Topic Significance Ranking of LDA Generative Models

Loulwah AlSumait[1] Daniel Barbará[1] James Gentle[2] Carlotta Domeniconi[1]

[1] Department of Computer Science, George Mason University, Fairfax VA 22030, USA
[2] Department of Computational and Data Sciences, George Mason University, Fairfax VA 22030, USA

Abstract. Topic models, like Latent Dirichlet Allocation (LDA), have been recently used to automatically generate text corpora topics, and to subdivide the corpus words among those topics. However, not all the estimated topics are of equal importance or correspond to genuine themes of the domain. Some of the topics can be a collection of irrelevant or background words, or represent insignificant themes. Current approaches to topic modeling perform manual examination of their output to find meaningful and important topics. This paper presents the first automated unsupervised analysis of LDA models to identify and distinguish junk topics from legitimate ones, and to rank the topic significance. The basic idea consists of measuring the distance between a topic distribution and a "junk distribution". In particular, three definitions of "junk distribution" are introduced, and a variety of metrics are used to compute the distances, from which an expressive figure of topic significance is implemented using a 4-phase Weighted Combination approach. Our experiments on synthetic and benchmark datasets show the effectiveness of the proposed approach in expressively ranking the significance of topics.

1 Introduction

Probabilistic Topic Modeling (PTM) is an emerging Bayesian approach to summarize data, such as text, in terms of (a small set of) latent variables that correspond (ideally) to the underlying themes or topics. It is a statistical generative model that represents documents as a mixture of probabilistic topics and topics as a mixture of words. Among the variety of topic models proposed, Latent Dirichlet Allocation (LDA) [4] is a truly generative model that is capable of generalizing the topic distributions so that it can be used to generate unseen documents as well. The completeness of the generative process for documents is achieved by considering Dirichlet priors on the document distributions over topics and on the topic distributions over words.

The setting of the number of latent variables K is extremely critical and directly effects the quality of the model and the interpretability of the estimated topics. Models with very few topics would result in broad topic definitions that could be a mixture of two or more distributions. On the other hand, models

with too many topics are expected to have very specific descriptions that are uninterpretable [8]. Since the actual number of underlying topics is unknown and there is no definite and efficient approach to accurately estimate it, the inferred topics of PTM does not always represent meaningful themes. For example, Table 1 lists five topics discovered by LDA when run on the Reuters-21578 dataset with K set to 50. It can be seen that the first three topics correspond to legitimate classes of the data. However, the last two topics are collections of insignificant words that are meaningless to the thematic structure of Reuters corpus.

Table 1. The Reuters: examples of topics estimated by LDA.

Class	**Top Words**
coffee	export, coffee , quota , product , market , price , Brazil
ship	ship , gulf , attack , Iran , American , oil , tanker , water
oil/crude	oil , price, barrel , crude (0.045), increase, product , petroleum , energy
Na	was , report , official, any , did , said , ask , told , made , comment , time
Na	two , on , three , five, six, four , month, seven, eight

Although LDA is heavily investigated and cited in the literature, none of the research provided an automatic analysis of the discovered topics to validate their importance and genuineness. Almost all the previous work manually examines the output to identify genuine topics in order to justify their work. Some work [10] computed the average distance of word distributions between all pairs of topics to measure how distinct they are. However, this figure evaluates the model in general and not the individual topics. In addition, the distance of a topic from the others does not provide any insight on the significance of the semantic content of the topic. Other approaches have used the probability of the topic as an indication of its importance [6, 1]. However, as will be seen later, some meaningless topics that consist of common words across documents with different content can have a high probability.

This paper introduces a novel approach to automatically rank the LDA topics based on their semantic importance and, eventually, identify junk and insignificant topics. The idea is to measure the amount of insignificance that an inferred topic carries in its distribution by measuring how "different" the topic distribution is from a "junk" distribution. In this work, three definitions of Junk and Insignificant (J/I) topics are introduced. To quantify the difference between an estimated topic and a J/I distribution, a number of distance measures are used. Based on a Weighted Combination of multi-criteria decision analysis, this paper introduces a novel unsupervised quantification of the topic significance. Our experiments on synthetic and benchmark datasets show the effectiveness of the proposed topic significance ranking, and its ability to identify junk and insignificant topics. To the best of our knowledge, this is the first attempt to evaluate and rank topic significance of PTM models.

The rest of this paper is organized as follows. An overview of the problem definition and notations including a brief description of the Latent Dirichlet Allocation (LDA) topic model is given in Section 2. Section 3 introduces three definitions of J/I distributions and lists three distance measures by which the

difference of the estimated topics of PTM from the J/I topics is computed. Then, the proposed Topic Significance Ranking (TSR) approach is defined in Section 4 followed by the experimental results that we obtained from applying the TSR on simulated and real data. Our final conclusions and future work are discussed in Section 6.

2 Problem Definition

LDA is a hierarchical Bayesian network that represents the generative model of a corpus of documents [4]. LDA assumes the standard bag-of-words representation, where each document d is represented as a vector of counts with W components, where W is the size of the dictionary. The documents of the corpus are modeled as mixtures over K topics each of which is a multinomial distribution over the dictionary of words. Each topic, $\boldsymbol{\phi}^{(k)}$, is drawn from a Dirichlet with parameter β, while each document, $\boldsymbol{\theta}^{(d)}$, is sampled from a Dirichlet with parameter α. For each word token i in document d, a topic assignment z_i is sampled from $\boldsymbol{\theta}_d$ which is introduced to represent the responsibility of a particular topic in using that word in the document. Then, the specific word x_i is drawn from $\boldsymbol{\phi}_{z_i}$. An exact estimation of $\boldsymbol{\phi}^{(k)}$ and $\boldsymbol{\theta}^{(d)}$ is found to be intractable [4], thus approximations such as Gibbs sampling [6] and variational inference [4] are used.

