Generating Synthetic Data Mimicking the Real Cancer Population

Sepideh Mosaferi

National Cancer Institute (NCI) Presentation
University of Maryland College Park

July 19, 2016

Intro

What Is Synthetic Data?

Synthetic data are “any production data applicable to a given situation that are not obtained by direct measurement”. The synthesized data might be directly obtained by perturbing the real data or based on some assumptions, distributions and frequency tables which might follow a particular real situation.

Why We Produce It?

- For the confidentiality purpose to protect disclosure risk
- For the Research Purpose
- For uniformity of evaluating some procedures or software

Synthetic Data for the De-duplication & Record Linkage Purpose

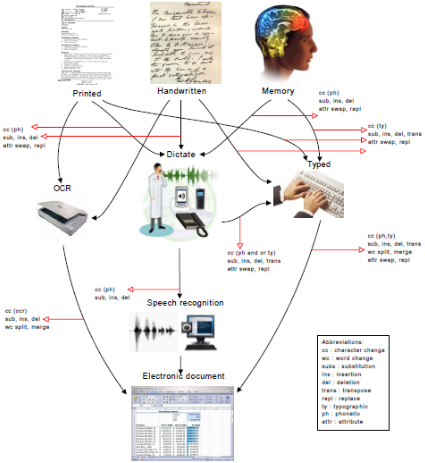
The real data set is dirty. Some entities exist multiple times. The formats of fields are not the same, and there might be variety kinds of typos. Therefore if we want to synthesize a real data for the de-duplication or record linkage purpose, we need to consider all of these situations.

Purpose & Goal of This Research:
Generating A Synthetic Data Mimicking the Real Cancer Population

- We need to consider different possibilities of errors that might happen in the real situation for the fields
- We need to incorporate frequency distributions or tables obtained from registered data or online reliable sources.

Cont'd Purpose & Goal of This Research:

- To have different scenarios we make fields dependent. For example a specific kind of cancer could be dependent to the gender. Or cancer and blood pressure could vary based on the age.
- We consider Zipf, Uniform, or Poisson distributions for randomly selecting units that need to be duplicated.
- Matching variables that we considered are: First name, Last name, SSN, DOB, State, Zipcode, Age, Phone Number, Cancer, Blood pressure, etc.



Abbreviations:
cc : character change
wc : word change
subs : substitution
ins : insertion
del : deletion
trans : transpose
repl : replace
ty : typographic
ph : phonetic
attr : attribute

Software

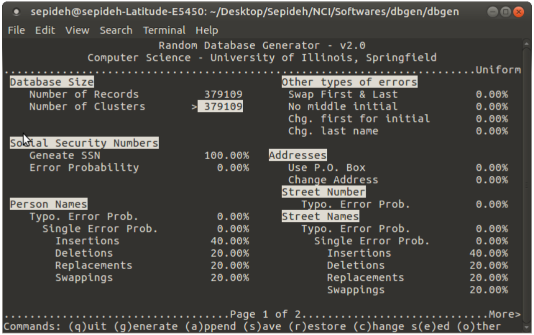
1. DBGen (C programming) developed around 1990s; producing US mailing address
2. GeCo (Python programming) developed from 2002-2008; for the special purpose of bio-medical studies

Both programs are capable of producing one synthesized data set with unique IDs, duplication and errors.
We should give them the frequency tables not the full raw data set.

DBGen

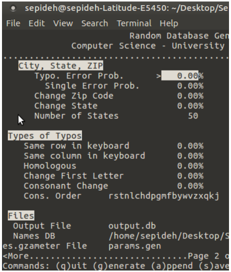
1. We cannot incorporate as many errors as we want
2. We can generate fields such as: First Name, Middle Initial, Last Name, Street Number, P.O. Box, Apartment Number, City, State, Zipcode.
3. We can consider Insertions, deletions, replacements, swapping, and other errors.
4. The program is old and not handy. For changing it, we need to write a lot of codes

Cont'd DBGen


sepideh@sepideh-Latitude-E5450: ~/Desktop/Sepideh/NCI/Softwares/dbgen/dbgen
File Edit View Search Terminal Help

```
                    Random Database Generator - v2.0
           Computer Science - University of Illinois, Springfield
.........................................................................Uniform
Database Size                              Other types of errors
   Number of Records          379109        Swap First & Last              0.00%
   Number of Clusters        > 379109       No middle initial              0.00%
                                            Chg. first for initial         0.00%
                                            Chg. last name                 0.00%

Social Security Numbers
   Geneate SSN                100.00%     Addresses
   Error Probability            0.00%       Use P.O. Box                   0.00%
                                            Change Address                 0.00%
                                           Street Number
Person Names                                   Typo. Error Prob.           0.00%
   Typo. Error Prob.            0.00%      Street Names
     Single Error Prob.         0.00%          Typo. Error Prob.           0.00%
       Insertions              40.00%            Single Error Prob.        0.00%
       Deletions               20.00%              Insertions             40.00%
       Replacements            20.00%              Deletions              20.00%
       Swappings               20.00%              Replacements           20.00%
                                                   Swappings              20.00%
..................................Page 1 of 2................................More>
Commands: (q)uit (g)enerate (a)ppend (s)ave (r)estore (c)hange s(e)ed (o)ther
```


sepideh@sepideh-Latitude-E5450: ~/Desktop/Se
File Edit View Search Terminal Help

```
                         Random Database Gen
                 Computer Science - University
...........................................
  City, State, ZIP
     Typo. Error Prob.        >    0.00%
       Single Error Prob.          0.00%
     Change Zip Code               0.00%
     Change State                  0.00%
     Number of States                50

 Types of Typos
   Same row in keyboard            0.00%
   Same column in keyboard         0.00%
   Homologous                      0.00%
   Change First Letter             0.00%
   Consonant Change                0.00%
   Cons. Order      rstnlchdpgmfbywvzxqkj

 Files
  Output File       output.db
  Names DB          /home/sepideh/Desktop/S
es.gzameter File    params.gen
<More.............................Page 2 o
Commands: (q)uit (g)enerate (a)ppend (s)ave
```

