IS 532 Data Cleaning Final Project

Directory of Open Access Journals

BOX LINK: https://uofi.box.com/s/rrrkdxzqfz0b3iyy6n835zbbowdcveat

Mansi Tripathi (mansit2)

Priyanshu Madan (pmadan3)

Simran Wig (swig2)

1. Introduction and Overview

The Directory of Open Access Journals (DOAJ) is a community-curated database that indexes high quality, peer reviewed Open Access research journals, periodicals and their articles' metadata. The DOAJ is maintained by the Infrastructure Services for Open Access.

The Directory aims to be comprehensive and cover all open access academic journals that use an appropriate quality control system and is not limited to particular languages, geographical regions, or subject areas. The Directory aims to increase the visibility and ease of use of open access academic journals—regardless of size and country of origin—thereby promoting their visibility, usage and impact.

Motivation and Purpose

Our primary motivation was to clean this data so that it can be used for further analysis - which could be helpful for industry personnels to find out about the research going on in their domain through these journal articles. Since it is an open source database it is not maintained by extensive support and hence we had all the more reason to work with this dataset and clean it.

We also wanted to explore the trends of these journals over years and evaluate how the journal publication has changed over the years.
Some of the trend which we wanted to explore are:

- Find the most widely used language in a journal
- Categorize the journals by the licenses used
- Find out the most common review processes for these journals
- Analyze the average time for journal publication with respect to the subject of the journal

2. Initial Assessment of the Dataset and Use Cases

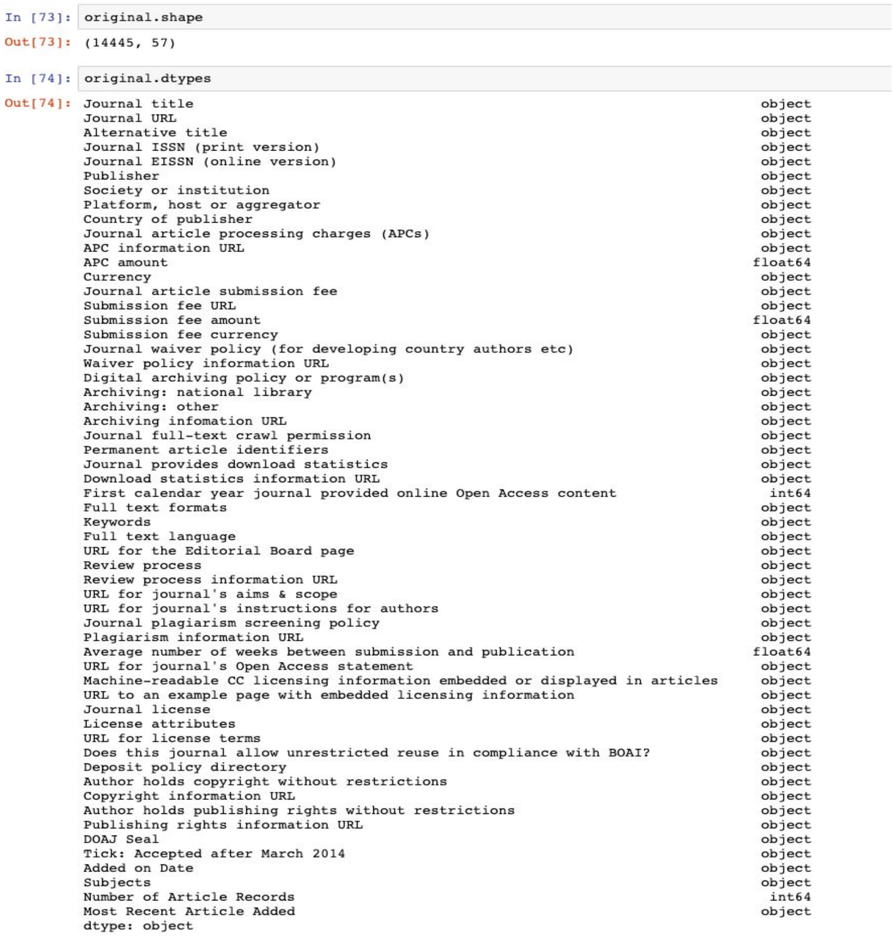
Dataset Description

The dataset is obtained from the DOAJ website. It is journal metadata for over 14000 Journals which includes information about the journal submission costs, the costs associated with article processing, information about the review process, the subjects of these journals and articles. Some of the columns in the dataset are:

- Journal Title
- Journal URL
- Publisher
- Review process
- Country of publisher
- Full text language

Shape of the Dataset

The dataset consists of over 14445 rows and 57 columns.



```
In [73]: original.shape
Out[73]: (14445, 57)

In [74]: original.dtypes
Out[74]: Journal title                                                                    object
         Journal URL                                                                      object
         Alternative title                                                                object
         Journal ISSN (print version)                                                     object
         Journal EISSN (online version)                                                   object
         Publisher                                                                        object
         Society or institution                                                           object
         Platform, host or aggregator                                                     object
         Country of publisher                                                             object
         Journal article processing charges (APCs)                                        object
         APC information URL                                                              object
         APC amount                                                                      float64
         Currency                                                                         object
         Journal article submission fee                                                   object
         Submission fee URL                                                               object
         Submission fee amount                                                           float64
         Submission fee currency                                                          object
         Journal waiver policy (for developing country authors etc)                       object
         Waiver policy information URL                                                    object
         Digital archiving policy or program(s)                                           object
         Archiving: national library                                                      object
         Archiving: other                                                                 object
         Archiving infomation URL                                                         object
         Journal full-text crawl permission                                               object
         Permanent article identifiers                                                    object
         Journal provides download statistics                                             object
         Download statistics information URL                                              object
         First calendar year journal provided online Open Access content                   int64
         Full text formats                                                                object
         Keywords                                                                         object
         Full text language                                                               object
         URL for the Editorial Board page                                                 object
         Review process                                                                   object
         Review process information URL                                                   object
         URL for journal's aims & scope                                                   object
         URL for journal's instructions for authors                                       object
         Journal plagiarism screening policy                                              object
         Plagiarism information URL                                                       object
         Average number of weeks between submission and publication                      float64
         URL for journal's Open Access statement                                          object
         Machine-readable CC licensing information embedded or displayed in articles      object
         URL to an example page with embedded licensing information                       object
         Journal license                                                                  object
         License attributes                                                               object
         URL for license terms                                                            object
         Does this journal allow unrestricted reuse in compliance with BOAI?              object
         Deposit policy directory                                                         object
         Author holds copyright without restrictions                                      object
         Copyright information URL                                                        object
         Author holds publishing rights without restrictions                              object
         Publishing rights information URL                                                object
         DOAJ Seal                                                                        object
         Tick: Accepted after March 2014                                                  object
         Added on Date                                                                    object
         Subjects                                                                         object
         Number of Article Records                                                         int64
         Most Recent Article Added                                                        object
         dtype: object
```

Data Quality Issues

Following are some of the data quality issues which we encountered while performing an exploratory analysis of our dataset:

- **Whitespaces**: Most of the textual columns have leading, trailing, and consecutive whitespaces
- **Number format**: Some of the columns with numbers are in string format which need to be converted to integer or float
- **Date/Time format**: The dates in this dataset are of Argo float format (example: 2004-04-23T21:31:00Z) which is difficult to understand. Since time is not an essential metric for our use case, we will trim the time and transform the date into a more recognizable format (YYYY/MM/DD)
- **Unwanted characters**: Most of the columns (Journal title,Publisher) have unwanted characters and parentheses which needs to be removed
- **Unnecessary columns**: Columns that have a lot of missing values (>80%) and add no value to our use case (URL's) need to be removed

Use Cases

Already good enough use cases

1. Which are the top 10 countries that publish journals?
2. Which is the most common review process in Journal publishing?
3. What are the common licenses used in Journal publication?

Middle-of-the-road use cases

1. **Researchers who want to find journals based on keywords and/or subjects** - to learn about the research going on in their domain. For instance, if a researcher wants to find information about the research going on for the vaccine of coronavirus they can do so by filtering the dataset using keywords and subject. To address this use case we cleaned the columns Keywords and Subjects. We explored the two columns and trimmed the whitespaces in both these columns as a first step to clean them. For the Keywords column, we used text facet and clustering to group similar rows which were different due to typos or punctuation differences. For the Subjects column, we split it in two columns to categorize it as the main subject (Medicine) and the sub subject (Pediatrics) for efficient filtering. The structure of the original column was main subject: sub subject.

2. **Researchers who want to find the average time for journal publication with respect to subject -** To address this use case we used the column - average number of weeks between submission and publication and Number of article records. We converted these columns which were in string format to number.

3. Data Cleaning Methods and Processes

To address the data quality issues mentioned above, we have used OpenRefine and Python for Data Cleaning.

A. OpenRefine

OpenRefine is an open source tool used to clean up the dirty data and transform it in a format required by the user which can also work with various spreadsheet formats. In this project, we are using a .csv file and cleaning the messy data from the DOAJ dataset. Through OpenRefine we plan to address data quality issues like trimming and collapsing whitespaces, converting data types, text facet and clustering same rows with different semantics (for instance - lowercase, uppercase difference. The reason we used OpenRefine for these tasks is because of the ease of use of OpenRefine as opposed to Python (the other data cleaning method we are using). It is relatively easier to handle regular expressions related issues in OpenRefine compared to Python.

