Bridging Text Data and Graph Data: Towards Semantics and Structure-aware Knowledge Discovery

Bowen Jin, Yu Zhang, Sha Li, Jiawei Han
Department of Computer Science
University of Illinois at Urbana-Champaign
Mar 4, 2024

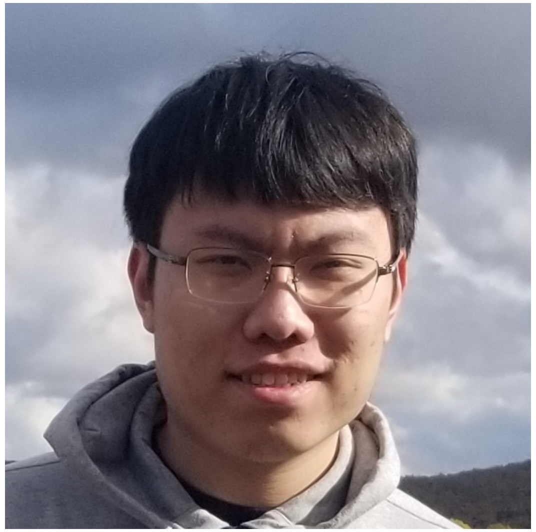
Tutorial Website:

Estimated Timeline for This Tutorial

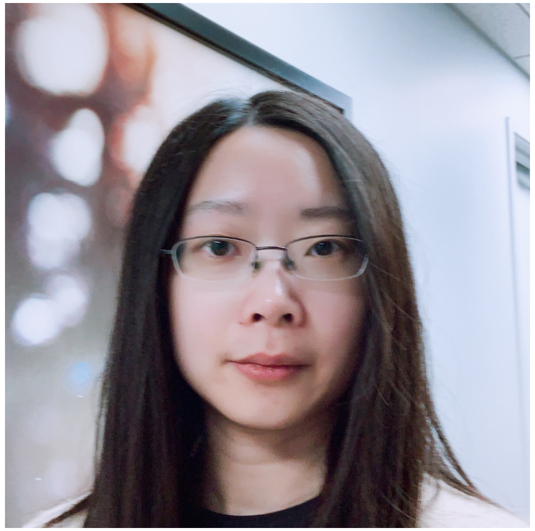
- Introduction: **15 mins (8:30 - 8:45 Bowen Jin)**
- Part I: Enhancing Text with Graph Structure: **45 mins (8:45 - 9:30 Sha Li)**
- Part II: Graph Mining with Large Language Models: **45 mins (9:30 - 10:15 Bowen Jin)**
- **Break: 15 mins (10:15 - 10:30)**
- Part III: Text Mining with Structured Information: **45 mins (10:30 - 11:15 Yu Zhang)**
- Part IV: Summary & Looking Forward: **15 mins (11:15 – 11:30 Bowen Jin)**

About Instructors

- Bowen Jin
- Ph.D. Candidate @ UIUC
- Apple PhD Fellowship (2024)



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



- Sha Li
- Ph.D. Candidate @ UIUC

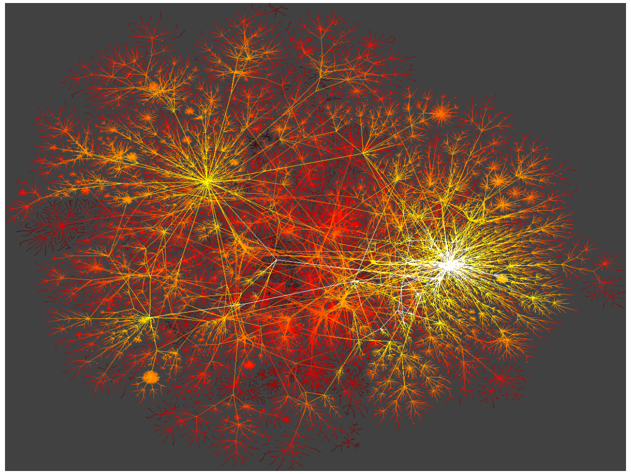
- Jiawei Han
- Michael Aiken Chair Professor @ UIUC
- ACM SIGKDD Innovation Award Winner (2024)

