~ L200921 (Aisha Muhammad Nawaz) ~

~ CS4080 Mining Massive Datasets (BSCS 8A Spring 2024) ~

~ Assignment 1 - Map Reduce Basics (Due Date: 12 Feb 2024) ~

SUBMISSION: Upload the Source code and the output file on Google Classroom in a zip file with your roll number.

INPUT FILE: You are given an input text file named citation.txt. It contains information regarding the research papers published in various journals. The complete file Citation-network V1 can be found at [https://cn.aminer.org/citation (https://cn.aminer.org/citation)](https://cn.aminer.org/citation). The format of the file is as follows:

* --- paperTitle

@ --- Authors

t ---- Year

c --- publication venue

index 00---- index id of this paper

QUESTION: Write an efficient MapReduce program for the following problems. To make your algorithm efficient, you should use combiners or in-mapper aggregation techniques that use arrays.

SOLUTION

Note: GC File is the file uploaded on Google Classroom assignment post

1. Process the citation.txt input file and output the number of papers published in each decade: 1970s, 1980s, 1990s, 2000s, 2010s, and 2020s.

In [1]:

```
%%file q1.py
#*NOTE: I am assuming the decades mentioned in the question are the only ones to output the count of.
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

class PapersPublishedEachDecade(MRJob):
    def configure_args(self):
        super(PapersPublishedEachDecade, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once

    def init_read_file(self):
        self.patternYear=re.compile(r"(#t[^\#]*)")
        self.papersPerDecade={
            '1970s':0,
            '1980s':0,
            '1990s':0,
            '2000s':0,
            '2010s':0,
            '2020s':0,
            'OTHERS':0
        }
    def get_papers_count(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if file:
            year=''.join(self.patternYear.findall(file[0])).replace('#t','').replace('\n','').replace(' ','')
            if(year):
                year=int(year)
                if((year>=1970) and (year<1980)):
                    year='1970s'
                elif((year>=1980) and (year<1990)):
                    year='1980s'
                elif((year>=1990) and (year<2000)):
                    year='1990s'
                elif((year>=2000) and (year<2010)):
                    year='2000s'
                elif((year>=2010) and (year<2020)):
                    year='2010s'
                elif((year>=2020) and (year<2030)):
                    year='2020s'
                else :
                    year='OTHERS'
                self.papersPerDecade[year]=self.papersPerDecade[year]+1

    def final_get_papers_count(self):
        for decade,val in self.papersPerDecade.items():
            yield decade,val

    def sum_decades(self,decade,counts):
        yield decade,sum(counts)

    def steps(self):
        return [
            MRStep(mapper_init=self.init_read_file,
                  mapper=self.get_papers_count,
                  mapper_final=self.final_get_papers_count,
                  combiner=self.sum_decades,
                  reducer=self.sum_decades)
        ]

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

```
Overwriting q1.py
```

Q1 GC FILE OUTPUT

```
In [2]: !python q1.py citation.txt
```

```
"1970s" 4
"1980s" 3
"1990s" 11
"2000s" 82
"2010s" 0
"2020s" 0
"OTHERS"        0

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240212.172204.817701
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240212.172204.817701\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240212.172204.817701\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q1.DELL.20240212.172204.817701...
```

2. Create an inverted index of the citation file. Your inverted index will output the year followed by the comma-separated list of the titles of the papers published in that year.

Sample Output format : Year1 -> PaperTitle, Paper Title Year2 -> Paper Title

In [3]:

```
%%file q2.py
from mrjob.job import MRJob
from mrjob.step import MRStep
from itertools import chain # To flatten the list before the final merge
import re

class InvertedIndexCitations(MRJob):
    def configure_args(self):
        super(InvertedIndexCitations, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternYear=re.compile(r"(#t[^\#]*)")
        self.papersByYear={}
    def get_papers_by_year(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                paperTitle=file[0].split('#@')[0]
                year=''.join(self.patternYear.findall(file[0])).replace('#t','').replace('\n','').replace('
','')
                if(year):
                    self.papersByYear.setdefault(year+" -> ",[]).append(paperTitle)

    def final_get_papers_by_year(self):
        for year,papers in self.papersByYear.items():
            yield year,papers

    def merge_papers_by_year(self,year,papers):
        yield year,list(chain.from_iterable(papers))


    def steps(self):
        return [
            MRStep(
                  mapper_init=self.init_read_file,
                  mapper=self.get_papers_by_year,
                  mapper_final=self.final_get_papers_by_year,
                  combiner=self.merge_papers_by_year,
                  reducer=self.merge_papers_by_year)
        ]
```

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

```
Overwriting q2.py
```

Q2 GC FILE OUTPUT

```
In [4]: !python q2.py citation.txt
```

```
"1973 -> "      ["Notes from industry"]
"1975 -> "      ["A control word model for detecting conflicts between microoperations "]
"1976 -> "      ["Microprogramming for the hardware engineer"]
"1978 -> "      ["Design team composition for high level language computer architectures "]
"1982 -> "      ["Review of \"Bit-Slice Microprocessor Design by John Mick and James Brick\", McGraw-Hill Boo
k Company, 1980 "]
"1985 -> "      ["Word Processing on Your MacIntosh "]
"1987 -> "      ["Type Graphics and MacIntosh"]
"1991 -> "      ["Tarski's World 3.0: Including the Macintosh TM Program (Center for the Study of Language an
d Information - Lecture Notes) "]
"1993 -> "      ["Hyperstat: Macintosh Hypermedia for Analyzing Data and Learning Statistics"]
"1994 -> "      ["At Ease With Performa","It's a Mad, Mad, Mad, Mad Mac\/Book and Disk ","Operations Researc
h: Macintosh Version (Business Statistics Series) "]
"1995 -> "      ["Internet and HTML Training on CD-ROM "]
"1996 -> "      ["Fast k-NN Classification Rule Using Metrics on Space-Filling Curves","A New Quadtree Decomp
osition Reconstruction Method"]
"1997 -> "      ["Multimedia Directory 1997","Elsevier's Dictionary of Wild and Cultivated Plants "]
"1999 -> "      ["Electronic Engineer's Handbook (Core Handbook CD-ROMs)"]
"2000 -> "      ["Exploring Macintosh Concepts in Visually Oriented Computing & Computing Concepts for End Us
ers ","Tips and Tuning Guide for MS Flight Simulator 2000 "]
"2001 -> "      ["Conker's Bad Fur Day (Prima's Official Strategy Guide)"]
"2002 -> "      ["ECDL Advanced "]
"2003 -> "      ["A+ Certification Core Hardware (Text & Lab Manual)","Start with a Digital Camera (Special E
dition) (2nd Edition) (Start with a)","Guia Visual de Microsoft Office 2000\/ Microsoft Office 2000 Visual Gu
ide (Guias Visuales)","KeyChamp 2.0 Macintosh Site License Package "]
"2004 -> "      ["Program Evaluation: Improving The Flow Of Information To The Congress","ExamInsight For MCP
\/ MCSE Certification: Installing, Configuring, and Administering Microsoft Windows XP Professional Exam 70-2
70 (ExamInsight) ","Ibook Fan Book: Smart and Beautiful to Boot (Ibook Fan Books) ","GO Series: Microsoft Exc
el 2003 Volume 2 (Go With Microsoft Office) ","Data Structures ","Adaptive Multimedia Retrieval: First Intern
ational Workshop, AMR 2003, Hamburg, Germany, September 15-16, 2003, Revised Selected and Invited Papers (Lec
ture Notes in Computer Science) ","The Prentice Hall Planner for Student Success ","Keno Winner: A Guide To W
inning At Video Keno ","Microsoft Powerpoint 2003 (Marquee Series) "]
"2005 -> "      ["Dude, You Can Do It! How to Build a Sweeet PC","Dependable Computing","Making the Digital C
ity: The Early Shaping of Urban Internet Space (Design & the Built Environment S.)","Linspire 5.0: The No Non
sense Guide! (No Nonsense Guide! series)","Federated Identity Management And Web Services Security With IBM T
ivoli Security Solutions","Open Process Frameworks: Patterns for the Adaptive e-Enterprise (Practitioner
s)","ASIS&T Thesaurus of Information Science, Technology, And Librarianship (Asist Monograph Series)","Call o
f Duty 2: Big Red One(tm) Official Strategy Guide (Official Strategy Guides)","Jocelyn Robert: Aucune de mes
mains ne fait mal","Special Edition Using Adobe Creative Suite 2 (Special Edition Using)","TCP\/IP Protocol S
uite, 3 edition","Computer Models of Musical Creativity ","Microsoft Word 2003 Advanced ","Gungrave: 2006 Wal
l Calendar ","Essentials for Design Adobe Illustrator CS 2 - Level 1 (2nd Edition) (Essentials for Design)
","The Game Producer's Handbook ","Visual Basic 2005 Demystified, 1 edition ","Oracle 10g PL\/SQL: Guia de Co
nsulta R\u00e1pida "]
```

