

# SAM 小组学生培养方案

## 1. SAM 小组的风格

团结、积极、愉快

## 2. SAM 小组的科研之路

科研是广大研究生的主要旋律，作为学术性研究生，必须培养好良好的科研素质。基本上，做好研究需要这几方面的素质：（1）坚韧不拔，良好心态：刚开始，科研中失败的实验一般远远多于成功，必须锻炼出良好的心态。自己成长就可以，不需要太看重结果；（2）严谨（细心），做研究必须非常细心，包括实验设置、写文章等等，决不能马虎，尽量做到完美；（3）坚持奋斗：做研究是个漫长的过程，相关能力的培养也需要不断努力，长期坚持才可以锻炼出来。

下面结合我这今年来的培养学生经验，建议下面具体的关于 NLP（Natural Language Processing）学生培养计划：

培养计划分为三个环节：（1）基本功练习；（2）辅助合作；（3）主导合作；

### （1）基本功练习

主要是英语、编程方面的基本功。英语是自己自由培养，这里不多阐述。编程方面的锻炼，主要是针对 NLP 一些基本方法的锻炼，例如机器学习编程，学习利用工具等。具体来讲，下面是学生必须尝试的练习：

- （a） 编写基于 K 近邻分类方法的情感分类系统 （完成时间 1-2 个月）
- （b） 编写 Naïve Bayes 分类方法的情感分类系统 （完成时间 0.5-1 个月）
- （c） 编写使用 SVM 工具的情感分类系统 （完成时间 0.5-1 个月）
- （d） 编写基于 CRF++工具的分词系统 （完成时间 1-2 个月）
- （e） 编写基于 IG 的特征提取方法用于文本向量降低维度（完成时间 0.5-1 个月）
- （f） 编写基于 LSTM 的情感分类系统（完成时间 1-2 个月）

**注：**（a）和（b）的分类方法是自己编程实现，（c）（d）的机器学习方法是已有的工具；（e）是特征提取方面的一个试探性研究方法学习，（f）是基于深度学习的方法，可以使用一些工具实现。

小组会提供基于最大熵工具的情感分类系统，用于学习。所有程序为 Python 程序。在完成上述系统的时候，还需要自己有一定的查找参考文献的能力。例如：对于（d），你首先要了解分词是什么样的任务，CRF++是什么的工具等等。

编程基本功是必须要锻炼的，不管你以后是否从事 NLP 方面的研究。编程能力都是学生必备的能力。

## （2） 辅助研究

进行完基本功练习以后，说明编程的基本功已经扎实。下面进入研究阶段，刚开始，可以同指导老师或者师兄师姐一起做一些研究，起得左右主要是帮忙编程序、分析语料等等。

作为指导老师，我会考察你第一环节（基本功练习的）的结果。如果效果好，可以同指导老师一起忙一些比较有挑战性的研究课题；如果效果不好，可以考虑同师兄师姐忙一些简单一些的研究课题；

## （3） 主导研究

辅助研究以后，可以开始进行自己为主导的研究课题（发表文章为第一作者）。下面是做自己研究课题的必须步骤：

（a）文献综述：根据我给同你说的研究大方向，查找相关文章，了解当前该方向研究状态，

（b）寻找 Idea：了解了当前研究状态后，寻找突破点，获得自己的研究点；

（c）实现并验证 Idea：编程实现自己的 idea，如果不成功，分析原因，继续改进或者放弃寻找新的 Idea

（d）如果 Idea 成功，开始进行文章写作

（e）成果汇报（参加学术会议），把你的文章在大会中进行汇报