

# **Part II: Embedding-Driven Topic Discovery**

**WWW 2023 Tutorial**

**Turning Web-Scale Texts to Knowledge: Transferring Pretrained Representations to Text Mining Applications**

**Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han**

**Computer Science, University of Illinois at Urbana-Champaign**

**April 30, 2023**

**Tutorial Website:**



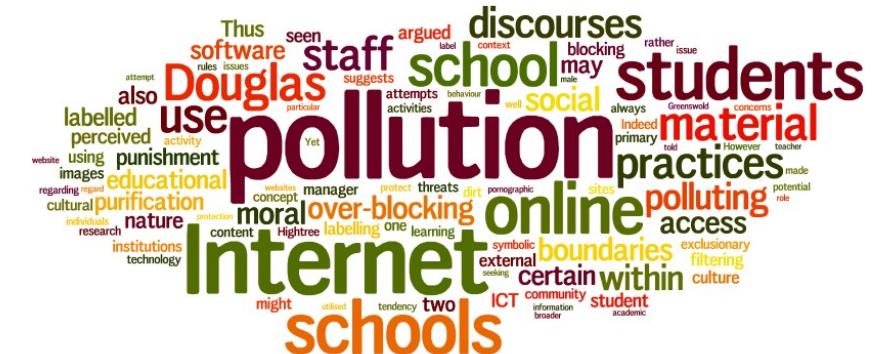
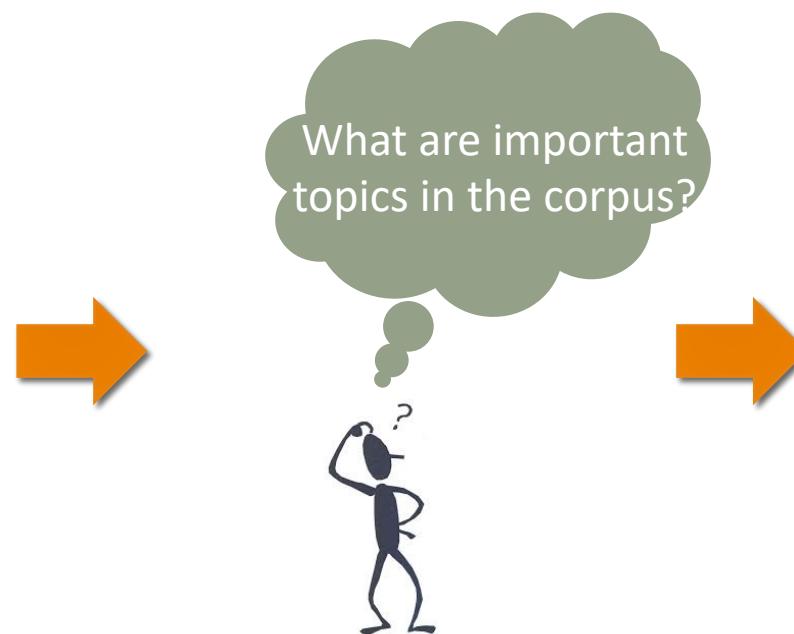
# Outline

---

- ❑ Unsupervised Topic Modeling 
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-Based Topic Discovery
- ❑ Discriminative Topic Mining

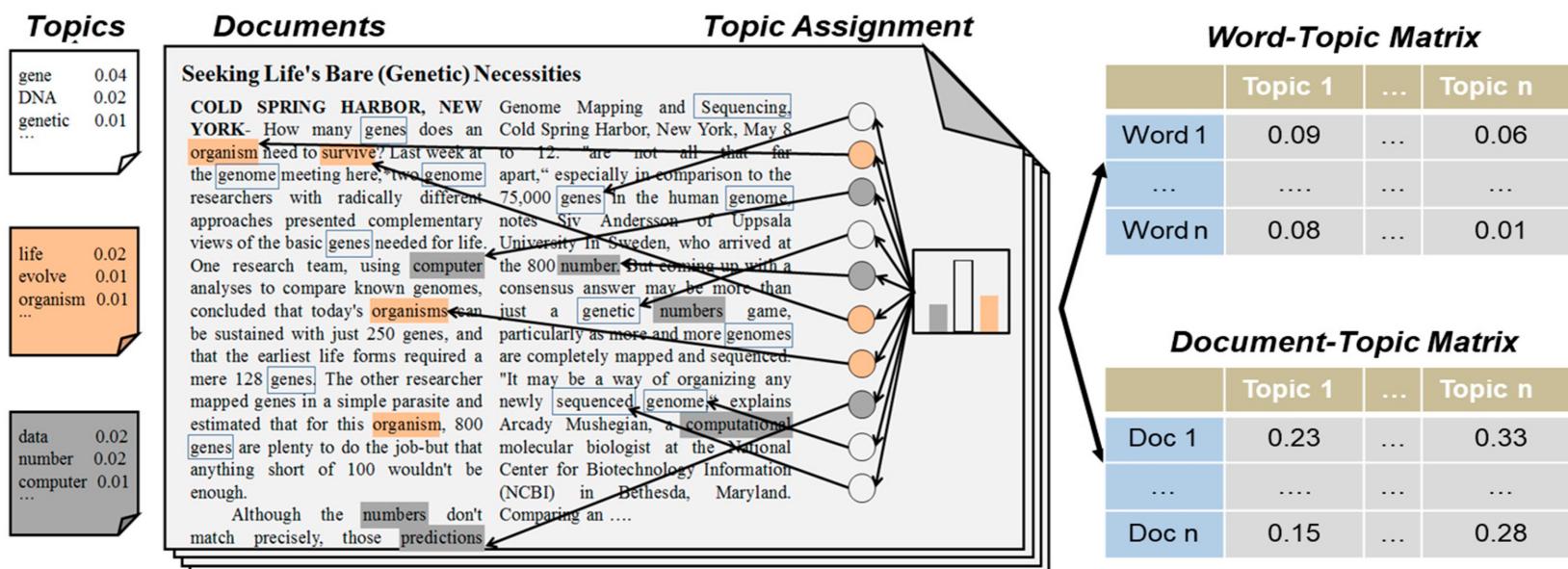
# Topic Modeling: Introduction

- ❑ How to effectively & efficiently comprehend a large text corpus?
  - ❑ Knowing what important topics are there is a good starting point!
  - ❑ Topic discovery facilitates a wide spectrum of applications
    - ❑ Document classification/organization
    - ❑ Document retrieval/ranking
    - ❑ Text summarization



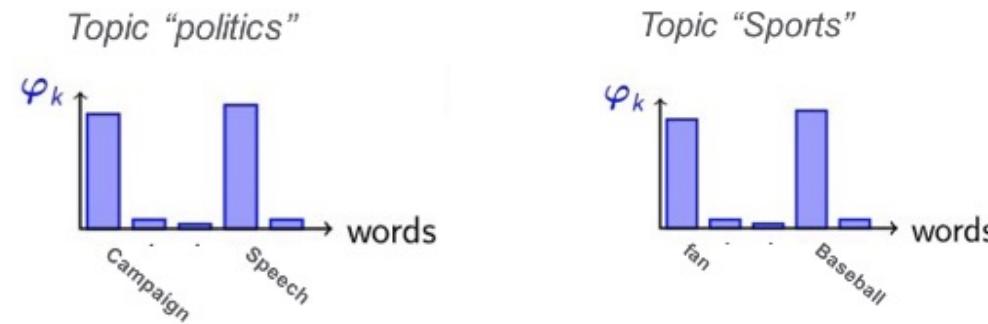
# Topic Modeling: Overview

- How to discover topics automatically from the corpus?
- By modeling the corpus statistics!
  - Each document has a latent topic distribution
  - Each topic is described by a different word distribution



# Latent Dirichlet Allocation (LDA): Overview

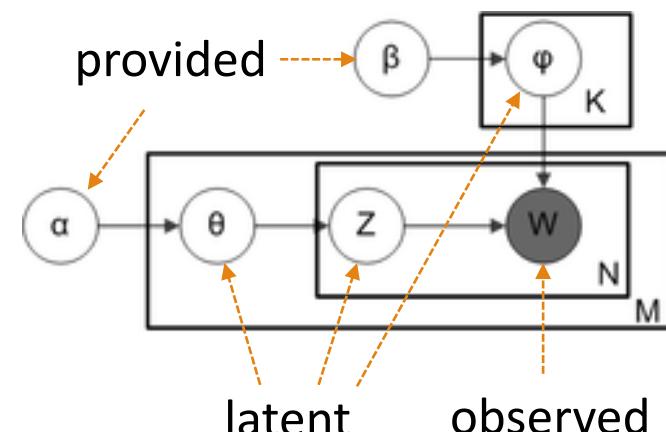
- Each document is represented as a mixture of various topics
  - Ex. A news document may be 40% on politics, 50% on economics, and 10% on sports
- Each topic is represented as a probability distribution over words
  - Ex. The distribution of “politics” vs. “sports” might be like:



- Dirichlet priors are imposed to enforce sparse distributions:
  - Documents cover only a small set of topics (sparse document-topic distribution)
  - Topics use only a small set of words frequently (sparse topic-word distribution)

# LDA: Inference

- Learning the LDA model (Inference)
- What need to be learned
  - Document-topic distribution  $\theta$  (for assigning topics to documents)
  - Topic-word distribution  $\varphi$  (for topic interpretation)
  - Words' latent topic  $z$
- How to learn the latent variables? – complicated due to intractable posterior
  - Monte Carlo simulation
  - Gibbs sampling
  - Variational inference
  - ...



# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-Based Topic Discovery
- ❑ Discriminative Topic Mining



# Issues with LDA

---

- ❑ LDA is completely unsupervised (i.e., users only input number of topics)
- ❑ Cannot take user supervision
- ❑ Ex. What if a user is specifically interested in some topics but LDA doesn't discover them?

