DAT450 Machine Learning for Natural Language Processing
Assignment 2 – Topic modelling

Alexander Bridgwater
Gustav Lundberg
Cecilia Nyberg
Rikard Radovac

November 16 2023

1 Introduction

In this report, we will use a Latent Dirichlet Allocation (LDA) for Topic Modelling [1][9]. Topic modeling is a natural language processing method for discovering latent semantic structures in documents. Latent here refers to the fact that the model strives to learn "hidden" thematic topics hidden in the data. This is done in an unsupervised manner, without any labeled data. Using LDA, topic modelling in done in a purely Bayesian way, by assuming priors for our distributions and updating them to model the latent variables. Telling by the name, LDA models this using a Dirichlet distribution [6] that takes the form of

$$f(\boldsymbol{\theta} \mid \boldsymbol{\alpha}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \tag{1}$$

where $\mathrm{B}(\alpha)$ is a multinomial beta function.

In topic modeling, we make use of text documents D and limit ourselves to K unspecified topics with a vocabulary size of V. One document is denoted $d \in D$, and a topic $k \in K$. To symbolize one word belonging to a document, we denote it as w_n^d and subscripted $w_{d,n}$ to represent the n-th word of the d-th document. Here d is the document past preprocessing, which removes stopwords, lemmatizes and tokenizes the text, typical steps used in natural language processing (NLP) preprocessing [8]. For the Dirichlet distribution, we have the prior parameters α and β seen in Equation 1. The α denotes the prior document topic density, and the β impacts how many words of a document are needed to characterize the document as belonging to a topic. The optimal prior parameters can depend on the use case and underlying data and are thus tuned empirically to achieve the best model performance. The latent variables or distributions that the model infers during the training process are...

- The document-topic distributions θ_d, where d indexes the documents. Each θ_d is a vector of length K, with θ_{dk} representing the probability of topic k in document d:

$$\theta_d = (\theta_{d1}, \theta_{d2}, \ldots, \theta_{dK}) \tag{2}$$

- The topic-word distributions ϕ_k, where k indexes the topics. Each ϕ_k is a vector of length V,

with ϕ_{kw} representing the probability of word w in topic k:

$$\phi_k = (\phi_{k1}, \phi_{k2}, \ldots, \phi_{kV}) \tag{3}$$

- The topic assignments $z_{d,n}$, where n indexes the word positions in document d. Each $z_{d,n}$ indicates the topic assignment for the n-th word in document d:

$$z_{d,n} \in \{1, 2, \ldots, K\} \tag{4}$$

In LDA the document-topic θ and topic-word ϕ distribution associations are modeled by Dirichlet distributions, which serve as priors to the multinomial distributions that we sample from to retrieve the observed words in documents. Initially, topic assignments for words are influenced by these Dirichlet priors. Through an iterative process, each word in a document is examined, and a new topic is assigned based on a conditional probability. This probability accounts for the current distribution of topics in the document and the prevalence of words across topics, modeled by the Dirichlet hyperparameters α and β. Techniques such as Gibbs sampling or Variational Inference are employed to draw samples from this conditional probability distribution, favoring a stochastic approach over a deterministic one to avoid local optima and ensure a more thorough exploration of the topic space. As the algorithm iterates, the topic assignments converge to a more stable distribution.

Another sampling technique that is often used is Collapsed Gibbs sampling. It is an extension to Gibbs sampling avoids conditioning on all the variables by integrating out of the sampling process. It simplifies the sampling process by integrating out certain variables from a probabilistic model. More concretely, in a probabilistic model with a set of variables $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, standard Gibbs Sampling involves iteratively sampling each variable from its conditional distribution given the others:

$$p(X_i | X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) \tag{5}$$

In Collapsed Gibbs Sampling, suppose we integrate out a subset of these variables, denoted as $\mathbf{Y} \subset \mathbf{X}$. The conditional distribution for a variable X_i is then given by:

$$p(X_i | \mathbf{X}_{-i}, \mathbf{Y}) = \int p(X_i | \mathbf{X}_{-i}, \mathbf{y}) p(\mathbf{y} | \mathbf{X}_{-i}) \, d\mathbf{y} \tag{6}$$

Where $\mathbf{X}_{-i}$ represents all variables in $\mathbf{X}$ except X_i, and the integration is over the space of the variables in $\mathbf{Y}$. This collapsing of variables $\mathbf{Y}$ simplifies the conditional distributions for the remaining variables, often leading to more efficient and faster convergence in high-dimensional spaces. Moreover, it is especially useful in Bayesian statistical models with numerous interdependent variables, such as in LDA for topic modeling [7].

