Part IV: Embedding-Driven Multi-Dimensional Text Analysis

KDD 2020 Tutorial

Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis

Yu Meng, Jiaxin Huang, Jiawei Han

Computer Science, University of Illinois at Urbana-Champaign

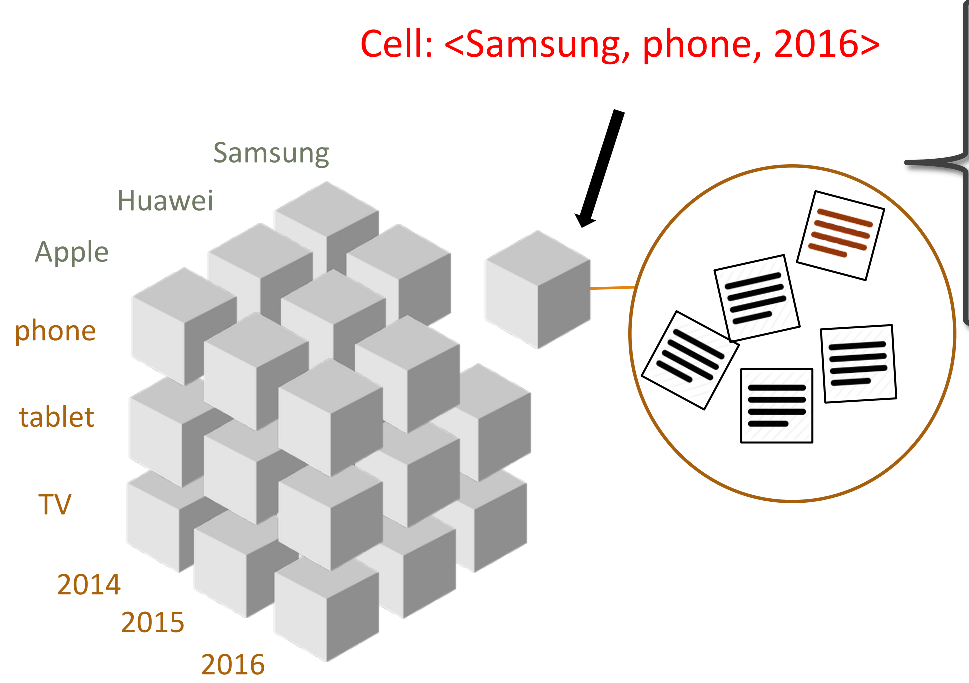
August 23, 2020

Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM'18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM'18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI'19]
  - Incorporating Metadata: MetaCat [SIGIR'20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

Multi-Dimensional Text Cube

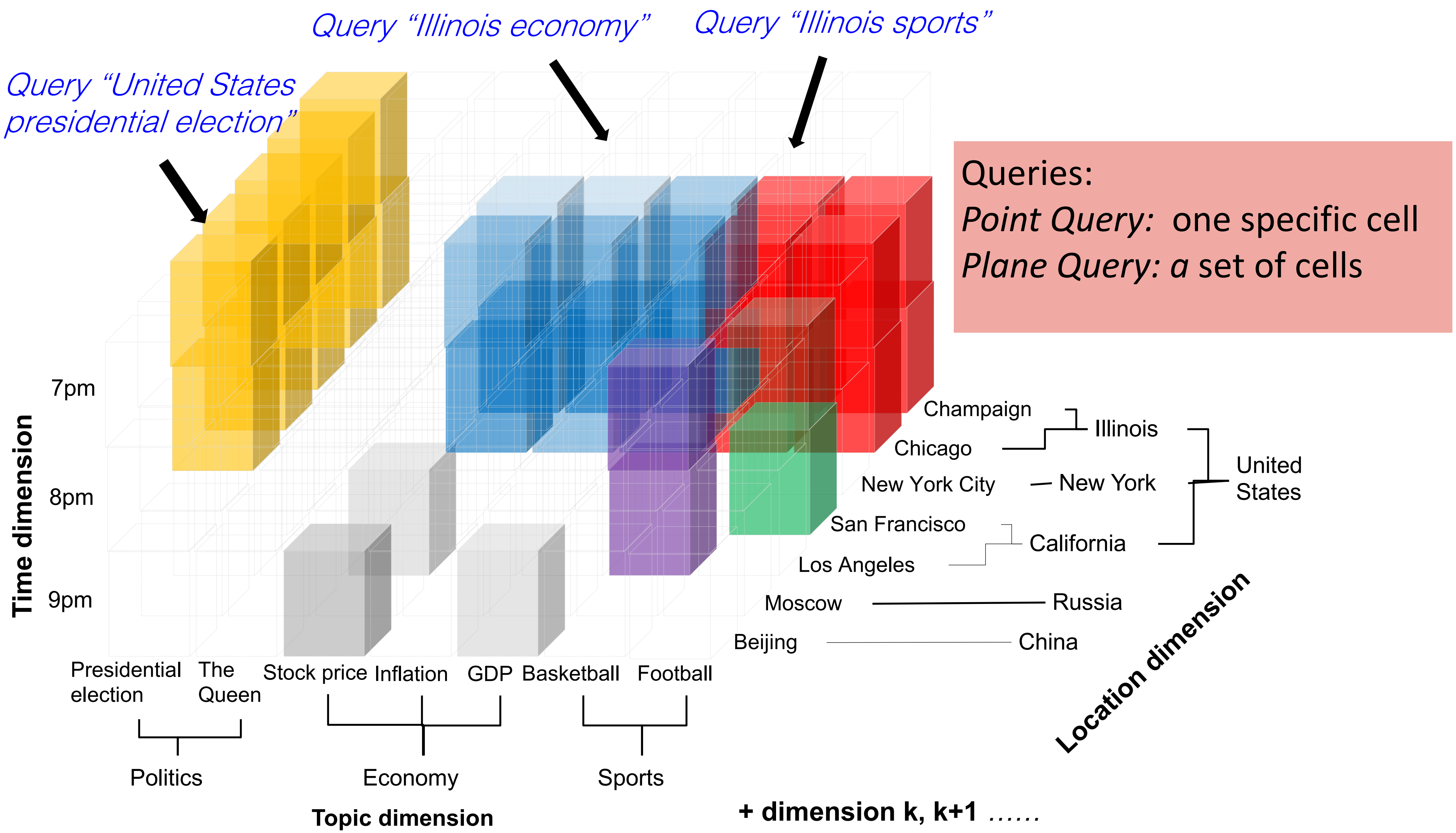
- Numerical data cube (each cell is a numerical value) has been extensively studied
  - Measures: Numerical aggregations as *sum* & *avg*.
- Text cube: Each cell contains a set of documents (e.g., Apple, TV, 2016>)
  - There is an imminent need to do OLAP analysis on text cubes



Dimensions:
Brand: Apple, Samsung, Huawei…
Product: phone, tablet, TV, laptop…
Field: IR, Machine Learning, NLP…

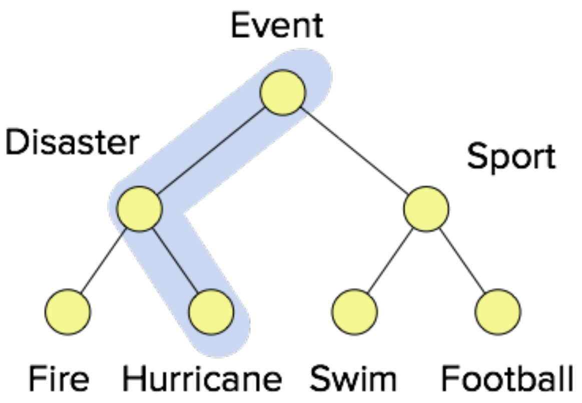
Text:
Aviation Safety Reports from NASA
Product Reviews from Amazon
Research Papers from DBLP

Multi-dimensional Text Cube with Queries & Hierarchies

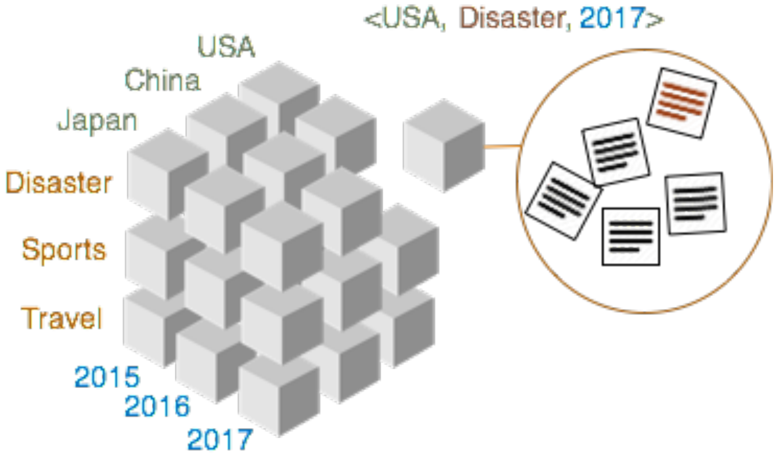


Text Cube Construction: Two Central Tasks

- 1. **Taxonomy Construction**
  - How to discover the taxonomy for each dimension?



- 2. **Document Allocation**
  - How to allocate documents into the cube?

