TOPIC MODELING IN THE HOUSE

Introduction

By the time you finish reading this sentence, over 1MB of data, per every person on earth, will be added to the world. That is over 7.5 billion MB of data added every second. That is petabytes more than the data being added two decades ago. However, very little about our lives have actually changed -- we still have only one brain and only two eyes. Our neurons can only fire so fast. The data is accumulating, but our ability to process it is staying roughly stagnant. Yes, headlines are shorter, articles contain more visualizations, microblogging is macro-popular and 15 second TikToks are teaching us how to cook, clean, dance and date, but the sheer volume of data is overwhelming.

Information inundation is equally wonderful and worrisome. When there is so much information, how do we know what to focus on? How do we know what should capture (and keep) our limited attention span? How do we know where to spend our still very limited minutes on earth? We have the same number of minutes in a day as our grandparents, and 100 times more information to process. How can we begin to break this

down into “things that deserve my attention” and “things that don’t”. What is “necessary” and “unnecessary”? Is one person’s “necessary” another person’s “necessary”?” How can we begin to solve this ever-pressing problem?

Enter **Topic Modeling.** The broadest and most basic definition of topic modeling is “fancy word counting.” Documents are collected, words are counted, some variety of math is performed and out come topics. Again, this is very basic. Pretty quickly, though, we have to leave the world of “basic” and ask ourselves a series of questions to better understand this collection of documents -- For example, what is a topic? How do we determine a topic from a document, a paragraph, a website or a book? Do topics belong to these documents? Do the topics belong to the words within those documents? Do documents belong to words? Do words belong do documents? Do topics belong to documents? Do topics belong to words? Do a certain number of words make up a topic? Do documents that contain these words automatically deserve to be in this topic? These are all questions that need to be answered, or better said, approximated.

Analysis & Models

ABOUT THE DATA

The data set is the text collection from the House floor debate of the 110th Congress. The data started out in four subfolders, divided first by political party and then by gender. The researchers merged these subsets together for the final model. The data was processed by removing stopwords, eventually adding to this stopword list. CountVectorizer was used to lowercase and vectorize the data, regex was used for additional tokenizing and the LDA model was applied to this data.

After a wild, wide-eyed moment, the researchers realized they were looking at the data all wrong -- the data was documents all smushed together **by person.** This would be useful if we were attempting to determine the topics **per person**, as in which topics did individual representatives talk about more/less. However, this was not our end goal. This required a hasty rewriting of the preprocessing. The final tests and graphs were done with this secondary processing.

ANALYSIS

A huge part of Topic Modeling is the end visualization. This is how resesrchers can communicate their findings to the stakeholders who ultimately make the business decisions to keep funding [insert dream project here]. Because these researchers aren't James Cameron and have yet to find a way (making movies) to fund our true passions (diving to the bottom of the ocean), they have to rely on these stake holders and these stakeholders rely on visualizations, ergo, the researchers rely on these visualizations.

The only issue is that the only Topic Modeling Visualization package readily available for use within python is actually a pythonic wrapper for an R package that converts R data into D3 for use in browsers. This means that for every topic the researchers would like to run, their python code gets translated into R code and that R code gets translated into JavaScript. This is not the fastest process for the researcher's laptop(s) and caused a considerable sticking point in the middle of the project.

One researcher decided it was time to put her JavaScript knowledge to the test and really unpack what was going on -- Could she recreate the Intertopic Distance Map using her own D3 knowledge? Would she even need to go this far or did this already exist somewhere in the package? To begin, this intrepid researcher cloned the original R package and began to explore the code behind the magic.

She first discovered that the majority of the magic was contained within the ldavis.js file. So she looked to see which other files called the main function within ldavis.js. This function, simply named LDAvis, was called in a couple of html files with two different parameters -- the first param was simply an ID to help with the visualizing after the magic happened. The second parameter was her second clue! The second parameter was a json file -- lda.json. Now where did this file come from? The researcher repeated her steps before but couldn't find anything that explicitly output 'lda.json' -- she only found multiple instances where it was called. She looked at one of the places where it was being called and found that the lda.json file and the file that called it were all in the same directory.

See below for a visual representation of all that's happening.

	**OVERALL INPUT:** **Unformatted Data** **OVERALL OUTPUT:** **D3 Visualization**	**Actual flow:** ***in pseudo code*** **Data = (Unformatted) data** **(this is in R)** **Json = CreateJSON(Data)** **(this is in JavaScript, D3)** **Vis = LDAvis(Json)** **HTML DISPLAYS THE VIS!**

Jeopardy.html	**createJSON()**	**index.html**	**ldavis.js**
Calls **createJson(data)**	Returns: **lda.json** Json of data formatted for future functions	Calls: LDAvis (via ldavis.js) With: **lda.json**	Returns: **Vis** used in index.html

Our intrepid researcher realized that if she could look into createJSON and unpack the output of that function, she, too could create her own lda.json and skip the whole laborious part of going from python to R to JavaScript and back again!

Here is the input/output of that file:

INPUT	`json <- with(Jeopardy, createJSON(phi, theta, doc.length, vocab, term.frequency))`

OUTPUT	`RJSONIO::toJSON(list(mdsDat = mds.df, tinfo = tinfo,`
	`token.table = token.table, R = R,`
	`lambda.step = lambda.step,`
	`plot.opts = plot.opts,`
	`topic.order = o))`

OK but what is the phi, theta, doc.length etc?

After more hunting, our fearless researcher found this:

The first two elements are Φ and θ
Both of these are matrices whose rows must sum to one, since their rows contain probability distributions over terms and topics, respectively.

