This pdf file encompasses the figures and images produced in the notebook, a.k.a., all the neat results, no code included.

Background

Freedom of Information Requests in the Region of Waterloo

One of the Region of Waterloo initiatives is Open Data. With it, the Region strives to be open, transparent, and accountable to citizens. It shares its data for everyone to use and republish with few restrictions. The data is provided in machine-readable format.

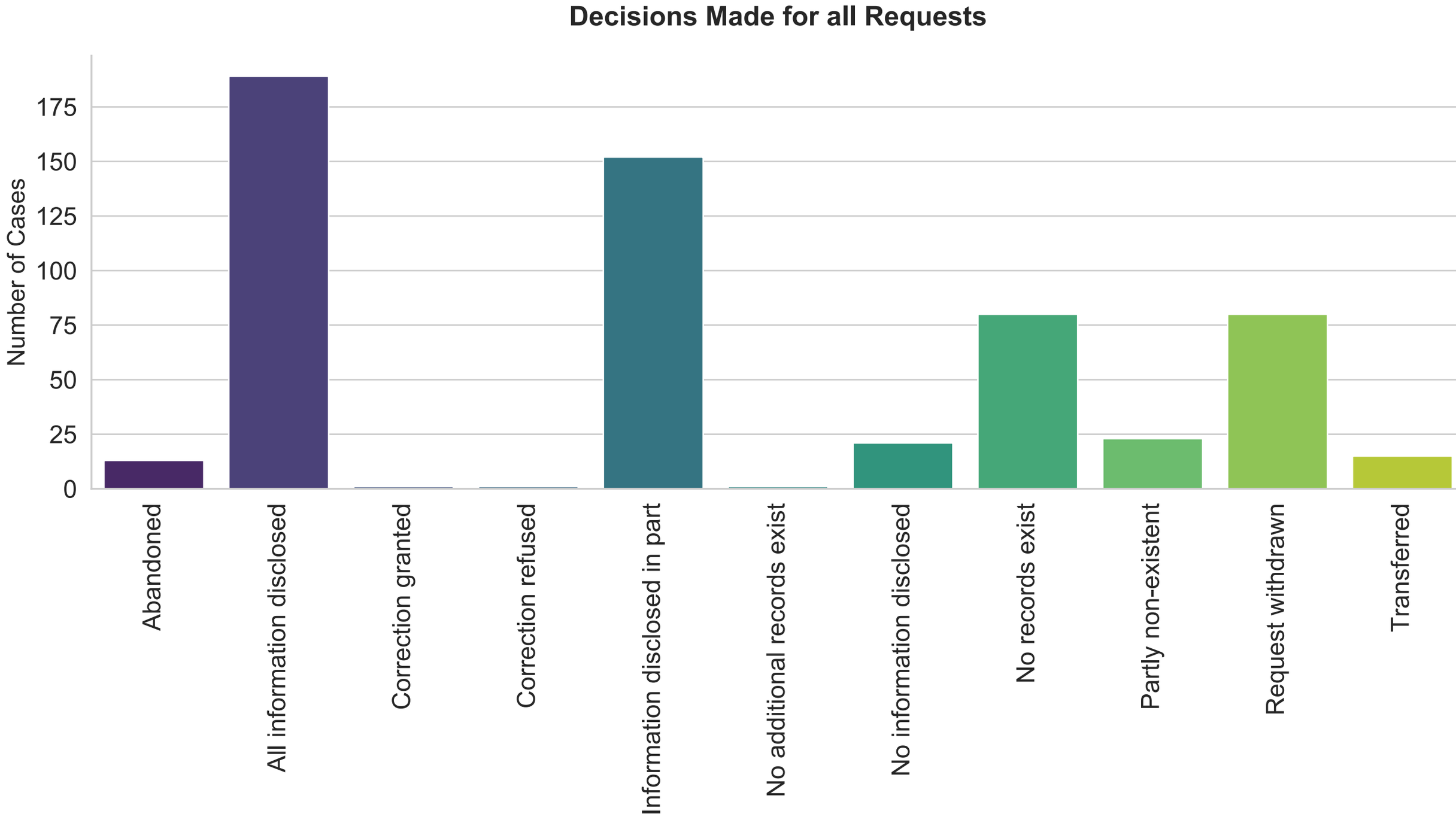
Searching the Region's Open Data Portal, one can find the Freedom of Information Requests (FOIR) data set. This data set spans 18 years (1999-2016).

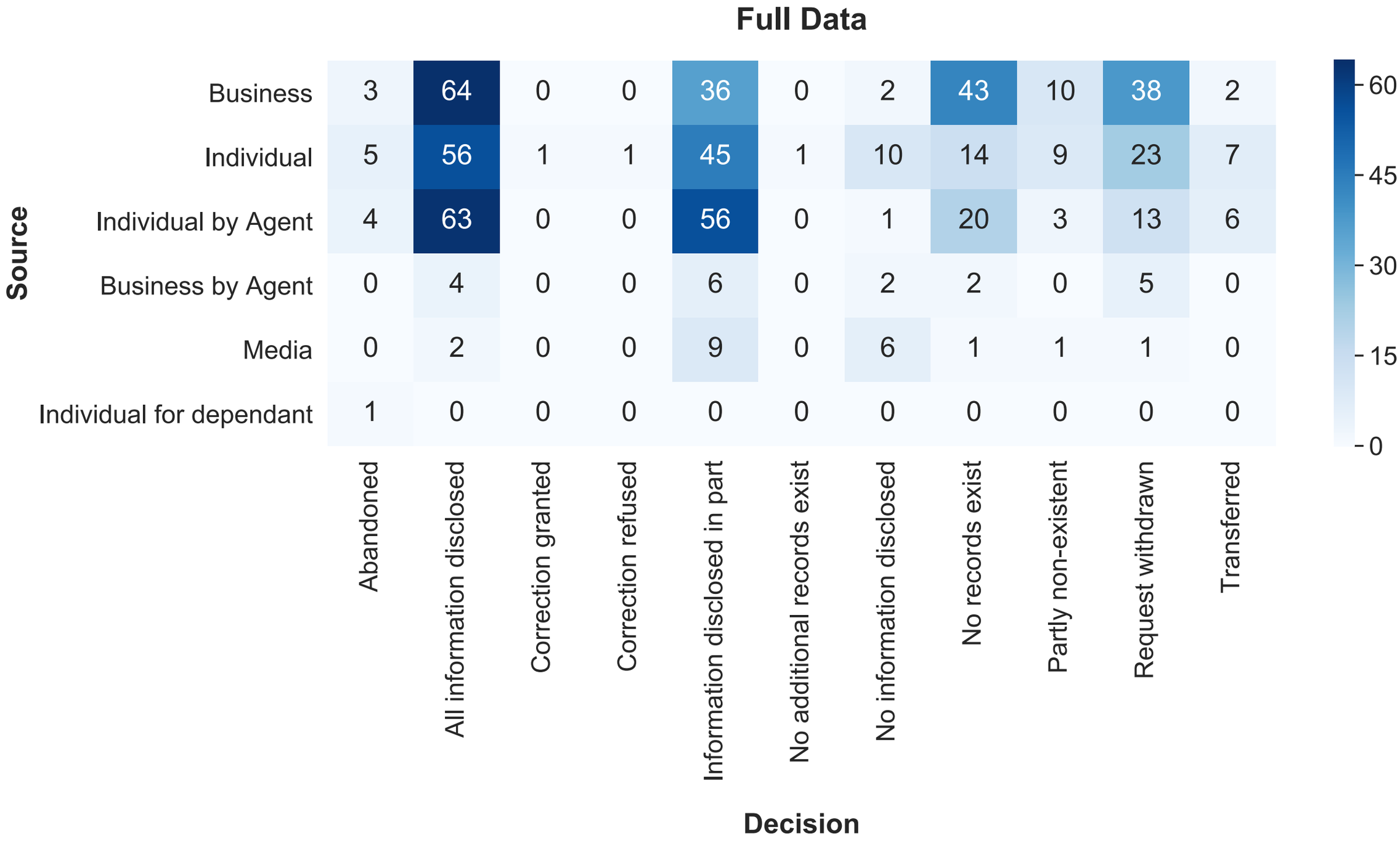
In the repository, you will find a jupyter notebook that does a thorough job at describing the FOIR set. Specially from the side of descriptive statistics, visualization, natural language processing (NLP), and topic modeling (LSA, LDA, LSI).

Overview of Notebook

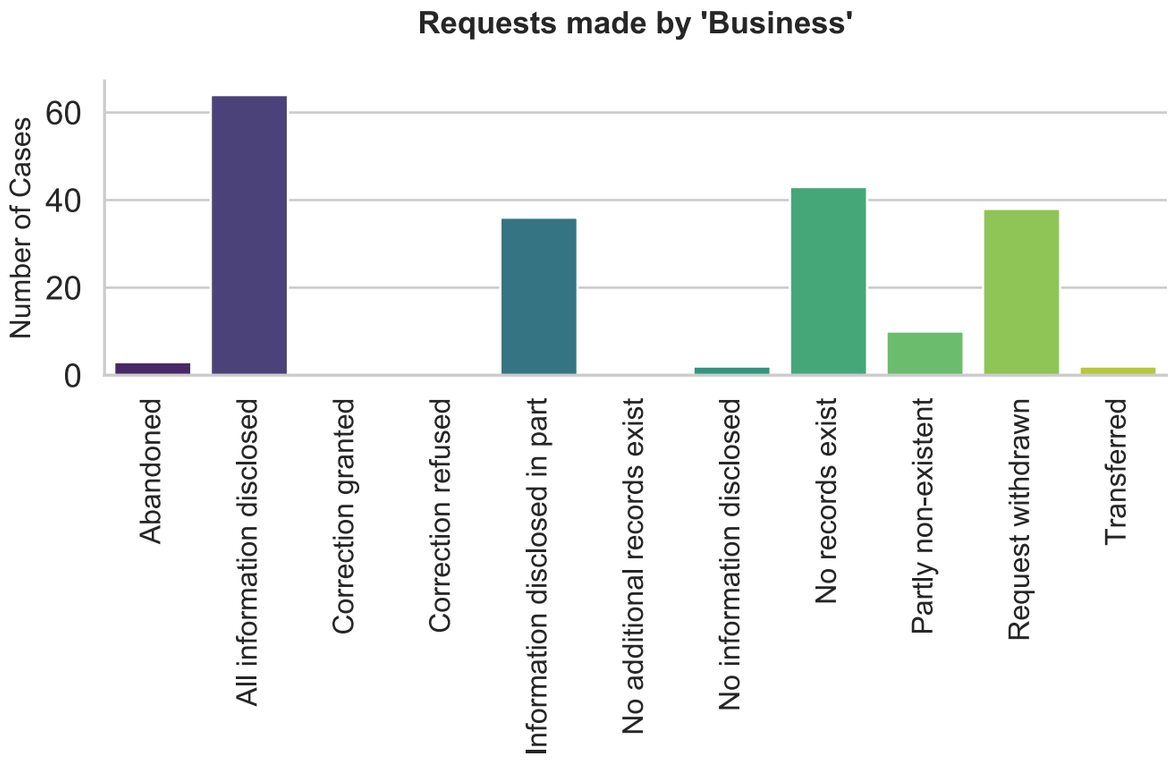
- Get to know the data
  - We look at the files, we merge them, and clean them if necessary.
  - We have five main columns: one is the request ID, three columns have categorical data, and one has the summary of the request itself (plain text).
  - Of those with categorical data, decision made, source (a.k.a., requester), and request type, can we somehow reduce the number of categories (spoiler alert: yes).
- Descriptive Analysis
  - There are about 11 types of decisions for six different types of sources.
    - How many requests per type of decision?
    - How are the decisions split based on the source? Are some sources more "lucky" than others? Spoiler alert: yes, though we don't know the reason.
    - How is each decision split among the sources?
    - For the main types of decisions (All information disclosed, Information disclosed in part, No information disclosed), how are they split among the sources?
- Natural Language Processing (NLP) - Analyzing the summary of request

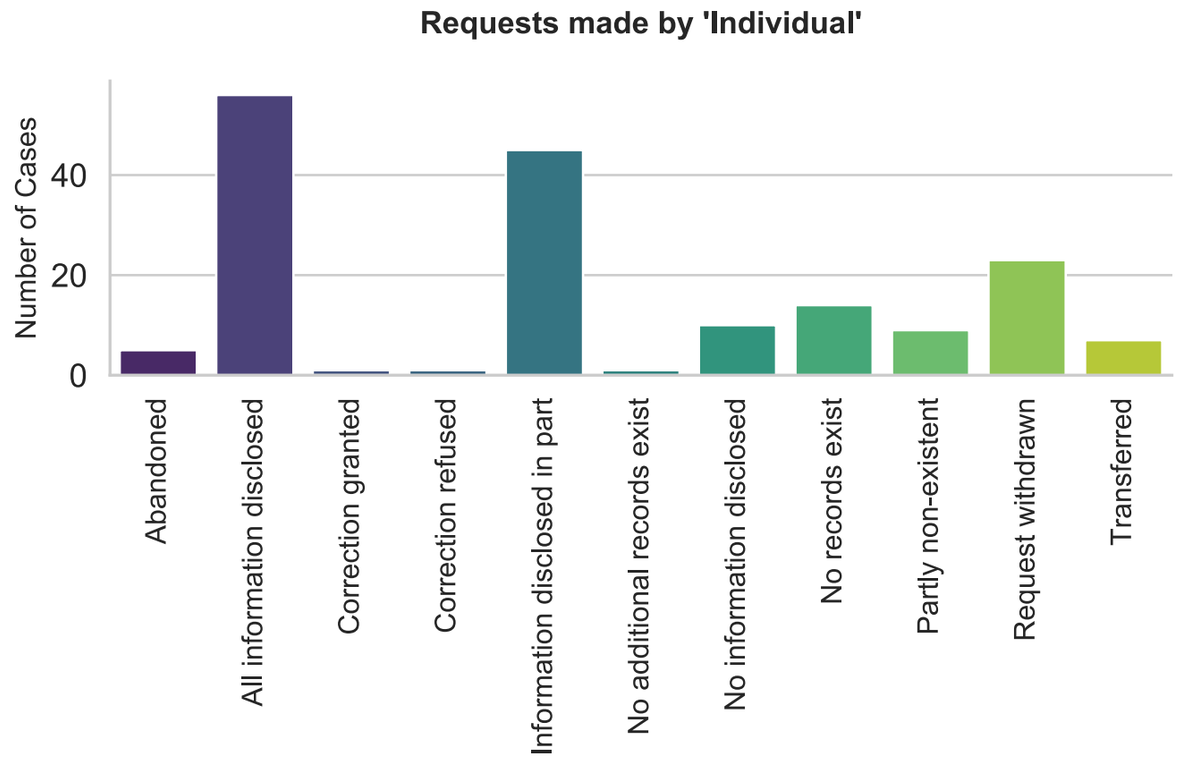
    - Before analyzing any text with NLP, one needs to go over some steps, which can vary depending on the goal:
        - Parse text, tokenize it, remove symbols, remove stopwords, remove punctuation, convert tokens to lowercase, remove short words, and lemmatize the tokens. We use both NLTK and spaCy.
    - n-grams. We take a look at the most frequent unigrams, bigrams, trigrams, and n-grams (4-5).
    - WordClouds. They are not only pretty, they are also useful. Why is a word/phrase so important!?
    - Let's take six of these frequent phrases and find if there is any correlation with the decision taken.
    - Summary of request statistics. With how much text are we working? How long are the requests before and after tokenization? (Spoiler alert: seven tokens is the median per request. Ouch!)
    - Part-of-Speech (POS) tagging of the requests. Nouns, verbs?
    - Topic Modeling
        - Here with do LSA and LDA Analysis using Bokeh, scikit-learn, and t-SNE.
        - We also try LDA Analysis using Gensim and pyLDAvis.
- Machine Learning (ML)
    - Here we compare the accuracies of eight classifiers, RandomForest, LinearSVC, MultinomialNB, LogisticRegression, SVC, KNeighbors, SGDC, and DecisionTree, using pipelines, GridSearchCV, confusion matrices, and classification reports.
    - We compare two vectorizers, Count vectorizer and tf-idf vectorizer.
    - Given that some of our decisions have less than 15 instances and that we also have an unbalanced case, we look at other ways to optimize this. For example, a) we keep decisions with over 15 instances, b) we merge our 11 type of decisions into three main bins (full, partial, or no info released), and c) we remove cases where no decision was made (withdrawn or abandoned, which we name it as the independent case).
    - And with all of this, our best score goes up to... 51%

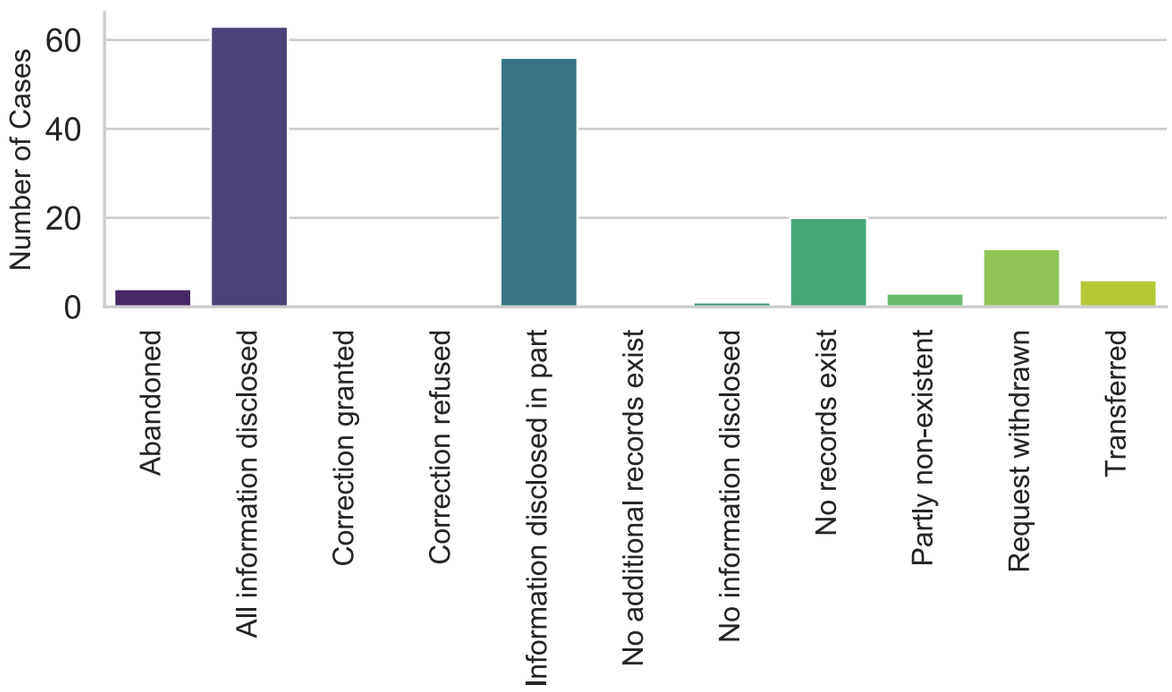

Decisions Made for all Requests
Number of Cases
0
25
50
75
100
125
150
175
Abandoned
All information disclosed
Correction granted
Correction refused
Information disclosed in part
No additional records exist
No information disclosed
No records exist
Partly non-existent
Request withdrawn
Transferred

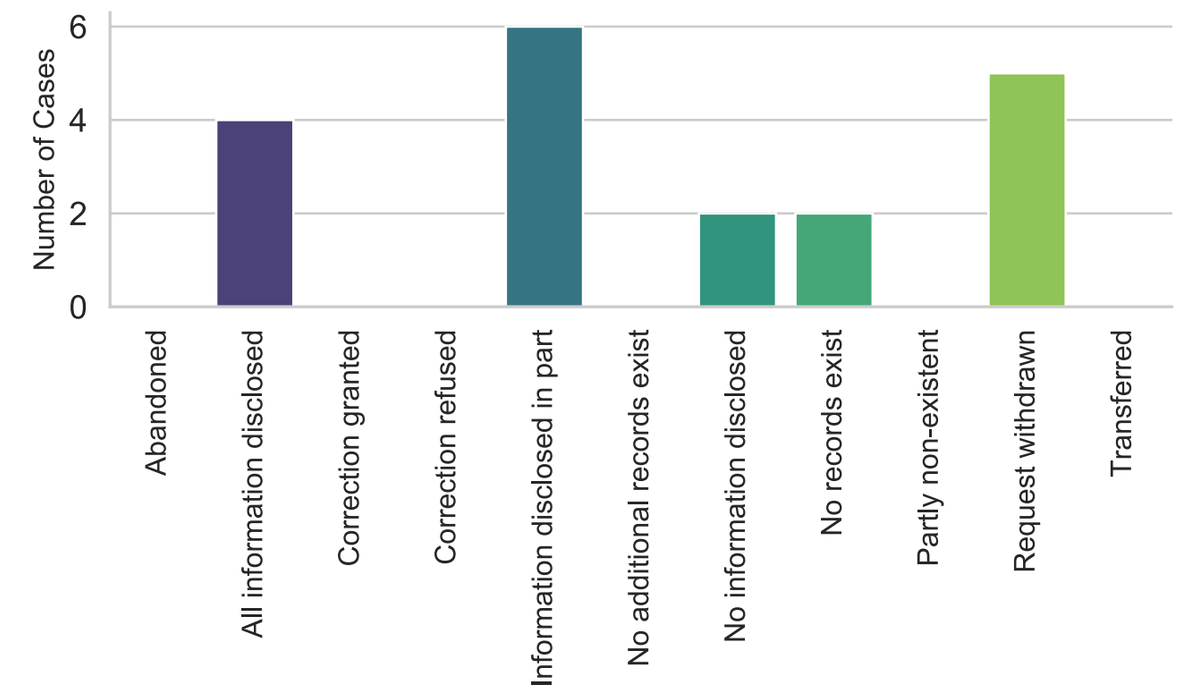

Full Data
Source
Business
Individual
Individual by Agent
Business by Agent
Media
Individual for dependant
3 64 0 0 36 0 2 43 10 38 2
5 56 1 1 45 1 10 14 9 23 7
4 63 0 0 56 0 1 20 3 13 6
0 4 0 0 6 0 2 2 0 5 0
0 2 0 0 9 0 6 1 1 1 0
1 0 0 0 0 0 0 0 0 0 0
Abandoned
All information disclosed
Correction granted
Correction refused
Information disclosed in part
No additional records exist
No information disclosed
No records exist
Partly non-existent
Request withdrawn
Transferred
60
45
30
15
0
Decision

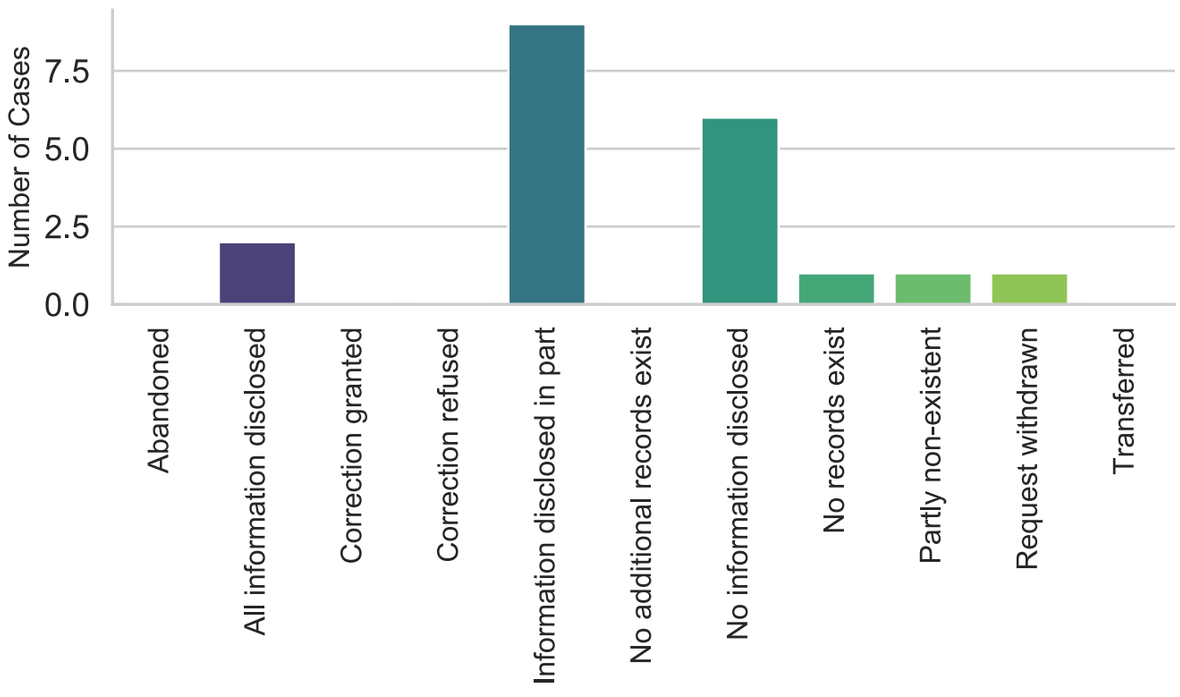
Number of cases for all type of decisions made for each of the sources