To numerically evaluate the model's generated topics, the Umass coherence score [2] may be used as a metric. The score is based on the semantic similarity between the most frequent words in each topic and thus gives an insight into the model's performance. The Umass coherence score is computed by the formula below,

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \left(\frac{D(v_m^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})} \right) \tag{7}$$

Where $C(t; V^{(t)})$ is the coherence score function, M is a parameter for how many of the most frequent words to account for, and $D(v_l^{(t)})$ is the document frequency which denotes how many documents has

least one token of type v_l for the topic t. Lastly, $D(v_m^{(t)}, v_l^{(t)})$ describes the co-document frequency and is the count of documents that contain at least one token of type v_m and v_l.

For reference, the generative process for LDA can be described as follows [9]:

For each topic k, where $k = 1, \ldots, K$:

(a) Draw the topic-word distribution: $\phi^{(k)} \sim \text{Dirichlet}(\beta)$

For each document d in the corpus D:

(a) Draw the document-topic distribution: $\theta_d \sim \text{Dirichlet}(\alpha)$

(b) For each word position i in document d:

i. Draw the topic assignment: $z_i \sim \text{Multinomial}(\theta_d)$

ii. Draw the word from the assigned topic: $w_i \sim \text{Multinomial}(\phi^{(z_i)})$

2 Method

This section describes our methodology in preprocessing the fetched data, how gibbs sampling was used for LDA, and the evaluation of different umass coherence scores.

2.1 Pre-processing

For the text corpus in our dataset, a news corpus was used from the 20NewsGroup. A subset of 3000 documents was used from the fetched corpus. Preprocessing of this subset of data was used to provide better results for analysis.

The preprocessing was comprised of three fundamental stages: firstly, cleaning the text by addressing whitespaces, followed by transforming all text to lowercase to ensure uniformity in word representation and remove all potential ambiguities. For instance, the words "Hello" and "hello" are considered identical, preventing different representations due to casing variations. Additionally, lemmatization was used to convert different conjugations and spelling variations into the same format.

Subsequently, tokenization was applied uniformly across the entire corpus to segment each word or short phrase into distinct components. Additionally, common stopwords, such as "who", "as", "can", and "I," etc . were removed using the spaCy library. The reason for this data cleaning is that the stopwords generally have minimal contribution to sentiment or text classification which makes them redundant [4].

Further preprocessing involved removing punctuations and special signs, ensuring that only textual data was included. Additionally, numerical values were excluded from the corpus, emphasizing the focus on textual content.

Finally, words occurring less than 10 times throughout the corpus were eliminated. This strategic decision aims to enhance the efficiency of classification, recognizing the challenges posed by infrequent word occurrences for effective model learning.

2.2 LDA with Collapsed Gibbs

To estimate the topic assignments for the words in the cleaned and tokenized documents, collapsed Gibbs sampling was used.

Before using collapsed Gibbs sampling, the documents must go through an initialization phase, where each word is assigned a random topic [3]. After this introduction phase, the collapsed Gibbs probability

equation (8) was used to assign a probability of each topic to a specific word in a specific document, based on the current state of the topic assignments to all other words in the same document, and to all other occurrences of the same words in the corpus. Then the topic with the highest probability was assigned to the specific word. This was done for each word in all documents, and the process was repeated over 150 iterations. This was repeated for different values of alpha and beta and different numbers of topics.

$$p(z_i|z(-i), w) \propto (n_{d,k}^{-i} + \alpha_k) \cdot \frac{n_{k,w}^{-i} + \beta_w}{\sum_{w'} n_{k,w'}^{-i} + \beta_{w'}} \tag{8}$$

For LDA for topic modeling, two latent distributions are searched for; portions θ_d and topic-word distributions ϕ_k. Since the latent distributions are conditional they could be derived by a sufficient statistic z for both distributions. This was done by calculating the topic index assignment for each word z_i. Sampling this z_i gave us the Gibbs sampler.

The implementation of collapsed Gibbs sampling closely followed the algorithm depicted in Algorithm 1 below. The process entailed establishing necessary count variables, initializing them randomly, and subsequently executing a loop for the specified number of iterations. In each iteration, a topic was sampled for every word instance in the corpus.

Algorithm 1 Algorithm for LDA Gibbs Sampling

Require: words w in documents d
Ensure: topic assignments z and counts $n_{d,k}$, $n_{k,w}$, and n_k
randomly initialize z and increment counters
for each iteration **do**
    for $i = 0$ to $N - 1$ **do**
        $word \leftarrow w[i]$
        $topic \leftarrow z[i]$
        $n_{d,\text{topic}} \leftarrow n_{d,\text{topic}} - 1$
        $n_{\text{word,topic}} \leftarrow n_{\text{word,topic}} - 1$
        $n_{\text{topic}} \leftarrow n_{\text{topic}} - 1$
        for $k = 0$ to $K - 1$ **do**
            $p(z = k) \leftarrow \frac{(n_{d,k}+\alpha)(n_{k,w}+\beta)}{(n_k+W\beta)}$
        end for
        $topic \leftarrow$ sample from $p(z)$
        $z[i] \leftarrow topic$
        $n_{d,\text{topic}} \leftarrow n_{d,\text{topic}} + 1$
        $n_{\text{word,topic}} \leftarrow n_{\text{word,topic}} + 1$
        $n_{\text{topic}} \leftarrow n_{\text{topic}} + 1$
    end for
end for
return z, $n_{d,k}$, $n_{k,w}$, n_k

To get the top twenty most common words for each topic by relative frequency, the topic-word distribution ϕ_k, was calculated for each topic. The ϕ_k distribution was also used to calculate the Umass coherence score.

2.3 Umass Coherence Score

The Umass coherence score was used to calculate a numerical assessment of how well topics aligned, helping the assessment and understanding of the topics produced by the LDA model. These scores were used as a tool to evaluate the logical understandability of the identified topics. Providing a reference point of how well the model performed for different model settings.

3 Results

The LDA model with collapsed Gibbs sampling was evaluated on the 20 News Groups dataset [5], with different values of the hyperparameters α, β, K. The results are presented below where tables 1, 2, 3, 4 represent the 20 most frequent words and 5, 6, 7, 8 show the coherence scores, for the four best topics with hyperparameters in the set:

- $\alpha = \beta : 0.1, 0.01$
- K: $10, 50$

Topic 1	**Topic 5**	**Topic 9**	**Topic 10**
window	image	people	game
use	jpeg	key	team
file	file	government	year
user	color	law	win
system	gif	right	player
program	format	president	play
mail	bit	think	good
include	program	know	van
application	version	state	season
run	available	gun	time
datum	display	fbi	de
server	quality	case	league
know	use	chip	hit
work	package	use	think
address	graphic	time	det
problem	free	number	hockey
post	convert	security	baseball
information	software	need	great
list	viewer	public	blue
message	pixel	weapon	vote

Table 1: Top 20 words for the best four topics with $\alpha = \beta = 0.1$ and $K = 10$

Topic 1	**Topic 4**	**Topic 5**	**Topic 8**
people	drive	file	god
government	card	image	jesus
say	work	jpeg	people
state	use	key	believe
president	problem	bit	say
law	like	gif	think
child	know	format	christian
kill	system	color	bible
gun	thank	program	know
israel	need	use	man
right	driver	version	love
ed	new	available	read
fbi	good	display	come
police	run	chip	mean
country	speed	code	church
year	think	software	time
force	car	quality	sin
arab	get	free	write
agent	disk	number	faith
report	buy	convert	day

Table 2: Top 20 words for the best four topics with $\alpha = \beta = 0.01$ and $K = 10$

Topic 2	**Topic 6**	**Topic 20**	**Topic 49**
group	conference	internet	drive
post	university	system	disk
article	paper	user	mb
list	information	use	monitor
newsgroup	research	information	system
discussion	science	email	card
issue	year	anonymous	mac
include	include	computer	controller
question	program	privacy	scsi
read	institute	address	port
provide	available	message	use
new	author	network	need
receive	additional	mail	bit
send	address	site	work
number	library	file	software
posting	june	access	hard
time	hotel	account	ram
request	send	anonymity	floppy
follow	member	usenet	pc
faq	rate	identity	ide

Table 3: Top 20 words for the best four topics with $\alpha = \beta = 0.1$ and $K = 50$

Topic 8	**Topic 10**	**Topic 27**	**Topic 38**
university	space	information	key
conference	orbit	internet	chip
school	earth	user	clipper
national	energy	system	encryption
history	launch	computer	number
research	moon	mail	phone
professor	constant	privacy	bit
book	mission	anonymous	algorithm
paper	probe	network	encrypt
include	satellite	use	escrow
institute	system	message	block
new	mile	site	agency
program	year	posting	use
april	cost	know	security
follow	velocity	anonymity	law
page	saturn	address	secure
international	titan	usenet	public
general	gravity	list	enforcement
receive	solar	account	system
author	mass	member	scheme

Table 4: Top 20 words for the best four topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 1	**Topic 5**	**Topic 9**	**Topic 10**	**Mean**
-470.98	-436.24	-444.29	-488.71	-415.08

Table 5: Umass coherence score for $\alpha = \beta = 0.1$ and $K = 10$

Topic 1	**Topic 4**	**Topic 5**	**Topic 8**	**Mean**
-521.11	-519.98	-433.71	-512.85	-444.35

Table 6: Umass coherence score for $\alpha = \beta = 0.01$ and $K = 10$

Topic 2	**Topic 6**	**Topic 20**	**Topic 49**	**Mean**
-474.77	-438.51	-515.88	-493.67	-446.36

Table 7: Umass coherence score for $\alpha = \beta = 0.1$ and $K = 50$

Topic 8	**Topic 10**	**Topic 27**	**Topic 38**	**Mean**
-409.97	-393.17	-458.43	-438.69	-441.80

Table 8: Umass coherence score for $\alpha = \beta = 0.01$ and $K = 50$

All topics with the top 20 words and their coherence scores can be found in appendix.

