Visualizing Topic Models

Allison J.B. Chaney and **David M. Blei**
Department of Computer Science
Princeton University, Princeton, NJ USA
achaney@cs.princeton.edu, blei@cs.princeton.edu

Abstract

Managing large collections of documents is an important problem for many areas of science, industry, and culture. Probabilistic topic modeling offers a promising solution. Topic modeling is an unsupervised machine learning method that learns the underlying themes in a large collection of otherwise unorganized documents. This discovered structure summarizes and organizes the documents. However, topic models are high-level statistical tools—a user must scrutinize numerical distributions to understand and explore their results. In this paper, we present a method for visualizing topic models. Our method creates a navigator of the documents, allowing users to explore the hidden structure that a topic model discovers. These browsing interfaces reveal meaningful patterns in a collection, helping end-users explore and understand its contents in new ways. We provide open source software of our method.

Understanding and navigating large collections of documents has become an important activity in many spheres. However, many document collections are not coherently organized and organizing them by hand is impractical. We need automated ways to discover and visualize the structure of a collection in order to more easily explore its contents.

Probabilistic topic modeling is a set of machine learning tools that may provide a solution (Blei and Lafferty 2009). Topic modeling algorithms discover a hidden thematic structure in a collection of documents; they find salient themes and represent each document as a combination of themes.

However, topic models are high-level statistical tools. A user must scrutinize numerical distributions to understand and explore their results; the raw output of the model is not enough to create an easily explored corpus.

We propose a method for using a fitted topic model to organize, summarize, visualize, and interact with a corpus. With our method, users can explore the corpus, moving between high level discovered summaries (the "topics") and the documents themselves, as Figure 1 illustrates.

Our design is centered around the idea that the model both summarizes and organizes the collection. Our method translates these representations into a visual system for exploring a collection, but visualizing this structure is not enough. The discovered structure induces relationships—between topics and articles, and between articles and articles—which lead to interactions in the visualization.

Thus, we have three main goals in designing the visualization: summarize the corpus for the user; reveal the relationships between the content and summaries; and, reveal the relationships across content. We aim to present these in a ways that are accessible and useful to a spectrum of users, not just machine learning experts.

Related Work

Several solutions to the problem of understanding large document corpora include Exemplar-based Visualization (Chen et al. 2009), ThemeRiver (Havre, Hetzler, and Nowell 2000), and FacetAtlas (Cao et al. 2010). These visualizations help users understand the corpus as a whole, but do not enable exploration of individual documents. Our visualization provides both a high-level summary of the corpus and links between the summary and individual documents.

Previous topic modeling research has focused on building new topic models and improving algorithms for fitting them. Researchers have typically used browsers to evaluate model algorithms (Newman et al. 2006; Gretarsson et al. 2011; Gardener et al. 2010). These browsers emphasize topics, giving little attention to documents. Further, they include little visual representation, relying mostly on numbers to convey meaning.

We present a way of using topic models to help learn about and discover items in a corpus. Our navigator presents the output of a topic model in an interface that illuminates a given corpus to non-technical users.

Probabilistic Topic Models

We review topic modeling, focusing on *latent Dirichlet allocation* (LDA) (Blei, Ng, and Jordan 2003), which is one of the simplest probabilistic topic models. LDA decomposes a collection of documents into *topics*—biased probability distributions over terms—and represents each document with a (weighted) subset of the topics. When fit to a set of documents, the topics are interpretable as themes in the collection, and the document representations indicate which themes each document is about. Thus, the learned topics summarize the collection, and the document representations organize the corpus into overlapping groups.

LDA is a statistical model of a collection of texts. *Hidden random variables* encode its thematic structure. The assumptions LDA makes can be articulated by its *probabilistic generative process*, the imaginary process by which the collection arose.

