MIDS-W261-HW-05-PHASE2-MASTER

June 22, 2017

1 MIDS - w261 Machine Learning At Scale

Course Lead: Dr James G. Shanahan (**email** Jimi via James.Shanahan *AT* gmail.com)

1.1 Assignment - HW5

Name1: Carlos Castro
StudentID1: 3032134613
Class1: MIDS w261 (Section 2, e.g., Fall 2016 Group 1)
Email1: carlosscastro@iSchool.Berkeley.edu
Name2: Chuqing He
StudentID2: 3032132858
Class1: MIDS w261 (Section 2, e.g., Fall 2016 Group 1)
Email1: chqngh@ischool.berkeley.edu
Name3: Nikki Haas
StudentID3: 3032161965
Class3: MIDS w261 (Section 2, e.g., Fall 2016 Group 1)
Email3: nhaas@ischool.berkeley.edu
Week: 5 **Due Time:** 2 Phases.

- **HW5 Phase 1** This can be done on a local machine (with a unit test on the cloud such as AltaScale's PaaS or on AWS) and is due Tuesday, Week 6 by 8AM (West coast time). It will primarily focus on building a unit/systems and for pairwise similarity calculations pipeline (for stripe documents)
- **HW5 Phase 2** This will require the AltaScale cluster and will be due Tuesday, Week 7 by 8AM (West coast time). The focus of HW5 Phase 2 will be to scale up the unit/systems tests to the Google 5 gram corpus. This will be a group exercise

```
In [1]: import sys
        sys.version
```

```
Out[1]: '2.7.13 | packaged by conda-forge | (default, May  2 2017, 12:48:11) \n[GCC
```

```
In [8]: pyArchive = 'hdfs:///user/nhaas/virtualenv/my_env.zip#my_env'
```

2 Table of Contents

1 Instructions Back to Table of Contents

MIDS UC Berkeley, Machine Learning at Scale
DATSCIW261 ASSIGNMENT #5
Version 2017-9-2

2.0.1 IMPORTANT

This homework must be completed in the cloud

2.0.2 === INSTRUCTIONS for SUBMISSIONS ===

Follow the instructions for submissions carefully.

Each student has a `HW-<user>` repository for all assignments.

Click this link to enable you to create a github repo within the MIDS261 Classroom: https://classroom.github.com/assignment-invitations/3b1d6c8e58351209f9dd865537111ff8 and follow the instructions to create a HW repo.

Push the following to your HW github repo into the master branch: * Your local HW5 directory. Your repo file structure should look like this:

```
HW-<user>
    --HW3
       |__MIDS-W261-HW-03-<Student_id>.ipynb
       |__MIDS-W261-HW-03-<Student_id>.pdf
       |__some other hw3 file
    --HW4
       |__MIDS-W261-HW-04-<Student_id>.ipynb
       |__MIDS-W261-HW-04-<Student_id>.pdf
       |__some other hw4 file
    etc..
```

2 Useful References Back to Table of Contents

- See async and live lectures for this week

3 HW Problems Back to Table of Contents

PHASE 2 ———-

3 HW 5.4

3.1 Full-scale experiment on Google N-gram data on the CLOUD

__ Once you are happy with your test results __ proceed to generating your results on the Google n-grams dataset.

3.2 3. HW5.4.0 Run systems tests on the CLOUD (PHASE 2)

Back to Table of Contents

Repeat HW5.3.0 (using the same small data sources that were used in HW5.3.0) on ** the cloud** (e.g., AltaScale / AWS/ SoftLayer/ Azure). Make sure all tests give correct results! Good luck out there!

3.3 MapReduce Classes

In [1]:
```
%%writefile buildStripes.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-

from __future__ import division
import re
import mrjob
import json
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import itertools
import collections
import logging
import time



class MRbuildStripes(MRJob):

    #START SUDENT CODE531_STRIPES

    MRJob.SORT_VALUES = True

    #def mapper_init(self):
    #    return self.start_time = time.time()

    def mapper(self, _, line):
        splits = line.rstrip("\n").split("\t")

        words = splits[0].lower().split()
        count = splits[1]
```

```
        H = {}
        for subset in itertools.combinations(sorted(set(words)), 2):

            # Process combinations in sorted order, i.e. "hello","tomorrow'
            if subset[0] not in H.keys():
                H[subset[0]] = {}
                H[subset[0]][subset[1]] = count
            elif subset[1] not in H[subset[0]]:
                H[subset[0]][subset[1]] = count
            else:
                H[subset[0]][subset[1]] += count

            # Obtain combinations in reverse order, to consider them both v
            # TODO: Should refactor this and the block above, shameless cop
            if subset[1] not in H.keys():
                H[subset[1]] = {}
                H[subset[1]][subset[0]] = count
            elif subset[0] not in H[subset[1]]:
                H[subset[1]][subset[0]] = count
            else:
                H[subset[1]][subset[0]] += count
        for key in H.keys():
            #print "%s\t%s" % (key, json.dumps(H[key]))
            yield key, H[key]

    def reducer(self, key, values):

        counter = {}

        for value in values:

            for k, v in value.iteritems():
                if k in counter:
                    counter[k] += int(v)
                else:
                    counter[k] = int(v)

        yield key, counter


    def steps(self):
        return [

            MRStep(#mapper_init=self.mapper_init
                   #,
                   mapper=self.mapper
                   ,
```

```
                            reducer=self.reducer
                        )
                    ]
          #END SUDENT CODE531_STRIPES

        if __name__ == '__main__':
            start_time = time.time()
            MRbuildStripes.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)
```

```
Overwriting buildStripes.py
```

In [2]:
```
%%writefile invertedIndex.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-


from __future__ import division
import collections
import re
import json
import math
import numpy as np
import itertools
import mrjob
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import json
import logging
import time

class MRinvertedIndex(MRJob):

    #START SUDENT CODE531_INV_INDEX

    def mapper(self, _, line):
        key, stripeJson = line.strip().split('\t')
        key = key.strip("\"")
        stripe = json.loads(stripeJson)

        for k, v in stripe.iteritems():
            yield k, [key, len(stripe)]
```

```
            def reducer(self, key, values):

                table = {}
                for value in values:
                    table[value[0]] = value[1]

                yield key, table

            #END SUDENT CODE531_INV_INDEX

        if __name__ == '__main__':
            start_time = time.time()
            MRinvertedIndex.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting invertedIndex.py


In [3]: %%writefile similarity.py
        #!~/anaconda2/bin/python
        # -*- coding: utf-8 -*-

        from __future__ import division
        import collections
        import re
        import json
        import math
        #import numpy as np
        import itertools
        import mrjob
        from mrjob.protocol import RawProtocol
        from mrjob.job import MRJob
        from mrjob.step import MRStep
        import time
        import logging

        class MRsimilarity(MRJob):

            #START SUDENT CODE531_SIMILARITY

            MRJob.SORT_VALUES = True

            def mapper(self, _, line):
```

```
            key, valuesJson = line.strip().split('\t')
            key = key.strip("\"")
            values = json.loads(valuesJson)

            for pair in itertools.combinations(sorted(set(values)), 2):
                yield pair, [values[pair[0]], values[pair[1]]]

        def reducer(self, key, values):
            intersection = 0
            count1 = None
            count2 = None

            cosine = 0.0

            # Iterate through the values
            for value in values:
                # Jaccard, get counts for the intersection, and for each set
                intersection += 1
                if count1 == None:
                    count1 = value[0]
                    count2 = value[1]

                # Cosine
                a = 1 / math.sqrt(value[0])
                b = 1 / math.sqrt(value[1])
                cosine += a * b

            jaccard = float(intersection) / float(count1 + count2 - intersectio

            overlap_coefficient = float(intersection) / min(count1, count2)

            dice_coefficient = float(2 * intersection) / (count1 + count2)

            average = (cosine + jaccard + overlap_coefficient + dice_coefficien

            yield average, [key[0] + ' - ' + key[1], cosine, jaccard, overlap_c

        #END SUDENT CODE531_SIMILARITY

    if __name__ == '__main__':
        start_time = time.time()
        MRsimilarity.run()
        elapsed_time = time.time() - start_time
        mins = elapsed_time/float(60)
        a = """Elapsed time: %s seconds
        In minutes: %s mins""" % (str(elapsed_time), str(mins))
        logging.warning(a)
```

```
Overwriting similarity.py
```

3.4 Mini-test data

```
In [4]: %%writefile googlebooks-eng-all-5gram-20090715-0-filtered-first-10-lines.tx
        A BILL FOR ESTABLISHING RELIGIOUS        59        59        54
        A Biography of General George        92        90        74
        A Case Study in Government        102        102        78
        A Case Study of Female        447        447        327
        A Case Study of Limited        55        55        43
        A Child's Christmas in Wales        1099        1061        866
        A Circumstantial Narrative of the        62        62        50
        A City by the Sea        62        60        49
        A Collection of Fairy Tales        123        117        80
        A Collection of Forms of        116        103        82

Overwriting googlebooks-eng-all-5gram-20090715-0-filtered-first-10-lines.txt


In [5]: %%writefile atlas-boon-systems-test.txt
        atlas boon        50        50        50
        boon cava dipped        10        10        10
        atlas dipped        15        15        15

Overwriting atlas-boon-systems-test.txt
```

3.5 Build stripes for mini-test data

```
In [7]: ##############################################################################
        # Make Stripes from ngrams for systems test 1
        ##############################################################################

        !hdfs dfs -rm -r systems_test_stripes_1
        !python buildStripes.py -r hadoop googlebooks-eng-all-5gram-20090715-0-filt
             --archive={pyArchive} \
            > systems_test_stripes_1

rm: `systems_test_stripes_1': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.chqngh.20170621.173926.821163
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
```

```
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1189
  Submitted application application_1497906899862_1189
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1189
  Job job_1497906899862_1189 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1189 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170621.1739
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
                Bytes Written=2406
        File System Counters
                FILE: Number of bytes read=1098
                FILE: Number of bytes written=401077
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1011
                HDFS: Number of bytes written=2406
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=18462720
                Total megabyte-milliseconds taken by all reduce tasks=11253760
```

```
                Total time spent by all map tasks (ms)=12020
                Total time spent by all maps in occupied slots (ms)=36060
                Total time spent by all reduce tasks (ms)=4396
                Total time spent by all reduces in occupied slots (ms)=21980
                Total vcore-milliseconds taken by all map tasks=12020
                Total vcore-milliseconds taken by all reduce tasks=4396
        Map-Reduce Framework
                CPU time spent (ms)=2810
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=82
                Input split bytes=448
                Map input records=10
                Map output bytes=3359
                Map output materialized bytes=1076
                Map output records=49
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1895075840
                Reduce input groups=49
                Reduce input records=49
                Reduce output records=28
                Reduce shuffle bytes=1076
                Shuffled Maps =2
                Spilled Records=98
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7731109888
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.201
Removing temp directory /tmp/buildStripes.chqngh.20170621.173926.821163...
WARNING:root:Elapsed time: 61.2851359844 seconds
    In minutes: 1.02141893307 mins
```

3.5.1 Stripes on Mini-Test Data Statistics:

Time

- *Run time: 61.29 seconds*
- *Run time: 1.02 minutes*

Input/Output statistics

- *Bytes Read: 563*
- *Bytes Written: 2046*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2810*

```
In [8]: !cat systems_test_stripes_1
```

```
"a"        {"limited": 55, "female": 447, "general": 92, "sea": 62, "in": 1201, "re
"bill"       {"a": 59, "religious": 59, "for": 59, "establishing": 59}
"biography"        {"a": 92, "of": 92, "george": 92, "general": 92}
"by"       {"a": 62, "city": 62, "the": 62, "sea": 62}
"case"       {"a": 604, "limited": 55, "government": 102, "of": 502, "study": 604,
"child's"        {"a": 1099, "wales": 1099, "christmas": 1099, "in": 1099}
"christmas"        {"a": 1099, "wales": 1099, "in": 1099, "child's": 1099}
"circumstantial"        {"a": 62, "of": 62, "the": 62, "narrative": 62}
"city"       {"a": 62, "the": 62, "by": 62, "sea": 62}
"collection"        {"a": 239, "forms": 116, "fairy": 123, "tales": 123, "of": 239}
"establishing"        {"a": 59, "bill": 59, "religious": 59, "for": 59}
"fairy"       {"a": 123, "of": 123, "tales": 123, "collection": 123}
"female"        {"a": 447, "case": 447, "study": 447, "of": 447}
"for"       {"a": 59, "bill": 59, "religious": 59, "establishing": 59}
"forms"       {"a": 116, "of": 116, "collection": 116}
"general"       {"a": 92, "of": 92, "george": 92, "biography": 92}
"george"       {"a": 92, "of": 92, "biography": 92, "general": 92}
"government"        {"a": 102, "case": 102, "study": 102, "in": 102}
"in"       {"a": 1201, "case": 102, "government": 102, "study": 102, "child's": 10
"limited"        {"a": 55, "case": 55, "study": 55, "of": 55}
"narrative"        {"a": 62, "of": 62, "the": 62, "circumstantial": 62}
"of"       {"a": 895, "case": 502, "circumstantial": 62, "george": 92, "limited":
"religious"        {"a": 59, "bill": 59, "for": 59, "establishing": 59}
"sea"       {"a": 62, "city": 62, "the": 62, "by": 62}
"study"       {"a": 604, "case": 604, "limited": 55, "government": 102, "of": 502,
"tales"       {"a": 123, "of": 123, "fairy": 123, "collection": 123}
"the"       {"a": 124, "city": 62, "circumstantial": 62, "of": 62, "sea": 62, "nar
"wales"       {"a": 1099, "in": 1099, "christmas": 1099, "child's": 1099}
```

```
In [9]: ##############################################################################
        # Make Stripes from ngrams for systems test 2
        ##############################################################################
```

```
        !hdfs dfs -rm -r systems_test_stripes_2
        !python buildStripes.py -r hadoop atlas-boon-systems-test.txt  \
             --archive={pyArchive} \
            > systems_test_stripes_2

rm: `systems_test_stripes_2': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.chqngh.20170621.174030.944052
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1192
  Submitted application application_1497906899862_1192
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1192
  Job job_1497906899862_1192 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1192 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170621.1740
Counters: 49
        File Input Format Counters
                Bytes Read=101
        File Output Format Counters
                Bytes Written=163
        File System Counters
                FILE: Number of bytes read=138
```

```
        FILE: Number of bytes written=399103
        FILE: Number of large read operations=0
        FILE: Number of read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=475
        HDFS: Number of bytes written=163
        HDFS: Number of large read operations=0
        HDFS: Number of read operations=9
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Rack-local map tasks=2
        Total megabyte-milliseconds taken by all map tasks=10801152
        Total megabyte-milliseconds taken by all reduce tasks=14988800
        Total time spent by all map tasks (ms)=7032
        Total time spent by all maps in occupied slots (ms)=21096
        Total time spent by all reduce tasks (ms)=5855
        Total time spent by all reduces in occupied slots (ms)=29275
        Total vcore-milliseconds taken by all map tasks=7032
        Total vcore-milliseconds taken by all reduce tasks=5855
Map-Reduce Framework
        CPU time spent (ms)=2490
        Combine input records=0
        Combine output records=0
        Failed Shuffles=0
        GC time elapsed (ms)=96
        Input split bytes=374
        Map input records=3
        Map output bytes=217
        Map output materialized bytes=173
        Map output records=7
        Merged Map outputs=2
        Physical memory (bytes) snapshot=1902125056
        Reduce input groups=7
        Reduce input records=7
        Reduce output records=4
        Reduce shuffle bytes=173
        Shuffled Maps =2
        Spilled Records=14
        Total committed heap usage (bytes)=5218762752
        Virtual memory (bytes) snapshot=7732584448
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
```

```
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/buildStripes.chqngh.2017
Removing temp directory /tmp/buildStripes.chqngh.20170621.174030.944052...
WARNING:root:Elapsed time: 58.2388238907 seconds
    In minutes: 0.970647064845 mins
```

3.5.2 Stripes on System Test 2 Statistics:

Time

- *Run time: 58.24 seconds*
- *Run time: 0.97 minutes*

Input/Output statistics

- *Bytes Read: 101*
- *Bytes Written: 163*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2490*

```
In [10]: !cat systems_test_stripes_2
```

```
"atlas"         {"dipped": 15, "boon": 50}
"boon"         {"atlas": 50, "dipped": 10, "cava": 10}
"cava"         {"dipped": 10, "boon": 10}
"dipped"         {"atlas": 15, "boon": 10, "cava": 10}
```

```
In [11]: ##########################################################################
         # Stripes for systems test 3 (given, no need to build stripes)
         ##########################################################################

         with open("systems_test_stripes_3", "w") as f:
             f.writelines([
                 '"DocA"\t{"X":20, "Y":30, "Z":5}\n',
                 '"DocB"\t{"X":100, "Y":20}\n',
                 '"DocC"\t{"M":5, "N":20, "Z":5, "Y":1}\n'
             ])
         !cat systems_test_stripes_3
```

```
"DocA"         {"X":20, "Y":30, "Z":5}
"DocB"         {"X":100, "Y":20}
"DocC"         {"M":5, "N":20, "Z":5, "Y":1}
```

3.6 Inverted indices for mini-test data

```
In [12]: !python invertedIndex.py -r hadoop systems_test_stripes_1 \
              --archive={pyArchive} \
             > systems_test_index_1
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.chqngh.20170621.174129.997537
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1196
  Submitted application application_1497906899862_1196
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1196
  Job job_1497906899862_1196 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1196 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.174
Counters: 49
        File Input Format Counters
                Bytes Read=3609
        File Output Format Counters
                Bytes Written=2192
        File System Counters
                FILE: Number of bytes read=1402
                FILE: Number of bytes written=400398
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3975
```

```
                HDFS: Number of bytes written=2192
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=25778688
                Total megabyte-milliseconds taken by all reduce tasks=49134080
                Total time spent by all map tasks (ms)=16783
                Total time spent by all maps in occupied slots (ms)=50349
                Total time spent by all reduce tasks (ms)=19193
                Total time spent by all reduces in occupied slots (ms)=95965
                Total vcore-milliseconds taken by all map tasks=16783
                Total vcore-milliseconds taken by all reduce tasks=19193
        Map-Reduce Framework
                CPU time spent (ms)=3800
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=200
                Input split bytes=366
                Map input records=28
                Map output bytes=3308
                Map output materialized bytes=1623
                Map output records=158
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1901072384
                Reduce input groups=28
                Reduce input records=158
                Reduce output records=28
                Reduce shuffle bytes=1623
                Shuffled Maps =2
                Spilled Records=316
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7758811136
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.2017
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.201
Removing temp directory /tmp/invertedIndex.chqngh.20170621.174129.997537...
WARNING:root:Elapsed time: 108.10663414 seconds
```

```
In minutes: 1.80177723567 mins
```

3.6.1 Inverted Indicies on Mini-Test Data Statistics:

Time

- *Run time: 108.11 seconds*
- *Run time: 1.8 minutes*

Input/Output statistics

- *Bytes Read: 3609*
- *Bytes Written: 2192*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 3080*

```
In [13]: !python invertedIndex.py -r hadoop systems_test_stripes_2 \
              --archive={pyArchive} \
             > systems_test_index_2

No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.chqngh.20170621.174318.388133
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1197
```

```
  Submitted application application_1497906899862_1197
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1197
  Job job_1497906899862_1197 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1197 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.174
Counters: 49
        File Input Format Counters
                Bytes Read=245
        File Output Format Counters
                Bytes Written=153
        File System Counters
                FILE: Number of bytes read=144
                FILE: Number of bytes written=397707
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=611
                HDFS: Number of bytes written=153
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=18461184
                Total megabyte-milliseconds taken by all reduce tasks=18065920
                Total time spent by all map tasks (ms)=12019
                Total time spent by all maps in occupied slots (ms)=36057
                Total time spent by all reduce tasks (ms)=7057
                Total time spent by all reduces in occupied slots (ms)=35285
                Total vcore-milliseconds taken by all map tasks=12019
                Total vcore-milliseconds taken by all reduce tasks=7057
        Map-Reduce Framework
                CPU time spent (ms)=2430
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=118
                Input split bytes=366
                Map input records=4
                Map output bytes=206
                Map output materialized bytes=190
                Map output records=10
```

```
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1906192384
                Reduce input groups=4
                Reduce input records=10
                Reduce output records=4
                Reduce shuffle bytes=190
                Shuffled Maps =2
                Spilled Records=20
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7757189120
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.2017
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.201
Removing temp directory /tmp/invertedIndex.chqngh.20170621.174318.388133...
WARNING:root:Elapsed time: 61.1126720905 seconds
    In minutes: 1.01854453484 mins
```

3.6.2 Inverted Indicies on System Test 2 Statistics:

Time

- *Run time: 61.11 seconds*
- *Run time: 1.02 minutes*

Input/Output statistics

- *Bytes Read: 245*
- *Bytes Written: 153*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2430*

```
In [14]: !python invertedIndex.py -r hadoop systems_test_stripes_3 \
              --archive={pyArchive} \
             > systems_test_index_3
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.chqngh.20170621.174419.784728
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1198
  Submitted application application_1497906899862_1198
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1198
  Job job_1497906899862_1198 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1198 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.20170621.174
Counters: 49
        File Input Format Counters
                Bytes Read=140
        File Output Format Counters
                Bytes Written=124
        File System Counters
                FILE: Number of bytes read=98
                FILE: Number of bytes written=397601
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=506
                HDFS: Number of bytes written=124
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
```

```
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=11292672
                Total megabyte-milliseconds taken by all reduce tasks=16296960
                Total time spent by all map tasks (ms)=7352
                Total time spent by all maps in occupied slots (ms)=22056
                Total time spent by all reduce tasks (ms)=6366
                Total time spent by all reduces in occupied slots (ms)=31830
                Total vcore-milliseconds taken by all map tasks=7352
                Total vcore-milliseconds taken by all reduce tasks=6366
        Map-Reduce Framework
                CPU time spent (ms)=2440
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=103
                Input split bytes=366
                Map input records=3
                Map output bytes=144
                Map output materialized bytes=130
                Map output records=9
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1906294784
                Reduce input groups=5
                Reduce input records=9
                Reduce output records=5
                Reduce shuffle bytes=130
                Shuffled Maps =2
                Spilled Records=18
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7773061120
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.2017
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/invertedIndex.chqngh.201
Removing temp directory /tmp/invertedIndex.chqngh.20170621.174419.784728...
WARNING:root:Elapsed time: 54.7973070145 seconds
    In minutes: 0.913288450241 mins
```