List containing the number of tokens in each document
List of the unique terms in the vocab (in the same order as the columns of Φ
List of the frequencies of the terms in the vocabulary

It was at this point that the fearless, intrepid researcher realized that this was just one giant distraction from actually writing the paper (thing that is actually due). So, unfortunately, this is where the researcher must stop on her quest to better understand the TOPIC_TERM matrix (phi) and the DOCUMENT_TOPIC matrix (theta). This is hopefully where she will pick up where she left off when she discovers how to create infinite time (but only for herself).

Results

These are the results from running the basic LDA code from scikit learn.

TEST 1: STARTING POINT

VECTORIZATION PARAMS: None

```
Topic 0:
[('the', 785.0256064179524), ('to', 645.6508076986827), ('of', 563.3439173262947), ('and',
542.7164636391508), ('that', 445.9161141427595), ('we', 267.91064217185743), ('in', 259.49073846656427),
('is', 246.60000754980436), ('this', 224.9582831133843), ('it', 190.17402853199783)]
Topic 1:
[('and', 209.6750086441383), ('the', 192.9853786586076), ('of', 170.17785974470274), ('to',
169.70137718746872), ('that', 144.5421140279574), ('in', 113.54346252014587), ('we', 83.96767998571751),
('for', 75.05958453892264), ('have', 73.0637394952233), ('this', 64.931179971705)]
Topic 2:
[('the', 350.10927536238313), ('to', 197.89358554078555), ('of', 146.67751590105448), ('in',
145.35814968113394), ('that', 125.02685414884509), ('and', 124.73104608809528), ('for',
87.16774469308824), ('this', 70.90221338618404), ('have', 68.46045981739343), ('are',
62.314452707063005)]
Topic 3:
[('the', 71961.31057183248), ('to', 44705.23814674544), ('of', 41455.018971298385), ('and',
39039.12153767716), ('that', 26890.762686276157), ('in', 24309.698791032286), ('we',
17147.456346668612), ('is', 16298.293847996309), ('this', 15457.141160161287), ('for',
13903.32208440825)]
```

NOTES:

Time to remove stopwords.

TEST 2: Removing Stopwords

VECTORIZATION PARAMS: Stopwords removed

```
Topic 0:
[('mr', 38.95977122241265), ('doc', 35.461180571443926), ('text', 35.192176634472226), ('docno',
27.38678249970673), ('speaker', 24.56293789783692), ('people', 18.966553005777442), ('house',
17.312176791748907), ('2007', 16.77449332650963), ('act', 15.612877215046574), ('time',
14.956023834203027)]
Topic 1:
[('mr', 128.38918439043834), ('docno', 83.85212846210575), ('doc', 68.02777280655773), ('text',
60.32011779947223), ('speaker', 50.849524221967066), ('house', 43.328654574876104), ('people',
40.864006714950264), ('representatives', 39.593509718240405), ('time', 36.98034955462443), ('2007',
31.85679823416929)]
Topic 2:
[('mr', 8125.642201754612), ('text', 6081.862273503348), ('doc', 6050.404128597955), ('docno',
6018.901790285164), ('house', 5339.693574140009), ('speaker', 4383.636468798018), ('people',
3665.3423830062825), ('representatives', 3324.6597456914546), ('time', 3248.646512778717), ('2007',
2947.958354987127)]
Topic 3:
[('mr', 117.81695194427587), ('house', 79.63197739057671), ('docno', 75.48274271004423), ('doc',
73.43090915754209), ('text', 63.611203640556376), ('speaker', 57.28250593468585), ('congress',
46.10025223808789), ('people', 45.438024956264975), ('representatives', 41.57706387515897), ('support',
39.78497806540509)]
```

NOTES:

Time to add additional stopwords

TEST 3: Adding to our Stopwords List

VECTORIZATION PARAMS: Stopwords removed, additional words removed

```
Topic 0:
[('doc', 102.84017881670427), ('people', 85.7726899202049), ('want', 74.49262514382508), ('going',
67.31439103661329), ('2007', 67.08573067054441), ('american', 58.89400600769927), ('think',
57.54101110269978), ('time', 56.762453785051264), ('today', 56.12320910936942), ('democrats',
53.09355258238547)]
Topic 1:
[('doc', 35.824042642400094), ('people', 31.092504270255617), ('2007', 26.989078751173572), ('going',
25.368688944958087), ('american', 22.206154769434434), ('time', 21.169358850356698), ('support',
19.628690495992913), ('act', 19.402222405938247), ('just', 19.255972932373407), ('know',
18.000960933703837)]
Topic 2:
[('doc', 49.18165626356072), ('time', 36.34333736609367), ('people', 32.95190753213756), ('act',
31.91442536357496), ('know', 28.942383779771742), ('today', 27.639313102464403), ('american',
26.690411396806347), ('think', 25.730719738938866), ('2007', 25.175119641175495), ('support',
24.748533308031238)]
Topic 3:
[('doc', 6039.48180392846), ('people', 3620.8475511114343), ('time', 3223.185187270864), ('2007',
2912.0472201031907), ('act', 2795.2776117549793), ('going', 2761.1539145507486), ('american',
```

```
2755.564252339808), ('support', 2467.6987076994433), ('want', 2406.7741553379014), ('today',
2344.7520738000485)]
```

NOTES:

Still not seeing topics. Need to add even more stopwords.

