Paul 7 - enrich do...

```
%pyspark                                                                          FINISHED

# Zeppelin notebook to enrich domain summaries (from Paul 5) with examples (from Paul 6)
# and topic metadata (from Tom 1)
# PJ - 3 November 2017

import boto
from pyspark.sql.types import *

# Load Domain Summaries DF in Gzip files (from Paul 5)
loadURL="s3://billsdata.net/CommonCrawl/domain_summaries6/" # Split into 20 files (as op
codec="org.apache.hadoop.io.compress.GzipCodec"
summary_df=spark.read.format('com.databricks.spark.csv').options(header='true', codec=co
summary_df.show(3)
summary_df.cache()
summary_df.rdd.getNumPartitions()
```

```
+--------------+--------+------------+---------+----------+-------------+------+----
--+
|payLevelDomain|numHosts|pldIsHostFlag|pldLinksIn|pldLinksOut|wasCrawledFlag|hcRank|prRa
nk|
+--------------+--------+------------+---------+----------+-------------+------+----
--+
|        37.ac|       1|        true|        1|      null|        false|49.503|47.7
23|
|     equal.ac|       1|        true|        3|         3|         true|69.571|68.0
98|
|     happy.ac|       1|        true|        2|         6|         true|19.030|69.9
99|
+--------------+--------+------------+---------+----------+-------------+------+----
--+
only showing top 3 rows
20
```

```
%pyspark                                                                          FINISHED

# Load examples dataframe (from Paul 6)
examplesURI="s3://billsdata.net/CommonCrawl/domain_examples4/"
example_df=spark.read.load(examplesURI)
example_df.show(10)
example_df.cache()
example_df.rdd.getNumPartitions()
```

```
+----+------------+-----------------+--------------------+
| PLD|exampleHosts|exampleInLinkPlds|  exampleOutLinkPlds|
+----+------------+-----------------+--------------------+
|null|        null|             null|     [instagram.com]|
|null|        null|             null|       [godaddy.com]|
|null|        null|             null|       [4.cn, 51.la]|
|null|        null|             null|[food.it, compro....|
|null|        null|             null|[web2printsoftwar...|
|null|        null|             null|[nu.nl, rtl.nl, g...|
|null|        null|             null|       [apdesign.be]|
|null|        null|             null|           [ovh.net]|
|null|        null|             null|[zonneschermenams...|
|null|        null|             null|    [googleapis.com]|
+----+------------+-----------------+--------------------+
only showing top 10 rows
204
```

```
%pyspark                                                    FINISHED

# Remove square brackets from list output and remove Nulls
from pyspark.sql.functions import udf
def remove_brackets(egs):
    return str(egs).replace('[', '').replace(']', '')
print(remove_brackets("[bla, bla]"))
udf_remove_brackets = udf(remove_brackets, StringType())

example_df2=example_df.filter(example_df.PLD.isNotNull()).withColumn("tmp", udf_remove_
example_df3=example_df2.withColumn("tmp", udf_remove_brackets("exampleInLinkPlds")).dro
example_df4=example_df3.withColumn("tmp", udf_remove_brackets("exampleHosts")).drop("ex
example_df4.show(10)
example_df4.count()
```

```
bla, bla
+-------------------+--------------------+--------------------+------------------+
|                PLD|        exampleHosts|   exampleInLinkPlds|exampleOutLinkPlds|
+-------------------+--------------------+--------------------+------------------+
|            0-0a.me|         dev.0-0a.me|     krijnhoetmer.nl|              None|
|    0-3-suikast.org|     0-3-suikast.org| list-of-domains.org|              None|
|         0-5-0.info|          0-5-0.info| list-of-domains.org|              None|
|            0-6.biz|             0-6.biz|     beyondwhois.com|              None|
|      0-6health.com|       0-6health.com|      allthecom.info|              None|
|            0-a.net|             0-a.net|      allthecom.info|              None|
|  0-apr-lifetime.cn|establish-credit-...|             fc2.com|              None|
|0-aprcreditcards.us| 0-aprcreditcards.us| list-of-domains.org|              None|
|     0-artlove.info|      0-artlove.info| list-of-domains.org|              None|
|         0-blog.com|     web2.0-blog.com|blogspot.com, seo...|              None|
+-------------------+--------------------+--------------------+------------------+
only showing top 10 rows
90839924
```

```
%pyspark                                                          FINISHED

# Join with Original summaries
example_df.unpersist()
example_summary_df=summary_df.join(example_df4, summary_df.payLevelDomain==example_df4.
