MIDS-W261-HW-05-Final_AnthonySpalvieriKruse

October 3, 2016

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

Name: Anthony Spalvieri-Kruse
Class: MIDS w261 Fall 2016 Group 1
Email: ask@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

2 Table of Contents

1 Instructions Back to Table of Contents
MIDS UC Berkeley, Machine Learning at Scale DATSCIW261 ASSIGNMENT #5
Version 2016-09-25
=== INSTRUCTIONS for SUBMISSIONS === Follow the instructions for submissions carefully.
https://docs.google.com/forms/d/1ZOr9RnIe_A06AcZDB6K1mJN4vrLeSmS2PD6Xm3eOiis/viewform?usp=send_form

2.0.1 IMPORTANT

HW4 can be completed locally on your computer

2.0.2 Documents:

- IPython Notebook, published and viewable online.
- PDF export of IPython Notebook.

2 Useful References Back to Table of Contents

- See async and live lectures for this week

HW Problems Back to Table of Contents

2.1 3. HW5.0

Back to Table of Contents

- What is a data warehouse? What is a Star schema? When is it used?

A data warehouse is a central repository for data from various sources, structured specifically for analytics as opposed to transactions like in a standard online transactional processing database. A star schema splits a set of data into facts and dimensions, where facts are the measurable, quantitative data, and dimensions are generally expressed as lookup tables that provided descriptive attributes related to a fact. Star schema's are typically used for data warehouses.

2.2 3. HW5.1

Back to Table of Contents

- In the database world What is 3NF? Does machine learning use data in 3NF? If so why?

```
3NF stands for third normal form, which is a subset of 1st and 2nd normal form.  It's characteristi
```

- In what form does ML consume data?

```
Generally ML uses data in the form of (label, features), which would be best expressed through denc
```

- Why would one use log files that are denormalized?

```
When we denormalize data we're adding redundant information back into a line of data, and this coul
```

2.3 3. HW5.2

Back to Table of Contents

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

Justify which table you chose as the Left table in this hashside join.

Please report the number of rows resulting from:

- (1) Left joining Table Left with Table Right
- (2) Right joining Table Left with Table Right
- (3) Inner joining Table Left with Table Right

In [18]:
```
%%writefile hashside_joins.py
#!/usr/bin/python

from mrjob.job import MRJob
from mrjob.step import MRStep
from collections import defaultdict
import itertools
import re

class HashsideJoin(MRJob):

    def configure_options(self):
        super(HashsideJoin, self).configure_options()
        self.add_passthrough_option("--join_type", type="str")
        self.add_passthrough_option("--right_table_length", type="int")
        self.add_file_option("--left_table")

    def __init__(self, *args, **kwargs):
        super(HashsideJoin, self).__init__(*args, **kwargs)
        self.join_type = self.options.join_type
        self.right_table_length = self.options.right_table_length

    def mapper_init(self):
        self.urlTable = {}
        self.keyMatch = {}
        with open(self.options.left_table, 'r') as f:
            for line in f:
                line = line.strip("\n").split(",")
                pageId = line[1]
                leftTableRow = line[:1] + line[2:]
                self.urlTable[pageId] = leftTableRow
                self.keyMatch[pageId] = False

    #Emit Only matches
    def mapper(self, _, line):
        line = line.strip("\n").split(",")
        pageId = line[1]
        rightTableRow = line[:1]+line[2:]

        if self.join_type == "inner":
```

```
                    if pageId in self.urlTable.keys():
                        value = self.urlTable[pageId] + rightTableRow
                        value = ",".join(value)
                        yield pageId,value
                if self.join_type == "right":
                    #Need to output the rightTableRow no matter what,
                    #i'm either padding with Nulls, or i'm tacking on the key match
                    if pageId in self.urlTable.keys():
                        value = self.urlTable[pageId] + rightTableRow
                        value = ",".join(value)
                    else:
                        value = ["null"]*len(self.urlTable.values()[0]) + rightTableRow
                        value = ",".join(value)
                    yield pageId, value
                if self.join_type == "left":
                    if pageId in self.urlTable.keys():
                        value = self.urlTable[pageId] + rightTableRow
                        value = ",".join(value)
                        self.keyMatch[pageId] = True
                        yield pageId,value

            def mapper_final(self):
                if self.join_type == "left":
                    for key in self.keyMatch.keys():
                        #If there were right table keys matching the left table key
                        if self.keyMatch[key] == False:
                            #Output Null padded rows
                            value = self.urlTable[key] + ["null"]*self.right_table_length
                            value = ",".join(value)
                            yield key, value

            def steps(self):
                return [MRStep(mapper_init=self.mapper_init, mapper=self.mapper, mapper_final=self.mapp

        if __name__=='__main__':
            HashsideJoin.run()

Overwriting hashside_joins.py
```

```
In [24]: !./hashside_joins.py anonymous-msweb-preprocessed.data -r hadoop --right_table_length 4 --join
         !./hashside_joins.py anonymous-msweb-preprocessed.data -r hadoop --right_table_length 4 --join
         !./hashside_joins.py anonymous-msweb-preprocessed.data -r hadoop --right_table_length 4 --join
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/hashside_joins.ask.20161004.000655.879208
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000655.879208/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0323
  Submitted application application_1473978660783_0323
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0323/
  Running job: job_1473978660783_0323
  Job job_1473978660783_0323 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
  Job job_1473978660783_0323 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000655.879208/output
Counters: 30
        File Input Format Counters
                Bytes Read=1756063
        File Output Format Counters
                Bytes Written=5868333
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=259124
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1756441
                HDFS: Number of bytes written=5868333
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=10
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=22821888
                Total time spent by all map tasks (ms)=14858
                Total time spent by all maps in occupied slots (ms)=44574
                Total time spent by all reduces in occupied slots (ms)=0
                Total vcore-milliseconds taken by all map tasks=14858
        Map-Reduce Framework
                CPU time spent (ms)=4270
                Failed Shuffles=0
                GC time elapsed (ms)=59
                Input split bytes=378
                Map input records=98654
                Map output records=98654
                Merged Map outputs=0
                Physical memory (bytes) snapshot=439418880
                Spilled Records=0
                Total committed heap usage (bytes)=1632108544
                Virtual memory (bytes) snapshot=4382457856
Streaming final output from hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000655.879208/output.
```

```
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000655.879208...
Removing temp directory /tmp/hashside_joins.ask.20161004.000655.879208...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/hashside_joins.ask.20161004.000750.987634
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000750.987634/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0324
  Submitted application application_1473978660783_0324
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0324/
  Running job: job_1473978660783_0324
  Job job_1473978660783_0324 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
  Job job_1473978660783_0324 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000750.987634/output
Counters: 30
        File Input Format Counters
                Bytes Read=1756063
        File Output Format Counters
                Bytes Written=5868333
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=259078
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1756441
                HDFS: Number of bytes written=5868333
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=10
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=22917120
                Total time spent by all map tasks (ms)=14920
                Total time spent by all maps in occupied slots (ms)=44760
                Total time spent by all reduces in occupied slots (ms)=0
```

```
                Total vcore-milliseconds taken by all map tasks=14920
        Map-Reduce Framework
                CPU time spent (ms)=3750
                Failed Shuffles=0
                GC time elapsed (ms)=93
                Input split bytes=378
                Map input records=98654
                Map output records=98654
                Merged Map outputs=0
                Physical memory (bytes) snapshot=511139840
                Spilled Records=0
                Total committed heap usage (bytes)=3135242240
                Virtual memory (bytes) snapshot=4389404672
Streaming final output from hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000750.987634/output.
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000750.987634...
Removing temp directory /tmp/hashside_joins.ask.20161004.000750.987634...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/hashside_joins.ask.20161004.000845.768116
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000845.768116/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0325
  Submitted application application_1473978660783_0325
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0325/
  Running job: job_1473978660783_0325
  Job job_1473978660783_0325 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
  Job job_1473978660783_0325 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000845.768116/output
Counters: 30
        File Input Format Counters
                Bytes Read=1756063
        File Output Format Counters
                Bytes Written=5871635
        File System Counters
                FILE: Number of bytes read=0
                FILE: Number of bytes written=259122
                FILE: Number of large read operations=0
```

```
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=1756441
                HDFS: Number of bytes written=5871635
                HDFS: Number of large read operations=0
                HDFS: Number of read operations=10
                HDFS: Number of write operations=4
        Job Counters
                Launched map tasks=2
                Rack-local map tasks=2
                Total megabyte-milliseconds taken by all map tasks=23510016
                Total time spent by all map tasks (ms)=15306
                Total time spent by all maps in occupied slots (ms)=45918
                Total time spent by all reduces in occupied slots (ms)=0
                Total vcore-milliseconds taken by all map tasks=15306
        Map-Reduce Framework
                CPU time spent (ms)=4020
                Failed Shuffles=0
                GC time elapsed (ms)=43
                Input split bytes=378
                Map input records=98654
                Map output records=98704
                Merged Map outputs=0
                Physical memory (bytes) snapshot=497541120
                Spilled Records=0
                Total committed heap usage (bytes)=1693450240
                Virtual memory (bytes) snapshot=4400062464
Streaming final output from hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000845.768116/output.
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/hashside_joins.ask.20161004.000845.768116...
Removing temp directory /tmp/hashside_joins.ask.20161004.000845.768116...
```

