Interpreting Topic Models in Byte-Pair Encoding Space

Jia Peng Lim
Singapore Management University
jiapeng.lim.2021@smu.edu.sg

Hady W. Lauw
Singapore Management University
hadywlauw@smu.edu.sg

Abstract

Byte-pair encoding (BPE) is pivotal for processing text into chunksize tokens, particularly in Large Language Model (LLM). From a topic modeling perspective, as these chunksize tokens might be mere parts of valid words, evaluating and interpreting these tokens for coherence is challenging. Most, if not all, of coherence evaluation measures are incompatible as they benchmark using valid words. We propose to interpret the recovery of valid words from these tokens as a ranking problem and present a model-agnostic and training-free recovery approach from the topic-token distribution onto a selected vocabulary space, following which we could apply existing evaluation measures. Results show that topic sets recovered from BPE vocabulary space are coherent.

1 Introduction

Byte-pair encoding (BPE) (Gage, 1994; Sennrich et al., 2016) is a popular method of tokenizing *valid* words from a given text onto a token space V_b with a predetermined fixed size, and handling out-of-vocabulary (OOV) words, breaking words into smaller tokens. Given the outsized attention and resources placed on LLM research (Kaddour et al., 2023), we can consider BPE tokens as the new 'meta' language and are likely integrated into future Natural Language Processing (NLP) applications. As LLMs train in a self-supervised manner in V_b and mechanistically investigated (Elhage et al., 2021; Geva et al., 2022) in V_b, therefore, it should also be feasible for topic models to interpret and learn BPE token distributions from V_b.

There are two motivations for investigating Topic Models in V_b. Firstly, topic models (Blei et al., 2003) may reap some practical benefits from BPE, enabling topic models to tackle existing and new challenges, such as vocabulary size constraints, text preprocessing requirements, and analyzing multi-lingual corpora. Additionally, previous downstream NLP applications (Lau et al., 2017; Wang et al., 2020) integrate topic modeling, and future downstream NLP applications may operate on BPE tokens instead of valid words. Since Blei (2012), there has been a concerted push towards qualifying the interpretability of probabilistic topic models, and the main barrier to adopting BPE in topic modeling stems from the difficulty in interpretation, which affects its subsequent evaluation. Being able to interpret BPE tokens will hopefully enable topic models to access previously inaccessible research areas.

Tokens (original)	▁polym formation ▁memb hes ▁force ▁dro ▁el ▁gel astic immer
Words (recovered)	droplet particle swimmer micro fluid deformation stress force elastic gel

Table 1: Real example: interpreting a token distribution where its top 10 tokens are unclear. Using our proposed approach, we recover its top 10 valid words, suggesting a physics-related topic. '▁' denote start of word.

The fundamental challenge of interpretation in V_b is that BPE tokens might be mere parts of, but not in and of themselves wholly, valid words. The conventional approach of evaluating the coherence of topic representations involves examining the likeliest words in the topic-word distribution using: corpus statistics (Mimno et al., 2011; Aletras and Stevenson, 2013; Lau et al., 2014; Röder et al., 2015), word-embeddings (Fang et al., 2016; Terragni et al., 2021b), intruder-based similarity (Thielmann et al., 2024), and prompt engineering (Stammbach et al., 2023). As such, due to their assumption of operating on valid words, conventional topic modeling evaluations are incompatible with BPE tokens from V_b .

Furthermore, conducting human evaluations will be challenging as we do not encounter these BPE

Proceedings of the 31st International Conference on Computational Linguistics, pages 10810–10838
January 19–24, 2025.

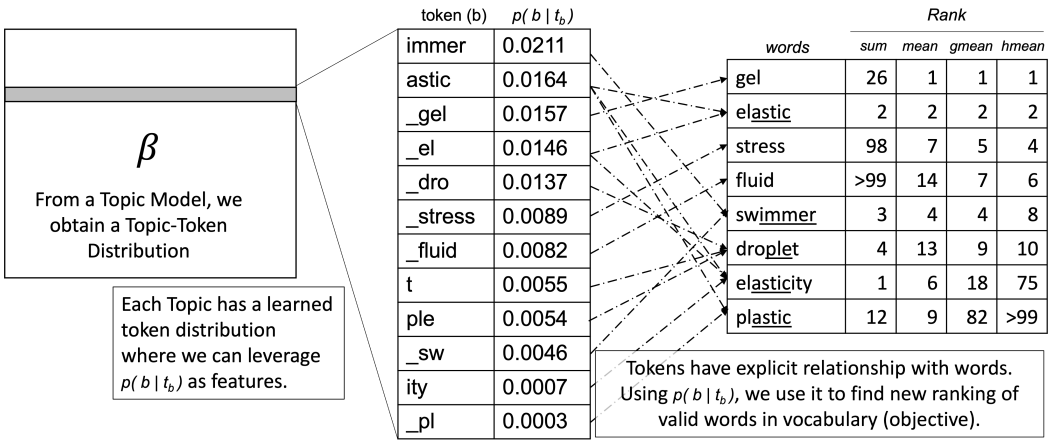


Figure 1: Overview of the proposed interpretation of the word recovery problem as a ranking problem. Different ranking approaches produce different orderings, influencing the coherence evaluation of the topic. This real example came from the same token distribution, seemingly depicting fluid dynamics, as shown in Table 1. Within words, we alternate underlining to denote distinct tokens. Tokens prefixed with '▁' are the start of a word.

tokens naturally. Using the example in Table 1, even if it is possible to assign a coherence score to the token representation, ascertaining the meaning of these tokens might not be possible. To overcome this problem, we propose a model-agnostic and training-free approach to recover valid words in original vocabulary space V_o from the topic-token distributions. With these valid words, we can employ existing methods to evaluate its quality.

This work's contribution consists of proposing a novel perspective of recovering valid words from topic-token distributions. We describe the challenges of interpretation via recovering valid words from V_o (see Section 2) and propose an efficient approach to recovery (see Section 4). From our experiments, we show that recovered topic representations from V_b, evaluated on existing methods, are coherent (see Sections 5, 7) and analogous (see Section 6) to topics learned directly from V_o. We emphasize our work does not seek to replace topic modeling on valid words but rather to explore the feasibility of interpretation in V_b to enable topic model research in new areas.

2 Interpretation Approach

2.1 Notations

We define several notations used in this work. A corpus with original vocabulary V_o can be tokenized to a corpus with BPE vocabulary V_b, with word $w \in V$ broken into its set of BPE tokens B_w. Topic model classes M_o and M_b are trained on corpora with V_o and V_b respectively. With a focus on M_b, its topic-token distribution set β assigns topic-conditional token probabilities $p(b|t_b)$ to every BPE token $b \in V_b$ for each token distribution $t_b \in \beta$. From each t_b, we recover top-ranked $w \in V_o$ based on ranking score $s(w|t_b)$ derived using the information in $p(b|t_b)$. Topic representation r in topics T is a set of top-ranked w and used for evaluation. We note that $s(w|t_b)$ is not a probability, although it can be normalized to fit within the laws of probability, such as using an approximation $\hat{p}(w|t_b) = s(w|t_b) / \sum_{w_j \in V_o} s(w_j|t_b)$ if required.

2.2 From Tokens to Vocabulary

Since LLMs learn concepts in the space of V_b, we should be able to recover coherent topic representations from t_b, utilizing $p(b|t_b)$ to recover and rank words $w \in V$. Motivated by plausible scenarios, we investigate a few methods to determine $s(w|t_b)$ by re-weighing $p(b|t_b)$ in B_w. We illustrate an example of the various approaches in Figure 1.

Information spread across tokens. As multiple words may share the same token, we expect $p(b|t_b)$ to contain information to help recover valid words representative of the topic. Aggregating $p(b|t_b)$ from B_w provides a baseline score for w (Equation 1). From our example, tokens belonging to 'elastic' (rank 2), 'swimmer' (rank 3), and 'droplet' (rank 4) have relatively high $p(b|t_b)$.

$$s_1(w|t_b) = \textit{sum}(w, t_b) = \sum_{b \in B_w} p(b|t_b) \quad (1)$$

Account for token size. A naïve summation as above may afford words with larger B_w an unfair advantage. From our example, when using

#	Type (Ease)	Top 25 Tokens (start/non-start)	Top 25 Words Recovered (*hmean*)
A1	Straightforward (Easy)	__simulation __apply __consider um __result __allow __dat __standard __different __good ing __simple __present __develop __compare __reproduce __prediction __use __base __describe __predict __include __fit __parameter __model (23/2)	
B1	More Start Tokens (Indirect)	cation __reve __behavior __complex ceptor acter __cort __exhib __cha empor __brain __gen ological ike __dynam __protein apt __dynamic __phen __neur __sp __gene __bi __cell ical (17/8)	understand reveal behavioral genetic exhibit chaotic bio pattern genome behavior biotic complex phenomenological cellular cha brain gen protein spike cortical biological dynamic gene dynamical cell
B2	Mixed Tokens (Indirect)	__ale ini ender ar __bott tail qu art __pint hol __alco __glass oda bon onic __co ka __mart __drink __wine pa ila __rum __liqu __whis (12/13)	cig craft mixer beck shot domestic beer marg cigar bev margarita draft tequila ale cock cocktail pint drinker martini alcohol glass mart drink wine rum
C1	Character-Heavyweight (Difficult)	__rotate __leng __ma __bell ough ao aro ony read ape __rim hi atto __fish una days apa ire ordo eless riv ues aco ast __t (7/18)	rotate leng bell treadmill tia mahi tasty tanning trivia tac rim tough fish tao tuesday taro tony tread tape tuna tuesdays tapa tire tasteless taco
C2	Starting Characters (Difficult)	__fal reek __hum oty pr __gy gh __c yy hh __g ww __z oo __x __n __h __q __k __r __o __a __u __i __e (17/8)	rationale eos npr aab aol eighth iowa ure ire hoo koo (not enough words)

Table 2: Real examples of topic representations with the top 25 words recovered from a token distribution with its top 25 tokens shown. Examples are in increasing difficulty of interpretation. See Appendix D for more examples.

sum, words with many tokens like 'elasticity' (rank 1) are ranked highly, while related words such as 'fluid' (rank >99), with a single token, rank poorly despite having a high $p(b|t_b)$. To manage larger B_w, we can use Arithmetic Mean to penalize large B_w containing b with low $p(b|t_b)$ (Equation 2).

$$s_2(w|t_b) = \textit{mean}(w, t_b) = \frac{\sum_{b \in B_w} p(b|t_b)}{|B_w|} \quad (2)$$

With $mean$, we get a more balanced view of these related words, with 'elasticity' (rank 6) and 'fluid' (rank 14) both ranked relatively highly.

Multiplicative property. Probabilities can be multiplicative, and using Geometric Mean's power $1/|B_w|$, we can avoid a vanishing value (Equation 3) while allowing us to minimize scenarios where words over-rely on a single token with strong $p(b|t_b)$. From our example, *sum* and *mean* score 'plastic' highly, as '*astic*' has a very high $p(b|t_b)$. Arguably, the high ranking of 'plastic' may raise doubts as it may have simply benefited from the higher-ranking 'elastic' sharing the same token 'astic'. Applying multiplication allow us to penalize words having tokens with extremely small $p(b|t_b)$, such as '__pl' ($p(b|t_b) = 0.0003$) in 'plastic'.

$$s_3(w|t_b) = \textit{gmean}(w, t_b) = \left(\prod_{b \in B_w} p(b|t_b)\right)^{\frac{1}{|B_w|}} \quad (3)$$

Higher importance on smaller $\mathbf{p(b|t_b)}$. As $b \in B_w$ might not be uniformly important where tokens of low $p(b|t_b)$ may be more important in signaling irrelevance. We can achieve this desired effect by employing Harmonic Mean (Equation 4) to account for the spread between $p(b|t_b)$ in B_w. From our example, 'elasticity' ranks highly in *sum*, *mean*, and *gmean*. Since 'elastic' is similar in rank, we can consider 'elasticity' as a duplicate, evidenced by the low $p(b|t_b)$ of token '*ity*'. Harmonic mean has a known relationship with Geometric and Arithmetic mean where $\textit{hmean} = \textit{gmean}^2/\textit{mean}$.

$$s_4(w|t_b) = \textit{hmean}(w, t_b) = \frac{|B_w|}{\sum_{b \in B_w} \frac{1}{p(b|t_b)}} \quad (4)$$

2.3 Types of Representations Recovered

From the recovered topics from our experiments (see Section 5), using Harmonic Mean on shortlisted top 100 tokens (see Section 4), we observe three groups categorized by the difficulty of initial interpretation and the meaningfulness of the recovered topic representations, exemplified in Table 2. We refer to tokens prefixed with '__' as *start* tokens, as the token signifies the start of a word.

Straightforward and meaningful. Words can exist as a *single* token, with some topics consisting primarily of such words. Interpreting these topics is straightforward; for example, A1 is a topic where we do not have to guess the meaning of the tokens.

Indirect but meaningful. Examples B1 and B2 present scenarios where interpretation is indirect. Usually, the top tokens consist of a mix of start

and non-start tokens. It is possible to guess the theme of the topic based on a few of the tokens, but we might be confused about the presence of other tokens. Our interpretation approach works best for this scenario, able to recover more contextual words from the tokens.

Difficult and subjective. It is possible to recover an incoherent topic representation. Example C1 presents a topic representation with many words starting with "__t", suggesting this start token is dominant in the token distribution. Example C2 cannot produce any valid words with its many single-character start tokens, indicating that this token distribution deals with the nature of BPE rather than concepts from the corpus. Evaluating the coherence of such topic representations is challenging as they are subjective and typically penalized in traditional evaluation methods.

3 Experimental Setup

3.1 Corpora Processing

We select four large corpora of different themes (Table 3) and process them along suggested guidelines in Hoyle et al. (2021). Refer to Appendix A.1 for additional processing details. **arXiv** Dataset[1] is a collection of abstracts, in areas of scientific research, submitted to the e-print archive. **UN** General Debate corpus (Jankin et al., 2017; Baturo et al., 2017) is a collection of statements on geopolitical issues at United Nations General Debates from 1970–2022. **Yahoo**! Answers, prepared by Zhang et al. (2015)[2], we use the best answers from the curated ten largest main categories. **Yelp** Review Full prepared by Zhang et al. (2015) from Yelp Dataset Challenge (2015)[3], comprises of reviews on businesses in metropolitan locations in the United States and Canada.

3.2 Tokenizer Selection

We use LLaMA's (Touvron et al., 2023) choice of tokenizer, SentencePiece's (Kudo and Richardson, 2018) BPE, as a large language model's tokenizer should be sufficient to reduce the vocabulary space, avoiding the need to create specific tokenizer models for different corpora. Table 3 describes the tokenized corpus, showing that vocabulary space required decreased substantially for all corpora.

3.3 Metrics Selection

Normalized Pointwise Mutual Information (NPMI) (Bouma, 2009). In Equation 5, $p(w_i, w_j)$ and $p(w)$ are probabilities of word(s) occurring in a sliding window. Using a sliding window of size 10, a document d consists of $\max(1, |d| - 10)$ windows. A small ϵ is included to prevent undefined logarithm zero when $p(w_i, w_j) = 0$.

$$\text{NPMI}(w_i, w_j) = -\frac{\log \frac{p(w_i,w_j)+\epsilon}{p(w_i)p(w_j)}}{\log p(w_i, w_j) + \epsilon} \quad (5)$$

The NPMI score of a topic representation r is the mean NPMI value between word pairs $w_i, w_j \in r$, measuring frequency of co-occurrence between words in r. Since NPMI based on Wikipedia corpus statistics (NPMI_W) has been found to correlate to human judgment (Lau et al., 2014; Röder et al., 2015; Lim and Lauw, 2023b), we can use NPMI_W as a proxy measure for the interpretability of r. We use the Wikipedia statistics[4] from Lim and Lauw (2023b, 2024) where the reported mean correlation to human judgement from several study groups is at $\bar{\rho} = 0.66$.

Topic Uniqueness (TU) (Dieng et al., 2020). A diverse set of topics will contain few duplicate words. The TU of a topic set T is the ratio of its unique words to words (Equation 6).

$$\text{TU}(T) = \frac{|\{w|w \in r, r \in T\}|}{\sum_{r \in T} |r|} \quad (6)$$

3.4 Topic Models Selection

We select two traditional topic models[5] and two neural topic models that are popular: Latent Dirichlet Allocation with Gibbs sampling (Blei et al., 2003; Newman et al., 2009) (**LDA**), Dirichlet-Multinomial Regression Topic Model (Mimno and McCallum, 2008) (**DMR**), autoencoder-based **ProdLDA**[6] (Srivastava and Sutton, 2017), and SBERT-embddings (Reimers and Gurevych, 2019) augmented autoencoder-based CombinedTM[7] (Bianchi et al., 2021) (**CTM**). Traditional topic models use inverse document frequency weighing scheme (Wilson and Chew, 2010).

[1] kaggle.com/datasets/Cornell-University/arxiv

[2] Using Yahoo! Answers Comprehensive Questions and Answers version 1.0 at webscope.sandbox.yahoo.com.

[3] kaggle.com/datasets/yelp-dataset/yelp-dataset

[4] Pre-processed NPMI downloadable from https://github.com/PreferredAI/topic-metrics/

[5] tomotopy implementation (Lee, 2022).

[6] Implementation by Carrow (2018)

[7] https://github.com/MilaNLProc/contextualized-topic-models

Corpus	Theme	Docs.	Words	Tokens (ratio)	<unk>	$\lvert V_o \rvert$	$\lvert V_b \rvert$	$\lvert V_b \cap V_o \rvert$ (%)	$\lvert V_b \cap V_w \rvert$
arXiv	science	2.31M	189.1M	249.2M (1.32)	3.170%	31713	10108	4792 (15.1%)	4649
UN	geopol.	320K	14.3M	21.4M (1.49)	1.120%	15828	8342	3664 (23.2%)	4097
Yahoo	news	880K	35.7M	45.8M (1.28)	5.290%	35019	11395	5574 (15.9%)	5229
Yelp	reviews	658K	39.0M	52.2M (1.34)	1.930%	33451	11189	5374 (16.1%)	4971

Table 3: Statistics of Corpora with respective vocabulary V_o and BPE vocabulary space V_b. Rare words are replaced by <unk> tokens. A % of corpus vocabulary V_o and Wikipedia vocabulary V_w exists as a single token in V_b.

	Method	sum			mean			gmean			hmean		
	↑ / ↑ / ↓	$m(T)$	$m(\hat{T})$	$\lvert \triangle \rvert$	$m(T)$	$m(\hat{T})$	$\lvert \triangle \rvert$	$m(T)$	$m(\hat{T})$	$\lvert \triangle \rvert$	$m(T)$	$m(\hat{T})$	$\lvert \triangle \rvert$
arXiv	LDA	-0.033	0.040	0.073	-0.007	0.044	0.051	0.054	0.055	0.001	**0.063**	**0.063**	**0**
	DMR	-0.037	0.034	0.071	-0.012	0.040	0.052	0.052	0.053	0.001	**0.058**	**0.058**	**0**
	ProdLDA	-0.024	**0.055**	0.079	-0.010	0.041	0.051	0.044	0.048	0.004	**0.053**	0.053	**0**
	CTM	-0.009	**0.084**	0.093	0.017	0.071	0.054	0.071	0.074	0.003	**0.074**	0.075	**0.001**
UN	LDA	-0.040	0.021	0.061	-0.016	0.030	0.046	**0.042**	0.040	0.002	0.041	**0.041**	**0**
	DMR	-0.042	0.023	0.065	-0.016	0.030	0.046	0.041	0.039	0.002	**0.043**	**0.043**	**0**
	ProdLDA	-0.057	0.043	0.100	0.039	0.047	0.008	0.051	0.048	**0.003**	**0.053**	**0.050**	**0.003**
	CTM	-0.071	**0.066**	0.137	0.058	0.062	0.004	**0.064**	0.063	**0.001**	0.063	0.062	**0.001**
Yahoo	LDA	-0.072	0.015	0.087	-0.014	0.030	0.044	0.042	0.042	**0**	**0.047**	**0.047**	**0**
	DMR	-0.075	0.013	0.088	-0.014	0.031	0.045	0.044	0.044	**0**	**0.049**	**0.049**	**0**
	ProdLDA	-0.070	-0.013	0.057	0.035	0.040	0.005	0.041	0.042	0.001	**0.044**	**0.044**	**0**
	CTM	-0.053	0.078	0.131	0.057	0.099	0.042	0.097	**0.100**	0.003	**0.100**	**0.100**	**0**
Yelp	LDA	-0.086	-0.025	0.061	-0.056	-0.006	0.050	0.012	0.013	0.001	**0.028**	**0.028**	**0**
	DMR	-0.086	-0.024	0.062	-0.054	-0.005	0.049	0.015	0.015	**0**	**0.028**	**0.028**	**0**
	ProdLDA	-0.092	-0.033	0.059	-0.039	-0.019	0.020	-0.016	-0.015	0.001	**-0.012**	**-0.012**	**0**
	CTM	-0.102	-0.020	0.082	-0.006	0.005	0.011	0.006	0.008	0.002	**0.010**	**0.010**	**0**

Table 4: Results comparing interpretation methods on topic sets produced using full vocabulary (T) and shortlisted vocabulary ($\hat{T}$), using the topic-token distributions producing $K = 100$ topics. $\lvert \triangle \rvert = m(T) - m(\hat{T})$, the difference of mean NPMI between topic sets T and $\hat{T}$, with 0 being the ideal result. NPMI calculated on respective corpus statistics on topic representations of size 10. The results displayed are the mean of 5 independent training runs.

4 Efficient Word Candidate Shortlist

Recovering topic representations from token distributions let us use existing metrics for evaluation. To determine the topic representation r of token distribution t_b, we must compute all $s(w|t_b)$ for all w in V_o. However, the purpose of applying BPE is to reduce the vocabulary space into a more controllable token space V_b, and hence computing all $s(w|t_b)$ is contradictory. We propose to shortlist candidate words to overcome this paradox and show its efficacy and potential pitfalls.

Methodology. Instead of using the entire V_o space, we shortlist a subset of vocabulary, and compute $s(w|t_b)$ for w if B_w is in the top 100 tokens with the largest $p(b|t_b)$ in t_b. To compare efficacy, we measure the performance gap $\lvert \triangle \rvert$ between mean NPMI of topic representations from full computation T and shortlisted candidates $\hat{T}$. $\lvert \triangle \rvert = 0$ implies that the shortlisted candidates produces the same result as a full computation. For each class of M_b, we train five independent models, using the same hyper-parameters optimized on M_o[8].

Results. From Table 4, across multiple models and corpora, using Harmonic Mean (Equation 4) recovers similar topic representations in both full computation and shortlisted candidates. Geometric Mean (Equation 3)'s results are slightly inferior to Harmonic Mean's. Sum (Equation 1) and Arithmetic Mean (Equation 2) have large differences in $\lvert \triangle \rvert$, recovering topic representations that occur more frequently together in shortlisted candidates compared to those from a full computation. On average, we exploit less than 1% of V_o when using shortlisted candidates.

Discussion. Within the proposed approaches, using Harmonic Mean to recover topic representations from the shortlisted candidates seems to be the best, with its $\lvert \triangle \rvert$ closest to 0. Its efficacy in T and $\hat{T}$ suggests that b with smaller $p(b|t_b)$ is important, and surfaces words with large $p(b|t_b)$ across its tokens in recovered r. Even though Sum and Arithmetic Mean can empirically produce moderate results for $\hat{T}$, the large $\lvert \triangle \rvert$ with T implies a potential pitfall in its interpretation process reliant on the shortlisting procedure and casting an illusion of improvement.

[8] Details in Appendix A.

	arXiv				UN				Yahoo				Yelp			
	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W
LDA	0.200	0.84	0.202	0.82	0.208	0.81	0.152	0.81	0.235	0.86	0.180	0.86	0.213	0.83	0.165	0.79
gmean	0.149	**0.86**	0.149	**0.86**	0.108	**0.84**	0.113	**0.88**	0.186	0.89	0.151	**0.90**	0.100	**0.86**	0.083	**0.88**
hmean	0.153	0.85	0.149	0.84	0.120	0.82	0.114	0.86	0.193	0.88	0.153	**0.90**	0.118	0.85	0.093	0.86
RPS	0.167	0.85	0.180	0.83	0.121	0.80	0.107	0.83	0.186	0.88	0.154	0.89	0.129	0.83	0.106	0.84
RST	**0.173**	0.85	**0.184**	0.84	**0.136**	0.83	**0.127**	0.87	**0.205**	**0.90**	**0.163**	**0.90**	**0.137**	0.85	**0.116**	0.86
DMR	0.199	0.85	0.200	0.82	0.204	0.80	0.149	0.80	0.248	0.85	0.192	0.85	0.213	0.82	0.161	0.80
gmean	0.144	**0.86**	0.145	**0.86**	0.106	**0.82**	0.123	**0.85**	0.185	**0.89**	0.152	0.89	0.099	**0.86**	0.082	**0.88**
hmean	0.149	0.85	0.147	0.85	0.122	0.81	0.124	0.84	0.193	0.88	0.154	0.88	0.118	0.85	0.093	0.85
RPS	0.167	0.84	0.180	0.83	0.122	0.78	0.117	0.82	0.188	0.88	0.157	0.88	0.126	0.84	0.107	0.84
RST	**0.174**	0.85	**0.186**	0.84	**0.138**	**0.82**	**0.138**	**0.85**	**0.206**	**0.89**	**0.166**	**0.90**	**0.134**	0.84	**0.116**	0.85
Prod.	0.200	0.97	0.173	0.97	0.178	0.95	0.113	0.90	0.166	0.97	0.126	0.95	0.133	0.98	0.074	0.93
gmean	0.125	**0.90**	0.117	**0.90**	0.105	0.89	0.090	0.82	0.116	0.89	0.095	0.90	0.055	0.91	0.038	0.89
hmean	0.131	**0.90**	0.115	0.89	0.110	0.89	0.090	0.82	0.120	0.89	0.097	0.90	0.060	0.90	0.039	0.89
RPS	0.147	**0.90**	0.137	0.88	0.125	**0.90**	0.083	**0.87**	0.120	0.91	0.099	**0.92**	0.074	0.90	0.043	0.89
RST	**0.150**	**0.90**	**0.140**	0.89	**0.127**	**0.90**	**0.098**	0.83	**0.128**	**0.92**	**0.104**	**0.92**	**0.078**	0.90	**0.049**	**0.90**
CTM	0.203	0.94	0.189	0.93	0.177	0.88	0.120	0.84	0.241	0.94	0.198	0.93	0.145	0.93	0.100	0.83
gmean	0.134	**0.87**	0.133	**0.85**	0.120	0.82	0.105	0.74	0.192	0.87	0.157	0.88	0.075	0.83	0.056	0.78
hmean	0.136	0.86	0.131	0.84	0.122	0.82	0.105	0.74	0.193	0.87	0.156	0.88	0.077	0.84	0.057	0.78
RPS	0.153	0.86	0.154	0.83	0.134	**0.84**	0.097	**0.79**	0.192	**0.88**	0.156	**0.89**	0.086	**0.86**	0.065	**0.80**
RST	**0.155**	0.86	**0.157**	0.83	**0.137**	**0.84**	**0.113**	0.76	**0.199**	**0.88**	**0.160**	**0.89**	**0.092**	**0.86**	**0.070**	**0.80**

Table 5: Evaluating topic sets of top 50 scoring representations from models with $K = 100$ trained on original corpora M_o (bolded), serving as benchmarks, and BPE-tokenized corpora M_b with respective recovery methods. Wikipedia reference corpus is used to calculate NPMI_W and shortlist top words in topic representations of size 10. Heuristics RPS and RST extends from hmean. Results are the mean of 15 independent runs. Standard deviation of NPMI is less than 0.01, and TU is less than 0.03. Bolded results denote best among recovery methods, and RST has the closest NPMI score to respective benchmarks. See Table 9 in Appendix B for $K = 200$ results.

Semantically Different Subsets. A closer examination of recovered r reveals the possibility of shortlisting semantically dissimilar words with similar token subsets. Consider Example B2, where most words describe consumables in a lounge/bar/pub (see Table 2), the words 'mart' and 'cock' seems out-of-place, with their inclusion attributed to 'martini' and 'cocktail' which are relevant to the implied topic. Hence, we can consider the semantic information provided in 'mart' (1 token) and 'cock' (2 tokens) as *duplicates*. To address duplication, we further propose two independent heuristics extensions to remove possible duplicates from the top 25 shortlisted words: (1) **remove proper subsets** (RPS), words that exist in another word, and (2) **remove single tokens** (RST) of duplicated short-length words ($|w| \leq 4$).

5 Topic Representation Evaluation

Methodology. We employ NPMI and TU (see Section 3.3) to evaluate the coherence and diversity of the topic representations r, comparing topic sets from topic models M_o and M_b trained on corpora with V_o and V_b respectively, 15 independent training runs each. For M_o, we use similar training configurations in Section 8. Since Geometric Mean's (gmean) results are similar to Harmonic Mean's (hmean), we further optimize the hyper-parameters of M_b for hmean results. We evaluate NPMI twice, using word statistics from the respective corpus (NPMI_C) and Wikipedia (NPMI_W). Some topics are difficult to evaluate (see Section 2.3), with subsets of respective corpus vocabulary not found in Wikipedia's vocabulary. Therefore, for each corpus statistic, we restrict evaluation to the top 50 out of K scoring r where $|r| = 10$, using TU_C and TU_W to denote their respective TU. We use M_o as a benchmark since it trains and evaluates in V_o space, serving as a plausible upper bound in performance for M_b, trained in V_b space, and thus disadvantaged in evaluation on V_o.

Results. From Table 5, amongst the recovery methods applied on M_b, RST extended on hmean has the closest NPMI scores to their respective M_o benchmarks across examined corpora. Their absolute NPMI_W scores also suggest that the recovered r is likely to be coherent, which we further investigate in Section 7. Diversity-wise, the examined recovery methods have similar TU scores in most experimental settings. For the subsequent sections, we use RST and hmean to recover r for evaluation.

Topic Sets from:	LDA$_o$	LDA$_b$	DMR$_o$	DMR$_b$	Prod.$_o$	Prod.$_b$	CTM$_o$	CTM$_b$
LDA$_o$	**4.9**	4.4	4.9	4.4	3.1	3.2	3.2	3.4
LDA$_b$	4.3	**4.9**	4.3	4.9	2.7	3.3	2.9	3.4
DMR$_o$	4.9	4.4	**4.9**	4.4	3.1	3.2	3.2	3.4
DMR$_b$	4.3	4.9	4.3	**4.9**	2.7	3.3	2.9	3.4
Prod.$_o$	3.1	2.8	3.1	2.8	**4.4**	3.1	3.9	3.0
Prod.$_b$	3.2	3.3	3.1	3.3	3.1	**4.9**	3.1	4.1
CTM$_o$	3.3	3.0	3.2	3.0	3.9	3.2	**4.6**	3.6
CTM$_b$	3.4	3.6	3.4	3.6	3.1	4.1	3.6	**4.9**

(a) Average maximum duplication of topic representation r from topic set T compared to topic set T' from target model class M'. Standard deviation of results is less than or around 0.2.

Top-50 Topics from:	LDA$_o$	LDA$_b$	DMR$_o$	DMR$_b$	Prod.$_o$	Prod.$_b$	CTM$_o$	CTM$_b$
LDA$_o$	**7.9**	7.0	8.0	7.0	5.3	5.3	5.4	5.6
LDA$_b$	7.1	**8.1**	7.1	8.1	4.8	5.5	5.0	5.8
DMR$_o$	7.9	7.0	**8.0**	7.0	5.2	5.2	5.3	5.5
DMR$_b$	7.1	8.0	7.0	**8.1**	4.9	5.6	4.9	5.8
Prod.$_o$	5.1	4.6	5.1	4.7	**7.2**	5.1	6.4	4.8
Prod.$_b$	5.3	5.5	5.3	5.6	5.4	**7.4**	5.5	6.6
CTM$_o$	5.3	4.9	5.2	4.9	6.6	5.2	**7.6**	5.7
CTM$_b$	5.5	5.8	5.5	5.8	5.0	6.4	5.8	**7.5**

(b) Average optimistic maximum duplication of Top-50 NPMI$_W$ topic representations from T compared to topic set T' from target model class M'. Standard deviation of results is less than or around 2.

Table 6: arXiv ($K = 100$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V. Results for UN, Yahoo, Yelp are similar (see Appendix B, Tables 10, 11, 12). Results for $K = 200$ are also similar (see Appendix B, Tables 13, 14, 15, 16).

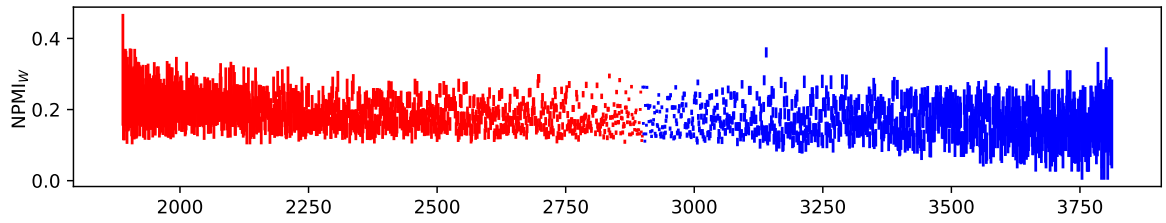


Figure 2: Visualizing NPMI difference between pairs of r with 5 common words from LDA ($K = 100$) on arXiv. For each r_b in each Top-50 $T \in \text{LDA}_b$, we select closest r_o in each $T' \in \text{LDA}_o$. Each vertical line is a pair of r, sorted by difference in NPMI$_W$; red lines on the left shows pairs where $r_o > r_b$, while blue lines on the right shows pairs where $r_b > r_o$. This visualization is consistent for other M_b/M_o across the other corpora (see Appendix C).

6 Similarity Between Topic Sets?

Previously, we used TU to measure the number of unique words in a topic set. In contrast, to determine the similarity between topic sets, the most straightforward approach is to find similar pairs of r across models containing common words.

Methodology. For a given r and target topic set T', we define maximum duplication (MD) as the largest possible subset of words between $r' \in T'$ and r (Equation 7). A high value indicates that r's concept exists in T'.

$$\text{MD}(r, T') = \max\{|r \bigcap r'| : r' \in T'\} \quad (7)$$

Since we have multiple training runs, we compute the mean of MD (MMD, Equation 8) of each $r \in T$, comparing topics sets across model classes $T \in M$ to $T' \in M'$.

$$\text{MMD}(T, T') = \frac{1}{|T|} \sum_{r \in T} \text{MD}(r, T') \quad (8)$$

However, some topics may only be present in some $T' \in M'$. Accounting for this scenario, we consider optimistic duplication (OD, Equation 9).

$$\text{OD}(T, M') = \tfrac{1}{|T|} \sum_{r \in T} \max\{\text{MD}(r, T') : T' \in M'\} \quad (9)$$

We compare across model class pairs, skipping models with similar indexes, and report the mean of the comparisons (Table 6). As pairs of r may have similar words but different NPMI scores, we visualize the quantitative difference (Figure 2).

Results. From both Table 6a (MMD results) and Table 6b (OD results), the similarity between M_o and M_b is no worse than the similarity between different classes of M_o. Visualizing NPMI$_W$ difference in Figure 2, comparing r_b from M_b and its closest r_o from M_o, shows that some $r_b > r_o$ and other $r_b < r_o$. Overall results suggest that modeling on different V_o and V_b captures analogous concepts from the corpus, where similar pairs of r with common words have a wide range of difference in NPMI scores.

7 User Study

While there is an expected difference in NPMI scores between M_o and M_b, the absolute NPMI score of M_b suggests that recovered r_b from M_b is coherent. We conduct a user study where we recruit and poll human participants for their opinions.

Q	Topic	NPMI_W
a	LDA_b: allergy bad **cooked** eat **food** **fry** greasy greed microwave **taste**	0.071
	LDA_o: chicken **cooked** dry flavor **food** **fry** order sauce **taste** wing	0.154
b	LDA_b: air density **energy** fuel **heat** liquid pressure **temperature** volume **water**	0.157
	LDA_o: atom carbon electron **energy** gas **heat** mole molecule **temperature water**	0.217

Table 7: Examples of topic pairs shown to participants.

User Study Design. Extending on Section 6, in each question, we present two different r, r_1 and r_2, with a common subset of four to five words, and three accompanying sub-questions[9]:

1. Is r_1 coherent?
2. Is r_2 coherent?
3. In terms of coherence, is $r_1 > r_2$, or $r_1 < r_2$, or $r_1 = r_2$?

We randomly sample 45 pairs of r, r_o from LDA_o and r_b from LDA_b, that fits our criteria with equal numbers from UN, Yahoo, and Yelp. We also generated *dummy* examples as substitutes, made from a subset of common words from its paired r_o or r_b and random words in V_o. There are three different kinds of pairs shown to our participants:

1. 30 r_o-r_b pairs where $r_o > r_b$ in terms of NPMI_W[10]. For a fair study, we account for the NPMI advantage of LDA_o, preventing possible sampling of r_b with higher NPMI.
2. 12 r-dummy pairs, with instances of r equally split between r_o and r_b.
3. 3 dummy-dummy pairs for verification.

Results. Treating responses as a poll, we aggregate the responses from 10 study participants[11]. Starting from questions with r_o-r_b pairs, using sub-question 1 and 2, 248/300 (82.6%) responded positively to r_b, compared to 274/300 (91.3%) for r_o. For sub-question 3, in terms of coherence, 51/300 (17%) responses for $r_b > r_o$, 116/300 (38.7%) responses for $r_o > r_b$, 123/300 (41%) responses consider the pair to be similar in coherence, with the remaining minority similarly incoherent. For questions with r-dummy pairs, in terms of coherence, 52/60 (86.7%) considers $r_b >$ dummy, and 54/60 (90%) considers $r_o >$ dummy. Overall results suggest that recovered r_b is coherent.

[9] See Appendix E for the exact phrasing, instructions, and r given to participants. Pairs are presented in a random order, i.e., r_1 can be r_b, r_o, or dummy.

[10] Some pairs have small NPMI_W difference (see App. E.3)

[11] They have graduate/post-graduate qualifications.

8 Discussion Relating to LLMs

While our work does not involve LLMs, only using their tokenizers, we expect questions related to LLMs because of their popularity.

Primarily, there are two key challenges to interpreting token distributions in LLMs. First, the observed phenomenon of superposition/polysemy of single neurons (Elhage et al., 2022) implies multiple 'topics' attributed to a lone neural activation. Second, tokenizers with a larger token per word performance are less interpretable, breaking words up into many more sub-words. Our work tackles the latter challenge. We show that if a BPE token distribution codes for one topic with a non-trivial token set, we can recover its word representations for the topic.

Application-wise, there is a desire to control an LLM's output, for reasons such as trust and safety. Apart from the many prompting strategies, there are architectures, such as Backpack Models (Hewitt et al., 2023), that allow intervention on the model's hidden state to shape the model's output. An area of future work may encompass incorporating topic modeling methodologies to reduce reliance on crafted prompts to guide LLM generation.

9 Related Work

Topic Models and applications. Advances in neural networks allowed for Neural Topic Models (NTM) (Miao et al., 2016) to replace traditional sampling approaches (Blei et al., 2003). Recent works on NTM utilizes additional techniques, such as clustering (Grootendorst, 2022), diffusion model (Xu et al., 2023), graph neural networks (Zhang et al., 2023), and incorporating LLMs (Sarkar et al., 2023). Topic modeling methodologies have been applied in various tasks such as stance detection (Arakelyan et al., 2023), dialogue summarization

(You and Ko, 2023), entity disambiguation (Xiao et al., 2023), multimodal relation extraction (Wu et al., 2023), and modeling document networks (Zhang and Lauw, 2020, 2022).

Word recovery. The focus of our work is on recovering the topic representation of t_b generated by a topic model discovered from some corpus in the V_b space. There are previous works that recover phrases from words. Blei and Lafferty (2009) visualize significant phrases from unigram topic models using co-occurrence statistics. Yu et al. (2013) and Li et al. (2019) propose specific phrase topic models for biomedicine and large corpora respectively.

10 Conclusion

This work seeks to recover, interpret, and evaluate topic representations from BPE token distributions, exploring a few straightforward methods. The core insight to interpretation lies in understanding how a word's semantic information spreads amongst its tokens (Section 2), with the meaningfulness of any approach stemming from the consistency of application (Section 4), and finally, deriving a ranking of valid words for conventional evaluation. From our evaluation (Section 5), we recommend using Harmonic Mean and RST to recover topic representations from token distributions from V_b, which are analagous to topic representations learnt from V_o (Section 6), and coherent to human judgment (Section 7). We hope that this work enables new exploration in topic model research areas.

Acknowledgments

This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-PhD-2023-08-055T).

Limitations

Tokenizer choice. We only use LLaMA's BPE tokenizer in our experiments, and our results may depend on our selection of tokenizer. However, other tokenizers might not be suitable, such as GPT-2's BPE tokenizer (Radford et al., 2019), as they may have a larger token space containing more valid words and thus unable to reduce the original vocabulary size by as much. Since it is more difficult to interpret coherent concepts from LLaMA's BPE token space compared to GPT-2 (Lim and Lauw, 2023a), we investigate the former. While fitting a custom BPE tokenizer on a corpus is possible, evaluating an existing tokenizer is much more convenient and meaningful for generalizability.

Projection to selected vocabulary space. We must curate a vocabulary space V_o to recover words from the token distributions. In our experiments, we shortlist words occurring above a certain threshold. If only rare words are selected, the topic representation recovered may just be as uninterpretable from the human perspective.

Language. Our corpora primarily contain English text and are evaluated against Wikipedia-EN. However, there are some documents containing non-English words, as we found topics consisting of German and French words, and further research is required for validation on non-English corpora.

Modeling small corpora in V_b may not be always beneficial. We repeat the experiments using similar models on smaller corpora: 20NewsGroup, BBC News, DBLP, and M10 from OCTIS (Terragni et al., 2021a), producing topic sets of size 20, from V_b with a similar size to V_o. For traditional topic models, LDA and DMR, we can recover coherent topics. However, neural topic models ProdLDA and CTM collapse to a few topics.

Ethics Statement

This work adheres to the ACL code of ethics. Our user study was approved by our Institutional Review Board, with its participants recruited via word of mouth, and paid US$15 for answering 45 questions with an estimated 0.5 hours of work. To the best of our abilities, we do not foresee any potential risk in our work. We use datasets widely used in academic settings; Wikipedia uses CC BY-SA 3.0 DEED, arXiv, and UN use CC0: Public Domain, while Yahoo and Yelp have customized licenses that allow for non-profit academic use. Software-wise, we use publicly available libraries and code repositories with an MIT License.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 13448–13464, Toronto, Canada. Association for Computational Linguistics.

Alexander Baturo, Niheer Dasandi, and Slava J. Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics*, 4(2):2053168017712821.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

David M. Blei and John D. Lafferty. 2009. Visualizing topics with multi-word expressions. *Preprint*, arXiv:0907.1013.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Biennial GSCL Conference 2009*.

Stephen Carrow. 2018. Pytorchavitm: Open source avitm implementation in pytorch. Github.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Preprint*, arXiv:2209.10652.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1057–1060, New York, NY, USA. Association for Computing Machinery.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

John Hewitt, John Thickstun, Christopher Manning, and Percy Liang. 2023. Backpack language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9103–9125, Toronto, Canada. Association for Computational Linguistics.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: The incoherence of coherence. In *Neural Information Processing Systems*.

Slava Jankin, Alexander Baturo, and Niheer Dasandi. 2017. United Nations General Debate Corpus 1946-2022.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *Preprint*, arXiv:2307.10169.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 355–365, Vancouver, Canada. Association for Computational Linguistics.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Minchul Lee. 2022. bab2min/tomotopy: 0.12.3.

Baoji Li, Wenhua Xu, Yuhui Tian, and Juan Chen. 2019. A phrase topic model for large-scale corpus. In *2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 634–639. IEEE.

Jia Peng Lim and Hady Lauw. 2023a. Disentangling transformer language models as superposed topic models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8646–8666, Singapore. Association for Computational Linguistics.

Jia Peng Lim and Hady Lauw. 2023b. Large-scale correlation analysis of automated metrics for topic models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13874–13898, Toronto, Canada. Association for Computational Linguistics.

Jia Peng Lim and Hady W. Lauw. 2024. Aligning human and computational coherence evaluations. *Computational Linguistics*, 50(3):893–952.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA.

David Mimno and Andrew McCallum. 2008. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, page 411–418, Arlington, Virginia, USA. AUAI Press.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

David Newman, Arthur Asuncion, Padhraic Smyth, and Max Welling. 2009. Distributed algorithms for topic models. *Journal of Machine Learning Research*, 10(62):1801–1828.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*, pages 399–408. ACM.

Souvika Sarkar, Dongji Feng, and Shubhra Kanti Karmaker Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16218–16233, Singapore. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Dominik Stammbach, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Revisiting automated topic model evaluation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9348–9357, Singapore. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021a. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270. Association for Computational Linguistics.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021b. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems*, pages 33–45, Cham. Springer International Publishing.

Anton Thielmann, Arik Reuter, Quentin Seifert, Elisabeth Bergherr, and Benjamin Säfken. 2024. Topics in the Haystack: Enhancing Topic Quality through Corpus Expansion. *Computational Linguistics*, pages 1–36.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Andrew T. Wilson and Peter A. Chew. 2010. Term weighting schemes for Latent Dirichlet Allocation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 465–473, Los Angeles, California. Association for Computational Linguistics.

Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. 2023. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, Toronto, Canada. Association for Computational Linguistics.

Zilin Xiao, Linjun Shou, Xingyao Zhang, Jie Wu, Ming Gong, and Daxin Jiang. 2023. Coherent entity disambiguation via modeling topic and categorical dependency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7480–7492, Singapore. Association for Computational Linguistics.

Weijie Xu, Wenxiang Hu, Fanyou Wu, and Srinivasan Sengamedu. 2023. DeTiME: Diffusion-enhanced topic modeling using encoder-decoder based LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9040–9057, Singapore. Association for Computational Linguistics.

Jaeah You and Youngjoong Ko. 2023. Topic-informed dialogue summarization using topic distribution and prompt-based modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5657–5663, Singapore. Association for Computational Linguistics.

Zhiguo Yu, Todd R Johnson, and Ramakanth Kavuluru. 2013. Phrase based topic modeling for semantic information processing in biomedicine. In *2013 12th International Conference on Machine Learning and Applications*, volume 1, pages 440–445. IEEE.

Ce Zhang and Hady W. Lauw. 2020. Topic modeling on document networks with adjacent-encoder. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6737–6745.

Delvin Ce Zhang and Hady Lauw. 2022. Dynamic topic models for temporal document networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 26281–26292. PMLR.

Delvin Ce Zhang, Rex Ying, and Hady W. Lauw. 2023. Hyperbolic graph topic modeling network with continuously updated topic tree. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 3206–3216, New York, NY, USA. Association for Computing Machinery.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

A Supplementary Model Details

A.1 Corpora Processing Details

We use spaCy[12] model "en_core_web_trf" and lemminflect[13] for lemmatization. We remove documents that have less than 10 words and lowercase words in the remaining document. We replace rare words, which occur below a certain threshold, with <unk> . For each corpus, to increase compatibility with Wikipedia's vocabulary, we use heuristics to break up portmanteau words, such as 'supermatrices' to 'super' and 'matrix'.

arXiv. We consider each abstract as a document. We replace rare words that occur less than 50 times.

UN. We consider each paragraph in a statement as a document. We replace rare words that occur less than 10 times.

Yahoo. We consider each best answer from a question as a document. We replace rare words that occur less than 20 times.

Yelp. We consider each review as a document. We replace rare words that occur less than 10 times.

A.2 Model Hyper-parameter Details

In Section 4, for all topic models used, we optimized the hyper-parameters on the original corpus, and used the same hyper-parameters to train on the BPE-tokenized corpus. For each original and modified corpus, we train five independent models.

In Section 5, we further optimize the hyper-parameters for topic models training on BPE-tokenized corpus. We also increase the number of training runs for all topic models from 5 to 15.

For traditional topic models, we train 1000 epochs, after which we find the increase in performance is marginal. We train models with two different initial term weights, inverse document frequency (IDF) and uniform, with IDF producing better results. Our findings also apply to models trained with uniform initial term weights. We optimize on α and η hyper-parameters. In the case of neural topic models, we train up to 200 epochs, early stopping when validation loss does not decrease for eight epochs. We optimize on dropout, number of layers, and neurons per layer. We use combinatorial linear search across a range of hyper-parameter values. We list the hyper-parameters used in Table 8.

		LDA_o		DMR_o		ProdLDA_o			CTM_o			LDA_b		DMR_b		ProdLDA_b			CTM_b		
Corpus	K	α	η	α	η	L	neu.	D	L	neu.	D	α	η	α	η	L	neu.	D	L	neu.	D
arXiv	100	1	1e-4	1	1e-4	1	4096	0	2	4096	0	1	1e-2	1	1e-2	1	3072	0	2	3072	0
	200	1	1e-4	1	1e-4	2	4096	0	2	4096	0.1	1	1e-4	1	1e-4	1	3072	0	2	2048	0.1
UN	100	1	1e-2	1	1e-2	1	4096	0.1	1	4096	0.1	1	1e-3	1	1e-2	2	1024	0.1	2	2048	0.1
	200	1	1e-2	1	1e-2	2	4096	0.1	2	2048	0.1	1	1e-2	1	1e-4	2	3072	0.1	1	2048	0.1
Yahoo	100	1	1e-3	0.5	1e-3	1	2048	0.1	2	2048	0	1	1e-3	1	1e-3	1	1024	0.1	2	1024	0
	200	1	1e-4	1	1e-4	2	4096	0.1	2	2048	0	1	1e-4	1	1e-3	2	512	0.2	1	512	0.1
Yelp	100	1	1e-2	1	1e-2	1	2048	0.1	2	2048	0.1	1	1e-2	1	1e-2	1	1536	0.1	2	2048	0.1
	200	1	1e-2	1	1e-2	1	4096	0.1	2	4096	0.1	1	1e-3	1	1e-2	2	2048	0.1	2	4096	0.1

Table 8: Hyper-parameter settings used in each model type. Abbrv: layers (L), neurons (neu.), and dropout (D).

A.3 Compute Environment

Experiments were run on machines configured with NVIDIA A40 GPUs, AMD EPYC 7763 CPUs, and 512GB of RAM. Training a model does not require more than a few hours.

[12] spacy.io

[13] pypi.org/project/lemminflect/

B Supplementary Tables

This section contains additional tables of information:

- Table 9 shows the main evaluation results for models with $K = 200$.
- Tables 10, 11, 12 shows Topic Similarity statistics for Models with $K = 100$.
- Tables 13, 14, 15, 16 shows Topic Similarity statistics for Models with $K = 200$.

	arXiv				UN				Yahoo				Yelp			
	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W	NPMI_O	TU_C	NPMI_W	TU_W
LDA	0.228	0.86	0.229	0.84	0.251	0.83	0.184	0.83	0.273	0.88	0.217	0.87	0.251	0.87	0.198	0.78
gmean	0.161	**0.88**	0.158	**0.89**	0.153	**0.82**	0.145	**0.88**	0.218	**0.89**	0.172	**0.91**	0.137	**0.85**	0.107	**0.88**
hmean	0.165	**0.88**	0.158	0.88	0.163	0.81	0.144	0.86	0.225	0.88	0.172	0.90	0.151	**0.85**	0.117	0.86
+RPS	0.182	0.87	0.188	0.86	0.169	0.79	0.134	0.83	0.218	0.88	0.175	0.90	0.159	0.83	0.127	0.83
+RST	**0.187**	**0.88**	**0.194**	0.87	**0.181**	0.81	**0.157**	0.87	**0.235**	**0.89**	**0.184**	**0.91**	**0.167**	0.84	**0.137**	0.84
DMR	0.227	0.86	0.228	0.83	0.252	0.84	0.184	0.83	0.272	0.88	0.216	0.88	0.251	0.87	0.198	0.77
gmean	0.160	**0.88**	0.155	**0.89**	0.138	**0.83**	0.117	**0.87**	0.228	0.88	0.179	**0.90**	0.138	**0.84**	0.107	**0.88**
hmean	0.165	0.87	0.158	0.88	0.149	**0.83**	0.121	0.86	0.235	0.87	0.180	0.89	0.155	**0.84**	0.121	0.87
+RPS	0.182	0.86	0.187	0.85	0.154	0.81	0.110	0.82	0.229	0.88	0.186	**0.90**	0.160	0.82	0.127	0.83
+RST	**0.188**	0.87	**0.191**	0.86	**0.162**	0.82	**0.128**	0.86	**0.245**	**0.89**	**0.194**	**0.90**	**0.168**	**0.84**	**0.141**	0.85
Prod.	0.214	0.94	0.186	0.92	0.187	0.87	0.115	0.82	0.171	0.97	0.145	0.97	0.151	0.96	0.091	0.89
gmean	0.139	**0.90**	0.132	**0.88**	0.124	0.86	0.104	0.75	0.106	0.93	0.093	0.93	0.088	0.84	0.051	0.80
hmean	0.145	0.89	0.133	0.86	0.127	0.86	0.102	0.76	0.105	0.93	0.091	0.93	0.090	0.84	0.054	0.79
+RPS	0.160	0.88	0.156	0.84	0.142	**0.88**	0.095	**0.82**	0.120	0.95	0.103	0.94	0.108	**0.87**	0.060	0.80
+RST	**0.161**	0.88	**0.158**	0.85	**0.140**	0.87	**0.110**	0.77	**0.124**	**0.95**	**0.106**	**0.95**	**0.110**	0.86	**0.064**	**0.82**
CTM	0.190	0.98	0.119	0.96	0.192	0.79	0.128	0.75	0.264	0.89	0.226	0.90	0.171	0.91	0.129	0.82
gmean	0.132	**0.95**	0.083	0.94	0.138	**0.84**	0.126	0.68	0.152	0.94	0.127	0.94	0.104	0.86	0.078	0.77
hmean	0.130	**0.95**	0.079	**0.93**	0.139	0.83	0.125	0.68	0.153	0.94	0.127	0.93	0.105	0.85	0.079	0.76
+RPS	**0.152**	**0.95**	**0.101**	**0.93**	0.156	**0.84**	0.119	**0.77**	0.161	**0.96**	0.135	**0.96**	0.118	0.88	0.089	0.76
+RST	0.151	**0.95**	0.100	**0.93**	**0.157**	**0.84**	**0.134**	0.70	**0.168**	**0.96**	**0.140**	0.95	**0.121**	**0.88**	**0.092**	**0.78**

Table 9: Evaluating topic sets of top 50 scoring representations from models with $K = 200$ trained on original corpora M_o (bolded), serving as benchmarks, and BPE-tokenized corpora M_b with respective recovery methods. Results are the mean of 15 independent runs. NPMI_W calculated using Wikipedia reference corpus statistics and used to shortlist top words in topic representations of size 10. Heuristics RPS and RST extends on hmean. Bolded results denote best among recovery methods, and RST has the closest NPMI score to respective benchmarks. Standard deviation of NPMI is less than 0.01, and TU is less than 0.03

Topic Sets from:	LDA_o	LDA_b	DMR_o	DMR_b	Prod._o	Prod._b	CTM_o	CTM_b
LDA_o	**5.8**	4.2	5.7	4.4	3.3	3.2	3.8	3.5
LDA_b	3.8	**4.7**	3.8	4.7	2.5	3.0	2.8	3.1
DMR_o	5.7	4.1	**5.7**	4.3	3.3	3.2	3.7	3.5
DMR_b	4.0	4.7	4.0	**5.0**	2.5	3.1	2.8	3.3
Prod._o	3.2	2.6	3.2	2.6	**3.6**	2.8	3.4	2.8
Prod._b	3.1	3.2	3.1	3.3	2.8	**4.2**	2.9	3.9
CTM_o	3.8	3.1	3.7	3.1	3.5	3.0	**4.1**	3.3
CTM_b	3.5	3.5	3.5	3.6	2.9	4.1	3.3	**4.5**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.2.

Top-50 Topics from:	LDA_o	LDA_b	DMR_o	DMR_b	Prod._o	Prod._b	CTM_o	CTM_b
LDA_o	**8.3**	6.4	8.3	6.5	5.5	5.1	5.9	5.4
LDA_b	6.2	**7.7**	6.2	7.7	4.7	5.5	5.0	5.7
DMR_o	8.2	6.4	**8.2**	6.5	5.5	5.0	5.9	5.4
DMR_b	6.4	7.9	6.4	**8.2**	4.7	5.6	5.0	5.9
Prod._o	4.9	4.3	5.0	4.2	**6.2**	4.6	5.6	4.5
Prod._b	4.5	5.0	4.5	5.0	4.6	**6.7**	4.6	6.2
CTM_o	5.4	4.6	5.4	4.5	5.6	4.7	**6.4**	5.0
CTM_b	4.9	5.4	4.9	5.4	4.6	6.2	5.0	**6.8**

(b) Average word similarity of each Top-50 NPMI_W topic representation and its most-similar topic representations from all independent models in target M. S.D ± 2.

Table 10: UN ($K = 100$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**5.3**	4.5	5.3	4.5	2.6	2.4	3.8	3.5
	LDA_b	4.3	**5.0**	4.3	5.0	2.0	2.4	3.1	3.5
	DMR_o	5.4	4.6	**5.6**	4.6	2.5	2.4	3.8	3.6
	DMR_b	4.3	5.0	4.4	**5.0**	2.0	2.4	3.2	3.5
	$Prod._o$	2.5	2.1	2.5	2.2	**3.0**	2.2	2.8	2.4
	$Prod._b$	2.4	2.5	2.4	2.5	2.2	**2.9**	2.4	2.7
	CTM_o	3.9	3.5	4.0	3.5	2.8	2.5	**4.8**	3.9
	CTM_b	3.7	3.9	3.7	3.9	2.5	2.8	4.0	**4.9**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.2.

Top-50 Topics from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**8.2**	7.0	8.2	7.0	4.6	4.1	6.3	5.7
	LDA_b	6.9	**8.2**	7.0	8.2	4.0	4.5	5.5	6.1
	DMR_o	8.1	6.9	**8.3**	6.8	4.4	4.2	6.2	5.7
	DMR_b	6.9	8.1	6.9	**8.2**	4.0	4.5	5.6	6.1
	$Prod._o$	4.0	3.7	4.0	3.6	**5.2**	3.8	4.4	3.8
	$Prod._b$	3.8	4.1	3.8	4.1	4.0	**5.3**	4.1	4.5
	CTM_o	6.0	5.4	6.1	5.5	4.5	4.2	**7.5**	6.0
	CTM_b	5.5	5.9	5.5	6.0	4.1	4.9	6.2	**7.6**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 2.

Table 11: Yahoo ($K = 100$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**6.1**	4.6	6.0	4.6	2.1	2.4	2.8	2.8
	LDA_b	4.4	**5.3**	4.3	5.3	1.6	2.4	2.2	2.6
	DMR_o	6.0	4.6	**6.0**	4.6	2.1	2.4	2.8	2.7
	DMR_b	4.4	5.4	4.4	**5.4**	1.6	2.4	2.2	2.6
	$Prod._o$	1.9	1.6	1.9	1.6	**3.9**	1.9	2.9	1.9
	$Prod._b$	2.3	2.4	2.3	2.4	1.9	**4.1**	2.2	3.4
	CTM_o	2.8	2.3	2.8	2.3	3.0	2.3	**3.9**	2.7
	CTM_b	2.9	2.9	2.9	2.9	2.1	3.6	2.8	**4.5**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.2.

Top-50 Topics from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**8.7**	6.8	8.7	6.8	3.7	3.8	4.6	4.4
	LDA_b	6.6	**8.2**	6.6	8.3	3.2	4.0	3.9	4.5
	DMR_o	8.7	6.7	**8.6**	6.7	3.6	3.7	4.5	4.3
	DMR_b	6.6	8.2	6.6	**8.3**	3.2	4.0	4.0	4.5
	$Prod._o$	3.4	2.9	3.4	2.8	**6.4**	3.7	5.0	3.4
	$Prod._b$	3.5	3.7	3.5	3.7	3.6	**6.4**	3.8	5.5
	CTM_o	4.2	3.7	4.2	3.7	5.0	3.9	**6.3**	4.5
	CTM_b	4.2	4.3	4.2	4.2	3.7	5.6	4.7	**6.9**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 2.

Table 12: Yelp ($K = 100$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**4.4**	3.6	4.4	3.6	2.7	2.8	1.3	2.0
	LDA_b	3.5	**3.9**	3.5	4.0	2.3	2.7	1.1	1.9
	DMR_o	4.4	3.6	**4.4**	3.6	2.7	2.8	1.3	2.0
	DMR_b	3.5	4.0	3.5	**4.0**	2.3	2.7	1.1	1.9
	$Prod._o$	3.0	2.6	3.0	2.6	**4.3**	3.0	1.7	1.9
	$Prod._b$	2.9	2.9	2.9	2.9	2.9	**4.4**	1.2	2.2
	CTM_o	1.1	1.0	1.1	1.0	1.6	0.9	**3.4**	1.4
	CTM_b	1.9	1.9	1.9	1.9	1.7	2.1	1.5	**3.9**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.1.

Top-50 Topics from:		LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
	LDA_o	**7.9**	6.7	7.9	6.7	5.2	5.2	2.4	3.9
	LDA_b	6.8	**7.5**	6.8	7.5	4.7	5.4	2.2	3.8
	DMR_o	8.0	6.7	**8.0**	6.7	5.3	5.2	2.5	3.9
	DMR_b	6.8	7.5	6.8	**7.5**	4.7	5.4	2.2	3.8
	$Prod._o$	5.0	4.5	5.0	4.4	**7.1**	5.0	3.2	3.6
	$Prod._b$	5.3	5.3	5.2	5.2	5.1	**6.7**	2.3	4.2
	CTM_o	2.3	1.9	2.4	1.9	3.2	1.7	**5.6**	2.6
	CTM_b	3.5	3.4	3.6	3.4	3.2	3.9	2.7	**5.9**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 1.5.

Table 13: arXiv ($K = 200$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**5.5**	4.1	5.4	3.7	3.3	3.0	3.6	3.2
LDA_b	3.7	**4.9**	3.7	4.0	2.4	2.8	2.5	2.8
DMR_o	5.4	4.1	**5.4**	3.6	3.2	2.9	3.5	3.2
DMR_b	3.3	3.9	3.3	**3.7**	2.3	2.5	2.4	2.6
$Prod._o$	3.5	2.8	3.5	2.8	**3.7**	2.9	3.7	3.0
$Prod._b$	3.0	3.2	3.0	2.9	2.8	**3.8**	2.9	3.7
CTM_o	3.9	3.3	3.9	3.1	3.8	3.1	**4.4**	3.5
CTM_b	3.4	3.5	3.4	3.1	3.0	3.9	3.4	**4.3**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.1.

Top-50 Topics from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**8.5**	6.6	8.5	6.2	5.1	4.8	5.5	5.3
LDA_b	6.5	**8.4**	6.5	7.4	4.5	5.5	4.8	5.7
DMR_o	8.4	6.6	**8.4**	6.1	5.1	4.7	5.5	5.2
DMR_b	6.3	7.3	6.3	**7.4**	4.7	5.2	4.9	5.4
$Prod._o$	5.0	4.2	5.0	4.3	**5.8**	4.4	5.7	4.5
$Prod._b$	4.3	4.9	4.3	4.5	4.2	**6.1**	4.3	6.0
CTM_o	5.5	4.7	5.4	4.7	5.6	4.5	**6.3**	5.0
CTM_b	4.5	5.2	4.5	4.6	4.3	6.1	4.6	**6.8**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 1.5.

Table 14: UN ($K = 200$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**4.7**	3.8	4.7	4.0	2.0	1.9	3.3	2.3
LDA_b	3.6	**4.1**	3.6	4.2	1.6	1.8	2.6	2.1
DMR_o	4.7	3.8	**4.7**	4.0	2.0	1.9	3.3	2.3
DMR_b	3.8	4.3	3.8	**4.6**	1.6	1.8	2.7	2.2
$Prod._o$	2.0	1.7	2.0	1.8	**2.8**	1.8	2.2	2.0
$Prod._b$	1.8	1.8	1.8	1.9	1.7	**2.9**	1.7	2.8
CTM_o	3.8	3.2	3.7	3.3	2.4	2.1	**4.6**	2.6
CTM_b	2.4	2.3	2.4	2.4	2.0	2.8	2.5	**3.5**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.1.

Top-50 Topics from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**8.1**	6.7	8.1	6.9	3.6	3.4	6.0	4.0
LDA_b	6.9	**7.8**	7.0	7.9	3.4	3.5	5.5	4.2
DMR_o	8.0	6.6	**8.0**	6.8	3.6	3.4	6.0	3.9
DMR_b	7.2	8.0	7.2	**8.3**	3.4	3.6	5.5	4.2
$Prod._o$	3.4	3.0	3.3	3.0	**4.9**	3.0	3.3	3.2
$Prod._b$	3.3	3.2	3.2	3.2	3.1	**5.2**	2.9	4.9
CTM_o	5.8	5.1	5.7	5.2	3.6	3.0	**7.2**	3.9
CTM_b	3.9	3.8	3.8	3.9	3.4	4.9	3.9	**6.0**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 1.5.

Table 15: Yahoo ($K = 200$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

Topic Sets from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**5.7**	4.0	5.6	4.2	2.3	2.5	2.8	2.7
LDA_b	3.8	**4.5**	3.7	4.6	1.7	2.3	2.1	2.4
DMR_o	5.6	4.0	**5.6**	4.2	2.3	2.5	2.7	2.6
DMR_b	4.0	4.6	4.0	**5.0**	1.7	2.4	2.2	2.4
$Prod._o$	2.3	1.9	2.3	1.9	**3.5**	2.0	2.8	2.1
$Prod._b$	2.6	2.7	2.6	2.7	2.0	**3.7**	2.4	3.5
CTM_o	3.0	2.5	2.9	2.5	2.9	2.4	**3.8**	2.8
CTM_b	2.9	2.9	2.9	2.9	2.2	3.6	2.8	**4.1**

(a) Average word similarity of each topic representation and its most-similar topic representations from each independent model in target M. S.D ± 0.1.

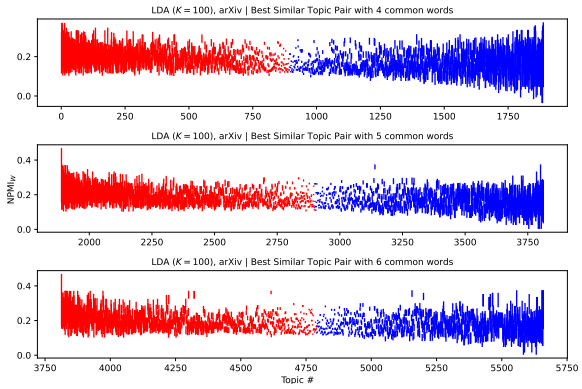
Top-50 Topics from:	LDA_o	LDA_b	DMR_o	DMR_b	$Prod._o$	$Prod._b$	CTM_o	CTM_b
LDA_o	**8.7**	6.5	8.7	6.7	3.9	4.0	4.8	4.2
LDA_b	6.5	**8.0**	6.5	7.9	3.3	4.3	4.1	4.5
DMR_o	8.7	6.6	**8.6**	6.8	3.9	4.0	4.8	4.3
DMR_b	6.7	8.1	6.7	**8.4**	3.3	4.2	4.1	4.5
$Prod._o$	3.9	3.3	3.9	3.3	**6.0**	3.5	4.9	3.6
$Prod._b$	4.0	4.3	4.0	4.2	3.6	**6.0**	4.1	5.7
CTM_o	4.4	3.9	4.4	3.8	4.8	3.7	**6.3**	4.4
CTM_b	4.2	4.4	4.2	4.3	3.7	5.5	4.6	**6.4**

(b) Average word similarity of each Top-50 $NPMI_W$ topic representation and its most-similar topic representations from all independent models in target M. S.D ± 1.5.

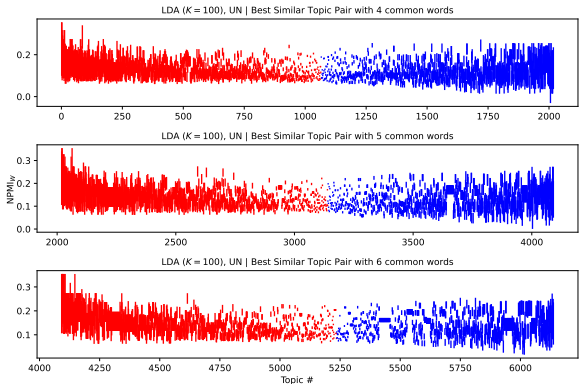
Table 16: Yelp ($K = 200$) Topic Similarity statistics showing duplication across different models M_o, trained on V_o, and M_b, trained on V_b. Bolded values on the diagonals serve as a benchmark comparing similar M trained on similar V.

C Supplementary Figures

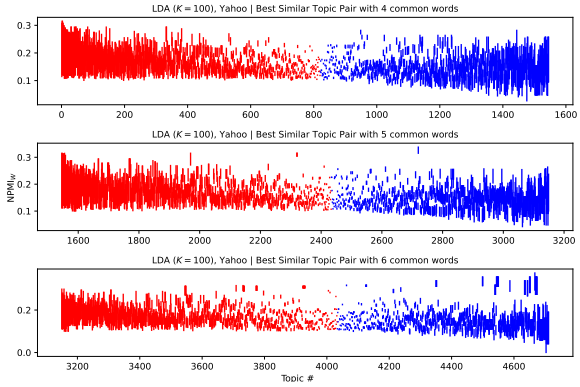
In each subfigure, we visualize NPMI difference between pairs of r with 4, 5, and 6 common words from **LDA** ($K = 100$). For each r_b in each Top-50 $T \in M_b$, we select closest r_o in each $T' \in M_o$. Each vertical line is a pair of r, sorted by difference in NPMI_W scores, where red lines on the left shows pairs where $r_o > r_b$, while blue lines on the right shows pairs where $r_b > r_o$. Visualization is similar for $K = 200$.



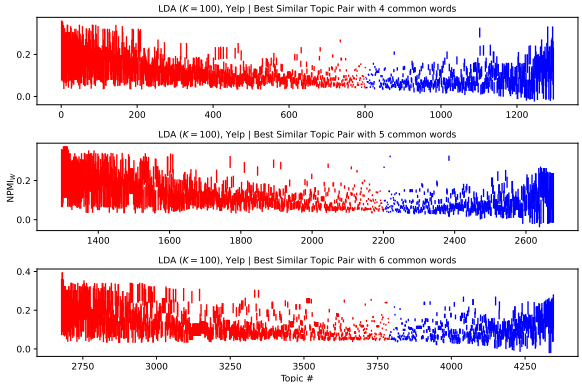
(a) arXiv corpus.



(b) UN corpus.

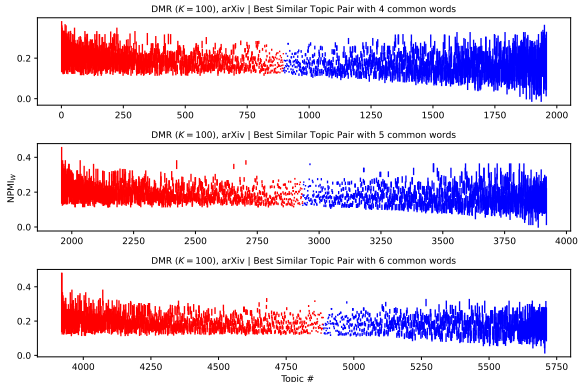


(c) Yahoo corpus.

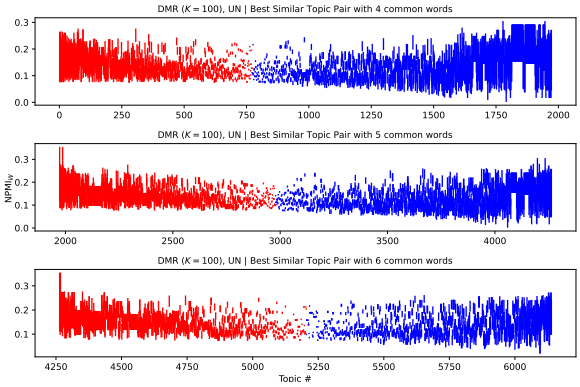


(d) Yelp corpus.

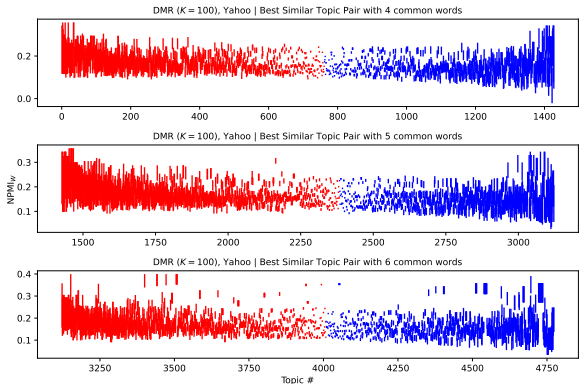
Figure 3: **LDA** ($K = 100$)



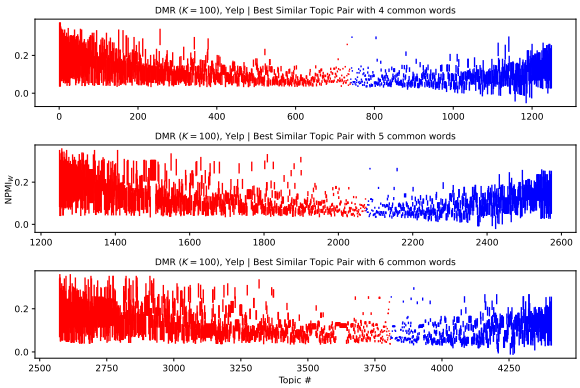
(a) arXiv corpus.



(b) UN corpus.

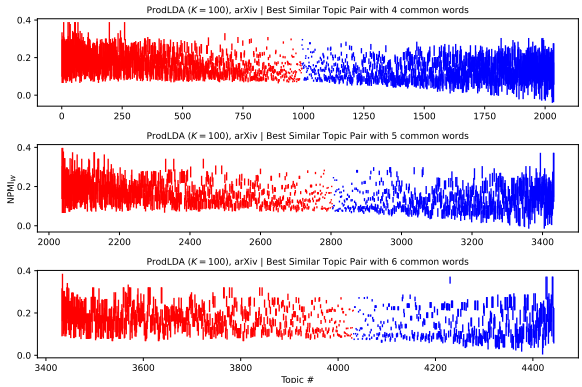


(c) Yahoo corpus.

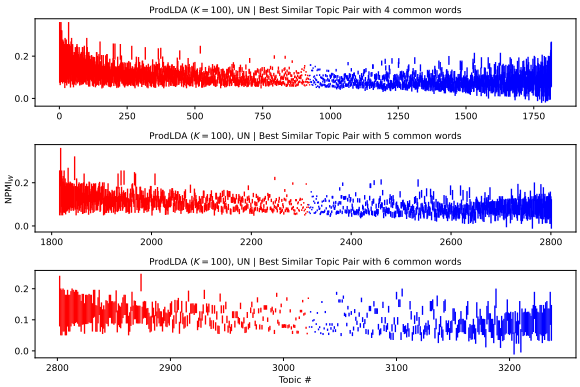


(d) Yelp corpus.

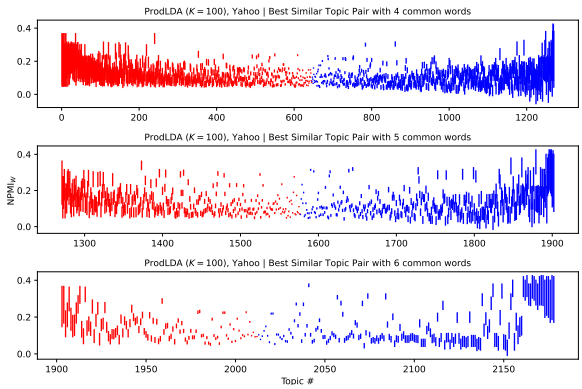
Figure 4: **DMR** ($K = 100$)



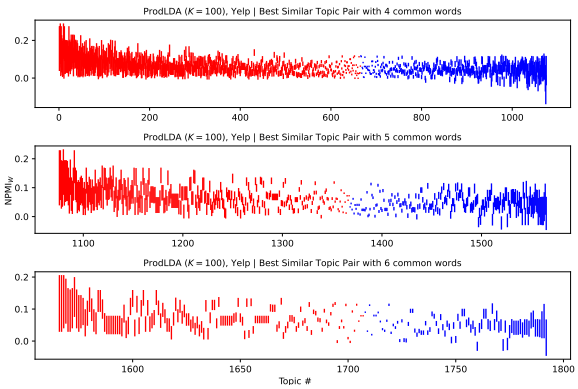
(a) arXiv corpus.



(b) UN corpus.

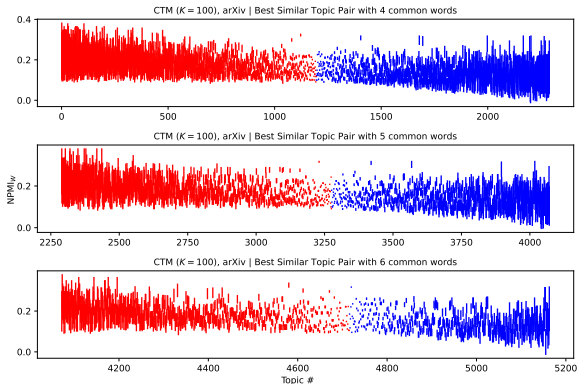


(c) Yahoo corpus.

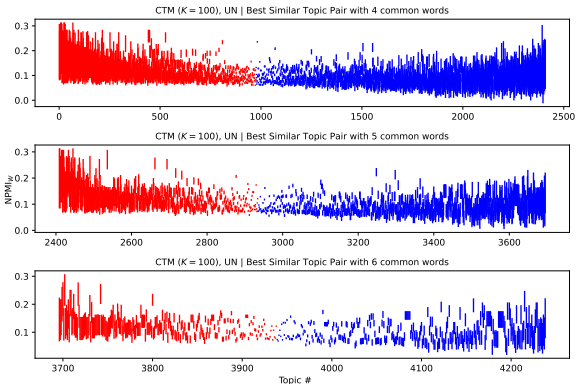


(d) Yelp corpus.

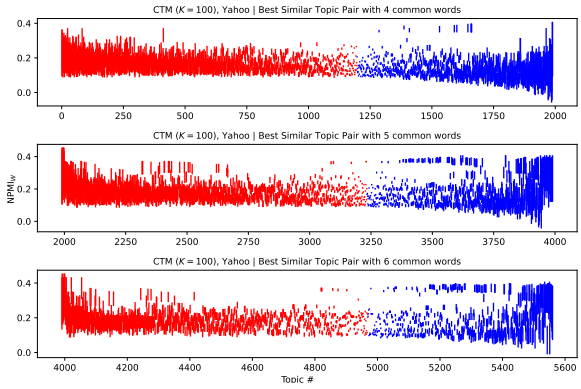
Figure 5: **ProdLDA** ($K = 100$)



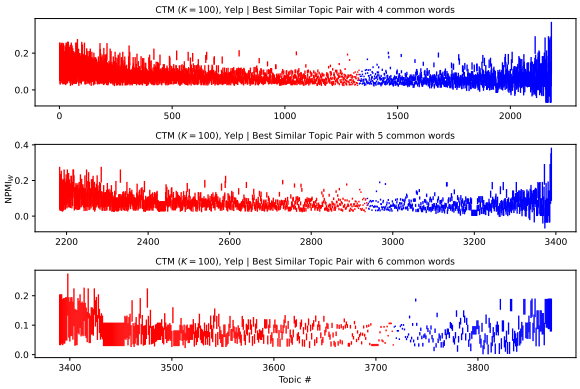
(a) arXiv corpus.



(b) UN corpus.



(c) Yahoo corpus.



(d) Yelp corpus.

Figure 6: **CTM** ($K = 100$)

D Supplementary Examples

These random examples are selected from random training runs' Top-50 NPMI$_W$.

#	Top 10 Words [Num. Tokens] Recovered (NPMI$_W$)
	LDA ($K = 100$)
1	error[1] fuzzy[3] problem[1] element[1] grid[1] equation[1] mesh[1] scheme[1] numerical[1] method[1] (0.106)
2	packing[2] viscous[3] sphere[1] emc[2] shear[2] particle[1] fluidity[2] fluid[1] granular[2] viscosity[3] (0.107)
3	cancer[1] clinical[2] brain[1] health[1] disease[1] diagnosis[2] segment[1] image[1] medical[1] patient[1] (0.144)
4	experimentation[2] violation[2] sterile[2] beta[1] decay[1] lepton[2] oscillation[2] theta[2] mixing[1] neutrino[3] (0.154)
5	membrane[2] nanoparticle[3] droplet[3] surface[1] droop[2] liquid[1] polymer[2] water[1] molecule[3] molecular[2] (0.157)
6	receiver[1] antenna[2] scheme[1] transmission[1] network[1] wireless[1] user[1] transmit[1] communication[1] channel[1] (0.162)
7	protocol[1] coherence[3] circuit[1] entangled[3] gate[1] entangle[2] state[1] qubit[2] entanglement[3] quantum[1] (0.178)
8	infrared[3] reciprocal[3] electron[1] light[1] ultra[2] optical[1] beam[1] laser[2] wavelength[3] pulse[2] (0.208)
9	cosmology[2] gravity[1] matter[1] inflationary[3] perturbation[2] scalar[1] dark[1] inflation[2] cosmological[2] universe[1] (0.214)

10	geodesic[3] diffeomorphism[3] riemann[2] surface[1] riemannian[3] space[1] symplectic[3] metric[1] curvature[2] manifold[1] (0.350)
	DMR ($K = 100$)
1	calculation[1] nucleon[2] discrepancy[4] isotope[3] energy[1] isp[2] shell[1] neutron[2] nuclear[1] nucleus[2] (0.134)
2	viscosity[3] shear[2] liquid[1] friction[2] surface[1] wall[1] force[1] particle[1] fluid[1] flow[1] (0.180)
3	elasticity[3] mechanical[1] micro[1] deformation[2] displacement[2] grain[2] material[1] stress[1] strain[2] elastic[2] (0.180)
4	virus[1] specie[1] infect[2] covid-19[5] epidemic[3] disease[1] pandemic[2] spread[1] population[1] covid[2] (0.181)
5	polytope[3] complex[1] sheaf[2] algebra[1] module[1] ideal[1] variety[1] homology[2] cohomology[3] ring[1] (0.187)
6	time[1] stochastic[2] monte-carlo[5] process[1] simulation[1] diffusion[1] chain[1] markov[2] monte[2] carlo[2] (0.188)
7	periodic[1] dynamic[1] harmonic[2] coherent[3] resonator[2] frequency[1] coherence[3] system[1] oscillation[2] oscillator[2] (0.205)
8	cyclic[2] automorphism[2] module[1] semigroup[3] finite[1] representation[1] lie[1] subgroup[1] algebra[1] group[1] (0.233)
9	spectrum[1] gamma[2] observation[1] kev[2] pulsar[3] gamma-ray[4] source[1] radio[1] emission[1] x-ray[3] (0.235)
10	homotopy[3] associative[2] theory[1] module[1] monoid[2] noncommutative[4] commutative[2] functor[2] algebra[1] category[1] (0.308)
	ProdLDA ($K = 100$)
1	boson[2] lhc[3] new[1] mhc[3] particle[1] halo[2] dark[1] matter[1] higgs[3] coupling[1] (0.076)
2	lense[2] large[1] power[1] use[1] lensing[3] redshift[2] density[1] lens[2] galaxy[2] find[1] (0.099)
3	observatory[2] sensitivity[2] survey[1] mission[1] resolution[1] detection[1] astro[2] instrument[1] detector[2] telescope[2] (0.117)
4	spectral[1] higgs[3] hilbert[3] eigenvector[2] halo[2] eigenvalue[2] operator[1] hamiltonian[3] hamilton[2] matrix[1] (0.117)
5	voltage[1] conductivity[2] conduct[1] conductor[2] junction[2] superconductivity[4] current[1] superconducte[4] superconductor[4] super[1] (0.132)
6	transition[1] state[1] chiral[2] degenerate[3] lattice[1] vortex[2] condensate[3] topological[1] symmetry[1] phase[1] (0.162)
7	hadron[2] meson[2] chiral[2] charm[1] plasma[2] qubit[2] qcd[2] strange[1] heavy[1] quark[2] (0.164)
8	calculation[1] obtain[1] density[1] coulomb[3] approximation[1] calculate[1] particle[1] interaction[1] energy[1] potential[1] (0.176)
9	symplectic[3] admit[1] legend[1] closed[1] holomorphic[2] hypersurface[4] metric[1] ricci[2] curvature[2] manifold[1] (0.191)
10	effect[1] treatment[1] cancer[1] diagnosis[2] result[1] disease[1] clinical[2] therapy[2] study[1] patient[1] (0.207)
	CTM ($K = 100$)
1	roll[1] vacation[2] fluctuation[4] perturb[1] scalar[1] inflationary[3] universe[1] cosmological[2] perturbation[2] inflation[2] (0.092)
2	execution[1] design[1] workload[2] software[1] application[1] cpu[1] code[1] hardware[1] bug[1] performance[1] (0.126)

3	dust[1] emission[1] glow[2] gamma-ray[4] gcr[2] gev[1] energy[1] cosmic[2] galactic[2] gamma[2] (0.140)
4	beam[1] wavefront[2] elect[1] wavelength[3] light[1] scatter[1] mode[1] frequency[1] optical[1] waveguide[2] (0.149)
5	boundary[1] topological[1] function[1] continuous[1] define[1] compact[1] map[1] point[1] measure[1] space[1] (0.151)
6	asymmetry[3] baseline[2] oscillation[2] experimentation[2] violation[2] mixing[1] lepton[2] decay[1] experiment[1] neutrino[3] (0.164)
7	experimenter[2] neutron[2] dark[1] high[1] electron[1] beam[1] energy[1] detector[2] experiment[1] particle[1] (0.167)
8	plasma[2] hydrodynamic[3] large[1] fluid[1] numerical[1] velocity[1] instability[2] turbulent[3] simulation[1] turbulence[3] (0.198)
9	encryption[1] internet[1] server[1] security[1] communication[1] authentication[1] protocol[1] client[1] privacy[2] secure[1] (0.220)
10	scattering[2] exciton[3] field[1] electron[1] fermion[2] interaction[1] boson[2] phonon[3] particle[1] photon[2] (0.319)

Table 17: Supplementary Examples from arXiv.

#	Top 10 Words [Num. Tokens] Recovered (NPMI_W)
	LDA ($K = 100$)
1	administer[2] west_germany[5] treat[1] poland[2] berlin[2] european[2] accompany[2] german[2] europe[1] germany[2] (0.065)
2	negate[2] mediation[2] settlement[1] cypriot[4] turkish[3] solution[1] party[1] turkey[2] cyprus[3] negotiation[3] (0.078)
3	impose[2] the_united_states[6] united[2] brussels[3] united_states[4] blockade[2] block[1] cuban[2] embargo[1] cuba[2] (0.082)
4	deserve[2] concern[1] solve[1] question[1] find[1] anxiety[2] situation[1] satisfactory[2] solution[1] problem[1] (0.089)
5	geographic[2] europe[1] geography[2] ukrainian[3] russian[2] georgian[3] geographical[2] russia[2] georgia[3] ukraine[2] (0.111)
6	island[1] barbados[3] devastating[3] disaster[2] catastrophic[4] caricom[3] hurricane[3] catastrophe[4] haiti[2] caribbean[3] (0.145)
7	rhodes[2] rhodesia[3] southern_africa[4] south_africa[4] zimbabwe[4] continent[1] africans[3] awe[2] africa[2] african[3] (0.148)
8	s_republic[4] republic_of_korea[7] the_democratic_republic_of_the_congo[15] the_democratic_republic_of_congo[13] south_korea[5] s_republic_of_korea[10] the_republic_of_korea[10] north_korea[5] korean[3] korea[3] (0.183)
9	costa_rica[3] el_salvador[5] guatemala[3] panama[2] central[1] central_america[4] latin_america[4] latin[1] american[2] america[2] (0.200)
10	argentina[2] taiwan[3] republic_of_china[6] the_people[3] the_chinese_people[6] chinese[2] s_republic[4] the_republic_of_china[9] s_republic_of_china[9] china[2] (0.239)
	DMR ($K = 100$)
1	principle[1] colonial[1] exercise[1] administer[2] determination[2] independence[1] territory[1] people[1] right[1] self-determination[5] (0.079)
2	geography[2] cia[2] slovakia[3] slovak[2] czechia[3] georgia[3] geographical[2] czech[2] czechoslovakia[5] czechoslovak[4] (0.096)
3	desert[1] devastating[3] upheaval[3] food[1] change[1] natural[1] climate[1] disaster[2] drought[2] catastrophe[4] (0.109)

4	universal[1] respect[1] freedom[1] protection[1] law[1] crime[1] humanity[2] rights[1] right[1] human[1] (0.143)
5	western_europe[4] co-operation[3] eastern_europe[4] the_european_community[7] the_european_union[7] japan[2] cooperation[2] european_union[4] european[2] europe[1] (0.154)
6	indian[2] asian[2] malaysian[3] philippines[4] indonesian[3] australia[2] malaysia[3] indonesia[3] india[2] asia[2] (0.180)
7	taiwan[3] the_dominican_republic[7] argentina[2] the_people[3] the_chinese_people[6] chinese[2] s_republic[4] the_republic_of_china[9] s_republic_of_china[9] china[2] (0.197)
8	the_middle_east[5] territory[1] palestinians[4] iraqi[3] iraq[3] arabian[2] palestine[3] israeli[3] palestinian[4] israel[2] (0.213)
9	fiji[2] netherlands[3] new_guinea[4] new_zealand[4] solomon[2] solomon_islands[5] papua[2] papua_new_guinea[7] pacific_islands[5] pacific[2] (0.219)
10	permanent[1] member[1] united_nations_security_council[11] resolution[1] council[1] the_united_nations_security_council[13] un_security_council[7] the_un_security_council[9] security_council[5] the_security_council[7] (0.289)

ProdLDA ($K = 100$)

1	internationalism[2] international[1] united[2] yearn[2] country[1] peace[1] world[1] united_nations[5] people[1] the_united_nations[7] (0.058)
2	the_west_bank[5] s_republic_of_china[9] the_democratic_republic_of_congo[13] the_people[3] the_democratic_republic_of_the_congo[15] the_republic_of_china[9] s_republic[4] the_republic_of_the_congo[11] the_world_bank[5] the_soviet_union[7] (0.062)
3	include[1] authentic[1] palestinians[4] people[1] sole[1] homeland[2] self[1] legitimate[2] palestine[3] palestinian[4] (0.068)
4	indians[2] asia[2] algeria[3] australia[2] pakistan[2] alia[2] armenia[3] indonesia[3] russia[2] india[2] (0.080)
5	partnership[2] senegal[3] dialog[1] dialogue[2] portugal[2] global[1] institutional[2] trilateral[4] bilateral[3] multilateral[4] (0.085)
6	kingdom[1] iraqi[3] syria[2] ian[2] iraq[3] yemen[2] saudi[2] sudan[2] iran[2] libyan[3] (0.121)
7	peace[1] people[1] international[1] country[1] united_nations[5] world[1] the_general_assembly[5] the_united_nations_general_assembly[11] the_un_general_assembly[7] the_united_nations[7] (0.134)
8	angola[2] the_central_african_republic[10] united_nations[5] people[1] south_africa[4] africa[2] the_general_assembly[5] the_united_nations_general_assembly[11] the_un_general_assembly[7] the_united_nations[7] (0.153)
9	19th[4] 6th[3] 26th[4] 20th[4] 12th[4] 27th[4] 7th[3] 38th[4] 9th[3] 30th[4] (0.162)
10	mutual[2] friendship[1] fruitful[2] co-operation[3] relation[1] co-existence[4] peace[1] peaceful[2] coexistence[3] cooperation[2] (0.165)

CTM ($K = 100$)

1	common[1] tie[1] region[1] grouping[1] economic[1] association[1] integration[1] co-operation[3] regional[1] cooperation[2] (0.075)
2	sanction[2] southern[1] apartheid[2] nelson[2] apart[1] zimbabwe[4] pretoria[2] regime[1] namibia[3] south[1] (0.075)
3	the_communist_party[7] the_republic_of_poland[9] the_federal_republic[8] west_germany[5] the_federal_republic_of_germany[14] ussr[2] beret[2] the_soviet_union[7] soviet_union[4] soviet[2] (0.084)
4	mass[1] extension[1] destruction[1] non-proliferation[6] treat[1] free[1] treaty[2] proliferation[4] nuclear[1] weapon[1] (0.097)

5	selfish[2] turkey[2] cypriots[4] territory[1] cyprus[3] sovereign[3] independence[1] territorial[1] integrity[1] sovereignty[5] (0.116)
6	secretary[1] delegation[2] the_united_nations_general_assembly[11] nobel[2] the_un_general_assembly[7] the_general_assembly[5] secretary-general[3] award[1] session[1] prize[1] (0.125)
7	papua_new_guinea[7] the_united_kingdom[7] papua[2] pacific[2] new_zealand[4] the_general_assembly[5] united_nations[5] the_un_general_assembly[7] the_united_nations_general_assembly[11] the_united_nations[7] (0.138)
8	the_far_east[5] the_security_council[7] the_people[3] the_un_security_council[9] the_united_nations[7] the_near_east[5] the_united_nations_general_assembly[11] the_middle_east[5] the_un_general_assembly[7] the_general_assembly[5] (0.139)
9	disease[1] vaccine[3] healthcare[2] covid[2] caribbean[3] hiv[2] medical[1] pandemic[2] health[1] virus[1] (0.165)
10	united_nations_security_council[11] the_un_security_council[9] the_united_nations_security_council[13] united_nations[5] international[1] the_people[3] the_united_nations[7] the_united_nations_general_assembly[11] the_un_general_assembly[7] the_general_assembly[5] (0.259)

Table 18: Supplementary Examples from UN.

#	Top 10 Words [Num. Tokens] Recovered (NPMI_W)
	LDA ($K = 100$)
1	guy[2] think[1] talk[1] thing[1] tell[1] want[1] know[1] like[1] good[1] friend[1] (0.122)
2	human[1] ribosome[3] egg[1] organism[2] organ[1] genetic[2] dome[2] dna[2] chromosome[3] genome[2] (0.122)
3	16th[4] 30th[4] 13th[4] 20th[4] 000th[5] 1000th[6] 100th[5] 12th[4] 10th[4] 11th[4] (0.124)
4	disease[1] psychic[2] medic[1] medication[2] diabetes[3] cause[1] psych[1] depression[2] symptom[2] disorder[2] (0.123)
5	sleeping[2] wake[2] ashe[2] awake[2] night[1] breath[1] horse[1] breathe[3] dream[1] sleep[1] (0.124)
6	counselor[3] hospital[1] patient[1] nurse[2] insist[2] health[1] medical[1] counsel[2] insurance[2] doctor[1] (0.134)
7	bia[2] hiv[2] antibacterial[4] bacteria[3] ini[2] antibiotic[4] infect[2] disease[1] virus[1] infection[2] (0.136)
8	bone[2] cancerous[2] cause[1] muscle[2] body[1] thyroid[2] cell[1] hone[2] hormone[3] cancer[1] (0.143)
9	cool[1] cold[1] ice[1] liquid[1] energy[1] pressure[1] air[1] heat[1] temperature[1] water[1] (0.187)
10	antivirus[3] computer[1] software[1] malware[2] program[1] download[1] free[1] spy[2] virus[1] spyware[3] (0.189)
	DMR ($K = 100$)
1	distance[1] point[1] height[1] area[1] circle[1] line[1] square[1] length[1] triangle[1] angle[1] (0.106)
2	constitution[1] election[1] congress[2] democrats[3] government[1] republic[1] bush[1] the_united_states[6] vote[1] president[1] (0.132)
3	tobacco[3] buddy[2] cancer[1] tote[2] quit[1] cigarette[3] smoker[2] smell[2] smoke[1] smoking[2] (0.149)
4	mexico[2] immigration[2] immigrant[3] mexicans[2] country[1] mexican[2] illegal[1] america[2] americans[2] american[2] (0.153)

5	absorb[2] particle[1] wavelength[3] radiation[1] frequency[1] blue[1] speed[1] color[1] energy[1] light[1] (0.157)
6	advertise[2] walmart[2] price[1] sale[1] product[1] item[1] advertising[2] store[1] advert[1] ebay[2] (0.158)
7	study[1] job[1] university[1] education[1] degree[1] teacher[1] class[1] student[1] college[1] school[1] (0.177)
8	number[1] server[1] password[1] e-mail[3] message[1] phone[1] account[1] send[1] address[1] email[1] (0.181)
9	acidic[2] health[1] nutrient[3] diet[2] ingest[2] intestine[3] digest[2] supplement[2] food[1] vitamin[2] (0.202)
10	therapy[2] antibiotic[4] disease[1] virus[1] infection[2] patient[1] cancer[1] treatment[1] surgery[2] doctor[1] (0.241)

ProdLDA ($K = 100$)

1	oil[1] bomb[1] invasion[1] japan[2] iraq[3] iraqi[3] afghan[2] ira[2] ian[2] iran[2] (0.050)
2	bet[1] meet[1] maybe[1] goodbye[2] wish[1] consider[1] how[1] choose[1] hope[1] luck[1] (0.051)
3	north_and[3] georgia[3] north_east[3] russia[2] west_virginia[5] south[1] asia[2] south_east[3] australia[2] india[2] (0.056)
4	bear[1] yes[1] birth[1] pregnancy[3] wife[1] married[1] baby[1] marry[1] young[1] old[1] (0.093)
5	grammar[1] translation[1] pronoun[2] phrase[1] voc[1] verb[1] spell[1] write[1] dictionary[1] word[1] (0.104)
6	player[1] spanish[2] british[2] english[2] tennis[1] rugby[1] sport[1] football[1] england[2] hockey[1] (0.102)
7	engineering[1] field[1] biology[2] master[1] degree[1] science[1] job[1] engineer[1] profession[1] assistant[1] (0.133)
8	amendment[3] lincoln[2] amend[2] supreme_court[4] declaration[1] the_united_states[6] the_united_states_supreme_court[11] the_white_house[5] the_supreme_court[6] rights[1] (0.123)
9	religion[1] believe[1] christianity[3] jesus[3] jew[2] jews[3] jus[2] christians[2] christ[1] christian[2] (0.160)
10	php[1] cities[1] pdf[1] website[1] asp[1] com[1] htm[2] edu[2] html[1] www[1] (0.276)

CTM ($K = 100$)

1	everyday[2] last[1] morning[1] minute[1] weekday[2] time[1] this[1] every[1] month[1] hour[1] (0.106)
2	alcoholic[3] help[1] tea[1] vitamin[2] medic[1] alcohol[2] blood[1] effect[1] drink[1] drug[1] (0.107)
3	stomach[3] allergy[2] digestion[2] aller[1] mouth[1] infection[2] eat[1] tract[1] fish[1] food[1] (0.107)
4	like[1] ask[1] interested[1] friendship[1] boyfriend[2] girlfriend[2] maybe[1] tell[1] talk[1] friend[1] (0.112)
5	money[1] stock[1] cheap[1] purchase[1] product[1] ebay[2] market[1] price[1] store[1] sell[1] (0.156)
6	azure[1] version[1] virus[1] firefox[1] program[1] software[1] windows[1] install[1] download[1] free[1] (0.126)
7	citizen[2] border[1] mexicans[2] country[1] illegal[1] immigrant[3] americans[2] mexican[2] immigration[2] american[2] (0.147)
8	democracy[2] president[1] democratic[3] democrats[3] election[1] vote[1] republicans[2] democrat[2] republic[1] bush[1] (0.207)

9	chinese[2] chocolate[3] soup[1] chicken[2] corn[1] egg[1] rice[1] fruit[1] bread[1] bean[1] (0.213)
10	medicine[1] disease[1] hospital[1] medic[1] pain[1] treatment[1] patient[1] surgery[2] medical[1] doctor[1] (0.222)

Table 19: Supplementary Examples from Yahoo.

#	Top 10 Words [Num. Tokens] Recovered (NPMI_W)
	LDA ($K = 100$)
1	receive[1] contact[1] message[1] tell[1] company[1] day[1] schedule[1] phone[1] email[1] appointment[1] (0.050)
2	east_coast[4] franco[2] the_west_coast[6] francisco[3] jang[2] the_east_coast[6] san_francisco_bay[7] los_angeles[4] san_diego[4] san_francisco[5] (0.065)
3	apology[2] ask[1] bad[1] service[1] tell[1] rude[2] unprofessional[3] customer[1] manager[1] horrible[2] (0.052)
4	dunkin[3] glaze[2] dunk[2] volcano[2] peanuts[3] dough[2] doughnut[3] peanut[3] butter[2] donut[2] (0.077)
5	potion[2] dion[2] crispy[3] cris[2] fry[2] dish[2] potato[2] chicken[2] flavor[2] sauce[2] (0.079)
6	lemon[2] sauce[2] leftover[4] lemonade[3] italian[2] spade[2] spaghetti[3] meatball[2] garlic[2] homemade[3] (0.110)
7	salami[2] chicken[2] lunch[2] turkey[2] cheese[2] lettuce[2] chad[2] bread[1] salad[2] sandwich[2] (0.184)
8	toast[2] ole[2] coffee[1] waffle[3] pancake[3] omelette[3] hash[1] bacon[2] egg[1] breakfast[1] (0.193)
9	11th[4] 5pm[3] 15pm[4] 30pm[4] 3pm[3] 2pm[3] 00pm[4] 10pm[4] 11pm[4] 1pm[3] (0.288)
10	vous[1] dans[1] qui[1] mais[1] pas[1] une[1] que[1] pour[1] les[1] est[1] (0.400)
	DMR ($K = 100$)
1	taj[2] taco[2] food[1] salsa[3] tequila[3] chip[1] tear[2] american[2] margarita[3] mexican[2] (0.049)
2	strip[1] the_east_end[5] the_east_coast[6] the_bay_area[5] san_francisco_bay[7] san_diego[4] vegas[2] the_las_vegas[6] san_francisco[5] las_vegas[4] (0.058)
3	concession[3] amc[2] screen[1] seater[2] film[1] theatre[1] ticket[1] popcorn[3] theater[2] movie[1] (0.086)
4	grand[1] adult[1] baby[1] mother[1] year[1] parent[1] family[1] daughter[1] child[1] kid[1] (0.086)
5	maple[2] cake[2] ole[2] breakfast[1] wake[2] velvet[2] waffle[3] syrup[2] cupcake[3] pancake[3] (0.109)
6	money[1] cost[1] extra[1] price[1] fee[1] tip[1] bill[1] pay[1] dollar[2] charge[1] (0.123)
7	plane[1] gate[1] bus[1] cab[1] car[1] driver[1] terminal[1] airline[2] flight[1] airport[2] (0.123)
8	roll[1] vietnamese[3] pork[2] vietnam[2] soup[1] ramen[2] broth[2] noodle[3] brood[2] bowl[2] (0.118)
9	brewery[3] ale[1] pretzel[2] bar[1] drinker[2] selection[1] brewer[2] brew[2] pub[1] beer[2] (0.206)
10	11th[4] 5pm[3] 15pm[4] 30pm[4] 3pm[3] 2pm[3] 00pm[4] 10pm[4] 11pm[4] 1pm[3] (0.288)
	ProdLDA ($K = 100$)
1	brisk[2] bone[2] lobster[3] rib[1] lob[2] link[1] bite[2] pull[1] bun[2] bbq[3] (0.015)

2	ric[1] thin[1] crisp[2] crispy[3] garlic[2] cris[2] pepper[2] mari[1] crust[2] sauce[2] (0.021)
3	stomach[3] stair[2] discomfort[3] filthy[2] smoker[2] dirty[1] bathroom[2] odell[2] smell[2] restroom[2] (0.031)
4	pittsburgh[3] rave[2] glove[2] prim[1] san_diego[4] los_angeles[4] crave[2] san_francisco[5] san_francisco_bay[7] las_vegas[4] (0.032)
5	pizzeria[4] pineapple[3] pine[2] dom[1] slice[1] pork[2] papa[1] pump[2] phoenix[3] pizza[2] (0.039)
6	menu[1] waitress[2] order[1] food[1] fountain[3] restaurant[1] fryer[3] flaw[2] french[2] fry[2] (0.045)
7	organic[2] product[1] market[1] checkout[1] vitamin[2] supply[1] grocery[3] organ[1] produce[1] bulk[1] (0.053)
8	weekday[2] cake[2] little[1] try[1] birthday[2] think[1] good[1] pretty[1] thing[1] like[1] (0.058)
9	driver[1] lot[1] street[1] route[1] bus[1] montreal[2] parking[2] rail[1] car[1] downtown[3] (0.126)
10	theatre[1] singing[1] magic[1] musical[1] performance[1] audience[1] production[1] actor[1] film[1] comedy[1] (0.142)
	CTM ($K = 100$)
1	organ[1] selection[1] supply[1] buy[1] find[1] item[1] produce[1] store[1] price[1] bulk[1] (0.042)
2	mall[2] thing[1] plane[1] airy[2] shy[2] think[1] way[1] look[1] people[1] like[1] (0.045)
3	know[1] order[1] food[1] want[1] fry[2] come[1] look[1] place[1] like[1] eat[1] (0.045)
4	navigate[1] sky[1] convenience[1] court[1] rail[1] traffic[1] convenient[1] location[1] lot[1] parking[2] (0.054)
5	fee[1] copy[1] account[1] credit[1] dispute[1] bank[1] bill[1] card[1] charge[1] statement[1] (0.059)
6	mus[1] gyro[2] salami[2] loma[2] lunch[2] sandwich[2] lettuce[2] soup[1] salmon[2] salad[2] (0.069)
7	hour[1] 11pm[4] 10pm[4] 00pm[4] 1pm[3] come[1] ask[1] time[1] order[1] minute[1] (0.093)
8	margarita[3] tab[1] friend[1] cocktail[3] cock[2] martini[2] bender[2] bartender[3] bart[2] drink[1] (0.079)
9	11pm[4] mondays[2] weekend[2] weekday[2] late[1] lend[2] monday[2] sundays[3] sunday[3] lunch[2] (0.157)
10	omelette[3] pancake[3] toast[2] scramble[3] brown[1] benedict[3] hash[1] eggs[1] bacon[2] breakfast[1] (0.187)

Table 20: Supplementary Examples from Yelp.

E User Study Information

E.1 Instructions

This section is a short primer of the tasks that will be presented in this study.

For each question, you will be presented with two groups of 10 alphabetically-sorted words. Between these Word Groups, the common words in both word groups are bolded. In this study, we wish to obtain your opinion on how coherent the Word Groups are. After examining the Word Groups, we will ask three simple sub-questions.

Usually, we consider a Word Group coherent when we can easily see how the different words relate to each another, and the context where they can be used together.

`Example A: apple berry durian grapes jackfruit lemon lime mango orange pineapple`
Example A is widely considered to be coherent as most people can easily relate to the theme of "fruits".

`Example B: apple berry citric lemon lime mango orange pineapple sour zest`
Example B is an example where there could be mixed opinions. Some may consider it as coherent as they can see how the different words are related to each other.

`Example C: apple black citric economics market orange quantum physics sour zest`
Example C is an example where it is very difficult to see how some words relate to other words, and thus considered to be incoherent.

For the last sub-question, both Word Groups can be coherent, but you may consider one to be more coherent than the other. The reverse may also apply, where you consider both Word Groups as incoherent, but perhaps one has more relatable words than the other, and you consider it as more coherent.

Some multi-word nouns are combined with underlines, such as "security_council". Some short words may also be lower-cased abbreviations, such as "irs".

E.2 Example Question

You are given two word groups:

`Word Group A:` **`apple berry`** `durian grapes jackfruit` **`lemon lime`** `mango orange pineapple`

`Word Group B:` **`apple berry`** `cart carrot economics fruit` **`lemon lime`** `market truck`

Please read every word in each group and answer the following sub-questions.

1. Is Word Group A coherent to you?
    - Yes
    - No
2. Is Word Group B coherent to you?
    - Yes
    - No
3. Comparing the two word groups' coherence, which statement best describe your opinion?
    - Both Word Groups are similar in coherence quality
    - Word Group A is more coherent than Word Group B
    - Word Group B is more coherent than Word Group A

E.3 Questions

Q	Topic representation pair (top r from LDA$_b$, bottom r from LDA$_o$)	NPMI$_W$	Votes
1	**court international** jurisdiction **law statute** the_international_atomic_energy_agency **the_international_court_of_justice** the_international_criminal_court the_international_monetary_fund the_world_bank	0.120	6
	court crime criminal **international** justice **law** legal rule **statute the_international_court_of_justice**	0.166	10
3	**america american brazil** central central_america guatemala **latin latin_america** mexico nicaragua	0.177	10
	america american bolivia **brazil** ecuador **latin latin_america** panama peru venezuela	0.238	10
4	atrocity **crime criminal** ethnic **genocide human** humanity propaganda rwanda uganda	0.135	8
	court **crime criminal genocide human** international law right statute violation	0.153	9
5	band concert event game **movie** stadium **theater theatre ticket** venue	0.083	9
	amc concession film **movie** popcorn screen seat **theater theatre ticket**	0.107	9
8	acknowledge **agency assistance** pledge **programme** special taiwan the_united_nations united **united_nations**	0.078	7
	agency aid **assistance** country development fund **programme** resource technical **united_nations**	0.111	9
10	allergy bad **cooked** eat **food fry** greasy greed microwave **taste**	0.071	7
	chicken **cooked** dry flavor **food fry** order sauce **taste** wing	0.154	10
11	air density **energy** fuel **heat** liquid pressure **temperature** volume **water**	0.157	8
	atom carbon electron **energy** gas **heat** mole molecule **temperature water**	0.217	9
13	arabs ish **islam** islamic israel jew jews **muslim muslims religion**	0.198	10
	allah belief christian church god **islam muslim muslims religion** religious	0.225	10
14	area **bar** beating **game** outdoor outside patio seating **sport watch**	0.038	7
	bar beer football **game** play screen **sport** stadium team **watch**	0.073	10
15	**game** goal **hit pitch pitcher player** shot team throw tooth	0.111	6
	ball baseball bat **game hit pitch pitcher** play **player** run	0.227	9
17	eritrea ethiopia moroccan **morocco** refer **referendum sahara** saharan satisfactory **western**	0.096	7
	african indonesia **morocco** netherlands oau **referendum sahara** self-determination spain **western**	0.119	5
18	buy **clothe** clothing find item **sale** selection **shirt shoe store**	0.104	10
	bike **clothe** dress pair **sale shirt shoe** shop **store** wear	0.131	9
19	**area building** car drive game **lot parking** parkway stadium **street**	0.063	9
	area building center locate location **lot** mall park **parking street**	0.152	10
20	asia asian **india indian indonesia** indonesian **malaysia** malaysian malian thailand	0.136	9
	india indian indonesia jammu kashmir **malaysia** nepal netherlands pakistan philippines	0.173	6
21	**air** area foot inch **mile** rain snow **weather wind** winter	0.117	7
	air cloud hour **mile** ocean sea storm water **weather wind**	0.12	8

22	diet **drink** fat **food fruit** health meal **sugar water** weight	0.152	10
	drink eat **food fruit** juice meat milk **sugar** tea **water**	0.242	10
23	**bacon benedict** biscuit **breakfast egg hash** mme omelette potato sausage	0.136	9
	bacon benedict breakfast brunch coffee **egg hash** pancake toast waffle	0.23	10
26	colombia **combat crime drug illicit** terror trafficker **trafficking** trans transit	0.102	9
	arm **combat crime drug** fight illegal **illicit** narcotic terrorism **trafficking**	0.193	10
27	**anxiety** depress depressed **depression disorder** medic **mental** psych psychic symptom	0.121	10
	anxiety brain **depression disorder** fear feel help **mental** physical stress	0.157	10
28	**democracy** democrat **democratic** democratization freedom institution **political society** systematic venezuela	0.091	9
	constitution **democracy democratic** election government law national **political** process **society**	0.105	10
30	**band** course guitar gun loud **music** play singe **song stage**	0.07	7
	act audience **band** cirque **music** performance performer **song stage** ticket	0.081	9
31	arabic **gulf** ian **iran** iranian **iraq iraqi** iraqis **kuwait** saudi	0.172	9
	arab **gulf iran iraq iraqi kuwait** resolution sanction security_council yemen	0.233	9
33	**about** ago day **million** month **more one than** time two	0.158	8
	about average less mile **million more one** percent **than** year	0.195	7
34	**doctor drug** medic **medication medicine** prescribe prescription survey therapist therapy	0.167	10
	doctor drug health hospital medical **medication medicine** patient surgery treatment	0.239	10
38	**album band** download **guitar** lime mp3 **music song** sound wire	0.109	6
	album band dance **guitar** lyric **music** play rock sing **song**	0.197	10
40	**displace** flee human **humanitarian** million **person** refuge **refugee return** thousand	0.086	9
	bosnia **displace** herzegovina home **humanitarian** kosovo **person refugee return** serbia	0.129	7
41	**animal** buy grocery **grooming hot** market **pet** produce **puppy** store	0.07	6
	animal cat dog groom **grooming hot pet** pup **puppy** vet	0.177	10
43	**bread** chad cheese **lunch** meat **order salad** salami **sandwich** turkey	0.137	9
	bread chicken food good **lunch order** place **salad sandwich** soup	0.138	10
44	**animal bird** dinosaur dog **fish** human mammal mph **specie** wild	0.145	8
	animal bird fish ocean plant river sea **specie** tree water	0.153	9
45	account bank **card charge credit debit** fee **gift** receipt tell	0.121	9
	card cash **charge** coupon **credit debit gift** pay store use	0.126	10

Table 21: In this table, we list the 30 question presented with pairs, containing common words (bolded), randomly drawn from LDA_o and LDA_b with equal numbers from UN, Yahoo, and Yelp corpus. The order of pairs within the question shown are random in the user study. 'Votes' denotes the number of study participants (out of 10) that think presented topic is coherent. Questions with trivial dummy topics omitted.