第一次会议：2018.5.20 9:00-10:00

本次会议分配期末项目初期任务以及讨论项目细节，具体情况如下：

1. 对三张表：voicetrain、smstrain、watrain分别做变量衍生、数据清洗：

**樊忠睿——voicetrain**（已接近完成，可灵活安排其他事宜）

**夏丹琳——smstrain**

**高星——watrain**

**注：每张表衍生出的变量，在/data文件夹中的变量衍生表中填上衍生变量的数据描述，每张表的衍生，在/目录下单独创建.py文件，命名格式为”表名\_数据格式名（RDD OR DF）.py”**

时间节点：5.27（本周末）

1. 变量衍生的具体细节：
2. voicetrain表和smstrain表中号码长度为11的记录有可能为手机或者直辖市座机，可通过号码前三位变量来区分，衍生变量时需要注意。
3. 手机号码前三位可以提取该号段是否为移动、联通、电信用户（已实现）。也可以判断该号段的使用年限（需要资料支持）。
4. 因为提供的数据是某45天内的，考虑采用时间戳判断是否有周期性（周末、工作日）。
5. wadata表中的网站/APP分类高达数十万，考虑采用文本挖掘方法对其进行分类处理。

3. 积极查阅相关资料，代码方面有问题请找樊忠睿讨论。