A Java implementation of a Gibbs Sampler for Latent Dirichlet Allocation for the Weka data mining framework

Leif Jonsson

January 18, 2013

Abstract

This paper describes a Java implementation of the Latent Dirichlet Allocation algorithm proposed by Blei. et. al. [1]. The implementation uses the Gibbs sampling technique described by Porteous et. al. [4]. Please observe that the implementation here is of the original Gibbs sampler described in Algorithm 3.1 in the paper and **not** the Fast version described in Algorithm 4.1. The nomenclature in the discussion below will follow that of Porteous et. al. [4] as close as possible. This implementation is adapted to be used in the Java based Weka [3] framework for data mining.

1 Introduction

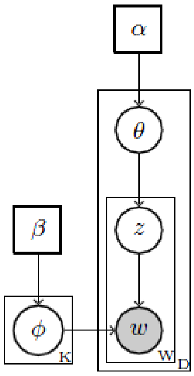
The Latent Dirichlet Allocation (LDA) and its derivatives has become a popular model for modeling text due to a couple of main reasons. First the model is a fully specified probabilistic model which is easy to specify in terms of a hierarchical model and easy to visualize in a probabilistic graphical model. It also gives quite good results in capturing clustering properties of text corpora.

2 LDA

Latent Dirichlet Allocation is a probabilistic hierarchical model of collections of discrete data, for instance text documents. The LDA model is a probabilistic generative model in three layers which models how, for instance, a text document is generated. But the LDA model has also been used to annotate images and many other applications.

The model has two hyper parameters α and β which controls the model. These hyper parameters can be either scalars or vector quantities. In this implementation we use the scalars for both α and β. The model is often

Figure 1: Graphical model of LDA



described as a *topic model* due to the fact that it tends to find topics in document collections. It is also an instance of a *clustering technique*. Further it can also be used as a *dimensionality reduction* technique. A graphical model of LDA can be seen in Fig. 1

The generative process for a document j in the LDA model is as follows:

```
1. Choose N_j ~ Poisson(ζ)
2. Choose θ ~ Dirichlet(α)
3. For each of the N_j words x_ij in document j:
        (a). Choose a topic z_ij ~ Multinomial(θ)
        (b). Choose a word x_ij ~ Multinomial(ϕ_{z_ij})
```

In step 1, we draw a scalar N_j from a Poisson distribution which controls how many words (N_j) document j should contain (this step is sometimes omitted, for instance in [4]). In step 2 we draw θ which represents the distribution over topics in the document. θ is drawn from a Dirichlet distribution parameterized by α. α can either be a scalar in which case we get a uniform Dirichlet or a vector of integer values in which case we get a non-uniform Dirichlet which can represent prior information that we have about the topic distribution.

Now, to generate the words in the document j we go to step 3. In 3.a we first draw a topic z_{ij} from a Multinomial distribution parametrized by the θ drawn in step 2. Since we just do one draw, we get a vector in which one element has the value 1 and the others are 0. This is a so called 1-of-K encoded vector where the topic z_{ij} drawn is represented by the index in the drawn vector which has its element set to one.

To generate the word given the topic z_{ij} and β we simply index the z_{ij}:th row in the ϕ matrix attained by sampling from a Dirichlet as many times as we have topics. The length of each of the Dirichlet samples will be as many words as we have in our dictionary. This means that if we have K topics and W words in our dictionary, ϕ will be a K x W matrix. Each topic is a probability distribution over the words in our dictionary. The word x_{ij} is generated by again drawing a 1-of-K vector from the Multinomial parametricized by Dirichlet sample in ϕ at row z_{ij}.

3 Gibbs Sampling for Inference in LDA

One of the first descriptions of a Gibbs sampler for LDA was by Griffiths et. al [2]. They describe a collapsed Gibbs sampler for LDA where the ϕ and θ is not sampled directly but are "collapsed" (integrated out) and are later calculated from samples from the posterior distribution of the topics p($z_{ij}|w$) where $w = x_{ij}$, which are sufficient statistics for ϕ and θ.

This algorithm in [2] is the same as Algorithm 3.1 described in Porteous et. al. [4].

