



Adapting Pretrained Representations for Text Mining

Yu Meng, Jiaxin Huang, Yu Zhang, Jiawei Han

Department of Computer Science

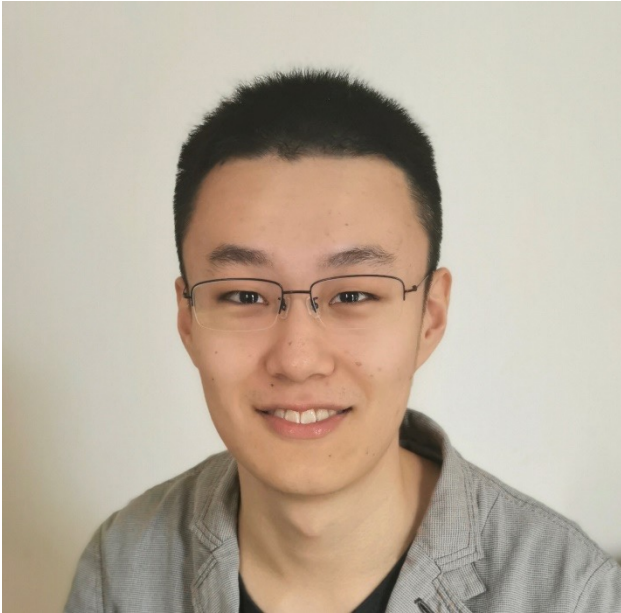
University of Illinois at Urbana-Champaign

Aug 14, 2022

Estimated Timeline for This Tutorial

- Introduction: **15 mins (13:00-13:15 Han)**
- Part I: Pretrained Language Models: **35 mins (13:15-13:50 Meng)**
- Part II: Revisiting Text Mining Fundamentals with Pretrained Language Models: **35 mins (13:50-14:25 Huang)**
- Break: **10 mins (14:25-14:35)**
- Part III: Embedding-Driven Topic Discovery: **35 mins (14:35-15:10 Yu Zhang, Yu Meng, Jiaxin)**
- Part IV: Weakly-Supervised Text Classification: Embeddings with Less Human Effort: **35 mins (15:10-15:45 Zhang)**
- Part V: Advanced Text Mining Applications: **10 mins (15:45-15:55 Han)**
- Summary and Future Directions: **5 mins (15:55-16:00)**

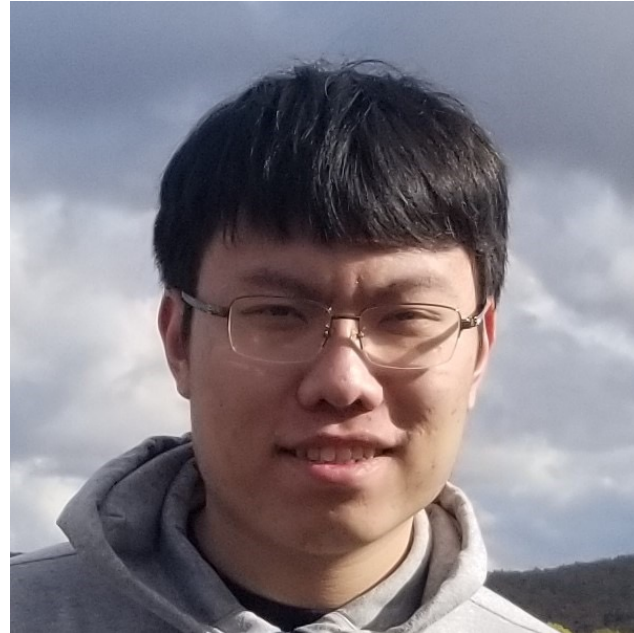
About Instructors



- ❑ Yu Meng
Ph.D. Candidate, UIUC
- ❑ Recipient of 2021
Google PhD Fellowship
in Structured Data and
Database Management



- ❑ Jiaxin Huang
Ph.D. Candidate, UIUC
- ❑ Recipient of 2021
Microsoft PhD
Fellowship



- ❑ Yu Zhang
Ph.D. Candidate, UIUC
- ❑ Recipient of 2022 Yunni
and Maxine Pao
Memorial Fellowship



- ❑ Jiawei Han
Michael Aiken Chair
Professor at UIUC
- ❑ ACM SIGKDD
Innovation Award
Winner (2004)

