MIDS - w261 Machine Learning At Scale

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

Assignment - HW5

Name: Chris Caudill
Class: MIDS w261 (Section Spring 2017 Group 1)
Email: cscaudill@iSchool.Berkeley.edu
StudentId 3032134574 **End of StudentId**
Week: 5

NOTE: please replace `1234567` with your student id above
Due Time: HW is due the Tuesday of the following week by 8AM (West coast time). I.e., Tuesday, Feb 14, 2017 in the case of this homework.

- **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, Feb 21 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

Table of Contents

1 Instructions

Back to Table of Contents

MIDS UC Berkeley, Machine Learning at Scale
DATSCIW261 ASSIGNMENT #5

Version 2017-9-2

IMPORTANT

This homework can be completed locally on your computer

=== INSTRUCTIONS for SUBMISSIONS ===

Follow the instructions for submissions carefully.

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

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

Push the following to your HW github repo into the master branch:

- Your local HW5 directory. Your repo file structure should look like this:

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

2 Useful References

Back to Table of Contents

- See async and live lectures for this week

HW Problems

Back to Table of Contents

3. HW5.0 data warehouse; star schema

Back to Table of Contents

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

A Data Warehouse is a repository of data from a variety of sources and is typically used in an enterprise setting as the basis for Business Intelligence and other types of data analysis. DWs can contain a range of data from structured to unstructured.

Star Schema is an approach used in Data Warehouses where a fact table is used as a compound reference of multiple dimension tables. This is used to create a single table that can represent the relation between two dimension tables. For example, if we own a business with multiple locations, our data repository may contain a table for product inventory, customers, and store location. Each of these 3 tables would be dimensions. The fact table would contain foreign keys that reference the associated product/customer/store from their respective table.

3. HW5.1 Databases: 3NF; denormalized

Back to Table of Contents

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

- 3NF (Third Normal Form) is format that data is typically stored within a database where data from multiple tables is represented in smaller tables by their keys. 3NF does not usually contain all of the fields from the referenced tables and therefore in Machine Learning, we will often need to denormalize the data to get the complete picture. If the 3NF table contains all of the necessary information, it may be used, but this is not always the case.
- ML can technically consume data in 1st, 2nd, or 3rd normal form. As mentioned above, if the normalized data contains the required content, denormalization may not be necessary. In an ideal scenario, ML would have all of the required content within each record (3NF), but due to technologies such as Hadoop, we can perform massive joins much more efficiently than we could with traditional databases.
- Denormalized log files are benefical because a typical log file can have a mass amount of rows. If we were required to reference other tables for additional data associated with each line item, we would be much less efficient than if we were to simply have all of the required content within each line.

3. HW5.2 Memory-backed map-side

Back to Table of Contents

Using MRJob, implement a hashside join (memory-backed map-side) for left, right and inner joins. Use the following tables for this HW and join based on the country code (third column of the transactions table and the second column of the Countries table:

```
transactions.dat
Alice Bob|$10|US
Sam Sneed|$1|CA
Jon Sneed|$20|CA
Arnold Wesise|$400|UK
Henry Bob|$2|US
Yo Yo Ma|$2|CA
Jon York|$44|CA
Alex Ball|$5|UK
Jim Davis|$66|JA

Countries.dat
United States|US
Canada|CA
United Kingdom|UK
Italy|IT
```

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 [251]:
```
%%writefile transactions.txt
Alice Bob|$10|US
Sam Sneed|$1|CA
Jon Sneed|$20|CA
Arnold Wesise|$400|UK
Henry Bob|$2|US
Yo Yo Ma|$2|CA
Jon York|$44|CA
Alex Ball|$5|UK
Jim Davis|$66|JA
```

```
Overwriting transactions.txt
```

In [252]:
```
%%writefile countries.txt
United States|US
Canada|CA
United Kingdom|UK
Italy|IT
```

```
Overwriting countries.txt
```

For the LEFT join, we output all of the rows from the TRANSACTIONS file, regardless of whether they have a match in the COUNTRIES file. If there is a match, we join the rows.

The Reason that I chose to use the TRANSACTIONS file as my LEFT table is because it logically made more sense. The table on the left will output ALL of its rows regardless of a match in the RIGHT table or not. In the scenario where we have a TRANSACTIONS file and a COUNTRIES file, it seemed to me that a typical use case would be more likely to be interested in joining the long description of the country onto each transaction, as opposed to the other way around.

Also, I chose to load my COUNTRIES file into memory as it was the smaller of the two tables.

In [14]:
```
%%writefile leftHashJoin.py
#!/usr/bin/env python

from mrjob.job import MRJob
from mrjob.step import MRStep
from mrjob.compat import jobconf_from_env
import os

class leftJoin(MRJob):

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

    def mapper_init(self):
        countries = open(str(os.path.dirname(os.path.realpath(__file__)))+"/
        self.left=0
        self.nomatch=0
        self.ct = {}
        for line in countries:
            ct_long, ct_short  = line.split('|',1)
            self.ct[ct_short.strip()] = ct_long.strip()

    def mapper(self, _, line):
        cust,price,country = line.split('|',2)

        # In this case, we are left joining with the transactions file on tl
        # Therefore, we will output 9 rows regardless of how many matches sl

        if country in self.ct:
            self.left += 1
            yield None, country+","+self.ct[country]+","+cust+","+price
        else:
            self.nomatch += 1
            yield None, country+",NULL,"+cust+","+price


    # Mapper_Final will output the total joined rows
    def mapper_final(self):
        yield None, "left-joined "+str(self.left)+" row(s)."
        yield None, "No left match for "+str(self.nomatch)+" row(s)."
        yield None, "Total Rows: "+str(self.left + self.nomatch)


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

```
Overwriting leftHashJoin.py
```

For the RIGHT join, we output all of the rows from the COUNTRIES file, whether or not there is a match in the TRANSACTIONS file. If there is a match, we join them.

In [15]:

```
%%writefile rightHashJoin.py
#!/usr/bin/env python

from mrjob.job import MRJob
from mrjob.step import MRStep
from mrjob.compat import jobconf_from_env
import os

class rightJoin(MRJob):

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

    def mapper_init(self):
        countries = open(str(os.path.dirname(os.path.realpath(__file__)))+"/
        self.right=0
        self.nomatch=0
        self.ct = {}
        self.matched_ct={}
        for line in countries:
            ct_long, ct_short  = line.split('|',1)
            self.ct[ct_short.strip()] = ct_long.strip()
#             self.matched_ct[ct_short.strip()] = ct_long.strip()

    def mapper(self, _, line):
        cust,price,country = line.split('|',2)

        # In this case, we are right joining with the transactions file on t
        # Therefore, we will output all of the rows from the countries table
        # from the transactions table that have a country match

        if country in self.ct:
            self.right += 1
            yield None, country+","+self.ct[country]+","+cust+","+price
            self.matched_ct[country] = self.ct[country]



    # Mapper_Final will output the total joined rows
    def mapper_final(self):
        for key,value in self.ct.iteritems():
            if key not in self.matched_ct:
                self.nomatch += 1
                yield None, key+","+value+",NULL,NULL"
        yield None, "right-joined "+str(self.right)+" row(s)."
        yield None, "No right match for "+str(self.nomatch)+" row(s)."
        yield None, "Total Rows: "+str(self.right + self.nomatch)

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

```
Overwriting rightHashJoin.py
```

For the INNER join, we output all of the records that match between the two files.

In [16]:
```
%%writefile innerHashJoin.py
#!/usr/bin/env python

from mrjob.job import MRJob
from mrjob.step import MRStep
from mrjob.compat import jobconf_from_env
import os

class innerJoin(MRJob):

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

    def mapper_init(self):
        countries = open(str(os.path.dirname(os.path.realpath(__file__)))+"/
        self.inner=0
        self.ct = {}
        for line in countries:
            ct_long, ct_short  = line.split('|',1)
            self.ct[ct_short.strip()] = ct_long.strip()

    def mapper(self, _, line):
        cust,price,country = line.split('|',2)
        if country in self.ct:
            self.inner += 1
            yield None, self.ct[country]+","+cust+","+price

    def mapper_final(self):
        yield None, "inner-joined "+str(self.inner)+" rows."


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

```
Overwriting innerHashJoin.py
```

In [17]:

```
!chmod a+x leftHashJoin.py
!./leftHashJoin.py --jobconf mapred.map.tasks=1 transactions.txt
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/leftHashJoin.root.20170214.034404.240845
Running step 1 of 1...
Streaming final output from /tmp/leftHashJoin.root.20170214.034404.24084
5/output...
null    "US,United States,Alice Bob,$10"
null    "CA,Canada,Sam Sneed,$1"
null    "CA,Canada,Jon Sneed,$20"
null    "UK,United Kingdom,Arnold Wesise,$400"
null    "US,United States,Henry Bob,$2"
null    "CA,Canada,Yo Yo Ma,$2"
null    "CA,Canada,Jon York,$44"
null    "UK,United Kingdom,Alex Ball,$5"
null    "JA,NULL,Jim Davis,$66"
null    "left-joined 8 row(s)."
null    "No left match for 1 row(s)."
null    "Total Rows: 9"
Removing temp directory /tmp/leftHashJoin.root.20170214.034404.240845...
```

In [18]:

```
!chmod a+x rightHashJoin.py
!./rightHashJoin.py --jobconf mapred.map.tasks=1 transactions.txt
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/rightHashJoin.root.20170214.034405.650385
Running step 1 of 1...
Streaming final output from /tmp/rightHashJoin.root.20170214.034405.65038
5/output...
null    "US,United States,Alice Bob,$10"
null    "CA,Canada,Sam Sneed,$1"
null    "CA,Canada,Jon Sneed,$20"
null    "UK,United Kingdom,Arnold Wesise,$400"
null    "US,United States,Henry Bob,$2"
null    "CA,Canada,Yo Yo Ma,$2"
null    "CA,Canada,Jon York,$44"
null    "UK,United Kingdom,Alex Ball,$5"
null    "IT,Italy,NULL,NULL"
null    "right-joined 8 row(s)."
null    "No right match for 1 row(s)."
null    "Total Rows: 9"
Removing temp directory /tmp/rightHashJoin.root.20170214.034405.650385...
```

```
In [19]: !chmod a+x innerHashJoin.py
         !./innerHashJoin.py --jobconf mapred.map.tasks=1 transactions.txt
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/innerHashJoin.root.20170214.034407.159333
Running step 1 of 1...
Streaming final output from /tmp/innerHashJoin.root.20170214.034407.15933
3/output...
null    "United States,Alice Bob,$10"
null    "Canada,Sam Sneed,$1"
null    "Canada,Jon Sneed,$20"
null    "United Kingdom,Arnold Wesise,$400"
null    "United States,Henry Bob,$2"
null    "Canada,Yo Yo Ma,$2"
null    "Canada,Jon York,$44"
null    "United Kingdom,Alex Ball,$5"
null    "inner-joined 8 rows."
Removing temp directory /tmp/innerHashJoin.root.20170214.034407.159333...
```

3. HW5.2.1 (OPTIONAL) Almost stateless reducer-side join

Back to Table of Contents

The following MRJob code, implements a reduce-side join for an inner join. The reducer is almost stateless, i.e., uses as little memory as possible. Use the tables from HW5.2 for this HW and join based on the country code (third column of the transactions table and the second column of the Countries table perform. Perform an left, right, inner joins using the code provided below and 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

Again make smart decisions about which table should be the left table (i.e., crosscheck the code).

Some notes on the code Here, the mapper receives its set of input splits either from the transaction table or from the countries table and makes the appropriate transformations: splitting the line into fields, and emitting a key/value. The key is the join key - in this case, the country code field of both sets of records. The mapper knows which file and type of record it is receiving based on the length of the fields. The records it emits contain the join field as the key, which acts as the partitioning key; We use the SORT_VALUES option, which ensures the values are sorted as well. Then, we employ a trick to ensure that for each join key, country records are seen always before transaction records. We achieve this by adding an arbitrary key to the front of the value: 'A' for countries, 'B' for customers. This makes countries sort before customers for each and every join/partition key. After that trick, the join is simply a matter of storing countries ('A' records) and crossing this array with each customer record.

In [ ]:

```
import sys, os, re
from mrjob.job import MRJob

class MRJoin(MRJob):

  # Performs secondary sort
  SORT_VALUES = True

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

    if len(splits) == 2: # country data
      symbol = 'A' # make country sort before transaction data
      country2digit = splits[1]
      yield country2digit, [symbol, splits]
    else: # person data
      symbol = 'B'
      country2digit = splits[2]
      yield country2digit, [symbol, splits]

  def reducer(self, key, values):
    countries = [] # should come first, as they are sorted on artificia key
    for value in values:
      if value[0] == 'A':
        countries.append(value)
      if value[0] == 'B':
        for country in countries:
          yield key, country[1:] + value[1:]

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

5.3 Pairwise similarity - PHASE 1

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

(1) Using the systems tests data sets, write mrjob code to build the stripes

(2) Write mrjob code to build an inverted index from the stripes

(3) Using two (symmetric) comparison methods of your choice (e.g., correlations, distances, similarities), pairwise compare all stripes (vectors), and output to a file.

==Design notes for (1)==
For this task you will be able to modify the pattern we used in HW 3.2 (feel free to use the solution as reference). To total the word counts across the n-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 n-gram. Treat these words as a basket, as we have in HW 3, but count all stripes or pairs in both orders, i.e., count both orderings: (word1,word2), and (word2,word1), to preserve symmetry in our output for (2).

==Design notes for (3)==
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.

Type *Markdown* and LaTeX: α^2

In [291]:

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

from __future__ import division
import re
import mrjob
import json
from mrjob.protocol import RawProtocol
from mrjob.job import MRJob
from mrjob.step import MRStep

class MRbuildStripes(MRJob):

    def mapper(self, _, line):
        ngram,count,page_count,book_count = line.split('\t',3)
        ngram = ngram.strip()
        count = int(count)

        # lowercase and parse out each word
        words = ngram.lower().split()

        d = {}

        # Create a dictionary within a dictionary
        # For example: d[biography] = {"a": 92, "of": 92, "george": 92, "ge

        for term in sorted(words):
            if term not in d.keys():
                d[term] = {}

            for term2 in sorted(words):
                if term != term2:
                    if term2 in d[term]:
                        d[term][term2] += count
                    else:
                        d[term][term2] = count

        # iterate through the dictionary and yield the top level term, the 
        # Example: "biography, (general, 92)"

        for k,v in d.iteritems():
            for k2,v2 in d[k].iteritems():
                yield k, (k2, v2)


    def reducer(self, key, line):

        red_d = {}
        term1 = key

        # Combine the various term cooccurrence counts into a single dictio

        for term,count in line:
            count = int(count)
            term2 = term
            if term1 not in red_d.keys():
                red_d[term1] = {}
            if term2 in red_d[term1]:
                red_d[term1][term2] += count
            else:
                red_d[term1][term2] = count

        for k,v in red_d.iteritems():
            yield k,v

  #END SUDENT CODE531_STRIPES
if __name__ == '__main__':
    MRbuildStripes.run()
```

```
Overwriting buildStripes.py
```

```
In [292]: %%writefile invertedIndex.py
          #!~/anaconda2/bin/python
          # -*- coding: utf-8 -*-


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

          class MRinvertedIndex(MRJob):

          #START STUDENT CODE531_INV_INDEX

              def mapper(self,_,line):

                  line = line.strip()
                  key_term, words = line.split("\t")

                  # 'words' are coming in with the structure of a dictionary, but form
                  # ast.literal_eval converts it to the dictionary that it should be
                  words = ast.literal_eval(words)
                  _len = len(words)

                  # for each word, output the cooccurring terms and the number of asso
                  for word in words:
                      yield word, (key_term, _len)

              def reducer(self,key,value):

                  d = collections.defaultdict(list)
                  for v in value:
                      d[key].append(v)
                  yield key,d[key]

          #END STUDENT CODE531_INV_INDEX

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

```
Overwriting invertedIndex.py
```

In [311]:

```
%%writefile similarity.py
#!~/anaconda2/bin/python
# -*- coding: utf-8 -*-

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

class MRsimilarity(MRJob):

  #START SUDENT CODE531_SIMILARITY

    MRJob.SORT_VALUES = True
    def steps(self):

        JOBCONF_STEP1 = {}
        JOBCONF_STEP2 = {
          ######### IMPORTANT: THIS WILL HAVE NO EFFECT IN -r local MODE. MU
            'mapreduce.job.output.key.comparator.class': 'org.apache.hadoop.
            'mapreduce.partition.keycomparator.options':'-k1,1nr',

        }

        return [MRStep(jobconf=JOBCONF_STEP1,
                    mapper=self.mapper_pair_sim,
                    reducer=self.reducer_pair_sim)
                ,
                MRStep(jobconf=JOBCONF_STEP2,
                    mapper=None,
                    reducer=self.reducer_sort)
                ]

    def mapper_pair_sim(self,_,line):
        line = line.strip()
        term,coterm = line.split("\t")
        coterm = json.loads(coterm)

        X = map(lambda x: x[0]+"."+str(x[1]) , coterm)

        # taking advantage of symmetry, output only (a,b), but not (b,a)
        # 'set' will output only the unique occurrences
        for subset in itertools.combinations(sorted(set(X)), 2):
            yield subset[0]+"."+subset[1], 1


    def reducer_pair_sim(self,key,value):
        Doc1, Doc1_len, Doc2, Doc2_len = key.split(".")
        doc1_len = int(Doc1_len)
        doc2_len = int(Doc2_len)
        t = sum(value)


        # calculate the similarity values
        jaccard = t / ( doc1_len + doc2_len - t )
        cosine = t * ((1/math.sqrt(doc1_len)) * (1/math.sqrt(doc2_len)))
        dice = (2*t) / (doc1_len + doc2_len)
        overlap = t / min(doc1_len, doc2_len)

        # Average the 4 similarities
        avg = sum([jaccard,cosine,dice,overlap]) / 4

        yield [avg,jaccard,cosine,overlap,dice], (Doc1+" - "+Doc2)


    def reducer_sort(self,key,value):
        for v in value:
            yield key,v

#END SUDENT CODE531_SIMILARITY

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

