Topic Modeling with Latent Dirichlet Allocation

Dylan Leeman

December 12, 2007

Abstract

The prevalent document-topic models are based on Latent Dirichlet Allocation (LDA). Stop words and words appearing with very low frequency in the corpus are generally removed before the corpus is sampled. Here, I investigate the effect of these types of noise on the inference process, in an attempt to identify the precise conditions under which they produce various effects. The model is applied with and without stop words to real and artificial corpora. The usefulness of the model is then analyzed with respect to the real data.

1 Introduction

In this paper I present a brief survey of statistical methods for topic modeling, as well as illustrating several motivations for using such techniques. I will examine briefly Probabilistic Latent Semantic Indexing (pLSI), as described in [Hof99]. In addition, the Latent Dirichlet Allocation method, as proposed by [BNJ03], will be described in some detail, as it is this model and derivatives thereof that are most commonly advanced as solutions to the problem of statistical topic modeling.

While the generative models of text corpora used in topic modeling are fairly straightforward, inferring the parametrization of those models, given a corpus is much more difficult. Various methods are therefore applied to the observed variables (documents distributed over words), in order to infer the parameters of unobserved variables (the distribution of documents over topics, and the distribution of topics over words). Methods of statistical inference most commonly used for this are expectation maximization, variational distribution, and Gibbs sampling, the last of which I employ here.

Like many solutions for statistical natural language processing, Gibbs sampling with Latent Dirichlet Allocation suffers from the influence of noise words. Stop words, or words that occur with very high frequency, tend to be syntactic, rather than semantic words, and have a potentially skewing effect on the topic model. Low-frequency words, by contrast, are likely to impart highly specialized semantic information; regrettably, these words are even more problematic, because their tendency is to distribute themselves amongst the topics without bias. Worse, if the words occur with very low frequency in the corpus, the individual instances of those words will not settle into a particular topic, but rather their topic assignments will vary from one iteration of the sampling algorithm to the next.

2 Uses of Topic Modeling

Topic modeling, and more generally, the field of data mining, has a broad range of applications, from its usefulness in understanding and describing neurological transactions during language learning,

to its capacity to enrich and refine search queries. A common, general task for topic models is, once trained, to be able to classify a given document with respect to the set of generated topics. This could be used to organize a collection of documents with respect to their contents, or it could be applied to meta-data concerning the documents, giving a more uniform set of terms by which to find relevant documents, as described in [NHCS07].

Classically, researchers wishing to study natural language and natural language acquisition have tried to create models whereby sentences (strings of symbols from the lexicon, in this case words) are categorized as grammatical or ungrammatical, according to their membership in the language generated by an innately human grammar. This approach, while undoubtedly both meaningful and intriguing, has two drawbacks: it is impossible for a machine to generalize to the correct grammar, given a finite subset of the strings generated by the grammar (a set of sentences on which to "learn" the language), and sentences that are rejected by this categorization are often considered perfectly acceptable by speakers of the language, furthermore, even malformed sentences can convey meaning. [MS99] In contrast, statistical language processing permits a greater latitude with respect to syntax, while simultaneously representing semantic content.

In order to demonstrate the utility of topic modeling, and to test and refine their entity-topic model, Newman et al. modeled a corpus of 330,000 New York Times articles taken from the period 2000-2002 [NCS06]. The resulting topics and distributions represent a kind of *Zeitgeist* of the period in question, in particular, it hints at the most significant entities (figures, places, and events), in a topical context.

Though the language about topic modeling comes almost exclusively from the language of verbal discourse, it is possible to apply topic modeling to non-verbal corpora. For example, [BL03][1] successfully modeled tagged images. Another possible application of topic-modeling to nonverbal corpora is the modeling of consumer purchase trends. Due to barcodes and the digitization of purchase transactions, retailers now maintain massive databases of information regarding the habits of consumers. The topic-model might be successfully applied to these transactions to determine a set of "topics", or more generalized purchase patterns of consumers, and to predict the purchases individual consumers are likely to make in the future.