In [25]:
```
%%bash

 wc -l inner.txt
 wc -l right.txt
 wc -l left.txt
 printf "\n"
 tail -10 inner.txt
 printf "\n"
 tail -10 right.txt
 printf "\n"
 tail -10 left.txt
```

```
98654 inner.txt
98654 right.txt
98704 left.txt

"1123"        "A,1,\"Germany\",\"/germany\",V,1,C,42708"
"1038"        "A,1,\"SiteBuilder Network Membership\",\"/sbnmember\",V,1,C,42708"
"1026"        "A,1,\"Internet Site Construction for Developers\",\"/sitebuilder\",V,1,C,42708"
"1041"        "A,1,\"Developer Workshop\",\"/workshop\",V,1,C,42708"
"1001"        "A,1,\"Support Desktop\",\"/support\",V,1,C,42709"
"1003"        "A,1,\"Knowledge Base\",\"/kb\",V,1,C,42709"
"1035"        "A,1,\"Windows95 Support\",\"/windowssupport\",V,1,C,42710"
"1001"        "A,1,\"Support Desktop\",\"/support\",V,1,C,42710"
```

```
"1018"        "A,1,\"isapi\",\"/isapi\",V,1,C,42710"
"1008"        "A,1,\"Free Downloads\",\"/msdownload\",V,1,C,42711"

"1123"        "A,1,\"Germany\",\"/germany\",V,1,C,42708"
"1038"        "A,1,\"SiteBuilder Network Membership\",\"/sbnmember\",V,1,C,42708"
"1026"        "A,1,\"Internet Site Construction for Developers\",\"/sitebuilder\",V,1,C,42708"
"1041"        "A,1,\"Developer Workshop\",\"/workshop\",V,1,C,42708"
"1001"        "A,1,\"Support Desktop\",\"/support\",V,1,C,42709"
"1003"        "A,1,\"Knowledge Base\",\"/kb\",V,1,C,42709"
"1035"        "A,1,\"Windows95 Support\",\"/windowssupport\",V,1,C,42710"
"1001"        "A,1,\"Support Desktop\",\"/support\",V,1,C,42710"
"1018"        "A,1,\"isapi\",\"/isapi\",V,1,C,42710"
"1008"        "A,1,\"Free Downloads\",\"/msdownload\",V,1,C,42711"

"1199"        "A,1,\"feedback\",\"/feedback\",null,null,null,null"
"1196"        "A,1,\"ie40\",\"/ie40\",null,null,null,null"
"1290"        "A,1,\"Activate the Internet Conference\",\"/devmovies\",null,null,null,null"
"1291"        "A,1,\"news\",\"/news\",null,null,null,null"
"1297"        "A,1,\"Central America\",\"/centroam\",null,null,null,null"
"1294"        "A,1,\"Bookshelf\",\"/bookshelf\",null,null,null,null"
"1248"        "A,1,\"Softimage \",\"/softimage\",null,null,null,null"
"1287"        "A,1,\"International AutoRoute\",\"/autoroute\",null,null,null,null"
"1289"        "A,1,\"Master Chef Product Information\",\"/masterchef\",null,null,null,null"
"1288"        "A,1,\"library\",\"/library\",null,null,null,null"
```

For this exercise I chose the URL only table as my left table, because it was the smaller of the two and thus the easiest one to store into memory. The inner and right joins have the same number of rows, which makes sense because the set of keys in the customer visit table is a subset of the keys in the url table. This is also why the left join had the greatest number of rows.

2.4 3. HW5.3 Systems tests on n-grams dataset (Phase1) and full experiment (Phase 2)

Back to Table of Contents

2.5 3. HW5.3.0 Run Systems tests locally (PHASE1)

Back to Table of Contents

A large subset of the Google n-grams dataset
https://aws.amazon.com/datasets/google-books-ngrams/
which we have placed in a bucket/folder on Dropbox and on s3:
https://www.dropbox.com/sh/tmqpc4o0xswhkvz/AACUifrl6wrMrlK6a3X3lZ9Ea?dl=0
s3://filtered-5grams/
In particular, this bucket contains (~200) files (10Meg each) in the format:

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

The next cell shows the first 10 lines of the googlebooks-eng-all-5gram-20090715-0-filtered.txt file.

DISCLAIMER: Each record is already a 5-gram. We should calculate the stripes cooccurrence data from the raw text and not from the 5-gram preprocessed data. Calculatating pairs on this 5-gram is a little corrupt as we will be double counting cooccurences. Having said that this exercise can still pull out some simialr terms.

1: unit/systems first-10-lines

```
In [7]: %%writefile googlebooks-eng-all-5gram-20090715-0-filtered-first-10-lines.txt
        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 Childs 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

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

For HW 5.4-5.5, unit test and regression test your code using the followings small test datasets:

- googlebooks-eng-all-5gram-20090715-0-filtered.txt [see above]
- stripe-docs-test [see below]
- atlas-boon-test [see below]

2: unit/systems atlas-boon

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

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

3: unit/systems stripe-docs-test Three terms, A,B,C and their corresponding stripe-docs of co-occurring terms

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

```
In [4]: ##############################################
        # Stripes for systems test 1 (predefined)
        ##############################################

        with open("mini_stripes.txt", "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 mini_stripes.txt

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

2.6 TASK: Phase 1

Complete 5.4 and 5.5 and systems test them using the above test datasets. Phase 2 will focus on the entire Ngram dataset.

