第八章 知识图谱

戴洪良

计算机科学与技术学院/人工智能学院

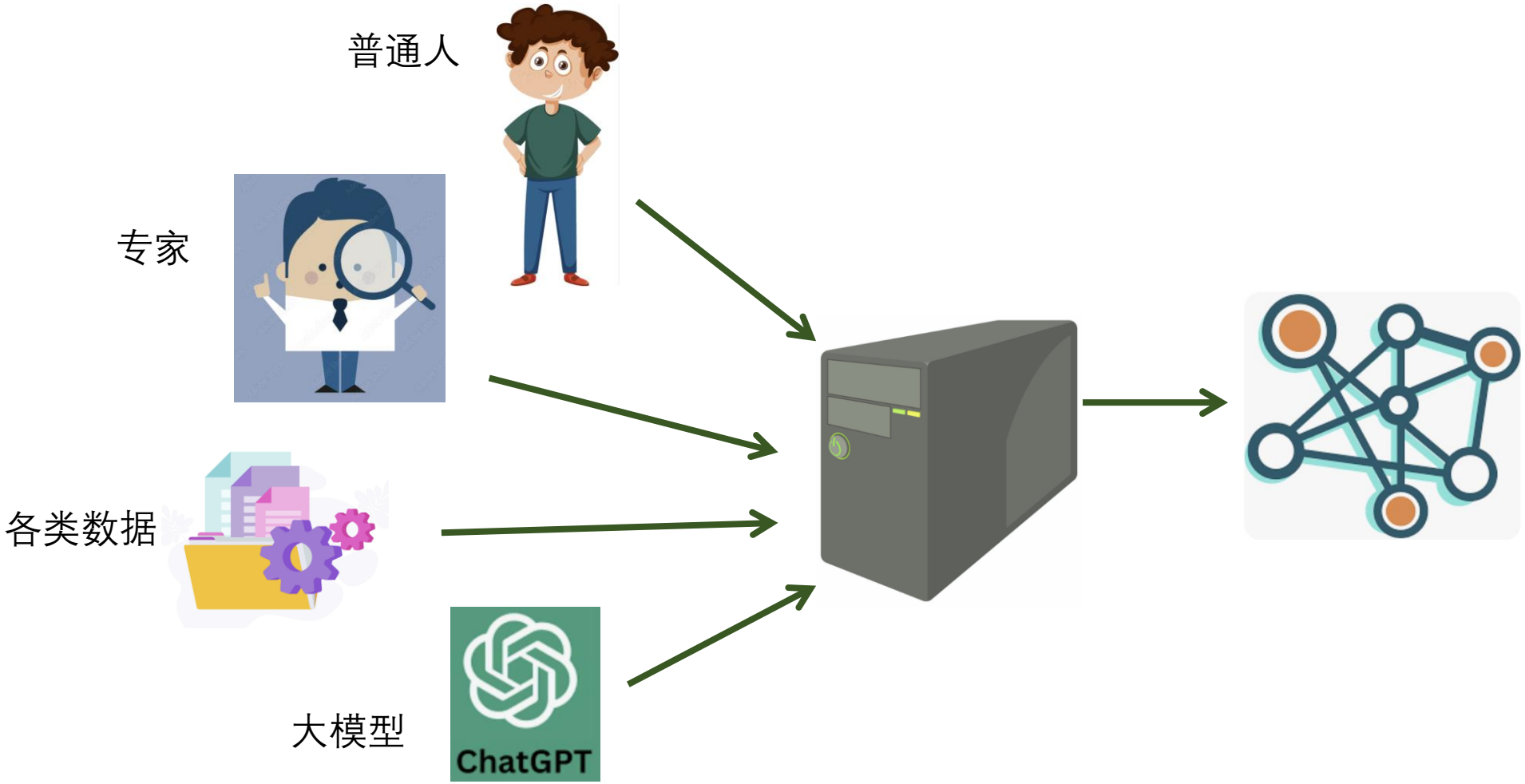
hongldai@nuaa.edu.cn

知识图谱构建

- 知识获取
  - 有知识才能构建知识图谱，知识获取是知识图谱构建的关键环节

知识获取

- 可从哪些途径获取知识来构建知识图谱?



