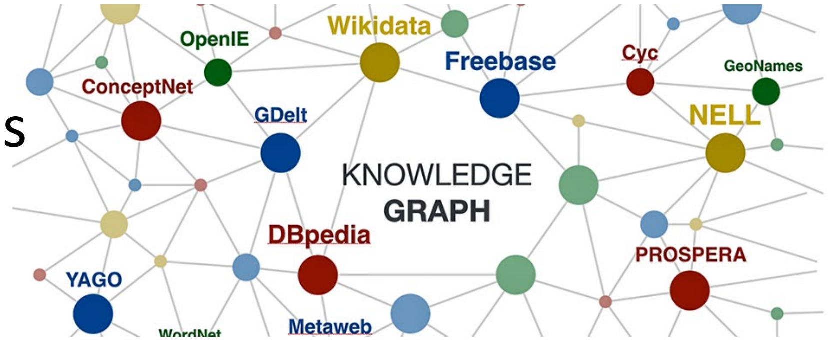
Construction of Multi-Dimensional Knowledge Graphs

JIAWEI HAN, MICHAEL-AIKEN CHAIR PROFESSOR
DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN
AUGUST 25, 2020

Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Why Multi-Dimensional Knowledge Graphs?



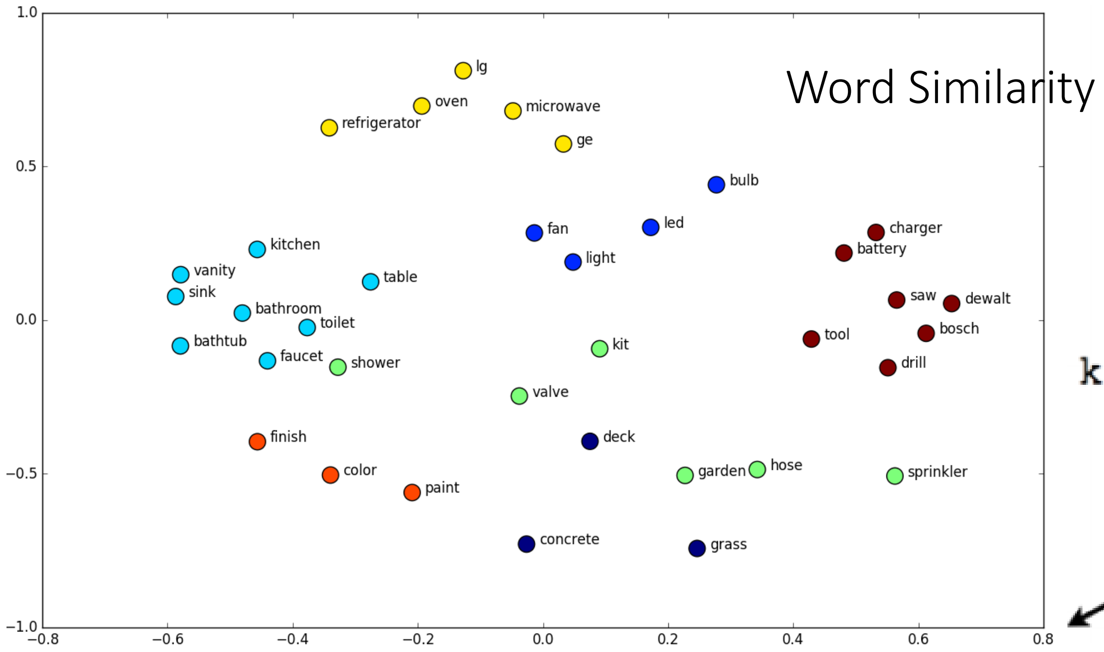
- General knowledge graphs
  - Many existing human-constructed knowledge graphs
  - Active research
    - Automated and incremental construction of KGs
- Why Multi-Dimensional Knowledge Graphs?
  - General KG could be too general to fit concrete problems
  - Properties/links could be conditional on time, location, situation, …
- Major challenges on construction/mining Multi-Dimensional KG
  - Data: 80%+ of big data is in the form of text/natural language/social media:
    - Unstructured, noisy, dynamic, …, but inter-related
  - Automatic construction of MD KG from text

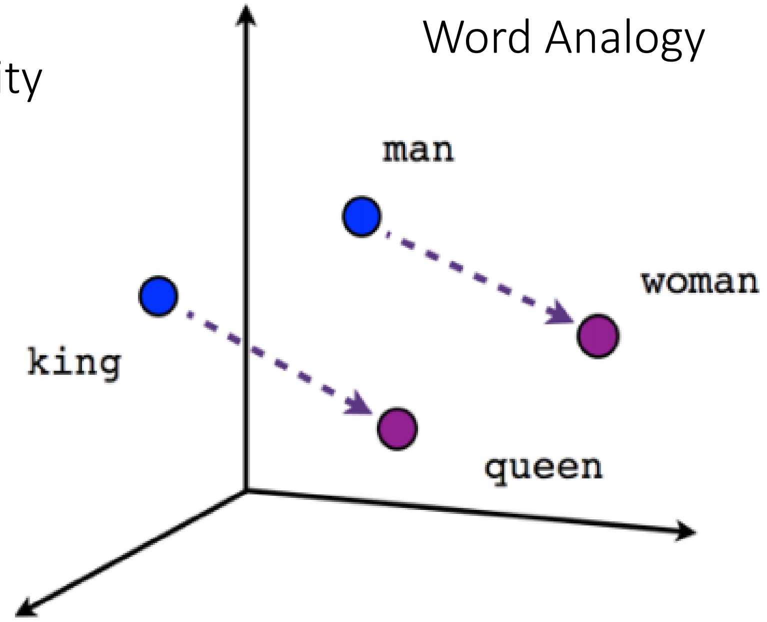
Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Text Embedding: Dimensionality Reduction in Text

- Embed one-hot vectors into lower-dimension space: Address “curse of dimensionality”
- Unsupervised learning of text representations: A milestone in NLP and ML
- Word embedding captures useful properties of word semantics
  - Word similarity: Words with similar meanings are embedded closer
  - Word analogy: Linear relationships between words (e.g., king − queen = man−woman)





Typical embedding methods:

Word2Vec

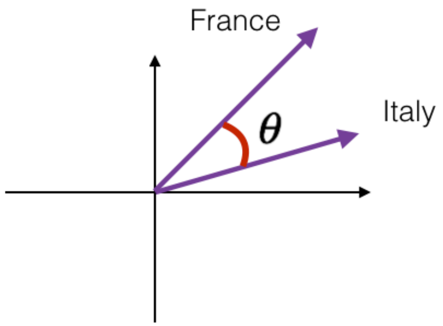
GloVe

fastText

Trained in Euclidean space

Spherical Text Embedding [NeurIPS'19]

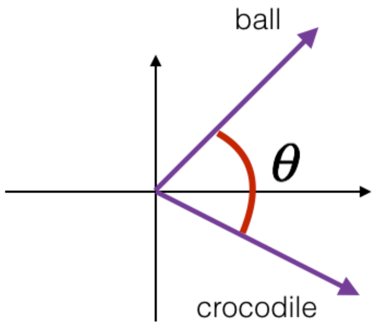
- Previous text embeddings (e.g., Word2Vec) are trained in the Euclidean space
  - But used on spherical space—Mostly directional similarity (i.e., cosine similarity)
  - Word similarity is derived using cosine similarity



