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:

- 26 files - years 1991 to 2016
- 902 requests in total
- All have the same six columns (amazingly!)

And it looks like this:

	Request_Number	Request_Type	Source	Summary_of_Request	Decision	OBJECTID
0	98001	Personal Information	Public	GWA file, specifically agreements, cheques and...	All disclosed	0
1	98002	General Information	Public	Records related to construction on {company na...	Partly non-existent	1
2	98003	General Information	Business	Information regarding damage to Regional facil...	Partly non-existent	2
3	98004	General Information	Individual by Agent	Identity of the Sunnyside Home employee who wi...	Nothing disclosed	3
4	98005	Personal Information	Public	Regional Solicitor's file for {name removed} r...	Partly exempted	4

The Columns & The Cleaning

Before

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

```
14
```

```
General Information                                439
Personal Information                               262
General                                             57
General Records                                     36
Personal                                            25
Personal                                            22
General                                             19
Personal Health Information/General Information     16
Personal information                                13
Correction                                           7
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                                551
Personal                               322
Personal Health Information/General     17
Correction                               7
Personal Health Information              3
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

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

```
13
```

```
Public                      376
Business                    214
Individual by Agent         149
Individual by agent          40
Individual                   26
Business by Agent            26
Media                        25
Individual by agent          19
Individual                   14
Business                      9
Business                      2
Media                         1
Individual for dependant      1
```

After

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

```
6
```

```
Individual                  416
Business                    225
Individual by Agent         208
Media                        26
Business by Agent            26
Individual for dependant      1
```

… even more cleaning

Before

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

```
24
```

All disclosed	265
Partly exempted	232
Withdrawn	123
No records exist	55
Information disclosed in part	50
Partly non-existent	36
Nothing disclosed	32
No record exists	21
Forwarded out	17
All Information disclosed	16
All information disclosed	13
Abandoned	13
No responsive records exist	11
Non-existent	3
Correction refused	3
Transferred to Region of Waterloo Public Health	2
Correction made	2
All disclosed	2
Request withdrawn	1
Transferred	1
No information disclosed	1
Correction granted	1
No additional records exist	1
Statement of disagreement filed	1

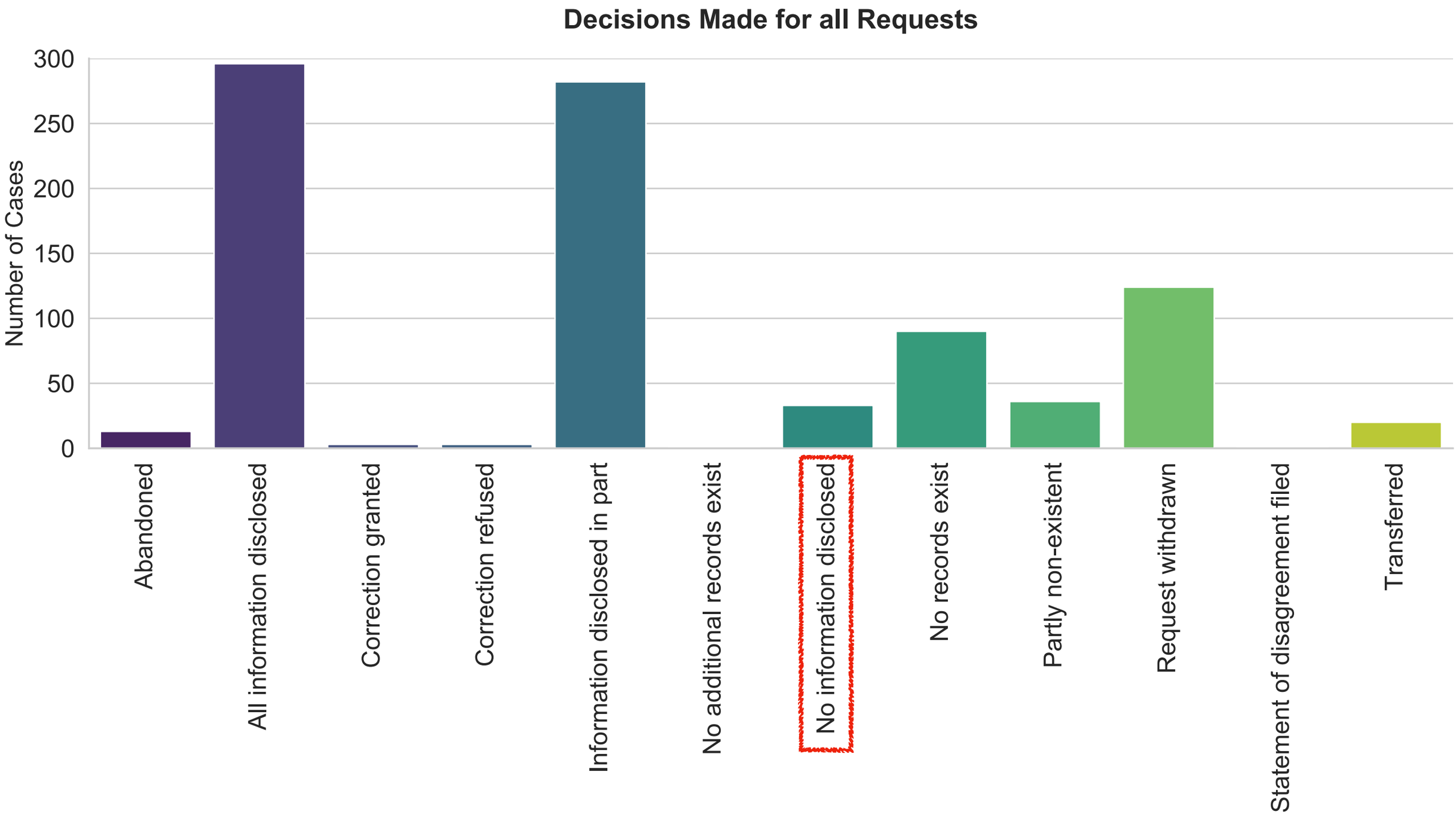
After

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

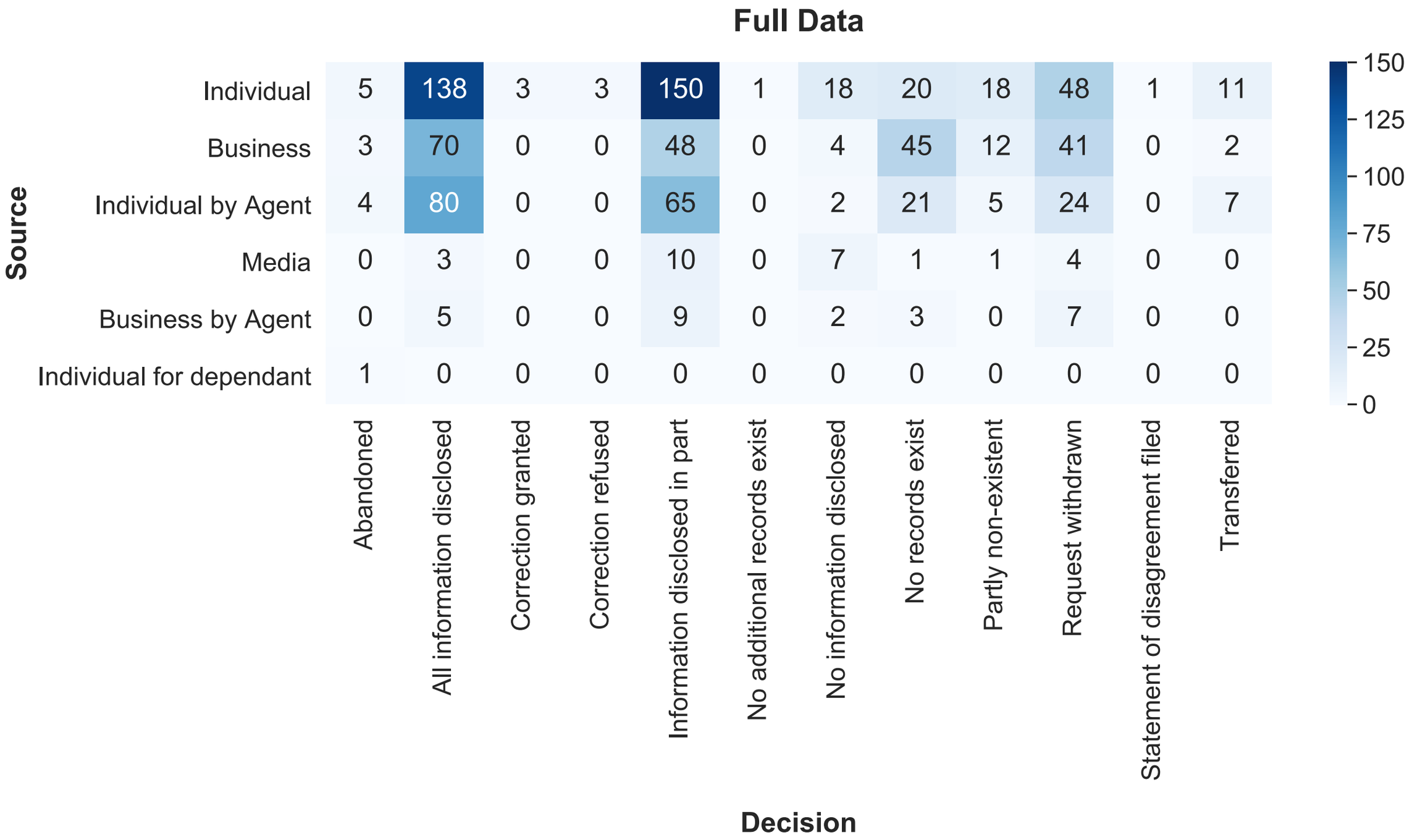
```
12
```

All information disclosed	296
Information disclosed in part	282
Request withdrawn	124
No records exist	90
Partly non-existent	36
No information disclosed	33
Transferred	20
Abandoned	13
Correction refused	3
Correction granted	3
No additional records exist	1
Statement of disagreement filed	1

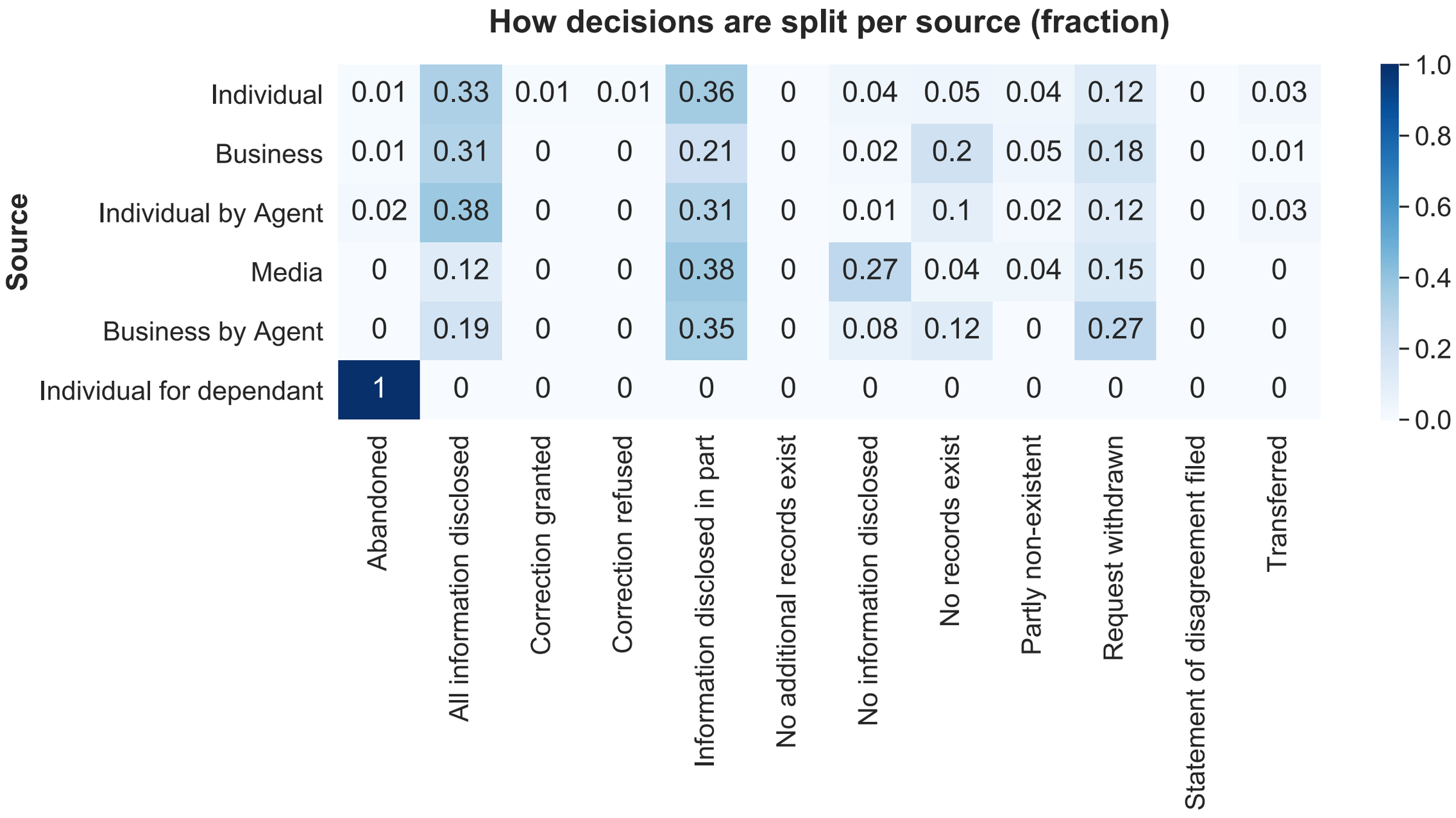
Descriptive Statistics: Univariate Analysis



Bivariate Analysis



Bivariate Analysis



And of those media requests:

Request_Number	Summary_of_Request	Decision
2006007	Quantity of pesticide used by Region of Waterloo including invoices for contracted application.	All information disclosed
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
96025	All records regarding the Family Awareness Centre or its four programs including Parents Are People Too.	All information disclosed
98012	Ontario Works participating community agencies and number of clients assigned to each.	No 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'.
- Lemmatize - 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
GWA file, specifically agreements, cheques and monetary income statements regarding "court order."	file specifically agreement cheque monetary income statement regard court order
Records related to construction on {company name removed} site in 1996 and 1997.	records relate construction company remove site 1996 1997
Information regarding damage to Regional facilities from a severe storm on May 20, 1996.	information regard damage regional facility severe storm 1996
Identity of the Sunnyside Home employee who witnessed a motor vehicle accident on February 20, 1990.	identity sunnyside home employee witness motor vehicle accident february 1990
Regional Solicitor's file for {name removed} regarding their employment with the Waterloo Regional Police Service.	regional solicitor file remove regard employment waterloo regional police service
Tender and contract for wastewater treatment operations including Request for Proposals; communications; evaluations; contract with {company name removed}.	tender contract wastewater treatment operation include request proposals communication evaluation contract company removed}.
Unit pricing (pages 24-28) of winning bid for HHW contract.	unit price page contract
GWA for {name removed} records from January 1, 1993 to present, particularly a letter regarding assets.	remove record january 1993 present particularly letter regard asset
Home Child Care Provider file for {name removed} from 1983 to present.	home child care provider file remove 1983 present
Income Maintenance file for {name removed} narrative notes from January 10-31, 1996; letter from {name removed} dated July 30, 1996.	income maintenance file remove narrative note january 1996 letter remove date july 1996
Various records regarding the voice radio system.	various record regard voice radio

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

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

```
No. of unique unigrams: 1339
('remove', 407)
('file', 295)
('removed}.', 201)
('regard', 146)
('record', 136)
('information', 136)
('waterloo', 134)
('copy', 132)
('address', 131)
('ontario', 122)
```

Out of about 8000 words/tokens, let's find the most common n-grams:

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

```
No. of unique bigrams: 4481
(('file', 'removed}.'), 122)
(('address', 'remove'), 113)
(('client', 'file'), 104)
(('ontario', 'works'), 103)
(('environmental', 'site'), 99)
(('site', 'assessment'), 98)
(('phase', 'environmental'), 98)
(('complete', 'copy'), 97)
(('assessment', 'address'), 83)
(('copy', 'ontario'), 81)
```

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}'

```
display_top_grams(trigrams_rm, 3, 35)
```

```
No. of unique trigrams: 5475
(('environmental', 'site', 'assessment'), 98)
(('phase', 'environmental', 'site'), 97)
(('copy', 'ontario', 'works'), 81)
(('complete', 'copy', 'ontario'), 74)
(('ontario', 'works', 'client'), 67)
(('works', 'client', 'file'), 66)
(('information', 'competition', 'file'), 46)
(('personal', 'information', 'competition'), 35)
(('ontario', 'works', 'file'), 33)
(('site', 'assessment', 'kitchener'), 30)
(('file', 'complete', 'copy'), 28)
(('home', 'child', 'care'), 23)
(('site', 'assessment', 'waterloo'), 22)
(('maintenance', 'client', 'file'), 22)
(('income', 'maintenance', 'client'), 22)
(('site', 'assessment', 'cambridge'), 22)
(('grand', 'river', 'transit'), 21)
(('general', 'information', 'competition'), 21)
(('kitchener', 'phase', 'environmental'), 20)
(('file', 'personal', 'information'), 19)
(('assessment', 'kitchener', 'phase'), 18)
(('public', 'health', 'inspection'), 18)
(('child', 'care', 'provider'), 18)
(('competition', 'file', 'personal'), 18)
(('complete', 'copy', 'income'), 17)
(('client', 'file', 'complete'), 17)
(('copy', 'income', 'maintenance'), 17)
(('care', 'provider', 'file'), 15)
(('competition', 'file', 'general'), 14)
(('rabies', 'control', 'investigation'), 14)
(('human', 'resources', 'personal'), 13)
(('food', 'bear', 'illness'), 13)
```

Trigrams

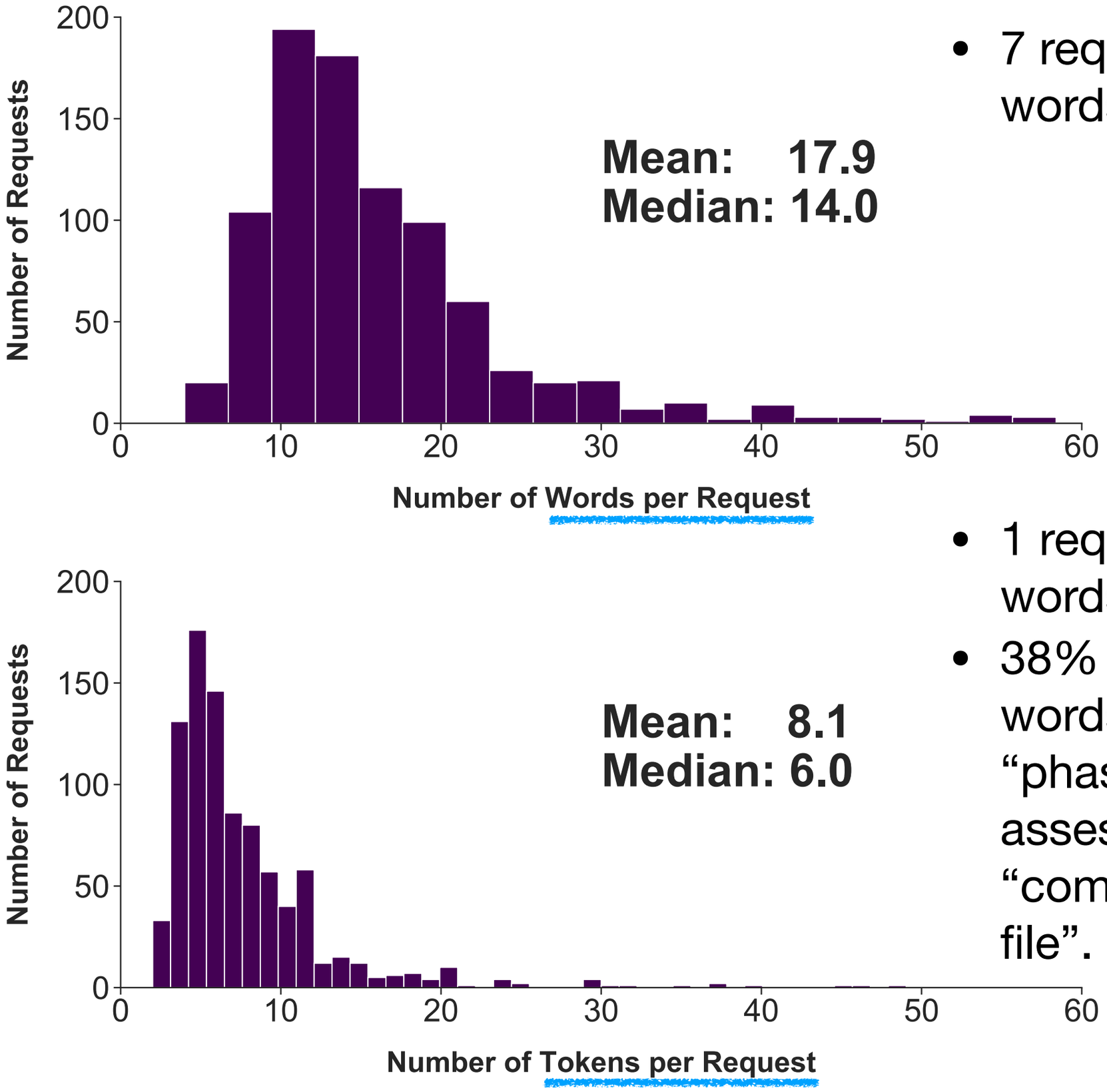
We see that there are common phrases:

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

How common are these phrases?

34% of the full data uses the following phrases.
For each phrase, here is how decisions are split (fraction).

Requests Statistics

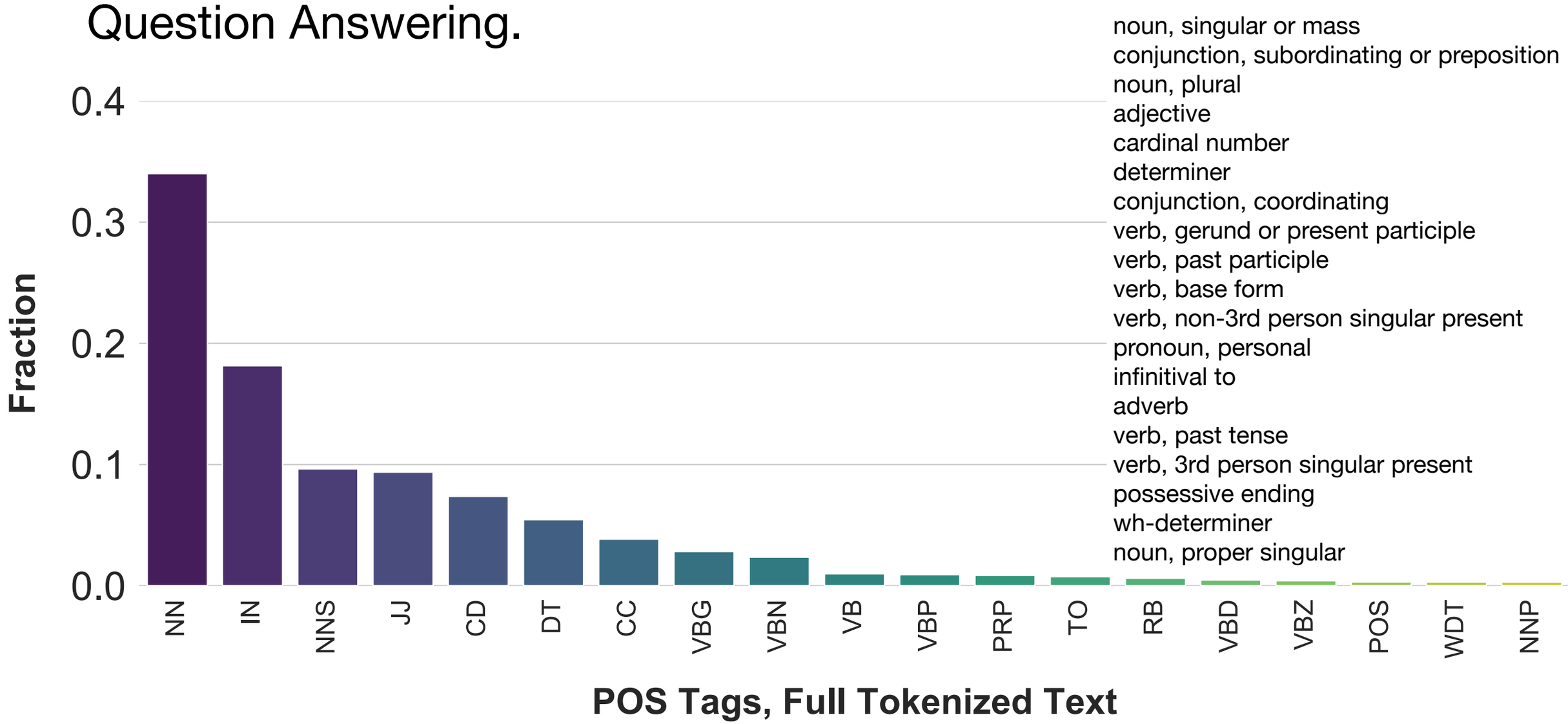


- 7 requests with more than 100 words.

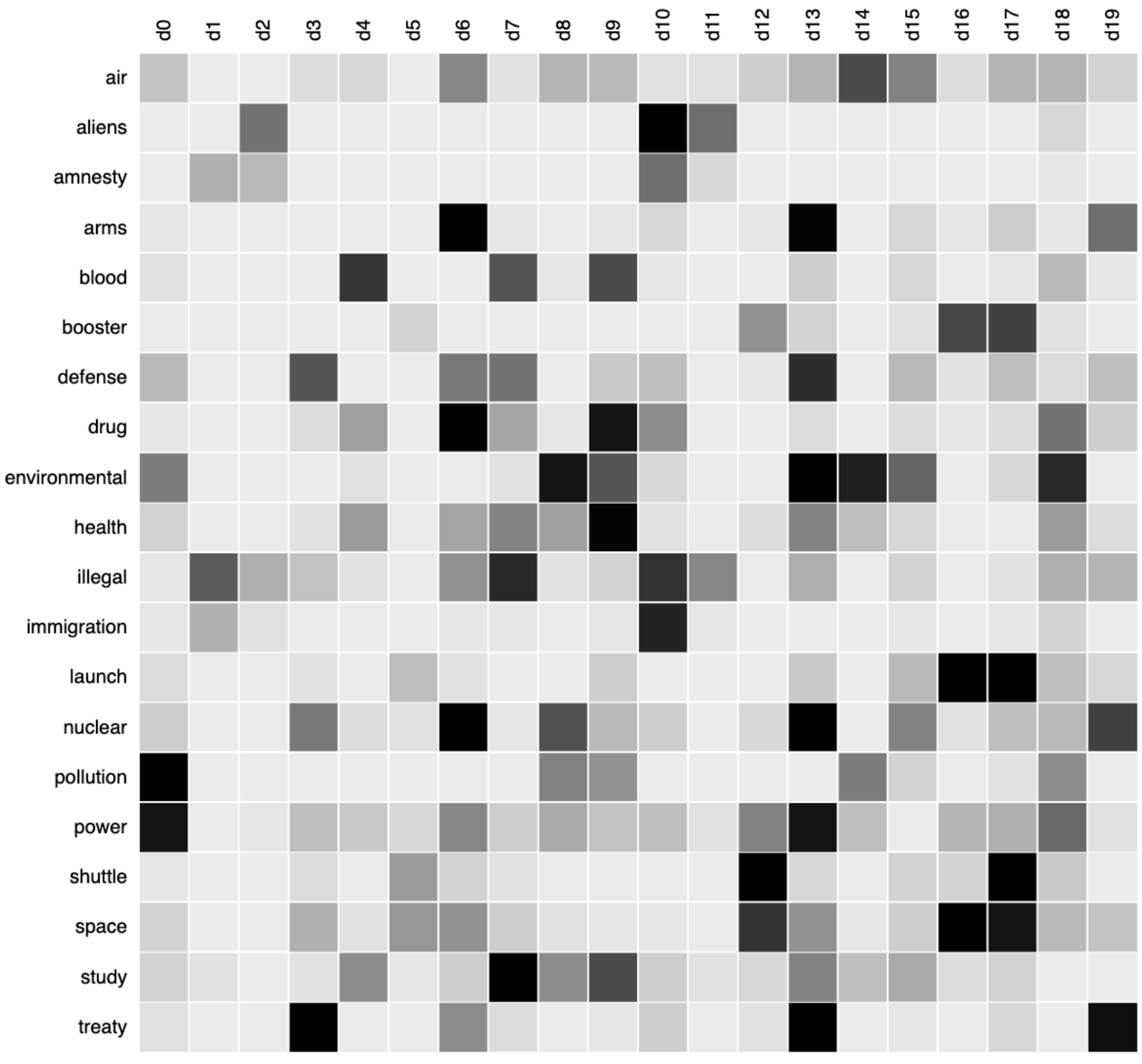
- 1 request with more than 100 words.
- 38% of the requests have 5 words or less:
  “phase environmental site assessment kitchener”
  “complete copy ontario works file”.

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.



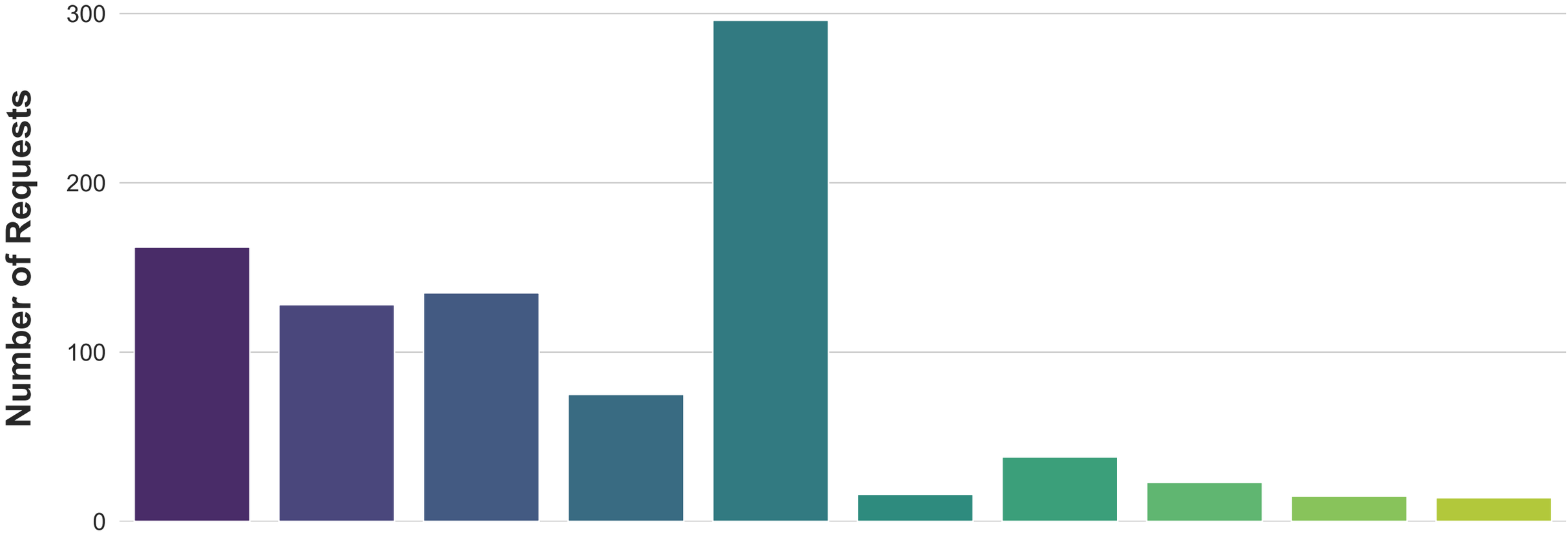
Topic Modeling



- 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

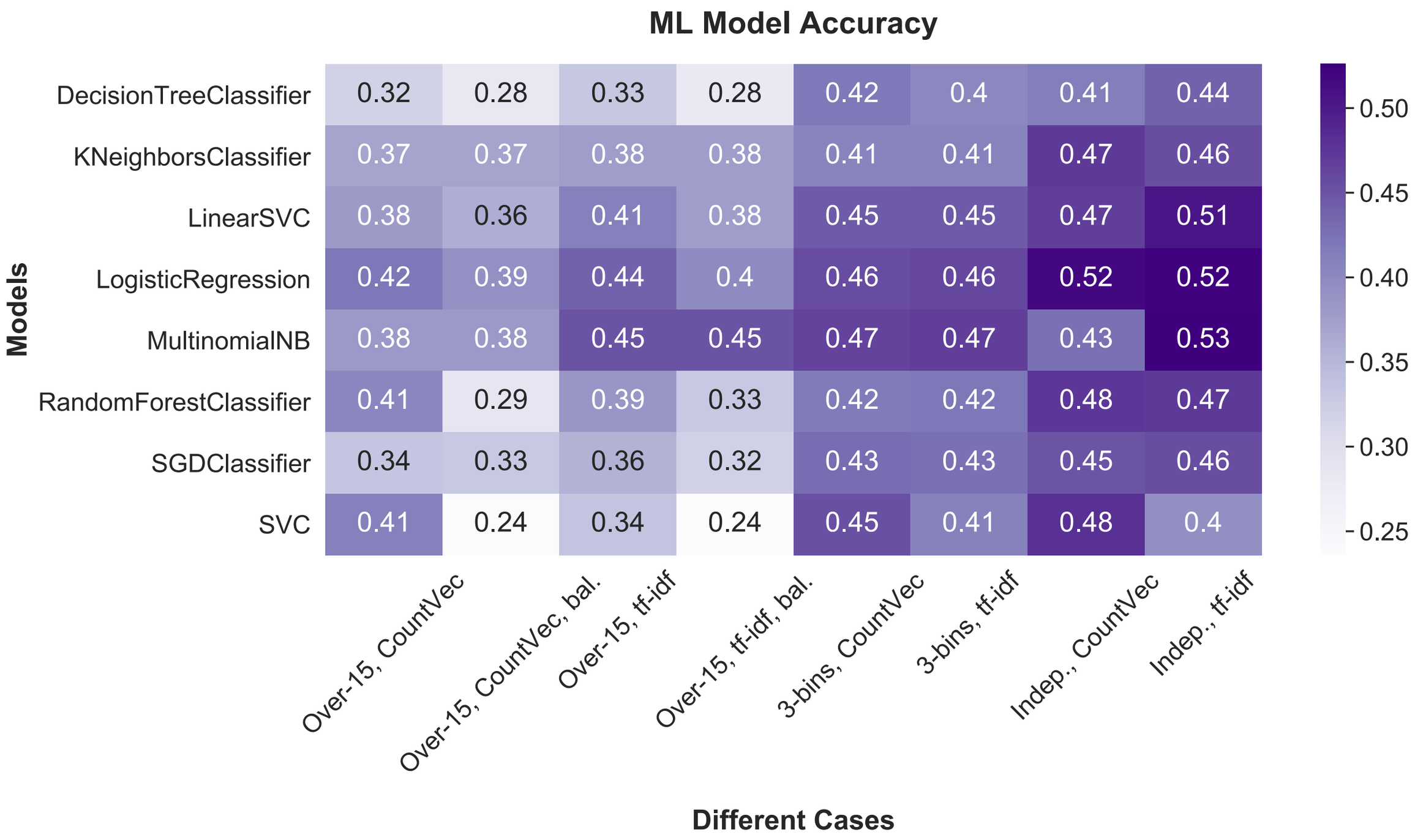
Title: Topic model scheme.webm
Author: Christoph Carl Kling
https://en.wikipedia.org/wiki/Latent_semantic_analysis#cite_note-3

LSA Topic Counts - tf-idf Vectorizer



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



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.