



中国科学院大学

University of Chinese Academy of Sciences

自然语言处理

第1讲 绪论

王石 资康莉 刘瑜

2026年春季课程

<https://ictkc.github.io/teaching/>

本课程考核方式

- 平时作业（共5个）： 50%
 - 阅读报告： 论文阅读报告
 - 动手实验： 提交设计思路、实现代码、实验报告
- 期末考试： 50%
- 面对面建群： 2026
- 课件： <https://ictkc.github.io/teaching/>

主讲教师



王石 正高级工程师

研究方向：知识工程；自然语言处理；认知计算

所属部门：认知智能系统研究中心、智能算法安全重点实验室

导师类别：硕导计算机软件与理论

联系方式：wangshi@ict.ac.cn

个人网页：https://ictkc.github.io

简 历：

王石，博士，正高级工程师，知识工程课题组负责人。其研究方向聚焦于知识工程、自然语言处理、认知计算等领域，先后主持国家重点研发计划课题、国防专项、国家信息安全计划、北京市科技新星计划等项目10余项，发表ACL、EMNLP、AAAI等人工智能、知识工程等领域高水平论文30余篇，参与编写国家标准1项，任人工智能学会心智计算专委会副秘书长、中国指挥与控制学会大模型与决策智能专委会常务委员等职务，获2022年度中国通讯学会科技进步一等奖。更多信息见个人主页：<https://ictkc.github.io>。

助教

照片

简介

照片

简介



第一讲 绪论



中国科学院大学
University of Chinese Academy of Sciences

目 录

1

人人都爱NLP

2

3

4

企业爱NLP

- 大模型：OpenAI、谷歌、阿里、DeepSeek
- 搜索：百度、谷歌、微软
- 推荐：字节、京东、阿里、亚马逊
- 情报：Palantir、大摩
- 问答：金融、电信、宇树、华为、小米
- 翻译：讯飞、百度、谷歌
-

美国 and 中国 市值最大的公司

美国 2025年6月26日 单位：亿美元 中国

1. NVIDIA \$3.763万亿	1. Tencent 腾讯 \$5928.9亿
2. Microsoft \$3.658万亿	2. ICBC 中国工商银行 \$3596.8亿
3. Apple \$3.010万亿	3. 中国农业银行 \$2937亿
4. amazon \$2.250万亿	4. Alibaba \$2808亿
5. Alphabet (Google) \$2.075万亿	5. 中国建设银行 \$2618.8亿
6. Meta Platforms (Facebook) \$1.781万亿	6. 贵州茅台酒股份有限公司 \$2488.9亿
7. BROADCOM \$1.244万亿	7. 中国银行 \$2440.9亿
8. TESLA \$1.055万亿	8. 中国移动 \$2423.4亿
9. BERKSHIRE HATHAWAY INC. \$1.049万亿	9. PetroChina \$2170.7亿
10. JPMorganChase \$7894.2亿	10. Xiaomi \$1870.2亿

来源: Companies Market Cap

企业爱NLP

□ 大模型：OpenAI、Google、阿里、DeepSeek



企业爱NLP

□ 大模型：OpenAI、Google、阿里、DeepSeek

以自然语言处理为核心技术的企业有哪些？

Qwen3.5-Plus 12:51:26

以自然语言处理（NLP）为核心技术的企业可以分为以下几类：**科技巨头**（拥有自研大模型和全栈能力）、**垂直领域AI独角兽**（专注NLP技术落地）、**传统软件转型企业**以及**开源生态领军者**。

结合2025-2026年的行业格局，以下是全球及中国范围内具有代表性的NLP核心企业：

一、中国领军企业

中国在NLP领域已形成“大厂引领 + 独角兽深耕”的格局，特别是在中文语义理解、大模型应用落地方面处于全球第一梯队。

1. 科技巨头（全栈能力）

- **百度 (Baidu)**
 - 核心技术：文心一言 (ERNIE Bot) 系列。百度是国内最早布局NLP的大厂，其ERNIE模型在中文理解上具有深厚积累。
 - 应用场景：智能搜索、自动驾驶（小度车载OS）、智能客服、文心一言API服务。
- **阿里巴巴 (Alibaba)**
 - 核心技术：通义千问 (Qwen) 系列。Qwen模型在开源社区（如Hugging Face）表现极佳，尤其在长文本处理和代码生成方面领先。
 - 应用场景：电商智能导购（淘宝/天猫）、钉钉智能助理、阿里云百炼平台、跨境翻译。
- **腾讯 (Tencent)**
 - 核心技术：混元大模型 (Hunyuan)。具备强大的多模态和自然语言处理能力，深度整合进微信、广告和游戏业务。
 - 应用场景：微信智能助手、腾讯云TI平台、游戏NPC对话生成、广告文案自动生成。

二、国际领军企业

全球NLP技术主要由美国科技巨头主导，同时涌现出一批专注于大模型基础设施的创新公司。

1. 全球科技巨头

- **OpenAI**
 - 地位：NLP领域的绝对领跑者，GPT系列（GPT-4o, o1等）定义了现代大语言模型的标准。
 - 核心能力：极强的通用推理、代码生成、多模态理解能力。
- **Google (DeepMind)**
 - 地位：Transformer架构的发源地之一，拥有Gemini系列模型和BERT等经典模型。
 - 核心能力：搜索增强、多语言翻译、学术研究与产业应用的深度结合。
- **Meta (Facebook)**
 - 地位：开源大模型的推动者，Llama系列（Llama 3/4）是全球最流行的开源基座模型，极大地降低了NLP技术的应用门槛。
- **Microsoft**
 - 地位：通过投资OpenAI并深度集成到Copilot体系中，将NLP能力植入Office、Windows和Azure云平台，是B端落地的典范。
- **Anthropic**
 - 地位：由前OpenAI高管创立，主打Claude系列模型，以“安全性”和“长上下文”著称，深受企业和开发者信赖。

企业爱NLP

□ 搜索：百度、谷歌、微软

Baidu 百度 自然语言处理(NLP)的应用

[自然语言处理 - 百度百科](#)

整体的NLP技术体系

适用领域：计算机、人工智能
简介：自然语言处理(Natural Language Processing, NLP)是人工智能领域的重要研究方向, 融合了语言学、计算机科学、机器学习、...[详情>](#)

[发展历史](#) [基本技术](#) [主要方法](#) [应用](#)

百度百科

[nlp应用 - 百度百科](#)

自然语言处理技术已广泛应用于电子邮件筛选、智能助手、搜索引擎、自动更正、在线翻译、智能客服及文本分析等领域。神经语言程序学则被应用于人际沟通、个人成长、心理咨询、企业管理及教育培训等多个方面。神经语言程序学的有效性不被主流心理学界所认可。基...

[基本介绍](#) [NLP讲师](#)

百度百科

[阿里云-自然语言理解-30天免费试用-数据可视化工具](#)

阿里云_自然语言理解,0元试用,强大的数据建模,丰富的数据可视化图表,快速搭建数据门户,自然语言理解,不仅是业务人员查看数据的工具,更是数据化运营的助推器。

服务行业: [互联网](#) [汽车](#) [电力能源](#) [零售](#) [更多>](#)

Microsoft Bing 图片 ...

国内版 国际版

出现这些迹象钱可能被转走

冰, 由内而外透出光

增值电信业务经营许可证: 合字B2-20090007 © 2026 Microsoft

京ICP备10036305号-7 京公网安备11010802047360号

[隐私与 Cookie](#) [法律声明](#) [广告](#) [关于我们的广告](#)

[帮助](#) [反馈](#)

Gmail 图片

Google

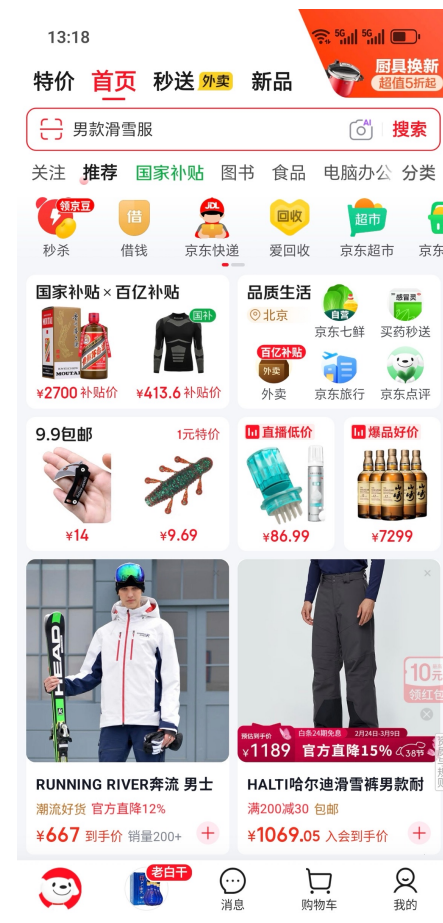
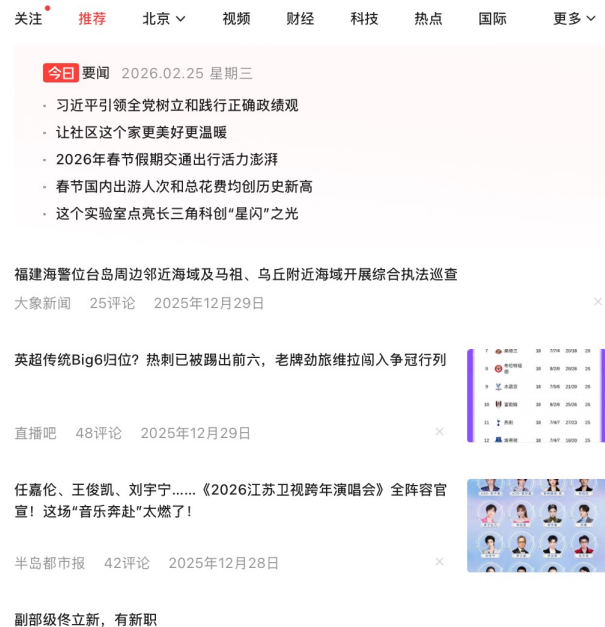
+

Google 搜索 手气不错

Google 提供: [日本語](#)

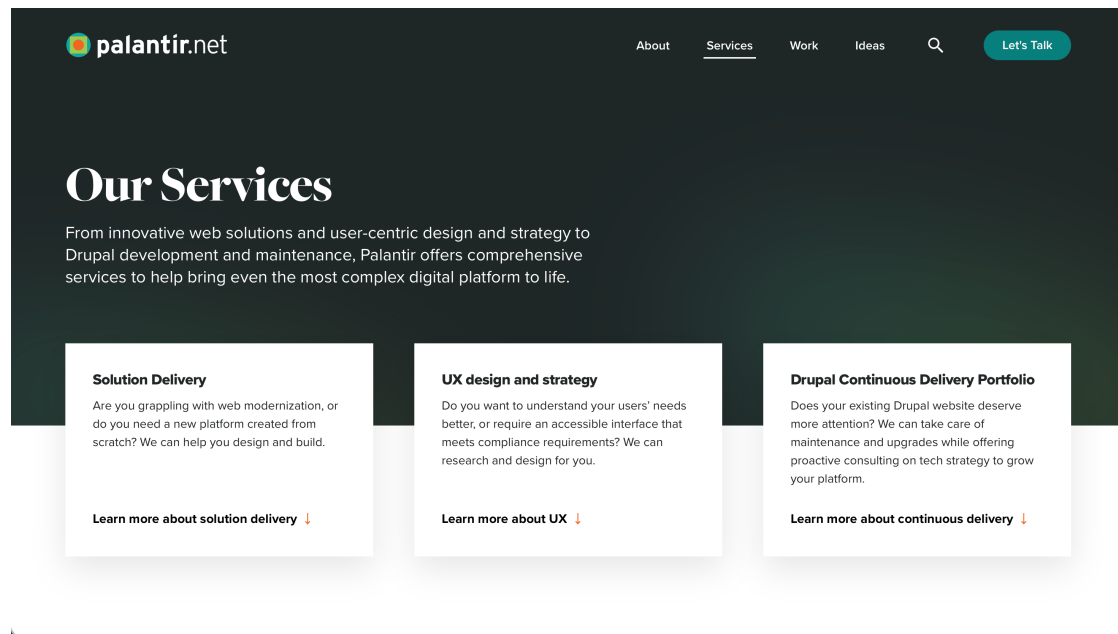
企业爱NLP

□ 推荐：字节、京东、阿里、亚马逊



企业爱NLP

□ 情报：Palantir、大摩



palantir.net

About Services Work Ideas Let's Talk

Our Services

From innovative web solutions and user-centric design and strategy to Drupal development and maintenance, Palantir offers comprehensive services to help bring even the most complex digital platform to life.

Solution Delivery

Are you grappling with web modernization, or do you need a new platform created from scratch? We can help you design and build.

