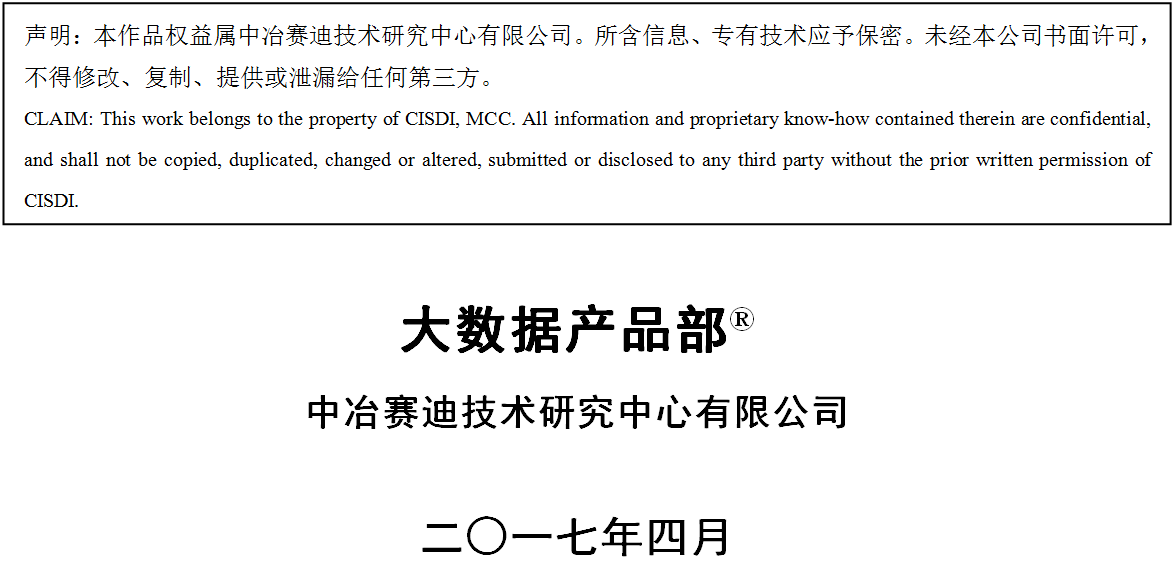
|  |
| --- |
| **CISDI中冶赛迪** |
| 测试题目——爬虫&NLP |
| 主 研 人：Elijah  参 研 人：Jacky |
| 审 核 人：Mark |
|  |



目 录

[1.概述 3](#_Toc480548682)

[2.测试题目 3](#_Toc480548683)

[2.1任务1—环评信息爬取 3](#_Toc480548684)

[2.2任务2—地点分级处理 4](#_Toc480548685)

[2.3任务3—环保资金信息抽取 5](#_Toc480548686)

[3.交付物要求 5](#_Toc480548687)

[4.提示 6](#_Toc480548688)

# 1.概述

本文档包含了NLP与爬虫方向实习生的测试题目，请各位同学使用Python，严格按照以下要求来完成测试。

# 2.测试题目

## 2.1任务1—环评信息爬取

1）任务描述

使用python的**Scrapy框架**爬取国家环评网中，“建设项目环境影响评价”栏目下，“拟审查项目公示”页面中发布日期在2016/1/1 到2017/4/5期间（有99个项目）的以下数据：“项目名称”、“建设地点”、“建设单位”、“项目概况”、“发布日期”。



**图2.1-1 需爬取信息网页截图**

注：国家环评网地址：<http://www.zhb.gov.cn/home/rdq/hjyxpj/>

2）输出要求：

将爬取结果以csv文件输出，文件名为“项目拟审核公示表.csv”。文件中包括5个字段，分别是“项目名称”、“建设地点”、“建设单位”、“项目概况”、“发布日期”。

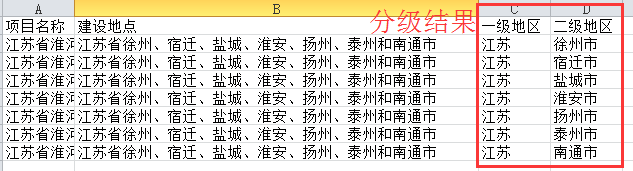
注：输出结果参见附件“参考\_项目拟审核公示表.csv”。

## 2.2任务2—地点分级处理

1）任务描述

对“项目拟审核公示表.csv”文件中“建设地点”中的地点信息进行分级处理。提取出每个地点中的一级地区和二级地区（见“省市级名称表.xlsx”）。

当一个项目的建设地点可能有多个时，其处理方式如图2.3-1所示。



**图2.3-1 地点分级结果示例图**

4）输出要求

将分级的结果以csv文件输出，文件名为“项目地点信息表.csv”。

注：参见附件“参考\_项目地点信息表.csv”所示。

## 2.3任务3—环保资金信息抽取

1）任务描述

提取“项目拟审核公示表.csv”（2.1小节得到的结果）文件中“项目概况”中各项目的总投资额、环保投资以及环保投资占比。

注：总投资、环保投资和环保占比的表述形式在各个项目中有所不同；若文中无总投资、环保投资或者环保占比，则单元格置空。

2）任务要求

本题分为以下三个层次的方法，方法依次递进，分值也依次递进。

* 基于正则表达式的抽取：要求准确度大且正则表达式尽量少。
* 基于NLP人工规则的抽取：要求准确度大且逻辑层次清晰。
* 基于NLP统计学习的抽取：不严格要求准确度，但要求一套可行的模型。（**至少提供一个严格可行的实现思路**）

注：**准确度**指的是抽取出的人为确认的**正确信息**占总信息的百分比。本题评判标准有三个，分别对应总投资额、环保投资和环保投资占比各自的准确度。

3）输出要求：

将抽取的结果以csv文件输出，文件名为“项目投资额信息表.csv”。文件中包括以下5个字段，分别是“项目名称”、“项目概况”、“总投资额”、“环保投资额”、“环保投资占比”。

注：输出结果参见附件“参考\_项目投资额信息表.csv”。

# 3.交付物要求

1）仔细参考《0 交付文档（标准格式-爬虫NLP方向）》，针对以上问题撰写一篇文档。

2）针对以上问题分别编写一个python程序代码（需有必要的代码注释），需注意代码的规范性、可读性、模块化等。

# 4.提示

Python中，中文NLP处理包有jieba，thulac，pyltp，snownlp等，请选择合适的包进行处理。