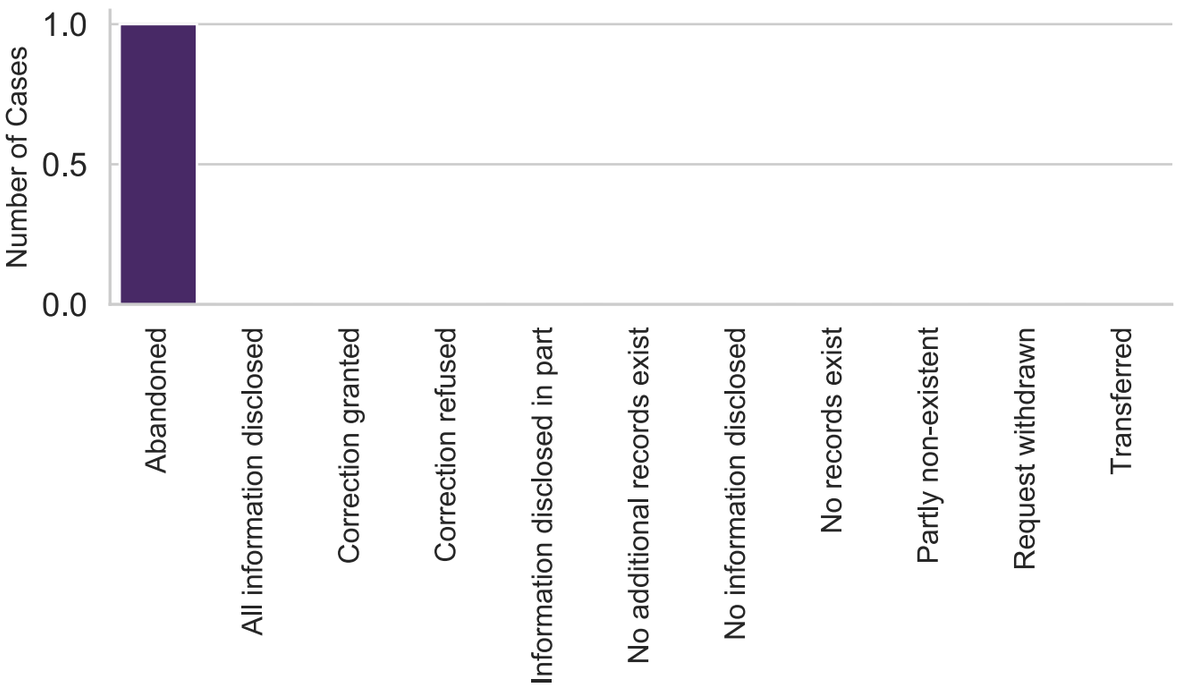




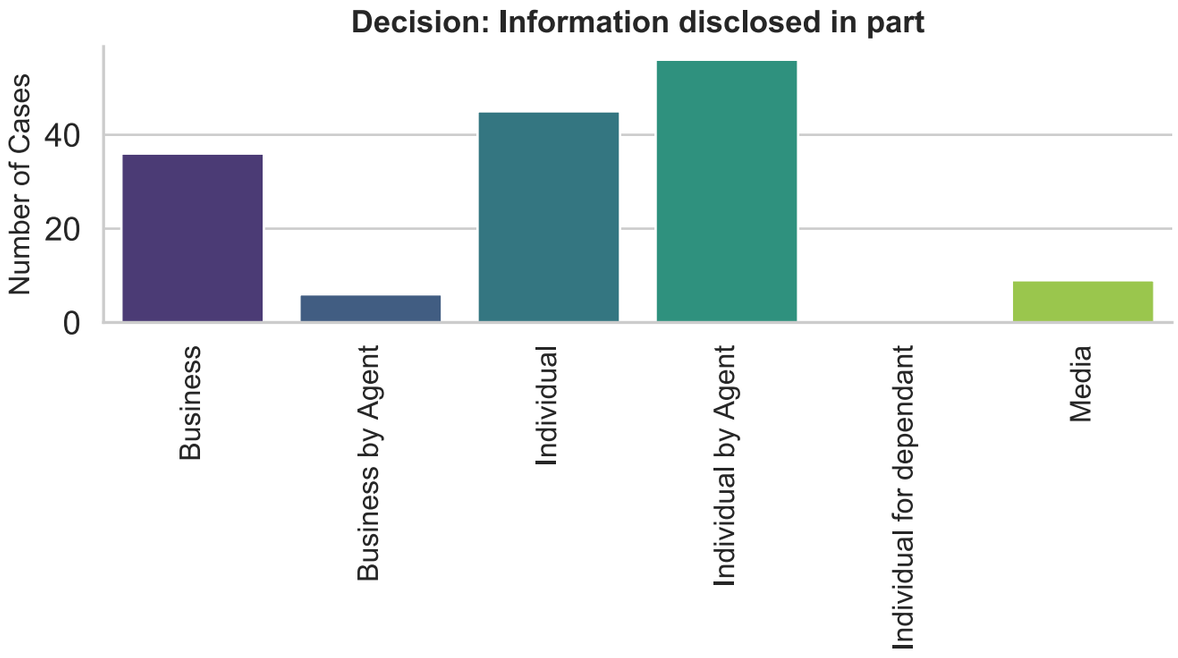


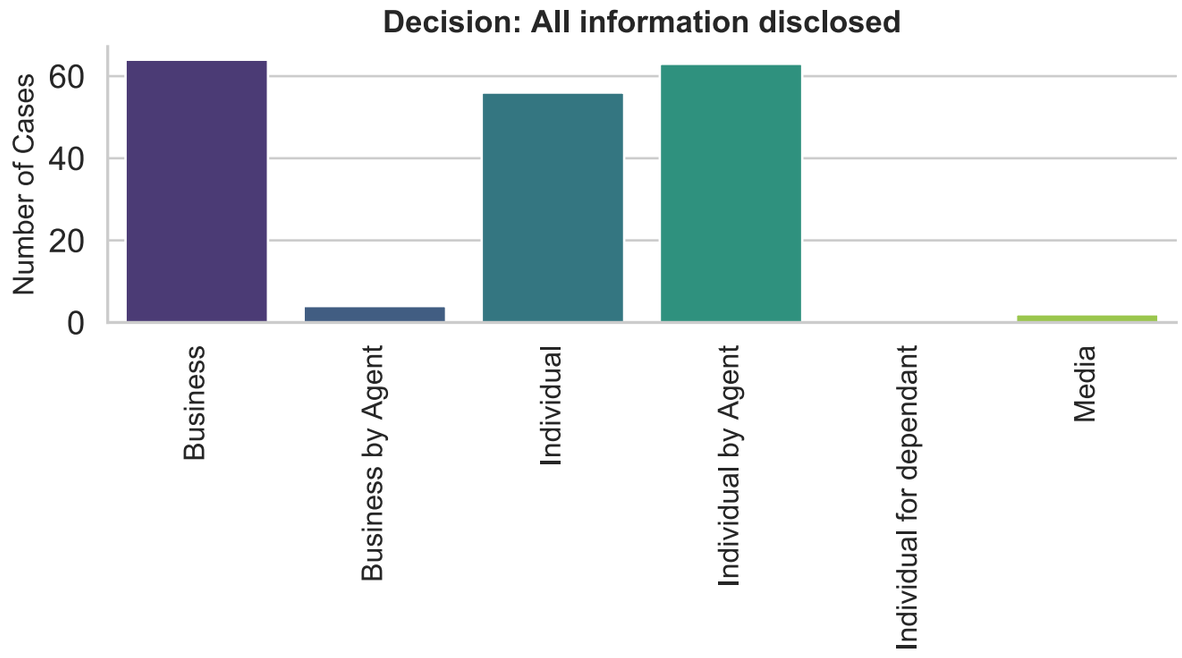


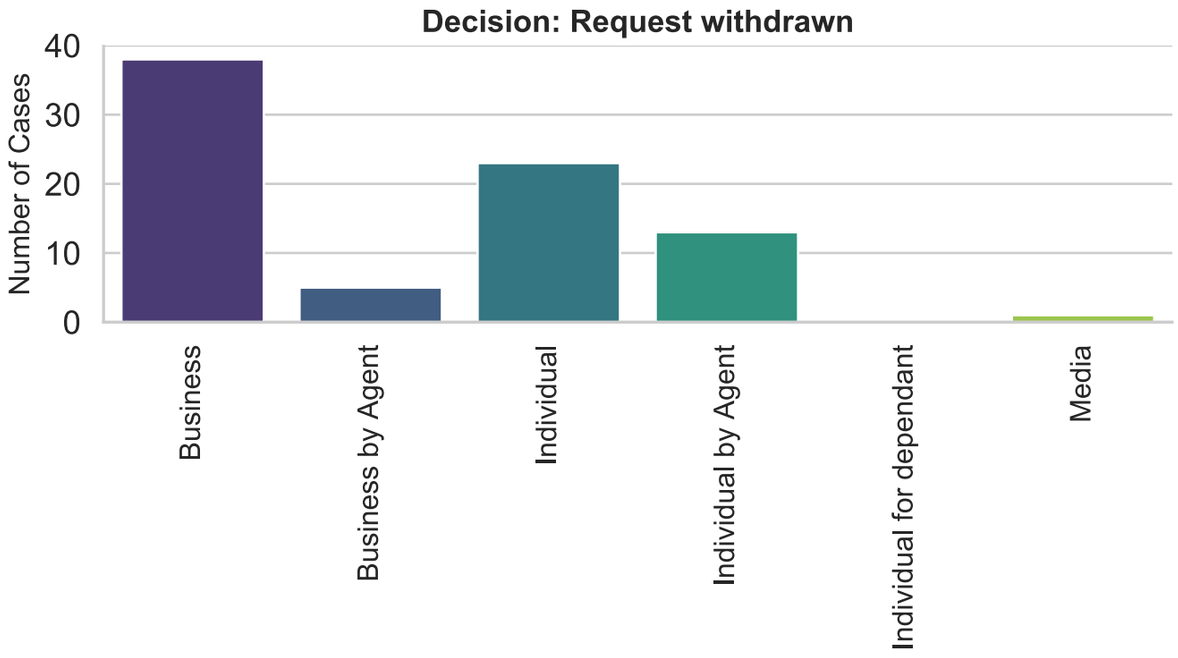


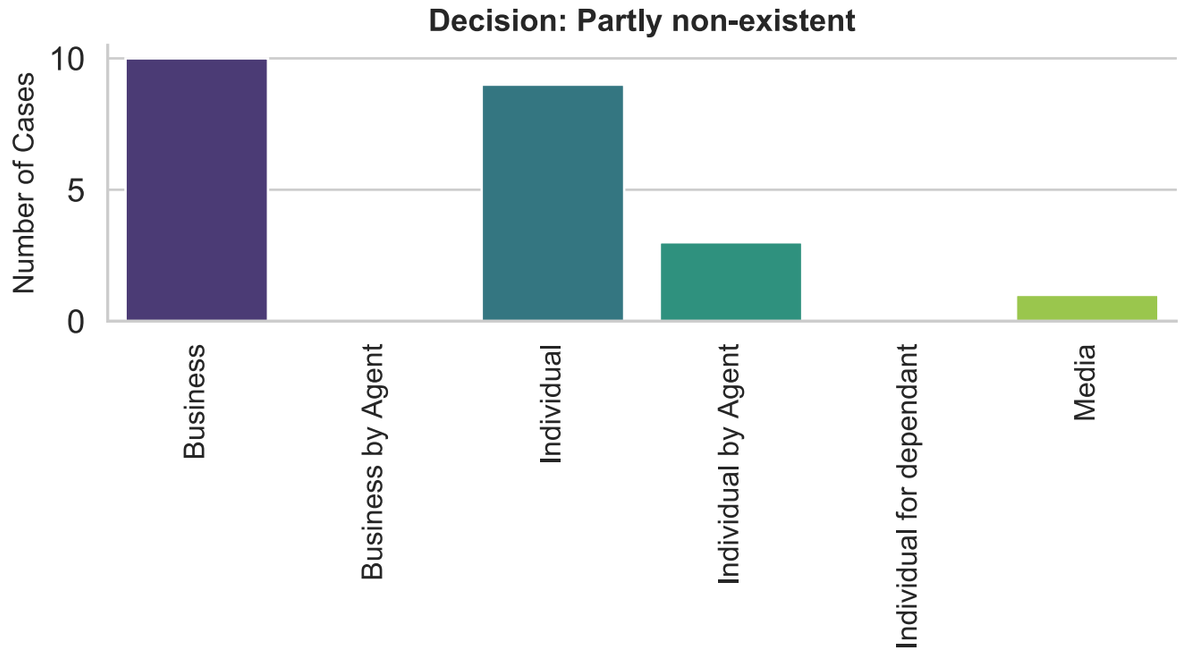


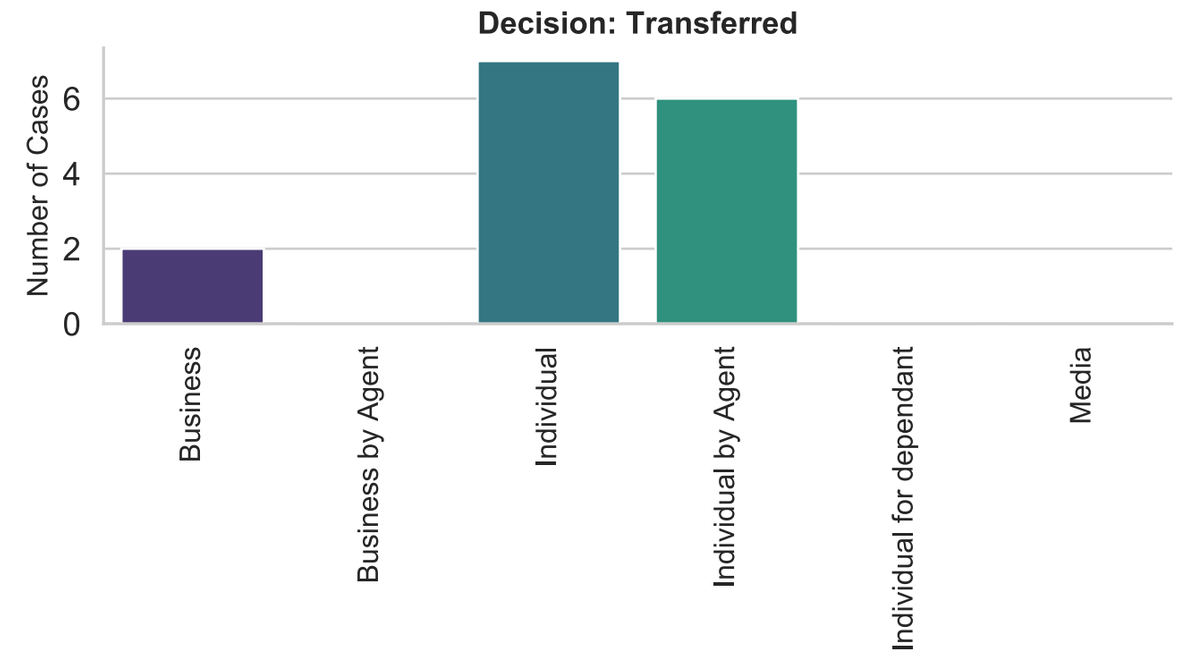
Number of cases for each type of decision made by sources