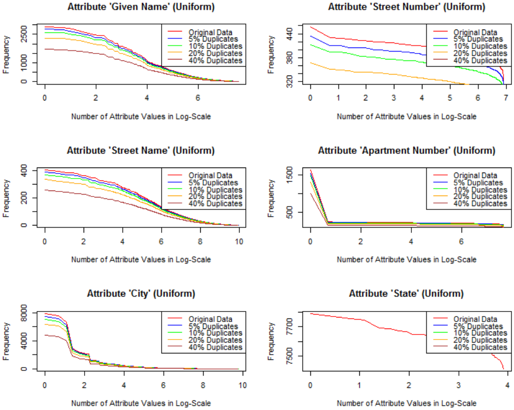
An Empirical Study (DBGen)

To test whether the DBGen (since it is old) is good for generating data, we implemented an experimental study. We assumed n (number of units)=c (number of clusters)
I generated an error free data set based on the frequency tables coming from Automated Vital Statistics System under the Institute for Social, Behavioral, and Economic Research at the University of California, Santa Barbara, CA 93106. Number of girls=240311 and boys=252972.

DBGen Evaluation

To evaluate the duplicated data set, we considered %5, %10, %20, and %40 duplication. Then we selected six attributes with their frequency distributions, Average Frequencies, and Standard Deviations.

DBGen Evaluation



Attribute 'Given Name' (Uniform)

Attribute 'Street Number' (Uniform)

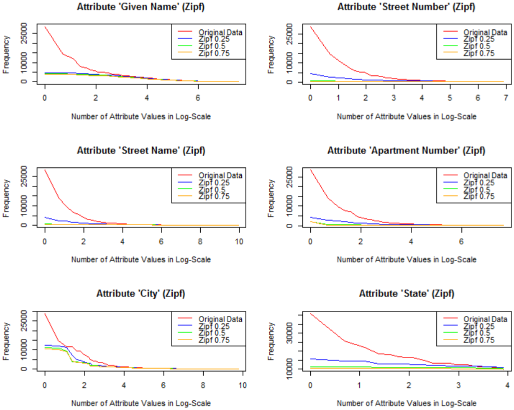
Attribute 'Street Name' (Uniform)

Attribute 'Apartment Number' (Uniform)

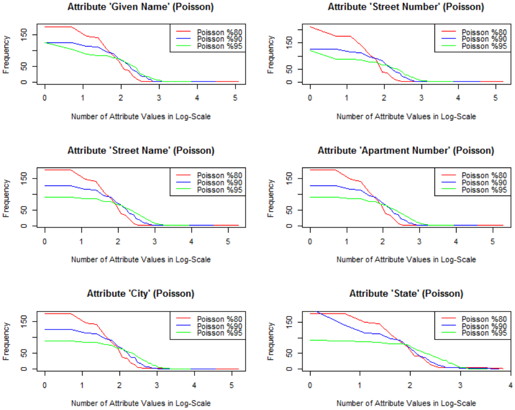
Attribute 'City' (Uniform)

Attribute 'State' (Uniform)

DBGen Evaluation



DBGen Evaluation


Attribute 'Given Name' (Poisson)

Attribute 'Street Number' (Poisson)

Attribute 'Street Name' (Poisson)

Attribute 'Apartment Number' (Poisson)

Attribute 'City' (Poisson)

Attribute 'State' (Poisson)

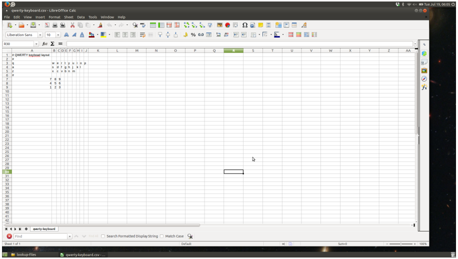
GeCo

1. Capable of incorporating variety kinds of errors, missing values, etc.
2. We can consider dependent attributes.
3. We need to define the number of duplication; otherwise, the program does not work.
4. Writing codes in Python is usually easier rather than C.

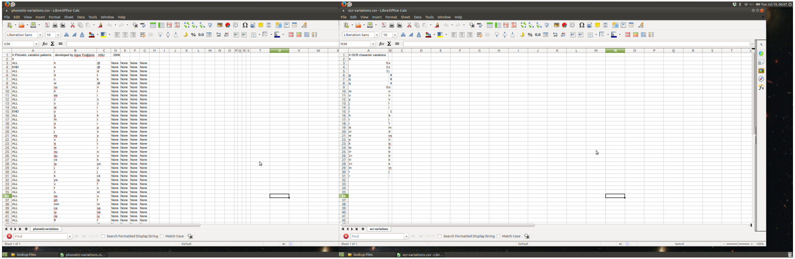
GeCo (As Our Final Focus)

We wrote codes and merged them with the original GeCo codes. I produced a big data set with variety of errors. Since the ids are unique just per data set not across the data set (generating the data is based on random process), we split the data set into 2 data set with different sizes after permutation for the purpose of record linkage. Unique IDs make it easy to evaluate the matching process since we have the true situations.