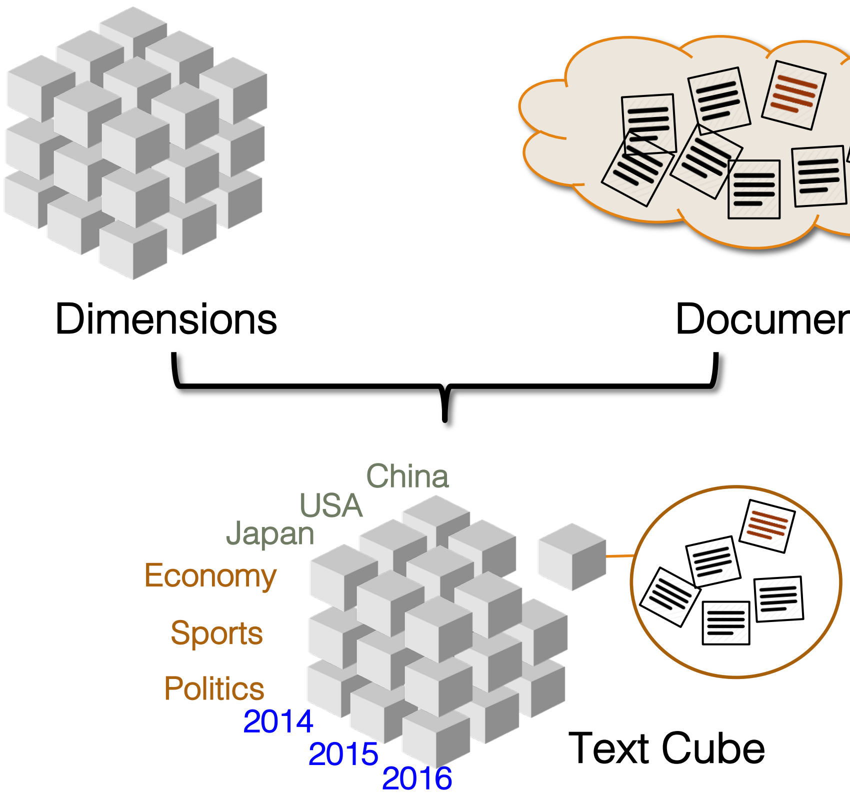


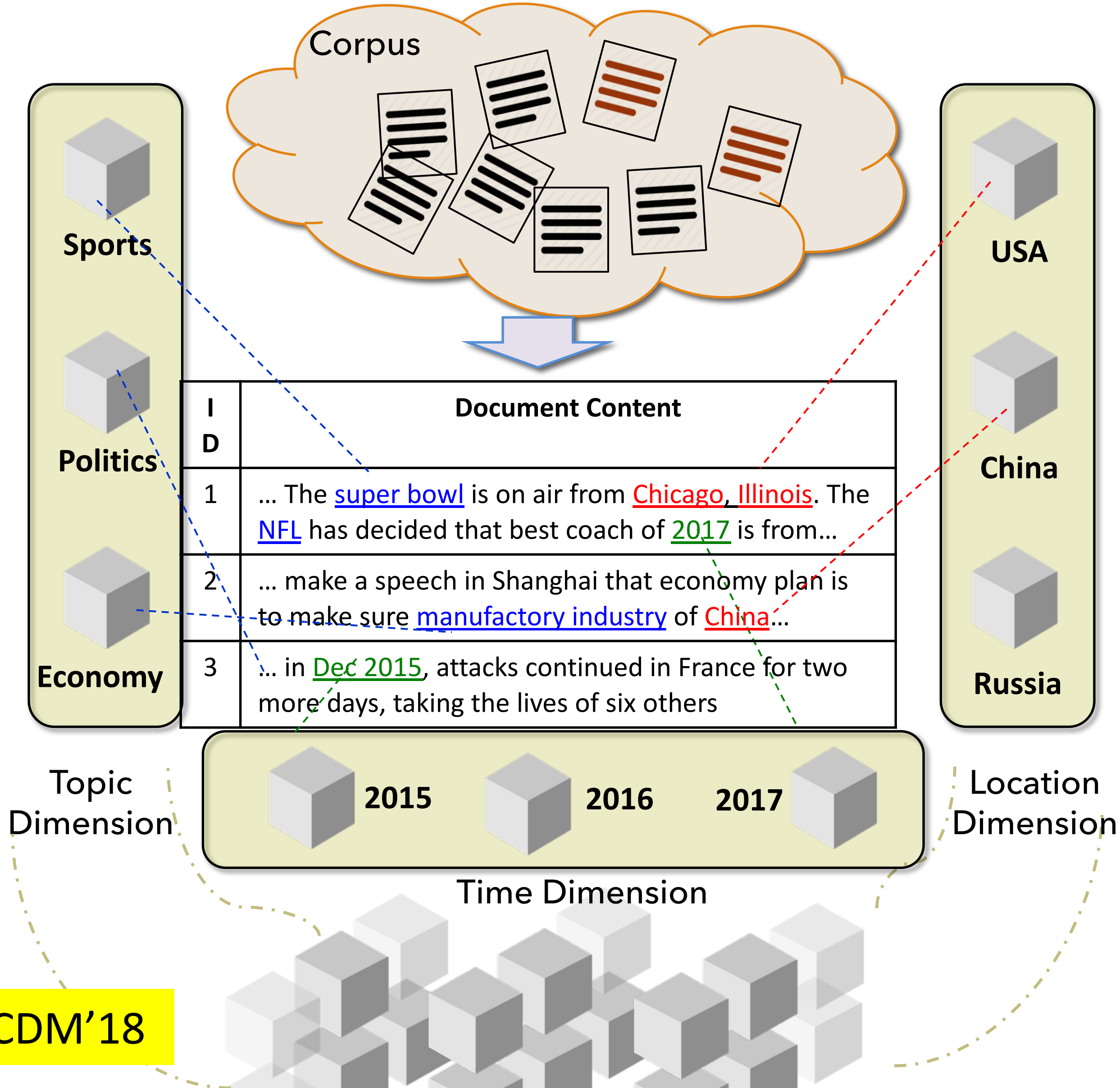
Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM'18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM'18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI'19]
  - Incorporating Metadata: MetaCat [SIGIR'20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

Cube Construction: Which Document Goes to Which Cell?

- Cell-based Document Allocation
  - Which document goes to which cell?

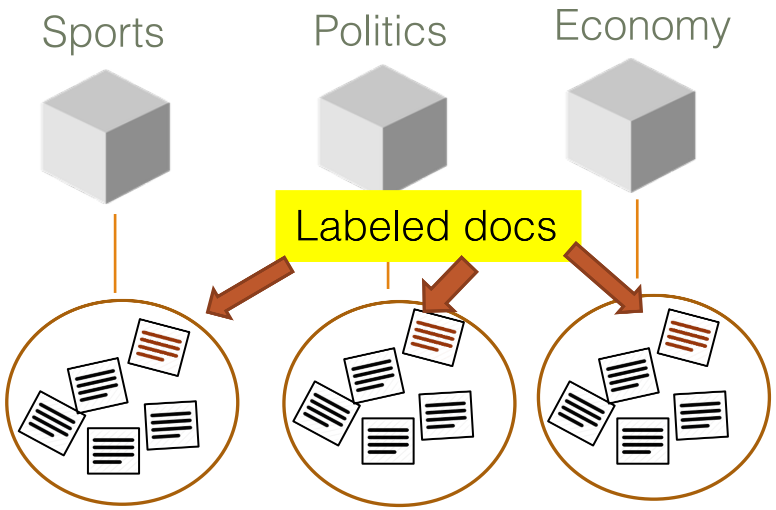


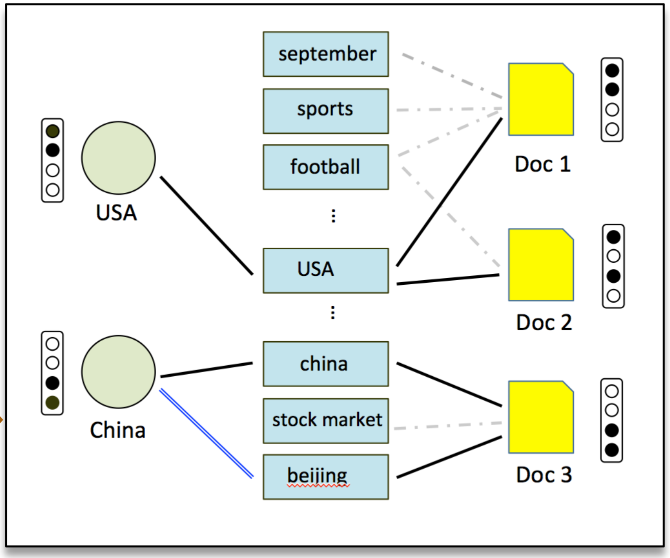


Doc2Cube: Constructing Cube from Massive Docs: ICDM'18

How to Put Documents into the Right Cube Cell?

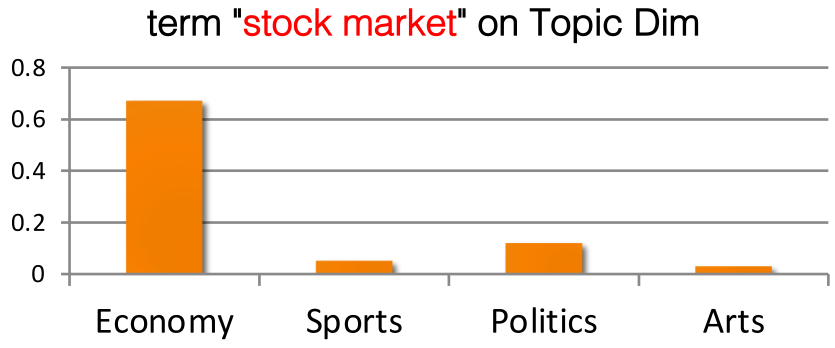
- Major challenges on putting docs into the right cell
  - Few would like label the “training sets”
    - So many cells, so many documents
  - Dimension values are often “under-represented”
    - E.g., Topic dimension: Sports, economy, politics, ….
  - Documents are often “over-represented” on single dimension
    - Ex. “ … … The super bowl is on air from Chicago, Illinois. The NFL has decided that best coach of 2017 is from …
- Our methodology: Dimension-aware joint embedding
  - Constructing an L-T-D (label-term-document) graph

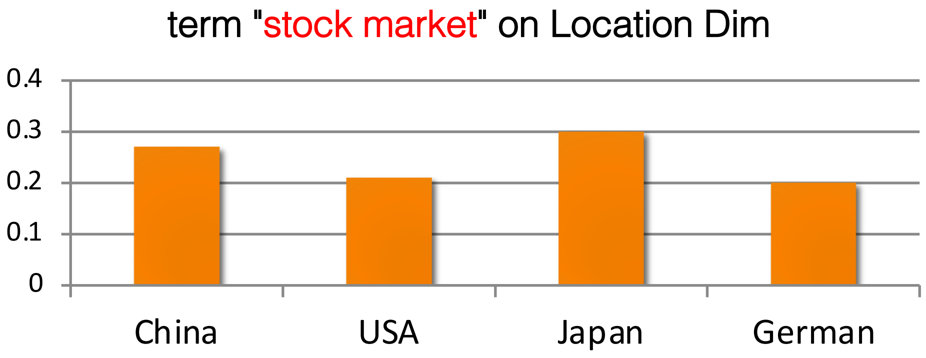




Constructing Text Cubes with Massive Data, Few Labels

- Dimension focusing—**Dimension-Focal Score**, a discriminative measure
  - A term *t* is "focal" to dimension L
    - The documents with t has very imbalanced labels (KL-divergence can be a good measure)
      - Ex.





- Label expansion: Combining two measures for seed expansion
    - Discriminativeness
      - Using focal score
    - Popularity
  - Example:

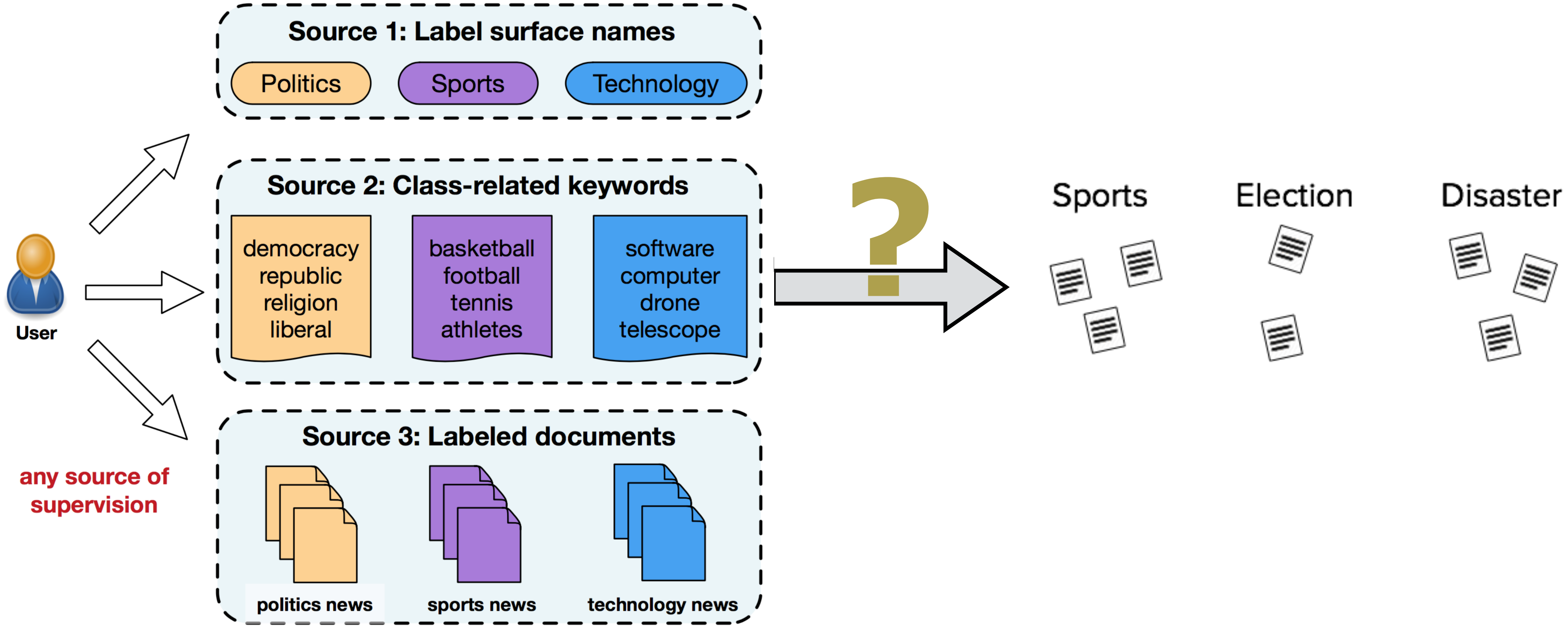
Dimension	Label	1st Expansion	2nd Expansion	3rd Expansion
Topic	*Movies*	films	director	hollywood
	Baseball	inning	hits	pitch
	Tennis	wimbledon	french open	grand slam
	Business	company	chief executive	industry
	Law Enforcement	litigation	law	county courthouse
Location	*Brazil*	brazilian	sao paulo	confederations cup
	Australia	sydney	australian	melbourne
	Spain	madrid	barcelona	la liga
	China	chinese	shanghai	beijing

Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM'18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM'18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI'19]
  - Incorporating Metadata: MetaCat [SIGIR'20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

