

# Limpeza de dados

Limpando dados com OpenRefine

# Limpeza de dados

## Introdução

Fazer o trabalho de limpeza como fizemos na aula anterior é um pouco cansativo e muitas vezes precisamos conhecer o site para realizar a limpeza dos dados corretamente.

Muitas vezes temos a tarefa de limpar os dados sem vê-los antes, desta forma, criar um script para limpeza é uma tarefa muito complicada.

# Limpeza de dados

## OpenRefine

Existem ferramentas para nos auxiliar nesta tarefa, uma delas é o OpenRefine, que pode, além de realizar limpeza dos dados de forma rápida e cômoda, permitir que os dados possam ser visualizados facilmente, até por quem não é programador.

# Limpeza de dados

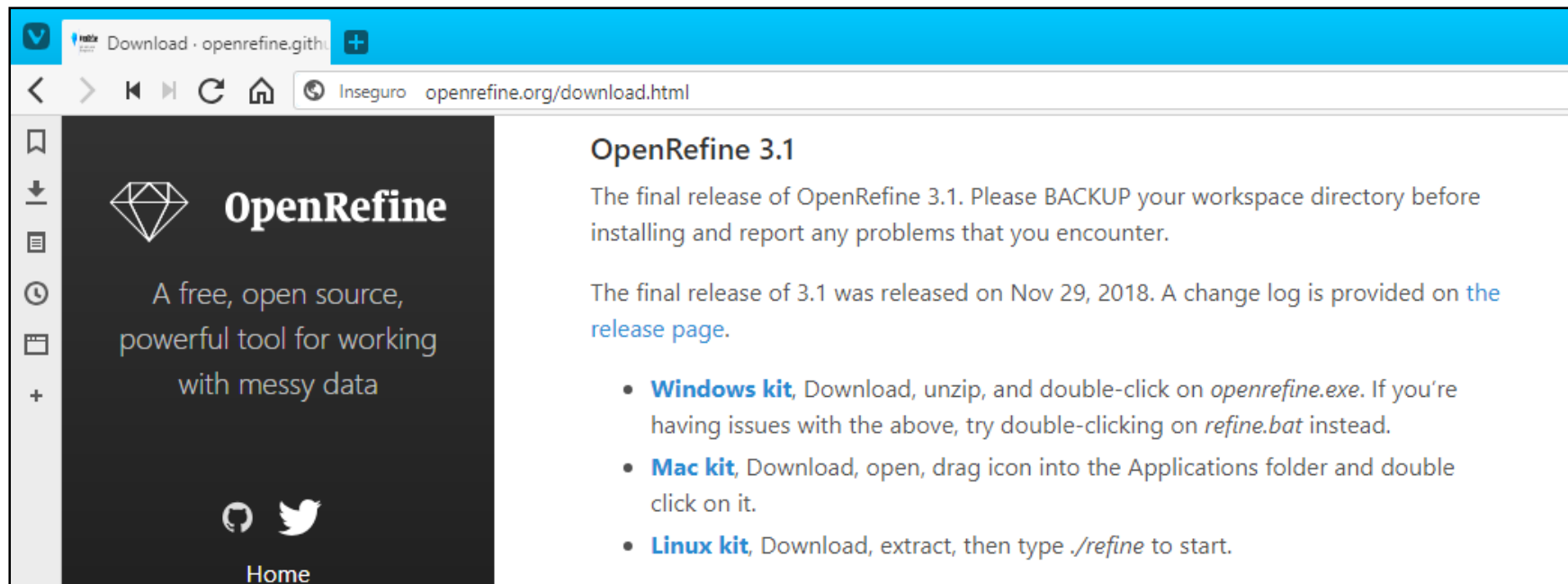
## OpenRefine

O OpenRefine, anteriormente chamado de Google Refine e, antes disso, o Freebase Gridworks, é um aplicativo de desktop (apesar de rodar em um navegador) de código aberto autônomo para limpeza de dados e transformação para outros formatos. É semelhante a aplicativos de planilhas (e pode trabalhar com formatos de arquivo de planilha). No entanto, ele se comporta mais como um banco de dados.

# Limpeza de dados

## Realizando download do OpenRefine

Para baixar o OpenRefine, acesse o endereço <http://openrefine.org/download.html>, escolha uma das opções de acordo com seu sistema operacional. Para esta aula vou utilizar a versão 3.1 para Windows (openrefine-win-3.1.zip).



# Limpeza de dados

## Instalando o OpenRefine

Para Windows, faça o download, descompacte o arquivo e execute o `openrefine.exe` para abrir o sistema.

Para Mac, faça o download, abra o arquivo e arraste o ícone para a pasta Applications e dê um duplo clique para abrir o sistema.

Para Linux, faça o download, descompacte e execute o arquivo `refine (./refine)` para abrir o sistema.

