阶段性工作小结（2020.2.1）

# 一、基础知识学习情况

## 1.NLP：主要参考资料《Python自然语言处理》

学习自然语言处理中一些基础的相关概念，安装了NLTK、numpy、Matplotlib、jieba等相关第三方库进行初步学习。

* 学习使用NLTK库进行简单文本分析（部分笔记如图）

图片包含 黑色, 监视器, 文字, 屏幕

描述已自动生成

NLTK中的基本对象为nltk.text（所谓的词链表），即Python中的列表对象，将分过词的文本存储在列表中，nltk中的其他函数均是对词链表进行种种分析，包括统计词频、寻找语法上相似的词等等。因而对于中文文本进行正确的分词后（jieba分词准确性较高），nltk库的相当一部分函数可以用于中文文本分析。而大量文本资料（nltk.text）对象涉及另一个重要概念：语料库（Corpus）

* 语料库（Corpus）

语料库在大多数情况下是自然语言处理不可或缺的基础。以下链接中我收集了语料库的一些相关资料并写了一些自己的思考

[https://github.com/JJYDXFS/JJYDXFS.github.io/blob/master/语料库.md](https://github.com/JJYDXFS/JJYDXFS.github.io/blob/master/%E8%AF%AD%E6%96%99%E5%BA%93.md)

链接中提出了当前遇到的主要问题：主题不明确。该问题将会在第二部分“问题”中再次提出。

## 2.爬虫：主要参考资料《Python网络爬虫权威指南》

学习爬虫基础原理与相关概念，安装了Requests、bs4、Scrapy、PyMySQL等第三方库进行初步学习。

目前仍在熟悉bs4库相关函数的用法，部分笔记如下

图片包含 监视器, 室内, 黑色, 网球

描述已自动生成

# 二、问题

1.主题不明确：

主题不明确将会使语料库中基本文本资料的收集没有针对性，从而寻找近义词等等问题将无法得到最有效解决，譬如同样的词在不同主题下有不同用法。当前初期阶段暂时通过人工解决该问题，但细化主题仍有助于提高效率。

//以下是一些废话

//如果文本量足够，可以参照早期新闻分类的方式，通过计算文本特征词汇相似度将收集到//的文本进行分类，此处则会面临计算机算力不够的问题（本项目后期应该也会需要进行分//布式集群的搭建以提高计算能力）。

综上，当务之急，从开展工作的角度，希望可以尽快细化主题。