Topic models

Thomas R. Bruce / Sara Frug
Legal Information Institute, Cornell University

Four questions we'll answer

- What is unsupervised topic modeling?
- What is it good for?
- What have we (and others) used it for?
- What else could we (or you) do with it?

An important new tool

- A tool for discovering “hidden” topics in large collections of documents
  - think: lots of (unpublished?) cases, large document collections of unknown composition, grey literature, historical collections of statutes and regulations
- Rose to prominence in digital humanities, especially history
- Part of a set of law-and-computation technologies that *everyone* needs to know

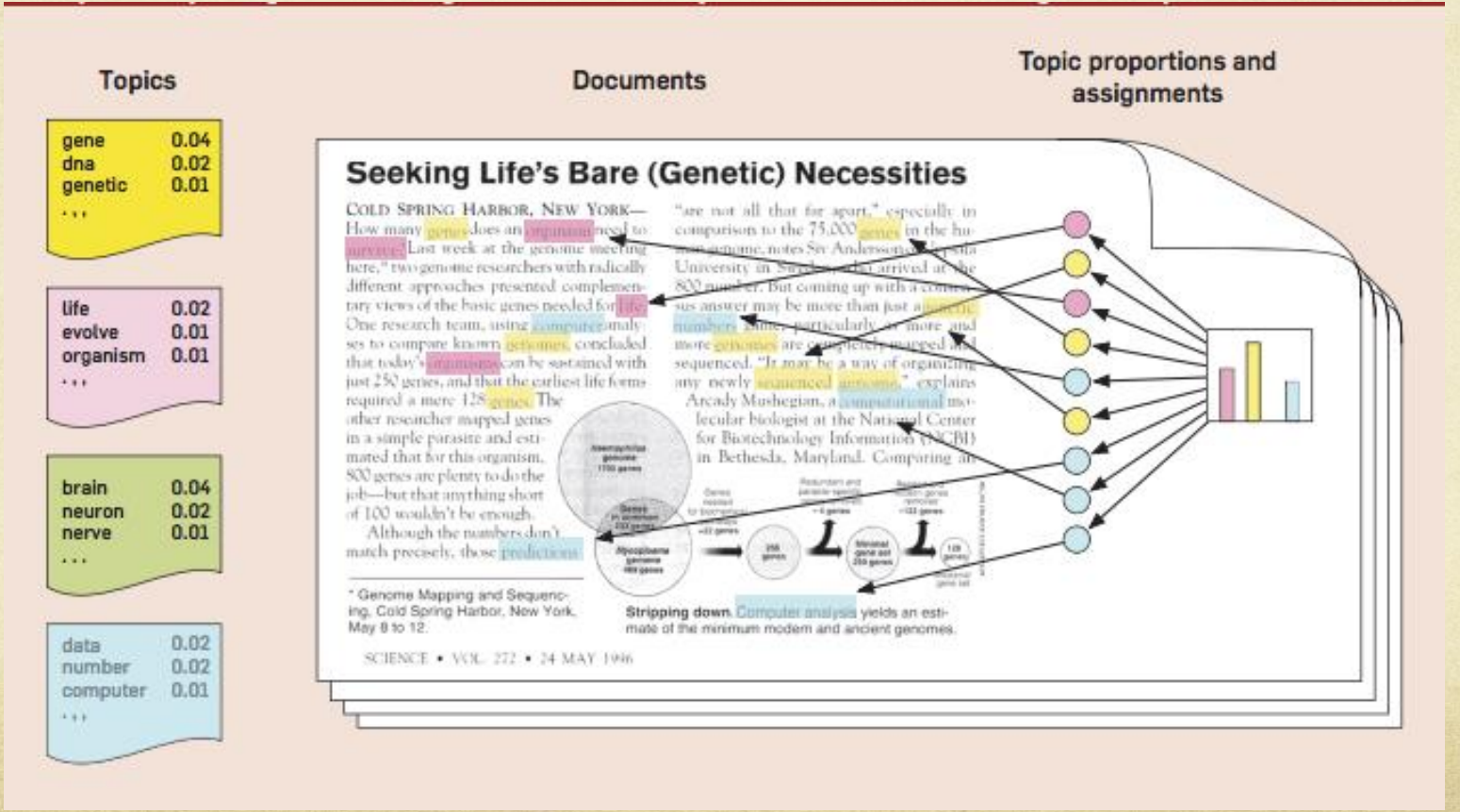
In other words:

- It does your skimming for you.
- It will sort hay into smaller, thematic haystacks.
- It will *not* find the needle in the haystack.

What is it? ....the basics

- Assume that a document collection represents a discourse...
- ... and that the discourse contains a number of “topics”
- For this purpose, a “topic” is represented as a cluster of associated keywords, often found together
- Guess the number of topics, and use fancy machine-learning based software to find the clusters by repeated guessing

What is it? ... in pictures



What is topic modeling good for?

- Finding out what's in a big collection of documents.
- Comparing discourse/topics between collections or between temporal snapshots of the same collection.
  - Looking at how developed the law is in different jurisdictions (eg. for business siting, planning law reform, etc.)
  - Comparing reactions to policies
- Lots of really weird stuff
  - identifying 19th C. trade routes
  - finding narrative patterns in IRS written responses

What's it look like?

- MALLET software (our tool of choice) creates 3 or 4 files of mathematical-looking gobbledegook.
- Easiest way to understand the model is via the keyword file
- Here's a labeled version

How do you make one?

- Get a collection of documents
- Guess at a “good” number of topics (usually 100)
- Run
- Tune stopwording
- Increase number of topics
- Lather, rinse, repeat
- Label the model

Preprocessing

- Terminology extraction
  - sometimes we want to work with phrases, not individual terms
  - collateral vs. collateral estoppel vs. estoppel vs. collateral damage
  - eg. adding war-crimes cases to debtor/creditor
- Stemming, lemmatization
- Normalization of terms
  - eg conforming “indigenous people” and “Native American” and “American Indian” to “indigenous people”
- Stopwording
  - tends to be labor intensive

Stopwording

- Puts focus on a single discourse in a collection that contains more than one
  - eg in caselaw, substantive topics vs. jurisdiction/geography
  - eg in most legal materials, procedure vs everything else
- Suppression of low-value (common) terms
- Plus the algorithm just derails every now and then.

How to know when you're done

- This is more like cookery than math
- We look for
  - topical coherence
  - topical cohesion
  - useful granularity of overall set of discovered topics
  - <10% junk topics
- Tuning and labeling can readily be done by law students

Streamlining model development

- It's early days for tools: still too project-specific and not well maintained
- Text files and spreadsheets
- Visualization for data exploration
  - stopword tuning
  - stemming, phrasing decisions
  - model sizing
  - gauging temporal stability of the model

Visualizer: topic list

CFR - Title 9 | Select Model | Import | **Overview** | Evaluate | Topic | Document | Word | Bibliography | Word index | Settings | About

Grid | Scaled | List | Years

click a column label to sort; click a row for more about a topic

topic	2000—2013	top words	proportion of corpus
1		program information redesignated official container collected years agriculture voluntary authorized blood equivalent sealed state certified	0.2%
2		inspector kind information line approved provisions amended service birds foreign commerce review slaughter factor contents	0.6%
3		fees test service fee aphis services person import payment provided inspection veterinary states diagnostic amended	0.6%
4		agents toxins aphis registration official responsible animal entity part listed individual cdc section exemption security	0.3%
5		act inspection section inspect post racking street disqualified licensing requirements excused representatives request paragraph period	0.2%
6		facilities states swine animal entry interstate fees cattle required importation product disease import quarantine permits	0.4%
7		imported entry section bird administrator requirements credit port listed national address paid plant canada zoo	0.5%
8		animal inspection production required release produced killed twenty redesignated medium requirements aphis filed source plate	0.6%
9		system drying air days lower information examinations continuous hours maintenance inches size method smoked top	0.3%
10		carcasses establishment carcass parts inspector viscera head test section slaughter paragraph number accordance bulls removed	0.4%
11		tubes official paragraph clean laboratory state eggs chick department flocks accordance units collect provided identification swab	0.2%
12		poultry birds permit end infected special eggs quarantined federal manure representative administrator accordance litter items	0.7%
13		veterinarian accredited state perform orientation accreditation duties application program veterinary applicant test category certificate written	0.2%
14		products requirements utensils product equipment subchapter accordance prepared sanitizing preparation custom provisions restaurant owner household	0.3%
15		process paragraph establishment official food maintained ensure minimum device sample venting operation cooked condensate plant	0.8%
16		food part meat subchapter regulations act livestock carcass substance term official article bears animal inspected	0.2%
17		inspection cattle cows hfrs testing official rates inspectors animal post table thereof washbasin made veterinary	0.2%
18		book secretary stud society record association animal england herd certificate breed st sheep flock canadian	0.4%

Visualizer: topic view

CFR - Title 9 | Select Model | Import | Overview | Evaluate | **Topic** | Document | Word | Bibliography | Word index | Settings | About

91 facility aphis ruminants area operator holding security representative approval privately minimum subpart official owned application

α = 0.001; 0.6% of corpus.

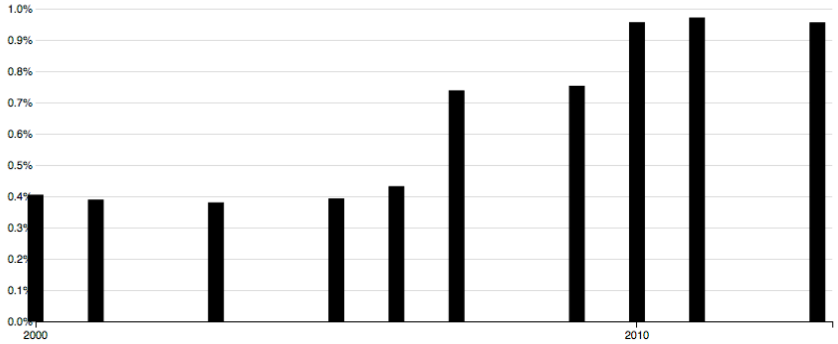
Top words

Word	Weight
facility	
aphis	
ruminants	
area	
operator	
holding	
security	
representative	
approval	
privately	
minimum	
subpart	
official	
owned	
application	
office	
movement	
live	
animals	

Yearly proportion of words in topic

Click a bar to limit documents to that year

clear selected year



Top documents

Document	%	Tokens
OFR, . "9 CFR Part 93 - IMPORTATION OF CERTAIN ANIMALS, BIRDS, FISH, AND POULTRY, AND CERTAIN ANIMAL, BIRD, AND POULTRY PRODUCTS; REQUIREMENTS FOR MEANS OF CONVEYANCE AND SHIPPING CONTAINERS." 0, no. 0 (January 2011): .	10.5%	1945
OFR, . "9 CFR Part 93 - IMPORTATION OF CERTAIN ANIMALS, BIRDS, FISH, AND POULTRY, AND CERTAIN ANIMAL, BIRD, AND POULTRY PRODUCTS; REQUIREMENTS FOR MEANS OF CONVEYANCE AND SHIPPING CONTAINERS." 0, no. 0 (January 2013): .	10.5%	1952
OFR, . "9 CFR Part 93 - IMPORTATION OF CERTAIN ANIMALS, BIRDS, FISH, AND POULTRY, AND CERTAIN ANIMAL, BIRD, AND POULTRY PRODUCTS;	10.3%	1913

Visualizer: document view

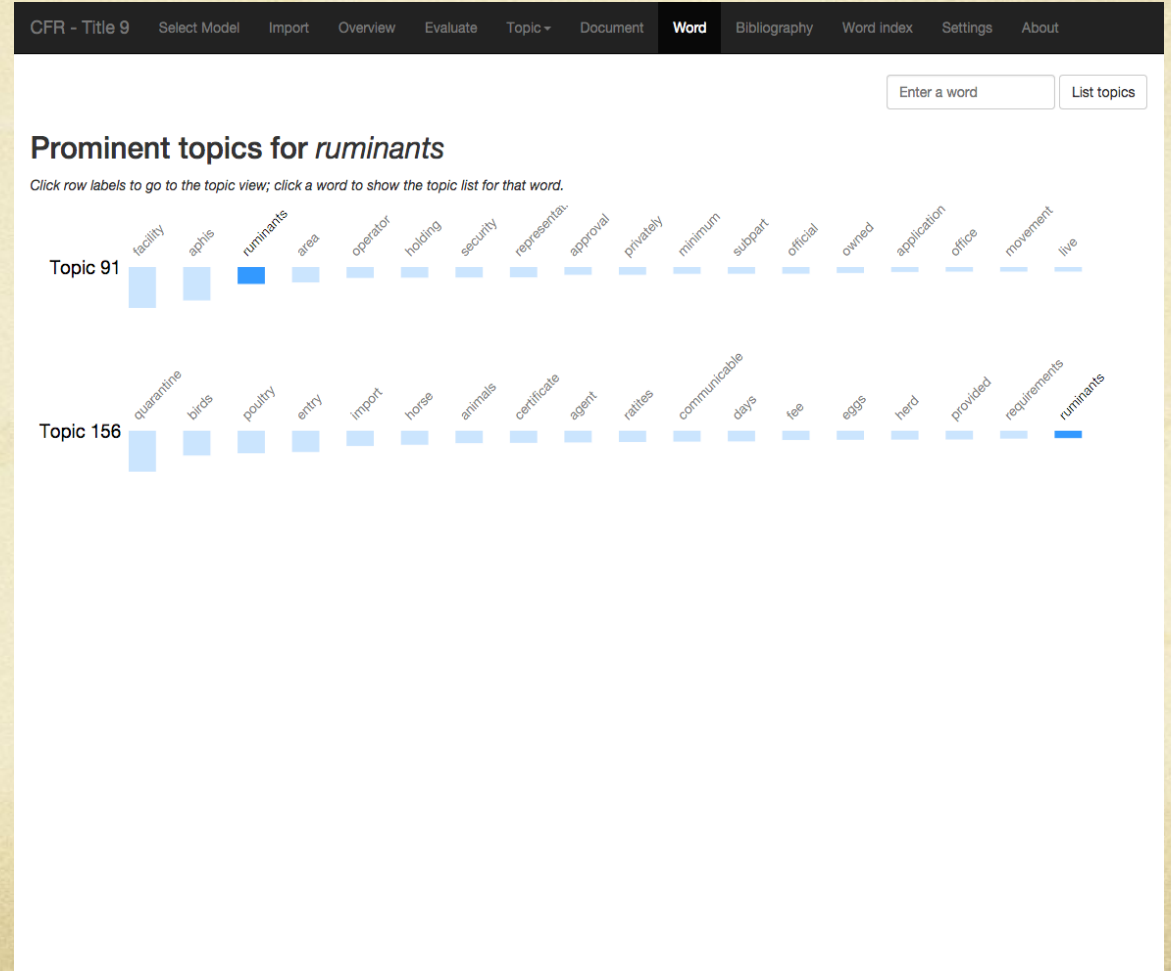
CFR - Title 9 | Select Model | Import | Overview | Evaluate | Topic | Document | Word | Bibliography | Word index | Settings | About

OFR, . "9 CFR Part 93 - IMPORTATION OF CERTAIN ANIMALS, BIRDS, FISH, AND POULTRY, AND CERTAIN ANIMAL, BIRD, AND POULTRY PRODUCTS; REQUIREMENTS FOR MEANS OF CONVEYANCE AND SHIPPING CONTAINERS." 0, no. 0 (January 2011): .

18494 tokens. View on JSTOR

Topic	%	Tokens
158 united states region disease importation inspection paragraph animal approved administrator accordance imported time part port	20.2%	3731
156 quarantine birds poultry entry import horse animals certificate agent ratites communicable days fee eggs herd	18.3%	3380
130 horses facility redesignated port cattle inspector amended provided veterinarian swine origin goats administrator part state	10.6%	1957
173 importer section amended date service canada permit port export free made inspection sheep mexico cem	10.6%	1953
91 facility aphis ruminants area operator holding security representative approval privately minimum subpart official owned application	10.5%	1945
70 facilities required premises accompanied applicable person horses supervision diseases handling issued sheep received livestock accredited	6.5%	1198
68 services number shipment period intended paragraph agreement regulations identification payment additional writing treated months cancellation	6.4%	1182
7 imported entry section bird administrator requirements credit port listed national address paid plant canada zoo	5.6%	1033
98 fish vhs regulated live state facility species virus water gametes health svc aphis federal testing	3.9%	724
29 permit issued application health transit imported animal states united service required amended permits plant sale	3.2%	595
142 animal health state states equipment veterinarian treatment section set containers united responsibility contact waste disinfectant	2.8%	523
1 program information redesignated official container collected years agriculture voluntary authorized blood equivalent sealed state certified	1.5%	273

Visualizer: word view



Visualizer: word list

CFR - Title 9 | Select Model | Import | Overview | Evaluate | Topic | Document | Word | Bibliography | **Word index** | Settings | About

All words prominent in any topic

Words not prominent in any topic are not listed

aboard
absorption
access
accompanied
accordance
account
accounts
accreditation
accredited
accuracy
acid
acquisition
act
action
actions
activities
activity
add
added
additional
address
adequate
adjacent
administration
administrator
admitted
adulterated
adulteration
advanced
advertising
affected
african
agar
age
agencies
agency
agent
agents
agglutination
chicks
children
chiller
chilling
circuit
city
claim
claims
class
classification
classified
clean
cleaned
cleaning
closing
clostridium
club
cm
code
collect
collected
collection
collectively
color
colored
combination
commerce
commercial
committee
communicable
competing
complaint
completed
compliance
complies
con
condemnation
condemned
condensate
due
duties
duty
eagle
eartag
ec
edible
edinburg
effective
efs
egg
eggs
electric
electrical
eligible
embryo
embryos
employed
employee
employees
enclosure
enclosures
encoding
end
england
enrichment
ensure
enteritidis
entity
entry
enzyme
equal
equine
equines
equipment
equivalent
eradication
established
establishment
imported
importer
inauguration
inch
inches
include
included
including
increment
incubate
indemnity
indicating
individual
individually
industry
inedible
infants
infected
infection
infested
influenza
informal
information
ingredients
initial
inject
injection
injury
inoculated
inoculation
insanitary
inspect
inspected
inspection
inspector
inspectors
intended
interest
interests
nutrient
nutrients
nutrition
nvsl
obligations
observed
obtained
occur
ocean
odors
offered
office
officer
offices
official
oleomargarine
opening
operating
operation
operations
operator
opportunity
oral
order
organ
organisms
organization
organs
orientation
origin
original
outbreaks
outline
overlap
overtime
owned
owner
owners
packaged
rabbits
racking
rate
rates
ratites
raw
rbc
reaccreditation
reactor
reactors
ready
reason
reasons
received
recognized
recommended
record
recorded
recording
records
redesignated
reference
refused
region
regional
regions
register
registered
registrants
registration
regulated
regulation
regulations
regulatory
release
removal
remove
removed
rendered
spread
square
st
stage
stalls
stanchions
standalone
standard
standardized
standards
state
statement
states
station
status
steam
sterile
stimulating
stockyard
stockyards
storage
stored
street
stud
stunning
style
subchapter
subgroup
subject
submitted
subpart
subsection
subserial
substance
subtypes
sufficient
sunday
supervision
supervisor

Visualizer: stopword evaluation

CFR - Title 9 | Select Model | Import | Overview | **Evaluate** | Topic | Document | Word | Bibliography | Word index | Settings | About

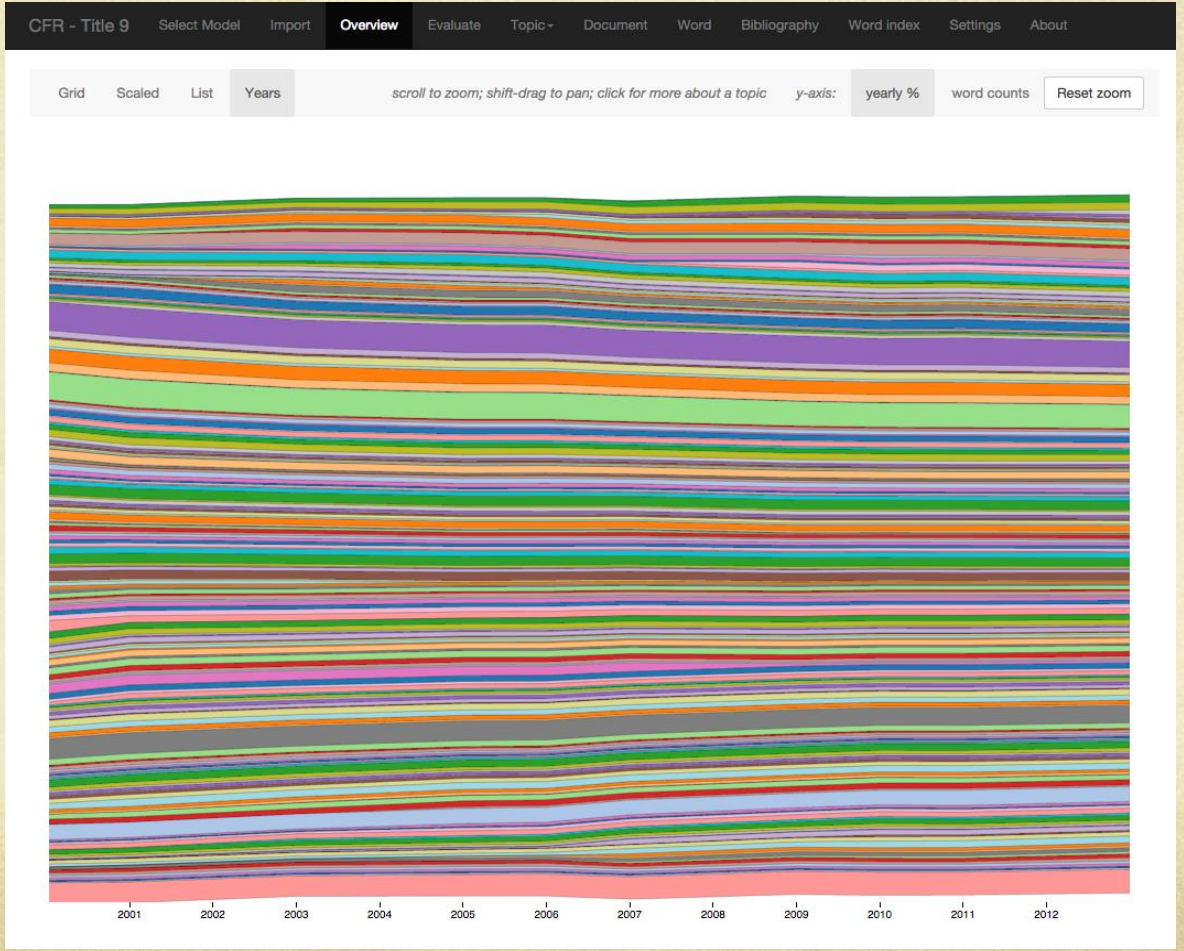
Stopwords

stopwords are highlighted in red

topic	top words
1	program information redesignated official container collected years agriculture voluntary authorized blood equivalent sealed state certified
2	inspector kind information line approved provisions amended service birds foreign commerce review slaughter factor contents
3	fees test service fee aphis services person import payment provided inspection veterinary states diagnostic amended
4	agents toxins aphis registration official responsible animal entity part listed individual cdc section exemption security
5	act inspection section inspect post racking street disqualified licensing requirements excused representatives request paragraph period
6	facilities states swine animal entry interstate fees cattle required importation product disease import quarantine permits
7	imported entry section bird administrator requirements credit port listed national address paid plant canada zoo
8	animal inspection production required release produced killed twenty redesignated medium requirements aphis filed source plate
9	system drying air days lower information examinations continuous hours maintenance inches size method smoked top
10	carcasses establishment carcass parts inspector viscera head test section slaughter paragraph number accordance bulls removed
11	tubes official paragraph clean laboratory state eggs chick department flocks accordance units collect provided identification swab
12	poultry birds permit end infected special eggs quarantined federal manure representative administrator accordance litter items
13	veterinarian accredited state perform orientation accreditation duties application program veterinary applicant test category certificate written
14	products requirements utensils product equipment subchapter accordance prepared sanitizing preparation custom provisions restaurant owner household
15	process paragraph establishment official food maintained ensure minimum device sample venting operation cooked condensate plant
16	food part meat subchapter regulations act livestock carcass substance term official article bears animal inspected
17	inspection cattle cows hfrs testing official rates inspectors animal post table thereof washbasin made veterinary
18	book secretary stud society record association animal england herd certificate breed st sheep flock canadian
19	products animals title animal amended authority version xml utf standalone encoding noted source section provided
20	water protocol chilling temperature system data retained retention equipment process chiller type time establishment testing
21	carcasses lesions condemned carcass food human cattle showing found tuberculosis livestock disease meat condition swine
22	container approved licensed amended enclosure biological production animal requirements section approval health prescribed including live
23	laboratory average designated paragraph determine date part based equal requirements provided greater standard procedures previously
24	product establishment sop sanitation adulteration insanitary equipment utensils rooms creation maintained fsis official procedures edible
25	program establishment carcasses employees selected establishments cooperative inspection shipment request administrator services interstate handling located operating
26	agency packers payment grower scales inspection coverage weights stockyards time custodial weighed test copy filed

Visualizer: scaled view

Visualizer: “years” view



Things we've done

- Categorization
  - caselaw, both general and bankruptcy
  - legal scholarship
- Discourse comparison
  - State statutes
  - CFR temporal snapshots

A final example: the Magic 8-Ball

- Inferencer… what's that?
- Trained on 25K cases from CourtListener
- Actually, 19K because we cut all cases under 3K bytes.
- 500 topics.
- "Site" implemented with JRuby, slightly hacked MALLET, Apache TIKA, and Tesseract (OCR) for input doc conversion
- Examples: kickboxing, whales, Roe, and Mabo.

Future work

- Automated stopword-list construction
  - eg. by named-entity recognition
- Automated labelling
- Hierarchical topic models
- Better visualizers

Us

- sara@liicornell.org
- tom@liicornell.org
- http://www.law.cornell.edu
- THANKS!

Resource guide: explanations

- Megan Brett's really easy explanation for non-technical people
- Blei article from Comm. of the ACM
- More Greek symbols than you ever want to see: Latent Dirichlet Allocation explained by Wikipedia.

Resource guide: survey papers

- David Blei's page
- David Mimno's bibliography

Resource guide: tools

- MALLET software tools for text mining
- David Mimno's blog
- DFR browser