When Data is Scarce…
Ways to Extract Valuable Insights

Bety E. Rodriguez-Milla

Waterloo Data Science and Data Engineering MeetUp May 9, 2019

The Dataset

Freedom of Information Requests of the Region of Waterloo via their Open Data project.

The Goal

Use Machine Learning (ML) to predict if a request will be approved or not.

The Outcome

ML does poorly due to the dataset size - see posts of Scott Jones on Medium @scottcurtisjones.

Solution: *Find more data!*

yet…

Data is gold *and* there are other ways extract its value,

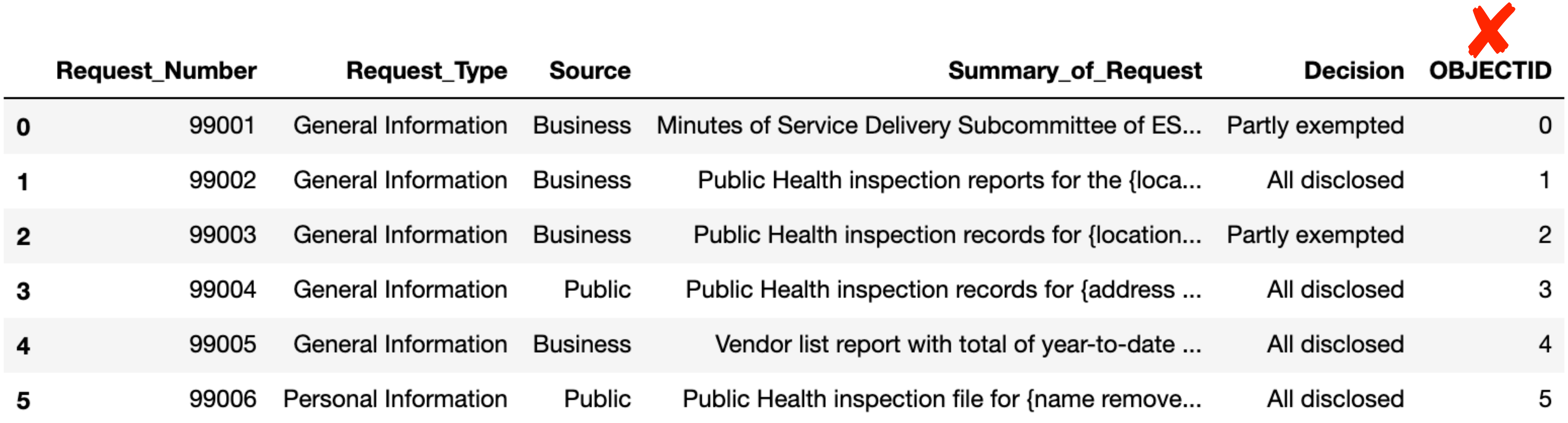
- Descriptive statistics
  - * Summarizes a sample, rather than the population.
  - * Univariate and multivariate analysis
- Exploratory Data Analysis (EDA)
  - * Explore the data, usually by visual methods, and possibly formulate hypotheses that could lead to new data collection and experiments.
- Natural Language Processing (NLP) techniques
  - * Process and analyze large amounts of natural language data.
  - * NLTK, spaCy, and scikit-learn

The Data

Freedom of Information Requests of the Region of Waterloo:

- 18 files - 1999 to 2016
- 576 requests in total
- All same six columns (amazingly!)

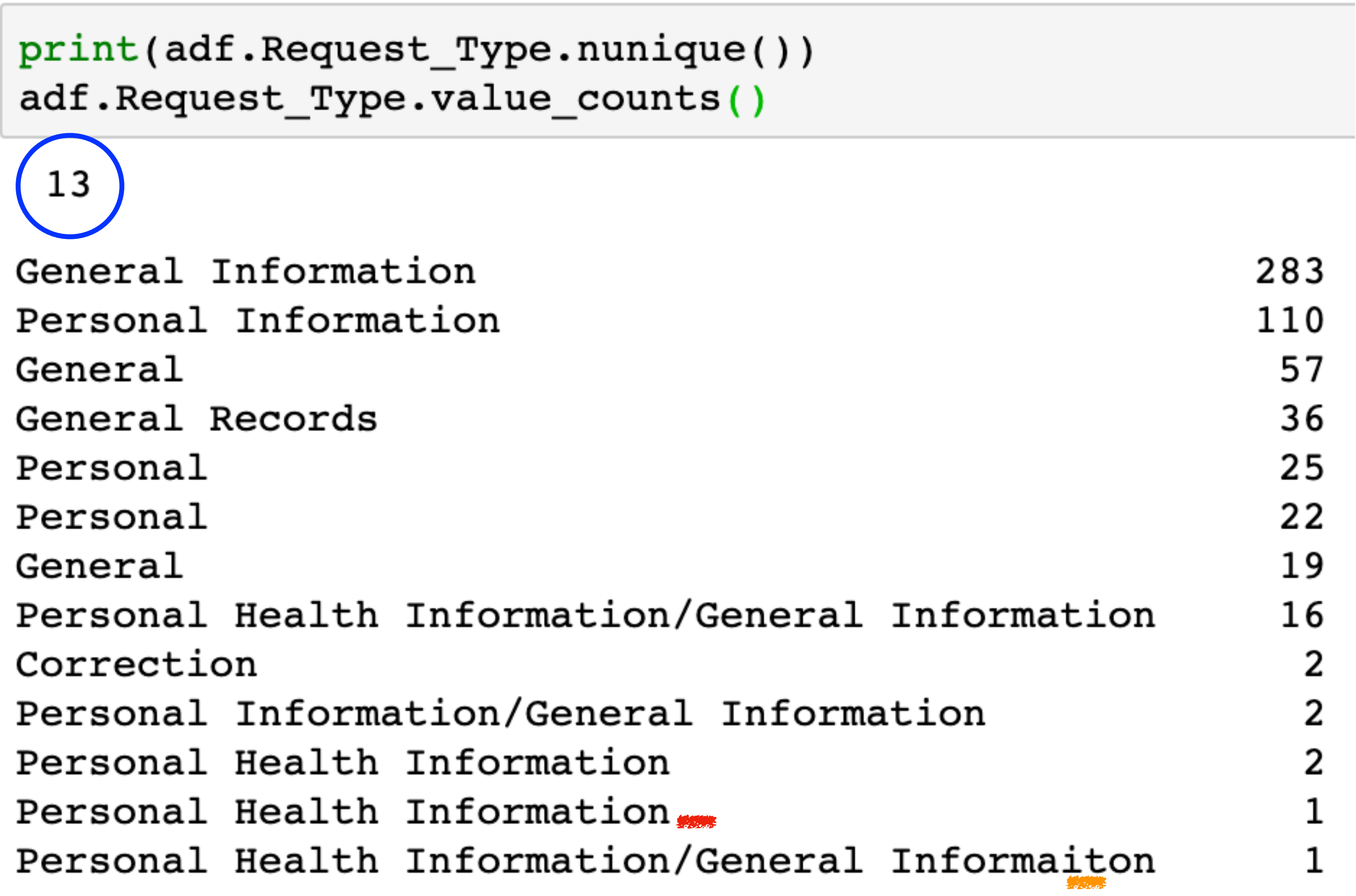
And it looks like this:



	Request_Number	Request_Type	Source	Summary_of_Request	Decision	OBJECTID
0	99001	General Information	Business	Minutes of Service Delivery Subcommittee of ES...	Partly exempted	0
1	99002	General Information	Business	Public Health inspection reports for the {loca...	All disclosed	1
2	99003	General Information	Business	Public Health inspection records for {location...	Partly exempted	2
3	99004	General Information	Public	Public Health inspection records for {address ...	All disclosed	3
4	99005	General Information	Business	Vendor list report with total of year-to-date ...	All disclosed	4
5	99006	Personal Information	Public	Public Health inspection file for {name remove...	All disclosed	5

The Columns & The Cleaning

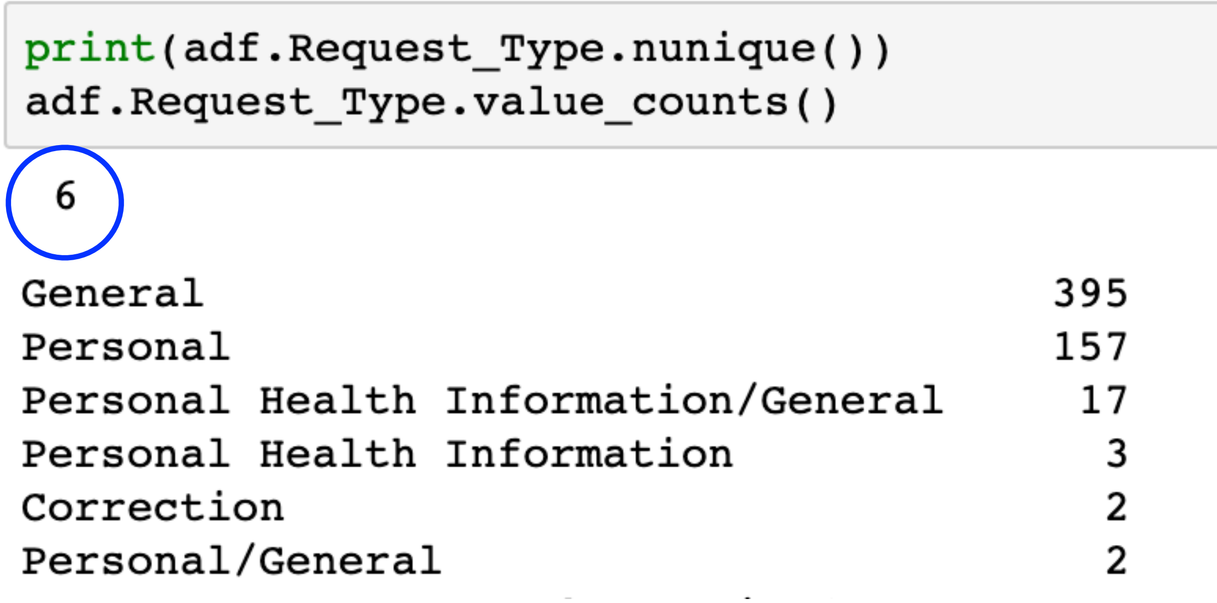
Before



```
print(adf.Request_Type.nunique())
adf.Request_Type.value_counts()
```

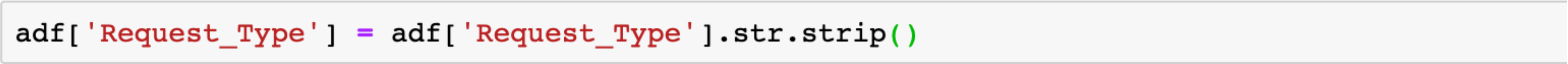
```
13

General Information                                   283
Personal Information                                  110
General                                                57
General Records                                        36
Personal                                               25
Personal                                               22
General                                                19
Personal Health Information/General Information        16
Correction                                              2
Personal Information/General Information                2
Personal Health Information                             2
Personal Health Information                             1
Personal Health Information/General Informaiton         1
```

After

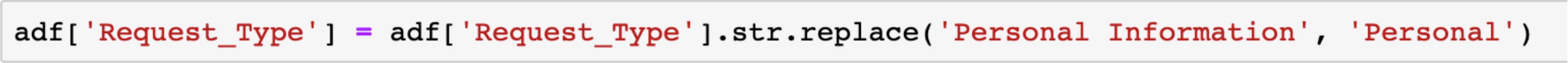


```
print(adf.Request_Type.nunique())
adf.Request_Type.value_counts()
```

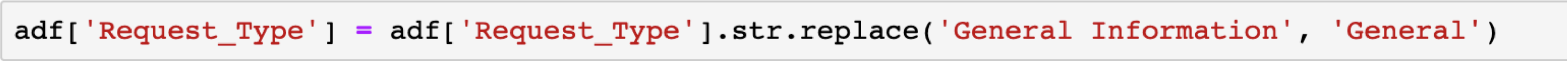
```
6

General                                        395
Personal                                       157
Personal Health Information/General             17
Personal Health Information                      3
Correction                                       2
Personal/General                                 2
```



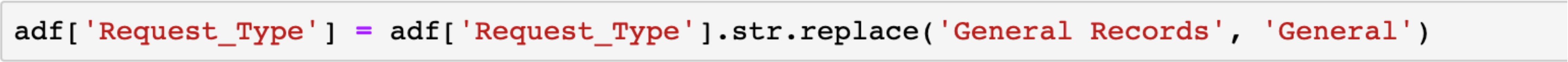
```
adf['Request_Type'] = adf['Request_Type'].str.strip()
```



```
adf['Request_Type'] = adf['Request_Type'].str.replace('Personal Information', 'Personal')
```