Following are the steps we took to clean our data:

1. Trim and collapse whitespaces

There were a lot of unnecessary white spaces in a few of the columns which make analysis difficult. We looked for white spaces in textual columns (string) - whitespaces that are present in the beginning and end of the string as well as consecutive whitespaces. While collapsing and trimming these whitespaces, we also found that a couple of textual columns recorded 0 cell changes - which meant they were already clean in terms of whitespaces. Following are the columns for which trimmed and collapsed whitespaces:

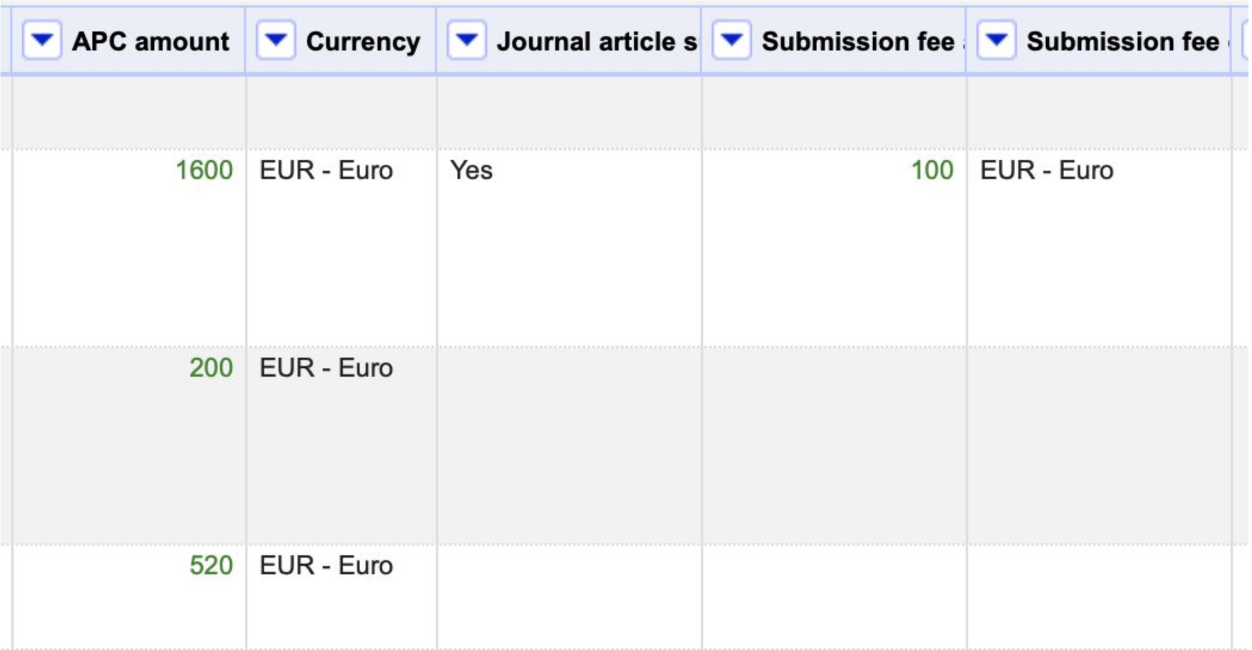
a. Journal title
b. Alternative title
c. Publisher
d. Society or Institution
e. Review process
f. Platform or host or aggregator
g. Country of Publisher
h. Keywords
i. Subjects

2. Number

The next thing we looked at was the numeric columns which were actually not numeric values, so we converted them to numeric values through the to number function. Following are the columns that we changed to number:

a. APC amount
b. Submission fee amount
c. The average number of weeks between submission and publication
d. Number of article records

OpenRefine displays these numeric values which are converted to number in green. We can see a couple of numeric columns in the screenshot below.

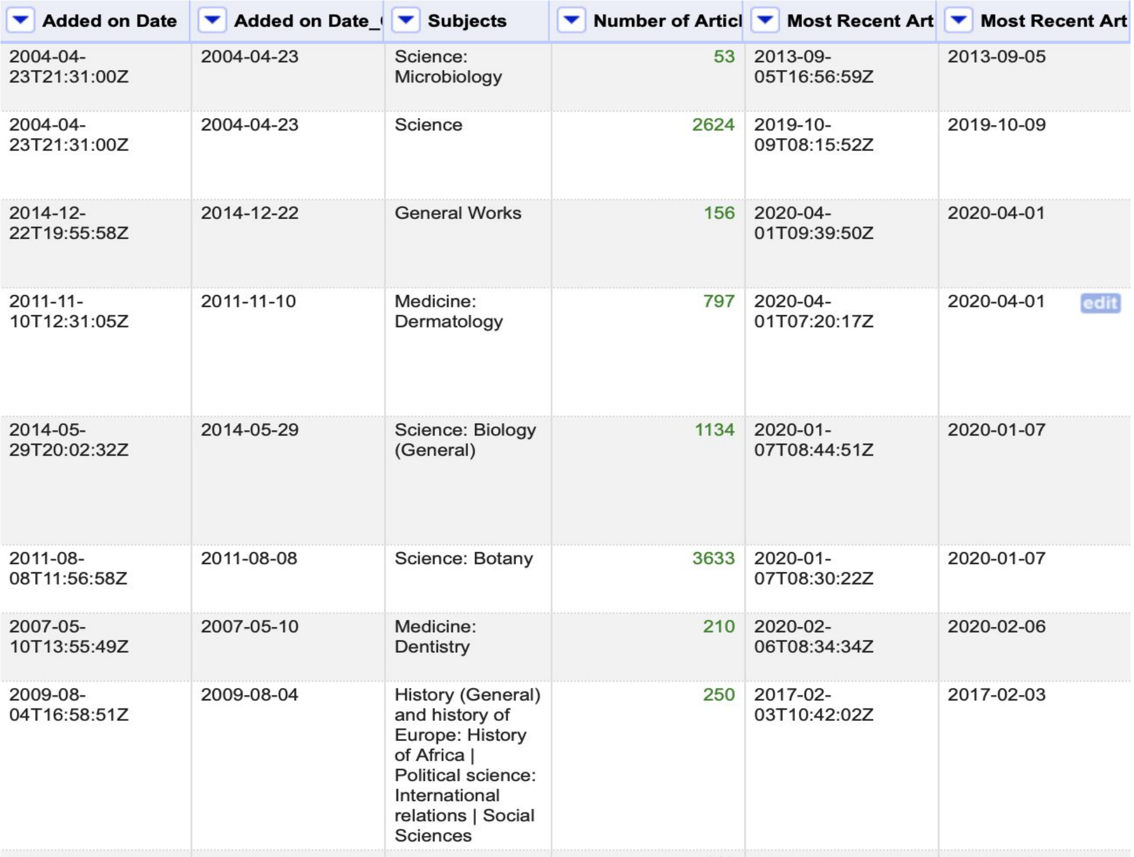


APC amount	Currency	Journal article s	Submission fee	Submission fee
1600	EUR - Euro	Yes	100	EUR - Euro
200	EUR - Euro			
520	EUR - Euro			

3. Date

In the given dataset, the dates are of Argo float format (e.g. 2004-04-23T21:31:00Z) including date as well as time. Since time is not an essential metric for our use case, we decided to trim the time and transform the date into a more recognizable format (YYYY/MM/DD). We used grel and used the function: toString(toDate(value),"yyyy-MM-dd"). We first used the inbuilt To Date function, but we did not witness any change in the column. Post that we used the grel function to convert it to date without time. Following are the columns on which we used this function:

a. The column “Added on Date” was converted to the required date format and added the same to a new column called “Added on Date_Clean”
b. The same was done for “Most Recent Article Added” and the new column was called “Most Recent Article Added_Clean”



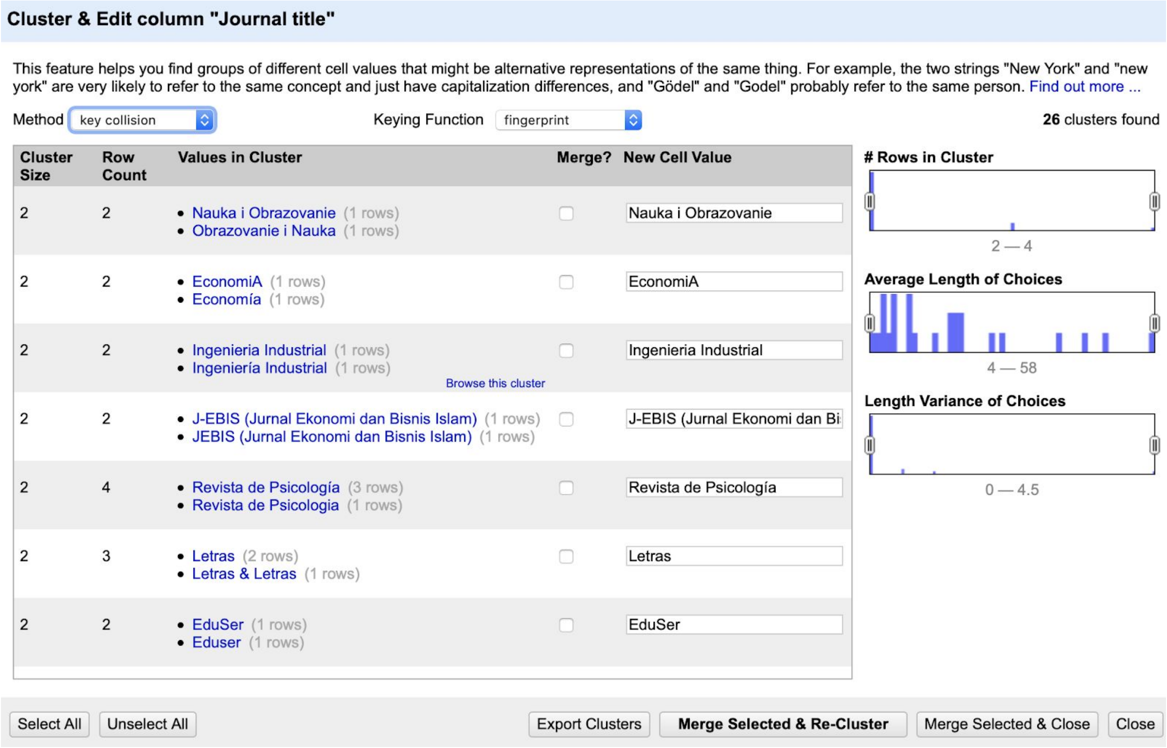
Added on Date	Added on Date_(	Subjects	Number of Articl	Most Recent Art	Most Recent Art
2004-04-23T21:31:00Z	2004-04-23	Science: Microbiology	53	2013-09-05T16:56:59Z	2013-09-05
2004-04-23T21:31:00Z	2004-04-23	Science	2624	2019-10-09T08:15:52Z	2019-10-09
2014-12-22T19:55:58Z	2014-12-22	General Works	156	2020-04-01T09:39:50Z	2020-04-01
2011-11-10T12:31:05Z	2011-11-10	Medicine: Dermatology	797	2020-04-01T07:20:17Z	2020-04-01
2014-05-29T20:02:32Z	2014-05-29	Science: Biology (General)	1134	2020-01-07T08:44:51Z	2020-01-07
2011-08-08T11:56:58Z	2011-08-08	Science: Botany	3633	2020-01-07T08:30:22Z	2020-01-07
2007-05-10T13:55:49Z	2007-05-10	Medicine: Dentistry	210	2020-02-06T08:34:34Z	2020-02-06
2009-08-04T16:58:51Z	2009-08-04	History (General) and history of Europe: History of Africa | Political science: International relations | Social Sciences	250	2017-02-03T10:42:02Z	2017-02-03

4. Text facet & Clustering

We created text facets to analyze specific textual columns and group similar items together. Clustering helps us group items that are different because of misspellings, typos, punctuation marks, difference in case of the words. We used the key collision method and fingerprint keying function for clustering the following columns.

a. Journal Title -> Journal Title_Clean:
   Method - key collision; Keying function - fingerprint: Using this method and keying function we detected 26 clusters for the Journal title column.