Figure 1

Wikipedia Topics — *Relative Presence of Topics in all Documents*

- {household, population, female}
- {film, series, show}
- {theory, work, human}
- {son, year, death}
- {war, force, army}
- {system, computer, user}
- {album, band, music}
- {government, party, election}
- {game, team, player}

{film, series, show}

words	related documents	related topics
film	The X-Files	{son, year, death}
series	Orson Welles	{work, book, publish}
show	Stanley Kubrick	{album, band, music}
character	B movie	{woman, child, man}
play	Mystery Science Theater 3000	{law, state, case}
make	Monty Python	{black, white, people}
episode	Doctor Who	{theory, work, human}
movie	Sam Peckinpah	{@card@, make, design}
good	Married... with Children	{war, force, army}
release	History of film	{god, call, give}

Stanley Kubrick

Stanley Kubrick (July 26, 1928 – March 7, 1999) was an American film director, writer, producer, and photographer who lived in England during most of the last four decades of his career. Kubrick was noted for the scrupulous care with which he chose his subjects, his slow method of working, the variety of genres he worked in, his technical perfectionism, and his reclusiveness about his films and personal life. He worked far beyond the confines of the Hollywood system, maintaining almost complete artistic control and making movies according to his

related topics: {film, series, show}, {theory, work, human}, {son, year, death}

related documents: Orson Welles, B movie, Mystery Science Theater 3000, Monty Python, Doctor Who, Sam Peckinpah, The A-Team, Pulp Fiction (film), Buffy the Vampire

{theory, work, human}

words	related documents	related topics
theory	Meme	{work, book, publish}
work	Intelligent design	{law, state, case}
human	Immanuel Kant	{son, year, death}
idea	Philosophy of mathematics	{woman, child, man}
term	History of science	{god, call, give}
study	Free will	{black, white, people}
view	Truth	{film, series, show}
science	Psychoanalysis	{war, force, army}
concept	Charles Peirce	{language, word, form}
form	Existentialism	{@card@, make, design}

Existentialism

Existentialism is a term applied to the work of a number of 19th and 20th-century philosophers who, despite profound doctrinal differences,[1][2] generally held that the focus of philosophical thought should be to deal with the conditions of existence of the individual person and his or her emotions, actions, responsibilities, and thoughts.[3][4] The early 19th century philosopher Søren Kierkegaard, posthumously regarded as the father of existentialism,[5][6] maintained that the individual is solely responsible for giving his or her own life

related topics: {theory, work, human}, {film, series, show}, {god, call, give}

related documents: Deconstruction, Charles Peirce, Social sciences, Idealism, Franz Boas, Karl Marx, Mysticism, Truth, Free will, Solipsism
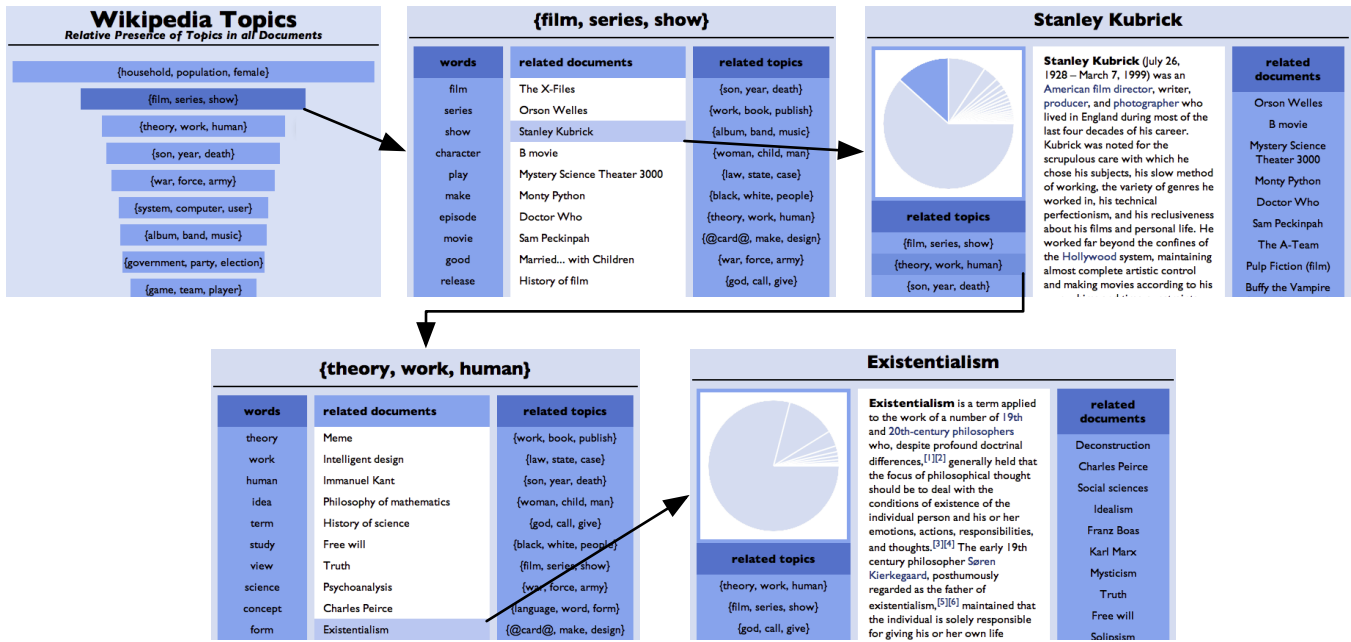
Figure 1: Navigating Wikipedia with a topic model. Beginning in the upper left, we see a set of topics, each of which is a theme discovered by a topic modeling algorithm. We click on a topic about film and television. We choose a document associated with this topic, which is the article about film director Stanley Kubrick. The page about this article includes its content and the topics that it is about. We explore a related topic about philosophy and psychology, and finally view a related article about Existentialism. This browsing structure—the themes and how the documents are organized according to them—is created by running a topic modeling algorithm on the raw text of Wikipedia and visualizing its output. This navigator can be found at http://bit.ly/wiki100.

