In [1]:
```
%%javascript
IPython.OutputArea.prototype._should_scroll = function(lines) {
    return false;
}
```

Raul G. Martinez

PID: A12461871

UCSD MAS DSE Cohort 6

DSE 203 - Fall 2020

Instructions for Assignment 2

Due Nov 1 by 11:59pm

Points 30

Submitting a file upload

Available Oct 23 at 12am - Nov 1 at 11:59pm 10 days

Data: citeseer.csv and dblp.csv (in Canvas/Files)

Your task:

Entity Resolution of the two tables using the two different methods shown in class (dedupe and Anhai's py_entitymatching)

Your submission should include the python scripts (includes SQL scripts) + results

Suggested Reference:

Jupyter Notebook I've uploaded: Entity Matching Complete Workflow.ipynb (Canvas/Files/Hands on/EM)

Anhai's Entity Matching Guide: https://sites.google.com/site/anhaidgroup/projects/magellan/py_entitymatching (https://sites.google.com/site/anhaidgroup/projects/magellan/py_entitymatching) (Links to an external site.)

dedupe Slides: Canvas/Files/Hands-on Session/dedupe

dedupe Examples (A how-to-use guide through example programs): https://github.com/dedupeio/dedupe-examples (https://github.com/dedupeio/dedupe-examples)

Notes:

(1) You have to install all the dependencies first before you can successfully install Anhai's py_entitymatching package

(2) If you try running the dedupe examples from this repo: https://github.com/dedupeio/dedupe-examples (https://github.com/dedupeio/dedupe-examples), (Links to an external site.) you might run into various issues. Read the dedupe Slides (in Canvas) to troubleshoot these.

Solution

1 Entity Resolution with Anahi's py_entitymatching

In [2]:
```
import py_entitymatching as em
import py_stringmatching as sm
import pandas as pd
import os, sys
```

In [3]:
```
# Display the versions
print('python version: ' + sys.version )
print('pandas version: ' + pd.__version__ )
print('magellan version: ' + em.__version__ )
```

```
python version: 3.7.9 (default, Aug 31 2020, 17:10:11) [MSC v.1916 64 bit (AMD64)]
pandas version: 1.1.3
magellan version: 0.3.2
```

1.1 Read input tables

In [4]:
```
# Read dataset
A = em.read_csv_metadata('../../Data for Assignments/citeseer.csv',low_memory=False) # setting the parameter low_memory to False  to speed up loading
B = em.read_csv_metadata('../../Data for Assignments/dblp.csv', low_memory=False)
```

```
Metadata file is not present in the given path; proceeding to read the csv file.
Metadata file is not present in the given path; proceeding to read the csv file.
```

In [5]:
```
len(A), len(B)
```

Out[5]:
```
(1823978, 2512927)
```

In [6]:
```
# use subset of original data
A = A[:100000]
B = B[:100000]
```

In [7]:
```
# find number of null values
A.isnull().sum()
```

Out[7]:
```
id                      0
title                   6
authors                 0
journal             99992
month              100000
year                80697
publication_type    99991
dtype: int64
```

In [8]:
```
# find number of null values
B.isnull().sum()
```

Out[8]:
```
id                      0
title                   0
authors                 0
journal            100000
month              100000
year                    0
publication_type        0
dtype: int64
```

In [9]:
```
A.head()
```

Out[9]:

	id	title	authors	journal	month	year	publication_type
0	1	An Arithmetic Analogue of Bezouts Theorem	David Mckinnon	NaN	NaN	NaN	NaN
1	2	Thompsons Group F is Not Minimally Almost Convex	James Belk, Kai-uwe Bux	NaN	NaN	2002.0	NaN
2	3	Cognitive Dimensions Tradeoffs in Tangible User Interface Design	Darren Edge, Alan Blackwell	NaN	NaN	NaN	NaN
3	4	ACTIVITY NOUNS, UNACCUSATIVITY, AND ARGUMENT MARKING IN YUKATEKAN SSILA meeting; Special Session...	J. Bohnemeyer, Max Planck, I. Introduction	NaN	NaN	2002.0	NaN
4	5	PS1-6 A6 ULTRASOUND-GUIDED HIFU NEUROLYSIS OF PERIPHERAL NERVES TO TREAT SPASTICITY AND	J. L. Foley, J. W. Little, F. L. Starr Iii, C. Frantz	NaN	NaN	NaN	NaN

In [10]:
```
B.head()
```

Out[10]:

	id	title	authors	journal	month	year	publication_type
0	1	Klaus Tschira Stiftung gemeinntzige GmbH, KTS	Klaus Tschira	NaN	NaN	2012	www
1	2	The SGML/XML Web Page	Robin Cover	NaN	NaN	2006	www
2	3	The Future of Classic Data Administration: Objects + Databases + CASE	Arnon Rosenthal	NaN	NaN	1998	www
3	4	XML Query Data Model	Mary F. Fernandez, Jonathan Robie	NaN	NaN	2001	www
4	5	The XML Query Algebra	Peter Fankhauser, Mary F. Fernndez, Ashok Malhotra, Michael Rys, Jrme Simon, Philip Wadler	NaN	NaN	2001	www

In [11]:
```
# drop all nan columns
A = A.drop(columns=['journal','month','year','publication_type'])
B = B.drop(columns=['journal','month','year','publication_type'])
```

In [12]:
```
# Set 'id' as the keys to the input tables
em.set_key(A, 'id')
em.set_key(B, 'id')
```

Out[12]:
```
True
```

In [13]:
```
A.head()
```

Out[13]:

	id	title	authors
0	1	An Arithmetic Analogue of Bezouts Theorem	David Mckinnon
1	2	Thompsons Group F is Not Minimally Almost Convex	James Belk, Kai-uwe Bux
2	3	Cognitive Dimensions Tradeoffs in Tangible User Interface Design	Darren Edge, Alan Blackwell
3	4	ACTIVITY NOUNS, UNACCUSATIVITY, AND ARGUMENT MARKING IN YUKATEKAN SSILA meeting; Special Session...	J. Bohnemeyer, Max Planck, I. Introduction
4	5	PS1-6 A6 ULTRASOUND-GUIDED HIFU NEUROLYSIS OF PERIPHERAL NERVES TO TREAT SPASTICITY AND	J. L. Foley, J. W. Little, F. L. Starr Iii, C. Frantz

In [14]:
```
B.head()
```

Out[14]:

	id	title	authors
0	1	Klaus Tschira Stiftung gemeinntzige GmbH, KTS	Klaus Tschira
1	2	The SGML/XML Web Page	Robin Cover
2	3	The Future of Classic Data Administration: Objects + Databases + CASE	Arnon Rosenthal
3	4	XML Query Data Model	Mary F. Fernandez, Jonathan Robie
4	5	The XML Query Algebra	Peter Fankhauser, Mary F. Fernndez, Ashok Malhotra, Michael Rys, Jrme Simon, Philip Wadler

1.2 Downsampling

In [15]:
```
# Downsample the datasets
sample_A, sample_B = em.down_sample(A, B, size=3000, y_param=1, show_progress=True, n_jobs=-1)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\sampler\down_sample.py:354: FutureWarning: The pandas.np module is depre
cated and will be removed from pandas in a future version. Import numpy directly instead
  sample_table_splits = pd.np.array_split(sample_table_b, n_jobs)
```

In [16]:
```
# Display the number of tuples in the sampled datasets
len(sample_A), len(sample_B)
```

Out[16]:
```
(2580, 3000)
```

In [17]:
```
# Show the metadata of sample_A, sample_B
em.show_properties(sample_A)
```

```
id: 2134681404552
key: id
```

In [18]:
```
em.show_properties(sample_B)
```

```
id: 2134681404168
key: id
```

In [19]:
```
sample_A.isnull().sum()
```

Out[19]:
```
id         0
title      0
authors    0
dtype: int64
```

In [20]:
```
sample_B.isnull().sum()
```

Out[20]:
```
id         0
title      0
authors    0
dtype: int64
```

In [21]:
```
sample_A.head()
```

Out[21]:

	id	title	authors
2	3	Cognitive Dimensions Tradeoffs in Tangible User Interface Design	Darren Edge, Alan Blackwell
3	4	ACTIVITY NOUNS, UNACCUSATIVITY, AND ARGUMENT MARKING IN YUKATEKAN SSILA meeting; Special Session...	J. Bohnemeyer, Max Planck, I. Introduction
6	7	A Methodology for the Enhancement of a Hypertext Version of a Textbook by the Automatic Insertio...	F. Crestani, M. Melucci
24584	24585	[4] A. Berman and J. Plemmons, Nonnegative Matrices in the Mathematical Sciences, Academic Press...	M. Aigner, G. M. Ziegler, Proofs From The Book, Nd Edition, Springer Verlag, H. Alt, C. Knauer, ...
8204	8205	Pacific Symposium on Biocomputing 14:75-86 (2009) CONTEXT-SPECIFIC GENE REGULATIONS IN CANCER GE...	Ina Sen, Michael P. Verdicchio, Sungwon Jung, Robert Trevino, Michael Bittner, Seungchan Kim

In [22]:
```
sample_B.head()
```

Out[22]:

	id	title	authors
1951	1952	Building Acceptable Classification Models.	David Martens, Bart Baesens
3284	3285	Narrative Interactive Multimedia Learning Environments: Achievements and Challenges.	Paul Brna
83153	83154	How to Design Good Educational Blogs in LMS?.	Ahmed Mohamed Fahmy Yousef, Guido Rling
30784	30785	Performance Analysis and Improvement Using LFSR in the Pipelined Key Scheduling Section of DES.	P. V. Sruthi, Prabaharan Poornachandran, A. S. Remya Ajai
29678	29679	An 800MS/s dual-residue pipeline ADC in 40nm CMOS.	Jan Mulder, Frank M. L. van der Goes, Davide Vecchi, Jan R. Westra, Emre Ayranci, Christopher M....