知识来源

- 不同类型的数据中包含了知识

张艺谋 [编辑]

条目 讨论 汉 漢 大陆简体

维基百科，自由的百科全书

张艺谋（1950年4月2日—），是一名中国导演、摄影师、演员。[1][2]他的电影经常取材自家乡陕西和母语关中话，代表作包括《红高粱》《菊豆》《秋菊打官司》《活着》《大红灯笼高高挂》《我的父亲母亲》《归来》。除电影外，他亦执导2008年北京奥运会与2022年北京冬奥会“双奥”开幕式，以及杭州G20峰会《最忆是杭州》文艺演出等大型活动。

C919圆满完成商业首航

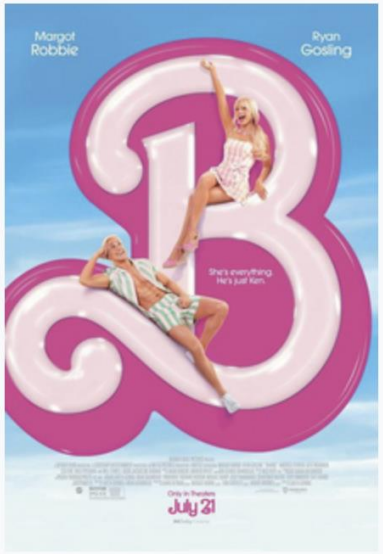
C919大型客机是我国首次按照国际通行适航标准自行研制、具有自主知识产权的喷气式干线客机。

“C919首次进入我国民航运输市场，展示了我国在航空交通领域的技术实力和市场潜力，提升了我国在国际民用航空市场上的话语权和影响力，为推动全球航空产业的多元化和平衡化作出了贡献。”邵荃说。

在邵荃看来，C919投入商业运营后将对中国的制造业产生两个重大“作用”：一是产业链龙头的“带动作用”，二是创新驱动的“引擎作用”。

邵荃进一步解释，目前，C919的国产化率已达60%，形成了规模化生产能力。“这必将带动我国航材、航电、航发、制造、设计、营销、售后等一系列产品和服务行业的崛起，也为5G/6G、大数据、云计算、区块链、人工智能等新技术新成果提供了新的应用场景，将促进数字化、智能化生产技术的革新，进一步提升生产效率和高质量发展能力。”

Barbie



Theatrical release poster

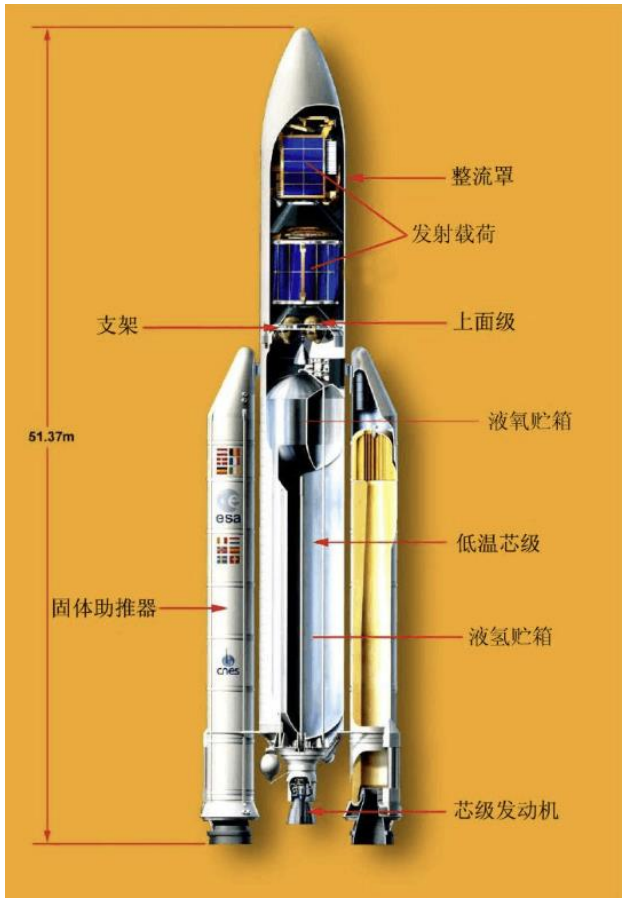
Directed by	Greta Gerwig
Written by	Greta Gerwig Noah Baumbach
Based on	Barbie by Mattel
Produced by	David Heyman Margot Robbie Tom Ackerley Robbie Brenner
Starring	Margot Robbie Ryan Gosling America Ferrera Kate McKinnon Issa Rae Rhea Perlman

