ENTREGABLES: Entregar vía correo electrónico, el día estipulado como fecha de entrega, lo siguiente:

1. Archivo que contenga las capturas de pantalla mostrando la evidencia de realización de todas y cada una de las actividades, así como de sus correspondientes resultados.

NOTA: Los archivos deben estar adjuntos al correo electrónico, NO deben ser parte del texto del mensaje.

Contexto: MixUp ha conseguido una base de datos de las cincuenta canciones más populares (archivo top50country.csv, archivo con 447 originales, total 1000 registros los demás sin título legible y 553 duplicados). Esta información está actualmente por país y el objetivo es conocer cuáles son las canciones más populares que puedan venderse más rápido en una nueva tienda en Panamá, en donde actualmente no hay tiendas similares, ni se conocen los gustos de la población.

Referencias:

https://help.talend.com/reader/jErhAENS5HA9L8lGuHSmsA/Z56119hIG_M9ffaX3izAIg

https://sourceforge.net/projects/febrl/

https://pypi.org/project/recordlinkage/

https://recordlinkage.readthedocs.io/en/latest/index.html

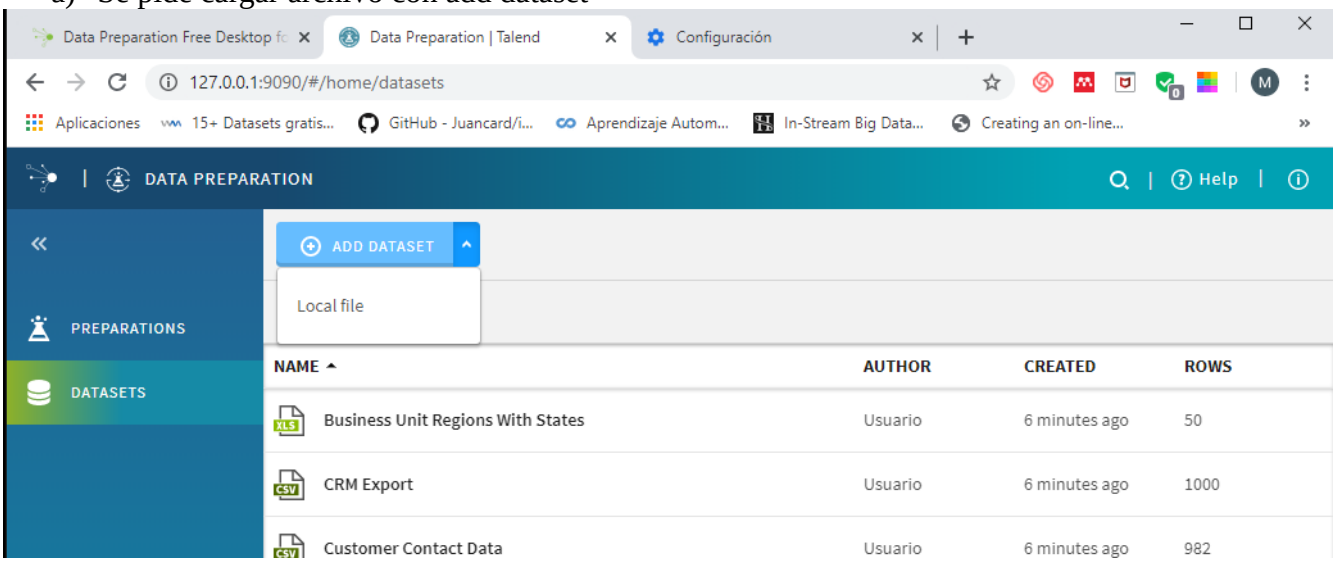
https://recordlinkage.readthedocs.io/en/latest/notebooks/data_deduplication.html