```
"2006 -> "	["Automated Deduction in Geometry ","What Every Programmer Needs to Know about Security (Adva
nces in information Security)","Interpreting Kullback-Leibler divergence with the Neyman-earson lemma","Digit
al Media: Transformations in Human Communication","Adaptive Hypermedia and Adaptive Web-Based Systems","Calcu
lus Early Transcendentals Single Variable","Approximating fluid schedules in crossbar packet-switches and Ban
yan networks","Fast and Efficient Context-Aware Services (Wiley Series on Communications Networking & Distrib
uted Systems)","Inside SQL Server 2005 Tools (Microsoft Windows Server System Series)","Inside Microsoft Dyna
mics AX 4.0 ","Wiley Plus\/Web CT Stand-alone to accompany Java Concepts (Wiley Plus Products)","Modeling met
hodology b: distributed simulation and the high level architecture","Beginning Ruby on Rails (Wrox Beginning 
Guides)","SUSE Linux Enterprise Server Administration (Course 3037)","An Integrative Modelling Approach for S
imulation and Analysis of Adaptive Agents","Computer Accounting with QuickBooks 2006","Class-specific feature
polynomial classifier for pattern classification and its application to handwritten numeral recognition","F.
E.A.R.: First Encounter Assault Recon (Prima Official Game Guide) ","Effects of reward expectancy on sequenti
al eye movements in monkeys ","Computer Concepts Illustrated Complete, Sixth Edition (Illustrated (Thompson L
earning)) ","Windows Vista All-in-One Desk Reference For Dummies (For Dummies (Computer\/Tech)) ","DV Filmmak
ing: From Start to Finish (O'Reilly Digital Studio)","Hands-On Guide to Video Blogging and Podcasting: Emergi
ng Media Tools for Business Communication (Hands-on Guide) ","Mage Knight(tm): Apocalypse Official Strategy G
uide (Official Strategy Guides) ","The Effect of Faults on Network Expansion ","Mesoscopic simulation of Ostw
ald ripening ","Gurus, Hired Guns, and Warm Bodies: Itinerant Experts in a Knowledge Economy ","Selected Area
s in Cryptography: 12th International Workshop, SAC 2005, Kingston, ON, Canada, August 11-12, 2005, Revised S
elected Papers (Lecture Notes in Computer Science) ","Real World Aperture (Real World) ","Making Projects Cri
tical (Management, Work and Organisations) ","Three-Level Caching for Efficient Query Processing in Large Web
Search Engines ","Creative Code: \u00c4sthetik und Programmierung am MIT Media Lab ","On an initial transient
deletion rule with rigorous theoretical support ","MICAI 2006: Advances in Artificial Intelligence: 5th Mexic
an International Conference on Artificial IntelligenceApizaco, Mexico, November 13-17, 2006Proceedings (Lectu
re Notes in Computer Science) ","Special issue: Dialog systems for health communications "]
"2007 -> "	["Performance engineering in industry: current practices and adoption challenges","TOPP---the
OpenMS proteomics pipeline","Webbots, Spiders, and Screen Scrapers","Podcasting for Profit: A Proven 10-Step 
Plan for Generating Income Through Audio and Video Podcasting","Introduction to Information Systems","Microso
ft Expression Web: Visual QuickStart Guide","A method to compute distance between two categorical values of s
ame attribute in unsupervised learning for categorical data set ","Database Modeling in Biology: Practices an
d Challenges ","The Internet: A Critical Introduction ","Java for Everyone ","CompTIA A+ Exam Cram (Exams 220
-602, 220-603, 220-604) (Exam Cram) "]

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q2.DELL.20240212.172206.992178
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q2.DELL.20240212.172206.992178\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q2.DELL.20240212.172206.992178\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q2.DELL.20240212.172206.992178...
```

3. Produce a list of co-authors of each author in the given input file.

Sample Output (Author -> List of Co -authors) David Jones -> Sam Nick, Ali Javed , Daniel Brown Sam Nick -> David Jones, Zan Jao, Ali Javed Ali Javed -> David Jones ,Sam Nick Zan Jao -> Sam Nick Daniel Brown -> David Jones

Consider the following citation given in the input file

*Automated Deduction in Geometry #@Hoon Hong,Dongming Wang#t2006#index0

In this citation Hoon Hong and Dongming Wang are coauthors of each other as they have written one paper together. So in other word if A has written a book with B then A is coauthor of B and B is coauthor of A

In [5]:

```
%%file q3.py
from mrjob.job import MRJob
from mrjob.step import MRStep
from itertools import chain # To flatten the list before the final merge
import re

class CitationsCoauthors(MRJob):
    def configure_args(self):
        super(CitationsCoauthors, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternAuthors=re.compile(r"\#\@[^\#\t]*")
        self.authorsCoauthors={}
    def get_authors_coauthors(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                paperTitle=file[0].split('#@')[0]
                authors=''.join(self.patternAuthors.findall(file[0])).replace('#@','').replace('\n','')
                authors=[isAuthor for isAuthor in authors.split(',') if len(isAuthor)>1]
                for author in authors:
                    author=author.strip()
                    for coAuthor in authors:
                        coAuthor=coAuthor.strip()
                        if(not(coAuthor==author)):
                            self.authorsCoauthors.setdefault(author+" - >",[]).append(coAuthor)

    def final_get_authors_coauthors(self):
        for author,coAuthor in self.authorsCoauthors.items():
            yield author,coAuthor

    def merge_authors_coauthors(self,author,coAuthor):
        yield author,list(chain.from_iterable(coAuthor))


    def steps(self):
        return [
            MRStep(
                  mapper_init=self.init_read_file,
                  mapper=self.get_authors_coauthors,
                  mapper_final=self.final_get_authors_coauthors,
                  combiner=self.merge_authors_coauthors,
                  reducer=self.merge_authors_coauthors)
        ]

