2. 动手实现的法律文书的预处理、分词、词性标注等过程。对其中涉及到的基本方法有了较为清晰的了解。

**a) 预处理**

把原始文本中的标点符号，括号，括号内的内容，以及无相关意义的词汇“本院认为”、“违反道路交通管理法规”、“驾驶机动车辆”、“因而”、“缓刑考.\*?计算”，“刑期.\*?止”、“依照《.\*?》、“第.\*?条”、“第.\*?款”、“的“、”了“、”其“、”另“、”已“、”且"。

**b) 分词**

用PYLTP库的segmentor进行分词。

除此之外，HMM分词、Jieba分词、SIGHAN分词、正向匹配法都是不错的分词方法。

**c) 词性标注**

利用哈工大的NLP工具LTP的Postagger进行词性标注。

总结：此过程主要是学习使用现有的工具，较为简单。

学习内容：

1. 看了几篇讲Phrase mining的论文、对NLP的一些重要概念初步了解：Apriori算法用于得到频繁项集和关联规则，马尔科夫链在文本生成器的应用，LSTM擅长于处理序列数据。
2. 初步了解词嵌入的知识：NLP中用向量表示词，词有one-hot和分布式表示两种常用表示方法。分布表示法还可以分为基于矩阵的、基于聚类的、基于神经网络的。词嵌入就是基于神经网络的分布表示。初步了解word2vec语言模型。

总结：要学习理解的概念和思想方法还有很多很多，学习的过程要注意联系和总结。

正在着手解决的问题

1. 对之前标注的数据进行清理，修订其中标注不当、标注错误的数据。

部分标点符号仍然没有去掉（比如省略号和一些键盘上没有的特殊符号），部分无意义词汇仍然没有去掉（比如“而”，“和”，“但”，“等”）。之前是基于个人经验的处理，但从实际效果来看、基于他人总结出来的常见案例来清洗数据可能效果更好、效率更高。

1. 根据案件内容进行判决结果的预测，方法是简单的线性回归。难点在于从语料库中提取出格式化的数据用于训练和预测。

希望实现的功能：输入构成罪名（比如：交通肇事罪）、责任（比如：刑事责任，民事责任）、事故（比如：交通事故）、积极因素（自愿认罪、赔偿受害人经济损失、取得被害人亲属谅解、具有悔罪表现、适用缓刑）、消极因素（比如：致被害人死亡、驾车逃逸），经过处理，可以得到输出的最终量刑（有期徒刑：？月，缓刑？月）的数字。

计划采用450个案例作为训练集、50个案例作为检测集。

计划由死亡人数、受伤人数、积极因素（自首、坦白、赔偿损失）、消极因素（逃逸、醉酒驾驶、有前科）等10个特征构成向量，由线性回归实现量刑月份的预测。目前遇到的问题是暂未找到合适的方法把上文提到了10个特征作为格式化的数据提取出来。

解决方法：① 借鉴one-hot的思想，把十个特征排成一个10维向量，然后根据训练结果给各个维度加权、直到最终准确率达到要求。

② 上网查找工具/算法，也许这个问题已经被解决了。