Collaborative Data Science within Gov’t of Canada

Development of R libraries for common tasks with Open Canada data

Jonathan Dench, Research Analyst, Results Division , Treasury Board of Canada Secretariat
Dmitry Gorodnichy, Research Data Scientist, Chief Data Office, Canada Border Services Agency
Patrick Little, Advisor, Open Government, Treasury Board of Canada Secretariat
Joseph Stinziano, Science Analyst, Canadian Food Inspection Agency

Slido: r4gc

Statistics Canada's 2021 International Methodology Symposium
29 October 2021, Ottawa

Outline

- Raison d'être
- Vision
- Why R (for data science collaboration)
- GC collaborative platforms (for growing technical knowledge)
- Key outputs (so far)
- What’s next
- Appendices: demos and technical details

Raison d'être

- In GoC, we are working on the same data science problems
  - Working with the same data (eg. Geospatial, StatCan, open.canada.ca)
  - Developing many similar visualizations, analyses and reporting tools
  - Addressing many of same data engineering and data mining challenges

- Challenges
  - Often, data scientists end up “reinventing the wheel”, and not able to catch-up with rapidly growing development of data science tools
  - Lack of collaboration and peer-reviewing creates the risk of being inefficient, producing suboptimal solutions

- Much more can be achieved, if we leverage each other’s work!
  - Discussed at GC Data 2021 Conf., Data Engineering workshop

Vision

To ensure standardized and consistent approaches to data science across the GoC, we need:

1. To grow and maintain our skills and knowledgebase
2. To build codes and tools for common data science problems
   - Contributed, reviewed and maintained by GoC data science community
   - Open & free - available to any data scientist who needs them

By leveraging what is the best and already available within GC:

1. Collaboration platforms: **gccode**, **gccollab**, **gcwiki, github**
2. Programming environment: **R**

Why R ?

1. Advanced graphics with *ggplot2* and its extensions
2. Automated report/tutorials/textbooks generation with *RMarkdown*
3. Streamlined package development with *devtools*
4. Streamlined Interactive interfaces and dashboards development and deployment with *Shiny*
5. “Best for geo-computation”
6. Common tidy design shared across packages
7. Curated peer-tested repo of packages at CRAN
8. RStudio IDE (Integrated Development Environment) on desktop and cloud (rstudio.cloud)
9. Full support and inter-operability with Python from the same IDE
10. Global RStudio-led movement for R education and advancement (rstudio.com)

https://geocompr.robinlovelace.net/intro.html#why-use-r-for-geocomputation

https://gccollab.ca/discussion/view/7404883/why-r

Collaborative Platforms

GC restricted:

- https://gccode.ssc-spc.gc.ca/r4gc/
- https://wiki.gccollab.ca/UseR!
- https://gccollab.ca/groups/profile/7391537/Use-R
    - 'Lunch and Learn' Data Science with R: Friday Meet-ups

Public facing:

- https://github.com/open-canada
    - UNCLASSIFIED material for Lunch and Learns
    - Apps (e.g. https://open-canada.github.io/Apps/atip)
- CRAN Views (ideal for finished packages)

GitLab Projects Groups More Search or jump to...

r4gc > Details

r4gc

Group ID: 7049

New subgroup | New project

Group of Data Scientists and collaborators working with R for sharing Codes and Knowledgebase. See GCcollab group for more details: https://gccollab.ca/groups/profile/7391537/Use-R

Subgroups and projects | Shared projects | Archived projects

Search by name | Last created

GC packages Owner
Packages we are building from "raw" codes. Join the effort ! See /packages101 and https://gccollab.ca/di...
0 5 1

Codes Owner
Various "raw" R codes contributed by GC community.
0 7 2

Resources Owner
Books, tutorials, presentations, blogs on R. NB: This folder contains subfolders that are visible only to log...
0 4 1

- **IntroSpatialAnalysis**
  Tutorials and Codes for Geo/Spatial coding and visualization in R. See https://gccol...
  0 — 1 day ago
- **howTo**
  This is a project that will be developed as a series of RMarkdown documents, typic...
  0 — 2 months ago
- **meetings**
  This is what we discuss at our weekly meet-ups
  0 — 4 months ago
- **gccode101**
  You have questions on how to use GCCode or git? - Here you'll find the answers!
  4 — 1 month ago

GCwiki

English Log in

Page Discussion Read View source View history Search wiki

GCwiki

Connecting people and ideas

Main Page
Browse categories
Random page
Help

Actions/Tools

Special pages

Tools

Related changes
Printable version
Permanent link

Print/export

Download as PDF
Printable version

GCaccount
GCcollab
GCmessage

UseR!

Data Science Communities of Practice - UseR!

This page provides the list of discussions organized by the GCcollab's Use R! group. Please consider contributing to those discussions by joining the Use R! group and participating in group's weekly "Lunch and Learn Data Science with R" meetups.