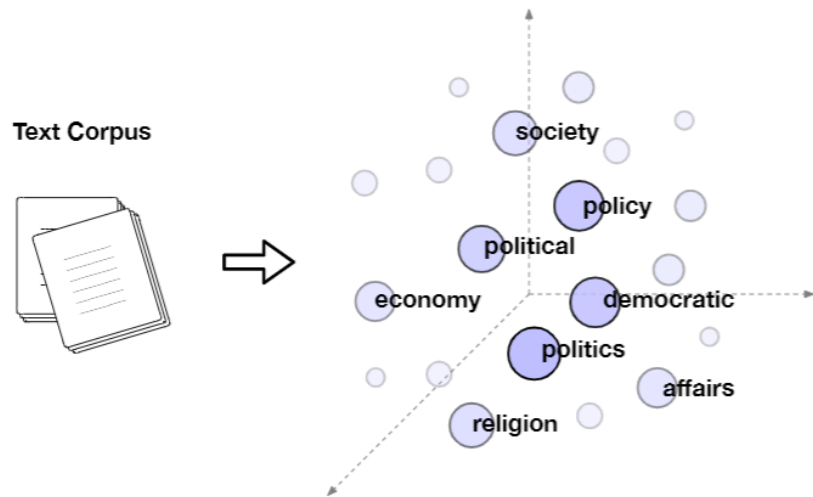
Over 80% of Big Data is Unstructured Text Data

- ❑ Ubiquity of big unstructured, text data
 - ❑ **Big Data**: Over 80% of our data is from text (e.g., news, papers, social media): unstructured/semi-structured, noisy, dynamic, inter-related, high-dimensional, ...
- ❑ How to mine/analyze such big data systematically?
 - ❑ **Text Representation** (i.e., computing vector representations of words/phrases/sentences)
 - ❑ **Basic Structuring** (i.e., phrase mining & transforming unstructured text into structured, typed entities/relationships)
 - ❑ **Advanced Structuring**: Discovering Hierarchies/taxonomies, exploring in multi-dimensional space



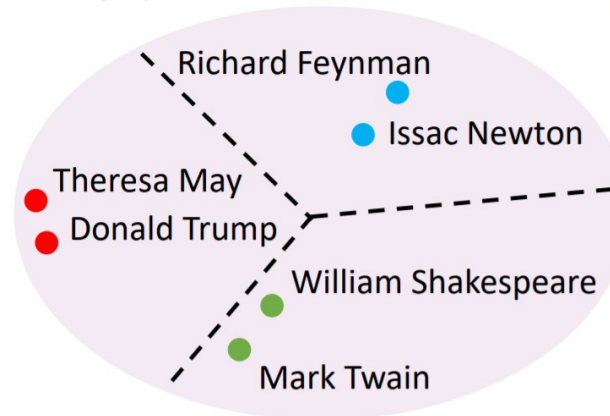
Text Representation: Embeddings & Language Models

- Word embeddings map words into a vector space which reflects semantic similarity

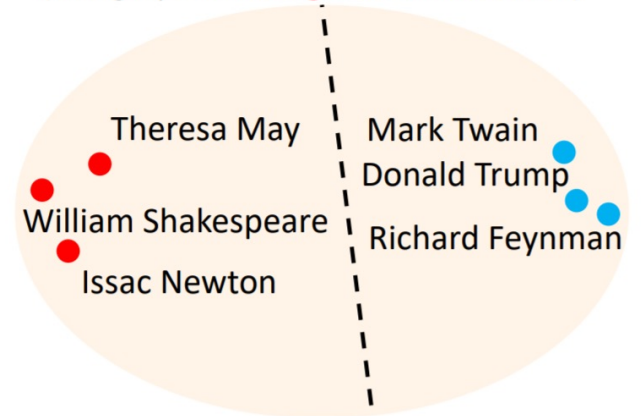


Unsupervised word embeddings:
learned from corpus statistics

Field Discriminative Embedding Space
(Category Name: **Politics**, **Science**, **Literature**)



