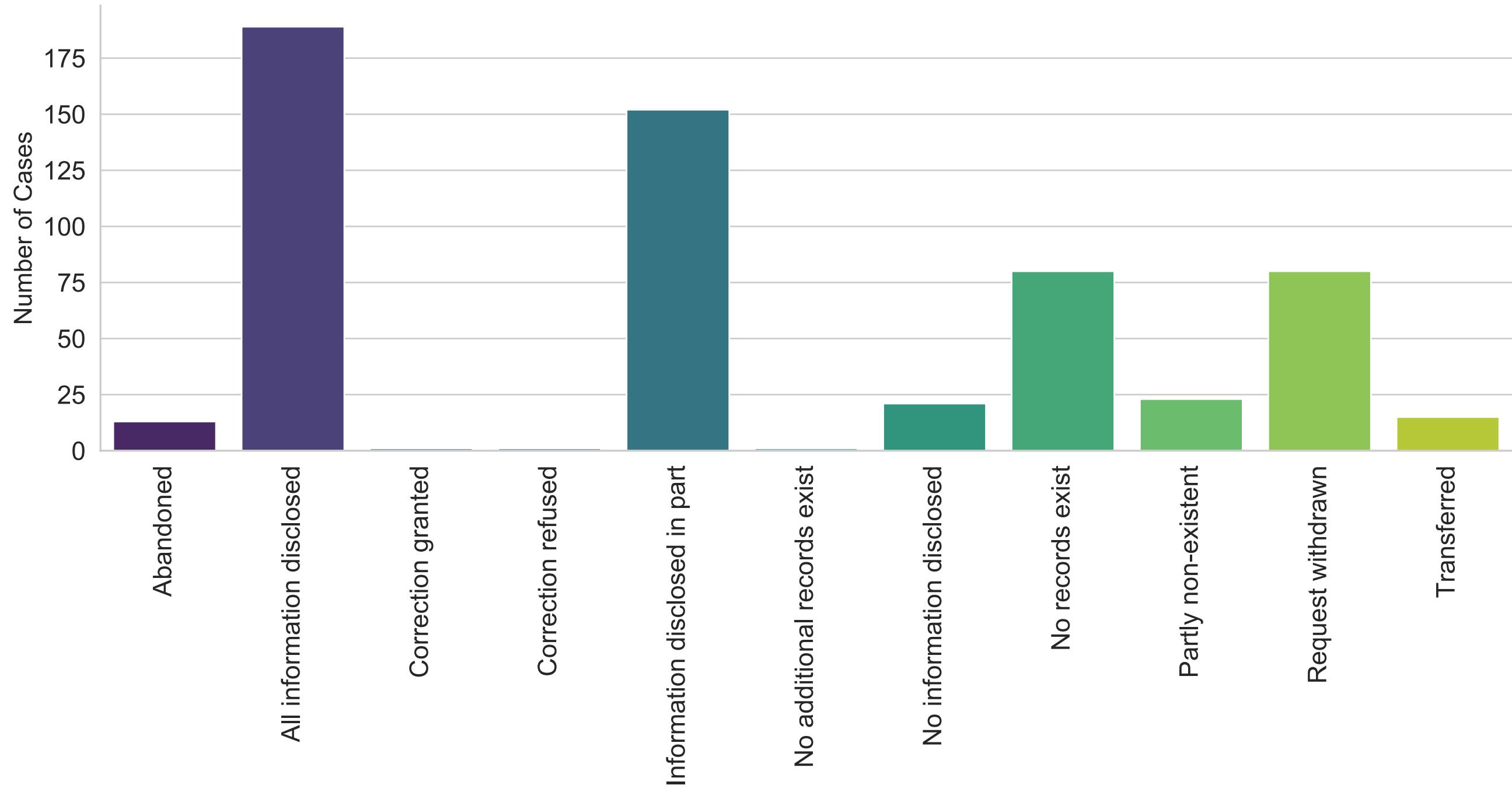
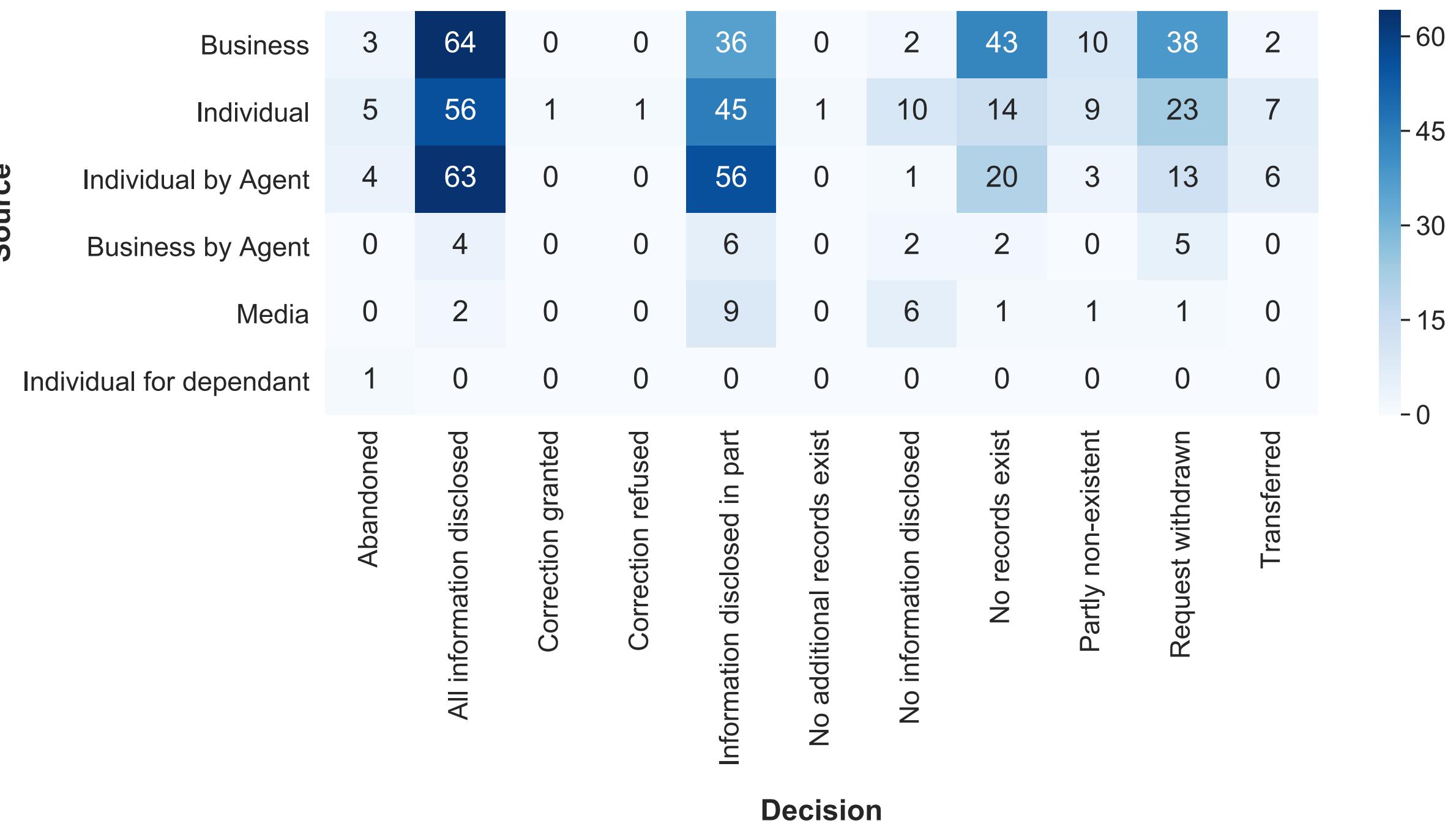


## Decisions Made for all Requests

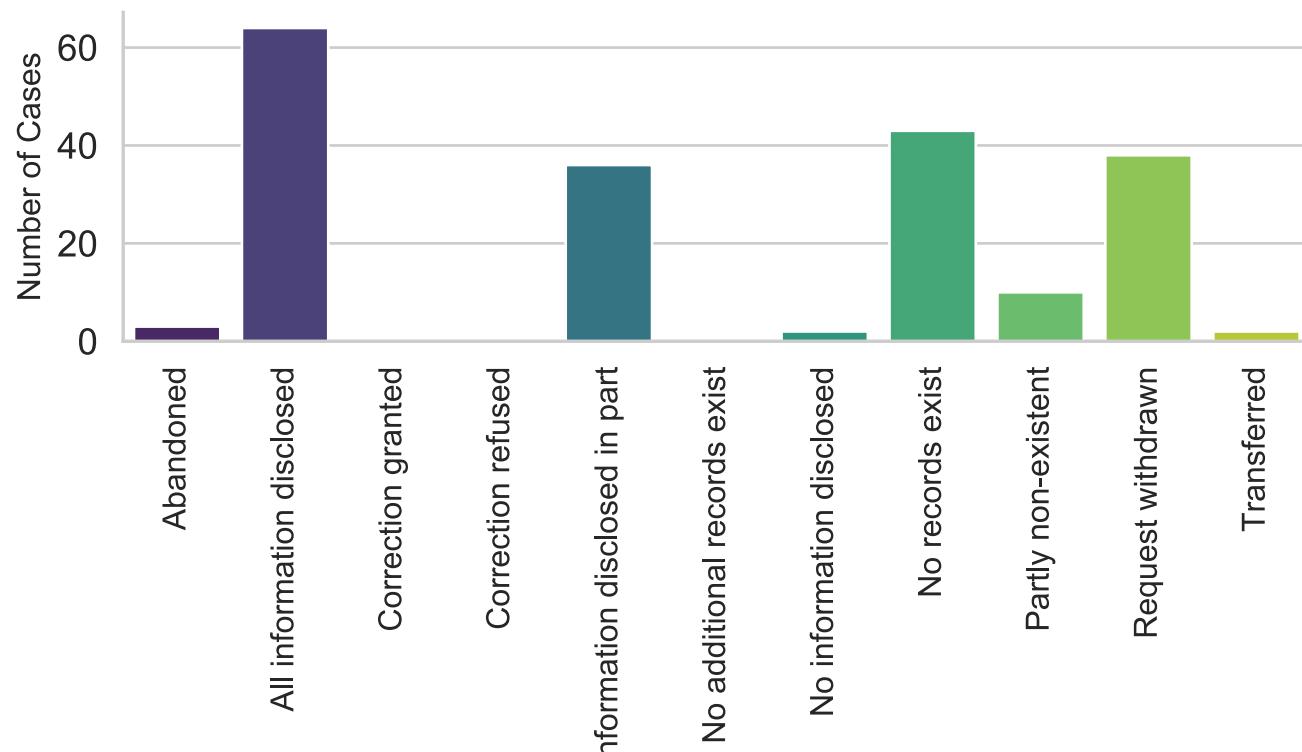


# Full Data

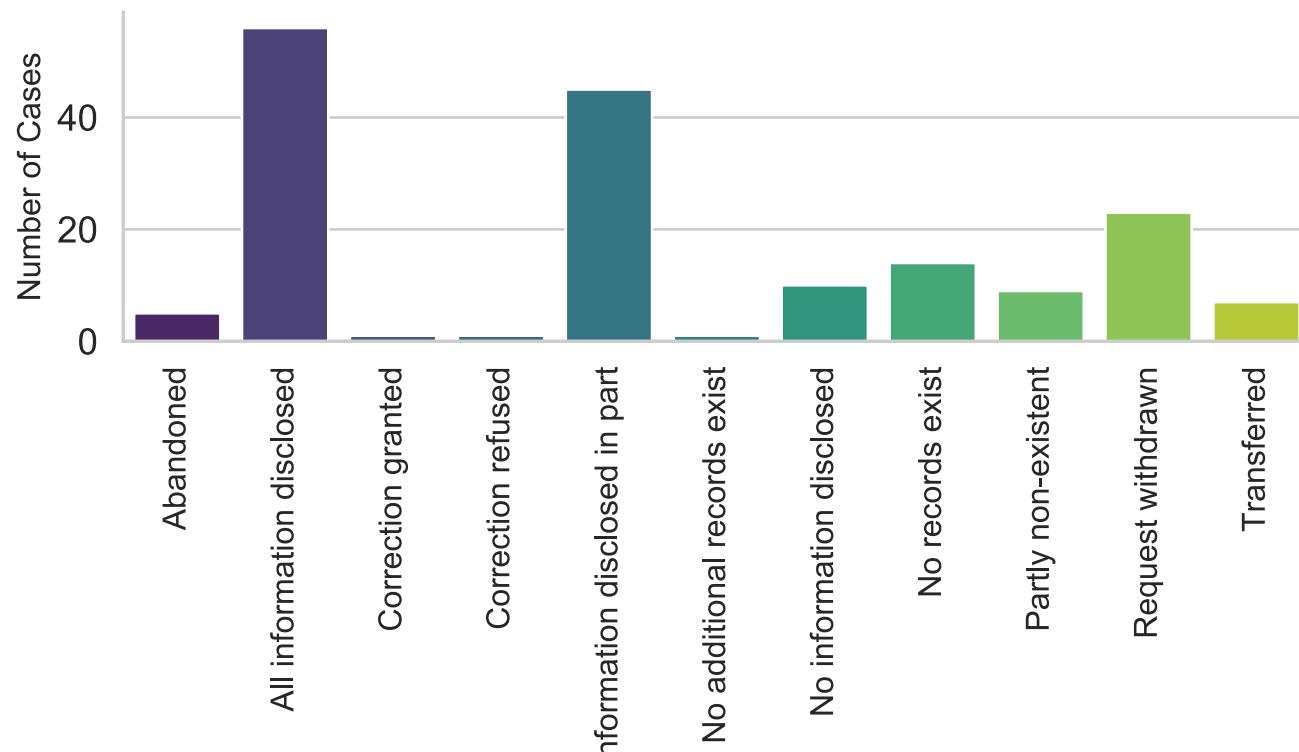


# Number of cases for all type of decisions made for each of the sources

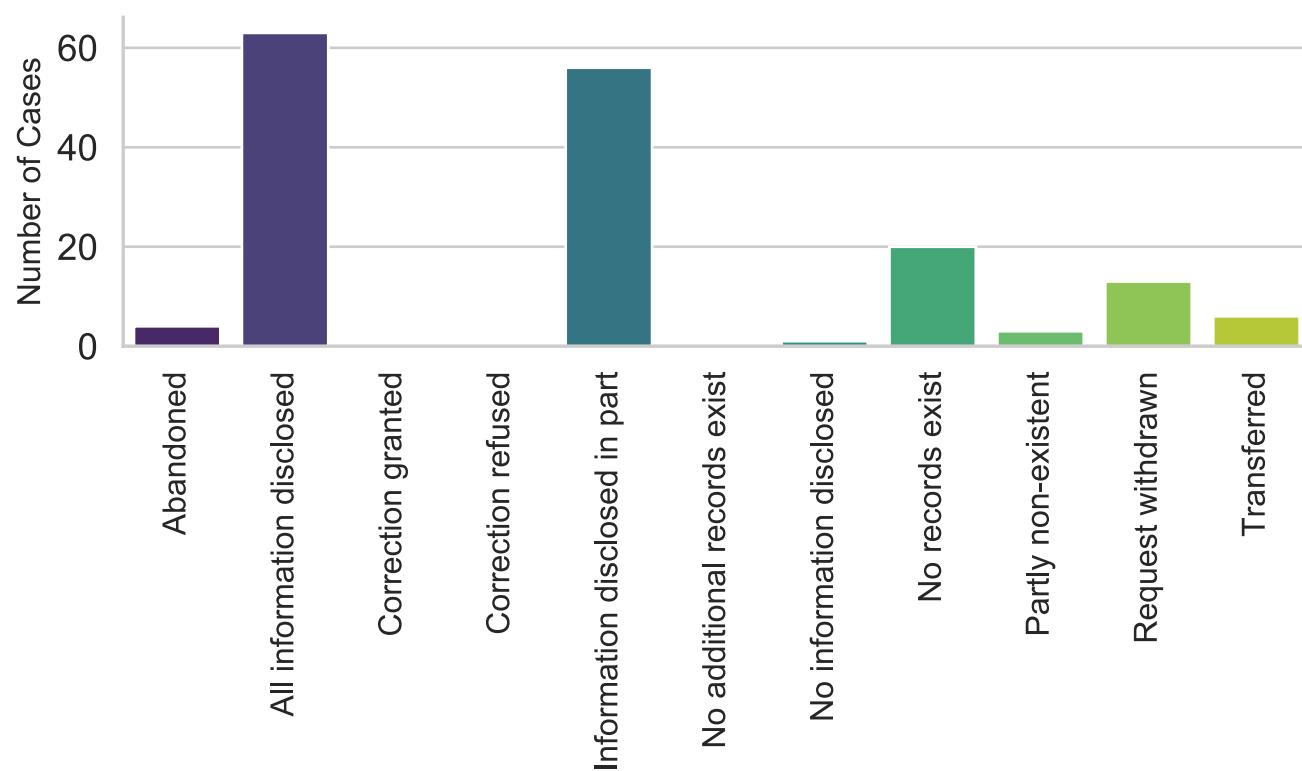
Requests made by 'Business'



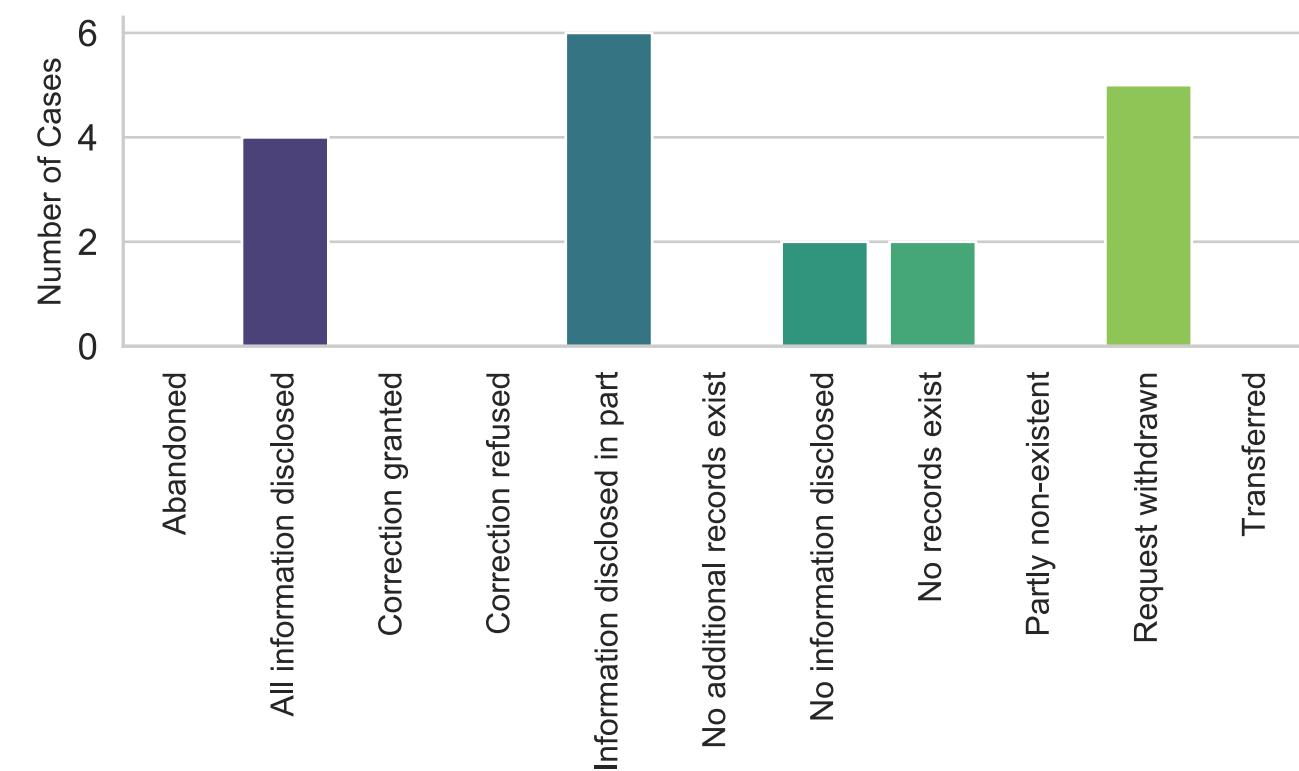
Requests made by 'Individual'



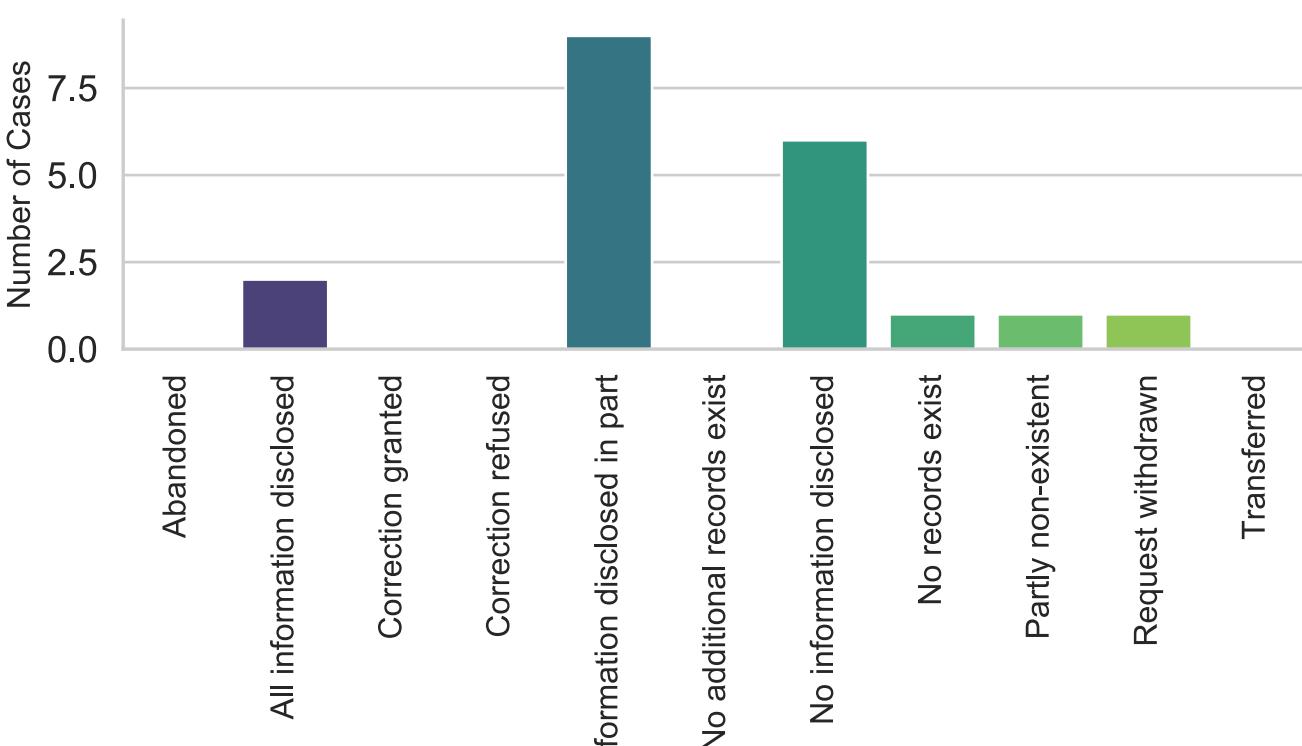
Requests made by 'Individual by Agent'



