Supplementary Material for Combinatorial Topic Models using Small-Variance Asymptotics

Ke Jiang
Ohio State University

Suvrit Sra
MIT

Brian Kulis
Boston University

1 Full Derivation of the SVA Objective

Recall the standard Latent Dirichlet Allocation (LDA) model:

- Choose $\theta_j \sim \text{Dir}(\alpha)$, where $j \in \{1, ..., M\}$.
- Choose $\psi_t \sim \text{Dir}(\beta)$, where $t \in \{1, ..., K\}$.
- For each word i in document j:
  - Choose a topic $z_{ji} \sim \text{Cat}(\theta_j)$.
  - Choose a word $w_{ji} \sim \text{Cat}(\psi_{z_{ji}})$.

Here α and β are scalar-valued (i.e., we are using a symmetric Dirichlet distribution). Denote $\mathbf{W}$ as the vector denoting all words in all documents, $\mathbf{Z}$ as the topic indicators of all words in all documents, $\boldsymbol{\theta}$ as the concatenation of all the θ_j variables, and $\boldsymbol{\psi}$ as the concatenation of all the ψ_t variables. Also let N_j be the total number of word tokens in document j. The θ_j vectors are each of length K, the number of topics. The ψ_t vectors are each of length V, the number of words in the dictionary. We can write down the full joint likelihood of the model as $p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\psi}|\alpha, \beta) =$

$$\prod_{t=1}^{K} p(\psi_t|\beta) \prod_{j=1}^{M} p(\theta_j|\alpha) \prod_{i=1}^{N_j} p(z_{ji}|\theta_j) p(w_{ji}|\psi_{z_{ji}}),$$

where each of the probabilities are given as specified in the above model. Now, following standard LDA manipulations, we can eliminate variables to simplify inference by integrating out $\boldsymbol{\theta}$ to obtain

$$p(\mathbf{Z}, \mathbf{W}, \boldsymbol{\psi}|\alpha, \beta) = \int_{\boldsymbol{\theta}} p(\mathbf{W}, \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\psi}|\alpha, \beta) d\boldsymbol{\theta}.$$

After simplification, we obtain $p(\mathbf{Z}, \mathbf{W}, \boldsymbol{\psi}|\alpha, \beta) =$

$$\left[\prod_{t=1}^{K} p(\psi_t|\beta) \prod_{j=1}^{M} \prod_{i=1}^{N_j} p(w_{ji}|\psi_{z_{ji}})\right] \times \left[\prod_{j=1}^{M} \frac{\Gamma(\alpha K)}{\Gamma(\sum_{t=1}^{K} n_{j\cdot}^{t} + \alpha K)} \prod_{t=1}^{K} \frac{\Gamma(n_{j\cdot}^{t} + \alpha)}{\Gamma(\alpha)}\right].$$

Here $n_{j\cdot}^{t}$ is the number of word tokens in document j assigned to topic t. Now, following (Broderick *et al.*, 2013), we can obtain the SVA objective by taking the log of this likelihood and observing what happens when the variance goes to zero. In order to do this, we must be able to scale the likelihood categorical distribution, which is not readily apparent. Here we use two facts about the categorical distribution. First, as discussed in (Banerjee *et al.*, 2005), we can equivalently express the distribution $p(w_{ji}|\psi_{z_{ji}})$ in its Bregman divergence form, which will prove amenable to SVA analysis. In particular, example 10 from (Banerjee *et al.*, 2005) details this derivation. In our case we have a categorical distribution, and thus we can write the probability of token w_{ji} as:

$$p(w_{ji}|\psi_{z_{ji}}) = \exp(-d_\phi(1, \psi_{z_{ji}, w_{ji}})). \quad (1)$$

d_ϕ is the unique Bregman divergence associated with the categorical distribution which, as detailed in example 10 from (Banerjee *et al.*, 2005), is the discrete KL divergence and $\psi_{z_{ji}, w_{ji}}$ is the entry of the topic vector associated with the topic indexed by z_{ji} at the entry corresponding to the word at token w_{ji}. This KL divergence will correspond to a single term of the form $x \log(x/y)$, where $x = 1$ since we are considering a single token of a word in a document. Thus, for a particular token, the KL divergence simply equals $-\log \psi_{z_{ji}, w_{ji}}$. Note that when plugging in $-\log \psi_{z_{ji}, w_{ji}}$ into (1), we obtain exactly the original probability for word token w_{ji} that we had in the original multinomial distribution. We will write the KL-divergence $d_\phi(1, \psi_{z_{ji}, w_{ji}})$ as $\text{KL}(\tilde{w}_{ji}, \psi_{z_{ji}})$, where $\tilde{w}_{ji}$ is an indicator vector for the word at token w_{ji}.

Although it may appear that we have gained nothing by this notational manipulation, there is a key advantage of expressing the categorical probability in terms of Bregman divergences. In particular, the second step is to parameterize the Bregman divergence by an additional variance parameter. As discussed in Lemma 3.1 of (Jiang *et al.*, 2012), we can introduce another parameter, which we will call η, that scales the variance in an exponential family while fixing the mean. This new distribution may be represented, using the Bregman divergence view, as proportional to $\exp(-\eta \cdot \text{KL}(\tilde{w}_{ji}, \psi_{z_{ji}}))$. As $\eta \to \infty$, the mean remains fixed while the variance goes to zero, which is precisely what we require to perform small-variance analysis.