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

```
Overwriting q3.py
```

Q3 GC FILE OUTPUT

In [6]:

```
!python q3.py citation.txt
```

```
"A. Krzyzak - >"        ["E. Skubalska-Rafajtowicz"]
"Ahmed Hassan - >"      ["Parminder Flora"]
"Alex Galis - >"        ["Danny Raz","Arto Tapani Juhola","Joan Serrat-Fernandez"]
"Alexander Gelbukh - >" ["Carlos Alberto Reyes-Garcia"]
"Alice Redmond-neal - >"        ["Marjorie M. K. Hlava"]
"Aline Maria Santos Andrade - >"        ["Carlos Alberto Maziero","Jo\u00e3o Gabriel Silva","Fl\u00e1vio Mora
is de Assis Silva"]
"Amir Ahmad - >"        ["Lipika Dey"]
"Amitabh Chaudhary - >" ["Amitabha Bagchi","Ankur Bhargava","David Eppstein","Christian Scheideler"]
"Amitabha Bagchi - >"   ["Ankur Bhargava","Amitabh Chaudhary","David Eppstein","Christian Scheideler"]
"Andreas N\u00fcrnberger - >"   ["Marcin Detyniecki"]
"Anita Kesavan - >"     ["Neil Daswani"]
"Ankur Bhargava - >"    ["Amitabha Bagchi","Amitabh Chaudhary","David Eppstein","Christian Scheideler"]
"Arthur Greef - >"      ["Michael Fruergaard Pontoppidan","Lars Dragheim Olsen","Palle Agermark","Hans J. Sko
vgaard"]
"Arto Tapani Juhola - >"        ["Danny Raz","Joan Serrat-Fernandez","Alex Galis"]
"Barry Smyth - >"       ["Vincent Wade","Helen Ashman"]
"Brenden Munnelly - >"  ["Paul Holden"]
"Bruce Shriver - >"     ["Ted Lewis"]
"Carlito Vicencio - >"  ["Darrel Creacy"]
"Carlos Alberto Maziero - >"    ["Jo\u00e3o Gabriel Silva","Aline Maria Santos Andrade","Fl\u00e1vio Morais d
e Assis Silva"]
"Carlos Alberto Reyes-Garcia - >"       ["Alexander Gelbukh"]
"Catholijn M. Jonker - >"       ["Tibor Bosse","Jan Treur"]
"Charles S. Wetherell - >"      ["Lyle A. Cox","Jr.","James R. McGraw"]
"Cheng-Lin Liu - >"     ["Hiroshi Sako"]
"Christian Scheideler - >"      ["Amitabha Bagchi","Ankur Bhargava","Amitabh Chaudhary","David Eppstein"]
"Clemens Gr\u00f6pl - >"        ["Oliver Kohlbacher","Knut Reinert","Eva Lange","Nico Pfeifer","Ole Schulz-Tr
ieglaff","Marc Sturm"]
"Constantine Caramanis - >"     ["Michael Rosenblum","Michel X. Goemans","Vahid Tarokh"]
"Daeyeol Lee - >"       ["Jeong-woo Sohn"]
"Damian Hodgson - >"    ["Svetlana Cicmil"]
"Damien Stolarz - >"    ["Lionel Felix"]
"Dan Oja - >"   ["June Jamrich Parsons"]
"Danny Raz - >" ["Arto Tapani Juhola","Joan Serrat-Fernandez","Alex Galis"]
"Darrel Creacy - >"     ["Carlito Vicencio"]
"David Eppstein - >"    ["Amitabha Bagchi","Ankur Bhargava","Amitabh Chaudhary","Christian Scheideler"]
"Denise Seguin - >"     ["Nita Hewitt Rutkosky"]
"Dima Sonkin - >"       ["Michael Raheem","Thierry D'Hers","Kami LeMonds"]
"Dongming Wang - >"     ["Hoon Hong"]
"E. Skubalska-Rafajtowicz - >"  ["A. Krzyzak"]
"Eva Lange - >" ["Oliver Kohlbacher","Knut Reinert","Clemens Gr\u00f6pl","Nico Pfeifer","Ole Schulz-Trieglaf
f","Marc Sturm"]
```

```
"Fenghui Zhang - >"     ["Jianer Chen"]
"Fl\u00e1vio Morais de Assis Silva - >" ["Carlos Alberto Maziero","Jo\u00e3o Gabriel Silva","Aline Maria Sant
os Andrade"]
"Gideon Kunda - >"      ["Stephen R. Barley"]
"Hans J. Skovgaard - >" ["Arthur Greef","Michael Fruergaard Pontoppidan","Lars Dragheim Olsen","Palle Agermar
k"]
"Helen Ashman - >"      ["Vincent Wade","Barry Smyth"]
"Hernan P. Awad - >"    ["Peter W. Glynn"]
"Hiroshi Sako - >"      ["Cheng-Lin Liu"]
"Hoon Hong - >" ["Dongming Wang"]
"J. Knipe - >"  ["X. Li"]
"Jake Chen - >" ["Zongmin Ma"]
"James R. McGraw - >"   ["Lyle A. Cox","Jr.","Charles S. Wetherell"]
"Jan Treur - >" ["Tibor Bosse","Catholijn M. Jonker"]
"Jeong-woo Sohn - >"    ["Daeyeol Lee"]
"Jianer Chen - >"       ["Fenghui Zhang"]
"Jo\u00e3o Gabriel Silva - >"   ["Carlos Alberto Maziero","Aline Maria Santos Andrade","Fl\u00e1vio Morais de
Assis Silva"]
"Joan Serrat-Fernandez - >"     ["Danny Raz","Arto Tapani Juhola","Alex Galis"]
"John Copas - >"        ["Shinto Eguchi"]
"John Etchemendy - >"   ["Jon Barwise"]
"John Maeda - >"        ["Red Burns"]
"John Preston - >"      ["Shelley Gaskin","Sally Preston"]
"Jon Barwise - >"       ["John Etchemendy"]
"Jr. - >"       ["Lyle A. Cox","James R. McGraw","Charles S. Wetherell"]
"June Jamrich Parsons - >"      ["Dan Oja"]
"Kami LeMonds - >"      ["Michael Raheem","Dima Sonkin","Thierry D'Hers"]
"Knut Reinert - >"      ["Oliver Kohlbacher","Clemens Gr\u00f6pl","Eva Lange","Nico Pfeifer","Ole Schulz-Trie
glaff","Marc Sturm"]
"Lars Dragheim Olsen - >"       ["Arthur Greef","Michael Fruergaard Pontoppidan","Palle Agermark","Hans J. Sk
ovgaard"]
"Lee Humphreys - >"     ["Paul Messaris"]
"Lionel Felix - >"      ["Damien Stolarz"]
"Lipika Dey - >"        ["Amir Ahmad"]
"Lyle A. Cox - >"       ["Jr.","James R. McGraw","Charles S. Wetherell"]
"Marc Sturm - >"        ["Oliver Kohlbacher","Knut Reinert","Clemens Gr\u00f6pl","Eva Lange","Nico Pfeife
r","Ole Schulz-Trieglaff"]
"Marcin Detyniecki - >" ["Andreas N\u00fcrnberger"]
"Marjorie M. K. Hlava - >"      ["Alice Redmond-neal"]
"Michael Fruergaard Pontoppidan - >"    ["Arthur Greef","Lars Dragheim Olsen","Palle Agermark","Hans J. Skovg
aard"]
"Michael Raheem - >"    ["Dima Sonkin","Thierry D'Hers","Kami LeMonds"]
"Michael Rosenblum - >" ["Constantine Caramanis","Michel X. Goemans","Vahid Tarokh"]
```

```
"Michael Schrenk - >"   ["Michael Shrenk"]
"Michael Shrenk - >"    ["Michael Schrenk"]
"Michel X. Goemans - >" ["Michael Rosenblum","Constantine Caramanis","Vahid Tarokh"]
"Neil Daswani - >"      ["Anita Kesavan"]
"Nico Pfeifer - >"      ["Oliver Kohlbacher","Knut Reinert","Clemens Gr\u00f6pl","Eva Lange","Ole Schulz-Trie
glaff","Marc Sturm"]
"Nita Hewitt Rutkosky - >"      ["Denise Seguin"]
"Ole Schulz-Trieglaff - >"      ["Oliver Kohlbacher","Knut Reinert","Clemens Gr\u00f6pl","Eva Lange","Nico Pf
eifer","Marc Sturm"]
"Oliver Kohlbacher - >" ["Knut Reinert","Clemens Gr\u00f6pl","Eva Lange","Nico Pfeifer","Ole Schulz-Trieglaf
f","Marc Sturm"]
"Palle Agermark - >"    ["Arthur Greef","Michael Fruergaard Pontoppidan","Lars Dragheim Olsen","Hans J. Skovg
aard"]
"Parminder Flora - >"   ["Ahmed Hassan"]
"Paul Holden - >"       ["Brenden Munnelly"]
"Paul Messaris - >"     ["Lee Humphreys"]
"Peter W. Glynn - >"    ["Hernan P. Awad"]
"Red Burns - >" ["John Maeda"]
"Sally Preston - >"     ["Shelley Gaskin","John Preston"]
"Shelley Gaskin - >"    ["John Preston","Sally Preston"]
"Shinto Eguchi - >"     ["John Copas"]
"Stephen R. Barley - >" ["Gideon Kunda"]
"Svetlana Cicmil - >"   ["Damian Hodgson"]
"Ted Lewis - >" ["Bruce Shriver"]
"Thierry D'Hers - >"    ["Michael Raheem","Dima Sonkin","Kami LeMonds"]
"Tibor Bosse - >"       ["Catholijn M. Jonker","Jan Treur"]
"Torsten Suel - >"      ["Xiaohui Long"]
"Vahid Tarokh - >"      ["Michael Rosenblum","Constantine Caramanis","Michel X. Goemans"]
"Vincent Wade - >"      ["Helen Ashman","Barry Smyth"]
"X. Li - >"     ["J. Knipe"]
"Xiaohui Long - >"      ["Torsten Suel"]
"Zongmin Ma - >"        ["Jake Chen"]

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q3.DELL.20240212.172209.172585
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q3.DELL.20240212.172209.172585\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q3.DELL.20240212.172209.172585\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q3.DELL.20240212.172209.172585...
```

