201709 evaluate C...

```
%pyspark

# Simple script to demonstrate evaluation of CommonCrawl-derived domain vectors by usin
# classify domains according to high-level topic in the DMOZ dataset. Currently configu
# Bill's domain hex feature vectors from the 'Bill 6' notebook.
# TODO: Should we really be trying to predict domain links instead?
# PJ - 14 Sept 2017

import csv
import boto
from pyspark.sql.types import *

# Import the DMOZ domain category dataset
# (downloaded from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910

dmoz_labels=sc.textFile('s3://billsdata.net/CommonCrawl/DMOZ/dmoz_domain_category.csv')
header = dmoz_labels.first() # extract header
dmoz_labels = dmoz_labels.filter(lambda row: row != header) # remove header row
dmoz_labels.take(1)
```

FINISHED

```
[u'"sdcastroverde.com","Top/World/Galego/regional/Galicia/Lugo/municipalities/Castroverd
e"']
```

```
%pyspark

# For now, collect all labels into a list on one node
# TODO: Could probably do this much faster using map-reduce!
dmoz_labels_list = dmoz_labels.collect()

# Take a look at one record
dmoz_labels_list[1]
```

FINISHED

```
u'"www.232analyzer.com","Top/Computers/Hardware/Test_Equipment/Analyzers"'
```

```
%pyspark

# Make a dictionary of short domains (without www.) to top-level category label, as per
# http://dmoztools.net
labels={}
prefix="www."

# TODO: Could probably do this much faster using map-reduce!
for row in dmoz_labels_list[1:700000]: # Sample initially for speed (increasing to 1M c
    row = row.replace('"','').split(',')
```

FINISHED

```
    fulldomain = row[0]
    shortdomain = fulldomain[len(prefix):] if fulldomain.startswith(prefix) else fulldoı
    label = row[1].split("/")[1].split("|")[0]
    labels[shortdomain]=label
    #print(shortdomain + " " + label)

# Take a look at the category for one domain from our dictionary
```

```
u'Computers'
```

```
%pyspark
```

FINISHED

```
# Summarize categories in the DMOZ data
from collections import Counter
Counter(labels.values())
```

```
Counter({u'World': 359715, u'Regional': 180492, u'Business': 42311, u'Society': 21918, u
'Arts': 18934, u'Shopping': 14576, u'Recreation': 13095, u'Computers': 12915, u'Sports':
9356, u'Science': 7543, u'Health': 6927, u'Reference': 6224, u'Games': 2892, u'Home': 20
07, u'News': 1089})
```

```
%pyspark
```

