State by State: Automated Alignment and Analysis of State Statutes

by

Zachary Hynes

S.B., Computer Science and Physics, M.I.T., 2013, M.Eng., Computer Science, M.I.T., 2014

Submitted to the Department of Electrical Engineering

and Computer Science

in Partial Fulfillment of the Requirements for the Degree of

Master of Engineering in Electrical Engineering and Computer Science

at the Massachusetts Institute of Technology

February, 2014

Author: Signature redacted

Department of Electrical Engineering and Computer Science January 31, 2014

Certified by: Signature redacted

Prof. Regina Barzilay, Thesis Supervisor January 31, 2014

Accepted by: Signature redacted

Prof. Albert R. Meyer, Chairman, Masters of Engineering Thesis Committee

Abstract

In this work, we explore text alignment with the context of the legal domain and outline several new tasks designed to make comparison and analysis of inhomogenous state statute hierarchies easier. We explore the unique features of the statute hierarchy dataset, apply several baseline text alignment algorithms, and address the issue of clustering evaluation when documents may belong to multiple clusters. We also explore pairwise alignment strategies and assess these in comparison to clustering methods.

Acknowledgements

The Masters' of Engineering program has been a unique experience for this one-time physics and chemistry student. I couldn't be more grateful to any number of friends and mentors along the way, but would like to especially mention my academic colleagues and fellow members of RBG. Tahira Naseem mentored me through a headfirst crash course in parsing, providing me with essential guidance as I developed some amount of skill in selecting, applying, and implementing algorithms for NLP. Karthik Narasimhan and Yonatan Belinkov served as invaluable friends as I navigated the ups and downs of the research process. My thesis advisor, Regina Barzilay, deserves my utmost thanks for giving me the freedom to explore the projects about which I was truly passionate, for showing me how to deconstruct problems in new domains, and for providing helpful guidance in gaining intuition about these problems. Along the way, I have collaborated with Rushi Ganmukhi and Alex Ratner, who helped this project inch forward from the initial painstaking collection of data to the development of what would hopefully become exciting new algorithms that can, in some way, help tackle the crucially important problem of civic engagement.

I would not have taken the leap to computer science in my senior year at MIT had it not been for the earnest support of my family. My dad's humility and work ethic, my mom's unwavering belief in me through countless late nights, and my sisters' support in keeping me grounded have all helped get me to the finish line. In navigating the last year of my M. Eng., Alex Hsu has stood by me throughout... and also made a victorious October 2013 that much more fun.

Chapter 1

Introduction

In order to understand and make comparisons between state laws, researchers need the capacity to identify and categorize efficiently the relevant areas of state statutes. Currently, resources dedicated to comparison of state laws are scattered across the internet in the form of single-issue resources created by public policy organizations, or else accessible from resources like Westlaw behind expensive pay walls. To identify relevant materials, researchers must comb through inhomogenous statute hierarchies. In cases where state hierarchies may not include titles, the task is still more difficult; for these reasons, open government organizations have begun work to annotate statutes in some states with titles. For example, at Maryland Decoded[1], the default title for a statute without a title is the first few words of the statute; users of the site are able to give input on what the proper title should be.