```
Overwriting similarity.py
```

HW5.3.1 Run Systems tests locally on small datasets (PHASE1)

Back to Table of Contents

Complete 5.3 and systems test using the below test datasets. Phase 2 will focus on the entire Ngram dataset.

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

Test datasets:

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

A large subset of the Google n-grams dataset

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

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

https://www.dropbox.com/sh/tmqpc4o0xswhkvz/AACUifrl6wrMrlK6a3X3lZ9Ea?dl=0 (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. In real life, we would calculate the stripes cooccurrence data from the raw text by windowing over the raw text and not from the 5-gram preprocessed data (as we are doing here). 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 [294]:

```
%%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 Child's Christmas in Wales    1099    1061    866
A Circumstantial Narrative of the   62  62  50
A City by the Sea   62  60  49
A Collection of Fairy Tales 123 117 80
A Collection of Forms of    116 103 82
```

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

2: unit/systems atlas-boon

In [295]:

```
%%writefile atlas-boon-systems-test.txt
atlas boon  50  50  50
boon cava dipped    10  10  10
atlas dipped    15  15  15
```

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

3: 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}

(1) build stripes for all the test data sets - run the commands and insure that your output matches the output below

In [296]:

```
###########################################################################
# Make Stripes from ngrams for systems test 1
###########################################################################

!hdfs dfs rm --recursive systems_test_stripes_1
!python buildStripes.py -r local googlebooks-eng-all-5gram-20090715-0-filter
```

```
rm: Unknown command
Did you mean -rm?  This command begins with a dash.
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.root.20170213.072821.757803
Running step 1 of 1...
Streaming final output from /tmp/buildStripes.root.20170213.072821.75780
3/output...
Removing temp directory /tmp/buildStripes.root.20170213.072821.757803...
```

In [297]:

```
!cat systems_test_stripes_1
```

```
"a"     {"limited":55,"sea":62,"general":92,"female":447,"in":1201,"relig
ious":59,"george":92,"biography":92,"city":62,"for":59,"tales":123,"gover
nment":102,"the":124,"forms":116,"wales":1099,"christmas":1099,"child's":
1099,"collection":239,"by":62,"case":604,"circumstantial":62,"of":1011,"s
tudy":604,"bill":59,"establishing":59,"narrative":62,"fairy":123}
"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,"fema
le":447,"in":102}
"child's"       {"a":1099,"wales":1099,"christmas":1099,"in":1099}
"christmas"     {"a":1099,"wales":1099,"in":1099,"child's":1099}
"circumstantial"        {"a":62,"of":62,"the":62,"narrative":62}
"city"  {"a":62,"the":62,"by":62,"sea":62}
"collection"    {"a":239,"forms":116,"fairy":123,"tales":123,"of":355}
"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":232,"collection":116}
"general"       {"a":92,"of":92,"george":92,"biography":92}
"george"        {"a":92,"of":92,"biography":92,"general":92}
"government"    {"a":102,"case":102,"study":102,"in":102}
"in"    {"a":1201,"case":102,"government":102,"study":102,"child's":109
9,"wales":1099,"christmas":1099}
"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,"the":62,"s
tudy":502,"collection":355,"general":92,"forms":232,"tales":123,"female":
447,"narrative":62,"fairy":123,"george":92,"biography":92}
"religious"     {"a":59,"bill":59,"for":59,"establishing":59}
"sea"   {"a":62,"city":62,"the":62,"by":62}
"study" {"a":604,"case":604,"limited":55,"of":502,"government":102,"femal
e":447,"in":102}
"tales" {"a":123,"of":123,"fairy":123,"collection":123}
"the"   {"a":124,"city":62,"circumstantial":62,"of":62,"sea":62,"narrativ
e":62,"by":62}
"wales" {"a":1099,"in":1099,"christmas":1099,"child's":1099}
```

```
"a"    {"limited": 55, "sea": 62, "general": 92, "female": 447, "i
n": 1201, "religious": 59, "george": 92, "biography": 92, "city": 6
2, "for": 59, "tales": 123, "child's": 1099, "forms": 116, "wales":
 1099, "christmas": 1099, "government": 102, "collection": 239, "b
y": 62, "case": 604, "circumstantial": 62, "fairy": 123, "of": 1011,
"study": 604, "bill": 59, "establishing": 59, "narrative": 62, "th
e": 124}
"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, "s
tudy": 604, "female": 447, "in": 102}
"child's"    {"a": 1099, "wales": 1099, "christmas": 1099, "in": 109
9}
"christmas"    {"a": 1099, "wales": 1099, "in": 1099, "child's": 109
9}
"circumstantial"    {"a": 62, "of": 62, "the": 62, "narrative": 62}
"city"    {"a": 62, "the": 62, "by": 62, "sea": 62}
"collection"    {"a": 239, "of": 355, "fairy": 123, "tales": 123, "f
orms": 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": 232, "collection": 116}
"general"    {"a": 92, "of": 92, "george": 92, "biography": 92}
"george"    {"a": 92, "of": 92, "biography": 92, "general": 92}
"government"    {"a": 102, "case": 102, "study": 102, "in": 102}
"in"    {"a": 1201, "case": 102, "government": 102, "study": 102, "c
hild's": 1099, "wales": 1099, "christmas": 1099}
"limited"    {"a": 55, "case": 55, "study": 55, "of": 55}
"narrative"    {"a": 62, "of": 62, "the": 62, "circumstantial": 62}
"of"    {"a": 1127, "case": 502, "circumstantial": 62, "george": 92,
"limited": 55, "tales": 123, "collection": 471, "general": 92, "form
s": 348, "female": 447, "narrative": 62, "study": 502, "fairy": 123,
"the": 62, "biography": 92}
"religious"    {"a": 59, "bill": 59, "for": 59, "establishing": 59}
"sea"    {"a": 62, "city": 62, "the": 62, "by": 62}
"study"    {"a": 604, "case": 604, "limited": 55, "government": 102,
"of": 502, "female": 447, "in": 102}
"tales"    {"a": 123, "of": 123, "fairy": 123, "collection": 123}
"the"    {"a": 124, "city": 62, "circumstantial": 62, "of": 62, "se
a": 62, "narrative": 62, "by": 62}
"wales"    {"a": 1099, "in": 1099, "christmas": 1099, "child's": 109
9}
```

In [298]:

```
###########################################################################
# Make Stripes from ngrams for systems test 2
###########################################################################

!hdfs dfs rm --recursive systems_test_stripes_2
!python buildStripes.py -r local atlas-boon-systems-test.txt > systems_test_
```

```
rm: Unknown command
Did you mean -rm?  This command begins with a dash.
No configs found; falling back on auto-configuration
Creating temp directory /tmp/buildStripes.root.20170213.072836.216996
Running step 1 of 1...
Streaming final output from /tmp/buildStripes.root.20170213.072836.21699
6/output...
Removing temp directory /tmp/buildStripes.root.20170213.072836.216996...
```

In [299]:

```
!cat systems_test_stripes_2
```

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

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

In [300]:

```
########################################################################
# Stripes for systems test 3 (given, no need to build stripes)
########################################################################

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

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

(2) Build Inverted Index - run the commands and insure that your output matches the output below

In [301]:

```
!python invertedIndex.py -r local systems_test_stripes_1 > systems_test_inde
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.root.20170213.072857.238594
Running step 1 of 1...
Streaming final output from /tmp/invertedIndex.root.20170213.072857.23859
4/output...
Removing temp directory /tmp/invertedIndex.root.20170213.072857.238594...
```

In [302]:

```
!python invertedIndex.py -r local systems_test_stripes_2 > systems_test_inde
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.root.20170213.072900.428320
Running step 1 of 1...
Streaming final output from /tmp/invertedIndex.root.20170213.072900.42832
0/output...
Removing temp directory /tmp/invertedIndex.root.20170213.072900.428320...
```

In [303]:

```
!python invertedIndex.py -r local systems_test_stripes_3 > systems_test_inde
```

```
No configs found; falling back on auto-configuration
Creating temp directory /tmp/invertedIndex.root.20170213.072903.458618
Running step 1 of 1...
Streaming final output from /tmp/invertedIndex.root.20170213.072903.45861
8/output...
Removing temp directory /tmp/invertedIndex.root.20170213.072903.458618...
```