3.6.3 Inverted Indicies on System Test 3 Statistics:

Time

- *Run time: 54.79 seconds*
- *Run time: 0.91 minutes*

Input/Output statistics

- *Bytes Read: 140*
- *Bytes Written: 124*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2440*

```
In [15]: ############################################################
         # Pretty print systems tests for generating Inverted Index
         ############################################################
         import json

         for i in range(1,4):
             print "--"*100
             print "Systems test ",i," - Inverted Index"
             print "--"*100
             with open("systems_test_index_"+str(i),"r") as f:
                 lines = sorted(f.readlines())
                 for line in lines:
                     line = line.strip()
                     word, doc_list = line.split("\t")
                     doc_dict = json.loads(doc_list)
                     stripe=[]
                     for doc in doc_dict:
                         stripe.append([doc, doc_dict[doc]])
                     stripe=sorted(stripe)
                     stripe.extend([["",""] for _ in xrange(3 - len(stripe))])

                     print "{0:>16} |{1:>16} |{2:>16} |{3:>16}".format(
                       (word), stripe[0][0]+" "+str(stripe[0][1]), stripe[1][0]+" '

--------------------------------------------------------------------------------
Systems test  1  - Inverted Index
--------------------------------------------------------------------------------
             "a" |          bill 4 |     biography 4 |            by 4
          "bill" |             a 27 |  establishing 4 |           for 4
     "biography" |             a 27 |       general 4 |        george 4
            "by" |             a 27 |          city 4 |           sea 4
          "case" |             a 27 |        female 4 |    government 4
```

```
       "child's" |            a 27 |     christmas 4 |            in 7
     "christmas" |            a 27 |       child's 4 |            in 7
"circumstantial" |            a 27 |     narrative 4 |          of 15
          "city" |            a 27 |            by 4 |           sea 4
    "collection" |            a 27 |         fairy 4 |         forms 3
  "establishing" |            a 27 |          bill 4 |           for 4
         "fairy" |            a 27 |    collection 5 |           of 15
        "female" |            a 27 |          case 7 |           of 15
           "for" |            a 27 |          bill 4 |  establishing 4
         "forms" |            a 27 |    collection 5 |           of 15
       "general" |            a 27 |     biography 4 |        george 4
        "george" |            a 27 |     biography 4 |       general 4
    "government" |            a 27 |          case 7 |            in 7
            "in" |            a 27 |          case 7 |       child's 4
       "limited" |            a 27 |          case 7 |           of 15
     "narrative" |            a 27 |circumstantial 4 |           of 15
            "of" |            a 27 |     biography 4 |          case 7
     "religious" |            a 27 |          bill 4 |  establishing 4
           "sea" |            a 27 |            by 4 |          city 4
         "study" |            a 27 |          case 7 |        female 4
         "tales" |            a 27 |    collection 5 |         fairy 4
           "the" |            a 27 |            by 4 |circumstantial 4
         "wales" |            a 27 |       child's 4 |     christmas 4
-------------------------------------------------------------------------------
Systems test  2  - Inverted Index
-------------------------------------------------------------------------------
         "atlas" |          boon 3 |        dipped 3 |
          "boon" |         atlas 2 |          cava 2 |        dipped 3
          "cava" |          boon 3 |        dipped 3 |
        "dipped" |         atlas 2 |          boon 3 |          cava 2
-------------------------------------------------------------------------------
Systems test  3  - Inverted Index
-------------------------------------------------------------------------------
             "M" |          DocC 4 |                 |
             "N" |          DocC 4 |                 |
             "X" |          DocA 3 |          DocB 2 |
             "Y" |          DocA 3 |          DocB 2 |          DocC 4
             "Z" |          DocA 3 |          DocC 4 |
```

3.7 Similarities for mini-test data

```
In [16]: !python similarity.py -r hadoop systems_test_index_1 \
              --archive={pyArchive} \
             > systems_test_similarities_1

No configs found; falling back on auto-configuration
Creating temp directory /tmp/similarity.chqngh.20170621.174514.852045
```

```
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1200
  Submitted application application_1497906899862_1200
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1200
  Job job_1497906899862_1200 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1200 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174514
Counters: 49
        File Input Format Counters
                Bytes Read=3288
        File Output Format Counters
                Bytes Written=35050
        File System Counters
                FILE: Number of bytes read=3756
                FILE: Number of bytes written=407377
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=3642
                HDFS: Number of bytes written=35050
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
```

```
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=10910208
                Total megabyte-milliseconds taken by all reduce tasks=17902080
                Total time spent by all map tasks (ms)=7103
                Total time spent by all maps in occupied slots (ms)=21309
                Total time spent by all reduce tasks (ms)=6993
                Total time spent by all reduces in occupied slots (ms)=34965
                Total vcore-milliseconds taken by all map tasks=7103
                Total vcore-milliseconds taken by all reduce tasks=6993
        Map-Reduce Framework
                CPU time spent (ms)=2910
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=97
                Input split bytes=354
                Map input records=28
                Map output bytes=19239
                Map output materialized bytes=4940
                Map output records=673
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1898192896
                Reduce input groups=378
                Reduce input records=673
                Reduce output records=378
                Reduce shuffle bytes=4940
                Shuffled Maps =2
                Spilled Records=1346
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7727140864
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.2017062
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.201706
Removing temp directory /tmp/similarity.chqngh.20170621.174514.852045...
WARNING:root:Elapsed time: 61.4911048412 seconds
    In minutes: 1.02485174735 mins
```

3.7.1 Similarities on Mini Test Data Statistics:

Time

- *Run time: 61.49 seconds*
- *Run time: 1.02 minutes*

Input/Output statistics

- *Bytes Read: 3288*
- *Bytes Written: 35050*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2910*

```
In [17]: !python similarity.py -r hadoop systems_test_index_2 \
             --archive={pyArchive} \
             > systems_test_similarities_2

No configs found; falling back on auto-configuration
Creating temp directory /tmp/similarity.chqngh.20170621.174616.581989
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
```

```
  Submitting tokens for job: job_1497906899862_1202
  Submitted application application_1497906899862_1202
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1202
  Job job_1497906899862_1202 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1202 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174616
Counters: 49
        File Input Format Counters
                Bytes Read=230
        File Output Format Counters
                Bytes Written=511
        File System Counters
                FILE: Number of bytes read=130
                FILE: Number of bytes written=399002
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=584
                HDFS: Number of bytes written=511
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=18029568
                Total megabyte-milliseconds taken by all reduce tasks=9786880
                Total time spent by all map tasks (ms)=11738
                Total time spent by all maps in occupied slots (ms)=35214
                Total time spent by all reduce tasks (ms)=3823
                Total time spent by all reduces in occupied slots (ms)=19115
                Total vcore-milliseconds taken by all map tasks=11738
                Total vcore-milliseconds taken by all reduce tasks=3823
        Map-Reduce Framework
                CPU time spent (ms)=2520
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=61
                Input split bytes=354
                Map input records=4
                Map output bytes=212
                Map output materialized bytes=191
```

```
                Map output records=8
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1908436992
                Reduce input groups=6
                Reduce input records=8
                Reduce output records=6
                Reduce shuffle bytes=191
                Shuffled Maps =2
                Spilled Records=16
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7748808704
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.2017062
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.201706
Removing temp directory /tmp/similarity.chqngh.20170621.174616.581989...
WARNING:root:Elapsed time: 55.6270561218 seconds
    In minutes: 0.92711760203 mins
```

3.7.2 Similarities on Systems Test 2 Statistics:

Time

- *Run time: 55.63 seconds*
- *Run time: 0.93 minutes*

Input/Output statistics

- *Bytes Read: 230*
- *Bytes Written: 511*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2520*

```
In [18]: !python similarity.py -r hadoop systems_test_index_3 \
             --archive={pyArchive} \
             > systems_test_similarities_3
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/similarity.chqngh.20170621.174712.462793
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1204
  Submitted application application_1497906899862_1204
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1204
  Job job_1497906899862_1204 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1204 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.20170621.174712
Counters: 49
        File Input Format Counters
                Bytes Read=186
        File Output Format Counters
                Bytes Written=327
        File System Counters
                FILE: Number of bytes read=80
                FILE: Number of bytes written=398872
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=540
                HDFS: Number of bytes written=327
```

```
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=18961920
                Total megabyte-milliseconds taken by all reduce tasks=15470080
                Total time spent by all map tasks (ms)=12345
                Total time spent by all maps in occupied slots (ms)=37035
                Total time spent by all reduce tasks (ms)=6043
                Total time spent by all reduces in occupied slots (ms)=30215
                Total vcore-milliseconds taken by all map tasks=12345
                Total vcore-milliseconds taken by all reduce tasks=6043
        Map-Reduce Framework
                CPU time spent (ms)=2480
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=99
                Input split bytes=354
                Map input records=5
                Map output bytes=125
                Map output materialized bytes=111
                Map output records=5
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1908244480
                Reduce input groups=3
                Reduce input records=5
                Reduce output records=3
                Reduce shuffle bytes=111
                Shuffled Maps =2
                Spilled Records=10
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7730655232
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.2017062
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/similarity.chqngh.201706
Removing temp directory /tmp/similarity.chqngh.20170621.174712.462793...
WARNING:root:Elapsed time: 57.4144351482 seconds
    In minutes: 0.956907252471 mins
```

3.7.3 Similarities on Systems Test 3 Statistics:

Time

- *Run time: 57.41 seconds*
- *Run time: 0.96 minutes*

Input/Output statistics

- *Bytes Read: 186*
- *Bytes Written: 327*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2480*

```
In [19]: ##########################################
         # Pretty print systems tests
         ##########################################

         import json
         for i in range(1,4):
           print '--'*110
           print "Systems test ",i," - Similarity measures"
           print '--'*110
           print "{0:>15} |{1:>15} |{2:>15} |{3:>15} |{4:>15} |{5:>15}".format(
                   "average", "pair", "cosine", "jaccard", "overlap", "dice")
           print '-'*110

           with open("systems_test_similarities_"+str(i),"r") as f:
               lines = f.readlines()
               for line in lines:
                   line = line.strip()
                   avg,stripe = line.split("\t")
                   stripe = json.loads(stripe)

                   print "{0:>15f} |{1:>15} |{2:>15f} |{3:>15f} |{4:>15f} |{5:>15f}
                       float(avg), stripe[0], float(stripe[1]), float(stripe[2]), f


--------------------------------------------------------------------------------
Systems test  1  - Similarity measures
```