In mathematical notation the joint probability of the LDA model becomes:

$$p(w, z, \theta, \phi|\alpha, \beta) = p(\phi|\beta)p(\theta|\alpha)p(z|\theta)p(w|\phi_z) \tag{1}$$

In the inference stage we are interested in the posterior probability of the parameters θ and ϕ given the observed words, α and β, that is:

$$p(\theta, \phi_z, z|w, \alpha, \beta) = \frac{p(w, z, \theta, \phi|\alpha, \beta)}{p(w|\alpha, \beta)} \tag{2}$$

The above posterior is intractable in closed form, but it is possible to construct approximate algorithms for it, one of which is the collapsed Gibbs sampler. From these equations and the formulations for the Dirichlet and the Multinomial and the fact that they are conjugate it is possible to derive the probability of a topic given the previous state of the Markov Chain, α, β and an observed word (x_{ij}).

$$p(z_{ij} = k|z^{\neg ij}, x_{ij}, \alpha, \beta) = \frac{\frac{(N_{kj}^{\neg ij}+\alpha)(N_{x_{ij}k}^{\neg ij}+\beta)}{(N_k^{\neg ij}+W\beta)}}{Z} \tag{3}$$

This equation is implemented in the "sample" method in the code listing.

Given the samples of z_{ij} we can calculate $\hat{\phi}$ and $\hat{\theta}$ using the following equations:

$$\hat{\phi}_{wk} = \frac{N_{wk} + \beta}{N_k + W\beta} \tag{4}$$

$$\hat{\theta}_{kj} = \frac{N_{kj} + \alpha}{N_j + K\alpha} \tag{5}$$

These equations are implemented in the "saveSample" method on the below code listing.

4 Weka Gibbs Sampler Implementation

In the first part of the implementation, the Gibbs Sampler is randomly initiated. The main work of the algorithm is then implemented in the "sample" and "saveSample" methods. The "sample" methods samples z_{ij} (which is denoted K_ij[i][j] in the below code) conditioned on the previous state of the chain, α and β. From the z_{ij} samples and the other counts we can then calculate the desired quantities $\hat{\phi}$ and $\hat{\theta}$, this is done in the "saveSample" method.

```
void sample() {
    // The following 3 loops represent (for i <- 1 to N in the paper)
    // Loop over all documents
    for( int j = 0; j < J; j++ ) {
        Instance instance_j = data.get(j);
        // Loop over all words in document
        int instanceWords = instance_j.numValues();
        int i = 0;
        // Fetch the iindex wordfrequency in the document
        for( int iindex = 0; iindex < instanceWords; iindex++) {
            int w = instance_j.index(iindex);
            int wordfreq = (int)Math.ceil(instance_j.value(instance_j.index(iindex)));
            // The word occurs wordfreq times
            for( int frequency = 0; frequency < wordfreq; frequency++) {
                // Fetch the current topic assignment for doc j
                int k = Kji[j][i];
                // Remove the topic (z_i) from the previous assignments
                // to get what is called Nkj^(-ij) in the paper
                Nk[k]--;
                Nwk[w][k]--;
                Nkj[k][j]--;
                // For each topic (for k <- 1 to K in paper)
                double PK = 0;
                double[] pks = new double[K];
                for(k = 0;k < K;k++) {
                    // Formula in Algoritm 3.1 in paper. (F1)
                    // Accumulate PK
                    PK += pks[k] =
                        ((double)Nkj[k][j] + alpha) * ((double)Nwk[w][k] + beta)
//                                 ------------------------------------------------
                                  / ((double)Nk[k] + ((double)W * beta) );
                }

                // This loop corresponds to P[k] <- P[k-1] + (F1)
                for (int k_tmp = 1; k_tmp < K; k_tmp++) {
                    pks[k_tmp] += pks[k_tmp - 1];
                }
                // Draw from Uniform[0,1]
                double u = randomizer.nextDouble();
                // Second (for k <- 1 to K in paper)
                for(k = 0;k < K;k++)
                    if( u < (pks[k]/PK))
                        break;
                if(k == K) k--;
                Nk[k]++;
                Nwk[w][k]++;
                Nkj[k][j]++;
                Kji[j][i] = k;
                i++; // On to the next word
            }
        }
    }
}


void saveSample() {
    // Calculate Phi sample according to formula in Ref. (1)
    for(int k = 0;k < K;k++) {
        for(int w = 0;w < W;w++) {
            Phi_wk[w][k] += ((double)Nwk[w][k] + beta)
//                          ---------------------------------------
                          / ((double)Nk[k] + (double) W * beta);
        }
    }
    // Calculate Theta sample according to formula in Ref. (1)
    for(int j = 0;j < J;j++) {
        for(int k = 0;k < K;k++) {
            Theta_kj[k][j] +=        ((double)Nkj[k][j] + alpha)
//                                ---------------------------------------
                                / ((double)Nj[j] + (double) K * alpha);
        }
    }
    draws++; // We have drawn another sample
}
```

