Frameworks: A Topic Modeling Project

Literature Review

Patricia J Goedecke, Jerry Duncan, Preston Provins

Supervised by Dr. Audris Mockus

COCS 545 Digital Archaeology

The University of Tennessee Knoxville

Fall 2018

Introduction

The challenge of modeling human language using artificial intelligence (AI) has been termed "AI-hard" or "AI-complete," meaning equivalent in difficulty to the task of making computers as intelligent as humans, a sort of holy grail in machine learning (Yampolskiy, 2013). This challenge implies that until computers surpass humans in human language use - as for example in chess - the art of perfecting AI representations of human language will continue.

Topic modeling using latent Dirichlet analysis (LDA) is a dimensionality reduction technique for large corpuses of text, clustering vocabulary terms into topics found to be distributed probabilistically among documents. Dimensionality, the number of topics (K) to model, is as yet qualitatively determined, though a number of methods have been suggested to quantitatively determine K. Methods proposed have included measures of perplexity and its difference, goodness of fit, and pointwise mutual information.

This project seeks to test another method of determining the appropriate level of K for a corpus. For this project, simulated corpora will be developed with various levels of vocabulary size, corpus size and topic distribution; these simulated corpora will have pre-determined levels of K. Topic models will be tested on these corpora in search of a mathematical relation between the quantitative measures of a corpus, and its appropriate dimensionality.

To test this mathematical relation on a real-world corpus, topic models will also be developed for a corpus of Java framework issues and comments scraped from GitHub and SourceForge sites. Here, best fit of model dimensionality will be measured as predictive ability of topic proportions applied as model features, predicting success of each framework.

Our literature review addresses text analysis methods leading up to LDA topic modeling, and subsequently developed methods. While recent methods are considered more advanced for their language parsing abilities and computational efficiency, they are generally "black box" methods providing little illumination as to the relevant content of text modeled.

1 Methods leading to Latent Dirichlet Allocation

1.1 Latent Semantic Analysis and singular value decomposition

Topic modeling was arguably begun by Deerwester et al. (1990) in an article introducing Latent Semantic Analysis (LSA) as an indexing method. Indexing methods originated in information retrieval with text data, developing criteria for measuring similarity between words and documents. The authors sought to overcome a limitation of earlier indexing methods, namely that search terms might not precisely match the vocabulary used in a document. Their insight was that an underlying probabilistic relationship between terms and documents could be imputed and then described using statistical methods. Their approach uses singular value decomposition (SVD).

In the SVD approach of Deerwester et al. (1990), both terms and documents are represented as vectors in a space of (some chosen value) K dimensions; similarity is measured by the dot product or cosine between points in the space. Any term can be represented in this multidimensional space, with synonyms presumably occupying nearly the same space. A search term will now retrieve not only the identical term but terms with similar meaning, and also terms used in a similar context. A polysemic term may occupy two or more locations in this vector space, each accompanied by terms of the appropriate context. For example, "tee" as a shirt may collocate with terms such as "shirt," "clothing," and "casual wear," while "tee" as a golf instrument may collocate near "putter" and "wood." This collocating space is considered the topic.

These vectors comprise a term by document matrix which is then decomposed into a set of orthogonal factors, from which the original matrix can be approximated by linear combination. The overall matrix is factored into two matrices, one representing the documents; the other, the topics. Any document now can be represented as a linear combination of topics.

1.2 Probabilistic Latent Semantic Analysis

Hofmann (1999) introduced Probabilistic Latent Semantic Analysis (pLSA), a step toward latent Dirichlet allocation. Probabilistic Latent Semantic Analysis imagines an automated document generating process. For

each word position i in document d, a topic z(d(i)) is first drawn probabilistically from a discrete set of topics, with probability θ(d). Given the topic z(d(i)), a word is then selected from the list associated with that topic. The words are probabilistically associated with each topic, with probability ϕ(z(d(i))). Each document is then represented as a linear combination of topics, and pLSA is likewise a linear topic model based on the factorization of the document-word matrix.