We will choose to scale α appropriately as well; this will ensure that the hierarchical form of the model is retained asymptotically. In particular, we will write $\alpha = \exp(-\lambda \cdot \eta)$. Now we consider the full negative log-likelihood:

$$-\log p(\mathbf{Z}, \mathbf{W}, \boldsymbol{\psi} | \alpha, \beta).$$

Let us first derive the asymptotic behavior arising from the Dirichlet-multinomial distribution part of the likelihood, for a given document j:

$$\frac{\Gamma(\alpha K)}{\Gamma(\sum_{t=1}^{K} n_{j\cdot}^{t} + \alpha K)} \prod_{t=1}^{K} \frac{\Gamma(n_{j\cdot}^{t} + \alpha)}{\Gamma(\alpha)}.$$

In particular, we will show the following lemma.

Lemma 1. *Consider the likelihood*

$$p(\boldsymbol{Z}|\alpha) = \left[\prod_{j=1}^{M} \frac{\Gamma(\alpha K)}{\Gamma(\sum_{t=1}^{K} n_{j\cdot}^{t} + \alpha K)} \prod_{t=1}^{K} \frac{\Gamma(n_{j\cdot}^{t} + \alpha)}{\Gamma(\alpha)} \right].$$

If $\alpha = \exp(-\lambda \cdot \eta)$*, then asymptotically as* $\eta \to \infty$ *we have*

$$-\log p(\boldsymbol{Z}|\alpha) \sim \eta\lambda \textstyle\sum_{j=1}^{M} (K_{j+} - 1).$$

Proof. Note that $N_j = \sum_{t=1}^{K} n_{j\cdot}^{t}$. Using standard properties of the Γ function, we have that the negative log of the above distribution is equal to

$$\sum_{n=0}^{N_j - 1} \log(\alpha K + n) - \sum_{i=1}^{K} \sum_{n=0}^{n_{j\cdot}^{t} - 1} \log(\alpha + n).$$

All of the logarithmic summands converge to a finite constant whenever they have an additional term besides α or αK inside. The only terms that asymptotically diverge are those of the form $\log(\alpha K)$ or $\log(\alpha)$, that is, when $n = 0$. The first term always occurs. Terms of the type $\log(\alpha)$ occur only when, for the corresponding t, we have $n_{j\cdot}^{t} > 0$. Recalling that $\alpha = \exp(-\lambda \cdot \eta)$, we can conclude that the negative log of the Dirichlet multinomial term becomes asymptotically $\eta\lambda(K_{j+} - 1)$, where K_{j+} is the number of topics t in document j where $n_{j\cdot}^{t} > 0$, i.e., the number of topics currently utilized by document j. (The maximum value for K_{j+} is K, the total number of topics.) □

The rest of the negative log-likelihood is straightforward. The $-\log p(\psi_t|\beta)$ terms vanish asymptotically since we are not scaling β (see the note below on scaling β). Thus, the remaining terms in the SVA objective are the ones arising from the word likelihoods which, after applying a negative logarithm, become

$$-\sum_{j=1}^{M} \sum_{i=1}^{N_j} \log p(w_{ji} | \psi_{z_{ji}}).$$

Using the Bregman divergence representation, we can conclude that the negative log-likelihood asymptotically yields the objective $-\log p(\mathbf{Z}, \mathbf{W}, \boldsymbol{\psi} | \alpha, \beta) \sim$

$$\eta \left[\sum_{j=1}^{M} \sum_{i=1}^{N_j} \mathrm{KL}(\tilde{w}_{ji}, \psi_{z_{ji}}) + \lambda \sum_{j=1}^{M} (K_{j+} - 1) \right],$$

where $f(x) \sim g(x)$ denotes that $f(x)/g(x) \to 1$ as $x \to \infty$. This leads to the objective function

$$\min_{\boldsymbol{Z}, \boldsymbol{\psi}} \sum_{j=1}^{M} \sum_{i=1}^{N_j} \mathrm{KL}(\tilde{w}_{ji}, \psi_{z_{ji}}) + \lambda \sum_{j=1}^{M} K_{j+}. \qquad (2)$$

We remind the reader that $\mathrm{KL}(\tilde{w}_{ji}, \psi_{z_{ji}}) = -\log \psi_{z_{ji}, w_{ji}}$. Thus we obtain a k-means-like term that says that any word should be "close" to its assigned topic in terms of KL-divergence under the word co-occurrence constraint enforced with reasonable λ value. Note that (2) reduces to the document-level K-means problem with $\lambda \to \infty$, and the token-level K-means with $\lambda \to 0$.

Note that we did not scale β, to obtain a simpler objective with only one parameter (other than the total number of topics), but let us say a few words about scaling β. A natural approach is to further integrate out $\boldsymbol{\psi}$ of the joint likelihood, as is done with the collapsed Gibbs sampler. One would obtain additional Dirichlet-multinomial distributions, and properly scaling as discussed above would yield a simple objective that places penalties on the number of topics per document as well as the number of words in each topic. Optimization would be performed only with respect to the topic assignment matrix. Future work would consider the effectiveness of such an objective function for topic modeling.

