Semantic Visualization for Spherical Representation

Tuan M. V. Le
School of Information Systems
Singapore Management University
vmtle.2012@phdis.smu.edu.sg

Hady W. Lauw
School of Information Systems
Singapore Management University
hadywlauw@smu.edu.sg

ABSTRACT

Visualization of high-dimensional data such as text documents is widely applicable. The traditional means is to find an appropriate embedding of the high-dimensional representation in a low-dimensional visualizable space. As topic modeling is a useful form of dimensionality reduction that preserves the semantics in documents, recent approaches aim for a visualization that is consistent with both the original word space, as well as the semantic topic space. In this paper, we address the semantic visualization problem. Given a corpus of documents, the objective is to simultaneously learn the topic distributions as well as the visualization coordinates of documents. We propose to develop a semantic visualization model that approximates L^2-normalized data directly. The key is to associate each document with three representations: a coordinate in the visualization space, a multinomial distribution in the topic space, and a directional vector in a high-dimensional unit hypersphere in the word space. We join these representations in a unified generative model, and describe its parameter estimation through variational inference. Comprehensive experiments on real-life text datasets show that the proposed method outperforms the existing baselines on objective evaluation metrics for visualization quality and topic interpretability.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; H.2.8 [**Database Applications**]: Data Mining

Keywords

semantic visualization; topic model; generative model; spherical space; spherical semantic embedding; dimensionality reduction; L2-normalized vector;

1. INTRODUCTION

Visualization is an important and widely applicable tool for exploratory analysis of high-dimensional data. There are various aspects to the study of visualization (e.g., interface, interactivity). Of special interest to data mining and machine learning is the dimensionality reduction aspect of visualization, i.e., finding a low-rank representation in two or three dimensions that preserves as much as possible the properties of the data. These low-rank representations are visualized on a scatterplot, a simple format to reveal the relationship structures among data points. The current state-of-the-art visualization approaches are formulated as finding coordinates whose distances in the visualization space "reflect" the corresponding distances in the original space [12].

While a scatterplot is useful for identifying visualizable structures among data points, it has relatively limited *explanatory* power, because the reduced dimensions have no prescribed semantics. Another type of dimensionality reduction that is focused more on interpretability is topic modeling [17, 5], where the objective is to reduce each document's original representation (e.g., word counts) into a probability distribution over Z (a user-defined quantity) topics. Each topic is associated with a distribution over words. Hence, the reduced dimensions (i.e., topics) have interpretable semantics, revealed by the most important words in each topic. However, topic model is not designed for visualization. In a 2D simplex, we can visualize topic distributions for only 3 topics, which is impractical, because Z is frequently much higher than that (though lower than the vocabulary size).

We are therefore interested in *semantic visualization*, defined as modeling visualization coordinates and topics in an integrated manner [19]. This integration has important benefits not available to either visualization or topic model on their own. On the one hand, it allows the infusion of the scatterplot visualization with topic modeling semantics. Each coordinate in the visualization space can now be associated with both a topic distribution, as well as a list of the most important words. This complements structural observations (e.g., clustering) with semantic explanations (the relevant topics and words). On the other hand, we envision that visualization may eventually serve as a user-friendly interface to explore and tune an underlying topic model, in a way that allows steering the topic model interactively.

In this paper, we propose a semantic visualization model for data with *spherical representation*. This refers to data whose instances can each be represented as a vector of unit length in a high-dimensional hypersphere [1], with dimensionality commensurate with the number of features. In other words, we are dealing with L^2-normalized feature vectors as input. One important category of such data that we focus on in this work is text document. A document can be

KDD'14, August 24–27, 2014, New York, NY, USA.

ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623620.

naturally represented as a normalized term vector, as done in the classical vector space model [32]. Stated more formally, the input to the problem is a corpus of documents $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$, where every d_n is represented by an L^2-normalized term vector ν_n. We seek to learn, for each d_n, a probability distribution θ_n over Z topics (*semantic*), and a coordinate x_n on a low-dimensional space (*visualization*). While we frame the discussion here in terms of documents and words, our technique is applicable to other data types for which both visualization and semantic interpretability are important, as long as they can be expressed in terms of spherical representation (i.e., L^2-normalized vectors).

Previous Approach. Jointly modeling topics and visualization coordinates is pioneered by PLSV [19] (reviewed briefly in Section 3). It is aimed at *dyadic data*, whereby every observation involves a couple (d, w) of word w's occurrence in document d. The observations for a document can be summarized as an integer vector of word counts in $\mathbb{N}^{|V|}$, where V is the vocabulary. Like its topic modeling predecessors [17, 5], PLSV uses the word count vectors to maximize the likelihood of generating individual words based on the learned latent multinomial distribution over words $\{\mathrm{P}(w|d_n)\}_{w \in V}$. Here, $\mathrm{P}(w|d_n)$ is obtained from topics' word distribution $\mathrm{P}(w|z)$ and document's topic distribution $\mathrm{P}(z|d_n)$, i.e., $\mathrm{P}(w|d_n) = \sum_{z=1}^{Z} \mathrm{P}(w|z)\mathrm{P}(z|d_n)$.

The stated aim of most visualization approaches is to recover a low-dimensional manifold embedded within the high-dimensional space of the original data [22, 31, 16, 12]. Key to manifold learning is the capacity for approximating the similarities and differences among data instances [2]. In this respect, multinomial modeling of dyadic data has a couple of downsides [30]. For one thing, it primarily models word presences, but does not directly model word absences. The likelihood of a document is defined over only words present in the document. For another thing, it is also sensitive to document lengths. If one document were to contain two copies of each word in another document, the two documents would have different likelihoods, even though the word distributions in the two documents are effectively identical.

Proposed Approach. Spherical representation could address the above-mentioned issues, leading towards better approximation of similarities among documents, and thus towards better manifold learning and visualization. In the spherical space, relationships between documents are measured as cosine similarity $\in [0, 1]$, which is the angular distance between two directional unit vectors. Firstly, two documents would have higher cosine similarity, not only if some words in common are present, but also if some other words in common are absent. Secondly, the normalization of all documents to unit vectors effectively neutralizes the impact of document lengths. Moreover, there is indicative evidence from the literature that a spherical approach will be promising in terms of dimensionality reduction. For instance, the spherical topic model SAM [30] performs significantly better than the multinomial topic model LDA [5], when used as a dimensionality reduction technique.

There are further advantages to spherical representation. For one thing, there is a greater degree of *flexibility* in admitting different L^2-normalized representations, e.g., term frequency *tf* or *tf-idf* or other feature vectors. For another thing, there is a greater degree of *expressiveness*, as an L^2-normalized vector can have both positive and negative elements, representing the degrees of word presences and absences respectively. Inspired by [30], this expressiveness engenders a change in the topic definition, from multinomial word distribution to a unit term vector. Given a topic, we no longer associate a word with a probability value, but rather with a real value that expresses the word's presence or absence (the sign) and relative importance (the weight).

Contributions. Our problem formulation is novel because to the best of our knowledge, we are the first to address semantic visualization for spherical representation (*first contribution*). We propose a generative model called SSE, which stands for *Spherical Semantic Embedding*. In Section 2.1, we develop the full generative process of SSE (*second contribution*). To learn its parameters, we describe an estimation based on variational inference in Section 2.2 (*third contribution*). In Section 3, we review related work in visualization and topic modeling. In Section 4, we validate SSE through experiments on publicly available real-life datasets, showing significant gains in visualization quality and topic interpretability (*fourth contribution*). We conclude in Section 5.

