Final Project Report of Text Mining in Public Policy

Shinya Takatani / UCID: 12147306

1. Introduction

In my project, I would like to propose a total package to improve each policy making process with text mining skills which we will learn in this class. I will apply the technique for the example of "Temporary Assistance for Needy Families (TANF)" . All of my work is stored in my GitHub repository, and the detailed codes are shown in TMproject.ipynb. You can see my presented slide in this repository.

2. Temporary Assistance for Needy Families(TANF)

At first, I will give a brief explanation for TANF. TANF is one of federal assistance program. It started in 1997, succeeding the previous program AFCD. The goal of the project is to help needy families to work and care children self-sufficiently. The block grants are given to state governments so that the program is designed and operated by themselves. The state governments have the authority to terminate payments on some conditions although the general rule is decided by the federal government. For example, the lifetime limit of receiving benefits is 60 months.

At the beginning of the program, it succeeded because it increased the number of working single mothers. However, there are many criticisms now. In fact, the volume of assistance decreased though the need continuously increased and the number of children under poverty line also increased. In addition, TANF funds are used for other purposes such as pre-k education or child care. Only 10% of the funds is used for work activities and work support. Some people say that the states government should invest more in education to enhance earning after employment. Budget constraint lifts up the hurdles of requirements or eligibility. This line is controversial, so local legislatures mainly discuss this type of topics. And TANF has been facing unequal resource allocation among states.

Trump administration cut the budget of TANF by 10% in 2018. It also has announced the order on welfare reform which indicates that the volume of social security or safety net would shrink. The stance has much influence on the reform of TANF, so the policy is a hot topic which should be studied actively.

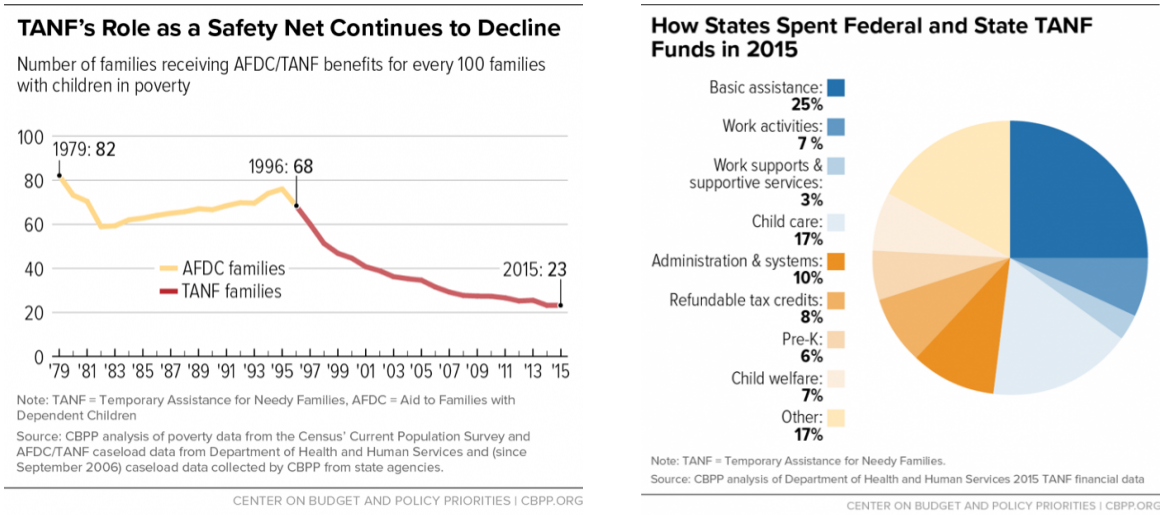


Figure1. Number of families receiving TANF benefits, How to use TANF funds

3. Text mining in public policy

Text Mining or Natural Language Processing has huge potential to be utilized in public policy. According to the review paper by E.W.T.Ngai and P.T.Y.Lee in 2016, they developed a conceptual framework which has four steps in policy making. It includes the policy making cycle similar to PDCA, which consists of 1. Agenda Setting 2. Policy Formulation and Decision Making 3. Policy Implementation and 4. Policy Evaluation. The text mining technique will enable public sectors to enhance the quality of the cycle and reduce the cost by saving time for the process.