TEST 4: Adding to our Stopwords List, Again

VECTORIZATION PARAMS: Stopwords removed, additional words removed

```
from sklearn.feature_extraction import text
additional_stopwords = ['mr', 'docno', 'house',
                        'speaker', 'text', 'congress', 'representatives',
                        'doc', 'time', 'want', 'today', '2007']
stop_words = text.ENGLISH_STOP_WORDS.union(additional_stopwords)

cv = CountVectorizer(stop_words=stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 4, cv)
```

```
Topic 0:
[('people', 3574.0098785086657), ('going', 2739.989587363846), ('american', 2673.3518173988336), ('act',
2618.145002233179), ('support', 2304.7299324003657), ('think', 2216.1489206816073), ('know',
2139.8417192883767), ('just', 2107.6026567017943), ('ms', 1988.4709208332206), ('president',
1983.7481564511195)]
Topic 1:
[('people', 48.837462253134454), ('going', 45.258909154612674), ('american', 37.44188880663209),
('just', 37.20963726495393), ('president', 33.513955585061744), ('act', 31.516219081092054), ('support',
28.1821694817231), ('know', 26.945353212378016), ('make', 24.555779872148328), ('think',
23.567999629257407)]
Topic 2:
[('energy', 286.3685702395726), ('act', 195.04208747826797), ('support', 175.4105727962995),
('nebraska', 162.44500783497375), ('national', 157.77286701216352), ('water', 133.65643102232963),
('new', 131.18457538707742), ('chairman', 127.10508864474131), ('texas', 126.15211767498214),
('research', 122.41941298100046)]
Topic 3:
[('support', 52.0569496256697), ('people', 47.313250670705415), ('american', 37.5521619256773), ('act',
37.114714857338), ('energy', 35.17280349231756), ('president', 31.439197643985697), ('new',
30.976080551633334), ('children', 30.037443249840443), ('think', 29.904548191194262), ('years',
29.00193722135195)]
```

NOTES:

Finally starting to see things that look like topics -- “energy”, “children”, “water”

TEST 5: Adding ngrams, removing even more stopwords

VECTORIZATION PARAMS: Stopwords removed, Ngrams added

```
from sklearn.feature_extraction import text
additional_stopwords = ['mr', 'docno', 'house',
                        'speaker', 'text', 'congress', 'representatives',
                        'doc', 'time', 'want', 'today', '2007', 'support', 'american',
                        'president', 'ms', 'mrs', 'going', 'think', 'just', 'know',
                        'make', 'people']
stop_words = text.ENGLISH_STOP_WORDS.union(additional_stopwords)

cv = CountVectorizer(ngram_range = (1,2), stop_words = stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 4, cv)
```

```
Topic 0:
[('act', 34.969804919444385), ('year', 26.135656545118305), ('thank', 23.062154855844906), ('2008',
21.978646648921558), ('new', 21.219386443943687), ('energy', 21.100566912438225), ('legislation',
20.87231503504356), ('children', 19.53688608911742), ('iraq', 19.482997534547945), ('important',
19.179211385730927)]
Topic 1:
[('act', 2762.0956842290097), ('important', 1955.2818093723074), ('country', 1915.1715753143494),
('new', 1866.4334561194892), ('2008', 1831.8577216477213), ('energy', 1807.8508090165776), ('iraq',
1787.8601064250947), ('need', 1785.7779078115782), ('year', 1779.0990604599117), ('legislation',
1735.799061137403)]
Topic 2:
[('act', 46.84340344693324), ('chairman', 36.942145235747944), ('health', 34.290439312935085),
('energy', 31.46686434566559), ('need', 30.959811624963685), ('country', 30.578121135208466), ('2008',
30.23676182058052), ('new', 29.35481225202963), ('important', 28.848632521775716), ('states',
28.774452282030154)]
Topic 3:
[('act', 37.896466410770195), ('important', 28.22514323590759), ('country', 28.030283807416126),
('iraq', 27.50894736587276), ('work', 26.021860866178336), ('new', 25.33982671486711), ('years',
24.620489594630246), ('children', 23.79324203866581), ('like', 23.492209447852574), ('members',
23.272752431636665)]
```

NOTES:

No bigrams seen. Will additional information come from forcing bigrams?

TEST 6: Forcing bigrams with original stopword list

VECTORIZATION PARAMS: Stopwords removed, Ngrams forced

```
cv = CountVectorizer(ngram_range = (2,2), stop_words = "english")
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 4, cv)
```

```
Topic 0:
[('mr speaker', 60.090122389847295), ('house representatives', 58.4674945677937), ('docno text',
50.84872389042602), ('text doc', 47.67863730666059), ('doc doc', 44.2330514998372), ('doc docno',
42.996473659282586), ('2007 docno', 28.435490884868514), ('mrs blackburn', 24.581769511981566), ('2008
```

```
docno', 23.251668296881206), ('text mr', 20.457613524926153)]
Topic 1:
[('house representatives', 145.77660848227626), ('doc docno', 139.11829880510373), ('doc doc',
138.58305315669585), ('mr speaker', 135.16150195902048), ('docno text', 122.73530688359799), ('text
doc', 122.38243688192), ('mr sali', 99.43324268710104), ('2007 docno', 87.72145955169452), ('text mr',
70.12707180604654), ('docno mr', 68.51839865406322)]
Topic 2:
[('house representatives', 1796.439606515149), ('mr speaker', 1736.904488827267), ('doc docno',
1693.124067777938), ('text doc', 1686.4648572425406), ('docno text', 1677.8377179185438), ('doc doc',
1657.6294133205213), ('2007 docno', 1047.0446681652443), ('united states', 702.7845295567759), ('docno
ms', 638.0080603943037), ('text ms', 634.000546181314)]
Topic 3:
[('house representatives', 1358.4655093461581), ('mr speaker', 1276.7403767844087), ('docno text',
1252.1971874461165), ('text doc', 1247.0387918466865), ('doc doc', 1231.0443629456659), ('doc docno',
1228.3676735958907), ('text mr', 925.5129660187531), ('docno mr', 924.582194421202), ('2007 docno',
854.630535587614), ('american people', 620.7018629447323)]
```

NOTES:

Stopwords are clearly necessary.