To help you through these tasks please verify that your code gives the following results (for stripes, inverted index, and pairwise similarities).

In [9]:
```
%%writefile buildStripes.py
#!/usr/bin/python

from mrjob.job import MRJob
from mrjob.step import MRStep
from collections import defaultdict
#from collections import Counter
import itertools
import re


#Goal: Take in n-gram file and output file w/ structure {Word1: {CoWord1: count1, CoWord2: coun
class BuildStripes(MRJob):

    def combine_dicts(a, b):
        return dict(a.items() + b.items() +
            [(k, a[k] + b[k]) for k in set(b) & set(a)])

    def mapper(self, _, line):
        ngram, count, page, book = line.strip("\n").split("\t")
        words = ngram.split()

        for word in words:
            #2.7 version: {coWord:int(count) for coWord in words if coWord != word}
            stripe = dict((coWord, int(count)) for coWord in words if coWord !=word)
            yield word, stripe

    def combiner(self,word, lines):
        #stripe = dict(reduce(lambda x,y: self.combine_dicts(x,y), line))
        stripe = reduce(lambda x,y: dict(x.items()+y.items()+ [(k, x[k] + y[k]) for k in set(x)
        yield word, stripe

    def reducer(self,word, lines):
        #stripe = dict(reduce(lambda x,y: Counter(x)+Counter(y), line))
        stripe = reduce(lambda x,y: dict(x.items()+y.items()+ [(k, x[k] + y[k]) for k in set(x)
        yield word, stripe

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

if __name__=='__main__':
    BuildStripes.run()
```

```
Overwriting buildStripes.py
```

In [11]:
```
!./buildStripes.py atlas-boon-systems-test.txt -r hadoop > atlasMiniStripesOutput.txt
 !./buildStripes.py googlebooks-eng-all-5gram-20090715-0-filtered-first-10-lines.txt -r hadoop
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.ask.20161004.023307.370251
```

```
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023307.370251/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0326
  Submitted application application_1473978660783_0326
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0326/
  Running job: job_1473978660783_0326
  Job job_1473978660783_0326 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0326 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023307.370251/output
Counters: 49
        File Input Format Counters
                Bytes Read=101
        File Output Format Counters
                Bytes Written=163
        File System Counters
                FILE: Number of bytes read=148
                FILE: Number of bytes written=389733
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=463
                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=20990976
                Total megabyte-milliseconds taken by all reduce tasks=19653120
                Total time spent by all map tasks (ms)=13666
                Total time spent by all maps in occupied slots (ms)=40998
                Total time spent by all reduce tasks (ms)=7677
                Total time spent by all reduces in occupied slots (ms)=38385
```

```
                Total vcore-milliseconds taken by all map tasks=13666
                Total vcore-milliseconds taken by all reduce tasks=7677
        Map-Reduce Framework
                CPU time spent (ms)=4100
                Combine input records=7
                Combine output records=6
                Failed Shuffles=0
                GC time elapsed (ms)=93
                Input split bytes=362
                Map input records=3
                Map output bytes=190
                Map output materialized bytes=168
                Map output records=7
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1728794624
                Reduce input groups=4
                Reduce input records=6
                Reduce output records=4
                Reduce shuffle bytes=168
                Shuffled Maps =2
                Spilled Records=12
                Total committed heap usage (bytes)=2478833664
                Virtual memory (bytes) snapshot=7760158720
        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/ask/tmp/mrjob/buildStripes.ask.20161004.023307.370251/output..
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023307.370251...
Removing temp directory /tmp/buildStripes.ask.20161004.023307.370251...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.ask.20161004.023407.516563
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023407.516563/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0327
```

```
  Submitted application application_1473978660783_0327
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0327/
  Running job: job_1473978660783_0327
  Job job_1473978660783_0327 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0327 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023407.516563/output
Counters: 49
        File Input Format Counters
                Bytes Read=561
        File Output Format Counters
                Bytes Written=2402
        File System Counters
                FILE: Number of bytes read=1045
                FILE: Number of bytes written=391669
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=997
                HDFS: Number of bytes written=2402
                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=20934144
                Total megabyte-milliseconds taken by all reduce tasks=10298880
                Total time spent by all map tasks (ms)=13629
                Total time spent by all maps in occupied slots (ms)=40887
                Total time spent by all reduce tasks (ms)=4023
                Total time spent by all reduces in occupied slots (ms)=20115
                Total vcore-milliseconds taken by all map tasks=13629
                Total vcore-milliseconds taken by all reduce tasks=4023
        Map-Reduce Framework
                CPU time spent (ms)=3920
                Combine input records=50
                Combine output records=31
                Failed Shuffles=0
                GC time elapsed (ms)=72
                Input split bytes=436
                Map input records=10
                Map output bytes=2970
                Map output materialized bytes=1096
                Map output records=50
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1745956864
                Reduce input groups=28
                Reduce input records=31
                Reduce output records=28
                Reduce shuffle bytes=1096
```

```
                Shuffled Maps =2
                Spilled Records=62
                Total committed heap usage (bytes)=2548039680
                Virtual memory (bytes) snapshot=7784456192
        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/ask/tmp/mrjob/buildStripes.ask.20161004.023407.516563/output..
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/buildStripes.ask.20161004.023407.516563...
Removing temp directory /tmp/buildStripes.ask.20161004.023407.516563...
```

In [12]:
```
%%bash

 cat atlasMiniStripesOutput.txt | sort -k1,1
 printf "\n\n"
 cat goog10lineStripes.txt | sort -k1,1
```

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


"A"        {"City": 62, "Tales": 123, "Forms": 116, "in": 1201, "Wales": 1099, "ESTABLISHING": 59, "Chri
"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, "Female": 447, "in"
"Childs"        {"A": 1099, "Wales": 1099, "Christmas": 1099, "in": 1099}
"Christmas"        {"A": 1099, "Wales": 1099, "Childs": 1099, "in": 1099}
"Circumstantial"        {"A": 62, "of": 62, "the": 62, "Narrative": 62}
"City"        {"A": 62, "the": 62, "by": 62, "Sea": 62}
"Collection"        {"A": 239, "of": 239, "Fairy": 123, "Tales": 123, "Forms": 116}
"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, "Childs": 1099, "Government": 102, "Study": 102, "Wales": 1099, "Ch
"Limited"        {"A": 55, "Case": 55, "Study": 55, "of": 55}
"Narrative"        {"A": 62, "of": 62, "the": 62, "Circumstantial": 62}
"of"        {"A": 1011, "Case": 502, "Circumstantial": 62, "Limited": 55, "Study": 502, "Tales": 123, "(
"RELIGIOUS"        {"A": 59, "BILL": 59, "FOR": 59, "ESTABLISHING": 59}
"Sea"        {"A": 62, "City": 62, "the": 62, "by": 62}
"Study"        {"A": 604, "Case": 604, "Government": 102, "of": 502, "Limited": 55, "Female": 447, "in"
"Tales"        {"A": 123, "of": 123, "Fairy": 123, "Collection": 123}
"the"        {"A": 124, "City": 62, "Circumstantial": 62, "of": 62, "Sea": 62, "Narrative": 62, "by": 62
"Wales"        {"A": 1099, "Childs": 1099, "Christmas": 1099, "in": 1099}
```