2 An Efficient Facility Location Algorithm for Improved Word Assignments

In this section, we describe an efficient $O(NK)$ algorithm based on facility location for obtaining the word assignments. Recall the algorithm, given for convenience above as Algorithm 1. Our first observation is that, for a fixed size of $\mathcal{W}$ and a given t, the best choice of $\mathcal{W}$ is obtained by selecting the $|\mathcal{W}|$ closest tokens to ψ_t in terms of the KL-divergence. Thus, as a first pass, we can obtain the correct points to mark by appropriately sorting KL-divergences of all tokens to all topics, and then searching over all sizes of $\mathcal{W}$ and topics t.

Next we make three observations about the sorting procedure. First, the KL-divergence between a word and a topic depends purely on counts of words within topics; recall that it is of the form $-\log \psi_{tw}$, where ψ_{tw} equals the number of occurrences of word w in topic t divided by the total number of word tokens assigned to t. Thus, for a given topic, the

Algorithm 1 Improved Word Assignments for **Z**

```
Input: Words: $\mathbf{W}$, Number of topics: $K$, Topic penalty: $\lambda$, Topics: $\boldsymbol{\psi}$
for every document $j$ do
    Let $f_t = \lambda$ for all topics $t$.
    Initialize all word tokens to be unmarked.
    while there are unmarked tokens do
        Pick the topic $t$ and set of unmarked tokens $\mathcal{W}$ that minimizes
            $\frac{f_t + \sum_{i \in \mathcal{W}} \mathrm{KL}(\tilde{w}_{ji}, \psi_t)}{|\mathcal{W}|}$.    (3)
        Let $f_t = 0$ and mark all tokens in $\mathcal{W}$.
        Assign $z_{ji} = t$ for all $i \in \mathcal{W}$.
    end while
end for
Output: Assignments $\mathbf{Z}$.
```

sorted words are obtained exactly by sorting word counts within a topic in decreasing order.

Second, because the word counts are all integers, we can use a linear-time sorting algorithm such as counting sort or radix sort to efficiently sort the items. In the case of counting sort, for instance, if we have n integers whose maximum value is k, the total running time is $O(n + k)$; the storage time is also $O(n + k)$. In our case, we perform many sorts. Each sort considers, for a fixed document d, sorting word counts to some topic t. Suppose there are n_{dt} tokens with non-zero counts to the topic, and the maximum word count is m_{dt}. Then the running time of this sort is $O(n_{dt} + m_{dt})$. Across the document, we do this for every topic, making the running time scale as $O(\sum_t (n_{dt} + m_{dt})) = O(N_d K)$, where N_d is the number of word tokens in document d. Across all documents this sorting then takes $O(NK)$ time.

Third, we note that we need only sort once per run of the algorithm. Once we have sorted lists for words to topics, if we mark some set $\mathcal{W}$, we can efficiently remove these words from the sorted lists and keep the updated lists in sorted order. Removing an individual word from a single sorted list can be done in constant time by maintaining appropriate pointers, for example using a doubly-linked list. Since each word token is removed exactly once during the algorithm, and must be removed from each topic, the total time to update the sorted lists during the algorithm is $O(NK)$.

At this point, we still do not have a procedure that runs in $O(NK)$ time. In particular, we must find the minimum of

$$\frac{f_t + \sum_{i \in \mathcal{W}} \mathrm{KL}(\tilde{w}_{ji}, \psi_t)}{|\mathcal{W}|}$$

at each round of marking. Naively this is performed by traversing the sorted lists and accumulating the value of the above score via summation. In the worst case, each round would take a total of $O(NK)$ time across all documents, so if there are R rounds on average across all the documents, the total running time would be $O(NKR)$. However, we can observe that we need not traverse entire sorted lists in general. Consider a fixed document, where we try to find the minimizer for some fixed topic t of the above expression. We can show that the value monotonically decreases until hitting the minimum value, and then monotonically increases afterward. We can formalize the monotonicity of the scoring function as follows:

Proposition 1. *Let s_{ni} be the value of the scoring function (3) for the best candidate set $\mathcal{W}$ of size n for topic t. If $s_{n-1,t} \leq s_{nt}$, then $s_{nt} \leq s_{n+1,t}$.*

Proof. Recall that the KL-divergence is equal to the negative logarithm of the number of occurrences of the corresponding word token divided by the total number occurrences of tokens in the topic. Write this as $\log n_t - \log c_{tj}$, where n_t is the number of occurrences of tokens in topic t and c_{tj} is the count of the j-th highest-count word j in topic t. Now, by assumption $s_{n-1,t} \geq s_{nt}$. Plugging the score functions into this inequality and cancelling the $\log n_t$ terms, we have

$$-\frac{1}{n-1} \sum_{j=1}^{n-1} \log c_{tj} + \frac{f_t}{n-1} \leq -\frac{1}{n} \sum_{j=1}^{n} \log c_{tj} + \frac{f_t}{n}.$$

