PLSA TOPIC MODELING

EE 527 PROJECT, JAN-MAY 2022

BUDDI KIRAN CHAITANYA
214161002
buddi.kiran@iitg.ac.in

Topic modeling is the extraction/identification of the latent themes in a corpus (texts/documents), given the observed (word, document) co-occurrence, typically specified in terms of term-document matrix.

Two involved tasks:

1. to discover the top-Z (a user-determined number) number of topics that are contained within the corpus and
2. to figure out to what extent a given document covers any of the extracted topics (a kind of soft membership of the documents in the corpus to the identified k-topics).

Terminology:

Inputs:

-d: { d_1, d_2, ... d_D }, a collection/corpus of D documents

Outputs:

-z: {z_1, z_2, … z_Z}, the set of Z topics

-coverage/membership of the topics for each document : p(z|d), the probability of a topic z being covered, conditioned on the document d.

$$\sum_{z \in topics} p(z|d) = 1$$

How to define a topic (z) :

One natural way, at the outset, to define a topic, is in terms of a word or a phrase. After we have a set of candidates for topics, we devise a scoring function to measure how good each term is as a topic.

Statistical strategies to design scoring function include:

- favor a representative term (high frequency preferable), while also avoiding words that are too frequent, ex: stopwords (the, a)
- TF-IDF weighting from retrieval is a useful tool.
- domain-specific heuristics are possible (e.g., favor title words, hashtags in tweets).

Using any such scoring function, we pick Z terms with highest scores, while at the same time trying to minimize redundancy, i.e., if it could be that multiple of the high scoring topic terms are closely related, semantically, then we pick one representative term from those while ignoring the others. A greedy algorithm like that of maximum marginal relevance ranking can be used to maximize coverage while minimizing redundancy among topics.

Issues with 'Term as Topic':

- Lack of expressive power: the major issue with using a single term/word or even a phrase for defining a topic is that they can only describe simple/general topics but sall short in capturing complex & specialized topics.
- Incompleteness in vocabulary coverage: variations of vocabulary cannot be captured. A single term does not make it clear what other terms/words might be related to the topic. This also results in the issue of ambiguity, as the term itself or the related terms can be ambiguous, and there is no means to resolve the ambiguity.

To address the expressivity issue, the intuitive way out is to use a collection/ensemble of words/terms to describe/define a topic.

The incompleteness in vocabulary coverage can be resolved by adding weights on the terms that make up the topic definition. This not only allows the distinguishing of subtle differences between topics but also allows the introduction of semantically related words in a fuzzy manner.

To address the issue of word ambiguity, we need to split ambiguous words so that we can disintegrate its topic.

It turns out that all these can be achieved by using a probabilistic topic model.

In this paradigm, the topic is defined as a (conditional) probability distribution over the vocabulary set. Each topic is thus defined, not by words/terms but by a word distribution.

What follows from such a definition of topic, is that topics, now, no longer have an explicit label, the label is latent/invisible and is expressed only via the distribution.

If the Vocabulary set is defined as w: { $w_1, w_2, \ldots w_v$ }, and topics set as {$z_1, z_2, \ldots z_Z$}, then, the topic is now defined as p (w | z), where w ∈ Vocabulary Set, and z ∈ Topic Set and

$$\sum_{w \in Vocab\ set} p(\,w \mid z\,) = 1, \qquad \forall z \in topic\ set$$

Vocabulary Set word (w)	z_1 [“Government”] $p(w\|z_1)$	z_2: [“Financial Markets”] $p(w\|z_2)$	z_3: [“Climate Change”] $p(w\|z_3)$	topic z p(w|z) z ∈ topic set
activism	0.0001	0.001	0.03	
assembly	0.02	0.000015	0.001	p(w|z), w ∈ Vocab set
bear	0.00001	0.01	0.01	
bill	0.015	0.005	0.05	
bull	0.000001	0.01	0.000002	
crash	0.0004	0.02	0.00003	
green house	0.005	0.000001	0.05	
growth	0.03	0.015	0.0002	
stock	0.000007	0.04	0.0001	
summit	0.0001	0.000006	0.009	
xenon	0.00000001	0.00000001	0.008	
zeppelin	0.000001	0.0000001	0.00001	
$\sum_{w\in vocab} p(w\|z\ )$	1.0	1.0	1.0	

In probabilistic topic modeling, a distribution over the entire vocabulary is used to define a topic, hence the scope for complex expressivity. And, as weights/probabilities are assigned to the words, it allows for modeling of subtle semantic similarities/differences between topics. This probabilistic weighting also achieves the task of bringing in the related words together. The ambiguity of a word is disintegrated by assigning different probabilities to the same word in different topics.

In this redefined topic paradigm, the inputs & outputs are:

Inputs:

- Corpus d: { d_1, d_2,....d_D } a set of D documents,
- Z (the predetermined number of topics to be mined), &
- a Vocabulary set w: { w_1, w_2, ...w_v }

Outputs:

-Z topics or Z conditional word distributions p(w|z): { z_1, z_2, … z_z } over the vocabulary W, one for each of topic

$$\sum_{w\in Vocab\ ser} p(\,w\,|\,z\,) = 1, \qquad \forall\, z \in topic\ set$$

-D conditional coverage/membership distributions, p(z|d), over the topics, one for each document, with the constraint that

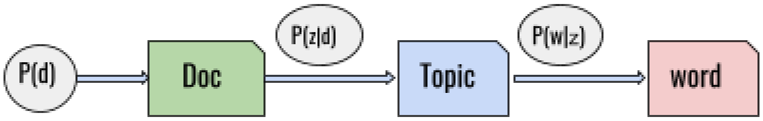
$$\sum_{z\in topic\ Set} p(\,z\,|\,d) = 1$$

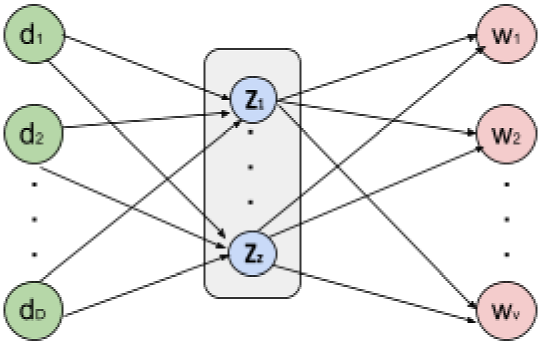
In sum, we need to output Z word-distributions p(w|z) over the V-vocabulary set and D coverage-distributions p(z|d) over the Z-topics. Thus the total number of parameters in this asymmetric generative probabilistic model are:

$$Z\,(V-1) + D\,(Z-1)$$

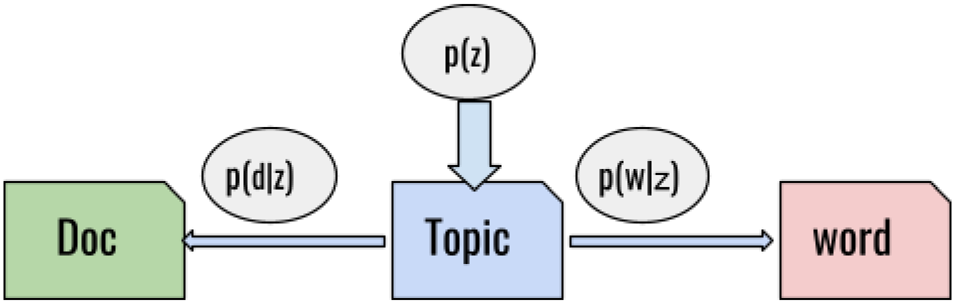
MODEL

The manner in which our corpus/documents could have have been generated is modeled as follows: we sample a document using the distribution p(d), the we look up the topic coverage distribution vector for the sampled document (say d), and sample a topic, say z, as per the distribution p(z|d). Next, we look up the word distribution of the sampled topic (z) and output a word (w) with as per the distribution p(w|z).





An alternative but equivalent symmetric parameterisation of the model is displayed below:



Model Assumptions:

- Bag-of-words: each document is regarded as an unordered collection of words, with the implication that the joint distribution of the observed data will factorize as a product.

$$P(corpus) \ = \ \prod_{(d,w)} p\,(d,\ w)$$

- Conditional independence: words and documents are conditionally independent given the topic.
$P(w, d \mid z) \ = \ P(w|z) \ * \ P(d|z) \ \Leftrightarrow \ P(w|\ d,\ z) \ = \ P(w|\ z).$

Using the above assumptions, in conjunction with product rule, the joint distribution can be expressed as below:

$$p\,(d,\ w) \ = \ p(d)\ p(w|d)$$

$$= \ p(d)\ \sum_{z \in Z} p(w|d)$$

$$= \ p(d)\ \sum_{z \in Z} p(w, z|d)$$

$$= \ p(d)\ \sum_{z \in Z} p(w|d, z)\, p(z|d)$$

$$= \ p(d)\ \sum_{z \in Z} p(w|z)\, p(z|d)$$

$$p\,(d,\ w) \ = \sum_{z \in Z} p(z)\, p(d|z)\, p(w|z)$$

The Likelihood of the corpus can expressed as below:

$$L = P(corpus) = \prod_{(d,w)} p\,(d,w)\ \alpha \prod_{(d,w)} p\,(w|d) = \prod_{d\in D}\prod_{w\in W} p(w|d)^{n(d,w)}$$

$$\mathcal{L} = Log\, L = \sum_{d\in D}\sum_{w\in W} n(d,w).\ log\ (\sum_{z\in Z} p(z)p(d|z)p(w|z)$$

The standard procedure for MLE in latent variable models is the Expectation Maximization (EM) Algorithm, which alternated two coupled step:

- an expectation (E) step where the posterior probabilities are computed for the latent variables. These probabilities can be thought of as the distribution of responsibility for a word in a document among all the topics.

$$p\,(z|\ d,w)\ \alpha\ p(z)\ p(d|z)\ p(w|z)$$

- a maximization (M) step , where the parameters are re-estimated using the responsibilities calculated in the E-step.

$$p\,(w|z)\ \alpha \sum_{d\in D} n(d,w)\ p(z|d,w)$$

$$p\,(d|z)\ \alpha \sum_{w\in W} n(d,w)\ p(z|d,w)$$

$$p\,(z)\ \alpha \sum_{d\in D}\sum_{w\in W} n(d,w)\ p(z|d,w)$$

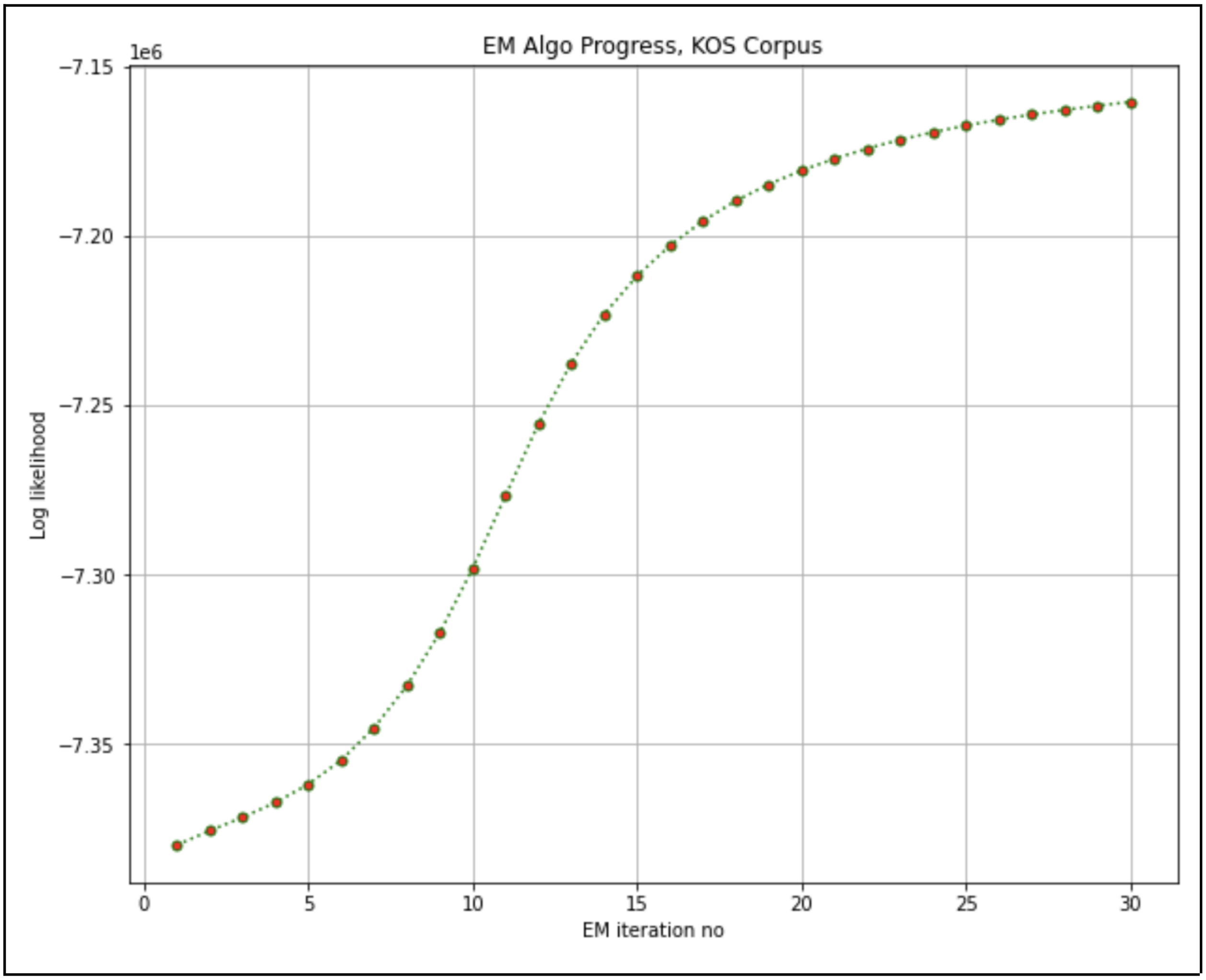
The EM algorithm is run as long as the increase in the $\mathcal{L}$ per iteration is above a certain threshold.

