Structure and Trends in a Selection of Academic Literature

CS 229, Autumn 2012

Hannah Hironaka
hannahh@stanford.edu

Rahul Suri
rahul.suri@stanford.edu

ABSTRACT

We investigate the use of unsupervised machine learning methods for discovery of latent structure in unstructured, unlabeled text data. We briefly present background material on two such methods, k-means clustering and Latent Dirichlet Allocation for topic modeling. We then describe a dataset of 677 text documents that we assembled and results from evaluating the described methods on this dataset. Lastly, we conclude with some general observations.

1. INTRODUCTION

Supervised learning is a primary focus in much of machine learning; in the context of text document input, supervised methods have been brought to bear on a wide variety of tasks including sentiment analysis, spam classification, and authorship attribution.

However, it is not always the case that categorization of input documents into predetermined buckets is truly the task of interest. Indeed, the unsupervised setting is often a more appropriate representation for problems whose input consists of unstructured text documents. Here, the goal is instead to cluster together similar documents in the hope that the grouping of documents reveals some latent structure underlying the data.

Moreover, unstructured, unlabeled text data are ubiquitous; books, magazines, and newspapers are traditional sources. The explosion of data in recent years has made many more sources readily available: webpages, blog posts, emails, tweets, transcripts of audio recordings, etc. While each of these sources could be shoehorned into a supervised learning problem in one way or another, a direct reckoning of the data with unsupervised methods can be equally satisfying, and the resultant insights can be used to summarize the data or to inform further exploration of it. To the extent that the volume of data now available to us is overwhelming, such automated methods for summarizing large text corpora are invaluable.

K-means is a fairly straightforward algorithm for clustering points in a vector space that can be applied to text documents. Latent Dirichlet Allocation (LDA) [5] is a Bayesian hierarchical model for discovering *topic models* from collections of text documents, with numerous extensions of the standard LDA framework appearing over the past decade. In the next sections, we briefly discuss k-means and LDA.

2. ALGORITHMS

2.1 K-means Clustering

K-means clustering groups a set of input vectors into k clusters based on similarity. It starts by initializing k cluster mean vectors with random values; at each iteration, some distance measure is used to group each input vector with the cluster whose mean it is closest to. Cluster mean vectors are then recalculated based on their newly-assigned vectors, and the algorithm loops until convergence.

In our analysis we made several decisions about the construction of the algorithm and its inputs. Each input vector represented one document. The vectors are indexed by words in the vocabulary, with the index's corresponding value representing the frequency of that word in the document. We initialized the cluster means with random input vectors. We chose cosine distance as a similarity measure because the input document vectors were not normalized; this addresses the scenario where two input vectors have similar directions (i.e. similar relative frequencies of words) but not magnitude (i.e. different document lengths). Finally, we adapted the algorithm to handle the case of zero vectors grouped with a mean after an iteration: in this event, the algorithm reuses the mean from the previous iteration.

2.2 Latent Dirichlet Allocation

Intuition. LDA assumes that the data exhibit underlying *topics*. Each topic is a multinomial[1] distribution over the vocabulary; the term "topic" reflects the fact that the estimated multinomials tend to place probability mass on words with thematic coherence (*atom*, *mass*, and *particle* for a topic such as "physics," e.g.).

Given these topics, the words in a single document are assumed to be drawn from a mixture over the topics. This gives a generative view of the set of all documents, with each document having its own mixing proportions.

Notation. Let β_i for $i = 1, \ldots, K$ be a set of K topics. Let θ_d for $d = 1, \ldots, D$ be a set of D documents. Let $w_{d,n}$ denote the n^{th} word in the d^{th} document, and let $z_{d,n} = k$ be an indicator that $w_{d,n}$ was generated by the k^{th} topic. For convenience, assume that each document contains N words.

[1]Technically, all "multinomial" distributions referred to in this document are better described as "categorical" distributions. However, we follow the relevant literature's standard convention of calling them multinomials.

Lastly, let α and η be Dirichlet-distributed priors for θ_d and β_k, respectively.