In pLSA, each document is a probabilistic mixture of topics; topics are modeled as occurring within documents according to a multinomial probability distribution. Blei et al. (2003) criticized pLSA for the following: 1) ill-defined document generative semantics, and 2) the number of parameters increasing linearly with the number of training documents, making the model susceptible to overfitting.

1.3 Non-Negative Matrix Factorization

In the same year as pLSA, Lee and Seung (1999) introduced Non-Negative Matrix Factorization (NMF), described as a neural network. NMF is contrasted with principle components analysis and vector quantization as representing parts-based rather than holistic representations. These constraints lead to a parts-based representation because they allow only additive, not subtractive, combinations. NMF is applied in representing facial images as well as semantic features of texts; it uses an iterative algorithm, starting from non-negative initial conditions. When used with text data, a set of H hidden variables are iterated with a column of W word frequencies, forming matrix V, constructing approximate factorizations.

Ding et al. (2008) showed NMF to be equivalent to pLSA, in that both optimize the same objective function (Pleple, 2013a):

$$F = \sum_{i=1}^{n} \sum_{\mu=1}^{m} [V_{i\mu} log(WH)_{i\mu} - (WH)_{i\mu}]$$

1.4 Latent Dirichlet Allocation and its reception

The contribution of Blei et al. (2003) distinguishing Latent Dirichlet Allocation (LDA) from pLSA was the Bayesian element of prior distributions on parameters. As with pLSA, Blei et al. imagine documents generated probabilistically from topics with multinomial probabilities, where a prior distribution is placed on the

parameter of the multinomial distribution. Every iteration in model development now improves on the prior model in a theoretically consistent manner, based on Bayesian prior and posterior distributions.

The LDA methodology of Blei et al. (2003) has been contested (Chang et al. 2009, Wallach et al. 2009, Stevens 2012) and perhaps improved upon. Mimno et al. (2011) provide three contributions to improve LDA modeling:

> (1) An analysis of the ways in which topics can be flawed; (2) an automated evaluation metric for identifying such topics that does not rely on human annotators or reference collections outside the training data; (3) a novel statistical topic model based on this metric that significantly improves topic quality in a large-scale document collection.

Nevertheless, LDA is the current standard in topic modeling. Topic modeling is recognized but still under scrutiny as an effective data mining tool. Teh et al. (2004, 2012) propose a hierarchical Dirichlet process which they find to outperform the simple LDA model. Wei and Croft (2006) compare LDA with information retrieval using cluster-based models and find LDA compares favorably. AlSumait el al. (2009) examine the topic significance ranking of LDA models. Musat et al. (2011) propose improving topic evaluation using conceptual knowledge. Niraula et al. (2013) compare LDA with LSA, finding LSA to be more effective regarding semantic similarity. Banjade et al. (2015) seek to combine a variety of methods for measuring word-to-word similarity. Among topic modeling researchers, rather than replacing LDA, researchers are now seeking ways to use LDA more effectively. In addition, the method is being applied to different types of big data, such as biomedical data sets (Pritchard et al., 2000, Falush et al., 2003, Zhao et al., 2014, Zhao et al., 2015).

2 Current trends in LDA topic modeling

While LDA is the current standard in topic modeling, and topic modeling is popular as a dimension reduction technique and clustering tool, the method may still be considered a work in progress. Two open areas of investigation are topic coherence and selecting the dimensionality of a model. Also of interest is the expansion of topic modeling into broader types of data mining, beyond text.

2.1 Expanding topic modeling applications

Topic modeling developed from a pursuit for effective information retrieval, in an age of ever-increasing data. Even prior to Blei et al.'s (2003) contribution of latent Dirichlet allocation to the method, Lee and Seung (1999) had applied non-negative matrix factorization to image data as well as text. In their studies, Gerber et al. (2007) and Zhao et al. (2014) effectively apply topic modeling to human gene expression data. Mimno et al. (2011) anticipate that the methods of topic modeling will need to incorporate ever-larger corpora of data.

2.2 Topic coherence

A troubling concern regarding topic modeling as a dimension reduction technique for data is that topics modeled may make little sense to human interpreters; or may appear flawed. A number of researchers have sought to evaluate or improve on the coherence of topics modeled.

