School of
Information Systems

SMU
SINGAPORE MANAGEMENT UNIVERSITY

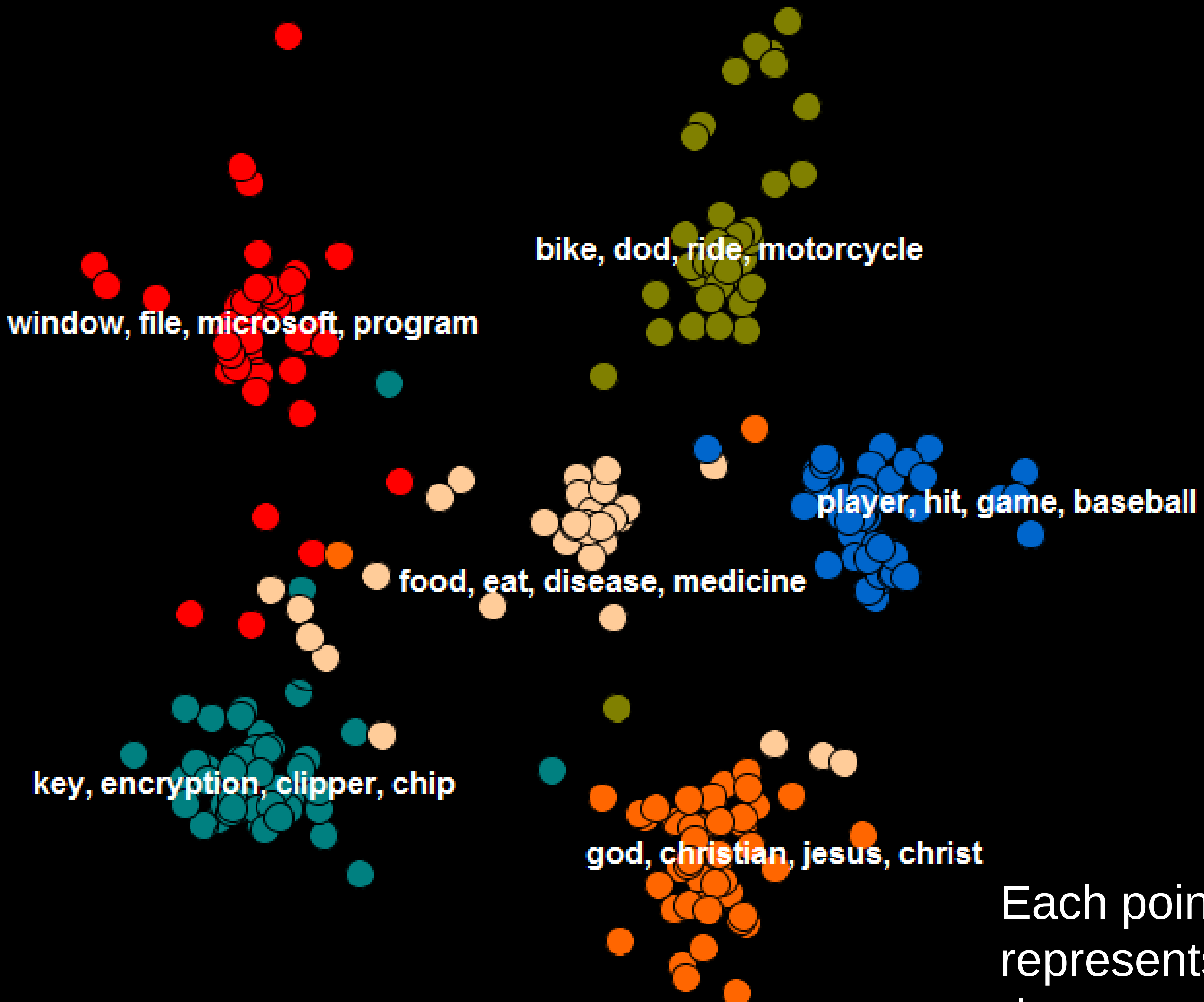
Semantic Visualization for Spherical Representation

Tuan M. V. Le and Hady W. Lauw

KDD 2014

Visualize Document Collections

- To present the contents/semantics/themes/etc of the documents.
- To show the **similarities** among documents in a collection.