[Learn more about solution delivery ↓](#)

UX design and strategy

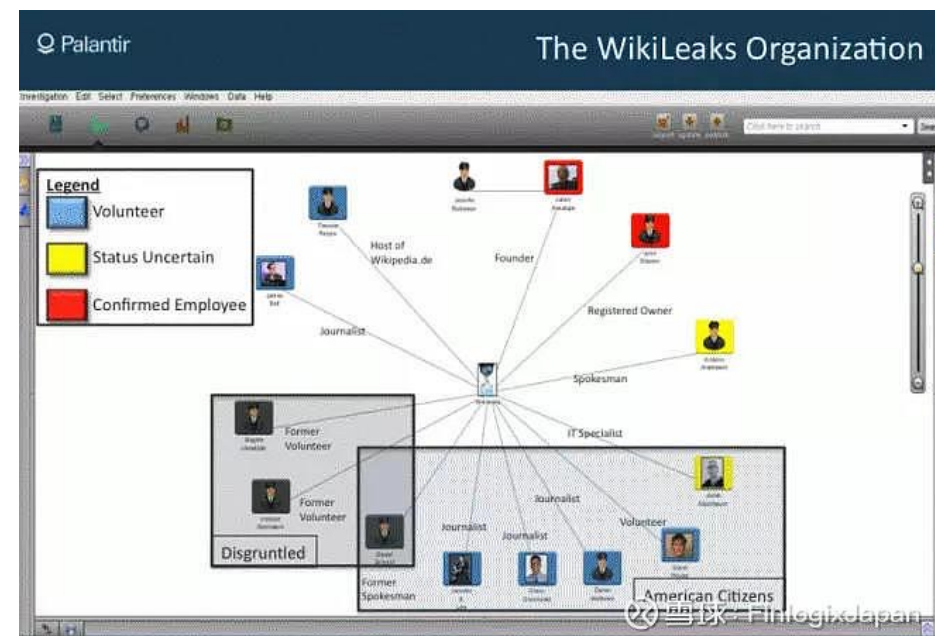
Do you want to understand your users' needs better, or require an accessible interface that meets compliance requirements? We can research and design for you.

[Learn more about UX ↓](#)

Drupal Continuous Delivery Portfolio

Does your existing Drupal website deserve more attention? We can take care of maintenance and upgrades while offering proactive consulting on tech strategy to grow your platform.

[Learn more about continuous delivery ↓](#)



企业爱NLP

□ 问答：金融、电信、宇树、华为、小米



企业爱NLP

□ 翻译：讯飞、百度、谷歌



科学爱NLP

VOL. LIX. No. 236.] [October, 1950

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING



1. *The Imitation Game.*

I PROPOSE to consider the question, ‘Can machines think?’

.....

We may hope that machines will eventually compete with men in all purely intellectual fields. But which are the best ones to start with? Even this is a difficult decision. Many people think that a very abstract activity, like the playing of chess, would be best. It can also be maintained that it is best to provide the machine with the best sense organs that money can buy, and then teach it to understand and speak English. This process could follow the normal teaching of a child. Things would be pointed out and named, etc. Again I do not know what the right answer is, but I think both approaches should be tried.

We can only see a short distance ahead, but we can see plenty there that needs to be done.



A.M. TURING AWARD WINNERS:

Barto, Andrew (2024)
Bengio, Yoshua (2018)
Feigenbaum, Edward A ("Ed") (1994)
Hinton, Geoffrey E (2018)
LeCun, Yann (2018)
McCarthy, John (1971)
Minsky, Marvin (1969)
Newell, Allen (1975)
Pearl, Judea (2011)
Reddy, Dabbala Rajagopal ("Raj") (1994)
Simon, Herbert ("Herb") Alexander (1975)
Sutton, Richard (2024)
Valiant, Leslie Gabriel (2010)

SELECT A RESEARCH SUBJECT:

Analysis & design of algorithms **Artificial intelligence**
Compilers Computer architecture **Cryptography & security** Databases
Automata theory
Graphics Human-computer interaction Communications Numerical computing
Operating systems **Programming languages**
Theoretical computer science
Programming methodology & software engineering Verification of hardware & software
Machine learning Hardware-software system design



在人工智能领域的13个图灵奖获得者中，有8位与NLP有密切关联

<https://amturing.acm.org/bysubject.cfm?cat=11>

政府爱NLP



中华人民共和国中央人民政府

www.gov.cn

发布日期： 2025年08月26日

国务院关于深入实施“人工智能+”行动的意见

国发〔2025〕11号

各省、自治区、直辖市人民政府，国务院各部委、各直属机构：

为深入实施“人工智能+”行动，推动人工智能与经济社会各行业各领域广泛深度融合，重塑人类生产生活范式，促进生产力革命性跃迁和生产关系深层次变革，加快形成人机协同、跨界融合、共创分享的智能经济和智能社会新形态，现提出如下意见。

一、总体要求

以习近平新时代中国特色社会主义思想为指导，完整准确全面贯彻新发展理念，坚持以人民为中心的发展思想，充分发挥我国数据资源丰富、产业体系完备、应用场景广阔等优势，强化前瞻谋划、系统布局、分业施策、开放共享、安全可控，以科技、产业、消费、民生、治理、全球合作等领域为重点，深入实施“人工智能+”行动，涌现一批新基础设施、新技术体系、新产业生态、新就业岗位等，加快培育发展新质生产力，使全体人民共享人工智能发展成果，更好服务中国式现代化建设。

到2027年，率先实现人工智能与6大重点领域广泛深度融合，新一代智能终端、智能体等应用普及率超70%，智能经济核心产业规模快速增长，人工智能在公共治理中的作用明显增强，人工智能开放合作体系不断完善。到2030年，我国人工智能全面赋能高质量发展，新一代智能终端、智能体等应用普及率超90%，智能经济成为我国经济发展的重要增长极，推动技术普惠和成果共享。到2035年，我国全面步入智能经济和智能社会发展新阶段，为基本实现社会主义现代化提供有力支撑。

https://www.gov.cn/zhengce/content/202508/content_7037861.htm

政府爱NLP

(一) “人工智能+”科学技术

1.加速科学发现进程。加快探索人工智能驱动的新型科研范式，加速“从0到1”重大科学发现进程。加快科学大模型建设应用，推动基础科研平台和重大科技基础设施智能化升级，打造开放共享的高质量科学数据集，提升跨模态复杂科学数据处理水平。强化人工智能跨学科牵引带动作用，推动多学科融合发展。

2.驱动技术研发模式创新和效能提升。推动人工智能驱动的技术研发、工程实现、产品落地一体化协同发展，加速“从1到N”技术落地和迭代突破，促进创新成果高效转化。支持智能化研发工具和平台推广应用，加强人工智能与生物制造、量子科技、第六代移动通信（6G）等领域技术协同创新，以新的科研成果支撑场景应用落地，以新的应用需求牵引科技创新突破。

3.创新哲学社会科学研究方法。推动哲学社会科学研究方法向人机协同模式转变，探索建立适应人工智能时代的新型哲学社会科学研究组织形式，拓展研究视野和观察视域。深入研究人工智能对人类认知判断、伦理规范等方面的深层次影响和作用机理，探索形成智能向善理论体系，促进人工智能更好造福人类。

政府爱NLP

(二) “人工智能+”产业发展

1.培育智能原生新模式新业态。鼓励有条件的企业将人工智能融入战略规划、组织架构、业务流程等，推动产业全要素智能化发展，助力传统产业改造升级，开辟战略性新兴产业和未来产业发展新赛道。大力发展智能原生技术、产品和服务体系，加快培育一批底层架构和运行逻辑基于人工智能的智能原生企业，探索全新商业模式，催生智能原生新业态。

2.推进工业全要素智能化发展。推动工业全要素智能联动，加快人工智能在设计、中试、生产、服务、运营全环节落地应用。着力提升全员人工智能素养与技能，推动各行业形成更多可复用的专家知识。加快工业软件创新突破，大力发展智能制造装备。推进工业供应链智能协同，加强自适应供需匹配。推广人工智能驱动的生产工艺优化方法。深化人工智能与工业互联网融合应用，增强工业系统的智能感知与决策执行能力。

3.加快农业数智化转型升级。加快人工智能驱动的育种体系创新，支持种植、养殖等农业领域智能应用。大力发展智能农机、农业无人机、农业机器人等智能装备，提高农业生产和加工工具的智能感知、决策、控制、作业等能力，强化农机农具平台化、智能化管理。加强人工智能在农业生产管理、风险防范等领域应用，帮助农民提升生产经营能力和水平。

4.创新服务业发展新模式。加快服务业从数字赋能的互联网服务向智能驱动的新型服务方式演进，拓展经营范围，推动现代服务业向智向新发展。探索无人服务与人工服务相结合的新模式。在软件、信息、金融、商务、法律、交通、物流、商贸等领域，推动新一代智能终端、智能体等广泛应用。

政府爱NLP

(三) “人工智能+”消费提质

1.拓展服务消费新场景。培育覆盖更广、内容更丰富的智能服务业态，加快发展提效型、陪伴型等智能原生应用，支持开辟智能助理等服务新入口。加强智能消费基础设施建设，提升文娱、电商、家政、物业、出行、养老、托育等生活服务品质，拓展体验消费、个性消费、认知和情感消费等服务消费新场景。

2.培育产品消费新业态。推动智能终端“万物智联”，培育智能产品生态，大力发展智能网联汽车、人工智能手机和电脑、智能机器人、智能家居、智能穿戴等新一代智能终端，打造一体化全场景覆盖的智能交互环境。加快人工智能与元宇宙、低空飞行、增材制造、脑机接口等技术融合和产品创新，探索智能产品新形态。

政府爱NLP

(四) “人工智能+”民生福祉

1.创造更加智能的工作方式。积极发挥人工智能在创造新岗位和赋能传统岗位方面的作用，探索人机协同的新型组织架构和管理模式，培育发展智能代理等创新型工作形态，推动在劳动力紧缺、环境高危等岗位应用。大力支持开展人工智能技能培训，激发人工智能创新创业和再就业活力。加强人工智能应用就业风险评估，引导创新资源向创造就业潜力大的方向倾斜，减少对就业的冲击。

2.推行更富成效的学习方式。把人工智能融入教育教学全要素、全过程，创新智能学伴、智能教师等人机协同教育教学新模式，推动育人从知识传授为重向能力提升为本转变，加快实现大规模因材施教，提高教育质量，促进教育公平。构建智能化情景交互学习模式，推动开展方式更灵活、资源更丰富的自主学习。鼓励和支持全民积极学习人工智能新知识、新技术。

3.打造更有品质的美好生活。探索推广人人可享的高水平居民健康助手，有序推动人工智能在辅助诊疗、健康管理、医保服务等场景的应用，大幅提高基层医疗健康服务能力和效率。推动人工智能在繁荣文化生产、增强文化传播、促进文化交流中展现更大作为，利用人工智能辅助创作更多具有中华文化元素和标识的文化内容，壮大文化产业。充分发挥人工智能对织密人际关系、精神慰藉陪伴、养老托育助残、推进全民健身等方面的重要作用，拓展人工智能在“好房子”全生命周期的应用，积极构建更有温度的智能社会。

政府爱NLP

(五) “人工智能+”治理能力

1.开创社会治理人机共生新图景。有序推动市政基础设施智能化改造升级，探索面向新一代智能终端发展的城市规划、建设与治理，提升城市运行智能化水平。加快人工智能产品和服务向乡村延伸，推动城乡智能普惠。深入开展人工智能社会实验。安全稳妥有序推进人工智能在政务领域应用，打造精准识别需求、主动规划服务、全程智能办理的政务服务新模式。加快人工智能在各类公共资源招标投标活动中的应用，提升智能交易服务和监管水平。

2.打造安全治理多元共治新格局。推动构建面向自然人、数字人、智能机器人等多元一体的公共安全治理体系，加强人工智能在安全生产监管、防灾减灾救灾、公共安全预警、社会治安管理等方面的应用，提升监测预警、监管执法、指挥决策、现场救援、社会动员等工作水平，增强应用人工智能维护和塑造国家安全的能力。加快推动人工智能赋能网络空间治理，强化信息精准识别、态势主动研判、风险实时处置等能力。

3.共绘美丽中国生态治理新画卷。提高空天地海一体化动态感知和国土空间智慧规划水平，强化资源要素优化配置。围绕大气、水、海洋、土壤、生物等多要素生态环境系统和全国碳市场建设等，提升人工智能驱动的监测预测、模拟推演、问题处置等能力，推动构建智能协同的精准治理模式。

政府爱NLP

(六) “人工智能+”全球合作

1.推动人工智能普惠共享。把人工智能作为造福人类的国际公共产品，打造平权、互信、多元、共赢的人工智能能力建设开放生态。深化人工智能领域高水平开放，推动人工智能技术开源可及，强化算力、数据、人才等领域国际合作，帮助全球南方国家加强人工智能能力建设，助力各国平等参与智能化发展进程，弥合全球智能鸿沟。