https://readthedocs.org/projects/recordlinkage/downloads/pdf/latest/

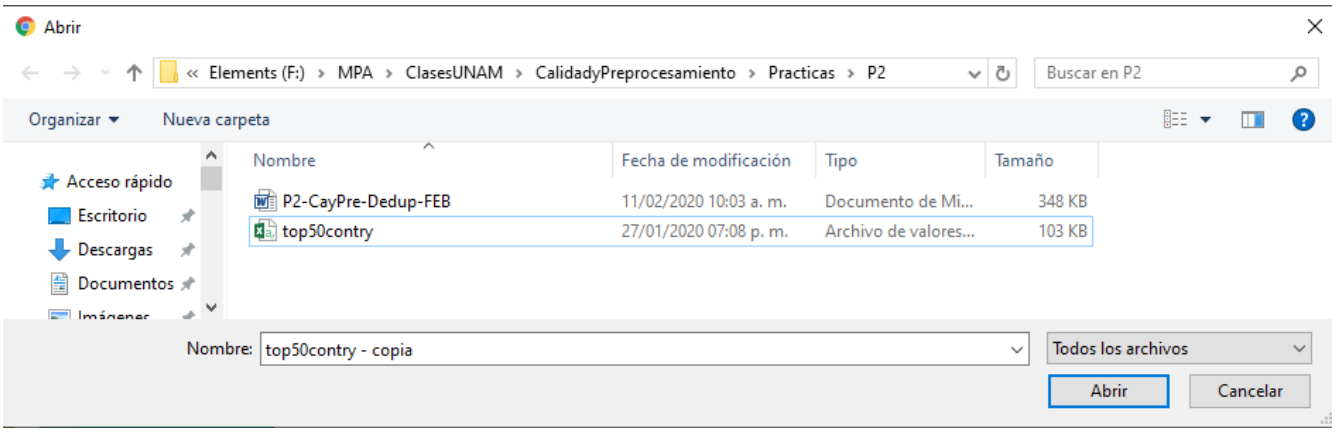
https://pypi.org/project/recordlinkage/#modal-close

Actividad 1: Explorar la información de las canciones e identificar aquellas que estén repetidas y sucias en la lista para posteriormente borrarlas. Utilice Talend Data Preparation para realizar esta actividad.

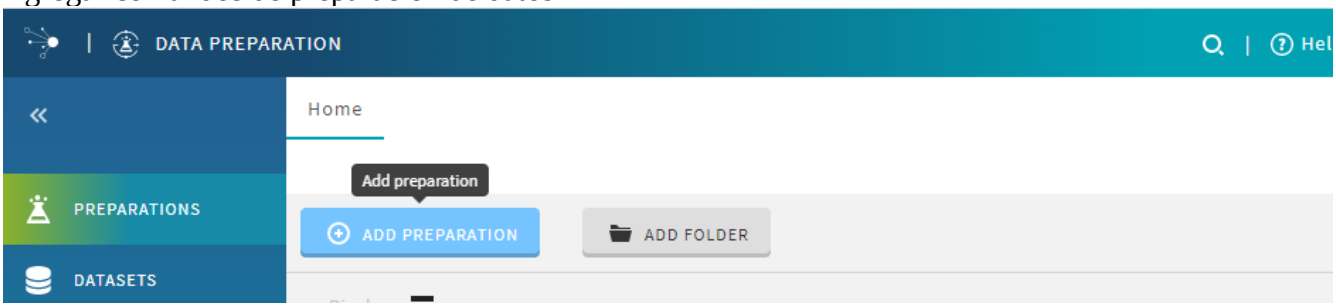
a) Se pide cargar archivo con add dataset



b) Se escoge el archivo top50country.csv



c) Agregar comandos de preparación de datos



d) Escoger el nombre por ejemplo top50countryPreparation

e) A nivel tab, escoger la categoría deduplication y posteriormente Remove duplicate rows

f) ¿Cuántos registros se eliminaron? ¿Por qué?

g) ¿Se realizó la misma actividad solo una vez o varias?

h) ¿Todas las columnas de los registros que se detectaron como duplicados eran iguales o muy similares?

i) A nivel columna titulo escoger find and group similar text

j) ¿Cuántos registros se eliminaron? ¿Por qué?

k) ¿Se realizó la misma actividad solo una vez o varias?

l) ¿Todas las columnas de los registros que se detectaron como duplicados eran iguales o muy similares?

m) ¿Cuál de las dos formas consideras que sea más efectiva?