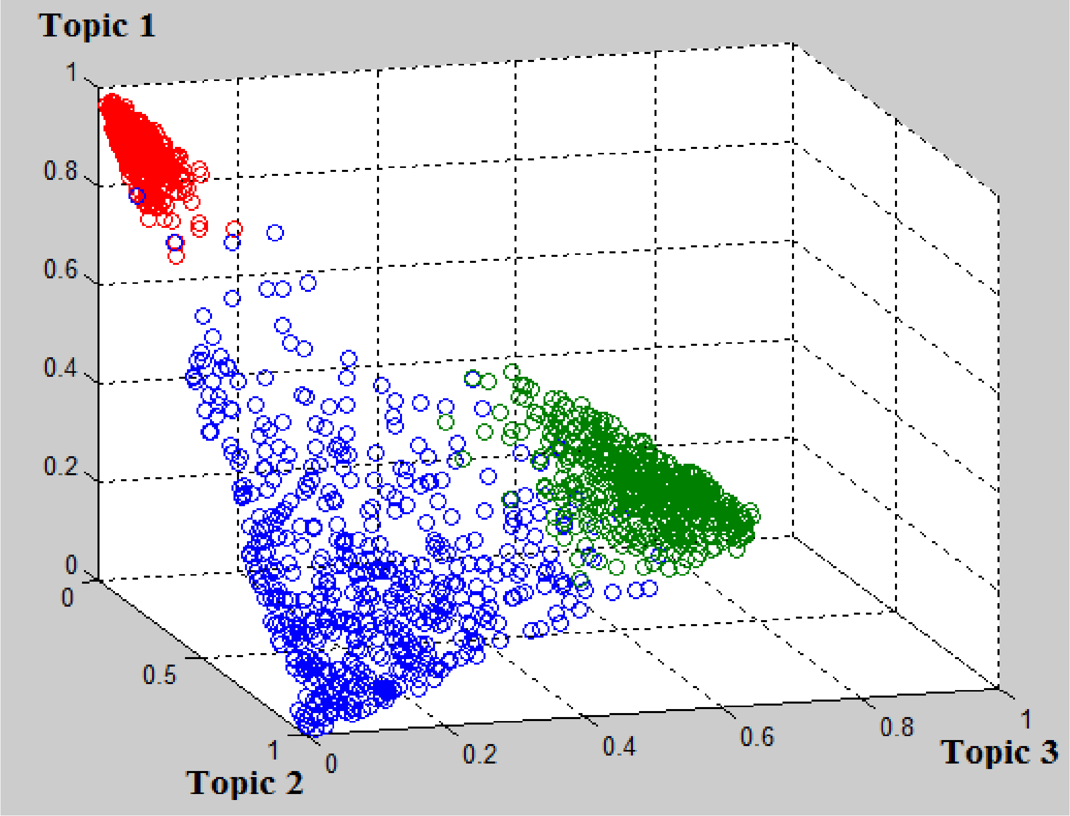


Each point represents a document.

Topic Model

"Arts"	"Budgets"	"Children"	"Education"
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI

Topic Model Not Intended for Visualization



Semantic Visualization Problem

Given a set of documents
(bag of words)

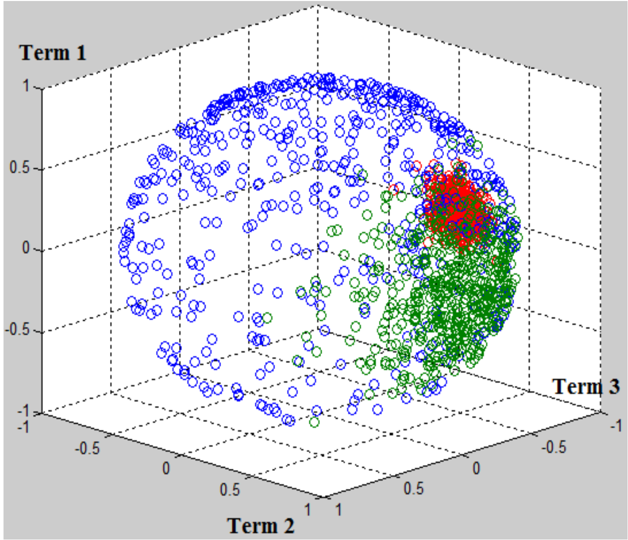
Find 2-dimensional visualization coordinates