3 Probabilistic Latent Semantic Analysis (pLSA)

Probabilistic Latent Semantic Analysis was first proposed by Thomas Hofmann [Hof99], and incorporates the basic principle by which the great majority of topic modeling is performed. Hofmann describes a model in which documents are distributed among an unobserved set of class variables (topics), each of which is distributed over the vocabulary of the corpus. This is symbolized in the following:

$$P(w \mid d) = \sum_{z \in Z} [P(w \mid z) P(z \mid d)] \tag{1}$$

where $P(w \mid d)$ is the probability of w (a word) given d (a document), in terms of the probability $P(w \mid z)$ (the probability of a word given a topic) and the probability $P(z \mid d)$ (the probability of a topic given a document). Therefore probability of a particular word w appearing in the document d is derived by summing the probabilities of that word appearing in the topic z multiplied by the probability of z appearing the document for all topics $z \in Z$. Notice that this model does allow for

[1] Cited in [NCS06].

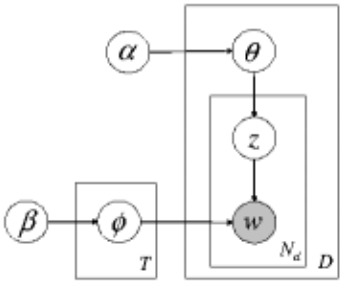

Figure 1: Plate Diagram of Basic LDA model [RZGSS04]

polysemy (multiplicity of meanings), as a word could be generated from any of several topics, as well as *synonymy* (multiple words with the same meaning), because multiple words with the same semantic meaning are generated from the same topic. Also, unlike many previous models, pLSA does not assume that a document is composed of only one topic. (To see why this assumption is problematic, take the case of a proposal to tax housecats. Clearly, the document is composed of at least two topics: taxation and cats, but a traditional model would coerce the document into a single topic, both obscuring the true nature of the document and skewing the topic assignments.)

Unfortunately, the definition of pLSA restricts its use to those documents in the training set, by virtue of its dependence on an index into the corpus to describe a document [BNJ03]. Clearly this is a severely limiting quality, as one of the most useful applications of a topic model is to be able to apply the model, once trained, to a new document in order to classify the document with respect to topics. [BNJ03] propose to eliminate this restriction by making sampling a distribution over topics for each document.

4 Latent Dirichlet Allocation (LDA)

4.1 Generative Model

Latent Dirichlet Allocation (LDA) is a generative model for document modeling detailed in [BNJ03] and seeking to overcome the limitations of pLSA. The selection of LDA as a reasonable means for document-topic modeling arises from the assumption of *exchangeability* made by a majority of topic-models. Exchangeability in this context refers both to exchangeability of words within a document, as well as exchangeability of *documents within a corpus*. Words are exchangeable within a document if any pair of words (w_i and w_j) can be swapped within the document (that is w_i becomes w_j and vice versa), and not alter the document. This property holds regardless of how many times the swapping is performed, that is, any permutation of the words in a given document yields the same document. The assumption of exchangeability of words is often referred to as the "bag of words" approach, while the corpus is a "bag of documents". It is the premise of exchangeability that permits LDA to generalize to new documents.[BNJ03]

The following high-level algorithm is used to generate a document under LDA (from [BNJ03]):

1. Choose $N \sim Poisson\,(\zeta)$.

2. Choose $\theta \sim Dir(\alpha)$.

3. For each of the N words w_n:

   (a) Choose a topic $z_n \sim Multinomial(\theta)$.

   (b) Choose a word w_n from $p(w_n \mid z_n, \beta)$, a multinomial probability conditioned on the topic z_n.