GeCo (Some Possible Errors)


qwerty-keyboard.csv - LibreOffice Calc
File Edit View Insert Format Sheet Data Tools Window Help
Liberation Sans
10
R30
qwerty-keyboard
Sheet 1 of 1
Default
lookup-files
qwerty-keyboard.csv - ...

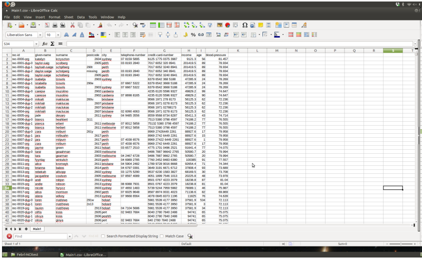
A	B	C	D	E	F	G	H	I	J
# QWERTY keyboad layout									
#									
q	w	e	r	t	y	u	i	o	p
a	s	d	f	g	h	j	k	l	
z	x	c	v	b	n	m			
#									
7	8	9							
4	5	6							
1	2	3							


phonetic-variations.csv - LibreOffice Calc
File Edit View Insert Format Sheet Data Tools Window Help
Liberation Sans
10
U36
phonetic-variations
Find
Search Formatted Display String
Match Case
Sheet 1 of 1
Default
lookup-files
phonetic-variations.cs...
ocr-variations.csv - LibreOffice Calc
N36
ocr-variations
ocr-variations.csv - Libr...

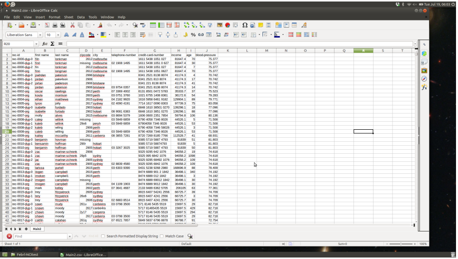
A	B	C	D	E	F	G
# Phonetic variation patterns	developed by Agus Pudjijono	ANU	2008			
#						
ALL	h	@	None	None	None	None
END	e	@	None	None	None	None
ALL	t	d	None	None	None	None
ALL	d	t	None	None	None	None
ALL	c	k	None	None	None	None
ALL	w	@	None	None	None	None
ALL	nn	n	None	None	None	None
ALL	ll	l	None	None	None	None
ALL	ee	i	None	None	None	None
ALL	z	s	None	None	None	None
ALL	s	z	None	None	None	None
ALL	ie	i	None	None	None	None
END	y	i	None	None	None	None
ALL	g	k	None	None	None	None
ALL	th	t	None	None	None	None
ALL	rr	r	None	None	None	None
ALL	b	p	None	None	None	None
ALL	j	s	None	None	None	None
ALL	ey	y	None	None	None	None
ALL	s	j	None	None	None	None
ALL	tt	t	None	None	None	None
ALL	ei	i	None	None	None	None
ALL	ou	o	None	None	None	None
ALL	au	o	None	None	None	None
ALL	ck	k	None	None	None	None
ALL	ia	ya	None	None	None	None
ALL	j	z	None	None	None	None
ALL	z	j	None	None	None	None
ALL	k	ck	None	None	None	None
ALL	ya	ia	None	None	None	None
ALL	v	f	None	None	None	None
ALL	f	v	None	None	None	None
ALL	s	st	None	None	None	None
ALL	oo	u	None	None	None	None
ALL	ph	f	None	None	None	None
ALL	mm	m	None	None	None	None
ALL	ce	se	None	None	None	None
ALL	ry	rie	None	None	None	None
ALL	rie	ry	None	None	None	None
ALL	ff	f	None	None	None	None

A	B
# OCR character variations	
#	
5	s
2	z
1	|
g	9
q	9
q	4
0	o
m	n
u	v
y	v
j	l
j	i
l	1
h	b
l	i
i	1
iii	m
cl	d
w	vv
ri	n
k	lc
lo	b
l>	b
1>	b
l<	k
1<	k
m	rn
l	|
i	:

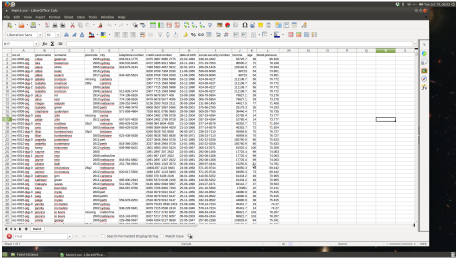
How the Produced Data Sets Look Like?


Main1.csv - LibreOffice Calc
File Edit View Insert Format Sheet Data Tools Window Help
Liberation Sans
10
S34
Main1
Find
Search Formatted Display String
Match Case
Sheet 1 of 1
Default
Febrl-NCItest
Main1.csv - LibreOffice...

