Mining Structures from Massive Texts by Exploring the Power of Pre-trained Language Models

Yu Zhang, Yunyi Zhang, and Jiawei Han

Department of Computer Science

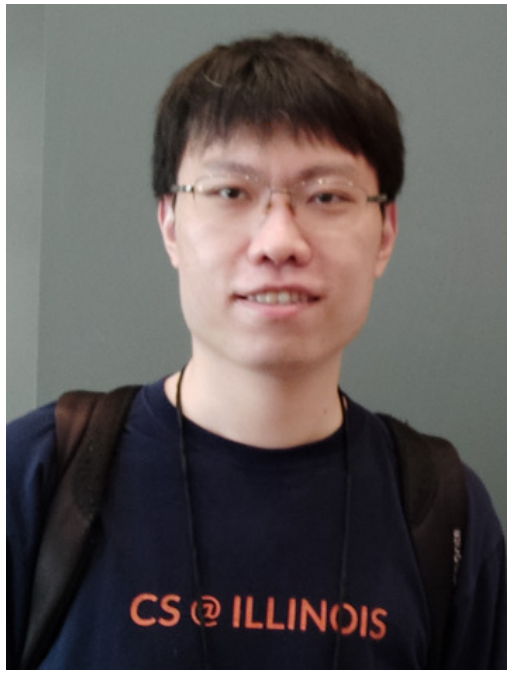
University of Illinois at Urbana-Champaign

Mar 29, 2023

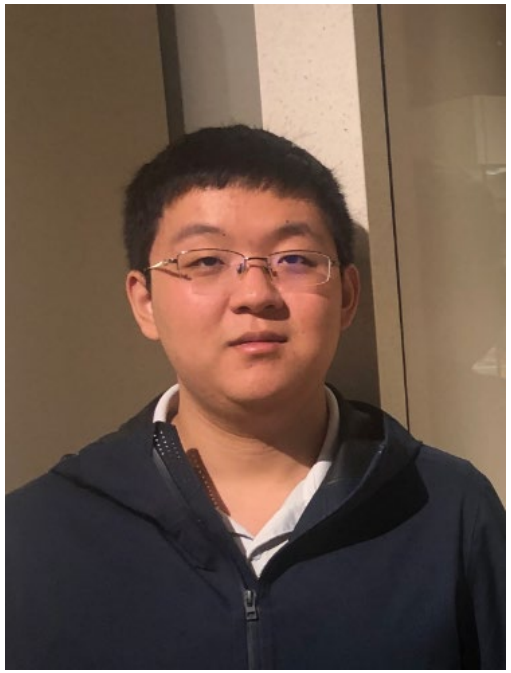
Estimated Timeline for This Tutorial

- Introduction:
  5 mins (11:00-11:05)
- Part I: Pre-trained Language Models:
  35 mins (11:05-11:40)
- Part II: Mining Topic Structures: Unsupervised and Seed-Guided Topic Discovery:
  35 mins (11:40-12:15)
- Part III: Mining Document Structures: Weakly Supervised Text Classification:
  35 mins (12:15-12:30, Break, 16:00-16:20)
- Part IV: Mining Entity Structures: Taxonomy and Knowledge Base Construction:
  60 mins (16:20-17:20)
- Towards an Integrated Information Processing Paradigm:
  10 mins (17:20-17:30)