Where N is the number of words in the document (here this is selected from the Poisson distribution, but better methods can be used depending on the corpus to be modeled.) Then the document-topic distribution θ is sampled from a Dirichlet prior α, and the set of words is generated from this distribution by repeatedly sampling a topic from θ and then sampling a word from the topic. The key feature distinguishing LDA from pLSA is the selection of θ (a distribution over topics) from $Dir(\alpha)$, instead of selecting a document index d from the corpus. This allows the LDA model to generalize to documents outside of the training set. In the case of pLSA, the only possible distributions of documents over topics are those found in the training set. LDA, by contrast, models the document-topic distribution over all possible real-valued distributions for the given number of topics. This is illustrated in Figure 2. Every possible sample point in the Dirichlet distribution for the topic set represents a distribution of topics over the words of the vocabulary, and it is this property which allows LDA to generalize beyond the training set.

4.2 Artificial Corpus Generation

Though the ultimate goal of topic modeling, and indeed all statistical language processing systems, is to model real samples of natural language, it is useful to have a mechanism for generating corpora that approximate real data, and whose parameters can be finely controlled. Here, the generative nature of the topic models proves extremely valuable. In order to perform experiments on the influence of classes of exceptional words, I created a simple generator that permits variation of three parameters: the size of the vocabulary, the number of topics to create, and the number of documents in the corpus. The generator can then be supplemented with tools to add stop words, low-frequency words, and other additional data with which one wishes to experiment.

In the initial design, the generator chose document lengths randomly, in magnitude proportional to the size of the vocabulary divided by the number of topics to be modeled. As the number of topics is always several orders of magnitude less than the size of the vocabulary for any reasonable corpus, the average length of each document was generally on a similar order of magnitude as the vocabulary size. This process was chosen with two concerns in mind: that documents should vary in length, and that documents should be sufficiently long to accurately express the topics which contribute most significantly to them. For comparison, the Earlham College Community Documents (ECCD) corpus has a vocabulary size of approximately 10,000 words after the removal of stop words and low-frequency words, while the average length of each document is on the order of 500. Because the results of experiments using this generator were unsatisfactory, I elected to alter the generator, to produce corpora more in line with the ratios I encountered in the ECCD corpus.