1.3 Generating features for manually

In [23]:
```
## Getting Attribute Types
atypes1 = em.get_attr_types(sample_A)
atypes2 = em.get_attr_types(sample_B)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\feature\attributeutils.py:191: FutureWarning: The pandas.np module is de
precated and will be removed from pandas in a future version. Import numpy directly instead
  if returned_type == bool or returned_type == pd.np.bool_:
```

In [24]:
```
# display attributes for each
atypes1['title'], atypes2['title'], atypes1['authors'], atypes2['authors']
```

Out[24]:
```
('str_gt_10w', 'str_bt_5w_10w', 'str_gt_10w', 'str_bt_5w_10w')
```

In [25]:
```
# make sure attributes match between columns 'title' and 'author'
atypes2['title'] = 'str_gt_10w'
atypes1['authors'] = 'str_bt_5w_10w'
```

In [26]:
```
# confirm they match
atypes1['title'], atypes2['title'], atypes1['authors'], atypes2['authors']
```

Out[26]:
```
('str_gt_10w', 'str_gt_10w', 'str_bt_5w_10w', 'str_bt_5w_10w')
```

In [27]:
```
## Getting Attribute Correspondences
block_c = em.get_attr_corres(sample_A, sample_B)
id(A), id(block_c['ltable']), id(B), id(block_c['rtable'])
```

Out[27]:
```
(2134599011016, 2134681404552, 2135555200840, 2134681404168)
```

In [28]:
```
block_c['corres']
```

Out[28]:
```
[('id', 'id'), ('title', 'title'), ('authors', 'authors')]
```

In [29]:
```
block_c.keys()
```

Out[29]:
```
dict_keys(['corres', 'ltable', 'rtable'])
```

In [30]:
```
# Getting tokenizers for blocking
tok = em.get_tokenizers_for_blocking()
tok
```

Out[30]:
```
{'qgm_2': <function py_entitymatching.feature.tokenizers._make_tok_qgram.<locals>.tok_qgram>,
 'qgm_3': <function py_entitymatching.feature.tokenizers._make_tok_qgram.<locals>.tok_qgram>,
 'wspace': <function py_entitymatching.feature.tokenizers.tok_wspace>,
 'alphabetic': <function py_entitymatching.feature.tokenizers.tok_alphabetic>,
 'alphanumeric': <function py_entitymatching.feature.tokenizers.tok_alphanumeric>,
 'dlm_dc0': <function py_entitymatching.feature.tokenizers._make_tok_delim.<locals>.tok_delim>}
```

In [31]:
```
# Getting Similarity Functions for blocking
sim = em.get_sim_funs_for_blocking()
sim
```

Out[31]:
```
{'affine': <function py_entitymatching.feature.simfunctions.affine>,
 'hamming_dist': <function py_entitymatching.feature.simfunctions.hamming_dist>,
 'hamming_sim': <function py_entitymatching.feature.simfunctions.hamming_sim>,
 'lev_dist': <function py_entitymatching.feature.simfunctions.lev_dist>,
 'lev_sim': <function py_entitymatching.feature.simfunctions.lev_sim>,
 'jaro': <function py_entitymatching.feature.simfunctions.jaro>,
 'jaro_winkler': <function py_entitymatching.feature.simfunctions.jaro_winkler>,
 'needleman_wunsch': <function py_entitymatching.feature.simfunctions.needleman_wunsch>,
 'smith_waterman': <function py_entitymatching.feature.simfunctions.smith_waterman>,
 'overlap_coeff': <function py_entitymatching.feature.simfunctions.overlap_coeff>,
 'jaccard': <function py_entitymatching.feature.simfunctions.jaccard>,
 'dice': <function py_entitymatching.feature.simfunctions.dice>,
 'monge_elkan': <function py_entitymatching.feature.simfunctions.monge_elkan>,
 'cosine': <function py_entitymatching.feature.simfunctions.cosine>,
 'exact_match': <function py_entitymatching.feature.simfunctions.exact_match>,
 'rel_diff': <function py_entitymatching.feature.simfunctions.rel_diff>,
 'abs_norm': <function py_entitymatching.feature.simfunctions.abs_norm>}
```

In [32]:
```
# Getting Features
feature_table = em.get_features(sample_A, sample_B, atypes1, atypes2, block_c, tok, sim)
feature_table.head()
```

Out[32]:

	feature_name	left_attribute	right_attribute	left_attr_tokenizer	right_attr_tokenizer	simfunction	function	function_source	is_auto_gen
0	id_id_exm	id	id	None	None	exact_match	<function id_id_exm at 0x000001F10002B5E8>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
1	id_id_anm	id	id	None	None	abs_norm	<function id_id_anm at 0x000001F10002BCA8>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
2	id_id_lev_dist	id	id	None	None	lev_dist	<function id_id_lev_dist at 0x000001F10002B558>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
3	id_id_lev_sim	id	id	None	None	lev_sim	<function id_id_lev_sim at 0x000001F100462048>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
4	title_title_jac_qgm_3_qgm_3	title	title	qgm_3	qgm_3	jaccard	<function title_title_jac_qgm_3_qgm_3 at 0x000001F100462168>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	

1.4 Block tables to get candidate set

In [33]:
```
# use rule-based blocker on 'title' column
rb = em.RuleBasedBlocker()
rb.add_rule(['title_title_jac_qgm_3_qgm_3(ltuple, rtuple) < 0.7'], feature_table)

C1 = rb.block_tables(sample_A, sample_B, l_output_attrs=['title', 'authors'], r_output_attrs=['title', 'authors'], show_progress=False)
len(C1)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_stringsimjoin\utils\converter.py:99: FutureWarning: The pandas.np module is deprecated
and will be removed from pandas in a future version. Import numpy directly instead
  if col_type == pd.np.object:
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_stringsimjoin\utils\validation.py:30: FutureWarning: The pandas.np module is deprecated
and will be removed from pandas in a future version. Import numpy directly instead
  if attr_type != pd.np.object:
```

Out[33]: 25

In [34]:
```
C1.head()
```

Out[34]:

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors
0	0	86543	93390	Query Optimization by Predicate Move-Around	Alon Y. Levy, Inderpal Singh Mumick, Yehoshua Sagiv	Query Optimization by Predicate Move-Around.	Alon Y. Levy, Inderpal Singh Mumick, Yehoshua Sagiv
1	1	47767	99906	3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions	Romer Rosales, Stan Sclaroff	3D Trajectory Recovery for Tracking Multiple Objects and Trajectory Guided Recognition of Actions.	Rmer Rosales, Stan Sclaroff
2	2	34416	8127	Botnet Detection Based on Network Behavior	W. Timothy Strayer, David Lapsely, Robert Walsh, Carl Livadas	Botnet Detection Based on Network Behavior.	W. Timothy Strayer, David E. Lapsley, Robert Walsh, Carl Livadas
3	3	29913	91546	Bridging the Application and DBMS Profiling Divide for Database Application Developers	Surajit Chaudhuri	Bridging the Application and DBMS Profiling Divide for Database Application Developers.	Surajit Chaudhuri, Vivek R. Narasayya, Manoj Syamala
4	4	47241	90913	Focused Crawling Using Context Graphs	M. Diligenti, F. M. Coetzee, S. Lawrence, C. L. Giles, M. Gori	Focused Crawling Using Context Graphs.	Michelangelo Diligenti, Frans Coetzee, Steve Lawrence, C. Lee Giles, Marco Gori

In [35]:
```
def title_title_function(x,y):

    # extract authors column
    x_author = x['authors']
    y_author = y['authors']

    # combine all last names in a string separated by a space
    x_lastnames = ' '.join([i.split(' ')[-1] for i in x_author.split(',')])
    y_lastnames = ' '.join([i.split(' ')[-1] for i in y_author.split(',')])

    # convert all last names to lower case
    x_lastnames = str(x_lastnames).lower()
    y_lastnames = str(y_lastnames).lower()

    # exclude when similarity score is lower than threshold
    if jac.get_raw_score(ws_tok_set.tokenize(x_lastnames), ws_tok_set.tokenize(y_lastnames)) < 0.7:
        return True
    else:
        return False

# Apply black box blocker on 'authors' column
ws_tok_set = sm.WhitespaceTokenizer(return_set=True)
jac = sm.Jaccard()

bb = em.BlackBoxBlocker()
bb.set_black_box_function(title_title_function)

C2 = bb.block_tables(sample_A, sample_B, l_output_attrs=['title', 'authors'], r_output_attrs=['title', 'authors'], show_progress=False)
len(C2)
```

Out[35]: 157

In [36]:
```
C2.head()
```

Out[36]:

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors
0	0	8205	84457	Pacific Symposium on Biocomputing 14:75-86 (2009) CONTEXT-SPECIFIC GENE REGULATIONS IN CANCER GE...	Ina Sen, Michael P. Verdicchio, Sungwon Jung, Robert Trevino, Michael Bittner, Seungchan Kim	Context-Specific Gene Regulations in Cancer Gene Expression Data.	Ina Sen, Michael P. Verdicchio, Sungwon Jung, Robert Trevino, Michael L. Bittner
1	1	65621	41771	ADAPTIVE CALIBRATION AND CONTROL OF CASCADE PROCESSES WITH UNKNOWN MEASUREMENT MODEL AND ACTUATO...	Zeyu Liu, Perry Li	A novel multi-band spectral subtraction method based on phase modification and magnitude compens...	Chao Li, Wen-Ju Liu
2	2	57491	35426	A Formal Framework for Image Indexing with Triples: Toward a Concept-Based Image Retrieval	Jae Dong Yang, Hyung Jeong Yang	Optimal Overlay Construction on Heterogeneous Live Peer-to-Peer Streaming Systems.	Min Yang, Yuanyuan Yang
3	3	57491	93599	A Formal Framework for Image Indexing with Triples: Toward a Concept-Based Image Retrieval	Jae Dong Yang, Hyung Jeong Yang	Better IT Governance for Organizations - A Model for Improving Flexibility and Capabilities of S...	Jungho Yang
4	4	202	18784	Field Modifiable Architecture with FPGAs and its Design/Verification/Debugging Methodologies	Masahiro Fujita, Satoshi Komatsu, Hiroshi Saito, Kenshu Seto, Thanyapat Sakunkonchak, Yoshihisa ...	Field Modifiable Architecture with FPGAs and its Design/Verification/Debugging Methodologies.	Masahiro Fujita, Satoshi Komatsu, Hiroshi Saito, Kenshu Seto, Thanyapat Sakunkonchak, Yoshihisa ...