About Instructors



- Yu Zhang
- Ph.D. Candidate at UIUC



- Yunyi Zhang
- Ph.D. Candidate at UIUC

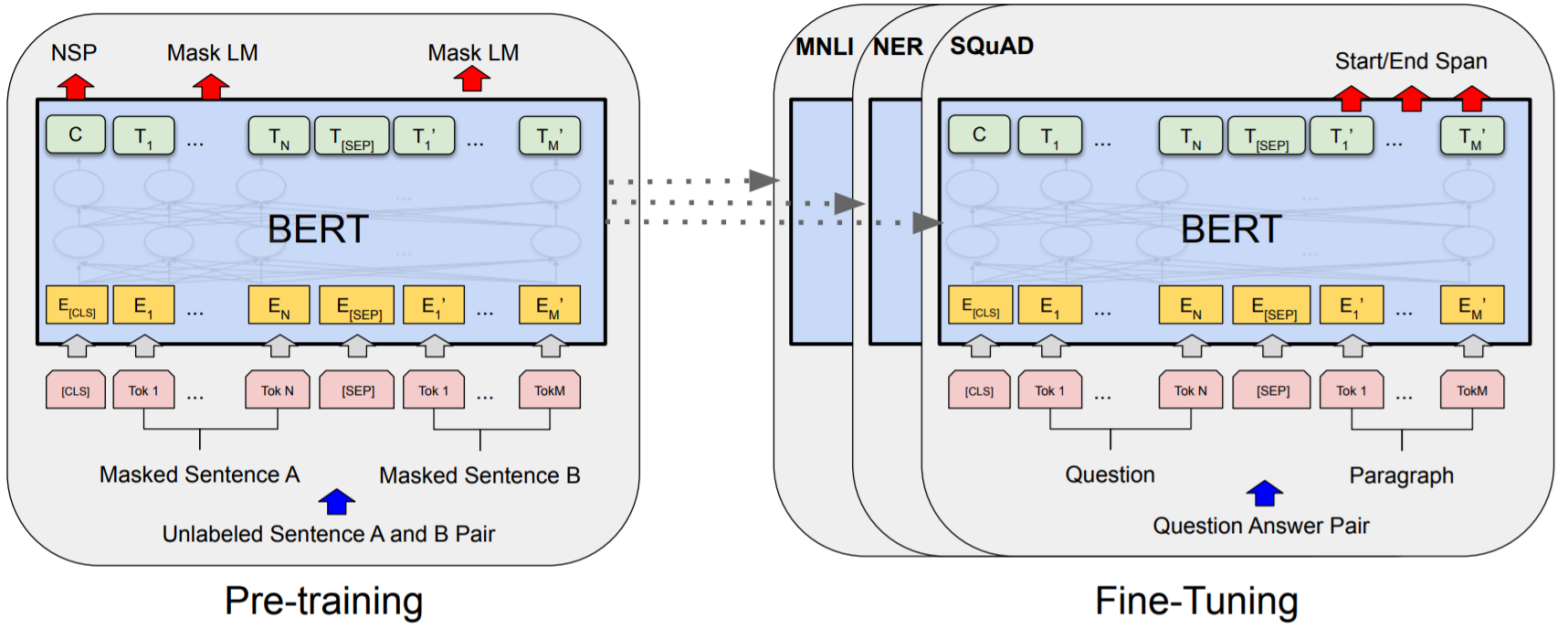
- Jiawei Han
- Michael Aiken Chair Professor at UIUC

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 such big data systematically?
  - **Representing Text** (i.e., computing vector representations of words/phrases/sentences)
  - **Mining Topic Structures** (i.e., topic discovery)
  - **Mining Document Structures** (i.e., text classification)
  - **Mining Entity Structures** (i.e., phrase mining, entity typing, taxonomy construction, relation extraction)

Pre-trained Language Models (PLMs)

- 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

Mining Topic Structures: Topic Discovery

- **Input**: (1) A large corpus. (2) User specifies a set of **seeds**.

