**网址内容爬取及二分类器设计**

**暑假项目进度**

G171118孙一宁，G171120田肇阳

简要的来说，本文预期完成以下几个任务：文章获取，即利用URLLIB模块从百度爬取搜索结果的网址；寻找关键词、对文本语义有影响的词，即将截取网站内容并利用JIEBA模块进行分词；试图复现他人的工作，利用WORDCLOUD模块提取文章特征，从而进行词的判别；利用TENSORFLOW模块人工分类一部分数据以后将剩余数据进行判别。本文以关键次为江歌案为例开展研究，但又不局限于这一个关键词，本文提出了一个文章负面信息过滤的方法。

截止目前，本文已经完成了文章获取及寻找关键词的工作。对于江歌案关联的文本二分类本文已经完成了这一关键词的初步结果，获取了关键词的判别数据，仍需要完成对于文章的判别工作。本文尚未开始WORDCLOUD词频统计工作。

论文撰写方面，本文已经完成了研究背景及文献综述的撰写工作，开始了文章获取及寻找关键词的撰写，尚未开始其余部分的撰写工作。