```
In [11]: ################################################
         # Make Stripes from ngrams for systems test 2
         ################################################

         !aws s3 rm --recursive s3://ucb261-hw5/hw5-4-stripes-mj
         !python buildStripes.py -r emr mini_stripes.txt \
             --cluster-id=j-1YW75NSU09AII \
             --output-dir=s3://ucb261-hw5/hw5-4-stripes-mj \
             --file=stopwords.txt \
             --file=mostFrequent/part-00000 \
         # Output suppressed
```

```
/bin/sh: 332aws: command not found
/bin/sh: 32python: command not found
```

Step 10 Build an cooccureence strips from the atlas-boon

```
In [ ]: #Using the atlas-boon systems test
        atlas boon        50        50        50
        boon cava dipped        10        10        10
        atlas dipped        15        15        15
```

Stripe documents for atlas-boon systems test

```
In [ ]: ################################################
        # Make Stripes from ngrams
        ################################################
        !aws s3 rm --recursive s3://ucb261-hw5/hw5-4-stripes-mj
        !python buildStripes.py -r emr mj_systems_test.txt \
            --cluster-id=j-2WHMJSLZDG \
            --output-dir=s3://ucb261-hw5/hw5-4-stripes-mj \
            --file=stopwords.txt \
            --file=mostFrequent/part-00000 \
        # Output suppressed
```

```
In [ ]: !mkdir stripes-mj
        !aws s3 sync s3://ucb261-hw5/hw5-4-stripes-mj/  stripes-mj/
        !cat stripes-mj/part-*
```

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

2.7 Building stripes execution 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
```

Step 20 create inverted index, and calculate pairwise similarity Solution 1:

Create an Inverted Index.

Use the output to calculate similarities.

Build custom partitioner, re-run the similarity code, and output total order sorted partitions.

In [15]:
```
%%writefile invertedIndex.py
#!/usr/bin/python

from mrjob.job import MRJob
from mrjob.step import MRStep
from collections import defaultdict
#from collections import Counter
import json
import itertools
import re


#Goal: Take in key {strip} file, and output inversion of {word: {doc1: wordsInDoc1, doc2: etc}
class InvertedIndex(MRJob):

    def mapper(self, _, line):
        doc, stripe = line.strip("\n").split("\t")
        stripe = json.loads(stripe)
        stripeLength = len(stripe)

        for word in stripe.keys():
            yield word, {doc.strip('"'): stripeLength}

    def combiner(self,word, lines):
        #A bit overkill because keys won't appear twice, but still combines it
        #stripe = dict(reduce(lambda x,y: Counter(x)+Counter(y), line))
        stripe = reduce(lambda x,y: dict(x.items()+y.items()+ [(k, x[k] + y[k]) for k in set(x
        yield word, stripe

    def reducer(self,word, lines):
        #stripe = dict(reduce(lambda x,y: Counter(x)+Counter(y), line))
        stripe = reduce(lambda x,y: dict(x.items()+y.items()+ [(k, x[k] + y[k]) for k in set(x
        yield word, stripe

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

if __name__=='__main__':
    InvertedIndex.run()
```

```
Overwriting invertedIndex.py
```

In [23]:
```
!./invertedIndex.py mini_stripes.txt -r hadoop > stripesInvertedOutput.txt
!./invertedIndex.py atlasMiniStripesOutput.txt -r hadoop --output-dir hdfs:///user/ask/tmp/mrj
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.ask.20161004.024521.843469
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/invertedIndex.ask.20161004.024521.843469/files/...
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
```

```
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0331
  Submitted application application_1473978660783_0331
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0331/
  Running job: job_1473978660783_0331
  Job job_1473978660783_0331 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0331 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/invertedIndex
Counters: 49
        File Input Format Counters
                Bytes Read=245
        File Output Format Counters
                Bytes Written=153
        File System Counters
                FILE: Number of bytes read=148
                FILE: Number of bytes written=389694
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=607
                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=19872768
                Total megabyte-milliseconds taken by all reduce tasks=18352640
                Total time spent by all map tasks (ms)=12938
                Total time spent by all maps in occupied slots (ms)=38814
                Total time spent by all reduce tasks (ms)=7169
                Total time spent by all reduces in occupied slots (ms)=35845
                Total vcore-milliseconds taken by all map tasks=12938
                Total vcore-milliseconds taken by all reduce tasks=7169
        Map-Reduce Framework
                CPU time spent (ms)=3530
                Combine input records=10
                Combine output records=8
                Failed Shuffles=0
                GC time elapsed (ms)=84
```

```
                Input split bytes=362
                Map input records=4
                Map output bytes=206
                Map output materialized bytes=186
                Map output records=10
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1734430720
                Reduce input groups=4
                Reduce input records=8
                Reduce output records=4
                Reduce shuffle bytes=186
                Shuffled Maps =2
                Spilled Records=16
                Total committed heap usage (bytes)=2548039680
                Virtual memory (bytes) snapshot=7771987968
        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/ask/tmp/mrjob/invertedIndex...
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/invertedIndex.ask.20161004.024521.843469...
Removing temp directory /tmp/invertedIndex.ask.20161004.024521.843469...
```

In [24]:
```
%%bash

cat stripesInvertedOutput.txt | sort -k1,1
printf "\n\n"
hdfs dfs -cat hdfs:///user/ask/tmp/mrjob/invertedIndex/part*
```

```
"M"       {"DocC": 4}
"N"       {"DocC": 4}
"X"       {"DocB": 2, "DocA": 3}
"Y"       {"DocB": 2, "DocC": 4, "DocA": 3}
"Z"       {"DocC": 4, "DocA": 3}


"atlas"       {"dipped": 3, "boon": 3}
"boon"       {"atlas": 2, "dipped": 3, "cava": 2}
"cava"       {"dipped": 3, "boon": 3}
"dipped"       {"atlas": 2, "boon": 3, "cava": 2}
```

2.7.1 Inverted Index

In [ ]:
```
Systems test mini_stripes - 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 |

 systems test atlas-boon - Inverted Index
```

```
"atlas" |         boon 3 |       dipped 3 |
"dipped" |        atlas 2 |         boon 3 |         cava 2
"boon" |         atlas 2 |         cava 2 |       dipped 3
"cava" |          boon 3 |       dipped 3 |
```

