## 本阶段成果

整个第一阶段，大概过去了一个多月，中间过了很久毕设题目才出来，我又是在公司实习，所以其实并没有做多少毕设上的事情。

整个第一阶段，大概做了一下几件事：

1. 根据毕设题目构思，规划毕设大致路线。
2. 完成开题报告。
3. 完成毕设的初级任务。

关于前面的规划问题以及开题报告，都已经在我的开题报告中阐述了，所以这里不再赘述。这篇阶段性报告主要就记录我第一阶段所做毕设的工作内容。

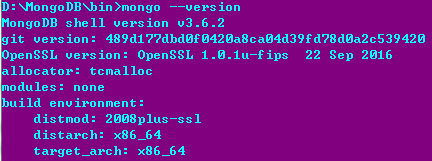
根据之前的规划，首先要做的事情就是搭建开发环境，已知可能要使用到的环境有Python、Node.js、MongoDB、Visual Studio Code，所以我在自己的电脑上安装并且配置好了这些开发环境，详情如下图：



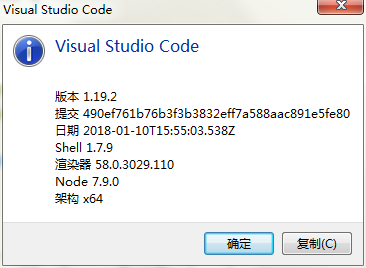
Python版本



Node版本



MongoDB版本



Visual Studio Code版本

环境已经基本搭建好了，接下来就是联系项目实际，着手准备开发。

根据任务书，第一条任务是“设计软件实现用户自定义收益率数据的收集和曲线的拟合”，这是一个比较模糊的概念，因为虽然我已经确定了数据的获取方式（Python爬虫）与获取目标（和讯网），但是我还不确定哪些数据是有用的，并且这是一个比较广泛的任务，是我最终应该做出来的产品，不应该在项目初期就开始做，所以我决定先完成第二个部分：“研究美国国债、LIBOR、SHIBOR、中国国债，使用PCA方法将曲线上的时点转化为较少的3-4个主因子，研究各主因子如何驱动收益率曲线形状的变化，以及对固定收益产品价值波动的贡献度”。目前而言主要是在了解PCA方法的原理以及应用步骤，并且取得了一定的进展，接下来将会对我取得的进展进行阐述。

根据我所学习到的知识，**PCA**（Principal Component Analysis）是一种常见的数据降维方法，可以将多维数据降到更少的维度，在信息量丢失最少的情况下，以最大的精度描述元数据所描述的事实。这是一种经典的将数学应用于数据分析与计算机的算法，现在最常见的应用是主成分分析（主元分析）、人工智能（图像识别），主要涉及到的知识有概率论与数理统计、线性代数。

**那么降维是什么，我们为什么要降维呢？**假设现在有一组数据用于描述我国部分地区的经济发展状况，这组数据包含GDP、人均GDP、农业增加值、工业增加值、第三产业增加值、社会销售总额等多个字段，我们如何通过这些字段来判断一个地区的经济发展的好与坏呢？这就需要我们使用PCA来进行数据降维了，所谓的维，就是指维度，可以理解为字段，一个维度代表一个属性，属性太多的时候，我们无法分辨每个属性对于某个指标的意义以及作用大小，所以我们需要进行主成分提取，提取出来的主成分可以帮助我们分辨哪个地区的经济状况更好，以及差距有多大。这是数据降维最大的作用，使用最少的数据表述最多的信息，这样既使得数据简单明了，而且数据量大大减小，更容易储存。

**降维是怎么实现的？**最简单的解释就是利用了线性代数中乘法的特性实现了降维。我们知道线性代数中，一个的矩阵与一个的矩阵相乘，结果矩阵为。假设我们的原始数据有12个维度（属性），共有172条数据记录，那么我们可以把数据集构造为一个的矩阵，我们让这个矩阵乘以一个的特殊矩阵，我们的数据集就可以简化为的矩阵，即维度减少到了6个。但是不是随便乘以一个矩阵就可以，因为我们降维的要求不仅仅是减少维度，最重要的是要使用较少的维度表述最多的元数据信息，所以我们怎么求这个特殊矩阵就成了一个问题，我们要确保信息的损失率减少到最低；这里就使用到了线性代数中特征值与特征向量的概念，以及概率论中协方差的概念。协方差和相关系数反映了两个随机变量X和Y的线性依赖关系，假设是X和Y的相关系数，当 > 0时，X和Y成正相关，及X增大Y也会随之增大；当 = 0时，可以认为X和Y两者没有关系；当 < 0时，可以认为X 和 Y成负相关，及X增大Y会随之减小。我们可以通过求原始数据集的协方差矩阵，得到各个变量之间的协方差，如果两个变量之间的协方差绝对值非常接近于1，那么可以认为两个变量之间具有很强的关联性，所以在降维时可以省去其中的一个以达到降维的目的。但是如果我们要通过编程实现取舍而非手动实现，那么我们应该怎么做呢？这时我们应该使用线性代数中的特征值和特征向量。我们可以求解协方差矩阵的特征值和特征向量，对于求出的特征值根据从大到小的顺序排列，选取前d项（d代表你想将元数据集所降至的维度数），然后选取对应的特征向量，得到我们想要的乘数因子。那么最后我们将原始数据集乘以我们现在得到的乘数因子（实际上是一个投影的过程），即可得到一个降维之后的数据集。至于为什么求协方差矩阵的前d个最大特征值得到的特征向量是最合适的乘数因子，需要深入了解线性代数，这里我没有过多地深入。

**降维的实现步骤：**

1. 将原始数据转化为一个的矩阵（m是数据量，n是数据维度数）。
2. 对数据矩阵进行归一化处理，即矩阵的每一项都减去平均值。
3. 对数据矩阵求协方差矩阵。
4. 求协方差矩阵的特征值和特征向量。
5. 特征值从大到小排列，根据所需要的维数d，选取前d项特征值，并选出对应的特征向量，组成特征矩阵。
6. 将数据矩阵乘以该特征矩阵，得到的结果即为降维后的结果。

## 下一阶段计划

下一阶段主要任务是接着第一阶段的计划，实现PCA主成分分析。

具体包括如下几步：

1. 编写爬虫程序，从中债网上爬取历年所有国债收益率数据，保存到本地。
2. 将所有.xlsx格式的数据转换为csv格式文件保存，方便Python程序读取数据。
3. 编写Python程序，实现PCA主成分分析（或者通过SPSS工具），提取出三个主因子，并分析贡献率。
4. 进一步分析美国国债、SHIBOR、LIBOR。
5. 研究蒙特卡洛模拟。