Requests made by 'Business by Agent'



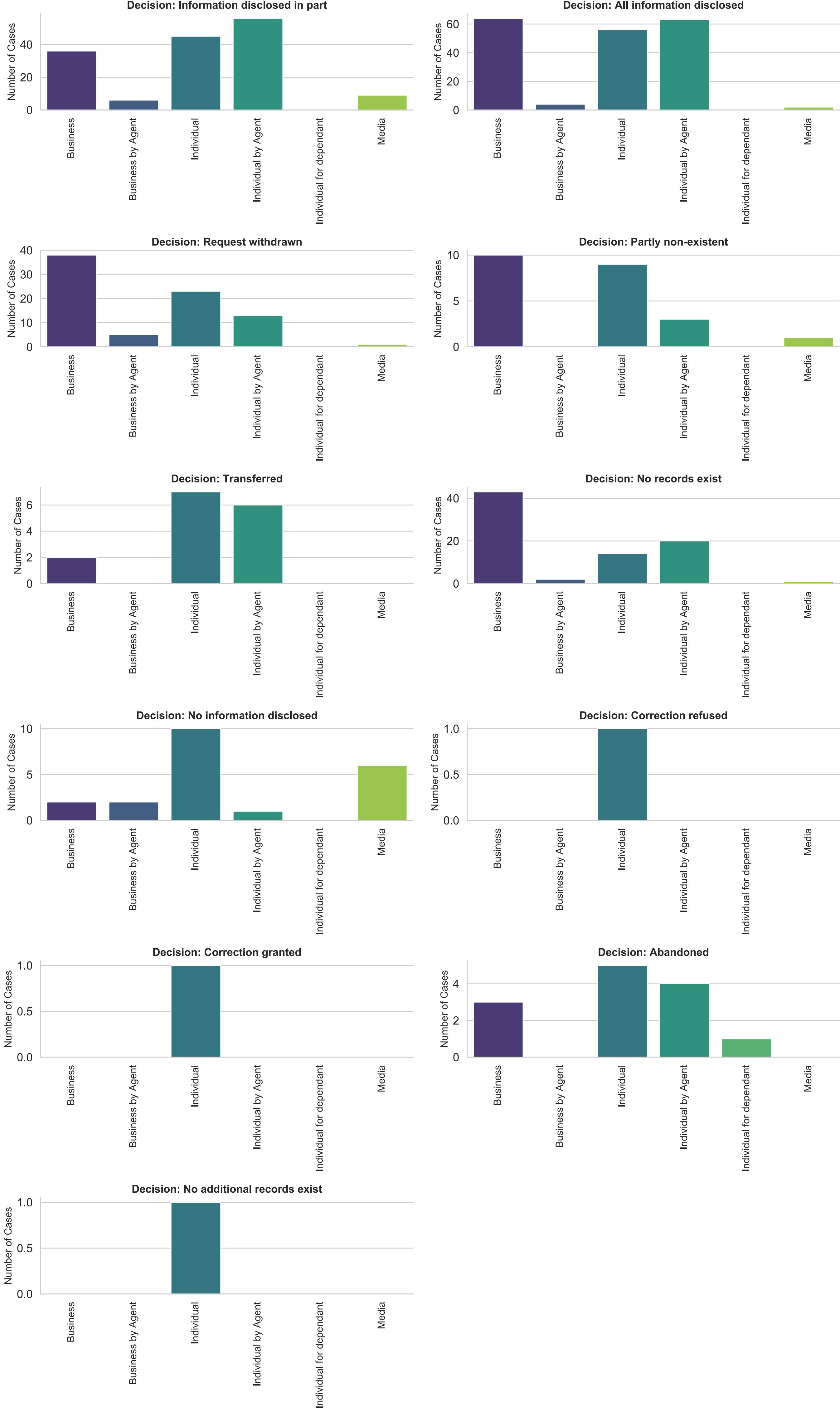
Requests made by 'Media'



Requests made by 'Individual for dependant'

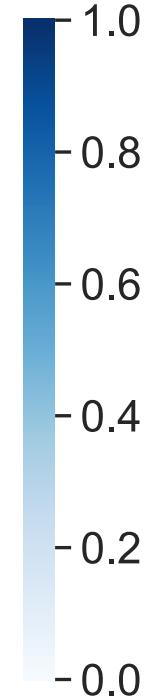


## Number of cases for each type of decision made by sources



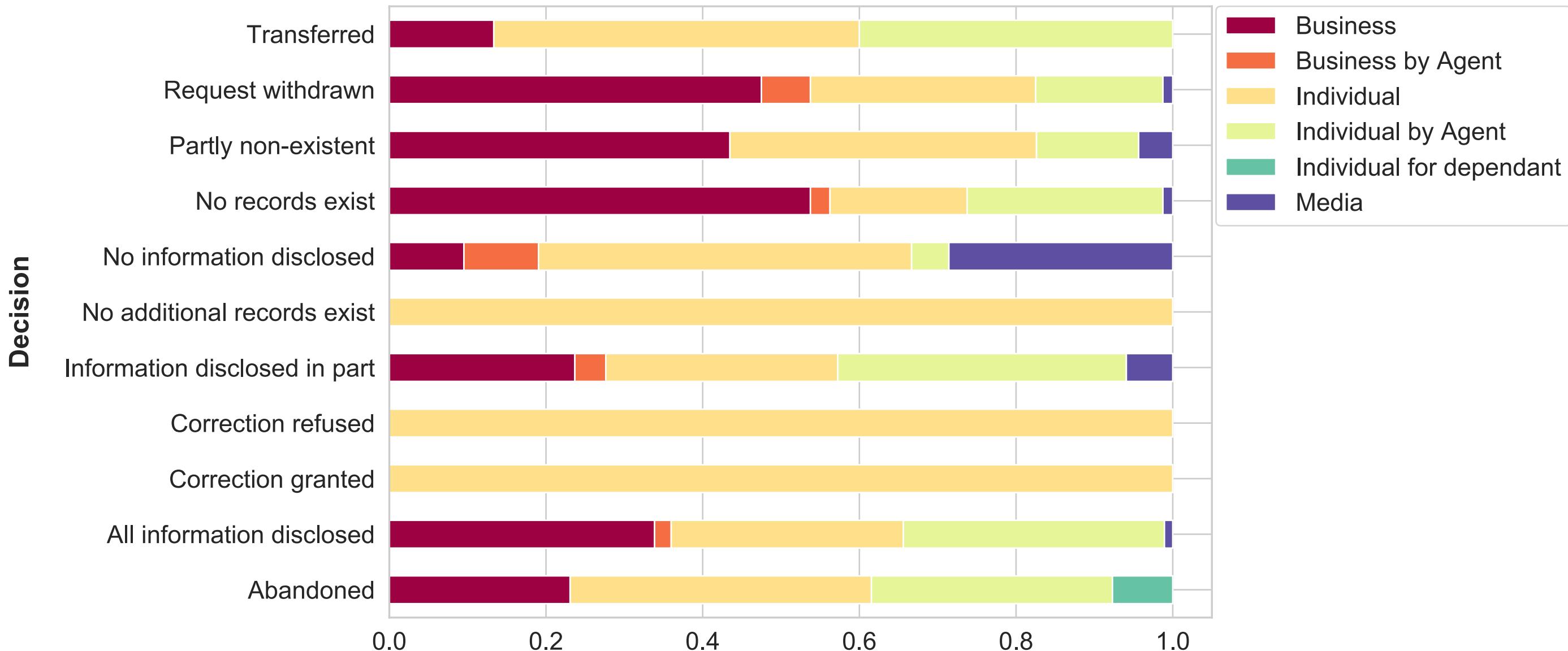
## How each decision is split among all sources (fraction)

Source	Decision										
	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	0.23	0.34	0	0	0.24	0	0.1	0.54	0.43	0.48	0.13
Individual	0.38	0.3	1	1	0.3	1	0.48	0.18	0.39	0.29	0.47
Individual by Agent	0.31	0.33	0	0	0.37	0	0.05	0.25	0.13	0.16	0.4
Business by Agent	0	0.02	0	0	0.04	0	0.1	0.02	0	0.06	0
Media	0	0.01	0	0	0.06	0	0.29	0.01	0.04	0.01	0
Individual for dependant	0.08	0	0	0	0	0	0	0	0	0	0

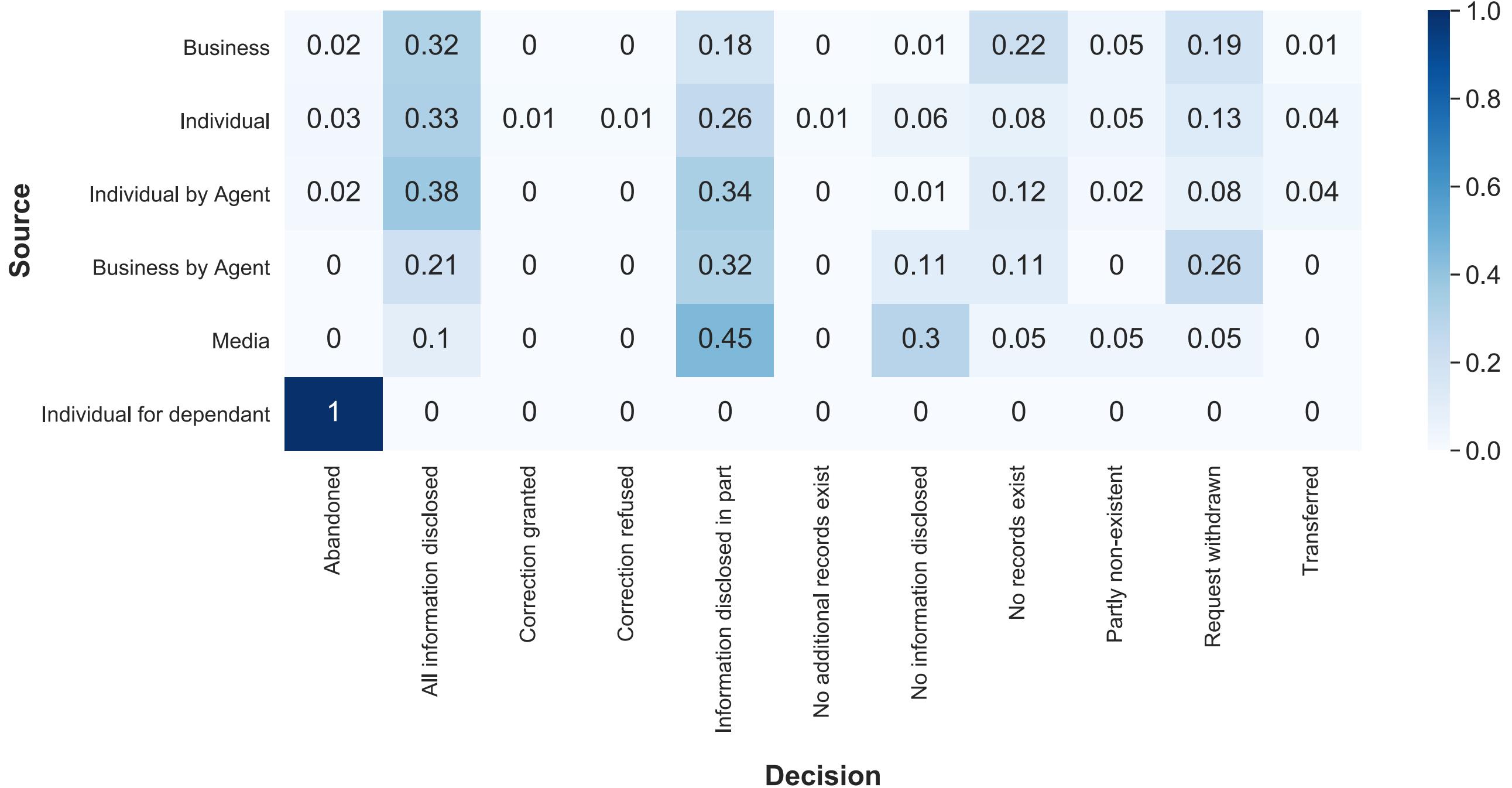


Decision

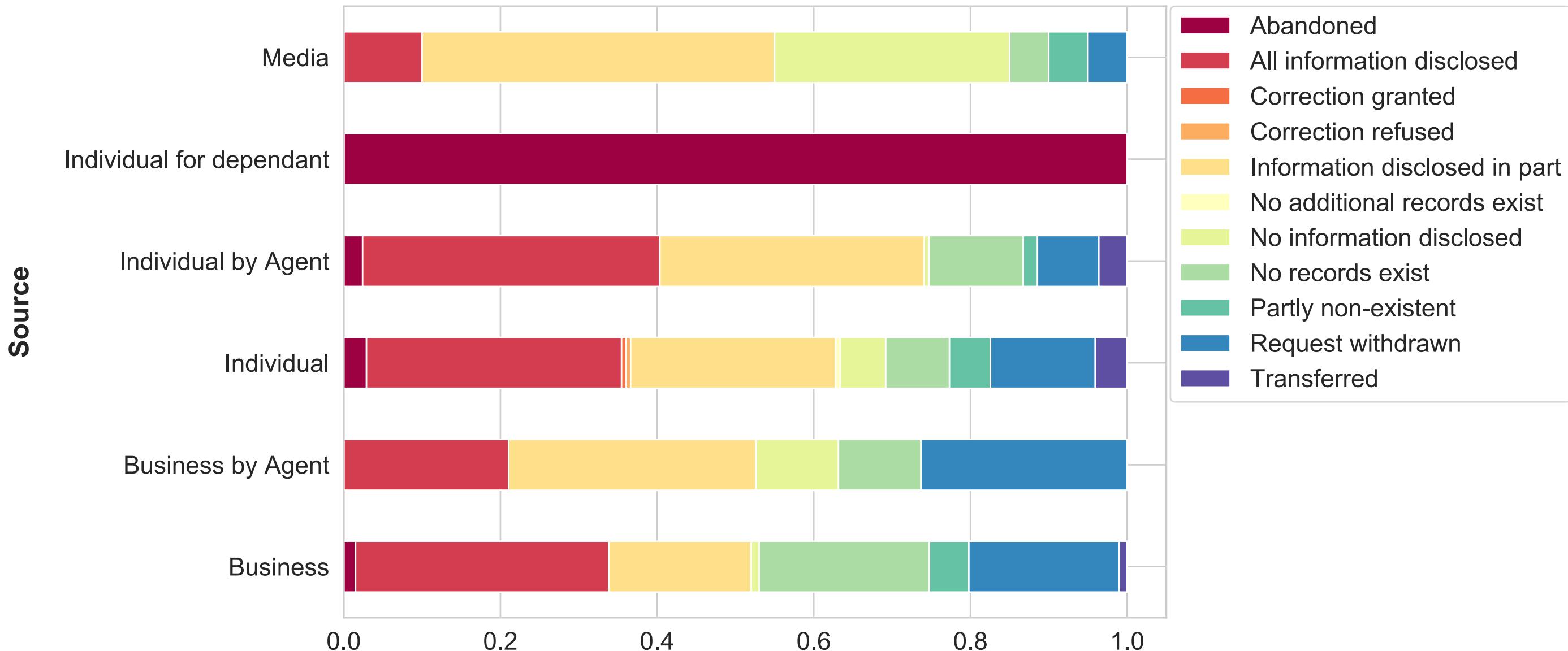
## Full data, how each decision is split per source



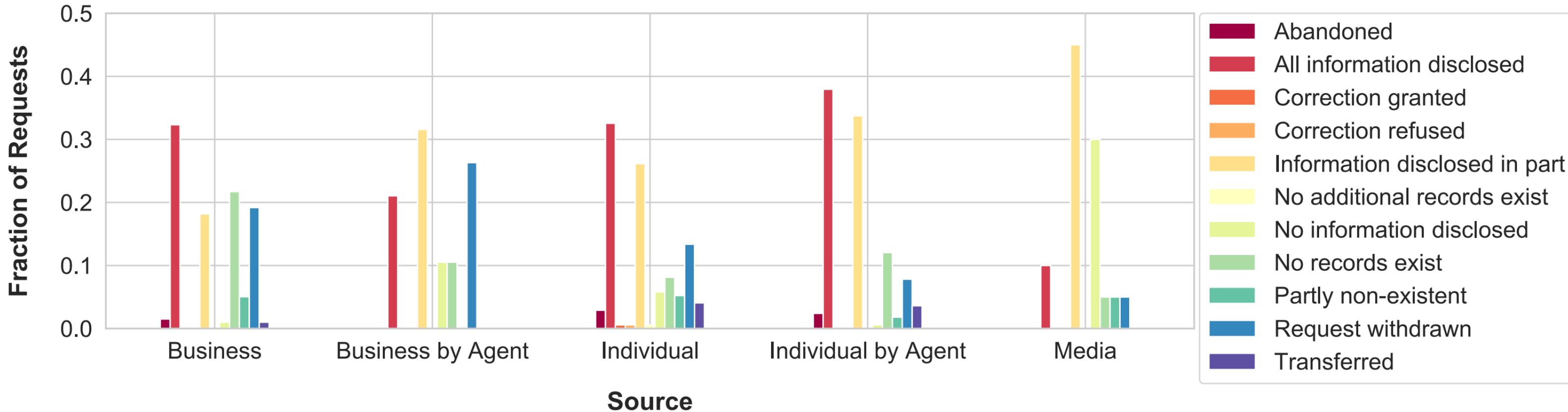
## How decisions are split per source (fraction)



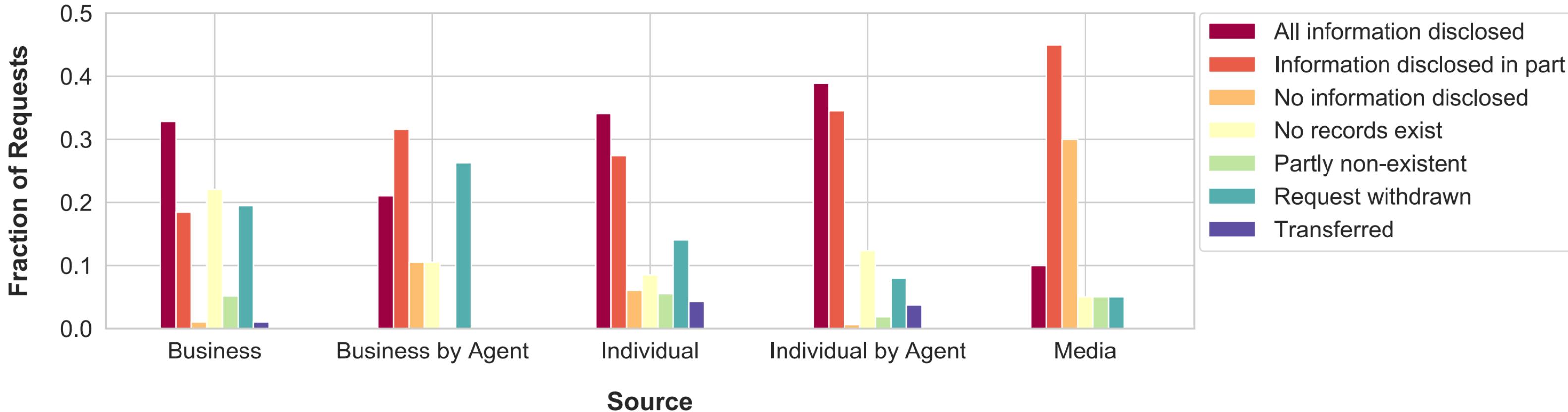
## Full data, fraction of decisions per source



## Full data, fraction of decisions per source



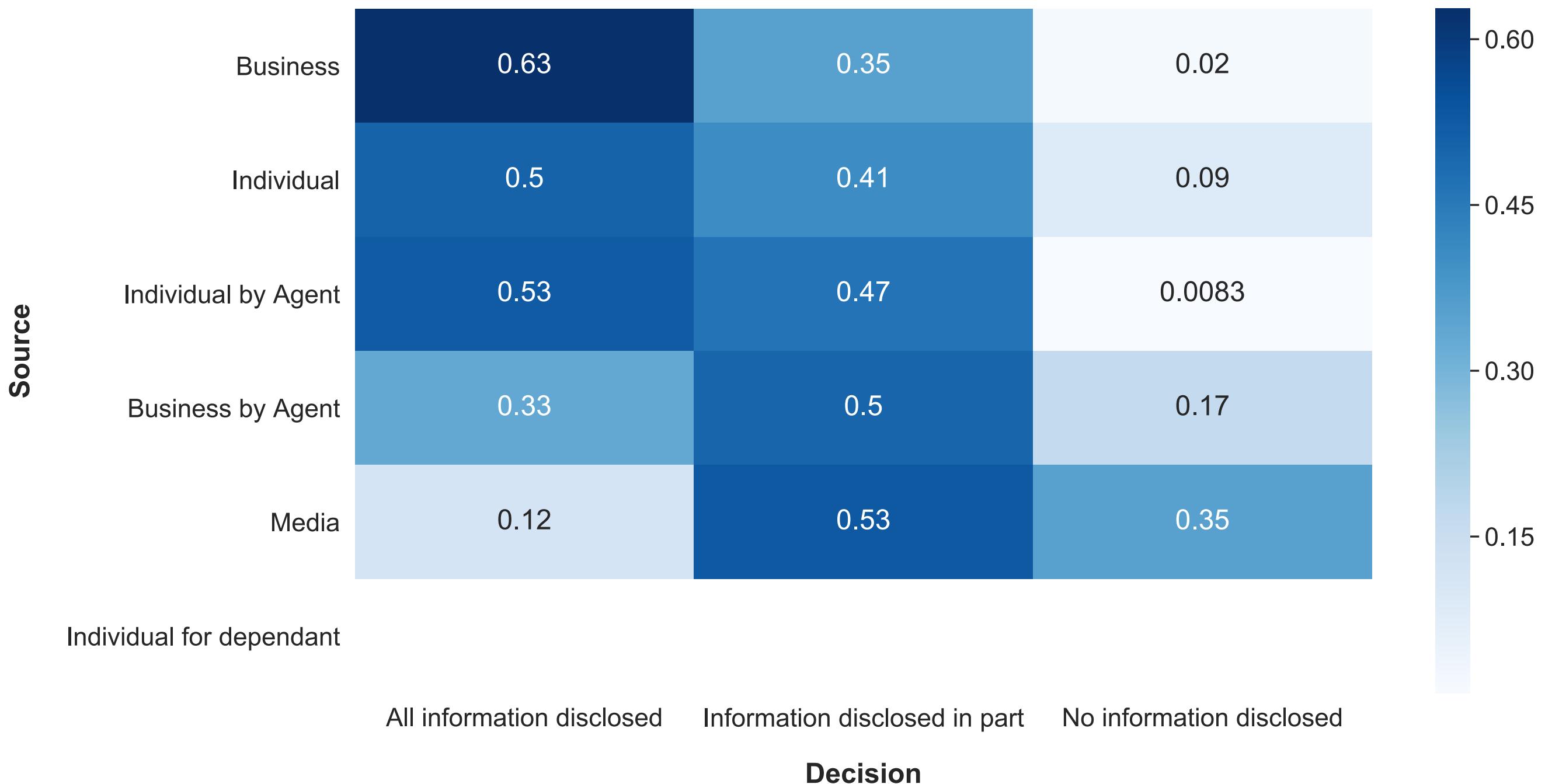
## Fraction of decisions per source, for decisions with more than 15 instances only



