自然语言处理应用实践

——暑期课程

南京大学软件学院
李传艺
费彝民楼917

自然语言处理是什么?

- 自然语言——人类语言
  - 符号系统、社会性、长期演化

- 自然语言处理（Natural Language Processing，NLP）指对人类语言进行自动的计算处理。

——《基于深度学习的自然语言处理》

- 自然语言处理就是以电子计算机为工具，对人类特有的书面形式和口头形式的自然语言的信息进行各种类型处理和加工的技术。

——《自然语言处理简明教程》

- 自然语言处理是研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法。

——百度百科

- Natural Language Processing is concerned with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data.

——Wikipedia

非自然语言 { 动物语言 / 人工语言

自然语言处理为什么重要?

- 自然语言的重要性
  - 是人类区别于其他动物的重要标志之一
  - 人类通过自然语言交流思想，相互了解，组成社会
  - 人类借助自然语言进行思维活动，认识事物的本质和规律，创造了人类的物质文明和精神文明

- 为了让计算机获得人类智能，首先要让它能够像人类一样理解和使用自然语言。

- 自然语言处理是人工智能的一个主要内容，是计算机模拟人类智能的一个重要方面，是研制智能化的电子计算的一项基础性工作。

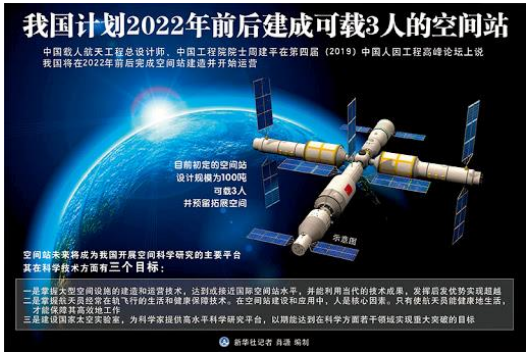
- “语言理解是**人工智能**领域 皇 冠 上的明珠” ——比尔·盖茨

你觉得人工智能是什么？

人工智能——定位

- 综合性很强的交叉学科
  - 计算机科学
  - 控制论
  - 信息论
