DATASCI W261, Machine Learning at Scale

Assignement: week #5

Lei Yang (mailto:leiyang@berkeley.edu) | Michael Kennedy (mailto:mkennedy@ischool.berkeley.edu) | Natarajan Krishnaswami (mailto:natarajan@krishnaswami.org)

Due: 2016-02-19, 8AM PST

INSTRUCTIONS for SUBMISSIONS

SPECIAL INSTURCTIONS: This weeks homework is a group exercise. Your team assignments for completing this HW are located at:

https://docs.google.com/spreadsheets/d/1ncFQl5Tovn-16slD8mYjP_nzMTPSfiGeLLzW8v_sMjg/edit?usp=sharing (https://docs.google.com/spreadsheets/d/1ncFQl5Tovn-16slD8mYjP_nzMTPSfiGeLLzW8v_sMjg/edit?usp=sharing)

See column Team assignment for Homeworks in tab "Teams for HW Assignments"

Please submit your homeworks (one per team) going forward via this form (and not thru the ISVC):

https://docs.google.com/forms/d/1ZOr9Rnle_A06AcZDB6K1mJN4vrLeSmS2PD6Xm3eOiis/viewform?usp=send_form (https://docs.google.com/forms/d/1ZOr9Rnle_A06AcZDB6K1mJN4vrLeSmS2PD6Xm3eOiis/viewform?usp=send_form)

Please follow the instructions for submissions carefully.

Week 5 ASSIGNMENTS

HW 5.0

1. What is a data warehouse?
2. What is a Star schema?
3. When is it used?

Answers:

1. A data warehouse is an enterprise's repository of all relevant information, be it structured, semi-structured, or unstructured, needed to monitor and predict business performance and needs.
2. A star schema is one where rows in "fact" tables connect together IDs from (flat) "dimension" tables, potentially along with some descriptive columns. It is a star since the fact table's columns fan out to each dimension table.
3. If the fact entries do not vary too much in the dimensions they need (e.g., varying by location in a hierarchy), a star schema can be a good fit, and permit straightforward, easy to optimize, queries, and natural visualization of the entity relationships. Snowflake schemas can be a good fit if more flexibility/variability in needed for the dimensions, at the cost of more complex processing.

HW 5.1

1. In the database world What is 3NF?
2. Does machine learning use data in 3NF?
   A. If so why?
3. In what form does ML consume data?
4. Why would one use log files that are denormalized?

Answers:

1. In Codd's hierarchy of normal forms, third normal form is a reduction of non-key column redundancy with and across rows, but not within superkeys. This avoidance of functional dependencies is (almost) sufficient to guarantee referential integrity during modification
2. ML algorithms do not generally use normalized data
3. ML algorithms generally use highly denormalized data transformed into a suitable feature space.
4. If log files are not normalized, the various pieces of a record would need to be located and processed in order (joined) to process each record

HW 5.2

Using MRJob, implement a hashside join (memory-backed map-side) for left, right and inner joins. Run your code on the data used in HW 4.4: (Recall HW 4.4: Find the most frequent visitor of each page using mrjob and the output of 4.2 (i.e., transfromed log file).
In this output please include the webpage URL, webpageID and Visitor ID.)

1. Justify which table you chose as the Left table in this hashside join.
2. Please report the number of rows resulting from:
   A. Left joining Table Left with Table Right
   B. Right joining Table Left with Table Right
   C. Inner joining Table Left with Table Right

In [24]:
```
%%writefile hw52.py
#!/opt/anaconda/bin/python
from mrjob.job import MRJob
from mrjob.step import MRStep

class HW52Job(MRJob):
    def configure_options(self):
        super(HW52Job,self).configure_options()
        self.add_passthrough_option("--join_type",type='string',help='[left|right|inner]', default='le
ft')

    urlmap={}
    seen=set()
    def load_urlmap(self):
        with open('hw4.2-urls.txt', 'r') as urls:
            for row in urls:
                fields=row.strip().split(',')
                self.urlmap[int(fields[1])]=fields[3].strip('"')
    @staticmethod
    def split_visit(line):
        fields=line.strip().split(',')
        if len(fields) > 4:
            return int(fields[1]), int(fields[4])

    def left_join(self, _, line):
        urlid, userid=self.split_visit(line)
        if urlid in self.urlmap:
            self.seen.add(urlid)
            yield urlid, (userid, self.urlmap[urlid])
    def left_join_final(self):
        for urlid in set(self.urlmap)-self.seen:
            yield urlid, (self.urlmap[urlid], None)

    def right_join(self, _, line):
        urlid, userid=self.split_visit(line)
        yield urlid, (userid, self.urlmap.get(urlid, None))

    def inner_join(self, _, line):
        urlid, userid=self.split_visit(line)
        url=self.urlmap.get(urlid, None)
        if url:
            yield urlid, (userid, url)


    def id_reducer(self, key, values):
        for value in values:
            yield key, value

    def steps(self):
        if self.options.join_type.lower()=='left':
            mapper=self.left_join
            mapper_final=self.left_join_final
        elif self.options.join_type.lower()=='right':
            mapper=self.right_join
            mapper_final=None
        elif self.options.join_type.lower()=='inner':
            mapper=self.inner_join
            mapper_final=None
        else:
            raise ValueError('Unknown join type '+ self.options.join_type)
        return [MRStep(
                mapper_init=self.load_urlmap,
```

```
                mapper=mapper,
                mapper_final=mapper_final,
                reducer=self.id_reducer
            )]

if __name__=="__main__":
    HW52Job().run()
```

```
Overwriting hw52.py
```

In [27]:
```
%%bash
export HADOOP_HOME=/opt/hadoop-2.7.1
export PATH=$HADOOP_HOME/bin:$PATH
export HADOOP_ROOT_LOGGER=INFO,console

for jointype in left right inner; do
    rm -rf hw5.2-${jointype}
    ./hw52.py --join_type=${jointype} \
      -r inline --no-output \
      --file ../hw4/hw4.2-urls.txt \
      --output=hw5.2-${jointype}-output \
      ../hw4/hw4.2-visits.txt
done
```

```
using configs in /home/nkrishna/.mrjob.conf
creating tmp directory /tmp/hw52.nkrishna.20160215.003154.042726
writing to /tmp/hw52.nkrishna.20160215.003154.042726/step-0-mapper_part-00000
Counters from step 1:
  (no counters found)
writing to /tmp/hw52.nkrishna.20160215.003154.042726/step-0-mapper-sorted
> sort /tmp/hw52.nkrishna.20160215.003154.042726/step-0-mapper_part-00000
writing to /tmp/hw52.nkrishna.20160215.003154.042726/step-0-reducer_part-00000
Counters from step 1:
  (no counters found)
Moving /tmp/hw52.nkrishna.20160215.003154.042726/step-0-reducer_part-00000 -> hw5.2-left-output/part-0
0000
removing tmp directory /tmp/hw52.nkrishna.20160215.003154.042726
using configs in /home/nkrishna/.mrjob.conf
creating tmp directory /tmp/hw52.nkrishna.20160215.003200.262725
writing to /tmp/hw52.nkrishna.20160215.003200.262725/step-0-mapper_part-00000
Counters from step 1:
  (no counters found)
writing to /tmp/hw52.nkrishna.20160215.003200.262725/step-0-mapper-sorted
> sort /tmp/hw52.nkrishna.20160215.003200.262725/step-0-mapper_part-00000
writing to /tmp/hw52.nkrishna.20160215.003200.262725/step-0-reducer_part-00000
Counters from step 1:
  (no counters found)
Moving /tmp/hw52.nkrishna.20160215.003200.262725/step-0-reducer_part-00000 -> hw5.2-right-output/part-
00000
removing tmp directory /tmp/hw52.nkrishna.20160215.003200.262725
using configs in /home/nkrishna/.mrjob.conf
creating tmp directory /tmp/hw52.nkrishna.20160215.003206.714690
writing to /tmp/hw52.nkrishna.20160215.003206.714690/step-0-mapper_part-00000
Counters from step 1:
  (no counters found)
writing to /tmp/hw52.nkrishna.20160215.003206.714690/step-0-mapper-sorted
> sort /tmp/hw52.nkrishna.20160215.003206.714690/step-0-mapper_part-00000
writing to /tmp/hw52.nkrishna.20160215.003206.714690/step-0-reducer_part-00000
Counters from step 1:
  (no counters found)
Moving /tmp/hw52.nkrishna.20160215.003206.714690/step-0-reducer_part-00000 -> hw5.2-inner-output/part-
00000
removing tmp directory /tmp/hw52.nkrishna.20160215.003206.714690
```

Answers:

1. I chose the URLs list for the left table since it is much, much smaller than the visits list. Thus it made sense for it to be the one to load into ram.

In [36]:
```
%%bash
echo "2. A:" $(< hw5.2-left-output/part-00000 wc -l)
echo "   B:" $(< hw5.2-right-output/part-00000 wc -l)
echo "   C:" $(< hw5.2-inner-output/part-00000 wc -l)
```

```
2. A: 98663
   B: 98654
   C: 98654
```

HW 5.3 For the remainder of this assignment you will work with two datasets:

1: unit/systems test data set: SYSTEMS TEST DATASET

Three terms, A,B,C and their corresponding strip-docs of co-occurring terms

```
DocA {X:20, Y:30, Z:5}
DocB {X:100, Y:20}
DocC {M:5, N:20, Z:5}
```

2: A large subset of the Google n-grams dataset

https://aws.amazon.com/datasets/google-books-ngrams/ (https://aws.amazon.com/datasets/google-books-ngrams/)

which we have placed in a bucket/folder on Dropbox on s3:

https://www.dropbox.com/sh/tmqpc4o0xswhkvz/AACUifrl6wrMrlK6a3X3lZ9Ea?dl=0
(https://www.dropbox.com/sh/tmqpc4o0xswhkvz/AACUifrl6wrMrlK6a3X3lZ9Ea?dl=0)

s3://filtered-5grams/

For each HW 5.3 -5.5 Please unit test and system test your code with with SYSTEMS TEST DATASET and show the results.
Please compute the expected answer by hand and show your hand calculations. Then show the results you get with you system.
Final show your results on the Google n-grams dataset

In particular, this bucket contains (~200) files (10Meg each) in the format:

```
(ngram) \t (count) \t (pages_count) \t (books_count)
```

Do some EDA on this dataset using mrjob, e.g.,

1. Longest 5-gram (number of characters)
2. Top 10 most frequent words (please use the count information), i.e., unigrams
3. Most/Least densely appearing words (count/pages_count) sorted in decreasing order of relative frequency (Hint: save to PART-000* and take the `head -n 1000`)
4. Distribution of 5-gram sizes (using counts info.) sorted in decreasing order of relative frequency.
5. OPTIONAL Question: Plot the log-log plot of the frequency distributuion of unigrams. Does it follow power law distribution?

For more background see:
https://en.wikipedia.org/wiki/Log%E2%80%93log_plot (https://en.wikipedia.org/wiki/Log%E2%80%93log_plot)
https://en.wikipedia.org/wiki/Power_law (https://en.wikipedia.org/wiki/Power_law)

In [29]:
```
with open('hw5.3-test.txt','w') as f:
    print >>f, 'DocA\t{X:20, Y:30, Z:5}'
    print >>f, 'DocB\t{X:100, Y:20}'
    print >>f, 'DocC\t{M:5, N:20, Z:5}'
```

In [101]:
```
%%writefile hw5.3-1.py
#!/usr/bin/env python

from collections import namedtuple
from mrjob.job import MRJob
from mrjob.step import MRStep

class HW53Job(MRJob):
    JOBCONF = {
        'mapred.output.key.comparator.class':
          'org.apache.hadoop.mapred.lib.KeyFieldBasedComparator',
        'mapred.text.key.comparator.options': '-k1,1nr',
    }
    Row=namedtuple('Row',['ngram', 'count', 'pages_count', 'books_count'])
    @staticmethod
    def split_line(line):
        fields=line.strip().split('\t')
        return HW53Job.Row(fields[0],*[int(field) for field in fields[1:]])

    # Find the longest 5-gram
    ## Mapper: keep track of/update the longest 5-gram.
    #          finally yield the longest one seen.
    def longest_map_init(self):
        self.longest=''
    def longest_map(self, _, line):
        row=HW53Job.split_line(line)
```

```
        row=HW53Job.split_line(line);
        if len(row.ngram) > len(self.longest):
            self.longest=row.ngram
    def longest_map_final(self):
        yield len(self.longest), self.longest

    ## Reducer: keep track of/update the longest word list.
    #          finally yield each of the longest words with their length.
    def longest_red_init(self):
        self.longest=[]
    def longest_red(self, key, values):
        if not self.longest or key > len(self.longest[0]):
            for val in values:
                self.longest.append(val)
    def longest_red_final(self):
        for val in self.longest:
            yield len(val), val

    def steps(self):
        return [MRStep(
                mapper_init=self.longest_map_init,
                mapper=self.longest_map,
                mapper_final=self.longest_map_final,
                reducer_init=self.longest_red_init,
                reducer=self.longest_red,
                reducer_final=self.longest_red_final,
            )]
if __name__=='__main__':
    HW53Job().run()
    exit(0)
```

```
Overwriting hw5.3-1.py
```