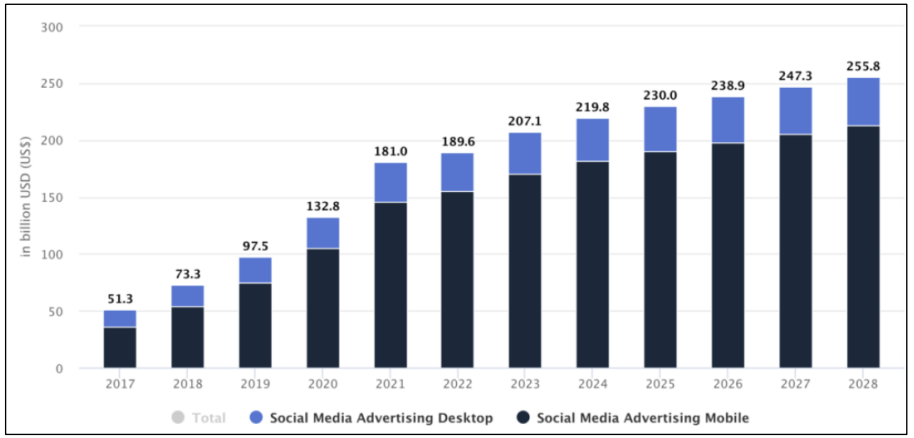
知识来源

- 不同类型的数据中包含了知识

我认为一份好吃的"蛋炒饭"应该要这样做。

我的世界：当我被困在荒岛上，该怎样生存呢?





知识获取

- 人工编写
  - 专家、普通人

准确度高，成本也高

- 各种类型的数据中抽取
  - 结构化数据
  - 半结构化数据
  - 非结构化数据
    - 文本、图片、视频等

能快速获取大量结构化的知识，成本低，准确度也低

知识获取

- 结构化数据
  - 以固定格式和标准组织和表示数据元素及它们间的关联
  - 如SQL数据库
- 半结构化数据
  - 有一定的数据元素组织和表示格式，但不同数据条目的格式可能不同
  - 如XML、HTML文档
- 非结构化数据
  - 没有将数据按一定结构进行组织
  - 如文本、图片

知识获取

- 结构化数据
- 半结构化数据
- 非结构化数据（考虑文本数据）

从关系数据库中抽取知识

- 抽取原理
  - 表(Table) - 类(Class)
  - 列(Column) - 属性(Property)
  - 行(Row) - 资源/实例(Resource/Instance)
  - 单元(Cell) - 属性值(Property Value)
  - 外键(Foreign Key) - 指代(Reference)