[1] http://marylandcode.org/

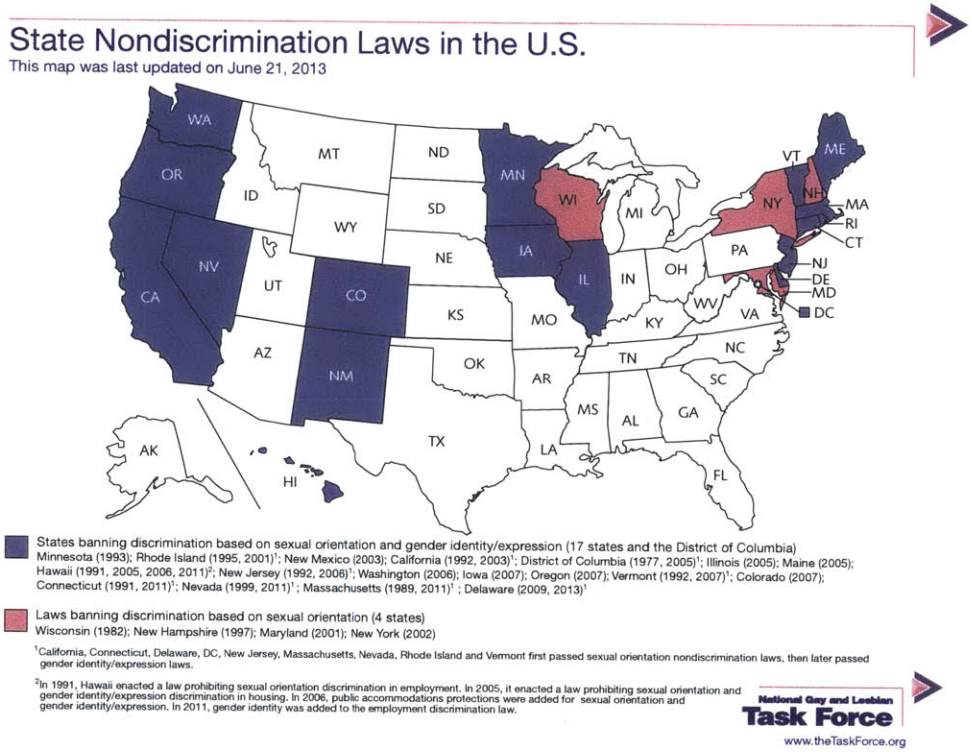
The long term goal of the project, toward which the work in this thesis serves as a first step, is to be able to automatically generate categorical comparisons of state laws. Given some question, like "What types of crimes are classified as hate crimes?"[2], and some finite set of answers, like the categories of bias that qualify a crime as a hate crime, a future iteration of our system will generate an automated summary of the differences across states, perhaps in the form a visual map like the one displayed in Figure 1.1. Such a task differs from free-form Q&A in that there is some finite set of candidate answers; moreover, we would like to ideally make joint decisions on the answers to those questions across multiple states, incorporating some understanding of the similarities and differences across states in both the philosophies their laws espouse and the structure through which those laws are expressed.

Our initial problem falls within the realm of organizing components of the law by domain and subdomain. Interest in this particular task, even outside of academic circles, came to the forefront with the debut of the Constitute project[3], developed by Google in collaboration with political scientists. This project allows users to search across different areas of world constitutions and compare world constitutions across various domains of the law. While our work is in a slightly different domain, many of the same challenges persist.

[2]An implied question in the chart at: http://www.adl.org/assets/pdf/combating-hate/state_hate_crime_laws.pdf

[3]https://www.constituteproject.org/#/

Figure 1.1: Non-discrimination state law map



Chapter 2

Data

2.1 Acquisition

In the acquisition phase, we scraped and parsed web data for both state statutes and categorical question and answer pairs. We retrieved state statutes by applying a simple "wget" scraper to `statutes.laws.com`. We obtained clean data from about 30 states before scraping restrictions prevented the acquisition of further data. While many states provide their statutes online for free access, few states provide them in any sort of uniform format, and some states limit bulk downloading as a whole.

For the longer term goals of the project, we obtained the kinds of categorical question and answer sets in regards to various characteristics of state statutes by scraping an array of public policy websites. We have obtained question and answer sets pertaining to a diverse set of subdomains of state

law, including divorce[1], hate crimes[2], and abuse[3].

2.2 Features

In each state, statutes are organized in terms of hierarchies. While state statutes are comprised of formal, legal language that may be unique to their domain, many free-text sources adhere to a similar, hierarchical structures. For example, textbook chapters and encyclopedia articles often consist of chapters, sections, and subsections, which each deal with subsequently more specific areas of text.

There are several distinguishing features of statutes organized in hierarchies as opposed to free-text documents; the unique features of these hierarchies present both challenges and additional information which our algorithms hope to exploit. Firstly, statute hierarchies are inhomogenous across states, as is shown in Table 2.1. Hierarchies vary in depth and breadth; correspondingly, leaf-level statutes can range in size from a sentence to several pages.

The semantic content of statutes will also vary, as states establish law within given domains in different ways. Moreover, some states may have entire title-level statutes devoted to areas not covered in other states. Alaska has a title devoted to mining; a state that does not possess the same natural

[1]http://www.americanbar.org/groups/family_law/resources/family_law_in_the_50_states.html
[2]http://archive.adl.org/learn/hate_crimes_laws/state_hate_crime_statutory_provisions_chart.pdf
[3]http://www.rainn.org/public-policy/laws-in-your-state

resources might address mining in one or two leaf-level statutes, if at all. In some cases, the regulations that govern a particular domain of law might be delegated to localities or else addressed in administrative codes.

Within the broader domains of law that all states cover, some states may choose to split up certain subjects that others choose to integrate. For example, child custody and parenting plans may be intertwined within the same statute, or they may be separated out into different statutes.

Leaf-level statutes will vary in size can range in size from a sentence to several pages. Unlike free-text documents, statute hierarchies exhibit several forms of unique structure. Not unlike web documents with hyperlinks, statutes may reference other statutes. While perhaps not as significant within the context of the initial task of text clustering, cross-statute references can cause problems within the context of the Q&A task. In our example pertaining to the types of bias tied to hate crimes, for example, some states might choose to address the definition of and the penalties for a hate crime all within the same statute. Others might choose to separate these, lumping the definition of a hate crime in with many other, unrelated definitions and the penalty for such a crime in with the penalties of many other crimes. The ability to parse these references and go beyond our current alignment scheme, to parse individual items and sentences within the statute, will most certainly be needed in order to answer fine-grained legal questions.

Table 2.1: Statute Statistics

Level	**Mean**	Median	**Min**	**Max**
Title	73	53	5	370
Leaf	30596	29825	18969	49019

Utah

Section 30-2-1 Grounds; jurisdiction for proceedings; divorce judgment awarded to both parties

(a) The circuit court has power to divorce persons from the bonds of matrimony, upon a complaint filed by one of the parties, entitled "In re the marriage of ____ and _____," for the causes following

(1) In favor of either party, when the other was, at the time of the marriage physically and incurably incapacitated from entering into the marriage state

(2) For adultery

(3) For voluntary abandonment from bed and board for one year next preceding the filing of the complaint

(4) Imprisonment in the penitentiary of this or any other state for two years, the sentence being for seven years or longer

(5) The comm... of the crime against nature, whether with mankind or beast, either before or after marriage

(6) For becoming addicted after marriage to habitual drunkenness or to habitual use of opium, morphine, cocaine or other like drug

...

(9) Upon application of either party, when the court finds there has been an irretrievable breakdown of the marriage and that further attempts at reconciliation are impractical or futile and not in the best interests of the parties or family. (10) In favor of the husband, when the wife was pregnant at the time of marriage, without his knowledge or agency

Massachusetts

Section 1: General Provisions

(1) A divorce from the bond of matrimony may be adjudged for adultery, impotency, utter desertion continued for one year next prior to the filing of the complaint, gross and confirmed habits of intoxication caused by voluntary and excessive use of intoxicating liquor, opium, or other drugs, cruel and abusive treatment, or, if a spouse being of sufficient ability, grossly or wantonly and cruelly refuses or neglects to pro-

vide suitable support and maintenance for the other spouse, or for an irretrievable breakdown of the marriage as provided in sections one A and B; provided, however, that a divorce shall be adjudged although both parties have cause, and no defense upon recrimination shall be entertained by the court.

Section 2: Confinement for Crime

(2) A divorce may also be adjudged if either party has been sentenced to confinement for life or for five years or more in a federal penal institution or in a penal or reformatory institution in this or any other state; and, after a divorce for such cause, no pardon granted to the party so sentenced shall restore such party to his or her conjugal rights.

Section 3 Absence; presumption of death

(3) A divorce may be adjudged for any of the causes allowed by sections one, one B, or two although the defendant has been continuously absent for such time and under such circumstances as would raise a presumption of death.

It is helpful to examine some of the contrasts in presentation of the grounds for divorce in Utah and Massachusetts in order to understand the optimal representation for this data within the setting of the alignment task. The in homogenous hierarchy issue is evidenced in this comparison, suggesting that attempts at one-to-one alignment could prove problematic. While some of the verbiage is the same, there are differences (e.g., 'intoxicating liquor' vs. 'habitual drunkenness' to address alcohol use) which suggest that some broader clustering of words should prove useful. The differences in legal syntax, with the Utah statute enumerating grounds in the form of a numbered list and the Massachusetts statute displaying them all in a single sentence, suggest that primitive paraphrasing approaches, which compare dependency trees, would be less helpful in this domain compared to free-text sentences.

