一共有三位老师授课，每位老师出了一部分题，第一部分是崔院长出的，占50分，鹿老师和李老师各出了25分的，特别标注一下，李老师有一章ppt没讲，所以没出那一部分的题，但是以后就不知道了，可能会出没讲的那一章的算法之类的。

第一部分

1.ETL概念

2.相似度距离举例

3.朴素贝叶斯算法的计算（给的题目很简单，只要明白算法概念，算起来就很简单）

4.knn计算（算相似度距离，然后看给的数据是哪一个分类）

5.k-mean算法的概念和优缺点

第二部分

1.根据给的图说明同质性计算的概念（图就是上课给的例子）

2.pagerank计算（题目给的很简单）

3.gsp vcg计算（也很简单，只要理解了就能做出来）

第三部分

1.hdfs冗余数据的原因

2.mapreduce的map shuffle 几个过程的具体内容

3.hdfs只有一个名称节点的局限性

4.mapreduce计算数据集中的最大值的过程