NLP
TOPIC EXTRACTION

Gresa, Sue, Ganguly

SUPREME COURT LANDMARKS | UNITED STATES COURTS

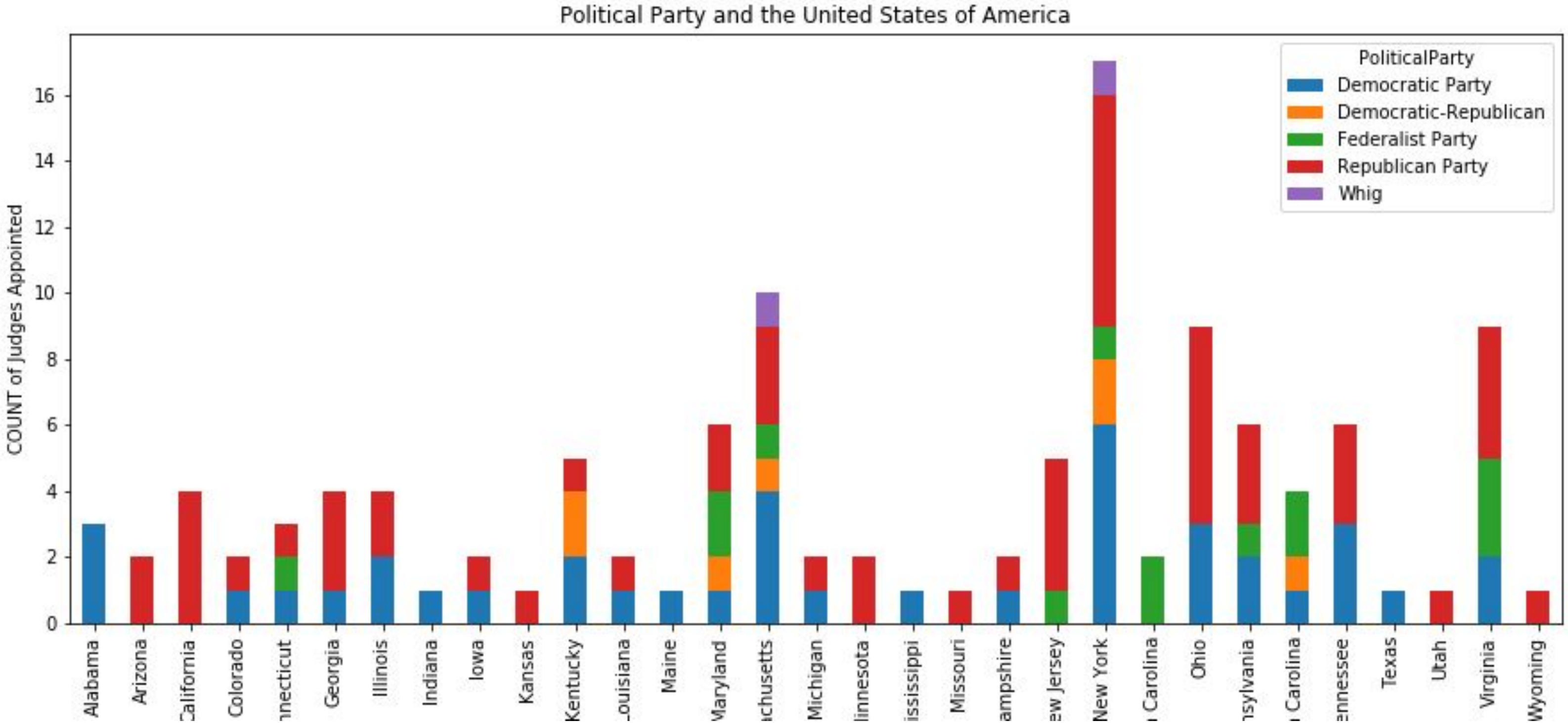
Challenges of NLP

- Ambiguity
- Common Knowledge
- Creativity
- Diversity Across Languages

Advantages of our Dataset

- Rules of Practice and Procedure
- Study of Operation and Effect
- Pending Amendments
- Glossary of Legal Terms

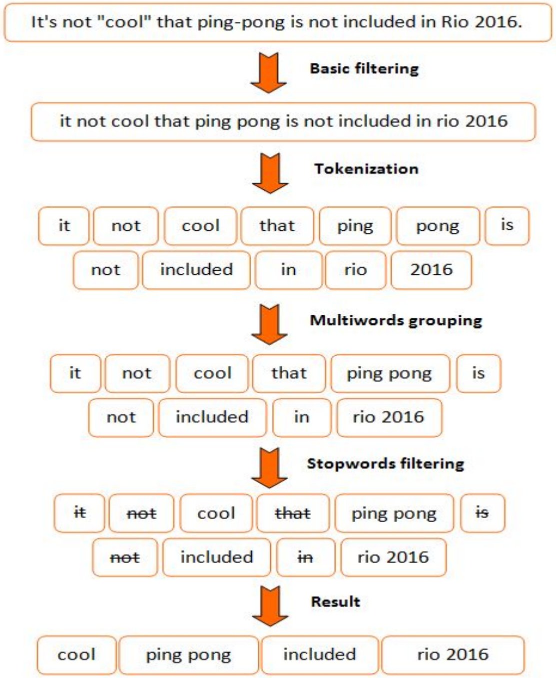
SUPREME COURT JUDGES and their STATE



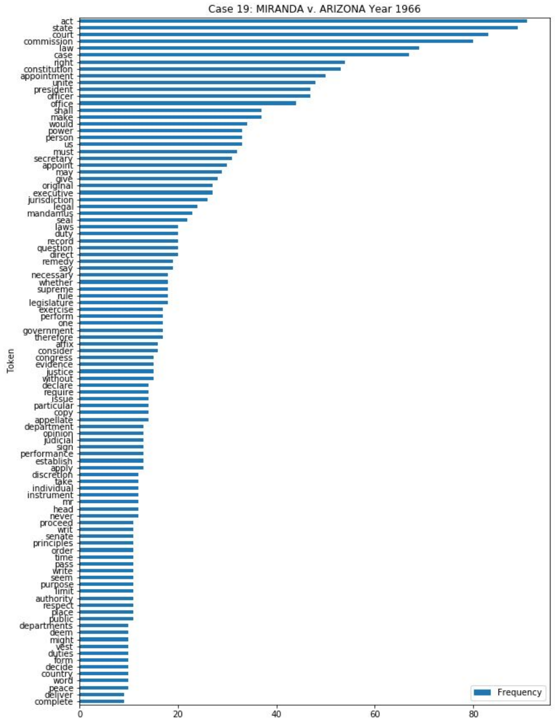
TOPIC EXTRACTION

Given a Landmark Case, we wondered if the main topics could be extracted from the case. Not only did we accomplish this task but Gresa worked on a model that gives us a high Coherence Score.

Tokens and their Weights (tf)



Case 19



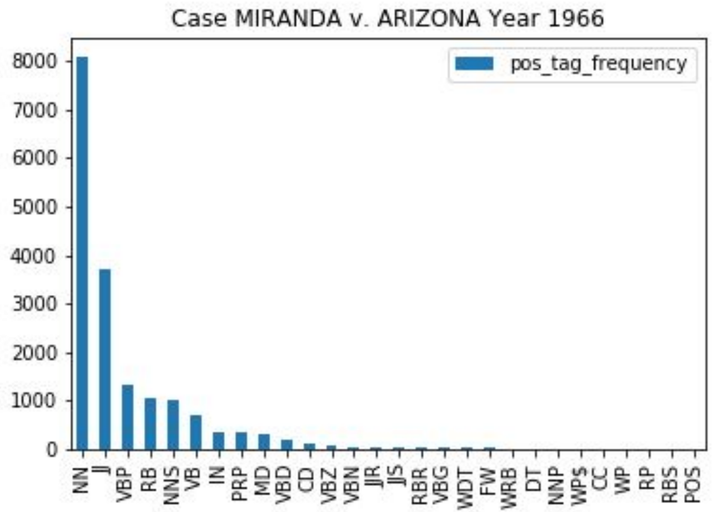
KNOWLEDGE GRAPH AND CONTEXT

By providing the full context for every piece of data extracted you will not only know "who said it?", "when did they say it?" and "what time period were they talking about?"

What we lose with tokens we gain with the Knowledge Graph from Sue.

POS Tagging (CONTEXT)

POS Frequency



Popular POS Tags: NN, JJ, VBP, RB, NNS

Tag	Description
CC	Coordinating conjunction
CD	Cardinal number
DT	Determiner
EX	Existential there
FW	Foreign word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List item marker
MD	Modal
NN	Noun, singular or mass
NNS	Noun, plural
NNP	Proper noun, singular
NNPS	Proper noun, plural
PDT	Predeterminer
POS	Possessive ending
PRP	Personal pronoun

Tag	Description
PRP$	Possessive pronoun
RB	Adverb
RBR	Adverb, comparative
RBS	Adverb, superlative
RP	Particle
SYM	Symbol
TO	to
UH	Interjection
VB	Verb, base form
VBD	Verb, past tense
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Whdeterminer
WP	Whpronoun
WP$	Possessive whpronoun
WRB	Whadverb

TF-IDF and LDA

TF-IDF (term frequency-inverse document frequency) can be thought of as a numerical metric that reflects how important a word is in a collection of corpus. Words that are frequent in a document but not across documents tend to have high TF-IDF score.

$$w_{x,y} = tf_{x,y} \times \log\left(\frac{N}{df_x}\right)$$

TF-IDF

Term x within document y

$tf_{x,y}$ = frequency of x in y

df_x = number of documents containing x

N = total number of documents

LDA Latent Dirichlet Allocation

The upper table shows words versus topics and the lower table shows documents versus topics.

Each column in the upper table and each row in the lower table must sum to 1.

LDA on the Texts of Harry Potter by Greg Rafferty

	Topic 0	Topic 1	Topic 2	Topic 3
harry	0.709	0.001	0.001	0.003
hermione	0.001	0.709	0.001	0.003
malfoy	0.001	0.001	0.709	0.003
magic	0.001	0.001	0.001	0.980
wand	0.284	0.001	0.001	0.003
robe	0.001	0.284	0.001	0.003
spell	0.001	0.001	0.284	0.003

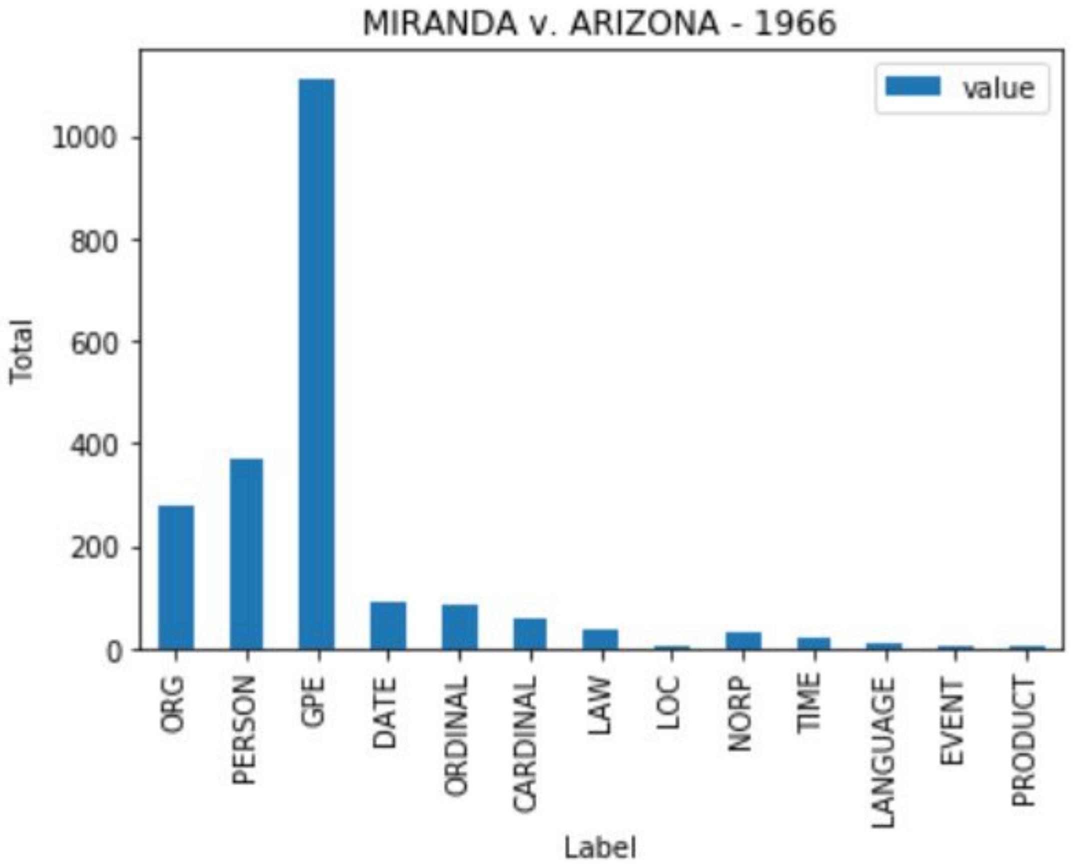
	Topic 0	Topic 1	Topic 2	Topic 3
Document 0	0.727	0.045	0.045	0.182
Document 1	0.045	0.727	0.045	0.182
Document 2	0.045	0.045	0.727	0.182
Document 3	0.318	0.318	0.318	0.045


2 3 4 5 6 7 8
E R T Y U I
D F G H J K

Named Entity Recognition NER

NER is a subtask of information extract that seeks to locate and classify named entities mentioned in the text into pre-defined categories such as person names, organizations, locations, time expressions, quantities, monetary values, etc.

	name	value
ORG		279
PERSON		368
GPE		1111
DATE		91
ORDINAL		83
CARDINAL		58
LAW		38
LOC		5
NORP		33
TIME		20
LANGUAGE		10
EVENT		5
PRODUCT		4



Parts Of Speech POS

A Part-Of-Speech Tagger (**POS** Tagger) is a piece of software that reads text in some language and assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc.

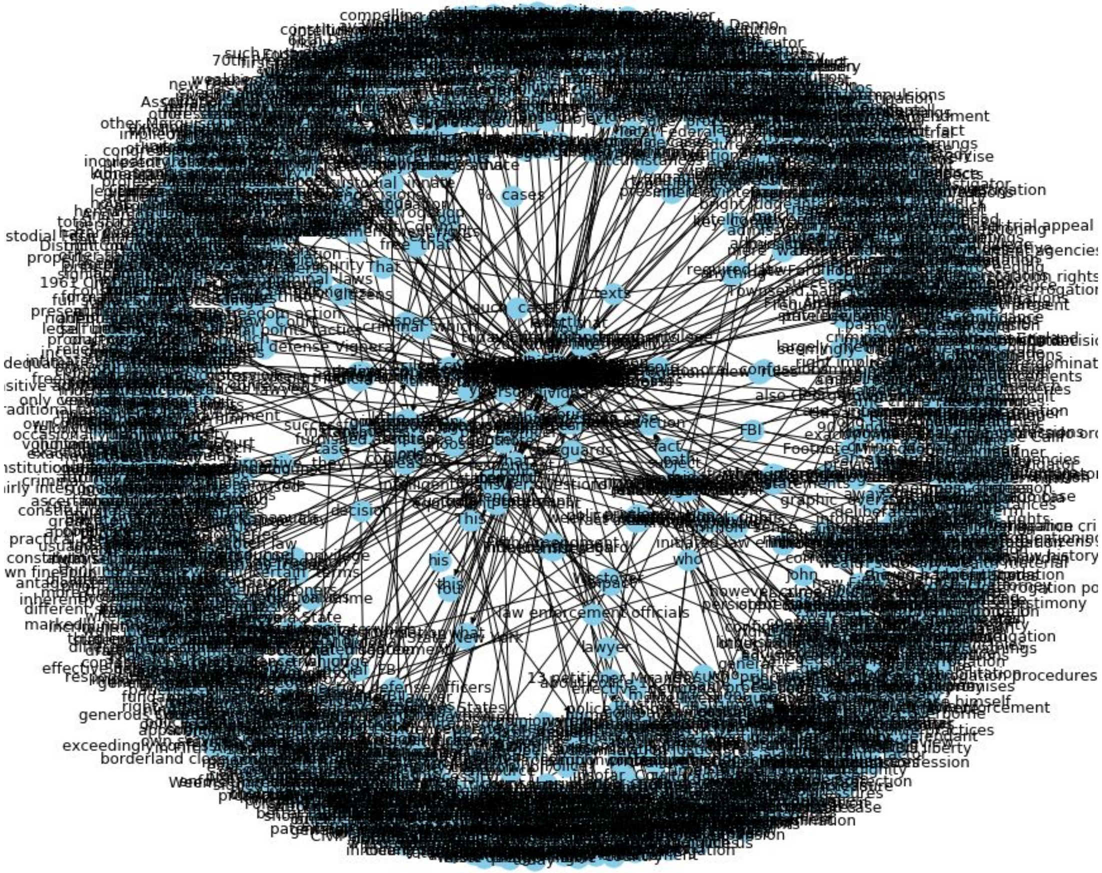
```python
doc = nlp("Such investigation may include inquiry persons not under restraint")

for tok in doc:
  print(tok.text, "...", tok.dep_)
```

```
Such ... amod
investigation ... nsubj
may ... aux
include ... ROOT
inquiry ... compound
persons ... dobj
not ... neg
under ... prep
restraint ... pobj
```

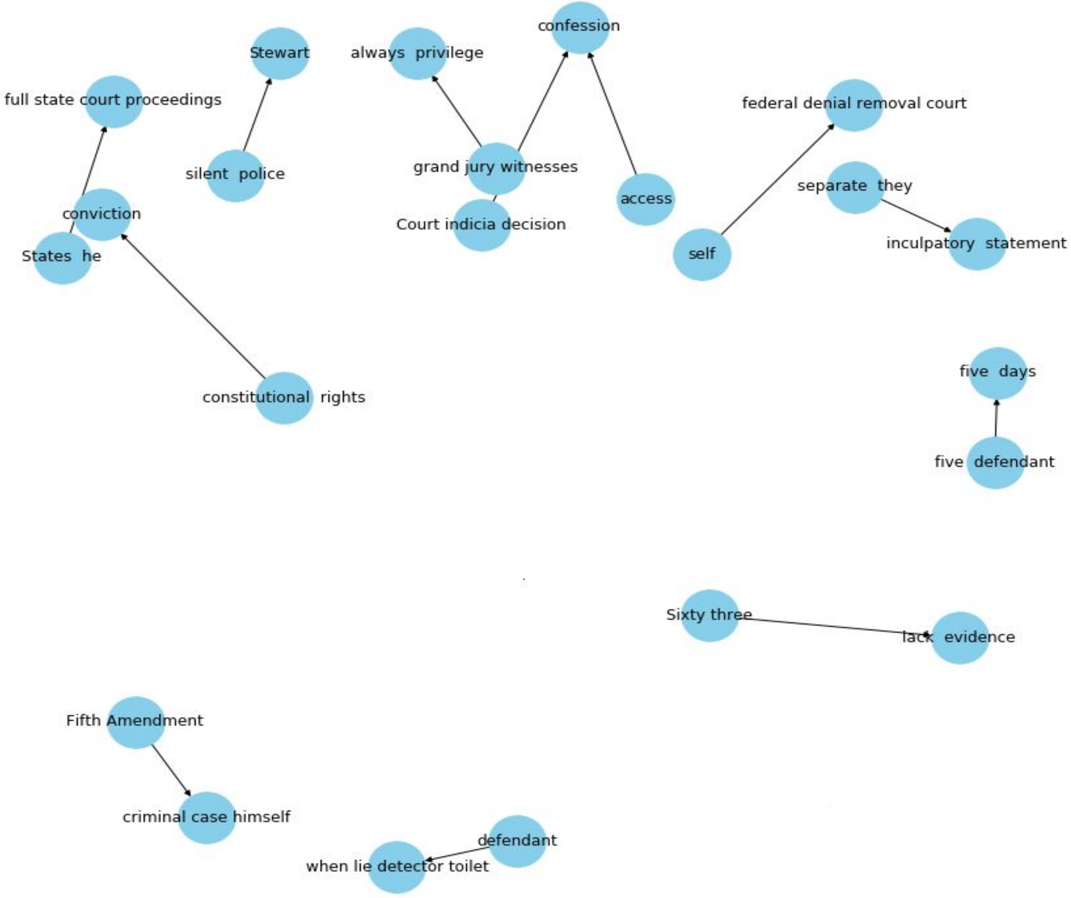
Knowledge Graph GK

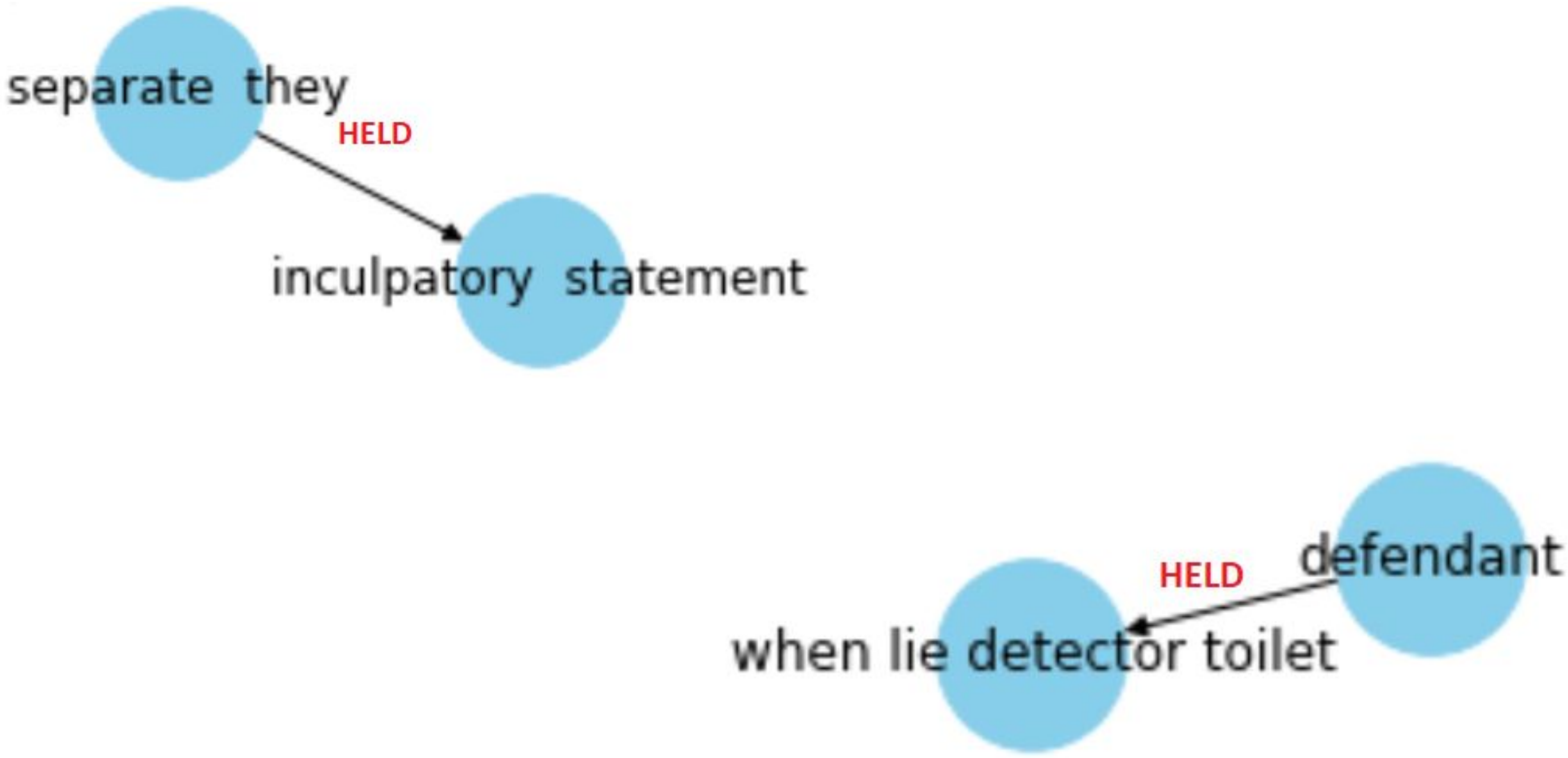
- This knowledge graph, a powerful foundation for a question-answer system, can then be traversed to provide answers.
- To build a KG from text, the machine must understand Natural Language(NLP)
- The program will go through the sentences and extract the subject and the object and when they are encountered – Relations(ROOT of the sentence)
- Facts about the case


law enforcement officials
FBI
lawyer
cases
texts
who
his
decision
John
Townsend
interrogation procedures

Most Frequent Relation:

HELD




separate they
HELD
inculpatory statement
defendant
HELD
when lie detector toilet

LDA Mallet Modeling Pipeline

Evaluating our model: Coherence Score and Range

Coherence Score assesses the quality of the topics by examining the degree of semantic similarity between each topic's top words.

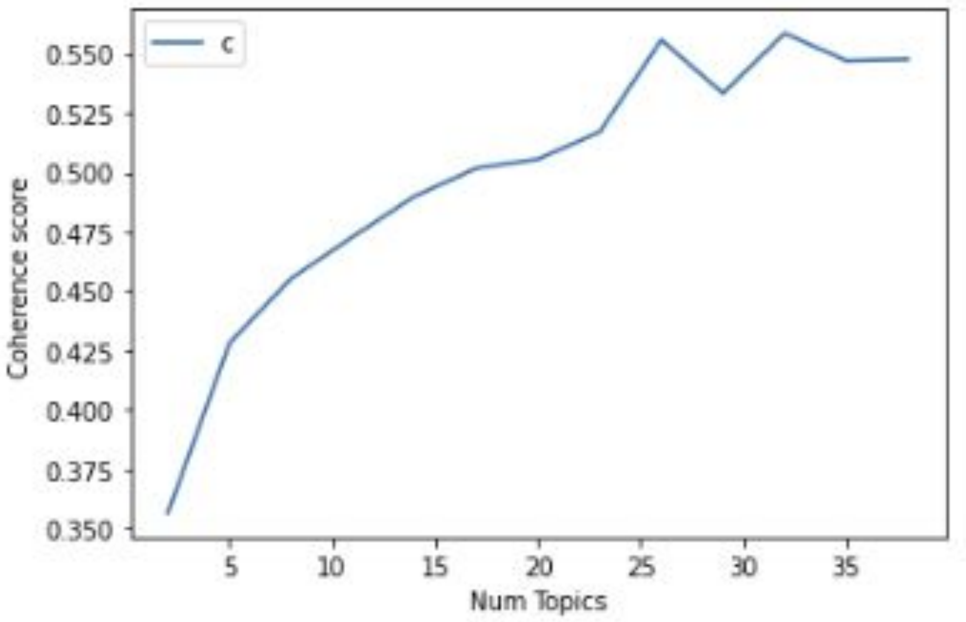
Ranges:

.3-.4 = Low

.5-.7 = Good

.8-.9 = Unlikely

Topic Modeling for Landmark Supreme Court Cases

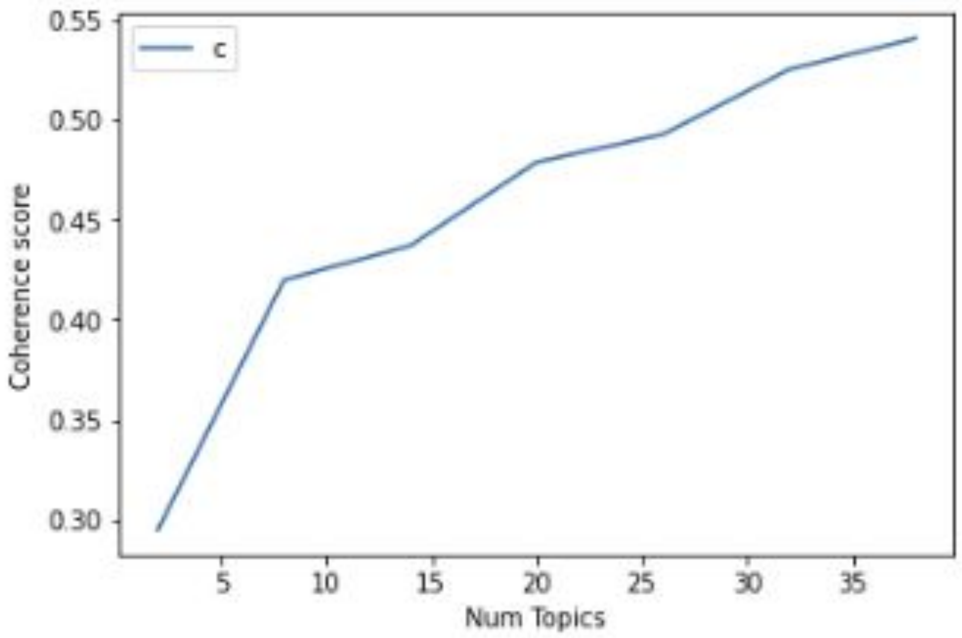


	num_topics	Coherence Score
10	32	0.558583
8	26	0.555781
12	38	0.547657
11	35	0.547029
9	29	0.533279
7	23	0.516981
6	20	0.505520
5	17	0.501700
4	14	0.489870
3	11	0.472870
2	8	0.455236
1	5	0.428044
0	2	0.356330

```
[(22,
  '0.030*"flag" + 0.015*"state" + 0.014*"unite" + 0.012*"government" + '
  '0.011*"expression" + 0.010*"statute" + 0.010*"american" + 0.009*"speech" + '
  '0.009*"conduct" + 0.008*"texas"'),
 (17,
  '0.027*"religious" + 0.022*"school" + 0.021*"children" + 0.017*"amish" + '
  '0.013*"education" + 0.012*"religion" + 0.011*"public" + 0.010*"parent" + '
  '0.010*"prayer" + 0.010*"age"'),
 (4,
  '0.048*"state" + 0.030*"court" + 0.024*"power" + 0.024*"unite" + '
  '0.023*"territory" + 0.021*"slave" + 0.017*"congress" + 0.016*"government" + '
  '0.013*"constitution" + 0.012*"citizens"'),
 (7,
  '0.019*"program" + 0.015*"race" + 0.014*"school" + 0.013*"title_vi" + '
  '0.012*"white" + 0.011*"discrimination" + 0.009*"action" + 0.008*"racial" + '
  '0.008*"federal" + 0.008*"negro"'),
 (12,
  '0.031*"public" + 0.014*"charge" + 0.014*"publish" + 0.011*"press" + '
  '0.010*"official" + 0.009*"publication" + 0.009*"libel" + 0.009*"warehouse" '
  '+ 0.008*"business" + 0.008*"government"'),
 (1,
  '0.034*"state" + 0.026*"law" + 0.014*"constitution" + 0.014*"act" + '
  '0.014*"make" + 0.009*"part" + 0.009*"time" + 0.008*"limit" + '
  '0.007*"exercise" + 0.007*"establish"'),
 (21,
  '0.024*"arm" + 0.020*"militia" + 0.019*"second_amendment" + 0.015*"state" + '
  '0.012*"bear_arm" + 0.011*"military" + 0.010*"amendment" + 0.010*"district" '
  '+ 0.010*"gun" + 0.009*"keep_bear"'),
 (9,
  '0.030*"state" + 0.028*"interest" + 0.023*"life" + 0.019*"treatment" + '
  '0.016*"medical" + 0.013*"patients" + 0.012*"patient" + 0.009*"person" + '
  '0.009*"evidence" + 0.008*"decision"'),
 (18,
  '0.024*"candidate" + 0.019*"candidates" + 0.018*"political" + '
  '0.017*"commission" + 0.016*"election" + 0.015*"party" + 0.013*"committee" + '
  '0.012*"congress" + 0.012*"contributions" + 0.012*"provision"'),
 (23,
  '0.058*"power" + 0.051*"state" + 0.031*"congress" + 0.019*"laws" + '
  '0.019*"commerce" + 0.018*"regulate" + 0.015*"exclusive" + 0.014*"grant" + '
  '0.014*"subject" + 0.012*"trade"')]
```

Topic Modeling for Miranda v. Arizona

	num_topics	Coherence Score
6	38	0.540782
5	32	0.525332
4	26	0.492798
3	20	0.478898
2	14	0.437135
1	8	0.419988
0	2	0.294958



Model topics and score prior to parameter tuning

```
/usr/local/lib/python3.6/dist-packages/smart_open/smart_open_lib.py:254: UserWar
  'See the migration notes for details: %s' % _MIGRATION_NOTES_URL
[(0,
  '0.082*"state" + 0.035*"unite" + 0.020*"federal" + 0.017*"crime" + '
  '0.015*"require" + 0.014*"criminal" + 0.014*"law_enforcement" + 0.013*"law" '
  '+ 0.012*"effective" + 0.012*"general"'),
 (1,
  '0.032*"accuse" + 0.030*"evidence" + 0.024*"constitutional" + '
  '0.016*"justice" + 0.014*"waiver" + 0.013*"rev" + 0.013*"prior" + '
  '0.013*"fbi" + 0.013*"person" + 0.013*"constitution"'),
 (2,
  '0.045*"question" + 0.026*"defendant" + 0.021*"time" + 0.020*"arrest" + '
  '0.018*"officer" + 0.018*"suspect" + 0.017*"compel" + 0.016*"obtain" + '
  '0.016*"fact" + 0.015*"subject"'),
 (3,
  '0.076*"interrogation" + 0.049*"counsel" + 0.044*"privilege" + 0.036*"warn" '
  '+ 0.026*"individual" + 0.025*"attorney" + 0.022*"fifth_amendment" + '
  '0.021*"present" + 0.017*"today" + 0.014*"practice"'),
 (4,
  '0.112*"court" + 0.018*"hold" + 0.018*"criminal" + 0.017*"make" + '
  '0.017*"decision" + 0.012*"circumstances" + 0.012*"point" + '
  '0.011*"california" + 0.010*"show" + 0.010*"law"')]

Coherence Score:  0.3727912590472383
```

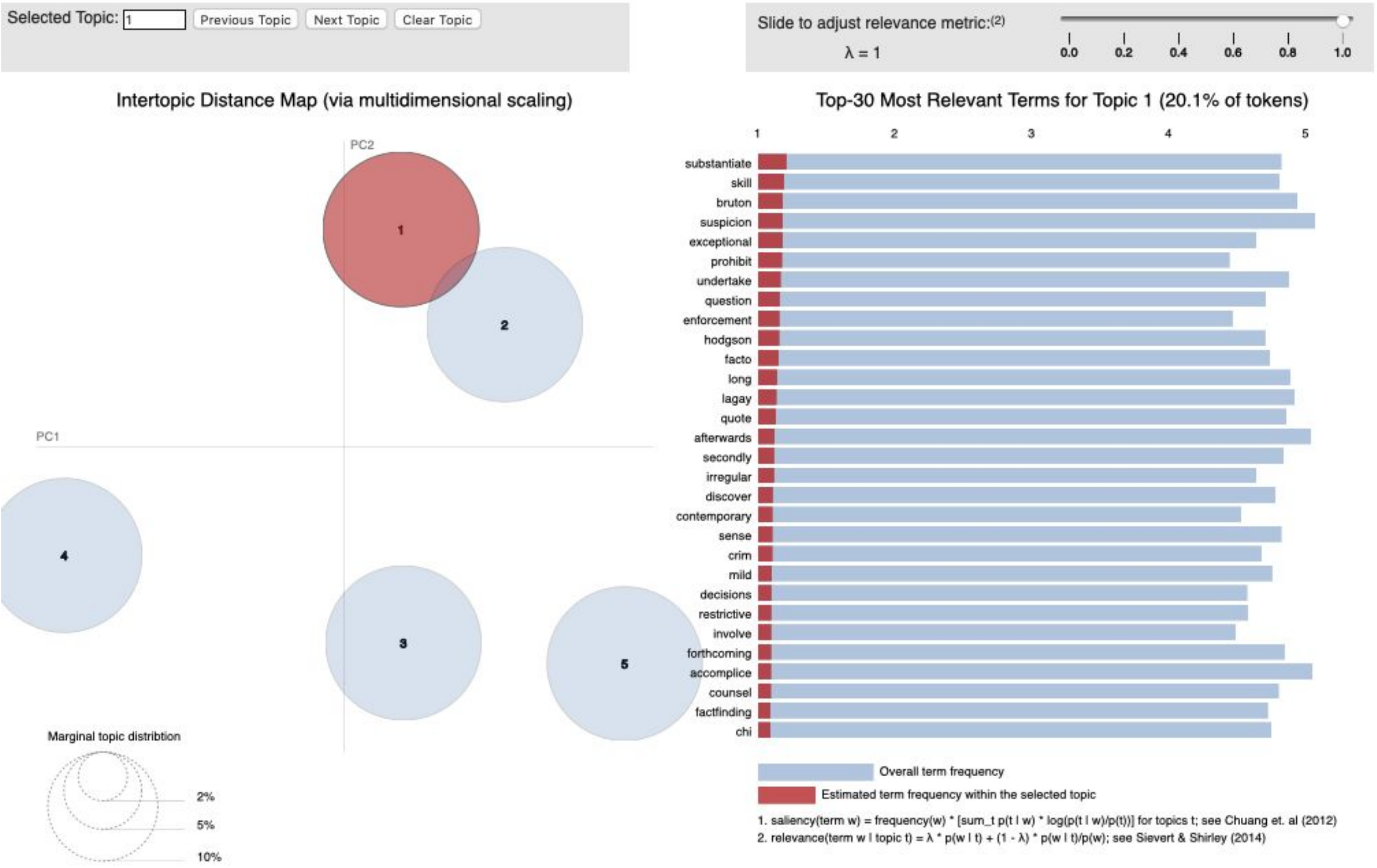
Model topics after parameter tuning

```
[(19,
  '0.082*"remain_silent" + 0.076*"interrogate" + 0.076*"lawyer" + '
  '0.059*"speak" + 0.059*"station" + 0.035*"country" + 0.029*"talk" + '
  '0.026*"result" + 0.021*"man" + 0.021*"guarantee"'),
 (14,
  '0.198*"make" + 0.060*"voluntary" + 0.049*"long" + 0.033*"custody" + '
  '0.030*"admissible" + 0.030*"absence" + 0.027*"establish" + '
  '0.027*"voluntarily" + 0.019*"influence" + 0.019*"basis"'),
 (24,
  '0.093*"fbi" + 0.072*"arrest" + 0.064*"suspect" + 0.061*"counsel" + '
  '0.061*"advise" + 0.040*"interview" + 0.040*"agents" + 0.037*"follow" + '
  '0.029*"escobedo_illinois" + 0.027*"offense"'),
 (6,
  '0.087*"general" + 0.049*"new_york" + 0.038*"haynes_washington" + '
  '0.026*"assistant" + 0.026*"leave" + 0.026*"attorney" + 0.026*"argue_cause" '
  '+ 0.026*"john" + 0.023*"arizona" + 0.018*"silent"'),
 (26,
  '0.252*"court" + 0.058*"years" + 0.049*"judicial" + 0.047*"district" + '
  '0.038*"amendment" + 0.027*"precedents" + 0.027*"sentence" + 0.027*"sixth" + '
  '0.025*"draw" + 0.022*"imprisonment"'),
 (10,
  '0.194*"interrogation" + 0.053*"authorities" + 0.031*"important" + '
  '0.031*"afford" + 0.031*"incommunicado" + 0.025*"judge" + '
  '0.022*"information" + 0.022*"procedures" + 0.019*"agencies" + '
  '0.019*"invoke"'),
 (1,
  '0.093*"effective" + 0.076*"exercise" + 0.059*"persons" + 0.039*"employ" + '
  '0.039*"safeguard" + 0.034*"measure" + 0.034*"silence" + 0.034*"fully" + '
  '0.031*"opportunity" + 0.031*"follow"'),
 (28,
  '0.223*"evidence" + 0.140*"trial" + 0.036*"wigmore" + 0.031*"prosecution" + '
  '0.025*"india" + 0.022*"event" + 0.020*"mcnaughton_rev" + '
  '0.017*"inculpatory" + 0.017*"produce" + 0.014*"cert"'),
 (0,
  '0.143*"compel" + 0.067*"suspect" + 0.064*"witness" + 0.030*"jury" + '
  '0.027*"deny" + 0.021*"remain_silent" + 0.021*"response" + 0.021*"seek" + '
  '0.021*"interrogators" + 0.021*"finally"'),
 (31,
  '0.189*"constitutional" + 0.048*"deal" + 0.045*"history" + '
  '0.033*"inadmissible" + 0.027*"sense" + 0.027*"observe" + 0.024*"issue" + '
  '0.024*"policy" + 0.021*"accord" + 0.015*"examine"'),
```

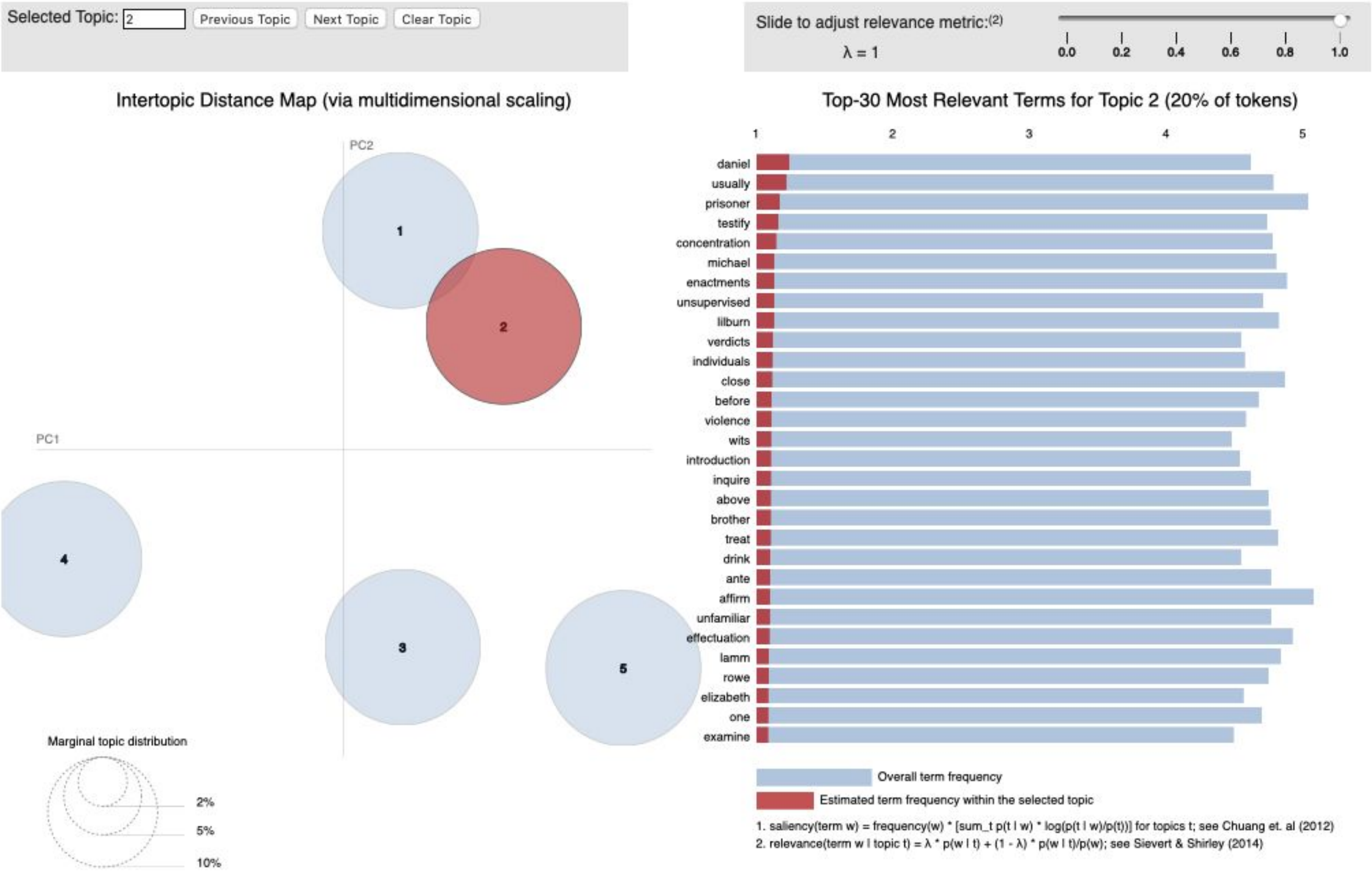
Verifying our topic modeling by using industry insights

Most Common words in MIRANDA V. ARIZONA case

The Court ruled that the ***Fifth Amendment*** to the U.S. Constitution prevents prosecutors from using a person's statements made in response to ***interrogation*** in police **custody** as evidence at their ***trial*** unless they can show that the person was informed of the right to consult with an attorney before and during questioning, and of the ***right*** against ***self-incrimination*** before police questioning, and that the ***defendant*** not only understood these rights, but ***voluntarily waived*** them.



Selected Topic: 1
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribtion
2%
5%
10%
Top-30 Most Relevant Terms for Topic 1 (20.1% of tokens)
1
2
3
4
5
substantiate
skill
bruton
suspicion
exceptional
prohibit
undertake
question
enforcement
hodgson
facto
long
lagay
quote
afterwards
secondly
irregular
discover
contemporary
sense
crim
mild
decisions
restrictive
involve
forthcoming
accomplice
counsel
factfinding
chi
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)



Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (20% of tokens)
1
2
3
4
5
daniel
usually
prisoner
testify
concentration
michael
enactments
unsupervised
lilburn
verdicts
individuals
close
before
violence
wits
introduction
inquire
above
brother
treat
drink
ante
affirm
unfamiliar
effectuation
lamm
rowe
elizabeth
one
examine
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

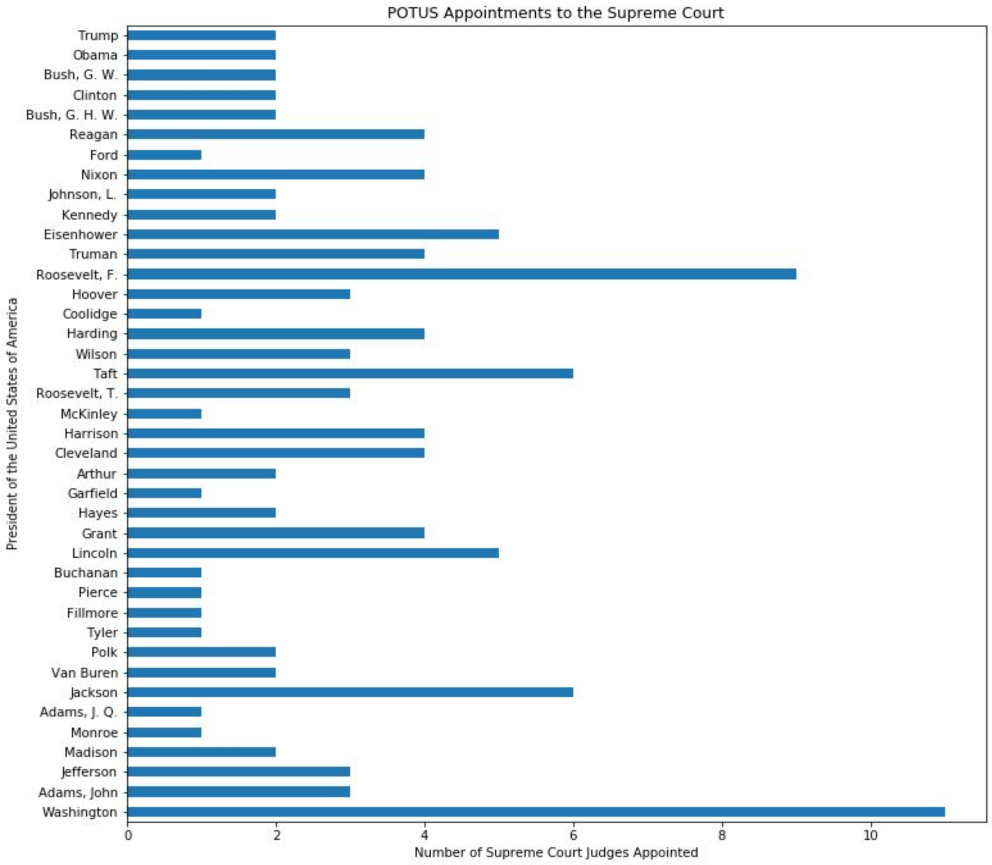
```
[(0,
  '0.038*"warn" + 0.027*"defendant" + 0.025*"attorney" + 0.022*"time" + '
  '0.018*"officer" + 0.013*"general" + 0.012*"california" + 0.011*"new_york" + '
  '0.010*"show" + 0.010*"consult"'),
 (1,
  '0.114*"court" + 0.033*"criminal" + 0.021*"federal" + 0.018*"crime" + '
  '0.016*"justice" + 0.016*"law_enforcement" + 0.013*"opinion" + '
  '0.011*"witness" + 0.009*"voluntary" + 0.009*"judicial"'),
 (2,
  '0.052*"counsel" + 0.033*"accuse" + 0.031*"evidence" + 0.025*"person" + '
  '0.024*"law" + 0.022*"arrest" + 0.016*"require" + 0.016*"practice" + '
  '0.015*"waiver" + 0.014*"rev"'),
 (3,
  '0.046*"question" + 0.024*"make" + 0.021*"present" + 0.020*"interrogation" + '
  '0.018*"suspect" + 0.018*"hold" + 0.017*"today" + 0.016*"fact" + '
  '0.016*"decision" + 0.016*"subject"'),
 (4,
  '0.058*"interrogation" + 0.044*"privilege" + 0.027*"individual" + '
  '0.025*"constitutional" + 0.023*"fifth_amendment" + 0.018*"obtain" + '
  '0.018*"compel" + 0.013*"selfincrimination" + 0.013*"effective" + '
  '0.013*"custody"')]
```

