

**НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»**

Работа по очистке набора данных при помощи Open Refine

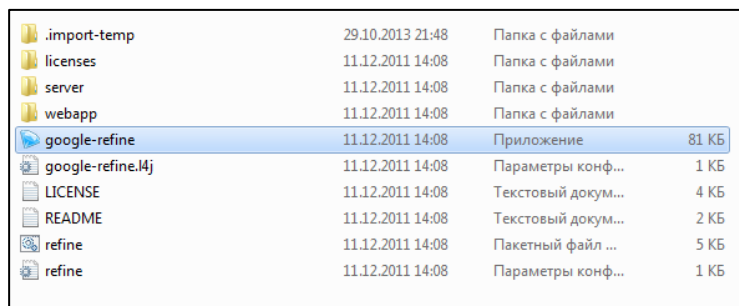
Содержание:

1. Загрузка Open Refine.....	2
2. Создание нового проекта и загрузка в него набора данных.....	2
3. Очистка данных в поле country.....	4
4. Очистка данных в поле numUndergrad и поле numStudents	8
5. Дополнительная функциональность Open Refine	11

1. Загрузка Open Refine

Зайдите в каталог **D:\GR**

Для того чтобы загрузить Open Refine, необходимо дважды щелкнуть мышью по файлу **google-refine** (рис. 1). Open Refine загрузится в браузере.



.import-temp	29.10.2013 21:48	Папка с файлами	
licenses	11.12.2011 14:08	Папка с файлами	
server	11.12.2011 14:08	Папка с файлами	
webapp	11.12.2011 14:08	Папка с файлами	
google-refine	11.12.2011 14:08	Приложение	81 КБ
google-refine.l4j	11.12.2011 14:08	Параметры конф...	1 КБ
LICENSE	11.12.2011 14:08	Текстовый докум...	4 КБ
README	11.12.2011 14:08	Текстовый докум...	2 КБ
refine	11.12.2011 14:08	Пакетный файл ...	5 КБ
refine	11.12.2011 14:08	Параметры конф...	1 КБ

Рис. 1 — Загрузка Open Refine

В появившемся окне необходимо закрыть Google Chrome Frame. Для этого нужно нажать на кнопку **Close** в правом верхнем углу (рис. 2).

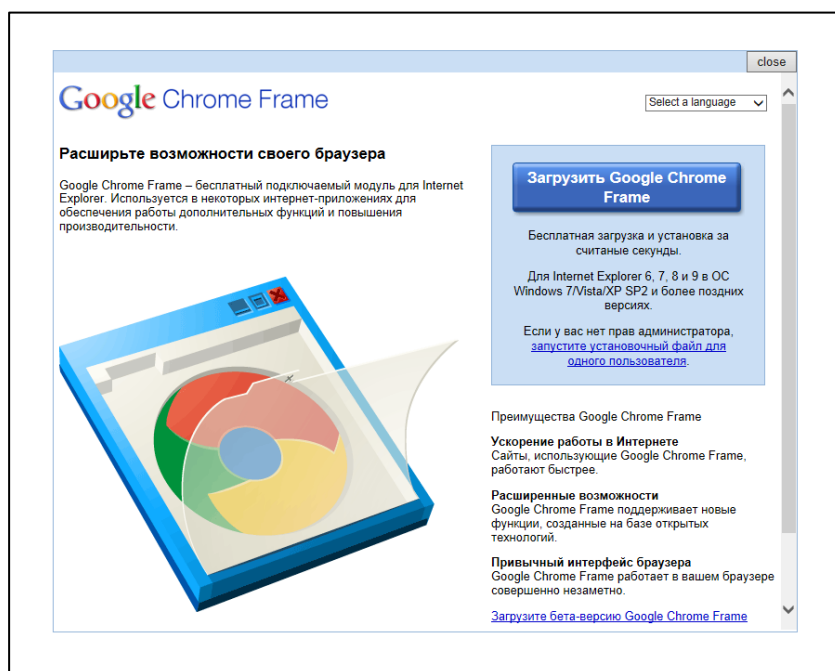


Рис. 2 — Закрыть Google Chrome Frame

2. Создание нового проекта и загрузка в него набора данных

Для того чтобы загрузить набор данных в новый проект удобно воспользоваться опцией **This Computer**, в которой необходимо указать расположение загружаемого набора данных (**D:**) при помощи стандартного

диалогового меню, которое отображается после нажатия на кнопку **Обзор...** (рис. 3 и рис. 4).