- General topics:
  - Why R?
  - Best way to start (and keep learning) R
  - Events and Forums for R users
  - From Excel to R
  - R with Python (and other languages/tools)
  - Efficient programming in R (coding style, memory-efficient coding, collaboration-ready codes, source control)
  - **data.table** for efficient data processing
  - Reading various kinds of data in R
  - Open R codes for GC: on GCcode and GitHub
  - RStudio news and tricks
- Specialized topics:
  - **ggplot2** and its extensions for data visualization
  - **Shiny** for Interactive Data Visualization, Analysis and Web App development
  - **R Markdown** for automated and reproducible data science
  - *Record Linking* and other Data Engineering tasks in R
  - *Geo/Spatial* coding and visualization in R
  - *Text Analysis* in R
  - *Machine Learning and Modeling* in R
- Webinars and Tutorials (NB: you need to join the "Lunch and Learn Data Science with R" meetups group to access recordings of these sessions)
  - 30 Jul 2021: Geospatial data tools in R (code)
  - 16 Jul 2021: Dual Coding - Python and R unite !
  - 9 Jul 2021: Exploring ggplots (recording, code)
  - 2 Jul 2021: Parsing GC Tables (code)
  - 25 Jun 2021: Using the Open Government Portal API within R (recording, code on github.com/open-canada)
  - 21 Apr 2021: Analyzing PSES results using R and Shiny
  - 16 Apr-15 May 2021: Building R packages (recording, code)

Use R!

Owner: Dmitry Gorodnichy Group members: 184

+ Invite

Activity Discussion Files Blog More

Discussion topics

Add discussion topic

Best way to start (and keep) learning R

The number of resources and ways to learn R is enormous. Some of us had tried many of them until we found the ones that we believe are the best ones. Share them here! Here's how my personal recommendation - qouted...

4 Replies

Started by Dmitry Gorodnichy 2021-03-05 22:52

3 likes

Shiny for Interactive Data Visualization, Analysis and Web App development

This discussion thread is dedicated to Shiny package - a RStudio-curated tool for developing and deploying Interactive Data Visualization and Analysis tools and applications. Share your experiences, tricks, tools and questions...

6 Replies

Started by Dmitry Gorodnichy 2021-03-05 23:22

Geo/Spatial coding and visualization in R

There's much effort to across many GC departments to link and visualize geo-data. This discussion is the place to share your results, ideas or problems related to the problem. Below is a great resource to start, which also...

4 Replies

Discussion

Reading (all sorts of) data in R - efficiently!

My favourite methods for reading / writing "regular" .csv files has been 'data.table::fread() / fwrite()' - the fastest and automated in many ways. Now there's another one - with package 'vroom' -...

2 Replies

Started by Dmitry Gorodnichy 2021-04-22 12:40

2 likes

Excel -> R

There was a keen interest expressed at last Friday meetup on transitioning from Excel to R. Incidentally, there was an RStudio Community Meet-up focused exactly on this topic: Meetup: Making the Shift from Excel to R:...

2 Replies

Started by Dmitry Gorodnichy 2021-04-14 16:38

R communities in GC

Roughly sorted by the level of group activity GCConnex (GC...

Sub-Group

'Lunch and Learn' Data Science with R: Friday Meet-ups

Owner: Dmitry Gorodnichy Group members: 59

+ Invite

Activity Discussion Files Blog More

'Lunch and Learn' Data Science with R: Friday Meet-ups's files

New file folder Upload a file

Main folder

- VIDEO & NOTES: 23 April-15 May, 2021. Building R packages - Sessions 1-4
  By Dmitry Gorodnichy - 17 May 2021 @ 3:33pm - Download
- VIDEO & NOTES: Meetup 28 May 2021. - Lookup table function w. data.table, delivering packages, new...
  By Dmitry Gorodnichy - 1 June 2021 @ 4:49pm - Download
- VIDEO & NOTES: Meetup 4 June 2021 - Utility functions and Converting Shiny to Exe
  By Dmitry Gorodnichy - 4 June 2021 @ 8:31pm - Download
- VIDEO & NOTES: Meetup 2021-06-11.- parse_gcTable(), api.canada.ca, shiny in aws, best PSES...
  By Dmitry Gorodnichy - 11 June 2021 @ 8:19pm - Download
- VIDEO & NOTES: Meetup 2021-06-18. How to dynamically assign()
  By Dmitry Gorodnichy - 18 June 2021 @ 5:34pm - Download
- VIDEO & NOTES: Meetup 2021-07-25 . Using API for working with Open Government Data within R
  By Dmitry Gorodnichy - 25 June 2021 @ 5:28pm - Download
- VIDEO: Lunch and Learn (2021-07-09). Automating advanced common visualizations with ggplot()
  By Jonathan Dench - 9 July 2021 @ 6:30pm - Download

Folder structure

Main folder

Did you know?

You can drag and drop files on to the folders to organize them!

Next steps

- The work is in progress (and will always be!)
- Much more ahead. We need your help!
  - curating data problems and public domain solutions (codes/papers)
  - curating public domain datasets
  - testing & benchmarking
  - tutorials, use cases
- Join the community: Join GCcollab / GCcode groups
- Contacts:
  - Jonathan.Dench@tbs-sct.gc.ca
  - Dmitry.Gorodnichy@cbsa-asfc.gc.ca
  - Patrick.Little@tbs-sct.gc.ca
  - Joseph.Stinziano@inspection.gc.ca

Appendices: key outputs (so far)

- GCcode 101 for GC employees: https://gccode.ssc-spc.gc.ca/r4gc/resources/gccode101
- R packages 101 for GC employees: https://gccode.ssc-spc.gc.ca/r4gc/gc-packages/packages101
- How To: Interactive *rmarkdown / learnr* built tutorials to various problems
- Geospatial analysis and visualization: *markdown* built use cases
- Data Engineering: package and App for fuzzy matching, record linking & deduplication - https://rCanada.shinyapps.io/demo
- Interactive Shiny Apps: for ATIP, PSES, COVID-19, Border Wait Times: https://open-canada.github.io/Apps/atip (~/pses, ~/covid, ~/border)
- Working with Open Government Portal API within R (using *ckanr* and *adobeanalyticsr)*
- Automating R scripts to run with GitHub Actions

Slides below will <u>not</u> be presented,
and are for reference only.