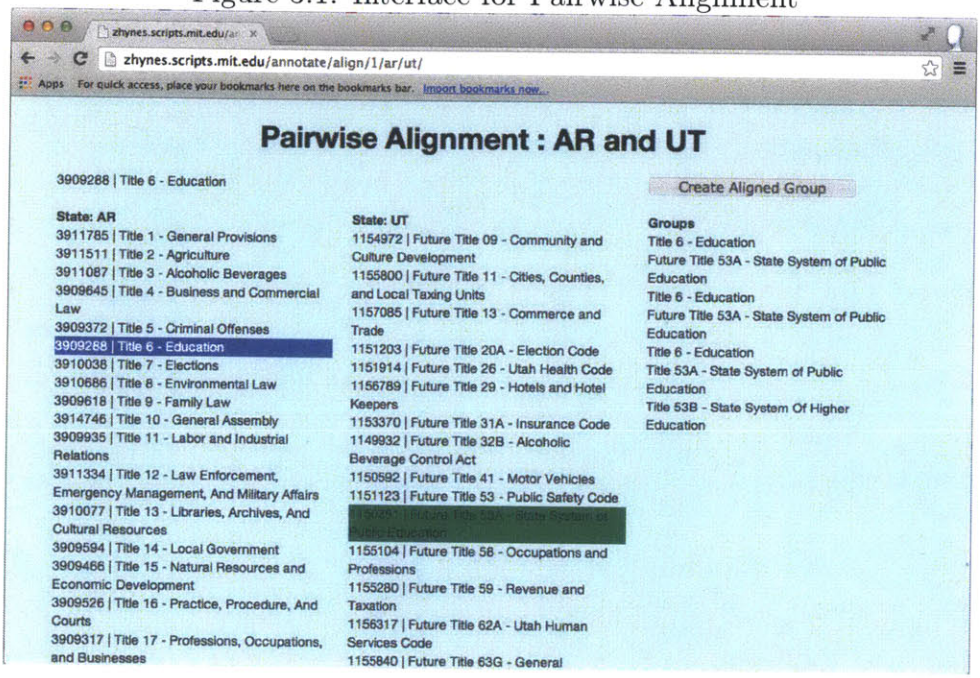
In considering these statutes, it is also worthwhile to note that of the three Massachusetts statutes tagged with "grounds for divorce" by our first annotator, not one of them includes the keyword "grounds" in the title, as is found in many other statute titles that address grounds for divorce. Indeed, the title "general provisions" would most likely be associated with a broader category, like the first annotator's "judicial proceedings" category, than anything else.

Chapter 3

Annotations

Annotators completed several tasks which provide the basis for evaluation. Initial attempts at pairwise alignment proved to be unwieldy, so we instead asked annotators to "tag" statutes in accordance with their semantic content. Annotators were allowed to tag as many statutes as needed.

Figure 3.1: Interface for Pairwise Alignment



Annotators tagged statutes at all levels of the hierarchy. At the top- and intermediate- levels, annotators were instructed to browse the statutes using the provided web interface in order to understand the domains covered under each statute. At the leaf statute level, instructors were asked to read or skim the statute to apply the appropriate tag. Annotators were allowed to supply as many tags as necessary, given the variance in tree structure described previously.

For the low-level task of divorce, annotators were given either a subtree or a set of subtrees within which the annotations should be made. Therefore, while not all states received tags addressing some specific issue, this does

not mean that the topic is not at all addressed in the entirety of the state statutes.

As in the evaluation, we have to choose from a number of imperfect measures when comparing the inter-annotator agreement on clustering.

Figure 3.2: Interface for Statute Exploration (showing Arkansas title-level statutes)

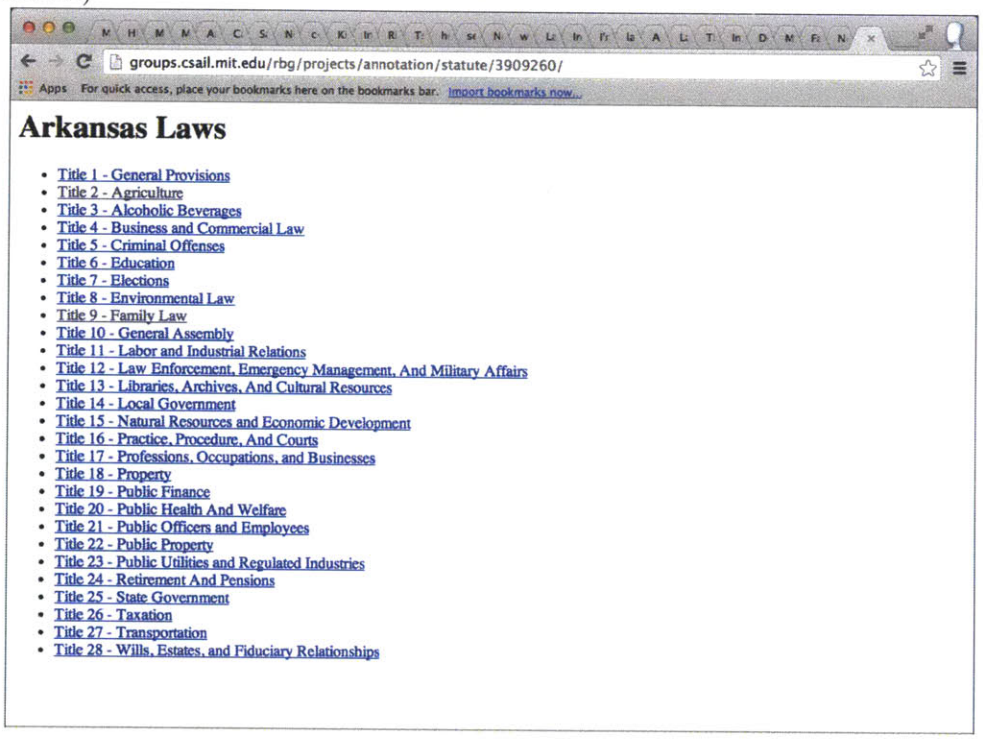


Table 3.1: Inter-annotator agreement

Annotation Set	Precision	Recall	F1
High-Level	0.386	0.392	0.389
Divorce	0.161	0.669	0.260

Our annotators broadly agreed on the high-level set, while exhibiting some greater difficulties in segmenting the divorce statutes. For the title-level statutes, Arkansas (featured in Figure 3.2) provided a good "template" for annotators in that it succinctly describes various domains of law which surface in Utah and Massachusetts. Utah's higher-level statutes, meanwhile, present some challenge due to the allocation of "Future Titles", typically much shorter in overall length and much narrower in domain than the standard titles. This is another example of a situation in which the title of a statute can be misleading. Massachusetts is unique in that the highest level is concerned with the broadest of divisions; most of the annotators' tags would be applied to one of these broad divisions.

As will be discussed later, the annotation task is hardly trivial in terms of time and labor expended. Annotators must often revise their internal notion of the proper segmentation of topics as they read through the statutes. Still, qualitative feedback from the annotators suggests that the task became easier once the segmentation of some domain of law into subdomains was established. While urged to read through the contents of statutes (or examine the lower-level statutes in the case of the high-level tags), titles can provide a useful guide. It is this intuition that we attempted to exploit in developing our models.

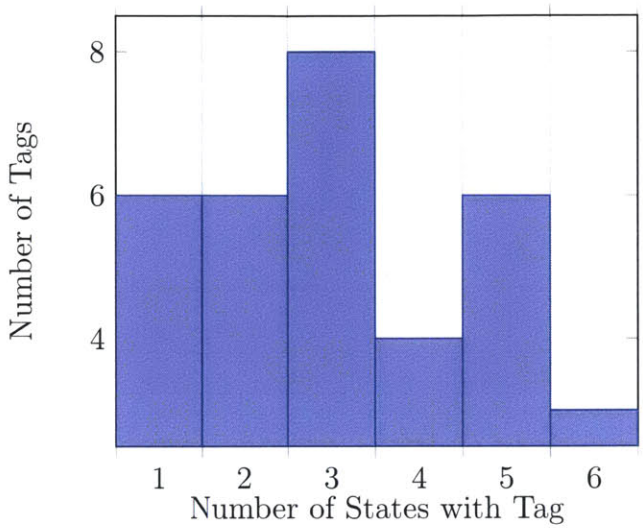


Figure 3.3: State By State Comparison of Divorce Statutes

In examining the annotations, it is clear that there are some sequential elements at the leaf-level statutes. As is suggested by Figure 3.3, permutation models, in which contiguous sets of statutes would each be assigned a unique label, are not appropriate for this dataset.

Chapter 4

Approaches

Text clustering is a classic task in natural language processing, but approaches are more commonly tested on datasets consisting of thousands of free-text documents, with some limited number of topics expected to span the corpus. For example, in [4], numerous state-of-the-art unsupervised clustering algorithms are evaluated on datasets of $\tilde{1}8{,}000$ documents spanning 20 categories and $\tilde{7}{,}000$ documents spanning 10 categories. By contrast, there are around $\tilde{2}000$ documents pertaining to divorce, marriage, and child custody across 30 states. These topics may inherit some properties of topics from other states, like penalties for illegal actions, descriptions of various court procedures, and functional sections like those for "definitions" and purposes, but there is also specific domain information within these statutes that would not be found in other statutes.

4.1 Clustering Alignments

Conventional approaches to document clustering assume some fixed representation of the data and proceed to the clustering stage. Perhaps the simplest possible representation of textual data is through unigrams. However, relying solely upon unigrams can be problematic due to issues of synonymy; two states might use different words to describe the same intrinsic phenomena. For example, some states refer to divorces caused due to "irreconcilable differences" while other states use "irretrievable breakdown" to describe the same possible grounds for divorce.

In order to address these limitations, there are several alternative representations of words to consider. Perhaps the most common representation of statute words are the topic or cluster assignments associated with words themselves. One classic representation is that of the topic model, Latent Dirichlet Allocation, which assigns each word in a document to a topic. As is shown in Table 4.4, topics formed around divorce can be seen as reasonably representing some of the subdomains of divorce law. Likewise, at the top levels of the hierarchy, the topics tend to organize around the given high-level subjects. We use the Gibbs sampling formulation described in [1] (with 2000 sampling iterations) in order to estimate the parameters of this model. Each statute is then represented as a vector of LDA topic counts, normalized to the length of the statute. The normalization is completed due to the fact that aligned statutes should not necessarily be closer in word count than any

other pair of statutes; given that a subject covered in one statute in one state could be covered by multiple statutes in another state, and the state clustering decisions are made independently, it is not reasonable to assume that two statutes on the same subject should be close in length.

Given some fixed representation of the data, whether word vectors [WV] or LDA topic counts, we develop the primary baselines by applying K-means clustering with initialization from KMeans++ [2].

Table 4.1: Performance on Title-Level Statutes

Algorithm	**Precision**	**Recall**	**F1**
WV + KMeans [Body]	0.056	0.269	0.092
WV + KMeans [Titles]	0.062	0.195	0.094
LDA+KMeans [Body]	0.136	0.271	0.181
LDA + KMeans [Titles]	0.039	0.428	0.072

Table 4.2: Top LDA Words by Topic [AR, UT, MA; title-level statutes]

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
chapter	person	school	court	licens	state	shall	person	employ	shall
subsect	shall	educ	shall	alcohol	fund	counti	corpor	shall	state
session	court	shall	person	pcomm...	refund	district	under	retir	water
general	section	board	section	person	shall	bond	state	employe	wast
section	trust	district	state	license	author	municip	secur	member	oper
under	estat	state	defend	depart	commiss	author	limit	benefit	use
amend	properti	student	law	beverag	revenu	board	shall	system	under
provid	interest	program	offic	shall	project	citi	chapter	servic	board
enact	appoint	year	offens	beer	use	properti	section	year	facil
requir	truste	public	counti	permit	issu	provid	good	under	mean
titl	order	fund	order	commiss	under	public	director	board	applic
part	such	institut	under	retail	land	under	right	contribut	depart
use	under	under	judg	liquor	amount	tax	file	fund	permit
servic	guardian	provid	justic	product	interest	land	member	state	section
public	proceed	higher	upon	state	money	within	interest	section	commiss
includ	time	teacher	attorney	sale	board	state	name	credit	develop
person	benefi...i	each	commit	applic	account	court	partner...	receiv	person
mean	repres	colleg	crimin	under	purpos	area	busi	paid	requir
divis	power	section	district	sell	construct	elect	provid	time	plan
act	sec	requir	one	issu	general	improv	mean	amount	author

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
tax	insur	shall	shall	vehicl	licens	such	shall	elect	state
shall	shall	child	properti	state	board	shall	such	shall	shall
state	comm...	health	person	shall	shall	section	person	vote	depart
counti	state	servic	owner	motor	person	chapter	use	ballot	offic
properti	under	care	notic	licens	state	under	public	counti	agenc
under	polici	parent	land	highway	applic	provid	state	offic	subchapt
section	section	depart	record	person	under	comm...	commiss	voter	fund
sale	provid	court	lien	under	chapter	depart	servic	candid	servic
year	person	person	file	oper	practic	town	regul	each	law
use	health	provid	interest	section	divis	one	dollar	person	public
commiss	compani	state	state	depart	requir	citi	requir	file	member
amount	contract	order	unit	driver	mean	comm...	upon	state	committe
incom	requir	under	time	plate	examin	offic	oper	report	general
credit	associ	facil	action	use	issu	year	one	name	provid
person	applic	divis	court	transport	certif	person	director	polit	author
taxpay	plan	medic	claim	issu	license	board	each	parti	board
provid	author	juvenil	right	registr	section	author	provis	clerk	inform
fuel	licens	support	parti	special	fee	provis	within	day	legisl
collect	benefit	program	real	requir	registr	hundr	provid	board	governor
subsect	coverag	mean	order	fee	rule	two	violat	general	employe

Table 4.3: Performance on Leaf-Level Divorce Statutes

Algorithm	Precision	Recall	F1
WV + KMeans [Body]	0.284	0.339	0.309
WV + KMeans [Titles]	0.268	0.486	0.346
LDA+KMeans [Body]	0.416	0.211	0.280
LDA + KMeans [Titles]	0.279	0.145	0.191
CTM + Titles	0.385	0.239	0.295

Table 4.4: Top LDA Words by Topic [27 states; statutes associated with marriage and divorce] , with most frequent words in selected corpus removed

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
domest	mediat	wife	ground	complaint	respond	financi	arbitr	individu	health
protect	program	former	code	motion	regist	insur	evid	term	new
violenc	particip	husband	year	temporari	registr	benefit	present	author	coverag
abus	famili	name	grant	final	contest	compens	concili	unit	insur
against	educ	statut	annul	counti	copi	institut	articl	defin	care
expens	concili	contract	mental	defend	petition	oblig	immun	establish	medic
famili	cours	marri	civil	district	sought	gross	between	entiti	plan
victim	judg	defend	wife	civil	document	account	wit	patern	expens
violat	attend	estat	defend	serv	valid	month	civil	temporari	avail
attorney	district	transfer	husband	clerk	confirm	worker	controversiniti		depend

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
best	friend	repeal	licens	reloc	depart	marit	address	appli	commenc
consid	appoint	abduct	contempt	appear	arrearag	divis	noncustod	uniform	exercis
circumst	attorney	civil	compli	chang	pay	asset	name	continu	forum
modif	report	januari	fail	princip	oblig	retir	number	foreign	appropri
plan	guardian	convent	certif	declar	oblige	real	social	communic	continu
joint	counti	intern	failur	disclosur	receiv	distribut	locat	articl	more
chang	litem	hagu	author	address	due	benefit	secur	given	stay
both	committe	countri	suspens	give	clerk	dure	telephon	over	warrant
educ	counsel	aspect	bond	plead	deduct	equit	testimoni	exclus	through
guidelin	investig	session	suspend	without	paid	acquir	avail	individu	pend

4.2 Joint Cluster Topic Models

While the topics shown appear to be semantically coherent, the LDA+KMeans model has several basic weaknesses. By assuming a fixed representation for the data, the LDA+KMeans model misses out on potential information that could be gained in response to an understanding of the cluster quality. More advanced models, such as the modified version of one of the cluster topic model (CTM) described in [3] that we implement here seeks to re-estimate the topic model along with the clustering model.

4.3 Integrating Title Information

However, intrinsic to our dataset is the presence of guiding "title" information. While there is no uniform methodology by which topic titles are granted (indeed, some states, such as Arizona, simply number the statutes), title information can be extremely predictive of the semantic contents, even more so than the actual text of the statutes themselves. Our goal is to enhance several standard generative clustering models by requiring that each cluster also generate the title words in addition to the body text.

In this case, we modify the generative process described in Section 4.2. The generative process now provides for a document to be drawn from several clusters.