Over 80% of Big Data is 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

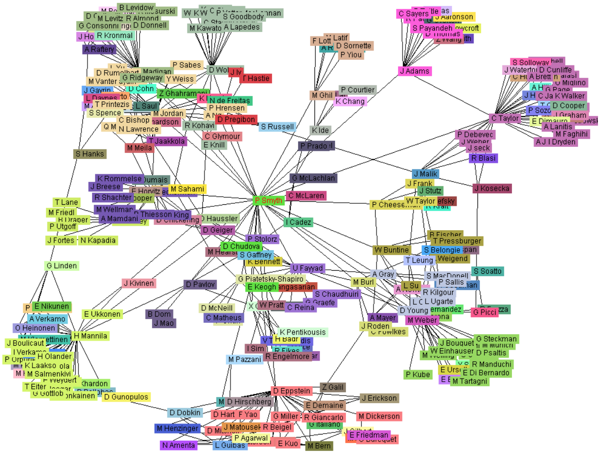
Graphs are Ubiquitous

- Graphs and substructures: Chemical compounds, visual objects, circuits, XML
- Biological networks
- Bibliographic networks: DBLP, ArXiv, PubMed, …
- Social networks: Facebook >100 million active users
- World Wide Web (WWW): > 3 billion nodes, > 50 billion arcs
- Cyber-physical networks



World-Wide Web

Yeast protein interaction network

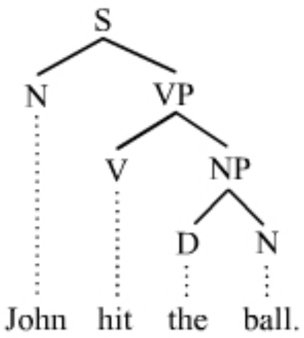


Co-author network

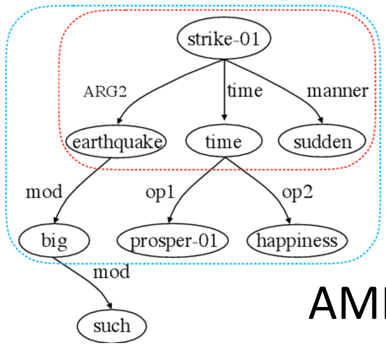
Social network sites

Text & Graph often appears simultaneously

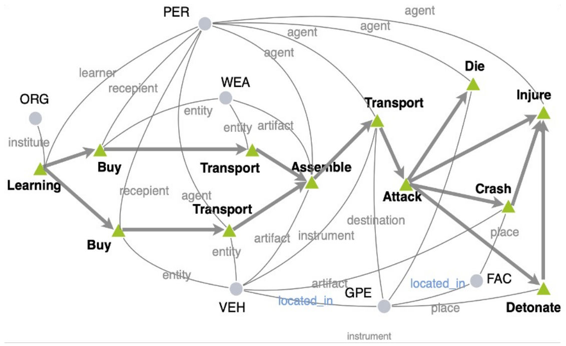
- Text sequence can be modeled as graph (AMR, information extraction, …).



Constituency parsing Graph



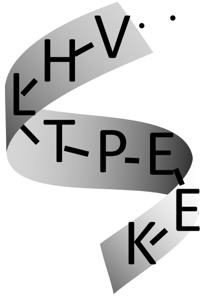
AMR Graph



IE Graph

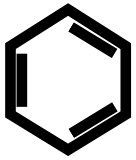
- Graphs are associated with text information.

Protein Graphs

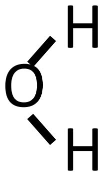


"Myoglobin holds oxygen in muscles."

Molecule Graphs

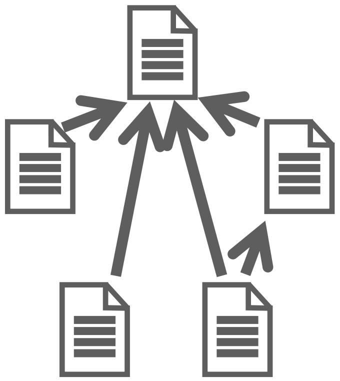


"Benzene is toxic"

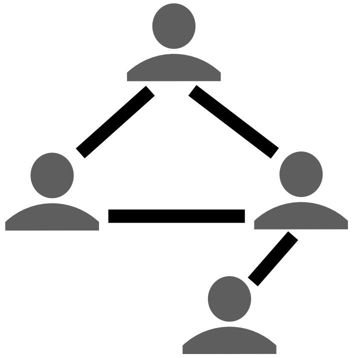


"Water is less toxic"