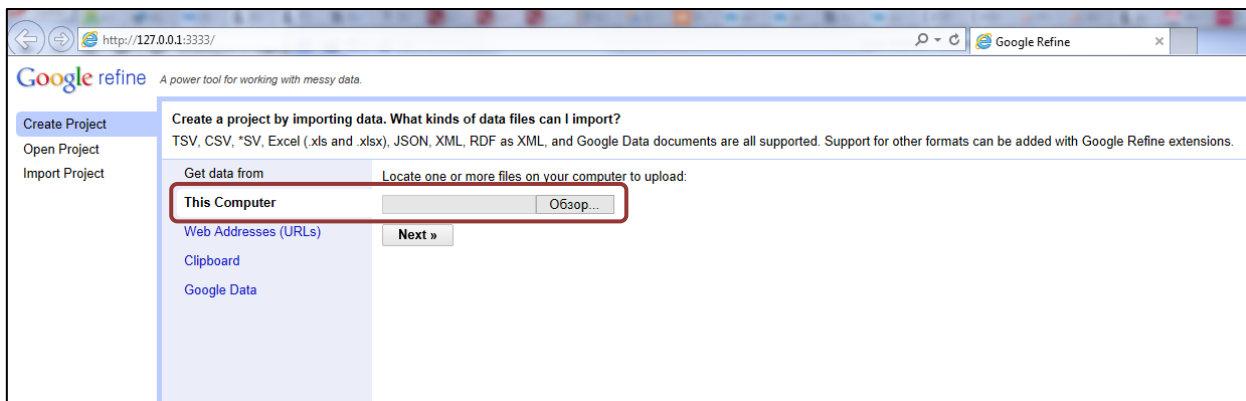


Рис. 3 — Создание нового проекта

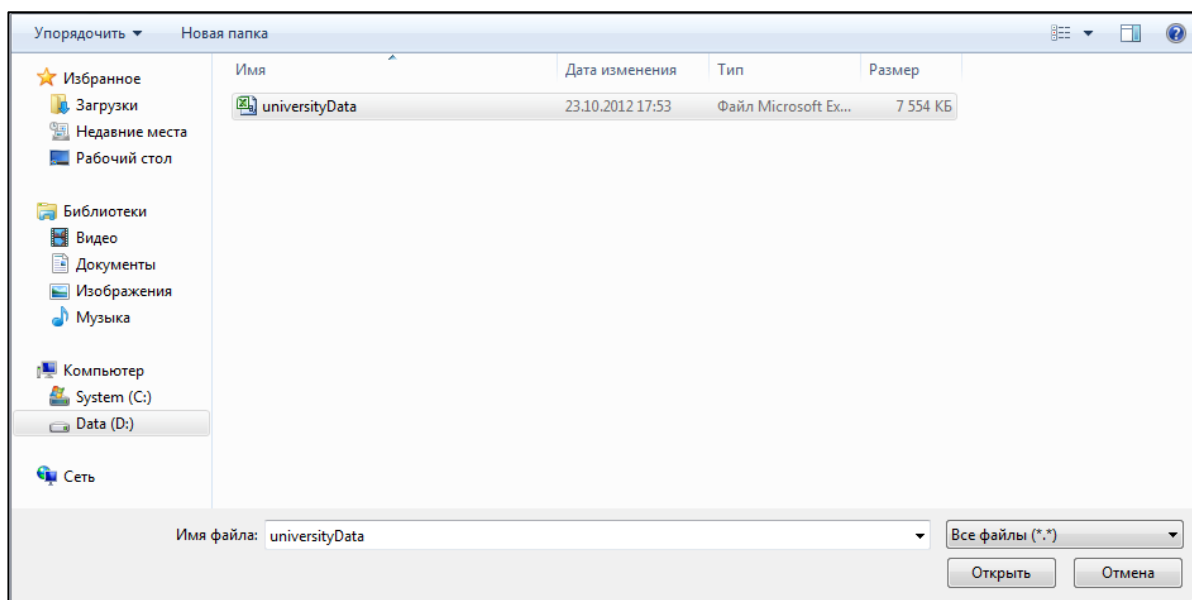


Рис. 4 — Выбор набора данных для загрузки

Выбрать загружаемый набор данных (**universityData.csv**) и затем надо нажать на кнопку **Next** (рис. 5).

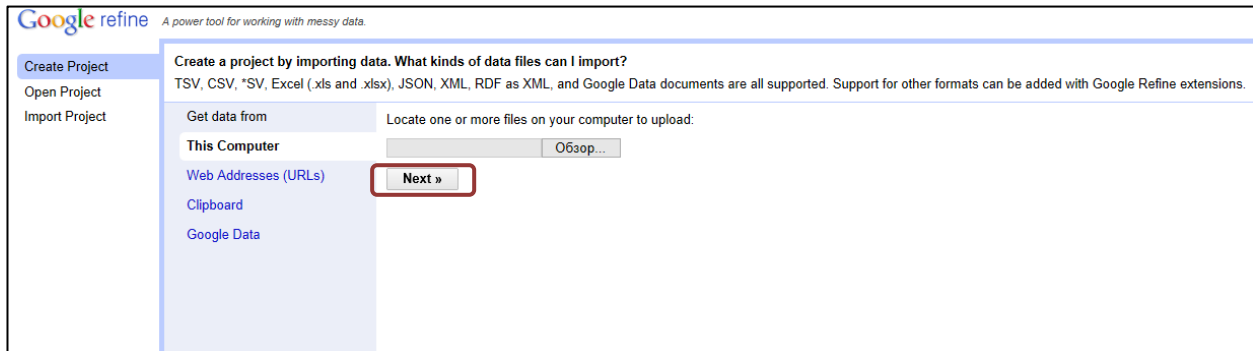


Рис. 5 — Загрузка набора данных

Произойдет загрузка набора данных в новый проект. И затем надо нажать на кнопку **Create Project >>** в верхнем правом углу браузера (рис. 6).