average	pair	cosine	jaccard	overlap
0.334842	a - bill	0.288675	0.107143	0.750000
0.334842	a - biography	0.288675	0.107143	0.750000
0.334842	a - by	0.288675	0.107143	0.750000
0.465201	a - case	0.436436	0.214286	0.857143
0.334842	a - child's	0.288675	0.107143	0.750000
0.334842	a - christmas	0.288675	0.107143	0.750000
0.334842	a - circumstantial	0.288675	0.107143	0.750
0.334842	a - city	0.288675	0.107143	0.750000
0.384281	a - collection	0.344265	0.142857	0.800000
0.334842	a - establishing	0.288675	0.107143	0.75000
0.334842	a - fairy	0.288675	0.107143	0.750000
0.334842	a - female	0.288675	0.107143	0.750000
0.334842	a - for	0.288675	0.107143	0.750000
0.273413	a - forms	0.222222	0.071429	0.666667
0.334842	a - general	0.288675	0.107143	0.750000
0.334842	a - george	0.288675	0.107143	0.750000
0.334842	a - government	0.288675	0.107143	0.750000
0.465201	a - in	0.436436	0.214286	0.857143
0.334842	a - limited	0.288675	0.107143	0.750000
0.334842	a - narrative	0.288675	0.107143	0.750000
0.698916	a - of	0.695666	0.500000	0.933333
0.334842	a - religious	0.288675	0.107143	0.750000
0.334842	a - sea	0.288675	0.107143	0.750000
0.465201	a - study	0.436436	0.214286	0.857143
0.334842	a - tales	0.288675	0.107143	0.750000
0.465201	a - the	0.436436	0.214286	0.857143
0.334842	a - wales	0.288675	0.107143	0.750000
0.223214	bill - biography	0.250000	0.142857	0.25000
0.223214	bill - by	0.250000	0.142857	0.250000
0.180200	bill - case	0.188982	0.100000	0.250000
0.223214	bill - child's	0.250000	0.142857	0.250000
0.223214	bill - christmas	0.250000	0.142857	0.25000
0.223214	bill - circumstantial	0.250000	0.142857	0.
0.223214	bill - city	0.250000	0.142857	0.250000
0.205207	bill - collection	0.223607	0.125000	0.2500
0.712500	bill - establishing	0.750000	0.600000	0.75
0.223214	bill - fairy	0.250000	0.142857	0.250000
0.223214	bill - female	0.250000	0.142857	0.250000
0.712500	bill - for	0.750000	0.600000	0.750000
0.268597	bill - forms	0.288675	0.166667	0.333333
0.223214	bill - general	0.250000	0.142857	0.250000
0.223214	bill - george	0.250000	0.142857	0.250000
0.223214	bill - government	0.250000	0.142857	0.2500
0.180200	bill - in	0.188982	0.100000	0.250000
0.223214	bill - limited	0.250000	0.142857	0.250000

```
0.223214 |bill - narrative |       0.250000 |       0.142857 |       0.25000
0.134980 |      bill - of |       0.129099 |       0.055556 |       0.250000
0.712500 |bill - religious |       0.750000 |       0.600000 |       0.75000
0.223214 |     bill - sea |       0.250000 |       0.142857 |       0.250000
0.180200 |   bill - study |       0.188982 |       0.100000 |       0.250000
0.223214 |   bill - tales |       0.250000 |       0.142857 |       0.250000
0.180200 |     bill - the |       0.188982 |       0.100000 |       0.250000
0.223214 |   bill - wales |       0.250000 |       0.142857 |       0.250000
0.223214 | biography - by |       0.250000 |       0.142857 |       0.250000
0.365956 |biography - case |       0.377964 |       0.222222 |       0.50000
0.223214 |biography - child's |       0.250000 |       0.142857 |       0.25
0.223214 |biography - christmas |       0.250000 |       0.142857 |       0.
0.458333 |biography - circumstantial |       0.500000 |       0.333333 |
0.223214 |biography - city |       0.250000 |       0.142857 |       0.25000
0.419343 |biography - collection |       0.447214 |       0.285714 |       0
0.223214 |biography - establishing |       0.250000 |       0.142857 |
0.458333 |biography - fairy |       0.500000 |       0.333333 |       0.5000
0.458333 |biography - female |       0.500000 |       0.333333 |       0.500
0.223214 |biography - for |       0.250000 |       0.142857 |       0.250000
0.553861 |biography - forms |       0.577350 |       0.400000 |       0.6666
0.712500 |biography - general |       0.750000 |       0.600000 |       0.75
0.712500 |biography - george |       0.750000 |       0.600000 |       0.750
0.223214 |biography - government |       0.250000 |       0.142857 |       0
0.180200 | biography - in |       0.188982 |       0.100000 |       0.250000
0.458333 |biography - limited |       0.500000 |       0.333333 |       0.50
0.458333 |biography - narrative |       0.500000 |       0.333333 |       0.
0.410147 | biography - of |       0.387298 |       0.187500 |       0.750000
0.223214 |biography - religious |       0.250000 |       0.142857 |       0.
0.223214 |biography - sea |       0.250000 |       0.142857 |       0.250000
0.365956 |biography - study |       0.377964 |       0.222222 |       0.5000
0.458333 |biography - tales |       0.500000 |       0.333333 |       0.5000
0.365956 |biography - the |       0.377964 |       0.222222 |       0.500000
0.223214 |biography - wales |       0.250000 |       0.142857 |       0.2500
0.180200 |      by - case |       0.188982 |       0.100000 |       0.250000
0.223214 |   by - child's |       0.250000 |       0.142857 |       0.250000
0.223214 | by - christmas |       0.250000 |       0.142857 |       0.250000
0.458333 |by - circumstantial |       0.500000 |       0.333333 |       0.50
0.712500 |      by - city |       0.750000 |       0.600000 |       0.750000
0.205207 |by - collection |       0.223607 |       0.125000 |       0.250000
0.223214 |by - establishing |       0.250000 |       0.142857 |       0.2500
0.223214 |     by - fairy |       0.250000 |       0.142857 |       0.250000
0.223214 |    by - female |       0.250000 |       0.142857 |       0.250000
0.223214 |       by - for |       0.250000 |       0.142857 |       0.250000
0.268597 |     by - forms |       0.288675 |       0.166667 |       0.333333
0.223214 |   by - general |       0.250000 |       0.142857 |       0.250000
0.223214 |    by - george |       0.250000 |       0.142857 |       0.250000
0.223214 |by - government |       0.250000 |       0.142857 |       0.250000
0.180200 |        by - in |       0.188982 |       0.100000 |       0.250000
```

```
0.223214 |   by - limited |       0.250000 |       0.142857 |       0.250000
0.458333 | by - narrative |       0.500000 |       0.333333 |       0.500000
0.271593 |        by - of |       0.258199 |       0.117647 |       0.500000
0.223214 | by - religious |       0.250000 |       0.142857 |       0.250000
0.712500 |       by - sea |       0.750000 |       0.600000 |       0.750000
0.180200 |     by - study |       0.188982 |       0.100000 |       0.250000
0.223214 |     by - tales |       0.250000 |       0.142857 |       0.250000
0.559350 |       by - the |       0.566947 |       0.375000 |       0.750000
0.223214 |     by - wales |       0.250000 |       0.142857 |       0.250000
0.365956 | case - child's |       0.377964 |       0.222222 |       0.500000
0.365956 |case - christmas |       0.377964 |       0.222222 |       0.50000
0.365956 |case - circumstantial |       0.377964 |       0.222222 |       0.
0.180200 |    case - city |       0.188982 |       0.100000 |       0.250000
0.317849 |case - collection |       0.338062 |       0.200000 |       0.4000
0.180200 |case - establishing |       0.188982 |       0.100000 |       0.25
0.365956 |   case - fairy |       0.377964 |       0.222222 |       0.500000
0.559350 |  case - female |       0.566947 |       0.375000 |       0.750000
0.180200 |     case - for |       0.188982 |       0.100000 |       0.250000
0.438276 |   case - forms |       0.436436 |       0.250000 |       0.666667
0.365956 | case - general |       0.377964 |       0.222222 |       0.500000
0.365956 |  case - george |       0.377964 |       0.222222 |       0.500000
0.559350 |case - government |       0.566947 |       0.375000 |       0.7500
0.389610 |      case - in |       0.428571 |       0.272727 |       0.428571
0.559350 | case - limited |       0.566947 |       0.375000 |       0.750000
0.365956 |case - narrative |       0.377964 |       0.222222 |       0.50000
0.386912 |      case - of |       0.390360 |       0.222222 |       0.571429
0.180200 |case - religious |       0.188982 |       0.100000 |       0.25000
0.180200 |     case - sea |       0.188982 |       0.100000 |       0.250000
0.830357 |   case - study |       0.857143 |       0.750000 |       0.857143
0.365956 |   case - tales |       0.377964 |       0.222222 |       0.500000
0.255952 |     case - the |       0.285714 |       0.166667 |       0.285714
0.365956 |   case - wales |       0.377964 |       0.222222 |       0.500000
0.712500 |child's - christmas |       0.750000 |       0.600000 |       0.75
0.223214 |child's - circumstantial |       0.250000 |       0.142857 |
0.223214 | child's - city |       0.250000 |       0.142857 |       0.250000
0.205207 |child's - collection |       0.223607 |       0.125000 |       0.2
0.223214 |child's - establishing |       0.250000 |       0.142857 |       0
0.223214 |child's - fairy |       0.250000 |       0.142857 |       0.250000
0.223214 |child's - female |       0.250000 |       0.142857 |       0.25000
0.223214 |  child's - for |       0.250000 |       0.142857 |       0.250000
0.268597 |child's - forms |       0.288675 |       0.166667 |       0.333333
0.223214 |child's - general |       0.250000 |       0.142857 |       0.2500
0.223214 |child's - george |       0.250000 |       0.142857 |       0.25000
0.458333 |child's - government |       0.500000 |       0.333333 |       0.5
0.559350 |   child's - in |       0.566947 |       0.375000 |       0.750000
0.223214 |child's - limited |       0.250000 |       0.142857 |       0.2500
0.223214 |child's - narrative |       0.250000 |       0.142857 |       0.25
0.134980 |   child's - of |       0.129099 |       0.055556 |       0.250000
```

```
0.223214 |child's - religious |       0.250000 |       0.142857 |       0.25
0.223214 |  child's - sea |       0.250000 |       0.142857 |       0.250000
0.365956 |child's - study |       0.377964 |       0.222222 |       0.500000
0.223214 |child's - tales |       0.250000 |       0.142857 |       0.250000
0.180200 |  child's - the |       0.188982 |       0.100000 |       0.250000
0.712500 |child's - wales |       0.750000 |       0.600000 |       0.750000
0.223214 |christmas - circumstantial |       0.250000 |       0.142857 |
0.223214 |christmas - city |       0.250000 |       0.142857 |       0.25000
0.205207 |christmas - collection |       0.223607 |       0.125000 |       0
0.223214 |christmas - establishing |       0.250000 |       0.142857 |
0.223214 |christmas - fairy |       0.250000 |       0.142857 |       0.2500
0.223214 |christmas - female |       0.250000 |       0.142857 |       0.250
0.223214 |christmas - for |       0.250000 |       0.142857 |       0.250000
0.268597 |christmas - forms |       0.288675 |       0.166667 |       0.3333
0.223214 |christmas - general |       0.250000 |       0.142857 |       0.25
0.223214 |christmas - george |       0.250000 |       0.142857 |       0.250
0.458333 |christmas - government |       0.500000 |       0.333333 |       0
0.559350 | christmas - in |       0.566947 |       0.375000 |       0.750000
0.223214 |christmas - limited |       0.250000 |       0.142857 |       0.25
0.223214 |christmas - narrative |       0.250000 |       0.142857 |       0.
0.134980 | christmas - of |       0.129099 |       0.055556 |       0.250000
0.223214 |christmas - religious |       0.250000 |       0.142857 |       0.
0.223214 |christmas - sea |       0.250000 |       0.142857 |       0.250000
0.365956 |christmas - study |       0.377964 |       0.222222 |       0.5000
0.223214 |christmas - tales |       0.250000 |       0.142857 |       0.2500
0.180200 |christmas - the |       0.188982 |       0.100000 |       0.250000
0.712500 |christmas - wales |       0.750000 |       0.600000 |       0.7500
0.458333 |circumstantial - city |       0.500000 |       0.333333 |       0.
0.419343 |circumstantial - collection |       0.447214 |       0.285714 |
0.223214 |circumstantial - establishing |       0.250000 |       0.142857 |
0.458333 |circumstantial - fairy |       0.500000 |       0.333333 |       0
0.458333 |circumstantial - female |       0.500000 |       0.333333 |
0.223214 |circumstantial - for |       0.250000 |       0.142857 |       0.2
0.553861 |circumstantial - forms |       0.577350 |       0.400000 |       0
0.458333 |circumstantial - general |       0.500000 |       0.333333 |
0.458333 |circumstantial - george |       0.500000 |       0.333333 |
0.223214 |circumstantial - government |       0.250000 |       0.142857 |
0.180200 |circumstantial - in |       0.188982 |       0.100000 |       0.25
0.458333 |circumstantial - limited |       0.500000 |       0.333333 |
0.712500 |circumstantial - narrative |       0.750000 |       0.600000 |
0.410147 |circumstantial - of |       0.387298 |       0.187500 |       0.75
0.223214 |circumstantial - religious |       0.250000 |       0.142857 |
0.458333 |circumstantial - sea |       0.500000 |       0.333333 |       0.5
0.365956 |circumstantial - study |       0.377964 |       0.222222 |       0
0.458333 |circumstantial - tales |       0.500000 |       0.333333 |       0
0.559350 |circumstantial - the |       0.566947 |       0.375000 |       0.7
0.223214 |circumstantial - wales |       0.250000 |       0.142857 |       0
0.205207 |city - collection |       0.223607 |       0.125000 |       0.2500
```

```
0.223214 |city - establishing |       0.250000 |       0.142857 |       0.25
0.223214 |   city - fairy |       0.250000 |       0.142857 |       0.250000
0.223214 |  city - female |       0.250000 |       0.142857 |       0.250000
0.223214 |     city - for |       0.250000 |       0.142857 |       0.250000
0.268597 |   city - forms |       0.288675 |       0.166667 |       0.333333
0.223214 | city - general |       0.250000 |       0.142857 |       0.250000
0.223214 |  city - george |       0.250000 |       0.142857 |       0.250000
0.223214 |city - government |       0.250000 |       0.142857 |       0.2500
0.180200 |      city - in |       0.188982 |       0.100000 |       0.250000
0.223214 | city - limited |       0.250000 |       0.142857 |       0.250000
0.458333 |city - narrative |       0.500000 |       0.333333 |       0.50000
0.271593 |      city - of |       0.258199 |       0.117647 |       0.500000
0.223214 |city - religious |       0.250000 |       0.142857 |       0.25000
0.712500 |     city - sea |       0.750000 |       0.600000 |       0.750000
0.180200 |   city - study |       0.188982 |       0.100000 |       0.250000
0.223214 |   city - tales |       0.250000 |       0.142857 |       0.250000
0.559350 |     city - the |       0.566947 |       0.375000 |       0.750000
0.223214 |   city - wales |       0.250000 |       0.142857 |       0.250000
0.205207 |collection - establishing |       0.223607 |       0.125000 |
0.646872 |collection - fairy |       0.670820 |       0.500000 |       0.750
0.419343 |collection - female |       0.447214 |       0.285714 |       0.50
0.205207 |collection - for |       0.223607 |       0.125000 |       0.25000
0.504099 |collection - forms |       0.516398 |       0.333333 |       0.666
0.419343 |collection - general |       0.447214 |       0.285714 |       0.5
0.419343 |collection - george |       0.447214 |       0.285714 |       0.50
0.205207 |collection - government |       0.223607 |       0.125000 |
0.156652 |collection - in |       0.169031 |       0.090909 |       0.200000
0.419343 |collection - limited |       0.447214 |       0.285714 |       0.5
0.419343 |collection - narrative |       0.447214 |       0.285714 |       0
0.477970 |collection - of |       0.461880 |       0.250000 |       0.800000
0.205207 |collection - religious |       0.223607 |       0.125000 |       0
0.205207 |collection - sea |       0.223607 |       0.125000 |       0.25000
0.317849 |collection - study |       0.338062 |       0.200000 |       0.400
0.646872 |collection - tales |       0.670820 |       0.500000 |       0.750
0.317849 |collection - the |       0.338062 |       0.200000 |       0.40000
0.205207 |collection - wales |       0.223607 |       0.125000 |       0.250
0.223214 |establishing - fairy |       0.250000 |       0.142857 |       0.2
0.223214 |establishing - female |       0.250000 |       0.142857 |       0.
0.712500 |establishing - for |       0.750000 |       0.600000 |       0.750
0.268597 |establishing - forms |       0.288675 |       0.166667 |       0.3
0.223214 |establishing - general |       0.250000 |       0.142857 |       0
0.223214 |establishing - george |       0.250000 |       0.142857 |       0.
0.223214 |establishing - government |       0.250000 |       0.142857 |
0.180200 |establishing - in |       0.188982 |       0.100000 |       0.2500
0.223214 |establishing - limited |       0.250000 |       0.142857 |       0
0.223214 |establishing - narrative |       0.250000 |       0.142857 |
0.134980 |establishing - of |       0.129099 |       0.055556 |       0.2500
0.712500 |establishing - religious |       0.750000 |       0.600000 |
```

```
0.223214 |establishing - sea |       0.250000 |       0.142857 |       0.250
0.180200 |establishing - study |       0.188982 |       0.100000 |       0.2
0.223214 |establishing - tales |       0.250000 |       0.142857 |       0.2
0.180200 |establishing - the |       0.188982 |       0.100000 |       0.250
0.223214 |establishing - wales |       0.250000 |       0.142857 |       0.2
0.458333 | fairy - female |       0.500000 |       0.333333 |       0.500000
0.223214 |    fairy - for |       0.250000 |       0.142857 |       0.250000
0.868292 |  fairy - forms |       0.866025 |       0.750000 |       1.000000
0.458333 |fairy - general |       0.500000 |       0.333333 |       0.500000
0.458333 | fairy - george |       0.500000 |       0.333333 |       0.500000
0.223214 |fairy - government |       0.250000 |       0.142857 |       0.250
0.180200 |     fairy - in |       0.188982 |       0.100000 |       0.250000
0.458333 |fairy - limited |       0.500000 |       0.333333 |       0.500000
0.458333 |fairy - narrative |       0.500000 |       0.333333 |       0.5000
0.410147 |     fairy - of |       0.387298 |       0.187500 |       0.750000
0.223214 |fairy - religious |       0.250000 |       0.142857 |       0.2500
0.223214 |    fairy - sea |       0.250000 |       0.142857 |       0.250000
0.365956 |  fairy - study |       0.377964 |       0.222222 |       0.500000
0.712500 |  fairy - tales |       0.750000 |       0.600000 |       0.750000
0.365956 |    fairy - the |       0.377964 |       0.222222 |       0.500000
0.223214 |  fairy - wales |       0.250000 |       0.142857 |       0.250000
0.223214 |   female - for |       0.250000 |       0.142857 |       0.250000
0.553861 | female - forms |       0.577350 |       0.400000 |       0.666667
0.458333 |female - general |       0.500000 |       0.333333 |       0.50000
0.458333 |female - george |       0.500000 |       0.333333 |       0.500000
0.712500 |female - government |       0.750000 |       0.600000 |       0.75
0.559350 |    female - in |       0.566947 |       0.375000 |       0.750000
1.000000 |female - limited |       1.000000 |       1.000000 |       1.00000
0.458333 |female - narrative |       0.500000 |       0.333333 |       0.500
0.410147 |    female - of |       0.387298 |       0.187500 |       0.750000
0.223214 |female - religious |       0.250000 |       0.142857 |       0.250
0.223214 |   female - sea |       0.250000 |       0.142857 |       0.250000
0.559350 | female - study |       0.566947 |       0.375000 |       0.750000
0.458333 | female - tales |       0.500000 |       0.333333 |       0.500000
0.365956 |   female - the |       0.377964 |       0.222222 |       0.500000
0.223214 | female - wales |       0.250000 |       0.142857 |       0.250000
0.268597 |    for - forms |       0.288675 |       0.166667 |       0.333333
0.223214 |  for - general |       0.250000 |       0.142857 |       0.250000
0.223214 |   for - george |       0.250000 |       0.142857 |       0.250000
0.223214 |for - government |       0.250000 |       0.142857 |       0.25000
0.180200 |       for - in |       0.188982 |       0.100000 |       0.250000
0.223214 |  for - limited |       0.250000 |       0.142857 |       0.250000
0.223214 |for - narrative |       0.250000 |       0.142857 |       0.250000
0.134980 |       for - of |       0.129099 |       0.055556 |       0.250000
0.712500 |for - religious |       0.750000 |       0.600000 |       0.750000
0.223214 |      for - sea |       0.250000 |       0.142857 |       0.250000
0.180200 |    for - study |       0.188982 |       0.100000 |       0.250000
0.223214 |    for - tales |       0.250000 |       0.142857 |       0.250000
```

```
0.180200 |      for - the |       0.188982 |       0.100000 |       0.250000
0.223214 |    for - wales |       0.250000 |       0.142857 |       0.250000
0.553861 |forms - general |       0.577350 |       0.400000 |       0.666667
0.553861 | forms - george |       0.577350 |       0.400000 |       0.666667
0.268597 |forms - government |       0.288675 |       0.166667 |       0.333
0.215666 |     forms - in |       0.218218 |       0.111111 |       0.333333
0.553861 |forms - limited |       0.577350 |       0.400000 |       0.666667
0.553861 |forms - narrative |       0.577350 |       0.400000 |       0.6666
0.328008 |     forms - of |       0.298142 |       0.125000 |       0.666667
0.268597 |forms - religious |       0.288675 |       0.166667 |       0.3333
0.268597 |    forms - sea |       0.288675 |       0.166667 |       0.333333
0.438276 |  forms - study |       0.436436 |       0.250000 |       0.666667
0.868292 |  forms - tales |       0.866025 |       0.750000 |       1.000000
0.438276 |    forms - the |       0.436436 |       0.250000 |       0.666667
0.268597 |  forms - wales |       0.288675 |       0.166667 |       0.333333
0.712500 |general - george |       0.750000 |       0.600000 |       0.75000
0.223214 |general - government |       0.250000 |       0.142857 |       0.2
0.180200 |   general - in |       0.188982 |       0.100000 |       0.250000
0.458333 |general - limited |       0.500000 |       0.333333 |       0.5000
0.458333 |general - narrative |       0.500000 |       0.333333 |       0.50
0.410147 |   general - of |       0.387298 |       0.187500 |       0.750000
0.223214 |general - religious |       0.250000 |       0.142857 |       0.25
0.223214 |  general - sea |       0.250000 |       0.142857 |       0.250000
0.365956 |general - study |       0.377964 |       0.222222 |       0.500000
0.458333 |general - tales |       0.500000 |       0.333333 |       0.500000
0.365956 |  general - the |       0.377964 |       0.222222 |       0.500000
0.223214 |general - wales |       0.250000 |       0.142857 |       0.250000
0.223214 |george - government |       0.250000 |       0.142857 |       0.25
0.180200 |    george - in |       0.188982 |       0.100000 |       0.250000
0.458333 |george - limited |       0.500000 |       0.333333 |       0.50000
0.458333 |george - narrative |       0.500000 |       0.333333 |       0.500
0.410147 |    george - of |       0.387298 |       0.187500 |       0.750000
0.223214 |george - religious |       0.250000 |       0.142857 |       0.250
0.223214 |   george - sea |       0.250000 |       0.142857 |       0.250000
0.365956 | george - study |       0.377964 |       0.222222 |       0.500000
0.458333 | george - tales |       0.500000 |       0.333333 |       0.500000
0.365956 |   george - the |       0.377964 |       0.222222 |       0.500000
0.223214 | george - wales |       0.250000 |       0.142857 |       0.250000
0.559350 |government - in |       0.566947 |       0.375000 |       0.750000
0.712500 |government - limited |       0.750000 |       0.600000 |       0.7
0.223214 |government - narrative |       0.250000 |       0.142857 |       0
0.410147 |government - of |       0.387298 |       0.187500 |       0.750000
0.223214 |government - religious |       0.250000 |       0.142857 |       0
0.223214 |government - sea |       0.250000 |       0.142857 |       0.25000
0.559350 |government - study |       0.566947 |       0.375000 |       0.750
0.223214 |government - tales |       0.250000 |       0.142857 |       0.250
0.180200 |government - the |       0.188982 |       0.100000 |       0.25000
0.458333 |government - wales |       0.500000 |       0.333333 |       0.500
```

```
0.559350 |   in - limited |       0.566947 |       0.375000 |       0.750000
0.180200 | in - narrative |       0.188982 |       0.100000 |       0.250000
0.287991 |        in - of |       0.292770 |       0.157895 |       0.428571
0.180200 | in - religious |       0.188982 |       0.100000 |       0.250000
0.180200 |       in - sea |       0.188982 |       0.100000 |       0.250000
0.389610 |     in - study |       0.428571 |       0.272727 |       0.428571
0.180200 |     in - tales |       0.188982 |       0.100000 |       0.250000
0.126374 |       in - the |       0.142857 |       0.076923 |       0.142857
0.559350 |     in - wales |       0.566947 |       0.375000 |       0.750000
0.458333 |limited - narrative |       0.500000 |       0.333333 |       0.50
0.410147 |   limited - of |       0.387298 |       0.187500 |       0.750000
0.223214 |limited - religious |       0.250000 |       0.142857 |       0.25
0.223214 |  limited - sea |       0.250000 |       0.142857 |       0.250000
0.559350 |limited - study |       0.566947 |       0.375000 |       0.750000
0.458333 |limited - tales |       0.500000 |       0.333333 |       0.500000
0.365956 |  limited - the |       0.377964 |       0.222222 |       0.500000
0.223214 |limited - wales |       0.250000 |       0.142857 |       0.250000
0.410147 | narrative - of |       0.387298 |       0.187500 |       0.750000
0.223214 |narrative - religious |       0.250000 |       0.142857 |       0.
0.458333 |narrative - sea |       0.500000 |       0.333333 |       0.500000
0.365956 |narrative - study |       0.377964 |       0.222222 |       0.5000
0.458333 |narrative - tales |       0.500000 |       0.333333 |       0.5000
0.559350 |narrative - the |       0.566947 |       0.375000 |       0.750000
0.223214 |narrative - wales |       0.250000 |       0.142857 |       0.2500
0.134980 | of - religious |       0.129099 |       0.055556 |       0.250000
0.271593 |       of - sea |       0.258199 |       0.117647 |       0.500000
0.386912 |     of - study |       0.390360 |       0.222222 |       0.571429
0.410147 |     of - tales |       0.387298 |       0.187500 |       0.750000
0.287991 |       of - the |       0.292770 |       0.157895 |       0.428571
0.134980 |     of - wales |       0.129099 |       0.055556 |       0.250000
0.223214 |religious - sea |       0.250000 |       0.142857 |       0.250000
0.180200 |religious - study |       0.188982 |       0.100000 |       0.2500
0.223214 |religious - tales |       0.250000 |       0.142857 |       0.2500
0.180200 |religious - the |       0.188982 |       0.100000 |       0.250000
0.223214 |religious - wales |       0.250000 |       0.142857 |       0.2500
0.180200 |    sea - study |       0.188982 |       0.100000 |       0.250000
0.223214 |    sea - tales |       0.250000 |       0.142857 |       0.250000
0.559350 |      sea - the |       0.566947 |       0.375000 |       0.750000
0.223214 |    sea - wales |       0.250000 |       0.142857 |       0.250000
0.365956 |  study - tales |       0.377964 |       0.222222 |       0.500000
0.255952 |    study - the |       0.285714 |       0.166667 |       0.285714
0.365956 |  study - wales |       0.377964 |       0.222222 |       0.500000
0.365956 |    tales - the |       0.377964 |       0.222222 |       0.500000
0.223214 |  tales - wales |       0.250000 |       0.142857 |       0.250000
0.180200 |    the - wales |       0.188982 |       0.100000 |       0.250000
-------------------------------------------------------------------------------
Systems test  2  - Similarity measures
-------------------------------------------------------------------------------
```

```
       average |           pair |         cosine |        jaccard |        overlap
-----------------------------------------------------------------------------------
      0.389562 |   atlas - boon |       0.408248 |       0.250000 |       0.500000
      1.000000 |   atlas - cava |       1.000000 |       1.000000 |       1.000000
      0.389562 | atlas - dipped |       0.408248 |       0.250000 |       0.500000
      0.389562 |    boon - cava |       0.408248 |       0.250000 |       0.500000
      0.625000 |  boon - dipped |       0.666667 |       0.500000 |       0.666667
      0.389562 |  cava - dipped |       0.408248 |       0.250000 |       0.500000
-----------------------------------------------------------------------------------
Systems test  3  - Similarity measures
-----------------------------------------------------------------------------------
       average |           pair |         cosine |        jaccard |        overlap
-----------------------------------------------------------------------------------
      0.820791 |    DocA - DocB |       0.816497 |       0.666667 |       1.000000
      0.553861 |    DocA - DocC |       0.577350 |       0.400000 |       0.666667
      0.346722 |    DocB - DocC |       0.353553 |       0.200000 |       0.500000
```

3.8 3. HW5.4.1 Full-scale experiment: EDA of Google n-grams dataset (PHASE 2)

Back to Table of Contents

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

- A. Longest 5-gram (number of characters)
- B. Top 10 most frequent words (please use the count information), i.e., unigrams
- C. 20 Most/Least densely appearing words (count/pages_count) sorted in decreasing order of relative frequency
- D. Distribution of 5-gram sizes (character length). E.g., count (using the count field) up how many times a 5-gram of 50 characters shows up. Plot the data graphically using a histogram.

3.8.1 HW5.4.1 - A. Longest 5-gram (number of characters)

In [20]:
```
%%writefile longest5gram.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-

import re

import mrjob
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import logging
import time

class longest5gram(MRJob):

    # START STUDENT CODE 5.4.1.A
```

```
MRJob.SORT_VALUES = True

def __init__(self, args):
    super(longest5gram, self).__init__(args)
    self.max_count = 0
    self.max_grams = []

def mapper(self, _, line):

    # Split line
    splits = line.rstrip("\n").split("\t")
    words = splits[0].lower().split()

    char_count = 0

    # Count characters
    for word in words:
        char_count += len(word)

    # Optimization: we track the max count local to the current mapper
    # have higher count than the max, we output them and update the ma
    # records that are smaller than the local max.
    # Even some non-max records are passed on, the good thing about th
    # in that it uses constant memory.
    if char_count > self.max_count:
        self.max_count = char_count
        yield (words), char_count
    elif char_count == self.max_count:
        yield (words), char_count


def combiner(self, ngram, char_counts):
    current_max = max(char_counts)

    # Optimization: we track the max count local to the current combir
    # have higher count than the max, we output them and update the ma
    # records that are smaller than the local max, drastically reducir
    # Even some non-max or local max records are passed on, the good t
    # memory efficient in that it uses constant memory (just 1 integer
    if current_max > self.max_count:
        self.max_count = current_max
        yield ngram, current_max
    elif current_max == self.max_count:
        yield ngram, current_max

def reducer(self, ngram, char_counts):
```

```
                current_count = max(char_counts)

                # Track in max_grams the n-grams with the max count of words
                if current_count > self.max_count:
                    self.max_count = current_count
                    self.max_grams = [(current_count, ngram)]
                elif current_count == self.max_count:
                    self.max_grams.append((current_count, ngram))

            def reducer_final(self):
                # Once
                for gram in self.max_grams:
                    yield gram[0], gram[1]


            def steps(self):

                # We need 1 reducer for this approach. However the optimizations i
                # help us ensure that a small percentage of records get to the red
                custom_jobconf = {
                    'stream.num.map.output.key.fields':'1',
                    'mapred.output.key.comparator.class': 'org.apache.hadoop.mapre
                    'mapred.text.key.comparator.options': '-k2,2nr',
                    'mapred.reduce.tasks': '1'
                }

                return [
                        MRStep(jobconf=custom_jobconf,
                            mapper=self.mapper,
                            reducer=self.reducer,
                            combiner = self.combiner,
                            reducer_final = self.reducer_final
                              )
                         ]

            # END STUDENT CODE 5.4.1.A

        if __name__ == '__main__':
            start_time = time.time()
            longest5gram.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting longest5gram.py
```

On test data set:

```
In [20]: !hdfs dfs -rm -r systems_test_stripes_5.4.1.a_1
         !python longest5gram.py -r hadoop googlebooks-eng-all-5gram-20090715-0-fil
             --archive={pyArchive} \
             > systems_test_stripes_5.4.1.a_1

rm: `systems_test_stripes_5.4.1.a_1': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/longest5gram.chqngh.20170621.174812.584718
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1206
  Submitted application application_1497906899862_1206
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1206
  Job job_1497906899862_1206 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1206 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170621.1748
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
```

```
        Bytes Written=106
File System Counters
        FILE: Number of bytes read=160
        FILE: Number of bytes written=401456
        FILE: Number of large read operations=0
        FILE: Number of read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=1011
        HDFS: Number of bytes written=106
        HDFS: Number of large read operations=0
        HDFS: Number of read operations=9
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Rack-local map tasks=2
        Total megabyte-milliseconds taken by all map tasks=19525632
        Total megabyte-milliseconds taken by all reduce tasks=10140160
        Total time spent by all map tasks (ms)=12712
        Total time spent by all maps in occupied slots (ms)=38136
        Total time spent by all reduce tasks (ms)=3961
        Total time spent by all reduces in occupied slots (ms)=19805
        Total vcore-milliseconds taken by all map tasks=12712
        Total vcore-milliseconds taken by all reduce tasks=3961
Map-Reduce Framework
        CPU time spent (ms)=2770
        Combine input records=3
        Combine output records=3
        Failed Shuffles=0
        GC time elapsed (ms)=140
        Input split bytes=448
        Map input records=10
        Map output bytes=154
        Map output materialized bytes=176
        Map output records=3
        Merged Map outputs=2
        Physical memory (bytes) snapshot=1909878784
        Reduce input groups=3
        Reduce input records=3
        Reduce output records=2
        Reduce shuffle bytes=176
        Shuffled Maps =2
        Spilled Records=6
        Total committed heap usage (bytes)=5218762752
        Virtual memory (bytes) snapshot=7766528000
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
```

```
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.2017
Removing temp directory /tmp/longest5gram.chqngh.20170621.174812.584718...
WARNING:root:Elapsed time: 56.3087100983 seconds
    In minutes: 0.938478501638 mins
```

3.8.2 Longest 5 Gram on Mini Test Set Statistics:

Time

- *Run time: 56.31 seconds*
- *Run time: 0.94 minutes*

Input/Output statistics

- *Bytes Read: 563*
- *Bytes Written: 106*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2770*

```
In [21]: !cat systems_test_stripes_5.4.1.a_1

29        ["a", "bill", "for", "establishing", "religious"]
29        ["a", "circumstantial", "narrative", "of", "the"]
```