# Limpeza de dados

## Java é necessário

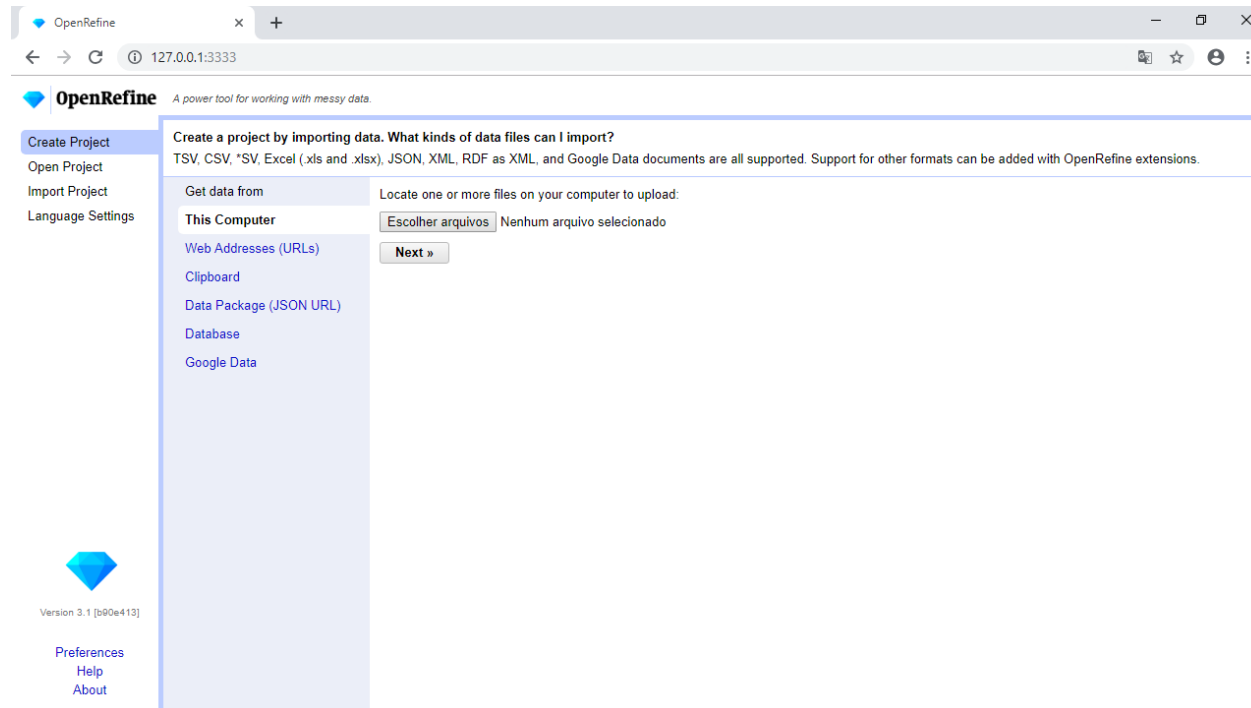
O OpenRefine precisa do Java, caso não tenha instalado ele vai abrir a página [https://www.java.com/pt\\_BR/download/](https://www.java.com/pt_BR/download/) para instalação.



# Limpeza de dados

## Executando o OpenRefine

Ao executar o OpenRefine, o mesmo será aberto no seu navegador padrão.



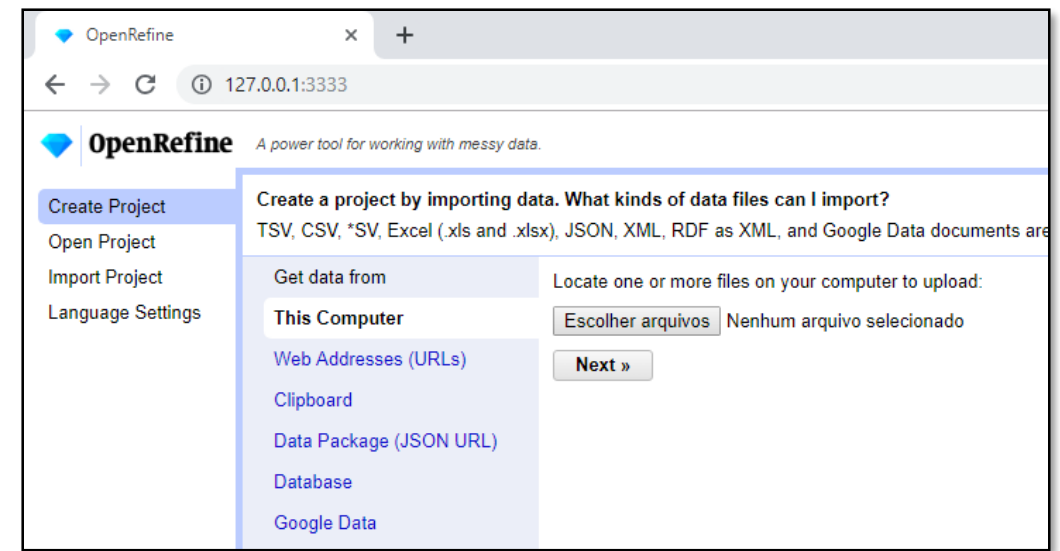


# Limpeza de dados

## Criando o projeto com o arquivo de exemplo

Clique no botão “Escolher arquivos” (na opção Create Project/ This Computer) e selecione o arquivo comparacao\_editores\_texto.csv (disponibilizado junto a aula) que contém uma lista de editores de texto disponível no endereço [https://en.wikipedia.org/wiki/Comparison\\_of\\_text\\_editors](https://en.wikipedia.org/wiki/Comparison_of_text_editors).

Tipos de arquivos que podem ser importados:  
TSV, CSV, \*SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, e documentos Google Data  
São todos suportados. Suporte a outros formatos, podem ser adicionados usando extensões.