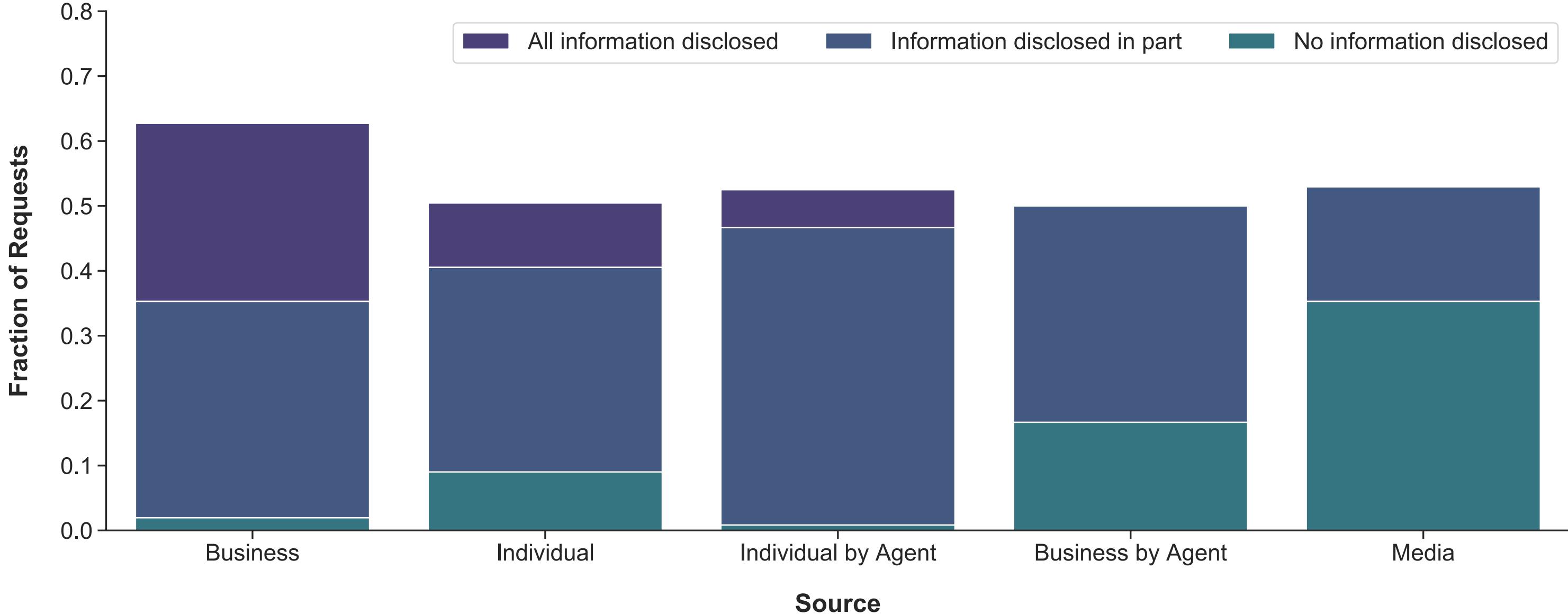
## Each of the three main decisions split among all the sources (fraction)



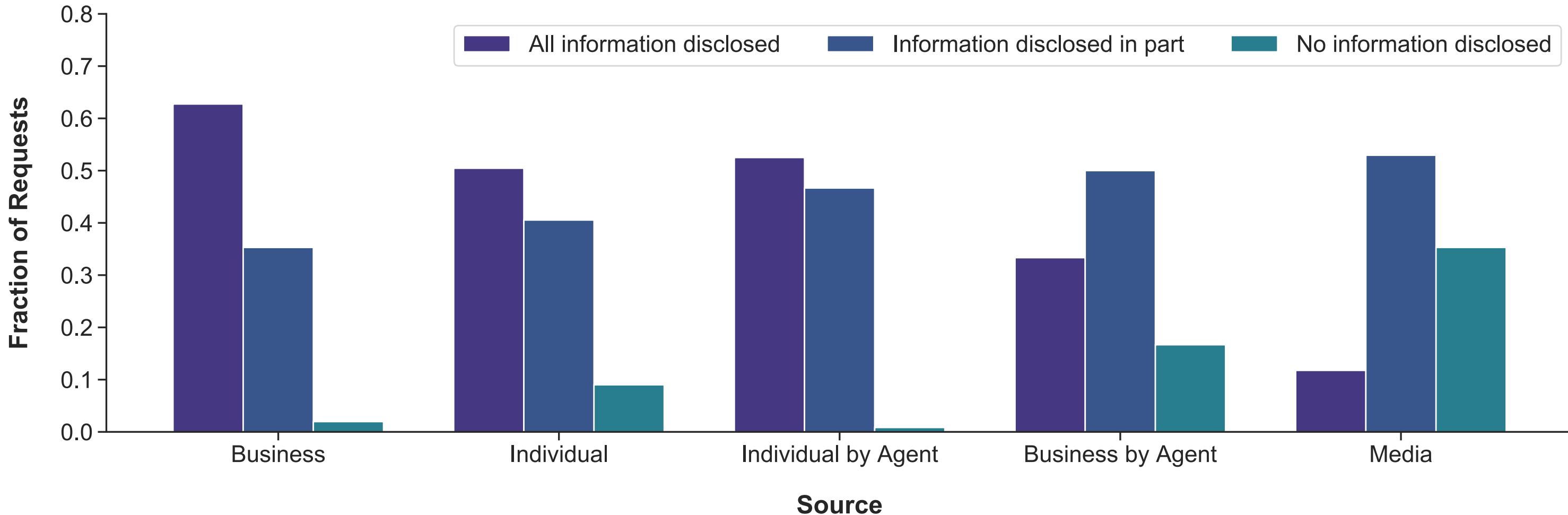
### Three main decisions split for each source (fraction)

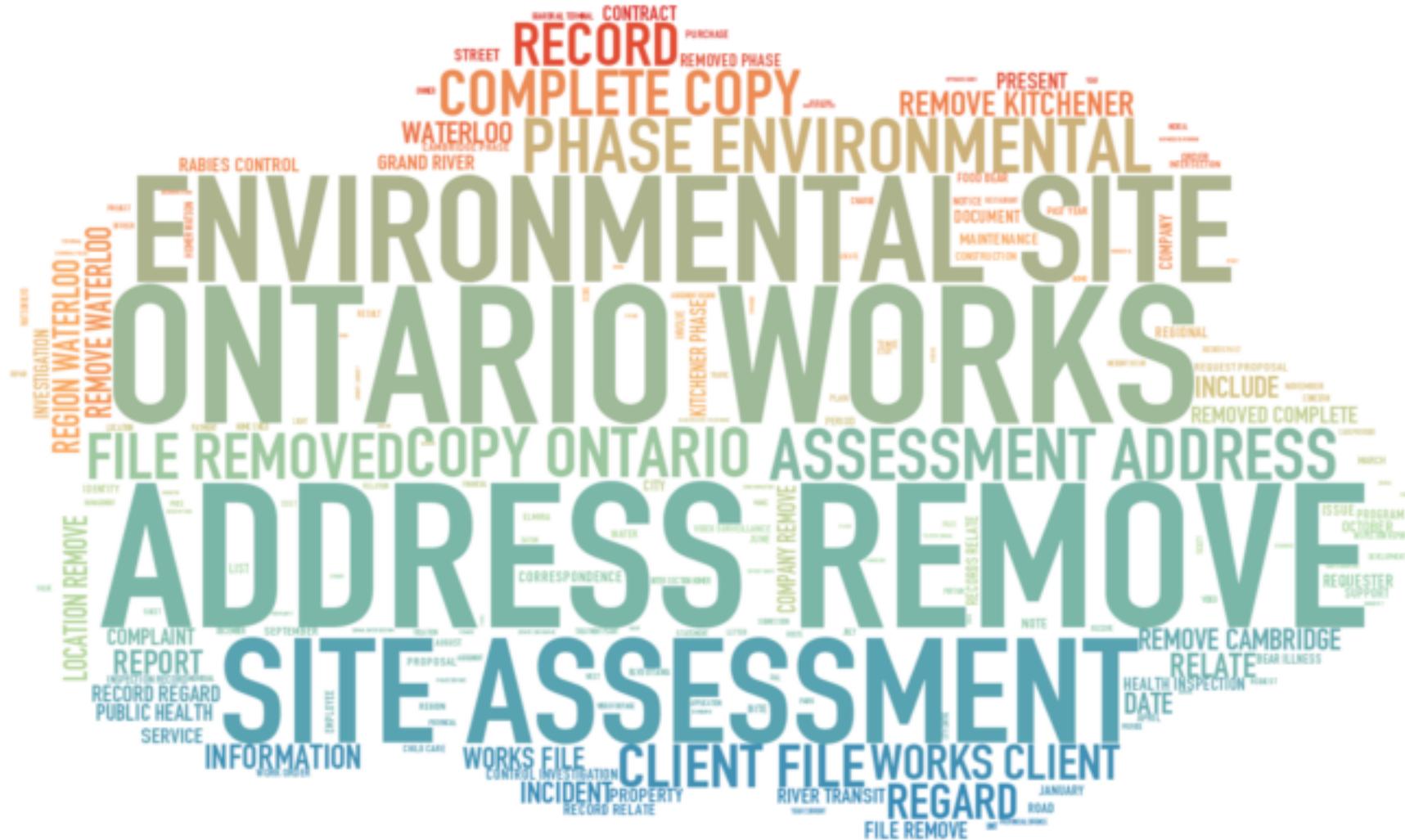


# Three main decisions only, fractions for each source add to 1



## Three main decisions only, fractions for each source add to 1





Top 200 unigrams/bigrams, full text



Top 200 unigrams, full text

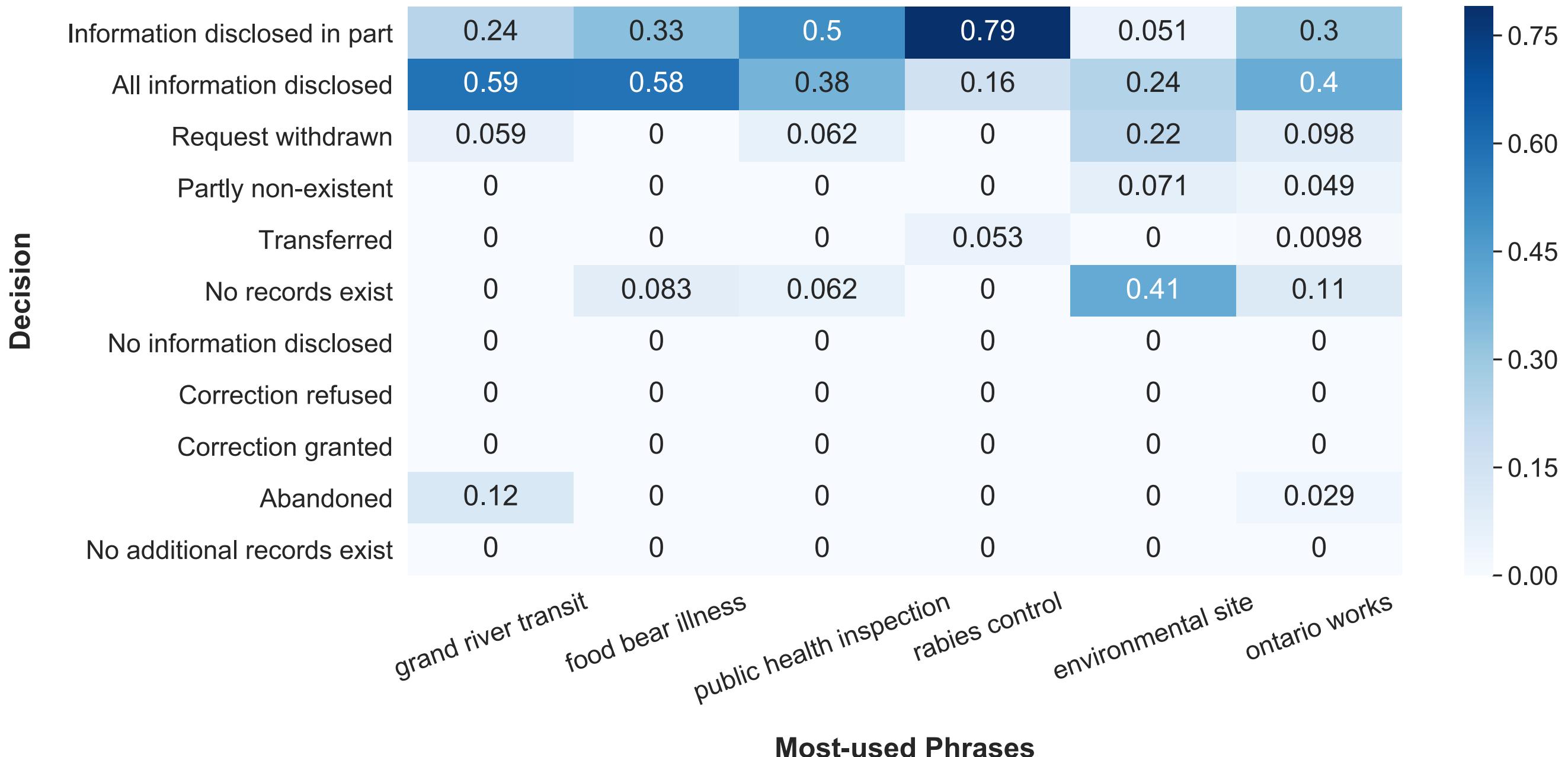


Top 200 unigrams/bigrams, full text without '{\* remove}'



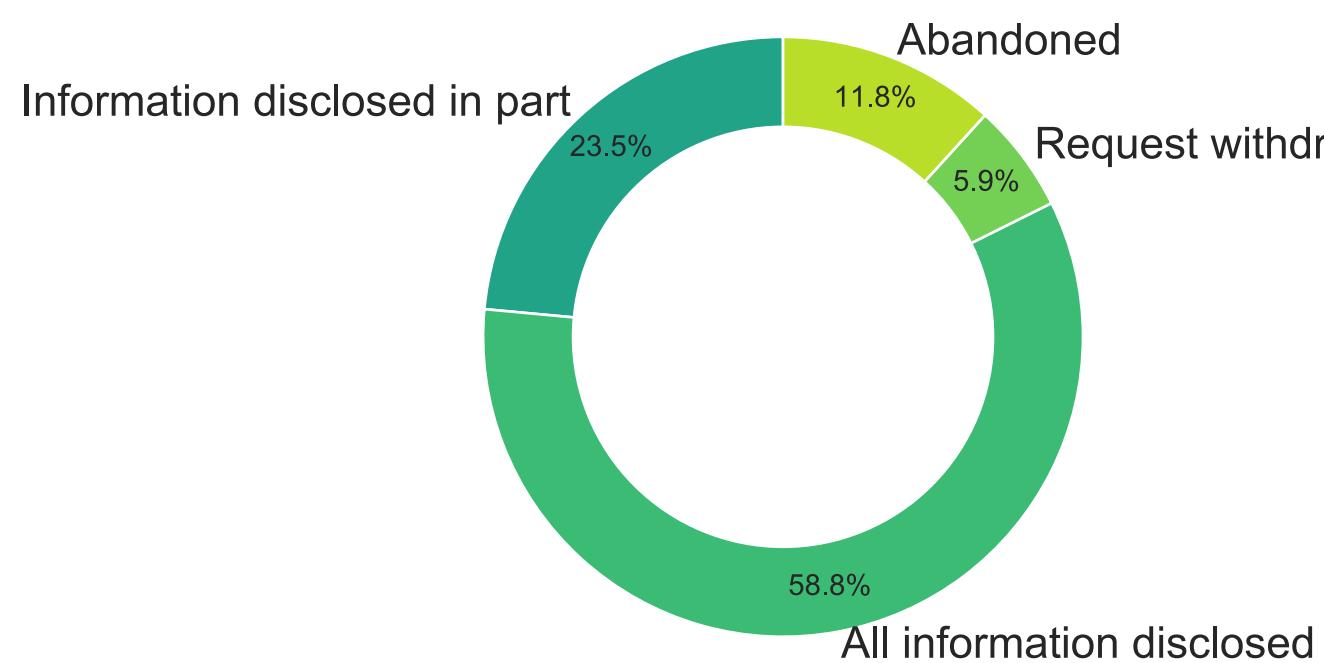
Top 200 unigrams, full text without '{\* remove}'

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

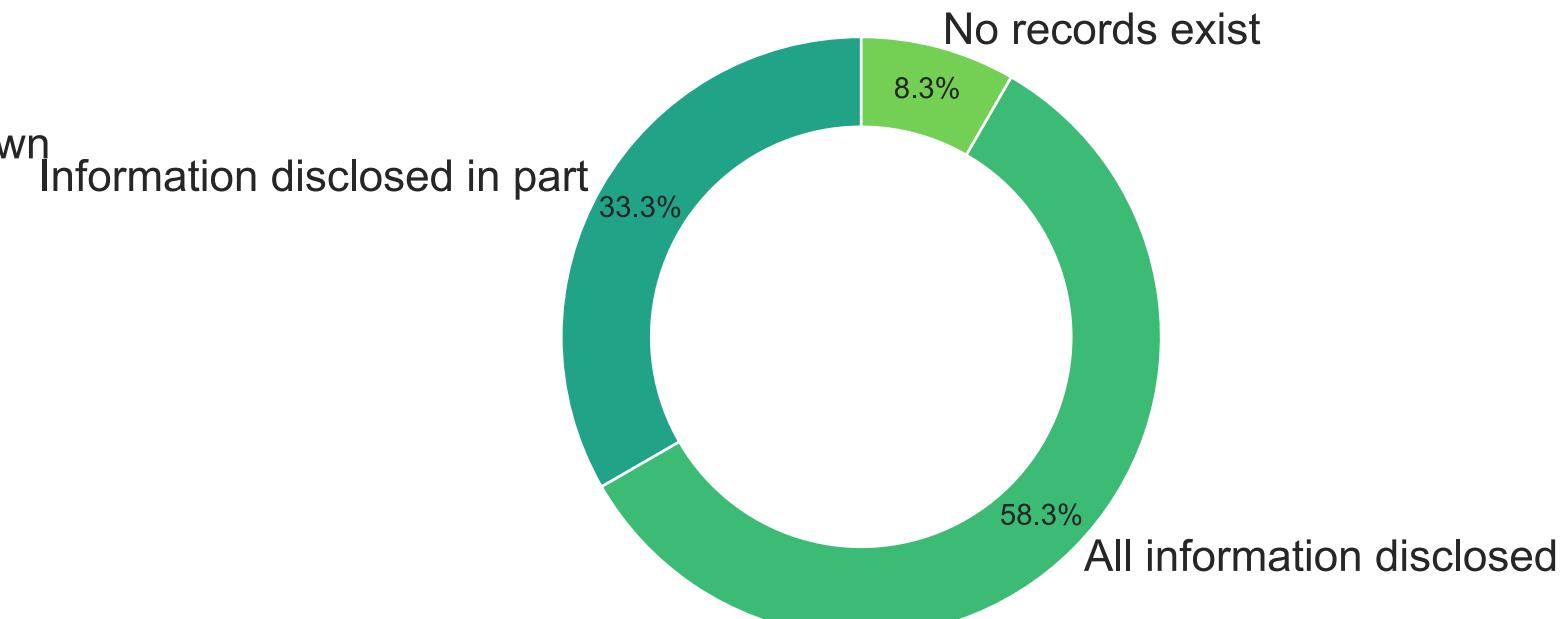


## Decision percentage for each n-gram

**grand river transit**

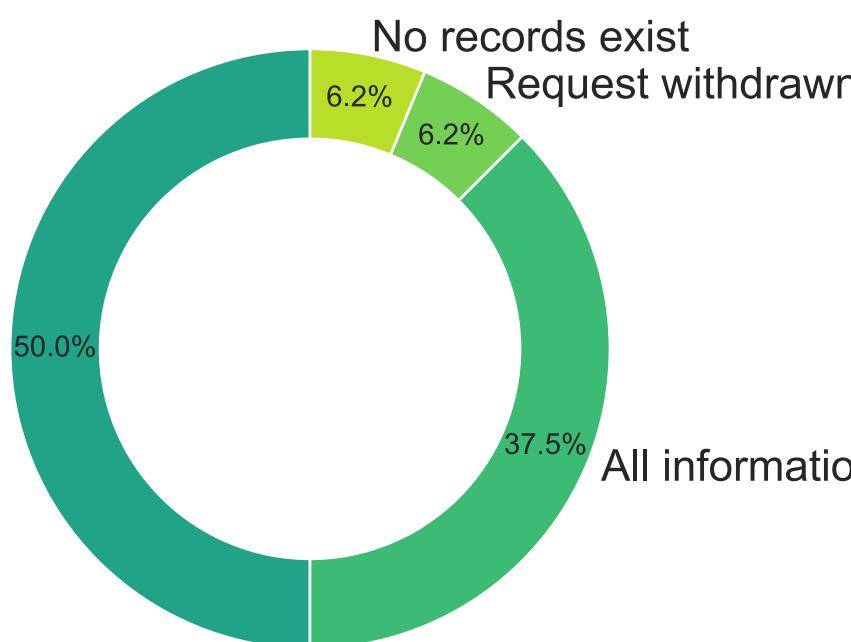


**food bear illness**



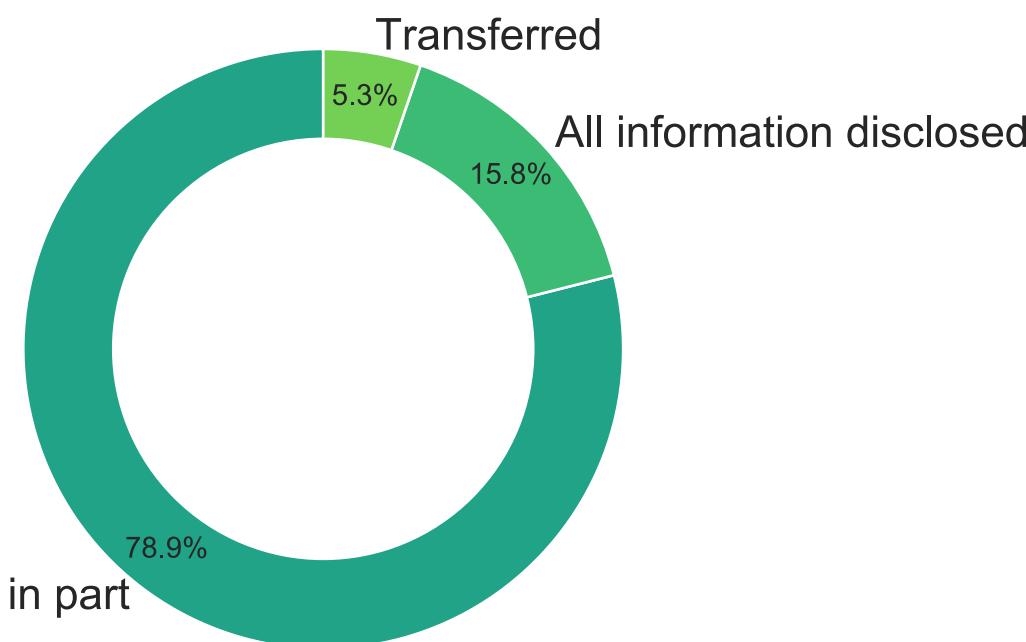
**public health inspection**

Information disclosed in part 50.0%

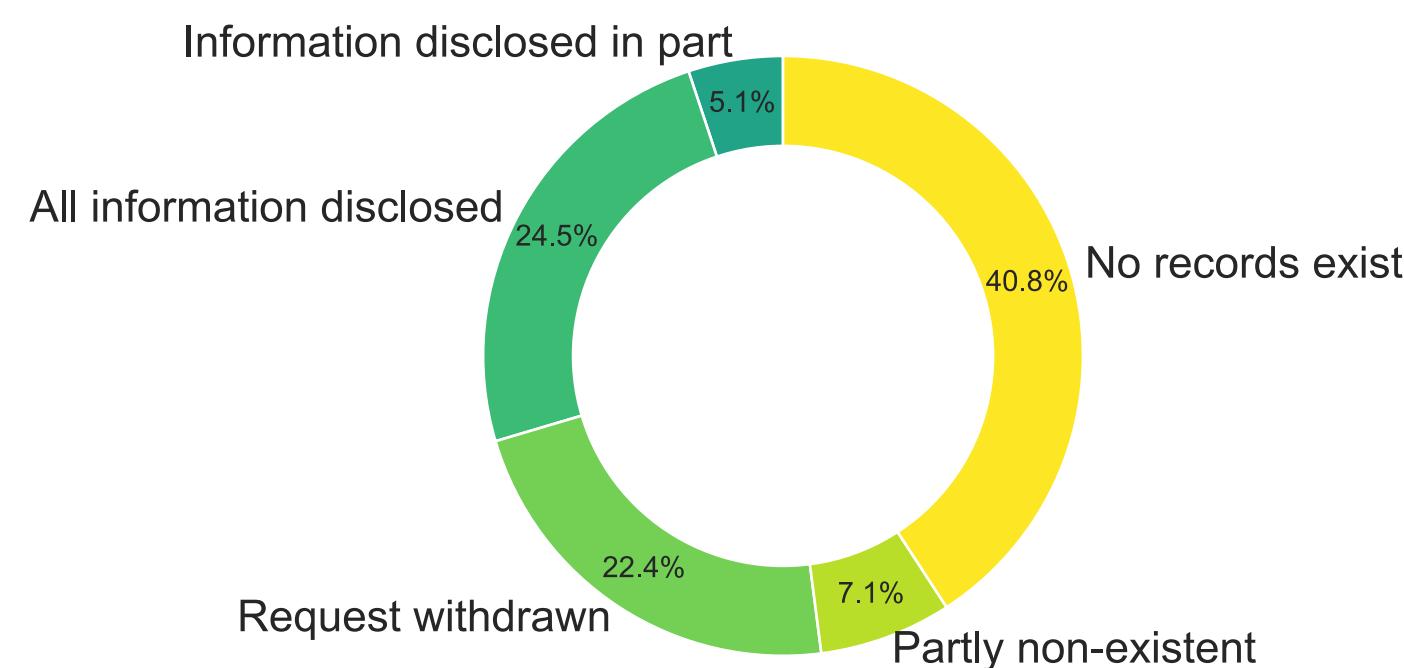


**rabies control**

Information disclosed in part



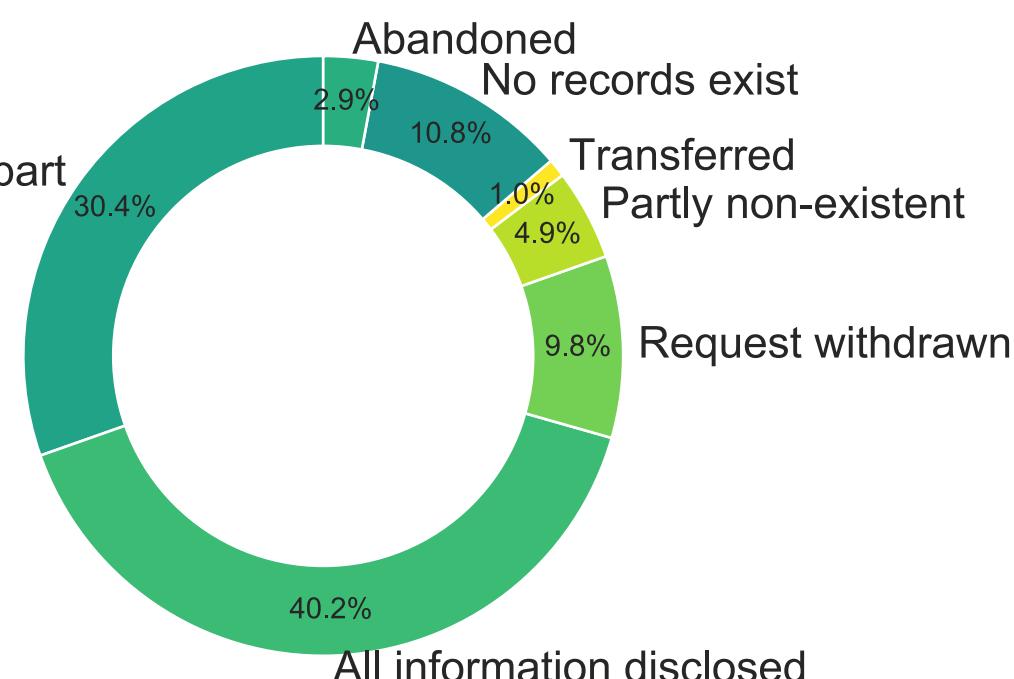
**environmental site**



Information disclosed in part

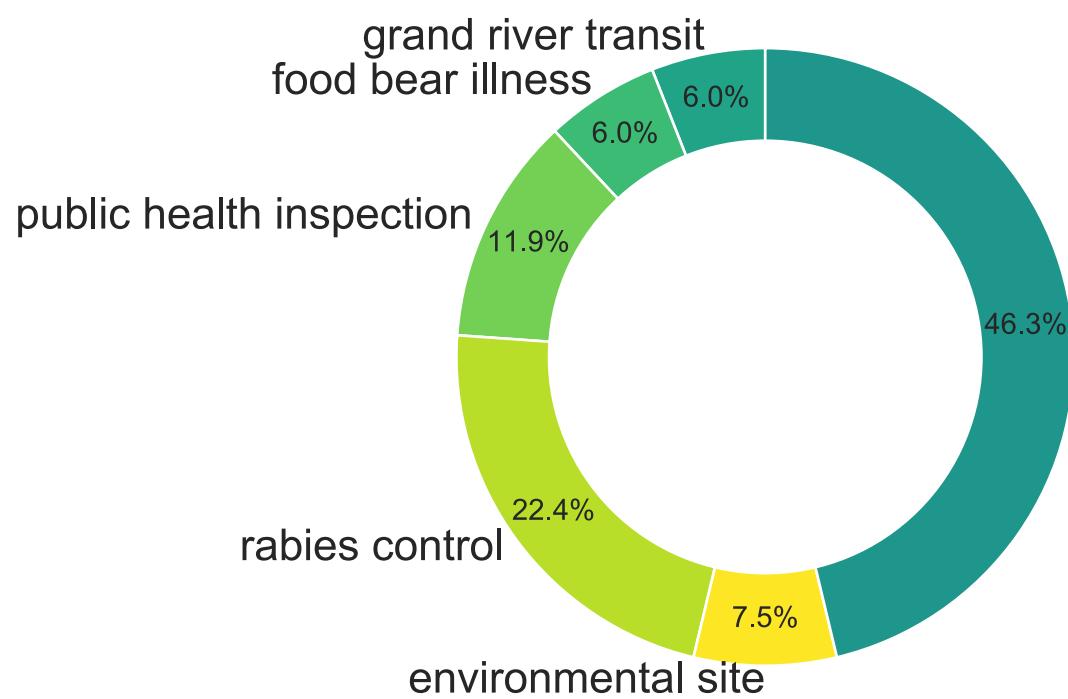
**ontario works**

All information disclosed

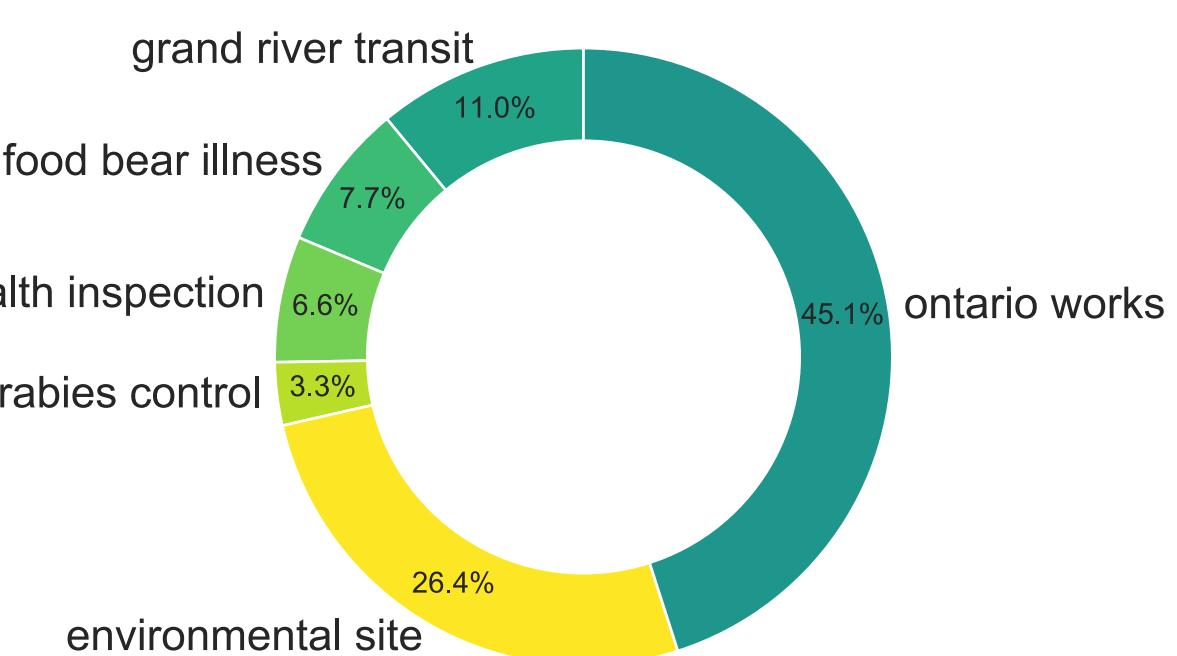


## For requests with the n-grams, n-gram percentage based on decision

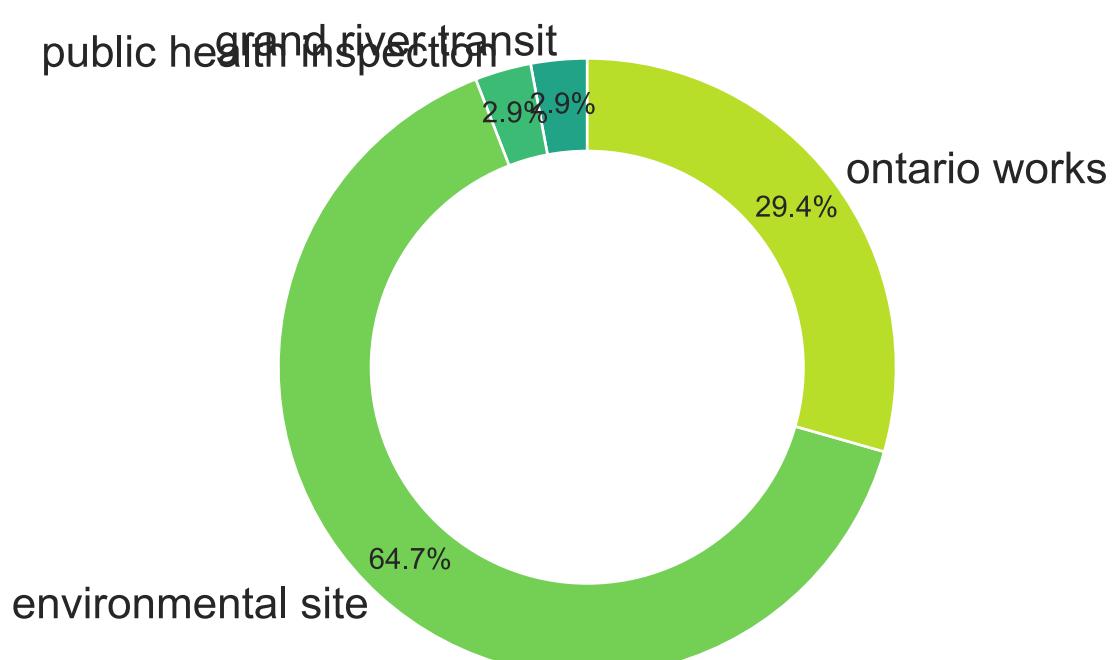
### Information disclosed in part



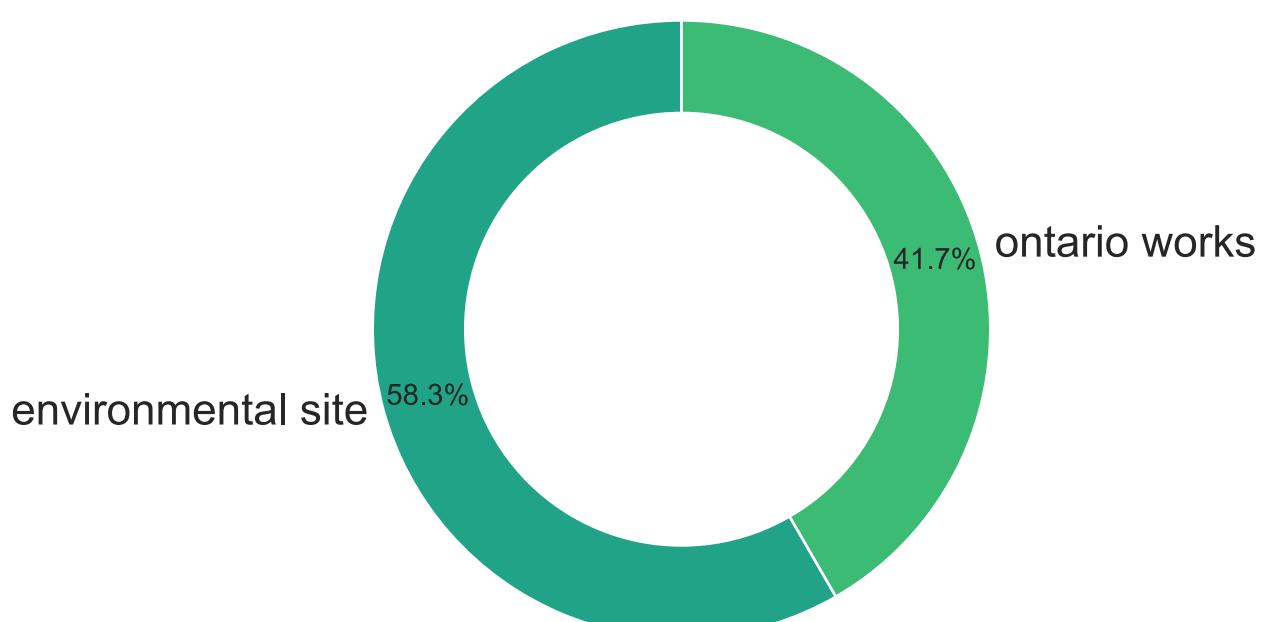
### All information disclosed



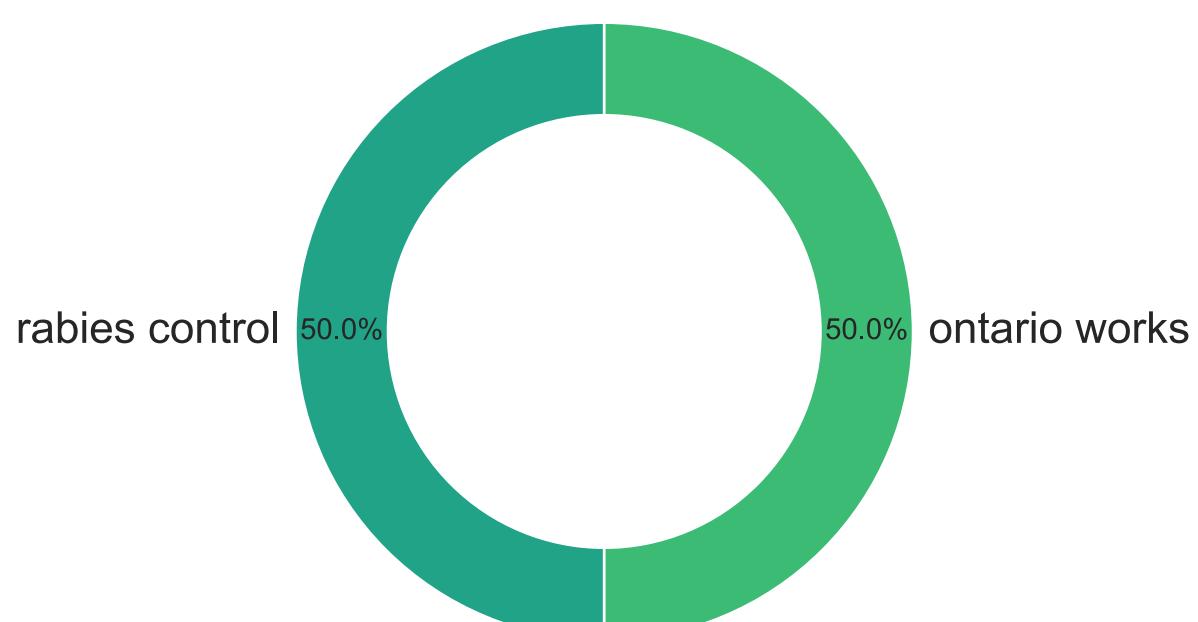
### Request withdrawn



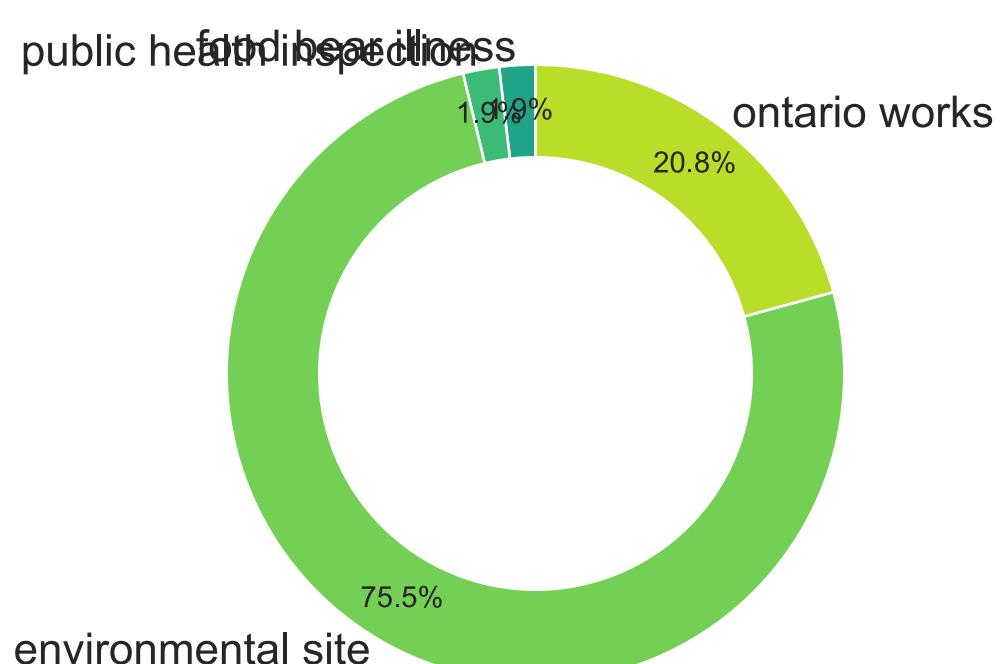
### Partly non-existent



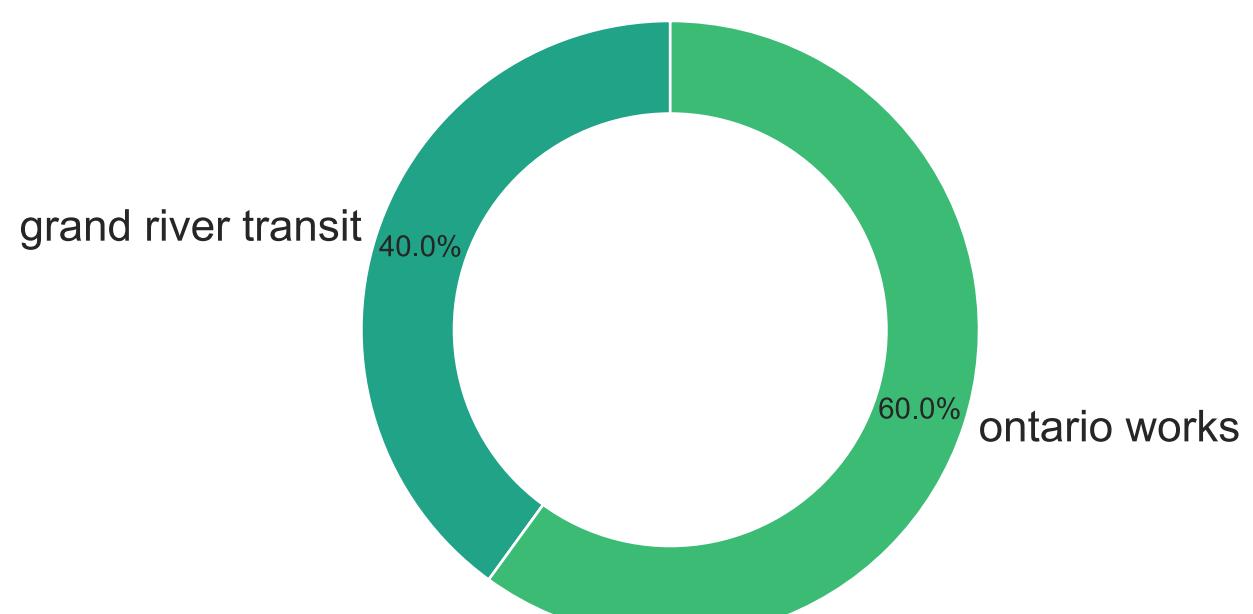
### Transferred



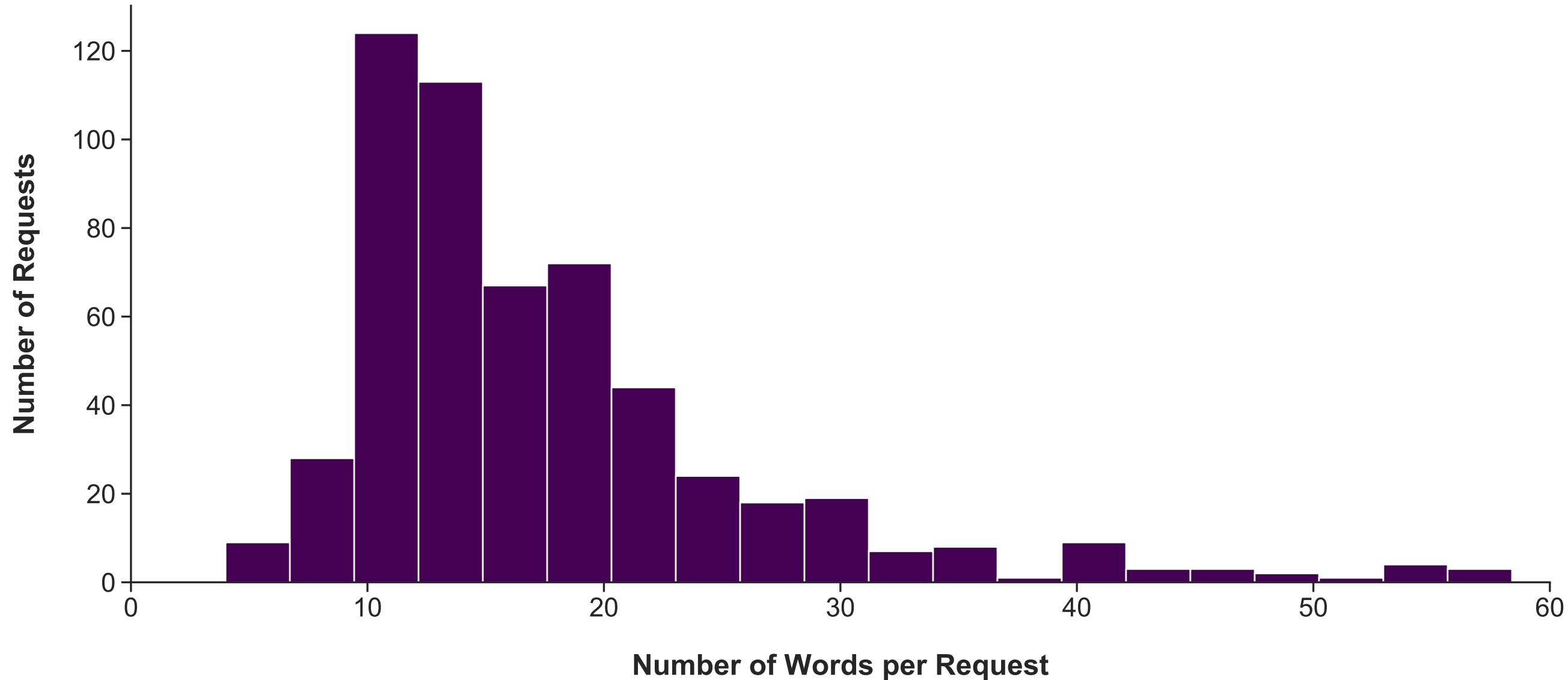
### No records exist



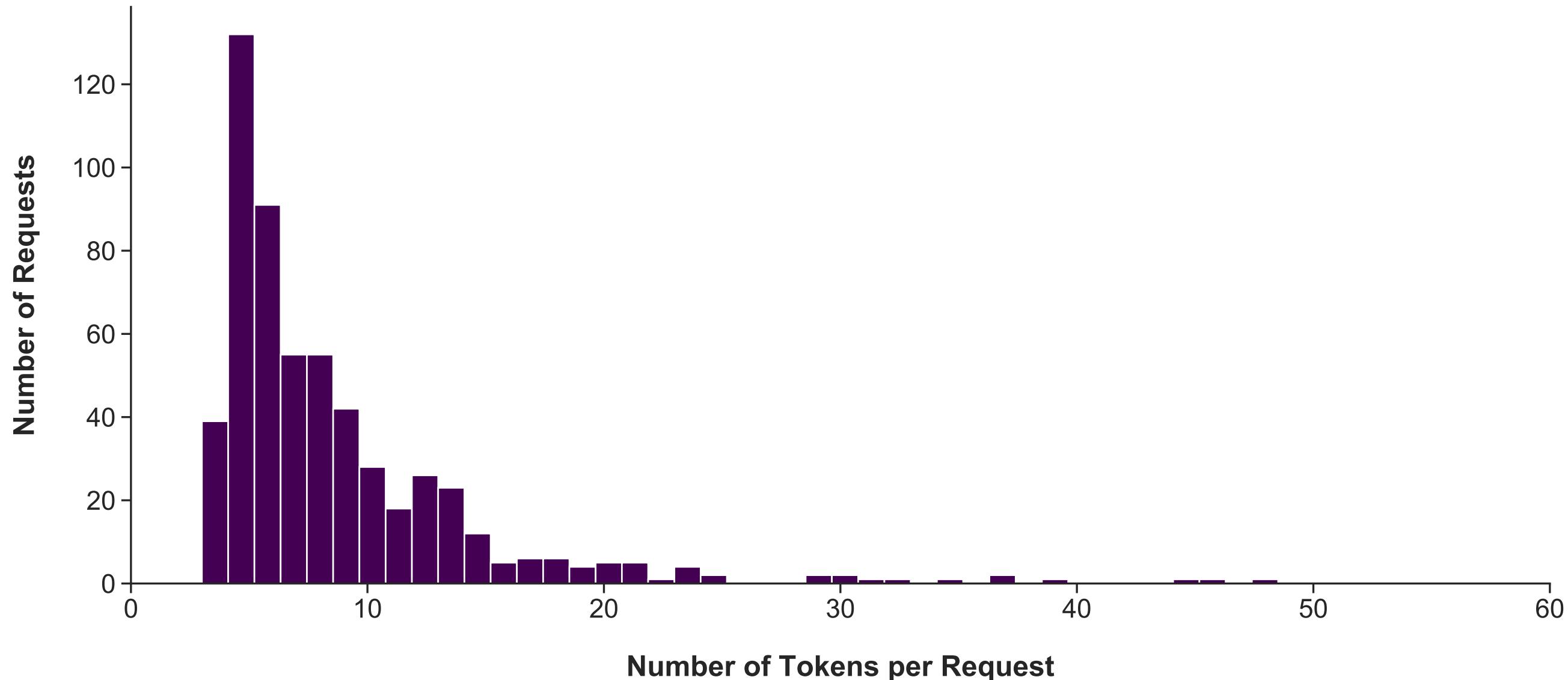
### Abandoned



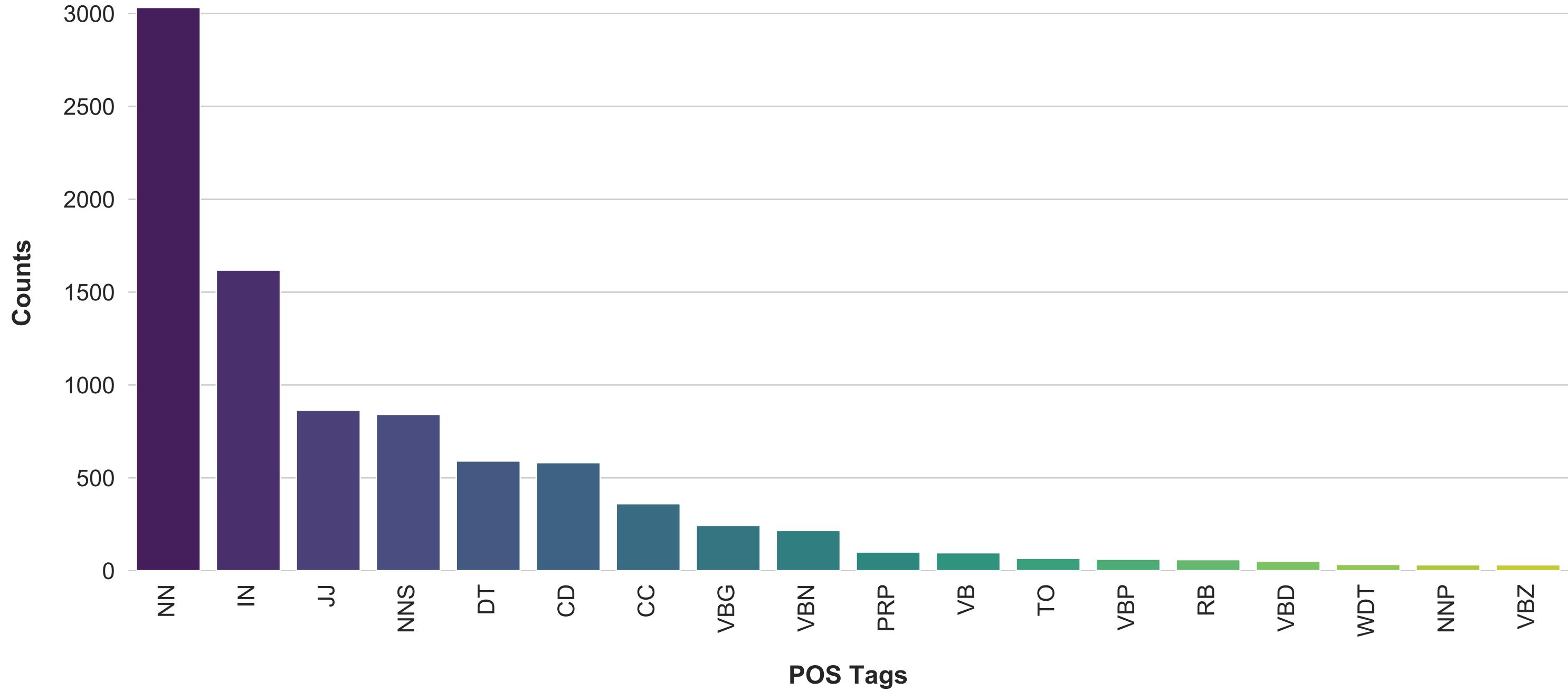
**Average number of words per request is 20.5, while the median is 15.0**



**Average number of tokens per request is 9.3, while the median is 7.0**

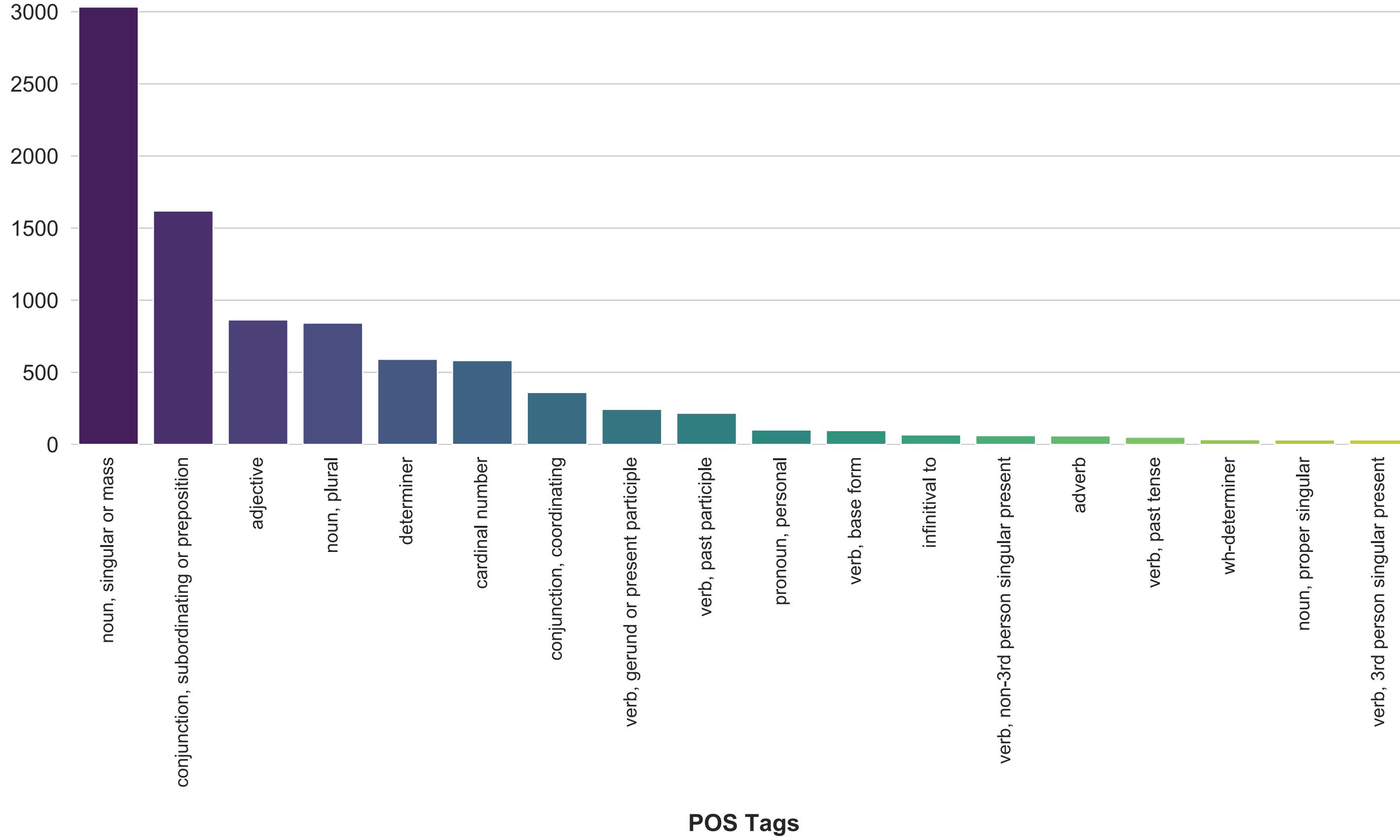


## Full Tokenized Text

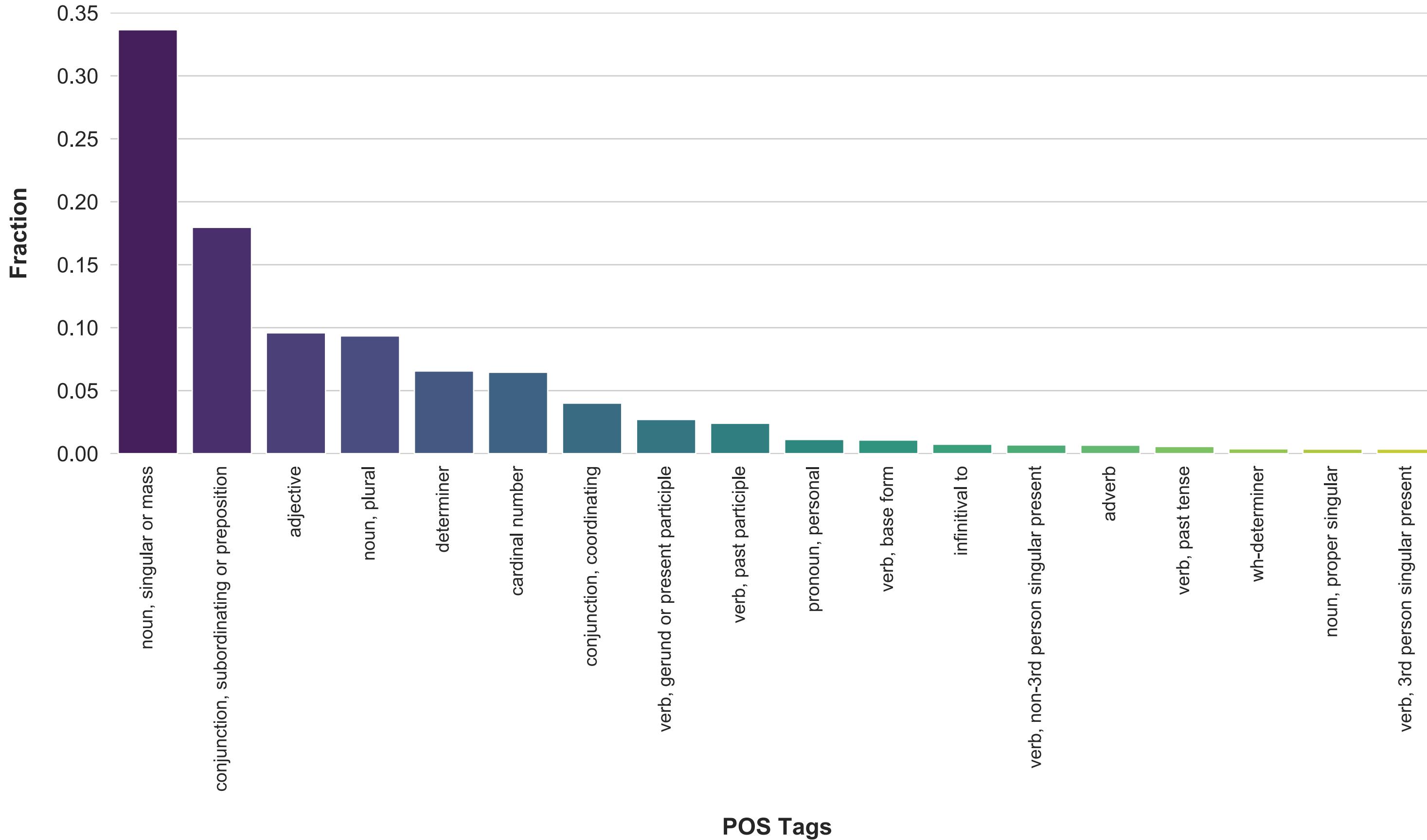


## Full Tokenized Text

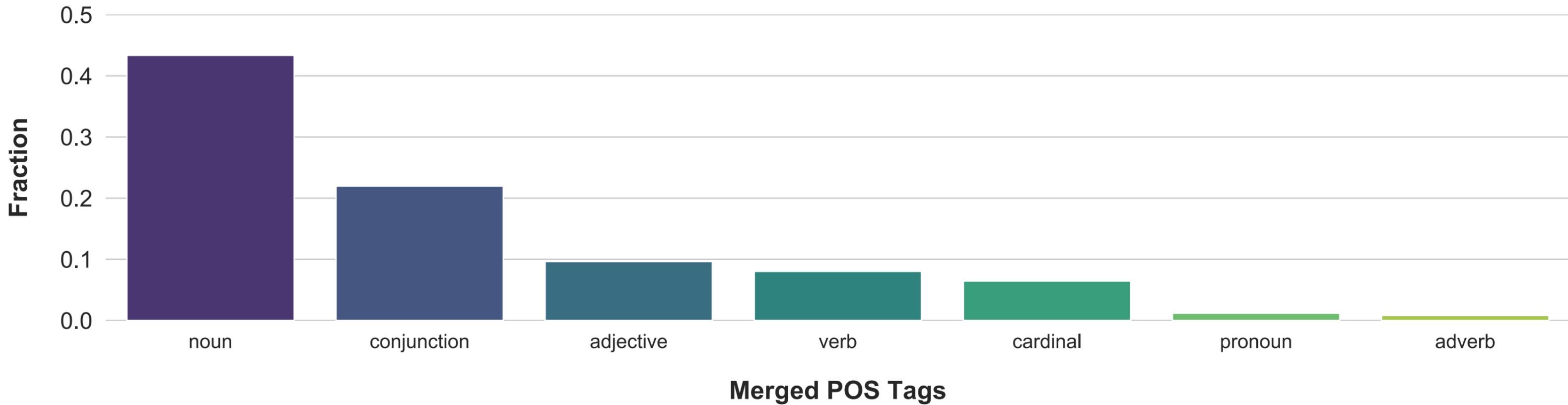
Counts



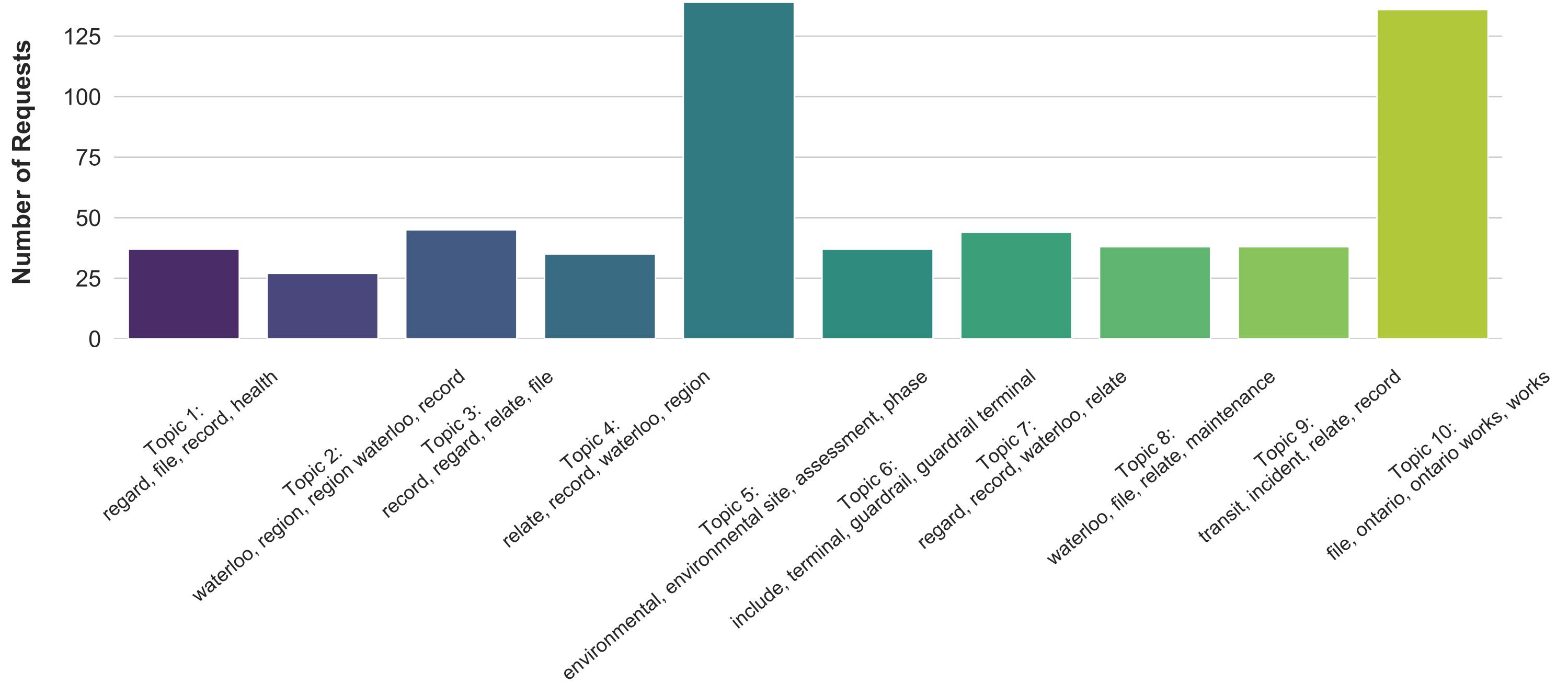
## Full Tokenized Text



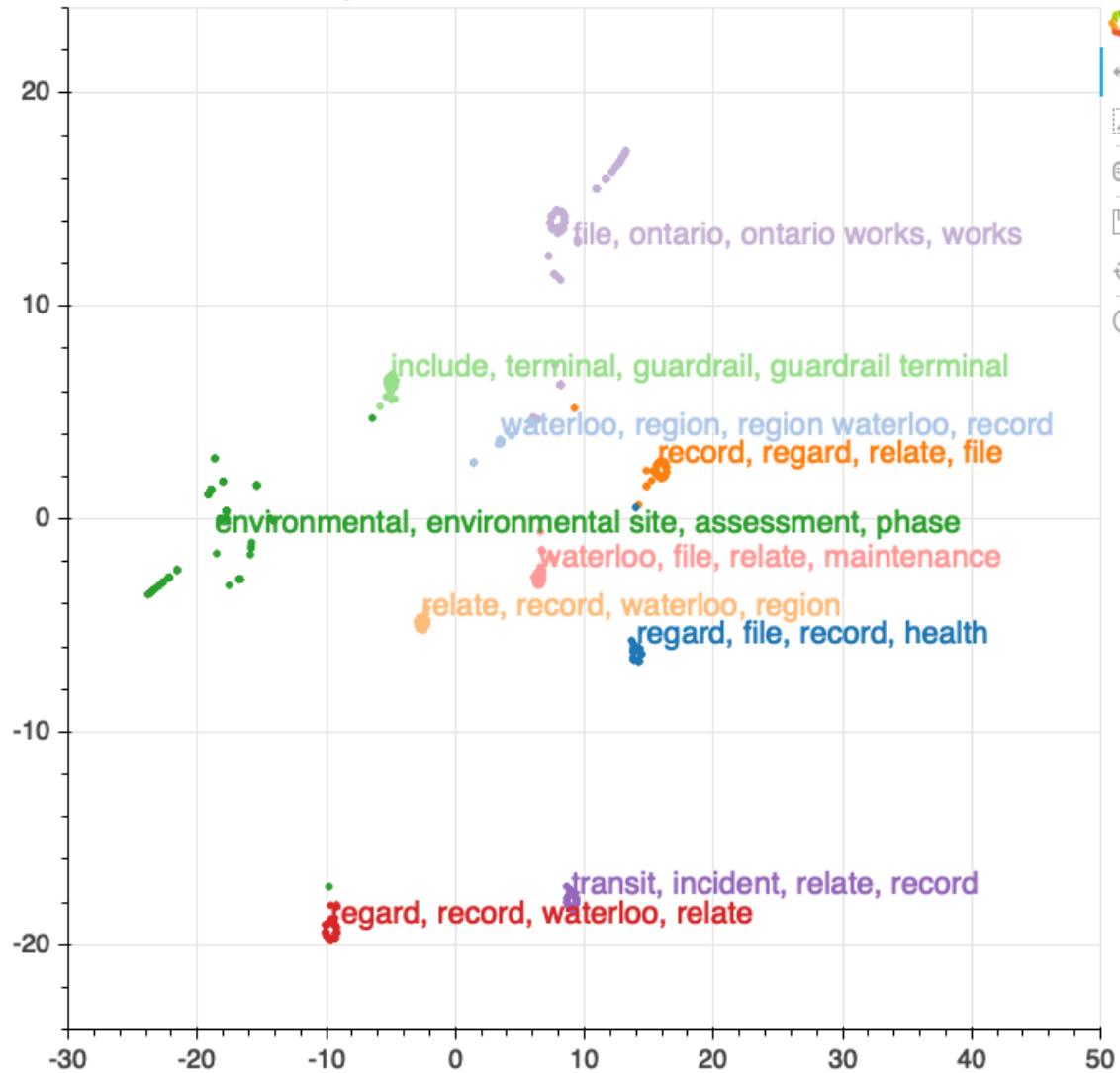
## Full Tokenized Text



## LDA Topic Counts - CountVectorizer

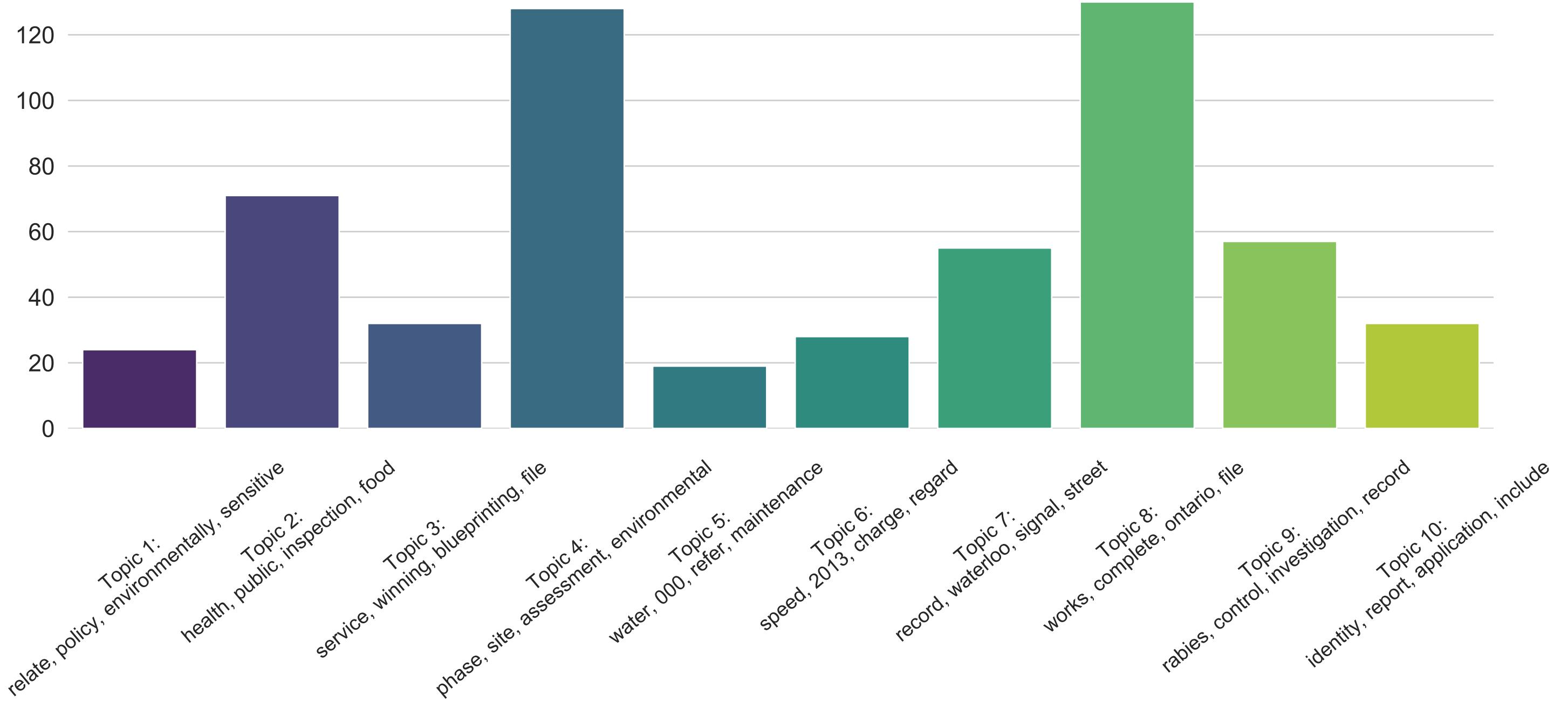


## t-SNE Clustering of 10 LDA Topics - CountVectorizer

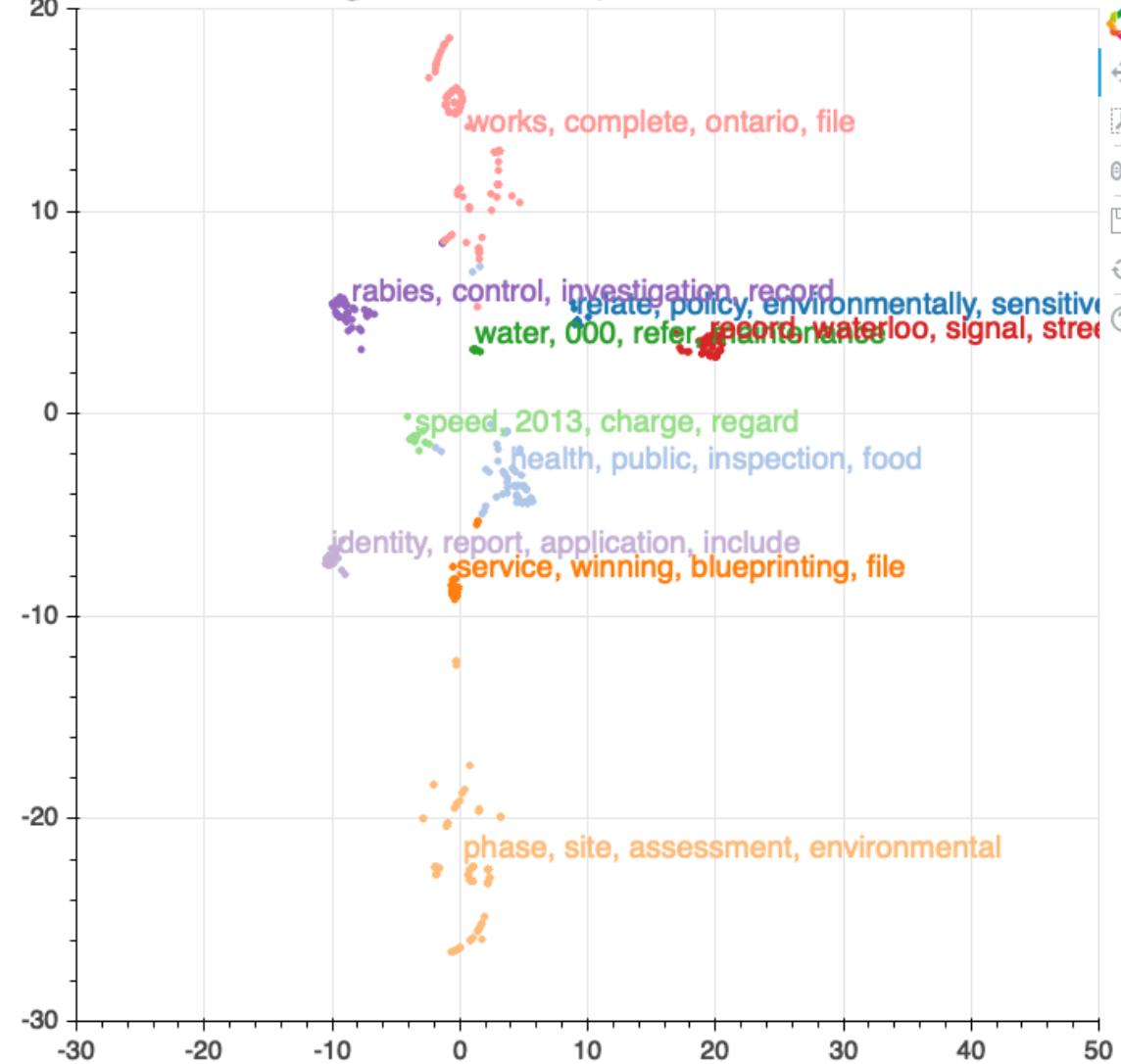


## LDA Topic Counts - tf-idf Vectorizer

Number of Requests



### t-SNE Clustering of 10 LDA Topics - tf-idf Vectorizer



## LSA Topic Counts - CountVectorizer

Number of Requests

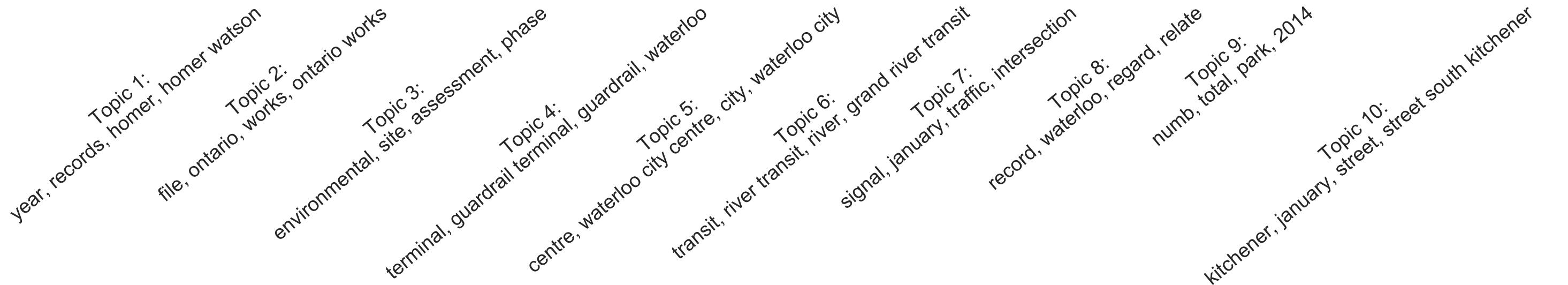
200

150

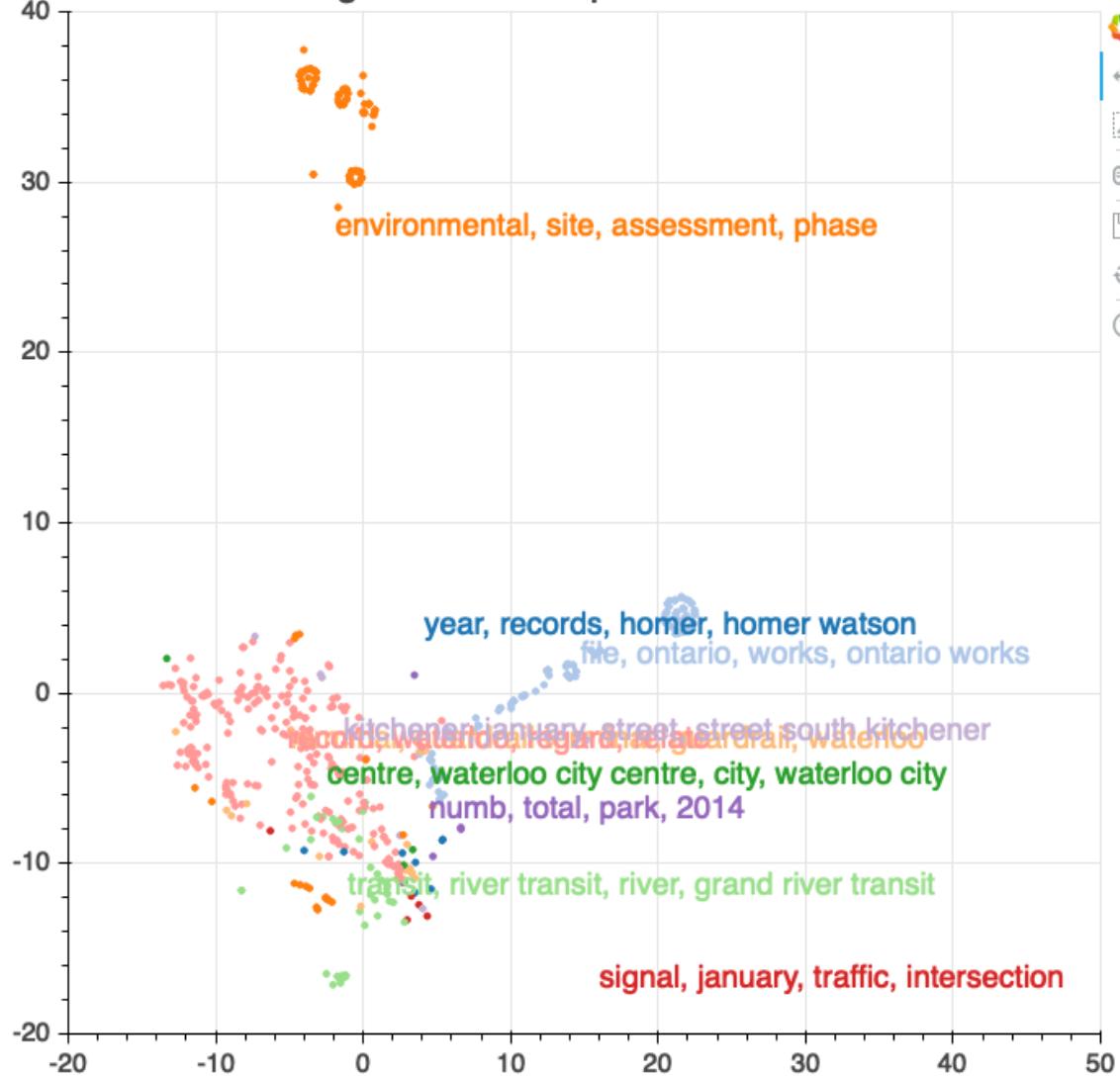
100

50

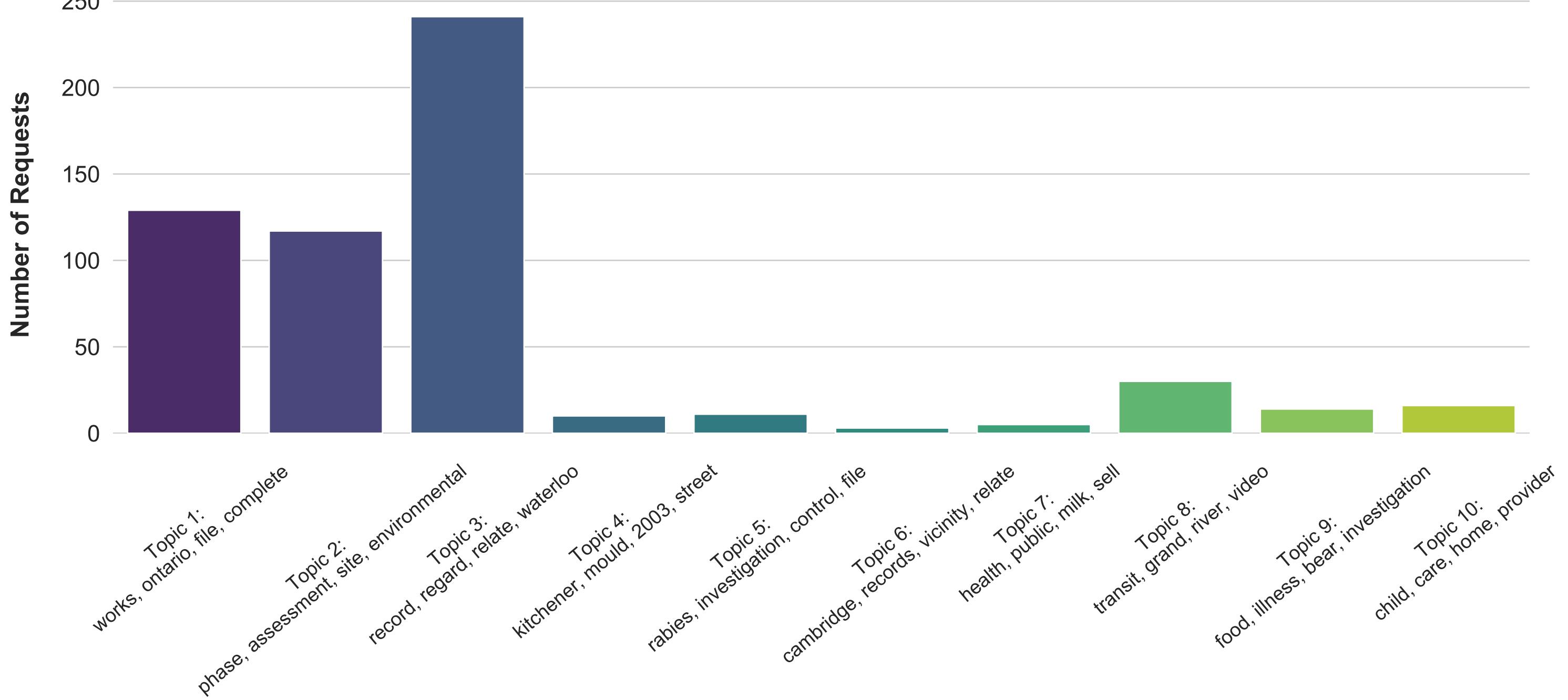