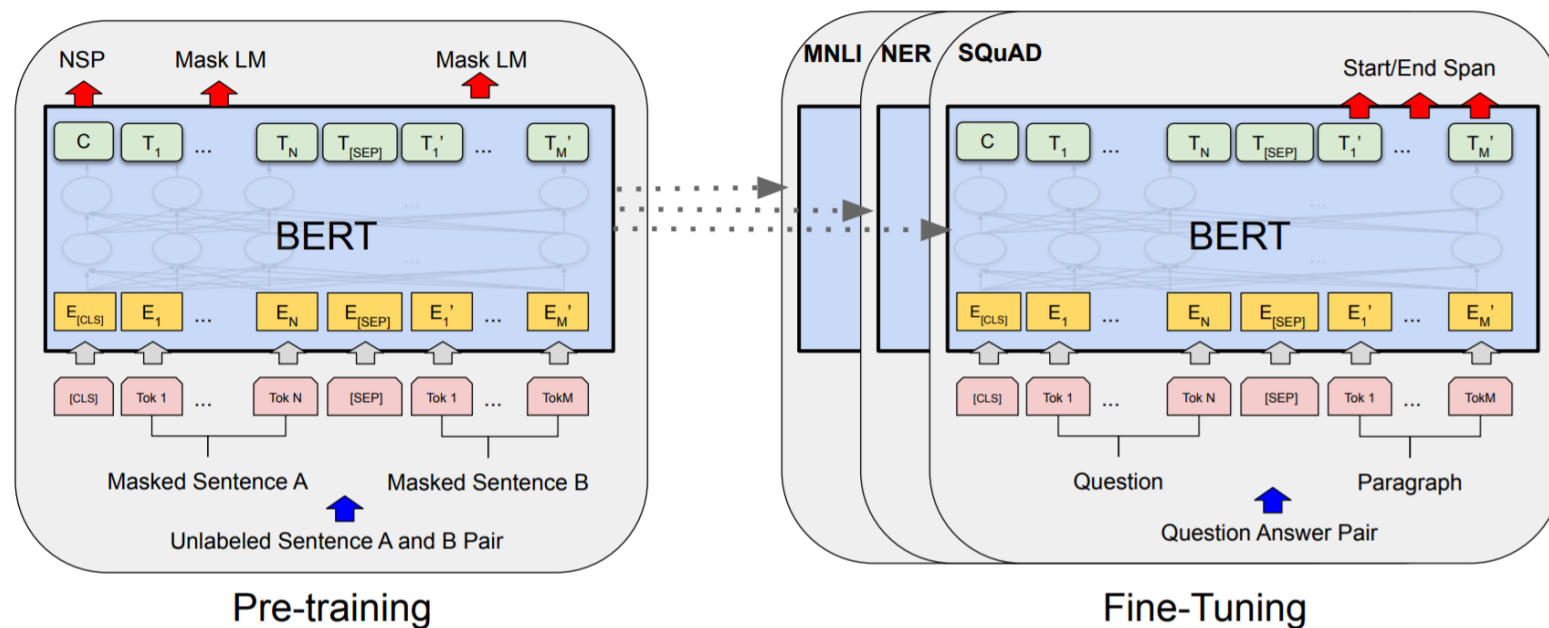
Location Discriminative Embedding Space
(Category Name: **England**, **United States**)



(Weakly-)supervised word embeddings:
learned from corpus statistics & user guidance

Text Representation: Embeddings & Language Models

- Language models are pre-trained on large-scale general-domain corpora to learn universal/generic language representations that can be transferred to downstream tasks via fine-tuning



Unsupervised/Self-supervised;
On large-scale general domain corpus

Task-specific supervision;
On target corpus

Basic Structuring: Phrase Mining and Information Extraction

Example: Finding “Interesting Hotel Collections”

The screenshot shows a search interface for New York City hotels. On the left, a sidebar titled "Collections" lists various categories: Walk to Penn Station (13), Times Square Views (9), Urban Oasis (12), Trendy Soho (11), Central Park Views (10), Art Deco Classic (12), Catch a Show (22), and Design Hotels (12). The "Catch a Show" collection is highlighted with a red box. Below the sidebar, there are sections for "Accommodation" (Hotels: 82, B&B and Inns: 45) and "Sort by" options (Just For You, Availability, Ranking, Price (low to high)). The main content area displays two hotel listings: Hyatt Times Square New York and Hilton Times Square, both with 2,576 reviews and "Great Location!" and "Loved our stay here" reviews.

Grouping hotels based on structured facts extracted from the review text

Different Dimensions of Information

Features for “Catch a Show” collection

- 1 Broadway shows
- 2 Beacon Theater
- 3 Broadway Dance Center
- 4 Broadway plays
- 5 David Letterman Show
- 6 Radio City Music Hall
- 7 Theatre shows

Features for “Near The High Line” collection

- 1 High Line Park
- 2 Chelsea Market
- 3 Highline Walkway
- 4 Elevated Park
- 5 Meatpacking District
- 6 West Side
- 7 Old Railway

Basic Structuring: Automated Named Entity Recognition & Typing

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 ...