1.5 Combine all blocker outputs

In [39]:
```
C = em.combine_blocker_outputs_via_union([C1, C2])
len(C)
```

Out[39]:
```
165
```

In [40]:
```
C.head()
```

Out[40]:

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors
0	0	202	18784	Field Modifiable Architecture with FPGAs and its Design/Verification/Debugging Methodologies	Masahiro Fujita, Satoshi Komatsu, Hiroshi Saito, Kenshu Seto, Thanyapat Sakunkonchak, Yoshihisa ...	Field Modifiable Architecture with FPGAs and its Design/Verification/Debugging Methodologies.	Masahiro Fujita, Satoshi Komatsu, Hiroshi Saito, Kenshu Seto, Thanyapat Sakunkonchak, Yoshihisa ...
1	1	449	90667	LBSs and Location Privacy From Data Privacy to Location Privacy Unique Challenges of Location...	Ling Liu	An Agent and Goal-Oriented Approach for Virtual Enterprise Modelling: A Case Study.	Zhi Liu, Lin Liu
2	2	1196	62720	A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode	Wenhua Jiao Hongxi Wang	Towards a measurement tool for verification and validation of simulation models.	Zhongshi Wang
3	3	1196	67994	A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode	Wenhua Jiao Hongxi Wang	Sample selection based on multiple incremental decision trees in BSP programming library.	Shuo Wang, Jian-Jian Wang, Yi Wang, Xuezheng Wang
4	4	1196	85023	A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode	Wenhua Jiao Hongxi Wang	A review of electronic signatures regulations: do they facilitate or impede international electr...	Minyan Wang

1.6 Sampling and labeling the candidate set

In [41]:
```
# Sample candidate set, in other words, rough evaluation by using Precision@K metric (where k = 50)
S = em.sample_table(C, 50)
len(S)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\sampler\single_table.py:103: FutureWarning: The pandas.np module is deprecated and will be removed from pandas in a future version. Import numpy directly instead
  sample_indices = pd.np.random.choice(len(table), sample_size,
```

Out[41]: 50

In [42]:
```
S.head()
```

Out[42]:

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors
3	3	1196	67994	A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode	Wenhua Jiao Hongxi Wang	Sample selection based on multiple incremental decision trees in BSP programming library.	Shuo Wang, Jian-Jian Wang, Yi Wang, Xuezheng Wang
10	10	5629	35844	LOOP SCHEDULING FOR MINIMIZING SCHEDULE LENGTH AND SWITCHING ACTIVITIES	Zili Shao, Qingfeng Zhuge, Edwin H. -m. Sha	Timing Optimization of Nested Loops Considering Code Size for DSP Applications.	Qingfeng Zhuge, Zili Shao, Edwin Hsing-Mean Sha
12	12	6029	78323	National Knowledge Service Business Plan and on the National Library for Health Development Plan...	J A Muir Gray	Introduction to silicon compilation.	John P. Gray
15	15	6542	83517	An Architecture of Game Grid Based on Resource Router	Yu Wang, Enhua Tan, Wei Li, Zhiwei Xu	A Semantic Web Based Peer to Peer Service Discovery Mechanism for Intelligent Business Process.	Desheng Li, Ruzhi Xu, Haiyang Wang
20	20	9736	84183	Pacific Symposium on Biocomputing 8:490-501(2003) ERRORS AND LINKAGE DISEQUILIBRIUM INTERACT MUL...	D. Gordon, M. A. Levenstien, S. J. Finch, J. Ott, D. Gordon, M. A. Levenstien, S. J. Finch, J. Ott	Errors and Linkage Disequilibrium Interact Multiplicatively When Computing Sample Sizes for Gene...	Derek Gordon, Mark A. Levenstien, Stephen J. Finch, Jrg Ott

In [43]:
```
# Label S
G = em.label_table(S, 'gold')
len(G)
```

```
Column name (gold) is not present in dataframe
```

Out[43]:
```
50
```

In [44]:
```
G
```

Out[44]:

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors	gold
3	3	1196	67994	A Service Flow Management Strategy for IEEE 802.16 Broadband Wireless Access Systems in TDD Mode	Wenhua Jiao Hongxi Wang	Sample selection based on multiple incremental decision trees in BSP programming library.	Shuo Wang, Jian-Jian Wang, Yi Wang, Xuezheng Wang	0
10	10	5629	35844	LOOP SCHEDULING FOR MINIMIZING SCHEDULE LENGTH AND SWITCHING ACTIVITIES	Zili Shao, Qingfeng Zhuge, Edwin H. -m. Sha	Timing Optimization of Nested Loops Considering Code Size for DSP Applications.	Qingfeng Zhuge, Zili Shao, Edwin Hsing-Mean Sha	0
12	12	6029	78323	National Knowledge Service Business Plan and on the National Library for Health Development Plan...	J A Muir Gray	Introduction to silicon compilation.	John P. Gray	0
15	15	6542	83517	An Architecture of Game Grid Based on Resource Router	Yu Wang, Enhua Tan, Wei Li, Zhiwei Xu	A Semantic Web Based Peer to Peer Service Discovery Mechanism for Intelligent Business Process.	Desheng Li, Ruzhi Xu, Haiyang Wang	0
20	20	9736	84183	Pacific Symposium on Biocomputing 8:490-501(2003) ERRORS AND LINKAGE DISEQUILIBRIUM INTERACT MUL...	D. Gordon, M. A. Levenstien, S. J. Finch, J. Ott, D. Gordon, M. A. Levenstien, S. J. Finch, J. Ott	Errors and Linkage Disequilibrium Interact Multiplicatively When Computing Sample Sizes for Gene...	Derek Gordon, Mark A. Levenstien, Stephen J. Finch, Jrg Ott	0
23	23	12962	13507	SimTester: A Controllable and Observable Testing Framework for Embedded Systems	Tingting Yu, Witawas Srisa-an, Gregg Rothermel	Testing Inter-layer and Inter-task Interactions in RTES Applications.	Ahyoung Sung, Witawas Srisa-an, Gregg Rothermel, Tingting Yu	0
33	33	18047	91948	Offering a Precision-Performance Tradeoff for Aggregation Queries over Replicated Data	Chris Olston, Jennifer Widom	Offering a Precision-Performance Tradeoff for Aggregation Queries over Replicated Data.	Chris Olston, Jennifer Widom	1
34	34	18073	66485	Freeflow: Mediating Between Representation and Action in Workflow Systems	Paul Dourish, Jim Holmes, Allan Maclean, Pernille Marqvardsen, Alex Zbyslaw	Freeflow: Mediating Between Representation and Action in Workflow Systems.	Paul Dourish, Jim Holmes, Allan MacLean, Pernille Marqvardsen, Alex Zbyslaw	0
35	35	18136	25039	Rapid Knowledge Work Visualization for Abstract Organizations	Markus Strohmaier, Stefanie Lindstaedt	KnowFlow - A Hybrid Approach to Identifying and Visualizing Distributed Knowledge Work Practices.	Markus Strohmaier, Stefanie N. Lindstaedt	0
37	37	19395	35426	Optimal and heuristic algorithms for quality-of-service routing with multiple constraints	Wen-lin Yang	Optimal Overlay Construction on Heterogeneous Live Peer-to-Peer Streaming Systems.	Min Yang, Yuanyuan Yang	0
38	38	19395	93599	Optimal and heuristic algorithms for quality-of-service routing with multiple constraints	Wen-lin Yang	Better IT Governance for Organizations - A Model for Improving Flexibility and Capabilities of S...	Jungho Yang	0
40	40	20643	35426	Phone Server: Design, Implementation and Performance Evaluation	Bo Yang	Optimal Overlay Construction on Heterogeneous Live Peer-to-Peer Streaming Systems.	Min Yang, Yuanyuan Yang	0