4. Find the average number of papers published each year.

In [7]:

```
%%file q4.py
#*NOTE: I am assuming the average number of papers published each year means Total papers published / Total Distinct Years.
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

class AvgPapersPublishedEachYear(MRJob):
    def configure_args(self):
        super(AvgPapersPublishedEachYear, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternYear=re.compile(r"(#t[^\#]*)")
        self.papersPerYearSum={}
    def get_papers_count(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if file:
            year=''.join(self.patternYear.findall(file[0])).replace('#t','').replace('\n','').replace(' ','')
            if(year):
                year=int(year)
                self.papersPerYearSum.setdefault(year,0)
                self.papersPerYearSum[year]=self.papersPerYearSum[year]+1

    def final_get_papers_count(self):
        for year,value in self.papersPerYearSum.items():
            yield year,(value,1)

    def sum_years_count(self,year,value):
        sumValue=0
        for val,count in value:
            sumValue=sumValue+val
        yield year,(sumValue,1)

    def sum_years_count_red(self,year,value):
        sumValue=0
        sumCount=1
        for val,count in value:
            sumValue=sumValue+val
        yield None,(sumValue,sumCount)

    def avg_years_count(self,year,value):
        sumValue=0
        sumCount=0
        for val,count in value:
            sumValue=sumValue+val
            sumCount=sumCount+count
        yield "Average Papers Published Each Year = ",(sumValue/sumCount)

    def steps(self):
        return [
            MRStep(mapper_init=self.init_read_file,
                  mapper=self.get_papers_count,
                  mapper_final=self.final_get_papers_count,
                  combiner=self.sum_years_count,
                  reducer=self.sum_years_count_red
                  ),
                  MRStep(reducer=self.avg_years_count)
        ]

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

```
Overwriting q4.py
```

Q4 GC FILE OUTPUT

In [8]:

```
!python q4.py citation.txt
```

```
"Average Papers Published Each Year = " 4.545454545454546

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q4.DELL.20240212.172211.380475
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\DELL\AppData\Local\Temp\q4.DELL.20240212.172211.380475\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q4.DELL.20240212.172211.380475\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q4.DELL.20240212.172211.380475...
```

5. List the names of authors who have written the maximum number of papers.

In [9]:

```
%%file q5.py
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

class CitationsAuthorsMax(MRJob):
    def configure_args(self):
        super(CitationsAuthorsMax, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternAuthors=re.compile(r"\#\@[^\#\t]*")
        self.authorsCount={}
    def get_authors_count(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                authors=''.join(self.patternAuthors.findall(file[0])).replace('#@','').replace('\n','')
                authors=[isAuthor for isAuthor in authors.split(',') if len(isAuthor)>1]
                for author in authors:
                    author=author.strip()
                    self.authorsCount.setdefault(author,0)
                    self.authorsCount[author]=self.authorsCount[author]+1

    def final_get_authors_count(self):
        for author,count in self.authorsCount.items():
            yield (author,count)

    def sum_authors_count_combiner(self,author,count):
        yield (author,sum(count))

    def sum_authors_count_reducer(self,author,count):
        yield (None,(sum(count),author))

    def max_authors_count(self,key,values):
        max_value = float('-inf')  # Initialize to the smallest possible value
        max_authors = []

        for count, author in values:
            if count > max_value:
                max_value = count
                max_authors = [(author, count)]
            elif count == max_value:
                max_authors.append((author, count))

        for author, count in max_authors:
              yield (author, count)

    def steps(self):
        return [
            MRStep(
                  mapper_init=self.init_read_file,
                  mapper=self.get_authors_count,
                  mapper_final=self.final_get_authors_count,
                  combiner=self.sum_authors_count_combiner,
                  reducer=self.sum_authors_count_reducer),
            MRStep(reducer=self.max_authors_count)
        ]