# Limpeza de dados

## Configurando opções de parse

Após selecionar o arquivo, clique em “Next” para visualizar a tela de configuração de parse. Nesta tela podemos configurar os detalhes de importação dos dados.

Nome do projeto

Project name: comparacao\_editores\_texto.csv

Create Project

	Name	Creator	First public release	Latest stable version	Latest Release Date	Programming language	Cost (US\$)	Software license	Open source	CLI available
1.	Acme	Rob Pike	1993	Plan 9 and Inferno		C	Free	LPL (OSI approved)	Yes	
2.	AkelPad	Alexey Kuznetsov, Alexander Shengalts	2003	4.9.8	18/07/2016	C	Free	BSD	Yes	
3.	Alphatk	Vince Darley	1999	8.3.3	10/12/2004		\$40	Proprietary, with BSD components	No	
4.	Aquamacs	David Reitter	2005	3.3	20/09/2016	C, Emacs Lisp	Free	GPL	Yes	
5.	Atom	GitHub	2014	1.31.1	28/09/2018	HTML, CSS, JavaScript, C++	Free	MIT	Yes	No
6.	BBEdit	Rich Siegel	1992	12.1.3	11/04/2018	Objective-C, Objective-C++	\$49.99	Proprietary	No	
7.	Bluefish	Bluefish Development Team	1999	2.2.10	27/01/2017	C	Free	GPL	Yes	

Parse data as

CSV / TSV / separator-based files

Line-based text files

Fixed-width field text files

PC-Axis text files

JSON files

MARC files

JSON-LD files

RDF/N3 files

RDF/N-Triples files

RDF/Turtle files

Character encoding

Columns are separated by

☐ commas (CSV)

☐ tabs (TSV)

☒ custom: ;

Escape special characters with \

☐ Column names (comma separated):

☐ 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

☒ Use character " to enclose cells containing column separators

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

☒ Store blank rows

☒ Store blank cells as nulls

☐ Store file source (file names, URLs) in each row

Update Preview

Ao término clique em Create Project

É possível modificar as configurações e clicar em Update Preview para atualizar a exibição dos dados.

# Limpeza de dados

## Usando o OpenRefine

### Tela do OpenRefine com os dados

The screenshot shows the OpenRefine web interface in a browser window. The address bar displays the URL `127.0.0.1:3333/project?project=2224279130030`. The interface includes a top navigation bar with 'Open...', 'Export', and 'Help' buttons. Below this, a sidebar on the left contains a 'Facet / Filter' section with a 'Using facets and filters' tip. The main area displays a table of 75 records, with the first 10 visible. The table has columns for 'All', 'Name', 'Creator', 'First public release', 'Latest stable version', 'Latest Release Date', 'Programming language', 'Cost (US\$)', and 'Software license'. The records are sorted by 'Name' and show details for various text editors like Acme, AkelPad, Alphasoft, Aquamacs, Atom, BBEdit, Bluefish, Brackets, Coda, and ConTEXT.

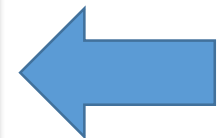
	All	Name	Creator	First public release	Latest stable version	Latest Release Date	Programming language	Cost (US\$)	Software license
1.	★	Acme	Rob Pike	1993	Plan 9 and Inferno		C	Free	LPL (OSI approved)
2.	★	AkelPad	Alexey Kuznetsov, Alexander Shengalts	2003	4.9.8	18/07/2016	C	Free	BSD
3.	★	Alphasoft	Vince Darley	1999	8.3.3	10/12/2004		\$40	Proprietary, with BSD components
4.	★	Aquamacs	David Reitter	2005	3.3	20/09/2016	C, Emacs Lisp	Free	GPL
5.	★	Atom	GitHub	2014	1.31.1	28/09/2018	HTML, CSS, JavaScript, C++	Free	MIT
6.	★	BBEdit	Rich Siegel	1992	12.1.3	11/04/2018	Objective-C, Objective-C++	\$49.99	Proprietary
7.	★	Bluefish	Bluefish Development Team	1999	2.2.10	27/01/2017	C	Free	GPL
8.	★	Brackets	Adobe Systems	2012	1.12	05/02/2018	HTML, CSS, JavaScript, C++	Free	MIT
9.	★	Coda	Panic	2007	2.6.6	05/06/2017	Objective-C	\$99	Proprietary
10.	★	ConTEXT	ConTEXT Project Ltd	1999	0.98.6	14/08/2009	Object Pascal (Delphi)	Free	BSD

# Limpeza de dados

## Usando o OpenRefine

Em todas as colunas podemos ver uma seta, onde existem opções para filtrar, classificar, transformar ou remover os dados.

