Exploring Topic Coherence over many models and many topics

Keith Stevens[1,2] **Philip Kegelmeyer**[3] **David Andrzejewski**[2] **David Buttler**[2]
[1]University of California Los Angeles; Los Angeles , California, USA
[2]Lawrence Livermore National Lab; Livermore, California, USA
[3]Sandia National Lab; Livermore, California, USA
{stevens35,andrzejewski1,buttler1}@llnl.gov
wpk@sandia.gov

Abstract

We apply two new automated semantic evaluations to three distinct latent topic models. Both metrics have been shown to align with human evaluations and provide a balance between internal measures of information gain and comparisons to human ratings of coherent topics. We improve upon the measures by introducing new aggregate measures that allows for comparing complete topic models. We further compare the automated measures to other metrics for topic models, comparison to manually crafted semantic tests and document classification. Our experiments reveal that LDA and LSA each have different strengths; LDA best learns descriptive topics while LSA is best at creating a compact semantic representation of documents and words in a corpus.

1 Introduction

Topic models learn bags of related words from large corpora without any supervision. Based on the words used within a document, they mine topic level relations by assuming that a single document covers a small set of concise topics. Once learned, these topics often correlate well with human concepts. For example, one model might produce topics that cover ideas such as government affairs, sports, and movies. With these unsupervised methods, we can extract useful semantic information in a variety of tasks that depend on identifying unique topics or concepts, such as distributional semantics (Jurgens and Stevens, 2010), word sense induction (Van de Cruys and Apidianaki, 2011; Brody and Lapata, 2009), and information retrieval (Andrzejewski and Buttler, 2011).

When using a topic model, we are primarily concerned with the degree to which the learned topics match human judgments and help us differentiate between ideas. But until recently, the evaluation of these models has been ad hoc and application specific. Evaluations have ranged from fully automated intrinsic evaluations to manually crafted extrinsic evaluations. Previous extrinsic evaluations have used the learned topics to compactly represent a small fixed vocabulary and compared this distributional space to human judgments of similarity (Jurgens and Stevens, 2010). But these evaluations are hand constructed and often costly to perform for domain-specific topics. Conversely, intrinsic measures have evaluated the amount of information encoded by the topics, where perplexity is one common example(Wallach et al., 2009), however, Chang et al. (2009) found that these intrinsic measures do not always correlate with semantically interpretable topics. Furthermore, few evaluations have used the same metrics to compare distinct approaches such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003), Latent Semantic Analysis (LSA) (Landauer and Dutnais, 1997), and Non-negative Matrix Factorization (NMF) (Lee and Seung, 2000). This has made it difficult to know which method is most useful for a given application, or in terms of extracting useful topics.

We now provide a comprehensive and automated evaluation of these three distinct models (LDA, LSA, NMF), for automatically learning semantic topics. While these models have seen significant improvements, they still represent the core differences between each approach to modeling topics. For our evaluation, we use two recent automated coherence measures (Mimno et al., 2011; Newman et al., 2010)

Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 952–961, Jeju Island, Korea, 12–14 July 2012. ©2012 Association for Computational Linguistics

originally designed for LDA that bridge the gap between comparisons to human judgments and intrinsic measures such as perplexity. We consider several key questions:

1. How many topics should be learned?
2. How many learned topics are useful?
3. How do these topics relate to often used semantic tests?
4. How well do these topics identify similar documents?

We begin by summarizing the three topic models and highlighting their key differences. We then describe the two metrics. Afterwards, we focus on a series of experiments that address our four key questions and finally conclude with some overall remarks.

2 Topic Models

We evaluate three latent factor models that have seen widespread usage:

1. Latent Dirichlet Allocation
2. Latent Semantic Analysis with Singular Value Decomposition
3. Latent Semantic Analysis with Non-negative Matrix Factorization

Each of these models were designed with different goals and are supported by different statistical theories. We consider both LSA models as topic models as they have been used in a variety of similar contexts such as distributional similarity (Jurgens and Stevens, 2010) and word sense induction (Van de Cruys and Apidianaki, 2011; Brody and Lapata, 2009). We evaluate these distinct models on two shared tasks (1) grouping together similar words while separating unrelated words and (2) distinguishing between documents focusing on different concepts.

We distill the different models into a shared representation consisting of two sets of learned relations: how words interact with topics and how topics interact with documents. For a corpus with $\mathcal{D}$ documents and $\mathcal{V}$ words, we denote these relations in terms of $\mathcal{T}$ topics as

(1) a $\mathcal{V} \times \mathcal{T}$ matrix, W, that indicates the strength each word has in each topic, and

(2) a $\mathcal{T} \times \mathcal{D}$ matrix, H, that indicates the strength each topic has in each document.

$\mathcal{T}$ serves as a common parameter to each model.