Mass edit 57 cells in column Journal title_Clean



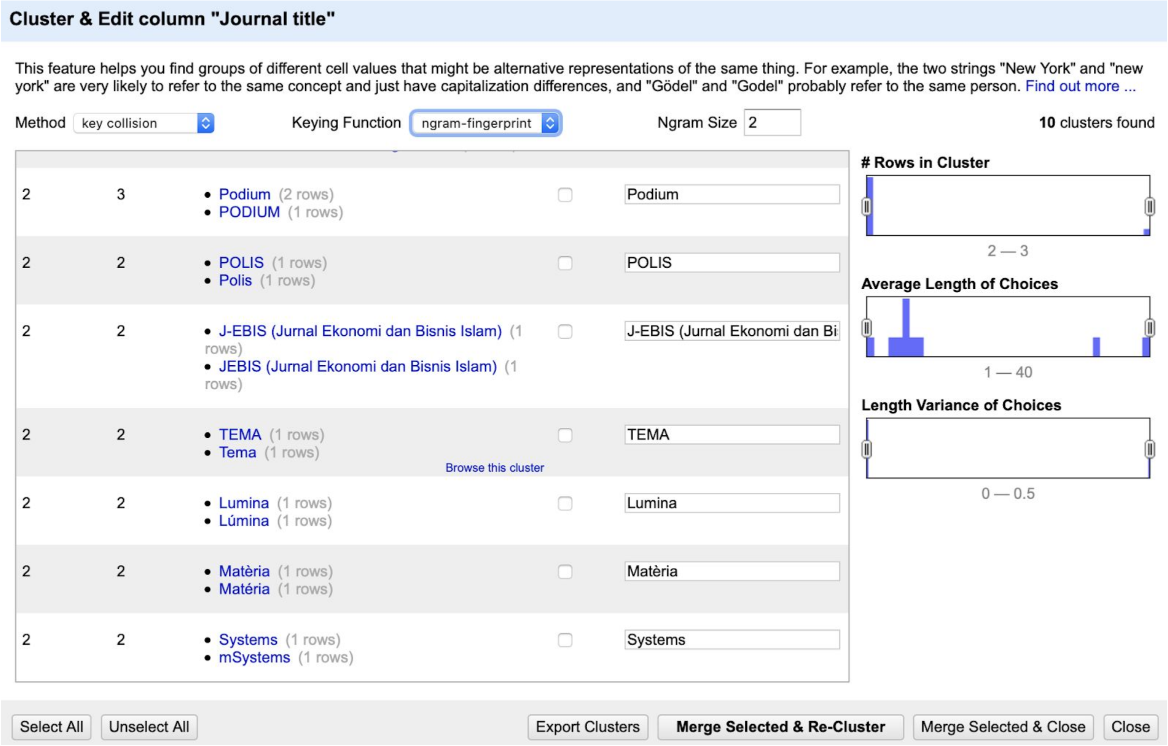
Cluster & Edit column "Journal title"

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 "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision Keying Function: fingerprint **26** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
2	2	- Nauka i Obrazovanie (1 rows) - Obrazovanie i Nauka (1 rows)	□	Nauka i Obrazovanie
2	2	- EconomiA (1 rows) - Economía (1 rows)	□	EconomiA
2	2	- Ingenieria Industrial (1 rows) - Ingeniería Industrial (1 rows) Browse this cluster	□	Ingenieria Industrial
2	2	- J-EBIS (Jurnal Ekonomi dan Bisnis Islam) (1 rows) - JEBIS (Jurnal Ekonomi dan Bisnis Islam) (1 rows)	□	J-EBIS (Jurnal Ekonomi dan Bi
2	4	- Revista de Psicología (3 rows) - Revista de Psicologia (1 rows)	□	Revista de Psicología
2	3	- Letras (2 rows) - Letras & Letras (1 rows)	□	Letras
2	2	- EduSer (1 rows) - Eduser (1 rows)	□	EduSer

Method - key collision; Keying function - ngram-fingerprint: By using the keying function ngram-fingerprint we detected fewer clusters for the same column (10 clusters). We therefore clustered using the fingerprint keying function instead of ngram-fingerprint.



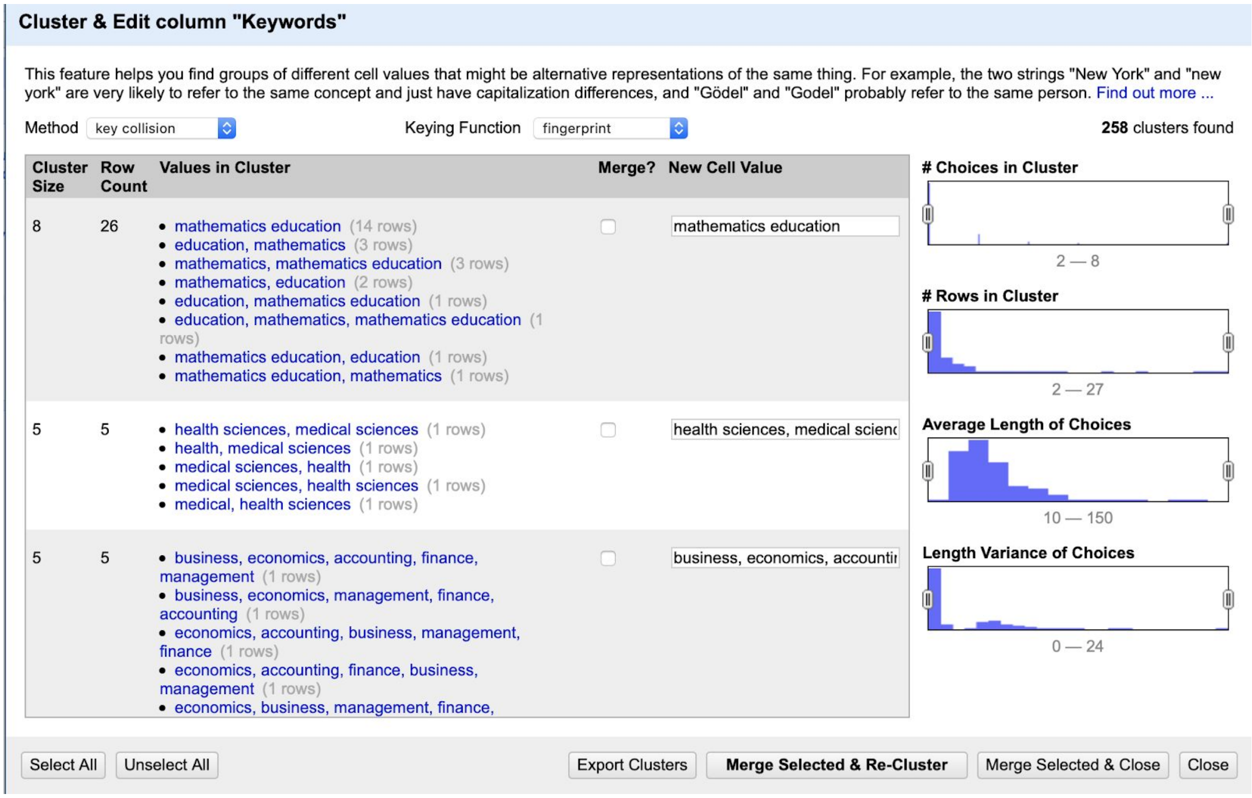
Cluster & Edit column "Journal title"

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 "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision Keying Function: ngram-fingerprint Ngram Size: 2 **10** clusters found

2	3	- Podium (2 rows) - PODIUM (1 rows)	□	Podium
2	2	- POLIS (1 rows) - Polis (1 rows)	□	POLIS
2	2	- J-EBIS (Jurnal Ekonomi dan Bisnis Islam) (1 rows) - JEBIS (Jurnal Ekonomi dan Bisnis Islam) (1 rows)	□	J-EBIS (Jurnal Ekonomi dan Bi
2	2	- TEMA (1 rows) - Tema (1 rows) Browse this cluster	□	TEMA
2	2	- Lumina (1 rows) - Lúmina (1 rows)	□	Lumina
2	2	- Matèria (1 rows) - Matéria (1 rows)	□	Matèria
2	2	- Systems (1 rows) - mSystems (1 rows)	□	Systems

b. Keywords -> Keywords_Clean:
Method - key collision; Keying function - fingerprint

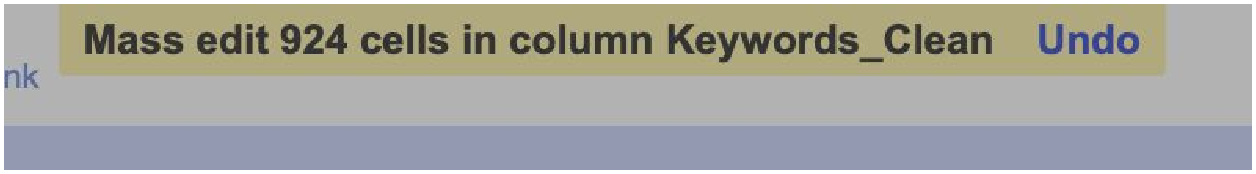


Cluster & Edit column "Keywords"

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 "Gödel" and "Godel" probably refer to the same person. Find out more ...