- 根据上述规则可将关系数据库转化为一个知识库。

知识获取

- 结构化数据
- 半结构化数据
- 非结构化数据（考虑文本数据）

从半结构化数据抽取知识

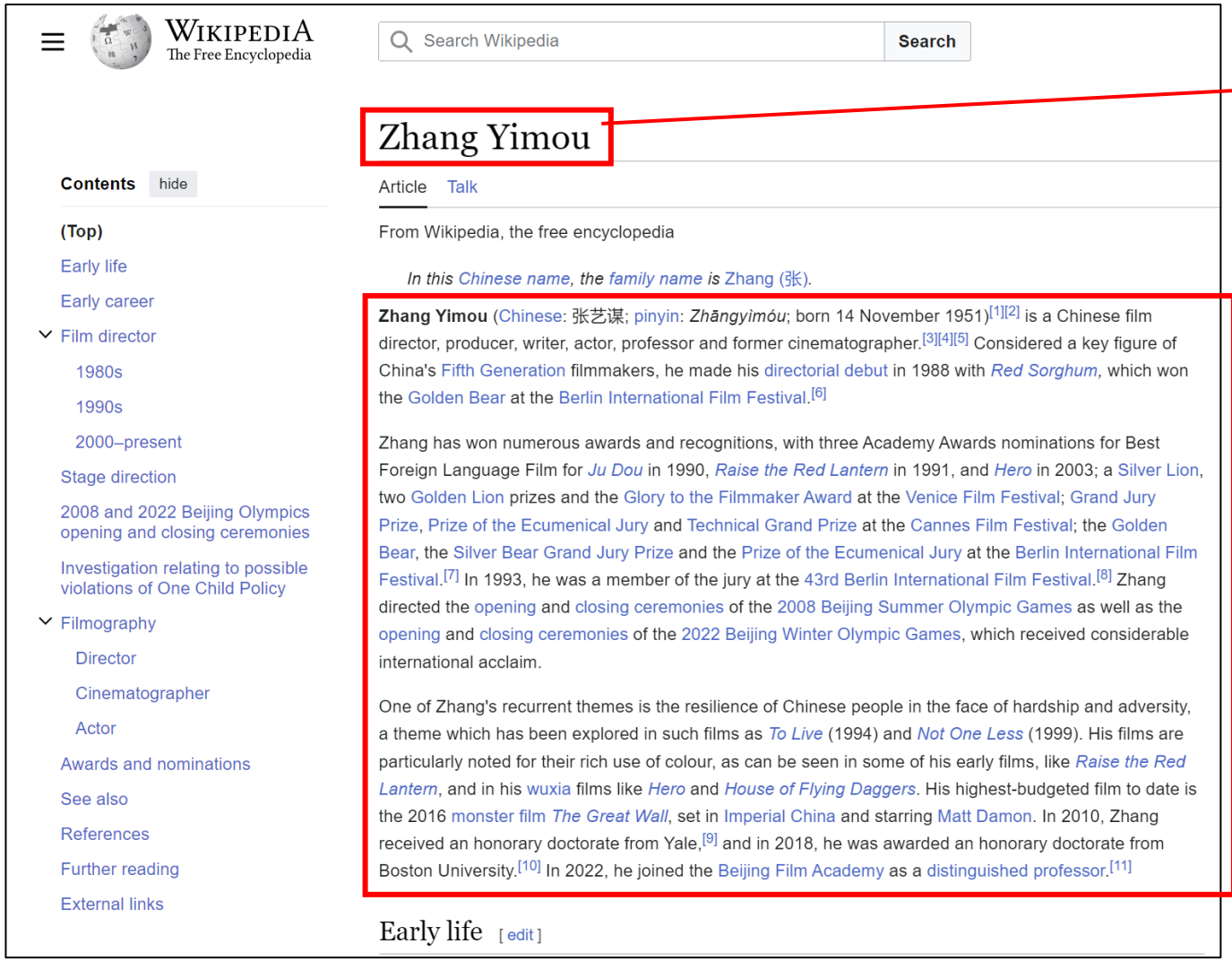
- 半结构化数据形式多样，缺少统一的抽取方法
- Wikipedia被广泛作为数据源应用于知识图谱构建
  - 如DBpedia、YAGO、BabelNet都利用了Wikipedia构建

以维基百科为例

- DBpedia从Wikipedia的infobox、摘要（开头几行）、类别（categories）、消歧连接（disambiguation links）等抽取了知识

Name	*Description*	*Example*
abstract	Extracts the first lines of the Wikipedia article.	`dbr:Berlin dbo:abstract "Berlin is the capital city of (...)".`
article categories	Extracts the categorization of the article.	`dbr:Oliver_Twist dc:subject dbr:Category:English_novels.`
category label	Extracts labels for categories.	`dbr:Category:English_novels rdfs:label "English novels".`
category hierarchy	Extracts information about which concept is a category and how categories are related using the SKOS Vocabulary.	`dbr:Category:World_War_II skos:broader dbr:Category:Modern_history.`
disambiguation	Extracts disambiguation links.	`dbr:Alien dbo:wikiPageDisambiguates dbr:Alien_(film).`
external links	Extracts links to external web pages related to the concept.	`dbr:Animal_Farm dbo:wikiPageExternalLink <http://books.google.com/?id=RBGmrDnBs8UC>.`
geo coordinates	Extracts geo-coordinates.	`dbr:Berlin georss:point "52.5006 13.3989".`
grammatical gender	Extracts grammatical genders for persons.	`dbr:Abraham_Lincoln foaf:gender "male".`
homepage	Extracts links to the official homepage of an instance.	`dbr:Alabama foaf:homepage <http://alabama.gov/>.`
image	Extracts the first image of a Wikipedia page.	`dbr:Berlin foaf:depiction <http://.../Overview_Berlin.jpg>.`
infobox	Extracts all properties from all infoboxes.	`dbr:Animal_Farm dbo:date "March 2010".`
interlanguage	Extracts interwiki links.	`dbr:Albedo dbo:wikiPageInterLanguageLink dbr-de:Albedo.`

以维基百科为例



WIKIPEDIA
The Free Encyclopedia

Search Wikipedia | Search

Zhang Yimou

Article | Talk

From Wikipedia, the free encyclopedia

In this Chinese name, the family name is Zhang (张).