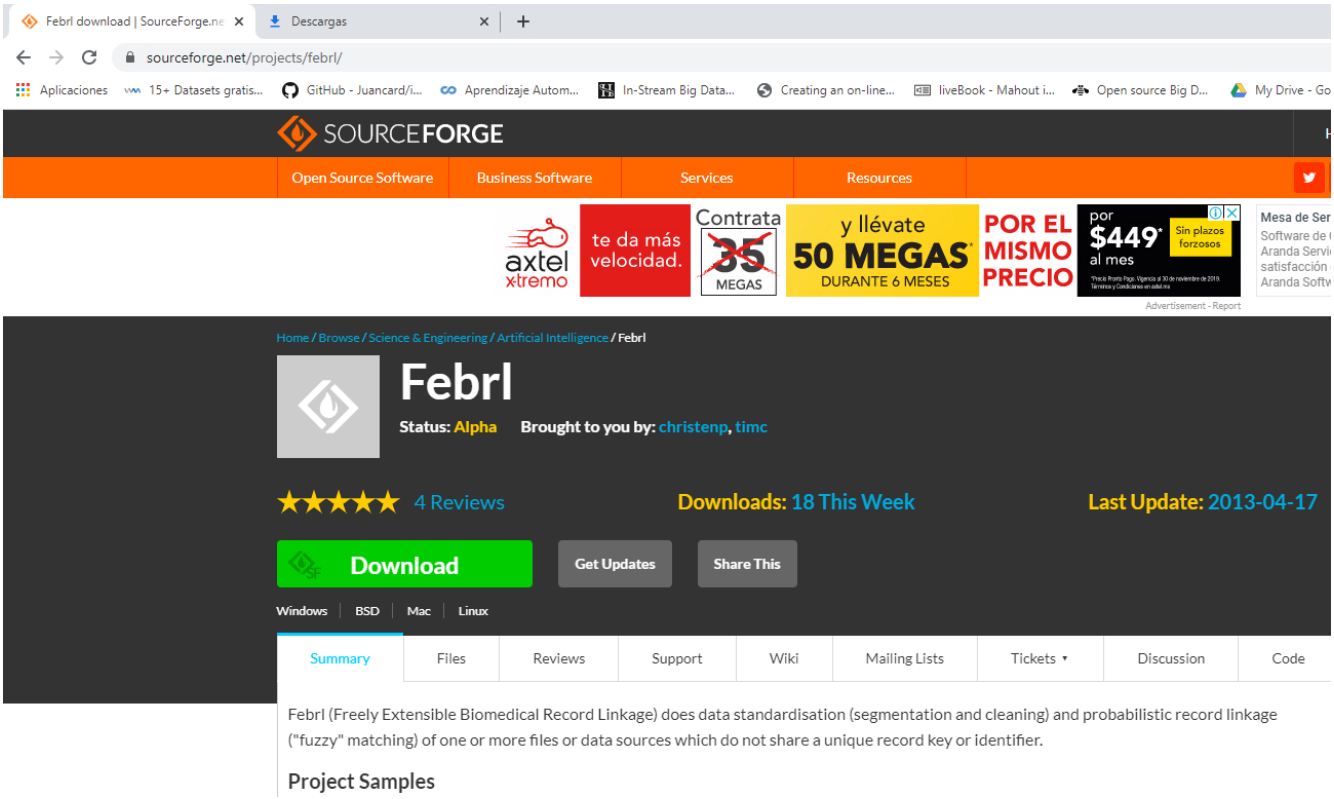
Actividad 2: Instalar aplicación o biblioteca para encontrar duplicados en un archivo de canciones