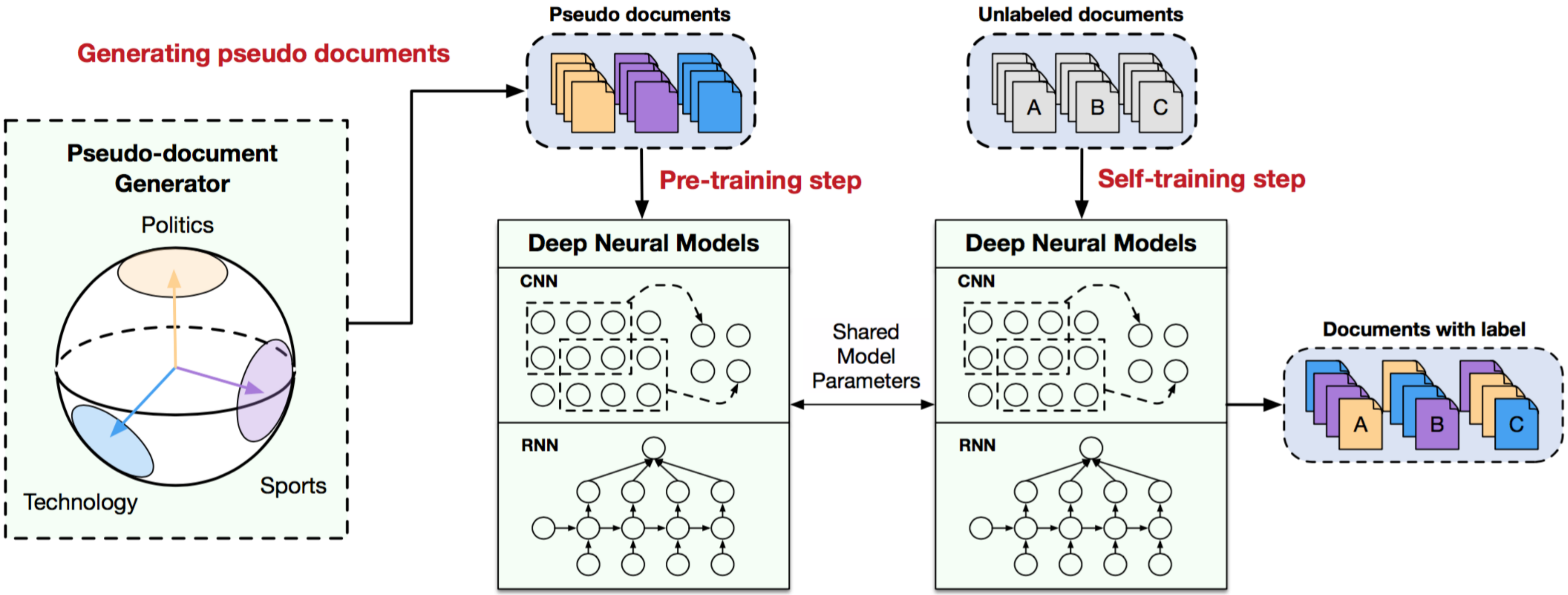
Weakly-Supervised Text Classification

- Require no training data, but a small amount of seed information
  - (1) label names, or (2) relevant keywords, or (3) a few labeled docs



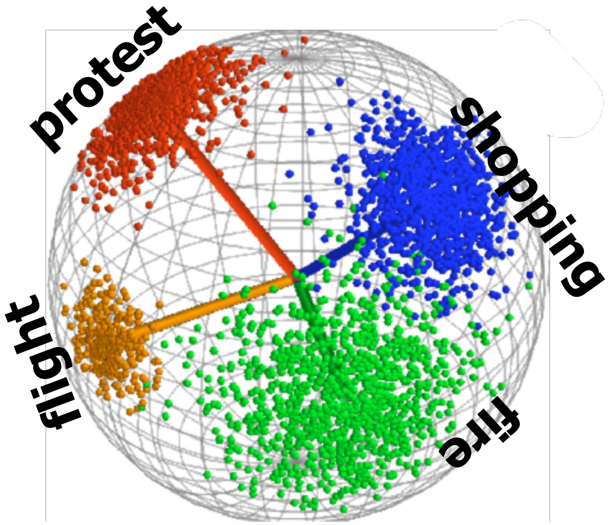
Pseudo Training Data + Self-Training

- Pseudo document generation: generate pseudo documents from seeds.
- Self-training: train deep neural nets (CNN, RNN) with bootstrapping.



Pseudo Document Generation

- Fit a von-Mishes Fisher distribution with the embeddings of seeds.
- Sample bag-of-keywords as pseudo documents for each class.



Mean direction

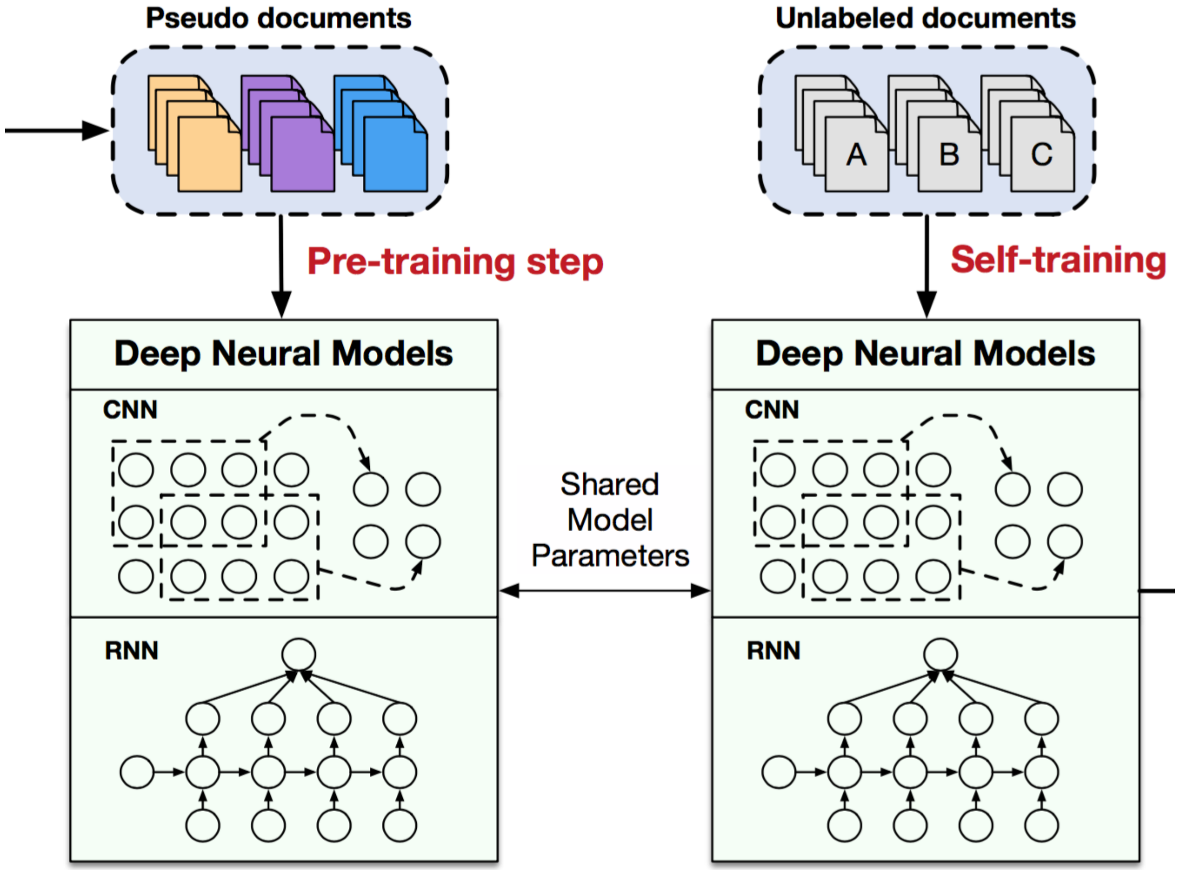
Concentration parameter

$$p(\mathbf{x}|\mu, \kappa) = C_D(\kappa) \exp(\kappa \mu^T \mathbf{x})$$

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

Self-Training Deep Neural Nets

- 1. **Pre-training**: Use pseudo documents to initialize DNNs (e.g., CNN, RNN)
- 2. **Self-training**: Iteratively refine DNNs in a self-boosting fashion.



new score of
label j for document i

$$l_{ij} = \frac{y_{ij}^2 / f_j}{\sum_{j'} y_{ij'}^2 / f_{j'}}$$

Overall Classification Performance

- Datasets: (1) NYT, (2) AG's News, (3) Yelp
- Evaluation: use different types of weak supervision and measure accuracies

Macro-F1 scores:

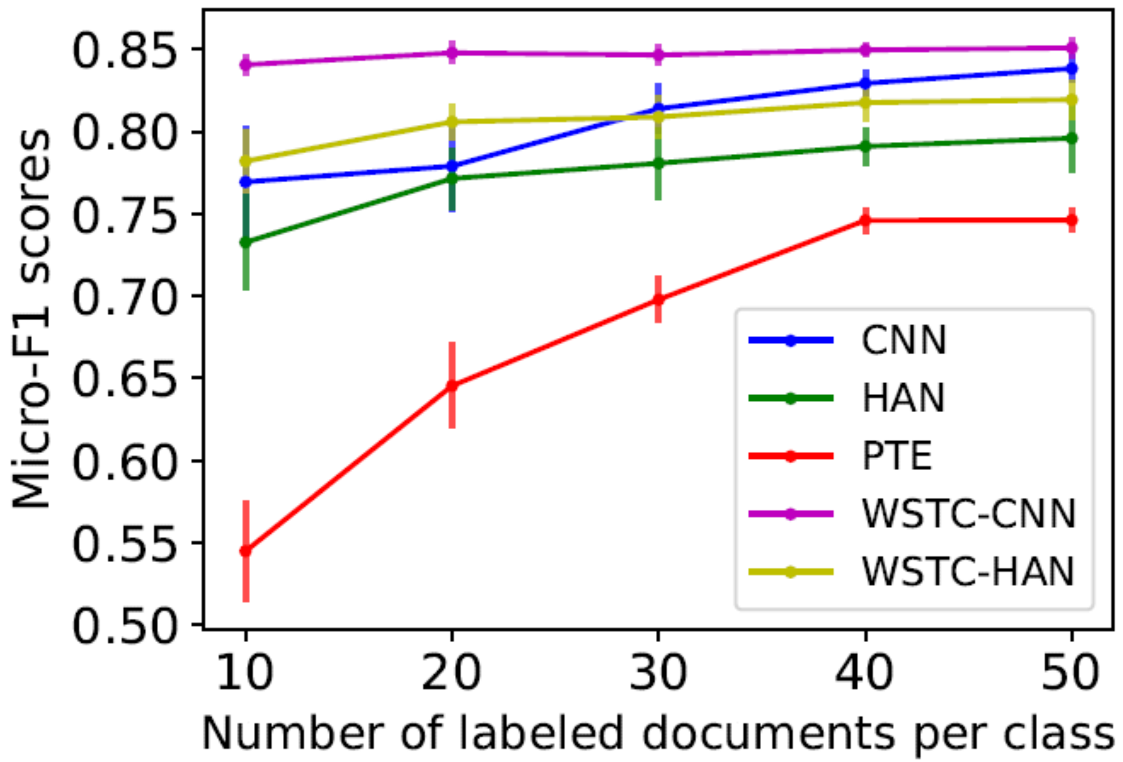
Methods	The New York Times			AG's News			Yelp Review		
	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS
IR with tf-idf	0.319	0.509	-	0.187	0.258	-	0.533	0.638	-
Topic Model	0.301	0.253	-	0.496	0.723	-	0.333	0.333	-
Dataless	0.484	-	-	0.688	-	-	0.337	-	-
UNEC	0.690	-	-	0.659	-	-	0.602	-	-
PTE	-	-	0.834 (0.024)	-	-	0.542 (0.029)	-	-	0.658 (0.042)
HAN	0.348	0.534	0.740 (0.059)	0.498	0.621	0.731 (0.029)	0.519	0.631	0.686 (0.046)
CNN	0.338	0.632	0.702 (0.059)	0.758	0.770	0.766 (0.035)	0.523	0.633	0.634 (0.096)
NoST-HAN	0.515	0.213	0.823 (0.035)	0.590	0.727	0.745 (0.038)	0.731	0.338	0.682 (0.090)
NoST-CNN	0.701	0.702	0.833 (0.013)	0.534	0.759	0.759 (0.032)	0.639	0.740	0.717 (0.058)
WeSTClass-HAN	0.754	0.640	0.832 (0.028)	0.816	0.820	0.782 (0.028)	**0.769**	0.736	0.729 (0.040)
WeSTClass-CNN	**0.830**	**0.837**	**0.835 (0.010)**	**0.822**	**0.821**	**0.839 (0.007)**	0.735	**0.816**	**0.775 (0.037)**

Micro-F1 scores:

Methods	The New York Times			AG's News			Yelp Review		
	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS
IR with tf-idf	0.240	0.346	-	0.292	0.333	-	0.548	0.652	-
Topic Model	0.666	0.623	-	0.584	0.735	-	0.500	0.500	-
Dataless	0.710	-	-	0.699	-	-	0.500	-	-
UNEC	0.810	-	-	0.668	-	-	0.603	-	-
PTE	-	-	0.906 (0.020)	-	-	0.544 (0.031)	-	-	0.674 (0.029)
HAN	0.251	0.595	0.849 (0.038)	0.500	0.619	0.733 (0.029)	0.530	0.643	0.690 (0.042)
CNN	0.246	0.620	0.798 (0.085)	0.759	0.771	0.769 (0.034)	0.534	0.646	0.662 (0.062)
NoST-HAN	0.788	0.676	0.906 (0.021)	0.619	0.736	0.747 (0.037)	0.740	0.502	0.698 (0.066)
NoST-CNN	0.767	0.780	0.908 (0.013)	0.553	0.766	0.765 (0.031)	0.671	0.750	0.725 (0.050)
WeSTClass-HAN	0.901	0.859	0.908 (0.019)	0.816	0.822	0.782 (0.028)	**0.771**	0.737	0.729 (0.040)
WeSTClass-CNN	**0.916**	**0.912**	**0.911 (0.007)**	**0.823**	**0.823**	**0.841 (0.007)**	0.741	**0.816**	**0.776 (0.037)**