TEST 7: Forcing bigrams with updated stopword list

VECTORIZATION PARAMS: Stopwords removed, Ngrams forced

```
from sklearn.feature_extraction import text
additional_stopwords = ['mr', 'docno', 'house',
                        'speaker', 'text', 'congress', 'representatives',
                        'doc', 'time', 'want', 'today', '2007', 'support', 'american',
                        'president', 'ms', 'mrs', 'going', 'think', 'just', 'know',
                        'make', 'people']
stop_words = text.ENGLISH_STOP_WORDS.union(additional_stopwords)

cv = CountVectorizer(ngram_range = (2,2), stop_words = stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 4, cv)
```

```
Topic 0:
[('united states', 683.055817521494), ('health care', 313.9339520494879), ('ros lehtinen',
284.7013314447599), ('act 2008', 268.6612449917461), ('jones ohio', 262.54217353270127), ('urge
colleagues', 226.6973150572445), ('yield consume', 213.94826690588957), ('men women',
186.51298900319526), ('reserve balance', 165.02050184967732), ('federal government',
160.90252519113545)]
Topic 1:
[('meek florida', 443.56753013297225), ('working group', 304.63703032616127), ('30 working',
297.47294749289836), ('health care', 291.93393204885683), ('united states', 261.3267881047336), ('men
women', 199.56341527159637), ('florida 30', 188.0260176992834), ('come floor', 151.1659476256355), ('new
direction', 136.49307066239703), ('making sure', 129.25142845029842)]
Topic 2:
[('united states', 163.30300517072715), ('health care', 149.9019912602803), ('thank gentleman',
147.87095329800187), ('urge colleagues', 121.01312355145309), ('like thank', 116.31827281413162), ('act
2008', 115.40035459026032), ('ranking member', 103.59363227212471), ('men women', 102.88514591504743),
('public housing', 98.7160217143333), ('federal government', 96.61511092471652)]
```

```
Topic 3:
[('united states', 141.15612723551357), ('davis california', 130.7378998779459), ('health care',
97.73717042642885), ('mccarthy california', 93.61007489203621), ('new york', 83.65262841965003), ('urge
colleagues', 76.49387056349194), ('men women', 75.17345300161934), ('act 2008', 73.6389191860809),
('thank gentleman', 48.94595821150284), ('reserve balance', 45.9051712021919)]
```

NOTES:

More topics are bubbling up -- “health care”, “reserve balance”

TEST 8: Expanding the number of topics from 4 to 10

VECTORIZATION PARAMS: Stopwords removed, Ngrams forced, topics = 10

```
cv = CountVectorizer(ngram_range = (1,2), stop_words = stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 10, cv)
```

```
Topic 0:
[('like', 8.99291291281483), ('country', 8.730439967309731), ('important', 8.667189783472503),
('children', 8.653958188862616), ('act', 8.512072765191212), ('year', 8.45043673213108), ('new',
8.141894461751193), ('energy', 8.105871053214859), ('way', 7.814583777027177), ('legislation',
7.788090810047982)]
Topic 1:
[('act', 2252.1593067558947), ('chairman', 1400.9581291980394), ('2008', 1335.3098627433324),
('legislation', 1312.4502982800914), ('year', 1282.2280733728733), ('new', 1265.2727846809694),
('energy', 1254.0301462096863), ('years', 1222.4175021013857), ('country', 1219.7436275710188),
('health', 1206.162804518337)]
Topic 2:
[('foxx', 516.8856988028886), ('democrats', 480.18444918930066), ('energy', 453.7777614138316),
('country', 417.7658043054481), ('act', 405.9356488379371), ('2008', 392.25793674036555), ('year',
360.97284427233546), ('new', 353.69091102752697), ('need', 350.168221589533), ('states',
313.96547917162377)]
Topic 3:
[('members', 751.625660948011), ('important', 639.6568541621323), ('say', 588.7395552438567), ('floor',
530.886817911088), ('florida', 526.849453194008), ('sure', 457.74166329109516), ('iraq',
455.86797228444743), ('meek', 448.56951721978766), ('meek florida', 438.8335695075969), ('republican',
424.8276378472788)]
Topic 4:
[('act', 15.737751283995939), ('years', 12.417490292265999), ('new', 12.010049541276496), ('country',
11.914576050339415), ('like', 11.869221918639232), ('work', 11.27221267238566), ('states',
10.621669151412371), ('important', 10.54066533148507), ('members', 10.53183431323579), ('year',
10.438593841455598)]
Topic 5:
[('act', 13.057140010730777), ('work', 12.705984667024572), ('need', 12.556365689552386), ('energy',
10.983158913580253), ('state', 10.176055372386692), ('year', 9.830731374593801), ('iraq',
9.79388885320985), ('health', 9.543173397719405), ('like', 9.403684584398407), ('children',
9.362578871942356)]
Topic 6:
[('act', 19.171098966447772), ('2008', 15.013404702037228), ('years', 14.062870119963213), ('country',
12.471296259970947), ('energy', 11.822472194354203), ('important', 11.533300954326446), ('new',
11.356711610684146), ('states', 10.98507958663112), ('chairman', 10.890272637301207), ('legislation',
10.779455205435607)]
```

```
Topic 7:
[('act', 13.595768051586218), ('legislation', 11.529476561677367), ('year', 11.511694277161785),
('2008', 11.467244365815048), ('thank', 11.10782694843333), ('important', 10.544812871895362), ('say',
10.070660612573448), ('energy', 10.062345361217297), ('new', 9.689224671192322), ('country',
9.639606026324595)]
Topic 8:
[('act', 9.230827994435014), ('new', 8.958313698719257), ('important', 7.860210024255094), ('country',
7.671501333938701), ('chairman', 7.537538865264406), ('need', 7.315253894206442), ('like',
6.769046302490408), ('year', 6.390871632595374), ('way', 6.269511445653796), ('colleagues',
6.161683677727667)]
Topic 9:
[('act', 14.16980969114261), ('thank', 13.4029971820924), ('energy', 13.38056785337938), ('need',
12.53118637238359), ('2008', 12.02376555145941), ('states', 11.335969593578705), ('work',
11.260430641182198), ('years', 10.948293115235245), ('new', 10.928365102684586), ('important',
10.607384192351462)]
```

NOTES:

Need to add more stopwords -- potentially it's time to add POS tagging?

