# 文本分析兴趣小组安排

## 一、整体目标

每个人在这一年内都要经历从数据获取、数据清洗、数据分析以及到论文撰写的一个完整过程。

但因为每个人的基础不一样，年级不同，咱们就分两条线来执行，但具体根据个人的进度来调整。所以如果找到了自己感兴趣的研究点，或者想早点做研究，可以提出来。

**主要分两条线：**一是技术导向，二是研究导向。（何时参与哪一项，可以自己随时有想法就跟我提出来，但建议前2个月要夯实一下技术功底，把数据源准备好，再展开研究）

## 二、重要时间节点

2019年9-12月：以学技术为导向，专攻技术难点

2020年1-9月：主要以研究为导向，技术作辅助

## 三、实现模式

**“知识分享+任务复现+实战演练”相结合的模式**（实践操作才能真正领会到内涵，并且运用到自己的领域，再讲出来，大家一起讨论后，才会最终成为自己的东西，谁也抢不走）

1. **知识分享**：包括技术分享、学术分享和生活分享。目的是快速学到更多的知识和技能。每周每人专攻一块，然后将这块分享给大家，大家就会同时学到了所有的知识。如果感兴趣，在接下来可以自己找时间继续复现。
2. 技术分享：分享你这个任务是如何实现的，是如何实现的，在实现的过程中遇到了什么问题，哪些坑以及自己的心得等，记录下来并分享给大家，这样能够加快大家的学习进度。
3. 学术分享：我会不定期的跟大家分享一些适用的知识，引导大家如何做文献回顾、如何管理文献、如何找研究思路、如何撰写论文等，即便不读博士，你们以后的本科毕设和硕士论文都需要涉及到写作。大家也可以参与分享，一起学习。
4. 生活分享：可以畅谈自己想分享的故事和一些烦恼或困惑等，一是为了心理疏解，二是大家能共同为你出出主意等。
5. **任务复现**：通过复现网上现成的例子学习，这个我可以分配或大家对自己感兴趣的自己找一些。
6. **实战演练**：我会根据你复现的任务，让你结合本领域的研究环境再次用此方法完成任务，最终变成自己的知识。

（具体如何执行任务：每周2个小任务，每个任务一周完成，如果任务比较难或者该周比较忙，可以两周完成。每个任务2-3人一起完成，这样方便遇到问题可以交流沟通，并且在技术分享的时候，能够互相补充。具体而言，第一周复现网上原有任务（我找给大家，或者大家自己找到感兴趣的提出来作为本周任务一起完成）；第二周按照复现的任务，完成与自己相关的感兴趣的任务。如：网上任务可能是微博数据情感分类，那第一周你就要复现这部分任务，第二周要完成与电子健康相关的情感分类任务。）

## 四、阶段性目标

1. 技术基础夯实阶段：爬虫
2. 数据清洗及文本分析阶段
3. 数据分析阶段：
4. 研究想法构思、论文撰写阶段

## 五、任务分组

1. 2大组3小组的模式（按照能够方便在一起讨论的模式）
2. 爬虫小组：方鹏和勇奇做小组长，做培训及指导。

方鹏：建涛、楚云

勇奇：紫月、文佳

1. 文本分析小组：方鹏、勇奇

（注：每个人都会走从数据获取到论文撰写的过程，只是每个人的时间节点不同）

## 六、个人规划

（一）近5年规划（需要详细一点），或说自己的总体目标

1. 以后想做什么？读博或工作？
2. 想做数据分析师，并在研究上有了解/深入了解/建树等
3. 想实现的目标：会爬虫，文本分析，计量模式分析数据，撰写论文（小论文/本科毕设/硕士论文）等

（二）近一年内的详细时间规划

2019.9-2020.9一年的详细规划，每个时间节点完成什么目标

比如方鹏，现在研一，那你研二开学后就要评奖学金，那可能需要在这之前出来一篇小论文，期刊或会议等。那就要在此基础上规划，什么时候写完，什么时候投出去，什么时候做好数据收集、处理及分析等。

## 七、文档管理

小组内**享受全部知识与技能共享的模式**，互相学习，才能更快的进步。每周的任务文档、技术难点、以及心得等大家觉得值的记忆的都放在共享文件上。放干货，供大家学习参考以及复现等。

每个人要养成好的习惯，做好版本控制与管理，QQ群就是咱们的交流群，我会每个小组建个文件夹，大家把自己每周的任务行文记录在一个Word里，然后上传到自己的文件夹。文档命名格式为：任务名-完成人姓名-时间戳-版本号（文档命名要简短易懂），如：文本分析兴趣小组安排-书晴20190909-V1.0

## 八、每周分享会时间

最好在周一-周五内选个可行的时间

## 九、任务列表

1. 文献数据爬取，包括title，abstract，journal，year，DV，IV，甚至可以做语义分析等。
2. 文献计量分析
3. 主题分析
4. 有监督学习等
5. …

## 十、本周开始的任务

1. 数据爬取任务：广医医科大学附属第一医院、广州第一人民医院、北京航空医院、北京航天医院 好大夫线上数据
2. 文本分析任务：主题分析，文献计量分析