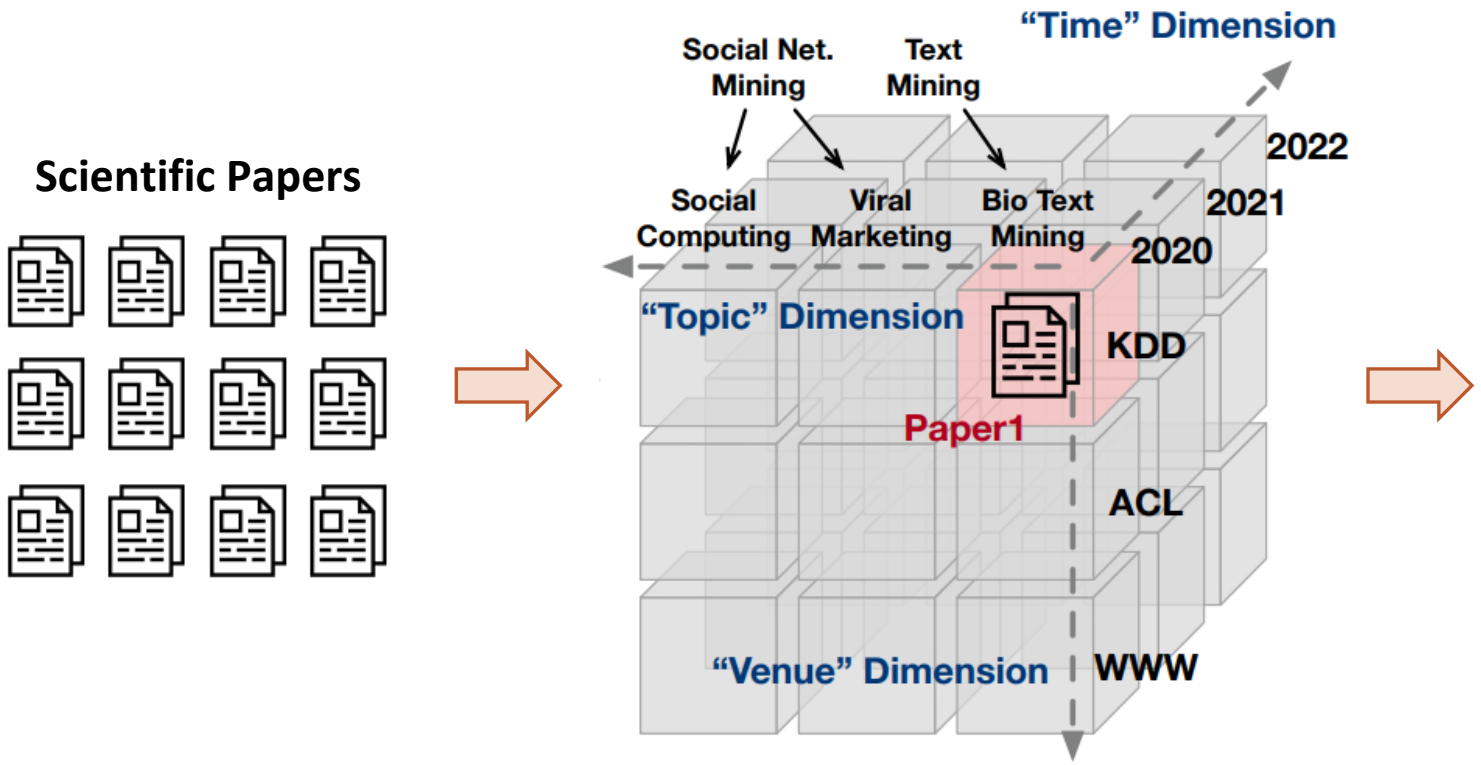
mathematics | physics | computer science

- **Output**: Find a set of terms under each category to form a coherent topic.

mathematics	physics	computer science
algebra geometry calculus …	mechanics electromagnetism optics …	data mining operating system mobile computing …

