商品关键词权重选择

前面我们分别使用TEXTRANK和TFIDF计算出了每件商品的关键词以及相应的权重，但发现两者好像都有一点瑕疵

textrank的权重

sku_id	industry	tag	weights
148	电子产品	高度	1.0
148	电子产品	终端	0.9883639010223446
148	电子产品	WPOS	0.9468601431427544
148	电子产品	触摸屏	0.86760348122463
148	电子产品	森锐	0.8641422469338305
148	电子产品	收银机	0.8604138186822986
148	电子产品	智能	0.8531525111008649
148	电子产品	业务	0.8492750654830078
148	电子产品	身份	0.7319843682220414
148	电子产品	包邮	0.7023074720861229
148	电子产品	读卡器	0.6095231846952228
148	电子产品	购物	0.5376711388516783
148	电子产品	正品	0.5362617954388065
148	电子产品	数码配件	0.4855585880270606
148	电子产品	数码	0.4839580467664243
463	电子产品	读卡器	1.0
463	电子产品	颜色	0.8520815042804352
463	电子产品	安卓	0.5855862956518749
463	电子产品	手机	0.47418584773791483
463	电子产品	电脑	0.43966846994965897

tfidf的权重

```
+------+----+------------------+
|sku_id| tag|           weights|
+------+----+------------------+
|   148|  智能|0.12920924724659527|
|   148|  数码|0.09416669674352422|
|   148| 读卡器|0.22691875231942266|
|   148|  正品| 0.1984394909665287|
|   148|数码配件| 0.2300917543361638|
|   148|  购物| 0.2408169399138654|
|   148|  包邮| 0.2749585054126381|
|   148| 触摸屏|0.31885875801848024|
|   148|  高度| 0.3896366762502419|
|   148| 收银机|0.45724967270727696|
|   148|  终端|  0.515996300178136|
|   148|  业务|  0.537514526329006|
|   148|  身份| 0.6107553455735467|
|   148|  森锐| 0.6107553455735467|
|   148|WPOS| 0.6672418695993603|
|   463|  颜色|0.08984603313709096|
|   463|  黑色|0.20849464181052813|
|   463|  电脑|0.17166530284651157|
|   463|  手机|0.20699498923484225|
|   463|  数码|0.05231483152418012|
+------+----+------------------+
```

解决方案：

- 最好的解决方法：优化分词效果，比如将部分不太相关词语设为停用词，如业务、高度等词
- 较次的解决方案：将tfidf权重和textrank权重加权平均，降低两者都导致的过高或过低的词概率
  - 比如148号sku商品，如果使用tfidf，那么发现“读卡器”，比“购物”这样的词权重还低；如果使用textrank，发现“数码”“数码配件”这样的词比“正品”“包邮”这样的还低，但如果相加求平均后，可以使用这个问题得到解决。

In [1]:

```
import os
# 配置pyspark和spark driver运行时 使用的python解释器
JAVA_HOME = '/root/bigdata/jdk'
PYSPARK_PYTHON = '/miniconda2/envs/py365/bin/python'
# 当存在多个版本时，不指定很可能会导致出错
os.environ['PYSPARK_PYTHON'] = PYSPARK_PYTHON
os.environ['PYSPARK_DRIVER_PYTHON'] = PYSPARK_PYTHON
os.environ['JAVA_HOME'] = JAVA_HOME
# 配置spark信息
from pyspark import SparkConf
from pyspark.sql import SparkSession

SPARK_APP_NAME = "mergeTagWeights"
SPARK_URL = "spark://192.168.58.100:7077"

conf = SparkConf()    # 创建spark config对象
config = (
    ("spark.app.name", SPARK_APP_NAME),    # 设置启动的spark的app名称，没有提供，将随
    ("spark.executor.memory", "2g"),    # 设置该app启动时占用的内存用量，默认1g，指一
    ("spark.master", SPARK_URL),    # spark master的地址
    ("spark.executor.cores", "2"),    # 设置spark executor使用的CPU核心数，指一台虚拟
    ("hive.metastore.uris", "thrift://localhost:9083"),    # 配置hive元数据的访问，否

    # 以下三项配置，可以控制执行器数量
#     ("spark.dynamicAllocation.enabled", True),
#     ("spark.dynamicAllocation.initialExecutors", 1),    # 1个执行器
#     ("spark.shuffle.service.enabled", True)
#   ('spark.sql.pivotMaxValues', '99999'),  # 当需要pivot DF，且值很多时，需要修改，默
)
# 查看更详细配置及说明：https://spark.apache.org/docs/latest/configuration.html

conf.setAll(config)

# 利用config对象，创建spark session
spark = SparkSession.builder.config(conf=conf).enableHiveSupport().getOrCreate()
```