In [306]:

```
#############################################################
# Pretty print systems tests for generating Inverted Index
#############################################################

import json

for i in range(1,4):
    print "—"*100
    print "Systems test ",i," - Inverted Index"
    print "—"*100
    with open("systems_test_index_"+str(i),"r") as f:
        lines = f.readlines()
        for line in lines:
            line = line.strip()
            word,stripe = line.split("\t")
            stripe = json.loads(stripe)
            stripe.extend([["",""] for _ in xrange(3 - len(stripe))])
            print "{0:>16} |{1:>16} |{2:>16} |{3:>16}".format(
                (word), stripe[0][0]+" "+str(stripe[0][1]), stripe[1][0]+" "
```

```
————————————————————————————————————————————————————————————————————————————
Systems test  1  - Inverted Index
————————————————————————————————————————————————————————————————————————————
             "a" |        "bill" 4 |   "biography" 4 |          "by" 4
          "bill" |          "a" 27 |"establishing" 4 |         "for" 4
     "biography" |          "a" 27 |     "general" 4 |      "george" 4
            "by" |          "a" 27 |        "city" 4 |         "sea" 4
          "case" |          "a" 27 |      "female" 4 |  "government" 4
       "child's" |          "a" 27 |   "christmas" 4 |          "in" 7
     "christmas" |          "a" 27 |     "child's" 4 |          "in" 7
"circumstantial" |          "a" 27 |   "narrative" 4 |         "of" 15
          "city" |          "a" 27 |          "by" 4 |         "sea" 4
    "collection" |          "a" 27 |       "fairy" 4 |       "forms" 3
  "establishing" |          "a" 27 |        "bill" 4 |         "for" 4
         "fairy" |          "a" 27 |  "collection" 5 |         "of" 15
        "female" |          "a" 27 |        "case" 7 |         "of" 15
           "for" |          "a" 27 |        "bill" 4 |"establishing" 4
         "forms" |          "a" 27 |  "collection" 5 |         "of" 15
       "general" |          "a" 27 |   "biography" 4 |      "george" 4
        "george" |          "a" 27 |   "biography" 4 |     "general" 4
    "government" |          "a" 27 |        "case" 7 |          "in" 7
            "in" |          "a" 27 |        "case" 7 |     "child's" 4
       "limited" |          "a" 27 |        "case" 7 |         "of" 15
     "narrative" |          "a" 27 |"circumstantial" 4 |         "of" 15
            "of" |          "a" 27 |   "biography" 4 |        "case" 7
     "religious" |          "a" 27 |        "bill" 4 |"establishing" 4
           "sea" |          "a" 27 |          "by" 4 |        "city" 4
         "study" |          "a" 27 |        "case" 7 |      "female" 4
         "tales" |          "a" 27 |  "collection" 5 |       "fairy" 4
           "the" |          "a" 27 |          "by" 4 |"circumstantial" 4
         "wales" |          "a" 27 |     "child's" 4 |   "christmas" 4
————————————————————————————————————————————————————————————————————————————
```

```
Systems test  2  - Inverted Index
──────────────────────────────────────────────────────────────────────────

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

Systems test  3  - Inverted Index
──────────────────────────────────────────────────────────────────────────

             "M" |         "DocC" 4 |                  |
             "N" |         "DocC" 4 |                  |
             "X" |         "DocA" 3 |         "DocB" 2 |
             "Y" |         "DocA" 3 |         "DocB" 2 |         "DocC" 4
             "Z" |         "DocA" 3 |         "DocC" 4 |
```

Inverted Index

In [ ]:

```
————————————————————————————————————————————————————————————————————
Systems test  1  - Inverted Index
————————————————————————————————————————————————————————————————————
             "a" |          bill 4 |     biography 4 |            by 4
          "bill" |            a 27 |  establishing 4 |           for 4
     "biography" |            a 27 |       general 4 |        george 4
            "by" |            a 27 |          city 4 |           sea 4
          "case" |            a 27 |        female 4 |    government 4
       "child's" |            a 27 |     christmas 4 |            in 7
     "christmas" |            a 27 |       child's 4 |            in 7
"circumstantial" |            a 27 |     narrative 4 |           of 15
          "city" |            a 27 |            by 4 |           sea 4
    "collection" |            a 27 |         fairy 4 |         forms 3
  "establishing" |            a 27 |          bill 4 |           for 4
         "fairy" |            a 27 |    collection 5 |           of 15
        "female" |            a 27 |          case 7 |           of 15
           "for" |            a 27 |          bill 4 |  establishing 4
         "forms" |            a 27 |    collection 5 |           of 15
       "general" |            a 27 |     biography 4 |        george 4
        "george" |            a 27 |     biography 4 |       general 4
    "government" |            a 27 |          case 7 |            in 7
            "in" |            a 27 |          case 7 |       child's 4
       "limited" |            a 27 |          case 7 |           of 15
     "narrative" |            a 27 |circumstantial 4 |           of 15
            "of" |            a 27 |     biography 4 |          case 7
     "religious" |            a 27 |          bill 4 |  establishing 4
           "sea" |            a 27 |            by 4 |          city 4
         "study" |            a 27 |          case 7 |        female 4
         "tales" |            a 27 |    collection 5 |         fairy 4
           "the" |            a 27 |            by 4 |circumstantial 4
         "wales" |            a 27 |       child's 4 |     christmas 4
————————————————————————————————————————————————————————————————————
Systems test  2  - Inverted Index
————————————————————————————————————————————————————————————————————
         "atlas" |          boon 3 |        dipped 3 |
          "boon" |         atlas 2 |          cava 2 |        dipped 3
          "cava" |          boon 3 |        dipped 3 |
        "dipped" |         atlas 2 |          boon 3 |          cava 2
————————————————————————————————————————————————————————————————————
Systems test  3  - Inverted Index
————————————————————————————————————————————————————————————————————
             "M" |          DocC 4 |                 |
             "N" |          DocC 4 |                 |
             "X" |          DocA 3 |          DocB 2 |
             "Y" |          DocA 3 |          DocB 2 |          DocC 4
             "Z" |          DocA 3 |          DocC 4 |
```

(3) Calculate similarities - run the commands and insure that your output matches the output below

NOTE: you must run in hadoop mode to generate sorted similarities

In [307]:

```
!python similarity.py -r hadoop systems_test_index_1 > systems_test_similari
```

```
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.6.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.ja
r
Creating temp directory /tmp/similarity.root.20170213.073031.946921
Copying local files to hdfs:///user/root/tmp/mrjob/similarity.root.201702
13.073031.946921/files/...
Detected hadoop configuration property names that do not match hadoop ver
sion 2.6.0:
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.
class
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.opt
ions
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.o
```

In [308]:

```
!python similarity.py -r hadoop systems_test_index_2 > systems_test_similari
```

```
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.6.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.ja
r
Creating temp directory /tmp/similarity.root.20170213.073151.795311
Copying local files to hdfs:///user/root/tmp/mrjob/similarity.root.201702
13.073151.795311/files/...
Detected hadoop configuration property names that do not match hadoop ver
sion 2.6.0:
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.
class
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.opt
ions
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.o
```