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

```
Overwriting q5.py
```

Q5 GC FILE OUTPUT

In [10]:

```
!python q5.py citation.txt
```

```
"Cay S. Horstmann"      2
"Charles J. Brooks"     2

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q5.DELL.20240212.172213.630325
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\DELL\AppData\Local\Temp\q5.DELL.20240212.172213.630325\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q5.DELL.20240212.172213.630325\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q5.DELL.20240212.172213.630325...
```

6. Find the names of authors who have written at most one paper in a year.

In [11]:

```
%%file q6.py
#*NOTE: I am assuming here 'at most one paper in a year' means the author has some year in which they wrote o
ne paper (This may or may not be their max in all years)
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

class CitationsAuthorsPerYearCountMaxOne(MRJob):
    def configure_args(self):
        super(CitationsAuthorsPerYearCountMaxOne, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternAuthors=re.compile(r"\#\@[^\#\t]*")
        self.patternYear=re.compile(r"(#t[^\#]*)")
        self.authorsCount={}
    def get_authors_count(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                year=''.join(self.patternYear.findall(file[0])).replace('#t','').replace('\n','').replace('
',''')
                if(year):
                    authors=''.join(self.patternAuthors.findall(file[0])).replace('#@','').replace('\n','')
                    authors=[isAuthor for isAuthor in authors.split(',') if len(isAuthor)>1]
                    for author in authors:
                        author=author.strip()
                        self.authorsCount.setdefault(author+'->'+year,0)
                        self.authorsCount[author+'->'+year]=self.authorsCount[author+'->'+year]+1

    def final_get_authors_count(self):
        for authorYear,count in self.authorsCount.items():
            yield authorYear,count

    def sum_authors_count_combiner(self,authorYear,count):
        yield authorYear,sum(count)

    def sum_authors_count_reducer(self,authorYear,count):
        valueSummed=sum(count)
        if(valueSummed==1):
            yield None,(valueSummed,authorYear)

    def maxOne_authors_count(self,key,value):
        for value,authorYear in value:
            authorYear=authorYear.split('->')
            yield 'Author: '+authorYear[0],' Year: '+authorYear[1]


    def steps(self):
        return [
            MRStep(
                  mapper_init=self.init_read_file,
                  mapper=self.get_authors_count,
                  mapper_final=self.final_get_authors_count,
                  combiner=self.sum_authors_count_combiner,
                  reducer=self.sum_authors_count_reducer),
            MRStep(reducer=self.maxOne_authors_count)
        ]

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

```
Overwriting q6.py
```

Q6 GC FILE OUTPUT

In [12]:

```
!python q6.py citation.txt
```

```
"Author: A. Krzyzak"    " Year: 1996"
"Author: Ahmed Hassan"  " Year: 2007"
"Author: Alessandro Aurigi"     " Year: 2005"
"Author: Alex Galis"    " Year: 2006"
"Author: Alexander Gelbukh"     " Year: 2006"
"Author: Alice Redmond-neal"    " Year: 2005"
"Author: Aline Maria Santos Andrade"    " Year: 2005"
"Author: Allan Hunkin"  " Year: 2007"
"Author: Amir Ahmad"    " Year: 2007"
"Author: Amitabh Chaudhary"     " Year: 2006"
"Author: Amitabha Bagchi"       " Year: 2006"
"Author: Andreas N\u00fcrnberger"       " Year: 2004"
"Author: Anita Kesavan" " Year: 2006"
"Author: Ankur Bhargava"        " Year: 2006"
"Author: Arthur Greef"  " Year: 2006"
"Author: Arto Tapani Juhola"    " Year: 2006"
"Author: Axel Bucker"   " Year: 2005"
"Author: Barry Smyth"   " Year: 2006"
"Author: Bart Preneel"  " Year: 2006"
"Author: Behrouz A. Forouzan"   " Year: 2005"
"Author: Ben Long"      " Year: 2006"
"Author: Brenden Munnelly"      " Year: 2002"
"Author: Bruce Shriver" " Year: 1975"
"Author: Carla Rose"    " Year: 1994"
"Author: Carlito Vicencio"      " Year: 2005"
"Author: Carlos Alberto Maziero"        " Year: 2005"
"Author: Carlos Alberto Reyes-Garcia"   " Year: 2006"
"Author: Catholijn M. Jonker"   " Year: 2006"
"Author: Cay S. Horstmann"      " Year: 2006"
"Author: Cay S. Horstmann"      " Year: 2007"
"Author: Celso H. Poderoso de Oliveira" " Year: 2005"
"Author: Charles J. Brooks"     " Year: 2003"
"Author: Charles J. Brooks"     " Year: 2007"
"Author: Charles S. Wetherell"  " Year: 1978"
"Author: Cheng-Lin Liu" " Year: 2006"
"Author: Christian Scheideler"  " Year: 2006"
"Author: Clemens Gr\u00f6pl"    " Year: 2007"
"Author: Constantine Caramanis" " Year: 2006"
"Author: Daeyeol Lee"   " Year: 2006"
"Author: Damian Hodgson"        " Year: 2006"
"Author: Damien Stolarz"        " Year: 2006"
"Author: Dan Irish"     " Year: 2005"
"Author: Dan Oja"       " Year: 2006"
```