France and Italy are quite similar

θ is close to 0°

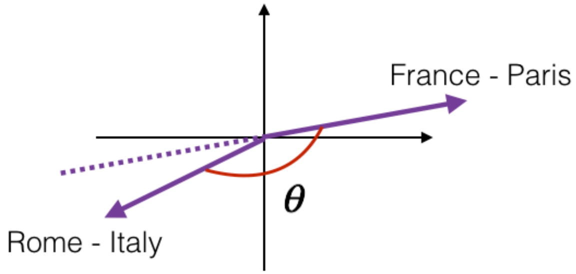
$\cos(\theta) \approx 1$



ball and crocodile are not similar

θ is close to 90°

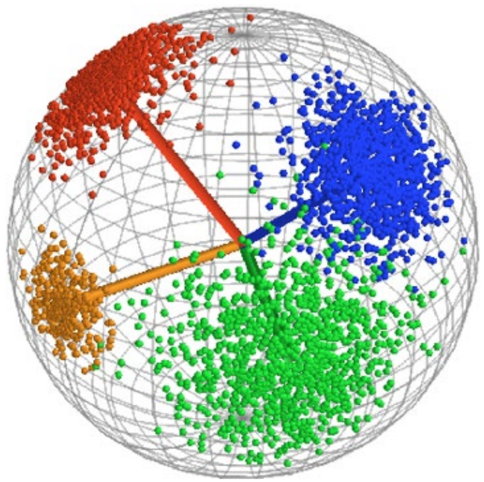
$\cos(\theta) \approx 0$



the two vectors are similar but opposite
the first one encodes (city - country)
while the second one encodes (country - city)

θ is close to 180°

$\cos(\theta) \approx -1$



  - Word clustering (e.g., TaxoGen) is performed on a sphere
  - Better document clustering performances when embeddings are normalized and spherical clustering algorithms are used

Joint Embedding: Integrating Local and Global Contexts

- Local contexts can only partly define word semantics in unsupervised word embedding learning

Local contexts of "harmful"

> If I hear someone screwing with my car (ie, setting off the **alarm**) and **taunting** me to come out, you can be very sure that my Colt Delta Elite will also be coming with me. It is not the screwing with the car that would get them **shot**, it is the potential physical **danger**. If they are **taunting** like that, it's very possible that they also intend to **rob** me and or do other physically *harmful* things. Here in Houston last year a woman heard the sound of someone …

- Design a generative model on the sphere that follows how humans write articles:
  - First a general idea of the paragraph/doc, then start to write down each word in consistent with not only the paragraph/doc, but also the surrounding words

$$p(u \mid d) \propto \exp(\cos(\boldsymbol{u}, \boldsymbol{d})) \qquad p(v \mid u) \propto \exp(\cos(\boldsymbol{v}, \boldsymbol{u}))$$

Document/ Paragraph ($\boldsymbol{d}$) → Center Word ($\boldsymbol{u}$) → Surrounding Word ($\boldsymbol{v}$)

JoSE: Performance & Case Studies

- Document classification

Table 3: Document classification evaluation using k-NN ($k = 3$).

Embedding	20 Newsgroup Macro-F1	20 Newsgroup Micro-F1	Movie Review Macro-F1	Movie Review Micro-F1
Avg. W2V	0.630	0.631	0.712	0.713
SIF	0.552	0.549	0.650	0.656
BERT	0.380	0.371	0.664	0.665
Doc2Vec	0.648	0.645	0.674	0.678
JoSE	**0.703**	**0.707**	**0.764**	**0.765**

- Training efficiency

Table 4: Training time (per iteration) on the latest Wikipedia dump.

Word2Vec	GloVe	fastText	BERT	Poincaré GloVe	**JoSE**
0.81 hrs	0.85 hrs	2.11 hrs	> 5 days	1.25 hrs	**0.73 hrs**

- Decoding acronyms

Table 5: Effect of Global Context on Interpreting Acronyms.

Acronyms	Global ($\lambda = \infty$)	Local ($\lambda = 0$)
CMU	**mellon**, **carnegie**, andrew, pa, pittsburgh	andrew, kfnjyea00uh, am2x, mr47, devineni
UIUC	**urbana**, **illinois**, uxa, **univ**, uchicago	uxa, ux4, ux1, mrcnext, cka52397
UNC	**chapel**, **carolina**, astro, images, usc	launchpad, gibbs, umr, lambada, jge
Caltech	**california**, gap, **institute**, keith, **technology**	juliet, jafoust, lmh, henling, bdunn
JHU	**johns**, camp, **hopkins**, nation, grand	pablo, hasch, iglesias, davidk, atlantis

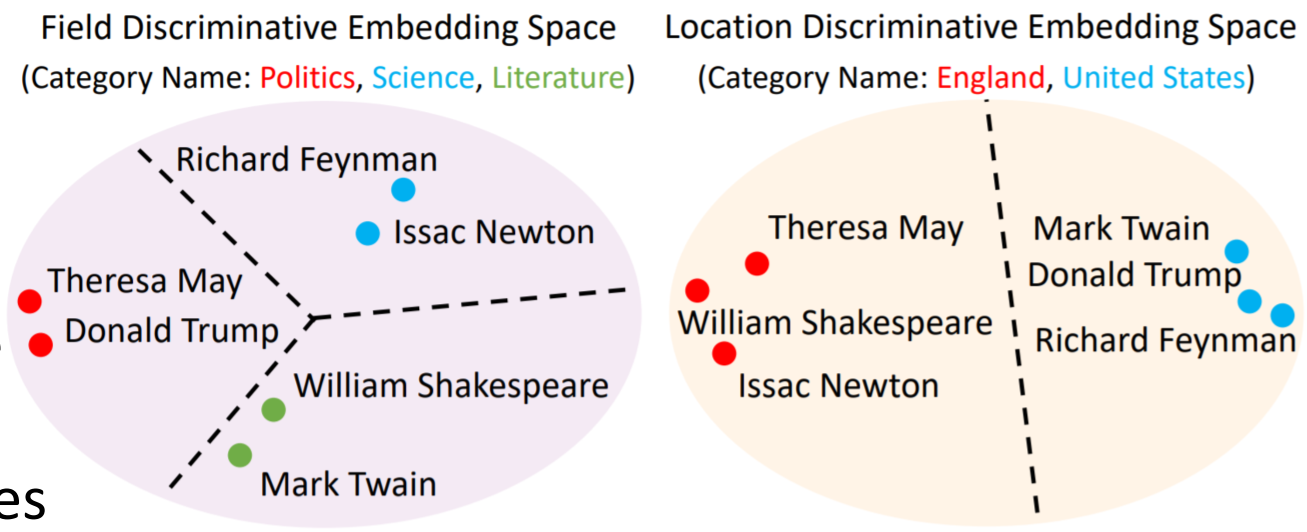
- Distinguishing antonyms with embedding

Table 6: Cosine Similarity of Antonym Embeddings Trained with Different Contexts.

Antonyms	Global ($\lambda = \infty$)	Local ($\lambda = 0$)
good - bad	0.3150	0.7127
happy - unhappy	0.3911	0.6178
large - small	0.4871	0.7265
increase - decrease	0.2663	0.7308
enter - exit	0.2756	0.5553
save - spend	-0.0388	0.4792

Discriminative Topic Mining via Category Name-Guided Embedding

- Traditional text embedding (e.g., Word2Vec, GloVe, fastText, JoSE)
  - Mapping words with similar local contexts closer in the embedding space
  - Not imposing particular assumptions on the type of data distributions
- **CatE: Category Name-guided Embedding [WWW'20]**
  - Weak guidance: leverages *category names* to learn word embeddings with discriminative power over the specific set of categories
- CatE: Inputs
  - Category names + Corpus
- CatE: Outputs
  - The same set of celebrities are embedded differently given different sets of category names



Comparative Evaluation of Discriminative Topic Mining

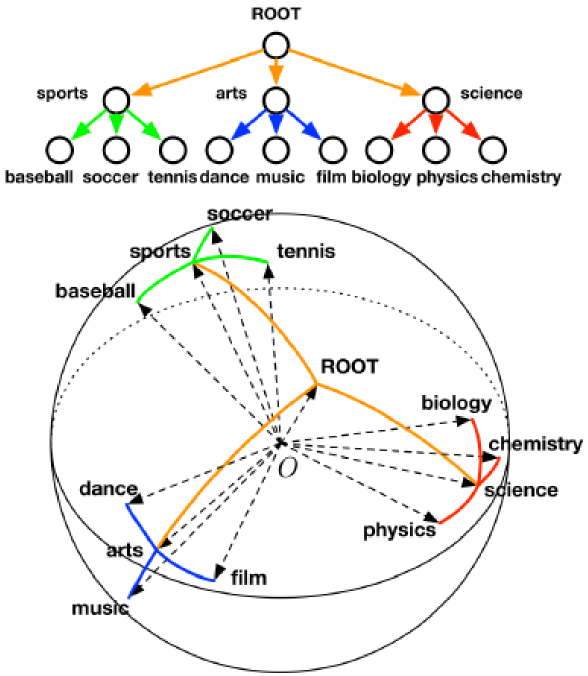
Methods	NYT-Location		NYT-Topic		Yelp-Food		Yelp-Sentiment	
	britain	canada	education	politics	burger	desserts	good	bad
LDA	company (×) companies (×) british shares (×) britain	percent (×) economy (×) canadian united states (×) trade (×)	school students city (×) state (×) schools	campaign clinton mayor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