Newman et al. (2010) tested automated techniques for evaluating topic coherence, including pointwise mutual information (PMI) and other lexical relatedness measures based on WordNet, Wikipedia and Google's search engine. They found several Wikipedia-based measures to show strong results, including one using PMI from Wikipedia source data. PMI measures the relatedness between any two words as the discrepancy between the relative frequency of their co-occurence $p(x, y)$ and the relative frequency predicted by their joint probabilities when independence is assumed $p(x)p(y)$.

Mimno et al. (2011) addressed the concern that topics derived from the automated LDA method frequently baffle human interpreters. They identified several specific types of topic flaws, presented a method for evaluating topic coherence automatedly, and provided a new statistical topic model which they call a Generalized Pólya Urn model. In their discussion, Mimno et al. (2011) express the concerns that going forward, improving semantic cohesion will be of primary concern to text modeling researchers, along with scaling to ever larger data sets.

3 Recent developments in text analysis

Blei et al.'s (2003) topic modeling with latent Dirichlet allocation (LDA), is generally considered to have been surpassed by models such as Mikolov et al.'s (2013a) more recently dominant word embedding in word2vec. Other recent developments include variable-length models in doc2vec; approaches merging global co-occurrence probability matrices with word embedding; and unsupervised recurrent neural networks (RNNs). While these methods provide superior computational efficiency as compared with topic modeling, and include sentence-level as well as document-level relationships between words, they are black box methods generally unrevealing of much regarding the content of the text analyzed.

3.1 WORD2VEC: Mikolov et al. (2013a); CBOW and Skip-Gram

Bengio et al. (2001) introduced word embedding as a dimension reduction of vocabularies from one dimension per term to a vector representation per term, in generally several hundred dimensions. Mikolov et al. (2013a) introduce word2vec in two simultaneously presented methods: Continuous-Bag-of-Words (CBOW) and Skip-Gram. These word embeddings derive predictions and hence vector weightings from the words immediately surrounding them in consecutive text. Mikolov et al. set themselves the challenge of increasing the size of text corpora to billions of words which can be incorporated in a neural net model. They improve upon earlier neural net models with an approximational method to reduce computational complexity.

Mikolov et al. first compare their method against a Feedforward Neural Net Language Model (NNLM; Bengio et al., 2003), explaining that the NNLM computational complexity Q could be defined as

$$Q = N * D + N * D * H + H * V$$

where N is the number of preceding terms included in the model (typically 10), N*D=P is the size of the projection layer, which might be 500-2000, H is the size of the hidden layer, often 500-1000, and V is the size of the vocabulary. This complexity is dominated by the H*V term, because of the typically large vocabulary size. Mikolov et al. explain that they reduce computational complexity by using hierarchical softmax, and representing the vocabulary with a Huffman binary tree.

Mikolov et al. then make comparison with Recurrent Neural Net Language Models (RNNLM; Bengio & LeCun, 2007; Mikolov et al., 2010). They explain that the RNNLM has no projection layer, but rather the hidden layer connects with itself using a time delay. This results in a computational complexity for RNNLM of:

$$Q = H * H + H * V$$

Here, word representations (D) have the same dimensionality as hidden layer H, and H * V can be reduced to H*log2(V) using hierarchical softmax.

Mikolov et al.'s (2013a) new architecture employs log-linear models and relies on earlier work, particularly Mikolov's 2007 master's thesis work. In this two-step method, continuous word vectors are first learned using a simple model; then an N-gram NNLM is trained on top of these. For the CBOW model, the current word is predicted from a window of N preceding and subsequent words; this research team found their best results with N=4. Computational complexity Q for CBOW is found as:

$$Q = N * D + D * \log2(V)$$

For the Skip-Gram model, the predictive order is reversed; the current word is used to predict words within its window. In both cases, it is not the prediction that is the desired output; rather, the weightings used to obtain predictions are the output of interest.