rec-id	given-name	surname	postcode	city	telephone-number	credit-card-number	income	age	blood-pressure
rec-0000-org	katelyn	krzyszton	2604	sydney	07 8159 5895	6125 1775 0375 3987	9121.3	56	81.457
rec-0001-dup-0	taylor-saig	scolberg	2905	pdrth	03 8193 2840	7017 6052 320 8941	201419.5	89	78.934
rec-0001-dup-1	tayloah-saige	schplberg	290t	perth		7017 6052 3240 8941	201419.5	89	78.934
rec-0001-dup-2	daylor-saige	scholperg	missing	perdh	03 8193 2840	7017 6052 340 8941	201419.5	89	78.934
rec-0001-org	taylor-saige	scholberg	2905	perth	03 8193 2840	7017 6052 3240 8941	201419.5	89	78.934
rec-0002-dup-0	isabella	tzouris	2903	sydney		6379 6542 368 5188	47285.9	24	78.269
rec-0002-dup-1	istabella	tzouris	290e		07 6667 5322	6379 6542 3668 5188	47285.9	24	78.269
rec-0002-org	isabella	tsouris	2903	sydney	07 6667 5322	6379 6542 3668 5188	47285.9	24	78.269
rec-0003-dup-0	carejse	muzolino	2902	canberra		4235 6120 5596 9327	49829.2	99	74.647
rec-0003-org	caresse	musolino	2902	canberra	07 8698 8165	4235 6120 5596 9327	49829.2	90	74.647
rec-0004-dup-0	mikaili	maklucas	260u	brisbane		9566 1971 278 8173	56125.3	62	72.236
rec-0004-dup-1	mikhali	maklucas	2607	pristane		9566 1971 0278 8173	56125.3	62	72.236
rec-0004-dup-2	mikhaili	maclukas	2607	brisbanw		9566 1971 02788173	56125.3	62	72.236
rec-0004-org	mikhaili	maclucas	2607	brisbane	02 6090 4063	9566 1971 0278 8173	56125.3	62	72.236
rec-0005-org	jade	desaro	2611	sydney	04 9485 3556	8559 9568 8734 8287	65411.3	43	74.714
rec-0006-dup-0	biancs	heahbert	2t11			7513 5380 3798 4597	74188.2	77	76.555
rec-0006-dup-1	pianca	erbert	2611	melboutje	07 8012 5858	75132 5380 3798 4597	74188.2	77	76.555
rec-0006-org	bianca	herbert	2611	melbourne	07 8012 5858	7513 5380 3798 4597	74188.2	77	76.555
rec-0007-dup-0	zara	milburri	261y	perth		8969 27426449 2261	88927.6	17	79.958
rec-0007-dup-1	jara	milbuahn	2617	perth		8969 2742 6449 2261	88927.6	15	79.958
rec-0007-dup-2	jara	milpurn	2617	perth	07 4038 6576	8969 27422 6449 2261	88927.6	1	79.958
rec-0007-org	zara	milburn	2617	perth	07 4038 6576	8969 2742 6449 2261	88927.6	17	79.958
rec-0008-org	jayme	green	2611	hobart	03 6577 2510	4775 1701 0496 2521	91641.4	77	74.075
rec-0009-dup-0	tarai	geaahman	2903	melbourne		9486 7887 96642 2765	50580.7	20	79.994
rec-0009-org	tara	gearman	2903	melbourne	04 2467 6728	9486 7887 9662 2765	50580.7	20	79.994
rec-0010-org	fyynlay	wintulich	2615	perth	04 6989 2785	7760 2452 8483 6380	100385	91	77.557
rec-0011-org	alice	kromwyk	2611	brisbane	04 5904 2482	1789 9728 9016 8668	92654.4	71	74.344
rec-0012-org	brianna	babos	2614	perth	04 6797 0391	3649 3191 6671 6712	37808.4	93	73.689
rec-0013-org	rebekah	allsopp	2602	sydney	03 1275 5290	9537 6230 1583 3827	68169.5	30	73.708
rec-0014-org	jacqueline	coulson	2905	melbourne	07 8597 4089	4051 1899 7546 1013	20225.8	48	73.978
rec-0015-dup-0	andi	robjon	2913	sydney		8501 0767 4223 2079	18238.8	87	81.04
rec-0015-org	andie	robson	2913	sydney	08 6088 7931	8501 0767 4223 2079	18238.8	81	81.04
rec-0016-org	nicole	hyrycz	2603	sydney	07 4650 1483	5739 5244 7959 5982	78089.1	46	75.987
rec-0017-org	alisa	morrison	2902	perth	07 6025 9648	8587 8974 0031 4023	71138.6	82	69.869
rec-0018-org	alana	wilkey	2913	sydney	07 9668 6564	4078 0845 8373 1196	11825	76	74.639
rec-0019-dup-0	loren	mattews	291w	hobart		5981 5539 4177 3850	37581.9	534	72.113
rec-0019-dup-1	loren	matthews	2o13	hobard		5981 5539 4177 3850	37581.9	3	72.113
rec-0019-org	lauren	matthews	2913	hobart	04 7104 5686	5981 5539 4177 3850	37581.9	34	72.113
rec-0020-dup-0	olfia	kioia	2605	pert	02 9483 7684	6040 2780 7840 2488	94741	65	75.075
rec-0020-dup-1	olivya	kioia	2606	peahth		6040 2780 7840 2488	94741	65	75.075
rec-0020-dup-2	olisia	gioya	2606	pert	02 9483 7684	640 2780 7840 2488	94741	65	75.075


Main2.csv - LibreOffice Calc
File Edit View Insert Format Sheet Data Tools Window Help
Liberation Sans
10
R20
Main2
Find
Search Formatted Display String
Match Case
Sheet 1 of 1
Default
Febrl-NCItest
Main2.csv - LibreOffice...
Tue Jul 19, 06:03