Multiplying by $n(n-1)$ and simplifying yields the inequality

$$f_t + n \log c_{tn} \leq \sum_{j=1}^{n} \log c_{tj}.$$

Now, assuming this holds for $s_{n-1,t}$ and $s_{n,t}$, we must show that this inequality also holds for $s_{n,t}$ and $s_{n+1,t}$, i.e. that

$$f_t + (n+1) \log c_{t,n+1} \leq \sum_{j=1}^{n+1} \log c_{tj}.$$

Simple algebraic manipulation and the fact that the counts are sorted, i.e., $\log c_{t,n+1} \leq \log c_{tn}$, shows the inequality to hold. □

In words, the above proof demonstrates that, once the scoring function stops decreasing, it will not decrease any further, i.e., the minimum score has been found. Thus, once the score function starts to increase as $\mathcal{W}$ gets larger, we can stop and the best score (i.e., the best set $\mathcal{W}$) for that topic t has been found. We do this for all topics t until we find the best set $\mathcal{W}$. Under the mild assumption that the size of the chosen minimizer $\mathcal{W}$ is similar (specifically, within a constant factor) to the average size of the best candidate sets $\mathcal{W}$ across the other topics (an assumption which holds in practice), then it follows that the total time to find all the sets $\mathcal{W}$ takes $O(NK)$ time.

Objective Value ($\times 10^6$)	SynthA	SynthB
`Basic`	5.07	5.45
`Word`	4.06	3.79
`Word+Refine`	3.98	3.61

Table 1: Optimized combinatorial topic modeling objective function values for different algorithms with $\lambda = 10$.

Putting everything together, all the steps of this algorithm combine to cost $O(NK)$ time.

3 Additional Experimental Results

Objective optimization. Table 1 shows the optimized objective function values for `Basic`, `Word` and `Word+Refine` algorithms. We can see that the `Word` algorithm significantly reduces the objective value when compared with the `Basic` algorithm, and the `Word+Refine` algorithm reduces further. As pointed out in (Yen *et al.*, 2015) in the context of other SVA models, the `Basic` algorithm is very sensitive to initializations and λ values. However, this is not the case for the `Word` and `Word+Refine` algorithms and they are quite robust to initializations. From the objective values, the improvement from `Word+Refine` to `Word` seems to be marginal, but the incorporation of the local refinement is crucial for learning good topic models.

Running time. See Table 2 for comparisons of our approach to `CGS`. The two most expensive steps of the `Word+Refine` algorithm are the word assignments via facility location and the local refinement step (the other steps of the algorithm are lower-order). The relative running times improve as the data set sizes gets larger and, on large data sets, an iteration of `Refine` is roughly equivalent to one Gibbs iteration while an iteration of `Word` is roughly equivalent to two Gibbs iterations. Since one typically runs thousands of Gibbs iterations (while ours runs in 10 iterations even on very large data sets, yielding a running time equivalent to approximately 30 Gibbs iterations), we can observe several orders of magnitude improvement in speed by our algorithm. Further, running time could be significantly enhanced by noting that the `Word` algorithm trivially parallelizes.

Topic reconstruction error. We look at the reconstruction error between the true topic-word distributions and the learned distributions. In particular, given a learned topic matrix $\hat{\psi}$ and the true matrix ψ, we use the Hungarian algorithm (Kuhn, 1955) to align topics, and then evaluate the ℓ_1 distance between each pair of topics. In addition to the fixed document-length synthetic datasets, we also consider generating documents with varied number of word tokens similar to (Podosinnikova *et al.*, 2015). We report the results of different spectral methods with both the LDA moments Anandkumar *et al.* (2012) and the discrete independent anal-

Method	Number of Documents				
	10k	50k	100k	500k	1M
`CGS` (s)	.321	1.96	4.31	23.36	55.69
`Word` (s)	.922	4.88	9.75	50.38	101.58
`Refine` (s)	.533	2.58	5.09	25.75	52.28
`W/CGS`	2.87	2.48	2.26	2.16	1.82
`R/CGS`	1.66	1.32	1.18	1.10	0.94

Table 2: Running time comparison per iteration (in secs) of `CGS` to the improved word algorithm (`Word`) and the local refinement algorithm (`Refine`).

	SynthA	
	same length	varied length
`Word`	0.220 (0.428)	0.283 (0.471)
`VB`	0.059 (0.010)	0.317 (0.543)
`Spectral-DICA`	0.557 (0.430)	0.475 (0.268)
`Spectral-LDA`	0.112 (0.020)	0.120 (0.041)
`JD-DICA`	0.259 (0.336)	0.153 (0.007)
`JD-LDA`	0.099 (0.005)	0.102 (0.006)
`TPM-DICA`	1.717 (0.091)	0.077 (0.004)
`TPM-LDA`	0.099 (0.005)	0.033 (0.002)
`Anchor`	0.102 (0.018)	0.103 (0.019)
`W+R`	0.080 (0.005)	0.080 (0.004)
`CGS`	0.197 (0.440)	0.059 (0.003)
	SynthB	
	same length	varied length
`Word`	0.504 (0.676)	0.363 (0.559)
`VB`	0.392 (0.663)	0.401 (0.662)
`Spectral-DICA`	0.707 (0.495)	0.635 (0.461)
`Spectral-LDA`	0.314 (0.199)	0.274 (0.170)
`JD-DICA`	0.115 (0.270)	0.113 (0.267)
`JD-LDA`	0.161 (0.022)	0.162 (0.022)
`TPM-DICA`	0.946 (0.868)	1.571 (0.735)
`TPM-LDA`	0.153 (0.017)	0.155 (0.017)
`Anchor`	0.112 (0.028)	0.111 (0.023)
`W+R`	0.105 (0.271)	0.051 (0.004)
`CGS`	0.276 (0.556)	0.160 (0.421)

