预习任务

为了使学生更好的掌握项目所涉及到的基础知识，对机器学习-文本挖掘有大概的认识和了解，更快地进入项目角色，完成项目内容，制定本预习方案，包含 4 周预习内容。

一．第一周

1. 配置 Java/matlab/Python 环境，推荐使用 python。

环境说明：python2.7 或 python3

下载地址：https://www.python.org/downloads/

python IDE：

pycharm，社区版已够用，下载地址：

http://www.jetbrains.com/pycharm/download/#section=windows

2. 小程序练习：

输入两个字符串，输出两个字符串的最长公共子串。

二． 第二周

1. 了解机器学习和数据挖掘的基本步骤，分为哪几类问题。

2. 了解中文分词与英文分词。学习并至少掌握一种分词工具，如结巴分词，ICTCLS 等。

3. 选择《威尼斯商人》剧本，即 shakespeare-merchant.trec 文件夹下任意一个文件，

（1）统计词项数量，文档数量，每个词在每篇文章中的出现次数。

（2）对于给定某个关键词，能够检索出在哪篇文章中出现过。

每个文档有如下格式：

```
<DOC>
<DOCNO>StringID</DOCID>
<title>The Title</title>
Content goes here
<speaker> Name </speaker> Speech
</DOC>
```

其中，DOC 标识文档起止位置，DOCNO 为文档字符串 ID，title 为标题。

三．第三周

1. 了解文本表示的方法（布尔表示，词频表示，tf-idf 表示，word2vec 等）

2. 将 videotitle 里面的文本用不同方法表示成向量。

3. 对于向量化之后的 videotitle 数据，输入一个视频标题，输出与其相似的前 10 个视频标题，比较在不同文本表示方法下输出结果的效果。

四．第四周

1. 了解爬虫原理。

2. 实现一个简单的爬虫程序或者跑通开源的爬虫代码。

3. 尝试爬取一些新闻网页，提取相关信息，能按时间、热度（需要自己定义）等属性对网页进行排序，最好有界面展示，输入一些 URL，能显示爬取结果。