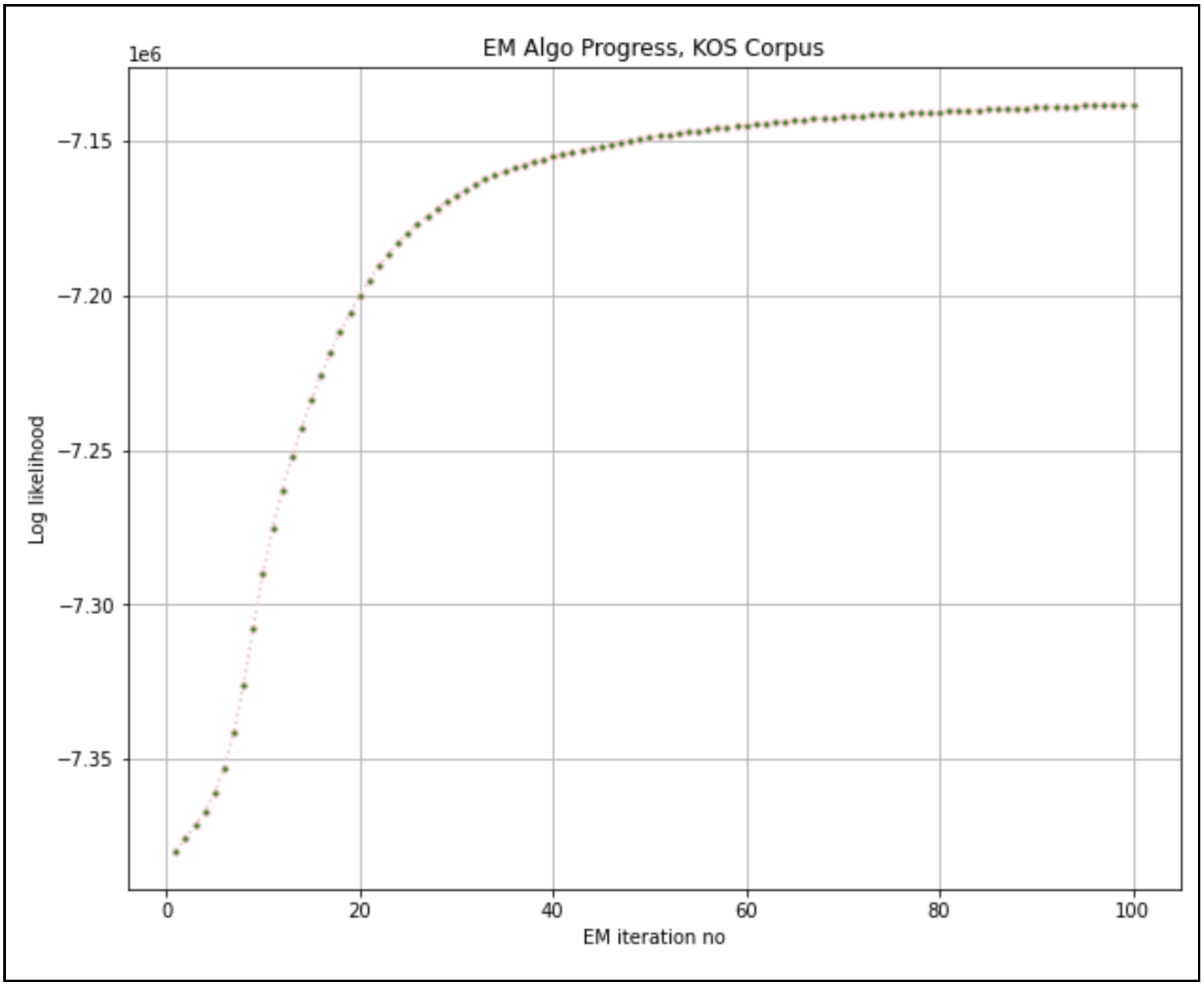
Experiments & Results:

Three datasets have been chosen for demonstration:

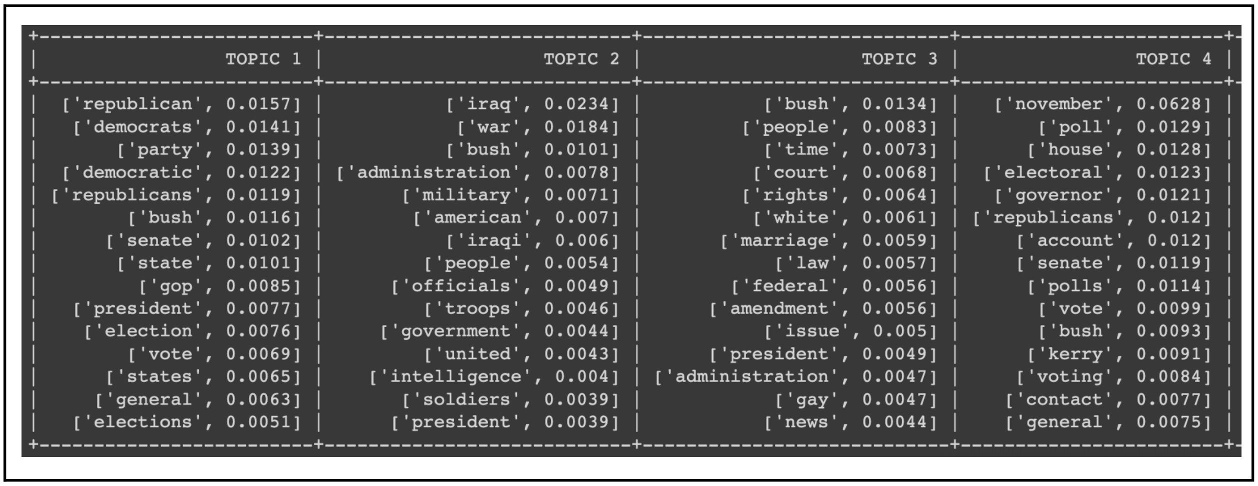
- KOS Blog Entries:
- NIPS Papers:
- NYT articles

KOS Blog Entries:

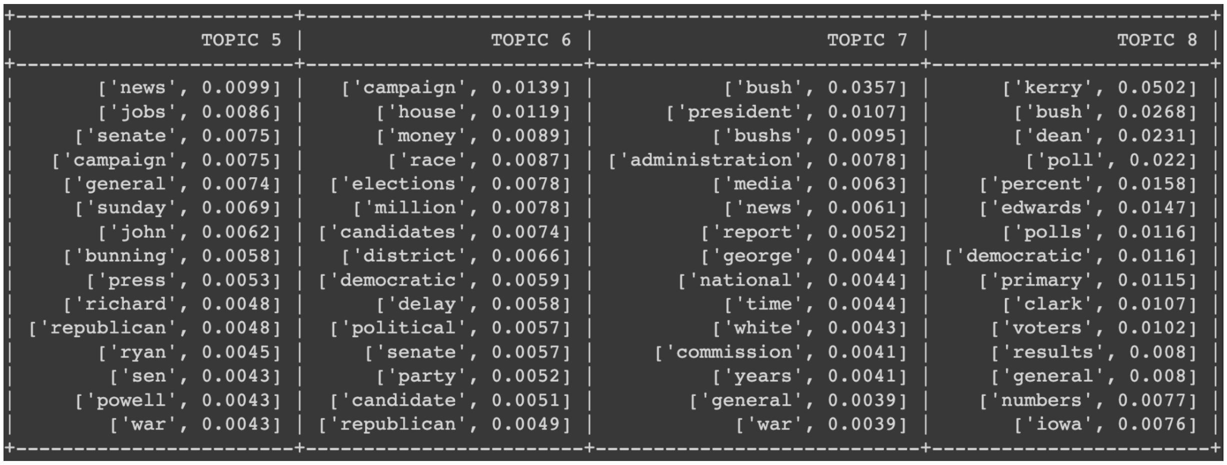




Topics for KOS Blog Corpus (Z=8, top-15 words):



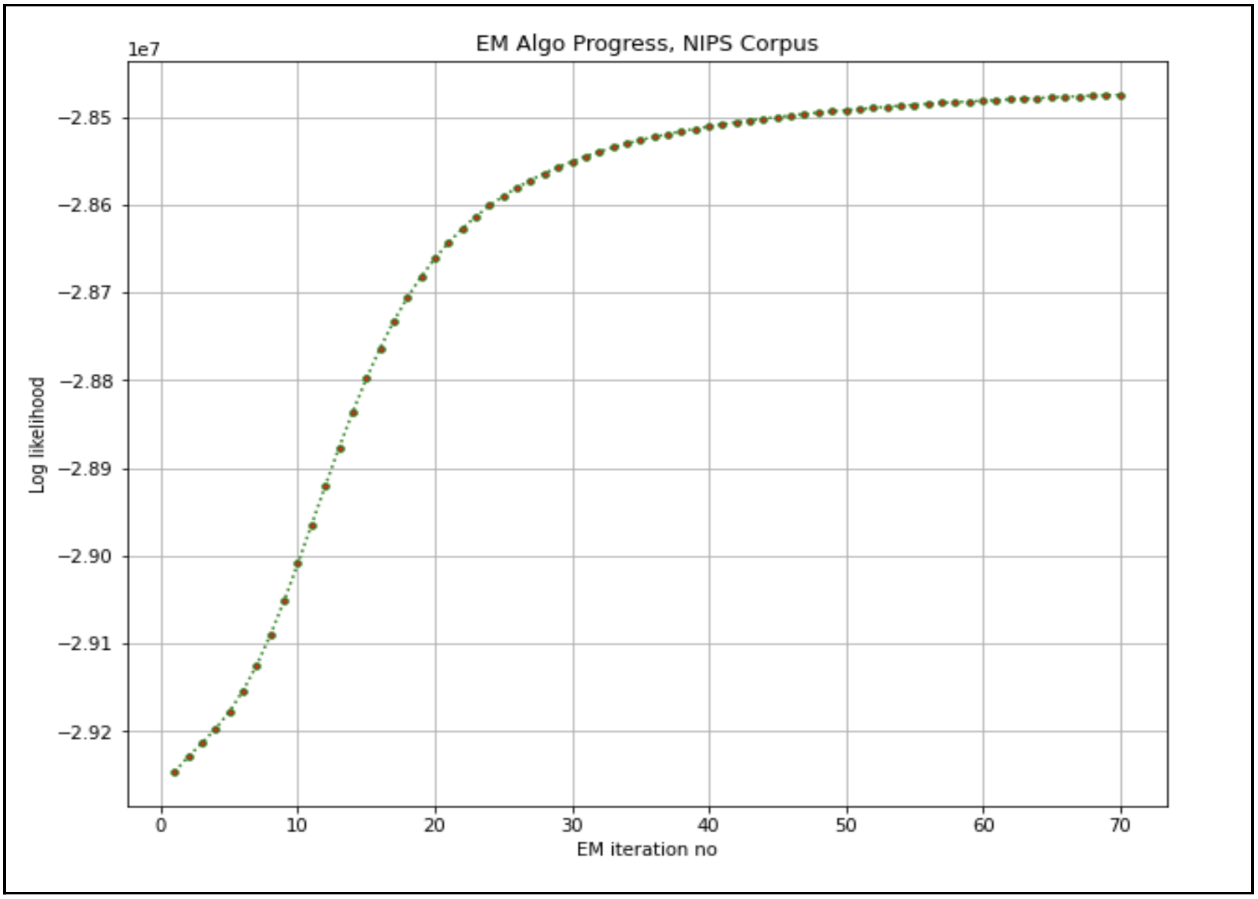
```
+-------------------------+----------------------------+----------------------------+------------------------+
|                 TOPIC 1 |                    TOPIC 2 |                    TOPIC 3 |                TOPIC 4 |
+-------------------------+----------------------------+----------------------------+------------------------+
|  ['republican', 0.0157] |           ['iraq', 0.0234] |           ['bush', 0.0134] |   ['november', 0.0628] |
|   ['democrats', 0.0141] |            ['war', 0.0184] |         ['people', 0.0083] |       ['poll', 0.0129] |
|       ['party', 0.0139] |           ['bush', 0.0101] |           ['time', 0.0073] |      ['house', 0.0128] |
|  ['democratic', 0.0122] | ['administration', 0.0078] |          ['court', 0.0068] |  ['electoral', 0.0123] |
| ['republicans', 0.0119] |       ['military', 0.0071] |         ['rights', 0.0064] |   ['governor', 0.0121] |
|        ['bush', 0.0116] |        ['american', 0.007] |          ['white', 0.0061] | ['republicans', 0.012] |
|      ['senate', 0.0102] |           ['iraqi', 0.006] |       ['marriage', 0.0059] |     ['account', 0.012] |
|       ['state', 0.0101] |         ['people', 0.0054] |            ['law', 0.0057] |     ['senate', 0.0119] |
|         ['gop', 0.0085] |      ['officials', 0.0049] |        ['federal', 0.0056] |      ['polls', 0.0114] |
|   ['president', 0.0077] |         ['troops', 0.0046] |      ['amendment', 0.0056] |       ['vote', 0.0099] |
|    ['election', 0.0076] |     ['government', 0.0044] |           ['issue', 0.005] |       ['bush', 0.0093] |
|        ['vote', 0.0069] |         ['united', 0.0043] |      ['president', 0.0049] |      ['kerry', 0.0091] |
|      ['states', 0.0065] |     ['intelligence', 0.004] | ['administration', 0.0047] |     ['voting', 0.0084] |
|     ['general', 0.0063] |       ['soldiers', 0.0039] |            ['gay', 0.0047] |    ['contact', 0.0077] |
|   ['elections', 0.0051] |      ['president', 0.0039] |           ['news', 0.0044] |    ['general', 0.0075] |
+-------------------------+----------------------------+----------------------------+------------------------+
```