3.2 DOC2VEC: Le and Mikolov (2014) Paragraph Vector

Le and Mikolov (2014) introduce a new methodology, pointing out two major weaknesses with the popular bag-of-words methodology (BOW, Harris, 1954). The first weakness of bag-of-words, which Le and Mikolov describe as obvious, is that word order is lost. For the second weakness they argue that in losing word order, semantics are lost. Le and Mikolov propose to address these weaknesses using what they term Paragraph Vector, an unsupervised algorithm for learning fixed-length feature representations of variable-length texts. They call these texts paragraphs, which may be of any length from phrase to sentence to document. ("Doc2vec" appears to be a moniker applied to this method after the fact.)

Le and Mikolov compare their method both with word-level vector methods and with methods aspiring toward phrase- or sentence-level modeling. They explain that in earlier work, simple models were found to be more effective than complex ones; however, with increasing capacity of CPUs, complex models are now becoming feasible. A simple document-level model, then, represents a document as the weighted average of the vectors of words in the document. At an intermediate level of complexity, Bengio et al. (2006) use the concatenation of several previous word vectors to form the input of a neural network, which then aims to predict the next word. A more complex model by Socher et al. (2011) parses sentences; Le and Mikolov admire Socher et al.'s work but contend that its primary weakness is that it can only operate at sentence length; a secondary weakness is that it requires parsing.

Specifically, Le and Mikolov propose a vector representation of the current paragraph, which is then concatenated with vector representations of words. The word representations are constant regardless of paragraph, while each paragraph representation is unique. The word vectors are then predicted within the given paragraph. That is, at prediction time, the paragraph vectors are inferred by fixing the word vectors and training the new paragraph vector until convergence.

3.3 Global matrices with word embedding

Pennington et al.'s (2014) Global vectors (GloVe), Moody's (2016) lda2vec and Bunk and Krestel's (2018) Word Embeddings with Latent Dirichlet Allocation (WELDA) each combine the strengths of topic models with those of word embedding. GloVe will be the focus here, as Pennington et al. make a clear logical, mathematical and transferable argument for combining the two methods. Pennington et al. use the phrase "global matrix factorization methods" to describe what may also be termed topic modeling; for word embedding, they use the term "local context window methods." They describe the strengths of these methods as such: topic modeling methods aptly leverage statistical measures of co-occurrence of words across a corpus, i.e., globally; word embedding methods develop a strong vector space structure, allowing effective analogy calculations.

Discussing meaning that can be found probabilistically, Pennington et al. reduce word embedding methods to the following: "they attempt to maximize the log probability as a context window scans over the corpus"

(p 1535). Pennington et al. argue that a more appropriate starting point for word vector learning should be ratios of co-occurrence probabilities, rather than the probabilities themselves. That is, for target words i and j, their relationship can be illuminated by exploring the ratio of co-occurrence with various probe words k. For example, if i=ice, j=steam, and k=solid, they expect a high co-occurrence ratio of Pik/Pjk; for k=gas, this ratio should be small. Pennington et al. argue that these ratios illuminate relevance and relationships between words. Pennington et al. go on to describe the global objective function J of word embedding models as an unweighted sum of the log probabilities of each word within each context window.

Considering this unweighted sum an unnecessarily burdensome oversimplification of global word relationships, Pennington et al. introduce a weighting function. While allowing that a variety of functions could meet these criteria, they argue for the following properties as of primary importance to the weighting: 1. f (0) = 0. If f is viewed as a continuous function, it should vanish as x-¿0 fast enough that the lim x-¿0 f (x) log2 x is finite. 2. f (x) should be non-decreasing so that rare co-occurrences are not overweighted. 3. f (x) should be relatively small for large values of x, so that frequent co-occurrences are not overweighted. (p 1535)

Moody (2016) and Bunk and Krestel (2018) notably each incorporate Blei et al.'s latent Dirichlet allocation (LDA) methodology into the topic modeling portion of their models. Similarly to Pennington et al., they merge topic modeling with word embedding. The specific strength of LDA is that it provides theoretical coherence for the updating of a topic model's probability distribution across iterations.

Pennington et al. succinctly report the effectiveness of the GloVe model: "The model produces a vector space with meaningful substructure, as evidenced by its performance of 75% on a recent word analogy task. It also outperforms related models on similarity tasks and named entity recognition" (p 1532).

