DAT450 - Assignment 2

Group CTH-15: Stefano Ribes, Cody Hesse, Apoorva Udayakumar

November 18, 2022

1 Introduction

Topic modeling is an unsupervised natural language processing technique used to extract topics to the words of a set of documents, *i.e.* a corpus. To do so, topic modeling categorises each word into one of K topics, which is used to analyse and understand the context of the document. For the rest of this report, we will use the term "word" to actually refer to the resulting tokens obtained by the cleaning and tokenization processes applied to the corpus onto which perform topic modeling.

A popular way to implement topic modelling is by utilizing the Latent Dirichlet Allocation (LDA) statistical model. LDA is able to learn both the topics distribution in each document within a corpus, as well as the words distribution associated to each topic in the entire corpus [1]. In particular, the topics distribution for each word in a document d is given by:

$$\theta_d|z, w \sim Dir(\alpha + n_d(1), \ldots, \alpha + n_d(K)), \tag{1}$$

where $n_d(k)$ represents the number of words in document d with topic k. The words distribution associated to each topic k in the corpus is instead given by:

$$\phi_k|z, w \sim Dir(\beta + m_k(1), \ldots, \beta + m_k(V)), \tag{2}$$

where $m_k(w)$ represents the number of positions in the corpus with topic k and word w. The values α and β are two model's hyperparameters which assign a non-zero probability of allocating a topic k to a specific word even when its n_d or m_k counts are zero.

The aim of LDA is to generate two suitable distributions ϕ_k and θ_d, from which one can sample the words that are highly associated to each topic k and the topics to which a document d likely belongs to. LDA considers the distribution of topics in the document and the distribution of words in a topic. Based on both these distributions, a topic can be assigned to a word. An approach that helps in achieving this is Gibbs sampling.

Gibbs sampling is used to sample a conditional distribution of variables whose distribution over a long run will converge to a true distribution. The probability that each document belongs to a specific topic is then calculated by finding the probability of each topic in the document and word belonging to that topic in that document conditioned on

θ and ϕ. Hence, for a given θ, ϕ, Gibbs sampling finds the probability of each word at position j in document d belongs to a topic k by sampling each word independently from the document d. For all (d, j) independently, we can draw samples according to Equation 3 as follows:

$$\mathbf{P}(Z_{d_j} = k, W_{d_j} = w_{d_j} | \theta, \phi) \propto \theta_d(k)\, \phi(w_{d_j}). \tag{3}$$

A more efficient implementation of Gibbs Sampling is referred to Collapsed Gibbs Sampling (CGS). CGS updates z_{d_j} *i.e.* the topic of a word at position j in document d by drawing a sample for a given (d, j) as per the conditional distribution of all the words and all topics k except the topic at (d, j) [3]. The probability of each word belonging to a topic is then proportional to:

$$q_k = \frac{(\alpha + n_d^{-d_j}(k))(\beta + m_k^{-dj}(w_{dj})}{V\beta + m^{-dj}} \tag{4}$$

and z_{d_j} is randomly chosen from a categorical distribution $z_{d_j} \sim \text{Cat}(p_1,...,p_K)$, where p_k is given by

$$p_k = \frac{q_k}{\sum_{j=1}^{K} q_j} \tag{5}$$

where $n_d^{-d_j}$ is the count of the each topic across the entire document, except the topic at position j in document d. The value $m_k^{-d_j}$ represents instead the count of the words with topic k across the entire corpus, except the topic of the word at that position j.

2 Methodology

In this section we briefly describe our followed methodology, which includes the corpus preprocessing and the implmentation of the traditional Gibbs sampling and CGS.

2.1 Corpus Preprocessing

The corpus preprocessing is performed in three steps: lower casing, tokenization and stopwords removal. Lower casing is applied throughout the corpus in order to guarantee consistency in word representation and avoid ambiguities. For instance, "corn" and "Corn" are the same word, but would have two different representations without adaptation of the casing. Tokenization is performed on the entire corpus in order to split each word, or short phase, into individual components. Moreover, common *stop words* are removed, such as "the", "is" and "are". Stop words are removed as they rarely contribute to the sentiment or classification of text, and thus take up a large place in memory without major contribution. Finally, words with less than 10 occurrences throughout the corpus are removed, as classification becomes difficult without recurrence.