2. SPHERICAL SEMANTIC EMBEDDING

2.1 Generative Model

Document Representations. We associate each document with representations in three different spaces. Table 1 provides a list of notations for reference.

- We model the *visualization space* as a Cartesian plane, where relationships can be visualized spatially in terms of Euclidean distances. This space is low-dimensional, and without loss of generality, we assume it has two dimensions (2D). Each document d_n is associated with 2D coordinates x_n. This is consistent with visualization techniques oriented towards dimensionality reduction [16, 12].

- We model the *topic space* as a $(Z-1)$-simplex, where Z is the number of topics. This is consistent with the practice in most topic models [17, 5, 30]. Each document d_n occupies a point θ_n in the simplex, which codes for a multinomial distribution over the topics $\{\mathrm{P}(z|d_n)\}_{z=1}^{Z}$.

- We model the *word space* as a $(|V|-1)$-dimensional unit sphere in $\mathbb{R}^{|V|}$, where V is the vocabulary. Each document d_n is associated with a directional, unit-length vector ν_n. For instance, ν_n could be a *tf-idf* vector, or other L^2-normalized vector. This is consistent with the vector space model [32], and spherical models [1, 30].

Of the three representations of d_n, only ν_n is observed, while x_n and θ_n are latent. A key step towards integrating visualization and topic modeling is to define a mapping between the spaces to ensure a consistency among the representations. In defining the mapping, we associate each topic z with representations in both the visualization space ϕ_z and the word space τ_z. The coordinate ϕ_z reveals where a topics is in the visualization space, allowing users to observe the relationships between documents and topics. The word vector τ_z reveals the topic semantics in terms of the relative importance of various words within τ_z.

Visualization Space to Topic Space. As both documents and topics have coordinates in the visualization space,

Notation	Description
d_n	a specific document
x_n	coordinate of d_n in the visualization space
θ_n	topic distribution of d_n
$\theta_{n,z}$	probability of topic z in document d_n
ν_n	the observed L^2-normalized word vector of d_n
z	a specific topic
ϕ_z	coordinate of topic z in the visualization space
τ_z	L^2-normalized word vector of topic z
V	the vocabulary (the set of words in the lexicon)
N	total number of documents in the corpus
Z	total number of topics (user-defined)

Table 1: Notations

their relationship can be expressed in terms of distances $||x_n - \phi_z||$. Intuitively, the closer is x_n to a topic's ϕ_z, the higher is $\theta_{n,z}$ or the probability of topic z for document d_n. One framework to relate variables based on distances is Radial Basis Function or RBF [7], which defines a function $\lambda(||x_n - \phi_z||)$ in terms of how far a data point (e.g., x_n) is from a center (e.g., ϕ_z). The function λ may take on various forms, e.g., Gaussian, multi-quadric, polyharmonic spline.

RBF network [3] is frequently used to build a function approximation. We use an RBF network as a "kernel" for the mapping between coordinates and topic distributions. To express θ_n as a function of x_n, we consider the normalized architecture of RBF network, with three layers. The input layer consists of one input node (x_n). The hidden layer consists of Z number of normalized RBF activation functions. Each is centered at ϕ_z and computes $\frac{\lambda(||x_n-\phi_z||)}{\sum_{z'=1}^{Z}\lambda(||x_n-\phi_{z'}||)}$. The linear output layer consists of Z output nodes. Each output node $y_z(x_n)$ corresponds to $\theta_{n,z}$, which is a linear combination of the RBF functions, as shown in Equation 1. Here, $w_{z,z'}$ is the weight of influence of the RBF function of z' on the $\theta_{n,z}$, with the constraint $\sum_{z'=1}^{Z} w_{z,z'} = 1$.

$$\theta_{n,z} = y_z(x_n) = \frac{\sum_{z'=1}^{Z} w_{z,z'} \cdot \lambda(||x_n - \phi_{z'}||)}{\sum_{z'=1}^{Z} \lambda(||x_n - \phi_{z'}||)} \tag{1}$$

While Equation 1 is the general form, to instantiate a specific mapping function, we need to determine both the assignment of $w_{z,z'}$ and the form of the function λ. In this work, we will experiment with a special case (λ is Gaussian and $w_{z,z'} = 1$ when $z = z'$ and 0 otherwise), which yields the function in Equation 2, where Φ refers to the collective set of ϕ_z's. This specific function has appeared previously in the baseline [19] that we will compare to, and this design decision helps to establish parity for comparative purposes. In future work, we will explore other function instantiations.

$$\theta_{n,z} = \mathrm{P}(z|x_n, \Phi) = \frac{\exp(\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{Z}\exp(\frac{1}{2}||x_n - \phi_{z'}||^2)} \tag{2}$$

Topic Space to Word Space. For d_n, we also need to bridge θ_n to its word space representation ν_n. As introduced previously, each topic z also has a word space representation τ_z. Because θ_n is essentially a topic distribution, we adopt a similar practice as in conventional topic model, which represents a document's word distribution as a weighted average (based on topic distribution) of the topics' word distributions. In our context, it means taking a weighted average

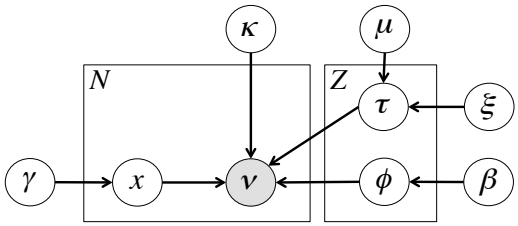


Figure 1: Graphical Model of SSE

of the topics' spherical unit vectors τ_z's, weighted by $\theta_{n,z}$, followed by L^2-normalization to return the mean vector to unit length, i.e., $\tau_n = \frac{\sum_{z=1}^{Z}\theta_{n,z}\cdot\tau_z}{||\sum_{z=1}^{Z}\theta_{n,z}\cdot\tau_z||}$.

To avoid overfitting, instead of equating ν_n to τ_n, we assume a probabilistic process where ν_n is drawn from a distribution centered at τ_n. Because ν_n and τ_n are both directional vectors, we turn to directional statistics [25]. In particular, von Mises-Fisher (vMF) distribution [24] was previously used to model documents [1, 30]. Equation 3 specifies the probability density function (p.d.f.) for a random unit vector ν, given mean directional vector μ, and concentration parameter κ. Note how the p.d.f. is parameterized by the cosine similarity $\mu^T\nu$ between the mean direction μ and ν, which is effectively the angular distance between the two unit vectors. The higher the κ, the more concentrated the distribution is around μ. The distribution is unimodal for $\kappa > 0$, and is uniform for $\kappa = 0$. C_D is the normalization constant, defined in Equation 4, where I_r denotes the modified Bessel function of the first kind and order r.

$$\mathrm{vMF}(\nu; \mu, \kappa) = C_D(\kappa)\exp(\kappa\mu^T\nu) \tag{3}$$

$$C_D(\kappa) = \frac{\kappa^{D/2-1}}{(2\pi)^{D/2}I_{D/2-1}(\kappa)} \tag{4}$$

We can then express ν_n as a draw from a vMF distribution with mean direction τ_n, i.e., $\nu_n \sim \mathrm{vMF}(\tau_n, \kappa)$.