2.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (Blei et al., 2003) learns the relationships between words, topics, and documents by assuming documents are generated by a particular probabilistic model. It first assumes that there are a fixed set of topics, $\mathcal{T}$ used throughout the corpus, and each topic z is associated with a multinomial distribution over the vocabulary Φ_z, which is drawn from a Dirichlet prior $Dir(\beta)$. A given document D_i is then generated by the following process

1. Choose $\Theta_i \sim Dir(\alpha)$, a topic distribution for D_i
2. For each word $w_j \in D_i$:
   (a) Select a topic $z_j \sim \Theta_i$
   (b) Select the word $w_j \sim \Phi_{z_j}$

In this model, the Θ distributions represent the probability of each topic appearing in each document and the Φ distributions represent the probability of words being used for each topic. These two sets of distributions correspond to our H and W matrices, respectively. The process above defines a generative model; given the observed corpus, we use collapsed Gibbs sampling implementation found in Mallet[1] to infer the values of the latent variables Φ and Θ (Griffiths and Steyvers, 2004). The model relies only on two additional hyper parameters, α and β, that guide the distributions.

2.2 Latent Semantic Analysis

Latent Semantic Analysis (Landauer and Dutnais, 1997; Landauer et al., 1998) learns topics by first forming a traditional term by document matrix used in information retrieval and then smoothing the counts to enhance the weight of informative words. Based on the original LSA model, we use the Log-Entropy transform. LSA then decomposes this smoothed, term by document matrix in order to generalize observed relations between words and documents. For both LSA models, we used implementations found in the S-Space package.[2]

Traditionally, LSA has used the Singular Value Decomposition, but we also consider Non-negative Matrix Factorization as we've seen NMF applied in similar situations (Pauca et al., 2004) and others

[1] http://mallet.cs.umass.edu/

[2] https://github.com/fozziethebeat/S-Space

Model	Label	Top Words	UMass	UCI
High Quality Topics				
LDA	interview	told asked wanted interview people made thought time called knew	-2.52	1.29
	wine	wine wines bottle grapes made winery cabernet grape pinot red	-1.97	1.30
NMF	grilling	grilled sweet spicy fried pork dish shrimp menu dishes sauce	-1.01	1.98
	cloning	embryonic cloned embryo human research stem embryos cell cloning cells	-1.84	1.46
SVD	cooking	sauce food restaurant water oil salt chicken pepper wine cup	-1.87	-1.21
	stocks	fund funds investors weapons stocks mutual stock movie film show	-2.30	-1.88
Low Quality Topics				
LDA	rates	10-yr rate 3-month percent 6-month bds bd 30-yr funds robot	-1.94	-12.32
	charity	fund contributions .com family apartment charities rent 22d children assistance	-2.43	-8.88
NMF	plants	stem fruitful stems trunk fruiting currants branches fence currant espalier	-3.12	-12.59
	farming	buzzards groundhog prune hoof pruned pruning vines wheelbarrow tree clematis	-1.90	-12.56
SVD	city	building city area buildings p.m. floors house listed eat-in a.m.	-2.70	-8.03
	time	p.m. system study a.m. office political found school night yesterday	-1.67	-7.02

Table 1: Top 10 words from several high and low quality topics when ordered by the UCI Coherence Measure. Topic labels were chosen in an ad hoc manner only to briefly summarize the topic's focus.

have found a connection between NMF and Probabilistic Latent Semantic Analysis (Ding et al., 2008), an extension to LSA. We later refer to these two LSA models simply as SVD and NMF to emphasize the difference in factorization method.

Singular Value Decomposition decomposes M into three smaller matrices

$$M = U\Sigma V^T$$

and minimizes Frobenius norm of M's reconstruction error with the constraint that the rows of U and V are orthonormal eigenvectors. Interestingly, the decomposition is agnostic to the number of desired dimensions. Instead, the rows and columns in U and V^T are ordered based on their descriptive power, i.e. how well they remove noise, which is encoded by the diagonal singular value matrix Σ. As such, reduction is done by retaining the first $\mathcal{T}$ rows and columns from U and V^T. For our generalization, we use $W = U\Sigma$ and $H = \Sigma V^T$. We note that values in U and V^T can be both negative and positive, preventing a straightforward interpretation as unnormalized probabilities

Non-negative Matrix Factorization also factorizes M by minimizing the reconstruction error, but with only one constraint: the decomposed matrices consist of only non-negative values. In this respect, we can consider it to be learning an unnormalized probability distributions over topics. We use the original Euclidean least squares definition of NMF[3]. Formally, NMF is defined as

$$M = WH$$

where H and W map directly onto our generalization. As in the original NMF work, we learn these unnormalized probabilities by initializing each set of probabilities at random and update them according to the following iterative update rules

$$W = W\frac{MH^T}{WHH^T} \quad H = H\frac{W^TM}{W^TWH}$$

3 Coherence Measures

Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artifacts of statistical inference, see Table 1 for examples ordered by the UCI measure. For our evaluations, we consider two new coherence measures designed for LDA, both of which have been shown to match well with human judgements of topic quality: (1) The UCI measure (Newman et al., 2010) and (2) The UMass measure (Mimno et al., 2011).