Table 3: Comparison of topic reconstruction errors of different algorithms where the number inside the parentheses is the standard deviation. Here, “same length” means all the generated documents have the same number of word tokens, while “varied length” means the generated documents have varied number of word tokens. The “same length” and “varied length” documents share the same mean length. Also, “LDA” is the latent Dirichlet allocation moments (Anandkumar *et al.*, 2012), and “DICA” is the discrete independent component analysis cumulants (Podosinnikova *et al.*, 2015).

ysis cumulants (Podosinnikova *et al.*, 2015).

Table 3 presents the mean reconstruction errors and standard deviations for 10K documents. We can see that the LDA-moments based spectral methods, the `Anchor` method and the proposed `Word+Refine` method are insensitive to the variation of the document lengths and perform consistently. However, the DICA-cumulants based spectral methods are quite sensitive and unstable. The Gibbs sampler can easily become trapped in a local optima area and needs many iterations to start to mix. On the other hand, the `Anchor` and proposed `Word+Refine` methods perform very nicely, where `Word+Refine` gives often better results and allows more flexibility.

Topics learned from NYTimes. Table 4 provides the full list learned from the NYTimes dataset.

References

A. Anandkumar, Y. Liu, D. J. Hsu, D. P. Foster, and S. M. Kakade. A spectral algorithm for latent Dirichlet allocation. In *NIPS*, pages 917–925, 2012.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005.

T. Broderick, B. Kulis, and M. I. Jordan. MAD-Bayes: MAP-based asymptotic derivations from Bayes. In *ICML*, 2013.

K. Jiang, B. Kulis, and M. I. Jordan. Small-variance asymptotics for exponential family Dirichlet process mixture models. In *NIPS*, 2012.

H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955.

A. Podosinnikova, F. Bach, and S. Lacoste-Julien. Rethinking LDA: moment matching for discrete ica. In *NIPS*, pages 514–522, 2015.

I. E. H. Yen, X. Lin, K. Zhang, P. Ravikumar, and I. S. Dhillon. A convex exemplar-based approach to MAD-Bayes Dirichlet process mixture models. In *ICML*, 2015.

CGS	Word+Refine
computer, zzz_microsoft, software, system, program, window, technology, user, zzz_government, data	computer, web, site, zzz_internet, internet, online, sites, information, mail, software
president, zzz_george_bush, zzz_bush, zzz_white_house, zzz_clinton, zzz_bill_clinton, administration, presidential, zzz_dick_cheney, zzz_washington	zzz_bush, president, administration, zzz_george_bush, government, zzz_white_house, zzz_al_gore, zzz_clinton, campaign, zzz_congress
zzz_olympic, games, team, gold, medal, sport, zzz_u_s, event, won, women	smiles, ghost, gang, gown, lip, screaming, lightning, foster, greet, lonely
card, store, stores, holiday, customer, zzz_christmas, item, shopping, gift, sales	laptop, disk, click, desktop, keyboard, mouse, printed, videos, zzz_sony, mode
family, father, son, died, wife, daughter, brother, marriage, married, husband	marriage, divorce, wedding, couples, divorced, marry, dating, phrase, royal, draft
guy, big, young, fun, kid, today, ago, show, kind, friend	zzz_right, strength, attitude, speaking, realize, columnist, bear, mention, guess, honest
public, history, personal, early, ago, close, role, career, brought, moment	creating, status, core, rank, consultant, promise, desire, zzz_west, highest, pointed
zzz_american, war, flag, zzz_america, history, american, nation, zzz_king, country, event	zzz_pearl_harbor, zzz_vietnam, zzz_navy, tragedy, japanese, flag, enemy, innocent, bombing, zzz_world_war_ii
attack, terrorist, security, zzz_fbi, official, government, terrorism, zzz_united_states, intelligence, agent	police, attack, official, zzz_fbi, security, officer, federal, information, agent, letter
season, player, draft, team, pick, round, play, career, guy, free	discrimination, zzz_reagan, defended, zzz_commission, zzz_richard_nixon, principles, credential, urged, courage, fairly
government, political, election, country, leader, minister, president, power, party, opposition	zzz_mexico, led, protest, leading, anti, press, french, view, attempt, conference
hour, night, left, morning, told, car, police, minutes, street, body	started, morning, felt, front, close, moment, wanted, couple, car, mind
feel, guy, hard, thought, put, bad, kind, head, happen, point	emotion, realize, sad, rarely, experienced, smile, totally, chose, shy, proud
campaign, political, mayor, governor, zzz_new_york, candidates, election, office, democratic, run	dignity, pen, wheelchair, zzz_red, zzz_rudolph_giuliani, harassment, residence, vintage, medal, zzz_new_yorker
union, worker, officer, police, labor, member, strike, zzz_union, official, gang	zzz_brown, zzz_davis, zzz_johnson, institute, exempt, zzz_kansas_city, compact, eligible, zzz_act, zzz_nevada
official, meeting, decision, leader, talk, night, told, conference, statement, plan	unable, respond, representative, wire, shortly, contacted, funeral, faced, locked, losing
art, artist, painting, museum, century, show, collection, history, french, exhibition	painting, exhibition, portrait, drawing, object, photograph, gallery, flag, artist, painted
company, zzz_enron, companies, million, deal, billion, firm, executives, chief, executive	company, million, percent, companies, business, money, plan, industry, high, part
zzz_china, zzz_united_states, zzz_japan, chinese, countries, japanese, european, zzz_europe, foreign, trade	zzz_united_states, government, zzz_u_s, country, zzz_american, group, zzz_china, countries, foreign, power
black, white, zzz_black, african, race, racial, hispanic, american, percent, zzz_african_american	zzz_king, zzz_black, zzz_african_american, zzz_civil_war, flag, whites, racial, zzz_south, racist, memorial
point, game, team, play, shot, zzz_laker, games, season, goal, final	zzz_laker, zzz_kobe_bryant, zzz_o_neal, rebound, zzz_phil_jackson, pointer, rocket, foul, zzz_nba, shooting