Adv. Structuring: Multidimensional Nature of Texts

- ❑ The same document can naturally describe things across multiple dimensions
- ❑ Example:
 - ❑ A technical review may cover
 - ❑ Brands
 - ❑ Products
 - ❑ Aspects
 - ❑ Years
 - ❑ ...

Apple's 10th anniversary iPhone X sets a new gold standard for the next decade of iPhones. Coming hot on the heels of the iPhone 8 and iPhone 8 Plus, the iPhone X stole the show despite sharing nearly identical internal hardware. The X (pronounced "ten," like the Roman numeral) is a beautiful, modern sculpture, and iPhone owners finally have a reason to show off their phones again. As we're now about four months from Apple's next iPhone launch, we're revisiting the iPhone X to see if it's still worth the high price tag.

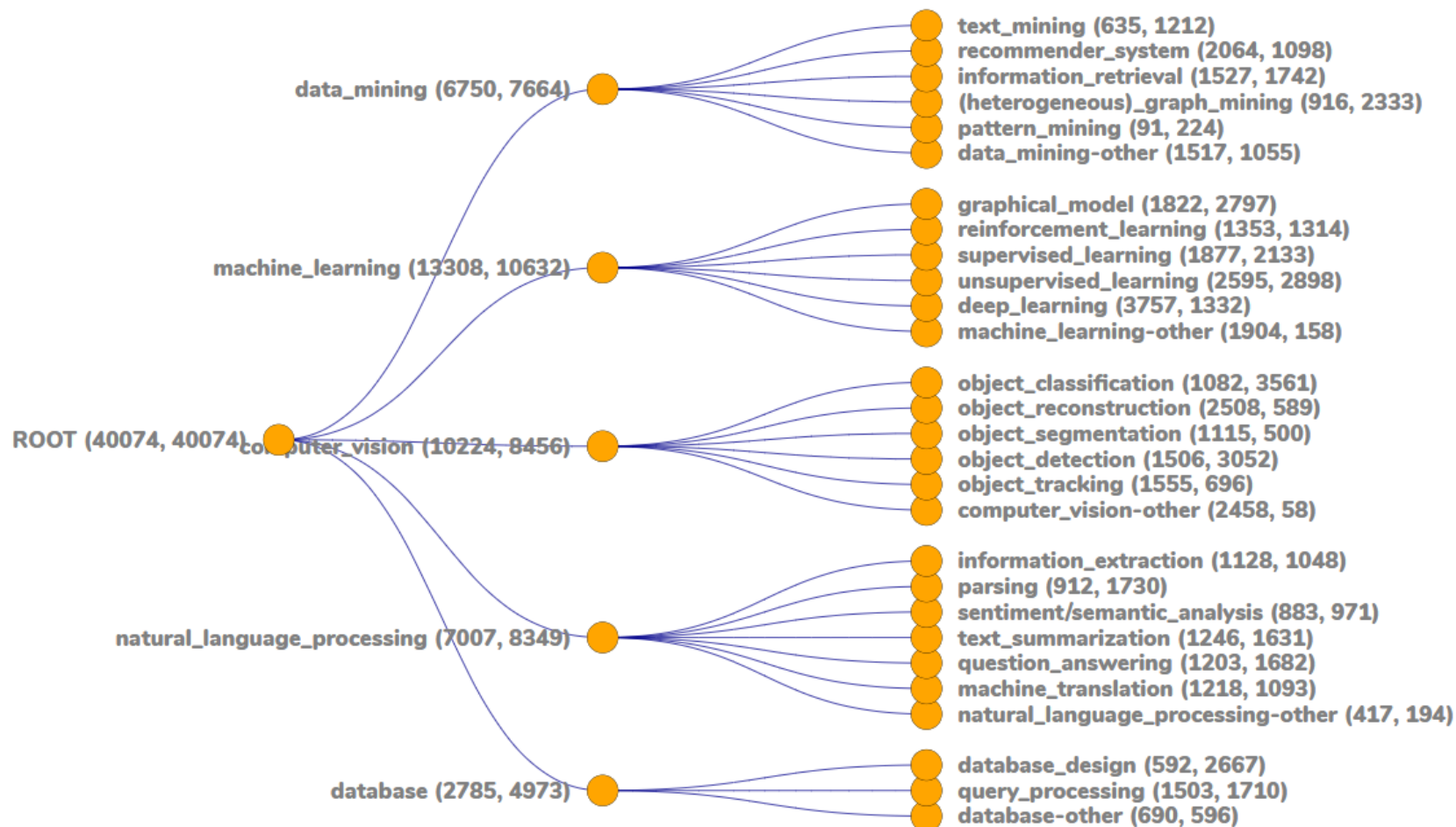
... ..

