

深度学习与中文自然语言处理

Deep Learning for Natural Language Processing

April-07-2018

Lecture 1

NLP and ML introduction, tf-idf ,
keywords

课程结构（基础篇）

- 1. 自然语言处理的基础知识
- 2. 人工智能及机器学习基础
- 3. Word Embedding (Word2Vec) 及其扩展
- 4. 神经网络原理及其结构
- 5. RNN原理、变体及其应用
- 6. CNN及其应用

课程结构 (进阶篇)

- 7. TreeRNN与知识图谱
- 8. 自动对话机器人 (Chatbots)
- 9. 机器阅读理解
- 10. 强化学习与自然语言处理
- 11. 文本自动摘要
- 12. Sequence2Sequence 模型
- 13. NLP领域的前沿问题

课程历史

- 毕业于浙江大学。长期从事机器学习与自然语言处理的问题。本科期间负责的机器人项目即获得中国教育部奖励，之后参与过国家保密局的舆情项目，知识图谱的科研项目。研究生期间独立创业，之后返校继续从事自然语言与机器学习的研究。2017年负责阿里天猫精灵以及Rokid的新闻自动摘要项目，2017 - 2018年负责支付宝犯罪文本挖掘的项目。目前为IBM中国总部NLP方向的数据科学家。
- 全球人工智能平台
- 蚂蚁金服与阿里云期间的问题
- 女朋友找工作
- 最后一次培训



课程目标

- 1. 具备建模能力，能够把实际问题抽象为机器学习（深度学习）问题
- 2. 理解自然语言常用方法，遇到问题能够将问题结构为自然语言处理的基本问题
 - 词汇 (word2vec, 新词发现, 词汇扩展, NER实体发现)
 - 句子 (句子 Embedding 相似度判断)
 - 文本 (文本聚类)
 - 知识库 (聊天机器人, 知识图谱)
- 3. 能应对BAT级别深度学习（自然语言理解）方向的面试。

课程要求

先修课程

- 1. 基础的编程能力
 - 函数, 循环, 列表, 模块
- 2. 微积分基础知识
 - 积分, 倒数, 偏导
- 3. 基础环境的使用
 - jupyter notebook, python 解释器, pip/conda, git

学习要求

- 1. 大量的编程练习
- 2. 积极得查阅资料
- 3. 手推公式 (*)
- 4. 基础模块的重新实现 (*)

课程特点

- 1. 真实案例驱动
- 2. 在线编程演示
- 3. 论文资料(Reference)

关于另外两门录播课程

- 1. *Design of Programming*;
- 2. *Algorithm Analysis and Design*;
- ==> They are so smart, but how could be?
 - Design of Programming, 7 units, 2 week each unit;
 - Algorithm Analysis and Design, 6 units, 3 week each unit;

Assignment-01

- Assignments
 - 1. 注册Kaggle账号
 - 2. 注册Github账号 (Github账号发送给我)
 - 3. 下载安装anaconda, jupyter notebook
 - 5. NIPS, ICML, ICLR, ACL, AAAI任选一个会议找到其accepted papers