  - 神经心理学
  - 哲学
  - 语言学

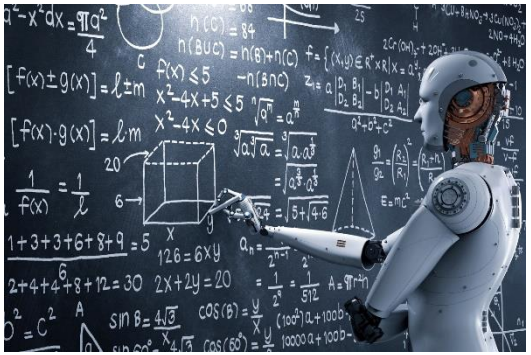
新思想、新观念
新理论、新技术
不断出现的新兴学科



20世纪三大科学技术成就

空间技术
原子能技术
人工智能技术

- 前三次工业革命
  - 扩展人手的功能，将人类从繁重的体力劳动中解放出来
- 而人工智能技术
  - 扩展人脑的功能，实现脑力劳动的自动化

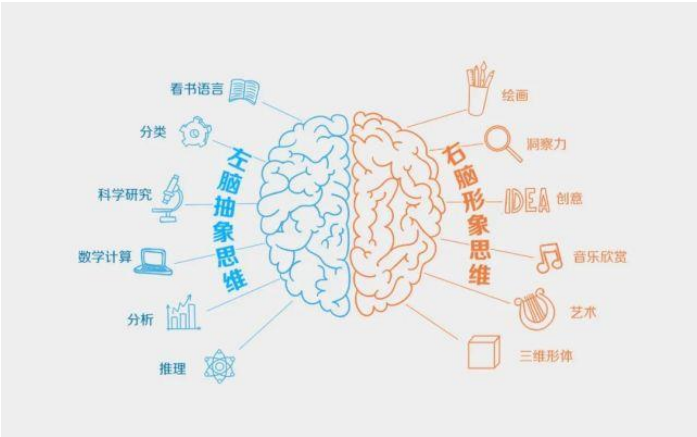


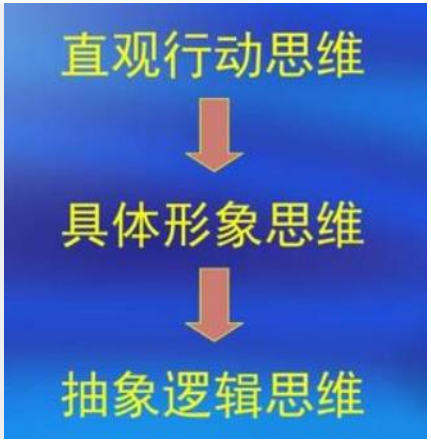
人工智能——概念（1）

- 智能的概念
  - 知识与智力的总和。知识是一切智能行为的基础，智力是获取知识并应用知识求解问题的能力。

- 智能的特征——人类
  - 感知能力：视觉、听觉、触觉、嗅觉、味觉
  - 记忆与思维能力——记忆与思维是不可分的
    - 记忆用于存储感知到的外部信息，和由思维产生的知识
    - 思维用于对记忆的信息进行处理，包括分析、计算、比较、判断、推理、联想及决策等
      - 逻辑思维（抽象思维）
      - 形象思维（直感思维）
      - 顿悟思维
  - 学习能力
  - 行为能力

知觉

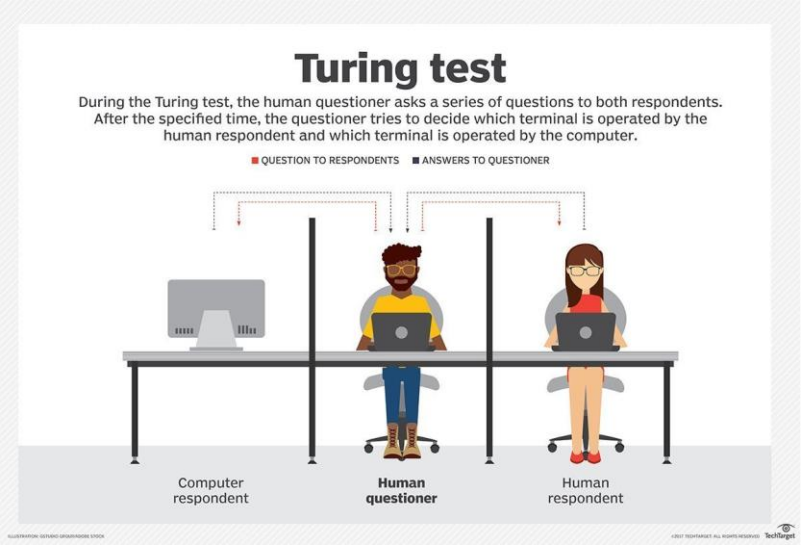




- 人工智能
  - 用机器实现人类的部分智能
  - 用人工的方法在机器（计算机）上实现的智能，也称为机器智能。

人工智能——概念（2）

- 图灵测试——1950年，英国数学家 图灵 A. M. Turing
  - 论文《计算机与智能（Computing Machinery and Intelligence）》
  - 形象地指出什么是人工智能及机器应该达到地智能标准
  - 一直被质疑：反映结果，没有涉及思维过程