In [309]:

```
!python similarity.py -r hadoop systems_test_index_3 > systems_test_similari
```

```
No configs found; falling back on auto-configuration
Looking for hadoop binary in $PATH...
Found hadoop binary: /usr/bin/hadoop
Using Hadoop version 2.6.0
Looking for Hadoop streaming jar in /home/hadoop/contrib...
Looking for Hadoop streaming jar in /usr/lib/hadoop-mapreduce...
Found Hadoop streaming jar: /usr/lib/hadoop-mapreduce/hadoop-streaming.ja
r
Creating temp directory /tmp/similarity.root.20170213.073316.826528
Copying local files to hdfs:///user/root/tmp/mrjob/similarity.root.201702
13.073316.826528/files/...
Detected hadoop configuration property names that do not match hadoop ver
sion 2.6.0:
The have been translated as follows
 mapred.output.key.comparator.class: mapreduce.job.output.key.comparator.
class
mapred.text.key.comparator.options: mapreduce.partition.keycomparator.opt
ions
mapred.text.key.partitioner.options: mapreduce.partition.keypartitioner.o
ptions
```

```
In [313]: ##############################################
          # Pretty print systems tests
          ##############################################

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

            with open("systems_test_similarities_"+str(i),"r") as f:
                  lines = f.readlines()
                  for line in lines:
                      line = line.strip()
                      sims,stripe = line.split("\t")
                      sims = ast.literal_eval(sims)
                      stripe = json.loads(stripe)
                      print "{0:>15f} |{1:>15} |{2:>15f} |{3:>15f} |{4:>15f} |{5:>15f
                          sims[0], stripe, sims[2], sims[1], sims[3], sims[4])
```

```
——————————————————————————————————————————————————————————————————————————

Systems test  1  - Similarity measures
——————————————————————————————————————————————————————————————————————————

        average |            pair |         cosine |         jaccard |
   overlap |            dice
---------------------------------------------------------------------------
-------------------------------------
       0.180200 |"the" - "wales" |       0.188982 |        0.100000 |
  0.250000 |       0.181818
       0.223214 |"tales" - "wales" |       0.250000 |        0.142857 |
    0.250000 |       0.250000
       0.365956 |"tales" - "the" |       0.377964 |        0.222222 |
  0.500000 |       0.363636
       0.365956 |"study" - "wales" |       0.377964 |        0.222222 |
    0.500000 |       0.363636
       0.255952 |"study" - "the" |       0.285714 |        0.166667 |
  0.285714 |       0.285714
       0.365956 |"study" - "tales" |       0.377964 |        0.222222 |
    0.500000 |       0.363636
       0.223214 |"sea" - "wales" |       0.250000 |        0.142857 |
  0.250000 |       0.250000
       0.559350 |  "sea" - "the" |       0.566947 |        0.375000 |
  0.750000 |       0.545455
       0.223214 |"sea" - "tales" |       0.250000 |        0.142857 |
  0.250000 |       0.250000
       0.180200 |"sea" - "study" |       0.188982 |        0.100000 |
  0.250000 |       0.181818
       0.223214 |"religious" - "wales" |       0.250000 |        0.142857
 |       0.250000 |       0.250000
       0.180200 |"religious" - "the" |       0.188982 |        0.100000 |
      0.250000 |       0.181818
       0.223214 |"religious" - "tales" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.180200 |"religious" - "study" |       0.188982 |       0.100000
 |       0.250000 |       0.181818
       0.223214 |"religious" - "sea" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.134980 | "of" - "wales" |       0.129099 |       0.055556 |
  0.250000 |       0.105263
       0.287991 |   "of" - "the" |       0.292770 |       0.157895 |
  0.428571 |       0.272727
       0.410147 | "of" - "tales" |       0.387298 |       0.187500 |
  0.750000 |       0.315789
       0.386912 | "of" - "study" |       0.390360 |       0.222222 |
  0.571429 |       0.363636
       0.271593 |   "of" - "sea" |       0.258199 |       0.117647 |
  0.500000 |       0.210526
       0.134980 |"of" - "religious" |       0.129099 |       0.055556 |
     0.250000 |       0.105263
       0.223214 |"narrative" - "wales" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.559350 |"narrative" - "the" |       0.566947 |       0.375000 |
      0.750000 |       0.545455
       0.458333 |"narrative" - "tales" |       0.500000 |       0.333333
 |       0.500000 |       0.500000
       0.365956 |"narrative" - "study" |       0.377964 |       0.222222
 |       0.500000 |       0.363636
       0.458333 |"narrative" - "sea" |       0.500000 |       0.333333 |
      0.500000 |       0.500000
       0.223214 |"narrative" - "religious" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.410147 |"narrative" - "of" |       0.387298 |       0.187500 |
     0.750000 |       0.315789
       0.223214 |"limited" - "wales" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.365956 |"limited" - "the" |       0.377964 |       0.222222 |
    0.500000 |       0.363636
       0.458333 |"limited" - "tales" |       0.500000 |       0.333333 |
      0.500000 |       0.500000
       0.559350 |"limited" - "study" |       0.566947 |       0.375000 |
      0.750000 |       0.545455
       0.223214 |"limited" - "sea" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"limited" - "religious" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.410147 |"limited" - "of" |       0.387298 |       0.187500 |
   0.750000 |       0.315789
       0.458333 |"limited" - "narrative" |       0.500000 |       0.33333
3 |       0.500000 |       0.500000
       0.559350 | "in" - "wales" |       0.566947 |       0.375000 |
  0.750000 |       0.545455
       0.126374 |   "in" - "the" |       0.142857 |       0.076923 |
  0.142857 |       0.142857
       0.180200 | "in" - "tales" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.389610 | "in" - "study" |       0.428571 |       0.272727 |
  0.428571 |       0.428571
```

```
       0.180200 |   "in" - "sea" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.180200 |"in" - "religious" |       0.188982 |       0.100000 |
     0.250000 |       0.181818
       0.287991 |    "in" - "of" |       0.292770 |       0.157895 |
  0.428571 |       0.272727
       0.180200 |"in" - "narrative" |       0.188982 |       0.100000 |
     0.250000 |       0.181818
       0.559350 |"in" - "limited" |       0.566947 |       0.375000 |
   0.750000 |       0.545455
       0.458333 |"government" - "wales" |       0.500000 |       0.333333
|       0.500000 |       0.500000
       0.180200 |"government" - "the" |       0.188982 |       0.100000 |
       0.250000 |       0.181818
       0.223214 |"government" - "tales" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.559350 |"government" - "study" |       0.566947 |       0.375000
|       0.750000 |       0.545455
       0.223214 |"government" - "sea" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"government" - "religious" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.410147 |"government" - "of" |       0.387298 |       0.187500 |
      0.750000 |       0.315789
       0.223214 |"government" - "narrative" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.712500 |"government" - "limited" |       0.750000 |       0.6000
00 |       0.750000 |       0.750000
       0.559350 |"government" - "in" |       0.566947 |       0.375000 |
      0.750000 |       0.545455
       0.223214 |"george" - "wales" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.365956 |"george" - "the" |       0.377964 |       0.222222 |
   0.500000 |       0.363636
       0.458333 |"george" - "tales" |       0.500000 |       0.333333 |
     0.500000 |       0.500000
       0.365956 |"george" - "study" |       0.377964 |       0.222222 |
     0.500000 |       0.363636
       0.223214 |"george" - "sea" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.223214 |"george" - "religious" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.410147 |"george" - "of" |       0.387298 |       0.187500 |
  0.750000 |       0.315789
       0.458333 |"george" - "narrative" |       0.500000 |       0.333333
|       0.500000 |       0.500000
       0.458333 |"george" - "limited" |       0.500000 |       0.333333 |
       0.500000 |       0.500000
       0.180200 |"george" - "in" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"george" - "government" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.223214 |"general" - "wales" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.365956 |"general" - "the" |       0.377964 |       0.222222 |
    0.500000 |       0.363636
       0.458333 |"general" - "tales" |       0.500000 |       0.333333 |
      0.500000 |       0.500000
       0.365956 |"general" - "study" |       0.377964 |       0.222222 |
      0.500000 |       0.363636
       0.223214 |"general" - "sea" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"general" - "religious" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.410147 |"general" - "of" |       0.387298 |       0.187500 |
   0.750000 |       0.315789
       0.458333 |"general" - "narrative" |       0.500000 |       0.33333
3 |       0.500000 |       0.500000
       0.458333 |"general" - "limited" |       0.500000 |       0.333333
 |       0.500000 |       0.500000
       0.180200 |"general" - "in" |       0.188982 |       0.100000 |
   0.250000 |       0.181818
       0.223214 |"general" - "government" |       0.250000 |       0.1428
57 |       0.250000 |       0.250000
       0.712500 |"general" - "george" |       0.750000 |       0.600000 |
      0.750000 |       0.750000
       0.268597 |"forms" - "wales" |       0.288675 |       0.166667 |
    0.333333 |       0.285714
       0.438276 |"forms" - "the" |       0.436436 |       0.250000 |
  0.666667 |       0.400000
       0.868292 |"forms" - "tales" |       0.866025 |       0.750000 |
    1.000000 |       0.857143
       0.438276 |"forms" - "study" |       0.436436 |       0.250000 |
    0.666667 |       0.400000
       0.268597 |"forms" - "sea" |       0.288675 |       0.166667 |
  0.333333 |       0.285714
       0.268597 |"forms" - "religious" |       0.288675 |       0.166667
 |       0.333333 |       0.285714
       0.328008 | "forms" - "of" |       0.298142 |       0.125000 |
  0.666667 |       0.222222
       0.553861 |"forms" - "narrative" |       0.577350 |       0.400000
 |       0.666667 |       0.571429
       0.553861 |"forms" - "limited" |       0.577350 |       0.400000 |
      0.666667 |       0.571429
       0.215666 | "forms" - "in" |       0.218218 |       0.111111 |
  0.333333 |       0.200000
       0.268597 |"forms" - "government" |       0.288675 |       0.166667
|       0.333333 |       0.285714
       0.553861 |"forms" - "george" |       0.577350 |       0.400000 |
     0.666667 |       0.571429
       0.553861 |"forms" - "general" |       0.577350 |       0.400000 |
      0.666667 |       0.571429
       0.223214 |"for" - "wales" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.180200 |  "for" - "the" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"for" - "tales" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.180200 |"for" - "study" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |  "for" - "sea" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.712500 |"for" - "religious" |       0.750000 |       0.600000 |
      0.750000 |       0.750000
```

```
          0.134980 |   "for" - "of" |       0.129099 |       0.055556 |
     0.250000 |       0.105263
          0.223214 |"for" - "narrative" |       0.250000 |       0.142857 |
         0.250000 |       0.250000
          0.223214 |"for" - "limited" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
          0.180200 |   "for" - "in" |       0.188982 |       0.100000 |
     0.250000 |       0.181818
          0.223214 |"for" - "government" |       0.250000 |       0.142857 |
          0.250000 |       0.250000
          0.223214 |"for" - "george" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
          0.223214 |"for" - "general" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
          0.268597 |"for" - "forms" |       0.288675 |       0.166667 |
     0.333333 |       0.285714
          0.223214 |"female" - "wales" |       0.250000 |       0.142857 |
        0.250000 |       0.250000
          0.365956 |"female" - "the" |       0.377964 |       0.222222 |
      0.500000 |       0.363636
          0.458333 |"female" - "tales" |       0.500000 |       0.333333 |
        0.500000 |       0.500000
          0.559350 |"female" - "study" |       0.566947 |       0.375000 |
        0.750000 |       0.545455
          0.223214 |"female" - "sea" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
          0.223214 |"female" - "religious" |       0.250000 |       0.142857
|       0.250000 |       0.250000
          0.410147 |"female" - "of" |       0.387298 |       0.187500 |
     0.750000 |       0.315789
          0.458333 |"female" - "narrative" |       0.500000 |       0.333333
|       0.500000 |       0.500000
          1.000000 |"female" - "limited" |       1.000000 |       1.000000 |
          1.000000 |       1.000000
          0.559350 |"female" - "in" |       0.566947 |       0.375000 |
     0.750000 |       0.545455
          0.712500 |"female" - "government" |       0.750000 |       0.60000
0 |       0.750000 |       0.750000