0



## t-SNE Clustering of 10 LSA Topics - CountVectorizer



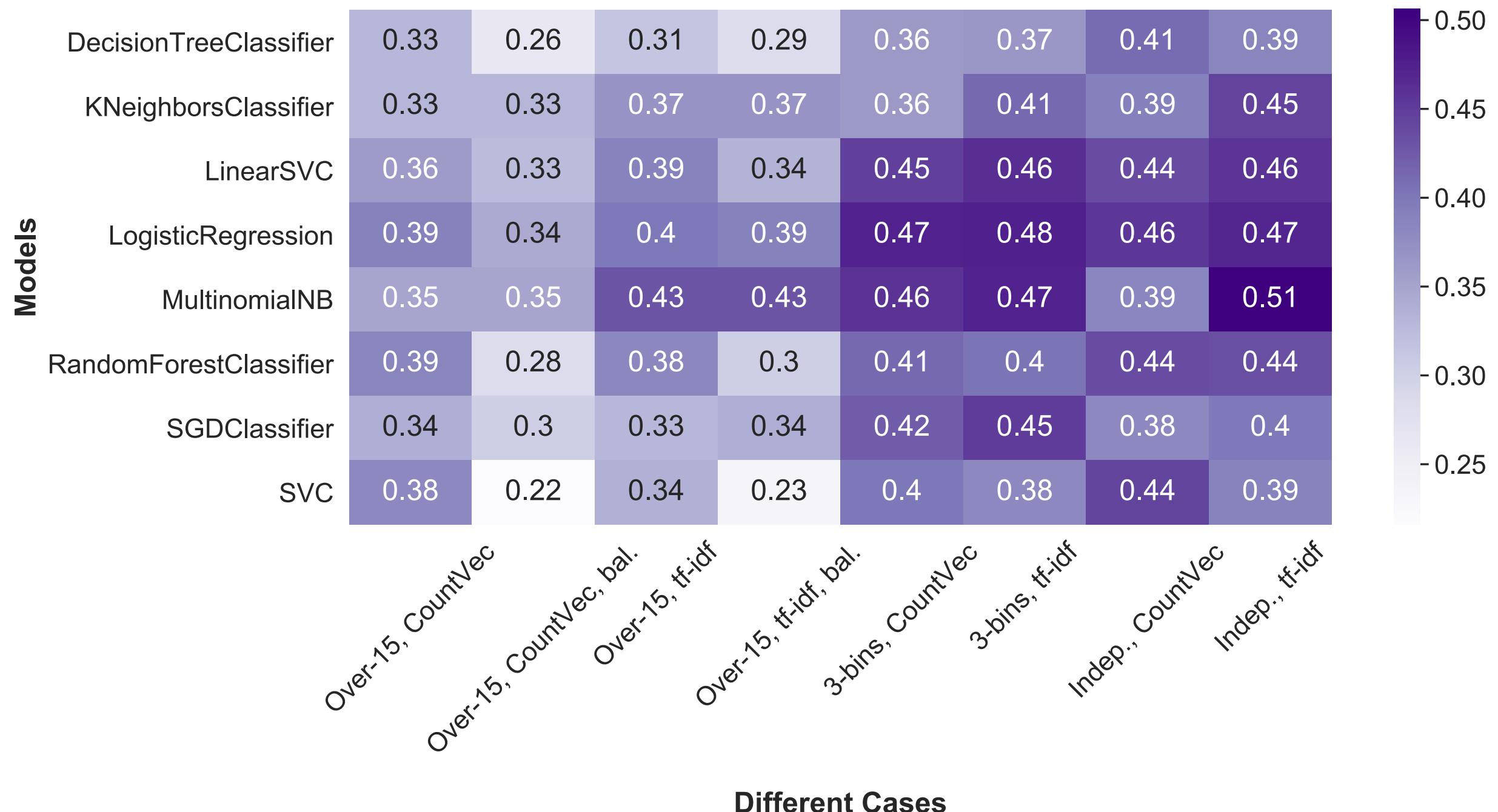
## LSA Topic Counts - tf-idf Vectorizer



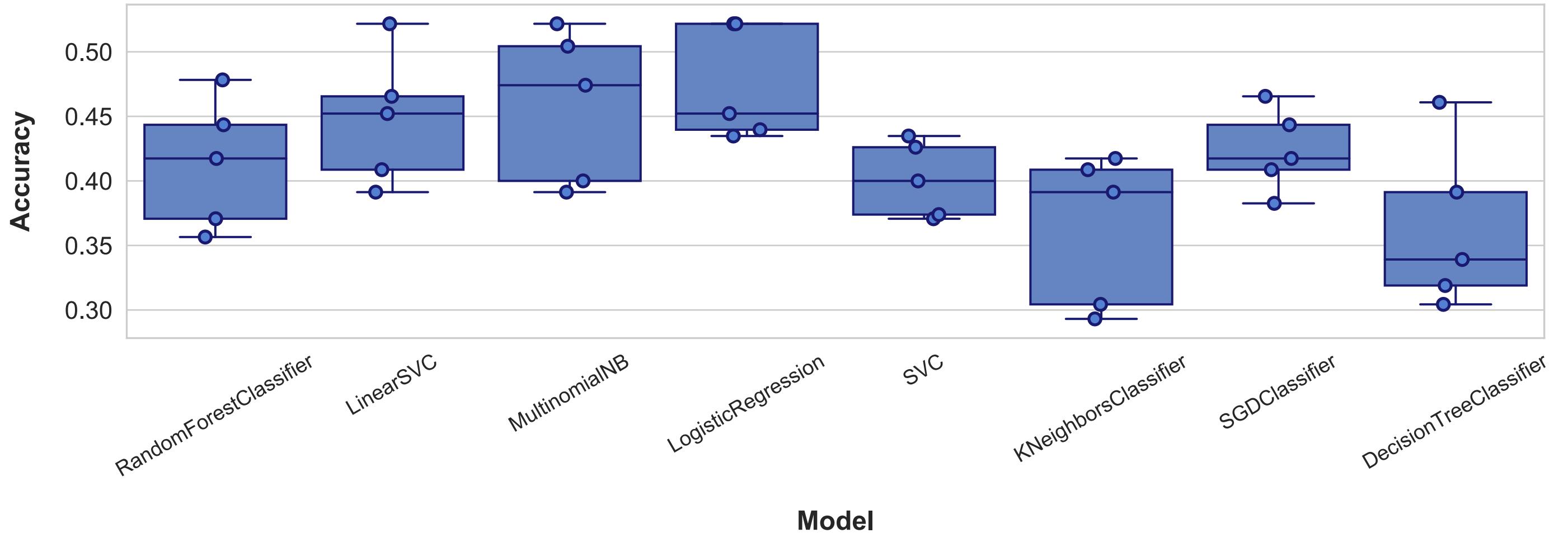
### t-SNE Clustering of 10 LSA Topics - tf-idf Vectorizer



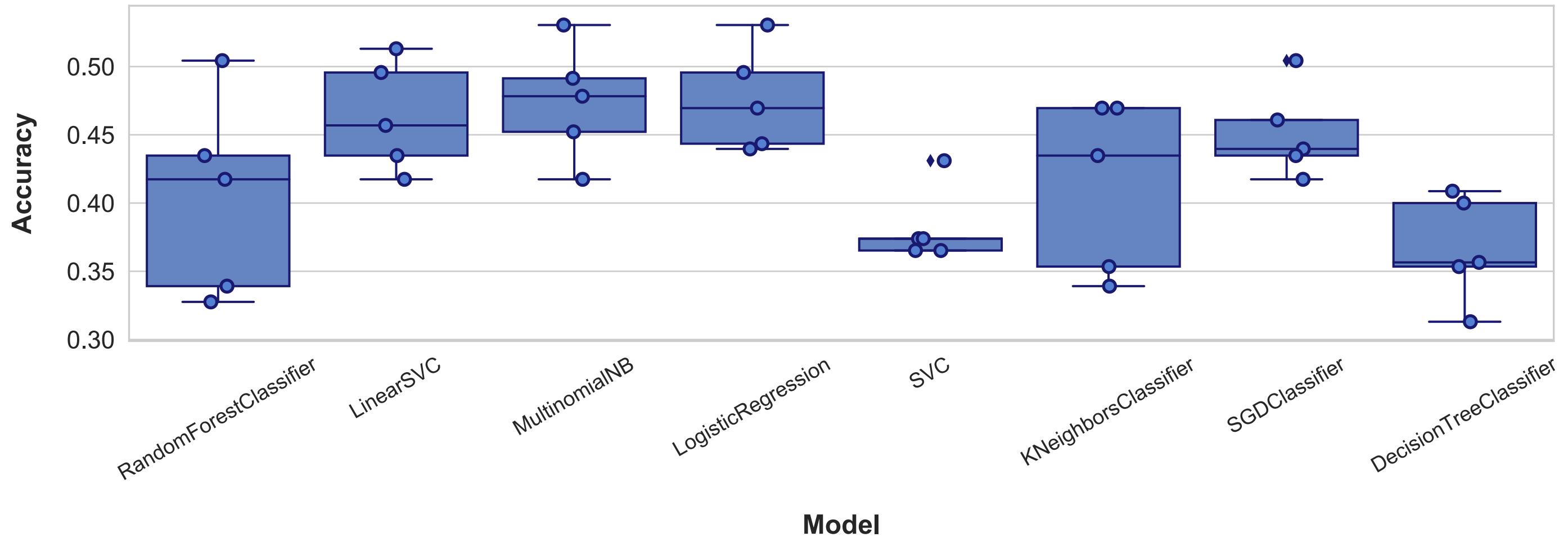
# ML Model Accuracy



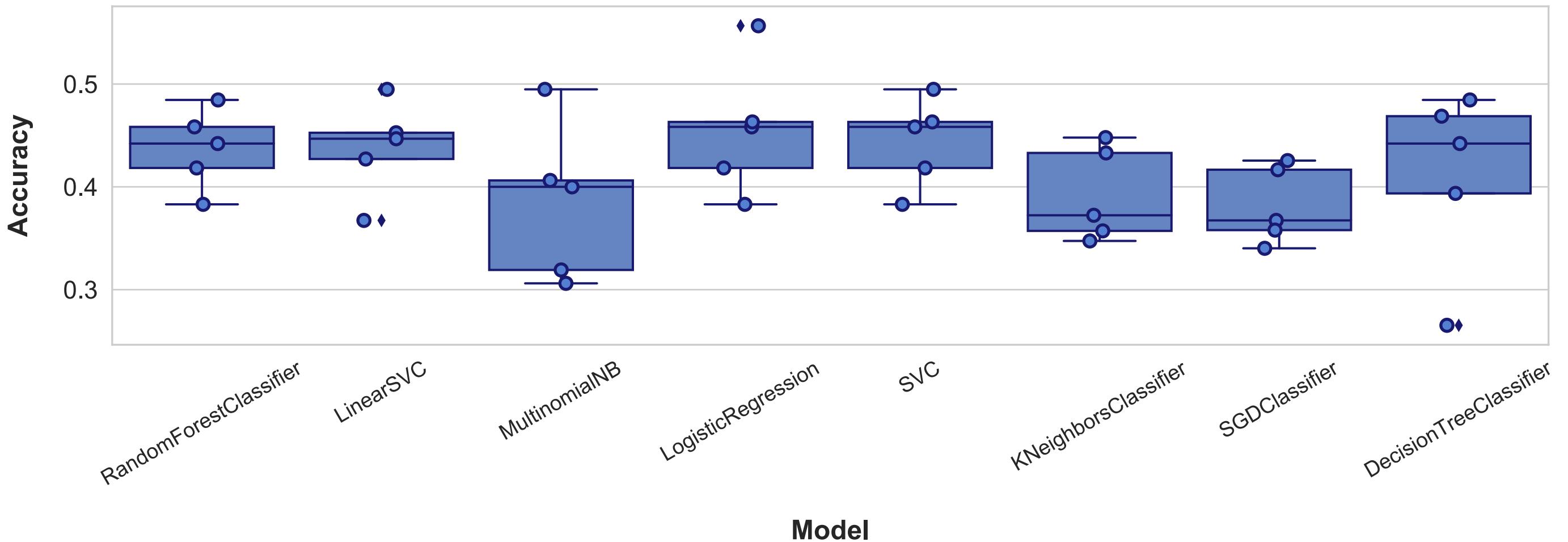
## Classifier comparison for the 3-bin case, using CountVectorizer



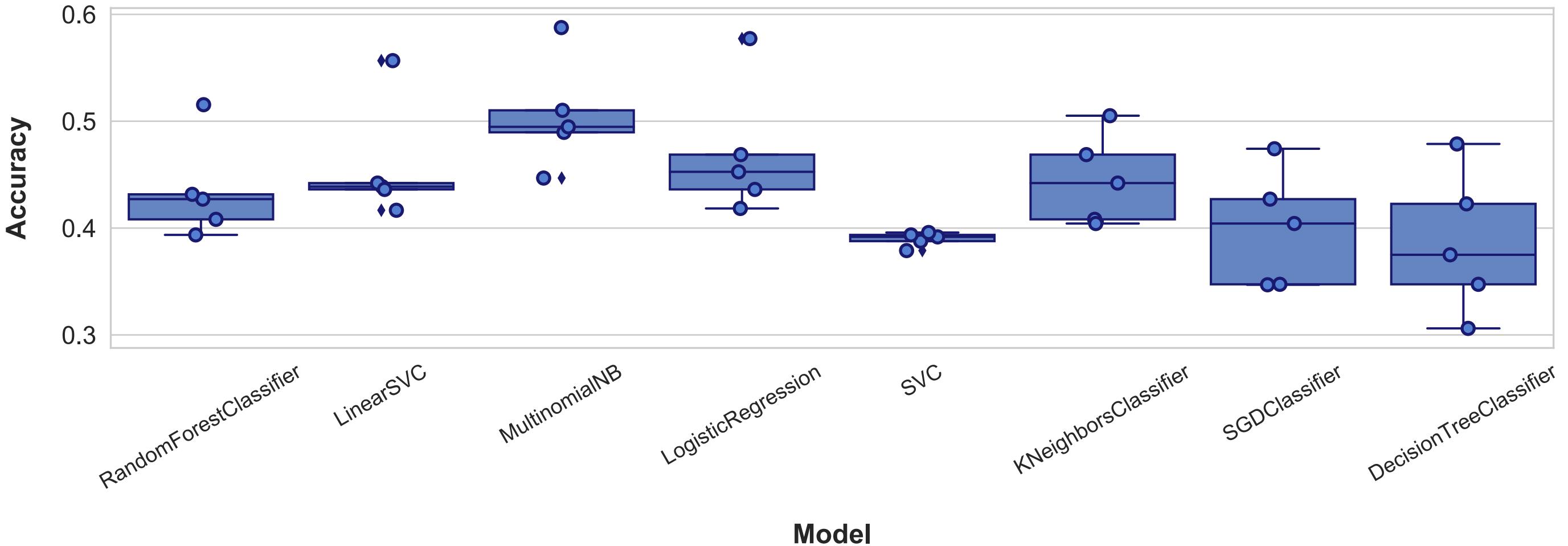
## Classifier comparison for the 3-bin case, using tf-idf



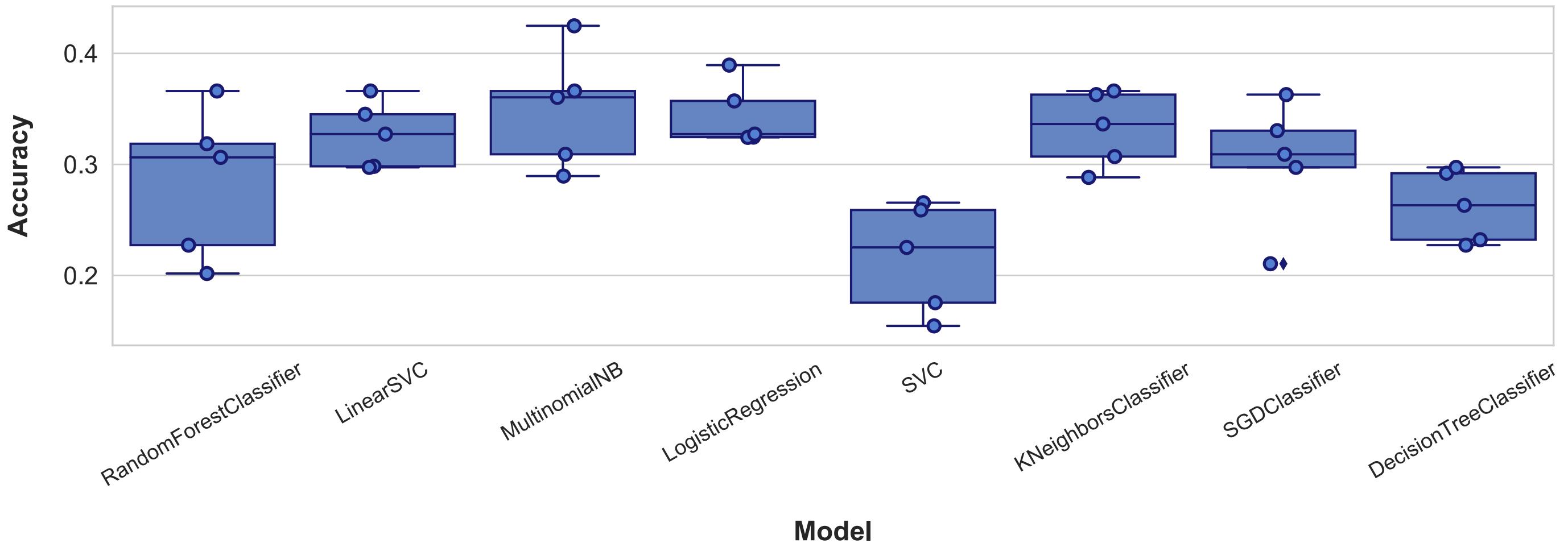
## Classifier comparison for the indep. case, using CountVectorizer



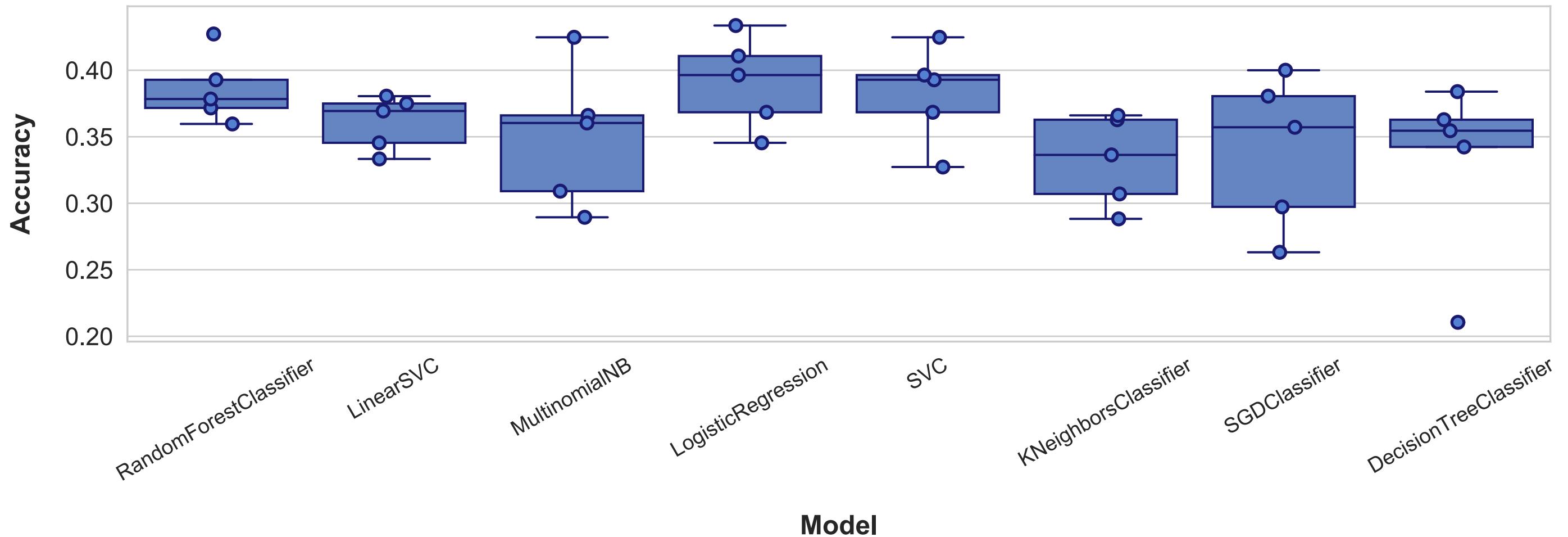
## Classifier comparison for the indep. case, using tf-idf



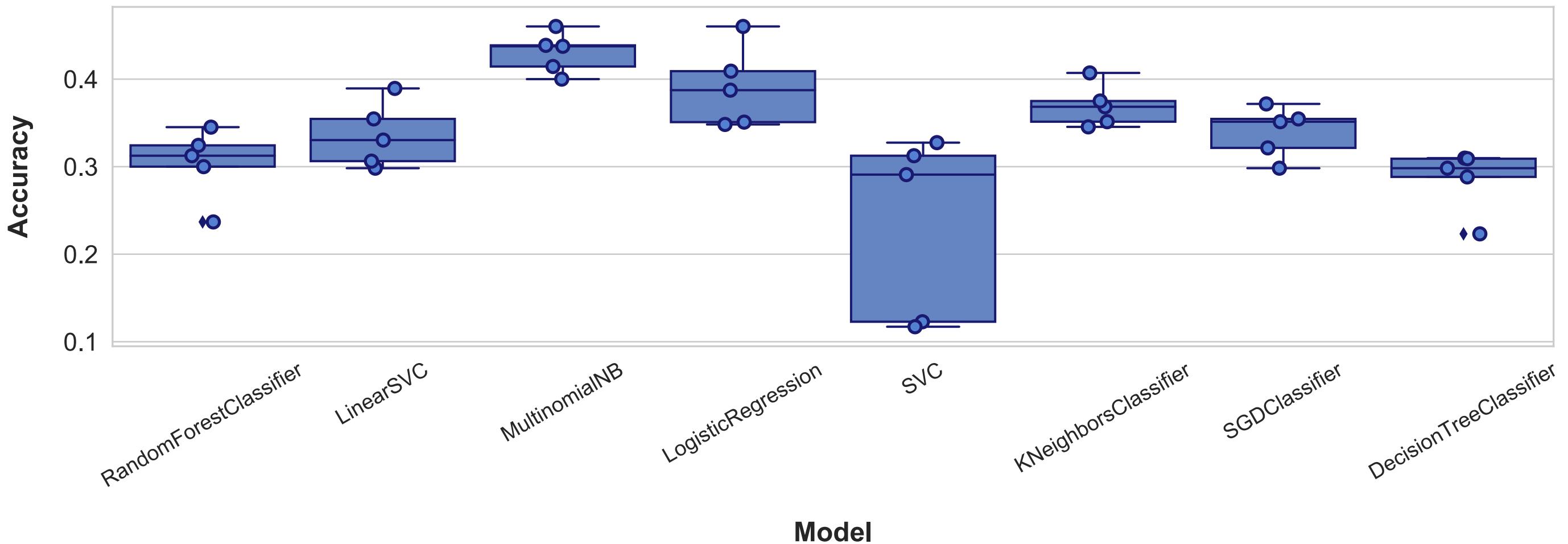
## Classifier comparison for the over 15 case, using CountVectorizer, Balanced



