

# Lab1 实验报告

## 1. 实验目的

- openrefine工具的数据过滤,聚类,文本过滤,数据分组,数据转换,数据输出以及正则表达式的应用等.

## 2. 实验环境

- win10 64位
- java8
- openrefine

## 3. 实验步骤

- 首先使用命令行检查本机是否安装了java.可以看到本机已经安装了java的10.0.1的版本.

```
Microsoft Windows [版本 10.0.17134.285]  
(c) 2018 Microsoft Corporation。保留所有权利。  
  
C:\Users\HASEE>java -version  
java version "10.0.1" 2018-04-17  
Java(TM) SE Runtime Environment 18.3 (build 10.0.1+10)  
Java HotSpot(TM) 64-Bit Server VM 18.3 (build 10.0.1+10, mixed mode)  
  
C:\Users\HASEE>_
```

- 前往openrefine官网下载最新版的openrefine3.0版本,解压后运行,发现依然弹出安装java的窗口.在检查下载没有问题之后,我安装了弹出窗口的java8.0版本.

```
Microsoft Windows [版本 10.0.17134.285]
(c) 2018 Microsoft Corporation。保留所有权利。

C:\Users\HASEE>java -version
java version "1.8.0_181"
Java(TM) SE Runtime Environment (build 1.8.0_181-b13)
Java HotSpot(TM) Client VM (build 25.181-b13, mixed mode)

C:\Users\HASEE>_
```

- 在重新安装完成后,根据群内的同学的反映,我替换掉了openrefine中的一个文件从而能够在win10的64位中文版系统中顺利打开openrefine界面.打开后,根据实验指南中的网址下载了数据集,并将数据集导入了openrefine中.

**OpenRefine** A power tool for working with messy data.

Create Project | Open Project | Import Project | Language Settings

Project name: universityData.csv | Tags: | Create Project »

18.	Idaho State University	40200750	838		USA	1269	1901 -	2661	12892	15553
19.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476	62801
20.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476	65234
21.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476	62801
22.	University of Milan	562000000	4210		Italy	2455	1924	4354	49476	65234
23.	Montserrat College of Art	N/A			United States		1970		-400	
24.	University of Seoul	N/A	372		South Korea	1229	1918-05-01	2974	12450	15424
25.	Toho University	N/A	705	154	Japan	3365	1925	454	4079	4533
26.	Korea National University of Education	N/A		274	South Korea	508	Established 1985	3477	2279	5756
27.	Korea National University of Education	N/A		274	South Korea	508	Chartered 1984	3477	2279	5756
28.	California State University%2C Monterey Bay	1.3E7	334		United States		1994	387	4795	
29.	Orange Coast College	1.0E7			United States		1947		18711	24424
30.	Stonehill College	3.5E8	255		USA		1948		2600	2426
31.	Creighton University	4.5E8	Total: 960		United States		1878	3577	4153	7730
32.	Northwest film school	0.0	4		USA		2004-09-30		23	
33.	Northwest Film School	0.0	4		United States		2004-09-30		23	
34.	New College of Florida	2.5E7	87		United States		1960		825	
35.	Saint Peter's University	2.5E7	Total: 284		United States		1872	643	2344	2987
36.	Southeast Missouri State University	2.8E7	400		United States		1873	1500	8977	10477
37.	New River Community College	7200.0	206		United States		1959		4345	
38.	Westminster Choir College	2.0E7	75		United States		1926	91	440	
39.	Lindsey Wilson College	5.3E7	113		United States of America		1903		1902	
40.	Lindsey Wilson College	5.3E7	123		United States of America		1903		1902	
41.	Lindsey Wilson College	5.3E7	113		United States of America		1903		2600	
42.	Lindsey Wilson College	5.3E7	123		United States of America		1903		2600	
43.	Seton Hill University	8.0E7			United States of America		1883		2014	
44.	Lamar University	8.0E7	600		United States	550	1923-09-17	4000	10500	13773
45.	Lamar University	8.0E7	600		United States	550	1923-09-17	4000	10500	14388
46.	Lamar University	8.0E7	600		United States	550	1923-09-17	4000	10500	14388
47.	Lamar University	8.0E7	600		United States	550	1923-09-17	4000	10500	14522
48.	Lamar University	8.0E7	600		United States	550	1923-09-17	4000	10500	14522

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 | RDF/JSON files | Wikitext | XML files | Plain Document Format

Character encoding: | Columns are separated by: ☐ commas (CSV) ☒ tabs (TSV) ☐ custom: \t | Escape special characters with \

☐ 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

## 使用openrefine打开数据集

- 首先使用text facet功能来对特定文本的值进行分组归类统计.在country列中执行text facet操作.