[WWW’20]

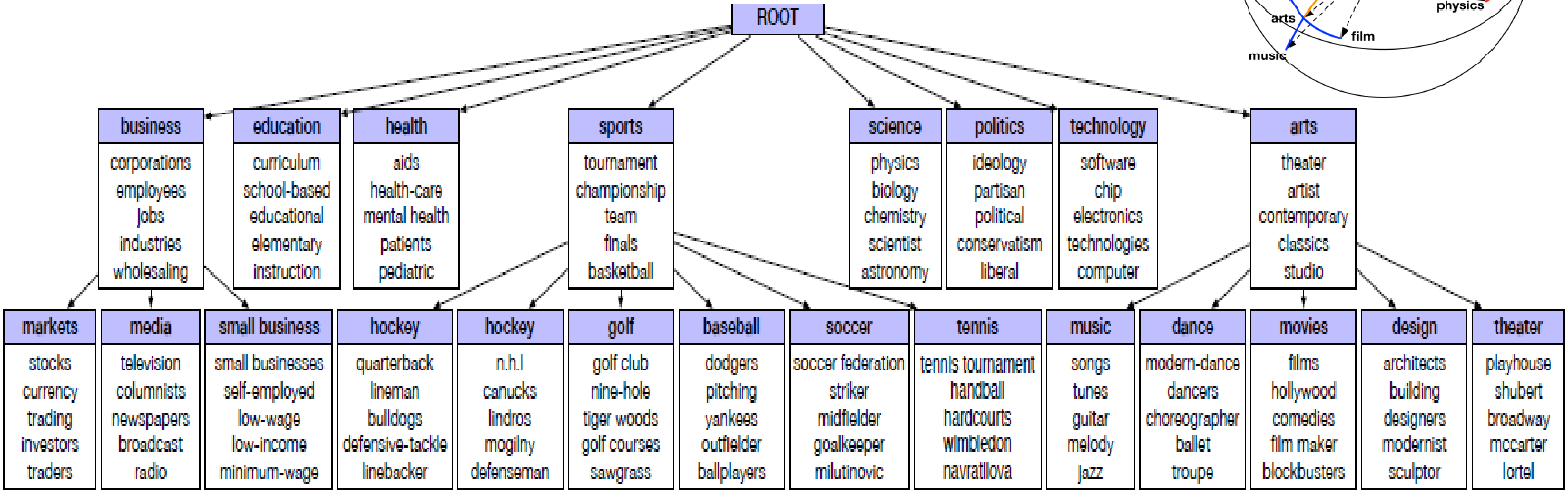
Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]



- JoSH: A joint tree and text embedding method
- Simultaneously modeling of the category tree structure in the spherical space
- Effective category representative term discovery

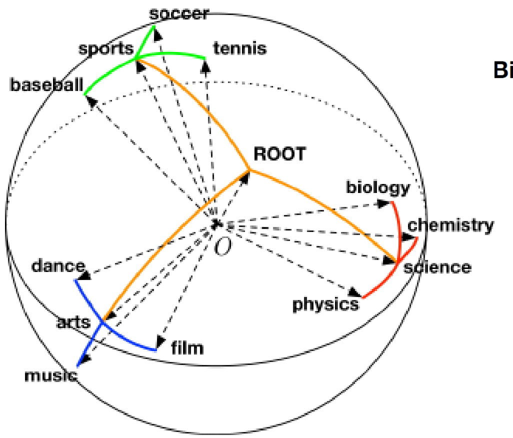


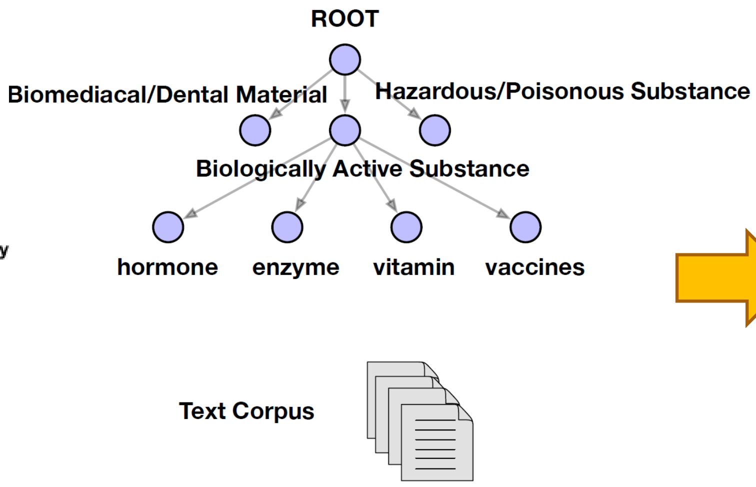
Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding [KDD'20]

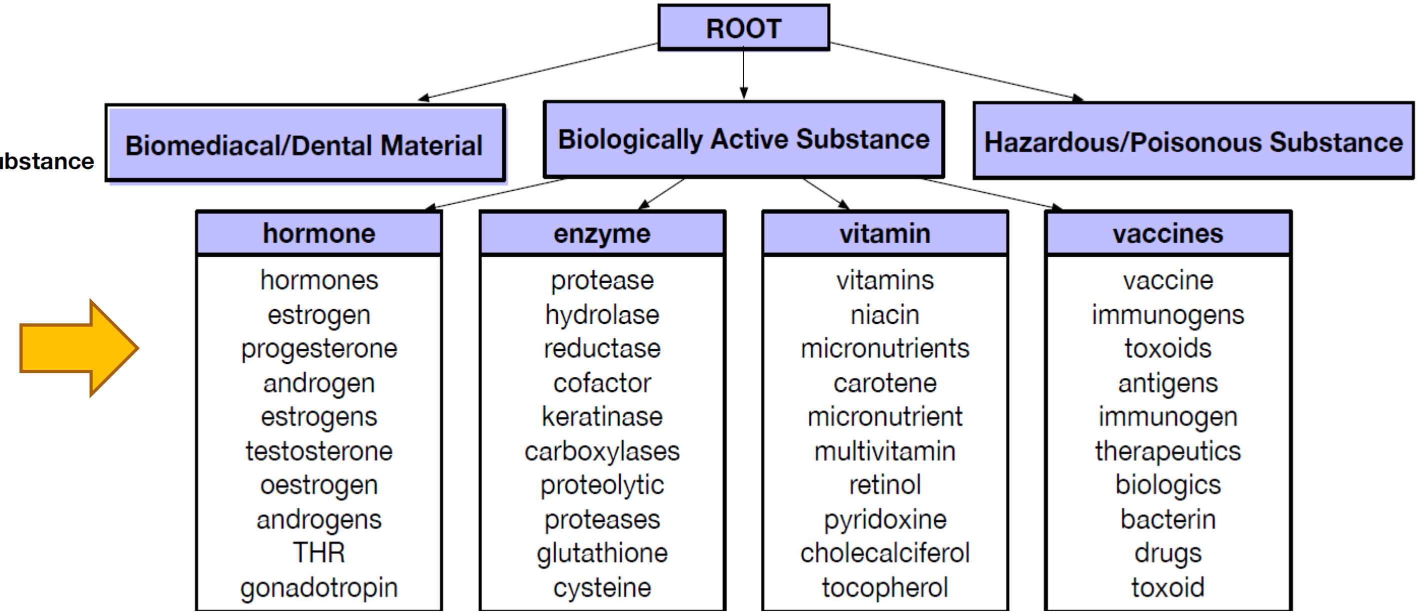
- **Automatic ontology construction from corpus**
  - Hierarchical embedding & discriminative topic mining
    - JoSE [NeurIPS'19], CatE [WWW'20], JoSH [KDD'20]
    - Multiple pieces collaborated with Google, MSR, Amazon
- **Weakly supervised Doc classification and knowledge-base construction**
  - WeSHClass [AAAI'19] and EvidenceMiner [ACL'20] (COVID'19 analysis)

