# Pretrained Language Representations for Text Understanding: A Weakly-Supervised Perspective

**Yu Meng, Jiaxin Huang, Yu Zhang, Yunyi Zhang, Jiawei Han**

Department of Computer Science
University of Illinois Urbana-Champaign

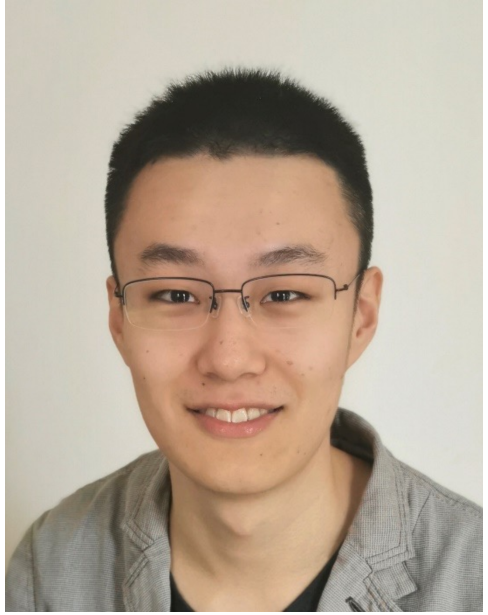
Aug 9, 2023

Tutorial Website:

# Estimated Timeline for This Tutorial

- ❑ Introduction: **15 mins (10:00-10:15 Jiawei Han)**
- ❑ Part I: Language Foundation Models for Text Analysis: **35 mins (10:15-10:50 Yu Meng)**
- ❑ Part II: Embedding-Driven Topic Discovery: **35 mins (10:50-11:25 Jiaxin Huang)**
- ❑ **Break: 10 mins (11:25-11:35)**
- ❑ Part III: Weakly-Supervised Text Classification: Embeddings with Less Human Effort: **35 mins (11:35-12:10 Yu Zhang**)
- ❑ Part IV: Language Models for Knowledge Base Construction: **35 mins (12:10-12:45 Jiawei Han)**
- ❑ Part V: Summary & Looking Forward: Advanced Text Mining Applications: **15 mins (12:45-13:00 Jiawei Han)**

# About Instructors



- Yu Meng
- Ph.D. Candidate @ UIUC
- Incoming Asst. Prof. @ UVA CS
- Google PhD Fellowship (2021)

- Jiaxin Huang
- Ph.D. Candidate @ UIUC
- Incoming Asst. Prof. @ WashU CSE
- Microsoft PhD Fellowship (2021)

- Yu Zhang
- Ph.D. Candidate @ UIUC
- Dissertation Completion Fellowship (2023)
- Yunni and Maxine Pao Memorial Fellowship (2022)