On full data set:

```
In [22]: !hdfs dfs -rm -r full_stripes_5.4.1.a
         !python longest5gram.py -r hadoop hdfs:///user/cendylin/filtered-5Grams/ \
             --archive={pyArchive} \
             > full_stripes_5.4.1.a

rm: `full_stripes_5.4.1.a': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/longest5gram.chqngh.20170621.174911.583647
```

```
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
  Submitting tokens for job: job_1497906899862_1208
  Submitted application application_1497906899862_1208
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1208
  Job job_1497906899862_1208 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 3% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
   map 6% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 11% reduce 0%
   map 14% reduce 0%
   map 15% reduce 0%
   map 16% reduce 0%
   map 17% reduce 0%
   map 18% reduce 0%
   map 21% reduce 0%
   map 23% reduce 0%
   map 25% reduce 0%
   map 26% reduce 0%
   map 29% reduce 0%
   map 30% reduce 0%
```

```
   map 32% reduce 0%
   map 33% reduce 0%
   map 34% reduce 0%
   map 36% reduce 0%
   map 37% reduce 0%
   map 38% reduce 0%
   map 39% reduce 0%
   map 44% reduce 0%
   map 47% reduce 0%
   map 48% reduce 0%
   map 50% reduce 0%
   map 55% reduce 0%
   map 58% reduce 0%
   map 61% reduce 0%
   map 62% reduce 0%
   map 64% reduce 0%
   map 68% reduce 0%
   map 70% reduce 0%
   map 71% reduce 0%
   map 74% reduce 0%
   map 80% reduce 0%
   map 88% reduce 0%
   map 92% reduce 0%
   map 93% reduce 0%
   map 94% reduce 0%
   map 96% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1208 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170621.1749
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=360
        File System Counters
                FILE: Number of bytes read=37681
                FILE: Number of bytes written=25622952
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=360
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
```

```
                Launched map tasks=191
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=189
                Total megabyte-milliseconds taken by all map tasks=4439741952
                Total megabyte-milliseconds taken by all reduce tasks=10885120
                Total time spent by all map tasks (ms)=2890457
                Total time spent by all maps in occupied slots (ms)=8671371
                Total time spent by all reduce tasks (ms)=4252
                Total time spent by all reduces in occupied slots (ms)=21260
                Total vcore-milliseconds taken by all map tasks=2890457
                Total vcore-milliseconds taken by all reduce tasks=4252
        Map-Reduce Framework
                CPU time spent (ms)=393900
                Combine input records=2699
                Combine output records=1108
                Failed Shuffles=0
                GC time elapsed (ms)=25012
                Input split bytes=32000
                Map input records=58682266
                Map output bytes=172790
                Map output materialized bytes=63827
                Map output records=2699
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154265554944
                Reduce input groups=1108
                Reduce input records=1108
                Reduce output records=2
                Reduce shuffle bytes=63827
                Shuffled Maps =190
                Spilled Records=2216
                Total committed heap usage (bytes)=299992875008
                Virtual memory (bytes) snapshot=421420744704
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/longest5gram.chqngh.2017
Removing temp directory /tmp/longest5gram.chqngh.20170621.174911.583647...
WARNING:root:Elapsed time: 603.28647089 seconds
    In minutes: 10.0547745148 mins
```

3.8.3 Longest 5 Gram on Full Set Statistics:

Time

- *Run time: 603.29 seconds*
- *Run time: 10.05 minutes*

Input/Output statistics

- *Bytes Read: 2156069116*
- *Bytes Written: 360*

Cluster Resources

- *Number of Mappers: 191*
- *Number of Reducers: 1*
- *CPU time spent: 393900*

```
In [23]: !cat full_stripes_5.4.1.a
```

```
155         ["roplezimpredastrodonbraslpklson", "yhroaclmparcheyxmmioudavesaurus", '
155         ["aiopjumrxuyvaslyhypsibemapodikr", "ufrydiuuolbigasuaurusrexlisnaye", '
```

3.9 Report Stats:

3.10 example:

3.11 Longest 5grams MR stats

```
ec2_instance_type: m3.xlarge
num_ec2_instances: 15
```

Step 1:

```
RUNNING for 107.0s ~= 2 minutes
Reduce tasks = 16
```

Step 2:

```
RUNNING for 108.8s ~= 2 minutes
Reduce tasks = 1
```

3.11.1 HW5.4.1 - B. Top 10 most frequent words

```
In [24]: %%writefile mostFrequentWords.py
         #!~/anaconda2/bin/python
         # -*- coding: utf-8 -*-

         import re

         import mrjob
         from mrjob.protocol import RawProtocol
         from mrjob.job import MRJob
         from mrjob.step import MRStep
         import logging
         import time

         class mostFrequentWords(MRJob):

             # START STUDENT CODE 5.4.1.B

             MRJob.SORT_VALUES = True

             def mapper(self, _, line):

                 # Split line
                 splits = line.rstrip("\n").split("\t")
                 words = splits[0].lower().split()
                 count = int(splits[1])

                 for word in words:
                     yield word, count


             def combiner(self, word, counts):
                 total = sum(count for count in counts)
                 yield word, total

             def reducer(self, word, counts):
                 total = sum(count for count in counts)
                 yield total, word

             def max_reducer(self, count, words):
                 for word in words:
                     yield word, count

             def steps(self):

                 custom_jobconf = {
                     'stream.num.map.output.key.fields':'2',
```

```
                    'mapred.output.key.comparator.class': 'org.apache.hadoop.mapre
                    'mapred.text.key.comparator.options': '-k1,1nr',
                    'mapred.reduce.tasks': '1'
                }

                return [
                        MRStep(mapper=self.mapper,
                            reducer=self.reducer,
                            combiner = self.combiner),
                        MRStep(jobconf=custom_jobconf,
                               reducer=self.max_reducer)
                         ]

            # END STUDENT CODE 5.4.1.B

        if __name__ == '__main__':
            start_time = time.time()
            mostFrequentWords.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)
```

```
Overwriting mostFrequentWords.py
```

On the test data set:

```
In [25]: !hdfs dfs -rm -r systems_test_stripes_5.4.1.b_1
         !python mostFrequentWords.py -r hadoop googlebooks-eng-all-5gram-20090715-
             --archive={pyArchive} \
             > systems_test_stripes_5.4.1.b_1
```

```
rm: `systems_test_stripes_5.4.1.b_1': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/mostFrequentWords.chqngh.20170621.175917.650248
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
```

```
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1212
  Submitted application application_1497906899862_1212
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1212
  Job job_1497906899862_1212 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1212 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170621
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
                Bytes Written=357
        File System Counters
                FILE: Number of bytes read=430
                FILE: Number of bytes written=401198
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1021
                HDFS: Number of bytes written=357
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=11819520
                Total megabyte-milliseconds taken by all reduce tasks=9710080
                Total time spent by all map tasks (ms)=7695
                Total time spent by all maps in occupied slots (ms)=23085
                Total time spent by all reduce tasks (ms)=3793
                Total time spent by all reduces in occupied slots (ms)=18965
                Total vcore-milliseconds taken by all map tasks=7695
```

```
                Total vcore-milliseconds taken by all reduce tasks=3793
        Map-Reduce Framework
                CPU time spent (ms)=2930
                Combine input records=50
                Combine output records=31
                Failed Shuffles=0
                GC time elapsed (ms)=105
                Input split bytes=458
                Map input records=10
                Map output bytes=602
                Map output materialized bytes=458
                Map output records=50
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1913602048
                Reduce input groups=28
                Reduce input records=31
                Reduce output records=28
                Reduce shuffle bytes=458
                Shuffled Maps =2
                Spilled Records=62
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7776137216
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1214
```

```
  Submitted application application_1497906899862_1214
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1214
  Job job_1497906899862_1214 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1214 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170621
Counters: 49
        File Input Format Counters
                Bytes Read=536
        File Output Format Counters
                Bytes Written=357
        File System Counters
                FILE: Number of bytes read=396
                FILE: Number of bytes written=400663
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=908
                HDFS: Number of bytes written=357
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=26038272
                Total megabyte-milliseconds taken by all reduce tasks=19911680
                Total time spent by all map tasks (ms)=16952
                Total time spent by all maps in occupied slots (ms)=50856
                Total time spent by all reduce tasks (ms)=7778
                Total time spent by all reduces in occupied slots (ms)=38890
                Total vcore-milliseconds taken by all map tasks=16952
                Total vcore-milliseconds taken by all reduce tasks=7778
        Map-Reduce Framework
                CPU time spent (ms)=3130
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=88
                Input split bytes=372
                Map input records=28
                Map output bytes=385
                Map output materialized bytes=437
                Map output records=28
```

```
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1897517056
                Reduce input groups=28
                Reduce input records=28
                Reduce output records=28
                Reduce shuffle bytes=437
                Shuffled Maps =2
                Spilled Records=56
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7729250304
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh
Removing temp directory /tmp/mostFrequentWords.chqngh.20170621.175917.650248...
WARNING:root:Elapsed time: 91.7200388908 seconds
    In minutes: 1.52866731485 mins
```

3.11.2 Most Frequent Words on Test Set Statistics:

Time

- *Run time: 91.72 seconds*
- *Run time: 1.53 minutes*

Input/Output statistics Step 1

- *Bytes Read: 536*
- *Bytes Written: 357*

Input/Output statistics Step 2

- *Bytes Read: 536*
- *Bytes Written: 357*

Cluster Resources Step 1

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2930*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 3130*

```
In [26]: !head -5 systems_test_stripes_5.4.1.b_1
```

```
"a"        2217
"in"        1201
"wales"        1099
"christmas"        1099
"child's"        1099
```

On the full data set:

```
In [27]: !hdfs dfs -rm -r full_mostFrequentWords_5.4.1.b
         !python mostFrequentWords.py -r hadoop hdfs:///user/cendylin/filtered-5Gra
              --archive={pyArchive} \
             > full_mostFrequentWords_5.4.1.b
```

```
rm: `full_mostFrequentWords_5.4.1.b': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/mostFrequentWords.chqngh.20170621.180052.137838
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
```

```
Submitting tokens for job: job_1497906899862_1216
Submitted application application_1497906899862_1216
The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
Running job: job_1497906899862_1216
Job job_1497906899862_1216 running in uber mode : false
 map 0% reduce 0%
 map 1% reduce 0%
 map 2% reduce 0%
 map 3% reduce 0%
 map 4% reduce 0%
 map 5% reduce 0%
 map 6% reduce 0%
 map 7% reduce 0%
 map 8% reduce 0%
 map 9% reduce 0%
 map 10% reduce 0%
 map 11% reduce 0%
 map 13% reduce 0%
 map 14% reduce 0%
 map 16% reduce 0%
 map 17% reduce 0%
 map 18% reduce 0%
 map 19% reduce 0%
 map 20% reduce 0%
 map 21% reduce 0%
 map 22% reduce 0%
 map 23% reduce 0%
 map 24% reduce 0%
 map 26% reduce 0%
 map 27% reduce 0%
 map 28% reduce 0%
 map 29% reduce 0%
 map 30% reduce 0%
 map 31% reduce 0%
 map 32% reduce 0%
 map 33% reduce 0%
 map 34% reduce 0%
 map 35% reduce 0%
 map 36% reduce 0%
 map 37% reduce 0%
 map 38% reduce 0%
 map 39% reduce 0%
 map 40% reduce 0%
 map 41% reduce 0%
 map 42% reduce 0%
 map 43% reduce 0%
 map 44% reduce 0%
 map 45% reduce 0%
```

```
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 68% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 77% reduce 0%
map 79% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
```

```
   map 98% reduce 0%
   map 99% reduce 0%
   map 100% reduce 0%
   map 100% reduce 67%
   map 100% reduce 68%
   map 100% reduce 69%
   map 100% reduce 70%
   map 100% reduce 71%
   map 100% reduce 72%
   map 100% reduce 73%
   map 100% reduce 74%
   map 100% reduce 75%
   map 100% reduce 76%
   map 100% reduce 77%
   map 100% reduce 78%
   map 100% reduce 79%
   map 100% reduce 80%
   map 100% reduce 81%
   map 100% reduce 82%
   map 100% reduce 83%
   map 100% reduce 84%
   map 100% reduce 85%
   map 100% reduce 86%
   map 100% reduce 87%
   map 100% reduce 88%
   map 100% reduce 89%
   map 100% reduce 90%
   map 100% reduce 91%
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1216 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170621
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=4158739
        File System Counters
                FILE: Number of bytes read=39014765
                FILE: Number of bytes written=138285192
                FILE: Number of large read operations=0
```

```
        FILE: Number of read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=2156101116
        HDFS: Number of bytes written=4158739
        HDFS: Number of large read operations=0
        HDFS: Number of read operations=573
        HDFS: Number of write operations=2
Job Counters
        Killed map tasks=1
        Launched map tasks=191
        Launched reduce tasks=1
        Other local map tasks=2
        Rack-local map tasks=189
        Total megabyte-milliseconds taken by all map tasks=44741812224
        Total megabyte-milliseconds taken by all reduce tasks=311226880
        Total time spent by all map tasks (ms)=29128784
        Total time spent by all maps in occupied slots (ms)=87386352
        Total time spent by all reduce tasks (ms)=121573
        Total time spent by all reduces in occupied slots (ms)=607865
        Total vcore-milliseconds taken by all map tasks=29128784
        Total vcore-milliseconds taken by all reduce tasks=121573
Map-Reduce Framework
        CPU time spent (ms)=15461130
        Combine input records=293411330
        Combine output records=6822745
        Failed Shuffles=0
        GC time elapsed (ms)=126154
        Input split bytes=32000
        Map input records=58682266
        Map output bytes=3430141090
        Map output materialized bytes=73800744
        Map output records=293411330
        Merged Map outputs=190
        Physical memory (bytes) snapshot=154715713536
        Reduce input groups=269339
        Reduce input records=6822745
        Reduce output records=269339
        Reduce shuffle bytes=73800744
        Shuffled Maps =190
        Spilled Records=13645490
        Total committed heap usage (bytes)=298018930688
        Virtual memory (bytes) snapshot=421244575744
Shuffle Errors
        BAD_ID=0
        CONNECTION=0
        IO_ERROR=0
        WRONG_LENGTH=0
        WRONG_MAP=0
```

```
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1221
  Submitted application application_1497906899862_1221
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1221
  Job job_1497906899862_1221 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 79%
   map 100% reduce 90%
   map 100% reduce 100%
  Job job_1497906899862_1221 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.20170621
Counters: 49
        File Input Format Counters
                Bytes Read=4176522
        File Output Format Counters
                Bytes Written=4158739
        File System Counters
                FILE: Number of bytes read=2953963
                FILE: Number of bytes written=6323053
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4176894
                HDFS: Number of bytes written=4158739
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
```

```
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=20636160
                Total megabyte-milliseconds taken by all reduce tasks=50790400
                Total time spent by all map tasks (ms)=13435
                Total time spent by all maps in occupied slots (ms)=40305
                Total time spent by all reduce tasks (ms)=19840
                Total time spent by all reduces in occupied slots (ms)=99200
                Total vcore-milliseconds taken by all map tasks=13435
                Total vcore-milliseconds taken by all reduce tasks=19840
        Map-Reduce Framework
                CPU time spent (ms)=16680
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=260
                Input split bytes=372
                Map input records=269339
                Map output bytes=4428078
                Map output materialized bytes=2969260
                Map output records=269339
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1948950528
                Reduce input groups=269339
                Reduce input records=269339
                Reduce output records=269339
                Reduce shuffle bytes=2969260
                Shuffled Maps =2
                Spilled Records=538678
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7762849792
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh.
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/mostFrequentWords.chqngh
Removing temp directory /tmp/mostFrequentWords.chqngh.20170621.180052.137838...
WARNING:root:Elapsed time: 443.990502119 seconds
    In minutes: 7.39984170198 mins
```

3.11.3 Most Frequent Words on Full Set Statistics:

Time

- *Run time: 443.99 seconds*
- *Run time: 7.40 minutes*

Input/Output statistics Step 1

- *Bytes Read: 2156069116*
- *Bytes Written: 4158739*

Input/Output statistics Step 2

- *Bytes Read: 4176522*
- *Bytes Written: 4158739*

Cluster Resources Step 1

- *Number of Mappers: 191*
- *Number of Reducers: 1*
- *CPU time spent: 15461130*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 16680*

```
In [28]: !head -5 full_mostFrequentWords_5.4.1.b
```

```
"the"        5490815394
"of"        3698583299
"to"        2227866570
"in"        1421312776
"a"        1361123022
```

3.12 Most frequent words MR stats

```
ec2_instance_type: m3.xlarge
num_ec2_instances: 15
```

Step 1:

```
RUNNING for 590.7s ~= 10 minutes
Launched map tasks=191
Launched reduce tasks=57
```

Step 2:

```
RUNNING for 76.6s
Launched map tasks=110
Launched reduce tasks=16
```

3.12.1 HW5.4.1 - C. 20 Most/Least densely appearing words

In [29]:
```
%%writefile mostLeastDenseWords.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-
from __future__ import division
import re
import numpy as np
import mrjob
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import time
import logging
import math

class mostLeastDenseWords(MRJob):

    # START STUDENT CODE 5.4.1.C

    MRJob.SORT_VALUES = True

    def __init__(self, args):
        super(mostLeastDenseWords, self).__init__(args)
        self.total_word_count = None

    def mapper(self, _, line):

        # Split line
        splits = line.rstrip("\n").split("\t")
        words = splits[0].lower().split()
        count = int(splits[1])
```

```
            for word in words:
                yield "*", count
                yield word, count


        def combiner(self, word, counts):
            total = sum(count for count in counts)
            yield word, total

        def reducer(self, word, counts):

            total = sum(count for count in counts)

            if word == "*":
                self.total_word_count = total
            else:
                yield math.log(float(total) / float(self.total_word_count)), w

        def max_reducer(self, count, words):
            for word in words:
                yield word, count

        def steps(self):

            custom_jobconf = {
                'stream.num.map.output.key.fields':'2',
                'mapred.output.key.comparator.class': 'org.apache.hadoop.mapre
                'mapred.text.key.comparator.options': '-g -k1,1nr',
                'mapred.reduce.tasks': '1'
            }

            return [
                    MRStep(
                        mapper=self.mapper,
                        reducer=self.reducer,
                        combiner = self.combiner),
                    MRStep(jobconf=custom_jobconf,
                           reducer=self.max_reducer)
                     ]

        # END STUDENT CODE 5.4.1.C

    if __name__ == '__main__':
        start_time = time.time()
        mostLeastDenseWords.run()
        elapsed_time = time.time() - start_time
        mins = elapsed_time/float(60)
```

```
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting mostLeastDenseWords.py
```

On the test data set:

```
In [30]: !hdfs dfs -rm -r systems_test_stripes_5.4.1.c_1
         !python mostLeastDenseWords.py -r hadoop googlebooks-eng-all-5gram-2009071
             --archive={pyArchive} \
             > systems_test_stripes_5.4.1.c_1

rm: `systems_test_stripes_5.4.1.c_1': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/mostLeastDenseWords.chqngh.20170621.180818.927917
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1223
  Submitted application application_1497906899862_1223
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1223
  Job job_1497906899862_1223 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1223 completed successfully
```

```
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201706
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
                Bytes Written=810
        File System Counters
                FILE: Number of bytes read=447
                FILE: Number of bytes written=401330
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1025
                HDFS: Number of bytes written=810
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=11593728
                Total megabyte-milliseconds taken by all reduce tasks=10785280
                Total time spent by all map tasks (ms)=7548
                Total time spent by all maps in occupied slots (ms)=22644
                Total time spent by all reduce tasks (ms)=4213
                Total time spent by all reduces in occupied slots (ms)=21065
                Total vcore-milliseconds taken by all map tasks=7548
                Total vcore-milliseconds taken by all reduce tasks=4213
        Map-Reduce Framework
                CPU time spent (ms)=2940
                Combine input records=100
                Combine output records=33
                Failed Shuffles=0
                GC time elapsed (ms)=103
                Input split bytes=462
                Map input records=10
                Map output bytes=1032
                Map output materialized bytes=477
                Map output records=100
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1911533568
                Reduce input groups=29
                Reduce input records=33
                Reduce output records=28
                Reduce shuffle bytes=477
                Shuffled Maps =2
                Spilled Records=66
```

```
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7780741120
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1224
  Submitted application application_1497906899862_1224
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1224
  Job job_1497906899862_1224 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1224 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201706
Counters: 49
        File Input Format Counters
                Bytes Read=1215
        File Output Format Counters
                Bytes Written=810
        File System Counters
                FILE: Number of bytes read=646
                FILE: Number of bytes written=401352
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
```

```
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1591
                HDFS: Number of bytes written=810
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=24503808
                Total megabyte-milliseconds taken by all reduce tasks=10785280
                Total time spent by all map tasks (ms)=15953
                Total time spent by all maps in occupied slots (ms)=47859
                Total time spent by all reduce tasks (ms)=4213
                Total time spent by all reduces in occupied slots (ms)=21065
                Total vcore-milliseconds taken by all map tasks=15953
                Total vcore-milliseconds taken by all reduce tasks=4213
        Map-Reduce Framework
                CPU time spent (ms)=3010
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=105
                Input split bytes=376
                Map input records=28
                Map output bytes=838
                Map output materialized bytes=783
                Map output records=28
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1900892160
                Reduce input groups=28
                Reduce input records=28
                Reduce output records=28
                Reduce shuffle bytes=783
                Shuffled Maps =2
                Spilled Records=56
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7742779392
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqng
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqn
```

```
Removing temp directory /tmp/mostLeastDenseWords.chqngh.20170621.180818.927917...
WARNING:root:Elapsed time: 93.3255689144 seconds
    In minutes: 1.55542614857 mins
```

3.12.2 Most/Least Dense Words on Test Set Statistics:

Time

- *Run time: 93.33 seconds*
- *Run time: 1.56 minutes*

Input/Output statistics Step 1

- *Bytes Read: 563*
- *Bytes Written: 810*

Input/Output statistics Step 2

- *Bytes Read: 1215*
- *Bytes Written: 810*

Cluster Resources Step 1

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2490*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 3010*

```
In [31]: !cat systems_test_stripes_5.4.1.c_1
```

```
"a"        -1.6094379124341003
"in"        -2.2224382999704284
"wales"        -2.3111921676467904
"christmas"        -2.3111921676467904
"child's"        -2.3111921676467904
"of"        -2.3946529030299404
"study"        -2.9097739241155969
"case"        -2.9097739241155969
"female"        -3.2107895274368428
```

```
"collection"        -3.8368845701189014
"the"        -4.4930665564453749
"tales"        -4.501163766677994
"fairy"        -4.501163766677994
"forms"        -4.5597579309440475
"government"        -4.6883753087661404
"george"        -4.7915595450013715
"general"        -4.7915595450013715
"biography"        -4.7915595450013715
"city"        -5.1862137370053203
"circumstantial"        -5.1862137370053203
"by"        -5.1862137370053203
"sea"        -5.1862137370053203
"narrative"        -5.1862137370053203
"religious"        -5.2358106781446923
"establishing"        -5.2358106781446923
"for"        -5.2358106781446923
"bill"        -5.2358106781446923
"limited"        -5.3060149368179408
```

