Kendrick Lo
klo@g.harvard.edu
CS181-S16

Assignment #5
Due: 5:00pm April 15, 2016

Collaborators: A. Lee, TA's (OHs)

Homework 5: EM for a Simple Topic Model

There is a mathematical component and a programming component to this homework. Please submit ONLY your PDF to Canvas, and push all of your work to your Github repository. If a question requires you to make any plots, please include those in the writeup.

Background: In this homework, you will implement a very simple kind of topic model. Latent Dirichlet allocation, as we discussed in class, is a topic model in which each document is composed of multiple topics. Here we will make a simplified version in which each document has just a single topic. As in LDA, the vocabulary will have V words and a topic will be a distribution over this vocabulary. Let's use K topics and the kth topic is a vector $\boldsymbol{\beta}_k$, where $\beta_{k,v} \geq 0$ and $\sum_v \beta_{k,v} = 1$. Each document can be described by a set of word counts $\boldsymbol{w}_d$, where $w_{d,v}$ is a nonnegative integer. Document d has N_d words in total, i.e., $\sum_v w_{d,v} = N_d$. Let's have the unknown overall mixing proportion of topics be $\boldsymbol{\theta}$, where $\theta_k \geq 0$ and $\sum_k \theta_k = 1$. Our generative model is that each of the D documents has a single topic $z_d \in \{1, \ldots, K\}$, drawn from $\boldsymbol{\theta}$; then, each of the words is drawn from $\boldsymbol{\beta}_{z_d}$.

Problem 1 (Complete Data Log Likelihood, 4 pts)
Write the complete-data log likelihood $\ln p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)$. It may be convenient to write z_d as a one-hot coded vector $\boldsymbol{z}_d$.

Solution

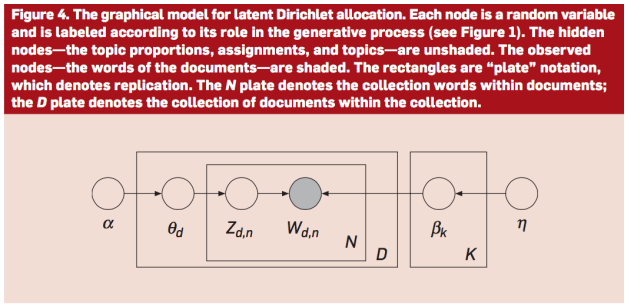

Figure 4. The graphical model for latent Dirichlet allocation. Each node is a random variable and is labeled according to its role in the generative process (see Figure 1). The hidden nodes—the topic proportions, assignments, and topics—are unshaded. The observed nodes—the words of the documents—are shaded. The rectangles are "plate" notation, which denotes replication. The *N* plate denotes the collection words within documents; the *D* plate denotes the collection of documents within the collection.

Source: Blei, D. "Probabilistic Topic Models"

First, we've imported a graphical model representation for Latent Dirichlet Allocation for reference. Our model is slightly different because it is simplified. For example, $z_{d,n}$ would simply be represented as z_d and appear outside of the box N because rather than having a topic assignment for each word in a document, we have assumed that each document contains words from only a single topic.

Furthermore, $\boldsymbol{\theta}_d$ can be represented merely as $\boldsymbol{\theta}$ and appear outside of the box D because rather than having per-document topic proportions, we only have an overall proportion of topics (i.e. the per-document topic proportion is always one). The node for $\boldsymbol{\theta}$ could, however, be enclosed in a plate K since $\boldsymbol{\theta}$ will take on a value for each of K topics.