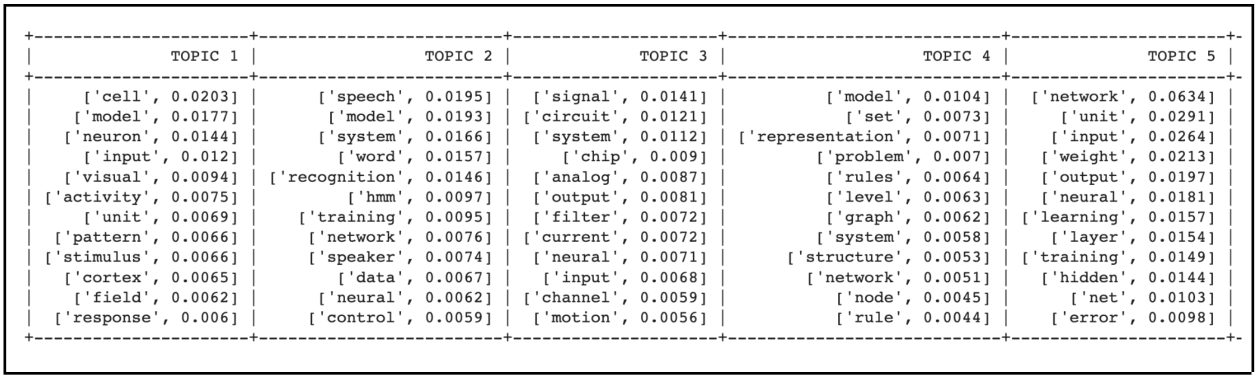
```
+------------------------+------------------------+----------------------------+------------------------+
|                TOPIC 5 |                TOPIC 6 |                    TOPIC 7 |                TOPIC 8 |
+------------------------+------------------------+----------------------------+------------------------+
|       ['news', 0.0099] |   ['campaign', 0.0139] |           ['bush', 0.0357] |      ['kerry', 0.0502] |
|       ['jobs', 0.0086] |      ['house', 0.0119] |      ['president', 0.0107] |       ['bush', 0.0268] |
|     ['senate', 0.0075] |      ['money', 0.0089] |          ['bushs', 0.0095] |       ['dean', 0.0231] |
|   ['campaign', 0.0075] |       ['race', 0.0087] | ['administration', 0.0078] |        ['poll', 0.022] |
|    ['general', 0.0074] |  ['elections', 0.0078] |          ['media', 0.0063] |    ['percent', 0.0158] |
|     ['sunday', 0.0069] |    ['million', 0.0078] |           ['news', 0.0061] |    ['edwards', 0.0147] |
|       ['john', 0.0062] | ['candidates', 0.0074] |         ['report', 0.0052] |      ['polls', 0.0116] |
|    ['bunning', 0.0058] |   ['district', 0.0066] |         ['george', 0.0044] | ['democratic', 0.0116] |
|      ['press', 0.0053] | ['democratic', 0.0059] |       ['national', 0.0044] |    ['primary', 0.0115] |
|    ['richard', 0.0048] |      ['delay', 0.0058] |           ['time', 0.0044] |      ['clark', 0.0107] |
| ['republican', 0.0048] |  ['political', 0.0057] |          ['white', 0.0043] |     ['voters', 0.0102] |
|       ['ryan', 0.0045] |     ['senate', 0.0057] |     ['commission', 0.0041] |     ['results', 0.008] |
|        ['sen', 0.0043] |      ['party', 0.0052] |          ['years', 0.0041] |     ['general', 0.008] |
|     ['powell', 0.0043] |  ['candidate', 0.0051] |        ['general', 0.0039] |    ['numbers', 0.0077] |
|        ['war', 0.0043] | ['republican', 0.0049] |            ['war', 0.0039] |       ['iowa', 0.0076] |
+------------------------+------------------------+----------------------------+------------------------+
```

As can be expected from a blog focussed on the politics of US Democratic Party, most of the topics are concerned with electoral campaigns, polls, Iraq invasion, bills, senate, critiquing republicans etc. Also can be expected is the relative ‘sameness’ between the topics as the blog’s main focus in the narrow domain of politics.

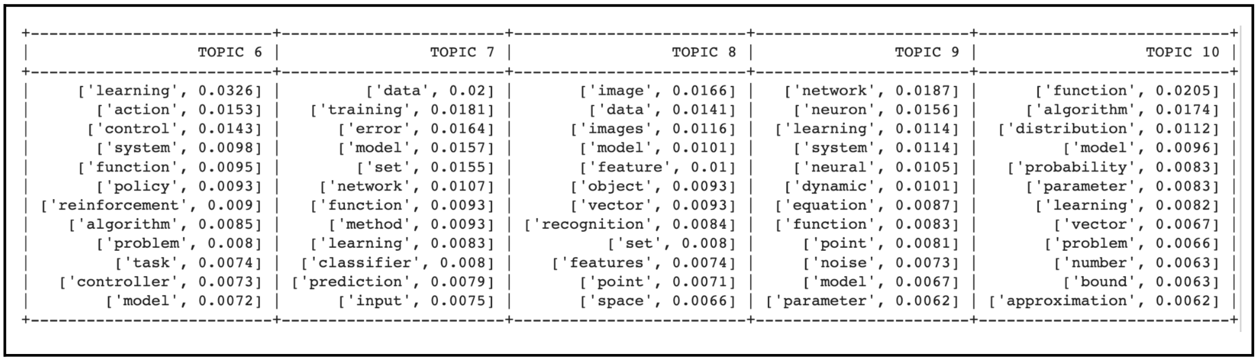
NIPS Papers:



NIPS Corpus Topics (Z=10, top-12 words):