summary_df.unpersist()
example_summary_df.dropDuplicates().sort("numHosts",ascending=False).show(100)
example_summary_df.count()
```

```
|       woxingma.com|     999|         true|       202|       230|         true|69.03
9|96.955|woxingma.com, 0xc...|butlercarcare.com...|yudingall.cn, wuz...|
|         eaek22.com|     999|         true|        37|         2|         true|89.66
8|89.051|eaek22.com, a.eae...|blogspot.com, yww...|yahoo.com, ticrf....|
|        jhjhkk6.com|     999|         true|         9|         4|         true|86.46
0|70.916|jhjhkk6.com, a.jh...|blogspot.com, ddf...|yahoo.com, aatk5....|
|          azbuz.com|     999|         true|       460|      null|        false|99.82
7|98.870|azbuz.com, 06ct.a...|blogspot.com, sec...|                None|
|         hkfr55.com|     999|         true|        16|       268|         true|89.65
1|79.461|hkfr55.com, a.hkf...|blogspot.com, klk...|cilis.net, kk69mm...|
|          kgfs1.com|     999|         true|        11|        18|         true|86.46
0|68.908|kgfs1.com, a.kgfs...|ddft1.com, kkyytt...|yahoo.com, ishow9...|
|          hhry1.com|     999|         true|        12|         3|         true|86.46
0|75.488|hhry1.com, a.hhry...|ddft1.com, kkyytt...|ygnm4.com, ticrf....|
|       uthome51.com|     999|         true|       109|         9|         true|89.60
3|91.288|uthome51.com, a1....|bb198.net, hwe7.c...|gogo176.com, macr...|
|          kgfs2.com|     999|         true|        12|        18|         true|89.94
3|69.972|kgfs2.com, a.kgfs...|blogspot.com, kte...|yahoo.com, ishow9...|
```

```
%pyspark                                                          FINISHED

# Save Example Summaries as GZIP files, approx 100MB each.
outURI="s3://billsdata.net/CommonCrawl/domain_summaries_withexamples4/"
codec="org.apache.hadoop.io.compress.GzipCodec"
example_summary_df.coalesce(40).write.format('com.databricks.spark.csv').options(header
```

```
%pyspark                                                          FINISHED

# Load topic labels (from Tom 1)
loadURL="s3://billsdata.net/CommonCrawl/topic_model_1024_files/cc_index_page_topic_labe
topic_df=spark.read.load(loadURL)
topic_df.show(3)
print(topic_df.count())
topic_df.cache()
topic_df.rdd.getNumPartitions()
```

```
+--------------------+--------------------+--------------------+------------------+-----
---------------+--------------------+--------------------+--------------------+
|                host|                 url|              topic1|            score1|
topic2|              score2|              topic3|              score3|
+--------------------+--------------------+--------------------+------------------+-----
---------------+--------------------+--------------------+--------------------+
|billives.typepad.com|http://billives.t...|one_may_january_l...|0.3004067571438536|news_
business_dat...|0.17487022035440172|science_research_...| 0.12825866906681413|
|     docs.oracle.com|http://docs.oracl...|nbsp_windows_quot...|0.7843980004795321|forum
s_member_lik...|0.14079996570227268|add_cart_buy_pric...|0.058564617674095515|
|     e.il.tripod.com|http://e.il.tripo...|park_john_james_d...|0.7109430667614804|page_
collection_l...|0.08726735225206302|state_church_coll...|  0.0481260760704955|
+--------------------+--------------------+--------------------+------------------+-----
---------------+--------------------+--------------------+--------------------+
only showing top 3 rows
53526
171
```

```
%pyspark
```

FINISHED

```
# TODO: Convert hosts to PLDs using convert_hostname function from Paul 5.

# TODO: If multiple rows per PLD, only use the one with the shortest URL.

# TODO: Consider putting the three topic columns into a single column.

# Join on host/PLD
enrich_summary_df=example_summary_df.join(topic_df, example_summary_df.payLevelDomain==
enrich_summary_df.dropDuplicates().sort("numHosts",ascending=False).show()
enrich_summary_df.count()
```

```
+--------------------+--------+-------------+----------+-----------+--------------+-----
-+------+--------------------+--------------------+--------------------+----------------
----+--------------------+--------------------+
|      payLevelDomain|numHosts|pldIsHostFlag|pldLinksIn|pldLinksOut|wasCrawledFlag|hcRan
k|prRank|        exampleHosts|   exampleInLinkPlds|  exampleOutLinkPlds|              to
pic1|              topic2|              topic3|
+--------------------+--------+-------------+----------+-----------+--------------+-----
-+------+--------------------+--------------------+--------------------+----------------
----+--------------------+--------------------+
|      uchastings.edu|      99|         true|      1986|       2131|          true|99.94
3|99.946|uchastings.edu, a...|anu.edu.au, backe...|coronertalk.com, ...|nbsp_windows_quo
t...|student_students_...|product_order_shi...|
|      uchastings.edu|      99|         true|      1986|       2131|          true|99.94
3|99.946|uchastings.edu, a...|anu.edu.au, backe...|coronertalk.com, ...|nbsp_windows_quo
t...|product_order_shi...|student_students_...|
|      uchastings.edu|      99|         true|      1986|       2131|          true|99.94
3|99.946|uchastings.edu, a...|anu.edu.au, backe...|coronertalk.com, ...|product_order_sh
i...|nbsp_windows_quot...|student_students_...|
```

```
%pyspark
```

READY