2.7.2 Pairwise Similairity

In [50]:
```
%%writefile pairwiseSimilarity.py
#!/usr/bin/python

from mrjob.job import MRJob
from mrjob.step import MRStep
from collections import defaultdict
#from collections import Counter
import json
import itertools
import re
import math


#Goal: Take in key {strip} file, and output inversion of {word: {doc1: wordsInDoc1, doc2: etc}
class PairwiseSimilarity(MRJob):

    #For future reference, if this is too large to store in memory
    #we can hack it.  Tack the union sum onto the end of the sorted key
    #and then parse it all out at the reducer stage
    #unions = defaultdict(int)

    def configure_options(self):
        super(PairwiseSimilarity, self).configure_options()
        self.add_passthrough_option("--similarity_measure", type="str")

    def __init__(self, *args, **kwargs):
        super(PairwiseSimilarity, self).__init__(*args, **kwargs)
        self.similarity_measure = self.options.similarity_measure

    def mapper(self, _, line):
        doc, stripe = line.strip("\n").split("\t")
        stripe = json.loads(stripe)
        stripeLength = len(stripe)

        if self.similarity_measure == "jaccard":
            pairs = map(dict, itertools.combinations(stripe.items(),2))
            for pair in pairs:
                #A hack for sure, but pretty efficient way of storing (A+B) value
                key = sorted(pair.keys()) + [str(sum(pair.values()))] # ",".join(sorted(pair.k
                #self.unions[",".join(key)]=sum(pair.values())
                yield key, 1
        if self.similarity_measure == "cosine":
            pairs = map(dict, itertools.combinations(stripe.items(),2))
            for pair in pairs:
                key = sorted(pair.keys()) # ",".join(sorted(pair.keys()))
                normProduct = reduce(lambda x,y: math.sqrt(x)*math.sqrt(y), pair.values())
```

```
                    yield key, float(1)/normProduct

            def combiner(self,key, values):

                if self.similarity_measure == "jaccard":
                    yield key, sum(values)
                if self.similarity_measure == "cosine":
                    yield key, sum(values)

            def reducer(self,key, values):
                totalCount = sum(values)
                if self.similarity_measure == "jaccard":
                    #similarity = float(totalCount)/(self.unions[",".join(key)] - totalCount) #float(co
                    similarity = float(totalCount)/(int(key[len(key)-1])-totalCount)
                    yield key[:-1], similarity
                if self.similarity_measure == "cosine":
                    yield key, totalCount

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

        if __name__=='__main__':
            PairwiseSimilarity.run()

Overwriting pairwiseSimilarity.py

In [51]: %%bash

        hdfs dfs -rm -r hdfs:///user/ask/tmp/mrjob/jaccardSimilarityStripes
        hdfs dfs -rm -r hdfs:///user/ask/tmp/mrjob/jaccardSimilarityAtlas
        hdfs dfs -rm -r hdfs:///user/ask/tmp/mrjob/cosineSimilarityStripes
        hdfs dfs -rm -r hdfs:///user/ask/tmp/mrjob/cosineSimilarityAtlas

        ./pairwiseSimilarity.py stripesInvertedOutput.txt -r hadoop --output-dir hdfs:///user/ask/tmp/m
        ./pairwiseSimilarity.py stripesInvertedOutput.txt -r hadoop --output-dir hdfs:///user/ask/tmp/m

        ./pairwiseSimilarity.py atlasInvertedOutput.txt -r hadoop --output-dir hdfs:///user/ask/tmp/mr
        ./pairwiseSimilarity.py atlasInvertedOutput.txt -r hadoop --output-dir hdfs:///user/ask/tmp/mr

Moved: 'hdfs://nn-ia.s3s.altiscale.com:8020/user/ask/tmp/mrjob/jaccardSimilarityStripes' to trash at: ho
Moved: 'hdfs://nn-ia.s3s.altiscale.com:8020/user/ask/tmp/mrjob/jaccardSimilarityAtlas' to trash at: hdfs
Moved: 'hdfs://nn-ia.s3s.altiscale.com:8020/user/ask/tmp/mrjob/cosineSimilarityStripes' to trash at: hdf
Moved: 'hdfs://nn-ia.s3s.altiscale.com:8020/user/ask/tmp/mrjob/cosineSimilarityAtlas' to trash at: hdfs

16/10/04 03:44:40 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 5760 mi
16/10/04 03:44:42 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 5760 mi
16/10/04 03:44:44 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 5760 mi
16/10/04 03:44:47 INFO fs.TrashPolicyDefault: Namenode trash configuration: Deletion interval = 5760 mi
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pairwiseSimilarity.ask.20161004.034447.901122
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034447.901122/files/.
Looking for Hadoop streaming jar in /opt/hadoop...
```

```
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0347
  Submitted application application_1473978660783_0347
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0347/
  Running job: job_1473978660783_0347
  Job job_1473978660783_0347 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0347 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/jaccardSimilarityStripes
Counters: 49
        File Input Format Counters
                Bytes Read=186
        File Output Format Counters
                Bytes Written=111
        File System Counters
                FILE: Number of bytes read=76
                FILE: Number of bytes written=390039
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=556
                HDFS: Number of bytes written=111
                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=20143104
                Total megabyte-milliseconds taken by all reduce tasks=17861120
                Total time spent by all map tasks (ms)=13114
                Total time spent by all maps in occupied slots (ms)=39342
                Total time spent by all reduce tasks (ms)=6977
                Total time spent by all reduces in occupied slots (ms)=34885
                Total vcore-milliseconds taken by all map tasks=13114
                Total vcore-milliseconds taken by all reduce tasks=6977
        Map-Reduce Framework
                CPU time spent (ms)=3440
                Combine input records=5
```

```
                Combine output records=4
                Failed Shuffles=0
                GC time elapsed (ms)=101
                Input split bytes=370
                Map input records=5
                Map output bytes=120
                Map output materialized bytes=112
                Map output records=5
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1770520576
                Reduce input groups=3
                Reduce input records=4
                Reduce output records=3
                Reduce shuffle bytes=112
                Shuffled Maps =2
                Spilled Records=8
                Total committed heap usage (bytes)=3649044480
                Virtual memory (bytes) snapshot=7767248896
        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/ask/tmp/mrjob/jaccardSimilarityStripes...
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034447.901122..
Removing temp directory /tmp/pairwiseSimilarity.ask.20161004.034447.901122...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pairwiseSimilarity.ask.20161004.034546.458287
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034546.458287/files/.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0349
  Submitted application application_1473978660783_0349
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0349/
  Running job: job_1473978660783_0349
  Job job_1473978660783_0349 running in uber mode : false
   map 0% reduce 0%
```

```
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0349 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/cosineSimilarityStripes
Counters: 49
        File Input Format Counters
                Bytes Read=186
        File Output Format Counters
                Bytes Written=111
        File System Counters
                FILE: Number of bytes read=109
                FILE: Number of bytes written=390130
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=556
                HDFS: Number of bytes written=111
                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=19998720
                Total megabyte-milliseconds taken by all reduce tasks=17881600
                Total time spent by all map tasks (ms)=13020
                Total time spent by all maps in occupied slots (ms)=39060
                Total time spent by all reduce tasks (ms)=6985
                Total time spent by all reduces in occupied slots (ms)=34925
                Total vcore-milliseconds taken by all map tasks=13020
                Total vcore-milliseconds taken by all reduce tasks=6985
        Map-Reduce Framework
                CPU time spent (ms)=3830
                Combine input records=5
                Combine output records=4
                Failed Shuffles=0
                GC time elapsed (ms)=114
                Input split bytes=370
                Map input records=5
                Map output bytes=185
                Map output materialized bytes=157
                Map output records=5
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1744343040
                Reduce input groups=3
                Reduce input records=4
                Reduce output records=3
                Reduce shuffle bytes=157
                Shuffled Maps =2
                Spilled Records=8
                Total committed heap usage (bytes)=2542796800
                Virtual memory (bytes) snapshot=7745228800
        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/ask/tmp/mrjob/cosineSimilarityStripes...
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034546.458287..
Removing temp directory /tmp/pairwiseSimilarity.ask.20161004.034546.458287...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pairwiseSimilarity.ask.20161004.034644.857016
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034644.857016/files/.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0351
  Submitted application application_1473978660783_0351
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0351/
  Running job: job_1473978660783_0351
  Job job_1473978660783_0351 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0351 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/jaccardSimilarityAtlas
Counters: 49
        File Input Format Counters
                Bytes Read=230
        File Output Format Counters
                Bytes Written=139
        File System Counters
                FILE: Number of bytes read=117
                FILE: Number of bytes written=390160
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=596
                HDFS: Number of bytes written=139
                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=20328960
                Total megabyte-milliseconds taken by all reduce tasks=17635840
                Total time spent by all map tasks (ms)=13235
                Total time spent by all maps in occupied slots (ms)=39705
                Total time spent by all reduce tasks (ms)=6889
                Total time spent by all reduces in occupied slots (ms)=34445
                Total vcore-milliseconds taken by all map tasks=13235
                Total vcore-milliseconds taken by all reduce tasks=6889
        Map-Reduce Framework
                CPU time spent (ms)=4460
                Combine input records=8
                Combine output records=8
                Failed Shuffles=0
                GC time elapsed (ms)=144
                Input split bytes=366
                Map input records=4
                Map output bytes=204
                Map output materialized bytes=179
                Map output records=8
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1697161216
                Reduce input groups=6
                Reduce input records=8
                Reduce output records=6
                Reduce shuffle bytes=179
                Shuffled Maps =2
                Spilled Records=16
                Total committed heap usage (bytes)=2478833664
                Virtual memory (bytes) snapshot=7771037696
        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/ask/tmp/mrjob/jaccardSimilarityAtlas...
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034644.857016..
Removing temp directory /tmp/pairwiseSimilarity.ask.20161004.034644.857016...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/pairwiseSimilarity.ask.20161004.034743.709730
Looking for hadoop binary in /opt/hadoop/bin...
Found hadoop binary: /opt/hadoop/bin/hadoop
Using Hadoop version 2.7.2
Copying local files to hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034743.709730/files/.
Looking for Hadoop streaming jar in /opt/hadoop...
Found Hadoop streaming jar: /opt/hadoop/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar
Running step 1 of 1...
  packageJobJar: [] [/opt/hadoop-2.7.2/share/hadoop/tools/lib/hadoop-streaming-2.7.2.jar] /tmp/streamjo
```

```
  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/10.251.255.108:10200
  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/10.251.255.108:10200
  Loaded native gpl library from the embedded binaries
  Successfully loaded & initialized native-lzo library [hadoop-lzo rev d62701d4d05dfa6115bbaf8d9dff002d
  Total input paths to process : 1
  number of splits:2
  Submitting tokens for job: job_1473978660783_0352
  Submitted application application_1473978660783_0352
  The url to track the job: http://rm-ia.s3s.altiscale.com:8088/proxy/application_1473978660783_0352/
  Running job: job_1473978660783_0352
  Job job_1473978660783_0352 running in uber mode : false
   map 0% reduce 0%
   map 100% reduce 0%
   map 100% reduce 100%
  Job job_1473978660783_0352 completed successfully
  Output directory: hdfs:///user/ask/tmp/mrjob/cosineSimilarityAtlas
Counters: 49
        File Input Format Counters
                Bytes Read=230
        File Output Format Counters
                Bytes Written=231
        File System Counters
                FILE: Number of bytes read=149
                FILE: Number of bytes written=390226
                FILE: Number of large read operations=0
                FILE: Number of read operations=0
                FILE: Number of write operations=0
                HDFS: Number of bytes read=596
                HDFS: Number of bytes written=231
                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=11844096
                Total megabyte-milliseconds taken by all reduce tasks=17548800
                Total time spent by all map tasks (ms)=7711
                Total time spent by all maps in occupied slots (ms)=23133
                Total time spent by all reduce tasks (ms)=6855
                Total time spent by all reduces in occupied slots (ms)=34275
                Total vcore-milliseconds taken by all map tasks=7711
                Total vcore-milliseconds taken by all reduce tasks=6855
        Map-Reduce Framework
                CPU time spent (ms)=4090
                Combine input records=8
                Combine output records=8
                Failed Shuffles=0
                GC time elapsed (ms)=75
```

```
                Input split bytes=366
                Map input records=4
                Map output bytes=308
                Map output materialized bytes=225
                Map output records=8
                Merged Map outputs=2
                Physical memory (bytes) snapshot=1716875264
                Reduce input groups=6
                Reduce input records=8
                Reduce output records=6
                Reduce shuffle bytes=225
                Shuffled Maps =2
                Spilled Records=16
                Total committed heap usage (bytes)=2481455104
                Virtual memory (bytes) snapshot=7745163264
        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/ask/tmp/mrjob/cosineSimilarityAtlas...
Removing HDFS temp directory hdfs:///user/ask/tmp/mrjob/pairwiseSimilarity.ask.20161004.034743.709730..
Removing temp directory /tmp/pairwiseSimilarity.ask.20161004.034743.709730...
```

