BOSTON UNIVERSITY GRADUATE SCHOOL
COLLEGE OF ART AND SCIENCE & COLLEGE OF ENGINEERING

CS506 Final Report
WGBH Reliability of Informant Cases

Changhao Liang U16843909
Xiang Liu U46406505
Qitong Wang U44926429
Chen Xin U72406425

Department of Computer Science, Boston University
Department of Systems Engineering, Boston University

May 1, 2020

Contents

List of Figures

1 Introduction

The WGBH Informant Project is an effort to create a comprehensive analysis of criminal cases which have been overturned due to the use of an informant in New Hampshire and Rhode Island over the past decade (2008-2018). The project aims to

- Uncover never before-seen patterns of judicial behavior and decision-making in criminal cases
- Recast and spur more aggressive news coverage of the court system and the use of informants
- Create a valuable understanding of the effectiveness of informant testimony
- Find relationship between overturned cases and presence of informant

In a criminal case, only the defendant has a right to an appeal in most states. Lower court cases can be challenged for any erroneous rulings by the judge on the evidence or on the law, the prosecutor's conduct during the case, or even the trial attorney's handling of the defense.

2 Data Collection

Criminal cases from lower courts can be reversed in full or reversed in parts. We want to collect both full and partial reversed decisions. To do this, we need to scrape for all cases that were sent to appellate courts of New Hampshire and Rhode Island.

2.1 Data Source

Our data source are the appellate cases posted on the supreme court websites of New Hampshire and Rhode Island. Each case is stored in pdf format.

- Cases from New Hampshire
  https://www.courts.state.nh.us/supreme/opinions/
- Cases from Rhode Island
  https://www.courts.ri.gov/Courts/SupremeCourt/Pages/Opinions%20and%20Orders%20Issued%20in%20Supreme%20Court%20Cases.aspx

Based on our research, we found that 9 of the 50 states in U.S. do not have an intermediate appellate court, and New Hampshire and Rhode Island are among these states. (https://en.wikipedia.org/wiki/List_of_state_intermediate_appellate_courts) Their appeals from lower courts go directly to their supreme courts.

2.2 Data Downloading

Selenium Webdriver is a collection of open source APIs which are used to automate the testing of a web application. To collect all case records from 2008 to 2018 in New Hampshire and Rhode Island, we used Selenium Webdrive to automate the download process of all the case files from their supreme court websites.

2.2.1 New Hampshire

The website of New Hampshire cases has a link for each year's cases from 2008 to 2018. Clicking on one of the links directs us to the web page that lists all the downloading links of case records in the respective year. Inside each of these web pages, clicking on one of the downloading links downloads the clicked case as a pdf file.

To automate the downloading process, we created and configured a Selenium Webdriver using a Python script. It first opened up the website of New Hampshire cases with a Chrome browser. Then we looked at the website's html and found the XPath of all links of years from 2008 to 2018. The web driver located all the links using the XPath and then clicked each of the links from 2008 to 2018 in a loop.

Year	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	total
# of cases	140	146	150	146	136	94	113	100	92	79	86	1282

Table 1: Number of files downloaded including both civil and criminal cases from New Hampshire supreme court

Each time after clicking on a specific year, the web driver located all the downloading links based on the XPath we found for case titles and downloading links in the website's html, and then it clicked through all the downloading links to download all the pdf case files in the clicked year. The pdf files were downloaded into a directory configured when creating the web driver. The pdfs of different years were automatically saved inside different folders inside the download directory.

The files contained both civil cases and criminal cases, and we only needed to look at criminal cases. We filtered all cases to extract only criminal cases using the text information inside the pdf files in the following data pre-processing step for cases in New Hampshire. Table 1

Downloaded pdf files: *data/pdf/pdf_nh*
Python script: *src/scraper/download_nh_cases.py*

2.2.2 Rhode Island

The website of Rhode Island cases also has a link for each year's cases from 2008 to 2018, which can be located by XPath from the website's html. Clicking on one of the links directs us to the web page that lists all the cases in the clicked year. For each case, it shows the case title, the case head notes and the downloading link for the case text as a pdf file.

The website has two lists of links, one lists the court opinions of each year and one lists the court orders of each year. To automate the downloading process, we again created and configured a Selenium Webdriver to open up the website. Then we extracted the XPath of the court opinions' year list, and the web driver used the XPath to locate all the years of case opinions and clicked each year from 2008 to 2018.

Each time after clicking on a specific year, we extracted the XPath containing the case title and the case's pdf downloading link from html and then let the web driver click the downloading links to download the pdfs using the XPaths. The XPath structures were different for 2008 $\sim$ 2011, 2011 $\sim$ 2017 and 2017 $\sim$ 2018, and we modified our code into 3 sections to adapt and download using different XPath structures. The pdfs of different year periods were downloaded into the separated folders.

Since text information of cases in Rhode Island does not contain clear indicator of whether a case is criminal or civil, we only downloaded cases that had case titles starting with "State v", which fit in the profile of a criminal case because all criminal cases had state as the prosecutor. And we skipped cases where the defendant was an organization instead of an individual to filter out lawsuits between state and non-human entities. In this way, we only downloaded criminal cases from all cases on the website for 2008 to 2018. Table 2.

Year Period	2008 $\sim$ 2009	2009 $\sim$ 2010	2010 $\sim$ 2011	2011 $\sim$ 2012	2012 $\sim$ 2013	2013 $\sim$ 2014	2014 $\sim$ 2015	2015 $\sim$ 2016	2016 $\sim$ 2017	2017 $\sim$ 2018	total
# of criminal cases	31	31	37	7	46	40	21	10	11	7	241

Table 2: Number of files downloaded including only criminal cases from Rhode Island supreme court

Python script: *src/scraper/download_ri_criminal_cases.ipynb*
Downloaded pdf files: *data/pdf/pdf_ri_crminal*

pdf file name	case title	case type (criminal or civil)	case decision	case text
string	string	criminal non-criminal	affirmed affirmed in part not affirmed	list of strings

Table 3: contents and format of a tuple

2.3 Data Pre-processing

Our goal of data pre-processing is to remove invalid data entries and then arrange data of each case into a tuple containing pdf file name, case title, case type, case decision and case text. Table 3

The list of tuple could be stored inside a csv file or a json file for further data analysis.

For both New Hampshire and Rhode Island, the pdf file name and case title were already recorded when downloading the cases. The case text of each case was extracted from the pdf file using **pdftotext** library.

2.3.1 New Hampshire

We extracted the case type of New Hampshire case by looking for the RSA number in the third paragraph of the case text from pdf. By the legal document RSA of RI, if the labeled RSA number of a case is between 625 and 652, then the case is criminal and we label the case as "criminal". If RSA number was out of that range or we do not find a RSA, we label the case as "non-criminal".

The case decision for New Hampshire was written at the end of each pdf file. We wrote a script to split each text into paragraphs which helped us with locating and extracting the case decision. Our case decisions were recorded as "affirmed", "affirmed in part" or "not affirmed" based on our extracted information.

Python script: *src/scraper/get_nh_cases.py*
Processed data files: *data/*

2.3.2 Rhode Island

The case type of Rhode Island cases were all "criminal", because we already only collected cases that were criminal when we downloaded the pdf files of the cases.

The case decision for New Hampshire was written in a section called conclusion in each pdf file. We again wrote a script to split each text into paragraphs to find a paragraph called "conclusion" and the case decision followed right after it. Same as New Hampshire, our case decisions were recorded as "affirmed", "affirmed in part" or "not affirmed" based on our extracted information.

Python script: *src/scraper/get_ri_cases.ipynb*
Processed data files: *data/*

3 Data Analysis

Our data representation is

	file name	title	type	decision	text
0	06-290.pdf	State v. Michael Tetreault, No. 06-290 (June 1...	criminal	affirmed	['', 'Supreme Court', 'No. 2006-290-C.A.', '(P...
1	08-27.pdf	State v. Thomas P. Byrne, No. 08-27 (June 19, ...	criminal	not affirmed	['', 'Supreme Court', 'No. 2008-27-C.A.', '(P2...
2	07-108.pdf	State v. Robert Collazo, No. 07-108 (April 3, ...	criminal	affirmed	['', 'Supreme Court', 'No. 2007-108-C.A.', '(P...
3	07-334.pdf	State v. Samuel Adewumi, No. 07-334 (March 17,...	criminal	affirmed	['', 'Supreme Court', 'No. 2007-334-C.A.', '(W...
4	07-123.pdf	State v. Phillip Jackson, No. 07-123 (March 20...	criminal	not affirmed	['', 'Supreme Court', 'No. 2007-123-C.A.', '(P...

3.1 Data Visualization

To find the relevance between confidential informant cases and the final decision of cases, we need to find the distribution of decision in confidential informant cases and cases which are not confidential informant cases.

There are three kinds of final decision: "affirmed", "partly affirmed" and "reversed". By extracting keywords that we need, the total number of all kinds of cases are shown in Table 4.

State	New Hampshire			Rhode Island		
Decision	affirmed	partly affirmed	reversed	affirmed	partly affirmed	reversed
C.I. cases	11	1	3	10	0	3
non-C.I. cases	183	8	63	196	5	26

Table 4: The number of all kinds of cases. In this table, "confidential informant" is abbreviated as "C.I.".

Moreover, we draw some pie graphs to show the distribution of cases. Details are shown in Figure 1 and Figure 2.