water, fish, bird, ship, boat, species, forest, fishing, sea, island	ship, beach, shark, port, boat, bird, golf, river, adventure, sea
family, friend, feel, child, lives, parent, children, feeling, home, relationship	psychologist, emotion, soccer, psychological, emotional, depression, italian, zzz_world_cup, mood, handed
zzz_tiger_wood, round, player, tour, tournament, shot, golf, play, par, major	shot, goal, king, minutes, ahead, break, beat, net, minute, set
money, million, fund, donation, pay, dollar, contribution, donor, raising, financial	fund, raising, contribution, donation, raised, donor, soft, raise, finance, foundation
newspaper, magazine, media, reporter, press, issue, journalist, article, public, wrote	correspondent, subscription, chat, zzz_washington_post, headlines, rumor, fee, appearances, generated, actual
trip, tour, www, travel, visitor, night, hotel, offer, ticket, hour	hotel, miles, water, holiday, visitor, guest, tour, flower, mountain, tourist
weather, rain, air, storm, snow, wind, water, zzz_new_england, cold, temperatures	weather, rain, sun, water, wind, storm, forest, trees, river, coast
job, worker, employees, company, companies, firm, manager, business, employer, executive	zzz_brazil, brazilian, recruiting, applicant, hiring, opportunities, attract, zzz_chronicle, innovation, boom
number, find, hand, fact, point, reason, big, line, hard, makes	sees, behavior, advice, telling, miss, likes, wait, write, guess, worse
food, eat, fat, meat, drink, chicken, diet, restaurant, product, meal	teaspoon, sauce, meat, pan, flavor, onion, cook, recipe, fruit, juice
school, student, teacher, children, education, test, district, parent, program, public	school, student, children, high, group, part, program, family, show, percent
plane, flight, airport, passenger, pilot, aircraft, crew, planes, air, jet	flight, plane, passenger, airport, pilot, airline, aircraft, jet, planes, airlines
scientist, human, research, researcher, science, stem, brain, found, genetic, scientific	drug, anthrax, research, human, researcher, scientist, virus, test, stem, infection
zzz_united_states, zzz_bush, zzz_u_s, zzz_american, administration, policy, zzz_washington, foreign, official, international	zzz_kosovo, sanction, zzz_beijing, zzz_yugoslavia, iraqi, zzz_south_korea, communist, korean, zzz_iran, diplomatic
children, mother, family, son, daughter, father, parent, child, home, husband	friend, women, family, father, son, mother, wife, told, house, woman
percent, economy, economic, rate, growth, rates, cut, economist, market, number	longer, expect, large, higher, quality, growing, period, areas, difficult, huge
book, writer, author, writing, word, read, wrote, write, history, character	theatrical, nomination, revival, charming, premiere, terrific, premise, variation, themes, acclaimed
police, prison, charges, officer, case, prosecutor, crime, criminal, arrested, arrest	suspected, identified, warrant, linked, plot, classified, gang, questioning, suspicion, hijacker
women, percent, study, found, group, survey, number, studies, research, high	suggest, research, scientist, studies, tend, natural, behavior, science, researcher, theory
question, asked, word, answer, talk, speak, language, interview, meeting, point	part, question, ago, called, problem, kind, show, early, making, half
plant, industry, environmental, water, farm, government, farmer, pollution, air, regulation	monument, coal, zzz_phoenix, mining, zzz_arizona, zzz_arizona_republic, mine, gold, abandoned, tunnel
priest, sexual, sex, church, abuse, gay, bishop, victim, cardinal, children	accounting, bankruptcy, lay, abuse, complaint, fraud, filing, transaction, partnership, scandal
zzz_internet, companies, customer, company, network, consumer, services, zzz_at, cable, zzz_aol	zzz_at, cable, telecommunication, subscriber, zzz_aol_time_warner, combined, distance, takeover, zzz_nasdaq, acquire