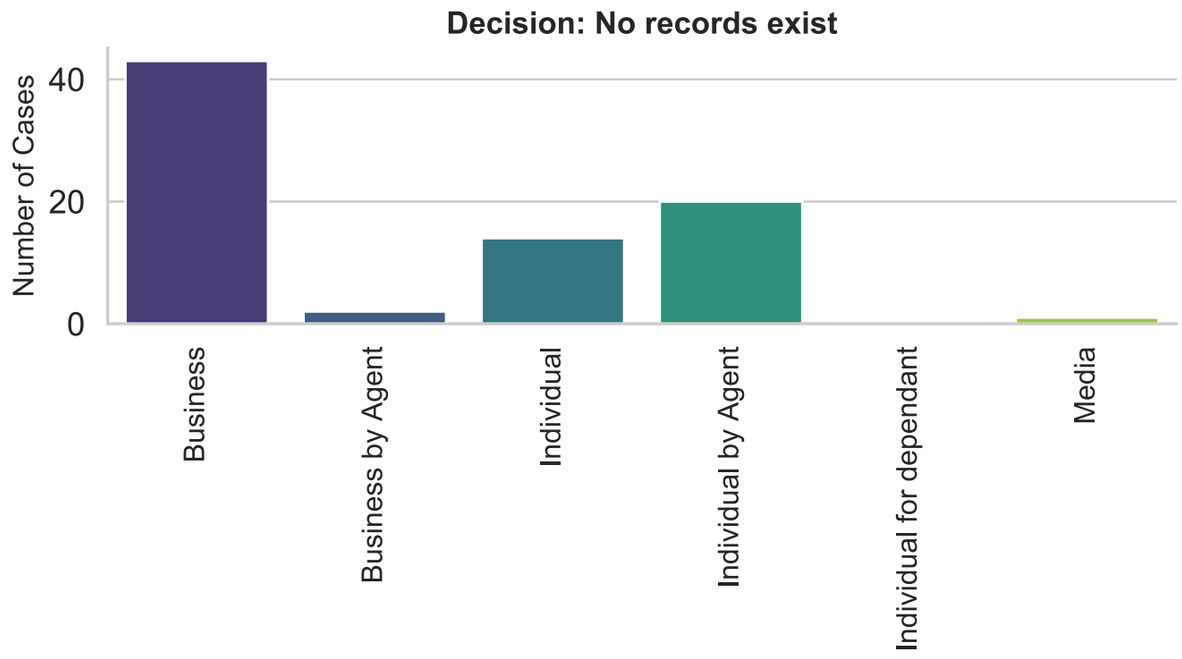


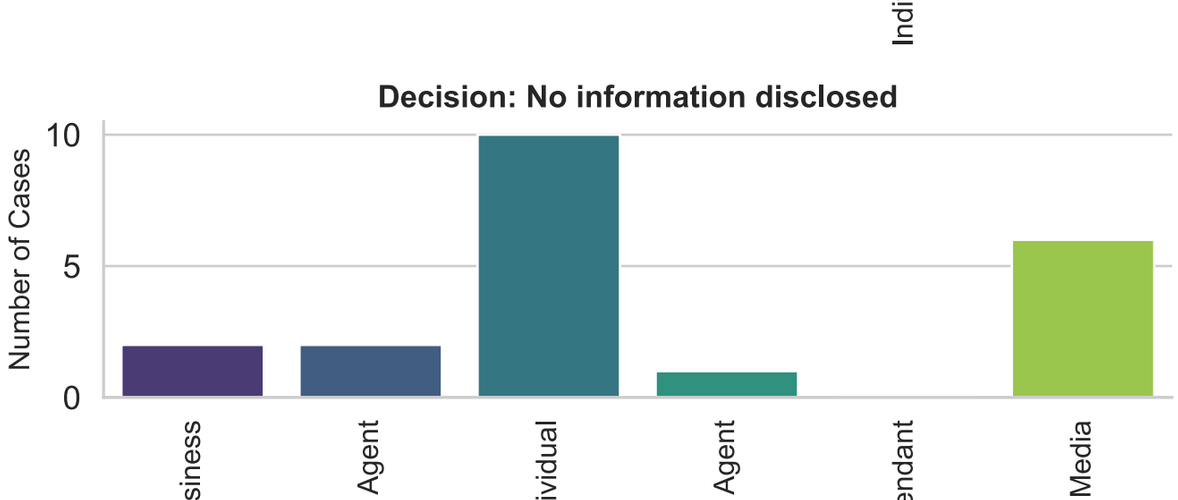


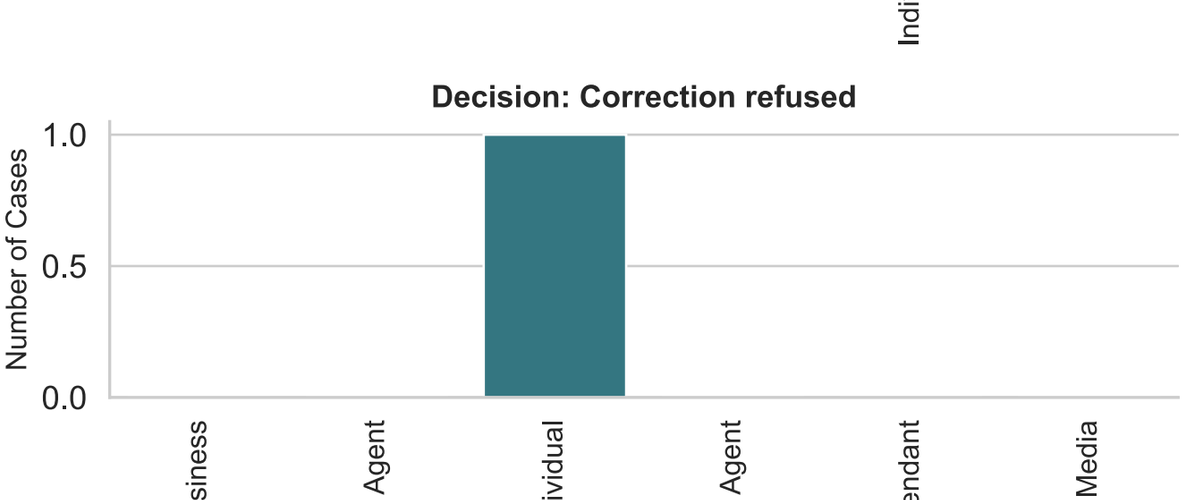


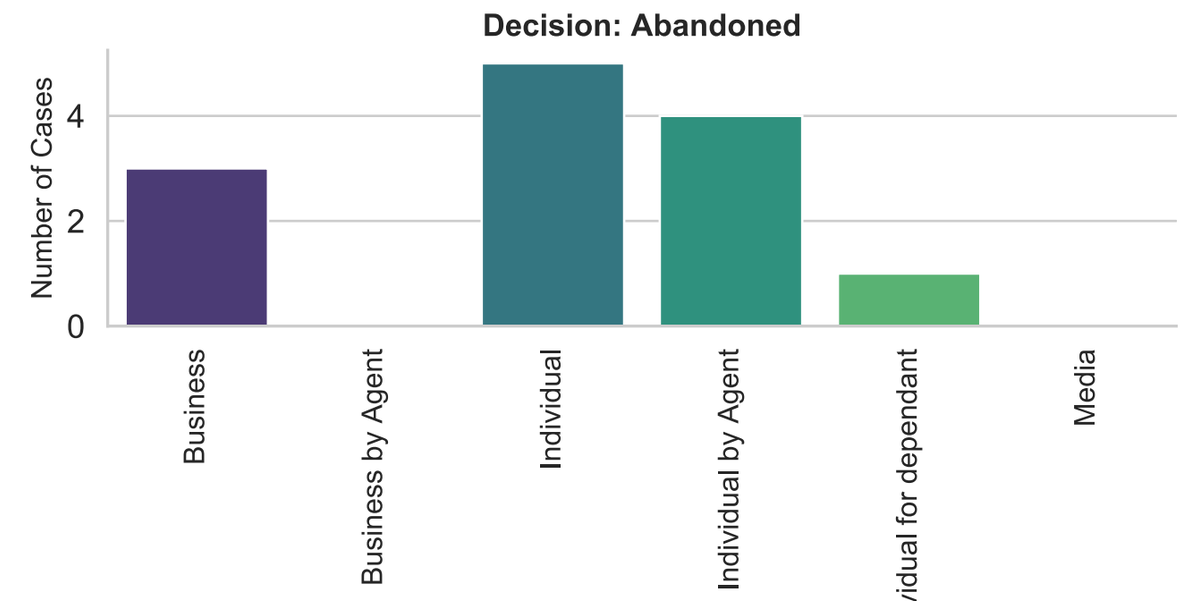


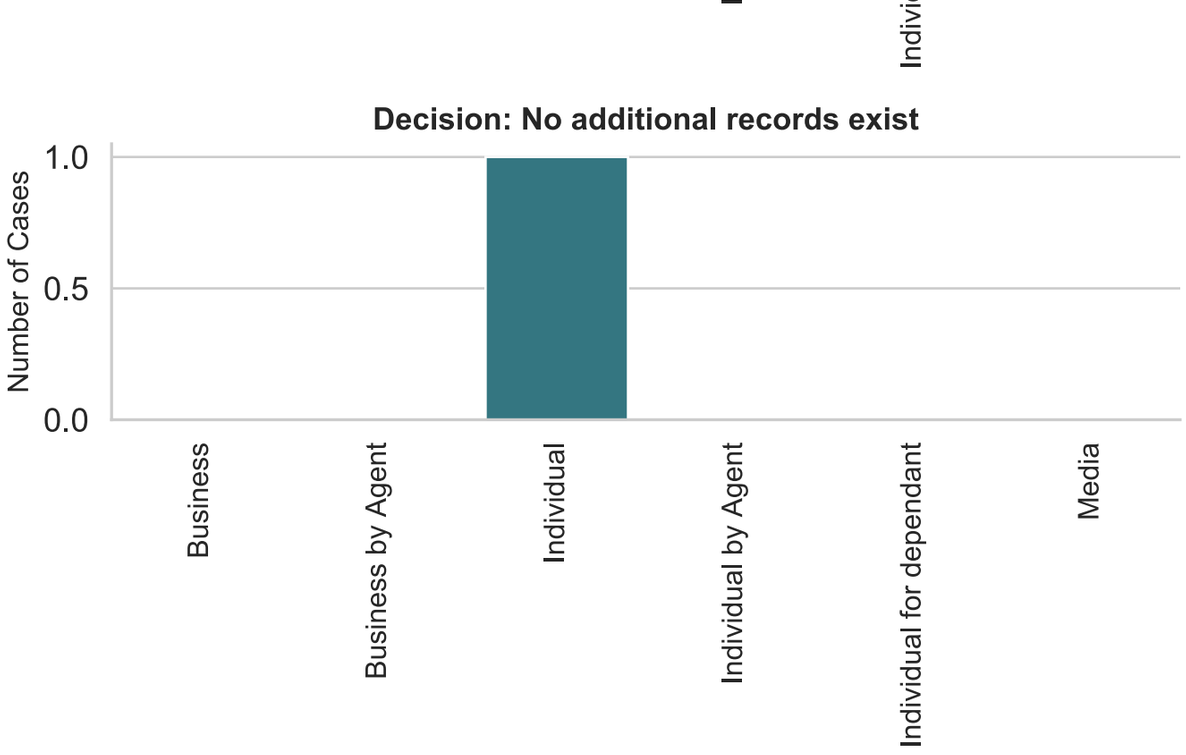


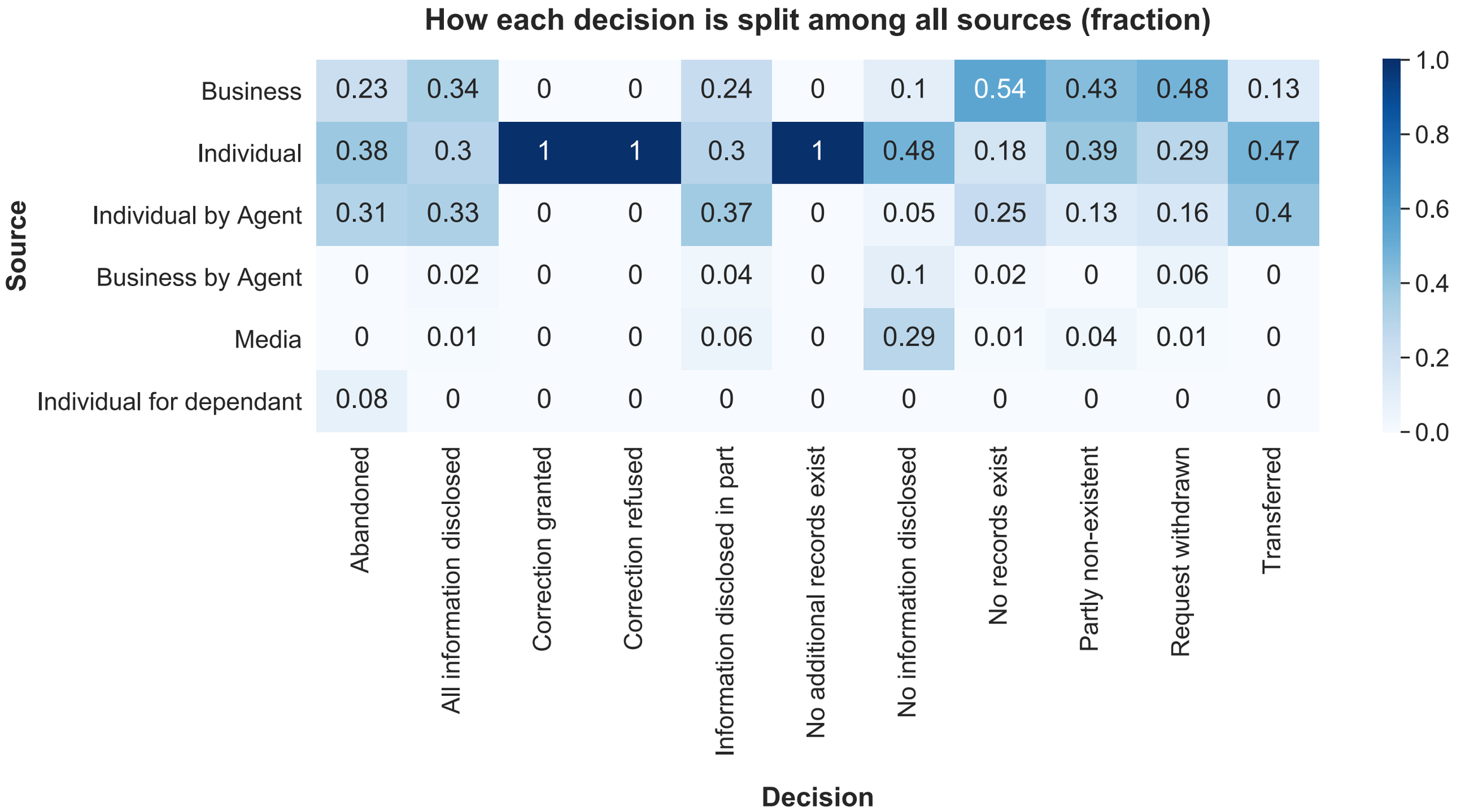


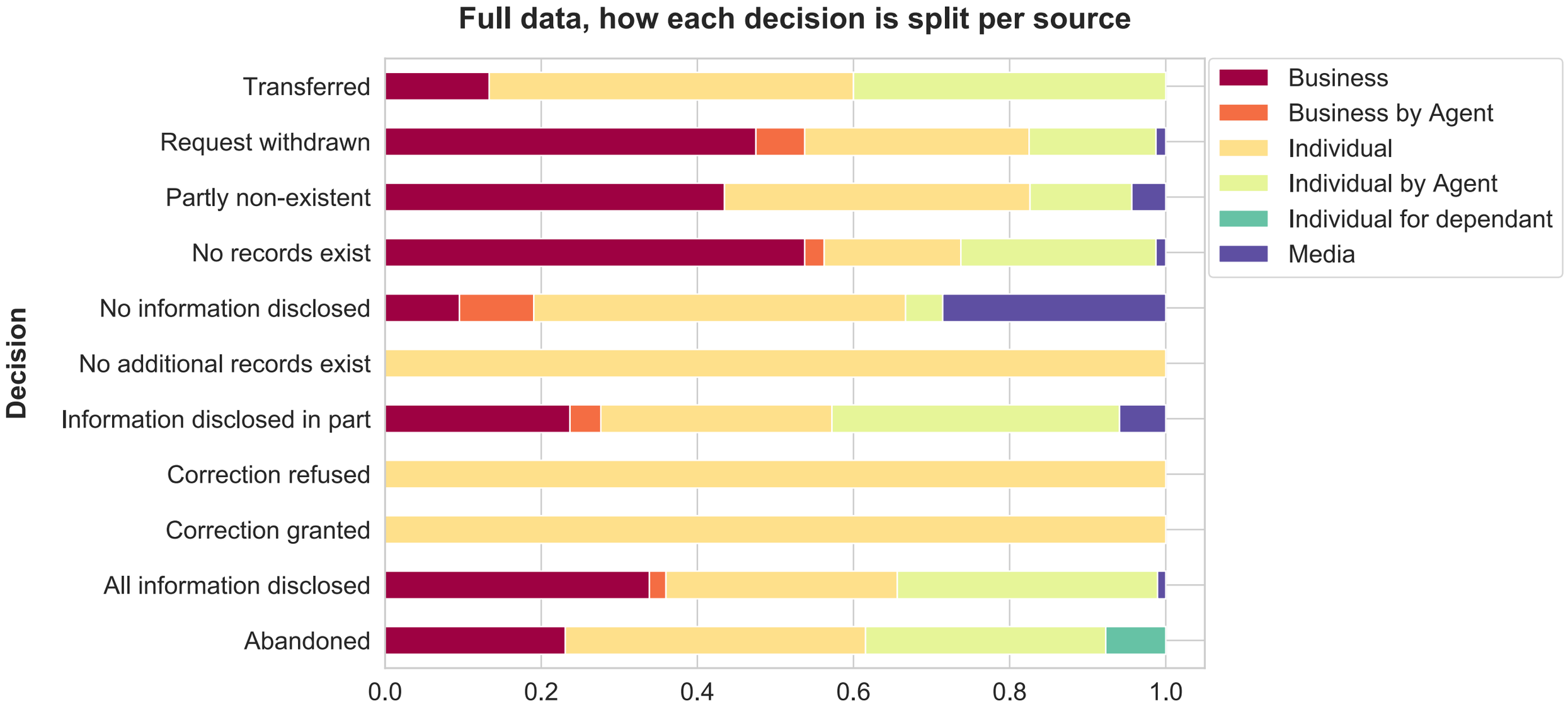


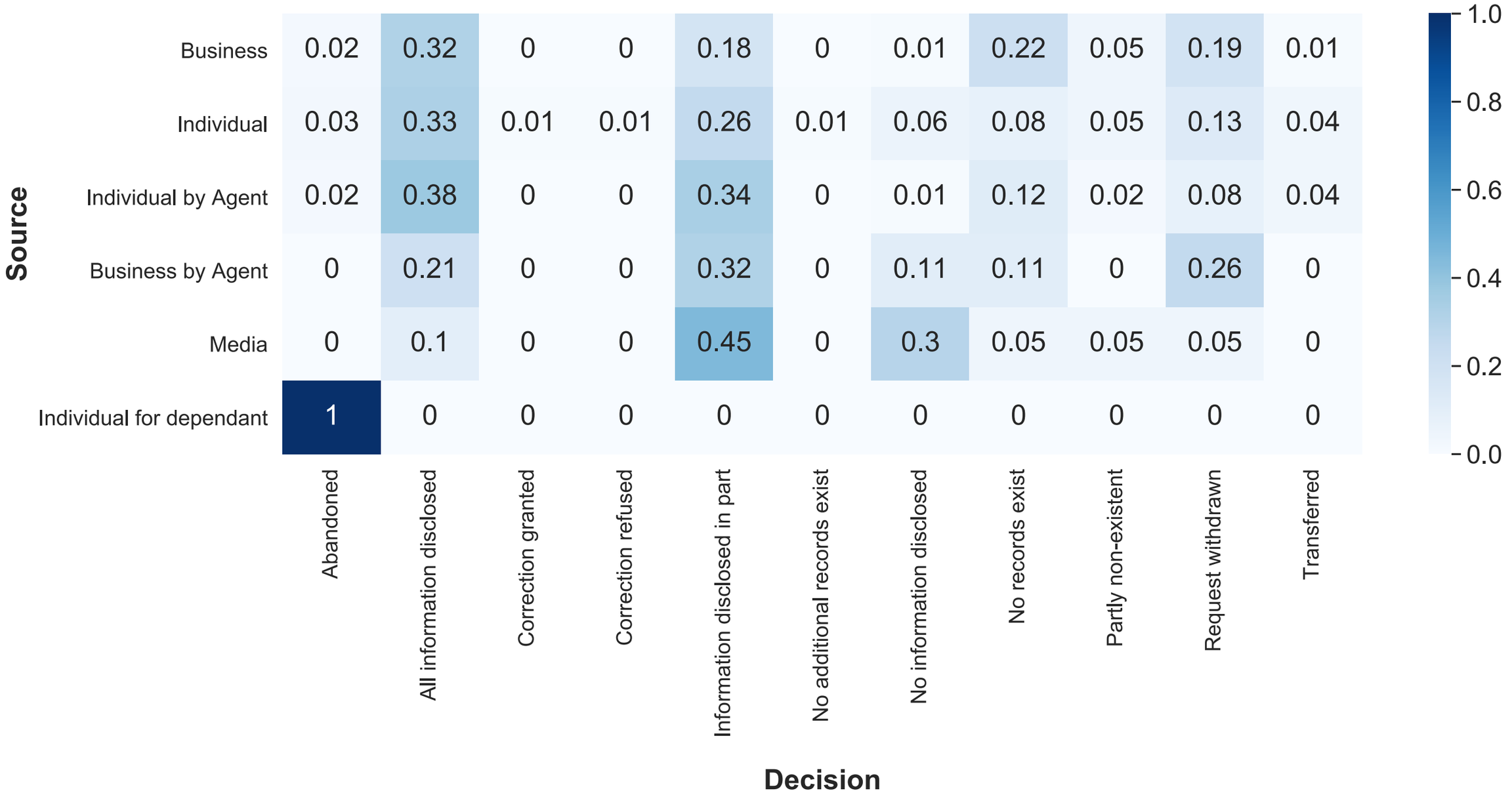


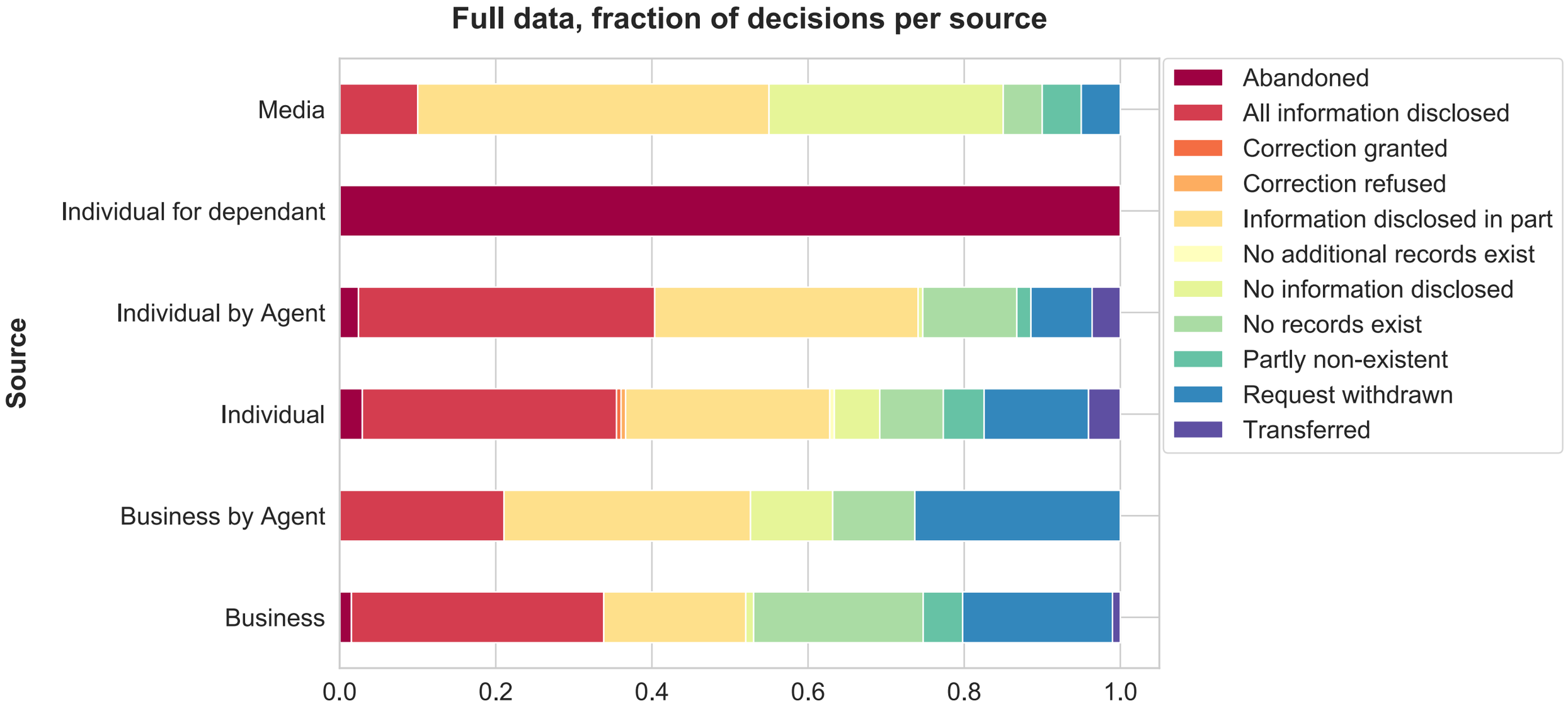


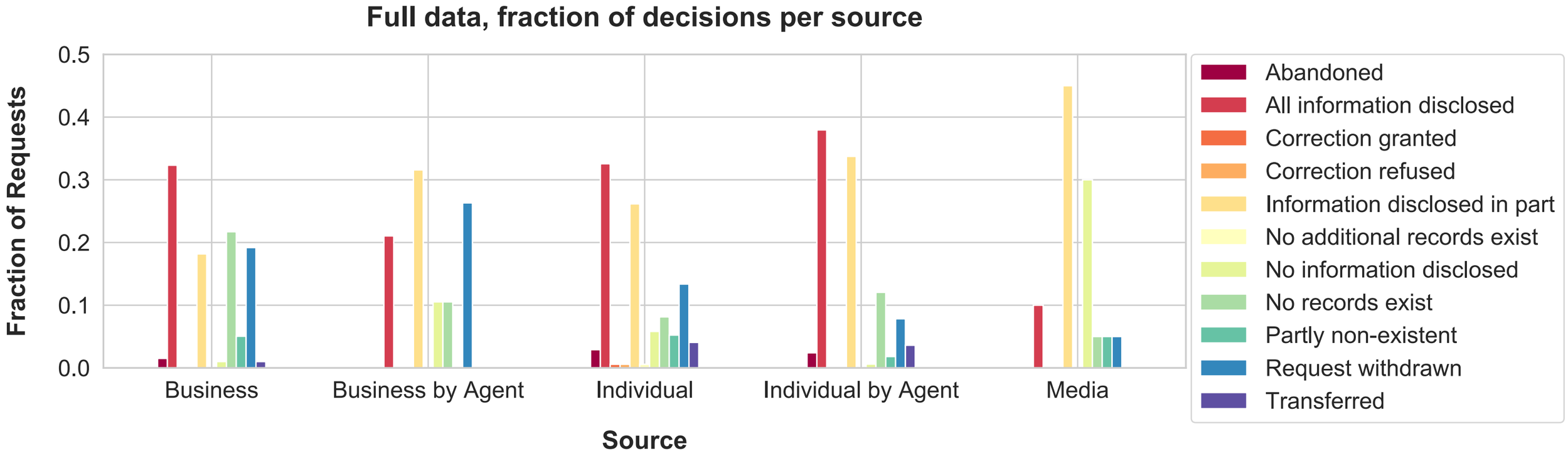


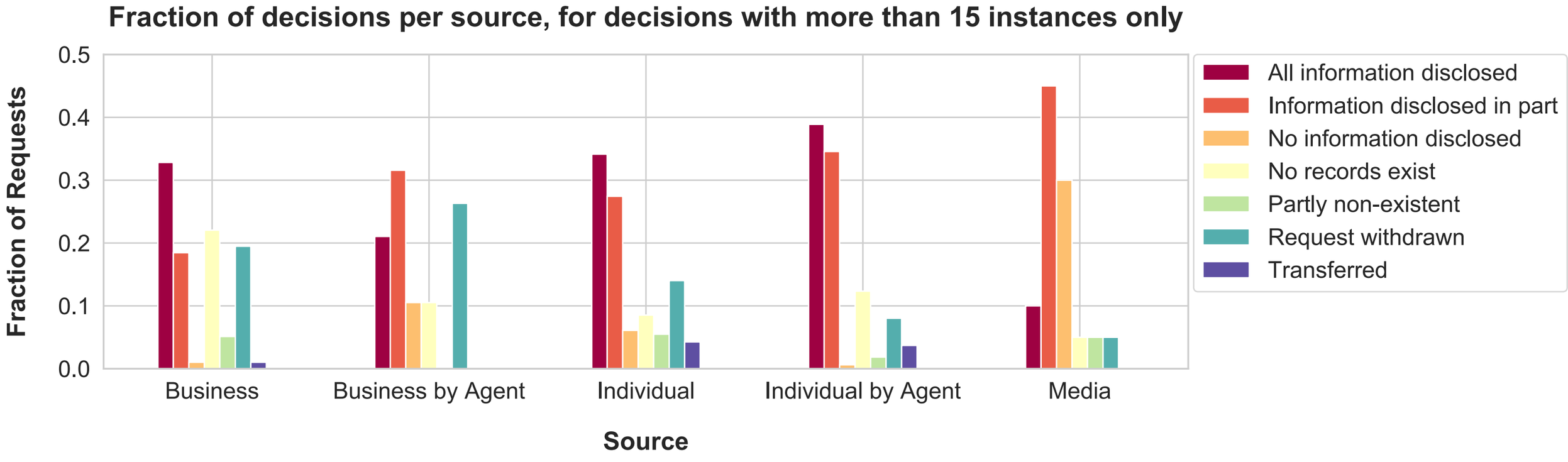

How each decision is split among all sources (fraction)
Source
Business
Individual
Individual by Agent
Business by Agent
Media
Individual for dependant
0.23 0.34 0 0 0.24 0 0.1 0.54 0.43 0.48 0.13
0.38 0.3 1 1 0.3 1 0.48 0.18 0.39 0.29 0.47
0.31 0.33 0 0 0.37 0 0.05 0.25 0.13 0.16 0.4
0 0.02 0 0 0.04 0 0.1 0.02 0 0.06 0
0 0.01 0 0 0.06 0 0.29 0.01 0.04 0.01 0
0.08 0 0 0 0 0 0 0 0 0 0
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
Decision
1.0
0.8
0.6
0.4
0.2
0.0

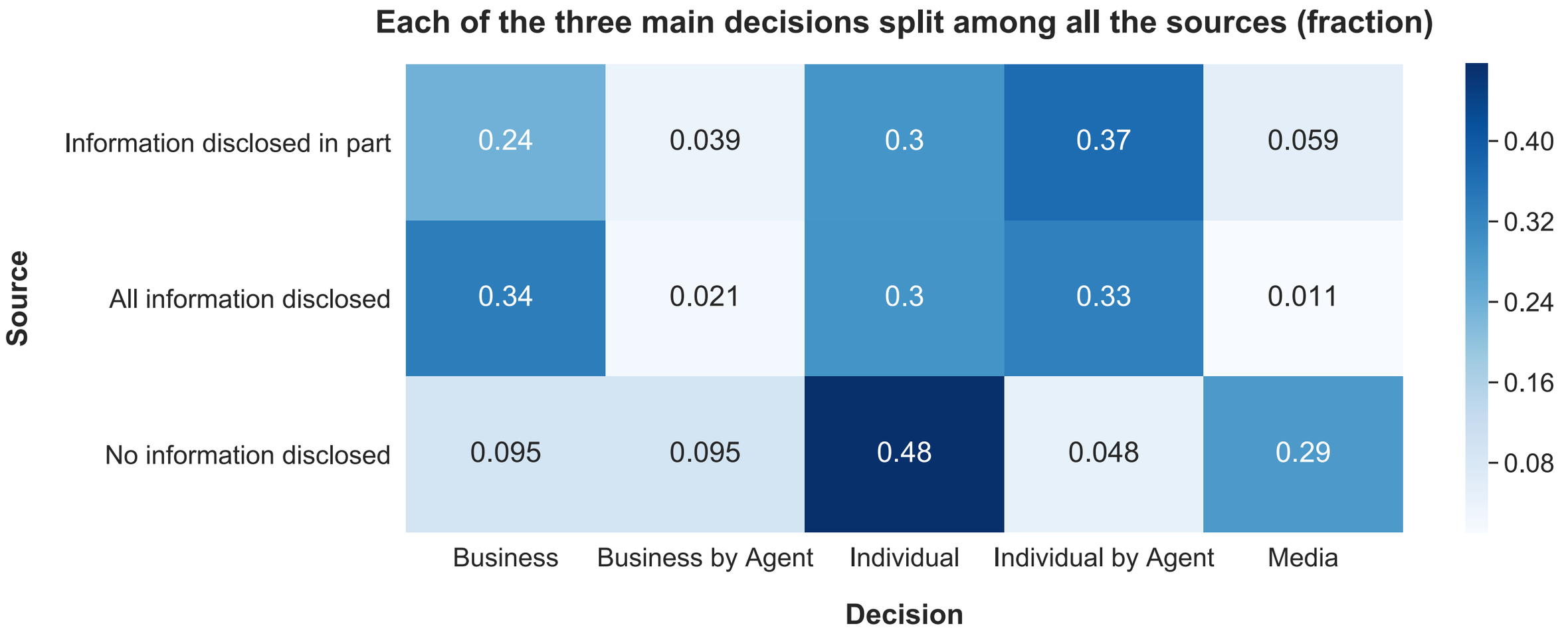

Full data, how each decision is split per source
Decision
Transferred
Request withdrawn
Partly non-existent
No records exist
No information disclosed
No additional records exist
Information disclosed in part
Correction refused
Correction granted
All information disclosed
Abandoned
0.0
0.2
0.4
0.6
0.8
1.0
Business
Business by Agent
Individual
Individual by Agent
Individual for dependant
Media

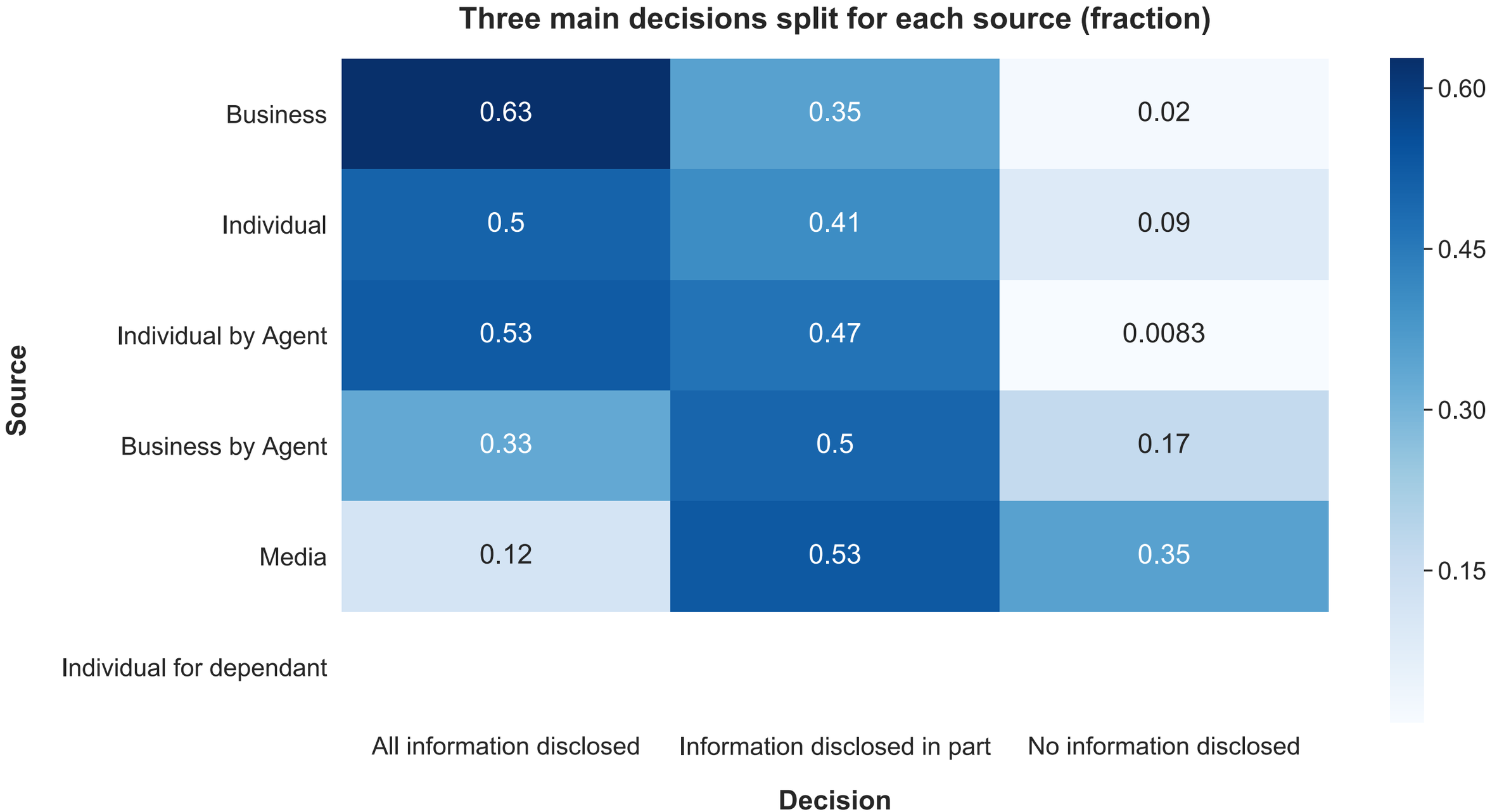

How decisions are split per source (fraction)
Source
Business
Individual
Individual by Agent
Business by Agent
Media
Individual for dependant
0.02 0.32 0 0 0.18 0 0.01 0.22 0.05 0.19 0.01
0.03 0.33 0.01 0.01 0.26 0.01 0.06 0.08 0.05 0.13 0.04
0.02 0.38 0 0 0.34 0 0.01 0.12 0.02 0.08 0.04
0 0.21 0 0 0.32 0 0.11 0.11 0 0.26 0
0 0.1 0 0 0.45 0 0.3 0.05 0.05 0.05 0
1 0 0 0 0 0 0 0 0 0 0
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
Decision
1.0
0.8
0.6
0.4
0.2
0.0

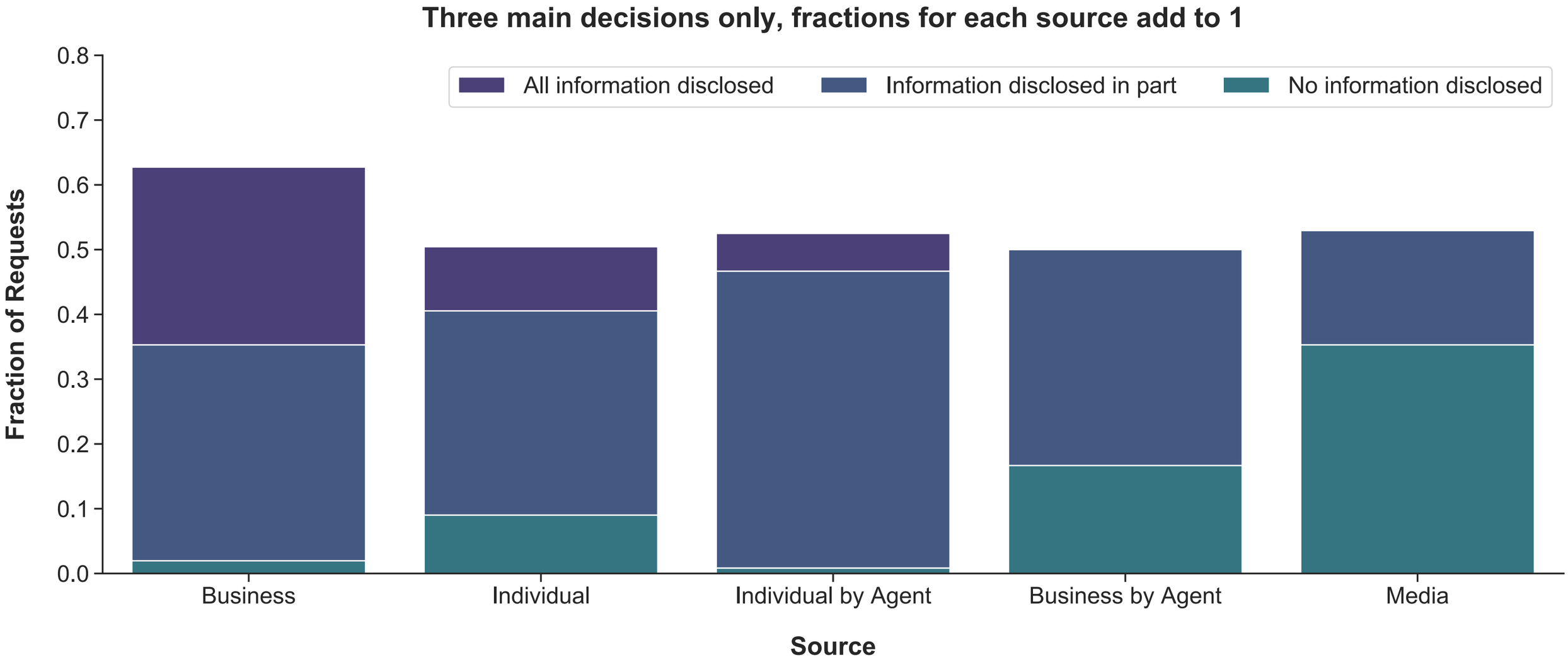

Full data, fraction of decisions per source
Source
Media
Individual for dependant
Individual by Agent
Individual
Business by Agent
Business
0.0
0.2
0.4
0.6
0.8
1.0
Abandoned
All information disclosed
Correction granted
Correction refused
Information disclosed in part
No additional records exist
No information disclosed
No records exist
Partly non-existent
Request withdrawn
Transferred

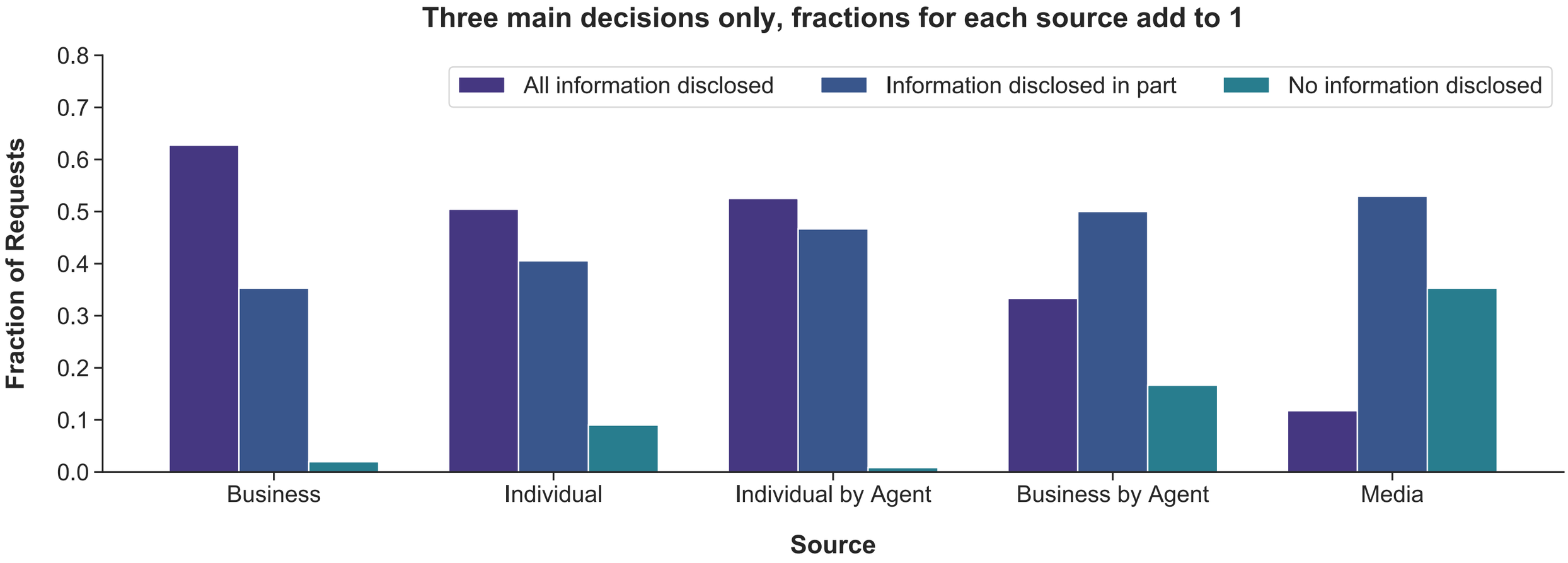

Full data, fraction of decisions per source
Fraction of Requests
0.5
0.4
0.3
0.2
0.1
0.0
Business
Business by Agent
Individual
Individual by Agent
Media
Source
Abandoned
All information disclosed
Correction granted
Correction refused
Information disclosed in part
No additional records exist
No information disclosed
No records exist
Partly non-existent
Request withdrawn
Transferred

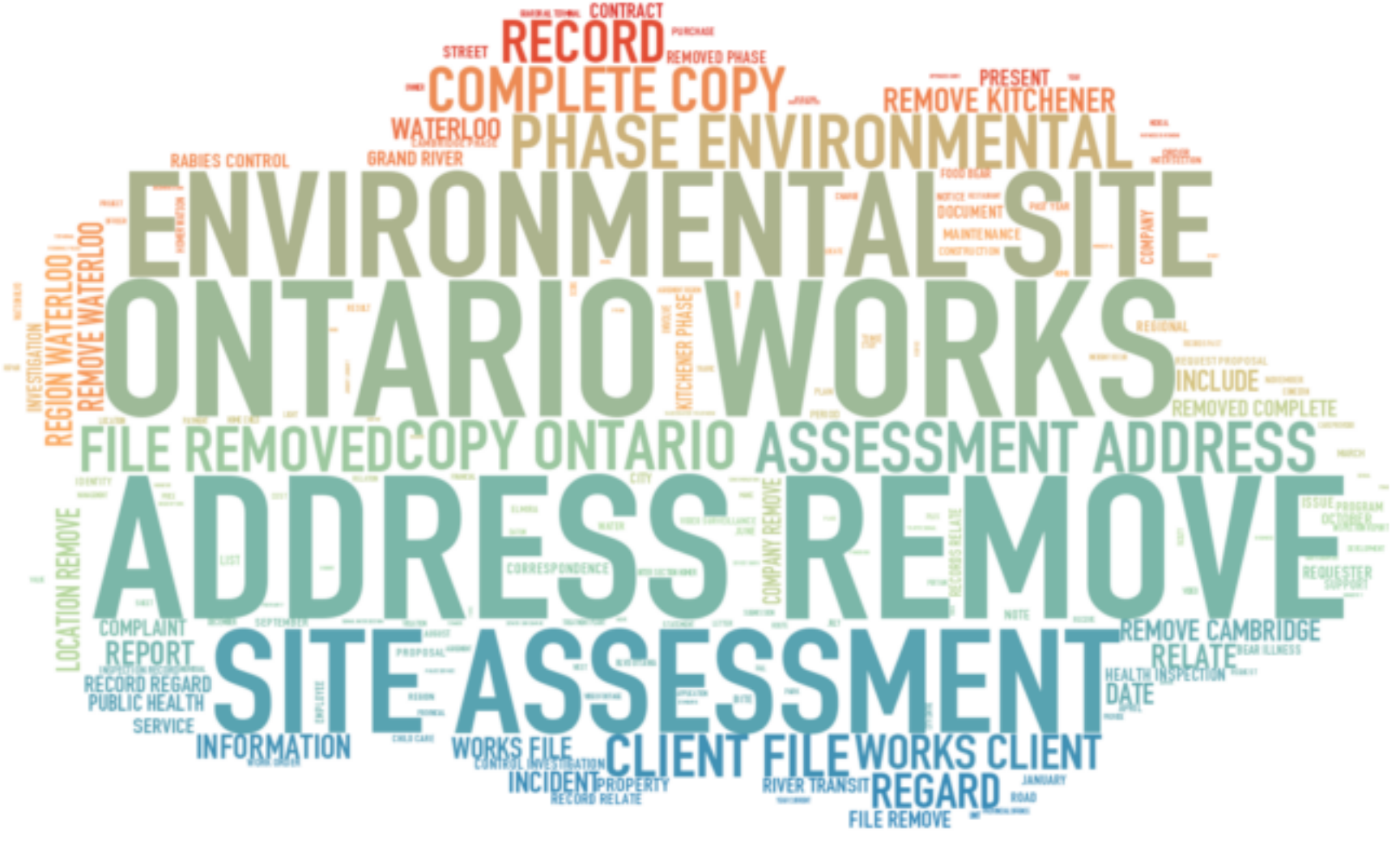

Fraction of decisions per source, for decisions with more than 15 instances only
Fraction of Requests
0.5
0.4
0.3
0.2
0.1
0.0
Business
Business by Agent
Individual
Individual by Agent
Media
Source
All information disclosed
Information disclosed in part
No information disclosed
No records exist
Partly non-existent
Request withdrawn
Transferred