Models	NYT		arXiv	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
WeSHClass	0.425	0.581	0.320	0.542
JoSH	0.429	0.600	0.367	0.610
WeSHClass + CatE	0.503	0.679	0.401	0.622
WeSHClass + JoSH	**0.582**	**0.703**	**0.412**	**0.673**

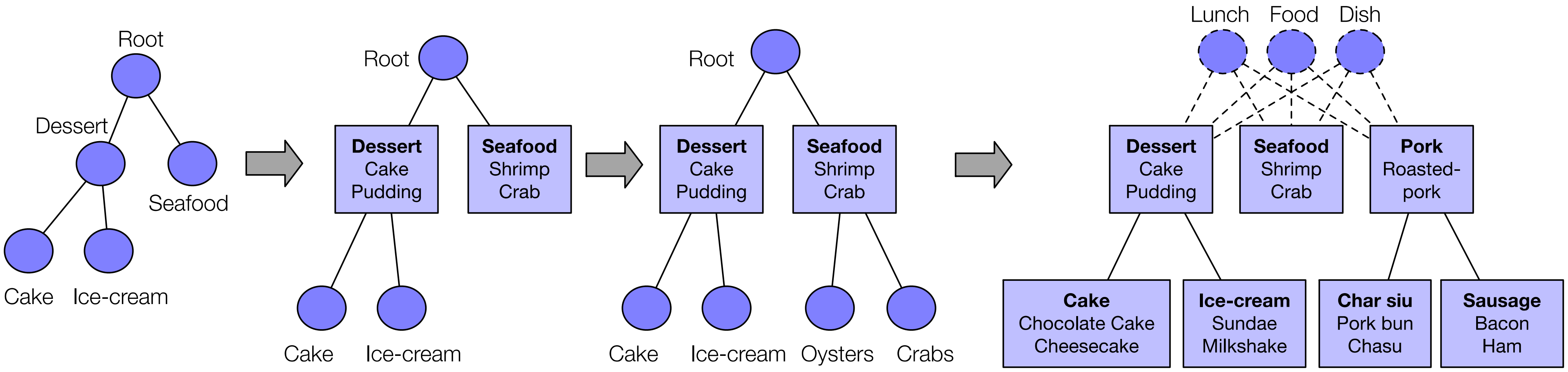
Weakly supervised classification







Seed-guided Taxonomy Construction: CoRel [KDD'20]



Step 1: Concept learning for major topic nodes

Step 2: Topic and relation aware subtopic finding

Step 3: New topic finding by root node discovery & concept learning for all nodes

- Step 1: Extract user-interested terms from the corpus to **enrich the semantics of each node**, and to **generate candidate** terms for subtopic finding. (Concept Learning)
- Step 2: Capture seed relations and transfers the relation to non-leaf nodes for **subtopic finding**. (Relation transferring)
- Step 3: **Discover common root concepts** by transferring the relation to top, and new topics can then be explored through potential root nodes. Enrich new nodes by representative terms. (Relation transferring + Concept Learning)

Qualitative/Quantitative Evaluation

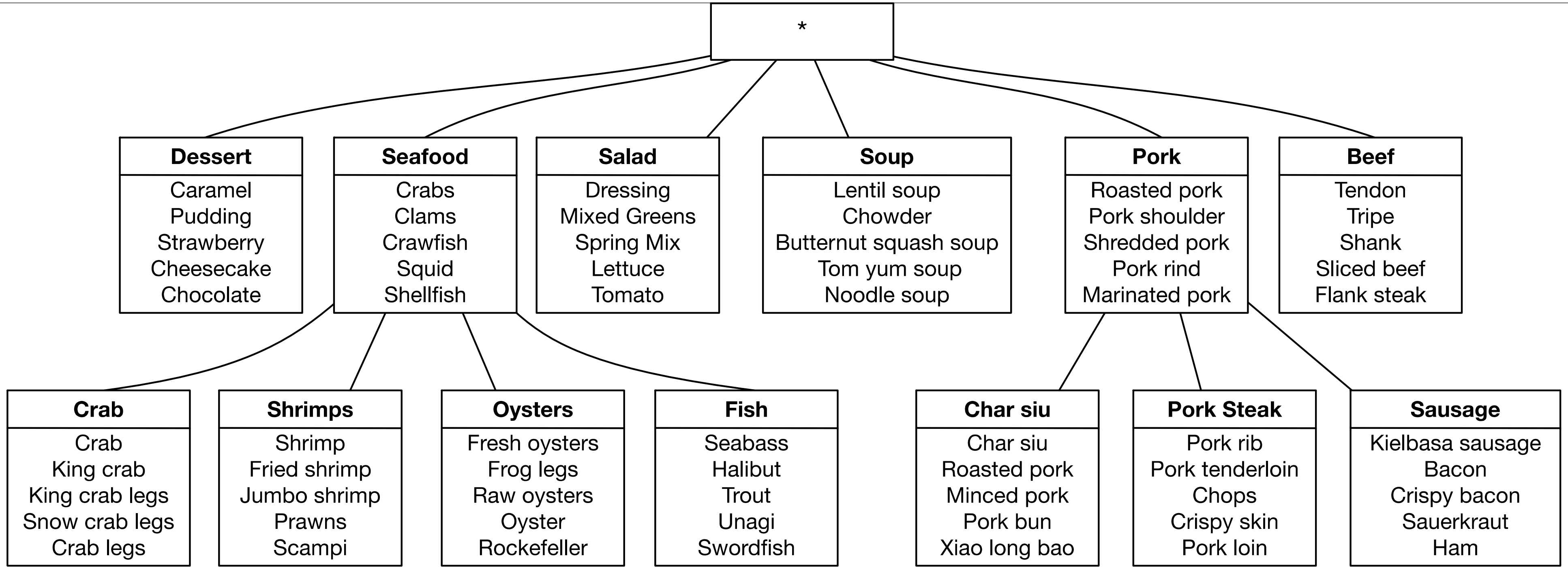


Table 5: Quantitative evaluation on topical taxonomies.

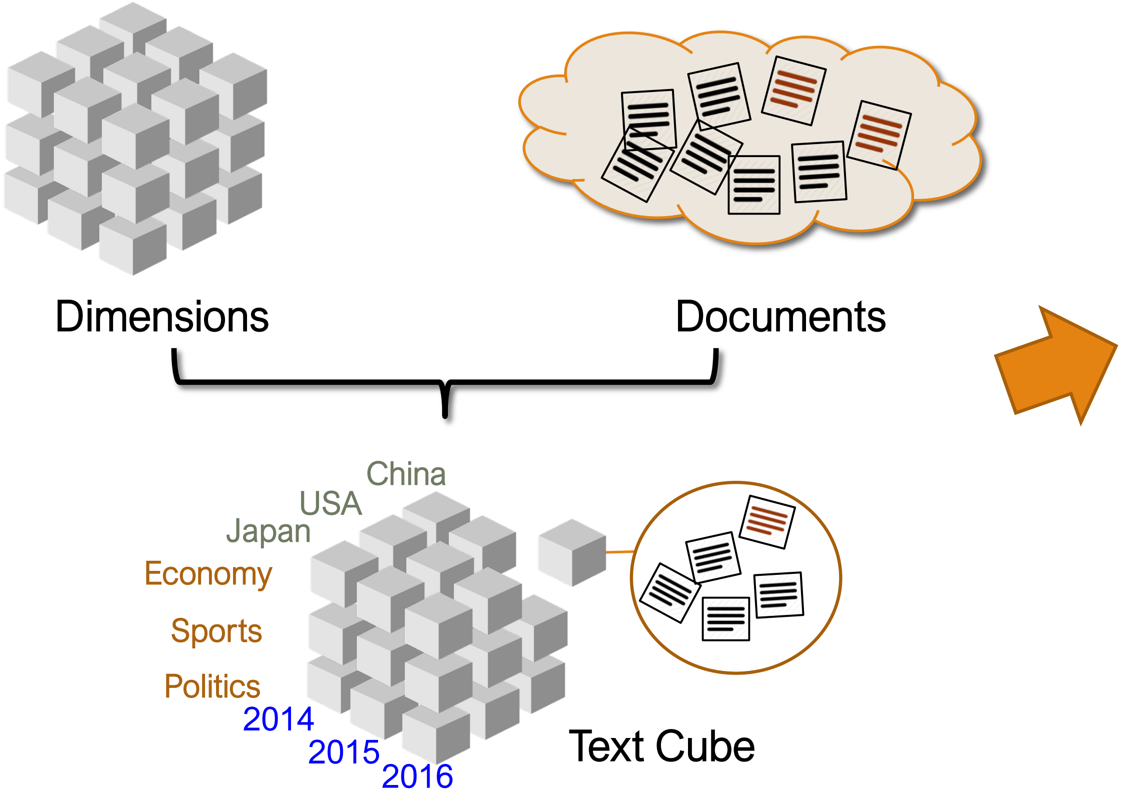
Methods	DBLP					Yelp				
	TC	SD	Precision$_r$	Recall$_r$	F1-score$_r$	TC	SD	Precision$_r$	Recall$_r$	F1-score$_r$
HLDA	0.582	0.981	0.188	0.577	0.283	0.517	0.991	0.135	0.387	0.200
HPAM	0.557	0.905	0.362	0.538	0.433	0.687	0.898	0.173	0.615	0.271
TaxoGen	0.720	0.979	0.450	0.429	0.439	0.563	0.965	0.267	0.381	0.314
Hi-Expan + CoL.	0.819	0.996	0.676	0.532	0.595	0.815	**1.000**	0.429	0.677	0.525
CoRel	**0.855**	**1.000**	**0.730**	**0.607**	**0.663**	**0.825**	**1.000**	**0.564**	**0.710**	**0.629**