R packages 101

- Key package
  - *devtools* has a series of key functions for setting up a package, especially directory and file structures
- Testing code
  - Writing tests is a key skill to ensuring robust, reproducible code
    - Goal is to ensure each step of a function works properly with a reproducible example
    - E.g. Is the output of function X a list?
  - testthat & testthis packages facilitate test writing
- Key considerations for GoC R packages
  - Licensing
  - What can be submitted to CRAN? What are the legal implications?

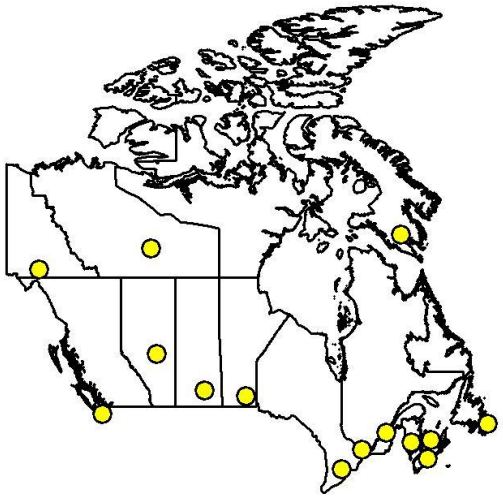
Geospatial Analysis in R

- Guidance & Tutorials

  - Applied Spatial Data Analysis with R (2008) Roger Bivand et al.

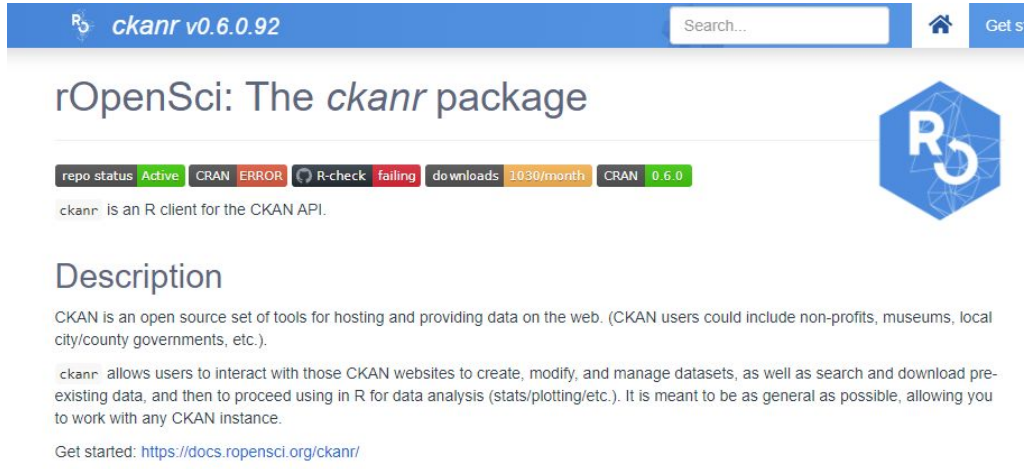
  - Geocomputation with R (2021) (https://geocompr.robinlovelace.net/)

  - Preparing series of workshops and guided code for the R4GC group.



Working with Open Government Portal API (1)

- CKAN is a very widely used software package for powering open data portal catalogues (data.gov, open.canada.ca, data.gov.uk, etc.)
- CKAN offers an API that can be used to retrieve datasets and metadata from the system, but also create, update, and manage datasets.
- Using the *ckanr* package offers a good developer experience for using the CKAN API within R.



Working with Open Government Portal API (2)
Using ckanr

What you can do with it

Function	API Command	CKANR function
Get information about the system	action/status_show	ckan_info()
List organizations that publish data	action/organization_list	organization_list()
Get a list of datasets on the portal	action/package_list	package_list()
Retrieve the metadata for a dataset	action/package_show/{id}	package_show()
Search for datasets	action/package_search?q={something-to-search-for}	package_search()
Create a new dataset	action/package_create	package_create()
Update an existing resource	action/resource_patch()	resource_patch()

Example Use Case *:* What datasets relating to COVID-19 are available on the portal?

- Web Browser: https://open.canada.ca/data/api/action/package_search?q="COVID“
- Batch: `curl --verbose https://open.canada.ca/data/api/action/package_search?q=“COVID”`
- ckanr:

```
library(ckanr)
ckanr_setup(url="https://open.canada.ca/data")
search_results<-package_search(q="COVID", as= "table")
View(search_results$results)
```

Web Analytics in R with Adobe Analytics

- The GC uses Adobe Analytics to measure usage on Canada.ca as well as several standalone web applications.
- The *adobeanalyticsr* package enables an analyst to pull in data from Adobe Analytics to create web analytics reports within R.
- This can be used to generate simple data extracts, but also to create Rmd reports, or power Shiny Applications.

adobeanalyticsr

R Client for Adobe Analytics API 2.0

Connect to the Adobe Analytics API v2.0, which powers Analysis Workspace. The package was developed with the analyst in mind and will continue to be developed with the guiding principles of iterative, repeatable, timely analysis. New features are actively being developed and we value your feedback and contribution to the process. Please submit bugs, questions, and enhancement requests as issues in this Github repository.