Both measures compute the coherence of a topic as the sum of pairwise distributional similarity

[3] We note that the alternative KL-Divergence form of NMF has been directly linked to PLSA (Ding et al., 2008)

scores over the set of topic words, V. We generalize this as

$$coherence(V) = \sum_{(v_i,v_j)\in V} score(v_i, v_j, \epsilon)$$

where V is a set of word describing the topic and ϵ indicates a smoothing factor which guarantees that $score$ returns real numbers. (We will be exploring the effect of the choice of ϵ; the original authors used $\epsilon = 1$.)

The UCI metric defines a word pair's score to be the pointwise mutual information (PMI) between two words, i.e.

$$score(v_i, v_j, \epsilon) = \log \frac{p(v_i, v_j) + \epsilon}{p(v_i)p(v_j)}$$

The word probabilities are computed by counting word co-occurrence frequencies in a sliding window over an external corpus, such as Wikipedia. To some degree, this metric can be thought of as an external comparison to known semantic evaluations.

The UMass metric defines the score to be based on document co-occurrence:

$$score(v_i, v_j, \epsilon) = \log \frac{D(v_i, v_j) + \epsilon}{D(v_j)}$$

where $D(x, y)$ counts the number of documents containing words x and y and $D(x)$ counts the number of documents containing x. Significantly, the UMass metric computes these counts over the *original corpus* used to train the topic models, rather than an external corpus. This metric is more intrinsic in nature. It attempts to confirm that the models learned data known to be in the corpus.

4 Evaluation

We evaluate the quality of our three topic models (LDA, SVD, and NMF) with three experiments. We focus first on evaluating aggregate coherence methods for a complete topic model and consider the differences between each model as we learn an increasing number of topics. Secondly, we compare coherence scores to previous semantic evaluations. Lastly, we use the learned topics in a classification task and evaluate whether or not coherent topics are equally informative when discriminating between documents.

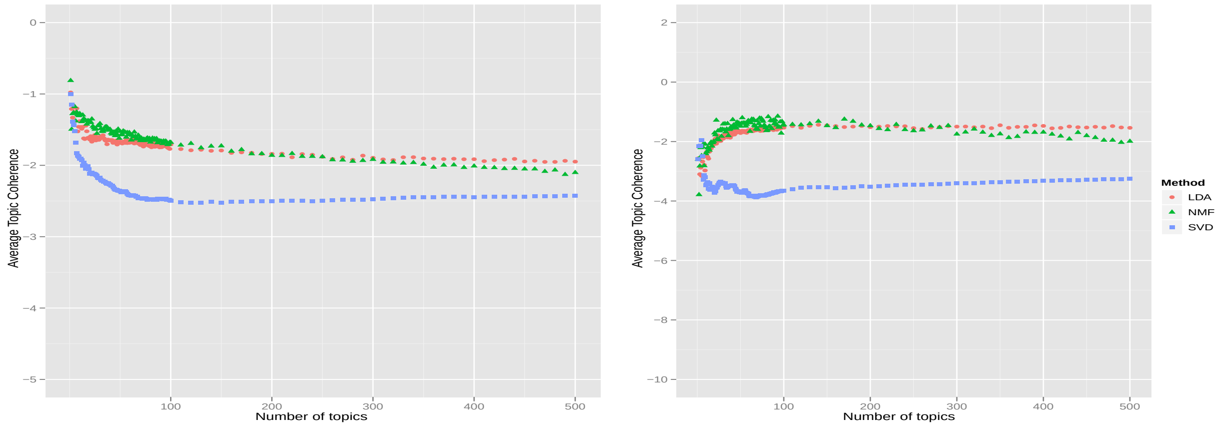
For all our experiments, we trained our models on 92,600 New York Times articles from 2003 (Sandhaus, 2008). For all articles, we removed stop words and any words occurring less than 200 times in the corpus, which left 35,836 unique tokens. All documents were tokenized with OpenNLP's MaxEnt[4] tokenizer. For the UCI measure, we compute the PMI between words using a 20 word sliding window passed over the WaCkypedia corpus (Baroni et al., 2009). In all experiments, we compute the coherence with the top 10 words from each topic that had the highest weight, in terms of LDA and NMF this corresponds with a high probability of the term describing the topic but for SVD there is no clear semantic interpretation.