award, won, zzz_oscar, winner, zzz_academy, dog, nomination, honor, win, contest	younger, realized, focused, youth, speaking, broke, conversation, older, shared, strength
yard, game, team, quarterback, season, play, defense, touchdown, zzz_nfl, defensive	yard, football, quarterback, defense, shot, zzz_nfl, defensive, round, offense, goal
building, house, project, space, office, neighborhood, home, zzz_new_york, center, estate	built, land, space, area, art, development, building, design, downtown, build
television, network, station, broadcast, show, radio, commercial, zzz_nbc, advertising, program	show, television, network, media, zzz_nbc, station, cable, zzz_abc, broadcast, zzz_cb
million, zzz_los_angeles, zzz_arizona, area, zzz_phoenix, local, zzz_california, cities, public, zzz_new_york	casino, gambling, zzz_san_antonio, zzz_las_vegas, zzz_austin, wake, forest, zzz_dallas, texan, zzz_houston
town, road, local, miles, land, resident, farm, small, country, ago	bridge, pet, nail, railroad, subway, atop, screaming, distant, suburb, lobby
group, member, board, director, public, meeting, committee, organization, agency, president	guidelines, application, overturned, precedent, zzz_commission, zzz_smith, submitted, intent, statute, enforce
home, family, job, friend, told, wanted, worked, knew, thought, wife	hard, put, kind, give, past, side, hand, feel, night, left
official, investigation, document, information, record, case, letter, agency, told, statement	donor, liver, pet, inspector, zzz_society, properly, tank, procedures, resulted, transplant
daily, question, statesman, palm, information, beach, american, austin, zzz_eastern, sport	daily, question, statesman, palm, american, information, beach, zzz_washington, austin, zzz_eastern
run, inning, hit, game, pitch, ball, home, field, season, lead	pitches, rookie, bullpen, plate, zzz_league, devil, passes, sixth, zzz_st_louis, pitched
season, team, player, yankees, baseball, game, zzz_met, games, zzz_red_sox, manager	corp, homework, bat, pitch, dance, puzzle, subtle, clue, pure, mine
zzz_florida, election, ballot, votes, vote, zzz_al_gore, recount, voter, count, result	recount, count, zzz_florida, votes, majority, voted, cast, counties, electoral, ballot
car, driver, truck, vehicles, vehicle, zzz_ford, seat, wheel, driving, drive	car, driver, vehicles, vehicle, truck, wheel, fuel, engine, drive, zzz_ford
stock, percent, market, fund, quarter, analyst, earning, share, company, investor	stock, market, billion, analyst, investment, quarter, investor, fund, prices, share
zzz_al_gore, campaign, zzz_george_bush, voter, republican, presidential, zzz_john_mccain, zzz_bush, poll, democratic	campaign, zzz_al_gore, election, zzz_george_bush, vote, voter, political, republican, democratic, presidential
religious, zzz_god, religion, christian, church, faith, jewish, jew, zzz_muslim, muslim	religious, church, priest, jewish, religion, zzz_god, jew, faith, christian, zzz_muslim
room, restaurant, hotel, dinner, wine, guest, bar, table, food, night	bathroom, bug, machine, boxes, stuck, wedding, wash, bite, soap, mold
women, fashion, wear, designer, show, clothes, shirt, wearing, dress, black	fashion, wear, shirt, designer, clothes, suit, dress, blue, wearing, jean
court, law, case, federal, decision, lawyer, legal, lawsuit, ruling, zzz_supreme_court	court, case, lawyer, law, legal, federal, decision, lawsuit, attorney, judge
zzz_taliban, zzz_afghanistan, military, forces, war, bin, laden, afghan, official, zzz_pakistan	afghan, fighter, commander, zzz_northern_alliance, zzz_kabul, qaida, zzz_osama, prisoner, zzz_united_nation, ethnic
music, song, band, album, musical, singer, record, concert, artist, pop	song, music, band, album, singer, pop, concert, artist, rock, dance
million, deal, contract, agent, agreement, free, offer, sign, money, pay	follow, agreement, negotiation, process, optional, agreed, sides, statement, union, continue