2.共建人工智能全球治理体系。支持联合国在人工智能全球治理中发挥主渠道作用，探索形成各国广泛参与的治理框架，共同应对全球性挑战。深化与国际组织、专业机构等交流合作，加强治理规则、技术标准等对接协调。共同研判、积极应对人工智能应用风险，确保人工智能发展安全、可靠、可控。

政府爱NLP



<https://www.ai.gov>



中国科学院大学
University of Chinese Academy of Sciences

目 录

1

人人都爱NLP

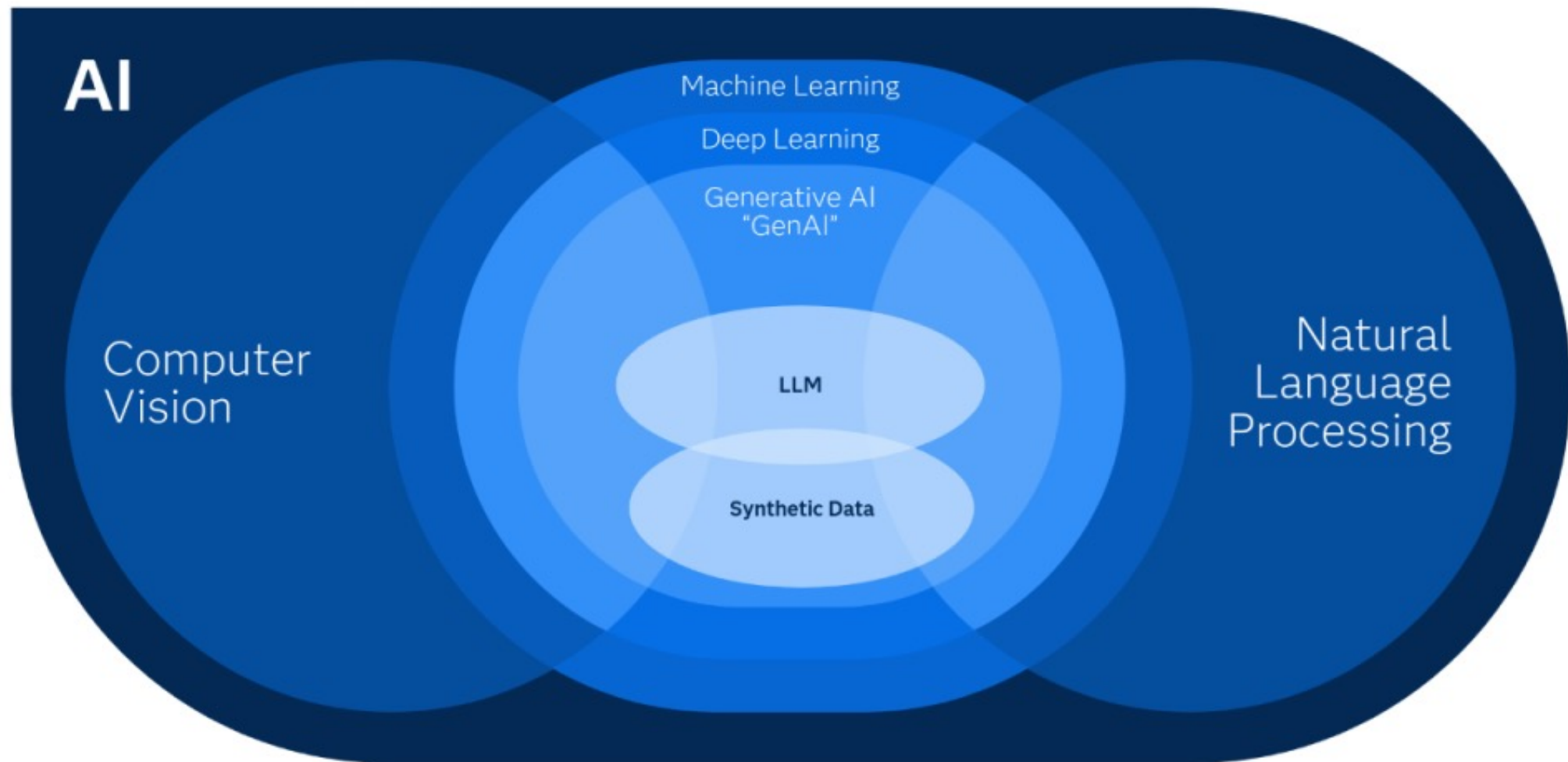
2

Hello, NLP!

3

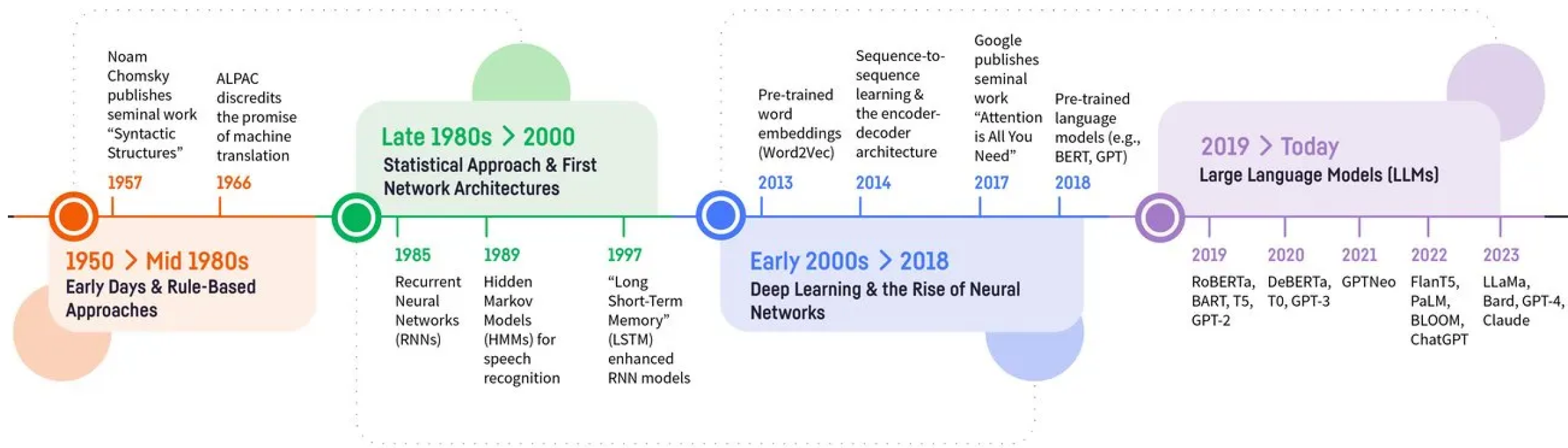
4

NLP在AI大家庭的位置





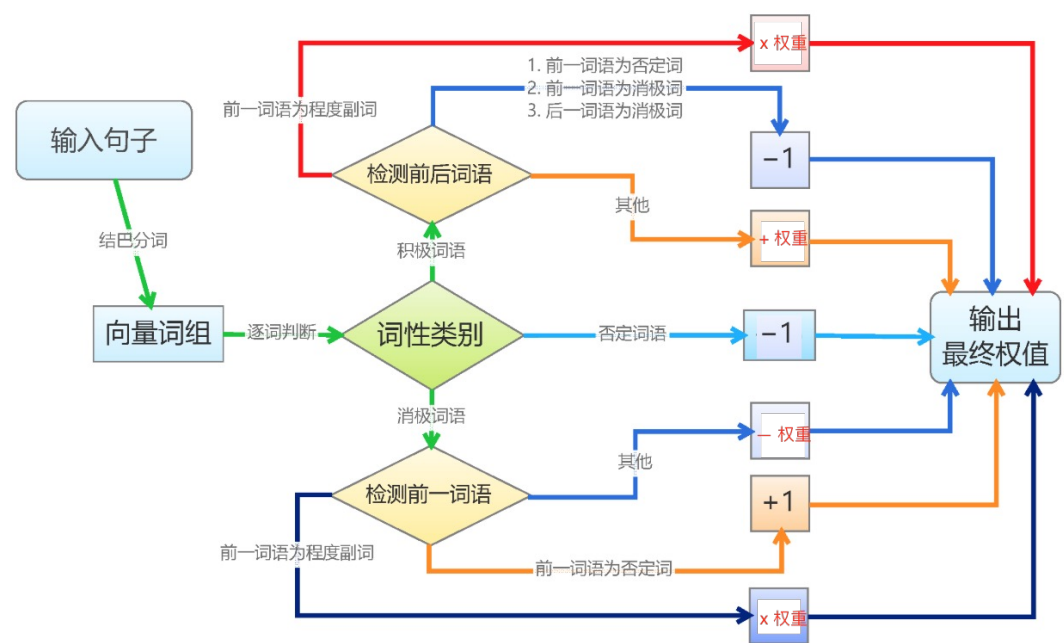
NLP的成长经历



幼儿：规则系统（有点傻）



正向词：好消息



基于情感词典的文本分类-程序框图

https://blog.csdn.net/weixin_41657760

负向词：狗

少年：机器学习（有点难）

雷军 25-12-14 07:10 发布于 北京 来自 Xiaomi 17 Pro

早！转发一个好消息：小米手机射频团队论文成功入选全球半导体与电子器件领域顶会 IEDM 2025! ❤️

51条回复

栽花shi者 25-12-14 08:42 来自浙江
:大疆老板成天在微博炫耀这些吗

PoppinYG 25-12-14 09:36 来自四川
:好

火星第四王朝 25-12-14 09:43 来自江苏
:这是花真金白银砸出来的好消息，这不是炫耀，连人工智能都懂你们不懂，天天喊着打打打，没有这些高科技技术支持，靠你的嘴打吗？

飞胜华 25-12-14 12:07 来自四川
:除了雷总还有哪个大老板天天哄着你啊？

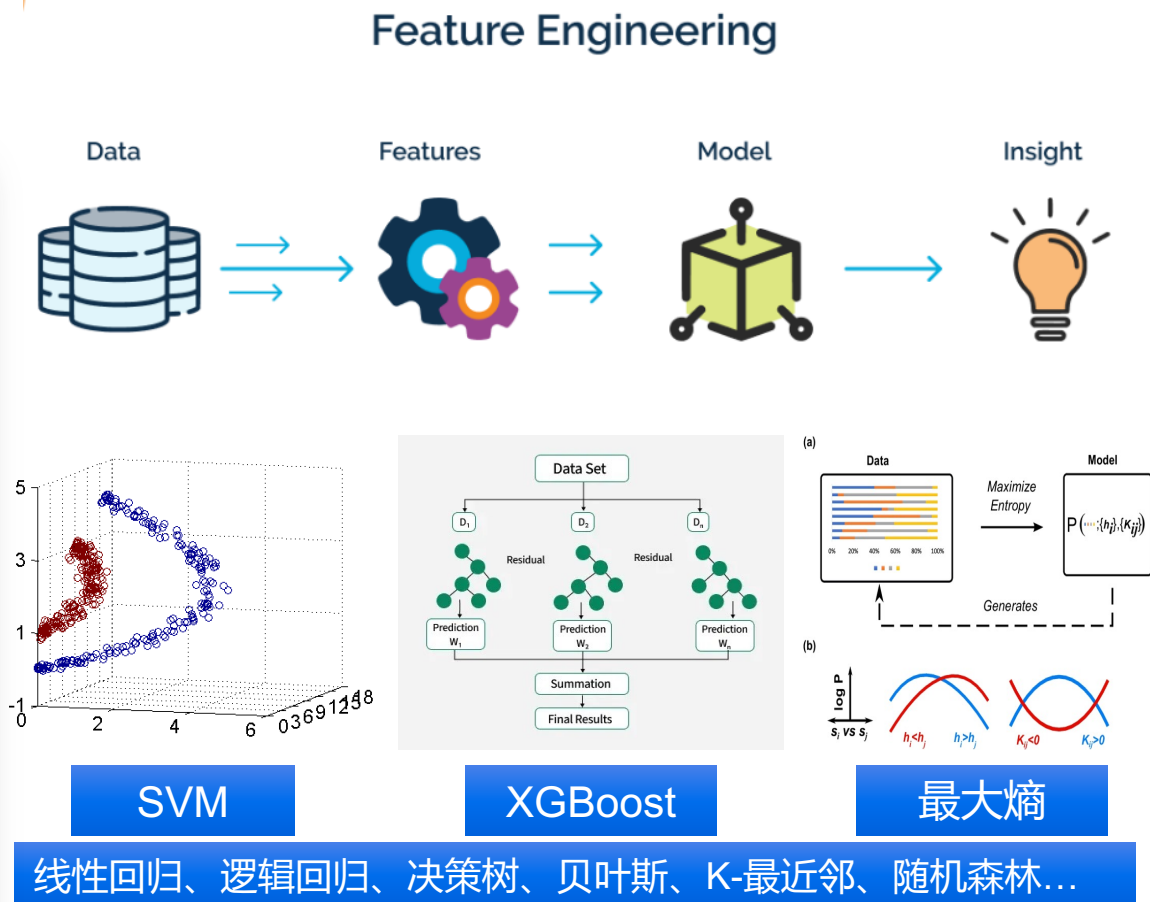
以der付人 25-12-14 10:46 来自上海
:大疆会天天像祥林嫂一样叫：我们被制裁了吗？