- Yunyi Zhang
- Ph.D. Candidate @ UIUC

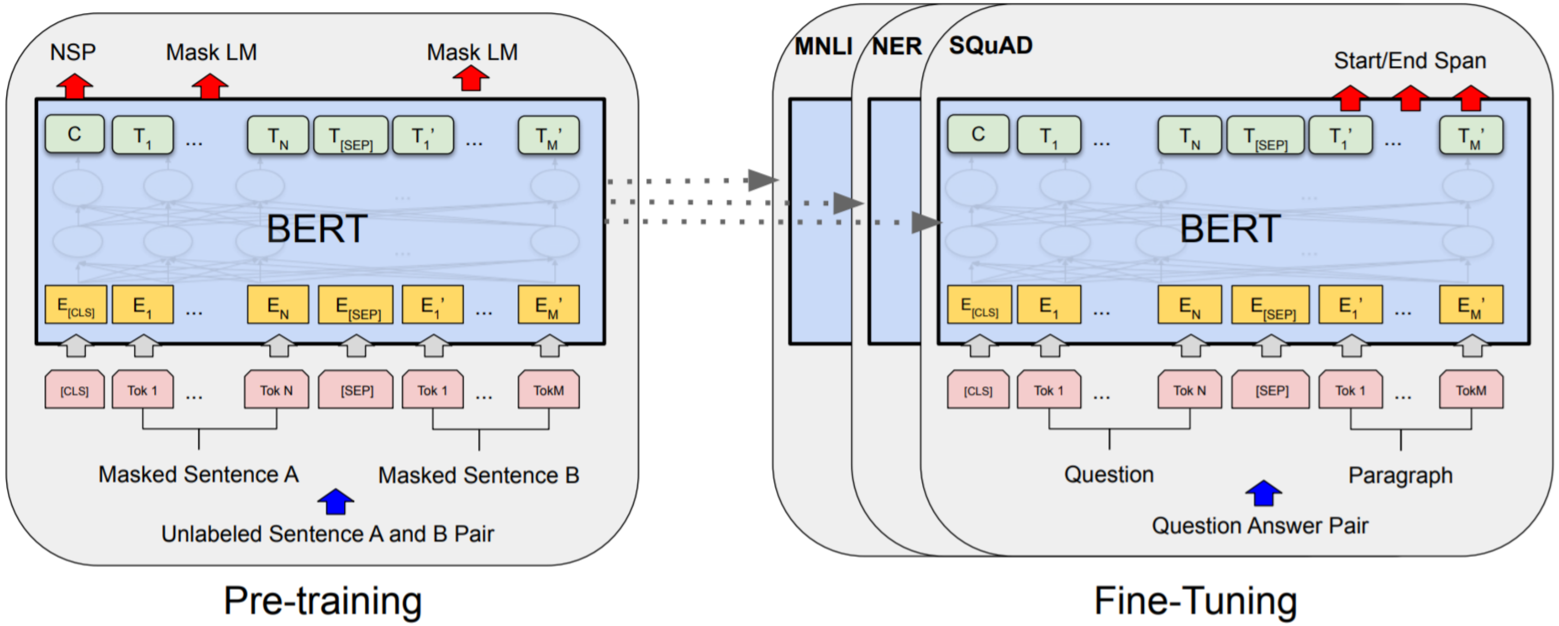
- Jiawei Han
- Michael Aiken Chair Professor @ UIUC
- ACM SIGKDD Innovation Award Winner (2004)

# Over 80% of Big (Web) 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., phase mining & transforming unstructured text into structured, typed entities/relationships)
  - **Advanced Structuring**: Discovering Hierarchies/taxonomies, exploring in multi-dimensional space

