

Taller #bdw14. Día 1

Fuentes y limpieza de datos



<https://github.com/hackbdw14/hackatonData>

hackbdw14@gmail.com
@hackbdw14

Licencia Universal: Puedes hacer con este material lo que quieras
<http://creativecommons.org/publicdomain/zero/1.0/deed.es>

Contenidos

Minería de Datos

- ▶ Motivación
- ▶ El Pipeline de datos
- ▶ Herramientas

Datasets para el hackaton

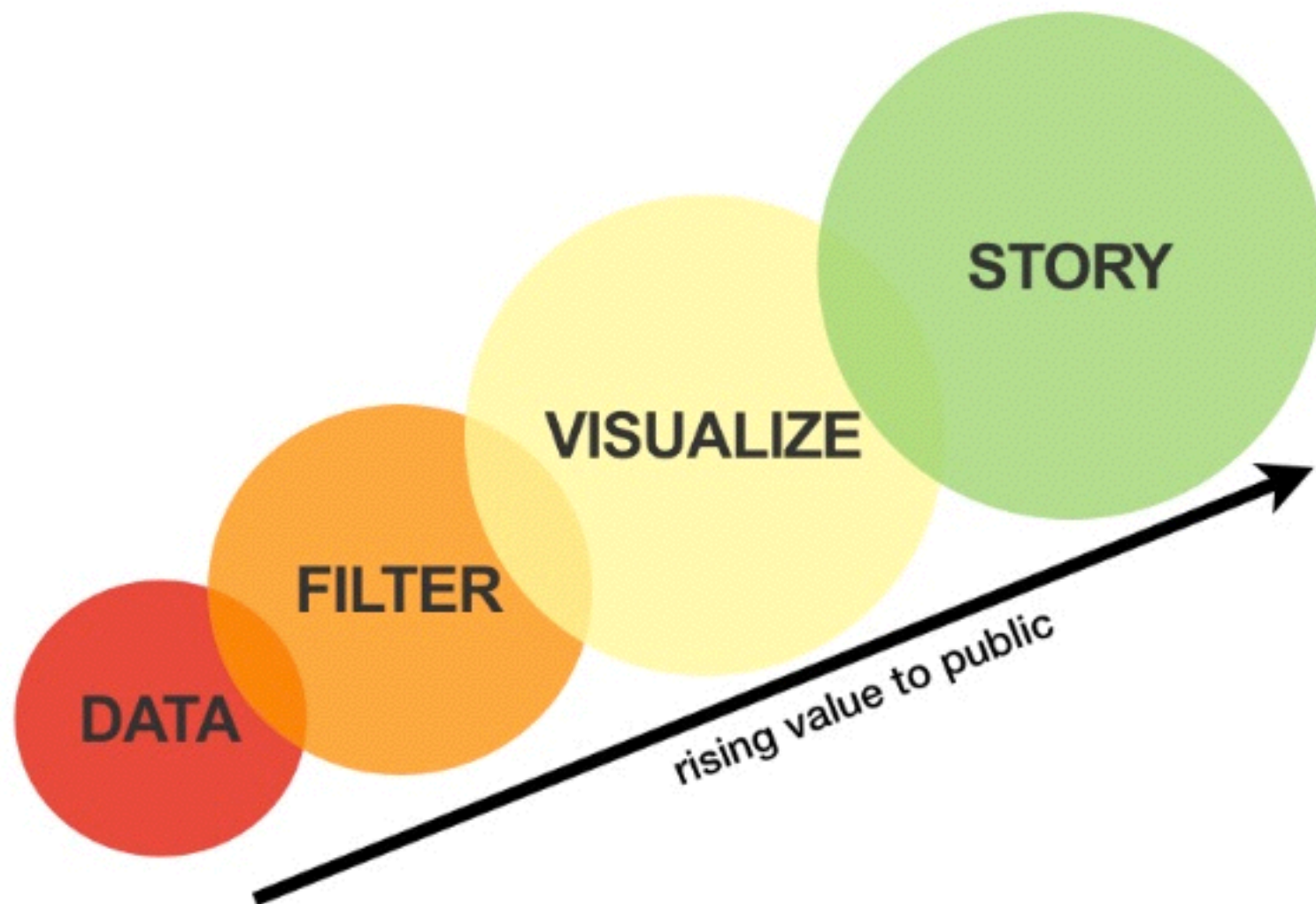
Adquisición de Datos

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

Open Refine

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

Minería de Datos



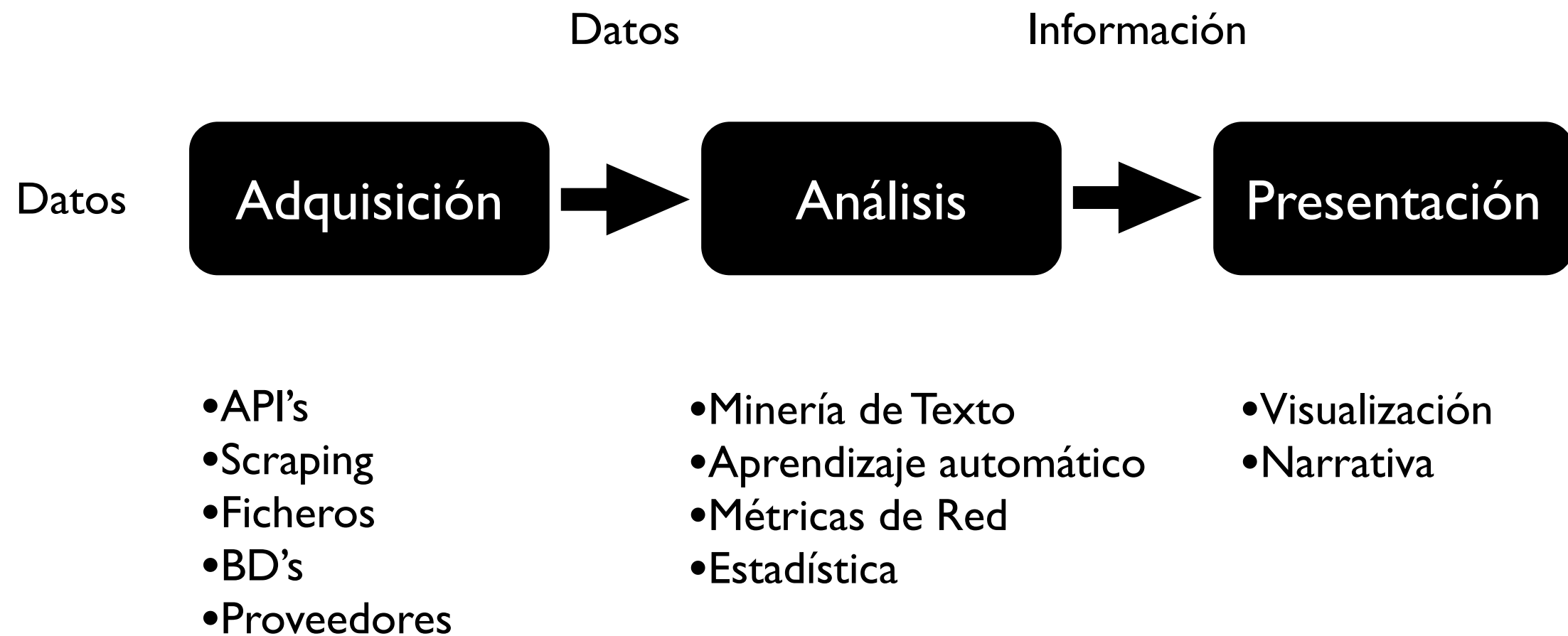
Attribution 2.0 Generic (CC BY 2.0)