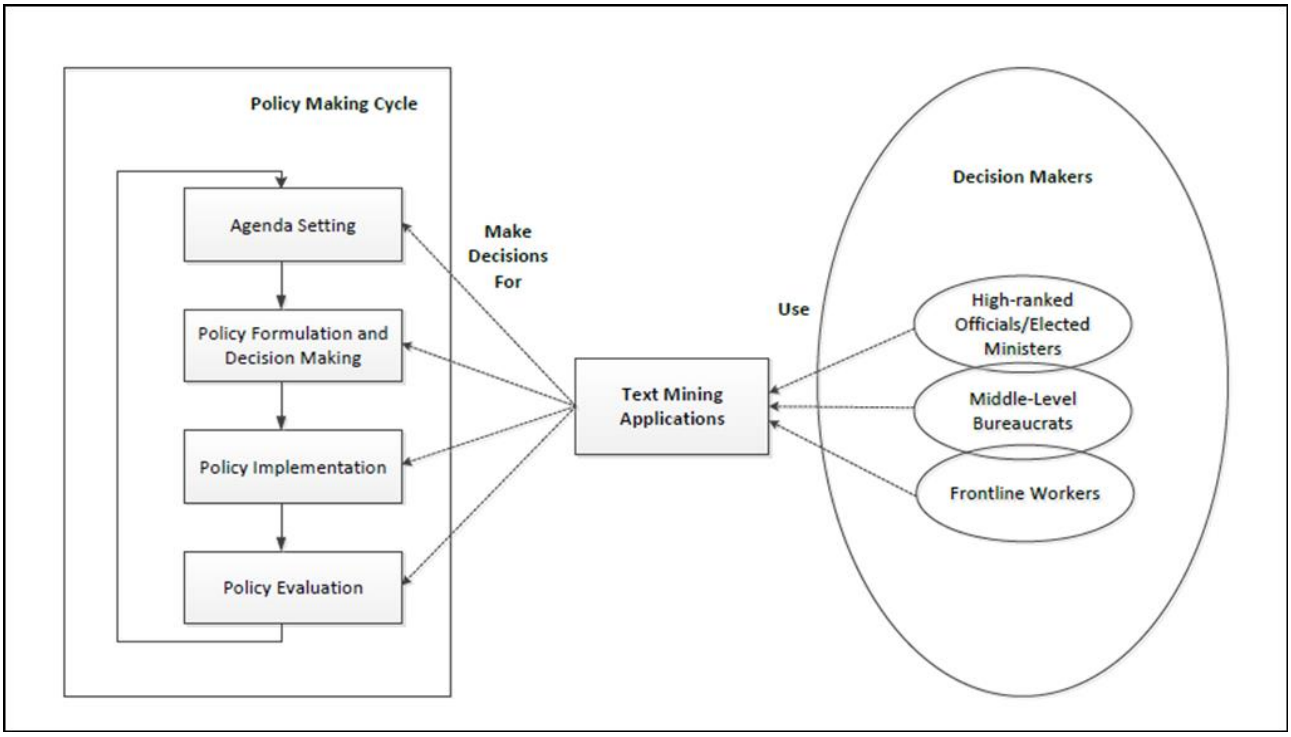


Figure2: Conceptual framework of text mining in public policy

3-1. Agenda Setting

At the first step, policy makers are interested in gauging the attention and sentiment of stakeholders including the public. I can know the trend or the forecast from news articles or SNS. The data could be easily scraped by the related words (for example, "TANF"), using Twitter API and NY Times API. Fortunately, some sentiment analysis tools are available for Python tough some of them do not seem free of charge. We could speed up analysis by leveraging them.

3-2. Policy Formulation and Decision Making

The government wants to collect and summarize the opinions or understanding of the policy by citizens or stakeholders in this process. For example, we can visualize and summarize the information from testimony documents or questionnaires. As for TANF, various types of documents are available, but it is dispersed and hard to collect in a smart way. We can use bills and corresponding testimony documents at state legislatures, state plans on TANF by state governments, bills and laws on TANF by Congress, research paper by academic professionals and so on.

3-3. Policy Implementation

The automatic screening and reply or the error detection for application documents is one of prospective public applications of text mining. Many frontline workers are involved in this kind of process, so we can cut the project cost if the process is automated. Unfortunately, I cannot have access to the real application document database of TANF because they include much personal information. However, we can create a simple mock-up which detects handwriting errors in the documents, and we can also check the function by a

sample filled by ourselves, using an official format. We can take advantage of one OCR application such as Google Cloud Vision API.

3-4. Policy Evaluation

It is valuable to pay attention to feedback comments of participants or frontline workers in existing programs. They might give some hints to decide performance measures of TANF which are mentioned in the past work by H. Hahn and P. Loprest.

As mentioned above, I can utilize text mining to make TANF more efficient. However, some processes of the package are more challenging than I expected, and I didn't have much time. Therefore, I decided to focus on the second step. My goal is to provide some important findings on what I can and cannot do with text mining skills covered in the class rather than take efforts to write well-developed codes or create just beautifully visualized graphs. I believe that this try would give us practical insights on text mining or NLP.

4. My work and results

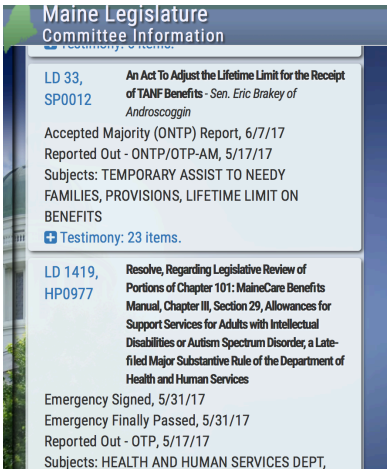
I chose the corpus of bill and testimony documents on TANF provided by state legislatures since these include opinions with clear stances. Research paper is too analytics and qualitative to apply text mining skills, and official documents such as state plans or laws provided by state governments or Congress describe rules or plans, so they are not suitable for analysis of comments considered at policy making process. Therefore, here I analyze the bill and testimony documents in state legislatures with text mining techniques.

4-1. Scraping dossiers

I tried to scrape documents from websites of more than 20 states. But their structures are totally different, so handling all of them is very time-consuming. And many of them have very few documents on TANF. Therefore, I focus on Maine legislature because the state has much more documents on TANF than other states. Furthermore, I added documents of Kansas as a training dataset to predict positions of the testimony documents of Maine.

Scraping code is almost the same as the first homework. It is easy to scrape documents of Maine because API is available on the website. I scraped only documents on TANF from Maine's last five legislatures (125th-129th). These bill documents have the word of "TANF" or "needy families" in their titles, and the testimony files are stored under the bill section. As a result, I got 18 bills documents and 177 testimony documents. Kansas has much less documents on TANF (I found only one documents for five years). So I scraped all the documents with information on positions in the pages of "Health and Human Service at House" and "Public Health and Welfare committee at Senate" rather than filtered by TANF. These documents are later used to predict the positions (proponent or opponent) of Maine testimony documents because Maine testimony documents do not have the information on positions. As a result, I downloaded 438 testimony documents from Kansas. Please note that most of them are irrelevant to TANF.

The acquired documents are PDF files. I converted them into text files with "pdftotext" command. Here, I was able to convert all of Maine documents, but some of Kansas documents were failures. This is because PDF files from Kansas do not have text information (they are like images). Finally, I got 279 text files of Kansas.



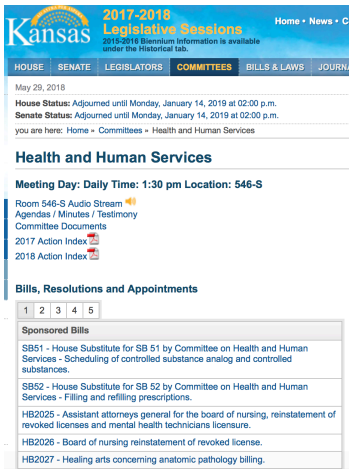


Figure3. Maine and Kansas legislature websites

4-2. Overview of the corpus

The first step of analysis is to visualize various types of information of the corpus to grasp the structure. I checked the length of each document (how many words does each document have) and the keywords on TANF discussed in the Maine legislature.

The following graphs show the lengths of testimony and bill documents. The average length of each type is 1082 and 1044 words.

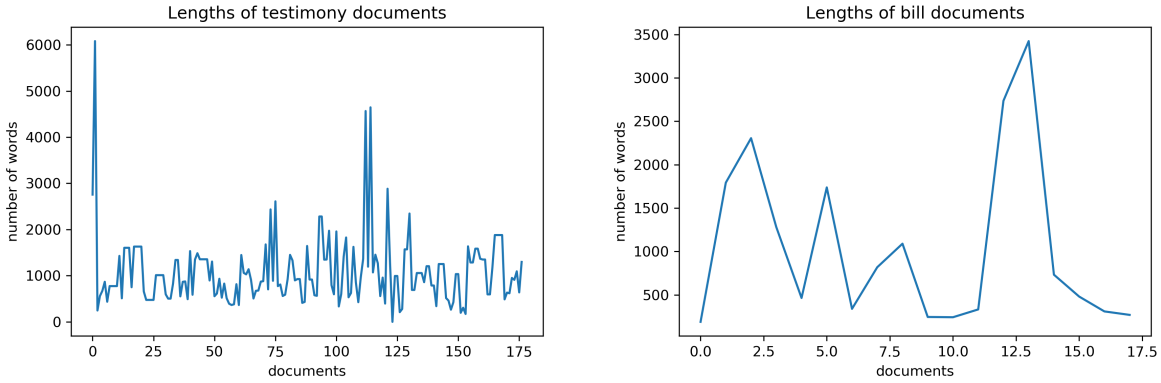


Figure4: how many words does each document have (n=177, 18)

Next, I depict a word cloud in which words are mapped with different colors and sizes for the frequency. We can see frequent words and bigrams in all the testimony documents of Maine. This picture shows us a good view of frequent words.

Figure5: Word cloud for testimony documents on TANF of Maine

The bill documents on TANF have summary paragraphs, so I extracted them. Looking at them, I can see that many of bills are related to requirements or eligibility of the assistance program. Therefore, I can assume that “requirement” “eligibility”, “limit”, “prohibit” , “drug” , “substance” and “violence” are expressing the important agendas in the legislatures. In the dispersion plot, I can assure that these words cover almost the entire corpus.

Be it enacted by the People of the State of Maine as follows:

Sec. 1. 22 MRSA §3762, sub-§21 is enacted to read:

21. Work search requirement. The department may not grant TANF assistance to an applicant who is job-ready as determined by the department unless the applicant has applied in writing for 3 separate advertised jobs and submitted verifiable documentation of the applications to the department. To satisfy this work search requirement, the applicant may submit the job applications at any time from one week prior to the date of the application to the department for TANF assistance to 2 weeks following that date.

SUMMARY

This bill creates a work search requirement for job-ready applicants to the Temporary Assistance for Needy Families program.

Figure6: example of bill documents' summary

```
dic
```

```
{'SP0181': 'This bill creates a work search requirement for job - ready applicants to the Temporary Assistance for Needy Families pro
gram .',
 'SP0136': 'This bill provides that a person who has exhausted the 60 - month lifetime limit on Temporary Assistance for Needy Famili
es program benefits is ineligible to receive municipal general assistance program benefits .',
 'SP0066': 'This bill provides that a person who has exhausted the 60 - month lifetime limit on Temporary Assistance for Needy Famili
es program benefits is ineligible to receive municipal general assistance program benefits except that a person who has been ineligib
le to receive benefits under the Temporary Assistance for Needy Families program for 5 or more years may be considered eligible and a
person who is in the process of seeking an extension of benefits under the Temporary Assistance for Needy Families program may be con
sidered eligible .',
 'SP0012': 'This bill changes the lifetime limit for receiving benefits through the Temporary Assistance for Needy Families program f
rom 60 months to 36 months .',
 'HP1324': 'This bill makes the following changes to the laws governing the Temporary Assistance for Needy Families program . 20 21 2
2 23 24 25 1 . It removes the provision that prohibits a person from being sanctioned under the Additional Support for People in Retr
aining and Employment - Temporary Assistance for Needy Families program or the Temporary Assistance for Needy Families program for fa
ilure to participate in the Additional Support for People in Retraining and EmploymentTemporary Assistance for Needy Families program
if that failure to participate is based on good cause . 26 27 28 2 . It removes the 24 - month limit on education , training and trea
tment for participants in the Additional Support for People in Retraining and Employment - Temporary Assistance for Needy Families pr
ogram . 29 3 . It eliminates the Parents as Scholars Program .',
 'HP1317': "This bill requires the Department of Health and Human Services to report annually by February 15th to the joint standing
committee of the Legislature having jurisdiction over health and human services matters and the joint standing committee of the Legis
lature having jurisdiction over financial affairs regarding actions taken by the department to investigate program integrity under th
e MaineCare , Temporary Assistance for Needy Families and food supplement programs , including the amount recovered , the cost of tho
se investigations and prosecutions , the number of personnel working on the investigations , the status of cases referred to the Atto
rney General ' s office , a description of the performance and activities of a vendor , contractor or other program integrity unit us
ed by the department to help recover overpayments , a description of the department ' s participation in federally mandated program i
ntegrity efforts , the results of federal audits , a description of defects , deficiencies or weaknesses in department systems , a de
scription of planned investments in technology and a description of policy changes or improvements implemented .",
 'HP1312': 'This bill prohibits benefits provided under the Temporary Assistance for Needy Families program from being expended on to
bacco , imitation liquor , liquor , gambling , lotteries or bail .',
```