In [135]:
```
%%bash
/usr/bin/ssh root@50.22.252.4 bash -xs <<'EOF'
cd hw5
prog=hw5.3-1
hdfs=hdfs://master:9000
input=$hdfs/filtered-5grams
output=$hdfs/$prog-output
HADOOP_ROOT_LOGGER=WARN,console
hdfs dfs -rm -r -f $output
time ./$prog.py -q -r hadoop --no-bootstrap-mrjob \
  --output $output \
  $input
EOF
```

```
Deleted hdfs://master:9000/hw5.3-1-output
159     "ROPLEZIMPREDASTRODONBRASLPKLSON YHROACLMPARCHEYXMMIOUDAVESAURUS PIOFPILOCOWERSURUASOGETSESNEG
CP TYRAVOPSIFENGOQUAPIALLOBOSKENUO OWINFUYAIOKENECKSASXHYILPOYNUAT"
159     "AIOPJUMRXUYVASLYHYPSIBEMAPODIKR UFRYDIUUOLBIGASUAURUSREXLISNAYE RNOONDQSRUNSUBUNOUGRABBERYAIR
TC UTAHRAPTOREDILEIPMILBDUMMYUVERI SYEVRAHVELOCYALLOSAURUSLINROTSR"
+ cd hw5
+ prog=hw5.3-1
+ hdfs=hdfs://master:9000
+ input=hdfs://master:9000/filtered-5grams
+ output=hdfs://master:9000/hw5.3-1-output
+ HADOOP_ROOT_LOGGER=WARN,console
+ hdfs dfs -rm -r -f hdfs://master:9000/hw5.3-1-output
16/02/16 23:24:39 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minu
tes, Emptier interval = 0 minutes.
+ ./hw5.3-1.py -q -r hadoop --no-bootstrap-mrjob --output hdfs://master:9000/hw5.3-1-output hdfs://mas
ter:9000/filtered-5grams

real    0m57.575s
user    0m25.311s
sys     0m1.572s
```

In [149]:
```
%%writefile hw5.3-2.py
#!/usr/bin/env python

from collections import namedtuple
from mrjob.job import MRJob
from mrjob.step import MRStep
import sys

class HW53Job(MRJob):
    Row=namedtuple('Row',['ngram', 'count', 'pages_count', 'books_count'])
    @staticmethod
    def split_line(line):
```

```
        fields=line.strip().split('\t')
        return HW53Job.Row(fields[0],*[int(field) for field in fields[1:]])

    """Find the top ten unigrams"""
    def get_wordcounts(self, _, line):
        """Mapper: split the 5-grams, and yield the count with each."""
        row=HW53Job.split_line(line)
        for word in row.ngram.split():
            print >>sys.stderr, "word:",word
            yield word, int(row.count)
    def sum_wordcounts(self, key, values):
        """Combiner: sum the counts as in usual word count"""
        yield key, sum(values)
    def sum_swap_wordcounts(self, key, values):
        """Reducer: sum the counts as in usual word count and swap key/val for sorting"""
        yield sum(values), key

    ## dummy map/red steps to cause another sort
    def map_id(self, key, val):
        yield key, val
    def red_id(self, key, vals):
        for x in vals:
            yield key, x

    def steps(self):
        return [
            MRStep(
                mapper=self.get_wordcounts,
                combiner=self.sum_wordcounts,
                reducer=self.sum_swap_wordcounts,
            ),
            MRStep(
                mapper=self.map_id,
                reducer=self.red_id,
                jobconf={
                    'mapred.output.key.comparator.class':
                       'org.apache.hadoop.mapred.lib.KeyFieldBasedComparator',
                    'mapred.text.key.comparator.options': '-k1,1nr',
                },
            ),
        ]

if __name__=='__main__':
    HW53Job().run()
    exit(0)
```

```
Overwriting hw5.3-2.py
```

In [156]:
```
%%bash
/usr/bin/ssh root@50.22.252.4 bash -xs <<'EOF'
cd hw5
prog=hw5.3-2
hdfs=hdfs://master:9000
input=${hdfs}/filtered-5grams
output=${hdfs}/${prog}-output
HADOOP_ROOT_LOGGER=INFO,console
hdfs dfs -rm -r -f ${output}
time ./${prog}.py -r hadoop --strict-protocols --no-bootstrap-mrjob \
  --no-output \
  --output ${output} \
   ${input}
hdfs dfs -cat ${output}/part-00000 | head -10
EOF
```

```
Deleted foo
5375699242      "the"
3691308874      "of"
2221164346      "to"
1387638591      "in"
1342195425      "a"
1135779433      "and"
798553959       "that"
756296656       "is"
688053106       "be"
481373389       "as"
+ cd hw5
+ cat out
+ cat err
16/02/17 08:24:14 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minu
```

```
tes, emptier interval = 0 minutes.
no configs found; falling back on auto-configuration
no configs found; falling back on auto-configuration
creating tmp directory /tmp/hw5.root.20160217.142415.557030
writing wrapper script to /tmp/hw5.root.20160217.142415.557030/setup-wrapper.sh
Using Hadoop version 2.7.2
Copying local files into hdfs:///user/root/tmp/mrjob/hw5.root.20160217.142415.557030/files/
HADOOP: packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/s
treamjob1443581881176190507.jar tmpDir=null
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Total input paths to process : 190
HADOOP: number of splits:190
HADOOP: Submitting tokens for job: job_1455677115242_0032
HADOOP: Submitted application application_1455677115242_0032
HADOOP: The url to track the job: http://master:8088/proxy/application_1455677115242_0032/
HADOOP: Running job: job_1455677115242_0032
HADOOP: Job job_1455677115242_0032 running in uber mode : false
HADOOP:  map 0% reduce 0%
HADOOP:  map 1% reduce 0%
HADOOP:  map 2% reduce 0%
HADOOP:  map 3% reduce 0%
HADOOP:  map 4% reduce 0%
HADOOP:  map 5% reduce 0%
HADOOP:  map 6% reduce 0%
HADOOP:  map 7% reduce 0%
HADOOP:  map 8% reduce 0%
HADOOP:  map 9% reduce 0%
HADOOP:  map 10% reduce 0%
HADOOP:  map 11% reduce 0%
HADOOP:  map 12% reduce 0%
HADOOP:  map 13% reduce 0%
HADOOP:  map 14% reduce 0%
HADOOP:  map 15% reduce 0%
HADOOP:  map 16% reduce 0%
HADOOP:  map 17% reduce 0%
HADOOP:  map 18% reduce 0%
HADOOP:  map 19% reduce 0%
HADOOP:  map 20% reduce 0%
HADOOP:  map 22% reduce 0%
HADOOP:  map 24% reduce 0%
HADOOP:  map 25% reduce 0%
HADOOP:  map 25% reduce 6%
HADOOP:  map 26% reduce 6%
HADOOP:  map 27% reduce 6%
HADOOP:  map 28% reduce 6%
HADOOP:  map 29% reduce 6%
HADOOP:  map 30% reduce 6%
HADOOP:  map 31% reduce 6%
HADOOP:  map 31% reduce 7%
HADOOP:  map 32% reduce 7%
HADOOP:  map 33% reduce 7%
HADOOP:  map 34% reduce 7%
HADOOP:  map 35% reduce 7%
HADOOP:  map 35% reduce 8%
HADOOP:  map 36% reduce 8%
HADOOP:  map 36% reduce 9%
HADOOP:  map 37% reduce 9%
HADOOP:  map 38% reduce 9%
HADOOP:  map 39% reduce 9%
HADOOP:  map 40% reduce 9%
HADOOP:  map 41% reduce 9%
HADOOP:  map 42% reduce 9%
HADOOP:  map 43% reduce 10%
HADOOP:  map 44% reduce 10%
HADOOP:  map 45% reduce 12%
HADOOP:  map 46% reduce 12%
HADOOP:  map 46% reduce 13%
HADOOP:  map 47% reduce 13%
HADOOP:  map 48% reduce 14%
HADOOP:  map 49% reduce 14%
HADOOP:  map 49% reduce 15%
HADOOP:  map 50% reduce 15%
HADOOP:  map 51% reduce 15%
HADOOP:  map 52% reduce 15%
HADOOP:  map 53% reduce 15%
HADOOP:  map 54% reduce 15%
HADOOP:  map 55% reduce 15%
HADOOP:  map 56% reduce 15%
HADOOP:  map 57% reduce 15%
```

```
HADOOP:  map 58% reduce 15%
HADOOP:  map 59% reduce 15%
HADOOP:  map 60% reduce 15%
HADOOP:  map 61% reduce 15%
HADOOP:  map 62% reduce 15%
HADOOP:  map 63% reduce 15%
HADOOP:  map 64% reduce 15%
HADOOP:  map 64% reduce 16%
HADOOP:  map 65% reduce 16%
HADOOP:  map 66% reduce 16%
HADOOP:  map 66% reduce 17%
HADOOP:  map 67% reduce 18%
HADOOP:  map 68% reduce 18%
HADOOP:  map 68% reduce 19%
HADOOP:  map 69% reduce 19%
HADOOP:  map 70% reduce 19%
HADOOP:  map 70% reduce 21%
HADOOP:  map 71% reduce 21%
HADOOP:  map 72% reduce 22%
HADOOP:  map 73% reduce 22%
HADOOP:  map 74% reduce 22%
HADOOP:  map 75% reduce 22%
HADOOP:  map 76% reduce 22%
HADOOP:  map 77% reduce 22%
HADOOP:  map 78% reduce 22%
HADOOP:  map 79% reduce 22%
HADOOP:  map 80% reduce 22%
HADOOP:  map 81% reduce 22%
HADOOP:  map 82% reduce 22%
HADOOP:  map 83% reduce 22%
HADOOP:  map 84% reduce 22%
HADOOP:  map 85% reduce 22%
HADOOP:  map 86% reduce 22%
HADOOP:  map 87% reduce 22%
HADOOP:  map 88% reduce 22%
HADOOP:  map 89% reduce 22%
HADOOP:  map 90% reduce 22%
HADOOP:  map 90% reduce 25%
HADOOP:  map 91% reduce 25%
HADOOP:  map 91% reduce 26%
HADOOP:  map 92% reduce 26%
HADOOP:  map 93% reduce 27%
HADOOP:  map 94% reduce 28%
HADOOP:  map 95% reduce 28%
HADOOP:  map 95% reduce 29%
HADOOP:  map 96% reduce 29%
HADOOP:  map 96% reduce 30%
HADOOP:  map 97% reduce 30%
HADOOP:  map 97% reduce 31%
HADOOP:  map 98% reduce 31%
HADOOP:  map 99% reduce 31%
HADOOP:  map 99% reduce 32%
HADOOP:  map 99% reduce 33%
HADOOP:  map 100% reduce 33%
HADOOP:  map 100% reduce 34%
HADOOP:  map 100% reduce 67%
HADOOP:  map 100% reduce 69%
HADOOP:  map 100% reduce 70%
HADOOP:  map 100% reduce 72%
HADOOP:  map 100% reduce 73%
HADOOP:  map 100% reduce 75%
HADOOP:  map 100% reduce 76%
HADOOP:  map 100% reduce 78%
HADOOP:  map 100% reduce 80%
HADOOP:  map 100% reduce 82%
HADOOP:  map 100% reduce 85%
HADOOP:  map 100% reduce 87%
HADOOP:  map 100% reduce 89%
HADOOP:  map 100% reduce 91%
HADOOP:  map 100% reduce 93%
HADOOP:  map 100% reduce 95%
HADOOP:  map 100% reduce 98%
HADOOP:  map 100% reduce 100%
HADOOP: Job job_1455677115242_0032 completed successfully
HADOOP: Counters: 52
HADOOP:         File System Counters
HADOOP:                 FILE: Number of bytes read=50683493
HADOOP:                 FILE: Number of bytes written=206122757
HADOOP:                 FILE: Number of read operations=0
```

```
HADOOP:                 FILE: Number of large read operations=0
HADOOP:                 FILE: Number of write operations=0
HADOOP:                 HDFS: Number of bytes read=2156095226
HADOOP:                 HDFS: Number of bytes written=5251252
HADOOP:                 HDFS: Number of read operations=573
HADOOP:                 HDFS: Number of large read operations=0
HADOOP:                 HDFS: Number of write operations=2
HADOOP:         Job Counters
HADOOP:                 Killed map tasks=4
HADOOP:                 Launched map tasks=194
HADOOP:                 Launched reduce tasks=1
HADOOP:                 Other local map tasks=2
HADOOP:                 Data-local map tasks=190
HADOOP:                 Rack-local map tasks=2
HADOOP:                 Total time spent by all maps in occupied slots (ms)=22263989
HADOOP:                 Total time spent by all reduces in occupied slots (ms)=408000
HADOOP:                 Total time spent by all map tasks (ms)=22263989
HADOOP:                 Total time spent by all reduce tasks (ms)=408000
HADOOP:                 Total vcore-milliseconds taken by all map tasks=22263989
HADOOP:                 Total vcore-milliseconds taken by all reduce tasks=408000
HADOOP:                 Total megabyte-milliseconds taken by all map tasks=22798324736
HADOOP:                 Total megabyte-milliseconds taken by all reduce tasks=417792000
HADOOP:         Map-Reduce Framework
HADOOP:                 Map input records=58682266
HADOOP:                 Map output records=293411330
HADOOP:                 Map output bytes=3136729760
HADOOP:                 Map output materialized bytes=132001630
HADOOP:                 Input split bytes=26110
HADOOP:                 Combine input records=298596684
HADOOP:                 Combine output records=8216908
HADOOP:                 Reduce input groups=343019
HADOOP:                 Reduce shuffle bytes=132001630
HADOOP:                 Reduce input records=3031554
HADOOP:                 Reduce output records=343019
HADOOP:                 Spilled Records=10953597
HADOOP:                 Shuffled Maps =190
HADOOP:                 Failed Shuffles=0
HADOOP:                 Merged Map outputs=190
HADOOP:                 GC time elapsed (ms)=82586
HADOOP:                 CPU time spent (ms)=21252850
HADOOP:                 Physical memory (bytes) snapshot=51913187328
HADOOP:                 Virtual memory (bytes) snapshot=406539476992
HADOOP:                 Total committed heap usage (bytes)=37774426112
HADOOP:         Shuffle Errors
HADOOP:                 BAD_ID=0
HADOOP:                 CONNECTION=0
HADOOP:                 IO_ERROR=0
HADOOP:                 WRONG_LENGTH=0
HADOOP:                 WRONG_MAP=0
HADOOP:                 WRONG_REDUCE=0
HADOOP:         File Input Format Counters
HADOOP:                 Bytes Read=2156069116
HADOOP:         File Output Format Counters
HADOOP:                 Bytes Written=5251252
HADOOP: Output directory: hdfs:///user/root/tmp/mrjob/hw5.root.20160217.142415.557030/step-output/1
Counters from step 1:
  (no counters found)
Detected hadoop configuration property names that do not match hadoop version 2.7.2:
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
HADOOP: packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/s
treamjob6811890661771153365.jar tmpDir=null
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Total input paths to process : 1
HADOOP: number of splits:2
HADOOP: mapred.text.key.comparator.options is deprecated. Instead, use mapreduce.partition.keycomparat
or.options
HADOOP: mapred.output.key.comparator.class is deprecated. Instead, use mapreduce.job.output.key.compar
ator.class
HADOOP: Submitting tokens for job: job_1455677115242_0033
HADOOP: Submitted application application_1455677115242_0033
HADOOP: The url to track the job: http://master:8088/proxy/application_1455677115242_0033/
HADOOP: Running job: job_1455677115242_0033
HADOOP: Job job_1455677115242_0033 running in uber mode : false
HADOOP:  map 0% reduce 0%
HADOOP:  map 42% reduce 0%
HADOOP:  map 62% reduce 0%
HADOOP:  map 100% reduce 0%
```

```
HADOOP:  map 100% reduce 0%
HADOOP:  map 100% reduce 78%
HADOOP:  map 100% reduce 84%
HADOOP:  map 100% reduce 91%
HADOOP:  map 100% reduce 96%
HADOOP:  map 100% reduce 100%
HADOOP: Job job_1455677115242_0033 completed successfully
HADOOP: Counters: 49
HADOOP:         File System Counters
HADOOP:                 FILE: Number of bytes read=5937296
HADOOP:                 FILE: Number of bytes written=12242556
HADOOP:                 FILE: Number of read operations=0
HADOOP:                 FILE: Number of large read operations=0
HADOOP:                 FILE: Number of write operations=0
HADOOP:                 HDFS: Number of bytes read=5255642
HADOOP:                 HDFS: Number of bytes written=5251252
HADOOP:                 HDFS: Number of read operations=9
HADOOP:                 HDFS: Number of large read operations=0
HADOOP:                 HDFS: Number of write operations=2
HADOOP:         Job Counters
HADOOP:                 Launched map tasks=2
HADOOP:                 Launched reduce tasks=1
HADOOP:                 Data-local map tasks=2
HADOOP:                 Total time spent by all maps in occupied slots (ms)=25505
HADOOP:                 Total time spent by all reduces in occupied slots (ms)=19097
HADOOP:                 Total time spent by all map tasks (ms)=25505
HADOOP:                 Total time spent by all reduce tasks (ms)=19097
HADOOP:                 Total vcore-milliseconds taken by all map tasks=25505
HADOOP:                 Total vcore-milliseconds taken by all reduce tasks=19097
HADOOP:                 Total megabyte-milliseconds taken by all map tasks=26117120
HADOOP:                 Total megabyte-milliseconds taken by all reduce tasks=19555328
HADOOP:         Map-Reduce Framework
HADOOP:                 Map input records=343019
HADOOP:                 Map output records=343019
HADOOP:                 Map output bytes=5251252
HADOOP:                 Map output materialized bytes=5937302
HADOOP:                 Input split bytes=294
HADOOP:                 Combine input records=0
HADOOP:                 Combine output records=0
HADOOP:                 Reduce input groups=41362
HADOOP:                 Reduce shuffle bytes=5937302
HADOOP:                 Reduce input records=343019
HADOOP:                 Reduce output records=343019
HADOOP:                 Spilled Records=686038
HADOOP:                 Shuffled Maps =2
HADOOP:                 Failed Shuffles=0
HADOOP:                 Merged Map outputs=2
HADOOP:                 GC time elapsed (ms)=221
HADOOP:                 CPU time spent (ms)=38540
HADOOP:                 Physical memory (bytes) snapshot=706281472
HADOOP:                 Virtual memory (bytes) snapshot=6388023296
HADOOP:                 Total committed heap usage (bytes)=542113792
HADOOP:         Shuffle Errors
HADOOP:                 BAD_ID=0
HADOOP:                 CONNECTION=0
HADOOP:                 IO_ERROR=0
HADOOP:                 WRONG_LENGTH=0
HADOOP:                 WRONG_MAP=0
HADOOP:                 WRONG_REDUCE=0
HADOOP:         File Input Format Counters
HADOOP:                 Bytes Read=5255348
HADOOP:         File Output Format Counters
HADOOP:                 Bytes Written=5251252
HADOOP: Output directory: hdfs:///hw5.3-2-output
Counters from step 2:
  (no counters found)
removing tmp directory /tmp/hw5.root.20160217.142415.557030
deleting hdfs:///user/root/tmp/mrjob/hw5.root.20160217.142415.557030 from HDFS

real    9m38.781s
user    0m33.362s
sys     0m2.039s
+ hdfs dfs -cat foo/part-00000
+ head
cat: Unable to write to output stream.
```