a) Instalar la aplicación o biblioteca necesaria

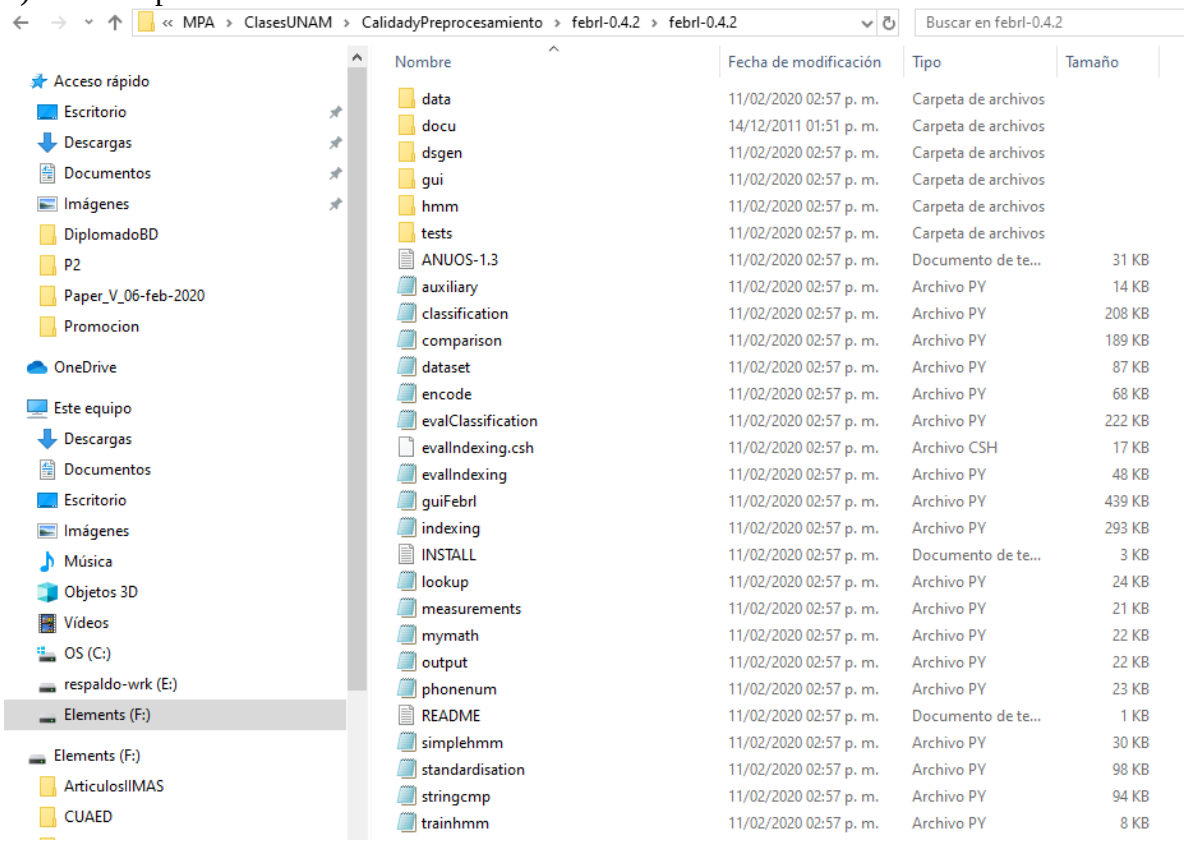
(puede hacer esta actividad con la biblioteca Python recordlinkage, biblioteca inspirada en febrl, pero usa numpy y pandas) o bien Instalar febrl-0.4.2, recordar que febrl corre en Python 2.6 o 2.7

i) En caso de instalar febrl, instalar Python 2.6/2.7 y entrar a la dirección

https://sourceforge.net/projects/febrl/https://sourceforge.net/projects/febrl/



a) Dar click en download
b) Descomprimir el archivo



ii) Leer los archivos INSTALL y README
iii) Probar que las bibliotecas requeridas se encuentren disponibles como lo indica INSTALL
Si no se encuentran entonces instalarlas con

```
pip install pygtk
pip install gtk
pip install svm
```

Es posible que se necesite instalar pygobject con pip install pygobject o conda install C:\Users\Usuario\Downloads/pygobject-3.30.5-py38h5e4a255_0.tar.bz2

c) Instalar la biblioteca recordlinkage en la versión que indica la referencia pip installrecordlinkage

Actividad 3: Encontrar duplicados en un archivo de canciones

a) Tome como entrada el archivo top50countryDos.csv (archivo duplicado con 447 originales y 29 duplicados total 476)
b) Leer el archivo y cargarlo como DataFrame
c) Recuerde que opcionalmente puede preprocesar codificando por soundex
d) Identifique que campos pueden ser relevantes para indexar, escoja su método de indexado y genere su conjunto de pares a comparar, muestre el tamaño y si desea el contenido
e) Compare los campos que considere relevantes a partir de los pares identificados, recuerde utilizar el comparador que corresponda al tipo de dato.
f) Realice el agrupamiento de aquellos registros que corresponda
g) Muestre el número de registros identificados como correspondientes, no correspondientes

Investigue sobre el programa generate.py de la biblioteca recordlinkage
¿Podría realizar aprendizaje supervisado con estos datos?
¿Qué pasos se necesitaría realizar para generar su modelo y después probarlo?

Actividad 4: Encontrar duplicados en un archivo de canciones

a) Tome como entrada el archivo top50countryUno.csv (archivo sin duplicados con 447 en total)
b) Leer el archivo y cargarlo como DataFrame
c) Recuerde que opcionalmente puede preprocesar codificando por soundex
d) Identifique que campos pueden ser relevantes para indexar, escoja su método de indexado y genere su conjunto de pares a comparar, muestre el tamaño y si desea el contenido
e) Compare los campos que considere relevantes a partir de los pares identificados, recuerde utilizar el comparador que corresponda al tipo de dato.
f) Realice el agrupamiento de aquellos registros que corresponda
g) Muestre el número de registros identificados como correspondientes, no correspondientes
h) Justifique su respuesta

