此为占位符文本。它可在您键入文本时，向您提供直观的页面文本视图。例如，您可以看到文本的大小、颜色和布局。轻点此文本，并开始键入。一旦开始键入，占位符文本就会被键入的文本所替换。此为占位符文本。它可在您键入文本时，向您提供直观的页面文本视图。

structure report

2021.06.24

Domineer

Draft Questions

我们获取的是什么数据（爬的是谁的或者哪个tag的数据以及哪个时间段）

我目前设想的是监测twiiter的：最近3个月甚至更长时间内#btc #eth #blockchain，还监测一些在虚拟货币社区可以呼风唤雨的大v：elon musk/jack Dorsey的近期（差不多1年的）所有tweets然后根据keyword过滤出来

具体关键字我也会在监测tweets数据集上跑最相关10个词的数据分析，然后对关键字进行维护

报告标题

2

获取后的数据是如何进行清洗和存储

清洗：肯定先去掉各种html标签、各种乱码和无法分析的部分。然后进行stopword分析，去掉没有语义价值的词汇。然后在进行分析

看数据量大小，如果几百M级别肯定就存在本地文件系统csv/json

如果更大，考虑进到hadoop集群

如何进行情感分析（情感词典还是机器学习，机器学习的话准备采用什么模型），

这块目前考虑采取深度学习方法，已经从各种开源渠道下载到一些标记好的数据集，然后尝试进行训练，神经网络模型层数和节点数应该会自己架构）同时还调查贝叶斯模型的分类感情的准确度

从哪里获取价格的数据，

（还在研究中，也许想分析的区块链付出一定的gas费用直接从区块链同步，或者从第三方api平台获取）

情绪和币值的模型准备采用哪些模型，

（这个目前打算使用线性回归分析，或者多项式逼近，或者逻辑回归，随即森林，也许是一个综合的池化模型，这个需要到时候前面步骤做好了之后，然后进行数据探索，然后再具体探索模型）

如何进行短线模拟交易（大概的做法），

（暂时就取最近的历史数据1个月，然后进行训练集/测试集划分，然后使用测试集上对短线交易进行模拟）

工程化自动化如何实现。就是整个项目在老师脑海里大概准备怎样逐步操作。

（自动化主要通过unix 管道和python脚本的方式，希望可以达到95%以上的流程自动化，然后模型和应用化，通过模型持久化，把训练好的模型提供给web服务进行使用，这样可以实现离线训练，在线使用）