4 Discussion and Analysis

When comparing the results for different settings, the mean value of the umass coherence score can be used. When comparing the results from this perspective, using $\alpha = \beta = 0.1$ and letting the number of topics be 10 had an average Umass coherence score of -415.08, which was closest to zero compared to the other settings. Therefore, when using the mean value of the coherence scores, $\alpha = \beta = 0.1$ and K=10 resulted in the most favorable outcome.

The results for the same α and β values but different numbers of topics show that the mean value of the UMass coherence score was better for using 10 topics when α and β was equal to 0.1 but using 50 topics was better when α and β was equal to 0.01. This makes it difficult to conclude whether a larger or lower amount of topics is most suitable for this dataset. There could potentially be a better value between 10 and 50. To potentially achieve better results, other numbers of topics could have been tested. Preferably numbers closer to 10 and with α and β close to, or higher than 0.1 since K = 10 and $\alpha = \beta = 0.1$ resulted in the best overall score. Additionally, the model could have been run on different numbers of iterations, possibly yielding a better result.

An essential factor in assessing the model's performance also involves evaluating the topics in relation to human understanding and comparing it to the Umass coherence score. When the topics were analyzed by us, the accuracy overall seemed good and most of the words in a topic seemed to have an obvious association to each other. In the result section, the topics with the highest degree of logical coherence were chosen. When comparing the umass coherence scores for these topics with the umass coherence score for the topics not chosen, it can be seen that the umass coherence scores do not align completely with the logical analysis. For example, topic two was eliminated when using $\alpha = \beta = 0.1$ and K = 10 since it was not considered one of the best topics. However, table 13 in the appendix shows this topic having the best Umass coherence score of all topics with these settings.

Furthermore, according to the Umass coherence score, K = 10 and $\alpha = \beta = 0.1$ resulted in the best overall score. When analyzed by us, the results from K = 10 and $\alpha = \beta = 0.01$ were considered to have the most distinct topic classifications. Choosing a higher number of topics was considered to result in less specific and less clear topic classifications compared to a lower number of topics.

Nevertheless, for some topics, the ranking based on the UMass score closely aligned with a logical ranking. This indicates that while the UMass score may not be a perfect representation of topic quality, it can be a useful indicator. Given the subjective nature of logical ranking, other people may perceive the accuracy of UMass coherence score rankings differently from these authors. To evaluate this further, a more extensive range of opinions from various people would have been necessary.

Another important factor to take into account when evaluating the model is that only the best resulting topics were picked and presented as a result. These topics were picked based on our judgment, and the fact that the badly generated topics are not presented may make the model seem better than it is. To further evaluate the model, both good and bad topics could be compared and possibly with more people to reduce the biased opinion from a single group.

A better result could have been achieved with better preprocessing. While having done several steps in our data cleaning, including lemmatization, removing stopwords, special characters, and more, there are still some tokens that preferably should not appear in any topic. In some topics, one-letter "words" appear without any specific reason to the human eye. For example in topic 28 in table 10 the word "o" appears without any intuitive relation to the other words in the topic. This type of flaw in the modeling could have been resolved with further lemmatization and data cleaning.

But one should also be careful about filtering out too many misspellings and stop words as these can be useful in other types of contexts and topics. Filtering out various correct spellings with lemmatization of the same word may result in the loss of information about linguistic differences, which could be

important in some cases. In topic 17 in table 11, the one-letter words actually have some type of relevance and relation to each other. Here they could represent different types of constants, or variables which could be used in maths or physics. While filtering these out would have led to cleaner data for other topics, it would also entail a complete loss of information specific to this particular topic. So, there is a balancing point to find, and a trade-off to consider, regarding how much information one is willing to lose in exchange for better-represented topics.

To be able to evaluate the model further, it could be beneficial to try it on other datasets as well.

References

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] E.Leenders A. McCallum D. Minmo, H. Wallach. Optimizing semantic coherence in topic models. *Association for Computing Machinery, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011.

[3] William M. Darling. *A Theoretical and Practical Implementation Tutorial on Topic Modeling and Gibbs Sampling.* 2011.

[4] Jacob Eisenstein. *Natural Language Processing.* 2018.

[5] Ken Lang. 20 newsgroups dataset, 2008.

[6] Jiayu Lin. On the dirichlet distribution. *Department of Mathematics and Statistics, Queens University*, pages 10–11, 2016.

[7] 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*, pages 569–577, 2008.

[8] S Vijayarani, Ms J Ilamathi, Ms Nithya, et al. Preprocessing techniques for text mining-an overview. *International Journal of Computer Science & Communication Networks*, 5(1):7–16, 2015.

[9] Wikipedia contributors. Latent dirichlet allocation — Wikipedia, the free encyclopedia, 2023. [Online; accessed 15-November-2023].

Appendix

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
window	space	armenian	like	image	card	x	god	people	game
use	w	people	go	jpeg	drive	do	people	key	team
file	m	say	know	file	system	test	think	government	year
user	appear	muslim	think	color	thank	value	jesus	law	win
system	s	ed	say	gif	use	int	believe	right	player
program	research	azerbaijan	get	format	work	define	know	president	play
mail	copy	israel	car	bit	disk	science	say	think	good
include	university	kuwait	good	program	good	remote	time	know	van
application	cost	war	time	version	mb	return	read	state	season
run	new	arab	look	available	need	objective	point	gun	time
datum	program	year	thing	display	problem	set	mean	fbi	de
server	book	government	come	quality	video	treatment	like	case	league
know	design	turkish	way	use	speed	mpc	man	chip	hit
work	art	new	people	package	buy	need	christian	use	think
address	station	state	problem	graphic	pc	include	bible	time	det
problem	system	turkey	year	free	chip	study	come	number	hockey
post	include	ar	see	convert	like	try	word	security	baseball
information	cover	university	old	software	bit	process	love	need	great
list	mission	history	want	viewer	know	scientific	good	public	blue
message	c	russian	find	pixel	monitor	world	thing	weapon	vote

Table 9: Top 20 words for all 10 topics with $\alpha = \beta = 0.1$ and $K = 10$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
people	mail	window	drive	file	x	know	god	year	armenian
government	address	do	card	image	space	think	jesus	go	muslim
say	information	include	work	jpeg	w	like	people	game	azerbaijan
state	list	use	use	key	m	people	believe	say	kuwait
president	send	x	problem	bit	s	way	say	good	turkish
law	post	file	like	gif	remote	thing	think	time	university
child	internet	application	know	format	mission	right	christian	like	turkey
kill	computer	server	system	color	station	want	bible	know	history
gun	appear	program	thank	program	launch	go	know	think	ar
israel	email	set	need	use	option	good	man	get	new
right	system	run	driver	version	orbit	time	love	team	russian
ed	group	widget	new	available	earth	point	read	look	armenia
fbi	e	font	good	display	shuttle	problem	come	come	population
police	copy	work	run	chip	program	use	mean	play	soviet
country	datum	resource	speed	code	t	try	church	win	dead
year	new	function	think	software	v	find	time	day	azeri
force	university	try	car	quality	cost	case	sin	player	baku
arab	anonymous	ftp	get	free	energy	work	write	car	york
agent	include	version	disk	number	slip	system	faith	see	republic
report	user	library	buy	convert	moon	tell	day	run	karabakh

Table 10: Top 20 words for all 10 topics with $\alpha = \beta = 0.01$ and $K = 10$

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**	**Topic 6**	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
bit	group	window	sin	sell	conference	insurance	bit	time	believe
read	post	program	law	pay	university	yes	order	think	think
oh	article	server	homosexual	buy	paper	power	number	get	true
say	list	resource	christian	price	information	company	output	good	argument
think	newsgroup	font	people	service	research	use	line	lot	mean
colour	discussion	application	point	want	science	amp	high	like	say
xv	issue	client	sex	water	year	energy	find	away	claim
hold	include	use	paul	new	include	low	work	big	evidence
see	question	run	homosexuality	plant	program	current	point	hit	point
change	read	xterm	christians	market	institute	like	level	year	reason
black	provide	set	man	company	available	private	range	dog	question
computer	new	display	commandment	business	author	small	value	ball	statement
know	receive	event	case	good	additional	cost	source	real	context
like	send	app	gay	cost	address	component	sound	long	assume
yes	number	try	child	start	library	supply	low	maybe	belief
stuff	posting	manager	act	cheap	june	thing	buffer	stick	agree
bad	time	time	sabbath	radio	hotel	health	light	right	understand
post	request	start	mean	purchase	send	rate	pixel	run	fact
problem	follow	problem	scripture	order	member	bell	test	go	jim
big	faq	memory	word	place	rate	speed	degree	way	article

Table 11: Top 20 words for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$

Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
car	space	president	say	datum	offer	w	game	driver	internet
bike	station	think	police	image	good	m	team	use	system
drive	mission	people	san	available	o	s	year	window	user
mile	launch	q	home	package	sell	b	play	slip	use
engine	orbit	states	anti	fax	sale	d	win	load	information
power	shuttle	united	car	graphic	original	c	player	port	email
speed	earth	say	francisco	information	include	t	time	phone	anonymous
ride	option	time	league	software	interested	g	season	modem	computer
oil	system	plan	work	format	ask	v	hockey	mouse	privacy
road	flight	question	spy	display	price	giz	good	do	address
temperature	moon	bill	maybe	application	book	mass	league	c	message
way	cost	clinton	adl	research	manual	fij	hit	run	network
go	design	country	shagen	model	condition	constant	get	irq	mail
look	satellite	good	shout	sgi	shipping	f	baseball	application	site
little	solar	mr	angeles	ftp	new	h	think	packet	file
owner	high	thing	friday	workstation	great	int	well	ip	access
plastic	redesign	administration	gerard	contact	come	l	fan	mode	account
tower	team	war	officer	user	email	k	lose	file	anonymity
head	spacecraft	general	south	set	model	r	bad	instruction	usenet
rear	nasa	force	los	inc	etc	cooper	score	work	identity

Table 11: Top 20 words for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
problem	people	armenian	book	azerbaijan	gun	case	god	det	do
printer	like	muslim	know	russian	weapon	study	love	vote	not
work	think	turkish	text	state	drug	effect	bible	period	mpc
try	know	university	read	national	kill	thing	believe	tor	cd
color	want	turkey	history	accord	death	science	christ	nyr	world
new	thing	history	write	russia	number	scientific	faith	john	microsoft
print	right	population	etc	baku	year	try	know	shot	system
thank	way	istanbul	idea	april	rate	idea	life	chi	family
line	go	turk	law	turan	handgun	good	people	van	ii
upgrade	good	new	fact	economic	attack	treatment	religion	la	master
colormap	tell	genocide	nazi	armenia	person	patient	come	blue	find
help	let	greek	people	president	people	evidence	christian	save	city
fix	point	government	find	oil	homicide	result	human	stl	business
change	lot	million	work	soldier	firearm	doctor	christianity	mtl	time
want	hear	ottoman	give	sign	crime	know	think	pt	place
appreciate	get	today	look	great	control	medical	thing	bos	great
mode	say	slaughter	evidence	appeal	carry	point	man	min	development
use	see	soviet	try	work	likely	experience	give	david	internet
buy	course	york	little	peace	vancouver	theory	way	pit	software
check	sure	ve	take	un	defense	kind	accept	pp	money

Table 11: Top 20 words for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
include	thank	like	jesus	go	kuwait	number	moral	appear	x
new	mail	tape	matthew	day	fbi	gas	think	copy	remote
error	know	use	say	tell	fire	non	church	art	rx
muslim	post	lock	speak	child	agent	package	system	cover	set
type	e	sound	prophecy	know	batf	like	mean	man	command
function	like	line	day	kill	koresh	know	objective	vs	program
washington	question	try	david	come	child	mask	morality	x	shell
use	look	find	passage	look	compound	precision	definition	new	size
class	help	signal	lord	happen	gas	actually	action	annual	echo
york	reply	way	prophet	leave	go	tear	term	wolverine	shar
street	appreciate	thing	tell	mean	al	c	catholic	ghost	return
warning	find	look	son	people	davidian	course	code	issue	try
dc	email	go	psalm	find	branch	b	life	rider	need
islam	address	cable	mary	let	warrant	kind	wrong	green	pass
parse	advance	switch	israel	get	state	type	standard	hulk	user
usa	hi	ground	time	long	start	multiple	matter	story	file
undeclared	info	sure	look	start	sheikh	music	society	comic_strip	host
news	send	need	gospel	like	die	effect	murder	punisher	test
ny	good	get	king	see	evidence	stuff	canon	hobgoblin	fi
pom	anybody	hole	claim	old	cult	come	orthodox	sabretooth	display

Table 11: Top 20 words for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 41	Topic 42	Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
armenian	key	year	right	israel	card	jpeg	include	drive	file
people	chip	school	government	ed	mhz	image	support	disk	program
ar	clipper	tax	people	arab	bus	gif	widget	mb	use
say	encryption	work	law	jewish	speed	color	ftp	monitor	window
dead	bit	fund	state	jews	fast	format	free	system	need
body	phone	go	court	israeli	cpu	file	use	card	work
kill	algorithm	money	case	palestinian	chip	quality	xview	mac	write
azeri	encrypt	job	human	de	video	bit	o	controller	pc
year	law	battery	allow	state	run	version	available	scsi	run
azerbaijan	escrow	university	power	people	cache	display	motif	port	time
town	security	billion	military	non	driver	viewer	code	use	know
mr	block	public	country	van	isa	convert	look	need	code
hojali	number	bring	political	palestinians	machine	available	version	bit	software
karabakh	agency	life	time	party	board	free	function	work	version
figure	government	pay	bear	territory	transfer	view	application	software	ftp
war	system	support	constitution	country	test	program	system	hard	find
woman	public	program	second	quote	performance	pixel	base	ram	look
report	secure	problem	protect	claim	processor	jfif	note	floppy	machine
massacre	scheme	center	citizen	jew	vlb	note	library	pc	graphic
man	enforcement	class	grant	het	frame	well	subject	ide	copy

Table 11: Top 20 words for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
ed	people	gun	know	god	jesus	know	university	appear	space
new	go	weapon	group	believe	matthew	think	conference	art	orbit
article	time	batf	post	thing	lord	like	school	copy	earth
point	say	de	etc	way	christian	people	national	cover	energy
design	good	fbi	think	human	church	good	history	vs	launch
mean	want	kill	book	say	say	want	research	x	moon
line	thing	agent	like	time	speak	time	professor	man	constant
question	know	firearm	kind	man	prophecy	right	book	new	mission
time	like	assault	question	question	scripture	way	paper	wolverine	probe
build	think	death	find	think	write	get	include	annual	satellite
detector	little	warrant	read	belief	passage	try	institute	rider	system
bad	take	davidian	good	nature	heaven	lot	new	hulk	mile
speed	doctor	van	thing	true	son	say	program	ghost	year
radar	give	handgun	thank	understand	david	go	april	issue	cost
field	get	use	look	mean	john	find	follow	comic_strip	velocity
current	hear	branch	try	truth	book	thing	page	punisher	saturn
way	open	homicide	treatment	see	word	wrong	international	black	titan
today	child	attack	bad	different	verse	let	general	hobgoblin	gravity
ago	work	shoot	experience	give	gospel	mean	receive	sabretooth	solar
engine	ask	rate	need	term	text	need	author	bag	mass

Table 12: Top 20 words for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**	**Topic 16**	**Topic 17**	**Topic 18**	**Topic 19**	**Topic 20**
people	space	year	god	card	package	car	jpeg	file	x
government	design	good	sin	driver	datum	use	image	program	include
state	research	hit	people	video	image	low	color	time	widget
war	center	player	life	monitor	model	like	gif	read	function
arab	provide	run	law	bus	graphic	power	file	faq	file
force	cost	look	know	system	sgi	air	format	number	set
israel	project	number	faith	color	code	control	bit	code	remote
number	study	like	think	window	line	problem	quality	disk	error
peace	use	well	come	vga	precision	change	display	version	int
fact	include	game	bible	work	inc	oil	free	list	try
support	nasa	battery	christ	problem	map	well	viewer	copy	display
claim	support	time	good	port	technology	pressure	convert	work	program
give	level	get	believe	use	dec	follow	pixel	need	use
russia	human	think	day	mode	tape	old	view	create	server
long	current	play	jesus	mb	vertex	way	program	write	resource
way	information	win	like	support	visualization	make	jfif	result	return
security	scientific	water	paul	fast	analysis	little	version	software	shell
authority	capability	high	say	connect	processing	level	quicktime	article	motif
clinton	system	problem	mean	know	new	part	well	know	define
party	science	come	christians	screen	work	year	setting	find	command

Table 12: Top 20 words for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$ (Continued)

Topic 21	**Topic 22**	**Topic 23**	**Topic 24**	**Topic 25**	**Topic 26**	**Topic 27**	**Topic 28**	**Topic 29**	**Topic 30**
love	go	use	w	good	work	information	o	government	think
god	say	driver	m	offer	thank	internet	card	pay	course
know	car	device	s	sell	like	user	machine	tax	real
event	come	slip	b	cd	look	system	good	system	american
window	like	company	c	include	think	computer	buy	private	people
like	know	work	d	sound	know	mail	thank	plant	police
quote	look	phone	t	price	help	privacy	sell	cost	like
man	kill	port	g	new	find	anonymous	cable	provide	come
pray	get	c	giz	sale	hear	network	mhz	money	anti
parent	way	high	v	power	idea	use	price	yes	carry
power	home	modem	f	interested	thing	message	board	insurance	live
reply	live	like	fij	great	good	site	like	business	officer
moses	tell	number	e	mail	need	posting	problem	service	club
ask	right	packet	k	shipping	try	know	chip	policy	office
refuse	happen	speed	l	manual	run	anonymity	new	fund	white
give	thing	ip	bit	condition	current	address	know	canada	look
take	head	load	h	ask	get	usenet	try	new	care
stephen	let	b	p	original	appreciate	list	couple	person	pretty
work	burn	read	r	rom	email	account	operational	sell	require
come	maybe	box	not	system	new	member	power	work	public

Table 12: Top 20 words for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$ (Continued)

Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
drive	year	game	thank	child	say	president	key	kuwait	problem
scsi	dog	team	work	fire	new	option	chip	det	run
disk	april	win	know	gas	people	station	clipper	tor	printer
bike	mark	play	want	know	time	q	encryption	al	font
car	start	hockey	feel	think	go	united	number	period	window
go	white	think	e	tell	bosnia	work	phone	nyr	program
mb	right	get	need	come	washington	year	bit	chi	use
speed	month	player	question	go	come	russian	algorithm	shot	memory
ide	great	lose	time	like	muslim	states	encrypt	van	work
controller	woman	season	mail	thing	york	job	escrow	la	mhz
tape	high	fan	like	fbi	war	administration	block	stl	time
hard	happen	league	appreciate	say	area	mr	agency	mtl	need
time	military	time	send	story	street	report	use	pt	cpu
use	area	playoff	help	day	agreement	american	security	bos	pc
mile	zoroastrian	good	try	compound	political	new	law	sheikh	print
device	rest	year	see	start	begin	today	secure	save	line
buy	state	go	problem	force	land	program	public	pit	application
engine	young	watch	advance	koresh	news	country	enforcement	min	fast
problem	time	score	get	want	dc	operation	system	pp	want
lock	hour	nhl	effect	hear	live	clinton	scheme	british	mac

Table 12: Top 20 words for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$ (Continued)

Topic 41	Topic 42	Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
armenian	server	list	people	azerbaijan	window	order	do	people	right
turkish	use	free	think	israel	file	mail	mpc	dead	law
ar	file	ftp	evidence	jewish	available	send	information	year	people
muslim	address	mail	mean	jews	system	address	san	say	state
armenia	email	include	point	history	ftp	e	mission	body	case
turkey	machine	vote	fact	palestinian	use	place	flight	kill	court
baku	client	note	claim	israeli	version	call	francisco	mr	government
population	post	send	homosexual	muslim	like	support	say	woman	order
karabakh	send	post	case	nazi	support	phone	league	man	power
turk	x	newsgroup	read	arab	program	information	adl	town	issue
republic	example	subject	word	country	user	blue	police	old	time
azeri	process	david	believe	territory	datum	way	gerard	azerbaijan	political
government	xterm	discussion	way	text	run	need	defamation	hojali	rule
russian	window	etc	agree	die	application	board	anti	figure	point
istanbul	problem	xview	consider	land	software	st	world	see	act
soviet	different	contact	reason	islam	graphic	hawk	los	child	believe
greek	mode	look	thing	non	base	time	bullock	take	call
massacre	command	john	TRUE	policy	write	pass	file	wife	constitution
genocide	message	group	sense	religious	source	message	search	day	allow
million	set	yes	support	palestinians	look	second	time	start	general

Table 12: Top 20 words for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$ (Continued)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
-470.98	-393.94	-449.58	-472.74	-436.24	-483.34	-425.34	-500.78	-444.29	-488.71

Table 13: Umass coherence score for all 10 topics with $\alpha = \beta = 0.1$ and $K = 10$

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
-474.08	-474.77	-393.7	-421.02	-495.46	-438.51	-478.51	-498.08	-477.39	-515.43

Table 14: Umass coherence score for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
-500.98	-390.89	-458.78	-420.17	-417.49	-421.82	-406.23	-412.56	-428.76	-515.88

Table 14: Umass coherence score for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
-478.25	-468.94	-432.06	-406.73	-374.46	-498.47	-432.48	-353.56	-373.97	-399.65

Table 14: Umass coherence score for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
-542.39	-503.57	-534.42	-416.1	-392.26	-410.53	-507.5	-547.1	-486.48	-389.52

Table 14: Umass coherence score for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 41	Topic 42	Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
-402.63	-448.54	-467.23	-464.14	-417.74	-459.3	-424.6	-460.61	-504.47	-493.67

Table 14: Umass coherence score for all 50 topics with $\alpha = \beta = 0.1$ and $K = 50$ (Continued)

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
-520.23	-526.16	-425.13	-536.11	-464.64	-428.66	-504.5	-409.97	-465.42	-393.17

Table 15: Umass coherence score for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 11	Topic 12	Topic 13	Topic 14	Topic 15	Topic 16	Topic 17	Topic 18	Topic 19	Topic 20
-465.35	-469.37	-498.77	-420.07	-474.41	-436.37	-492.29	-440.3	-464.47	-410.61

Table 15: Umass coherence score for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 21	Topic 22	Topic 23	Topic 24	Topic 25	Topic 26	Topic 27	Topic 28	Topic 29	Topic 30
-489.17	-463.86	-484.93	-364.1	-435.23	-499.36	-458.43	-466.37	-430.02	-491.17

Table 15: Umass coherence score for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 31	Topic 32	Topic 33	Topic 34	Topic 35	Topic 36	Topic 37	Topic 38	Topic 39	Topic 40
-501.27	-435.66	-440.74	-519.29	-457.76	-507.72	-392.88	-438.69	-314.98	-487.92

Table 15: Umass coherence score for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$

Topic 41	Topic 42	Topic 43	Topic 44	Topic 45	Topic 46	Topic 47	Topic 48	Topic 49	Topic 50
-315.45	-526.27	-515.31	-438.16	-359.55	-494.48	-534.94	-357.43	-465.78	-431.51

Table 15: Umass coherence score for all 50 topics with $\alpha = \beta = 0.01$ and $K = 50$