In [52]:
```
%%bash

 printf "Jaccard Similarity Measure\n\n"
 cat jaccardAtlSimilarity.txt
 printf "\n\n"
 cat jaccardSimilarity.txt
 printf "\n\nCosine Similarity Measure\n\n"
 cat cosineAtlSimilarity.txt
 printf "\n\n"
 cat cosineSimilarity.txt
```

```
Jaccard Similarity Measure

["atlas", "boon"]        0.25
["atlas", "cava"]        1.0
["atlas", "dipped"]        0.25
["boon", "cava"]        0.25
["boon", "dipped"]        0.5
["cava", "dipped"]        0.25


["DocA", "DocB"]        0.66666666666666663
["DocA", "DocC"]        0.40000000000000002
["DocB", "DocC"]        0.20000000000000001


Cosine Similarity Measure

["atlas", "boon"]        0.40824829046386296
```

```
["atlas", "cava"]        0.99999999999999978
["atlas", "dipped"]        0.40824829046386296
["boon", "cava"]        0.40824829046386296
["boon", "dipped"]        0.66666666666666674
["cava", "dipped"]        0.40824829046386296


["DocA", "DocB"]        0.81649658092772592
["DocA", "DocC"]        0.57735026918962584
["DocB", "DocC"]        0.35355339059327373
```

```
In [53]: %%bash

         hdfs dfs -cat hdfs:///user/ask/tmp/mrjob/jaccardSimilarityStripes/part*
         printf "\n"
         hdfs dfs -cat hdfs:///user/ask/tmp/mrjob/jaccardSimilarityAtlas/part*
         printf "\n"
         hdfs dfs -cat hdfs:///user/ask/tmp/mrjob/cosineSimilarityStripes/part*
         printf "\n"
         hdfs dfs -cat hdfs:///user/ask/tmp/mrjob/cosineSimilarityAtlas/part*
```

```
["DocA", "DocB"]        0.66666666666666663
["DocA", "DocC"]        0.40000000000000002
["DocB", "DocC"]        0.20000000000000001

["atlas", "boon"]        0.25
["atlas", "cava"]        1.0
["atlas", "dipped"]        0.25
["boon", "cava"]        0.25
["boon", "dipped"]        0.5
["cava", "dipped"]        0.25

["DocA", "DocB"]        0.81649658092772592
["DocA", "DocC"]        0.57735026918962584
["DocB", "DocC"]        0.35355339059327373

["atlas", "boon"]        0.40824829046386296
["atlas", "cava"]        0.99999999999999978
["atlas", "dipped"]        0.40824829046386296
["boon", "cava"]        0.40824829046386296
["boon", "dipped"]        0.66666666666666674
["cava", "dipped"]        0.40824829046386296
```