5 Evaluation

To evaluate the sampler we use a popular data set called the “20 newsgroups” data set which contains news postings from 20 USENET newsgroups. This data set is available from `http://qwone.com/~jason/20Newsgroups/`.

The program is run with the following commandline:

```
java -jar CollapsedLDA.jar 10 50.0 0.01 100 1 500 20news-bydate-test
```

The main implementation code is in the file:

```
code/src/weka/clusterers/CollapsedGibbsSamplerLDA.java
```

We load the news postings from the newsgroups with the Weka class “TextDirectoryLoader”. In this process the “classlabel” of the document is assigned the name of the directory the document is stored in. For instance if a document is in the folder “sci.med” the class label of that document will be “sci.med”.

The text instances are then converted to word frequencies and stop words are removed. We then run our Gibbs sampler on the resulting data set. The setting for the sampler are as follows:

α	β	No. topics	No. burn-in samples	No. samples	Lag
$\frac{50}{topics}$	0.01	20	4000	20 000	1

“Lag” means the number of iterations of the chain between each sample is taken, this is to avoid correlated samples.

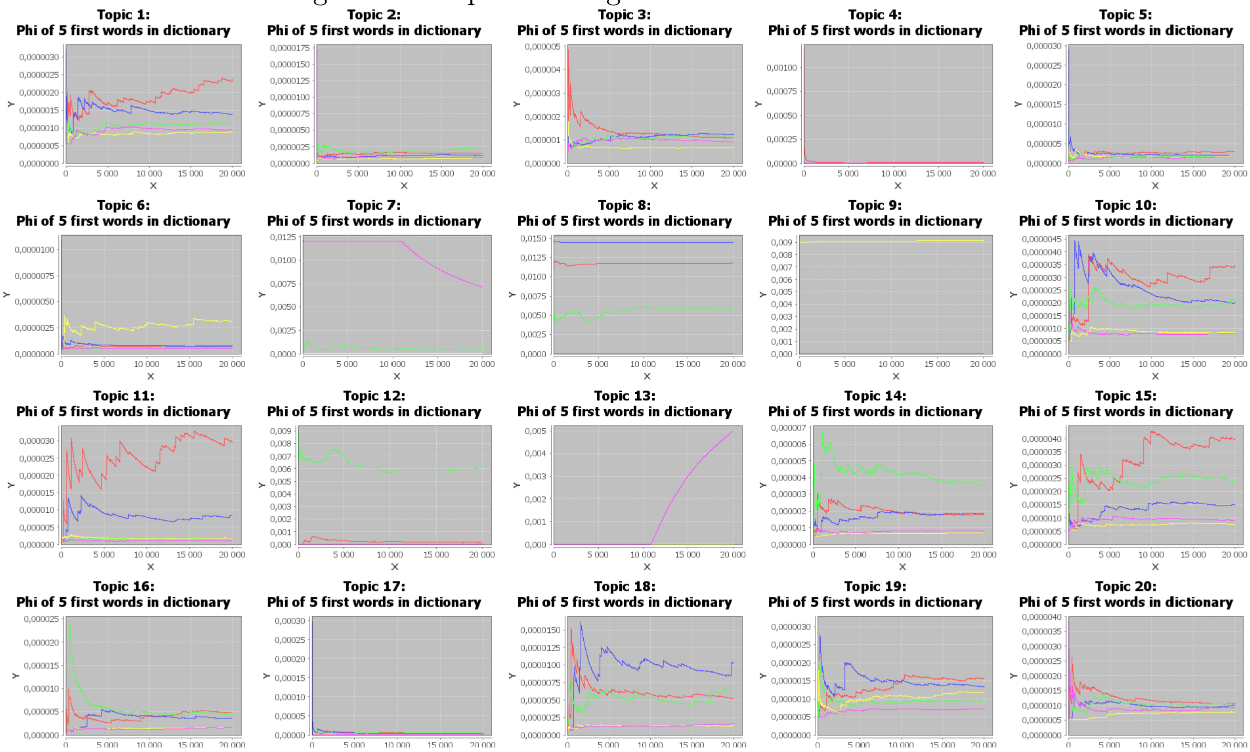
Due to the high dimensionality of the sampled quantities $\hat{\theta}$ (K x J = number of topics times number of documents) and $\hat{\phi}$ (W x K = number of words in dictionary times number of topics) it is too memory demanding to save the samples of $\hat{\phi}$ and $\hat{\theta}$ for study. What we do to get some feeling for the convergence of the sampler is that we study the average (over the number of draws) posterior probability of $\hat{\phi}$ of five words in each of the selected topics.