Academic Networks

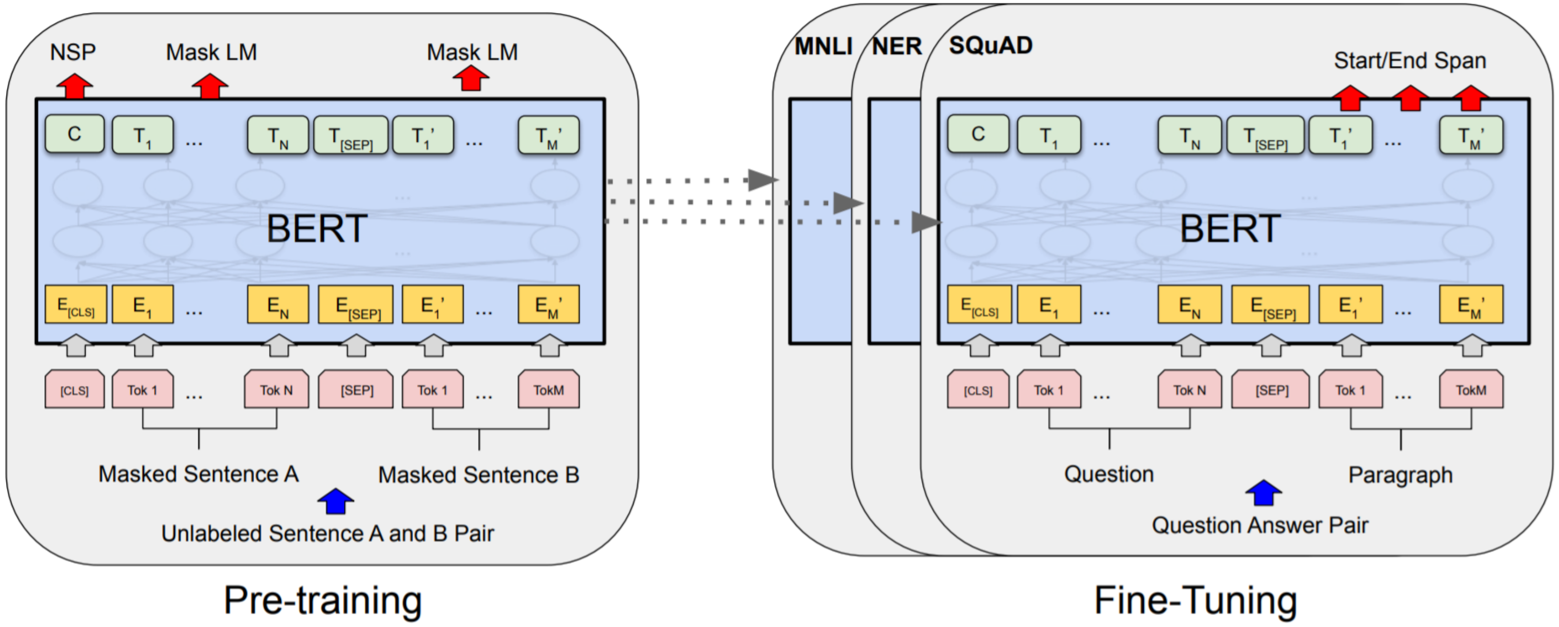


Social Networks



Foundation for Text Analysis: (Large) 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