---

1. For K topics, choose each topic distribution β_k.
   (Each β_k is a distribution over the vocabulary.)

2. For each document in the collection:

   (a) Choose a distribution over topics θ_d.
       (The variable θ_d is a distribution over K elements.)

   (b) For each word in the document

      i. Choose a topic assignment z_n from θ_d.
         (Each z_n is a number from 1 to K.)

      ii. Choose a word w_n from the topic distribution β_{z_n}.
         (Notation β_{z_n} selects the z_nth topic from step 1.)

Notice that the same set of topics $\{\beta_1, \ldots, \beta_K\}$ is used for every document, but that each document exhibits those topics with different proportions θ_d. This captures heterogeneity in documents: LDA can model that some articles are about "sports" and "business", others are about "sports" and "health", and that the topic of "sports" is similar in each. In statistics, this is known as a *mixed membership model*, where each data point exhibits multiple components.

Given a set of documents, the central problem for LDA is *posterior inference* of the variables, or determining the hidden thematic structure that best explains an observed collection of documents. This is akin to "reversing" the generative process to find the topics, topic proportions, and topic assignments that best explain the observed documents. Researchers have developed several algorithms for LDA inference, including Markov chain Monte Carlo sampling (Steyvers and Griffiths 2006) and optimization-based variational inference (Blei, Ng, and Jordan 2003). These algorithms take a collection of documents as input and return how the hidden topic variables decompose it.

Inference results in a set of interpretable topics $\{\beta_1, \ldots, \beta_K\}$ and a representation of each document of the collection according to those topics $\{\theta_1, \ldots, \theta_D\}$.[1] One of the main applications of topic models is for exploratory data analysis, that is, to help browse, understand, and summarize otherwise unstructured collections This is the application that motivates our work.

Visualizing a Topic Model

Our goals are to use the topic model to summarize the corpus, reveal the relationships between documents and the discovered summary, and reveal the relationships between the documents themselves. We applied our method to 100,000 Wikipedia articles, which we will use as a running example.

Our visualization uses both the observed data from a corpus and the inferred topic model variables.[2] The topic model variables are the topics β_k, each of which is a distribution over a vocabulary, and the topic proportions θ_d, one for each document and each of which is a distribution over the topics.

We use multiple views to illuminate this complex structure and created a basic navigator that fully represents a corpus through the lens of an LDA analysis. In this section, we explain our design choices.

---

[1]That the topics are interpretable is an empirical finding; the name "topic model" is retrospective. Topic models correlate to curated thesauri (Steyvers and Griffiths 2006) and to human judgements of interpretability (Chang et al. 2009).