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors	gold
43	43	20743	60902	Improving Spatial Reuse through Tuning Transmit Power, Carrier Sense Threshold, and Data Rate in...	Tae-suk Kim	Proposed Methodology for Comparing Schedule Generation Schemes in Construction Resource Scheduling.	Jin-Lee Kim	0
45	45	21154	51447	BAYESIAN MOTION BLUR IDENTIFICATION USING BLUR PRIORI *	Xuezheng Liu, Mingjing Li, Hongjiang Zhang, Dingxing Wang	Design of an Expandable Website Platform for Quality Course Cluster.	Shuoping Wang, Gaoyan Zhang, Jun Liu	0
46	46	22133	26444	LANGUAGE MODEL ADAPTATION FOR AUTOMATIC SPEECH RECOGNITION AND STATISTICAL MACHINE TRANSLATION	Woosung Kim	A Study on Developing CRM Scorecard.	Hyung-Su Kim, Young-Gul Kim	0
48	48	22148	13626	Granular Computing on Binary Relations II Rough Set Representations and Belief Functions	T. Y. Lin	Model Checking Value-Passing Processes.	Huimin Lin	0
51	51	23233	51636	NIOS: A Distributed Service Oriented Architecture for Business Process Execution ABSTRACT	Guoli Li	Concept Granular System and Granular Concept Lattice.	Hong Li	0
56	56	25994	4163	structures and fuzzy concept networks	Kyung-joong Kim, Sung-bae Cho	Evolutionary Algorithms for Board Game Players with Domain Knowledge.	Kyung-Joong Kim, Sung-Bae Cho	0
60	60	26587	19443	Special Issue Editorial Building Parallel and Distributed Systems	Paddy Nixon, Vinny Cahill, Fethi Rabhi	Software Engineering for Distrinuted Systems.	Paddy Nixon, Vinny Cahill, Fethi A. Rabhi	0
64	64	28412	18324	Commitments and Conventions: The Foundation of Coordination in Multi- Agent Systems	Nick R. Jennings	Symbolic incompletely specified functions for correct evaluation in the presence of indeterminat...	Glenn Jennings	0
69	69	30333	10331	Abstract H-Animator: A Visual Tool for Modeling, Reuse and Sharing of X3D Humanoid Animations	Fabio Buttussi, Luca Chittaro, Daniele Nadalutti	Filtering Fitness Trail Content Generated by Mobile Users.	Fabio Buttussi, Luca Chittaro, Daniele Nadalutti	0
77	77	34811	57009	Grid Service Monitor Performance Monitoring and Measurement of Grid Services Using Peer-to-Peer ...	Yin Chen	A fast simulation approach for tandem queueing systems.	Liang Chen, Chien-Liang A. Chen	0
81	81	36543	7194	Challenges in Visual Data Analysis	Daniel A. Keim, Florian Mansmann, Jrn Schneidewind, Hartmut Ziegler	Visual Analytics: Scope and Challenges.	Daniel A. Keim, Florian Mansmann, Jrn Schneidewind, Jim Thomas, Hartmut Ziegler	0
86	86	43111	17598	An Unfolding Algorithm for Synchronous Products of Transition Systems	Javier Esparza, Stefan Romer	An Unfolding Algorithm for Synchronous Products of Transition Systems.	Javier Esparza, Stefan Rmer	1
88	88	43692	3999	Abstract Domain Decomposition Methods for Linear-Quadratic Elliptic Optimal Control Problems	Hoang Q. Nguyen, Hoang Q. Nguyen	Cloud-Based Data Warehousing Application Framework for Modeling Global and Regional Data Managem...	Thanh Binh Nguyen	0
91	91	43831	16565	Resolving Implementation Convolution in Middleware Systems	Charles Zhang, Hans-arno Jacobsen	Efficiently mining crosscutting concerns through random walks.	Charles Zhang, Hans-Arno Jacobsen	0

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors	gold
93	93	44183	37337	MetaFluxNet, a Program Package for Metabolic Pathway Construction and Analysis, and Its Use in L...	Sang Yup Lee, Dong-yup Lee, Soon Ho Hong, Tae Yong Kim	A Real-Time Communication Method for Wormhole Switching Networks.	Byungjae Kim, Jong Kim, Sung Je Hong, Sunggu Lee	0
95	95	44416	30646	An Interaction Control Architecture for Large Chairperson-controlled Conferences Over the Internet	Lukas Ruf, Thomas Walter, Bernhard Plattner	An Interaction Control Architecture for Large Chairperson-Controlled Conferences over the Internet.	Lukas Ruf, Thomas Walter, Bernhard Plattner	1
98	98	45083	93216	Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community	Gladys A. Cotter, Barbara T. Bauldock	Biodiversity Informatics Infrastructure: An Information Commons for the Biodiversity Community.	Gladys A. Cotter, Barbara T. Bauldock	1
103	103	45965	27465	Using Constraint Programming and Local Search Methods to Solve Vehicle Routing Problems	Paul Shaw	The Role of Identification in the Privacy Decisions of Information System Students.	Thomas Shaw	0
104	104	46229	11715	Abstract Access to Italian legal literature: Integration between Structured Repositories and Web...	E. Francesconi, G. Peruginelli	Searching and retrieving legal literature through automated semantic indexing.	Enrico Francesconi, Ginevra Peruginelli	0
108	108	48058	10412	Generalized Knapsack Solvers for Multi-Unit Combinatorial Auctions: Analysis and Application to ...	Terence Kelly	A Longitudinal, Naturalistic Study of Information Search & Use Behavior as Implicit Feedback for...	Diane Kelly	0
111	111	50060	97312	Converting Legacy Relational Database into XML Database through Reverse Engineering	Chunyan Wang, Anthony Lo, Reda Alhajj, Ken Barker	Converting Legacy Relational Database into XML Database through Reverse Engineering.	Chunyan Wang, Anthony Chiu Wa Lo, Reda Alhajj, Ken Barker	1
124	124	65850	54467	Modularization and automatic composition of Object- Role Modeling (ORM) schemes	Mustafa Jarrar	Modularization and Automatic Composition of Object-Role Modeling (ORM) Schemes.	Mustafa Jarrar	1
127	127	66232	57009	The Analysis of Different Production Planning Decision Models in the Supply Chain Network	Yin-yann Chen	A fast simulation approach for tandem queueing systems.	Liang Chen, Chien-Liang A. Chen	0
132	132	70320	74019	On the Minimization of SOPs for Bi-Decomposable Functions	Tsutomu Sasao, Jon T. Butler	Hardware Index to Set Partition Converter.	Jon T. Butler, Tsutomu Sasao	0
137	137	76990	37547	PAPER Evaluation of Two Load-Balancing Primary-Backup Process Allocation Schemes	Heejo Lee, Jong Kim, Sung Je Hong	DTN: A New Partitionable Torus Topology.	SangHo Chae, Jong Kim, Dongseung Kim, Sung Je Hong, Sunggu Lee	0
138	138	78425	75967	Low Power Design of Memory Intensive Functions Case Study: Vector Quantization	David B. Lidsky, Jan M. Rabaey	Early Power Exploration - A World Wide Web Application.	David Lidsky, Jan M. Rabaey	0
139	139	80863	60953	Permuted Estimators for Regenerative Simulations	James M. Calvin, Marvin K. Nakayama	Output analysis: a comparison of output-analysis methods for simulations of processes with multi...	James M. Calvin, Marvin K. Nakayama	0
141	141	81155	91111	Adaptive Commitment for Distributed Real-Time Transactions	Nandit Soparkar, Eliezer Levy, Henry F. Korth, Avi Silberschatz	Triggered Real-Time Databases with Consistency Constraints.	Henry F. Korth, Nandit Soparkar, Abraham Silberschatz	0

	_id	ltable_id	rtable_id	ltable_title	ltable_authors	rtable_title	rtable_authors	gold
144	144	82965	37455	PROGRAM ANALYSIS FOR CACHE COHERENCE: BEYOND PROCEDURAL BOUNDARIES	Lynn Choi, Pen-chung Yew	Program Analysis for Cache Coherence: Beyond Procedural Boundaries.	Lynn Choi, Pen-Chung Yew	1
145	145	83921	89103	Theory and Practice of I/O-Efficient Algorithms for Multidimensional Batched Searching Problems ...	Lars Arge, Octavian Procopiuc, Sridhar Ramaswamy, Torsten Suel, Jeffrey Scott Vitter	A Unified Approach for Indexed and Non-Indexed Spatial Joins.	Lars Arge, Octavian Procopiuc, Sridhar Ramaswamy, Torsten Suel, Jan Vahrenhold, Jeffrey Scott Vi...	0
154	154	96995	51636	A computation offloadingscheme on handheld devices	Zhiyuan Li	Concept Granular System and Granular Concept Lattice.	Hong Li	0
155	155	96995	85299	A computation offloadingscheme on handheld devices	Zhiyuan Li	Dynamic nature of trust in e-commerce.	Nan Li	0
157	157	97468	73784	A Prototype Content-based Image Retrieval System for Spine X-rays	L. Rodney Long, Sameer K. Antani, George R. Thoma	Unsupervised Grow-Cut: Cellular Automata-Based Medical Image Segmentation.	Payel Ghosh, Sameer Antani, L. Rodney Long, George R. Thoma	0
158	158	97733	92305	A Native Extension of SQL for Mining Data Streams	Chang Luo, Hetal Thakkar, Haixun Wang, Carlo Zaniolo	ATLAS: A Small but Complete SQL Extension for Data Mining and Data Streams.	Haixun Wang, Carlo Zaniolo, Chang Luo	0
160	160	98084	78514	Abstract Practical Iterated Fill Synthesis for CMP Uniformity	Yu Chen, Andrew B. Kahng, Gabriel Robins, Er Zelikovsky	Practical iterated fill synthesis for CMP uniformity.	Yu Chen, Andrew B. Kahng, Gabriel Robins, Alexander Zelikovsky	0
161	161	98124	32911	Evolving Modular Recursive Sorting Algorithms	Ros Agapitos, Simon M. Lucas	Evolving Modular Recursive Sorting Algorithms.	Alexandros Agapitos, Simon M. Lucas	1
162	162	99327	6271	Link Bandwidth Detection for Multimedia Streaming in a Distributed Server Environment	Liu Yin	The Self-Organizing Maps: Background, Theories, Extensions and Applications.	Hujun Yin	0
163	163	99496	31095	The Influence of Social Dependencies on Decision-Making: Initial Investigations with a New Game	Barbara J. Grosz	TEAM: A Transportable Natural-Language Interface System.	Barbara J. Grosz	0

1.7 Splitting the labeled data into development and evaluation set

In [45]:
```
# Split S into development set (I) and evaluation set (J)
IJ = em.split_train_test(G, train_proportion=0.8, random_state=0)
I = IJ['train']
J = IJ['test']
len(I), len(J)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\matcher\matcherutils.py:98: FutureWarning: The pandas.np module is depre
cated and will be removed from pandas in a future version. Import numpy directly instead
  idx_values = pd.np.array(labeled_data.index.values)
```

Out[45]:
```
(40, 10)
```

1.8 Creating features

In [46]:
```
# use entire dataset to create features
feature_table = em.get_features_for_matching(A, B, validate_inferred_attr_types=False)
feature_table.head()
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\feature\attributeutils.py:191: FutureWarning: The pandas.np module is de
precated and will be removed from pandas in a future version. Import numpy directly instead
  if returned_type == bool or returned_type == pd.np.bool_:
```

Out[46]:

	feature_name	left_attribute	right_attribute	left_attr_tokenizer	right_attr_tokenizer	simfunction	function	function_source	is_auto_gen
0	id_id_exm	id	id	None	None	exact_match	<function id_id_exm at 0x000001F1044F70D8>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
1	id_id_anm	id	id	None	None	abs_norm	<function id_id_anm at 0x000001F1044F7048>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
2	id_id_lev_dist	id	id	None	None	lev_dist	<function id_id_lev_dist at 0x000001F1044F7F78>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
3	id_id_lev_sim	id	id	None	None	lev_sim	<function id_id_lev_sim at 0x000001F1044F7E58>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	
4	title_title_jac_qgm_3_qgm_3	title	title	qgm_3	qgm_3	jaccard	<function title_title_jac_qgm_3_qgm_3 at 0x000001F1142E4708>	from py_entitymatching.feature.simfunctions import *\nfrom py_entitymatching.feature.tokenizers ...	