Each of the three main decisions split among all the sources (fraction)
Source
Information disclosed in part
All information disclosed
No information disclosed
0.24
0.039
0.3
0.37
0.059
0.34
0.021
0.3
0.33
0.011
0.095
0.095
0.48
0.048
0.29
Business
Business by Agent
Individual
Individual by Agent
Media
Decision
0.40
0.32
0.24
0.16
0.08


Three main decisions split for each source (fraction)
Business
Individual
Individual by Agent
Business by Agent
Media
Individual for dependant
0.63
0.35
0.02
0.5
0.41
0.09
0.53
0.47
0.0083
0.33
0.5
0.17
0.12
0.53
0.35
0.60
0.45
0.30
0.15
All information disclosed
Information disclosed in part
No information disclosed
Source
Decision

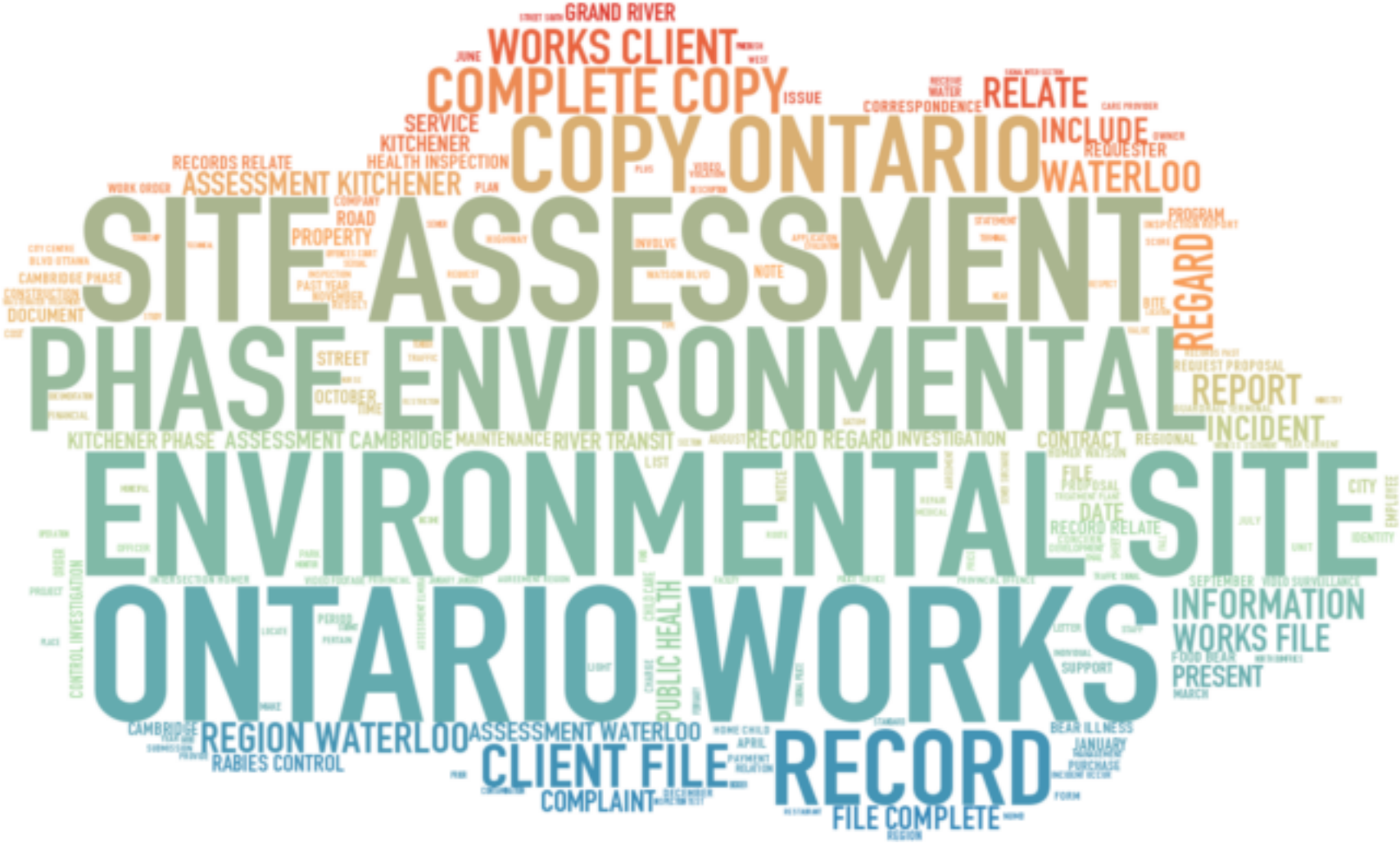

Three main decisions only, fractions for each source add to 1
All information disclosed
Information disclosed in part
No information disclosed
Fraction of Requests
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Business
Individual
Individual by Agent
Business by Agent
Media
Source


Three main decisions only, fractions for each source add to 1
All information disclosed
Information disclosed in part
No information disclosed
Fraction of Requests
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Business
Individual
Individual by Agent
Business by Agent
Media
Source



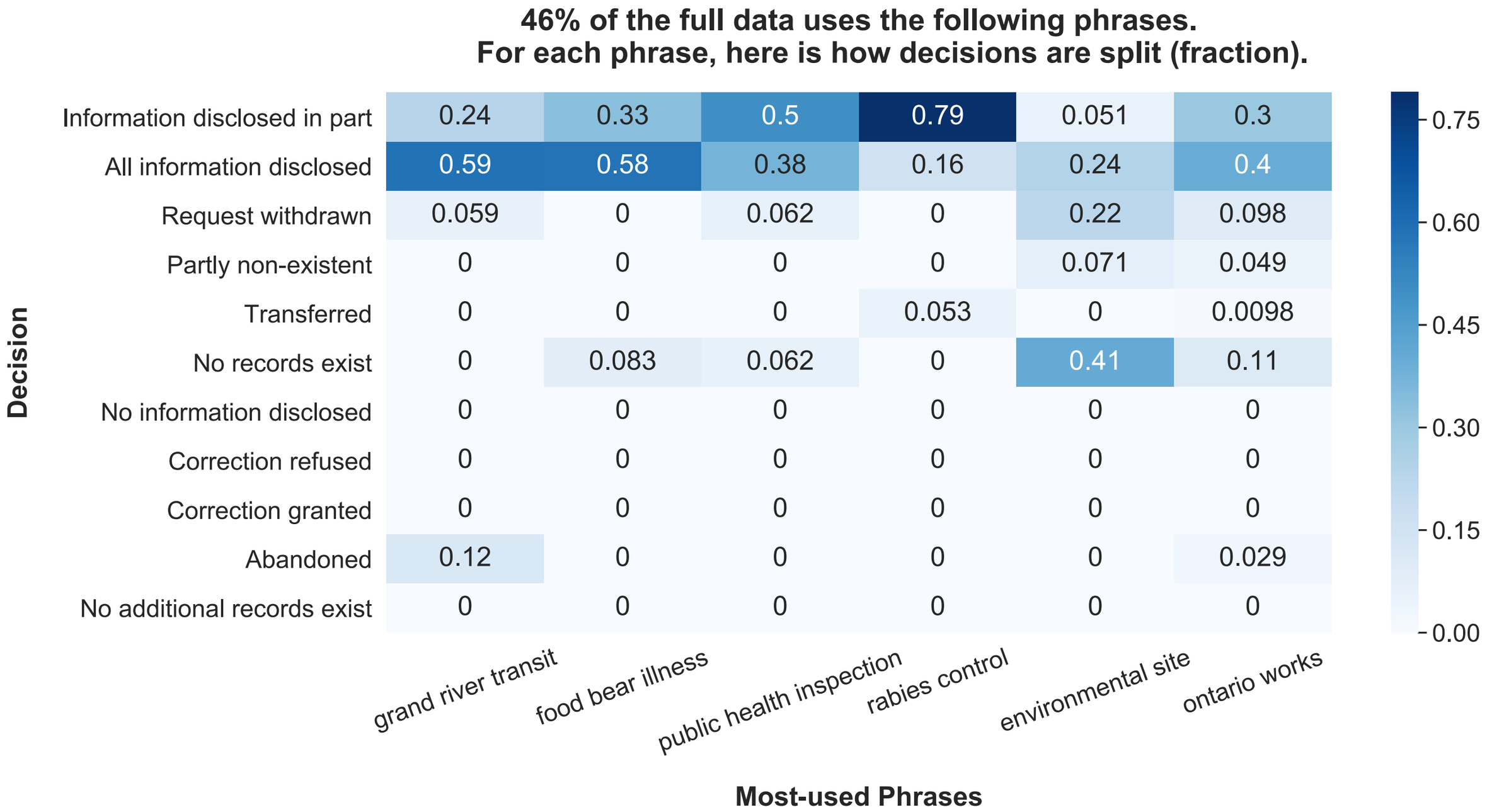
Top 200 unigrams/bigrams, full text

Top 200 unigrams, full text



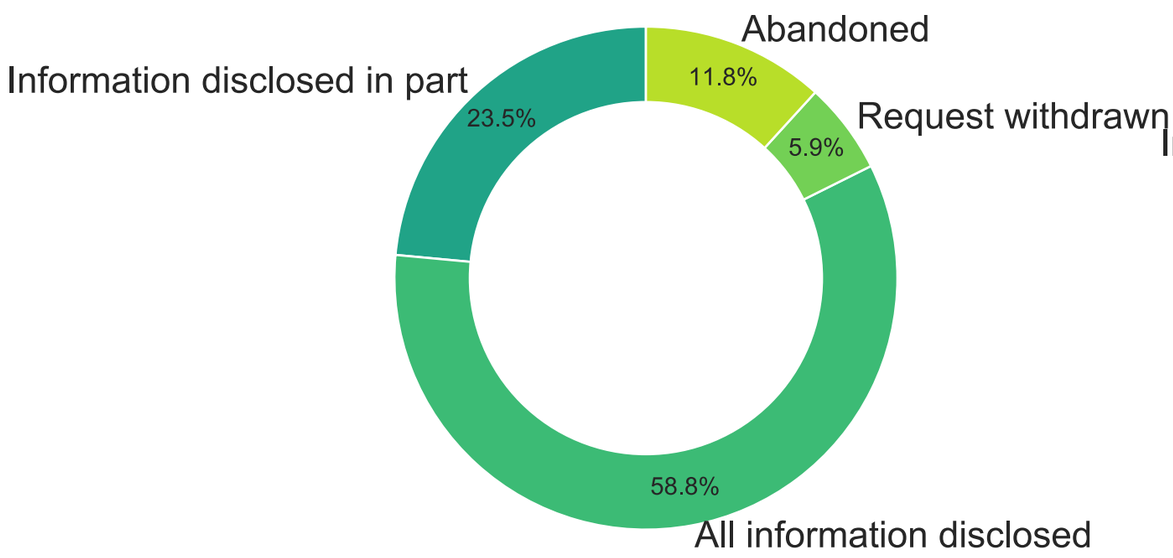
Top 200 unigrams/bigrams, full text without '{* remove}'

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

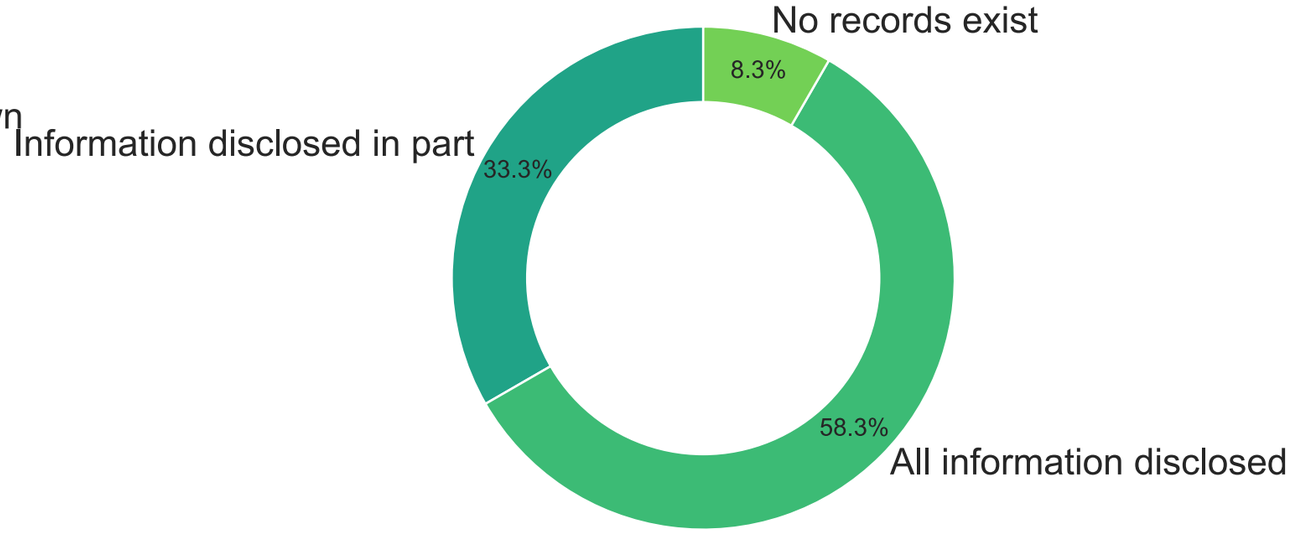

46% of the full data uses the following phrases.
For each phrase, here is how decisions are split (fraction).
Decision
Information disclosed in part
All information disclosed
Request withdrawn
Partly non-existent
Transferred
No records exist
No information disclosed
Correction refused
Correction granted
Abandoned
No additional records exist
0.24 0.33 0.5 0.79 0.051 0.3
0.59 0.58 0.38 0.16 0.24 0.4
0.059 0 0.062 0 0.22 0.098
0 0 0 0 0.071 0.049
0 0 0 0.053 0 0.0098
0 0.083 0.062 0 0.41 0.11
0 0 0 0 0 0
0 0 0 0 0 0
0 0 0 0 0 0
0.12 0 0 0 0 0.029
0 0 0 0 0 0
0.75
0.60
0.45
0.30
0.15
0.00
grand river transit
food bear illness
public health inspection
rabies control
environmental site
ontario works
Most-used Phrases

Decision percentage for each n-gram

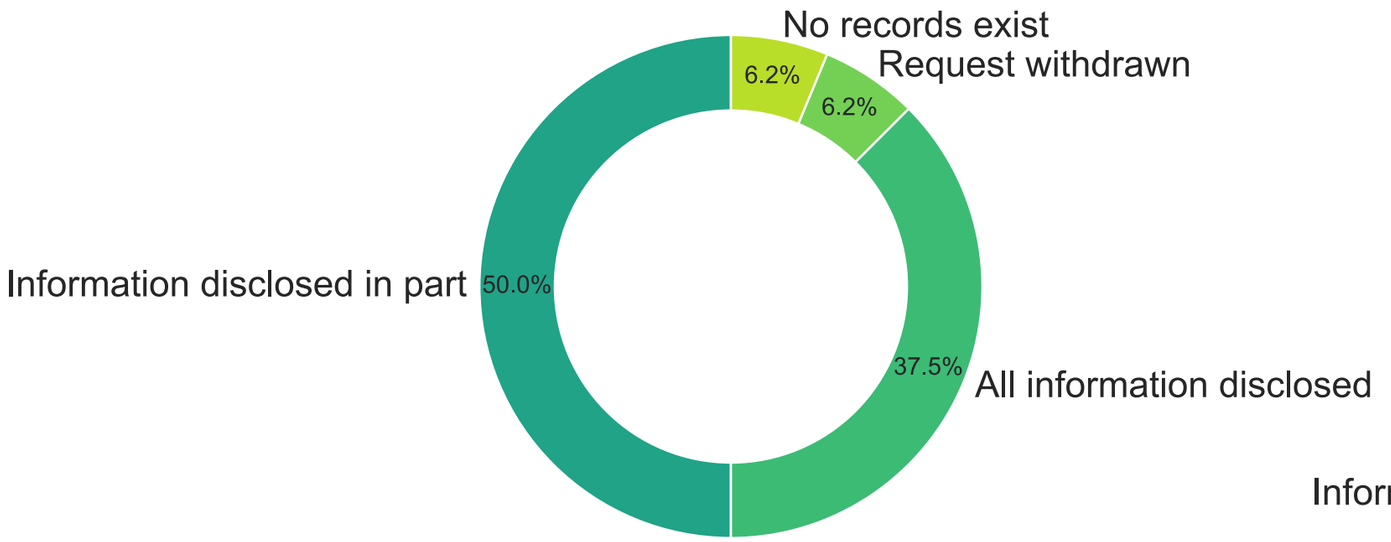
grand river transit



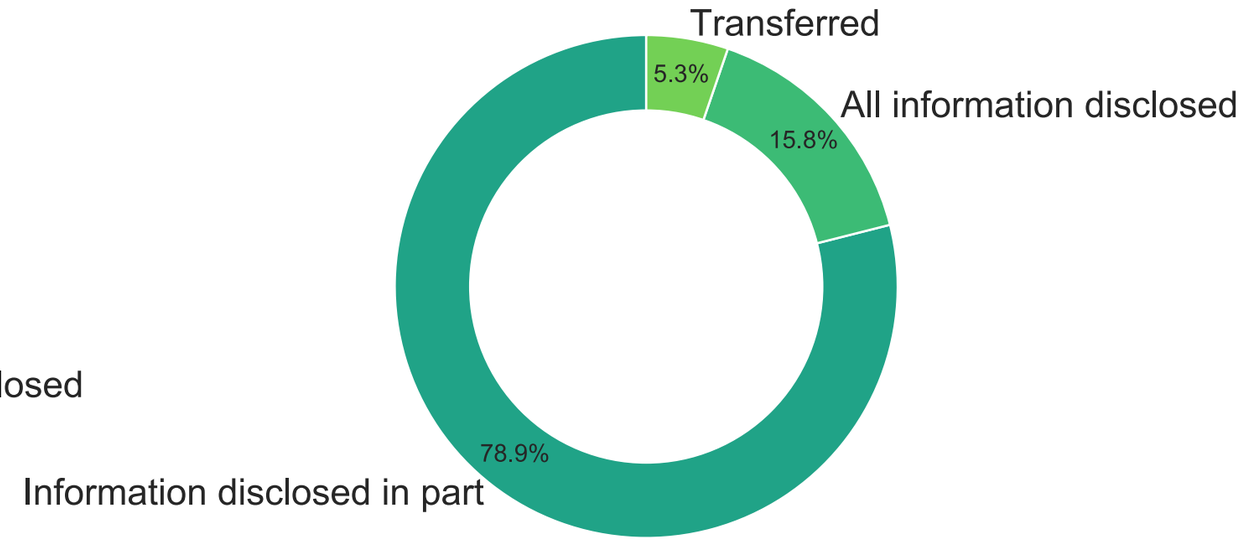
food bear illness



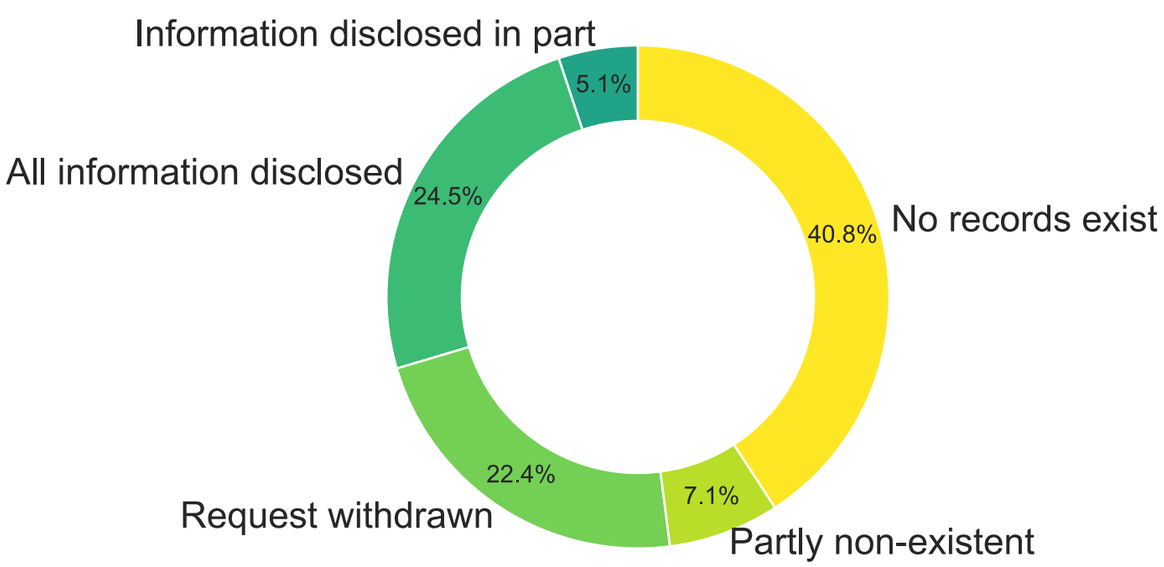
public health inspection



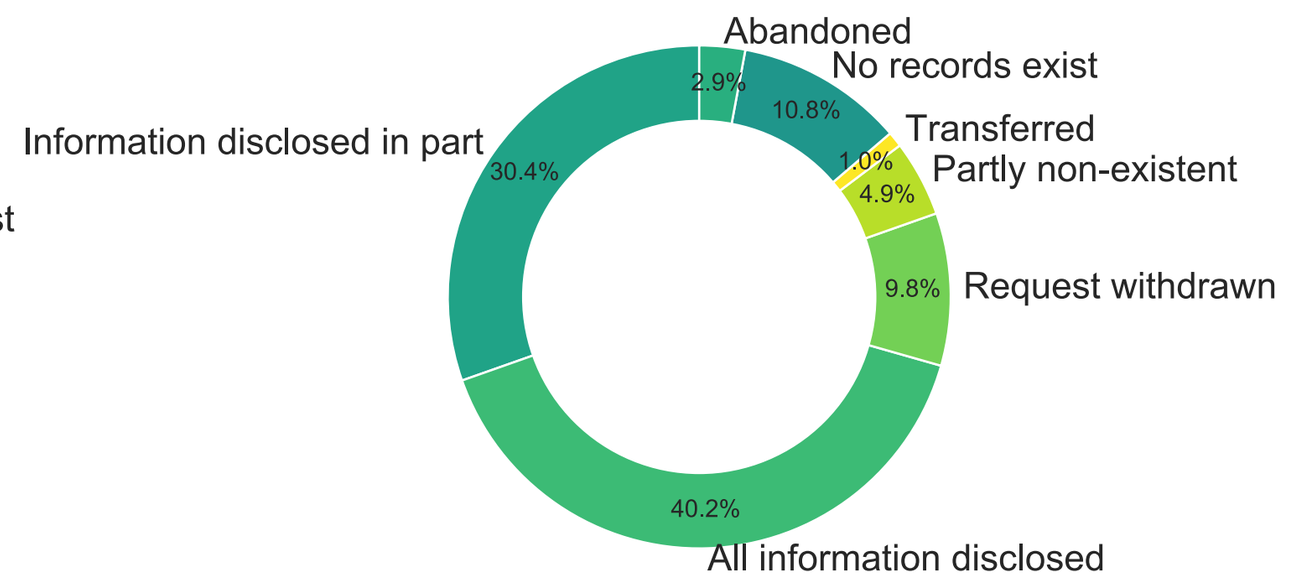
rabies control



environmental site

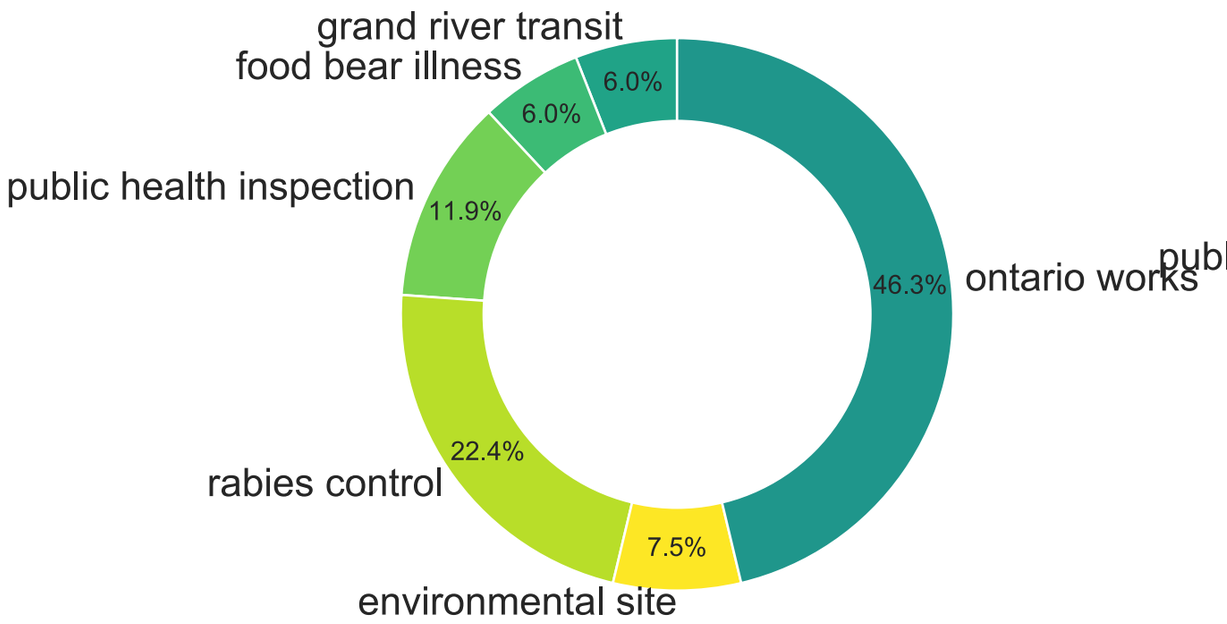


ontario works

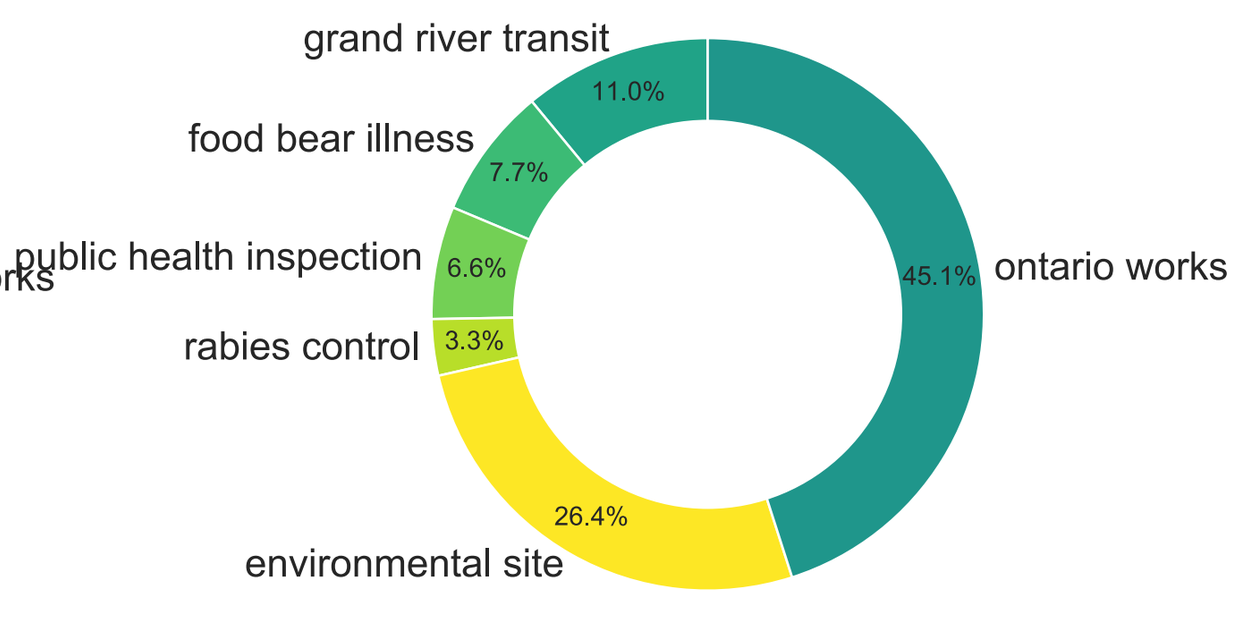


For requests with the n-grams, n-gram percentage based on decision

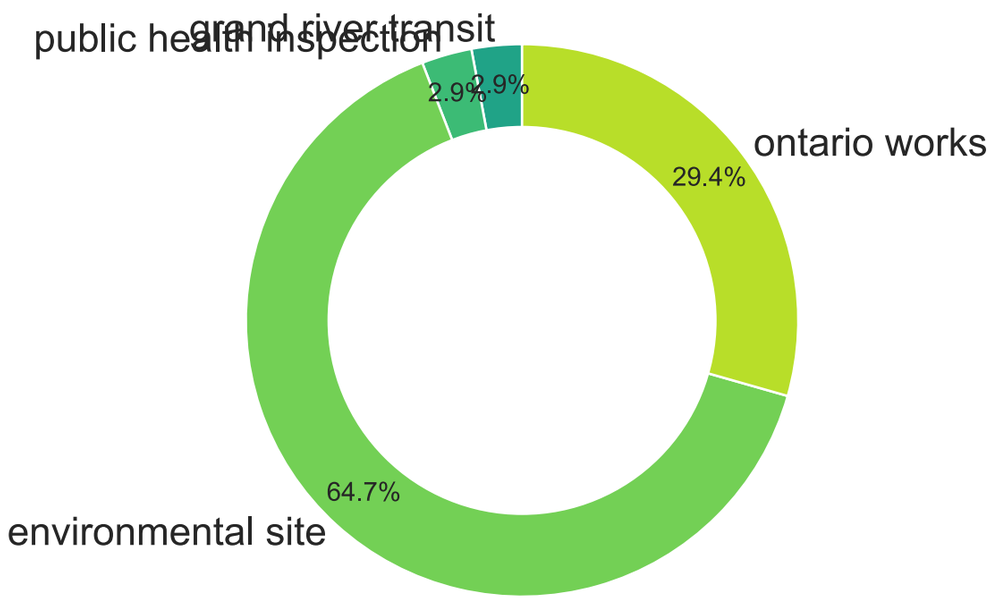
Information disclosed in part



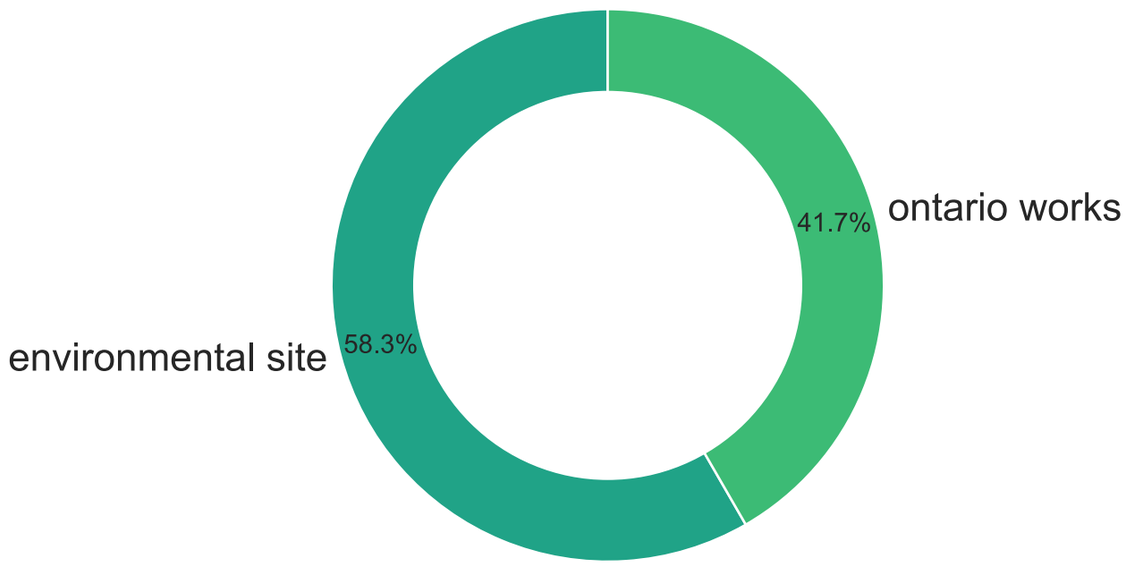
All information disclosed



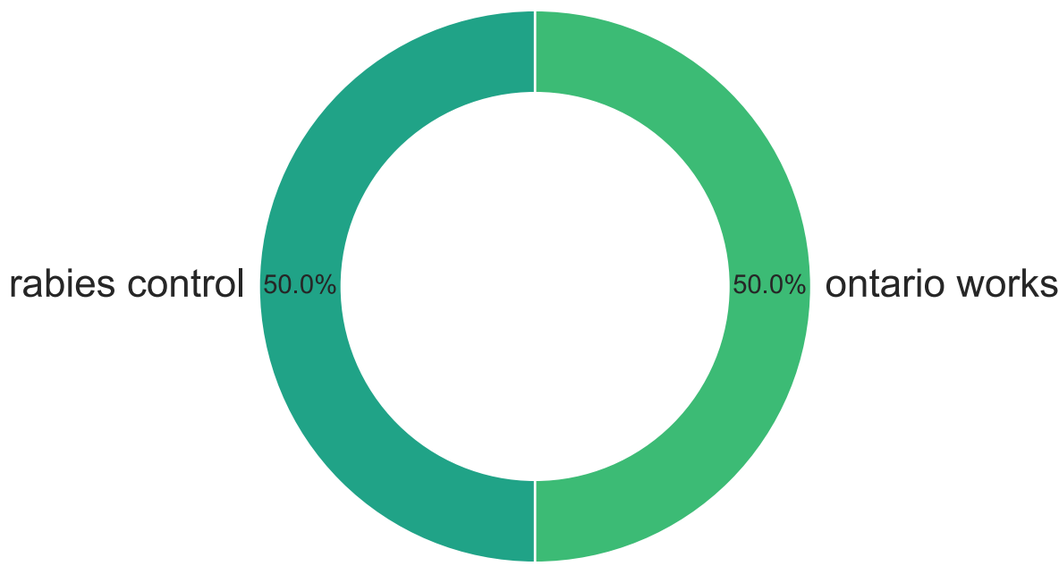
Request withdrawn



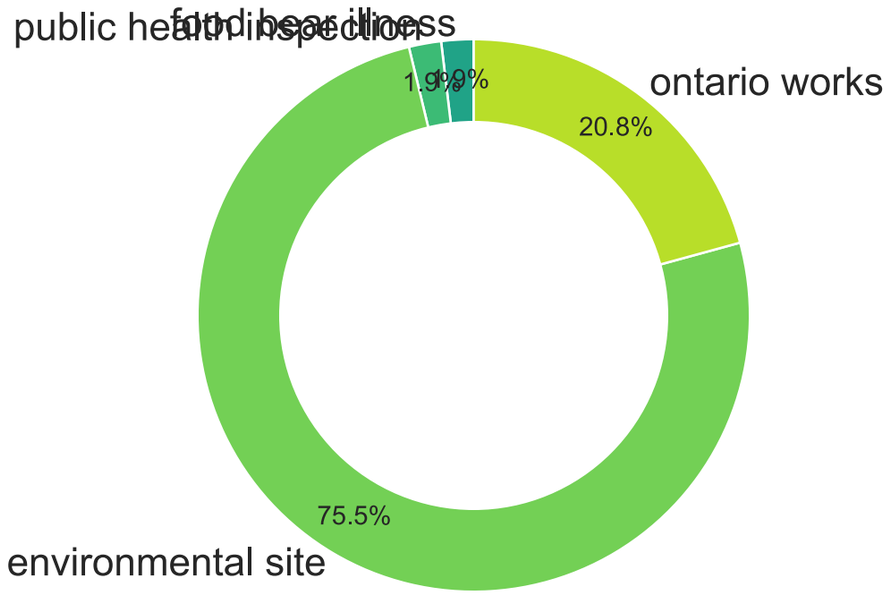
Partly non-existent



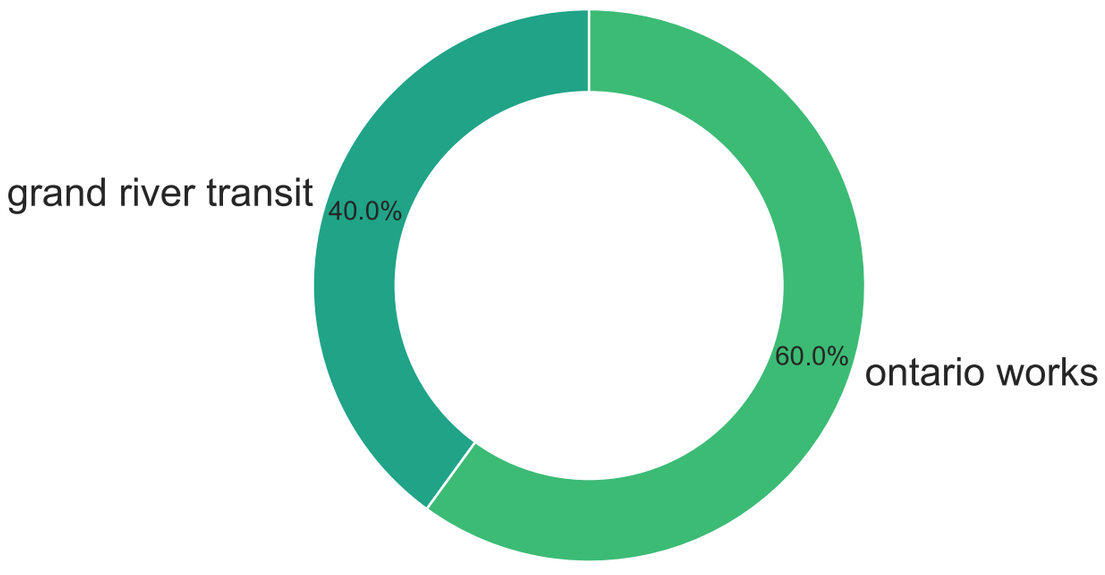
Transferred

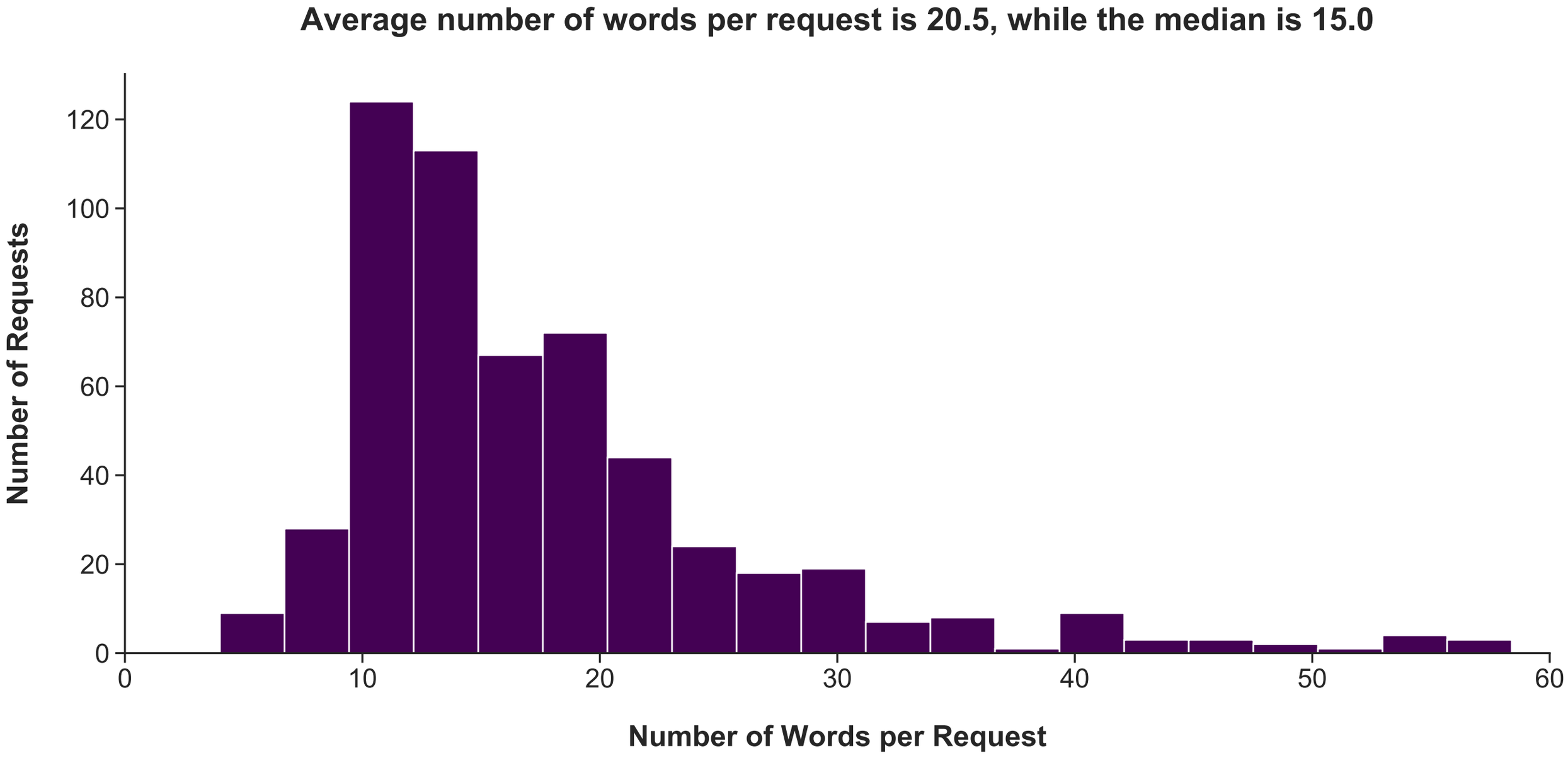


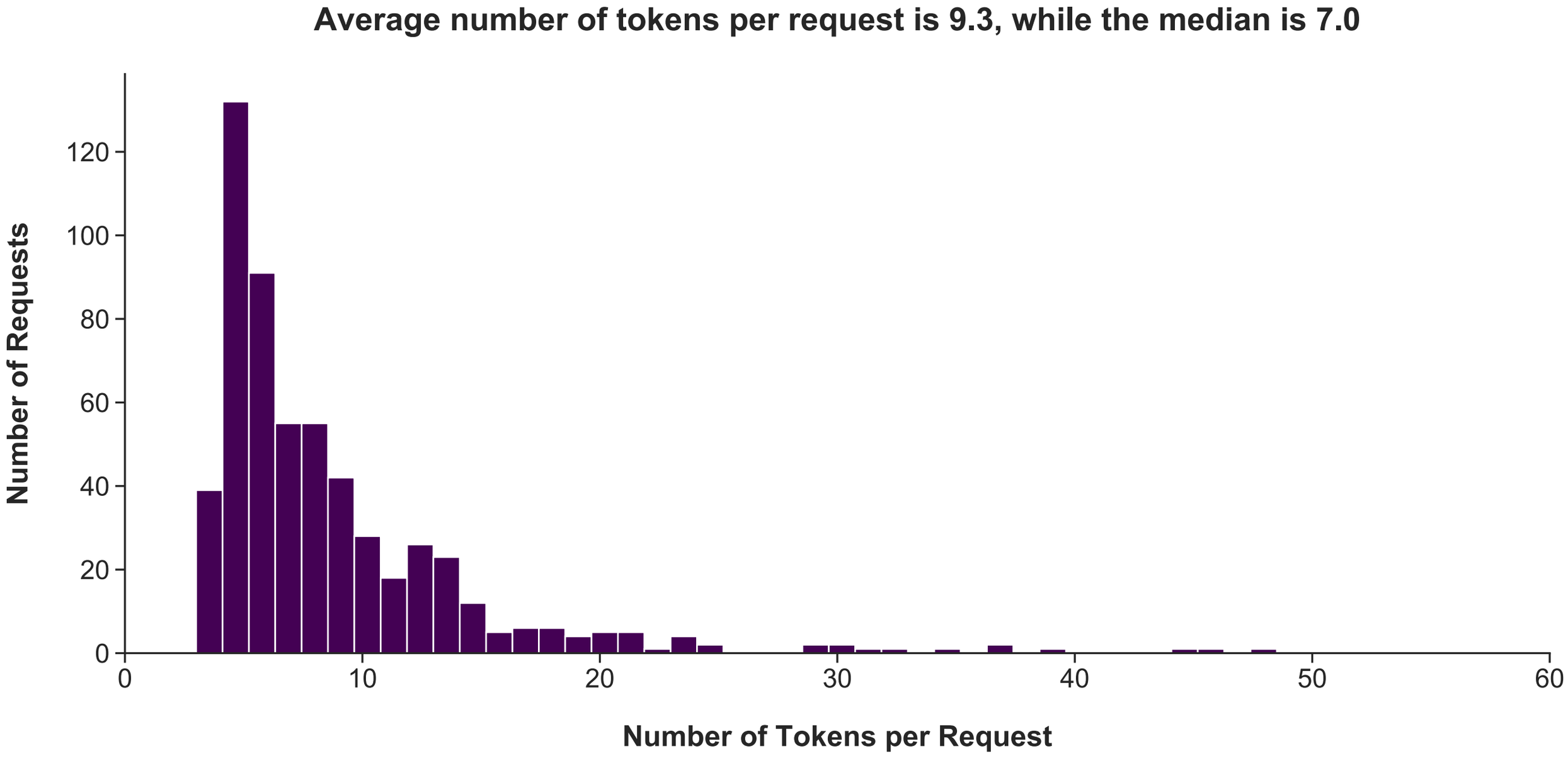
No records exist

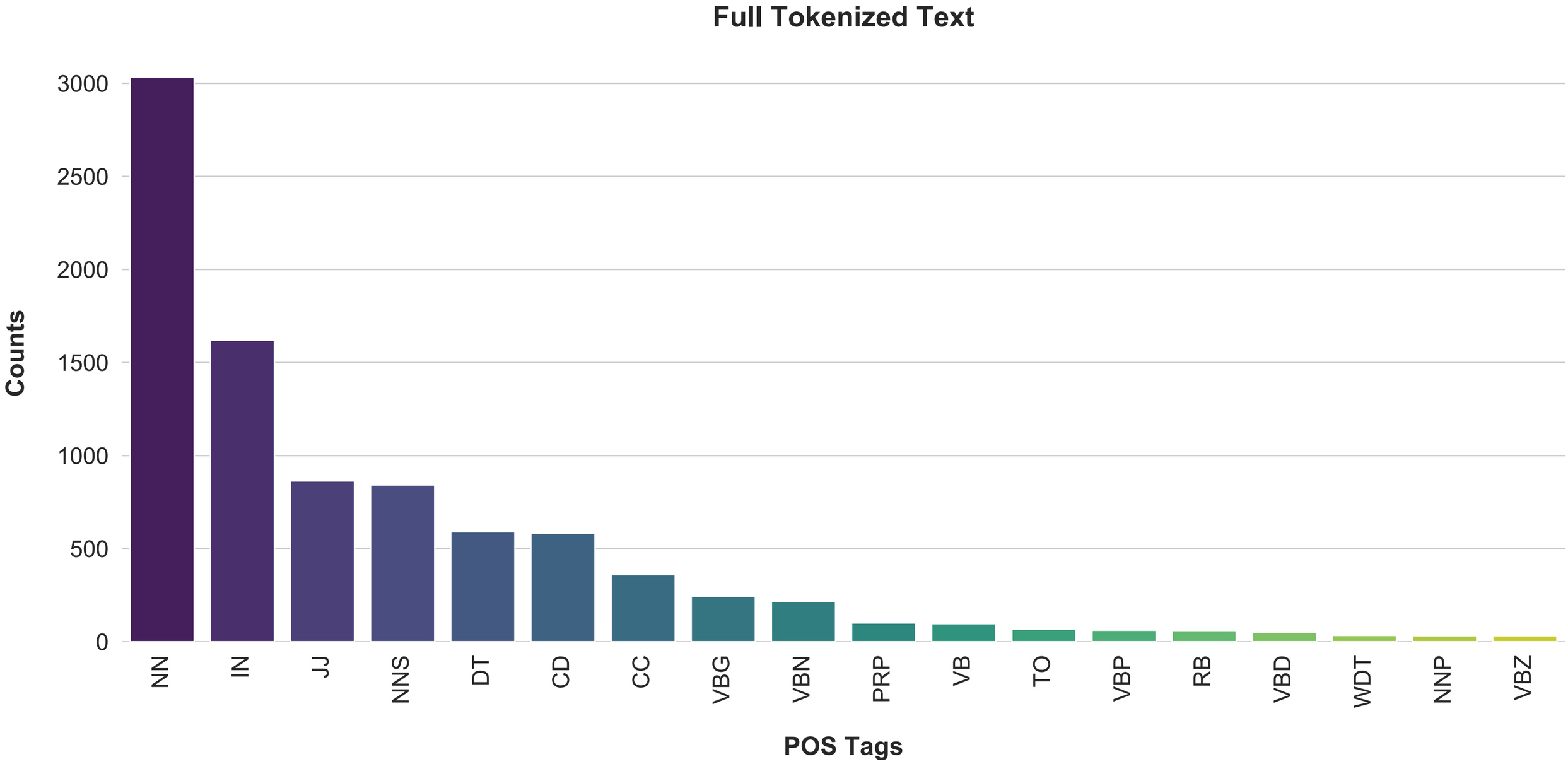


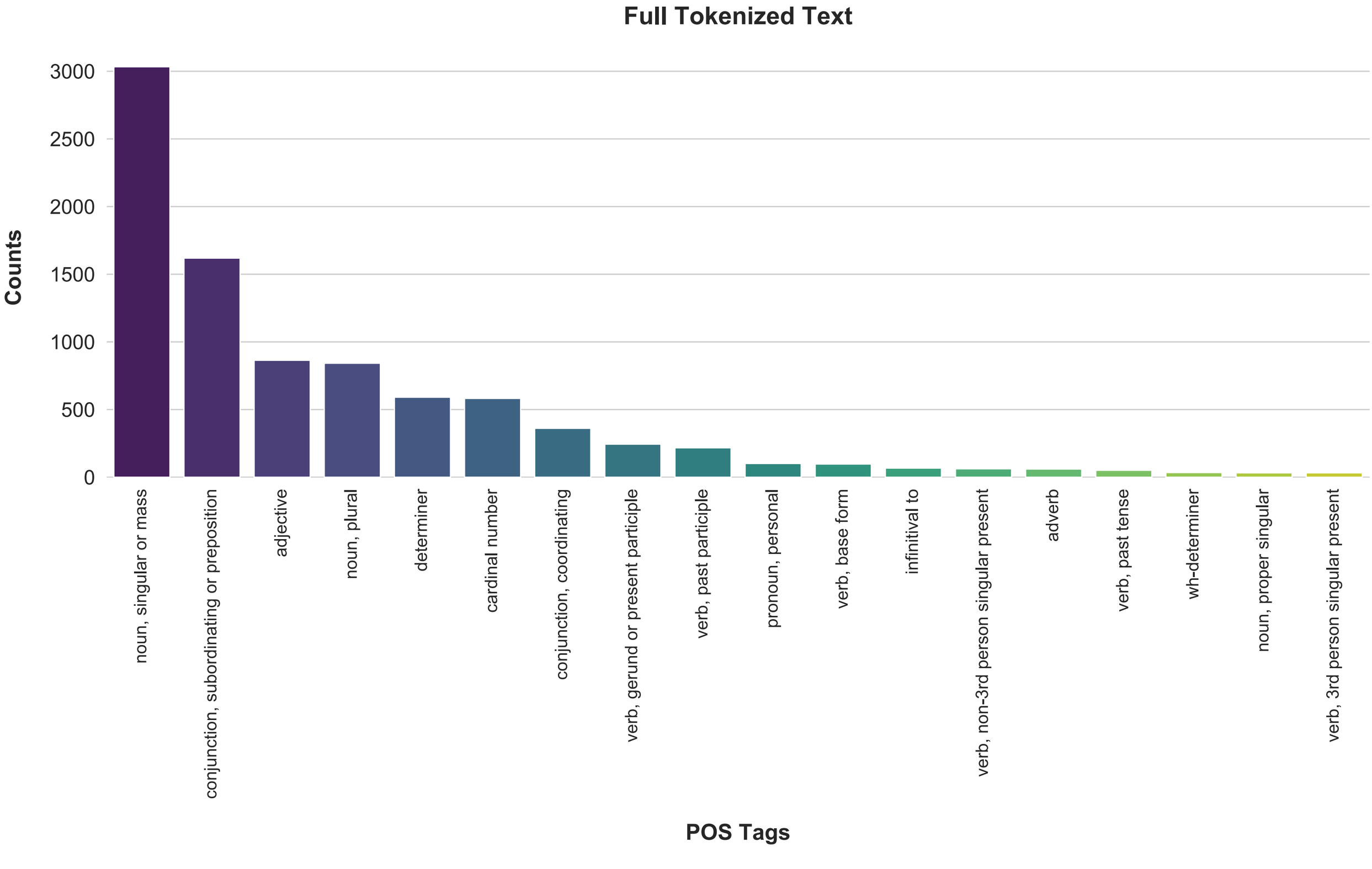
Abandoned

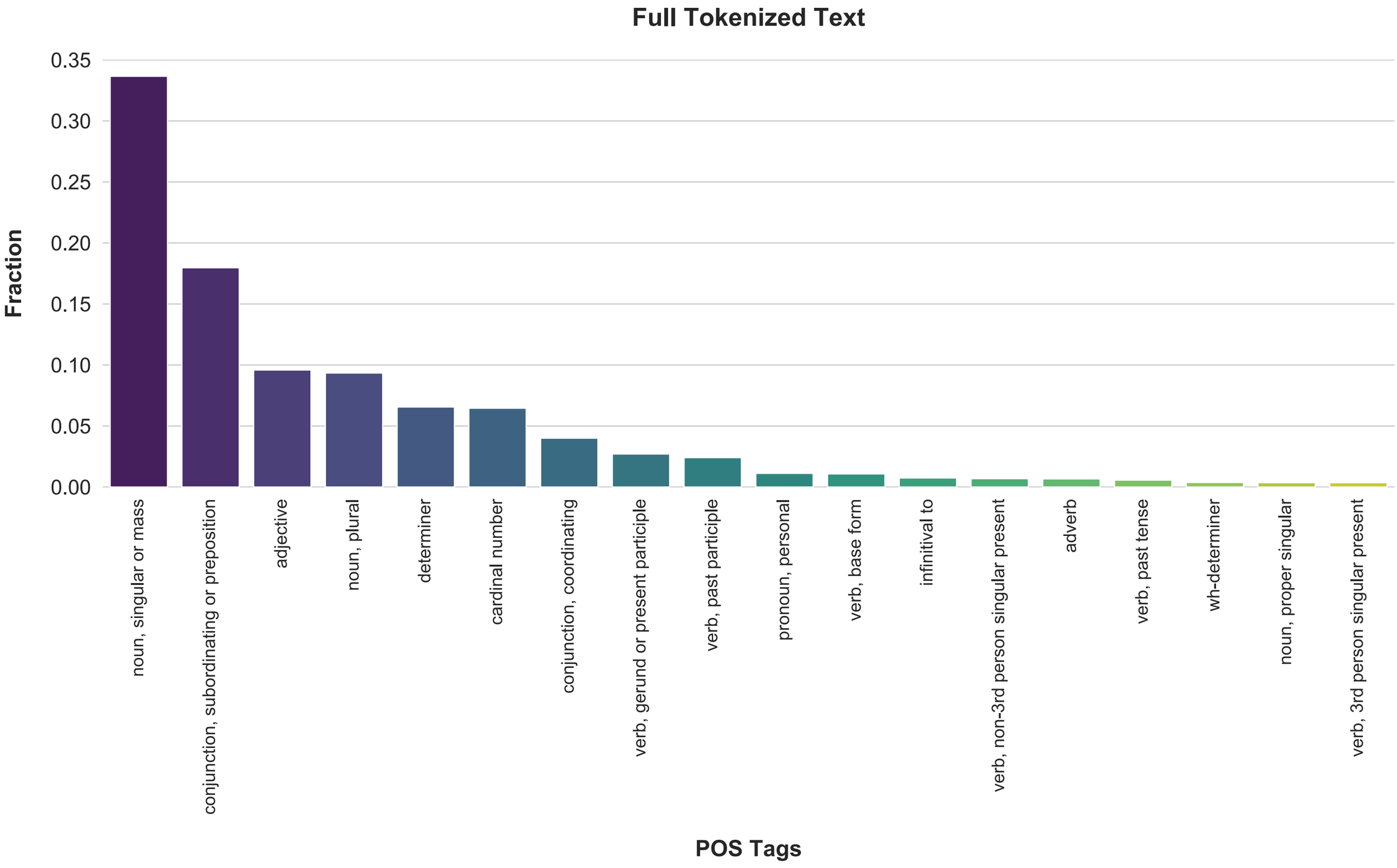


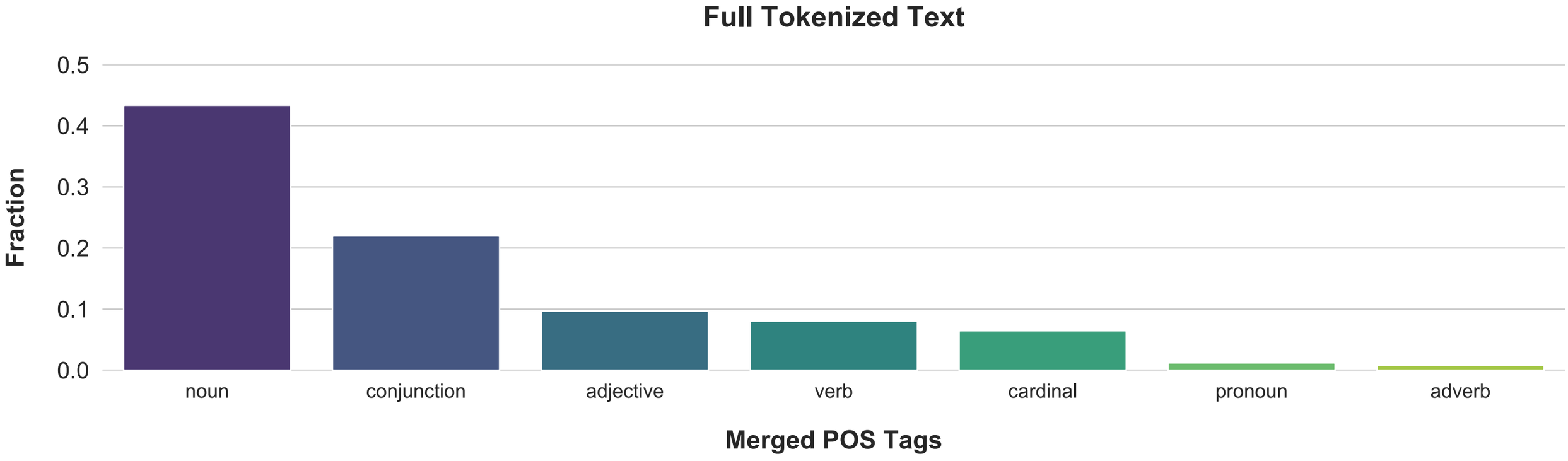

Average number of words per request is 20.5, while the median is 15.0
Number of Requests
0
20
40
60
80
100
120
0
10
20
30
40
50
60
Number of Words per Request

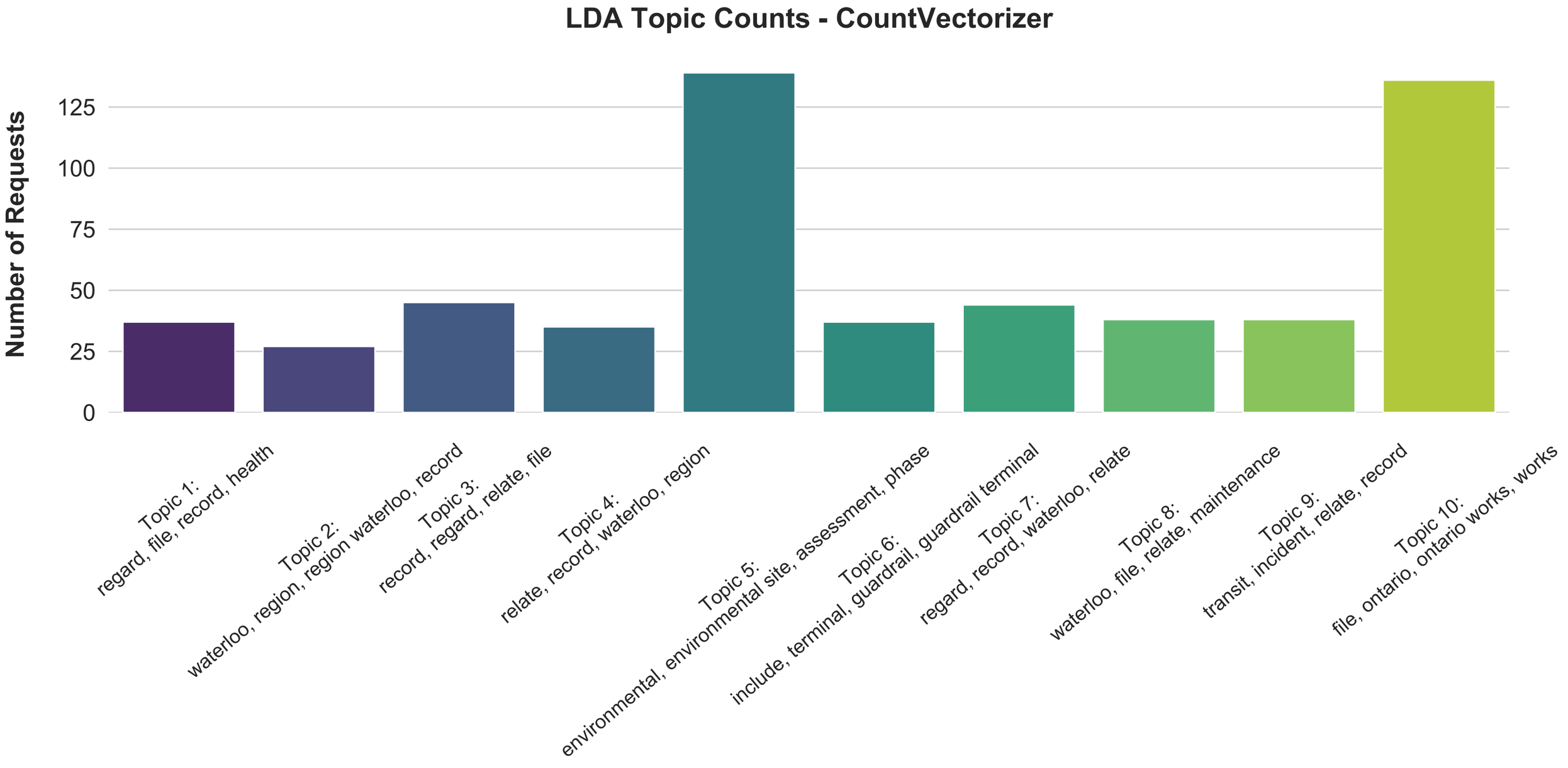

Average number of tokens per request is 9.3, while the median is 7.0
Number of Requests
0
20
40
60
80
100
120
0
10
20
30
40
50
60
Number of Tokens per Request

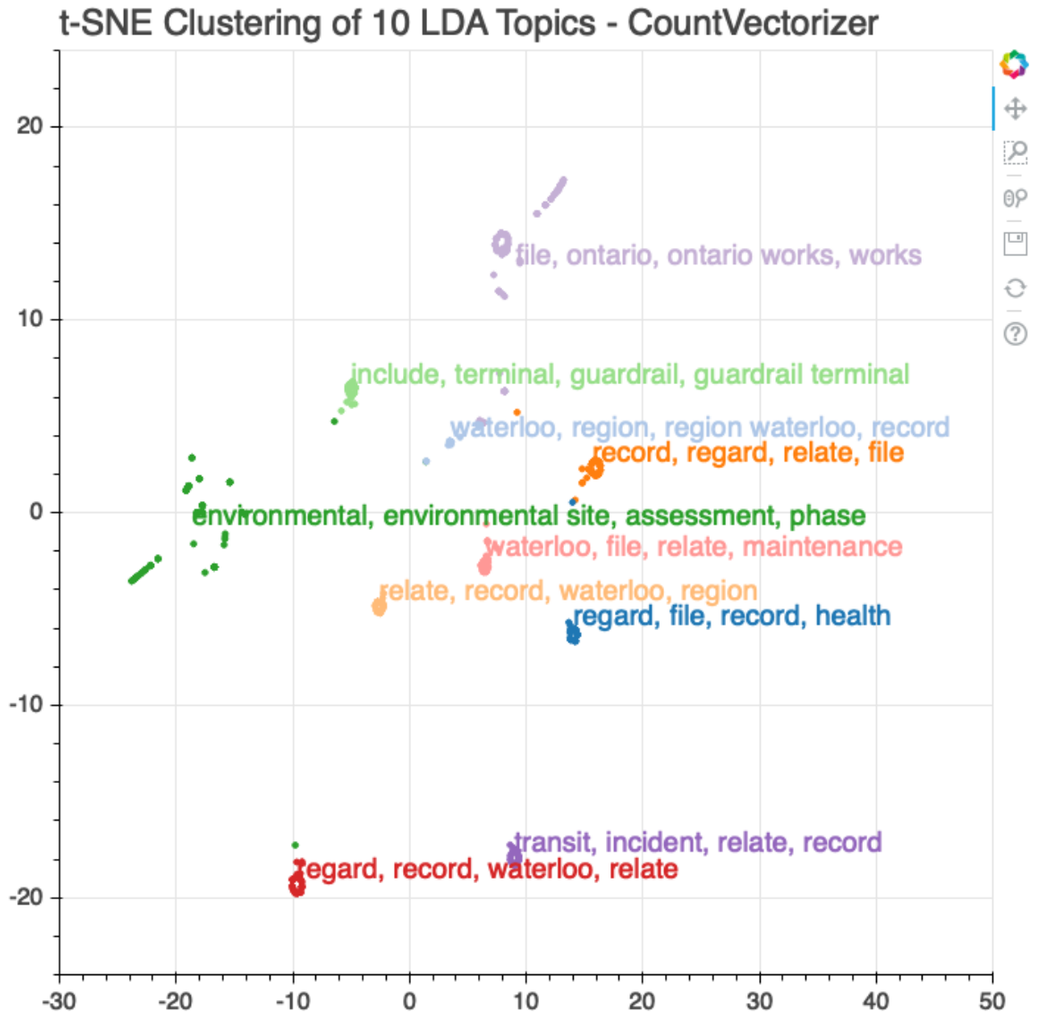

Full Tokenized Text
Counts
3000
2500
2000
1500
1000
500
0
NN
IN
JJ
NNS
DT
CD
CC
VBG
VBN
PRP
VB
TO
VBP
RB
VBD
WDT
NNP
VBZ
POS Tags

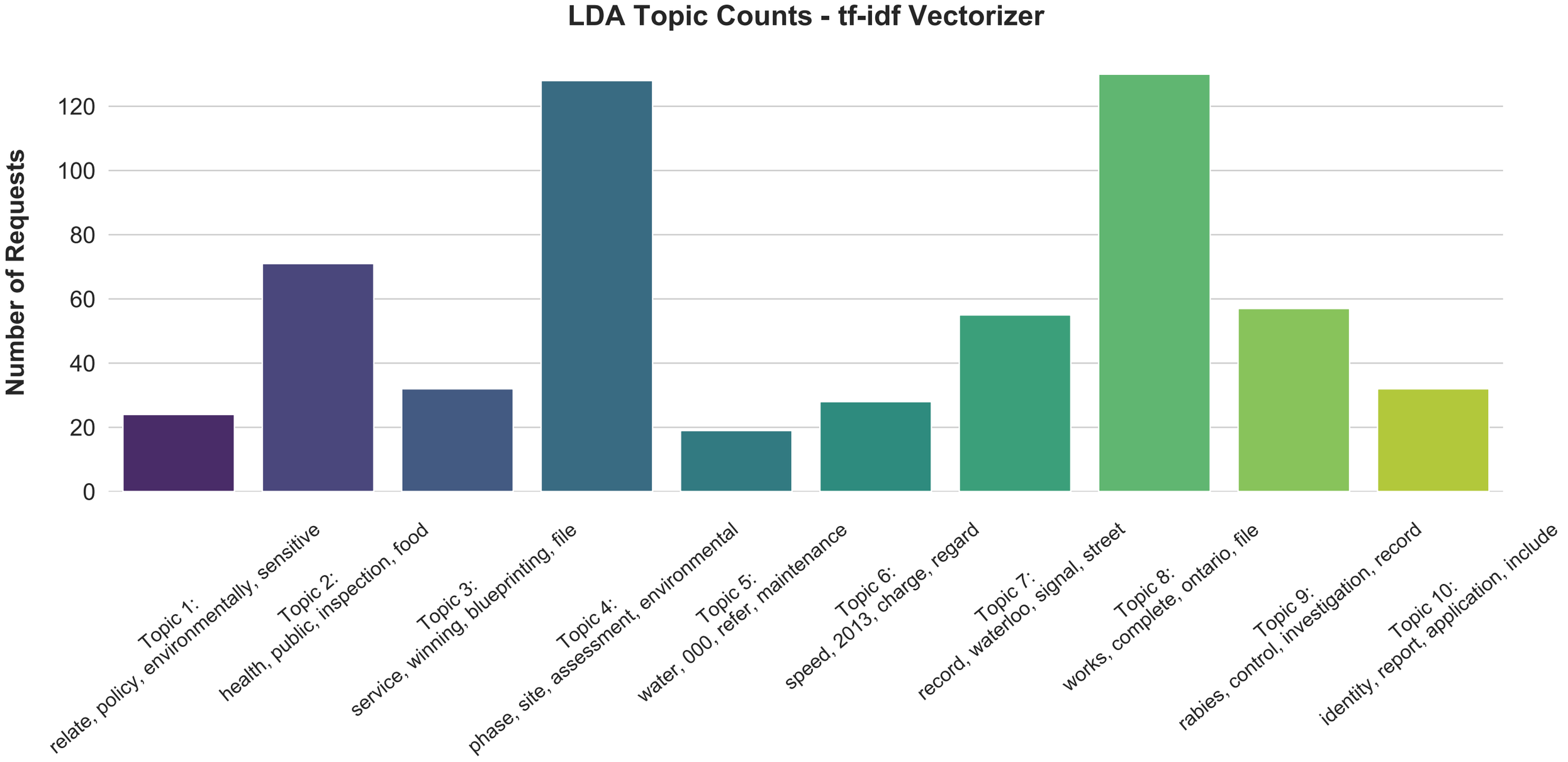

Full Tokenized Text
Counts
3000
2500
2000
1500
1000
500
0
noun, singular or mass
conjunction, subordinating or preposition
adjective
noun, plural
determiner
cardinal number
conjunction, coordinating
verb, gerund or present participle
verb, past participle
pronoun, personal
verb, base form
infinitival to
verb, non-3rd person singular present
adverb
verb, past tense
wh-determiner
noun, proper singular
verb, 3rd person singular present
POS Tags

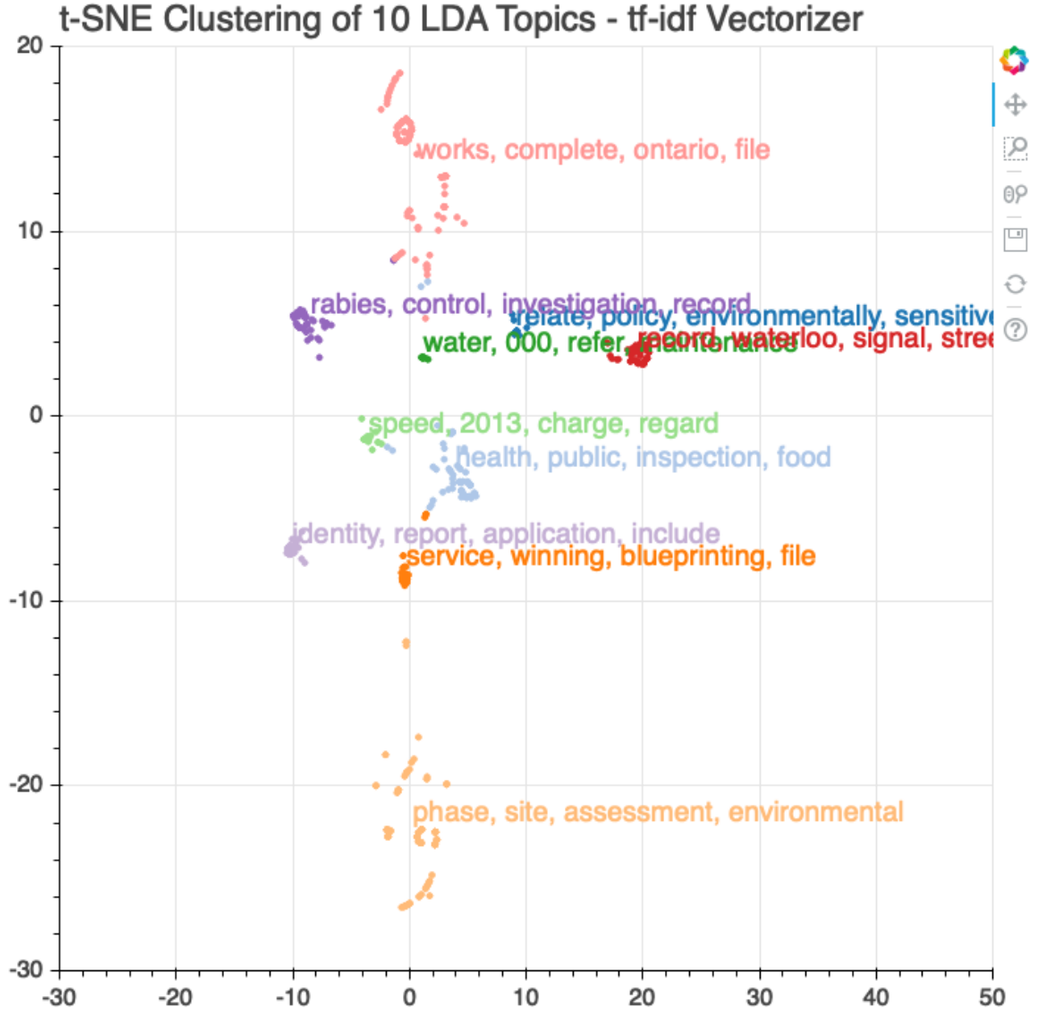

Full Tokenized Text
Fraction
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
noun, singular or mass
conjunction, subordinating or preposition
adjective
noun, plural
determiner
cardinal number
conjunction, coordinating
verb, gerund or present participle
verb, past participle
pronoun, personal
verb, base form
infinitival to
verb, non-3rd person singular present
adverb
verb, past tense
wh-determiner
noun, proper singular
verb, 3rd person singular present
POS Tags

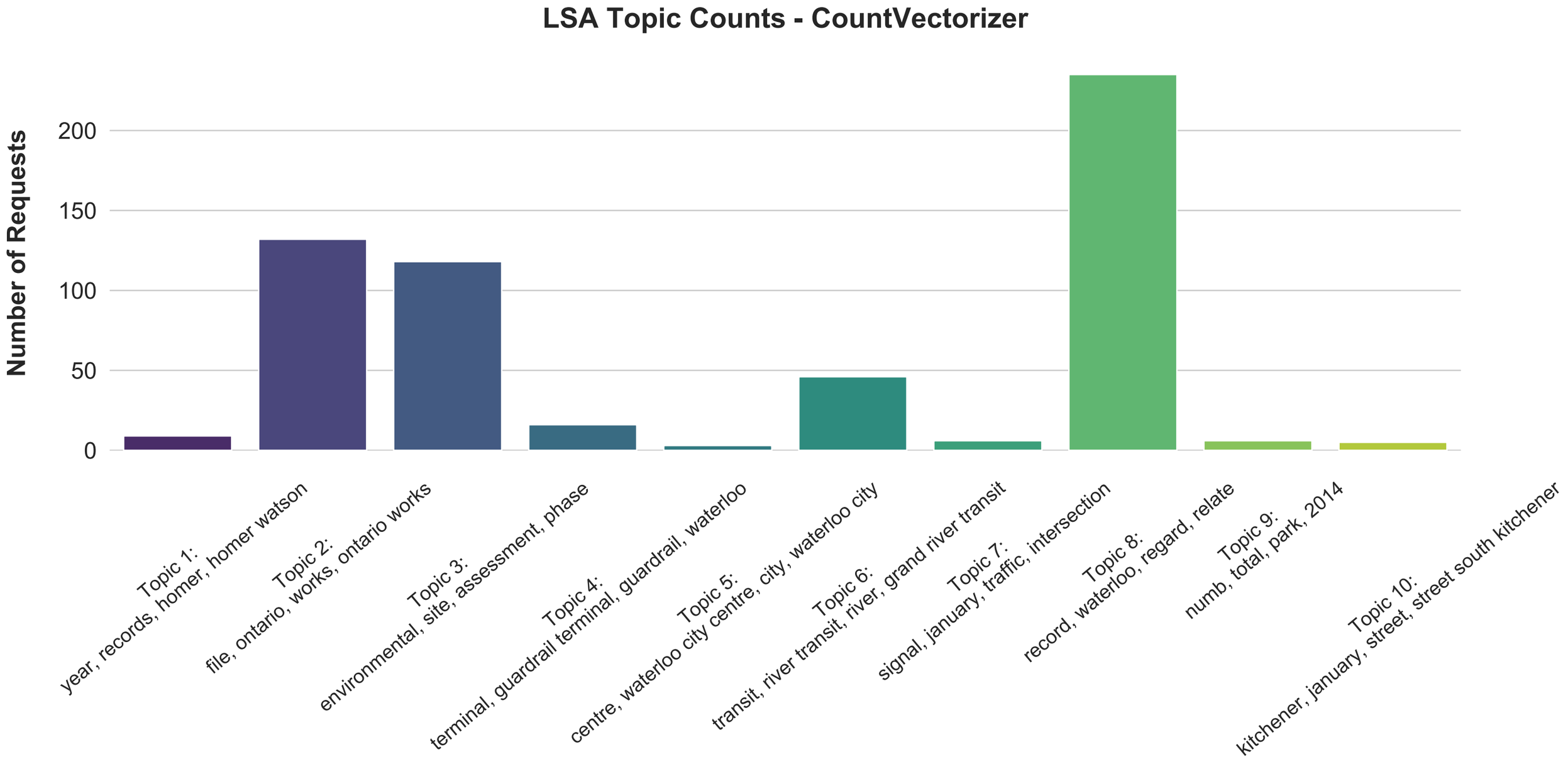

Full Tokenized Text
Fraction
0.5
0.4
0.3
0.2
0.1
0.0
noun
conjunction
adjective
verb
cardinal
pronoun
adverb
Merged POS Tags

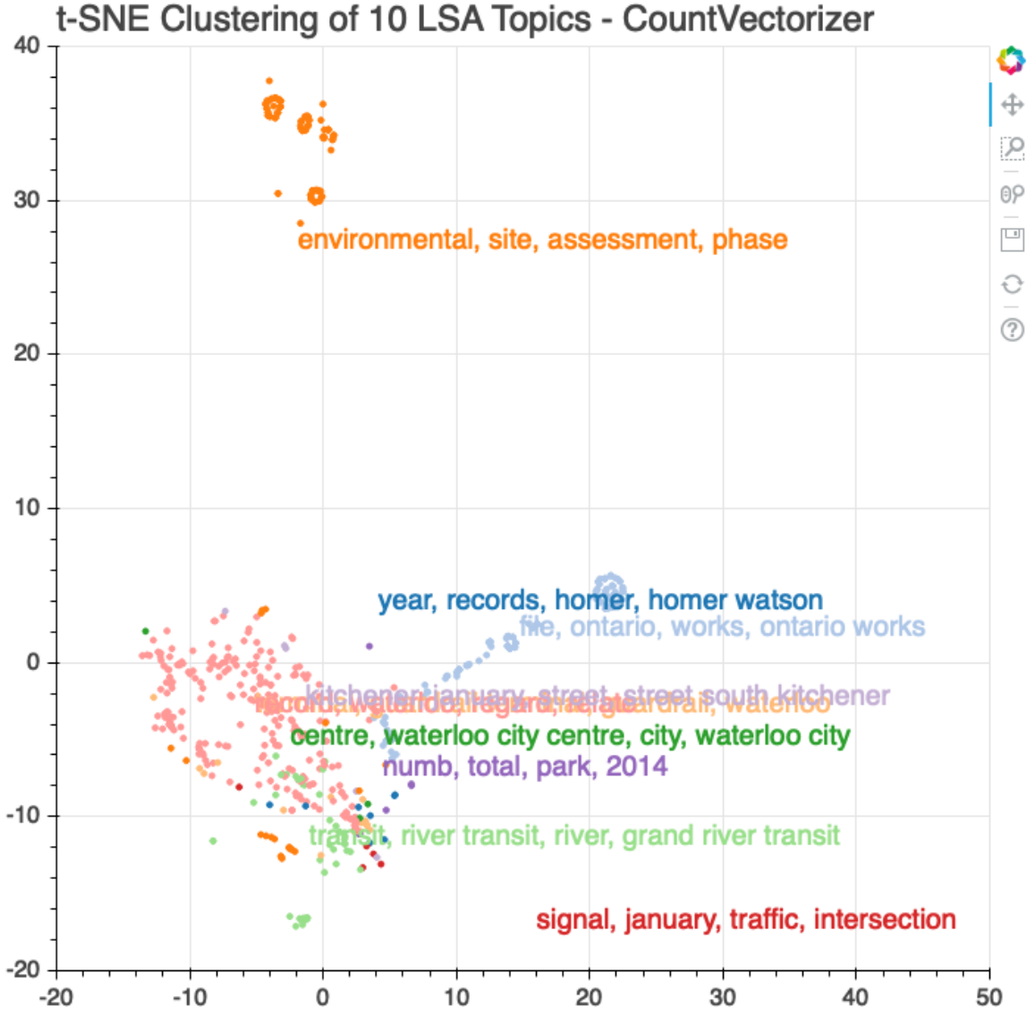

LDA Topic Counts - CountVectorizer
Number of Requests
0
25
50
75
100
125
Topic 1:
regard, file, record, health
Topic 2:
waterloo, region, region waterloo, record
Topic 3:
record, regard, relate, file
Topic 4:
relate, record, waterloo, region
Topic 5:
environmental, environmental site, assessment, phase
Topic 6:
include, terminal, guardrail, guardrail terminal
Topic 7:
regard, record, waterloo, relate
Topic 8:
waterloo, file, relate, maintenance
Topic 9:
transit, incident, relate, record
Topic 10:
file, ontario, ontario works, works

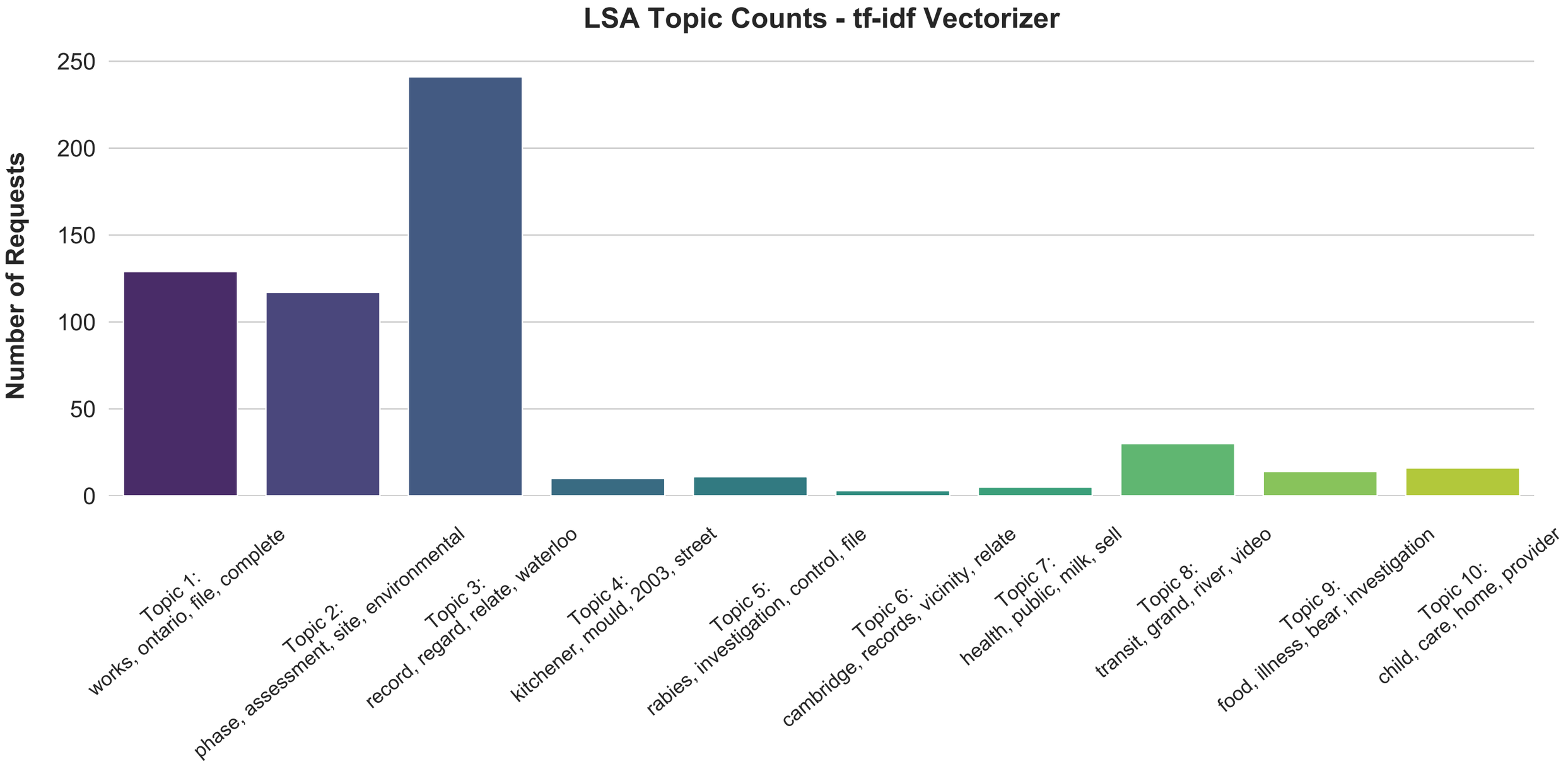

t-SNE Clustering of 10 LDA Topics - CountVectorizer
file, ontario, ontario works, works
include, terminal, guardrail, guardrail terminal
waterloo, region, region waterloo, record
record, regard, relate, file
environmental, environmental site, assessment, phase
waterloo, file, relate, maintenance
relate, record, waterloo, region
regard, file, record, health
transit, incident, relate, record
egard, record, waterloo, relate
20
10
0
-10
-20
-30
-20
-10
0
10
20
30
40
50