Mirko Lorenz, 2010

Motivación



El Pipeline de Datos



El Pipeline de Datos: Herramientas

Adquisición de Datos

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

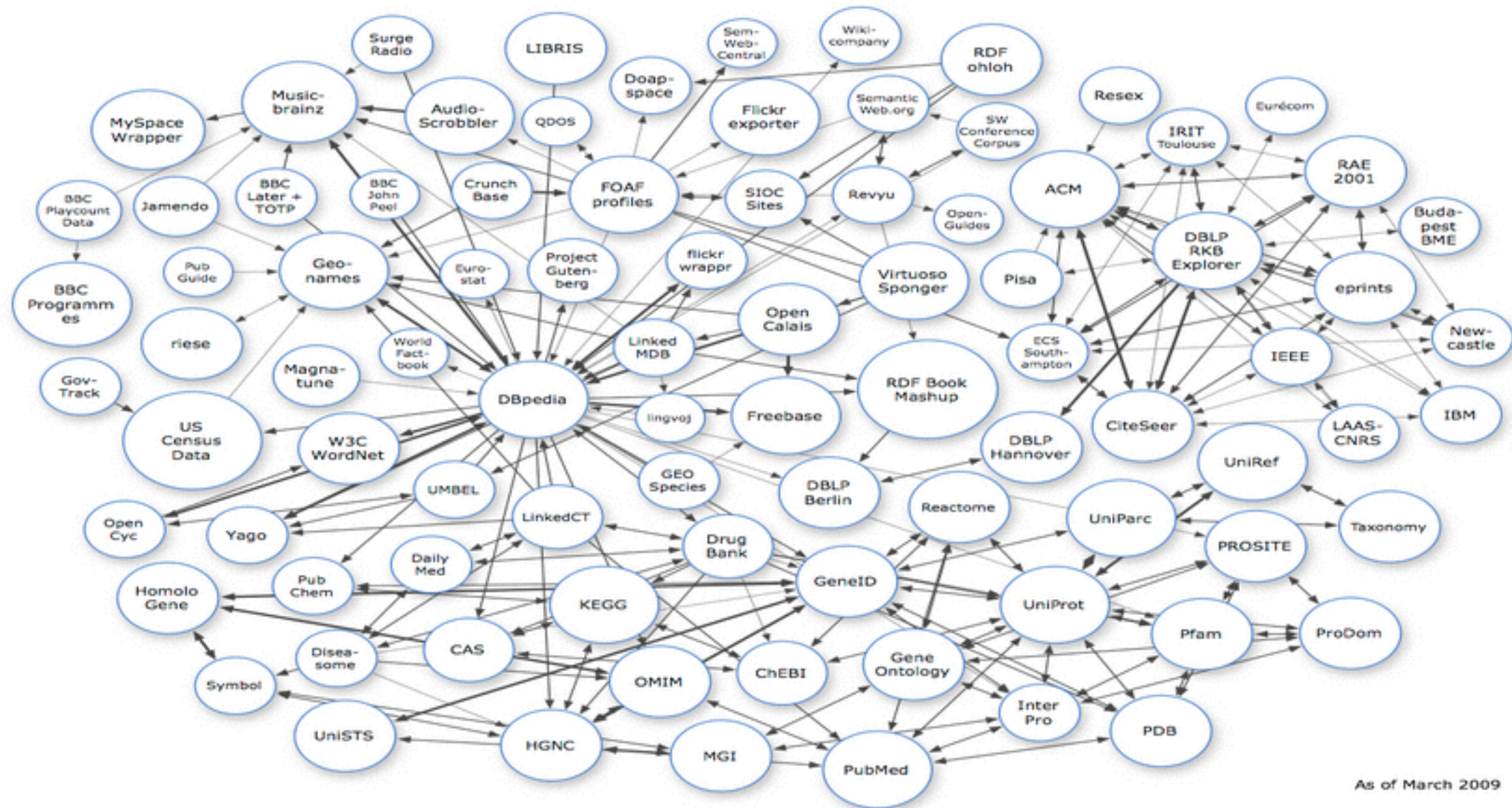
Análisis

- ▶ Open Refine
- ▶ Google Spreadsheet
- ▶ Outliers Data Tools

Visualización de Datos

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

Adquisición de Datos



As of March 2009

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

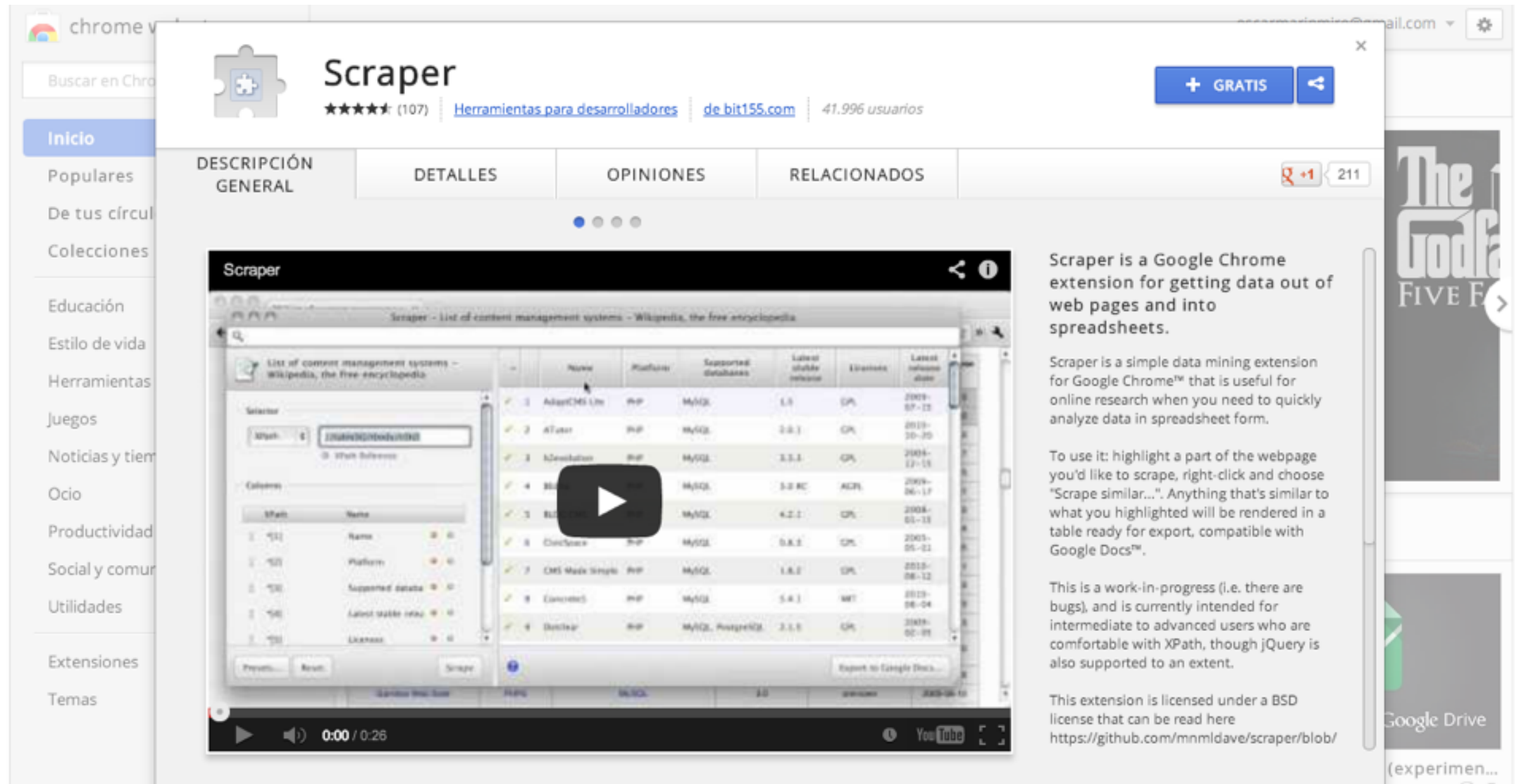
http://epp.eurostat.ec.europa.eu/portal/page/portal/region_cities/city_urban/

<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



Scraper
★★★★★ (107) | [Herramientas para desarrolladores](#) | [de bit155.com](#) | 41.996 usuarios

DESCRIPCIÓN GENERAL | DETALLES | OPINIONES | RELACIONADOS

Scraper

Scraper is a Google Chrome extension for getting data out of web pages and into spreadsheets.

Scraper is a simple data mining extension for Google Chrome™ that is useful for online research when you need to quickly analyze data in spreadsheet form.

To use it: highlight a part of the webpage you'd like to scrape, right-click and choose "Scrape similar...". Anything that's similar to what you highlighted will be rendered in a table ready for export, compatible with Google Docs™.