In [171]:
```
%%writefile hw5.3-3.py
#!/usr/bin/env python

from collections import namedtuple
```

```
from mrjob.job import MRJob
from mrjob.step import MRStep
import sys

class HW53Job(MRJob):
    Row=namedtuple('Row',['ngram', 'count', 'pages_count', 'books_count'])
    @staticmethod
    def split_line(line):
        fields=line.strip().split('\t')
        return HW53Job.Row(fields[0],*[int(field) for field in fields[1:]])

    """Produce the word densities and sort them"""
    def get_counts(self, _, line):
        """Mapper: split the 5-grams, and yield the count with each."""
        row=HW53Job.split_line(line)
        for word in row.ngram.split():
            yield word, (row.count, row.pages_count)
    def sum_counts(self, key, values):
        """Combiner: sum the counts as in usual word count"""
        count, page_count = 0,0
        for val in values:
            count+=val[0]
            page_count+=val[1]
        yield key, (count, page_count)
    def calc_freqs(self, key, values):
        """Reducer: sum the counts as in usual word count and swap key/val for sorting"""
        for _,(count, page_count) in self.sum_counts(key, values):
            yield 1.0*count/page_count, key

    ## dummy map/red steps to cause another sort
    def map_id(self, key, val):
        yield key, val
    def red_id(self, key, vals):
        for x in vals:
            yield key, x

    def steps(self):
        return [
            MRStep(
                mapper=self.get_counts,
                combiner=self.sum_counts,
                reducer=self.calc_freqs,
            ),
            MRStep(
                mapper=self.map_id,
                reducer=self.red_id,
                jobconf={
                    'mapred.output.key.comparator.class':
                       'org.apache.hadoop.mapred.lib.KeyFieldBasedComparator',
                    'mapred.text.key.comparator.options': '-k1,1nr',
                },
            ),
        ]

if __name__=='__main__':
    HW53Job().run()
    exit(0)
```

```
Overwriting hw5.3-3.py
```

In [176]:

```
%%bash
/usr/bin/ssh root@50.22.252.4 bash -xs <<'EOF'
cd hw5
prog=hw5.3-3
hdfs=hdfs://master:9000
input=$hdfs/filtered-5grams
output=$hdfs/$prog-output
HADOOP_ROOT_LOGGER=WARN,console
hdfs dfs -rm -r -f $output
time ./$prog.py -q -r hadoop --no-bootstrap-mrjob \
  --no-output --output $output \
  $input
hdfs dfs -cat ${output}/* | head -1000
EOF
```

```
11.557291666666666      "xxxx"
10.161726044782885      "NA"
8.074159907300116       "blah"
7.533333333333333       "nnn"
6.561143644505684       "nd"
```

5.40736428467472 "ND"
4.921875 "ooooooooooooooo"
4.7272727272727275 "PIC"
4.511627906976744 "llll"
4.349498327759197 "LUTHER"
4.207237859573151 "oooooo"
4.0908402725208175 "NN"
3.9492846924177396 "ooooo"
3.9313725490196076 "OOOOOO"
3.7877030162412995 "IIII"
3.7624521072796937 "lillelu"
3.6570701447431206 "OOOOO"
3.6065625 "Sc"
3.576923076923077 "Pfeffermann"
3.576923076923077 "Madarassy"
3.56 "Meteoritical"
3.536491677336748 "Undecided"
3.505639097744361 "Lib"
3.5 "xxxxxxxx"
3.4791318864774623 "ri"
3.375068493150685 "Vir"
3.2390171258376768 "DREAM"
3.229038854805726 "beep"
3.188679245283019 "Latha"
3.188317505823329 "MARTIN"
3.1699346405228757 "Lis"
3.1147458480120784 "Ac"
3.037142857142857 "OUTPUT"
3.022222222222222 "HENNESSY"
3.0 "ALLIS"
2.9191176470588234 "IYENGAR"
2.869891270467005 "ft"
2.8432451923076925 "Adapted"
2.825 "counterfeiteth"
2.81981981981982 "nonmorular"
2.81981981981982 "nonsquamous"
2.8085106382978724 "RHYME"
2.74468085106383 "YOUTHS"
2.7264957264957266 "Poing"
2.7 "Kuhl"
2.6748466257668713 "Sirignano"
2.673469387755102 "METRES"
2.673469387755102 "YARDS"
2.66414686825054 "Illl"
2.660377358490566 "Neophytos"
2.660377358490566 "Edelby"
2.640311804008909 "CHOH"
2.627005347593583 "Vocht"
2.5973377703826954 "Inactive"
2.595744680851064 "plappern"
2.595744680851064 "schwatzen"
2.58974358974359 "Horida"
2.575 "undrinkable"
2.561606725829052 "Phys"
2.5521472392638036 "Radiat"
2.5166666666666666 "Nambo"
2.5133995784402288 "Sag"
2.490549791191286 "Oncol"
2.476923076923077 "Wadiar"
2.4754098360655736 "Villanelle"
2.467065868263473 "ingrossed"
2.456140350877193 "USF"
2.4508196721311477 "HUMBER'S"
2.4508196721311477 "cillum"
2.418472063854048 "pryntinge"
2.418472063854048 "lycense"
2.406764168190128 "FLEISCHER"
2.4 "midell"
2.3991416309012874 "Chimeric"
2.394578313253012 "Compilations"
2.3846153846153846 "rauens"
2.3734177215189876 "Brosman"
2.369565217391304 "Bursch"
2.357142857142857 "Scheffauer"
2.3553750966744005 "CHO"
2.3421727972626174 "Bl"
2.3333333333333335 "Vacchagotta"
2.3316326530612246 "RAVI"
2.3114754098360657 "Carnwarth"

