2018 年第一周工作进展汇报

课题任务：
(1) 阅读文献，理解和掌握现有的各种事件抽取方法和模型；

通过搜索相关资料，了解事件抽取的定义及基本概念，查阅相关的文献，理解事件抽取的一般过程和应用方向，在中国知网上下载相应的论文准备进行学习。

阅读论文：事件抽取技术研究综述
认识清楚事件抽取的概念，了解基于模式匹配的元事件抽取和基于机器学习的元事件抽取。
更进一步，大致理解基于事件框架的主题事件抽取和基于本体的主题事件抽取。
查阅其他资料，帮助理解相关的涉及技术。

阅读论文：基于事件抽取的网络新闻多文档自动摘要
基于模式匹配的事件抽取知识表示直观、便于推理，但过于依赖具体领域，可移植性差，性价比不高。提出了一种基于事件实例聚类的事件抽取方法，以单句作为事件的基本抽取单位，通过二元分类器辨析出事件句和非事件句，通过对事件句聚类，得到同一主题文档集中所包含的不同事件集合，完成事件抽取。

阅读论文：面向微博文本的事件抽取
这篇论文是一个完整的事件抽取实现的体系，是一个能够学习的范例。
LDA(Latent DirichletAllocation)模型，用一个服从 Dirichlet 分布的 K 维隐含随机变量表示文档的主题概率分布，试图模拟文档的产生过程。
对于 Dirichlet 分布的含义，有的地方还没搞清楚，但大体使用方法算是理解了。
给定原始的微博数据，先进行了命名实体识别(Named Entity Recogllition，NER)和时间信息抽取和推断，将命名实体和日期信息从微博中抽取出来。然后再对关键词进行抽取，确定事件的类型。
原始数据处理，命名实体识别(NER)和时间信息抽取和推断，然后进行下一步操作。
使用名为隐事件分类模型(Latent Event&Category Model)的贝叶斯模型来抽取事件并把他们分为不同的类。
了解了基于 LECM 的事件抽取和基于 LECM-d 的事件抽取的差异和相似点，预处理的必要与否需要考虑。

阅读论文：中文事件抽取技术研究
阅读论文：基于主题的中文事件抽取技术研究及应用
阅读论文：基于信息单元融合的新闻原子事件抽取
大致阅览了以上列出的几篇论文以及其他论文的相关内容，准备接下来进一步研究。

总结：
这周大体上了解了事件抽取的基础概念，跟一部分抽取方法和模型。
事件抽取首先需要搞清楚 NER 和时间信息抽取，在法律案例上分析也有一定借鉴之处。
对于下一阶段的学习方向有了一定的想法。需要进一步学习不同的抽取方法和模型，至少需要对不同方法基础的应用有初步了解。