1．课题回顾
   a) 存储墙问题
      i. 计算速度远大于 I/O 速度
      ii. 能否用计算弥补？
   b) 老师上课 ppt
   c) 考虑分布式文件系统
      i. 类似本地文件系统
      ii. 读取过程更加冗长，能否提前读？
      iii. 需要预取
      iv. 预取可以是使用高速缓存，或其他简便的预测
      v. 能否使用机器学习达到更好的精确度？
   d) 我们的方案
      i. 分成两个模块预测与预取
      ii. 预测是使用机器学习，利用历史信息预测未来趋势
      iii. 预取是结合具体的文件系统实践，这里选择的是 HDFS
      iv. 两个模块分离是为了有更好的适用性
2．预测部分
   a) 有哪些历史信息
      i. 很多很多
      ii. 大部分信息量很少，提取出来不划算
      iii. 比如文件名：可能需要用到简单的自然语言处理
   b) 然而
      i. 只有文件 id 顺序价值量最大
      ii. 研究的时候其他的也试探着做了
   c) 数据来源
      i. Garbage in, garbage out，数据对于机器学习问题的研究非常重要
      ii. 但是关于 DFS 文件存取信息的数据很难找到
      iii. 很少相关论文，其中绝大部分使用了 20 年前的数据，而这些数据的解析代码没法用
      iv. 自行构建
      v. Strace，从某篇论文获得启发，使用本地的文件系统来模拟，记录系统调用，本实验记录的是 busybox 的 make 过程
   d) 静态预测
      i. 预测器不随时间变化
      ii. 使用自行构建的数据

e) 动态预测
    i. 预测器一直在动态改变
    ii. 使用 strace 得到的数据
f) 动态预测为主
    i. 由于时间限制，仅仅讲动态预测，关于静态预测感兴趣的可以参考我们的文档
    ii. 模块的添加也是以动态预测为主
    iii. 关于模型的细节，在后面介绍

3．添加模块部分
a) 结构示意图
    i. 预测与预取分离
    ii. 便于协作与适用性
b) 两个方案
c) 使用脚本
d) 部分细节
e) 问题与改进方案

4．本地测试结果
a) 数据预处理
    i. 为了简化结果，使用在 busybox 目录内部的文件，而且只使用二级目录下的
b) 数据探索
    i. 按文件出现的顺序对其进行编码，也包括对目录重新编码
    ii. 得到目录的趋势如左，得到 id 的趋势如右，每个图的上方为局部放大图
        1. 可以明显看到的是某种规律，趋势，特别是放大图
    iii. 得到 id 的 delta 的分布为右上，做一下说明，总体的 delta 序列有 14441 个，需要编码的 delta 有 966 个，但是从图中可以看到，这里的分布比较集中，为了缩小搜索空间，可以只取出现最频繁的。我们取了 30 个最频繁出现的 delta,这 30 已经囊括了 70.5%的 delta 序列。
c) LSTM 模型
    i. 一种循环神经网络，具体的如果有时间在最后解释
    ii. 模型结构如右上，比较简单
        1. 为了缩小搜索空间，使用的是 delta 而不是文件本身的 id 作为数据值
        2. 将窗口设为 5，意思是根据之前的 5 个 delta 值预测出下一个 delta

值

iii. 测试结果为左下，baseline 结果为右下，baseline 的原理是以上一次的 delta 作为本次的预测值，因为考虑到 delta 的连续性

1. Precision 指的是预测的项中正确的
2. Recall 指应该预测的项中预测准确的
3. F1 是两者的调和平均

d) 倘若去掉已知 30 频繁项呢？

i. 左下为上个模型的效果，右下为 baseline，此处的 baseline 是使用上一次读取的时候，本文件的下一个文件

ii. 对比上一个，相当于折半。

1. 这说明 delta 频繁项对于预测至关重要
2. 实际的运用场景：delta 的分布会有某种规律
3. 具体场景下，通过定期调整 delta 来满足需要

5．其他模型

a) GRU

i. 在使用 dropout=0.2 的微调之后，获得了和 LSTM 相当的效果
ii. GRU 的计算量小于 LSTM

b) Bidirectional

i. LSTM+Bidirectional
ii. GRU+Bidirectional
iii. 不论哪一种，效果都不如直接使用原 RNN

c) Conv1D

i. 一维卷积，计算量小于 RNN
ii. 卷积与 RNN 层叠，或卷积层叠
iii. 其他的类似，这里只放两层卷积的
iv. 结果比 RNN 差，这部分还没有完成超参调整

d) 使用这些模型的原因

i. 使用更小的计算量来获得同样的效果

6．延伸

a) 回到起点

i. 存储墙问题
ii. 计算弥补存储缓慢

b) 实际效果

i. 目前的计算时间远远超过系统所能承受的地步
ii. 只具有学术价值，不具有实践价值

c) 延伸

i. 使用更轻便的 NN 或者加速
ii. 使用普通的机器学习手段
    1. 使用普通机器学习的部分由于时间问题没有完成

7．附录

a) RNN
b) Bidirectional
c) LSTM
    i. 解决梯度消失与梯度爆炸问题
    ii. 细胞状态，隐藏状态
    iii. 输入门，输出门，遗忘门
d) GRU
    i. 遗忘门和输入门合并成为单一的“更新门
    ii. 将细胞状态和隐状态合并
e) CNN 与其他更加详细内容线下交流