Automated Mining of Structured Knowledge from Text in the Era of Large Language Models

Yunyi Zhang, Ming Zhong, Siru Ouyang, Yizhu Jiao, Sizhe Zhou, Linyi Ding, Jiawei Han

Siebel School of Computing and Data Science

University of Illinois Urbana-Champaign

Aug 25, 2024

Tutorial Website:

1

About Instructors
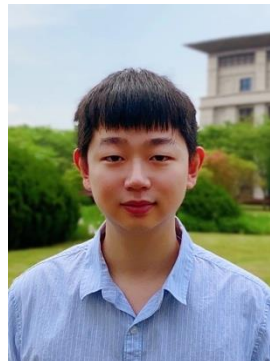

- Yunyi Zhang
- Ph.D. Candidate @ UIUC

- Siru Ouyang
- Ph.D. Candidate @ UIUC

- Jiawei Han
- Michael Aiken Chair Professor @ UIUC

- Ming Zhong
- Ph.D. Candidate @ UIUC

- Yizhu Jiao
- Ph.D. Candidate @ UIUC

- Sizhe Zhou
- M.S. Student @ UIUC

- Linyi Ding
- M.S. Student @ UIUC

Why Do We Still Need to Mine Structured Text Knowledge?

❑ LLM has revolutionized the field of NLP and Text Mining

❑ Some fundamental NLP/Text Mining tasks still cannot be replaced by LLM

    ❑ Ex. Fine-grained information extraction, automated fine-grained text classification, structured knowledge mining, …

❑ Structured Knowledge Mining may empower LLM applications

    ❑ Ex. RAG (retrieval augmented generation), natural language understanding

❑ LLM will empower automated structured knowledge mining from text

    ❑ Almost every task shown here gets help from LMs

❑ Automated structured knowledge mining from text—A key to both successful LLM applications and NL understanding

What Will Be Covered in This Tutorial?

❑ **An Introduction to Large Language Models**

  ❑ **LLM architecture, training + fine tuning, prompting, retrieval augmented generation, …**

❑ **Taxonomy Construction and Enrichment**

  ❑ **Why is taxonomy critical in text mining? Methods for Taxonomy Construction, Expansion and Enrichment**

❑ **Weakly-Supervised Text Classification**

  ❑ **Why weakly-supervised text classification? Methods for Weakly-supervised flat and hierarchical text classification**

❑ **Weakly-Supervised Information Extraction**

  ❑**Methods for Entity Typing, Relation Extraction, and Comprehensive Knowledge Structuring**

❑ **Conclusions**

❑ **Q/A**

Estimated Timeline for This Tutorial

❑ **Introduction: 5 mins (10:00-10:05 Han)**

❑ **Part I: Language Foundation Models for Text Analysis: 45 mins (10:05-10:50 Han)**

❑ **Part II: Taxonomy Construction and Enrichment: 35 mins (10:50-11:25 Zhang)**

❑ **Break: 10 mins (11:25-11:35)**

❑ **Part III: Weakly-Supervised Text Classification: 30 mins (11:35-12:05 Zhang)**

❑ **Part IV: Weakly-Supervised Information Extraction: 45 mins (12:05-12:50 Ouyang)**

❑ **Part V: Conclusions plus Q/A: 10 min (12:50-13:00 All)**