We can also study the actual word to topic assignments to see if the topics seem to be reasonable.

The convergence properties of the sampler can be seen in Fig. 2. The figures shows the probabilities of the first five words for each topic after 20 000 samples after a burn in period of 4000 and a lag of 1. It seems from visual inspection of Fig. 2 that the Markov Chain is converging.

In the tables in the following pages we show the 20 topics and under the topic column we list the top 30 words from each topic. In the column “Sample classes” we show the document class of the 30 documents with the highest probability under that particular topic.

Figure 2: Sampler convergence



Topic 1	Sample classes	Topic 2	Sample classes	Topic 3	Sample classes
file	comp.graphics	windows	comp.sys.mac.hardware	people	soc.religion.christian
image	comp.graphics	drive	misc.forsale	evidence	talk.religion.misc
version	comp.windows.x	card	comp.sys.ibm.pc.hardware	religion	soc.religion.christian
program	comp.graphics	dos	comp.windows.x	true	alt.atheism
files	comp.graphics	pc	comp.os.ms-windows.misc	claim	alt.atheism
ftp	comp.os.ms-windows.misc	apple	comp.sys.ibm.pc.hardware	agree	alt.atheism
graphics	comp.windows.x	mac	comp.sys.ibm.pc.hardware	paul	soc.religion.christian
jpeg	comp.graphics	software	comp.sys.ibm.pc.hardware	homosexuality	alt.atheism
images	comp.graphics	disk	comp.sys.ibm.pc.hardware	reason	soc.religion.christian
software	comp.os.ms-windows.misc	hard	comp.sys.ibm.pc.hardware	wrong	alt.atheism
format	comp.graphics	run	comp.os.ms-windows.misc	argument	talk.religion.misc
display	comp.windows.x	running	comp.graphics	homosexual	alt.atheism
color	comp.graphics	video	comp.os.ms-windows.misc	word	soc.religion.christian
window	comp.graphics	monitor	misc.forsale	sex	alt.atheism
programs	comp.windows.x	scsi	comp.sys.ibm.pc.hardware	words	alt.atheism
user	comp.graphics	memory	comp.os.ms-windows.misc	question	alt.atheism
server	comp.sys.mac.hardware	network	comp.sys.ibm.pc.hardware	religious	soc.religion.christian
comp	comp.os.ms-windows.misc	board	comp.sys.ibm.pc.hardware	statement	alt.atheism
gif	comp.graphics	cpu	comp.sys.ibm.pc.hardware	moral	alt.atheism
free	comp.graphics	machine	comp.sys.ibm.pc.hardware	means	talk.politics.mideast
code	comp.graphics	driver	misc.forsale	term	alt.atheism
windows	comp.windows.x	bus	comp.sys.mac.hardware	meaning	alt.atheism
manager	comp.os.ms-windows.misc	modem	comp.sys.mac.hardware	sense	soc.religion.christian
application	comp.windows.x	microsoft	comp.os.ms-windows.misc	simply	alt.atheism
applications	comp.windows.x	drivers	comp.sys.ibm.pc.hardware	gay	talk.religion.misc
run	comp.windows.x	mouse	comp.sys.ibm.pc.hardware	makes	soc.religion.christian