Advanced Structuring: Automatic Taxonomy Generation

Automatically Generated Taxonomy Visualization

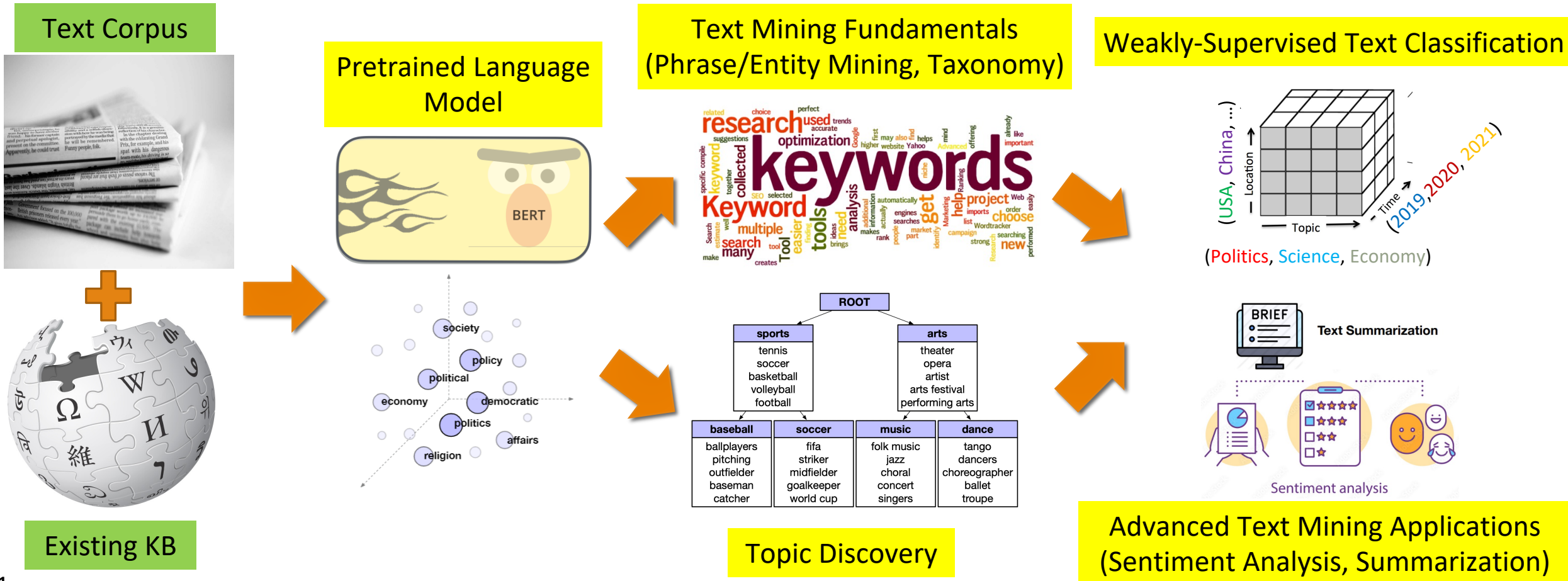
Current Selected: ROOT

Numbers in () from left to right represents the number of main papers and the number of secondary papers respectively.