Part-2 人工智能简史

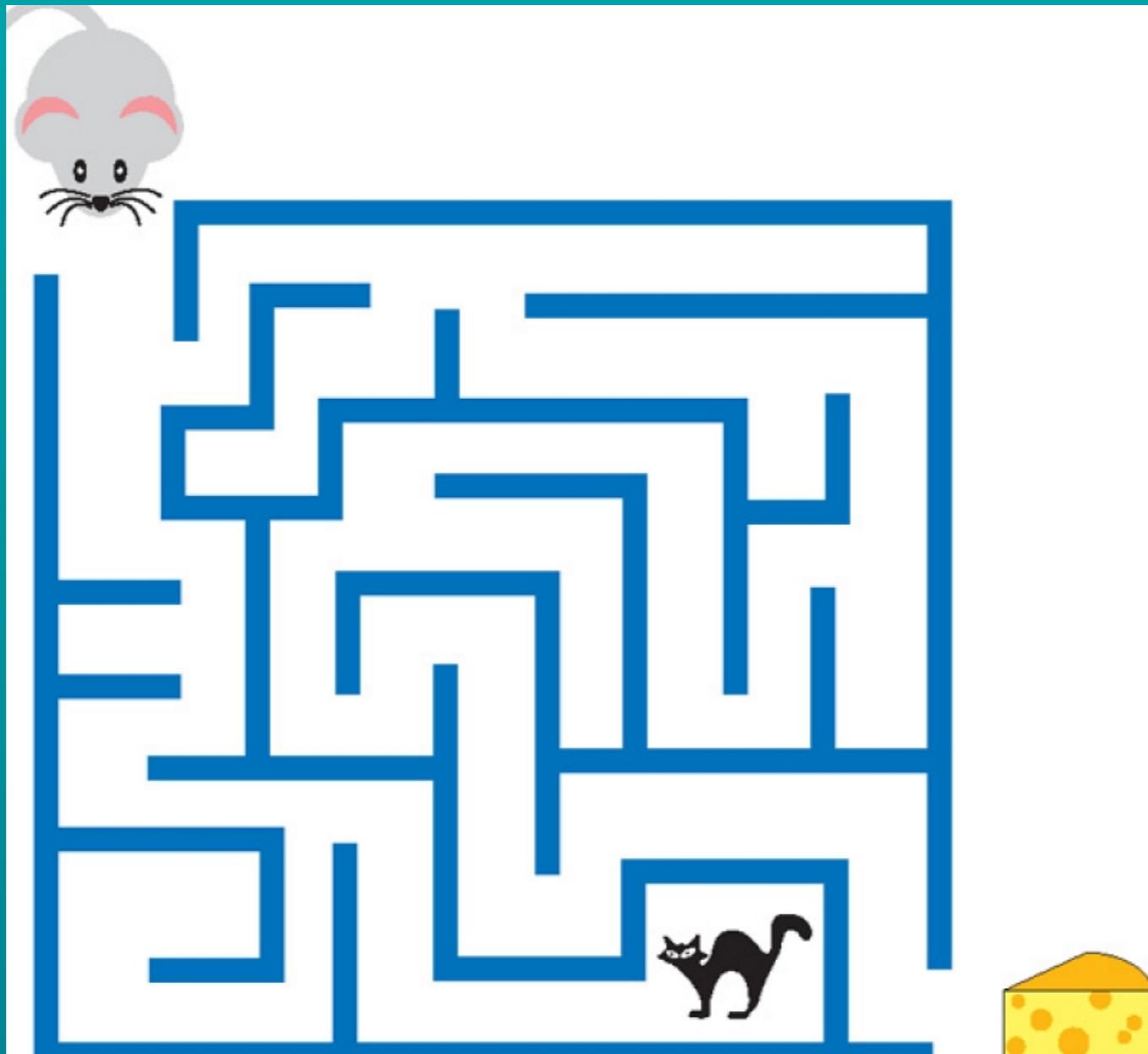
- 1. 对机器学习， 人工智能的历史和演化具有基本认识
- 2. 对深度学习与自然语言处理具有基本认识

What'AI

	Think	Action
Human	Chatbot	机械机器人, Boston Dynamic
Rational	AlphaZero, Assistant System	Drones

Remarkable Different

- 1. AI Agent and the usually Programming.
 - [] A. AI Agent Programming is more difficult;
 - [] B. AI Agent Programming focus on general problem solving;
 - [] C. AI Agent Programming need solve problems which met in future.
 - [] D. AI Agent Programming Need mathematical foundation;



迷宫问题

猫、老鼠和奶酪

例子-1

- 走迷宫
- 策略1，沿着能走通的路一直走，知道走不通，返回一步，继续走；
- 策略2，将这个口子连着的所有路口进行一下探测，然后再将所有的路口标为可以坐；然后在标过的路口，选择一个，重复上述过程。
- 下棋？