In [9]:
```
sql='''
select a.sku_id,a.industry,a.tag,a.weights textrank,b.weights tfidf,(a.weights + b.wei
from sku_tag_weights a
join
(select sku_id,tag,weights from sku_tag_tfidf b)
on a.sku_id=b.sku_id and a.tag=b.tag
'''
df = spark.sql(sql)
df.show()
```

```
+------+--------+------+------------------+------------------+------------------+
|sku_id|industry|   tag|          textrank|             tfidf|           weights|
+------+--------+------+------------------+------------------+------------------+
|    85|  电子产品|  数码|0.37932517906026586|0.07434212900804543|0.22683365403415
565|
|   172|  电子产品|  USB2| 0.5258851966532786| 0.4902682570089824| 0.508076726831130
5|
|   182|  电子产品|  型号| 0.8279531788256792| 0.4417105502575884| 0.6348318645416
338|
|   190|  电子产品|  雷克沙|0.18932069864134093|0.25239101355870736|0.22085585610000
2414|
|   271|  电子产品|  数码配件|0.18629417960331815|0.16435125309725987|  0.1753227163
50289|
|   282|  电子产品|  香槟色|0.20939011016154468|  0.450895660720322|0.33014288544409
3334|
|   305|  电子产品|  星空|0.25201085335696144|0.32303206716234273| 0.2875214602596
521|
|   312|  电子产品|  SONY| 0.6888028719429866| 0.4221730662623697| 0.555487969102678
2|
|   326|  电子产品|  川宇| 0.3939664586744098| 0.8599491424894472| 0.6269578005819
285|
|   334|  电子产品| TOPSSD|0.40044920086809455| 0.9027798790978679| 0.651614539982981
2|
|   334|  电子产品|  天硕|0.33249685251643607| 0.9027798790978679|  0.617638365807
152|
|   351|  电子产品|  数码配件| 0.3284858938211362|0.43142203938030715|0.37995396660000
72167|
|   370|  电子产品|  功能|0.17728834473611424|0.20335407147331752|0.19032120810471
587|
|   403|  电子产品|  数码配件| 0.3543503711641944| 0.2465268796458898|0.30043862540505
04207|
|   410|  电子产品|MicroSD| 0.3341093557715508|0.35569624565338087| 0.344902800712465
8|
|   414|  电子产品|  颜色| 0.1632912209999795|0.03465489849573508|0.09897305974785
729|
|   441|  电子产品|  数码|0.09662172868610486|0.04556453068235043|0.07109312968422
765|
|   450|  电子产品|  数码配件| 0.4159506040114997|0.21571101969015358| 0.3158308118508266|
|   456|  电子产品|  数码|0.26235279061473155|0.06726192624537444|  0.1648073584300
53|
|   458|  电子产品|  发货|0.07446956339855329| 0.1804562973837209| 0.1274629303911
371|
+------+--------+------+------------------+------------------+------------------+
only showing top 20 rows
```

In [12]:
```
df.registerTempTable('tempTable')
```

In [13]:
```
sql='''
create table if not exists sku_tag_merge_weights(
sku_id int,
industry string,
tag string,
textrank double,
tfidf double,
weights double
)
'''
spark.sql(sql)
```

Out[13]: DataFrame[]

In [14]:
```
spark.sql('insert into sku_tag_merge_weights select * from tempTable')
```

Out[14]: DataFrame[]

In [15]:
```
spark.sql('select * from sku_tag_merge_weights where sku_id=148').show()
```

```
+------+--------+----+-----------------+-----------------+------------------+
|sku_id|industry| tag|          textrank|             tfidf|           weights|
+------+--------+----+-----------------+-----------------+------------------+
|   148|  电子产品|  智能|0.8531525111008649|0.12920924724659527| 0.4911808791737301|
|   148|  电子产品|WPOS|0.9468601431427544| 0.6672418695993603| 0.8070510063710573|
|   148|  电子产品|  包邮|0.7023074720861229| 0.2749585054126381| 0.4886329887493805|
|   148|  电子产品|  购物|0.5376711388516783| 0.2408169399138654|0.38924403938277186|
|   148|  电子产品|  终端|0.9883639010223446|  0.515996300178136| 0.7521801006002403|
|   148|  电子产品|  身份|0.7319843682220414| 0.6107553455735467|  0.671369856897794|
|   148|  电子产品|  正品|0.5362617954388065| 0.1984394909665287| 0.3673506432026676|
|   148|  电子产品| 触摸屏| 0.86760348122463|0.31885875801848024| 0.5932311196215552
|
|   148|  电子产品|  森锐|0.8641422469338305| 0.6107553455735467| 0.7374487962536886|
|   148|  电子产品|  数码|0.4839580467664243|0.09416669674352422|0.28906237175497423|
|   148|  电子产品| 收银机|0.8604138186822986|0.45724967270727696| 0.6588317456947878
|
|   148|  电子产品|  高度|              1.0| 0.3896366762502419|  0.694818338125121|
|   148|  电子产品| 读卡器|0.6095231846952228|0.22691875231942266| 0.4182209685073227
|
|   148|  电子产品|  业务|0.8492750654830078|  0.537514526329006| 0.6933947959060069|
|   148|  电子产品|数码配件|0.4855585880270606| 0.2300917543361638|0.35782517118161217|
+------+--------+----+-----------------+-----------------+------------------+
```