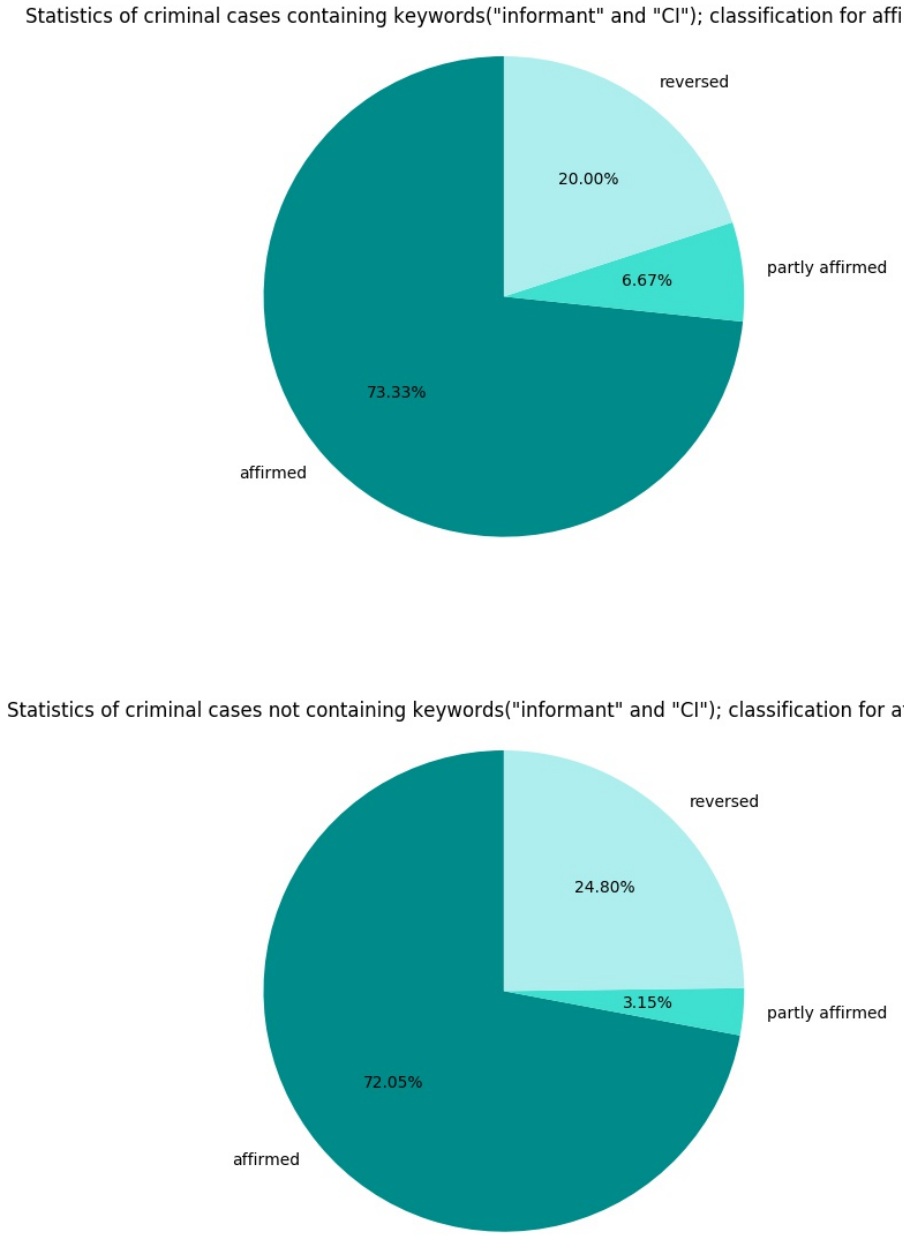

Figure 1: Distribution of cases in New Hampshire State

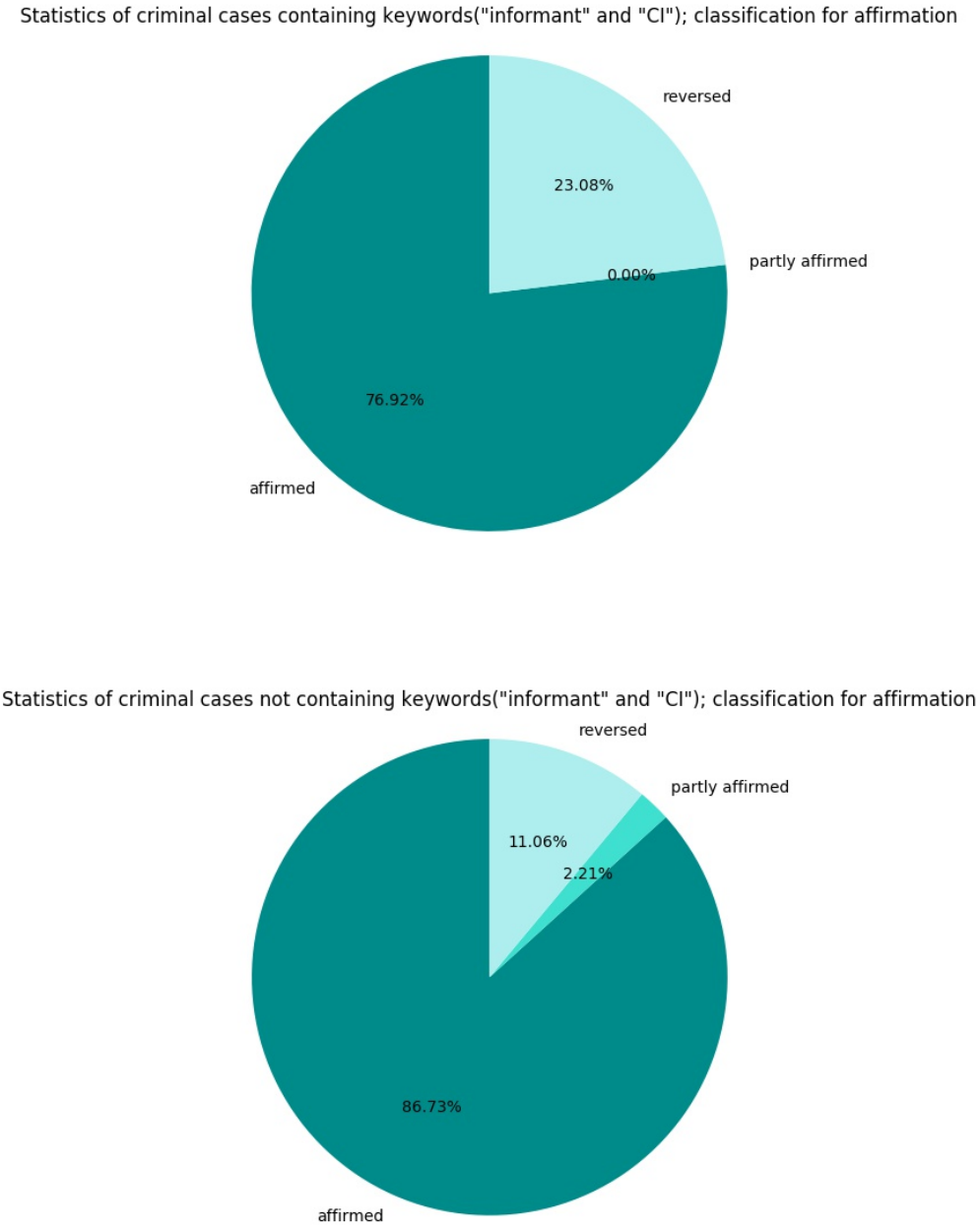

Figure 2: Distribution of cases in Rhode Island State

According to these pic pictures, we find:

• In New Hampshire State, the final decision of cases is not relevant to the confidential informant in cases.

• However, in Rhode Island State, confidential informant cases are more likely to be reversed than that of non-confidential informant cases. So when there is more effective evidences from confidential informant, case appeal is less likely to succeed.

As is stated above, the circumstances of the case may differ greatly in different states, which may be caused by different laws in different states; or the confidential informant in different states provides different information for the case; perhaps the information provided by confidential informant in RI state is more useful than confidential informant in NH state. This is what we think causing the above phenomenon (see Figure 1 and Figure 2).

3.2 Data Cleaning

Use Tf-idf to delete common words that appear many times in every case. For example, in the New Hampshire's data, 'New', 'Hampshire', 'court', 'cases' will appear many times. But these words are not important. To solve this, we input all cases into Tf-idf model to get the most frequent 20 words. Then, we delete this words from our cases' text.

Remove the number, punctuation and stop words. Stop words refer to meaningless words, like 'I','a','be'. These text are not important to use, so we delete them.

3.3 Tf-idf

We use Tf-idf to get the most frequent features in criminal cases, affirmed cases, reversed cases and partial affirmed cases. We found in NH's data, reversed cases have two frequent features: sexual and child. These two features are not appearing in affirmed frequent features. Figure 3. We found in RI's data, there is no clear pattern. Figure 4.

Figure 3: Affirmed and Reversed Cases in New Hampshire

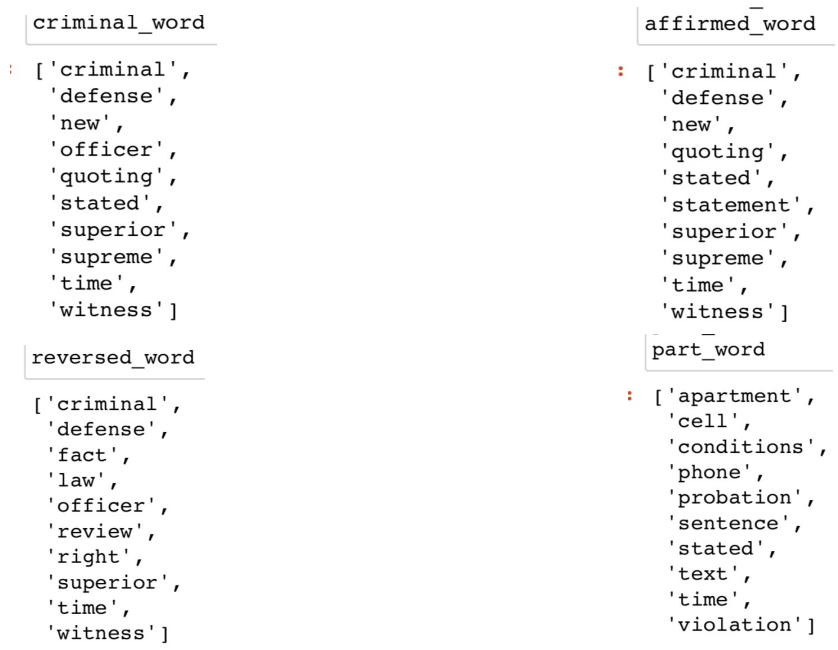

Figure 4: Affirmed and Reversed Cases in Rhode Island

We also use Tf-idf to get the most frequent features in criminal cases with informants and without informants. There is no clear pattern in New Hampshire data with informants and without. Figure 5. For Rhode Island, we found 'murder' feature in criminal cases with informants. But 'murder' does not appeal in non-informant frequent features. We may get conclusion that cases with informants are more likely related to 'murder'. Figure 6.