3.4 Recurrent neural networks (RNNs) and future directions

LeCun et al. (2015) see the future of AI natural language processing in unsupervised recurrent neural networks (RNNs). As these three authors are Manning's (2015) declared "giant[s] of modern Deep Learning" (p 701), they would seem to speak with some authority on the matter. Notable among the accomplishments of RNN in recent years, LeCun et al. point out that RNNs have succeeded in predicting the next character in a text

(Sutskever et al., 2011); the next word in a sequence (Mikolov et al., 2013b); and have in their words encoded the "thought" of a sentence in one language, so that it may be decoded in another (Sutskever et al., 2014).

When Manning (2015) expresses his wishes for future developments of natural language processing with AI, he suggests a return to "the interesting linguistic and cognitive issues that motivated noncategorical representations and neural network approaches" (p 704). He hopes that rather than striving to inch down the error percentage on benchmark tasks, researchers will look to the content matter of language for inspiration. As an example of an intrigue built into the English language, Manning observes the categorical ambiguity of the gerund V-ing form. He provides examples of this form being used as an adjective, verb, noun or preposition.

While Manning might like to see the depths of this ambiguity explored, LeCun et al. would leap from this to the genius of neural networks. They argue that rather than requiring painstakingly derived rules, "neural networks just use big activity vectors, big weight matrices and scalar non-linearities to perform the type of fast 'intuitive' inference that underpins effortless commonsense reasoning." That is to say, LeCun el al. argue that neural nets already learn the way humans do, i.e., from observation rather than instruction

Bibliography

AlSumait, L., Barbará, D., Gentle, J., and Domeniconi, C. (2009, September). Topic significance ranking of LDA generative models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 67-82. Springer Berlin Heidelberg.

Banjade, R., Maharjan, N., Niraula, N. B., Rus, V., and Gautam, D. (2015, April). Lemon and tea are not similar: Measuring word-to-word similarity by combining different methods. *International Conference on Intelligent Text Processing and Computational Linguistics*, 335-346. Springer International Publishing, Chicago.

Bengio, Y., Ducharme, R. & Vincent, P. (2001). A neural probabilistic language model. In Proc. Advances in Neural Information Processing Systems 13, 932–938.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003). A neural probabilistic language model. Journal of Machine Learning Research, 3(Feb), 1137-1155. Blei, D. M., & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes.*Journal of Machine Learning Research*, 12(Aug), 2461-2488.

Blei, D.M., Ng, A.Y. and Jordan, M.I., 2003. Latent dirichlet allocation. Journal of Machine Learning Research, 3(Jan), 993-1022.

Bunk, S. and Krestel, R. (2018). WELDA: Enhancing Topic Models by Incorporating Local Word Context. In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, 293-302. ACM.

Chang, J., Boyd-Graber, J. L., Gerrish, S., Wang, C., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems (NIPS)*, 31, 1-9.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6), 391.

Ding, C., Li, T. and Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics and Data Analysis*, 52(8), 3913-3927.

Falush, D., Stephens, M., and Pritchard, J. K. (2003). Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics*, 164(4), 1567-1587.

Gerber, G. K., Dowell, R. D., Jaakkola, T. S., and Gifford, D. K. (2007). Automated discovery of functional generality of human gene expression programs. *PLOS Computational Biology*, 3(8), e148.

Goldberg, Y. and Levy, O. (2014). Word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. arXiv preprint arXiv:1402.3722. Harris, Z.S., 1954. Distributional structure. Word, 10(2-3), 146-162.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 50-57. ACM.

Kiros, R., Zhu, Y., Salakhutdinov, R.R., Zemel, R., Urtasun, R., Torralba, A. and Fidler, S. (2015). Skip-thought vectors. In Advances in Neural Information Processing Systems, 3294-3302.

Lau, J.H. and Baldwin, T. (2016). An empirical evaluation of doc2vec with practical insights into document embedding generation. arXiv preprint arXiv:1607.05368.

Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In International Conference on Machine Learning, 1188-1196.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. Nature, 521(7553), 436.