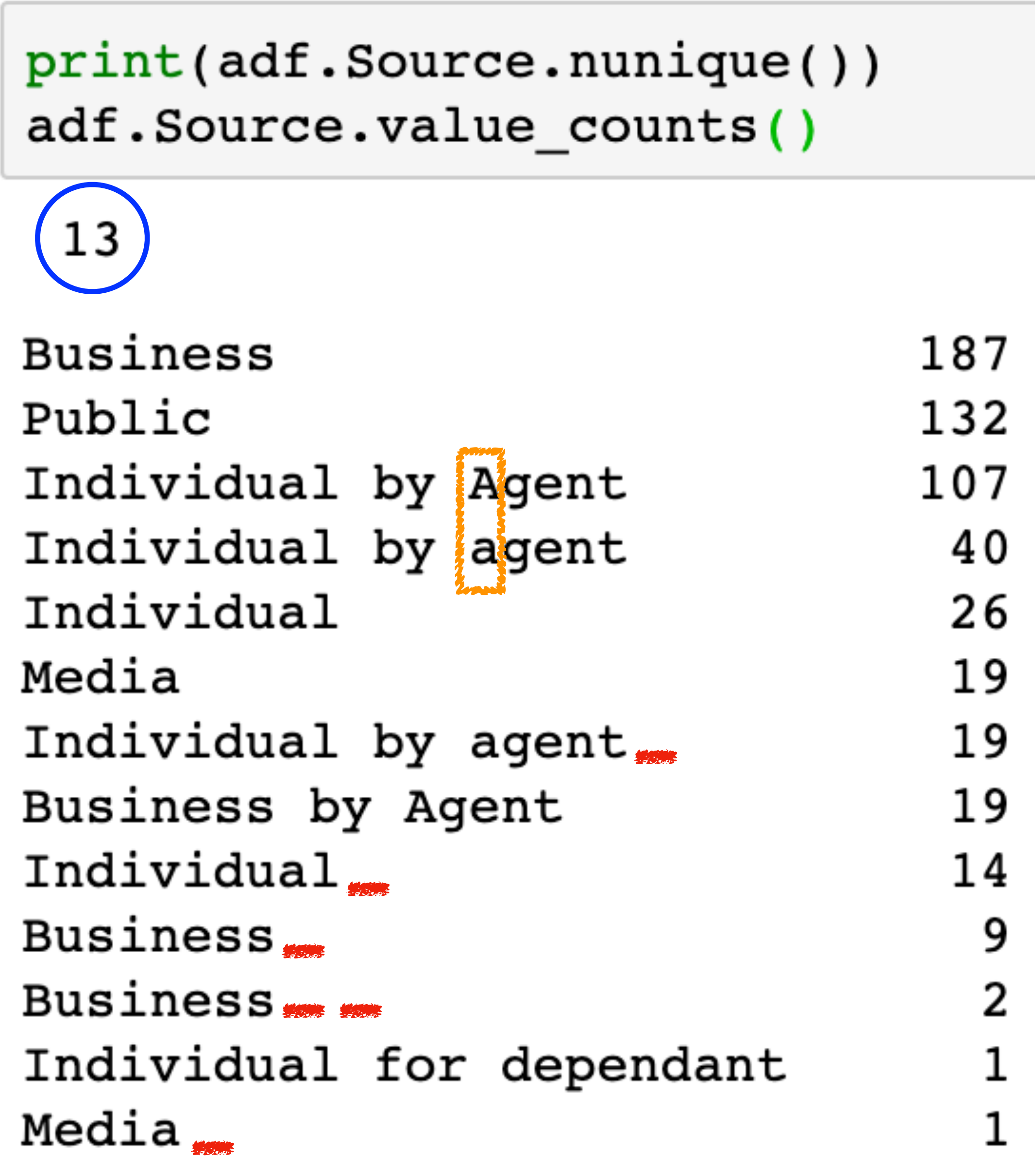
```
adf['Request_Type'] = adf['Request_Type'].str.replace('General Information', 'General')
```



```
adf['Request_Type'] = adf['Request_Type'].str.replace('General Records', 'General')
```

… more cleaning

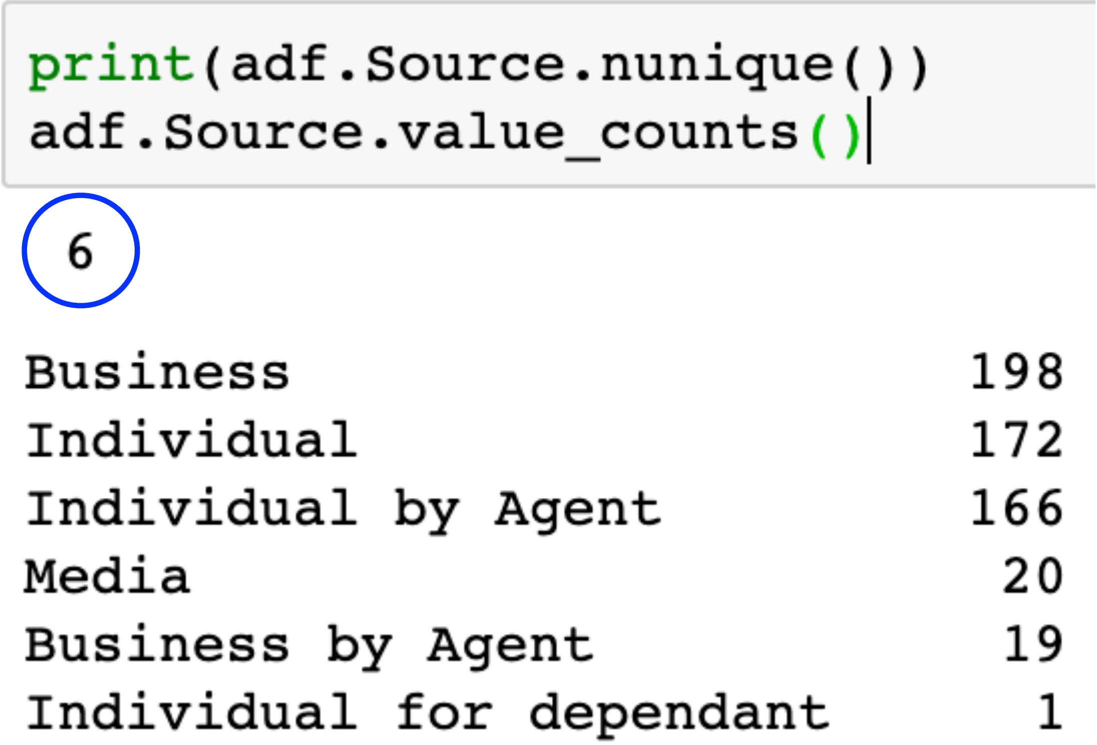
Before



```python
print(adf.Source.nunique())
adf.Source.value_counts()
```

```
13

Business                    187
Public                      132
Individual by Agent         107
Individual by agent          40
Individual                   26
Media                        19
Individual by agent          19
Business by Agent            19
Individual                   14
Business                      9
Business                      2
Individual for dependant      1
Media                         1
```

After

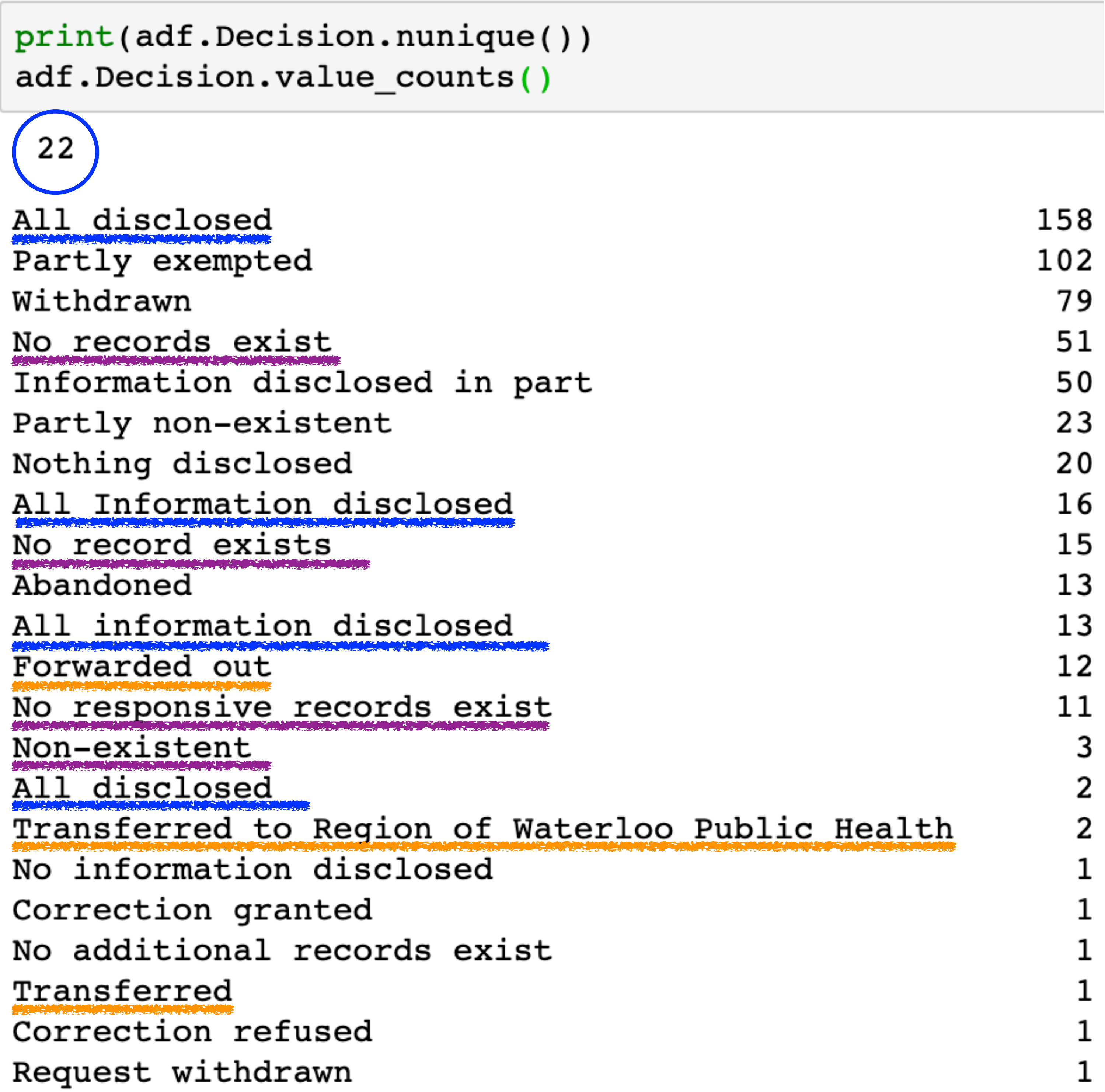


```python
print(adf.Source.nunique())
adf.Source.value_counts()
```

```
6

Business                    198
Individual                  172
Individual by Agent         166
Media                        20
Business by Agent            19
Individual for dependant      1
```

… even more cleaning

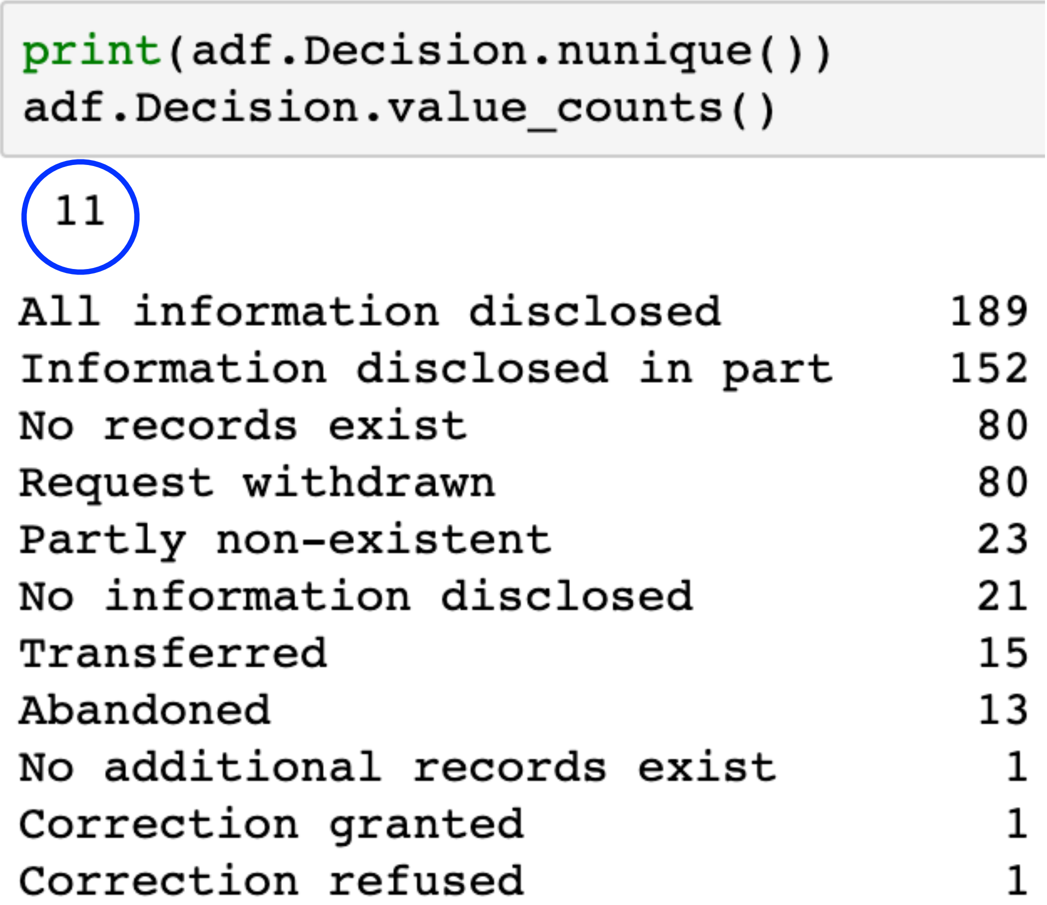
Before



```
print(adf.Decision.nunique())
adf.Decision.value_counts()
```

```
22
All disclosed                                   158
Partly exempted                                 102
Withdrawn                                        79
No records exist                                 51
Information disclosed in part                    50
Partly non-existent                              23
Nothing disclosed                                20
All Information disclosed                        16
No record exists                                 15
Abandoned                                        13
All information disclosed                        13
Forwarded out                                    12
No responsive records exist                      11
Non-existent                                      3
All disclosed                                     2
Transferred to Region of Waterloo Public Health   2
No information disclosed                          1
Correction granted                                1
No additional records exist                       1
Transferred                                       1
Correction refused                                1
Request withdrawn                                 1
```

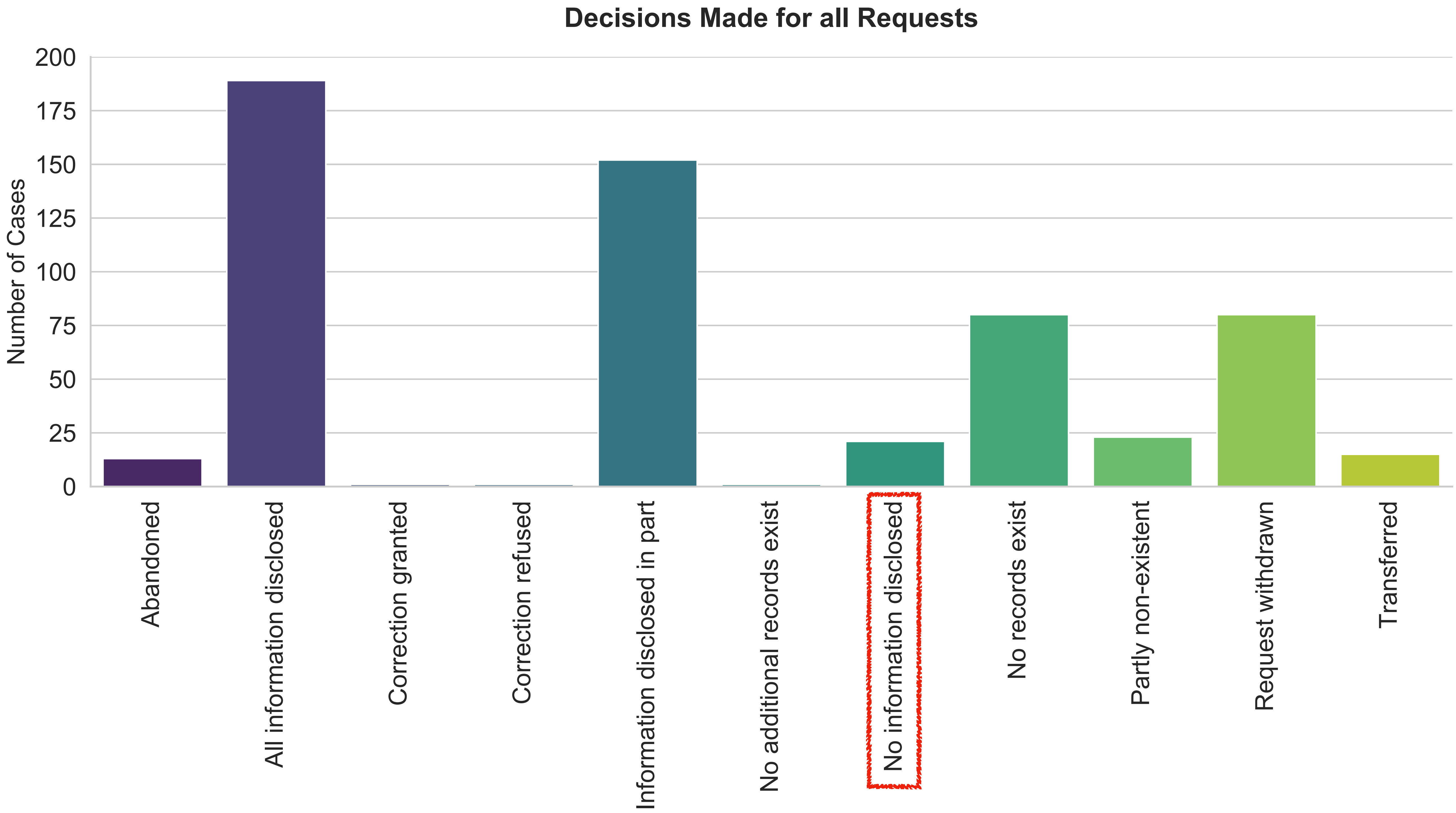
After



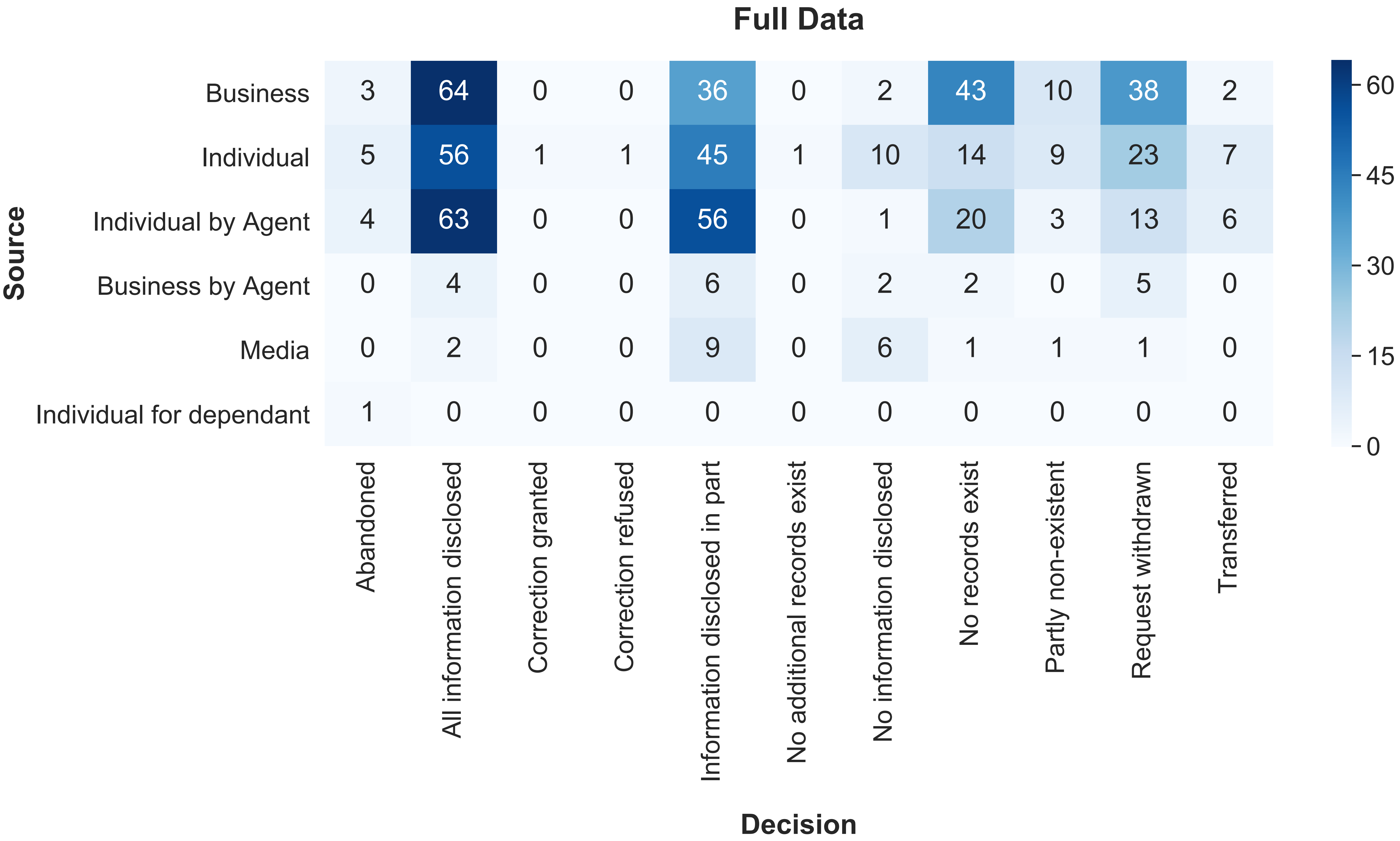
```
print(adf.Decision.nunique())
adf.Decision.value_counts()
```

```
11
All information disclosed         189
Information disclosed in part     152
No records exist                   80
Request withdrawn                  80
Partly non-existent                23
No information disclosed           21
Transferred                        15
Abandoned                          13
No additional records exist         1
Correction granted                  1
Correction refused                  1
```

Descriptive Statistics: Univariate Analysis



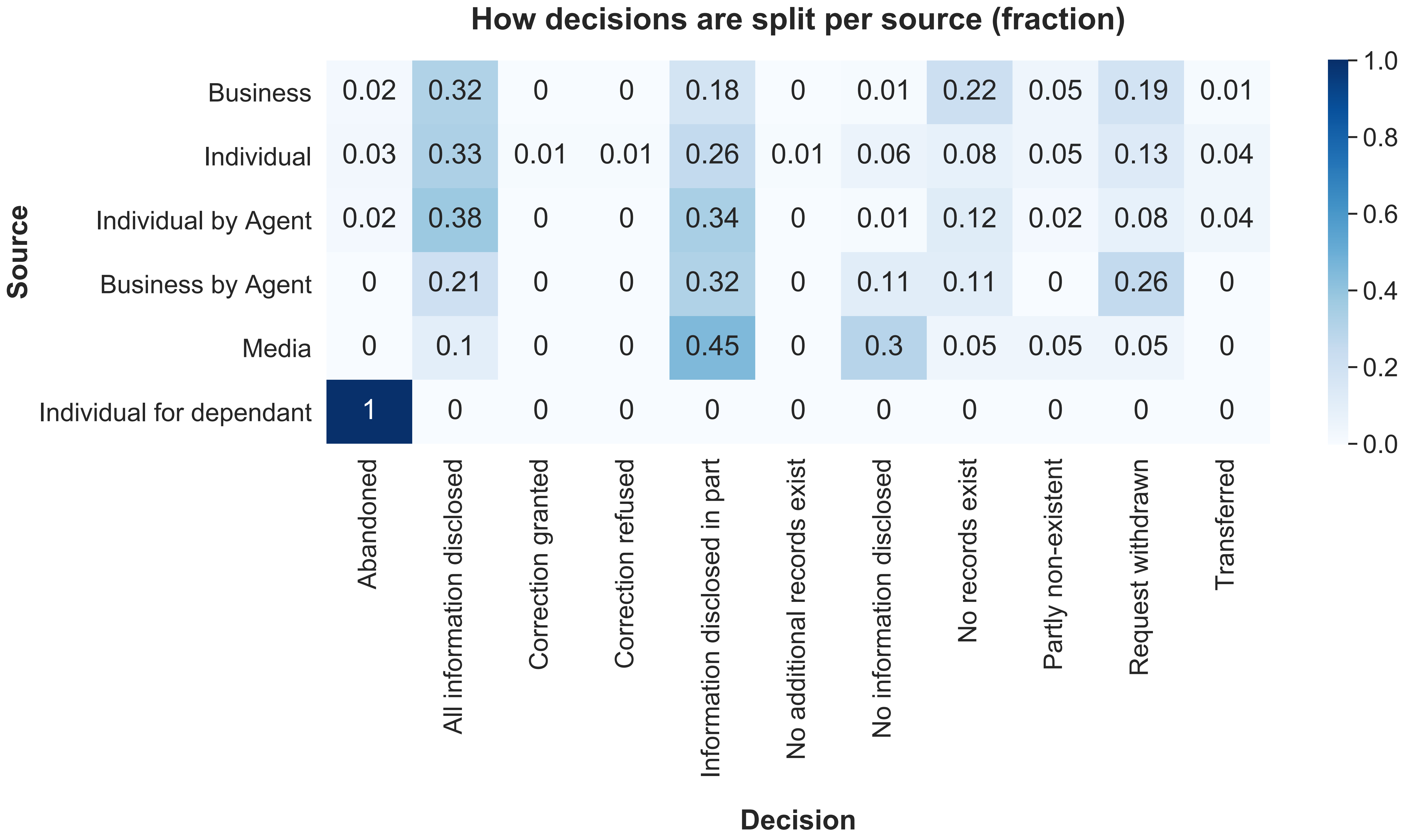
Bivariate Analysis



Full Data

	Abandoned	All information disclosed	Correction granted	Correction refused	Information disclosed in part	No additional records exist	No information disclosed	No records exist	Partly non-existent	Request withdrawn	Transferred
Business	3	64	0	0	36	0	2	43	10	38	2
Individual	5	56	1	1	45	1	10	14	9	23	7
Individual by Agent	4	63	0	0	56	0	1	20	3	13	6
Business by Agent	0	4	0	0	6	0	2	2	0	5	0
Media	0	2	0	0	9	0	6	1	1	1	0
Individual for dependant	1	0	0	0	0	0	0	0	0	0	0

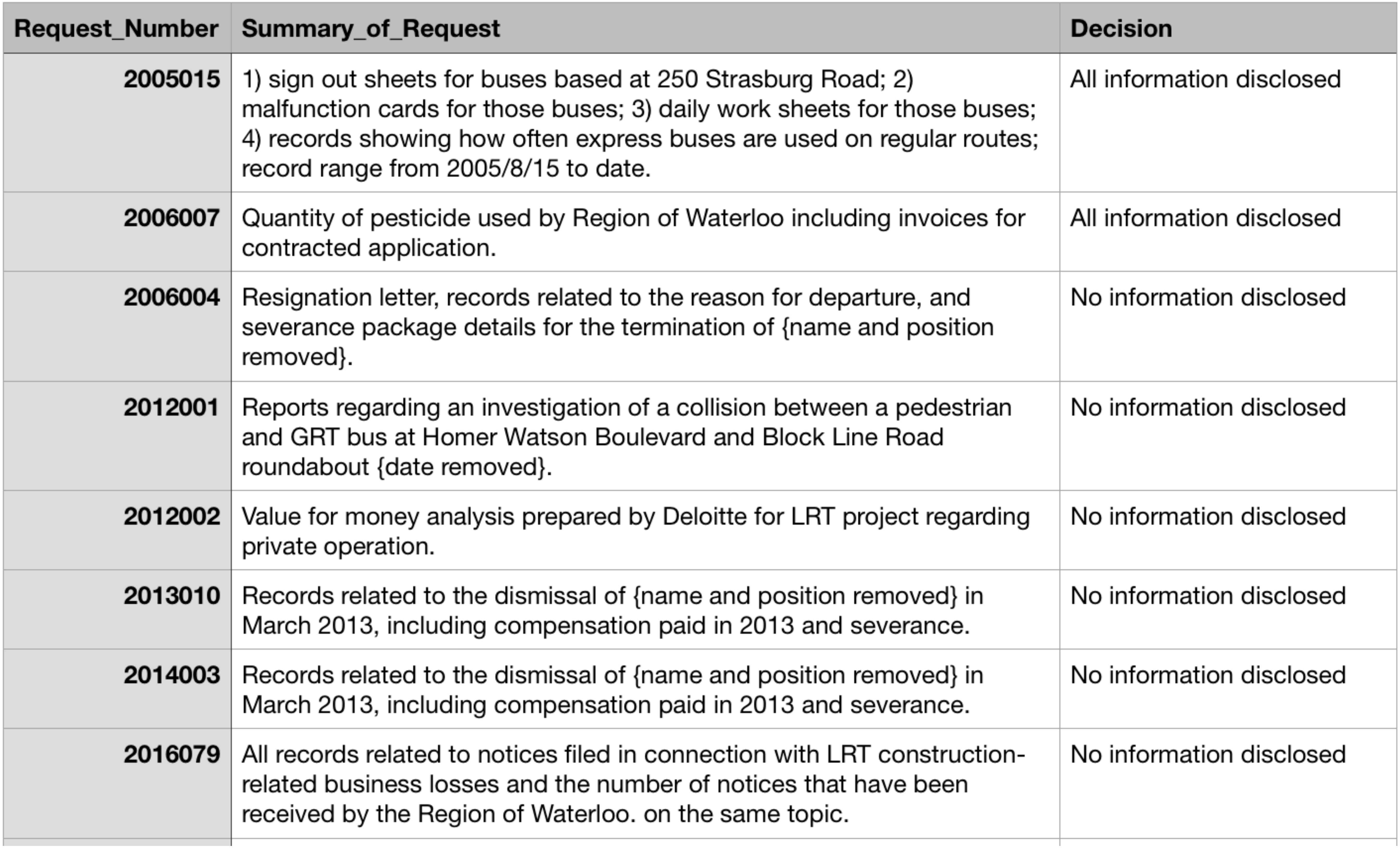
Bivariate Analysis



How decisions are split per source (fraction)

Source	Abandoned	All information disclosed	Correction granted	Correction refused	Information disclosed in part	No additional records exist	No information disclosed	No records exist	Partly non-existent	Request withdrawn	Transferred
Business	0.02	0.32	0	0	0.18	0	0.01	0.22	0.05	0.19	0.01
Individual	0.03	0.33	0.01	0.01	0.26	0.01	0.06	0.08	0.05	0.13	0.04
Individual by Agent	0.02	0.38	0	0	0.34	0	0.01	0.12	0.02	0.08	0.04
Business by Agent	0	0.21	0	0	0.32	0	0.11	0.11	0	0.26	0
Media	0	0.1	0	0	0.45	0	0.3	0.05	0.05	0.05	0
Individual for dependant	1	0	0	0	0	0	0	0	0	0	0

And of those media requests:



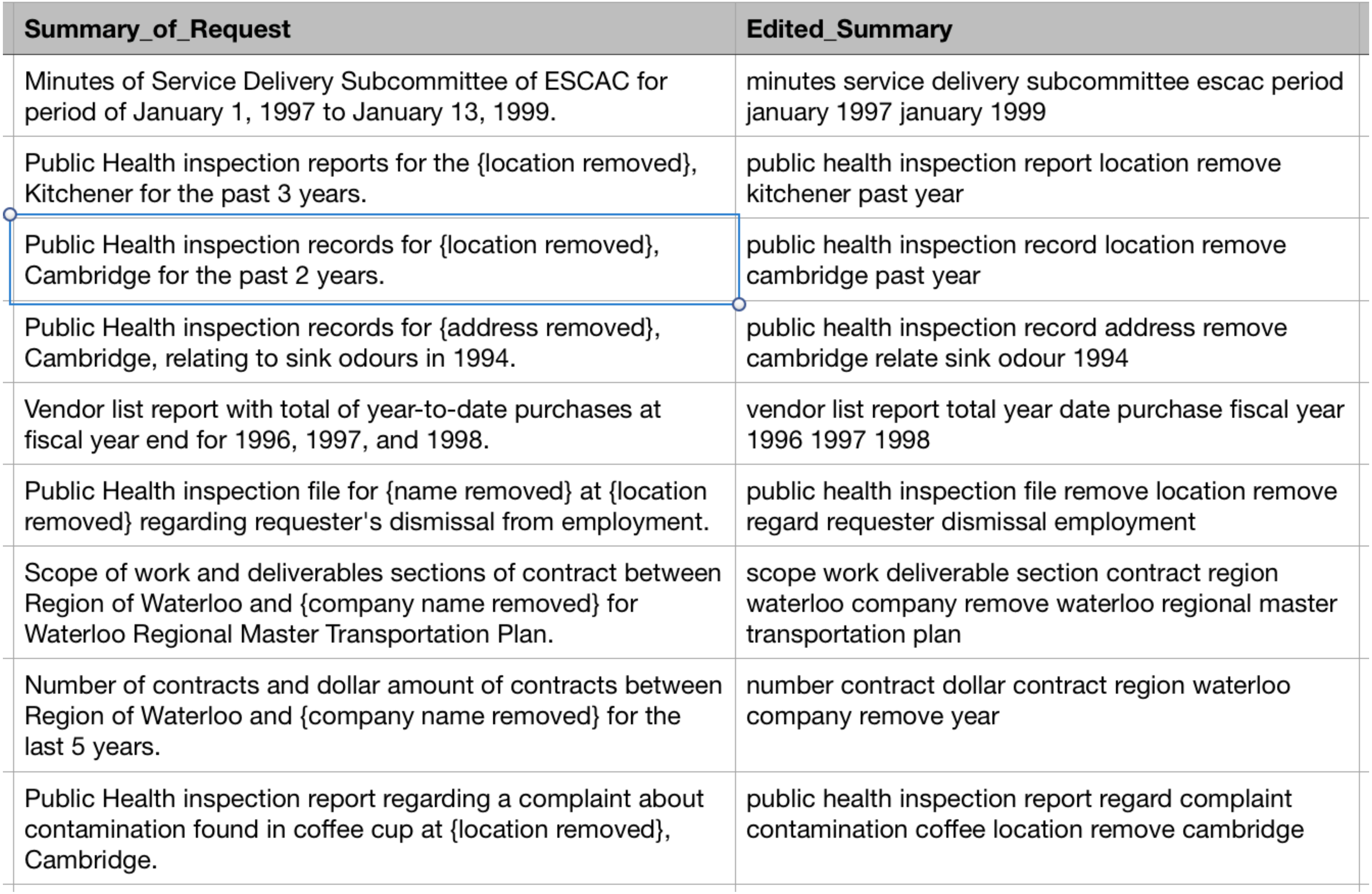
Request_Number	Summary_of_Request	Decision
2005015	1) sign out sheets for buses based at 250 Strasburg Road; 2) malfunction cards for those buses; 3) daily work sheets for those buses; 4) records showing how often express buses are used on regular routes; record range from 2005/8/15 to date.	All information disclosed
2006007	Quantity of pesticide used by Region of Waterloo including invoices for contracted application.	All information disclosed
2006004	Resignation letter, records related to the reason for departure, and severance package details for the termination of {name and position removed}.	No information disclosed
2012001	Reports regarding an investigation of a collision between a pedestrian and GRT bus at Homer Watson Boulevard and Block Line Road roundabout {date removed}.	No information disclosed
2012002	Value for money analysis prepared by Deloitte for LRT project regarding private operation.	No information disclosed
2013010	Records related to the dismissal of {name and position removed} in March 2013, including compensation paid in 2013 and severance.	No information disclosed
2014003	Records related to the dismissal of {name and position removed} in March 2013, including compensation paid in 2013 and severance.	No information disclosed
2016079	All records related to notices filed in connection with LRT construction-related business losses and the number of notices that have been received by the Region of Waterloo. on the same topic.	No information disclosed

NLP - Summary of Requests

Broadly generalizing, there are few steps one needs to do before analyzing any text:

- Tokenize the text - break the text in single words, i.e., tokens.
- Remove any unwanted characters (\n), and punctuation ("-", "...", “””.)
- Remove URLs or replace them with a word, say, "URL".
- Remove screen names or replace the '@'.
- Remove capitalization of words.
- Remove words with less than *n* characters (*n* = 4?)
- Remove *stop words* - examples are words such as 'a', 'the', 'and'.
- Lemmatization - group together the inflected forms of a word so they can be analyzed as a single item, identified by the word's lemma, or dictionary form.

After preparing the text:



Summary_of_Request	Edited_Summary
Minutes of Service Delivery Subcommittee of ESCAC for period of January 1, 1997 to January 13, 1999.	minutes service delivery subcommittee escac period january 1997 january 1999
Public Health inspection reports for the {location removed}, Kitchener for the past 3 years.	public health inspection report location remove kitchener past year
Public Health inspection records for {location removed}, Cambridge for the past 2 years.	public health inspection record location remove cambridge past year
Public Health inspection records for {address removed}, Cambridge, relating to sink odours in 1994.	public health inspection record address remove cambridge relate sink odour 1994
Vendor list report with total of year-to-date purchases at fiscal year end for 1996, 1997, and 1998.	vendor list report total year date purchase fiscal year 1996 1997 1998
Public Health inspection file for {name removed} at {location removed} regarding requester's dismissal from employment.	public health inspection file remove location remove regard requester dismissal employment
Scope of work and deliverables sections of contract between Region of Waterloo and {company name removed} for Waterloo Regional Master Transportation Plan.	scope work deliverable section contract region waterloo company remove waterloo regional master transportation plan
Number of contracts and dollar amount of contracts between Region of Waterloo and {company name removed} for the last 5 years.	number contract dollar contract region waterloo company remove year
Public Health inspection report regarding a complaint about contamination found in coffee cup at {location removed}, Cambridge.	public health inspection report regard complaint contamination coffee location remove cambridge

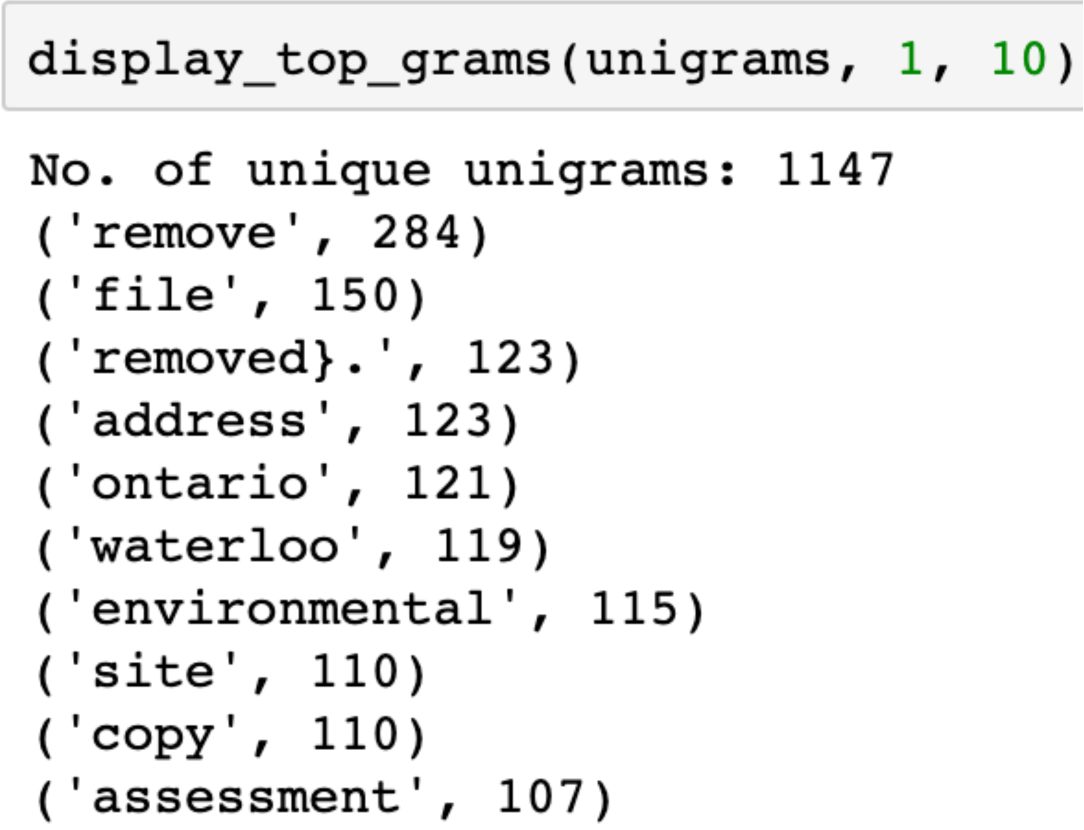
NLTK n-grams

n-grams are sets of co-occuring words within a given window, typically moving one word forward.

* unigrams - single words

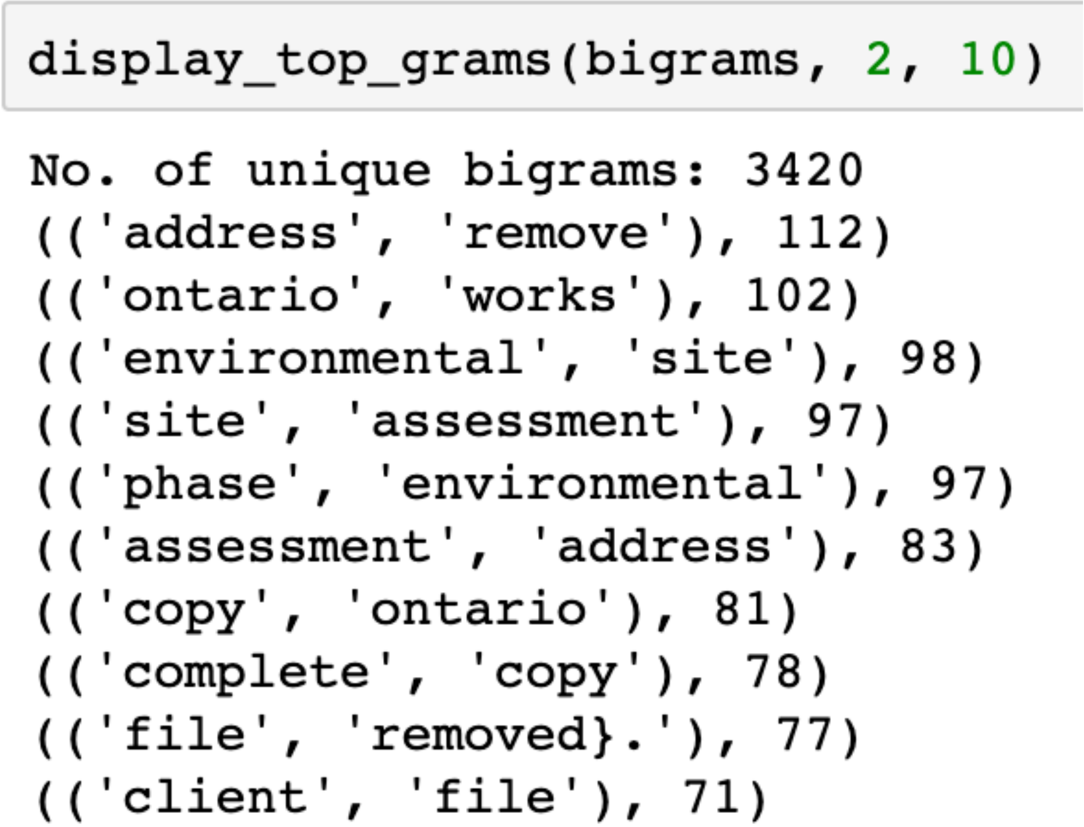
* bigrams - sets of two words

Out of about 9000 words/tokens, let’s find the most common n-grams:



```
display_top_grams(unigrams, 1, 10)
```

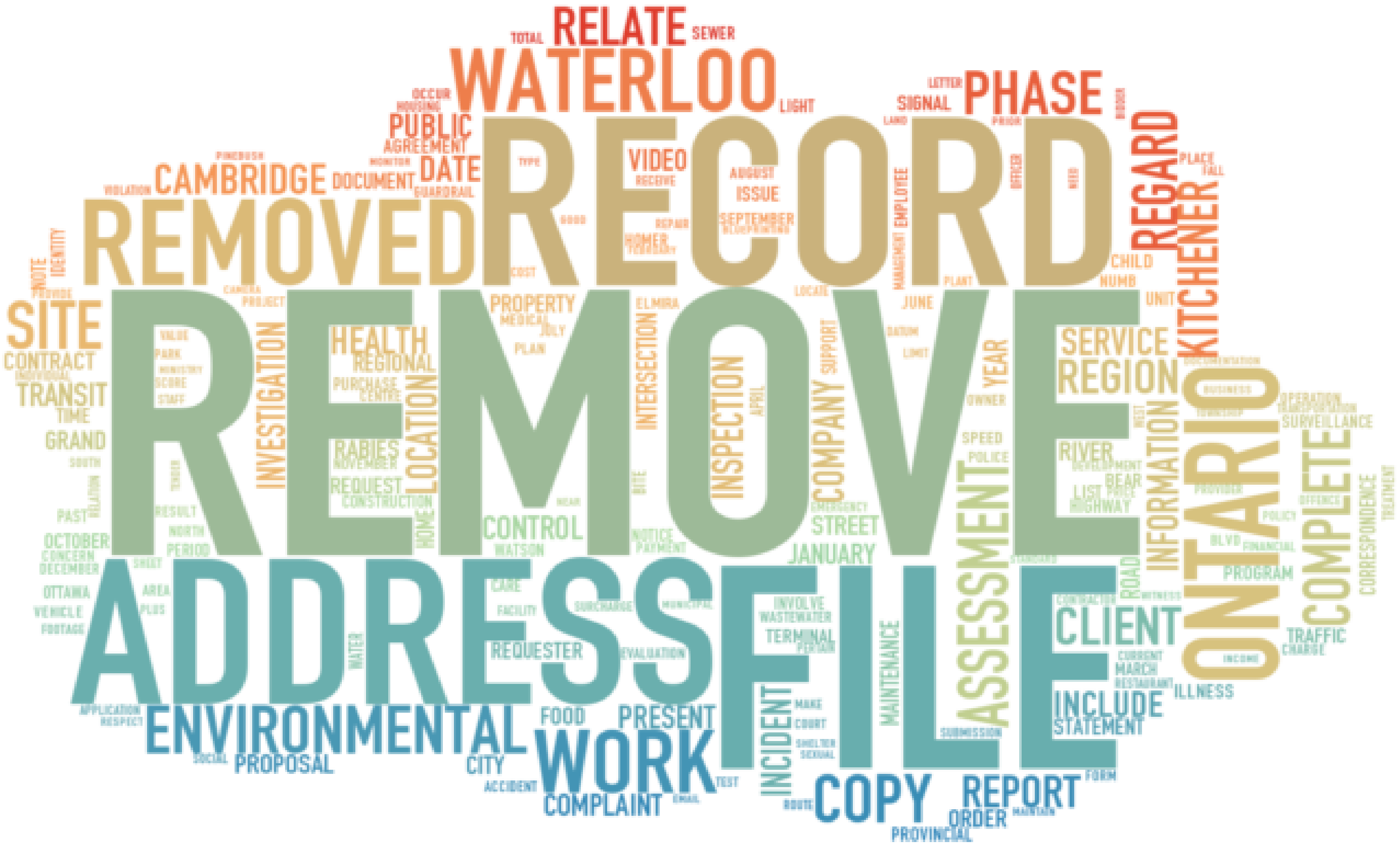
```
No. of unique unigrams: 1147
('remove', 284)
('file', 150)
('removed}.', 123)
('address', 123)
('ontario', 121)
('waterloo', 119)
('environmental', 115)
('site', 110)
('copy', 110)
('assessment', 107)
```



```
display_top_grams(bigrams, 2, 10)
```

```
No. of unique bigrams: 3420
(('address', 'remove'), 112)
(('ontario', 'works'), 102)
(('environmental', 'site'), 98)
(('site', 'assessment'), 97)
(('phase', 'environmental'), 97)
(('assessment', 'address'), 83)
(('copy', 'ontario'), 81)
(('complete', 'copy'), 78)
(('file', 'removed}.'), 77)
(('client', 'file'), 71)
```

EDA: Word Clouds



Top 200 unigrams, full text

“Remove”?!

Many of these requests have names of people or locations that needed to be removed for privacy reasons:

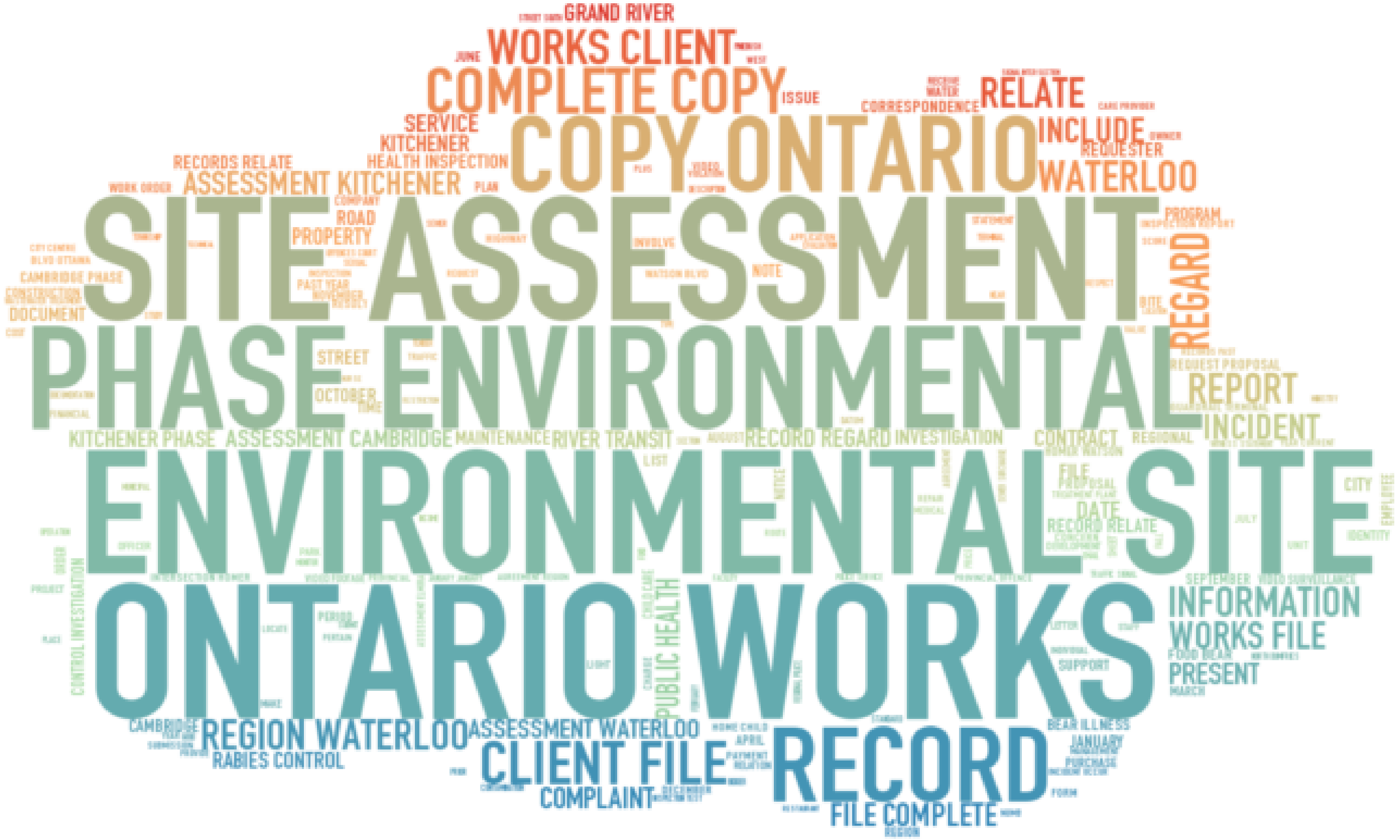
{address removed}, {name removed}, {location removed}, {company name removed}, {intersection removed}, …

Reprocessing the text using regEx:

```
regex_phrase = r'(?:\{\w+\s*\w*\s*\w*\s*\w*\s*\w*\s*\w*\}|\
                \(\w+\s*\w*\s*\w*\s*\w*\s*\w*\s*\w*\}|\
                \{\w+\s*\w*\s*\w*\s*\w*\s*\w*\s*\w*\)|\
                \{\w+\s*\w*\s*\w*\s*\w*\s*\w*\s*\w*\{|
                \(\w+\s*removed\)|\
                  )’
```

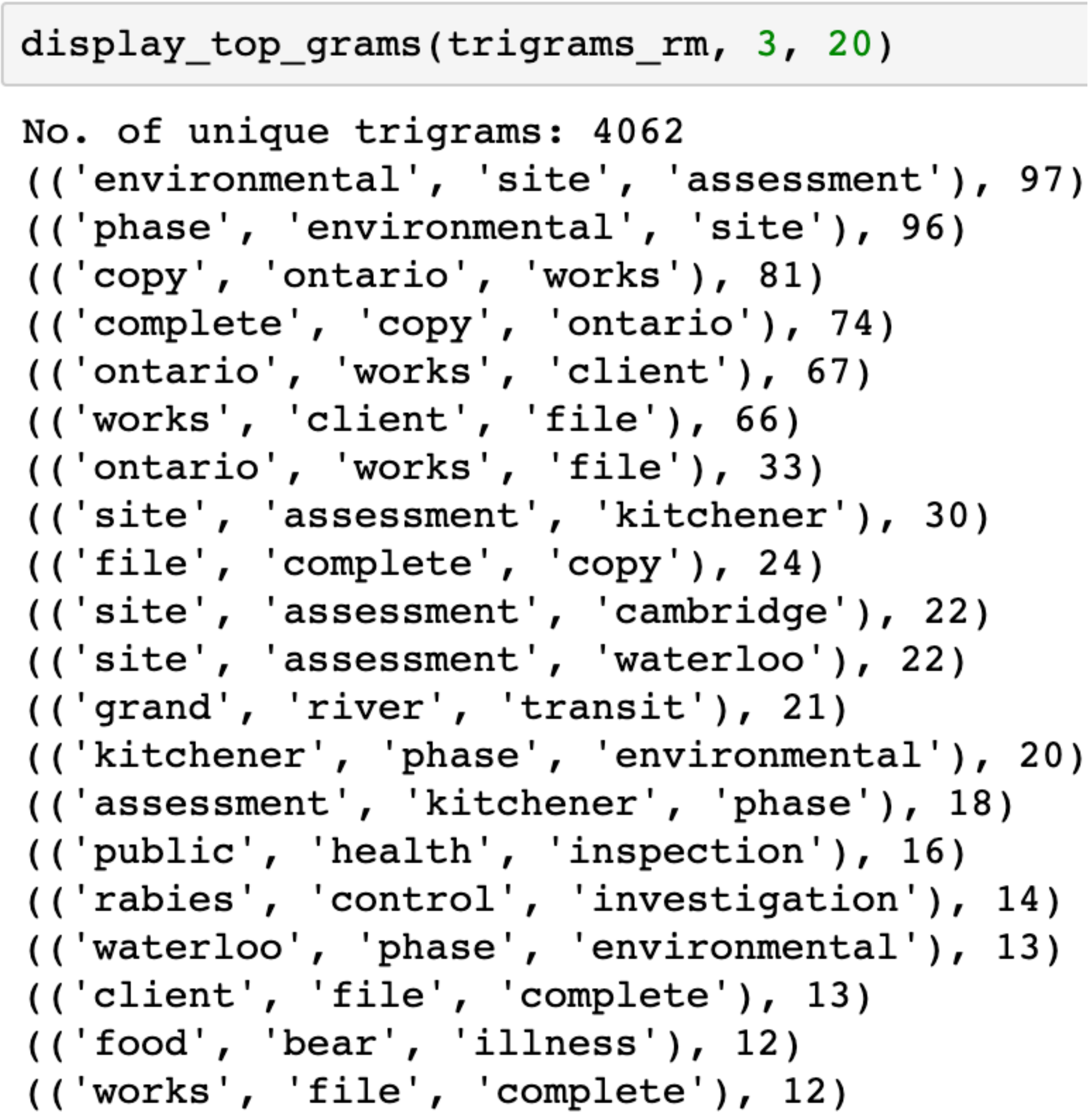
More than 33 variations, about 2% of the text.

… and allowing bigrams in the Word Cloud



Top 200 unigrams/bigrams, full text without '{* remove}'

Trigrams



```
display_top_grams(trigrams_rm, 3, 20)
```

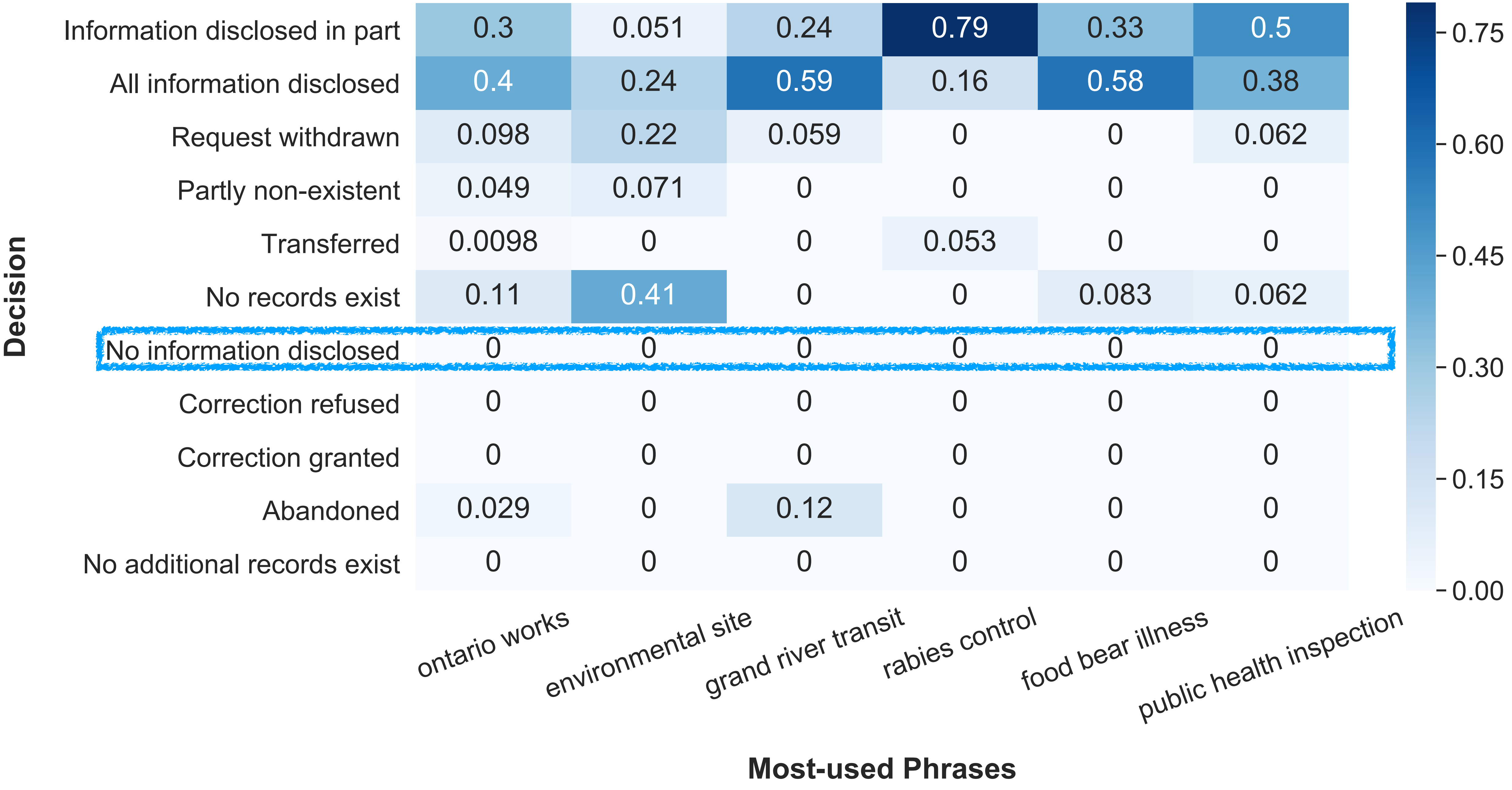
```
No. of unique trigrams: 4062
(('environmental', 'site', 'assessment'), 97)
(('phase', 'environmental', 'site'), 96)
(('copy', 'ontario', 'works'), 81)
(('complete', 'copy', 'ontario'), 74)
(('ontario', 'works', 'client'), 67)
(('works', 'client', 'file'), 66)
(('ontario', 'works', 'file'), 33)
(('site', 'assessment', 'kitchener'), 30)
(('file', 'complete', 'copy'), 24)
(('site', 'assessment', 'cambridge'), 22)
(('site', 'assessment', 'waterloo'), 22)
(('grand', 'river', 'transit'), 21)
(('kitchener', 'phase', 'environmental'), 20)
(('assessment', 'kitchener', 'phase'), 18)
(('public', 'health', 'inspection'), 16)
(('rabies', 'control', 'investigation'), 14)
(('waterloo', 'phase', 'environmental'), 13)
(('client', 'file', 'complete'), 13)
(('food', 'bear', 'illness'), 12)
(('works', 'file', 'complete'), 12)
```

We see that there are common phrases:

- ‘ontario works’,
- ‘environmental site’
- ‘grand river transit’
- ‘rabies control’
- ‘public health inspection’
- ‘food bear illness’ (as in *‘food borne illness’*

How common are these phrases?

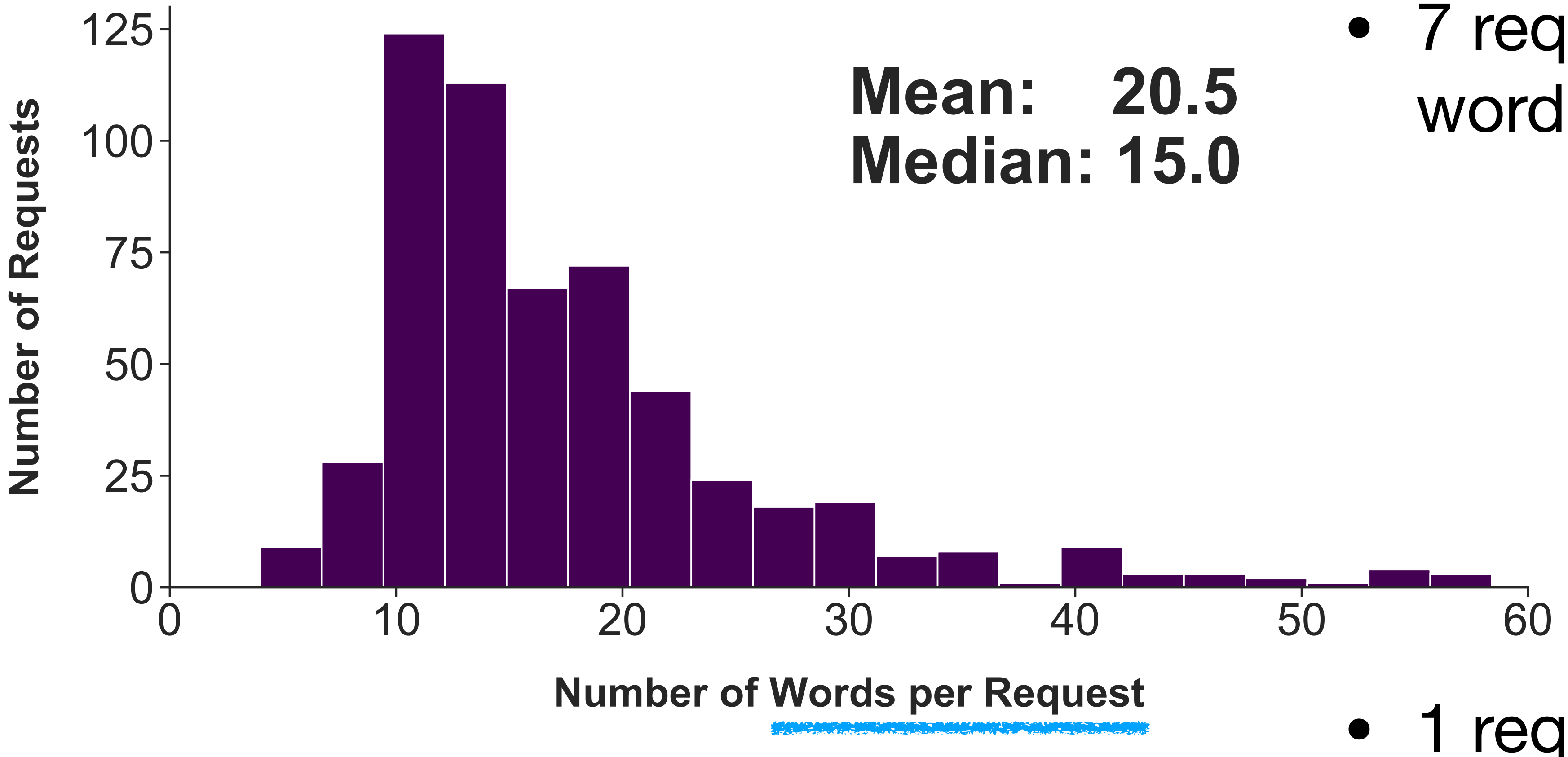
46% of the full data uses the following phrases.
For each phrase, here is how decisions are split (fraction).



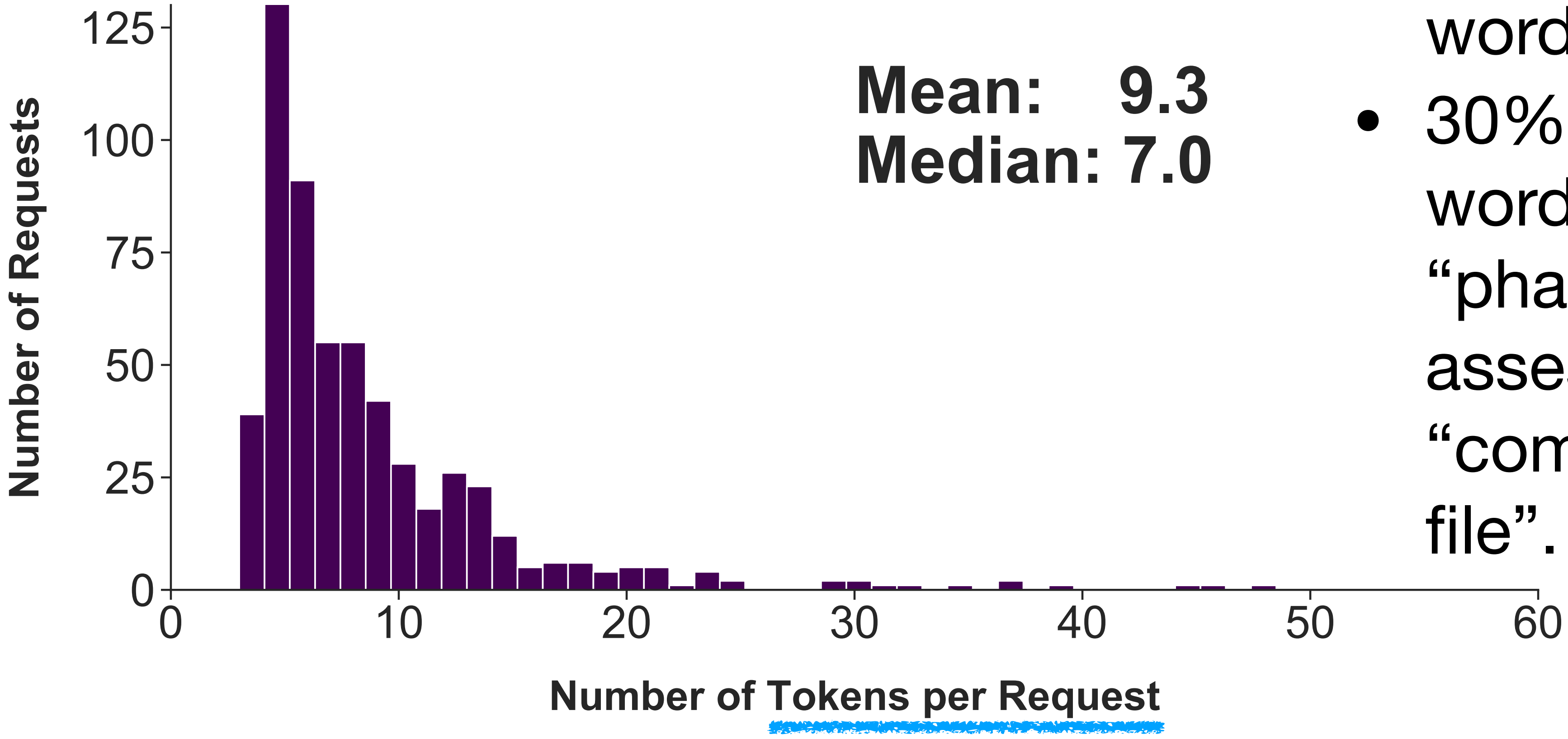
Decision	ontario works	environmental site	grand river transit	rabies control	food bear illness	public health inspection
Information disclosed in part	0.3	0.051	0.24	0.79	0.33	0.5
All information disclosed	0.4	0.24	0.59	0.16	0.58	0.38
Request withdrawn	0.098	0.22	0.059	0	0	0.062
Partly non-existent	0.049	0.071	0	0	0	0
Transferred	0.0098	0	0	0.053	0	0
No records exist	0.11	0.41	0	0	0.083	0.062
No information disclosed	0	0	0	0	0	0
Correction refused	0	0	0	0	0	0
Correction granted	0	0	0	0	0	0
Abandoned	0.029	0	0.12	0	0	0
No additional records exist	0	0	0	0	0	0

Most-used Phrases

Requests Statistics



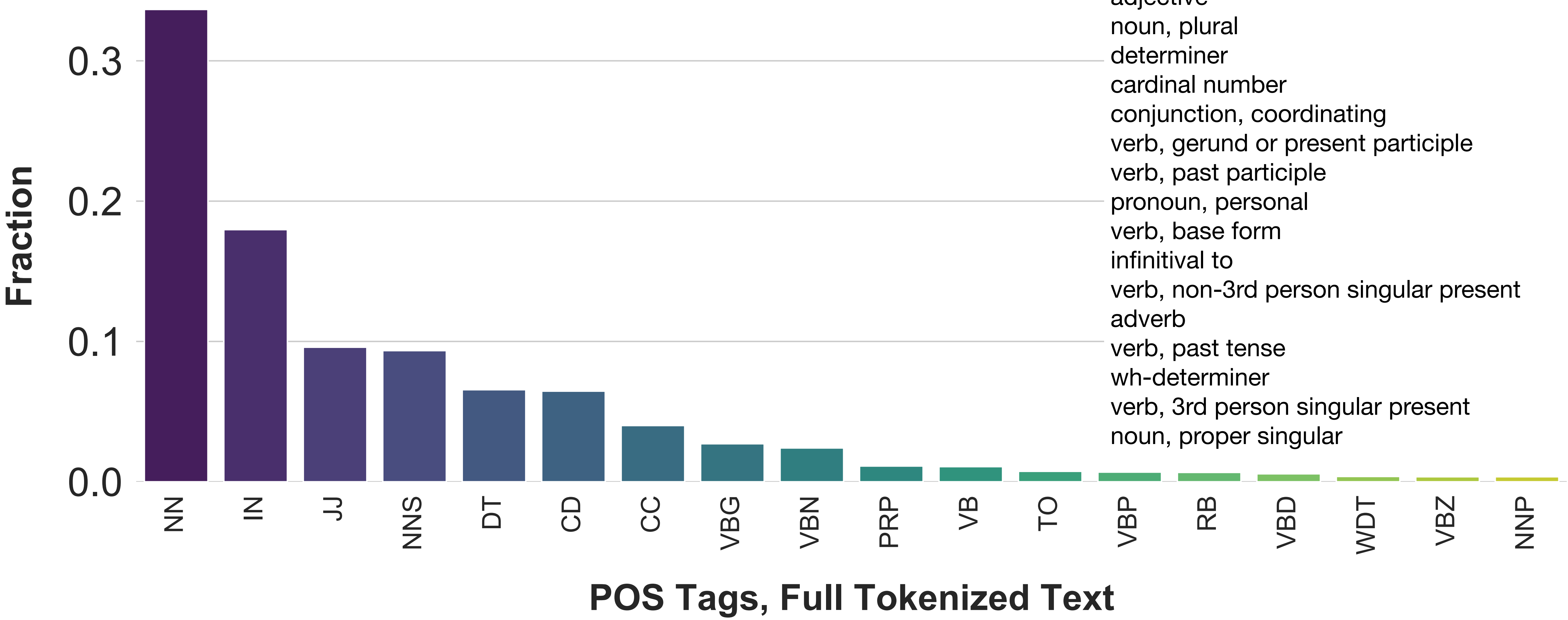
- 7 requests with more than 100 words.



- 1 request with more than 100 words.
- 30% of the requests have 5 words or less:
  “phase environmental site assessment kitchener”
  “complete copy ontario works file”.

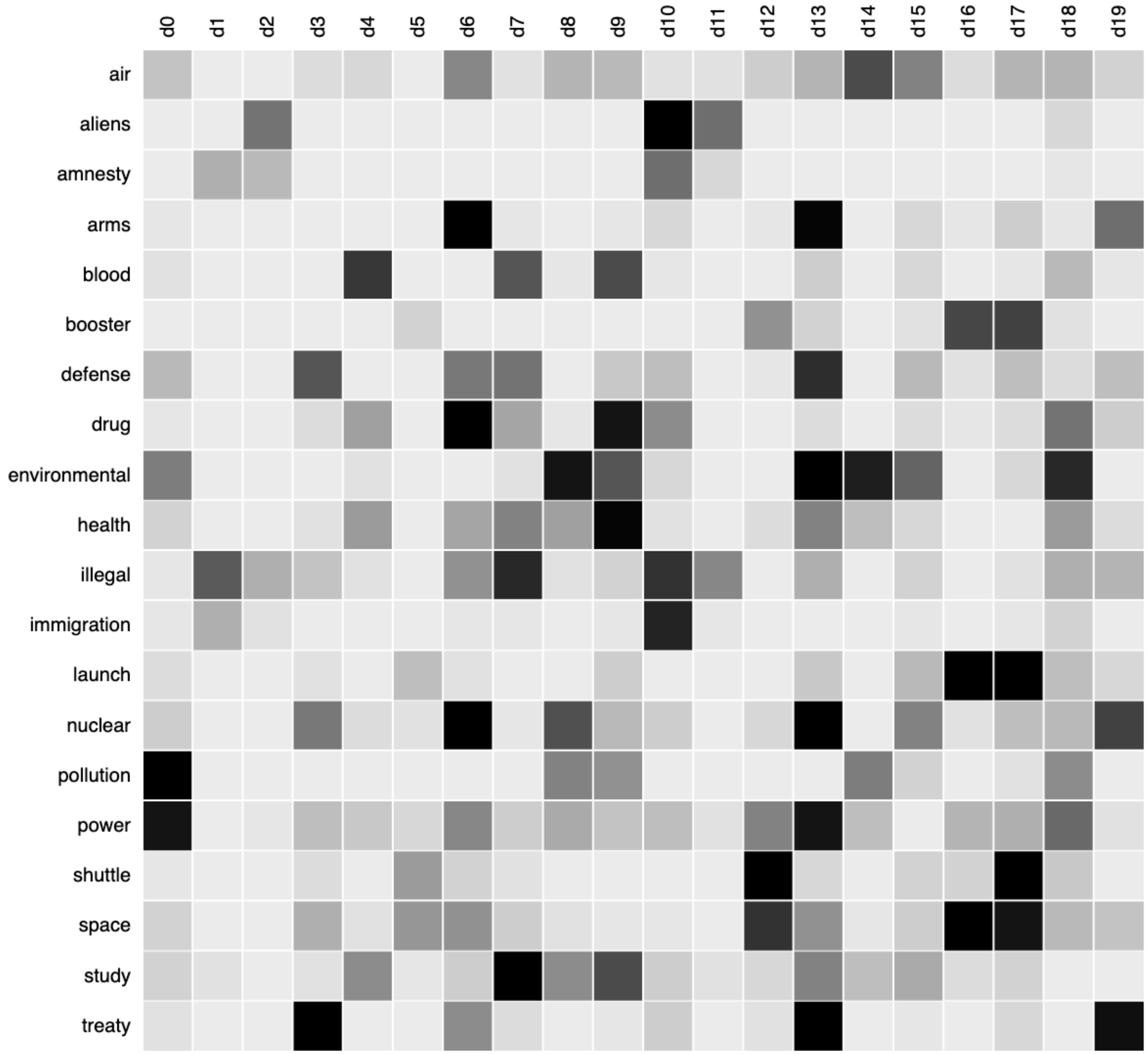
Part-of-Speech (POS) Tagging

- POS tagging is the identification of words as nouns, verbs, adjectives, adverbs, etc., based on its definition and context.
- Used as features to build parse trees, which can be used for Named Entity Resolution, Coreference Resolution, Sentiment Analysis and Question Answering.



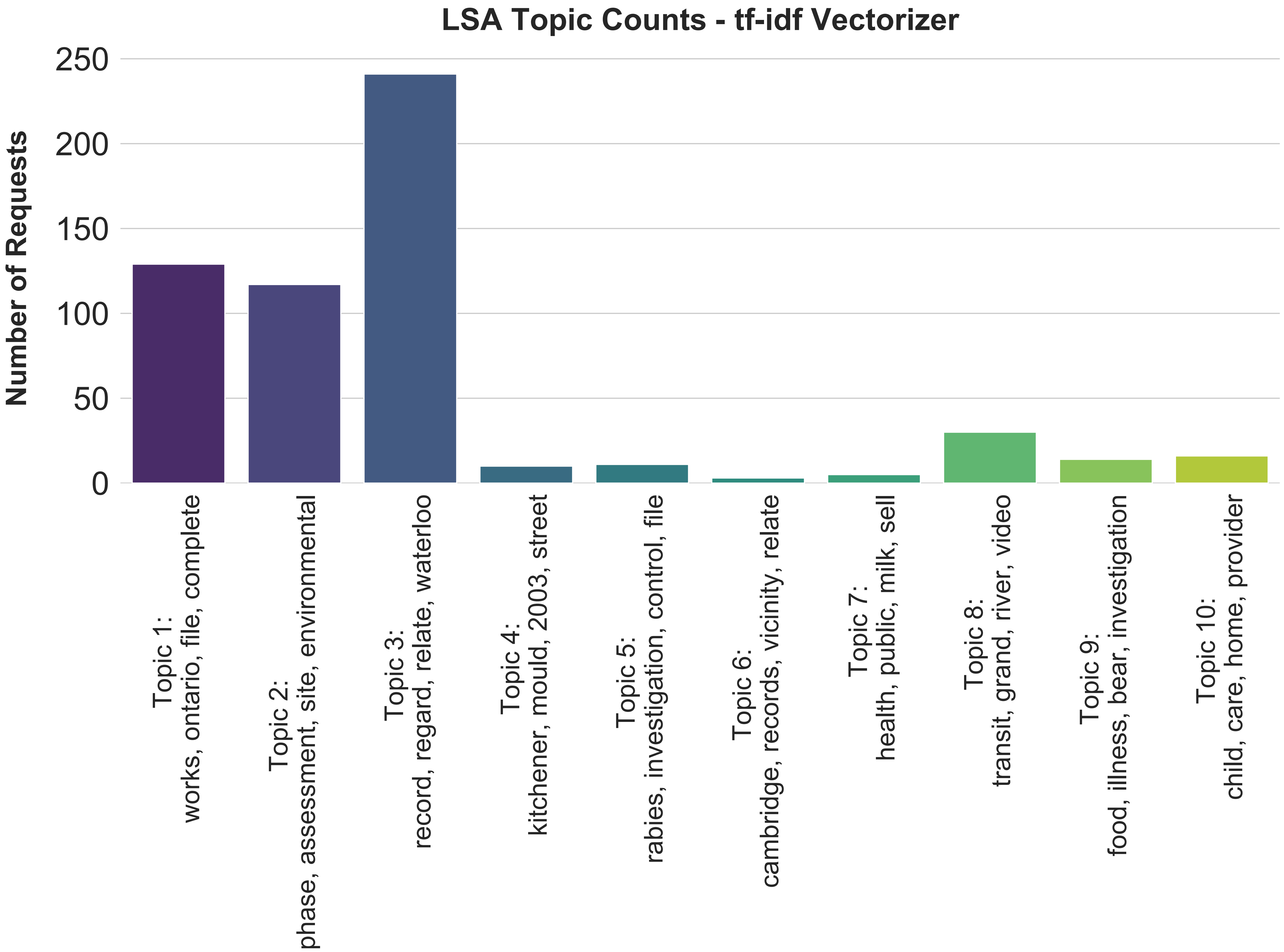
noun, singular or mass
conjunction, subordinating or preposition
adjective
noun, plural
determiner
cardinal number
conjunction, coordinating
verb, gerund or present participle
verb, past participle
pronoun, personal
verb, base form
infinitival to
verb, non-3rd person singular present
adverb
verb, past tense
wh-determiner
verb, 3rd person singular present
noun, proper singular

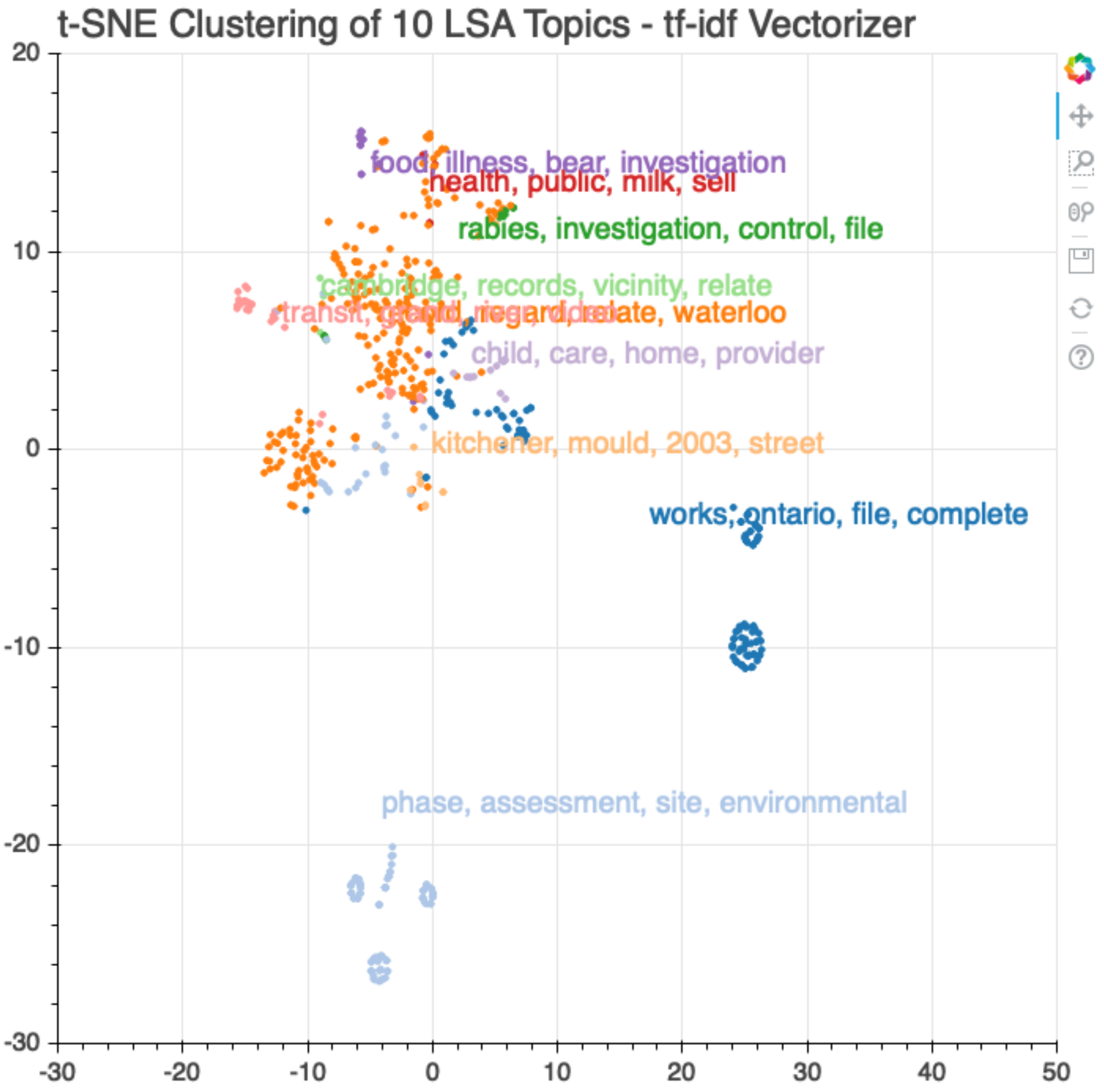
Topic Modeling



- Statistical model and text mining tool for discovering the abstract "topics" that occur in a collection of documents.
- Latent Dirichlet allocation (LDA)
- Latent Semantic Analysis (LSA)
- Vectorizers: Count and tf-idf

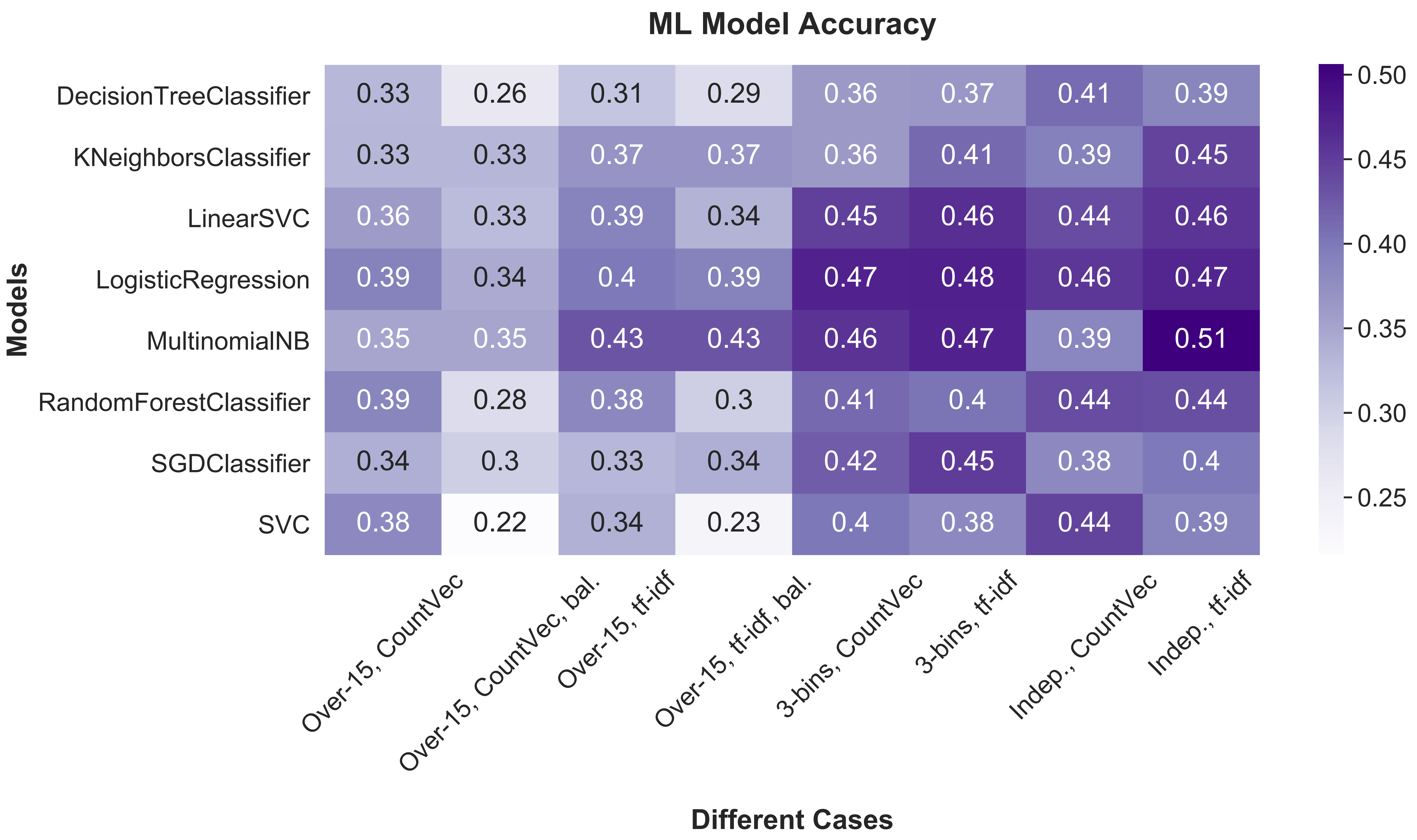
Title: Topic model scheme.webm
Author: Christoph Carl Kling
https://en.wikipedia.org/wiki/Latent_semantic_analysis#cite_note-3


LSA Topic Counts - tf-idf Vectorizer
Number of Requests
250
200
150
100
50
0
Topic 1:
works, ontario, file, complete
Topic 2:
phase, assessment, site, environmental
Topic 3:
record, regard, relate, waterloo
Topic 4:
kitchener, mould, 2003, street
Topic 5:
rabies, investigation, control, file
Topic 6:
cambridge, records, vicinity, relate
Topic 7:
health, public, milk, sell
Topic 8:
transit, grand, river, video
Topic 9:
food, illness, bear, investigation
Topic 10:
child, care, home, provider



T-distributed Stochastic Neighbor Embedding (t-SNE): machine learning algorithm for visualization. It is a nonlinear dimensionality reduction technique for embedding high-dimensional data for visualization in a low-dimensional space of two or three dimensions.

Machine Learning



ML Model Accuracy

Models	Over-15, CountVec	Over-15, CountVec, bal.	Over-15, tf-idf	Over-15, tf-idf, bal.	3-bins, CountVec	3-bins, tf-idf	Indep., CountVec	Indep., tf-idf
DecisionTreeClassifier	0.33	0.26	0.31	0.29	0.36	0.37	0.41	0.39
KNeighborsClassifier	0.33	0.33	0.37	0.37	0.36	0.41	0.39	0.45
LinearSVC	0.36	0.33	0.39	0.34	0.45	0.46	0.44	0.46
LogisticRegression	0.39	0.34	0.4	0.39	0.47	0.48	0.46	0.47
MultinomialNB	0.35	0.35	0.43	0.43	0.46	0.47	0.39	0.51
RandomForestClassifier	0.39	0.28	0.38	0.3	0.41	0.4	0.44	0.44
SGDClassifier	0.34	0.3	0.33	0.34	0.42	0.45	0.38	0.4
SVC	0.38	0.22	0.34	0.23	0.4	0.38	0.44	0.39

Summary

ML fails in this case because we don’t have enough data.

But do not despair, not everything is lost!

There are other tools we can use to extract valuable information and insights.

- Descriptive Statistics
- Exploratory Data Analysis
- (text) Natural Language Processing tools:
  Macro understanding: n-grams, topic modeling, word clouds, …
  Micro understanding: POS-tagging, Name Entity Recognition and Resolution, …

Remember, *understanding your data* should always be the first step towards ML.