Joint PDF. The joint density of the variables under the model is given by

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}|\alpha, \eta) = \prod_{i=1}^{K} p(\beta_i|\eta) \prod_{d=1}^{D} p(\theta_d|\alpha) \left(\prod_{n=1}^{N} p(z_{d,n}|\theta_d) p(w_{d,n}|\beta_{1:K}, z_{d,n}) \right).$$

Posterior. The posterior of the parameters given the data is

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}|w_{1:D,1:N}, \alpha, \eta) = \frac{p(\beta_{1:K}, \theta_{1:D}, z_{1:D,1:N}, w_{1:D,1:N}|\alpha, \eta)}{p(w_{1:D,1:N}|\alpha, \eta)},$$

where the numerator is the joint PDF and the denominator is the probability of the evidence.

The two main approaches for numerically approximating the posterior are Gibbs sampling [7, 12] and variational methods [10, 14].

Dynamic Topic Modeling and Extensions to LDA. The LDA hierarchical model as presented here has often been used as a building block for constructing more sophisticated approaches by relaxing its modeling assumptions [2, 8, 15, 13]. In particular, dynamic topic models [3] take inputs which include discrete-valued temporal data; intuitively, a topic model is learned for each timestep, and the topics themselves evolve from one timestep to the next as a random walk. Specifically, if we let $\beta_{t,i}$ denote the i^{th} topic at time t, then $\beta_{t,i}|\beta_{t-1,i} \sim \mathcal{N}(\beta_{t-1,i}, \sigma^2 I)$. Dynamic topic models thus allow us to track the evolution of topics over time.

3. DATA AND METHODOLOGY

3.1 Data

The text corpus for this project was downloaded from the CS 229 website containing students' final project reports [6]. It contains 677 documents from 2005 to 2011. After removing stopwords, numbers, and punctuation, the corpus contains occurrences of 47121 unique tokens. The median number of (non-unique) word occurrences per document is 1058, with an interquartile range of 515.

3.2 Methodology

To perform topic modeling analyses on the corpus described above, we used open-source topic modeling software [9] to stem the tokens and to shrink the vocabulary size from 47121 to 13562 by retaining only the highest-scoring tokens, as measured by their term frequency-inverse document frequency (tf.idf) scores [11]. The size of the shrunken vocabulary was selected via informal experimentation; the selected settings are those which appeared to reveal the most interesting patterns and which made execution of dynamic topic modeling manageable.

For k-means clustering, we used the open-source Natural Language Toolkit (NLTK) [1] for preprocessing: we standardized to lowercase; removed punctuation, non-alphanumeric characters, and stopwords; and lemmatized the remaining words. The resulting vocabulary size was 37059.

The results of these preprocessing steps produced document-term matrices, which were then used as input to our analyses.

4. RESULTS

4.1 K-means

After informal experimentation, we chose $k = 9$ as the parameter for the k-means algorithm; the clusters it produced seemed most coherent, and no clusters were empty. We ran k-means using each year as a separate dataset, in addition to running it once with the entire corpus as its input.

Table 1 displays results from the single-dataset analysis. Despite the small sample of titles from each cluster and the minimal detail about each one's full content, several patterns are evident. For example, cluster 1 groups papers related to game play, and cluster 4 groups object manipulation papers. Some clusters, such as the third one, remain incoherent, though.

Table 2 shows a few results from clustering within an individual year; several characteristics of the evolution of project topics are evident in the output. Object manipulation and robotics were common topics throughout all years, but were more prevalent in earlier years, as was image processing and detection. A cohesive computational biology cluster appears in 2010. News recommendation and Twitter sentiment analysis first appear in 2010 and 2011.

Both variations of k-means clustering suffer from several limitations. The small size of the corpus attenuated the "closeness" of the clusters around their means. This was most evident for 2005, 2006, and 2007 which had 69, 68, and 66 project papers, respectively – the smallest numbers. Clusters from these years were often incoherent or contained many outliers. Furthermore, running k-means multiple times produced noticeably different clusters, indicating that this dataset was particularly likely to cause the algorithm to converge to local optima.

4.2 LDA

The results shown below are for a model with 15 topics. As with choosing the exact size of the vocabulary, the number of topics was chosen based on informal experimentation (and on a rule of thumb suggesting roughly $\sqrt{N/2}$ clusters for a dataset with N points).