Actividad 5: Encontrar duplicados en dos archivos de canciones

a) Tome como entrada los archivos top50countryDos.csv y top50countryTres.csv (archivo con 447 originales duplicados 29 (canciones con T) y 39 (canciones con B) total de registros 515 y 68 duplicados)
b) Leer cada archivo y cargarlo como DataFrame
c) Recuerde que opcionalmente puede pre-procesar codificando por soundex
d) Identifique que campos pueden ser relevantes para indexar y escoja su método de indexado) y genere su conjunto de pares a comparar, muestre el tamaño y si desea el contenido
e) Compare los campos que considere relevantes a partir de los pares identificados, recuerde utilizar el comparador que corresponda al tipo de dato.

f) Realice el agrupamiento de aquellos registros que corresponda
g) Muestre el número de registros identificados como correspondientes
h) Justifique su respuesta

Actividad 6: Encontrar duplicados en dos archivos de canciones

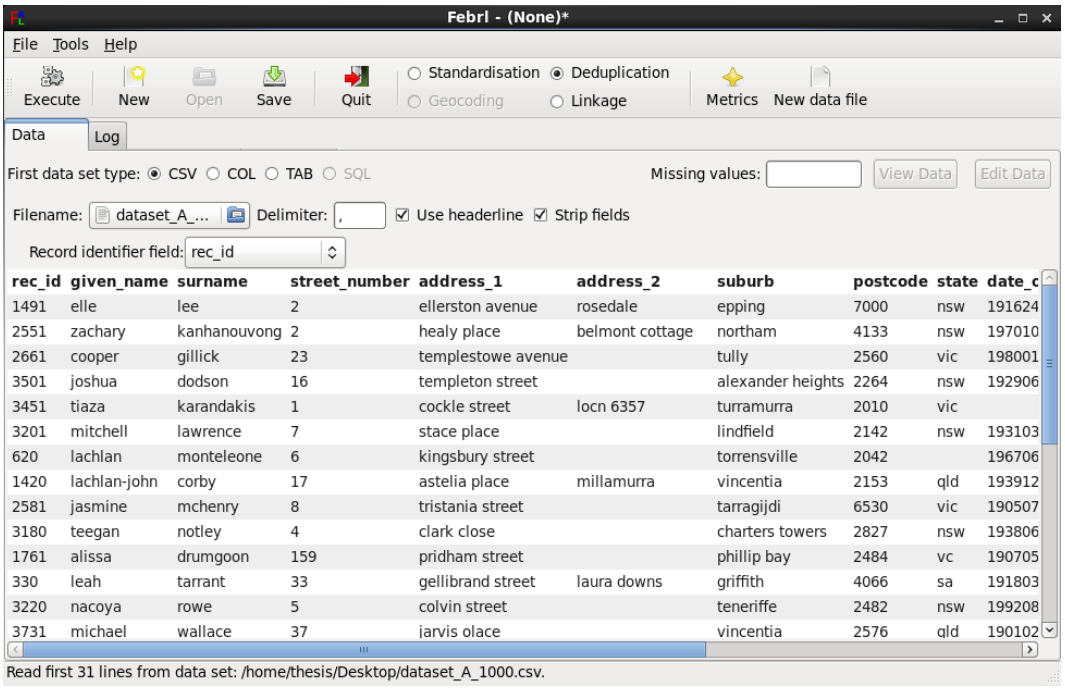
a) Tome como entrada los archivos top50countryUno.csv y top50countryTres.csv
b) Leer cada archivo y cargarlo como DataFrame
c) Recuerde que opcionalmente puede pre-procesar codificando por soundex
d) Identifique que campos pueden ser relevantes para indexar y escoja su método de indexado) y genere su conjunto de pares a comparar, muestre el tamaño y si desea el contenido
e) Compare los campos que considere relevantes a partir de los pares identificados, recuerde utilizar el comparador que corresponda al tipo de dato.
f) Realice el agrupamiento de aquellos registros que corresponda
g) Muestre el número de registros identificados como correspondientes
h) Justifique su respuesta

Actividad 6 OPCIONAL: Genere un archivo de nombre input.csv de registros duplicados con el programa correspondiente de la biblioteca recordlinkage. Considere las siguientes características para el archivo a generar.

Número de registros totales	2000
Número de registros originales	1800
Número de registros duplicados	200
Número máximo de registros duplicados para un registro original que posea duplicados	3
Número máximo de diferencias entre el registro original y el registro duplicado por atributo	2
Número máximo de diferencias entre el registro original y el registro duplicado	8
Atributos de los datos	“rec_id”, “given_name”, “surname”, “street_number”, “address_1”, “address_2”, “suburb”, “postcode”, “state”, “date_of_birth”, “age”, “phone_number”, “soc_sec_id”

Características de los datos de entrada para la de-duplicación

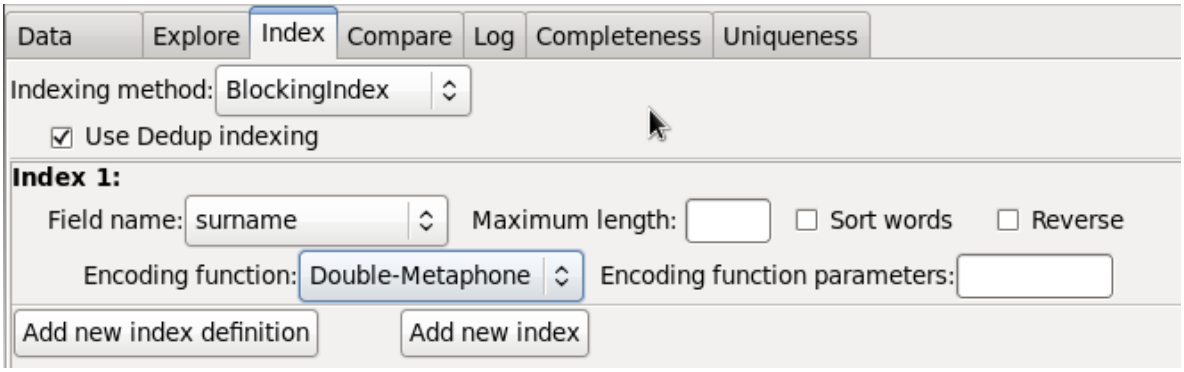
Actividad 7 EXTRA-OPCIONAL SI INSTALA FEBRL O CON LA BIBLIOTECA recordlinkage: Arranque la aplicación FEBRL y cargue el archivo posicionándose en la ruta donde se generó input.csv. Y realice los pasos que se indican con las imágenes con FEBRL o la biblioteca



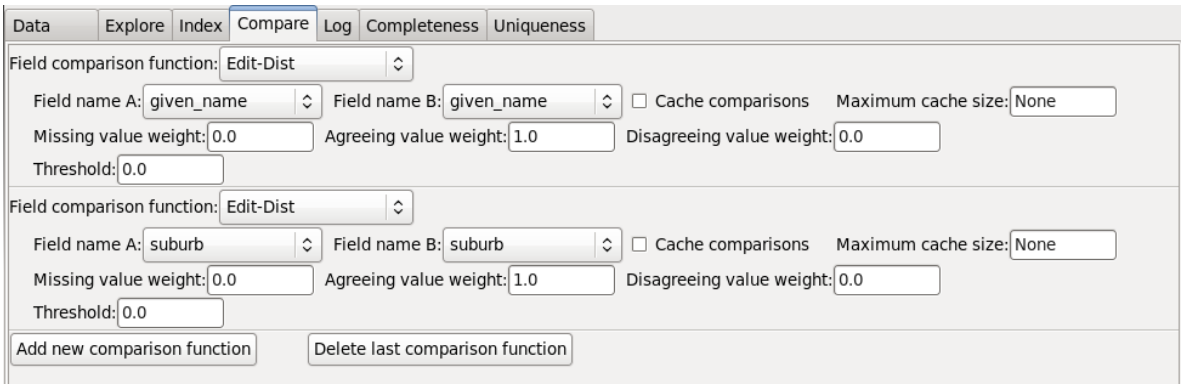
rec_id	given_name	surname	street_number	address_1	address_2	suburb	postcode	state	date_c
1491	elle	lee	2	ellerston avenue	rosedale	epping	7000	nsw	191624
2551	zachary	kanhanouvong	2	healy place	belmont cottage	northam	4133	nsw	197010
2661	cooper	gillick	23	templestowe avenue		tully	2560	vic	198001
3501	joshua	dodson	16	templeton street		alexander heights	2264	nsw	192906
3451	tiaza	karandakis	1	cockle street	locn 6357	turramurra	2010	vic	
3201	mitchell	lawrence	7	stace place		lindfield	2142	nsw	193103
620	lachlan	monteleone	6	kingsbury street		torrensville	2042		196706
1420	lachlan-john	corby	17	astelia place	millamurra	vincentia	2153	qld	193912
2581	jasmine	mchenry	8	tristania street		tarragijdi	6530	vic	190507
3180	teegan	notley	4	clark close		charters towers	2827	nsw	193806
1761	alissa	drumgoon	159	pridham street		phillip bay	2484	vc	190705
330	leah	tarrant	33	gellibrand street	laura downs	griffith	4066	sa	191803
3220	nacoya	rowe	5	colvin street		teneriffe	2482	nsw	199208
3731	michael	wallace	37	jarvis place		vincentia	2576	qld	190102

