参考文档：单独用每天新闻头条标题来预测当天的道琼斯开盘价与前一天开盘价之间的差异。

目标：利用股票价格和相关新闻对股票开盘价进行一个预测。

1. 单纯利用价格数据进行时间序列的预测波动率
2. 用新闻语料来进行波动率的预测

2.1 认识word2vec

Word2vec是一个词向量表征工具，目的是将语料中的所有单词都分配到一个向量，而这个词与词之间的相似性关系，相关性关系，能够通过词向量之间的向量距离来表现。

Word2Vec这个模型的原理是一个神经网络，其神经网络的结构只包含一个输入层、隐藏层和一个输出层。输入层对应一个输入单词的onehot向量，输出层是与输入单词同时出现的单词的概率分布。

要理解word2Vec就得先理解huffman树，huffman树是指这样一种树，其每个节点带一个权值，这种树的结构使得所有节点的带权路径之和最小。那么采用huffman树进行编码能够使得整体编码的长度最小（如果用词的词频作为节点的长度，树节点（词）的深度就是当前这个词编码的长度，因为叶子节点没有子节点，因此恰好避免了出现要给词的编码是另一个词的前缀这种情况）。在word2vec中有两个约定：1、权值较大的节点作为左孩子，权值较小的节点作为右孩子。2、左孩子编码为1，右孩子编码为0。

2.2 用word2vec可以获取每个词对应的词向量