To identify genuine topics from the LDA estimated topics, the following learning setting is considered. Given a dataset of D documents with a total of N token words and W unique terms, a topic model $\mathcal{T}$ is generated from fitting its parameters, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, to the dataset assuming that the number of topics is set to K. The matrix $\boldsymbol{\theta}$ is a $D \times K$ parameter matrix in which each row $\boldsymbol{\theta}^{(d)}$ is the multinomial distribution of document d. The matrix $\boldsymbol{\phi}$ consists of $W \times K$ parameters in which each column $\boldsymbol{\phi}^{(k)}$ represents the multinomial distribution of topic j [3]. Thus, the parameters in $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ indicate the relative importance of words in topics (i.e. $\phi_{w,k} = p(w|k)$) and the relative importance of topics in documents (i.e. $\theta_{d,k} = p(k|d)$), respectively.

In practice, a topic model $\mathcal{T}$ includes different sets of "*Junk and Insignificant*" (J/I) topics. A junk topic is an "uninterpretable topic that picks out idiosyncratic word combinations" [8]. An insignificant topic is a topic that consists of general words, known as "background words", which are commonly used in general or across a broad range of documents within each corpus/domain [5]. For domain experts and text miners, the content of these topics is low in significance and often meaningless.

To identify J/I topics, the approach is to define a decision criterion $\mathcal{C}$ as the distance D of the topic from a common J/I topic description Ω. If the distance is large, then this would provide a fair indication of the topic significance. However, if the distance of a topic to the J/I distribution is small, then the topic is more likely to be irrelevant to the domain structure.

[3] The notation $\boldsymbol{\phi}^{(k)}$ ($\boldsymbol{\theta}^{(d)}$) is used to indicate the topic (document) distribution. To refer to a particular probability value, this is noted by $\phi_{w,k}$ ($\theta_{d,k}$)

3 Junk/Insignificance Based Decision Criteria

This section introduces the J/I topic definitions and the distance measures that are used to evaluate the significance of a topic distribution.

3.1 Junk/Insignificance Topic Definitions

Uniform Distribution Over Words (W-Uniform) Aligning with the Zipf law for words [7], a genuine topic is expected to be modeled by a distribution that is skewed toward a small set of words, called "*salient words*", out of the total dictionary. A topic distribution under which a large number of terms are highly probable is more likely to be insignificant or "junk".

To illustrate this, the number of salient terms of topics estimated by LDA on the 20-Newsgroups dataset is computed. These words are the ones that have the highest conditional probability under a topic k. For each estimated topic, the number of salient words for which the total conditional probability is equal to some percentage, X, of the topic probability is counted. Then, these counts are averaged over all the topics. Table 2 lists the average and percentage of salient words for X ranged from 60% to 100%. The values are reported for experiments done with K set to 40 components.

It can be seen that most of the topic density corresponds to less than 3% of the total vocabulary. In fact, when $X = 100\%$, the average value was biased toward a set of extreme topics that have nonzero probability for the whole dictionary. When these topics were excluded from the average, the percentage of words dropped from 52% to 3.9%. These values are given in the table between parenthesis. Such topics are more likely to be junk topics that are irrelevant to the domain.

Table 2. 20-Newsgroups: average and percentage of terms and documents that hold X percent of the topic density estimated by LDA.

$D \times W$	X	Average # of terms	%age of dictionary	Average # of docs	%age of total docs
11269 × 53795	60%	227.25	0.42%	544.25	4.82%
	70%	342.7	0.64%	856.2	7.6%
	80%	521.35	0.97%	1382.9	12.27%
	90%	846.3	1.6%	2442.95	21.7%
	100%	28026.5 (2085.1)	52% (3.9%)	8538.8 (5202)	75.8% (46.1%)

Under this frame, an extreme version of a junk topic will take the form of a uniform distribution over the dictionary. This topic, which is named *W-Uniform*, is the first junk definition in this paper. Formally, W-Uniform is a junk topic, $\boldsymbol{\Omega}^{\mathcal{U}}$, in which all the terms of the dictionary are equally probable

$$P(w_i|\boldsymbol{\Omega}^{\mathcal{U}}) = \frac{1}{W}, \forall i \in \{1, 2, \ldots, W\} \tag{1}$$

The degree of "*uniformity*", $\mathcal{U}$, of an estimated topic, $\phi^{(k)}$, can be quantified by computing its distance from the W-Uniform junk distribution, $\boldsymbol{\Omega}^{\mathcal{U}}$. The

computed distance will provide a reasonable figure of the topic significance. The larger the distance is, i.e. the farther a topic description is from the uniform distribution over the dictionary, the higher its significance is, and vise versa. Other definitions of junk topics are given next.

The Vacuous Semantic Distribution (W-Vacuous) The empirical distribution (the total word frequencies of the whole sample) is a convex combination of the probability distributions of the underlying themes that reveals no significant information if taken as a whole. A distribution of a real topic is expected to have a unique characteristic rather than a mixture model. Thus, the closer the topic distribution is to the empirical distribution of the sample, the less its significance is expected to be.

