周工作总结2019/04/22-2019/04/30

这个周是五一前的八个工作日算一周了，完成的工作有：simhash重复效果验证，规则补充开发及上线，然后是看点这边启动了文章主题分类，输出了三种方案的模型。

技术积累

1. HDFS

   在使用公司submarine系统训练模型时用到了HDFS的命令：

   查看集群文件：

   ```
   hadoop fs -ls <url>
   ```

   获取集群文件：

   ```
   hadoop fs -get hdfs://nn-cluster/user/XXX/tensorflow/submarine
   ```

   删除集群文件：

   ```
   hadoop fs -rm -r hdfs://nn-
   cluster/user/strategy/tensorflow/submarine/yolov3/trained_weight
   s_16_120_0.01_0313.h5
   ```

2. jieba分词

   了解到jieba分词的三种模式：

   - 精确模式，试图将句子最精确地切开，适合文本分析；
   - 全模式，把句子中所有的可以成词的词语都扫描出来, 速度非常快，但是不能解决歧义；
   - 搜索引擎模式，在精确模式的基础上，对长词再次切分，提高召回率，适合用于搜索引擎分词。

   Example：

```
# encoding=utf-8
import jieba
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))  # 全模式
seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))  # 精确模式
seg_list = jieba.cut("他来到了网易杭研大厦")  # 默认是精确模式
print(", ".join(seg_list))
seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")  # 搜索引擎模式
print(", ".join(seg_list))
```

Output:

```
【全模式】：我/ 来到/ 北京/ 清华/ 清华大学/ 华大/ 大学

【精确模式】：我/ 来到/ 北京/ 清华大学

【新词识别】：他，来到，了，网易，杭研，大厦　　（此处，“杭研”并没有在词典中，但是也被Viterbi算法识别出来了）

【搜索引擎模式】：　小明，硕士，毕业，于，中国，科学，学院，科学院，中国科学院，计算，计算所，后，在，日本，京都，大学，日本京都大学，深造
```

3. 正则

   这周在查上周那个正则匹配出[“”,“”,“”]的问题时，学到了一个新的知识点：

   ```
   当该字符紧跟在任何一个其他限制符（*,+,?，{*n*}，{*n*,}，{*n*,*m*}）后面时，匹配模式是非贪婪的。非贪婪模式尽可能少地匹配所搜索的字符串，而默认的贪婪模式则尽可能多地匹配所搜索的字符串。例如，对于字符串 “oooo”，“o+” 将尽可能多地匹配 “o”，得到结果[“oooo”]，而“o+?” 将尽可能少地匹配“o”，得到结果['o', 'o', 'o', 'o’]
   ```

4. shell压缩与解压缩

   zip：

   ```
   Zip -r myfile.zip ./*
   ```

   unzip:

```
unzip -o -d /home/sunny myfile.zip  把myfile.zip文件解压到/home/sunny/
-o:不提示的情况下覆盖文件
-d:-d /home/sunny 指明将文件解压缩到/home/sunny目录下
```

tar.gz和tar.bz2:

```
tar -zxvf ×××.tar.gz
tar -jxvf ×××.tar.bz2
```

5. pandas

   a.安装：

   pandas需要依赖处理Excel的xlrd模块，所以我们需要提前安装这个，安装命令是：pip install xlrd。下一步就是pip install pandas。

   b.生成dataframe，写入excel：

```
import pandas as pd

# create some Pandas DateFrame from some data
df1=pd.DataFrame({'Data1':[1,2,3,4,5,6,7]})
df2=pd.DataFrame({'Data2':[8,9,10,11,12,13]})
df3=pd.DataFrame({'Data3':[14,15,16,17,18]})
All=[df1,df2,df3]
# create a Pandas Excel writer using xlswriter
writer=pd.ExcelWriter('test.xlsx')

df1.to_excel(writer,sheet_name='Data1',startcol=0,index=False)
df2.to_excel(writer,sheet_name='Data1',startcol=1,index=False)
df3.to_excel(writer,sheet_name='Data3',index=False)
```

异常问题

1. 用python版的fasttext训练预测文章类别的时候，输出不是一个类别，而是很长的一个列表，看起来像是没个词都有一个类别,就是类似[‘000’,’109’,’000’]这样。

   解决：经过看官方文档，predict的输入应该是列表类型，里面放的是要要预测的字符串，而不是直接输入字符串。

2. 在用pycharm生成文件时，发现with open一个文件，文件明明就在，但是就是报IO错误，复制了一遍还是不行

解决：在折腾的时候发现设置interpreter的页面，有一个working directory，改成当前目录就好了。

个人反思

1. 对fasttext的原理了解太少，所以在调整参数的时候没有方向，所以要把论文和一些好的笔记博客都看一下。
2. 对于数据分析的工具pandas和matplotlib需要学习一下，写一篇学习笔记。