

Measuring Correlation Between Linguists' Judgments and Latent Dirichlet Allocation Topics

Ari Chanen and Jon Patrick

School of Information Technologies

University of Sydney

Sydney, Australia, 2007

{ari, jonpat}@it.usyd.edu.au

Abstract

Data that has been annotated by linguists is often considered a gold standard on many tasks in the NLP field. However, linguists are expensive so researchers seek automatic techniques that correlate well with human performance. Linguists working on the ScamSeek project were given the task of deciding how many and which document classes existed in this previously unseen corpus. This paper investigates whether the document classes identified by the linguists correlate significantly with Latent Dirichlet Allocation (LDA) topics induced from that corpus. Monte-Carlo simulation is used to measure the statistical significance of the correlation between LDA models and the linguists' characterisations. In experiments, more than 90% of the linguists' classes met the level required to declare the correlation between linguistic insights and LDA models is significant. These results help verify the usefulness of the LDA model in NLP and are a first step in showing that the LDA model can replace the efforts of linguists in certain tasks like subdividing a corpus into classes.

1 Introduction

Since linguists are expensive to employ, there is a preference in most NLP projects to use automatic processes especially where it can be shown that the automatic process approaches the performance of the linguists. Several linguists were used on the ScamSeek project (Patrick, 2006). ScamSeek was created for the Australian Securities and Investments

Commission (ASIC) government agency to identify financial scam websites based on the linguistic properties of the webpage content. A major task they performed by the project linguists was to partition the corpus into classes. Besides defining the classes in terms of the documents assigned to them, the linguists also identified phrases they believed were indicative of each class.

The LDA corpus model (Blei, 2004) can automatically generate a likely set of corpus topics and subdivide the corpus words among those topics. We will show that there are similarities between the task the LDA performs and the tasks the ScamSeek linguists performed. This paper attempts to determine to what degree LDA topics correlate with the judgments of linguists in partitioning a corpus into document classes.

Formally, we set a null hypothesis, H_0 , to claim that the relationship between the linguists' document classes and LDA topics is random. The alternative hypothesis, H_a , claims those document classes and the topics have a significant amount of correspondence or correlation between them. In order to measure how significant the correlation is, principled methods of measuring the statistical significance of the correlation must be found. If the p -value for the correlation between a document class and the **best** correlating topic for that class is less than $\alpha = 0.05$, then H_0 will be rejected in favor of H_a . The determination of the p -values are discussed in the Methods section.

2 Background

2.1 LDA Model

The LDA is a Bayesian, generative corpus model which posits a corpus wide set of k topics from which the words of each document are generated. In this model, a topic is a multinomial distribution over terms. According to the LDA model, an author first determines, through a random process, the topic proportions of a new document. Thereafter, the author chooses a topic for the next word and then draws that word randomly according to the chosen topic distribution.

The LDA model can be represented as a graphical model as shown in figure 1. Graphical models represent the dependencies between probabilistic model hyper-parameters and variables. A good introduction can be found in (Buntine, 1995). The LDA model includes two hyper-parameters, α and β as well as three random variables (RV's), $\theta_{1:D}$, z and w , where D is the number of corpus of documents.

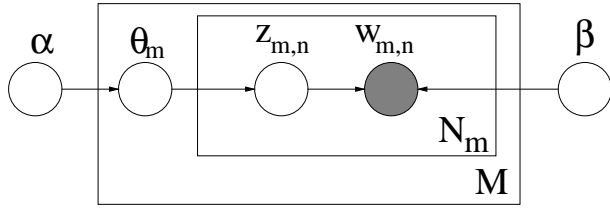


Figure 1: The LDA graphical model

α takes a scalar value that affects the amount of smoothing of the symmetric Dirichlet (*dir*) distribution that produces the multinomial (*multi*) distributed θ_m , representing the topic proportions for document m . The hyper-parameter β is a $k \times V$ matrix of probabilities where V is the size of the corpus vocabulary. Each row of β is a topic multinomial where $\beta_{ij} = p(w = j | z = i)$. The RV z is an index variable that indicates which topic was chosen for each document word (Steyvers and Griffiths, 2005)(Blei, 2004).