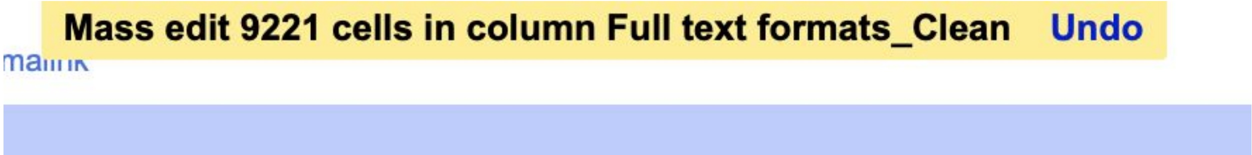
Method: key collision — Keying Function: fingerprint — **258** clusters found

Cluster Size	Row Count	Values in Cluster	Merge?	New Cell Value
8	26	• mathematics education (14 rows) • education, mathematics (3 rows) • mathematics, mathematics education (3 rows) • mathematics, education (2 rows) • education, mathematics education (1 rows) • education, mathematics, mathematics education (1 rows) • mathematics education, education (1 rows) • mathematics education, mathematics (1 rows)	□	mathematics education
5	5	• health sciences, medical sciences (1 rows) • health, medical sciences (1 rows) • medical sciences, health (1 rows) • medical sciences, health sciences (1 rows) • medical, health sciences (1 rows)	□	health sciences, medical scienc
5	5	• business, economics, accounting, finance, management (1 rows) • business, economics, management, finance, accounting (1 rows) • economics, accounting, business, management, finance (1 rows) • economics, accounting, finance, business, management (1 rows) • economics, business, management, finance,	□	business, economics, accountir



Mass edit 924 cells in column Keywords_Clean Undo

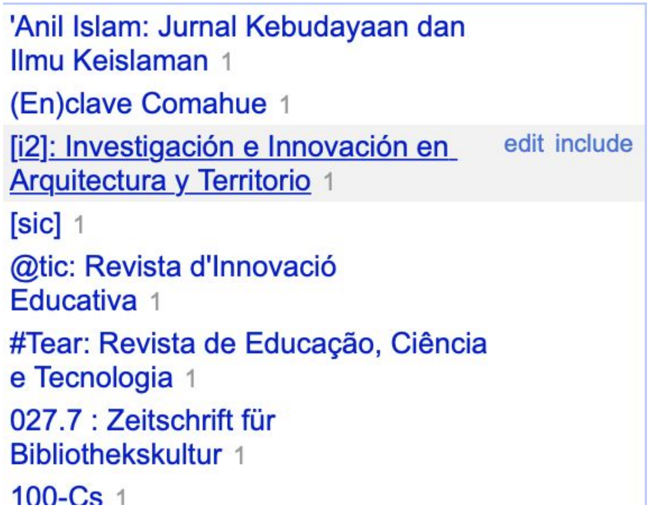
c. Full text formats -> Full text formats_Clean



Mass edit 9221 cells in column Full text formats_Clean Undo

5. Removed parenthesis

Some of the columns had parentheses and those parentheses needed to be removed. Thus they were removed from “Journal title_Clean” and added to a new column “Journal title_parantheses”. Below is a screenshot of the parentheses in some of the rows of the columns.



'Anil Islam: Jurnal Kebudayaan dan Ilmu Keislaman 1

(En)clave Comahue 1

[i2]: Investigación e Innovación en Arquitectura y Territorio 1

[sic] 1

@tic: Revista d'Innovació Educativa 1

#Tear: Revista de Educação, Ciência e Tecnologia 1

027.7 : Zeitschrift für Bibliothekskultur 1

100-Cs 1

6. Remove columns

We removed columns that consisted of URLs since these columns were not required for future analysis or any of the use cases of our project. Following are the columns which we removed:

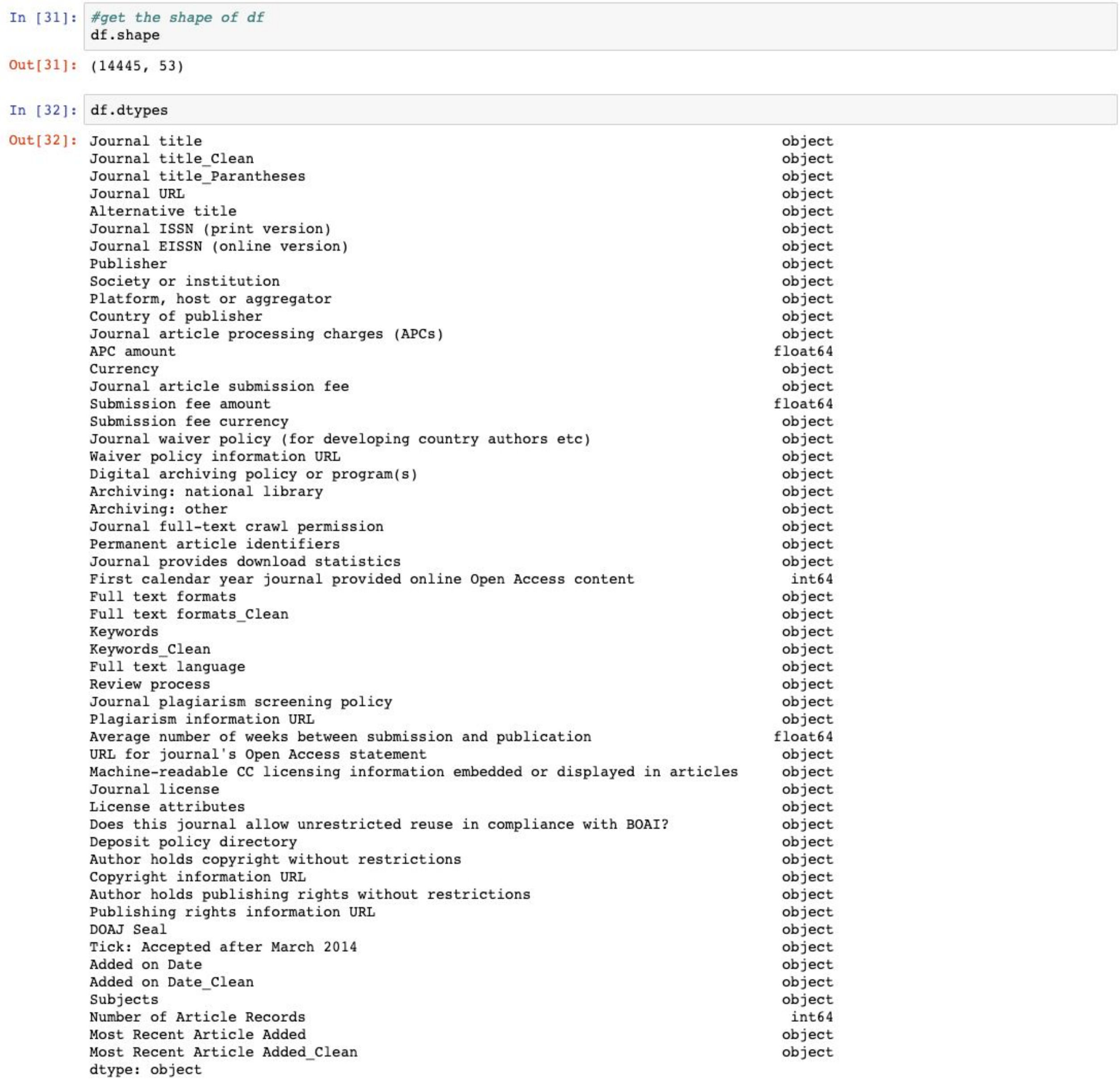
a. Submission fee URL
b. APC Information URL
c. Archiving Information URL
d. Download statistics information URL
e. URL for the Editorial Board page
f. Review process information URL
g. URL for journal's aims & scope
h. URL for journal's instruction for authors
i. URL to an example page with embedded licensing information
j. URL for license terms

B. Python Cleaning

1. The shape of the dataset

The shape of the dataset and datatypes of columns: Now that we have covered the cleaning process in open refine, the first step was to verify and analyze the dataset for further cleaning. We used the "shape" function to find the number of rows and columns and the "dtypes" function to find the data type of each column.

The dataset has 14445 rows and 53 columns with the following data types.



```
In [31]: #get the shape of df
         df.shape

Out[31]: (14445, 53)

In [32]: df.dtypes

Out[32]: Journal title                                                                object
         Journal title_Clean                                                          object
         Journal title_Parantheses                                                    object
         Journal URL                                                                  object
         Alternative title                                                            object
         Journal ISSN (print version)                                                 object
         Journal EISSN (online version)                                               object
         Publisher                                                                    object
         Society or institution                                                       object
         Platform, host or aggregator                                                 object
         Country of publisher                                                         object
         Journal article processing charges (APCs)                                    object
         APC amount                                                                  float64
         Currency                                                                     object
         Journal article submission fee                                               object
         Submission fee amount                                                       float64
         Submission fee currency                                                      object
         Journal waiver policy (for developing country authors etc)                   object
         Waiver policy information URL                                                object
         Digital archiving policy or program(s)                                       object
         Archiving: national library                                                  object
         Archiving: other                                                             object
         Journal full-text crawl permission                                           object
         Permanent article identifiers                                                object
         Journal provides download statistics                                         object
         First calendar year journal provided online Open Access content               int64
         Full text formats                                                            object
         Full text formats_Clean                                                      object
         Keywords                                                                     object
         Keywords_Clean                                                               object
         Full text language                                                           object
         Review process                                                               object
         Journal plagiarism screening policy                                          object
         Plagiarism information URL                                                   object
         Average number of weeks between submission and publication                  float64
         URL for journal's Open Access statement                                      object
         Machine-readable CC licensing information embedded or displayed in articles     object
         Journal license                                                              object
         License attributes                                                           object
         Does this journal allow unrestricted reuse in compliance with BOAI?          object
         Deposit policy directory                                                     object
         Author holds copyright without restrictions                                  object
         Copyright information URL                                                    object
         Author holds publishing rights without restrictions                          object
         Publishing rights information URL                                            object
         DOAJ Seal                                                                    object
         Tick: Accepted after March 2014                                              object
         Added on Date                                                                object
         Added on Date_Clean                                                          object
         Subjects                                                                     object
         Number of Article Records                                                     int64
         Most Recent Article Added                                                    object
         Most Recent Article Added_Clean                                              object
         dtype: object
```

2. **Removing unnecessary columns and the unclean versions of the clean columns**

After analyzing the dataset we decided to remove all the unnecessary columns that won't add any value to the future analysis. These columns were mostly URLs referring to the components of the journals. Also removing unclean columns that were cleaned using openrefine.