Actividad -Explore: Visualice el archivo generado y obtenga su estadística básica

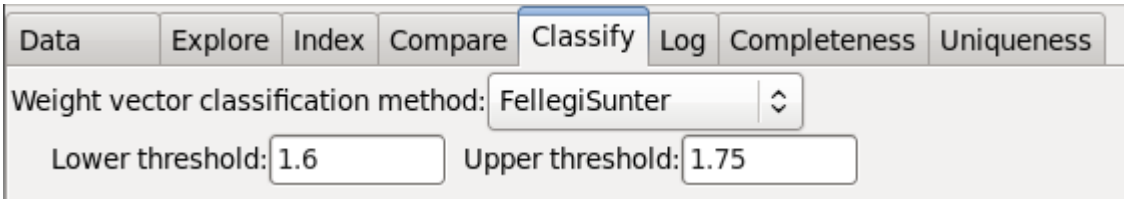
Actividad -Index: La siguiente actividad a realizar es la selección de atributos y métodos para la fase de indexado del proceso de Vinculación de Registros. La siguiente figura muestra la selección del método de indexado "Double Metaphone" a ejecutarse sobre el atributo "surname" para la realización del indexado por bloques.



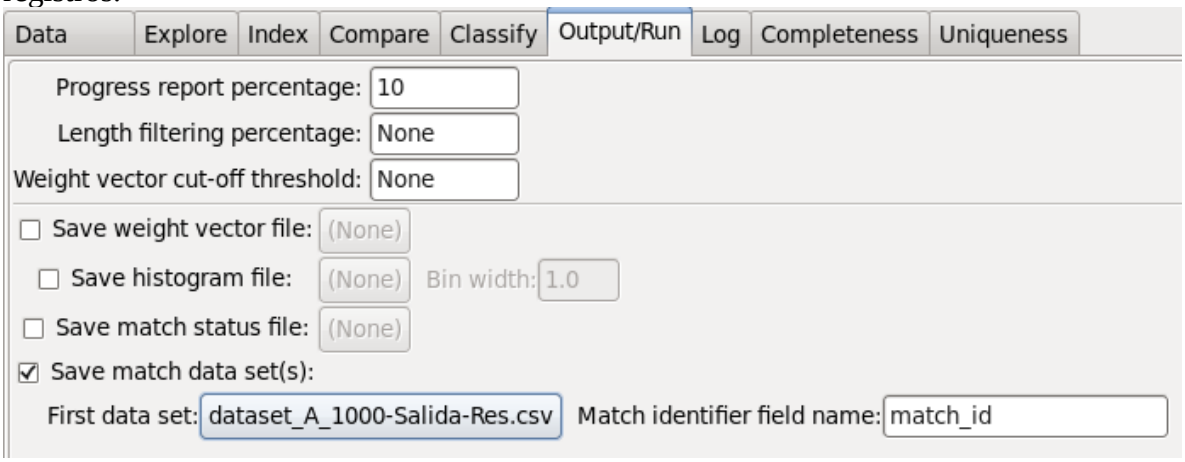
Actividad -Comparación: Posteriormente se accede al menú correspondiente a la fase de comparación para seleccionar los atributos que serán comparados y las funciones que realizarán tales comparaciones. La siguiente figura muestra la selección de los atributos "given_name" y "suburb" para ser comparados mediante la función "Edit-Dist", la cual corresponde la métrica de Similitud por Distancia de Levenshtein.



Actividad -clasificación; Febrl solicita al usuario que seleccione un método de clasificación y los parámetros que dicho método requiera. La Figura siguiente figura muestra la selección del método de clasificación "FellegiSunter" con valores de 1.6 para el umbral inferior y de 1.75 para el umbral superior. El método de clasificación mostrado en Febrl como "FellegiSunter" corresponde al clasificador por umbral de similitud sumada descrito en el Capítulo III del libro Data Matching de Peter Christen.