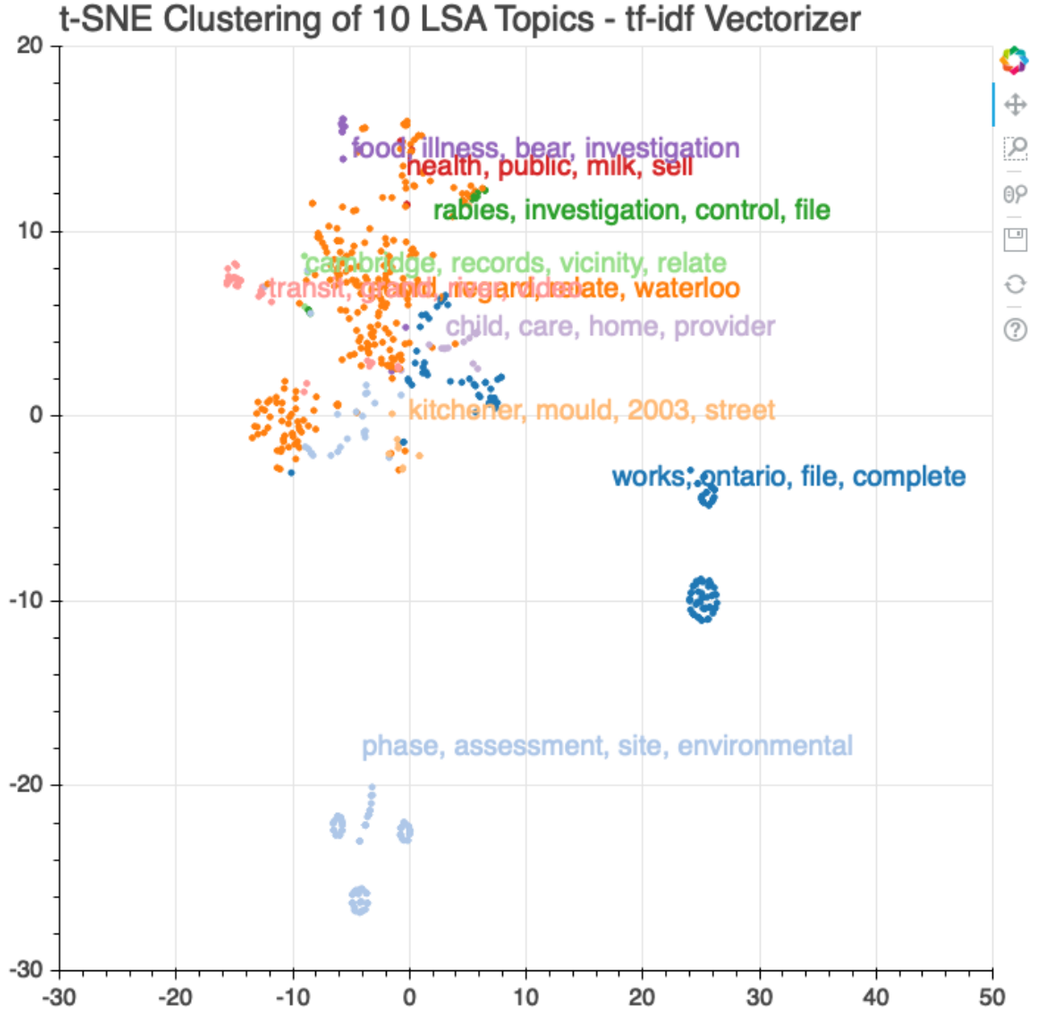

LDA Topic Counts - tf-idf Vectorizer
Number of Requests
0
20
40
60
80
100
120
Topic 1:
relate, policy, environmentally, sensitive
Topic 2:
health, public, inspection, food
Topic 3:
service, winning, blueprinting, file
Topic 4:
phase, site, assessment, environmental
Topic 5:
water, 000, refer, maintenance
Topic 6:
speed, 2013, charge, regard
Topic 7:
record, waterloo, signal, street
Topic 8:
works, complete, ontario, file
Topic 9:
rabies, control, investigation, record
Topic 10:
identity, report, application, include

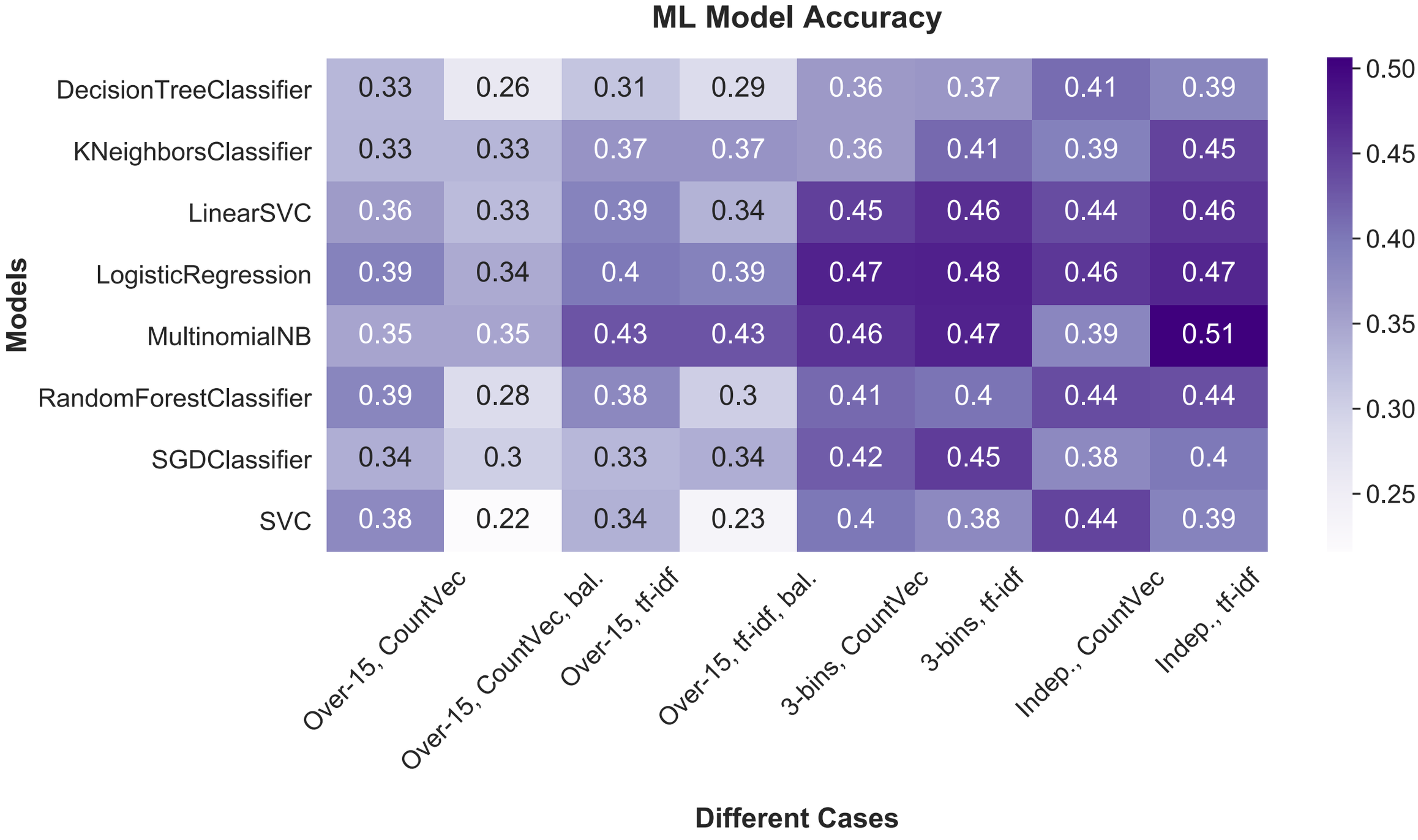

t-SNE Clustering of 10 LDA Topics - tf-idf Vectorizer
works, complete, ontario, file
rabies, control, investigation, record
water, 000, refer, record, waterloo, signal, stree
speed, 2013, charge, regard
health, public, inspection, food
identity, report, application, include
service, winning, blueprinting, file
phase, site, assessment, environmental
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
40
50

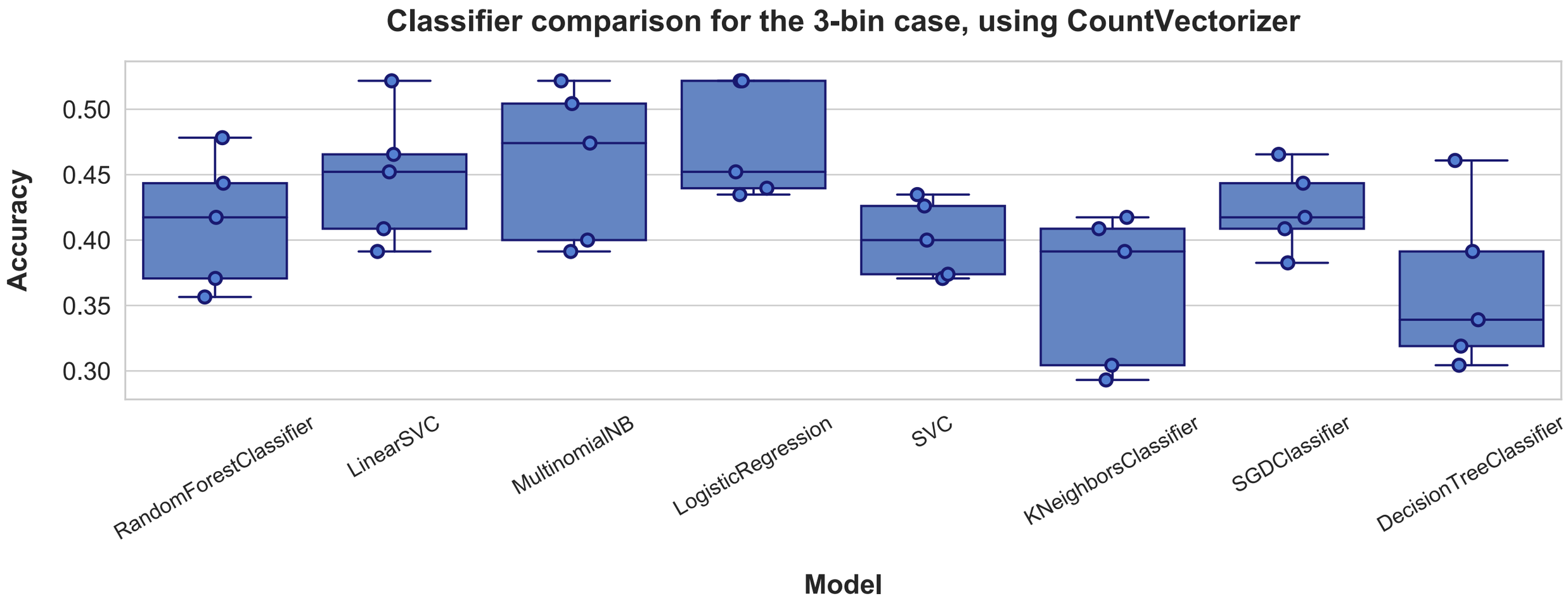

LSA Topic Counts - CountVectorizer
Number of Requests
200
150
100
50
0
Topic 1:
year, records, homer, homer watson
Topic 2:
file, ontario, works, ontario works
Topic 3:
environmental, site, assessment, phase
Topic 4:
terminal, guardrail terminal, guardrail, waterloo
Topic 5:
centre, waterloo city centre, city, waterloo city
Topic 6:
transit, river transit, river, grand river transit
Topic 7:
signal, january, traffic, intersection
Topic 8:
record, waterloo, regard, relate
Topic 9:
numb, total, park, 2014
Topic 10:
kitchener, january, street, street south kitchener

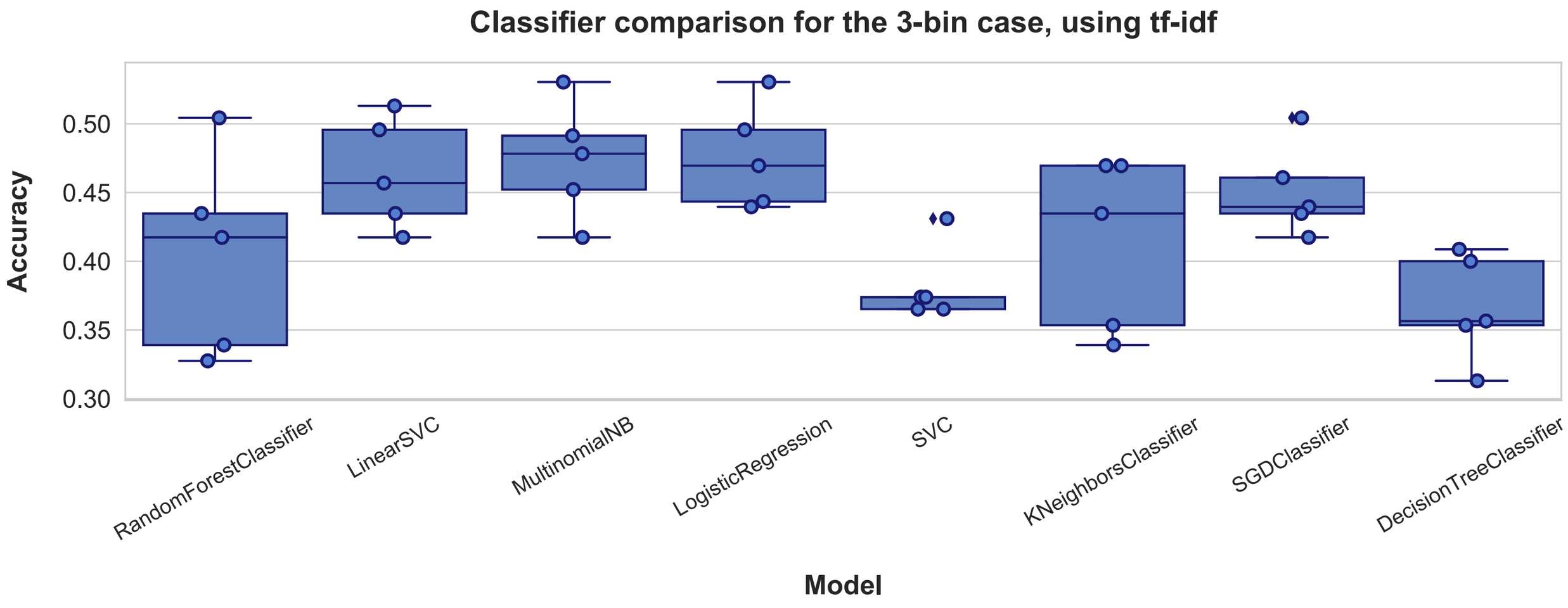

t-SNE Clustering of 10 LSA Topics - CountVectorizer
environmental, site, assessment, phase
year, records, homer, homer watson
centre, waterloo city centre, city, waterloo city
numb, total, park, 2014
transit, river transit, river, grand river transit
signal, january, traffic, intersection
40
30
20
10
0
-10
-20
-20
-10
0
10
20
30
40
50

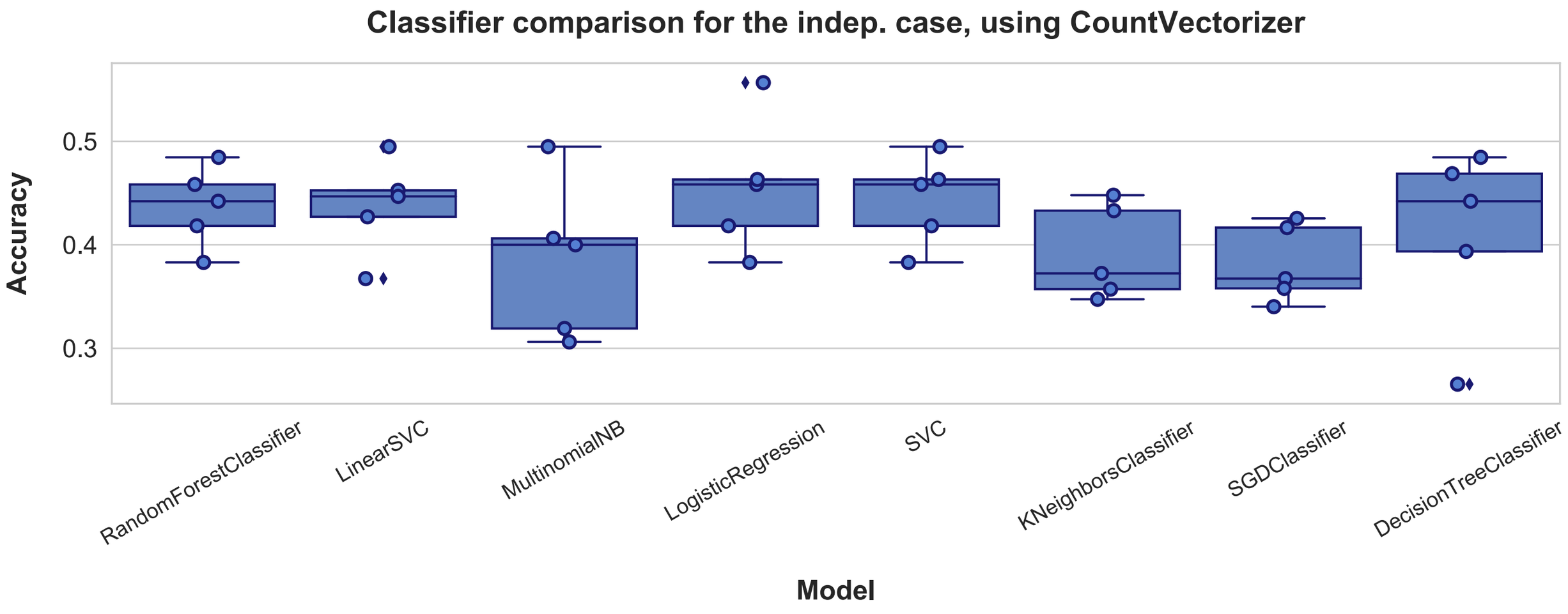

LSA Topic Counts - tf-idf Vectorizer
Number of Requests
250
200
150
100
50
0
Topic 1:
works, ontario, file, complete
Topic 2:
phase, assessment, site, environmental
Topic 3:
record, regard, relate, waterloo
Topic 4:
kitchener, mould, 2003, street
Topic 5:
rabies, investigation, control, file
Topic 6:
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

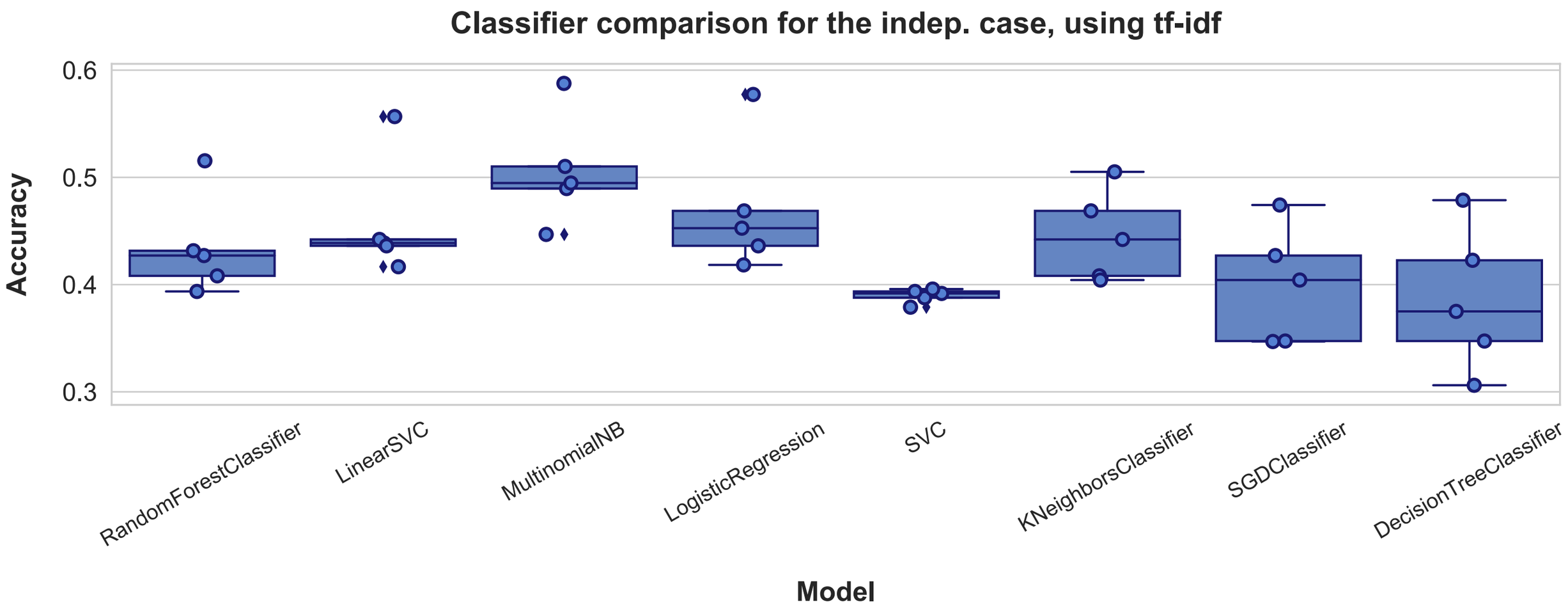

t-SNE Clustering of 10 LSA Topics - tf-idf Vectorizer
food, illness, bear, investigation
health, public, milk, sell
rabies, investigation, control, file
cambridge, records, vicinity, relate
transit, guelph, regard, relate, waterloo
child, care, home, provider
kitchener, mould, 2003, street
works, ontario, file, complete
phase, assessment, site, environmental
20
10
0
-10
-20
-30
-30
-20
-10
0
10
20
30
40
50

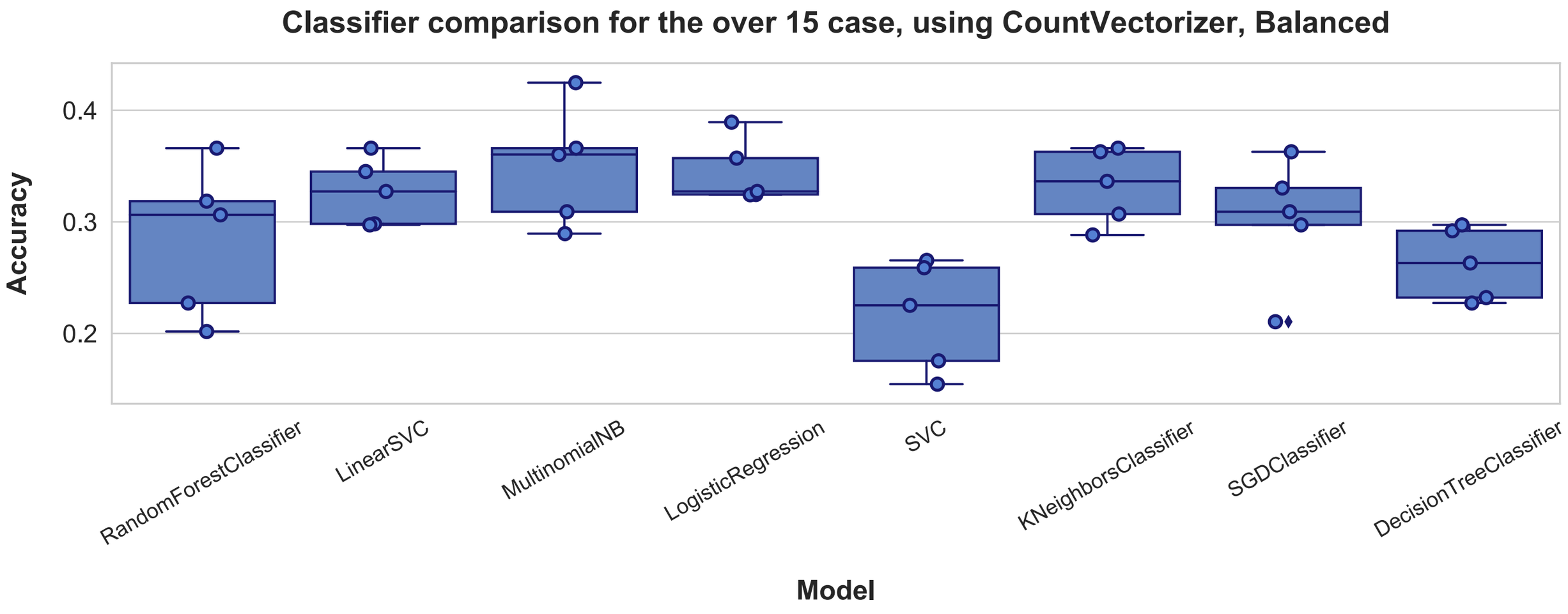

ML Model Accuracy
Models
DecisionTreeClassifier
KNeighborsClassifier
LinearSVC
LogisticRegression
MultinomialNB
RandomForestClassifier
SGDClassifier
SVC
0.33 0.26 0.31 0.29 0.36 0.37 0.41 0.39
0.33 0.33 0.37 0.37 0.36 0.41 0.39 0.45
0.36 0.33 0.39 0.34 0.45 0.46 0.44 0.46
0.39 0.34 0.4 0.39 0.47 0.48 0.46 0.47
0.35 0.35 0.43 0.43 0.46 0.47 0.39 0.51
0.39 0.28 0.38 0.3 0.41 0.4 0.44 0.44
0.34 0.3 0.33 0.34 0.42 0.45 0.38 0.4
0.38 0.22 0.34 0.23 0.4 0.38 0.44 0.39
0.50
0.45
0.40
0.35
0.30
0.25
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
Different Cases

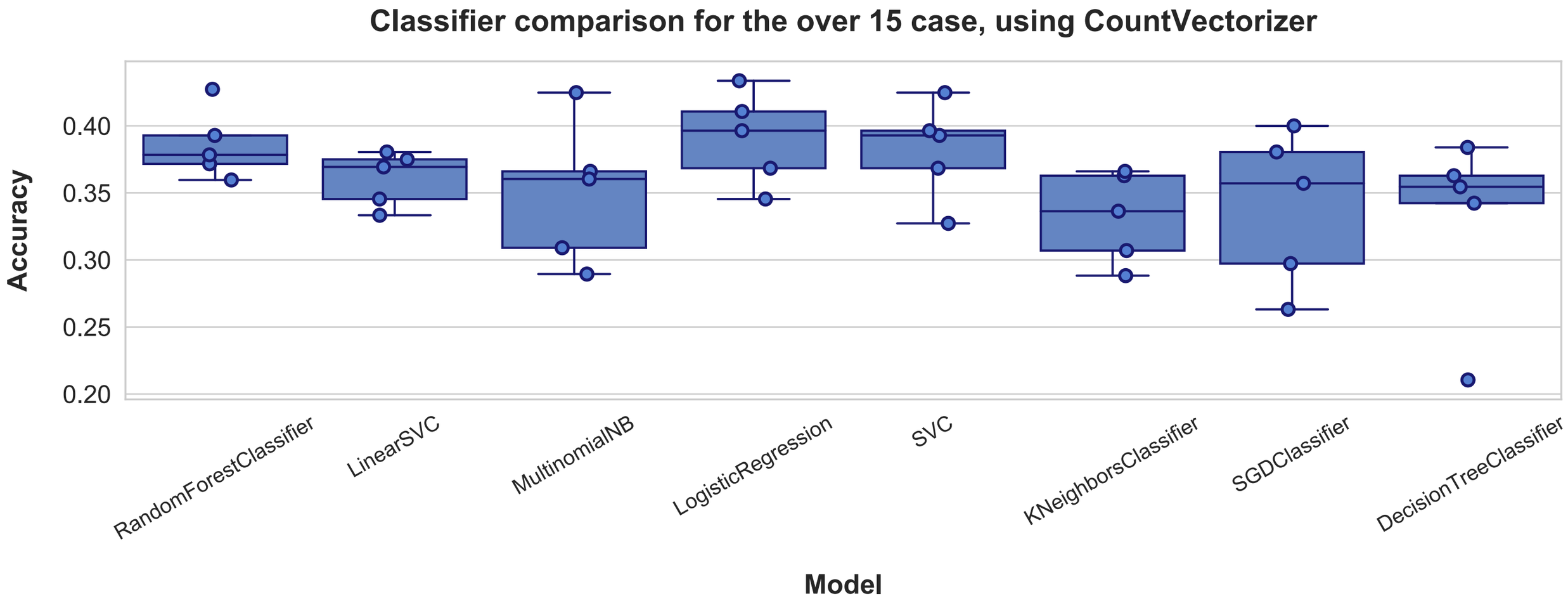

Classifier comparison for the 3-bin case, using CountVectorizer
Accuracy
0.50
0.45
0.40
0.35
0.30
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

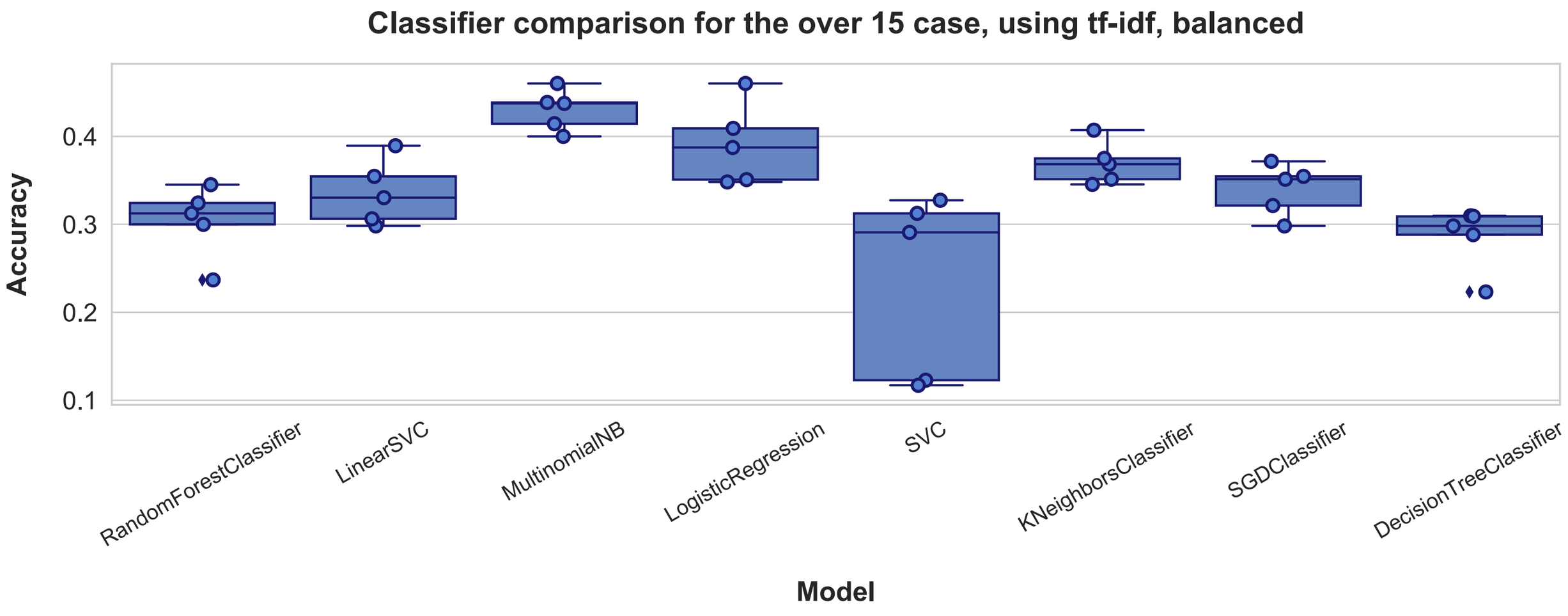

Classifier comparison for the 3-bin case, using tf-idf
Accuracy
0.50
0.45
0.40
0.35
0.30
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