	Topic 1	Weight	Topic 2	Weight	Topic 3	Weight	Topic 4	Weight	Topic 5	Weight
0	life	0.018076	father	0.059603	official	0.017620	case	0.021908	art	0.010555
1	man	0.017714	graduate	0.048363	force	0.015388	law	0.020698	open	0.010413
2	woman	0.016657	son	0.042746	military	0.014587	court	0.019967	room	0.010363
3	book	0.010486	mrs	0.041379	war	0.011381	lawyer	0.016935	house	0.009002
4	family	0.010382	daughter	0.037156	government	0.010564	state	0.014501	building	0.008722
5	young	0.009896	mother	0.034542	troop	0.008949	judge	0.012487	artist	0.008264
6	write	0.009493	receive	0.029211	attack	0.008886	legal	0.011141	design	0.008162
7	child	0.009460	marry	0.029038	leader	0.008082	rule	0.009854	floor	0.008034
8	live	0.008819	yesterday	0.024107	peace	0.006835	decision	0.009261	museum	0.007917
9	love	0.007814	degree	0.022899	soldier	0.006562	file	0.008289	exhibition	0.007222

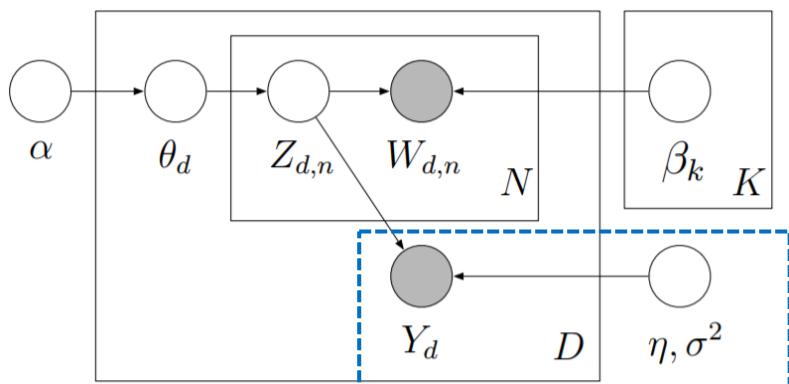
  

	Topic 6	Weight	Topic 7	Weight	Topic 8	Weight	Topic 9	Weight	Topic 10	Weight
0	group	0.051052	market	0.024976	serve	0.010918	change	0.007661	city	0.021776
1	member	0.040683	stock	0.024874	add	0.010185	system	0.007233	area	0.014865
2	meeting	0.016390	share	0.020583	minute	0.009301	problem	0.006835	build	0.014361
3	issue	0.014988	price	0.018141	pepper	0.009235	power	0.005400	building	0.014326
4	official	0.013069	sell	0.016564	oil	0.008976	create	0.005056	home	0.013632
5	support	0.011994	buy	0.015415	cook	0.008711	research	0.004712	resident	0.013483
6	leader	0.011799	company	0.015249	food	0.008689	produce	0.004574	community	0.012479
7	organization	0.011135	investor	0.015062	cup	0.008682	far	0.004447	local	0.010686
8	meet	0.010235	yesterday	0.012813	sauce	0.008209	result	0.004280	live	0.010661
9	effort	0.008479	analyst	0.010768	small	0.007864	kind	0.004166	project	0.010459

10 topics generated by LDA on The New York Times dataset

# Supervised LDA (sLDA)

- Allow users to provide document annotations/labels
- Incorporate document labels into the generative process
  - For the  $i$ th document, choose  $\theta_i \sim \text{Dir}(\alpha)$  document's topic distribution
  - For the  $j$ th word in the  $i$ th document,
    - choose topic  $z_{i,j} \sim \text{Categorical}(\theta_i)$  word's topic
    - choose a word  $w_{i,j} \sim \text{Categorical}(\beta_{z_{i,j}})$
  - For the  $i$ th document, choose  $y_i \sim N(\eta^\top \bar{z}_i, \sigma^2)$ ,  $\bar{z}_i = \frac{1}{L} \sum_{j=1}^L z_{i,j}$



generate document's label

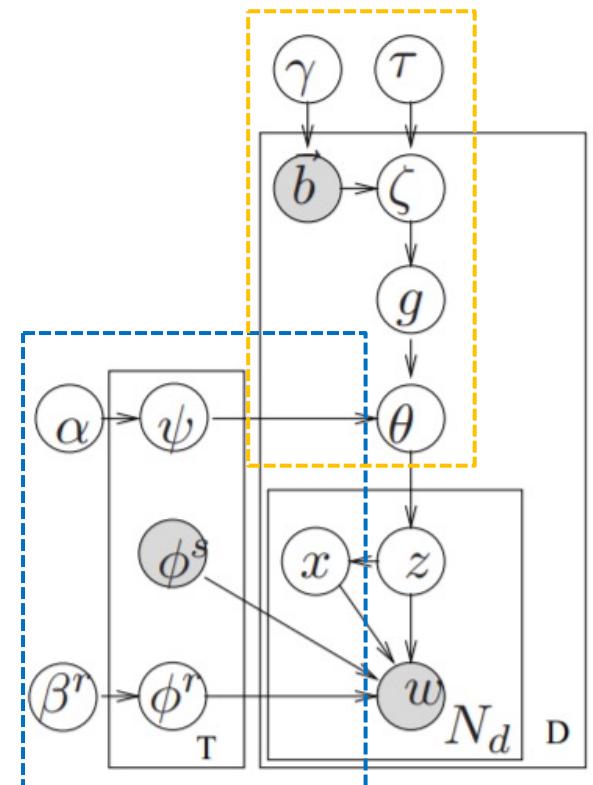
# Seeded LDA: Guided Topic-Word Distribution

- ❑ Another form of user supervision: several seed words for each topic

1. For each  $k=1 \dots T$ ,
  - (a) Choose regular topic  $\phi_k^r \sim \text{Dir}(\beta_r)$ .
  - (b) Choose seed topic  $\phi_k^s \sim \text{Dir}(\beta_s)$ .
  - (c) Choose  $\pi_k \sim \text{Beta}(1, 1)$ .
2. For each seed set  $s = 1 \dots S$ ,
  - (a) Choose group-topic distribution  $\psi_s \sim \text{Dir}(\alpha)$ .
3. For each document  $d$ ,
  - (a) Choose a binary vector  $\vec{b}$  of length  $S$ .
  - (b) Choose a document-group distribution  $\zeta^d \sim \text{Dir}(\tau \vec{b})$ .
  - (c) Choose a group variable  $g \sim \text{Mult}(\zeta^d)$ .
  - (d) Choose  $\theta_d \sim \text{Dir}(\psi_g)$ . // of length  $T$
  - (e) For each token  $i = 1 \dots N_d$ :
    - i. Select a topic  $z_i \sim \text{Mult}(\theta_d)$ .
    - ii. Select an indicator  $x_i \sim \text{Bern}(\pi_{z_i})$ .
    - iii. if  $x_i$  is 0
      - Select a word  $w_i \sim \text{Mult}(\phi_{z_i}^r)$ .
    - iv. if  $x_i$  is 1
      - Select a word  $w_i \sim \text{Mult}(\phi_{z_i}^s)$ .

Seed topics used to improve the document-topic distribution:  
Group-topic distribution = seed set distribution over regular topics  
Group-topic distribution used as prior to draw document-topic distribution

Seed topics used to improve the topic-word distribution:  
Each word comes from either “regular topics” with a distribution over all word like in LDA, or “seed topics” which only generate words from the seed set



# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-Based Topic Discovery 
- ❑ TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
- ❑ Discriminative Topic Mining

# Clustering-Based Topic Discovery

---

- ❑ Topic modeling frameworks use **bag-of-words** features (i.e., only word counts in documents matter; word ordering is ignored)
- ❑ In Part I of the tutorial, we introduced distributed text representations (text embeddings and language models) that better model sequential information in text
- ❑ Can we take advantage of those advanced text representations for the topic discovery task, as an alternative to topic modeling?

# Word Embedding + Clustering

---

- Cast “topics” as clusters of word types — similar to taking the top-ranked words from each topic’s distribution in topic modeling
- How to obtain word clusters? Run clustering algorithms on word embeddings
- Since the text embedding space captures word semantic similarity (i.e., high vector similarity implies high semantic similarity), using distance-based clustering algorithms (like K-means) will naturally group semantically similar words into the same cluster

# **Clustering-Based Topic Discovery: A benchmark study**

---

- ❑ Clustering algorithms:
  - ❑ k-means (KM)
  - ❑ Gaussian Mixture Models (GMM)
- ❑ Embeddings:
  - ❑ Word2Vec
  - ❑ GloVe
  - ❑ fastText
  - ❑ Spherical text embedding
  - ❑ ELMo
  - ❑ BERT

Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! EMNLP

# Clustering-Based Topic Discovery: Word Frequency

---

- One thing to consider is that text embeddings do not explicitly encode frequency information, which is important for topic discovery (i.e., more frequent words in the corpus may be more representative)
- Two ways to incorporate frequency information
  - Weighted clustering: Frequent words weigh more when computing cluster centroids
  - Rerank words in clusters: Rerank terms by frequency in each cluster when selecting representative terms

# Clustering-Based Topic Discovery: Results

- Using k-means (KM)/Gaussian Mixture Models (GMM) as clustering algorithm and using Spherical text embedding/BERT as representations leads to comparable results with LDA
- Future work
  - More advanced clustering algorithms?
  - Joint modeling of document-topic distribution via clustering?

	Reuters						20 Newsgroups									
	KM $\diamond$ GMM		KM $\diamond^w$ GMM		KM $\diamond_r$ GMM		KM $\diamond^w$ GMM		KM $\diamond$ GMM		KM $\diamond^w$ GMM		KM $\diamond_r$ GMM		KM $\diamond^w_r$ GMM	
Word2vec	-0.39	-0.47	-0.21	-0.09	0.02	0.01	0.03	0.08	-0.21	-0.10	-0.11	0.13	0.18	0.16	0.19	0.20
ELMo	-0.73	-0.55	-0.43	0.00	-0.10	-0.08	-0.02	0.06	-0.56	-0.13	-0.38	0.18	0.13	0.14	0.16	0.19
GloVe	-0.67	-0.59	-0.04	0.01	-0.27	-0.03	0.01	0.05	-0.18	-0.12	0.06	0.24	0.22	0.23	0.23	0.23
Fasttext	-0.68	-0.70	-0.46	-0.08	0.00	0.00	0.06	0.11	-0.32	-0.20	-0.18	0.21	0.24	0.23	0.25	0.24
Spherical	-0.53	-0.65	-0.07	0.09	0.01	-0.05	0.10	0.12	-0.05	-0.24	0.24	0.23	0.25	0.22	0.26	0.24
BERT	-0.43	-0.19	-0.07	0.12	0.00	-0.01	0.12	0.15	0.04	0.14	0.25	0.25	0.17	0.19	0.25	0.25
average	-0.57	-0.52	-0.21	0.01	-0.06	-0.03	0.05	0.10	-0.21	-0.11	-0.02	0.21	0.20	0.20	0.23	0.23
std. dev.	0.14	0.18	0.19	0.09	0.12	0.03	0.05	0.04	0.21	0.13	0.25	0.05	0.04	0.04	0.04	0.02

Table 1: NPMI Results (higher is better) for pre-trained word embeddings and k-means (KM), and Gaussian Mixture Models (GMM).  $\diamond^w$  indicates weighted and  $\diamond_r$  indicates reranking of top words. For Reuters (left table), LDA has an NPMI score of 0.12, while GMM $^w_r$  BERT achieves 0.15. For 20NG (right), both LDA and KM $^w_r$  Spherical achieve a score of 0.26. All results are averaged across 5 random seeds.

# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-Based Topic Discovery
  - ❑ TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22] 
- ❑ Discriminative Topic Mining

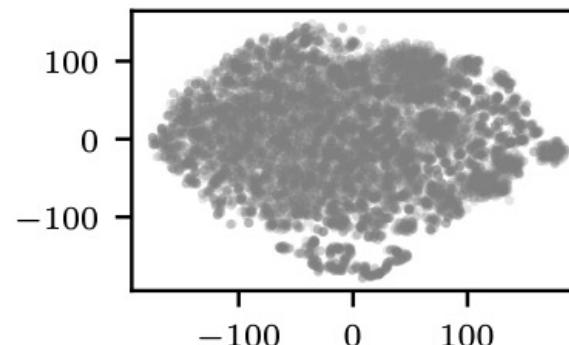
# Motivation

---

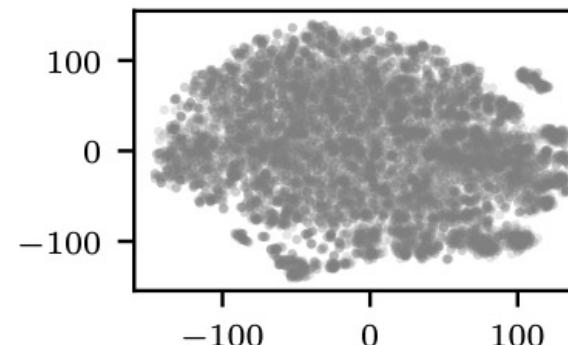
- Recently, pre-trained language models (LMs) have achieved enormous success in lots of tasks
  - They employ Transformer as the backbone architecture for capturing the **long-range, high-order** semantic dependency in text sequences, yielding superior representations
  - They are pre-trained on large-scale text corpora like Wikipedia, they carry **generic linguistic features** that can be generalized to almost any text-related applications
- Given the strong representation power of the contextualized embeddings, it is natural to consider simply **clustering** them as an alternative to topic models
- Topics are essentially interpreted via clusters of semantically coherent and meaningful words
- Interestingly, such an attempt has not been reported successful yet

# The Challenges

- Why not naively cluster pre-trained embeddings?
- Visualization: The embedding spaces do not exhibit clearly separated clusters
- Applying K-means with a typical K (e.g., K=100) to these spaces leads to low-quality and unstable clusters



(a) New York Times.



(b) Yelp Review.

**Figure 1: Visualization using t-SNE of 10,000 randomly sampled contextualized word embeddings of BERT on (a) NYT and (b) Yelp datasets, respectively. The embedding spaces do not have clearly separated clusters.**

# The Challenges

---

- Theoretically, such embedding space structure is due to **too many clusters**
- **Theorem:** The MLM pre-training objective of BERT assumes that the learned contextualized embeddings are generated from a Gaussian Mixture Model (GMM) with  $|V|$  mixture components where  $|V|$  is the vocabulary size of BERT.
- **Mismatch** between the number of clusters in the pre-trained LM embedding space and the number of topics to be discovered
  - If a smaller  $K$  ( $K \ll |V|$ ) is used, the resulting partition will not fit the original data well, resulting in unstable and low-quality clusters
  - If a bigger  $K$  ( $K \approx |V|$ ) is used, most clusters will contain only one unique term, which is meaningless for topic discovery

# The Latent Space Model

---

- We propose to project the original embedding space into a latent space with K clusters of words corresponding to K latent topics
- We assume that the latent space is **lower-dimensional** and **spherical**, with the following preferable properties:
  - **Spherical latent space** employs angular similarity between vectors to capture word semantic correlations, which works better than Euclidean metrics
  - **Lower-dimensional space** mitigates the “curse of dimensionality”
  - Projection from high-dimension to lower-dimension space forces the model to discard the information that is not helpful for forming topic clusters (e.g., syntactic features, “play”, “plays” and “playing” should not represent different topics)

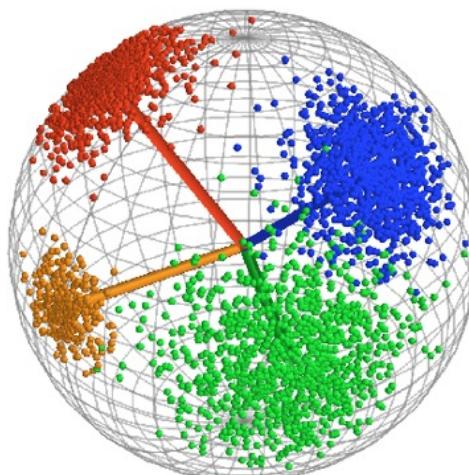
# Latent Topic Space

---

- We propose a generative model for the joint learning

$$t_k \sim \text{Uniform}(K), \mathbf{z}_i \sim \text{vMF}_{d'}(t_k, \kappa), \mathbf{h}_i = g(\mathbf{z}_i).$$

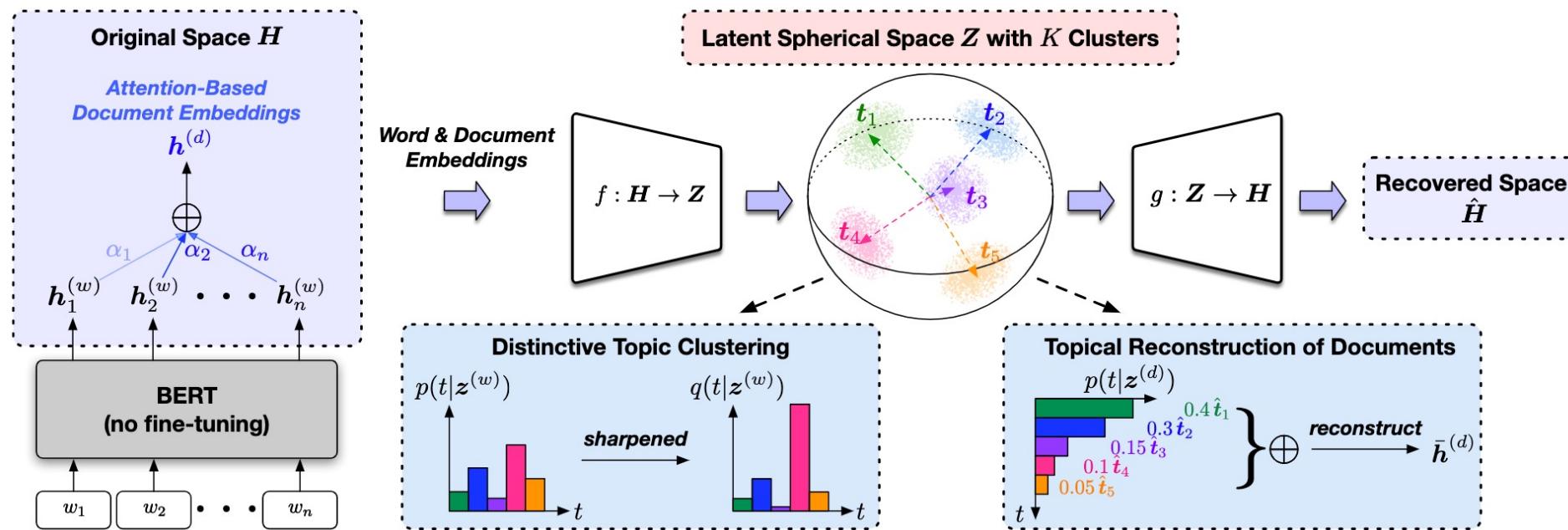
- A topic  $t$  is sampled from a uniform distribution over the  $K$  topics
- A latent embedding  $z$  is generated from the vMF distribution associated with topic  $t$



# The Latent Space Model

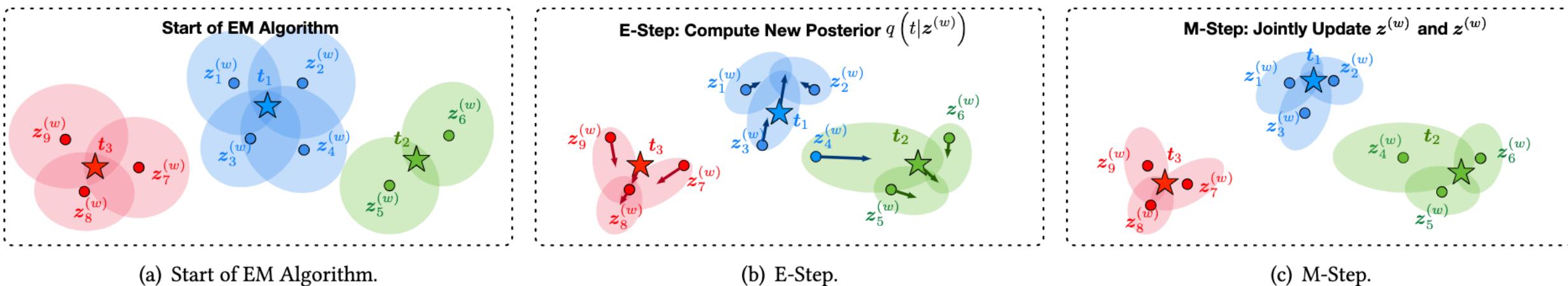
## □ How to train the generative model?

- A preservation loss that encourages the latent space to preserve the semantics of the original pre-trained LM induced embedding space (**preservation of original PLM embeddings**)
- A reconstruction loss to ensure the learned latent topics are meaningful summaries of the documents (**Topic reconstruction of documents**)
- A clustering loss that enforces separable cluster structures in the latent space for distinctive topic learning (**clustering**)



# The Clustering Loss

- An EM algorithm, analogous to K-means
  - The E-step estimates a new cluster assignment of each word based on the current parameters
  - The M-step updates the model parameters given the cluster assignments



(a) Start of EM Algorithm.

(b) E-Step.

(c) M-Step.

