# Bill 5 - understandi...

FINISHED

```
%pyspark

import boto
from boto.s3.key import Key
from gzipstream import GzipStreamFile
from pyspark.sql.types import *
import warc
import ujson as json
import urlparse

watlist = sc.textFile("s3://commoncrawl/crawl-data/CC-MAIN-2017-04/wat.paths.gz")
watlist.cache()

def unpack(uri):
    conn = boto.connect_s3(anon=True, host='s3.amazonaws.com')
    bucket = conn.get_bucket('commoncrawl')
    key_ = Key(bucket, uri)
    file_ = warc.WARCFile(fileobj=GzipStreamFile(key_))
    return file_

def extract_json(id_, iterator):
    for uri in iterator:
        file = unpack(uri)
        for record in file:
            if record['Content-Type'] == 'application/json':
                try:
                    content = json.loads(record.payload.read())
                    yield content['Envelope']
                except:
                    yield None

def parse_urls(record):
    url_list = []
    try:
        page_url = record['WARC-Header-Metadata']['WARC-Target-URI']
        x = urlparse.urlparse(page_url)
        url_list += [(x.netloc, x.path)]
    except:
        pass
    try:
        links = record['Payload-Metadata']['HTTP-Response-Metadata']['HTML-Metadata']['Lin
        for url in links:
            x = urlparse.urlparse(url['url'])
            url_list += [(x.netloc, x.path)]
    except:
        pass

    return url_list
```

Took 31 sec. Last updated by anonymous at September 09 2017, 5:17:16 PM.

READY

# Parse URLs from JSON: Records RDD

```
%pyspark
```
FINISHED

```
from __future__ import print_function

nfiles = 1
files = sc.parallelize(watlist.take(nfiles))

json_rdd = files.mapPartitionsWithIndex(extract_json)
json_rdd.cache()

print("Nr json records:", json_rdd.count())

records = json_rdd\
        .flatMap(parse_urls)\
        .filter(lambda x: x[0] is not "")\
        .groupByKey()\
        .map(lambda x: (x[0], set(x[1])))

records.cache()
json_rdd.unpersist()

record_count = records.map(lambda x: (x[0], len(x[1]))).sortBy(lambda x: -x[1]).collect()
```

```
Nr json records: 162874
(u'www.facebook.com', 10872)
(u'twitter.com', 10241)
(u'www.newslocker.com', 5784)
(u'artodyssey1.blogspot.com', 5366)
(u'www.youtube.com', 5305)
(u'plus.google.com', 4337)
(u'www.socarrao.com.br', 3551)
(u'4chanarchives.cu.cc', 3249)
(u'www.price4all.ru', 3079)
(u'akulagi.com', 3034)
```

Took 3 min 7 sec. Last updated by anonymous at September 09 2017, 5:20:29 PM.

```
%pyspark
```
READY

```
from __future__ import print_function

ex = records.filter(lambda x: len(x[1])==10).takeSample(False,1)[0]
print("Domain:", ex[0])
print("Pages:")
for y in ex[1]: print(y)
```

```
Domain: wikileaks.org
Pages:
/-Leaks-.html
/the-gifiles.html
/-About-66-.html
/-Partners-.html
/-News-.html
/static/gfx/WL_Hour_Glass_small.png
/static/gfx/gifiles.jpg
/talk
/gifiles/docs/11/1197857_budget-sri-lanka-tigers-getting-their-ass-handed-to-them-.html
```

READY

We next define a string encoding of domains.
The idea will be to choose this so that domain structure (as contained in its URIs) can be learnt be an RNN.

```
%pyspark
```
FINISHED

```
import re
from __future__ import print_function

def hexify(c):
    try:
        s = c.encode("utf-8").encode("hex")
    except UnicodeDecodeError:
        s = 0
    n = len(s)
    if n <= 2: return s
    a = ' '.join([s[i:i+2]+' -' for i in range(0,n,2)])
    return a[:-1]

def hexalise(str):
    return ' '.join([hexify(c) for c in str]) + ' . '

def domain_string(domain, path_set):
    out = hexalise(domain)
    for p in path_set: out += hexalise(p)
    return out
```

Took 0 sec. Last updated by anonymous at September 09 2017, 5:20:39 PM.