# Foundation for Text Analysis: 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

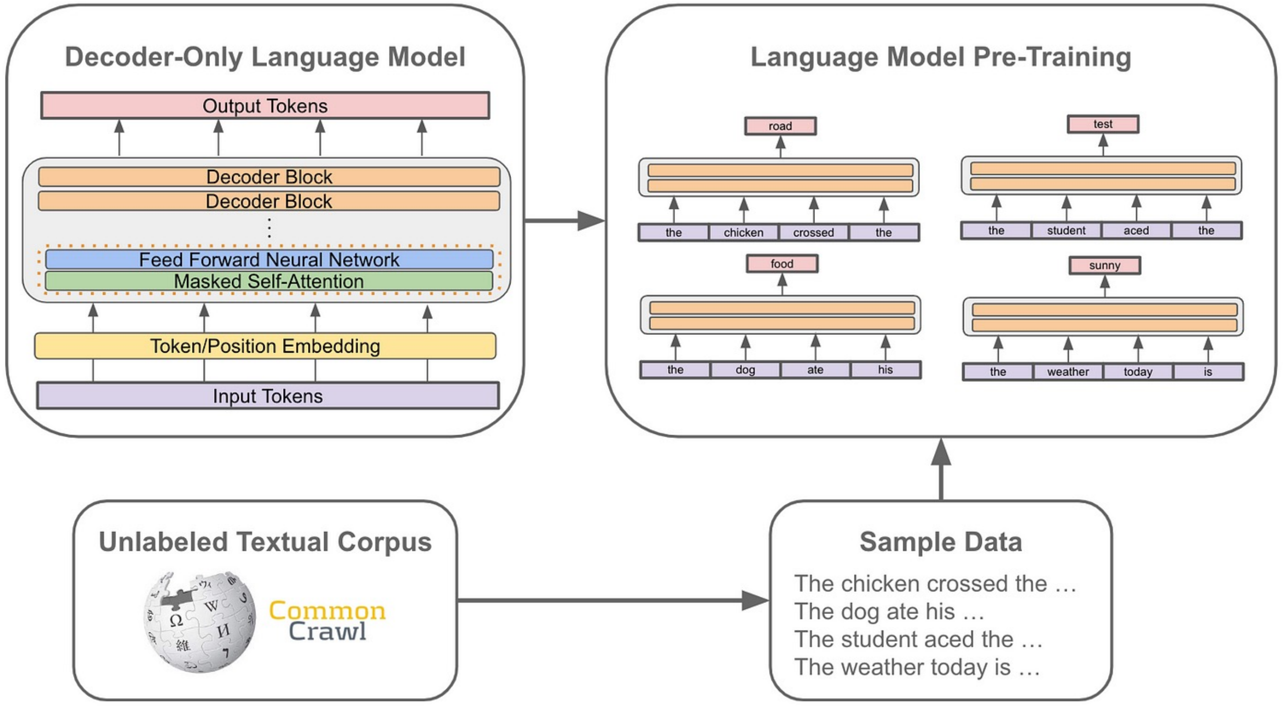


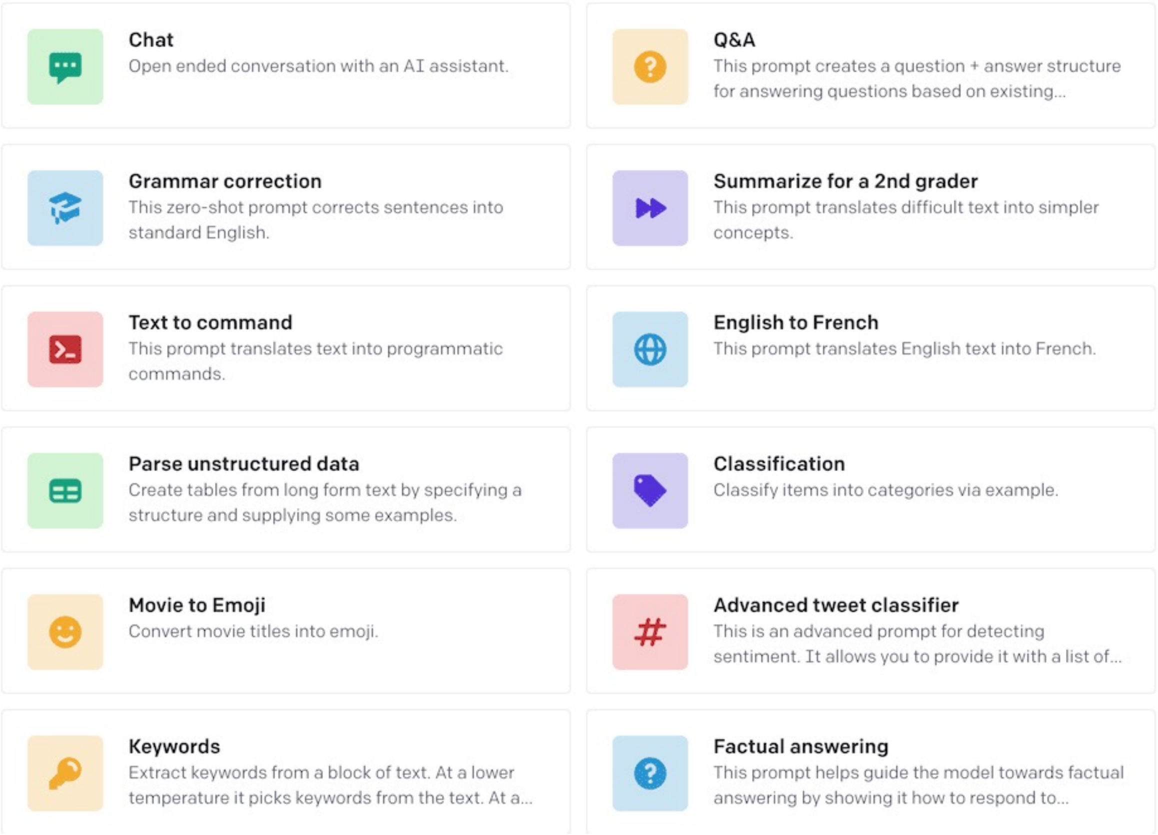
Unsupervised/Self-supervised;
On large-scale general domain corpus