彤彤彤彤彤彤 25-12-14 11:46 来自四川
:果然是曾智慧用户

裸K交易元 25-12-14 12:25 来自河北
:大疆老板有雷总一半努力，也不至于就这么点钱

冰是睡着的水的空间 25-12-14 21:02 来自河南
:纯血海狗

度鸟奔云 25-12-14 13:21 来自江苏
:华为的营销:给手机搞走秀



青年：深度学习（有点累）

雷军 25-12-14 07:10 发布于 北京 来自 Xiaomi 17 Pro

早！转发一个好消息：小米手机射频团队论文成功入选全球半导体与电子器件领域顶会 IEDM 2025! ❤️

51条回复

栽花shi者 25-12-14 08:42 来自浙江
大疆老板成天在微博炫耀这些吗

PoppinYG 25-12-14 09:36 来自四川
好

火星第四王朝 25-12-14 09:43 来自江苏
这是花真金白银砸出来的好消息，这不是炫耀，连人工智能都懂你们不懂，天天喊着打打打那，没有这些高科技技术支持，靠你的嘴打吗？

飞胜华 25-12-14 12:07 来自四川
除了雷总还有哪个大老板天天哄着你啊？

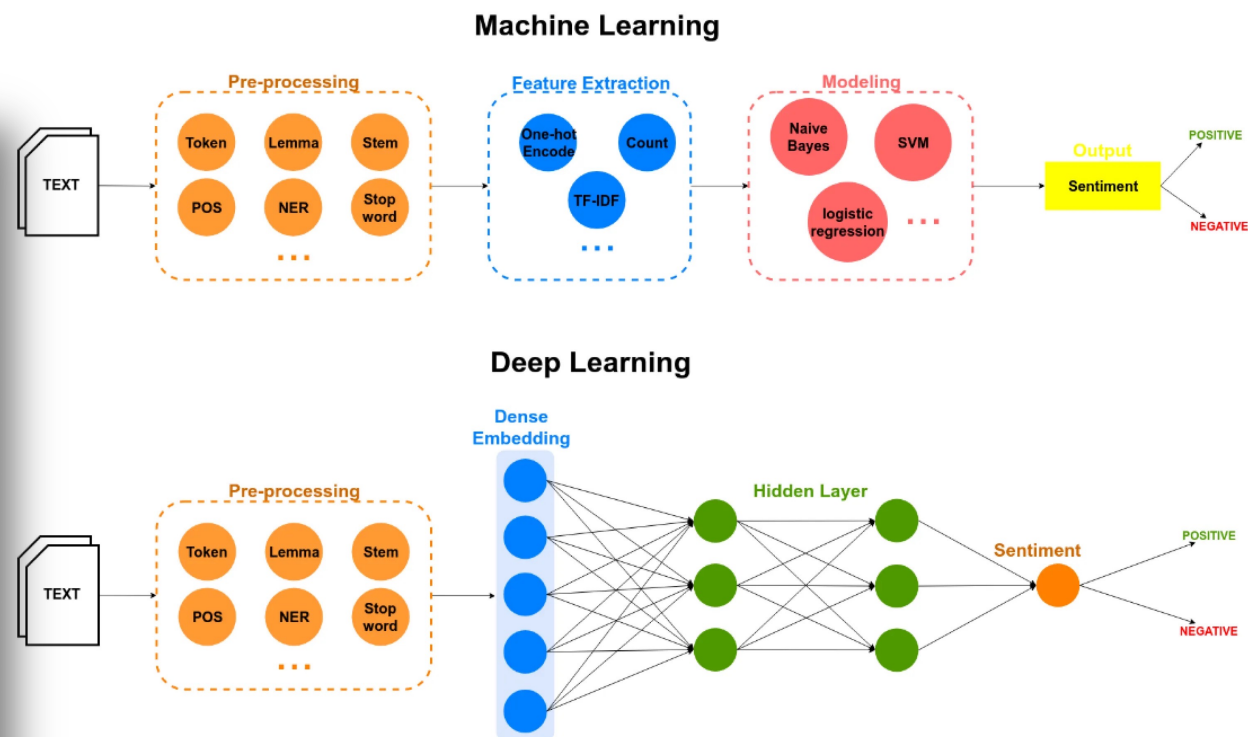
以der付人 25-12-14 10:46 来自上海
大疆会天天像祥林嫂一样叫：我们被制裁了吗？

彤彤彤彤彤 25-12-14 11:46 来自四川
果然是曾智慧用户

裸K交易元 25-12-14 12:25 来自河北
大疆老板有雷总一半努力，也不至于就这么点钱

冰是睡着的水的空间 25-12-14 21:02 来自河南
纯血海狗

度鸟奔云 25-12-14 13:21 来自江苏
华为的营销：给手机搞走秀



无需特征工程，但对标注数据量要求高

壮年：预训练模型（有点戏）



这句话对雷总是支持还是反对：“大疆老板有雷总一半努力，也不至于就这么点钱”

已思考（用时 13 秒）

要准确分析这句话的情感倾向，关键在于识别其核心是通过抬高一方来贬低另一方的“拉踩”式反讽。

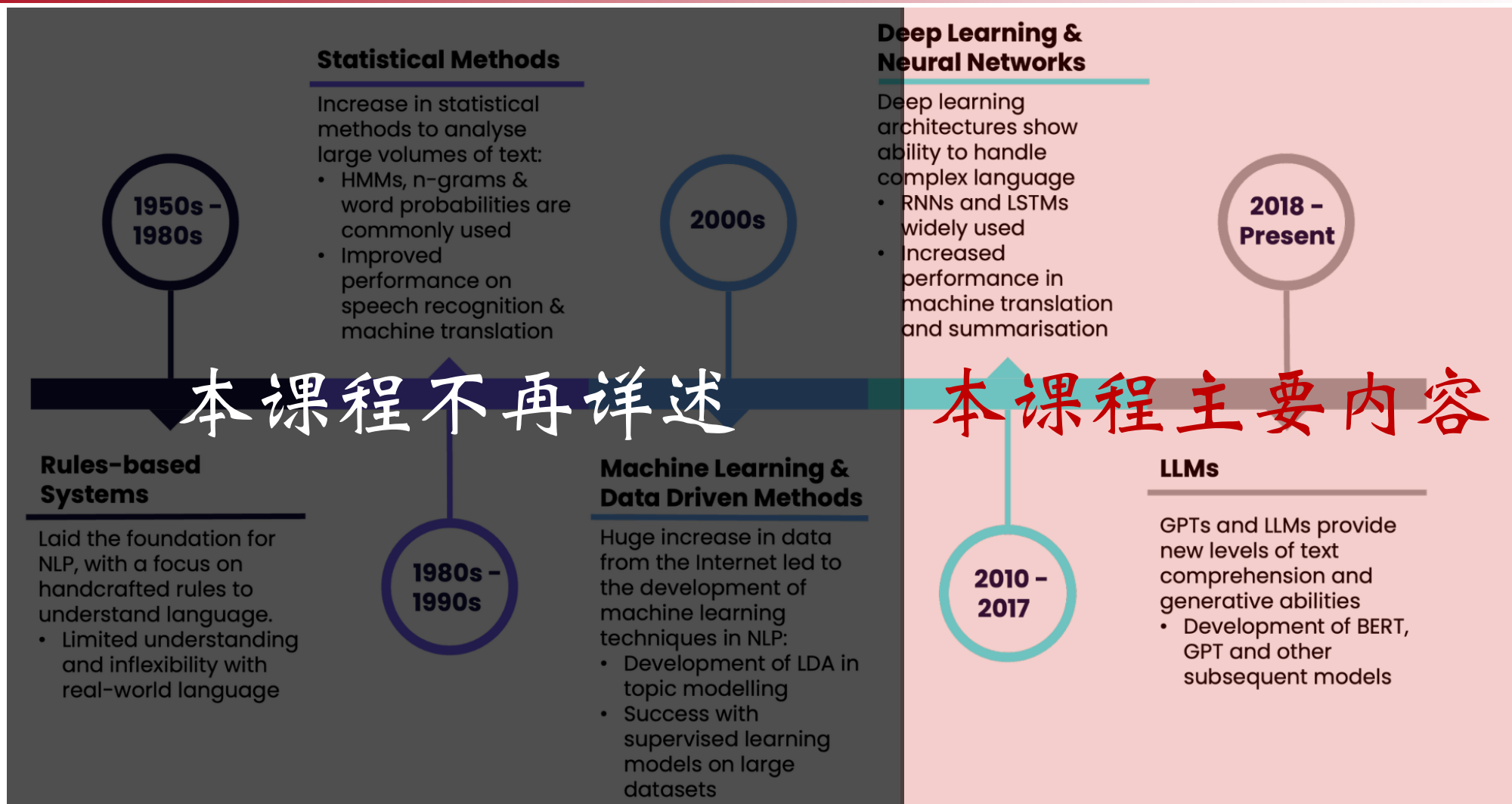
情感判定：支持雷总

这句话对雷军（雷总）的情感倾向是明确的赞扬和支持，具体分析如下：

分析维度	具体解释	情感指向
直接表达	“有雷总一半努力”直接将雷军树立为“努力”的标杆和榜样。	支持雷军
反讽逻辑	句子真正想表达的是：大疆老板不够努力 → 所以成就（“钱”）不够大。而“努力”的参照物正是雷军，这反向强化了雷军“极其努力”的正面形象。	支持雷军
潜在预设	这句话默认听众都认同“雷军的巨大成功源于他的超人努力”这个前提，是在巩固和传播这一正面叙事。	支持雷军

⚠ 注意：支持的复杂性

本课程内容





目 录

1

人人都爱NLP

2

Hello, NLP!

3

NLP任务概览

4

NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation		Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation	Clause boundary detection	Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

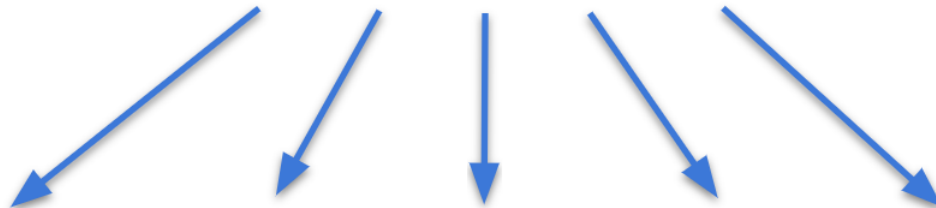
1. Tokenization

Tokenization是NLP的基本任务，把一段连续的文本序列（如句子、段落等）切分为一个字符串序列（如单词、短语、字符、标点等），其中的元素称为token或词语。**Token是NLP任务的原子计算单元**

“ Tokenization is essential in NLP! ”



Tokenization



“Tokenization”

“is”

“essential”

“in”

“NLP”

1. Tokenization

1.词粒度Tokenization

词粒度Tokenization是最直观的分词方式，即是指将文本按照词汇words进行切分。例如：

1. The quick brown fox jumps over the lazy dog.
- 2.
3. 词粒度Tokenized结果:
- 4.
5. ['The', 'quick', 'brown', 'fox', 'jumps', 'over', 'the', 'lazy', 'dog', '.']