Zhang Yimou (Chinese: 张艺谋; pinyin: *Zhāngyìmóu*; born 14 November 1951)[1][2] is a Chinese film director, producer, writer, actor, professor and former cinematographer.[3][4][5] Considered a key figure of China's Fifth Generation filmmakers, he made his directorial debut in 1988 with *Red Sorghum*, which won the Golden Bear at the Berlin International Film Festival.[6]

Zhang has won numerous awards and recognitions, with three Academy Awards nominations for Best Foreign Language Film for *Ju Dou* in 1990, *Raise the Red Lantern* in 1991, and *Hero* in 2003; a Silver Lion, two Golden Lion prizes and the Glory to the Filmmaker Award at the Venice Film Festival; Grand Jury Prize, Prize of the Ecumenical Jury and Technical Grand Prize at the Cannes Film Festival; the Golden Bear, the Silver Bear Grand Jury Prize and the Prize of the Ecumenical Jury at the Berlin International Film Festival.[7] In 1993, he was a member of the jury at the 43rd Berlin International Film Festival.[8] Zhang directed the opening and closing ceremonies of the 2008 Beijing Summer Olympic Games as well as the opening and closing ceremonies of the 2022 Beijing Winter Olympic Games, which received considerable international acclaim.

One of Zhang's recurrent themes is the resilience of Chinese people in the face of hardship and adversity, a theme which has been explored in such films as *To Live* (1994) and *Not One Less* (1999). His films are particularly noted for their rich use of colour, as can be seen in some of his early films, like *Raise the Red Lantern*, and in his wuxia films like *Hero* and *House of Flying Daggers*. His highest-budgeted film to date is the 2016 monster film *The Great Wall*, set in Imperial China and starring Matt Damon. In 2010, Zhang received an honorary doctorate from Yale,[9] and in 2018, he was awarded an honorary doctorate from Boston University.[10] In 2022, he joined the Beijing Film Academy as a distinguished professor.[11]

Early life [edit]

Contents hide

(Top)
Early life
Early career
Film director
1980s
1990s
2000–present
Stage direction
2008 and 2022 Beijing Olympics opening and closing ceremonies
Investigation relating to possible violations of One Child Policy
Filmography
Director
Cinematographer
Actor
Awards and nominations
See also
References
Further reading
External links

维基页面文章标题
可在KG中作为节点label

rdfs:label
- Zhang Yimou (en)
- 张艺谋 (zh)