On the full data set:

```
In [32]: !hdfs dfs -rm -r full_mostLeastDenseWords_5.4.1.c
         !python mostLeastDenseWords.py -r hadoop hdfs:///user/cendylin/filtered-50
             --archive={pyArchive} \
             > full_mostLeastDenseWords_5.4.1.c

rm: `full_mostLeastDenseWords_5.4.1.c': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/mostLeastDenseWords.chqngh.20170621.180955.056202
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
```

```
Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
Loaded native gpl library from the embedded binaries
Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
Total input paths to process : 190
number of splits:190
Submitting tokens for job: job_1497906899862_1226
Submitted application application_1497906899862_1226
The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
Running job: job_1497906899862_1226
Job job_1497906899862_1226 running in uber mode : false
 map 0% reduce 0%
 map 1% reduce 0%
 map 2% reduce 0%
 map 3% reduce 0%
 map 4% reduce 0%
 map 5% reduce 0%
 map 6% reduce 0%
 map 7% reduce 0%
 map 8% reduce 0%
 map 9% reduce 0%
 map 10% reduce 0%
 map 11% reduce 0%
 map 12% reduce 0%
 map 13% reduce 0%
 map 14% reduce 0%
 map 15% reduce 0%
 map 16% reduce 0%
 map 17% reduce 0%
 map 18% reduce 0%
 map 19% reduce 0%
 map 20% reduce 0%
 map 21% reduce 0%
 map 22% reduce 0%
 map 23% reduce 0%
 map 24% reduce 0%
 map 25% reduce 0%
 map 26% reduce 0%
 map 27% reduce 0%
 map 28% reduce 0%
 map 29% reduce 0%
 map 30% reduce 0%
 map 31% reduce 0%
 map 32% reduce 0%
 map 33% reduce 0%
 map 34% reduce 0%
 map 35% reduce 0%
 map 36% reduce 0%
```

```
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 68% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 80% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
```

```
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 63%
map 100% reduce 67%
map 100% reduce 68%
map 100% reduce 69%
map 100% reduce 70%
map 100% reduce 71%
map 100% reduce 72%
map 100% reduce 73%
map 100% reduce 74%
map 100% reduce 75%
map 100% reduce 76%
map 100% reduce 77%
map 100% reduce 78%
map 100% reduce 79%
map 100% reduce 80%
map 100% reduce 81%
map 100% reduce 82%
map 100% reduce 83%
map 100% reduce 84%
map 100% reduce 85%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 88%
map 100% reduce 89%
map 100% reduce 90%
map 100% reduce 91%
map 100% reduce 92%
map 100% reduce 93%
map 100% reduce 94%
map 100% reduce 95%
map 100% reduce 96%
map 100% reduce 97%
```

```
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1226 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201706
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=8397356
        File System Counters
                FILE: Number of bytes read=39016573
                FILE: Number of bytes written=138296103
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=8397356
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=3
                Launched map tasks=193
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=191
                Total megabyte-milliseconds taken by all map tasks=71139557376
                Total megabyte-milliseconds taken by all reduce tasks=595397120
                Total time spent by all map tasks (ms)=46314816
                Total time spent by all maps in occupied slots (ms)=138944448
                Total time spent by all reduce tasks (ms)=232577
                Total time spent by all reduces in occupied slots (ms)=1162885
                Total vcore-milliseconds taken by all map tasks=46314816
                Total vcore-milliseconds taken by all reduce tasks=232577
        Map-Reduce Framework
                CPU time spent (ms)=29446900
                Combine input records=586822660
                Combine output records=6822933
                Failed Shuffles=0
                GC time elapsed (ms)=166218
                Input split bytes=32000
                Map input records=58682266
                Map output bytes=5878243690
                Map output materialized bytes=73804117
                Map output records=586822660
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154857852928
```

```
                Reduce input groups=269340
                Reduce input records=6822933
                Reduce output records=269339
                Reduce shuffle bytes=73804117
                Shuffled Maps =190
                Spilled Records=13645866
                Total committed heap usage (bytes)=295568408576
                Virtual memory (bytes) snapshot=421216260096
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1241
  Submitted application application_1497906899862_1241
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1241
  Job job_1497906899862_1241 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 73%
   map 100% reduce 80%
   map 100% reduce 87%
   map 100% reduce 95%
   map 100% reduce 100%
  Job job_1497906899862_1241 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqngh.201706
Counters: 49
```

```
File Input Format Counters
        Bytes Read=8524054
File Output Format Counters
        Bytes Written=8397356
File System Counters
        FILE: Number of bytes read=3650721
        FILE: Number of bytes written=7815546
        FILE: Number of large read operations=0
        FILE: Number of read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=8524430
        HDFS: Number of bytes written=8397356
        HDFS: Number of large read operations=0
        HDFS: Number of read operations=9
        HDFS: Number of write operations=2
Job Counters
        Launched map tasks=2
        Launched reduce tasks=1
        Rack-local map tasks=2
        Total megabyte-milliseconds taken by all map tasks=13779456
        Total megabyte-milliseconds taken by all reduce tasks=84121600
        Total time spent by all map tasks (ms)=8971
        Total time spent by all maps in occupied slots (ms)=26913
        Total time spent by all reduce tasks (ms)=32860
        Total time spent by all reduces in occupied slots (ms)=164300
        Total vcore-milliseconds taken by all map tasks=8971
        Total vcore-milliseconds taken by all reduce tasks=32860
Map-Reduce Framework
        CPU time spent (ms)=23090
        Combine input records=0
        Combine output records=0
        Failed Shuffles=0
        GC time elapsed (ms)=258
        Input split bytes=376
        Map input records=269339
        Map output bytes=8666695
        Map output materialized bytes=3764905
        Map output records=269339
        Merged Map outputs=2
        Physical memory (bytes) snapshot=1949208576
        Reduce input groups=269339
        Reduce input records=269339
        Reduce output records=269339
        Reduce shuffle bytes=3764905
        Shuffled Maps =2
        Spilled Records=538678
        Total committed heap usage (bytes)=5218762752
        Virtual memory (bytes) snapshot=7756877824
```

```
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqng
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/mostLeastDenseWords.chqn
Removing temp directory /tmp/mostLeastDenseWords.chqngh.20170621.180955.056202...
WARNING:root:Elapsed time: 1049.24184608 seconds
    In minutes: 17.4873641014 mins
```

3.12.3 Most/Least Dense Words on Full Set Statistics:

Time

- *Run time: 1049.24 seconds*
- *Run time: 17.49 minutes*

Input/Output statistics Step 1

- *Bytes Read: 2156069116*
- *Bytes Written: 8397356*

Input/Output statistics Step 2

- *Bytes Read: 8524054*
- *Bytes Written: 8397356*

Cluster Resources Step 1

- *Number of Mappers: 193*
- *Number of Reducers: 1*
- *CPU time spent: 29446900*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 23090*

3.13 Highest frequency words

```
In [33]: !head -20 full_mostLeastDenseWords_5.4.1.c
```

```
"the"        -2.1997146286591951
"of"        -2.5948415424832834
"to"        -3.1017469641694286
"in"        -3.5512104619731759
"a"        -3.5944812861760083
"and"        -3.7633969330906387
"that"        -4.1222901645139665
"is"        -4.1794296162291786
"be"        -4.2757305602957727
"as"        -4.6117218523743988
"it"        -4.6218779855265648
"was"        -4.6570322907820954
"for"        -4.6735590741386348
"not"        -4.8176584973674252
"with"        -4.8807319433270129
"on"        -4.9304593136462938
"by"        -4.9607205141927553
"he"        -5.0417326420258641
"have"        -5.0502224848469091
"which"        -5.1681749927840777
```

3.14 Lowest frequency words

```
In [34]: !tail -20 full_mostLeastDenseWords_5.4.1.c
         #TODO revert order probably
```

```
"appartenants"        -20.937177779481853
"imbodiment"        -20.937177779481853
"hadrianopolin"        -20.937177779481853
"anruf"        -20.937177779481853
"apposito"        -20.937177779481853
"ambrosiano"        -20.937177779481853
"bashfulest"        -20.937177779481853
"dtes"        -20.937177779481853
"dtcd"        -20.937177779481853
"falaba"        -20.937177779481853
"karakaya"        -20.937177779481853
"ampurdan"        -20.937177779481853
"broster"        -20.937177779481853
"greatgrandmother"        -20.937177779481853
"croissy"        -20.937177779481853
"greatlie"        -20.937177779481853
"chawing"        -20.937177779481853
"gottesheim's"        -20.937177779481853
```

```
"boivant"        -20.937177779481853
"anteponia"        -20.937177779481853
```

3.15 Word density MR stats

```
ec2_instance_type: m3.xlarge
num_ec2_instances: 15
```

Step 1:

```
RUNNING for 649.2s  ~= 10 minutes
Launched map tasks=190
Launched reduce tasks=57
```

Step 2:

```
RUNNING for 74.4s  ~= 1 minute
Launched map tasks=110
Launched reduce tasks=20
```

3.15.1 HW5.4.1 - D. Distribution of 5-gram sizes (character length)

```
In [35]: %%writefile distribution.py
         #!~/anaconda2/bin/python
         # -*- coding: utf-8 -*-

         import mrjob
         from mrjob.protocol import RawProtocol
         from mrjob.job import MRJob
         from mrjob.step import MRStep
         import time
         import logging

         class distribution(MRJob):

             # START STUDENT CODE 5.4.1.D

             MRJob.SORT_VALUES = True

             def mapper(self, _, line):

                 # Split line
                 splits = line.rstrip("\n").split("\t")
                 words = splits[0].lower().split()

                 char_count = 0

                 # Count characters
```

```
                for word in words:
                    char_count += len(word)

                yield char_count, 1


            def combiner(self, ngram_size, counts):
                yield ngram_size, sum(count for count in counts)

            def reducer(self, ngram_size, counts):
                yield ngram_size, sum(count for count in counts)

            def steps(self):

                custom_jobconf = {
                    'stream.num.map.output.key.fields':'2',
                    'mapred.output.key.comparator.class': 'org.apache.hadoop.mapre
                    'mapred.text.key.comparator.options': '-k1,1n',
                    'mapred.reduce.tasks': '1'
                }

                return [
                        MRStep(
                            jobconf=custom_jobconf,
                            mapper=self.mapper,
                            reducer=self.reducer,
                            combiner = self.combiner)
                ]

                # END STUDENT CODE 5.4.1.D

        if __name__ == '__main__':
            start_time = time.time()
            distribution.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)
```

```
Overwriting distribution.py
```

On the test data set:

```
In [36]: !hdfs dfs -rm -r 5.3distributions_test/part-00000
         !python distribution.py -r hadoop googlebooks-eng-all-5gram-20090715-0-fil
             --archive={pyArchive} \
             > 5.3distributions_test
```

```
rm: `5.3distributions_test/part-00000': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/distribution.chqngh.20170621.182727.140813
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1242
  Submitted application application_1497906899862_1242
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1242
  Job job_1497906899862_1242 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1242 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170621.1827
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
                Bytes Written=45
        File System Counters
                FILE: Number of bytes read=68
                FILE: Number of bytes written=401273
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
```

```
                HDFS: Number of bytes read=1011
                HDFS: Number of bytes written=45
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=40777728
                Total megabyte-milliseconds taken by all reduce tasks=23802880
                Total time spent by all map tasks (ms)=26548
                Total time spent by all maps in occupied slots (ms)=79644
                Total time spent by all reduce tasks (ms)=9298
                Total time spent by all reduces in occupied slots (ms)=46490
                Total vcore-milliseconds taken by all map tasks=26548
                Total vcore-milliseconds taken by all reduce tasks=9298
        Map-Reduce Framework
                CPU time spent (ms)=3420
                Combine input records=10
                Combine output records=10
                Failed Shuffles=0
                GC time elapsed (ms)=101
                Input split bytes=448
                Map input records=10
                Map output bytes=60
                Map output materialized bytes=88
                Map output records=10
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1919008768
                Reduce input groups=9
                Reduce input records=10
                Reduce output records=9
                Reduce shuffle bytes=88
                Shuffled Maps =2
                Spilled Records=20
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7778889728
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.2017
Removing temp directory /tmp/distribution.chqngh.20170621.182727.140813...
```

```
WARNING:root:Elapsed time: 68.6133317947 seconds
    In minutes: 1.14355552991 mins
```

3.15.2 5 Gram Disttribution on Test Set Statistics:

Time

- *Run time: 68.61 seconds*
- *Run time: 1.14 minutes*

Input/Output statistics

- *Bytes Read: 563*
- *Bytes Written: 45*

Cluster Resources

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 3420*

```
In [37]: !cat 5.3distributions_test
```

```
13        1
18        1
19        1
20        1
22        1
23        1
24        1
25        1
29        2
```

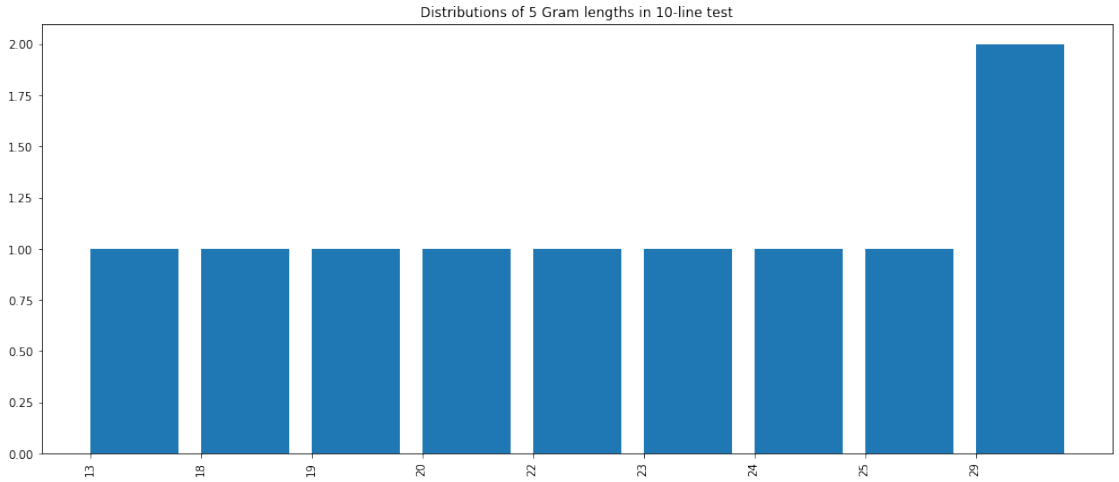
3.16 Test Histogram 10-line test

```
In [38]: %matplotlib inline
         import numpy as np
         import pylab as pl

         results_A = []
         for line in open("5.3distributions_test").readlines():
             line = line.strip()
             X,Y = line.split("\t")
             results_A.append([int(X),int(Y)])
```

```
items = (np.array(results_A)[::-1].T)
fig = pl.figure(figsize=(17,7))
ax = pl.subplot(111)
width=0.8
ax.bar(range(len(items[0])), items[1], width=width)
ax.set_xticks(np.arange(len(items[0])) + width/2)
ax.set_xticklabels(items[0], rotation=90)
ax.invert_xaxis()


pl.title("Distributions of 5 Gram lengths in 10-line test")
pl.show()
```



On the full data set:

```
In [39]: !hdfs dfs -rm -r full_distribution_5.4.1.d
         !python distribution.py -r hadoop hdfs:///user/cendylin/filtered-5Grams/ \
             --archive={pyArchive} \
             > full_distribution_5.4.1.d

rm: `full_distribution_5.4.1.d': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/distribution.chqngh.20170621.182839.411034
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170621.1
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
```

```
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
  Submitting tokens for job: job_1497906899862_1243
  Submitted application application_1497906899862_1243
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1243
  Job job_1497906899862_1243 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 3% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
   map 6% reduce 0%
   map 7% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 11% reduce 0%
   map 12% reduce 0%
   map 13% reduce 0%
   map 14% reduce 0%
   map 16% reduce 0%
   map 17% reduce 0%
   map 18% reduce 0%
   map 19% reduce 0%
   map 21% reduce 0%
   map 22% reduce 0%
   map 23% reduce 0%
   map 25% reduce 0%
   map 26% reduce 0%
   map 27% reduce 0%
   map 29% reduce 0%
   map 30% reduce 0%
   map 31% reduce 0%
   map 32% reduce 0%
```

```
map 34% reduce 0%
map 35% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 58% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 67% reduce 0%
map 68% reduce 0%
map 70% reduce 0%
map 72% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 78% reduce 0%
map 80% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
```

```
   map 99% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1243 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170621.1828
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=619
        File System Counters
                FILE: Number of bytes read=29346
                FILE: Number of bytes written=25629819
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=619
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=3
                Launched map tasks=193
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=191
                Total megabyte-milliseconds taken by all map tasks=15436655616
                Total megabyte-milliseconds taken by all reduce tasks=12277760
                Total time spent by all map tasks (ms)=10049906
                Total time spent by all maps in occupied slots (ms)=30149718
                Total time spent by all reduce tasks (ms)=4796
                Total time spent by all reduces in occupied slots (ms)=23980
                Total vcore-milliseconds taken by all map tasks=10049906
                Total vcore-milliseconds taken by all reduce tasks=4796
        Map-Reduce Framework
                CPU time spent (ms)=3812340
                Combine input records=58682266
                Combine output records=9172
                Failed Shuffles=0
                GC time elapsed (ms)=74981
                Input split bytes=32000
                Map input records=58682266
                Map output bytes=352088828
                Map output materialized bytes=79220
                Map output records=58682266
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154564587520
```

```
                Reduce input groups=80
                Reduce input records=9172
                Reduce output records=80
                Reduce shuffle bytes=79220
                Shuffled Maps =190
                Spilled Records=18344
                Total committed heap usage (bytes)=299992875008
                Virtual memory (bytes) snapshot=421097193472
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.20170
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/distribution.chqngh.2017
Removing temp directory /tmp/distribution.chqngh.20170621.182839.411034...
WARNING:root:Elapsed time: 178.160407066 seconds
    In minutes: 2.96934011777 mins
```

3.16.1 5 Gram Distribution on Full Set Statistics:

Time

- *Run time: 178.16 seconds*
- *Run time: 2.97 minutes*

Input/Output statistics

- *Bytes Read: 2156069116*
- *Bytes Written: 619*

Cluster Resources

- *Number of Mappers: 193*
- *Number of Reducers: 1*
- *CPU time spent: 3812340*

3.17 Distribution MRJob stats

Step 1:

```
RUNNING for 157.8s ~= 2.6 minutes
Launched map tasks=191
Launched reduce tasks=16
```

Step 2:

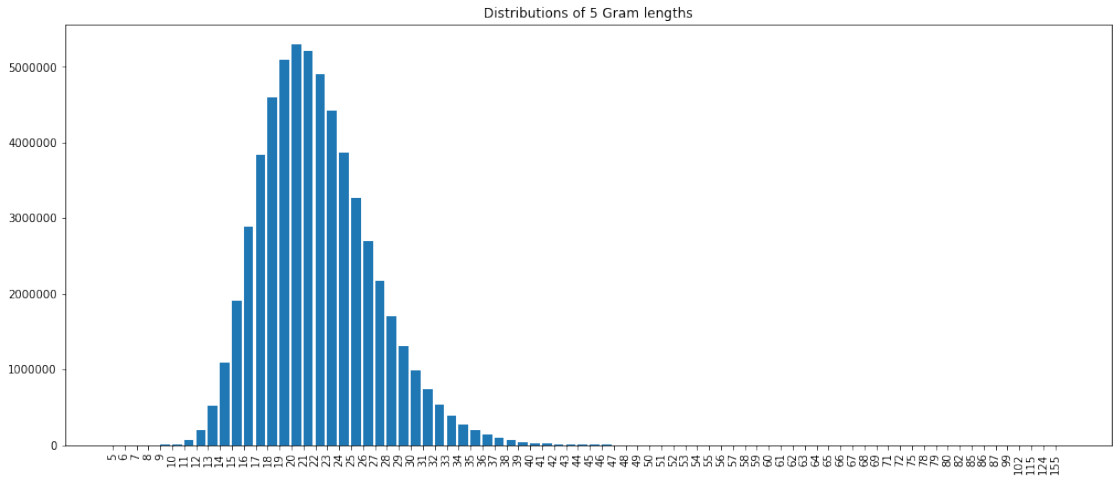
```
RUNNING for 115.0s ~= 2 minutes
Launched map tasks=139
Launched reduce tasks=1
```

```
In [40]: %matplotlib inline
         import numpy as np
         import pylab as pl

         results_A = []
         for line in open("full_distribution_5.4.1.d").readlines():
             line = line.strip()
             X,Y = line.split("\t")
             results_A.append([int(X),int(Y)])

         items = (np.array(results_A)[::-1].T)
         fig = pl.figure(figsize=(17,7))
         ax = pl.subplot(111)
         width=0.8
         ax.bar(range(len(items[0])), items[1], width=width)
         ax.set_xticks(np.arange(len(items[0])) + width/2)
         ax.set_xticklabels(items[0], rotation=90)
         ax.invert_xaxis()


         pl.title("Distributions of 5 Gram lengths")
         pl.show()
```



3.18 3. HW5.4.2 OPTIONAL Question: log-log plots (PHASE 2)

Back to Table of Contents

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/Power_law

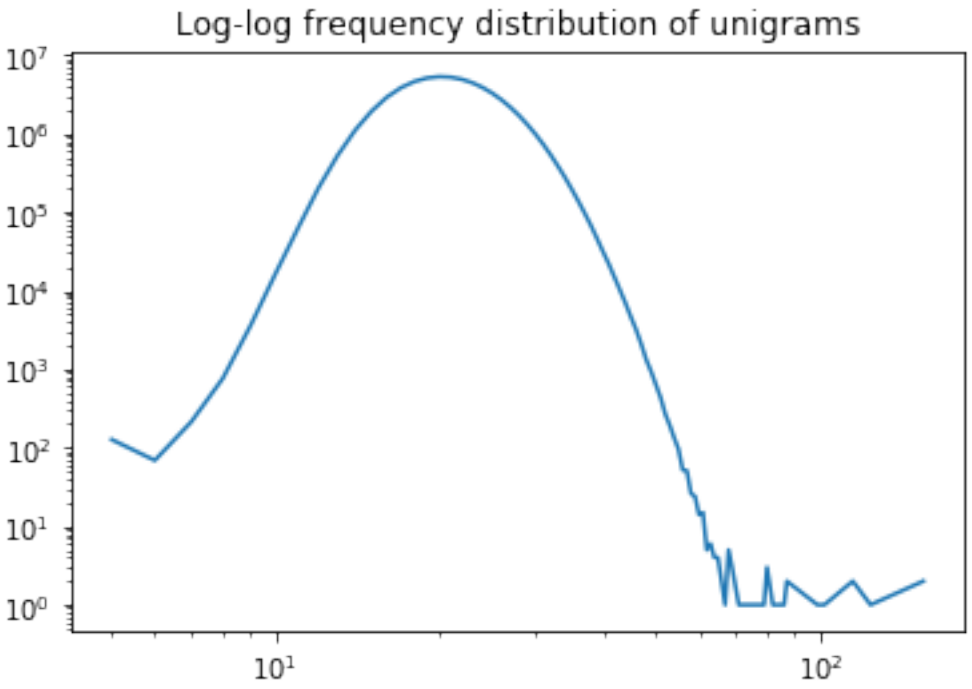
In [41]:
```
%matplotlib inline
import numpy as np
import pylab as plt

X = []
Y = []

for line in open("full_distribution_5.4.1.d").readlines():
    line = line.strip()
    x, y  = map(int, line.split("\t"))
    X.append(x)
    Y.append(y)

plt.loglog(X, Y)


plt.title("Log-log frequency distribution of unigrams")
plt.show()
```



3.19 3. HW5.5 Synonym detection over 2Gig of Data with extra Preprocessing steps (HW5.3 plus some preprocessing) (Phase 2)

Back to Table of Contents

For the remainder of this assignment please feel free to eliminate stop words from your analysis

> 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']

3.19.1 2: A large subset of the Google n-grams dataset as was described above

For each HW 5.4 -5.5.1 Please unit test and system test your code with respect to SYSTEMS TEST DATASET and show the results. Please compute the expected answer by hand and show your hand calculations for the SYSTEMS TEST DATASET. Then show the results you get with your system.