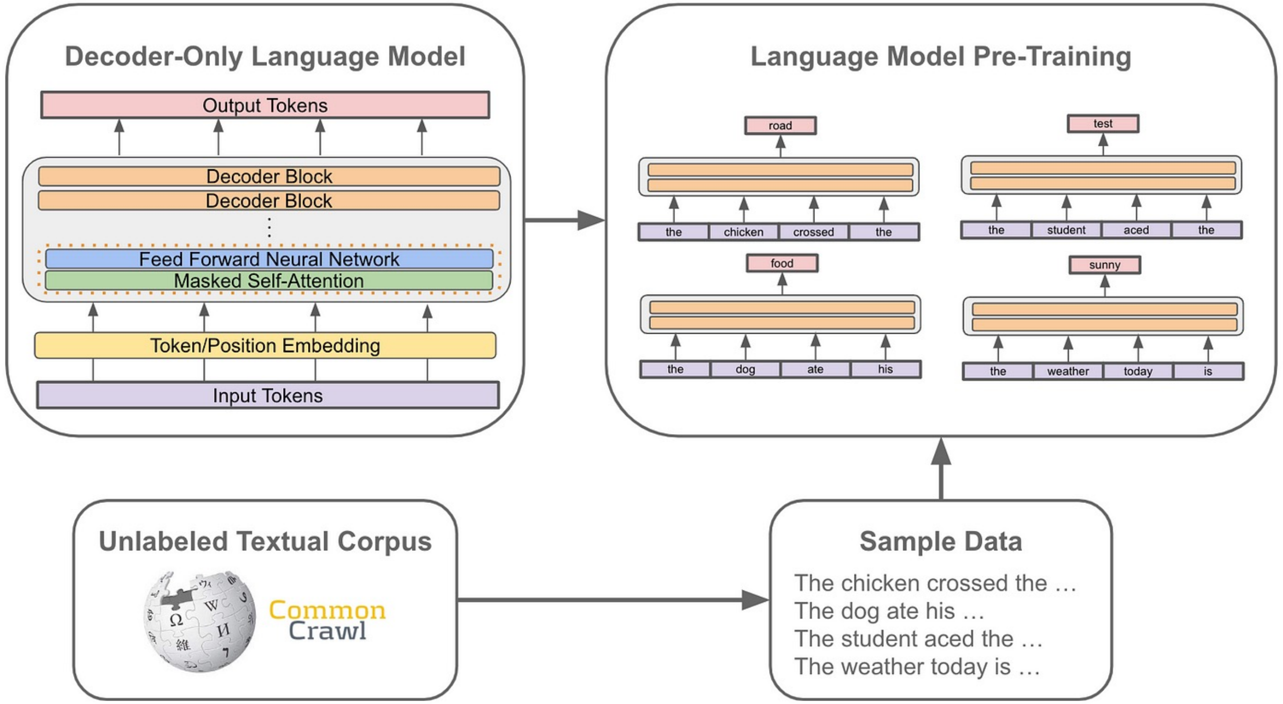


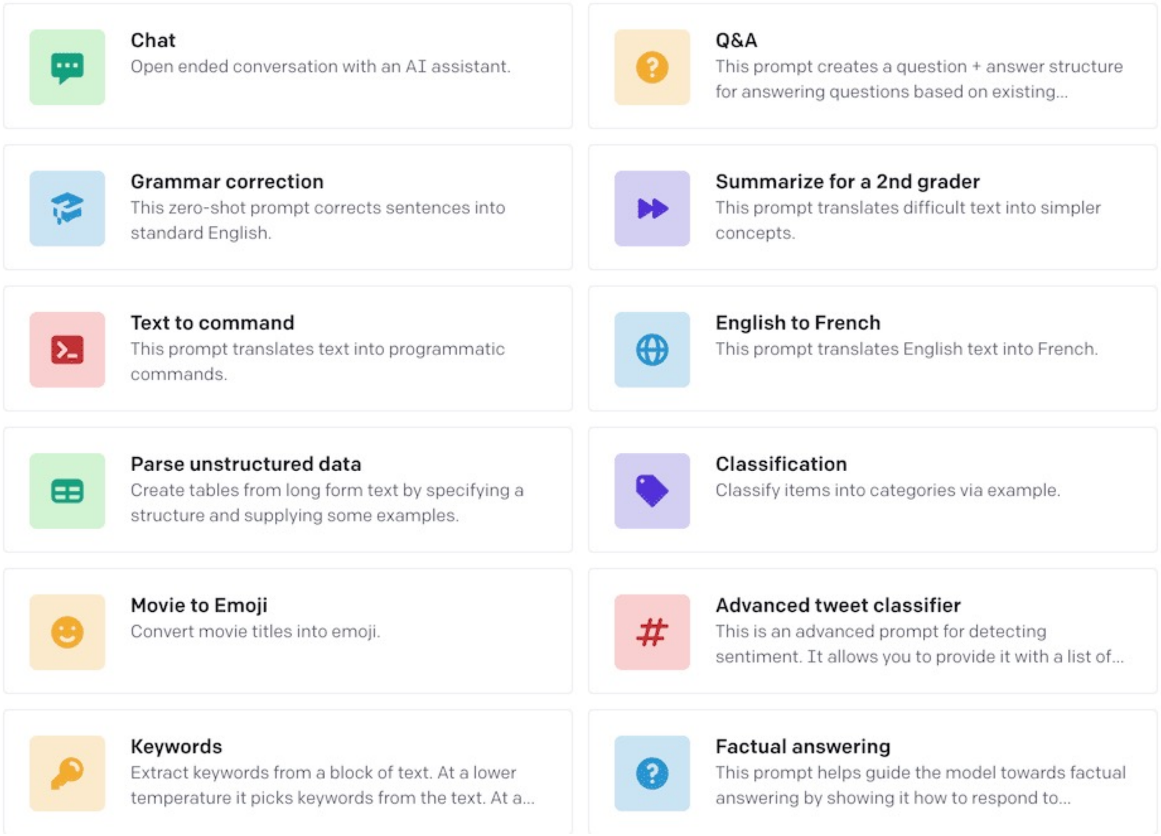
Unsupervised/Self-supervised;
On large-scale general domain corpus

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





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:

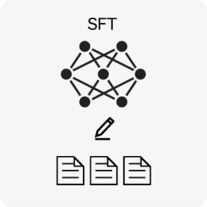
A prompt is sampled from our prompt dataset.

Explain reinforcement learning to a 6 year old.

A labeler demonstrates the desired output behavior.

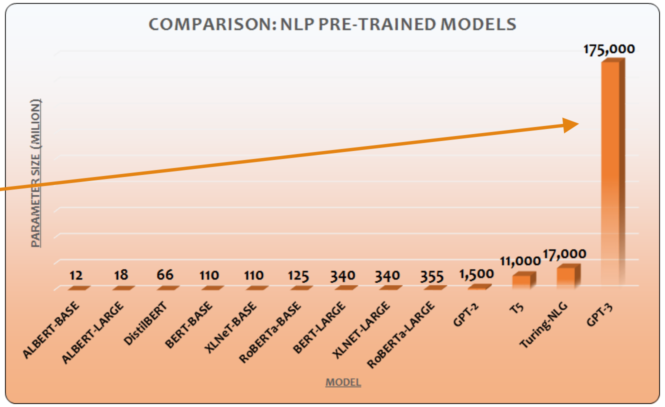
We give treats and punishments to teach...