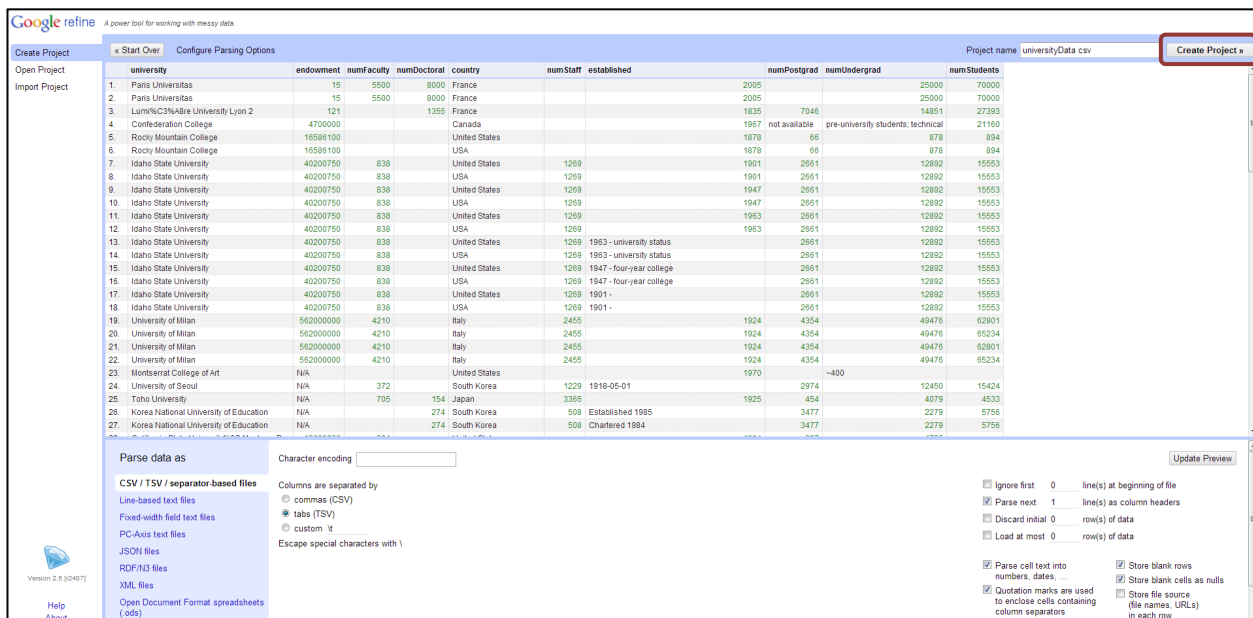



Рис. 6 — Завершение загрузки набора данных

3. Очистка данных в поле country

Данные в колонке **country** (Название страны) содержат различные варианты названия стран. Для того чтобы привести их к единому виду, необходимо щелкнуть мышью по кнопке , которая находится слева от названия колонки **country** и затем выбрать последовательно опции **Edit cells** → **Cluster and edit...** (рис. 7)