3 Calculations By Hand

Jaccard Scratch Notes:

```
docA & docB = {x + y}
docA | docB = {x + y + z}
A&B/A|B = .6666

docA & docC = {z + y}
docA | docC = {x + y + z+ m + n}
A&C/A|C = .4
```

```
docB & docC = {y}
docB | docC = {y + x + z + n + m}
C&B/C|B = .2

So jaccard literally is just using binary counts.

For each line i can spit out the combinations of docs/words, and also the sum of their attached counts

In the reduce phase we effectively get a free look at A&B, and we include A+B in the tuple, so then its

For words/docs that have no co-occurrence I would need to keep a set in memory of every term/document, a
```

Cosine Notes:

```
docA*docB = {1*1 + 1*1 + 1*0} = 2
|docA||docB| = 1/sqrt(2) * 1/sqrt(3) = 1/sqrt(6)
A&B/A|B = .81

docA*docC = {1*0 + 1*1 + 1*1 + 1*0 + 1*0} = 2
docA | docC = 1/sqrt(3) * 1/sqrt(4) = 1/sqrt(12)
A&C/A|C = .57

docB*docC = {1*0 + 1*1 + 1*0 + 1*0 + 1*0} = 1
docB | docC = 1/sqrt(2) * 1/sqrt(4) = 1/sqrt(8)
C&B/C|B = .35
```

In [ ]:
```
Systems test mini_stripes - 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 |

 systems test atlas-boon - Inverted Index
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
      "atlas" |         boon 3 |       dipped 3 |
     "dipped" |        atlas 2 |         boon 3 |         cava 2
       "boon" |        atlas 2 |         cava 2 |       dipped 3
       "cava" |         boon 3 |       dipped 3 |

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

In [ ]:
```
Systems test mini_stripes - Similarity measures
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
      average |           pair |         cosine |        jaccard |        overlap |
------------------------------------------------------------------------------------------
     0.741582 |    DocA - DocB |       0.816497 |       0.666667 |       1.000000 |       0.8
     0.488675 |    DocA - DocC |       0.577350 |       0.400000 |       0.666667 |       0.5
     0.276777 |    DocB - DocC |       0.353553 |       0.200000 |       0.500000 |       0.3
------------------------------------------------------------------------------------------
Systems test atlas-boon 2 - Similarity measures
```

```
        average |           pair |         cosine |        jaccard |        overlap |
-------------------------------------------------------------------------------------
       1.000000 |    atlas - cava |       1.000000 |       1.000000 |       1.000000 |       1.0
       0.625000 |   boon - dipped |       0.666667 |       0.500000 |       0.666667 |       0.6
       0.389562 |   cava - dipped |       0.408248 |       0.250000 |       0.500000 |       0.4
       0.389562 |     boon - cava |       0.408248 |       0.250000 |       0.500000 |       0.4
       0.389562 | atlas - dipped |       0.408248 |       0.250000 |       0.500000 |       0.4
       0.389562 |    atlas - boon |       0.408248 |       0.250000 |       0.500000 |       0.4
-------------------------------------------------------------------------------------
```

3.1 3. HW5.3.1 Run systems tests on the CLOUD (PHASE 1)

Back to Table of Contents

Repeat HW5.3.0 on the cloud (AltaScale / AWS/ SoftLayer/ Azure). Make sure all tests give correct results

4 PHASE 2: Full-scale experiment on Google N-gram data

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

4.1 3. HW5.3.2 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.,

- Longest 5-gram (number of characters)
- Top 10 most frequent words (please use the count information), i.e., unigrams
- 20 Most/Least densely appearing words (count/pages_count) sorted in decreasing order of relative frequency
- 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.

In [22]:
```
%%writefile ngramEDA.py
#!/usr/bin/env python

from mrjob.job import MRJob
from mrjob.step import MRStep
from collections import defaultdict
import itertools
import re

class NgramEDA(MRJob):



    def configure_options(self):
        super(NgramEDA, self).configure_options()
        self.add_passthrough_option("--feature_type", type="str")
        self.add_passthrough_option("--topN", type="int")

    def __init__(self, *args, **kwargs):
```

```
        super(NgramEDA, self).__init__(*args, **kwargs)
        self.feature_type = self.options.feature_type
        self.topN = self.options.topN
        self.ngram = ["nada" for i in range(self.topN)]
        self.frequencies = [0 for i in range(self.topN)]

    def mapper(self, key, line):
        title, count, pages, books = line.strip("\n").split("\t")
        words = title.split()
        numChar = len(title)

        if self.feature_type == "length":
            yield None, numChar
        if self.feature_type == "frequency":
            for word in words:
                yield word, int(count)
        if self.feature_type == "density":
            for word in words:
                yield word, (int(count),int(pages))
        if self.feature_type == "distribution":
            yield str(numChar), 1

    def reducer(self, key, counts):
        if self.feature_type == "length":
            yield "Max Length", max(counts)
        if self.feature_type == "frequency":
            total = sum(counts)
            ix = -1
            for i in range(len(self.frequencies)):
                if total > self.frequencies[i]:
                    ix = i
                else:
                    break
            if ix >= 0:
                self.frequencies.insert(ix+1,total)
                self.ngram.insert(ix+1,key)
                self.frequencies = self.frequencies[1:(1+len(self.frequencies))]
                self.ngram = self.ngram[1:(1+len(self.frequencies))]
            #yield key, total
        if self.feature_type == "density":
            count, pages = map(sum,zip(*counts))
            yield key, float(count)/pages
        if self.feature_type == "distribution":
            yield key, sum(counts)

    def reducer_final(self):
        if self.feature_type == "frequency":
            self.frequencies.reverse()
            self.ngram.reverse()
            print "The top 10000 pages are:"
            for i in range(self.topN):
                yield self.ngram[i] , self.frequencies[i]

    def steps(self):
```

```
                return [MRStep(mapper=self.mapper, reducer=self.reducer, reducer_final=self.reducer_fi