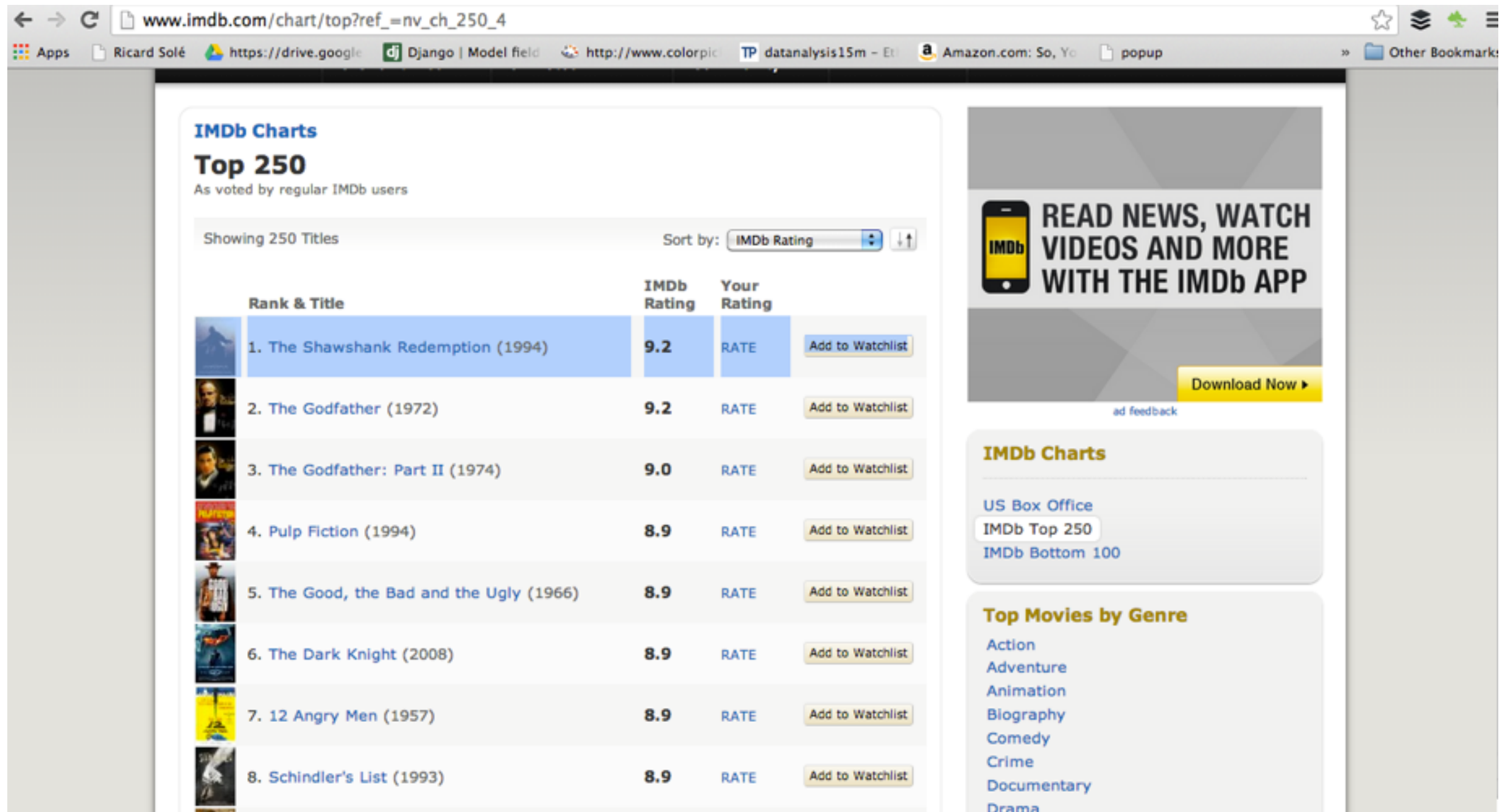
This is a work-in-progress (i.e. there are bugs), and is currently intended for intermediate to advanced users who are comfortable with XPath, though jQuery is also supported to an extent.

This extension is licensed under a BSD license that can be read here <https://github.com/mnmdave/scraper/blob/>

Name	Platform	Supported databases	Latest stable release	License	Latest release date
1. AdaptCMS Lite	PHP	MySQL	1.0	GPL	2009-07-13
2. A7user	PHP	MySQL	2.0.1	GPL	2010-10-20
3. A7evolution	PHP	MySQL	3.3.1	GPL	2009-12-13
4. Blog	PHP	MySQL	3.0 RC	AGPL	2009-06-17
5. Blog	PHP	MySQL	4.2.1	GPL	2008-01-13
6. ConQuest	PHP	MySQL	0.8.1	GPL	2009-05-21
7. CMS Made Simple	PHP	MySQL	1.8.2	GPL	2010-08-12
8. Concrete5	PHP	MySQL	5.4.1	MIT	2010-08-04
9. DedeCMS	PHP	MySQL, PostgreSQL	2.1.0	GPL	2009-02-01

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

Scrape w/ Chrome Scraper



IMDb Charts
Top 250
As voted by regular IMDb users

Showing 250 Titles Sort by: IMDb Rating

Rank & Title	IMDb Rating	Your Rating	
1. The Shawshank Redemption (1994)	9.2	RATE	Add to Watchlist
2. The Godfather (1972)	9.2	RATE	Add to Watchlist
3. The Godfather: Part II (1974)	9.0	RATE	Add to Watchlist
4. Pulp Fiction (1994)	8.9	RATE	Add to Watchlist
5. The Good, the Bad and the Ugly (1966)	8.9	RATE	Add to Watchlist
6. The Dark Knight (2008)	8.9	RATE	Add to Watchlist
7. 12 Angry Men (1957)	8.9	RATE	Add to Watchlist
8. Schindler's List (1993)	8.9	RATE	Add to Watchlist

READ NEWS, WATCH VIDEOS AND MORE WITH THE IMDb APP
Download Now

IMDb Charts

US Box Office
IMDb Top 250
IMDb Bottom 100

Top Movies by Genre

Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama

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

Scrape w/ Chrome Scraper

IMDb Charts

Top 250

As voted by regular IMDb users

Showing 250 Titles

Sort by: IMDb Rating

Rank & Title	IMDb Rating	Your Rating
1. The Shawshank Redemption (1994)	9.2	RATE...
2. The Godfather (1972)	8.9	RATE...
3. The Godfather: Part II (1974)	8.9	RATE...
4. Pulp Fiction (1994)	8.9	RATE...
5. The Good, the Bad and the Ugly (1966)	8.9	RATE
6. The Dark Knight (2008)	8.9	RATE
7. 12 Angry Men (1957)	8.9	RATE
8. Schindler's List (1993)	8.9	RATE