Mining Document Structures: Text Classification

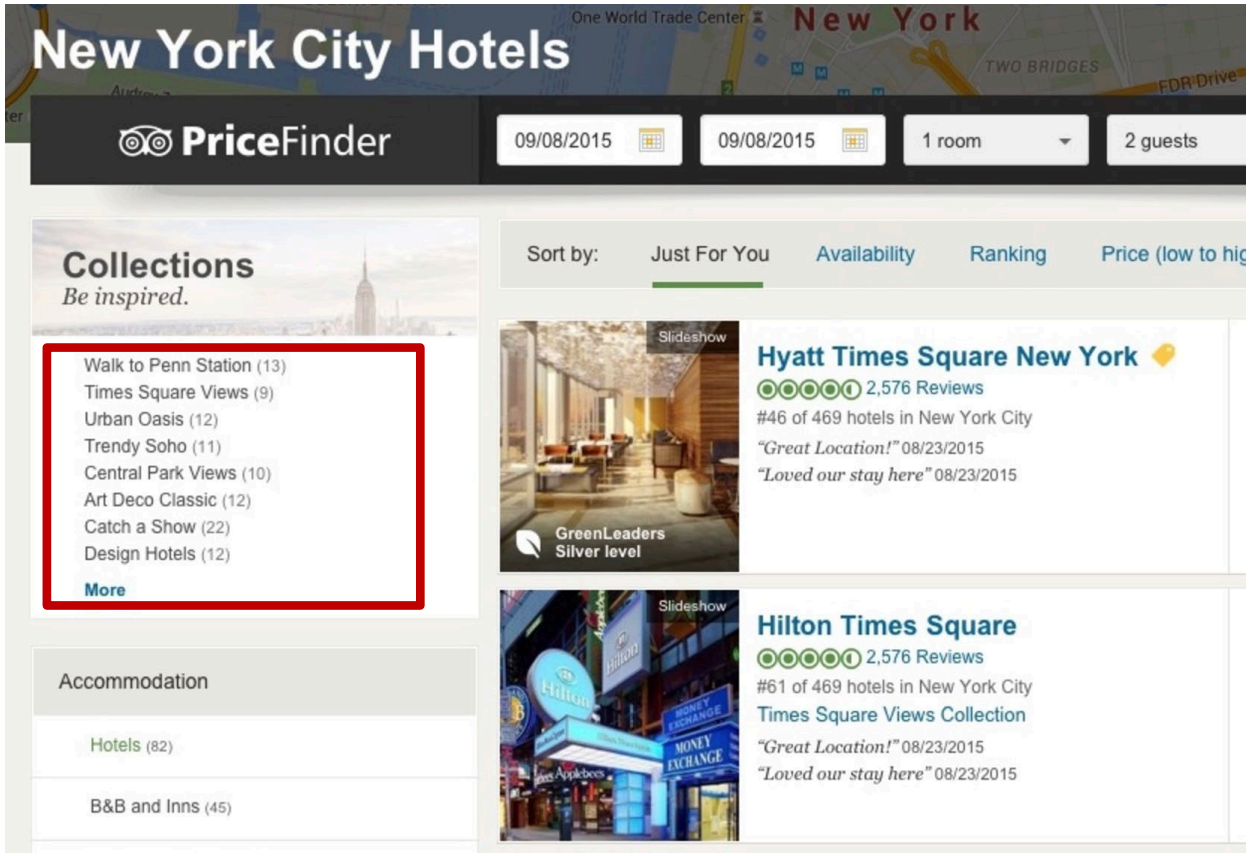
- Organize documents according to multiple criteria (e.g., topic, time) so that users can track their interested information instead of drowning in the whole literature.