Let's start with the probability term:

$$
\begin{aligned}
&p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) \\
&= p(\{\boldsymbol{w}_d\}_{d=1}^D \mid \{z_d\}_{d=1}^D, \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)\, p(\{z_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) && \text{product rule (Bishop 1.11)} \\
&= p(\{\boldsymbol{w}_d\}_{d=1}^D \mid \{z_d\}_{d=1}^D, \{\boldsymbol{\beta}_k\}_{k=1}^K)\, p(\{z_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) && \boldsymbol{w}_d \text{ is conditionally independent of } \boldsymbol{\theta} \\
&= p(\{\boldsymbol{w}_d\}_{d=1}^D \mid \{z_d\}_{d=1}^D, \{\boldsymbol{\beta}_k\}_{k=1}^K)\, p(\{z_d\}_{d=1}^D \mid \boldsymbol{\theta}) && z_d \text{ is conditionally independent of } \boldsymbol{\beta}_k \\
&= \prod_{d=1}^D \prod_{k=1}^K p(w_d \mid z_d = k, \beta_k)^{z_{d,k}} \prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}} && \text{one hot encoded vector, exponent trick} \\
&\propto \prod_{d=1}^D \prod_{k=1}^K \prod_{v=1}^V \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} \prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}} && \text{assume probability each word occurs is independent}
\end{aligned}
$$

We now take the logarithm of the above expression:

$$
\begin{aligned}
\log p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) &\propto \log \left[\prod_{d=1}^D \prod_{k=1}^K \prod_{v=1}^V \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} \prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}} \right] \\
&= \log \left[\prod_{d=1}^D \prod_{k=1}^K \prod_{v=1}^V \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} \right] + \log \left[\prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}} \right] + C && \text{where C is a constant} \\
&= \sum_{d=1}^D \sum_{k=1}^K \sum_{v=1}^V \log \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} + \sum_{d=1}^D \sum_{k=1}^K \log \theta_k^{z_{d,k}} + C && \text{log of products = sum of logs} \\
&= \sum_{d=1}^D \sum_{k=1}^K \sum_{v=1}^V z_{d,k} \log \beta_{k,v}^{w_{d,v}} + \sum_{d=1}^D \sum_{k=1}^K z_{d,k} \log \theta_k + C \\
&= \sum_{d=1}^D \sum_{k=1}^K \sum_{v=1}^V z_{d,k}\, w_{d,v} \log \beta_{k,v} + \sum_{d=1}^D \sum_{k=1}^K z_{d,k} \log \theta_k + C
\end{aligned}
$$

Problem 2 (Expectation Step, 5pts)

Introduce estimates $q(\boldsymbol{z}_d)$ for the posterior over the hidden variables $\boldsymbol{z}_d$. What did you choose and why? Write down how you would determine the parameters of these estimates, given the observed data $\{\boldsymbol{w}_d\}_{d=1}^D$ and the parameters $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^K$.

Solution

There is a bit of confusion between the use of Q in Bishop Chapter 9.3, and q as used here. We assume the questions asks us to derive the posterior: i.e. $q(\boldsymbol{z}_d) = (p(\{z_d\}_{d=1}^D \mid \{\boldsymbol{w}_d\}_{d=1}^D, \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)$. As noted in the "General EM Algorithm" in Bishop (pp. 440-441), evaluating this posterior distribution is part of the "E step".

Note that we can apply the product rule again to the expression we started with at the beginning of Problem 1:

$$
\begin{aligned}
&p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) \\
&= p(\{z_d\}_{d=1}^D \mid \{\boldsymbol{w}_d\}_{d=1}^D, \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)\, p(\{\boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K) \quad \text{product rule (Bishop 1.11)}
\end{aligned}
$$

Rearranging this expression, and substituting for $q(\boldsymbol{z}_d)$, we get:

$$
\begin{aligned}
q(\boldsymbol{z}_d) &= \frac{p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)}{p(\{\boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K)} \\
&\propto p(\{z_d, \boldsymbol{w}_d\}_{d=1}^D \mid \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K), \qquad \{\boldsymbol{w}_d\}_{d=1}^D, \boldsymbol{\theta}, \{\boldsymbol{\beta}_k\}_{k=1}^K \text{ all known or fixed} \\
&\propto \prod_{d=1}^D \prod_{k=1}^K \prod_{v=1}^V \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} \prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}} \quad \text{from Problem 1}
\end{aligned}
$$

We use the expression for the probability derived in Problem 1 to estimate q. We choose starting values for our parameters $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^K$, and subsequently we use updated parameters obtained from the M step of our EM algorithm. We use this posterior distribution to find the expectation of the complete-data log likelihood (e.g. Bishop 9.30) which we will derive in Problem 3.