4.1 Aggregate methods for topic coherence

Before we can compare topic models, we require an aggregate measure that represents the quality of a complete model, rather than individual topics. We consider two aggregates methods: (1) the average coherence of all topics and (2) the entropy of the coherence for all topics. The average coherence provides a quick summarization of a model's quality whereas the entropy provides an alternate summarization that differentiates between two interesting situations. Since entropy measures the complexity of a probability distribution, it can easily differentiate between uniform distributions and multimodal, distributions. This distinction is relevant when users prefer to have roughly uniform topic quality instead of a wide gap between high- and low-quality topics, or vice versa. We compute the entropy by dropping the log and ϵ factor from each scoring function.

Figure 1 shows the average coherence scores for each model as we vary the number of topics. These average scores indicate some simple relationships between the models: LDA and NMF have approximately the same performance and both models are consistently better than SVD. All of the models quickly reach a stable average score at around 100 topics. This initially suggests that learning more

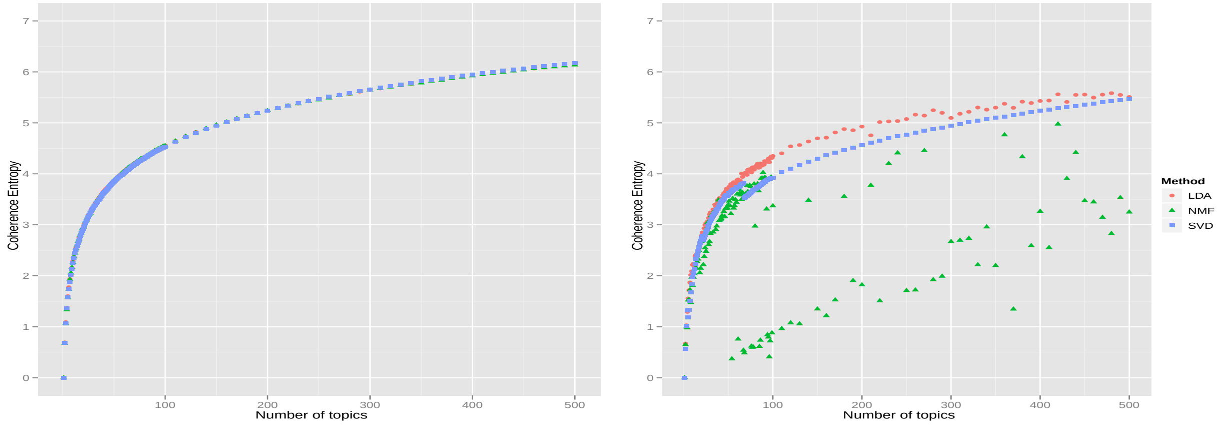
[4] http://incubator.apache.org/opennlp/



(a) UMass

(b) UCI

Figure 1: Average Topic Coherence for each model



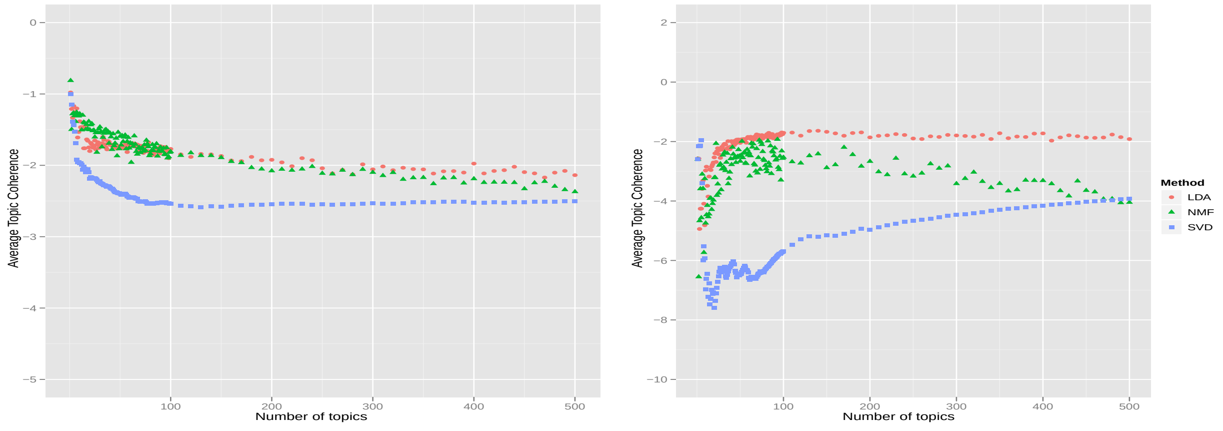
(a) UMass

(b) UCI

Figure 2: Entropy of the Topic Coherence for each model

topics neither increases or decreases the quality of the model, but Figure 2 indicates otherwise. While the entropy for the UMass score stays stable for all models, NMF produces erratic entropy results under the UCI score as we learn more topics. As entropy is higher for even distributions and lower for all other distributions, these results suggest that the NMF is learning topics with drastically different levels of quality, i.e. some with high quality and some with very low quality, but the average coherence over all topics do not account for this.