Effect of # Labeled Documents

- Compare the performances of five methods on the AG's News dataset by varying the number of labeled documents per class and

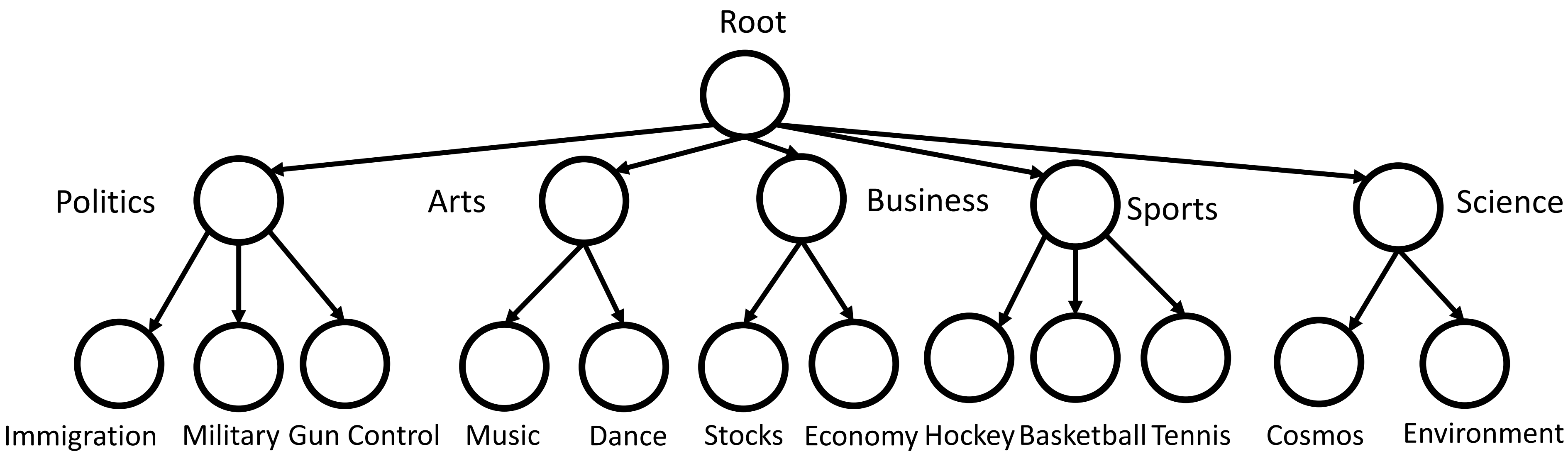


Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM’18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM’18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI’19]
  - Incorporating Metadata: MetaCat [SIGIR’20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

Weakly-Supervised Hierarchical Text Classification

- Class Hierarchy (toy example):



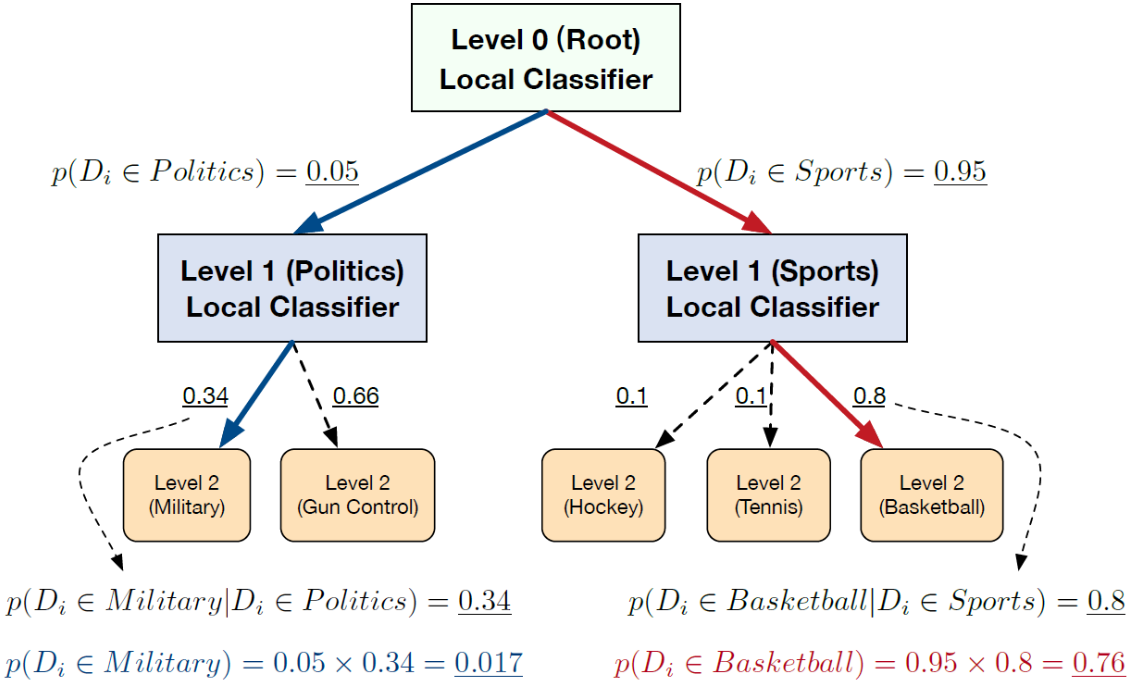
- What if we have a taxonomy and aim to allocate documents into categories in the taxonomy?

Hierarchical Classification Model

- Local Classifier Pre-training
  - We generate β pseudo documents per class to pre-train the local classifier;
  - A naive way of creating the label for a pseudo document D_i^*:
    - Directly use the associated class label it is generated from; one-hot encodings;
    - Problem: classifier overfitting to pseudo documents
  - Instead, use pseudo labels:
    - $l_{ij} = \begin{cases} (1-\alpha) + \alpha/m & D_i^* \text{ is generated from class } j \\ \alpha/m & \text{otherwise} \end{cases}.$
    - α accounts for the "noises" in pseudo documents; it is evenly split into all m classes
  - Pre-training is performed by minimizing KL divergence loss to pseudo labels

Hierarchical Classification Model

- Global Classifier Per Level
  - At each level k in the class taxonomy, we construct a global classifier by ensembling all local classifiers from root to level k
  - Use unlabeled documents to bootstrap the global classifier



Hierarchical Classification Model

- Global Classifier Construction
  - The multiplication operation can be explained by the conditional probability formula:

$$p(D_i \in C_{child}) = p(D_i \in C_{child} \mid D_i \in C_{parent})p(D_i \in C_{parent})$$

  - All local classifiers from root to to level k are fine-tuned simultaneously via back-propagation during self-training; misclassifications at higher levels can be corrected
- Global Classifier Self-training
  - Step 1: Use the pre-trained global classifier to classify all unlabeled documents in the corpus;
  - Step 2: Compute pseudo labels based on current predictions:

$l_{ij} = \frac{y_{ij}^2/f_j}{\sum_{j'} y_{ij'}^2/f_{j'}}$ where $f_j = \sum_i y_{ij}$ and y_{ij} is the current prediction.

  - Step 3: Minimize KL divergence loss to pseudo labels.
  - Iterate between Steps 2 and 3 until less than $\delta\%$ of documents in the corpus have class assignment changes

Hierarchical Classification Model

- Blocking Mechanism
  - Some documents should be classified into internal classes because they are more related to general topics rather than specific topics;
  - When a document D_i is classified into an internal class C_j, we use the output q of C_j's local classifier to determine whether or not D_i should be blocked at the current class:
    - If q is close to a one-hot vector, D_i should be classified into the corresponding child;
    - If q is close to uniform distribution, D_i should be blocked at current class;
    - Use normalized entropy as measure for blocking, i.e. block D_i if

$$-\frac{1}{\log m}\sum_{i=1}^{m} q_i \log q_i > \gamma$$

Overall Classification Performance

- Datasets:
  - New York Times; arXiv; Yelp Review
- Evaluation: Micro-F1 and Macro-F1 among all classes

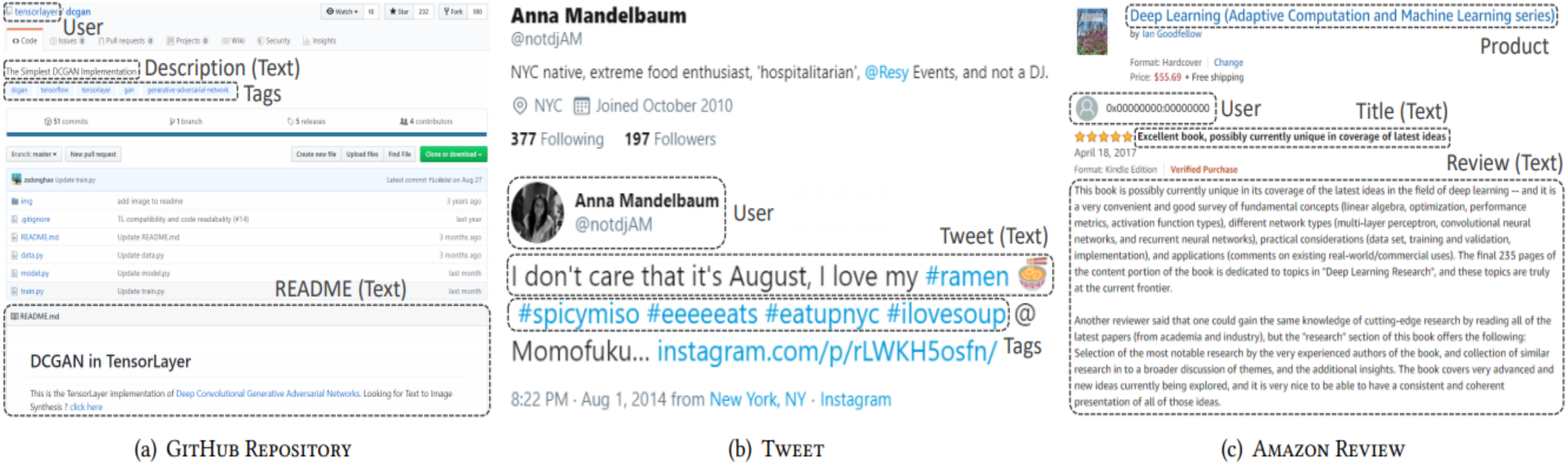
Methods	NYT				arXiv				Yelp Review			
	KEYWORDS		DOCS		KEYWORDS		DOCS		KEYWORDS		DOCS	
	Macro	Micro	Macro Avg. (Std.)	Micro Avg. (Std.)	Macro	Micro	Macro Avg. (Std.)	Micro Avg. (Std.)	Macro	Micro	Macro Avg. (Std.)	Micro Avg. (Std.)
Hier-Dataless	0.593	0.811	-	-	0.374	0.594	-	-	0.284	0.312	-	-
Hier-SVM	-	-	0.142 (0.016)	0.469 (**0.012**)	-	-	0.049 (**0.001**)	0.443 (**0.006**)	-	-	0.220 (0.082)	0.310 (0.113)
CNN	-	-	0.165 (0.027)	0.329 (0.097)	-	-	0.124 (0.014)	0.456 (0.023)	-	-	0.306 (0.028)	0.372 (0.028)
WeSTClass	0.386	0.772	0.479 (0.027)	0.728 (0.036)	0.412	0.642	0.264 (0.016)	0.547 (0.009)	0.348	0.389	0.345 (0.027)	0.388 (0.033)
No-global	0.618	0.843	0.520 (0.065)	0.768 (0.100)	0.442	0.673	0.264 (0.020)	0.581 (0.017)	0.391	0.424	0.369 (0.022)	0.403 (0.016)
No-vMF	0.628	0.862	0.527 (0.031)	0.825 (0.032)	0.406	0.665	0.255 (0.015)	0.564 (0.012)	0.410	0.457	0.372 (0.029)	0.407 (0.015)
No-self-train	0.550	0.787	0.491 (0.036)	0.769 (0.039)	0.395	0.635	0.234 (0.013)	0.535 (0.010)	0.362	0.408	0.348 (0.030)	0.382 (0.022)
Our method	**0.632**	**0.874**	**0.532** (0.015)	**0.827** (0.012)	**0.452**	**0.692**	**0.279** (0.010)	**0.585** (0.009)	**0.423**	**0.461**	**0.375** (0.021)	**0.410** (0.014)

Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM’18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM’18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI’19]
  - Incorporating Metadata: MetaCat [SIGIR’20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

MetaCat: Incorporating Metadata for Categorization

- Metadata is prevalent in many text sources, especially social media platforms
  - GitHub Repositories: User, Tags; Tweets: User, Hashtags; Amazon Reviews: User, Product
- How to leverage these heterogenous signals in the categorization process?



(a) GitHub Repository (b) Tweet (c) Amazon Review

Figure 1: Three examples of documents with metadata.

