Journal of Artificial Intelligence Research 55 (2016) 1091-1133 Submitted 09/2015; published 04/2016

Semantic Visualization with Neighborhood Graph Regularization

Tuan M. V. Le vmtle.2012@phdis.smu.edu.sg
Hady W. Lauw hadywlauw@smu.edu.sg
School of Information Systems
Singapore Management University
80 Stamford Road, Singapore 178902

Abstract

Visualization of high-dimensional data, such as text documents, is useful to map out the similarities among various data points. In the high-dimensional space, documents are commonly represented as bags of words, with dimensionality equal to the vocabulary size. Classical approaches to document visualization directly reduce this into visualizable two or three dimensions. Recent approaches consider an intermediate representation in topic space, between word space and visualization space, which preserves the semantics by topic modeling. While aiming for a good fit between the model parameters and the observed data, previous approaches have not considered the local consistency among data instances. We consider the problem of semantic visualization by jointly modeling topics and visualization on the intrinsic document manifold, modeled using a neighborhood graph. Each document has both a topic distribution and visualization coordinate. Specifically, we propose an unsupervised probabilistic model, called Semafore, which aims to preserve the manifold in the lower-dimensional spaces through a neighborhood regularization framework designed for the semantic visualization task. To validate the efficacy of Semafore, our comprehensive experiments on a number of real-life text datasets of news articles and Web pages show that the proposed methods outperform the state-of-the-art baselines on objective evaluation metrics.

1. Introduction

Text documents come in various flavors, such as Web pages, news articles, blog posts, emails, or messages on social media such as Twitter. While much is in English, there are increasing amounts of content in various languages as well. With the backdrop of the growth in volume, diversity, and complexity of various corpora, we need more useful tools to analyze the wealth of text content. One form of analysis which we will look into in this paper is visualization. There are different types of visualizations, be it of the temporal or longitudinal, networked, or other natures. What we are interested in is a form of visualization where we can represent a collection of documents as coordinates on the same low-dimensional space, so as to learn of the similarities and differences among documents based on their distances on the visualization space.

Visualization of high-dimensional data is an important exploratory data analysis task, which is actively studied by various academic communities. While the HCI community is interested in the presentation of information, as well as other interface aspects (Chi, 2000), the machine learning community is interested in the quality of dimensionality reduction

(Van der Maaten & Hinton, 2008), i.e., how to transform the high-dimensional representation into a lower-dimensional representation that can be shown on a scatterplot. This visualization form is simple, and widely applicable across various domains.

Consider therefore the problem of visualizing documents on a scatterplot. Commonly, a document is represented as a bag of words, i.e., a vector of word counts. This high-dimensional representation would be reduced into coordinates on a visualizable 2D (or 3D) space. One pioneering technique is Multidimensional Scaling (MDS) (Kruskal, 1964). The goal is to preserve the *distances* in the high-dimensional space in the low-dimensional embedding. When applied to documents, a visualization technique for generic high-dimensional data, e.g., MDS, may not necessarily preserve the topical semantics. Words are often ambiguous, with issues such as *polysemy*, when the same word carries multiple senses, and *synonymy*, when different words carry the same sense. Because the dimensions in the original representation (which are words) may not accurately capture this ambiguity, this affects the quality of the reduced representation (which is the visualization space) as well.

To model semantics in documents in a way that can resolve some of this ambiguity, the current popular approach is by topic modeling, such as PLSA (Hofmann, 1999) or LDA (Blei, Ng, & Jordan, 2003). Each document is associated with a probability distribution over a set of topics. Each topic is a probability distribution over words in the vocabulary. In this way, polysemous words can be separated into different topics, and synonymous words can be grouped into the same topic.

Topic modeling itself is another form of dimensionality reduction: from word space to topic space. The word space refers to a document's original representation, which is usually a bag of words. The topic space refers to the simplex of topic distributions. A document's probability distribution over topics is effectively the representation of this document in this topic space. However, a topic model by itself is not designed for visualization. While one possible visualization is to plot documents' topic distributions on a simplex, a 2D visualization space could express only three topics, which is very limiting.

Given its success in modeling semantics in documents, we therefore ask the question of whether and how best to do both forms of dimensionality reductions (visualization and topic modeling) for documents. The end goal is to arrive at a visualization of documents that is consistent with both the semantic representation (topics), as well as the original representation (words). This coupling is a distinct task from topic modeling or visualization respectively, as it enables novel capabilities. For one thing, topic modeling helps to create a richer visualization, as we can now associate each coordinate on the visualization space with both topic and word distributions, providing semantics to the visualization space. For another, the tight integration potentially allows the visualization to serve as a way to explore and tune topic models, allowing users to introduce feedback (Hu, Boyd-Graber, Satinoff, & Smith, 2014) to the model through a visual interface (Choo, Lee, Reddy, & Park, 2013). These capabilities support several use case scenarios. One potential use case is a document organizer system. The visualization could potentially help in assigning categories to documents, by showing how closely related documents have been labeled. Another is an augmented retrieval system. Given a query, the results may include not just relevant documents, but also other similar documents (neighbors in the visualization).

1.1 Problem Statement

We refer to the task of jointly modeling topics and visualization as *semantic visualization*. The input is a set of documents $\mathcal{D}$. For a specified number of topics Z and visualization dimensionality (assumed to be 2D, without losing any generality), the goal is to derive, for every document in $\mathcal{D}$, a latent coordinate on the visualization space, and a probability distribution over the Z topics. While we focus on documents in our description, the same approach would apply to visualization of other data types for which latent factor modeling, i.e., topic model, makes sense.

A straightforward way is to undergo two-step reductions. In the first reduction, the original representation for documents are reduced into topic distributions using topic modeling. In the second reduction, documents' topic distributions are further reduced into visualization coordinates. This approach may have some value compared to direct reduction from word space to visualization space. However, it is not ideal, because the disjoint reductions could mean that errors may propagate from the first to the second reduction, and the resulting visualization may not faithfully capture the original representation.

A better way to solve this problem is to join up the two reductions into a single, joint process that produces both topic distributions and visualization coordinates. This approach was first pioneered by PLSV (Iwata, Yamada, & Ueda, 2008), which also showed that the joint approach outperformed the disjoint approach. PLSV derives the latent parameters by maximizing the likelihood of observing the documents. This goal is concerned with the "error" between the model and the observation.

In the literature, it is found that algorithms that ensure "smoothness" tend to perform better at learning tasks (Zhou, Bousquet, Lal, Weston, & Schölkopf, 2004). Smoothness concerns preserving the observed proximity between documents. This objective arises naturally from the assumption that the intrinsic geometry of the data is a low-rank, non-linear subspace within the high-dimensional space. Therefore, preserving neighborhood structure is important for learning tasks. This assumption is well-accepted in the machine learning community (Lafferty & Wasserman, 2007), and finds application in both supervised and unsupervised learning (Belkin & Niyogi, 2003; Zhou et al., 2004; Zhu, Ghahramani, Lafferty, et al., 2003). Recently, there is a preponderance of evidence that this assumption also applies to text data in particular (Cai, Mei, Han, & Zhai, 2008; Cai, Wang, & He, 2009; Huh & Fienberg, 2012). We therefore propose to incorporate this assumption into a new unsupervised, semantic visualization model.

1.2 Overview

We propose an unsupervised probabilistic model that jointly derives topic distributions and visualization coordinates on the intrinsic geometry of the data. Our proposed model is called Semafore, which stands for SEmantic visualization with MAniFOld REgularization. We build a neighborhood regularization framework into a semantic visualization model. The framework involves new issues to resolve, including the regularization function, and the space in which regularization should take place.

The model is evaluated on a series of real-life, publicly available datasets, which are also benchmark datasets used in document classification task. An advantage of a statistical method, such as ours, is that it is not dependent on a specific language. Two of the datasets are in English, and one is in Brazilian Portuguese. While our model is unsupervised (class label is neither required nor used in learning), to objectively quantify the visualization quality, we leverage on the class label information. It is a common assumption that documents of the same class are expected to be neighbors on the original space (Belkin, Niyogi, & Sindhwani, 2006; Zhou et al., 2004; Zhu et al., 2003), which suggests that they should also be close on the visualization space. We investigate the effectiveness of Semafore in placing documents of the same class nearby on the visualization space, and systematically compare it to existing baselines without one or more of our properties, namely: joint modeling of topic and visualization, or neighborhood regularization.

1.3 Contributions

While visualization and topic modeling are, separately, well-studied problems, the interface between the two, semantic visualization, is a relatively new problem, with very few previous work. In this work, we make the following contributions.

- We propose incorporating neighborhood structure in semantic visualization. In this respect, we propose a probabilistic model Semafore, with two integrated components. One is a kernelized semantic visualization model, enabling the substitution of the kernel functions that relate visualization coordinates to topic distributions (see Section 3.3). The other is a neighborhood graph regularization framework for semantic visualization as described in Section 4.1.
- Realizing the neighborhood graph regularization involves an exploration of how to incorporate the appropriate forms of the neighborhood structure. In this respect, we investigate the effects of neighborhood graph construction techniques such as k-nearest neighbors (k-NN), ϵ-ball, and disjoint minimum spanning trees (DMST), as well as different edge weight estimations such as heat-kernel (see Section 4.2) in the context of semantic visualization.
- In Section 5, we describe the requisite learning algorithms based on maximum a posteriori (MAP) estimation using expectation-maximization (EM), in order to fit the parameters for the various regularization functions and kernels that we propose.
- Our final contribution is the evaluation of Semafore's effectiveness on a series of real-life, public datasets described in Section 6, which shows that Semafore outperforms existing baselines on a well-established and objective visualization metric.

In our prior work (Le & Lauw, 2014b), we proposed the problem and described the preliminary model. In this extended article, there are significant technical changes that provide a significantly more comprehensive discussion of the model. For instance, we now discuss the Student-t kernel, in addition to the previously introduced Gaussian kernel. Furthermore, we investigate the efficacies of different neighborhood graph constructions, including the ϵ-ball and DMST graphs, in addition to the previously introduced kNN graph. The graph weights are also enhanced through investigation of heat kernel, in addition to the simple-minded binary scheme previously. As discussed in Section 6.3, these enhancements collectively result in statistically significant improvements over the previous model. Beyond the technical enhancements, we also provide more comprehensive model analysis and empirical validation, including richer quantitative and qualitative discussions of the visualizations and the resulting topic models, as well as a metric to measure topic interpretability based on pairwise mutual information.

2. Related Work

In this section, we discuss the different aspects of our work, identify the related papers in the literature, and point out the key conceptual differences.

2.1 Visualization and Dimensionality Reduction

One way to perform visualization is by using a *generic* dimensionality reduction technique. Such techniques come in several flavors, depending on the objective. Principal component analysis (PCA) (Jolliffe, 2005) identifies the components that explain most of the variance in the data. Related to PCA is singular value decomposition (SVD) (Golub & Van Loan, 2012). Comparatively, independent component analysis (ICA) (Comon, 1994) identifies the components that are independent of one another, whereas linear discriminant analysis (Fisher's LDA) (Fisher, 1936) identifies the components that most discriminate between known class labels. Being generic, these techniques are more frequently applied to feature extraction, as they are not optimized for visualization. They focus more on the properties of the components (e.g., orthogonality, independence) rather than on the intrinsic relationship among data instances. Furthermore, as they are based on linear projections, they may not capture non-linearities in the data well.

Another category of techniques, which is more directly related to visualization, is the *embedding* approach. It aims to preserve the high-dimensional similarities or differences in the low-dimensional embedding. One pioneering such work is multidimensional scaling (MDS) (Kruskal, 1964). Given a set of pairwise distances δ_{ij} between data points i and j, MDS determines coordinates x_i and x_j respectively, such that the embedded visualization distance $||x_i - x_j||$ approximates δ_{ij} as much as possible. For MDS, the distance to be preserved δ_{ij} is frequently the linear distance, measuring the distance along a straight line between two points in the input space. Instead of this linear distance, Isomap (Tenenbaum, De Silva, & Langford, 2000) seeks to preserve the geodesic distance, by finding shortest paths in a graph with edges connecting neighboring data points. LLE (Roweis & Saul, 2000) seeks to preserve linear distances, but only among the neighboring points and avoiding the need to estimate pairwise distances between widely separated data points. Recently there are also works applying a similar concept to embedding but using probabilistic modeling, such as PE (Iwata, Saito, Ueda, Stromsten, Griffiths, & Tenenbaum, 2007), SNE (Hinton & Roweis, 2002), t-SNE (Van der Maaten & Hinton, 2008), and GTM (Bishop, Svensén, & Williams, 1998). Yet others are based on semi-definite programming (Shaw & Jebara, 2007, 2009). Alternatively, several embedding techniques do not aim to preserve relationship among data instances, but rather other properties such as local minima (Kim & Torre, 2010). Importantly, all these techniques are not optimized for *semantic* visualization, as they do not model topics at all. The coordinates do not reflect any semantic meaning, other than reflecting the optimization objective.

There are only a few related works so far that seek to address the semantic visualization task directly. The closest previous work that does both topic modeling and visualization in a single generative process is Probabilistic Latent Semantic Visualization (PLSV) (Iwata et al., 2008), which also shows that a joint approach outperforms a separate approach. Just as PLSV builds upon the foundation of the topic modeling technique Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) by incorporating visualization coordinates, so do we build upon the foundation of PLSV by incorporating RBF kernels (Section 3.3) and neighborhood structure (Section 4).

There are also related works that share a similar objective, but do not share the same paradigm of visualization or topic modeling. For instance, LDA-SOM (Millar, Peterson, & Mendenhall, 2009) first conducts topic modeling using Latent Dirichlet Allocation (LDA) (Blei et al., 2003), and then separately embeds the documents' topic distributions on a Self-Organizing Map (SOM) (Kohonen, 1990). However, this is not a joint model, and SOM uses a different visualization space than the Euclidean space that we are interested in. For another instance, SSE (Le & Lauw, 2014a) builds on the Spherical Admixture Model (SAM) (Reisinger, Waters, Silverthorn, & Mooney, 2010) belonging to the class of spherical topic models targeted at spherical (unit vector) reprepresentations of topics and documents, which are not directly comparable or equivalent with the simplex representation and multinomial modeling (probability distribution over words) adopted in this work as well as PLSV.

By semantic visualization, we refer to the task of joining visualization and topic modeling. A related, but different, task is topic visualization, where the objective is to visualize the topics, in terms of which keywords are dominant for each topic (Chaney & Blei, 2012; Chuang, Manning, & Heer, 2012), which topics are dominant in a corpus (Wei, Liu, Song, Pan, Zhou, Qian, Shi, Tan, & Zhang, 2010), and how topics are related to one another (Gretarsson, O'donovan, Bostandjiev, Höllerer, Asuncion, Newman, & Smyth, 2012).

2.2 Topic Modeling

Topic model involves statistical modeling of text (documents and words) in order to discover some abstract concepts or "topics" that occur in a corpus. Beginning with latent semantic indexing (Dumais, Furnas, Landauer, Deerwester, Deerwester, et al., 1995), topic model evolves into the modern probabilistic treatments, such as Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Intuitively, a topic captures a collection of words that tend to co-occur because they describe the same concept. This has the appeal of producing highly interpretable statistical models that let users make semantic sense of the corpus. Other than text-only document corpora, topic models have also been applied to cases where links are observed in addition to text (McCallum, Wang, & Corrada-Emmanuel, 2007).

Meanwhile, the assumption that the intrinsic geometry of the data is a non-linear low dimensional subspace within the high-dimensional space finds application in both supervised and unsupervised (Belkin & Niyogi, 2003) learning algorithms. It is especially prevalent in semi-supervised learning (Zhou et al., 2004; Zhu et al., 2003) as a way to bridge labeled and unlabeled data. Regularization as a technique to realize this assumption has a long history (Belkin et al., 2006). The specific form of the regularization function varies among applications. The study of this assumption for unsupervised topic models begins with LapPLSI (Cai et al., 2008), which introduces regularization to PLSA (Hofmann, 1999), by minimizing the Euclidean distance between neighboring documents' topic distributions. Follow-up work introduce other distance functions (Cai et al., 2009; Wu, Bu, Chen, Zhu, Zhang, Liu, Wang, & Cai, 2012). While these previous work focus on maintaining proximity of similar documents, DTM (Huh & Fienberg, 2012) adds a new criterion to also maintain the distance among different documents. Our work is different in that we also need to contend with the visualization aspects, and not just topic modeling.

2.3 Semantic Similarity

Other than topic models, there are alternative mechanisms to learn the semantic relationship between documents. One way is by measuring the semantic similarity among documents or words. For instance, in vector space model, documents may be represented as a term vector, and their similarity may be expressed in terms of cosine similarity (Turney, Pantel, et al., 2010). Other than word occurrences alone, there could also be additional signals of semantic similarity. For instance, working with Wikipedia corpus, the categories and links are also took into account to determine the similarity among articles (Gabrilovich & Markovitch, 2009; Ponzetto & Strube, 2007). Our work differs from these in several important respects. First, our objective is not in the similarity value per se, but rather in determining lower-dimensional embedding coordinates, which would allow visualization as one application. Second, our method is based on probabilistic modeling of latent variables, akin to topic modeling, instead of operating on the vector space model representation of documents.

3. Semantic Visualization

We introduce the problem formulation for semantic visualization in Section 3.1. Our focus in this paper is on the effects of the neighborhood graph structure on the semantic visualization task. We figure that the clearest way to showcase these effects is to design a neighborhood preservation framework over and above an existing generative process, such as PLSV (Iwata et al., 2008), which we will review in Section 3.2. In Section 3.3, we describe an innovation over the semantic visualization model, which is an abstraction of the mapping between the topic space and the visualization space using radial basis function (RBF) kernels. This allows the exploration of various kernels, of which we identify two for further exploration. For ease of following the discussion, we include a table of notations in Table 1.

3.1 Problem

For the task of semantic visualization, the input is a corpus of documents $\mathcal{D} = \{d_1, \ldots, d_N\}$. Every d_n is a bag of words, and w_{nm} denotes the m^{th} word in d_n. The total number of words in d_n is M_n. The objective is to learn, for each d_n, a latent distribution over Z topics $\{\mathrm{P}(z|d_n)\}_{z=1}^{Z}$. Each topic z is associated with a parameter θ_z, which is a probability distribution $\{\mathrm{P}(w|\theta_z)\}_{w \in W}$ over words in the vocabulary W. The words with the highest probabilities for a given topic capture the semantic of that topic.

Notation	Description
d_n	a specific document
x_n	latent coordinate of d_n in the visualization space
M_n	number of words in document d_n
z	a specific topic
ϕ_z	coordinate of topic z in the visualization space
θ_z	word distribution of topic z
W	the vocabulary (the set of words in the lexicon)
N	total number of documents in the corpus
Z	total number of topics (user-defined)
χ	the collection of x_n's for all documents
Φ	the collection of ϕ_z's for all topics
Θ	the collection of θ_z's for all topics
Ψ	the collective set of parameters $\{\chi, \Phi, \Theta\}$

Table 1: Notations.

In semantic visualization, there is an additional objective for semantic visualization, which is to learn, for each document d_n, its latent coordinate x_n on a low-dimensionality visualization space. Similarly, each topic z is associated with a latent coordinate ϕ_z on the visualization space. A document d_n's topic distribution is then expressed in terms of the Euclidean distance between its coordinate x_n and the different topic coordinates $\Phi = \{\phi_z\}_{z=1}^{Z}$. Intuitively, the closer is x_n to a topic's ϕ_z, the higher is $\mathrm{P}(z|d_n)$ or the probability of topic z for document d_n.

In the following sections, we systematically describe the various components of our solution. The generative process that links the latent variables (coordinates) and the words in the documents is described in Section 3.2. The specific relationship between documents and topics' coordinates constitutes a specific mapping function, which we model as an RBF kernel in Section 3.3. In the following Section 4, we discuss how to incorporate neighborhood structure into semantic visualization.

3.2 Generative Process

We now describe the generative process of documents based on both topics and visualization coordinates. Below we review PLSV whose graphical model is shown in Figure 1. Our eventual complete model is a generalization of this model, involving enhancements through kernelization (Section 3.3) and neighborhood structure preservation (Section 4).

The generative process is as follows:

1. For each topic $z = 1, \ldots, Z$:
   (a) Draw z's word distribution: $\theta_z \sim \text{Dirichlet}(\alpha)$
   (b) Draw z's coordinate: $\phi_z \sim \text{Normal}(0, \beta^{-1}I)$
2. For each document d_n, where $n = 1, \ldots, N$:
   (a) Draw d_n's coordinate: $x_n \sim \text{Normal}(0, \gamma^{-1}I)$

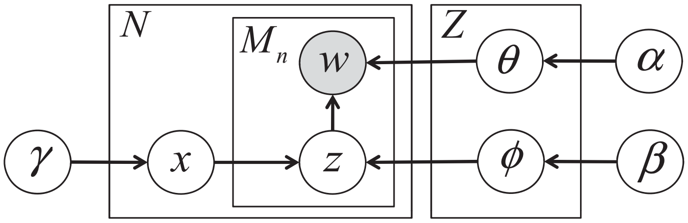


Figure 1: Graphical model of PLSV.

(b) For each word $w_{nm} \in d_n$:

i. Draw a topic: $z \sim \text{Multi}(\{\text{P}(z|x_n, \Phi)\}_{z=1}^{Z})$

ii. Draw a word: $w_{nm} \sim \text{Multi}(\theta_z)$

Here, α is a Dirichlet prior, I is the identity matrix, β and γ control the variance of the Normal distributions. The parameters $\chi = \{x_n\}_{n=1}^{N}$, $\Phi = \{\phi_z\}_{z=1}^{Z}$, $\Theta = \{\theta_z\}_{z=1}^{Z}$, collectively denoted as $\Psi = \langle \chi, \Phi, \Theta \rangle$, are learned from documents $\mathcal{D}$ based on maximum a posteriori estimation. The log likelihood function is shown in Equation 1.

$$\mathcal{L}(\Psi|\mathcal{D}) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log \sum_{z=1}^{Z} \text{P}(z|x_n, \Phi)\text{P}(w_{nm}|\theta_z) \tag{1}$$

We reiterate that our focus here is on incorporating neighborhood graph structure into semantic visualization. By building a neighborhood graph regularization framework into an existing generative process, i.e., PLSV, we can clearly observe that any improvement over PLSV arises from the neighborhood graph regularization. In this sense, our work is in the tradition of introducing neighborhood graph regularization to probabilistic topic modeling (Huh & Fienberg, 2012; Cai et al., 2008, 2009), where the contributions relate to the neighborhood graph regularization, rather than the generative process. That said, there is one significant difference to PLSV, which is our flexibility in allowing various kernel functions, which we will discuss next.

3.3 RBF Kernels

In the Step 2(b)i of the above generative process, the topic z of a word is drawn from the distribution $\{\text{P}(z|x_n, \Phi)\}_{z=1}^{Z}$. This distribution relates the coordinates of topics in the visualization space $\Phi = \{\phi_z\}_{z=1}^{Z}$ and the coordinate x_n of a document d_n with the document's topic distribution $\{\text{P}(z|d_n)\}_{z=1}^{Z}$.

This relationship can be formulated as a mapping problem where we want to find a function $\mathcal{G}$ which maps a point in visualization space to a point in the topic space. However, the form of $\mathcal{G}$ cannot be known exactly because both visualization space and topic space are latent spaces and $\mathcal{G}$ may be different across different domains. Therefore, to compute the topic distributions, we need a way to approximate $\mathcal{G}$.

To build a function approximation of the unknown function $\mathcal{G}$, we use the abstraction of Radial Basis Function (RBF) neural networks (Bishop, 1995) because feedforward multilayered RBF neural networks with one hidden layer can serve as a universal approximator

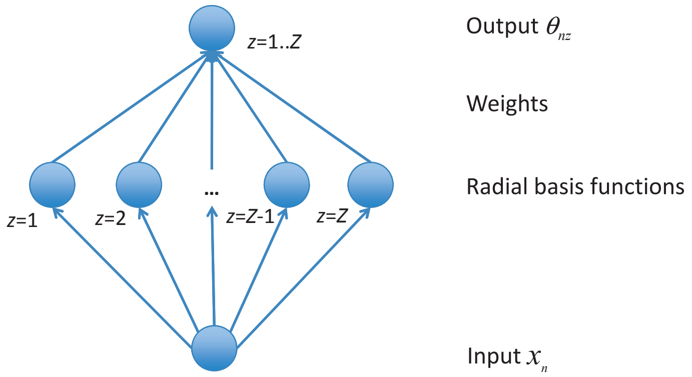


Figure 2: Topic distribution is expressed as a function of visualization coordinates using Radial Basis Function (RBF) network.

to arbitrary continuous functions (Park & Sandberg, 1991). This property provides the confidence that the model would have the ability to approximate any existing relationship between visualization space and topic space with arbitrary precision. Unlike PLSV (Iwata et al., 2008) that defined a specific mapping function, our approach generalizes the semantic visualization model by defining the mapping problem in terms of kernelization, which admits several mapping functions within the family of RBF kernels.

In our context, Radial Basis Function (Buhmann, 2000) will relate coordinate variables based on distances which defines a kernel 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 kernel function Λ may take on various forms, e.g., Gaussian, multi-quadric, inverse quadratic, polyharmonic spline. To express $\mathrm{P}(z|d_n)$ as a function of x_n, we consider the normalized architecture of RBF network, with three layers as shown in Figure 2. 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 $\mathrm{P}(z|d_n)$, which is a linear combination of the RBF functions, as shown in Equation 2. Here, $w_{z,z'}$ is the weight of influence of the RBF function of z' on the $\mathrm{P}(z|d_n)$, with the constraint $\sum_{z'=1}^{Z} w_{z,z'} = 1$.

$$\mathrm{P}(z|d_n) = y_z(x_n) = \frac{\sum_{z'=1}^{Z} w_{z,z'} \cdot \Lambda(||x_n - \phi_{z'}||)}{\sum_{z'=1}^{Z}\Lambda(||x_n - \phi_{z'}||)} \tag{2}$$

While Equation 2 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. For $w_{z,z'}$, we will experiment with a special case $w_{z,z'} = 1$ when $z = z'$ and 0 otherwise.

For the kernel function Λ, one variation we consider is Gaussian, which yields the function in Equation 3, where Φ refers to the collective set of ϕ_z's. Note that here we set variance of Gaussian to 1. However, its true value is not really important because a different variance value just produces a re-scaled visualization with the scaling factor equal to that variance.

$$\mathrm{P}(z|d_n)_{Gaussian} = \mathrm{P}(z|x_n, \Phi)_{Gaussian} = \frac{\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{Z} \exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)} \tag{3}$$

Another variation of Λ being considered is Student-t. This distribution is also used by t-SNE (Van der Maaten & Hinton, 2008) in the context of non-semantic, direct embedding to mitigate the effects of crowding. Due to mismatched dimensionalities, the points are crunched together in the center of the visualization, which prevents gaps from forming between the clusters. Therefore, we hypothesize that using Student-t as radial basis function, which yields the function in Equation 4, can help to improve the performance of our model if crowding becomes an issue. Note that the Student-t distribution with one degree of freedom yields a radial basis function having the form similar to the inverse quadratic.

$$\mathrm{P}(z|d_n)_{Student-t} = \mathrm{P}(z|x_n, \Phi)_{Student-t} = \frac{(1 + ||x_n - \phi_z||^2)^{-1}}{\sum_{z'=1}^{Z} (1 + ||x_n - \phi_{z'}||^2)^{-1}} \tag{4}$$

The Gaussian function (Equation 3) was also used previously in the baseline PLSV (Iwata et al., 2008) that we will compare to. Its inclusion helps to establish parity for comparative purposes, both to investigate the effectiveness of the alternative Student-t kernel (described above), as well as that of the neighborhood regularization (described in the next section).

4. Neighborhood Graph Regularization Framework

There are recent works (Cai et al., 2008, 2009; Huh & Fienberg, 2012) trying to preserve the local neighborhood structure when learning low-dimensional topic representations of documents. These works assume that documents are sampled from a nonlinear low-dimensional subspace that are embedded in a high-dimensional space. Therefore, the local neighborhood structure is important for revealing the hidden topics of documents and should be preserved when learning topic representations of documents (Bai, Guo, Lan, & Cheng, 2014). In the generative process for semantic visualization described in Section 3, the document parameters are sampled independently, and may not necessarily reflect the underlying local neighborhood structure. We therefore seek to realize this assumption for semantic visualization. In particular, we assume that when two documents d_i and d_j are close in the original space, then their parameters ψ_i and ψ_j of the low-rank representation are similar as well. Coupled with the kernelized semantic visualization model described in Section 3, the neighborhood preservation approach described in this section constitutes our proposed model, Semafore, which stands for SEmantic visualization with MAniFOld REgularization.

4.1 Neighborhood Regularization

The neighborhood structure can be represented by a neighborhood graph. Given a set of data points in the Euclidean space, a neighborhood graph is constructed with the input data points as vertices. By definition, edges are symmetric, i.e., $\omega_{ij} = \omega_{ji}$, and weighted. The collection of edge weights are collectively denoted as $\Omega = \{\omega_{ij}\}$.

For the moment, we will assume that we have the neighborhood graph, and address the issue of how this neighborhood graph may be incorporated into our semantic visualization framework. In actuality, the neighborhood graph construction itself is an important component, whose construction is described in detail in Section 4.2.

One effective means to incorporate a neighborhood structure into a learning model is through a regularization framework (Belkin et al., 2006). This leads to a re-design of the log-likelihood function in Equation 1 into a new *regularized* function $\mathbf{L}$ (Equation 5), where Ψ consists of the parameters (visualization coordinates and topic distributions), and $\mathcal{D}$ and Ω are the documents and neighborhood structure.

$$\mathbf{L}(\Psi|\mathcal{D}, \Omega) = \mathcal{L}(\Psi|\mathcal{D}) + \lambda \cdot \mathcal{R}(\Psi|\Omega) \tag{5}$$

The first component $\mathcal{L}$ is the log-likelihood function in Equation 1, which reflects the fit between the latent parameters Ψ and the observation $\mathcal{D}$. The second component $\mathcal{R}$ is a regularization function, which reflects the consistency between the latent parameters Ψ of neighboring documents in the neighborhood structure Ω. λ is the regularization parameter, commonly found in neighborhood based algorithms (Belkin et al., 2006; Cai et al., 2008, 2009), which controls the extent of regularization (we will experiment with different λ's in experiments).

4.1.1 Proposed Regularization Function

We now turn to the definition of the $\mathcal{R}$ function. The intuition is that the data points that are close in the high-dimensional space, should also be close in their low-rank representations, i.e., local consistency, also known as smoothness. One function that satisfies this is $\mathcal{R}_+$ in Equation 6. Here, $\mathcal{F}$ is a distance function that operates on the low-rank space. Minimizing $\mathcal{R}_+$ leads to minimizing the distance $\mathcal{F}(\psi_i, \psi_j)$ between neighbors ($\omega_{ij} = 1$).

$$\mathcal{R}_+(\Psi|\Omega) = \sum_{i,j=1; i \neq j}^{N} \omega_{ij} \cdot \mathcal{F}(\psi_i, \psi_j) \tag{6}$$

The above level of local consistency is still insufficient, because it does not regulate how *non*-neighbors (i.e., $\omega_{ij} = 0$) behave. For instance, it does not prevent *non*-neighbors from having similar low-rank representations. Another valid objective in visualization is to keep *non*-neighbors apart, which is satisfied by another objective function $\mathcal{R}_-$ in Equation 7. $\mathcal{R}_-$ is minimized when two *non*-neighbors d_i and d_j (i.e., $\omega_{ij} = 0$) are distant in their low-rank representations. The addition of 1 to $\mathcal{F}$ is to prevent division-by-zero error.

$$\mathcal{R}_-(\Psi|\Omega) = \sum_{i,j=1; i \neq j; \omega_{ij}=0}^{N} \frac{1 - \omega_{ij}}{\mathcal{F}(\psi_i, \psi_j) + 1} \tag{7}$$

We hypothesize that neither objective is effective on its own. A more complete objective would capture the spirits of both keeping neighbors close, and keeping *non*-neighbors apart. Therefore, we put Equation 6 and Equation 7 together using summation and maximize the objective function as shown in Equation 8. Note that the coefficient $\frac{1}{2}$ in Equation 8 is for simplifying the formula of the derivative of $\mathcal{R}_*(\Psi|\Omega)$.

$$\mathcal{R}_*(\Psi|\Omega) = -\frac{1}{2}(\mathcal{R}_+(\Psi|\Omega) + \mathcal{R}_-(\Psi|\Omega)) \tag{8}$$

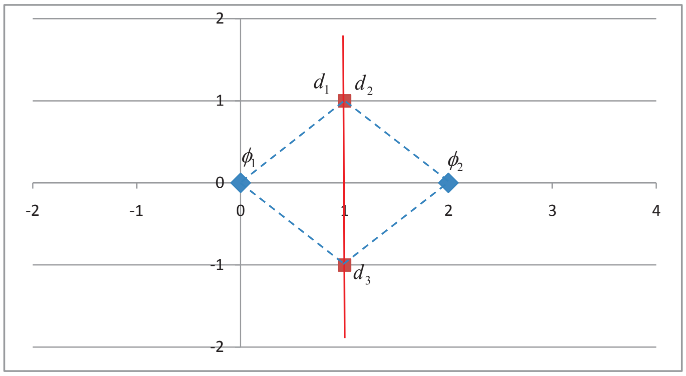


Figure 3: Example of how the same topic distribution may have different visualization coordinates. Any points on the red line have same topic distributions.

Summation preserves the absolute magnitude of the distance, and helps to improve the visualization task by keeping *non*-neighbors separated on a visualizable Euclidean space. Taking the product is unsuitable, because it constrains the *ratio* of distances between neighbors to distances between *non*-neighbors. This may result in the crowding effect, where many documents are clustered together, because the relative ratio may be maintained, but the absolute distances on the visualization space could be too small.

Other than the proposed regularization function above, it is also possible to consider other regularization functions. For instance, we have also experimented with modifying the regularization function adapted from Discriminative Topic Model (DTM) (Huh & Fienberg, 2012), which addressed topic modeling but not semantic visualization. Note that while in the original DTM formulation, the distance function $\mathcal{F}(\psi_i, \psi_j)$ operates in the topic space, we adapt it for semantic visualization by redefining the distance function $\mathcal{F}(\psi_i, \psi_j)$ so that it will operate in the visualization space instead. This modified DTM formulation is shown to underperform the proposed regularization function above (Le & Lauw, 2014b).

4.1.2 Enforcing Neighborhood Structure: Visualization vs. Topic Space

We now turn to the definition of $\mathcal{F}(\psi_1, \psi_2)$. In neighborhood-based models (Belkin et al., 2006; Cai et al., 2008, 2009), there is only one low-rank representative space. For semantic visualization, there are two: topic and visualization spaces. We look into where and how to enforce the neighborhood graph structure.

At first glance, they seem equivalent. After all, they are representations of the same documents. However, this is not necessarily the case. Consider a simple example of two topics z_1 and z_2 with visualization coordinates $\phi_1 = (0, 0)$ and $\phi_2 = (2, 0)$ respectively. Meanwhile, there are three documents $\{d_1, d_2, d_3\}$ with coordinates $x_1 = (1, 1)$, $x_2 = (1, 1)$, and $x_3 = (1, -1)$. If two documents have the same coordinates, they will also have the same topic distributions. In this example, x_1 and x_2 are both equidistant from ϕ_1 and ϕ_2, and therefore according to Equation 3, they have the same topic distribution $\mathrm{P}(z_1|d_1) = \mathrm{P}(z_1|d_2) = 0.5$, and $\mathrm{P}(z_2|d_1) = \mathrm{P}(z_2|d_2) = 0.5$. If two documents have the same topic distributions, they may not necessarily have the same coordinates. d_3 also has the same topic distribution as d_1 and d_2, but a different coordinate. In fact, any coordinate of the form $(1, ?)$ will have the same topic distribution. This example is illustrated in Figure 3.

This suggests that enforcing neighborhood structure on the topic space may not necessarily lead to having data points closer on the visualization space. We postulate that regularizing the visualization space is more effective. There are also advantages in computational efficiency to doing so, which we will describe further shortly. Therefore, we define $\mathcal{F}(\psi_i, \psi_j)$ as the squared Euclidean distance $||x_i - x_j||^2$ between the corresponding visualization coordinates.

4.2 Neighborhood Graph

We discuss how the neighborhood graph may be approximated, which concerns the two issues of how the graph edges are defined, as well as how they are weighted. The neighborhood graph is constructed in the original data space where we represent each document as a tf-idf vector (Manning, Raghavan, Schütze, et al., 2008). We also experiment with different vector representations, including word counts and term frequencies, and find tf-idf to give the best results. The distance between two document vectors is measured using Euclidean distance.

4.2.1 Graph Construction

There have been research studies on the properties and methods for construction of neighborhood graphs (Zemel & Carreira-Perpiñán, 2004; Carey & Mahadevan, 2014). Since the construction of neighborhood graph is a critical step that may affect the performance of various graph-based algorithms, this problem itself is a research issue of independent interest. Our scope is in exploring how some well-established graph construction techniques may apply to the case of semantic visualization. We will investigate these various graph construction methods empirically in Section 6.

In the following, we briefly review two categories of graph construction methods.

1. *Neighborhood-based Graphs.* In this formulation, edges are formed between data points that are deemed to be sufficiently close to each other. This admits different definitions of "sufficient closeness". The most common definitions found in the literature include the two below.

   (a) ϵ-ball: The neighborhood graph contains an edge connecting two documents d_i and d_j, if d_i and d_j have a distance less than a threshold ϵ.

   (b) k-nearest neighbors (k-NN) graph: The neighborhood graph contains an edge connecting two documents d_i and d_j, if d_i is in the set $\mathcal{N}_k(d_j)$ of the $k-$nearest neighbors of d_j, or d_j is in the set $\mathcal{N}_k(d_i)$.

   ϵ-ball and k-NN both have strongly data-dependent parameters (i.e., ϵ and k) and it is not straightforward to choose the best value for these parameters. Neither guarantees that the graph would be connected. They also need to be carefully selected or tuned, as to some extent they also affect the "balance" between the contribution of neighbors $\mathcal{R}_+$ and non-neighbors $\mathcal{R}_-$ to the neighborhood regularization $\mathcal{R}_*$ in Equation 8. In

Appendix A, we explore empirically how these graph parameters can help to maintain this balance within the neighborhood regularization function.

ϵ-ball suffers from another issue that it tends to produce many edges for the points located at high-density regions, and thus has little restriction on the maximum degree of a vertex. k-NN does not suffer from that problem and is one of the most commonly used types of graphs.

In our subsequent development and experiments, we will experiment with both ϵ-ball and k-NN graph as there may be some variance in the performance of different graph construction techniques for different datasets (Hein, Audibert, & Luxburg, 2007; Ting, Huang, & Jordan, 2010; Coifman & Lafon, 2006).

2. *Minimum Spanning Tree-based Graphs.* While ϵ-ball and k-NN are quite sensitive to noise and sparsity, graph construction based on combining multiple minimum spanning trees can help to reduce sensitivity to noise of the output graph (Zemel & Carreira-Perpiñán, 2004). There are two variations based on this approach.

   (a) Perturbed Minimum Spanning Trees (PMST): PMST builds a neighborhood graph by generating $T > 1$ perturbed copies of the whole dataset according to the local noise model and fitting an MST to each perturbed copy. A weight $e_{ij} \in [0, 1]$ will be assigned to the edge between points x_i and x_j equal to the average number of times that edge appears on the trees.

   (b) Disjoint Minimum Spanning Trees (DMST): DMST produces a neighborhood graph by finding a deterministic collection of r minimum spanning trees that satisfies the property that no tree in the collection uses any edge of other trees. The neighborhood graph is the union of all edges of trees and contains $r(N-1)$ edges.

   As the representative of this category, we use DMST, which is deterministic and easier to construct than PMST while showing similar efficacies.

4.2.2 Graph Weighting

The next issue is how to assign weights to the edges in the neighborhood graph. In this respect, we consider two variations of edge weights.

1. *Simple Minded*:

$$\omega_{ij} = \begin{cases} 1, & \text{if only if } d_i \text{ and } d_j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

   This is the simplest approach where we use binary weighting to assign the weights to the edges. However, this approach to assign uniform weights to edges can be sensitive to errors, because of the “cliff effect” from 1 immediately to 0. Moreover, since the weights are not smoothed, it could result in some loss of information. We hypothesize that among the connected nodes, there may still be some differences in terms of degrees of similarity, which are expressed by their mutual distances. This motivates the second approach below.

2. *Heat Kernel*:

$$\omega_{ij} = \begin{cases} \exp(-\frac{||d_i - d_j||^2}{\tau}), & \text{if only if } d_i \text{ and } d_j \text{ are connected,} \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

An alternative approach is using the Heat Kernel function (Belkin & Niyogi, 2001; Jebara, Wang, & Chang, 2009). Heat Kernel has the advantage over Simple Minded by allowing smoother weights for the edges, which helps address the issues of sensitivity and loss of information. However, while Simple Minded is not parameterized, Heat Kernel has one parameter that needs to be determined (i.e., τ). Note that for $\tau = \infty$, Heat Kernel degenerates into Simple Minded, i.e., the former is the more general formulation. The exact value of τ is not important in our model because it would effectively be absorbed by the regularization parameter. For simplicity, we set $\tau = 2$.

5. Model Fitting

We now discuss how the parameters of the model described in Sections 3 and 4 can be learned. One well-accepted framework to learn model parameters using maximum a posteriori (MAP) estimation is the Expectation-Maximization or EM algorithm (Dempster, Laird, & Rubin, 1977).

For our model, the regularized conditional expectation of the complete-data log likelihood in MAP estimation with priors is:

$$\begin{aligned} \mathcal{Q}(\Psi|\hat{\Psi}) = & \sum_{n=1}^{N} \sum_{m=1}^{M_n} \sum_{z=1}^{Z} \mathrm{P}(z|n, m, \hat{\Psi}) \log \left[\mathrm{P}(z|x_n, \Phi) \mathrm{P}(w_{nm}|\theta_z) \right] \\ & + \sum_{n=1}^{N} \log(\mathrm{P}(x_n)) + \sum_{z=1}^{Z} \log(\mathrm{P}(\phi_z)) + \sum_{z=1}^{Z} \log(\mathrm{P}(\theta_z)) \\ & + \lambda \cdot \mathcal{R}(\Psi|\Omega), \end{aligned}$$

where $\hat{\Psi}$ is the current estimate. $\mathrm{P}(z|n, m, \hat{\Psi})$ is the class posterior probability of the n^{th} document and the m^{th} word in the current estimate. $\mathrm{P}(\theta_z)$ is a symmetric Dirichlet prior with parameter α for word probability θ_z. $\mathrm{P}(x_n)$ and $\mathrm{P}(\phi_z)$ are Gaussian priors with a zero mean and a spherical covariance for the document coordinates x_n and topic coordinates ϕ_z. We set the hyper-parameters to $\alpha = 0.01$, $\beta = 0.1N$ and $\gamma = 0.1Z$ following PLSV (Iwata et al., 2008).

In the E-step, $\mathrm{P}(z|n, m, \hat{\Psi})$ is updated as follows:

$$\mathrm{P}(z|n, m, \hat{\Psi}) = \frac{\mathrm{P}(z|\hat{x}_n, \hat{\Phi}) \mathrm{P}(w_{nm}|\hat{\theta}_z)}{\sum_{z'=1}^{Z} \mathrm{P}(z'|\hat{x}_n, \hat{\Phi}) \mathrm{P}(w_{nm}|\hat{\theta}_{z'})}.$$

In the M-step, by maximizing $\mathcal{Q}(\Psi|\hat{\Psi})$ w.r.t θ_{zw}, the next estimate of word probability θ_{zw} is as follows:

$$\theta_{zw} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M_n} \mathrm{I}(w_{nm} = w) \mathrm{P}(z|n, m, \hat{\Psi}) + \alpha}{\sum_{w'=1}^{W} \sum_{n=1}^{N} \sum_{m=1}^{M_n} \mathrm{I}(w_{nm} = w') \mathrm{P}(z|n, m, \hat{\Psi}) + \alpha W},$$

where I(.) is the indicator function. ϕ_z and x_n cannot be solved in a closed form, and are estimated by maximizing $\mathcal{Q}(\Psi|\hat{\Psi})$ using quasi-Newton (Liu & Nocedal, 1989).

The computation fo the gradients of $\mathcal{Q}(\Psi|\hat{\Psi})$ w.r.t ϕ_z and x_n depend on the specific kernel used (see Section 3.3).

- For the Gaussian kernel, we have the following gradients:

$$\frac{\partial \mathcal{Q}(\Psi|\hat{\Psi})}{\partial \phi_z} = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\big(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\big)(\phi_z - x_n) - \beta\phi_z,$$

$$\frac{\partial \mathcal{Q}(\Psi|\hat{\Psi})}{\partial x_n} = \sum_{m=1}^{M_n}\sum_{z=1}^{Z}\big(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\big)(x_n - \phi_z) - \gamma x_n + \lambda\cdot\frac{\partial \mathcal{R}(\Psi|\Omega)}{\partial x_n}.$$

- For the Student-t kernel, we have the following gradients:

$$\frac{\partial \mathcal{Q}(\Psi|\hat{\Psi})}{\partial \phi_z} = \sum_{n=1}^{N}\sum_{m=1}^{M_n}\frac{2\big(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\big)(\phi_z - x_n)}{1 + ||x_n - \phi_z||^2} - \beta\phi_z,$$

$$\frac{\partial \mathcal{Q}(\Psi|\hat{\Psi})}{\partial x_n} = \sum_{m=1}^{M_n}\sum_{z=1}^{Z}\frac{2\big(\mathrm{P}(z|x_n,\Phi) - \mathrm{P}(z|n,m,\hat{\Psi})\big)(x_n - \phi_z)}{1 + ||x_n - \phi_z||^2} - \gamma x_n + \lambda\cdot\frac{\partial \mathcal{R}(\Psi|\Omega)}{\partial x_n}.$$

The gradient of $\mathcal{R}(\Psi|\Omega)$ w.r.t. x_n is computed depending on the form of the regularization function $\mathcal{R}(\Psi|\Omega)$. When we use the proposed regularization function $\mathcal{R}_*(\Psi|\Omega)$ described in Section 4.1.1, we have the following gradient:

$$\begin{aligned}\frac{\partial \mathcal{R}(\Psi|\Omega)}{\partial x_n} &= \frac{\partial \mathcal{R}_*(\Psi|\Omega)}{\partial x_n}\\ &= -\frac{1}{2}\sum_{j=1;j\neq n}\big(4\omega_{nj}(x_n - x_j)\big) - \sum_{j=1;j\neq n}\big(4(1-\omega_{nj})\frac{(x_n - x_j)}{(\mathcal{F}(\psi_n,\psi_j)+1)^2}\big).\end{aligned}$$

As mentioned earlier, there is an efficiency advantage to regularizing on the visualization space. $\mathcal{R}(\Psi|\Omega)$ does not contain the variable ϕ_z if we do regularization on visualization space. The complexity of computing all $\frac{\partial \mathcal{R}(\Psi|\Omega)}{\partial x_n}$ is $O(N^2)$. In contrast, if we do regularization on topic space, we have to take the gradient of $\mathcal{R}(\Psi|\Omega)$ w.r.t to ϕ_z. That contributes towards a greater complexity of $O(Z^2 \times N^2)$ to compute all $\frac{\partial \mathcal{R}(\Psi|\Omega)}{\partial \theta_z}$. Therefore, regularization on topic space would run much slower than on visualization space.

6. Experiments

The main objective of our experiments is to evaluate the effectiveness of neighborhood regularization for semantic visualization model. After describing the experimental setup, we first examine the different design choices of the model relating to kernel, graph construction, and regularization function. Thereafter, we compare Semafore against the baseline methods that also aim to address both visualization and topic modeling, quantitatively and qualitatively, first in terms of visualization and then in terms of topic modeling.

6.1 Experimental Setup

In this section, we give a description of benchmark datasets as well as suitable metrics that are used for evaluation.

6.1.1 Datasets

We use three real-life, publicly available datasets (Cardoso-Cachopo, 2007) for evaluation.

- $20News$ contains newsgroup articles (in English) from 20 classes.
- $Reuters8$ contains newswire articles (in English) from 8 classes.
- $Cade12$ contains web pages (in Brazilian Portuguese) classified into 12 classes.

These are benchmark datasets used for document classification. While our task is fully unsupervised, the ground-truth class labels are useful for an objective evaluation. We create balanced classes by sampling fifty documents from each class, following the practice in PLSV (Iwata et al., 2008). This results in, for one sample, 1000 documents for $20News$, 400 for $Reuters8$, and 600 for $Cade12$. The vocabulary sizes are 5.4K for $20News$, 1.9K for $Reuters8$, 7.6K for $Cade12$. As the algorithms are probabilistic, we generate five samples for each dataset. For each sample, we conduct five independent runs. Therefore, the result reported for each setting is the average over a total of 25 runs.

6.1.2 Metrics

For a suitable metric, we return to the fundamental principle that a good visualization should preserve the relationship between documents (in high-dimensional space) in the lower-dimensional visualization space. User studies, even when well-designed, could be overly subjective and may not be repeatable across different users reliably. Therefore, for a more objective evaluation, we rely on two types of quantitative analysis:

- *Classification*: This evaluation relies on the ground-truth class labels found in the datasets. This is a well-established practice in many clustering and visualization works in machine learning. The basis for this evaluation is the reasonable assumption that documents of the same class are more related than documents of different classes. Therefore a good visualization would place documents of the same class as neighbors on the visualization.

  For each document d_n, we hide its true class c_n, and generate a prediction for its class $\hat{\mathcal{C}}_t(n)$ by taking the majority class among its t-nearest neighbors, as determined by Euclidean distance on the visualization space. Classification accuracy $Classification\_Acc(t)$ is defined as the fraction of documents whose predicted class $\hat{\mathcal{C}}_t(n)$ matches the true class c_n. More specifically, we have:

$$Classification\_Acc(t) = \frac{1}{N}\sum_{n=1}^{N}\delta(\hat{\mathcal{C}}_t(n) = c_n),$$

where δ is the delta function that equals 1 if the prediction matches and 0 otherwise.

The same metric is used in PLSV (Iwata et al., 2008). While accuracy is computed based on documents' coordinates, the same trends will be produced if computed based on topic distributions (due to their coupling through the kernels described in Section 3.3).

- *Neighborhood Preservation*: This evaluation does not rely on the ground-truth class labels but on the local neighborhood structure in the input data. The assumption is that a good visualization would be able to preserve the local structure in the input data as much as possible. If two documents are neighbors in the input data, they should still be neighbors in the visualization space.

  For every document d_n, we compute sets of t-nearest neighbors $\mathcal{Y}_t(n)$ and $\mathcal{X}_t(n)$ of document d_n in the input data and the visualization respectively. The neighborhood preservation accuracy $Preservation\_Acc(t)$ is then defined as the average fraction of the overlap size of $\mathcal{Y}_t(n)$ and $\mathcal{X}_t(n)$ over the size of $\mathcal{Y}_t(n)$ (i.e. t), where $n = 1, \ldots, N$. More specifically, we have:

$$Preservation\_Acc(t) = \frac{1}{N} \sum_{n=1}^{N} \frac{|\mathcal{Y}_t(n) \cap \mathcal{X}_t(n)|}{t},$$

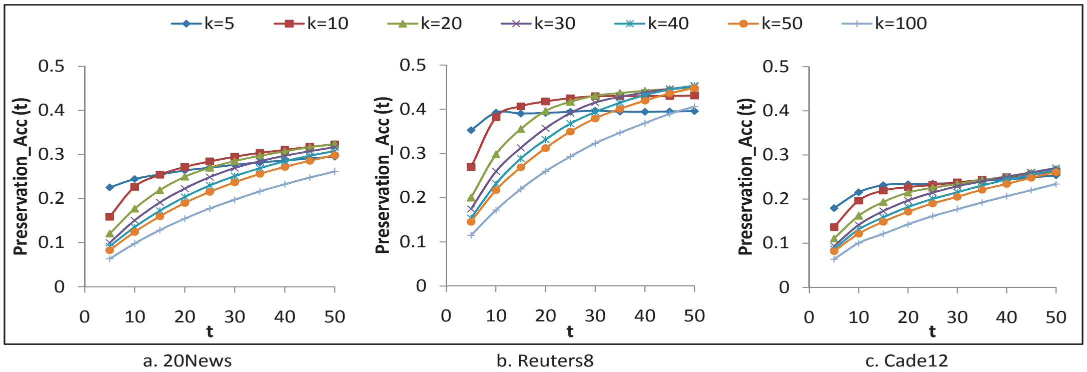
  where $|\mathcal{Y}_t(n) \cap \mathcal{X}_t(n)|$ is the size of the overlap set $\mathcal{Y}_t(n) \cap \mathcal{X}_t(n)$.

  A similar measure can be found in the literature (Akkucuk & Carroll, 2006), where it is called the "rate of agreement in local structure" or "agreement rate" and is used to measure how well the local structure is preserved between the input data and the low dimensional embedding. It is also used for tuning the parameters of a non-linear dimensionality reduction method (Chen & Buja, 2009).

In the subsequent experiments, we let t vary in the range $[5, 50]$ with the step size 5 and report the accuracies. Since different methods may behave differently at different t's, choosing a specific t for comparison may be unfair for some methods. Moreover, a method that consistently does well for different t's would also have a "smoother" local structure. Therefore, when comparing various methods, we present the preservation or classification accuracies averaged across $t \in [5, 50]$, denoted $Preservation\_Acc(Avg)$ and $Classification\_Acc(Avg)$ respectively.

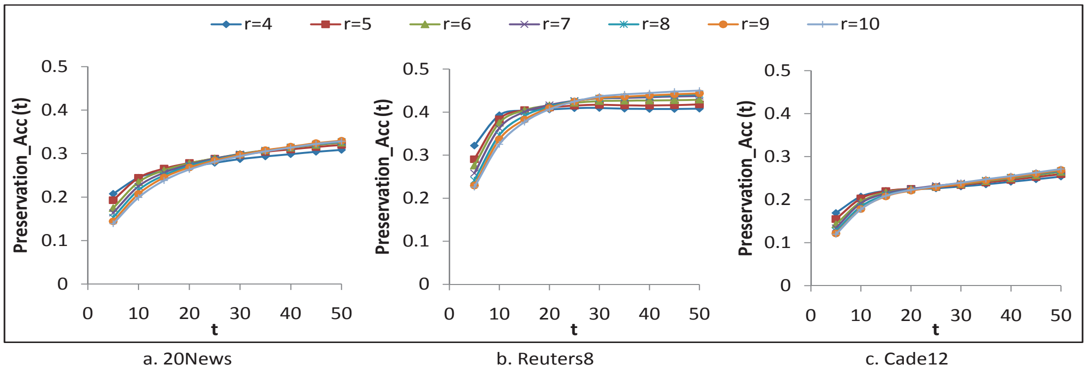
6.2 Parameter Study

In this section, we study the effects of graph parameters on our model. Specifically, the parameters concern the graph construction, including the number of neighbors k in k-NN graph, the distance threshold ϵ in ϵ-ball graph, and the number of minimum spanning trees r in DMST. For each type of graph, we use the Simple Minded weight. For the following figures, the regularization function is $\mathcal{R}_*$ with $\lambda = 10$ and the number of topics $Z = 20$. We use neighborhood preservation accuracy $Preservation\_Acc(t)$ to show the effects of graph parameters because this metric does not need ground-truth class labels, which are not always available for tuning these graph parameters.



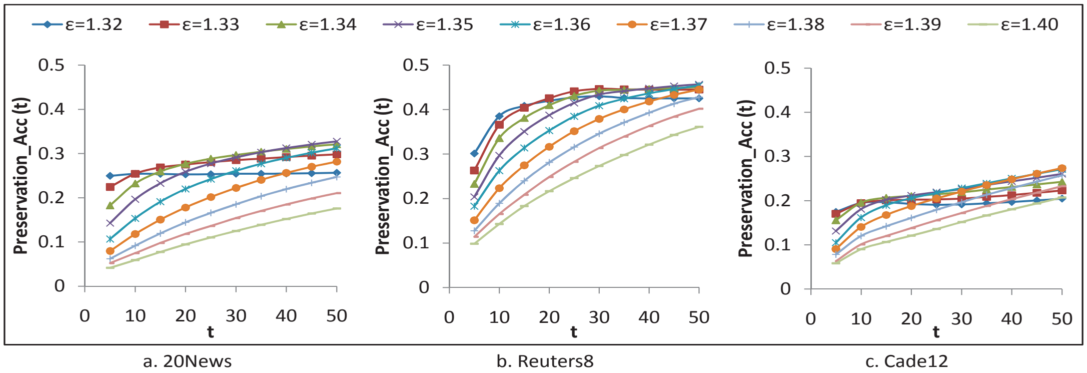
a. 20News b. Reuters8 c. Cade12

Figure 4: Preservation accuracy of Semafore when using k-NN graph with different neighborhood size k for (a) 20$News$, (b) $Reuters$8, and (c) $Cade$12.



a. 20News b. Reuters8 c. Cade12

Figure 5: Preservation accuracy of Semafore when using DMST graph with different number of minimum spanning trees r for (a) 20$News$, (b) $Reuters$8, and (c) $Cade$12.



a. 20News b. Reuters8 c. Cade12

Figure 6: Preservation accuracy of Semafore when using ϵ-ball graph with different values of distance threshold ϵ for (a) 20$News$, (b) $Reuters$8, and (c) $Cade$12.

In Figure 4, we show the performance of our model with different neighborhood size k in k-NN graph for different datasets. For every k, we vary t and plot the $Preservation\_Acc(t)$. Figure 4 shows that the optimum k for $20News$, $Reuters8$, and $Cade12$ is 10, 10, and 5 respectively. We compute the average accuracy $Preservation\_Acc(Avg)$ and it confirms that the optima are indeed at those k values. From now on, we will use k=10 for $20News$ and $Reuters8$, and k=5 for $Cade12$ when k-NN graph is used.

For DMST graph, we plot the $Preservation\_Acc(t)$ for different number of minimum spanning trees r with different datasets in Figure 5. It is difficult to see which r is the best in the figure because the differences between them are not much. The $Preservation\_Acc(Avg)$ is computed and it shows that for all three datasets, the optimum is about at r=5,6,7. Subsequently, we will use r=6 for DMST graphs for all three datasets.

For ϵ-ball graph, in Figure 6 we plot the $Preservation\_Acc(t)$ for different values of ϵ in the range $[1.32, 1.40]$. We choose that range because ϵ=1.32 and ϵ=1.40 roughly give an average number of neighbors of 5 and 100 respectively. The $Preservation\_Acc(Avg)$ shows that the optimum ϵ for $20News$, $Reuters8$, and $Cade12$ is 1.34, 1.35, and 1.33 respectively.

6.3 Model Analysis

In this section, we study the various design choices involved in designing the Semafore model, before finally concluding on the eventual synthesis of design choices to be used for comparison against the baselines. To keep the discussion focused and organized, in each of the following sub-section, we vary a single design choice, in order to isolate its effects. When unvaried, the model has the following setup by default: the number of topics is $Z = 20$, the graph construction method is k-NN, the graph weighting method is simple minded, the RBF kernel is Gaussian, and the regularization function is $\mathcal{R}_*$ with $\lambda = 10$.

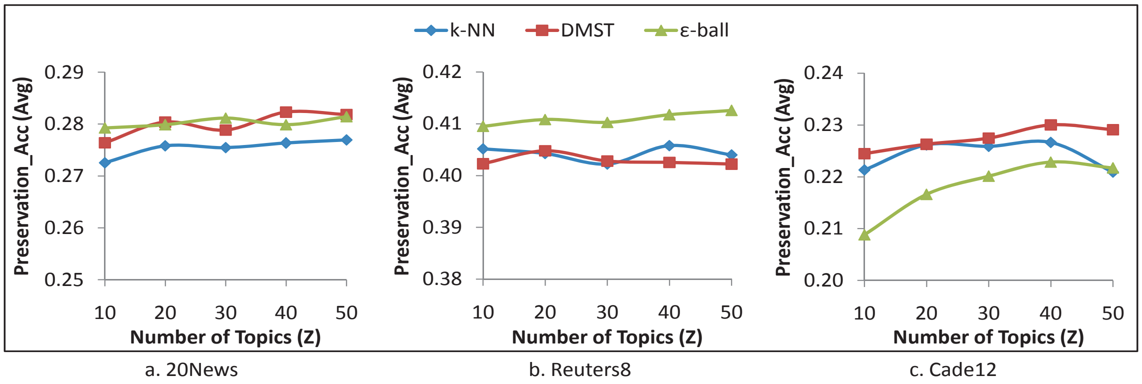
6.3.1 Neighborhood Graph Construction

We investigate three graph construction methods: k-NN, ϵ-ball and DMST, which are representatives of neighborhood-based and minimum spanning tree-based methods respectively. For each graph, its parameter is tuned as shown in Section 6.2. For the regularization parameter λ, we try different settings of λ on each dataset. It so happens that $\lambda = 10$ performs the best for all the graph construction methods across the three datasets.

In Figure 7, we run Semafore with different types of graph on the three datasets and report the $Preservation\_Acc(Avg)$ at different number of topics Z. The results show that different types of graph behave differently with different datasets. In $20News$, ϵ-ball and DMST give our model highest performance. Since the difference between the two are not statistically significant, we choose to use DMST for subsequent experiments on $20News$. For $Reuters8$, since ϵ-ball outperforms the others (significant at 0.05 level), it is going to be the default choice for subsequent experiments. For $Cade12$, the choice is DMST, which is slightly better than k-NN (statistically significant for $Z = 10, 40, 50$).

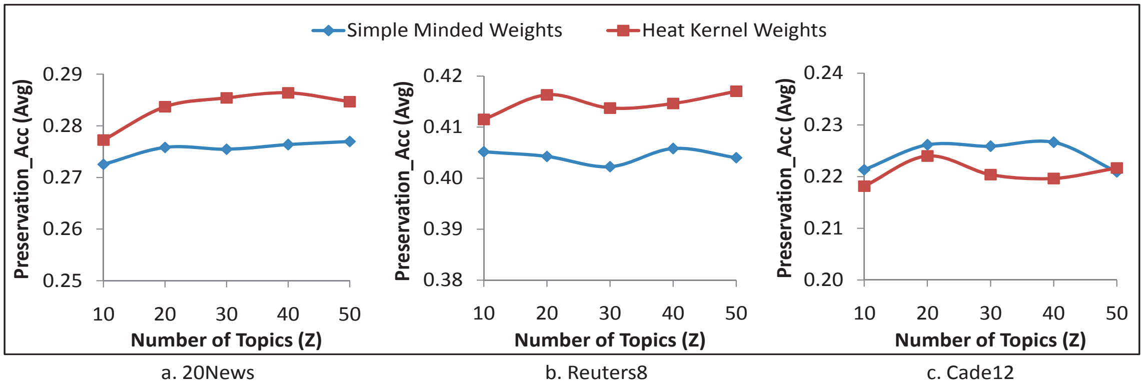
6.3.2 Neighborhood Graph Weighting

We now compare two variations of graph weighting methods, namely: Simple Minded and Heat Kernel methods. In this experiment, we use k-NN graph with specific k's for different



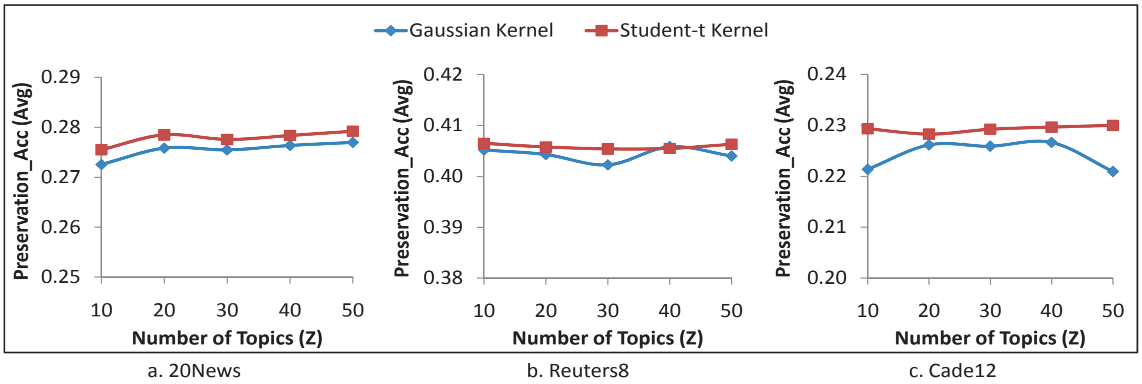
a. 20News b. Reuters8 c. Cade12

Figure 7: The effects of different graph construction methods on our model's performance.



a. 20News b. Reuters8 c. Cade12

Figure 8: The effects of different graph weighting schemes on our model's performance. The graph used in this experiment is k-NN graph with specific k's for different datasets as studied in Section 6.2.



a. 20News b. Reuters8 c. Cade12

Figure 9: The effects of Gaussian and Student-t RBF kernels on our model's performance.

	$20News$	$Reuters8$	$Cade12$
Regularization function	$\mathcal{R}_*$	$\mathcal{R}_*$	$\mathcal{R}_*$
Graph construction	DMST	ϵ-ball	DMST
Graph weighting	Heat Kernel	Heat Kernel	Simple Minded
RBF kernel	Student-t	Student-t	Student-t

Table 2: Synthesized Model for Each Dataset.

datasets as studied in Section 6.2. The regularization parameter λ is set to 10 after trying various settings and picking the best one.

In Figure 8, we compare Simple Minded method and Heat Kernel method to see their influences on our model at different number of topics Z. We observe that Heat Kernel is significantly and consistently better than Simple Minded method across all the cases in $20News$ and $Reuters8$. The difference is statistically significant at 0.01 level. One explanation is that Heat Kernel assigns smoother weights to the graph edges, and thus is more robust than Simple Minded. For $Cade12$, Simple Minded is slightly better, though the differences are statistically significant at 0.05 level only for $Z = 40$. Subsequently, we will use Heat Kernel for $20News$ and $Reuters8$, and Simple Minded for $Cade12$ as part of the final synthesis.

6.3.3 RBF Kernel

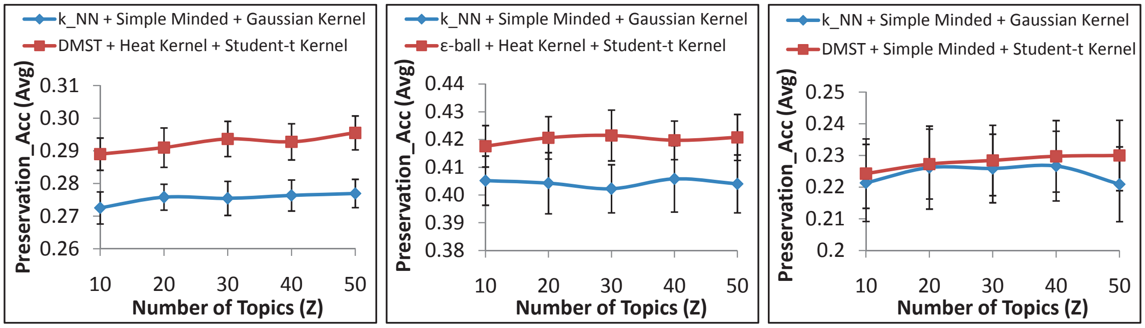
As described in Section 3.3, we express topic distributions as a function of visualization coordinates using RBF network as an abstraction. In this section, we show how different RBF kernels affect our model's performance. The two kernels we are exploring are Gaussian (Equation 3) and Student-t (Equation 4). We tune the regularization term λ for each kernel and see that the best one for the two kernels are $\lambda = 10$.

Figure 9 shows the results for different number of topics Z. Student-t kernel has a slight edge over Gaussian kernel consistently across different number of topics. The difference is small, but is statistically significant (at 0.05 level) in a majority of the cases (for $20News$ at $Z = 10, 20, 30, 50$, for $Reuters8$ at $Z = 30$, and for $Cade12$ at $Z = 10, 30, 50$). The slight improvement could be a sign that crowding problem does exist in the model. Student-t kernel would be even more useful when there is more extreme crowding issues, such as when the number of documents to be visualized is even larger. Subsequently, due to its slight edge, we will use Student-t as part of the final synthesis. As we will see shortly, using Student-t within the synthesized model results in a significant improvement overall.

6.3.4 Synthesised Semafore Model

Based on the model analysis in the preceding paragraphs, we combine the design choices into a final synthesis model called Semafore. The synthesized model is slightly different for different datasets, as listed in Table 2.

We now conduct another set of experiments to verify that those synthesized models would produce a noticeable improvement over the earlier version (kNN + Simple Minded + Gaussian Kernel) that appeared in our earlier work (Le & Lauw, 2014b), underlining



a. 20News b. Reuters8 c. Cade12

Figure 10: Our synthesized models with different properties compared to the earlier version (kNN + Simple Minded + Gaussian Kernel) that appeared in our earlier work (Le & Lauw, 2014b).

the utility of the subsequent enhancements. Figure 10 shows that this is indeed the case. Based on the standard deviations shown in the figures, the improvements are very clear in $20News$ and $Reuters8$ but not so clear in $Cade12$. Paired samples t-test indicate that the improvement is significant at 0.05 level or lower in all cases, except for the cases where $Z = 10, 20$ in $Cade12$. We will use these synthesized models in the comparisons against the baseline methods in the following section.

6.4 Comparison of Visualizations

We now compare our proposed model with several baselines. First, we outline the set of comparative methods. Thereafter, we discuss quantitative evaluation (in terms of accuracy), as well as qualitative evaluation (in terms of example visualizations). Finally, we will show that the gains in visualization quality does not come at the expense of topic modeling.

As semantic visualization seeks to ensure consistency between topic model and visualization, the comparison focuses on methods producing *both* topics and visualization coordinates which are listed in Table 3.

- Semafore is our proposed method that incorporates neighborhood structure into semantic visualization.

- PLSV (Iwata et al., 2008) is the state-of-the-art, representing the joint approach without neighborhood structure preservation.

- PE (LDA) represents the pipeline approach involving topic modeling with LDA (Blei et al., 2003), followed by visualizing documents' topic distributions with PE (Iwata et al., 2007). This pipeline is better than the LDA/MDS that appeared in our earlier work (Le & Lauw, 2014b). There are other pipeline methods, shown inferior to PLSV (Iwata et al., 2008), which are not reproduced here to avoid duplication.

	Visualization	Topic model	Joint model	Neighborhood
Semafore	✓	✓	✓	✓
PLSV	✓	✓	✓	
PE (LDA)	✓	✓		
t-SNE (LDA)	✓	✓		

Table 3: Comparative Methods.

- t-SNE (LDA) is another pipeline approach that first uses LDA (Blei et al., 2003) to learn topic model and then use t-SNE (Van der Maaten & Hinton, 2008) to visualize documents' topic distributions.

For completeness, we also conduct experiments for comparing our method with t-SNE and Laplacian EigenMaps (LE) (Belkin & Niyogi, 2003) (direct visualization, without topic modeling). To keep the discussion focused, we show them in Appendix B, as we do not consider t-SNE and LE as comparative baselines because these two methods only model visualization, but not topics.

6.4.1 Accuracy

In this section, we compare our model with several baselines in terms of classification accuracy (Figure 11) and neighborhood preservation accuracy (Figure 12). In the two figures, only the standard deviations for Semafore are shown.

Classfication Accuracy. Figure 11(a), 11(c) and 11(e) show the $Classfication\_Acc(t)$ at different t's for $Z = 20$ for $20News$, $Reuters8$, and $Cade12$ respectively. At any t, the comparison shows outperformance by Semafore over the baselines consistently. All four methods show the same behavior that their performances decrease when t increases. As t increases, they may lose accuracy in predicting labels for documents near to the border of each "cluster".

Now, we vary the number of topics Z. In Figure 11(b), we show the performance in $Classfication\_Acc(Avg)$ on $20News$. Figure 11(d) and 11(f) show the same for $Reuters8$ and $Cade12$ respectively. From these figures, we draw the following observations about the comparative methods:

- Semafore performs the best on all datasets across various numbers of topics (Z). Semafore beats PLSV by 25% to 51% on $20News$, by 6–13% on $Reuters8$, and by 22–32% on $Cade12$. These margins of performance with respect to PLSV are statistically significant at 0.01 significant level or lower in all cases. This effectively showcases the utility of neighborhood regularization in enhancing the quality of visualization. By preserving local consistency, Semafore achieves a good accuracy even at small number of topics (e.g., 10).

- PLSV performs better than PE (LDA) and t-SNE (LDA), which shows that there is utility to having a *joint*, instead of separate, modeling of topics and visualization. PE (LDA) and t-SNE (LDA) are worse than PLSV because it embeds documents by using two-step reductions that optimize separately two different objective functions.

Therefore, the errors from the previous step may propagate to the next, without an opportunity for correction. This may cause distortions in the visualization.

- In some cases, PLSV, PE (LDA) and t-SNE (LDA) tend to have decreasing accuracies when the number of topics increases. This may be because when number of topic increases, the topic distributions and the word probabilities may overfit the data and thus the accuracy is reduced. In contrast, Semafore shows a quite stable performance across different numbers of topics. This may be explained by the utility of neighborhood regularization, which helps to prevent overfitting when the number of topics increases.

Neighborhood Preservation Accuracy. While having better classification accuracy, Semafore also preserves well the local structure of the input data in the visualization space. The $Preservation\_Acc(t)$ results in Figure 12(a), 12(c) and 12(e) show that Semafore is consistently better than the other baselines in terms of neighborhood preservation across different t's and different datasets. In Figure 12(b), 12(d) and 12(f), we vary the number of topics Z and report the $Preservation\_Acc(Avg)$ results. Semafore beats PLSV by 41% to 76% on $20News$, by 24–36% on $Reuters8$, and by 29–45% on $Cade12$ in terms of neighborhood preservation accuracy. The improvements of Semafore over PLSV are statistically significant at 0.01 significant level or lower in all cases.

The above accuracy results are based on visualization coordinates. We have also computed accuracies based on topic distributions, which have similar trends.

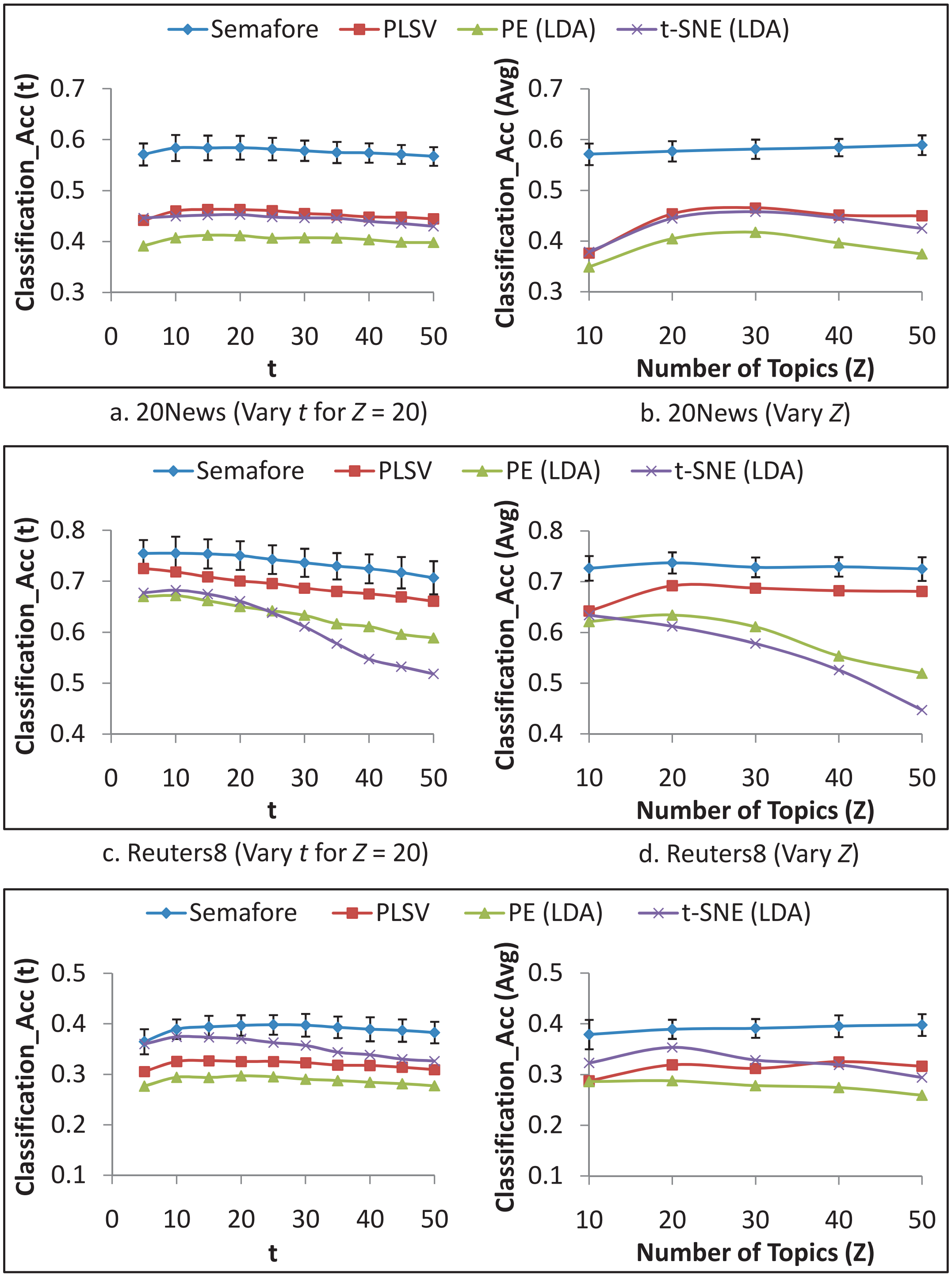
6.4.2 Visualizations

To provide an intuitive appreciation, we briefly describe a qualitative comparison of visualizations. For each method on each dataset, a visualization is shown as a scatterplot (best seen in color). Each document has a coordinate, and is assigned a shape and color based on its class. Each topic also has a coordinate, drawn as a black, hollow circle. A legend is provided, mapping each symbol to the corresponding class label.

Note that this is an illustrative, rather than a comparative discussion, as an objective evaluation should not rely on eyeballing alone. However, as we have shown the quantitative results in the preceding section, in this section, we focus on the qualitative study of the output visualizations.

20News. Figure 13 shows a visualization of $20News$ dataset. Semafore's Figure 13(a) shows that the different classes are well separated. There are distinct clusters of blue squares and purple diamonds at the top for hockey and baseball classes respectively, clusters of orange triangles and pink asterisks at the bottom for cryptography and medicine, etc. Beyond individual classes, the visualization also places related classes nearby. Computer-related classes are found on the lower left. Politics and religion are on the lower right.

Comparatively, Figure 13(b) by PLSV shows crowding at the center. For instance, motorcycle (green dashes) and autos (red dashes) are mixed at the center without a good separation. Figure 13(c) by PE (LDA) is worse. PE (LDA) does not give good separation for not similar classes. It mixes autos (red dashes) and space (green circles) together at the center. Medicine (pink asterisks) is also mixed with other classes in PE (LDA) while Semafore and PLSV give a good separation for it. Figure 13(d) is visualization by t-SNE (LDA). Although t-SNE (LDA) can separate well hockey (blue squares) and baseball



a. 20News (Vary *t* for *Z* = 20)

b. 20News (Vary *Z*)

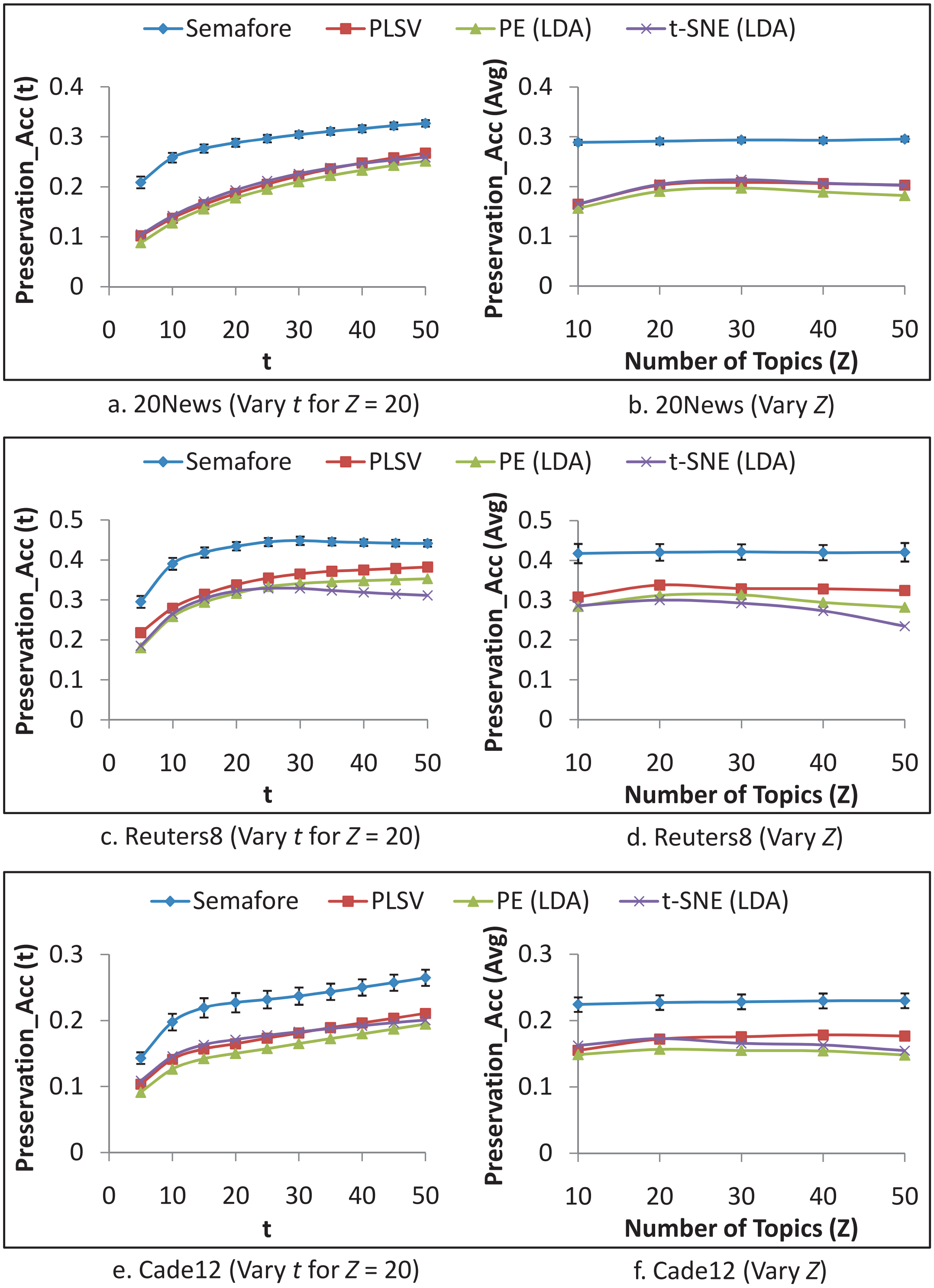
c. Reuters8 (Vary *t* for *Z* = 20)

d. Reuters8 (Vary *Z*)

e. Cade12 (Vary *t* for *Z* = 20)

f. Cade12 (Vary *Z*)

Figure 11: Classification Accuracy Comparison.



a. 20News (Vary *t* for *Z* = 20)

b. 20News (Vary *Z*)

c. Reuters8 (Vary *t* for *Z* = 20)

d. Reuters8 (Vary *Z*)

e. Cade12 (Vary *t* for *Z* = 20)

f. Cade12 (Vary *Z*)

Figure 12: Preservation Accuracy Comparison.

(purple diamonds) classes, it is not able to detect their semantic similarities (as baseball and hockey are both about sports). In addition, it still mixes documents of different classes together at the center and on the upper right.

Reuters8. Figure 14 shows the visualization outputs for *Reuters*8 dataset. Semafore in Figure 14(a) is better at separating the eight classes into distinct clusters. In an anticlockwise direction from the top, we have navy blue diamonds (*money-fx*), red dashes (*interest*), red squares (*crude*), light blue pluses (*earn*), green triangles (*acq*), purple crosses (*ship*), blue asterisks (*grain*), and finally orange circles (*trade*).

In comparison, PLSV in Figure 14(b) shows that several classes are intermixed at the center, including red dashes (*interest*), orange circles (*trade*), and navy blue diamonds (*money-fx*). PE (LDA) in Figure 14(c) is also worse when it mixes differentiated classes such as red dashes (*interest*) and navy blue diamonds (*money-fx*) together. t-SNE (LDA) in Figure 14(d) seems have better cluster separation but still mix documents with different classes together such as red squares (*crude*) and green triangles (*acq*) on the upper right. Green triangles (*acq*) also mix with light blue pluses (*earn*) on the left in the visualization by t-SNE (LDA).

Cade12. Figure 15 shows the visualization outputs for *Cade*12. This is the most challenging dataset. Even so, Semafore in Figure 15(a) still achieves a better separation between the classes, as compared to PLSV in Figure 15(b). Particularly, Semafore gives better separation for esportes (green triangles) as well as compras-on-line (orange circles) than PLSV and PE (LDA). t-SNE (LDA) shows quite good clusters for esportes (green triangles) as well as compras-on-line (orange circles) but it also merges many different classes together as in the clusters on the right and on the upper right.

6.5 Comparison of Topic Models

One question is whether Semafore's gain in visualization quality over the closest baseline PLSV is at the expense of the quality of its topic model. To investigate this, we will compare the topic models of Semafore and PLSV, which share a core generative process. For parity, in this comparison, we only include the joint models, whereby the visualization coordinates affect the topic models as well.

The metric we use to measure the quality of topic models is pairwise mutual information or PMI. It measures topic interpretability, based on coocurrence frequencies of the top words in each topic in a large external corpus. Although other metrics such as perplexity or held-out likelihood can show the generalization ability of a learned topic model on unseen test data, these traditional metrics do not capture whether topics are coherent (Chang, Gerrish, Wang, Boyd-Graber, & Blei, 2009). Therefore, in this comparison, we rely on PMI, which can measure the quality of topic words in terms of their interpretability to a human. To human subjects, interpretability is closely related to coherence (Newman, Lau, Grieser, & Baldwin, 2010), i.e., how much the top keywords in each topic are "associated" with each other. After an extensive study of evaluation methods for coherence, Newman et al. (2010) identify Pointwise Mutual Information (PMI) as the best measure, in terms of having the greatest correlation with human judgments.

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

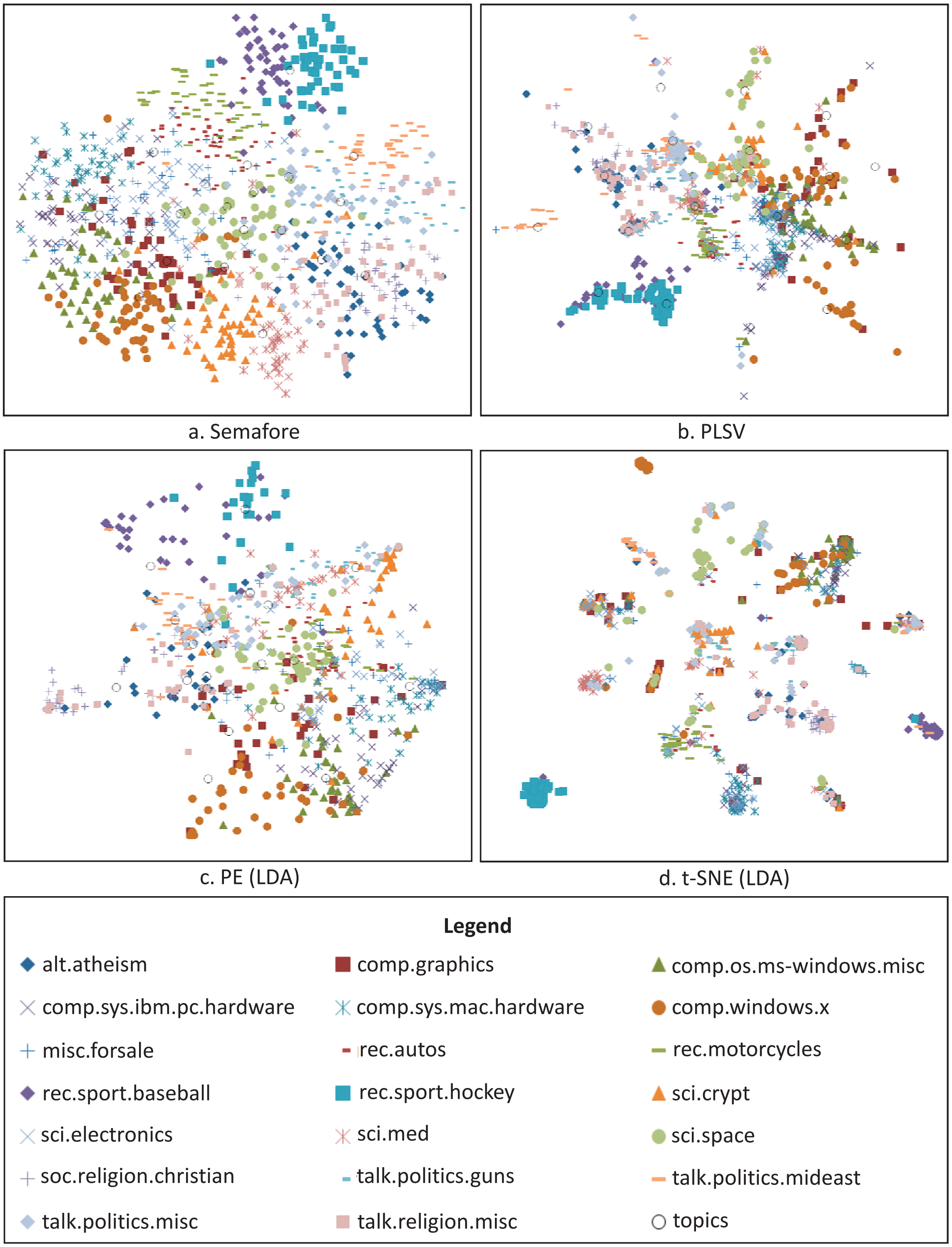
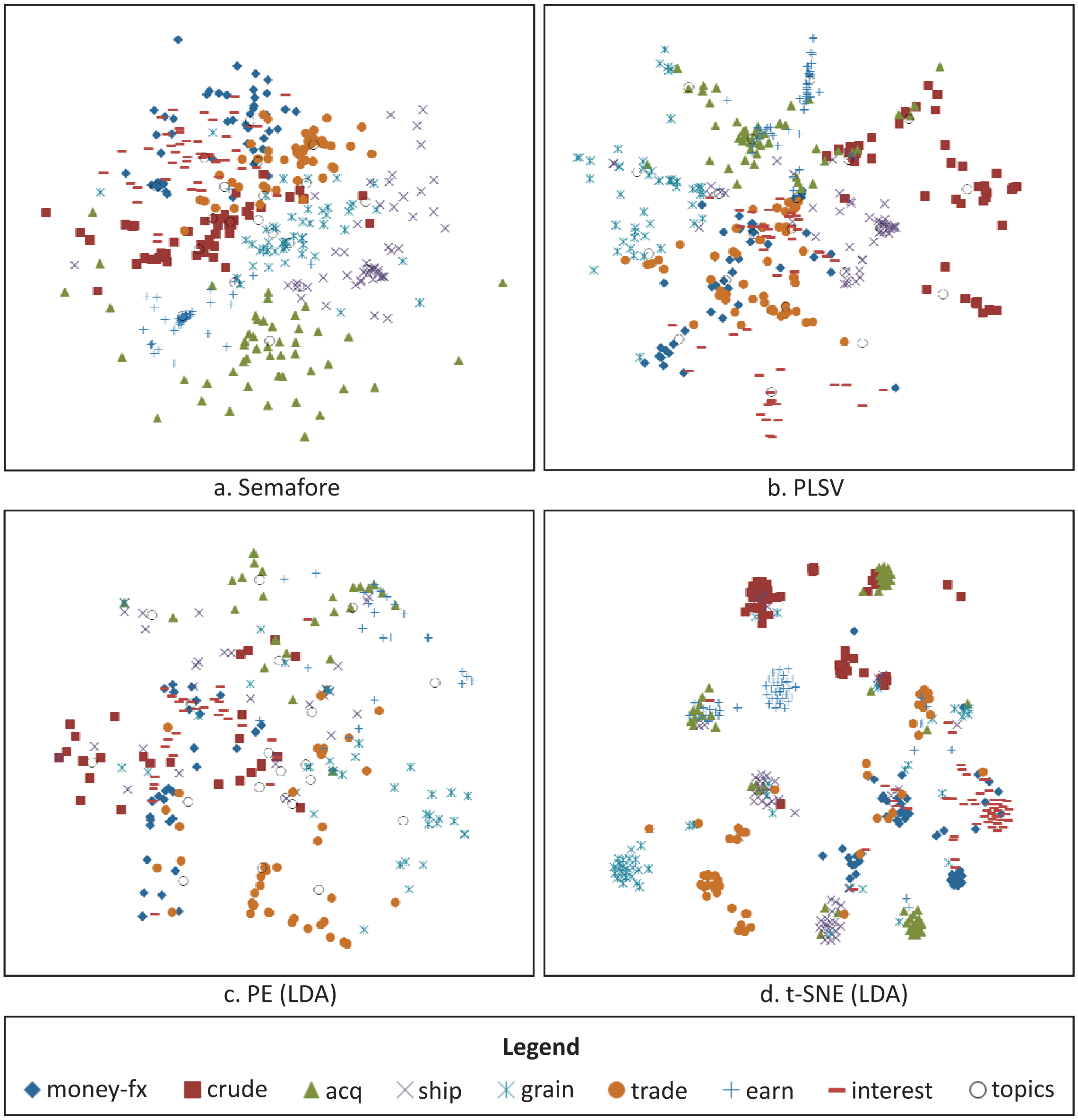


Figure 13: Visualization of documents in 20*News* for number of topics $Z = 20$. Each point represents a document and the shape and color represent document class. Each topic is drawn as a black, hollow circle.



a. Semafore

b. PLSV

c. PE (LDA)

d. t-SNE (LDA)

Figure 14: Visualization of documents in *Reuters*8 for number of topics $Z = 20$. Each point represents a document and the shape and color represent document class. Each topic is drawn as a black, hollow circle.

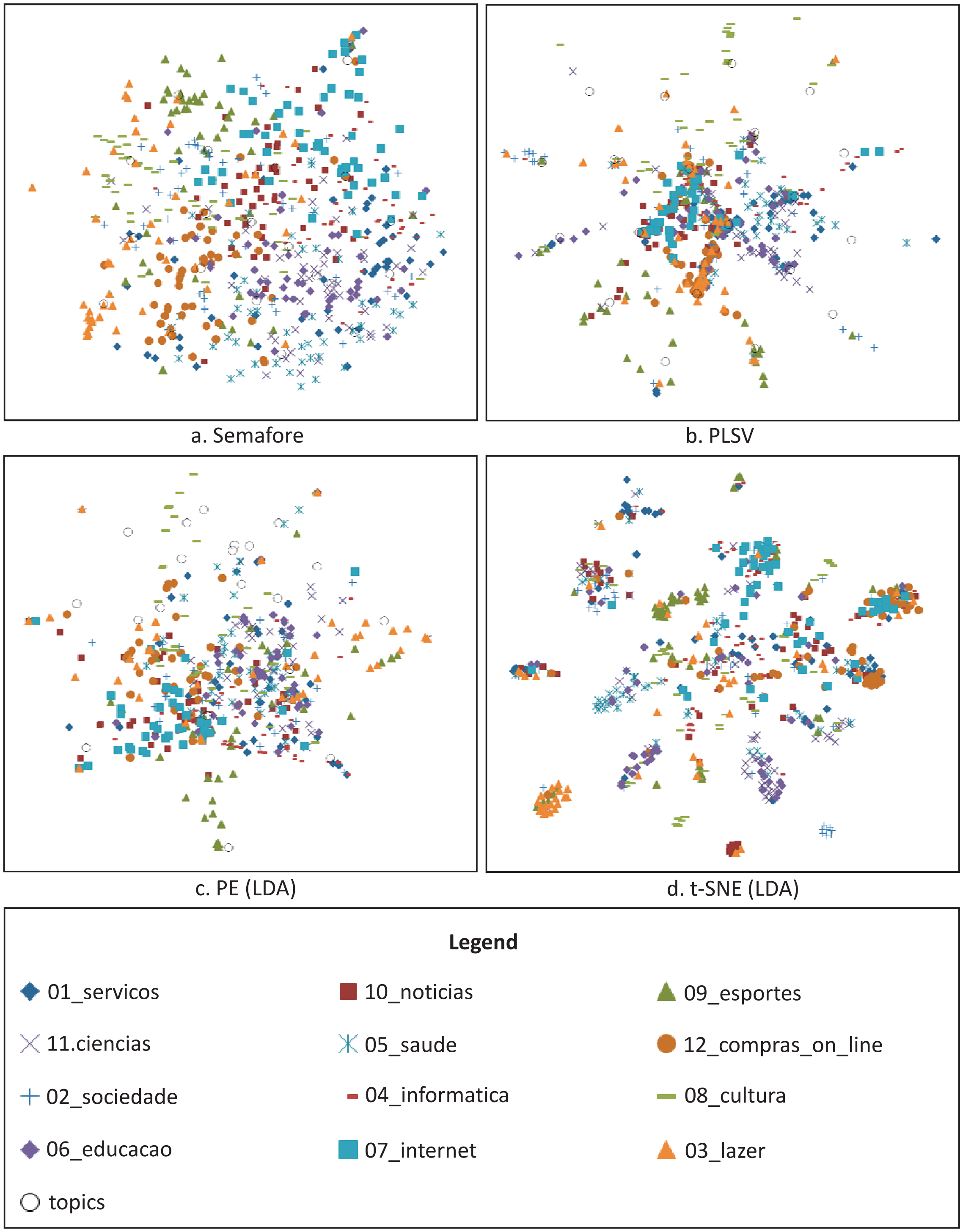
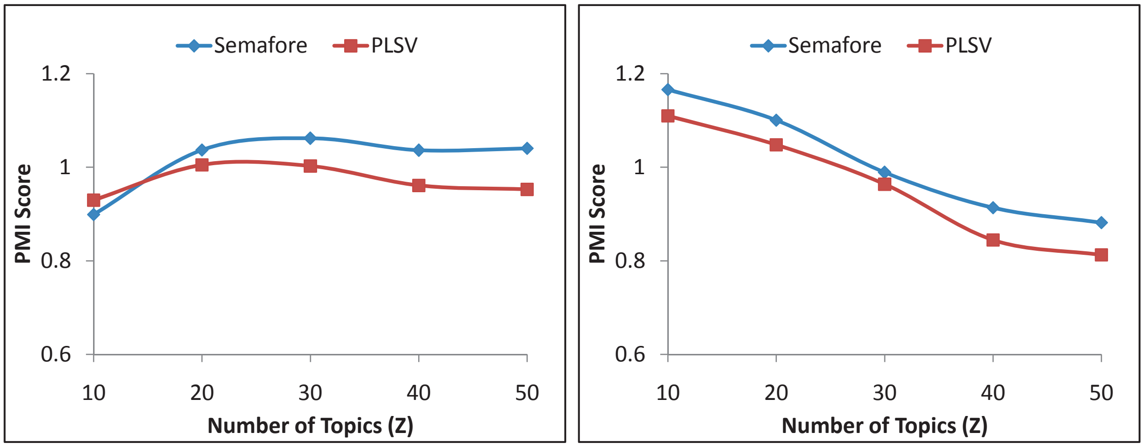


Figure 15: Visualization of documents in *Cade*12 for number of topics $Z = 20$. Each point represents a document and the shape and color represent document class. Each topic is drawn as a black, hollow circle.



a. 20News

b. Reuters8

Figure 16: Topic Interpretability of Semafore and PLSV in terms of PMI Score (higher is better).

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 Newman et al. (2009), we use *Google Web 1T 5-gram Version 1* (Brants & Franz, 2006), a huge corpus of n-grams generated from 1 trillion word tokens. $p(w_i)$ is estimated from the frequencies of 1-grams. As recommended by Newman et al., $p(w_i, w_j)$ is estimated from the frequencies of 5-grams. We obtain PMI for the English-based $20News$ and $Reuters8$, but not for $Cade12$ because we do not possess a large-scale n-gram corpus specifically for Brazilian Portuguese.

In Figure 16, we plot the PMI score for various number of topics Z. Semafore performs better than PLSV across most of the topics settings. In Figure 16(a) for 20News, except for the case at $Z = 10$, all cases of Semafore's outperformance are significant at 0.05 level or lower. In Figure 16(b) for Reuters8, all cases of Semafore's outperformance are significant at 0.05 level or lower except for the $Z = 30$. These results show that Semafore improves visualization while not sacrificing the topic interpretability of learned topics.

For a greater appreciation of the quality of the output topic models, in Appendix C, we show several examples of topic models for $Z = 20$, for both Semafore and PLSV, in terms of the top keywords with the highest probabilities for each topic.

7. Conclusion

In this paper, we address the semantic visualization problem, which jointly conducts topic modeling and visualization of documents. We propose a new framework to incorporate neighborhood structure within a probabilistic semantic visualization model called Semafore. The model is carefully designed to reflect the context of semantic visualization, leading to a number of design choices related to the RBF kernel for mapping topic and visualization spaces, the approximation of neighborhood graph through construction and weighting, as well as the appropriate regularization functions and spaces. Experiments on real-life datasets show that Semafore significantly outperforms the baselines in terms of visualization quality and accuracy, while having a similar, if not slightly better topic model. This provides evidence that neighborhood structure, together with joint modeling of topics and visualization, is important for semantic visualization.

Appendix A. Balancing Contributions of Neighbors and Non-neighbors to Regularization

As mentioned in Section 4.2, the balance between the contribution of neighbors $\mathcal{R}_+$ and non-neighbors $\mathcal{R}_-$ to the neighborhood regularization $\mathcal{R}_*$ in Equation 8 may require careful tuning of the graph parameters (i.e., ϵ or k). For example, in the case of using k-NN graph and N total number of documents, we would have kN terms in the neighbor regularization $\mathcal{R}_+$, and $(N-k)N$ terms in the non-neighbor regularization $\mathcal{R}_-$. Supposing that N increases significantly, there might be imbalance if k were to remain unchanged. Therefore, as N changes, k should also be tuned accordingly to maintain this balance. For a simplistic point, the ratio between kN and $(N-k)N$ would remain roughly the same if both N and k grow by similar factors. In practice, we recommend tuning k carefully.

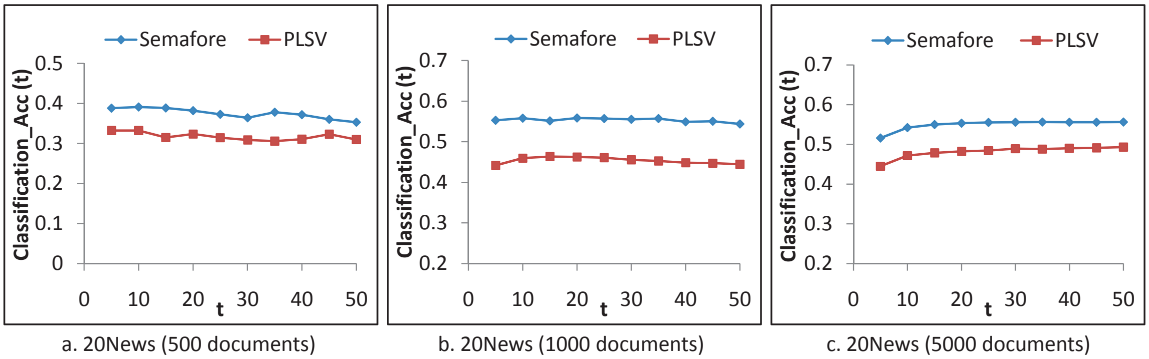
We run additional experiments to validate the above argument on the $20News$ dataset. Our basic point is that as N changes, k can be tuned to still show significant improvement due to the neighborhood graph regularization. The closest baseline is PLSV, both empirically in terms of classification accuracy, as well as conceptually as PLSV shares a similar generative process but with a different kernel and without neighborhood regularization. Hence, we compare the performance of our method Semafore (with k-NN graph, heat kernel weighting, and Student-t kernel) to PLSV on various data sizes at $Z = 20$ topics.

- Figure 17(a) is for dataset of size $N = 500$, and Semafore runs with $k = 10$.
- Figure 17(b) is for dataset of size $N = 1000$, and Semafore runs with $k = 10$.
- Figure 17(c) is for dataset of size $N = 5000$, and Semafore runs with $k = 50$.

We note that there is a 10X difference between the smallest and the largest datasets. Yet the relative outperformance of Semafore over PLSV by around 15% to 20% is evident across the three datasets. This supports the case that k can be tuned to produce a positive effect using neighborhood graph regularization.

Appendix B. Additional Comparisons

As mentioned in Section 6.4, for completeness, we include here additional comparisons to visualization methods that do not also aim at topic modeling. In particular, we include two methods. First, we include t-SNE (Van der Maaten & Hinton, 2008), which is also used in the composite t-SNE (LDA). Second, we include Laplacian EigenMaps (LE) (Belkin & Niyogi, 2003), which takes as input the neighborhood graph. Figure 18 and Figure 19 show the classification accuracy and preservation accuracy of Semafore , t-SNE and LE

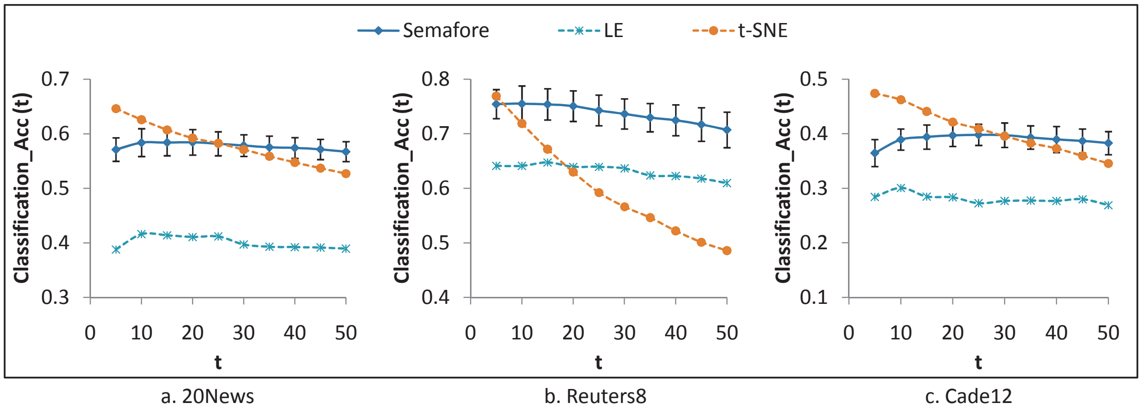


a. 20News (500 documents)

b. 20News (1000 documents)

c. 20News (5000 documents)

Figure 17: Classification accuracy comparison on 20*News* with various data sizes ($Z = 20$).

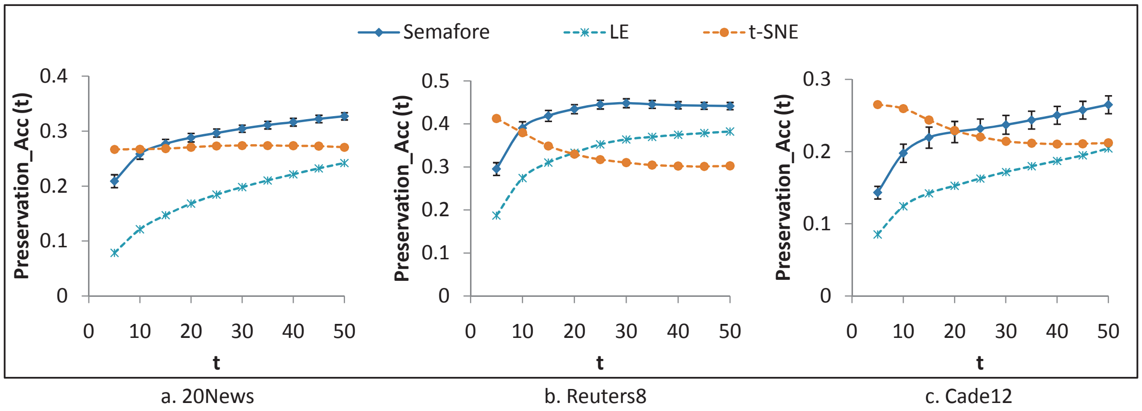


a. 20News

b. Reuters8

c. Cade12

Figure 18: Classification accuracy comparison.



a. 20News

b. Reuters8

c. Cade12

Figure 19: Preservation accuracy comparison.

when varying t. Semafore outperforms LE in all cases. For t-SNE, Semafore outperforms t-SNE for *Reuters*8. However, for 20*News* and *Cade*12, it is more difficult to tell whether Semafore or t-SNE is better. t-SNE tends to have decreasing accuracy as t increases. This is expected because t-SNE is known to focus on preserving the local structure (Van der Maaten & Hinton, 2008). When t is small, we basically consider only the local structure of the visualization. When t increases, we consider the more global structure of the visualization and Semafore outperforms t-SNE significantly. Overall, Semafore is more stable than t-SNE as t changes, which indicates that Semafore tries to balance preserving the local and global structure better than t-SNE. We emphasize that this comparison is for information purpose only, as we do not regard t-SNE and LE as comparative baselines.

Appendix C. Topic Model Examples

We showcase the topic models derived by Semafore and PLSV. For 20*News*, Table 4 shows the topics of Semafore, and Table 5 shows the topics of PLSV. For *Reuters*8, Table 6 shows the topics of Semafore, and Table 7 shows the topics of PLSV. For *Cade*12, Table 8 shows the topics of Semafore, and Table 9 shows the topics of PLSV.

For each method, we show the list of twenty topics. For each topic, we produce the top ten words with the highest probabilities. As shown by the top words, the topics do correspond strongly to some of the classes. For example, topic s_{19} in Table 4 for 20*News* is about Christianity, which corresponds to the *soc.religion.christian* class. Topic s_4 is about cars and motorcycles, corresponding to *rec.autos* and *rec.motorcycles*. Topic s_{12} is probably concerning the categories of *rec.sport.baseball* and/or *rec.sport.hockey*.

Overall, we observe that the quality of topic words are comparable across the comparative methods. Note that there is no direct correspondence between the topics of different methods (e.g., the first topic of Semafore may not correspond to the first topic of PLSV). Through manual inspection, we can see that there are some related topics, e.g., s_4 and p_6, or s_{12} and p_7. However, the sets of topics and the set of keywords for each topic are not identical. This is borne out in the slight difference in terms of PMI scores.

This qualitative study helps to show that Semafore improves the visualization quality, while still maintaining at least the same quality of topic words, if not better. This supports the conclusion reached by the quantitative comparisons in the main manuscript.

Table 4: Semafore's Topic Model for 20*News* (for 20 topics)

Topic ID	Top 10 Words
s_0	space, system, -rcb-, book, computer, university, list, post, price, science
s_1	article, year, good, write, guy, well, time, head, question, leave
s_2	gun, law, kuwait, people, death, fbus, article, control, weapon, child
s_3	window, file, program, widget, application, type, will, resource, call, function
s_4	car, bike, speed, engine, drive, lock, turn, mile, front, change
s_5	will, power, place, work, rate, write, sound, lead, good, interested
s_6	write, article, thing, time, people, better, start, problem, will, good
s_7	write, time, people, friend, pay, public, article, tax, opinion, money
s_8	people, claim, write, system, person, moral, evidence, objective, read, state
s_9	image, datum, graphic, send, file, format, package, software, mail, include
s_{10}	armenian, fire, jew, child, kill, start, people, turkish, door, israel
s_{11}	system, board, will, datum, time, work, tape, test, copy, command
s_{12}	game, team, year, player, win, play, will, hit, season, hockey
s_{13}	will, post, space, good, time, include, cost, option, launch, people
s_{14}	drive, card, window, appear, disk, ram, driver, memory, work, color
s_{15}	mr., president, stephanopoulo, state, group, consider, party, question, issue, press
s_{16}	write, article, well, will, thing, work, point, include, time, help
s_{17}	key, article, chip, food, write, people, government, encryption, thing, algorithm
s_{18}	price, buy, apple, computer, dealer, fit, model, problem, sell, monitor
s_{19}	god, jesus, will, christian, religion, faith, truth, bible, belief, church

Table 5: PLSV's Topic Model for 20*News* (for 20 topics)

Topic ID	Top 10 Words
p_0	write, people, christian, belief, time, faith, god, religion, life, will
p_1	god, will, jesus, kuwait, atheist, church, christian, man, religion, sin
p_2	armenian, appear, art, turkish, tartar, 1st, village, armenia, 1.40, genocide
p_3	will, key, write, time, article, government, system, thing, chip, hit
p_4	mr., stephanopoulo, president, will, party, state, door, time, meeting, open
p_5	write, fire, article, gun, system, -rcb-, start, people, fbus, claim
p_6	car, will, bike, engine, drive, well, dealer, battery, change, front
p_7	game, win, year, will, team, play, season, good, goal, playoff
p_8	player, team, write, hockey, game, fan, article, year, will, guy
p_9	space, system, datum, will, april, nasa, security, university, computer, list
p_{10}	graphic, image, file, ftp, send, format, package, system, datum, object
p_{11}	image, datum, program, window, version, file, software, tool, support, user
p_{12}	drive, jumper, master, ndet_loop, slave, rate, gun, function, crime, set
p_{13}	window, file, card, will, program, color, driver, support, disk, bit
p_{14}	people, write, state, article, law, government, country, rights, jew, will
p_{15}	write, article, thing, people, good, will, time, lot, year, day
p_{16}	work, drive, tape, scsus, problem, simm, controller, write, memory, article
p_{17}	widget, -rcb-, window, -lcb-, application, resource, set, visual, type, file
p_{18}	price, will, write, system, computer, article, apple, chip, monitor, board
p_{19}	will, vote, comp, newsgroup, suit, problem, os2, sco, post, mail

Table 6: Semafore's Topic Model for *Reuters*8 (for 20 topics)

Topic ID	Top 10 Words
s_0	company, pipeline, raise, crude, march, spokesman, refinery, capacity, corp, post
s_1	pct, bank, day, stg, today, reuter, money, market, mln, bill
s_2	offer, share, company, board, group, acquire, stock, dlr, acquisition, receive
s_3	exchange, currency, dollar, west, finance, baker, monetary, germany, continue, interest
s_4	share, reuter, dlr, mln, buy, company, corp, pay, stock, group
s_5	price, opec, market, bpd, official, february, month, output, saudus, january
s_6	rate, bank, pct, cut, fund, prime, point, reserve, issue, lower
s_7	billion, foreign, import, increase, dlr, trade, economic, export, will, country
s_8	bank, billion, market, government, fall, stock, economy, rise, surplus, deficit
s_9	will, company, sell, pct, vessel, operation, week, billion, shipping, unit
s_{10}	strike, port, union, spokesman, cargo, employer, worker, sector, redundancy, court
s_{11}	oil, export, dlr, industry, year, pct, future, company, report, price
s_{12}	reuter, pct, report, national, week, brazil, today, increase, pay, april
s_{13}	trade, japan, japanese, reagan, state, tariff, unite, market, washington, official
s_{14}	grain, mln, soviet, crop, tonne, year, usda, production, fall, analyst
s_{15}	trade, talk, gulf, gatt, bill, yeutter, round, reuter, call, negotiation
s_{16}	certificate, reuter, cost, government, program, agreement, agriculture, will, study, loan
s_{17}	year, official, import, will, state, price, government, china, land, rise
s_{18}	mln, ct, loss, net, shr, dlr, profit, qtr, reuter, year
s_{19}	oil, mln, will, barrel, dlr, crude, source, level, petroleum, day

Table 7: PLSV's Topic Model for *Reuters*8 (for 20 topics)

Topic ID	Top 10 Words
p_0	will, oil, company, reuter, industry, canada, price, shell, raise, sell
p_1	rate, currency, dollar, exchange, baker, west, will, bank, reuter, treasury
p_2	bank, pct, day, import, year, rate, export, february, expect, reuter
p_3	share, company, corp, offer, stock, board, will, reuter, dlr, buy
p_4	rate, bank, pct, prime, cut, point, interest, market, lower, savings
p_5	market, bank, stock, price, japan, ministry, rise, official, gulf, bond
p_6	reuter, pct, week, report, year, march, mark, american, commission, figure
p_7	mln, ct, loss, net, dlr, shr, year, profit, qtr, reuter
p_8	mln, pct, billion, stg, dlr, reuter, market, january, revise, rise
p_9	billion, dlr, rate, market, surplus, currency, reserve, trading, dollar, foreign
p_{10}	oil, opec, price, bpd, pipeline, mln, crude, official, dlr, output
p_{11}	crude, dlr, barrel, corp, capacity, refinery, oil, company, offer, group
p_{12}	reuter, official, state, cut, gulf, government, today, action, force, tell
p_{13}	oil, government, indonesium, price, foreign, bank, billion, reserve, company, industry
p_{14}	certificate, company, mln, year, grain, cooperative, program, dlr, government, cost
p_{15}	year, trade, agriculture, reuter, grain, agreement, gatt, yeutter, financial, agricultural
p_{16}	strike, port, union, spokesman, employer, brazil, cargo, worker, redundancy, sector
p_{17}	trade, japan, japanese, reagan, tariff, unite, washington, state, nakasone, semiconductor
p_{18}	grain, mln, crop, tonne, soviet, year, official, china, pct, offer
p_{19}	trade, country, minister, talk, state, meeting, economic, exchange, issue, baldrige

Table 8: Semafore's Topic Model for *Cade*12 (for 20 topics)

Topic ID	Top 10 Words
s_0	sp, aulas, tecnologia, rj, sao, area, janeiro, particulares, areas, fisica
s_1	terra, jun, gif, busca, virtual, brasil, forum, tempo, noticias, revistas
s_2	trabalho, seguranca, saude, medicina, ocupacional, prevencao, ppra, pcmso, imagem, imagens
s_3	peixes, cade, lazer, pesca, agua, rio, praia, hotel, sao, doce
s_4	agar, vida, personal, fisica, base, tratamento, tem, pode, sistema, trainer
s_5	sao, br, rio, sul, criancas, www, escola, mail, http, atendimento
s_6	links, page, home, fotos, pagina, dicas, download, tenis, informacoes, jogos
s_7	internet, informatica, acesso, mg, br, servicos, provedor, mail, revista, horizonte
s_8	servicos, sao, paulo, entregas, entrega, sp, cesta, express, empresa, servico
s_9	pesca, sp, grupo, brasil, eventos, video, mg, informacoes, turismo, danca
s_{10}	astronomia, pagina, jose, foi, bem, espaco, tem, veja, filosofia, correio
s_{11}	mp, banda, musicas, rock, musica, page, letras, bandas, pagina, site
s_{12}	historia, cultura, mundo, site, page, brasil, informacoes, rs, livro, arte
s_{13}	noticias, jornal, cidade, sp, sao, regiao, demolay, ordem, rio, capitulo
s_{14}	empresas, informacoes, informacao, dados, atraves, textos, mail, equipe, unicamp, centro
s_{15}	engenharia, servicos, projetos, empresa, consultoria, quimica, instituto, pesquisa, rio, manutencao
s_{16}	site, informacoes, brasil, associacao, educacao, pagina, organizacao, centro, brasileira, direitos
s_{17}	software, web, empresa, sistemas, sistema, br, marketing, desenvolvimento, windows, dados
s_{18}	virtual, online, venda, produtos, cade, shopping, internet, loja, compras, cursos
s_{19}	futebol, informacoes, fotos, clube, historia, paulo, sao, quake, pagina, cade

Table 9: PLSV's Topic Model for *Cade*12 (for 20 topics)

Topic ID	Top 10 Words
p_0	engenharia, projetos, servicos, trabalho, empresa, consultoria, seguranca, sp, medicina, sao
p_1	sao, ong, rio, instituto, personal, educacao, organizacao, sp, paulo, fins
p_2	sao, br, desenvolvimento, sistema, tratamento, mail, sistemas, clientes, informacoes, empresa
p_3	aulas, formula, quimica, particulares, informacoes, matematica, pilotos, fotos, fisica, site
p_4	jornal, tenis, noticias, esportes, sp, informacoes, sao, esporte, fotos, links
p_5	musica, page, rock, bandas, links, home, pagina, musicas, music, fotos
p_6	pesca, demolay, sp, peixes, sao, fotos, ordem, capitulo, paulo, jitsu
p_7	mp, musicas, nacionais, agar, internacionais, rock, formato, site, page, pagina
p_8	pesquisa, tecnologia, informacoes, cade, ciencia, geografia, pesquisas, area, instituto, pagina
p_9	site, pagina, internet, mail, clique, veja, br, pode, foi, links
p_{10}	astronomia, informacoes, cultura, site, pagina, brasil, home, page, fotos, historia
p_{11}	banda, fotos, rock, letras, page, musicas, pagina, site, home, mp
p_{12}	internet, provedor, acesso, mg, informatica, software, servicos, belo, horizonte, manutencao
p_{13}	futebol, clube, sao, paulo, campeonato, historia, informacoes, pagina, turismo, tricolor
p_{14}	noticias, terra, internet, brasil, informatica, online, jornal, virtual, servicos, busca
p_{15}	links, page, quake, home, pagina, fotos, dicas, mp, download, informacoes
p_{16}	grupo, banda, karate, pagina, page, informacoes, fotos, home, rio, historia
p_{17}	produtos, virtual, shopping, cade, venda, online, sao, rio, loja, compras
p_{18}	br, sao, informacoes, marketing, mail, empresa, internet, www, fax, site
p_{19}	vida, dia, sao, foi, terra, panico, jose, tem, planetas, grande

References

Akkucuk, U., & Carroll, J. D. (2006). PARAMAP vs. Isomap: a comparison of two nonlinear mapping algorithms. *Journal of Classification*, *23*(2), 221–254.

Bai, L., Guo, J., Lan, Y., & Cheng, X. (2014). Local Linear Matrix Factorization for Document Modeling. In *Advances in Information Retrieval*, pp. 398–411. Springer.

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 14, pp. 585–591.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, *15*(6), 1373–1396.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research (JMLR)*, *7*, 2399–2434.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bishop, C. M., Svensén, M., & Williams, C. K. (1998). GTM: The generative topographic mapping. *Neural Computation*, *10*(1), 215–234.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research (JMLR)*, *3*, 993–1022.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Linguistic Data Consortium, Philadelphia.

Buhmann, M. D. (2000). Radial basis functions. *Acta Numerica 2000*, *9*.

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*.

Cai, D., Wang, X., & He, X. (2009). Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Cardoso-Cachopo, A. (2007). Improving Methods for Single-label Text Categorization. PhD Thesis, Instituto Superior Tecnico, Universidade Tecnica de Lisboa.

Carey, C., & Mahadevan, S. (2014). Manifold Spanning Graphs. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Chaney, A. J.-B., & Blei, D. M. (2012). Visualizing Topic Models. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pp. 288–296.

Chen, L., & Buja, A. (2009). Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Journal of the American Statistical Association*, *104*(485), 209–219.

Chi, E. H.-h. (2000). A taxonomy of visualization techniques using the data state reference model. In *Proceedings of the IEEE Symposium on Information Visualization (InfoVis)*, pp. 69–75.

Choo, J., Lee, C., Reddy, C. K., & Park, H. (2013). UTOPIAN: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, *19*(12), 1992–2001.

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: visualization techniques for assessing textual topic models. In *Proceedings of the International Working Conference on Advanced Visual Interfaces (AVI)*, pp. 74–77.

Coifman, R. R., & Lafon, S. (2006). Diffusion maps. *Applied and Computational Harmonic Analysis*, *21*(1), 5 – 30.

Comon, P. (1994). Independent component analysis, a new concept?. *Signal Processing*, *36*(3), 287–314.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *39*(1), 1–38.

Dumais, S., Furnas, G., Landauer, T., Deerwester, S., Deerwester, S., et al. (1995). Latent semantic indexing. In *Proceedings of the Text Retrieval Conference*.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188.

Gabrilovich, E., & Markovitch, S. (2009). Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research (JAIR)*, *34*(2), 443.

Golub, G. H., & Van Loan, C. F. (2012). *Matrix Computations*, Vol. 3. JHU Press.

Gretarsson, B., O'donovan, J., Bostandjiev, S., Höllerer, T., Asuncion, A., Newman, D., & Smyth, P. (2012). TopicNets: Visual analysis of large text corpora with topic modeling. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *3*(2), 23.

Hein, M., Audibert, J.-y., & Luxburg, U. V. (2007). Graph Laplacians and their Convergence on Random Neighborhood Graphs. In *Journal of Machine Learning Research*, pp. 1325–1368.

Hinton, G. E., & Roweis, S. T. (2002). Stochastic neighbor embedding. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 833–840.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 50–57.

Hu, Y., Boyd-Graber, J., Satinoff, B., & Smith, A. (2014). Interactive topic modeling. *Machine Learning*, *95*(3), 423–469.

Huh, S., & Fienberg, S. E. (2012). Discriminative topic modeling based on manifold learning. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, *5*(4), 20.

Iwata, T., Saito, K., Ueda, N., Stromsten, S., Griffiths, T. L., & Tenenbaum, J. B. (2007). Parametric embedding for class visualization. *Neural Computation*, *19*(9), 2536–2556.

Iwata, T., Yamada, T., & Ueda, N. (2008). Probabilistic latent semantic visualization: topic model for visualizing documents. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 363–371.

Jebara, T., Wang, J., & Chang, S.-F. (2009). Graph construction and b-matching for semi-supervised learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 441–448. ACM.

Jolliffe, I. (2005). *Principal Component Analysis*. Wiley Online Library.

Kim, M., & Torre, F. (2010). Local minima embedding. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 527–534.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Lafferty, J. D., & Wasserman, L. (2007). Statistical Analysis of Semi-Supervised Regression. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 801–808.

Le, T., & Lauw, H. W. (2014a). Semantic visualization for spherical representation. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1007–1016. ACM.

Le, T. M., & Lauw, H. W. (2014b). Manifold learning for jointly modeling topic and visualization. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*, 503–528.

Manning, C. D., Raghavan, P., Schütze, H., et al. (2008). *Introduction to Information Retrieval*, Vol. 1. Cambridge University Press Cambridge.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email.. *Journal of Artificial Intelligence Research (JAIR)*, *30*, 249–272.

Millar, J. R., Peterson, G. L., & Mendenhall, M. J. (2009). Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps. In *FLAIRS Conference*, Vol. 21, pp. 69–74.

Newman, D., Karimi, S., & Cavedon, L. (2009). External evaluation of topic models. In *Australasian Document Computing Symposium (ADCS)*.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, *3*(2), 246–257.

Ponzetto, S. P., & Strube, M. (2007). Knowledge derived from Wikipedia for computing semantic relatedness.. *Journal of Artificial Intelligence Research (JAIR)*, *30*, 181–212.

Reisinger, J., Waters, A., Silverthorn, B., & Mooney, R. J. (2010). Spherical topic models. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 903–910.

Roweis, S. T., & Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*(5500), 2323–2326.

Shaw, B., & Jebara, T. (2007). Minimum volume embedding. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 460–467.

Shaw, B., & Jebara, T. (2009). Structure preserving embedding. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 937–944. ACM.

Tenenbaum, J. B., De Silva, V., & Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, *290*(5500), 2319–2323.

Ting, D., Huang, L., & Jordan, M. I. (2010). An Analysis of the Convergence of Graph Laplacians. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Turney, P. D., Pantel, P., et al. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research (JAIR)*, *37*(1), 141–188.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research (JMLR)*, *9*(2579-2605), 85.

Wei, F., Liu, S., Song, Y., Pan, S., Zhou, M. X., Qian, W., Shi, L., Tan, L., & Zhang, Q. (2010). Tiara: a visual exploratory text analytic system. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 153–162.

Wu, H., Bu, J., Chen, C., Zhu, J., Zhang, L., Liu, H., Wang, C., & Cai, D. (2012). Locally discriminative topic modeling. *Pattern Recognition*, *45*(1), 617–625.

Zemel, R. S., & Carreira-Perpiñán, M. Á. (2004). Proximity graphs for clustering and manifold learning. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 225–232.

Zhou, D., Bousquet, O., Lal, T. N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. *Advances in Neural Information Processing Systems (NIPS)*, *16*(16).

Zhu, X., Ghahramani, Z., Lafferty, J., et al. (2003). Semi-supervised learning using Gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, Vol. 3, pp. 912–919.