2.3112582781456954 "Logotherapy"
2.3 "aliquyam"
2.298611111111111 "Bev"
2.2954545454545454 "Realschulen"
2.276190476190476 "irure"
2.276190476190476 "reprehenderit"
2.2441176470588236 "Intraocul"
2.234375 "Jerico"
2.2339449541284404 "JAF"
2.2339449541284404 "Roodbergen"
2.2325581395348837 "Sanatana"
2.232142857142857 "Fouilles"
2.2262351921409995 "tO"
2.2222222222222223 "douff"
2.207920792079208 "Cynddylan"
2.2051282051282053 "DOLLS"
2.193548387096774 "venireman"
2.1932515337423313 "Snorkeling"
2.181858567909212 "Pathol"
2.1814321874744835 "Adolesc"
2.1706827309236947 "ff"
2.1597938144329896 "mumblin"
2.1564625850340136 "PINT"
2.153960029607698 "Aq"
2.1475409836065573 "deserunt"
2.1475409836065573 "mollit"
2.140625 "Triturus"
2.14 "defalt"
2.1372549019607843 "Kevo"
2.136842105263158 "trigram"
2.135135135135135 "gefahrlicher"
2.1325301204819276 "Multiconference"
2.1221826809015423 "JPEN"
2.1221826809015423 "Parenter"
2.1144619602467443 "COM"
2.0991166077738517 "ESTATE"
2.0943775100401605 "MOV"
2.091549295774648 "MARJORIE"
2.087719298245614 "Talich"
2.083565459610028 "cun"
2.082666666666667 "tempor"
2.082666666666667 "incididunt"
2.08125 "herrin"
2.0806451612903225 "eiusmod"
2.075187969924812 "LEKYTHOI"
2.063960639606396 "Pogue"
2.0588235294117645 "Meseguer"
2.0588235294117645 "Goguen"
2.0491803278688523 "Blissful"
2.0472560975609757 "tcp"
2.0466101694915255 "Mekhilta"
2.0412451361867703 "laoreet"
2.04 "Phelan's"
2.0363475177304964 "Ub"
2.0357142857142856 "Pos"
2.0357142857142856 "Neg"
2.030075187969925 "pericranial"
2.028013965906757 "Surg"
2.0268456375838926 "nonperfection"
2.026143790849673 "MAZEL"
2.025189283634246 "CL"
2.0217606330365974 "IDIOMS"
2.016872160934458 "Neusner"
2.010204081632653 "Hampartumian"
2.009625668449198 "Lorem"
2.008873596522999 "Intern"
2.0051730998806208 "SP"
2.0 "GROWlNG"
2.0 "Chamrajendra"
2.0 "miming"
1.9909638554216869 "ullamco"
1.9852941176470589 "Aposentadoria"
1.9852941176470589 "Pensoes"
1.978021978021978 "PELL"
1.9765625 "Waigal"
1.9704704704704705 "amet"
1.969450101832994 "euismod"
1.969450101832994 "tincidunt"
1.969450101832994 "nibh"
1.969450101832994 "nonummy"

1.9685336689741977 "LESTER"
1.9660869565217391 "Expense"
1.965034965034965 "Pis"
1.9644268774703557 "Berenstain"
1.962962962962963 "Ventriloquist"
1.9610004936646372 "Artif"
1.9591836734693877 "Eidhnech"
1.9591836734693877 "Cluain"
1.9583333333333333 "Branover"
1.9583333333333333 "Yeshajahu"
1.957142857142857 "Beforderung"
1.9545454545454546 "Phytochemicals"
1.9505813953488371 "microbiologica"
1.9505813953488371 "pathologica"
1.9461538461538461 "pariatur"
1.9461538461538461 "fugiat"
1.9397590361445782 "HEUVEN"
1.9387755102040816 "Charuhas"
1.9366197183098592 "Aggadic"
1.9363636363636363 "Breakfasts"
1.9298245614035088 "Cropanzano"
1.9285714285714286 "arthropodan"
1.9268292682926829 "Civilian's"
1.92580724673404 "Disagree"
1.925 "pickety"
1.9157894736842105 "LEMESHOW"
1.9157894736842105 "HOSMER"
1.912736265677442 "Gem"
1.9078947368421053 "Indemnification"
1.9059829059829059 "Reprocessing"
1.90234375 "posie"
1.8945544554455445 "dolore"
1.893939393939394 "HOLOCAUST"
1.8933333333333333 "Oakton"
1.8907917383820998 "Cl"
1.8904428904428905 "Redonnet"
1.888 "VOLTAGE"
1.888 "NODE"
1.8877679697351828 "consequat"
1.8870967741935485 "Thleary"
1.8859223300970873 "Tuberc"
1.885782475281282 "ti"
1.881720430107527 "starcher"
1.88135593220339 "wisi"
1.8801213960546281 "aliquip"
1.8796895213454075 "VUNAKIS"
1.8793103448275863 "UMN"
1.8793103448275863 "LMN"
1.879120879120879 "IOOO"
1.8789473684210527 "pistol's"
1.878787878787879 "nostrud"
1.8780487804878048 "wedres"
1.8780487804878048 "wintres"
1.8780487804878048 "overshake"
1.8775510204081634 "histochemischen"
1.8775510204081634 "Sichtbarmachung"
1.8775510204081634 "Fluoreszenzmethoden"
1.8764044943820224 "OMe"
1.875 "Unease"
1.875 "McEwan's"
1.872340425531915 "BURLINGAME"
1.8708708708708708 "CALENDAR"
1.8683035714285714 "lobortis"
1.8652857042203903 "Biol"
1.8613861386138615 "Mandelbrote"
1.859922178988327 "LILLEY"
1.8596491228070176 "overseership"
1.8571428571428572 "Wagrushka"
1.85678391959799 "ullamcorper"
1.85678391959799 "exerci"
1.85678391959799 "suscipit"
1.8562091503267975 "Fonda's"
1.853846153846154 "Moggallana"
1.8535469107551488 "veniam"
1.8514851485148516 "Berubari"
1.8512396694214877 "aliquam"
1.8469750889679715 "diam"
1.8444444444444446 "Archaeol"
1.84375 "isteach"

```
1.84375 "phost"
1.8414634146341464      "curiousest"
1.8412887828162292      "commodo"
1.8369098712446352      "Ackermann's"
1.836734693877551       "BUSHY"
1.8339041095890412      "Optom"
1.8314606741573034      "Inflorescence"
1.8297872340425532      "Portlock's"
1.8296543107038734      "lournal"
1.828125        "Thelephoraceae"
1.8269230769230769      "gostak"
1.8269230769230769      "distimmed"
1.8268150896722326      "Med"
1.826086956521739       "Crocombe"
1.8251748251748252      "Corson's"
1.8245614035087718      "becam"
1.8229166666666667      "Primov"
1.8214285714285714      "Obeisance"
1.8181818181818181      "Immobil"
1.8181818181818181      "Substit"
1.8179611650485437      "scandinavica"
1.8166666666666667      "haerede"
1.8166563082660037      "iiii"
1.8148148148148149      "tation"
1.8109028960817717      "Topchiev"
1.8012820512820513      "supposit"
1.8     "propios"
1.8     "Vibrating"
1.7966101694915255      "laboris"
1.7964071856287425      "TONY"
1.7946428571428572      "Darrow's"
1.794392523364486       "ACCOUNTANCY"
1.793103448275862       "Mikrofossilien"
1.793103448275862       "Silurs"
1.793103448275862       "baltischen"
1.793103448275862       "mault"
1.7923076923076924      "PONY"
1.7916666666666667      "Immunoisolation"
1.7892813641900123      "cadastre"
1.7882352941176471      "Scanty"
1.7872340425531914      "Maarten"
1.78    "Midship"
1.7752808988764044      "TRE"
1.775   "Eval"
1.775   "Carcinog"
1.7748303516347934      "Endod"
1.7745098039215685      "TEXTURES"
1.7720694645441388      "dolor"
1.7719298245614035      "Pickpocket"
1.7695652173913043      "Lightwave"
1.7692307692307692      "HERCULES"
1.7692307692307692      "leive"
1.7647058823529411      "Hollandais"
1.7622504537205081      "thalweg"
1.7619047619047619      "Lavu"
1.76    "Fearest"
1.7570093457943925      "Nonpassive"
1.7553191489361701      "Tanzi"
1.7549019607843137      "Ferratas"
1.7547169811320755      "Intl"
1.7524613220815752      "Tra"
1.7521008403361344      "Cryogenic"
1.751269035532995       "molestie"
1.7501040366208906      "Strongly"
1.75    "Marcia's"
1.75    "Auricle"
1.7467248908296944      "BATSFORD"
1.7452229299363058      "chou"
1.742873541228238       "Expiration"
1.7407302173604204      "Radiol"
1.74    "catechetics"
1.7384615384615385      "warrantee"
1.7370689655172413      "EMBASSY"
1.7354596622889307      "BR"
1.7349081364829397      "Digumarti"
1.7339246119733924      "punkin"
1.7337733773377337      "RVs"
1.732484076433121       "Pellegrino"
1.73    "Misadventure"
1.729166666666667       "inflexional"
```

1.7291666666666667 "Inflexional"
1.728395061728395 "Gratuitos"
1.728395061728395 "Texto"
1.7272727272727273 "RUBIN"
1.7258883248730965 "prelabour"
1.7241379310344827 "Jur"
1.7183098591549295 "lungless"
1.7154471544715446 "Recalcitrant"
1.712676056338028 "Mosq"
1.7122302158273381 "Nubar"
1.7117971334068358 "INTERNAL"
1.7117117117117118 "Gallons"
1.711111111111111 "surviver"
1.7092198581560283 "Oto"
1.7076882007474639 "CENRO"
1.707142857142857 "dignissim"
1.7065868263473054 "volke"
1.7065868263473054 "miscalle"
1.704 "Desalination"
1.7035398230088497 "Kareah"
1.7031924072476272 "SUFFRAGE"
1.703125 "Implanted"
1.7030303030303031 "iusto"
1.7026143790849673 "Aeroengine"
1.702020202020202 "ARCHETYPES"
1.7001321003963012 "vulputate"
1.6978417266187051 "Medineh"
1.6970338983050848 "exercitation"
1.696969696969697 "jinns"
1.6964285714285714 "Buu"
1.6956521739130435 "GOSCH"
1.6949924127465856 "Opportunistic"
1.6933019976498238 "Spectrom"
1.6923076923076923 "verrucose"
1.6923076923076923 "initdefault"
1.690909090909091 "Straf"
1.6901408450704225 "Salomos"
1.6890756302521008 "Tahle"
1.6890756302521008 "Stahle"
1.6868686868686869 "prenumeraty"
1.6833333333333333 "typologique"
1.6829268292682926 "Unicellular"
1.6818181818181819 "dientes"
1.6818181818181819 "parientes"
1.6766169154228856 "Retrieves"
1.6763085399449036 "Biographer"
1.676056338028169 "Leaming"
1.6744186046511629 "Larras"
1.674074074074074 "Deuterium"
1.672627235213205 "Waitotara"
1.671641791044776 "TORRANCE"
1.6691365979381443 "Nutr"
1.6688741721854305 "JONES'S"
1.66875 "Greenwillow"
1.6666666666666667 "Klingel"
1.6666666666666667 "Deluxe"
1.6666666666666667 "lmages"
1.6623376623376624 "Axson"
1.6615720524017468 "FPGAs"
1.6612903225806452 "Shatin"
1.6611721611721613 "Ormazd"
1.6610169491525424 "IACHR"
1.6608695652173913 "Resolusies"
1.6603773584905661 "MeO"
1.659722222222223 "noninferior"
1.6586345381526104 "Psychosoc"
1.656326932545544 "Auditor's"
1.65625 "REITER"
1.655 "Borisov"
1.6544117647058822 "voluptate"
1.654320987654321 "Wyes"
1.6538461538461537 "Sukhavati"
1.6534090909090908 "MEDAL"
1.653061224489796 "Percycross"
1.6527777777777777 "Mischling"
1.6515151515151516 "LYING"
1.6507936507936507 "Venation"
1.6501457725947521 "praesent"
1.6501457725947521 "blandit"
1.6481481481481481 "Trucks"

1.647887323943662 "occupacion"
1.6470588235294117 "transactors"
1.6470588235294117 "Fert"
1.6458333333333333 "Onc"
1.6442953020134228 "unnat"
1.6431492842535786 "Cents"
1.6428571428571428 "capellet"
1.6428571428571428 "lanceam"
1.6425891181988743 "Bodywork"
1.6415094339622642 "Briefcase"
1.6407185628742516 "Nurs"
1.639344262295082 "Literarische"
1.6385542168674698 "Hortic"
1.6376811594202898 "Shaying"
1.6363636363636365 "Morhidity"
1.6355140186915889 "Microelectromechanical"
1.6354938271604937 "Retroviruses"
1.6349041901321149 "Acad"
1.631704410011919 "Fluidized"
1.6294416243654823 "Giovane"
1.6281859070464768 "MATTOS"
1.6279294983536703 "cn"
1.6274509803921569 "anticariogenic"
1.6271186440677967 "nmol"
1.6271186440677967 "kwh"
1.6262626262626263 "cowkeeper"
1.6246096189881325 "MISCELLANEOUS"
1.6229508196721312 "hayfork"
1.6226415094339623 "reviewest"
1.6219512195121952 "createst"
1.6212653778558874 "velit"
1.6206896551724137 "endurethfor"
1.6203703703703705 "Virologists"
1.620123203285421 "STAMP"
1.619718309859155 "promisedst"
1.618805849889089 "Natl"
1.6179520801048992 "YES"
1.6177303782702883 "vellum"
1.6157635467980296 "Patently"
1.6143344709897611 "birk"
1.6140350877192982 "fquares"
1.6132075471698113 "SIB"
1.6129032258064515 "bec"
1.6122448979591837 "microcode"
1.6122448979591837 "Littledale's"
1.6115702479338843 "Finlande"
1.6101694915254237 "Birket"
1.609907120743034 "hauke"
1.609628217349857 "Aa"
1.609375 "bizniss"
1.6074074074074074 "Rabutin"
1.606936416184971 "Lubbub"
1.6046511627906976 "gootther"
1.6013071895424837 "Taffy's"
1.6011016913876595 "Proc"
1.6006933501418217 "ipsum"
1.6 "Instituta"
1.5996938384998087 "keratomileusis"
1.598802395209581 "Jee"
1.5975609756097562 "Pilcher"
1.5974781765276431 "Mitchenson"
1.5974025974025974 "Armorer"
1.5968992248062015 "Hwe"
1.5957446808510638 "fossile"
1.5952380952380953 "Didelphys"
1.5930232558139534 "cyanophytes"
1.5925925925925926 "Guvener"
1.59247889485802 "Implant"
1.5924242424242425 "meritos"
1.5921787709497206 "Baussiere"
1.5919003115264798 "doy"
1.5914634146341464 "DESTROYED"
1.5909090909090908 "Autohiographical"
1.5899390243902438 "Bequeath"
1.5893333333333333 "CARL"
1.588623732952658 "Sci"
1.587378640776699 "Tue"
1.5866666666666667 "Zerahites"
1.5853658536585367 "Mitchelson"