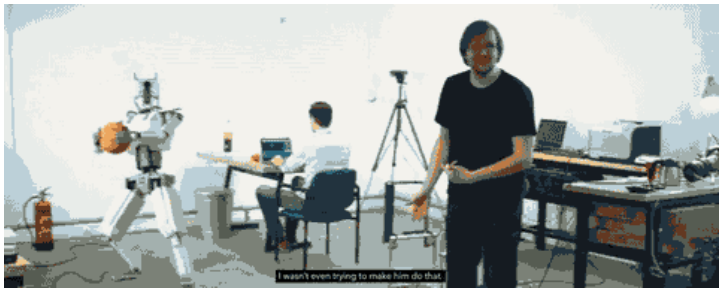


- 中文屋思想实验——1980年，哲学家 约翰·希尔勒 John Searle
  - 一个按照规则执行的计算机程序并未真正理解其输入和输出的意义

Reverse Turing Test

人工智能——研究内容

- 现状——“人工智障”案例



- 基本内容
  - 知识表示
    - 符号表示法（如一阶谓词逻辑知识表示）
    - 连接机制表示法（如语义网络、知识图谱）
  - 机器感知
    - 计算机视觉和计算机听觉
  - 机器思维
    - 对感知到的外部信息和内部工作信息进行有目的的处理——最重要、最关键
  - 机器学习：如何使计算机具有人的学习能力，自动地获取知识——不同于狭义的“机器学习”算法
  - 机器行为：表达能力，说、写、画、行等能力

人工智能——研究领域

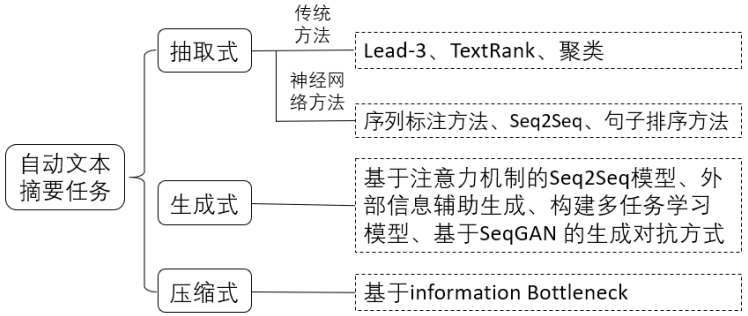
- 主要的研究领域
  - 自动定理证明
  - 博弈
  - 模式识别
  - 机器视觉
  - 自然语言理解
  - 智能信息检索
  - 数据挖掘与知识发现
  - 专家系统
  - 自动程序设计
  - 机器人
  - 组合优化问题
  - 人工神经网络
  - 分布式人工智能与多智能体
  - 智能控制
  - 智能仿真
  - 智能CAD
- 智能ICAI
- 智能管理与智能决策
- 智能多媒体系统
- 智能操作系统
- 智能计算机系统
- 智能通信
- 智能网络系统
- 人工生命

QA SYSTEMS
The Software Quality Company

理解和回答问题

研究能够实现人与计算机之间用自然语言进行通信的理论与方法

生成摘要

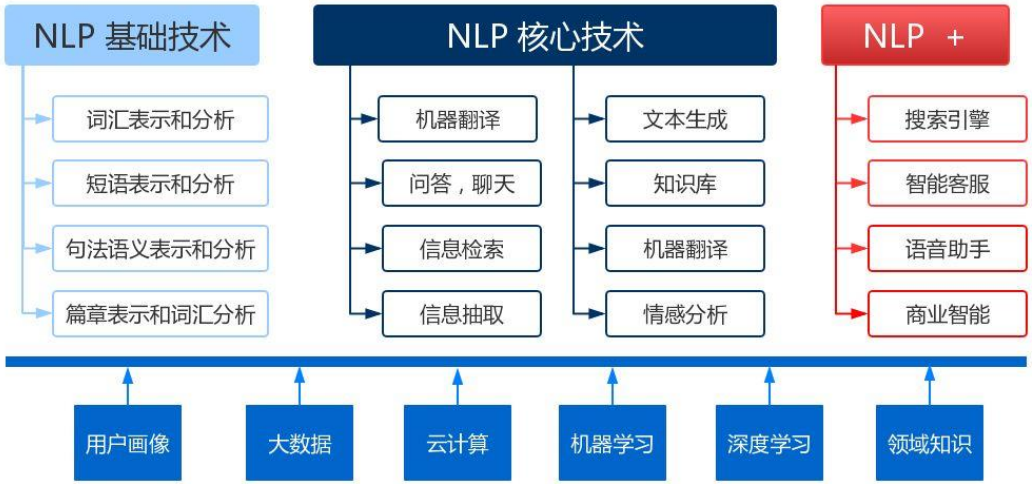


自动翻译

Google Translate

自然语言处理能干什么?