The Underlying Model: A Generative Process

- Two categories of metadata:
  - **Global metadata**: user/author, product
    - “Causes” the generation of documents. (E.g., User -> Document)
  - **Local metadata**: tag/hashtag
    - “Describes” the documents. (E.g., Document -> Tag)
  - We can also say “label” is global, and “words” are local

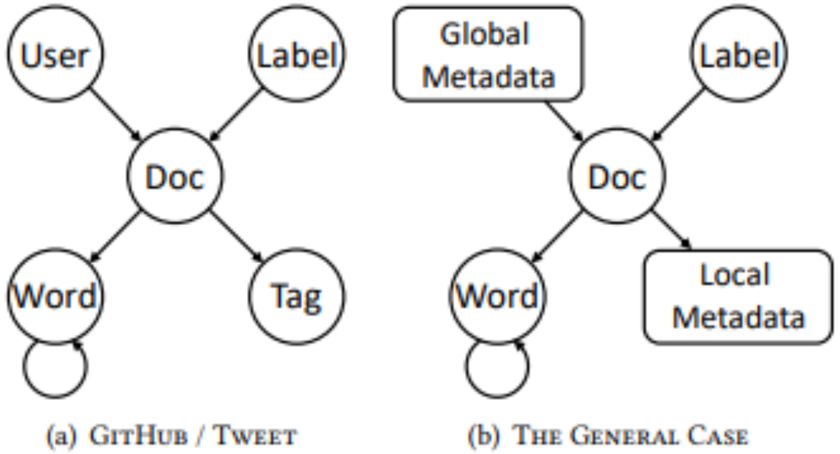


Figure 2: The generative process of text and metadata. The self loop of “Word” represents the step of words generating contexts.

The underlying model: A generative process

- We use GitHub/Tweet as a specific example to illustrate the process.

- **Step 1: User (Global Metadata) & Label -> Document**

$$p(d|u,l) \propto \exp(\boldsymbol{e}_d^T \boldsymbol{e}_u) \cdot \exp(\boldsymbol{e}_d^T \boldsymbol{e}_l).$$

- **Step 2: Document -> Word**

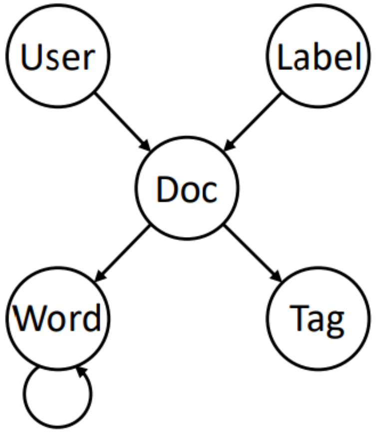
$$p(w|d) \propto \exp(\boldsymbol{e}_w^T \boldsymbol{e}_d).$$

- **Step 3: Document -> Tag (Local Metadata)**

$$p(t|d) \propto \exp(\boldsymbol{e}_t^T \boldsymbol{e}_d).$$

- **Step 4: Word -> Context**

$$p(C(w_i,h)|w_i) \propto \prod_{w_j \in C(w_i,h)} \exp(\boldsymbol{e'}_{w_j}^T \boldsymbol{e}_{w_i}).$$

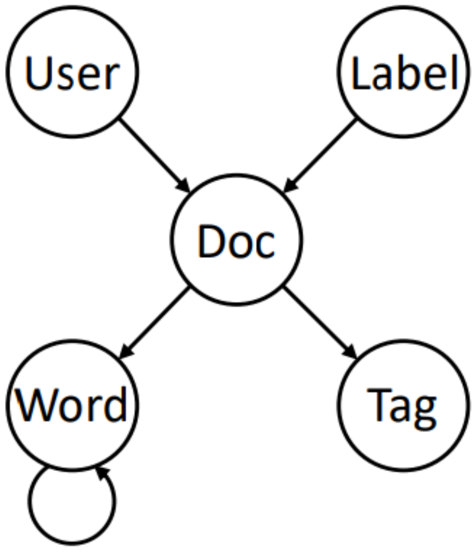


(a) GitHub / Tweet

How do we use this underlying model?

- **Embedding** Learning Module:
  - All embedding vectors e_u, e_l, e_d, e_t, e_w are parameters of the generative process.
  - We can learn the embedding vectors through maximizing the likelihood of observing all text and metadata.

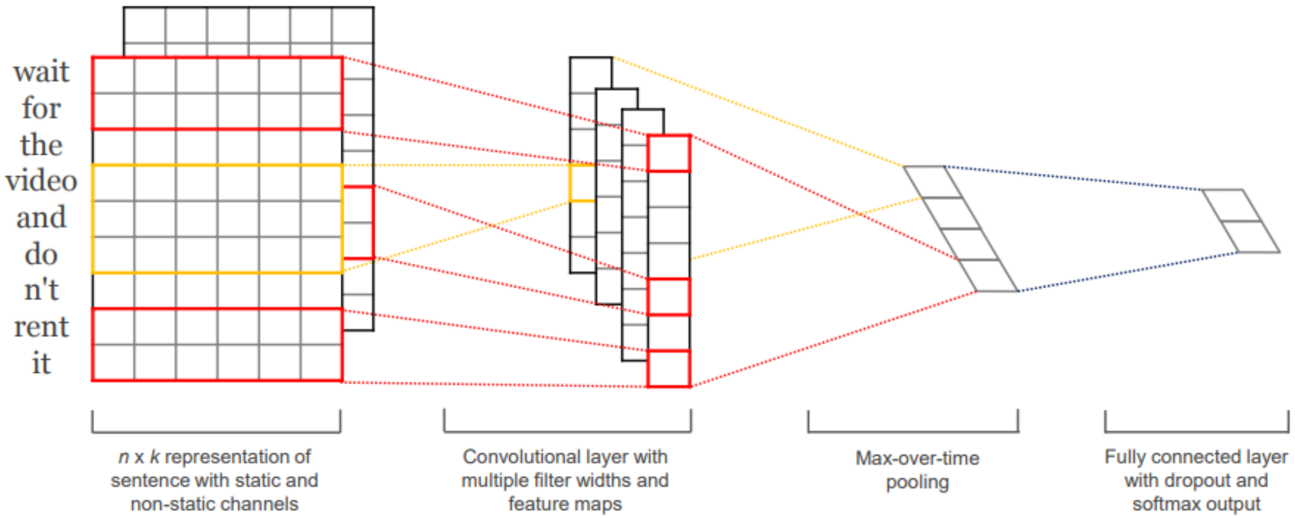
- Training Data **Generation** Module:
  - We have learned e_u, e_l, e_d, e_t, e_w .
  - Given a label l, we can generate d, w and t according to the generative process.



(a) GitHub / Tweet

Train a text classifier

- After the embedding and generation steps, what do we have?
  - A set of word embeddings which considers label and metadata information
  - For each category, we have a small set of "real" training data and a large set of synthesized training data
- Using both "real" and synthesized training data to train a text classifier; taking the pre-trained embeddings as input features
- We use CNN as the text classifier; may be replaced by other architectures



Outline

- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM'18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM'18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI'19]
  - Incorporating Metadata: MetaCat [SIGIR'20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

Language Models for Weakly-Supervised Classification

- The previous approaches only use the local corpus
- Fail to take advantage of the general knowledge source (e.g. Wikipedia)
- Why general knowledge?
  - Humans can classify texts with general knowledge
  - Common linguistic features to understand texts better
  - Compensate for potential data scarcity of the local corpus
- How to use general knowledge?
  - Neural language models (e.g. BERT) are pre-trained on large-scale general knowledge texts
  - Their learned semantic/syntactic features can be transferred to downstream tasks

Find Similar Meaning Words with Label Names

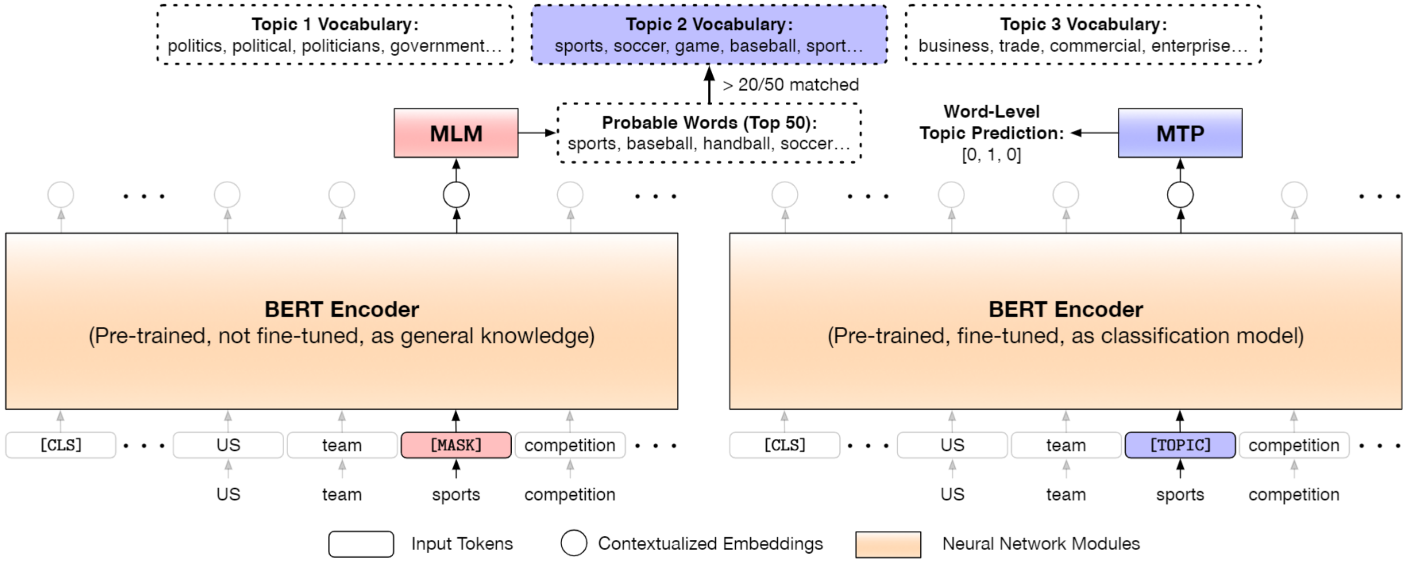
- Find topic words based on label names
  - Overcome the low semantic coverage of label names
- Use language models to predict what words can replace the label names
  - Interchangeable words are likely to have similar meanings

Sentence	Language Model Prediction
The oldest annual US team **sports** competition that includes professionals is not in baseball, or football or basketball or hockey. It's in soccer.	sports, baseball, handball, soccer, basketball, football, tennis, sport, championship, hockey, ...
Samsung's new SPH-V5400 mobile phone **sports** a built-in 1-inch, 1.5-gigabyte hard disk that can store about 15 times more data than conventional handsets, Samsung said.	has, with, features, uses, includes, had, is, contains, featured, have, incorporates, requires, offers, ...

Table 1: BERT language model prediction (sorted by probability) for the word to appear at the position of "sports" under different contexts. The two sentences are from *AG News* corpus.