Task-specific supervision;
On target corpus

# Generative Large Language Models: The GPT Series

- ❑ GPT models: Large language models (LLMs) trained for text generation
- ❑ Applicable to a wide range of tasks





**Chat**
Open ended conversation with an AI assistant.

**Grammar correction**
This zero-shot prompt corrects sentences into standard English.

**Text to command**
This prompt translates text into programmatic commands.

**Parse unstructured data**
Create tables from long form text by specifying a structure and supplying some examples.

**Movie to Emoji**
Convert movie titles into emoji.

**Keywords**
Extract keywords from a block of text. At a lower temperature it picks keywords from the text. At a...

**Q&A**
This prompt creates a question + answer structure for answering questions based on existing...

**Summarize for a 2nd grader**
This prompt translates difficult text into simpler concepts.

**English to French**
This prompt translates English text into French.

**Classification**
Classify items into categories via example.

**Advanced tweet classifier**
This is an advanced prompt for detecting sentiment. It allows you to provide it with a list of...

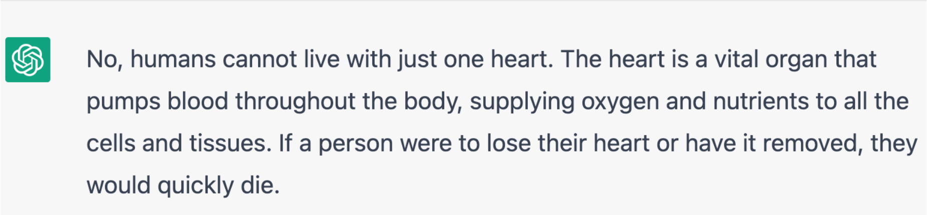
**Factual answering**
This prompt helps guide the model towards factual answering by showing it how to respond to...

# Challenges of Large Language Models

- Not factually guaranteed: May generate wrong information

Can humans live with one heart?

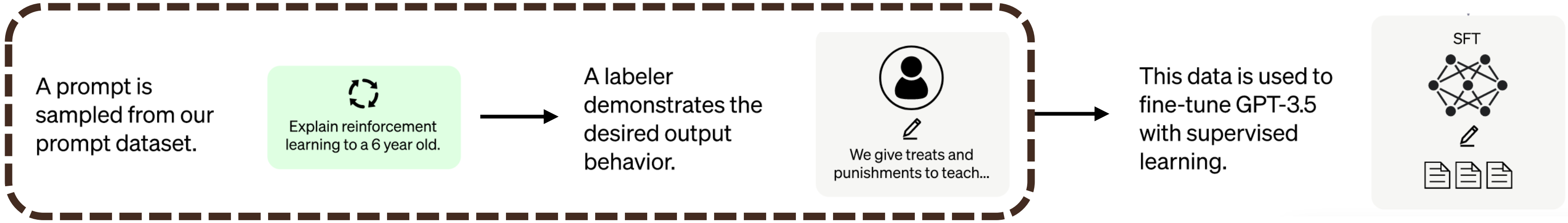
**ChatGPT:**



No, humans cannot live with just one heart. The heart is a vital organ that pumps blood throughout the body, supplying oxygen and nutrients to all the cells and tissues. If a person were to lose their heart or have it removed, they would quickly die.

- Heavy supervision required: Trained on massive annotated data

**Manually annoated:**



A prompt is sampled from our prompt dataset.

Explain reinforcement learning to a 6 year old.