[2]Note that the we use variables to indicate their posterior expectations. This is to make the notation simple.
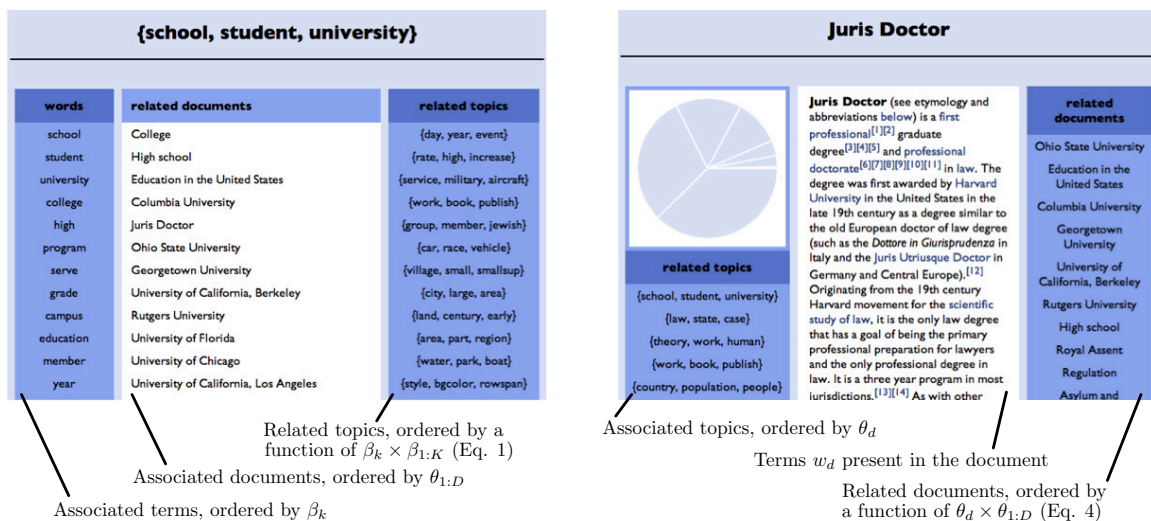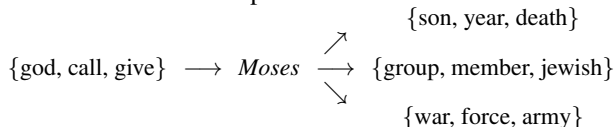
{school, student, university}		
words	**related documents**	**related topics**
school	College	{day, year, event}
student	High school	{rate, high, increase}
university	Education in the United States	{service, military, aircraft}
college	Columbia University	{work, book, publish}
high	Juris Doctor	{group, member, jewish}
program	Ohio State University	{car, race, vehicle}
serve	Georgetown University	{village, small, smallsup}
grade	University of California, Berkeley	{city, large, area}
campus	Rutgers University	{land, century, early}
education	University of Florida	{area, part, region}
member	University of Chicago	{water, park, boat}
year	University of California, Los Angeles	{style, bgcolor, rowspan}

Juris Doctor

Juris Doctor (see etymology and abbreviations below) is a first professional[1][2] graduate degree[3][4][5] and professional doctorate[6][7][8][9][10][11] in law. The degree was first awarded by Harvard University in the United States in the late 19th century as a degree similar to the old European doctor of law degree (such as the *Dottore in Giurisprudenza* in Italy and the Juris Utriusque Doctor in Germany and Central Europe).[12] Originating from the 19th century Harvard movement for the scientific study of law, it is the only law degree that has a goal of being the primary professional preparation for lawyers and the only professional degree in law. It is a three year program in most jurisdictions.[13][14] As with other

related topics
{school, student, university}
{law, state, case}
{theory, work, human}
{work, book, publish}
{country, population, people}

related documents
Ohio State University
Education in the United States
Columbia University
Georgetown University
University of California, Berkeley
Rutgers University
High school
Royal Assent
Regulation
Asylum and

Related topics, ordered by a function of $\beta_k \times \beta_{1:K}$ (Eq. 1)

Associated documents, ordered by $\theta_{1:D}$

Associated terms, ordered by β_k

Associated topics, ordered by θ_d

Terms w_d present in the document

Related documents, ordered by a function of $\theta_d \times \theta_{1:D}$ (Eq. 4)

Figure 2: A topic page and document page from the navigator of Wikipedia. We have labeled how we compute each component of these pages from the output of the topic modeling algorithm.

Visualizing the Elements of a Topic Model

The navigator has two main types of pages: one for displaying discovered topics and another for the documents. There are also overview pages, which illustrate the overall structure of the corpus; they are a launching point for browsing.

These pages display the corpus and the discovered structure. But this is not sufficient—we also use the topic model inference to find connections between these visualizations. With these connections, a user can move between summary and document-level presentations.

Hence, in our visualization every element on a page links a user to a new view. With these links, a user can easily traverse the network of relationships in a given corpus. For example, from a topic page a user can link to view a specific document. This document might link to several topics, each of which the user can explore:

{son, year, death}

{god, call, give} $\longrightarrow$ *Moses* $\longrightarrow$ {group, member, jewish}

{war, force, army}

We illustrated another navigation example in Figure 1.

An advantage of this design is that every type of relationship has a representation and an interaction. This illuminates the structure of corpus to a user and helps her navigate that structure. Further, any variable may occur in multiple views; all relationships are many-to-many.

Topic Pages Topics summarize the corpus. In the output of an inference algorithm, they are probability distributions over the vocabulary. But topics tend to be sparse, so a good representation is as a set of words that have high probability. Given such a set, users can often conceive meaning in a topic model (Chang et al. 2009). For example, one can intuitively glean from the three words {*film, series, show*} (Figure 1) that this topic is about film and television. We illustrate example topic pages in Figure 2. In these pages, the terms are represented as a list of words in the left column, ordered by their topic-term probability β_{kv}.

The center column of the view lists documents that exhibit the topic, ordered by inferred topic proportion θ_{dk}. Document titles links to the corresponding document pages. We can see that the list of documents related to {*school, student, university*} (Figure 2) are tied to education and academics:

College

{school, student, university} $\longrightarrow$ *Columbia University*

Juris Doctor
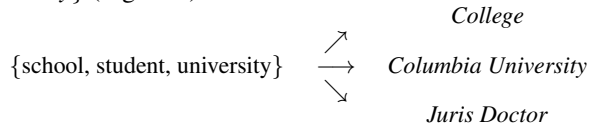
Finally, related topics are also listed with corresponding links, allowing a user to explore the high-level topic space. Topic similarity is not inferred directly with LDA, but can be computed from the topic distributions that it discovers. Related topics are shown in the right column of the topic page by pairwise topic dissimilarity score

$$\xi_{ij} = \sum_{v \in V} \mathbf{1}_{\mathbb{R}_{\neq 0}}(\beta_{iv}) \mathbf{1}_{\mathbb{R}_{\neq 0}}(\beta_{jv}) \left| \log(\beta_{iv}) - \log(\beta_{jv}) \right| \quad (1)$$

where the indicator function $\mathbf{1}_A(x)$ is defined as 1 if $x \in A$ and 0 otherwise. This is related to the average log odds ratio of the probability of each term in the two topics. This metric finds topics that have similar distributions.

Continuing with the topic from Figure 1, this metric scores the following topics highly.

{son, year, death}

{film, series, show} $\longrightarrow$ {work, book, publish}

{album, band, music}
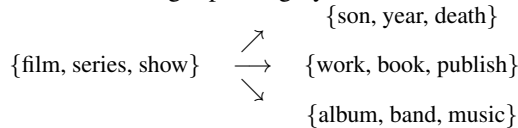
The original topic relates to film and television. The related topics cover a spectrum of concepts from other forms of media to human relationships and life events.

Document Pages Document pages render the original corpus. We show the document and the topics that it exhibits, ordered by their topic proportions θ_{dk}. We also display these topics in a pie chart, showing their respective proportions within the document. For example, Figure 2 illustrates that

Figure 3: Topic overviews from a visualization of Wikipedia (left), the New York Times (center), and US Federal cases (right). All of these navigators are online (see the *Implementation and Study* section).

the *Juris Doctor* article is a third about academia, a third about law, and a third about other topics. Each rendering of a topic links to its respective page.

Finally, documents are associated with similar documents. Document similarity is not inferred directly with LDA, but is defined by the topic proportions:

$$\sigma_{ij} = \sum_{k \in K} \mathbf{1}_{\mathbb{R}_{\neq 0}}(\theta_{ik}) \mathbf{1}_{\mathbb{R}_{\neq 0}}(\theta_{jk}) \left| \log(\theta_{ik}) - \log(\theta_{jk}) \right|. \quad (2)$$

This metric says that a document is similar to other documents that exhibit a similar combination of topics.

Overview Pages Overview pages are the entry points to exploring the corpus. In the simplest of these pages, we rank the topics by their relative presence in the corpus and display each in a bar with width proportional to the topic's presence score p_k: the sum of the topic proportions for a given topic over all documents,

$$p_k = \sum_{d \in D} \theta_{dk}. \quad (3)$$

Examples of this view can be found in Figure 3. From this figure, we see that many documents are related to the topic {*household, population, female*}; this is consistent with our observations of the corpus, which includes many Wikipedia articles on individual cities, towns, and townships.

Implementation and Study

We provide an open source implementation of the topic modeling visualization. There are three steps in applying our method to visualizing a corpus: (1) run LDA inference on the corpus to obtain posterior expectations of the latent variables (2) generate a database and (3) create the web pages to navigate the corpus.

Any open-source LDA package can be used; we used LDA-C.[3] We implemented the remainder of the pipeline in python. It can be found at http://code.google.com/p/tmve.

We created three examples of navigators using our visualization. We analyzed 100,000 Wikipedia articles with a 50-topic LDA model (http://bit.ly/wiki100). We analyzed 61,000 US Federal Cases[4] with a 30-topic model (http://bit.ly/case-demo). We analyzed 3,000 New York Times articles with a 20-topic model (http://bit.ly/nyt-demo). A page from each of these three demos can be seen in Figure 3. One week after we released the source code, we received links to a navigator of arXiv (a large archive of scientific preprints) that was generated using our code; it is at http://bit.ly/arxiv-demo.

Preliminary User Study We conducted a preliminary user study on seven individuals, asking for qualitative feedback on the Wikipedia navigator. The reviews were positive, all noting the value of presenting the high-level structure of a corpus with its low-level content. One reviewer felt it organized similar to how he thinks.

Six individuals responded that they discovered connections that would have remained obscure by using Wikipedia traditionally. For example, one user explored articles about economics and discovered countries with inflation or deflation problems of which he had previously been unaware.

Acknowledgements

References

Blei, D., and Lafferty, J. 2009. Topic models. In Srivastava, A., and Sahami, M., eds., *Text Mining: Theory and Applications*. Taylor and Francis.

Blei, D.; Ng, A.; and Jordan, M. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Cao, N.; Sun, J.; Lin, Y.-R.; Gotz, D.; Liu, S.; and Qu, H. 2010. FacetAtlas: Multifaceted Visualization for Rich Text Corpora. *IEEE Transactions on Visualization and Computer Graphics* 16(6):1172 – 1181.

Chang, J.; Boyd-Graber, J.; Wang, C.; Gerrish, S.; and Blei, D. 2009. Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Chen, Y.; Wang, L.; Dong, M.; and Hua, J. 2009. Exemplar-based visualization of large document corpus. *IEEE Transactions on Visualization and Computer Graphics* 15(6):1161–1168.

Gardener, M. J.; Lutes, J.; Lund, J.; Hansen, J.; Walker, D.; Ringger, E.; and Seppi, K. 2010. The topic browser: An interactive tool for browsing topic models. In *Proceedings of the Workshop on Challenges of Data Visualization (in conjunction with NIPS)*.

Gretarsson, B.; O'Donovan, J.; Bostandjiev, S.; Asuncion, A.; Newman, D.; Smyth, P.; and Hllerer, T. 2011. Topicnets: Visual analysis of large text corpora with topic modeling. In *ACM Transactions on Intelligent Systems and Technology*, 1–26.

Havre, S.; Hetzler, B.; and Nowell, L. 2000. Themeriver(tm): In search of trends, patterns, and relationships. In *Proc. IEEE Symposium on Information Visualization (InfoVis)*, 115–123.

Newman, D.; Asuncion, A.; Chemudugunta, C.; Kumar, V.; Smyth, P.; and Steyvers, M. 2006. Exploring large document collections using statistical topic models. In *KDD-2006 Demo Session*.

Steyvers, M., and Griffiths, T. 2006. Probabilistic topic models. In Landauer, T.; McNamara, D.; Dennis, S.; and Kintsch, W., eds., *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.

---

[3] http://www.cs.princeton.edu/ blei/lda-c

[4] http://www.infochimps.com/datasets/text-of-us-federal-cases