# Web Science

# Course Project

please take it seriously!

# Tasks1

- Constructing a knowledge graph for a specific domain
  - Domain: Military (军事)
  - Source data: crawling data from Web (wiki and news)
    从百度百科/维基百科爬取半结构化数据
    从新闻网站爬取非结构化数据
  - Two group collaboration
    允许两个组协作完成，分别负责半结构化数据和非结构化数据
    在报告中明确每个成员的贡献

# Tasks1

- Constructing a knowledge graph for a specific domain
  - Ontology: knowledge representation, concepts, properties, …
    人工+半自动构建本体，确定概念、属性、公理
  - Knowledge graph:
    - knowledge extraction (entity, relations)
      本结构化数据：可以用规则进行抽取
      非结构化数据：利用深度学习模型抽取主要类型的实体和关系
    - knowledge fusion/integration
      融合2个组构建的知识，解决本体层异构和实例层异构

# Tasks1

- Constructing a knowledge graph for a specific domain
  - Knowledge graph:
    - knowledge storage
      利用图数据库如neo4j等存储知识图谱，非结构化数据用MongoDB存储
    - knowledge embedding (optional)
  - Knowledge graph sizes:
    - >100,000 entities
    - >1,000,000 triples

# Tasks1

- Intelligent applications based on the knowledge graph
  - Semantic search
  - Question answer
  - Visualization (optional)
  - Mining and analyzing knowledge graph (optional)
  - Any other exciting applications (use your imagination)

# Tasks1: How to do it

1. Problem direction, context, outline of algorithm and evaluation

- Foundation of knowledge graph(知识图谱基础学习)
  - online open course: https://github.com/npubird/KnowledgeGraphCourse
  - book: 《知识图谱：方法、实践与应用》，电子工业出版社，2019
- make clear knowledge graph domains and applications
  - domains: data sources (明确知识图谱领域和数据源)
  - applications:  techniques (明确基于知识图谱的应用和涉及的技术)
- Preliminary design(概要设计)
  - System architecture(系统架构)
  - key modules and their functions(模块及其功能)
  - key techniques and challenges (关键技术及挑战)

# Tasks1: How to do it

2. Formulation, algorithm, data, preliminary results

- Collecting data (数据采集)
  - Crawling data from Web: wikipedia, baike, news sites, social networks, online forum, …
  - data type: databased, text, image, video, …
  - Reference: https://github.com/npubird/KnowledgeGraphCourse/blob/master/pub-5知识抽取-数据获取.pdf

- Ontology building (本体构建)
  - Reference:

Noy N F, McGuinness D L. Ontology Development 101: A Guide to Creating Your First Ontology. another version

# Tasks1: How to do it

2. Formulation, algorithm, data, preliminary results

- Knowledge extraction (知识抽取)
  - named entity recognition
  - relation extraction
  - Reference:

Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. KDD2014: 601-610.

Auer S, Bizer C, Kobilarov G, et al. Dbpedia: A nucleus for a web of open data. ISWC2007: 722-735.

Suchanek F M, Kasneci G, Weikum G. Yago: a core of semantic knowledge. WWW2007: 697-706.

# Tasks1: How to do it

2. Formulation, algorithm, data, preliminary results

- Knowledge fusion (知识融合)
  - ontology matching
  - instance matching
  - Reference:
    第五章 知识融合，in《知识图谱：方法、实践与应用》，电子工业出版社，2019

- Knowledge storage (知识存储)
  - graph database
  - Reference:
    https://github.com/npubird/KnowledgeGraphCourse/blob/master/pub-11知识存储.pdf

# Tasks1: How to do it

3. Additional theory/methods and results, applications

- Intelligence applications (智能应用)
  - applications: search, QA, visualization, mining, reasoning
  - theories/methods/algorithms
  - Reference:

    《聊天机器人技术原理与应用》，中国工信出版集团，2019

# Tasks2

- . Reproducing the SOTA models or methods
  - WWW, Web Search, Social Network, Knowledge Graph, NLP
  - 2020~2021 models on top conferences or leading journals
  - Focusing on one problem
  - Reproducing 1-2 models
  - Evaluations
  - Comparison

# Tasks3

- . Evaluating the SOTA models or methods
  - WWW, Web Search, Social Network, Knowledge Graph, NLP
  - 2020~2021 models on top conferences or leading journals
  - Focusing on one problem
  - Datasets
  - Reproducing or running at least 5 models
  - Evaluations
  - Comparison

# 参考的主题

- 数据增强方法评估/数据增强在文本分类(某问题)的评估
- BERT模型压缩方法评估
- 模型压缩方法评估
- 社交网络嵌入方法评估
- 实体识别评估
- 关系抽取评估
- 图数据库/知识图谱存储评估
- 一定要有编码和实验的工作量！

# Presentation and Report

- Final presentation
  - 15 mins presentation + 5 mins demo and questions

- Final report
  - Detailed writeup (latex, <30 pages)
  - Github site: source code & data set

# Timeline

| Tasks | Important Date |
| --- | --- |
| checkpoint1 | Septemper 30 |
| checkpoint2 | October 31 |
| checkpoint3 | November 30 |
| Final Presentation | December 30 |
| Final Report | December 30 |
| | |


धन्यवाद
Hindi
Hindi
多謝
Traditional Chinese
ขอบคุณ
Thai
Спасибо
Russian
Thank You
English
Gracias
Spanish
شكرا
Arabic
Grazie
Italian
Danke
German
Obrigado
Brazilian Portuguese
Merci
French
多谢
Simplified Chinese
நன்றி
Tamil
Tamil
ありがとうございました
Japanese
감사합니다
Korean