Text Mining – Assignment #3

Roger Cuscó, Matthew Sudmann-Day and Miquel Torrens

Exercise 1

Using 'State of the Union' speeches since 1975, we implemented a Gibbs Sampler over LDA with a topic count of $K = 5$.

See the code in:

`https://github.com/m-sudmann-day/BGSE-text-mining/tree/master/week3_hw/week3_hw.py`

In particular, the `Gibbs_LDA()` function allows custom analysis with a variety of parameters:

```
Gibbs_LDA(self, K, eta, theta, max_iterations, convergence_cutoff, output_path)
```

In the execution shown below, we look at the top 3 topics in each speech, and the top 10 words in each speech. Observing the topics subjectively, we can characterize them in the following way:

1. Topic 0 is touchy-feely.
2. Topic 1 is hard to characterize.
3. Topic 2 is economics.
4. Topic 3 is social.
5. Topic 4 is administrative.

Because speeches cover a great variety of topics, it is not too surprising that we do not see a huge difference in the topic focus by president or by party with the exception of Topic 0, "touchy-feely". This topic seems to be used far more often by Republican presidents, possibly to overcome an uncaring reputation, whereas Democrats possibly choose to avoid this topic for fear of appearing weak.

Below you will find the topics with their words and corresponding probabilities, and the topics with their corresponding topics and probabilities.

```
-------------------------
Topic 0
    Word: tonight (0.0138425328211)
    Word: unit (0.0142222453789)
    Word: effort (0.0148904741422)
    Word: support (0.0154579344854)
    Word: develop (0.0185808486204)
    Word: tax (0.0210238023105)
    Word: countri (0.0228463935966)
    Word: help (0.0262175800777)
    Word: world (0.0291946052189)
    Word: will (0.0497781881195)
-------------------------
```

```
Topic 1
    Word: million (0.0127524341615)
    Word: energi (0.0154751633326)
    Word: last (0.0160296948337)
    Word: propos (0.0163084229484)
    Word: continu (0.0181626211934)
    Word: must (0.0216611108564)
    Word: govern (0.0250912595654)
    Word: ha (0.0301336650546)
    Word: can (0.0323575547207)
    Word: american (0.0354875666911)
------------------------
Topic 2
    Word: economi (0.0122057518505)
    Word: polici (0.0124203635527)
    Word: care (0.0130456056266)
    Word: peac (0.0132504110008)
    Word: legisl (0.0145024802232)
    Word: budget (0.0152426669561)
    Word: job (0.0169291454836)
    Word: peopl (0.0316100146858)
    Word: congress (0.0317810009397)
    Word: thi (0.0809484308419)
------------------------
Topic 3
    Word: increas (0.0145624219651)
    Word: everi (0.0155762478113)
    Word: administr (0.0167316022607)
    Word: health (0.0174221107324)
    Word: feder (0.0187254211091)
    Word: secur (0.0190818292299)
    Word: need (0.0225778587459)
    Word: state (0.0243698198396)
    Word: america (0.0315127881759)
    Word: work (0.0339331955569)
------------------------
Topic 4
    Word: live (0.00954792748775)
    Word: system (0.0108633316491)
    Word: know (0.0120801907984)
    Word: it (0.0124684197955)
    Word: provid (0.0127560996449)
    Word: time (0.0150398115198)
    Word: program (0.0267241510243)
    Word: nation (0.0375147356445)
    Word: will (0.0408382083471)
    Word: year (0.0515177156739)
------------------------
Document 0, Ford 1975
    Topic 2 (5.21710997128e-05)
```

```
    Topic 4 (2.89606481994e-07)
    Topic 1 (6.93364619211e-23)
-------------------------
Document 1, Ford 1976
    Topic 4 (2.24757814555e-28)
    Topic 2 (7.25716271294e-30)
    Topic 1 (5.41876288071e-28)
-------------------------
Document 2, Ford 1977
    Topic 2 (3.52002722843e-10)
    Topic 3 (1.00994773181e-09)
    Topic 1 (1.90062129185e-61)
-------------------------
Document 3, Carter 1978
    Topic 2 (5.9636300199e-06)
    Topic 1 (4.32313518299e-22)
    Topic 4 (7.16023601172e-88)
-------------------------