没有小节标题的前几段
是对实体的简单整体描述
rdfs:comment只取第一段
dbo:abstract取全部段落

https://en.wikipedia.org/wiki/Zhang_Yimou

以维基百科为例

infobox模板
可作为实体类别
rdf:type

Professor
Zhang Yimou
张艺谋

Zhang in 2023

Born	14 November 1951 (age 72) Xi'an, Shaanxi, China
Alma mater	Beijing Film Academy
Occupation(s)	Film director, producer, cinematographer and actor
Notable work	*Full River Red* *House of Flying Daggers* *Cliff Walkers* *The Flowers of War*
Spouses	Xiao Hua (肖华) (m. 1978–1988) Chen Ting (陈婷) (m. 2011)
Children	Zhang Mo Zhang Yinan

```
{infobox person
| honorific_prefix   = [[Professor]]
| name               = Zhang Yimou
| image              = Zhang Yimou from "Full River Red" at Red Carpet of the Tokyo
International Film Festival 2023 (53347207442) (cropped).jpg
| alt                = cropped headshot of Zhang at Tokyo International Film Festival
2023
| caption            = Zhang in 2023
| native_name        = {{nobold|张艺谋}}
| native_name_lang   = zh
| birth_date         = {{Birth date and age|df=yes|1951|11|14}}
| birth_place        = [[Xi'an]], [[Shaanxi]], China
| alma_mater         = [[Beijing Film Academy]]
| occupation         = [[Film director]], [[Film producer|producer]],
[cinematographer]] and [[actor]]
| notable_works      = ''[[Full River Red]]''<br>''[[House of Flying
Daggers]]''<br>''[[Cliff Walkers]]''<br>''[[The Flowers of War]]''
| spouse             = {{plainlist|
* {{marriage|Xiao Hua ({{lang|zh-Hans|肖华}})|1978|1988}}
* {{marriage|Chen Ting ({{lang|zh-Hans|陈婷}})|2011}}
}}
| children           = [[Zhang Mo (director)|Zhang Mo]]<br />Zhang Yinan<br />Zhang
Yiding<br />Zhang Yijiao
| parents            = Zhang Bingjun<br />Zhang Xiaoyou
| family             = Zhang Weir
| awards             = [[BAFTA Av                    e English Language|BAFTA
Best Film Not in the English Lan                    the Red Lantern]]''<br
```

可作为实体属性，需对内容进行解析

以维基百科为例

- 内部超链接（anchor links）
  - 构建dbo:wikiPageWikiLink关联

Zhang has won numerous awards and recognitions, with three Academy Awards nominations for Best Foreign Language Film for *Ju Dou* in 1990, *Raise the Red Lantern* in 1991, and *Hero* in 2003; a Silver Lion, two Golden Lion prizes and the Glory to the Filmmaker Award at the Venice Film Festival; Grand Jury Prize, Prize of the Ecumenical Jury and Technical Grand Prize at the Cannes Film Festival; the Golden Bear, the Silver Bear Grand Jury Prize and the Prize of the Ecumenical Jury at the Berlin International Film Festival.[7] In 1993, he was a member of the jury at the 43rd Berlin International Film Festival.[8] Zhang directed the opening and closing ceremonies of the 2008 Beijing Summer Olympic Games as well as the opening and closing ceremonies of the 2022 Beijing Winter Olympic Games, which received considerable international acclaim.

以维基百科为例

- Categories
  - Categories are intended to group together pages on similar subjects
  - 在dbpedia中用于构建dcterms:subject属性

Categories: Zhang Yimou | Filmmakers who won the Best Foreign Language Film BAFTA Award | Beijing Film Academy alumni | Chinese cinematographers | Film directors from Shaanxi | Artists from Xi'an | 1951 births | Living people | Directors of Golden Bear winners | Directors of Golden Lion winners | Chevaliers of the Ordre des Arts et des Lettres | Chinese film directors | Members of the 9th Chinese People's Political Consultative Conference | Members of the 10th Chinese People's Political Consultative Conference | Members of the 11th Chinese People's Political Consultative Conference | Asia Game Changer Award winners | Writers from Xi'an | Male actors from Xi'an | 20th-century Chinese male actors | 21st-century Chinese male actors

以维基百科为例

- 一些Category间有层次关系
  - DBpedia中用来构建了skos:broader关联

Category:Chinese film directors

Category Talk Read Edi

From Wikipedia, the free encyclopedia

Classification: People: By occupation: Filmmakers / Directors: Film directors: By nationality: **Chinese**
Also: China: People: By occupation: Filmmakers / Directors: **Film directors**

See also: Category:Taiwanese film directors

Contents

Top · A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

Subcategories

This category has the following 7 subcategories, out of 7 total.

▶ Chinese film directors by province (30 C)

.

▶ Hong Kong film directors (3 C, 205 P)

+

▶ Chinese women film directors (1 C, 48 P)

A

▶ Chinese animated film directors (14 P)

D

▶ Chinese documentary film directors (12 P)

S

▶ Chinese silent film directors (26 P)

Σ

▶ Chinese film director stubs (107 P)

以维基百科为例

- 消歧页（disambiguation page）和重定向页（redirect page）
  - 可作为实体的别名
  - DBpedia中用重定向页构建dbo:wikiPageRedirects属性

https://en.wikipedia.org/wiki/Yimou_Zhang

知识获取

- 结构化数据
- 半结构化数据
- 非结构化数据（考虑文本数据）

数据采集

- 根据数据类型不同，采集和处理数据的方式也不同，如：
  - Wikipedia：直接下载网站提供的dump文件
    - https://dumps.wikimedia.org/
  - 网页数据：爬虫
  - Word、PDF等格式文档：相应的文档读取/处理库，如PyPDF2
  - 扫描得到的文档：使用OCR技术得到文本