- 相关应用产品
  - 闲聊机器人
  - 智能音箱：Amazon Alexa、天猫精灵
  - 智能客服：京东客服、支付宝客服
  - 垃圾邮件检测：邮箱自带服务
  - 文本纠错：Office Word、Grammarly
  - 知识问答机器人
  - 语音点餐机器人
  - 推荐系统
  - 搜索引擎
  - ……

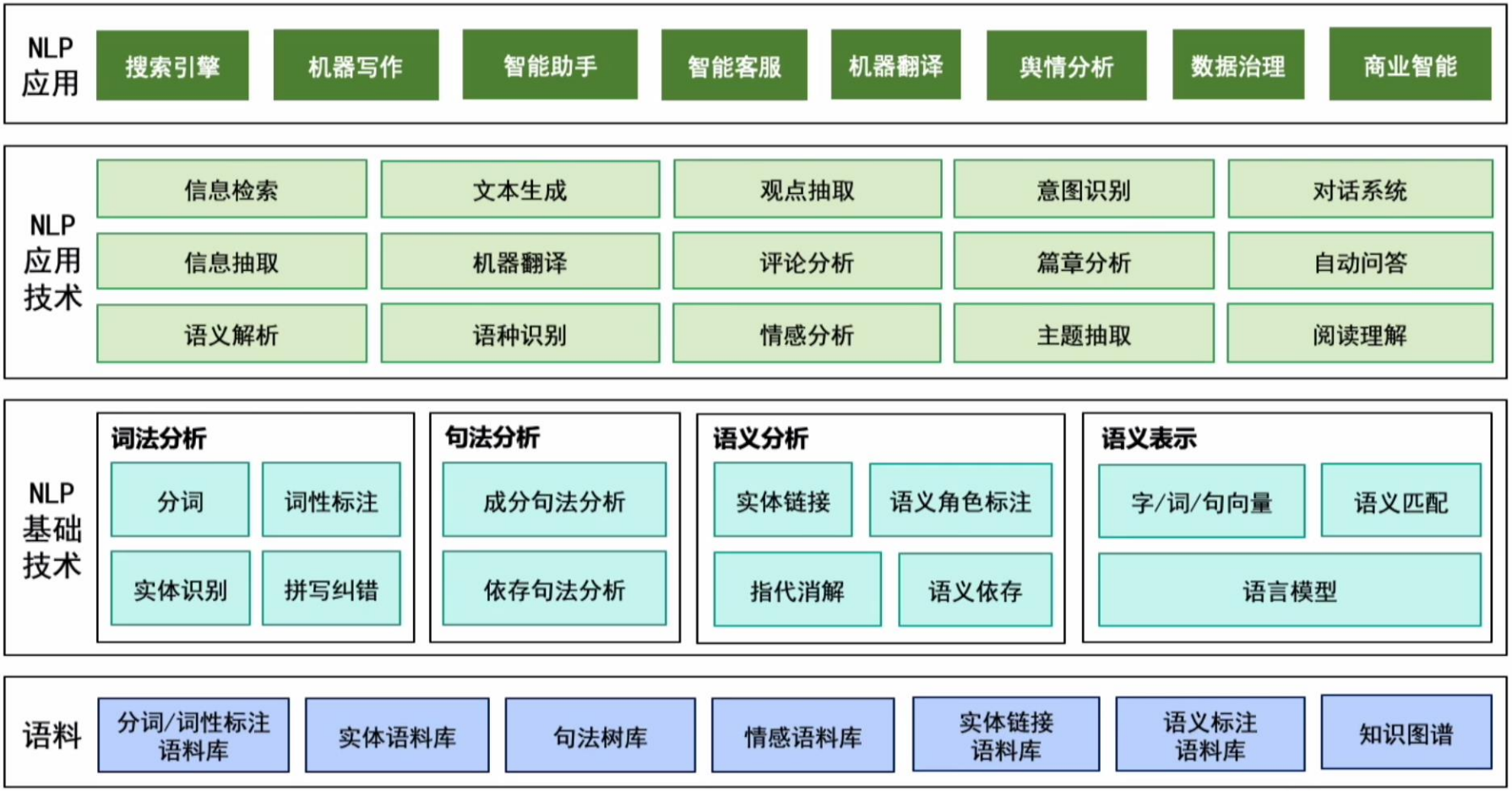


- 自然语言处理拟解决的两大核心任务
  - 自然语言理解：Natural Language Understanding
  - 自然语言生成：Natural Language Generation