```
"Author: Danny Raz"     " Year: 2006"
"Author: Darrel Creacy" " Year: 2005"
"Author: David A. Marca"        " Year: 2005"
"Author: David Cope"    " Year: 2005"
"Author: David Eppstein"        " Year: 2006"
"Author: David J. Horntrop"     " Year: 2006"
"Author: David M. Lane" " Year: 1993"
"Author: Dean Bagley"   " Year: 2005"
"Author: Deborah Timmons"       " Year: 2004"
"Author: Denise Seguin" " Year: 2004"
"Author: Derrick Story" " Year: 2004"
"Author: Dima Sonkin"   " Year: 2006"
"Author: Donald Christiansen"   " Year: 1999"
"Author: Dongming Wang" " Year: 2006"
"Author: Donna Ulmer"   " Year: 2006"
"Author: E. Skubalska-Rafajtowicz"      " Year: 1996"
"Author: Eric Grebler"  " Year: 2005"
"Author: Eva Lange"     " Year: 2007"
"Author: Fl\u00e1vio Morais de Assis Silva"     " Year: 2005"
"Author: Gene Orwell"   " Year: 1994"
"Author: Gideon Kunda"  " Year: 2006"
"Author: Hans J. Skovgaard"     " Year: 2006"
"Author: Helen Ashman"  " Year: 2006"
"Author: Hernan P. Awad"        " Year: 2006"
"Author: Hiroshi Sako"  " Year: 2006"
"Author: Hoon Hong"     " Year: 2006"
"Author: Howard A. Anton"       " Year: 2006"
"Author: Ian David Aronson"     " Year: 2006"
"Author: J. Knipe"      " Year: 1996"
"Author: Jake Chen"     " Year: 2007"
"Author: James R. McGraw"       " Year: 1978"
"Author: Jan Treur"     " Year: 2006"
"Author: Jason Eckert"  " Year: 2006"
"Author: Jeff Kent"     " Year: 2005"
"Author: Jeong-woo Sohn"        " Year: 2006"
"Author: Jo\u00e3o Gabriel Silva"       " Year: 2005"
"Author: Joan Serrat-Fernandez" " Year: 2006"
"Author: Jocelyn Robert"        " Year: 2005"
"Author: John Blaint"   " Year: 1987"
"Author: John Copas"    " Year: 2006"
"Author: John Etchemendy"       " Year: 1991"
"Author: John Maeda"    " Year: 2006"
"Author: John Odam"     " Year: 2003"
```

```
"Author: John Preston"  " Year: 2004"
"Author: John R. Mick"  " Year: 1976"
"Author: Jon Barwise"   " Year: 1991"
"Author: Jose Pedro Llamazares" " Year: 2003"
"Author: Jr."   " Year: 1978"
"Author: June Jamrich Parsons"  " Year: 2006"
"Author: Kami LeMonds"  " Year: 2006"
"Author: Ken Abernethy" " Year: 2000"
"Author: Knut Reinert"  " Year: 2007"
"Author: Korinna Patelis"       " Year: 2007"
"Author: Lars Dragheim Olsen"   " Year: 2006"
"Author: Lee Humphreys" " Year: 2006"
"Author: Lionel Felix"  " Year: 2006"
"Author: Lipika Dey"    " Year: 2007"
"Author: Lyle A. Cox"   " Year: 1978"
"Author: Marc Sturm"    " Year: 2007"
"Author: Marcin Detyniecki"     " Year: 2004"
"Author: Marjorie M. K. Hlava"  " Year: 2005"
"Author: Michael Cloran"        " Year: 2003"
"Author: Michael Fruergaard Pontoppidan"        " Year: 2006"
"Author: Michael Raheem"        " Year: 2006"
"Author: Michael Rosenblum"     " Year: 2006"
"Author: Michael Schrenk"       " Year: 2007"
"Author: Michael Shrenk"        " Year: 2007"
"Author: Michael Smick" " Year: 2005"
"Author: Michael T. Goodrich"   " Year: 2004"
"Author: Michel X. Goemans"     " Year: 2006"
"Author: Neil Daswani"  " Year: 2006"
"Author: Nico Pfeifer"  " Year: 2007"
"Author: Nightow Yoshiro"       " Year: 2005"
"Author: Nita Hewitt Rutkosky"  " Year: 2004"
"Author: Nolan Hester"  " Year: 2007"
"Author: Ole Schulz-Trieglaff"  " Year: 2007"
"Author: Oliver Kohlbacher"     " Year: 2007"
"Author: Palle Agermark"        " Year: 2006"
"Author: Pamela W. Adams"       " Year: 2005"
"Author: Parminder Flora"       " Year: 2007"
"Author: Paul Holden"   " Year: 2002"
"Author: Paul Messaris" " Year: 2006"
"Author: Peter W. Glynn"        " Year: 2006"
"Author: R. Kelly Rainer"       " Year: 2007"
"Author: Red Burns"     " Year: 2006"
"Author: Ron Dulin"     " Year: 2006"
```

```
"Author: Rudolph Langer"        " Year: 1985"
"Author: Sally Preston" " Year: 2004"
"Author: Shelley Gaskin"        " Year: 2004"
"Author: Shinto Eguchi" " Year: 2006"
"Author: Stanley Habib" " Year: 1973"
"Author: Stephen R. Barley"     " Year: 2006"
"Author: Steve Holzner" " Year: 2006"
"Author: Svetlana Cicmil"       " Year: 2006"
"Author: Ted Lewis"     " Year: 1975"
"Author: Thierry D'Hers"        " Year: 2006"
"Author: Tibor Bosse"   " Year: 2006"
"Author: Tom Collins"   " Year: 2004"
"Author: Torsten Suel"  " Year: 2006"
"Author: Vahid Tarokh"  " Year: 2006"
"Author: Vincent Wade"  " Year: 2006"
"Author: W. E. Clason"  " Year: 1997"
"Author: Wayne L. Winston"      " Year: 1994"
"Author: William J. Tracz"      " Year: 1982"
"Author: Woody Leonhard"        " Year: 2006"
"Author: X. Li" " Year: 1996"
"Author: Xiaohui Long"  " Year: 2006"
"Author: Zongmin Ma"    " Year: 2007"

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q6.DELL.20240212.172215.925612
Running step 1 of 2...
Running step 2 of 2...
job output is in C:\Users\DELL\AppData\Local\Temp\q6.DELL.20240212.172215.925612\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q6.DELL.20240212.172215.925612\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q6.DELL.20240212.172215.925612...
```

7. Find the title of papers such that their venue is not mentioned in the input file.

In [13]:

```
%%file q7.py
from mrjob.job import MRJob
from mrjob.step import MRStep
import re
#Note: I am assuming we have to find both those entries that have #c but nothing follows that and those that dont even have #c.
class PapersWithoutVenue(MRJob):
    def configure_args(self):
        super(PapersWithoutVenue, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternVenue=re.compile(r"(\#c[^\#]*)")
        self.papersWithVenueMissing=[]
    def get_papers_without_venue(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                paperTitle=file[0].split('#@')[0].strip()
                venue=''.join(self.patternVenue.findall(file[0])).replace('#c','').replace('\n','').strip()
                if(len(venue)<=1):
                    self.papersWithVenueMissing.append(paperTitle)

    def final_get_papers_without_venue(self):
        for paper in self.papersWithVenueMissing:
            yield "-> Paper : ",paper