Formally, each document m is assumed to be formed by the following generative steps:

1. Choose proportions $\theta_m | \alpha \sim Dir(\alpha)$.
2. For $n \in \{1, \dots, N_m\}$:
 - (a) Choose topic $z_{m,n} \sim Multi(\theta_m)$

- (b) Choose word $w_{m,n}$ from $p(w_{m,n} | z_{m,n}, \beta_{z_{m,n}})$

where N_m is the number of words in document m .

Under graphical model notation, shaded elements are observed and unshaded elements are latent. Thus, the circle denoting the w element, representing the words of a document, is the only observed element. The other elements are latent. In order for the LDA model to be useful in practical settings, these latent RV's and hyper-parameters need to be estimated.

If α and β are assumed fixed, then the posterior probability w.r.t. θ and z can be expressed as follows:

$$\begin{aligned} p(\theta, \mathbf{z} | \mathbf{w}, \alpha, \beta) &= \frac{p(\mathbf{w} | \theta, \mathbf{z}, \alpha, \beta) p(\theta, \mathbf{z} | \alpha, \beta)}{p(\mathbf{w} | \alpha, \beta)} \\ &= \frac{p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta)}{\int_{\theta} \sum_{\mathbf{z}} p(\theta, \mathbf{z}, \mathbf{w} | \alpha, \beta) d\theta} \end{aligned}$$

Unfortunately, this posterior probability is intractable to calculate due to the integral over the Dirichlet variable. There are several methods for approximating θ and z . The LDA topic data used in this research was induced using the mean field variational method which is an iterative algorithm that converges on estimates of θ and z for each document and each word in those documents. Once these estimates have been obtained, then estimates for α and β can be obtained by holding the values of θ and z fixed and using an empirical bayes estimation technique. By alternating between the mean field variational estimation and the empirical bayes estimate the values of the latent elements are guaranteed to eventually converge to stable values. For further details on this latent element estimation technique see (Blei, 2004).

In the experiment section the topic proportions $\theta_{1:D}$ of each document and the topic rows of β will be compared to similar data produced by linguists.

Table 1 shows the 25 top terms from four sample topics induced from the ScamSeek corpus for a 64 topic model. The top terms are constructed by sorting a topic's multinomial terms by term probability in descending order. The first row of the table shows the name of the linguists' document class that

is most correlated¹ with the topic terms shown in the rows below. The last row shows the cumulative probability mass that the top 25 topic words account for. Three of these example topics are most associated with scam classes. Only the topic most associated with the **Licensed Operator** class is a non-scam class. A good indicator of this is that the word, “risk”, is one of the most probable terms.

Nigerian scam	Mail scam	Licensed operator	Online betting
i	you	investment	online
my	i	investments	casino
you	your	you	betting
your	will	invest	gambling
me	name	your	casinos
am	post	risk	games
all	money	investing	sport
will	now	funds	vegas
we	make	can	las
would	list	returns	odds
not	newsgroups	investors	sportsbook
thanks	only	shares	free
thank	just	their	sports
course	if	fund	internet
one	my	return	betted
work	message	over	best
some	all	australian	book
money	article	more	wagering
good	step	managed-funds	guide
just	more	portfolio	sports-betting
now	made	not	gaming
well	me	property	football
time	letter	investor	line
great	people	cash	your
0.31	0.30	0.27	0.41

Table 1: The 25 top terms from four sample topics induced from the ScamSeek corpus for a 64 topic model.

