周工作总结20190603-20190606

这周先是想优化看点的文章主题分类，后面就是优化去重方案。

技术积累

1. python3里bytes和str互转

   logging模块在记录日志的时候，输出是utf8编码的字节流，也就是形如"\x...\x...\x..."的形式，因此对于post请求里的bytes格式的字节流要做转码，这里要注意的一点是：**在bytes和str的互相转换过程中，实际就是编码解码的过程，必须显式地指定编码格式。**比如bytes转str，可以用`str(s, encoding="utf8")`,如果直接用`str(s)`就会生成"\\xe4\\xb8"这种。示例代码如下：

   ```
   >>> b
   b'\xe4\xb8\xad\xe6\x96\x87'
   >>> s1 = str(b)
   >>> s1
   "b'\\xe4\\xb8\\xad\\xe6\\x96\\x87’"
   s1 = str(b, encoding='utf-8')
   >>> s1
   '中文'
   ```

2. 在写try，except的时候要验证一下try条件不满足的时候的报错，否则真的遇到的时候，except不能正常执行，服务就挂了。

3. 十进制转二进制的位数问题。

   我是在处理simhash算法生成的hash值时，由于中间记录的时候int型整数存储的格式比较容易存储，所以中间会把二进制01串转成int型。但是在转回来的时候，本来是64位的二进制，就不一定能转成64位了，需要自己写函数在前面补齐。

异常问题

1. 问题表现：文本中出现"\u0002"去不掉。

   解决思路：问题没有解决，问题出现的情况是这样的，在原文里并没有任何字符，但是分词结果出来的就有"\u0002"，而且用re.sub去不掉。比如这段原文：

   ```
   去年的140亿美元，而这一指标2016年曾经达到236亿美元的峰值
   ```

分词结果为：

```
去年 的 140 亿 美元 , 而 这\u0002一\u0002指 标 2016 年 曾经 达到 236 亿 美元 的 峰值
```

我查了一些博客，没有什么结论，只能先记录一下，下周再找文彬一起讨论一下吧。记录是：

```
用 Python 写就是 "\u0001", 用 mysql 則是 char(1), VIM 會用 ^A 表示 (\u0002 則用 ^B), 最妙的是, Firefox 會用一個小方格在裡面填入 0001 表示它, 所以寫到資料庫後, 用 phpMyAdmin 看也不會有問題, 真是太貼心了。
```

个人总结

1. 现在已经基本能养成及时检验自己代码的习惯了，但是总感觉不够系统，想学习一下有没有什么框架能用来验证代码效果，或者是自动写case那种。这算是一个todo吧。在之后的工作中要多留意。
2. 本周想优化下fasttext的分类效果，结果发现除了去下停用词竟然没有任何思路，所以还是要从原理上多了解用的模型，所以这周就主要花精力看了下fasttext的原理。