1.5846153846153845 "Angoysses"
1.5846153846153845 "douloureuses"
1.5846153846153845 "procedent"
1.5844155844155845 "costilla"
1.5820895522388059 "Namburbi"
1.5820895522388059 "Scarpa"
1.58203125 "Cruikshank's"
1.5818181818181818 "Aichi"
1.5818181818181818 "Actto"
1.5809768637532133 "wunt"
1.5802707930367506 "Gyrate"
1.5802469135802468 "Cotran"
1.5801526717557253 "Ufficiale"
1.5795454545454546 "mearrah"
1.5791139240506329 "luptatum"
1.5791139240506329 "zzril"
1.5789473684210527 "SelectionMode"
1.5781725888324873 "Polym"
1.5777777777777777 "BALEK"
1.5777777777777777 "Episcopale"
1.5775075987841944 "Taile"
1.5769230769230769 "engorda"
1.576470588235294 "Observatorium"
1.576470588235294 "Meteorologisch"
1.576470588235294 "Magnetisch"
1.576086956521739 "Metaplastic"
1.5757804252752223 "Agree"
1.5757575757575757 "Terracing"
1.5717488789237668 "SOMEBODY"
1.5714285714285714 "Rogerian"
1.5714285714285714 "visaed"
1.5689655172413792 "distolingual"
1.568445475638051 "Instructs"
1.5681818181818181 "Echinoidea"
1.5681063122923589 "HELEN"
1.5680473372781065 "Nisir"
1.5675675675675675 "Gord"
1.5668789808917198 "AAMFT"
1.5666666666666667 "Commutative"
1.5662650602409638 "Helminthic"
1.5660814839350778 "char"
1.565217391304348 "ofEnology"
1.565217391304348 "Superelastic"
1.5641646489104115 "Autographa"
1.5619136960600375 "Specify"
1.5612244897959184 "Calliactis"
1.56 "Heliozoa"
1.5593220338983051 "dicast"
1.5588235294117647 "Biblesoft"
1.558455114822547 "wapentake"
1.5573770491803278 "Endodontics"
1.5572916666666667 "Paleontol"
1.5555555555555556 "LONSDALE"
1.5546875 "phred"
1.5545454545454545 "GHD"
1.553763440860215 "laissie"
1.553191489361702 "WAGON"
1.5524861878453038 "Scatophaga"
1.5511811023622046 "shocK"
1.5510204081632653 "FMI"
1.5510204081632653 "Codicils"
1.55 "Krishnamurti"
1.55 "Myxoid"
1.5493827160493827 "halachah"
1.5490196078431373 "DAI"
1.5489534098582038 "GILBERT"
1.5472972972972974 "BiCMOS"
1.5465116279069768 "Coenzyme"
1.5460526315789473 "Samanas"
1.5447619047619048 "Akaroa"
1.5436893203883495 "PLATH"
1.5436893203883495 "HOMAGE"
1.5423728813559323 "JerseyRobert"
1.5419847328244274 "Thur"
1.5416666666666667 "chemosurgery"
1.5407725321888412 "hendrerit"
1.5406859448554135 "Chieh"
1.5405405405405406 "previllous"
1.5404310186037944 "Unlawful"
1.5394190871369295 "Mischlinge"

1.5392670157068062 "Siboga"
1.538812785388128 "Spillius"
1.5384615384615385 "bidis"
1.5373953219751662 "MOLDAVE"
1.537313432835821 "matchsticks"
1.536470588235294 "Stenographisches"
1.5363636363636364 "Feyl"
1.5363636363636364 "Chavkin"
1.5357142857142858 "Psychoanalyzing"
1.53475935828877 "procreatis"
1.5347222222222223 "Audited"
1.5340236686390532 "WOLD"
1.5330012453300124 "REGARDLESS"
1.5304232804232805 "SHOCK"
1.5294117647058822 "Nieves"
1.5294117647058822 "Recode"
1.5291734197730957 "Hyg"
1.5284552845528456 "Oholibamah"
1.528301886792453 "urodelean"
1.5276595744680852 "unsorted"
1.5274725274725274 "Maja"
1.5272727272727273 "Bezoar"
1.5272257230570367 "Trop"
1.527027027027027 "Videha"
1.5268817204301075 "brances"
1.5260208166533227 "Drilling"
1.5255198487712665 "Rangitikei"
1.5227272727272727 "Katzeff"
1.5225643748340856 "Min"
1.5223880597014925 "Ibs"
1.521594684385382 "Mycological"
1.5215759849906192 "Isot"
1.521505376344086 "Monkmeyer"
1.521172638436482 "Memher"
1.5211267605633803 "NeuroBloc"
1.5209834675710046 "Geoscience"
1.5207100591715976 "Repelita"
1.52 "Appealed"
1.5191815856777493 "canakin"
1.5189831898318984 "Bone"
1.5180722891566265 "Intermediaire"
1.5179487179487179 "Friedhoff"
1.5178571428571428 "Parme's"
1.5173237753882916 "Urogynecol"
1.5172413793103448 "intergrown"
1.5169491525423728 "Toroidal"
1.5161290322580645 "progressiva"
1.5161290322580645 "harse"
1.5158730158730158 "roily"
1.5156794425087108 "equivocated"
1.5151515151515151 "salticid"
1.5147058823529411 "trinomial"
1.5144927536231885 "Ultrasonics"
1.5142857142857142 "Codicibus"
1.5141776937618148 "Minelike"
1.5141752577319587 "Cetacean"
1.514018691588785 "Averments"
1.5138888888888888 "Aportaciones"
1.5127919911012235 "Flaws"
1.5121951219512195 "Durzan"
1.5121951219512195 "Moats"
1.5121951219512195 "Bonga"
1.5120481927710843 "Adjournments"
1.5119617224880382 "Amarillo"
1.5119047619047619 "Besonderheiten"
1.51171875 "mustached"
1.5116279069767442 "ffreemen"
1.5114155251141552 "Ebeling"
1.5106382978723405 "Gado"
1.5104166666666667 "Panhellenion"
1.5090991810737033 "Benthological"
1.5087719298245614 "Nonfinancial"
1.5084745762711864 "Multiclass"
1.5078534031413613 "Kheta"
1.5076923076923077 "Ostdeutschland"
1.5076923076923077 "Anpassungsprozesse"
1.5076923076923077 "unternehmerische"
1.5076923076923077 "ingresos"
1.5059920106524634 "Mending"

1.5058823529411764 "Enabled"
1.5034122842232036 "Volcanology"
1.5031847133757963 "ampholytes"
1.5030345471521942 "Rosen"
1.501466275659824 "Harnack's"
1.5011961722488039 "TEIXEIRA"
1.5 "ANNING"
1.5 "SWEEPSTAKES"
1.5 "Koniunktur"
1.5 "Ingbar's"
1.5 "Cen"
1.5 "Okeanografii"
1.5 "Khozyaistva"
1.5 "Rybnogo"
1.5 "Handlu"
1.5 "Arkham"
1.5 "Aupres"
1.5 "interganglionic"
1.5 "WorId's"
1.5 "lmprovement"
1.5 "tetanized"
1.5 "nigrum"
1.499222395023328 "Tackle"
1.498949894989499 "Bit"
1.4978723404255319 "DAVIDS"
1.4965299040620534 "Chem"
1.496 "Hola"
1.4959349593495934 "Herpetic"
1.4954954954954955 "sitae"
1.4952380952380953 "AUTHORSHIP"
1.4949367088607595 "servicios"
1.49486301369863 "Phe"
1.4945054945054945 "GAMBIA"
1.4944567627494456 "Geothermal"
1.4943683675414627 "Oral"
1.494296577946768 "SITTING"
1.4939271255060729 "USDS"
1.4936708860759493 "Antiga"
1.4933333333333334 "holomorphy"
1.492537313432836 "nightes"
1.4925187032418952 "Sakha"
1.4915254237288136 "JPB"
1.4912240184757506 "MATTER"
1.4911417322834646 "odio"
1.490566037735849 "Korkut"
1.490566037735849 "Tarratines"
1.490566037735849 "Dede"
1.490066225165563 "Regionalisms"
1.4900542495479205 "IRC"
1.4893617021276595 "obligatoire"
1.4893617021276595 "ngo"
1.488888888888889 "PIPELINE"
1.4885057471264367 "Slumber's"
1.4882005319915614 "Hearing"
1.4881516587677726 "int"
1.4880952380952381 "Soloukhin"
1.4875283446712018 "SABATO"
1.4872389791183294 "Lonergan"
1.4862385321100917 "bequeathe"
1.4851485148514851 "masculis"
1.484375 "Cranio"
1.4814814814814814 "electroencephalic"
1.481012658227848 "RESUMES"
1.4808743169398908 "herede"
1.480357142857143 "Vide"
1.48 "Avantgarde"
1.4790419161676647 "CATTERMOLE"
1.4786324786324787 "Manshu"
1.4786324786324787 "kaisha"
1.4786324786324787 "tetsudo"
1.4786324786324787 "kabushiki"
1.4785714285714286 "loumal"
1.4782608695652173 "flavicollis"
1.4782608695652173 "Kalotermes"
1.4782608695652173 "Choudhry"
1.4782608695652173 "Moorad"
1.4782608695652173 "Termite"
1.4774980930587338 "Tu"
1.4761904761904763 "Moraba"
1.476027397260274 "emendata"

1.4760273972602740 "emendata"
1.4759825327510918 "Coverdell"
1.475177304964539 "GLAZE"
1.475 "Cusps"
1.4745529573590097 "Azincourt"
1.4741935483870967 "Turneaure's"
1.473529411764706 "SPE"
1.472972972972973 "Embalmed"
1.472972972972973 "Urheberrecht"
1.4728571428571429 "PENR"
1.4728571428571429 "PAGANO"
1.4727272727272727 "harmoniser"
1.4727272727272727 "YOZAN"
1.4723926380368098 "Viticultural"
1.4723404255319148 "Mendeleev"
1.4712041884816753 "Gig"
1.4695259593679457 "Vasc"
1.4695259593679457 "Endovasc"
1.4690265486725664 "herpetology"
1.4688686408504177 "KIVIE"
1.4680851063829787 "ROUTINE"
1.4678899082568808 "Innocently"
1.4678228165482659 "Respir"
1.467032967032967 "TODD"
1.466992665036675 "Outdoors"
1.4666666666666666 "parc"
1.4666666666666666 "geta"
1.4659090909090908 "CRUSH"
1.4655172413793103 "Anni"
1.4651162790697674 "metaphosphoric"
1.4649122807017543 "Yeb"
1.4646464646464648 "epiphysiolysis"
1.4646017699115044 "TSCA"
1.4644194756554307 "Willebois"
1.463855421686747 "Politieke"
1.463768115942029 "Cadran"
1.4636363636363636 "Thacher"
1.4636363636363636 "obiit"
1.4631357207342763 "SIDNEY"
1.4630872483221478 "Endophthalmitis"
1.4628297362110312 "Fanfare"
1.4625 "diversicolor"
1.4623655913978495 "Chicorel"
1.4615384615384615 "Paddock's"
1.4615384615384615 "Cygni"
1.4615384615384615 "Explode"
1.4612546125461254 "GROSSMAN"
1.4607843137254901 "Alimentos"
1.4606741573033708 "Yonson"
1.4602076124567474 "BARTHOLOMEW"
1.4588235294117646 "Jumps"
1.4583333333333333 "Ultracentrifugation"
1.4583333333333333 "literaturas"
1.4580152671755726 "Ihram"
1.4578005115089514 "Gethsemani"
1.4577259475218658 "Alinari"
1.457543281121187 "betroth"
1.4566929133858268 "Vaasa"
1.4565217391304348 "homocytotropic"
1.4563881043362685 "Enteral"
1.4563106796116505 "Succulents"
1.456140350877193 "bolshevism"
1.456 "Tathagatas"
1.455621301775148 "Tiong"
1.4546952224052718 "Historiador"
1.4545454545454546 "Dharmaraksa"
1.4532019704433496 "Magn"
1.4526315789473685 "Verhal"
1.4523809523809523 "Demoniacks"
1.4517709957674314 "GRAMMAR"
1.4503937007874015 "Petrochemical"
1.45016077170418 "officia"
1.45 "stercoraria"
1.45 "TRIBUNALS"
1.4496644295302012 "oldkyndighed"
1.4482758620689655 "Rappard"
1.448051948051948 "buffler"
1.4475138121546962 "Mesnil"
1.4470588235294117 "Axworthy"
1.4464285714285714 "Frankiel"