**OpenRefine** universityData.csv [Permalink](#)

Open...
Export
Help

Facet / Filter
Undo / Redo 0 / 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](#)

75043 rows

Extensions: Wikidata

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

numDoctoral	country	numStaff	established	numPostgrad	numUndergr
					25000
					25000
5				7046	14851
				not available	pre-university
				66	878
				66	878
				2661	12892
				2661	12892
		1269	1947	2661	12892
		1269	1947	2661	12892
	United States	1269	1963	2661	12892
	USA	1269	1963	2661	12892
	United States	1269	1963 - university status	2661	12892
	USA	1269	1963 - university status	2661	12892
	United States	1269	1947 - four-year college	2661	12892
	USA	1269	1947 - four-year college	2661	12892
	United States	1269	1901 -	2661	12892
	USA	1269	1901 -	2661	12892
	Italy	2455	1924	4354	49476
	Italy	2455	1924	4354	49476
	Italy	2455	1924	4354	49476
	Italy	2455	1924	4354	49476
	United States		1970		~400
	South Korea	1229	1918-05-01	2974	12450
	Japan	3365	1925	454	4079
	South Korea	508	Established 1985	3477	2279
	South Korea	508	Chartered 1984	3477	2279
	United States		1994	387	4795
	United States		1947		18711
	USA		1948		2600
	United States		1878	3577	4153
	USA		2004-09-30		23
	United States		2004-09-30		23
	United States		1960		825
	United States		1872	643	2344
	United States		1873	1500	8977
	United States		1959		4345
	United States		1926	91	440
	United States of America		1903		1902

text facet

- 对country列进行text facet操作后,openrefine自动将country列中的国别进行了聚类操作,并将结果显示在了左边栏中.



text facet结果

- 在聚类完成后,发现类别中对于美国国名的描述出现了不同种类的偏差.为了将描述进行统一,按照实验指南,我对这些类别进行了merge操作.

## 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 "New York" and "New York City" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person.

Method

Keying Function

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

Select All Unselect All

Export Clusters

Merge Selected & Re-Cluster

修改类别标签

- 接着对样本进行文本过滤.假设我们不需要来自美国大学的信息.我使用text filter的功能,输入united states,并使用edit rows中的remove all matching rows来删除匹配到的所有行.

OpenRefine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 1 / 1

Refresh Reset All Remove All

**country** change

2 choices Sort by: name count Cluster

United States 45431  
United States of America 615  
Facet by choice counts

**country** invert reset

United States

☐ case sensitive ☐ regular expression

**46046 matching rows** (75043 total) Extensions: Wikidata

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

	All	university	endowment	numFaculty	numDoctoral	country	num
Transform	Mountain	16586100				United States	
Facet	Mountain	16586100				United States	
Edit rows	Star rows		38			United States	1269
Edit columns	Unstar rows		38			United States	1269
View	Flag rows		38			United States	1269
	Unflag rows		38			United States	1269
	Remove all matching rows		38			United States	1269
9.	Idaho S						
10.	Idaho S						
11.	Idaho State University	40200750	838			United States	1269
12.	Idaho State University	40200750	838			United States	1269
13.	Idaho State University	40200750	838			United States	1269
14.	Idaho State University	40200750	838			United States	1269
15.	Idaho State University	40200750	838			United States	1269
16.	Idaho State University	40200750	838			United States	1269
17.	Idaho State University	40200750	838			United States	1269
18.	Idaho State University	40200750	838			United States	1269
23.	Montserrat College of Art	N/A				United States	
28.	California State University%2C Monterey Bay	1.3E7	334			United States	
29.	Orange Coast College	1.0E7				United States	
30.	Stonehill College	3.5E8	255			United States	
31.	Creighton University	4.5E8	Total: 960			United States	
32.	Northwest film school	0.0	4			United States	
33.	Northwest Film School	0.0	4			United States	
34.	New College of Florida	2.5E7	87			United States	
35.	Saint Peter's University	2.5E7	Total: 284			United States	
36.	Southeast Missouri State University	2.8E7	400			United States	

javascript:[]

删除某类样本

- 接着使用numeric facet对数值型数据进行分组.由于numStudents中的数据是使用文本字符串存储的数值型数据,因此我们使用common transform中的to number功能将此类数据转换为数值型数据.