        if __name__=='__main__':
            NgramEDA.run()
```

```
Overwriting ngramEDA.py
```

```
In [23]: !./ngramEDA.py google5gram0Top10.txt --feature_type "length" --topN 20 > top10Length.txt
         !./ngramEDA.py google5gram0Top10.txt --jobconf mapred.reduce.tasks=1 --feature_type "frequency
         !./ngramEDA.py google5gram0Top10.txt --feature_type "density" --topN 20  > top10Density.txt
         !./ngramEDA.py google5gram0Top10.txt --feature_type "distribution" --topN 20  > top10Distributi
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/ngramEDA.cloudera.20161002.171514.216981
Running step 1 of 1...
Streaming final output from /tmp/ngramEDA.cloudera.20161002.171514.216981/output...
Removing temp directory /tmp/ngramEDA.cloudera.20161002.171514.216981...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/ngramEDA.cloudera.20161002.171514.696650
Running step 1 of 1...
Streaming final output from /tmp/ngramEDA.cloudera.20161002.171514.696650/output...
Removing temp directory /tmp/ngramEDA.cloudera.20161002.171514.696650...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/ngramEDA.cloudera.20161002.171515.068401
Running step 1 of 1...
Streaming final output from /tmp/ngramEDA.cloudera.20161002.171515.068401/output...
Removing temp directory /tmp/ngramEDA.cloudera.20161002.171515.068401...
No configs found; falling back on auto-configuration
Creating temp directory /tmp/ngramEDA.cloudera.20161002.171515.479685
Running step 1 of 1...
Streaming final output from /tmp/ngramEDA.cloudera.20161002.171515.479685/output...
Removing temp directory /tmp/ngramEDA.cloudera.20161002.171515.479685...
```

```
In [ ]:
```

4.2 3. HW5.4 Synonym detection over 2Gig of Data

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']

4.2.1 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!

```
In [ ]:
```

```
In [ ]: 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}".format(
        stripe[0], float(stripe[1]), float(stripe[2]), float(stripe[3]), float(stripe[4]), floa

print '|'*117

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

In [ ]:
```
Top/Bottom 20 results - Similarity measures - sorted by cosine
(From the entire data set)
```

pair	cosine	jaccard	overlap	di
cons - pros	0.894427	0.800000	1.000000	0.888
forties - twenties	0.816497	0.666667	1.000000	0.800
own - time	0.809510	0.670563	0.921168	0.802
little - time	0.784197	0.630621	0.926101	0.773
found - time	0.783434	0.636364	0.883788	0.777
nova - scotia	0.774597	0.600000	1.000000	0.750
hong - kong	0.769800	0.615385	0.888889	0.761
life - time	0.769666	0.608789	0.925081	0.756
time - world	0.755476	0.585049	0.937500	0.738
means - time	0.752181	0.587117	0.902597	0.739
form - time	0.749943	0.588418	0.876733	0.740
infarction - myocardial	0.748331	0.560000	1.000000	0.717
people - time	0.745788	0.573577	0.923875	0.729
angeles - los	0.745499	0.586207	0.850000	0.739
little - own	0.739343	0.585834	0.767296	0.738
life - own	0.737053	0.582217	0.778502	0.735
anterior - posterior	0.733388	0.576471	0.790323	0.731
power - time	0.719611	0.533623	0.933586	0.695
dearly - install	0.707107	0.500000	1.000000	0.666
found - own	0.704802	0.544134	0.710949	0.704
arrival - essential	0.008258	0.004098	0.009615	0.008
governments - surface	0.008251	0.003534	0.014706	0.007
king - lesions	0.008178	0.003106	0.017857	0.006
clinical - stood	0.008178	0.003831	0.011905	0.007
till - validity	0.008172	0.003367	0.015625	0.006
evidence - started	0.008159	0.003802	0.012048	0.007
forces - record	0.008152	0.003876	0.011364	0.007
primary - stone	0.008146	0.004065	0.009091	0.008
beneath - federal	0.008134	0.004082	0.008403	0.008
factors - rose	0.008113	0.004032	0.009346	0.008
evening - functions	0.008069	0.004049	0.008333	0.008
bone - told	0.008061	0.003704	0.012346	0.007
building - occurs	0.008002	0.003891	0.010309	0.007
company - fig	0.007913	0.003257	0.015152	0.006

chronic - north	0.007803	0.003268	0.014493	0.006
evaluation - king	0.007650	0.003030	0.015625	0.006
resulting - stood	0.007650	0.003663	0.010417	0.007
agent - round	0.007515	0.003289	0.012821	0.006
afterwards - analysis	0.007387	0.003521	0.010204	0.007
posterior - spirit	0.007156	0.002660	0.016129	0.005

4.3 3. HW5.5 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.

4.3.1 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_Testing

In [ ]:
```
nltk.download()
```

In [61]:
```
''' Performance measures '''
#Partial-Author: Anthony Spalvieri-Kruse
#I modified this script and used it for my synonym analysis
from __future__ import division
import numpy as np
import json
import nltk
from nltk.corpus import wordnet as wn
import sys
import re

#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()
hits = []

TP = 0
FP = 0

TOTAL = 0
flag = False # so we don't double count, but at the same time don't miss hits

## For this part we can use one of three outputs. They are all the same, but were generated di
# 1. the top 1000 from the full sorted dataset -> sortedSims[:1000]
# 2. the top 1000 from the partial sort aggragate file -> sims2/top1000sims
# 3. the top 1000 from the total order sort file -> head -1000 sims_parts/part-00004

f1 = open("jaccardAtlSimilarity.txt","r")
f2 = open("jaccardSimilarity.txt","r")
f3 = open("cosineAtlSimilarity.txt","r")
f4 = open("cosineSimilarity.txt", "r")

f1 = f1.readlines()
f2 = f2.readlines()
f3 = f3.readlines()
f4 = f4.readlines()

f1 = [i.strip("\n").split("\t") for i in f1]
f2 = [i.strip("\n").split("\t") for i in f2]
f3 = [i.strip("\n").split("\t") for i in f3]
f4 = [i.strip("\n").split("\t") for i in f4]

top1000sims = f1+f2+f3+f4
#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

    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])
        print words
        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 synonyms in wordnet
    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
```

```
["atlas", "boon"]
["atlas", "cava"]
["atlas", "dipped"]
["boon", "cava"]
["boon", "dipped"]
["cava", "dipped"]
["DocA", "DocB"]
["DocA", "DocC"]
["DocB", "DocC"]
["atlas", "boon"]
["atlas", "cava"]
["atlas", "dipped"]
["boon", "cava"]
["boon", "dipped"]
["cava", "dipped"]
["DocA", "DocB"]
["DocA", "DocC"]
["DocB", "DocC"]
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Number of Hits: 0 out of top 18
Number of words without synonyms: 36
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Precision        nan
Recall                nan
F1                nan
||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Words without synonyms:
----------------------------------------------------------------------------------------------------
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
```

```
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
[] [
[] "
```

```
/opt/anaconda2/envs/py27/lib/python2.7/site-packages/numpy/core/_methods.py:59: RuntimeWarning: Mean of
  warnings.warn("Mean of empty slice.", RuntimeWarning)
/opt/anaconda2/envs/py27/lib/python2.7/site-packages/numpy/core/_methods.py:70: RuntimeWarning: invalid
  ret = ret.dtype.type(ret / rcount)
```

4.3.2 Sample output

In [ ]:

```
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
[]
.......
```

In [ ]:

In [ ]: