Information Extraction from Public Health Statutes

Jaromir Savelka
(advisor: Kevin Ashley)

Intelligent Systems Program
University of Pittsburgh

jas438@pitt.edu

April 11, 2014

Overview

Overview

Useful Questions

- What obligations does a particular PHS agent have in a situation of emergency?
- With which other PHS agents must he/she coordinates his/her efforts?
- Do obligations of PHS agents differ across individual states?
- What about their interactions? Are they defined consistently across the US?
- Which agents are the key components in emergency preparedness and response?

Data Acquisition

Coding Scheme

Specialized coding scheme (codebook) was developed. This scheme specifies (i) which types of information should be extracted from (ii) which kind of statutory texts and (iii) how should they be encoded.

Manual Coding

The annotators are experts that follow the instructions of the codebook and manually code (excel sheet) relevant statutory provisions (word documents).

Coding Components

Coding Scheme

- **Relevance**
- Citation
- Public Health System Agent (acting)
- Prescription
- Action
- Goal
- Purpose
- Type of Emergency Disaster
- Public Health System Agent (receiving)
- Timeframe
- Condition

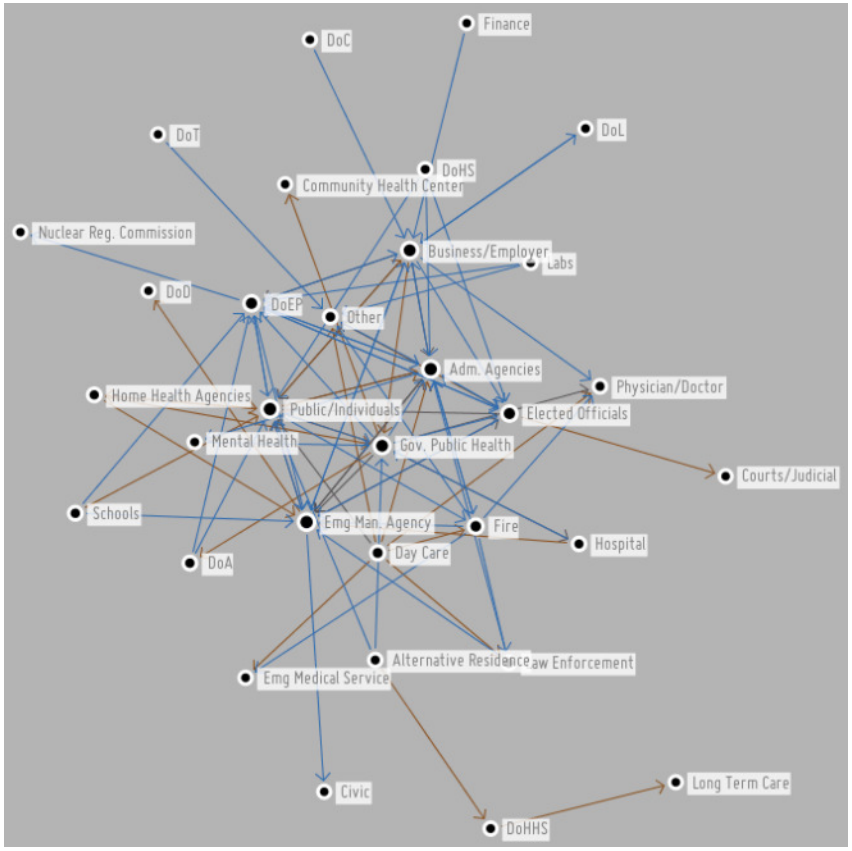
Coding Components II

Statutory provisions

The number of patients admitted to any area of the hospital shall not exceed the number for which the area is designed, equipped, and staffed except in cases of emergency, and then only in accordance with the emergency or disaster plan of the hospital. (28 Pa. Code para 101.172)

Coded statutory provisions

28 Pa. Code para 101.172; Hospital (14); Must Do (2); Suspend (29); Rule/Regulations/ Restrictions (4); For Emergency Response (2); Non-specified Disaster/Emergency (5); Public/Individuals (27); Silent (0); Silent (0)



(Sweeney, Ashley et al. 2013)