          0.458333 |"female" - "george" |       0.500000 |       0.333333 |
         0.500000 |       0.500000
          0.458333 |"female" - "general" |       0.500000 |       0.333333 |
          0.500000 |       0.500000
          0.553861 |"female" - "forms" |       0.577350 |       0.400000 |
        0.666667 |       0.571429
          0.223214 |"female" - "for" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
          0.223214 |"fairy" - "wales" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
          0.365956 |"fairy" - "the" |       0.377964 |       0.222222 |
     0.500000 |       0.363636
          0.712500 |"fairy" - "tales" |       0.750000 |       0.600000 |
       0.750000 |       0.750000
          0.365956 |"fairy" - "study" |       0.377964 |       0.222222 |
       0.500000 |       0.363636
          0.223214 |"fairy" - "sea" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
          0.223214 |"fairy" - "religious" |       0.250000 |       0.142857
```

```
 |       0.250000 |       0.250000
       0.410147 | "fairy" - "of" |       0.387298 |       0.187500 |
  0.750000 |       0.315789
       0.458333 |"fairy" - "narrative" |       0.500000 |       0.333333
 |       0.500000 |       0.500000
       0.458333 |"fairy" - "limited" |       0.500000 |       0.333333 |
      0.500000 |       0.500000
       0.180200 | "fairy" - "in" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"fairy" - "government" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.458333 |"fairy" - "george" |       0.500000 |       0.333333 |
     0.500000 |       0.500000
       0.458333 |"fairy" - "general" |       0.500000 |       0.333333 |
      0.500000 |       0.500000
       0.868292 |"fairy" - "forms" |       0.866025 |       0.750000 |
    1.000000 |       0.857143
       0.223214 |"fairy" - "for" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.458333 |"fairy" - "female" |       0.500000 |       0.333333 |
     0.500000 |       0.500000
       0.223214 |"establishing" - "wales" |       0.250000 |       0.1428
57 |       0.250000 |       0.250000
       0.180200 |"establishing" - "the" |       0.188982 |       0.100000
|       0.250000 |       0.181818
       0.223214 |"establishing" - "tales" |       0.250000 |       0.1428
57 |       0.250000 |       0.250000
       0.180200 |"establishing" - "study" |       0.188982 |       0.1000
00 |       0.250000 |       0.181818
       0.223214 |"establishing" - "sea" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.712500 |"establishing" - "religious" |       0.750000 |       0.
600000 |       0.750000 |       0.750000
       0.134980 |"establishing" - "of" |       0.129099 |       0.055556
 |       0.250000 |       0.105263
       0.223214 |"establishing" - "narrative" |       0.250000 |       0.
142857 |       0.250000 |       0.250000
       0.223214 |"establishing" - "limited" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.180200 |"establishing" - "in" |       0.188982 |       0.100000
 |       0.250000 |       0.181818
       0.223214 |"establishing" - "government" |       0.250000 |
 0.142857 |       0.250000 |       0.250000
       0.223214 |"establishing" - "george" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.223214 |"establishing" - "general" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.268597 |"establishing" - "forms" |       0.288675 |       0.1666
67 |       0.333333 |       0.285714
       0.712500 |"establishing" - "for" |       0.750000 |       0.600000
|       0.750000 |       0.750000
       0.223214 |"establishing" - "female" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.223214 |"establishing" - "fairy" |       0.250000 |       0.1428
57 |       0.250000 |       0.250000
       0.205207 |"collection" - "wales" |       0.223607 |       0.125000
|       0.250000 |       0.222222
```

```
       0.317849 |"collection" - "the" |       0.338062 |       0.200000 |
       0.400000 |       0.333333
       0.646872 |"collection" - "tales" |       0.670820 |       0.500000
|       0.750000 |       0.666667
       0.317849 |"collection" - "study" |       0.338062 |       0.200000
|       0.400000 |       0.333333
       0.205207 |"collection" - "sea" |       0.223607 |       0.125000 |
       0.250000 |       0.222222
       0.205207 |"collection" - "religious" |       0.223607 |       0.12
5000 |       0.250000 |       0.222222
       0.477970 |"collection" - "of" |       0.461880 |       0.250000 |
      0.800000 |       0.400000
       0.419343 |"collection" - "narrative" |       0.447214 |       0.28
5714 |       0.500000 |       0.444444
       0.419343 |"collection" - "limited" |       0.447214 |       0.2857
14 |       0.500000 |       0.444444
       0.156652 |"collection" - "in" |       0.169031 |       0.090909 |
      0.200000 |       0.166667
       0.205207 |"collection" - "government" |       0.223607 |       0.1
25000 |       0.250000 |       0.222222
       0.419343 |"collection" - "george" |       0.447214 |       0.28571
4 |       0.500000 |       0.444444
       0.419343 |"collection" - "general" |       0.447214 |       0.2857
14 |       0.500000 |       0.444444
       0.504099 |"collection" - "forms" |       0.516398 |       0.333333
|       0.666667 |       0.500000
       0.205207 |"collection" - "for" |       0.223607 |       0.125000 |
       0.250000 |       0.222222
       0.419343 |"collection" - "female" |       0.447214 |       0.28571
4 |       0.500000 |       0.444444
       0.646872 |"collection" - "fairy" |       0.670820 |       0.500000
|       0.750000 |       0.666667
       0.205207 |"collection" - "establishing" |       0.223607 |
 0.125000 |       0.250000 |       0.222222
       0.223214 |"city" - "wales" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.559350 | "city" - "the" |       0.566947 |       0.375000 |
  0.750000 |       0.545455
       0.223214 |"city" - "tales" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.180200 |"city" - "study" |       0.188982 |       0.100000 |
   0.250000 |       0.181818
       0.712500 | "city" - "sea" |       0.750000 |       0.600000 |
  0.750000 |       0.750000
       0.223214 |"city" - "religious" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.271593 |  "city" - "of" |       0.258199 |       0.117647 |
  0.500000 |       0.210526
       0.458333 |"city" - "narrative" |       0.500000 |       0.333333 |
       0.500000 |       0.500000
       0.223214 |"city" - "limited" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.180200 |  "city" - "in" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"city" - "government" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.223214 |"city" - "george" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"city" - "general" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.268597 |"city" - "forms" |       0.288675 |       0.166667 |
   0.333333 |       0.285714
       0.223214 | "city" - "for" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"city" - "female" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"city" - "fairy" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.223214 |"city" - "establishing" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.205207 |"city" - "collection" |       0.223607 |       0.125000
 |       0.250000 |       0.222222
       0.223214 |"circumstantial" - "wales" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.559350 |"circumstantial" - "the" |       0.566947 |       0.3750
00 |       0.750000 |       0.545455
       0.458333 |"circumstantial" - "tales" |       0.500000 |       0.33
3333 |       0.500000 |       0.500000
       0.365956 |"circumstantial" - "study" |       0.377964 |       0.22
2222 |       0.500000 |       0.363636
       0.458333 |"circumstantial" - "sea" |       0.500000 |       0.3333
33 |       0.500000 |       0.500000
       0.223214 |"circumstantial" - "religious" |       0.250000 |
 0.142857 |       0.250000 |       0.250000
       0.410147 |"circumstantial" - "of" |       0.387298 |       0.18750
0 |       0.750000 |       0.315789
       0.712500 |"circumstantial" - "narrative" |       0.750000 |
 0.600000 |       0.750000 |       0.750000
       0.458333 |"circumstantial" - "limited" |       0.500000 |       0.
333333 |       0.500000 |       0.500000
       0.180200 |"circumstantial" - "in" |       0.188982 |       0.10000
0 |       0.250000 |       0.181818
       0.223214 |"circumstantial" - "government" |       0.250000 |
0.142857 |       0.250000 |       0.250000
       0.458333 |"circumstantial" - "george" |       0.500000 |       0.3
33333 |       0.500000 |       0.500000
       0.458333 |"circumstantial" - "general" |       0.500000 |       0.
333333 |       0.500000 |       0.500000
       0.553861 |"circumstantial" - "forms" |       0.577350 |       0.40
0000 |       0.666667 |       0.571429
       0.223214 |"circumstantial" - "for" |       0.250000 |       0.1428
57 |       0.250000 |       0.250000
       0.458333 |"circumstantial" - "female" |       0.500000 |       0.3
33333 |       0.500000 |       0.500000
       0.458333 |"circumstantial" - "fairy" |       0.500000 |       0.33
3333 |       0.500000 |       0.500000
       0.223214 |"circumstantial" - "establishing" |       0.250000 |
   0.142857 |       0.250000 |       0.250000
       0.419343 |"circumstantial" - "collection" |       0.447214 |
0.285714 |       0.500000 |       0.444444
       0.458333 |"circumstantial" - "city" |       0.500000 |       0.333
333 |       0.500000 |       0.500000
       0.712500 |"christmas" - "wales" |       0.750000 |       0.600000
 |       0.750000 |       0.750000
```

```
       0.180200 |"christmas" - "the" |       0.188982 |       0.100000 |
      0.250000 |       0.181818
       0.223214 |"christmas" - "tales" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.365956 |"christmas" - "study" |       0.377964 |       0.222222
 |       0.500000 |       0.363636
       0.223214 |"christmas" - "sea" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.223214 |"christmas" - "religious" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.134980 |"christmas" - "of" |       0.129099 |       0.055556 |
     0.250000 |       0.105263
       0.223214 |"christmas" - "narrative" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.223214 |"christmas" - "limited" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.559350 |"christmas" - "in" |       0.566947 |       0.375000 |
     0.750000 |       0.545455
       0.458333 |"christmas" - "government" |       0.500000 |       0.33
3333 |       0.500000 |       0.500000
       0.223214 |"christmas" - "george" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.223214 |"christmas" - "general" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.268597 |"christmas" - "forms" |       0.288675 |       0.166667
 |       0.333333 |       0.285714
       0.223214 |"christmas" - "for" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.223214 |"christmas" - "female" |       0.250000 |       0.142857
|       0.250000 |       0.250000
       0.223214 |"christmas" - "fairy" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.223214 |"christmas" - "establishing" |       0.250000 |       0.
142857 |       0.250000 |       0.250000
       0.205207 |"christmas" - "collection" |       0.223607 |       0.12
5000 |       0.250000 |       0.222222
       0.223214 |"christmas" - "city" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"christmas" - "circumstantial" |       0.250000 |
 0.142857 |       0.250000 |       0.250000
       0.712500 |"child's" - "wales" |       0.750000 |       0.600000 |
      0.750000 |       0.750000
       0.180200 |"child's" - "the" |       0.188982 |       0.100000 |
    0.250000 |       0.181818
       0.223214 |"child's" - "tales" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.365956 |"child's" - "study" |       0.377964 |       0.222222 |
      0.500000 |       0.363636
       0.223214 |"child's" - "sea" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"child's" - "religious" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.134980 |"child's" - "of" |       0.129099 |       0.055556 |
   0.250000 |       0.105263
       0.223214 |"child's" - "narrative" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.223214 |"child's" - "limited" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.559350 |"child's" - "in" |       0.566947 |       0.375000 |
   0.750000 |       0.545455
       0.458333 |"child's" - "government" |       0.500000 |       0.3333
33 |       0.500000 |       0.500000
       0.223214 |"child's" - "george" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"child's" - "general" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.268597 |"child's" - "forms" |       0.288675 |       0.166667 |
      0.333333 |       0.285714
       0.223214 |"child's" - "for" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"child's" - "female" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"child's" - "fairy" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.223214 |"child's" - "establishing" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.205207 |"child's" - "collection" |       0.223607 |       0.1250
00 |       0.250000 |       0.222222
       0.223214 |"child's" - "city" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.223214 |"child's" - "circumstantial" |       0.250000 |       0.
142857 |       0.250000 |       0.250000