In this part of the assignment we will focus on developing methods for detecting synonyms, using the Google 5-grams dataset. At a high level:

1. remove stopwords
2. get 10,0000 most frequent
3. get 1000 (9001-10000) features
4. build stripes

To accomplish this you must script two main tasks using MRJob:

TASK (1) Build stripes for the most frequent 10,000 words using cooccurence information based on the words ranked from 9001,-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).

TASK (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 TASK (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 TASK (2).

Design notes for *TASK (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.

Please report the size of the cluster used and the amount of time it takes to run for the index construction task and for the synonym calculation task. How many pairs need to be processed (HINT: use the posting list length to calculate directly)? Report your Cluster configuration!

3.20 Example MR stats: (report times!)

```
took ~11 minutes on 5 m3.xlarge nodes
Data-local map tasks=188
Launched map tasks=190
Launched reduce tasks=15
Other local map tasks=2
```

```
In [43]: # START STUDENT CODE 5.5
         # ADD OR REMOVE CELLS AS NEEDED
```

3.21 Frequency ranking

```
In [2]: %%writefile frequencies5_5.py
        #!~/anaconda2/bin/python
        # -*- coding: utf-8 -*-
```

```
import re

import mrjob
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import time
import logging

class frequencies(MRJob):

    # START STUDENT CODE 5.4.1.B

    MRJob.SORT_VALUES = True

    def __init__(self, args):
        super(frequencies, self).__init__(args)
        self.current_rank = 0

    def configure_options(self):
        super(frequencies, self).configure_options()
        self.add_passthrough_option('--min_rank', dest='min_rank', type='ir
        self.add_passthrough_option('--max_rank', dest='max_rank', type='ir

    def mapper(self, _, line):

        # Split line
        splits = line.rstrip("\n").split("\t")
        words = splits[0].lower().split()
        count = int(splits[1])

        for word in words:
            yield word, count


    def combiner(self, word, counts):
        total = sum(count for count in counts)
        yield word, total

    def reducer(self, word, counts):
        total = sum(count for count in counts)
        yield total, word

    def max_reducer(self, count, words):

        # Words come in frequency descending order here
        # Only yield the words that are within the min and max frequency ra
```

```
                for word in words:
                    self.current_rank += 1

                    if self.current_rank >= self.options.min_rank and self.current_
                        yield word, count

            def steps(self):

                custom_jobconf = {
                    'stream.num.map.output.key.fields':'2',
                    'mapred.output.key.comparator.class': 'org.apache.hadoop.mapred
                    'mapred.text.key.comparator.options': '-k1,1nr',
                    'mapred.reduce.tasks': '1'
                }

                return [
                        MRStep(mapper=self.mapper,
                            reducer=self.reducer,
                            combiner = self.combiner),
                        MRStep(jobconf=custom_jobconf,
                               reducer=self.max_reducer)
                         ]

            # END STUDENT CODE 5.4.1.B

        if __name__ == '__main__':
            start_time = time.time()
            frequencies.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting frequencies5_5.py
```

3.21.1 Frequency ranking on 10-line test dataset

```
In [3]: !hdfs dfs -rm -r frequencies_test5.5
        !python frequencies5_5.py --min_rank 2 --max_rank 4 -r hadoop googlebooks-e
            --archive={pyArchive} \
            > frequencies_test5.5

rm: `frequencies_test5.5': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/frequencies5_5.chqngh.20170622.011039.181672
Looking for hadoop binary in /opt/hadoop/bin...
```

```
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1482
  Submitted application application_1497906899862_1482
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1482
  Job job_1497906899862_1482 running in uber mode : false
   map 0% reduce 0%
   map 50% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1482 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622.01
Counters: 49
        File Input Format Counters
                Bytes Read=563
        File Output Format Counters
                Bytes Written=357
        File System Counters
                FILE: Number of bytes read=430
                FILE: Number of bytes written=401288
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1015
                HDFS: Number of bytes written=357
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
```

```
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=25688064
                Total megabyte-milliseconds taken by all reduce tasks=18977280
                Total time spent by all map tasks (ms)=16724
                Total time spent by all maps in occupied slots (ms)=50172
                Total time spent by all reduce tasks (ms)=7413
                Total time spent by all reduces in occupied slots (ms)=37065
                Total vcore-milliseconds taken by all map tasks=16724
                Total vcore-milliseconds taken by all reduce tasks=7413
        Map-Reduce Framework
                CPU time spent (ms)=2970
                Combine input records=50
                Combine output records=31
                Failed Shuffles=0
                GC time elapsed (ms)=112
                Input split bytes=452
                Map input records=10
                Map output bytes=602
                Map output materialized bytes=458
                Map output records=50
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1902063616
                Reduce input groups=28
                Reduce input records=31
                Reduce output records=28
                Reduce shuffle bytes=458
                Shuffled Maps =2
                Spilled Records=62
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7756341248
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
```

```
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1484
  Submitted application application_1497906899862_1484
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1484
  Job job_1497906899862_1484 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1497906899862_1484 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622.01
Counters: 49
        File Input Format Counters
                Bytes Read=536
        File Output Format Counters
                Bytes Written=40
        File System Counters
                FILE: Number of bytes read=396
                FILE: Number of bytes written=400615
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=902
                HDFS: Number of bytes written=40
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=17522688
                Total megabyte-milliseconds taken by all reduce tasks=36666880
                Total time spent by all map tasks (ms)=11408
                Total time spent by all maps in occupied slots (ms)=34224
                Total time spent by all reduce tasks (ms)=14323
                Total time spent by all reduces in occupied slots (ms)=71615
                Total vcore-milliseconds taken by all map tasks=11408
```

```
                Total vcore-milliseconds taken by all reduce tasks=14323
        Map-Reduce Framework
                CPU time spent (ms)=2740
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=98
                Input split bytes=366
                Map input records=28
                Map output bytes=385
                Map output materialized bytes=437
                Map output records=28
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1911013376
                Reduce input groups=28
                Reduce input records=28
                Reduce output records=3
                Reduce shuffle bytes=437
                Shuffled Maps =2
                Spilled Records=56
                Total committed heap usage (bytes)=5218762752
                Virtual memory (bytes) snapshot=7750164480
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.201
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20
Removing temp directory /tmp/frequencies5_5.chqngh.20170622.011039.181672...
WARNING:root:Elapsed time: 125.631541014 seconds
    In minutes: 2.0938590169 mins
```

3.21.2 Frequencies on Test Set Statistics:

Time

- *Run time: 125.63 seconds*
- *Run time: 2.09 minutes*

Input/Output statistics Step 1

- *Bytes Read: 536*
- *Bytes Written: 357*

Input/Output statistics Step 2

- *Bytes Read: 536*
- *Bytes Written: 40*

Cluster Resources Step 1

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2970*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 2740*

```
In [2]: !cat frequencies_test5.5
```

```
"in"        1201
"wales"        1099
"christmas"        1099
```

3.21.3 Frequency ranking on full dataset

We run 2 jobs, one to calculate the top 10k frequencies, and a separate job to calculate a specific range. In this item we calculate the range 9001:10000 for the second job, but question 3.7 reuses the second job to calculate other custom ranges.

3.21.4 Top 10000

```
In [ ]: !hdfs dfs -rm -r frequencies5.5_top10k
        !python frequencies5_5.py --min_rank 1 --max_rank 10000 -r hadoop hdfs:///u
            --archive={pyArchive} \
            > frequencies5.5_top10k
```

```
rm: `frequencies5.5_top10k': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/frequencies5_5.chqngh.20170622.011247.508724
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
```

```
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
  Submitting tokens for job: job_1497906899862_1486
  Submitted application application_1497906899862_1486
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1486
  Job job_1497906899862_1486 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 3% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
   map 6% reduce 0%
   map 7% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 10% reduce 0%
   map 11% reduce 0%
   map 12% reduce 0%
   map 13% reduce 0%
   map 14% reduce 0%
   map 15% reduce 0%
   map 16% reduce 0%
   map 17% reduce 0%
   map 18% reduce 0%
   map 19% reduce 0%
   map 20% reduce 0%
   map 21% reduce 0%
   map 23% reduce 0%
   map 24% reduce 0%
   map 26% reduce 0%
   map 27% reduce 0%
```

```
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 32% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 67% reduce 0%
map 68% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
```

```
map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 80% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 87% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 42%
map 100% reduce 51%
map 100% reduce 58%
map 100% reduce 65%
map 100% reduce 67%
map 100% reduce 68%
map 100% reduce 69%
map 100% reduce 70%
map 100% reduce 71%
map 100% reduce 72%
map 100% reduce 73%
map 100% reduce 74%
map 100% reduce 75%
map 100% reduce 76%
map 100% reduce 77%
map 100% reduce 78%
map 100% reduce 79%
map 100% reduce 80%
map 100% reduce 81%
map 100% reduce 82%
map 100% reduce 83%
map 100% reduce 84%
map 100% reduce 85%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 88%
```

```
   map 100% reduce 89%
   map 100% reduce 90%
   map 100% reduce 91%
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1486 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622.01
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=4158739
        File System Counters
                FILE: Number of bytes read=39013260
                FILE: Number of bytes written=138292091
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=4158739
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
                Launched map tasks=192
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=190
                Total megabyte-milliseconds taken by all map tasks=46834326528
                Total megabyte-milliseconds taken by all reduce tasks=873692160
                Total time spent by all map tasks (ms)=30491098
                Total time spent by all maps in occupied slots (ms)=91473294
                Total time spent by all reduce tasks (ms)=341286
                Total time spent by all reduces in occupied slots (ms)=1706430
                Total vcore-milliseconds taken by all map tasks=30491098
                Total vcore-milliseconds taken by all reduce tasks=341286
        Map-Reduce Framework
                CPU time spent (ms)=15898760
                Combine input records=293411330
                Combine output records=6822745
```

```
                Failed Shuffles=0
                GC time elapsed (ms)=125872
                Input split bytes=32000
                Map input records=58682266
                Map output bytes=3430141090
                Map output materialized bytes=73800744
                Map output records=293411330
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154705190912
                Reduce input groups=269339
                Reduce input records=6822745
                Reduce output records=269339
                Reduce shuffle bytes=73800744
                Shuffled Maps =190
                Spilled Records=13645490
                Total committed heap usage (bytes)=297894674432
                Virtual memory (bytes) snapshot=421219143680
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
```

3.21.5 Frequencies on Full Set (Top 10000) Statistics:

Time

- *Run time: 477.21 seconds*
- *Run time: 8.14 minutes*

Input/Output statistics Step 1

- *Bytes Read: 2156069116*
- *Bytes Written: 4158739*

Input/Output statistics Step 2

- *Bytes Read: 2156069116*

- *Bytes Written: 4158739*

Cluster Resources Step 1

- *Number of Mappers: 192*
- *Number of Reducers: 1*
- *CPU time spent: 15898760*

Cluster Resources Step 2

- *Number of Mappers: 191*
- *Number of Reducers: 1*
- *CPU time spent: 15443020*

```
In [3]: !head frequencies5.5_top10k
```

```
"the"        5490815394
"of"        3698583299
"to"        2227866570
"in"        1421312776
"a"        1361123022
"and"        1149577477
"that"        802921147
"is"        758328796
"be"        688707130
"as"        492170314
```

3.21.6 Ranking 9001 : 10000

```
In [8]: !hdfs dfs -rm -r frequencies5.5
        !python frequencies5_5.py -r hadoop hdfs:///user/cendylin/filtered-5Grams/
            --archive={pyArchive} \
            > frequencies5.5
```

```
rm: `frequencies5.5': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/frequencies5_5.chqngh.20170622.020221.223750
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
  Submitting tokens for job: job_1497906899862_1526
  Submitted application application_1497906899862_1526
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1526
  Job job_1497906899862_1526 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 3% reduce 0%
   map 4% reduce 0%
   map 5% reduce 0%
   map 6% reduce 0%
   map 7% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 10% reduce 0%
   map 11% reduce 0%
   map 12% reduce 0%
   map 13% reduce 0%
   map 14% reduce 0%
   map 15% reduce 0%
   map 16% reduce 0%
   map 17% reduce 0%
   map 18% reduce 0%
   map 19% reduce 0%
   map 20% reduce 0%
   map 21% reduce 0%
   map 22% reduce 0%
   map 23% reduce 0%
   map 24% reduce 0%
   map 25% reduce 0%
   map 26% reduce 0%
   map 27% reduce 0%
```

```
map 28% reduce 0%
map 29% reduce 0%
map 30% reduce 0%
map 31% reduce 0%
map 32% reduce 0%
map 33% reduce 0%
map 34% reduce 0%
map 35% reduce 0%
map 36% reduce 0%
map 37% reduce 0%
map 38% reduce 0%
map 39% reduce 0%
map 40% reduce 0%
map 41% reduce 0%
map 42% reduce 0%
map 43% reduce 0%
map 44% reduce 0%
map 45% reduce 0%
map 46% reduce 0%
map 47% reduce 0%
map 48% reduce 0%
map 49% reduce 0%
map 50% reduce 0%
map 51% reduce 0%
map 52% reduce 0%
map 53% reduce 0%
map 54% reduce 0%
map 55% reduce 0%
map 56% reduce 0%
map 57% reduce 0%
map 58% reduce 0%
map 59% reduce 0%
map 60% reduce 0%
map 61% reduce 0%
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 66% reduce 0%
map 68% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 72% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
```

```
map 79% reduce 0%
map 80% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 59%
map 100% reduce 67%
map 100% reduce 68%
map 100% reduce 69%
map 100% reduce 70%
map 100% reduce 71%
map 100% reduce 72%
map 100% reduce 73%
map 100% reduce 74%
map 100% reduce 75%
map 100% reduce 76%
map 100% reduce 77%
map 100% reduce 78%
map 100% reduce 79%
map 100% reduce 80%
map 100% reduce 81%
map 100% reduce 82%
map 100% reduce 83%
map 100% reduce 84%
map 100% reduce 85%
map 100% reduce 86%
map 100% reduce 87%
map 100% reduce 88%
map 100% reduce 89%
map 100% reduce 90%
map 100% reduce 91%
```

```
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1526 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622.02
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=4158739
        File System Counters
                FILE: Number of bytes read=39014765
                FILE: Number of bytes written=138276597
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=4158739
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=1
                Launched map tasks=191
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=189
                Total megabyte-milliseconds taken by all map tasks=42921259008
                Total megabyte-milliseconds taken by all reduce tasks=588021760
                Total time spent by all map tasks (ms)=27943528
                Total time spent by all maps in occupied slots (ms)=83830584
                Total time spent by all reduce tasks (ms)=229696
                Total time spent by all reduces in occupied slots (ms)=1148480
                Total vcore-milliseconds taken by all map tasks=27943528
                Total vcore-milliseconds taken by all reduce tasks=229696
        Map-Reduce Framework
                CPU time spent (ms)=15822650
                Combine input records=293411330
                Combine output records=6822745
                Failed Shuffles=0
                GC time elapsed (ms)=115023
                Input split bytes=32000
```

```
                Map input records=58682266
                Map output bytes=3430141090
                Map output materialized bytes=73800744
                Map output records=293411330
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154402918400
                Reduce input groups=269339
                Reduce input records=6822745
                Reduce output records=269339
                Reduce shuffle bytes=73800744
                Shuffled Maps =190
                Spilled Records=13645490
                Total committed heap usage (bytes)=297668706304
                Virtual memory (bytes) snapshot=421226057728
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1539
  Submitted application application_1497906899862_1539
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1539
  Job job_1497906899862_1539 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 70%
   map 100% reduce 76%
```

```
   map 100% reduce 84%
   map 100% reduce 92%
   map 100% reduce 100%
  Job job_1497906899862_1539 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20170622.02
Counters: 49
        File Input Format Counters
                Bytes Read=4176522
        File Output Format Counters
                Bytes Written=17944
        File System Counters
                FILE: Number of bytes read=2953963
                FILE: Number of bytes written=6322927
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4176888
                HDFS: Number of bytes written=17944
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=61499904
                Total megabyte-milliseconds taken by all reduce tasks=123589120
                Total time spent by all map tasks (ms)=40039
                Total time spent by all maps in occupied slots (ms)=120117
                Total time spent by all reduce tasks (ms)=48277
                Total time spent by all reduces in occupied slots (ms)=241385
                Total vcore-milliseconds taken by all map tasks=40039
                Total vcore-milliseconds taken by all reduce tasks=48277
        Map-Reduce Framework
                CPU time spent (ms)=21190
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=831
                Input split bytes=366
                Map input records=269339
                Map output bytes=4428078
                Map output materialized bytes=2969260
                Map output records=269339
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1956659200
                Reduce input groups=269339
                Reduce input records=269339
```

```
                Reduce output records=1000
                Reduce shuffle bytes=2969260
                Shuffled Maps =2
                Spilled Records=538678
                Total committed heap usage (bytes)=5200936960
                Virtual memory (bytes) snapshot=7761235968
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.201
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/frequencies5_5.chqngh.20
Removing temp directory /tmp/frequencies5_5.chqngh.20170622.020221.223750...
WARNING:root:Elapsed time: 767.667668819 seconds
    In minutes: 12.794461147 mins
```

3.21.7 Frequencies on Full Set (Ranking 9001:10000) Statistics:

Time

- *Run time: 767.67 seconds*
- *Run time: 12.79 minutes*

Input/Output statistics Step 1

- *Bytes Read: 2156069116*
- *Bytes Written: 4158739*

Input/Output statistics Step 2

- *Bytes Read: 4176522*
- *Bytes Written: 17944*

Cluster Resources Step 1

- *Number of Mappers: 191*
- *Number of Reducers: 1*
- *CPU time spent: 15822650*

Cluster Resources Step 2

- *Number of Mappers: 2*
- *Number of Reducers: 1*
- *CPU time spent: 21190*

```
In [4]: !head frequencies5.5
```

```
"surveys"        169333
"jungle"        169314
"lacked"        169282
"correlate"        169273
"boxes"        169237
"escort"        169220
"disclosed"        169132
"shepherd"        169114
"commend"        169081
"zenith"        169049
```

3.22 Stripes

```
In [10]: import nltk
         nltk.download('wordnet')
         nltk.download('stopwords')

         from nltk.corpus import stopwords


         with open("stopwords", "w") as f:
             for word in stopwords.words('english'):
                 f.write(word + '\n')
```

```
[nltk_data] Downloading package wordnet to /home/chqngh/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package stopwords to /home/chqngh/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
In [5]: !head stopwords
```

```
i
me
my
myself
we
our
ours
```

```
ourselves
you
your
```