FINISHED

```
# Load Bill's domain feature vectors from s3, in the following format:
# (u'www.angelinajolin.com', [4.30406509320417, 0.02702702702702703, 0.0, 0.13513513513

nfiles=128

# Load feature vectors from WAT files (from 'Bill 6' notebook):
inputURI = "s3://billsdata.net/CommonCrawl/domain_hex_feature_vectors_from_%d_WAT_files
features_rdd = sc.textFile(inputURI).map(eval)
features_rdd.cache()
print("Nr domains:", features_rdd.count())
print(features_rdd.take(1))
```

```
('Nr domains:', 2626203)
[(u'www.iggl.de', [3.6375861597263857, 0.5, 0.0, 0.0, 0.02564102564102564, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.358974358974359, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.05128205128205128, 0.0, 0.0, 0.
0, 0.05128205128205128, 0.0, 0.02564102564102564, 0.02564102564102564, 0.153846153846153
85, 0.20512820512820512, 0.0, 0.02564102564102564, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.
0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.025641
02564102564, 0.0, 0.05128205128205128, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
```

```
%pyspark                                                                        FINISHED

# Convert to a python dictionary for ease-of-use initially
# TODO: Find a better Spark way to do this!
features_sample = features_rdd.sample(0, 0.07, seed=42) # TODO: Investigate memory erro
features_dict = features_sample.collectAsMap()
#features_dict['232analyzer.com']
print(len(features_dict.keys()))
features_dict.itervalues().next() # Output one vector for testing
```

```
183530
[4.770684624465665, 0.5338983050847458, 0.0, 0.0, 0.008403361344537815, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.33613445378151263, 0.0, 0.0, 0.
0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.01680672268907563, 0.0, 0.0336134
4537815126, 0.025210084033613446, 0.058823529411764705, 0.01680672268907563, 0.008403361
344537815, 0.03361344537815126, 0.21008403361344538, 0.11764705882352941, 0.0, 0.0252100
84033613446, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.03361344537815126, 0.01680672268907563, 0.02
5210084033613446, 0.03361344537815126, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0
, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0
.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0,
```

```
%pyspark                                                                        FINISHED

# Filter embeddings for only those vectors that have entries in the DMOZ dictionary (i.
```

```
new_vec_ids=[]
new_vec_embs=[]
ground_truth=[]

def intersect(a, b):
     return list(set(a) & set(b))

common_domains=intersect(features_dict.keys(), labels.keys())
print(len(common_domains))
print(common_domains[1])

# Iterate over all the domain IDs for which we also have a vector embedding
for domain in common_domains:

    new_vec_ids.append(domain)
    new_vec_embs.append(features_dict[domain])
    ground_truth.append(labels[domain])

# Verify lengths of each list
```

```
1092
dailynews.com
1092 1092 1092
```

```
%pyspark
```

FINISHED

```
# Split into training and test sets
from sklearn.cross_validation import train_test_split
X_train, X_test, y_train, y_test = train_test_split(new_vec_embs, ground_truth, test_si
```

```
%pyspark
```

FINISHED

```
# Summarize labels in our test data
Counter(y_test)
```

```
Counter({u'World': 197, u'Regional': 99, u'Computers': 42, u'Arts': 33, u'Society': 32,
u'Reference': 25, u'Business': 24, u'Science': 24, u'Recreation': 15, u'Health': 13, u'G
ames': 12, u'Shopping': 11, u'Home': 9, u'Sports': 7, u'News': 3})
```

```
%pyspark
```

FINISHED

```
# Fit classifiers to the training data
from sklearn.neighbors import KNeighborsClassifier
from sklearn.ensemble import RandomForestClassifier

neigh = KNeighborsClassifier(n_neighbors=3, metric='cosine', algorithm='brute')
neigh.fit(X_train, y_train)

rf = RandomForestClassifier(max_depth=2, random_state=0)
rf.fit(X_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
            max_depth=2, max_features='auto', max_leaf_nodes=None,
            min_impurity_decrease=0.0, min_impurity_split=None,
            min_samples_leaf=1, min_samples_split=2,
            min_weight_fraction_leaf=0.0, n_estimators=10, n_jobs=1,
            oob_score=False, random_state=0, verbose=0, warm_start=False)
```

```
%pyspark

# Attempt to classify all test points using nearest neighbours
from sklearn.metrics import classification_report
print(classification_report(y_test, neigh.predict(X_test)))
print(classification_report(y_test, rf.predict(X_test)))
```

FINISHED

```
             precision    recall  f1-score   support
       Arts       0.03      0.06      0.04        33
   Business       0.03      0.08      0.05        24
  Computers       0.06      0.07      0.07        42
      Games       0.00      0.00      0.00        12
     Health       0.00      0.00      0.00        13
       Home       0.04      0.11      0.06         9
       News       0.00      0.00      0.00         3
 Recreation       0.00      0.00      0.00        15
  Reference       0.08      0.12      0.09        25
   Regional       0.24      0.21      0.23        99
    Science       0.00      0.00      0.00        24
   Shopping       0.00      0.00      0.00        11
    Society       0.30      0.09      0.14        32
     Sports       0.00      0.00      0.00         7
      World       0.48      0.37      0.41       197
avg / total       0.25      0.20      0.21       546
             precision    recall  f1-score   support
```

```
%pyspark
```

READY