rec-id	first name	last name	zipcode	city	telephone-number	credit-card-number	income	age	blood-pressure
rec-0000-dup-0	fiin	lonkman	2612	melbourbe		1611 5438 1051 027	81647.4	70	75.377
rec-0000-dup-1	finn	lonkman	missing	melbourne	02 1908 1495	1611 5438 1051 0 627	81647.4	30	75.377
rec-0000-dup-2	fin	lonkman	2612	melbourne		1611 5438 1051 027	81647.4	7	75.377
rec-0000-org	finn	longman	2612	melbourne	02 1908 1495	1611 5438 1051 0627	81647.4	70	75.377
rec-0001-dup-0	joahdan	paterson	2996	brisbgne		6341 2521 8138 8074	41174.3	4	70.742
rec-0001-dup-1	jordan	pate4son	2906			6341 2521 813 8074	41174.3	17	70.742
rec-0001-dup-2	jortan	paderson	2906	brisbane		6341 221 8138 8074	41174.3	41	70.742
rec-0001-org	jordan	paterson	2906	brisbane	03 9754 0357	6341 2521 8138 8074	41174.3	14	70.742
rec-0002-org	oscar	rawlings	2912	perth	07 3369 4002	6131 8501 0472 5783	35333.7	37	75.523
rec-0003-org	koula	morrison	2603	perth	03 0751 3760	1021 6755 1408 8361	36172.6	54	79.283
rec-0004-org	ethan	matthews	2615	sydney	04 2182 9920	1616 5958 6461 9182	129904.1	89	79.771
rec-0005-org	tyron	jolly	2617	sydney	02 4090 4191	7714 1617 0090 8303	97739.3	75	83.058
rec-0006-dup-0	isabelle	furtado	2903	hobart		6848 1610 38501 0270	139298.1	18	77.096
rec-0006-org	isabelle	furtado	2902	hobart	08 9081 6383	6848 1610 3851 0270	139298.1	88	77.096
rec-0007-org	molly	alves	2615	melbourne	03 8694 5379	1606 0600 2351 7854	59794.8	106	80.136
rec-0008-dup-0	calep	willink	missing		03 5949 6859	6790 4058 7346 8026	44526.1	5	71.508
rec-0008-dup-1	kaleb	willink	29o6	peryh	03 5949 6859	67904058 7346 8026	44526.1	53	71.508
rec-0008-dup-2	kaleb	wiling	2906	perth		6790 4058 7346 58026	44526.1	3	71.508
rec-0008-org	caleb	willing	2906	perth	03 5949 6859	6790 4058 7346 8026	44526.1	53	71.508
rec-0009-org	kailey	mccarthy	2611	canberra	08 3855 7281	8720 7269 8185 7766	112528.7	41	68.931
rec-0010-dup-0	benjamin	hovman	missing			9385 5719 5887 4793	91839	51	81.803
rec-0010-dup-1	bensamin	hoffman	290r	hobart		9385 5719 58874793	91839	5	81.803
rec-0010-org	benjamin	hoffman	2905	hobart	03 3267 3025	9385 5719 5887 4793	91839	50	81.803
rec-0011-dup-0	zac	mariner-schrenk	2906			9325 9295 642 1076	94058.2	19	74.618
rec-0011-dup-1	zac	mariner-schrenk	29p5	sydny		9325 995 6842 1076	94058.2	1098	74.618
rec-0011-dup-2	jac	mariner-zchrenk	2905	sydndy		9325 9295 68492 1076	94058.2	109	74.618
rec-0011-org	zac	mariner-schrenk	2905	sydney	02 8839 4580	9325 9295 6842 1076	94058.2	109	74.618
rec-0012-org	larissa	purtell	2615	perth	03 6303 9390	0431 5238 9268 2980	168696.9	88	78.409
rec-0013-dup-0	iiogen	campbell	2615	perth		8474 6889 901 2 1842	36498.1	340	74.192
rec-0013-dup-1	imoken	campbel1	2615	perth		8474 6889 012 1842	36498.1	3	74.192
rec-0013-dup-2	imogen	campbelo	missing			8474 6889 69012 1842	36498.1	30	74.192
rec-0013-org	imogen	campbell	2615	perth	04 1109 1903	8474 6889 9012 1842	36498.1	30	74.192
rec-0014-org	mark	kelley	2602	perth	07 3641 4687	2133 9489 6362 5705	206105	63	77.361
rec-0015-dup-0	trey	fitzpatrock	2605	sydney		8915 6407 54241 2556	66725.7	39	74.709
rec-0015-dup-1	tery	fitzpatrick	26o6	sydbey		8915 6407 4241 2556	66725.7	3	74.709
rec-0015-org	trey	fitzpatrick	2606	sydney	02 6883 8514	8915 6407 4241 2556	66725.7	30	74.709
rec-0016-dup-0	sawn	mudy	261u	canbeera	03 0798 3500	571 8146 5435 5519	15697.5	29	82.718
rec-0016-dup-1	snawn	moody	2617	canbe4ra		5717 81465435 5519	15697.5	429	82.718
rec-0016-dup-2	zhawn	moody	2y17	canperra		5717 8146 5435 5519	15697.5	294	82.718
rec-0016-org	shawn	moody	2617	canberra	03 0798 3500	5717 8146 5435 5519	15697.5	29	82.718
rec-0017-dup-0	caitlin	calahan	261q	sydhey	07 6521 7857	5849 5837 6796 8878	96788.7	91	72.754

