**自然语言处理设计报告 - 项目计划**

概要：每个学校有自己的特色，例如东莞理工学院有“北街”、“知行课”等。我们希望知道每个学校的特色是什么。为了实现这个，我们可以利用贴吧的语料信息：首先选取多个不同学校，然后爬取每个学校的的贴吧内容，代表该学校的语料。接着，对于每个学校的语料，我们使用语言学的统计方法进行无词库抽词，得到该学校的所有词语，以及每个词语的打分信息。最后，我们将设计一个算法对比不同学校的词语信息，选举出每个学校的特色词。

本项目共由三部分组成。

1. **第一部分：贴吧爬虫**

文件由三部分组成：(1) 爬虫的任务列表 (2) 爬下来的语料数据 (3) 爬虫的代码

1. 爬虫的任务列表（crawler\_tasks.txt）

表示计划完成的爬虫任务总量。格式多行，每行格式是“贴吧名称 总字数要求”。（“贴吧名称”和“总字数要求”中间用空格隔开，可以是多个空格）。例子如下：

---- crawler\_tasks.txt ----

东莞理工学院 100000

上海大学 50000

山东大学 80000

==== crawler\_tasks.txt ====

注1：每次爬虫启动的时候要统计已经爬出的所有txt的总字数，然后判断是否需要继续爬。

注2：每个帖子最多爬10页，以防止单个帖子权重过大。

1. 爬下来的语料数据

使用如下目录结构来储存语料（project是代码和crawler\_tasks.txt所在的目录）：

- project

--- corpus

----- 东莞理工学院

------- 9359076910.txt

------- 9359027226.txt

------- 9359015994.txt

----- 上海大学

------- 9358773558.txt

------- 9359429578.txt

注意txt文件的数字代表帖子的id（在网址里面有）。

1. 爬虫的代码

共一个文件：tieba\_crawler.py

包含以下函数：

1. tiebaPage2PostIDs

参数：

tieba\_name 字符串，贴吧名称

page\_no 整数，贴吧页号

返回值：

一个字符串列表，包含贴吧该页面的所有帖子的id。例如[‘9359076910’, ‘9359027226’，...]

1. PostID2Text

参数：

post\_id 字符串，帖子的id，如’9359076910’。

page\_no 整数，帖子的页号

返回值：

一个字符串，包含帖子该页的所有用户生成内容。

1. TextQuantityCount

参数：

tieba\_name 贴吧名称

返回值：

一个整数，表示该贴吧已爬取的字符数。

1. fetch\_tasks

无参数

返回值：

一个字典，表示crawler\_tasks中指定的需要爬的贴吧名称和每个贴吧的总字数要求。例如：[‘东莞理工学院’ : 100000, ‘上海大学’ : 50000, ‘山东大学’ : 80000]。

1. is\_already\_fetched

参数：

tieba\_name，字符串，贴吧名称

post\_id，字符串，帖子的id，如’9359076910’。

返回值：

一个bool值，表示该帖子是否已经爬过（已经有语料记录）。

1. clean\_text

参数：

text，字符串

返回值：一个字符串，为text去掉所有非中文字符后的字符串。

1. save\_text

参数：

text，字符串，只允许包含中文字符

tieba\_name，字符串，贴吧名称

post\_id，整数，帖子的id

无返回值

1. if \_\_name\_\_=='\_\_main\_\_': 部分

进行爬虫任务。流程：

1. 获取任务列表。
2. 获取每个需要爬的贴吧的已完成字数。
3. 遍历任务列表，逐个完成任务。
4. 对于每个任务，不断地随机从1-200选一个贴吧页号，然后获取该页面的所有帖子id，然后爬每个帖子。每爬完一个帖子，就更新该贴吧的字数信息，如果达到要求了就停止该任务并进行下一个任务。每个帖子的字符串合并并去除非中文字符后保存到txt文件。
5. 所有任务完成后退出程序。
6. **第二部分：无词库抽词**