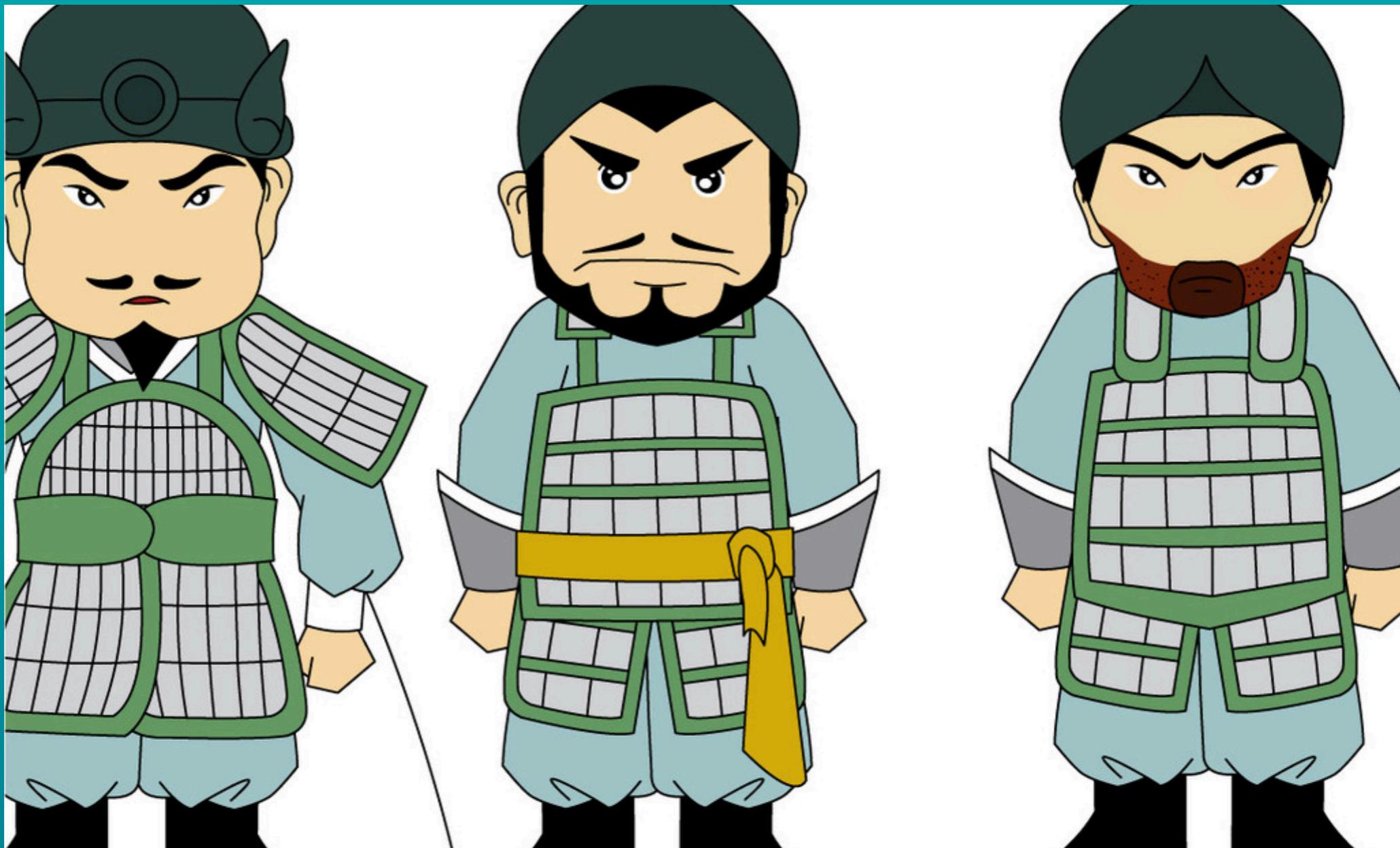


Mouse & Maze

人工规则?

例子-2

- 食堂打饭 不标价钱， 问， 每种菜多少钱？
 - 第一次 茄子*2 + 西瓜*1 + 馒头 * 1 : 7.8元
 - 第二次 茄子*1 + 西瓜*1 + 馒头 * 2 : 6.5元
 - 第三次 玉米*1 + 菠菜*2 + 馒头 * 1: 5.6元
 - ...
- 每种价格？