In [12]:
```
%%writefile stripes5_5.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-

from __future__ import division
import re
import mrjob
import json
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep
import itertools
import collections
import time
import logging


class stripes(MRJob):

    #START SUDENT CODE531_STRIPES

    MRJob.SORT_VALUES = True

    def __init__(self, args):
        # Words in the range we want to build stripes within the frequency
        self.valid_words = set()

        # Top 10k words in the frequency ranking. We don't consider words
        self.top10k_words = set()

        # Stop words to be ignored
        self.stop_words = set()
        super(stripes, self).__init__(args)

    def mapper_init(self):
        # Load the left table in the init so all mappers get this info
        with open("frequencies5.5", 'r') as f:
            for line in f.readlines():
                x = line.strip().split("\t")
                self.valid_words.add(x[0].strip("\""))

        with open("frequencies5.5_top10k", 'r') as f:
            for line in f.readlines():
```

```
                x = line.strip().split("\t")
                self.top10k_words.add(x[0].strip("\""))

        with open("stopwords", 'r') as f:
            for line in f.readlines():
                x = line.strip('\n')
                self.stop_words.add(x)

    def mapper(self, _, line):

        splits = line.rstrip("\n").split("\t")

        words = splits[0].lower().split()
        count = splits[1]

        H = {}
        for subset in itertools.combinations(sorted(set(words)), 2):


            # If many words go through here, later on the similarity job w
            # because it needs to calculate for the combinations of the it
            # performance hit. So we also remove stopwords here
            if subset[1] != subset[0] and subset[1] not in self.stop_words

                # We also want to make sure that both words belong to the
                if subset[1] in self.top10k_words and subset[0] in self.to

                    # Logic from slack discussion, with Sharmila and Kyle:
                    # If a given word is a valid word, then we build a str
                    # its co-ocurrence to the valid word. If both words ar
                    # to stripes in a symmetric way
                    if subset[1] in self.valid_words:

                        if subset[0] not in H.keys():
                            H[subset[0]] = {}
                            H[subset[0]][subset[1]] = count
                        elif subset[1] not in H[subset[0]]:
                            H[subset[0]][subset[1]] = count
                        else:
                            H[subset[0]][subset[1]] += count

                    if subset[0] in self.valid_words:

                        if subset[1] not in H.keys():
                            H[subset[1]] = {}
                            H[subset[1]][subset[0]] = count
                        elif subset[0] not in H[subset[1]]:
                            H[subset[1]][subset[0]] = count
```

```
                                else:
                                    H[subset[1]][subset[0]] += count

                    for key in H.keys():
                        yield key, H[key]

                def reducer(self, key, values):

                    counter = {}

                    for value in values:

                        for k, v in value.iteritems():
                            if k in counter:
                                counter[k] += int(v)
                            else:
                                counter[k] = int(v)

                    yield key, counter

              #END SUDENT CODE531_STRIPES

            if __name__ == '__main__':
                start_time = time.time()
                stripes.run()
                elapsed_time = time.time() - start_time
                mins = elapsed_time/float(60)
                a = """Elapsed time: %s seconds
                In minutes: %s mins""" % (str(elapsed_time), str(mins))
                logging.warning(a)

Overwriting stripes5_5.py


In [13]: !hdfs dfs -rm -r stripes5.5
         !python stripes5_5.py --file frequencies5.5_top10k --file frequencies5.5 -
             --archive={pyArchive} \
             > stripes5.5

rm: `stripes5.5': No such file or directory
No configs found; falling back on auto-configuration
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Creating temp directory /tmp/stripes5_5.chqngh.20170622.021512.576460
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/stripes5_5.chqngh.20170622.021
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
```

```
Detected hadoop configuration property names that do not match hadoop version 2.7.3
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
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 190
  number of splits:190
  Submitting tokens for job: job_1497906899862_1546
  Submitted application application_1497906899862_1546
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1546
  Job job_1497906899862_1546 running in uber mode : false
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 4% reduce 0%
   map 8% reduce 0%
   map 9% reduce 0%
   map 12% reduce 0%
   map 16% reduce 0%
   map 18% reduce 0%
   map 19% reduce 0%
   map 20% reduce 0%
   map 21% reduce 0%
   map 22% reduce 0%
   map 23% reduce 0%
   map 25% reduce 0%
   map 27% reduce 0%
   map 30% reduce 0%
   map 34% reduce 0%
   map 37% reduce 0%
   map 40% reduce 0%
   map 43% reduce 0%
   map 45% reduce 0%
   map 48% reduce 0%
   map 49% reduce 0%
   map 51% reduce 0%
   map 54% reduce 0%
```

```
map 57% reduce 0%
map 60% reduce 0%
map 64% reduce 0%
map 68% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 80% reduce 0%
map 81% reduce 0%
map 83% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 66%
map 100% reduce 67%
map 100% reduce 68%
map 100% reduce 69%
map 100% reduce 70%
map 100% reduce 71%
map 100% reduce 72%
map 100% reduce 73%
map 100% reduce 74%
map 100% reduce 75%
map 100% reduce 76%
map 100% reduce 77%
map 100% reduce 78%
map 100% reduce 79%
map 100% reduce 80%
map 100% reduce 81%
map 100% reduce 82%
map 100% reduce 83%
map 100% reduce 84%
map 100% reduce 85%
map 100% reduce 86%
```

```
   map 100% reduce 87%
   map 100% reduce 88%
   map 100% reduce 89%
   map 100% reduce 90%
   map 100% reduce 91%
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1546 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/stripes5_5.chqngh.20170622.021512
Counters: 51
        File Input Format Counters
                Bytes Read=2156069116
        File Output Format Counters
                Bytes Written=9441916
        File System Counters
                FILE: Number of bytes read=13105756
                FILE: Number of bytes written=61116398
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=2156101116
                HDFS: Number of bytes written=9441916
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=573
                HDFS: Number of write operations=2
        Job Counters
                Killed map tasks=2
                Launched map tasks=192
                Launched reduce tasks=1
                Other local map tasks=2
                Rack-local map tasks=190
                Total megabyte-milliseconds taken by all map tasks=10660137984
                Total megabyte-milliseconds taken by all reduce tasks=380282880
                Total time spent by all map tasks (ms)=6940194
                Total time spent by all maps in occupied slots (ms)=20820582
                Total time spent by all reduce tasks (ms)=148548
                Total time spent by all reduces in occupied slots (ms)=742740
                Total vcore-milliseconds taken by all map tasks=6940194
                Total vcore-milliseconds taken by all reduce tasks=148548
        Map-Reduce Framework
                CPU time spent (ms)=1707340
```

```
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=44411
                Input split bytes=32000
                Map input records=58682266
                Map output bytes=48662007
                Map output materialized bytes=22381856
                Map output records=1647512
                Merged Map outputs=190
                Physical memory (bytes) snapshot=154249535488
                Reduce input groups=1500812
                Reduce input records=1647512
                Reduce output records=9855
                Reduce shuffle bytes=22381856
                Shuffled Maps =190
                Spilled Records=3295024
                Total committed heap usage (bytes)=299992875008
                Virtual memory (bytes) snapshot=421315768320
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/stripes5_5.chqngh.2017062
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/stripes5_5.chqngh.201706
Removing temp directory /tmp/stripes5_5.chqngh.20170622.021512.576460...
WARNING:root:Elapsed time: 314.092614889 seconds
    In minutes: 5.23487691482 mins
```

3.22.1 Frequencies 5_5 on Full Set Statistics:

Time

- *Run time: 314.09 seconds*
- *Run time: 5.23 minutes*

Input/Output statistics

- *Bytes Read: 2156069116*
- *Bytes Written: 9441916*

Cluster Resources

- *Number of Mappers: 192*

- *Number of Reducers: 1*
- *CPU time spent: 1707340*

```
In [6]: # Verify stripe count
        !wc -l stripes5.5
```

```
9855 stripes5.5
```

```
In [3]: !head stripes5.5
```

```
"ab"        {"sterling": 62, "conveying": 273, "chord": 1006, "spectators": 89, "ho
"abandon"        {"misfortunes": 46, "ridicule": 148, "defenders": 45, "unconstitut
"abandoned"        {"restless": 63, "humiliation": 46, "lutheran": 42, "defenders":
"abandonment"        {"wept": 47, "conditional": 92, "forts": 145, "hasty": 76, "wh
"abbey"        {"warrior": 182, "nova": 56, "illuminated": 45, "vault": 129, "solem
"abdomen"        {"swell": 74, "penetrating": 845, "kicked": 212, "centered": 265,
"abdominal"        {"examinations": 159, "sac": 208, "oblique": 2343, "penetrating"
"abide"        {"codes": 180, "resolving": 71, "wills": 128, "frost": 112, "bacon":
"abilities"        {"rating": 43, "temperament": 96, "precarious": 51, "proportiona
"ability"        {"rating": 47, "surveys": 50, "groundwork": 41, "defenders": 53, "
```

3.23 Inverted Index

```
In [16]: %%writefile index5_5.py
         #!~/anaconda2/bin/python
         # -*- coding: utf-8 -*-


         from __future__ import division
         import collections
         import re
         import json
         import math
         import itertools
         import mrjob
         from mrjob.protocol import RawProtocol
         from mrjob.job import MRJob
         from mrjob.step import MRStep
         import json
         import time
         import logging

         class index(MRJob):

             #START SUDENT CODE531_INV_INDEX
```

```
            def mapper(self, _, line):
                key, stripeJson = line.strip().split('\t')
                key = key.strip("\"")
                stripe = json.loads(stripeJson)

                for k, v in stripe.iteritems():
                    yield k, [key, len(stripe)]

            def reducer(self, key, values):

                table = {}
                for value in values:
                    table[value[0]] = value[1]

                yield key, table

            #END SUDENT CODE531_INV_INDEX

        if __name__ == '__main__':
            start_time = time.time()
            index.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting index5_5.py


In [17]: !hdfs dfs -rm -r index5.5
         !python index5_5.py -r hadoop stripes5.5 \
              --archive={pyArchive} \
             > index5.5

rm: `index5.5': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/index5_5.chqngh.20170622.022029.624899
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/index5_5.chqngh.20170622.02202
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
```

```
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1497906899862_1554
  Submitted application application_1497906899862_1554
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1554
  Job job_1497906899862_1554 running in uber mode : false
   map 0% reduce 0%
   map 43% reduce 0%
   map 66% reduce 0%
   map 100% reduce 0%
   map 100% reduce 73%
   map 100% reduce 77%
   map 100% reduce 82%
   map 100% reduce 86%
   map 100% reduce 91%
   map 100% reduce 95%
   map 100% reduce 100%
  Job job_1497906899862_1554 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/index5_5.chqngh.20170622.022029.6
Counters: 49
        File Input Format Counters
                Bytes Read=9570622
        File Output Format Counters
                Bytes Written=8518090
        File System Counters
                FILE: Number of bytes read=6409140
                FILE: Number of bytes written=13029025
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=9570952
                HDFS: Number of bytes written=8518090
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=9
                HDFS: Number of write operations=2
        Job Counters
                Launched map tasks=2
                Launched reduce tasks=1
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=48110592
                Total megabyte-milliseconds taken by all reduce tasks=76206080
```

```
                Total time spent by all map tasks (ms)=31322
                Total time spent by all maps in occupied slots (ms)=93966
                Total time spent by all reduce tasks (ms)=29768
                Total time spent by all reduces in occupied slots (ms)=148840
                Total vcore-milliseconds taken by all map tasks=31322
                Total vcore-milliseconds taken by all reduce tasks=29768
        Map-Reduce Framework
                CPU time spent (ms)=48460
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=234
                Input split bytes=330
                Map input records=9855
                Map output bytes=15274194
                Map output materialized bytes=6222773
                Map output records=563656
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1920004096
                Reduce input groups=1000
                Reduce input records=563656
                Reduce output records=1000
                Reduce shuffle bytes=6222773
                Shuffled Maps =2
                Spilled Records=1127312
                Total committed heap usage (bytes)=5280104448
                Virtual memory (bytes) snapshot=7754145792
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/index5_5.chqngh.20170622.
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/index5_5.chqngh.20170622
Removing temp directory /tmp/index5_5.chqngh.20170622.022029.624899...
WARNING:root:Elapsed time: 93.5285439491 seconds
    In minutes: 1.55880906582 mins
```

3.23.1 Index 5_5 on Stripes5.5 Set Statistics:

Time

- *Run time: - 93.53 seconds*
- *Run time: - 1.56 minutes*

Input/Output statistics

- *Bytes Read: - 9570622*
- *Bytes Written: - 8518090*

Cluster Resources

- *Number of Mappers: - 2*
- *Number of Reducers: - 1*
- *CPU time spent: - 48460*

```
In [8]: # Verify index line count
        !wc -l index5.5
```

```
1000 index5.5
```

```
In [2]: !head index5.5
```

```
"abnormalities"        {"limited": 180, "evidence": 229, "skeleton": 18, "urinary":
"abyss"        {"issued": 87, "souls": 71, "hath": 172, "hanging": 62, "go": 413, "
"accent"        {"remarkable": 84, "cheerful": 32, "seemed": 369, "whose": 432, "nu
"accepts"        {"limited": 180, "concept": 175, "customary": 39, "sacrifice": 71,
"accompaniment"        {"needful": 19, "writings": 75, "customary": 39, "dance": 44
"accrue"        {"sector": 54, "mind": 453, "surplus": 21, "proposed": 128, "profit
"accumulate"        {"managed": 63, "leads": 118, "thirst": 23, "skin": 125, "hate"
"accusation"        {"partial": 75, "consider": 192, "chinese": 127, "caused": 218,
"acetic"        {"saturated": 22, "chain": 92, "less": 540, "caused": 218, "detecti
"acknowledgments"        {"customary": 39, "grateful": 45, "preface": 25, "go": 413
```

3.24 Similarity

```
In [3]: %%writefile similarity5_5.py
        #!~/anaconda2/bin/python
        # -*- coding: utf-8 -*-

        from __future__ import division
        import collections
        import re
        import json
        import math
        #import numpy as np
        import itertools
        import mrjob
        from mrjob.protocol import RawProtocol
        from mrjob.job import MRJob
```

```
from mrjob.step import MRStep
import time
import logging

class similarity(MRJob):

    #START SUDENT CODE531_SIMILARITY

    MRJob.SORT_VALUES = True

    def mapper(self, _, line):
        key, valuesJson = line.strip().split('\t')
        key = key.strip("\"")
        values = json.loads(valuesJson)

        for pair in itertools.combinations(sorted(set(values)), 2):
            yield pair, [values[pair[0]], values[pair[1]], 1]
    '''
    def combiner(self, key, values):
        intersection = 0
        count1 = None
        count2 = None

        # Iterate through the values
        for value in values:
            # Jaccard, get counts for the intersection
            # value[0] and value[1] are the same for all occurrences
            intersection += int(value[2])
            if count1 == None:
                count1 = value[0]
                count2 = value[1]

        yield key, [count1, count2, intersection]
    '''

    def reducer(self, key, values):
        intersection = 0
        count1 = None
        count2 = None

        cosine = 0.0

        # Iterate through the values
        for value in values:
            # Jaccard, get counts for the intersection, and for each set
            intersection += value[2]
            if count1 == None:
                count1 = value[0]
```

```
            count2 = value[1]

    for i in range (0, intersection):
        # Cosine
        a = 1 / math.sqrt(value[0])
        b = 1 / math.sqrt(value[1])
        cosine += a * b

    jaccard = float(intersection) / float(count1 + count2 - intersectio

    overlap_coefficient = float(intersection) / min(count1, count2)

    dice_coefficient = float(2 * intersection) / (count1 + count2)

    average = (cosine + jaccard + overlap_coefficient + dice_coefficien

    yield average, [key[0] + ' - ' + key[1], cosine, jaccard, overlap_c


def max_reducer(self, average, records):
    for record in records:
        yield average, record

def steps(self):
    custom_conf = {
        "mapred.map.tasks":'300',
        "mapred.reduce.tasks" : '15'
    }
    custom_jobconf = {
        'mapred.map.tasks': '100',
        'stream.num.map.output.key.fields':'2',
        'mapred.output.key.comparator.class': 'org.apache.hadoop.mapred
        'mapred.text.key.comparator.options': '-k1,1nr',
        'mapred.reduce.tasks': '1'
    }

    return [
            MRStep(jobconf=custom_conf,
                mapper=self.mapper,
                reducer=self.reducer#,
                   #combiner = self.combiner
                ),
            MRStep(jobconf=custom_jobconf,
                   reducer=self.max_reducer)
             ]

#END SUDENT CODE531_SIMILARITY
```

```
        if __name__ == '__main__':
            start_time = time.time()
            similarity.run()
            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

Overwriting similarity5_5.py


In [4]: !hdfs dfs -rm -r similarity5.5

        !python similarity5_5.py -r hadoop index5.5 \
             --archive={pyArchive} \
            > similarity5.5

rm: `similarity5.5': No such file or directory
No configs found; falling back on auto-configuration
Creating temp directory /tmp/similarity5_5.chqngh.20170622.082618.178612
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.3
Copying local files to hdfs:///user/chqngh/tmp/mrjob/similarity5_5.chqngh.20170622.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.map.tasks: mapreduce.job.maps
mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 1 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 1
  number of splits:300
  Submitting tokens for job: job_1497906899862_1809
  Submitted application application_1497906899862_1809
```

```
The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
Running job: job_1497906899862_1809
Job job_1497906899862_1809 running in uber mode : false
 map 0% reduce 0%
 map 1% reduce 0%
 map 2% reduce 0%
 map 3% reduce 0%
 map 4% reduce 0%
 map 5% reduce 0%
 map 6% reduce 0%
 map 7% reduce 0%
 map 9% reduce 0%
 map 12% reduce 0%
 map 13% reduce 0%
 map 15% reduce 0%
 map 16% reduce 0%
 map 17% reduce 0%
 map 20% reduce 0%
 map 24% reduce 0%
 map 29% reduce 0%
 map 31% reduce 0%
 map 33% reduce 0%
 map 34% reduce 0%
 map 35% reduce 0%
 map 36% reduce 0%
 map 37% reduce 0%
 map 38% reduce 0%
 map 39% reduce 0%
 map 40% reduce 0%
 map 41% reduce 0%
 map 42% reduce 0%
 map 43% reduce 0%
 map 44% reduce 0%
 map 45% reduce 0%
 map 46% reduce 0%
 map 47% reduce 0%
 map 48% reduce 0%
 map 50% reduce 0%
 map 51% reduce 0%
 map 52% reduce 0%
 map 53% reduce 0%
 map 54% reduce 0%
 map 55% reduce 0%
 map 56% reduce 0%
 map 57% reduce 0%
 map 58% reduce 0%
 map 60% reduce 0%
 map 61% reduce 0%
```

```
map 62% reduce 0%
map 63% reduce 0%
map 64% reduce 0%
map 65% reduce 0%
map 67% reduce 0%
map 68% reduce 0%
map 69% reduce 0%
map 70% reduce 0%
map 71% reduce 0%
map 72% reduce 0%
map 73% reduce 0%
map 74% reduce 0%
map 75% reduce 0%
map 76% reduce 0%
map 77% reduce 0%
map 78% reduce 0%
map 79% reduce 0%
map 81% reduce 0%
map 82% reduce 0%
map 83% reduce 0%
map 84% reduce 0%
map 85% reduce 0%
map 86% reduce 0%
map 87% reduce 0%
map 88% reduce 0%
map 89% reduce 0%
map 90% reduce 0%
map 91% reduce 0%
map 92% reduce 0%
map 93% reduce 0%
map 94% reduce 0%
map 95% reduce 0%
map 96% reduce 0%
map 97% reduce 0%
map 98% reduce 0%
map 99% reduce 0%
map 100% reduce 0%
map 100% reduce 5%
map 100% reduce 6%
map 100% reduce 9%
map 100% reduce 14%
map 100% reduce 15%
map 100% reduce 22%
map 100% reduce 35%
map 100% reduce 38%
map 100% reduce 42%
map 100% reduce 46%
map 100% reduce 47%
```

```
   map 100% reduce 49%
   map 100% reduce 53%
   map 100% reduce 55%
   map 100% reduce 59%
   map 100% reduce 61%
   map 100% reduce 64%
   map 100% reduce 65%
   map 100% reduce 66%
   map 100% reduce 67%
   map 100% reduce 68%
   map 100% reduce 69%
   map 100% reduce 70%
   map 100% reduce 71%
   map 100% reduce 72%
   map 100% reduce 73%
   map 100% reduce 74%
   map 100% reduce 75%
   map 100% reduce 76%
   map 100% reduce 77%
   map 100% reduce 78%
   map 100% reduce 79%
   map 100% reduce 80%
   map 100% reduce 81%
   map 100% reduce 82%
   map 100% reduce 83%
   map 100% reduce 84%
   map 100% reduce 85%
   map 100% reduce 86%
   map 100% reduce 87%
   map 100% reduce 88%
   map 100% reduce 89%
   map 100% reduce 90%
   map 100% reduce 91%
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1809 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/similarity5_5.chqngh.20170622.082
Counters: 50
        File Input Format Counters
                Bytes Read=39006886
        File Output Format Counters
```

```
        Bytes Written=4304774605
File System Counters
        FILE: Number of bytes read=693422285
        FILE: Number of bytes written=2976767336
        FILE: Number of large read operations=0
        FILE: Number of read operations=0
        FILE: Number of write operations=0
        HDFS: Number of bytes read=39057286
        HDFS: Number of bytes written=4304774605
        HDFS: Number of large read operations=0
        HDFS: Number of read operations=945
        HDFS: Number of write operations=30
Job Counters
        Killed map tasks=1
        Launched map tasks=301
        Launched reduce tasks=15
        Rack-local map tasks=301
        Total megabyte-milliseconds taken by all map tasks=30922023936
        Total megabyte-milliseconds taken by all reduce tasks=48313536000
        Total time spent by all map tasks (ms)=20131526
        Total time spent by all maps in occupied slots (ms)=60394578
        Total time spent by all reduce tasks (ms)=18872475
        Total time spent by all reduces in occupied slots (ms)=94362375
        Total vcore-milliseconds taken by all map tasks=20131526
        Total vcore-milliseconds taken by all reduce tasks=18872475
Map-Reduce Framework
        CPU time spent (ms)=27358240
        Combine input records=0
        Combine output records=0
        Failed Shuffles=0
        GC time elapsed (ms)=189563
        Input split bytes=50400
        Map input records=1000
        Map output bytes=6146158982
        Map output materialized bytes=2241363126
        Map output records=169873726
        Merged Map outputs=4500
        Physical memory (bytes) snapshot=257860091904
        Reduce input groups=28856437
        Reduce input records=169873726
        Reduce output records=28856437
        Reduce shuffle bytes=2241363126
        Shuffled Maps =4500
        Spilled Records=339747452
        Total committed heap usage (bytes)=496724082688
        Virtual memory (bytes) snapshot=710183120896
Shuffle Errors
        BAD_ID=0
```

```
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Detected hadoop configuration property names that do not match hadoop version 2.7.3
The have been translated as follows
 mapred.map.tasks: mapreduce.job.maps
mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.class
mapred.reduce.tasks: mapreduce.job.reduces
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.options
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.options
Running step 2 of 2...
  packageJobJar: [] [/opt/hadoop-2.7.3/share/hadoop/tools/lib/hadoop-streaming-2.7.
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Timeline service address: http://rm-ia.s3s.altiscale.com:8188/ws/v1/timeline/
  Connecting to ResourceManager at rm-ia.s3s.altiscale.com/10.251.255.108:8032
  Connecting to Application History server at rm-ia.s3s.altiscale.com/10.251.255.10
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05c
  Total input paths to process : 15
  Adding a new node: /default-rack/10.251.236.214:50010
  Adding a new node: /default-rack/10.251.237.214:50010
  Adding a new node: /default-rack/10.251.237.66:50010
  Adding a new node: /default-rack/10.251.253.174:50010
  Adding a new node: /default-rack/10.251.237.158:50010
  Adding a new node: /default-rack/10.251.240.186:50010
  Adding a new node: /default-rack/10.251.240.106:50010
  Adding a new node: /default-rack/10.251.235.97:50010
  Adding a new node: /default-rack/10.251.249.182:50010
  Adding a new node: /default-rack/10.251.253.170:50010
  Adding a new node: /default-rack/10.251.236.126:50010
  Adding a new node: /default-rack/10.251.235.85:50010
  Adding a new node: /default-rack/10.251.253.202:50010
  Adding a new node: /default-rack/10.251.249.146:50010
  Adding a new node: /default-rack/10.251.253.214:50010
  Adding a new node: /default-rack/10.251.240.178:50010
  Adding a new node: /default-rack/10.251.235.206:50010
  Adding a new node: /default-rack/10.251.249.90:50010
  Adding a new node: /default-rack/10.251.249.154:50010
  number of splits:105
  Submitting tokens for job: job_1497906899862_1815
  Submitted application application_1497906899862_1815
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1
  Running job: job_1497906899862_1815
  Job job_1497906899862_1815 running in uber mode : false
```

```
   map 0% reduce 0%
   map 1% reduce 0%
   map 2% reduce 0%
   map 7% reduce 0%
   map 18% reduce 0%
   map 20% reduce 0%
   map 22% reduce 0%
   map 24% reduce 0%
   map 29% reduce 0%
   map 33% reduce 0%
   map 37% reduce 0%
   map 42% reduce 0%
   map 45% reduce 0%
   map 47% reduce 0%
   map 50% reduce 0%
   map 52% reduce 0%
   map 58% reduce 0%
   map 62% reduce 0%
   map 65% reduce 0%
   map 69% reduce 0%
   map 73% reduce 0%
   map 75% reduce 0%
   map 77% reduce 0%
   map 78% reduce 0%
   map 79% reduce 0%
   map 81% reduce 0%
   map 83% reduce 0%
   map 87% reduce 0%
   map 88% reduce 0%
   map 89% reduce 0%
   map 90% reduce 0%
   map 93% reduce 0%
   map 95% reduce 0%
   map 98% reduce 0%
   map 100% reduce 0%
   map 100% reduce 15%
   map 100% reduce 23%
   map 100% reduce 32%
   map 100% reduce 33%
   map 100% reduce 36%
   map 100% reduce 40%
   map 100% reduce 44%
   map 100% reduce 49%
   map 100% reduce 55%
   map 100% reduce 61%
   map 100% reduce 67%
  Task Id : attempt_1497906899862_1815_r_000000_0, Status : FAILED
Container [pid=1819,containerID=container_1497906899862_1815_01_000110] is running
```

```
Dump of the process-tree for container_1497906899862_1815_01_000110 :
        |- PID PPID PGRPID SESSID CMD_NAME USER_MODE_TIME(MILLIS) SYSTEM_TIME(MILLI
        |- 4697 4695 1819 1819 (python) 340 4 115060736 4027 python similarity5_5.p
        |- 4695 1828 1819 1819 (sh) 0 0 9949184 707 /bin/sh -ex setup-wrapper.sh py
        |- 1828 1819 1819 1819 (java) 20123 1758 3335307264 651609 /usr/java/defaul
        |- 1819 1817 1819 1819 (bash) 0 1 9490432 512 /bin/bash -c /usr/java/defaul

Container killed on request. Exit code is 143
Container exited with a non-zero exit code 143

   map 100% reduce 0%
   map 100% reduce 15%
   map 100% reduce 18%
   map 100% reduce 23%
   map 100% reduce 29%
   map 100% reduce 32%
   map 100% reduce 34%
   map 100% reduce 37%
   map 100% reduce 41%
   map 100% reduce 46%
   map 100% reduce 52%
   map 100% reduce 59%
   map 100% reduce 66%
   map 100% reduce 67%
   map 100% reduce 68%
   map 100% reduce 69%
   map 100% reduce 70%
   map 100% reduce 71%
   map 100% reduce 72%
   map 100% reduce 73%
   map 100% reduce 74%
   map 100% reduce 75%
   map 100% reduce 76%
   map 100% reduce 77%
   map 100% reduce 78%
   map 100% reduce 79%
   map 100% reduce 80%
   map 100% reduce 81%
   map 100% reduce 82%
   map 100% reduce 83%
   map 100% reduce 84%
   map 100% reduce 85%
   map 100% reduce 86%
   map 100% reduce 87%
   map 100% reduce 88%
   map 100% reduce 89%
   map 100% reduce 90%
   map 100% reduce 91%
```

```
   map 100% reduce 92%
   map 100% reduce 93%
   map 100% reduce 94%
   map 100% reduce 95%
   map 100% reduce 96%
   map 100% reduce 97%
   map 100% reduce 98%
   map 100% reduce 99%
   map 100% reduce 100%
  Job job_1497906899862_1815 completed successfully
  Output directory: hdfs:///user/chqngh/tmp/mrjob/similarity5_5.chqngh.20170622.082
Counters: 51
        File Input Format Counters
                Bytes Read=4311500695
        File Output Format Counters
                Bytes Written=4304774605
        File System Counters
                FILE: Number of bytes read=753535584
                FILE: Number of bytes written=2052087260
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=4311519805
                HDFS: Number of bytes written=4304774605
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=318
                HDFS: Number of write operations=2
        Job Counters
                Failed reduce tasks=1
                Killed map tasks=1
                Launched map tasks=106
                Launched reduce tasks=2
                Rack-local map tasks=106
                Total megabyte-milliseconds taken by all map tasks=6893274624
                Total megabyte-milliseconds taken by all reduce tasks=10007943680
                Total time spent by all map tasks (ms)=4487809
                Total time spent by all maps in occupied slots (ms)=13463427
                Total time spent by all reduce tasks (ms)=3909353
                Total time spent by all reduces in occupied slots (ms)=19546765
                Total vcore-milliseconds taken by all map tasks=4487809
                Total vcore-milliseconds taken by all reduce tasks=3909353
        Map-Reduce Framework
                CPU time spent (ms)=5347030
                Combine input records=0
                Combine output records=0
                Failed Shuffles=0
                GC time elapsed (ms)=69041
                Input split bytes=19110
```

```
                Map input records=28856437
                Map output bytes=4362231833
                Map output materialized bytes=1284427979
                Map output records=28856437
                Merged Map outputs=105
                Physical memory (bytes) snapshot=88092209152
                Reduce input groups=28856437
                Reduce input records=28856437
                Reduce output records=28856437
                Reduce shuffle bytes=1284427979
                Shuffled Maps =105
                Spilled Records=57712874
                Total committed heap usage (bytes)=166263783424
                Virtual memory (bytes) snapshot=234258374656
        Shuffle Errors
                BAD_ID=0
                CONNECTION=0
                IO_ERROR=0
                WRONG_LENGTH=0
                WRONG_MAP=0
                WRONG_REDUCE=0
Streaming final output from hdfs:///user/chqngh/tmp/mrjob/similarity5_5.chqngh.2017
Removing HDFS temp directory hdfs:///user/chqngh/tmp/mrjob/similarity5_5.chqngh.201
Removing temp directory /tmp/similarity5_5.chqngh.20170622.082618.178612...
WARNING:root:Elapsed time: 5738.15179205 seconds
    In minutes: 95.6358632008 mins
```