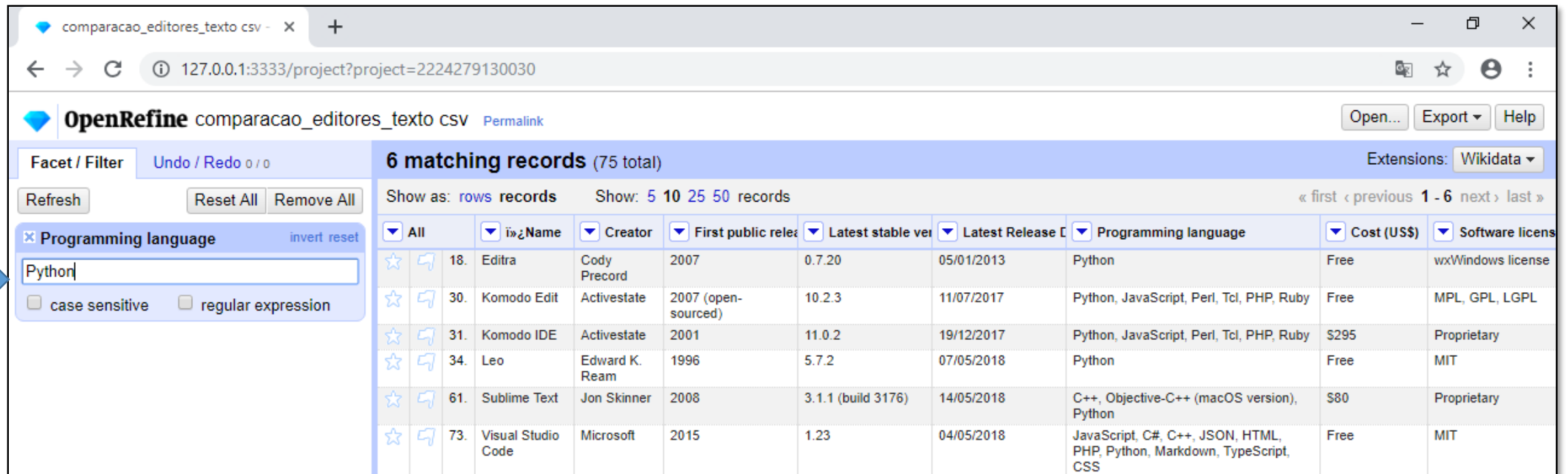
75 records							
Show as: <b>rows</b> records		Show: 5 10 25 50 records					
<input type="checkbox"/> All	<input type="checkbox"/> i»zName	<input type="checkbox"/> Creator	<input type="checkbox"/> First public relea	<input type="checkbox"/> Latest stable ver	<input type="checkbox"/> Latest Release C	<input type="checkbox"/>	<input type="checkbox"/>
☆	1.	Acme	Rob Pike	1993	Plan 9 and Inferno		C
☆	2.	AkelPad	Alexey Kuznetsov, Alexander Shengalts	2003	4.9.8	18/07/2016	C
☆	3.	Alphatk	Vince Darley	1999	8.3.3	10/12/2004	
☆	4.	Aquamacs	David Reitter	2005	3.3	20/09/2016	C, I
☆	5.	Atom	GitHub	2014	1.31.1	28/09/2018	HT, Jav
☆	6.	BEdit	Rich Siegel	1992	12.1.3	11/04/2018	Obj, Obj
☆	7.	Bluefish	Bluefish Development Team	1999	2.2.10	27/01/2017	C
☆	8.	Brackets	Adobe Systems	2012	1.12	05/02/2018	HT



# Limpeza de dados

## Filtragem

Podemos usar filtros e facetar para filtrar os dados apresentados. Na elaboração de um filtro, podemos utilizar expressões regulares, por exemplo, podemos mostrar apenas os registros que contenham Python na coluna “Programming Language”.



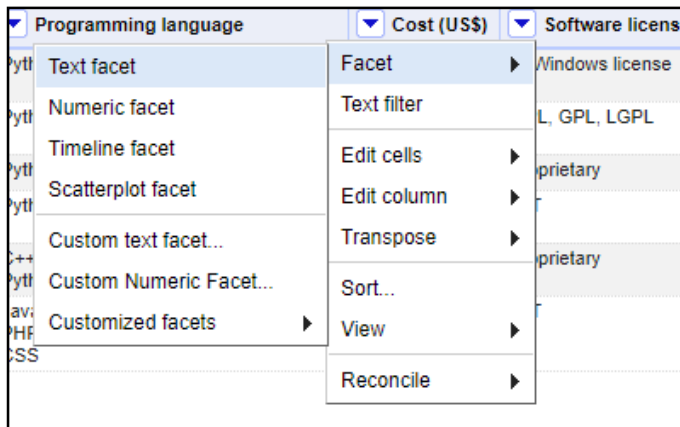
The screenshot shows the OpenRefine web interface. The browser address bar displays `127.0.0.1:3333/project?project=2224279130030`. The interface title is "comparacao\_editores\_texto csv". On the left, the "Facet / Filter" panel shows a filter for "Programming language" with the value "Python" entered. Below the input field are checkboxes for "case sensitive" and "regular expression". The main table area shows "6 matching records (75 total)". The table has columns: "All", "i»zName", "Creator", "First public release", "Latest stable version", "Latest Release Date", "Programming language", "Cost (US\$)", and "Software license". The table lists 6 records, each with a star icon, a speech bubble icon, and a row number. A blue arrow points to the "Python" filter input.