Document 4, Carter 1978
    Topic 3 (0.000473572054718)
    Topic 1 (1.36105690374e-18)
    Topic 4 (1.03733728925e-44)
-------------------------
Document 5, Carter 1979
    Topic 2 (2.06819659894e-19)
    Topic 1 (2.84688299977e-07)
    Topic 4 (2.75223292674e-09)
-------------------------
Document 6, Carter 1979
    Topic 3 (8.75211421926e-06)
    Topic 1 (2.1239611115e-26)
    Topic 4 (1.68397649742e-09)
-------------------------
Document 7, Carter 1980
    Topic 3 (1.01179691175e-14)
    Topic 0 (1.52068260584e-14)
    Topic 1 (3.52461569328e-11)
-------------------------
Document 8, Carter 1980
    Topic 3 (2.41619036686e-05)
    Topic 1 (4.47496515514e-57)
    Topic 4 (4.71536513793e-06)
-------------------------
Document 9, Carter 1981
    Topic 1 (2.03499486422e-07)
    Topic 3 (2.89358543428e-14)
    Topic 4 (1.392826838e-31)
-------------------------
Document 10, Reagan 1981
    Topic 1 (2.56604476926e-12)
```

```
    Topic 4 (4.09323726979e-05)
    Topic 3 (2.8662040351e-09)
------------------------
Document 11, Reagan 1982
    Topic 3 (3.49521404043e-45)
    Topic 1 (1.82983344374e-33)
    Topic 4 (6.61374648361e-05)
------------------------
Document 12, Reagan 1983
    Topic 3 (5.23950560703e-49)
    Topic 1 (1.06277226809e-14)
    Topic 4 (8.83379502204e-08)
------------------------
Document 13, Reagan 1984
    Topic 3 (8.3529000745e-05)
    Topic 2 (9.37950227013e-07)
    Topic 1 (9.14393663195e-27)
------------------------
Document 14, Reagan 1985
    Topic 0 (2.14796137215e-21)
    Topic 1 (0.000104238063118)
    Topic 4 (1.32571709297e-25)
------------------------
Document 15, Reagan 1986
    Topic 0 (5.83531244441e-31)
    Topic 3 (1.72677579139e-06)
    Topic 4 (1.94019890484e-09)
------------------------
Document 16, Reagan 1987
    Topic 1 (7.37081279104e-06)
    Topic 3 (1.11723711405e-14)
    Topic 4 (5.82751752384e-08)
------------------------
Document 17, Reagan 1988
    Topic 1 (8.38691197031e-08)
    Topic 4 (1.46190341751e-26)
    Topic 3 (1.29564938771e-08)
------------------------
Document 18, Bush 1989
    Topic 3 (1.13975725205e-37)
    Topic 1 (4.5998319138e-12)
    Topic 4 (2.99632431197e-05)
------------------------
Document 19, Bush 1990
    Topic 2 (0.000287750939327)
    Topic 3 (2.71315808847e-23)
    Topic 4 (1.52540906851e-08)
------------------------
Document 20, Bush 1991
    Topic 1 (3.94024711564e-05)
```

```
    Topic 3 (2.2341768174e-06)
    Topic 4 (3.4911094629e-19)
-------------------------
Document 21, Bush 1992
    Topic 0 (5.66667385105e-06)
    Topic 1 (3.05066077685e-16)
    Topic 4 (8.6683463161e-16)
-------------------------
Document 22, Clinton 1993
    Topic 2 (2.08838675255e-42)
    Topic 1 (0.000678064600232)
    Topic 4 (2.14539547923e-13)
-------------------------
Document 23, Clinton 1994
    Topic 2 (2.23088883035e-14)
    Topic 4 (1.47651770591e-07)
    Topic 3 (1.42053237253e-09)
-------------------------
Document 24, Clinton 1995
    Topic 3 (1.29538123415e-21)
    Topic 1 (0.0250912595654)
    Topic 4 (3.86306608397e-10)
-------------------------
Document 25, Clinton 1996
    Topic 4 (1.89846613181e-39)
    Topic 3 (4.94424415518e-05)
    Topic 1 (0.00016459265926)
-------------------------
Document 26, Clinton 1997
    Topic 1 (1.89533870596e-45)
    Topic 4 (3.03677634506e-36)
    Topic 3 (1.16829269583e-42)
-------------------------
Document 27, Clinton 1998
    Topic 3 (1.93070041358e-10)
    Topic 4 (2.05159216671e-07)
    Topic 1 (2.35592735424e-30)
-------------------------
Document 28, Clinton 1999
    Topic 1 (2.69574844895e-06)
    Topic 4 (2.46067694194e-12)
    Topic 3 (0.000510643699604)
-------------------------
Document 29, Clinton 2000
    Topic 2 (9.10522218273e-15)
    Topic 4 (4.63632442885e-06)
    Topic 1 (2.12785100987e-78)
-------------------------
Document 30, Bush 2001
    Topic 4 (2.59005396828e-19)
```