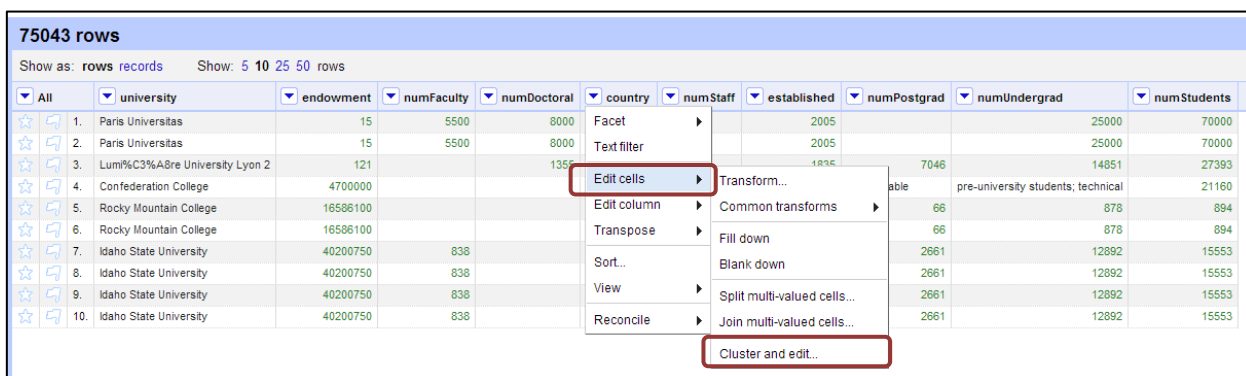


Рис. 7 — Выбор элементов меню для очистки данных в поле **country**

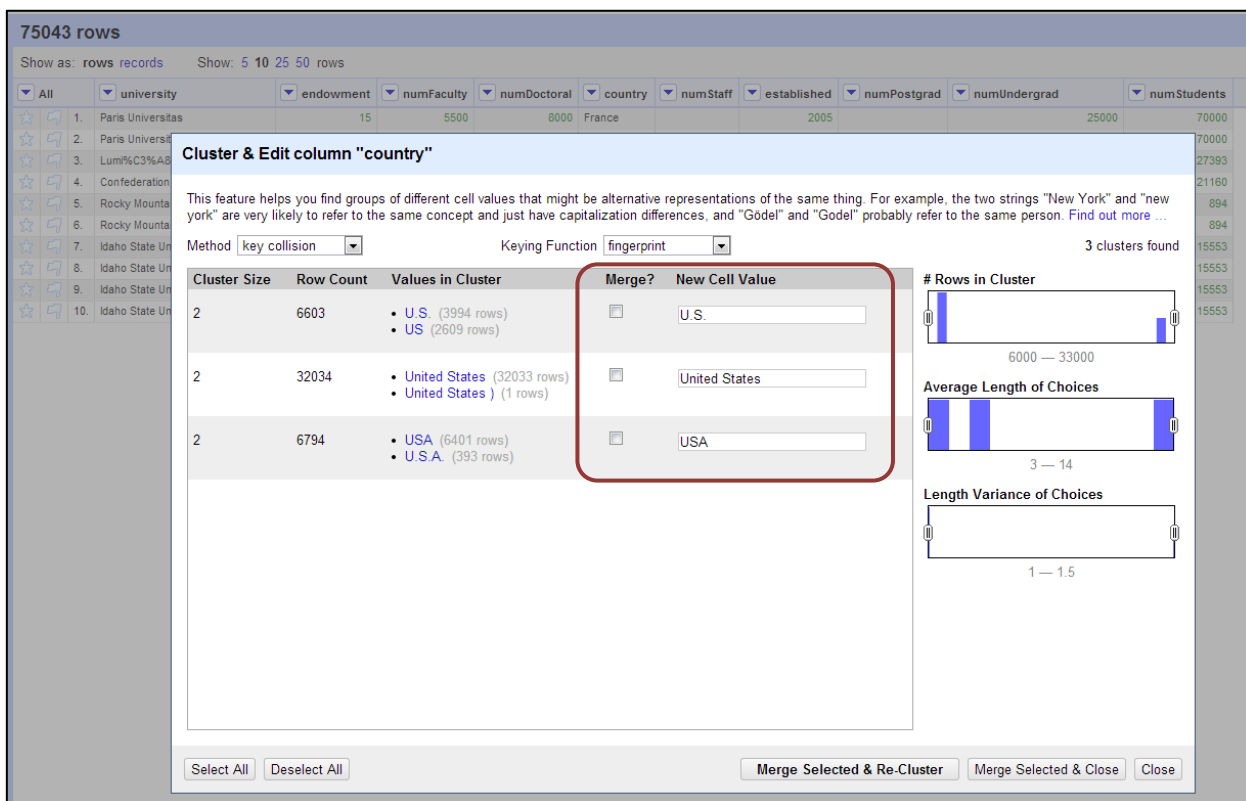


Рис. 8 — Очистка данных в поле **country**

75043 rows

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

▼ All ▼ university ▼ endowment ▼ numFaculty ▼ numDoctoral ▼ country ▼ numStaff ▼ established ▼ numPostgrad ▼ numUndergrad ▼ numStudents

1. Paris Universit 15 5500 8000 France 2005 25000 70000

2. Paris Univers 70000

3. Lum%C3%A8 27393

4. Confederation 21160

5. Rocky Mounta 894

6. Rocky Mounta 894

7. Idaho State Un 15553

8. Idaho State Un 15553

9. Idaho State Un 15553

10. Idaho State Un 15553

Cluster & Edit column "country"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Godel" and "Godel" probably refer to the same person. [Find out more ...](#)

Method: key collision Keying Function: fingerprint 3 clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	6603	<ul style="list-style-type: none"> U.S. (3994 rows) US (2609 rows) 	<input checked="" type="checkbox"/>	USA
2	32034	<ul style="list-style-type: none"> United States (32033 rows) United States) (1 rows) 	<input checked="" type="checkbox"/>	USA
2	6794	<ul style="list-style-type: none"> USA (6401 rows) U.S.A. (393 rows) 	<input checked="" type="checkbox"/>	USA

Rows in Cluster: 6000 — 33000

Average Length of Choices: 3 — 14

Length Variance of Choices: 1 — 1.5

Select All Deselect All Merge Selected & Re-Cluster Merge Selected & Close Close

Рис. 9 — Выбор элементов меню для очистки данных в поле **country**

Для наиболее полного поиска наименований, которые могут содержать различные варианты стран, надо выбрать **cologne-phonetic** в выпадающем списке **Key Function** (обратите внимание, что результат зависит от выбранного алгоритма: **fingerprint** и **cologne-phonetic**). И написать новое унифицированное название страны, которое требуется по смыслу (в данном примере: **USA**, **Russia** и **USA**).