directory	comp.os.ms-windows.misc	ram	comp.os.ms-windows.misc	definition	soc.religion.christian
quality	comp.graphics	ide	comp.sys.ibm.pc.hardware	based	comp.windows.x
screen	comp.windows.x	drives	comp.os.ms-windows.misc	exist	alt.atheism
xv	comp.windows.x	port	comp.sys.ibm.pc.hardware	belief	talk.politics.misc
Topic 4	Sample classes	Topic 5	Sample classes	Topic 6	Sample classes
children	talk.politics.misc	god	talk.religion.misc	data	misc.forsale
fbi	talk.politics.misc	jesus	soc.religion.christian	information	comp.graphics
people	talk.politics.misc	christian	soc.religion.christian	list	comp.graphics
president	talk.politics.misc	love	alt.atheism	systems	comp.graphics
gun	talk.politics.mideast	church	soc.religion.christian	send	rec.motorcycles
koresh	talk.politics.guns	christ	soc.religion.christian	mit	comp.windows.x
police	talk.politics.misc	life	soc.religion.christian	mail	comp.graphics
started	talk.politics.mideast	bible	soc.religion.christian	info	comp.os.ms-windows.misc
stratus	talk.politics.guns	sin	soc.religion.christian	address	comp.graphics
waco	talk.politics.guns	faith	soc.religion.christian	internet	comp.graphics
house	talk.politics.guns	christians	soc.religion.christian	fax	comp.graphics
batf	talk.politics.misc	lord	talk.religion.misc	e	comp.windows.x
clinton	talk.politics.guns	christianity	soc.religion.christian	unix	comp.windows.x
told	talk.politics.guns	rutgers	talk.religion.misc	contact	comp.graphics
happened	talk.politics.guns	people	talk.religion.misc	access	comp.sys.ibm.pc.hardware
day	talk.politics.guns	word	soc.religion.christian	command	comp.graphics
didn	talk.politics.guns	hate	soc.religion.christian	source	comp.sys.ibm.pc.hardware
guns	talk.politics.guns	father	soc.religion.christian	email	comp.windows.x
ms	talk.politics.guns	mary	talk.religion.misc	box	comp.graphics
home	talk.politics.guns	paul	soc.religion.christian	analysis	comp.windows.x
compound	talk.politics.guns	death	soc.religion.christian	package	comp.graphics
door	talk.religion.misc	son	soc.religion.christian	message	sci.crypt
don	talk.politics.misc	truth	soc.religion.christian	based	comp.sys.ibm.pc.hardware
kill	talk.politics.guns	heaven	soc.religion.christian	software	comp.graphics
david	talk.politics.misc	human	soc.religion.christian	remote	comp.windows.x
atf	talk.religion.misc	day	talk.religion.misc	include	rec.motorcycles
gas	talk.politics.guns	die	soc.religion.christian	conference	sci.crypt
decision	talk.politics.guns	may	talk.religion.misc	set	sci.space
building	talk.politics.misc	original	soc.religion.christian	digital	comp.windows.x
myers	talk.politics.guns	true	talk.religion.misc	including	comp.graphics