Copy
Search Google for '1. The Shawshank Redemption (1994) 9.2 RATE...'
Print...
Buffer Selected Text
Scrape similar...
Inspect Element
Look Up in Dictionary
Speech



READ NEWS, WATCH
VIDEOS AND MORE
WITH THE IMDb APP

Download Now

Top Movies by Genre

Action
Adventure
Animation
Biography
Comedy
Crime
Documentary
Drama

Scrape w/ Chrome Scraper

The screenshot displays the Chrome Scraper extension interface, titled "Scraper - IMDb Top 250 - IMDb". The interface is divided into several sections:

- Selector:** A dropdown menu set to "XPath" with a text input field containing the XPath expression `//div[3]/div[1]/div/table/tbody/tr`. Below the input is a link for "XPath Reference".
- Columns:** A table with two columns: "XPath" and "Name". It lists four columns to be scraped:

XPath	Name
*[1]	Name (optional)
*[2]	Rank & Title
*[3]	IMDb Rating
*[4]	Your Rating
- Buttons:** "Presets...", "Reset", "Scrape", and "Export to Google Docs..."
- Table:** A table showing the top 8 movies from the IMDb Top 250 list. Each row includes a rank, title, IMDb rating, and a "Your Rating" section with a "RATE" button and a grid of 10 stars. An "Add to Watchlist" button is also present for each row.

The table data is as follows:

Rank	Title	IMDb Rating	Your Rating	Action
1	1. The Shawshank Redemption (1994)	9.2	RATE 1 2 3 4 5 6 7 8 9 10 9.3/10 X	Add to Watchlist
2	2. The Godfather (1972)	9.2	RATE 1 2 3 4 5 6 7 8 9 10 9.2/10 X	Add to Watchlist
3	3. The Godfather: Part II (1974)	9.0	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
4	4. Pulp Fiction (1994)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
5	5. The Good, the Bad and the Ugly (1966)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
6	6. The Dark Knight (2008)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 9/10 X	Add to Watchlist
7	7. 12 Angry Men (1957)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 8.9/10 X	Add to Watchlist
8	8. Schindler's List (1993)	8.9	RATE 1 2 3 4 5 6 7 8 9 10 8.9/10 X	Add to Watchlist

Ejercicio

Scrapear con Chrome Scraper la siguiente url:

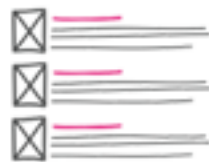
<http://www.bcn.cat/estadistica/catala/dades/economia/renda/rdfamiliar/a2012/rfbarris.htm>

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

Scrape w/ import.io



Our browser is quite new and we're still working out the kinks. [Click here for a list of our known issues.](#)



Create Data Set

+ New Data Source

Create a new source to bring data into import.io



Connector

Use a search box to retrieve one or more pages of results



Extractor

Take a single page and extract all of the data from it



Crawler

Get data from multiple similar pages on the same site



My Data

<http://import.io/>

PDF Extraction: Tabula



Tabula

Tabula is a tool for liberating data tables trapped inside PDF files.

[View the Project on GitHub](#)
jazzido/tabula

Download for
Windows

Download for
Mac

View source on
GitHub

Current Version: [0.9.1 \(archive\)](#)



Tabula

Tabula helps you liberate data tables trapped inside PDFs



0:00 / 0:30



Using Tabula

1. Upload a file with tables you would like to copy.
2. Draw a box around the area of the table you would like to copy.
(Note: currently, Tabula can't select tables over multiple pages)
3. You will be given the option to copy the table as a CSV (comma-separated values) file or download the CSV or TSV (tab separated values). If you notice any errors in the table, you can make text edits to the selected text before copying or

<http://tabula.nerdpower.org/>

PDF Extraction: Tabula

Tabula is experimental software Home About

Example. Mr. and Mrs. Brown are filing a joint return. Their taxable income on Form 1040, line 43, is \$25,300. First, they find the \$25,300-\$25,350 taxable income line. Next, they find the column for married filing jointly and read down the column. The amount shown where the taxable income line and filing status column meet is \$2,929. This is the tax amount they should enter on Form 1040, line 44.

Load	Less Than	Single	Married Jointly	Married Separately	Head of Household
25,200	25,250	3,349	2,914	3,349	3,164
25,250	25,300	3,350	2,921	3,350	3,171
25,300	25,350	3,364	2,929	3,364	3,179
25,350	25,400	3,371	2,936	3,371	3,186

Page 1

Page 2

Page 3

Loading...