如若是中文文本，则通常是按照词典收录的标准词汇或是通过分词算法识别出的短语、成语、专有名词等进行切分。例如：

- | | | | |
|----|-----------------------------|--------------------|---------------------|
| 1. | 我喜欢吃苹果。 | 他/ 说/ 的/ 确实/ 在理/ 。 | 门/ 把/ 手/ 弄/ 坏/ 了/ 。 |
| 2. | | | |
| 3. | 词粒度Tokenized结果: | 他/ 说/ 的确/ 实在/ 理/ 。 | 门/ 把手/ 弄/ 坏/ 了/ 。 |
| 4. | | | |
| 5. | ['我', '喜欢', '吃', '苹果', '.'] | | |

1. Tokenization

2. 字符粒度Tokenization

字符粒度Tokenization将文本分割成最小的字符单元，即每个字符被视为一个单独的token。例如：

1. Hello, world!
- 2.
3. 字符粒度Tokenized结果:
- 4.
5. ['H', 'e', 'l', 'l', 'o', ',', ' ', 'w', 'o', 'r', 'l', 'd', '!']

字符粒度Tokenization在中文中是将文本按照每个独立的汉字进行切分。

1. 我喜欢吃苹果。
- 2.
3. 字符粒度Tokenized结果:
- 4.
5. ['我', '喜', '欢', '吃', '苹', '果', '。']

1. Tokenization

3.subword粒度Tokenization

subword粒度Tokenization介于词粒度和字符粒度之间，它将文本分割成介于单词和字符之间的子词（subwords）作为token。常见的subword Tokenization方法包括Byte Pair Encoding (BPE)、WordPiece等。这些方法通过统计文本数据中的子串频率，自动生成一种分词词典，能够有效应对未登录词（OOV）问题，同时保持一定的语义完整性。

```
1. helloworld
```

[复制](#)

假设经过BPE算法训练后，生成的子词词典包含以下条目：

```
1. h, e, l, o, w, r, d, hel, low, wor, orld
```

[复制](#)

子词粒度Tokenized结果：

```
1. ['hel', 'low', 'orld']
```

[复制](#)

这里，“helloworld”被切分为三个子词“hel”，“low”，“orld”，这些都是词典中出现过的高频子串组合。这种切分方式既能处理未知词汇（如“helloworld”并非标准英语单词），又保留了一定的语义信息（子词组合起来能还原原始单词）。

1. Tokenization

3.subword粒度Tokenization

subword粒度Tokenization介于词粒度和字符粒度之间，它将文本分割成介于单词和字符之间的子词（subwords）作为token。常见的subword Tokenization方法包括Byte Pair Encoding (BPE)、WordPiece等。这些方法通过统计文本数据中的子串频率，自动生成一种分词词典，能够有效应对未登录词（OOV）问题，同时保持一定的语义完整性。

在中文中，subword粒度Tokenization同样是将文本分割成介于汉字和词语之间的子词作为token。例如：

1. 我喜欢吃苹果

复制

假设经过BPE算法训练后，生成的子词词典包含以下条目：

1. 我，喜，欢，吃，苹，果，我喜欢，吃苹果

复制

子词粒度Tokenized结果：

1. ['我', '喜欢', '吃', '苹果']

复制

2. Named Entity Recognition (命名实体识别)

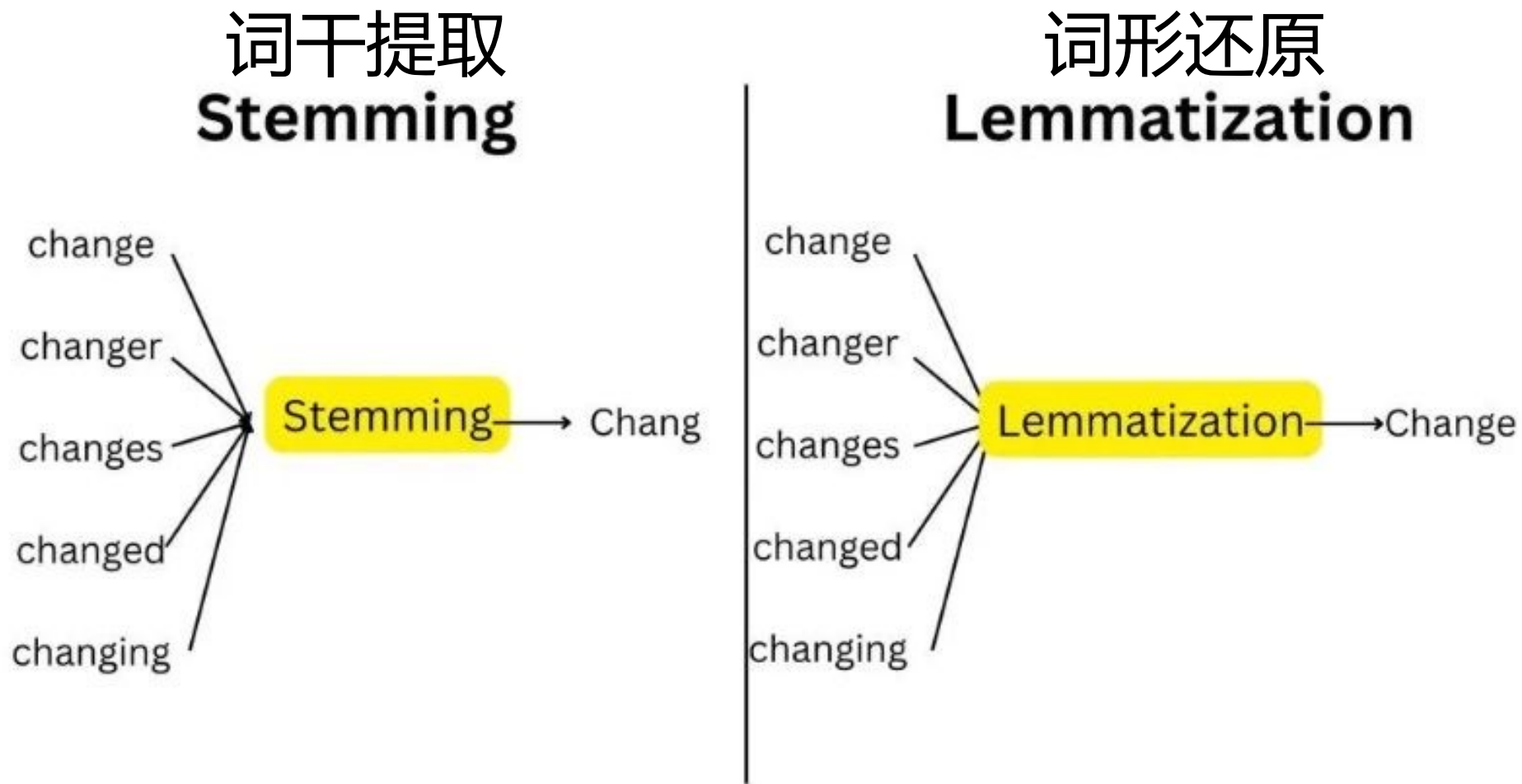


3. Part-of-speech Tagging (词性标注)

我/r 爱/v 自然语言/n 处理/v? n?

名词(n)、时间词(t)、处所词(s)、方位词(f)、数词(m)、量词(q)、区别词(b)、
代词(r)、动词(v)、形容词(a)、状态词(z)、副词(d)、介词(p)、连词(c)、助
词(u)、语气词(y)、叹词(e)、拟声词(o)、成语(i)、习用语(l)、简称(j)、前
接成分(h)、后接成分(k)、语素(g)、非语素字(x)、标点符号(w)

4. Lemmatization / stemming



5. Word Sense Disambiguation (词义消歧)

苹果 [píng guǒ]

这是一个多义词，请在下列义项上选择浏览（共25个义项）

- 苹果：蔷薇科苹果属植物
- 苹果：苹果产品公司
- 苹果：韩国2008年康理贯执导电影
- 苹果：动漫《男子高中生的日常》中角色
- 苹果：谢和弦、E-SO演唱歌曲
- 苹果：安与骑兵演唱歌曲
- 苹果：伊朗1998年莎米拉·玛克玛尔巴夫执导电影
- 苹果：Minecraft游戏中的食物类物品
- 苹果：邓丽欣演唱歌曲
- 苹果：2007年李玉执导电影
- 苹果：网游《天堂梦》中人物
- 苹果：中药
- 苹果：廖俊涛演唱歌曲
- 苹果：2010年法国喜剧电影
- 苹果：游戏《和平精英》中的投掷物
- 苹果：游戏《集合啦！动物森友会》中的岛民角色
- 苹果：2020年赫里斯托斯·尼寇执导电影
- 苹果：2005年中国农业大学出版社出版的图书
- 苹果：2008年北方文艺出版社出版的图书
- 苹果：游戏《仙境传说RO》中的物品道具
- 苹果：蒋先贵(Giorgio)演唱歌曲
- 苹果：南惠萍所作的诗歌
- 苹果：2020年甘肃科学技术出版社出版的图书
- 苹果：2018年甘肃科学技术出版社出版的图书
- 苹果：2020年中信出版社出版的图书

直到2021年12月，科技媒体The Information 发布了一份的报告，介绍了库克为**苹果**在中国发展所做的努力，从而为苹果在中国亮眼的成绩铺平了道路。

可可西里的青稞、洛川的**苹果**、秭归的脐橙……9月，走进南京农副产品物流配送中心，来自西部8个省区市的消费扶贫产品琳琅满目，企业家和消费者在这里边品尝边洽谈选购。

反映都市底层小人物命运与现实生活的艺术片《**苹果**》，11月30日在香港以三级片身份上映，国内则公映了删节版。

苹果产品公司

蔷薇科苹果属植物

2007年李玉执导电影

6. Keyword Extraction

关键词抽取结果:

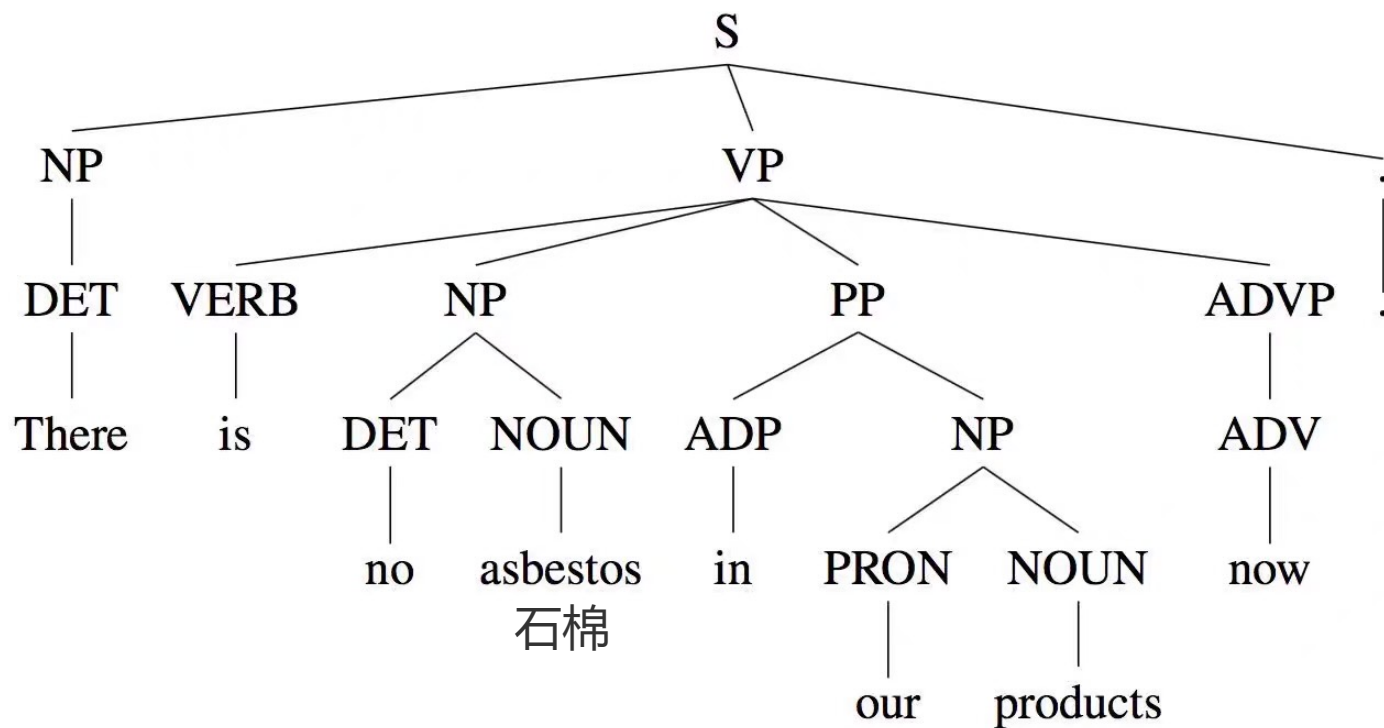
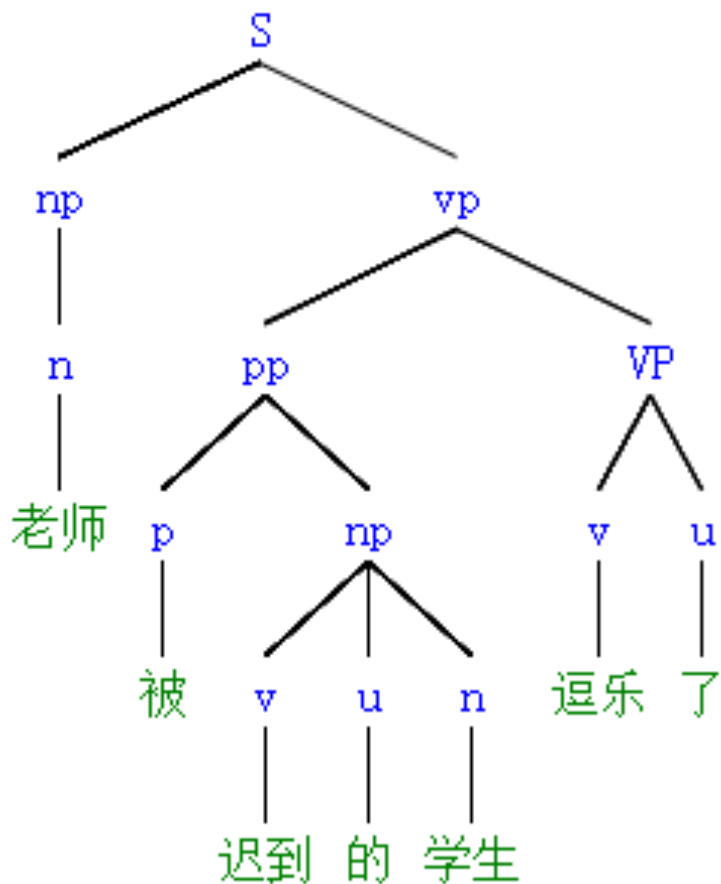
AI 生成 keyword 视频是利用 人工智能 keyword 技术 生成 keyword 视频的过程, 通常, 这种技术基于图像识别、语音 合成 keyword 等算法, 从各种不同来源的数据中获取信息, 然后自动 生成 keyword 视频。 AI 生成 keyword 视频技术的优势是可以快速 生成 keyword 高质量 keyword 视频, 同时还可以自动 完成 keyword 各种冗长的编辑工作, 提高工作效率。此外, 还可以实现一些人类不可能 完成 keyword 的任务, 例如在极端天气、危险环境下 进行 keyword 拍摄。

