不同公司不同年份的报告。
2. 将所有报告导入Python，分词并去除图表、数字、特殊字符、标点符号和停用词。
3. 下载每份报告次年上市公司股票在当年最后一个交易日的收盘价及第一个交易日的开盘价并导入Python。例如，对2020年的年报，需要下载2021年最后一个交易日的收盘价及第一个交易日的开盘价。
4. 根据3中的收盘价和开盘价，计算股票的年度收益率。
5. 任意选择400份报告作为训练样本，使用词袋模型预测4中的股票收益率，预测模型为线性回归，并以剩余100份报告作为测试样本，计算词袋模型的样本外均方误差。
6. 将5中的词袋模型换成Word2Vec模型，生成每个词的词向量（词向量维度为100），以每份报告中所有词向量的均值预测股票收益率，并计算样本外均方误差。
7. 将5中的词袋模型换成Doc2Vec模型，生成每份报告的文本向量（文本向量维度为100），以文本向量预测股票收益率，并计算样本外均方误差。
8. 比较5、6、7，哪个模型预测效果较好？为什么？
9. 将预测目标改为每份报告两年以后的股票收益率（即对2020年的年报，预测目标改为2022年股票收益率），重复5-8，与预测次年收益率相比，样本外预测误差如何变化？试解释导致这种变化的原因？