Let's call our estimates $\gamma = q(\boldsymbol{z}_d)$ after being evaluated, using our fixed parameters. Since z is of dimension d x k, then so is q, and γ. Thus:

$$
\gamma \propto \prod_{d=1}^D \prod_{k=1}^K \prod_{v=1}^V \left(\beta_{k,v}^{w_{d,v}}\right)^{z_{d,k}} \prod_{d=1}^D \prod_{k=1}^K \theta_k^{z_{d,k}}
$$

.

Problem 3 (Maximization Step, 5pts)
With the $q(\boldsymbol{z}_d)$ estimates in hand from the E-step, derive an update for maximizing the expected complete data log likelihood in terms of $\boldsymbol{\theta}$ and $\{\boldsymbol{\beta}_k\}_{k=1}^K$.

(a) Derive an expression for the expected complete data log likelihood for fixed γ's.

(b) Find a value of θ that maximizes the expected complete data log likelihood derived in (a). You may find it helpful to use Lagrange multipliers in order to force the constraint $\sum \theta_k = 1$. Why does this optimized θ make intuitive sense?

(c) Apply a similar argument to find the value of $\beta_{k,v}$ that maximizes the expected complete data log likelihood.

Solution

Using the complete data log likelihood derived in Problem 1, the expected complete data log likelihood (Q) is:

$$
\begin{aligned}
&E\left[\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} z_{d,k}\, w_{d,v} \log \beta_{k,v} + \sum_{d=1}^{D}\sum_{k=1}^{K} z_{d,k} \log \theta_k + C\right] \\
&= E[\sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} z_{d,k}\, w_{d,v} \log \beta_{k,v}] + E[\sum_{d=1}^{D}\sum_{k=1}^{K} z_{d,k} \log \theta_k] + C \quad \text{linearity} \\
&= \sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} \gamma_{d,k}\, w_{d,v} \log \beta_{k,v} + \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{d,k} \log \theta_k + C \qquad \text{previous parameters fixed, replace } 0/1 \text{ with probabilities}
\end{aligned}
$$

We now optimize to get an updated θ_k in the above expression:

$$
\begin{aligned}
L &= \sum_{d=1}^{D}\sum_{k=1}^{K}\sum_{v=1}^{V} \gamma_{d,k}\, w_{d,v} \log \beta_{k,v} + \sum_{d=1}^{D}\sum_{k=1}^{K} \gamma_{d,k} \log \theta_k + \lambda(\sum_{k=1}^{K} \theta_k - 1) + C \quad \text{using Lagrange multiplier} \\
\frac{\partial L}{d\theta_k} &= \sum_{d=1}^{D} \frac{\gamma_{d,k}}{\theta_k} + \lambda = 0 \\
\lambda &= -\frac{1}{\theta_k} \sum_{d=1}^{D} \gamma_{d,k} \\
\theta_k &= \frac{-1}{\lambda} \sum_{d=1}^{D} \gamma_{d,k}
\end{aligned}
$$

So by summing both sides over k and applying the constraint:

$$
\begin{aligned}
\sum_{k=1}^{K} \theta_k &= \frac{-1}{\lambda} \sum_{k=1}^{K}\sum_{d=1}^{D} \gamma_{d,k} \\
1 &= \frac{-1}{\lambda} \sum_{k=1}^{K}\sum_{d=1}^{D} \gamma_{d,k} = \frac{-D}{\lambda} \quad \text{D rows of probabilities over K classes, sums to 1} \\
\lambda &= -D
\end{aligned}
$$

Therefore, the optimal $\theta_k = \frac{\sum_{d=1}^{D} \gamma_{d,k}}{D}$, which means the best new estimate for the overall proportion of topic k is the average posterior probability that a document relates to topic K averaged over all D documents.