	All	i»zName	Creator	First public release	Latest stable version	Latest Release Date	Programming language	Cost (US\$)	Software license
18.	☆	Editra	Cody Precord	2007	0.7.20	05/01/2013	Python	Free	wxWindows license
30.	☆	Komodo Edit	Activestate	2007 (open-sourced)	10.2.3	11/07/2017	Python, JavaScript, Perl, Tcl, PHP, Ruby	Free	MPL, GPL, LGPL
31.	☆	Komodo IDE	Activestate	2001	11.0.2	19/12/2017	Python, JavaScript, Perl, Tcl, PHP, Ruby	\$295	Proprietary
34.	☆	Leo	Edward K. Ream	1996	5.7.2	07/05/2018	Python	Free	MIT
61.	☆	Sublime Text	Jon Skinner	2008	3.1.1 (build 3176)	14/05/2018	C++, Objective-C++ (macOS version), Python	\$80	Proprietary
73.	☆	Visual Studio Code	Microsoft	2015	1.23	04/05/2018	JavaScript, C#, C++, JSON, HTML, PHP, Python, Markdown, TypeScript, CSS	Free	MIT

# Limpeza de dados

## Filtragem

As facetas são utilizadas na inclusão ou exclusão dos dados com base no conteúdo inteiro da coluna. Clique na seta da coluna “Software License”, escolha “Facet” e “Text facet” conforme abaixo.







2 matching records (75 total)

Extensions: Wikidata

Show as: rows records

Show: 5 10 25 50 records

« first < previous 1 - 2 next > last »

<input type="checkbox"/> All	<input type="checkbox"/> Name	<input type="checkbox"/> Creator	<input type="checkbox"/> First public release	<input type="checkbox"/> Latest stable version	<input type="checkbox"/> Latest Release Date	<input type="checkbox"/> Programming language	<input type="checkbox"/> Cost (US\$)	<input type="checkbox"/> Software license
  34.	Leo	Edward K. Ream	1996	5.7.2	07/05/2018	Python	Free	MIT
  73.	Visual Studio Code	Microsoft	2015	1.23	04/05/2018	JavaScript, C#, C++, JSON, HTML, PHP, Python, Markdown, TypeScript, CSS	Free	MIT

# Limpeza de dados

## Limpeza

A transformação de dados é executada no OpenRefine com o uso da OpenRefine Expression Language (GREL – Google Refine Expression Language). Essa linguagem é usada para criar funções lambda curtas que transformam os valores das células de acordo com regras simples.

# Limpeza de dados

## Limpeza

Veja por exemplo que a coluna “First public release” em alguns casos não possuem apenas 4 dígitos para o ano.

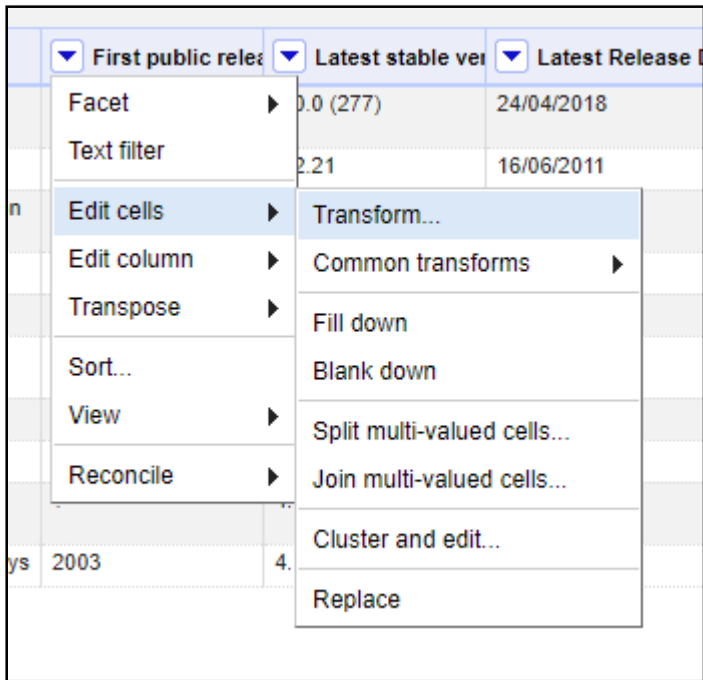
75 records						
Show as: <a href="#">rows</a> <a href="#">records</a>			Show: <a href="#">5</a> <a href="#">10</a> <a href="#">25</a> <a href="#">50</a> records			
<input type="checkbox"/> All	<input type="checkbox"/> Name	<input type="checkbox"/> Creator	<input type="checkbox"/> First public release	<input type="checkbox"/> Latest		
<input type="checkbox"/>	<input type="checkbox"/>	51. PSPad	Jan Fiala	2002	5.0.0 (277)	
<input type="checkbox"/>	<input type="checkbox"/>	52. Q10	Baara Estudio	2007	1.2.21	
<input type="checkbox"/>	<input type="checkbox"/>	53. RJ TextEd	Rickard Johansson	2004	13.10	
<input type="checkbox"/>	<input type="checkbox"/>	54. RText	Fifesoft	2003	2.6.3	
<input type="checkbox"/>	<input type="checkbox"/>	55. Sam	Rob Pike	1980s (early)	stable	
<input type="checkbox"/>	<input type="checkbox"/>	56. SciTE	Neil Hodgson	1999	4.0.5	
<input type="checkbox"/>	<input type="checkbox"/>	57. SlickEdit	SlickEdit, Inc.	1988	23.0.0	
<input type="checkbox"/>	<input type="checkbox"/>	58. Smultron	Peter Borg	2004	9.2.3	
<input type="checkbox"/>	<input type="checkbox"/>	59. Source Insight	Source Dynamics	?	4.0.0084	
<input type="checkbox"/>	<input type="checkbox"/>	60. SubEthaEdit	TheCodingMonkeys	2003	4.1	



# Limpeza de dados

# Limpeza

Clique na seta da coluna “First public release” e escolha “Edit cells” e “Tranform”. Na janela que se abre temos a caixa “Expression” para escrevermos a expressão a ser utilizada. Enquanto digitamos o resultado é exibido automaticamente em “Preview”.