1.4464285714285714	"Greenfeld"
1.4462809917355373	"Brieg"
1.4462540716612378	"candesartan"
1.4461966604823748	"Moncton"
1.4461538461538461	"Sarig"
1.4461538461538461	"Roni"
1.4461538461538461	"warranta"
1.4454545454545455	"Koyane"
1.4453125	"Americanised"
1.4452054794520548	"noncontiguous"
1.4444444444444444	"Dodonov"
1.4444444444444444	"Shostak"
1.4444444444444444	"Gomberg"
1.4440959912735205	"Seismological"
1.4438502673796791	"Geologica"
1.443298969072165	"Groveton"
1.443254817987152	"Sphingolipid"
1.4430538172715894	"HMSO"
1.4430379746835442	"Aveiro"
1.4428571428571428	"Observes"
1.44279176201373	"SD"
1.4426229508196722	"MALES"
1.4423076923076923	"ACCELERATED"
1.4416243654822336	"ametropia"
1.4411764705882353	"Cavities"
1.4409722222222223	"Assayer"
1.4408945686900958	"HHI"
1.4406779661016949	"Sanga"
1.4403669724770642	"JUMPS"
1.4401041666666667	"Timaru"
1.4397163120567376	"Suppliant"
1.4390243902439024	"nutbrown"
1.4384615384615385	"bechance"
1.4383561643835616	"Hondureno"
1.4375	"Mansell"
1.4375	"CAUALLERO"
1.4375	"ZIFAR"
1.4374778604321643	"Th"
1.4372093023255814	"Sinusoidal"
1.4369747899159664	"vueil"
1.4366197183098592	"nonmelanocytic"
1.4363636363636363	"responsiones"
1.4363636363636363	"Augustino"
1.4362969752520622	"POSTAGE"
1.436105476673428	"benzol"
1.435897435897436	"hait"
1.4357798165137614	"Envelopes"
1.4356435643564356	"CANR"
1.435483870967742	"Nonpeptide"
1.435483870967742	"Gorlach"
1.4352119094637703	"unsigned"
1.4349442379182156	"Republished"
1.4342322032428683	"thankfully"
1.4342105263157894	"DIALECTIC"
1.4339622641509433	"Falconara"
1.4339622641509433	"Camarines"
1.4339622641509433	"Parapsychological"
1.4339622641509433	"Lundensia"
1.4339622641509433	"Archaeologica"
1.4337349397590362	"Morskogo"
1.4336283185840708	"MCCARTHY"
1.4328621908127208	"subduplicate"
1.4328358208955223	"balanoides"
1.4328358208955223	"Balanus"
1.4328358208955223	"Culham"
1.4323232323232322	"OSA"
1.4320388349514563	"dogstar"
1.4319526627218935	"cavernes"
1.4319444444444445	"eft"
1.4318181818181819	"HRAFlex"
1.4317825405122844	"Outlay"
1.430939226519337	"Brustein"
1.4309178743961353	"Wycliffite"
1.430622009569378	"Tiss"
1.4305555555555556	"pluye"
1.4305555555555556	"froidure"
1.4301470588235294	"Monogr"
1.4294975688816856	"Folder"
1.4289710208714588	"yes"

1.4285714285714286 "canso"
1.4285714285714286 "constantem"
1.4285714285714286 "sithen"
1.4285714285714286 "Disapprove"
1.4285714285714286 "Conical"
1.4285714285714286 "Pratyekabuddhas"
1.4282700421940928 "Virchows"
1.4275618374558303 "Pushpa"
1.4273504273504274 "farden"
1.4270531400966184 "Anal"
1.425531914893617 "Irrigators"
1.425531914893617 "cypionate"
1.4253731343283582 "JLG"
1.4253731343283582 "Forense"
1.425 "liberatio"
1.4245283018867925 "juridictions"
1.4239244491080798 "OUTLAY"
1.423841059602649 "slumber's"
1.423728813559322 "tribewn"
1.4230769230769231 "aith"
1.4230769230769231 "Cullian"
1.4230769230769231 "Kelleram"
1.4225806451612903 "pseudarthrosis"
1.422343324250681 "ICs"
1.4222972972972974 "achylia"
1.4222222222222223 "Pursuivants"
1.4220183486238531 "ALR"
1.4220183486238531 "Electrocatalysis"
1.4216867469879517 "Basicranium"
1.4216867469879517 "pilage"
1.4215686274509804 "Powerscourt"
1.421383647798742 "Dioxygen"
1.4210526315789473 "Christlich"
1.420807453416149 "Supracondylar"
1.420183486238532 "nontransitory"
1.4201497735980495 "LA"
1.42 "Colleer"
1.42 "Acholeplasma"
1.4195402298850575 "fascismo"
1.419431279620853 "caske"
1.4193548387096775 "Khok"
1.4193194291986828 "minim"
1.4191919191919191 "Burgher"
1.4190193164933136 "carryovers"
1.4189189189189189 "Peintures"
1.4188235294117648 "VICTORIA"
1.4188034188034189 "iDisk"
1.4183006535947713 "melanocytoma"
1.418200408997955 "Prematurity"
1.4181818181818182 "HARVEY'S"
1.4178921568627452 "NARA"
1.4173228346456692 "grossier"
1.4166666666666667 "postsubiculum"
1.4166666666666667 "Prets"
1.4166666666666667 "Escompte"
1.4166666666666667 "Klundert"
1.4161490683229814 "counterpositive"
1.4157303370786516 "gzhi"
1.415680473372781 "Serv"
1.4156626506024097 "Dermatomal"
1.4153846153846155 "rereward"
1.4152046783625731 "jostice"
1.4152046783625731 "plet"
1.4148936170212767 "locomotory"
1.4146341463414633 "ANCHORA"
1.4145077720207253 "CCPCC"
1.4138461538461538 "HUNTER"
1.4137931034482758 "Quasars"
1.4137931034482758 "Commonality"
1.4137205603160756 "Soc"
1.4131274131274132 "Econ"
1.4130434782608696 "grotches"
1.4130434782608696 "pallial"
1.4130434782608696 "voods"
1.4125 "Contempory"
1.4125 "Combattimento"
1.4118993135011442 "Occurs"
1.411764705882353 "Oogenesis"
1.4115755627009647 "gastrica"
1.4114852762851962 "Yes"

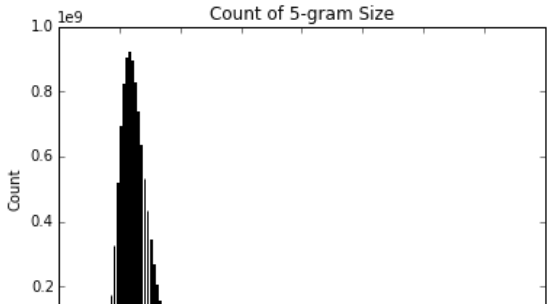
```
1.4113475177304964      "presentada"
1.4112903225806452      "WPK"
1.4111111111111112      "sectionized"
1.4109589041095891      "Deliberazioni"
1.4107275133881407      "Per"
1.4103269172013444      "Geelong"
1.4098360655737705      "BERE"
1.4098360655737705      "BAGHOT"
1.4098360655737705      "Neuvieme"
1.4098360655737705      "lnc"
1.4095744680851063      "plantagineum"
1.4094076655052266      "nonadverse"
1.4091176470588236      "Bourke"
1.4090909090909092      "Informazioni"
1.4090909090909092      "lookahead"
1.4090909090909092      "mul"
1.4088888888888889      "McGrade"
1.408856183836819       "interrogatory"
1.4086956521739131      "vocalics"
1.408450704225352       "Ance"
1.4083769633507854      "Biotelemetry"
1.4083094555873925      "Evol"
1.4081632653061225      "CFC's"
1.4081632653061225      "Lilawati"
1.4070524787388508      "Lecture"
1.4069767441860466      "unassembled"
1.4069592263206492      "Agreed"
1.4067796610169492      "Norraikow"
1.40625 "Kiukiang"
1.4062362758014932      "Assoc"
1.4059405940594059      "Beltwide"
1.4059405940594059      "MATILDE"
1.4057142857142857      "neurula"
1.4055555555555554      "organi"
1.4054054054054055      "JUMPED"
1.4054054054054055      "Schillebeeckx"
1.4051724137931034      "Inj"
1.4050632911392404      "Clennams"
1.4049586776859504      "Trenartha"
1.4048442906574394      "coparcener"
1.4038117165515478      "Cent"
1.4036697247706422      "Opt"
+ cd hw5
+ prog=hw5.3-3
+ hdfs=hdfs://master:9000
+ input=hdfs://master:9000/filtered-5grams
+ output=hdfs://master:9000/hw5.3-3-output
+ head -1000
+ hdfs dfs -cat 'hdfs://master:9000/hw5.3-3-output/*'
cat: Unable to write to output stream.
```

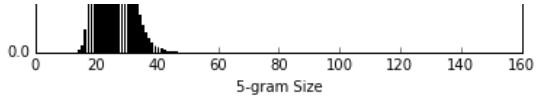
5-gram size distribution

In [1]:
```
%matplotlib inline
import matplotlib.pyplot as plt
from numpy import loadtxt

mat = loadtxt('Dist5Gram')
plt.bar(mat[:,0],mat[:,1],0.4)

plt.ylabel('Count')
plt.xlabel('5-gram Size')
plt.title('Count of 5-gram Size')
plt.show()
```





HW 5.4 (over 2Gig of Data)

In this part of the assignment we will focus on developing methods for detecting synonyms, using the Google 5-grams dataset. To accomplish this you must script two main tasks using MRJob:

1. Build stripes for the most frequent 10,000 words using cooccurence informationa based on the words ranked from 1001,-10,000 as a basis/vocabulary (drop stopword-like terms), and output to a file in your bucket on s3 (bigram analysis, though the words are non-contiguous).
2. Using two (symmetric) comparison methods of your choice (e.g., correlations, distances, similarities), pairwise compare all stripes (vectors), and output to a file in your bucket on s3.

Design notes for (1):
For this task you will be able to modify the pattern we used in HW 3.2 (feel free to use the solution as reference). To total the word counts across the 5-grams, output the support from the mappers using the total order inversion pattern:

```
<word,count>
```

to ensure that the support arrives before the cooccurrences.

In addition to ensuring the determination of the total word counts, the mapper must also output co-occurrence counts for the pairs of words inside of each 5-gram. Treat these words as a basket, as we have in HW 3, but count all stripes or pairs in both orders, i.e., count both orderings: `(word1,word2)`, and `(word2,word1)`, to preserve symmetry in our output for (2).

Design notes for (2):
For this task you will have to determine a method of comparison. Here are a few that you might consider:

- Jaccard
- Cosine similarity
- Spearman correlation
- Euclidean distance
- Taxicab (Manhattan) distance
- Shortest path graph distance (a graph, because our data is symmetric!)
- Pearson correlation
- Kendall correlation

However, be cautioned that some comparison methods are more difficult to parallelize than others, and do not perform more associations than is necessary, since your choice of association will be symmetric.

Please use the inverted index (discussed in live session #5) based pattern to compute the pairwise (term-by-term) similarity matrix.