So, the second junk topic, namely the vacuous semantic distribution (W-Vacuous), is defined to be the empirical distribution of the sample set. It is equivalent to the marginal distribution of words over the latent variables. The probability of each term w_i under the W-Vacuous ($\boldsymbol{\Omega}^{\mathcal{V}}$) topic is given by

$$p(w_i|\boldsymbol{\Omega}^{\mathcal{V}}) = \sum_{k=1}^{K} p(w_i|k)p(k) \tag{2}$$

where $p(w_i|k) = \phi_{i,k}$, by definition, and $p(k)$ is the probability of the topics under the PTM which can be computed from

$$p(k) = \frac{\sum_{d=1}^{D} \theta_{d,k}}{N} \tag{3}$$

In order to detect junk topics, the "*vacuousness*" decision criterion of a topic, $\mathcal{V}$, is measured by computing the distance between the estimated distribution and the W-Vacuous. Lower $\mathcal{V}$ distances corresponds to distributions with probability mixture models that represent insignificant topics.

The Background Distribution (D-BGround) The previous two definitions of junk topics are characterized by their distribution over words. However, investigating the distribution of topics over documents would identify another class of insignificant topics. In real datasets, well defined topics are usually covered in a subset (not all) of the documents. If a topic is estimated to be responsible of generating words in a wide range of documents, or all documents in the extreme case, then it is far from having a definite and authentic identity. Such topics are most likely to be constructed of the background terms, which are irrelevant to the domain structure.

Table 2 also provides the average and percentage of documents in which X percent of the topic density appears. In general, topics are inclined to appear heavily in a small subset of documents. Yet, nearly half of the topics are estimated to appear in a much larger fraction of documents, and in the extreme, in the whole the dataset. Examples of such topics are given in Table 3, in addition to examples of "normal" topics that appear in fewer documents.

Table 3. 20-Newsgroup: examples of background and legitimate topics.

TopicID (Class)	Top Words
9(NA)	edu writes article cs apr cc michael andrew bitnet colorado cmu ohio acs cwru au
36(NA)	university information research national april center washington san california dr
4(space)	space nasa gov earth launch moon orbit satellite shuttle henry lunar flight mission
5(Crypt)	encryption government clipper chip technology key law phone security escrow

To show reasonable significance for consideration, a topic is required to be far (enough) from being a "*background topic*", which can be defined as a topic that has a nonzero weight in all the documents. In the extreme case, the background topic (D-BGround) is found equally probable in all the documents. Formally, under the D-BGround topic, $\boldsymbol{\Omega}^{\mathcal{B}}$, the probability of each document d_m is given by

$$p(d_m|\boldsymbol{\Omega}^{\mathcal{B}}) = \frac{1}{D}, m \in \{1, 2, \ldots, D\}. \tag{4}$$

The distance between a topic and the D-BGround topic would determine how much "*background*" does it carry and, ultimately, grade the significance of the topic. Thus, given a topic k, defined as a distribution over documents

$$\boldsymbol{\vartheta}^{(k)} = (\theta_{1,k} \ldots \theta_{d,k} \ldots \theta_{D,k}), \tag{5}$$

then the background, $\mathcal{B}$, of a topic is measured by computing the distance of the topic distribution over documents from the D-BGround.

3.2 Distance Measures

Kullback-Leibler (KL) Divergence The KL-Divergence D_{KL} is a distance measure that is constructed based on the KL-divergence (or relative entropy) [2]. Thus, using D_{KL}, the distance of the topic distribution over words $\boldsymbol{\phi}^{(k)}$ from W-Uniform $\boldsymbol{\Omega}^{\mathcal{U}}$ and W-Vacuous $\boldsymbol{\Omega}^{\mathcal{V}}$, and the distance of the topic distribution over documents $\boldsymbol{\vartheta}^{(k)}$ from D-BGround $\boldsymbol{\Omega}^{\mathcal{B}}$ can be computed as follows

$$\mathcal{U}_k^{\mathrm{KL}} = D_{\mathrm{KL}}(\boldsymbol{\phi}^{(k)}, \boldsymbol{\Omega}^{\mathcal{U}}) \tag{6}$$

$$\mathcal{V}_k^{\mathrm{KL}} = D_{\mathrm{KL}}(\boldsymbol{\phi}^{(k)}, \boldsymbol{\Omega}^{\mathcal{V}}) \tag{7}$$

$$\mathcal{B}_k^{\mathrm{KL}} = D_{\mathrm{KL}}(\boldsymbol{\vartheta}^{(k)}, \boldsymbol{\Omega}^{\mathcal{B}}) \tag{8}$$

Cosine Dissimilarity The cosine dissimilarity D_{COS} is a distance measure that is constructed based on the cosine similarity [9]. Similar to the D_{KL} in Equations (6), (7), and (8), the cosine distance is used to measure the uniformity ($\mathcal{U}_k^{\mathrm{COS}}$), vacuousness ($\mathcal{V}_k^{\mathrm{COS}}$), and background ($\mathcal{B}_k^{\mathrm{COS}}$) of topic k based on the cosine angle between the inferred topic vector and the W-Uniform, W-Vacuous, and D-BGround vectors, respectively. A cosine distance of value 0 (1) corresponds to completely related (unrelated) topics.

Correlation Coefficient The correlation coefficient distance measure D_{COR} is a dissimilarity measure that is based on the correlation coefficient statistic [9]. The uniformity ($\mathcal{U}_k^{\mathrm{COR}}$), vacuousness ($\mathcal{V}_k^{\mathrm{COR}}$), and background ($\mathcal{B}_k^{\mathrm{COR}}$) of topic k under the correlation-based distance is computed by measuring the correlation between the topic distribution and the W-Uniform, W-Vacuous, and D-BGround vectors, respectively. The distance is bounded by the closed interval $[0, 2]$, where independent and negatively related topics will result in distances greater than or equal to one. This fits with the definition of our problem since semantic relatedness between topics is evinced by positive correlations only.

