目录

* 标注语料类型分析
* Json字段含义
* Json格式示例
* 正文抽取处理示例

注 该数据用于百度-北大合作项目的研究，未经同意不得用于其他目的

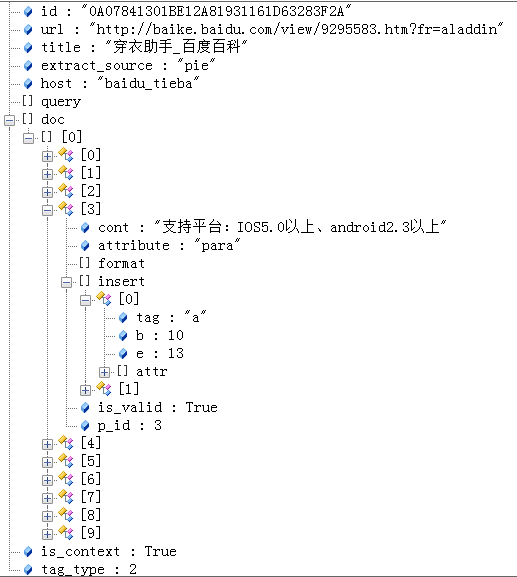
【标注语料类型分析】

|  |  |
| --- | --- |
| 类型 | 占比 |
| 无结构 | 60.7% |
| 显式 | 10.6% |
| 半显式 | 15.2% |
| 隐式 | 13.5% |

【Json字段含义】

|  |  |  |
| --- | --- | --- |
| Tag | 意义 | 数据格式 |
| url | 网页网址 | string |
| ID | 网页内容的唯一KEY，是url的md5值 | stirng |
| title | 网页标题（hao lib识别标题） | string |
| extract\_source | 用来区分网页正文的抽取来源 | string，取值为pie（规则抽取）或hao（模型抽取），PIE目前的覆盖：百度知道、百度贴吧、百度文库、新浪微博、腾讯微博、百度经验等 |
| host | 用来标注网页对应的网站 | string，目前仅覆盖了pie能召回的网站 |
| is\_context | 表示该网页是否为内容型，值由后处理模块输出 | boolean |
| tag\_TYPE | 标识内容型文档的具体类型 | int，0表示短文类，1表示社交类，2表示篇章类 |
| query(列表) | 大搜结果中展示此url的query集合 | 列表，每个元素标注一个query，包含query，pos两个Key |
| query（string） | query\_list中的某个query | string |
| Key | url在query展示结果中排序位置 | string |
| p\_id | 段落唯一编号 | int |
| doc | 正文内容（hao lib识别正文） | 双层列表，首层每个元素为一个正文块，内层每个元素为正文中一段 |
| cont | 段落正文 | string |
| format | 段落中特殊格式的标注 | 列表，每个元素为段落中一个特殊格式的标注，每个元素包含b，e，tag三个Key |
| insert | 段落中插入内容的标注 | 列表，每个元素标注为段落中的插入内容，每个元素包含b，e，tag三个Key |
| b | 表示特殊格式/插入内容的开始字符编号 | int |
| e | 表示特殊格式/插入内容的结束字符编号 | int |
| tag | 表示特殊格式/插入内容的具体内容 | string |
| is\_valid | 表示该段是否属于正文，值由后处理模块输出 | boolean |
| p\_item | 表示该段是否是两个主题间的切分段,字段值为1表示为切分点，其他情况表示非切分点，该字段为人工标注字段 | 1/0 |
| doc\_type | 表示该篇章的类型，1：无结构，2：显式3：半显式 4：隐式，该字段为人工标注字段 | 1/2/3/4 |

【Json格式示例】



插入内容列表 格式内容列表格式类似

内层正文 表示每段内容  
分段标准以页面显示为准

外层正文 表示文本块

如贴吧中的多层楼

双层结构正文

【正文抽取处理示例】以python代码为例

**# 设置环境**

reload(sys)

sys.setdefaultencoding( "utf-8" )

**# json解析，系统输出json为gb18030格式**

uline = line.decode('gb18030').encode('utf8')

diJson = {}

try:

diJson = json.loads(uline)

except:

continue

**#获取正文**

liDoc = diJson['doc'] #外层正文

for aDoc in liDoc:

for aPara in aDoc: # 内层正文

sentId = aPara['p\_id']

sent = aPara['cont']

if len(sent)>0:

list.append('PARA'+str(sentId)+'\t'+sent) # 输出格式 *PARA*+*正文长度*+*TAB*+*正文内容*