As the examples below show, we've chosen this encoding with the following constraints in mind: READY

- All symbols should be separated by spaces in order to parse at RNN training time.
- As well as hex symbols we include '.' to delimit different URIs.
- We include '-' as a limiter within non-Latin unicode characters. This will allow the RNN to distinguish Chinese characters, say, from sequences of Latin characters.
- Distinct domains will be delimited by '\n' at RNN training time.

```
%pyspark
```
FINISHED

```
from __future__ import print_function

ex = records.filter(lambda x: len(x[1]) > 10 and len(x[1]) < 100).takeSample(False, 100)

for dom in ex:
    print("-----------------------------------------")
    print("Domain:", dom[0])
    print("URIs:")
    print('\n'.join(list(dom[1])))
```

```
/wirtschaft/boersenkurse/
/469260
/sport/fussball/WM-Qualifikation/
/sport/fussball/uefa/

/sport/formel1/Statistik/Rennkalender-article14692191.html
/ratgeber/
/mediathek/videos/sport/Ice-Skater-jagen-mit-bis-zu-80-km-h-durch-den-Eiskanal-article19558
531.html
```

```
/thema/
/495196
/14595871
/panorama/
/14737826
/sport/fussball/1bundesliga/
/sport/
/sport/fussball/redelings_nachspielzeit/
/mediathek/bilderserien/sport/Ronaldo-uebertrumpft-Messi-und-Ribery-article12058206.html
/wirtschaft/
```

Took 2 sec. Last updated by anonymous at September 09 2017, 5:30:28 PM.

```
%pyspark

from __future__ import print_function

ex = records.filter(lambda x: len(x[1])==10).take(2)

for dom in ex:
    print("-----------------------------------------")
    print("Domain:", dom[0])
    print("Page string:")
    print(domain_string(dom[0], dom[1]))
```

READY

```
 67 73 72 65 63 68 74 2e 70 68 70 . 2f 4d 6f 64 75 6c 65 2f 56 65 72 6b 65 68 72 73 73 74
6 69 6c 72 65 63 68 74 2e 70 68 70 . 2f 54 65 78 74 65 2f 52 73 70 72 32 31 38 37 2e 70 68
 70 . 2f 4d 6f 64 75 6c 65 2f 56 65 72 6b 65 68 72 73 73 74 72 61 66 73 61 63 68 65 6e 2e 7
0 68 70 . 2f 4c 65 78 69 6b 6f 6e 2e 70 68 70 . 2f 49 6d 70 72 65 73 73 75 6d 2e 70 68 70 .

-----------------------------------------
Domain: www.charityblossom.org
Page string:
77 77 77 2e 63 68 61 72 69 74 79 62 6c 6f 73 73 6f 6d 2e 6f 72 67 .  . 2f 64 69 72 65 63 74
 6f 72 79 2f 46 4c 2f 4f 72 6c 61 6e 64 6f 2f 33 32 38 31 31 2f . 2f 64 69 72 65 63 74 6f 7
2 79 2f 4b 53 2f 54 6f 77 61 6e 64 61 2f 63 61 74 65 67 6f 72 79 2f 70 75 62 6c 69 63 2d 73
 61 66 65 74 79 2d 64 69 73 61 73 74 65 72 2d 70 72 65 70 61 72 65 64 6e 65 73 73 2d 72 65
 6c 69 65 66 2d 6d 2f 6d 61 6e 61 67 65 6d 65 6e 74 2d 74 65 63 68 6e 69 63 61 6c 2d 61 73
 73 69 73 74 61 6e 63 65 2d 6d 30 32 2f . 2f 6e 6f 6e 70 72 6f 66 69 74 2f 61 6d 65 72 69 6
3 61 6e 2d 6c 65 67 69 6f 6e 2d 64 75 6e 6b 69 72 6b 2d 6e 79 2d 31 34 30 34 38 2d 65 64 6d
 75 6e 64 2d 66 2d 67 6f 75 6c 64 2d 6a 72 2d 31 36 30 37 32 30 31 36 33 2f . 2f 6e 6f 6e 7
0 72 6f 66 69 74 2f 74 65 63 68 6e 6f 6c 6f 67 79 2d 72 65 76 69 65 77 2d 69 6e 63 2d 63 61
 6d 62 72 69 64 67 65 2d 6d 61 2d 30 32 31 34 32 2d 6a 61 6d 65 73 2d 63 6f 79 6c 65 2d 39
```

The following count shows the motivation for encoding domains in this way.