TEST 9: Removing even more stopwords

VECTORIZATION PARAMS: Stopwords removed, Ngrams forced, topics = 10

```
cv = CountVectorizer(ngram_range = (1,2), stop_words = stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 10, cv)
```

```
Topic 0:
[('energy', 26.45203979878982), ('national', 19.174487491887685), ('tax', 17.680044282958956),
('united', 17.009604297571887), ('years', 15.869550475696053), ('iraq', 15.692976371452394), ('america',
15.436655699227973), ('nebraska', 14.232064094110019), ('gentleman', 13.843279330108452), ('health',
13.41593935409385)]
Topic 1:
[('years', 10.384290912644355), ('care', 9.503106775301964), ('way', 7.644332029395534), ('america',
7.602365776946815), ('iraq', 7.592166906926408), ('said', 7.525794800235866), ('children',
7.520714851016946), ('percent', 7.488249834183103), ('colleagues', 7.400238394907784), ('families',
7.234841027023669)]
Topic 2:
[('united', 13.573242083063887), ('years', 12.27169326037616), ('colleagues', 12.089443010387491),
('iraq', 11.83322909266544), ('energy', 11.456752784305028), ('national', 10.997317849570354), ('tax',
10.40908407736109), ('resolution', 10.350847332288607), ('percent', 9.794071861281383), ('federal',
9.494898032463931)]
Topic 3:
[('energy', 11.504736022281454), ('years', 11.169703342813758), ('iraq', 10.316873503984128),
('children', 9.258717783349924), ('tax', 9.212507197333736), ('national', 8.400996714583723), ('health',
7.926660319871598), ('way', 7.7858511874107155), ('federal', 7.658001678503496), ('good',
7.259085476296372)]
Topic 4:
[('children', 10.667263549889466), ('national', 8.262420316789648), ('united', 7.45256971353309),
('years', 7.271225717355232), ('iraq', 7.250910793159072), ('vote', 7.245170588074981), ('energy',
7.007849355279482), ('health', 6.978489989401828), ('nation', 6.9263034714472), ('leadership',
6.745830853745591)]
```

```
Topic 5:
[('years', 21.747108696881014), ('energy', 20.245774643916654), ('blumenauer', 19.068841865984147),
('health', 18.66812671076742), ('iraq', 17.231323912875308), ('care', 17.18147415606492), ('colleagues',
14.993069441894251), ('children', 14.701828451728037), ('able', 14.613871031042247), ('working',
14.469949111263135)]
Topic 6:
[('energy', 11.166818593388882), ('years', 9.867236505299081), ('health', 9.843439302225852), ('iraq',
9.65216976455357), ('care', 9.08690427450303), ('united', 8.69406693641062), ('children',
8.53034636682573), ('tax', 8.005993719026627), ('national', 7.6208662811558945), ('percent',
7.471190998148229)]
Topic 7:
[('energy', 1763.0380968680051), ('iraq', 1758.3739081712854), ('years', 1676.3502425909553), ('health',
1580.2272541674724), ('children', 1482.6933961057653), ('colleagues', 1349.4589566241875), ('tax',
1340.7656926724987), ('united', 1317.3761529318742), ('said', 1298.3273810804014), ('care',
1288.1634889864938)]
Topic 8:
[('energy', 16.592698521091414), ('united', 12.837529337598774), ('iraq', 12.375705960182884),
('health', 11.996652188054123), ('tax', 11.585722268881785), ('colleagues', 11.573755292751674),
('resolution', 11.202374813513934), ('years', 10.944448990822481), ('nation', 10.428831912283577),
('america', 9.844654690207799)]
Topic 9:
[('iraq', 13.714333999285259), ('years', 11.738123905526717), ('children', 10.345699811272844),
('health', 9.377517020320012), ('energy', 9.036022536836267), ('resolution', 8.016060907501412),
('care', 7.878769067173163), ('budget', 7.642686717197339), ('way', 7.393842722058604), ('vote',
6.993927229206308)]
```

NOTES:

Need to add more stopwords -- potentially it's time to add POS tagging?