telegram, zzz_texas, zzz_mexico, mexican, immigrant, visit, services, www, web, zzz_world_wide	fax, syndicate, zzz_mexico, con, zzz_vicente_fox, mexican, zzz_paris, article, zzz_canada, purchased
zzz_russia, zzz_iraq, weapon, zzz_russian, nuclear, missile, russian, defense, zzz_moscow, zzz_vladimir_putin	zzz_russia, nuclear, zzz_russian, military, russian, missile, defense, weapon, zzz_moscow, arm
company, companies, business, industry, product, technology, market, million, firm, billion	initial, antitrust, technologies, machines, innovation, creating, compete, text, zzz_at, monopoly
death, trial, lawyer, penalty, case, court, jury, judge, execution, prosecutor	prison, prosecutor, police, criminal, charges, crime, lawyer, victim, court, investigation
film, movie, play, character, actor, director, movies, zzz_hollywood, minutes, theater	film, movie, actor, movies, play, character, zzz_hollywood, minutes, director, theater
system, problem, plan, change, important, effort, put, difficult, making, number	headquarter, highly, signal, task, institution, begun, sources, established, increasing, possibility
friend, home, night, told, asked, house, ago, wife, family, thought	irish, horses, loud, circle, rush, figured, occasional, silence, youngest, accent
patient, doctor, hospital, medical, care, cancer, treatment, health, disease, blood	patient, doctor, hospital, medical, care, health, disease, treatment, cancer, women
cup, minutes, tablespoon, add, teaspoon, oil, pepper, butter, cream, sugar	cup, food, water, minutes, add, tablespoon, oil, restaurant, hot, large
zzz_america, power, economic, political, country, zzz_american, american, problem, today, nation	irish, enormous, lies, notion, tap, threatening, master, observer, extreme, ordinary
zzz_republican, zzz_party, republican, zzz_senate, democratic, vote, democrat, abortion, political, zzz_democrat	opposed, marriage, conservatives, differences, ban, gay, voucher, supported, allowing, politically
team, player, coach, sport, game, fan, football, league, basketball, zzz_nba	team, game, season, player, play, games, point, run, won, win
drug, anthrax, disease, test, virus, zzz_fda, cases, testing, infection, infected	drug, food, product, disease, animal, health, zzz_fda, scientist, consumer, research
book, zzz_brown, zzz_schuster, zzz_warner, zzz_simon, sales, woman, author, bookstores, memoir	memoir, zzz_schuster, fiction, biography, zzz_simon, bookstores, volume, zzz_warner, publisher, ranking
war, military, rebel, zzz_india, con, government, troop, zzz_colombia, zzz_pakistan, army	carpet, diamond, seed, jewelry, pile, garden, purple, branch, weed, planted
boy, gun, child, girl, teen, father, children, zzz_miami, parent, kid	zzz_valley, courage, wanting, bother, practically, trademark, launching, inform, pen, certainty
program, million, government, zzz_aid, aid, group, care, money, percent, poor	welfare, poverty, subsidies, assistance, minorities, eligible, gap, wage, household, employment
sales, million, market, price, company, sell, consumer, percent, sold, prices	car, sales, price, sell, market, sold, buy, store, product, cost
student, school, college, program, high, professor, campus, zzz_university, zzz_harvard, class	reunion, zzz_stanford, zzz_princeton, zzz_yale, coin, zzz_harvard, zzz_south_florida, cheer, accomplished, sounded
bill, zzz_congress, zzz_bush, legislation, zzz_senate, law, proposal, administration, federal, zzz_white_house	limiting, safeguard, damaging, accountable, mechanism, premature, array, discourage, pen, preventing
game, games, fight, video, player, zzz_dvd, computer, digital, screen, play	fight, bigger, ring, challenge, doubt, junior, deep, successful, fighting, ability
syndicate, fax, article, information, contact, separate, buy, art, zzz_u_s, sales	article, information, art, mail, purchasing, contact, separate, syndicate, buy, word
plant, water, flower, garden, light, hand, floor, wood, house, skin	metal, wood, paint, repair, machine, steel, concrete, wire, clean, roof
web, site, www, sites, mail, online, internet, information, find, offer	www, web, telegram, visit, site, information, room, hour, book, show

game, team, season, coach, play, zzz_ucla, tournament, games, zzz_usc, zzz_ncaa	gloves, beam, instrument, gravity, variation, oxygen, earliest, elephant, span, ray
oil, energy, gas, power, prices, zzz_california, electricity, fuel, price, market	retailer, sale, electricity, auction, supply, stores, production, zzz_california, gas, item
show, series, season, network, television, zzz_nbc, zzz_abc, zzz_fox, episode, zzz_cb	smart, guest, drama, zzz_tony, sitcom, tonight, episodes, imagine, zzz_west, wing
airline, travel, carrier, flight, industry, airlines, zzz_delta, zzz_american, zzz_united, business	inflation, zzz_fed, forecast, portfolio, unemployment, rising, slowdown, layoff, percentage, zzz_nasdaq
tax, cut, billion, taxes, plan, government, pay, income, zzz_social_security, benefit	bill, cut, billion, cost, tax, benefit, proposal, spending, health, taxes
group, police, protest, killed, camp, government, street, protester, violence, killing	protester, zzz_lebanon, suicide, zzz_syria, zzz_authority, zzz_jerusalem, cease, zzz_gaza, jewish, occupation
attack, zzz_new_york, fire, zzz_world_trade_center, firefighter, building, worker, disaster, terrorist, tower	zzz_world_trade_center, terrorist, firefighter, fire, attack, victim, driver, disaster, tower, zzz_new_york_city
palestinian, zzz_israel, zzz_israeli, zzz_yasser_arafat, israeli, peace, israelis, zzz_west_bank, zzz_arab, leader	palestinian, attack, military, zzz_israel, terrorist, official, zzz_afghanistan, war, zzz_united_states, zzz_taliban
race, racing, won, track, horse, win, races, horses, winner, lap	race, zzz_olympic, zzz_tiger_wood, tour, racing, track, car, medal, driver, gold

Table 4: Full list of topics learned from the NYTimes dataset.