+

Find K-dimensional topic distributions

Our Approach for Semantic Visualization

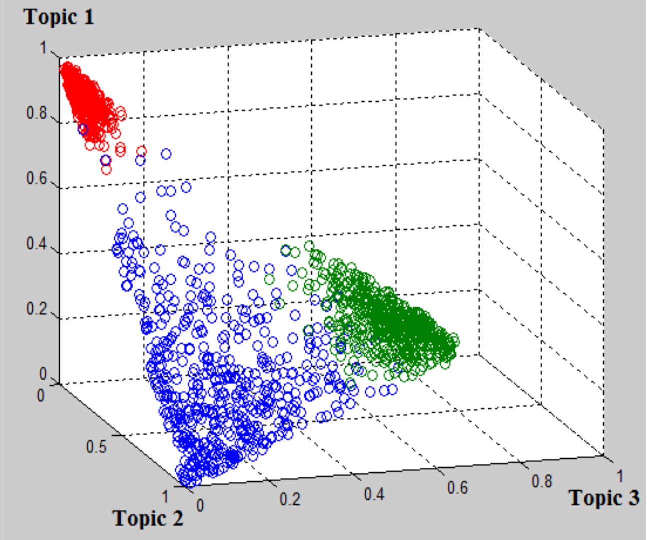
Spherical



L^2-normalized vector

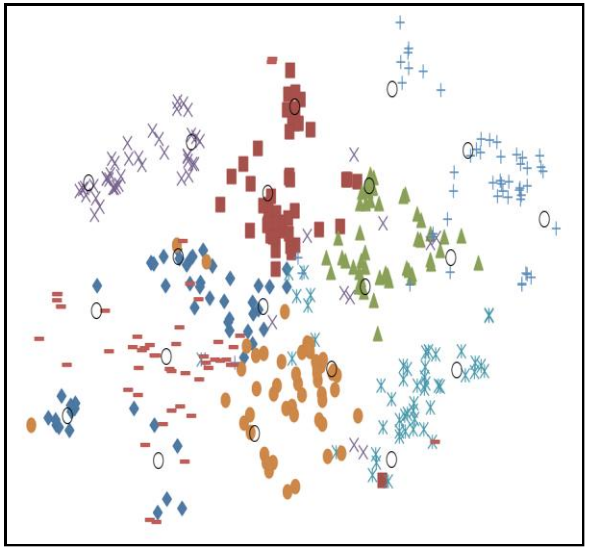
+ Document: v_n
+ Topic: τ_z

Semantic



Topic distribution

Embedding



Visualization coordinates

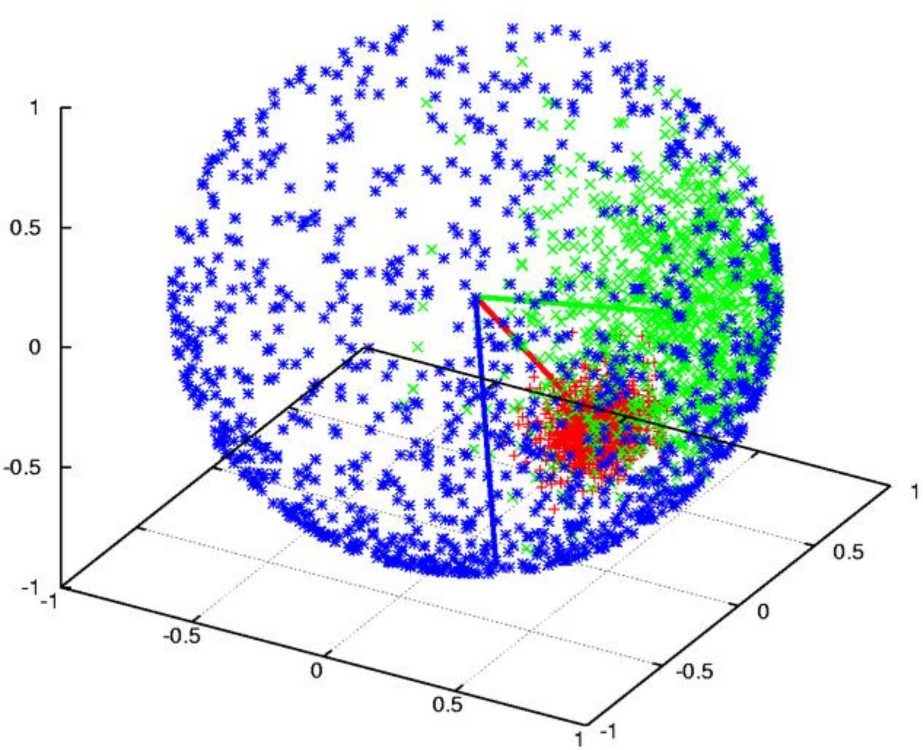
+ Document: x_n
+ Topic: ϕ_z

School of
Information Systems

SMU
SINGAPORE MANAGEMENT UNIVERSITY

Data Representation – Word Space

Spherical



- Richer feature representations
  - ➢ tf, tf-idf,…
- Model directly absences of word.
- Similarity is based on cosine distance:
  - ➢ Not sensitive to document length.

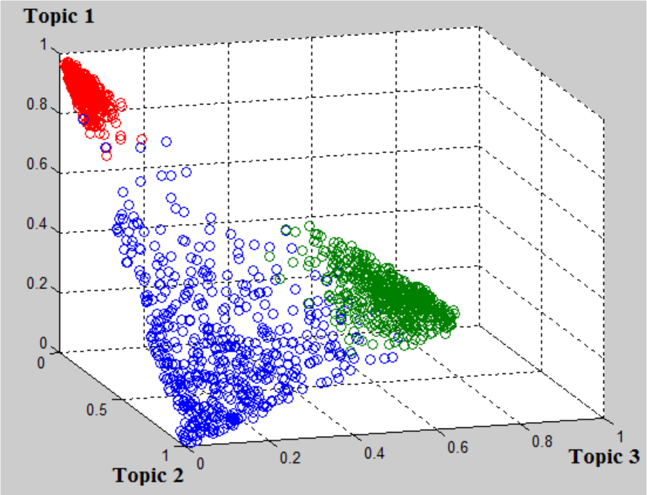
Von Mises–Fisher distribution (vMF):

$$f_p(x;\mu,\kappa) = C_p(\kappa)\exp(\kappa\mu^T x)$$

Data Representation – Topic Space

Each document is represented as a point on the topic simplex

Semantic



Topic distribution

Topic Representation

- Each topic is represented as a point on the sphere (word space)

Multinomial (Sum up to 1)

	Probability
Word 1	0.7
Word 2	0.2
Word 3	0.1

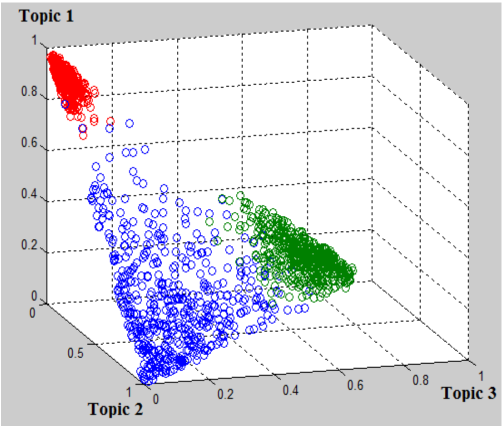
Spherical (Unit length)

	Weight
Word 1	0.95
Word 2	0.27
Word 3	0.14

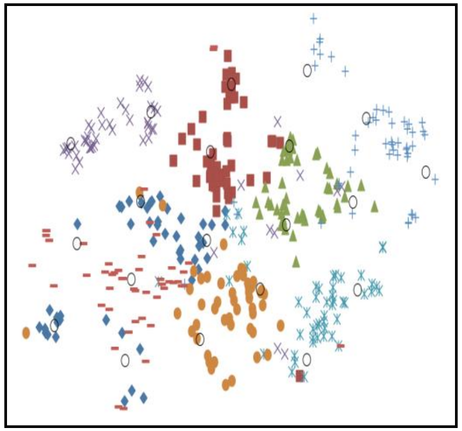
Relationship between Topic Distributions and Visualization Coordinates

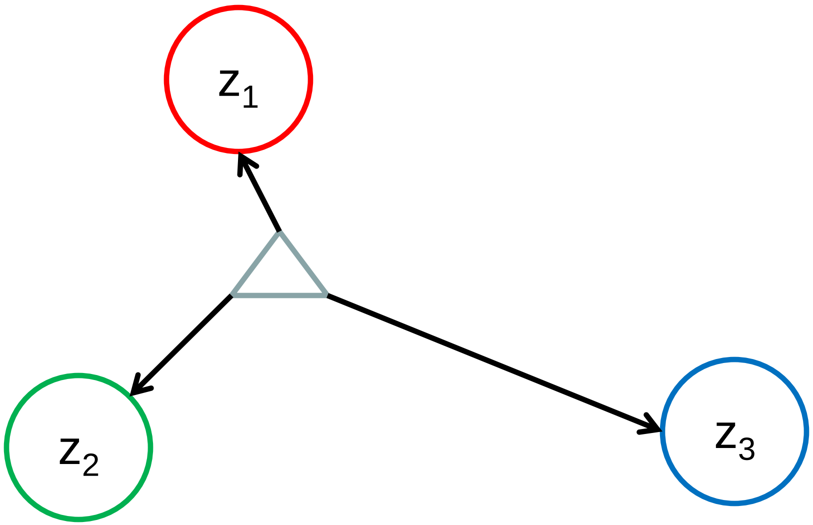
Semantic **E**mbedding

$P(z_1 | d) > P(z_2 | d) > P(z_3 | d)$



Topic distribution


Visualization coordinates
+ Document: x_n
+ Topic: ϕ_z



Shorter visualization distance means greater topic probability.

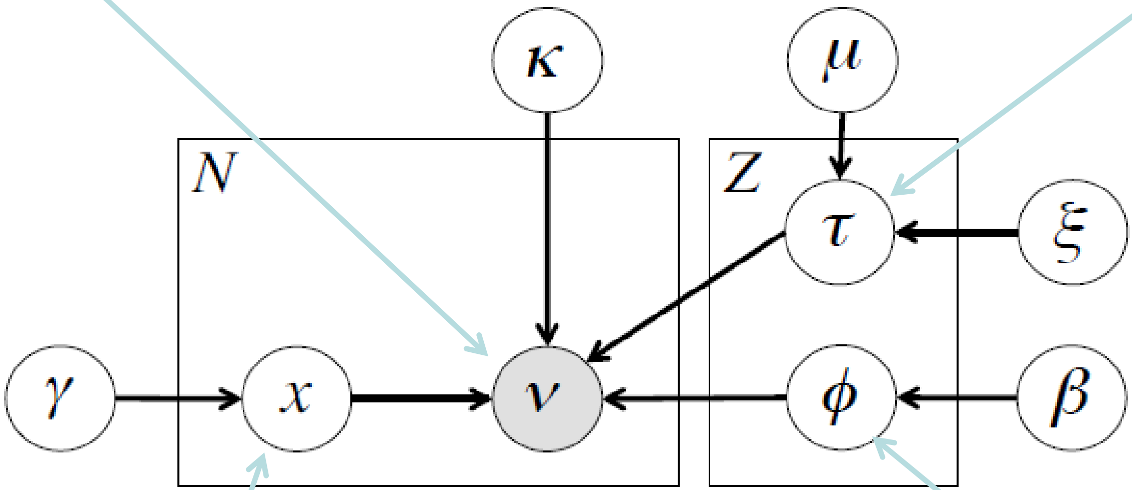
$$\mathrm{P}(z|d_n) = \mathrm{P}(z|x_n, \Phi) = \frac{\exp(-\frac{1}{2}||x_n - \phi_z||^2)}{\sum_{z'=1}^{Z} \exp(-\frac{1}{2}||x_n - \phi_{z'}||^2)}$$

This formula is also used in PLSV, Iwata et al., KDD 2008

Spherical Semantic Embedding

The observed L2-normalized word vector of d_n

L2-normalized word vector of topic z



Visualization coordinate of document d_n

Visualization coordinate of topic z

Generative Process

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

Parameter Estimation

- Variational EM with MAP estimation.

Comparative Methods

Visualization

t-SNE

Semantic Visualization

Joint

Pipeline

Spherical

SSE (Ours)

Multinomial

PLSV

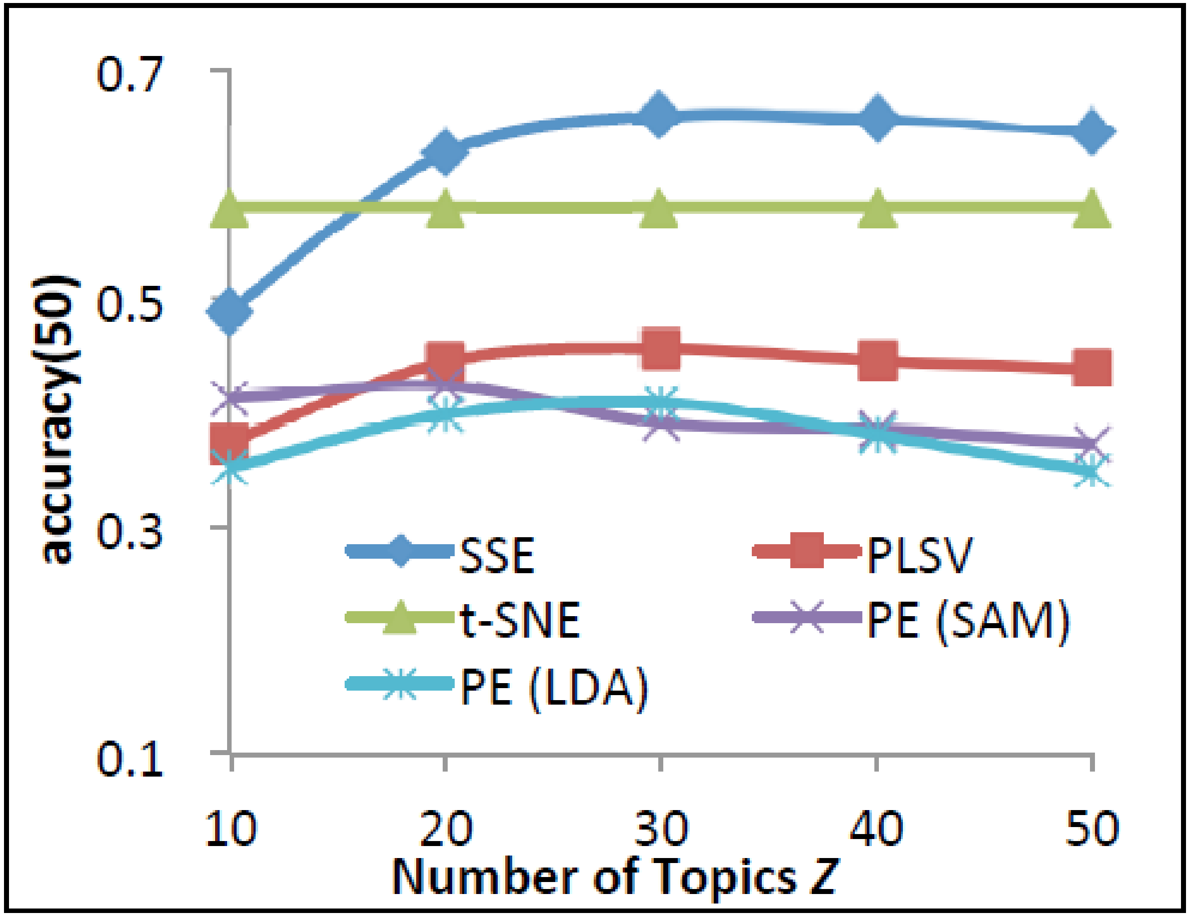
Spherical

PE (SAM)

Multinomial

PE (LDA)

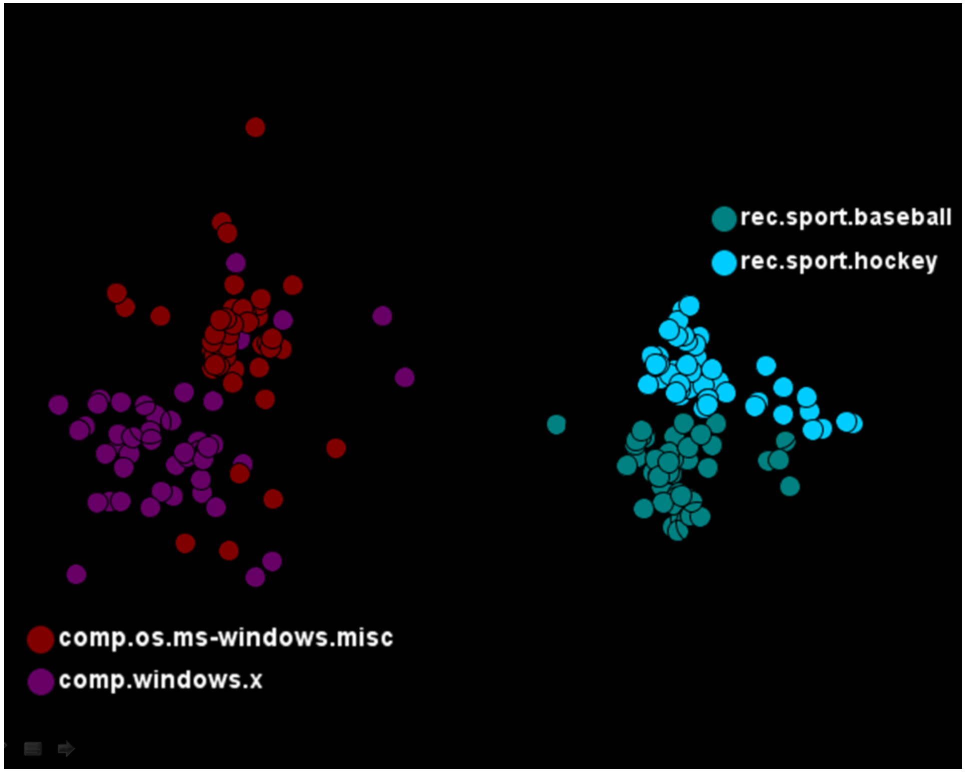
20News Dataset (20 Categories)



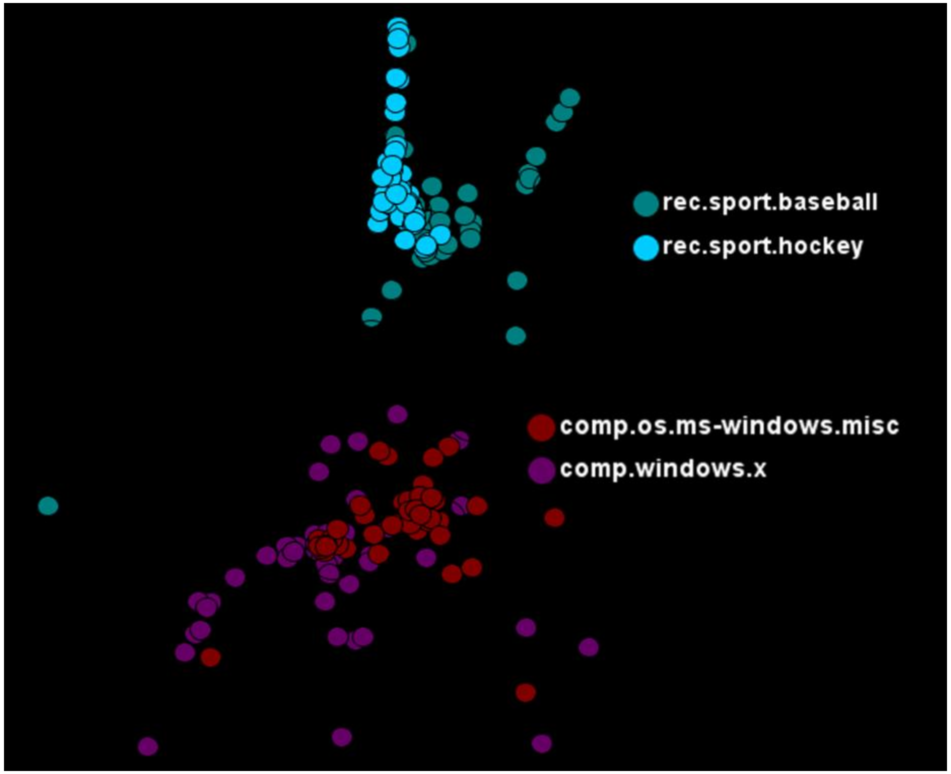
a. 20*News*

Visualization Comparison

SSE (Spherical)



PLSV (Multinomial)



Conclusion

- Spherical Semantic Embedding (SSE) is designed for data with spherical representation.

- Promising applications for integrated modeling:
  - semantic-rich visualizations
  - assigning categories to documents