If line 43 (taxable income) is—		And you are—				If line 43 (taxable income) is—		And you are—				If line 43 (taxable income) is—		And you are—			
At least	But less than	Single	Married filing jointly	Married filing separately	Head of a household	At least	But less than	Single	Married filing jointly	Married filing separately	Head of a household	At least	But less than	Single	Married filing jointly	Married filing separately	Head of a household
1,000																	
0	5	0	0	0	0	1,000	1,025	101	101	101	101	2,000	2,025	201	201	201	201
5	15	1	1	1	1	1,025	1,050	104	104	104	104	2,025	2,050	204	204	204	204
15	25	2	2	2	2	1,050	1,075	106	106	106	106	2,050	2,075	206	206	206	206
25	35	4	4	4	4	1,075	1,100	109	109	109	109	2,075	2,100	209	209	209	209
35	45	6	6	6	6	1,100	1,125	111	111	111	111	2,100	2,125	211	211	211	211
45	55	8	8	8	8	1,125	1,150	114	114	114	114	2,125	2,150	214	214	214	214
55	65	10	10	10	10	1,150	1,175	116	116	116	116	2,150	2,175	216	216	216	216
65	75	12	12	12	12	1,175	1,200	119	119	119	119	2,175	2,200	219	219	219	219
75	85	14	14	14	14	1,200	1,225	121	121	121	121	2,200	2,225	221	221	221	221
85	95	16	16	16	16	1,225	1,250	124	124	124	124	2,225	2,250	224	224	224	224
95	105	18	18	18	18	1,250	1,275	126	126	126	126	2,250	2,275	226	226	226	226
105	115	20	20	20	20	1,275	1,300	129	129	129	129	2,275	2,300	229	229	229	229
115	125	22	22	22	22	1,300	1,325	131	131	131	131	2,300	2,325	231	231	231	231
125	135	24	24	24	24	1,325	1,350	134	134	134	134	2,325	2,350	234	234	234	234
135	145	26	26	26	26	1,350	1,375	136	136	136	136	2,350	2,375	236	236	236	236
145	155	28	28	28	28	1,375	1,400	139	139	139	139	2,375	2,400	239	239	239	239
155	165	30	30	30	30	1,400	1,425	141	141	141	141	2,400	2,425	241	241	241	241
165	175	32	32	32	32	1,425	1,450	144	144	144	144	2,425	2,450	244	244	244	244
175	185	34	34	34	34	1,450	1,475	146	146	146	146	2,450	2,475	246	246	246	246
185	195	36	36	36	36	1,475	1,500	149	149	149	149	2,475	2,500	249	249	249	249
195	205	38	38	38	38	1,500	1,525	151	151	151	151	2,500	2,525	251	251	251	251
205	215	40	40	40	40	1,525	1,550	154	154	154	154	2,525	2,550	254	254	254	254
215	225	42	42	42	42	1,550	1,575	156	156	156	156	2,550	2,575	256	256	256	256
225	235	44	44	44	44	1,575	1,600	159	159	159	159	2,575	2,600	259	259	259	259
235	245	46	46	46	46	1,600	1,625	161	161	161	161	2,600	2,625	261	261	261	261
245	255	48	48	48	48	1,625	1,650	164	164	164	164	2,625	2,650	264	264	264	264
255	265	50	50	50	50	1,650	1,675	166	166	166	166	2,650	2,675	266	266	266	266
265	275	52	52	52	52	1,675	1,700	169	169	169	169	2,675	2,700	269	269	269	269
275	285	54	54	54	54	1,700	1,725	171	171	171	171	2,700	2,725	271	271	271	271
285	295	56	56	56	56	1,725	1,750	174	174	174	174	2,725	2,750	274	274	274	274
295	305	58	58	58	58	1,750	1,775	176	176	176	176	2,750	2,775	276	276	276	276
305	315	60	60	60	60	1,775	1,800	179	179	179	179	2,775	2,800	279	279	279	279
315	325	62	62	62	62	1,800	1,825	181	181	181	181	2,800	2,825	281	281	281	281
325	335	64	64	64	64	1,825	1,850	184	184	184	184	2,825	2,850	284	284	284	284
335	345	66	66	66	66	1,850	1,875	186	186	186	186	2,850	2,875	286	286	286	286
345	355	68	68	68	68	1,875	1,900	189	189	189	189	2,875	2,900	289	289	289	289
355	365	70	70	70	70	1,900	1,925	191	191	191	191	2,900	2,925	291	291	291	291
365	375	72	72	72	72	1,925	1,950	194	194	194	194	2,925	2,950	294	294	294	294
375	385	74	74	74	74	1,950	1,975	196	196	196	196	2,950	2,975	296	296	296	296
385	395	76	76	76	76	1,975	2,000	199	199	199	199	2,975	3,000	299	299	299	299

(Continued)

<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 mentioned 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

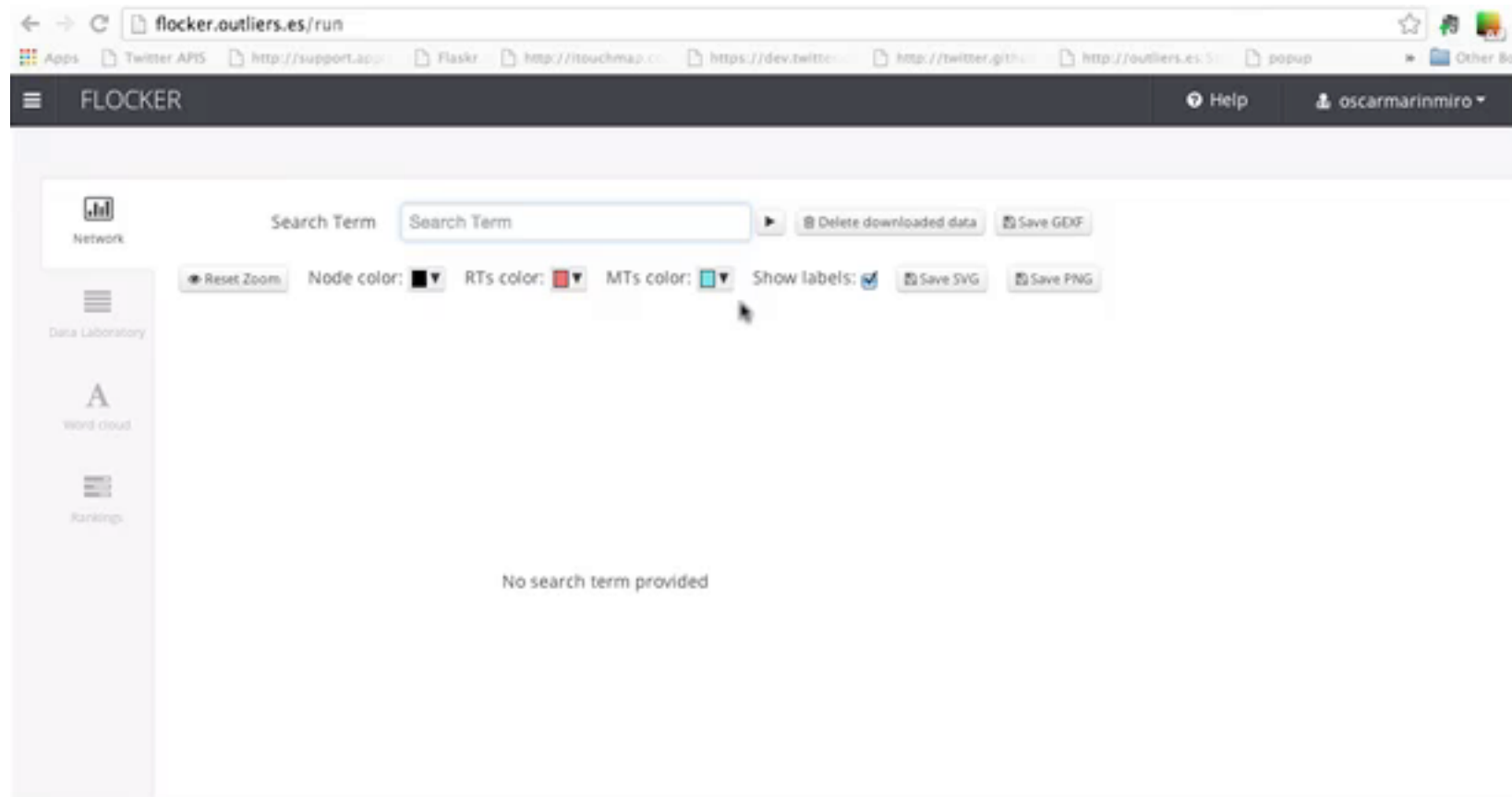
The chart on the right shows the percentage of features we have currently developed.