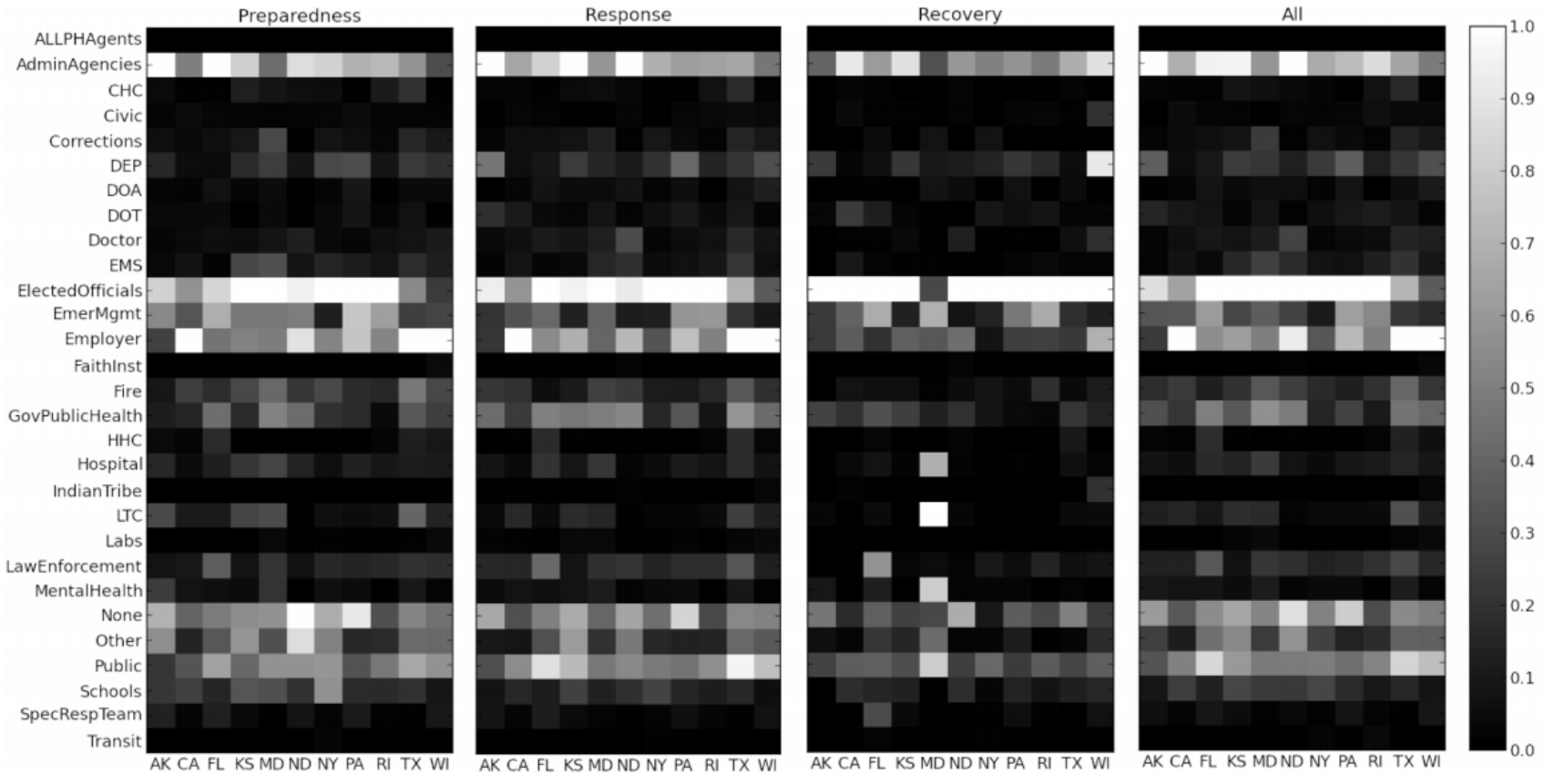
Comparison of legislatively created legal networks in Florida and Pennsylvania for the purpose of radiological/nuclear incident surveillance.

Circles: PHS actors and partners directed by law in both FL and PA.

Tan links: relationships present in both states.

Blue links: relationships present in PA but not in FL.

Heatmap of Agent Strength



(Sweeney, Ashley et al. 2013)

Overview

Definition of the Research Problem

Automatize the coding process in such a way that a computer presented with statutory texts would be able to distinguish between relevant and irrelevant provisions and assign the correct codes to the provisions recognized as relevant.

We use supervised classification to facilitate the task.

1. We have data (statutory texts and the codebook).
2. We have target labels (codes).

Overview

Features

For features generation we use the text chunks.

Term Frequency – Inverse Document Frequency

$$tfidf_{ij} = \frac{Count(w_i \ in \ d_j)}{Size(d_j)} \times \frac{N}{C(d_j \ with \ w_i)}$$

Indexing based on Latent Semantic Analysis

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^*$$

Latent Dirichlet Allocation

$$p(\mathbf{w}|\alpha,\beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N} p(W_n|\theta,\beta) \right) d\theta$$

Labels

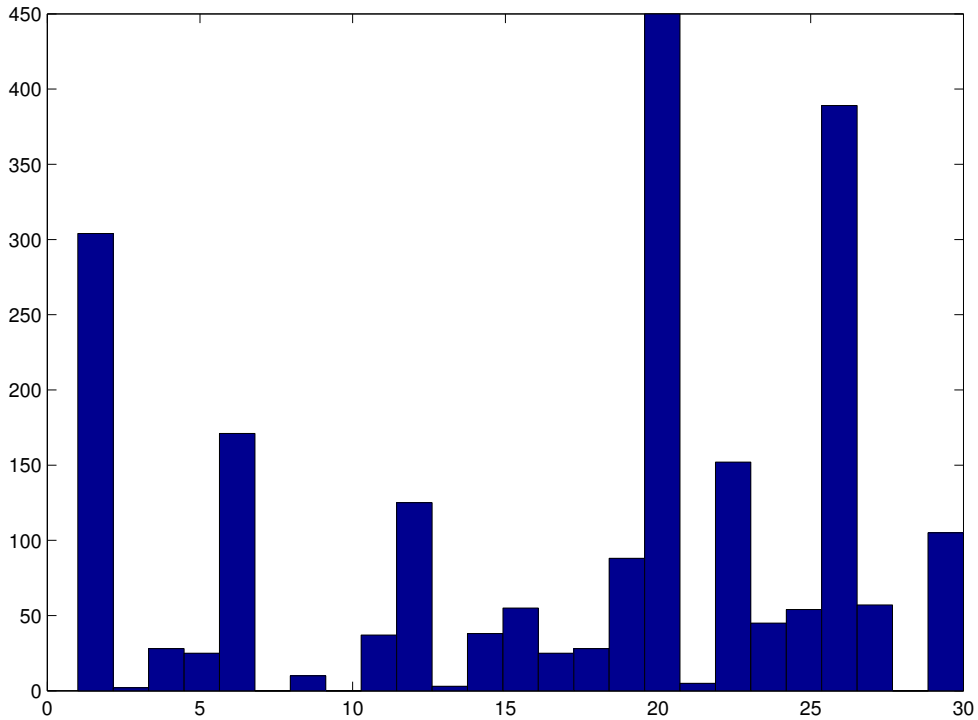


Figure: The histogram shows distribution of PHS agent (acting) class labels.

PHAa	PHAr	PHAa count	PHAr count	Co-occur	P
30	27	39	235	19	0.49
26	6	203	165	50	0.25
14	12	22	53	8	0.36
6	20	128	222	46	0.36
4	4	17	10	8	0.47
5	12	14	53	5	0.36

Table: The table shows possible dependencies between PHS agent (acting) and PHS agent (receiving).

Overview

Decision Trees

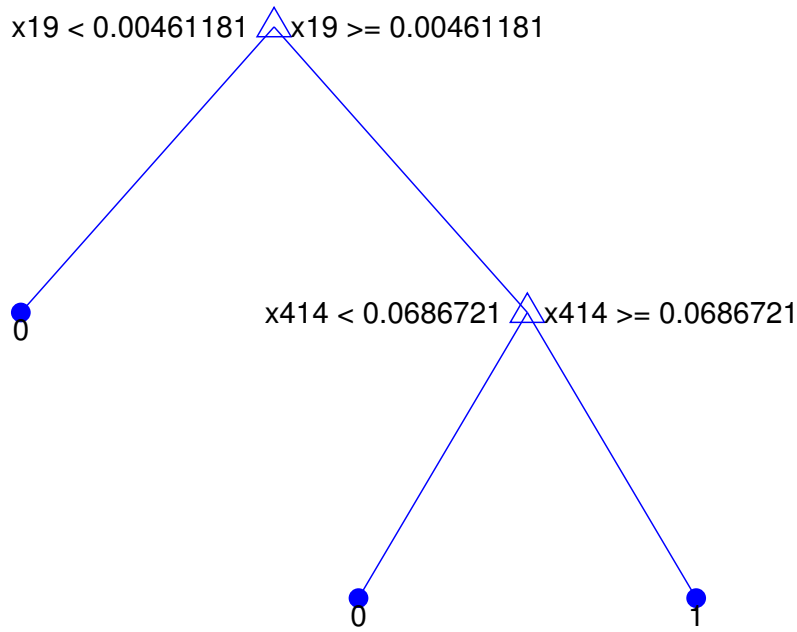


Figure: The decision tree classifier that was pruned in a way to retain only two levels.

mme = 0.14; Precision = 0.94; Recall = 0.29

Decision Trees

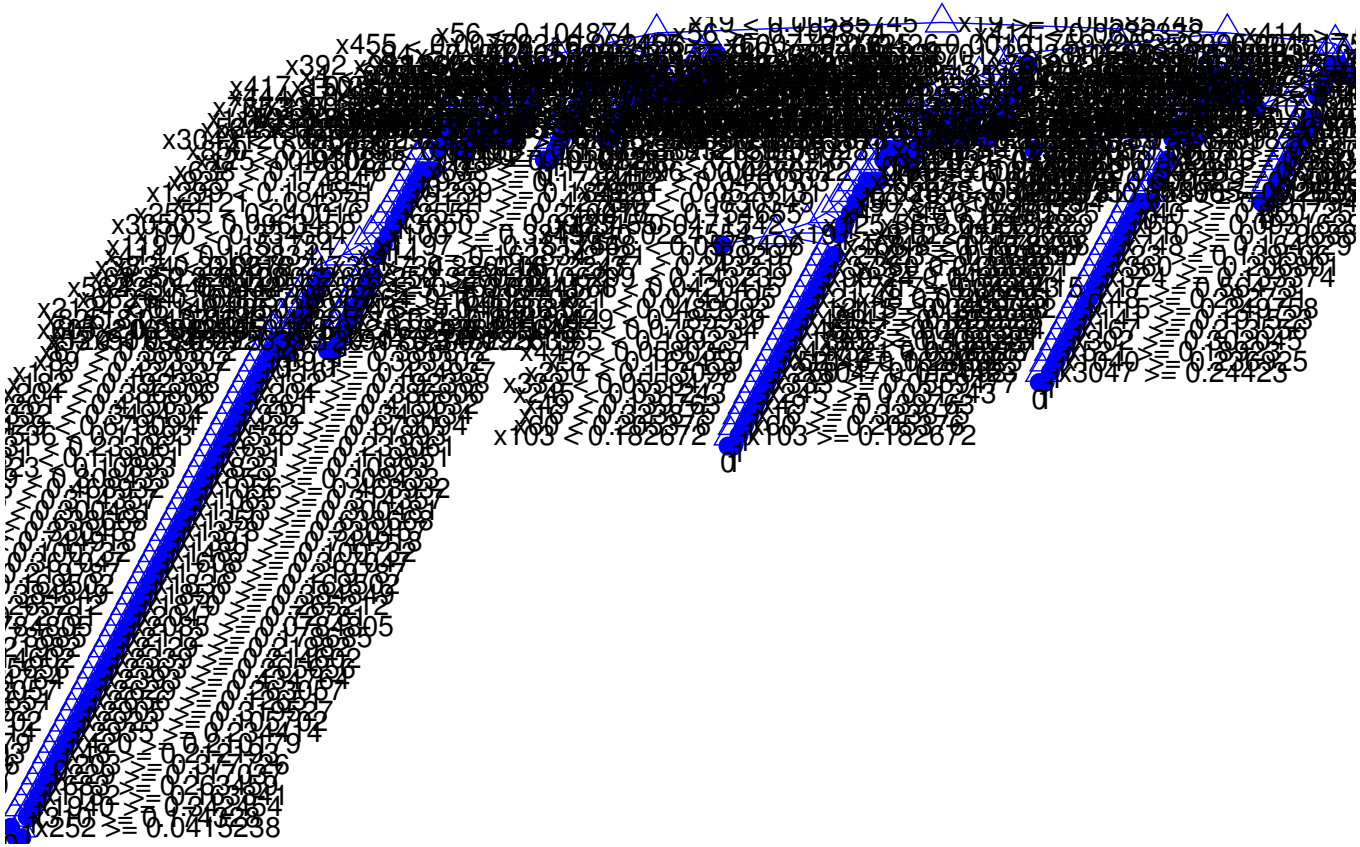


Figure: Part of the full decision tree classifier with 483 nodes.