We now optimize to get an updated $\beta_{k,v}$ in the expression from the first part, noting a similar constraint, $\sum_{v=1}^{V} \beta_{k,v} = 1$, can be set up where the sum of probabilities that a word comes from a particular class sums to one over the entire vocabulary V:

$$L = \sum_{d=1}^{D} \sum_{k=1}^{K} \sum_{v=1}^{V} \gamma_{d,k}\, w_{d,v} \log \beta_{k,v} + \sum_{d=1}^{D} \sum_{k=1}^{K} \gamma_{d,k} \log \theta_k + \lambda\Big(\sum_{v=1}^{V} \beta_{k,v} - 1\Big) + C \quad \text{using Lagrange multiplier}$$

$$\frac{\partial L}{d\beta_{k,v}} = \sum_{d=1}^{D} \frac{\gamma_{d,k}\, w_{d,v}}{\beta_{k,v}} + \lambda = 0$$

$$\lambda = -\frac{1}{\beta_{k,v}} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

$$\beta_{k,v} = \frac{-1}{\lambda} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

So by summing both sides over v and applying the constraint:

$$\sum_{v=1}^{V} \beta_{k,v} = \frac{-1}{\lambda} \sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

$$1 = \frac{-1}{\lambda} \sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

$$\lambda = -\sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

Therefore, the optimal $\beta_{k,v}$ is given by substituting this value of λ:

$$\beta_{k,v} = \frac{-1}{\lambda} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}$$

$$= \frac{\sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}}{\sum_{v=1}^{V} \sum_{d=1}^{D} \gamma_{d,k}\, w_{d,v}}$$

Problem 4 (Implementation, 10pts)

Implement this expectation maximization algorithm and try it out on some text data. In order for the EM algorithm to work, you may have to do a little preprocessing.

The starter code loads the text data as a numpy array that is 5224951×3 in size. As shown below, the first number in the numpy array represents the document_id, the second number represents a word_id, and the third number is the count the word appears.

[doc_id, word_id, count]

A dictionary of the mappings between word_ids and words is also provided. The full dataset description can be found at http://kdd.ics.uci.edu/databases/nsfabs/nsfawards.data.html.

Plot the objective function as a function of iteration and verify that it never increases. Try different numbers of topics and report what topics you find by, e.g., listing the most likely words.

Solution

We chose to minimize an objective function that is the negative of the expected complete data log likelihood function derived in Problem 3. Here are a few plots of the objective function as a function of iteration number, for a few different number of topics. We also provide a listing of the most likely words. However, what would be more interesting for future investigation is to find a number of topics that reasonably look somewhat distinct, and then use our model to determine the most likely topic for each document (but we were not asked to do this).