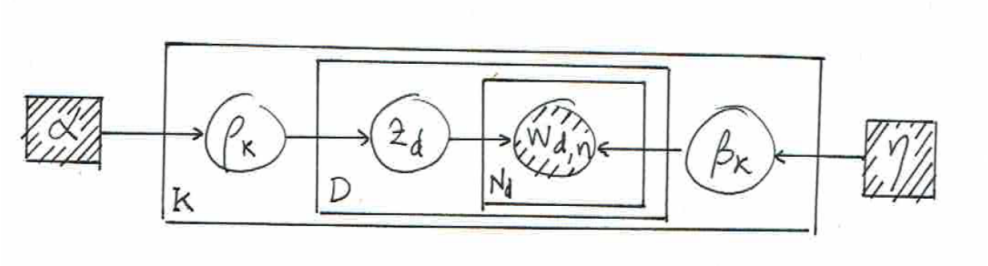
```
    Topic 3 (3.32873596812e-50)
    Topic 1 (8.79959585506e-05)
------------------------
Document 31, Bush 2002
    Topic 4 (5.41525140838e-54)
    Topic 1 (2.8936992362e-11)
    Topic 3 (3.81049578036e-11)
------------------------
Document 32, Bush 2003
    Topic 1 (3.69967520921e-05)
    Topic 4 (1.2165506659e-07)
    Topic 0 (1.66704697814e-05)
------------------------
Document 33, Bush 2004
    Topic 3 (3.30593654347e-07)
    Topic 0 (9.80323138259e-06)
    Topic 1 (5.61621785833e-43)
------------------------
Document 34, Bush 2005
    Topic 1 (8.99100510642e-12)
    Topic 4 (2.54064266393e-19)
    Topic 0 (3.91494149536e-05)
------------------------
Document 35, Bush 2006
    Topic 0 (0.00120116780105)
    Topic 1 (6.62006434249e-31)
    Topic 4 (1.85198785851e-05)
------------------------
Document 36, Bush 2007
    Topic 4 (6.73154830198e-18)
    Topic 1 (1.26544051713e-23)
    Topic 2 (3.33707240699e-08)
------------------------
Document 37, Bush 2008
    Topic 1 (2.87796095088e-08)
    Topic 4 (6.11987090709e-10)
    Topic 2 (9.84494017518e-07)
------------------------
Document 38, Obama 2009
    Topic 2 (9.19889926742e-59)
    Topic 1 (0.000792131365914)
    Topic 4 (6.32597706248e-26)
------------------------
Document 39, Obama 2010
    Topic 4 (9.87631108148e-27)
    Topic 2 (5.64013946291e-09)
    Topic 1 (1.77542615851e-20)
------------------------
Document 40, Obama 2011
    Topic 2 (9.51221029442e-64)
```

```
    Topic 1 (0.000134032178743)
    Topic 4 (7.71538760439e-24)
-------------------------
Document 41, Obama 2012
    Topic 3 (3.1998517289e-24)
    Topic 1 (1.00508994226e-11)
    Topic 4 (6.16792357698e-06)
-------------------------
Document 42, Obama 2013
    Topic 4 (2.29885716764e-14)
    Topic 3 (1.24108408638e-54)
    Topic 1 (3.80317661336e-27)
-------------------------
Document 43, Obama 2014
    Topic 2 (0.0152426669561)
    Topic 1 (1.54795221569e-45)
    Topic 3 (1.91810631821e-24)
```

Exercise 2

Part (a)

The directed graph is the following:



Part (b)

The Markov blankets of these elemets of the model can be expressed as follows:

- Words in document d: topic assignments z_d (parent) and topics β_k (parent).
- Topic assignment z_d: topic probabilities ρ_k (parent), the set of words $w_{d,n}$ (children) and topics β_k (children's parent).
- Topics β_k: hyperparameter η (parent), the set of words $w_{d,n}$ (children) and topic assignment z_d (children's parent).

Part (c)

An uncollapsed Gibbs algorithm could be the following:

1. Set values for $\boldsymbol{\eta} \in \mathbb{R}^V$ and $\boldsymbol{\alpha} \in \mathbb{R}^K$
2. Draw for each topic $k \in \{1, \ldots, K\}$ a sample $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta}) \in \Delta^{V-1}$
3. Draw a sample $\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\alpha}) \in \Delta^{K-1}$ that specifies the likelihood of each topic
4. Draw for each document $d \in \{1, \ldots, D\}$ a sample $z_d \sim \text{multinom}\,(\boldsymbol{\rho})$
5. Draw for each word $n \in \{1, \ldots, N_d\}$ in document d the word $w_{d,n} \sim \text{multinom}\left(\boldsymbol{\beta}_{z_d}\right)$
6. Update for each k the vector $\boldsymbol{\beta}_k \sim \text{Dir}(\boldsymbol{\eta} + \mathbf{m}_k) \in \Delta^{V-1}$, where element v of vector $\mathbf{m}_k \in \mathbb{R}^V$ is $m_{k,v}$, the number of times topic k generates word v.
7. Update the vector $\boldsymbol{\rho} \sim \text{Dir}(\boldsymbol{\alpha} + \boldsymbol{\delta}) \in \Delta^{K-1}$, where element k of vector $\boldsymbol{\delta} \in \mathbb{R}^K$ is δ_k, the number of documents that are assigned topic k.
8. Return to step 4 and repeat until convergence.