网页数据采集 – 爬虫

- 爬虫的一般流程
  1. 获得目标数据的URL
  2. 向对应URL提交HTTP请求
  3. 解析HTTP响应
  4. 存储解析结果

网页数据采集 – 爬虫

- 例：用python requests爬取网页内容

```
import requests

# 发送一个GET请求
r = requests.get('https://xiyouji.fandom.com/zh/wiki/%E7%8C%AA%E5%85%AB%E6%88%92')

# 查看HTTP状态码
print(r.status_code)

# 打印请求到的内容
print(r.content)
```

爬取后，需对得到的数据进行清洗。如用正则表达式或Python的BeautifulSoup抽取出其中需要的内容。

网页数据采集 – 爬虫

- 例：用python的BeautifulSoup4包处理html数据

```
<html>
<head><title>Example HTML</title></head>
<body>
<div class="section">
<p>猪八戒原是天庭玉皇大帝手下的天蓬元帅，主管天河 ...</p>
</div>
<div class="section">
<p>猪八戒身上既有人的吃苦耐劳、憨厚率直的品质 ...</p>
</div>
</body>
</html>
```

```
from bs4 import BeautifulSoup

with open('FILE_PATH', encoding='utf-8') as f:
    html_data = f.read()

soup = BeautifulSoup(html_data, features="html.parser")

# 找div tag，且class属性为'section'
sec_div = soup.findAll('div', {'class': 'section'})
for div in sec_div:
    p = div.find('p')
    print(p.text)
```

网页数据采集 – 爬虫

- robots.txt
  - 位于网站根目录，用来告诉爬虫哪些URL可以访问

例：

```
User-agent: GPTBot
Disallow: /

User-agent: *
Allow: /zh/wiki/Special:CreateNewWiki
Allow: /zh/wiki/Special:AllMaps
Noindex: /zh/wiki/Template:
Noindex: /zh/wiki/Template_talk:
Noindex: /zh/wiki/Help:
Disallow: /zh/wiki/Special:
Disallow: /zh/wiki/User_talk:
```

爬虫合法性的相关阅读：https://www.zhihu.com/question/291554395

从已有文档中读取文本

- 例：使用PyPDF2读PDF文件

```
import PyPDF2

# creating a pdf reader object
reader = PyPDF2.PdfReader('PATH_TO_PDF_FILE')

# print the number of pages in pdf file
print(len(reader.pages))

# print the text of the first page
print(reader.pages[0].extract_text())
```

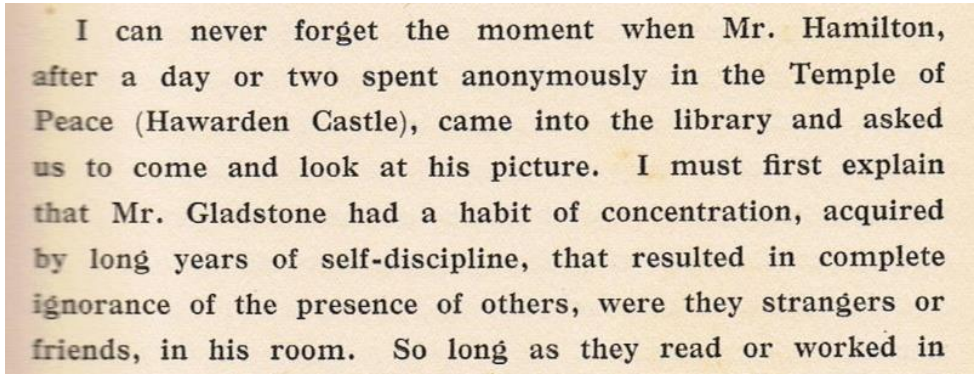
使用OCR获取文本

- 光学字符识别(Optical character recognition, OCR) 是指将文本图像转换为机器可读文本格式的流程

```
from PIL import Image
import pytesseract

# If you don't have tesseract executable in your PATH, include the following:
pytesseract.pytesseract.tesseract_cmd = r'PATH_TO_TESSERACT_EXECUTABLE'

# Simple image to string
print(pytesseract.image_to_string(Image.open('PATH_TO_IMAGE')))
```


