周工作总结20190513-20190517

本周主要的工作是在做simhash去重的线上服务，遇到了一些实际的问题，然后都解决了，这一点还挺开心。

技术积累

1. 计算两个向量的Jaccard距离

```
def jaccard_sim(a, b):
    unions = len(set(a).union(set(b)))
    print(unions)
    intersections = len(set(a).intersection(set(b)))
    print(intersections)
    return 1. * intersections / unions
```

2. Jmeter工具，可以用java发送请求。
3. ES中的filters的使用方法

   最外层一定要是and，里面一层只有一个条件的时候要用or。

```
"filters":
    {"and":
        [
            {"or":
                [{"field": "task_id",
                "action": "not",
                "value": "18010635"
                }]
            }
        ]
    }
```

   比如上面这个case找到的结果就是task_id不等于"18010635"的记录。

4. hive中**regexp_replace**方法的使用

在hive中查询内容要解析json的时候，可能需要替换字符，比如说"\t"会影响文件的分割，所以要把"\t"替换成一个其他的符号。如果有两个字符同时被替换成某个符号，则可以在第一个参数中使用"|"符号，比如regexp_replace("foobar", "oo|ar", "") returns 'fb'.

异常问题

1. ValueError: Invalid control character at: line 13862 column 32 (char 151679)

   出现原因：在python读取的一个文件里有一个特殊字符，打印出来是^B(类似吧)

   解决方法：**添加strict=False非严格模式**

   ```
   json.loads(jsonstr,strict=False)  # 使用的是python3.6
   ```

2. 'utf-8' codec can't decode byte invalid start byte

   出错原因：这个问题是浩宇发现说他用Jmeter请求了我给的case，出错了，出错在decode("utf8")那一句。我用postman和python程序请求都没有问题，然后推测是Java和python的编码方式不同吧，所以找业务方用Java接口调用了一遍也没问题。最后发现原来是因为Jmeter默认的编码为ISO-8859-1，不是utf8的，我们decode成utf8当然会出错。

   解决方法：中间有一个错误的尝试，就是给decode加了参数"ignore"，虽然没有报错，但是结果是不一致的，比如python接口返回的是

   ```
   {"code": 0, "msg": "success", "data": {"result": {"code": 0,
   "prob_pass": 0.5779587883808025, "prob_fail":
   0.4220412116191974, "rule": 0}, "detail": []}, "request_id":
   "J6t3kQb3JI903C94j45PR32FEkktn6X220190515101603"}
   ```

   而Jmeter返回的结果是：

   ```
   {"code": 0, "msg": "success", "data": {"result": {"code": 1,
   "prob_pass": 0.353832181296122, "prob_fail": 0.646167818703878,
   "rule": 0}, "detail": []}, "request_id": "23456789"}
   ```

   看起来是正常返回了，但是结果出错了，如果没有用Python的比对，很难发现其实是有问题的，所以decode的ignore参数要慎用！！

   最终的解决方案其实是修改Jmeter那条请求的content encoding参数，或者修改Jmeter的默认编码格式。

个人反思

1. 现在写代码的感觉确实比之前好一些了，写类，还有写服务没有那么费劲了。有一点不太好的地方是提前对于结构的设计还不够，总是中间调整函数调用，或者增删函数等等，还是应该一开始先想清楚怎么写，再动笔。
2. 这个周在查"异常问题2"的时候花费了得有半天的时间，虽然最后确认了问题在哪儿，但是花费的时间有点太多了。所以在**解决问题的思路上还是有些问题**。一开始一直卡在为什么转utf8失败了，怎么让它成功的转过来，然后各种尝试decode的不同方法，什么bytes转str啊，decode加参数啊什么的，那既然是java接口的报错，是不是应该先对比一下两个不同的java程序调用的结果，就很容易发现问题在哪儿呢，所以这一次算是一个借鉴吧。

3. 这个周很开心的一个地方，赵群也遇到了hive里取数据被"\t"把数据截断的情况，我把regexp_replace函数推荐给他，第一次感觉自己的积累还是有点用的。嘿嘿。继续努力！