Contextualized Word-Level Topic Prediction

- Context-free matching of topic words is inaccurate
  - "Sports" does not always imply the topic "sports"
- Contextualized topic prediction:
  - Predict a word's implied topic under specific contexts
  - We regard a word as "topic indicative" only when its top replacing words have enough overlap with the topic vocabulary



High-Quality Weakly-Supervised Classification

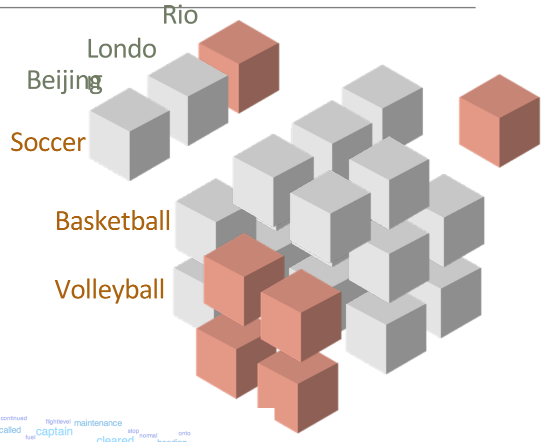
- Achieve around 90% accuracy on four benchmark datasets by only using at most 3 words (1 in most cases) per class as the label name
  - Outperforming previous weakly-supervised approaches significantly
  - Comparable to state-of-the-art semi-supervised models

Supervision Type	Methods	AG News	DBPedia	IMDB	Amazon
Weakly-Sup.	Dataless (Chang et al., 2008)	0.696	0.634	0.505	0.501
	WeSTClass (Meng et al., 2018)	0.823	0.811	0.774	0.753
	BERT w. simple match	0.752	0.722	0.677	0.654
	Ours w/o. self train	0.822	0.850	0.844	0.781
	Ours	**0.864**	**0.889**	**0.894**	**0.906**
Semi-Sup.	UDA (Xie et al., 2019)	0.869	0.986	0.887	0.960
Supervised	char-CNN (Zhang et al., 2015)	0.872	0.983	0.853	0.945
	BERT (Devlin et al., 2019)	0.944	0.993	0.937	0.972

Outline

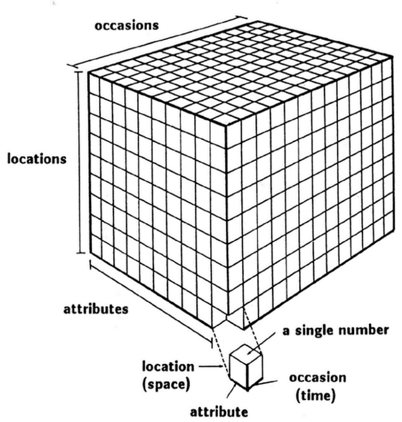
- Why Multi-Dimensional Text Analysis?
- Automatic Document Allocation for Text Cube construction
  - Weakly-Supervised Embedding-Based Classification: Doc2Cube [ICDM'18]
  - Weak-Supervised Neural Text Classification: WeSTClass [CIKM'18]
  - Weakly-Supervised Hierarchical Document Classification: WeSHClass [AAAI'19]
  - Incorporating Metadata: MetaCat [SIGIR'20]
  - Using Neural Language Models for Weakly-Supervised Classification
- Cube-based Multidimensional Analysis

Exploration of Text Cube—Semantic Analysis

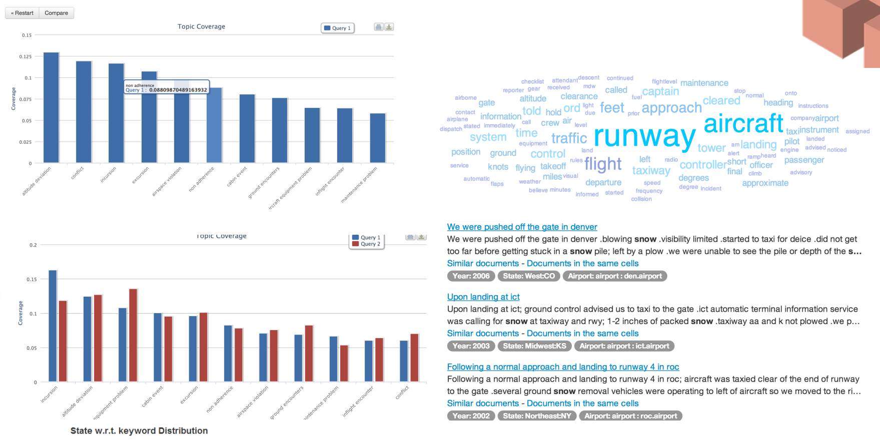


- EventCube [KDD’13 demo]: Point Query
  - Simple summary to support keyword/document search
- CASeOLAP [EngBul’16]: Plane Query
  - Comparative summary/mining

Cube Structure

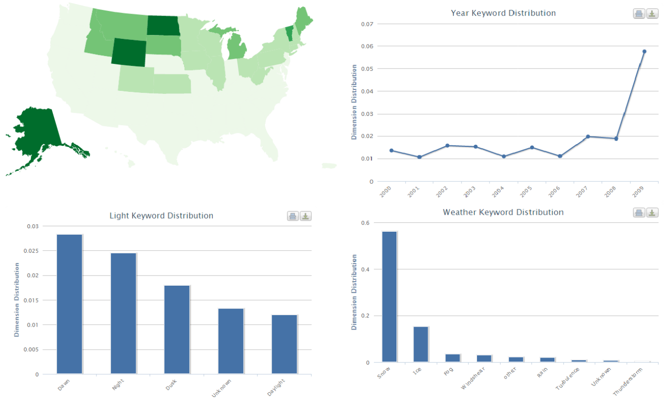


Slice
Roll-up
Drill-down
Dice
…



Textual Analysis

Text Data



Structural Analysis

EventCube [KDD'13 demo]

- Multiple functions supported by EventCube (on Avi. Safety Report System DataSet)

Similar Document Finding: based on Contextual Search

Keyword Frequency Distribution

Top 20 similar documents

Where we have learned that at a particular point in the service we will have to break for t...

All or most passenger were seated .flight attendants were gathered at the back of the aircraft preparing service carts for movement into the aisles .slight bumps occurred; whic...

Similar documents

Year: 2003 | Weather: Unknown | State: South:TX

Of particular concern is that a flight attendant chose to ignore the turbulence and was inj...

Prior to flt; passenger notified to expect some turbulence enroute .when descent commenced weather radar turned on due to thunderstorm in ohio valley .prior to fl180; a...

Similar documents

Year: 2001 | Weather: Thunderstorm | State: Unknown

Cabin crew picked up remaining service items when rough turbulence started approxima...

Captain called back to cabin to inform flight attendants it would be turbulent descending into dfw; so i suggested we go ahead and prepare cabin for landing .we had been experi...

Similar documents

Year: 2004 | Weather: Thunderstorm | State: South:TX

A flight attendant told one of our 2 jumpseating captains that a seat was now available in...

At departure time; my cabin crew advised me of an open seat .i sent one of my 2 jumpseat riders back .after pushback; i was told by my cabin crew that a deadheading flight attend...

Similar documents

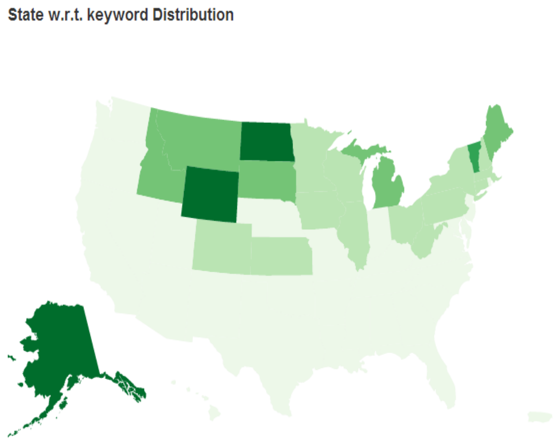
Year: 2003 | Weather: Unknown | State: South:TX

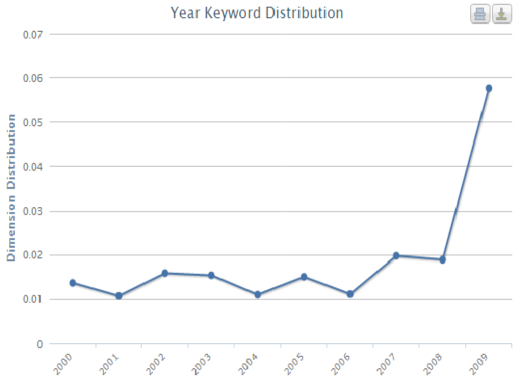
Chosen document

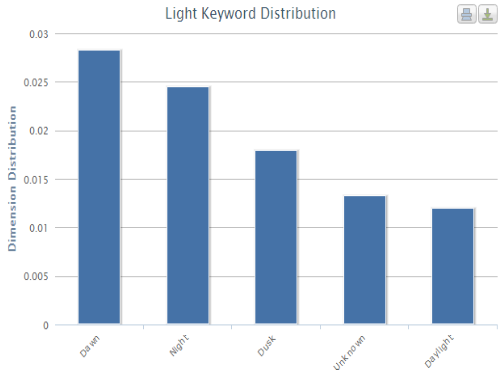
No title

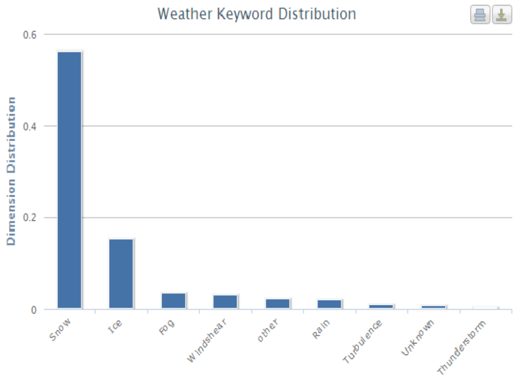
All or most passenger were seated .flight attendants were gathered at the back of the aircraft preparing service carts for movement into the aisles .slight bumps occurred; whic...

Event Anomaly: cabin event:other | Weather: Unknown | Year: 2003 | State: South:TX | Airport: atc facility : zhu.artcc | Make Model: Boeing:B767-300 and 300 ER | Resolutory Action: none taken : insufficient time | Detector: flight attendant : on duty | Light: Daylight | Problem Area: Weather | Flight Phase: cruise : level



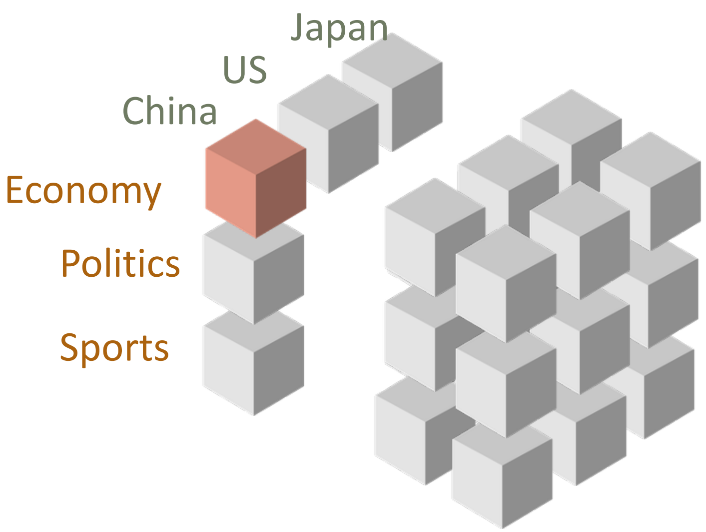




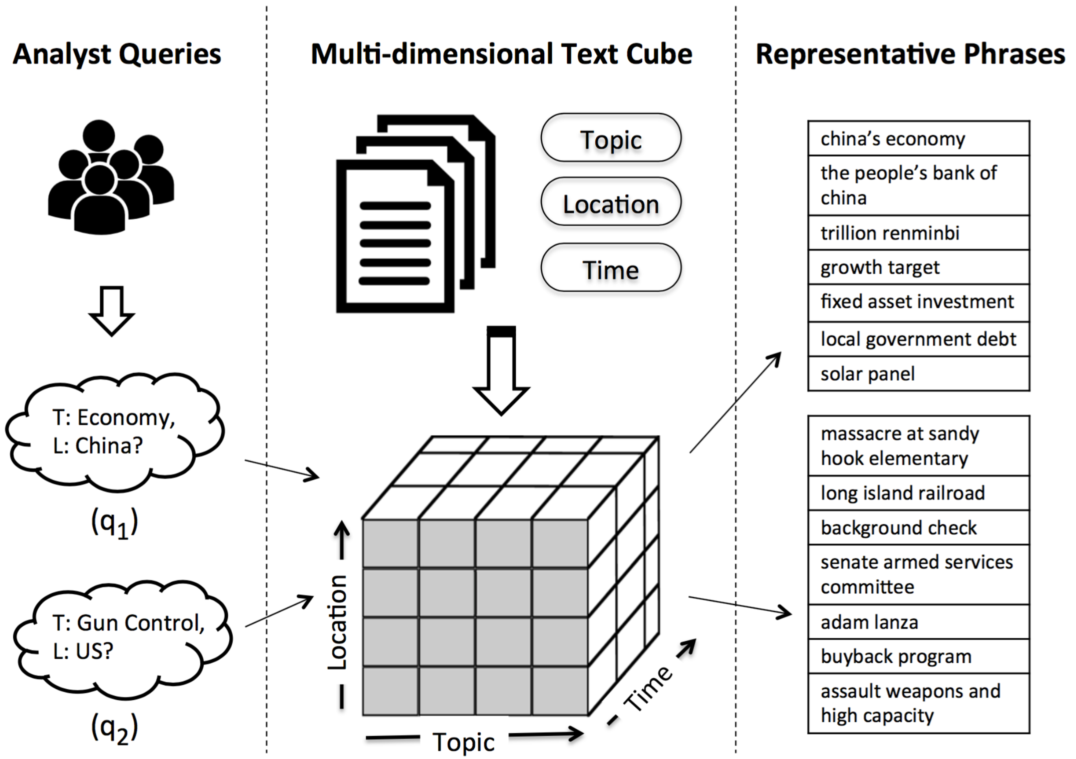


CASE (Context-Aware SEmantic) OLAP

- A cell has comparative context
- Comparative study is meaningful
  - Given a query <**China, Economy**>
  - Target documents have frequent phrases
  - Be specific to "China"+"Economy"