Figure 5: Informant in New Hampshire

Figure 6: Informant in Rhode Island

For reversed cases, we select the most frequent 10 words in cases with informant and without. In New Hampshire, we found 'sexual' feature in reversed cases without informants. Figure 7. We may get conclusion

that the reversed cases related to sex don't have informants. In Rhode Island, we found 'child' in reversed cases with informants. Figure 8. We may get conclusion that the reversed cases related to child may have informants.

```
info_r_word
```

```
['circumstances',
 'confidentiality',
 'detective',
 'home',
 'officer',
 'officers',
 'promise',
 'requirement',
 'search',
 'warrant']
```

```
non_info_r_word
```

```
['assault',
 'conduct',
 'criminal',
 'defense',
 'issue',
 'motion',
 'person',
 'quotation',
 'reasonable',
 'sexual']
```

Figure 7: Informant in Reversed Cases For New Hampshire

```
info_r_word
```

```
['baccaire',
 'child',
 'defense',
 'det',
 'error',
 'information',
 'laforest',
 'phillip',
 'right',
 'said']
```

```
non_info_r_word
```

```
['criminal',
 'defense',
 'fact',
 'issue',
 'law',
 'officer',
 'review',
 'superior',
 'time',
 'witness']
```

Figure 8: Informant in Reversed Cases For Rhode Island

3.4 Word2Vec

In this section, we want to find the similarity between the 'reversed' and other words in reversed cases. First, we put all reversed cases together, that is, put all words appeared in reversed cases into a list. Then we use Word2vec to find the similarity. We found that the word with the highest similarity to 'reversed' is 'remanded' in New Hampshire. Figure 9. We found that the word with the highest similarity to 'reversed' is 'remand' in Rhode Island. Figure 10. We can not get any useful conclusion from the 10 most similar words.

```
word2vec_model.wv.most_similar('reversed')
```

```
[('remanded', 0.9986464977264404),
 ('dissented', 0.9959361553192139),
 ('superior', 0.9943211674690247),
 ('broderick', 0.9942866563796997),
 ('conflict', 0.9933634996414185),
 ('appointed', 0.993209958076477),
 ('johnson', 0.9921435713768005),
 ('dugas', 0.9921237826347351),
 ('green', 0.9920369386672974),
 ('respondent', 0.9920048117637634)]
```

Figure 9: the Similarity Between 'reversed' and Other Words in Reversed Cases for NH

```
word2vec_model.wv.most_similar('reversed')
```

```
[('remand', 0.9991968274116516),
 ('district', 0.9991153478622437),
 ('instant', 0.99909508228302),
 ('specifically', 0.999093770980835),
 ('new', 0.9990875720977783),
 ('pepper', 0.9990874528884888),
 ('conclusion', 0.9990872144699097),
 ('shower', 0.9990850687026978),
 ('including', 0.9990758895874023),
 ('close', 0.9990744590759277)]
```

Figure 10: the Similarity Between 'reversed' and Other Words in Reversed Cases for RI

3.5 N-gram

Because using word2vec method cannot offer useful results of words similarity, we try the N-gram method to analysis text data. An N-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n-1)$ order Markov model.

Using N-gram, we can extract important language contexts which has high weights from 'text' data. For example, if some language sequence often appears in many cases, we can get those sequence and try to find the relationship between these sequences and other things such as the case decision or if the case has informant.

3.6 Visualization of Ngram results

To make the model better fits our data, we visualize the n-gram results into a graph by utilizing *networkx* package. to find the best parameters. We found that if we give a high weight factor for each sequence, the plot will be more obvious, there will be some Aggregation and sparse place in the graph, it can help us find the relationship among sequence more quickly.

In the graph, each node is a word, we add edges by n-gram results. All the nodes are sorted by weights, we only add top 100 nodes into graph, this can avoid analysing useless text information.

Besides, we give different color for graph nodes by node degree. If a node degree is high, it means that it might be more important because it connects with many other nodes. If a node degree is low, it means that it might be less important or it has its own context and far away from other aggregation parts.