TEST 10: Using the entire corpus and 40 topics

VECTORIZATION PARAMS: Stopwords removed, Ngrams used, topics = 40

```
cv = CountVectorizer(ngram_range = (1,2), stop_words = stop_words)
lda_model, lda, lda_vec, cv = run_lda(all_data_sm, 10, cv)
```

```
Topic 0:
[('tsongas', 13.715657118695905), ('whitfield kentucky', 5.782357920326937), ('barrett farm',
5.07083192090729), ('tsongas rise', 4.381167138454782), ('health', 3.0261532079624742), ('years',
2.970111474896366), ('children', 2.9300387713317635), ('iraq', 2.9172148934909816), ('sergeant jimenez',
2.8618232829707124), ('minute man', 2.673874961848008)]
Topic 1:
[('iraq', 2.2894835921450607), ('years', 2.062678856285181), ('energy', 1.9055327739851224), ('program',
1.8622197689747775), ('nation', 1.8008970082863252), ('percent', 1.778994146259842), ('united',
1.7478949629226386), ('colleagues', 1.744229359722898), ('security', 1.703181685323808), ('national',
1.6219923610442637)]
Topic 2:
[('years', 2.40164591440504), ('iraq', 2.0261060711887082), ('national', 1.8142571227245055), ('energy',
1.7644255180718207), ('care', 1.6870240226984834), ('health', 1.6812100948808983), ('children',
1.6602717900927628), ('war', 1.6337031628254644), ('united', 1.5493519120300447), ('program',
1.485070742051341)]
```

```
Topic 3:
[('years', 2.047914014809786), ('health', 1.858571553852879), ('energy', 1.819256075396485), ('war',
1.7315544983332878), ('nation', 1.7025432109558398), ('iraq', 1.6620027989915471), ('percent',
1.6212197822510794), ('federal', 1.5707752754664268), ('national', 1.5367879823304544), ('children',
1.530079657108243)]
Topic 4:
[('energy', 1.9895836311290682), ('texas', 1.9186492210345925), ('america', 1.8687004150493067),
('federal', 1.8455637169599026), ('million', 1.8275924346561345), ('children', 1.7574511017966115),
('iraq', 1.6950513265414429), ('war', 1.6937429130420976), ('nation', 1.6913009954617253), ('health',
1.6681765263648518)]
Topic 5:
[('years', 2.070217099030579), ('health', 1.9511411787018735), ('iraq', 1.8605985683208102), ('energy',
1.5501645233503523), ('national', 1.539492246315864), ('war', 1.5306557479986305), ('colleagues',
1.5134883785882822), ('children', 1.46246005779672), ('families', 1.4539408795317263), ('america',
1.4362808255864123)]
Topic 6:
[('trona', 8.052185731589802), ('mcgee', 4.842060294694177), ('gale mcgee', 3.798751363749579), ('iraq',
2.247297907133224), ('nation', 2.2281058465804313), ('laramie', 2.2123091136243773), ('trona air',
2.202913913298052), ('6901', 2.0369594586566526), ('colleagues', 1.8996599469809434), ('security',
1.80040482136311)]
Topic 7:
[('years', 1.7392149546089974), ('iraq', 1.7214467569113392), ('children', 1.4959141003171634),
('energy', 1.4319334993958277), ('million', 1.3407484863141055), ('health', 1.3017598991331925),
('nation', 1.2803737073871984), ('veterans', 1.2752393673728575), ('amendment', 1.2738740373027782),
('national', 1.2689000178781162)]
Topic 8:
[('health', 3.841402978196184), ('years', 2.238874347403486), ('children', 2.113250876269056),
('nation', 2.0881101030278346), ('national', 1.7603330544985514), ('care', 1.6920637059252577), ('iraq',
1.6563316274753053), ('troops', 1.615196852579845), ('security', 1.6123640342077625), ('million',
1.5957972967819865)]
Topic 9:
[('iraq', 2.801316891983262), ('energy', 2.655821061535527), ('children', 2.4763329099558233), ('war',
2.3897708534837987), ('meek florida', 2.2633996533020015), ('years', 2.1618002330883983), ('health',
2.1367007602167654), ('america', 2.032178246884775), ('united', 1.8074338136804478), ('nation',
1.765148808227784)]
Topic 10:
[('iraq', 3.7764065564892295), ('trojans', 3.319577699605524), ('rcwd', 2.8139996190110157), ('years',
2.770871091850797), ('bono rise', 2.7336012716784657), ('united', 2.697736234347712), ('program',
2.4521185167821153), ('amendment', 2.363089727378448), ('health', 2.2501690941276946), ('million',
2.2373967921193114)]
Topic 11:
[('iraq', 2.8119359660587384), ('energy', 1.980604407300287), ('national', 1.7643793937628982), ('care',
1.709307849836053), ('colleagues', 1.6821479932608914), ('children', 1.674619777487679), ('security',
1.6613849863790002), ('united', 1.6086394802355901), ('years', 1.5322122090680437), ('health',
1.4293334827277546)]
Topic 12:
[('iraq', 2.3739987649918306), ('children', 1.5691793893869965), ('national', 1.5587832552024143),
('energy', 1.5393125857860783), ('colleagues', 1.5311276117855634), ('united', 1.5271535783463235),
('years', 1.5260766290682746), ('nation', 1.4626343348829032), ('health', 1.4188780350231138),
('veterans', 1.2755985290150176)]
Topic 13:
[('health', 2.7669515844342256), ('iraq', 2.702934395552306), ('years', 2.304371103045542),
('colleagues', 2.299760340632374), ('program', 2.153565931727933), ('united', 2.098739184012054),
('nation', 1.9593702462308757), ('help', 1.955072996834885), ('energy', 1.9084997330395634), ('war',
1.8753185089314817)]
Topic 14:
```

```
[('years', 2.845380452551978), ('children', 2.794941552872071), ('million', 2.351901240758951),
('america', 2.3463898186439853), ('national', 2.30134519953638), ('health', 2.157511509469691),
('amendment', 2.0477484636388836), ('united', 1.957628110884377), ('nation', 1.91902965578512), ('iraq',
1.918246969384595)]
Topic 15:
[('iraq', 2.339005501608942), ('energy', 1.7283745581161976), ('years', 1.635585680708884), ('united',
1.4965042819064234), ('america', 1.4903213612477415), ('health', 1.472533725321335), ('national',
1.4096686310562838), ('amendment', 1.4051078441192397), ('nation', 1.362985378350064), ('war',
1.36094343904717)]
Topic 16:
[('children', 1.702785449897993), ('years', 1.589509010259555), ('iraq', 1.499901414037624),
('security', 1.4856018813193637), ('united', 1.4729158510434532), ('care', 1.4099495571796774),
('nation', 1.3148122797952682), ('amendment', 1.2797397468120868), ('said', 1.2644165328141883),
('energy', 1.2202589626801332)]
Topic 17:
[('iraq', 1.5863240375789556), ('america', 1.5811391200028664), ('waite florida', 1.5262677006527525),
('ratify business', 1.5100256468176931), ('vote ratify', 1.506632031756399), ('bailout foreign',
1.5044796220768608), ('health', 1.4427887420913006), ('class homeowner', 1.4350162053797069), ('care',
1.3745167475073368), ('years', 1.3495611980329232)]
Topic 18:
[('iraq', 2.3591908117870712), ('health', 1.8598284148509117), ('national', 1.7834890183362917),
('united', 1.7643129938833058), ('years', 1.7107962156442118), ('colleagues', 1.6899668125063076),
('energy', 1.5712564651686936), ('nation', 1.5408334906916026), ('way', 1.5014805304671817), ('tax',
1.4856135620169133)]
Topic 19:
[('price north', 153.8303808442232), ('carolina department', 46.202871656521936), ('north carolina',
29.89404336636968), ('14 price', 22.869309296043713), ('myrick', 21.727269000187768), ('larsen
washington', 15.115062272099573), ('carolina', 13.459654938451353), ('north', 13.0175340351649),