In [229]:
```
%%writefile hw5.4-1.py
#!/usr/bin/env python

from collections import namedtuple, defaultdict, Counter
from mrjob.job import MRJob
from mrjob.step import MRStep
import sys

class HW54Job(MRJob):
    Row=namedtuple('Row',['ngram', 'count', 'pages_count', 'books_count'])
    @staticmethod
    def split_line(line):
        fields=line.strip().split('\t')
        return HW54Job.Row(fields[0],*[int(field) for field in fields[1:]])

    """Produce the word densities and sort them"""
    def coocurrence_init(self):
        # I produced the vocabulary in a prior step, so I don't actually need the
        # total counts for ranking/fitlering in the reducer.
        with open('hw5.4-terms-1k-10k.txt') as vocab_file:
            self.vocab=set(term.strip('\n"') for term in vocab_file)
    def coocurrence(self, _, line):
        """Mapper: split the 5-grams, and yield the coocurrence stripes for each."""
        row=HW54Job.split_line(line)
        # here I filter both co-occuring terms to be in rank 1k to rank 10k.
        # This corresponds to option "B" in the sync session
        terms=[term for term in row.ngram.split() if term in self.vocab]
```

```
        # Since terms dont CO-occur with themselves, I use slices to omit the
        # term under consideration from the inner loop(s)
        for idx,term in enumerate(terms):
            counts=defaultdict(int)
            for co in terms[:idx]:
                counts[co]+=row.count
            for co in terms[idx+1:]:
                counts[co]+=row.count
            if counts:
                yield term, [row.count,counts]
    def sum_counts(self, key, values):
        """Combiner: sum the counts as in usual word count"""
        counts=Counter()
        total=0
        for value in values:
            total+=value[0]
            counts.update(value[1])
        yield key, [total, counts]

    def steps(self):
        return [
            MRStep(
                mapper_init=self.coocurrence_init,
                mapper=self.coocurrence,
                combiner=self.sum_counts,
                reducer=self.sum_counts,
            ),
        ]

if __name__=='__main__':
    HW54Job().run()
    exit(0)
```

```
Overwriting hw5.4-1.py
```

In [230]:

```
%%bash
(cd ../prov; vagrant rsync master)
/usr/bin/ssh root@50.22.252.4 bash -xs <<'EOF'
cd hw5
freqs=hdfs://master:9000/hw5.3-3-output
hdfs dfs -cat ${freqs}/\* | cut -f2 | head -10000 | tail -9000 > hw5.4-terms-1k-10k.txt

prog=hw5.4-1
hdfs=hdfs://master:9000
input=${hdfs}/filtered-5grams
output=${hdfs}/${prog}-output
HADOOP_ROOT_LOGGER=INFO,console
hdfs dfs -rm -r -f ${output}
time ./${prog}.py -r hadoop \
  --strict-protocols --no-bootstrap-mrjob \
  --no-output --output ${output} \
  --file  hw5.4-terms-1k-10k.txt \
  --hadoop-arg -Dmapreduce.job.maps=56 \
  --hadoop-arg -Dmapreduce.job.reduces=56 \
  ${input}
hdfs dfs -cat ${output}/\* | head -100
EOF
```

```
==> master: Rsyncing folder: /media/sf_berkeley/w261/hw/hw5/hw5/ => /root/hw5
==> master: Rsyncing folder: /media/sf_berkeley/w261/hw/hw5/prov/ => /vagrant
Deleted hdfs://master:9000/hw5.4-1-output
"ADV"   [159, {"Router": 159, "Seq": 159}]
"Adjutant"      [137, {"Received": 137}]
"Agronomic"     [94, {"Chania": 94}]
"Ar"    [1944, {"aq": 47, "Tau": 408, "Cap": 448, "abs": 47, "Ar": 2270}]
"Arxiu" [174, {"Ciutat": 174}]
"Beverlacense"  [55, {"Dunelmense": 55, "Sanctuarium": 110}]
"Biosensors"    [190, {"Sensors": 190}]
"Bowen" [280, {"J": 280}]
"Chichimec"     [499, {"Ripples": 499}]
"Clinics"       [18195, {"Respiratory": 67, "Immunology": 546, "Nursing": 156, "Maxillofacial": 226,
"Multidisciplinary": 88, "Otolaryngologic": 162, "Allergy": 546, "Anesthesiology": 244, "Radiologic":
1607, "Gynecology": 447, "Radiological": 273, "Orthopedic": 302, "Pediatric": 12112, "Medicine": 1788,
"Surgery": 226, "Metabolic": 86, "Obstetrics": 447, "Mortality": 91}]
"Colliding"     [99, {"Instrumentation": 99, "Beam": 99}]
"Constitutio"   [53, {"Criminalis": 53}]
"Copenhague"    [1381, {"Cercle": 1381}]
"Croat" [930, {"Slovene": 930}]
"Derate"        [124, {"Dissipation": 124, "Total": 124}]
"Dewanee"       [86, {"Adawlut": 86, "Sudder": 86}]
```

```
"Diccionarios"  [67, {"Vocabularios": 67}]
"Easie" [2895, {"Readie": 2895}]
"Electrodiagnostic"     [688, {"Medicine": 688}]
"FEB"   [7771, {"MAR": 6008, "MAY": 6063, "JUN": 5929, "JAN": 1763, "APR": 5929, "OCT": 1708}]
"Finanze"       [3112, {"dello": 3112, "Amministrazione": 1768}]
"Fundamentales" [136, {"Agricultura": 136, "Tropical": 136}]
"Hannum"        [290, {"Hannum": 290}]
"Hematology"    [1332, {"Medicine": 112, "Internal": 160, "Pediatric": 571, "Oski's": 489}]
"Hortalez"      [106, {"Roderigue": 106}]
"Integr"        [1994, {"Regul": 1994, "J": 1994, "Am": 1994, "Physiol": 1994}]
"Jun"   [2893, {"Ocl": 212, "Aug": 2422, "Sep": 395, "Apr": 1492, "Jul": 1276, "Nov": 139, "Total": 7
8, "Oct": 1632}]
"Kissa" [99, {"Kursi": 99}]
"McDouall's"    [226, {"Journal": 226}]
"Microbiology"  [13395, {"Immunology": 10562, "Chromatography": 78, "Journal": 606, "Infections": 412,
"Microbiology": 116, "Automation": 352, "Cell": 222, "Clinical": 446, "Health": 245, "Medicine": 472,
"Toxicology": 208, "Experimental": 243}]
"Molar" [225, {"Molar": 450}]
"Mushtari"      [83, {"Parwin": 83}]
"Nadotti"       [70, {"Giuliana": 70}]
"Nephrology"    [261, {"Medicine": 65, "Urology": 65, "Nursing": 51, "Dialysis": 145}]
"Number"        [103244, {"Psychology": 65, "USA": 53, "Percent": 51833, "Acres": 212, "Internal": 50,
"Percentage": 16991, "IEEE": 100, "Phone": 159, "Pounds": 74, "df": 418, "Items": 68, "Vehicles": 63,
"Refractory": 102, "Amount": 2949, "Printing": 63, "Publications": 60, "Variance": 243, "Shops": 71,
"Male": 198, "Whole": 501, "Mean": 824, "Colleges": 106, "Code": 75, "sq": 98, "Testing": 156, "Numbe
r": 126690, "Rank": 1011, "Marital": 80, "Tons": 58, "Computer": 100, "Registration": 424, "Total": 81
29}]
"Oficial"       [207, {"Federacidn": 85, "Boletin": 122, "Ministerio": 122}]
"PIARE" [411, {"LALL": 525, "KUREEL": 411, "URF": 297, "TALIB": 297}]
"Pambansa"      [1454, {"Batasang": 1454}]
"Parishads"     [760, {"Samithis": 760}]
"Passengers"    [1488, {"Luggage": 1488}]
"Petraglia"     [106, {"MD": 106}]
"Pichler"       [225, {"Gasparikova": 225}]
"Pidgin"        [2567, {"Journal": 2567}]
"Pigmented"     [760, {"Pigmented": 256, "villonodular": 760, "synovitis": 461}]
"Prados"        [167, {"Escosura": 167}]
"Quantities"    [248, {"Materials": 91, "Abnormal": 157}]
"Ravishing"     [284, {"Lol": 284}]
"Referred"      [76, {"Notes": 76}]
"Registration"  [2466, {"Register": 568, "Forensic": 147, "Deeds": 116, "Number": 424, "Architects": 2
67, "Unification": 146, "Cavite": 124, "Registration": 148, "CAVITE": 451, "Dummies": 150}]
"SAI"   [928, {"NARHARI": 928, "PRASAD": 928, "SUKHDEO": 928, "SHRI": 928}]
"SIGMOD"        [1311, {"ACM": 1311}]
"STATUTES"      [1384, {"AFFECTING": 1384}]
"Sarasin"       [406, {"Rive": 406}]
"Sponsors"      [40, {"WOMEN'S": 40}]
"Statistisches" [75, {"Jahrbuchfur": 75}]
"Tentang"       [91, {"Dasar": 91, "Negara": 91}]
"Unnamable"     [135, {"Molloy": 135}]
"Upload"        [95, {"Notes": 95}]
"Vakil" [143, {"Brahmananda": 143}]
"Vaticani"      [196, {"Oecumenici": 196, "Sacrosancti": 196, "Synodalia": 196}]
"Vehicular"     [405, {"IEEE": 405, "Semiannual": 78}]
"Veterinary"    [21580, {"Journal": 8362, "Residues": 368, "Homoeopathic": 80, "Services": 621, "Clini
cal": 420, "Human": 315, "Medicine": 11276, "Toxicology": 93, "Notes": 231, "Experimental": 65}]
"anos"  [877, {"Asi": 463, "menos": 334, "pasen": 463, "umbral": 80}]
"cana"  [329, {"azucar": 329}]
"dissecans"     [635, {"Osteochondritis": 635}]
"geeigneten"    [102, {"Bekanntmachung": 102, "Verhandlungen": 102}]
"heredibus"     [1482, {"presentes": 394, "successoribus": 679, "assignatis": 409}]
"piccola"       [81, {"aree": 81}]
"revenu"        [113, {"impot": 113}]
"secretaria"    [123, {"Memoria": 123}]
"senderos"      [691, {"bifurcan": 691}]
"septimanas"    [158, {"septimanis": 158}]
"shutoff"       [250, {"OFF": 250}]
"soberana"      [124, {"decretos": 124}]
"stratiform"    [85, {"stratabound": 85}]
"tratare"       [323, {"reina": 323}]
"Anat"  [678, {"Klin": 317, "Anat": 294, "Abt": 120, "Physiol": 504}]
"Audits"        [686, {"Accounts": 321, "Health": 365}]
"Brachionus"    [71, {"calyciflorus": 71, "rotifer": 71}]
"Cap"   [4585, {"Tropical": 59, "Cap": 6188, "Ar": 448}]
"Chihuahuense"  [367, {"Sociedad": 367}]
"Christoph"     [424, {"Hardin": 424}]
"Courtesy"      [3494, {"Reproduced": 694, "Hagley": 478, "Journal": 936, "Register": 107, "Illuminati
ng": 95, "Health": 255, "Services": 571, "Starrett": 223, "Bettmann": 135}]
"Datagrams"     [84, {"Received": 84}]
"Ditto" [138, {"Ditto": 138}]
```

```
"Ditto" [138, {"Ditto": 138}]
"ENERGY"        [10371, {"MINISTER": 2594, "MINISTRY": 7777}]
"Equalization"  [168, {"Opportunity": 168}]
"Forening"      [321, {"Geologisk": 225, "Dansk": 225, "Meddelelser": 321}]
"Gayre" [234, {"Gayre": 234}]
"Genitourinary" [192, {"Medicine": 134, "Prostatic": 58}]
"Hiking"        [244, {"Hiking": 244}]
"Hydroids"      [144, {"Gymnoblastic": 144, "Tubularian": 144}]
"Intraocular"   [684, {"J": 684, "Am": 510}]
"J"     [119715, {"Tse": 125, "Plastic": 81, "Collected": 1035, "Trans": 1792, "Reproduced": 254, "Sym
p": 613, "Mol": 14809, "Pfliigers": 110, "VAN": 357, "Notes": 1069, "Elem": 191, "GA": 44, "Dis": 262
0, "Embryo": 130, "Reconstr": 391, "Crii": 138, "Raffel": 111, "Pharm": 114, "Crit": 374, "Sport": 55
6, "Occup": 1435, "Commun": 1681, "Radiation": 1057, "Lond": 51, "Sports": 294, "Biomater": 404, "App
l": 2962, "Biomed": 197, "Pert": 130, "Afr": 262, "Infect": 367, "Pub": 154, "Ion": 364, "Abuse": 237
8, "Mich": 170, "Pediatr": 254, "Gynaecol": 2094, "Integr": 1994, "Exp": 622, "Biotechnol": 264, "iner
tia": 1269, "Chir": 68, "Radium": 5509, "Hufbauer": 2822, "Plast": 356, "Physiol": 12095, "Clin": 462
2, "Ophthalmol": 254, "Scand": 5411, "Cell": 15956, "Circ": 3808, "Kerkvliet": 109, "Internal": 72, "H
ealth": 7207, "Rad": 158, "Eur": 7155, "Roentgenol": 5509, "Fr": 68, "Regul": 1994, "Bowen": 280, "N
Z": 114, "Pract": 195, "Inst": 205, "Amer": 141, "Symptom": 114, "AFFAIRS": 82, "Sourcebook": 65, "Mic
robiol": 1163, "J": 3748, "CG": 42, "Emp": 1404, "Introduced": 51, "Int": 13024, "Motil": 956, "McDona
ld": 43, "Lung": 2133, "KV": 111, "Intraocular": 684, "Allergy": 161, "Ment": 1257, "Tria": 109, "Indu
st": 105, "Nucl": 8034, "Subchapter": 72, "MINISTER": 44, "IEEE": 44, "Anim": 137, "Chromatography": 6
2, "Endocrinol": 154, "Syst": 1006, "Pollut": 382, "Ther": 8400, "Gastrointest": 2018, "Forensic": 234
3, "MINES": 278, "Ortho": 145, "Environ": 5395, "Behav": 256, "Radial": 243, "Pelvic": 1270, "Disord":
82, "Am": 59681, "Cormier": 368, "Immunol": 452}]
"Jing"  [394, {"Nong": 269, "Su": 125}]
"Kadish"        [103, {"Ruoff": 103}]
"Landsmal"      [68, {"Folkliv": 68, "Svenska": 68}]
"Lefkandi"      [87, {"Toumba": 87}]
"Lokalisation"  [420, {"histologischen": 420, "Grosshirnrinde": 209}]
"ME"    [4578, {"ME": 15240, "VA": 97, "MD": 193, "EDWARDS": 255, "YS": 46, "Reprinted": 102}]
+ cd hw5
+ freqs=hdfs://master:9000/hw5.3-3-output
+ hdfs dfs -cat 'hdfs://master:9000/hw5.3-3-output/*'
+ cut -f2
+ head -10000
+ tail -9000
cat: Unable to write to output stream.
+ prog=hw5.4-1
+ hdfs=hdfs://master:9000
+ input=hdfs://master:9000/filtered-5grams
+ output=hdfs://master:9000/hw5.4-1-output
+ HADOOP_ROOT_LOGGER=INFO,console
+ hdfs dfs -rm -r -f hdfs://master:9000/hw5.4-1-output
16/02/18 00:16:05 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 0 minu
tes, Emptier interval = 0 minutes.
+ ./hw5.4-1.py -r hadoop --strict-protocols --no-bootstrap-mrjob --no-output --output hdfs://master:90
00/hw5.4-1-output --file hw5.4-terms-1k-10k.txt --hadoop-arg -Dmapreduce.job.maps=56 --hadoop-arg -Dma
preduce.job.reduces=56 hdfs://master:9000/filtered-5grams
no configs found; falling back on auto-configuration
no configs found; falling back on auto-configuration
Using Hadoop version 2.7.2
Copying local files into hdfs:///user/root/tmp/mrjob/hw5.root.20160218.061606.405897/files/
HADOOP: packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/s
treamjob7845335100281926203.jar tmpDir=null
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Connecting to ResourceManager at master/10.108.114.214:8032
HADOOP: Total input paths to process : 190
HADOOP: number of splits:190
HADOOP: Submitting tokens for job: job_1455677115242_0055
HADOOP: Submitted application application_1455677115242_0055
HADOOP: The url to track the job: http://master:8088/proxy/application_1455677115242_0055/
HADOOP: Running job: job_1455677115242_0055
HADOOP: Job job_1455677115242_0055 running in uber mode : false
HADOOP:  map 0% reduce 0%
HADOOP:  map 2% reduce 0%
HADOOP:  map 5% reduce 0%
HADOOP:  map 11% reduce 0%
HADOOP:  map 17% reduce 0%
HADOOP:  map 18% reduce 0%
HADOOP:  map 19% reduce 0%
HADOOP:  map 20% reduce 0%
HADOOP:  map 22% reduce 0%
HADOOP:  map 24% reduce 0%
HADOOP:  map 25% reduce 0%
HADOOP:  map 29% reduce 0%
HADOOP:  map 31% reduce 0%
HADOOP:  map 34% reduce 0%
HADOOP:  map 36% reduce 0%
HADOOP:  map 39% reduce 0%
```

```
HADOOP:  map 39% reduce 1%
HADOOP:  map 40% reduce 1%
HADOOP:  map 41% reduce 1%
HADOOP:  map 42% reduce 1%
HADOOP:  map 43% reduce 1%
HADOOP:  map 46% reduce 2%
HADOOP:  map 46% reduce 3%
HADOOP:  map 48% reduce 4%
HADOOP:  map 49% reduce 4%
HADOOP:  map 51% reduce 4%
HADOOP:  map 52% reduce 5%
HADOOP:  map 55% reduce 5%
HADOOP:  map 56% reduce 6%
HADOOP:  map 57% reduce 6%
HADOOP:  map 58% reduce 7%
HADOOP:  map 59% reduce 8%
HADOOP:  map 60% reduce 8%
HADOOP:  map 61% reduce 8%
HADOOP:  map 62% reduce 8%
HADOOP:  map 65% reduce 9%
HADOOP:  map 66% reduce 9%
HADOOP:  map 67% reduce 10%
HADOOP:  map 69% reduce 10%
HADOOP:  map 70% reduce 10%
HADOOP:  map 71% reduce 10%
HADOOP:  map 72% reduce 10%
HADOOP:  map 73% reduce 11%
HADOOP:  map 74% reduce 11%
HADOOP:  map 74% reduce 12%
HADOOP:  map 75% reduce 12%
HADOOP:  map 76% reduce 12%
HADOOP:  map 78% reduce 12%
HADOOP:  map 79% reduce 12%
HADOOP:  map 81% reduce 12%
HADOOP:  map 82% reduce 12%
HADOOP:  map 84% reduce 12%
HADOOP:  map 85% reduce 12%
HADOOP:  map 86% reduce 13%
HADOOP:  map 87% reduce 13%
HADOOP:  map 88% reduce 13%
HADOOP:  map 88% reduce 14%
HADOOP:  map 89% reduce 14%
HADOOP:  map 91% reduce 14%
HADOOP:  map 93% reduce 14%
HADOOP:  map 94% reduce 14%
HADOOP:  map 95% reduce 14%
HADOOP:  map 97% reduce 14%
HADOOP:  map 97% reduce 15%
HADOOP:  map 98% reduce 15%
HADOOP:  map 99% reduce 15%
HADOOP:  map 99% reduce 16%
HADOOP:  map 100% reduce 18%
HADOOP:  map 100% reduce 19%
HADOOP:  map 100% reduce 34%
HADOOP:  map 100% reduce 86%
HADOOP:  map 100% reduce 93%
HADOOP:  map 100% reduce 98%
HADOOP:  map 100% reduce 100%
HADOOP: Job job_1455677115242_0055 completed successfully
HADOOP: Counters: 53
HADOOP:         File System Counters
HADOOP:                 FILE: Number of bytes read=681099
HADOOP:                 FILE: Number of bytes written=31550684
HADOOP:                 FILE: Number of read operations=0
HADOOP:                 FILE: Number of large read operations=0
HADOOP:                 FILE: Number of write operations=0
HADOOP:                 HDFS: Number of bytes read=2156095226
HADOOP:                 HDFS: Number of bytes written=226669
HADOOP:                 HDFS: Number of read operations=738
HADOOP:                 HDFS: Number of large read operations=0
HADOOP:                 HDFS: Number of write operations=112
HADOOP:         Job Counters
HADOOP:                 Killed map tasks=1
HADOOP:                 Killed reduce tasks=1
HADOOP:                 Launched map tasks=190
HADOOP:                 Launched reduce tasks=56
HADOOP:                 Other local map tasks=2
HADOOP:                 Data-local map tasks=187
HADOOP:                 Rack-local map tasks=1
```

```
HADOOP:                 Total time spent by all maps in occupied slots (ms)=2996125
HADOOP:                 Total time spent by all reduces in occupied slots (ms)=1889011
HADOOP:                 Total time spent by all map tasks (ms)=2996125
HADOOP:                 Total time spent by all reduce tasks (ms)=1889011
HADOOP:                 Total vcore-milliseconds taken by all map tasks=2996125
HADOOP:                 Total vcore-milliseconds taken by all reduce tasks=1889011
HADOOP:                 Total megabyte-milliseconds taken by all map tasks=3068032000
HADOOP:                 Total megabyte-milliseconds taken by all reduce tasks=1934347264
HADOOP:         Map-Reduce Framework
HADOOP:                 Map input records=58682266
HADOOP:                 Map output records=22921
HADOOP:                 Map output bytes=840755
HADOOP:                 Map output materialized bytes=744603
HADOOP:                 Input split bytes=26110
HADOOP:                 Combine input records=22921
HADOOP:                 Combine output records=15415
HADOOP:                 Reduce input groups=3557
HADOOP:                 Reduce shuffle bytes=744603
HADOOP:                 Reduce input records=15415
HADOOP:                 Reduce output records=3557
HADOOP:                 Spilled Records=30830
HADOOP:                 Shuffled Maps =10640
HADOOP:                 Failed Shuffles=0
HADOOP:                 Merged Map outputs=10640
HADOOP:                 GC time elapsed (ms)=38502
HADOOP:                 CPU time spent (ms)=657200
HADOOP:                 Physical memory (bytes) snapshot=60454170624
HADOOP:                 Virtual memory (bytes) snapshot=523736563712
HADOOP:                 Total committed heap usage (bytes)=45622493184
HADOOP:         Shuffle Errors
HADOOP:                 BAD_ID=0
HADOOP:                 CONNECTION=0
HADOOP:                 IO_ERROR=0
HADOOP:                 WRONG_LENGTH=0
HADOOP:                 WRONG_MAP=0
HADOOP:                 WRONG_REDUCE=0
HADOOP:         File Input Format Counters
HADOOP:                 Bytes Read=2156069116
HADOOP:         File Output Format Counters
HADOOP:                 Bytes Written=226669
HADOOP: Output directory: hdfs://master:9000/hw5.4-1-output
Counters from step 1:
  (no counters found)
deleting hdfs:///user/root/tmp/mrjob/hw5.root.20160218.061606.405897 from HDFS

real    1m58.853s
user    0m22.961s
sys     0m1.442s
+ hdfs dfs -cat 'hdfs://master:9000/hw5.4-1-output/*'
+ head -100
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
```

```
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
cat: Unable to write to output stream.
```