READY

We would like (for later use, when we model the string using an RNN) the alphabet of symbols in the representation to be reliably bounded. If we use the raw (unicode) string concatenation of the path URIs, then this is not the case because we get an explosion of possibilities from various languages. Here's a histogram of the symbols, together with their hex encodings:

```
%pyspark

from collections import Counter

char_count = records.map(lambda x: Counter('.'.join(list(x[1]))))\
                    .aggregate(Counter(),
                            lambda acc, value: acc + value,
```

READY

```
                              lambda acc1, acc2: acc1 + acc2)
char_count = dict(char_count)

# examine:
print("Nr characters:", len(char_count.keys()))
for key, value in sorted(char_count.iteritems(), key=lambda (k,v): (-v,k)):
    print "%8d %4s %16s" % (value, key, hexify(key))
```

```
('Nr characters:', 2083)
 5123801    /               2f
 4146432    e               65
 3690910    a               61
 2983947    -               2d
 2879741    t               74
 2783207    i               69
 2766669    s               73
 2707176    o               6f
 2475434    .               2e
 2433279    r               72
 2270142    n               6e
 2081952    l               6c
 1763606    c               63
 1636562    d               64
 1569923    m               6d
 1536649    p               70
 1478606    0               30
```

Compare this with the distribution after hexification. The number of symbols is bounded by 256 + 2 = 258 this time it's more informative to sort by key:

READY

```
%pyspark

from collections import Counter

hex_count = records.map(lambda x: Counter(domain_string(x[0], x[1]).split()))\
                   .aggregate(Counter(),
                              lambda acc, value: acc + value,
                              lambda acc1, acc2: acc1 + acc2)
hex_count = dict(hex_count)

# examine:
print("Nr hex characters:", len(hex_count.keys()))
for key, value in sorted(hex_count.iteritems(), key=lambda (k,v): k):
    print "%2s %8d" % (key, value)
```

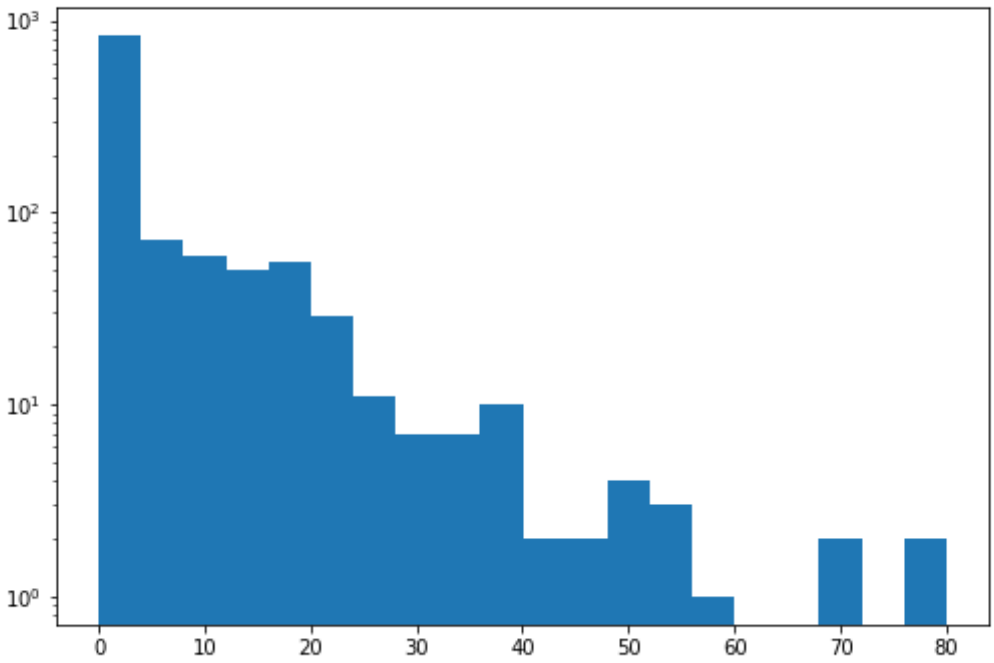
```
('Nr hex characters:', 199)
 -   252648
 .  1950605
03        1
09      413
0a      573
0b        1
0d      414
20    25473
21     1845
22       23
24     1291
25  1122548
26     3063
```

```
27      750
28     3561
29     3541
```