OpenRefine universityData.csv Permalink

Open... Export Help

Facet / Filter Undo / Redo 2 / 2

Refresh Reset All Remove All

country change

61 choices Sort by: name count Cluster

2

Albania 8

Argentina 1

Australia 335

Bangladesh 27

Botswana 2

Brazil 21

Bulgaria 2

Canada 12907

Canada B1P 6L2 576

Canada C1A 4P3 Telephone: 902-566-0439 Fax: 902-566-0795 1

country invert reset

case sensitive regular expression

28997 rows Extensions: Wikidata

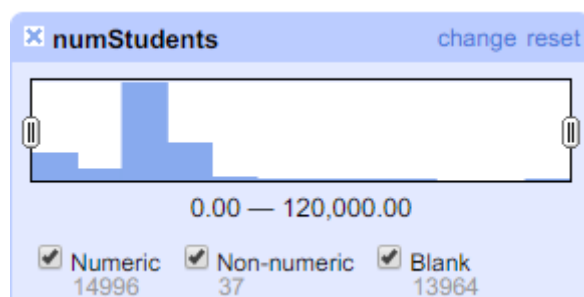
Show as: rows records Show: 5 10 25 50 rows « first < previous 1 - 50 next > last »

country	numStaff	established	numPostgrad	numUndergrad	numStudents
France		2005			
France		2005			
France		1835			
South Korea	508	Chartered 1984	3477	2279	5756
England, UK		1881	8138	24334	32472
England, UK		1881	8138	24334	40829
England, UK		1881	10031	24334	32472
England, UK		1881	10031	24334	40829
England, UK		1881	10060	24334	32472
England, UK		1881	10060	24334	40829
England, UK		1881	8138	25115	32472
England, UK		1881	8138	25115	40829
England, UK		1881	10031	25115	32472
England, UK		1881	10031	25115	40829
England, UK		1881	10060	25115	32472
England, UK		1881	10060	25115	40829

javascript:()

将文本转换成数字

- 此时再使用numeric facet功能对数据集中的数据进行分析。



对转换后的数据进行统计

- 可以看到数据集中出现了一些转换后依然是非数值型的数据.取消勾选Numeric以及Blank选项.并使用edit cell中的transform功能将修改这些数据使其能够符合格式要求.

The screenshot shows the OpenRefine interface with a dataset named 'universityData.csv'. A facet is applied to the 'country' column. The 'numStudents' column is selected for a custom text transform. The dialog box shows the expression `value.replace("+", "")` and a preview table with the following data:

row	value	value.replace("+", "")
70.	900+	900
93.	~50,000	~50,000
94.	~50,000	~50,000
97.	~50,000	~50,000
98.	~50,000	~50,000
101.	~50,000	~50,000

The dialog also includes options for 'On error' (keep original, set to blank, store error) and a checkbox for 'Re-transform up to 10 times until no change'.

### 修改不合规规范的数据

- 修改后重新进行类数据格式的转换,将依然不符合格式要求的非数值型数据删除.
- openrefine1还有很方便的撤销操作,在左边栏切换到undo/redo标签即可通过点击操作列表回退操作

0. Create project
1. Mass edit 45431 cells in column country
2. Remove 46046 rows
3. Text transform on 14996 cells in column numStudents: value.toNumber()
4. Edit single cell on row 70, column numStudents
5. Edit single cell on row 70, column numStudents
6. Text transform on 1 cells in column numStudents: grel:value.replace("+", "")
7. Edit single cell on row 93, column numStudents
8. Text transform on 8 cells in column numStudents: grel:value.replace("~", ""); value.replace",", ""
9. Text transform on 1 cells in column numStudents: value.toNumber()
10. Remove 36 rows

方便的撤销操作列表

## 4. 实验总结

在本次实验中我接触了一个可以对数据集进行一定程度预处理的软件openrefine,在以前我对数据进行预处理的时候都是使用python自带的pandas包进行的,由于我本人的操作不熟练导致效率相当的低下.而相对的,在openrefine中进行简答的数据预处理则非常方便.不过,这同时也要求我对这款软件进行更加深刻的了解.