Topic 7	Sample classes	Topic 8	Sample classes	Topic 9	Sample classes
world	talk.politics.mideast	medical	sci.med	writes	rec.autos
israel	talk.politics.mideast	april	sci.med	article	sci.med
war	talk.politics.mideast	research	sci.med	edu	talk.politics.mideast
jews	talk.politics.mideast	care	sci.med	in	rec.autos
armenian	talk.politics.mideast	national	sci.med	i	alt.atheism
history	talk.politics.mideast	health	sci.med	the	talk.politics.mideast
jewish	talk.politics.mideast	american	sci.med	com	talk.politics.mideast
people	talk.politics.mideast	information	talk.politics.misc	to	talk.politics.misc
armenians	talk.politics.mideast	drug	talk.politics.misc	and	rec.sport.baseball
muslims	talk.politics.mideast	school	talk.politics.misc	of	rec.motorcycles
muslim	talk.politics.mideast	study	talk.politics.misc	a	alt.atheism
peace	talk.politics.mideast	public	sci.med	umd	rec.autos
turkish	talk.politics.mideast	months	sci.med	1993apr21	talk.religion.misc
york	talk.politics.mideast	test	talk.politics.misc	that	talk.religion.misc
uucp	talk.politics.mideast	cancer	talk.politics.misc	it	rec.autos
arab	talk.politics.mideast	insurance	talk.politics.misc	you	rec.autos
killed	talk.politics.mideast	effect	talk.politics.misc	if	talk.religion.misc
europe	talk.politics.mideast	disease	sci.med	i	talk.religion.misc
genocide	talk.politics.mideast	private	sci.med	1993apr22	rec.autos
armenia	talk.politics.mideast	drugs	sci.med	is	sci.crypt
party	talk.politics.mideast	women	sci.med	this	sci.crypt
israeli	talk.politics.mideast	treatment	sci.med	remember	talk.politics.mideast
population	talk.politics.mideast	results	talk.politics.misc	1993apr23	rec.autos
city	talk.politics.mideast	news	talk.politics.misc	but	talk.politics.misc
russian	talk.politics.mideast	patients	sci.med	in	rec.autos
political	talk.politics.mideast	medicine	sci.med	what	rec.autos
turkey	talk.politics.mideast	page	talk.politics.misc	doubt	rec.autos
army	talk.politics.mideast	total	sci.med	guess	talk.politics.mideast
dead	talk.politics.mideast	increase	sci.med	arizona	talk.religion.misc
government	talk.politics.mideast	effects	sci.med	1993apr20	rec.autos
Topic 10	Sample classes	Topic 11	Sample classes	Topic 12	Sample classes
game	rec.sport.hockey	read	misc.forsale	subject	sci.space
team	rec.sport.hockey	time	misc.forsale	space	sci.space
games	rec.sport.hockey	post	alt.atheism	gov	sci.space
win	rec.sport.hockey	book	misc.forsale	net	sci.space
won	rec.sport.hockey	questions	misc.forsale	nasa	sci.space
play	rec.sport.baseball	question	soc.religion.christian	technology	sci.space
buffalo	rec.sport.baseball	copy	misc.forsale	institute	sci.space
hit	rec.sport.baseball	books	alt.atheism	research	sci.space
hockey	rec.sport.baseball	reading	misc.forsale	center	sci.space
baseball	rec.sport.hockey	answer	misc.forsale	brian	sci.space
series	rec.sport.baseball	posting	soc.religion.christian	distribution	sci.space
boston	rec.sport.baseball	change	sci.electronics	toronto	sci.space
players	rec.sport.hockey	write	sci.med	access	sci.space
fans	rec.sport.hockey	paper	soc.religion.christian	design	sci.space
player	rec.sport.hockey	posted	misc.forsale	mil	sci.space
clutch	rec.sport.hockey	newsgroup	soc.religion.christian	pat	sci.space
lost	rec.sport.hockey	written	comp.windows.x	mission	sci.space
fan	rec.sport.baseball	original	alt.atheism	sci	sci.space
period	rec.sport.hockey	response	rec.sport.baseball	digex	sci.space
league	rec.sport.hockey	note	soc.religion.christian	caltech	sci.space
time	rec.sport.hockey	idea	sci.electronics	cost	sci.space
st	rec.sport.baseball	special	alt.atheism	acs	sci.space
san	rec.sport.baseball	discussion	comp.sys.ibm.pc.hardware	hst	sci.space
espn	rec.sport.hockey	news	sci.electronics	shuttle	sci.space
season	rec.sport.hockey	alt	sci.electronics	gatech	sci.space
night	rec.sport.baseball	mentioned	alt.atheism	ohio	sci.space
mike	rec.sport.hockey	text	soc.religion.christian	laboratory	sci.space
cornell	rec.sport.hockey	issue	alt.atheism	sky	sci.space
home	rec.sport.baseball	comments	talk.religion.misc	il	sci.space
average	rec.sport.hockey	similar	misc.forsale	lab	sci.space

Topic 13	Sample classes	Topic 14	Sample classes	Topic 15	Sample classes
ca	comp.sys.mac.hardware	writes	comp.os.ms-windows.misc	subject	comp.windows.x
subject	rec.sport.hockey	subject	rec.autos	ve	rec.motorcycles
distribution	rec.sport.hockey	article	rec.autos	cc	rec.sport.baseball
usa	rec.sport.hockey	edu	rec.motorcycles	org	rec.motorcycles
sale	misc.forsale	ibm	talk.politics.guns	sun	rec.motorcycles
apr	rec.motorcycles	com	talk.politics.guns	world	rec.motorcycles
price	rec.sport.hockey	hp	talk.politics.guns	heard	comp.windows.x
att	misc.forsale	uiuc	talk.politics.guns	distribution	comp.sys.mac.hardware
canada	misc.forsale	opinions	talk.politics.guns	sound	rec.motorcycles
gmt	misc.forsale	steve	sci.space	bike	rec.motorcycles
university	rec.motorcycles	mark	sci.electronics	stuff	rec.motorcycles
cwru	rec.motorcycles	cso	talk.politics.guns	columbia	rec.motorcycles
offer	rec.autos	james	talk.politics.guns	bob	rec.motorcycles
se	misc.forsale	jim	sci.crypt	ed	rec.motorcycles
sell	misc.forsale	indiana	sci.space	dave	rec.motorcycles
email	misc.forsale	mike	rec.autos	robert	sci.med
na	comp.os.ms-windows.misc	scott	rec.autos	rochester	rec.sport.baseball
buy	rec.autos	usa	sci.space	dod	talk.politics.guns
germany	rec.sport.baseball	corporation	sci.space	type	rec.motorcycles
sgi	sci.electronics	corp	rec.sport.baseball	friend	comp.sys.ibm.pc.hardware
thomas	rec.sport.hockey	illinois	sci.crypt	ma	rec.autos
cb	misc.forsale	disclaimer	talk.politics.guns	pretty	rec.motorcycles
wondering	alt.atheism	austin	sci.crypt	couple	rec.motorcycles
e	comp.sys.mac.hardware	ucs	comp.os.ms-windows.misc	error	rec.motorcycles
in	misc.forsale	mot	alt.atheism	std	rec.motorcycles
western	comp.windows.x	george	sci.crypt	road	misc.forsale
ca	misc.forsale	convex	talk.politics.misc	advice	rec.motorcycles
jon	sci.crypt	stephen	sci.space	haven	rec.motorcycles
cd	misc.forsale	expressed	talk.politics.misc	tv	rec.motorcycles
at	sci.electronics	necessarily	talk.politics.misc	bmw	comp.graphics
Topic 16	Sample classes	Topic 17	Sample classes	Topic 18	Sample classes
don	talk.religion.misc	people	talk.politics.misc	car	sci.space
doesn	talk.religion.misc	government	talk.politics.misc	power	talk.religion.misc
ll	talk.religion.misc	law	talk.politics.misc	time	comp.windows.x
ve	rec.sport.hockey	rights	alt.atheism	light	sci.electronics
lot	rec.sport.baseball	person	talk.religion.misc	earth	sci.electronics
isn	sci.med	free	talk.politics.mideast	day	sci.electronics
time	rec.autos	public	talk.religion.misc	black	alt.atheism
didn	talk.religion.misc	time	talk.politics.misc	left	sci.electronics
real	rec.sport.baseball	society	talk.politics.misc	miles	rec.autos
bad	rec.sport.baseball	live	talk.politics.misc	hours	rec.autos
guess	rec.sport.hockey	idea	talk.religion.misc	green	rec.autos
deal	talk.politics.guns	pay	talk.politics.misc	air	rec.autos
feel	talk.religion.misc	human	sci.crypt	engine	rec.autos
wouldn	rec.motorcycles	laws	talk.politics.guns	speed	sci.space
mind	rec.motorcycles	legal	talk.politics.mideast	field	sci.space
hard	talk.religion.misc	media	alt.atheism	cars	sci.space
hand	alt.atheism	money	sci.crypt	water	sci.electronics
pretty	rec.sport.hockey	business	sci.crypt	energy	comp.graphics
expect	talk.politics.guns	freedom	sci.crypt	blue	sci.electronics
won	rec.sport.hockey	wrong	sci.space	battery	sci.space
reason	sci.med	matter	alt.atheism	oil	rec.autos
remember	rec.sport.baseball	actions	talk.politics.mideast	heat	sci.space
understand	rec.sport.baseball	court	sci.crypt	planet	rec.motorcycles
difference	talk.religion.misc	situation	talk.politics.guns	low	sci.electronics
aren	sci.space	choice	talk.politics.guns	ago	sci.space
fine	rec.autos	talk	alt.atheism	theory	sci.electronics
worth	rec.sport.hockey	set	talk.politics.misc	red	sci.electronics
makes	rec.autos	simply	sci.crypt	physical	sci.electronics
talking	sci.electronics	illegal	talk.politics.guns	temperature	sci.electronics
nice	rec.sport.baseball	life	sci.space	called	sci.electronics