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



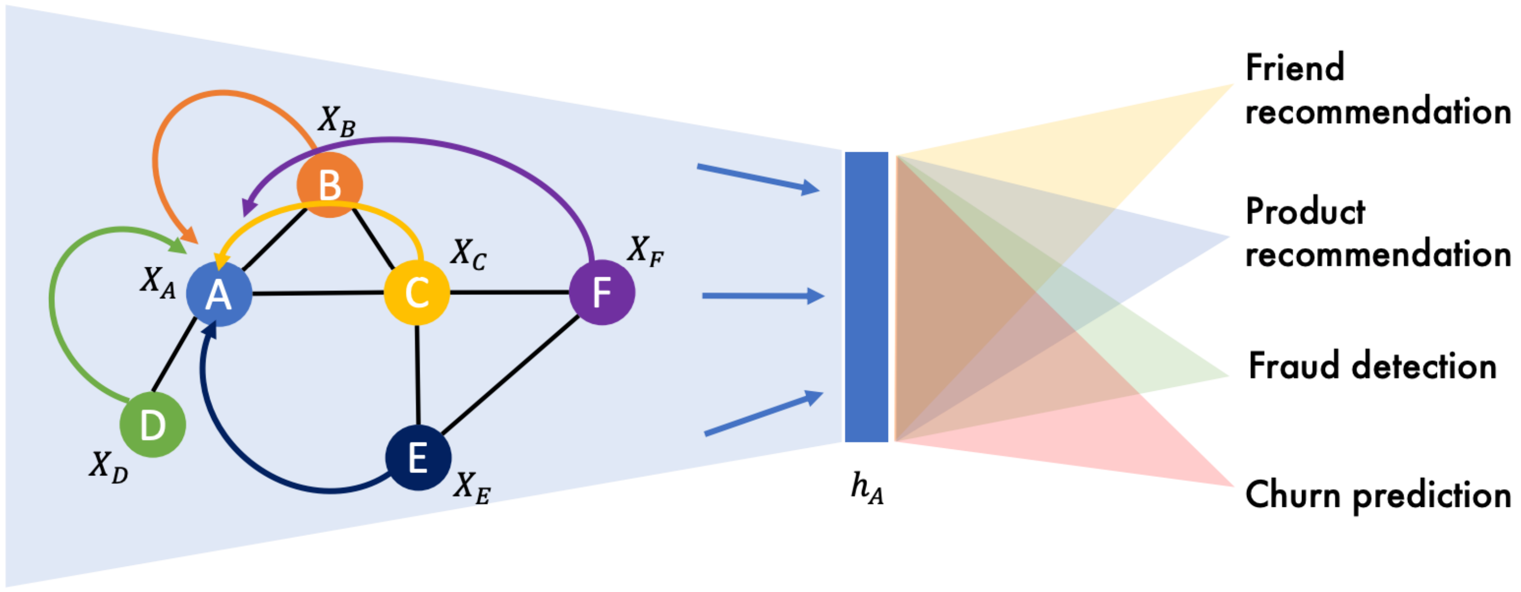
- Costly & Inefficient: Too large to be used in many applications

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



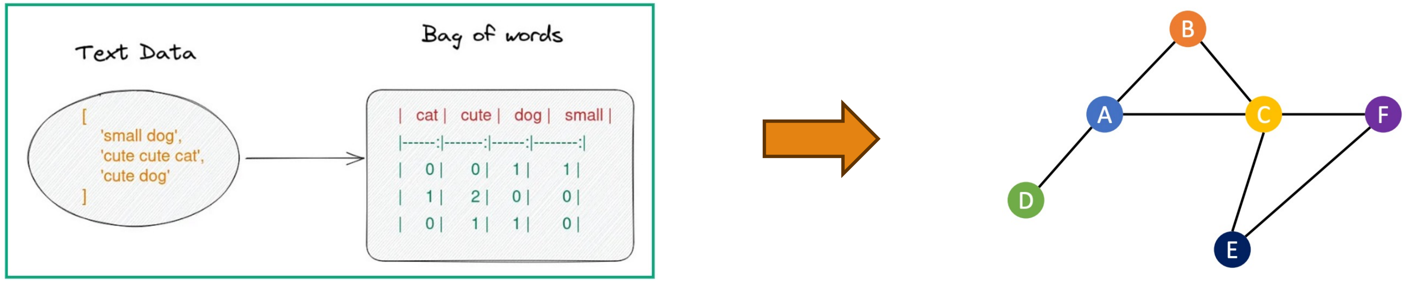
Foundation for Graph Analysis: Graph Neural Networks

- Graph Neural Network
  - Propagation & Aggregation.
  - Applied for various downstream tasks.