I can never forget the moment when Mr. Hamilton, after a day or two spent anonymously in the Temple of Peace (Hawarden Castle), came into the library and asked us to come and look at his picture. I must first explain that Mr. Gladstone had a habit of concentration, acquired by long years of self-discipline, that resulted in complete ignorance of the presence of others, were they strangers or friends, in his room. So long as they read or worked in

```
I can never forget the moment when Mr. Hamilton,
after a day or two spent anonymously in the Temple of
Peace (Hawarden Castle), came into the library and asked
ws to come and look at his picture. I must first explain
that Mr. Gladstone had a habit of concentration, acquired
by long years of self-discipline, that resulted in complete
ignorance of the presence of others, were they strangers or

friends, in his room. So long as they read or worked in
```

从文本中抽取知识

3月20日消息，微软旗下语音识别子公司 Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

从以上文本中，我们能获取哪些知识？

从文本中抽取知识

3月20日消息，微软旗下语音识别子公司 Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

实体间关系

头实体	关系	尾实体
Nuance	是子公司	微软
Nuance	有产品	DAX Express

实体类别

实体	类别
Nuance	公司
DAX Express	软件

事件

类别	*产品发布*
发布公司	Nuance
发布产品	DAX Express
时间	3月20日
地点	NULL

如何自动抽取？

命名实体识别 (Named Entity Recognition, NER)

3月20日消息，微软旗下语音识别子公司Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

目标输出：

实体提及[1]	位置	类别
3月20日	(0, 4)	日期
微软	(8, 9)	机构
Nuance	(20, 25)	机构
DAX Express	(46, 56)	产品

- 识别文本中的实体（如人名、地名、机构名、产品名）并分类
  - 要求1：确定实体的位置（Span）
  - 要求2：实体分类
- 要识别的实体类别根据实际应用场景预先设定

[1] 英文为Entity Mention，简称mention, 在没有歧义的情况下也可简称为实体

细粒度实体分类（Fine-grained Entity Typing, FET）

3月20日消息，微软旗下语音识别子公司Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

目标输出

实体提及[1]	位置	类别
3月20日	(0, 4)	/日期
微软	(8, 9)	/机构, /机构/公司, /机构/公司/科技公司
Nuance	(20, 25)	/机构, /机构/公司, /机构/公司/科技公司
DAX Express	(46, 56)	/产品, /产品/软件

- 给定一个细粒度实体类别体系，对文本中已识别出的实体进行分类
- 先用NER先识别出实体位置，再用FET对它们分类

关系抽取 (Relation Extraction)

3月20日消息，微软旗下语音识别子公司 Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

目标输出：

头实体	关系类别	尾实体
Nuance	是子公司	微软
Nuance	有产品	DAX Express

- 基于预定义好的关系类别，抽取文本中实体间的关系
- 基本实现流程
  - 识别实体
  - 结合上下文将每组实体对分类

注：有时只执行了分类部分也称“关系抽取”

事件抽取 (Event Extraction)

Trigger

3月20日消息，微软旗下语音识别子公司 Nuance今日发布一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

目标输出：

触发词	发布
类别	*产品发布*
发布公司	Nuance
发布产品	DAX Express
时间	3月20日
地点	NULL

- ***事件***是由一个**事件触发词（Event Trigger）**和一些对应的**事件元素（Event Arguments）**所组成的结构。
  - 事件触发词: 最明确地表示了事件的发生的词。
  - 事件元素：与事件相关的实体。

事件抽取 (Event Extraction)

Trigger

3月20日消息，微软旗下语音识别子公司 Nuance今日**发布**一款 AI 临床笔记软件，命名为DAX Express，主要面向医护人员。

- 两个子任务
  - 事件识别
    - 识别事件触发词（Event Trigger）
    - 事件分类
  - 事件元素（Event Argument）抽取
    - 元素识别
    - 元素角色分类

触发词	发布
类别	*产品发布*
发布公司	Nuance
发布产品	DAX Express
时间	3月20日
地点	NULL

从文本中抽取知识

- 命名实体识别
- 细粒度实体分类
- 关系抽取
- 事件抽取