1.9 Converting the development set to feature vectors

In [47]:
```
# Convert the I into a set of feature vectors using F
H = em.extract_feature_vecs(I,
                            feature_table=feature_table,
                            attrs_after=['gold'],
                            show_progress=False)
len(H)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\feature\extractfeatures.py:157: FutureWarning: The pandas.np module is d
eprecated and will be removed from pandas in a future version. Import numpy directly instead
  c_splits = pd.np.array_split(candset, n_procs)
```

Out[47]: 40

In [48]:
```
H.head()
```

Out[48]:

	_id	ltable_id	rtable_id	id_id_exm	id_id_anm	id_id_lev_dist	id_id_lev_sim	title_title_jac_qgm_3_qgm_3	title_title_cos_dlm_dc0_dlm_dc0	title_title_mel	title_title_lev_dist	title_title
124	124	65850	54467	0	0.827137	5	0.0	0.688889	0.527046	0.910414	5	
132	132	70320	74019	0	0.950026	4	0.2	0.030612	0.000000	0.497494	46	
93	93	44183	37337	0	0.845054	5	0.0	0.066298	0.081111	0.582108	122	
127	127	66232	57009	0	0.860747	5	0.0	0.035971	0.000000	0.592197	69	
60	60	26587	19443	0	0.731297	5	0.0	0.206522	0.000000	0.664487	36	

1.10 Converting the evaluation set to feature vectors

In [49]:
```
L = em.extract_feature_vecs(J, feature_table=feature_table,
                            attrs_after='gold', show_progress=False)
```

```
C:\Users\rmartinez4\Anaconda3\envs\py37\lib\site-packages\py_entitymatching\feature\extractfeatures.py:157: FutureWarning: The pandas.np module is d
eprecated and will be removed from pandas in a future version. Import numpy directly instead
  c_splits = pd.np.array_split(candset, n_procs)
```

1.11 Train, predict, and evaluate different matchers

In [50]:
```
# Create a set of ML-matchers, test only these 3 but ideally we would evaluate all available methods
dt = em.DTMatcher(name='DecisionTree', random_state=0)
svm = em.SVMMatcher(name='SVM', random_state=0)
rf = em.RFMatcher(name='RF', random_state=0)

for matcher in [dt, svm, rf]:
    print('Evaluation Summary for ---> {}'.format(matcher.name))

    # Train using feature vectors from I
    dt.fit(table=H,exclude_attrs=['_id', 'ltable_id', 'rtable_id', 'gold'],target_attr='gold')

    # Predict on L
    predictions = dt.predict(table=L, exclude_attrs=['_id', 'ltable_id', 'rtable_id', 'gold'],
                  append=True, target_attr='predicted', inplace=False)

    # Evaluate the predictions
    eval_result = em.eval_matches(predictions, 'gold', 'predicted')
    em.print_eval_summary(eval_result)
    print('\n')
```

```
Evaluation Summary for ---> DecisionTree
Precision : 50.0% (1/2)
Recall : 50.0% (1/2)
F1 : 50.0%
False positives : 1 (out of 2 positive predictions)
False negatives : 1 (out of 8 negative predictions)


Evaluation Summary for ---> SVM
Precision : 50.0% (1/2)
Recall : 50.0% (1/2)
F1 : 50.0%
False positives : 1 (out of 2 positive predictions)
False negatives : 1 (out of 8 negative predictions)


Evaluation Summary for ---> RF
Precision : 50.0% (1/2)
Recall : 50.0% (1/2)
F1 : 50.0%
False positives : 1 (out of 2 positive predictions)
False negatives : 1 (out of 8 negative predictions)
```

2 Entity Resolution with Dedupe

In [51]:
```
import os
import csv
import re
import logging
import optparse

import dedupe
from unidecode import unidecode
```

2.1 Define functions

In [52]:
```
def preProcess(column):
    """
    Do a little bit of data cleaning with the help of Unidecode and Regex.
    Things like casing, extra spaces, quotes and new lines can be ignored.
    """
    column = unidecode(column)
    column = re.sub('  +', ' ', column)
    column = re.sub('\n', ' ', column)
    column = column.strip().strip('"').strip("'").lower().strip()
    # If data is missing, indicate that by setting the value to `None`
    if not column:
        column = None
    return column


def readData(filename):
    """
    Read in our data from a CSV file and create a dictionary of records,
    where the key is a unique record ID and each value is dict

    Note: Source code was modified to only read the first 10,000 rows
    """

    data_d = {}
    with open(filename) as f:
        reader = csv.DictReader(f)
        count=0
        for row in reader:
            clean_row = [(k, preProcess(v)) for (k, v) in row.items()]
            row_id = int(row['id'])
            data_d[row_id] = dict(clean_row)

            count+=1
            if count >= 10000: break

    return data_d
```

2.2 Read CSV files

In [53]:
```
# Read dataset, first 10,000 rows only
A1 = readData('../../Data for Assignments/citeseer.csv')
B1 = readData('../../Data for Assignments/dblp.csv')
```

In [54]:
```
len(A1), len(B1)
```

Out[54]:
```
(10000, 10000)
```

In [55]:
```
# show column names
print([v.keys() for k,v in A1.items()][0])
print([v.keys() for k,v in B1.items()][0])
```

```
dict_keys(['id', 'title', 'authors', 'journal', 'month', 'year', 'publication_type'])
dict_keys(['id', 'title', 'authors', 'journal', 'month', 'year', 'publication_type'])
```