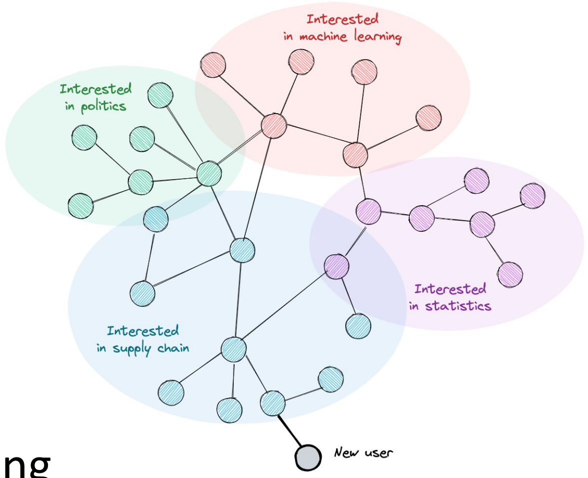


Challenges of Graph Neural Networks

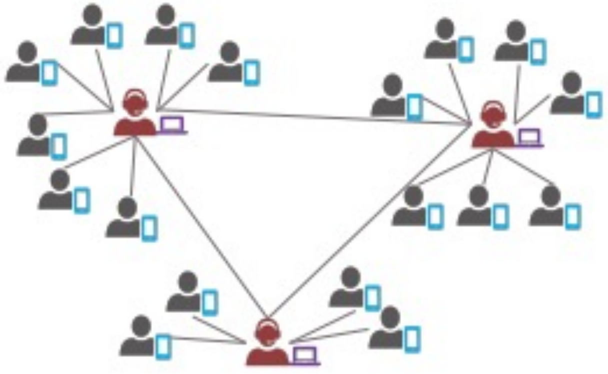
- Not able to capture rich contextualized text info with nodes/edges.
  - Need to transfer to BOW features or context-free embeddings.



- Suffers from over-smoothing and heterophily issues.



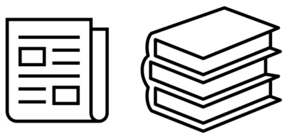
Over-Smoothing



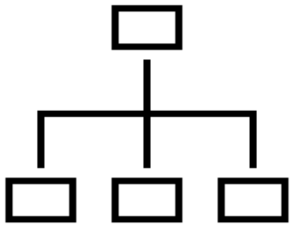
Heterophilic Graph

Bridging Text Data and Graph data

- Towards Semantics and Structure-aware Knowledge Discovery



Unstructured massive text data:
papers, news, …

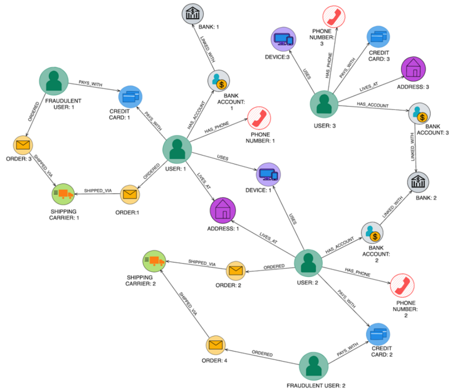


Structured graphs

Part I: Enhancing Text with Graph Structure

Graph signal

Text signal



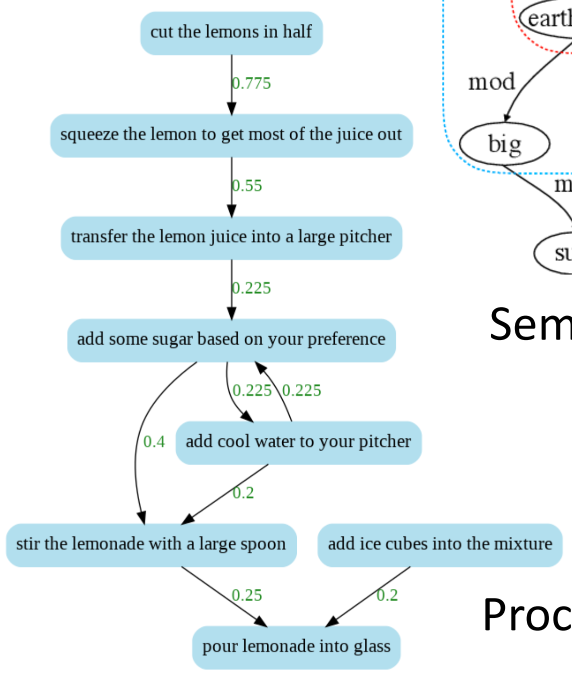
Part II: Graph Mining with Large Language Model