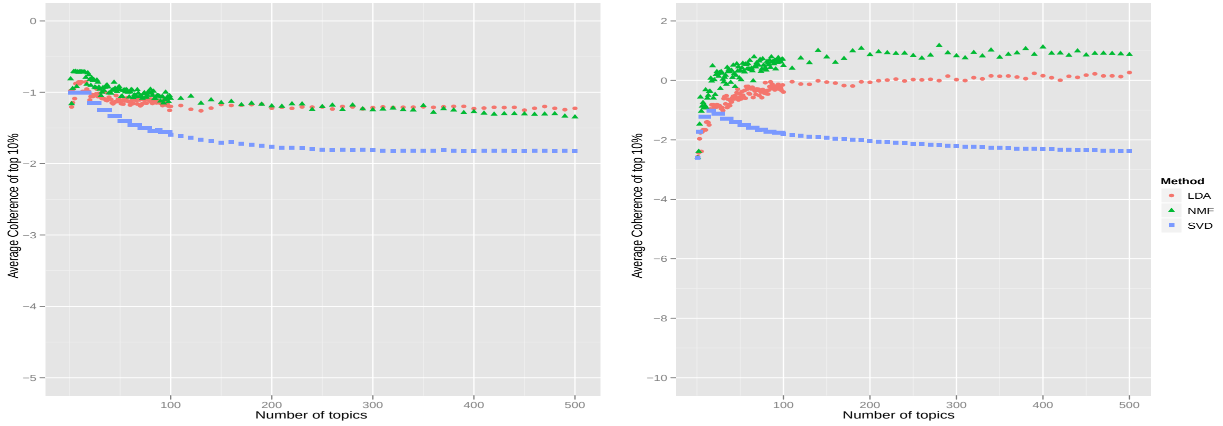
Low quality topics may be composed of highly unrelated words that can't be fit into another topic, and in this case, our smoothing factor, ϵ, may be artificially increasing the score for unrelated words. Following the practice of the original use of these metrics, in Figures 1 and 2 we set $\epsilon = 1$. In Figure 3, we consider $\epsilon = 10^{-12}$, which should significantly reduce the score for completely unrelated words. Here, we see a significant change in the performance of NMF, the average coherence decreases dramatically as we learn more topics. Similarly, performance of SVD drops dramatically and well below the other models. In figure 4 we lastly compute the average coherence using only the top 10% most coherence topics with $\epsilon = 10^{-12}$. Here, NMF again performs on par with LDA. With the top 10% topics still having a high average coherence but the full set



(a) UMass

(b) UCI

Figure 3: Average Topic Coherence with $\epsilon = 10^{-12}$



(a) UMass

(b) UCI

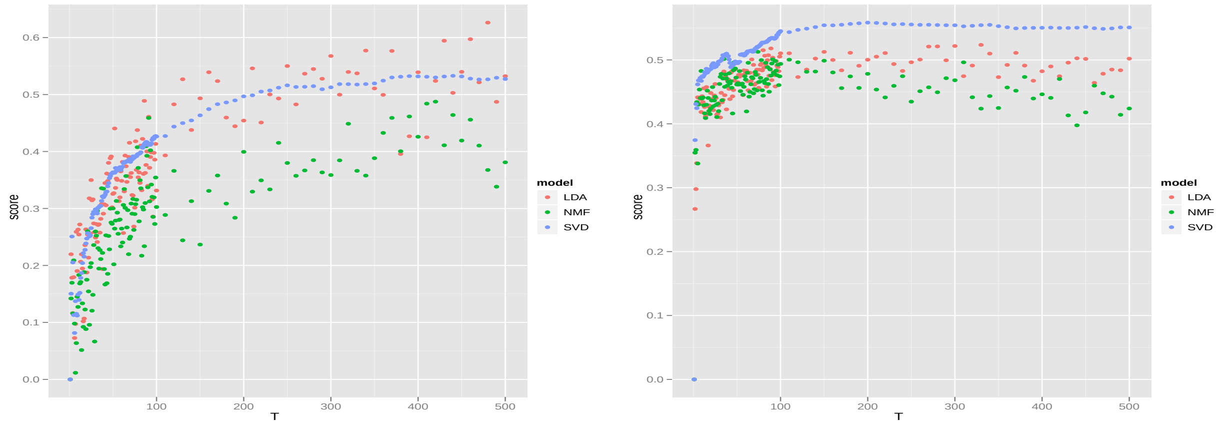
Figure 4: Average Topic Coherence of the top 10% topics with $\epsilon = 10^{-12}$

of topics having a low coherence, NMF appears to be learning more low quality topics once it's learned the first 100 topics, whereas LDA learns fewer low quality topics in general.