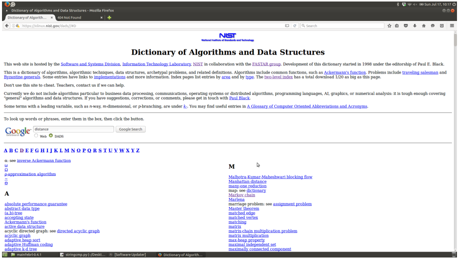

Main3.csv - LibreOffice Calc
File Edit View Insert Format Sheet Data Tools Window Help
Liberation Sans
10
R17
Main3
Find
Search Formatted Display String
Match Case
Sheet 1 of 1
Default
Febrl-NCItest
Main3.csv - LibreOffice...
Tue Jul 19, 06:03

rec-id	given-name	surname	postcode	city	telephone-number	credit-card-number	date-of-birth	social-security-number	income	age	blood-pressure
rec-0000-org	chloe	gearman	2906	sydney	603-612-1775	0375 3987 9965 2775	22-02-1984	188-16-4302	53725.7	38	86.626
rec-0001-org	lara	bowerman	2903	sydney	930-532-8445	0472 3395 9011 5864	16-11-1941	371-19-7053	96500.2	72	78.168
rec-0002-org	alana	white	2905	melbourne	918-679-3144	7489 5383 4657 5912	20-01-1974	266-29-1426	129182.8	37	70.934
rec-0003-dup-0	abbie	keatc	2616	zydney		9200 9709 7304 2030	11-06-1950	539-03-6090	88723	54	73.851
rec-0003-org	abbie	keatch	2617	sydney	843-420-5624	9200 9709 7304 2030	11-06-1950	539-03-6090	88723	54	73.851
rec-0004-dup-0	ijabella	morrison	missing	canberra		2007 7715 1583 5998	16-12-1990	424-39-4227	111138.7	56	70.772
rec-0004-dup-1	isabella	moerison	290y	canbrra		2007 7715 1583 5998	16-12-1990	424-39-4227	111138.7	56	70.772
rec-0004-dup-2	isabslla	moahrison	2906	canberr		2007 7715 1583 5998	16-12-1990	424-39-4227	111138.7	56	70.772
rec-0004-org	isabella	morrison	2906	canberra	512-828-1474	2007 7715 1583 5998	16-12-1990	424-39-4227	111138.7	56	70.772
rec-0005-dup-0	aliza	rid	2914	sydney	774-108-0626	9479 6678 5677 409	19-09-1956	006-79-5954	79827.1	39	73.276
rec-0005-org	alisa	reid	2913	sydney	774-108-0626	9479 6678 5677 4095	19-09-1956	006-79-5954	79827.1	39	73.276
rec-0006-org	megan	walpole	2906	melbourne	209-252-0443	5136 2593 7619 2311	30-05-1954	131-99-1450	44617.5	77	71.408
rec-0007-org	isabella	green	2602	perth	872-468-3479	8608 0557 3087 5366	08-09-2001	575-96-2765	25179.3	26	74.193
rec-0008-org	stephanie	sydenham	2603	brisbane	571-858-4864	7539 6831 0785 0680	29-09-1969	509-28-7760	38446.4	77	70.735
rec-0009-dup-0	paig	zilm	missing	syney		5904 2482 1789 9728	28-11-2004	037-19-4284	10785.4	14	73.777
rec-0009-org	paige	zilm	2611	sydney	607-507-4820	5904 2482 1789 9728	28-11-2004	037-19-4284	10785.4	14	73.777
rec-0010-dup-0	amy	mattews	26o3	melourne	865-629-5156	0498 084 8694 4929	21-10-1968	577-14-6579	46352.7	72	71.924
rec-0010-org	amy	matthews	2603	melbourne	865-629-5156	0498 0584 8694 4929	21-10-1968	577-14-6579	46352.7	72	71.924
rec-0011-dup-0	lilian	humbercross	26p3	brisbane		6260 9528 782 8836	06-05-1971	236-23-7110	45959.8	75	76.727
rec-0011-org	lilian	humberdross	2603	brisbane	623-438-5638	6260 9528 7892 8836	06-05-1971	236-23-7110	45959.8	75	76.727
rec-0012-dup-0	arapella	junderland	2611	pert		3237 3846 2964 0728	13-01-1985	192-22-5258	165760.9	86	75.633
rec-0012-org	arabella	sunderland	2612	perth	818-366-2284	3237 3846 2964 0728	13-01-1985	192-22-5258	165760.9	86	75.633
rec-0013-org	tansy	hinksman	2612	sydney	609-966-8422	1601 4960 1910 5424	22-03-1967	006-13-2571	81925.4	106	70.359
rec-0014-dup-0	sayme	reid	2605			1091 2897 307 2622	22-03-1981	292-08-1308	17725.4	99	74.063
rec-0014-dup-1	jayme	reid	2606	melbou5ne		1091 897 1307 2622	22-03-1981	292-08-1308	17725.4	99	74.063
rec-0014-org	jayme	reid	2605	melbourne	843-061-6862	1091 2897 1307 2622	22-03-1981	292-08-1308	17725.4	99	74.063
rec-0015-org	juliana	dalli	2615	melbourne	251-794-9924	4784 3584 1319 3570	06-08-1944	289-97-4434	15255.8	61	74.781
rec-0016-dup-0	ashton	mkcloskey	26q2	melbourne		19481397 1123 9483	24-08-1958	571-25-9744	94950.3	[illegible]	69.442
rec-0016-org	ashton	mccloskey	2612	melbourne	620-617-5500	1948 1397 1123 9483	24-08-1958	571-25-9744	94950.3	73	69.442
rec-0017-dup-0	kathkeen	grin	2611	canberra		6350 370 6108 2106	08-01-1984	410-50-5252	61434.2	48	70.965
rec-0017-org	kathleen	green	2611	canberra	985-900-2843	6350 5370 6108 2106	08-01-1984	410-50-5252	61434.2	48	70.965
rec-0018-org	makayla	savas	2614	melbourne	512-962-7736	9146 4882 0994 5867	25-06-1990	243-07-1671	82142.7	37	78.503
rec-0019-org	kiera	bocchino	2612	perth	863-087-6766	0055 3706 8062 7293	25-06-1978	221-16-5300	176951	22	72.211
rec-0020-dup-0	paig	mylej	2603	pert		2518 0079 5012 6147	25-11-1950	032-19-9502	44880.9	38	75.815
rec-0020-dup-1	paig	mylest	2604	pe5th		2518 0079 5012 6147	25-11-1950	032-19-9502	44880.9	38	75.815
rec-0020-org	paige	myles	2603	perth	559-978-8253	2518 0079 5012 6147	25-11-1950	032-19-9502	44880.9	38	75.815
rec-0021-dup-0	jamilla	mcmellon	2602	sydnsy		8979 75215 0536 2419	15-09-1940	576-14-7224	45441.7	19	74.27
rec-0021-org	jamilla	mcmellon	2602	sydney	938-228-5841	8979 7215 0536 2419	15-09-1940	576-14-7224	45441.7	14	74.27
rec-0022-dup-0	jessica	le lievre	missing	melbo7rne		8027 5717 2742 6057	05-09-2003	496-63-2434	68921.7	103	78.207
rec-0022-org	jessica	le lievre	2600	melbourne	616-143-9783	8027 5717 2742 6057	05-09-2003	496-63-2434	68921.7	103	78.207
rec-0023-org	emily	george	2905	perth	229-468-5937	0469 4204 5127 0839	22-05-1947	257-80-2188	133628.6	84	75.341