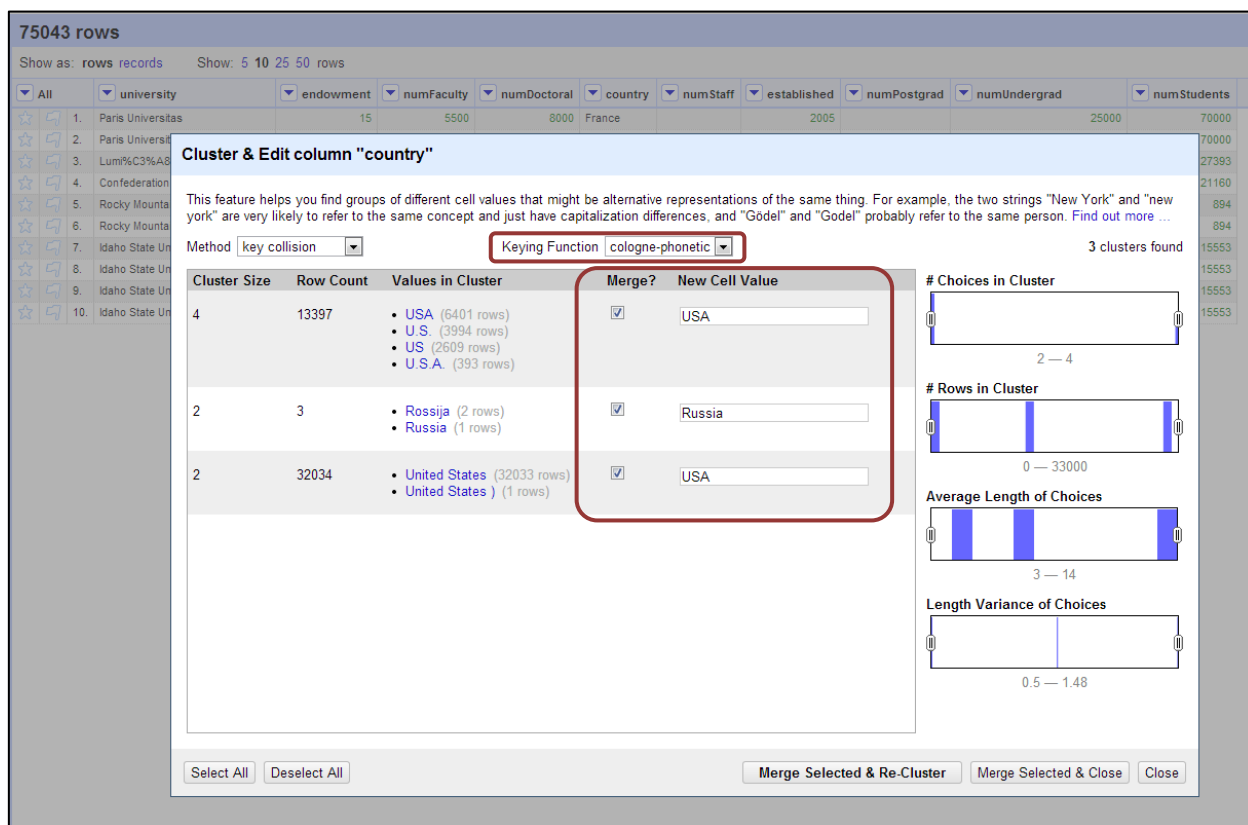


Рис. 10 — Выбор элементов меню для очистки данных в поле **country**

Затем необходимо нажать мышью на кнопку **Select All**, чтобы выделить все поля в колонке **Merge?** И после этого нажать на кнопку **Merge Selected & Close**.