将人类自然语言文本作为输入

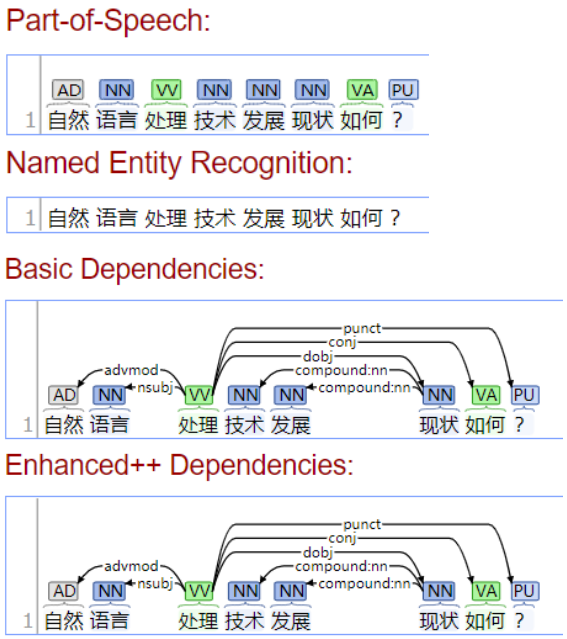
产生自然语言的文本作为输出

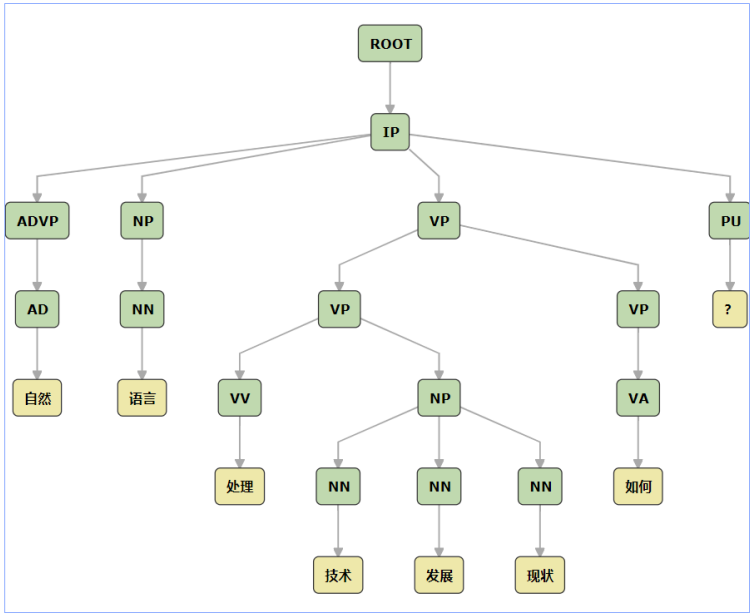
NLP技术和应用框架


NLP
应用
搜索引擎
机器写作
智能助手
智能客服
机器翻译
舆情分析
数据治理
商业智能
NLP
应用
技术
信息检索
文本生成
观点抽取
意图识别
对话系统
信息抽取
机器翻译
评论分析
篇章分析
自动问答
语义解析
语种识别
情感分析
主题抽取
阅读理解
NLP
基础
技术
词法分析
分词
词性标注
实体识别
拼写纠错
句法分析
成分句法分析
依存句法分析
语义分析
实体链接
语义角色标注
指代消解
语义依存
语义表示
字/词/句向量
语义匹配
语言模型
语料
分词/词性标注
语料库
实体语料库
句法树库
情感语料库
实体链接
语料库
语义标注
语料库
知识图谱

NLP基础技术

- 词法分析
  - 分词、词性标注
  - 实体识别
  - 文本纠错
- 句法分析
  - 成分句法分析：
  - 依存句法分析：
- 语义分析
  - 实体链接、语义角色标注、指代消解、语义依存
- 语义表示
  - 字词句向量、语义模型、语义匹配





一阶谓词演算　概念依存图

语义网络　基于框架的表示法

NLP问题解决程度现状

- 基本解决
  - 词法分析（分词、词性标注、命名实体识别等）
  - 简单分类应用：垃圾邮件检测
- 有令人振奋的进展
  - 情感分析：判断商品评论是称赞还是诟病、判断对商品不同方面的细粒度评价