Conclusions:

- LDA gives a coherent output of hidden topics in collections of documents
- You can find insights about the semantics of those documents
- Being equipped with industry insights gives you an advantage to find the optimal topics

Congress Decides

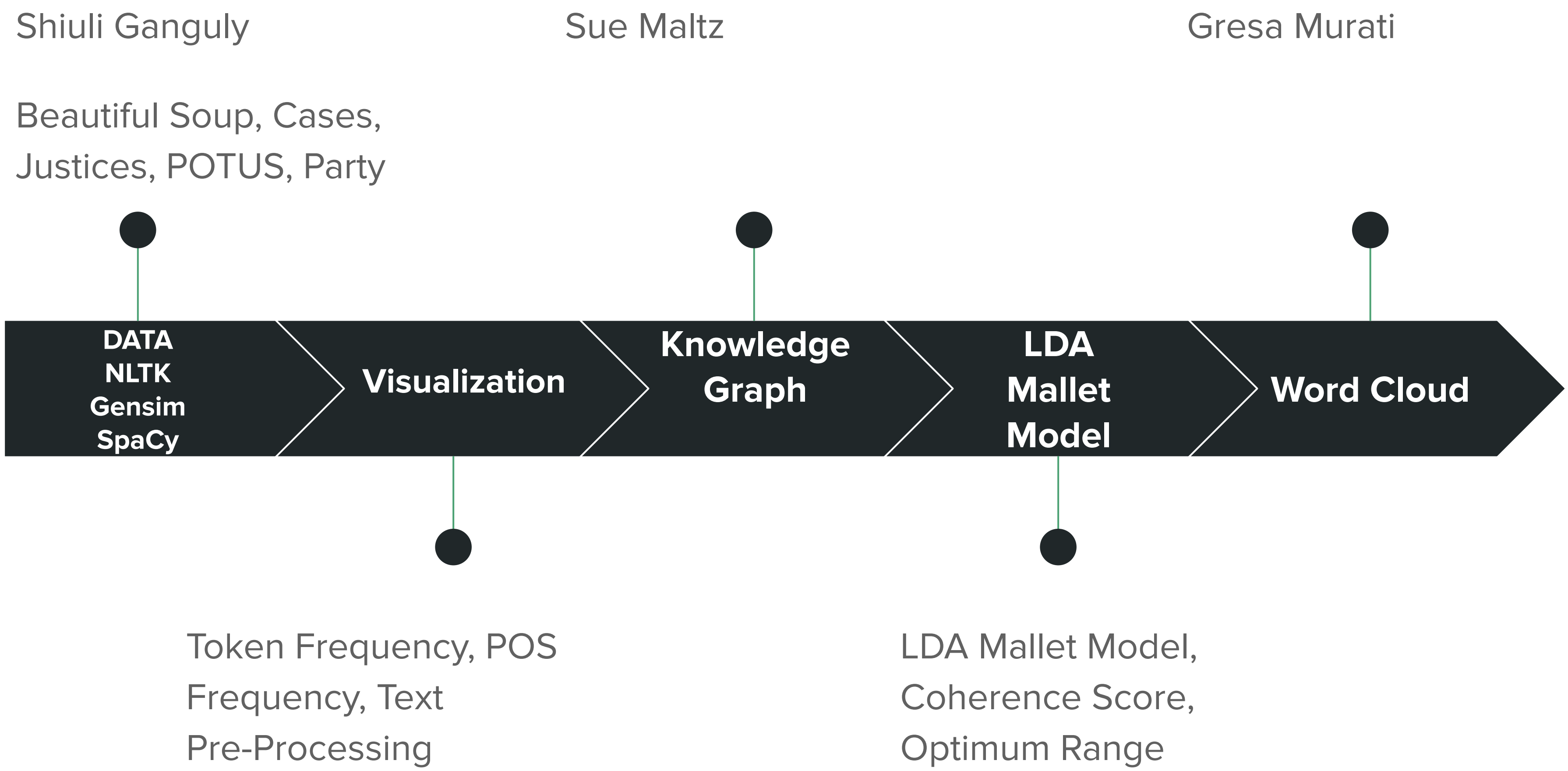
- 1789 (6)
- 1807 (increased to 7)
- 1837 (increased to 9)
- 1863 (increased to 10)
- 1866 (Reduced to 7) Prevented Andrew Jackson from Nominating anyone to the Supreme Court.
- 1869 (Increased to 9)
- 1937 (+1 for every over 70)*



Natural Language Processing

Building Blocks of The Human Language

- Phonemes : cats, bats
- Morphemes : cat
- Lexemes : un-**break**-able
- Syntax : rules


Shiuli Ganguly
Sue Maltz
Gresa Murati
Beautiful Soup, Cases,
Justices, POTUS, Party
DATA
NLTK
Gensim
SpaCy
Visualization
Knowledge
Graph
LDA
Mallet
Model
Word Cloud
Token Frequency, POS
Frequency, Text
Pre-Processing
LDA Mallet Model,
Coherence Score,
Optimum Range