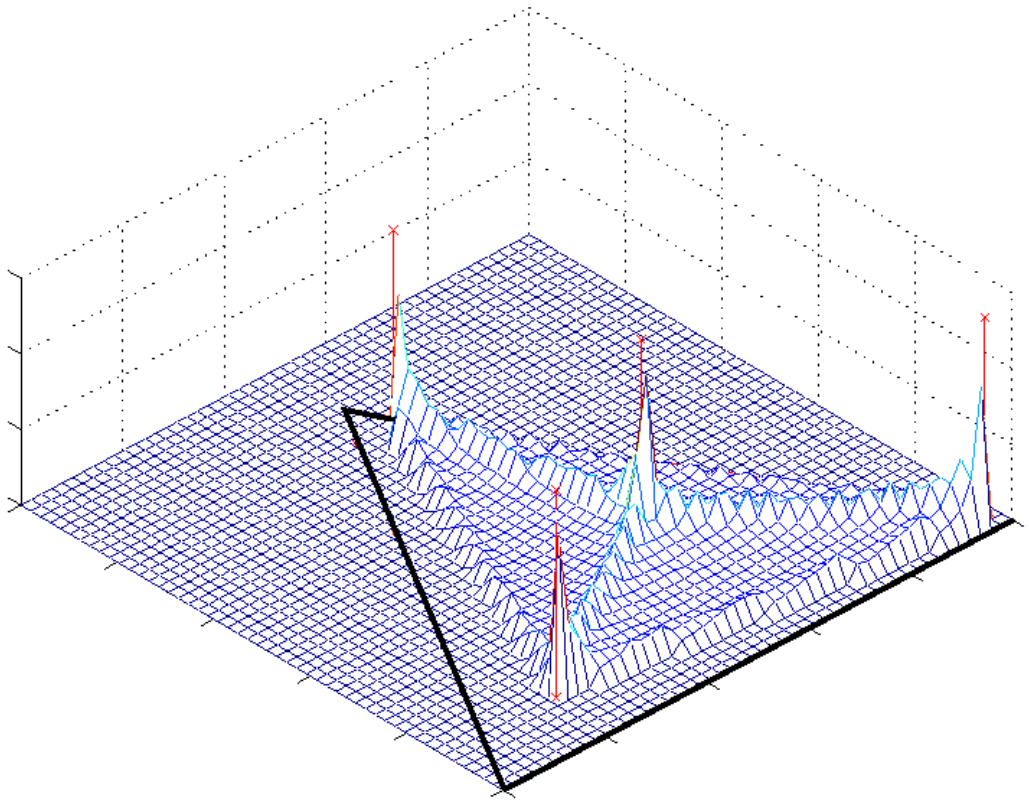

Figure 2: Density Distributions $p(w \mid \theta, \beta)$ from [BNJ03] The points of the triangle represent words, while locations marked with x are topics.

5 Bayesian Inference / Updating

Thus far, I have presented only the generative models of documents and corpora used in topic modeling; however, topic modeling seeks to identify properties of an *existing* corpus. Statistical inference is the process by which the distributions of documents over topics and of topics over words are discovered, given a corpus. A number of algorithms have been proposed for this purpose, but they all rely heavily on an extension of Bayes' theorem (Eq. 2)

$$P(B \mid A) = \frac{P(A \mid B)\, P(B)}{P(A)} \tag{2}$$

called Bayesian updating, which is an iterative application of the above equality, to successively refine estimates of B, given an initial B_0 and a series of events $a \in A$. For example, given an initial estimate of the distribution of a particular topic over words **w**, and an infinite series of drawn words from that topic, the estimate **w'** can be refined to reflect the distribution used to generate the words, to an arbitrary precision. Of course, this problem is somewhat more complicated in topic modeling, where the only known distribution is the joint distribution of documents over words (joint because it is the observed result of the unobserved selection of a topic in the document and a word in the selected topic), from which **w** and **z** must be inferred.

The assumption that a document is composed primarily of just a few topics is crucial to our ability to infer both θ (the probability of a topic given a document) and ϕ (the probability of a word given a topic). On the other hand, the smoothing effect of the Dirichlet priors means that the topic model is unlikely to become mired in a self-fulfilling prophecy of sorts, in which all words from all documents in the corpus are generated from the same topic. Intuitively, the process of Bayesian updating under the LDA model works as follows:

1. For each $d \in D$:
   (a) Select a word w_i from d.
   (b) Calculate the probability of w_i having been generated by each of the topics in d.
   (c) Probabilistically assign w_i to a topic based on those probabilities.
   (d) Update document-topic and topic-word distributions based on the above selection.

The likelihood of assigning a word token to a particular topic grows with the number of times other tokens of the same word have been assigned to that topic, also different words from the same document will have a greater likelihood of being assigned to the same topic.

Take a simple example of three documents and two topics. The topics are ¡cats, taxes¿. Document A has the mixture $\langle 0.5, 0.5 \rangle$, document B has the mixture $\langle 1, 0 \rangle$, and document C has the mixture $\langle 0, 1 \rangle$. On the first iteration of the inference algorithm, distribution of words in document A with respect to topics will be random, in other words, the distribution over topics will be correct, even though the distribution of topics over words will be grossly inaccurate. When document B is sampled, the words will group naturally into one topic (which will ultimately become the "cats" topic), since the first meaningful word will be placed in one topic or the other, marginally increasing the likelihood that subsequent word tokens from the document will be placed in the same topic. The same will happen for C, although the words in C will tend to be assigned to the "taxes" topic, particularly since the topic assignments from sampling B will have greatly decreased the likelihood of words in C being assigned to the "cats" topic. After a large number of iterations, the distribution of words in the two topics will closely resemble the mixture from the generative model used to create the documents.