3.6.1 New Hampshire

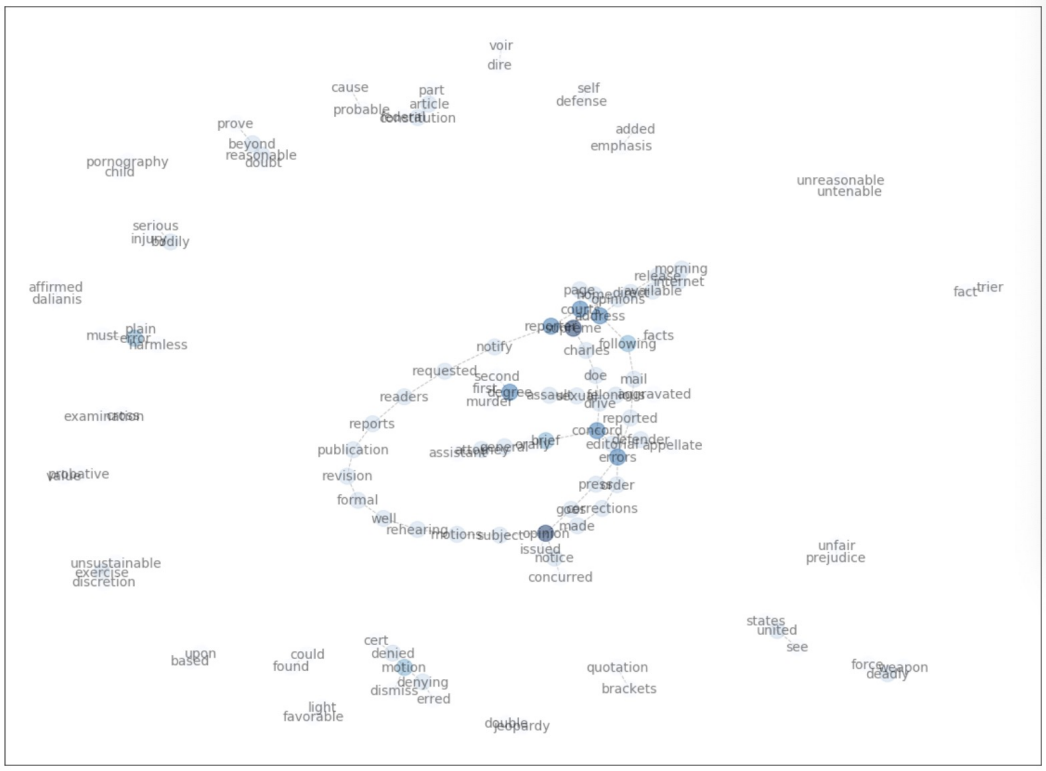

Figure 11: New Hampshire criminal cases which has affirm decision

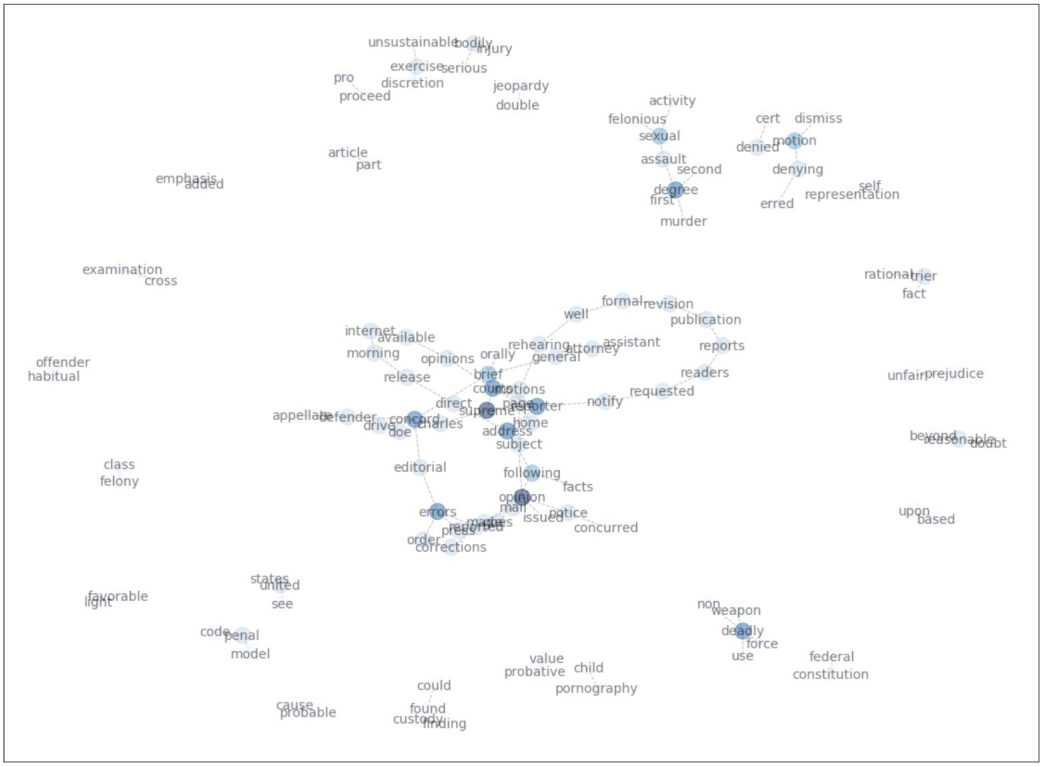

Figure 12: New Hampshire criminal cases which has reverse decision

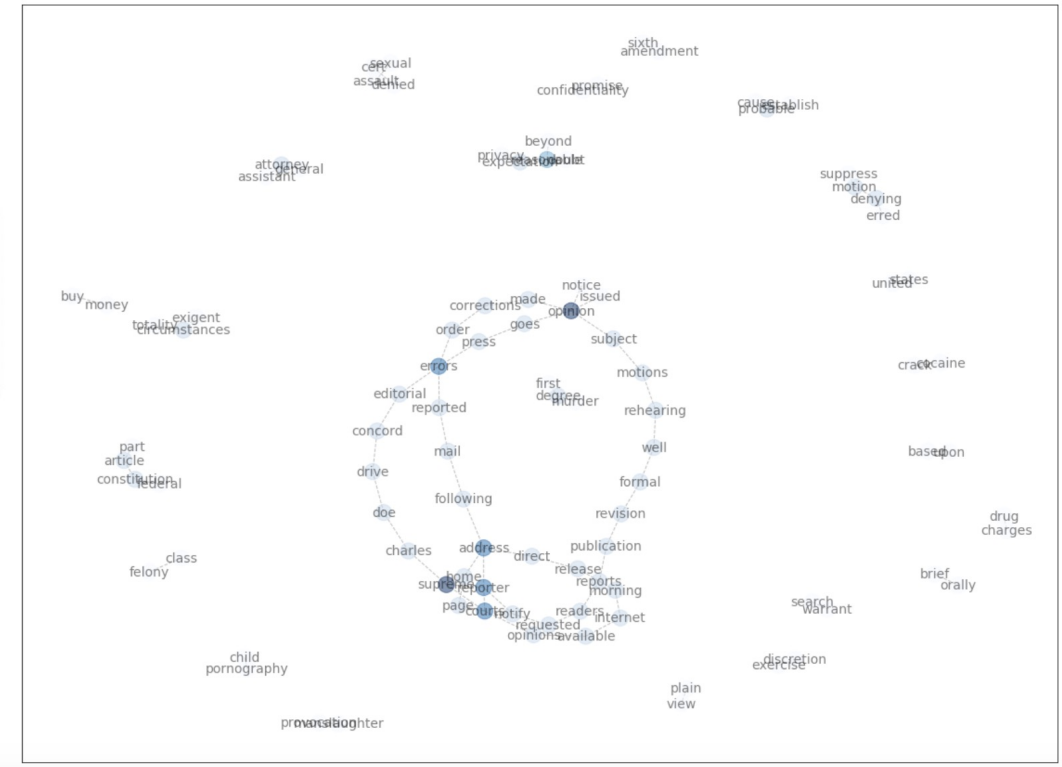

Figure 13: New Hampshire cases with informant

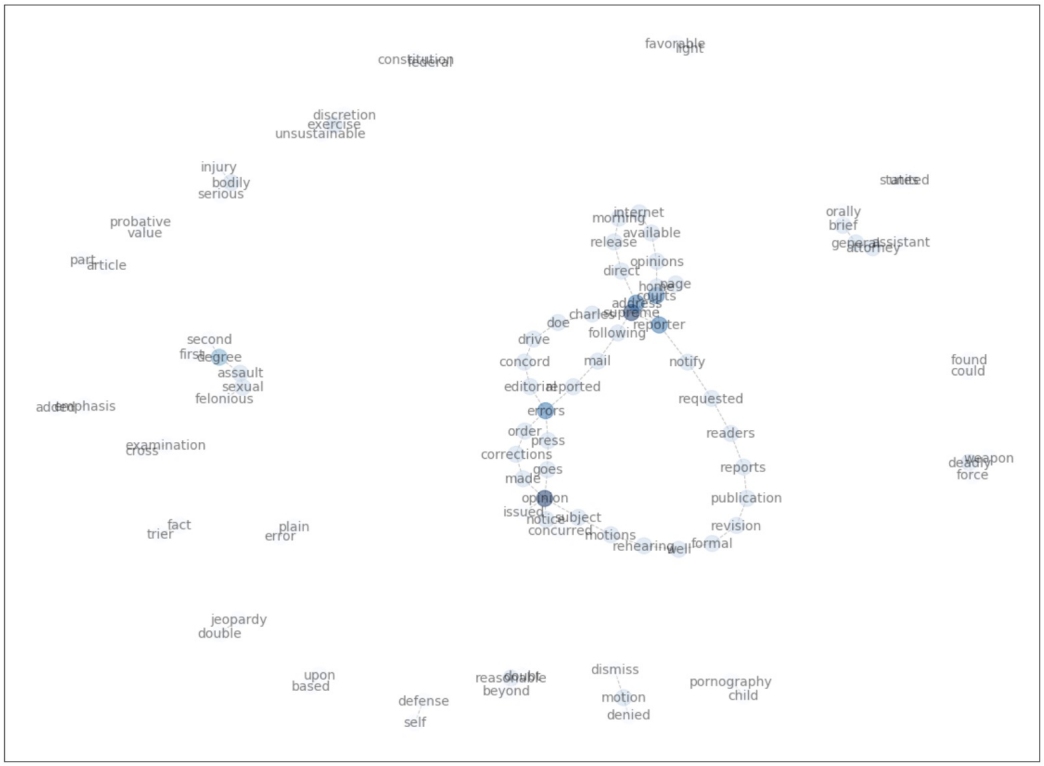

Figure 14: New Hampshire cases without no informant

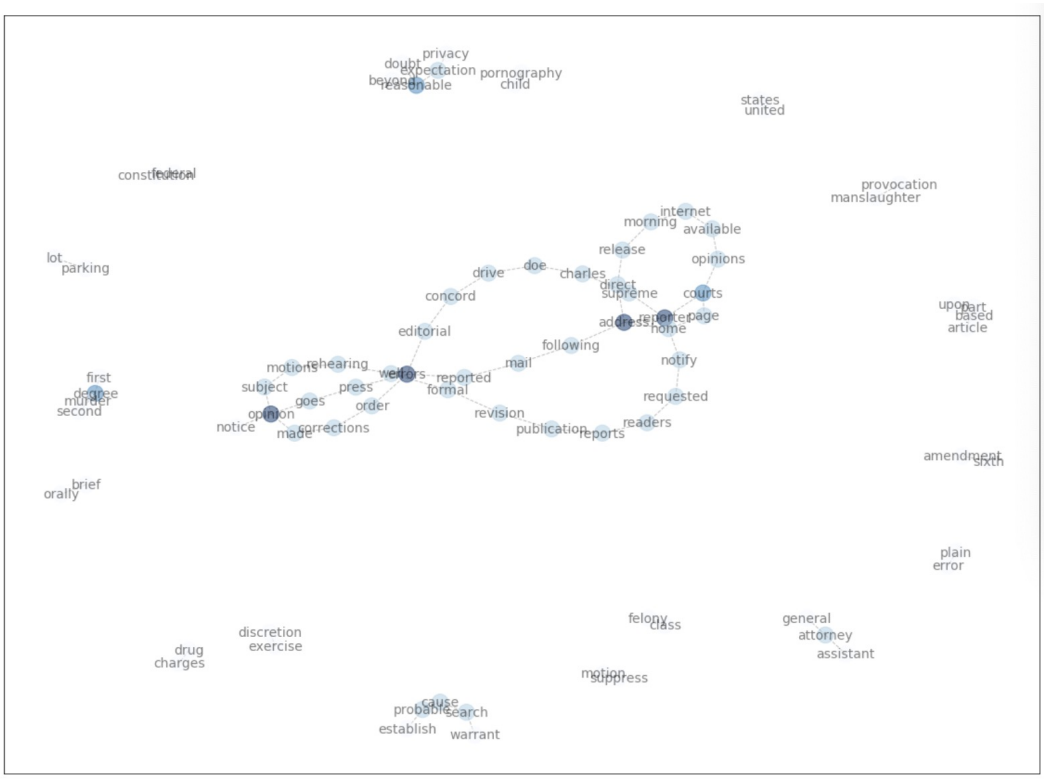

Figure 15: New Hampshire cases with informant and decision is affirm

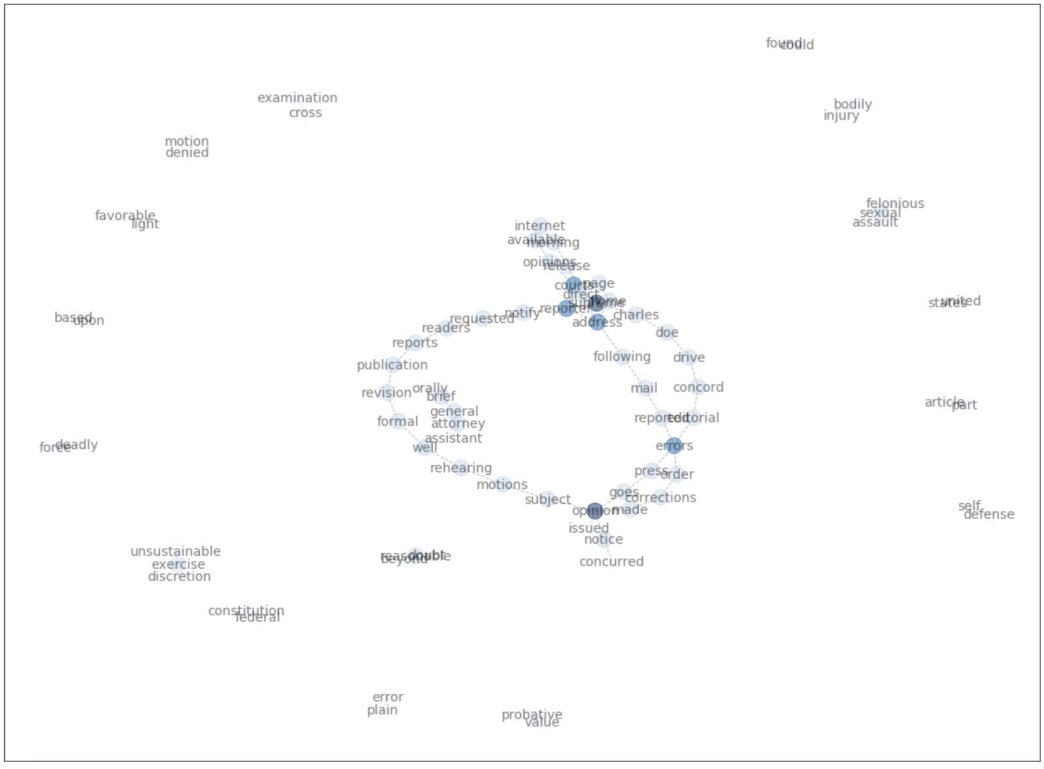

Figure 16: New Hampshire cases which without informant and decision is affirmed

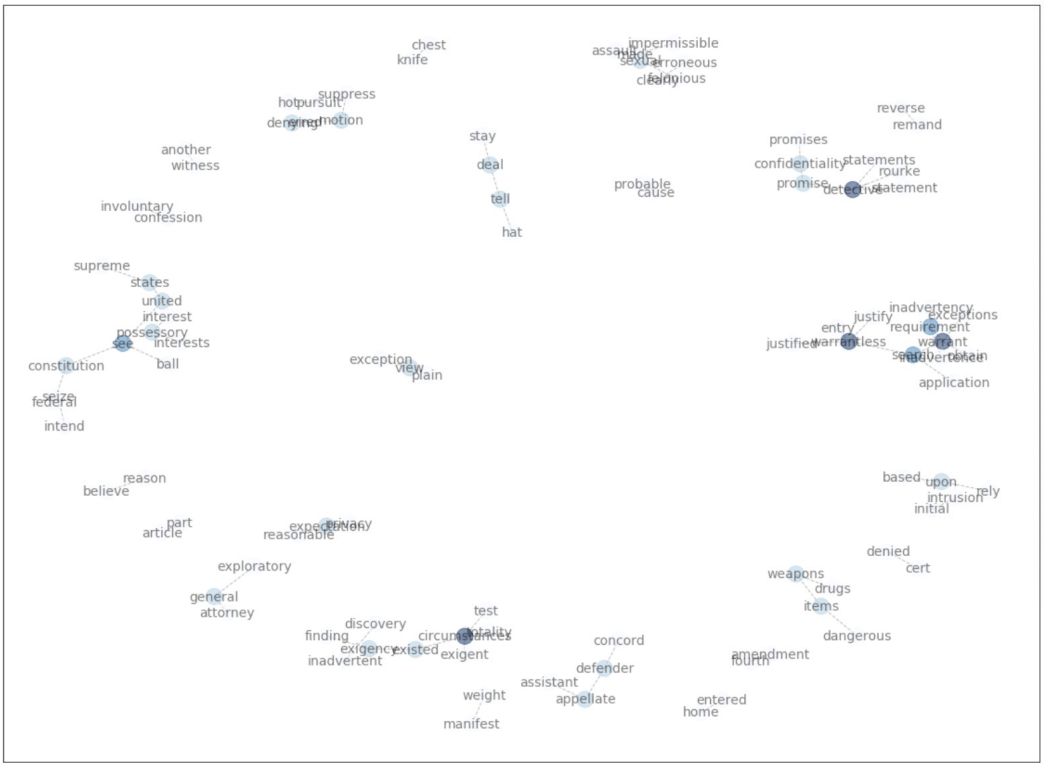

Figure 17: New Hampshire cases with informant and decision is reversed