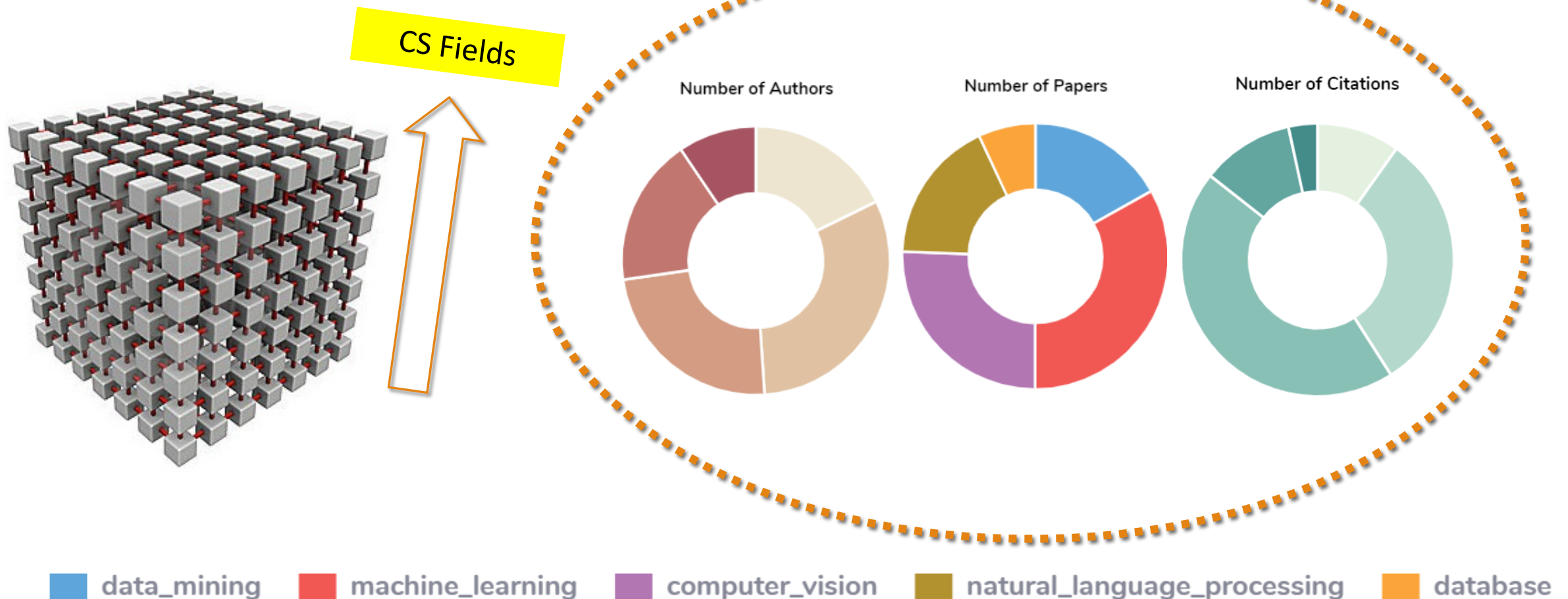
Adv. Structuring: Multi-Dimensional Text Cube Construction

- ❑ Understand and Extract Information from Massive Text Corpora
- ❑ Organize and Analyze Information using **Multidimensional** Text Analysis