关键词: 完成; 生成; 人工智能; 进行; 合成; 高质量;

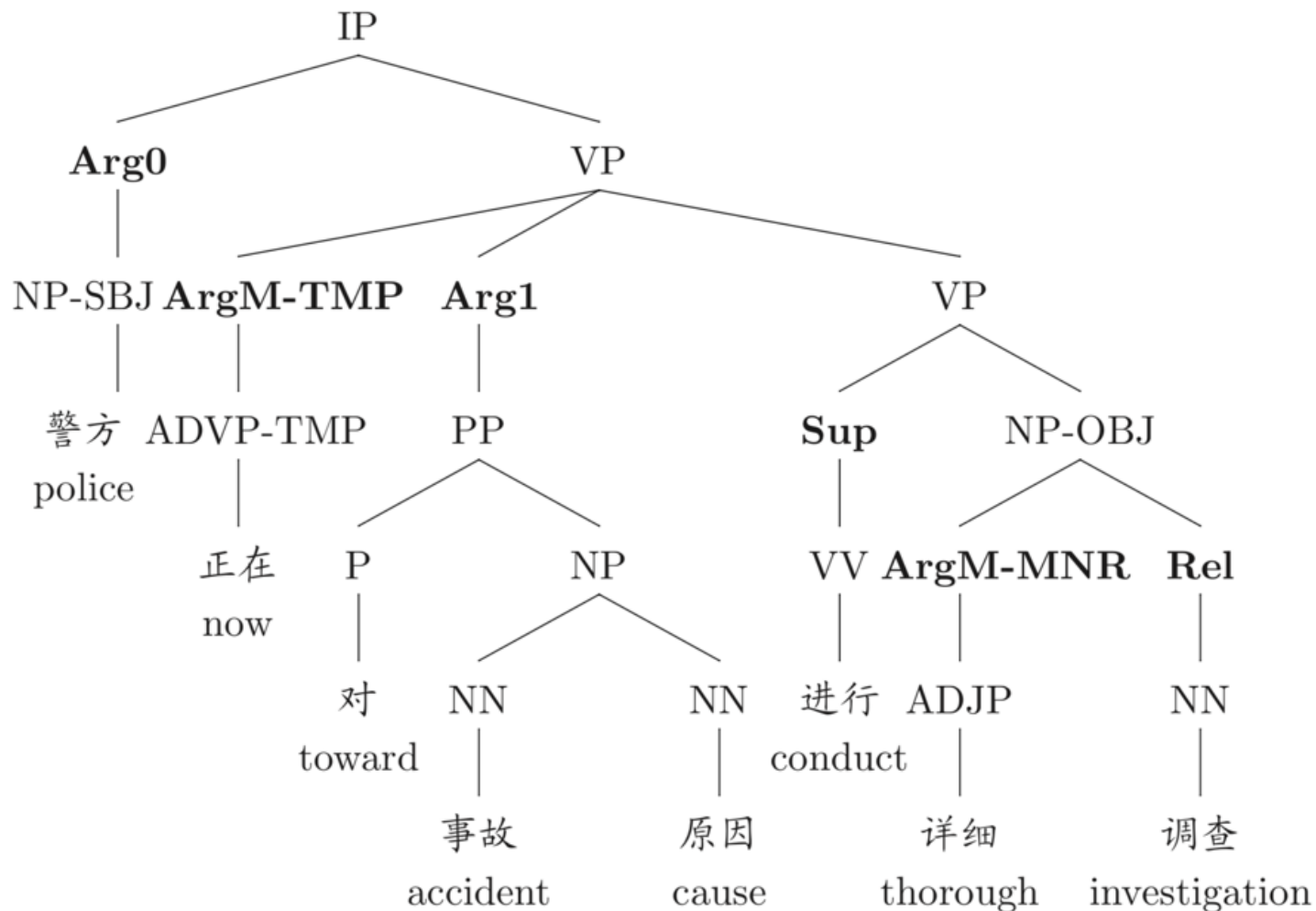
NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation		Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

1. Constituency parsing (句法成分分析)

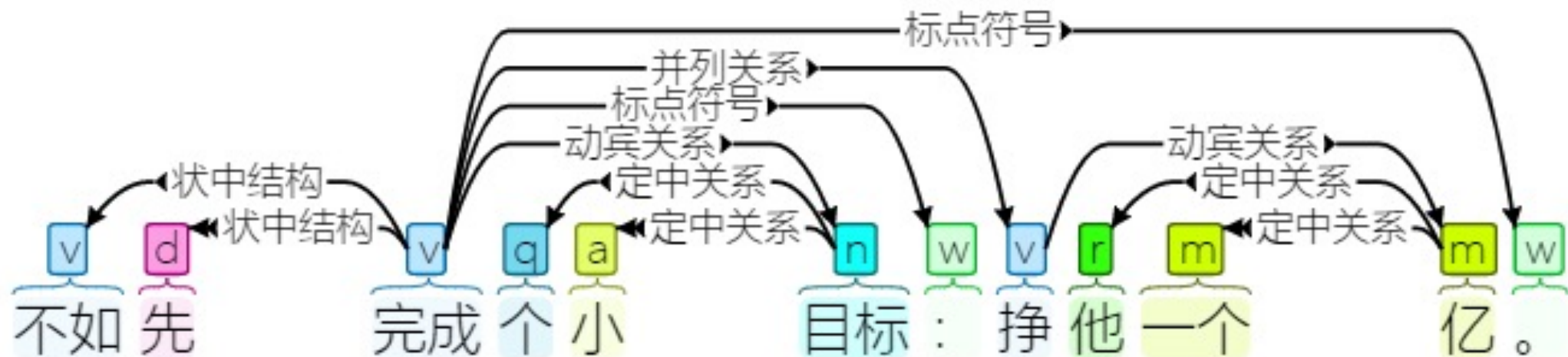


2. semantic labeling (语义标注)



"The police are conducting a thorough investigation of the cause of the accident."

3. Dependency Parsing (依存句法分析)



4. Coreference Parsing (共指消解)



我姐送我她的猫。
我很喜欢它。

NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation	Clause boundary detection	Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

1. Sentiment analysis (情感分析)



My experience
so far has been
fantastic!