Febrl (for de-duplication and record linkage)- Beyond the Contract

Febrl is a program in Python which needs different packages and features on the computer for its implementation. Its installation is time-consuming and different per PC. It has different versions. I considered Febrl 0.4.1 (Dec 2008). It also has a data generator module called dsgen which is not error free and handy.
It contains a GUI. But we are considering the main Python codes to add some other criteria.
1. Indexing 2. Standardization 3. De-duplication, 4. Record Linkage, 5. Evaluation

Febrl (For Adding Features)


Sun Jul 17, 10:17
Dictionary of Algorithms and Data Structures - Mozilla Firefox
Dictionary of Algorith... ×
404 Not Found ×
https://xlinux.nist.gov/dads//#D
Search
NIST
National Institute of Standards and Technology
Google
Google Search
mainfebrl-0.4.1
stringcmp.py (~/Deskt...
[Software Updater]
Dictionary of Algorith...

Dictionary of Algorithms and Data Structures

This web site is hosted by the Software and Systems Division, Information Technology Laboratory, NIST in collaboration with the FASTAR group. Development of this dictionary started in 1998 under the editorship of Paul E. Black.

This is a dictionary of algorithms, algorithmic techniques, data structures, archetypal problems, and related definitions. Algorithms include common functions, such as Ackermann's function. Problems include traveling salesman and Byzantine generals. Some entries have links to implementations and more information. Index pages list entries by area and by type. The two-level index has a total download 1/20 as big as this page.

Don't use this site to cheat. Teachers, contact us if we can help.

Currently we do not include algorithms particular to business data processing, communications, operating systems or distributed algorithms, programming languages, AI, graphics, or numerical analysis: it is tough enough covering "general" algorithms and data structures. If you have suggestions, corrections, or comments, please get in touch with Paul Black.

Some terms with a leading variable, such as *n*-way, *m*-dimensional, or *p*-branching, are under *k*-. You may find useful entries in A Glossary of Computer Oriented Abbreviations and Acronyms.

To look up words or phrases, enter them in the box, then click the button.

distance

○ Web ● DADS

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

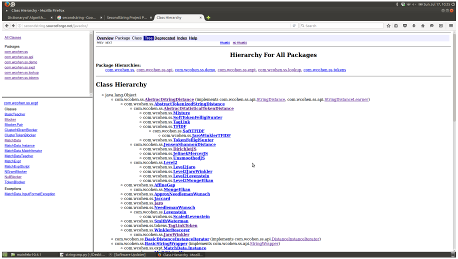
α: see inverse Ackermann function
ω
Ω
ρ-approximation algorithm
~
Θ

A

absolute performance guarantee
abstract data type
(a,b)-tree
accepting state
Ackermann's function
active data structure
acyclic directed graph: see directed acyclic graph
acyclic graph
adaptive heap sort
adaptive Huffman coding
adaptive k-d tree

M

Malhotra-Kumar-Maheshwari blocking flow
Manhattan distance
many-one reduction
map: see dictionary
Markov chain
Marlena
marriage problem: see assignment problem
Master theorem
matched edge
matched vertex
matching
matrix
matrix-chain multiplication problem
matrix multiplication
max-heap property
maximal independent set
maximally connected component


Sun Jul 17, 10:25
Class Hierarchy - Mozilla Firefox
Dictionary of Algorith... ×
secondstring - Goo... ×
SecondString Project P... ×
Class Hierarchy ×
secondstring.sourceforge.net/javadoc/
Search
Overview Package Class Tree Deprecated Index Help
PREV NEXT
FRAMES NO FRAMES
mainfebrl-0.4.1
stringcmp.py (~/Deskt...
[Software Updater]
Class Hierarchy - Mozill...

