

# NLP 期末个人项目总结报告

## 一、项目要求

文本分类、聚类项目：在搜索引擎中输入自己的姓名，将搜索结果的前几十页中所有链接列表中的网页抓取下来。

抽取网页文章正文，保留真正有用的描述人的文本。

对抽取出来的文章进行分类、聚类，找出这些文章中哪些是描述的同一个人，并尽可能准确的将不同的人及其简单描述抽取出来。比如：

可理解为：爬取网页正文内容+文本分类+文本聚类

## 二、总体思路

1. 爬取网页正文内容存为 txt 文件
2. 对每一个 txt 文件进行聚类分类的分析

## 三、项目分析

1. 爬虫的知识上课有过讲解并且网上有许多参考资料。在期末的小组项目中也进行了对网站国学大师的爬取。因为对爬虫有一定的实操经验。但问题在于，国学大师的爬取局限在一个网站内，只要搞清楚其框架即可。但是搜索引擎下面各个网页的框架不一，不能采用上述方法。因而查找到参考资料：  
<https://github.com/chrislinan/cx-extractor-python>  
其基本思路是根据整个网页的文字分布判断哪一部分是正文，然后抽取出来。

2. 文本聚类、分类分析的相关知识课上一时难以完全理解清楚，课下查找了一些资料（见下网址）


[https://github.com/AimeeLee77/keyword\\_extraction](https://github.com/AimeeLee77/keyword_extraction)（这个对聚类讲的真的很详细）

<https://blog.csdn.net/u011587401/article/details/78323706>

<https://sspai.com/post/49121>（这个也很有用，sklearn 做分类）

<https://cloud.tencent.com/developer/article/1082154>

<https://yq.aliyun.com/articles/73677>

（网上的参考资料和方法很多，看来看去还是有些摸不着头脑  后来请教了一位本科计算机专业的同学，最后才搞清楚文本处理这一部分的思路）

LDA 聚类得到每篇文章的描述关键词，然后转化成词向量建立分类模型，用 RNN 做分类。

LDA 相关知识：<https://baike.baidu.com/item/LDA/13489644>

## 四、 项目实践

在项目实践中用到许多第三方 python 库。比如比较著名的中文分词组件 jieba, gensim(用于主题模型、文档索引和大型语料相似度索引), pandas (用于高效处理大型数据集、执行数据分析任务) 至于 sklearn 这类不是非常了解的内容, 除了网上参考代码之外, 一直在请教更为专业的同学。

## 五、 个人感想

- Python 真的非常强大, 学好 python 高效率高质量地能做很多事情。
- Python 的第三方模块能解决许多问题, 一定要多查资料, 多看例子, 多思考。
- 平时的代码练习非常重要, 有许多内容课上提到过但是当时没有深入理解, 课后就要花更多时间。
- 分解问题, 理清逻辑很重要, 不懂的地方多请教更加专业的同学。
- 千万不要懒, 千万不要拖延!