基本原理：根据

<https://www.matrix67.com/blog/archives/5044>

进行无词库抽词，并且得到每个词的词频。

文件由两部分组成：(1) 词语统计结果 (2) 词语统计代码。

1. 词语统计结果

使用以下目录结构：

- project

--- words

----- 东莞理工学院.txt

----- 上海大学.txt

----- 山东大学.txt

每个txt文件由多行组成，每行格式为“词语 词频”，其中词语是一个中文字符串，词频是一个浮点数，是该单词出现次数/文本总字数。例子：

---- 东莞理工学院.txt ----

鼠鼠 0.00233245

北街 0.00092335

是 0.21342

[后面省略多行]

==== 东莞理工学院.txt ====

1. 词语统计代码

共一个文件words\_extract.py

包含以下函数：

1. extract\_words

参数：

text，字符串，read\_text返回值

返回值：一个字典，表示所有潜在的词和每个词的打分信息。打分公式：score = 词频 \* log2(凝聚程度)^3 \* 灵活程度^3。类似[‘鼠鼠’ : 0.00233245, ‘北街’ : 0.00092335, ‘是’ : 0.21342]

1. calc\_entropy

参数：

char\_list，字符列表，可能是左邻字或右邻字的列表

返回值：

一个浮点数，表示这个列表的信息熵

1. read\_text

参数：

tieba\_name，字符串，贴吧名称

返回值：一个字符串，由该贴吧所有已爬取语料连接形成

1. save\_result

参数：

tieba\_name，贴吧名称

result，一个字典，为extract\_words的返回值

无返回值

1. fetch\_tieba\_names

无参数

返回值：一个字符串列表，是根据已爬取语料的文件夹名称得到的所有贴吧名称。

1. if \_\_name\_\_=='\_\_main\_\_': 部分

进行词语抽取和统计。步骤：

1. 获取所有已爬取语料的贴吧名称
2. 对于每个贴吧的语料，读取、抽词、保存。
3. **第三部分：特色词提取**

文件由两部分组成：(1) 特色词报告 (2) 特色词提取代码

1. 特色词报告

一个report.txt文件，在project目录下。格式如下：

---- report.txt ----

东莞理工学院吧特色词汇：北街、巴萨克、莞博。

上海大学吧特色词汇：首日教育、宝山。

==== report.txt ====

1. 特色词提取代码

共一个文件：feature\_word\_extract.py。

1. fetch\_tieba\_names

无参数

返回值：一个字符串列表，是根据已爬取语料的文件夹名称得到的所有贴吧名称。

1. read\_word\_result

参数：

tieba\_name，字符串，贴吧名称

返回值：

一个字典，表示从“words/贴吧名称.txt”读取的词和每个词的打分信息。类似[‘鼠鼠’ : 0.0233245, ‘北街’ : 0.092335, ‘莞专’ : 0.21342]

1. extract\_feature\_words

参数：

all\_word\_result：一个字典，字典的key是贴吧名称，字典的value是这个贴吧的read\_word\_result的返回值。

返回值：

一个字典，字典的key是贴吧名称，字典的value是一个字符串列表，包含该贴吧所有的特色词语。

功能：使用score = 词语分数 / 文档频率^3 给每个词的特色度打分，然后对于每个贴吧的语料，获得排位最高的10个单词，根据score的值从大到小排列。

1. save\_report

参数：

feature\_words，一个字典，extract\_feature\_words的返回值。

无返回值

功能：保存报告到report.txt

1. show\_report

无参数和返回值

功能：读取report.txt并输出到控制台。

1. if \_\_name\_\_=='\_\_main\_\_': 部分

进行特征词提取。流程：

1. 获取所有已爬取语料的贴吧名称。
2. 读取所有贴吧在第二部分的词语统计结果。
3. 提取特色词。
4. 保存报告。
5. 在控制台输出报告。