All Classes

Packages
com.wcohen.ss
com.wcohen.ss.api
com.wcohen.ss.demo
com.wcohen.ss.expt
com.wcohen.ss.lookup
com.wcohen.ss.tokens

com.wcohen.ss.expt
Classes
BasicTeacher
Blocker
Blocker.Pair
ClusterNGramBlocker
ClusterTokenBlocker
MatchData
MatchData.Instance
MatchData.MatchIterator
MatchDataTeacher
MatchExpt
MatchExptScript
NGramBlocker
NullBlocker
TokenBlocker
Exceptions
MatchData.InputFormatException

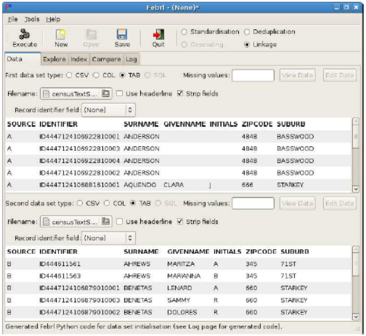
Hierarchy For All Packages

Package Hierarchies:
com.wcohen.ss, com.wcohen.ss.api, com.wcohen.ss.demo, com.wcohen.ss.expt, com.wcohen.ss.lookup, com.wcohen.ss.tokens

Class Hierarchy

- java.lang.Object
  - com.wcohen.ss.**AbstractStringDistance** (implements com.wcohen.ss.api.StringDistance, com.wcohen.ss.api.StringDistanceLearner)
    - com.wcohen.ss.**AbstractTokenizedStringDistance**
      - com.wcohen.ss.**AbstractStatisticalTokenDistance**
        - com.wcohen.ss.**Mixture**
        - com.wcohen.ss.**SoftTokenFelligiSunter**
        - com.wcohen.ss.**TagLink**
        - com.wcohen.ss.**TFIDF**
          - com.wcohen.ss.**SoftTFIDF**
            - com.wcohen.ss.**JaroWinklerTFIDF**
        - com.wcohen.ss.**TokenFelligiSunter**
      - com.wcohen.ss.**JensenShannonDistance**
        - com.wcohen.ss.**DirichletJS**
        - com.wcohen.ss.**JelinekMercerJS**
        - com.wcohen.ss.**UnsmoothedJS**
      - com.wcohen.ss.**Level2**
        - com.wcohen.ss.**Level2Jaro**
        - com.wcohen.ss.**Level2JaroWinkler**
        - com.wcohen.ss.**Level2Levenstein**
        - com.wcohen.ss.**Level2MongeElkan**
    - com.wcohen.ss.**AffineGap**
      - com.wcohen.ss.**MongeElkan**
    - com.wcohen.ss.**ApproxNeedlemanWunsch**
    - com.wcohen.ss.**Jaccard**
    - com.wcohen.ss.**Jaro**
    - com.wcohen.ss.**NeedlemanWunsch**
      - com.wcohen.ss.**Levenstein**
        - com.wcohen.ss.**ScaledLevenstein**
    - com.wcohen.ss.**SmithWaterman**
    - com.wcohen.ss.tokens.**TagLinkToken**
    - com.wcohen.ss.**WinklerRescorer**
      - com.wcohen.ss.**JaroWinkler**
  - com.wcohen.ss.**BasicDistanceInstanceIterator** (implements com.wcohen.ss.api.DistanceInstanceIterator)
  - com.wcohen.ss.**BasicStringWrapper** (implements com.wcohen.ss.api.StringWrapper)
    - com.wcohen.ss.expt.**MatchData.Instance**

Febrl (GUI)



○ Standardisation ○ Deduplication
○ Geocoding ● Linkage

First data set type: ○ CSV ○ COL ● TAB ○ SQL Missing values:

Filename: censusTextS... □ Use headerline ☑ Strip fields

Record identifier field: (None)

SOURCE	IDENTIFIER	SURNAME	GIVENNAME	INITIALS	ZIPCODE	SUBURB
A	ID4447124106922810001	ANDERSON			4848	BASSWOOD
A	ID4447124106922810003	ANDERSON			4848	BASSWOOD
A	ID4447124106922810004	ANDERSON			4848	BASSWOOD
A	ID4447124106922810002	ANDERSON			4848	BASSWOOD
A	ID4447124106881610001	AQUENDO	CLARA	J	666	STARKEY

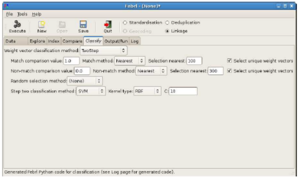
Second data set type: ○ CSV ○ COL ● TAB ○ SQL Missing values:

Filename: censusTextS... □ Use headerline ☑ Strip fields

Record identifier field: (None)

SOURCE	IDENTIFIER	SURNAME	GIVENNAME	INITIALS	ZIPCODE	SUBURB
B	ID444611561	AHREWS	MARITZA	A	345	71ST
B	ID444611563	AHREWS	MARIANNA	B	345	71ST
B	ID4447124106879010001	BENETAS	LENARD	A	660	STARKEY
B	ID4447124106879010003	BENETAS	SAMMY	R	660	STARKEY
B	ID4447124106879010002	BENETAS	DOLORES	R	660	STARKEY

Generated Febrl Python code for data set initialisation (see Log page for generated code).


Febrl - (None)*
File Tools Help
Execute New Open Save Quit
Data Explore Index Compare Classify Output/Run Log

○ Standardisation ○ Deduplication
○ Geocoding ● Linkage

Weight vector classification method: TwoStep

Match comparison value: 1.0 Match method: Nearest Selection nearest: 100 ☑ Select unique weight vectors

Non-match comparison value: 0.0 Non-match method: Nearest Selection nearest: 300 ☑ Select unique weight vectors

Random selection method: (None)

Step two classification method: SVM Kernel type: RBF C: 10

Generated Febrl Python code for classification (see Log page for generated code).

Some Future Steps

- Considering more other specific scenarios for the cancer population
- Availability of real data set or frequency tables for some cancer population
- Febrl applications (Beyond the contract)
- Time and Energy

Thank You!

Questions: smosafer@umd.edu