# Experiments

## □ Topic Discovery

### Quantitative

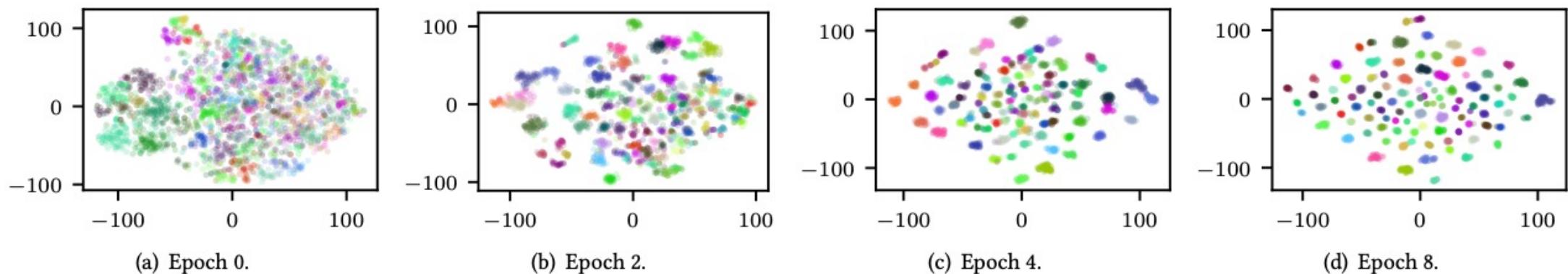
Methods	NYT				Yelp			
	UMass	UCI	Int.	Div.	UMass	UCI	Int.	Div.
LDA	-3.75	-1.76	0.53	0.78	-4.71	-2.47	0.47	0.65
CorEx	-3.83	-0.96	0.77	-	-4.75	-1.91	0.43	-
ETM	-2.98	-0.98	0.67	0.30	-3.04	-0.33	0.47	0.16
BERTopic	-3.78	-0.51	0.70	0.61	-6.37	-2.05	0.73	0.36
TopClus	<b>-2.67</b>	<b>-0.45</b>	<b>0.93</b>	<b>0.99</b>	<b>-1.35</b>	<b>-0.27</b>	<b>0.87</b>	<b>0.96</b>

### Qualitative

Methods	NYT					Yelp				
	Topic 1 (sports)	Topic 2 (politics)	Topic 3 (research)	Topic 4 (france)	Topic 5 (japan)	Topic 1 (positive)	Topic 2 (negative)	Topic 3 (vegetables)	Topic 4 (fruits)	Topic 5 (seafood)
LDA	olympic	<u>mr</u>	<u>said</u>	french	japanese	amazing	loud	spinach	mango	fish
	<u>year</u>	bush	report	<u>union</u>	tokyo	<u>really</u>	awful	carrots	strawberry	<u>roll</u>
	<u>said</u>	president	evidence	<u>germany</u>	<u>year</u>	<u>place</u>	<u>sunday</u>	greens	<u>vanilla</u>	salmon
	games	white	findings	<u>workers</u>	matsui	phenomenal	<u>like</u>	salad	banana	<u>fresh</u>
	team	house	defense	paris	<u>said</u>	pleasant	slow	<u>dressing</u>	<u>peanut</u>	<u>good</u>
CorEx	baseball	house	possibility	french	japanese	great	<u>even</u>	garlic	strawberry	shrimp
	championship	white	challenge	<u>italy</u>	tokyo	friendly	bad	tomato	<u>caramel</u>	<u>beef</u>
	playing	support	reasons	<u>paris</u>	<u>index</u>	<u>atmosphere</u>	mean	onions	<u>sugar</u>	crab
	<u>fans</u>	<u>groups</u>	<u>give</u>	francs	osaka	love	cold	<u>toppings</u>	fruit	<u>dishes</u>
	league	<u>member</u>	planned	jacques	<u>electronics</u>	favorite	<u>literally</u>	<u>slices</u>	mango	<u>salt</u>
ETM	olympic	government	approach	french	japanese	nice	disappointed	avocado	strawberry	fish
	league	national	problems	<u>students</u>	<u>agreement</u>	worth	cold	<u>greek</u>	mango	shrimp
	<u>national</u>	<u>plan</u>	experts	paris	tokyo	<u>lunch</u>	<u>review</u>	salads	<u>sweet</u>	lobster
	basketball	public	<u>move</u>	<u>german</u>	<u>market</u>	recommend	<u>experience</u>	spinach	<u>soft</u>	crab
	athletes	support	<u>give</u>	<u>american</u>	<u>europen</u>	friendly	bad	tomatoes	<u>flavors</u>	<u>chips</u>
BERTopic	swimming	bush	researchers	french	japanese	awesome	horrible	tomatoes	strawberry	lobster
	freestyle	democrats	scientists	paris	tokyo	<u>atmosphere</u>	<u>quality</u>	avocado	mango	crab
	<u>popov</u>	white	cases	lyon	ufj	friendly	disgusting	<u>soups</u>	<u>cup</u>	shrimp
	gold	bushs	<u>genetic</u>	<u>minister</u>	<u>company</u>	<u>night</u>	disappointing	kale	lemon	oysters
	olympic	house	study	<u>billion</u>	yen	good	<u>place</u>	cauliflower	banana	<u>amazing</u>
TopClus	athletes	government	hypothesis	french	japanese	good	tough	potatoes	strawberry	fish
	medalist	ministry	methodology	seine	tokyo	best	bad	onions	lemon	octopus
	olympics	bureaucracy	possibility	toulouse	osaka	friendly	painful	tomatoes	apples	shrimp
	tournaments	politicians	criteria	marseille	hokkaido	cozy	frustrating	cabbage	grape	lobster
	quarterfinal	electoral	assumptions	paris	yokohama	casual	brutal	mushrooms	peach	crab

# Experiments

## □ Visualization



**Figure 5: Visualization using t-SNE of 10,000 randomly sampled latent embeddings during the course of TopClus training. Embeddings assigned to the same cluster are denoted with the same color. The latent space gradually exhibits distinctive and balanced cluster structure.**

# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-based Topic Discovery
- ❑ Discriminative Topic Mining
  - ❑ Introduction of the Task 
  - ❑ CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - ❑ JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
  - ❑ SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]

# Limitations of Unsupervised Topic Discovery

- **Cannot incorporate user guidance:** Topic models tend to retrieve the most general and prominent topics from a text collection
  - may not be of a user's particular interest
  - provide a skewed and biased summarization of the corpus
- **Cannot enforce distinctiveness among retrieved topics:** Topic models do not impose discriminative constraints
  - E.g., three retrieved topics from the New York Times annotated corpus via LDA

**Table 1: LDA retrieved topics on NYT dataset. The meanings of the retrieved topics have overlap with each other.**

Topic 1	Topic 2	Topic 3
canada, united states canadian, economy	sports, united states olympic, games	united states, iraq government, president



Difficult to clearly define the meaning of the three topics due to an overlap of their semantics (e.g., the term “united states” appears in all 3 topics)

# Seed-Guided, Discriminative Topic Mining

---

- ❑ **Discriminative Topic Mining:** Given a text corpus and a set of **category names**, retrieve a set of terms that **exclusively belong to** each category
  - ❑ E.g., given  $c_1$ : “The United States”,  $c_2$ : “France”,  $c_3$ : “Canada”
    - ❑ Yes to “Ontario” under  $c_3$ : (a province in Canada and exclusively belongs to Canada)
    - ❑ No to “North America” under  $c_3$ : (a continent and does not belong to any countries (**reversed belonging relationship**))
    - ❑ No to “English” under  $c_3$ : (English is also the national language of the United States (**not discriminative**))
  - ❑ Difference from topic modeling
    - ❑ requires **a set of user provided category names** and only focuses on retrieving terms belonging to the given categories
    - ❑ imposes strong discriminative requirements that each retrieved term under the corresponding category must **belong to and only belong to** that category semantically

# Outline

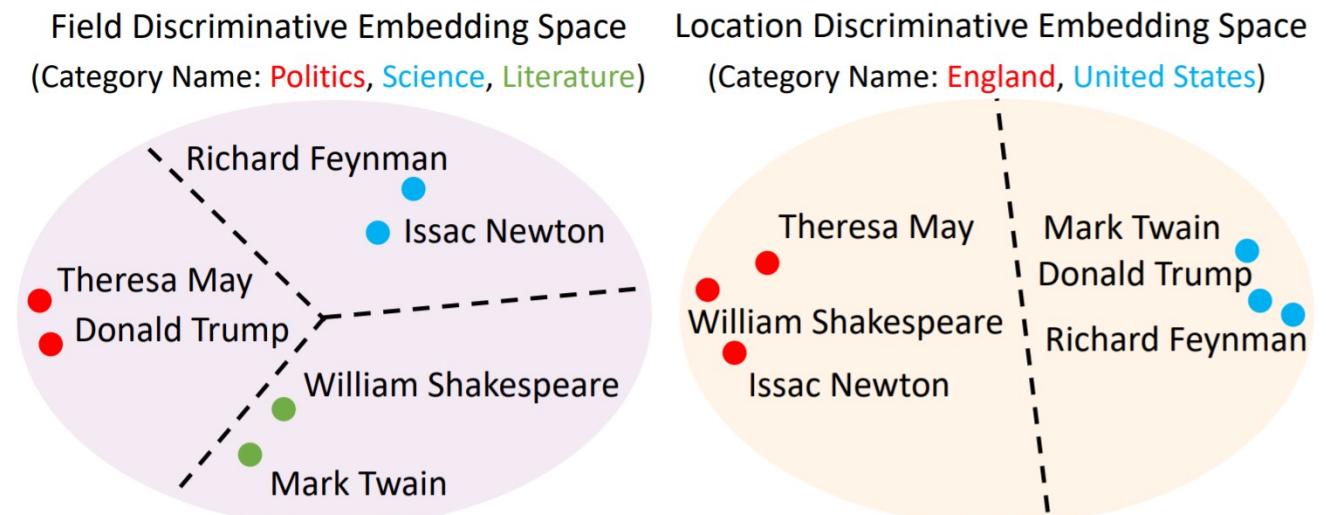
---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-based Topic Discovery
- ❑ Discriminative Topic Mining
  - ❑ Introduction of the Task
  - ❑ CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - ❑ JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
  - ❑ SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]



# Discriminative Topic Mining via CatE

- ❑ Word embeddings capture word semantic correlations via the distributional hypothesis
  - ❑ captures local context similarity
  - ❑ not exploit document-level statistics (global context)
  - ❑ not model topics
- ❑ **CatE: Category Name-guided Embedding:** leverages *category names* to learn word embeddings with discriminative power over the specific set of categories
- ❑ CatE: Inputs
  - ❑ Category names + Corpus
- ❑ CatE: Outputs (see figure)
  - ❑ The same set of celebrities are embedded differently given different sets of category names



# CatE Embedding: Text Generation Modeling

---

- Modeling text generation under user guidance
- A three-step process:
  1. A document  $d$  is generated conditioned on one of the  $n$  categories      [1. Topic assignment](#)
  2. Each word  $w_i$  is generated conditioned on the semantics of the document  $d$       [2. Global context](#)
  3. Surrounding words  $w_{i+j}$  in the local context window of  $w_i$  are generated conditioned on the semantics of the center word  $w_i$       [3. Local context](#)
- Compute the likelihood of corpus generation conditioned on user-given categories

# CatE Embedding: Objective

## □ Objective: negative log-likelihood

$$P(\mathcal{D} | C) = \prod_{d \in \mathcal{D}} p(d | c_d) \prod_{w_i \in d} p(w_i | d) \prod_{\substack{w_{i+j} \in d \\ -h \leq j \leq h, j \neq 0}} p(w_{i+j} | w_i)$$

1. Topic assignment    2. Global context    3. Local context

$p(d | c_d) \propto p(c_d | d)p(d) \propto p(c_d | d) \propto \prod_{w \in d} p(c_d | w),$  Decompose into word-topic distribution

## □ Introducing specificity

**Definition 2** (Word Distributional Specificity). We assume there is a scalar  $\kappa_w \geq 0$  correlated with each word  $w$  indicating how specific the word meaning is. The bigger  $\kappa_w$  is, the more specific meaning word  $w$  has, and the less varying contexts  $w$  appears in.

- E.g., “seafood” has a higher word distributional specificity than “food”, because seafood is a specific type of food

# Category Representative Word Retrieval

- Ranking Measure for Selecting Class Representative Words:
- We find a representative word of category  $c_i$  and add it to the set  $S$  by

Prefer words having high embedding cosine similarity with the category name

Prefer words with low distributional specificity (more general)

$$w = \arg \min_w \text{rank}_{sim}(w, c_i) \cdot \text{rank}_{spec}(w)$$

$$\text{s.t. } w \notin S \quad \text{and} \quad \kappa_w > \kappa_{c_i}.$$

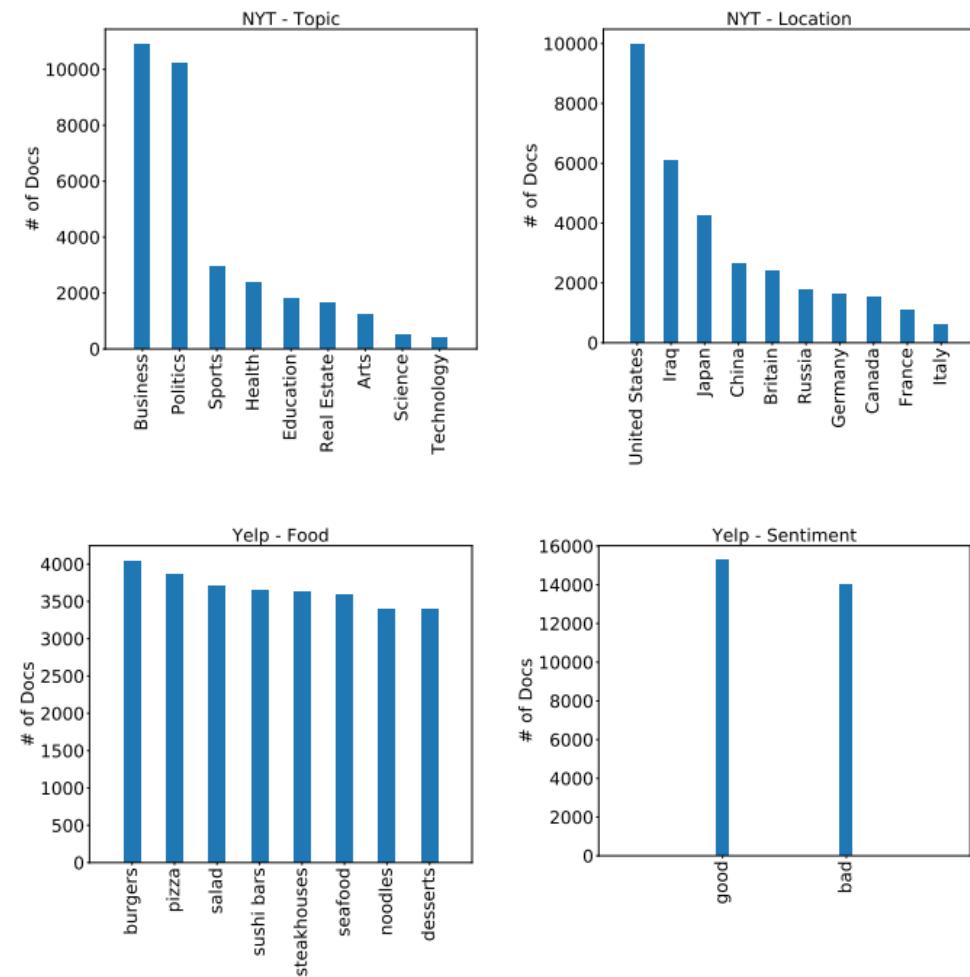
w hasn't been a representative word

w must be more specific than the category name

# Quantitative Results

- Two datasets:
  - New York Times annotated corpus (NYT)
    - Two categories: topic and location
  - Recently released Yelp Dataset Challenge (Yelp)
    - Two categories: food type and sentiment

Methods	NYT-Location		NYT-Topic		Yelp-Food		Yelp-Sentiment	
	TC	MACC	TC	MACC	TC	MACC	TC	MACC
LDA	0.007	0.489	0.027	0.744	-0.033	0.213	-0.197	0.350
Seeded LDA	0.024	0.168	0.031	0.456	0.016	0.188	0.049	0.223
TWE	0.002	0.171	-0.011	0.289	0.004	0.688	-0.077	0.748
Anchored CorEx	0.029	0.190	0.035	0.533	0.025	0.313	0.067	0.250
Labeled ETM	0.032	0.493	0.025	0.889	0.012	0.775	0.026	0.852
CatE	<b>0.049</b>	<b>0.972</b>	<b>0.048</b>	<b>0.967</b>	<b>0.034</b>	<b>0.913</b>	<b>0.086</b>	<b>1.000</b>



Dataset stat: # of docs by category name

# Qualitative Results

Methods	NYT-Location		NYT-Topic		Yelp-Food		Yelp-Sentiment	
	britain	canada	education	politics	burger	desserts	good	bad
LDA	company (x)	percent (x)	school	campaign	fatburger	ice cream	great	valet (x)
	companies (x)	economy (x)	students	clinton	dos (x)	chocolate	place (x)	peter (x)
	british	canadian	city (x)	mayor	liar (x)	gelato	love	aid (x)
	shares (x)	united states (x)	state (x)	election	cheeseburgers	tea (x)	friendly	relief (x)
	great britain	trade (x)	schools	political	bearing (x)	sweet	breakfast	rowdy
Seeded LDA	british	city (x)	state (x)	republican	like (x)	great (x)	place (x)	service (x)
	industry (x)	building (x)	school	political	fries	like (x)	great	did (x)
	deal (x)	street (x)	students	senator	just (x)	ice cream	service (x)	order (x)
	billion (x)	buildings (x)	city (x)	president	great (x)	delicious (x)	just (x)	time (x)
	business (x)	york (x)	board (x)	democrats	time (x)	just (x)	ordered (x)	ordered (x)
TWE	germany (x)	toronto	arts (x)	religion	burgers	chocolate	tasty	subpar
	spain (x)	osaka (x)	fourth graders	race	fries	complimentary (x)	decent	positive (x)
	manufacturing (x)	booming (x)	musicians (x)	attraction (x)	hamburger	green tea (x)	darned (x)	awful
	south korea (x)	asia (x)	advisors	era (x)	cheeseburger	sundae	great	crappy
	markets (x)	alberta	regents	tale (x)	patty	whipped cream	suffered (x)	honest (x)
Anchored CorEx	moscow (x)	sports (x)	republican (x)	military (x)	order (x)	make (x)	selection (x)	did (x)
	british	games (x)	senator (x)	war (x)	know (x)	chocolate	prices (x)	just (x)
	london	players (x)	democratic (x)	troops (x)	called (x)	people (x)	great	came (x)
	german (x)	canadian	school	baghdad (x)	fries	right (x)	reasonable	asked (x)
	russian (x)	coach	schools	iraq (x)	going (x)	want (x)	mac (x)	table (x)
Labeled ETM	france (x)	canadian	higher education	political	hamburger	pana	decent	horrible
	germany (x)	british columbia	educational	expediency (x)	cheeseburger	gelato	great	terrible
	canada (x)	britain (x)	school	perceptions (x)	burgers	tiramisu	tasty	good (x)
	british	quebec	schools	foreign affairs	patty	cheesecake	bad (x)	awful
	europe (x)	north america (x)	regents	ideology	steak (x)	ice cream	delicious	appallingly
CatE	england	ontario	educational	political	burgers	dessert	delicious	sickening
	london	toronto	schools	international politics	cheeseburger	pastries	mindful	nasty
	britons	quebec	higher education	liberalism	hamburger	cheesecakes	excellent	dreadful
	scottish	montreal	secondary education	political philosophy	burger king	scones	wonderful	freaks
	great britain	ottawa	teachers	geopolitics	smash burger	ice cream	faithful	cheapskates

# Case Study: Effect of Distributional Specificity

- Coarse-to-fine topic presentation on NYT-Topic

Range of $\kappa$	Science ( $\kappa_c = 0.539$ )	Technology ( $\kappa_c = 0.566$ )	Health ( $\kappa_c = 0.527$ )
$\kappa_c < \kappa < 1.25\kappa_c$	scientist, academic, research, laboratory	machine, equipment, devices, engineering	medical, hospitals, patients, treatment
$1.25\kappa_c < \kappa < 1.5\kappa_c$	physics, sociology, biology, astronomy	information technology, computing, telecommunication, biotechnology	mental hygiene, infectious diseases, hospitalizations, immunizations
$1.5\kappa_c < \kappa < 1.75\kappa_c$	microbiology, anthropology, physiology, cosmology	wireless technology, nanotechnology, semiconductor industry, microelectronics	dental care, chronic illnesses, cardiovascular disease, diabetes
$\kappa > 1.75\kappa_c$	national science foundation, george washington university, hong kong university, american academy	integrated circuits, assemblers, circuit board, advanced micro devices	juvenile diabetes, high blood pressure, family violence, kidney failure

- The table lists the most similar words/phrases with each category (measured by embedding cosine similarity) from different ranges of distributional specificity
- When  $\kappa$  is smaller, the retrieved words have wider semantic coverage
- In our model design, if not imposing constraints on the  $\kappa$ , the retrieved words might be too general and do not belong to the category

# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-based Topic Discovery
- ❑ Discriminative Topic Mining
  - ❑ Introduction of the Task
  - ❑ CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - ❑ JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
  - ❑ SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23]



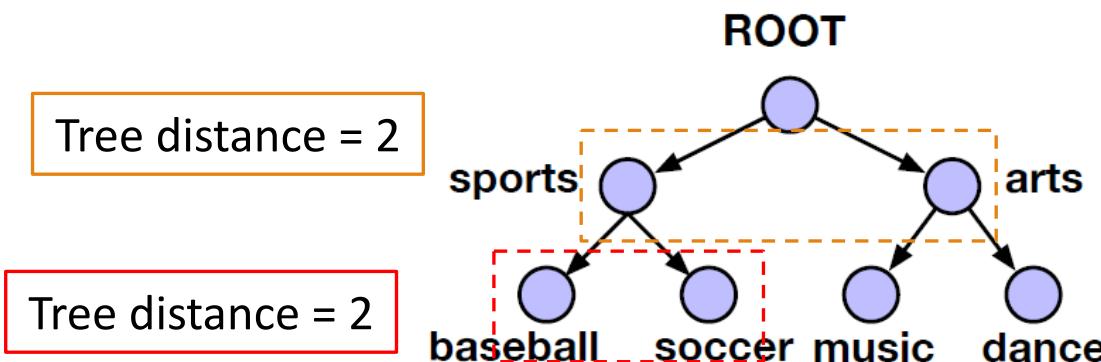
# Motivation: Hierarchical Topic Mining

---

- Mining a set of meaningful topics organized into a **hierarchy** is intuitively appealing and has broad applications
  - Coarse-to-fine topic understanding
  - Hierarchical corpus summarization
  - Hierarchical text classification
  - ...
- Hierarchical topic models discover topic structures from text corpora via modeling the text generative process with a latent hierarchy

# JoSH Embedding

- Difference from hyperbolic models (e.g., Poincare, Lorentz)
  - Hyperbolic embeddings preserve absolute tree distance (similar embedding distance => similar tree distance)
  - We do not aim to preserve the absolute tree distance, but rather use it as a relative measure



Although  $d_{\text{tree}}(\text{sports}, \text{arts}) = d_{\text{tree}}(\text{baseball}, \text{soccer})$ , “baseball” and “soccer” should be embedded closer than “sports” and “arts” to reflect semantic similarity.

Use tree distance in a relative manner: Since  $d_{\text{tree}}(\text{sports}, \text{baseball}) < d_{\text{tree}}(\text{baseball}, \text{soccer})$ , “baseball” and “sports” should be embedded closer than “baseball” and “soccer”.

# JoSH Text Embedding

---

- Modeling Text Generation Conditioned on the Category Tree (Similar to CatE)
- A three-step process:

1. A document  $d_i$  is generated conditioned on one of the  $n$  categories

1. Topic assignment

$$p(d_i | c_i) = \text{vMF}(d_i; c_i, \kappa_{c_i}) = n_p(\kappa_{c_i}) \exp(\kappa_{c_i} \cdot \cos(d_i, c_i))$$

2. Each word  $w_j$  is generated conditioned on the semantics of the document  $d_i$

2. Global context

$$p(w_j | d_i) \propto \exp(\cos(u_{w_j}, d_i))$$

3. Surrounding words  $w_{j+k}$  in the local context window of  $w_i$  are generated conditioned on the semantics of the center word  $w_i$

3. Local context

$$p(w_{j+k} | w_j) \propto \exp(\cos(v_{w_{j+k}}, u_{w_j}))$$

# JoSH Tree Embedding

- **Intra-Category Coherence:** Representative terms of each category should be highly semantically relevant to each other, reflected by high directional similarity in the spherical space

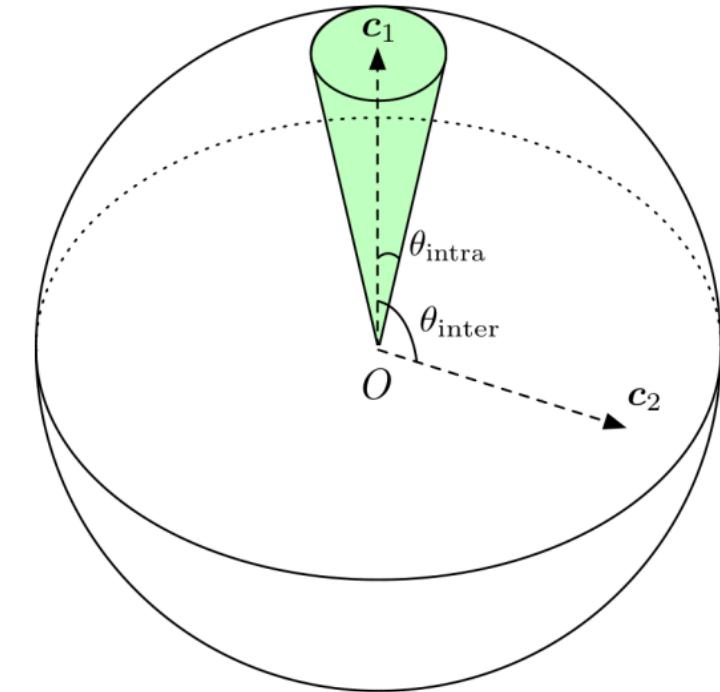
$$\mathcal{L}_{\text{intra}} = \sum_{c_i \in \mathcal{T}} \sum_{w_j \in C_i} \min(0, \mathbf{u}_{w_j}^\top c_i - m_{\text{intra}}),$$

- **Inter-Category Distinctiveness:** Encourage distinctiveness across different categories to avoid semantic overlaps so that the retrieved terms provide a clear and distinctive description

$$\mathcal{L}_{\text{inter}} = \sum_{c_i \in \mathcal{T}} \sum_{c_j \in \mathcal{T} \setminus \{c_i\}} \min(0, 1 - c_i^\top c_j - m_{\text{inter}}).$$

$$\theta_{\text{intra}} \leq \arccos(m_{\text{intra}})$$

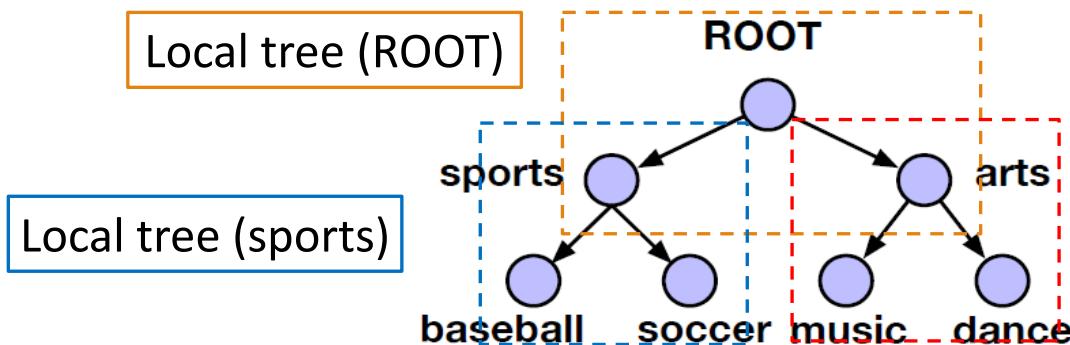
$$\theta_{\text{inter}} \geq \arccos(1 - m_{\text{inter}})$$



(a) Intra- & Inter-Category Configuration.

# JoSH Tree Embedding

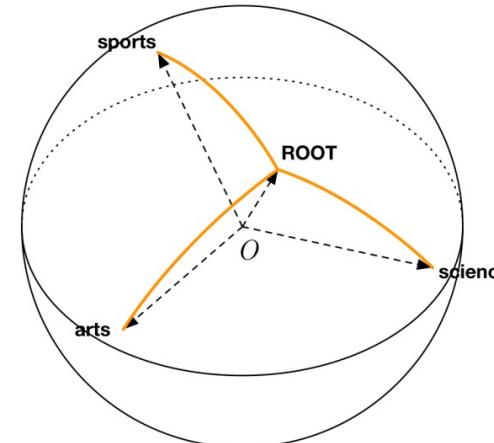
- **Recursive Local Tree Embedding:** Recursively embed local structures of the category tree onto the sphere



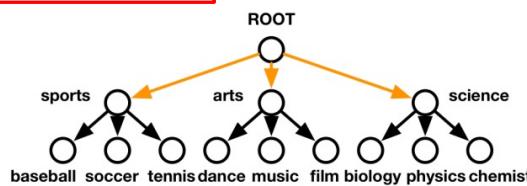
Local tree: A local tree  $T_r$  rooted at node  $c_r \in T$  consists of node  $c_r$  and all of its direct children

- **Preserving Relative Tree Distance within Local Trees:** A category should be closer to its parent category than to its sibling categories in the embedding space

$$\mathcal{L}_{\text{inter}} = \sum_{c_i \in \mathcal{T}_r} \sum_{c_j \in \mathcal{T}_r \setminus \{c_r, c_i\}} \min(0, c_i^\top c_r - c_i^\top c_j - m_{\text{inter}}),$$



(b) Embed First-Level Local Tree.



(c) Embed Second-Level Local Trees.

# Experiments: Qualitative Results on NYT

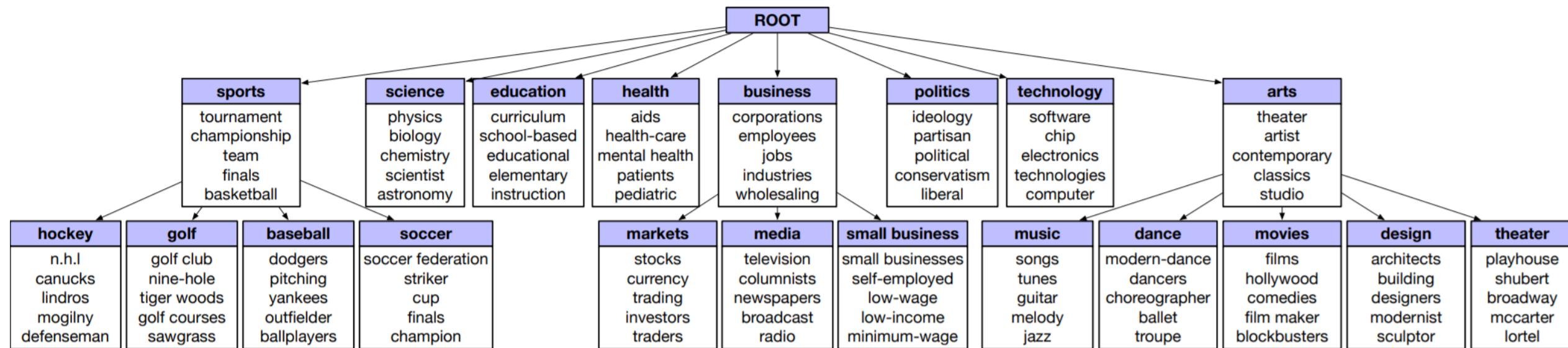
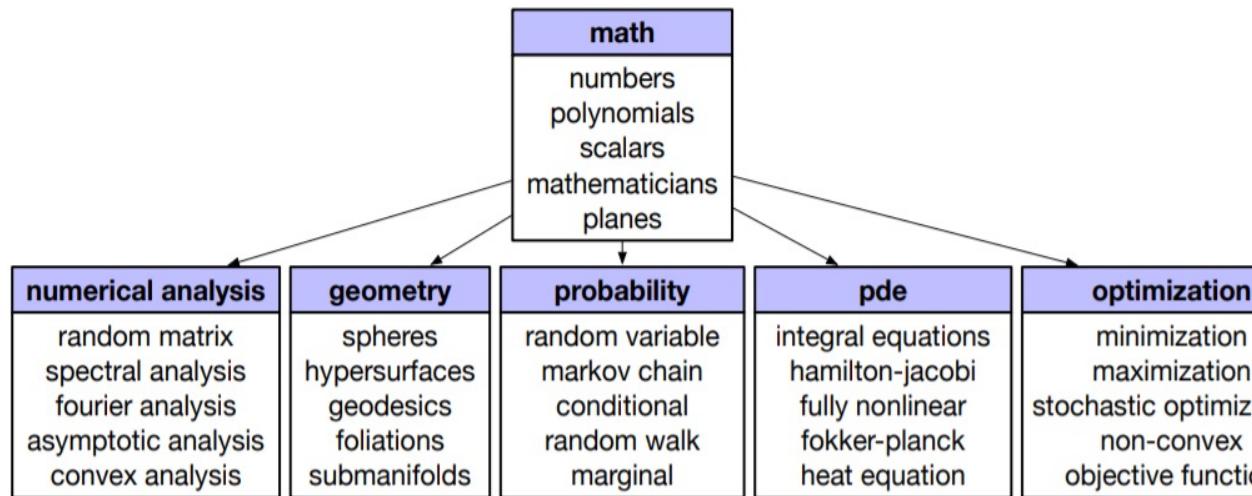
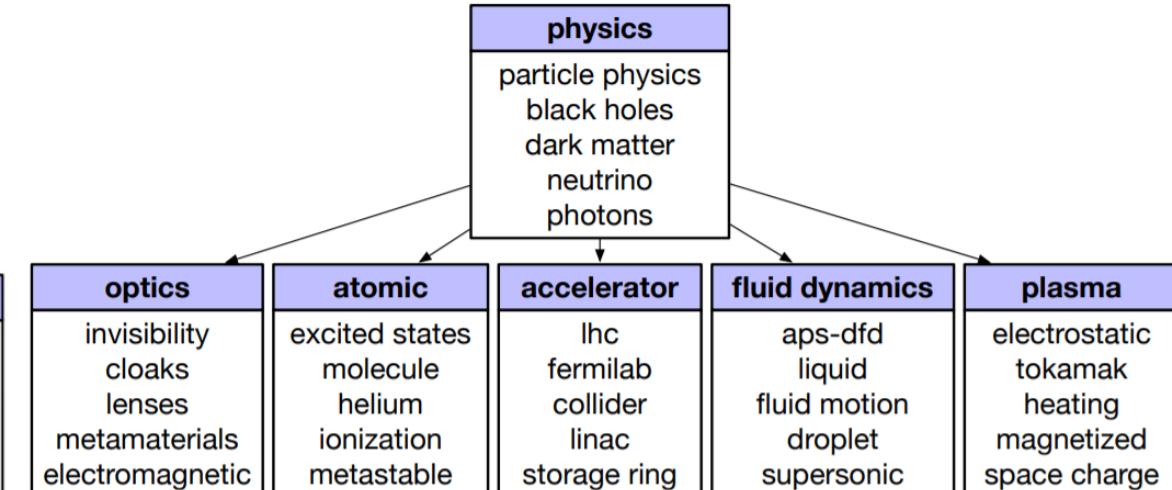


Figure 3: Hierarchical Topic Mining results on NYT.

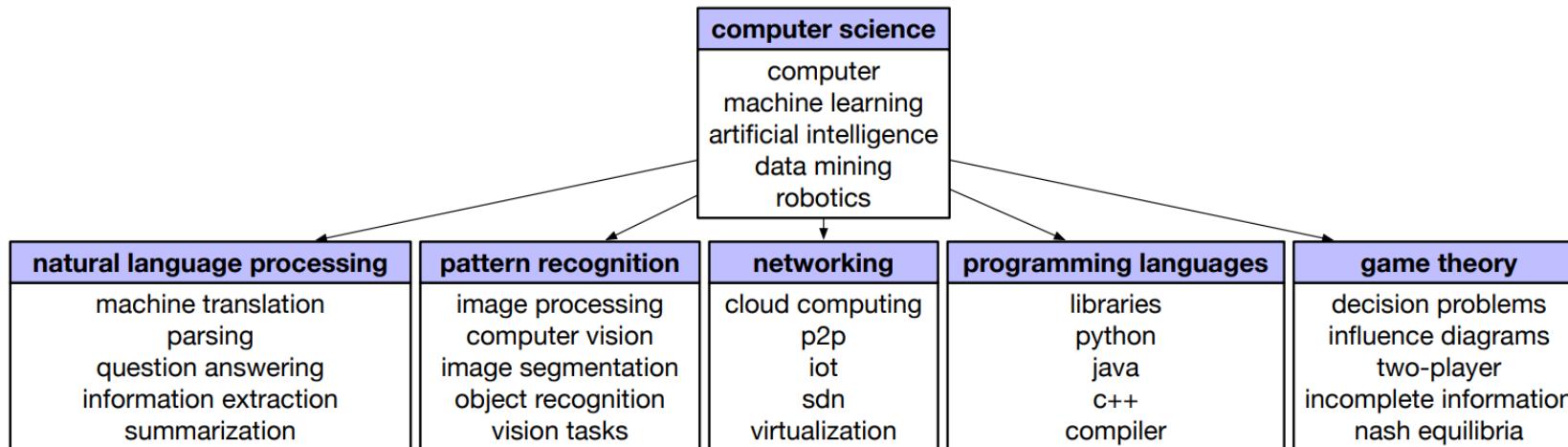
# Experiments: Qualitative Results on ArXiv and Quantitative Results



(a) “Math” subtree.



(b) “Physics” subtree.



(c) “Computer Science” subtree.

Models	NYT		arXiv	
	TC	MACC	TC	MACC
hLDA	-0.0070	0.1636	-0.0124	0.1471
hPAM	0.0074	0.3091	0.0037	0.1824
JoSE	0.0140	0.6818	0.0051	0.7412
Poincaré GloVe	0.0092	0.6182	-0.0050	0.5588
Anchored CorEx	0.0117	0.3909	0.0060	0.4941
CatE	0.0149	0.9000	0.0066	0.8176
JoSH	<b>0.0166</b>	<b>0.9091</b>	<b>0.0074</b>	<b>0.8324</b>

# Outline

---

- ❑ Unsupervised Topic Modeling
- ❑ Supervised & Seed-Guided Topic Modeling
- ❑ Clustering-based Topic Discovery
- ❑ Discriminative Topic Mining
  - ❑ Introduction of the Task
  - ❑ CatE: Discriminative Topic Mining via Category-Name Guided Text Embedding [WWW'20]
  - ❑ JoSH: Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]
  - ❑ SeedTopicMine: Integrating Multiple Types of Contexts [WSDM'23] 

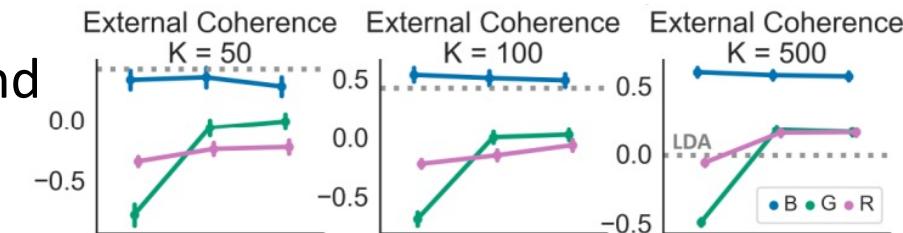
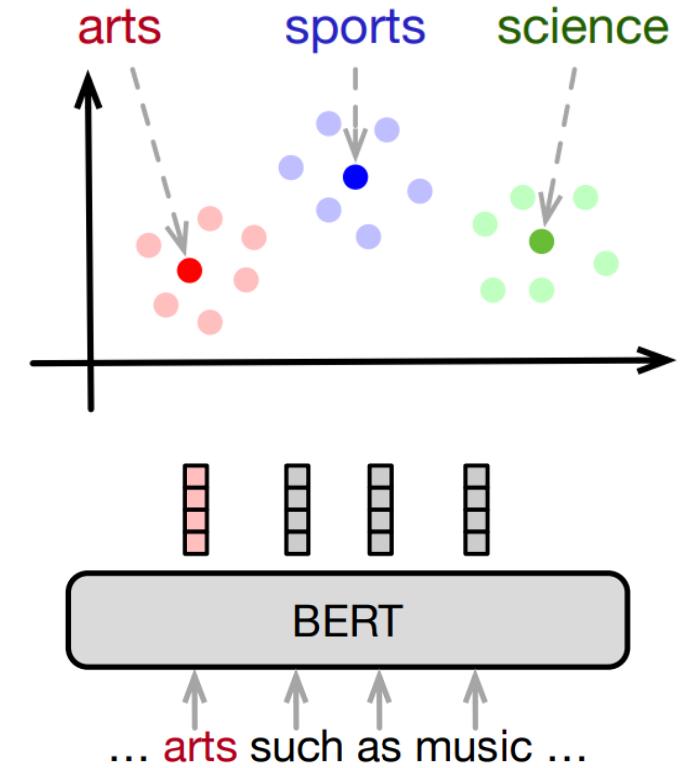
# Commonly Used Context Information

## ❑ Context Type I - Skip-Gram Embeddings

- ❑ Previous slides have shown that clustering skip-gram embeddings underperforms clustering output representations of contextualized language models such as BERT in unsupervised topic modeling.

## ❑ Context Type II - Pre-trained Language Model Representations

- ❑ Previous slides have shown that BERT representations suffer from the curse of dimensionality and may not form clearly separated clusters
- ❑ Thompson and Mimno [1] find that GPT-2 representations work well only if the outputs of certain layers are taken, and RoBERTa-induced topics are consistently of poor quality.

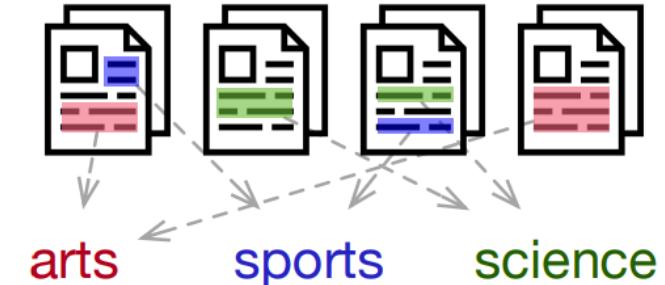


[1] Thompson, L., and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv.

# Commonly Used Context Information

## ❑ Context Type III - Topic-Indicative Documents

- ❑ Supervised topic models [1] propose to leverage document-level training data. However, such information relies on **massive human annotation**, which is not available under the seed-guided setting.
- ❑ A document may be **too broad** to be viewed as a context unit because each document can be relevant to multiple topics simultaneously.

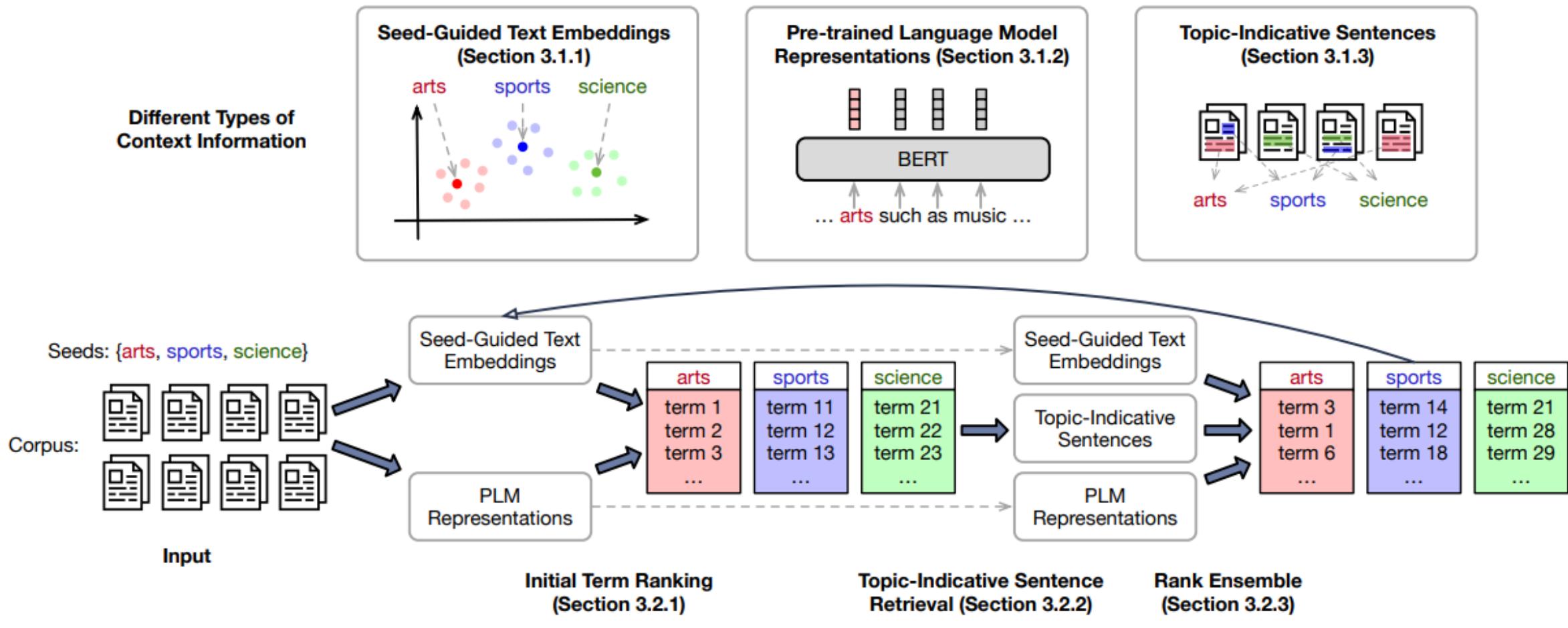


## ❑ Each type of context signals has its specific advantages and disadvantages.

- ❑ A topic discovery method purely relying on one type of context information may not be robust across different datasets or seed dimensions.
- ❑ Meanwhile, the three types of contexts strongly **complement each other**.

[1] Blei, D., and McAuliffe, J. (2007). Supervised topic models. NIPS.

# SeedTopicMine: Overview

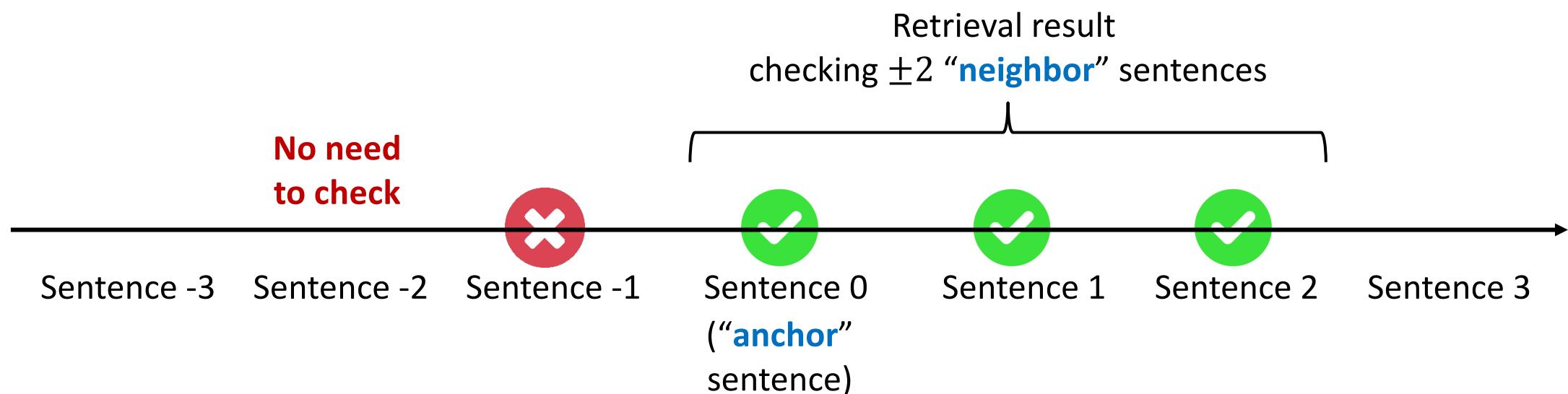


**Figure 1: Overview of the SEEDTOPICMINE framework.**

Zhang, Y., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.

# SeedTopicMine: Topic-Indicative Sentence Retrieval

- The sentences containing many topic-indicative terms from one category and do not contain any topic-indicative term from other categories should be topic-indicative sentences. We call such sentences “**anchor**” sentences.
- The “**neighbor**” sentences of topic-indicative “anchor” sentences should be included in topic-indicative sentences as well if they do not contain topic-indicative terms from other categories.



# Quantitative Results

**Table 2: NPMI, P@20, and NDCG@20 scores of compared algorithms. NPMI measures topic coherence; P@20 and NDCG@20 measure term accuracy.**

Method	NYT-Topic			NYT-Location			Yelp-Food			Yelp-Sentiment		
	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20	NPMI	P@20	NDCG@20
SeededLDA [15]	0.0841	0.2389	0.2979	0.0814	0.1050	0.1873	0.0504	0.1200	0.2132	0.0499	0.1700	0.2410
Anchored CorEx [10]	0.1325	0.2922	0.3627	0.1283	0.2040	0.3003	0.1204	0.3725	0.4531	0.0627	0.1200	0.1997
KeyETM [13]	0.1254	0.1589	0.2342	0.1146	0.0700	0.1676	0.0578	0.1788	0.2940	0.0327	0.4250	0.4994
CatE [27]	0.1941	0.8067	0.8306	0.2165	0.7480	0.7840	<b>0.2058</b>	0.6812	0.7312	<b>0.1509</b>	0.7150	0.7713
SEEDTOPICMINE	<b>0.1947</b>	<b>0.9456</b>	<b>0.9573</b>	<b>0.2176</b>	<b>0.8360</b>	<b>0.8709</b>	0.2018	<b>0.7912</b>	<b>0.8379</b>	0.0922	<b>0.9750</b>	<b>0.9811</b>

Method	Yelp-Food		Yelp-Sentiment	
	P@20	NDCG@20	P@20	NDCG@20
SEEDTOPICMINE	<b>0.7912</b>	<b>0.8379</b>	<b>0.9750</b>	<b>0.9811</b>
SEEDTOPICMINE-NoEmb	0.4488	0.5335	0.9550	0.9646
SEEDTOPICMINE-NoPLM	0.6962	0.7602	0.7550	0.8029
SEEDTOPICMINE-NoSntn	0.7488	0.8029	0.9500	0.9631

- Three types of contexts all have positive contribution.
- Even for the same dataset (i.e., Yelp), the contribution of a certain type of context information varies significantly with the input seeds. Therefore, it becomes necessary to **integrate them together** to make the framework more robust.

# Qualitative Results

Table 3: Top-5 terms retrieved by different algorithms. ×: At least 3 of the 5 annotators judge the term as irrelevant to the seed.

Method	NYT-Topic		NYT-Location		Yelp-Food		Yelp-Sentiment	
	health	business	france	canada	sushi	desserts	good	bad
SeededLDA	said (x)	said (x)	said (x)	new (x)	roll	food (x)	place (x)	food (x)
	dr (x)	percent (x)	new (x)	city (x)	good (x)	us (x)	food (x)	service (x)
	new (x)	company	state (x)	said (x)	place (x)	order (x)	great	us (x)
	would (x)	year (x)	would (x)	building (x)	food (x)	service (x)	like (x)	order (x)
	hospital	billion (x)	dr (x)	mr (x)	rolls	time (x)	service (x)	time (x)
Anchored CorEx	case (x)	employees	school (x)	market (x)	rolls	also (x)	definitely (x)	one (x)
	court (x)	advertising	students (x)	percent (x)	roll	really (x)	prices (x)	would (x)
	patients	media (x)	children (x)	companies (x)	sashimi	well (x)	strip (x)	like (x)
	cases (x)	businessmen	education (x)	billion (x)	fish (x)	good (x)	selection (x)	could (x)
	lawyer (x)	commerce	schools (x)	investors (x)	tempura	try (x)	value (x)	us (x)
KeyETM	team (x)	percent (x)	city (x)	people (x)	sashimi	food (x)	great	food (x)
	game (x)	japan (x)	state (x)	year (x)	rolls	great (x)	delicious	place (x)
	players (x)	year (x)	york (x)	china (x)	roll	place (x)	amazing	service (x)
	games (x)	japanese (x)	school (x)	years (x)	fish (x)	good (x)	excellent	time (x)
	play (x)	economy	program (x)	time (x)	japanese	service (x)	tasty	restaurant (x)
CatE	public health	diversifying (x)	french	alberta	freshest fish (x)	delicacies (x)	tasty	unforgivable
	health care	clients (x)	corsica	british columbia	sashimi	sundaes	delicious	frustrating
	medical	corporate	spain (x)	ontario	nigiri	savoury (x)	yummy	horrible
	hospitals	investment banking	belgium (x)	manitoba	ayce sushi	pastries	chilaquiles (x)	irritating
	doctors	executives	de (x)	canadian	rolls	custards	also (x)	rude
SEEDTOPICMINE	medical	companies	french	canadian	maki rolls	cheesecakes	great	terrible
	hospitals	businesses	paris	quebec	sashimi	croissants	excellent	horrible
	hospital	corporations	philippe (x)	montreal	ayce sushi	pastries	fantastic	awful
	public health	firms	french state	toronto	revolving sushi	bread (x)	delicious	lousy
	patients	corporate	frenchman	ottawa	nigiri	cheesecake	amazing	shitty

# References

---

- Blei, D. M., Griffiths, T. L., Jordan, M. I., & Tenenbaum, J. B. (2003). Hierarchical topic models and the nested Chinese restaurant process. NIPS.
- Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. NIPS.
- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research.
- Mimno, D., Li, W., & McCallum, A. (2007). Mixtures of hierarchical topics with pachinko allocation. ICML.
- Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. EACL.
- Meng, Y., Huang, J., Wang, G., Wang, Z., Zhang, C., Zhang, Y., & Han, J. (2020). Discriminative topic mining via category-name guided text embedding. WWW.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. (2020). Hierarchical topic mining via joint spherical tree and text embedding. KDD.
- Meng, Y., Zhang, Y., Huang, J., Zhang, Y., & Han, J. (2022). Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations. WWW.
- Sia, S., Dalmia, A., & Mielke, S. J. (2020). Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too! EMNLP.
- Zhang, Y., Zhang, Y., Michalski, M., Jiang, Y., Meng, Y., & Han, J. (2023). Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts. WSDM.