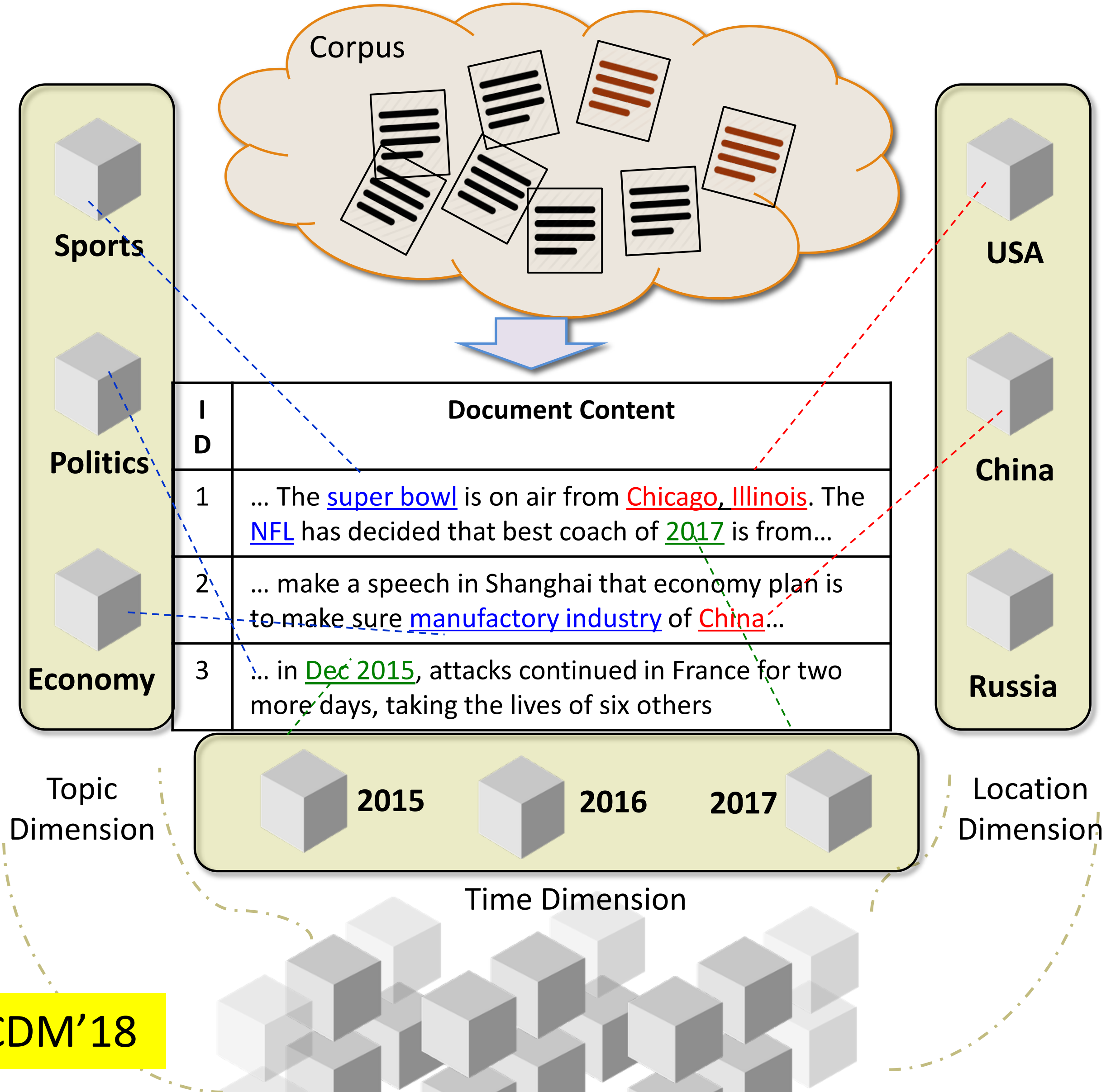
Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Cube Construction: Which Document Goes to Which Cell?

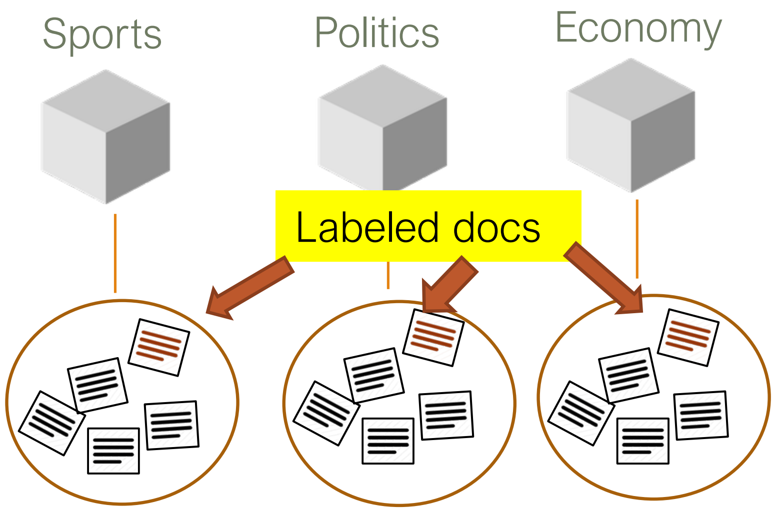
- Cell-based Document Allocation
  - Which document goes to which cell?



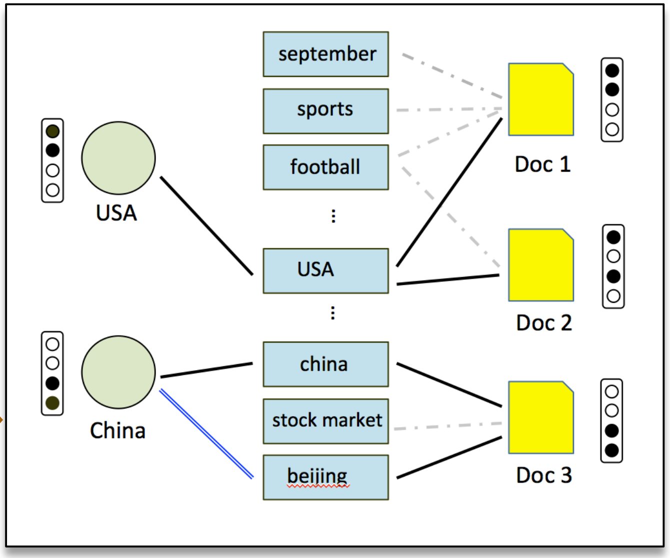


Doc2Cube: Constructing Cube from Massive Docs: ICDM’18

How to Put Documents into the Right Cube Cell?