4 Topic Significance Ranking

Due to the uncertainty that surround the data and the statistical modeling, it is very appealing to have an expressive quantitative measure of the topic significance that can assist in discriminating genuine topics from J/I ones.

In this paper, three different categories of topic significance criteria are defined each of which is quantified by a variety of distance measures. The objective is to construct a qualitatively representative figure of the topic significance by combining the information from these "*multi-criteria measures*" to form a single index of evaluation based on a "*Weighted Linear Combination*" (WLC) decision strategy [3].

WLC is a simple technique that is widely used in the area of multi-criteria decision analysis [3]. The simplest form of WLC evaluates each topic by the following formula

$$A_k = \sum_{m=1}^{N_m} \Psi_m \mathcal{S}_{m,k} \tag{9}$$

where N_m is the number of distance measures to be combined, and Ψ_m is the weight of the J/I criterion in the total score, and $\mathcal{S}_{m,k}$ is the score of the k^{th} topic with respect to the m^{th} measure. Because of the different scales upon which these criteria are measured, it is necessary that the measures be standardized before combination.

In this work, a 4-phase weighted combination approach is introduced. The idea is to use the computed measurements to construct both the scores $\mathcal{S}_{m,k}$ and weights Ψ_m of the different criteria. To do so, two "*standardization procedures*" are performed in the first phase to transfer each distance measure from its true value into two standardized scores, one is a relative score of the distances and the other is a weight value between 0 and 1. Then, the standardized measurements of each topic within each J/I definition are combined into a single figure during the intra-criterion phase. In the third phase, two different techniques of "*Weighted Combination*" (WC) are performed to combine the J/I scores to construct a weight and a total score for each topic from which the final rank of the topic significance is computed. The following subsections describe each phase of the TSR in further details.

4.1 Standardization Procedure

Given the distance measures m, where $m \in \{KL, COR, COS\}$, under each J/I definition criterion $\mathcal{C}$, where $\mathcal{C} \in \{\mathcal{U}, \mathcal{V}, \mathcal{B}\}$, and for each topic k, the first phase is concerned with linearly transforming each distance value into a standardized score, denoted $\acute{\mathcal{C}}_k^m$, that maintains the relative order of distance magnitude with respect to the other topics instead of the original raw value.

To construct both the scores $\mathcal{S}_{m,k}$ and weights Ψ_m for each of the criteria, two standardization procedures are used. The first, re-scales the scores based on the weight of each score with respect to the total score over all topics. This is given in the form

$$\acute{\mathcal{C}}1_k^m = \mathcal{C}_k^m \times \frac{\sum_{j=1, j\neq k}^{K} \mathcal{C}_j^m}{\sum_{j=1}^{K} \mathcal{C}_j^m}. \tag{10}$$

The fraction in Equation (10) is a normalized weight of the topic distance.

The second standardization procedure is the score range procedure that uses the minimum and maximum values as scaling points for standardization. This is given by

$$\acute{\mathcal{C}}2_k^m = \frac{\mathcal{C}_k^m - \mathcal{C}_{\min}^m}{\mathcal{C}_{\max}^m - \mathcal{C}_{\min}^m} \tag{11}$$

where $\mathcal{C}_{\min}^m$ ($\mathcal{C}_{\max}^m$) is the minimum (maximum) distance value measured by the distance measure m under the criterion $\mathcal{C}$. While the first standardization rescales the raw measures to a relatively smaller range, the score range procedure bounds the resulted scores between zero and one. The former will be used as the topic score in the final TSR while the latter is used as the weight.

4.2 Intra-Criterion Weighted Linear Combination

Before computing the rank of topic significance, it is required to combine the different distance measures within each J/I criterion into a single figure. Thus, the second phase of the topic ranking performs a Weighted Linear Combination (WLC) of the standardized scores of the distance measures as given in Equation (9). The weights Ψ_m in the equation determine the contributions of the distance measures in the total score. In this work, all the attributes under each criterion are assumed to weigh equally. Thus, the intra-criterion WLC is given by the mean score of the three distance measures.

So, given the standardized scores of the three distance measures under criterion $\mathcal{C}$ for topic k, i.e. $\acute{\mathcal{C}}_k^{KL}$, $\acute{\mathcal{C}}_k^{COR}$, and $\acute{\mathcal{C}}_k^{COS}$, then, the WLC score of the criterion $\mathcal{C}$ for topic k is given by

$$\mathcal{S}_k^c = \frac{\acute{\mathcal{C}}_k^{\mathrm{KL}} + \acute{\mathcal{C}}_k^{\mathrm{COR}} + \acute{\mathcal{C}}_k^{\mathrm{COS}}}{3}. \tag{12}$$

Substituting the two standardized scores $\acute{\mathcal{C}}1_k^m$ and $\acute{\mathcal{C}}2_k^m$ in Equation (12) results in two scores $\mathcal{S}1_k^c$ and $\mathcal{S}2_k^c$. Under each standardized procedure, a topic will have

three intra-WLC scores $\mathcal{S}_k^u$, $\mathcal{S}_k^v$, and $\mathcal{S}_k^b$ based on the uniformity, vacuousness, and the background criteria, respectively.

4.3 Inter-Criterion Weighted Combination

In this phase, a Weighted Combination[4] (WC) is performed over the scores computed in phase two. This involves assigning a weight, $\hat{\Psi}_c$, to each criterion $\mathcal{C}$ in order to adjust its contribution in the final ranking. However, two different WC techniques are used to combine the scores and the weights.

The first WC technique is based on Equation (10) and uses the standardized score of the background criterion as a weight for the uniformity and vacuousness scores as follows

$$\hat{S}_k = \hat{\mathcal{S}1}_k^b \left(\acute{\Psi}_u \mathcal{S}1_k^u + \acute{\Psi}_v \mathcal{S}1_k^v \right) \tag{13}$$

where $\acute{\Psi}_u$ ($\acute{\Psi}_v$) is the weight of the Uniformity (Vacuousness) criterion in the score and $\hat{\mathcal{S}1}_k^b$ is the rank (or weight) of the topic background. The rank is computed by substituting the intra-criterion score of topic background, $\acute{\mathcal{S}1}_k^b$, for $\acute{\mathcal{C}}_k^m$ in Equation (11). So, the background indicator is used to weigh the uniformity and vacuousness of the topic's word distribution by the uniformity of its distribution over the documents.

The second technique is performed over the intra-criteria scores that are based on the score range standardization procedure (Equation 11). This is done by a simple application of the WLC in Equation (9) as follows

$$\hat{\Psi}_k = \Psi_u \mathcal{S}2_k^u + \Psi_v \mathcal{S}2_k^v + \Psi_b \mathcal{S}2_k^b \tag{14}$$

where Ψ_c is the weight of the criterion $\mathcal{C}$ in the score $\hat{\Psi}_k$. These weights are assumed to sum to 1 so that the total score remains bounded between zero and one.

4.4 The Final Topic Significance Score

To compute the final rank, the score in Equation (13) is considered the total topic score while the normalized weight in Equation (14) is used as the weight of the topic score. Thus, the final rank of the topic significance is given by

$$TSR_k = \hat{\Psi}_k \times \hat{S}_k. \tag{15}$$

[4] Since the weights and scores are constructed from the computed distances, this phase (and the one that follows) are no longer linear.

Table 4. Topic distributions of the simulated data.

TopicID	k1	k2	k3
TopicName	River	Bank	Factory
Topic %age	33%	34%	33%
↓Dictionary	$p(w_i\|k1)$	$p(w_i\|k2)$	$p(w_i\|k3)$
river	0.37	0	0
stream	0.41	0	0
bank	0.22	0.28	0
money	0	0.3	0.07
loan	0	0.2	0
debt	0	0.12	0
factory	0	0	0.33
product	0	0	0.25
labor	0	0	0.25
news	0.05	0.05	0.05
reporter	0.05	0.05	0.05

5 Experimental Design

The proposed post analysis to rank the significance of the topics in probabilistic models is evaluated on synthetic and real data. An LDA Gibbs sampler topic model is first used to learn the model parameters using the corpus of documents. The resulted topics are evaluated against the ground truth for the simulated data and the 20Newsgroups, and subjectively by investigating the topics and checking their significance for all the datasets. All experiments were implemented using a modified version of the "Matlab Topic Modeling Toolbox", authored by Mark Steyvers and Tom Griffiths[5].

The weights of the different criteria in Equations (13) and (14) are first tuned using four different sets of documents that were generated from the synthetic data under different settings of K. The weights that resulted in the best ranking based on the ground truth are then fixed for the real datasets. The weights of the topic uniformity and vacuousness in Equation (13) ($\acute{\Psi}_u$ and $\acute{\Psi}_v$) are set to 0.6 and 0.4, respectively, while they are assigned to equal values, $\Psi_u, \Psi_v = 0.25$, in Equation (14) and the background weight Ψ_b is set to 0.5.

There are four datasets used in the experiments.

Simulated Data The synthetic dataset consists of 6 samples of 16 documents that have been generated from three static equally weighted topic distributions. On average, the document size was 16 words. Table 4 shows the dictionary and topic distributions of the data. The dataset is configured such that shared words (background words) exists between subsets of topics, e.g. money and bank, and among all the topics, e.g. news and reporters. Given each sample of documents, LDA was run to estimate the topics.

In some experiments, fake junk topics were deliberately injected before computing the topic ranks. These topics were randomly sampled from the J/I topic distributions $\boldsymbol{\Omega}^{\mathcal{U}}$ and $\boldsymbol{\Omega}^{\mathcal{V}}$ and are denoted as Utopic and Vtopic, respectively.

[5] The Topic Modeling Toolbox is available at: psiexp.ss.uci.edu/research/programs_data/toolbox.htm

Table 5. TSR of simulated data without (left) and with (right) injected J/I topics.

Without injected J/I topics				
ID	Topic Distribution	$\hat{S}$	$\hat{\Psi}$	TSR
3	river stream	3.263	0.766	2.491
4	factory labor production	3.602	0.653	2.377
1	money loan debt	2.260	0.486	1.09
5	reporter bank	0.502	0.3673	0.191
2	bank news	0.212	0.246	0.05

With injected J/I topics				
ID	Topic Distribution	$\hat{S}$	Ψ	TSR
4	river stream	3.4	0.8	2.7
5	production factory labor	3.3	0.8	2.7
3	money loan debt	2.4	0.7	1.6
2	bank	1.5	0.5	0.8
1	reporter news	0.9	0.7	0.7
	Vtopic	1.7	0.2	0.3
	Utopic	0.14	0.03	0

20Newsgroup The 20 Newsgroups data set is a collection of approximately 20,000 newsgroup documents, partitioned across 20 different newsgroups[6]. Some of the newsgroups are very closely related to each other (e.g. comp.sys.ibm.pc.hard ware / comp.sys.mac.hardware), while others are highly unrelated (e.g misc.forsale / soc.religion.christian). Data preprocessing included the removal of stop and rare words. The final dataset consisted of 1,359,612 word tokens and a dictionary size of 46191 terms.

NIPS Proceedings The NIPS set consists of the full text of the 13 years of proceedings from 1988 to 2000 Neural Information Processing Systems (NIPS) Conferences[7]. The data was preprocessed for down-casing, removing stopwords and numbers, and removing the words appearing less than five times in the corpus. The data set contains 1,740 research papers, 13,649 unique words, and 2,301,375 word tokens in total.

5.1 Experimental Results

The Topic Significance Ranking algorithm was first evaluated on the simulated data. Table 5 lists the topics discovered by LDA with K set to 5 along with their total TSR rankings. The listed TSR is the average rank of the topic over six different samples. The topics are ordered by their significance index. It can be seen that the proposed ranking method is able to properly rank the topics based on their true significance. Both fake and true junk topics such as Topic 5 and 2 had the lowest ranks, while legitimate topics such as Topics 3 (k1), 4 (k3) and 1 (k2) have gained the highest ranks.

The rank of the topic, in general, depends on the amounts of background words that its distribution carries. For example, Topic k2 (money bank loan) is ranked lower than the other topics because both the “money” and ”bank” terms are shared between more than one theme.

The proposed TSR is also tested using the 20Newsgroups dataset. Table 6 lists the distribution of topics that received the highest (lowest) TSR index. To determine the class of each topic and compare the results with the ground

[6] The dataset is available at http://people.csail.mit.edu/jrennie/20Newsgroups/.

[7] The original dataset is available at the NIPS Online Repository. http://nips.djvuzone.org/txt.html.

truth, an F1 measure (pF1) is computed based on a "probabilistic Contingency Table" (pCT) of size $C \times K$, where C is the number of classes. The table is constructed based on the document-topic distribution and the document labels. By considering the X topics with the highest probability under each document, an entry $pCT(i, j)$ in the table is the average probability that a topic j appeared in documents of class i. Then, based on the contingency table pCT, the F1 measure is computed. A class is then assigned to the topic that has the highest pF1 measure . Table 6 lists the classes under which each topic had the highest pCT entry. The class that is assigned to the topic, i.e. the pF1 measure of the class document under that topic is the highest, are marked by *. Nearly half of the topics, e.g. 10, 32, and 34, did not get a high pF1 index for any class, while three topics (7, 12, and 23) had the highest pF1 measures for, and hence assigned to, more than one class.

It can be seen that the TSR rank matches the pF1 measure in 6 of the 10 highest ranked topics and 6 of the lowest ranked topics. For example, the distribution of topic 4 is focused on the "space" theme which matches the pF1 measure of the class "space" that has assigned the highest index for the class under that particular topic. Furthermore, the TSR is able to highly rank topics that better represent a class even though the corresponding pF1 index is not the highest. Topics 11 and 37 are examples of such topics. Although the classes "talk.politics.misc" and "comp.graphics" were assigned to topics 35 and 25, respectively, based on the pF1 measure, the word-distribution of topics 11 and 37 better represent the class semantics, see Table 6. Thus, the TSR does a better unsupervised judgment based on the topic distributions only.

When examining the lowest ranked topics, the most insignificant topics had a very low and approximately identical pF1 indexes for most, and sometimes all, of the classes. The word distribution of these topics clearly include a large set of background words, e.g. email header terminology (re, edu, ca), greeting words (topics 32 and 10), verbs (topic 16), and names of people and organizations (topic 34). The rest of the list included topics that had a high pF1 measure for one or more class, like topics 18, 19, and 21, or had been assigned a class, like topics 25 and 35. While an explanation regarding the latter topics was given earlier, the former topics illustrate another interesting observations. First, topic 18 contains the background words of the two religion related classes. As the TSR is high for topic 12 which better describes the underlying theme in more specific terms, it correctly identifies topic 18 as a background topic by assigning a low rank to it. The same explanation can be given for topic 25.

On the other hand, the class "misc.forsale" introduced a different behavior. First, the class is clearly heterogenous by its nature and involves a lot of shared words with other classes, particularly autos, electronics and computer-related classes. Thus, the topic is expected to have a large variance and heavy tailed distribution which makes it dominated for lower significance ranking. In addition, it can be seen that topic 19 provides a closer description for the class than topic 35. In fact, 294 documents (50.5%) of the class had topic 19 as the highest topic, compared to 11 documents for topic 35. However, the pF1 measure of topic 19

Table 6. The 20Newsgroups: distribution and class of the 10 highest and lowest ranked topics.

ID	Topic	class (pF1 measure)	TSR
Highest Ranked Topics			
12	god jesus christ christian bible christians hell faith lord paul believe	soc.religion.christian*(0.121) talk.religion.misc*(0.112)	19.06
11	president money think going stephanopoulos tax don insurance pay care working clinton jobs bill	talk.politics.misc(0.1)	16.76
39	file output program entry section check line build ok read	comp.windows.x(0.073) comp.sys.ibm.pc.hw(0.071)	16.1
37	edu software image graphics ftp version pub data images package	comp.graphics(0.074) comp.windows.x(0.072)	14.17
20	turkish armenian armenians war turkey armenia soviet today greek genocide history	talk.politics.mideast*(0.198)	13.05
27	window server display widget mit application motif set manager sun	comp.windows.x*(0.176)	12.99
17	la period st play pts power pp chicago gm flyers buffalo van mon	rec.sport.hockey(0.156)	12.91
6	god believe say true truth question exist reason evidence religion existence argument atheism atheists	alt.atheism*(0.106) soc.religion.christian(0.111)	12.57
4	space nasa gov earth launch moon orbit satellite shuttle henry lunar	sci.space*(0.144)	12.40
23	drive scsi mb disk hard card system bit mac drives speed bus mhz apple	comp.sys.ibm.pc.hw*(0.134) comp.sys.mac.hw*(0.105)	12.146
Lowest Ranked Topics			
32	thanks mail uk ac help advance fax university looking email hi appreciated		2.89
10	edu writes article ca apr news uiuc think don cso heard sorry		3.53
24	com writes article apr netcom hp opinions att ibm mark wrote		4.02
34	org edu chris david john scott mil navy ed jeff robert		4.06
16	don think re want going say things ll thing let ve doesnot maybe		4.14
35	black white edu virginia sex article sexual cover gay writes	misc.forsale*(0.128) talk.politics.misc*(0.118)	4.27
18	church think catholic thought true mean christian order group religion	talk.religion.misc(0.097) soc.religion.christian(0.086)	4.33
25	problem problems find line try ve help tried lines don	comp.graphics*(0.185) comp.sys.mac.hw(0.108)	4.407
19	price buy offer sale sell interested cd shipping printer asking sound condition apple cost computer	misc.forsale(0.084) comp.sys.mac.hardware(0.074)	4.47
21	writes science think system theory objective moral don morality article	sci.med(0.111) alt.atheism(0.106)	4.59

is less than the pF1 of topic 35 because the average relative importance of the latter topic in the class documents (0.5) is higher than the former topic (0.2). Although topic 35 is focused on political themes, the words "black" and "white" could be responsible of attracting the "forsale" documents into this topic. As a result, the topic's vacuous significance index is clearly affected.

The proposed TSR showed similar outcomes when tested on NIPS dataset. Table 7 lists the NIPS topics that gained the highest and lowest 10 indexes. The most significant topics clearly correspond to genuine themes of NIPS. Examples include reinforcement learning (30), speech recognition (41), image processing (43), and neuroscience (10). On the other hand, common terms across NIPS publications have been grouped by LDA in distinguished topics and have received the lowest significance rankings, see Table 7.

Table 7. The NIPS: distribution and ranks of the 10 highest and lowest ranked topics.

ID	Topic	$\hat{S}$	Ψ	TSR
	Highest Ranked Topics			
30	state action policy function reinforcement actions optimal time algorithm	23.4	0.7	16.8
10	firing spike cell cells neurons time potential membrane rate neuron	22.9	0.7	16.6
41	speech recognition word system training hmm words context speaker acoustic	23.0	0.7	16.2
17	cells cell visual orientation cortex receptive cortical spatial field fields	21.6	0.7	16.1
31	analog circuit chip figure current output vlsi voltage input circuits	22.9	0.7	15.8
29	control motor trajectory arm forward feedback movement inverse hand	20.7	0.7	15.1
43	image images visual pixel vision pixels figure edge features texture	19.6	0.7	14.3
44	motion direction velocity field moving flow directions eeg time optical	18.7	0.7	13.9
24	node nodes tree rules rule trees structure set representation connectionist	18.7	0.7	13.5
18	neurons synaptic input activity synapses connections inhibitory figure excitatory	18.9	0.6	12.7
	Lowest Ranked Topics			
27	case order general simple form work theory fact section terms	3.7	0.2	0.7
11	rate convergence results values number large random size constant fixed	9.5	0.33	3.2
34	method problem function optimal methods estimation solution parameter based	9.8	0.3	3.5
46	performance set training results test table number data method experiments	11.3	0.3	3.7
35	system time data systems real block large applications computer user	10.9	0.3	3.9
14	network neural net systems information architecture processing work	10.3	0.5	4.8
21	input output layer inputs training weights outputs network back hidden	11.9	0.5	6.2
16	noise information distribution correlation variance gaussian function density	12.5	0.5	6.3
12	local space figure points map point dimensional regions global region	12.1	0.5	6.3
32	learning algorithm weight gradient error weights descent time update	15.0	0.5	7.1

To verify the proposed approach to compute the ranks based on the judgments of a variety of distance measures, TSR rankings based on individual distance measures are constructed and compared to the proposed TSR. This is achieved by ignoring the intra-criterion WLC phase and directly combining the standardized distances for each of the individual measures separately. The resulted TRS ranks are called the TSR-KL, the TSR-COS, and the TSR-COR for the KL-divergence, the cosine dissimilarity, and the coefficient correlation based ranks, respectively. Table 8 shows the TSR rankings for the topics of the simulated data with injected J/I topics. The topics are ordered by the TSR-KL rank. Given the true densities of the topics (Table 4), it can be seen that ranks from individual measures do not always provide the correct judgments regarding the semantic significance of the topics. In fact, the injected J/I topics "Vtopic" and "Utopic" have received the highest ranks under TSR-COS (not shown), while topic "Utopic" was the highest ranked topic under TSR-COR. In addition, based on the TSR-KL, topic 5, which corresponds to the genuine theme $k3$, was ranked lower than other insignificant topics, topic 1 (reporter news) and topic 2 (bank). Clearly, the intra-criterion WLC of the distance measures strengthens the judgments of the individual measures and provides a better representation of the topics' semantic significance.

Similarly, testing on the 20Newsgroup reveals similar findings. Table 9 lists the 10 highest significant topics from the 20Newsgroups based on the TSR-KL rank. The table also shows the order of these topics under the cosine dissimilarity (TSR-COS) ranking and the proposed TSR ranking. It can be seen that the TSR-KL have introduced three topics to the list that are clearly not significant based on their distribution and the pF1 measure. The TSR-COS agrees with the TSR-KL in two of these J/I topics and introduces additional insignificant topic

Table 8. Synthetic data: the TSR based on individual distance measures compared to the combined TSR.

ID	Topic	**TSR-KL**	TSR- COS	TSR- COR	TSR
3	money loan debt	**2.05**	0.32	0.71	1.60
4	river stream	**2.01**	0.37	0.60	2.70
2	bank	**1.90**	0.36	0.58	0.80
1	reporter news	**1.83**	0.34	1.48	0.70
5	production factory labor	**1.82**	0.39	0.89	2.70
	Gtopic	**0.85**	0.45	0.46	0.30
	Utopic	**0.40**	0.42	0.89	0.00

Table 9. The 20Newsgroups: the 10 highest ranked topics based on the TSR-KL.

ID	Topic	class(pF1 measure)	TSR	TSR-COS
37	edu software image graphics ftp version pub data images package	comp.graphics(0.126) comp.windows.x(0.110)	4	8
39	file output program entry section check line build ok read	comp.sys.ibm.pc.hw(0.99) comp.windows.x(0.97)	3	1
23	drive scsi mb disk hard card system bit mac drives speed bus mhz memory controller pc board ram data apple	comp.sys.ibm.pc.hw*(0.208) comp.sys.mac.hw*(0.159)	10	9
31	list mail internet send information posting group email faq news address message usenet		12	3
20	turkish armenian armenians war turkey armenia soviet today greek genocide history	talk.politics.mideast*(0.360)	5	14
27	window server display widget mit application motif set manager sun	comp.windows.x*(0.252)	6	4
16	don think re want going say things ll thing let ve doesnot maybe		36	39
11	president money think going stephanopoulos tax don insurance pay care working clinton jobs bill	talk.politics(0.181)	2	11
12	god jesus christ christian bible christians hell faith lord paul believe	talk.religion.misc*(0.206) soc.religion.christian*(0.195)	1	18
8	db didn told home saw say don says going took re started happened building room wanted wife		15	6

(topic 33: de ma pa um em ei el rs sg mu di) to the list. In addition, topics such as 5, 12, and 16 illustrate how the combined rank provide a better judgment about the topic significance when compared to the individual measures.

6 Conclusion

In order to overcome the uncertainty that surrounds the outcome of a generative model, this paper presents a novel unsupervised analysis of Probabilistic Topic Models (PTM) for Topic Significance Ranking (TSR) to automatically distinguish genuine topics from Junk and Insignificant (J/I) topics. The proposed solution measures the distance of a topic from a set of J/I topic distributions using three different distance measures. A descriptive Topic Significance Ranking is constructed by applying 4 levels of Weighted Combination decision strategy. To the best of our knowledge, this work is the first attempt to automatically evaluate the inferred topics of PTM to judge their semantic significance.

The proposed ranking approach was evaluated on simulated and real datasets. The results are evaluated against the ground truth, when exists, and subjectively by examining the topic distribution. The outcomes confirm the potential of the

proposed method as it is able to correctly highly rank the true topics while J/I topics received low figures. The approach was also verified against less complex ranking systems that depend on the judgment of a single distance measure.

To extend this work, we plan first to investigate the sensitivity of the approach to the applied combination techniques and to the weight settings. In addition, analyzing the effect of the number of components on the resulted rank is also considered. Consequently, the rank can be extended to be used as an indicator to adjust the setting of this critical parameter. The use of other J/I definition criteria and/or distance measures is also under consideration. In addition, further analysis of the use of the TSR in visualizing the evolution of topics in streaming text is planned.

References

1. L. AlSumait, D. Barbará, and C. Domeniconi, Online LDA: Adaptive Topic Model for Mining Text Streams with Application on Topic Detection and Tracking, *Proceedings of IEEE International Conference on Data Mining (ICDM08)*, (2008).
2. C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
3. D. Bouyssou, T. Marchant, M. Pirlot, A. Tsoukias, and P. Vincke. *Evaluation and Decision Models with Multiple Criteria*. Springer, 2006.
4. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *The Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.
5. C. Chemudugunta, P. Smyth, and M. Steyvers, "Modeling General and Specific Aspects of Documents with a Probabilistic Topic Model," *Proceedings of Neural Information Processing Systems*, 2007.
6. T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceeding of the National Academy of Sciences*, pp. 5228–5235, 2004.
7. T. Joachims, A Statistical Learning Model of Text Classification with Support Vector Machines. Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), 2001.
8. M. Steyvers and T. L. Griffiths, "Probabilistic Topic Models," In T. Landauer, D McNamara, S. Dennis, and W. Kintsch (ed), *Latent Semantic Analysis: A Road to Meaning*, Laurence Erlbaum, 2005.
9. P. Tan, M. Steinbach, V. Kumar, *Introduction to Data Mining,* Pearson Addison Wesley, 2006.
10. X. Wang and A. McCallum, "Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends," *ACM SIGKDD international conference on Knowledge discovery in data mining*, 2006.