3 Simplifying the column names (CSV header)

The CSV types and their columns have a Brazilian Portuguese description in SciELO's public reports documentation page[1].

However, the column names have some issues:

- Sometimes they're way too long;
- Almost always they have some whitespace or other non-alphanumeric character;
- Some names have trailing whitespaces;
- There are multiple languages in the `journals_kbart.csv`, following the `Brazilian Portuguese Name (english_name)` format;
- They might include redundant/ambiguous/misleading parts.

In summary, they're difficult to deal with when we're performing some exploratory data analysis or otherwise using them in the middle of some source code, in almost any language. Our goal is to simplify that to keep it similar to a `snake_case` format.

In [1]:

```python
import csv
from glob import glob
```

In [2]:

```python
import numpy as np
import pandas as pd
pd.options.display.max_rows = 200 # Default is 60
```

3.1 Current rows

From all the column titles from every CSV file, there are a lot of names that appear more than once. The first 5 columns for every CSV type is:

In [3]:

```python
for fname in glob("tabs_network/*.csv"):
    with open(fname) as f:
        cr = csv.reader(f)
        print(next(cr)[:5])
```

```
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['Título do Periódico (publication_title)', 'ISSN impresso (print_identifier)', 'ISSN
online (online_identifier)', 'Data do primeiro fascículo (date_first_issue_online)',
'volume do primeiro fascículo (num_first_vol_online)']
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
['extraction date', 'study unit', 'collection', 'ISSN SciELO', "ISSN's"]
```

Joining the column headers from every CSV, the full set of names we obtain is:

In [4]:

```python
names = set()
for fname in glob("tabs_network/*.csv"):
    with open(fname) as f:
        cr = csv.reader(f)
        names.update(next(cr))
np.array(sorted(names))
```

[1] http://docs.scielo.org/projects/scielo-processing/pt/latest/public_reports.html