2.3 Set up data in dedupe and also run entity matching measures

In [59]:
```
# Use RecordLink to join both datasets
fields = [
            {'field': 'title', 'type': 'String'},
            {'field': 'authors', 'type': 'String'}]

#Create a new linker object and pass our data model to it.
linker = dedupe.RecordLink(fields)

# run entity matching measures
linker.prepare_training(A1, B1, sample_size=1000)
```

```
INFO:dedupe.canopy_index:Removing stop word  c
INFO:dedupe.canopy_index:Removing stop word an
INFO:dedupe.canopy_index:Removing stop word ch
INFO:dedupe.canopy_index:Removing stop word he
INFO:dedupe.canopy_index:Removing stop word is
INFO:dedupe.canopy_index:Removing stop word r
INFO:dedupe.canopy_index:Removing stop word st
INFO:dedupe.canopy_index:Removing stop word  l
INFO:dedupe.canopy_index:Removing stop word do
INFO:dedupe.canopy_index:Removing stop word i
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word li
INFO:dedupe.canopy_index:Removing stop word on
INFO:dedupe.canopy_index:Removing stop word  d
INFO:dedupe.canopy_index:Removing stop word  t
INFO:dedupe.canopy_index:Removing stop word da
INFO:dedupe.canopy_index:Removing stop word o
INFO:dedupe.canopy_index:Removing stop word ti
INFO:dedupe.canopy_index:Removing stop word  a
INFO:dedupe.canopy_index:Removing stop word  h
INFO:dedupe.canopy_index:Removing stop word  j
INFO:dedupe.canopy_index:Removing stop word  p
INFO:dedupe.canopy_index:Removing stop word al
INFO:dedupe.canopy_index:Removing stop word ca
INFO:dedupe.canopy_index:Removing stop word e
INFO:dedupe.canopy_index:Removing stop word en
INFO:dedupe.canopy_index:Removing stop word ie
INFO:dedupe.canopy_index:Removing stop word io
INFO:dedupe.canopy_index:Removing stop word lo
INFO:dedupe.canopy_index:Removing stop word na
INFO:dedupe.canopy_index:Removing stop word ol
```

```
INFO:dedupe.canopy_index:Removing stop word ra
INFO:dedupe.canopy_index:Removing stop word s
INFO:dedupe.canopy_index:Removing stop word vi
INFO:dedupe.canopy_index:Removing stop word  k
INFO:dedupe.canopy_index:Removing stop word ar
INFO:dedupe.canopy_index:Removing stop word ei
INFO:dedupe.canopy_index:Removing stop word es
INFO:dedupe.canopy_index:Removing stop word ho
INFO:dedupe.canopy_index:Removing stop word k
INFO:dedupe.canopy_index:Removing stop word l
INFO:dedupe.canopy_index:Removing stop word le
INFO:dedupe.canopy_index:Removing stop word mi
INFO:dedupe.canopy_index:Removing stop word n
INFO:dedupe.canopy_index:Removing stop word va
INFO:dedupe.canopy_index:Removing stop word dr
INFO:dedupe.canopy_index:Removing stop word nd
INFO:dedupe.canopy_index:Removing stop word be
INFO:dedupe.canopy_index:Removing stop word or
INFO:dedupe.canopy_index:Removing stop word ro
INFO:dedupe.canopy_index:Removing stop word  g
INFO:dedupe.canopy_index:Removing stop word am
INFO:dedupe.canopy_index:Removing stop word h
INFO:dedupe.canopy_index:Removing stop word ic
INFO:dedupe.canopy_index:Removing stop word rt
INFO:dedupe.canopy_index:Removing stop word t
INFO:dedupe.canopy_index:Removing stop word te
INFO:dedupe.canopy_index:Removing stop word as
INFO:dedupe.canopy_index:Removing stop word ba
INFO:dedupe.canopy_index:Removing stop word ka
INFO:dedupe.canopy_index:Removing stop word m
INFO:dedupe.canopy_index:Removing stop word ne
INFO:dedupe.canopy_index:Removing stop word ad
INFO:dedupe.canopy_index:Removing stop word ge
INFO:dedupe.canopy_index:Removing stop word d
INFO:dedupe.canopy_index:Removing stop word os
INFO:dedupe.canopy_index:Removing stop word se
INFO:dedupe.canopy_index:Removing stop word pa
INFO:dedupe.canopy_index:Removing stop word hi
INFO:dedupe.canopy_index:Removing stop word th
INFO:dedupe.canopy_index:Removing stop word ll
INFO:dedupe.canopy_index:Removing stop word a
INFO:dedupe.canopy_index:Removing stop word for
INFO:dedupe.canopy_index:Removing stop word based
```

```
INFO:dedupe.canopy_index:Removing stop word and
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word the
INFO:dedupe.canopy_index:Removing stop word of
INFO:dedupe.canopy_index:Removing stop word  a
INFO:dedupe.canopy_index:Removing stop word  f
INFO:dedupe.canopy_index:Removing stop word  p
INFO:dedupe.canopy_index:Removing stop word a
INFO:dedupe.canopy_index:Removing stop word al
INFO:dedupe.canopy_index:Removing stop word as
INFO:dedupe.canopy_index:Removing stop word ba
INFO:dedupe.canopy_index:Removing stop word ch
INFO:dedupe.canopy_index:Removing stop word co
INFO:dedupe.canopy_index:Removing stop word ed
INFO:dedupe.canopy_index:Removing stop word er
INFO:dedupe.canopy_index:Removing stop word fo
INFO:dedupe.canopy_index:Removing stop word io
INFO:dedupe.canopy_index:Removing stop word le
INFO:dedupe.canopy_index:Removing stop word ne
INFO:dedupe.canopy_index:Removing stop word oc
INFO:dedupe.canopy_index:Removing stop word or
INFO:dedupe.canopy_index:Removing stop word pr
INFO:dedupe.canopy_index:Removing stop word r
INFO:dedupe.canopy_index:Removing stop word ro
INFO:dedupe.canopy_index:Removing stop word rt
INFO:dedupe.canopy_index:Removing stop word ss
INFO:dedupe.canopy_index:Removing stop word ti
INFO:dedupe.canopy_index:Removing stop word  l
INFO:dedupe.canopy_index:Removing stop word an
INFO:dedupe.canopy_index:Removing stop word ct
INFO:dedupe.canopy_index:Removing stop word e
INFO:dedupe.canopy_index:Removing stop word en
INFO:dedupe.canopy_index:Removing stop word g
INFO:dedupe.canopy_index:Removing stop word im
INFO:dedupe.canopy_index:Removing stop word is
INFO:dedupe.canopy_index:Removing stop word me
INFO:dedupe.canopy_index:Removing stop word nc
INFO:dedupe.canopy_index:Removing stop word ni
INFO:dedupe.canopy_index:Removing stop word pe
INFO:dedupe.canopy_index:Removing stop word si
INFO:dedupe.canopy_index:Removing stop word st
INFO:dedupe.canopy_index:Removing stop word ta
INFO:dedupe.canopy_index:Removing stop word un
```

```
INFO:dedupe.canopy_index:Removing stop word vi
INFO:dedupe.canopy_index:Removing stop word  i
INFO:dedupe.canopy_index:Removing stop word ab
INFO:dedupe.canopy_index:Removing stop word da
INFO:dedupe.canopy_index:Removing stop word iv
INFO:dedupe.canopy_index:Removing stop word ll
INFO:dedupe.canopy_index:Removing stop word ol
INFO:dedupe.canopy_index:Removing stop word ra
INFO:dedupe.canopy_index:Removing stop word th
INFO:dedupe.canopy_index:Removing stop word  m
INFO:dedupe.canopy_index:Removing stop word de
INFO:dedupe.canopy_index:Removing stop word he
INFO:dedupe.canopy_index:Removing stop word l
INFO:dedupe.canopy_index:Removing stop word nd
INFO:dedupe.canopy_index:Removing stop word tr
INFO:dedupe.canopy_index:Removing stop word  d
INFO:dedupe.canopy_index:Removing stop word  t
INFO:dedupe.canopy_index:Removing stop word f
INFO:dedupe.canopy_index:Removing stop word mo
INFO:dedupe.canopy_index:Removing stop word of
INFO:dedupe.canopy_index:Removing stop word rk
INFO:dedupe.canopy_index:Removing stop word so
INFO:dedupe.canopy_index:Removing stop word tw
INFO:dedupe.canopy_index:Removing stop word wo
INFO:dedupe.canopy_index:Removing stop word ca
INFO:dedupe.canopy_index:Removing stop word ge
INFO:dedupe.canopy_index:Removing stop word iz
INFO:dedupe.canopy_index:Removing stop word ob
INFO:dedupe.canopy_index:Removing stop word op
INFO:dedupe.canopy_index:Removing stop word sy
INFO:dedupe.canopy_index:Removing stop word us
INFO:dedupe.canopy_index:Removing stop word  b
INFO:dedupe.canopy_index:Removing stop word  r
INFO:dedupe.canopy_index:Removing stop word hi
INFO:dedupe.canopy_index:Removing stop word lt
INFO:dedupe.canopy_index:Removing stop word mu
INFO:dedupe.canopy_index:Removing stop word pa
INFO:dedupe.canopy_index:Removing stop word ul
INFO:dedupe.canopy_index:Removing stop word gr
INFO:dedupe.canopy_index:Removing stop word og
INFO:dedupe.canopy_index:Removing stop word  g
INFO:dedupe.canopy_index:Removing stop word gi
INFO:dedupe.canopy_index:Removing stop word ia
```

```
INFO:dedupe.canopy_index:Removing stop word ut
INFO:dedupe.canopy_index:Removing stop word ag
INFO:dedupe.canopy_index:Removing stop word no
INFO:dedupe.canopy_index:Removing stop word lo
INFO:dedupe.canopy_index:Removing stop word ev
INFO:dedupe.canopy_index:Removing stop word ts
INFO:dedupe.canopy_index:Removing stop word  u
INFO:dedupe.canopy_index:Removing stop word ci
INFO:dedupe.canopy_index:Removing stop word bi
INFO:dedupe.canopy_index:Removing stop word am
INFO:dedupe.canopy_index:Removing stop word m
INFO:dedupe.canopy_index:Removing stop word mp
INFO:dedupe.canopy_index:Removing stop word ie
INFO:dedupe.canopy_index:Removing stop word fi
INFO:dedupe.canopy_index:Removing stop word rm

INFO:dedupe.canopy_index:Removing stop word c
INFO:dedupe.canopy_index:Removing stop word pl
INFO:dedupe.canopy_index:Removing stop word lu
INFO:dedupe.canopy_index:Removing stop word o
INFO:dedupe.canopy_index:Removing stop word ex
INFO:dedupe.canopy_index:Removing stop word er
INFO:dedupe.canopy_index:Removing stop word ng
INFO:dedupe.canopy_index:Removing stop word ri
INFO:dedupe.canopy_index:Removing stop word  w
INFO:dedupe.canopy_index:Removing stop word g
INFO:dedupe.canopy_index:Removing stop word ia
INFO:dedupe.canopy_index:Removing stop word  r
INFO:dedupe.canopy_index:Removing stop word ha
INFO:dedupe.canopy_index:Removing stop word  f
INFO:dedupe.canopy_index:Removing stop word  m
INFO:dedupe.canopy_index:Removing stop word ma
INFO:dedupe.canopy_index:Removing stop word re
INFO:dedupe.canopy_index:Removing stop word de
INFO:dedupe.canopy_index:Removing stop word la
INFO:dedupe.canopy_index:Removing stop word mo
INFO:dedupe.canopy_index:Removing stop word nt
INFO:dedupe.canopy_index:Removing stop word il
INFO:dedupe.canopy_index:Removing stop word sh
INFO:dedupe.canopy_index:Removing stop word  n
INFO:dedupe.canopy_index:Removing stop word sa
INFO:dedupe.canopy_index:Removing stop word  b
INFO:dedupe.canopy_index:Removing stop word ni
```

```
INFO:dedupe.canopy_index:Removing stop word ta
INFO:dedupe.canopy_index:Removing stop word  c
INFO:dedupe.canopy_index:Removing stop word  n
INFO:dedupe.canopy_index:Removing stop word ac
INFO:dedupe.canopy_index:Removing stop word ap
INFO:dedupe.canopy_index:Removing stop word at
INFO:dedupe.canopy_index:Removing stop word d
INFO:dedupe.canopy_index:Removing stop word el
INFO:dedupe.canopy_index:Removing stop word h
INFO:dedupe.canopy_index:Removing stop word la
INFO:dedupe.canopy_index:Removing stop word n
INFO:dedupe.canopy_index:Removing stop word on
INFO:dedupe.canopy_index:Removing stop word pp
INFO:dedupe.canopy_index:Removing stop word re
INFO:dedupe.canopy_index:Removing stop word se
INFO:dedupe.canopy_index:Removing stop word t
INFO:dedupe.canopy_index:Removing stop word  s
INFO:dedupe.canopy_index:Removing stop word ar
INFO:dedupe.canopy_index:Removing stop word ea
INFO:dedupe.canopy_index:Removing stop word il
INFO:dedupe.canopy_index:Removing stop word it
INFO:dedupe.canopy_index:Removing stop word mi
INFO:dedupe.canopy_index:Removing stop word ng
INFO:dedupe.canopy_index:Removing stop word ri
INFO:dedupe.canopy_index:Removing stop word su
INFO:dedupe.canopy_index:Removing stop word y
INFO:dedupe.canopy_index:Removing stop word  w
INFO:dedupe.canopy_index:Removing stop word ms
INFO:dedupe.canopy_index:Removing stop word s
INFO:dedupe.canopy_index:Removing stop word ve
INFO:dedupe.canopy_index:Removing stop word em
INFO:dedupe.canopy_index:Removing stop word ic
INFO:dedupe.canopy_index:Removing stop word ma
INFO:dedupe.canopy_index:Removing stop word ns
INFO:dedupe.canopy_index:Removing stop word  o
INFO:dedupe.canopy_index:Removing stop word et
INFO:dedupe.canopy_index:Removing stop word na
INFO:dedupe.canopy_index:Removing stop word rs
INFO:dedupe.canopy_index:Removing stop word ur
INFO:dedupe.canopy_index:Removing stop word ec
INFO:dedupe.canopy_index:Removing stop word li
INFO:dedupe.canopy_index:Removing stop word od
INFO:dedupe.canopy_index:Removing stop word te
```

```
INFO:dedupe.canopy_index:Removing stop word  e
INFO:dedupe.canopy_index:Removing stop word ou
INFO:dedupe.canopy_index:Removing stop word ig
INFO:dedupe.canopy_index:Removing stop word ry
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinSearchPredicate: (3, authors), SimplePredicate: (commonTwoTokens, title))
```