       0.712500 |"child's" - "christmas" |       0.750000 |       0.60000
0 |       0.750000 |       0.750000
       0.365956 |"case" - "wales" |       0.377964 |       0.222222 |
   0.500000 |       0.363636
       0.255952 | "case" - "the" |       0.285714 |       0.166667 |
  0.285714 |       0.285714
       0.365956 |"case" - "tales" |       0.377964 |       0.222222 |
   0.500000 |       0.363636
       0.830357 |"case" - "study" |       0.857143 |       0.750000 |
   0.857143 |       0.857143
       0.180200 | "case" - "sea" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.180200 |"case" - "religious" |       0.188982 |       0.100000 |
       0.250000 |       0.181818
       0.386912 |  "case" - "of" |       0.390360 |       0.222222 |
  0.571429 |       0.363636
       0.365956 |"case" - "narrative" |       0.377964 |       0.222222 |
       0.500000 |       0.363636
       0.559350 |"case" - "limited" |       0.566947 |       0.375000 |
     0.750000 |       0.545455
       0.389610 |  "case" - "in" |       0.428571 |       0.272727 |
  0.428571 |       0.428571
       0.559350 |"case" - "government" |       0.566947 |       0.375000
 |       0.750000 |       0.545455
       0.365956 |"case" - "george" |       0.377964 |       0.222222 |
    0.500000 |       0.363636
       0.365956 |"case" - "general" |       0.377964 |       0.222222 |
     0.500000 |       0.363636
       0.438276 |"case" - "forms" |       0.436436 |       0.250000 |
   0.666667 |       0.400000
       0.180200 | "case" - "for" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
```

```
       0.559350 |"case" - "female" |       0.566947 |       0.375000 |
    0.750000 |       0.545455
       0.365956 |"case" - "fairy" |       0.377964 |       0.222222 |
   0.500000 |       0.363636
       0.180200 |"case" - "establishing" |       0.188982 |       0.10000
0 |       0.250000 |       0.181818
       0.317849 |"case" - "collection" |       0.338062 |       0.200000
 |       0.400000 |       0.333333
       0.180200 |"case" - "city" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.365956 |"case" - "circumstantial" |       0.377964 |       0.222
222 |       0.500000 |       0.363636
       0.365956 |"case" - "christmas" |       0.377964 |       0.222222 |
       0.500000 |       0.363636
       0.365956 |"case" - "child's" |       0.377964 |       0.222222 |
     0.500000 |       0.363636
       0.223214 | "by" - "wales" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.559350 |   "by" - "the" |       0.566947 |       0.375000 |
  0.750000 |       0.545455
       0.223214 | "by" - "tales" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.180200 | "by" - "study" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.712500 |   "by" - "sea" |       0.750000 |       0.600000 |
  0.750000 |       0.750000
       0.223214 |"by" - "religious" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.271593 |    "by" - "of" |       0.258199 |       0.117647 |
  0.500000 |       0.210526
       0.458333 |"by" - "narrative" |       0.500000 |       0.333333 |
     0.500000 |       0.500000
       0.223214 |"by" - "limited" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.180200 |    "by" - "in" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"by" - "government" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.223214 |"by" - "george" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"by" - "general" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.268597 | "by" - "forms" |       0.288675 |       0.166667 |
  0.333333 |       0.285714
       0.223214 |   "by" - "for" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"by" - "female" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 | "by" - "fairy" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"by" - "establishing" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.205207 |"by" - "collection" |       0.223607 |       0.125000 |
      0.250000 |       0.222222
       0.712500 |  "by" - "city" |       0.750000 |       0.600000 |
  0.750000 |       0.750000
       0.458333 |"by" - "circumstantial" |       0.500000 |       0.33333
```

```
3 |       0.500000 |       0.500000
       0.223214 |"by" - "christmas" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.223214 |"by" - "child's" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.180200 |  "by" - "case" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"biography" - "wales" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.365956 |"biography" - "the" |       0.377964 |       0.222222 |
      0.500000 |       0.363636
       0.458333 |"biography" - "tales" |       0.500000 |       0.333333
 |       0.500000 |       0.500000
       0.365956 |"biography" - "study" |       0.377964 |       0.222222
 |       0.500000 |       0.363636
       0.223214 |"biography" - "sea" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.223214 |"biography" - "religious" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.410147 |"biography" - "of" |       0.387298 |       0.187500 |
     0.750000 |       0.315789
       0.458333 |"biography" - "narrative" |       0.500000 |       0.333
333 |       0.500000 |       0.500000
       0.458333 |"biography" - "limited" |       0.500000 |       0.33333
3 |       0.500000 |       0.500000
       0.180200 |"biography" - "in" |       0.188982 |       0.100000 |
     0.250000 |       0.181818
       0.223214 |"biography" - "government" |       0.250000 |       0.14
2857 |       0.250000 |       0.250000
       0.712500 |"biography" - "george" |       0.750000 |       0.600000
|       0.750000 |       0.750000
       0.712500 |"biography" - "general" |       0.750000 |       0.60000
0 |       0.750000 |       0.750000
       0.553861 |"biography" - "forms" |       0.577350 |       0.400000
 |       0.666667 |       0.571429
       0.223214 |"biography" - "for" |       0.250000 |       0.142857 |
      0.250000 |       0.250000
       0.458333 |"biography" - "female" |       0.500000 |       0.333333
|       0.500000 |       0.500000
       0.458333 |"biography" - "fairy" |       0.500000 |       0.333333
 |       0.500000 |       0.500000
       0.223214 |"biography" - "establishing" |       0.250000 |       0.
142857 |       0.250000 |       0.250000
       0.419343 |"biography" - "collection" |       0.447214 |       0.28
5714 |       0.500000 |       0.444444
       0.223214 |"biography" - "city" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.458333 |"biography" - "circumstantial" |       0.500000 |
 0.333333 |       0.500000 |       0.500000
       0.223214 |"biography" - "christmas" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.223214 |"biography" - "child's" |       0.250000 |       0.14285
7 |       0.250000 |       0.250000
       0.365956 |"biography" - "case" |       0.377964 |       0.222222 |
       0.500000 |       0.363636
       0.223214 |"biography" - "by" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
```

```
       0.223214 |"bill" - "wales" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.180200 | "bill" - "the" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"bill" - "tales" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.180200 |"bill" - "study" |       0.188982 |       0.100000 |
   0.250000 |       0.181818
       0.223214 | "bill" - "sea" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.712500 |"bill" - "religious" |       0.750000 |       0.600000 |
       0.750000 |       0.750000
       0.134980 |  "bill" - "of" |       0.129099 |       0.055556 |
  0.250000 |       0.105263
       0.223214 |"bill" - "narrative" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"bill" - "limited" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.180200 |  "bill" - "in" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |"bill" - "government" |       0.250000 |       0.142857
 |       0.250000 |       0.250000
       0.223214 |"bill" - "george" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"bill" - "general" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.268597 |"bill" - "forms" |       0.288675 |       0.166667 |
   0.333333 |       0.285714
       0.712500 | "bill" - "for" |       0.750000 |       0.600000 |
  0.750000 |       0.750000
       0.223214 |"bill" - "female" |       0.250000 |       0.142857 |
    0.250000 |       0.250000
       0.223214 |"bill" - "fairy" |       0.250000 |       0.142857 |
   0.250000 |       0.250000
       0.712500 |"bill" - "establishing" |       0.750000 |       0.60000
0 |       0.750000 |       0.750000
       0.205207 |"bill" - "collection" |       0.223607 |       0.125000
 |       0.250000 |       0.222222
       0.223214 |"bill" - "city" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"bill" - "circumstantial" |       0.250000 |       0.142
857 |       0.250000 |       0.250000
       0.223214 |"bill" - "christmas" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.223214 |"bill" - "child's" |       0.250000 |       0.142857 |
     0.250000 |       0.250000
       0.180200 |"bill" - "case" |       0.188982 |       0.100000 |
  0.250000 |       0.181818
       0.223214 |  "bill" - "by" |       0.250000 |       0.142857 |
  0.250000 |       0.250000
       0.223214 |"bill" - "biography" |       0.250000 |       0.142857 |
       0.250000 |       0.250000
       0.334842 |  "a" - "wales" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.465201 |    "a" - "the" |       0.436436 |       0.214286 |
  0.857143 |       0.352941
       0.334842 |  "a" - "tales" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.465201 |  "a" - "study" |       0.436436 |       0.214286 |
  0.857143 |       0.352941
       0.334842 |    "a" - "sea" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |"a" - "religious" |       0.288675 |       0.107143 |
    0.750000 |       0.193548
       0.698916 |     "a" - "of" |       0.695666 |       0.500000 |
  0.933333 |       0.666667
       0.334842 |"a" - "narrative" |       0.288675 |       0.107143 |
    0.750000 |       0.193548
       0.334842 |"a" - "limited" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.465201 |     "a" - "in" |       0.436436 |       0.214286 |
  0.857143 |       0.352941
       0.334842 |"a" - "government" |       0.288675 |       0.107143 |
     0.750000 |       0.193548
       0.334842 | "a" - "george" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |"a" - "general" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.273413 |  "a" - "forms" |       0.222222 |       0.071429 |
  0.666667 |       0.133333
       0.334842 |    "a" - "for" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 | "a" - "female" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |  "a" - "fairy" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |"a" - "establishing" |       0.288675 |       0.107143 |
       0.750000 |       0.193548
       0.384281 |"a" - "collection" |       0.344265 |       0.142857 |
     0.800000 |       0.250000
       0.334842 |   "a" - "city" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |"a" - "circumstantial" |       0.288675 |       0.107143
|       0.750000 |       0.193548
       0.334842 |"a" - "christmas" |       0.288675 |       0.107143 |
    0.750000 |       0.193548
       0.334842 |"a" - "child's" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.465201 |   "a" - "case" |       0.436436 |       0.214286 |
  0.857143 |       0.352941
       0.334842 |     "a" - "by" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
       0.334842 |"a" - "biography" |       0.288675 |       0.107143 |
    0.750000 |       0.193548
       0.334842 |   "a" - "bill" |       0.288675 |       0.107143 |
  0.750000 |       0.193548
_________________________________________________________________________

Systems test  2  - Similarity measures
_________________________________________________________________________

        average |           pair |         cosine |        jaccard |
   overlap |           dice
-------------------------------------------------------------------------
```

```
-------------------------------------
       0.389562 |"cava" - "dipped" |       0.408248 |       0.250000 |
    0.500000 |       0.400000
       0.625000 |"boon" - "dipped" |       0.666667 |       0.500000 |
    0.666667 |       0.666667
       0.389562 |"boon" - "cava" |       0.408248 |       0.250000 |
  0.500000 |       0.400000
       0.389562 |"atlas" - "dipped" |       0.408248 |       0.250000 |
     0.500000 |       0.400000
       1.000000 |"atlas" - "cava" |       1.000000 |       1.000000 |
   1.000000 |       1.000000
       0.389562 |"atlas" - "boon" |       0.408248 |       0.250000 |
   0.500000 |       0.400000
————————————————————————————————————————————————————————————————————————

Systems test  3  - Similarity measures
————————————————————————————————————————————————————————————————————————

        average |           pair |         cosine |         jaccard |
   overlap |           dice
-------------------------------------------------------------------------
-------------------------------------
       0.346722 |"DocB" - "DocC" |       0.353553 |       0.200000 |
  0.500000 |       0.333333
       0.553861 |"DocA" - "DocC" |       0.577350 |       0.400000 |
  0.666667 |       0.571429
       0.820791 |"DocA" - "DocB" |       0.816497 |       0.666667 |
  1.000000 |       0.800000
```