3.24.1 Similarity 5_5 on Index 5.5 Set Statistics:

Time

- *Run time: - 5738.15 seconds*
- *Run time: - 95.64 minutes*

Input/Output statistics Step 1

- *Bytes Read: - 39006886*
- *Bytes Written: - 4304774605*

Input/Output statistics Step 2

- *Bytes Read: - 4311500695*
- *Bytes Written: - 4304774605*

Cluster Resources Step 1

- *Number of Mappers: - 301*
- *Number of Reducers: - 1*
- *CPU time spent: - 27358240*

Cluster Resources Step 2

- *Number of Mappers: - 106*
- *Number of Reducers: - 1*
- *CPU time spent: - 5347030*

```
In [6]:  !head similarity5.5
```

```
0.94644813148501816        ["may - one", 0.95225195659193151, 0.90845070422535212,
0.93908388451535063        ["one - time", 0.94280333419215168, 0.8904665314401623,
0.92367375447215494        ["one - well", 0.92926353279135865, 0.86633165829145731,
0.91781670099350054        ["one - would", 0.92061831133187233, 0.84959349593495936
0.9143083417507526        ["first - one", 0.91727050749090477, 0.84365482233502542,
0.90903394628645784        ["may - well", 0.92092827339961603, 0.85306122448979593,
0.90015031810239354        ["one - part", 0.90442739978532816, 0.82157258064516125,
0.89742767905867249        ["may - time", 0.91183344591858018, 0.83770161290322576,
0.89393870632282468        ["one - upon", 0.89684548994238211, 0.80749746707193515,
0.89053012293395373        ["may - would", 0.90135059121609029, 0.81901840490797551
```

```
In [ ]:
```

```
In [1]: !head similarity5.5
```

```
0.94644813148501816        ["may - one", 0.95225195659193151, 0.90845070422535212,
0.93908388451535063        ["one - time", 0.94280333419215168, 0.8904665314401623,
0.92367375447215494        ["one - well", 0.92926353279135865, 0.86633165829145731,
0.91781670099350054        ["one - would", 0.92061831133187233, 0.84959349593495936
0.9143083417507526        ["first - one", 0.91727050749090477, 0.84365482233502542,
0.90903394628645784        ["may - well", 0.92092827339961603, 0.85306122448979593,
0.90015031810239354        ["one - part", 0.90442739978532816, 0.82157258064516125,
0.89742767905867249        ["may - time", 0.91183344591858018, 0.83770161290322576,
0.89393870632282468        ["one - upon", 0.89684548994238211, 0.80749746707193515,
0.89053012293395373        ["may - would", 0.90135059121609029, 0.81901840490797551
```

```
In [2]: import json

        sortedSims = []
        with open("similarity5.5","r") as f:
            for line in f.readlines():
```

```
                line = line.strip()
                avg,lisst = line.split("\t")
                lisst = json.loads(lisst)
                lisst.append(avg)
                sortedSims.append(lisst)
```

```
In [3]: !mkdir sims2
        !head -1000 similarity5.5 > sims2/top1000sims
```

```
mkdir: cannot create directory `sims2': File exists
```

```
In [6]: # END STUDENT CODE 5.5
```

```
In [4]: print "\nTop/Bottom 20 results - Similarity measures - sorted by cosine"
        print "(From the entire data set)"
        print '--'*117
        print "{0:>30} |{1:>15} |{2:>15} |{3:>15} |{4:>15} |{5:>15}".format(
                "pair", "cosine", "jaccard", "overlap", "dice", "average")
        print '-'*117

        for stripe in sortedSims[:20]:
            print "{0:>30} |{1:>15f} |{2:>15f} |{3:>15f} |{4:>15f} |{5:>15f}".forma
                stripe[0], float(stripe[1]), float(stripe[2]), float(stripe[3]), fl

        print '--'*117

        for stripe in sortedSims[-20:]:
            print "{0:>30} |{1:>15f} |{2:>15f} |{3:>15f} |{4:>15f} |{5:>15f}".forma
                stripe[0], float(stripe[1]), float(stripe[2]), float(stripe[3]), fl
```

```
Top/Bottom 20 results - Similarity measures - sorted by cosine
(From the entire data set)
-------------------------------------------------------------------------------------
                          pair |         cosine |        jaccard |        overlap |
-------------------------------------------------------------------------------------
                       may - one |       0.952252 |       0.908451 |       0.973060 |
                      one - time |       0.942803 |       0.890467 |       0.981006 |
                      one - well |       0.929264 |       0.866332 |       0.970721 |
                     one - would |       0.920618 |       0.849593 |       0.982374 |
                     first - one |       0.917271 |       0.843655 |       0.981110 |
                      may - well |       0.920928 |       0.853061 |       0.941441 |
                      one - part |       0.904427 |       0.821573 |       0.972554 |
                      may - time |       0.911833 |       0.837702 |       0.928492 |
                      one - upon |       0.896845 |       0.807497 |       0.977914 |
                     may - would |       0.901351 |       0.819018 |       0.941246 |
                     great - one |       0.892185 |       0.798985 |       0.980075 |
                      made - one |       0.891130 |       0.797776 |       0.975278 |
```

time - would	0.901776	0.820647	0.924794
must - one	0.882379	0.782389	0.976010
may - must	0.884164	0.787942	0.957071
one - two	0.880712	0.780020	0.972327
first - time	0.890120	0.801448	0.914994
first - may	0.886557	0.794742	0.927981
may - upon	0.881947	0.785861	0.941104
may - part	0.883368	0.789260	0.929594
glass - serious	0.008862	0.004444	0.009346
husband - thin	0.008701	0.004348	0.009615
captain - cells	0.008461	0.004167	0.010309
arrival - basic	0.008851	0.004444	0.009009
assumption - door	0.008403	0.004132	0.010309
bottom - execution	0.008361	0.004098	0.010417
response - sons	0.007734	0.003390	0.013158
cells - declared	0.008375	0.004132	0.010101
bright - department	0.008480	0.004219	0.009709
asked - characterized	0.008357	0.004149	0.009709
fundamental - stood	0.007945	0.003759	0.011236
fallen - limits	0.008524	0.004274	0.009009
community - threw	0.007767	0.003584	0.011765
basic - glass	0.008548	0.004292	0.008696
characterized - thou	0.008211	0.004065	0.009709
population - window	0.007753	0.003704	0.010638
agreement - evening	0.008015	0.004000	0.008929
created - floor	0.008100	0.004065	0.008333
arms - characterized	0.007579	0.003690	0.009709
bank - severe	0.007536	0.003774	0.008065

Top/Bottom 20 results - Similarity measures - sorted by cosine (From the entire data set)

pair | cosine | jaccard | overlap | dice | average

cons - pros | 0.894427 | 0.800000 | 1.000000 | 0.888889 | 0.895829 forties - twenties | 0.816497 | 0.666667 | 1.000000 | 0.800000 | 0.820791 own - time | 0.809510 | 0.670563 | 0.921168 | 0.802799 | 0.801010 little - time | 0.784197 | 0.630621 | 0.926101 | 0.773473 | 0.778598 found - time | 0.783434 | 0.636364 | 0.883788 | 0.777778 | 0.770341 nova - scotia | 0.774597 | 0.600000 | 1.000000 | 0.750000 | 0.781149 hong - kong | 0.769800 | 0.615385 | 0.888889 | 0.761905 | 0.758995 life - time | 0.769666 | 0.608789 | 0.925081 | 0.756829 | 0.765091 time - world | 0.755476 | 0.585049 | 0.937500 | 0.738209 | 0.754058 means - time | 0.752181 | 0.587117 | 0.902597 | 0.739854 | 0.745437 form - time | 0.749943 | 0.588418 | 0.876733 | 0.740885 | 0.738995 infarction - myocardial | 0.748331 | 0.560000 | 1.000000 | 0.717949 | 0.756570 people - time | 0.745788 | 0.573577 | 0.923875 | 0.729010 | 0.743063 angeles - los | 0.745499 | 0.586207 | 0.850000 | 0.739130 | 0.730209 little - own | 0.739343 | 0.585834 | 0.767296 | 0.738834 | 0.707827 life - own | 0.737053 | 0.582217 | 0.778502 | 0.735951 | 0.708430 anterior

- posterior | 0.733388 | 0.576471 | 0.790323 | 0.731343 | 0.707881 power - time | 0.719611 | 0.533623 | 0.933586 | 0.695898 | 0.720680 dearly - install | 0.707107 | 0.500000 | 1.000000 | 0.666667 | 0.718443 found - own | 0.704802 | 0.544134 | 0.710949 | 0.704776 | 0.666165

---

arrival - essential | 0.008258 | 0.004098 | 0.009615 | 0.008163 | 0.007534 governments - surface | 0.008251 | 0.003534 | 0.014706 | 0.007042 | 0.008383 king - lesions | 0.008178 | 0.003106 | 0.017857 | 0.006192 | 0.008833 clinical - stood | 0.008178 | 0.003831 | 0.011905 | 0.007634 | 0.007887 till - validity | 0.008172 | 0.003367 | 0.015625 | 0.006711 | 0.008469 evidence - started | 0.008159 | 0.003802 | 0.012048 | 0.007576 | 0.007896 forces - record | 0.008152 | 0.003876 | 0.011364 | 0.007722 | 0.007778 primary - stone | 0.008146 | 0.004065 | 0.009091 | 0.008097 | 0.007350 beneath - federal | 0.008134 | 0.004082 | 0.008403 | 0.008130 | 0.007187 factors - rose | 0.008113 | 0.004032 | 0.009346 | 0.008032 | 0.007381 evening - functions | 0.008069 | 0.004049 | 0.008333 | 0.008065 | 0.007129 bone - told | 0.008061 | 0.003704 | 0.012346 | 0.007380 | 0.007873 building - occurs | 0.008002 | 0.003891 | 0.010309 | 0.007752 | 0.007489 company - fig | 0.007913 | 0.003257 | 0.015152 | 0.006494 | 0.008204 chronic - north | 0.007803 | 0.003268 | 0.014493 | 0.006515 | 0.008020 evaluation - king | 0.007650 | 0.003030 | 0.015625 | 0.006042 | 0.008087 resulting - stood | 0.007650 | 0.003663 | 0.010417 | 0.007299 | 0.007257 agent - round | 0.007515 | 0.003289 | 0.012821 | 0.006557 | 0.007546 afterwards - analysis | 0.007387 | 0.003521 | 0.010204 | 0.007018 | 0.007032 posterior - spirit | 0.007156 | 0.002660 | 0.016129 | 0.005305 | 0.007812

3.25 3. HW5.6 Evaluation of synonyms that your discovered

Back to Table of Contents

In this part of the assignment you will evaluate the success of you synonym detector (developed in response to HW5.4). Take the top 1,000 closest/most similar/correlative pairs of words as determined by your measure in HW5.4, 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

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.

3.25.1 Calculate performance measures:

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * (precision * recall)}{precision + recall}$$

We calculate Precision by counting the number of hits and dividing by the number of occurances in our top1000 (opportunities)
We calculate Recall by counting the number of hits, and dividing by the number of synonyms in wordnet (syns)

Other diagnostic measures not implemented here: https://en.wikipedia.org/wiki/F1_score#Diagnostic_Testi

In [5]:
```
''' Performance measures '''
from __future__ import division
import numpy as np
import json
import nltk
from nltk.corpus import wordnet as wn
import sys
import time
import logging

#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()

def evaluate():
    hits = []

    TP = 0
    FP = 0

    TOTAL = 0
    flag = False # so we don't double count, but at the same time don't mis
    start_time = time.time()
    top1000sims = []
    with open("sims2/top1000sims","r") as f:
        for line in f.readlines():

            line = line.strip()
            avg,lisst = line.split("\t")
            lisst = json.loads(lisst)
            lisst.append(avg)
            top1000sims.append(lisst)
```

```
measures = {}
not_in_wordnet = []

for line in top1000sims:
    TOTAL += 1

    pair = line[0]
    words = pair.split(" - ")

    for word in words:
        if word not in measures:
            measures[word] = {"syns":0,"opps": 0,"hits":0}
        measures[word]["opps"] += 1

    syns0 = synonyms(words[0])
    measures[words[1]]["syns"] = len(syns0)
    if len(syns0) == 0:
        not_in_wordnet.append(words[0])

    if words[1] in syns0:
        TP += 1
        hits.append(line)
        flag = True
        measures[words[1]]["hits"] += 1



    syns1 = synonyms(words[1])
    measures[words[0]]["syns"] = len(syns1)
    if len(syns1) == 0:
        not_in_wordnet.append(words[1])

    if words[0] in syns1:
        if flag == False:
            TP += 1
            hits.append(line)
            measures[words[0]]["hits"] += 1

    flag = False

precision = []
recall = []
f1 = []

for key in measures:
    p,r,f = 0,0,0
    if measures[key]["hits"] > 0 and measures[key]["syns"] > 0:
        p = measures[key]["hits"]/measures[key]["opps"]
```

```
                    r = measures[key]["hits"]/measures[key]["syns"]
                    f = 2 * (p*r)/(p+r)

                # For calculating measures, only take into account words that have
                if measures[key]["syns"] > 0:
                    precision.append(p)
                    recall.append(r)
                    f1.append(f)


            # Take the mean of each measure
            print "--"*110
            print "Number of Hits:",TP, "out of top",TOTAL
            print "Number of words without synonyms:",len(not_in_wordnet)
            print "--"*110
            print "Precision\t", np.mean(precision)
            print "Recall\t\t", np.mean(recall)
            print "F1\t\t", np.mean(f1)
            print "--"*110

            print "Words without synonyms:"
            print "-"*100

            for word in not_in_wordnet:
                print synonyms(word),word

            elapsed_time = time.time() - start_time
            mins = elapsed_time/float(60)
            a = """Elapsed time: %s seconds
            In minutes: %s mins""" % (str(elapsed_time), str(mins))
            logging.warning(a)

        evaluate()
```

```
WARNING:root:Elapsed time: 3.32309103012 seconds
    In minutes: 0.055384850502 mins
```

```
--------------------------------------------------------------------------------
Number of Hits: 10 out of top 1000
Number of words without synonyms: 239
--------------------------------------------------------------------------------
Precision        0.00536217841434
Recall                0.0152932994678
F1                0.00608268078938
--------------------------------------------------------------------------------
Words without synonyms:
--------------------------------------------------------------------------------
```

```
[] would
[] upon
[] would
[] would
[] upon
[] could
[] would
[] could
[] upon
[] would
[] would
[] would
[] upon
[] would
[] could
[] would
[] could
[] hong
[] kong
[] would
[] would
[] upon
[] without
[] upon
[] would
[] could
[] would
[] would
[] angeles
[] los
[] could
[] upon
[] would
[] upon
[] without
[] would
[] without
[] upon
[] could
[] could
[] upon
[] would
[] could
[] would
[] would
[] could
[] would
[] without
```

[] would
[] upon
[] would
[] would
[] could
[] would
[] upon
[] could
[] upon
[] would
[] without
[] would
[] francisco
[] san
[] would
[] upon
[] could
[] would
[] upon
[] upon
[] shall
[] upon
[] shall
[] would
[] upon
[] upon
[] could
[] shall
[] could
[] would
[] upon
[] without
[] upon
[] without
[] without
[] upon
[] would
[] could
[] would
[] per
[] could
[] upon
[] upon
[] upon
[] shall
[] without
[] without
[] upon

[] would
[] would
[] upon
[] would
[] upon
[] would
[] upon
[] without
[] could
[] upon
[] would
[] would
[] could
[] could
[] could
[] could
[] would
[] shall
[] could
[] would
[] would
[] could
[] could
[] without
[] would
[] would
[] could
[] among
[] upon
[] would
[] could
[] could
[] would
[] would
[] upon
[] upon
[] would
[] would
[] could
[] shall
[] upon
[] would
[] would
[] shall
[] shall
[] upon
[] would
[] would

[] shall
[] could
[] shall
[] could
[] upon
[] shall
[] would
[] would
[] upon
[] upon
[] upon
[] upon
[] without
[] without
[] could
[] upon
[] would
[] without
[] would
[] upon
[] would
[] among
[] upon
[] would
[] upon
[] could
[] without
[] upon
[] could
[] upon
[] upon
[] upon
[] would
[] without
[] would
[] among
[] upon
[] upon
[] would
[] would
[] without
[] would
[] upon
[] upon
[] among
[] would
[] could
[] upon

[] without
[] would
[] among
[] would
[] would
[] could
[] could
[] could
[] could
[] shall
[] would
[] would
[] would
[] shall
[] would
[] could
[] could
[] shall
[] could
[] could
[] without
[] could
[] would
[] shall
[] among
[] without
[] upon
[] would
[] upon
[] upon
[] among
[] upon
[] without
[] would
[] upon
[] would
[] could
[] without
[] would
[] upon
[] could
[] upon
[] upon
[] upon
[] upon
[] without
[] upon

3.25.2 Sample output

——————————————————————— Number of Hits: 31 out of top 1000 Number of words without synonyms: 67 ————————————————

————————— Precision 0.0280214404967 Recall 0.0178598869579 F1 0.013965517619

——————————————————————— Words without synonyms: ——————————————————————————— [] scotia [] hong [] kong [] angeles [] los [] nor [] themselves [] . . . .

3.26 3. HW5.7 OPTIONAL: using different vocabulary subsets

Back to Table of Contents

Repeat HW5 using vocabulary words ranked from 8001,-10,000; 7001,-10,000; 6001,-10,000; 5001,-10,000; 3001,-10,000; and 1001,-10,000; Dont forget to report you Cluster configuration.

Generate the following graphs: – vocabulary size (X-Axis) versus CPU time for indexing – vocabulary size (X-Axis) versus number of pairs processed – vocabulary size (X-Axis) versus F1 measure, Precision, Recall

3.27 3. HW5.8 OPTIONAL: filter stopwords

Back to Table of Contents

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']

```
In [ ]:
```

3.28 3. HW5.9 OPTIONAL

Back to Table of Contents

There are many good ways to build our synonym detectors, so for this 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.

```
In [ ]:
```

3.29 3. HW5.10 OPTIONAL

Back to Table of Contents

Once again, benchmark your top 10,000 associations (as in 5.5), this time for your results from 5.6. Has your detector improved?

```
In [ ]:
```