Lee, D. D., and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755), 788-791.

Manning, C.D. (2015). Computational linguistics and deep learning. Computational Linguistics, 41(4), 701-707.

McCallum, A. K. (2002). Mallet: A machine learning for language toolkit.

Mekala, D., Gupta, V., Paranjape, B. and Karnick, H. (2016). SCDV: Sparse Composite Document Vectors using soft clustering over distributional representations. arXiv preprint arXiv:1612.06778.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013a). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S. and Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, 3111-3119.

Mikolov, T., Yih, W.T. and Zweig, G. (2013c). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 746-751.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262-272. Association for Computational Linguistics.

Moody, C.E. (2016). Mixing dirichlet topic models and word embeddings to make lda2vec. arXiv preprint arXiv:1605.02019.

Musat, C., Velcin, J., Trausan-Matu, S., and Rizoiu, M. A. (2011). Improving topic evaluation using conceptual knowledge. In *22nd International Joint Conference on Artificial Intelligence* (IJCAI), 3, 1866-1871.

Newman, D., Lau, J. H., Grieser, K., and Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100-108. Association for Computational Linguistics.

Niraula, N., Banjade, R., Ştefănescu, D., and Rus, V. (2013). Experiments with semantic similarity measures based on LDA and LSA. In *Statistical Language and Speech Processing*, 188-199. Springer Berlin Heidelberg.

Papadimitriou, C.H., Tamaki, H., Raghavan, P. and Vempala, S. (1998). Latent semantic indexing: A probabilistic analysis. In Proceedings of the Seventeenth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (159-168). ACM.

Pathik, N. And Shukla, P. (2017). An ample analysis on extended LDA models for aspect based review analysis. International Journal of Computer Science & Applications, 14(2).

Pennington, J., Socher, R. and Manning, C. (2014). Glove: Global vectors for word representation. In

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532-1543.

Pleple, Q. (2013). Topic modeling bibliography. Retrieved from http://qpleple.com/bib/

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155(2), 945-959.

Socher, R., Lin, C.C., Ng, A., and Manning, C. (2011). Parsing natural scenes and natural language with recursive neural networks. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), 129–136.

Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. (2012). Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 952-961. Association for Computational Linguistics.

Sutskever, I., Martens, J. and Hinton, G. E. (2011). Generating text with recurrent neural networks. In Proc. 28th International Conference on Machine Learning, 1017–1024.

Sutskever, I. Vinyals, O. and Le. Q. V. (2014). Sequence to sequence learning with neural networks. In Proc. Advances in Neural Information Processing Systems 27, 3104–3112.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2004). Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes. In *Neural Information Processing Systems* (NIPS), 1385-1392.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2012). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* (JASA), 1566-1581.

Van Landeghem, J. (2016). A Survey of Word Embedding Literature.

Wei, X., and Croft, W. B. (2006, August). LDA-based document models for ad-hoc retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 178-185. ACM.

Wallach, H. M., Murray, I., Salakhutdinov, R., and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1105-1112. ACM.

Yampolskiy, R.V. (2013). Turing test as a defining feature of AI-completeness. In Artificial Intelligence, Evolutionary Computing and Metaheuristics, 3-17. Springer, Berlin, Heidelberg.

Yu, M. and Dredze, M. (2014). Improving lexical embeddings with semantic knowledge. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) 2, 545-550.

Yu, M. and Dredze, M. (2015). Learning composition models for phrase embeddings. Transactions of the Association for Computational Linguistics, 3, 227-242.

Yu, M., Gormley, M. and Dredze, M. (2014). Factor-based compositional embedding models. In NIPS Workshop on Learning Semantics, 95-101.

Zhao, W., Zou, W. and Chen, J.J. (2014). Topic modeling for cluster analysis of large biological and medical

datasets. *BMC Bioinformatics*, 15, Suppl 11:511.

Zhao, W., Chen, J. J., Perkins, R., Liu, Z., Ge, W., Ding, Y., & Zou, W. (2015). A heuristic approach to determine an appropriate number of topics in topic modeling. *BMC Bioinformatics*, 16(13), S8.