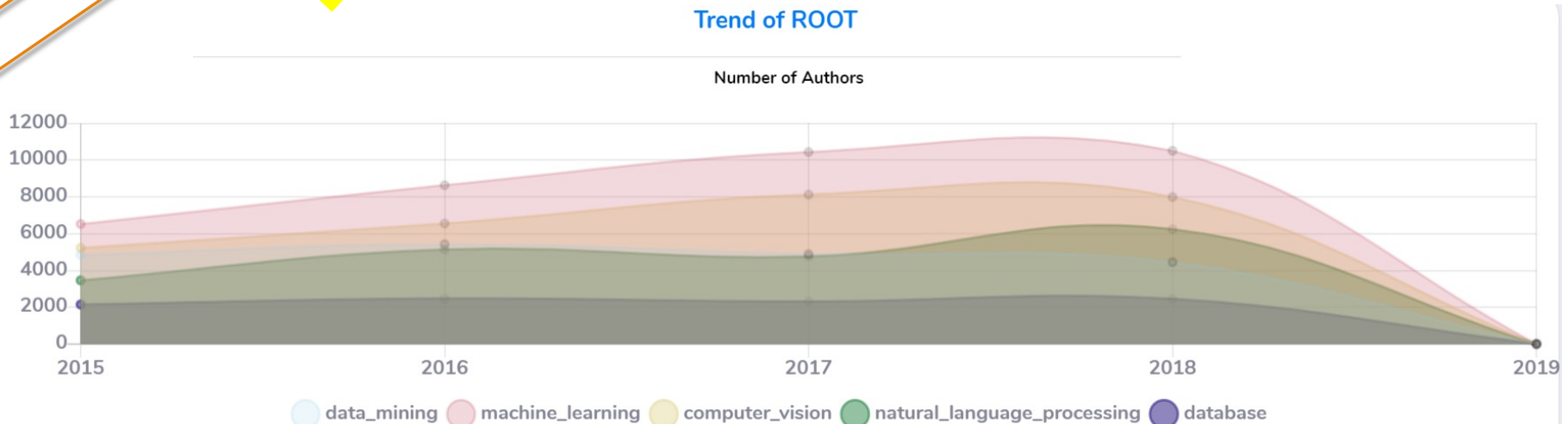
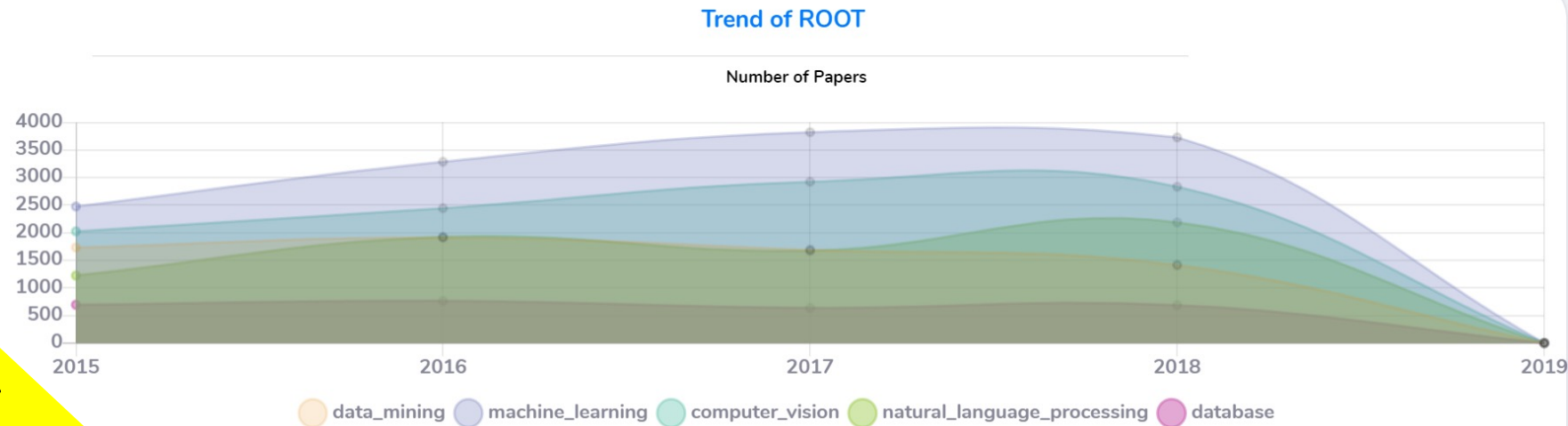
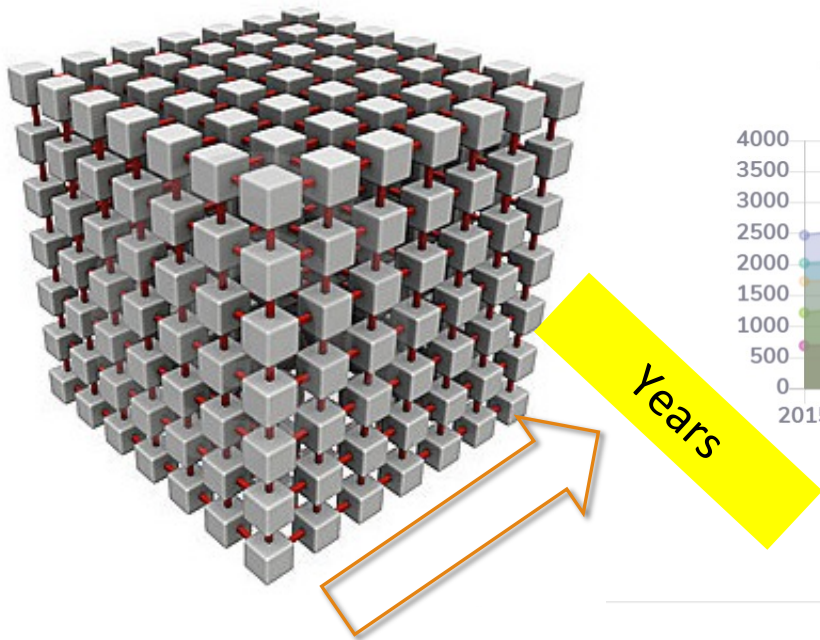
Application: DBLP—Automatic Paper Categorization

- Multidimensional text categorization and exploration across different CS fields