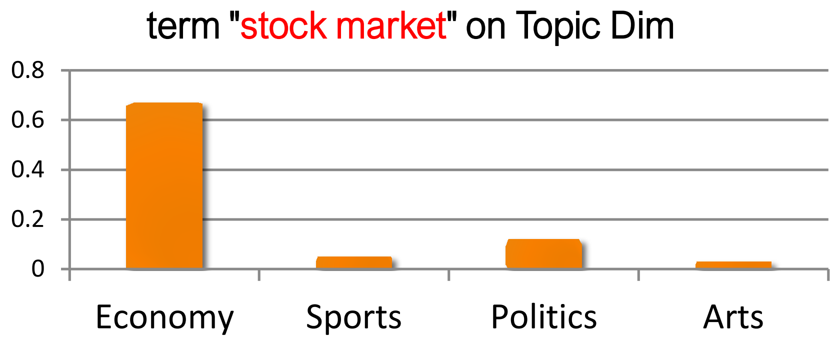


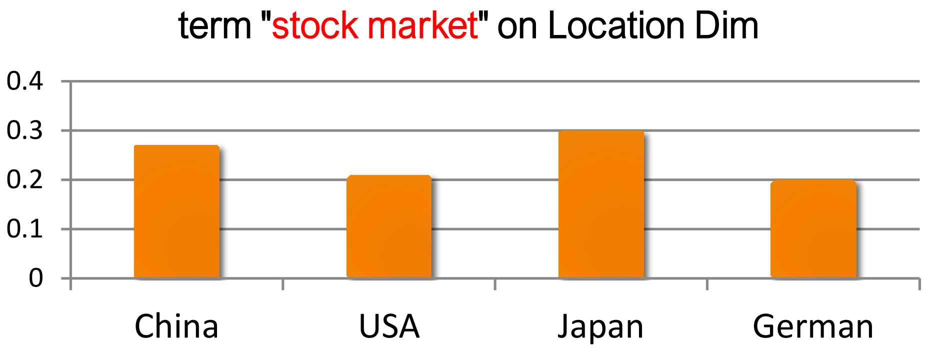
- Major challenges on putting docs into the right cell
  - Few would like label the "training sets"
    - So many cells, so many documents
  - Dimension values are often "under-represented"
    - E.g., Topic dimension: Sports, economy, politics, ….
  - Documents are often "over-represented" on single 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.



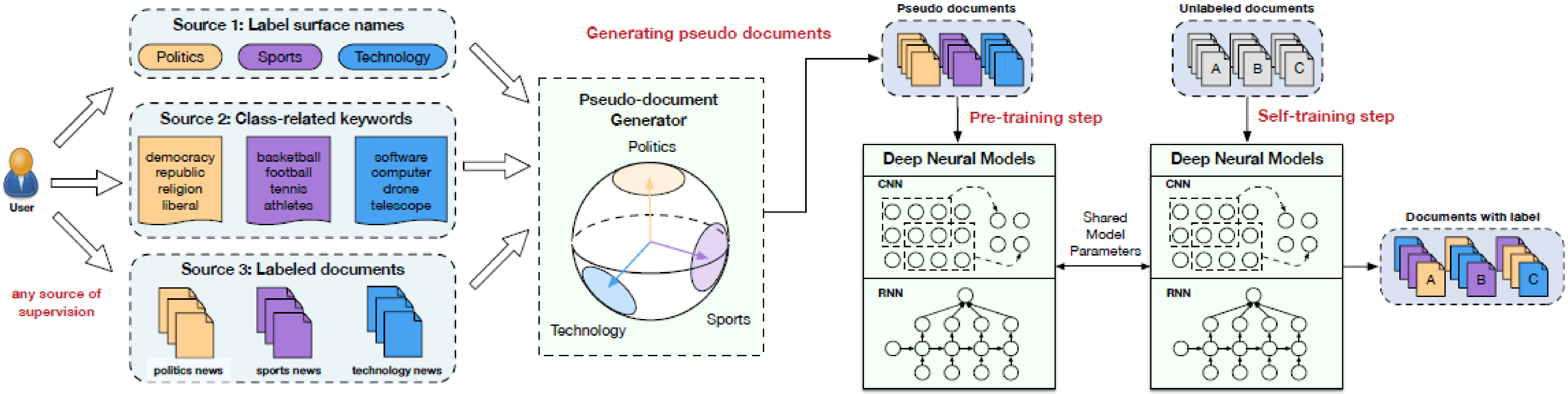


- Label expansion: Combining two measures for seed expansion
    - Discriminativeness
      - Using focal score
    - Popularity
  - Example:

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

WeSTClass: Weakly Supervised Text Classification

- Modeling class distribution in word2vec embedding space
  - Word2vec embedding captures **skip-gram (local) similarity** (i.e., words with similar local context windows are expected to have similar meanings)



WeSTClass (Weakly Supervised Text Classification): CIKM'18
WeSHClass (Weakly Supervised Hierarchical Text Classification): AAAI'19

WeSTClass: 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:

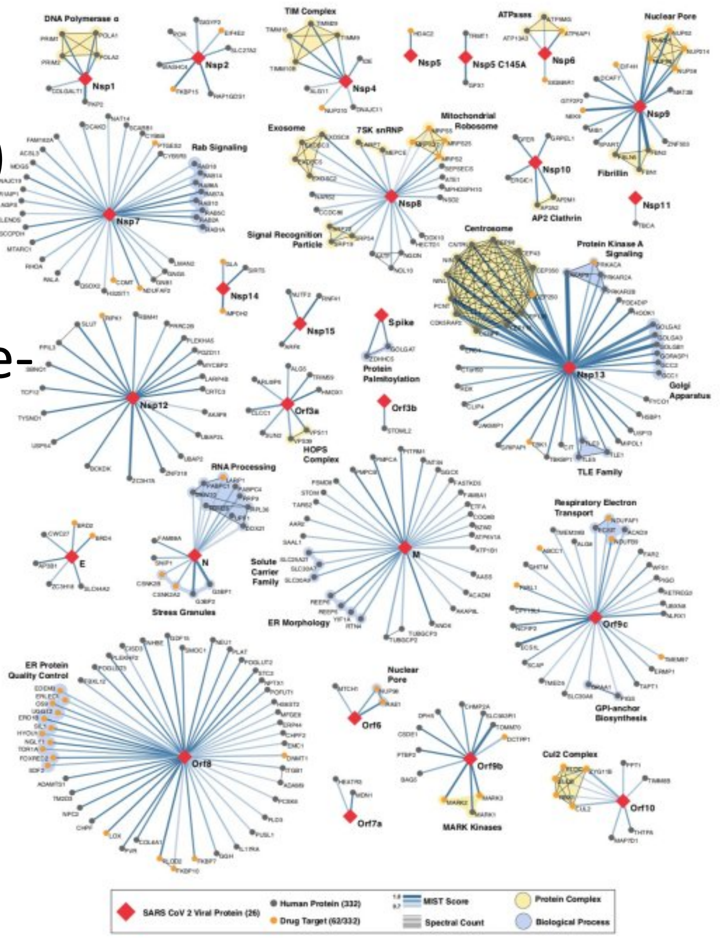
Methods	The New York Times			AG's News			Yelp Review		
	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS	LABELS	KEYWORDS	DOCS
IR with tf-idf	0.240	0.346	-	0.292	0.333	-	0.548	0.652	-
Topic Model	0.666	0.623	-	0.584	0.735	-	0.500	0.500	-
Dataless	0.710	-	-	0.699	-	-	0.500	-	-
UNEC	0.810	-	-	0.668	-	-	0.603	-	-
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)**

Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Construction of Multi-Dimensional Knowledge Graphs

- Layout structure of MD Knowledge graphs
  - Essentials: dimensions, levels and properties (attributes)
  - COVID-19: dimensions (e.g., time, location, situation, …)
  - To study different aspects on COVID-19, construct theme-based KGs (e.g., origin, spread, medication, vaccine, …)
- (Weighted) nodes/edges extracted from text mining
  - Constructed dynamically or materialized into multi-graphs
  - Context-based knowledge graph construction
- Graph mining and embedding methods can then be explored on such graphs
  - A wide-open area for research



Cube Demo | Dataset: Hong Kong Protest | Time: 2019.08.12-08.18 | Category: transportation | City: Islands

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

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.