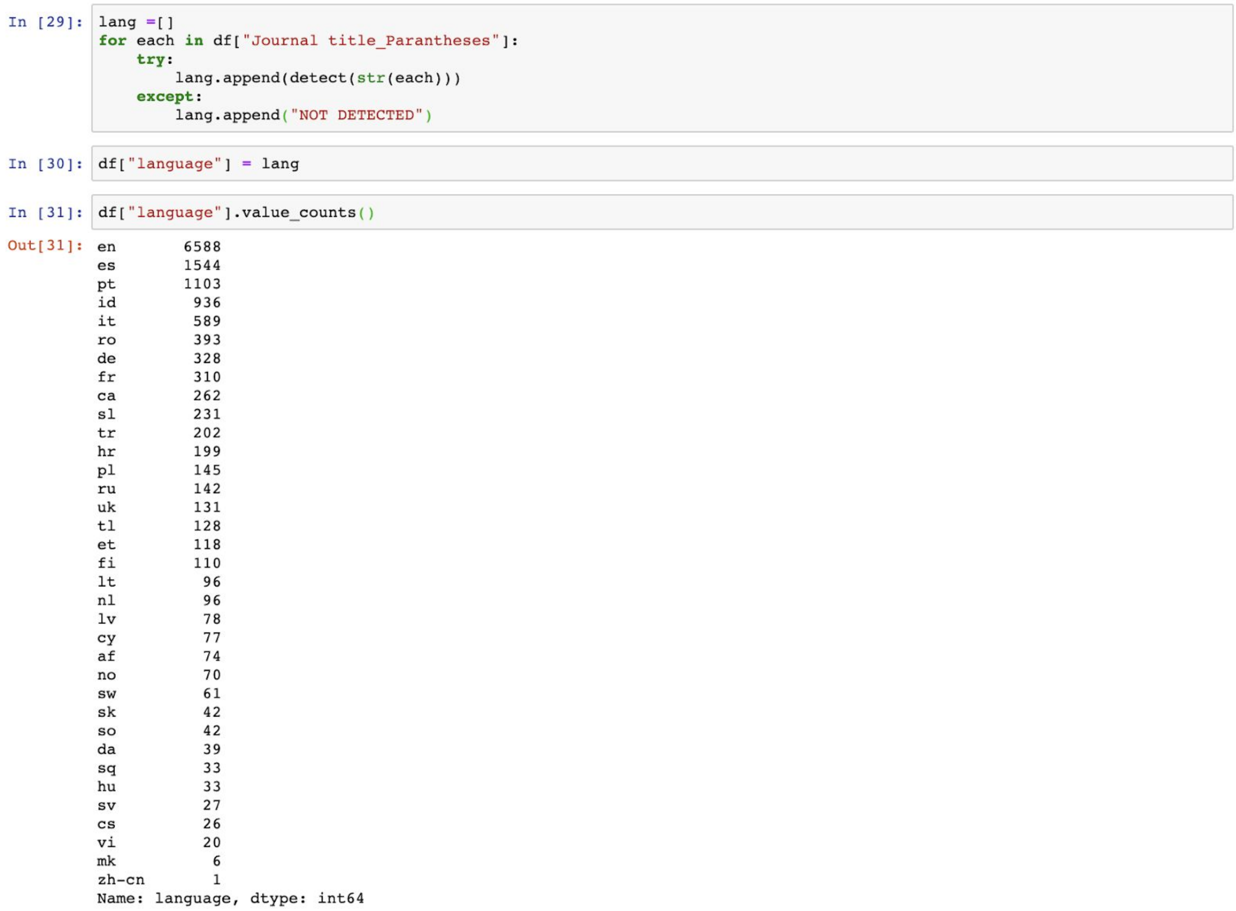
Drop function was used to remove the following columns:

- Plagiarism information URL
- Waiver policy information URL
- Publishing rights information URL
- Copyright information URL
- Full-text formats

- Added on Date
- Most Recent Articles Added
- Keywords
- Journal article submission fee
- Journal title
- Journal title_clean

Removing rows with different languages:
Most of the rows have texts that are in different languages that are read (right to left) RTL and languages that may cause encoding issues. So it is necessary to remove these rows. Also removed those that were just numbers for journal title

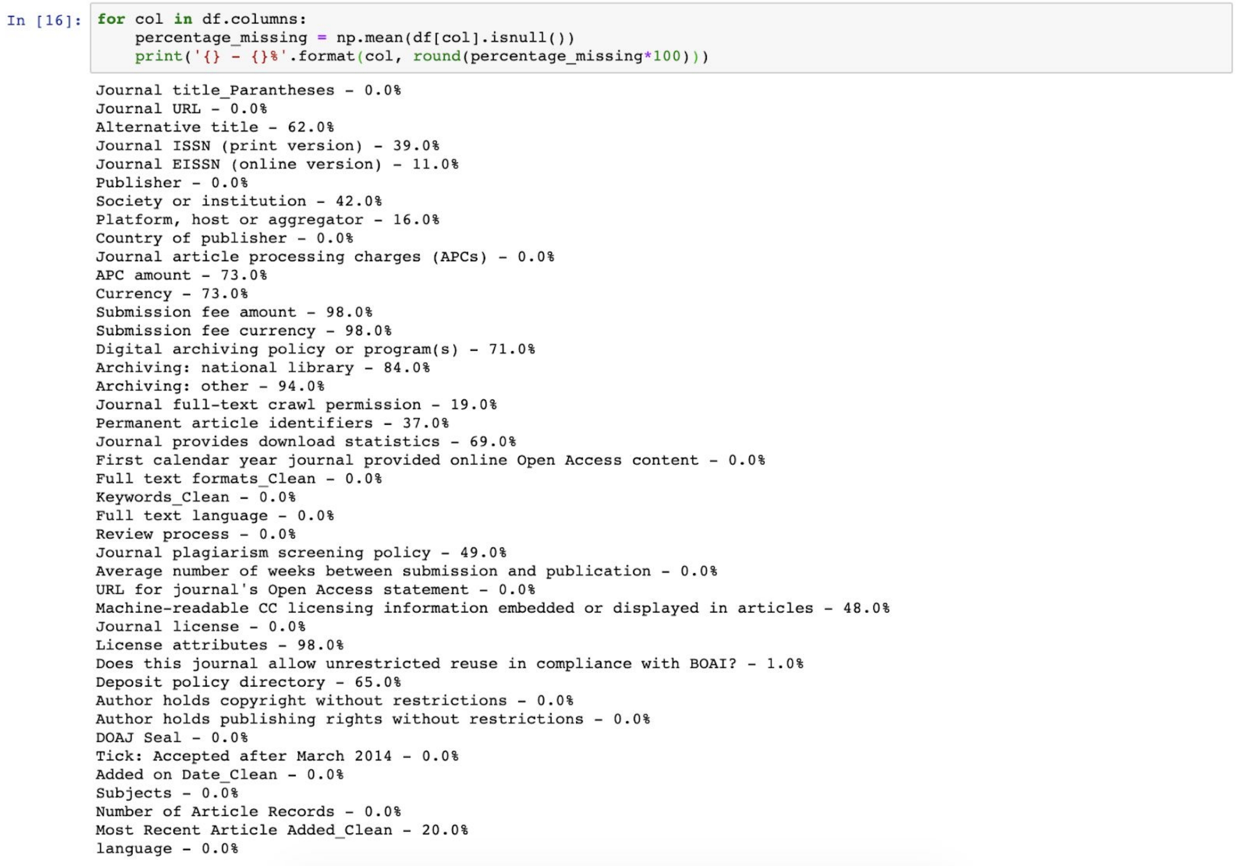
“langdetect” python package was used to detect the languages of journal titles. The screenshot below shows the code and the distribution of the languages.
Languages removed were Arabic, Bulgarian, Persian, Korean, Ukrainian, Urdu, Macedonian, Russian, Chinese.



```
In [29]: lang =[]
         for each in df["Journal title_Parantheses"]:
             try:
                 lang.append(detect(str(each)))
             except:
                 lang.append("NOT DETECTED")

In [30]: df["language"] = lang

In [31]: df["language"].value_counts()

Out[31]: en       6588
         es       1544
         pt       1103
         id        936
         it        589
         ro        393
         de        328
         fr        310
         ca        262
         sl        231
         tr        202
         hr        199
         pl        145
         ru        142
         uk        131
         tl        128
         et        118
         fi        110
         lt         96
         nl         96
         lv         78
         cy         77
         af         74
         no         70
         sw         61
         sk         42
         so         42
         da         39
         sq         33
         hu         33
         sv         27
         cs         26
         vi         20
         mk          6
         zh-cn       1
         Name: language, dtype: int64
```

3. Checking missing values in each column

Most of the columns in this dataset had many missing values. So it was necessary to find the columns with the maximum number of missing values and remove them. The Image below shows the code used and the percentage of missing values in each column.



```
In [16]: for col in df.columns:
             percentage_missing = np.mean(df[col].isnull())
             print('{} - {}%'.format(col, round(percentage_missing*100)))
```