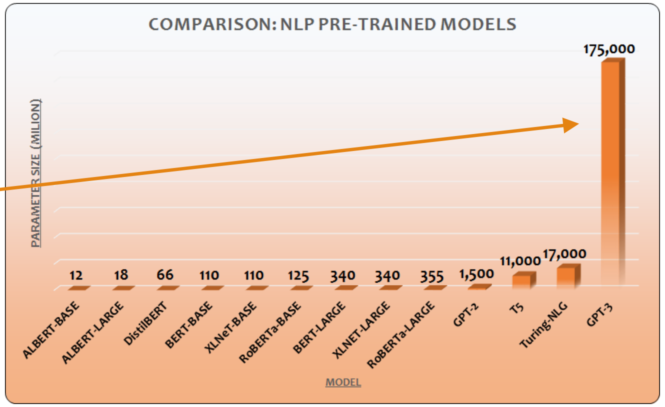
A labeler demonstrates the desired output behavior.

We give treats and punishments to teach...

This data is used to fine-tune GPT-3.5 with supervised learning.

- Costly & Inefficient: Too large to be used in many applications

**GPT3 has 175B parameters (ChatGPT/GPT-4 may have more!)**



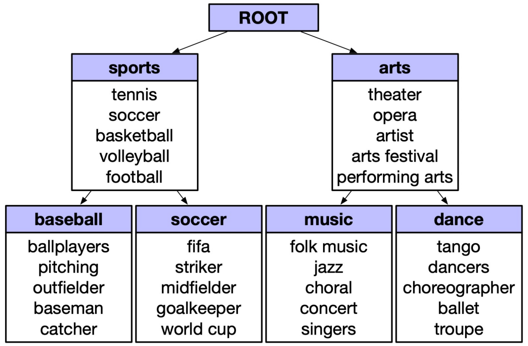
# Towards Factual, Automatic, and Efficient Text Mining

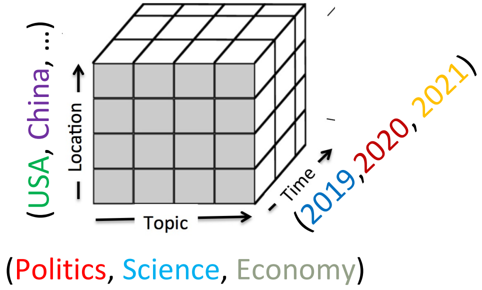
- ❑ Understand and Extract Information from Massive Text Corpora
- ❑ Organize and analyze texts in a **factual, efficient and effective** way

Text Corpus

Language Foundation Models

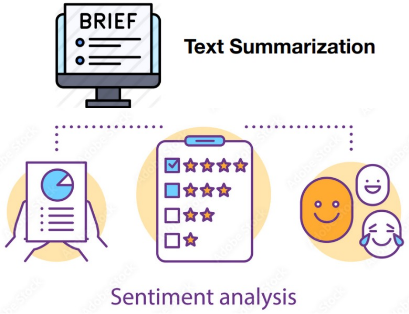
Topic Discovery





Weakly-Supervised Text Classification

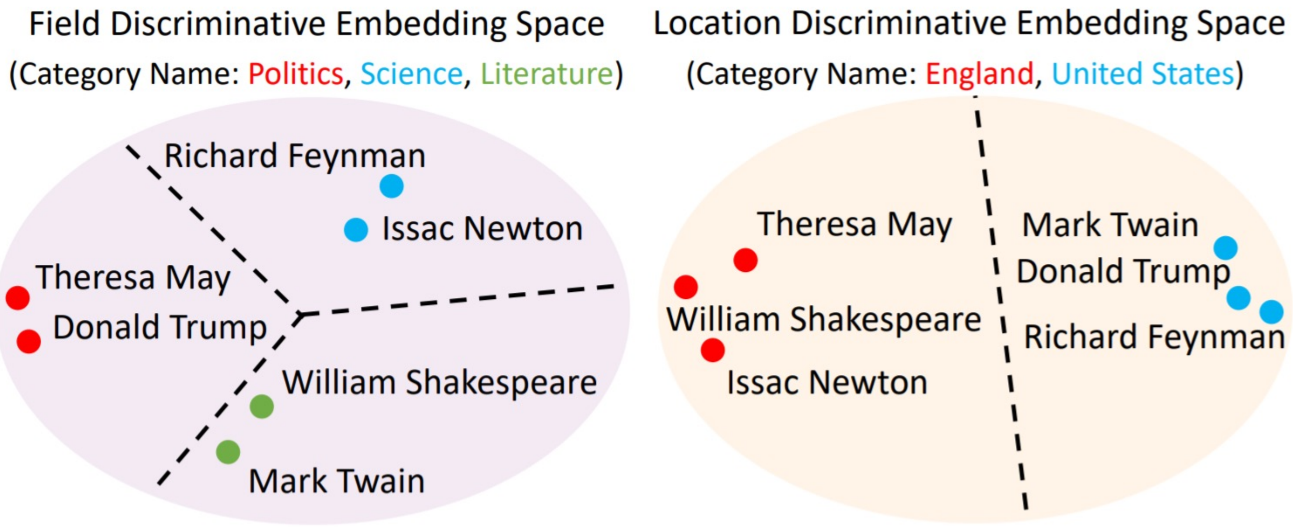
Knowledge Base Construction (Entity, Relation & Event)