4.2 Word Similarity Tasks

The initial evaluations for each coherence measure asked human judges to directly evaluate topics (Newman et al., 2010; Mimno et al., 2011). We expand upon this comparison to human judgments by considering word similarity tasks that have often been used to evaluate distributional semantic spaces (Jurgens and Stevens, 2010). Here, we use the learned topics as generalized semantics describing our knowledge about words. If a model's topics generalize the knowledge accurately, we would expect similar words, such as "cat" and "dog", to be represented with a similar set of topics. Rather than evaluating individual topics, this similarity task considers the knowledge within the entire set of topics, the topics act as more compact representation for the known words in a corpus.

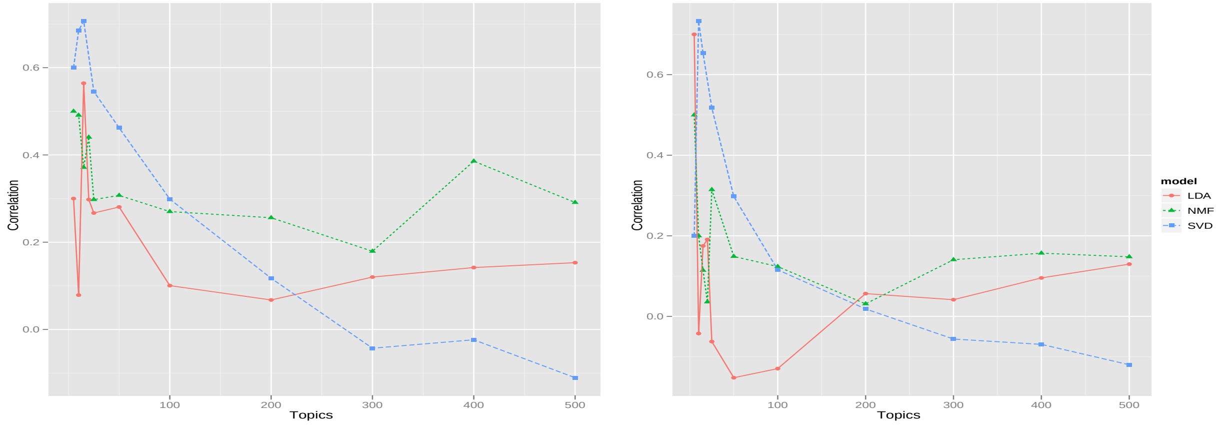
We use the Rubenstein and Goodenough (1965) and Finkelstein et al. (2002) word similarity tasks. In each task, human judges were asked to evaluate the similarity or relatedness between different sets of word pairs. Fifty-One Evaluators for the Rubenstein and Goodenough (1965) dataset were given 65 pairs



(a) Rubenstein & Goodenough

(b) Wordsim 353/Finklestein et. al.

Figure 5: Word Similarity Evaluations for each model



(a) UMass

(b) UCI

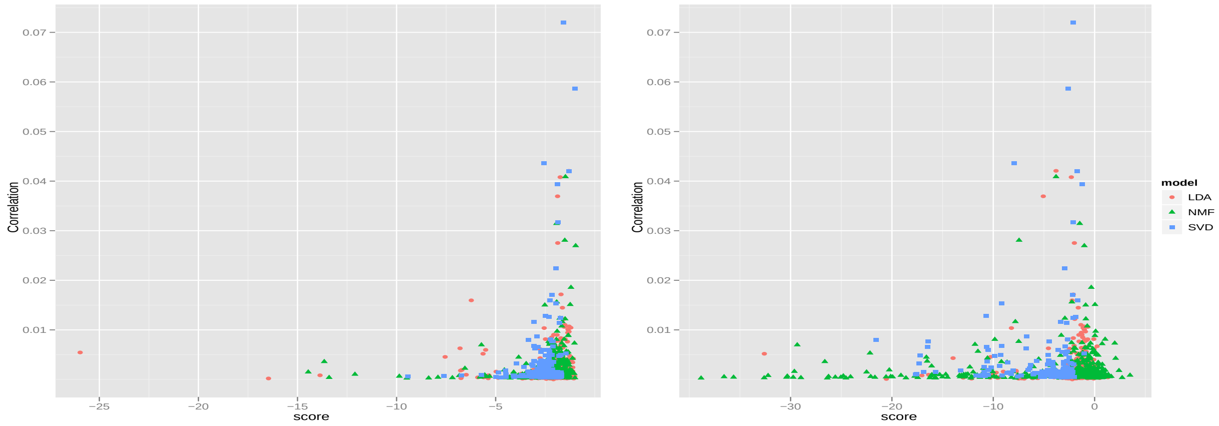
Figure 7: Correlation between topic coherence and topic ranking in classification

of words and asked to rate their similarity on a scale from 0 to 4, where a higher score indicates a more similar word pair. Finkelstein et al. (2002) broadens the word similarity evaluation and asked 13 to 16 different subjects to rate 353 word pairs on a scale from 0 to 10 based on their relatedness, where relatedness includes similarity and other semantic relations. We can evaluate each topic model by computing the cosine similarity between each pair of words in the evaluate set and then compare the model's ratings to the human ratings by ranked correlation. A high correlation signifies that the topics closely model human judgments.