('162f', 10.869077328248816), ('folio 162f', 10.713158448831907)]
Topic 20:
[('years', 2.627845640339843), ('health', 2.4952791899345605), ('program', 2.2276457313387614),
('veterans', 2.2256526374568777), ('iraq', 2.1185795216488468), ('nation', 2.0794450741344885),
('energy', 2.0014586718287455), ('amendment', 1.9264848999128696), ('national', 1.8184150806170762),
('children', 1.8052977595170236)]
Topic 21:
[('24 cramer', 2.8660880723786772), ('city huntsville', 2.8399898022451957), ('colleagues',
2.4276538195430213), ('cramer rise', 2.1592115261796576), ('national', 2.0442999109629363), ('years',
2.021245569401965), ('amendment', 1.7335909045844453), ('health', 1.632671039334483), ('america',
1.6022312118083013), ('children', 1.5285190684419974)]
Topic 22:
[('energy', 31374.486533943185), ('years', 28341.54440842587), ('oil', 23955.331502879915), ('united',
21686.89941125787), ('iraq', 21523.62368825468), ('amendment', 20440.414207615435), ('percent',
20377.74755399727), ('said', 19967.33994940072), ('national', 19672.758938301475), ('way',
19500.957436573524)]
Topic 23:
[('iraq', 2.513356794336642), ('energy', 2.3822287442574406), ('nation', 2.1183554497298624), ('united',
2.0786745343600046), ('years', 2.0164290757626397), ('health', 1.9904987292280598), ('day',
1.8438747834434854), ('world', 1.8328112633439557), ('america', 1.6191165262521188), ('americans',
1.6166380999162844)]
Topic 24:
[('health', 2.242243863881511), ('energy', 1.647916753401699), ('united', 1.6214320662678745),
('children', 1.6176834257231905), ('care', 1.5572495698370816), ('nation', 1.535022778458392), ('iraq',
1.4848634489493817), ('years', 1.417520738683851), ('percent', 1.401367665388492), ('amendment',
1.3786995535371835)]
Topic 25:
```

```
[('national', 1.470514532225469), ('iraq', 1.4455033451032182), ('energy', 1.4247525993447179),
('united', 1.422126424874869), ('program', 1.3091682865856646), ('america', 1.2357127399113192),
('federal', 1.1422684513266643), ('resolution', 1.0857471549271447), ('million', 1.080281915263105),
('health', 1.077217478951349)]
Topic 26:
[('iraq', 2.0279272575419482), ('nation', 1.7782055179729255), ('amendment', 1.697702710586682),
('national', 1.6700412745276834), ('war', 1.63654310794989), ('years', 1.606645012800417), ('america',
1.4170306220024096), ('million', 1.343561275306094), ('colleagues', 1.3305248849580429), ('world',
1.3233813077409404)]
Topic 27:
[('schmidt', 46.64319236614091), ('schmidt rise', 15.51892781664844), ('joni', 7.746863874501813),
('hahn', 6.256322094694558), ('gnep', 5.801051411412758), ('schmidt energy', 4.875761428465828), ('title
grantees', 4.6118030435221895), ('maupin', 4.526344945186737), ('patricia corbett', 4.323086749774799),
('galen', 4.238941496542439)]
Topic 28:
[('lee texas', 1119.46522586448), ('jackson lee', 1072.761177194754), ('jackson', 1006.3721950923417),
('lee', 957.1592948250438), ('iraq', 565.7039166065728), ('woolsey', 261.3081025895303), ('kaptur',
259.9271404593028), ('war', 246.76592700535832), ('ros', 244.2167740644297), ('ros lehtinen',
241.33230038531468)]
Topic 29:
[('care', 2.305764798869615), ('nation', 2.2155046618041276), ('children', 2.150761993175409),
('energy', 2.1275979938723415), ('colleagues', 2.116854004770791), ('national', 2.1079843709252937),
('veterans', 2.0758035207074017), ('americans', 1.8571798837384526), ('years', 1.8423530780663364),
('program', 1.8123622786645344)]
Topic 30:
[('years', 2.3706745465233077), ('health', 2.0776336956002495), ('united', 2.077612655581257), ('iraq',
2.0029442874335057), ('nation', 1.8135843383115586), ('security', 1.8082422716478674), ('war',
1.7626671079441059), ('care', 1.671245093221841), ('veterans', 1.6522624931144148), ('national',
1.6276372529293075)]
Topic 31:
[('iraq', 1.9441008738503789), ('health', 1.8985495548419835), ('united', 1.8368113765323077), ('years',
1.6703934713721431), ('national', 1.4457487143298462), ('children', 1.4031366062522646), ('security',
1.367844079244229), ('colleagues', 1.345240797794196), ('amendment', 1.3312876384944943), ('energy',
1.3216587811344676)]
Topic 32:
[('frank massachusetts', 538.1266993651593), ('oberstar', 523.1544308181512), ('inslee',
214.38743019968496), ('massachusetts', 163.23638775858507), ('latourette', 161.2604434923527), ('coast
guard', 135.64838496813164), ('transportation', 135.21740070023316), ('frank', 117.53184294157734),
('gentleman', 91.0716280257499), ('amtrak', 89.0758463264207)]
Topic 33:
[('iraq', 2.21157775765804), ('health', 2.1162458267253643), ('nation', 1.980636965920637), ('program',
1.9773070058119573), ('children', 1.9611262796501536), ('colleagues', 1.8976048153833132), ('united',
1.896176138773581), ('energy', 1.8917430777483486), ('war', 1.8415629550330572), ('security',
1.8045473001109693)]
Topic 34:
[('energy', 2.522380429443432), ('children', 2.407712463731839), ('nation', 1.8769454872556208),
('iraq', 1.8613307710239613), ('health', 1.8495778939042395), ('america', 1.727520415568534),
('national', 1.7207525119887652), ('care', 1.7021290983988555), ('years', 1.6767816722346953),
('amendment', 1.633709834673332)]
Topic 35:
[('years', 2.6444196535742193), ('nation', 2.2946398374291235), ('iraq', 2.1636183271658727), ('energy',
1.8663360779492328), ('united', 1.8526548560424796), ('women', 1.810859797370497), ('percent',
1.7960025223506064), ('health', 1.763413916361885), ('war', 1.6925033318962424), ('care',
1.6050143985735754)]
Topic 36:
[('children', 2.6238721734165598), ('iraq', 2.445895490202885), ('years', 2.382922061790647), ('health',
```

```
2.2918796767734437), ('war', 2.251332420859546), ('colleagues', 1.9577513253787615), ('care',
1.8831501824770216), ('nation', 1.845547283979052), ('america', 1.8247660956083185), ('come',
1.758766342579622)]
Topic 37:
[('south mississippi', 5.095640116112494), ('iraq', 3.4410076917739354), ('years', 3.2767425066226368),
('million', 2.427859152856377), ('children', 2.3112088947744827), ('health', 2.2789479116760067),
('united', 2.252671649469959), ('national', 2.196787172430396), ('security', 2.183669638906338),
('taylor mississippi', 2.1125782450372097)]
Topic 38:
[('iraq', 2.2993673888246207), ('nation', 2.088694867447548), ('years', 2.008439158852034), ('war',
1.891964963894034), ('care', 1.7241405189031314), ('percent', 1.6248277022441668), ('energy',
1.6089619977831118), ('colleagues', 1.5960046316316554), ('america', 1.5906557184089392), ('united',
1.4974616364051545)]
Topic 39:
[('millender mcdonald', 9.827911349920484), ('mcdonald congressional', 5.779332792047013), ('22
millender', 5.576098417042082), ('millender', 5.5307368997446815), ('mcdonald', 5.193404996101821),
('pension forfeiture', 4.383573658236061), ('balance millender', 3.602255845380467), ('iraq',
2.9408951505709275), ('19 millender', 2.8436503408196536), ('08 millender', 2.820447475597192)]
```