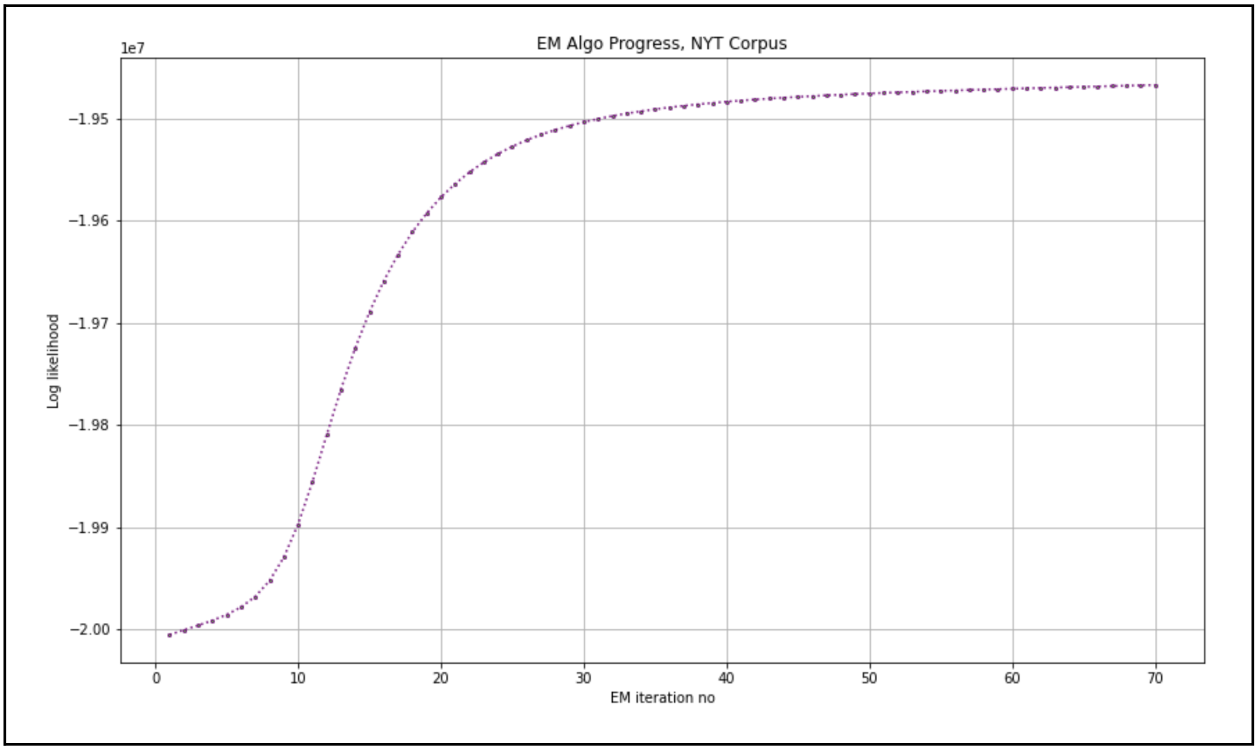
```
+----------------------+-------------------------+---------------------+----------------------------+----------------------+-
|              TOPIC 1 |                 TOPIC 2 |             TOPIC 3 |                    TOPIC 4 |              TOPIC 5 |
+----------------------+-------------------------+---------------------+----------------------------+----------------------+-
|     ['cell', 0.0203] |      ['speech', 0.0195] |  ['signal', 0.0141] |          ['model', 0.0104] |  ['network', 0.0634] |
|    ['model', 0.0177] |       ['model', 0.0193] | ['circuit', 0.0121] |            ['set', 0.0073] |     ['unit', 0.0291] |
|   ['neuron', 0.0144] |      ['system', 0.0166] |  ['system', 0.0112] | ['representation', 0.0071] |    ['input', 0.0264] |
|     ['input', 0.012] |        ['word', 0.0157] |     ['chip', 0.009] |         ['problem', 0.007] |   ['weight', 0.0213] |
|   ['visual', 0.0094] | ['recognition', 0.0146] |  ['analog', 0.0087] |          ['rules', 0.0064] |   ['output', 0.0197] |
| ['activity', 0.0075] |         ['hmm', 0.0097] |  ['output', 0.0081] |          ['level', 0.0063] |   ['neural', 0.0181] |
|     ['unit', 0.0069] |    ['training', 0.0095] |  ['filter', 0.0072] |          ['graph', 0.0062] | ['learning', 0.0157] |
|  ['pattern', 0.0066] |     ['network', 0.0076] | ['current', 0.0072] |         ['system', 0.0058] |    ['layer', 0.0154] |
| ['stimulus', 0.0066] |     ['speaker', 0.0074] |  ['neural', 0.0071] |      ['structure', 0.0053] | ['training', 0.0149] |
|   ['cortex', 0.0065] |        ['data', 0.0067] |   ['input', 0.0068] |        ['network', 0.0051] |   ['hidden', 0.0144] |
|    ['field', 0.0062] |      ['neural', 0.0062] | ['channel', 0.0059] |           ['node', 0.0045] |      ['net', 0.0103] |
|  ['response', 0.006] |     ['control', 0.0059] |  ['motion', 0.0056] |           ['rule', 0.0044] |    ['error', 0.0098] |
+----------------------+-------------------------+---------------------+----------------------------+----------------------+-
```