# Limpeza de dados

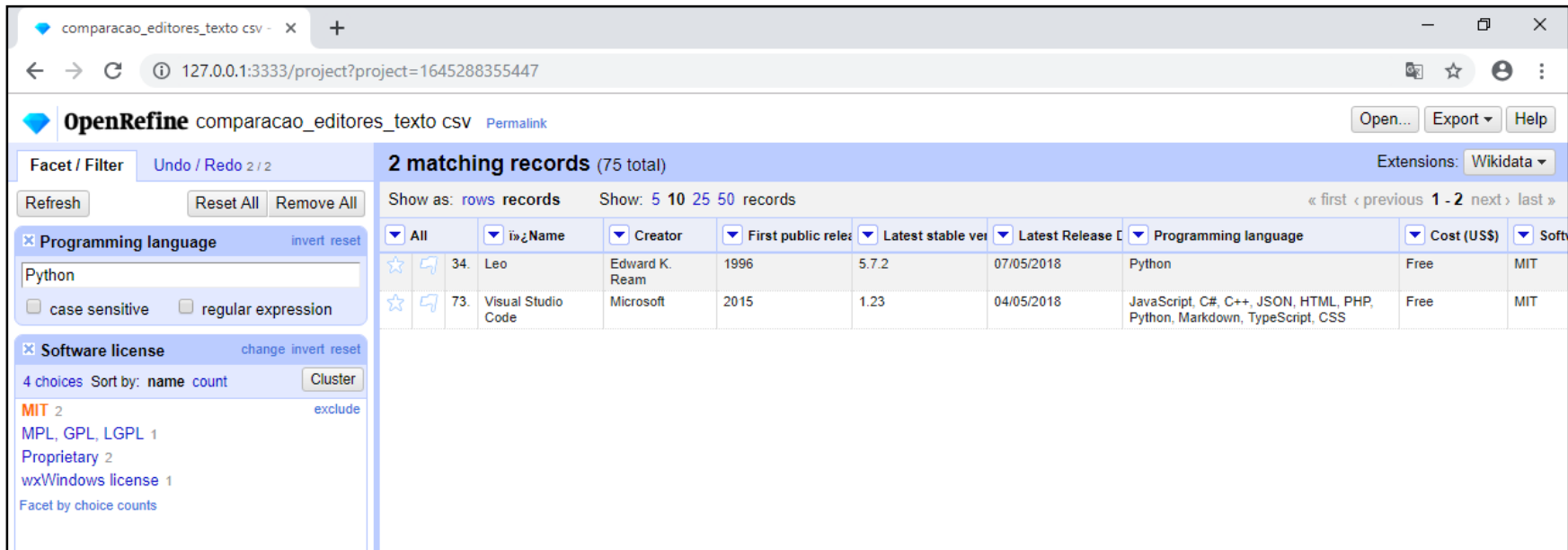
## Limpeza

Utilize a seguinte expressão:

```
value.match(".*([0-9]{4}).*").get(0)
```

Esta expressão encontra todas ocorrências sucessivas de quatro decimais e retorna a primeira.

Quando não encontrar uma correspondência, retorna null.



comparacao\_editores\_texto csv

127.0.0.1:3333/project?project=1645288355447

OpenRefine comparacao\_editores\_texto csv

Open... Export Help

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

2 matching records (75 total)

Show as: rows records Show: 5 10 25 50 records « first < previous 1 - 2 next > last »

		Name	Creator	First public release	Latest stable version	Latest Release Date	Programming language	Cost (US\$)	Software license
34.	Leo	Edward K. Ream	1996	5.7.2	07/05/2018	Python	Free	MIT	
73.	Visual Studio Code	Microsoft	2015	1.23	04/05/2018	JavaScript, C#, C++, JSON, HTML, PHP, Python, Markdown, TypeScript, CSS	Free	MIT	