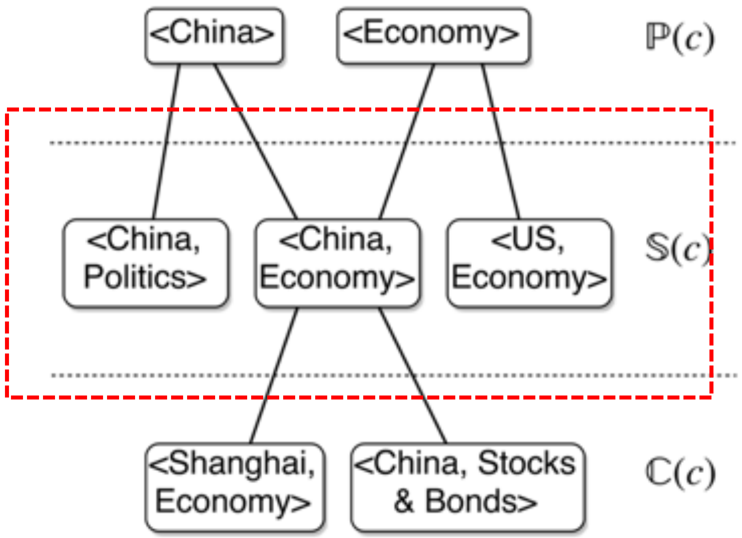
hong kong
united states
prime minister
double digit
communist party
economic growth
the united states
retail sales
G.D.P
monetary policy

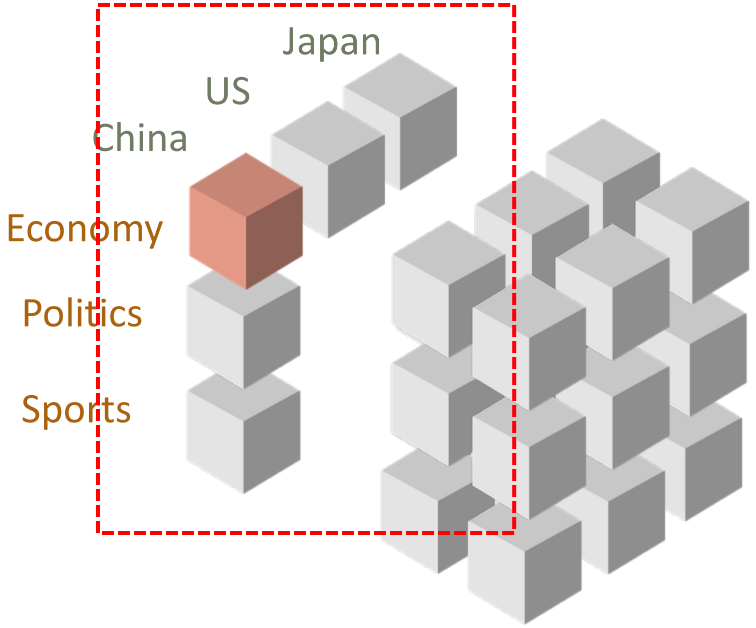


Fangbo Tao, Honglei Zhuang, Chi Wang Yu, Qi Wang, Taylor Cassidy, Lance Kaplan, Clare Voss, Jiawei Han, "Multi-Dimensional, Phrase-Based Summarization in Text Cubes", Data Eng. Bull. 39(3), Sept. 2016

Design Question I: Which Comparative Groups to Pick?

- Option 1: User-specified (too much burden to users): undesirable
- Option 2: **Sibling cells** in every dimension (comparable cells)





Design Question II: How to Score Important Phrases?

- Three ingredients
  - Integrity: **meaningful, high-quality phrase**
    - Using SegPhrase as score (>0.7)
  - Popularity: large # of occurrences in the cell

$$pop(p, c) = \frac{\log(tf(p, c) + 1)}{\log cntP(c)} \qquad (2)$$

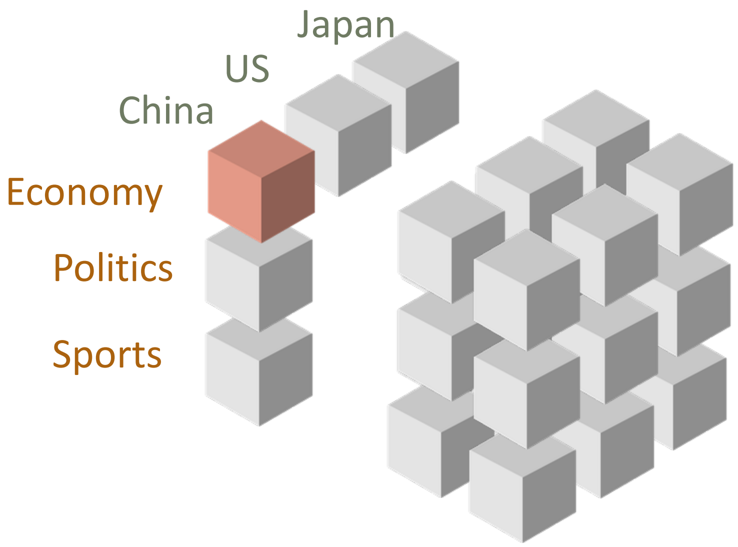
  - Distinctness: distinguish the target cell from context cells
    - A key to have a crisp definition
- Combining with geometric mean:

$$r(p, c) = \sqrt[3]{int(p, c) \cdot pop(p, c) \cdot disti(p, c)} \qquad (1)$$

How to Find or Evaluate Distinct Phrases in a Cell?

- Judge if a phrase p is distinct in cell c: Transform it into a dual problem
  - *Original problem: Find distinctive phrases for cell c, compared to sibling cells*
  - *Transformed problem: Classify phrases into one of the most relevant cell*
- For a distinct phrase p, if we measure relevance(p, c) for all c
  - rel(p, c*) >> rel(p, sibling)
- Adopt Softmax function as

$$disti(p,c) = \frac{e^{rel(p,c)}}{\underbrace{1 +}_{\text{Smoothing}} \underbrace{\sum_{c' \in \mathbb{S} \bigcup \{c\}, p \in c'} e^{rel(p,c')}}_{\text{Sibling relevance}}} \qquad (4)$$



How to Design Relevance Score for a Phrase to a Cell?

- Normalized Term Frequency
  - Treat each cell as a **super document**
  - Apply **BM25**

$$ntf(p,c) = \frac{tf(p,c)\cdot(k_1+1)}{tf(p,c)+k_1\cdot\left(1-b+b\cdot\boxed{\frac{cntP(c)}{avgCP(c)}}\right)} \quad (5)$$

Balance cell size

- Normalized Document Frequency

$$ndf(p,c) = \frac{\log(1+\boxed{df(p,c)})}{\log(1+maxDF(c))} \quad (6)$$

Guarantee spread out!

- Combine:

$$rel(p,c) = ndf(p,c)\cdot ntf(p,c) \quad (7)$$

CaseOLAP on Real-World Datasets

Distinct phrases on 2016 news data

Top-10 representative phrases for five example queries

⟨US, Gun Control⟩	⟨US, Immigration⟩	⟨US, Domestic Politics⟩	⟨US, Law and Crime⟩	⟨US, Military⟩
gun laws	immigration debate	gun laws	district attorney	sexual assault in the military
the national rifle association	border security	insurance plans	shot and killed	military prosecutors
gun rights	guest worker program	background check	federal court	armed services committee
background check	immigration legislation	health coverage	life in prison	armed forces
gun owners	undocumented immigrants	tax increases	death row	defense secretary
assault weapons ban	overhaul of the nation's immigration laws	the national rifle association	grand jury	military personnel
mass shootings	legal status	assault weapons ban	department of justice	sexually assaulted
high capacity magazines	path to citizenship	immigration debate	child abuse	fort meade
gun legislation	immigration status	the federal exchange	plea deal	private manning
gun control advocates	immigration reform	medicaid program	second degree murder	pentagon officials

PubMed Abstracts: Distinct relationships between subcategories of cardiovascular diseases and proteins

Table 2: Top representative phrases for 6 cardiac diseases

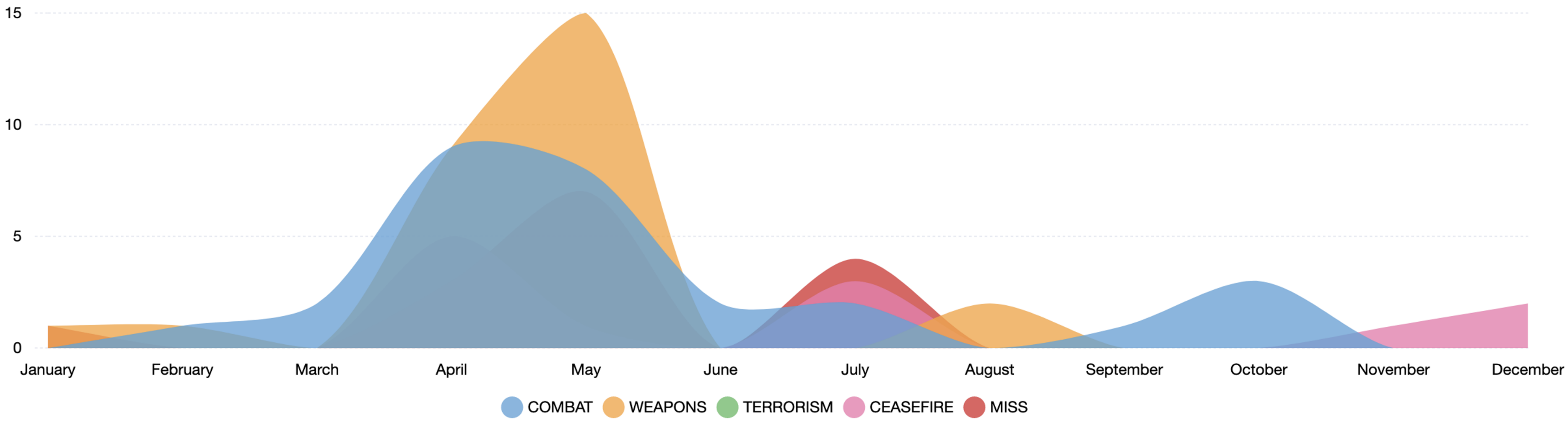
⟨Cerebrovascular Accident⟩	⟨Ischemic Heart Disease⟩	⟨Cardiomyopathy⟩	⟨Arrhythmia⟩	⟨Valve Dysfunction⟩	⟨Congenital Heart Disease⟩
alpha-galactosidase a	Cholesteryl ester transfer protein	Interferon gamma	Methionine synthase	Mineralocorticoid receptor	fibrillin-1
brain neurotrophic factor	apolipoprotein a-I	interleukin-4	ryanodine receptor 2	tropomyosin alpha-1 chain	plakophilin-2
tissue-type activator	integrin alpha-iib	interleukin-17a	potassium v.g. h member 2	elastin	tyrosine-protein type 11
apolipoprotein e	adiponectin	titin	inward rectifier channel 2	beta-2-glycoprotein 1	arachidonate 5-l-a protein
neurogenic l.n.h.p. 3	p2y purinoceptor 12	tumor necrosis factor	beta-2-glycoprotein 1	myosin-binding protein c	catechol o-methyltransferase

MissionCube: Ukraine-Russia Crisis & Hong Kong Demonstration Analysis @ NSCTA Demo 2019

- Demo Scenario: Ukraine-Russia Crisis & Hong Kong Demonstration
  - Data Source: News text data and images crawled from multiple news agencies
  - Dimensions included in the Cube
    - Time, location, and topic mentioned in the news

LINE CHART VISUALIZATION (KIEV CITY, 2014, MILITARY)

THE LINE CHART REPRESENTS THE DISTRIBUTION OF NEWS OF DIFFERENT CATEGORIES IN DIFFERENT MONTHS.



Images, Top-K Keywords and Summary

Cube Demo   Time: 2014-07   Category: infrastructure   PROVINCE NAME   UPDATE   CURRENT: CHERKASY

IMAGE & TOP-K KEYWORDS & SUMMARY

IT SHOWS THE RELATED IMAGE AND KEYWORDS.

‹ PREV   NEXT ›

SHOT DOWN	PASSENGER JET
PLANE CRASH	MISSILE FIRED
BLACK BOX	CIVIL AVIATION
TOP PRIORITY	AIR TRAFFIC CONTROL
AIR TRAFFIC	REBEL LEADER

Malaysia Airlines flight MH17 crash: 'Nine Britons, 23 Americans and 80 children' feared dead after Boeing passenger jet is 'shot down' near Ukraine-Russia border Rescuers stand on the site of the crash of a Malaysian airliner near the town of Shaktarsk, in rebel-held east Ukraine Nine Britons, 23 US citizens and 80 children are reported to be among the 298 people killed when a Malaysia Airlines jet crashed near the eastern Ukraine border on Thursday.