1. Sample a set of documents from some set of clusters, $\vec{c}$

2. For each word in the statute body
    - Sample a cluster, c_i from $\vec{c}$
    - Sample a topic z_i from θ_{c_i}
    - Sample a word from β_{z_i}
3. For each word in the statute title
    - Sample a topic z_i from θ_{c_i}
    - Sample a word from β_{z_i}

In order to learn our model, we initialize the word topic assignments of the body text with LDA, and then run Gibbs sampling. The assignments used in the evaluation represent the 10th sample from our sampling procedure.

In examining the output of this method, it appears that some clusters carry broader themes consisting of multiple tags, but they are unable to make the types of fine-grained distinctions that the annotators were ostensibly capable of making. These qualitative realizations are evidenced in the tag-frequency counts shown in Table 4.5.

4.4 Evaluation

We evaluate the clustering results along several pairwise and clustering metrics. For pairwise alignment, we consider precision, recall, and F1 scores. Given that multiple tags can be applied by any one annotator, we define a 'true positive' to be a pair of statutes both clustered together and sharing a gold tag.

Table 4.5: Selected coherent statute clusters

Cluster: Parenting + Custody	**Cluster: Court Procedures**	**Cluster: Alimony, Property, and other Settlement Issues**
child custody:11	alimony:1	
relocating:1	legal separation:6	child support:10
parenting plan:10	temporary separation:1	dissolution of marriage:1
	alimony:4	child custody:2
	dissolution of marriage:2	
	illegal divorce procedures:5	
	child custody:2	health insurance:3
	misc:8	property:1
	remarriage:6	
	annulment:3	
	judical proceedings:1	
	grounds for divorce:12	
	name change:2	
	general divorce law:8	
	waiting period:1	
	name:1	

4.5 Dimension Reduction

In tagging statutes, it becomes immediately clear that the presence of certain key words and phrases, especially within the title, are all that is needed to distinguish between statutes tagged in different ways. Therefore, we explore ways to refine the representation of the statutes so as to diminish noise that can make distinguishing between statutes more difficult.

As is standard practice in many natural language processing methods, stop words [such as articles] are removed from the text; as none of our models employ syntactic information, these words serve little purpose in our analysis. We also eliminate from consideration any words which occur in fewer than 5 states; while 5 is an arbitrary limitation, it is designed to ensure that the words used in clustering can be utilized in clustering, rather than just serving as extra noise.

4.6 Pairwise Alignments

We also explored direct alignment between two states using a simple, greedy TF-IDF cosine matching algorithm. Despite being incapable of learning the underlying topics, the pairwise alignment manages to outperform the clustering alignments by some metrics as is shown in Table 4.6.

Table 4.6: Performance on Pairwise Alignment Task: Divorce Statutes, Arkansas and Utah

Algorithm	**Precision**	**Recall**	**F1**
WV + KMeans [Body]	0.344	0.285	0.312
WV + KMeans [Titles]	0.249	0.776	0.377
LDA+KMeans [Body]	0.392	0.095	0.153
LDA + KMeans [Titles]	0.341	0.086	0.137
CTM + Titles	0.310	0.083	0.131
TF-IDF cosine [Body]	0.471	0.157	0.235

Chapter 5

Future Work

With additional annotations, one could imagine understanding in greater detail the universality of various tags. With this knowledge, it would become possible to impose constraints on the cluster composition to reflect how frequently tags are observed across states.

This work is intended to provide the foundation for further development of comparison tools between state laws. News and public policy organizations frequently publish categorical summaries of the differences between state laws on a particular issue. While alignment, or identification of relevant domains of law across different states, is a first step toward automatic production of these categorical summaries, significant work remains to be done at the intersection of information retrieval, classification, and syntactic analysis. Our initial empirical investigations of this unique, categorical Q&A task suggested that it would first be necessary to determine which parts of

the statutes pertained to one another; while we have considered fixed alignment solutions (in which the alignment of statutes is based on some notion of 'semantically-related' clusters that can be reasonably expected to have some overlap in content, these alignments could serve merely as a precursor for dynamic alignments for specific questions. Depending on the type of question at hand and the organization of the relevant information across statutes.

Additional data within or around these domains may alleviate some of the issues observed with projections of our data to lower dimensions. While the clusters appeared to be coherent at the highest-level, we were unable to obtain significant gains with a different representation.