Out [4]:
```
array(['+6 authors', '0 authors', '1 author', '2 authors', '3 authors',
       '4 authors', '5 authors',
       'Data do primeiro fascículo (date_first_issue_online)',
       'Data do último fascículo publicado (date_last_issue_online)',
       'ID de publicação pai (parent_publication_title_id)',
       'ID de publicação prévia (preceding_publication_title_id)',
       'ID do periódico no SciELO (title_id)', 'ISSN SciELO',
       'ISSN impresso (print_identifier)',
       'ISSN online (online_identifier)', "ISSN's",
       'Título do Periódico (publication_title)', 'accesses to abstract',
       'accesses to epdf', 'accesses to html', 'accesses to pdf',
       'accesses year', 'alpha frequency', 'altmetrics url', 'authors',
       'citable documents', 'citable documents at 2013',
       'citable documents at 2014', 'citable documents at 2015',
       'citable documents at 2016', 'citable documents at 2017',
       'citable documents at 2018', 'cobertura (coverage_depth)',
       'collection',
       'data de publicação monográfica impressa (date_monograph_published_print)',
       'data de publicação monográfica online (date_monograph_published_online)',
       'date of the first document', 'date of the last document',
       'document accepted at', 'document accepted at day',
       'document accepted at month', 'document accepted at year',
       'document affiliation city', 'document affiliation country',
       'document affiliation country ISO 3166',
       'document affiliation instituition', 'document affiliation state',
       'document author', 'document author affiliation city',
       'document author affiliation country',
       'document author affiliation state', 'document author institution',
       'document en', 'document es', 'document is citable',
       'document languages', 'document license',
       'document other languages', 'document pt',
       'document published as ahead of print at',
       'document published as ahead of print at day',
       'document published as ahead of print at month',
       'document published as ahead of print at year',
       'document published at', 'document published at day',
       'document published at month', 'document published at year',
       'document published in SciELO at',
       'document published in SciELO at day',
       'document published in SciELO at month',
       'document published in SciELO at year',
       'document publishing ID (PID SciELO)', 'document publishing year',
       'document reviewed at', 'document reviewed at day',
       'document reviewed at month', 'document reviewed at year',
       'document submitted at', 'document submitted at day',
       'document submitted at month', 'document submitted at year',
       'document type', 'document updated in SciELO at',
       'document updated in SciELO at day',
       'document updated in SciELO at month',
       'document updated in SciELO at year', 'documents at 2013',
       'documents at 2014', 'documents at 2015', 'documents at 2016',
       'documents at 2017', 'documents at 2018',
       'edição de monografia (monograph_edition)',
       'english documents at 2013 ', 'english documents at 2014 ',
       'english documents at 2015 ', 'english documents at 2016 ',
       'english documents at 2017 ', 'english documents at 2018 ',
       'extraction date', 'google scholar h5 2013 ',
       'google scholar h5 2014 ', 'google scholar h5 2015 ',
       'google scholar h5 2016 ', 'google scholar h5 2017 ',
       'google scholar h5 2018 ', 'google scholar m5 2013 ',
       'google scholar m5 2014 ', 'google scholar m5 2015 ',
       'google scholar m5 2016 ', 'google scholar m5 2017 ',
       'google scholar m5 2018 ', 'inclusion year at SciELO',
       'informação de embargo (embargo_info)',
       'informação sobre cobertura (coverage_notes)',
       'issue of the first document', 'issue of the last document',
       'issues at 2013', 'issues at 2014', 'issues at 2015',
       'issues at 2016', 'issues at 2017', 'issues at 2018',
       'nome do publicador (publisher_name)',
       'numeric frequency (in months)',
       'número do primeiro fascículo (num_first_issue_online)',
       'número do último fascículo publicado (num_last_issue_online)',
       'other language documents at 2013 ',
       'other language documents at 2014 ',
       'other language documents at 2015 ',
       'other language documents at 2016 ',
       'other language documents at 2017 ',
       'other language documents at 2018 ', 'pages',
       'portuguese documents at 2013 ', 'portuguese documents at 2014 ',
       'portuguese documents at 2015 ', 'portuguese documents at 2016 ',
       'portuguese documents at 2017 ', 'portuguese documents at 2018 ',
       'primeiro autor (first_author)', 'primeiro editor (first_editor)',
       'publisher name', 'publishing year', 'references',
       'regular issues at 2013', 'regular issues at 2014',
       'regular issues at 2015', 'regular issues at 2016',
       'regular issues at 2017', 'regular issues at 2018', 'score',
       'short title ISO', 'short title SciELO',
       'spanish documents at 2013 ', 'spanish documents at 2014 ',
       'spanish documents at 2015 ', 'spanish documents at 2016 ',
       'spanish documents at 2017 ', 'spanish documents at 2018 ',
       'status change date', 'status change day', 'status change month',
       'status change reason', 'status change year', 'status changed to',
       'stopping reason', 'stopping year at SciELO', 'study unit',
       'tipo de acesso (access_type)',
       'tipo de publicação (publication_type)', 'title + subtitle SciELO',
       'title PubMed', 'title at SciELO', 'title current status',
       'title is agricultural sciences',
       'title is applied social sciences', 'title is biological sciences',
       'title is engineering', 'title is exact and earth sciences',
       'title is health sciences', 'title is human sciences',
       'title is linguistics, letters and arts',
       'title is multidisciplinary', 'title thematic areas',
       'total accesses', 'total of documents', 'total of issues',
       'total of regular issues', 'url de fascículos (title_url)',
       'use license', 'volume de monografia (monograph_volume)',
       'volume do primeiro fascículo (num_first_vol_online)',
       'volume do último fascículo publicado (num_last_vol_online)',
       'volume of the first document', 'volume of the last document'],
      dtype='<U72')
```

Sometimes the columns names have some misleading stuff we can fix, like:

- Extra trailing whitespace
- Distinct/mixed letter cases
- Redundant parentheses structure like `Plain text column description in Portuguese (snake_case_descr_in_english)`
- Symbols like `'` and `+`

We can fix that by keeping only the parenthesized code, removing some less meaningful common words, shortening some lenghty words, and performing some replacements:

In [5]:

```
def normalize_column_title(name):
    import re
    name_unbracketed = re.sub(r".*\((.*)\)", r"\1",
                              name.replace("(in months)", "in_months"))
    words = re.sub("[^a-z0-9+_ ]", "", name_unbracketed.lower()).split()
    ignored_words = ("at", "the", "of", "and", "google", "scholar", "+")
    replacements = {
        "document": "doc",
        "documents": "docs",
        "frequency": "freq",
        "language": "lang",
    }
    return "_".join(replacements.get(word, word)
                    for word in words if word not in ignored_words) \
              .replace("title_is", "is")
```

With Pandas, its use should be straightforward.

In [6]:

```
network_journals = pd.read_csv("tabs_network/journals.csv") \
                     .rename(columns=normalize_column_title)
network_journals.columns
```

Out [6]:

```
Index(['extraction_date', 'study_unit', 'collection', 'issn_scielo', 'issns',
       'title_scielo', 'title_thematic_areas', 'is_agricultural_sciences',
       'is_applied_social_sciences', 'is_biological_sciences',
       'is_engineering', 'is_exact_earth_sciences', 'is_health_sciences',
       'is_human_sciences', 'is_linguistics_letters_arts',
       'is_multidisciplinary', 'title_current_status', 'title_subtitle_scielo',
       'short_title_scielo', 'short_iso', 'title_pubmed', 'publisher_name',
       'use_license', 'alpha_freq', 'numeric_freq_in_months',
       'inclusion_year_scielo', 'stopping_year_scielo', 'stopping_reason',
       'date_first_doc', 'volume_first_doc', 'issue_first_doc',
       'date_last_doc', 'volume_last_doc', 'issue_last_doc', 'total_issues',
       'issues_2018', 'issues_2017', 'issues_2016', 'issues_2015',
       'issues_2014', 'issues_2013', 'total_regular_issues',
       'regular_issues_2018', 'regular_issues_2017', 'regular_issues_2016',
       'regular_issues_2015', 'regular_issues_2014', 'regular_issues_2013',
       'total_docs', 'docs_2018', 'docs_2017', 'docs_2016', 'docs_2015',
       'docs_2014', 'docs_2013', 'citable_docs', 'citable_docs_2018',
       'citable_docs_2017', 'citable_docs_2016', 'citable_docs_2015',
       'citable_docs_2014', 'citable_docs_2013', 'portuguese_docs_2018',
       'portuguese_docs_2017', 'portuguese_docs_2016', 'portuguese_docs_2015',
       'portuguese_docs_2014', 'portuguese_docs_2013', 'spanish_docs_2018',
       'spanish_docs_2017', 'spanish_docs_2016', 'spanish_docs_2015',
       'spanish_docs_2014', 'spanish_docs_2013', 'english_docs_2018',
       'english_docs_2017', 'english_docs_2016', 'english_docs_2015',
       'english_docs_2014', 'english_docs_2013', 'other_lang_docs_2018',
       'other_lang_docs_2017', 'other_lang_docs_2016', 'other_lang_docs_2015',
       'other_lang_docs_2014', 'other_lang_docs_2013', 'h5_2018', 'h5_2017',
       'h5_2016', 'h5_2015', 'h5_2014', 'h5_2013', 'm5_2018', 'm5_2017',
       'm5_2016', 'm5_2015', 'm5_2014', 'm5_2013'],
      dtype='object')
```

The map of names is:

In [7]:

```
name_map = pd.DataFrame(pd.Series({name: normalize_column_title(name)
                                   for name in names})
                        .rename("simple_name"))
name_map.sort_values("simple_name")
```

Out [7]:

	simple_name
+6 authors	+6_authors
0 authors	0_authors
1 author	1_author
2 authors	2_authors
3 authors	3_authors
4 authors	4_authors
5 authors	5_authors
tipo de acesso (access_type)	access_type
accesses to abstract	accesses_to_abstract
accesses to epdf	accesses_to_epdf
accesses to html	accesses_to_html
accesses to pdf	accesses_to_pdf
accesses year	accesses_year
alpha frequency	alpha_freq
altmetrics url	altmetrics_url
authors	authors
citable documents	citable_docs
citable documents at 2013	citable_docs_2013
citable documents at 2014	citable_docs_2014
citable documents at 2015	citable_docs_2015
citable documents at 2016	citable_docs_2016
citable documents at 2017	citable_docs_2017
citable documents at 2018	citable_docs_2018
collection	collection
cobertura (coverage_depth)	coverage_depth
informação sobre cobertura (coverage_notes)	coverage_notes
date of the first document	date_first_doc
Data do primeiro fascículo (date_first_issue_o...	date_first_issue_online
date of the last document	date_last_doc
Data do último fascículo publicado (date_last_...	date_last_issue_online
data de publicação monográfica online (date_mo...	date_monograph_published_online
data de publicação monográfica impressa (date_...	date_monograph_published_print
document accepted at	doc_accepted
document accepted at day	doc_accepted_day
document accepted at month	doc_accepted_month
document accepted at year	doc_accepted_year
document affiliation city	doc_affiliation_city
document affiliation country	doc_affiliation_country
document affiliation country ISO 3166	doc_affiliation_country_iso_3166
document affiliation instituition	doc_affiliation_instituition
document affiliation state	doc_affiliation_state
document author	doc_author
document author affiliation city	doc_author_affiliation_city
document author affiliation country	doc_author_affiliation_country
document author affiliation state	doc_author_affiliation_state
document author institution	doc_author_institution
document en	doc_en
document es	doc_es
document is citable	doc_is_citable

Continued on next page

	simple_name
document languages	doc_languages
document license	doc_license
document other languages	doc_other_languages
document pt	doc_pt
document published at	doc_published
document published as ahead of print at	doc_published_as_ahead_print
document published as ahead of print at day	doc_published_as_ahead_print_day
document published as ahead of print at month	doc_published_as_ahead_print_month
document published as ahead of print at year	doc_published_as_ahead_print_year
document published at day	doc_published_day
document published in SciELO at	doc_published_in_scielo
document published in SciELO at day	doc_published_in_scielo_day
document published in SciELO at month	doc_published_in_scielo_month
document published in SciELO at year	doc_published_in_scielo_year
document published at month	doc_published_month
document published at year	doc_published_year
document publishing year	doc_publishing_year
document reviewed at	doc_reviewed
document reviewed at day	doc_reviewed_day
document reviewed at month	doc_reviewed_month
document reviewed at year	doc_reviewed_year
document submitted at	doc_submitted
document submitted at day	doc_submitted_day
document submitted at month	doc_submitted_month
document submitted at year	doc_submitted_year
document type	doc_type
document updated in SciELO at	doc_updated_in_scielo
document updated in SciELO at day	doc_updated_in_scielo_day
document updated in SciELO at month	doc_updated_in_scielo_month
document updated in SciELO at year	doc_updated_in_scielo_year
documents at 2013	docs_2013
documents at 2014	docs_2014
documents at 2015	docs_2015
documents at 2016	docs_2016
documents at 2017	docs_2017
documents at 2018	docs_2018
informação de embargo (embargo_info)	embargo_info
english documents at 2013	english_docs_2013
english documents at 2014	english_docs_2014
english documents at 2015	english_docs_2015
english documents at 2016	english_docs_2016
english documents at 2017	english_docs_2017
english documents at 2018	english_docs_2018
extraction date	extraction_date
primeiro autor (first_author)	first_author
primeiro editor (first_editor)	first_editor
google scholar h5 2013	h5_2013
google scholar h5 2014	h5_2014
google scholar h5 2015	h5_2015
google scholar h5 2016	h5_2016
google scholar h5 2017	h5_2017
google scholar h5 2018	h5_2018
inclusion year at SciELO	inclusion_year_scielo
title is agricultural sciences	is_agricultural_sciences
title is applied social sciences	is_applied_social_sciences

Continued on next page

	simple_name
title is biological sciences	is_biological_sciences
title is engineering	is_engineering
title is exact and earth sciences	is_exact_earth_sciences
title is health sciences	is_health_sciences
title is human sciences	is_human_sciences
title is linguistics, letters and arts	is_linguistics_letters_arts
title is multidisciplinary	is_multidisciplinary
ISSN SciELO	issn_scielo
ISSN's	issns
issue of the first document	issue_first_doc
issue of the last document	issue_last_doc
issues at 2013	issues_2013
issues at 2014	issues_2014
issues at 2015	issues_2015
issues at 2016	issues_2016
issues at 2017	issues_2017
issues at 2018	issues_2018
google scholar m5 2013	m5_2013
google scholar m5 2014	m5_2014
google scholar m5 2015	m5_2015
google scholar m5 2016	m5_2016
google scholar m5 2017	m5_2017
google scholar m5 2018	m5_2018
edição de monografia (monograph_edition)	monograph_edition
volume de monografia (monograph_volume)	monograph_volume
número do primeiro fascículo (num_first_issue_...	num_first_issue_online
volume do primeiro fascículo (num_first_vol_on...	num_first_vol_online
número do último fascículo publicado (num_last...	num_last_issue_online
volume do último fascículo publicado (num_last...	num_last_vol_online
numeric frequency (in months)	numeric_freq_in_months
ISSN online (online_identifier)	online_identifier
other language documents at 2013	other_lang_docs_2013
other language documents at 2014	other_lang_docs_2014
other language documents at 2015	other_lang_docs_2015
other language documents at 2016	other_lang_docs_2016
other language documents at 2017	other_lang_docs_2017
other language documents at 2018	other_lang_docs_2018
pages	pages
ID de publicação pai (parent_publication_title...	parent_publication_title_id
document publishing ID (PID SciELO)	pid_scielo
portuguese documents at 2013	portuguese_docs_2013
portuguese documents at 2014	portuguese_docs_2014
portuguese documents at 2015	portuguese_docs_2015
portuguese documents at 2016	portuguese_docs_2016
portuguese documents at 2017	portuguese_docs_2017
portuguese documents at 2018	portuguese_docs_2018
ID de publicação prévia (preceding_publication...	preceding_publication_title_id
ISSN impresso (print_identifier)	print_identifier
Título do Periódico (publication_title)	publication_title
tipo de publicação (publication_type)	publication_type
publisher name	publisher_name
nome do publicador (publisher_name)	publisher_name
publishing year	publishing_year
references	references
regular issues at 2013	regular_issues_2013

Continued on next page

	simple_name
regular issues at 2014	regular_issues_2014
regular issues at 2015	regular_issues_2015
regular issues at 2016	regular_issues_2016
regular issues at 2017	regular_issues_2017
regular issues at 2018	regular_issues_2018
score	score
short title ISO	short_iso
short title SciELO	short_title_scielo
spanish documents at 2013	spanish_docs_2013
spanish documents at 2014	spanish_docs_2014
spanish documents at 2015	spanish_docs_2015
spanish documents at 2016	spanish_docs_2016
spanish documents at 2017	spanish_docs_2017
spanish documents at 2018	spanish_docs_2018
status change date	status_change_date
status change day	status_change_day
status change month	status_change_month
status change reason	status_change_reason
status change year	status_change_year
status changed to	status_changed_to
stopping reason	stopping_reason
stopping year at SciELO	stopping_year_scielo
study unit	study_unit
title current status	title_current_status
ID do periódico no SciELO (title_id)	title_id
title PubMed	title_pubmed
title at SciELO	title_scielo
title + subtitle SciELO	title_subtitle_scielo
title thematic areas	title_thematic_areas
url de fascículos (title_url)	title_url
total accesses	total_accesses
total of documents	total_docs
total of issues	total_issues
total of regular issues	total_regular_issues
use license	use_license
volume of the first document	volume_first_doc
volume of the last document	volume_last_doc

There's no overlap in these new names:

In [8]:
```
name_map.shape[0] == len(names)
```

Out [8]:
```
True
```