Table 3 shows results produced by LDA. Each topic is illustrated by a list of the ten most-probable words from that topic. We observe that topic 2 has mostly grouped terms relating to natural language processing together, topic 10 has grouped together terms about news recommendation projects, and topic 15 has grouped terms concerning projects about image processing. Similar conclusions can be stated for the remaining topics; audio processing, image processing, game play, robotics, bioinformatics, and stock market prediction emerge as some of the major underlying themes of CS 229 class projects. So, LDA has provided a way for automatically discovering latent thematic structure, grouping words from relevant subjects using only the unstructured source documents.

Cluster 1
Beat CAL: Machine Learning in Football Play-calling
MLB Prediction and Analysis
A Machine Learning Approach to Opponent Modeling in General Game Playing
Cluster 2
Applying Synthetic Images to Learning Grasping Orientation from Single Monocular Images
Image Processing for Bubble Detection in Microfluidics
Value Iteration and DDP for an Inverted Pendulum
Cluster 3
RoboChef: Automatic Recipe Generation "byte-sized recipes"
Person Following On STAIR
Travel Time Estimation Using Floating Car Data
Cluster 4
STAIR Subcomponent: Learning to Manipulate Objects from Simulated Images
Door Handle Detection for the Stanford AI Robot (STAIR)
Learning To Pick Up a Novel Object
Cluster 7
Supervised Learning - Stock Trend Classifier
Finding Optimal Hardware Configurations For C Code
Statistical Analysis and Application of Ensemble Method on the Netflix Challenge
Cluster 8
Clustering WordNet Senses Utilizing Modified and Novel Similarity Metrics
Classification of Amazon Reviews
Predicting Dow Jones Movement with Twitter
Cluster 9
CRF Based Point Cloud Segmentation
Clustering Autism Cases on Social Functioning
Whos in Charge Here? : Using Clustering Algorithms to Infer Association of Putative Regulatory Elements and Genes

Table 1: Representative documents for a few clusters, for a 9-cluster model.

4.3 Dynamic Topic Modeling

Figures 1 through 3 depict results from running dynamic topic modeling on our dataset. Each figure corresponds to a single topic from a 15-topic dynamic model. Lines are shown for the five tokens whose probability of appearing in that topic displayed the highest variance across timesteps; the lines plot each token's probability of occurring in that topic against time.

For topic 3 (Figure 1), the 10 tokens which appear with highest probability (marginalizing across timesteps) are user, cluster, movi, articl, recommend, feed, rmse, read, stori, and kmean. This information can be used to inform directed exploration of the original dataset to conclude that this topic has grouped together terms relating to recommender system projects. From the figure, we see signs of the diminishing popularity of Netflix-inspired movie recommendation projects (which were scored using rmse and often included some form of clustering, typically kmeans). Moreover, we see an increase in tokens related to news article recommendation

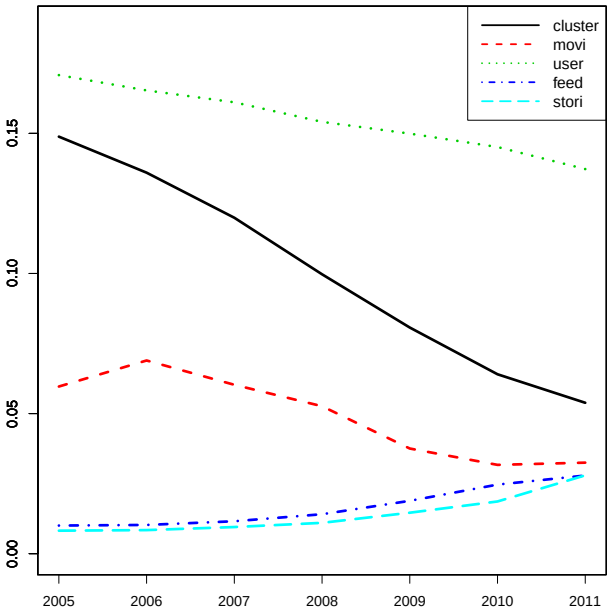

Figure 1: Word probabilities in topic 3 across time.

projects starting in 2009, in which students used data from Pulse to suggest news stories to users in their news feeds.

For topic 4 (Figure 2), the 10 tokens which appear with highest probability are stock, price, day, market, trade, tweet, portfolio, forecast, twitter, and network; the thematic content of this topic is rather obvious. Encouragingly, the figure clearly depicts the sharp increase in mentions of tweets and twitter beginning in 2009, as many projects have used Twitter data as a springboard for stock prediction. Moreover, the figure also reflects an increase in terms like stock and price. If we were looking at these graphs to try to understand latent structure in the dataset without any prior knowledge, Figure 2 would have alerted us to one of the major trending themes in an entirely automated manner.

Lastly, we consider topic 9 (Figure 3), whose top 10 tokens are layer, imag, network, video, deep, frame, fig, roi, templat, and reconstruct. It appears that this topic groups together terms related to projects on neural networks. The figure seems to indicate an upward trend in use of neural networks, which is consonant with the recent increase in popularity of deep learning with neural networks for image analysis.

5. DISCUSSION

One of the unexpected challenges that arose with these analyses was that of simple data manipulation; scraping text from PDF documents often introduced strange encoding errors that negatively impacted results and took time to hunt down.

Additionally, identical tokens sometimes appeared in distinct contexts: light appeared both in the context of traffic light control policies and also in image processing; k-means clustering was particularly susceptible to conflating the two meanings. Moreover, the stemming we used to preprocess words at times caused otherwise distinct tokens to be con-

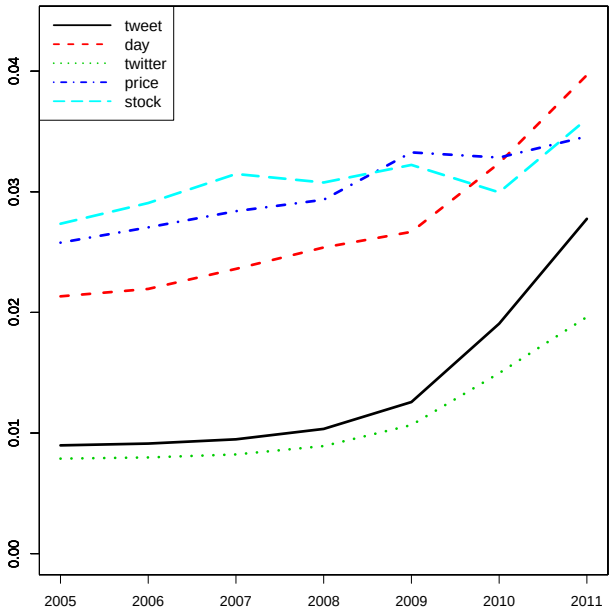

Figure 2: Word probabilities in topic 4 across time.

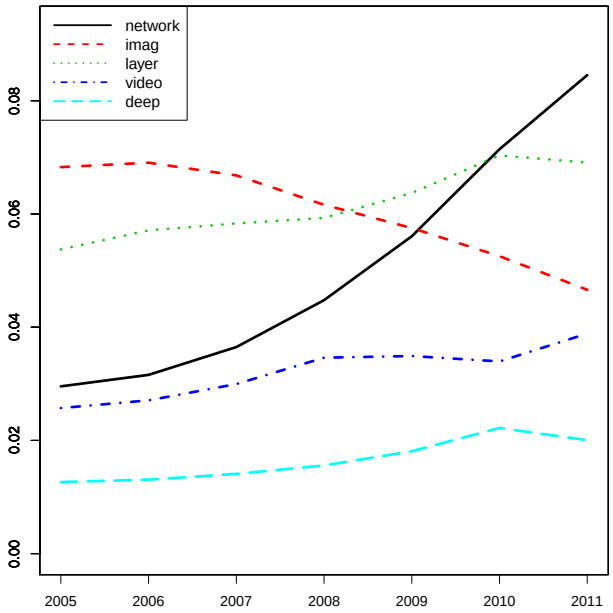

Figure 3: Word probabilities in topic 9 across time.

flated – movie and moving were both stemmed to movi even though the former concerned film recommendations while the latter concerned video analysis.

Lastly, we note that most algorithmic parameters (k for k-means, tf.idf cutoffs) were chosen via informal experimentation. While more rigorous approaches to selecting these values (cross validation, e.g.) could have been used, we decided to focus our efforts elsewhere, as we suspected that the marginal gains to be achieved from optimizing these values would be relatively modest, particularly when the ultimate evaluation of the results was largely qualitative.

6. CONCLUSIONS

We have presented k-means clustering, latent Dirichlet allocation, and dynamic topic modeling, which are three approaches to unsupervised learning from unlabeled data. We applied these methods to a novel collection of text documents and showed how the automatic discovery of latent structure from free text can be used to highlight interesting patterns and trends in the data.

7. REFERENCES

[1] S. Bird. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[2] D. M. Blei. Introduction to probabilistic topic models. *Communications of the ACM*, 2011.

[3] D. M. Blei and J. D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 113–120, New York, NY, USA, 2006. ACM.

[4] D. M. Blei and J. D. Lafferty. *Topic Models*. CRC Press, 2009.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[6] CS 229 final project reports. `http://cs229.stanford.edu`, 2005-2011.

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

[8] T. L. Griffiths, M. Steyvers, D. M. Blei, and J. B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, 2005.

[9] B. Grun and K. Hornik. topicmodels: An r package for fitting topic models. *Journal of Statistical Software*, 40(13):1–30, 5 2011.

[10] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999. 10.1023/A:1007665907178.

[11] C. Manning and H. Schütze. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

[12] D. Mimno, H. Wallach, and A. McCallum. Gibbs Sampling for Logistic Normal Topic Models with Graph-Based Priors. In *NIPS Workshop on Analyzing Graphs, 2008*, 2008.

[13] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[14] M. J. Wainwright and M. I. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Now Publishers Inc., Hanover, MA, USA, 2008.

[15] H. M. Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 977–984, New York, NY, USA, 2006. ACM.

2005 – Cluster 1
STAIR Subcomponent: Learning to Manipulate Objects from Simulated Images
Door Handle Detection for the Stanford AI Robot (STAIR)
Re-Learning to Walk: Adding Force Feedback Control to the Quadruped Robot
2005 – Cluster 2
Learning Depth in Light Field Images
Explicit Image Filter
Learning Traffic Light Control Policies
2005 – Cluster 3
Splice Site Prediction using Multiple Sequence Alignment
CS229 Project: Musical Alignment Discovery
Prostate Detection Using Principal Component Analysis
2009 – Cluster 1
Automatic graph classification
Automatic Fatigue Detection
Automatic Beat Alignment
2009 – Cluster 2
Discovering Visual Hierarchy through Unsupervised Learning
Spoken Language Identification With Hierarchical Temporal Memories
2009 – Cluster 3
Learning to splash
Stock Forecasting using Hidden Markov Processes
Recognizing Informed Option Trading
2011 – Cluster 1
Complex Sentiment Analysis using Recursive Autoencoders
Predicting Rating with Sentiment Analysis
Sentiment Based Model for Reputation Systems in Amazon
2011 – Cluster 2
Pulse News Preference Prediction
Predicting Preferences: Analyzing Reading Behavior and News Preferences
Reddit Recommendation System
Pulse Project: User-Interest-based News Prediction
2011 – Cluster 3
Attentional Based Multiple-Object Tracking
Learning Unsupervised Features from Objects
Unsupervised Learning Of Temporally Coherent Features For Action Recognition

Table 2: Selected documents and clusters for clustering by years.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
music	word	game	song	day
instrument	document	player	audio	stock
signal	queri	team	music	price
fig	cluster	win	pitch	tweet
mixtur	review	agent	speech	market
genr	sentenc	oppon	mfcc	trade
facial	corpus	games	energi	twitter
autoencod	topic	action	network	node
ica	charact	outcom	accent	word
spectral	semant	bet	ppca	network
Topic 6	**Topic 7**	**Topic 8**	**Topic 9**	**Topic 10**
imag	user	network	robot	user
brain	movi	beat	gestur	cluster
patient	cluster	signal	action	articl
roi	student	neuron	sensor	feed
voxel	rmse	onset	aircraft	stori
student	friend	snps	finger	read
fmri	social	fig	reward	word
pixel	emot	music	mous	recommend
neuron	kmean	patent	flight	day
tumor	recommend	diseas	movement	news
Topic 11	**Topic 12**	**Topic 13**	**Topic 14**	**Topic 15**
cluster	tag	network	robot	imag
gene	price	layer	pca	pixel
cell	portfolio	node	cluster	edg
kmean	trade	car	topic	video
signal	stock	sensor	entiti	depth
transcript	market	lane	terrain	segment
protein	econom	robot	motion	frame
rna	compani	signal	lda	cluster
damag	asset	cascad	channel	camera
genes	risk	polici	gpr	patch

Table 3: Most probable words for each topic, for a 15-topic topic model.