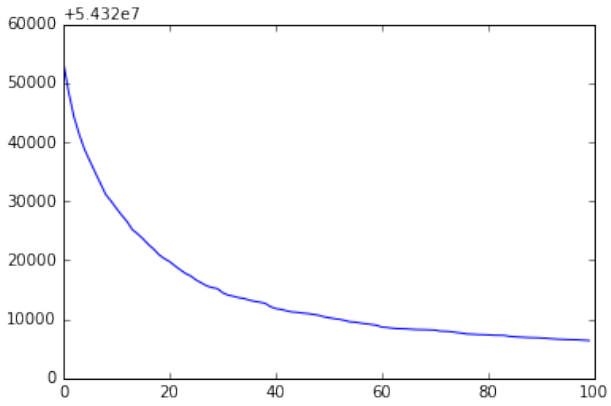

Figure 1: 10 topics, plot of negative expected complete data log likelihood

```
topic 0
['protein', 'cell', 'cells', 'proteins', 'gene',
'research', 'genes', 'molecular', 'dna', 'studies',
'specific', 'function', 'system', 'expression', 'development']
topic 1
['species', 'research', 'study', 'genetic', 'studies',
'populations', 'plant', 'understanding', 'important', 'project',
'provide', 'plants', 'data', 'population', 'dna']
topic 2
['students', 'research', 'science', 'project', 'program',
```

```
'laboratory', 'university', 'engineering', 'computer', 'mathematics',
'teachers', 'faculty', 'education', 'school', 'courses']
topic 3
['theory', 'research', 'problems', 'study', 'equations',
'systems', 'project', 'mathematical', 'work', 'methods',
'analysis', 'geometry', 'differential', 'mathematics', 'time']
topic 4
['research', 'project', 'study', 'data', 'social',
'economic', 'important', 'understanding', 'information', 'political',
'model', 'work', 'provide', 'policy', 'analysis']
topic 5
['research', 'university', 'support', 'project', 'science',
'data', 'program', 'conference', 'scientists', 'dr',
'workshop', 'provide', 'international', 'award', 'scientific']
topic 6
['research', 'materials', 'high', 'properties', 'chemistry',
'project', 'study', 'systems', 'phase', 'program',
'surface', 'studies', 'metal', 'chemical', 'university']
topic 7
['data', 'study', 'research', 'project', 'ocean',
'provide', 'studies', 'ice', 'processes', 'water',
'climate', 'time', 'understanding', 'models', 'work']
topic 8
['research', 'systems', 'design', 'system', 'project',
'data', 'control', 'models', 'based', 'problems',
'time', 'algorithms', 'methods', 'performance', 'model']
topic 9
['research', 'data', 'study', 'project', 'model',
'studies', 'university', 'high', 'program', 'system',
'processes', 'understanding', 'measurements', 'models', 'field']
```

There are some discernible groupings: classes relating to genes, academic courses, mathematical analysis, chemistry, climate, modelling, and so on. Research comes up as a top word in every category; these documents might be related to different aspects or types of research.

We reran the algorithm for K=5 and K=15 classes.

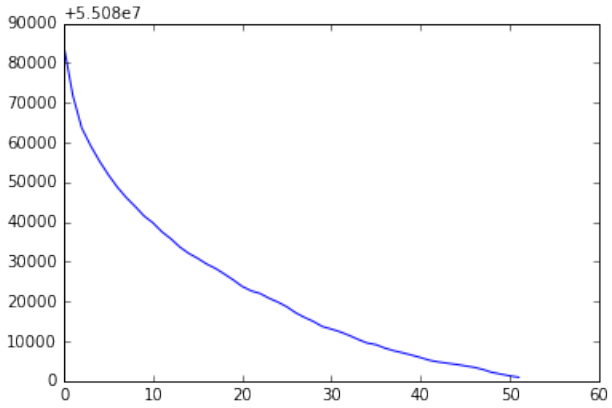

Figure 2: 5 topics, plot of negative expected complete data log likelihood

```
topic 0
['research', 'students', 'project', 'science', 'program',
'university', 'support', 'data', 'laboratory', 'engineering',
'development', 'provide', 'computer', 'education', 'study']
topic 1
['research', 'materials', 'high', 'project', 'chemistry',
'study', 'properties', 'university', 'program', 'systems',
'studies', 'phase', 'structure', 'chemical', 'surface']
topic 2
['research', 'theory', 'problems', 'systems', 'project',
'study', 'methods', 'models', 'analysis', 'work',
'design', 'system', 'data', 'algorithms', 'control']
topic 3
['research', 'species', 'protein', 'cell', 'cells',
'studies', 'study', 'gene', 'proteins', 'molecular',
understanding', 'plant', 'dna', 'genes', 'project']
topic 4
['research', 'data', 'project', 'study', 'provide',
'time', 'ocean', 'system', 'studies', 'model',
'models', 'processes', 'field', 'important', 'work']
```