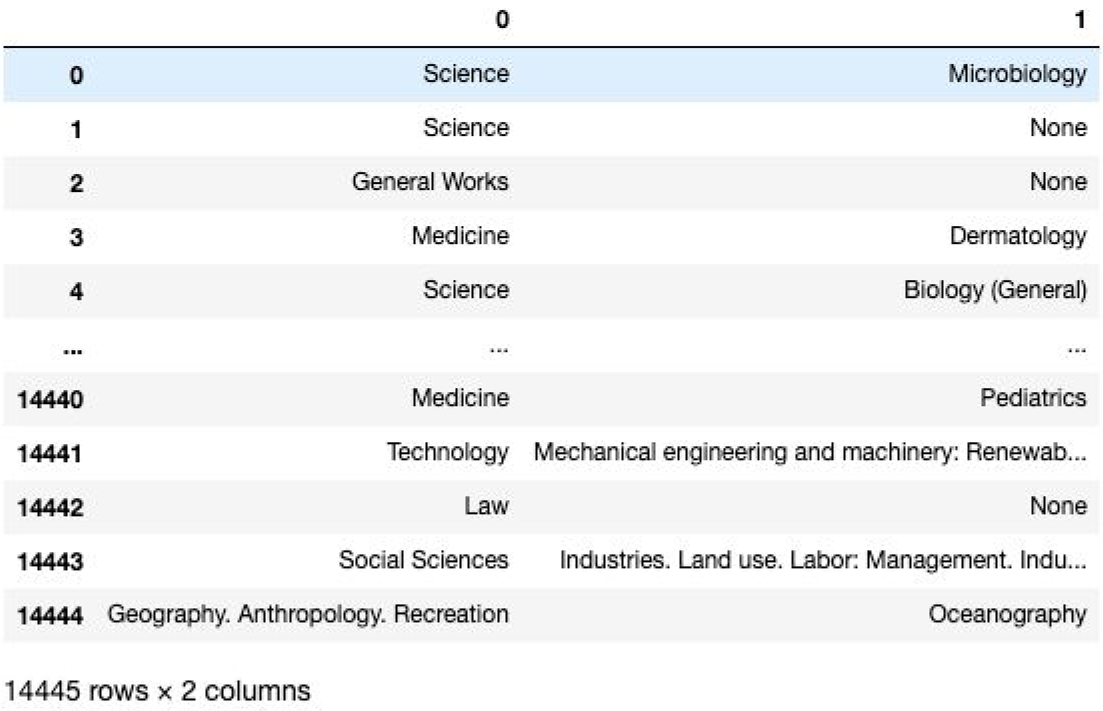
```
Journal title_Parantheses - 0.0%
Journal URL - 0.0%
Alternative title - 62.0%
Journal ISSN (print version) - 39.0%
Journal EISSN (online version) - 11.0%
Publisher - 0.0%
Society or institution - 42.0%
Platform, host or aggregator - 16.0%
Country of publisher - 0.0%
Journal article processing charges (APCs) - 0.0%
APC amount - 73.0%
Currency - 73.0%
Submission fee amount - 98.0%
Submission fee currency - 98.0%
Digital archiving policy or program(s) - 71.0%
Archiving: national library - 84.0%
Archiving: other - 94.0%
Journal full-text crawl permission - 19.0%
Permanent article identifiers - 37.0%
Journal provides download statistics - 69.0%
First calendar year journal provided online Open Access content - 0.0%
Full text formats_Clean - 0.0%
Keywords_Clean - 0.0%
Full text language - 0.0%
Review process - 0.0%
Journal plagiarism screening policy - 49.0%
Average number of weeks between submission and publication - 0.0%
URL for journal's Open Access statement - 0.0%
Machine-readable CC licensing information embedded or displayed in articles - 48.0%
Journal license - 0.0%
License attributes - 98.0%
Does this journal allow unrestricted reuse in compliance with BOAI? - 1.0%
Deposit policy directory - 65.0%
Author holds copyright without restrictions - 0.0%
Author holds publishing rights without restrictions - 0.0%
DOAJ Seal - 0.0%
Tick: Accepted after March 2014 - 0.0%
Added on Date_Clean - 0.0%
Subjects - 0.0%
Number of Article Records - 0.0%
Most Recent Article Added_Clean - 20.0%
language - 0.0%
```

Removing columns with more than 80% missing values:

- Submission fee amount
- Submission fee currency
- Archiving: national library
- Archiving: other
- License attributes

4. Split columns

Splitting “Subjects” into 2 columns and dropping “Subjects” column: Main_subject, Sub_subject



	0	1
0	Science	Microbiology
1	Science	None
2	General Works	None
3	Medicine	Dermatology
4	Science	Biology (General)
...	...	...
14440	Medicine	Pediatrics
14441	Technology	Mechanical engineering and machinery: Renewab...
14442	Law	None
14443	Social Sciences	Industries. Land use. Labor: Management. Indu...
14444	Geography. Anthropology. Recreation	Oceanography

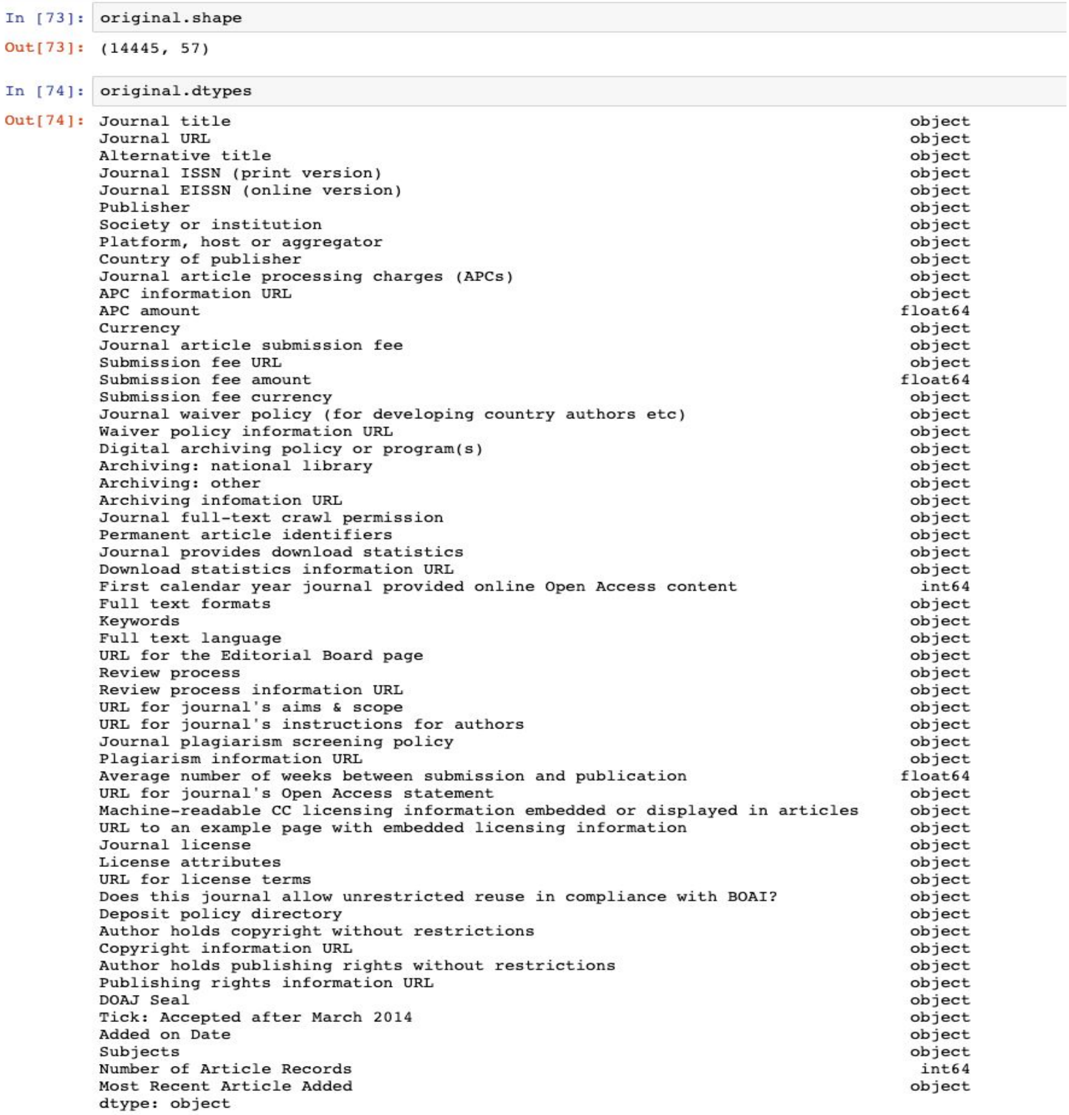
14445 rows × 2 columns

The shape of the dataset after cleaning : (14000, 39)