CATEGORY REPRESENTATIVE PHRASES

IT SHOWS RELEVANT WORDS OF DIFFERENT CATEGORIES;

category names and three examples from the experts

POLITICAL	MILITARY	ECONOMIC	SOCIAL	INFORMATION	CIVILIAN
Political power	Military forces	Employment	Demographic	Infowars	Urban areas
Dictator	Infantry	Economic activity	Ethnic	Information warfare	Residential area
Anarchy	Insurgents	Market	Population	Radio	Utilities
Pro government	Combatants	Finance	Language	Information security	Transportation
Neo nazi	National guard	European union	Ethnic russians	Ekho moskvy	Nuclear power plants
Viktor yanukovych	Armored vehicles	Foreign policy	Soviet union	Ukraine http empr	Power plants
Right sector	Special forces	Sergei ivanov	Western ukraine	Social media	Nuclear fuel
Pro russian	Self defense	Interior ministry	Russian language	News media	Crash site
Opposition politicians	Armored personnel	Economic sanctions	Police state	Novaya gazeta	Civil aviation
Maidan movement	Pro russian separatists	Rinat akhmetov	Anglo zionist empire	Ria novosti	Surface to air missile
Pro western	Donetsk oblast	Billion dollars	Maidan supporters	Rfe rl	Contaminated water
Kulikovo pole	Heavy fighting	Right sector	The vast majority	Mainstream media	Main entrance
Communist party	Peoples militia	Closer ties	Social media	Main st	
Civil war	Automatic rifles	Magnitsky act	Martial law	Intellig commu	

Category representative phrases generated automatically

IMAGE & TOP-K KEYWORDS & SUMMARY

IT SHOWS THE RELATED IMAGE AND KEYWORDS.

ALLEGEDLY SHOT | EYE PATCHES
TEAR GAS INSIDE | PATCHES
AIRPORTS | AIRPORT SECURITY
CHASING PROTESTERS | CHARGED PROTESTERS
BEANBAG ROUND | NEWS FOOTAGE

Text and Visual Summarization for Hong Kong Protests @ 2019

Demonstrators don eye patches at Lantau Island hub, one of the world's busiest international airports, in anger that a girl allegedly shot with a police beanbag round could lose an eye \n Sit-in comes after night of escalated violence inside subway stations \n Demonstrators don eye patches at Lantau Island hub, one of the world's busiest international airports, in anger that a girl allegedly shot with a police beanbag round could lose an eye.

Phrase-based Topic Mining on the corpus

CATEGORY REPRESENTATIVE PHRASES

IT SHOWS RELEVANT WORDS OF DIFFERENT CATEGORIES;

POLITICAL	POLICE	ECONOMIC	INFORMATION	INFRASTRUCTURE
pro democracy	tear gas	financial crisis	cbc news	hong kong university
pro beijing	hong kong police	economic downturn	cbs news	transportation
hong kong government	riot police	economic growth	fox news	international aiport
Chief executive	Water cannon	Infrastructure	Chinese state media	Mass transit railway
Mainland china	Pepper spray	Real estate	Bbc news	Lantau link
Pro establishment	Petrol bombs	Affordable housing	Global times	Flight cancellations
Mainland chinese	Hong kong government	Trade war	News media	Victoria harbour
Chief executive carrie lam	Beanbag rounds	The united states	Sina weibo	Rail operator
Carrie lam	Firing tear gas	Financial secretary	Internet censorship	Busiest airports
The chinese government	Tsuen wan	Global financial	Local media	Public transport

EvidenceMiner: retrieving related documents

COVID-19, remdesivir

COVID 19 | Cancer And Heart Disease

"COVID-19, remdesivir" (Total: 10000, Took: 3ms)

~ At most 10 results are shown per page ~

Remdesivir as a possible therapeutic option for the COVID-19 Title

Evidence Score 31.90 | 2020 | Travel Medicine and Infectious Disease | Al-Tawfiq, Jaffar A.

Clinical trials on promising regimens for COVID-19, such as remdesivir, lopinavir, and chloroquine phosphate are ongoing, which shed light on conquering the COVID-19 epidemic 12 . Context

Evidence Score 31.12 | No Date | No journal info

Title: No Title

4 28 In addition, a case report showed that remdesivir, an adenosine analogue, has shown survival benefits in one severe COVID-19 pneumonia patient 29 . Context

Evidence Score 25.78 | No Date | No journal info

Title: No Title

On 13 February, Clifford Lane went to a Washington, D.C.–area airport to catch a flight to Japan, where he would help launch a study of an experimental drug, remdesivir, against coronavirus disease 2019 (COVID-19). Context

Evidence Score 25.02 | 2020 | Science | Cohen, Jon

Title: Quarantined at home now, U.S. scientist describes his visit to China's hot zone

Clinical trials of remdesivir for treatment of COVID-19 just started on Feb. 5th and 12th, 2020 in Wuhan and Beijing, respectively, and the experimental results remain unclear [85, 86]. Context

Evidence Score 23.62 | 2020 | Viruses | Xu, Jiabao

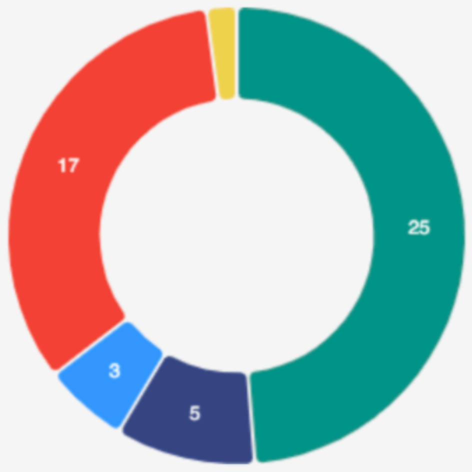
Title: Systematic Comparison of Two Animal-to-Human Transmitted Human Coronaviruses: SARS-CoV-2 and SARS-CoV

Several drugs such as chloroquine, arbidol, remdesivir, and favipiravir are currently undergoing clinical studies to test their efficacy and safety in the treatment of coronavirus disease 2019 (COVID-19) in China; some promising results have been achieved thus far. Context

Evidence Score 22.94 | 2020 | Drug Discoveries & Therapeutics | Dong, Liying

Title: Discovering drugs to treat coronavirus disease 2019 (COVID-19)

Label Coloring & Entity Counts



1 PHYSICAL OBJECT
4 subtypes, 25 entities

2 SPACY TYPE
5 subtypes, 17 entities

3 PHENOMENON OR PROCESS
2 subtypes, 5 entities

CORONAVIRUS cause DISEASEORSYNDROME

COVID 19 | Cancer And Heart Disease

"CORONAVIRUS cause DISEASEORSYNDROME" (Total: 10000, Took: 5ms)

~ At most 10 results are shown per page ~

HCoV-OC43, HCoV-229E, HCoV-HKU1, and HCoV-NL63 cause mild, self-limiting upper respiratory tract infections. Context

Evidence Score 19.00 | 2019 Jan 16 | Viruses | PMID30654597 | Yan, Bingpeng

Title: Characterization of the Lipidomic Profile of Human Coronavirus-Infected Cells: Implications for Lipid Metabolism Remodeling upon Coronavirus Replication

BACKGROUND: Coronavirus causes respiratory infections in humans. Context

Evidence Score 18.34 | 2016 Aug 26 | Springerplus | PMID27625974 | Soonnarong, Rapeepun

Title: Molecular epidemiology and characterization of human coronavirus in Thailand, 2012–2013

BACKGROUND: Porcine deltacoronavirus (PDCoV) is a novel coronavirus that can cause diarrhea in nursing piglets. Context

Evidence Score 18.30 | 2019 Apr 16 | BMC Vet Res | PMID30992015 | Wu, Jiao L.

Title: Expression profile analysis of 5-day-old neonatal piglets infected with porcine Deltacoronavirus

Feline infectious peritonitis (FIP), caused by virulent feline coronavirus, is the leading infectious cause of death in cats. Context

Evidence Score 18.09 | 2019 Dec 30 | Viruses | PMID31905881 | Chen, Si

Title: Feline Infectious Peritonitis Virus Nsp5 Inhibits Type I Interferon Production by Cleaving NEMO at Multiple Sites

The SARS coronavirus causes lung injury and inflammation in part through actions on the nonclassical renin angiotensin pathway. Context

Evidence Score 17.69 | No Date | No journal info | Hendrickson, Carolyn M.

Title: Viral Pathogens and Acute Lung Injury: Investigations Inspired by the SARS Epidemic and the 2009 H1N1 Influenza Pandemic

21 Choi et al. Context

Evidence Score 17.57 | 2019 Dec 17 | Infect Drug Resist | PMID31908501 | Czyzewski, Krzysztof

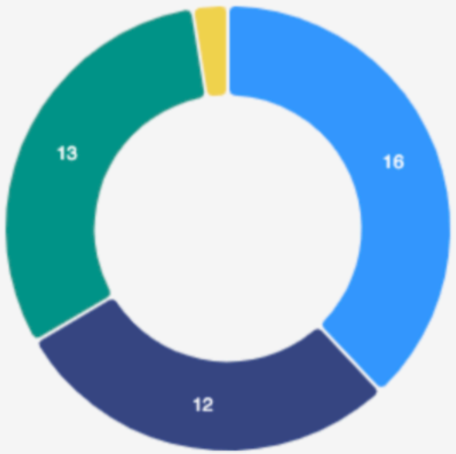
Title: Epidemiology, Outcome and Risk Factors Analysis of Viral Infections in Children and Adolescents Undergoing Hematopoietic Cell Transplantation: Antiviral Drugs Do Not Prevent Epstein–Barr Virus Reactivation

M iddle East respiratory syndrome coronavirus (MERS-CoV) is a novel coronavirus that can cause severe lower respiratory tract infection in humans (1,2). Context

Evidence Score 17.36 | 2014 Aug | Emerg Infect Dis | PMID25075761 | Raj, V. Stalin

Title: Isolation of MERS Coronavirus from a Dromedary Camel, Qatar, 2014

Label Coloring & Entity Counts



1 NEW TYPE
3 subtypes, 16 entities

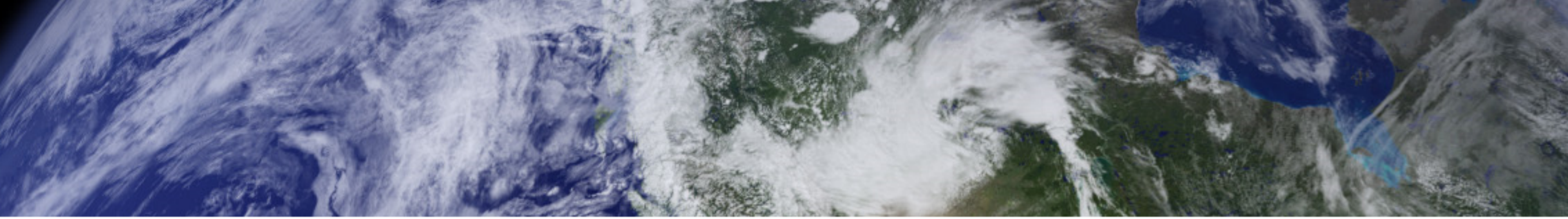
2 PHYSICAL OBJECT
6 subtypes, 13 entities

3 PHENOMENON OR PROCESS
1 subtype, 12 entities

4 ACTIVITY
1 subtype, 1 entity

References

- W. Feng, C. Zhang, W. Zhang, J. Han, J. Wang, C. Aggarwal, J. Huang, "STREAMCUBE: Hierarchical Spatio-temporal Hashtag Clustering for Event Exploration over the Twitter Stream", ICDE’15.
- Meng, Y., Shen, J., Zhang, C., & Han, J. “Weakly-supervised neural text classification”, CIKM’18
- Meng, Y., Shen, J., Zhang, C., & Han, J. “Weakly-Supervised Hierarchical Text Classification”, AAAI’19
- F. Ravat, O. Teste, R. Tournier, and G. Zurfluh. “Top_keyword: An aggregation function for textual document OLAP”. In Data Warehousing and Knowledge Discovery, pages 55–64, 2008.
- F. Tao, K. H. Lei, J. Han, C. Zhai, X. Cheng, M. Danilevsky, N. Desai, B. Ding, J. Ge, H. Ji, R. Kanade, A. Kao, Q. Li, Y. Li, C. X. Lin, J. Liu, N. C. Oza, A. N. Srivastava, R. Tjoelker, C. Wang, D. Zhang, and B. Zhao. Eventcube: multi-dimensional search and mining of structured and text data. KDD’13 (demo)
- F. Tao, C. Zhang, X. Chen, M. Jiang, T. Hanratty, L. Kaplan, and J. Han. “Doc2cube: Automated document allocation to text cube via dimension-aware joint embedding”. ICDM’18
- F. Tao, H. Zhuang, C. Wang, Q. Wang, T. Cassidy, L. Kaplan, C. Voss, J. Han, "Multi-Dimensional, Phrase-Based Summarization in Text Cubes", Data Eng. Bulletin 39(3):74-84, 2016
- Zhang, Y., Meng, Y., Huang, J., Xu, F.F., Wang, X., & Han, J. “Minimally Supervised Categorization of Text with Metadata”, SIGIR’20
- Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping, Jiawei Han, “EVIDENCEMINER: Textual Evidence Discovery for Life Sciences”, ACL’20 (demo)



Q&A