  - 共指消解：代词指代的具体实体
  - 歧义消解
  - 句法分析（成分句法分析、依存句法分析）
  - 信息抽取：文本结构化、知识图谱构建
  - 机器翻译
- 仍然很难
  - 问答系统（question answering）
  - 文本解释（Paraphrase）
  - 摘要
  - 对话

NLU

NLG

自然语言处理的困难之处?

自然语言处理的困难之处?

- 歧义——人类理解困难
  - 乒乓球拍卖完了 VS. 乒乓球拍买完了——分词歧义
  - [咬死猎人]的狗 VS. 咬死[猎人的狗]——短语歧义
  - 你真讨厌！——语用歧义
  - 词汇歧义——一词多义
- 病构
  - 不规范的语言表达
    - 他非常Man。（中英文混用，名词不应该被程度副词修饰）
  - 新造词/旧词新用法
    - 吃鸡，给力，奥利给
    - A:“看电影去不去？”，B:“我去，不去。”
- 外部知识——机器理解困难
  - 冬天能穿多少穿多少；夏天能穿多少穿多少

从语言的结构和语义角度理解：

1. 结构复杂多样：名词短语、动词短语等
2. 语义千变万化：歧义、多义、同义
3. 结构-语义不存在一一对应关系，单独处理

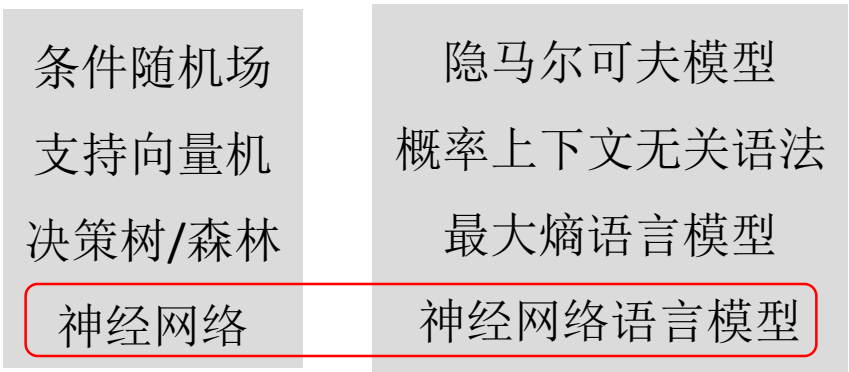
从自然语言特性角度理解：

1. 离散性
2. 组合性
3. 数据稀疏性

自然语言处理核心技术的变迁

- 基于句法—语义规则（符号操作）的理性主义
  - 哲学基础是逻辑实证主义：智能的基本单位是符号，认知过程就是在符号的表征下进行符号运算，因此，思维就是符号运算