NOTES:

Need to add more stopwords -- potentially it's time to add POS tagging?

There is a lot of iteration in topic modeling. The researchers manually did the iterating this time -- generating topics, removing stopwords, generating again, removing new stopwords that bubbled up etc -- however, it would be most effective for the researchers to look into ways of automating such tasks. Are there commonalities among the removed stopwords? Would it be efficient to remove all shared words in the documents before topic modeling?

Conclusion

Topic modeling helped to shed light on the many topics that crossed the house floor of the 110th Congress. Health care was talked about, the iraq war was debated, education was discussed as well as energy, transportation, and safety. Some topics bled into other topics (topics like "American" and "United"), while other topics, such as porn and the internet, were fairly siloed into their own groups. This is useful information for both those in government and those being governed -- are these the topics the people would like their representatives spending time on? How do these topics compare to past topics?

There is still a lot of room for growth in the field of topic modeling. How do we know which words to count? How do we know how to weight these words? Should we simply take the

word that occurs most frequently in a document and call that the topic? Should the word “and” be a topic? Probably not. However, should we disregard the word “and” entirely? No, we shouldn’t. If an online article mentions the word “cat” 19 times, it might be about cats. If a book mentions the word “cat” 19 times, it might be about cats. However, if we learned that the article is only 50 words and the book is actually 500 pages, we’d probably reach different conclusions. The book could be “History of Domesticating Animals” and the article could be “Why Everyone Needs a Cat.” Those extra words like “and” are important to our counting and that’s how we get things like term document frequencies and document term frequencies.

In a world with ever-increasing information overload, things like topic modeling are more necessary than ever. Topic modeling can help distill and direct everything from classroom learning to topics on the congress floor. It can help news websites tag articles for faster discovery and more efficient (and personalized) user experiences. It can help archivers categorize the backlog of analog data. It can help Natural Language Processing researchers better understand social media data. There is almost no limit to the usefulness of topic modeling.

REFERENCES

(1) https://www.socialmediatoday.com/news/how-much-data-is-generated-every-minute-infographic-1/525692/