POSITIVE



The product is
ok I guess

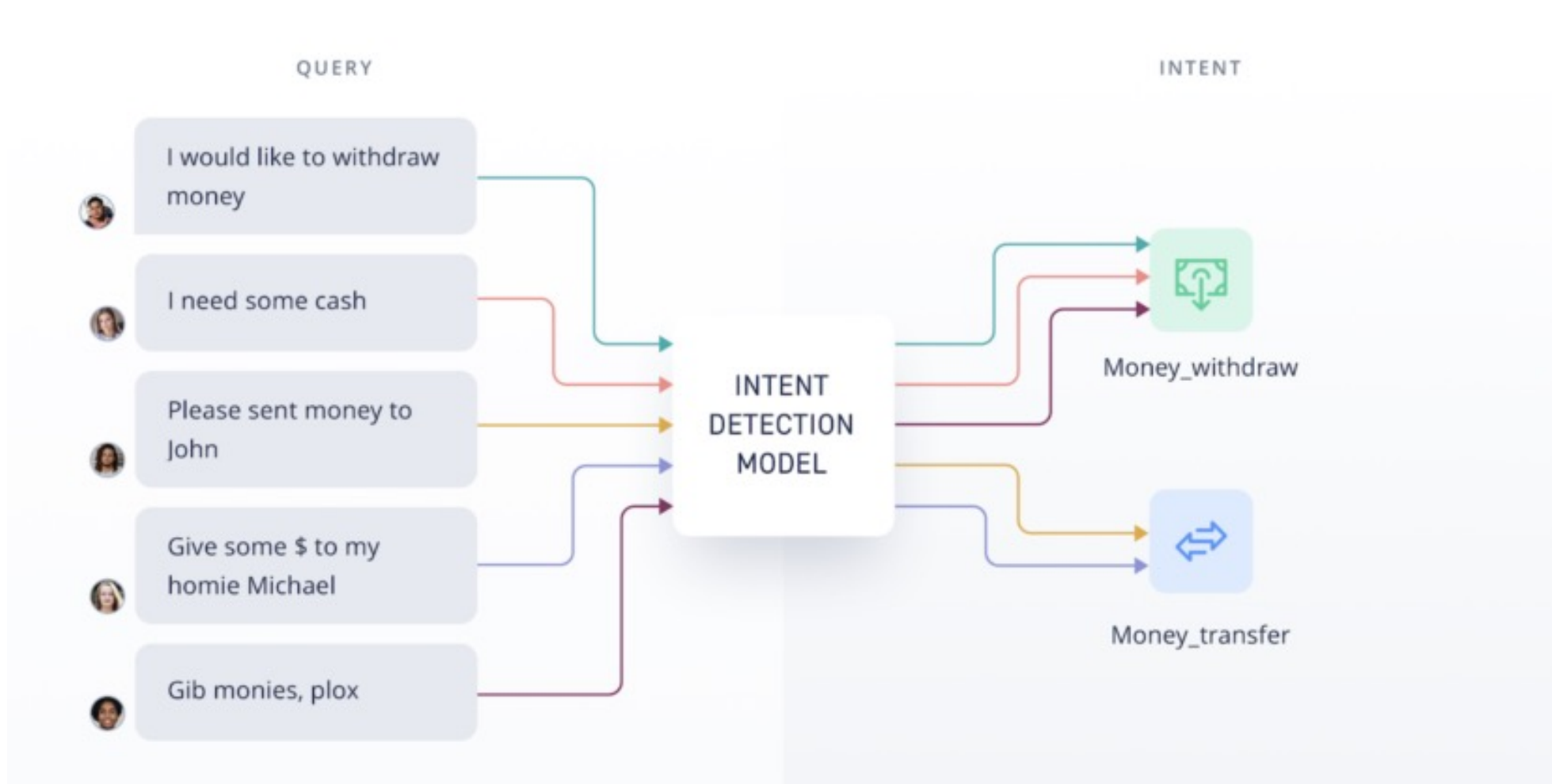
NEUTRAL



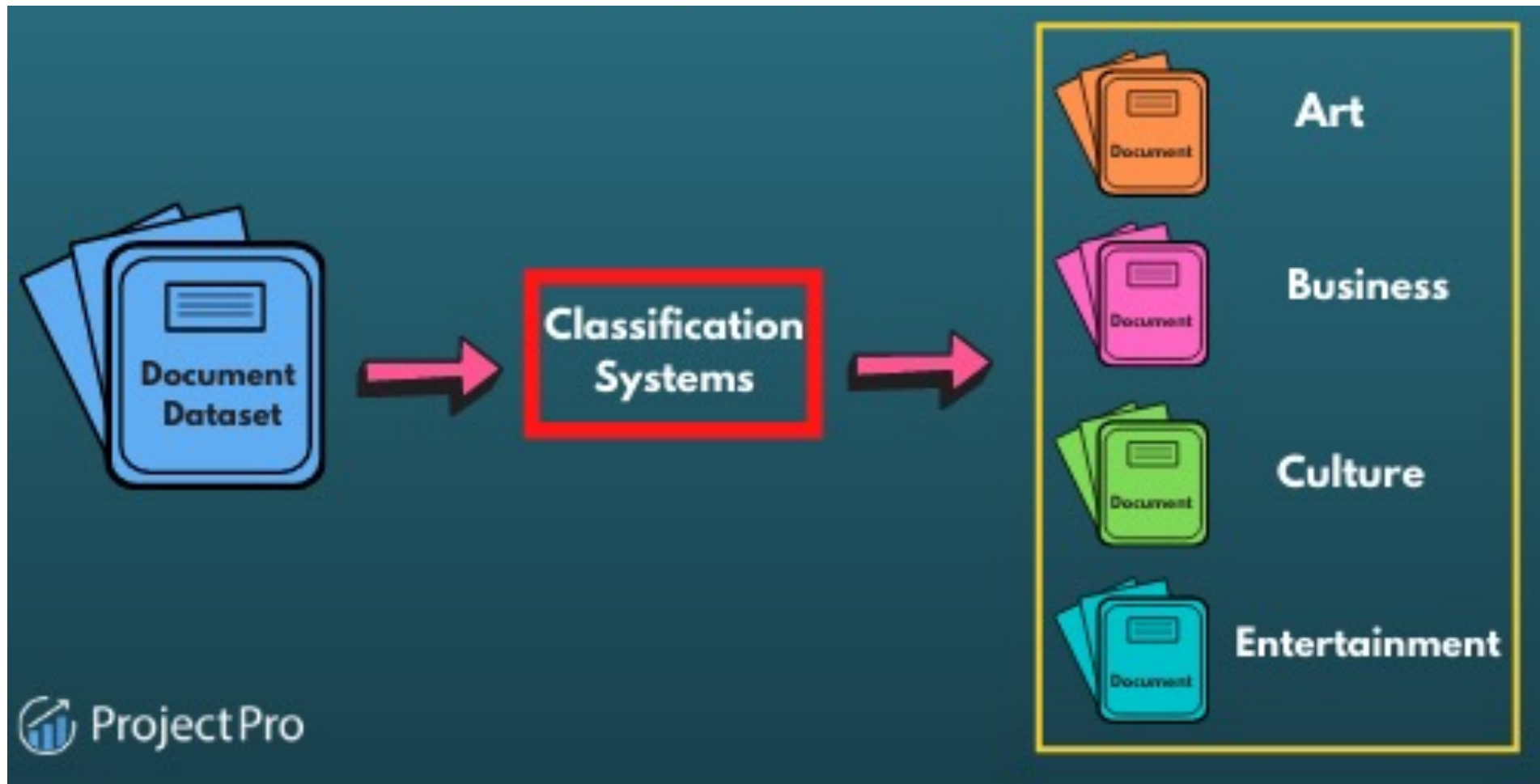
Your support team
is useless

NEGATIVE

2. Intent Detection (意图识别)



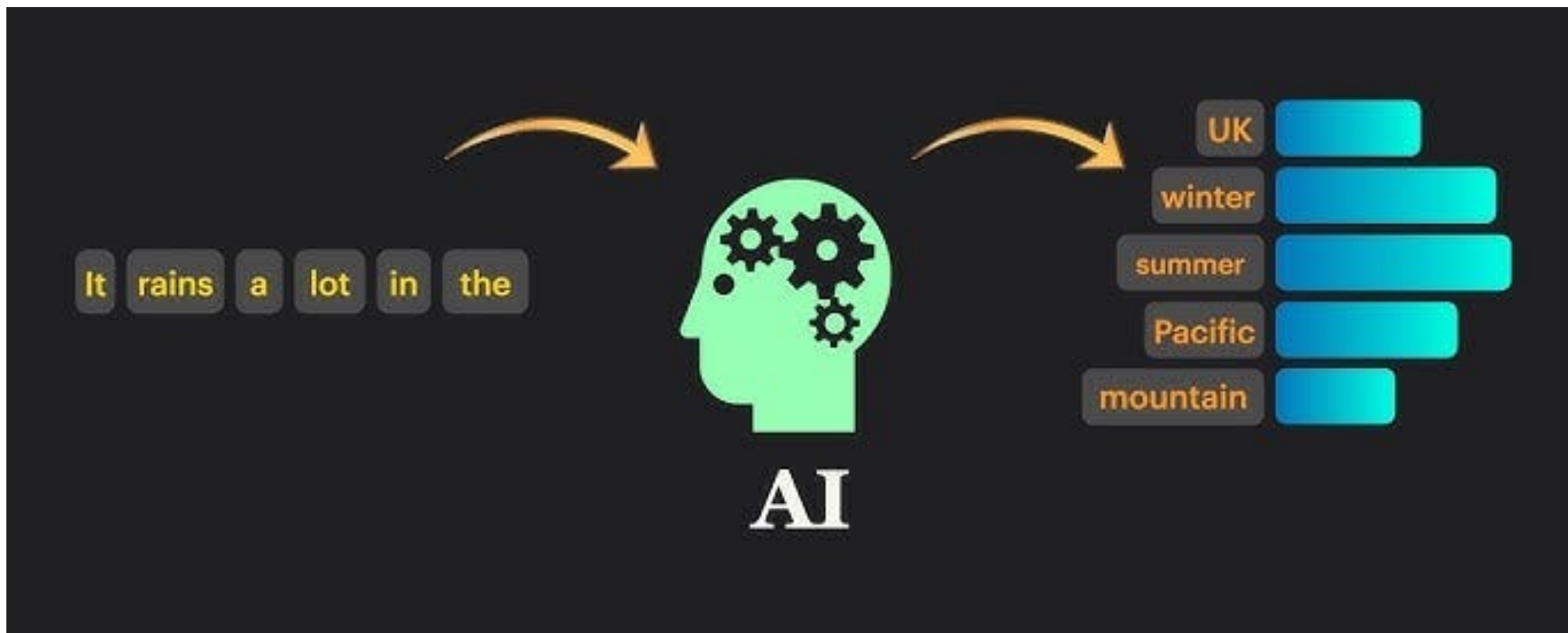
3. Topic Classification (主题分类)



NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation	Clause boundary detection	Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

1. Generative Text Modeling (文本生成)



2. Machine Translation (机器翻译)

机器翻译做什么？

Source:

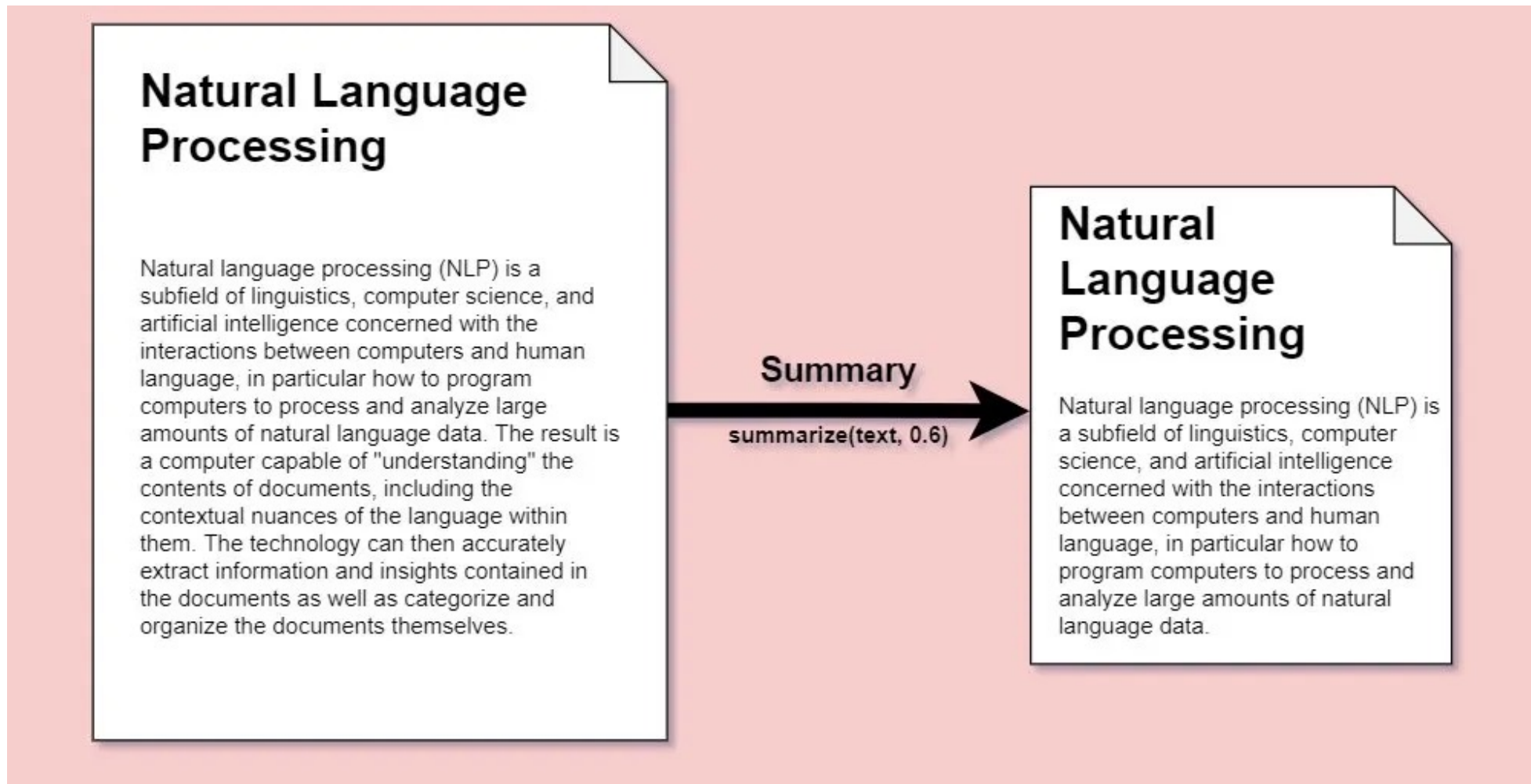
我 在 周日 看 了 一 本 书



Target:

I read a book on Sunday

3. Summarization (摘要)



4. Dialogue Systems (对话)



The image displays a user interface for a dialogue system, likely a chatbot. It is divided into two main sections: a chat window on the left and a knowledge base on the right.

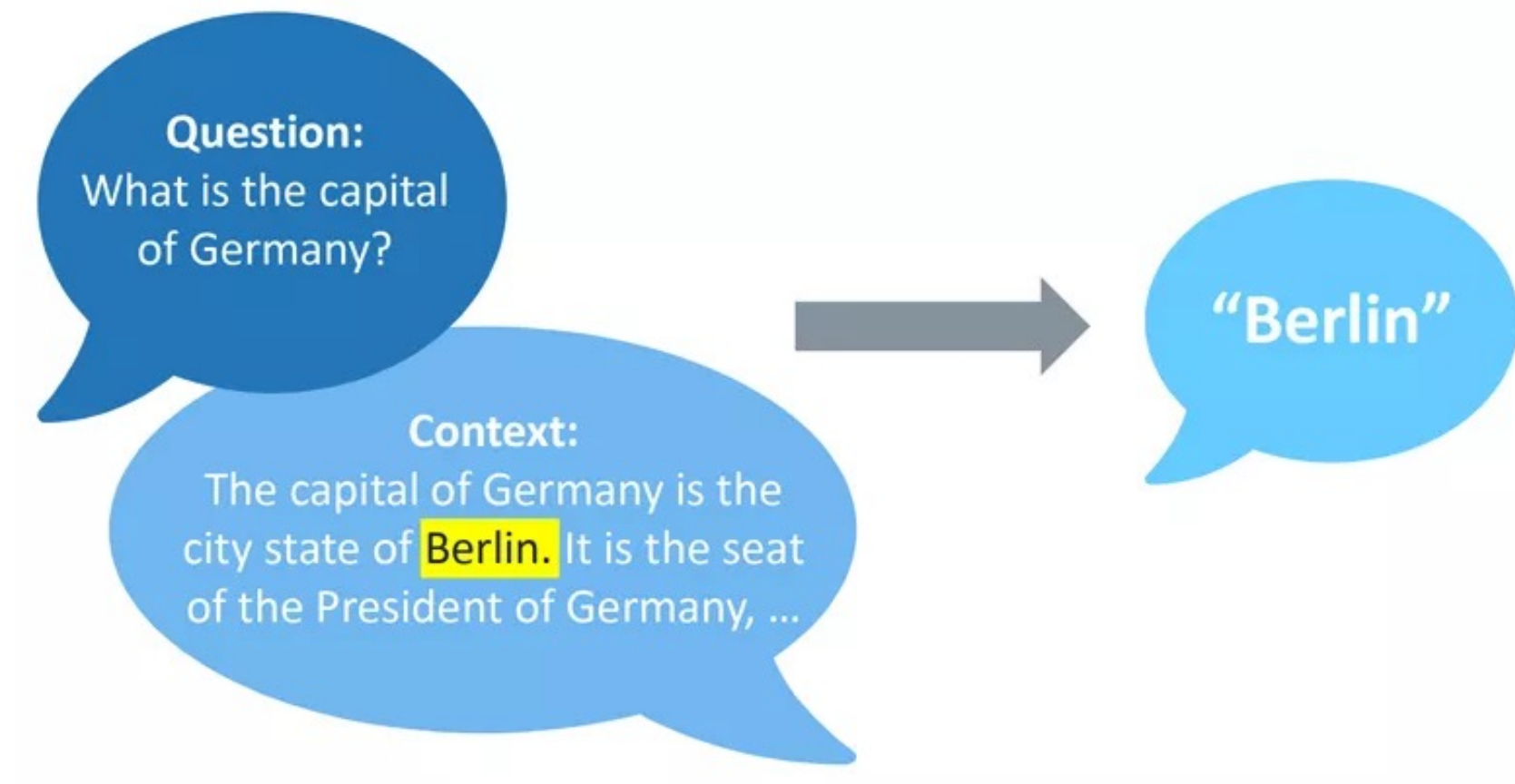
Chat Window:

- At the top, there are status indicators: "空闲" (Idle) with a dropdown arrow, "在线排队 99+" (Online Queue 99+), and "呼叫排队" (Call Queue) with a user icon.
- The chat history shows:
 - User: 您好! (Hello!)
 - Bot: 您好, 请问有什么能够帮您? (Hello, what can I help you with?)
 - Bot (highlighted): 识别客户真实意图"如何查看订单详情" (Identify the user's true intent: "How to view order details")
 - User: 我怎么看订单呀? (How do I view my order?)
 - Bot: 在手机App我的页面, 找到"我的订单"选项, 然后点击进入。 (In the mobile App, go to my page, find the "My Order" option, and then click to enter.)
 - User: 好的, 感谢哈 (Okay, thank you)
- At the bottom, there is a toolbar with icons for emojis, images, voice, contacts, documents, a magnifying glass, a close button, and a text input field.