- 基于大规模语料库分析的经验主义
  - 引入概率理论+数据驱动
  - 机器学习方法得到越来越多的应用
    - 建立带标记的语料库
    - 统计机器学习算法的应用，语言统计模型得到重视



- 词汇主义
  - 语法知识库、语义知识库
  - 语义网络
  - 知识图谱

国内高校NLP研究团队不完全统计

学校	代表人物	实验室	研究方向
清华大学	孙茂松、朱小燕、马少平、李涓子、刘知远、刘洋、黄民烈等	清华大学自然语言处理与社会人文计算实验室	研究领域的涵盖面正逐步从计算语言学的核心问题扩展到社会计算和人文计算。
		清华大学智能技术与系统国家重点实验室	
北京大学	王厚峰、李素建、穗志方、万小军、孙栩、严睿等	北京大学语言计算与互联网挖掘研究组	语义理解(语义分析系统)、机器写作(自动文摘、自然语言生成)、情感计算(高精度情感、立场与幽默)、人机对话技术等。
		北京大学计算语言学教育部重点实验室	中文计算的基础理论与模型；大规模多层次语言知识库构建的方法；国家语言资源整理与语音数据库建设；海量文本内容分析与动态监控；多语言信息处理和机器翻译。
哈工大	赵铁军、刘挺、车万翔、秦兵、刘秉权、孙承杰、徐睿峰、王晓龙等	哈工大社会计算与信息检索研究中心	语言分析,信息抽取,情感分析,问答系统,社会媒体处理和用户画像
中国科学院	刘群、宗成庆、赵军、孙乐、张家俊、刘康、王斌、韩先培等	中科院计算所自然语言处理研究组	机器翻译、人机对话、多语言词法分析、句法分析和网络信息挖掘
		中科院模式识别国家重点实验室	自然语言处理基础、机器翻译、信息抽取和问答系统等研究工作
苏州大学	张民、周国栋、陈文亮、李正华、熊德意、李军辉、洪宇等	苏大计算机科学与技术学院自然语言处理课题组	自然语言理解、中文信息处理、机器翻译和自然语言认知
复旦大学	黄萱菁、邱锡鹏、巍忠钰等人	复旦自然语言处理研究组	主要研究方向包括：自然语言处理、非规范化文本分析、语义计算、信息抽取、倾向性分析、文本挖掘等方面。
东北大学	朱靖波、肖桐、任飞亮等人	东北大学自然语言处理实验室	语言分析和机器翻译

南京大学NLP研究组

- 陈家骏，戴新宇，黄书剑

- 成果
  - 先后承担过该领域的18项国家科技攻关项目、863项目、国家自然科学基金和江苏省自然科学基金以及多项对外合作项目的研制
  - 在自然语言处理顶级国际会议ACL、EMNLP、NAACL和人工智能顶级国际会议IJCAI和AAAI上发表论文三十余篇，相关系统在机器翻译、中文分词、命名实体识别、情感计算等多个国际国内评测中名列前茅。

- 研究内容
  - 中文分词、词性标注、命名实体识别、句法分析、指代消解等自然语言处理基本任务的研究工作
  - 基于端到端的机器翻译、句法分析、自动问答等方面的研究
  - 基于规则的机器翻译研究
  - 基于统计的机器翻译研究
  - 智能问答系统：面向问答的知识库构建、基于知识库检索和深度神经网络的问答系统构建
  - 更广更深的推荐系统
  - 多层次多粒度的情感分析
  - 其他：古汉语、中文认知库、网页过滤系统、论文标题自动缩写等