MissionCube: Analysis of Different News Data Sets: HK Protests

Analysis of Hong Kong Protests

Category representative phrases generated automatically

category names and three examples from the experts

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

Discriminative Topic Mining on COVID-19 Data

- Dataset: Covid-19 Literature Data
  - Link provide in Whitehouse call
  - # of Documents: 9,000
  - # of Tokens: 46,540,112
  - Vocabulary size (unique token): 30,223

On Given Categories:

Origin	Evolution	Symptom	Examination
natrual_host	molecular_evolution	sore_throat	clinical_diagnosis
host_species	mutation_rate	respiratory_illness	emergency_room
homo_sapiens	natural_selection	common_cold	physical_examination
egyptian_fruit_bat	population_dynamics	dry_cough	primary_care
african_green_monkey	interspecies_transmission	acute_illness	observation_period
chinese_horseshoe_bat	genetic_drift	nasal_congestion	chest_x_ray
rhesus_macaque	host_adaptation	chest_pain	chest_radiographs

Distinct Topic Mining on Virus Type:

MERS	SARS	COVID-19	Ebola
middle_east	severe_acute_respiratory_syndrome	hubei_province	west_africa
saudi_arabia	guangdong_province	mainland_china	marburg_virus
renal_failure	murine_coronavirus	2020	evd_outbreak
united_arab_emirates	carboxy_terminal	global_spread	sierra_leone
dipeptidyl_peptidase	nonstructural_protein	incubation_periods	lassa_fever
mers_cov_spike	horseshoe_bats	unknown_etiology	nonhuman_primates
nosocomial_transmission	hcov_229e	imported_cases	ebov_infection

On Age Group:

Infant	Adult	Elderly
early_childhood	healthy_adult	chronic_diseases
blood_transfusion	median_age	older_individuals
cord_blood	adult_male	vulnerable_populations
vertical_transmission	wild_cautght	elderly_patients
maternal_antibodies	household_contact	economic_burden
mammary_gland	organ_transplantation	vulnerable_groups
birth_defects	sexual_transmission	long_term_care

NER Result Visualization: Scientific Literature

*Angiotensin-converting enzyme 2 **GENE_OR_GENOME** (ACE2 **GENE_OR_GENOME**) as a SARS-CoV-2 **CORONAVIRUS** receptor **CHEMICAL**: molecular mechanisms and potential therapeutic target.*

SARS-CoV-2 **CORONAVIRUS** has been sequenced [3] . A phylogenetic **EVOLUTION** analysis [3 , 4] found a bat **WILDLIFE** origin for the SARS-CoV-2 **CORONAVIRUS** . There is a diversity of possible intermediate hosts **NORP** for SARS-CoV-2 **CORONAVIRUS** , including pangolins **WILDLIFE** , but not mice **EUKARYOTE** and rats **EUKARYOTE** [5] . There are many similarities of SARS-CoV-2 **CORONAVIRUS** with the original SARS-CoV **CORONAVIRUS** . Using computer modeling , Xu et al **PERSON**. [6] found that the spike proteins **GENE_OR_GENOME** of SARS-CoV-2 **CORONAVIRUS** and SARS-CoV **CORONAVIRUS** have almost identical 3-D structures in the receptor binding domain that maintains Van der Waals forces **PHYSICAL_SCIENCE** . SARS-CoV spike proteins **GENE_OR_GENOME** has a strong binding affinity **DISEASE_OR_SYNDROME** to human ACE2 **GENE_OR_GENOME** , based on biochemical interaction studies and crystal structure analysis [7] . SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** share identity in amino acid sequences and , importantly, the SARS-CoV-2 **CORONAVIRUS** and SARS-CoV spike proteins **GENE_OR_GENOME** have a high degree of homology [6, 7] . Wan et al **PERSON**. [4] reported that residue 394 **CARDINAL** (glutamine **CHEMICAL**) in the SARS-CoV-2 **CORONAVIRUS** receptor-binding domain ....

	Gene			Chemical			Disease			Total		
	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1	Prec	Rec	F1
SciSpacy (BIONLP13CG)	**91.48**	**82.06**	**86.51**	64.66	39.81	49.28	8.11	2.75	4.11	76.36	53.59	62.98
SciSpacy (BC5CDR)	-	-	-	**86.97**	51.86	64.69	**80.31**	59.65	68.46	**82.40**	54.57	65.66
Ours	82.14	74.68	78.23	82.93	**75.22**	**78.89**	75.73	**68.42**	**71.89**	81.29	**73.65**	**77.28**

Xuan Wang, Xiangchen Song, Bangzheng Li, Yingjun Guan and Jiawei Han, “Comprehensive Named Entity Recognition on CORD-19 with Distant or Weak Supervision”, 2020 Intelligent Systems for Molecular Biology (ISMB’20), Abstracts (oral and poster), July 2020

EvidenceMiner: Theme-Specific Text [ACL'20]

UV, Ultraviolet, kill, sars-cov-2

Example : NSCLC is treated with nivolumab, HCC is treated with sorafenib, prostate cancer is tre

- Xuan Wang, Weili Liu, Aabhas Chauhan, Yingjun Guan and Jiawei Han, "Automatic Textual Evidence Mining in COVID-19 Literature", 2020 Intelligent Systems for Molecular Biology (ISMB'20), Abstracts (poster), July 2020
- Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping and Jiawei Han, "EVIDENCEMINER: Textual Evidence Discovery for Life Sciences", ACL'20 (System demo), July 2020

Sentence   Analytics

"UV, Ultraviolet, kill, sars-cov-2" (Total: 10000, Took: 7ms)
~ At most 10 results are shown per page ~

Ultraviolet-C (UV-C) radiation represents an alternative to chemical inactivation methods [21] . Context

Evidence Score 24.42 | 2019 Feb 21 | PLoS One | PMID30789926

['Blázquez, Elena', 'Rodríguez, Carmen', 'Ródenas, Jesús', 'Navarro, Núria', 'Riquelme, Cristina', 'Rosell, Rosa', 'Campbell, Joy', 'Crenshaw, Joe', 'Segalés, Joaquim', 'Pujols, Joan', 'Polo, Javier']

Title: Evaluation of the effectiveness of the SurePure Turbulator ultraviolet-C irradiation equipment on inactivation of different enveloped and non-enveloped viruses inoculated in commercially collected liquid animal plasma

Microscopy was performed using an IMT-2 Olympus microscope equipped with ultraviolet light (UV) and an OM-4 camera. Context

Evidence Score 24.37 | 2012 Dec 5 | Cells | PMID24710549

['Zhang, Qing', 'Miller, Christopher', 'Bible, Jesse', 'Li, Jiliang', 'Xu, Xiaoqing', 'Mehta, Nozer', 'Gilligan, James', 'Vignery, Agnès', 'Scholz, Jodi A Carlson']

Title: Additive Effects of Mechanical Marrow Ablation and PTH Treatment on de Novo Bone Formation in Mature Adult Rats

We discuss 2 such modalities, respirators (face masks) and ultraviolet (UV) light. Context

Evidence Score 23.61 | 0 | No journal info | PMID | ['Weiss, Martin Meyer', 'Weiss, Peter D.', 'Weiss, Danielle E.', 'Weiss, Joseph B.']