Knowledge Base:

- Section: 知识搜索 (Knowledge Search)
- Filters: 全部 (All), 问答库 (Q&A Library), 文档库 (Document Library)
- Search Input: 如何查看订单详情? (How to view order details?)
- Section: 知识库 (Knowledge Base)
- Items:
 - 1、查看订单详情 (View Order Details) - expanded, showing the instruction: "在手机App我的页面, 找到'我的订单'选项, 然后点击进入。"
 - 2、分享订单详情 (Share Order Details) - collapsed
 - 3、删除订单 (Delete Order) - collapsed

5. Question Answering (问答)



NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation	Clause boundary detection	Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	

1. Semantic Textual Similarity (语义相似度)

Source Sentence

你好吗

你怎么样

0.653

我吃了一个苹果

0.041

你过的好吗

0.539

你还好吗

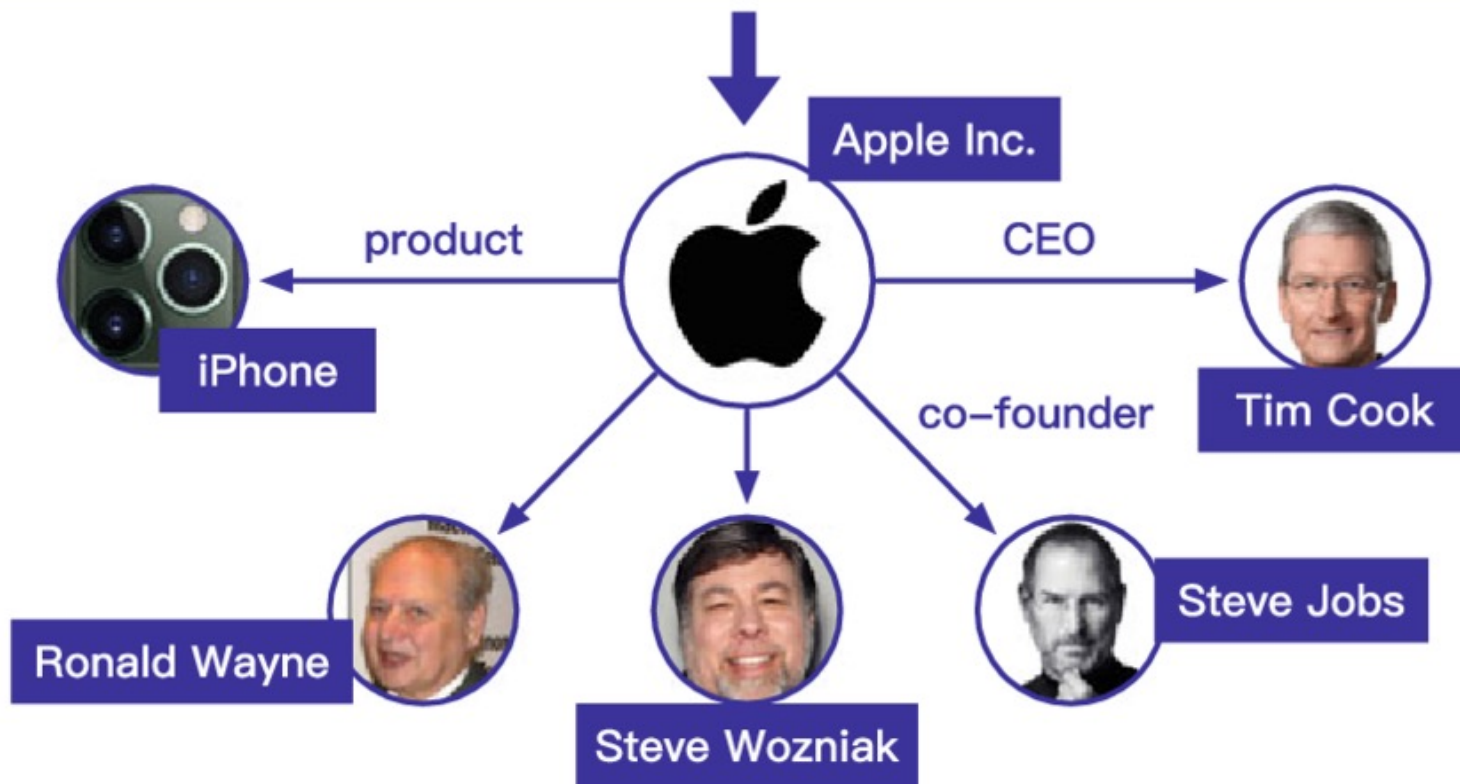
0.791

2. Natural Language Inference (文本蕴含)

Premise	Relation	Hypothesis
A turtle danced.	entails	A turtle moved.
turtle	contradicts	linguist
Every reptile danced.	neutral	A turtle ate.
Some turtles walk.	contradicts	No turtles move.
James Byron Dean refused to move without blue jeans.	entails	James Dean didn't dance without pants.
Mitsubishi Motors Corp's new vehicle sales in the US fell 46 percent in June.	contradicts	Mitsubishi's sales rose 46 percent.

3. Relation Extraction (关系抽取)

Apple Inc. is a technology company founded by Steve Jobs, Steve Wozniak and Ronald Wayne. Its current CEO is Tim Cook. Apple is well known for its product iPhone.



NLP任务概览

WORD TAGGING	SENTENCE PARSING	TEXT CLASSIFICATION	TEXT GENERATION	TEXT PAIR MATCHING
Tokenization	Constituency parsing	Sentiment analysis	Generative text modeling	Semantic textual similarity
Named entity recognition (NER)	Semantic labeling	Intent detection	Machine translation	Natural language inference (NLI)
Part-of-speech tagging	Dependency parsing	Topic classification	Summarization	Relation extraction
Lemmatization/ Stemming	Coreference parsing	Fake news detection	Personalized dialogue systems	
Word sense disambiguation		Email classification	Report generation	
Keyword extraction		Customer feedback analysis	Question answering (QA)	



中国科学院大学
University of Chinese Academy of Sciences

目 录

- 1 人人都爱NLP
- 2 Hello, NLP!
- 3 NLP任务概览
- 4 NLP学习资源

学术会议

主要国际会议

- ACL (Association of Computational Linguistics)
- EMNLP (Conference on Empirical Methods in Natural language Processing)
- NAACL (The North American Chapter of the Association for Computational Linguistics)
- Coling (International Conference on Computational Linguistics)
- EACL(European Chapter of ACL)
- IJCNLP(International Joint Conference on Natural language Processing)
- SIGIR(SIG Information Retrieval)
- TREC(Text REtrieval Conference)

.....

主要国内会议

- JSCL(全国计算语言学联合学术会议)

学习资源

学校/机构	课程名	网址
CMU	Natural Language Processing	http://demo.clab.cs.cmu.edu/NLP/
MIT	Natural Language Processing	http://web.mit.edu/6.863/www/fall2012/
Stanford University	Natural Language Processing	https://web.stanford.edu/class/cs224n/
Columbia University	Natural Language Processing	http://www.cs.columbia.edu/~cs4705/
Princeton	Understanding Large Language Models	https://www.cs.princeton.edu/courses/archive/fall22/cos597G/
Amazon	Dive into Deep Learning	https://d2l.ai/
Westlake University	Natural Language Processing – A Machine Learning Perspective	https://westlakenlp.github.io/nlpml/

本节复习

□ NLP的价值

□ NLP在AI中的位置

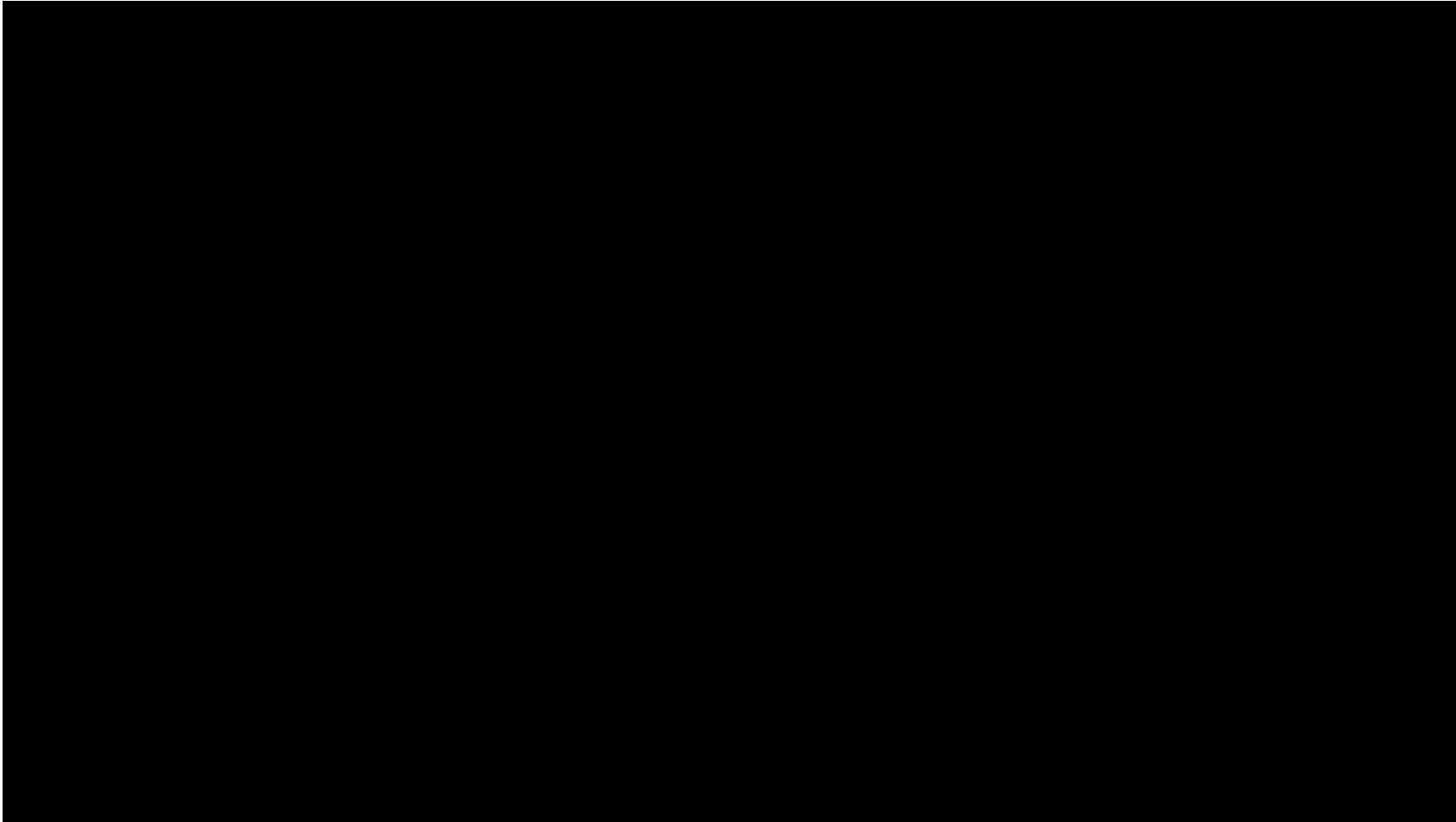
□ NLP的成长经历和各阶段特点

□ NLP的任务

□ Tagging、Parsing、Classification、Generation、Matching

参考文献

- Turing, Alan M. "Computing machinery and intelligence." *Creative Computing* 6.1 (1980): 44-53.



致 谢

- 胡玥、曹亚男、方芳：国科大《自然语言处理基础》
- 曹亚男、任昱冰：国科大《深度学习与自然语言处理概述》





THANKS