国内知名企业NLP研究团队与产品

- 百度NLP：AI Lab
  - 学术：https://nlp.baidu.com/homepage/index
  - 产品：https://ai.baidu.com/tech/nlp_basic

https://ai.baidu.com/tech/nlp_basic/lexical

- 华为NLP：诺亚方舟实验室
  - 学术：http://www.noahlab.com.hk/#/home
  - 产品：https://support.huaweicloud.com/nlp/index.html

面向非终端用户的产品
基于算法、算力的服务

- 腾讯NLP：AI Lab
  - 学术：https://ai.tencent.com/ailab/nlp/en/index.html
  - 产品：https://cloud.tencent.com/product/nlp

- 阿里NLP：达摩院-语言技术实验室
  - 学术：https://damo.alibaba.com/labs/language-technology
  - 产品：https://helpcdn.aliyun.com/product/60058.html

- 微软亚研NLC
  - https://www.microsoft.com/en-us/research/group/natural-language-computing/

本课程内容

- 自然语言处理基础理论与技术
  - 自然语言特性
  - 词汇与结构
  - 正则表达式、最小编辑距离、最大公共子串（文本相似度）
  - 中文分词
  - 文本特征表示
- 机器学习算法简介
  - 支持向量机SVM
  - 隐马尔可夫模型、条件随机场CRF
  - 神经网络、神经网络语言模型
- 自然语言处理应用研究Pipeline
  - 应用技术类型
  - 共享的流水线
- 自然语言处理应用实践——情感分析
  - 分类问题：夸张句检测（单文本二分类）——NB、SVM算法
  - 序列标注：夸张成分识别——Bi-LSTM+CRF算法
  - 文本生成：对给定的夸张句生成元组——基于Bert的Seq2Seq实现

需要大量课外时间的阅读和学习

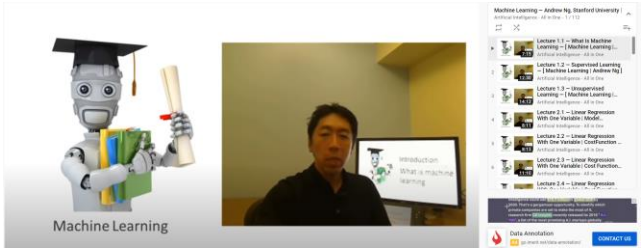
《统计学习方法》，李航

《统计自然语言处理》，宗成庆

《自然语言处理简明教程》，冯志伟

《机器学习》，周志华

《人工智能导论》，王万良



要求

- 自带笔记本电脑，要能够访问互联网
- 完成课外阅读
- 完成实验并提交报告
- 不可以迟到早退

20级暑期课程-NLP应用实践

该二维码7天内(7月19日前)有效，重新进入将更新

参考与阅读

- 《人工智能导论》，王万良 编著，高等教育出版社，P1-21
  - 智能的特征、人工智能的发展简史、主要研究领域介绍
- Turing, Alan (1950). “Computing Machinery and Intelligence”
  - https://www.csee.umbc.edu/courses/471/papers/turing.pdf
- Saygin, A. P. (2000). "Turing Test: 50 years later “
  - https://link.springer.com/content/pdf/10.1023/A:1011288000451.pdf

> The British mathematician Alan Turing[1] proposed the Turing Test (TT) as a replacement for the question "Can machines think?" in his 1950 *Mind* article ‘Computing Machinery and Intelligence’ (Turing, 1950). Since then, Turing’s ideas have been widely discussed, attacked, and defended over and over. At one extreme, Turing’s paper has been considered to represent the "beginning" of artificial intelligence (AI) and the TT has been considered its ultimate goal. At the other extreme, the TT has been called useless, even harmful. In between are arguments on consciousness, behaviorism, the ‘other minds’ problem, operational definitions of intelligence, necessary and sufficient conditions for intelligence-granting, and so on.

- 《自然语言处理简明教程》，冯志伟，上海外语教育出版社，P37-53
  - 自然语言的7个特性

谢谢！