Pairwise Similairity

In [ ]:

```
────────────────────────────────────────────────────────────────────────────────
Systems test  1  - Similarity measures
────────────────────────────────────────────────────────────────────────────────
   average |                 pair |          cosine |          jaccard |
--------------------------------------------------------------------------------
  1.000000 |     female - limited |        1.000000 |         1.000000 |        1
  0.868292 |         fairy - forms |        0.866025 |         0.750000 |        1
  0.868292 |         forms - tales |        0.866025 |         0.750000 |        1
  0.830357 |          case - study |        0.857143 |         0.750000 |        0
  0.712500 | bill - establishing |        0.750000 |         0.600000 |        0
  0.712500 |    christmas - wales |        0.750000 |         0.600000 |        0
  0.712500 |circumstantial - narrative |        0.750000 |         0.600000 |
  0.712500 |              by - sea |        0.750000 |         0.600000 |        0
  0.712500 |             by - city |        0.750000 |         0.600000 |        0
  0.712500 |       child's - wales |        0.750000 |         0.600000 |        0
  0.712500 |   biography - george |        0.750000 |         0.600000 |        0
  0.712500 | child's - christmas |        0.750000 |         0.600000 |        0
  ...

────────────────────────────────────────────────────────────────────────────────
Systems test  2  - Similarity measures
────────────────────────────────────────────────────────────────────────────────
   average |                 pair |          cosine |          jaccard |
--------------------------------------------------------------------------------
  1.000000 |          atlas - cava |        1.000000 |         1.000000 |        1
  0.625000 |         boon - dipped |        0.666667 |         0.500000 |        0
  0.389562 |         cava - dipped |        0.408248 |         0.250000 |        0
  0.389562 |           boon - cava |        0.408248 |         0.250000 |        0
  0.389562 |        atlas - dipped |        0.408248 |         0.250000 |        0
  0.389562 |          atlas - boon |        0.408248 |         0.250000 |        0
────────────────────────────────────────────────────────────────────────────────
Systems test  3  - Similarity measures
────────────────────────────────────────────────────────────────────────────────
   average |                 pair |          cosine |          jaccard |
--------------------------------------------------------------------------------
  0.820791 |           DocA - DocB |        0.816497 |         0.666667 |        1
  0.553861 |           DocA - DocC |        0.577350 |         0.400000 |        0
  0.346722 |           DocB - DocC |        0.353553 |         0.200000 |        0
```

=== END OF PHASE 1 ===

In [ ]: