

Quantifying Morphology in Latent Dirichlet Analysis Topic Models

Virginia Partridge

University of Massachusetts Amherst

vcpartridge@umass.edu

Abstract

Latent Dirichlet allocation (LDA) is a popular approach for probabilistic topic modeling, frequently applied in many disciplines for exploring themes and trends in large document collections. Because LDA assumes a bag-of-words approach, stemming or lemmatization are common pre-processing steps in preparing corpora for topic modeling, despite little evidence that these improve topic quality for English. Recent work has suggested post-processing topics so that they are more interpretable by end users may be a better approach.

There is more motivation to apply stemming or lemmatization for topic modeling on languages with complex inflectional morphology, due to concerns that rare word forms will either be randomly assigned to topics or cause lexemes that are important to a topic's interpretability to be obscured. We present several metrics designed to quantify the morphological and lexical complexity of topics learned by LDA, with a focus on identifying topics that may benefit from post-processing. We then use these metrics to analyze LDA topic models trained with Gibbs sampling on Russian and German corpora, comparing the effects of different stemmers and lemmatizers.

1 Introduction

Latent Dirichlet allocation (LDA) is a widely adopted approach for unsupervised topic modeling and has been used across disciplines for exploring themes and trends in large document collections. LDA has been applied to explore the ever-growing variety of text from online platforms and to analyze language changes in academic fields over time (Koltsova and Koltsov, 2013; McFarland et al., 2013; Vogel and Jurafsky, 2012; Mitrofanova, 2015). Assuming a bag-of-words approach, LDA produces latent topics as multinomial distributions over words and each topic is viewed as being generated by a mixture of topics (Blei et al., 2003; Steyvers and Griffiths, 2007).

However, what happens when words in this bag-of-words approach are themselves complex? Stemming and lemmatization treatments are typical text preprocessing steps for topic modeling, even for English, which has relatively little inflectional morphology, but there is a lack of empirical evidence that these treatments improve the models from the perspective of human interpretability or quantitative measures of topic quality (Schofield and Mimno, 2016). To understand the effects of these treatments on languages with more inflectional morphology, we train topic models on the German TIGER corpus¹ (Brants et al., 2004) and the Russian National corpus² (RNC) (Apresjan et al., 2006). These corpora have high quality morphological and syntactic annotations, which allow for analysis based on gold standard morphological analyses and lemmatization.

Following Schofield and Mimno's work, we use Variation of information (VOI) to show how stemming and lemmatization change topic assignments under various pre-processing treatments. The stability of topic assignments under each treatment show how pre-processing affects experimental reproducibility. Additionally, we adapt measures of morphological complexity to analyzing LDA topics produced by Gibbs sampling in order to quantify the ways in which inflectional morphology influence topic models and identify topics where morphology most strongly complicates interpretability. So long as reliable morphological analyses are available, these methods can be applied cross-linguistically, which we demonstrate using topic models for TIGER and RNC. We also find evidence that stemming in pre-processing hurts topic stability, but

¹<https://www.ims.uni-stuttgart.de/en/research/resources/corpora/tiger/>

²<https://ruscorpora.ru/old/en/corpora-morph.html>

2 Related Work

The proposal for applying stemming in post-processing comes from work comparing the effects of various stemming approaches on English (Schofield and Mimno, 2016). After comparing the relative strengths, qualitative and quantitative impacts of rule-based and context-based stemmers for English, it was concluded that stemming in pre-processing does not empirically improve LDA topic models and may hurt topic stability. Post-processing may still be valuable from the perspective of topic interpretability, avoiding repeating different surface forms of the same lexeme in topics’ key word lists and presenting users with concise results.

Probabilistic topic modeling has been applied on Russian text data from academic fields, social media, and Wikipedia articles (Mitrofanova, 2015; Koltsova and Koltsov, 2013; May et al., 2016). Prior to the work on Wikipedia, little attention was given to the role of lemmatization on topic modeling in Russian, and corpora were lemmatized by default. In studying Russian Wikipedia, May et al. (2016) address the impact of lemmatization on topic interpretability via a word intrusion evaluation task, finding that lemmatization may be beneficial. However, they also suggest measuring the effects of lemmatization and do not rule out that lemmatizing in post-processing would also be effective.

Although we found little prior work on the effects of stemming on topic modeling for German, Rieger et al. present methods for improving the stability of LDA and detail results of experiments on a German newspaper corpus (Rieger et al., 2020). Their focus is on increasing the reproducibility of LDA topic models by choosing the initial token allocations for Gibbs sampling after comparing multiple LDA models and they limit pre-processing to removal of stopwords and punctuation. Schofield and Mimno also explicitly measured stability of token allocations by using VOI to compare the stemming treatments, finding that certain types of stemming could increase the impact of random initialization, hurting reproducibility.

3 Background

3.1 Latent Dirichlet Analysis

LDA uses the observed frequencies of vocabulary terms within documents to infer the *latent*, or hid-

den, distributions of topics over words and topic assignments for each document. Once a number of topics T is selected, the multinomial distributions ϕ_1, \dots, ϕ_T define the distribution of each topic t over the vocabulary terms. Each ϕ_t is drawn from with a Dirichlet prior with concentration parameter β . Each document d also has a multinomial distribution θ_d over the terms in the vocabulary, also drawn from a Dirichlet prior with concentration parameter α . Viewing LDA as a generative process with a joint distribution of the observed and latent variables, find the ϕ_t and θ_d that maximize the likelihood of the corpus if you were to assign tokens to documents using the marginal distributions over topic assignments for the terms in each document. Gibbs Sampling allows estimation of the posterior for the joint topic distribution conditioned on the observed term frequencies by directly assigning topics to each token in the corpus, iteratively sampling topics and updating topic assignments (Stein and Griffiths, 2007; Blei et al., 2003; Schofield and Mimno, 2016).

Following Wallach et. al (2002), we will use a symmetric prior for β and an asymmetric prior for α with the MALLET’s Gibbs Sampling implementation to train topic models (Wallach et al., 2009; McCallum, 2002). These parameters are optimized every 20 iterations after the first 50, the burn-in period. The Gibbs sampling implementation in MALLET allows us to directly inspect the topic assignments at the level of each token in a document.

3.2 Framework for Morphological Complexity

We will first clarify terms for discussing Russian’s morphological paradigms, following frameworks for quantifying morphological complexity used in linguistics and computational linguistics (Baerman et al., 2015b; Ackerman and Malouf, 2013; Cotterell et al., 2019). We draw a distinction between *derivational* morphology, the process by which new words are formed through changing meaning or part-of-speech, and *inflectional* morphology, which can be simplistically understood as verb paradigms to capture subject-verb agreement or noun declensions for case and grammatical gender. For our purposes here, we are primarily interested in the equivalence classes formed by normalizing inflectional morphology, to use an English example, conflating “respond” and “responds”, rather than “respond” and “responsiveness”, although aggressive

stemming methods will do both types of conflation.

In the word-based morphology framework, inflection is captured by triples consisting of the surface form (also called wordform) w , a lexeme signifying the meaning and a slot σ , which can be understood as a set of “atomic” units of morphological meaning, also called inflectional features (Aronoff, 1976; Sylak-Glassman et al., 2015; Cotterell et al., 2019). A lemma is the surface form used to look up the lexeme in a dictionary, such as the infinitive verb form. Measurements of the size of a lexeme’s morphological paradigm capture *enumerative complexity*, the number of distinct surface forms for a particular part-of-speech (Ackerman and Malouf, 2013). A lexeme’s mapping between slots and the surface forms is not always straightforward outside the context of a sentence, as multiple slots may be realized with a single surface form. This type of morphological complexity is called *syncretism* and is common in Russian noun and adjective declensions (Baerman et al., 2015a; Milizia, 2015). German also demonstrates syncretism in adjective agreement for noun case and gender (Crysmann, 2005).

4 Corpora

5 Methods

5.1 Stemmers and Lemmatization Treatments

Following Schofield and Mimno (2016), we distinguish between rule-based stemmers, which are deterministic, but only remove endings and do not map to lemmas, and context-based lemmatizers, which rely on a dictionary of word forms paired with outputs from a part-of-speech tagger to produce lemmas (Schofield and Mimno, 2016; Sharoff and Nivre, 2011). Rule-based methods make no distinction between inflectional and derivational morphological processes, leading to word types, conflation classes of terms, whose original surface forms may cover several lemmas.

Oracle:

Truncation: This simple baseline method trims surface forms to the first n characters (Schofield and Mimno, 2016). We truncate with $n = 5$.

Snowball Stemmer: This stemmer was introduced as a rigorous framework for implementing stemming algorithms for a variety of languages. We utilize the NLTK implementation³ with the

³<https://www.nltk.org/api/nltk.stem.html>

original rules for Russian⁴ and German⁵ (Porter, 2001).

Mystem: This Yandex-owned tool is the most popular Russian lemmatizer and can be used without part-of-speech tags. Pairing a finite state machine algorithm for stemming with the Zalizniak grammatical dictionary for morphological tags, this system outputs a list of possible lemmas and slots for a given token input. The system also produces probabilities for each lemma and slot based on word frequency statistics, although the source corpus for these probabilities is not clear (Segalovich, 2003). This is not truly a context-based lemmatizer, as it does not use part-of-speech tags to disambiguate between lemmas or to assign a single slot to a syncretic surface form, but the word frequencies do represent some kind of contextual prior. We use the python wrapper for Mystem, pymystem3⁶. Notably, Mystem is as fast as the Snowball stemmer, while producing a normalized lemma form that is more interpretable for users.

SpaCy: TODO

Stanza: This toolkit implements full neural pipelines for processing raw text, including tagging morphological features using bidirectional long short-term memory networks and lemmatizing an ensemble of dictionary based and seq2seq methods (Qi et al., 2020). Because each step of the pipeline depends on the output of the previous step, in order to use Stanza for morphological tagging and lemmatization, we also use its tokenization and sentence-splitting. We use the Stanza model trained on the SynTagRus treebank⁷. Unlike Mystem, Stanza always produces a single lemma and morphological slot, the disambiguation step is included within the model.

5.2 Evaluation metrics

When it comes to topic modeling on Russian, we would like to quantify the trade-offs between topic interpretability and loss of information that is linked to a surface form’s morphology. We can apply Mystem or Stanza to retrieve the most likely morphological analysis for a surface form w in the vocabulary V to find a lemma ℓ_w and slot σ_w . Us-

⁴<http://snowball.tartarus.org/algorithms/russian/stemmer.html>

⁵<http://snowball.tartarus.org/algorithms/german/stemmer.html>

⁶pythonhosted.org/pymystem3/pymystem3.html

⁷https://universaldependencies.org/treebanks/ru_syntagrus/index.html

ing the token-level topic assignments from Gibbs Sampling as our surface form w , we follow Thompson and Mimno (2018) in viewing single topic assignments for each surface form as a data table with columns: surface form w , topic assignment k , slot σ , lemma ℓ . For a given topic k , we obtain the joint count of the slots for the topic $N(\sigma, k)$, the counts of the lemmas for a topic $N(\ell, k)$ and the marginal count variable for a topic $N(k)$. Also note that $\operatorname{argmax}_{w \in V} N(w, k)$ denotes the top key words or surface forms for the topic.

5.2.1 Entropy-based morphology measurements

Morphological slot entropy: The goal of this metric is to measure the concentration of slots within a given topic, a proxy for the enumerative complexity of the topic. Does a topic have a concentration of only a few morphological features or does it have a wide spread of the language’s inventory of features? This metric is similar to Author Entropy discussed in Thompson and Mimno (2018), where the morphology of the language is the metadata we are attempting to capture, rather than the author of a document (Thompson and Mimno, 2018). Topics that have low slot entropy would contain wordforms with the same grammatical features, for example different verbs conjugated in the first-person singular form or nominative case masculine nouns.

$$\begin{aligned} H(\sigma|k) &= \sum_{\sigma} P(\sigma|k) \log_2 P(\sigma|k) \\ &= \sum_{\sigma} \frac{N(\sigma, k)}{N(k)} \log_2 \frac{N(\sigma, k)}{N(k)} \end{aligned} \quad (1)$$

Lemma entropy: Similarly, we want to know when a topic is dominated by a single lexeme, containing many grammatical forms of a single lexeme, but few other lexemes. For example, a topic may have many counts of different surface forms for each declension of a particular noun, its nominative, accusative, dative, etc... forms or even high counts for a single surface form, but relatively low counts of surface forms for any other lemma. Topics with very low lemma entropy may not be particularly useful to end users, as they reflect lexical and grammatical information known to every speaker of the language, but may not provide specific information about the corpus, other than the

presence of a particular lexeme.

$$\begin{aligned} H(\ell|k) &= \sum_{\ell} P(\ell|k) \log_2 P(\ell|k) \\ &= \sum_{\ell} \frac{N(\ell, k)}{N(k)} \log_2 \frac{N(\ell, k)}{N(k)} \end{aligned} \quad (2)$$

5.2.2 Counting-based morphology measurements

In practice, topics are often identified by keywords, the most frequently allocated terms to a topic. However, without pre-processing it’s possible that the set of top n keywords consists of many surface forms of the same lexeme, obscuring forms of other lexemes that could be useful to identifying the topic. Similarly, it’s possible that a lexeme’s allocations to a topic are spread across many word forms, such that no forms of the lexeme appear in the keywords for the topic, even though this may be the most frequent lemma allocated for the topic. Both of these problems occurring simultaneously for many topics would suggest a need for post-processing treatment.

Lemmas expressed by top n key terms: This set is targeted at understanding how concise the presentation of a topic’s key terms is to a user. When its size is close to n , each key term presented to the user represents a unique lexeme or multiple lexemes in cases of lexical ambiguity. If the set’s size is closer to 1, different forms of the same lexeme are repeated in the keywords.

$$K_{\ell}(k) = \{\ell_w | w \in \{n \text{ largest } N(w, k)\}\} \quad (3)$$

Top n lemmas: A topic’s most frequent lexemes may not always overlap with the lexemes of its most frequent surface forms. Comparing the differences between the most a topic’s most frequent lexemes, $L(k)$, defined below, and $K_{\ell}(k)$, will reveal topics where morphology impacts interpretability.

$$L(k) = \{\ell | \ell \in \{n \text{ largest } N(\ell, k)\}\} \quad (4)$$

Exclusivity: Exclusivity quantifies the relative uniqueness of the top keywords in a topic. It is high when the terms most frequently generated by a topic are rarely generated by other topics in the model (Bischof and Airolidi, 2012). This metric can also be modified to quantify the relative uniqueness of lemmas to a topic, which we call *lemma exclusivity*.

5.2.3 Strength of treatment measurements

These measurements quantify the aggressiveness of stemming or lemmatization.

Type-token ratio: Following Schofield and Mimno (2016), this corpus-level metric measures a stemmer or lemmatizer’s conflation strength. It is found by taking the ratio of the number of word-type equivalence classes produced by the treatment (the post-treatment vocabulary size $|V|$) to the token counts for the corpus (Schofield and Mimno, 2016).

Character-token ratio: This metric, also from Schofield and Mimno (2016), measures the aggressiveness of stemmers in trimming surface forms to a root form. It measures the average length of the tokens in the corpus after the stemming treatment. Because lemmatizers map surface forms to a normalized lemma instead, this metric isn’t as meaningful for lemmatization.

5.2.4 Variation of Information

6 Results

6.1 Counting top lexemes

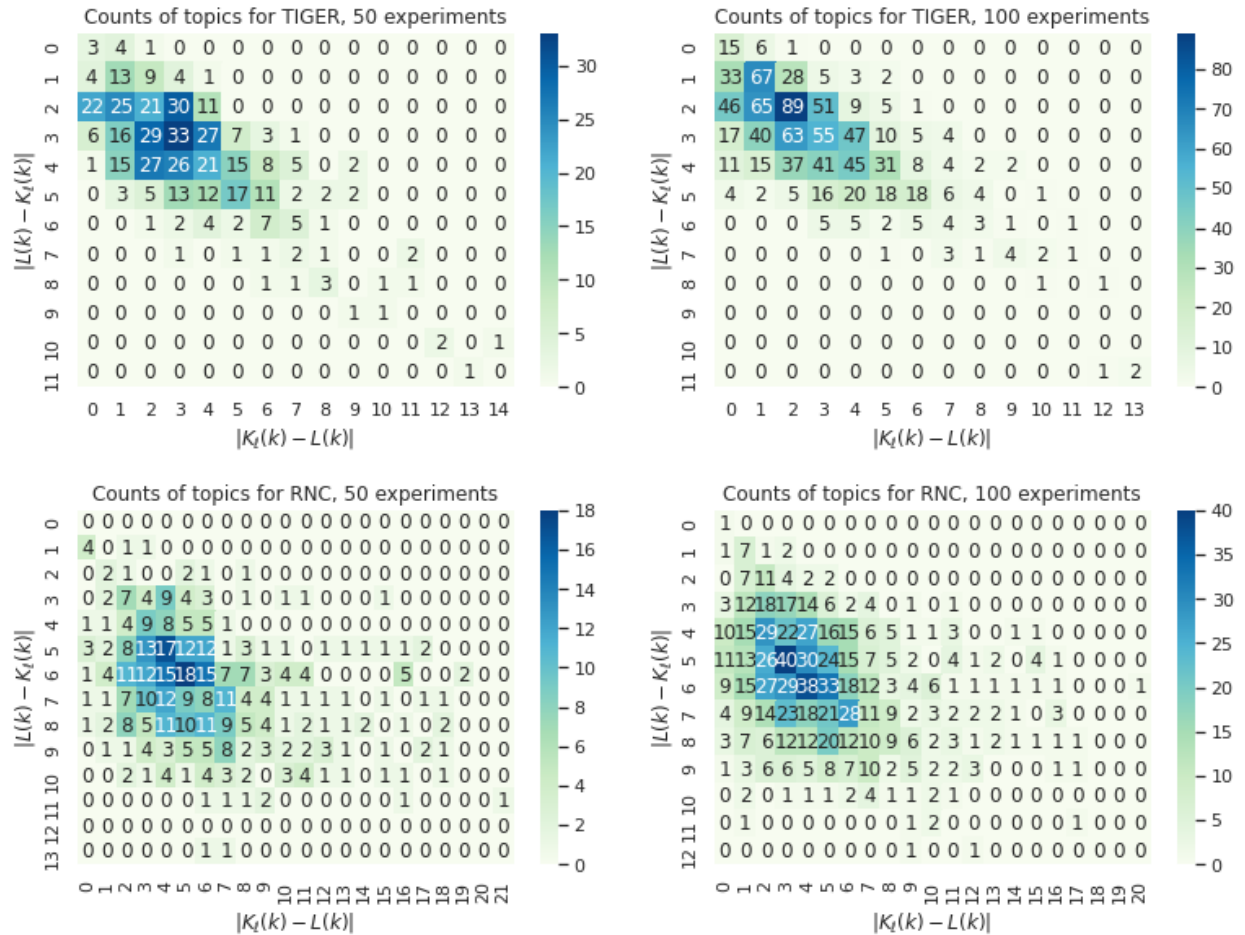
6.2 VOI and topic stability

7 Future Work

References

- Farrell Ackerman and Robert Malouf. 2013. Morphological organization: The low conditional entropy conjecture. *Language*, 89:429 – 464.
- Jurij D. Apresjan, Igor Boguslavsky, Boris Iomdin, Leonid L. Iomdin, Andrei Sannikov, and Victor G. Sizov. 2006. A syntactically and semantically tagged corpus of Russian: State of the art and prospects. In *LREC*.
- M. Aronoff. 1976. *Word Formation in Generative Grammar*. Linguistic inquiry monographs. MIT Press.
- Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015a. *Understanding and Measuring Morphological Complexity*. Oxford University Press, USA.
- Matthew Baerman, Dunstan Brown, and Greville G Corbett. 2015b. Understanding and measuring morphological complexity: An introduction. In Matthew Baerman, Dunstan Brown, and Greville G Corbett, editors, *Understanding and Measuring Morphological Complexity*, chapter 1. Oxford University Press, USA.
- Jonathan M. Bischof and Edoardo M. Airolidi. 2012. Summarizing topical content with word frequency and exclusivity. In *Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML’12*, page 9–16, Madison, WI, USA. Omnipress.
- David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.
- Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. Tiger: Linguistic interpretation of a german corpus. *Research on Language and Computation*, 2:597–620.
- Ryan Cotterell, Christo Kirov, Mans Hulden, and Jason Eisner. 2019. On the complexity and typology of inflectional morphological systems. *Transactions of the Association for Computational Linguistics*, 7:327–342.
- Berthold Crysman. 2005. Syncretism in german: A unified approach to underspecification, indeterminacy, and likeness of case. *Proceedings of the International Conference on Head-Driven Phrase Structure Grammar*.
- Olessia Koltsova and Sergei Koltsov. 2013. Mapping the public agenda with topic modeling: The case of the Russian livejournal. *Policy & Internet*, 5.
- Chandler May, Ryan Cotterell, and Benjamin Van Durme. 2016. An analysis of lemmatization on topic models of morphologically rich language.
- Andrew Kachites McCallum. 2002. *Mallet: A machine learning for language toolkit*.
- Daniel A McFarland, Daniel Ramage, Jason Chuang, Jeffrey Heer, Christopher D Manning, and Daniel Jurafsky. 2013. Differentiating language usage through topic models. *Poetics*, 41(6):607–625.
- Paolo Milizia. 2015. Patterns of syncretism and paradigm complexity: The case of old and middle indic declension. In Matthew Baerman, Dunstan Brown, and Greville G Corbett, editors, *Understanding and Measuring Morphological Complexity*, chapter 8. Oxford University Press, USA.
- Olga Mitrofanova. 2015. Probabilistic topic modeling of the Russian text corpus on musicology. In *International Workshop on Language, Music, and Computing*, pages 69–76. Springer.
- Martin F. Porter. 2001. *Snowball: A language for stemming algorithms*.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. *Stanza: A Python natural language processing toolkit for many human languages*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- J. Rieger, Jörg Rahnenführer, and Carsten Jentsch. 2020. Improving latent dirichlet allocation: On reliability of the novel method ldaprototype. *Natural Language Processing and Information Systems*, 12089:118 – 125.

Figure 1: Comparison of the top 20 lemmas for each topic and the lemmas covered by the topic’s top 20 key terms, over 10 experiments on the untreated corpus. The cells show the number of topics. Values in the upper left indicate topics with high overlap in lemmas of the top terms and the topic’s most frequent lemmas. Values in the lower left are topics with Values in the lower right indicate large mismatches between those sets, a challenge for topic interpretability.



Alexandra Schofield and David Mimno. 2016. [Comparing apples to Apple: The effects of stemmers on topic models](#). *Transactions of the Association for Computational Linguistics*, 4:287–300.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In *MLMTA*.

Serge Sharoff and Joakim Nivre. 2011. [The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge](#). pages 657–670, Moscow, Russia. Dialogue: Computational Linguistics and Intellectual Technologies.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

John Sylak-Glassman, Christo Kirov, David Yarowsky, and Roger Que. 2015. [A language-independent feature schema for inflectional morphology](#). In *Proceedings of the 53rd Annual Meeting of the Association*

for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 674–680, Beijing, China. Association for Computational Linguistics.

Laure Thompson and David Mimno. 2018. [Authorless topic models: Biasing models away from known structure](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Adam Vogel and Dan Jurafsky. 2012. [He said, she said: Gender in the ACL Anthology](#). In *Proceedings of the ACL-2012 Special Workshop on Rediscovering 50 Years of Discoveries*, pages 33–41, Jeju Island, Korea. Association for Computational Linguistics.

Hanna M Wallach, David M Mimno, and Andrew McCallum. 2009. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, pages 1973–1981.

Figure 2: Variation of information between pre-processing treatments averaged over pairwise comparison of 10 experiments for each treatment. ‘Raw’ indicates that no stemming or lemmatization pre-processing was performed.

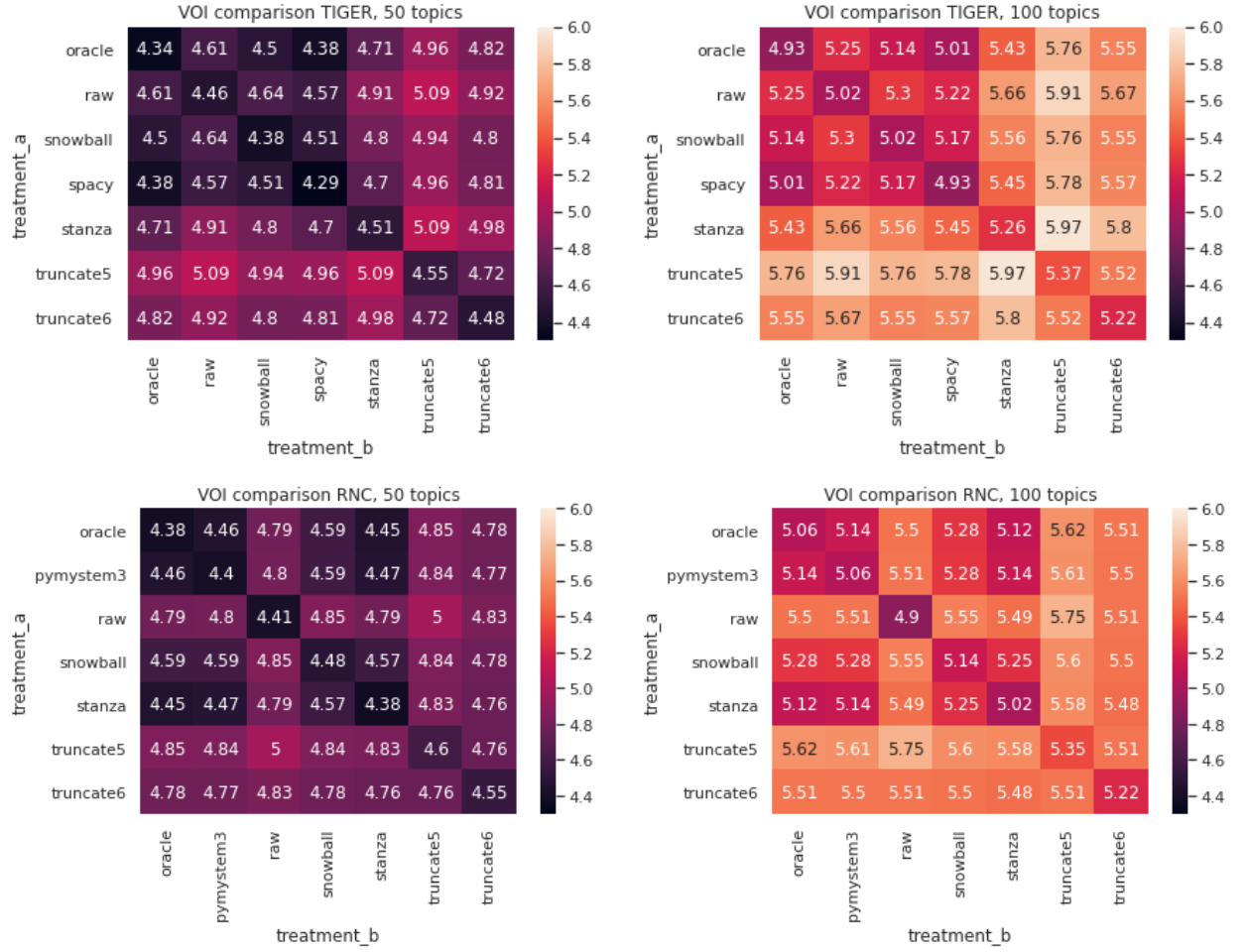


Table 1: These are sample topics from 50 topic models trained on the untreated corpora demonstrating how the difference between $K_\ell(k)$ and $L(k)$ can be used to identify topics as candidates for post-stemming.

| $K_\ell(k)$ | $L(k)$ | $ L(k) - K_\ell(k) $ | $ K_\ell(k) - L(k) $ | Comment |
|-------------|--------|----------------------|----------------------|---------|
|-------------|--------|----------------------|----------------------|---------|