2.2 Monte-Carlo Simulation

In this research, we want to measure the strength of the correlations between classes and topics. One challenge of this task is that the classes and the topics are in different forms and the topics are non-parametric distributions. We achieve this aim by utilizing one form of the Monte Carlo Simulation method where a number of random pseudo-LDA models are produced. The correlations between the linguists classes and both the real LDA model as

¹The correlation measure used to determine the most correlated class is the distributional intersection (DI) measure which is described later in the methods section.

well as the pseudo-models are measured. The correlation scores between all the pseudo-models and the linguists classes are sorted and the real model’s correlation score is ranked against the pseudo-models. The percentage of pseudo-model scores that the real model score beats is taken to be the significance level of the real correlation.

Let the correlation between the classes and the LDA topics be called the **real** correlation. From the ranking of the real correlation within all the random correlations an approximate p -value is derived. Let r be the number of random correlations that are the same or better than the real correlation and let n be the number of random models. Then²:

$$p\text{-value} = r/n$$

(B. V. North and Sham, 2002) report that using Monte-Carlo procedures to calculate empirical p -values has become commonplace in statistical analysis and give three major motivating factors:

1. Many test statistics do not have a standard asymptotic distribution.
2. Even if such a distribution does exist, it may not be reliable in realistic sample sizes.
3. Calculation of the exact sampling distribution through exhaustive enumeration of all possible samples may be too computationally intensive.

Reason #1 definitely applies to the case of trying to find a distribution for possible LDA models. The LDA estimation algorithm is nonparametric itself so there is no reason to think it would produce topic multinomials that fit a parametric distribution. Reason #2 does not apply. Reason #3 is a major factor for using Monte-Carlo techniques in the case of this research. Each randomised topic has $N = 18,000$ terms. To randomise a LDA model each topic has its terms and probabilities shuffled in a pseudorandom manner. There are $N!$ different shuffles for each topic which is for all practical purposes infinite in this case.

²There is some dispute as to whether r/n or $(r+1)/(n+1)$ is the better p -value estimator. (Ewens, 2003) and (Broman and Caffo, 2003) prove that $(r+1)/(n+1)$ is biased so we use r/n here.

3 Similarities and differences between document classes and LDA topics

The LDA generative corpus model assumes that every corpus document draws its terms from κ topics, where κ is a parameter of the LDA model. One of the products of the LDA model estimation process is a γ -vector for each document which gives the estimated distribution of a document’s terms over the topics. Normalizing this vector by dividing by the total number of document terms gives the document topic proportions which is the same information that the LDA model’s θ_m RV represents for a given document m .

Unlike topics, the document classes the linguists constructed are meant to be mutually exclusive; a document may belong to one and only one of those classes. Although this is a significant difference between topics and these document classes, in practice the two are not too dissimilar. An analysis of all the normalised γ -vectors shows that, on average, each document devotes around 60% of its terms to a major topic, and allocates between 4-20% of its remaining content to each of four or five minor topics, leaving only small amounts of the topic mass to the rest of the topics. This pattern seems to hold irrespective of the number of topics used to generate the LDA model, as table 2 shows. Since most documents have a single topic with more than a majority of the topic mass, we will assume that topics can approximate the behavior of document classes.

Topics	Topic rank						
	1st	2nd	3rd	4th	5th	6th	7th
8	61.0	21.7	10.0	4.5	1.8	0.7	0.2
16	58.8	20.7	9.8	5.0	2.7	1.4	0.7
32	55.4	19.7	10.0	5.8	3.5	2.1	1.3
64	57.5	17.8	9.2	5.6	3.5	2.2	1.4
128	61.6	16.8	8.2	4.7	2.9	1.8	1.1
256	69.1	14.7	6.6	3.7	2.2	1.3	0.8
mean	60.6	18.6	9.0	4.9	2.8	1.6	0.9

Table 2: The average percentage of the 7 top ranked topics from each document in six different LDA models.

In addition to creating document classes, the linguists also created **motif** classes to embody certain qualities of documents that transcend the document classes. In this way, the motifs are closer to topics than document classes. The linguists identified char-

acteristic phrases for the motif classes just as they did for the document classes. An example of a motif class is one called the **persuasion** class which has indicative phrases that are common to many scams in which a scammer tries to persuade victims to do something. Many of the scam documents exhibit some of these persuasion phrases. Unfortunately, exact phrases cannot be revealed because parts of the ScamSeek project are proprietary.