In [ ]:

```
%%writefile hw5.4-2.py
#!/usr/bin/env python

from collections import defaultdict, Counter
from mrjob.job import MRJob
from mrjob.step import MRStep
from mrjob.protocol import JSONProtocol
import sys

class HW54Job(MRJob):
    """Compare words using coocurrences"""
    def coocurrence_init(self):
        self.seen=set()
    def coocurrence(self, key, stripe):
        """Mapper: yield each member of the stripe"""
        for co in stripe[1]:
            # if we have seen this word as a key already,
            #
            if co not in self.seen:
                yield co, [key, stripe[1]]
        yield ' ', [key, stripe[1]]
        self.seen.add(key)

    def sum_counts(self, key, values):
        """Combiner: sum the counts as in usual word count"""
        counts=Counter()
        total=0
        for value in values:
            total+=value[0]
            counts.update(value[1])
        yield key, [total, counts]

    def steps(self):
        return [
            MRStep(
                mapper_init=self.coocurrence_init,
                mapper=self.coocurrence,
                combiner=self.sum_counts,
                reducer=self.sum_counts,
            ),
        ]

if __name__=='__main__':
    HW54Job().run()
    exit(0)
```

In [ ]:

```
%%bash
(cd ../prov; vagrant rsync master)
/usr/bin/ssh root@50.22.252.4 bash -xs <<'EOF'
cd hw5
prog=hw5.4-2
hdfs=hdfs://master:9000
input=${hdfs}/hw5.4-1-output
```

```
output=${hdfs}/${prog}-output
HADOOP_ROOT_LOGGER=INFO,console
hdfs dfs -rm -r -f ${output}
time ./${prog}.py -r hadoop \
  --strict-protocols --no-bootstrap-mrjob \
  --no-output --output ${output} \
  --file  hw5.4-terms-1k-10k.txt \
  --hadoop-arg -Dmapreduce.job.maps=56 \
  --hadoop-arg -Dmapreduce.job.reduces=56 \
  ${input}
hdfs dfs -cat ${output}/\* | head -100
EOF
```

HW 5.5

In this part of the assignment you will evaluate the success of you synonym detector. Take the top 1,000 closest/most similar/correlative pairs of words as determined by your measure in (2), and use the synonyms function in the accompanying python code:

nltk_synonyms.py

Note: This will require installing the python nltk package:

http://www.nltk.org/install.html (http://www.nltk.org/install.html)

and downloading its data with `nltk.download()`.

For each `(word1,word2)` pair,

- check to see if `word1` is in the list, `synonyms(word2)`, and vice-versa.
- If one of the two is a synonym of the other, then consider this pair a 'hit', and then report the precision, recall, and F1 measure of your detector across your 1,000 best guesses.
- Report the macro averages of these measures.

In [ ]:
```
#!/usr/bin/python2.7
''' pass a string to this funciton ( eg 'car') and it will give you a list of
words which is related to cat, called lemma of CAT. '''
import nltk
from nltk.corpus import wordnet as wn
import sys
import numpy as np
#print all the synset element of an element
def synonyms(string):
    syndict = {}
    for i,j in enumerate(wn.synsets(string)):
        syns = j.lemma_names()
        for syn in syns:
            syndict.setdefault(syn,1)
    return syndict.keys()
```

In [ ]:
```
text_file = open("HW5_4_results.txt", "r")
lines = text_file.readline().split('\r')
hits = []

for line in lines:
    pair = line.split('\t')[0].split('__')
    if len(pair) == 2:
        hits.append(pair[0] in synonyms(pair[1]) or pair[1] in synonyms(pair[0]))
print 'Precision of detection: %.4f' %(sum(hits[:1000])/1000.0)
#Precision of detection: 0.0050
```

HW 5.5.1 (optional)

There is also a corpus of stopwords, that is, high-frequency words like "the", "to" and "also" that we sometimes want to filter out of a document before further processing. Stopwords usually have little lexical content, and their presence in a text fails to distinguish it from other texts. Python's nltk comes with a prebuilt list of stopwords (see below). Using this stopword list filter out these tokens from your analysis and rerun the experiments in 5.5 and disucuss the results of using a stopword list and without using a stopword list.

```
>> from nltk.corpus import stopwords
>>> stopwords.words('english')
```

```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours',
'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers',
'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves',
'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are',
'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does',
'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until',
'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into',
'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down',
'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here',
'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more',
'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so',
'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'now']
```

HW 5.6 (optional)

There are many good ways to build our synonym detectors, so for optional homework, measure co-occurrence by (left/right/all) consecutive words only, or make stripes according to word co-occurrences with the accompanying 2-, 3-, or 4-grams (note here that your output will no longer be interpretable as a network) inside of the 5-grams.