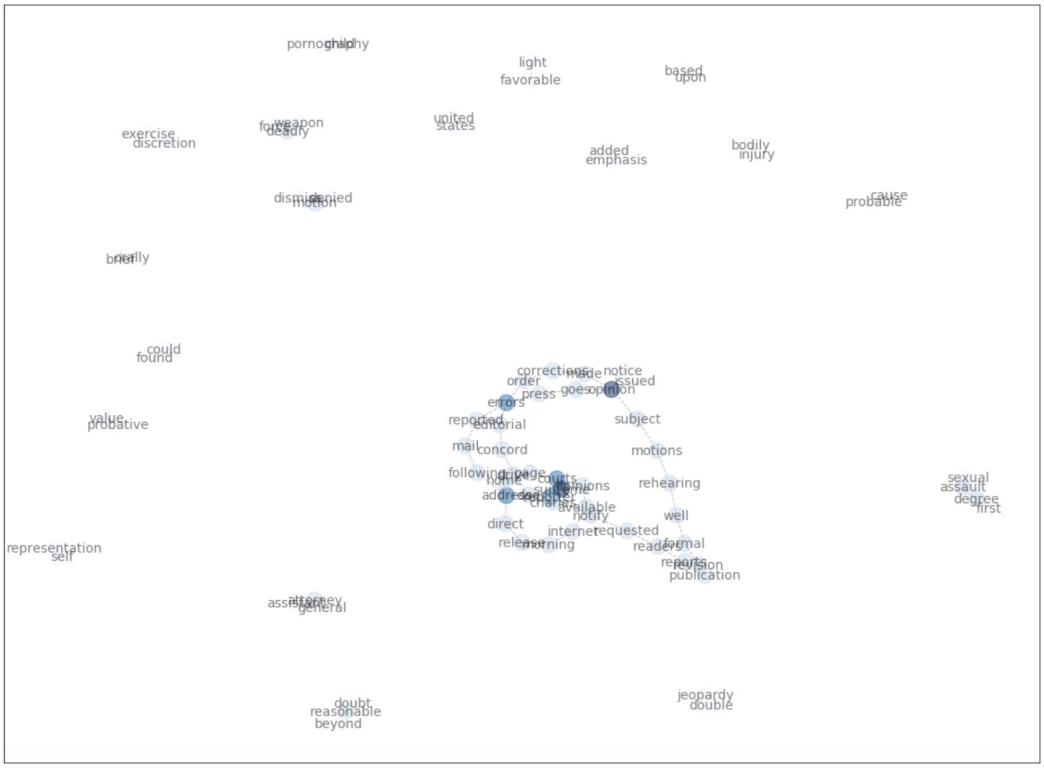

Figure 18: New Hampshire cases without informant and decision is reversed

3.6.2 Rhode Island

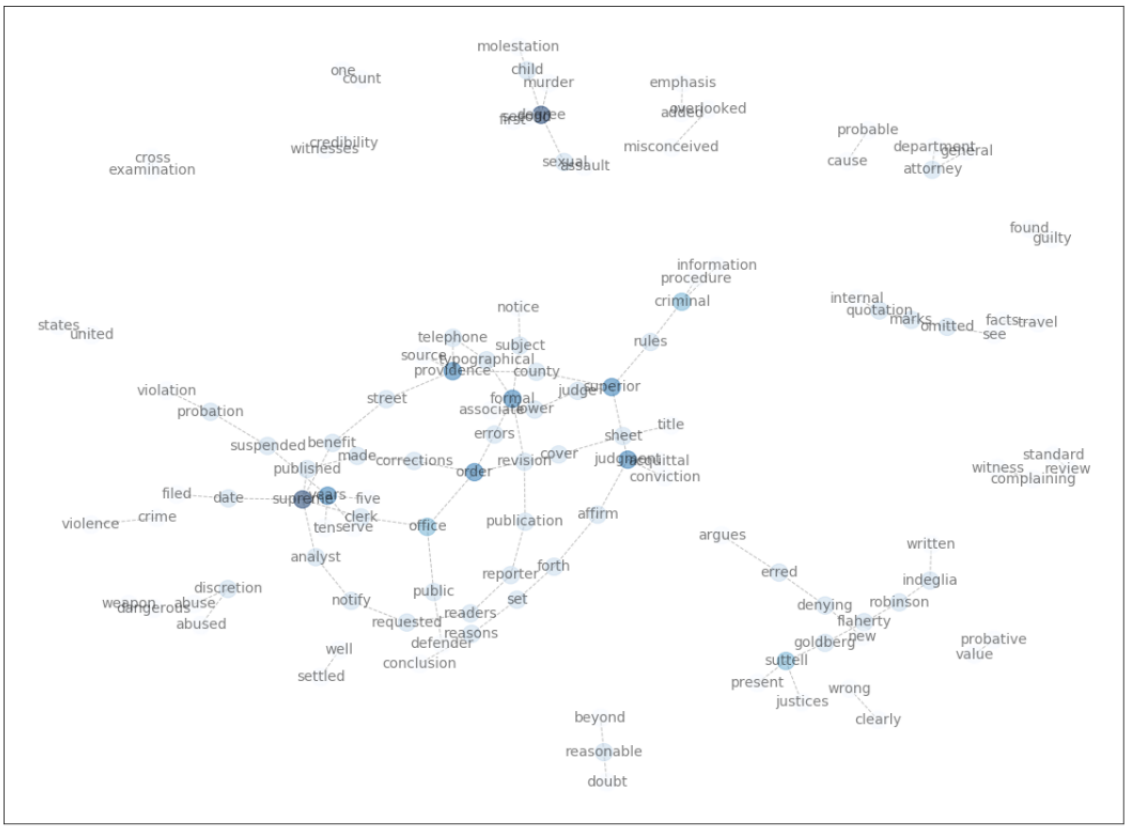

Figure 19: Rhode Island criminal cases which has affirmed decision

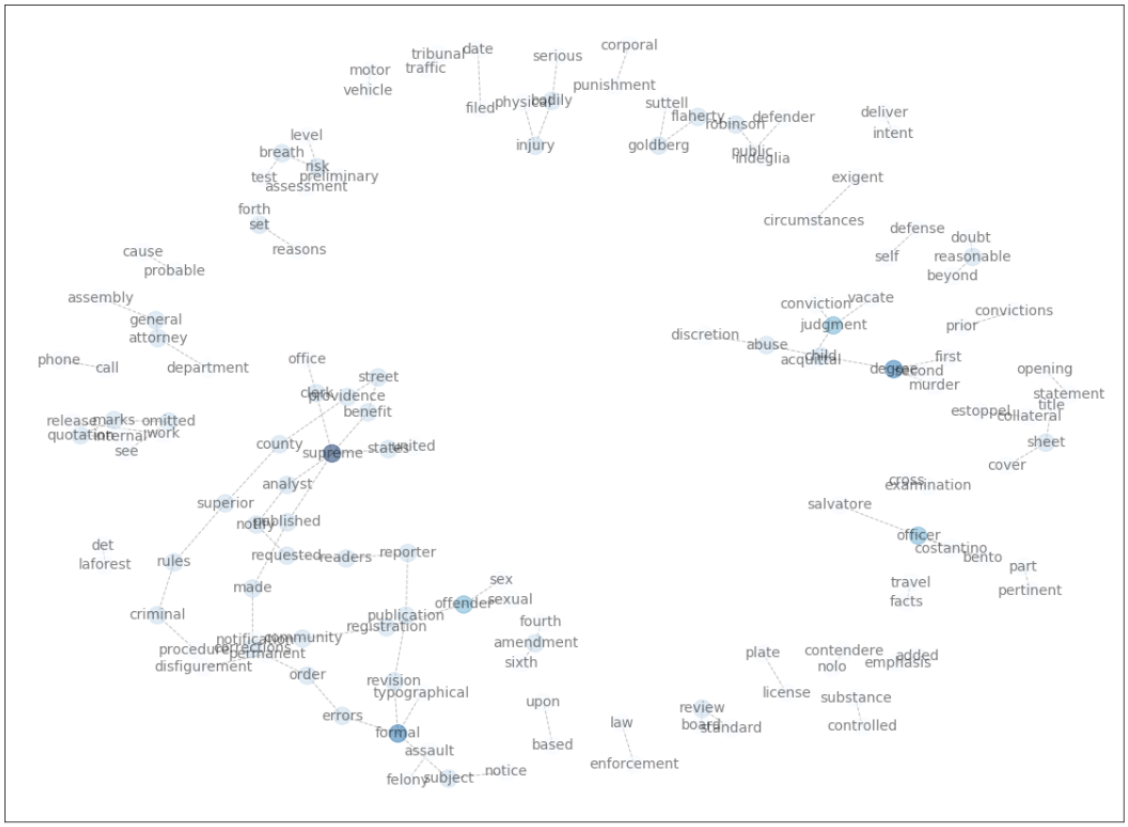

Figure 20: Rhode Island criminal cases which has reverse decision

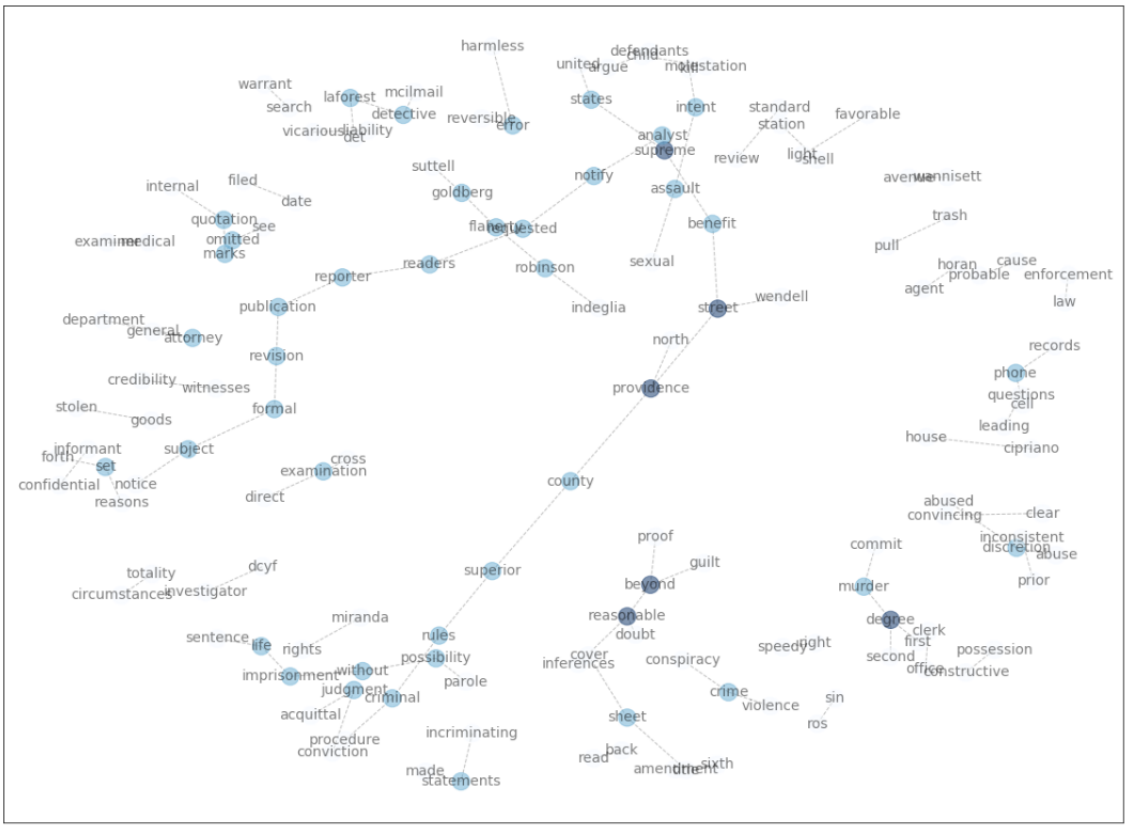

Figure 21: Rhode Island criminal cases with informant

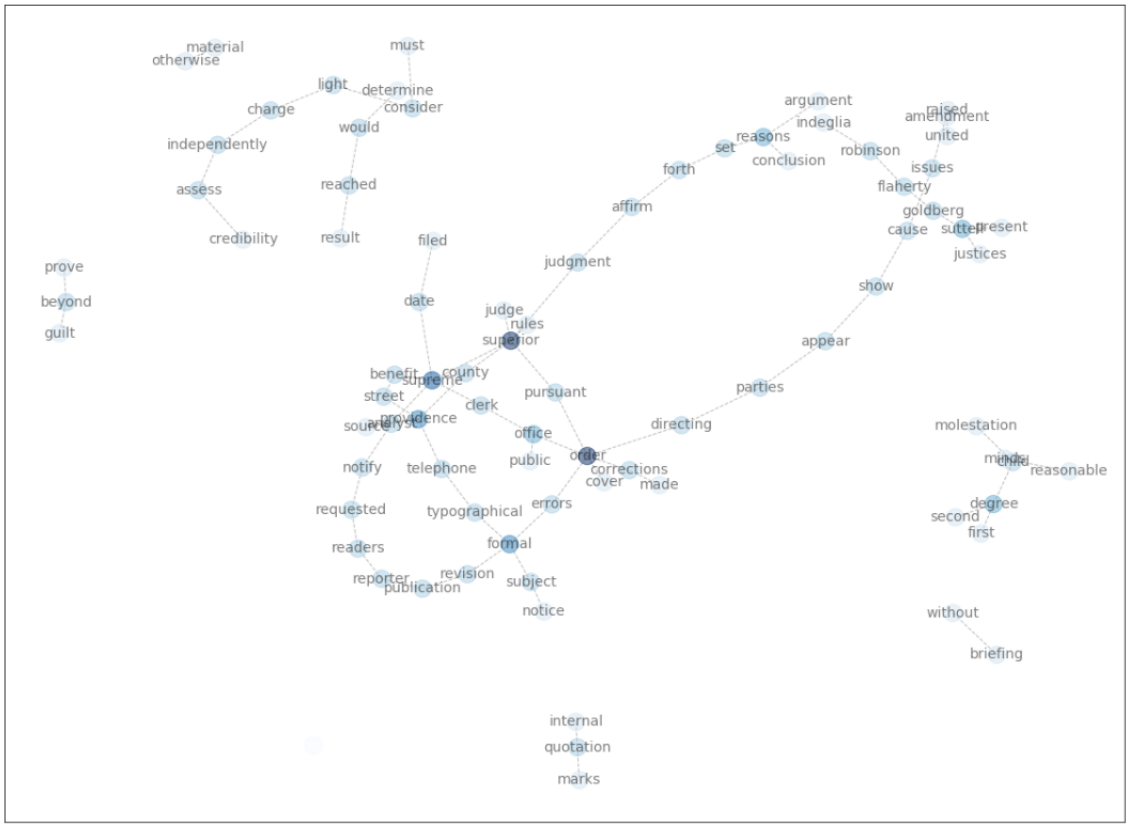

Figure 22: Rhode Island criminal cases without informant

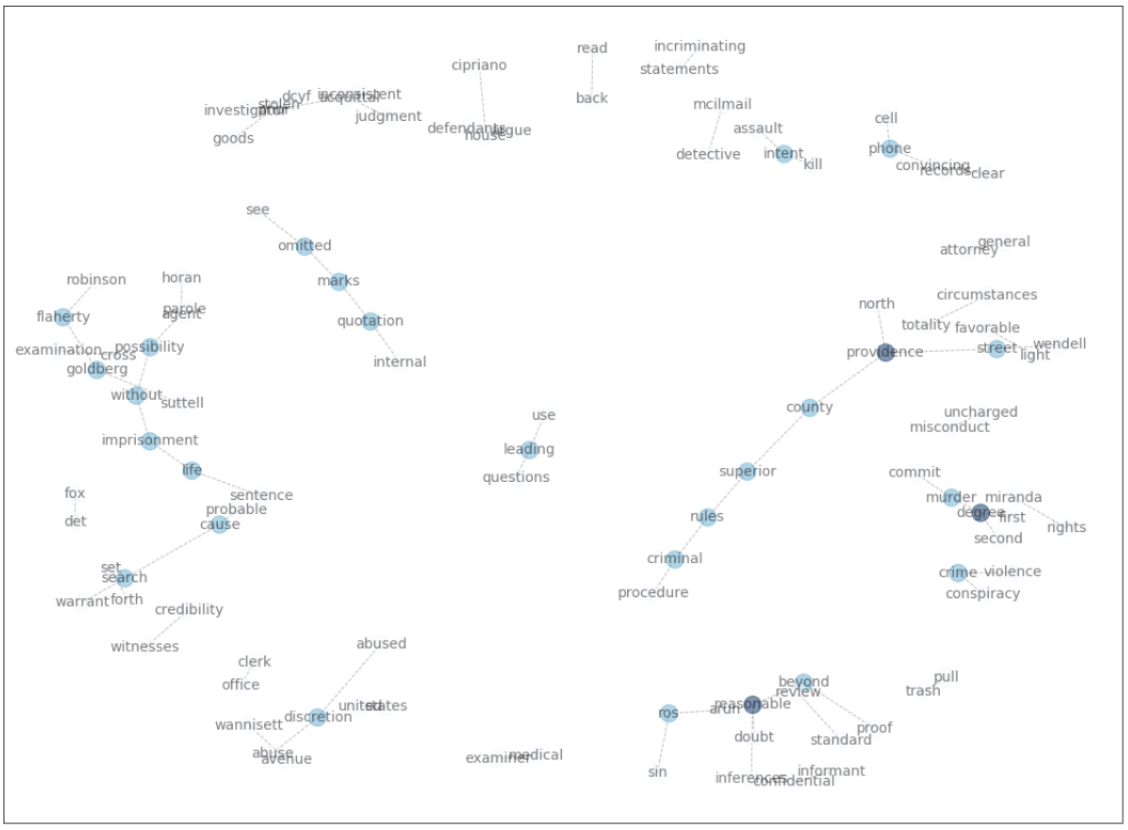

Figure 23: Rhode Island criminal cases with informant and decision is affirm

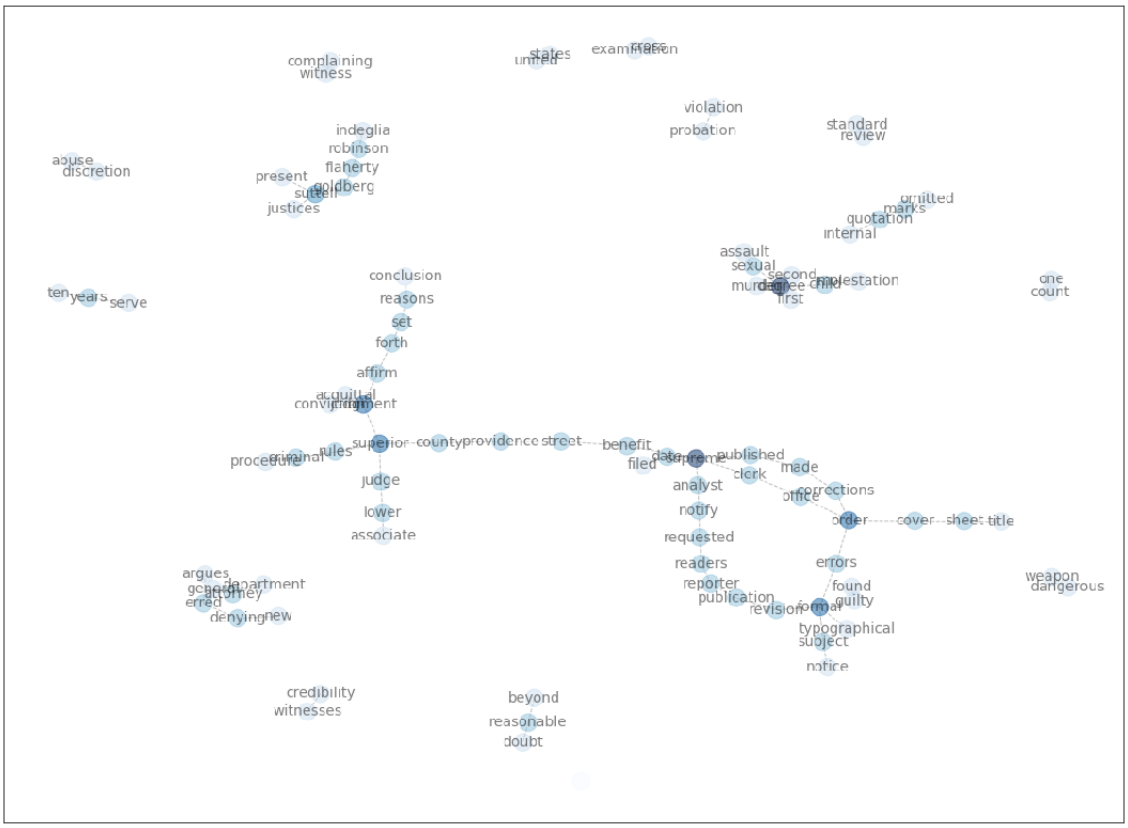

Figure 24: Rhode Island criminal cases which without informant and decision is affirm