4. Results

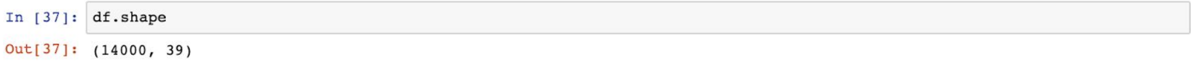
Overall changes

Before - Columns: 57, Rows: 14445

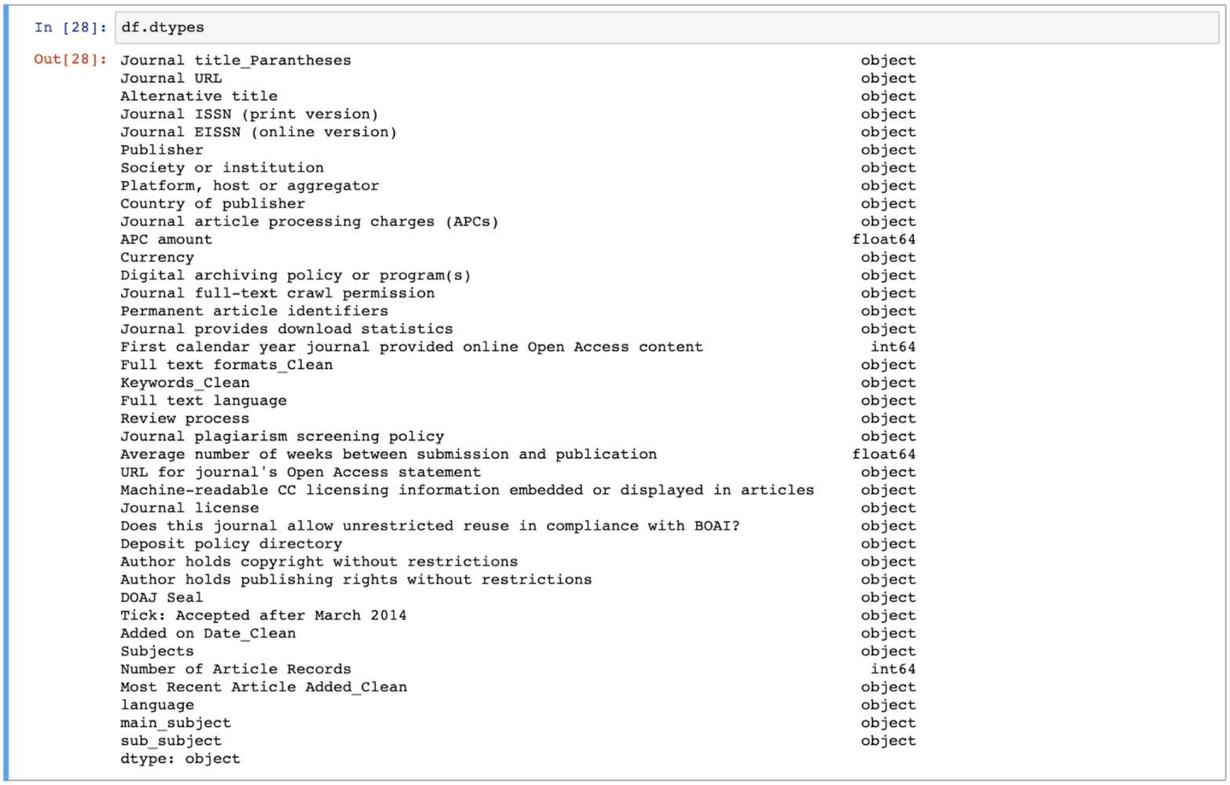


```
In [73]: original.shape
Out[73]: (14445, 57)

In [74]: original.dtypes
Out[74]: Journal title                                                                     object
         Journal URL                                                                       object
         Alternative title                                                                 object
         Journal ISSN (print version)                                                      object
         Journal EISSN (online version)                                                    object
         Publisher                                                                         object
         Society or institution                                                            object
         Platform, host or aggregator                                                      object
         Country of publisher                                                              object
         Journal article processing charges (APCs)                                         object
         APC information URL                                                               object
         APC amount                                                                       float64
         Currency                                                                          object
         Journal article submission fee                                                    object
         Submission fee URL                                                                object
         Submission fee amount                                                            float64
         Submission fee currency                                                           object
         Journal waiver policy (for developing country authors etc)                        object
         Waiver policy information URL                                                     object
         Digital archiving policy or program(s)                                            object
         Archiving: national library                                                       object
         Archiving: other                                                                  object
         Archiving infomation URL                                                          object
         Journal full-text crawl permission                                                object
         Permanent article identifiers                                                     object
         Journal provides download statistics                                              object
         Download statistics information URL                                               object
         First calendar year journal provided online Open Access content                    int64
         Full text formats                                                                 object
         Keywords                                                                          object
         Full text language                                                                object
         URL for the Editorial Board page                                                  object
         Review process                                                                    object
         Review process information URL                                                    object
         URL for journal's aims & scope                                                    object
         URL for journal's instructions for authors                                        object
         Journal plagiarism screening policy                                               object
         Plagiarism information URL                                                        object
         Average number of weeks between submission and publication                       float64
         URL for journal's Open Access statement                                           object
         Machine-readable CC licensing information embedded or displayed in articles       object
         URL to an example page with embedded licensing information                        object
         Journal license                                                                   object
         License attributes                                                                object
         URL for license terms                                                             object
         Does this journal allow unrestricted reuse in compliance with BOAI?               object
         Deposit policy directory                                                          object
         Author holds copyright without restrictions                                       object
         Copyright information URL                                                         object
         Author holds publishing rights without restrictions                               object
         Publishing rights information URL                                                 object
         DOAJ Seal                                                                         object
         Tick: Accepted after March 2014                                                   object
         Added on Date                                                                     object
         Subjects                                                                          object
         Number of Article Records                                                          int64
         Most Recent Article Added                                                         object
         dtype: object
```

After - Columns: 39, Rows: 14000



```
In [37]: df.shape
Out[37]: (14000, 39)
```



```
In [28]: df.dtypes
Out[28]: Journal title_Parantheses                                                       object
         Journal URL                                                                     object
         Alternative title                                                               object
         Journal ISSN (print version)                                                    object
         Journal EISSN (online version)                                                  object
         Publisher                                                                       object
         Society or institution                                                          object
         Platform, host or aggregator                                                    object
         Country of publisher                                                            object
         Journal article processing charges (APCs)                                       object
         APC amount                                                                     float64
         Currency                                                                        object
         Digital archiving policy or program(s)                                          object
         Journal full-text crawl permission                                              object
         Permanent article identifiers                                                   object
         Journal provides download statistics                                            object
         First calendar year journal provided online Open Access content                  int64
         Full text formats_Clean                                                         object
         Keywords_Clean                                                                  object
         Full text language                                                              object
         Review process                                                                  object
         Journal plagiarism screening policy                                             object
         Average number of weeks between submission and publication                     float64
         URL for journal's Open Access statement                                         object
         Machine-readable CC licensing information embedded or displayed in articles     object
         Journal license                                                                 object
         Does this journal allow unrestricted reuse in compliance with BOAI?             object
         Deposit policy directory                                                        object
         Author holds copyright without restrictions                                     object
         Author holds publishing rights without restrictions                             object
         DOAJ Seal                                                                       object
         Tick: Accepted after March 2014                                                 object
         Added on Date_Clean                                                             object
         Subjects                                                                        object
         Number of Article Records                                                        int64
         Most Recent Article Added_Clean                                                 object
         language                                                                        object
         main_subject                                                                    object
         sub_subject                                                                     object
         dtype: object
```

Table of Changes

Data Cleaning Step	**Column Changed**	**Number of Cells Changed**
value.trim()	● Journal title ● Society or Institution	● 660 ● 495
Drop consecutive whitespaces	● Journal title ● Society or Institution ● Keywords	● 14 ● 29 ● 24
value.toNumber()	● APC Amount ● Submission Fee Amount ● Number of Article Records	● 3992 ● 305 ● 14445
toString(toDate)	● Added on date_Clean ● Most Recent Article Added_Clean	● 14445 ● 11522

Clustering	● Journal title_Clean ● Full text formats_Clean ● Keywords_Clean	● 57 ● 9221 ● 924
Remove	● Submission fee URL ● APC Information URL ● Archiving Information URL ● Download statistics information URL ● URL for the Editorial Board page ● Review process information URL ● URL for journal’s aims & scope ● URL for journal’s instruction for authors ● URL to an example page with embedded licensing information ● URL for license terms ● Submission fee amount ● Submission fee currency ● Archiving: national library ● Archiving: other ● License attributes	

Integrity Constraints

a. Relational Schema:

SQL Exporter functionality of OpenRefine was used to get the final schema of our dataset into data_cleaning_final_dataset-csv.sql and then we explored the dataset in SQL and checked for Integrity Constraints.

b. Dataset description:

Journal title_Parantheses VARCHAR(255) NULL,
Journal URL VARCHAR(255) NOT NULL,

```
Alternative title VARCHAR(255) NULL,
Journal ISSN (print version) VARCHAR(255) NULL,
Journal EISSN (online version) VARCHAR(255) NULL,
Publisher VARCHAR(255) NULL,
Society or institution VARCHAR(255) NULL,
Platform, host or aggregator VARCHAR(255) NULL,
Country of publisher VARCHAR(255) NULL,
Journal article processing charges (APCs) VARCHAR(255) NULL,
APC amount NUMERIC NULL,
Currency VARCHAR(255) NULL,
Digital archiving policy or program(s) VARCHAR(255) NULL,
Journal full-text crawl permission VARCHAR(255) NULL,
Permanent article identifiers VARCHAR(255) NULL,
Journal provides download statistics VARCHAR(255) NULL,
First calendar year journal provided online Open Access content VARCHAR(255) NULL,
Full text formats_Clean VARCHAR(255) NULL,
Keywords_Clean VARCHAR(255) NULL,
Full text language VARCHAR(255) NULL,
Review process VARCHAR(255) NULL,
Journal plagiarism screening policy VARCHAR(255) NULL,
Average number of weeks between submission and publication VARCHAR(255) NULL,
URL for journal's Open Access statement VARCHAR(255) NULL,
Machine-readable CC licensing information embedded or displayed in articles VARCHAR(255) NULL,
Journal license VARCHAR(255) NULL,
Does this journal allow unrestricted reuse in compliance with BOAI? VARCHAR(255) NULL,
Deposit policy directory VARCHAR(255) NULL,
Author holds copyright without restrictions VARCHAR(255) NULL,
Author holds publishing rights without restrictions VARCHAR(255) NULL,
DOAJ Seal VARCHAR(255) NULL,
Tick: Accepted after March 2014 VARCHAR(3) NULL,
Added on Date_Clean DATE NULL,
Subjects VARCHAR(255) NULL,
Number of Article Records INT NULL,
Most Recent Article Added_Clean DATE NULL,
language VARCHAR(2) NULL,
main_subject VARCHAR(255) NULL,
sub_subject VARCHAR(255) NULL
```

- Every record in this dataset can be uniquely identified by the Journal URL column and therefore it is the PRIMARY KEY for our schema which is NOT NULL and UNIQUE.
- For this particular dataset we can use the following entities to create relationships:
    - Journal URL
    - Publisher
    - Society or institution
    - Country of publisher
    - Main_subject
    - Number of Article Records

  Here Journal URL is the foreign key and can be used to uniquely identify publisher, institution, country of publisher, main_subject and number of articles. This relationship brings all the necessary primary information together and makes querying easy. New entities can be added if required.

c. Integrity Constraints:

We identified some integrity constraints for our dataset.

SQL Queries that violate the ICs:

1)

INSERT INTO clean_data_set_lang(Journal URL, Journal title_Clean, Tick: Accepted after March 2014) values (www.pmadan.com, priyanshu madan, none)

According to our ICs in this data, the length of "Tick: Accepted after March 2014" can only be 3 or less but we have given an input more than 3 which violates the IC.