Programming language invert reset

Python

case sensitive regular expression

Software license change invert reset

4 choices Sort by: name count Cluster

MIT 2 exclude

MPL, GPL, LGPL 1

Proprietary 2

wxWindows license 1

Facet by choice counts

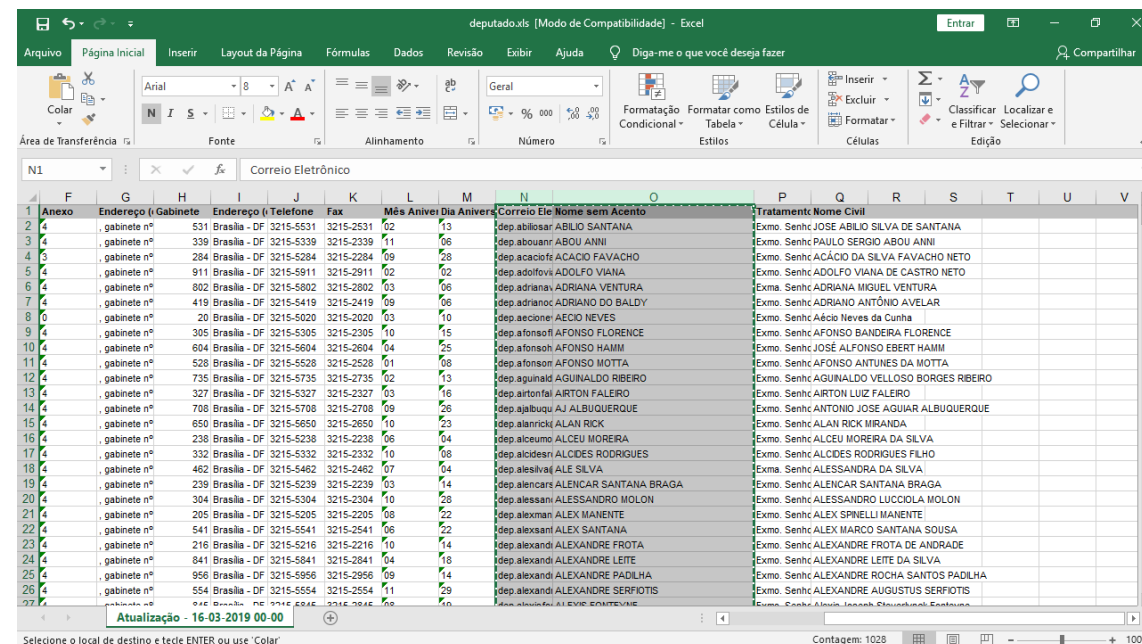
# Limpeza de dados

## Criando projeto com dados da área de transferência (clipboard)

Vamos ao Google procurar uma planilha em sites .gov.br informando “site:gov.br type:xls”.

O arquivo encontrado com a pesquisa e utilizado é o arquivo a seguir (que será disponibilizado junto a aula): <http://www.camara.gov.br/Internet/Deputado/deputado.xls>

Abra este arquivo, selecione as colunas “Correio eletrônico” e “nome sem acento” e copie seu conteúdo.