➔ 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

Google refine *A power tool for working with messy data.*

Create Project « Start Over Configure Parsing Options Project name All_In_One.csv Create Project »

Open Project
Import Project

16.	Quora	2011	4	547777	37163	4	2835	77	NA	11000000	NA
17.	Quora	2012	NA	NA	NA	4	792514	180	NA	61000000	NA
18.	Quora	2013	NA	NA	NA	NA	NA	NA	NA	NA	NA
19.	Evernote	2008	NA	500	NA	NA	4000	NA	NA	NA	NA
20.	Evernote	2009	4	2096030	319	4	3531	783	NA	6500000	100
21.	Evernote	2010	4	6003130	186	4	201308	470	1	9058860	39
22.	Evernote	2011	4	20000000	233	4	750	273	3	18000000	99
23.	Evernote	2012	4	41000000	105	4	1500000	100	4	80000000	344
24.	Evernote	2013	2	65000000	59	NA	NA	NA	NA	NA	NA
25.	Research Gate	2008	NA	NA	NA	NA	NA	NA	NA	NA	NA
26.	Research Gate	2009	NA	140	100	NA	40	NA	NA	NA	NA
27.	Research Gate	2010	4	500	257	2	60	50	NA	NA	NA
28.	Research Gate	2011	4	1300000	160	4	27000000	44900	NA	NA	NA

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

RDF/N3 files

XML files

Open Document Format spreadsheets (.ods)

RDF/XML files

Character encoding

Columns are separated by

☒ commas (CSV)

☐ tabs (TSV)

☐ custom ,

Escape special characters with \

Update Preview

☐ Ignore first 0 line(s) at beginning of file

☒ Parse next 1 line(s) as column headers

☐ Discard initial 0 row(s) of data

☐ Load at most 0 row(s) of data

☒ Parse cell text into numbers, dates, ...

☒ Store blank rows

☒ Store blank cells as nulls

Version 2.5 (j2407)

Help

About

<https://github.com/OpenRefine>

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

Open Refine: Cleaning Data

Google refine All_In_One.csv Permalink

Facet / Filter Undo / Redo 0

426 rows

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

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](#)

			Company	Year	UsersPattern	UsersAbsolute	UsersRelative	ImpactPattern	
☆	1.		Instagram	2008	NA	NA	NA	NA	NA
☆	2.		Instagram	2009	NA	NA	NA	NA	NA
☆	3.		Instagram	2010	NA	1000000	100	NA	NA
☆	4.		Instagram	2011	4	10000000	900	NA	NA
☆	5.		Instagram	2012	4	10000000	900	NA	NA
☆	6.		Instagram	2013	2	15000000	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

Google refine All_In_One.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 0

Refresh Reset All Remove All

426 rows

Extensions: Freebase

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

Company 71 choices Sort by: name count Cluster

23AndMe 6
37signals 6
BCoupon 6
AirBNB (2008) 6
Alibaba Group 6
Alpe d'Huizes 6
Ancestry 6
Android 6
Arduino 6
Athena Health 6
Baidu 6
RiaRiaCar (2006) 6

1. Facet
2. Text filter
3. Edit cells
4. Edit column
5. Transpose
6. Sort...
7. View
8. Reconcile

Text facet
Numeric facet
Timeline facet
Scatterplot facet
Custom text facet...
Custom numeric facet...
Customized facets

	All	Company	Year	UsersPattern	UsersAbsolute	UsersRelative	ImpactPattern	ImpactAbsolute	ImpactRelative	RevenuePattern
1.	Facet					NA	NA	NA	NA	NA
2.	Text filter					NA	NA	NA	NA	NA
3.	Edit cells				1000000	100	NA	NA	NA	NA
4.	Edit column				0000000	900	NA	NA	NA	NA
5.	Transpose				0000000	900	NA	NA	NA	NA
6.	Sort...				0000000	50	NA	NA	NA	NA
7.	View				7000000	NA	NA	250	NA	NA
8.	Reconcile				10000000	43	4	650	160	4
9.					2	50	4	3000000	362	4
10.					15000000					

<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

Refresh Reset All Remove All

426 rows

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

Extensions: Freebase

« first < previous 1 - 10 next > last »

Company 71 choices Sort by: name count Cluster

23AndMe 6
37signals 6
8Coupon 6
Airbnb (2008) 6
Alibaba Group 6
Alpe d'Huzes 6
Ancestry 6
Android 6
Arduino 6
Athena Health 6
Baidu 6
BlaBlaCar (2006) 6

Company

☐ case sensitive ☐ regular expression

1. Facet
2. Text filter
3. Edit cells
4. Edit column
5. Transpose
6. Sort...
7. View
8. Reconcile

Transform...
Common transforms
Fill down
Blank down
Split multi-valued cells...
Join multi-valued cells...
Cluster and edit...