Topic 19	Sample classes	Topic 20	Sample classes
ac	sci.crypt	university	talk.politics.guns
subject	sci.crypt	subject	rec.sport.baseball
uk	sci.crypt	john	comp.sys.mac.hardware
david	sci.crypt	michael	rec.sport.baseball
phone	sci.crypt	science	sci.space
key	sci.crypt	dept	talk.politics.guns
chip	sci.crypt	cmu	rec.sport.hockey
wrote	sci.crypt	washington	rec.sport.baseball
au	sci.crypt	department	rec.sport.baseball
bit	sci.crypt	edu	comp.windows.x
line	sci.crypt	andrew	talk.politics.mideast
tin	sci.crypt	engineering	talk.politics.mideast
message	sci.crypt	colorado	rec.sport.hockey
clipper	sci.crypt	stanford	rec.motorcycles
version	sci.crypt	berkeley	comp.windows.x
company	sci.crypt	california	rec.sport.hockey
encryption	sci.crypt	virginia	misc.forsale
keys	sci.crypt	college	misc.forsale
fi	sci.crypt	eric	comp.windows.x
chris	sci.crypt	purdue	comp.windows.x
security	sci.crypt	utexas	talk.politics.mideast
algorithm	sci.crypt	texas	comp.sys.mac.hardware
uk	sci.crypt	duke	rec.sport.baseball
simple	sci.crypt	univ	sci.electronics
des	sci.crypt	advance	talk.politics.mideast
code	sci.crypt	math	misc.forsale
chips	sci.crypt	pa	rec.motorcycles
australia	sci.crypt	student	talk.politics.mideast
voice	comp.graphics	pittsburgh	misc.forsale
method	sci.crypt	keywords	talk.politics.mideast

We can see from inspection of the above tables that the topics already quite well captures consistent “themes”. It is interesting to note that some topics completely captures documents from a specific class. For instance in “Topic 4” the 30 most probable documents are almost completely from the “talk.politics.guns” class. Similarly in “Topic 7” all of the top 30 documents are from the “talk.politics.mideast” class, and correspondingly for “Topic 12”. “Topic 10” is also interesting, it has captured a “sports” theme, but it cannot separate between “hockey” and “baseball”! In other topics we can see themes that are cross-class as for instance “Topic 20” which seems to capture “cities” and “university”.

6 Conclusion

We have implemented a Gibbs sampler for LDA in Java to be used in the Weka data mining framework. The sampler has been evaluated on a popular data set for evaluating classification and clustering techniques of text. By inspection of the posterior probability of a subset of the desired quantities it seems as if the sampler does eventually converge. The generated word to topic assignments has also been manually inspected and they seem to capture consistent “themes” or “topics”.

References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update. *SIGKDD Explorations*, 2009.

[4] Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 569–577, New York, NY, USA, 2008. ACM.