Figure 5 displays the results. SVD and LDA both surpass NMF on the Rubenstein & Goodenough test while SVD is clearly the best model on the Finklestein et. al test. While our first experiment showed that SVD was the worst model in terms of topic coherence scores, this experiment indicates that SVD provides an accurate, stable, and reliable approximation to human judgements of similarity and relatedness between word pairs in comparison to other topic models.

4.3 Coherence versus Classification

For our final experiment, we examine the relationship between topic coherence and classification accuracy for each topic model. We suspect that highly



(a) UMass

(b) UCI

Figure 8: Comparison between topic coherence and topic rank with 500 topics

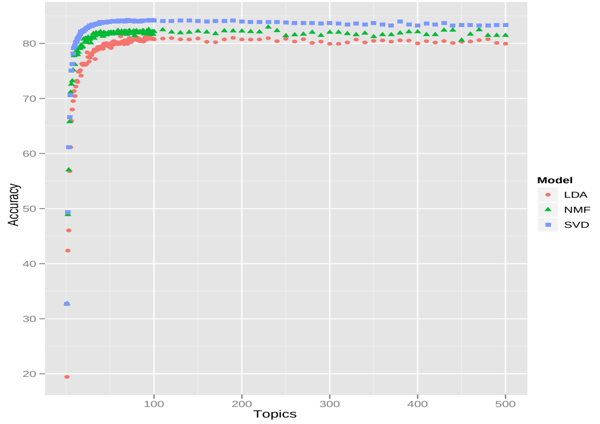


Figure 6: Classification accuracy for each model

coherent topics, and coherent topic models, will perform better for classification. We address this question by performing a document classification task using the topic representations of documents as input features and examine the relationship between topic coherence and the usefulness of the corresponding feature for classification.

We trained each topic model with all 92,600 New York Times articles as before. We use the section labels provided for each article as class labels, where each label indicates the on-line section(s) under which the article was published and should thus be related to the topics contained in each article. To reduce the noise in our data set we narrow down the articles to those that have only one label and whose label is applied to at least 2000 documents. This results in 57,696 articles with label distributions listed in Table 2. We then represent each document using columns in the topic by document matrix H learned for each topic model.

Label	Count	Label	Count
New York and Region	11219	U.S.	3675
Paid Death Notices	11152	Arts	3437
Opinion	8038	World	3330
Business	7494	Style	2137
Sports	7214		

Table 2: Section label counts for New York Times articles used for classification

For each topic model trained on N topics, we performed stratified 10-fold cross-validation on the 57,696 labeled articles. In each fold, we build an automatically-sized bagged ensemble of unpruned CART-style decision trees(Banfield et al., 2007) on 90% of the dataset[5], use that ensemble to assign labels to the other 10%, and measure the accuracy of that assignment. Figure 6 shows the average classification accuracy over all ten folds for each model. Interestingly, SVD has slightly, but statistically significantly, higher accuracy results than both NMF and LDA. Furthermore, performance quickly increases

[5] The precise choice of the classifier scheme matters little, as long as it is accurate, speedy, and robust to label noise; all of which is true of the choice here.

and plateaus with well under 50 topics.

Our bagged decision trees can also determine the importance of each feature during classification. We evaluate the strength of each topic during classification by tracking the number of times each node in our decision trees observe each topic, please see (Caruana et al., 2006) for more details. Figure 8 plot the relationship between this feature ranking and the topic coherence for each topic when training LDA, SVD, and NMF on 500 topics. Most topics for each model provide little classification information, but SVD shows a much higher rank for several topics with a relatively higher coherence score. Interestingly, for all models, the most coherent topics are not the most informative. Figure 7 plots a more compact view of this same relationship: the Spearman rank correlation between classification feature rank and topic coherence. NMF shows the highest correlation between rank and coherence, but none of the models show a high correlation when using more than 100 topics. SVD has the lowest correlation, which is probably due to the model's overall low coherence yet high classification accuracy.