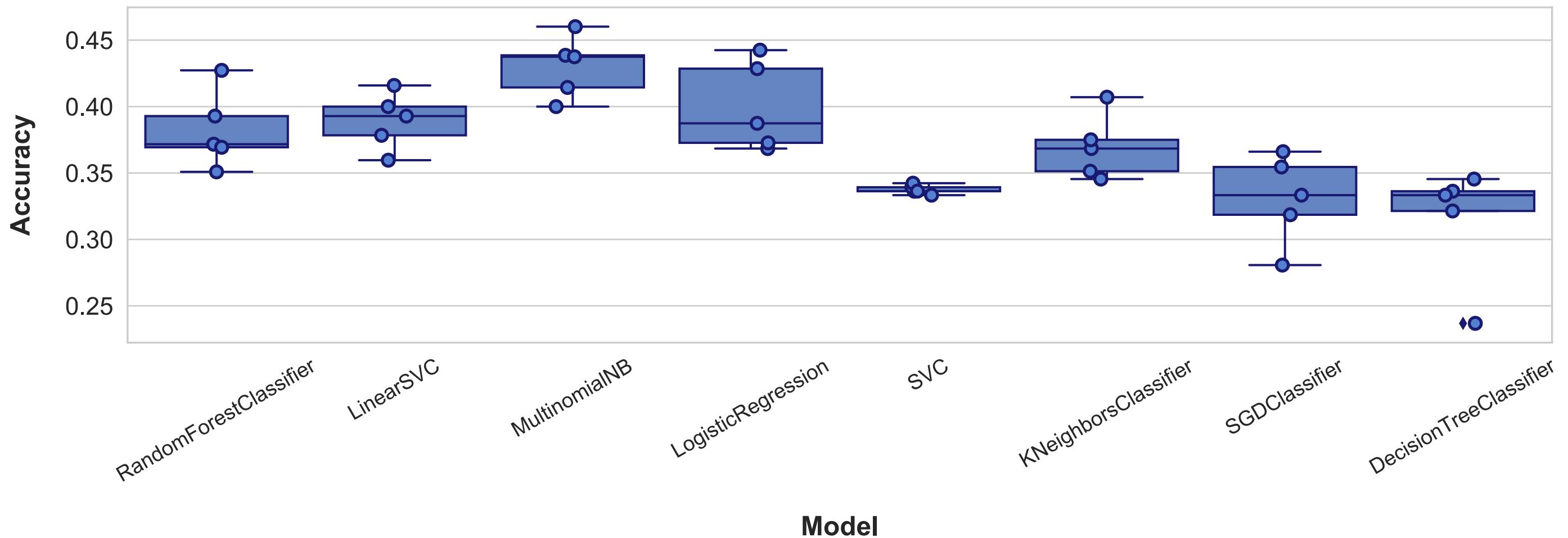
## Classifier comparison for the over 15 case, using CountVectorizer



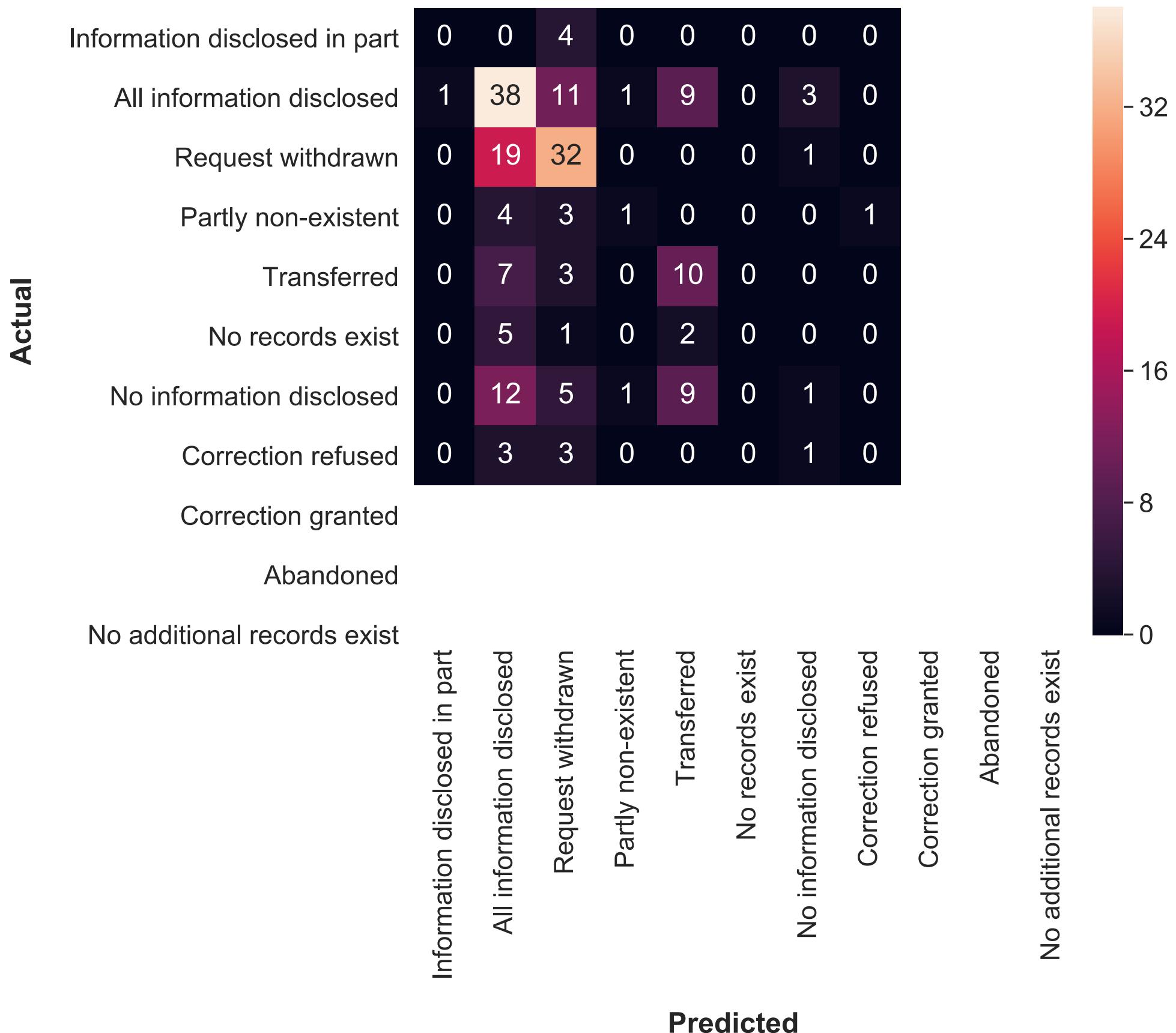
## Classifier comparison for the over 15 case, using tf-idf, balanced



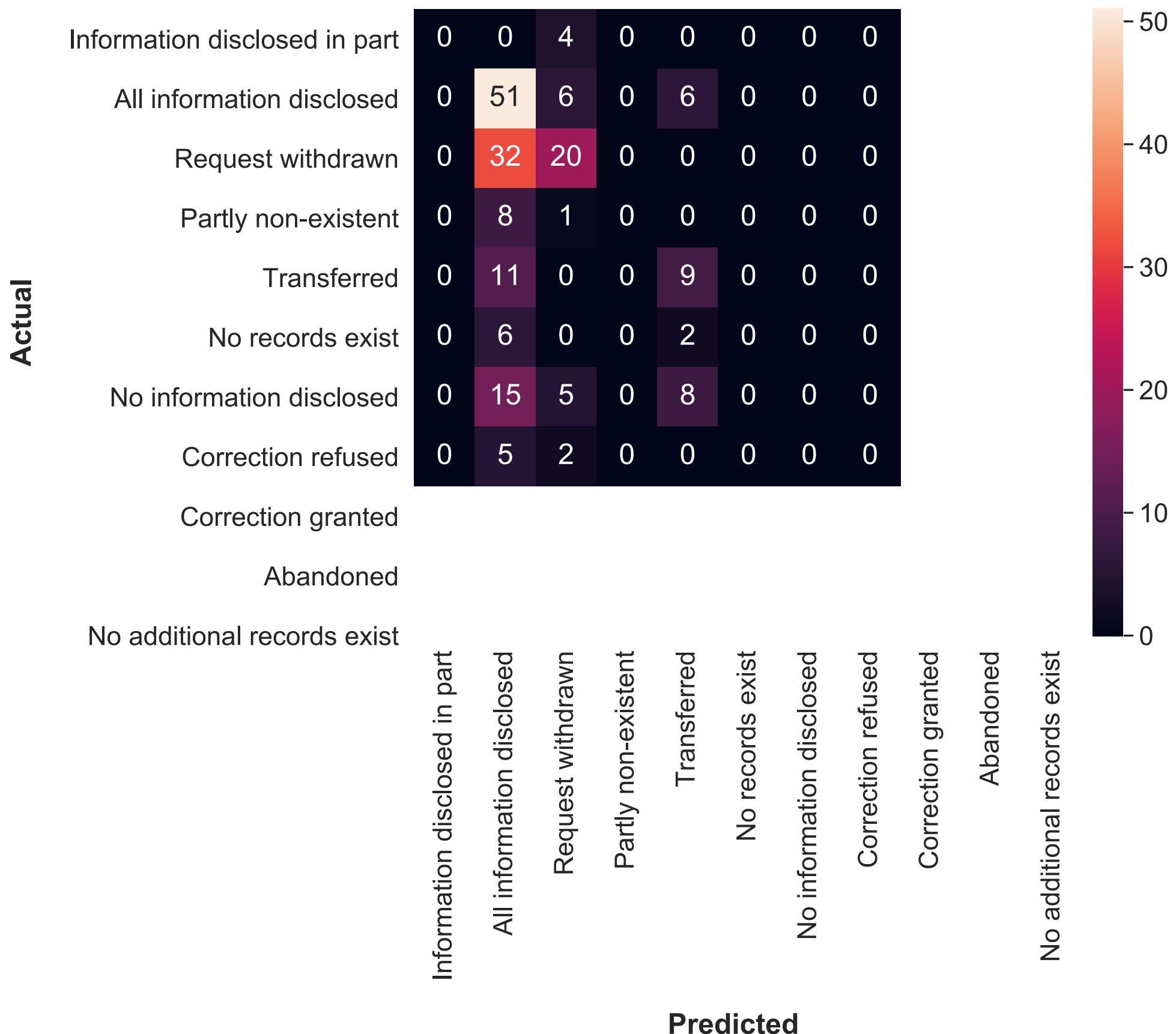
## Classifier comparison for the over 15 case, using tf-idf



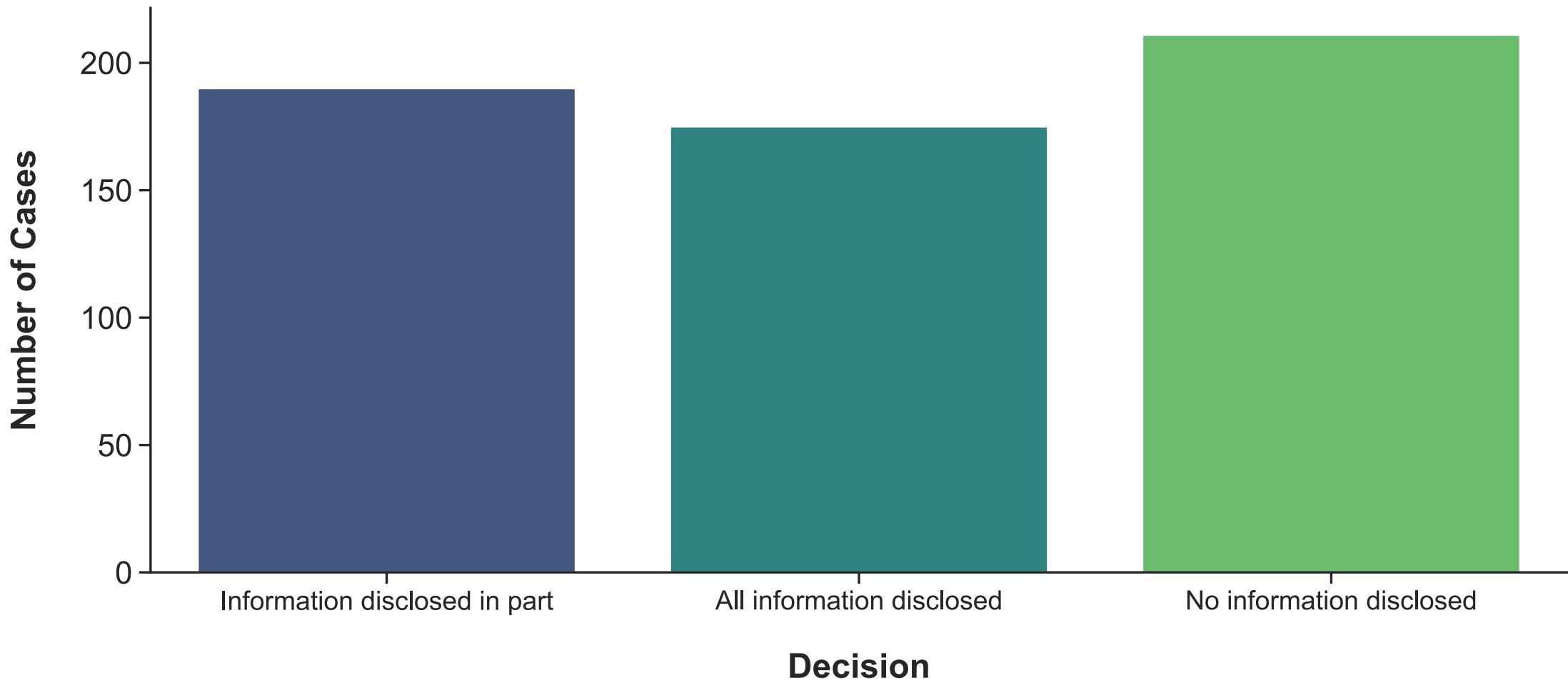
# MultinomialNB, CountVectorizer, full set



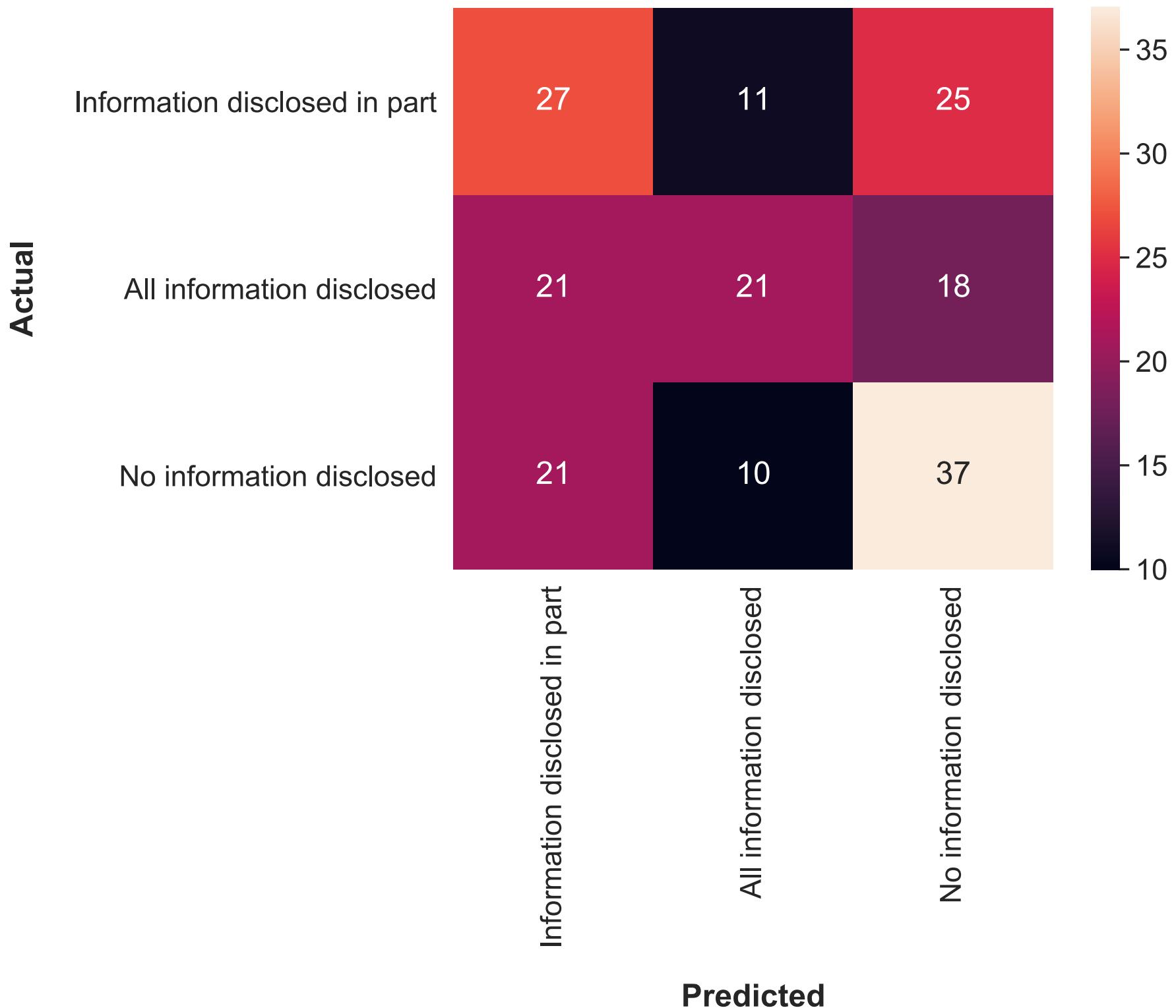
## MultinomialNB, tf-idf, full set



## **Full data split into three categories only**

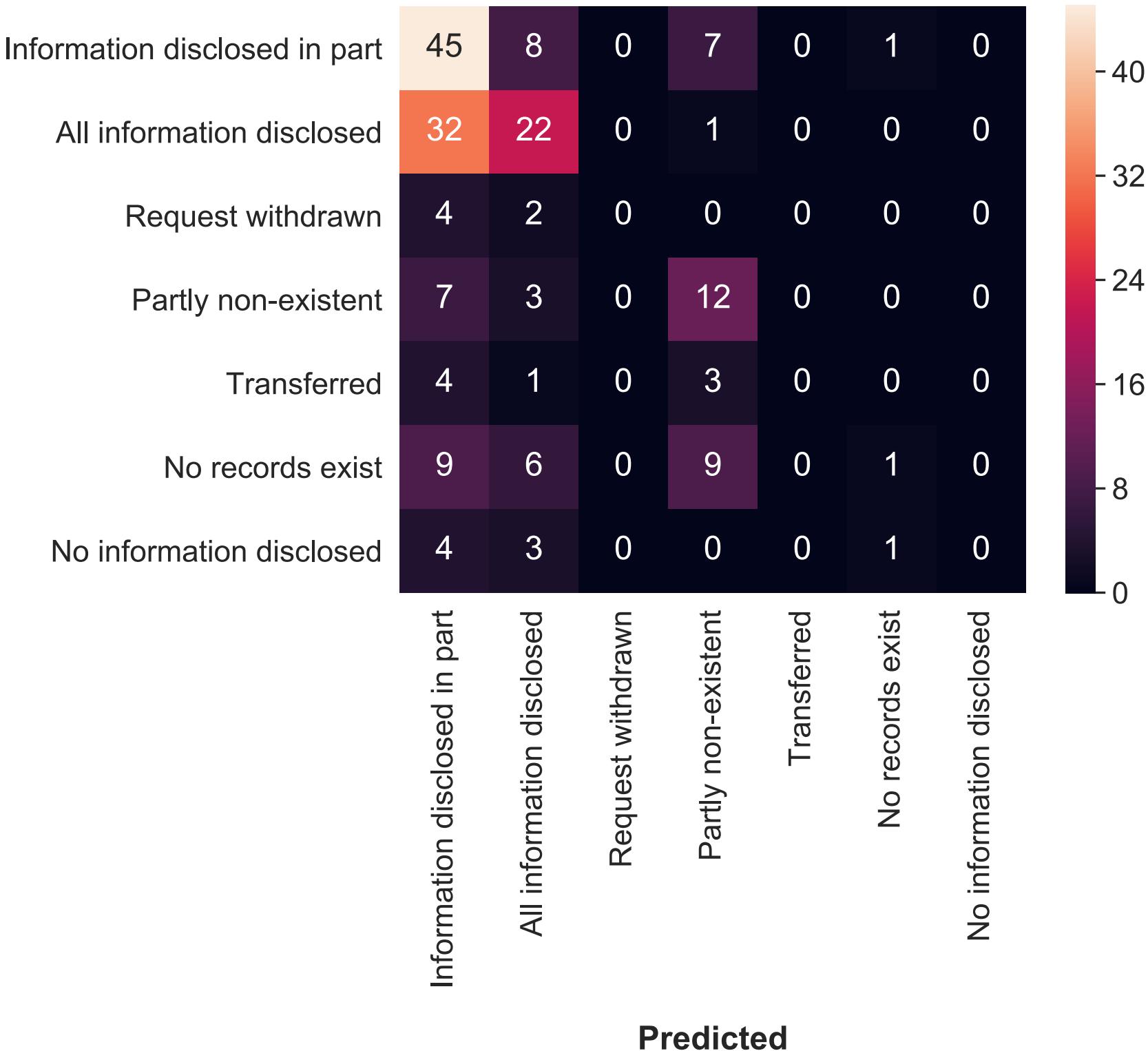


## LogisticRegression, tf-idf, 3 bins



# MultinomialNB, tf-idf, over 15

Actual



## MultinomialNB, tf-idf, indep.