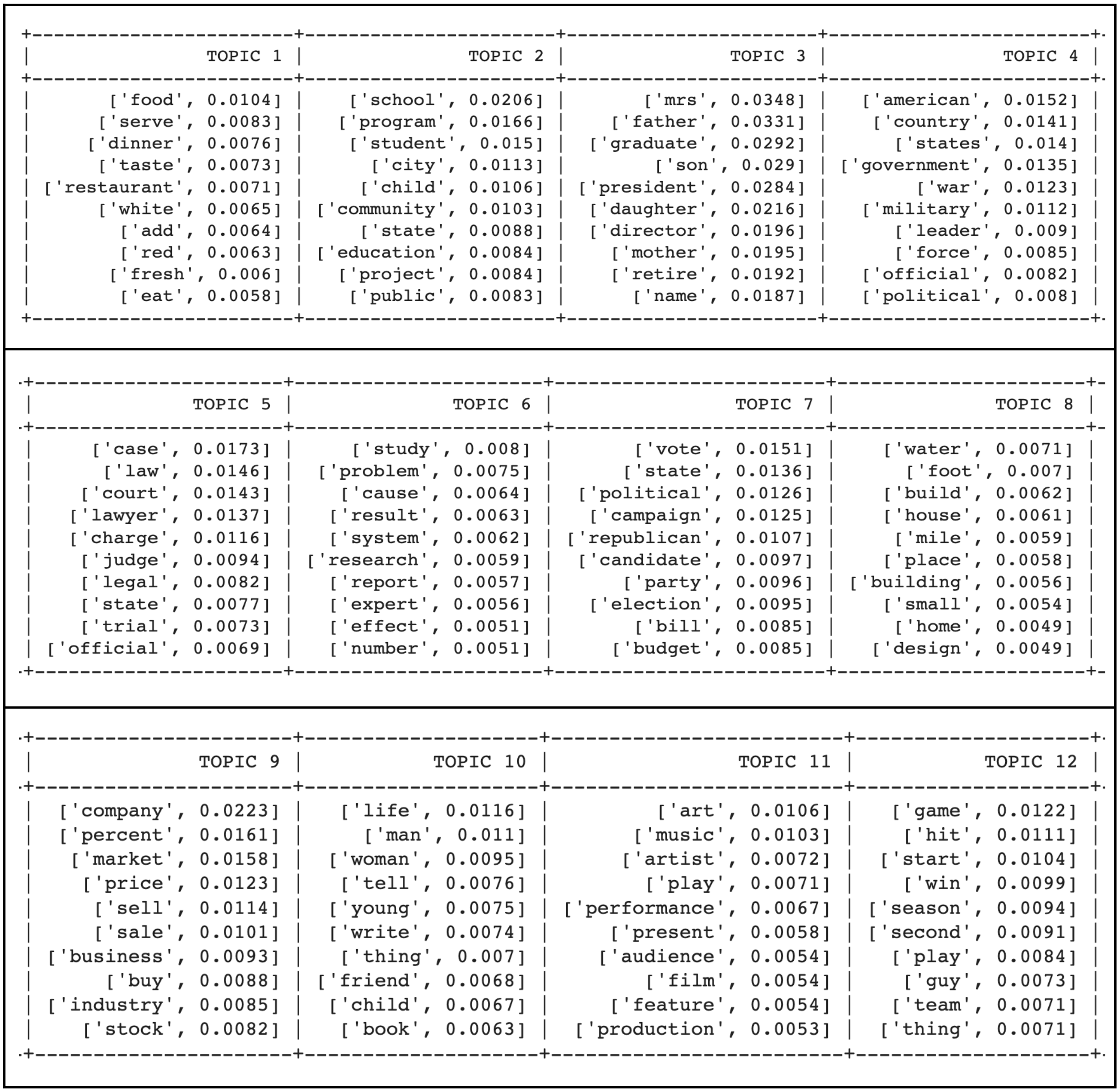
```
+--------------------------+------------------------+-------------------------+-----------------------+---------------------------+
|                  TOPIC 6 |                TOPIC 7 |                 TOPIC 8 |               TOPIC 9 |                  TOPIC 10 |
+--------------------------+------------------------+-------------------------+-----------------------+---------------------------+
|     ['learning', 0.0326] |         ['data', 0.02] |       ['image', 0.0166] |   ['network', 0.0187] |      ['function', 0.0205] |
|       ['action', 0.0153] |   ['training', 0.0181] |        ['data', 0.0141] |    ['neuron', 0.0156] |     ['algorithm', 0.0174] |
|      ['control', 0.0143] |      ['error', 0.0164] |      ['images', 0.0116] |  ['learning', 0.0114] |  ['distribution', 0.0112] |
|       ['system', 0.0098] |      ['model', 0.0157] |       ['model', 0.0101] |    ['system', 0.0114] |         ['model', 0.0096] |
|     ['function', 0.0095] |        ['set', 0.0155] |       ['feature', 0.01] |    ['neural', 0.0105] |   ['probability', 0.0083] |
|       ['policy', 0.0093] |    ['network', 0.0107] |      ['object', 0.0093] |   ['dynamic', 0.0101] |     ['parameter', 0.0083] |
| ['reinforcement', 0.009] |   ['function', 0.0093] |      ['vector', 0.0093] |  ['equation', 0.0087] |      ['learning', 0.0082] |
|    ['algorithm', 0.0085] |     ['method', 0.0093] | ['recognition', 0.0084] |  ['function', 0.0083] |        ['vector', 0.0067] |
|       ['problem', 0.008] |   ['learning', 0.0083] |          ['set', 0.008] |     ['point', 0.0081] |       ['problem', 0.0066] |
|         ['task', 0.0074] |  ['classifier', 0.008] |    ['features', 0.0074] |     ['noise', 0.0073] |        ['number', 0.0063] |
|   ['controller', 0.0073] | ['prediction', 0.0079] |       ['point', 0.0071] |     ['model', 0.0067] |         ['bound', 0.0063] |
|        ['model', 0.0072] |      ['input', 0.0075] |       ['space', 0.0066] | ['parameter', 0.0062] | ['approximation', 0.0062] |
+--------------------------+------------------------+-------------------------+-----------------------+---------------------------+
```

As can be expected from the papers from the Neural Information Processing Systems (NIPS) conference the topics are at the intersection of biology and Machine Learning. On the one end we have Topic 1 which is largely in the domain of 'neuro-biology', while at the other end is Topic 10 which is concerning algorithms & ML. Topic-2 seems to deal with audio-signal processing while Topic-8 seems to be focussed on neural processing of visual stimuli.