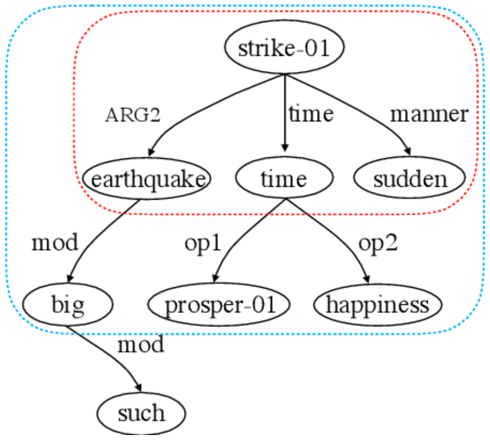
Part III: Text Mining with Structure Information

Part IV: Summary & Looking Forward

Overview of Enhancing Text with Graph Structure

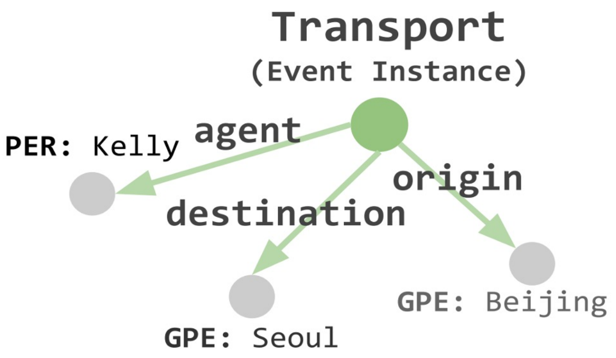
- Text can be converted to (or augmented with) graph structures at various different levels
  - How should we perform the conversion and when is it useful?



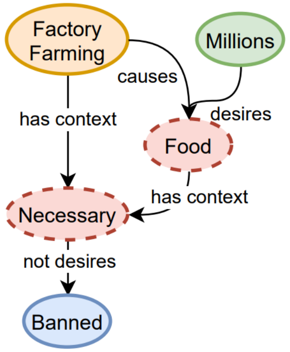


Semantic Parse Graphs

Procedure Graphs



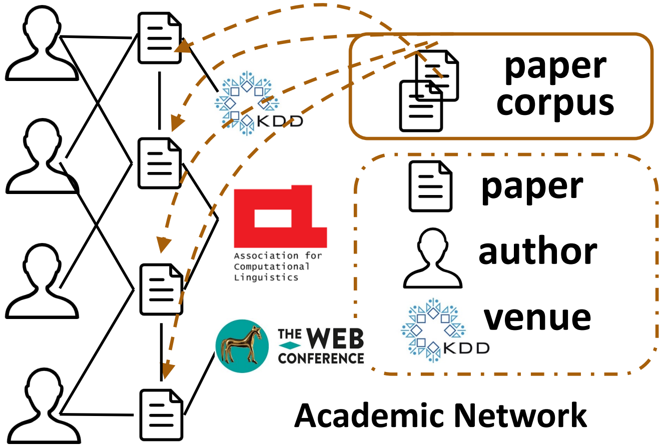
Information Extraction Graphs



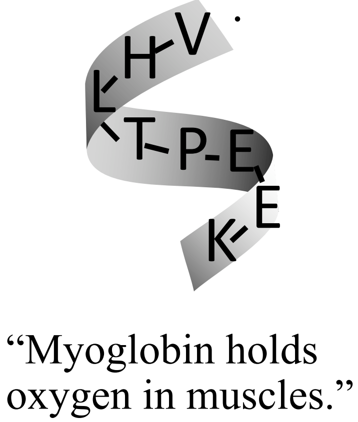
Reasoning Graphs

Overview of Graph Mining with Large Language Models

- In real world, text and graph appears simultaneously.
  - Text data are associated with rich structure information in the form of graphs.
  - Graph data are captioned with rich textual information.