Advanced Text Mining Applications

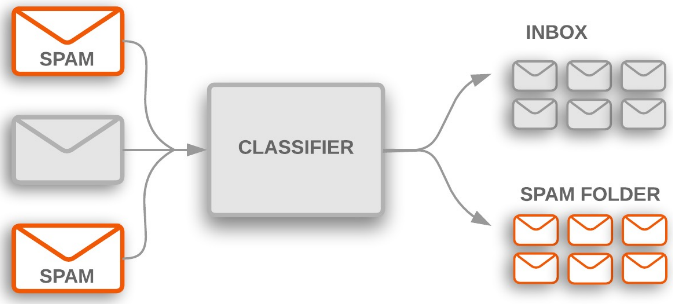
# Overview of Seed-Guided Topic Discovery

- Mining topic structures from massive corpora is crucial for text understanding
- The same set of concepts/topics/entities may be organized via different aspects
- How to incorporate user interests/preferences?
  - Manually labeling documents requires non-trivial human efforts and is hard to scale
  - Use seed words instead to guide topic discovery!


Field Discriminative Embedding Space
(Category Name: Politics, Science, Literature)
Richard Feynman
Issac Newton
Theresa May
Donald Trump
William Shakespeare
Mark Twain
Location Discriminative Embedding Space
(Category Name: England, United States)
Theresa May
William Shakespeare
Issac Newton
Mark Twain
Donald Trump
Richard Feynman

# Overview of Weakly-Supervised Text Classification

- Text classification is a core task for document organization and understanding
- Text classifiers are typically trained on massive manually-labeled data
- How to build text classifiers with fewer human annotations?
- Weakly-supervised text classification: Use label names & keywords as weak supervision



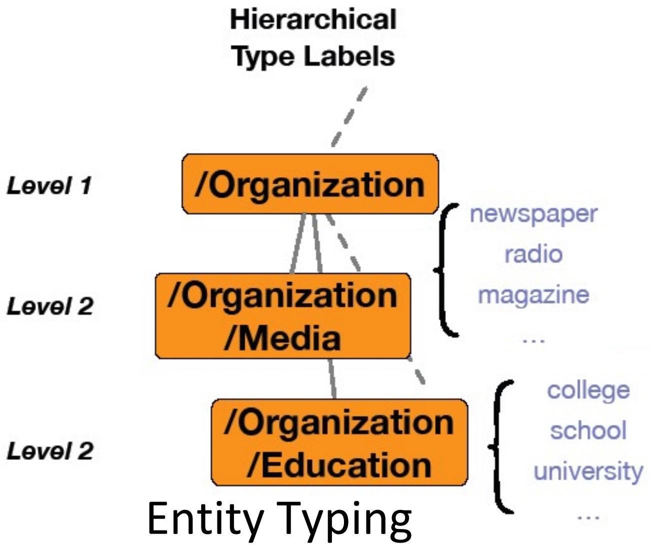
Text classifiers

The picture can't be displayed.