6 Gibbs Sampling

The Gibbs sampling algorithm on the Latent Dirichlet Allocation topic model is scarcely more complex than the high-level description presented in section 5. It follows:

1. Randomly assign a topic to each word token in the corpus.
2. For each $d \in D$:
   (a) Select a word w_i from d.
   (b) Remove the influence of w_i from the model.
   (c) Calculate the probability of w_i having been generated by each of the topics z_j:
      i. Calculate the probability of topic z_j in d.
      ii. Calculate the probability of word w_i in z_j.

(d) Probabilistically assign w_i to a topic based on the resulting distribution.

The Gibbs sampling algorithm terminates after some number n of iterations, determined before sampling begins, rather than attempting to test for convergence. It is, however, customary to save the state of the sampler at periodic intervals after some "burn-in" period, during which the word-topic and topic-document distributions are assumed to be too random to be meaningful.

The sampling algorithm has time complexity $O(I * W * 2T)$, where I is the number of iterations (typically on the order of 10^3), W is the total number of *word tokens* in the corpus, and T is the number of topics to be inferred. Unfortunately, due to the progressive nature of the sampler, which relies on repeated updating of probabilities and on random selection, it is difficult, though not impossible, to parallelize.[2]

7 Stop Words

7.1 Syntactic or Semantic?

The notion of a class of semantic – "meaning" – words is crucial to topic modeling, since a "topic" in a statistical probabilistic topic model is a distribution over words, whose individual and collective meanings make up the topic. According to [GSBT05], "[a] word can appear in a sentence for two reasons: because it serves a syntactic *function*, or because it provides semantic *content*."[3] They go on to state that "[syntactic] constraints result in relatively short-range dependencies. ... Semantic constraints result in long-range dependencies: different sentences within a document are likely to have similar content, and use similar words."[4] Given these notions of syntactic and semantic classes of words, it is useful to ask, which words are semantic words? Which words are syntactic words? How does membership in these classes vary from corpus to corpus? From discourse to discourse?

7.2 Nature of Stop Words

Noise words have the potential to skew topic-document and word-topic distributions in a statistical topic model like Latent Dirichlet Allocation. The most commonly treated type of noise is stop words. Stop words are words which appear with great frequency in a corpus, or indeed, in any corpus in the language in which the corpus is written. Stop words are usually function or syntax words, which have no semantic meaning of their own. Examples of English-language stop words include articles such as "a" and "the", and prepositions such as "of", "on" and "beyond".

A natural representation of stop words in the generative view of corpora assumed by topic models like LDA is for all stop words to be generated from the same topic, which comprises a large proportion of every document in the corpus. This view of stop words effectively models them as members of a single topic in which membership means only that the word is meaningless without context, that is, a semantic class of innately unsemantic words. In spite of any tendency in human interpretation to view stop words in this manner, there is no compelling intuitive reason to believe that stop words cluster in a topic model. Therefore, in an effort to determine the actual behavior of stop words in a corpus, and the conditions under which they behave in particular ways,

[2]For more on this: `http://www.ics.uci.edu/~newman/parallel_topic_model_newman_20060721.doc`

[3]Emphasis added.

[4]It should be reiterated that it is precisely this assumption that permits topic models such as Latent Dirichlet Allocation.

I conducted several experiments with artificially generated corpora. Then, I modeled the Earlham College Community Documents with and without stop words.

There exists another class of noise words, similar in nature to stop words, which apply not to natural languages, but to discourses. Such an example is the decision by [NHCS07] to model the topics of the OAIster catalog[5] with words that were either "topically too broad, or topically too specific. Examples of topically broad words include *january*, *february* (months of the year) and *result*, *paper*, *study* (words often used in research articles). Examples of topically specific words include *santa_ana* (city in California) and *ladies_repository* (the name of a historic period)." While the first type of words are similar in nature to stop words, the second are examples of very low frequency words, which I discuss briefly in section **??**.