For the remainder of the paper, the term **classes** will be used to signify both document classes and motif classes.

4 Methods

Two types of methods were employed to estimate a p -value for the correlation between the linguists classes and the LDA topics: categorical and term-based. The categorical method attempts to measure the randomness in the relationship between the topics and the linguists’ document classes. The term-based methods measure correlations between word distributions in the LDA topics and the linguists’ class characteristic phrases.

LDA models were generated on 1917 documents from the ScamSeek corpus. Eight models were induced with the following numbers of topics: 2, 4, 8, 16, 64, 128, 256. These models are referred to as the “real” models to differentiate them from the random LDA models introduced below.

4.1 Using the χ^2 test

The χ^2 test (Devore, 1999) can be used to test if two categorical variables are statistically independent. A contingency table is used to show the counts of some entity for every possible pairing of categories, one from each of the two variables. The empirical counts are compared to the counts that would be expected if the two variables were independent.

The χ^2 experiments described in this section only utilise the document classes and not the motif classes.

The raw LDA γ -vectors give a document’s term count for each topic therefore topics are categorical in this context. To make a document class into a categorical variable, the γ -vectors for all the documents in the same document class can be summed so that each cell contains the total term count for one topic

over all the documents in that class. Then, each cell (i, j) of the χ^2 contingency table will hold the total number of words from document class i that were assigned to topic j .

There is one problem with using the χ^2 test in this setting. Completely correct usage of the χ^2 test requires that each joint event from the contingency table is independent of all the others. However, according to (Blei, 2004, pg. 20), under LDA, the terms of the document are exchangeable, meaning that their order does not matter. This implies the terms are not independent of each other but rather *conditionally independent* with respect to the latent topics. Because of this potential problem, any results must be viewed with some caution.

The χ^2 statistic was calculated using each of the eight LDA models to determine the relationship between the document classes and the topics. These tests all indicated that the relationship was highly significant with a p -value of zero.

To verify this result, control experiments were performed where 10 random test sets were generated by shuffling the documents assigned to each class. The χ^2 test was run on each of the randomised sets. For the random sets, the χ^2 statistic was much lower than the value obtained from the real class assignments. Unexpectedly, the calculated p -value was still zero, indicating that even the randomised tests were highly significant.

We concluded that this method of applying the χ^2 test was not appropriate for the task of rejecting H_0 , and that the most likely reason is that the document words are not completely independent.

4.2 Using Monte-Carlo Simulation

Next, we turn to a term-based method of trying to verify the H_a hypothesis, using word distribution correlations between topics and classes rather than a categorical analysis. To test this hypothesis Monte-Carlo simulation was used as described in the Background section 2.2. Further details are provided in there section.

Again in this method, an approximate p -value is calculated from the ranking of real correlations within a sorted list of pseudo-correlations. The real correlations are between the words of the linguists' class characteristic phrases and real LDA topics while the pseudo-correlations are between

those phrase words and a set of randomly generated pseudo-topics.

4.2.1 Forming the random LDA models

To begin with, for each of the eight real models (models with 2, 4, 8, 16, 32, 64, 128, 256 topics), one hundred randomised models were generated. Real LDA models have topics that concentrate most of their probability mass on a relatively small number of terms compared to the total number of terms in the distribution. The method of randomization was chosen so as to maintain the same level of probabilistic "clumpiness" in the random topics. To form a pseudo-random LDA model from a real model, for each real topic, the terms and their probabilities are separated. To form a pseudo-topic, the terms are shuffled and assigned to one of the pre-existing multinomial probabilities from the real model's corresponding topic.