- **Find the most recent papers on Large Language Models.**
- **Analyze the trend of the volume of COVID-19 papers.**
- **Explore the topic shift of the WWW conference.**

Mining Entity Structures: Phrase Mining

Grouping hotels based on structured facts extracted from the review text



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

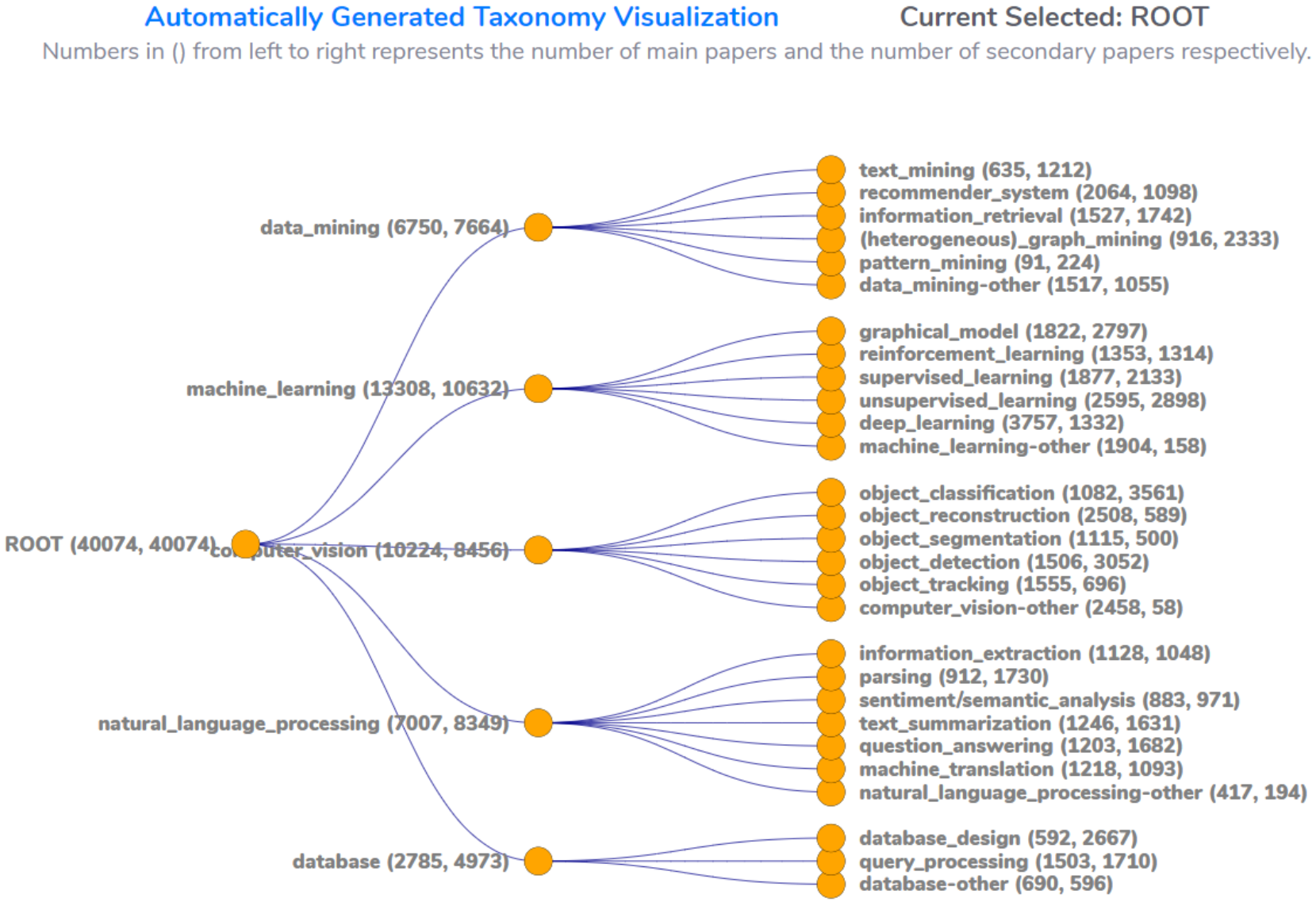
Mining Entity Structures: Entity Typing

Automatic Recognition of 75 entity types in COVID-19 Biomedical 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 ….

Mining Entity Structures: Taxonomy Construction

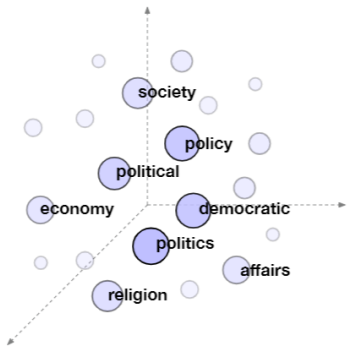


Our Roadmap of This Tutorial

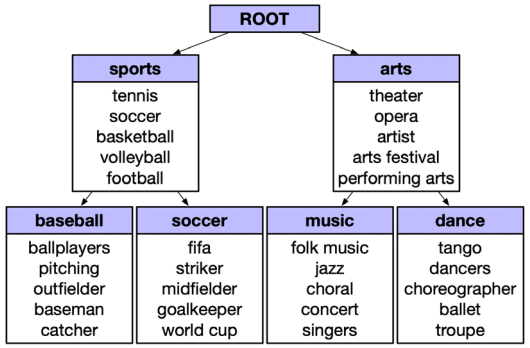
Text Corpus

Existing KB

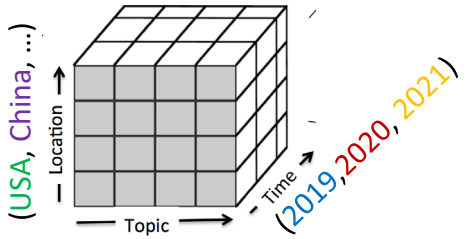
Part I: Pretrained Language Models



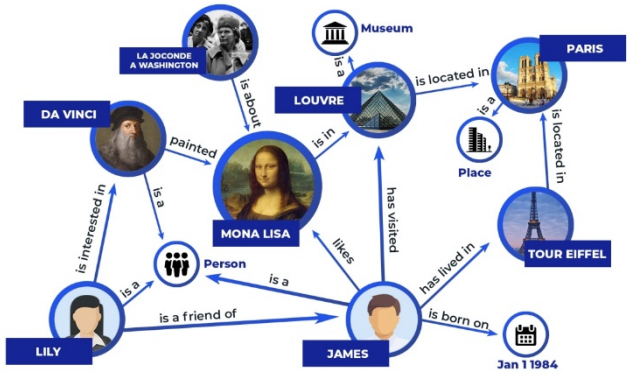
Part II: Mining Topic Structures



Part III: Mining Document Structures



(Politics, Science, Economy)



Part IV: Mining Entity Structures

Tutorial Outline

- Introduction:
  5 mins (11:00-11:05)
- Part I: Pre-trained Language Models:
  35 mins (11:05-11:40)
- Part II: Mining Topic Structures: Unsupervised and Seed-Guided Topic Discovery:
  35 mins (11:40-12:15)
- Part III: Mining Document Structures: Weakly Supervised Text Classification:
  35 mins (12:15-12:30, Break, 16:00-16:20)
- Part IV: Mining Entity Structures: Taxonomy and Knowledge Base Construction:
  60 mins (16:20-17:20)
- Towards an Integrated Information Processing Paradigm:
  10 mins (17:20-17:30)