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

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

```
Overwriting q7.py
```

Q7 GC FILE OUTPUT

In [14]:

```
!python q7.py citation.txt
```

```
"-> Paper : "   "Automated Deduction in Geometry"
"-> Paper : "   "A+ Certification Core Hardware (Text & Lab Manual)"
"-> Paper : "   "Performance engineering in industry: current practices and adoption challenges"
"-> Paper : "   "Dude, You Can Do It! How to Build a Sweeet PC"
"-> Paper : "   "What Every Programmer Needs to Know about Security (Advances in information Security)"
"-> Paper : "   "Interpreting Kullback-Leibler divergence with the Neyman-earson lemma"
"-> Paper : "   "Digital Media: Transformations in Human Communication"
"-> Paper : "   "TOPP---the OpenMS proteomics pipeline"
"-> Paper : "   "Type Graphics and MacIntosh"
"-> Paper : "   "Adaptive Hypermedia and Adaptive Web-Based Systems"
"-> Paper : "   "Dependable Computing"
"-> Paper : "   "Calculus Early Transcendentals Single Variable"
"-> Paper : "   "Webbots, Spiders, and Screen Scrapers"
"-> Paper : "   "Making the Digital City: The Early Shaping of Urban Internet Space (Design & the Built Envir
onment S.)"
"-> Paper : "   "Linspire 5.0: The No Nonsense Guide! (No Nonsense Guide! series)"
"-> Paper : "   "Podcasting for Profit: A Proven 10-Step Plan for Generating Income Through Audio and Video P
odcasting"
"-> Paper : "   "Federated Identity Management And Web Services Security With IBM Tivoli Security Solutions"
"-> Paper : "   "Start with a Digital Camera (Special Edition) (2nd Edition) (Start with a)"
"-> Paper : "   "Open Process Frameworks: Patterns for the Adaptive e-Enterprise (Practitioners)"
"-> Paper : "   "Fast and Efficient Context-Aware Services (Wiley Series on Communications Networking & Distr
ibuted Systems)"
"-> Paper : "   "Multimedia Directory 1997"
"-> Paper : "   "ASIS&T Thesaurus of Information Science, Technology, And Librarianship (Asist Monograph Seri
es)"
"-> Paper : "   "On product covering in 3-tier supply chain models: natural complete problems for W[3] and W
[4]"
"-> Paper : "   "Inside SQL Server 2005 Tools (Microsoft Windows Server System Series)"
"-> Paper : "   "Electronic Engineer's Handbook (Core Handbook CD-ROMs)"
"-> Paper : "   "Call of Duty 2: Big Red One(tm) Official Strategy Guide (Official Strategy Guides)"
"-> Paper : "   "Inside Microsoft Dynamics AX 4.0"
"-> Paper : "   "Wiley Plus\/Web CT Stand-alone to accompany Java Concepts (Wiley Plus Products)"
"-> Paper : "   "Beginning Ruby on Rails (Wrox Beginning Guides)"
"-> Paper : "   "Introduction to Information Systems"
"-> Paper : "   "SUSE Linux Enterprise Server Administration (Course 3037)"
"-> Paper : "   "Hyperstat: Macintosh Hypermedia for Analyzing Data and Learning Statistics"
"-> Paper : "   "Computer Accounting with QuickBooks 2006"
"-> Paper : "   "Program Evaluation: Improving The Flow Of Information To The Congress"
"-> Paper : "   "Jocelyn Robert: Aucune de mes mains ne fait mal"
"-> Paper : "   "Special Edition Using Adobe Creative Suite 2 (Special Edition Using)"
"-> Paper : "   "At Ease With Performa"
"-> Paper : "   "Guia Visual de Microsoft Office 2000\/ Microsoft Office 2000 Visual Guide (Guias Visuales)"
"-> Paper : "   "Microsoft Expression Web: Visual QuickStart Guide"
"-> Paper : "   "TCP\/IP Protocol Suite, 3 edition"
"-> Paper : "   "Conker's Bad Fur Day (Prima's Official Strategy Guide)"
"-> Paper : "   "Exploring Macintosh Concepts in Visually Oriented Computing & Computing Concepts for End Use
rs"
"-> Paper : "   "KeyChamp 2.0 Macintosh Site License Package"
"-> Paper : "   "Word Processing on Your MacIntosh"
"-> Paper : "   "F.E.A.R.: First Encounter Assault Recon (Prima Official Game Guide)"
"-> Paper : "   "ExamInsight For MCP \/ MCSE Certification: Installing, Configuring, and Administering Micros
oft Windows XP Professional Exam 70-270 (ExamInsight)"
"-> Paper : "   "Computer Models of Musical Creativity"
"-> Paper : "   "Computer Concepts Illustrated Complete, Sixth Edition (Illustrated (Thompson Learning))"
"-> Paper : "   "Tips and Tuning Guide for MS Flight Simulator 2000"
"-> Paper : "   "Windows Vista All-in-One Desk Reference For Dummies (For Dummies (Computer\/Tech))"
"-> Paper : "   "DV Filmmaking: From Start to Finish (O'Reilly Digital Studio)"
"-> Paper : "   "Ibook Fan Book: Smart and Beautiful to Boot (Ibook Fan Books)"
"-> Paper : "   "GO Series: Microsoft Excel 2003 Volume 2 (Go With Microsoft Office)"
"-> Paper : "   "Data Structures"
"-> Paper : "   "Hands-On Guide to Video Blogging and Podcasting: Emerging Media Tools for Business Communica
tion (Hands-on Guide)"
"-> Paper : "   "Mage Knight(tm): Apocalypse Official Strategy Guide (Official Strategy Guides)"
"-> Paper : "   "ECDL Advanced"
"-> Paper : "   "Database Modeling in Biology: Practices and Challenges"
"-> Paper : "   "Microsoft Word 2003 Advanced"
"-> Paper : "   "Adaptive Multimedia Retrieval: First International Workshop, AMR 2003, Hamburg, Germany, Sep
tember 15-16, 2003, Revised Selected and Invited Papers (Lecture Notes in Computer Science)"
"-> Paper : "   "Gungrave: 2006 Wall Calendar"
"-> Paper : "   "Gurus, Hired Guns, and Warm Bodies: Itinerant Experts in a Knowledge Economy"
"-> Paper : "   "Selected Areas in Cryptography: 12th International Workshop, SAC 2005, Kingston, ON, Canada,
August 11-12, 2005, Revised Selected Papers (Lecture Notes in Computer Science)"
"-> Paper : "   "The Prentice Hall Planner for Student Success"
"-> Paper : "   "Keno Winner: A Guide To Winning At Video Keno"
"-> Paper : "   "Real World Aperture (Real World)"
"-> Paper : "   "Making Projects Critical (Management, Work and Organisations)"
"-> Paper : "   "The Internet: A Critical Introduction"
"-> Paper : "   "It's a Mad, Mad, Mad, Mad Mac\/Book and Disk"
"-> Paper : "   "Java for Everyone"
"-> Paper : "   "Essentials for Design Adobe Illustrator CS 2 - Level 1 (2nd Edition) (Essentials for Desig
n)"
"-> Paper : "   "Operations Research: Macintosh Version (Business Statistics Series)"
"-> Paper : "   "Creative Code: \u00c4sthetik und Programmierung am MIT Media Lab"
"-> Paper : "   "The Game Producer's Handbook"
"-> Paper : "   "Visual Basic 2005 Demystified, 1 edition"
"-> Paper : "   "Oracle 10g PL\/SQL: Guia de Consulta R\u00e1pida"
"-> Paper : "   "CompTIA A+ Exam Cram (Exams 220-602, 220-603, 220-604) (Exam Cram)"
"-> Paper : "   "Microsoft Powerpoint 2003 (Marquee Series)"
"-> Paper : "   "MICAI 2006: Advances in Artificial Intelligence: 5th Mexican International Conference on Artificial IntelligenceApizaco, Mexico, November 13-17, 2006Proceedings (Lecture Notes in Computer Science)"
"-> Paper : "   "Tarski's World 3.0: Including the Macintosh TM Program (Center for the Study of Language and Information - Lecture Notes)"
"-> Paper : "   "Internet and HTML Training on CD-ROM"
"-> Paper : "   "Elsevier's Dictionary of Wild and Cultivated Plants"

No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q7.DELL.20240212.172218.190333
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q7.DELL.20240212.172218.190333\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q7.DELL.20240212.172218.190333\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q7.DELL.20240212.172218.190333...
```

(BONUS!) 8. Find the title of papers such that their venue IS mentioned in the input file.