Trim leading and trailing whitespace
Collapse consecutive whitespace
Unescape HTML entities
To titlecase
To uppercase
To lowercase
To number
To date
To text
Blank out cells

	All	Company	Year	UsersPattern	UsersAbsolute	UsersRelative	ImpactPattern	ImpactAbsolute	ImpactRelative	RevenuePattern
1.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
2.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
3.	NA	NA	NA	NA	1000000	100	NA	NA	NA	NA
4.	NA	NA	NA	NA	0000000	900	NA	NA	NA	NA
5.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
6.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
7.	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
8.	NA	NA	NA	NA	NA	NA	NA	250	NA	NA
9.	NA	NA	NA	NA	NA	NA	NA	650	160	4
10.	NA	NA	NA	NA	NA	NA	NA	3000000	362	4

<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

Refresh Reset All Remove All

Company change
1 choices Sort by: name count Cluster
Instagram 6
Facet by choice counts

Company
Instagram
☐ case sensitive ☐ regular expression

6 matching rows (426 total) Extensions: Freebase ▾

Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 6 next > last »

		▼ All	▼ Company	▼ Year	▼ UsersPattern	▼ UsersAbsolute	▼ UsersRelative	▼ ImpactPattern	▼ ImpactAbsolute	▼ ImpactRelative	▼ RevenuePattern
☆	1.		Instagram	2008	NA	NA	NA	NA	NA	NA	NA
☆	2.		Instagram	2009	NA	NA	NA	NA	NA	NA	NA
☆	3.		Instagram	2010	NA	1000000	100	NA	NA	NA	NA
☆	4.		Instagram	2011	4	10000000	900	NA	NA	NA	NA
☆	5.		Instagram	2012	4	100000000	900	NA	NA	NA	NA
☆	6.		Instagram	2013	2	150000000	50	NA	NA	NA	NA

Open Refine: Cleaning Data

Custom text transform on column Company

Expression Language Google Refine Expression Language (GREL)

`replace(value,"tra","ta")` No syntax error.

Preview History Starred Help

row	value	replace(value,"tra","ta")
1.	Instagram	Instagram
2.	Instagram	Instagram
3.	Instagram	Instagram
4.	Instagram	Instagram
5.	Instagram	Instagram
6.	Instagram	Instagram

On error ☒ keep original ☐ set to blank ☐ store error ☐ Re-transform up to times until no change

<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

1085 rows Extensions: Freebase

Show as: **rows** records Show: 5 10 25 50 rows « first < previous 1 - 10 next > last »

▼ All	▼ university	▼ endowment	▼ numFaculty	▼ numDoctoral	▼ country	▼ numStaff	▼ established	▼ numPostg
★ 1.	Facet	Scad145 million		NA	Canada	220	1873	
★ 2.	Text filter	5270000000			Denmark	11000	1928	16
★ 3.	Edit cells	40000000			Canada	211	1838	
★ 4.	Edit column	86000000		NA	United States	956	1896	3
★ 5.	Transpose		82		United States		1889	
★ 6.	Sort...		550		India		1956	
★ 7.	View		550		India		2012	
★ 8.	Reconcile		550		India		2012	
★ 9.			550		India		2012	
★ 10.			550		India		2012	

- Split into several columns...
- Add column based on this column...
- Add column by fetching URLs...
- Add columns from Freebase ...
- Rename this column
- Remove this column
- Move column to beginning
- Move column to end
- Move column left
- Move column right

Open Refine: Ejercicio Guiado

Add column by fetching URLs based on column university

New column name Throttle delay milliseconds

On error ☒ set to blank ☐ store error

Formulate the URLs to fetch:

Expression Language

```
"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 milliseconds

On error ☒ set to blank ☐ store error

Formulate the URLs to fetch:

Expression Language

```
"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 based on column geoResponse

New column name

On error

☒ set to blank ☐ store error ☐ copy value from original column

Expression

Language

Google Refine Expression Language (GREL)

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

No syntax error.

Preview

History

Starred

Help

9. { "results" : [{
 "address_components" : [{
 "long_name" : "AIIMS Raipur",
 "short_name" : "AIIMS Raipur",
 "types" : ["establishment"] }, {
 "long_name" : "AIIMS Campus",
 "short_name" : "AIIMS
Campus", "types" : [

21.2567797, 81.5791788

OK

Cancel

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

Open Refine:Ejercicio Guiado

Split column into several columns

How to Split Column

☒ by separator
Separator ☐ regular expression
Split into columns at most (leave blank for no limit)

☐ by field lengths

List of integers separated by commas, e.g., 5, 7, 15

After Splitting

☒ Guess cell type
☒ Remove this column

OK Cancel

Open Refine:Ejercicio Guiado

▼ All		▼ university	▼ geoResponse	▼ latlng 1	▼ latlng 2
★	💬	1.	%C3%89cole Polytechnique de Montr%C3%A9al	latlng 1	
★	💬	2.	Aarhus University	56.1702228	10.2000197

Licencia Universal: Puedes hacer con este material lo que quieras
<http://creativecommons.org/publicdomain/zero/1.0/deed.es>