Let's use a filter on '-' to find all domains with non-Latin URIs: READY

```
%pyspark

import matplotlib.pyplot as plt

def detector(dom):
    """
    Appends the average number of '-'s per URI.
    """
    n = len(dom[1])
    m =  domain_string(dom[0], dom[1]).count('-')
    return (dom[0], dom[1], float(m)/n)

nonlatin = records.map(detector).filter(lambda dom: dom[2] > 0).collect()

plt.hist([dom[2] for dom in nonlatin], bins=20)
plt.yscale("log")
plt.show()
```

FINISHED




Took 52 sec. Last updated by anonymous at September 09 2017, 5:26:08 PM. (outdated)

For example: READY

```
%pyspark

from __future__ import print_function
```

FINISHED

```
for dom in nonlatin:
    if dom[2] > 20:
        print("----------------------------------------")
        print("%s (%g)" % (dom[0], dom[2]))
        for uri in dom[1]:
            print(uri)
```

لوازم-خانگی-و-مبلمان/مبلمان-منزل-و-سرویس-خواب/پوف/
فرهنگ-،-هنر-و-سرگرمی/فیلم-و-سریال/فیلم-و-مجموعه-کودک-و-نوجوان/
لوازم-خانگی-و-مبلمان/مبلمان-منزل-و-سرویس-خواب/پوف/
پوشاک-،-کیف-و-کفش/پوشاک-و-کفش-نوزاد--دخترانه/لباس-راحتی-و-خواب/
/tags/خرید-اینترنتی-Bart-Simpson-Cover-For-Samsung-Galaxy-Note-4--Type-2
پوشاک-،-کیف-و-کفش/پوشاک-و-کفش-نوجوان-دخترانه/کاپشن-و-پالتو/
/tags/-كاسه سراميكي-
لوازم-کودک-و-اسباب-بازی/بهداشت-و-حمام/شامپو-کودک-و-نوزاد/
/Samsung-Galaxy-Note-4-N910H-32GB-Limited-Edition-Pack/گوشي-موبايل-سامسونگ-گلکسي-نوت-پک-کام
4-ل--N910H
لوازم-خانگی-و-مبلمان/مبلمان-اداری/میز-کارگروهی-و-سایت/
/tags/مناسب-براي-سامسونگ-گلکسي-نوت-4--طرح-Bart-Simpson-2-کاور-گوشي-موبايل-مدل
لوازم-شخصی/ساعت/
پوشاک-،-کیف-و-کفش/پوشاک-زنانه/دامن/
دیجیتال/قطعات-کامپیوتر/کارت-گرافیک/
/tags/لیست-قیمت-Samsung-Galaxy-Note-4-Baseus-Primary-Case
دیجیتال/ماشین-های-اداری/بارکد-خوان/

ورزش-و-سلامت/ورزش-های-آبی/

Output exceeds 102400. Truncated.

Took 1 sec. Last updated by anonymous at September 09 2017, 5:27:21 PM.

```
%pyspark

records.unpersist()
```

READY

PythonRDD[52] at RDD at PythonRDD.scala:48

READY

# Save to S3

The end-to-end process:

READY

```
%pyspark

nfiles = 128

files = sc.parallelize(watlist.take(nfiles))
json_rdd = files.mapPartitionsWithIndex(extract_json)
domains_rdd = json_rdd\
            .flatMap(parse_urls)\
            .filter(lambda x: x[0] is not "")\
            .groupByKey()\
            .map(lambda x: {'domain': x[0], 'path_set': set(x[1])})


# make sure the following S3 directory is deleted first:
```

READY

```
outputURI = "s3://billsdata.net/CommonCrawl/domain_paths_from_%d_WAT_files" % nfiles
codec = "org.apache.hadoop.io.compress.GzipCodec"
domains_rdd.saveAsTextFile(outputURI, codec)
```

Timings: READY

| Cluster | nr WAT files | time | output size (gzip) |
|---|---|---|---|
| 16 x m4.2xlarge | 128 | 7 min 24 sec | 944.6 MiB |
| 16 x m4.2xlarge | 256 | 10 min 16 sec | 1.7 GiB |
| 16 x m4.2xlarge | 512 | 19 min 31 sec | 3.1 GiB |
| 16 x m4.2xlarge | 1024 | 40 min 43 sec | 5.7 GiB |

To find output size:

```
$ aws s3 ls —human-readable —summarize
s3://billsdata.net/CommonCrawl/domain_paths_from_256_WAT_files/ | grep Total
```

```
%pyspark
```

READY