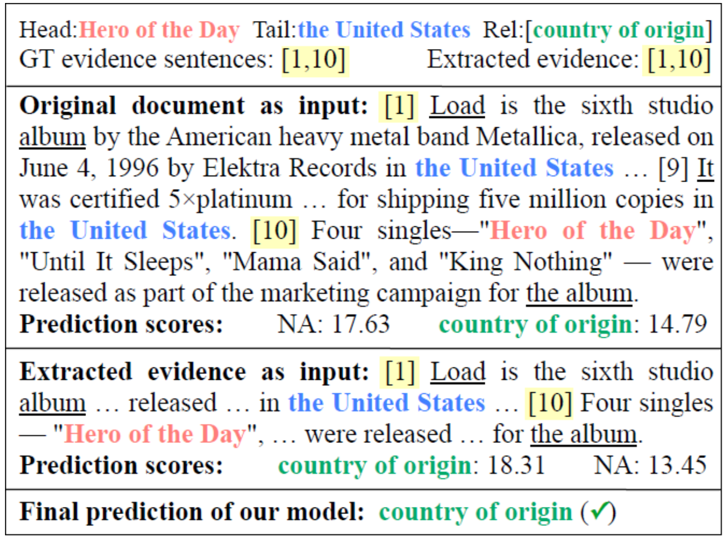
Weakly-supervised text classification only leverages label names as supervision

# Overview of Knowledge Base Construction

Phrase Mining



Entity Typing



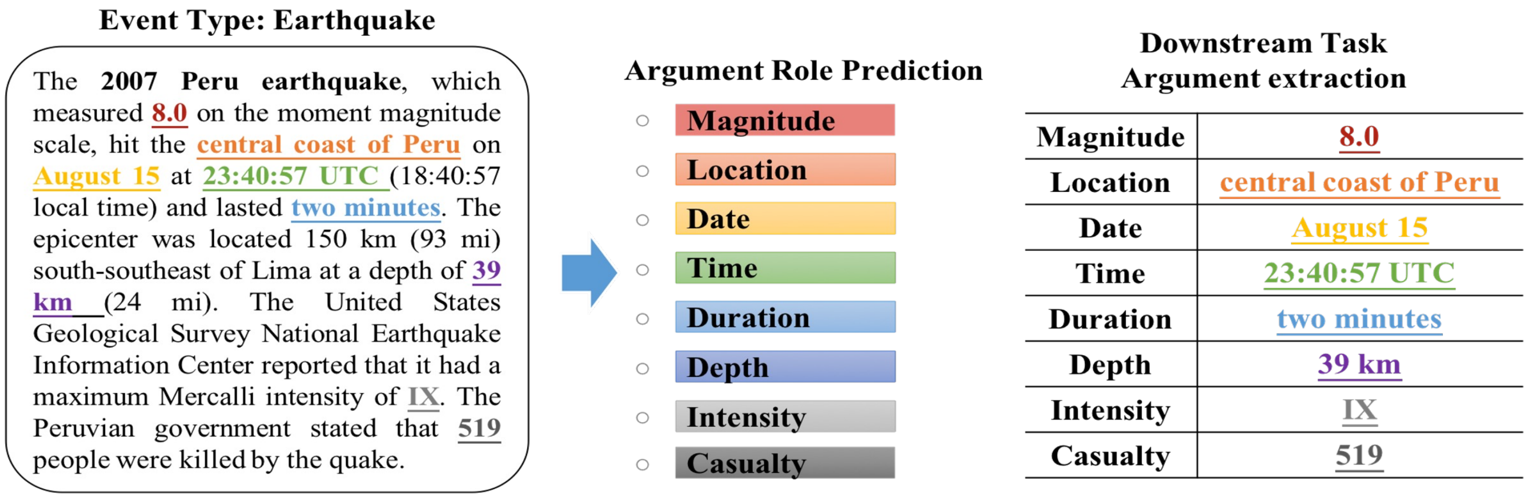
Head:Hero of the Day Tail:the United States Rel:[country of origin]
GT evidence sentences: [1,10] Extracted evidence: [1,10]