## 5. 操作习题



## 举例说明正则表达式?

```
^\d+(\.\d+)?
```

# 用来匹配个位数的实数

# `^\d`定义了以单个数字开始

# `?` 设置括号内的选项是可选的

# `\.` 匹配 `"."`

# `\d` 表示在`"."`后面接数字

## 简述正则表达式的作用

正则表达式是用来描述或者匹配一系列符合某个句法规则的字符串的单个字符串.它通常被用来检索和/或替换那些符合某个模式的文本内容.

## 通过操作,举例说明openrefine中正则表达式的使用

使用正则表达式选取大学创始在2000年到2006年之间的学校

1993 matching rows (75043 total)											
Extensions Wikidata											
Show as: rows records Show: 5 10 25 50 rows											
Facet / Filter Undo / Redo 0 / 0											
Refresh Reset All Remove All											
Invert reset											
case sensitive regular expression											
established											
200[0 1 2 3 4 5 6]											
All university endowment numFaculty numDoctoral country numStaff established numPostgrad numUndergrad numStudents											
1.	Paris Universita	15	5500	8000	France	2005		25000	70000		
2.	Paris Universita	15	5500	8000	France	2005		25000	70000		
32.	Northwest film school	0.0	4		USA	2004-09-30		23			
33.	Northwest Film School	0.0	4		United States	2004-09-30		23			
145.	University of Arkansas Honors College	2.0E8			United States	2002					
2161.	University of Bolton	160000.0			England, UK	2004	287	1317	2200		
2185.	University of Lincoln	424000.0			England	2001	1520	7731	9251		
2196.	Bath Spa University	155000.0			England	2005	1355	10367	11722		
2518.	Liverpool Hope University	179783.0			England, UK	2005	2605	4505	7110		
2519.	Liverpool Hope University	179783.0			England, UK	1844. Gained full university status in 2005 University Status	1675	6190	7885		
2747.	University of Bedfordshire	1400000.0			United Kingdom	2006	1675	6190	7885		
2748.	University of Bedfordshire	1400000.0			United Kingdom	2006	6801	13480	23859		
2749.	University of Bedfordshire	1400000.0			United Kingdom	2006	6801	13480	23930		
2750.	University of Bedfordshire	1400000.0			United Kingdom	2006	10379	13480	23859		
2751.	University of Bedfordshire	1400000.0			United Kingdom	2006	10379	13480	23930		
2752.	University of Bedfordshire	1400000.0			United Kingdom	2006	6801	17129	23859		
2753.	University of Bedfordshire	1400000.0			United Kingdom	2006	6801	17129	23930		
2754.	University of Bedfordshire	1400000.0			United Kingdom	2006	10379	17129	23859		
2755.	University of Bedfordshire	1400000.0			United Kingdom	2006	10379	17129	23930		
2756.	University of Bedfordshire	1400000.0			UK	2006	6801	13480	23859		
2757.	University of Bedfordshire	1400000.0			UK	2006	6801	13480	23930		
2758.	University of Bedfordshire	1400000.0			UK	2006	10379	13480	23859		
2759.	University of Bedfordshire	1400000.0			UK	2006	10379	13480	23930		
2760.	University of Bedfordshire	1400000.0			UK	2006	6801	17129	23859		
2761.	University of Bedfordshire	1400000.0			UK	2006	6801	17129	23930		
2762.	University of Bedfordshire	1400000.0			UK	2006	10379	17129	23859		
2795.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	10379	17129	23930		
2796.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	6801	13480	23859		
2797.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	6801	13480	23930		
2798.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	10379	13480	23859		
2799.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	10379	13480	23930		
2800.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	6801	17129	23859		
2801.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	6801	17129	23930		
2802.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	10379	17129	23859		
2803.	University of Bedfordshire	1400000.0			United Kingdom	2006 - University of Bedfordshire	10379	17129	23930		
2804.	University of Bedfordshire	1400000.0			UK	2006 - University of Bedfordshire	6801	13480	23859		
2805.	University of Bedfordshire	1400000.0			UK	2006 - University of Bedfordshire	6801	13480	23930		
2806.	University of Bedfordshire	1400000.0			UK	2006 - University of Bedfordshire	10379	13480	23859		
2807.	University of Bedfordshire	1400000.0			UK	2006 - University of Bedfordshire	10379	13480	23930		
2808.	University of Bedfordshire	1400000.0			UK	2006 - University of Bedfordshire	6801	17129	23859		