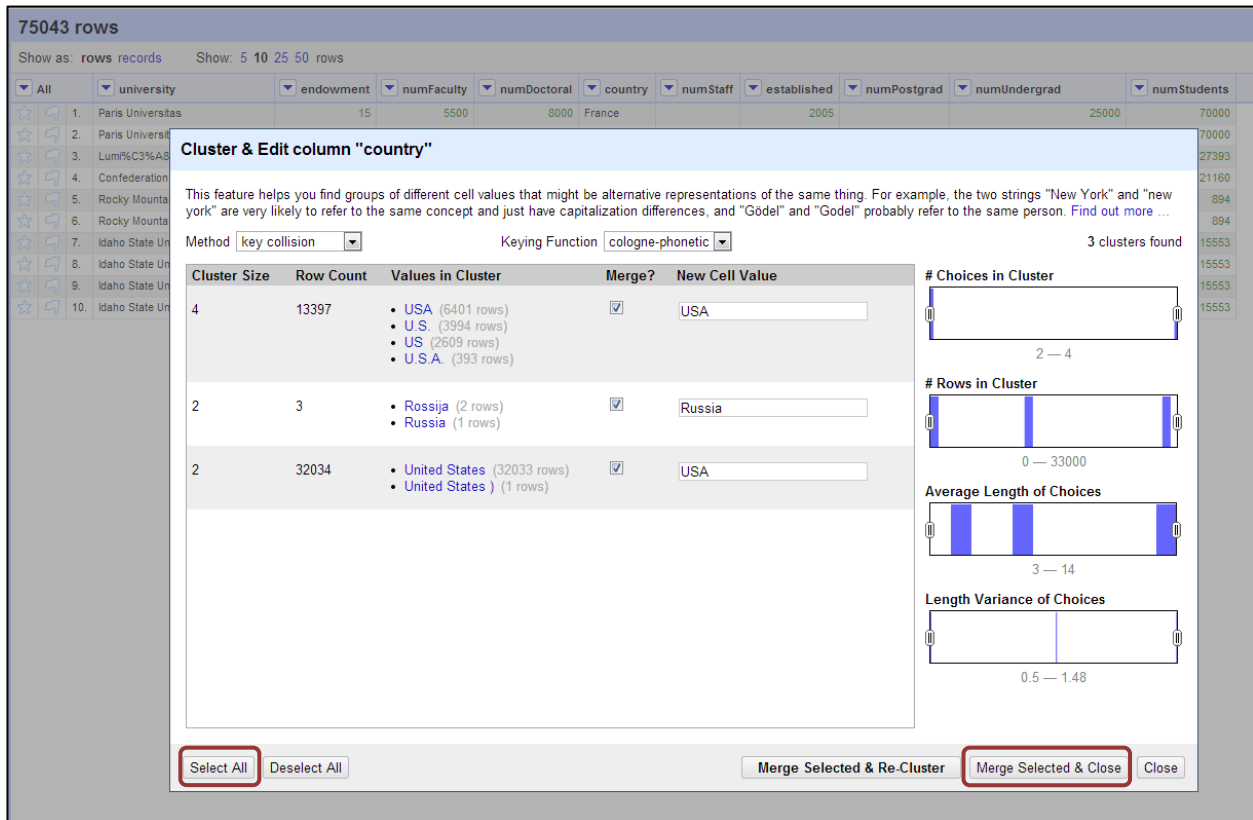


Рис. 11 — Очистка данных в поле **country**

Все различающиеся названия стран стали унифицированы.

4. Очистка данных в поле numUndergrad и поле numStudents

В этой колонке не все данные имеют числовой формат, многие поля содержат также текст в дополнение к числовым значениям. Эти данные необходимо исправить и привести к единому числовому виду.

Для этого нужно щелкнуть мышью по кнопке ▼, слева от названия колонки **numUndergrad** и затем выбрать последовательно опции **Facet -> NumericFacet** (рис. 12).

75043 rows

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

	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Paris Universitat	15	5500	8000	France		2005			
2.	Paris Universitat	15	5500	8000	France		2005			
3.	Lum%C3%A8re University Lyon 2	121		1355	France		1835	7048		
4.	Confederation College	4700000			Canada		1967	not available		
5.	Rocky Mountain College	16586100			USA		1878	66		
6.	Rocky Mountain College	16586100			USA		1878	66		
7.	Idaho State University	40200750	838		USA	1269	1901	2661		
8.	Idaho State University	40200750	838		USA	1269	1901	2661		
9.	Idaho State University	40200750	838		USA	1269	1947	2661		
10.	Idaho State University	40200750	838		USA	1269	1947	2661		

Facet menu for numUndergrad:

- Facet
 - Text facet
 - Numeric facet
 - Timeline facet
 - Scatterplot facet
 - Custom text facet...
 - Custom numeric facet...
 - Customized facets
- Text filter
- Edit cells
- Edit column
- Transpose
- Sort...
- View
- Reconcile

Рис. 12 — Выбор элементов меню для очистки данных в поле **numUndergrad**

На гистограмме слева отобразилось количество **числовых (Numeric)** и **нечисловых (Non-numeric)** записей (рис. 13).

Facet / Filter Undo / Redo

Refresh Reset All Remove All

numUndergrad change reset

0.00 — 3,200,000,000.00

☒ Numeric 69670 ☒ Non-numeric 5373 ☐ Blank 0 ☐ Error 0

75043 rows

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

	university	endowment	numFaculty	numDoctoral	country	numStaff	established	numPostgrad	numUndergrad	numStudents
1.	Paris Universitat	15	5500	8000	France		2005		25000	70000
2.	Paris Universitat	15	5500	8000	France		2005		25000	70000
3.	Lum%C3%A8re University Lyon 2	121		1355	France		1835	7048	14851	27393
4.	Confederation College	4700000			Canada		1967	not available	pre-university students, technical	21160
5.	Rocky Mountain College	16586100			USA		1878	66	878	894
6.	Rocky Mountain College	16586100			USA		1878	66	878	894
7.	Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
8.	Idaho State University	40200750	838		USA	1269	1901	2661	12892	15553
9.	Idaho State University	40200750	838		USA	1269	1947	2661	12892	15553
10.	Idaho State University	40200750	838		USA	1269	1947	2661	12892	15553

Рис. 13 — Очистка данных в поле **numUndergrad**

Для того чтобы избавиться от знаков «+» и «~» в поле **numUndergrad** необходимо последовательно выбрать опции в выпадающем меню **Edit cells -> Transform**. (рис. 14)

Custom text transform on column numUndergrad

Expression Language Google Refine Expression Language (GREL)

value No syntax error.