**Original document as input:** [1] Load is the sixth studio album by the American heavy metal band Metallica, released on June 4, 1996 by Elektra Records in the United States … [9] It was certified 5×platinum … for shipping five million copies in the United States. [10] Four singles—"Hero of the Day", "Until It Sleeps", "Mama Said", and "King Nothing" — were released as part of the marketing campaign for the album.
**Prediction scores:** NA: 17.63 country of origin: 14.79

**Extracted evidence as input:** [1] Load is the sixth studio album … released … in the United States … [10] Four singles — "Hero of the Day", … were released … for the album.
**Prediction scores:** country of origin: 18.31 NA: 13.45

**Final prediction of our model:** country of origin (✓)

Relation Extraction



**Event Type: Earthquake**

The **2007 Peru earthquake**, which measured 8.0 on the moment magnitude scale, hit the central coast of Peru on August 15 at 23:40:57 UTC (18:40:57 local time) and lasted two minutes. The epicenter was located 150 km (93 mi) south-southeast of Lima at a depth of 39 km (24 mi). The United States Geological Survey National Earthquake Information Center reported that it had a maximum Mercalli intensity of IX. The Peruvian government stated that 519 people were killed by the quake.

**Argument Role Prediction**

- ○ Magnitude
- ○ Location
- ○ Date
- ○ Time
- ○ Duration
- ○ Depth
- ○ Intensity
- ○ Casualty

**Downstream Task Argument extraction**

| | |
|---|---|
| Magnitude | 8.0 |
| Location | central coast of Peru |
| Date | August 15 |
| Time | 23:40:57 UTC |
| Duration | two minutes |
| Depth | 39 km |
| Intensity | IX |
| Casualty | 519 |

Event Extraction

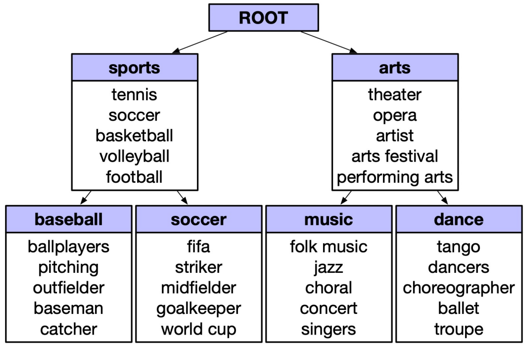
# Tutorial Outline

- Introduction
- Part I: Language Foundation Models for Text Analysis
- Part II: Embedding-Driven Topic Discovery
- Part III: Weakly-Supervised Text Classification
- Part IV: Language Models for Knowledge Base Construction
- Part V: Summary & Looking Forward: Advanced Text Mining Applications

# Our Roadmap of This Tutorial

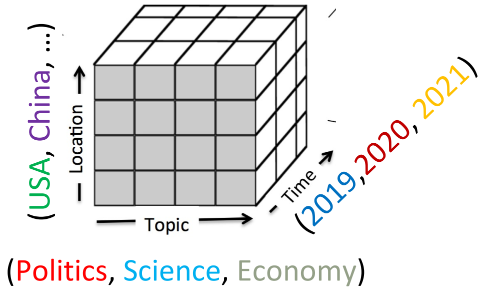
Text Corpus

Part I: Language Foundation Models



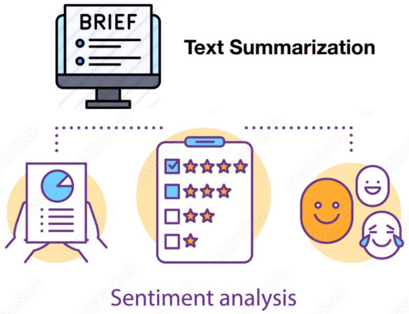
- ROOT
  - sports: tennis, soccer, basketball, volleyball, football
    - baseball: ballplayers, pitching, outfielder, baseman, catcher
    - soccer: fifa, striker, midfielder, goalkeeper, world cup
  - arts: theater, opera, artist, arts festival, performing arts
    - music: folk music, jazz, choral, concert, singers
    - dance: tango, dancers, choreographer, ballet, troupe

Part II: Topic Discovery



Part III: Weakly-Supervised Text Classification

Part IV: Knowledge Base Construction (Entity, Relation & Event)



Part V: Advanced Text Mining Applications