NYT Articles Dataset



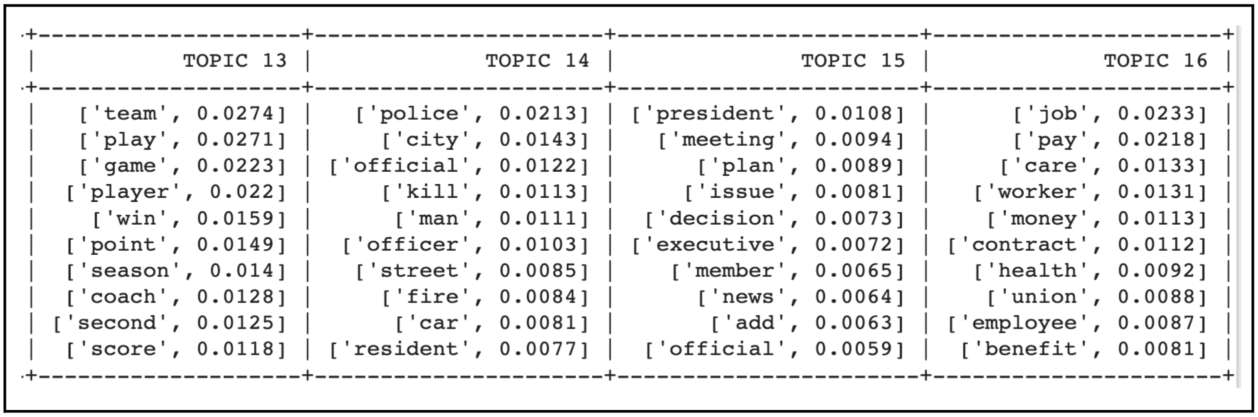
NYT Corpus Topics (Z=16, top-10 words)



```
+------------------------+-----------------------+-----------------------+------------------------+
|                TOPIC 1 |               TOPIC 2 |               TOPIC 3 |                TOPIC 4 |
+------------------------+-----------------------+-----------------------+------------------------+
|       ['food', 0.0104] |    ['school', 0.0206] |       ['mrs', 0.0348] |   ['american', 0.0152] |
|      ['serve', 0.0083] |   ['program', 0.0166] |    ['father', 0.0331] |    ['country', 0.0141] |
|     ['dinner', 0.0076] |    ['student', 0.015] |  ['graduate', 0.0292] |      ['states', 0.014] |
|      ['taste', 0.0073] |      ['city', 0.0113] |        ['son', 0.029] | ['government', 0.0135] |
| ['restaurant', 0.0071] |     ['child', 0.0106] | ['president', 0.0284] |        ['war', 0.0123] |
|      ['white', 0.0065] | ['community', 0.0103] |  ['daughter', 0.0216] |   ['military', 0.0112] |
|        ['add', 0.0064] |     ['state', 0.0088] |  ['director', 0.0196] |      ['leader', 0.009] |
|        ['red', 0.0063] | ['education', 0.0084] |    ['mother', 0.0195] |      ['force', 0.0085] |
|       ['fresh', 0.006] |   ['project', 0.0084] |    ['retire', 0.0192] |   ['official', 0.0082] |
|        ['eat', 0.0058] |    ['public', 0.0083] |      ['name', 0.0187] |   ['political', 0.008] |
+------------------------+-----------------------+-----------------------+------------------------+
```

```
+----------------------+----------------------+------------------------+----------------------+
|              TOPIC 5 |              TOPIC 6 |                TOPIC 7 |              TOPIC 8 |
+----------------------+----------------------+------------------------+----------------------+
|     ['case', 0.0173] |     ['study', 0.008] |       ['vote', 0.0151] |    ['water', 0.0071] |
|      ['law', 0.0146] |  ['problem', 0.0075] |      ['state', 0.0136] |      ['foot', 0.007] |
|    ['court', 0.0143] |    ['cause', 0.0064] |  ['political', 0.0126] |    ['build', 0.0062] |
|   ['lawyer', 0.0137] |   ['result', 0.0063] |   ['campaign', 0.0125] |    ['house', 0.0061] |
|   ['charge', 0.0116] |   ['system', 0.0062] | ['republican', 0.0107] |     ['mile', 0.0059] |
|    ['judge', 0.0094] | ['research', 0.0059] |  ['candidate', 0.0097] |    ['place', 0.0058] |
|    ['legal', 0.0082] |   ['report', 0.0057] |      ['party', 0.0096] | ['building', 0.0056] |
|    ['state', 0.0077] |   ['expert', 0.0056] |   ['election', 0.0095] |    ['small', 0.0054] |
|    ['trial', 0.0073] |   ['effect', 0.0051] |       ['bill', 0.0085] |     ['home', 0.0049] |
| ['official', 0.0069] |   ['number', 0.0051] |     ['budget', 0.0085] |   ['design', 0.0049] |
+----------------------+----------------------+------------------------+----------------------+
```

```
+----------------------+--------------------+-------------------------+--------------------+
|              TOPIC 9 |           TOPIC 10 |                TOPIC 11 |           TOPIC 12 |
+----------------------+--------------------+-------------------------+--------------------+
|  ['company', 0.0223] |   ['life', 0.0116] |         ['art', 0.0106] |   ['game', 0.0122] |
|  ['percent', 0.0161] |     ['man', 0.011] |       ['music', 0.0103] |    ['hit', 0.0111] |
|   ['market', 0.0158] |  ['woman', 0.0095] |      ['artist', 0.0072] |  ['start', 0.0104] |
|    ['price', 0.0123] |   ['tell', 0.0076] |        ['play', 0.0071] |    ['win', 0.0099] |
|     ['sell', 0.0114] |  ['young', 0.0075] | ['performance', 0.0067] | ['season', 0.0094] |
|     ['sale', 0.0101] |  ['write', 0.0074] |     ['present', 0.0058] | ['second', 0.0091] |
| ['business', 0.0093] |   ['thing', 0.007] |    ['audience', 0.0054] |   ['play', 0.0084] |
|      ['buy', 0.0088] | ['friend', 0.0068] |        ['film', 0.0054] |    ['guy', 0.0073] |
| ['industry', 0.0085] |  ['child', 0.0067] |     ['feature', 0.0054] |   ['team', 0.0071] |
|    ['stock', 0.0082] |   ['book', 0.0063] |  ['production', 0.0053] |  ['thing', 0.0071] |
+----------------------+--------------------+-------------------------+--------------------+
```



```
+--------------------+----------------------+-----------------------+----------------------+
|           TOPIC 13 |             TOPIC 14 |              TOPIC 15 |             TOPIC 16 |
+--------------------+----------------------+-----------------------+----------------------+
|   ['team', 0.0274] |   ['police', 0.0213] | ['president', 0.0108] |      ['job', 0.0233] |
|   ['play', 0.0271] |     ['city', 0.0143] |   ['meeting', 0.0094] |      ['pay', 0.0218] |
|   ['game', 0.0223] | ['official', 0.0122] |      ['plan', 0.0089] |     ['care', 0.0133] |
|  ['player', 0.022] |     ['kill', 0.0113] |     ['issue', 0.0081] |   ['worker', 0.0131] |
|    ['win', 0.0159] |      ['man', 0.0111] |  ['decision', 0.0073] |    ['money', 0.0113] |
|  ['point', 0.0149] |  ['officer', 0.0103] | ['executive', 0.0072] | ['contract', 0.0112] |
|  ['season', 0.014] |   ['street', 0.0085] |    ['member', 0.0065] |   ['health', 0.0092] |
|  ['coach', 0.0128] |     ['fire', 0.0084] |      ['news', 0.0064] |    ['union', 0.0088] |
| ['second', 0.0125] |      ['car', 0.0081] |       ['add', 0.0063] | ['employee', 0.0087] |
|  ['score', 0.0118] | ['resident', 0.0077] |  ['official', 0.0059] |  ['benefit', 0.0081] |
+--------------------+----------------------+-----------------------+----------------------+
```

As the corpus is much more diverse compared to the earlier ones, the topics too are more diverse and distinctive. If we were to summarize these topic clusters we would probably see something on the lines expressed below:

Topic 1: food & cuisine
Topic 2: public education & policy
Topic 4: war & politics
Topic 5: judiciary & legal System
Topic 6: studies & reports
Topic 7: polls & campaigns
Topic 8: housing & public utilities
Topic 9: economy & finance
Topic 10: family & lifestyle
Topic 11: arts & culture
Topic 12: sports
Topic 14: law enforcement
Topic 15: government activities
Topic 16: employment & jobs