7.3 Modeling the Effect of Stop words on Artificially-Generated Corpora

In the first series of experiments, I created a set of corpora using the generator described in 4.2. Each corpus contained 1500 documents. The vocabulary size was varied in ten roughly equal-sized increments between 375 words and 2283 words ($1250 \pm 70\%$), while the topics remained constant at 15. A second set of corpora were created by fixing the size of the vocabulary at 1250 and varying the number of topics in 10 roughly equal-sized increments, between 7 and 25 ($15 \pm 50\%$). Stop words were added from a list of 50 English-language stop words with proportion $\frac{1}{5}$ to copies of each of the twenty corpora, yielding twenty experimental corpora with twenty corresponding control corpora. The increased document length was not taken into account in measuring the effect of the stop words on the corpora. I conducted a second set of experiments, with parameters almost identical to the first, but with the second incarnation of the generator (see section 4.1) and a stop words list of only 25 words. Furthermore, the second set of corpora consisted of only 300 documents each. In both cases, the stop words distributed themselves roughly evenly among the topics.

7.4 Stop words in Real Corpora

I was able to observe the effect of stop words on a real corpus by modeling the Earlham College Community Documents corpus both with and without stop words. In the first run, an aggressive stop words policy was employed, under which English-language stop words were removed.[6] The topic set that emerged from this run of the sampler was extraordinarily promising. Topics were easily identifiable, with the exception of some small amount of noise that I had anticipated. The topics generated from the second and third runs of the Gibbs sampling algorithm were semantically not as useful. As in the case with the artificial corpora, the stopwords tended to distribute across several topics, although of note is the fact that they were not distributed "evenly" across *all* topics. There was a single topic that occurred with high probability in nearly all documents, and several others which were also composed primarily of stop words that occurred frequently in high proportion in many documents. While the prominent semantic words in these topics did not have (subjectively) a high degree of semantic correlation, the topics from the other, less frequent, lower distribution topics *did* exhibit strong coherency. For instance, a strong topic from this category was the "ECS" topic, which included such words as "software" and "systems", as well as entities such as "Moodle", and "Tom Steffes" (the current Director of Computing Services). Though this

[5] See `http://www.oaister.org`

[6] The list of stop words used can be found at `http://www.dcs.gla.ac.uk/idom/ir_resources/linguistic_utils/stop_words`

second class of topics did contain noise words, they were less common, and overall the topics were more focused.

8 Earlham College Community Documents (ECCD)

The Earlham College Community Documents corpus is a collection of approximately 800 documents, the bulk of which are committee meeting minutes from meetings between the years 2003 and 2007. Forty-four committees contribute to the Community Documents corpus, and this provides a natural number of topics to infer when running the Gibbs sampling algorithm on the ECCD corpus.

I sampled only a portion of the full ECCD corpus, due to a bug in the initial implementation of the code, which prevented me from using the full corpus. The reduced corpus had 400 documents, instead of the approximately 800 documents of the full corpus. After tokenization and removal of English-language stop words and low-frequency words, the reduced corpus has a vocabulary of about 11,000 unique words. The sampling algorithm was allowed to run twice for 1,000 and 5,000 iterations respectively, given Dirichlet priors for both α and β of .05. The sampler was directed to infer 45 topics.

Perhaps the most significant outcome of these runs was the production, in both cases, of a topic which contributed significantly to all documents in the corpus (making up approximately 1/5 of the entire corpus), whose most prominent members seemed to be *ECCD stop words*. For example, the words "committee", "faculty", and "meeting" were the most common words in this topic, while the only two entities that appeared with any significant frequency were "Earlham College" and "Doug Bennett" (the president of Earlham College during the period that the ECCD corpus covers). This can be seen in Figure 3.