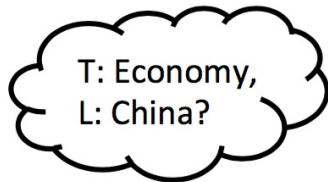
Application: DBLP—Trending Analysis

- Trending analysis on CS field development



Application: Comparative Summarization

Analyst Queries



(q₁)



(q₂)

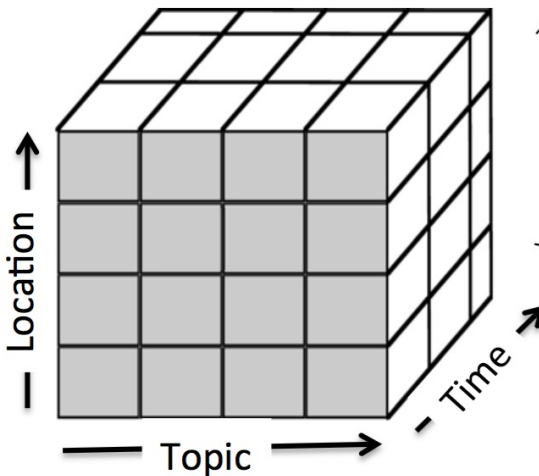
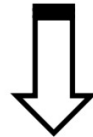
Multi-dimensional Text Cube



Topic

Location

Time



Representative Phrases

china's economy
the people's bank of china
trillion renminbi
growth target
fixed asset investment
local government debt
solar panel

massacre at sandy hook elementary
long island railroad
background check
senate armed services committee
adam lanza
buyback program
assault weapons and high capacity

Tutorial Outline

- ❑ Introduction
- ❑ Part I: Pretrained Language Models
- ❑ Part II: Revisiting Text Mining Fundamentals with Pretrained Language Models
- ❑ Part III: Embedding-Driven Topic Discovery
- ❑ Part IV: Weakly-Supervised Text Classification: Embeddings with Less Human Effort
- ❑ Part V: Advanced Text Mining Applications Empowered by Pretrained Embeddings
- ❑ Summary and Future Directions

Our Roadmap of This Tutorial

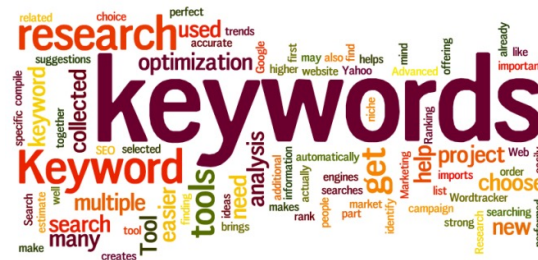
Text Corpus



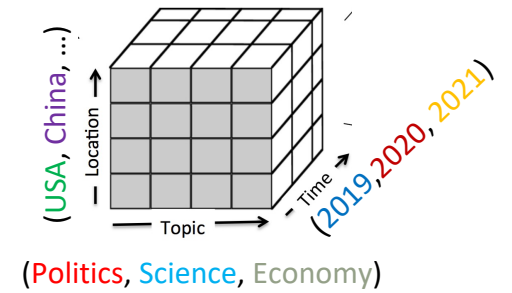
Part I: Pretrained Language Model



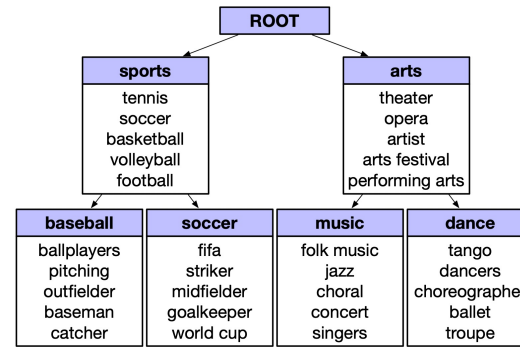
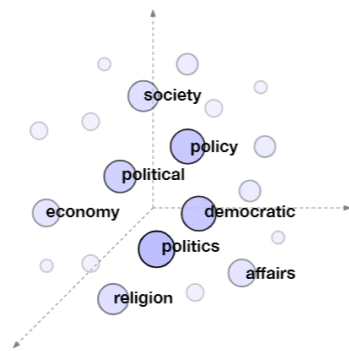
Part II: Text Mining Fundamentals (Phrase/Entity Mining, Taxonomy)



Part IV: Weakly-Supervised Text Classification



Existing KB



Part III: Topic Discovery

Part V: Advanced Text Mining Applications (Sentiment Analysis, Summarization)