mme = 0.14; Precision = 0.56; Recall = 0.53

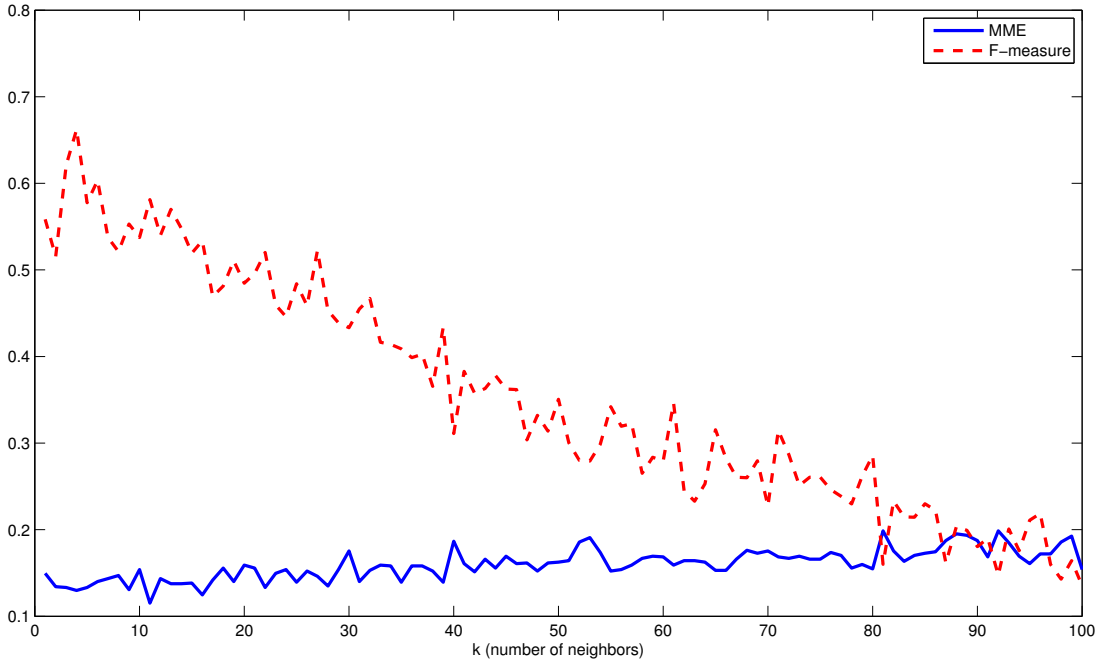


Figure: Progress of the F-measure and mean misclassification error of the kNN classifier as I increased the number of k (neighbors in kNN model).

Evaluation

$$Precision_{model} = \frac{P \cap Y}{|P|}$$

$$Recall_{model} = \frac{P \cap Y}{|Y|}$$

$$F_{model} = \frac{2 \times Precision_{model} \times Recall_{model}}{(Precision_{model} + Recall_{model})}$$

Precision

	tfidf.0	tfidf.15	tfidf.3	tfidf.4	LSA	LDA
MFC	0	0	0	0	0	0
MFC+	0.43	0.43	0.43	0.43	0.43	0.43
DT	0.56	0.53	0.58	0.60	0.66	0.59
SVM	0.42	0.61	**0.89**	0.61	~	~
kNN	0.67	0.70	0.73	0.78	0.69	0.62

Table: The table shows precision of different classifiers on different representations of data.

Recall

	tfidf.0	tfidf.15	tfidf.3	tfidf.4	LSA	LDA
MFC	0	0	0	0	0	0
MFC+	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**	**0.78**
DT	0.53	0.39	0.31	0.21	0.48	0.45
SVM	0.03	0.04	0.09	0.22	~	~
kNN	0.45	0.44	0.43	0.43	0.59	0.46

Table: The table shows recall of different classifiers on different representations of data.

F-Measure

	tfidf.0	tfidf.15	tfidf.3	tfidf.4	LSA	LDA
MFC	0	0	0	0	0	0
MFC+	0.55	0.55	0.55	0.55	0.55	0.55
DT	0.55	0.44	0.42	0.30	0.56	0.51
SVM	0.05	0.07	0.16	0.32	~	~
kNN	0.54	0.54	0.54	0.55	**0.62**	0.53

Table: The table shows F-measure of different classifiers on different representations of data.

Future Work and Conclusions

The results show that, even with very simple models, the latent factor based representations enable the classifiers to retain high precision while not sacrificing recall as much as with the previously used simple TF-IDF representations.

Coded statutory provisions

- test the developed sets of features with the same classifiers on other parts of the codes
- experiment with ensemble methods
- implement an overlay framework for multi-dimensional classification

Thank you for your Attention!