Preview History Starred Help

row	value	value
1.	25000	25000
2.	25000	25000
3.	14851	14851
4.	pre-university students, technical	pre-university students, technical
5.	878	878
6.	878	878

On error ☒ keep original ☐ set to blank ☐ store error

☐ Re-transform up to 10 times until no change

OK Cancel

Рис. 14 — Очистка данных в поле **numUndergrad**

Необходимо написать следующее выражение в текстовое поле **Expression:**

value.replace("+", "")

как это продемонстрировано на рис. 15. И затем щелкнуть мышью на кнопке **OK**.

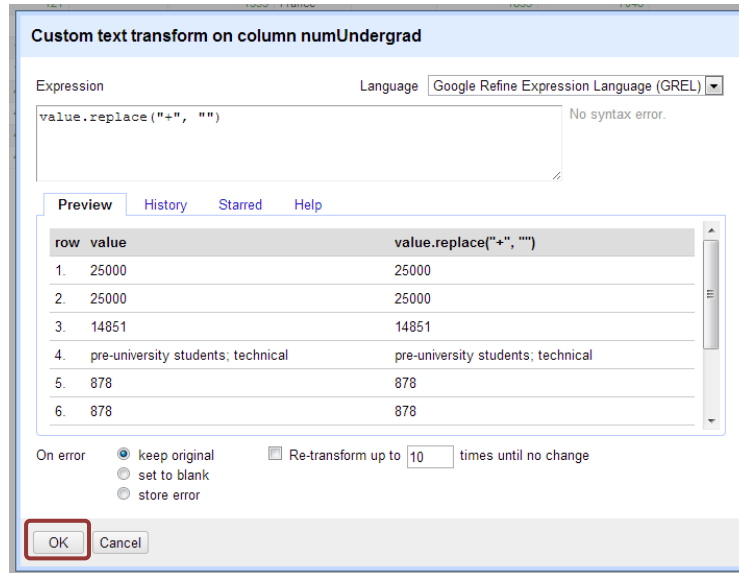


Рис. 15 — Очистка данных в поле **numUndergrad**

Эту же последовательность действий можно выполнить для того, чтобы очистить поле **numUndergrad** от знака ~.

Необходимо написать следующее выражение в текстовое поле **Expression:**

value.replace("~", "").replace(",","")

Необходимо также сконвертировать все нечисловые значения в числовые это можно сделать при помощи последовательного выбора следующих опций меню: **Edit cells -> Common transforms -> To number** (рис. 16).

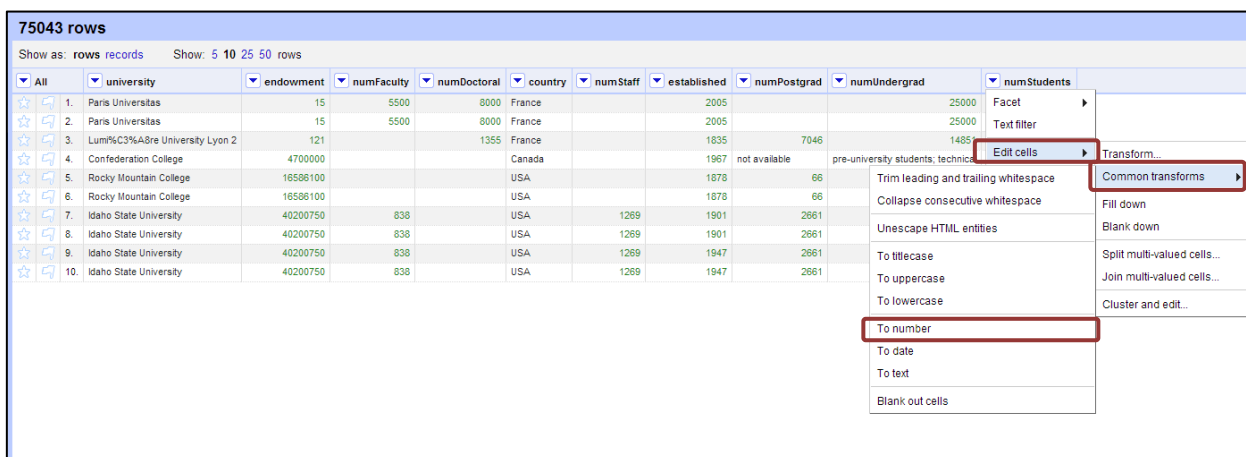


Рис. 16 — Очистка данных в поле **numStudents**

Таким образом, можно привести все значения поля **numStudent** к цифровому виду.