Adobeanalyticsr – basic usage

- Authenticate into Adobe Analytics using an OAuth token using function *aw_token()*
- Use the function *aw_freeform_table* to create a report based on parameters you supply
- Functions *aw_get_metrics, aw_get_dimensions, aw_get_segments* can be used to get available parameters.
- Analyze or Visualize your data within R

```
topPages<-aw_freeform_table(
  date_range = c("2021-04-01", "2021-04-28"),
  company_id = Sys.getenv("AW_COMPANY_ID"),
  rsid = Sys.getenv("AW_REPORTSUITE_ID"),
  dimensions = c("prop65","evar11"),
  metrics = c("pageviews","visits","event25"),
  search = "MATCH 'OG-GO'",
  top = c(20)
  )
```

```
## Estimated runtime: 16.8sec./0.28min.
```

```
## 1 of 21 possible data requests complete. Starting the next 1 requests.
```

```
## A total of 20 rows have been pulled.
```

```
names(topPages)<-c("App Name","Page Name","pageviews","visits","downloads")
kable(topPages)
```

App Name	Page Name	pageviews	visits	downloads
OG-GO	Open Government Portal	76222	24853	3
OG-GO	Open Government	19418	15608	0
OG-GO	Search Grants and Contributions	15383	3950	0
OG-GO	Search Government Contracts over $10,000	14397	3159	0
OG-GO	Completed Access to Information Requests	12511	2898	0
OG-GO	blank page title	12187	5943	0
OG-GO	Canada Base Map Transportation (CBMT) - Open Government Portal	11204	9643	968

Automating R scripts to run in GitHub Actions

- GitHub Actions is a free workflow driven platform designed for automating software development tasks such as CI/CD.
- GitHub actions uses docker containers that can be configured to run a myriad of different operating systems and software packages, including R.
- This allows a user to run an R script based on a cron schedule, or other events such as a change to the script.
- GitHub actions is very useful for automating reports or other R workloads.

How to run an R script in GitHub Actions

Actions | Projects | Wiki | Security | Insights | Settings

master | OpenGov_R_Scripts / .github / workflows / main.yml | View runs | Go to file | ···

The workflow is defined in a YAML file

PatLittle Update main.yml ✕ | Latest commit bfa7251 on Feb 3 | History

1 contributor

25 lines (22 sloc) | 590 Bytes | Raw | Blame

```yaml
on:
  push:
  schedule:
    # * is a special character in YAML so you have to quote this string
    - cron:  '0 11 * * *'
jobs:
  build:
    runs-on: ubuntu-latest
    steps:
      - uses: actions/checkout@v2
      - run: |
          echo $PWD

      - uses: r-lib/actions/setup-r@master
        with:
          r-version: '4.0.2'
        run: |
          Rscript PD-Count.r
          git config user.name github-actions
          git config user.email github-actions@github.com
          git add .
          git commit -m "generated"
          git push
```

On push – runs every time you change the code in the repository

On schedule – runs on cron times

r-lib setup-r will setup your container with Ubuntu and install the version of R you specify

Rscript is the R script you want to run

git . add
git commit
git push

git .add, commit, push will save the output of any files you created/modified into the GitHub repository

Data Engineering
Records cleaning, deduplication and linking

https://rCanada.shinyapps.io/demo

Leverages the work of CBSA, various R packages for data cleaning and linking, and RStudio's Shiny framework

Included use cases:

- Web crawling: …/demo/#section-web-crawling
  - Dates extraction
  - Finding nicknames and names variants

Record linking challenges

- Dates : ‘20210820’ vs. ‘dob 20 Aug 2021’
- Names: ‘Dmitry Gorodnichy’ vs. ‘Dimitri Horodnytchyyi’
- Business Names: AC, AirCanada, Air Canada Corp.
- Geographic Names: Ottawa, Orleans, Orléans

- General Text : “<tag> ca$h 4 u ! Sooo… C O O L! Cant believe it ☹️ ”
- Postal: “klo 0O1” vs “K100o1”
- Text matching: Phrase matching, topics/keywords detection

Data Engineering Testbed | Intro | Single-variable tasks | Multiple-variable tasks | Use cases | Test it! | Info | rCanada

`text2date()` : converts text to a date using various decision logics.

Test it:

Enter dates, any way you want, and observe how they get automatically converted to `YY MM DD` format.

7jul 35

Reset table

Result:

7 jul 35 --> 2035-07-07

text	YY	MM	DD
7jul35	2035	7	7
1935.08..7	1935	8	7
DOB 12/26/2010...	2010	12	26
26/12/1930	1930	12	26
7.VI.35	2035	6	7
7 jul35	2035	7	7
7 jul 35	2035	7	7

`text2timestamp()` : extracts automatically timestamp from free-form text

Test it:

Enter a timestamp any way you want and observe how it gets converted to the same canonical timestamp `YY-MM-DD hh:mm:ss` format.

2021-03-17 19:14:08

Result:

2021-03-17 19:14:08 --> 2021-03-17 19:14:08

text	TIMESTAMP
2010-04-14 22:00	2020-10-04 14:22:00
2010-04-14 10pm	2020-10-04 14:10:00
2010-04-14-04-35-59	2010-04-14 04:35:59
2010-04-01-12-00-00	2010-04-01 12:00:00
20/2/06 11:16:16.683	2020-02-06 11:16:16
20100101120101	2010-01-01 12:01:01
2009-01-02 12-01-02	2009-01-02 12:01:02
2009.01.03 12:01:03	2009-01-03 12:01:03
2009-1-4 12-1-4	2009-01-04 12:01:04
2009-1, 5 12:1, 5	2009-01-05 12:01:05
200901-08 1201-08	2009-01-08 12:01:08
20090107 120107	2009-01-07 12:01:07
10-01-10 10:01:10 and p format: AM	2010-01-10 10:01:10
Created on 10-01-11 at 10:01:11 PM	2010-01-11 22:01:11

Cleaning Dates

Data Engineering Testbed | Intro | Single-variable tasks | Multiple-variable tasks | Use cases | Test it! | Info | rCanada

Dates
Postal
Names

`searchName(name)` : find similar names

Find similar names, using a variety of string similarity me for definitions of all metrics.

Type a name: Dmitry

String similarity metric: jaccard

Metric threshold

0 0.45 1

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

Result | Search and Save:

	Name	osa	lv	hamming	lcs	qgram	cosine	jaccard	jw	soundex
	<char>	<num>	<num>	<num>	<num>	<num>	<num>	<num>	<num>	<num>
1:	myrtie	5	5	5	8	2	0.167	0.286	0.306	1
2:	myrtis	5	5	5	8	2	0.167	0.286	0.306	1
3:	timmy	4	4	Inf	7	3	0.228	0.333	0.411	1
4:	demetria	4	4	Inf	6	4	0.355	0.375	0.278	0
5:	demetrice	5	5	Inf	7	5	0.473	0.375	0.296	0
6:	meredith	7	7	Inf	8	4	0.355	0.375	0.403	1
7:	merideth	7	7	Inf	8	4	0.355	0.375	0.403	1
8:	meridith	7	7	Inf	8	4	0.225	0.375	0.403	1
9:	myrtice	6	6	Inf	9	3	0.228	0.375	0.337	1
10:	armida	5	5	6	8	4	0.423	0.429	0.444	1
11:	marita	5	5	6	6	4	0.423	0.429	0.347	1
12:	marti	4	5	Inf	7	3	0.270	0.429	0.261	1
13:	marty	3	4	Inf	5	3	0.270	0.429	0.261	1
14:	mertie	5	5	5	8	4	0.423	0.429	0.306	1
15:	mindy	3	3	Inf	5	3	0.270	0.429	0.300	1
16:	mirta	3	4	Inf	5	3	0.270	0.429	0.261	1
17:	misty	3	3	Inf	3	3	0.270	0.429	0.178	1
18:	myriam	6	6	6	8	4	0.278	0.429	0.444	1
19:	myrta	4	5	Inf	7	3	0.270	0.429	0.411	1
20:	trinity	5	5	Inf	7	5	0.261	0.429	0.357	1
21:	trudi	6	6	Inf	7	3	0.270	0.429	0.544	1
22:	trudy	5	5	Inf	5	3	0.270	0.429	0.544	1
23:	yadira	5	5	5	6	4	0.423	0.429	0.333	1
24:	demetrius	5	5	Inf	7	5	0.385	0.444	0.296	0
	Name	osa	lv	hamming	lcs	qgram	cosine	jaccard	jw	soundex

Approximate (fuzzy/probabilistic) name matching

Record deduplication

Data Engineering Testbed | Intro | Single-variable tasks | Multiple-variable tasks | Use cases | Test it! | Info | rCanada

Upload your CSV file or choose a preloaded one from the menu below. Then select a task to perform, choose parameters, and press "Start!".

* Use Uploaded File *

Browse... names_example

Upload complete

Columns to process (Click 'Del' to remove):

lastname firstname address sex postcode

Choose the task to perform:

○ View ○ Search
◉ Deduplicate ○ Link

Choose Columns to block:

☐ Measure processing time

Start!

Tuning parameters

String similarity metric:

jaccard

Decision Threshold

0 2.8 5

0 0.5 1 1.5 2 2.5 3 3.5 4 4.5 5

Max string distance (%)

0 0.15 0.5

0 0.05 0.1 0.15 0.2 0.25 0.3 0.35 0.4 0.45 0.5

☐ Use speed-optimized matching (experimental)

Input | Search and Save | Summarize

○ Show first and last row ○ ... top / bottom three rows ◉ ... entire table

◉ Show first file ○ ... second file

	lastname	firstname	address	sex	postcode
1	Smith	Anna	12 Mainstr	F	1234 AB
2	Smith	George	12 Mainstr	M	1234 AB
3	Johnson	Charles	61 Mainstr	M	1234 AB
4	Johnson	Charly	61 Mainstr	M	1234 AB
5	Schwartz	Ben	1 Eaststr	M	6789 XY

Output | Search and Save | Interim results | Log

lastname	firstname	address	sex	postcode	duplicate_groups
Smith	Anna	12 Mainstr	F	1234 AB	1
Smith	George	12 Mainstr	M	1234 AB	1
Johnson	Charles	61 Mainstr	M	1234 AB	3
Johnson	Charly	61 Mainstr	M	1234 AB	3
Schwartz	Ben	1 Eaststr	M	6789 XY	5

Record linking

Data Engineering Testbed | Intro | Single-variable tasks | Multiple-variable tasks | Use cases | Test it! | Info | rCanada

Upload your CSV file or choose a preloaded one from the menu below. Then select a task to perform, choose parameters, and press "Start!".

* Use Uploaded File *

Browse... names_example

Upload complete

Columns to process (Click 'Del' to remove):

lastname firstname address sex postcode

Choose the task to perform:

○ View ○ Search ○ Deduplicate ◉ Link

Choose Columns to block:

postcode

Upload second CSV file, or choose a preloaded one from menu below:

* Use Uploaded File *

Browse... names_example

Upload complete

☐ Measure processing time

Start!

Tuning parameters

String similarity metric:

jaccard

Decision Threshold

0 3.15 5

Input | Search and Save | Summarize

○ Show first and last row ○ ... top / bottom three rows ◉ ... entire table

◉ Show first file ○ ... second file

	lastname	firstname	address	sex	postcode
1	Smith	Anna	12 Mainstr	F	1234 AB
2	Smith	George	12 Mainstr	M	1234 AB
3	Johnson	Charles	61 Mainstr	M	1234 AB
4	Johnson	Charly	61 Mainstr	M	1234 AB
5	Schwartz	Ben	1 Eaststr	M	6789 XY

Output | Search and Save | Interim results | Log

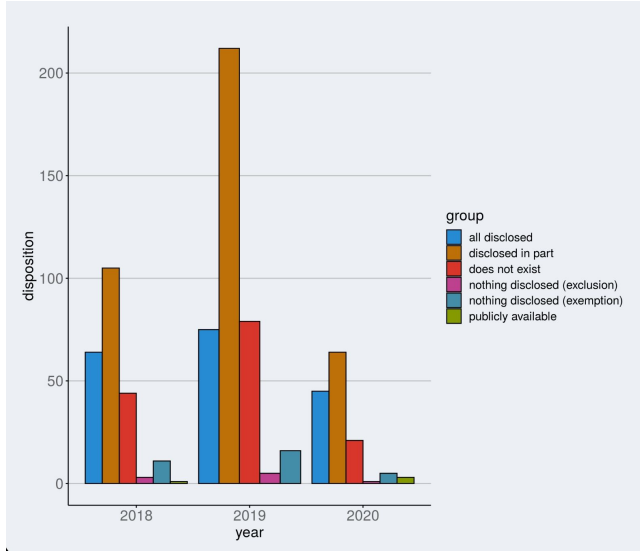
lastname.x	firstname.x	address.x	sex.x	postcode.x	lastname.y	firstname.y	address.y	sex.y	postcode.y
Smith	George	12 Mainstr	M	1234 AB	Smith	Gearge	12 Mainstreet		1234 AB
Johnson	Charles	61 Mainstr	M	1234 AB	Johnson	Charles	61 Mainstr	F	1234 AB
Johnson	Charly	61 Mainstr	M	1234 AB	Johnson	Charles	61 Mainstr	F	1234 AB
Smith	Anna	12 Mainstr	F	1234 AB	NA	NA	NA	NA	NA
Schwartz	Ben	1 Eaststr	M	6789 XY	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	Schwartz	Ben	1 Main	M	6789 XY
NA	NA	NA	NA	NA	Schwartz	Anna	1 Eaststr	F	6789 XY

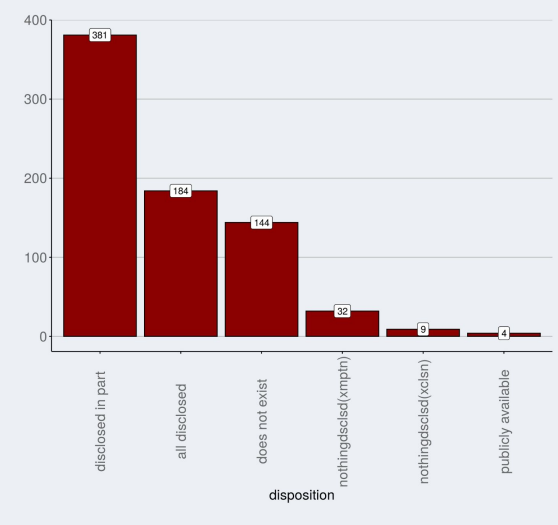
NLP topic modeling in TBS ATIP data

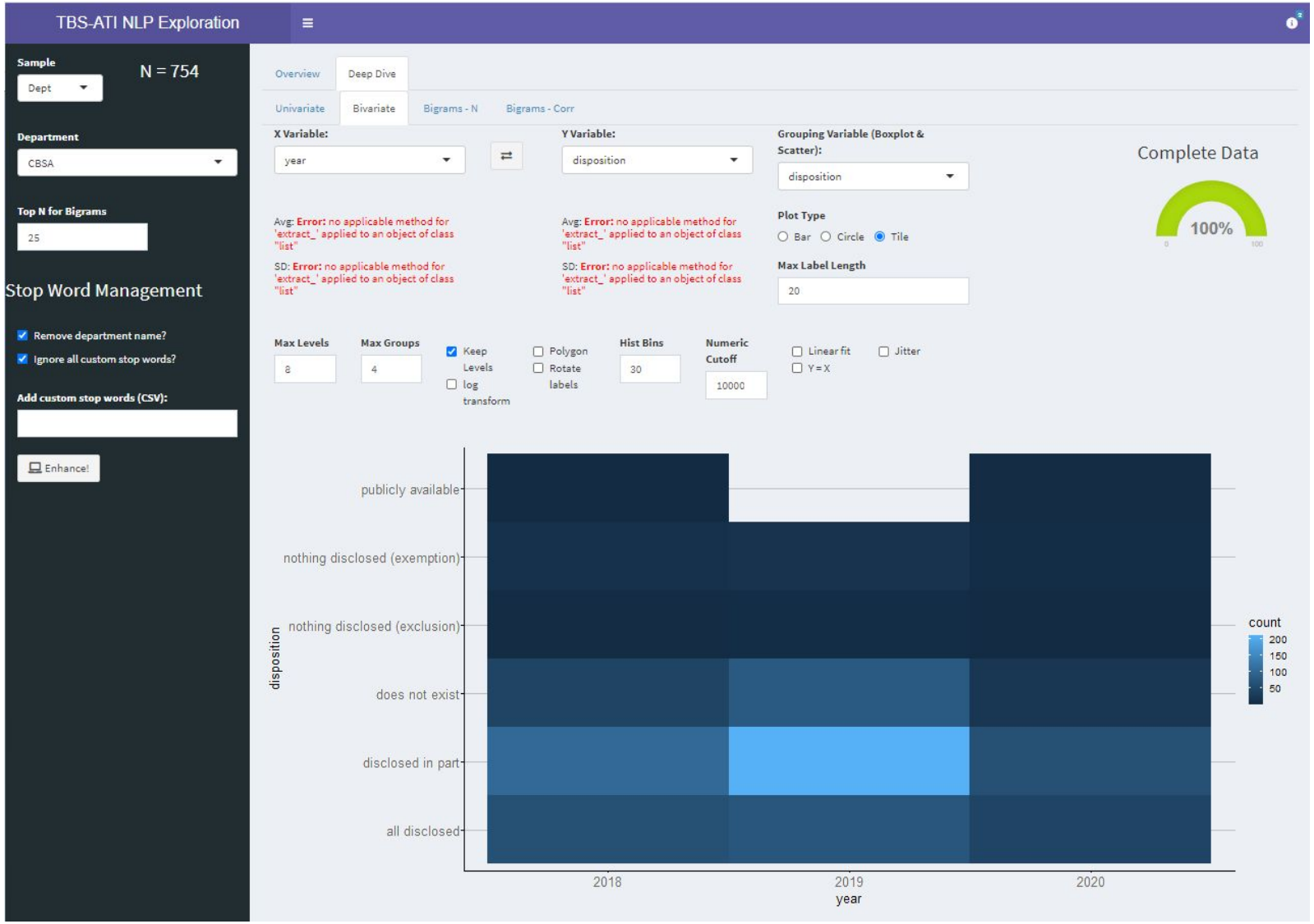
https://open-canada.github.io/Apps/atip

Leverages the work of TBS, various R packages for text mining, and RStudio's Shiny framework

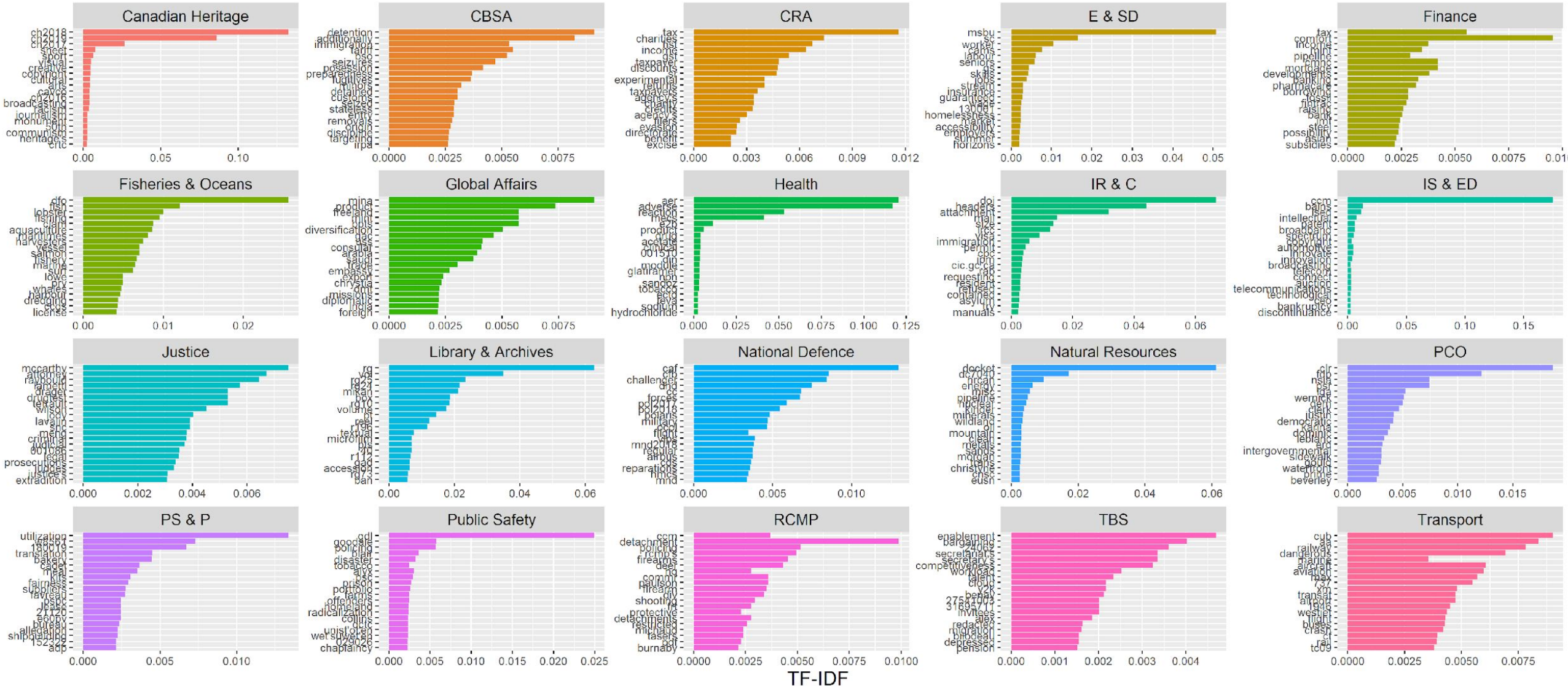
Univariate and bivariate analysis of dataset variables



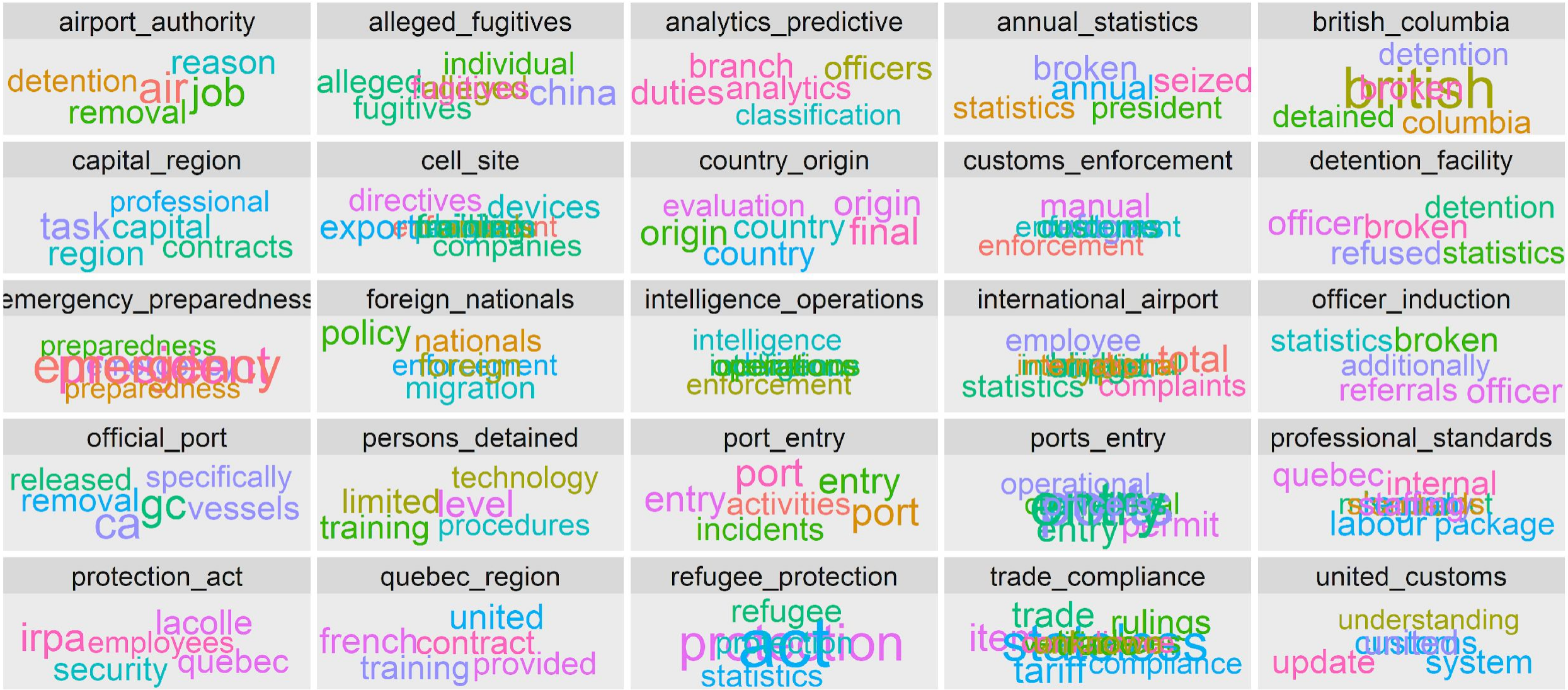




1-grams (single words)



Topic modeling (30 main topics): wordcloud



Topic modelling:

Graph / Network view

TBS-ATI NLP Exploration

Sample

Dept

N = 754

Department

CBSA

Top N for Bigrams

15

Stop Word Management

- [x] Remove department name?
- [x] Ignore all custom stop words?

Add custom stop words (CSV):

Enhance!

Overview | Deep Dive

Univariate | Bivariate | Bigrams - N | Bigrams - Corr

Min Corr

0.3

Show 18 entries Search:

	item1	item2	correlation
1	preparedness	emergency	0.971129452586342
2	emergency	preparedness	0.971129452586342
3	fugitives	alleged	0.936519829277792
4	alleged	fugitives	0.936519829277792
5	fugitives	control	0.909927934323004
6	control	fugitives	0.909927934323004
7	arrangements	supply	0.885947712418301
8	standing	supply	0.885947712418301
9	supply	arrangements	0.885947712418301
10	supply	standing	0.885947712418301
11	alleged	control	0.851958886147721
12	control	alleged	0.851958886147721
13	china	control	0.844085451931497
14	control	china	0.844085451931497
15	fugitives	individual	0.832662742876612
16	individual	fugitives	0.832662742876612
17	china	fugitives	0.831695559343427
18	fugitives	china	0.831695559343427

Showing 1 to 18 of 308 entries

Previous 1 2 3 4 5 ... 18 Next