2.4 Active learning part

In [60]:
```
dedupe.console_label(linker)
```

```
title : node localization using mobile robots in delay-tolerant sensor networks
authors : pubudu n pathirana, nirupama bulusu, andrey v savkin, sanjay jha

title : node localization using mobile robots in delay-tolerant sensor networks.
authors : pubudu n. pathirana, nirupama bulusu, andrey v. savkin, sanjay jha, thanh dang

0/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished

y

title : a fast 2-approximation algorithm for the minimum manhattan network problem
authors : zeyu guo, he sun, hong zhu

title : a fast 2-approximation algorithm for the minimum manhattan network problem.
authors : zeyu guo, he sun 0001, hong zhu

1/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(LevenshteinSearchPredicate: (4, title), SimplePredicate: (oneGramFingerprint, title))
title : protection of database security via collaborative inference detection 1 abstract
authors : yu chen, wesley w. chu

title : protection of database security via collaborative inference detection.
authors : yu chen, wesley w. chu

2/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : 2010 eleventh brazilian symposium on neural networks evolution strategies with q-gaussian mutation for dynamic optimization problems
authors : renato tins, shengxiang yang

title : evolution strategies with q-gaussian mutation for dynamic optimization problems.
```

```
authors : renato tins, shengxiang yang

3/10 positive, 0/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious


n

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (commonThreeTokens, authors), TfidfTextSearchPredicate: (0.4, title))
title : contents
authors : chris brown

title : portrait identification in digitized paintings on the basis of a face detection system.
authors : christos-nikolaos anagnostopoulos, ioannis anagnostopoulos, ilias maglogiannis, dimitris vergados

3/10 positive, 1/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : education
authors : dr. sylvia ratnasamy

title : education.
authors : paola salomoni, silvia mirri, stefano ferretti, marco roccetti

3/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : action rules mining
authors : angelina a. tzacheva, zbigniew w. ras

title : action rules mining
authors : agnieszka dardzinska

3/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
y
title : introduction
authors : globalized economy, dr. wesley cragg, director gardiner, programme business ethics

title : introduction.
authors : hamid r. tizhoosh, mario ventresca

4/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(SimplePredicate: (doubleMetaphone, title), SimplePredicate: (firstTokenPredicate, title))
title : implementation
authors : david short, malcolm dellow

title : implementation of learning objects using j2me: putting into practice the concept of m-learning in brazil.
authors : leandro ramos de oliveira, roseclea duarte medina

4/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : introduction to and report from the symposium on management of geodetic data,
authors : c. boucher, k. poder, c. r. schwarz, c. c. tscherning

title : introduction.
authors : steve cunningham, roger j. hubbold

4/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : notice
authors : john benner, satyen deb, robert mcconnell, electrochemical society, john benner, satyen deb, robert mcconnell, electrochemical society

title : getalife - an artificial life environment for the evaluation of agent-based systems and evolutionary algorithms for reinforcement learnin
g.
```

authors : daniel machado, miguel rocha

4/10 positive, 2/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
n

title : 6. the value of evaluation through the local implementation
authors : suzanne hoverman

title : 6th.
authors : damien perritaz, christophe salzmann, denis gillet, olivier naef, jacques bapst, frdric barras, elena mugellini, omar abou khaled

4/10 positive, 3/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : summary
authors : kenneth a. taylor, holger schmitz, mary c. reedy, yale e. goldman, clara franzini-armstrong, hiroyuki sasaki, richard t. tregear, kate poo
le, carmen lucaveche, robert j. edwards, li fan chen, hanspeter winkler, michael k. reedy, cambridge cb qh, united kingdom, abteilung biophysik

title : sequence jobs and assign due dates with uncertain processing times and quadratic penalty functions.
authors : yu xia 0004, bintong chen, jinfeng yue

4/10 positive, 4/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : preliminaries
authors : prime contractor, nexant inc, southern african power pool (sapp

title : preliminaries.
authors : andrzej p. wierzbicki, yoshiteru nakamori

4/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : introduction
authors : globalized economy, dr. wesley cragg, director gardiner, programme business ethics

title : introduction.
authors : olga pombo, juan manuel torres, john symons

4/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : editorial
authors : werner hlzl, andreas reinstaller

title : editorial.
authors : martin hgele, paul-gerhard plger

4/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

u

title : authors
authors : of ips, pieter van der wolf, tomas henriksson, alistair bruce, axel jantsch, mikael millberg, zhonghai lu, alain clouard

title : automatic generation of semantic metadata as basis for user modeling and adaptation.
authors : kees van der sluijs, geert-jan houben

4/10 positive, 5/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : plans
authors : james ramsay

title : transactional memory
authors : james r. larus, ravi rajwar

4/10 positive, 6/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : and
authors : hans lindblad

title : visualisation of cluster analysis results.
authors : hans-joachim mucha, hans-georg bartel, carlos morales-merino

4/10 positive, 7/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : www.fmre-gske.be
authors : koningin elisabeth, reine elisabeth, geneeskundige stichting, koningin elisabeth, inleiding verslag, activiteiten gske fmre

title : how to engineer robotic organisms and swarms? - bio-inspiration, bio-mimicry, and artificial evolution in embodied self-organized systems.
authors : thomas schmickl

4/10 positive, 8/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : and
authors : peter r. taylor, charles w. bauschlicher, david w. schwenke

title : experience-centered design: designers, users, and communities in dialogue
authors : peter c. wright, john c. mccarthy

4/10 positive, 9/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : contents
authors : peter d. karp

title : behaviorally founded recommendation algorithm for browsing assistance systems.
authors : peter gczy, noriaki izumi, shotaro akaho, kiti hasida

4/10 positive, 10/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : contents
authors : james s. plank

title : incongruence detection in audio-visual processing.
authors : michal havlena, jan heller, hendrik kayser, jrg-hendrik bach, jrn anemller, toms pajdla

4/10 positive, 11/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : physical layers
authors : pierre boulet, aurlien gron, andrew a. vladimirov, konstantin v. gavrilenko, andrei a. mikhailovsky, mac layer

title : c.s. peirce and artificial intelligence: historical heritage and (new) theoretical stakes.
authors : pierre steiner

4/10 positive, 12/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : the design and performance of dynamic and static configuration mechanisms in component middleware for distributed real-time and embedded sys
tems abstract
authors : venkita subramonian, liang-jui shen, christopher gill, nanbor wang

title : products of automata
authors : ferenc gcseg

4/10 positive, 13/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : problem formulation
authors : wiwek deshmukh, advisor dr. yingshu li

```
title : the problem of determinacy of infinite games from an intuitionistic point of view.
authors : wim veldman

4/10 positive, 14/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
n

title : contents
authors : matthew flatt, robert bruce findler, john clements, i windowing toolbox

title : machine learning techniques for multimedia - case studies on organization and retrieval
authors : matthieu cord, pdraig cunningham

4/10 positive, 15/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : 3
authors : hallie quinn brown

title : blogging from the top: a survey of higher education leaders' use of web 2.0 technologies.
authors : david c. wyld

4/10 positive, 16/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : abstract
authors : alexander rakhlin, ambuj tewari, peter bartlett, all rights reserved, alexander rakhlin, peter l. bartlett, ambuj tewari

title : knowledge annotation: making implicit knowledge explicit
authors : alexiei dingli

4/10 positive, 17/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n
```

title : www.publish.csiro.au/journals/sh sexual health, 2010, 7, 3134 misclassification bias: diversity in conceptualisations about having had sex
authors : stephanie a. s, ers a, on j. hill a, william l. yarber a, cynthia a. graham a, richard a. crosby a, robin r. milhausen a

title : vagueness and logic.
authors : stewart shapiro

4/10 positive, 18/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : programs with common sense
authors : john mccarthy

title : how to describe and propagate uncertainty when processing time series: metrological and computational challenges, with potential application
s to environmental studies.
authors : christian servin, martine ceberio, aline jaimes, craig e. tweedie, vladik kreinovich

4/10 positive, 19/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : chairperson
authors : vijay tirumalai, william a. stapleton, ph. d, keith a. woodbury, ph. d, david j. jackson, ph. d, kenneth g. ricks, ph. d

title : evaluation and certifications for component packages software.
authors : haeng-kon kim

4/10 positive, 20/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : 1
authors : anca muscholl, doron peled

title : dsp for matlab and labview iii: digital filter design
authors : forester w. isen

4/10 positive, 21/10 negative

Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : editorial issue number one hundred and prospect values education: the australian experience open file: learningto livetogether through thete
aching of history and geography part one: the duty, ability and desire for peaceful co-existence
authors : quartely review of, yves andr, antoine bailly, yves andr, zuzana wienerova, clarence edward, beeby w. l. renwick

title : bioinformatics: an introduction, second edition
authors : jeremy j. ramsden

4/10 positive, 22/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : a a abb
authors : b bcc, c caa

title : security and privacy challenges of a digital government.
authors : james b. d. joshi, arif ghafoor, walid g. aref, eugene h. spafford

4/10 positive, 23/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : audio
authors : dr. hans stork

title : time-domain beamforming and blind source separation - speech input in the car environment
authors : julien bourgeois, wolfgang minker

4/10 positive, 24/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : table of contents
authors : jonathan henke, shannon lawrence, ian miller, irene perciali, ph. d, david nasatir, ph. d., charis kaskiris, cara bautista

title : routing protocols for next-generation networks inspired by collective behaviors of insect societies: an overview.
authors : muddassar farooq, gianni a. di caro