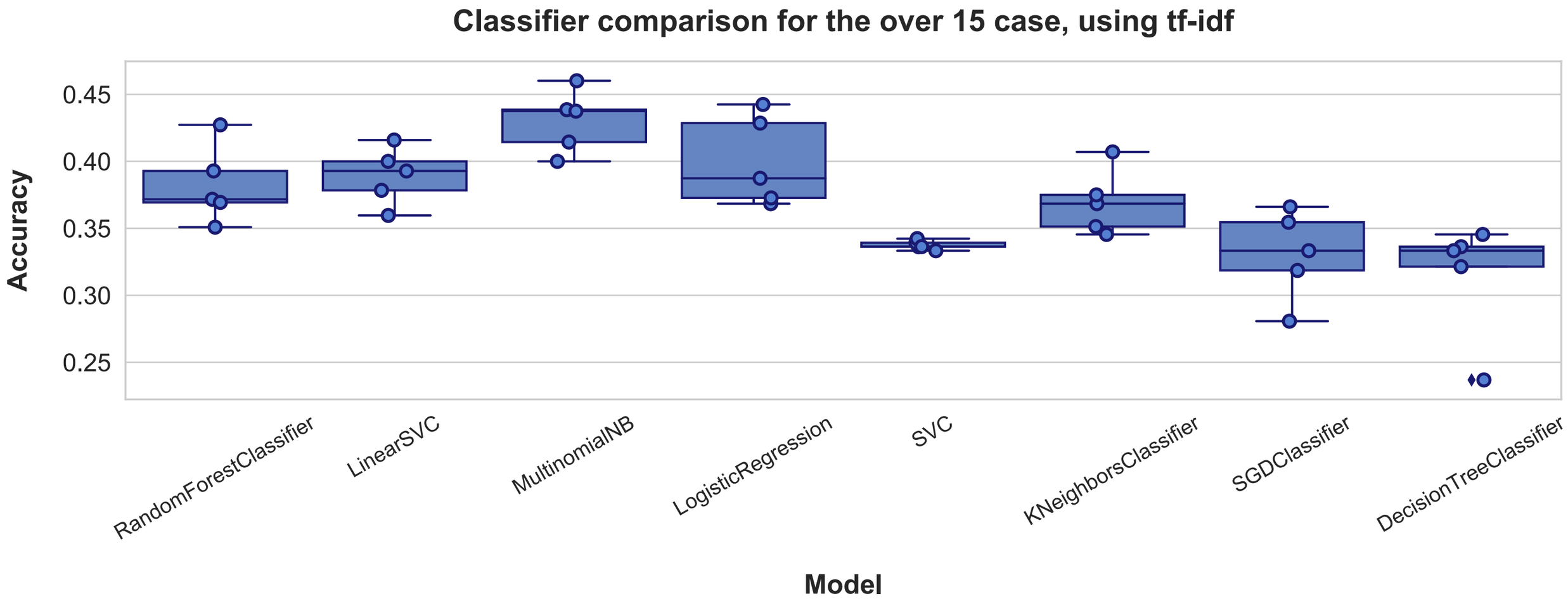

Classifier comparison for the indep. case, using CountVectorizer
Accuracy
0.5
0.4
0.3
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

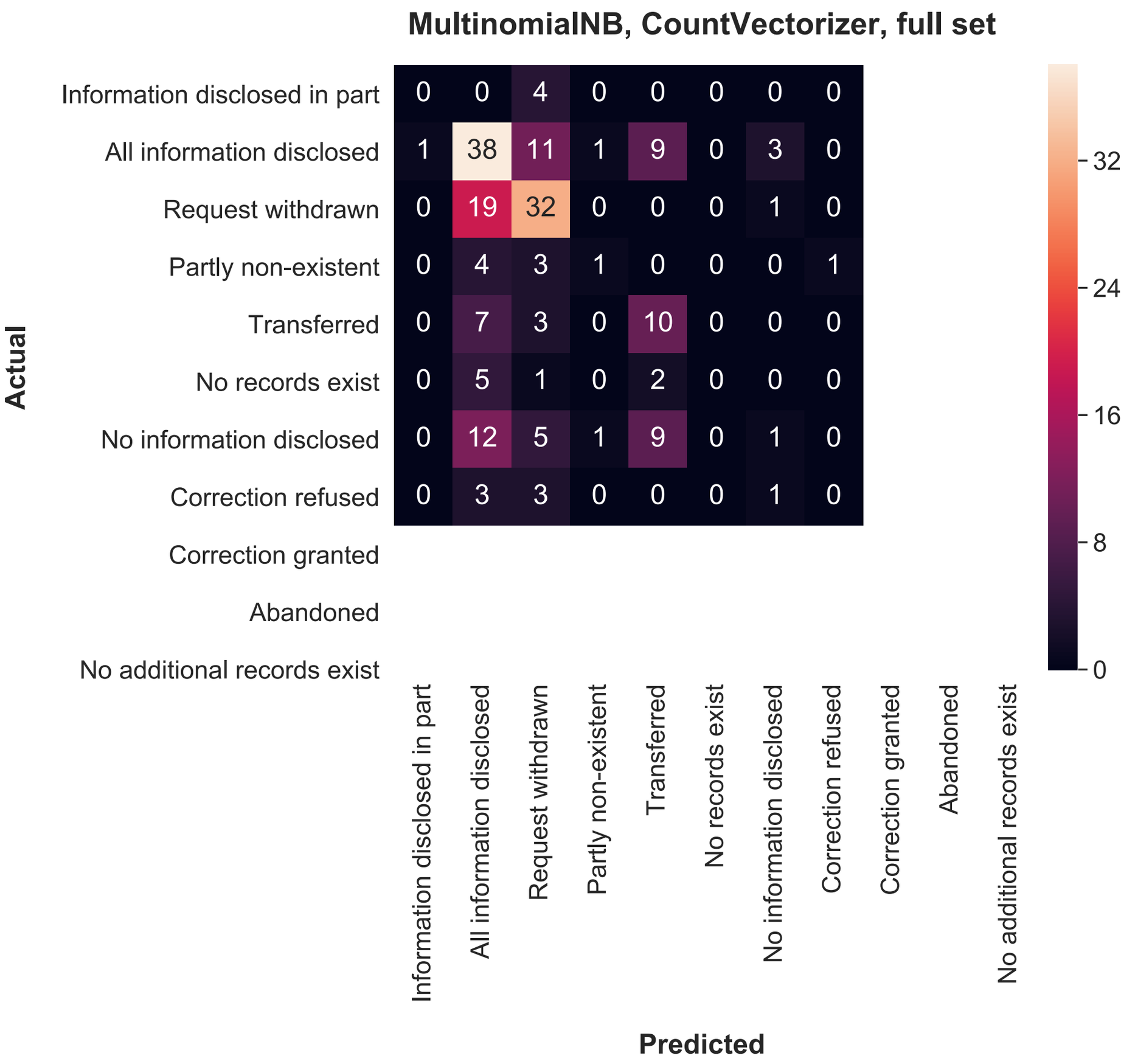

Classifier comparison for the indep. case, using tf-idf
Accuracy
0.6
0.5
0.4
0.3
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

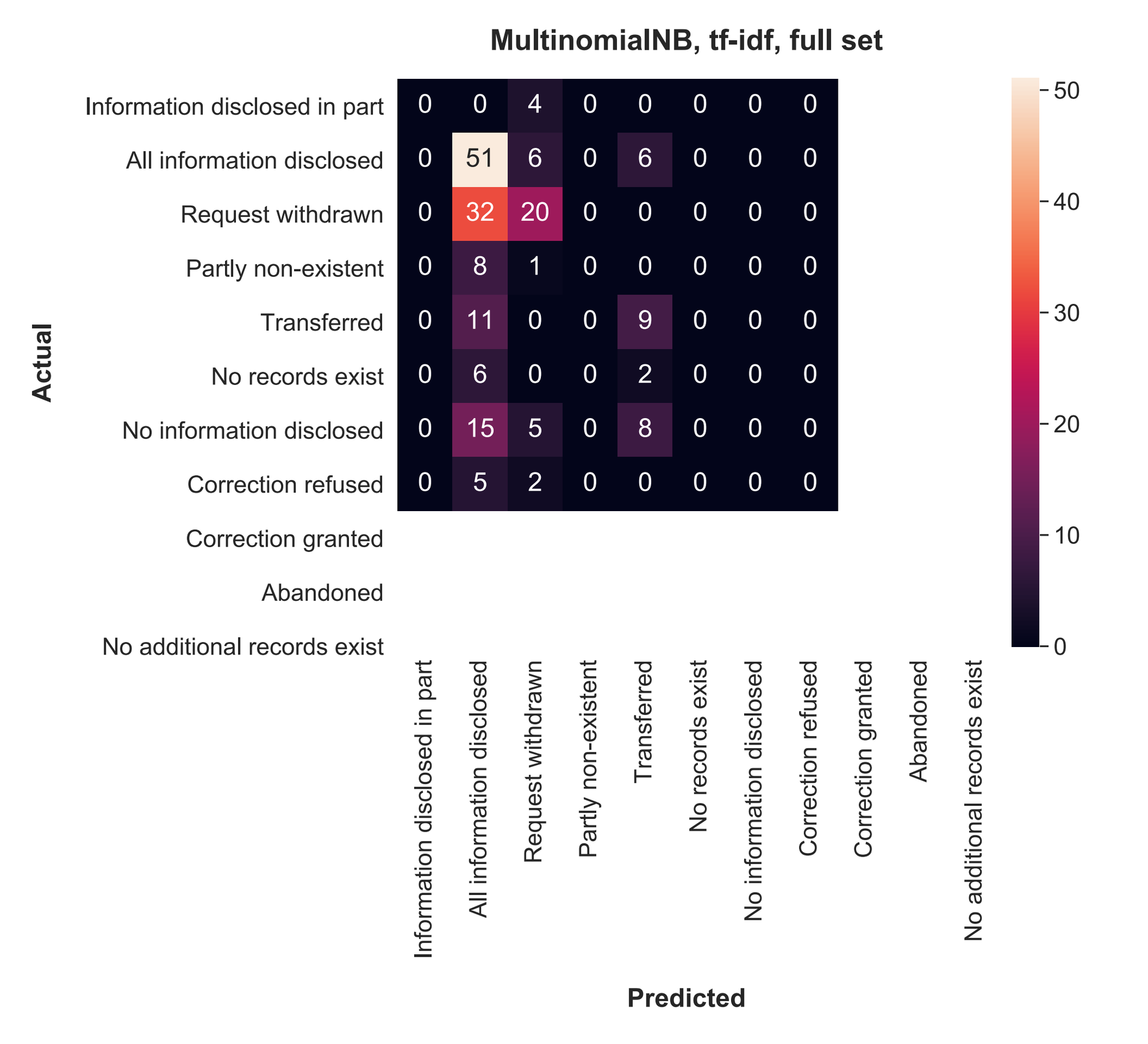

Classifier comparison for the over 15 case, using CountVectorizer, Balanced
Accuracy
0.4
0.3
0.2
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

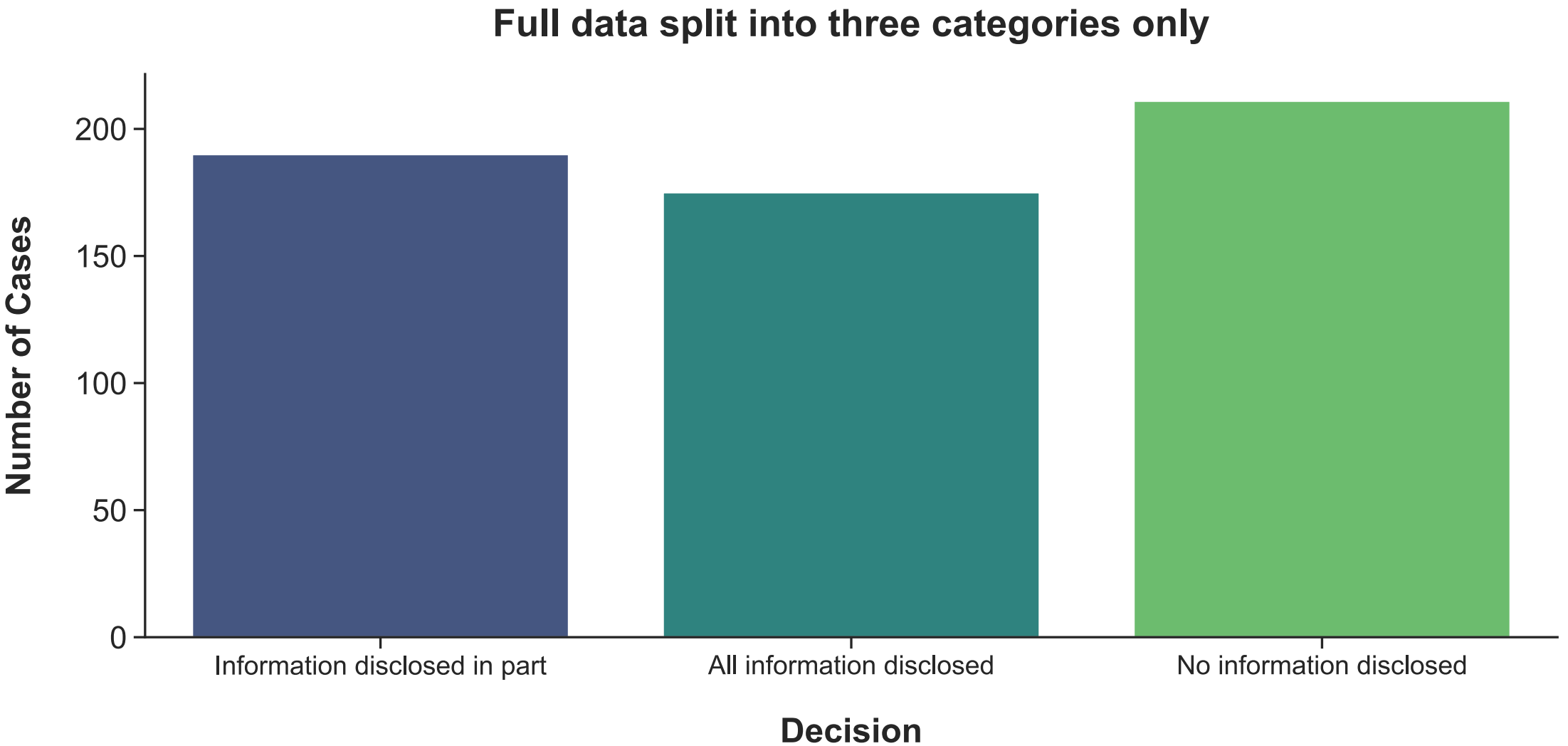

Classifier comparison for the over 15 case, using CountVectorizer
Accuracy
0.40
0.35
0.30
0.25
0.20
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

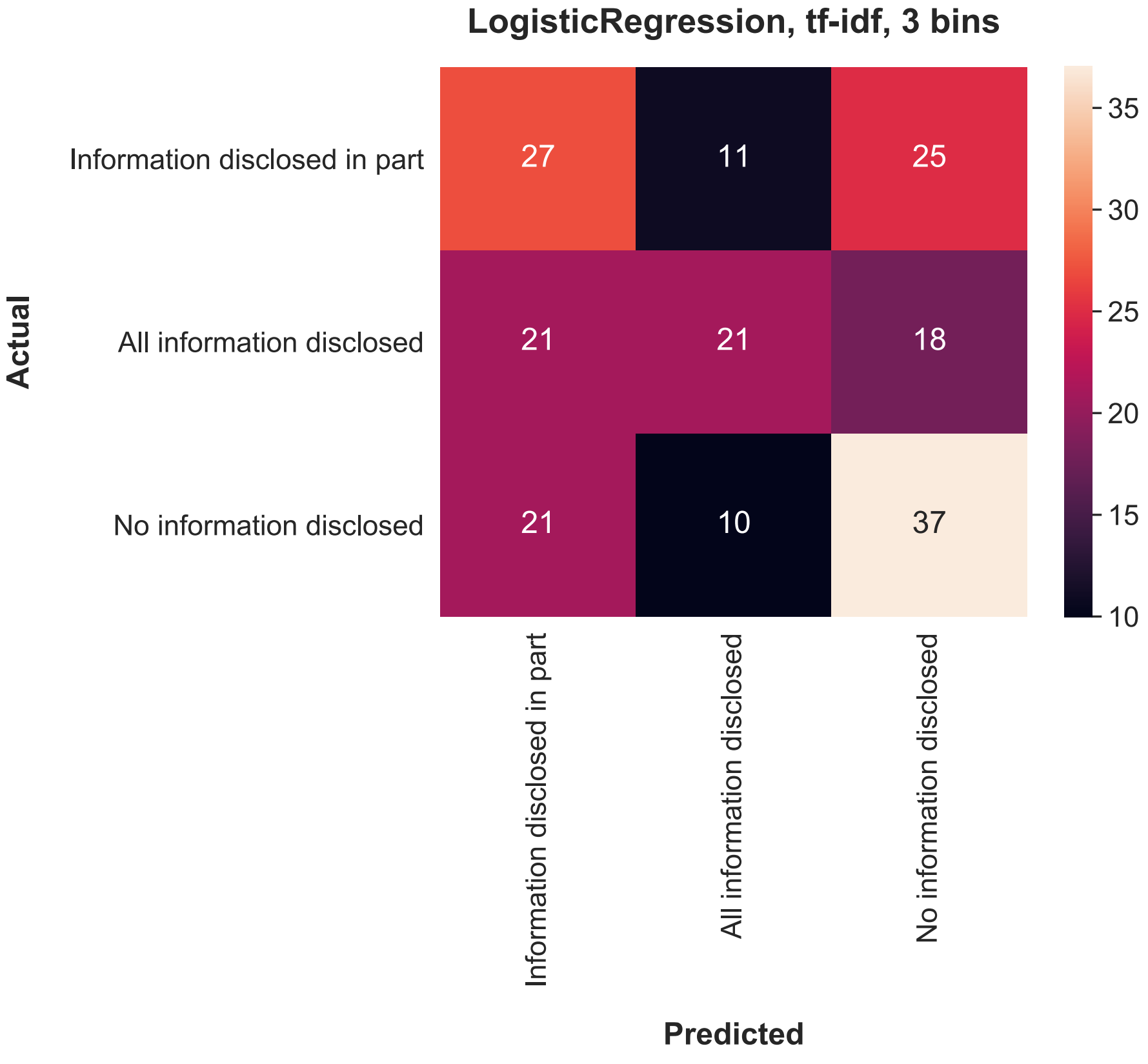

Classifier comparison for the over 15 case, using tf-idf, balanced
Accuracy
0.4
0.3
0.2
0.1
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

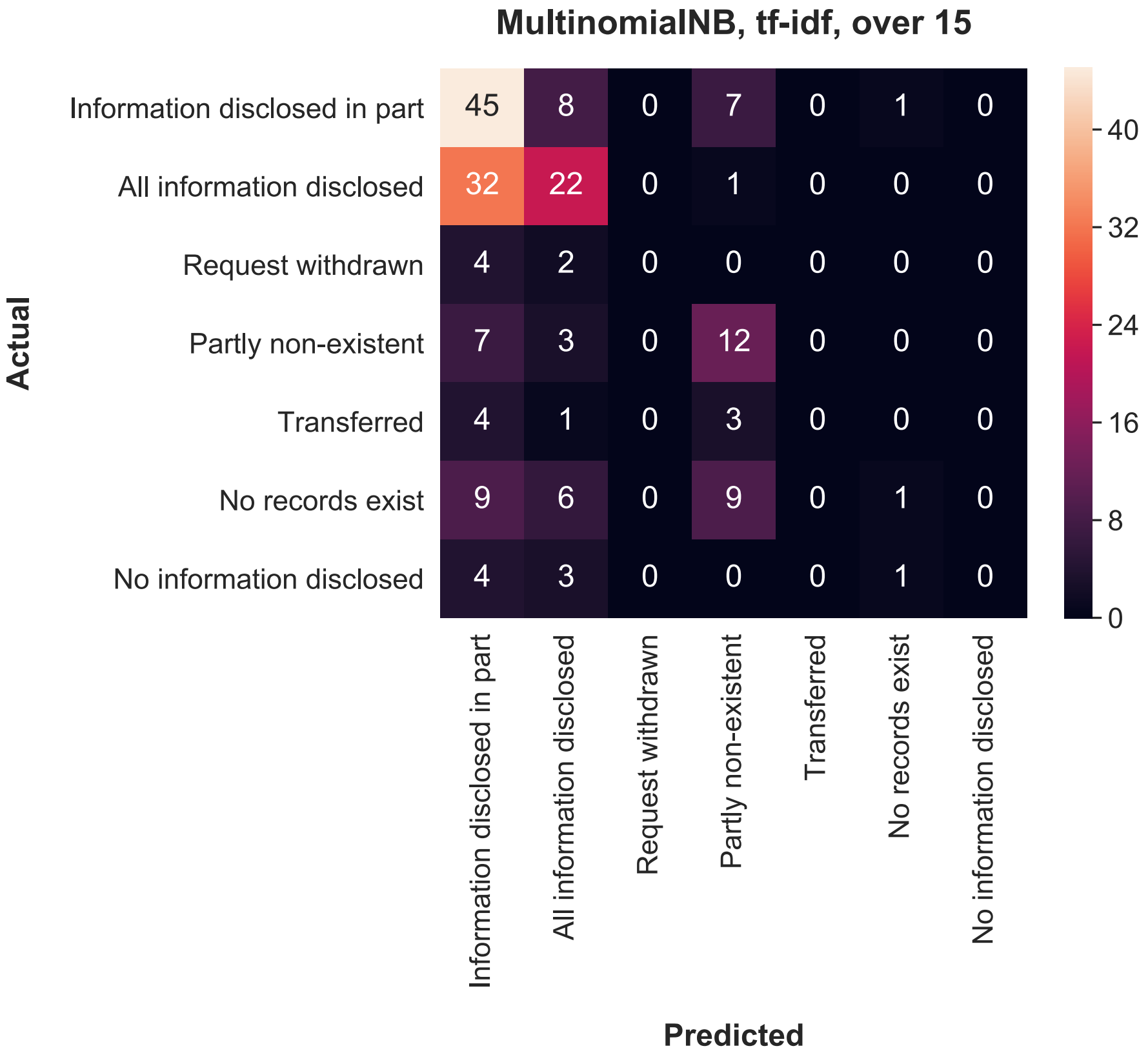

Classifier comparison for the over 15 case, using tf-idf
Accuracy
0.45
0.40
0.35
0.30
0.25
RandomForestClassifier
LinearSVC
MultinomialNB
LogisticRegression
SVC
KNeighborsClassifier
SGDClassifier
DecisionTreeClassifier
Model

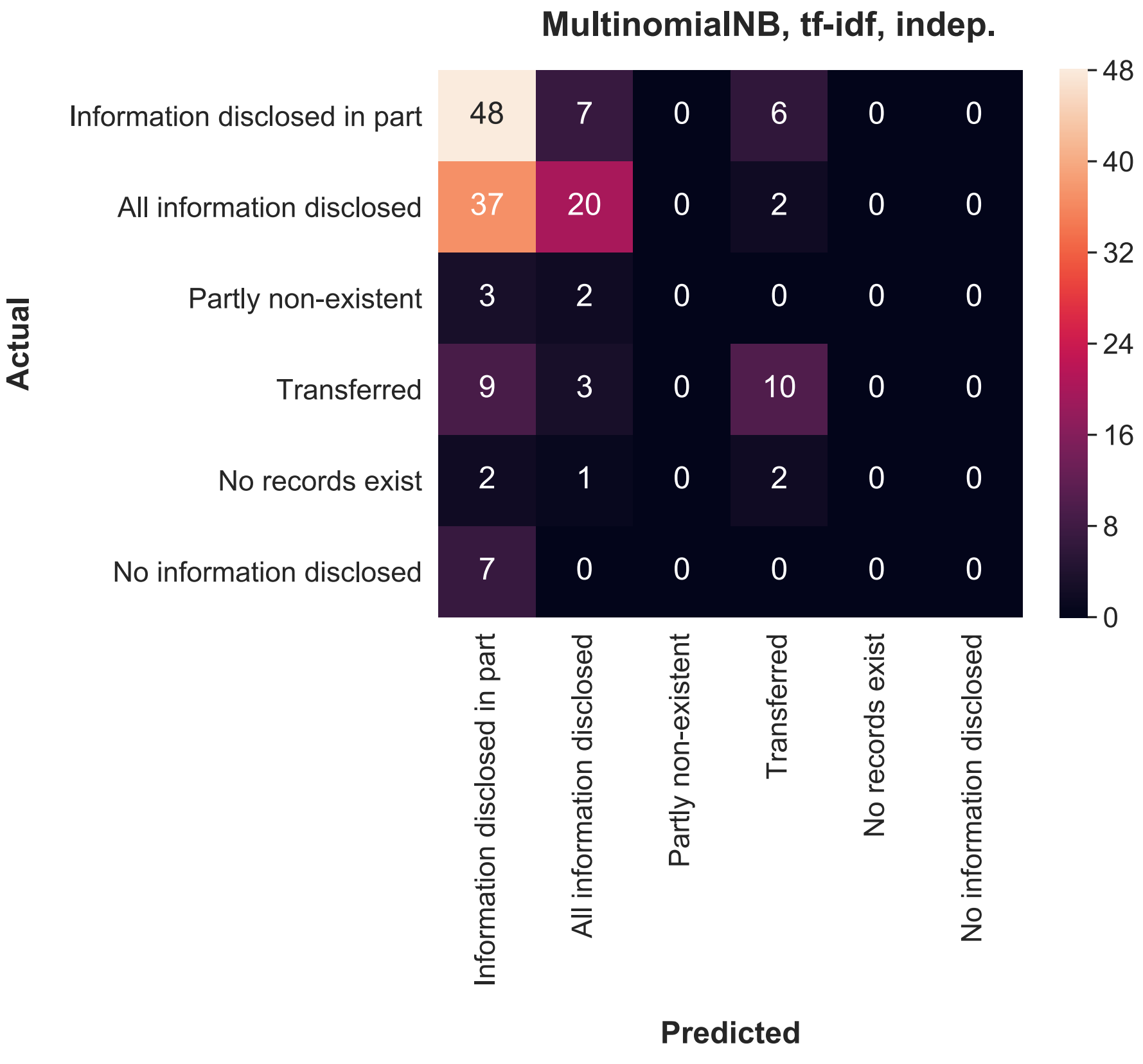

MultinomialNB, CountVectorizer, full set
Actual
Predicted
Information disclosed in part
All information disclosed
Request withdrawn
Partly non-existent
Transferred
No records exist
No information disclosed
Correction refused
Correction granted
Abandoned
No additional records exist
0 0 4 0 0 0 0 0
1 38 11 1 9 0 3 0
0 19 32 0 0 0 1 0
0 4 3 1 0 0 0 1
0 7 3 0 10 0 0 0
0 5 1 0 2 0 0 0
0 12 5 1 9 0 1 0
0 3 3 0 0 0 1 0
32
24
16
8
0


MultinomialNB, tf-idf, full set
Actual
Predicted
Information disclosed in part
All information disclosed
Request withdrawn
Partly non-existent
Transferred
No records exist
No information disclosed
Correction refused
Correction granted
Abandoned
No additional records exist
0 0 4 0 0 0 0 0
0 51 6 0 6 0 0 0
0 32 20 0 0 0 0 0
0 8 1 0 0 0 0 0
0 11 0 0 9 0 0 0
0 6 0 0 2 0 0 0
0 15 5 0 8 0 0 0
0 5 2 0 0 0 0 0
50
40
30
20
10
0


Full data split into three categories only
Number of Cases
200
150
100
50
0
Information disclosed in part
All information disclosed
No information disclosed
Decision


LogisticRegression, tf-idf, 3 bins
Information disclosed in part
All information disclosed
No information disclosed
27
11
25
21
21
18
21
10
37
Actual
Information disclosed in part
All information disclosed
No information disclosed
Predicted
35
30
25
20
15
10


MultinomialNB, tf-idf, over 15
Actual
Information disclosed in part
All information disclosed
Request withdrawn
Partly non-existent
Transferred
No records exist
No information disclosed
45 8 0 7 0 1 0
32 22 0 1 0 0 0
4 2 0 0 0 0 0
7 3 0 12 0 0 0
4 1 0 3 0 0 0
9 6 0 9 0 1 0
4 3 0 0 0 1 0
40
32
24
16
8
0
Information disclosed in part
All information disclosed
Request withdrawn
Partly non-existent
Transferred
No records exist
No information disclosed
Predicted


MultinomialNB, tf-idf, indep.
Actual
Information disclosed in part
All information disclosed
Partly non-existent
Transferred
No records exist
No information disclosed
48 7 0 6 0 0
37 20 0 2 0 0
3 2 0 0 0 0
9 3 0 10 0 0
2 1 0 2 0 0
7 0 0 0 0 0
Information disclosed in part
All information disclosed
Partly non-existent
Transferred
No records exist
No information disclosed
Predicted
48
40
32
24
16
8
0