9 Conclusions

The Latent Dirichlet Allocation topic model is clearly a useful device with many significant and worthwhile applications; however, its use must be supplemented with careful interpretation. Though topic models are currently no substitute for the human eye, they can serve as a valuable assistive technology, particularly when the corpus to be examined is relatively large in comparison with that portion of it that is useful to the reader. The standard removal of stop words and low-frequency words, while trivial to execute, may be anything but trivial in effect. It is difficult to predict precisely the behavior of the topic model, given two corpora which differ only in their degree of preprocessing. Furthermore, the lack of any mechanism for handling structure, indicated primarily by syntactic words (which are removed before processing begins), is troubling in that it denies the consumer of the information produced by the topic model vital semantic information conveyed by the structure of the data.

10 Future Work

10.1 Low-frequency Words

I had originally intended to perform experiments on very low-frequency words similar to those that I performed for stop words, because these words are also commonly removed from corpora before

sampling. For example, [NCS06] removed all words that did not appear in at least 10 documents of a corpus of 330,000 *New York Times* articles. While these words are undoubtedly noise to the model – they occur so infrequently that the kinds of broad correlations inferred by LDA are impossible to infer with any accuracy, they are also valuable semantic words, expressing extremely specific ideas and information. In general, removal of these words would severely limit the qualitative efficacy of a topic model; however, as models of this type do not accurately reflect the semantics of these words in any case, their removal is no great tragedy, but merely a symptom of a deeper problem.

10.2 Document Focus

The "bag of words" view of documents is useful in topic modeling applications, because it allows models to treat only those words that are taken to be *semantically* meaningful, that is, those words with long-range dependencies. Any human author (or indeed reader), would be appalled at the suggestion that documents are mere "bags of words", as structure is an important component of all levels of documents. It would seem useful, then, for topic models to account in some fashion for structure in human-authored documents. To determine if this is feasible or useful, I had hoped to conduct experiments regarding the focus of documents with respect to portions thereof. I intended to measure focus as the largest portion of the document composed of some fixed number of topics, that is, what is the maximum portion of the document that can be accounted for by summing the distributions for the largest n topics in the document? Regrettably, I did not have enough time to complete this portion of my exploration.

References

[BGJT04] D. Blei, T. Gri, M. Jordan, and J. Tenenbaum. Hierarchical topic models and the nested chinese restaurant process, 2004.

[BL03] D. Blei and J Lafferty. Modeling annotated data. In *Proceedings of the Annual Conference on Research and Development in Information Retrieval (SIGIR03)*, 2003.

[BNJ03] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[CG92] George Casella and Edward I. George. Explaining the gibbs sampler. *The American Statistician*, 46(3):167–174, aug 1992.

[CH01] David Cohn and Thomas Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity, 2001.

[GS07] Thomas Griffiths and Mark Steyvers. Probabilistic topic models. In T. Landauer, D. S. McNamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum, 2007.

[GSBT05] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In Lawrence K. Saul, Yair Weiss, and Leon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, Cambridge, MA, 2005.

[Hof99] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM Press.

[MS99] Christopher D. Manning and Hinrich Schtze. *Foundations of Statistical Natural Language Processing*. The MIT Press, June 1999.

[NCS06] David Newman, Chaitanya Chemudugunta, and Padhraic Smyth. Statistical entity-topic models. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686, New York, NY, USA, 2006. ACM Press.

[NHCS07] David Newman, Kat Hagedorn, Chaitanya Chemudugunta, and Padhraic Smyth. Subject metadata enrichment using statistical topic models. In *JCDL '07: Proceedings of the 2007 conference on Digital libraries*, pages 366–375, New York, NY, USA, 2007. ACM.

[RZGSS04] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, Arlington, VA, USA, 2004. AUAI Press.

[TJBB03] Y. Teh, M. Jordan, M. Beal, and D. Blei. Hierarchical dirichlet processes, 2003.

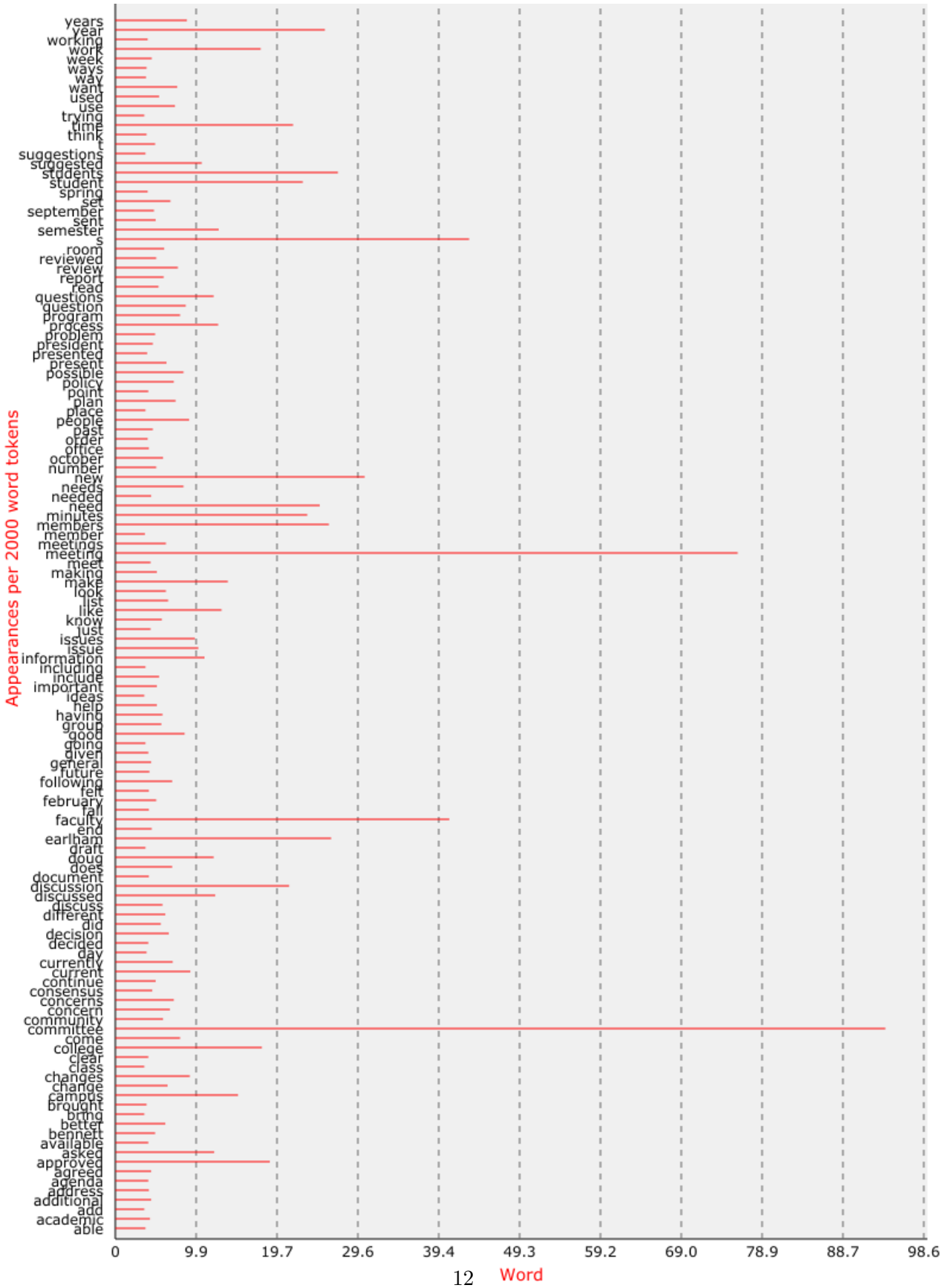

Figure 3: "Topic 11" from the ECCD corpus with stop words and low-frequency words removed.