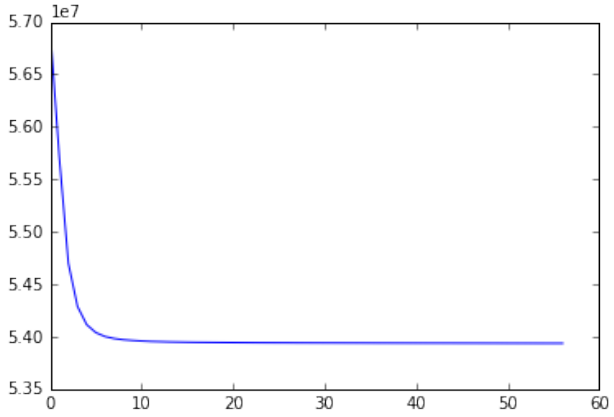

Figure 3: 15 topics, plot of negative expected complete data log likelihood

```
topic 0
['study', 'mantle', 'data', 'project', 'rocks',
'results', 'research', 'processes', 'models', 'crust',
'provide', 'studies', 'evolution', 'seismic', 'work']
topic 1
['protein', 'gene', 'dna', 'research', 'genes',
'molecular', 'proteins', 'cell', 'expression', 'plant',
'biology', 'studies', 'analysis', 'rna', 'specific']
topic 2
['research', 'data', 'project', 'study', 'ocean',
'provide', 'support', 'university', 'water', 'program',
'time', 'work', 'system', 'important', 'ice']
topic 3
['research', 'project', 'data', 'study', 'social',
'information', 'important', 'time', 'models', 'support',
'months', 'economic', 'understanding', 'model', 'development']
topic 4
['species', 'research', 'study', 'plant', 'project',
'studies', 'populations', 'important', 'understanding', 'effects',
'plants', 'growth', 'provide', 'determine', 'population']
topic 5
['research', 'materials', 'high', 'project', 'phase',
'properties', 'optical', 'system', 'process', 'applications',
'devices', 'development', 'systems', 'study', 'surface']
topic 6
['cells', 'cell', 'protein', 'proteins', 'system',
'research', 'studies', 'function', 'understanding', 'development',
'specific', 'mechanisms', 'brain', 'important', 'molecular']
topic 7
['research', 'university', 'students', 'program', 'science',
'support', 'project', 'award', 'engineering', 'conference',
'workshop', 'scientists', 'provide', 'faculty', 'national']
topic 8
['research', 'data', 'system', 'systems', 'project',
'design', 'network', 'performance', 'software', 'information',
```

```
'computer', 'high', 'based', 'development', 'parallel']
topic 9
['species', 'research', 'study', 'data', 'dna',
'genetic', 'project', 'molecular', 'studies', 'relationships',
'evolution', 'genes', 'important', 'group', 'evolutionary']
topic 10
['students', 'science', 'project', 'laboratory', 'research',
'program', 'teachers', 'mathematics', 'computer', 'courses', '
school', 'education', 'engineering', 'university', 'year']
topic 11
['research', 'data', 'model', 'study', 'models',
'project', 'flow', 'numerical', 'field', 'ice',
'waves', 'understanding', 'processes', 'surface', 'scale']
topic 12
['research', 'chemistry', 'study', 'project', 'studies',
'properties', 'program', 'systems', 'university', 'reactions',
'molecular', 'materials', 'structure', 'chemical', 'molecules']
topic 13
['theory', 'study', 'research', 'problems', 'equations',
'geometry', 'work', 'project', 'groups', 'systems',
'analysis', 'differential', 'mathematical', 'algebraic', 'mathematics']
topic 14
['research', 'systems', 'problems', 'control', 'design',
'methods', 'algorithms', 'project', 'system', 'models', '
theory', 'time', 'analysis', 'model', 'techniques']
```

The larger number of topics is much more interesting! We can really start to see some clearer separation here between different areas of research: geology, biology, oceanography, plants, materials, computers, mathematics, and so on. Very cool!

Problem 5 (Calibration, 1pt)
Approximately how long did this homework take you to complete? 30 hours