Title: Disrupting the Transmission of Influenza A: Face Masks and Ultraviolet Light as Control Measures

Ultraviolet light-inactivated TGEV (UV-TGEV) were developed by irradiating TGEV stocks under ultraviolet light at a dose of 100 mJ/cm 2 . Context

Evidence Score 23.31 | 2018 Oct 29 | Virulence | PMID30322331

['Wang, Li', 'Qiao, Xinyuan', 'Zhang, Sijia', 'Qin, Yue', 'Guo, Tiantian', 'Hao, Zhenye', 'Sun, Li', 'Wang, Xiaona', 'Wang, Yanan', 'Jiang, Yanping', 'Tang, Lijie', 'Xu, Yigang', 'Li, Yijing']

Title: Porcine transmissible gastroenteritis virus nonstructural protein 2 contributes to inflammation via NF-κB activation

Boeing is also exploring a prototype self-sanitizing lavatory that uses ultraviolet light to kill 99.99% of pathogens 48 . Context

Evidence Score 22.29 | 2019-01-26 | No journal info | PMID | Nicolaides, C.

Title: Hand-hygiene mitigation strategies against globaldisease spreading through the air transportationnetwork

In addition, UV-inactivated SARS CoV also activates immature DCs [44] . Context

Evidence Score 21.75 | 2008 Jul 16 | PLoS One | PMID18628832

['Bai, Bingke', 'Hu, Qinxue', 'Hu, Hui', 'Zhou, Peng', 'Shi, Zhengli', 'Meng, Jin', 'Lu, Baojing', 'Huang, Yi', 'Mao, Panyong', 'Wang, Hanzhong']

Title: Virus-Like Particles of SARS-Like Coronavirus Formed by Membrane Proteins from Different Origins Demonstrate Stimulating Activity in Human Dendritic Cells

Whole UV-inactivated SARS-CoV (UV-V), bearing multiple epitopes and proteins, is a candidate vaccine against this virus. Context

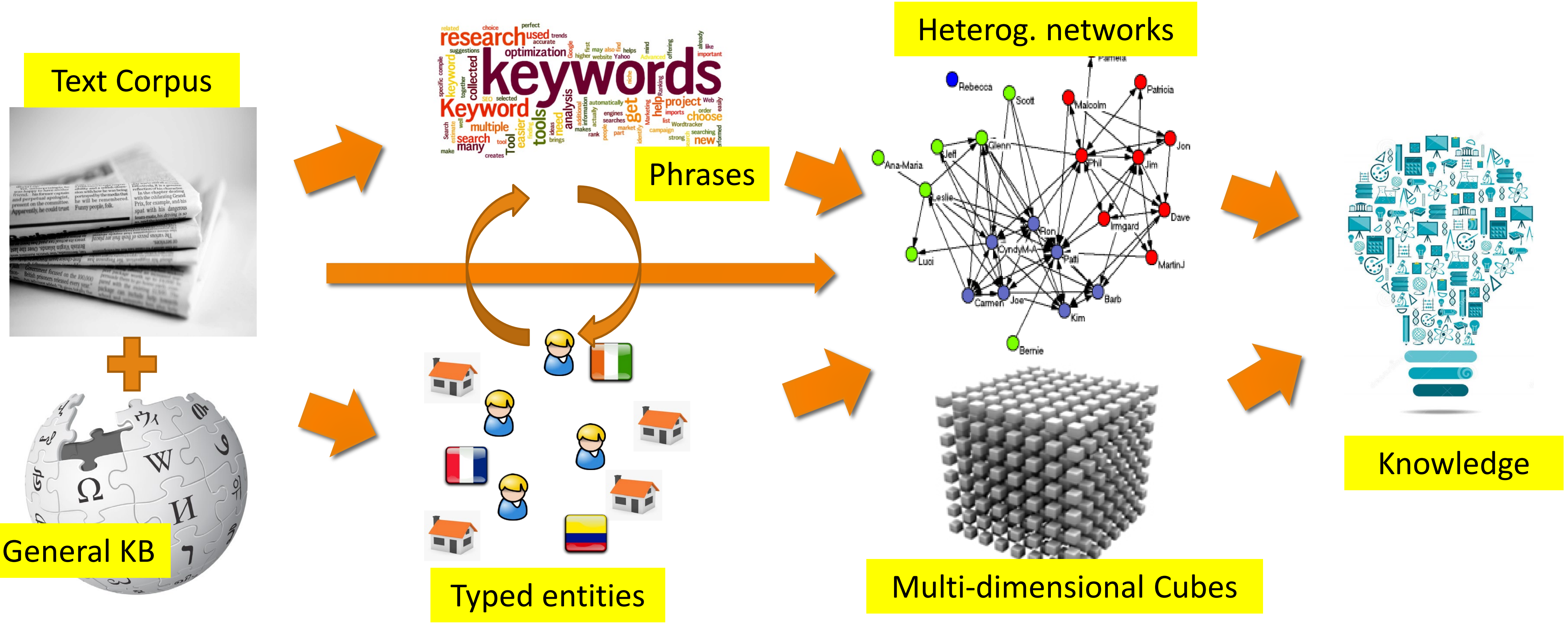
Evidence Score 21.74 | 0 | No journal info | PMID

Outline

- A Knowledge Graph in Need is a Good Knowledge Graph Indeed
- Mining Unstructured Text for Multi-Dimensional Structured Information
  - Understanding Semantics: Text Embedding and Discriminative Topic Analysis
  - Organizing Data in Hierarchical Conceptual Space: Hierarchical Topic Mining
  - Organizing Documents in Multi-Dimensional Space: Text-Cubes
- Construction of Multi-Dimensional Knowledge Graphs
- Looking Forward

Looking Forward: From Massive Text to Multi-D KGs

- To make KG solve real problems: We need multi-dimensional, situation-based knowledge graphs
- Key challenge: Automatically construct such knowledge graphs from text data



References

- Yu Meng, Jiaxin Huang, Jiawei Han, “Embedding-Driven Multi-Dimensional Topic Mining and Text Analysis”, (Conference tutorial), 2020 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’20), August 2020
- Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang and Jiawei Han, “CoRel: Seed-Guided Topical Taxonomy Construction by Concept Learning and Relation Transferring”, in Proc. of 2020 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’20), August 2020
- Yu Meng, Yunyi Zhang, Jiaxin Huang, Yu Zhang, Chao Zhang and Jiawei Han, “Hierarchical Topic Mining via Joint Spherical Tree and Text Embedding”, in Proc. of 2020 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’20), August 2020
- Dawei Zhou, Lecheng Zheng, Jiawei Han and Jingrui He, “A Data Driven Graph Generative Model for Temporal Interaction Networks”, in Proc. of 2020 ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD’20), August 2020
- Xuan Wang, Yingjun Guan, Weili Liu, Aabhas Chauhan, Enyi Jiang, Qi Li, David Liem, Dibakar Sigdel, John Caufield, Peipei Ping and Jiawei Han, “EVIDENCEMINER: Textual Evidence Discovery for Life Sciences”, in Proc. 2020 Annual Conf. of the Association for Computational Linguistics (ACL’20) (System demo), July 2020
- Honglei Zhuang, Fang Guo, Chao Zhang, Liyuan Liu and Jiawei Han, “Joint Aspect-Sentiment Analysis with Minimal User Guidance", in Proc. 2020 ACM SIGIR Int. Conf. on Research and development in Information Retrieval (SIGIR'20), July 2020
- Jiaxin Huang, Yiqing Xie, Yu Meng, Jiaming Shen, Yunyi Zhang and Jiawei Han, “Guiding Corpus-based Set Expansion by Auxiliary Sets Generation and Co-Expansion”, in Proc. 2020 Int. World Wide Web Conf. (WWW’20), Apr. 2020
- Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang and Jiawei Han, “Discriminative Topic Mining via Category-Name Guided Text Embedding”, in Proc. 2020 The Wide Web Conf. (WWW’20), Apr. 2020
- Jiaming Shen, Zhihong Shen, Chenyan Xiong, Chi Wang, Kuansan Wang and Jiawei Han, “TaxoExpan: Self-supervised Taxonomy Expansion with Position-Enhanced Graph Neural Network”, in Proc. 2020 The Web Conf. (WWW’20), Apr. 2020
- Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan and Jiawei Han, "Spherical Text Embedding", in Proc. 2019 Conf. on Neural Information Processing Systems (NeurIPS’19), Vancouver, Canada, Dec. 2019

Acknowledgements

- Thanks for the research support from: ARL/NSCTA, NIH, NSF, DARPA, DTRA, ……