2)

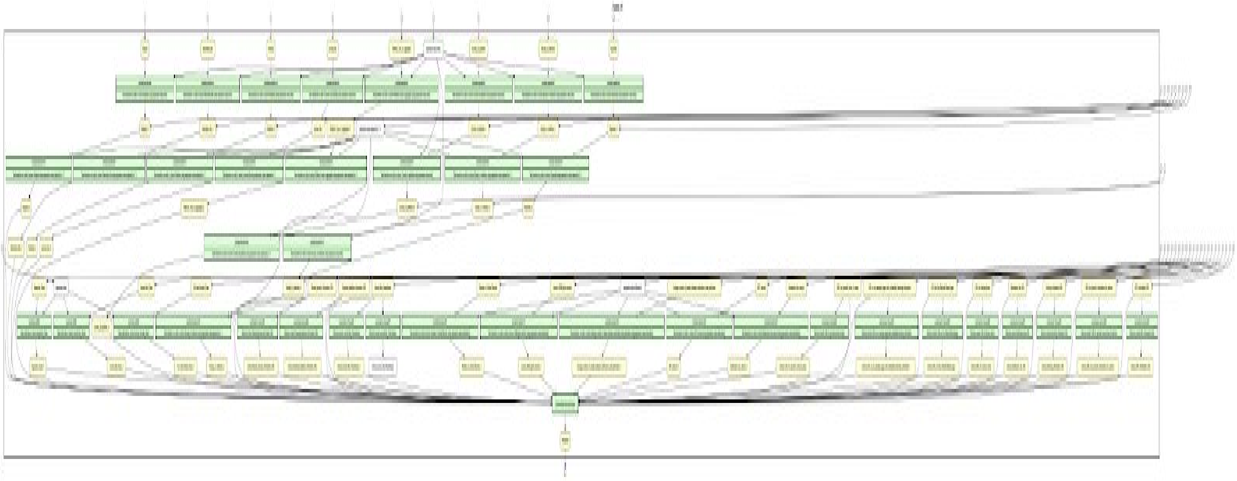
INSERT INTO clean_data_set_lang(Journal URL, Journal title_Clean, Tick: Accepted after March 2014, Number of Article Records) values (www.pmadan.com, priyanshu madan, Yes, 18.0)

According to our ICs in this data, "Number of Article Records" can only have integers but we have given a float input which violates the IC.

3) INSERT INTO clean_data_set_lang(Journal URL, Journal title_Clean, Tick: Accepted after March 2014, Number of Article Records) values (NA, priyanshu madan, Yes, 18)

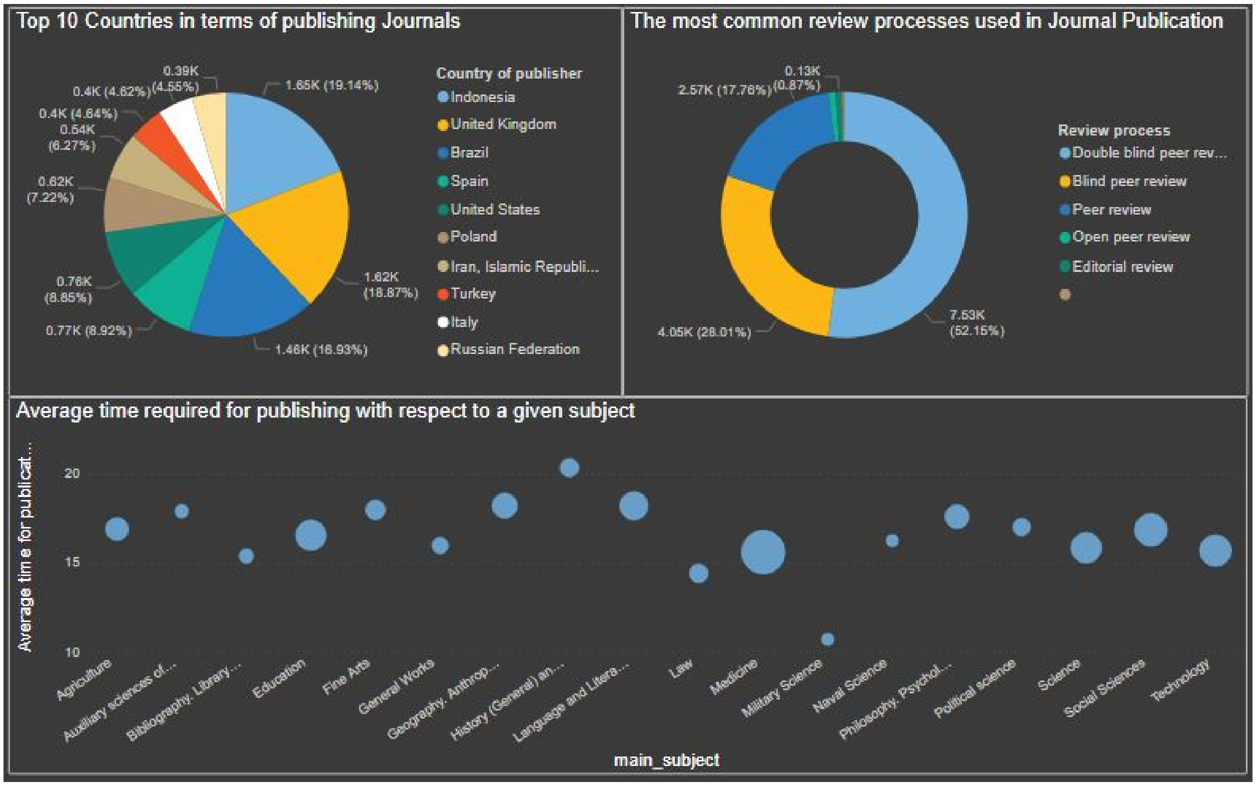
According to our ICs in this data, "Journal URL" cannot be NA as it is a primary key but we have given a null value input which violates the IC.

Data Cleaning Workflow



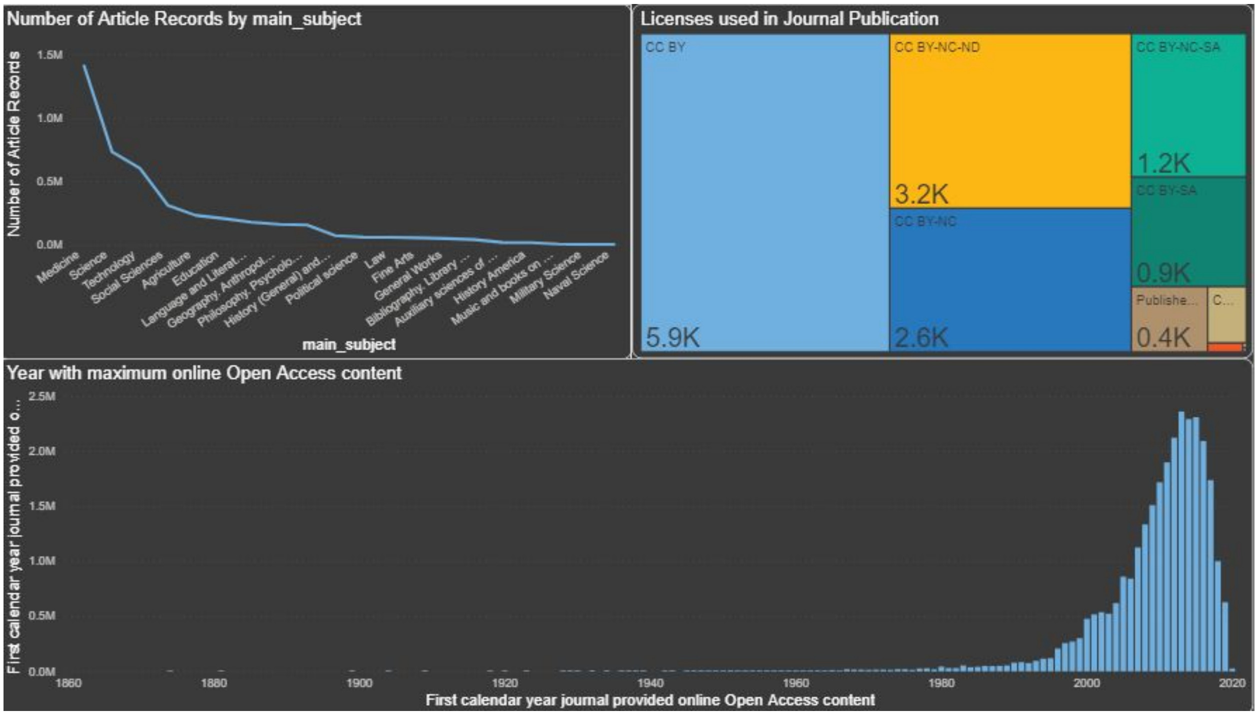
PowerBI Dashboard

After thoroughly cleaning the dataset, we wanted to gather insights from the dataset. The following insights were generated from the clean dataset.



According to the data from the DOAJ website, Indonesia, United Kingdom, and Brazil happen to be the top 3 countries in terms of publishing journals.

The visualization in the bottom suggests the average time required in publishing wrt a given subject. The maximum amount of time is taken by journals with topics relating to the general history and history of Europe while the minimum is taken by the journals of military science.



The visualization on the top left corner indicates that maximum articles according to the dataset are written about medicine followed by science and technology.

Talking about publishing licenses, CC BY happens to be the most common one followed by CC BY-NC-ND to the close second.

Lastly, since 1993 there has been a gradual increase in the number of journals opting for Open Access content and was seen to be a maximum 20 years later in the year 2013.

5. Conclusion and Future Work

This project helped us apply the learnings of this course on a real data analysis assignment. When we read the phrase that “data cleansing takes 90% of the time in the real world undertaking” first time on the internet we thought that it is an exaggeration, however now that we've experienced it first-hand, we modified our thoughts.

There were certain challenges we faced while working with the dataset and performing the data cleaning process including dealing with special characters and with data in different languages in some of the cells of the workbook. Dealing with such problems helped improve our decision making and analytical thinking.

One of the possibilities of future work in this project could be to utilise other data cleaning tools and do a comparative study of the tools in order to understand the difference and also similarities that the same offer.