2.2 LDA with Traditional Gibbs Sampling

Our implementation of LDA with Gibbs sampling is based on two matrices N and M. Matrix N is a $D \times K$ matrix, where D is the number of documents in the corpus and K is the number of topics in our model. The element at position (d, k) represents the number of words $w \in V$ in document $d \in D$ which are assigned topic $k \in K$, where V is the set of unique words in the corpus. On the other hand, the matrix $M \in \mathbb{N}^{\mathbf{V} \times \mathbf{K}}$ stores the counts of words belonging to a certain topic within the entire corpus. Hence, element (w, k) gives the count of the words w belonging to topic k.

At each iteration, N and M are recounted and multiplied for each document, topic and word to generate a vector of probabilities p as stated in Equation 3. Each word is then allocated a new topic by sampling from the distribution $\mathbf{P}(Z_{d_j} = k, W_{d_j} = w_{d_j} | \theta, \phi)$.

2.3 LDA with collapsed Gibbs Sampling

Collapsed Gibbs Sampling is performed similarly to traditional Gibbs Sampling, excluding the explicit calculations of θ and ϕ. Instead, Collapsed Gibbs Sampling sweeps each document and word in the corpus and continuously updates the count matrices N and M in order to generate a topic distribution, as shown in Equations 4 and 5.

Our implementation leverages the Numba Python library [2] to accelerate the count matrixes topic distribution updates.

3 Evaluation

The resulting models outputs were compared and evaluated based on their U_{MASS} coherence score, which is defined according to:

$$C(t; V^{(t)}) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{\mathbf{D}(v_m^{(t)}, v_l^{(t)}) + 1}{\mathbf{D}(v_l^{(t)})} \tag{6}$$

where $\mathbf{D}(v_m^{(t)}, v_l^{(t)})$ represents the number of documents d which contain at least one each of the words v_m and v_l for topic t [4]. The integer value M represents the number of top-words used to calculate the coherence score. In the collected results, we always consider the top 20 words. In this report, the top-words were selected based on their frequency of words based on raw count per topic.

3.1 LDA with Collapsed Gibbs Sampling

The Collapsed Gibbs Sampling algorithm was evaluated on multiple settings according to the following set of parameters:

- Corpora: Reuters, 20 News Group
- α: 0.10

- β: 0.10, 0.01
- K: 10, 50
- Iterations: 100, 200

The Collapsed Gibbs sampling algorithm was then run over the entire parameter set. The coherence scores of LDA, parameterized with the aforementioned the combinations, are listed in Table 3.1, for the Reuters and 20 News Group corpora. The U_{MASS} coherence scores are listed as the average of the coherence scores calculated over the top-20 words of each topic.

Table 1: Measured U_{MASS} Coherence score across parameter-space for the Reuters and 20 News Groups corpora respectively, generated with Collapsed Gibbs Sampling

Corpus	N. Topics	Iterations	Alpha	Beta	Avg U_{MASS} Score
reuters	10	100	0.1	0.1	-504.11
reuters	10	100	0.1	0.01	-516.45
reuters	10	200	0.1	0.1	-502.42
reuters	10	200	0.1	0.01	-507.97
reuters	50	100	0.1	0.1	-558.46
reuters	50	100	0.1	0.01	-627.40
reuters	50	200	0.1	0.1	-534.13
reuters	50	200	0.1	0.01	-616.86
20newsgroups	10	100	0.1	0.1	-503.68
20newsgroups	10	100	0.1	0.01	-516.20
20newsgroups	10	200	0.1	0.1	-500.99
20newsgroups	10	200	0.1	0.01	-538.99
20newsgroups	50	100	0.1	0.1	-487.59
20newsgroups	50	100	0.1	0.01	-488.09
20newsgroups	50	200	0.1	0.1	-482.74
20newsgroups	50	200	0.1	0.01	-483.11

After the specified iterations were complete, the resulting word-topic lists were manually searched for interpretable topic allocations. This manual search is summarised in tables 2, 3, which list 5 examples of labeled topic lists for the Reuters and 20 News Group corpora respectively.

3.2 LDA with traditional Gibbs Sampling

The traditional Gibbs Sampling algorithm was evaluated on the same set of parameters as the CGS model, but limited on the Reuters corpus, resulting in the following set of parameters:

- Corpora: Reuters
- α: 0.10
- β: 0.10, 0.01

Table 2: Manually labeled topics generated with CGS on Reuters news corpus with $K = 50$, $\alpha = 0.1$, $\beta = 0.01$, Iterations $= 200$.

War:	Business:	Oil/Gas:	Shipping:	Random:
gulf	assets	oil	tonnes	rise
oil	products	six	due	february
iranian	exports	statement	sources	economy
monday	finance	crude	report	economic
american	fall	effective	tender	ministry
iran	japan	petroleum	july	product
attack	plans	raised	already	inflation
response	house	according	traders	consumer
kuwaiti	committee	raises	august	statistics
war	budget	subsidiary	details	among
asked	reporters	texas	sugar	index
military	think	contract	tonne	forecast
news	nations	gas	french	base
forces	political	barrel	shipment	gross
reagan	miyazawa	natural	exporters	rising
shipping	paris	light	wheat	development
ship	countries	posted	china	south
platform	kiichi	barrels	buys	improvement
waters	germany	postings	rejected	unemployment
defense	parliament	capacity	shipments	number

Table 3: Manually labeled topics generated with CGS on 20 News Group corpus with $K = 50$, $\alpha = 0.1$, $\beta = 0.1$, Iterations $= 200$.

War:	Sports:	Religion:	Tech/Nationality:	Random:
history	game	also	email	would
war	series	god	subject	take
source	win	real	pc	try
near	lost	others	american	run
press	player	life	earth	send
muslims	hit	care	israel	list
muslim	season	body	bill	local
women	hockey	taking	period	statement
report	looked	church	generally	parts
million	chance	jesus	jews	places
men	players	faith	development	pass
population	final	neither	private	printer
died	fans	christ	jewish	box
genocide	fan	seeing	happens	bought
arms	watch	deleted	arab	usual
soviet	played	bible	killed	recommend
former	goal	accept	product	split
argic	hell	sin	friends	drawn
europe	showed	books	totally	foreign
serdar	Canada	evil	truth	lab

- K: 10, 50
- Iterations: 100, 200

The Gibbs sampling algorithm was again run over the entire parameter set. The parameters and the resulting coherence scores are listed in Table 4 for the Reuters. Again, the U_{MASS} coherence scores are listed as the mean of all coherence scores for each topic top-words.

Table 4: Measured U_{MASS} Coherence score across parameter-space for the Reuters corpus, generated with traditional Gibbs Sampling

	N. Topics	Iterations	Alpha	Beta	Avg U_{MASS} Score
0	10	100	0.1	0.1	-396.52
1	10	200	0.1	0.1	-403.82
2	10	100	0.1	0.01	-392.38
3	10	200	0.1	0.01	-408.00
4	50	100	0.1	0.1	-466.21
5	50	200	0.1	0.1	-447.90
6	50	100	0.1	0.01	-459.48
7	50	200	0.1	0.01	-448.23

4 Discussion

Table 3.1 shows the U_{MASS} coherence score for the Reuters and 20 News Groups corpora. Analysing the Reuters values indicates that an increase in number of topics K leads to a decrease in average U_{MASS} coherence, implying that topic allocations become worse for larger K. This is further confirmed by the similarities found in table 4, which lists the resulting U_{MASS} scores on the Reuters dataset produced by traditional Gibbs Sampling. Conversely, the opposite results are seen in the 20 News Group data, where an increase in topics K leads to a decrease in U_{MASS} coherence score, indicating that a larger K gives a more coherent topic allocation. Notably, when analysing the text outputs for the two corpora it is clear that choosing $K = 50$ often results in more intuitive labels, such as the examples shown in tables 2, 3. When selecting $K = 10$, however, there is rarely an obvious theme attached. Please see Appendix 5 for the full output of the best performing $K = 10$ model on the Reuters corpus. We conclude that improvements in U_{MASS} coherence score must not correlate directly with improvements in intuitive word clustering.

Comparing the Reuters U_{MASS} scores in Table 3.1 with the scores in table 4, we note that there is a significant difference despite handling the same dataset. Initially, this was thought to be caused by the stochasticity in the sampling. However, similar differences in score could be observed over multiple runs. Moreover, the resulting topics generated by CGS seem to continuously contain a larger number of intuitive topics than the topics generated by the traditional Gibbs Sampling.

Table 5: Manually labeled topics generated with traditional Gibbs sampling on Reuters news corpus with $K = 50$, $\alpha = 0.1$, $\beta = 0.01$, Iterations = 200.

Trade	Forecasting	Business	Random
forecasts	fluctuations	Minneapolis	union
analyst	ranges	economist	adjusted
funds	experts	china	already
rising	permits	chevron	largely
indicated	5	move	failure
consistent	000	decline	president
need	bottomed	conditions	charged
ranged	higher	spending	nine
since	stewart	continuing	businesses
overseas	nation	annual	protectionist
southmark	profits	protectionist	discussions
oils	1986/87	statistics	fears
often	source	one	lot
letter	farm	asian	raised
level	semiconductors	budget	7.7
levels	measured	staff	comments
&	increases	32	surpluses
liberty	looking	stabilize	economic
less	lme	officials	12.
lifting	deputy	processors	subject

5 Conclusions

In this report we implemented two versions of Gibbs Sampling to perform LDA on the Reuters and 20 News Groups corpora. The models were trained and evaluated on multiple parameter settings and successfully generated interpretable and intuitive word-topics across all settings. Notably, the models performed best when assuming a small topic-count on the Reuters corpus, while they performed best on large topic-counts on the 20 News Group corpus according to U_{MASS} coherence score. Contradictory to the U_{MASS} score, we found that the results were consistently more intuitive when evaluated on large topic-counts.

References

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

[2] Siu Kwan Lam, Antoine Pitrou, and Stanley Seibert. Numba: A LLVM-based Python JIT compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, pages 1–6, 2015.

[3] Jun S. Liu. The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 1994.

[4] David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. EMNLP '11, USA, 2011. Association for Computational Linguistics.

Appendix A

Resulting output from the best performing $K = 10$ topic model on the Reuters corpus:

Table 6: Output for Reuters news corpus with $K = 10$, $\alpha = 0.1$, $\beta = 0.01$, Iterations = 200

said	pct	year	dlrs	revs	cts	would	two	vs	group
mln	note	stock	corp	billion	shr	share	last	net	trade
company	may	sales	market	also	qtr	one	shrs	told	loss
prices	mths	per	june	new	inc	bank	march	six	five
three	shares	avg	total	could	nine	government	expected	rate	years
current	oil	price	dlr	due	april	added	ltd	japan	however
four	exchange	today	week	ended	co	month	end	states	american
purchase	profit	tax	months	assets	first	tonnes	rise	includes	higher
made	investment	interest	acquisition	sale	agreement	major	next	take	industry
financial	international	business	meeting	buy	quarter	earlier	agreed	markets	shareholders
full	cash	february	increase	officials	unit	reported	united	sources	early
period	record	federal	says	subsidiary	make	board	around	dividend	industries
domestic	common	rose	part	lower	securities	president	central	plan	stake
time	based	exports	offer	pay	west	national	products	industrial	move
operating	official	ago	average	operations	eight	report	imports	gain	compared
raised	outstanding	minister	acquire	announced	production	terms	world	high	policy
country	dollar	figures	already	foreign	july	companies	set	much	commission
estimated	since	include	sell	spokesman	fall	general	analysts	development	action
canada	acquired	finance	previously	january	held	levels	recent	japanese	talks
reached	economic	including	seven	agriculture	export	bought	likely	level	european