Figure7: summary of each bill document

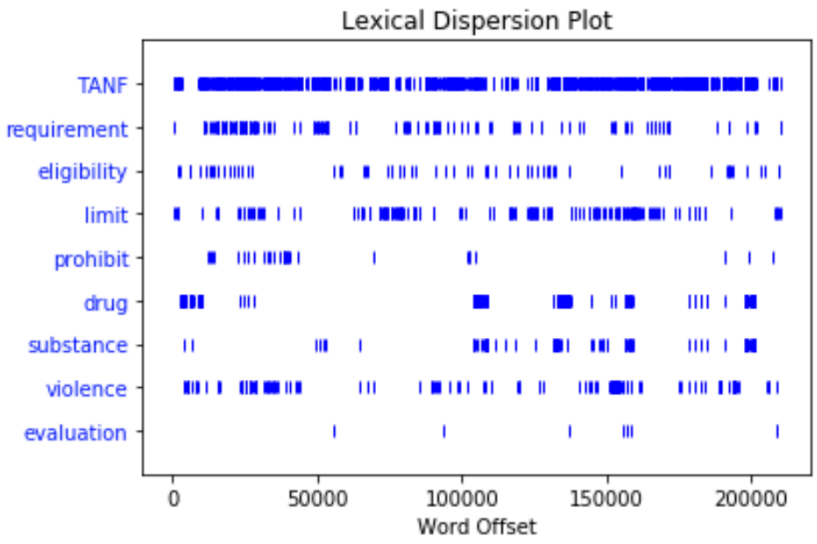


Figure8: Dispersion plot of testimony documents of Maine for possible keywords

4-3. Prediction of the positions

As mentioned above, Maine testimony documents have no labels of whether they are proponents or opponents. I would like to predict the positions of claims with machine learning techniques. Here I use the method of Naive Bayes Classification. I trained and tested the binomial model with the data of 279 testimony documents from Kansas. Among 279 documents, 219 files are labeled as proponents, and I used 25 of them as out-of-samples for testing. Here, the response variable is proponent or opponent. The features are 200 most frequent words in the entire corpus of Kansas. The accuracy of the trained model is 0.88. Looking at the informative features, I can find the word of "support". If a document has a word of "support", it is more likely to have the label of "proponent", and vice versa. However, other words are related to medical ore healthcare policy. These topics are mainly discussed in Kansas legislature, so it is slightly different form public assistance such as TANF.

```
all_words = nltk.FreqDist(w.lower() for w in documents_k.words()).most_common()
```

```
word_features = [w for (w, n) in all_words if w not in stopwords if len(w)>2][:200]
```

```
def document_features(document):
    document_words = set(document)
    features = {}
    for word in word_features:
        features['contains({})'.format(word)] = (word in document_words)
    return features
```

```
featuresets = [(document_features(d), c) for (d,c) in kansas_subset]
train_set, test_set = featuresets[25:], featuresets[:25]
classifier = nltk.NaiveBayesClassifier.train(train_set)
```

```
print(nltk.classify.accuracy(classifier, test_set))
```

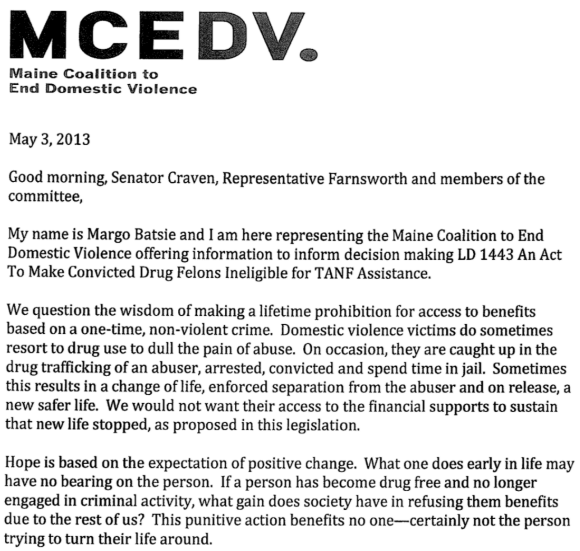
```
0.88
```

```
classifier.show_most_informative_features(7)
```

```
Most Informative Features
    contains(palliative) = True           False : True   =      5.6 : 1.0
    contains(anesthesia) = True           False : True   =      4.6 : 1.0
       contains(support) = False          False : True   =      3.9 : 1.0
       contains(kancare) = True           False : True   =      3.3 : 1.0
       contains(dentist) = True            True : False  =      3.0 : 1.0
     contains(therapist) = True            True : False  =      3.0 : 1.0
       contains(support) = True            True : False  =      2.9 : 1.0
```

Figure9: prediction model of the positions (proponent or opponent) of testimony documents

I applied the prediction model to the corpus of Maine. I got the predictions for all the 177 documents, but I cannot check whether they are correct or not. So I show just two examples. Both are predicted as proponent although the first one is actually opponent and the second is proponent. This might be because the second document has the word of "support" in the first paragraph.



MCEDV.

Maine Coalition to
End Domestic Violence

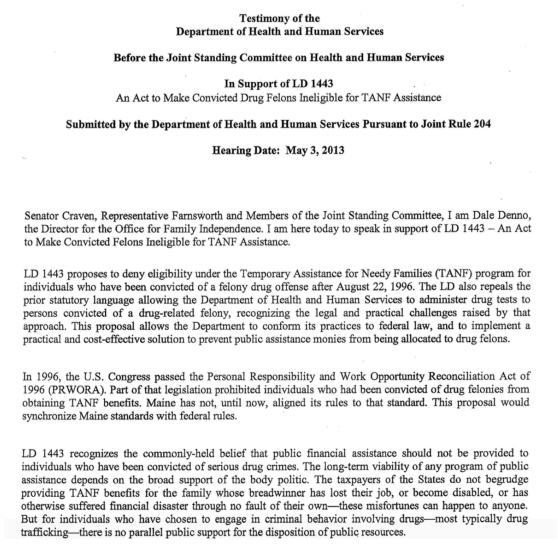
May 3, 2013

Good morning, Senator Craven, Representative Farnsworth and members of the committee,

My name is Margo Batsie and I am here representing the Maine Coalition to End Domestic Violence offering information to inform decision making LD 1443 An Act To Make Convicted Drug Felons Ineligible for TANF Assistance.

We question the wisdom of making a lifetime prohibition for access to benefits based on a one-time, non-violent crime. Domestic violence victims do sometimes resort to drug use to dull the pain of abuse. On occasion, they are caught up in the drug trafficking of an abuser, arrested, convicted and spend time in jail. Sometimes this results in a change of life, enforced separation from the abuser and on release, a new safer life. We would not want their access to the financial supports to sustain that new life stopped, as proposed in this legislation.

Hope is based on the expectation of positive change. What one does early in life may have no bearing on the person. If a person has become drug free and no longer engaged in criminal activity, what gain does society have in refusing them benefits due to the rest of us? This punitive action benefits no one—certainly not the person trying to turn their life around.



Testimony of the
Department of Health and Human Services

Before the Joint Standing Committee on Health and Human Services

In Support of LD 1443
An Act to Make Convicted Drug Felons Ineligible for TANF Assistance

Submitted by the Department of Health and Human Services Pursuant to Joint Rule 204

Hearing Date: May 3, 2013

Senator Craven, Representative Farnsworth and Members of the Joint Standing Committee, I am Dale Denno, the Director for the Office for Family Independence. I am here today to speak in support of LD 1443 – An Act to Make Convicted Felons Ineligible for TANF Assistance.

LD 1443 proposes to deny eligibility under the Temporary Assistance for Needy Families (TANF) program for individuals who have been convicted of a felony drug offense after August 22, 1996. The LD also repeals the prior statutory language allowing the Department of Health and Human Services to administer drug tests to persons convicted of a drug-related felony, recognizing the legal and practical challenges raised by that approach. This proposal allows the Department to conform its practices to federal law, and to implement a practical and cost-effective solution to prevent public assistance monies from being allocated to drug felons.

In 1996, the U.S. Congress passed the Personal Responsibility and Work Opportunity Reconciliation Act of 1996 (PRWORA). Part of that legislation prohibited individuals who had been convicted of drug felonies from obtaining TANF benefits. Maine has not, until now, aligned its rules to that standard. This proposal would synchronize Maine standards with federal rules.

LD 1443 recognizes the commonly-held belief that public financial assistance should not be provided to individuals who have been convicted of serious drug crimes. The long-term viability of any program of public assistance depends on the broad support of the body politic. The taxpayers of the States do not begrudge providing TANF benefits for the family whose breadwinner has lost their job, or become disabled, or has otherwise suffered financial disaster through no fault of their own—these misfortunes can happen to anyone. But for individuals who have chosen to engage in criminal behavior involving drugs—most typically drug trafficking—there is no parallel public support for the disposition of public resources.

Figure10: two examples of testimony documents on TANF (10881 and 10882)

4.4. Summarization of each document

Summarization is a very important skill of text mining, which is applied in a variety of fields. We studied a simple summarization in an assignment, but there are many summarization algorithms with more accuracy. Generally speaking, the summarization algorithms are divided into two types according to a brief summary by Mehdi Allahyari in 2017. They are Extractive and Abstractive summarization. The first one is to choose an important subset of the sentences in the original text. The second one is to interpret and examine the text using advanced natural language techniques. The abstractive approach is advanced one, so I use the extractive approach. The extractive summarization makes use of intermediate representation which is categorized into two methods, topic and indicator representation. The topic representation gets sentence scores with topic words, TFIDF, latent Dirichlet analysis while the indicator representation computes sentence scores with characteristics such as sentence length, having certain words or position in the document. I took advantage of the two representations and tried three hybrid summarization methods.

The first one uses sentences tokenized by documents.sents and computes the scores by sum of frequency of words of the sentences and bill documents. The second one uses sentences tokenized by another tokenizer because I found that the first tokenizer creates some failures of sentence tokenization. At scoring, the first and second one add the indicator of agenda words used in averaged bill documents' summary.
The third one computes the score by sum of tfidf. The last one's execution is very slow, so I applied it for just first three documents. The result is as below. Each algorithm gives two sentences with highest scores. Looking at the first and second outputs for one document(10881 shown in Figure10), one is the combination of the title and the first sentence, and the other is too specific to be a part of summary. At the third model, I used TFIDF instead of just word frequency. In the result, the second sentence is the most important part in the document which is the first sentence of the second paragraph. It seems that I can improve the summarization method, but there is huge room to make it more accurate.

1st model: Sentence tokenized by (corpus).sents(fileid)
Score for each sentence = sum(word freq in the document) + sum(word freq in the bill documents)/18

```
Summary of 10881.pdf.txt in testimony on TANF:
May 3 , 2013 Good morning , Senator Craven , Representative Farnsworth and members of
the committee , (sentence number: 2 , score: 60.666666666666664 )
On occasion , they are caught up in the drug trafficking of an abuser , arrested , con
victed and spend time in jail . (sentence number: 13 , score: 67.11111111111111 )
```

2nd model: Sentence tokenized by nltk.sent_tokenize(raw text)
Score for each sentence = same as above

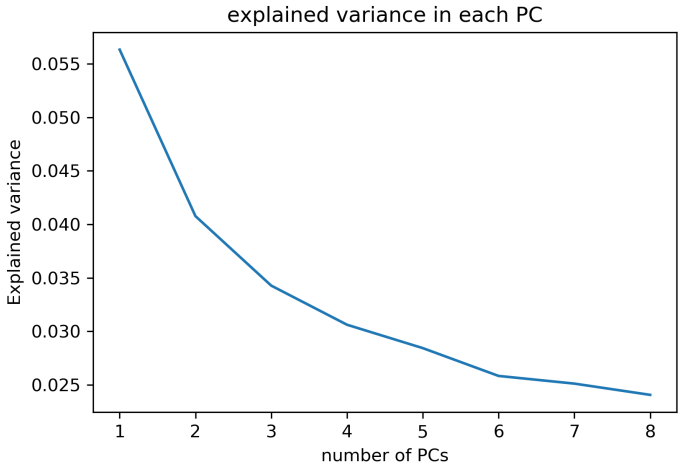
```
Summary of 10881.pdf.txt in testimony on TANF:
MCE  Maine Coalition to End Domestic Violence  May 3, 2013 Good morning, Senator Crave
n, Representative Farnsworth and members of the committee,  My name is Margo  Batsie a
nd  I  am here representing the Maine Coalition to End  Domestic Violence offering inf
ormation to inform decision making LD 1443 An Act  To Make Convicted Drug Felons  Inel
igible for  TANF Assistance. (sentence number: 0 , score: 138.5 )
On occasion, they are caught up in the drug trafficking of an abuser, arrested, convic
ted and spend time in jail. (sentence number: 3 , score: 67.11111111111111 )
```

