201708 common c...

FINISHED

Playing with Common Crawl files - WARC

Browsing the S3 bucket:

```
aws s3 ls s3://commoncrawl/crawl-data/
```

Took 0 sec. Last updated by anonymous at August 28 2017, 1:59:26 PM.

READY

```
%pyspark

import boto
from boto.s3.key import Key
from gzipstream import GzipStreamFile
from pyspark.sql.types import *
import warc

import json

warclist = sc.textFile("s3://commoncrawl/crawl-data/CC-MAIN-2017-04/warc.paths.gz")
warclist.cache()
warclist.count()
```

```
57800
```

READY

```
%pyspark

filename1 = warclist.take(1)[0]
print(filename1)

conn = boto.connect_s3(anon=True, host='s3.amazonaws.com')
bucket = conn.get_bucket('commoncrawl')

def unpack(filename):
    key_ = Key(bucket, filename)
    file_ = warc.WARCFile(fileobj=GzipStreamFile(key_))
    return file_
```

```
crawl-data/CC-MAIN-2017-04/segments/1484560279169.4/warc/CC-MAIN-20170116095119-00000-ip-10
-171-10-70.ec2.internal.warc.gz
```

READY

```
%pyspark

from __future__ import print_function
from collections import Counter

file = unpack(filename1)
ct = [record['Content-Type'] for record in file]

for x in zip(Counter(ct).keys(), Counter(ct).values()): print(x)
```

```
('application/warc-fields', 54292)
('application/http; msgtype=response', 54291)
('application/http; msgtype=request', 54291)
```

Counting the records on one WARC file has taken just under a minute. For the complete set of 57800 files, one node will take ~1000 hours.
So to proceed further on < 1000 nodes we should work with just a small sample of files.

READY

```
%pyspark

def mapper(filename):
    file = unpack(filename)
    return [record['Content-Type'] for record in file]

smalllist = sc.parallelize(warclist.take(12))

ct = smalllist.flatMap(mapper)
ct.cache()
ct.count()
```

```
1967037
```

READY

```
%pyspark

ct_list = ct.collect()
for x in zip(Counter(ct_list).keys(), Counter(ct_list).values()): print(x)
```

```
('application/warc-fields', 655687)
('application/http; msgtype=response', 655675)
('application/http; msgtype=request', 655675)
```

READY

```
%pyspark
```