5 Discussion and Conclusion

Through our experiments, we made several exciting and interesting discoveries. First, we discovered that the coherence metrics depend heavily on the smoothing factor ϵ. The original value, 1.0 created a positive bias towards NMF from both metrics even when NMF generated incoherent topics. The high smoothing factor also gave a significant increase to SVD scores. We suspect that this was not an issue in previous studies with the coherence measures as LDA prefers to form topics from words that co-occur frequently, whereas NMF and SVD have no such preferences and often create low quality topics from completely unrelated words. Therefore, we suggest a smaller ϵ value in general.

We also found that the UCI measure often agreed with the UMass measure, but the UCI-entropy aggregate method induced more separation between LSA, SVD, and NMF in terms of topic coherence. This measure also revealed the importance of the smoothing factor for topic coherence measures.

With respects to human judgements, we found that coherence scores do not always indicate a better representation of distributional information. The SVD model consistently out performed both LDA and NMF models, which each had higher coherence scores, when attempting to predict human judgements of similarity.

Lastly, we found all models capable of producing topics that improved document classification. At the same time, SVD provided the most information during classification and outperformed the other models, which again had more coherent topics. Our comparison between topic coherence scores and feature importance in classification revealed that relatively high quality topics, but not the most coherent topics, drive most of the classification decisions, and most topics do not affect the accuracy.

Overall, we see that each topic model paradigm has it's own strengths and weaknesses. Latent Semantic Analysis with Singular Value Decomposition fails to form individual topics that aggregate similar words, but it does remarkably well when considering all the learned topics as similar words develop a similar topic representation. These topics similarly perform well during classification. Conversely, both Non Negative Matrix factorization and Latent Dirichlet Allocation learn concise and coherent topics and achieved similar performance on our evaluations. However, NMF learns more incoherent topics than LDA and SVD. For applications in which a human end-user will interact with learned topics, the flexibility of LDA and the coherence advantages of LDA warrant strong consideration. All of code for this work will be made available through an open source project.[6]

6 Acknowledgments

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 (LLNL-CONF-522871) and by Sandia National Laboratory under Contract DE-AC04-94AL85000.

References

David Andrzejewski and David Buttler. 2011. Latent topic feedback for information retrieval. In *Proceed-*

[6] https://github.com/fozziethebeat/TopicModelComparison

ings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, pages 600–608, New York, NY, USA. ACM.

Robert E. Banfield, Lawrence O. Hall, Kevin W. Bowyer, and W. Philip Kegelmeyer. 2007. A comparison of decision tree ensemble creation techniques. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(1):173–180, January.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky wide web: A collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

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

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 103–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rich Caruana, Mohamed Elhawary, Art Munson, Mirek Riedewald, Daria Sorokina, Daniel Fink, Wesley M. Hochachka, and Steve Kelling. 2006. Mining citizen science data to predict orevalence of wild bird species. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '06, pages 909–915, New York, NY, USA. ACM.

Jonathan Chang, Sean Gerrish, Chong Wang, and David M Blei. 2009. Reading tea leaves : How humans interpret topic models. *New York*, 31:1–9.

Chris Ding, Tao Li, and Wei Peng. 2008. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Comput. Stat. Data Anal.*, 52:3913–3927, April.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: the concept revisited. *ACM Trans. Inf. Syst.*, 20:116–131, January.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Suppl. 1):5228–5235, April.

David Jurgens and Keith Stevens. 2010. The s-space package: an open source package for word space models. In *Proceedings of the ACL 2010 System Demonstrations*, ACLDemos '10, pages 30–35, Stroudsburg, PA, USA. Association for Computational Linguistics.

Thomas K Landauer and Susan T. Dutnais. 1997. A solution to platos problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, pages 211–240.

Thomas K. Landauer, Peter W. Foltz, and Darrell Laham. 1998. An Introduction to Latent Semantic Analysis. *Discourse Processes*, (25):259–284.

Daniel D. Lee and H. Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. In *In NIPS*, pages 556–562. MIT 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 Emperical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association of Computational Linguistics.

David Newman, Youn Noh, Edmund Talley, Sarvnaz Karimi, and Timothy Baldwin. 2010. Evaluating topic models for digital libraries. In *Proceedings of the 10th annual joint conference on Digital libraries*, JCDL '10, pages 215–224, New York, NY, USA. ACM.

V Paul Pauca, Farial Shahnaz, Michael W Berry, and Robert J Plemmons, 2004. *Text mining using nonnegative matrix factorizations*, volume 54, pages 452–456. SIAM.

Herbert Rubenstein and John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM*, 8:627–633, October.

Evan Sandhaus. 2008. The New York Times Annotated Corpus.

Tim Van de Cruys and Marianna Apidianaki. 2011. Latent semantic word sense induction and disambiguation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 1476–1485, Stroudsburg, PA, USA. Association for Computational Linguistics.

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. 2009. Evaluation methods for topic models. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*. Omnipress.