4/10 positive, 25/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : petition to list the scalloped hammerhead shark (sphyrna lewini) under the u.s. endangered species act either worldwide or as one or more di
stinct population segments
authors : frank burek, national oceanic, atmospheric administration

title : evolutionary tolerance.
authors : lus moniz pereira

4/10 positive, 26/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : component design
authors : katherine c. morris, david sauder, sandy ressler, barbara h. franklin, john w. lyons, katherine c. morris, david sauder, sandy ressler

title : partially distributed emergency teams: considerations of decision support for virtual communities of practice.
authors : linda plotnick, murray turoff, connie white

4/10 positive, 27/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : authors
authors : leonello serva apat, marino sorriso-valvo irpi-cnr, j. wasowski

title : gait analysis and human motion tracking.
authors : huiyu zhou

4/10 positive, 28/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : support vector machines and kernel methods
authors : geoff gordon

title : estimating quality of support vector machines learning under probabilistic and interval uncertainty: algorithms and computational complexit
y.
authors : canh hao nguyen, tu bao ho, vladik kreinovich

4/10 positive, 29/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : references
authors : li ding, aniko sabo, nicolas berkowicz, rekha r. meyer, yoram shotl, mark r. johnson, kymberlie h, richard k. wilson, john spieth, email a
lerting, li ding, aniko sabo, nicolas berkowicz, rekha r. meyer, yoram shotl, mark r. johnson, kymberlie h. pepin, richard k. wilson, john spieth

title : an user-driven tool for interactive retrieval of non annotated videos.
authors : maria ngeles mendoza, toms arnau, isabel gracia, filiberto pla, nicolas prez de la blanca

4/10 positive, 30/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : summary
authors : doris wagner, frank wellmer, kieran dilks, dilusha william, michael r. smith, prakash p. kumar, jose luis riechmann, andrew j. greenl, ell
iot m. meyerowitz

title : bandwidth extension of speech using perceptual criteria
authors : visar berisha, steven sandoval, julie liss

4/10 positive, 31/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : this article has been accepted for publication in a future issue of this journal, but has not been fully edited. content may change prior
to final publication. ieee transactions on circuits and systemsii: express briefs analysis of the bridged t-coil cir
authors : h paramesh, student member, david j. allstot

```
title : evaluating the next generation of multimedia software.
authors : ray adams

4/10 positive, 32/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : contents
authors : mary l. mcnabb, gilbert valdez, jeri nowakowski, mark hawkes

title : design methods for fluid construction grammar.
authors : luc steels

4/10 positive, 33/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : keeping things simple: finding frequent item sets by recursive elimination
authors : christian borgelt

title : simple algorithms for frequent item set mining.
authors : christian borgelt

4/10 positive, 34/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : contributors
authors : user group all, american english, american english, colleen cook, amy hoseth, fred heath, martha kyrillidou, brucethompson jonathan, d. so
usa, duane webster

title : gesture, gaze and persuasive strategies in political discourse.
authors : isabella poggi, laura vincze

5/10 positive, 34/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious
```

```
n

INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(TfidfTextSearchPredicate: (0.2, title), TfidfTextSearchPredicate: (0.8, authors))
INFO:dedupe.training:(SimplePredicate: (doubleMetaphone, title), SimplePredicate: (firstTokenPredicate, title))
title : data center evolution
authors : krishna kant

title : configuration management security in data center environments.
authors : krishna kant

5/10 positive, 35/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : distributed ( + 1)-coloring in linear (in ) time
authors : leonid barenboim, michael elkin

title : distributed graph coloring: fundamentals and recent developments
authors : leonid barenboim, michael elkin

6/10 positive, 35/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : generalized patterns in words and permutations
authors : sergey kitaev

title : patterns in permutations and words
authors : sergey kitaev

7/10 positive, 35/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : musical instrument classification and duet analysis employing music information retrieval techniques
authors : bozena kostek

title : perception-based data processing in acoustics: applications to music information retrieval and psychophysiology
```

```
authors : bozena kostek

8/10 positive, 35/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : splittability of bilexical context-free grammars is undecidable
authors : mark-jan nederhof, giorgio satta

title : probabilistic parsing.
authors : mark-jan nederhof, giorgio satta

9/10 positive, 35/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : keeping things simple: finding frequent item sets by recursive elimination
authors : christian borgelt

title : network creation: overview.
authors : christian borgelt

9/10 positive, 36/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

n

title : mathematical morphology on bipolar fuzzy sets
authors : isabelle bloch

title : bipolar fuzzy spatial information: geometry, morphology, spatial reasoning.
authors : isabelle bloch

9/10 positive, 37/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

y

title : modular reuse of ontologies: theory and practice
```

```
authors : bernardo cuenca grau, ian horrocks, yevgeny kazakov, ulrike sattler

title : extracting modules from ontologies: a logic-based approach.
authors : bernardo cuenca grau, ian horrocks, yevgeny kazakov, ulrike sattler

10/10 positive, 37/10 negative
Do these records refer to the same thing?
(y)es / (n)o / (u)nsure / (f)inished / (p)revious

f

Finished labeling
```

2.5 Train the classifier

In [61]:
```
linker.train()
```

```
INFO:rlr.crossvalidation:using cross validation to find optimum alpha...
INFO:rlr.crossvalidation:optimum alpha: 0.000010, score 0.4359232627630916
INFO:dedupe.training:Final predicate set:
INFO:dedupe.training:(TfidfTextSearchPredicate: (0.2, title), TfidfTextSearchPredicate: (0.8, authors))
INFO:dedupe.training:(SimplePredicate: (doubleMetaphone, title), SimplePredicate: (firstTokenPredicate, title))
```

2.6 Get matching pairs

In [62]:
```
# options join(), pairs(), score(), one_to_one(), many_to_one()

# use join() and pairs() to illustrate the matching, display the first 10 results
join = linker.join(A1, B1, threshold=0.0)
print('Using join() method, number of matches = {}'.format(len(join)))
print(join[:10])

pairs = linker.pairs(A1, B1)
pairs_list = list(pairs)
print('\nUsing pairs() method, number of matches = {}'.format(len(pairs_list)))
print(pairs_list[:10])
```

```
INFO:dedupe.canopy_index:Removing stop word a
INFO:dedupe.canopy_index:Removing stop word for
INFO:dedupe.canopy_index:Removing stop word based
INFO:dedupe.canopy_index:Removing stop word and
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word the
INFO:dedupe.canopy_index:Removing stop word of

Using join() method, number of matches = 28
[((7208, 3414), 0.8717715), ((7600, 8114), 0.7836728), ((9320, 8507), 0.7299457), ((8803, 9361), 0.69728917), ((2648, 4778), 0.624293), ((9139, 803
7), 0.605023), ((9206, 7820), 0.60468143), ((8589, 8164), 0.59365), ((5033, 9184), 0.5700371), ((9045, 9963), 0.5240481)]

INFO:dedupe.canopy_index:Removing stop word a
INFO:dedupe.canopy_index:Removing stop word for
INFO:dedupe.canopy_index:Removing stop word based
INFO:dedupe.canopy_index:Removing stop word and
INFO:dedupe.canopy_index:Removing stop word in
INFO:dedupe.canopy_index:Removing stop word the
INFO:dedupe.canopy_index:Removing stop word of


Using pairs() method, number of matches = 61
[((774, {'id': '774', 'title': 'numerical algorithms based on', 'authors': 'pj. ponenti, j. liandrat, biorthogonal wavelets, j. liandrap', 'journ
al': None, 'month': None, 'year': '1996', 'publication_type': None}), (4017, {'id': '4017', 'title': 'numerical prediction of friction, wear, hea
t generation and lubrication in case of sliding rubber components.', 'authors': 'tibor j. goda', 'journal': None, 'month': None, 'year': '2009',
'publication_type': 'incollection'})), ((1046, {'id': '1046', 'title': 'identifying idiomatic expressions using automatic word-alignment', 'autho
rs': 'begoa villada moirn, jrg tiedemann', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (8508, {'id': '8508', 'titl
e': 'bitext alignment', 'authors': 'jrg tiedemann', 'journal': None, 'month': None, 'year': '2011', 'publication_type': 'book'})), ((1079, {'id':
'1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': Non
e, 'month': None, 'year': None, 'publication_type': None}), (546, {'id': '546', 'title': 'introduction.', 'authors': 'yuliang zheng', 'journal':
```

None, 'month': None, 'year': '2010', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (791, {'id': '791', 'title': 'introduction.', 'authors': 'william sims bainbridge', 'journal': None, 'month': None, 'year': '2010', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (827, {'id': '827', 'title': 'introduction.', 'authors': 'nikolaos dimakis, john soldatos, lazaros polymenakos', 'journal': None, 'month': None, 'year': '2009', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (861, {'id': '861', 'title': 'introduction.', 'authors': 'ian douglas, zhengjie liu', 'journal': None, 'month': None, 'year': '2011', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (883, {'id': '883', 'title': 'introduction.', 'authors': 'emmanuel dubois 0001, philip d. gray, laurence nigay', 'journal': None, 'month': None, 'year': '2010', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (925, {'id': '925', 'title': 'introduction.', 'authors': 'fabio patern', 'journal': None, 'month': None, 'year': '2011', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (959, {'id': '959', 'title': 'introduction.', 'authors': 'gustavo rossi, daniel schwabe, luis olsina, oscar pastor', 'journal': None, 'month': None, 'year': '2008', 'publication_type': 'incollection'})), ((1079, {'id': '1079', 'title': 'introduction', 'authors': 'globalized economy, dr. wesley cragg, director gardiner, programme business ethics', 'journal': None, 'month': None, 'year': None, 'publication_type': None}), (1120, {'id': '1120', 'title': 'introduction.', 'authors': 'georges g. grinstein, haim levkowitz', 'journal': None, 'month': None, 'year': '1995', 'publication_type': 'incollection'}))]

In [ ]: