

# Part II: Text Representation Enhanced Topic Discovery

KDD 2023 Tutorial Pretrained Language Representations for Text Understanding: A Weakly-Supervised Perspective Yu Meng, Jiaxin Huang, Yu Zhang, Yunyi Zhang, Jiawei Han Computer Science, University of Illinois Urbana-Champaign Aug 9, 2023

### Outline

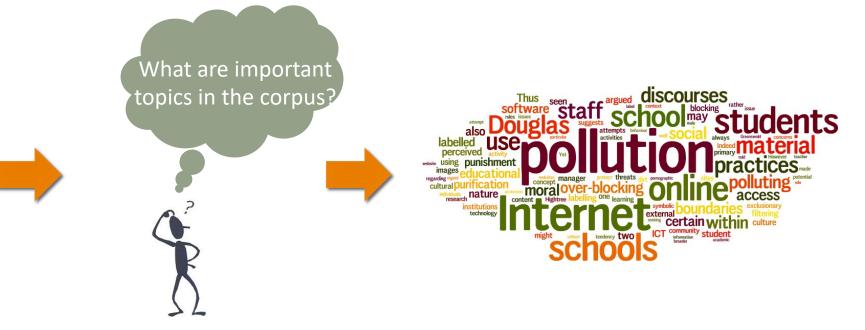
### Traditional Topic Models

- Embedding-Based Discriminative Topic Mining
- **D** Topic Discovery with PLMs

# **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





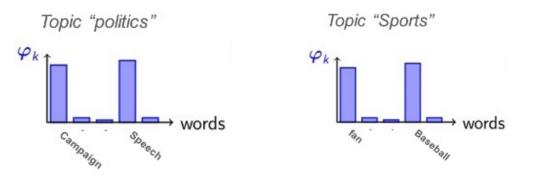
### **Topic Modeling: Overview**

- 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

Topics	Documents	Topic Assignn	nent	и	/ord-Topi	с Ма	atrix	
gene 0.04 DNA 0.02	Seeking Life's Bare (Genetic) Necess				Topic 1		Topic n	
genetic 0.01	COLD SPRING HARBOR, NEW Gence YORK- How many genes does an Cold organism need to survive? Last week at to	Spring Harbor, New York, May 8		Word 1	0.09		0.06	
	the genome meeting here, "two genome apart researchers with radically different 75,00	t," especially in comparison to the	h 1					
110 0.00	approaches presented complementary notes views of the basic genes needed for life. Univ	Siv Andersson of Uppsala	//	Word n	0.08		0.01	
life 0.02 evolve 0.01 organism 0.01	One research team, using computer the 8 analyses to compare known genomes, conse	300 number. But coming up with a						
	concluded that today's organisms ← just be sustained with just 250 genes, and partic that the earliest life forms required a are c	cularly as more and more genomes		Document-Topic Matrix				
	mere 128 genes. The other researcher "It n mapped genes in a simple parasite and newl	nay be a way of organizing any	\		Topic 1		Topic n	
data 0.02 number 0.02	estimated that for this organism, 800 Arca genes are plenty to do the job-but that mole			Doc 1	0.23		0.33	
computer 0.01		BI) in Bethesda, Maryland.						
	Although the numbers don't Com match precisely, those predictions	paring an		Doc n	0.15		0.28	
	Ч							

# Latent Dirichlet Allocation (LDA): Overview

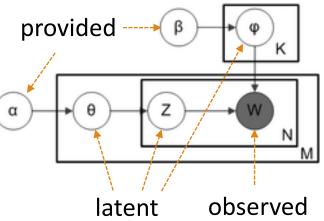
- Each document is represented as a mixture of various topics
  - Ex. A news document may be 40% on politics, 50% on economics, and 10% on sports
- Each topic is represented as a probability distribution over words
  - Ex. The distribution of "politics" vs. "sports" might be like:



- 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
  - $\Box$  Document-topic distribution  $\theta$  (for assigning topics to documents)
  - $\Box$  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



### **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

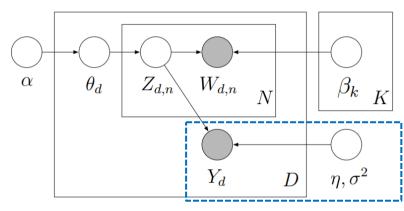
them?

	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	mambaav		1-	0.004074						
	member	0.040683	stock	0.024874	add	0.010185	system	0.007233	area	0.014865
2	meeting	0.040683	stock	0.024874	add minute	0.010185	system problem	0.007233	area build	0.014865
2				0			-			
	meeting	0.016390	share	0.020583	minute	0.009301	problem	0.006835	build	0.014361
3	meeting issue	0.016390	share price	0.020583	minute pepper	0.009301	problem power	0.006835	build building	0.014361 0.014326
3	meeting issue official	0.016390 0.014988 0.013069	share price sell	0.020583 0.018141 0.016564	minute pepper oil	0.009301 0.009235 0.008976	problem power create	0.006835 0.005400 0.005056	build building home	0.014361 0.014326 0.013632
3 4 5	meeting issue official support	0.016390 0.014988 0.013069 0.011994	share price sell buy	0.020583 0.018141 0.016564 0.015415	minute pepper oil cook	0.009301 0.009235 0.008976 0.008711	problem power create research	0.006835 0.005400 0.005056 0.004712	build building home resident	0.014361 0.014326 0.013632 0.013483
3 4 5	meeting issue official support leader	0.016390 0.014988 0.013069 0.011994 0.011799	share price sell buy company	0.020583 0.018141 0.016564 0.015415 0.015249	minute pepper oil cook food	0.009301 0.009235 0.008976 0.008711 0.008689	problem power create research produce	0.006835 0.005400 0.005056 0.004712 0.004574	build building home resident community	0.014361 0.014326 0.013632 0.013483 0.012479

#### 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,
    - $\Box$  choose topic  $z_{i,j} \sim ext{Categorical}( heta_i)$  word's topic
    - $\Box$  choose a word  $w_{i,j} \sim \text{Categorical}(\beta_{z_{i,j}})$
  - For the *i*th document, choose  $y_i \sim N(\eta^\top \bar{z}_i, \sigma^2)$ ,  $\bar{z}_i = \frac{1}{L} \sum_{i=1}^{L} z_{i,j}$



generate document's label

8

# **Seeded LDA: Guided Topic-Word Distribution**

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

- 1. For each  $k=1\cdots T$ ,
  - (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  $s = 1 \cdots S$ ,
  - (a) Choose group-topic distribution  $\psi_s \sim \text{Dir}(\alpha)$ .
- 3. For each document d,

```
(a) Choose a binary vector b of length S.
(b) Choose a document-group distribution ζ<sup>d</sup> ~ Dir(τb).
(c) Choose a group variable g ~ Mult(ζ<sup>d</sup>).
(d) Choose θ<sub>d</sub> ~ Dir(ψ<sub>g</sub>). // of length T
(e) For each token i = 1 · · · N<sub>d</sub>:

i. Select a topic z<sub>i</sub> ~ Mult(θ<sub>d</sub>).
ii. Select an indicator x<sub>i</sub> ~ Bern(π<sub>zi</sub>).
iii. if x<sub>i</sub> is 0

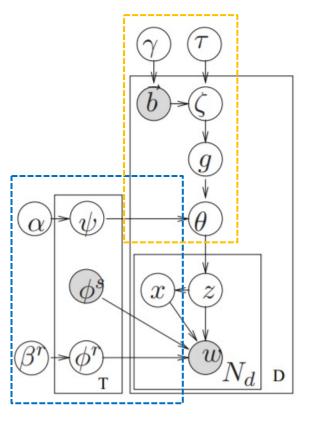
Select a word w<sub>i</sub> ~ Mult(φ<sup>r</sup><sub>zi</sub>).

iv. if x<sub>i</sub> is 1

Select a word w<sub>i</sub> ~ Mult(φ<sup>s</sup><sub>zi</sub>).
```

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

Traditional Topic Models

Embedding-Based 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]

Topic Discovery with PLMs

# **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	sports, united states	united states, iraq	
canadian, economy	olympic, games	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

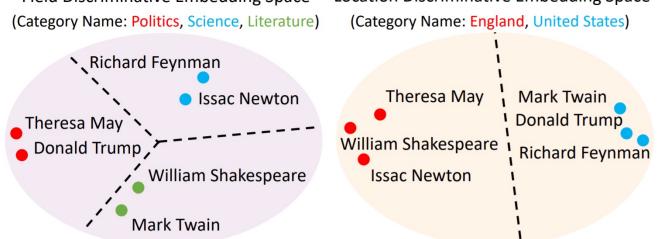
- Discriminative Topic Mining: Given a text corpus and a set of category names, retrieve a set of terms that exclusively belong to each category
  - □ E.g., given  $c_1$ : "The United States",  $c_2$ : "France",  $c_3$ : "Canada"
    - $\Box$  Yes to "Ontario" under  $c_3$ : (a province in Canada and exclusively belongs to Canada)
    - No to "North America" under c<sub>3</sub>: (a continent and does not belong to any countries (reversed belonging relationship))
    - No to "English" under c<sub>3</sub>: (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

- Traditional Topic Models
- Embedding-Based 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]
- Topic Discovery with PLMs

### **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 Field Discriminative Embedding Space Location Discriminative Embedding Space (Category Name: Politics, Science, Literature) Category names + Corpus Richard Feynman CatE: Outputs (see figure) Issac Newton Theresa May The same set of celebrities are Donald Trump
  - embedded differently given different sets of category names



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.

# **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
  - 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 usergiven categories

### **CatE Embedding: Objective**

Objective: negative log-likelihood

$$P(\mathcal{D} \mid C) = \prod_{d \in \mathcal{D}} p(d \mid c_d) \prod_{w_i \in d} p(w_i \mid d) \prod_{\substack{w_{i+j} \in d \\ -h \leq j \leq h, j \neq 0}} p(w_{i+j} \mid w_i)$$
  
1. Topic assignment 2. Global context 3. Local context  
$$p(d \mid c_d) \propto p(c_d \mid d)p(d) \propto p(c_d \mid d) \propto \prod_{w \in d} p(c_d \mid w), \text{ Decompose into word-topic distribution}$$

Introducing specificity

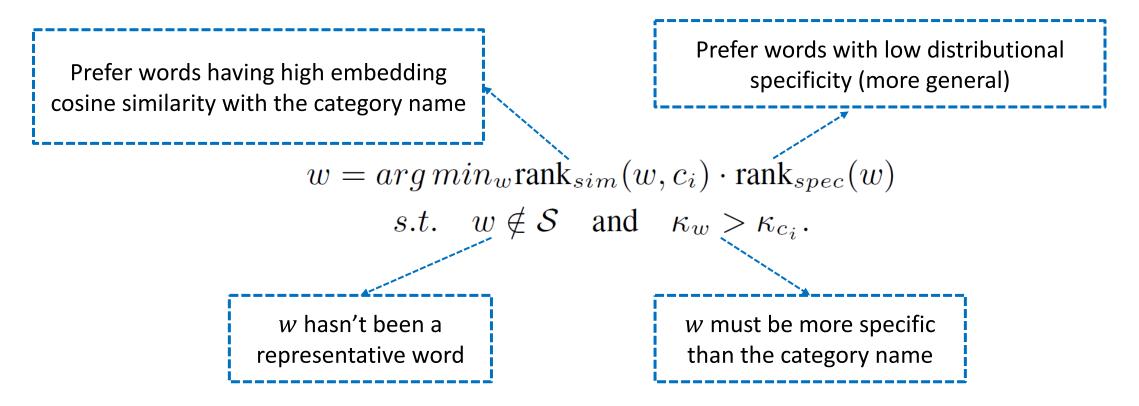
**Definition 2** (Word Distributional Specificity). We assume there is a scalar  $\kappa_w \ge 0$  correlated with each word *w* indicating how specific the word meaning is. The bigger  $\kappa_w$  is, the more specific meaning word *w* has, and the less varying contexts *w* appears in.

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:

 $\Box$  We find a representative word of category  $c_i$  and add it to the set S by



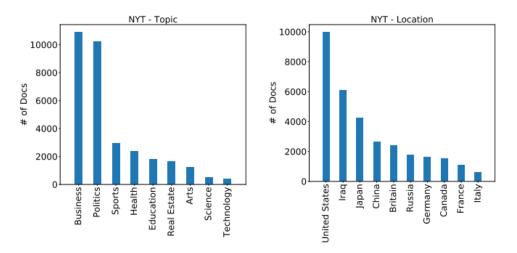
### **Quantitative Results**

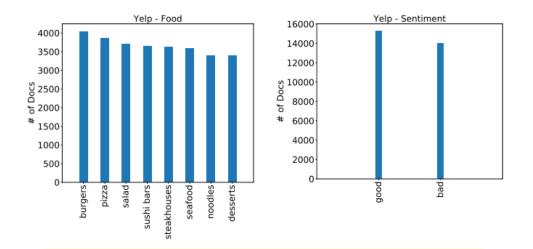
#### Two datasets:

- 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-Sentimen	
Wiethous	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	0.049	0.972	0.048	0.967	0.034	0.913	0.086	1.000





Dataset stat: # of docs by category name

### **Qualitative Results**

Methods         NYT-Location britain         NYT-Topic canada         Yelp-Food education         Yelp-Food politics         Yelp-Sentiment good         Yelp-Sentiment good         Yelp-Sentiment good         Yelp-Sentiment good         Yelp-Sentiment bad           LDA         companic (X) british shares (X)         economy (X) canadian         sthool cumpanics (X)         sthool cumpanics (X)         faburger cumpanics (X)         gerat british         valet (X) curve aid (X)           Seeded         british LDA         city (X)         state (X) schools         republican         like (X) president         great (X) presid (X)         great (X)										
DiffaincanadaeducationpointespurgerdessertsgoodbadLDAcompanies (x)economy (x)schoolcampaignfatburgerice creamgreatvalet (x)britishcanadiancity (x)mayorliar (x)gelatoloveaid (x)great britishcanadiancity (x)mayorbearing (x)sweetloveaid (x)great britishcity (x)schoolspoliticalbearing (x)sweetbreakfastrowlygeededbritishcity (x)state (x)republicanlike (x)great (x)greatgreatdid (x)deal (x)studentssenatorgreat (x)great (x)greatgreatdid (x)service (x)builting (x)building (x)studentssenatorgreat (x)delicious (x)great (x)did (x)builting (x)building (x)city (x)presidentgreat (x)delicious (x)great (x)ordered (x)germany (x)torontoarts (x)reigionfriesmomentaretastysubparsouth korea (x)asia (x)advisorsera (x)pattywhiped creansuffered (x)warfulanchoredlondonplayer (x)sentor(x)multistry (x)pattywhiped creansuffered (x)homest (x)german (x)canadianschoolsiraq (x)maltaregreatcame (x)suffered (x)homest (x)german (x)sports	Mathada	NYT-Lo	ocation	NYT	-Topic	Ye	lp-Food	Yelp-Sentiment		
LDAcompanies (x)economy (x)studentsclintondos (x)chocolateplace (x)peter (x)britishshares (x)united states (x)state (x)mayorlowlowaid (x)great britaintrade (x)schoolspoliticalbearing (x)sweetbreakfastrowdybritishcity (x)state (x)republicanlike (x)great (x)<	- Methods	britain	canada	education	politics	burger	desserts	good	bad	
LDAcompanies (x)economy (x) canadianstudentsclinton city (x)dos (x) mayorchocolate liar (x)glace (x) gelatopeter (x) loveSeeded LDAbritishcity (x)state (x) schoolsrepublical schoolslike (x) politicalglato friesgreat (x) like (x)glato (x) service (x) great (x)glato (x) service (x) great (x)glate (x) service (x) order (x)Seeded LDAbritishcity (x) deal (x) building (x) buildings (x)state (x) schoolrepublican politicallike (x) great (x) great (x)glace (x) service (x) great (x)germany (x) building (x) buildings (x)state (x) studentsrepublican schoollike (x) great (x)glace (x) great (x)TWE south korea (x) anakets (x) ankets (x)deronto asia (x) asia (x) asia (x) asia (x) asia (x) asia (x)eats (x) regultant (x)religion musicans (x) attraction (x)british green ta (x) cheeseburgertasty suffered (x) order (x) aradetasty awful arade (x) awful darad (x) suffered (x) order (x) suffered (x) schoolsselection (x) awful darad (x) schoolstasty suffered (x) awful darad (x) schoolsTWE Corefxmoscow (x) sports (x) greman (x) canada (x)sports (x) senator (x) schoolsregultan (x) mata (x) schoolsmate (x) mata (x) schoolsmate (x) awful schoolsselection (x) mata (x) schoolstasty great (x) mata (x)Anchored Corefx		company (×)	percent (×)	school	campaign	fatburger	ice cream	great	valet (×)	
LDAbritish shares (x)canadian united states (x)city (x) state (x)mayor electionliar (x) cheeseburgersgelato the electionlove friendlyrelid (x) relid (x)Seeded LDAbritishcity (x) building (x)state (x) schoolrepublican politicallike (x) politicalgreat (x) friesgreat (x) great (x) great (x)great (x) great (x) great (x)great (x) great (x) just (x)place (x) service (x)service (x) order (x)Seeded LDAbillion (x) building (x) business (x)street (x) york (x)students board (x)senator democratslike (x) great (x) delicious (x) just (x)great (x) itie (x) order (x)place (x) service (x) order (x)TWEgermany (x) manufacturing (x) booming (x) manufacturing (x) booming (x) asia (x) albertareligion fourth graders regentsburgers race fries pathyburgers market (x)chocolate green tea (x) homest (x) konest (x)tasty deecent positive (x) dered (x) homest (x)Anchored CorExmoscow (x) sports (x) greman (x) canadian canadian canadian schoolmilitary (x) schoolmilitary (x) market (x)military (x) democratic (x)mow (x) troops (x) gring (x)make (x) mark (x) gring (x)selection (x) mark (x) schoolselection (x) schoolselection (x) prices (x) gring (x)Anchored CorExmoscow (x) schoolsports (x) schoolsenator (x) regentsmilitary (x) troops (x) g		companies (×)	economy (×)	students	clinton	_	chocolate	place (×)	peter (×)	
shares (x) great britiainunited states (x) trade (x)state (x) schoolselection politicalcheeseburgers bearing (x)tea (x) sweetfriendly breakfastrelief (x) rowdySeeded LDAbritish deal (x)city (x) street (x)state (x) schoolrepublican politicallike (x) great (x)great (x) delcious (x) just (x)place (x) service (x) order (x)students service (x) order (x)Mathematic DDAbuilding (x) business (x)students order (x)senator service (x) democratsjust (x) itim (x)idelcious (x) just (x)place (x) service (x) order (x) isst (x)service (x) order (x) advisorsTWEgermany (x) south korea (x) markets (x)atorono albertareligion fourth graders race musicians (x) advisorsburgers attraction (x) hamburger pattyburgers chocolate fries complimentary (x) decenttasty subpar decent positive (x) dared (x) advisorsAnchored CorExmoscow (x) greman (x) canadian russian (x) canadian canadian canadian schoolrepublican (x) mator (x) schoolmilitary (x) war (x) troops (x) schoolorder (x) mark (x) schoolselection (x) dared (x)Labeled ETMfrace (x) canada (x)canadian british columbia schoolschool schoolpolitical resonable schoolhamburger schoolmake (x) going (x) war (x) troops (x) fries schoolgreat resonable tracefries resonable tracegreat terpiblic	LDA		canadian	city (×)	mayor	liar (×)	gelato	love	aid (×)	
Seeded LDAbritish industry (x) deal (x) buildings (x) deal (x) street (x) buildings (x) deal (x)state (x) schoolrepublican political senator political just (x) ice creamlike (x) great (x) just (x) ice creamplace (x) great (x) great (x) great (x) great (x) great (x) just (x)place (x) service (x) ordered (x) ordered (x)TWEgermany (x) spain (x) oosaka (x) manufacturing (x) border (x) alk (x)toronto arts (x) advisors great (x)arts (x) fourth graders musicians (x) advisors era (x)religion time (x) attraction (x) attraction (x) attraction (x) active (x) advisors german (x)toronto arts (x) advisors great tale (x)burgers fries complimentary (x) hamburger great tale (x)tasty decent avaful great crappy suffered (x) honest (x)Anchored CorExmoscow (x) sports (x) german (x) canadian russian (x) canadian canada (x)republican (x) senator (x)military (x) war (x) troops (x) frage (x) schoolsorder (x) marke (x) great cane (x) regentsselection (x) did (x) prices (x) great cane (x) regentsLabeled ETMfrance (x) canadian canada (x) british canada (x) british canada (x) british canada (x) british canada (x) british canada (x) british canada (x) british canada (x) british canada (x)higher education schoolspolitical expediency (x) schools preceptions (x) schoolshamburger pana canada (x) preceptions (x) schoolspana canad (x) british colece cane<		shares (×)	united states (×)		-	cheeseburgers	tea (×)	friendly	relief (×)	
Seeded LDAindustry (x) deal (x) builting (x) built		great britain	trade (×)	schools	political	bearing (×)	sweet	breakfast	rowdy	
Seeded LDAdeal (x)street (x)studentssenatorjust (x)ice creamservice (x)order (x)billion (x)buildings (x)buildings (x)board (x)democratsgreat (x)delicious (x)just (x)ordered (x)TWEgermany (x)torontoarts (x)religionburgerschocolatetastysubparmanufacturing (x)booming (x)musicians (x)attraction (x)hamburgergreen tea (x)derect (x)derect (x)south korea (x)asia (x)advisorsera (x)attraction (x)musicians (x)attraction (x)musicians (x)attraction (x)manufacturing (x)booming (x)advisorsera (x)avfulgreatcrappysouth korea (x)asia (x)advisorsera (x)war(x)chocolategreatcrappymarkets (x)albertaregentstale (x)war (x)know (x)chocolategreatcrappyfriesgreman (x)canadiansenator (x)war (x)chocolategreatcrappygerman (x)canadianschoolbaghdad (x)friesright (x)mac (x)tale (x)german (x)canadianschoolsiraq (x)going (x)want (x)mac (x)tale (x)german (x)canadianhigher educationalexpeciency (x)burgersdecenthorriblecorefgerman (x)canadian (x)british columbiaschoolsforeign affairspath		british	city (×)	state (×)	republican	like (×)	great (×)	place (×)	service (×)	
Seeded LDAdeal (x)street (x)studentssenatorjust (x)ice creamservice (x)order (x)billion (x)buildings (x)buildings (x)board (x)presidentgreat (x)delicious (x)just (x)order (x)TWEgermany (x)torontoarts (x)religionburgerschocolatetastysubparTWEspain (x)osaka (x)fourth gradersracefriescomplimentary (x)dered (x)awfulsouth korea (x)asia (x)advisorsera (x)eheeseburgersundaegreat (x)darned (x)awfulmanufacturing (x)booming (x)advisorsera (x)eheeseburgersundaegreat (x)darned (x)awfulsouth korea (x)asia (x)advisorsera (x)military (x)order (x)make (x)greatcrappymarkets (x)albertaregentstale (x)morta (x)military (x)order (x)make (x)selection (x)did (x)forthgerman (x)canadianschoolbaghdad (x)friesright (x)greatcrame (x)german (x)canadianschoolsiraq (x)going (x)want (x)mac (x)table (x)forthgerman (x)canadianschoolsprecepinos (x)stared (x)mac (x)table (x)forthgerman (x)canadianschoolsprecepinos (x)stared (x)mac (x)table (x)forthgerman (x)canadian </td <td>6 1 1</td> <td>industry (×)</td> <td>building (×)</td> <td>school</td> <td>political</td> <td>fries</td> <td>like (×)</td> <td>great</td> <td>did (×)</td>	6 1 1	industry (×)	building (×)	school	political	fries	like (×)	great	did (×)	
LDAbillion (x) business (x) york (x)buildings (x) board (x)city (x) board (x)president democratsgreat (x) time (x)delicious (x) just (x)just (x)time (x) ordered (x)TWEgermany (x) manufacturing (x) south korea (x) markets (x)toronto asia (x) albertaarts (x) fourth graders advisors regentsreligion race talk (x)burgers fries era (x) pattyburgers sundae sundae sundae sundae green tea (x) suffered (x)tasty decent positive (x) darned (x) awful great sundae greattasty darned (x) awful great great sundae great suffered (x)tasty decent houset (x) darned (x)Anchored CorExmoscow (x) german (x) canadian canadian canadian canadian schoolregublican (x) senator (x)military (x) war (x) troops (x) schoolsorder (x) make (x) going (x)make (x) make (x) people (x) great great crasonable asked (x) mac (x)Labeled ETMfrance (x) eanda(x)canadian british quebec schoolshigher education perceptions (x) school schoolspolitical perceptions (x) school perceptions (x)hamburger path steak (x)pana decentpana decentdecent horrible great crasonable asked (x)Labeled ETMfrance (x) canada (x)canadian britin (x) britishhigher education gerent (x)political perceptions (x) schoolshamburger path pathpana decentdecent horrible great trasonable tasked (x) </td <td></td> <td>deal (×)</td> <td>street (×)</td> <td>students</td> <td>senator</td> <td>just (×)</td> <td>ice cream</td> <td>service (×)</td> <td>order (×)</td>		deal (×)	street (×)	students	senator	just (×)	ice cream	service (×)	order (×)	
TWEgermany (x)torontoarts (x)religionTWEgermany (x)torontoarts (x)fourth gradersraceburgerschocolatetastysubparTWEmanufacturing (x)booming (x)asia (x)advisorsera (x)attraction (x)burgerschocolatetastytastysubparSouth korea (x)asia (x)albertaregentstale (x)military (x)order (x)make (x)gelection (x)did (x)Anchoredmoscow (x)sports (x)republican (x)military (x)order (x)make (x)selection (x)did (x)Joint Koreagerman (x)canadianschoolbaghdad (x)friesright (x)reasonableasked (x)Anchoredfrance (x)canadianschoolbaghdad (x)friesright (x)greatcanadianJoint Labeledfrance (x)canadianhigher educationpoliticalhamburgerganadecenthorribleETMgermany (x)britishquebecschoolsforeign affairsjattycheesecakeda(x)avfulLabeledgermany (x)britishquebecschoolsforeign affairsjattycheesecakeda(x)avfulLabeledforton schoolsinternational politicalhamburgergeastdecenthorribleLabeledfortishquebecschoolsforeign affairsjattycheesecakes <t< td=""><td>LDA</td><td>billion (×)</td><td>buildings (×)</td><td>city (×)</td><td>president</td><td></td><td>delicious (×)</td><td>just (×)</td><td>time (×)</td></t<>	LDA	billion (×)	buildings (×)	city (×)	president		delicious (×)	just (×)	time (×)	
TWEspain (x) manufacturing (x) south korea (x) markets (x)osaka (x) booming (x) asia (x)fourth graders musicians (x) advisorsrace attraction (x) attraction (x) attraction (x) advisorsfries era (x) pattycomplimentary (x) hamburger cheeseburgerdecent sundae green tea (x) sundaedecent positive (x) advisorspositive (x) advisorsAnchored CorExmoscow (x) perman (x) crasian (x)sports (x) senator (x)republican (x) senator (x)military (x) war (x)order (x) market (x)make (x) people (x) greatselection (x) did (x)did (x) prices (x)Anchored CorExmoscow (x) german (x) crasian (x)sports (x) canadian schoolrepublican (x) schoolsmilitary (x) trops (x) schoolsorder (x) trops (x) trops (x)make (x) know (x)selection (x) did (x) prices (x) just (x) great called (x)selection (x) trops (x) greatdid (x) prices (x) just (x) great called (x)Labeled ETMfrance (x) gremany (x) british canada (x)higher education political regentspolitical trops (x) schoolshamburger pana tract (x)pana decentdecent horrible great tasty good (x)Labeled ETMfrance (x) gremany (x) british canada (x)bigher education political schoolspolitical foreign affairs ideologyhamburger pana cheeseburgerpana gelato great tasty good (x)decent terrible tasty good (x)Labeled ETMfrance (x)<		business (×)	york (×)	board (×)	democrats	time (×)	just (×)	ordered (×)	ordered (×)	
TWEmanufacturing (x) south korea (x) markets (x)booming (x) asia (x) albertamusicians (x) advisorsattraction (x) era (x) regentshamburger tale (x)green tea (x) sundae whipped creamdarned (x) sundae whipped creamdarned (x) sundae greatawful greatAnchored CorExmoscow (x) britishsports (x) german (x)republican (x) senator (x)military (x) war (x)order (x) know (x)make (x) chocolate greatselection (x) did (x)did (x) prices (x) just (x)Anchored CorExmoscow (x) britishsports (x) canadian russian (x)republican (x) canadian schoolmilitary (x) war (x)order (x) know (x)make (x) chocolate greatselection (x) did (x) prices (x) just (x)Labeled ETMfrance (x) germany (x)canadian british olumbia canada (x)higher education educational educational educational educational schoolspolitical educational expediency (x) perceptions (x)hamburger mac (x)pana cheeseburger gelatodecent decenthorrible great table (x)Labeled ETMfrance (x) portish quebec europe (x) north america (x)higher education regentspolitical ideologyhamburger transinal political burgershamburger transit pathypana cheesecakes bad (x)decent bad (x) awful delicious appalinglyLabeled ETMfrance (x) portish quebec uonge (x) north america (x)higher education regentspolitical tideolog		germany (×)	toronto	arts (×)	religion	burgers	chocolate	tasty	subpar	
south korea (x) markets (x)asia (x) albertaadvisors regentsera (x) tale (x)cheeseburger pattysundae whipped creamgreat suffered (x)crappy honest (x)Anchored CorExmoscow (x) britishsports (x) garman (x)sports (x) canadianrepublican (x) senator (x)military (x) war (x)order (x) know (x)make (x) chocolate called (x)selection (x) prices (x) just (x)Anchored CorExlondon german (x) russian (x)canadian coachrepublican (x) senator (x)military (x) war (x)order (x) know (x)make (x) chocolate called (x)selection (x) prices (x) just (x) greatdid (x) greatLabeled ETMfrance (x) british canada (x)canadian british columbia british quebechigher education educational schoolspolitical perceptions (x) greatshamburger cheeseburger burgerspana cheesecake burgersdecent terrible tasty good (x) bad (x) awfulLabeled ETMfrance (x) british quebeccanadian political schoolspolitical regentshamburger ideologypana cheesecake burgersdecent terrible tasty good (x) bad (x) awfulLabeled ETMfrance (x) prices (x) prices (x)north america (x)regentsideologyhamburger cheesecakespana cheesecakesdecent decinus appallinglyLabeled ETMfrance (x) prices (x)north america (x)regentsideologyhamburger cheesecakesg		spain (×)	osaka (×)	fourth graders	race	fries	complimentary (×)	decent	positive (×)	
markets (x)albertaregentstale (x)pattywhipped creamsuffered (x)honest (x)Anchored CorExbritishgames (x)republican (x)military (x)order (x)make (x)selection (x)did (x)Jondonplayers (x)german (x)canadiansenator (x)war (x)know (x)chocolateprices (x)just (x)german (x)canadianschoolbaghdad (x)going (x)want (x)mac (x)table (x)Labeledfrance (x)canadianschoolsiraq (x)going (x)want (x)mac (x)table (x)Labeledfrance (x)canadianhigher educationpoliticalhamburgerpanadecenthorribleETMfrance (x)canadianschoolsperceptions (x)schoolsperceptions (x)schoolsperceptions (x)pattycheesecakebad (x)awfulLabeledeurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyCatEenglandontarioeducationalpoliticalburgersdesertdelicioussickeningbritonsquebecschoolsinternational politicsburgersdesertdelicioussickeningLabeledbritonsquebecschoolsinternational politicsburgersdesertdeliciousawfulLabeledfriesscotishmontrealsecondary educationpoliticalburgersde	TWE	manufacturing (×)	booming $(\times)$	musicians (×)	attraction $(\times)$	hamburger	green tea (×)	darned (×)	awful	
markets (x)albertaregentstale (x)pattywhipped creamsuffered (x)honest (x)Anchored CorExbritish britishgames (x) londonsenator (x)war (x) senator (x)order (x) war (x)make (x) know (x)selection (x)did (x) prices (x)Anchored CorExlondon german (x) russian (x)canadian coachschoolbaghdad (x) schoolsfries iraq (x)rigon (x)want (x)selection (x)did (x) prices (x)Labeled ETMfrance (x) britishcanadian canada (x)higher education british columbia schoolspolitical educationalhamburger expediency (x) schoolspana cheeseburger burgersdecent terrible trasty good (x)Labeled ETMfrance (x) britishcanadian quebechigher education regentspolitical ideologyhamburger prices (x) going (x)pana want (x)decent terrible trasty good (x)Labeled ETMfrance (x) britishcanadian quebechigher education regentspolitical ideologyhamburger steak (x)pana ice creamdecent decenthorrible greatLabeled ETMengland londonontario torontoeducational schoolspolitical ideologyhamburger steak (x)pana ice creamdecent delicioushorible greatLabeled ETMengland londonontario londoneducational schoolspolitical international political higher educationburgers polit		south korea (×)	asia (×)	advisors	era (×)	cheeseburger	sundae	great	crappy	
Anchored CorExbritish londongames (x) players (x) canadiansenator (x) democratic (x)war (x) troops (x) baghdad (x) iraq (x)know (x) called (x) people (x)chocolate prices (x) greatprices (x) came (x) reasonableLabeled ETMfrance (x) germany (x)canadian british canada (x)higher education british canadianpolitical educationalhamburger expediency (x) schoolpana burgersdecent decenthorrible greatLabeled ETMfrance (x) britishcanadian britishhigher education educationalpolitical expediency (x) schoolshamburger perceptions (x) perceptions (x)pana burgersdecent tasty good (x)Labeled ETMfrance (x) britishcanadian quebechigher education schoolspolitical ideologyhamburger steak (x)pana cheeseburgerdecent decenthorrible greatLabeled ETMfrance (x) britishcanadian quebechigher education schoolspolitical ideologyhamburger steak (x)pana cheesecakedecent decenthorrible decentLabeled ETMfrance (x) britishcanadian quebechigher educational schoolspolitical ideologyhamburger steak (x)pana cheesecakedecent decenthorrible decentLabeled ETMfrance (x) politicalnorth america (x)regentsideologyideologyisteak (x)ideologyCatEengland britons scottisho		markets (×)	alberta	regents	tale (×)	patty	whipped cream	suffered (×)		
Anchored CorExIondon german (x) russian (x)players (x) canadian coachdemocratic (x) school schooltroops (x) baghdad (x) iraq (x)called (x) fries going (x)people (x) right (x) going (x)great reasonable asked (x) mac (x)Labeled ETMfrance (x) germany (x) british canadia (x)canadian british columbia british columbia britishhigher education educational schoolpolitical expediency (x) perceptions (x) britishhamburger pana canada (x)pana decentdecent horrible great table (x)Labeled ETMfrance (x) germany (x) britishcanadian political schoolhigher education schoolpolitical perceptions (x) schoolshamburger pana expediency (x) perceptions (x) pattypana cheesecake burgersdecent tasty good (x) bad (x) awful delicious appallinglyCatEengland londonontario toronto schoolseducational regentspolitical international politics higher education liberalism secondary education political philosophyburgers cheeseburger pastriesdelicious mindful masty excellent delicious wonderful freaks		moscow (×)	sports (×)	republican (×)	military (×)	order (×)	make (×)	selection (×)	did (×)	
CorExIondonplayers (x)democratic (x)troops (x)called (x)people (x)greatcame (x)german (x)canadianschoolbaghdad (x)friesright (x)reasonableasked (x)russian (x)coachschoolsiraq (x)going (x)want (x)mac (x)table (x)Labeledfrance (x)canadianhigher educationpoliticalhamburgerpanadecenthorriblegermany (x)british columbiaeducationalexpediency (x)cheeseburgergelatogreatterriblecanada (x)britain (x)schoolperceptions (x)burgerstiramisutastygood (x)britishquebecschoolsforeign affairspattycheesecakebad (x)awfuleurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyCatEbritonsquebechigher educationliberalismburgersdesertdelicioussickeningbritonsquebecschoolsinternational politicalburgersdesertdelicioussickeningcheeseburgerpastriesmindfulnastycatebritonsquebechigher educationliberalismhamburgercheesecakesexcellentdeliciousschoolsinternational politicalburgersdesertdelicioussickeningfordontorontoschoolsinternational politicalburgers <td>A</td> <td>british</td> <td>games (×)</td> <td>senator (×)</td> <td>war (×)</td> <td>know (×)</td> <td>chocolate</td> <td>prices (×)</td> <td>just (×)</td>	A	british	games (×)	senator (×)	war (×)	know (×)	chocolate	prices (×)	just (×)	
german (x) russian (x)canadian coachschoolbaghdad (x) iraq (x)fries going (x)right (x) want (x)reasonable mac (x)asked (x) table (x)Labeled ETMfrance (x) germany (x)canadian british columbia canada (x)higher education british columbia schoolpolitical perceptions (x) perceptions (x)hamburger burgerspana terrible decentdecent horrible greathorrible terrible greatLabeled ETMfrance (x) germany (x)canadian british columbia otada (x)higher educational schoolpolitical perceptions (x)hamburger burgerspana terrible tiramisudecent terrible greathorrible terribleLabeled ETMfrance (x) perceptions (x)british outperceptions (x)political schoolshamburger pattypana cheesecakedecent decenthorrible tastyLabeled ETMgermany (x) britishbritain (x) quebecschools schoolsforeign affairs ideologypatty steak (x)ice creamdecent deliciousawful deliciousengland londonontario torontoeducational schoolspolitical international politicsburgers cheeseburgerdessert pastriesdelicious mindful mastyCatEbritons scottishquebec montrealhigher education secondary educationpolitical philosophyburger king burger kingsconeswonderful freaks		london	players (×)	democratic (×)	troops (×)	called (×)	people (×)	great	came (×)	
russian (x)coachschoolsiraq (x)going (x)want (x)mac (x)table (x)Labeled ETMfrance (x)canadianhigher educationpoliticalhamburgerpanadecenthorriblegermany (x) canada (x)british columbia britishbritain (x)schoolperceptions (x)cheeseburgergelatogreatterriblebritish europe (x)onth america (x)schoolsforeign affairspattycheesecakebad (x)awfuleurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyCatEbritons scottishquebechigher educationliberalismburgerscheesecakes scondary educationburgersdessertdelicioussickening mindfulScottishmontrealsecondary educationpolitical politicalburgerscheesecakes pastriesexcellentdreadful mondfulfrance (x)guebechigher educationliberalism politicalburger kingsconeswonderfulfreaks	CorEx	german (×)	canadian	school	baghdad (×)	fries	right (×)	reasonable	asked (×)	
Labeled ETMgermany (x) canada (x) britishbritish columbia britain (x) quebeceducational schoolexpediency (x) perceptions (x) ideologycheeseburger burgersgelato tiramisugreat tastyterrible tastyETMbritish quebecoutpace schoolsschool foreign affairs ideologypattycheesecake bad (x)bad (x) appallinglyeurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyengland londonontario torontoeducational schoolspolitical international politicsburgers cheeseburgerdessert pastriesdelicioussickening mindfulCatEbritons scottishquebec montrealhigher education secondary educationliberalism political philosophyburger kingsconesexcellent wonderfuldreadful freaks			coach	schools	iraq (×)	going (×)	want (×)	mac (×)	table (×)	
Labeled ETMcanada (x)britain (x)schoolperceptions (x)burgerstiramisutastygood (x)britishquebecschoolsforeign affairspattycheesecakebad (x)awfuleurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyenglandontarioeducationalpoliticalburgersdessertdelicioussickeninglondontorontoschoolsinternational politicscheeseburgerpastriesmindfulnastyCatEbritonsquebechigher educationliberalismhamburgercheesecakesexcellentdreadfulscottishmontrealsecondary educationpolitical philosophyburger kingsconeswonderfulfreaks		france (×)	canadian	higher education	political	hamburger	pana	decent	horrible	
ETMcanada (x)britain (x)schoolperceptions (x)burgerstiramisutastygood (x)britishquebecschoolsforeign affairspattycheesecakebad (x)awfuleurope (x)north america (x)regentsideologysteak (x)ice creamdeliciousappallinglyenglandontarioeducationalpoliticalburgersdessertdelicioussickeninglondontorontoschoolsinternational politicscheeseburgerpastriesmindfulnastyCatEbritonsquebechigher educationliberalismhamburgercheesecakesexcellentdreadfulscottishmontrealsecondary educationpolitical philosophyburger kingsconeswonderfulfreaks	Labeled	germany (×)	british columbia	educational	expediency (×)	cheeseburger	gelato	great	terrible	
britishquebecschoolsforeign affairspattycheesecakebad (×)awfuleurope (×)north america (×)regentsideologysteak (×)ice creamdeliciousappallinglyenglandontarioeducationalpoliticalburgersdessertdelicioussickeninglondontorontoschoolsinternational politicscheeseburgerpastriesmindfulnastyCatEbritonsquebechigher educationliberalismhamburgercheesecakesexcellentdreadfulscottishmontrealsecondary educationpolitical philosophyburger kingsconeswonderfulfreaks		canada (×)	britain (×)	school	perceptions (×)	burgers	tiramisu	tasty	good (×)	
CatEengland londonontario torontoeducational schoolspolitical international politicsburgers cheeseburgerdessert pastriesdelicious mindfulsickening nastyCatEbritons scottishquebec montrealhigher education secondary educationliberalism political philosophyburgers burger kingdessert sconesdelicious mindfulsickening nasty	LIM	british	quebec	schools	foreign affairs	patty	cheesecake	bad (×)	awful	
CatEIondontorontoschoolsinternational politicscheeseburgerpastriesmindfulnastyCatEbritonsquebechigher educationliberalismhamburgercheesecakesexcellentdreadfulscottishmontrealsecondary educationpolitical philosophyburger kingsconeswonderfulfreaks		europe (×)	north america (×)	regents	ideology	steak (×)	ice cream	delicious	appallingly	
CatEbritonsquebechigher educationliberalismhamburgercheesecakesexcellentdreadfulscottishmontrealsecondary educationpolitical philosophyburger kingsconeswonderfulfreaks		england	ontario	educational		burgers	dessert	delicious	sickening	
scottish montreal secondary education political philosophy burger king scones wonderful freaks		london	toronto	schools	international politics	cheeseburger	pastries	mindful	nasty	
	CatE	britons	quebec	higher education	liberalism	hamburger	cheesecakes	excellent	dreadful	
great britain ottawa teachers geopolitics smash burger ice cream faithful cheapskates		scottish	montreal	secondary education		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	cientist, academic, research, laboratory machine, equipment, devices, engineering 1			
$1.25\kappa_c < \kappa < 1.5\kappa_c$	physics, sociology,	information technology, computing,	mental hygiene, infectious diseases,		
	biology, astronomy	telecommunication, biotechnology	hospitalizations, immunizations		
1 5 1 1 1 75 1	microbiology, anthropology,	wireless technology, nanotechnology,	dental care, chronic illnesses,		
$1.5\kappa_c < \kappa < 1.75\kappa_c$	physiology, cosmology	semiconductor industry, microelectronics	cardiovascular disease, diabetes		
	national science foundation,	integrated circuits,	juvenile diabetes,		
$\kappa > 1.75\kappa_c$	george washington university,	assemblers,	high blood pressure,		
	hong kong university,	circuit board,	family violence,		
	american academy	advanced micro devices	kidney failure		

The table lists the most similar words/phrases with each category (measured by embedding cosine similarity) from different ranges of distributional specificity

 $\Box$  When  $\kappa$  is smaller, the retrieved words have wider semantic coverage

In our model design, if not imposing constraints on the κ, the retrieved words might be too general and do not belong to the category

### Outline

- Traditional Topic Models
- Embedding-Based 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]
- Topic Discovery with PLMs

### **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

• ...

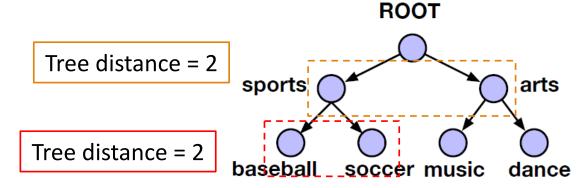
Hierarchical topic models discover topic structures from text corpora via modeling the text generative process with a latent hierarchy

Meng, Y., Zhang, Y., Huang, J., Zhang, Y., Zhang, C., & Han, J. (2020). Hierarchical topic mining via joint spherical tree and text embedding. KDD.

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

Use tree distance in a relative manner: Since  $d_{tree}$ (sports, baseball)  $< d_{tree}$ (baseball, 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

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

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

$$p(w_j \mid d_i) \propto \exp(\cos(\boldsymbol{u}_{w_j}, \boldsymbol{d}_i))$$

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

 $p(w_{j+k} \mid w_j) \propto \exp(\cos(\boldsymbol{v}_{w_{j+k}}, \boldsymbol{u}_{w_j}))$ 

3. Local context

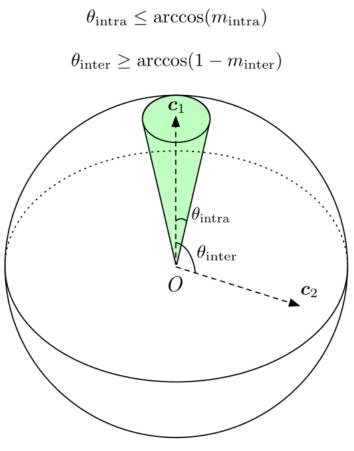
## **JoSH Tree Embedding**

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

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

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

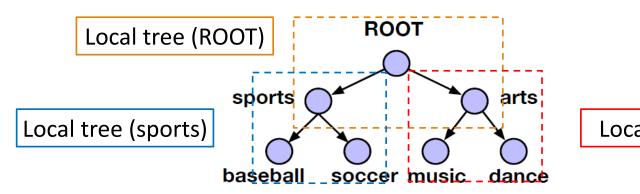
$$\mathcal{L}_{\text{inter}} = \sum_{c_i \in \mathcal{T}} \sum_{c_j \in \mathcal{T} \setminus \{c_i\}} \min(0, 1 - c_i^{\top} c_j - m_{\text{inter}}).$$



(a) Intra- & Inter-Category Configuration.

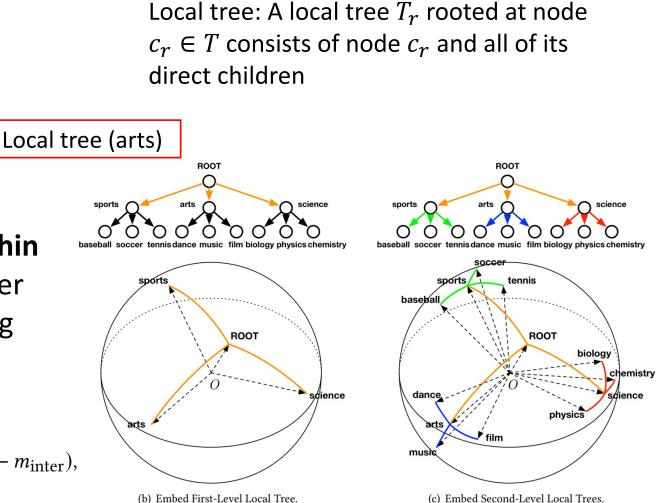
# **JoSH Tree Embedding**

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



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

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



### **Experiments: Qualitative Results on NYT**

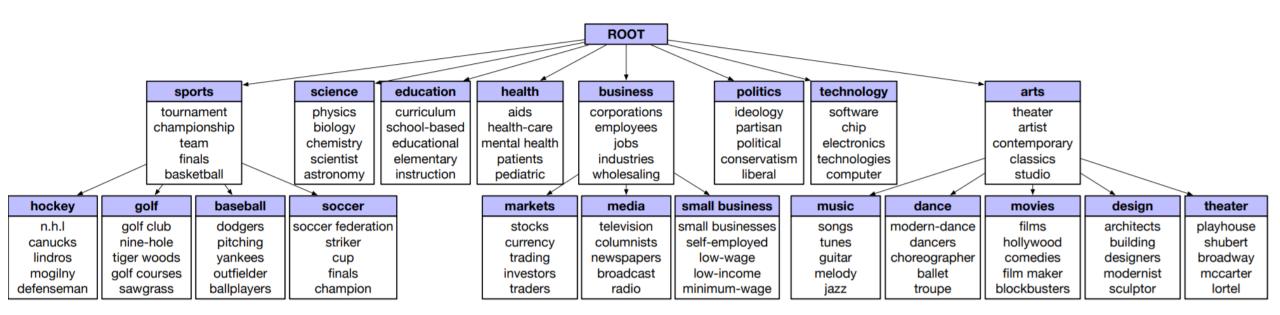
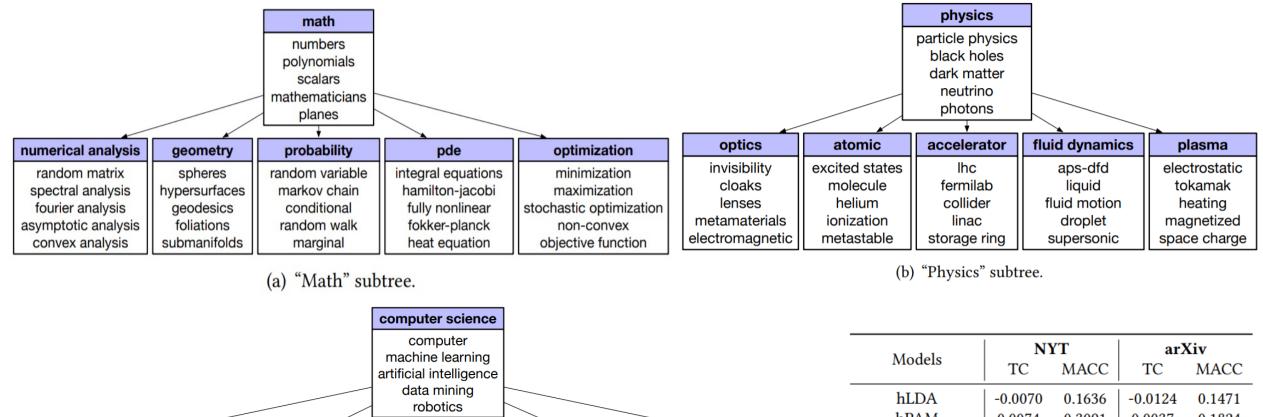


Figure 3: Hierarchical Topic Mining results on NYT.

### Experiments: Qualitative Results on ArXiv and Quantitative Results



					hPAM	0.0074	0.3091	0.0037	0.1824
natural language processing	pattern recognition	networking	programming languages	game theory	JoSE	0.0140	0.6818	0.0051	0.7412
machine translation	image processing	cloud computing	libraries	decision problems	Poincaré GloVe	0.0092	0.6182	-0.0050	0.5588
parsing	computer vision	p2p	python	influence diagrams	Anchored CorEx	0.0117	0.3909	0.0060	0.4941
question answering	image segmentation	iot	java	two-player	CatE	0.0149	0.9000	0.0066	0.8176
information extraction	object recognition	sdn	C++	incomplete information	JoSH	0.0166	0.9091	0.0074	0.8324
summarization	vision tasks	virtualization	compiler	nash equilibria		0.0100	0.9091	0.0074	0.0524

(c) "Computer Science" subtree.

### Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM'23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD'22]

### 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 lowquality and unstable clusters

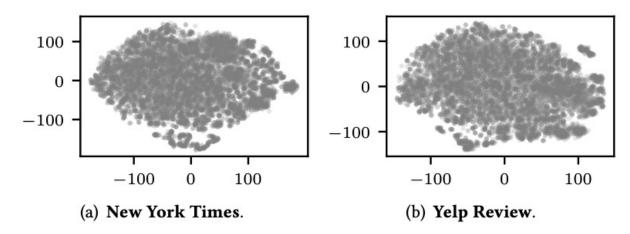


Figure 1: Visualization using t-SNE of 10,000 randomly sampled contextualized word embeddings of BERT on (a) NYT and (b) Yelp datasets, respectively. The embedding spaces do not have clearly separated clusters.

# **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 << |V|) is used, the resulting partition will not fit the original data well, resulting in unstable and low-quality clusters</p>
  - If a bigger K (K ≈ |V|) is used, most clusters will contain only one unique term, which is meaningless for topic discovery

# **The Latent Space Model**

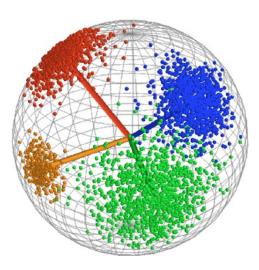
- We propose to project the original embedding space into a latent space with K clusters of words corresponding to K latent topics
- We assume that the latent space is **lower-dimensional** and **spherical**, with the following preferable properties:
  - Spherical latent space employs angular similarity 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

 $t_k \sim \text{Uniform}(K), \ \boldsymbol{z}_i \sim \text{vMF}_{d'}(\boldsymbol{t}_k, \kappa), \ \boldsymbol{h}_i = g(\boldsymbol{z}_i).$ 

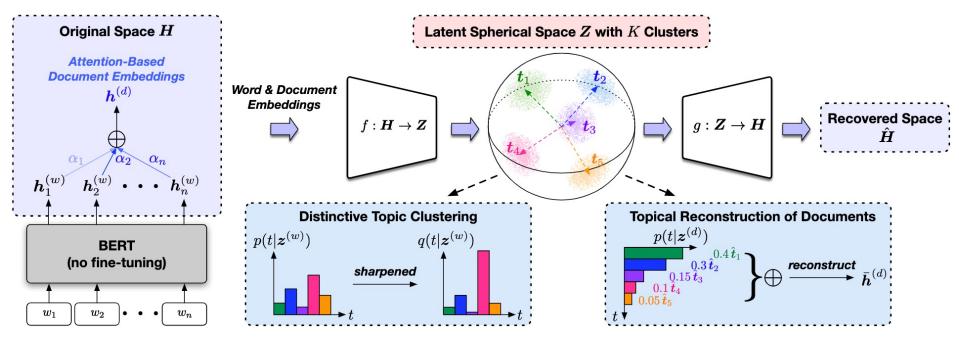
- □ A topic *t* is sampled from a uniform distribution over the K topics
- A latent embedding z is generated from the vMF distribution associated with topic t



# **The Latent Space Model**

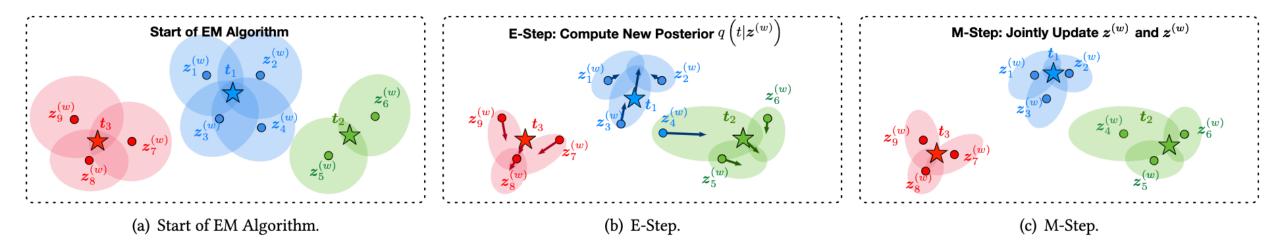
### How to train the generative model?

- □ A preservation loss that encourages the latent space to preserve the semantics of the original pretrained 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



## Experiments

### Topic Discovery

#### Quantitative

Methods		NY	Г		Yelp UMass UCI Int. Div.					
	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	-2.67	-0.45	0.93	0.99	-1.35	-0.27	0.87	0.96		

	Qualitative											
	NYT						Yelp					
Methods	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5		
	(sports)	(politics)	(research)	(france)	(japan)	(positive)	(negative)	(vegetables)	(fruits)	(seafood		
	olympic	<u>mr</u>	<u>said</u>	french	japanese	amazing	loud	spinach	mango	fish		
	<i>year</i>	bush	report	<u>union</u>	tokyo	<i>really</i>	awful	carrots	strawberry	<u>roll</u>		
LDA	<u>said</u>	president	evidence	germany	<u>year</u>	<i>place</i>	sunday	greens	<u>vanilla</u>	salmon		
	games	white	findings	workers	matsui	phenomenal	<u>like</u>	salad	banana	<i>fresh</i>		
	team	house	defense	paris	said	pleasant	slow	dressing	peanut	good		
	baseball	house	possibility	french	japanese	great	<u>even</u>	garlic	strawberry	shrimp		
	championship	white	challenge	<i>italy</i>	tokyo	friendly	bad	tomato	<u>caramel</u>	beef		
CorEx	playing	support	reasons	paris	<u>index</u>	atmosphere	mean	onions	sugar	crab		
	<i>fans</i>	groups	give	francs	osaka	love	cold	toppings	fruit	dishes		
	league	member	planned	jacques	electronics	favorite	literally	slices	mango	salt		
	olympic	government	approach	french	japanese	nice	disappointed	avocado	strawberry	fish		
	league	national	problems	<u>students</u>	agreement	worth	cold	greek	mango	shrimp		
ETM	<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>	german	<u>market</u>	recommend	experience	spinach	<u>soft</u>	crab		
	athletes	support	give	american	european	friendly	bad	tomatoes	flavors	chips		
	swimming	bush	researchers	french	japanese	awesome	horrible	tomatoes	strawberry	lobster		
	freestyle	democrats	scientists	paris	tokyo	atmosphere	<u>quality</u>	avocado	mango	crab		
BERTopic	<u>popov</u> gold	white bushs	cases genetic	lyon minister	ufj company	friendly night	disgusting disappointing	soups kale	<u>cup</u> lemon	shrimp oysters		
	olympic	house	study	billion	yen	good	place	cauliflower	banana	amazin		
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

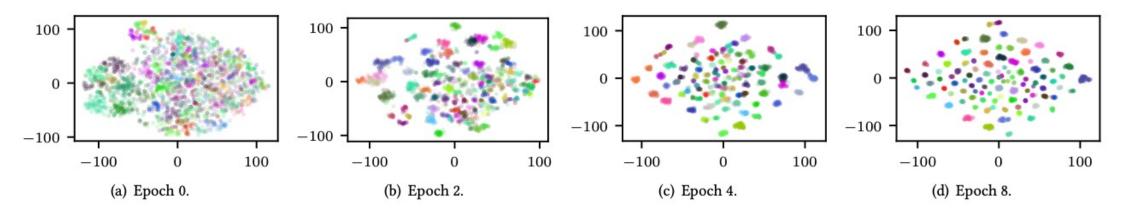


Figure 5: Visualization using t-SNE of 10,000 randomly sampled latent embeddings during the course of TopClus training. Embeddings assigned to the same cluster are denoted with the same color. The latent space gradually exhibits distinctive and balanced cluster structure.

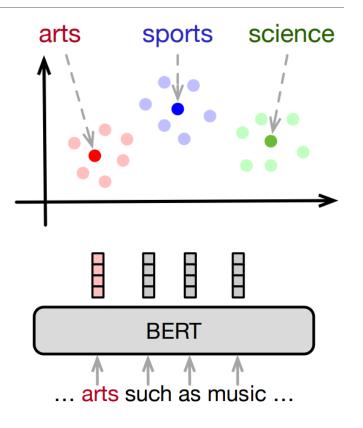
### Outline

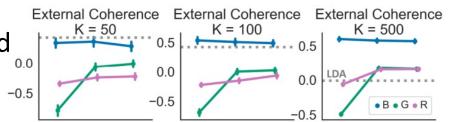
- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM'23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD'22]

# **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.



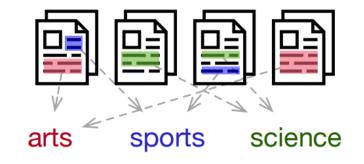


[1] Thompson, L., and Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv.

# **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.

## SeedTopicMine: Overview

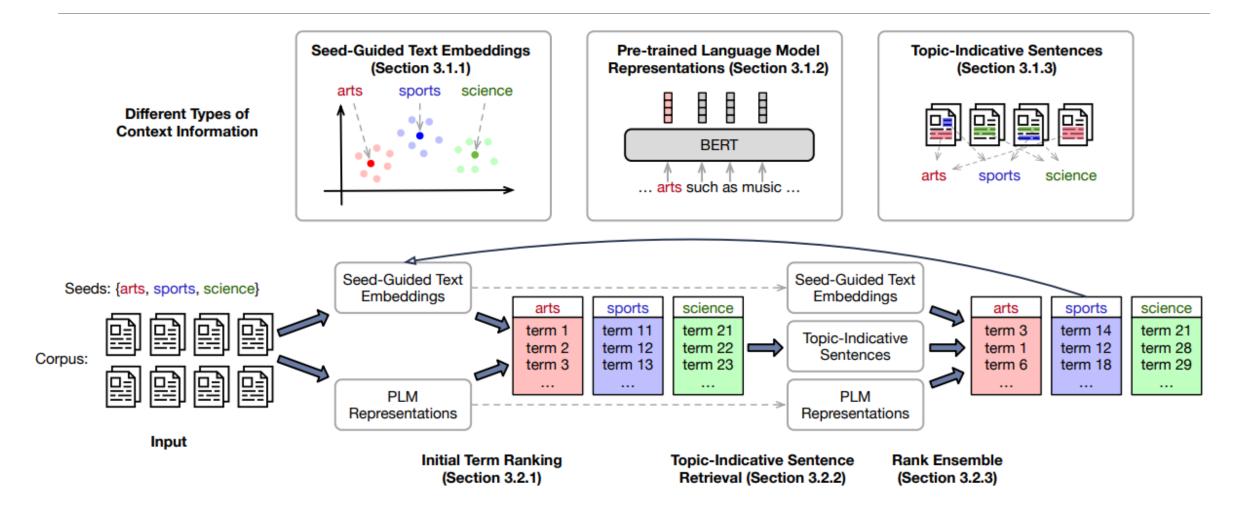
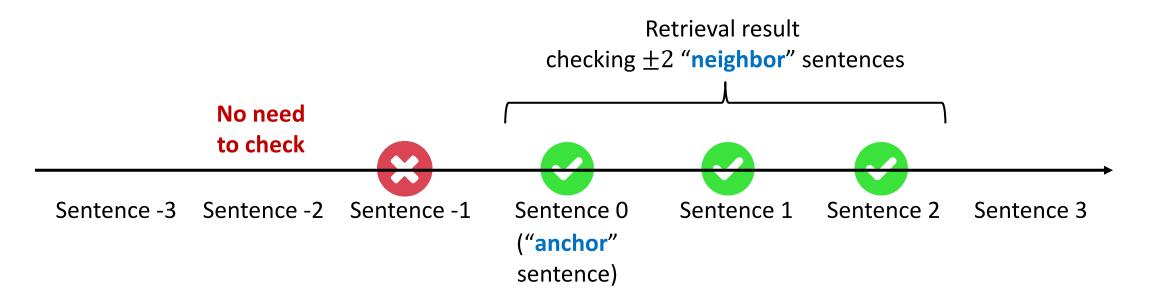


Figure 1: Overview of the SEEDTOPICMINE framework.

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 topicindicative 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 topicindicative 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	0.2058	0.6812	0.7312	0.1509	0.7150	0.7713
SeedTopicMine	0.1947	0.9456	0.9573	0.2176	0.8360	0.8709	0.2018	0.7912	0.8379	0.0922	0.9750	0.9811

Method	Yel	<b>p</b> -Food	Yelp-Sentiment		
Wiethou	P@20	NDCG@20	P@20	NDCG@20	
SeedTopicMine	0.7912	0.8379	0.9750	0.9811	
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**

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

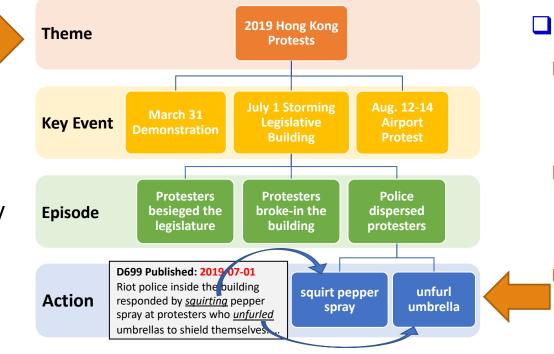
Method	N	YT-Topic	NYT-	Location	Yelp-F	ood	Yelp-Sentiment		
Method	health	business	france	canada	sushi	desserts	good	bad	
	said (×)	said (×)	said (×)	new (×)	roll	food (×)	place (×)	food (×)	
	dr (×)	percent (×)	new (×)	city (×)	good (×)	us (×)	food (×)	service (×)	
SeededLDA	new (×)	company	state (×)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

### Outline

- Traditional Topic Models
- Embedding-Based Discriminative Topic Mining
- Topic Discovery with PLMs
  - TopClus: Topic Discovery via Latent Space Clustering of Pretrained Language Model Representations [WWW'22]
  - SeedTopicMine: Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts [WSDM'23]
  - EvMine: Unsupervised Key Event Detection from Massive Text Corpora [KDD'22]

## **Motivation & Related Works**

- Real-world events are naturally organized in a hierarchical way
  - "Big events" have more general themes and may last longer
  - "Small events" have more concrete topics and may last shorter
- Topic Detection and Tracking
  - Detects themes from a corpus as document clusters
  - Themes are often topically distinct and thus easy to separate
  - These methods cannot distinguish key events of similar theme



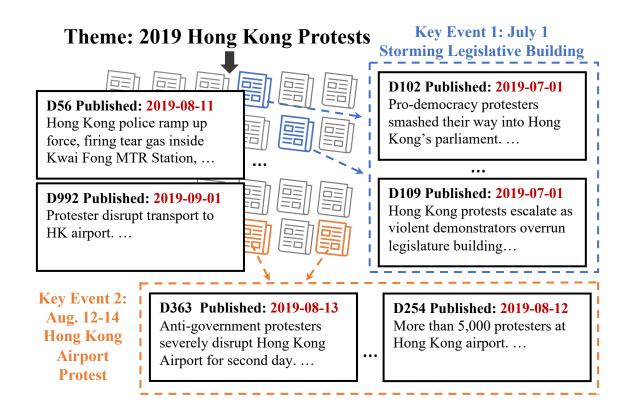
#### Action Extraction

- Extracts mention-level actions with triggers and arguments
- Rely on human curated schema and labeled training data
  - Too fine-grained to get an overall picture of an event

## **A New task: Key Event Detection**

#### Goal:

- Detects key events given a news corpus about one general theme
- Key events: non-overlapping documents clusters that not necessarily exhaust the corpus
- Challenges:
  - Key events are thematically similar and temporally closer
  - Impractical to label documents for model training



Zhang, Y., Guo, F., Shen, J., & Han, J. (2022). Unsupervised Key Event Detection from Massive Text Corpora. KDD.

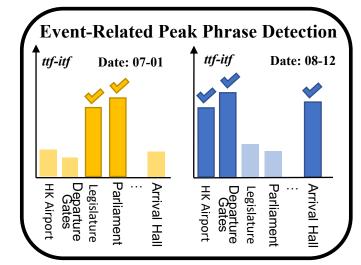
## **EvMine: Event-related Peak Phrase Detection**

- We introduce the idea of <u>temporal term frequency</u> <u>inverse time frequency</u>
- □ Temporal term frequency (**ttf**):
  - Measures how frequent a phrase is on a day
  - Aggregates frequencies from later days with decreasing weights to for delays and back referencing in news articles

$$ttf(p,t) = \frac{1}{n_t} \sum_{i=0}^{n_t-1} \left(1 - \frac{i}{n_t}\right) freq_{t+i}(p),$$

- Inverse time frequency (itf):
  - An event-indicative phrase will only be mentioned frequently around the event happening time

$$\operatorname{itf}(p) = \frac{\max \mathcal{T} - \min \mathcal{T} + 1}{|\{t \in \mathcal{T} | freq_t(p) > 0\}|},$$

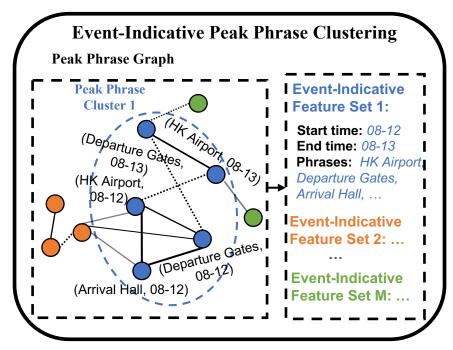


#### **Top-ranked Peak Phrases**

Phrase	Time
Hong Kong airport	2019-08-12
Victoria park	2019-08-18
legislative council	2019-07-01
Hong Kong airport	2019-08-13

## **EvMine: Event-indicative Peak Phrase Clustering**

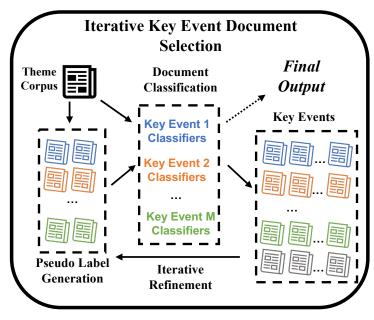
- A graph-based method to combine <u>textual</u> and <u>temporal</u> information
- Peak phrase graph construction:
  - **D** Each node is a peak phrase  $n_i = (p_i, t_i)$
  - Two types of edges:
    - Same-day peak phrases: edge weights are combination of NPMI for documentlevel thematic similarity and PLM-based phrase embedding similarity for semantic closeness.
    - Same-phrase consecutive-day peak phrases: connected with a constant edge weight (> 1)
- Form event-indicative feature sets with Louvain community detection algorithm



# **EvMine: Iterative Key Event Document Selection**

#### Pseudo Label Generation:

- Select top-ranked documents by their number of times matched with event-indicative phrases
- **Classifier Training: sampling and ensemble** 
  - Observation: much more negatives than positives in the corpus for each key event
  - For each key event, train multiple binary SVM classifiers by each time randomly sampling negative documents from the corpus
- Pseudo label refinement:
  - Remove current pseudo labels whose prediction score is negative
  - Enrich pseudo labels with top-n selected documents



## **Experiments: Quantitative Results**

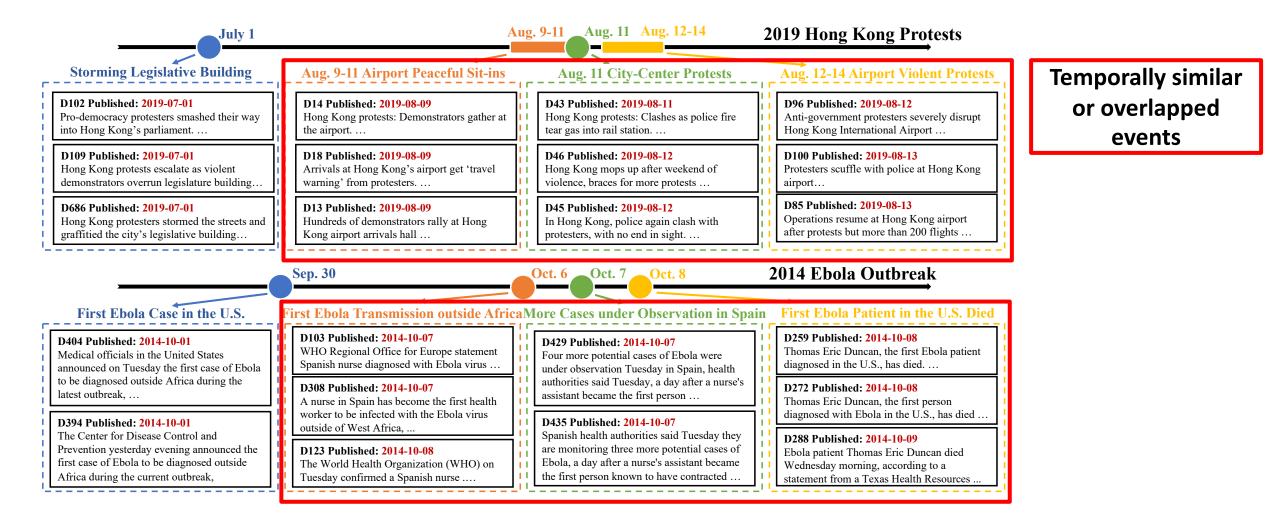
- Datasets:
  - **HK Protest**: We retrieve news articles about the theme "2019 Hong Kong protest".
  - Ebola: We collect from English part of a multilingual news clustering dataset that about the theme "2014 Ebola Outbreak". <u>Table 1: Datasets statistics.</u>

Dataset	# Docs	# Sents/Doc	# Words/Doc	# Events	# Docs/Event
HK Protest	1675	32.8	653.4	36	14.0
Ebola	741	25.2	554.4	17	43.6

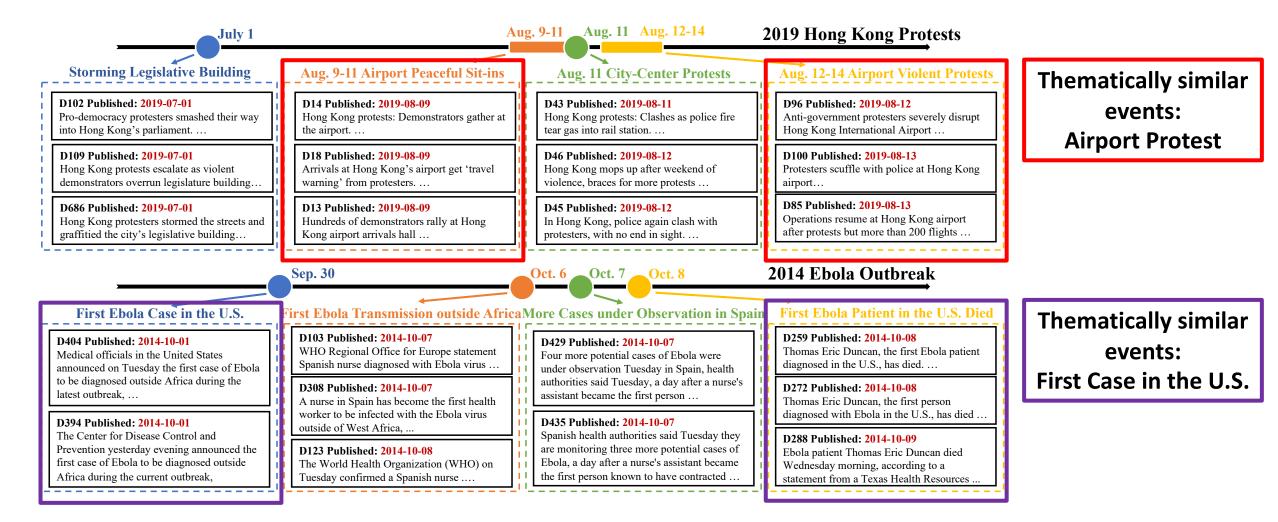
#### Quantitative results: key event detection ability based on top-k documents of each event

Methods	Ebola						HK Protest					
	5-prec	5-recall	5-F1	10-prec	10-recall	10-F1	5-prec	5-recall	5-F1	10-prec	10-recall	10-F1
newsLens [19]	0.481	0.765	0.591	0.524	0.647	0.579	0.352	0.886	0.504	0.571	0.343	0.429
Miranda et al. [21]	0.444	0.706	0.545	0.733	0.647	0.688	0.481	0.371	0.419	0.286	0.057	0.095
Staykovski et al. [35]	0.414	0.706	0.522	0.688	0.647	0.667	0.442	0.657	0.529	0.444	0.114	0.182
S-BERT	0.545	0.706	0.615	0.833	0.588	0.689	0.522	0.657	0.582	0.500	0.257	0.340
EvMine-NoClass	0.799	0.612	0.693	0.764	0.494	0.600	0.750	0.583	0.656	0.750	0.417	0.536
EvMine-COOC	0.846	0.647	0.733	0.909	0.588	0.714	0.815	0.611	0.698	0.807	0.431	0.561
EvMine-NoLM	0.784	0.659	0.715	0.865	0.635	0.732	0.905	0.608	0.728	0.942	0.453	0.612
<b>EvMine-Single</b>	0.814	0.671	0.735	0.872	0.635	0.734	0.916	0.636	0.751	0.958	0.458	0.620
EvMine	0.829	0.682	0.748	0.883	0.653	0.751	0.934	0.664	0.776	0.960	0.464	0.625

## **Experiments: Qualitative Results**



## **Experiments: Qualitative Results**



## 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., Guo, F., Shen, J., & Han, J. (2022). Unsupervised Key Event Detection from Massive Text Corpora. KDD.
- 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.