4.2.2 Correlating one class with one LDA topic

Again, we are trying to rank the **best** correlation of a real topic with a class among the correlations of that class with the best correlations among all the pseudo-topics in each randomised LDA model. This section defines some notation needed in discussing these class/topic correlations. This notation assumes a specific model (defined by the number of topics) and a specific correlation measure have been chosen. Different kinds of correlation measures will be explained below.

Below, classes are referred to with the index i . Topics are referred to with the index k . An index of r refers to the one real model while an integer index j refers to one of the 100 random models.

In our notation, C_{irk} , refers to the correlation of the i th class and real model's k th topic and C_{ijk} refers to the correlation of the i th class and j th random model's k th topic.

In order to obtain the p -value for each class, correlation measures are calculated for each pairing of class and topic, both real and random. First C_{irk} is calculated for the one real model. Next, C_{ijk} is calculated for each of the hundred random LDA models. The real topic that shows the best correlation score with class i is, \widehat{C}_{ir} . Next, the procedure is performed on each of the 100 random LDA models so a correlation C_{ijk} between the class and each

pseudo-topic k in each random model j is calculated. The best correlation for each random model \widehat{C}_{ij} is found. The best correlations for each random model are sorted from least correlated to most correlated. Then the rank of the best real topic correlation is found within the sorted list of random best correlations. Since our criteria for significance is $\alpha = 0.05$ then for a given number of topics, type of correlation measure and class i , if:

$$\widehat{C}_{ir} > \widehat{C}_{ij}$$

for 95 of the 100 random models then we would take this as sufficient evidence that the null hypothesis can be rejected in favor of the alternative hypothesis.

The following subsections first define a method for forming multinomial distributions from class indicative phrase and next specifies three correlation measures defined on two multinomial distributions over the same range of terms.

4.2.3 A distribution from class phrases

The LDA topics are multinomial distributions over 18,000 terms. One way to correlate a class with these topics is to form a multinomial distribution from the class. The phrases that the linguists generated as being characteristic of the class can be used to achieve this goal. All the phrases are treated as though they came from a single document and processed in the same way the corpus documents were processed before the LDA models were built from them. This means joining terms together into multiword expressions (MWE) where appropriate and eliminating stopwords. Next, a histogram is formed with the terms and MWE's as elements. Finally, the count for each element is normalised by the total number of elements, thus yielding a probability distribution.

4.2.4 The vector cosine correlation measure

Now that a distribution, \mathbb{C}_i , has been formed for each class i , we can correlate them with each topic distribution, \mathbb{T}_k . One way to do this is by treating the two distributions as vectors in term space. The cosine of the angle between these two vectors can be seen as a measure of how similar the two distributions are. If the angle is zero then the two distributions are the same whereas if they are perpendicular

they are maximally dissimilar. The cosine of the angle, θ , between \mathbb{C}_i and \mathbb{T}_k can be gotten from the formula:

$$\cos \theta = \frac{\mathbb{C} \cdot \mathbb{T}}{\|\mathbb{C}\| \|\mathbb{T}\|}$$

This measure will vary in the range $[0, 1]$ where 1 indicates the two distributions are identical.

4.2.5 The hypergeometric distribution correlation measure

The hypergeometric distribution (HD) (Devore, 1999, pg. 122) is often associated in with the probability of drawing lottery numbers that match the winning numbers. In the way the HD is used here, the winning lotto numbers are analogous to the words of the class indicated phrases and the most probable terms in a topic are analogous to the numbers on the lotto ticket. The HD assumes the following:

1. There is a population of size N to be sampled from.
2. Each member of the population can either be a *success* or a *failure*. There are M successes in the population.
3. A sample of size n is drawn in an independent and identically distributed manner.

$N = 18,000$ is the total number of terms in both the class and topic multinomial distributions. For a given class, \mathbb{C} , a term is defined as a *success* if it matches one of the terms from the class characteristic phrases, for a total of M possible successes. \mathbb{C}_M denotes the set of those success terms. Now, given the k th topic \mathbb{T}_k , $\mathbb{T}_{k,\hat{M}}$ is the set of the M most probable terms in that topic. Let \mathbb{I}_k be the number of elements in the intersection set $\mathbb{C}_M \cap \mathbb{T}_{k,\hat{M}}$. The probability of \mathbb{I}_k , for the subset of hypergeometric distributions where $n = M$ is:

$$P(\mathbb{I}_k | N, M) = \frac{\binom{M}{\mathbb{I}_k} \binom{N-M}{M-\mathbb{I}_k}}{\binom{N}{M}}$$

The lower the above probability is the greater the chance of correlation between \mathbb{C}_i and \mathbb{T}_k . Since this probability can be extremely small, log probabilities are used to express it. Therefore, the range of this correlation measure is $(-\infty, 0)$.

4.2.6 The distribution intersection correlation measure

Another simpler measure of distribution correlation is the amount of probability mass the two distributions share. The formula for calculating this measure for the distributions of the i th class \mathbb{C}_i and the k th topic \mathbb{T}_k is:

$$DI(\mathbb{C}_i, \mathbb{T}_k) = \sum_{j=1}^N \min(\mathbb{C}_i[j], \mathbb{T}_k[j])$$

where DI stands for “distribution intersection”, N is the total number of terms in each distribution. This measure also has the range of $[0, 1]$ with 1 meaning the two distributions are the same.

5 Results

5.1 Results for using 100 random models

To reiterate the problem definition, we seek to determine if there is enough evidence to reject the null hypothesis in favor of the alternative hypothesis. Since we set $\alpha = 0.05$, this means that the real model must have a better correlation should score then 95% of the random models, for a given model and type of correlation measure. The final performance results are measured in terms of the percentage of classes where the H_0 could be rejected. In many cases, the real model did better than all 100 of the pseudo-models so results are also provided for the case where we had set our H_0 rejection threshold to $\alpha = 0.01$.

Three different correlation measures were used: vector cosine (VC), distribution intersection (DI), and hypergeometric distribution (HD). Table 3 shows the results for the models of various numbers of topics and for the three correlation measures. The table gives the percentage of classes that have p -values less than 0.05 and 0.01.

For the DI correlation measure, there was enough evidence to reject H_0 at $\alpha = 0.05$ for comfortably over 90% of the classes for all eight LDA models classes and this was nearly true at $\alpha = 0.01$.

The results for the VC correlation measure are less significant where only five out of eight of the models could claim to reject H_0 for more than 90% of the classes for $\alpha = 0.05$. Also, the correlation level fell off for the models with higher num-

Topics	Vector Cosine		Distrib. Intersection		Hypergeometric	
	%<0.01	%<0.05	%<0.01	%<0.05	%<0.01	%<0.05
2	79.6	91.8	89.8	100	83.7	87.8
4	77.6	95.9	91.8	100	85.7	91.8
8	79.6	95.9	91.8	95.9	85.7	91.8
16	77.6	93.9	91.8	95.9	85.7	87.8
32	73.5	91.8	93.9	95.9	85.7	87.8
64	63.3	85.7	89.8	93.9	91.8	93.9
128	61.2	79.6	91.8	98	91.8	95.9
256	69.4	75.5	87.8	93.9	98	98

Table 3: The %’s of classes having $p < 0.01$ and $p < 0.05$ for 3 different correlation measures using 100 random LDA models for the Monte-Carlo simulation .

bers of topics (64, 128, 256) for $\alpha = 0.05$ and there was much larger gap between the correlations at $\alpha = 0.05$ and $\alpha = 0.01$ compared to the much smaller gap for the DI results. One problem with the VC measure is that the angle between the problem with \mathbb{C}_i and \mathbb{T}_k vectors is only measuring differences in the terms that have nonzero probabilities. Therefore, this measure is less restrictive, allowing for a greater chance that a random topic may have the right combination of terms so that its correlation with a class will be better than the corresponding real model’s best correlation.

The HP measure was the worst that $\alpha = 0.05$ but in the middle for $\alpha = 0.01$. one interesting trend is that it does much better then the VC measure for high topic models (128, 256.)

The DI correlation measure shows the generally higher correlation scores which does not necessarily mean it is the best measure for our purpose. Yet, it is a straightforward measure of the correlation between two distributions and it is the most straightforward to calculate.

5.2 Results for using 1000 random models

The evidence that LDA topics may mirror certain parts of linguistic instincts looks fairly convincing from the tests using 100 random LDA models. To add weight to these results more Monte-Carlo simulations were run using 1000 completely different random LDA models. The results are shown in table 4.

Notice that the column reporting the results for the DI correlation measurement and with $\alpha = 0.05$, has the exact same values as those for hundred

Topics	Vector Cosine		Distrib. Intersection		Hypergeometric	
	%<0.01	%<0.05	%<0.01	%<0.05	%<0.01	%<0.05
2	85.7	93.9	93.9	100	85.7	91.8
4	81.6	95.9	91.8	100	85.7	91.8
8	81.6	93.9	93.9	95.9	87.8	91.8
16	79.6	93.9	91.8	95.9	87.8	89.8
32	77.6	91.8	91.8	95.9	87.8	87.8
64	73.5	85.7	91.8	93.9	93.9	93.9

Table 4: The %’s of classes having $p < 0.01$ and $p < 0.05$ for 3 different correlation measures using 1000 random LDA models for the Monte-Carlo simulation .

model simulation in table 4.³ If average the percentages for the 2,4, 8,16 ,32 and 64 topic models for the hundred and thousand models test for each column from tables 3 and 4 then three of the columns are exactly the same and two have a change of 1% are less. That the change from the hundred model simulation to the thousand model simulation was minimal is a good sign that this technique of measuring the correlation is stable and adds weight to its validity.

6 Conclusion

Real LDA models and the judgments of the linguists in classifying the corpus do appear to be significantly well correlated when compared to random LDA models.

The distribution intersection correlation is used successfully here as a simple yet effective way of measuring the correspondence between the phrases that the linguists came up with to characterise classes and the words of the topics. The hypergeometric distribution and vector cosine correlation measures also showed significant correlation strengths but to a lesser degree than the DI measure.

The results reported on here should add to the confidence of the NLP field that the LDA corpus model, even though it is only an approximate statistical model, can correspond to human judgments as to what the salient features of a document corpus are.

³To have the exact same values may seem strange at first but these are percentages of classes that beat more than 5% of the random models. Some of the classes that did well in the hundred model test did not meet the significance cut off in the thousand model test and vice versa but the end result was the same.

Acknowledgements

The authors would like to thank Dr. Alex Smola and Dr. Sanjay Chawla for their input into this research.

References

- D. Curtis B. V. North and P. C. Sham. 2002. Letter to the editor on “simulation-based p values: Response to north et al.”. *Am J Hum Genet*, 71:439–440.
- David Blei. 2004. *Probabilistic models of text and images*. Ph.D. thesis, U.C. Berkeley.
- Karl W. Broman and Brian S. Caffo. 2003. Letter to the editor on “simulation-based p values: Response to north et al.”. *Am J Hum Genet*, 72:496.
- Wray L. Buntine. 1995. Graphical models for discovering knowledge. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. S. Uthurasamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 59–83. MIT Press.
- Jay L. Devore. 1999. *Probability and Statistics for Engineering and the Sciences*. Duxbury.
- Warren J. Ewens. 2003. Letter to the editor on “on estimating p-values by monte carlo methods”. *Am J Hum Genet*, 72:496–497.
- Jon Patrick. 2006. The scamseek project - text mining for financial scams on the internet. In *Selected Papers from AusDM*, pages 295–302.
- M. Steyvers and T. Griffiths. 2005. Probabilistic topic models. In T. Landauer, D. Mcnamara, S. Dennis, and W. Kintsch, editors, *Latent Semantic Analysis: A Road to Meaning*. Laurence Erlbaum.