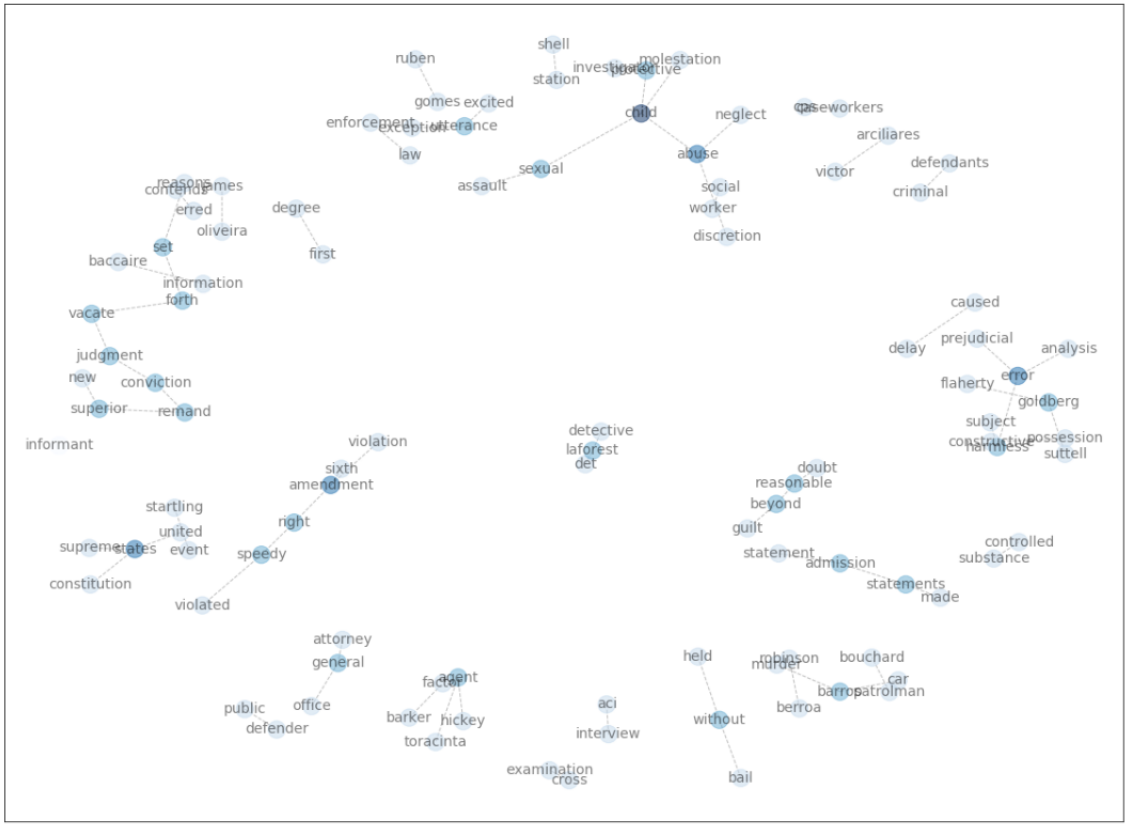

Figure 25: Rhode Island criminal cases with informant and decision is reversed

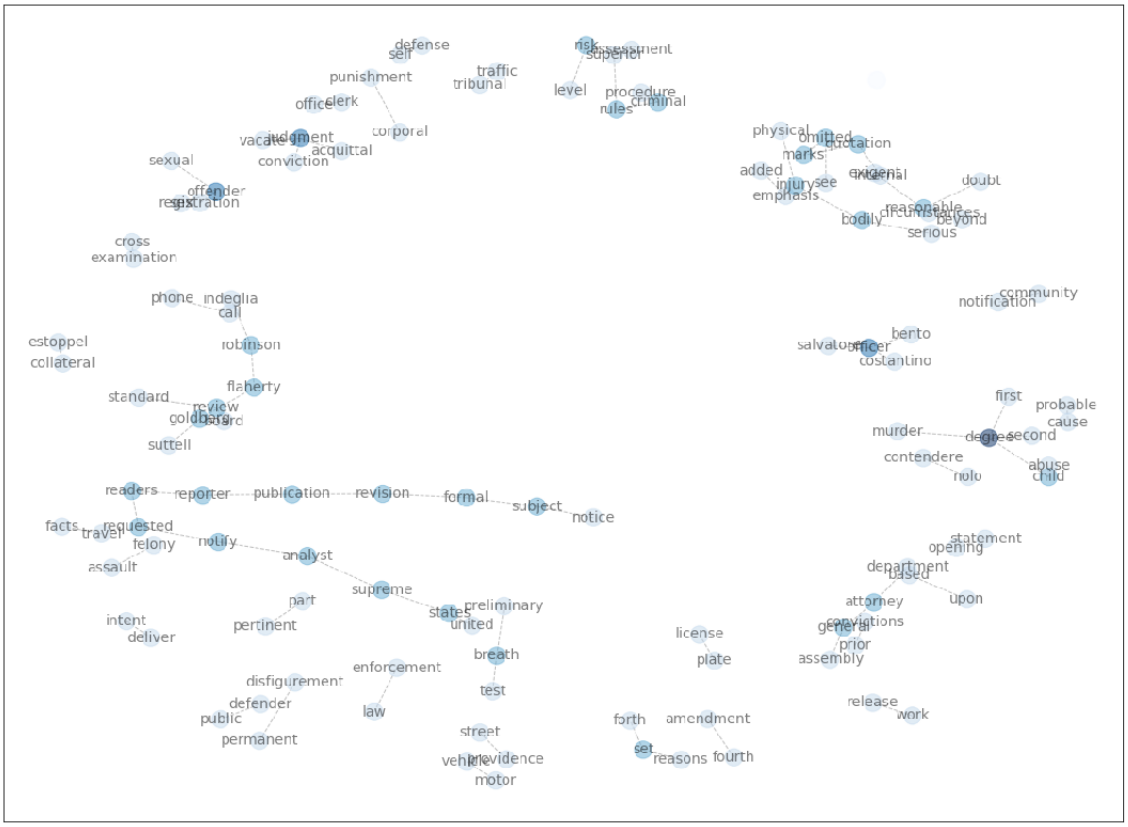

Figure 26: Rhode Island criminal cases which without informant and decision is reversed

4 Conclusion

In this project, we worked with the data of criminal cases from appellate courts in both New Hampshire and Rhode Island with steps of data collection, pre-processing and detailed analysis using Tfidf and N-gram.

4.1 NH conclusion

4.1.1 Relationship between Case Text and Criminal Cases Decision

The difference is shown in the table.

Decision	NH Criminal Cases
affirmed	self defense
reversed	self representation, felony

We found that in NH state,

* criminal cases about self representation cases can be more likely to be reversed. According to the information we search online, we know that in many areas, the defendant will be provided with a prosecution lawyer when appealing. This will not incur costs, so the self represented defendant can choose to appeal, which has no loss to him.

* criminal cases about self defense cases can be more likely to be affirmed. We can imagine a case that defendant appeal because he think he is self defense but actually he is not.

4.1.2 Relationship between Case Text and Informant/Non-informant Cases

The difference is shown in the table.

Type	NH Criminal Cases
with informant	cocaine, drug charges sixth amendment search warrant exigent circumstance probable cause
without informant	deadly weapon body injury

We found that in NH state, criminal cases having informant are

* more related to **drug problems** because of "cocaine", "drug charges" keywords. We think it is because drug cases often need informant to get case details. Normal people are hard to get this details.
* more related to **search warrant problem** because of "search warrant", " probable cause" keywords.
* more related to protect defendant's right because of "sixth amendment" keyword.

4.1.3 Relationship of Affirmed Criminal Case with or without Informant Cases

The difference is shown in the table.

Type	NH Affirmed Criminal Cases
with informant	child pornography drug charges provocation manslaughter sixth amendment search warrant felony probable cause
without informant	felonious sexual assault bodily injury self defense

We found that in NH state, affirmed criminal cases having informant are

* more related to **drug problems** because of "drug charges" keywords. We think it is because drug cases often need informant to get case details. Normal people are hard to get this details.
* more related to **search warrant problem** because of "search warrant", " probable cause" keywords.
* more related to protect defendant's rights because of "sixth amendment" keyword.

4.1.4 Relationship of Reversed Criminal Case with or without Informant Cases

The difference is shown in the table.

Type	NH Reversed Criminal Cases
with informant	dangerous items, drugs another witness felony assault sexual confidentiality, promise deal search warrant, enter home, probable cause, exigent circumstance involuntary confession justified warrantless entry, detective inadvertent, Totality of the Circumstances fourth amendment
without informant	self representation exercise of discretion

We found that in NH state, reversed criminal cases having informant are

* more related to **drugs and dangerous problems**, this is the same as the conclusion we get before.

* more related to **search warrant problem**, but for reversed cases, we found "enter home" keyword. This might be because entering home to search is a sensitive topic to people. However, if the informant can provide valid information, the police have probable cause to apply for a search warrant. In both affirmed and reversed cases, we found "search warrant" and "probable cause" keywords, hence, we can only have a conclusion that informants are related to offer probable cause which makes police have search warrant. Warrant validity is important but this cannot be learned from n-gram. Informant **cannot have direct influence on the case decision**. Meanwhile, we found that in exigent circumstance, police can also enter home to search, but they need to apply a warrant after searching.

* more related to "deal" problem. sometimes, useful information is got under some special deal by informant.

* "involuntary confession" keyword. Reversed cases appear these keywords may represent the defendant compelled to confess guilty in the first-instance judgment, and later the case appealed.The criminal justice system relies on confessions by defendants to help prove guilt at trial or to induce a guilty plea, which is controversial.

* "another witness" might be related to helping defendant to prove his innocent which made the appeal is more likely to be reversed.

4.2 RI conclusion

4.2.1 Relationship of Criminal Cases Decision

The difference is shown in the table.

Decision	RI Criminal Cases
affirmed	witness credibility, found guilty
reversed	sixth/fourth amendment, self defense Nolo contendere corporal punishment self-defense, exigent circumstance motor vehicle, license plate

We found that in RI state,

* Affirmed cases have the key word "witness credibility". In the presence of witnesses, the case is less likely to be overturned. This may be because there are more evidence.

* In reversed case, we found the key word "Nolo contendere'. This may be because the defendant applied nolo contendere, and the court made a judgment. However, the defendant did not agree with the judgment and therefore appealed. Final, the Supreme Court reversed the original sentence.

4.2.2 Relationship between Case Text and Informant/Non-informant Cases

The difference is shown in the table.

Type	RI Criminal Cases
with informant	probable cause, search warrant confidential informant life sentence life imprisonment without the possibility of parole (first degree murder) credibility witness sixth amendment conspiracy crime of violence beyond reasonable doubt
without informant	-

We found that in RI state, criminal cases having informant are

* more related to very serious cases because of "life imprisonment", "life sentence", "beyond reasonable doubt" keywords.

* more related to search warrant problems, which is similar to the results we analyze in NH state.

* informants might be more helpful for the courts' conviction of 'conspiracy crime of violence'.

* more related to protect defendant's rights.

4.2.3 Relationship of Affirmed Criminal Case with or without Informant Cases

The difference is shown in the table.

Type	RI Affirmed Criminal Cases
with informant	conspiracy crime of violence search warrant, probable cause uncharged misconduct incriminating statement intent of assaults and kill life sentence life imprisonment without the possibility of parole (first degree murder) detective inconsistent judgement
without informant	weapon dangerous violation probation sexual ten years serve abuse of discretion

We found that in RI state, affirmed criminal cases having informant are

* more involved with affirming cases of sentences and imprisonment for life, but cases affirmed without informant tend to have lighter prison sentence of only ten years serve.

* From "uncharged misconduct" keyword, we learn about **Uncharged Misconduct Under Rule 404(b): The Admissibility of Inextricably Intertwined Evidence**. Informants can provide evidences of other crimes, wrongs, or acts for case affirmation. Had the informants avoided mentioning the uncharged act, his testimony would have been awkward, and the narrative would have sounded less credible.

* less involved with sexual, weapon crimes.

4.2.4 Relationship of Reversed Case Text with or without Informant Cases Decision

The difference is shown in the table.

Type	RI Reversed Criminal Cases
with informant	child molestation, sexual assaults against child child protective service investigation sixth amendment beyond reasonable doubt held without bail (dangerous) detective
without informant	nolo contendere first and second degree murder motor vehicle, license plate

We found that in RI state reversed criminal cases having informant are

* more involved with child crimes and child protective service investigation, but reversed cases without informant are more involved with first and second degree of murder.

* less involved with cases using evidence of motor vehicle and license plates

* happening more often with defendants held without bail and cases without informants had more nolo contendere (you admit no guilt for the crime, but the court can determine the punishment.)

5 Challenge

As we were working on the project, we hoped that there could be more data to be fed into our analysis methods. Using more data of criminal cases to continuously analyse the relation between confidential informant and final decision results can be our future work.

The websites of our data source are poorly formatted. For some year periods, the website used different html structures and we had to create different web scrapers accordingly to relocate and validate the pdf files for each year.

We tried lots of methods to read PDF as string, many methods such as pypdf, pdfminor, Tabula-py and etc. They cannot extract PDF text well, sometimes they lost many important words or paragraph which might be difficult for us to analyze data. And we finally used **pdftotex** because of its best performance.

When we extract case title, case type and case decision, we have to keep improving and debugging our code to accommodate include all different structures of each case record, since the case records were written in different year periods and not necessarily in the same patterns. In this way, we could make sure the key information was located correctly in the texts.

We hope our analysis for criminal cases in New Hampshire and Rhode Island can help courts and reporters rethinking what can cause obvious effect to the final decision for criminal cases and help courts and confidential informants make better contribution to the criminal cases' decisions in the future.

A major challenge is that we lack legal knowledge and do not understand many legal knowledge and legal provisions. At the beginning, due to insensitivity to some keywords, such as, nolo contendere and involuntary confession, we missed a lot of information. As we have invested a lot of time and energy, we have dug more useful information, which makes our conclusion more perfect.