Taller #bdw14. Día 1
Fuentes y limpieza de datos

BIG DATAWEEK
5 - 11 May 2014

https://github.com/hackbdw14/hackatonData

hackbdw14@gmail.com
@hackbdw14

Contenidos

Minería de Datos

- Motivación
- El Pipeline de datos
- Herramientas

Datasets para el hackaton

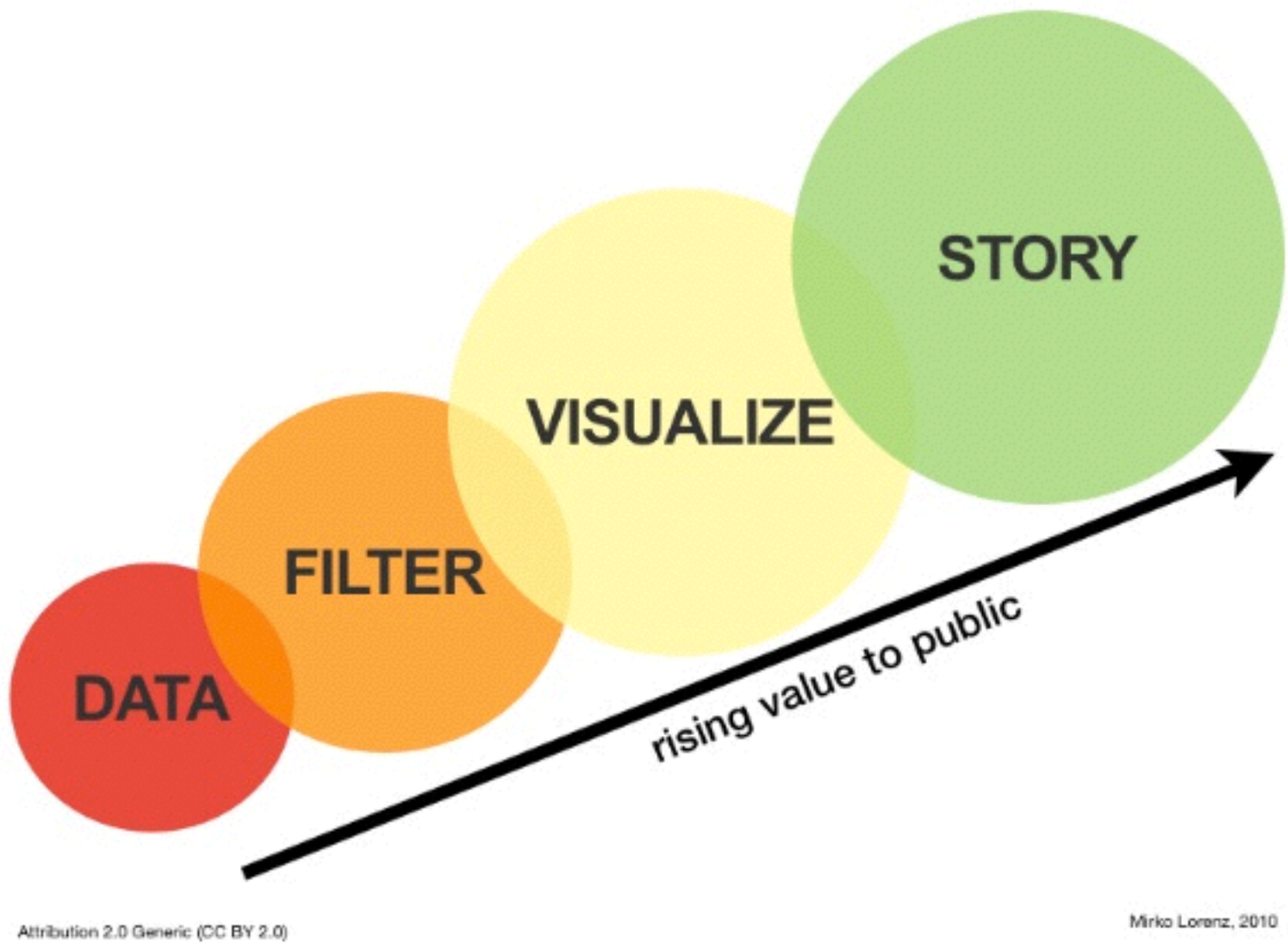
Adquisición de Datos

- Formato de Datos
- Fuentes de Datos
- Scraping: Chrome scraper
- Ejercicio
- Flocker
- PDF

Open Refine

- Introducción
- Ejercicios

Minería de Datos



http://en.wikipedia.org/wiki/Data-driven_journalism

Motivación

Complejidad Cognitiva

Sabiduría

- No transferible
- Aplicación ética del conocimiento

Conocimiento

- Asimilación personal
- Aplicación de la información: Acción
- 'Insights'

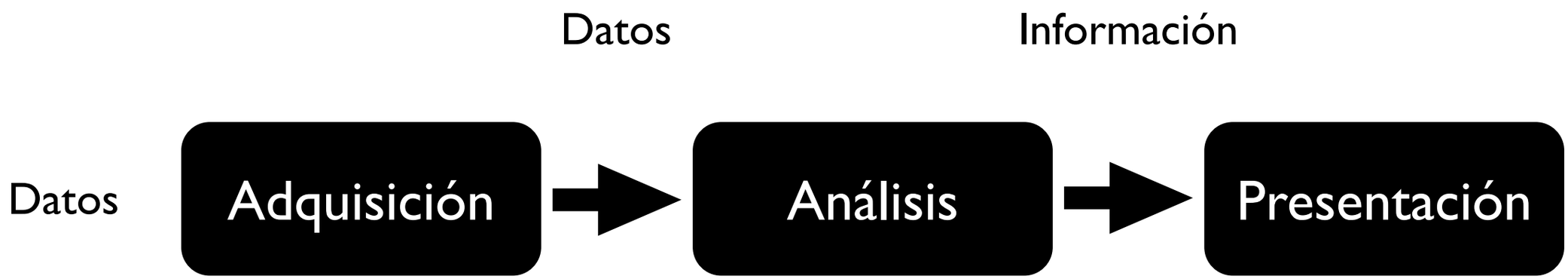
Información

- Datos Estructurados
- Semántica
- Narrativa
- Datos espacializados

Datos

- Registro de Observaciones

El Pipeline de Datos



- API's
- Scraping
- Ficheros
- BD's
- Proveedores

- Minería de Texto
- Aprendizaje automático
- Métricas de Red
- Estadística

- Visualización
- Narrativa

El Pipeline de Datos: Herramientas

Adquisición de Datos

- Chrome Scraper
- Import.io
- Google Docs
- Flocker
- Google Refine

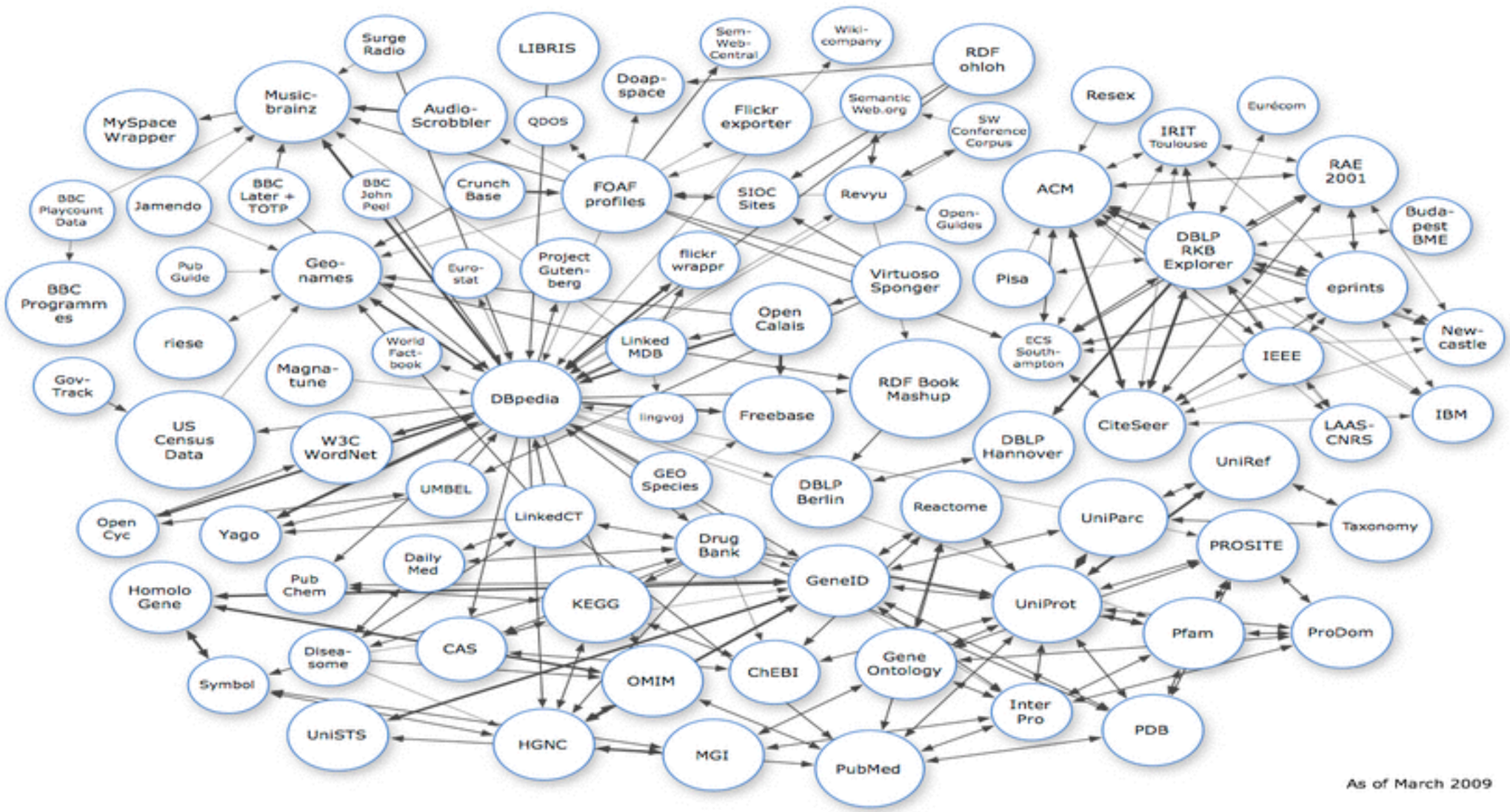
Análisis

- Open Refine
- Google Spreadsheet
- Outliers Data Tools

Visualización de Datos

- CartoDB
- Gephi
- Google Maps + Fusion Tables
- Mapbox
- Google Charts

Adquisición de Datos



Formatos de Datos

- JSON - Dev
- CSV, XLS - Tabular
- XML - Viejo, pero sigue usándose
- Web (HTML5?)

Formatos de Datos

- JSON (Javascript Object Notation) - Dev

```
{
    hey: "guy",
    anumber: 243,
  - anobject: {
        whoa: "nuts",
      - anarray: [
            1,
            2,
            "thr<h1>ee"
        ],
        more: "stuff"
    },
    awesome: true,
    bogus: false,
    meaning: null,
    japanese: "明日がある。",
    link: http://jsonview.com,
    notLink: "http://jsonview.com is great"
}
```

Formatos de Datos

- CSV (Comma separated Value), TSV, etc..

```
"EMPNO","ENAME","JOB","MGR","HIREDATE","SAL","COMM","DEPTNO"
7369,"SMITH","CLERK",7902,17-DEC-80 12.00.00,800,,20
7499,"ALLEN","SALESMAN",7698,20-FEB-81 12.00.00,1600,300,30
7521,"WARD","SALESMAN",7698,22-FEB-81 12.00.00,1250,500,30
7566,"JONES","MANAGER",7839,02-APR-81 12.00.00,2975,,20
7654,"MARTIN","SALESMAN",7698,28-SEP-81 12.00.00,1250,1400,30
7698,"BLAKE","MANAGER",7839,01-MAY-81 12.00.00,2850,,30
7782,"CLARK","MANAGER",7839,09-JUN-81 12.00.00,2450,,10
7788,"SCOTT","ANALYST",7566,19-APR-87 12.00.00,3000,,20
7839,"KING","PRESIDENT",,17-NOV-81 12.00.00,5000,,10
7844,"TURNER","SALESMAN",7698,08-SEP-81 12.00.00,1500,0,30
7876,"ADAMS","CLERK",7788,23-MAY-87 12.00.00,1100,,20
7900,"JAMES","CLERK",7698,03-DEC-81 12.00.00,950,,30
7902,"FORD","ANALYST",7566,03-DEC-81 12.00.00,3000,,20
7934,"MILLER","CLERK",7782,23-JAN-82 12.00.00,1300,,10
```

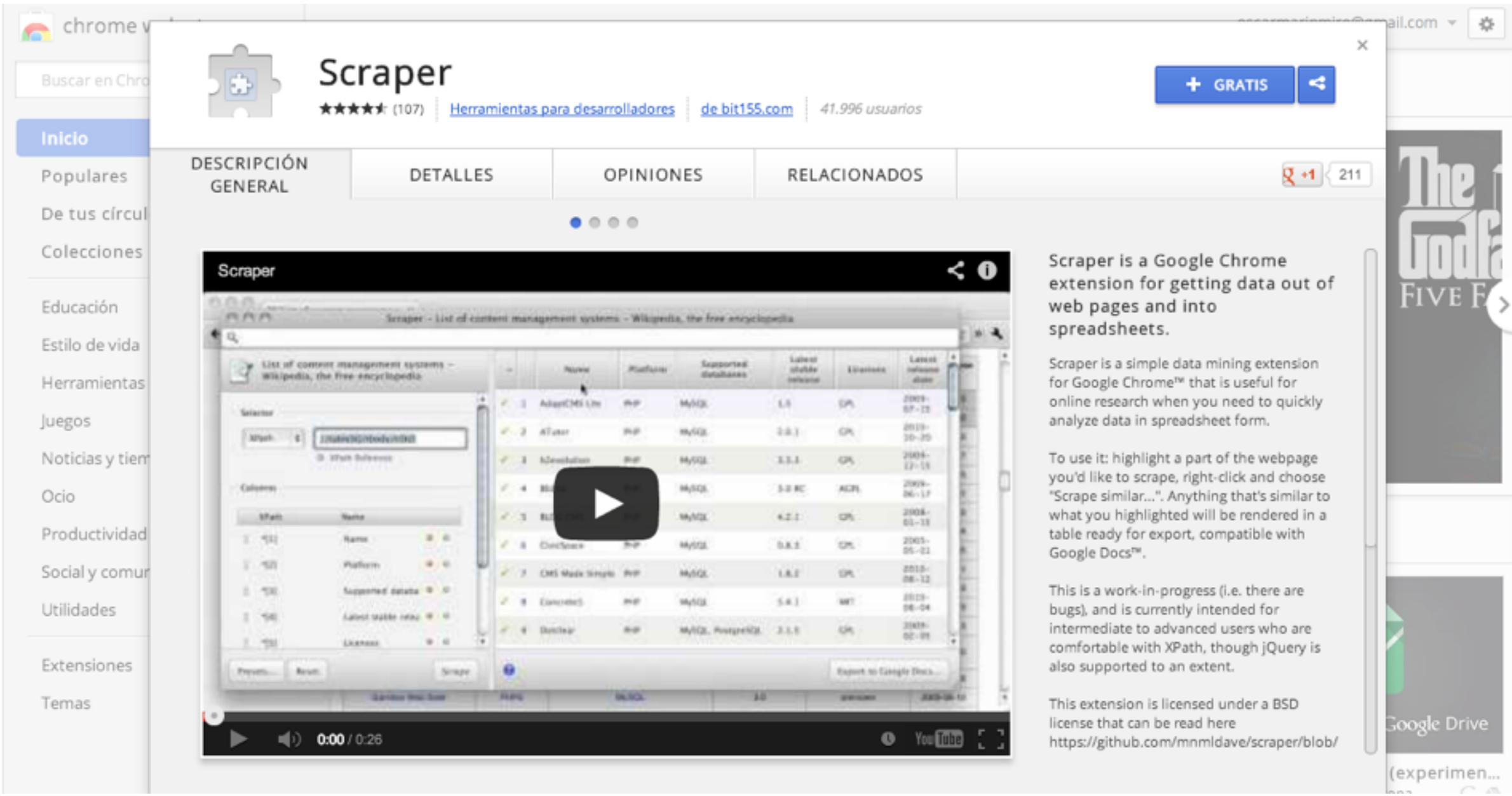
Técnicas de adquisición

- Descarga desde Repositorios de Datos
- Pregunta al usuario: Forms
- Web data: Scrape
- Pregunta al proveedor: API access

Fuentes de Datos sugeridas

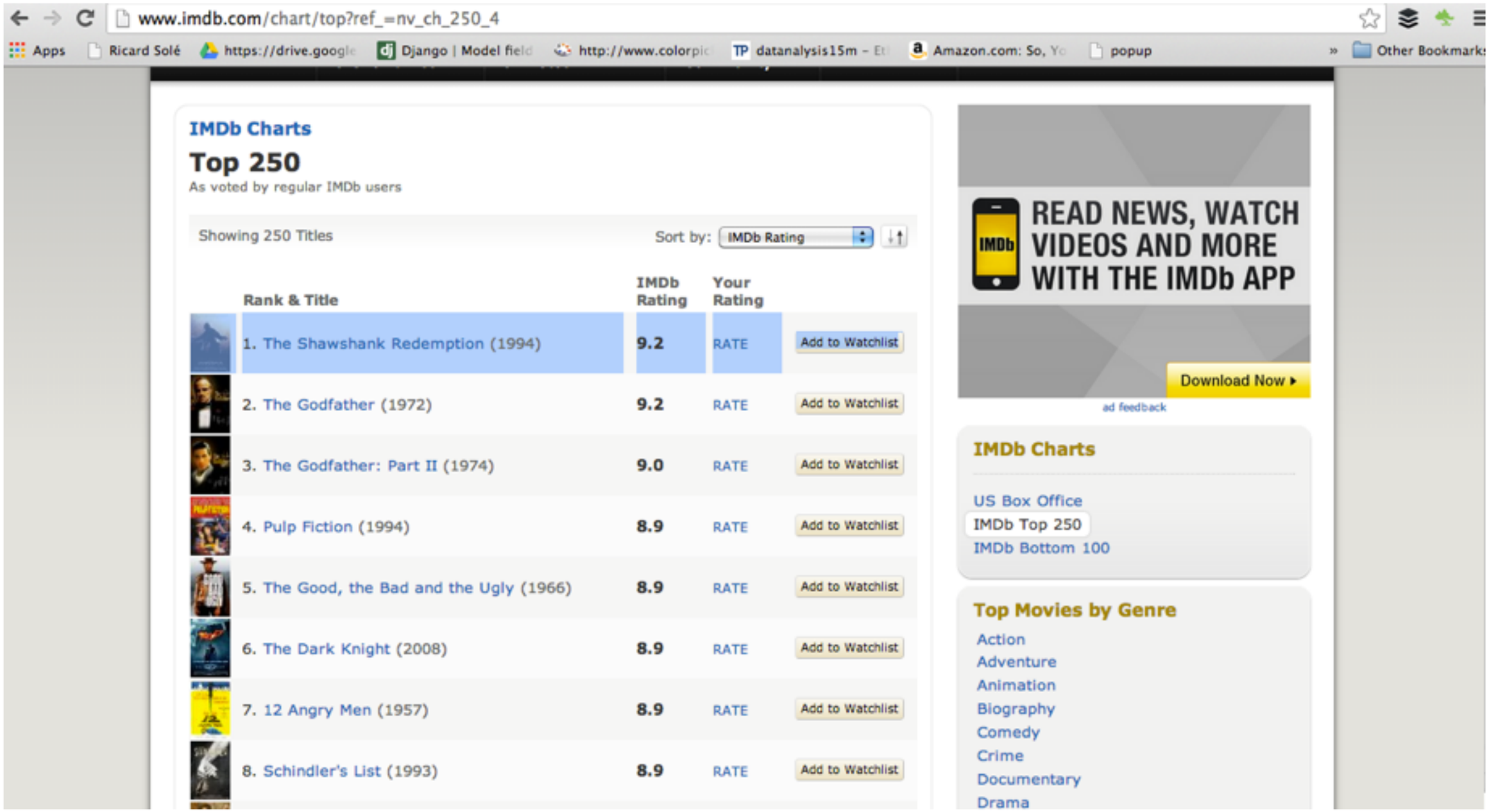
http://epp.eurostat.ec.europa.eu/portal/page/portal/region_cities/city_urban/
http://opendata.bcn.cat/opendata/ca
http://www.bcn.cat/estadistica/catala/index.htm
http://w20.bcn.cat/cartobcn/default.aspx?lang=en

Scrape w/ Chrome Scraper



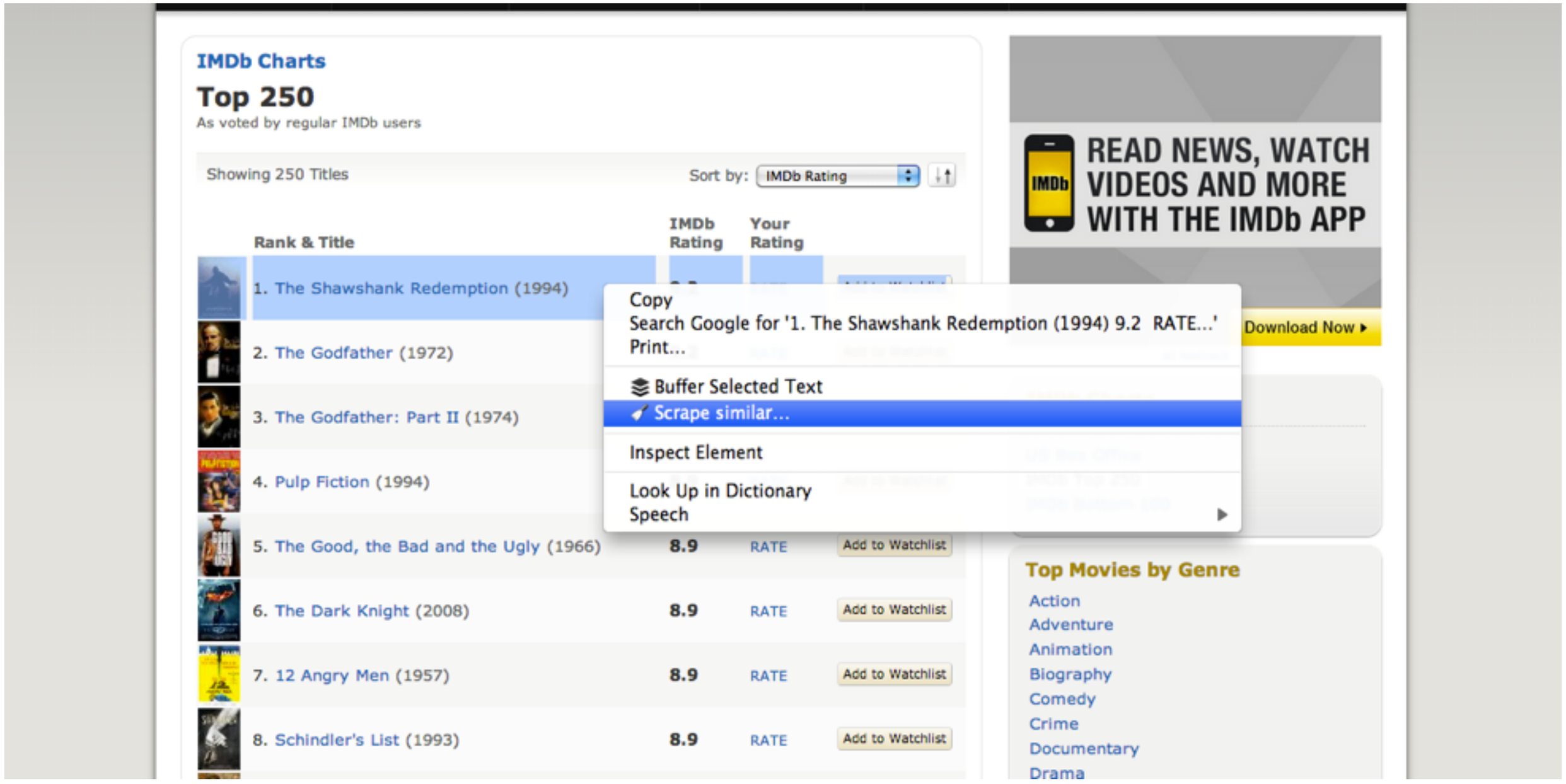
https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffohmbkdlecaccepngjd

Scrape w/ Chrome Scraper

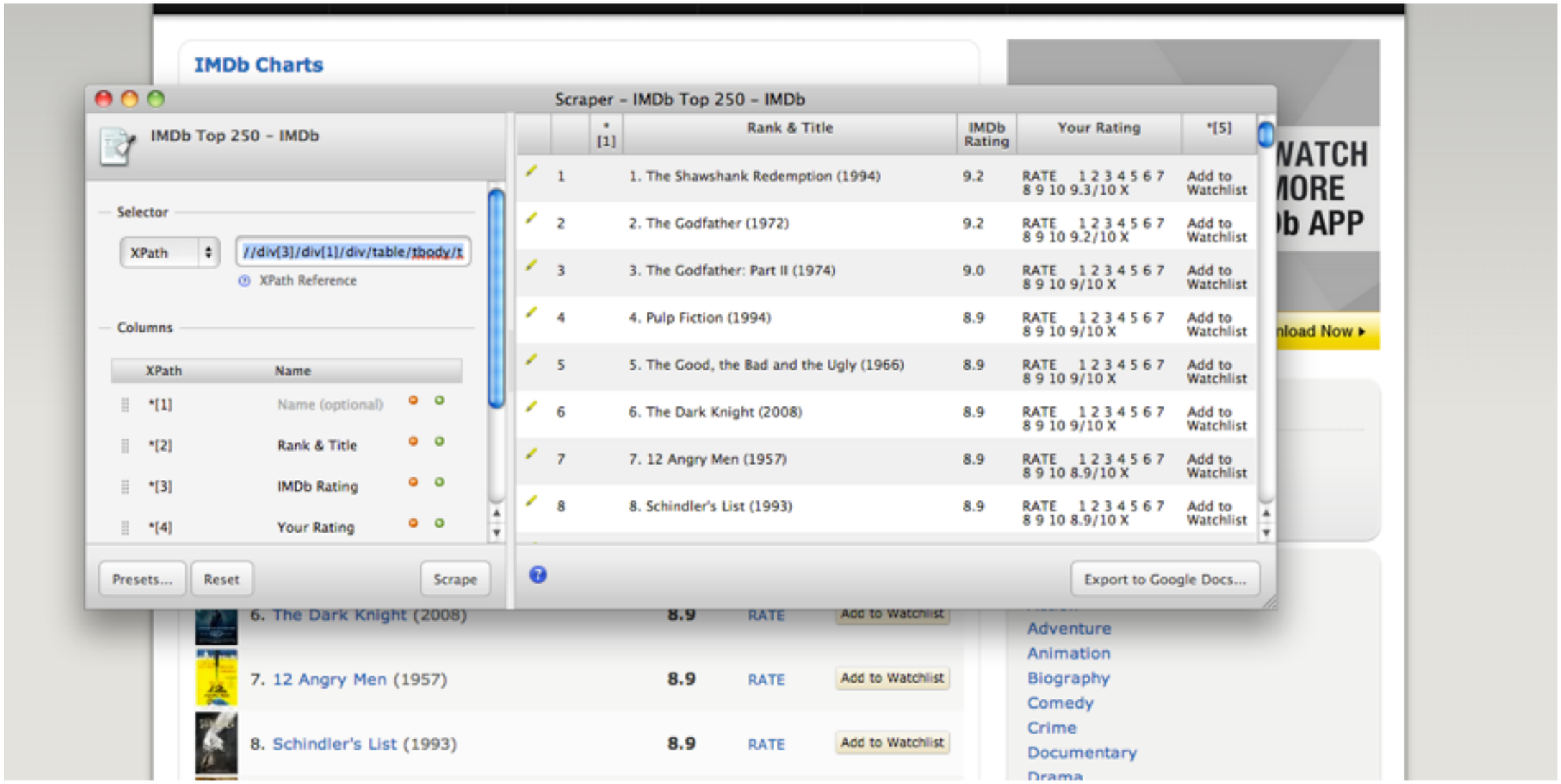


http://www.imdb.com/chart/top?sort=ir,desc

Scrape w/ Chrome Scraper



Scrape w/ Chrome Scraper



http://dataist.wordpress.com/2012/10/12/get-started-with-screenscraping-using-google-chromes-scraper-extension/

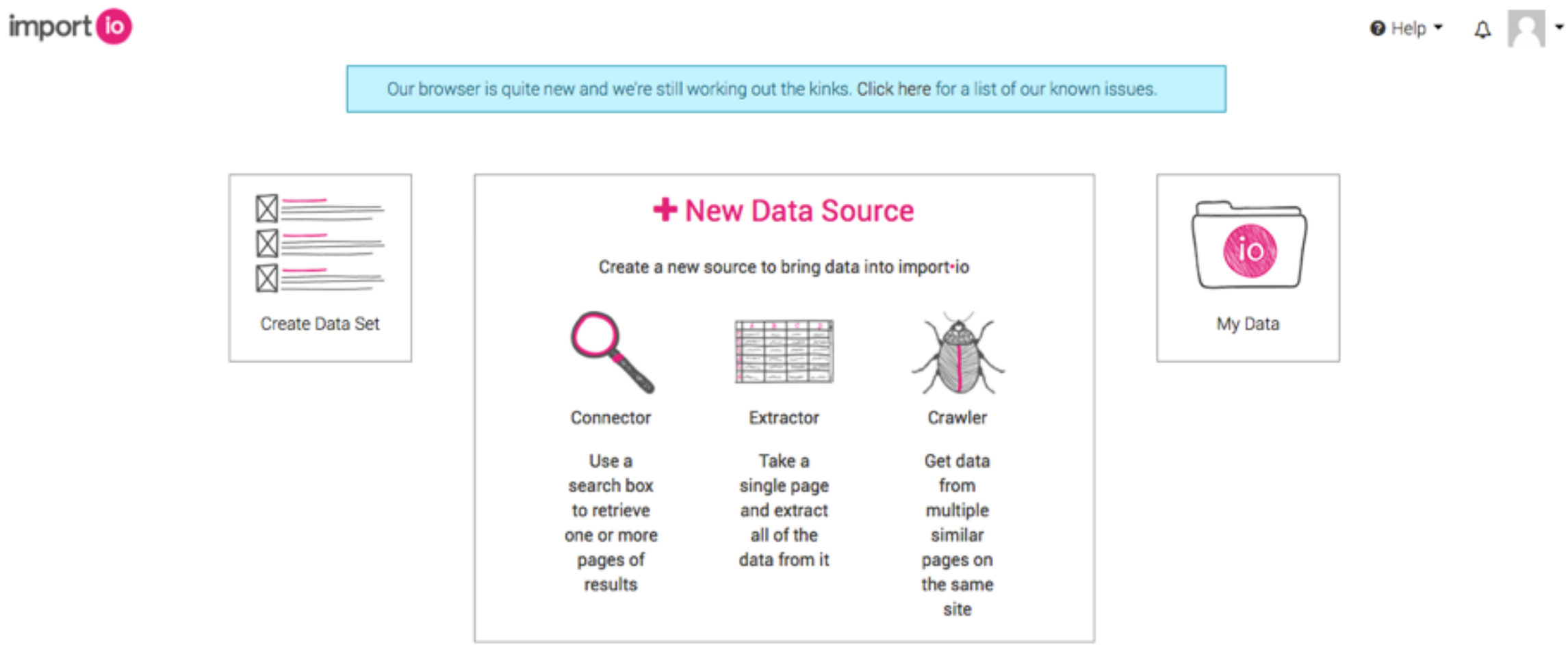
Ejercicio

Scrapear con Chrome Scraper la siguiente url:

http://www.bcn.cat/estadistica/catala/dades/economia/renda/rdfamiliar/a2012/rfbarris.htm

(guardarla como Google Docs y luego limpiarla en la misma spreadsheet)

Scrape w/ import.io

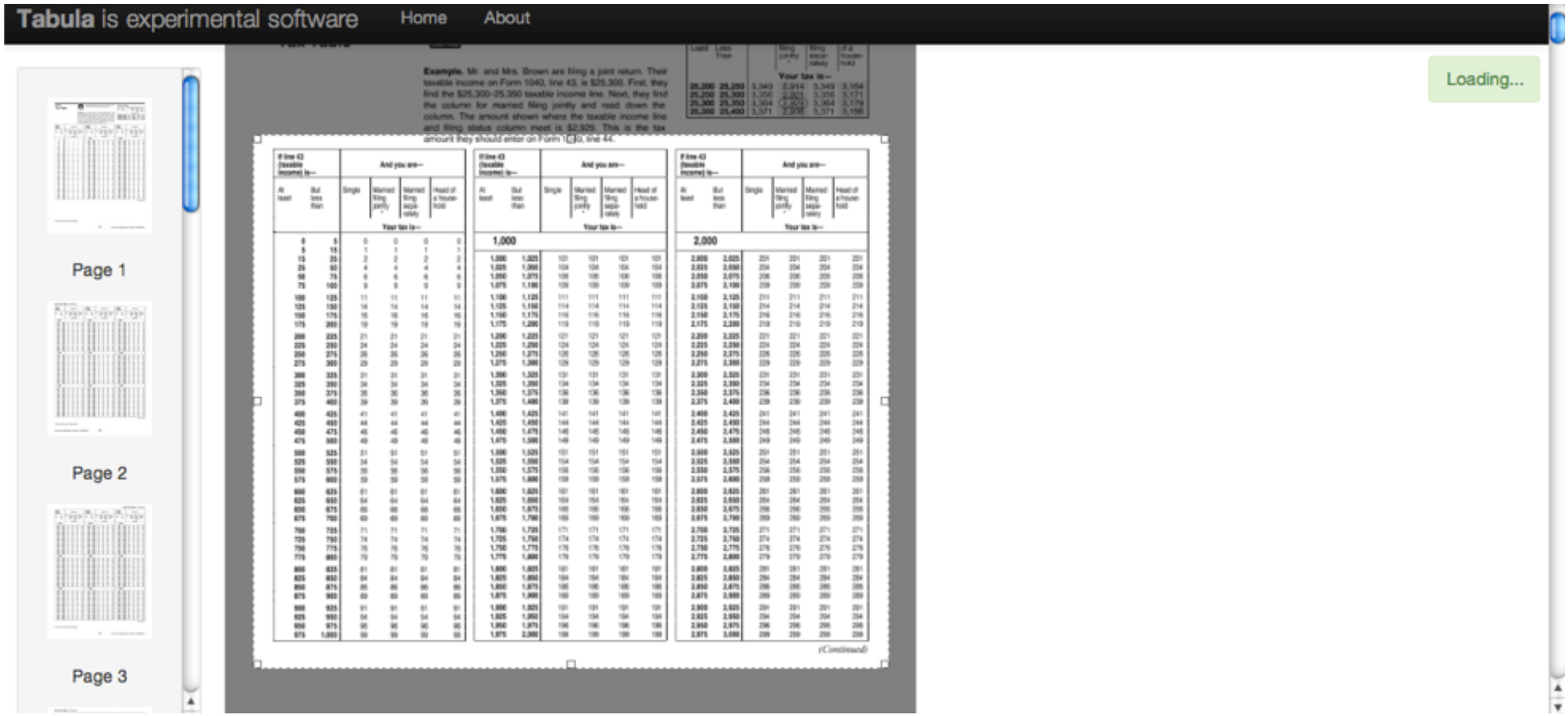


http://import.io/

PDF Extraction: Tabula

http://tabula.nerdpower.org/

PDF Extraction: Tabula



http://tabula.nerdpower.org/

PDF Extraction: Tabula

Extracted tabular data

3,000					
3,0003,050	3,0503,100	303308	303308	303308	303308
3,1003,150	3,1503,200	313318	313318	313318	313318
3,2003,250	3,2503,300	323328	323328	323328	323328
3,3003,350	3,3503,400	333338	333338	333338	333338
3,4003,450	3,4503,500	343348	343348	343348	343348
3,5003,550	3,5503,600	353358	353358	353358	353358
3,6003,650	3,6503,700	363368	363368	363368	363368
3,7003,750	3,7503,800	373378	373378	373378	373378
3,8003,850	3,8503,900	383388	383388	383388	383388
3,900	3,950	393	393	393	393

Use row/columns separators

Close | Copy to clipboard as CSV | Download data

http://tabula.nerdpower.org/

Network Data: Flocker

FLOCKER A Twitter real-time monitor

What?

FLOCKER is a Twitter real-time retweets networks builder.

Why?

Twitter is nowadays the fastest way to access and spread information. There are tools and services offering the possibility to monitor Twitter's stream. There are also tools offering the possibility to build networks based on retweets and mentions from a given dataset. But we haven't found any tool combining both functionalities (except Gephi's plugin Retweet Monitor).

Some of us worked in the menctioned plugin for Gephi and abandoned it. Gephi, although very useful and complete, is a complicate tool for both users and developers. Based on our experience we are trying to provide FLOCKER with the features most requested/used in Gephi by people analyzing Twitter.

Who?

FLOCKER is a project developed by Outliers.

Current status

Currently, FLOCKER is under development. At this moment you can:

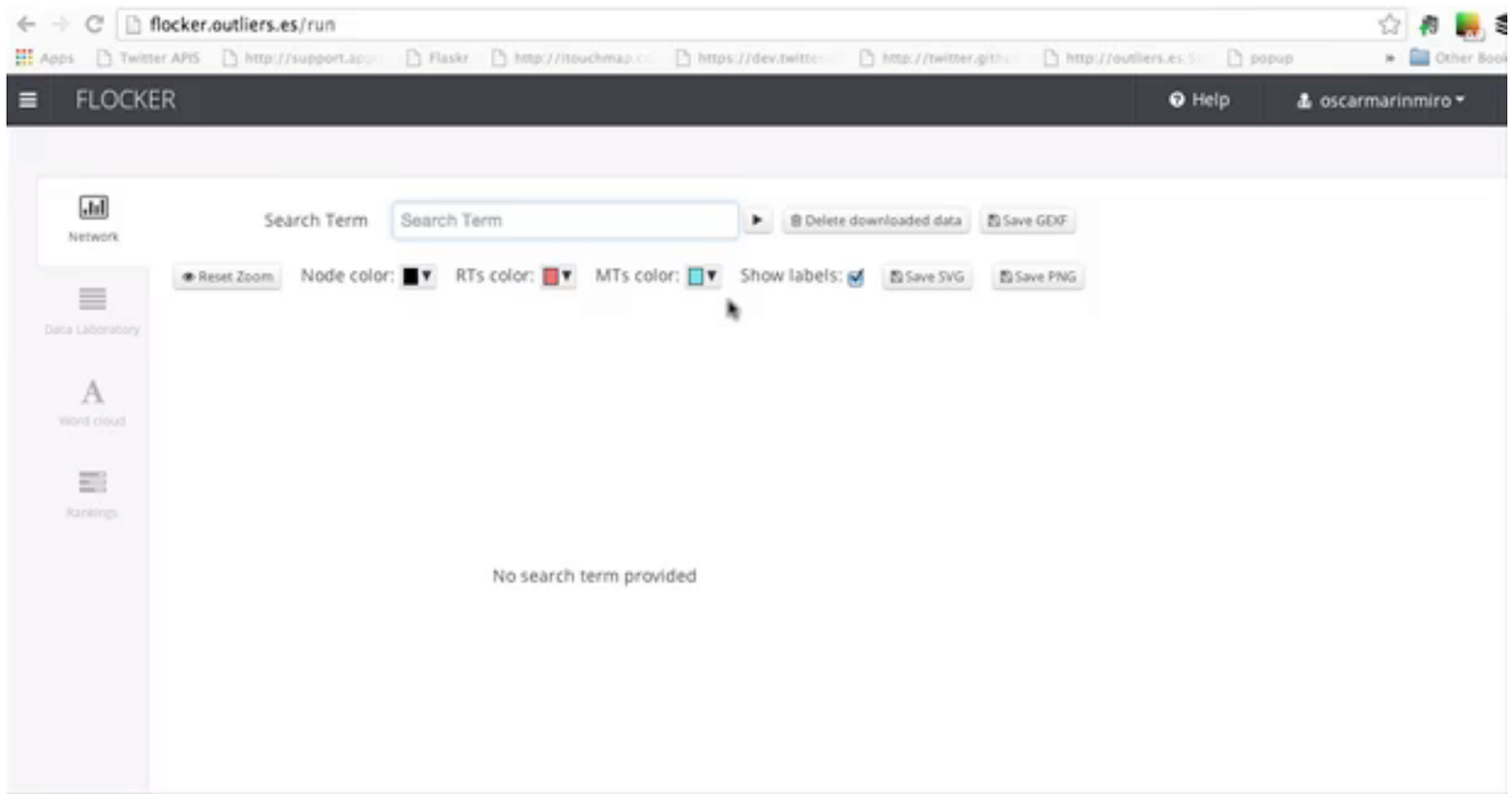
- Login using your Twitter's account
- Filter the stream using terms, hashtags or Twitter's usernames
- See how the retweets network is dinamically built
- Explore the data using the *data laboratory*
- Change the colors used to display nodes and edges
- Export the generated graph as GEXF
- Export the generated graph as PNG
- Export the generated graph as SVG

The chart on the right shows the percentage of features we have currently developed.

Start using FLOCKER!

http://flocker.outliers.es/

Network Data: Flocker



http://flocker.outliers.es/

Datasets para el hackaton

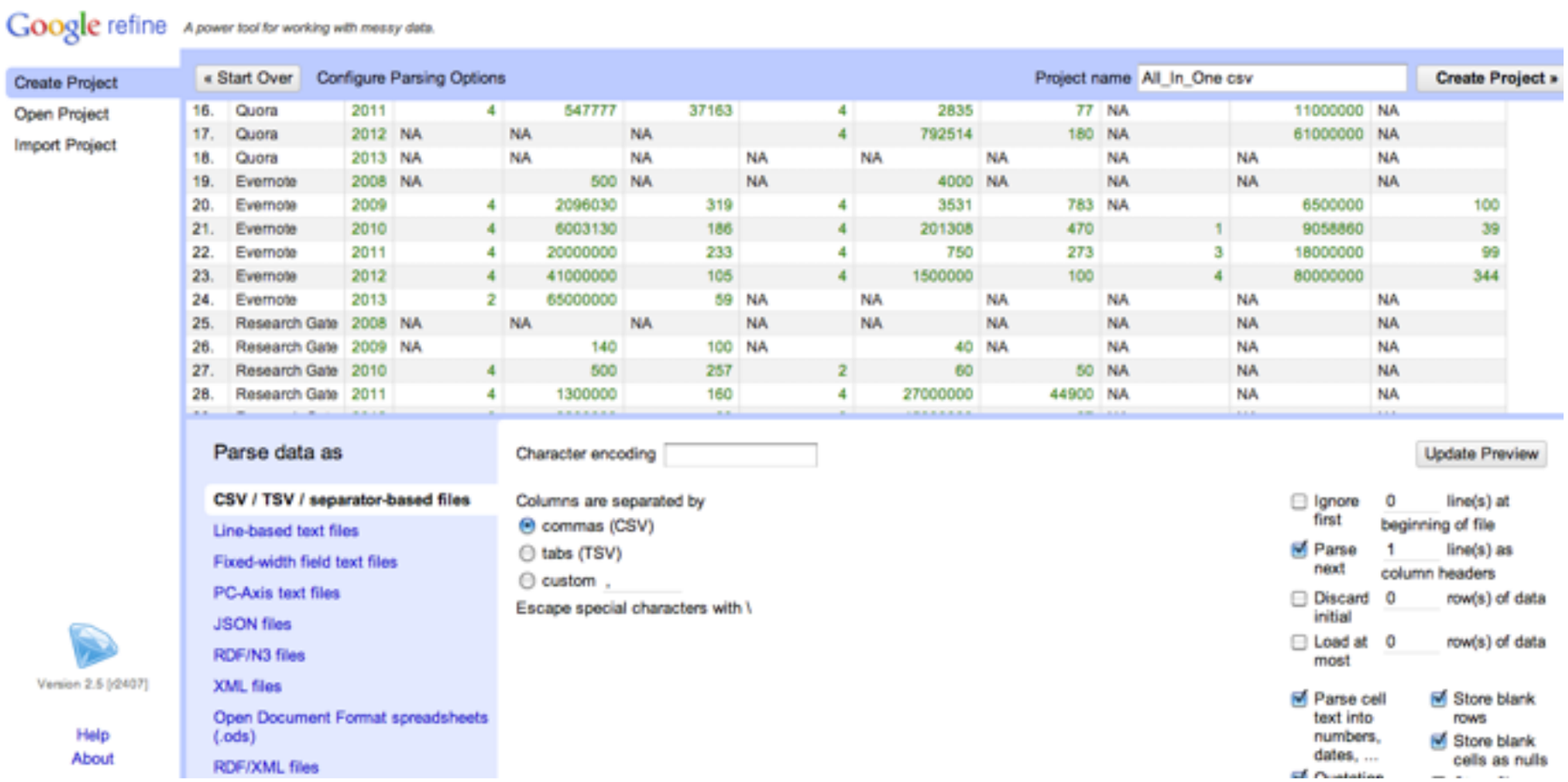
https://github.com/hackbdw14/hackatonData/tree/master/datasets

Open Refine: Cleaning Data

https://github.com/OpenRefine

https://code.google.com/p/google-refine/downloads/list

Open Refine: Cleaning Data



https://github.com/OpenRefine

https://code.google.com/p/google-refine/downloads/list

Open Refine: Cleaning Data

Google refine All_In_One csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 0

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
Watch these screencasts

426 rows

Show as: **rows** records Show: 5 **10** 25 50 rows

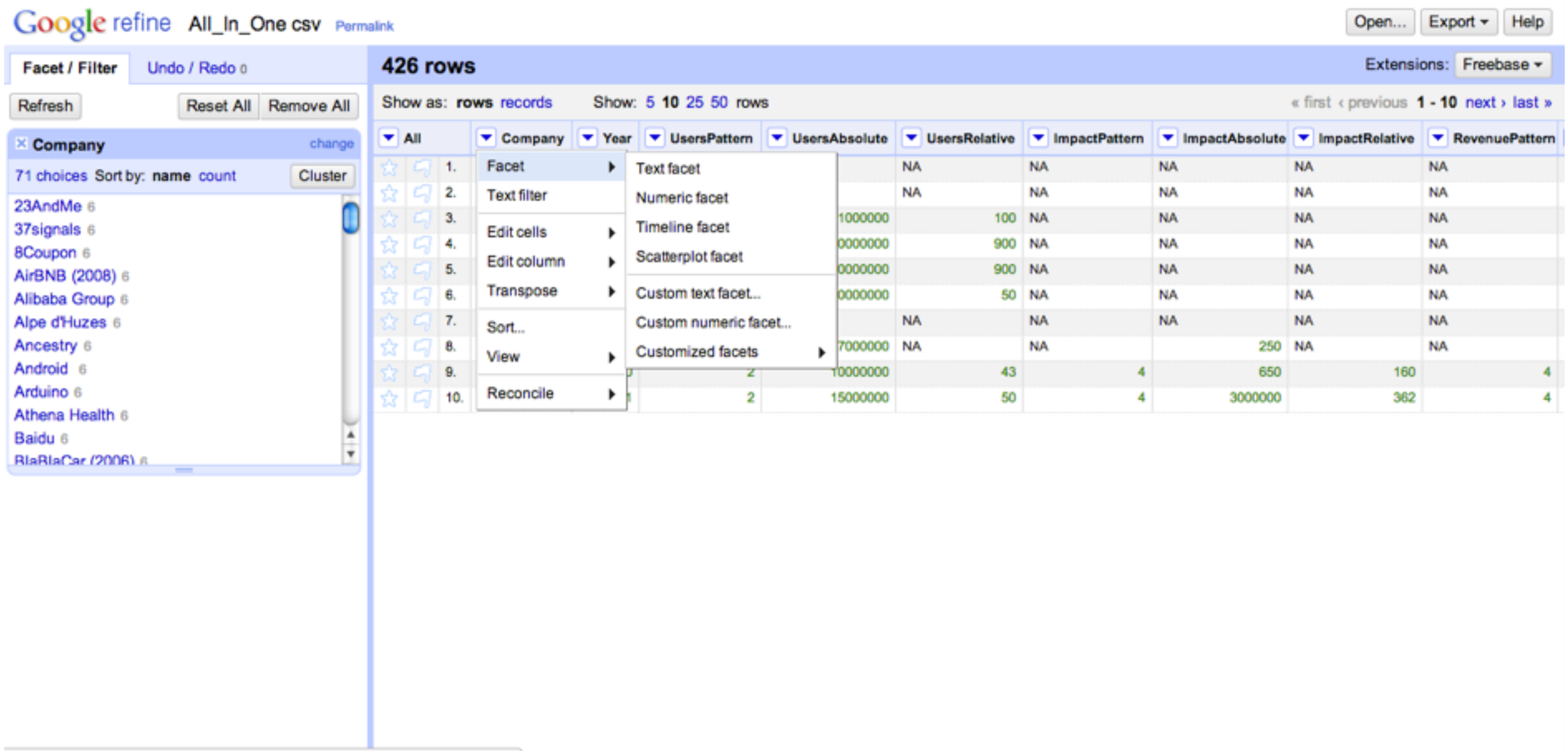
All			Company	Year	UsersPattern	UsersAbsolute	UsersRelative	ImpactPattern			
		1.	Instragram	2008	NA	NA	NA	NA	NA		
		2.	Instragram	2009	NA	NA	NA	NA	NA		
		3.	Instragram	2010	NA	1000000	100	NA	NA		
		4.	Instragram	2011	4	10000000	900	NA	NA		
		5.	Instragram	2012	4	100000000	900	NA	NA		
		6.	Instragram	2013	2	150000000	50	NA	NA		
		7.	Spotify	2008	NA	NA	NA	NA	NA		
		8.	Spotify	2009	NA	7000000	NA	NA			
		9.	Spotify	2010	2	10000000	43	4	650	160	4
		10.	Spotify	2011	2	15000000	50	4	3000000	362	4

Export project
Tab-separated value
Comma-separated value
HTML table
Excel
ODF spreadsheet
Triple loader
MQLWrite
Custom tabular exporter...
Templating...

https://github.com/OpenRefine

https://code.google.com/p/google-refine/downloads/list

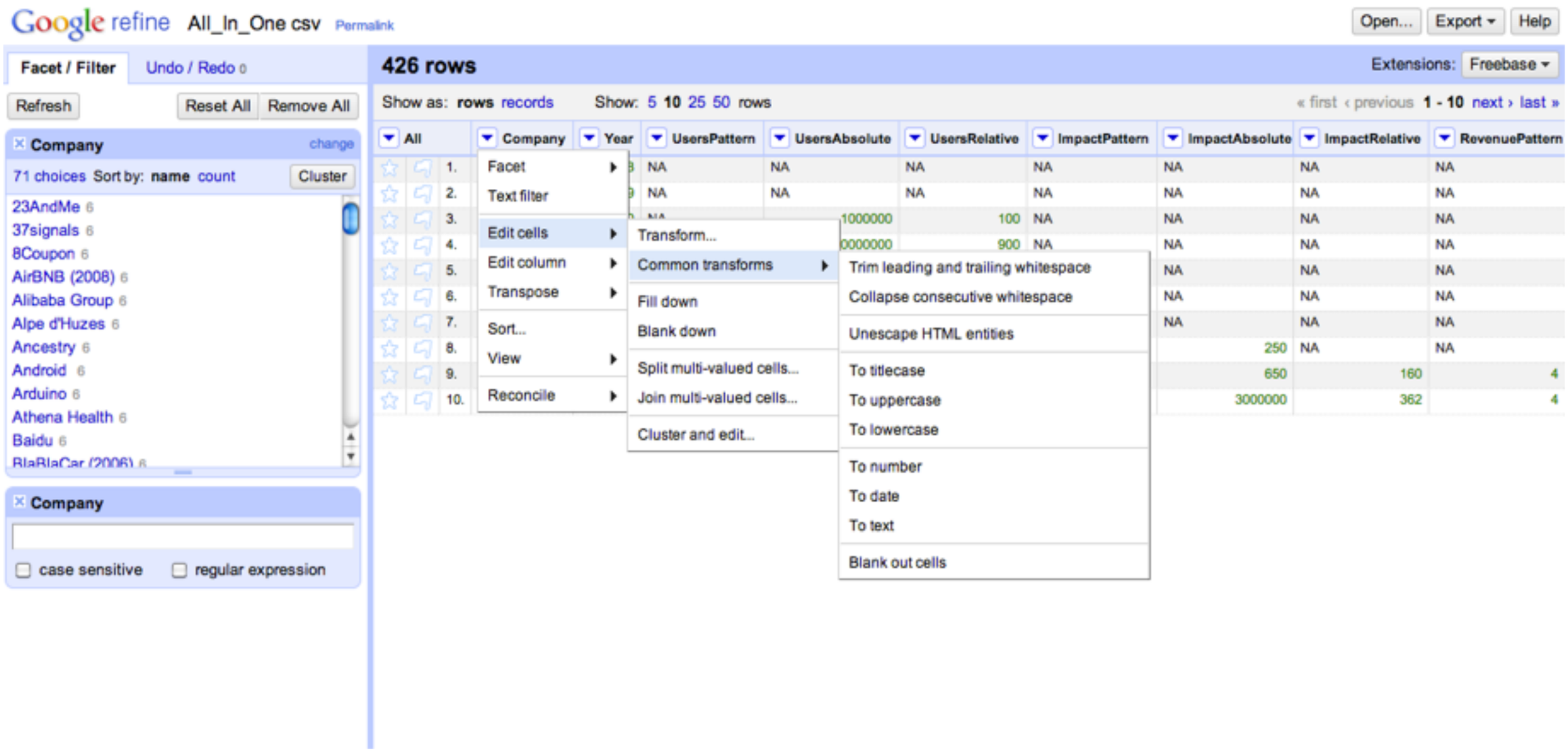
Open Refine: Cleaning Data



https://github.com/OpenRefine

https://code.google.com/p/google-refine/downloads/list

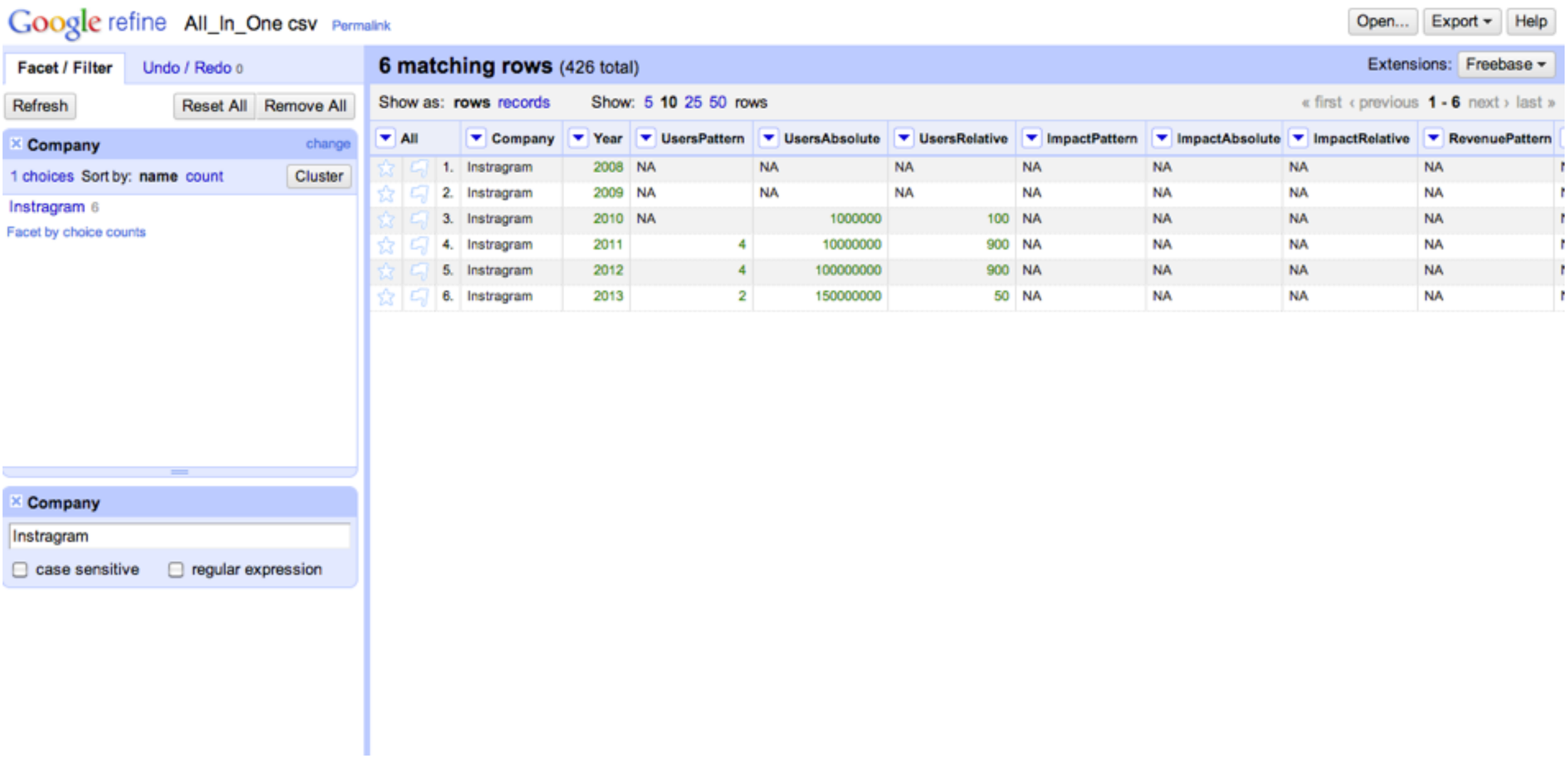
Open Refine: Cleaning Data



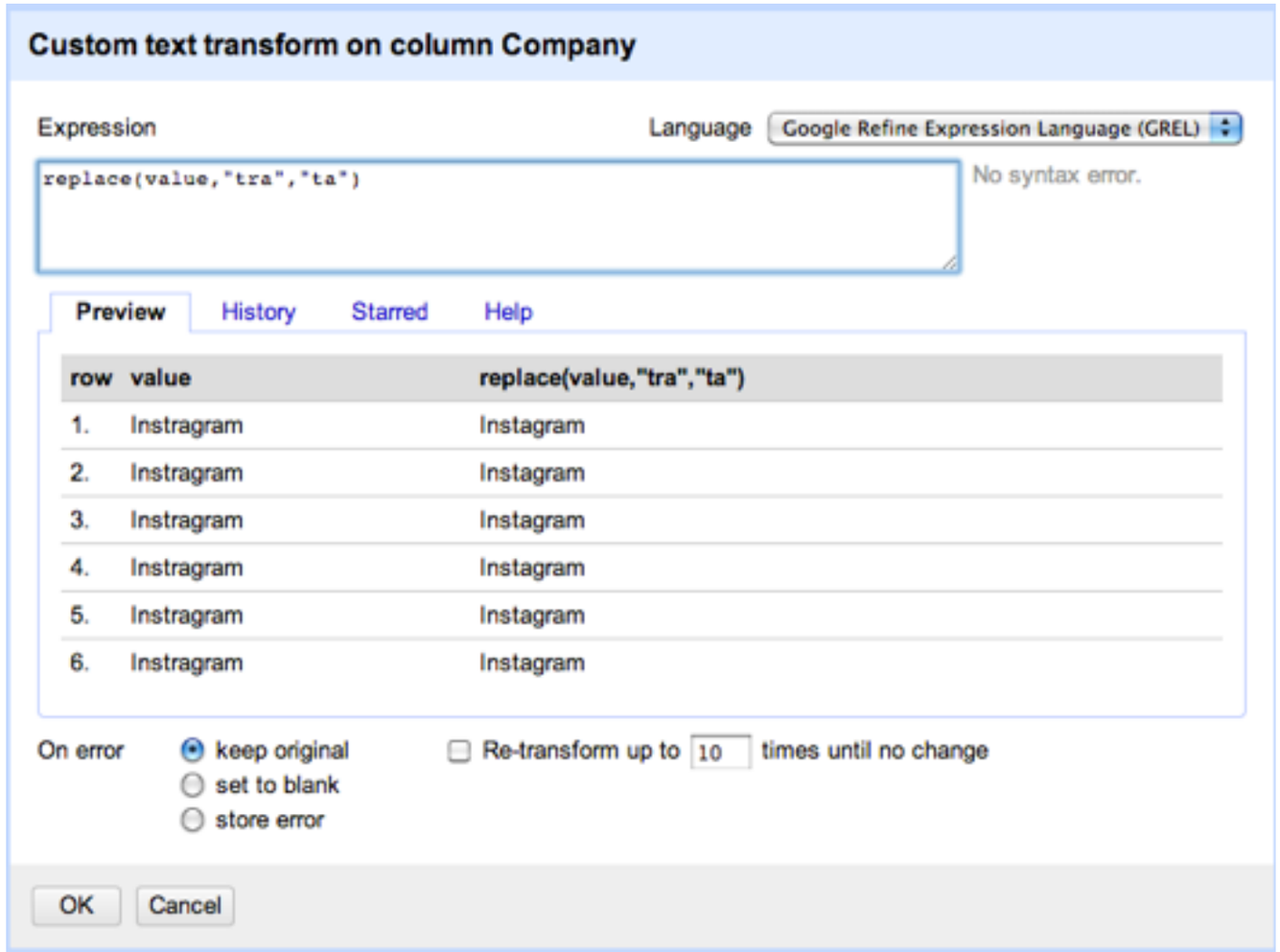
https://github.com/OpenRefine

https://code.google.com/p/google-refine/downloads/list

Open Refine: Cleaning Data



Open Refine: Cleaning Data



https://github.com/OpenRefine/OpenRefine/wiki/GREL-Functions

https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions

Open Refine:Ejercicio guiado

- Descargar Google Refine del enlace de abajo

https://code.google.com/p/google-refine/downloads/list

Open Refine:Ejercicio guiado

- Descargar el Dataset del enlace de abajo

https://raw.githubusercontent.com/hackbdw14/hackatonData/master/datasets/facilities/data/allotjament.csv

Open Refine:Ejercicio guiado

- Crear proyecto y abrir el dataset
- Filtrar por nombre de calle (búsqueda)
- Facetar por distrito y eliminar basura
- Facetar por categoría
- Facetar por distrito y eliminar tildes y diéresis
- Undo al principio y facetado numérico. Luego custom text facet
- Salvar el fichero

Open Refine:Ejercicio

https://raw.githubusercontent.com/hackbdw14/hackatonData/master/datasets/facilities/data/centres_informacio.csv

- Crear proyecto y abrir el dataset
- Encontrar todos los centros en ‘Av Diagonal’ (cualquier número)
- ¿Qué distritos ‘toca’?
- ¿Cuántos centros hay de cada categoría?
- Resetear todo y Filtrar por distrito ‘Eixample’. Corregir los barrios.
- Filtrar con el facetado numérico cualquier punto con lat <41.25
- Ídem con text faceting booleano. Exportar el fichero

Open Refine: id-distrito

https://raw.githubusercontent.com/hackbdw14/hackatonData/master/datasets/estadistica/cohesionSocial/C.csv

- Edit column/split into different columns

Open Refine: otras funciones

- value.replace("cadena1", "cadena2")
- value.contains("")
- toNumber(“value")
- cells[“columna”].value
- cell.cross("project", “column").cells["column"].value
- toNumber(value.replace(".","").match(/([\d\.]+).*/)[0])

Open Refine:Ejercicio Guiado

Vamos a llamar a un servicio externo de geoencoding

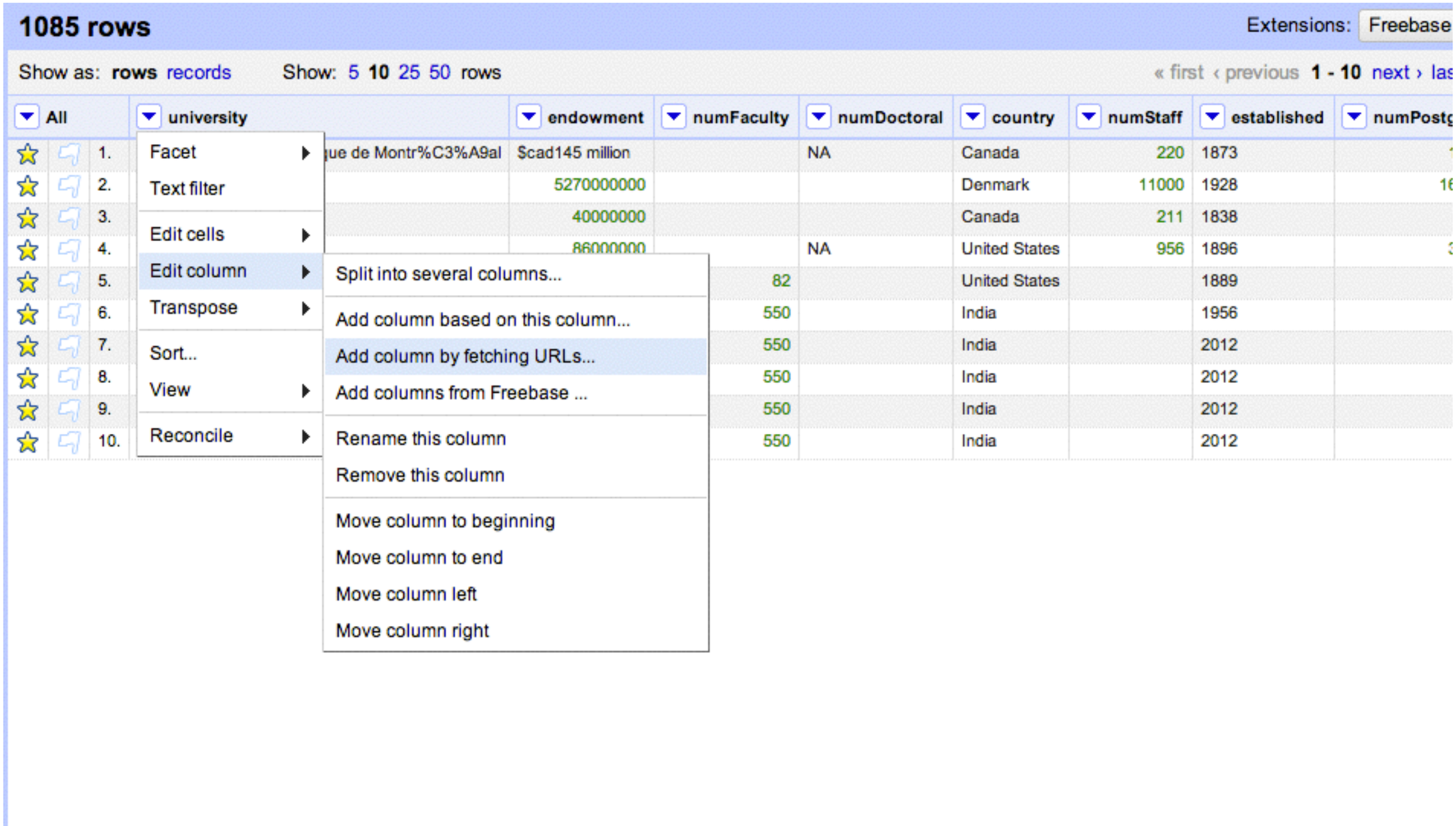
(ejemplo de http://enipedia.tudelft.nl/enipedia/images/f/ff/UniversityData.zip)

I) Marcamos las 10 primeras universidades con una estrella

II) Facetamos por estrella

III) Nos quedamos con las 'true'

Open Refine:Ejercicio Guiado



Open Refine:Ejercicio Guiado

Add column by fetching URLs based on column university

New column name

Throttle delay 5000 milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression

Language Google Refine Expression Language (GREL)

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" +
escape(value, "url")
```

No syntax error.

Preview History Starred Help

row	value	"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
1.	%C3%89cole Polytechnique de Montr%C3%A9al	http://maps.google.com/maps/api/geocode/json?sensor=false&address=%25C3%2589cole+Polytechnique+de+Montr%
2.	Aarhus University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Aarhus+University
3.	Acadia University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Acadia+University
4.	Adelphi University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Adelphi+University

OK Cancel

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
```

Open Refine:Ejercicio Guiado

Add column by fetching URLs based on column university

New column name

Throttle delay 5000 milliseconds

On error set to blank store error

Formulate the URLs to fetch:

Expression

Language Google Refine Expression Language (GREL)

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" +
escape(value, "url")
```

No syntax error.

Preview History Starred Help

row	value	"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
1.	%C3%89cole Polytechnique de Montr%C3%A9al	http://maps.google.com/maps/api/geocode/json?sensor=false&address=%25C3%2589cole+Polytechnique+de+Montr%
2.	Aarhus University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Aarhus+University
3.	Acadia University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Acadia+University
4.	Adelphi University	http://maps.google.com/maps/api/geocode/json?sensor=false&address=Adelphi+University

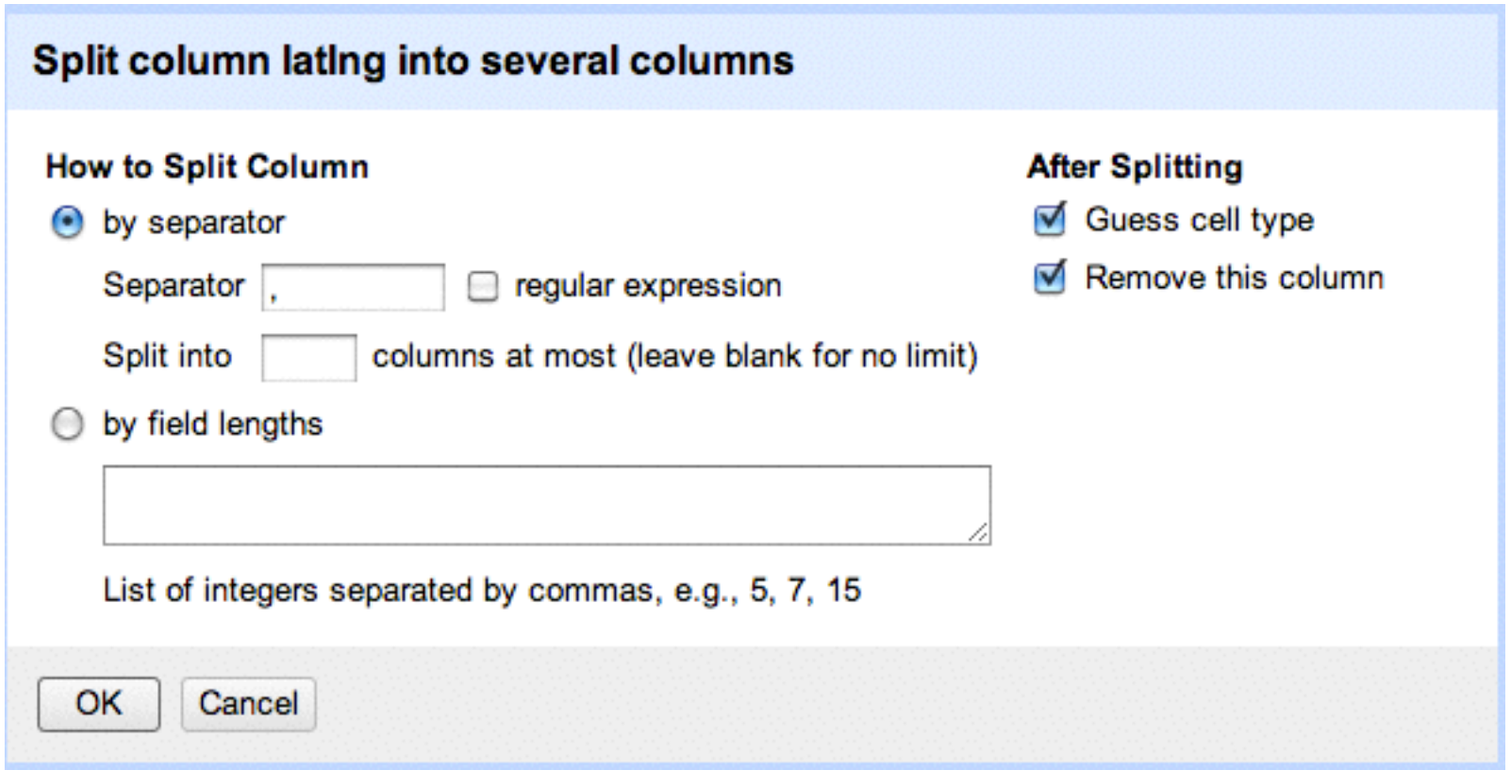
OK Cancel

```
"http://maps.google.com/maps/api/geocode/json?sensor=false&address=" + escape(value, "url")
```

Open Refine:Ejercicio Guiado


```
with(value.parseJson().results[0].geometry.location, pair, pair.lat +", " + pair.lng)
```

Open Refine:Ejercicio Guiado



Open Refine:Ejercicio Guiado

All			university	geoResponse	latlng 1	latlng 2
		1.	%C3%89cole Polytechnique de Montr%C3%A9al	{ "results" : [], "status" : "ZERO_RESULTS" }	latlng 1	
		2.	Aarhus University	{ "results" : [{ "address_components" : [{ "long_name" : "Aarhus University", "short_name" : "Aarhus University", "types" : ["point_of_interest", "establishment"] }, { "long_name" : "Aarhus University", "short_name" : "Aarhus University", "types" : ["establishment"] }, { "long_name" : "3", "short_name" : "3", "types" : ["street_number"] }, { "long_name" : "Victor Albecks Vej", "short_name" : "Victor Albecks Vej", "types" : ["route"] }, { "long_name" : "Aarhus Universitet", "short_name" : "Aarhus Universitet", "types" : ["sublocality", "political"] }, { "long_name" : "Midtbyen", "short_name" : "Midtbyen", "types" : ["sublocality", "political"] }, { "long_name" : "Aarhus C", "short_name" : "Aarhus C", "types" : ["sublocality", "political"] }	56.1702228	10.2000197