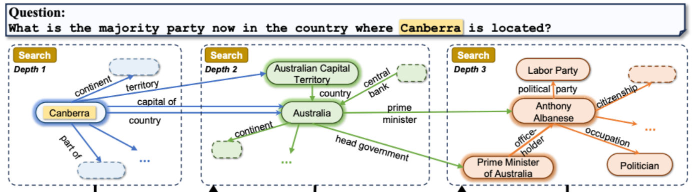


Academic Network



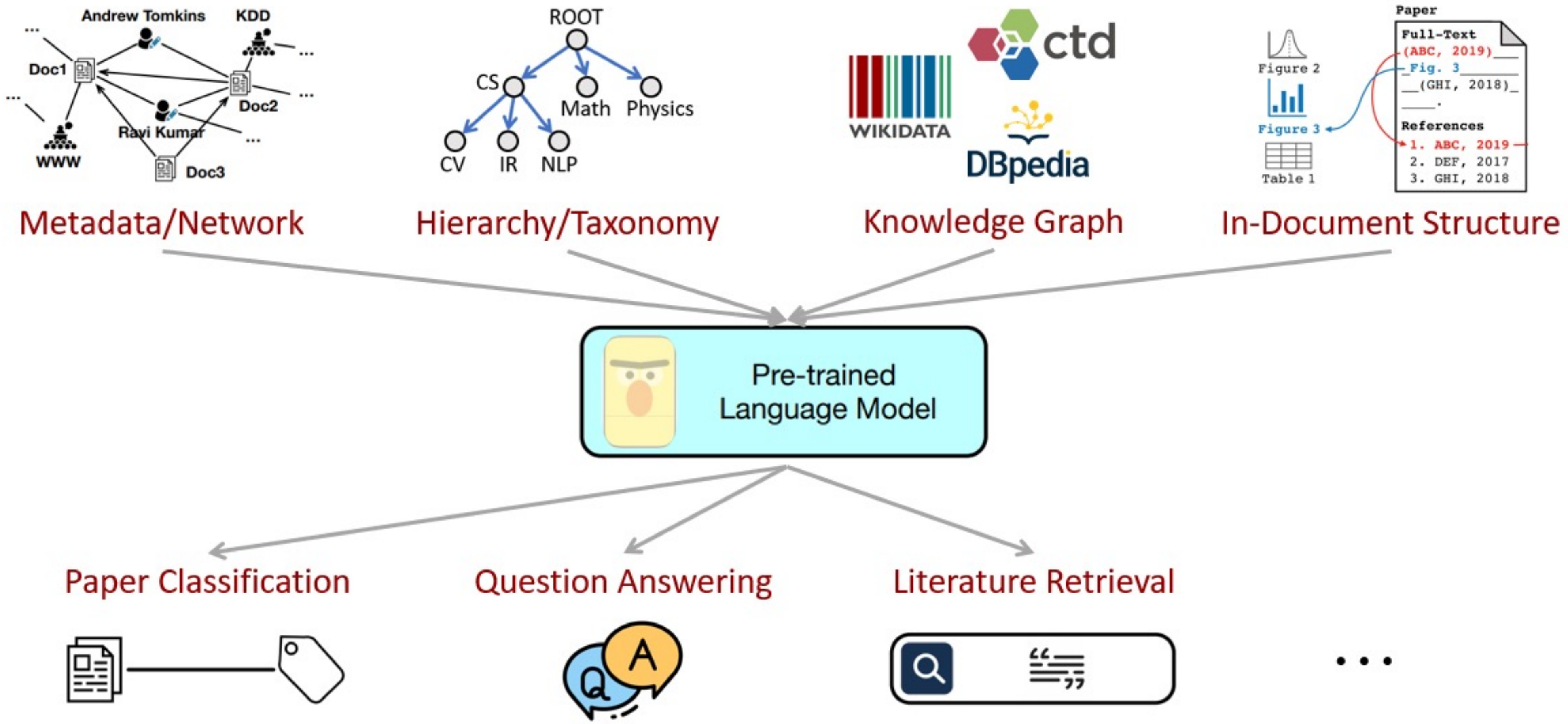
Protein Graphs

- Although LLMs have shown remarkable text reasoning ability, it is underexplored whether such ability can be generalized to graph scenarios.
- How can we adopt LLMs on graphs?

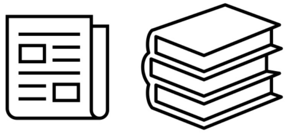


Overview of Text Mining with Structured Information

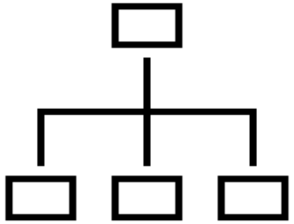
- Text data are often associated with or accompanied by structured information.
  - How to inject structured information into pre-trained language models for various text mining tasks?



Our Roadmap of This Tutorial



Unstructured massive text data:
papers, news, …

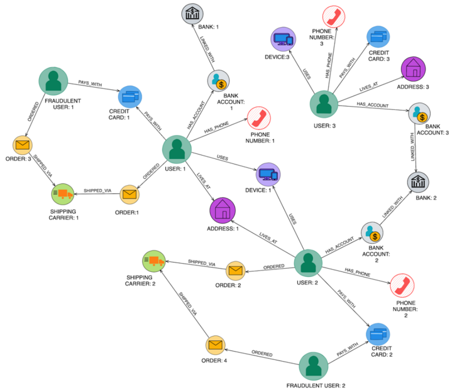


Structured graphs

Part I: Enhancing Text with Graph Structure

Graph signal

Text signal



Part II: Graph Mining with Large Language Model

Part III: Text Mining with Structure Information

Part IV: Summary & Looking Forward

16