Generative Process. We join the three representations into a generative model, with graphical representation as in Figure 1. The generative process of SSE is as follows:

1. Draw the corpus mean direction: $\mu \sim \mathrm{vMF}(m, \kappa_0)$
2. For each topic $z = 1, \ldots, Z$:
   - Draw z's coordinate: $\phi_z \sim \mathrm{Normal}(0, \beta^{-1}I)$
   - Draw z's spherical direction: $\tau_z \sim \mathrm{vMF}(\mu, \xi)$
3. For each document d_n, where $n = 1, \ldots, N$:
   - Draw d_n's coordinate: $x_n \sim \mathrm{Normal}(0, \gamma^{-1}I)$
   - Derive d_n's topic distribution: $\theta_{n,z} = \mathrm{P}(z|x_n, \Phi) = \frac{\exp(-\frac{1}{2}||x_n-\phi_z||^2)}{\sum_{z'=1}^{Z}\exp(-\frac{1}{2}||x_n-\phi_{z'}||^2)}$
   - Derive d_n's spherical average: $\tau_n = \frac{\sum_{z=1}^{Z}\theta_{n,z}\cdot\tau_z}{||\sum_{z=1}^{Z}\theta_{n,z}\cdot\tau_z||}$
   - Draw d_n's spherical direction: $\nu_n \sim \mathrm{vMF}(\tau_n, \kappa)$

In Step 1, we draw the corpus mean direction μ. In Step 2, we draw, for each topic, a visualization coordinate ϕ_z and a spherical direction τ_z. In Step 3, we draw, for each document, a visualization coordinate x_n, which we use to

compute topic distribution θ_n as a function of document and topics' coordinates. θ_n together with different topics' τ_z's are used to compute the weighted average of topics' directions, denoted τ_n. After normalizing τ_n to a unit-length vector, we draw ν_n from a vMF with mean τ_n. Though the observed ν_n is usually positive (e.g., *tf-idf*), the latent τ_n may contain negative elements, which reflect unlikely words.

2.2 Parameter Estimation

To estimate the parameters in SSE, we employ variational EM with maximum a posteriori (MAP) estimation. The unknown parameters are the coordinates for documents (collectively $\chi = \{x_n\}$) and for topics (collectively $\Phi = \{\phi_z\}$), the directional vectors for topics (collectively $\mathcal{T} = \{\tau_z\}$) and the hyperparameters ξ, m. Given a corpus $\mathcal{D}$, which are represented as L^2-normalized term vectors $\mathcal{V} = \{\nu_n\}_{n=1}^N$, we infer the posterior distribution $\mathrm{P}(\mathcal{T}, \mu|\mathcal{V}, \chi, \Phi, \beta, \gamma, \xi, m, \kappa_0, \kappa)$ of the directional vectors for topics (collectively $\mathcal{T} = \{\tau_z\}$) and the corpus mean direction μ.

We approximate the posterior using the following variational distribution:

$$q(\mathcal{T}, \mu|\tilde{\mu}, \xi) = q(\mathcal{T}|\tilde{\mu}, \xi)q(\mu|\tilde{m}, \kappa_0)$$

where $q(\tau_z) = \mathrm{vMF}(\tau_z|\tilde{\mu}, \xi)$, $q(\mu_z) = \mathrm{vMF}(\mu_z|\tilde{m}_z, \kappa_0)$ and the variational parameters are $\tilde{\mu}, \tilde{m}$. Given this variational distribution q, we have a lower bound $\mathcal{L}(\tilde{\mu}, \tilde{m})$ on the log likelihood with priors over the document and topic visualization coordinate x_n, ϕ_z, as follows:

$$\begin{aligned}
\mathcal{L}(\tilde{\mu}, \tilde{m}) &= \mathbb{E}_q\big[\log p(\mathcal{V}, \mathcal{T}, \mu)\big] - \mathbb{E}_q\big[\log q(\mathcal{T}, \mu|\tilde{\mu}, \xi)\big] \\
&\quad + \sum_{n=1}^{N} \log p(x_n) + \sum_{z=1}^{Z} \log p(\phi_z) \\
&= \mathbb{E}_q\big[\log p(\mathcal{V}|\mathcal{T}, \chi, \Phi)\big] + \mathbb{E}_q\big[\log p(\mathcal{T}|\mu, \xi)\big] \\
&\quad + \mathbb{E}_q\big[\log p(\mu)\big] - \mathbb{E}_q\big[\log p(\mathcal{T}|\tilde{\mu}, \xi)\big] - \mathbb{E}_q\big[\log p(\mu|\tilde{m}, \kappa_0)\big] \\
&\quad + \sum_{n=1}^{N} \log p(x_n) + \sum_{z=1}^{Z} \log p(\phi_z)
\end{aligned}$$

In the E-step, we optimize the lower bound $\mathcal{L}(\tilde{\mu}, \tilde{m})$ with respect to the variational parameters $\tilde{\mu}, \tilde{m}$. In the M-step, the lower bound is optimized with respect to the parameters χ, Φ, ξ, m. We alternate E and M-steps until some appropriate convergence criterion is reached. We use gradient-based numerical optimization method such as the quasi-Newton method to update $\tilde{\mu}, \chi, \Phi, \xi$.

E-step. Let $\rho_n = \mathbb{E}\big[\tau_n\big]^T \nu_n$ where $n \in \{1...N\}$ ranges over the documents. Taking the gradients of $\mathcal{L}(\tilde{\mu}, \tilde{m})$ w.r.t $\tilde{\mu}$, we have:

$$\bigtriangledown_{\tilde{\mu}_z} \mathcal{L} = A_V(\xi) A_V(\kappa_0) \xi \tilde{m}_z + \kappa \sum_{n=1}^{N} \bigtriangledown_{\tilde{\mu}_z} \rho_z$$

where $A_p(c)$ denotes the mean resultant length of a vMF distribution of dimension p with concentration c. Since $\mathbb{E}\big[\tau_n\big]$ does not have a closed form, following [30] we approximate it as:

$$\mathbb{E}\big[\tau_n\big] \approx \mathbb{E}\big[\sum_{z=1}^{Z} \theta_{n,z} \cdot \tau_z\big] \mathbb{E}\big[||\sum_{z=1}^{Z} \theta_{n,z} \cdot \tau_z||^2\big]^{-1/2}$$

We refer to $\mathbb{E}\big[||\sum_{z=1}^{Z} \theta_{n,z} \cdot \tau_z||^2\big]$ as S_n. ρ_n will be approximated as:

$$\rho_n \approx A_V(\xi) S_n^{-1/2} (\tilde{\mu}\theta_n)^T \nu_n$$

where

$$S_n = (1 - A_V(\xi)^2) \sum_z \theta_{nz}^2 + A_V(\xi)^2 ||\tilde{\mu}\theta_n||^2$$

Taking the gradients of ρ_n w.r.t $\tilde{\mu}_j$, yields:

$$\bigtriangledown_{\tilde{\mu}_j} \rho_n = A_V(\xi)\big(\frac{\theta_{n,j}\nu_n}{\sqrt{S_n}} - \frac{(\tilde{\mu}\theta_n)^T \nu_n}{2S_n^{3/2}} \cdot \bigtriangledown_{\tilde{\mu}_j} S_n\big)$$

where

$$\bigtriangledown_{\tilde{\mu}_j} S_n = 2A_V(\xi)^2 \theta_{n,j} \tilde{\mu}\theta_n$$

The variational corpus mean $\tilde{m}$ has a closed form update rule:

$$\tilde{m} \propto \kappa_0 m + A_V(\xi) \xi \sum_{z=1}^{Z} \tilde{\mu}_z$$

M-step. In the M-step, taking gradients of $\mathcal{L}(\tilde{\mu}, \tilde{m})$ w.r.t ξ, we have:

$$\bigtriangledown_{\xi} \mathcal{L} = (\bigtriangledown_{\xi} A_V(\xi)\xi + A_V(\xi))(A_V(\kappa_0)\tilde{m}^T \sum_z \tilde{\mu}_z - Z) + \kappa \sum_{n=1}^{N} \bigtriangledown_{\xi} \rho_n$$

where

$$\bigtriangledown_{\xi} \rho_n = \big(\bigtriangledown_{\xi} A_V(\xi) S_n^{-1/2} - \frac{1}{2} A_V(\xi) S_n^{-3/2} \bigtriangledown_{\xi} S_n\big)(\tilde{\mu}\theta_n)^T \nu_n$$

and

$$\bigtriangledown_{\xi} S_n = 2A_V(\xi) \bigtriangledown_{\xi} A_V(\xi)(||\tilde{\mu}\theta_n||^2 - \sum_z \theta_{nz}^2)$$

The corpus mean m has a closed form update rule as follows:

$$m \propto \sum_z \tilde{\mu}_z$$

Taking the gradients of $\mathcal{L}(\tilde{\mu}, \tilde{m})$ w.r.t x_n, we have:

$$\bigtriangledown_{x_n} \mathcal{L} = \kappa A_V(\xi)\big(-\frac{\bigtriangledown_{x_n} S_n}{2S_n^{3/2}} \tilde{\mu}\theta_n + \frac{\tilde{\mu} \bigtriangledown_{x_n} \theta_n}{\sqrt{S_n}}\big)^T \nu_n - \gamma x_n$$

where

$$\bigtriangledown_{x_n} S_n = 2(1 - A_V(\xi)^2) \sum_z \bigtriangledown_{x_n} \theta_{nz} \theta_{nz} + 2A_V(\xi)^2 \theta_n^T \tilde{\mu}^T \tilde{\mu} \bigtriangledown_{x_n} \theta_n$$

Taking the gradients of $\mathcal{L}(\tilde{\mu}, \tilde{m})$ w.r.t ϕ_z, we have:

$$\bigtriangledown_{\phi_z} \mathcal{L} = \kappa A_V(\xi)\big(-\frac{\bigtriangledown_{\phi_z} S_n}{2S_n^{3/2}} \tilde{\mu}\theta_n + \frac{\tilde{\mu} \bigtriangledown_{\phi_z} \theta_n}{\sqrt{S_n}}\big)^T \nu_n - \beta\phi_z$$

where

$$\bigtriangledown_{\phi_z} S_n = 2(1 - A_V(\xi)^2) \sum_{z'} \bigtriangledown_{\phi_z} \theta_{nz'} \theta_{nz'} + 2A_V(\xi)^2 \theta_n^T \tilde{\mu}^T \tilde{\mu} \bigtriangledown_{\phi_z} \theta_n$$

3. RELATED WORK

Visualization. The *dimensionality reduction* aspect of visualization is related to such techniques as PCA [20], ICA [11], and Fisher's Linear Discriminant [14], which are frequently used for feature selection. However, they are not designed specifically for visualization, and are concerned more with the relationship between the dimensions (orthogonality or independence), rather than the relationship between instances. Moreover, due to linear projection, PCA and variants do not capture intrinsic non-linearities well, such as

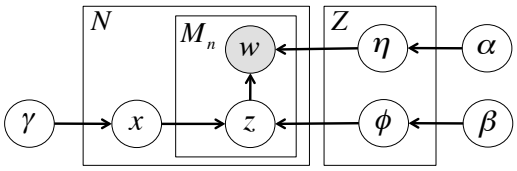


Figure 2: Graphical Model of PLSV

when the data is embedded as a low-dimensional non-linear manifold [2] (frequently-made assumption in visualization).

Therefore, visualization is often formulated as *manifold embedding*, where the objective is to preserve the relationship among data instances in the low-dimensional representation. In most cases, this relationship is expressed in terms of pairwise distance, such as in MDS [22], LLE [31], and Isomap [33]. Recent approaches employ probabilistic formulations, such as PE [18] and t-SNE [12], which we use as baselines. Yet, rather than distances, others seek to preserve the neighborhood information (e.g., SOM [21], GTM [4]).

The related work mentioned above has not incorporated the intermediate topic space. The problem of *semantic visualization* is introduced by PLSV [19]. We briefly review PLSV, whose graphical model is shown in Figure 2. The generative process of PLSV is as follows. For each topic z, we draw its word distribution η_z from a Dirichlet with parameter α, as well as its coordinate ϕ_z from Normal distribution with mean 0 and variance β^{-1}. In turn, for each document d_n, we draw its coordinate from Normal with mean 0 and variance γ^{-1}. To generate each of the M_n words in d_n, we draw a topic z based on Equation 2, and then draw a word from the selected topic's word distribution η_z. The key difference between SSE and PLSV is the representation of a document. Where SSE models the generation of an L^2-normalized vector, PLSV models the multinomial generation of words w. We compare to PLSV in Section 4.

We also describe a few other works in semantic visualization that are related, though not directly comparable. LDA-SOM [26] is a pipeline of LDA [5] followed by SOM [21], whose output is a topographic map not directly comparable to our scatterplot. Semafore [23] introduces manifold regularization to semantic visualization, which is orthogonal to the direction pursued in this paper (spherical representation), as manifold regularization could be applicable to both multinomial as well as spherical representations. CCG [29] is a topic model based on a latent grid space. It seeks to improve topic models, and is not designed specifically for document visualization. For one thing, the grid cells are discrete (unlike PLSV or SSE with continuous visualization space). For another, each document is associated with multiple grid cells, and it is not clear how to visualize such documents.

While semantic visualization deals with visualizing the relationship of documents based on topic modeling, another orthogonal direction is to visualize the topics themselves, such as the prevalence of topics in a corpus [34, 15], or the dominant keywords in topics [9, 10]. These works tend to be on the HCI aspects, such as user interfaces [13], rather than on dimensionality reduction or statistical modeling.

Topic Model. Probabilistic topic modeling is popularized by PLSA [17], and eventually by LDA [5], which provides a fully Bayesian generative model. These probabilistic models associate each document with a multinomial distribution over topics, and indirectly a multinomial distribution over words, effectively a simplex representation.

Recognizing the usefulness of L_2-normalized representations, SAM [30] introduces a topic model, which associates each document with a multinomial distribution of topics, and a directional unit vector in a spherical word space. With SAM, SSE shares a similar modeling of topics in the spherical space. The key difference is that SAM models only topics, whereas SSE also needs to model visualization in addition to topics. This requires a fundamental change in how a document's topic distribution is derived. Unlike SAM, in SSE the topic distribution θ_n is not drawn from a Dirichlet. Instead, to reflect the visualization objective, θ_n is expressed as a function of visualization coordinates (see Equation 1).

4. EXPERIMENTS

We conduct comprehensive experiments to evaluate the effectiveness of SSE, in terms of the quality of its outputs (primarily visualization, but also topic model).

4.1 Experimental Setup

Datasets. We rely on three publicly-available[1], real-life datasets [8]. $20News$ consists of newsgroup documents (in English) belonging to twenty classes. $Reuters8$ consists of newswire articles (in English) from eight classes. $Cade12$ consists of web pages (in Brazilian Portuguese) classified into twelve classes. These datasets are chosen because they represent benchmark datasets for document clustering and classification tasks. Each document in the dataset has a known class label. Because the semantic visualization task is unsupervised, these labels are not required for learning. However, they represent an objective ground truth, which we would use to evaluate visualization quality. In addition, they cover diverse document types, and different languages.

Following the practice in [19], we create balanced datasets by randomly sampling 50 documents from each class, resulting in, *for each sample*, 1000 documents for $20News$, 400 for $Reuters8$, and 600 for $Cade12$. These are of comparable sizes to those used in [19]. Moreover, because the algorithms are statistical, we draw five independent samples from each dataset, and run each sample five times. Hence, for each setting, the reported result is an average of 25 runs. Vocabulary sizes are similar among samples of the same dataset, with a maximum of 5455 for $20News$, 1972 for $Reuters8$, 7622 for $Cade12$. These are the dimensionalities of the word space.

L^2-normalized Representation. SSE admits different options for the L^2 representation of a document. The option that is most well-recognized in the information retrieval literature is *tf-idf*. We experimented with several alternatives, such as word count or term frequency (*tf*), and found *tf-idf* to give the best results. This echoes the finding in [30], which concluded that *tf-idf* was a better document representation than *tf*. Thus, we will use *tf-idf* in the experiments.

4.2 Comparative Methods

The comparative methods, and their attributes, are summarized in Table 2. **SSE** is our proposed method. A proper comparison is to another approach that jointly models visualization and topics, i.e., PLSV [19], which we use as the primary baseline. For completeness, we include other baselines

[1] `http://web.ist.utl.pt/acardoso/datasets/`

	Visualization	Topic model	Joint model	Spherical representation
SSE	✓	✓	✓	✓
PLSV	✓	✓	✓	
t-SNE	✓			✓
PE (SAM)	✓	✓		✓
PE (LDA)	✓	✓		

Table 2: Comparative Methods

in visualization (t-SNE, PE). While not direct competitors, they allow us to highlight certain aspects of our model.

PLSV [19] is a semantic visualization method based on multinomial modeling for dyadic data. Therefore, it is the proper baseline to SSE, allowing us to investigate the effects of SSE's modeling of *spherical representation*. For PLSV, we use the same settings as in the original paper [19] ($\beta = 0.1N$ and $\gamma = 0.1Z$, which we apply to SSE as well). We implement PLSV on our own (its authors have not made their implementation available), and verify that the results are similar to those reported in the original paper [19].

t-SNE [12] stands for t-distributed Stochastic Neighbor Embedding. It is one of the state-of-the-art approaches in visualizing high-dimensional data. Its input are feature vectors in the original dimensions, which in our context are the L^2-normalized *tf-idf* vectors. The idea behind t-SNE is to preserve the pairwise distances in the high dimensions in the visualization space. In addition to benchmarking against direct visualization, including t-SNE allows us to investigate the effects of *topic model* on visualization, by comparing SSE against an approach with the same input (*tf-idf* vectors) and output (visualization), but which does not have an intermediate topic space. We use the R implementation[2] of t-SNE with default settings and perplexity 40 as in [12].

PE [18] stands for Parameteric Embedding. It is also one of the state-of-the-art approaches in visualization, but is aimed at visualizing discrete probability distributions (e.g., class or topic distributions). PE cannot stand alone, as it needs to be coupled with a method that produces topic distributions. Including PE allows us to investigate the effects of modeling visualization and topic model *jointly*, as opposed to obtaining topic model separately before feeding it into PE. To produce the topic distributions, we experiment with two other topic models, as follows. **PE (LDA)** couples PE with LDA [5], which operates in the simplex word space. For LDA, we use the implementation[3] by its first author D. Blei. **PE (SAM)** couples PE with SAM [30], which operates in the spherical word space. For SAM, we use the implementation[4] by an author A. Waters with default settings ($\kappa_0 = 10, \kappa = 5000$, which we apply to SSE as well). [19] showed that PE with PLSA [17] is inferior to PLSV.

For visualization, we will be comparing SSE against PLSV, t-SNE, PE (SAM) and PE (LDA). We also investigate the topic models, comparing SSE against PLSV, and the two topic models used with PE, i.e., SAM and LDA. As input, for models with spherical representation (see Table 2), we use *tf-idf* vector (as explained in Section 4.1). For the multinomial models, we use their regular inputs (word counts).

[2]http://cran.r-project.org/web/packages/tsne/
[3]http://www.cs.princeton.edu/ blei/lda-c
[4]`https://github.com/austinwaters/py-sam`

4.3 Visualization Quality

Metric. The utility of a scatterplot visualization is in allowing the user to perceive similarities between documents through their distances in the visualization space. Our emphasis is on the strength of the dimensionality reduction, rather than on the user interface aspect. Evaluating dimensionality reduction through user studies is hard on the evaluator, may be overly subjective and not repeatable across evaluators. On the other hand, there exists established metrics to measure dimensionality reduction objectively.

One such approach is to rely on the available class labels as ground truth. Intuitively, documents of the same class are more likely to be similar than documents from different classes. A good visualization will "encode" this intuition, by placing documents of the same class nearby, and documents of different classes apart in the visualization space. Since dimensionality reduction means that the lower-dimensional representation still preserves the "properties" of the data, we can measure how well a visualization output reflects this intuition, by employing each document's visualization coordinates as a reduced "feature vector" in a classification task.

The choice of the classification method is not germane, because it is the feature vector that is being evaluated. In favor of simplicity, we employ kNN classification. For each document, we hide its class label, and predict a label by majority voting among its k-nearest neighbors as determined by Euclidean distance on the visualization space. The accuracy at k or $accuracy(k)$ is the fraction of documents whose predicted label based on kNN matches the true label. The higher the accuracy, the better is a visualization at encoding the class information. 1 is the highest possible accuracy, and 0 the lowest. The same metric was also used in [19].

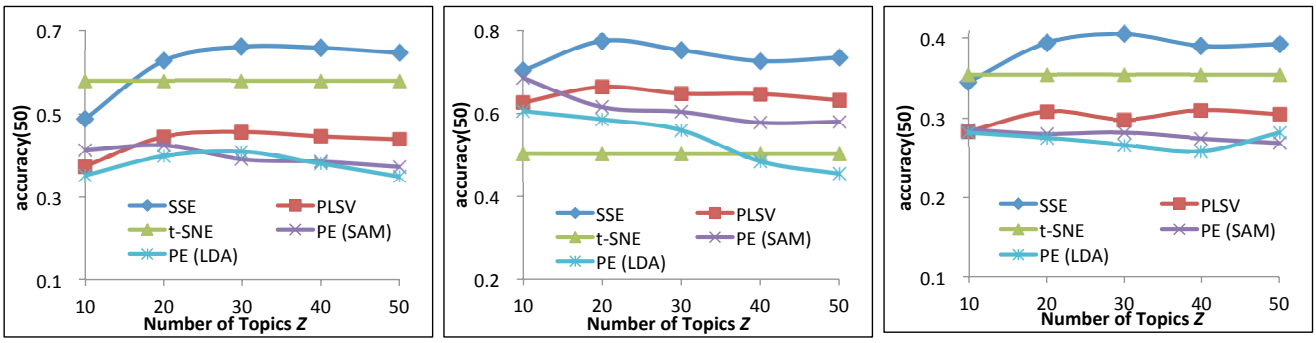
For relative comparison, we set $k = 50$, i.e., measuring $accuracy(50)$, which is appropriate, as the datasets contain 50 documents from each class. Setting $k << 50$ may not sufficiently penalize a visualization that splits documents of the same class into multiple small clusters in different localities.

Vary Number of Topics Z. We now compare the performance of various methods. In Figure 3, we plot the $accuracy(50)$ as we vary the number of topics Z from 10 to 50. The three sub-plots (a), (b), and (c) correspond to the three datasets $20News$, $Reuters8$, and $Cade12$ respectively.

In terms of SSE's performance as the number of topics varies: (#1) As the number of topics Z increases, initially there is an improvement in accuracy, most notably between $Z = 10$ and $Z = 30$. Thereafter, accuracies either remain flat or drop slightly as Z increases further. The best performance by SSE is 0.66 on $20News$ (at $Z = 30$), 0.77 on $Reuters8$ (at $Z = 20$), and 0.41 on $Cade12$ (at $Z = 30$).

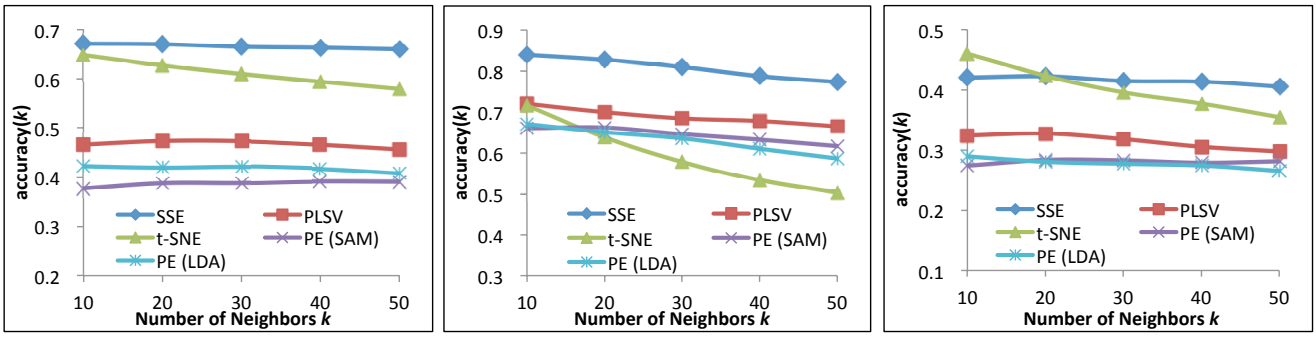
(#2) SSE achieves a drastic reduction in dimensionality from thousands (vocabulary size) to two (visualization), while preserving the relationship between data points. The above accuracies as measured in the reduced dimensionality (visualization) approach closely the accuracies of kNN when using the full dimensionality (i.e., *tf-idf* input vectors), which are 0.73 on $20News$, 0.85 on $Reuters8$, and 0.52 on $Cade12$. This shows that SSE's low-dimensional representation has high approximation ratios of 90% for $20News$ and $Reuters8$ and 78% for $Cade12$ in kNN accuracies, underlining the quality of dimensionality reduction achieved.

(#3) The varying accuracies across datasets indicate their relative difficulties, with $20News$ in between $Reuters8$ (the least difficult) and $Cade12$ (the most difficult).



a. 20*News* b. *Reuters*8 c. *Cade*12

Figure 3: Visualization Quality: Vary Number of Topics Z



a. 20*News* (Z = 30) b. *Reuters*8 (Z = 20) c. *Cade*12 (Z = 30)

Figure 4: Visualization Quality: Vary Number of Neighbors k

In terms of SSE's comparison to baselines: (#1) SSE has significantly higher accuracies than PLSV (the main baseline). In relative terms, SSE improves upon PLSV's accuracy by 30–48% on 20$News$, by 12–16% on $Reuters$8, and by 22–36% on $Cade$12. This indicates that *spherical representation* of word space helps to improve the visualization.

(#2) SSE outperforms the visualization method t-SNE that also takes in *tf-idf* vectors as input. t-SNE's accuracy is not affected by the number of topics. A direct visualization technique, t-SNE is competitive, outperforming the other baselines for 20$News$ and $Cade$12. However, it performs worst for $Reuters$8 (more on this later). SSE shows significantly higher accuracies than t-SNE in the majority of cases (except for very low number of topics $Z = 10$), with improvements up to 14% on 20$News$, 54% on $Reuters$8, and 14% on $Cade$12. Since t-SNE shares the spherical representation of documents but does not model topics, the outperformance by SSE could be attributed in part to the approach of *modeling topics* with visualization.

(#3) SSE also outperforms PE (SAM) by a large margin. Since SSE and SAM share a spherical representation of topics in the word space, this outperformance by SSE can be attributed to *jointly* modeling topics and visualization. This is further suported by how PLSV (which also jointly models topics and visualization) outperforms PE (LDA), even as they share multinomial modeling of topic words.

Vary Number of Neighbors k. In Figure 4 we investigate the effects of different neighborhood size k's at specific settings of topics ($Z = 30$ for 20$News$, $Z = 20$ for $Reuters$8, and $Z = 30$ for $Cade$12). These are Z settings where SSE performs best, but similar observations can be drawn for other Z settings. The focus here is on the number of neighbors, rather than on the relative comparison against the baselines again, so we apply the same Z for all methods.

(#1) As k increases from 10 to 50, the $accuracy(k)$ tends to decrease. This is expected because a small k is very conservative, where we are only concerned with the immediate neighbors, which tend to be very similar. As k increases, the neighborhood considered in the kNN is larger, with a higher chance of having neighbors of a different class.

(#2) The gradients of the decrease vary among methods. Most methods, such as SSE, are relatively stable. This stability across different k's is a good sign, indicating that documents of the same class are placed in the same general locality. The most affected is t-SNE, with the greatest difference in accuracies between $k = 10$ and $k = 50$. The sharp difference indicates that t-SNE may splinter documents of the same class into several clusters in different localities, such that neighbors at low k are still of the same class, but neighbors at higher k are of different classes. This is indeed the case, as seen in the qualitative comparison (Section 4.5).

In summary, the experiments show that SSE overall produces a significant gain in visualization quality over the baselines, as measured in terms of its accuracy in kNN classification with coordinates as features.

4.4 Topic Interpretability

We also investigate whether the gain in visualization comes at the expense of the topic model. We compare SSE with baselines PLSV, LDA, and SAM in terms of topic model.

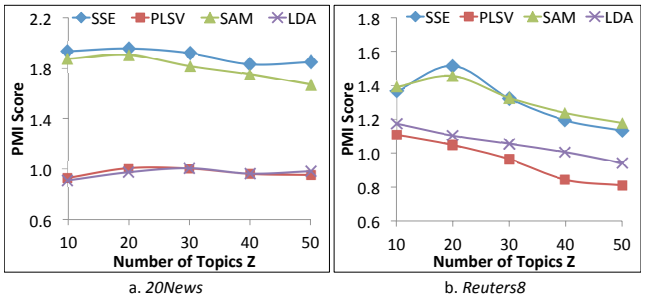


Figure 5: Topic Interpretability (PMI Score)

Metric. There are several evaluation methods for topic models proposed in the literature. One is perplexity [5], which measures the log-likelihood on unseen test data. Perplexity is *intrinsic*, i.e., dependent on the specific probability model, and may be inappropriate when comparing models with drastically different probability models, e.g., PLSV or LDA that uses multinomial models, versus SSE or SAM that uses vMF distributions. We thus need an *extrinsic* evaluation that compares these models using external validation.

In our setting, interpretability is important, because the topic model serves to provide semantics to the visualization of the data at hand. To human subjects, interpretability is closely related to coherence [28], i.e., how much the top keywords in each topic are "associated" with each other. After an extensive study of evaluation methods for coherence, [28] identifies Pointwise Mutual Information (PMI) as the best measure, in terms of having the greatest correlation with human judgments. We therefore adopt PMI as a metric.

PMI is based on term cooccurrences. For a pair of words w_i and w_j, PMI is defined as $\log \frac{p(w_i,w_j)}{p(w_i)p(w_j)}$. For a topic, we average the pairwise PMI's among the top 10 words of that topic. For a topic model, we average PMI across the topics. Intuitively, PMI is higher (better), if each topic features words that are highly correlated with one another.

Key to PMI is the use of an external corpus to estimate $p(w_i, w_j)$ and $p(w_i)$. Following [27], we use *Google Web 1T 5-gram Version 1* [6], a corpus of n-grams generated from 1 trillion word tokens. $p(w_i)$ is estimated from the frequencies of 1-grams. $p(w_i, w_j)$ is estimated from the frequencies of 5-grams, as recommended in [27]. We show the PMI for the English-based $20News$ in Figure 5(a) and $Reuters8$ in Figure 5(b). $Cade12$ is not included because we do not possess a large-scale n-gram corpus for Brazilian Portuguese.

Vary Number of Topics Z. From Figure 5, we draw the following observations on topic interpretability. **(#1)** SSE outperforms PLSV, and SAM outperforms LDA, in terms of PMI scores, across various topic settings, on $20News$ and $Reuters8$. It indicates that spherical models (SSE and SAM) produce topics that are more coherent and interpretable than multinomial models (PLSV and LDA). This is consistent with the conclusion reached in [30], which conducts an evaluation of coherence using human judges. This concurrence helps to show that our automatic evaluation on an external corpus is consistent with human judgments.

(#2) SSE performs similarly to SAM, with slightly higher PMI scores on $20News$, but comparable scores on $Reuters8$. This can be explained by their common modeling of topics in the spherical space. Since SSE also needs to deal with visualization constraints, it is notable that the gains in visualization quality have not hurt, and have even sometimes helped the topic model.

(#3) PLSV performs similarly to LDA on $20News$, but slightly worse on $Reuters8$, which is not surprising since they both share a similar multinomial modeling of topics but PLSV also faces constraints to fit the visualization task.

In summary, the experiments show that by incorporating spherical representation, SSE's significant gain in visualization does not come at the expense of the topic model.

4.5 Qualitative Comparison

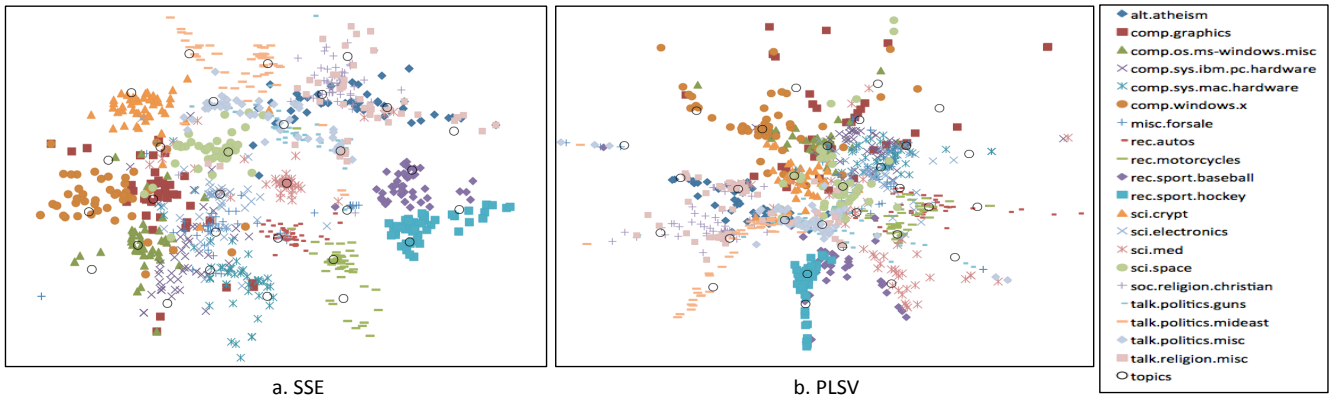
To gain a sense of the visualization quality, we show example visualization outputs for $20News$ and $Reuters8$. $Cade12$ is not shown here due to space constraint.

20News. The visualizations for $20News$ are shown in Figure 6 for $Z = 30$ (best viewed in color). Each document has a coordinate in the scatterplot. To aid identification, documents are drawn with a colored marker based on their class (see legend). Topics are drawn as black, hollow circles.

SSE's visualization in Figure 6(a) shows better separation of different classes. For instance, there are distinct blue cluster and purple cluster on the right for *rec.sport.hockey* and *rec.sport.baseball* classes respectively, green and red clusters on the lower right for *rec.motorcycles* and *rec.autos*, etc. Interestingly, not only are documents of the same class placed nearby, but related classes are also neighboring one another, with recreational classes *rec.** on the lower right, computer classes *comp.** on the lower left, science clases *sci.** at the center and upper left, while classes related to politics and religion are on the upper right. Comparatively, PLSV in Figure 6(b) suffers from greater crowding at the mid-section, while t-SNE in Figure 6(c) from splintering of some classes into multiple clusters (e.g., pink square documents denoting *talk.politics.mideast*). PE (LDA) in Figure 6(d) and PE (SAM) in in Figure 6(e) are weaker. The relative ranking in visualization quality largely mirrors the earlier finding on quantitative accuracy, with SSE being the best, followed by t-SNE, PLSV, and then the two PE approaches.

To show that the SSE's visualization is backed by a good topic model, we show some topic words in Table 3. One property of spherical representation is that each topic may have both positive and negative words. We show the five most positive words, and the five most negative words. Only 10 topics out of 30 are shown due to space constraint. Looking at the positive words, we see that the topics cover some classes very well, such as hockey, motorcycle, car, windows, apple, christianity, religion, middle eastern politics, medicine, and space. Looking at the negative words, we see that the topics also define what classes they are not. Topic 0 is about hockey, and not baseball. Topic 1 is about motorcycles, and not cars. Topic 3 is about software, and not hardware.

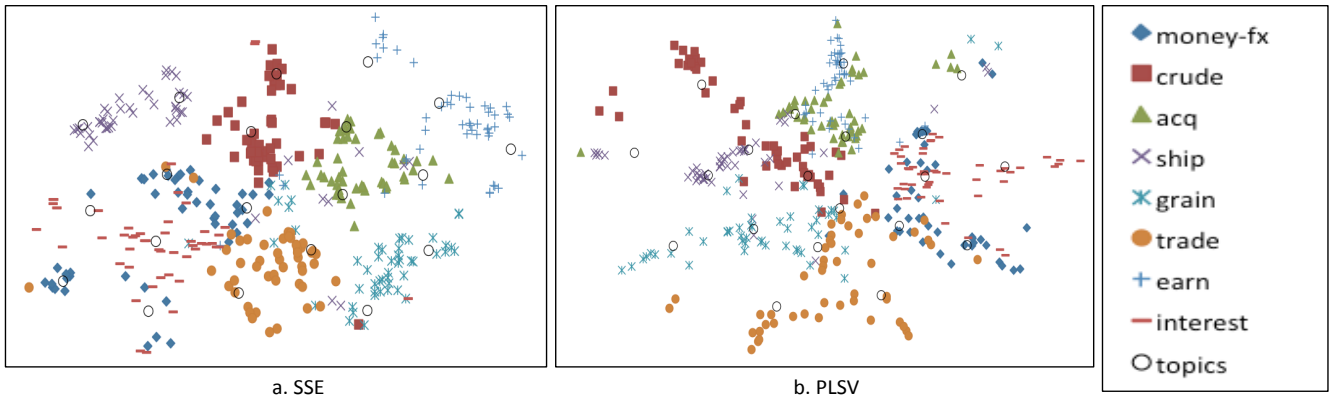
Reuters8. The visualizations for $Reuters8$ are shown in Figure 7 for $Z = 20$. Generally, it is an easier dataset, and most methods perform better than for $20News$. Comparatively, SSE still produces the clearest separation between classes, and similar observations apply as before. In particular, the splintering issue with t-SNE is even clearer in Figure 7(c). For instance, the *money-fx* class is splintered into two navy-blue diamond clusters (lower left and upper right) separated by other classes. This helps to explain why t-SNE performs even worse (in relative terms) on $Reuters8$ than on other datasets.



a. SSE b. PLSV

c. t-SNE d. PE (SAM) c. PE (LDA)

Figure 6: Visualization of $20News$ for $Z = 30$ topics (best viewed in color)



a. SSE b. PLSV

c. t-SNE d. PE (SAM) c. PE (LDA)

Figure 7: Visualization of $Reuters8$ for $Z = 20$ topics (best viewed in color)

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
5 Most Positive Weights									
hockey	bike	car	window	apple	jesus	god	israel	doctor	space
team	dod	engine	software	sale	christ	religion	israeli	patient	launch
cup	motorcycle	mile	product	monitor	christian	truth	arab	treatment	moon
playoff	ride	ford	price	computer	god	belief	jew	medicine	flight
nhl	rider	mustang	user	price	sin	existence	jewish	symptom	nasa
5 Most Negative Weights									
pitch	ford	circuit	video	scsi-2	scholar	reporter	encryption	objective	algorithm
pitcher	detector	amp	bus	scsi-1	addition	government	armenian	religion	file
inning	oort	board	slot	burst	wingate	livesey	algorithm	jew	driver
bullpen	sensor	lady	wiretap	scsi	sea	corruption	science	key	nice
giant	firearm	1983	ide	16-bit	livesey	theological	armenia	god	motorcycle

Table 3: Positive and Negative Words in Each Topic for $20News$ by SSE for $Z = 30$ (a selection of 10)

5. CONCLUSION

In this work, we address the problem of semantic visualization that jointly models visualization and topics. Our model, Spherical Semantic Embedding or SSE is designed for data with spherical representation, i.e., L^2-normalized term vectors. Its generative model associates each document with a triplet of representations, namely: a coordinate in the Euclidean visualization space, a multinomial topic distribution in the topic space, as well as a normalized term vector in the spherical word space. Comprehensive experiments on benchmark datasets show that SSE shows significantly improved performance when compared to existing state-of-the-art baselines in terms of visualization quality, as well as topic interpretability.

6. REFERENCES

[1] A. Banerjee, I. S. Dhillon, J. Ghosh, and S. Sra. Clustering on the unit hypersphere using von Mises-Fisher distributions. In *JMLR*, 2005.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6), 2003.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[4] C. M. Bishop, M. Svensén, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1), 1998.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3, 2003.

[6] T. Brants and A. Franz. Web 1T 5-gram Version 1. 2006.

[7] M. D. Buhmann. Radial basis functions. *Acta Numerica 2000*, 9, 2000.

[8] A. Cardoso-Cachopo. Improving Methods for Single-label Text Categorization. PdD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa, 2007.

[9] A. J.-B. Chaney and D. M. Blei. Visualizing topic models. In *ICWSM*, 2012.

[10] J. Chuang, C. D. Manning, and J. Heer. Termite: visualization techniques for assessing textual topic models. In *AVI*, 2012.

[11] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3), 1994.

[12] L. V. der Maaten and G. Hinton. Visualizing data using t-SNE. *JMLR*, 9, 2008.

[13] W. Dou, X. Wang, R. Chang, and W. Ribarsky. ParallelTopics: A probabilistic approach to exploring document collections. In *VAST*, 2011.

[14] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2), 1936.

[15] B. Gretarsson, J. O'donovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth. TopicNets: Visual analysis of large text corpora with topic modeling. *TIST*, 3(2), 2012.

[16] G. E. Hinton and S. T. Roweis. Stochastic neighbor embedding. In *NIPS*, 2002.

[17] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, 1999.

[18] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. L. Griffiths, and J. B. Tenenbaum. Parametric embedding for class visualization. *Neural Computation*, 19(9), 2007.

[19] T. Iwata, T. Yamada, and N. Ueda. Probabilistic latent semantic visualization: topic model for visualizing documents. In *KDD*, 2008.

[20] I. Jolliffe. *Principal Component Analysis*. Wiley Online Library, 2005.

[21] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9), 1990.

[22] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1), 1964.

[23] T. M. V. Le and H. W. Lauw. Manifold learning for jointly modeling topic and visualization. In *AAAI*, 2014.

[24] K. V. Mardia. Distribution theory for the von Mises-Fisher distribution and its application. In *A Modern Course on Statistical Distributions in Scientific Work*. Springer, 1975.

[25] K. V. Mardia and P. E. Jupp. *Directional Statistics*, volume 494. Wiley.com, 2009.

[26] J. R. Millar, G. L. Peterson, and M. J. Mendenhall. Document clustering and visualization with latent dirichlet allocation and self-organizing maps. In *FLAIRS*, 2009.

[27] D. Newman, S. Karimi, and L. Cavedon. External evaluation of topic models. In *ADCS*, 2009.

[28] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic coherence. In *NAACL HLT*, 2010.

[29] A. Perina, N. Jojic, M. Bicego, and A. Truski. Documents as multiple overlapping windows into grids of counts. In *NIPS*, 2013.

[30] J. Reisinger, A. Waters, B. Silverthorn, and R. J. Mooney. Spherical topic models. In *ICML*, 2010.

[31] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290, 2000.

[32] G. Salton, A. Wong, and C.-S. Yang. A vector space model for automatic indexing. *CACM*, 18(11), 1975.

[33] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, 2000.

[34] F. Wei, S. Liu, Y. Song, S. Pan, M. X. Zhou, W. Qian, L. Shi, L. Tan, and Q. Zhang. Tiara: a visual exploratory text analytic system. In *KDD*, 2010.