Actividad -salida: Posteriormente, Febrl permite la selección de la salida del proyecto de Vinculación de Registros en los archivos que el usuario seleccione para este fin. La salida puede ser un archivo que muestre los identificadores de registro junto con su estado de clasificación o el conjunto de datos inicial más una columna que indique su estado final de clasificación. La siguiente Figura muestra la selección del archivo "dataset_A_1000-Salida-Res.csv" para almacenar la salida del proyecto como el conjunto de datos inicial más la adición de un atributo "match_id" que mostrará el estado de clasificación de los registros.



Actividad -Ejecución: Una vez que el proyecto se haya ejecutado, la salida se almacena de acuerdo a la configuración provista por el usuario del sistema. La Figura 4.6 muestra el archivo de salida "dataset_A_1000-Salida-Res.csv". Como se mencionó con anterioridad, tal archivo contiene al conjunto de datos original más la adición del atributo "match_id" que indica el estado de clasificación

de los registros correspondientes; si no hay texto en dicho campo el registro no posee duplicados, en caso contrario el registros es duplicado de aquellos que compartan sus valores de “match_id”.

	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	surname	street_numb	address_1	address_2	suburb	postcode	state	date_of_birt	age	phone_num	soc_sec_id	blocking_nu	match_id	
2	lee	2	ellerston ave	rosedale	epping	7000	nsw	19162429	32	03 51056140	7478073	4	mid077;mid080;mid081	
3	kanhanouvo	2	healy place	belmont cott	northam	4133	nsw	19701031	16	07 65472979	9818217	5	mid231	
4	gillick	23	templestowe	avenue	tully	2560	vic	19800117	22	02 20717061	2796886	0	mid246	
5	dodson	16	templeton	street	alexander he	2264	nsw	19290630	34	08 95104881	9608919	9	mid352	
6	karandakis	1	cockle street	locn 6357	turramurra	2010	vic		29	03 22432781	7453487	1	mid347	
7	lawrence	7	stace place		lindfield	2142	nsw	19310310	33	02 25382252	9164231	6	mid323	
8	monteleone	6	kingsbury	street	torrensville	2042		19670625	36	08 90056511	6517771	0		
9	corby	17	astelia place	millamurra	vincentia	2153	qld	19391228	33	02 48225823	8261357	2	mid067	
10	mchenry	8	tristania	street	tarragijdi	6530	vic	19050718	38	03 93219558	7386221	4	mid238	
11	notley	4	clark close		charters tow	2827	nsw	19380601	35	03 67373058	2516426	1	mid320	
12	drumgoon	159	pridham	street	phillip bay	2484	vc	19070516	25	04 65400025	4003461	1	mid123	
13	tarrant	33	gellibrand st	laura downs	griffith	4066	sa	19180330	25	04 60808817	1585432	1	mid332	
14	rowe	5	colvin	street	teneriffe	2482	nsw	19920825	26	02 73759706	3075307	9	mid325	
15	wallace	37	jarvis	olace	vincentia	2576	qld	19010226	21	02 02951328	4122210	0	mid377;mid380;mid381	
16	hall	19	boldrewoodq	street	ferntree gull	5008	vic	19901220	30	03 50980836	6008124	3	mid493	
17	ludlow	320			baulkham hi	3842	wa	19731225	30	08 05597019	7501257	2	mid248	
18	leifheit	9	tinderry	circuit	millicent	2010	vic	19260812		04 91893541	9767580	2	mid487	
19	thompson	41	roebuck	stre pretoria	atherton	2230	vic	19260607		02 73041390	5179371	9	mid280;mid283;mid284	
20	hoffman	7	challinor	crescent	quinns rocks	3021	nsw	19060618		07 80154535	5425935	2	mid022;mid023	
21	elding	21	marsden	street	rowville	2641	vic	19390504	35	03 98308727	5165520	0	mid130;mid133;mid134	
22	collins	91	dwyer	street	bonny hills	4812	qld	19410719		02 18041653	9006774	7	mid504	
23		12	milford	street	dakabin	2227	vic	19600115	22	04 81631418	4841193	6		
24	van boom	9	birchall stre	springflat	penrith	3150	vic		31	07 55769472	8653551	5	mid509	
25	hookway	8	mountain cr	rsde 261	westleigh	3133	tas	19611101	29	08 23841560	7515588	1	mid107	

Actividad Graficar los resultados de la clasificación

Actividad Cambie algoritmo de comparación para los mismos campos y reflexione si cambia la salida

Actividad Cambie algoritmo de clasificación para los mismos campos y reflexione si cambia la salida