5. Дополнительная функциональность Open Refine

При необходимости можно изменить количество показываемых строк в таблице при помощи строки меню **Show:** (рис. 17).

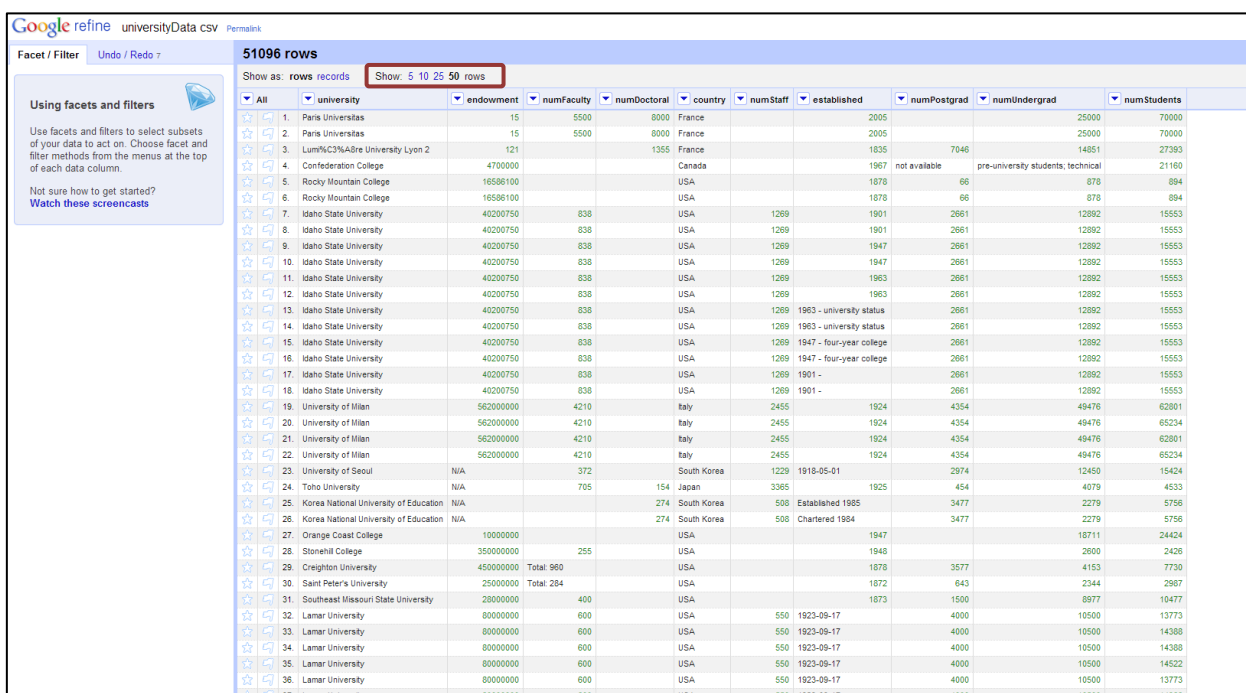



Рис. 17 — Строка меню **Show:**

В Open Refine также предусмотрена возможность сортировки записей при помощи выбора опции **Sort...** в выпадающем меню, которое вызывается при помощи щелчка мыши на кнопке  (рис. 18).

51096 records

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

▼ All	▼ university	▼ endowment	▼ numFaculty	▼ numDoctoral	▼ country	▼ numStaff	▼ est
★	2412. Defiance College	Facet	14700000	86	USA		
★	2420. Defiance College	Text filter		86	USA		
★	2437. University of Saskatchewan College of Agriculture and Bioresources	Edit cells		available	Canada	350	
★	4663. Epoka University	Edit column	3000	50	Albania	25	
★	4665. Epoka University	Transpose	3000	50	Albania	25	
★	26984. Defiance College	Sort...	14700000	86	USA		
★	27573. Defiance College	View		86	USA		
★	49737. Epoka University	Reconcile	nations by businesses affiliated with the Gf'jen movement.	50	NA	Albania	25
★	49739. Epoka University		nations by businesses affiliated with the Gf'jen movement.	50	NA	Albania	25
★	3093. WikiProject Universities			300	100	Utopia	300
★	32508. WikiProject Universities	US\$123,456,789		300	100	Utopia	300
★	26936. University of Wisconsin%E2%80%93La Crosse	US \$41,617,510		531		USA	
★	26937. University of Wisconsin%E2%80%93La Crosse	US \$41,617,510		551		USA	
★	1565. Purdue University Calumet	7900000	224			USA	
★	993. Corban University	US\$3 million				USA	
★	992. Morehead State University	\$22.5M				USA	
★	1081. Hanover College	121600000	100			USA	
★	1083. Hanover College	121600000	100			USA	

Рис. 17 — Сортировка записей