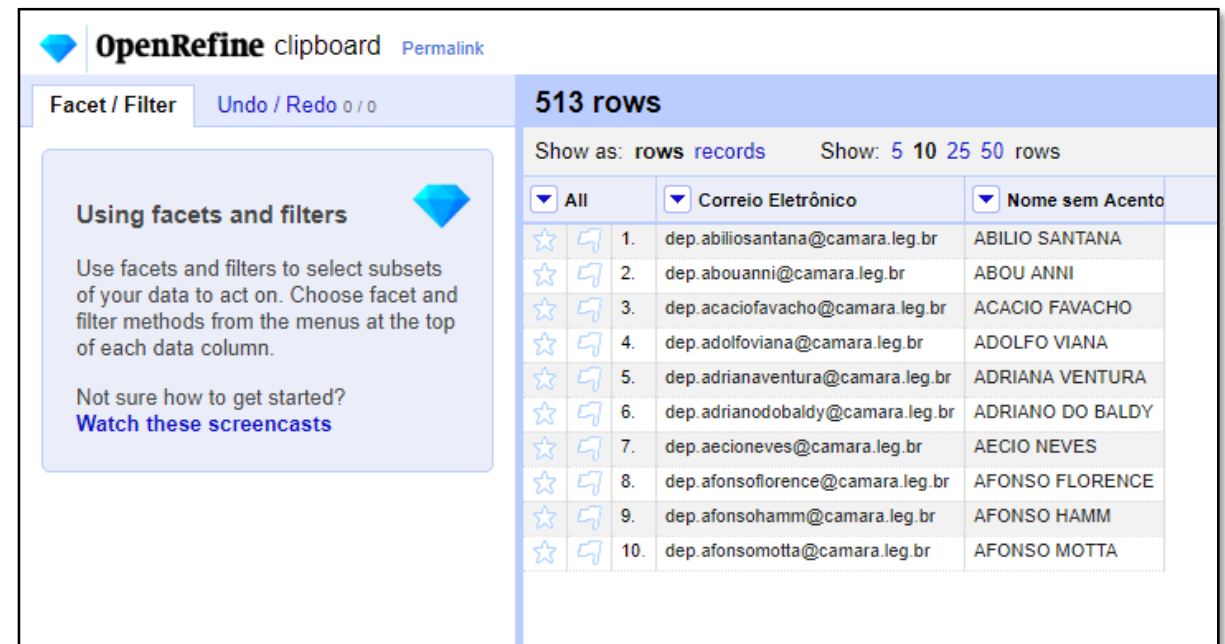


	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V
	Anexo	Endereço (Gabinete)	Endereço (Telefone)	Fax	Mês Aniversário	Nome sem Acento	Correio Eletrônico	Tratamento	Nome Civil								
1																	
2	4	.gabinete nº	531 Brasília - DF	3215-5531	02	13	dep.abilioar ABLIO SANTANA	Exmo. Senh	JOSE ABLIO SILVA DE SANTANA								
3	4	.gabinete nº	339 Brasília - DF	3215-5339	11	06	dep.abouanr ABOU ANNI	Exmo. Senh	PAULO SERGIO ABOU ANNI								
4	3	.gabinete nº	284 Brasília - DF	3215-5284	09	28	dep.acaciof ACACIO FAVACHO	Exmo. Senh	ACÁCIO DA SILVA FAVACHO NETO								
5	4	.gabinete nº	911 Brasília - DF	3215-5911	02	02	dep.adolfov ADOLFO VIANA	Exmo. Senh	ADOLFO VIANA DE CASTRO NETO								
6	4	.gabinete nº	802 Brasília - DF	3215-5802	03	06	dep.adrianav ADRIANA VENTURA	Exma. Senh	ADRIANA MIGUEL VENTURA								
7	4	.gabinete nº	419 Brasília - DF	3215-5419	09	06	dep.adrianoc ADRIANO DO BALDY	Exmo. Senh	ADRIANO ANTÔNIO AVELAR								
8	0	.gabinete nº	20 Brasília - DF	3215-5020	03	10	dep.aecionei AECIO NEVES	Exmo. Senh	Aécio Neves da Cunha								
9	4	.gabinete nº	305 Brasília - DF	3215-5305	02	15	dep.afonsoff AFONSO FLORENCE	Exmo. Senh	AFONSO BANDERA FLORENCE								
10	4	.gabinete nº	604 Brasília - DF	3215-5604	04	25	dep.afonsoff AFONSO HAMM	Exmo. Senh	JOSÉ ALFONSO EBERT HAMM								
11	4	.gabinete nº	528 Brasília - DF	3215-5528	01	08	dep.afonsof AFONSO MOTA	Exmo. Senh	AFONSO ANTUNES DA MOTA								
12	4	.gabinete nº	735 Brasília - DF	3215-5735	02	13	dep.aguinald AGUINALDO RIBEIRO	Exmo. Senh	AGUINALDO VELLOSO BORGES RIBEIRO								
13	4	.gabinete nº	327 Brasília - DF	3215-5327	03	16	dep.airetonf AIRTON FALERO	Exmo. Senh	AIRTON LUIZ FALERO								
14	4	.gabinete nº	708 Brasília - DF	3215-5708	09	26	dep.albuquaj AJ ALBUQUERQUE	Exmo. Senh	ANTONIO JOSE AGUIAR ALBUQUERQUE								
15	4	.gabinete nº	650 Brasília - DF	3215-5650	10	03	dep.alanrick ALAN RICK	Exmo. Senh	ALAN RICK MIRANDA								
16	4	.gabinete nº	238 Brasília - DF	3215-5238	06	04	dep.alceum ALCEU MOREIRA	Exmo. Senh	ALCEU MOREIRA DA SILVA								
17	4	.gabinete nº	332 Brasília - DF	3215-5332	10	08	dep.alcidesr ALCIDES RODRIGUES	Exmo. Senh	ALCIDES RODRIGUES FILHO								
18	4	.gabinete nº	462 Brasília - DF	3215-5462	07	04	dep.alesivaj ALE SILVA	Exma. Senh	ALESSANDRA DA SILVA								
19	4	.gabinete nº	239 Brasília - DF	3215-5239	03	14	dep.alencars ALENCAR SANTANA BRAGA	Exmo. Senh	ALENCAR SANTANA BRAGA								
20	4	.gabinete nº	304 Brasília - DF	3215-5304	10	28	dep.alessani ALESSANDRO MOLON	Exmo. Senh	ALESSANDRO LUCIOLA MOLON								
21	4	.gabinete nº	205 Brasília - DF	3215-5205	08	22	dep.alexman ALEX MANENTE	Exmo. Senh	ALEX SPINELLI MANENTE								
22	4	.gabinete nº	541 Brasília - DF	3215-5541	06	22	dep.alexsan ALEX SANTANA	Exmo. Senh	ALEX MARCO SANTANA SOUSA								
23	4	.gabinete nº	216 Brasília - DF	3215-5216	10	14	dep.alexandri ALEXANDRE FROTA	Exmo. Senh	ALEXANDRE FROTA DE ANDRADE								
24	4	.gabinete nº	841 Brasília - DF	3215-5841	04	18	dep.alexandri ALEXANDRE LEITE	Exmo. Senh	ALEXANDRE LEITE DA SILVA								
25	4	.gabinete nº	956 Brasília - DF	3215-5956	09	14	dep.alexandri ALEXANDRE PADILHA	Exmo. Senh	ALEXANDRE ROCHA SANTOS PADILHA								
26	4	.gabinete nº	554 Brasília - DF	3215-5554	11	29	dep.alexandri ALEXANDRE SERFIOTIS	Exmo. Senh	ALEXANDRE AUGUSTUS SERFIOTIS								
27	4	.gabinete nº	316 Brasília - DF	3215-5316	06	10	dep.alexandri ALEXANDRE SERFIOTIS	Exmo. Senh	ALEXANDRE AUGUSTUS SERFIOTIS								

# Limpeza de dados

## Criando projeto com dados da área de transferência (clipboard)

No OpenRefine, em “Clipboard”, cole o conteúdo e clique em “Next”, depois em “Create Project”.



# Limpeza de dados

## Mais informações

Mais informações sobre o OpenRefine podem ser obtidas em <https://github.com/OpenRefine/OpenRefine/wiki/Documentation-For-Users>

# FIM