3rd model: Sentence tokenized by the same way as above
Score for each sentence = sum(word tfidf)

```
Summary of 10881.pdf.txt in testimony on TANF:
MCE  Maine Coalition to End Domestic Violence  May 3, 2013 Good morning, Senator Crave
n, Representative Farnsworth and members of the committee,  My name is Margo  Batsie a
nd  I  am here representing the Maine Coalition to End  Domestic Violence offering inf
ormation to inform decision making LD 1443 An Act  To Make Convicted Drug Felons  Inel
igible for  TANF Assistance. (sentence number: 0 , score: 9.52826760439 )
We question the wisdom of making a lifetime prohibition for access to benefits based o
n a one-time, non-violent crime. (sentence number: 1 , score: 9.52826760439 )
```

4.5 Principal Component Analysis

Finally, I tried to apply topic model for the testimony documents of Maine. I select PCA as a dimension reduction method, using three clusters. And I looked into cosine similarity between documents and topic vectors for the components. The explained variance gradually decreases, so it is difficult to reach to the best choice of the number of topics. Here, I made three clusters. The cosine similarity indicates that the document of 10880 is similar to the document of 10881. In fact, these documents are claims on the same bill (HP1037) and the both are opponents to the bill. This result is reasonable. However, I am perplexed at the words with highest values in topic vectors. It is hard to interpret the clustering intuitively.



	10762	10880	10881	10882	10883	16070
10762	1.000	0.531	0.387	0.727	0.472	-0.278
10880	0.531	1.000	0.863	0.117	-0.496	-0.677
10881	0.387	0.863	1.000	0.340	-0.501	-0.227
10882	0.727	0.117	0.340	1.000	0.622	0.440
10883	0.472	-0.496	-0.501	0.622	1.000	0.418
16070	-0.278	-0.677	-0.227	0.440	0.418	1.000

```
drugs                0.295900
offering             0.270913
entire               0.129186
brunswick            0.104185
square               0.104185
month                0.103954
paving               0.099111
track                0.095154
disability-related   0.087257
initial              0.079361
homes                0.066933
struggled            0.065071
h1epowerofsocicr     0.064636
focus                0.063636
aspects              0.063472
Name: topic0, dtype: float64
```

```
offering             0.354143
entire               0.282921
track                0.197417
disability-related   0.192764
_xjg                 0.165400
result               0.145653
h1epowerofsocicr     0.143785
struggled            0.133537
aspects              0.127249
oftheir              0.126662
initial              0.116523
burdened             0.115803
experiments          0.094424
fundamentally        0.092366
unavailable          0.087157
Name: topic1, dtype: float64
```

```
lgf                  0.231542
checks               0.207125
executives           0.160663
_xjg                 0.152851
longperiods          0.138081
0ry                  0.128219
due                  0.098624
jcomart@mejp.org     0.085213
hisjob               0.085092
circuit              0.081275
2001                 0.079514
increased            0.079290
conceit              0.077351
ibuprofen            0.074218
estimated            0.073301
Name: topic2, dtype: float64
```

Figure11: PCA results (explained variance, cosine similarity, top 15 words in topic vectors)

5. Conclusion and limitations

- I depicted a comprehensive picture to apply text mining for public policy, and analyzed the corpora of legislature documents, using a lot of techniques which we learnt in this class.
- Some of them are useful, but others are not because there are some limitations.
- I had very few documents (177 testimony documents on TANF of just one legislature). This is because the file sources are dispersed to 50 state legislature websites whose structures are totally different. Scraping all the documents on one concerned topic (TANF) on all the platforms is very time-consuming. In addition, some states make bills on TANF frequently, and others not. If I had more documents on the policy, I could easily apply the methods that I established in this project to expanded corpora.
- The testimony documents are dirtier than the official documents such as laws or resolutions since they are submitted by a variety of people. The formats of the documents are different from one another. The unnecessary information such as title, header and footer prevents us from getting a clean text. The pdf files are not sometimes readable because they are just images, not text. And I experienced lost or added information at conversion to text.
- Nevertheless, I believe that the text mining skills which we learnt in this class are essential to improving public policy making process. There is also huge potential of further research. For example, there might be much room to improve summarization algorithm, and we could find change of issues or policy needs with more clean time-series corpora.