In [15]:

```
%%file q8.py
from mrjob.job import MRJob
from mrjob.step import MRStep
import re

class PapersWithVenue(MRJob):
    def configure_args(self):
        super(PapersWithVenue, self).configure_args()
        self.add_file_arg('--filename', help='Path to the input file') #TO make sure file opens only once
    def init_read_file(self):
        self.patternVenue=re.compile(r"(\#c[^\#]*)")
        self.papersWithVenue={}
    def get_papers_with_venue(self,_,line):
        file=[file for file in line.split('#*') if len(file)>0]
        if (file):
            if("#@" in file[0]):
                paperTitle=file[0].split('#@')[0].strip()
                venue=''.join(self.patternVenue.findall(file[0])).replace('#c','').replace('\n','').strip()
                if(len(venue)>1):
                    self.papersWithVenue[paperTitle]=venue

    def final_get_papers_with_venue(self):
        for paper,venue in self.papersWithVenue.items():
            yield "-> Paper: "+paper,"-> Venue: "+venue

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

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

```
Overwriting q8.py
```

Q8 GC FILE OUTPUT

In [16]:

```
!python q8.py citation.txt
```

```
No configs found; falling back on auto-configuration
No configs specified for inline runner
Creating temp directory C:\Users\DELL\AppData\Local\Temp\q8.DELL.20240212.172220.119516
Running step 1 of 1...
job output is in C:\Users\DELL\AppData\Local\Temp\q8.DELL.20240212.172220.119516\output
Streaming final output from C:\Users\DELL\AppData\Local\Temp\q8.DELL.20240212.172220.119516\output...
Removing temp directory C:\Users\DELL\AppData\Local\Temp\q8.DELL.20240212.172220.119516...
```

```
"-> Paper: Fast k-NN Classification Rule Using Metrics on Space-Filling Curves" "-> Venue: Proceedings of the
13th International Conference on Pattern Recognition - Volume 2"
"-> Paper: Approximating fluid schedules in crossbar packet-switches and Banyan networks"       "-> Venue: IE
EE\/ACM Transactions on Networking (TON)"
"-> Paper: Modeling methodology b: distributed simulation and the high level architecture"      "-> Venue: Pr
oceedings of the 38th conference on Winter simulation"
"-> Paper: An Integrative Modelling Approach for Simulation and Analysis of Adaptive Agents"    "-> Venue: Pr
oceedings of the 39th annual Symposium on Simulation"
"-> Paper: Notes from industry" "-> Venue: ACM SIGMICRONewsletter"
"-> Paper: A New Quadtree Decomposition Reconstruction Method"  "-> Venue: Proceedings of the 13th Internatio
nal Conference on Pattern Recognition - Volume 2"
"-> Paper: Microprogramming for the hardware engineer"  "-> Venue: ACM SIGMICRO Newsletter"
"-> Paper: A control word model for detecting conflicts between microoperations"        "-> Venue: ACM SIGMIC
RO Newsletter"
"-> Paper: Class-specific feature polynomial classifier for pattern classification and its application to han
dwritten numeral recognition"   "-> Venue: Pattern Recognition"
"-> Paper: Effects of reward expectancy on sequential eye movements in monkeys" "-> Venue: Neural Networks"
"-> Paper: A method to compute distance between two categorical values of same attribute in unsupervised lear
ning for categorical data set"  "-> Venue: Pattern Recognition Letters"
"-> Paper: Review of \"Bit-Slice Microprocessor Design by John Mick and James Brick\", McGraw-Hill Book Compa
ny, 1980"       "-> Venue: ACM SIGMICRO Newsletter"
"-> Paper: The Effect of Faults on Network Expansion"   "-> Venue: Theory of Computing Systems"
"-> Paper: Mesoscopic simulation of Ostwald ripening"   "-> Venue: Journal of Computational Physics"
"-> Paper: Design team composition for high level language computer architectures"      "-> Venue: ACM SIGARC
H Computer Architecture News"
"-> Paper: Three-Level Caching for Efficient Query Processing in Large Web Search Engines"      "-> Venue: Wo
rld Wide Web"
"-> Paper: On an initial transient deletion rule with rigorous theoretical support"     "-> Venue: Proceeding
s of the 38th conference on Winter simulation"
"-> Paper: Special issue: Dialog systems for health communications"     "-> Venue: Journal of Biomedical Info
rmatics"
```

-> Rough Work (The csv file generated was used to check accuracy of outputs above)

FOR GC FILE

In [17]:

```
import re
import pandas as pd

patterns={
'Paper Title':'#@',
'Authors':re.compile(r"\#\@[^\#\t]*"),
'Year':re.compile(r"(#t[^\#]*)"),
'Publication Venue':re.compile(r"(\#c[^\#]*)"),
'Index ID':re.compile(r"(index.)[^\n]*")
}
citations={
'Paper Title':[],
'Authors':[],
'Year':[],
'Publication Venue':[],
'Index ID':[]
}
fileOpened=open('citation.txt','r')
file=[file for file in fileOpened.read().split('#*') if len(file)>0]
for word in file:
    citations['Paper Title'].append(''.join(word.split(patterns['Paper Title'])[0]).strip())
    citations['Authors'].append(''.join(patterns['Authors'].findall(word)).replace('#@','').replace('\n','').strip())
    citations['Year'].append(int(''.join(patterns['Year'].findall(word)).replace('#t','').replace('\n','')))
    citations['Publication Venue'].append(''.join(patterns['Publication Venue'].findall(word)).replace('#c','').replace('\n','').strip())
    citations['Index ID'].append(int(''.join(patterns['Index ID'].findall(word)).replace('index','')))


citations=pd.DataFrame(citations)
citations.to_csv('Citations.csv')
citations.sample(5)
```

Out[17]:

	Paper Title	Authors	Year	Publication Venue	Index ID
79	Keno Winner: A Guide To Winning At Video Keno	Tom Collins	2004		8
82	The Internet: A Critical Introduction	Korinna Patelis	2007		8
4	What Every Programmer Needs to Know about Secu...	Neil Daswani,Anita Kesavan	2006		4
23	ASIS&T Thesaurus of Information Science, Techn...	Alice Redmond-neal,Marjorie M. K. Hlava	2005		2
32	Introduction to Information Systems	R. Kelly Rainer	2007		3