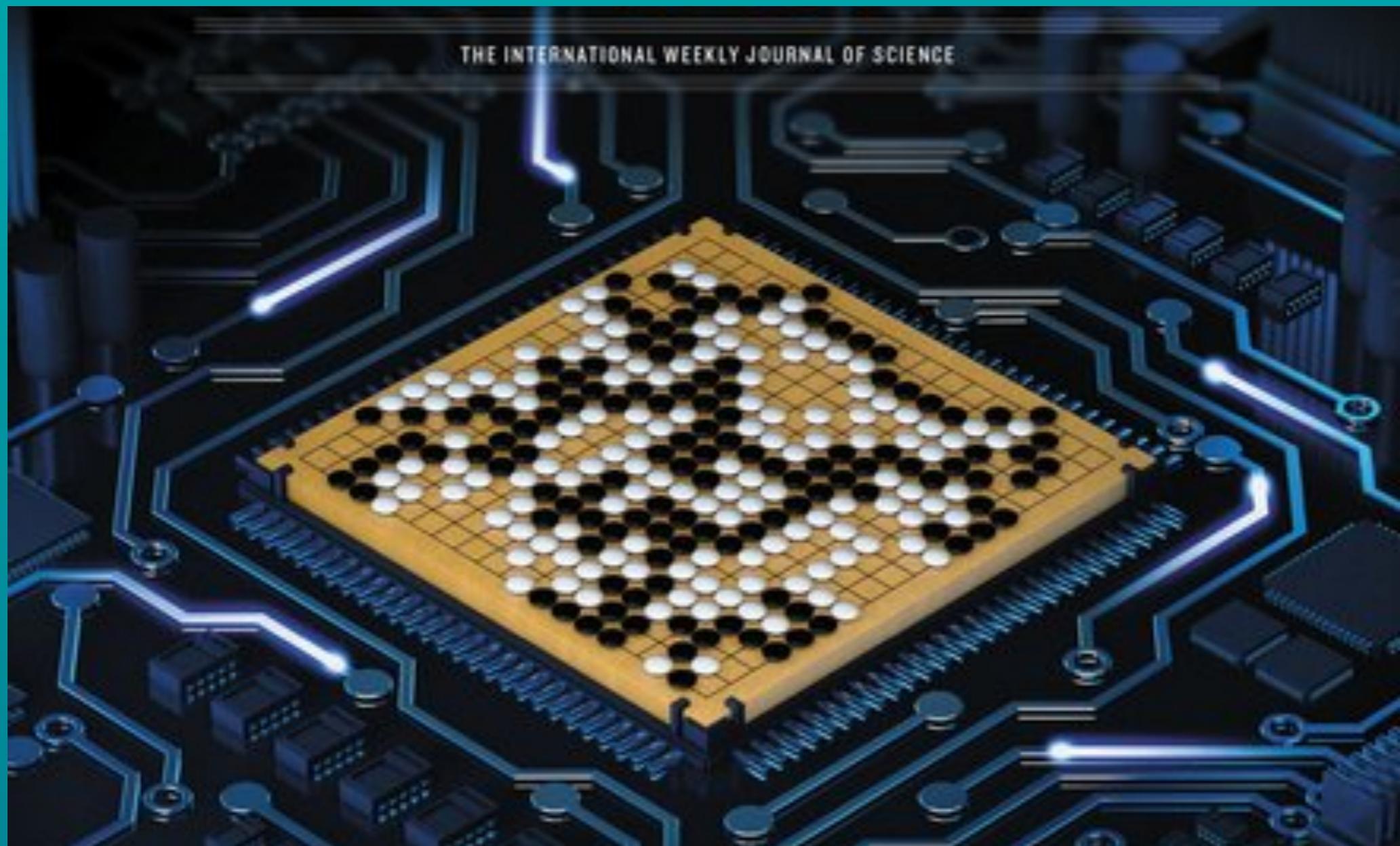


例子-3

找出奸细?



例子-4



AlphaGo

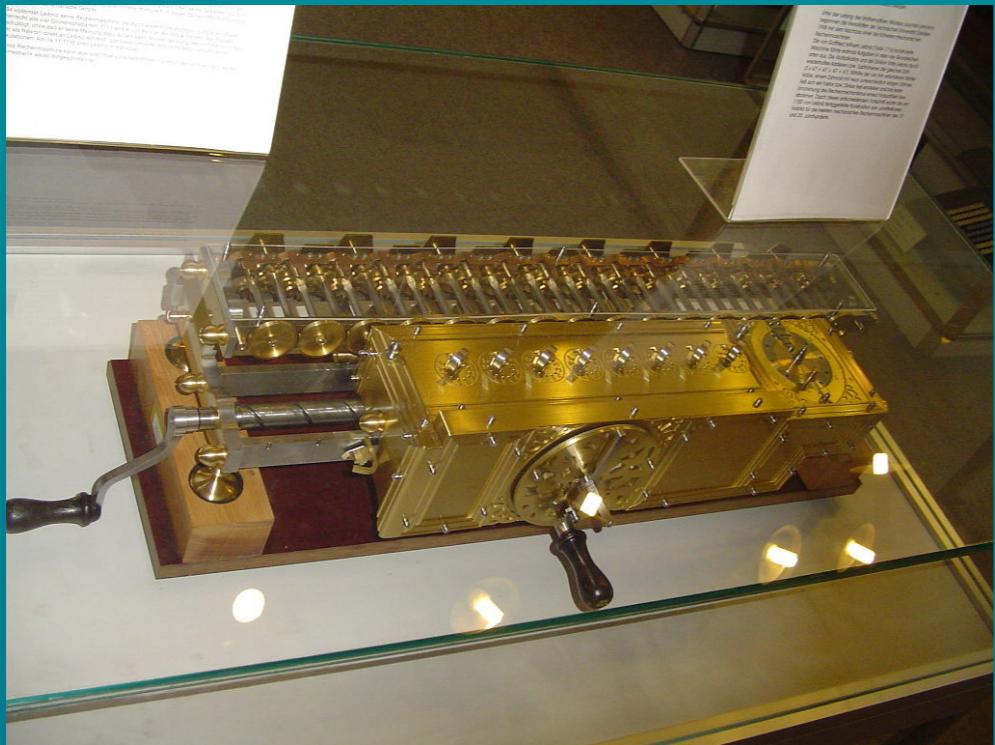


Boston Dynamic

Gottfried.W.Leibniz

莱布尼兹

- 1670s, 自动乘法机

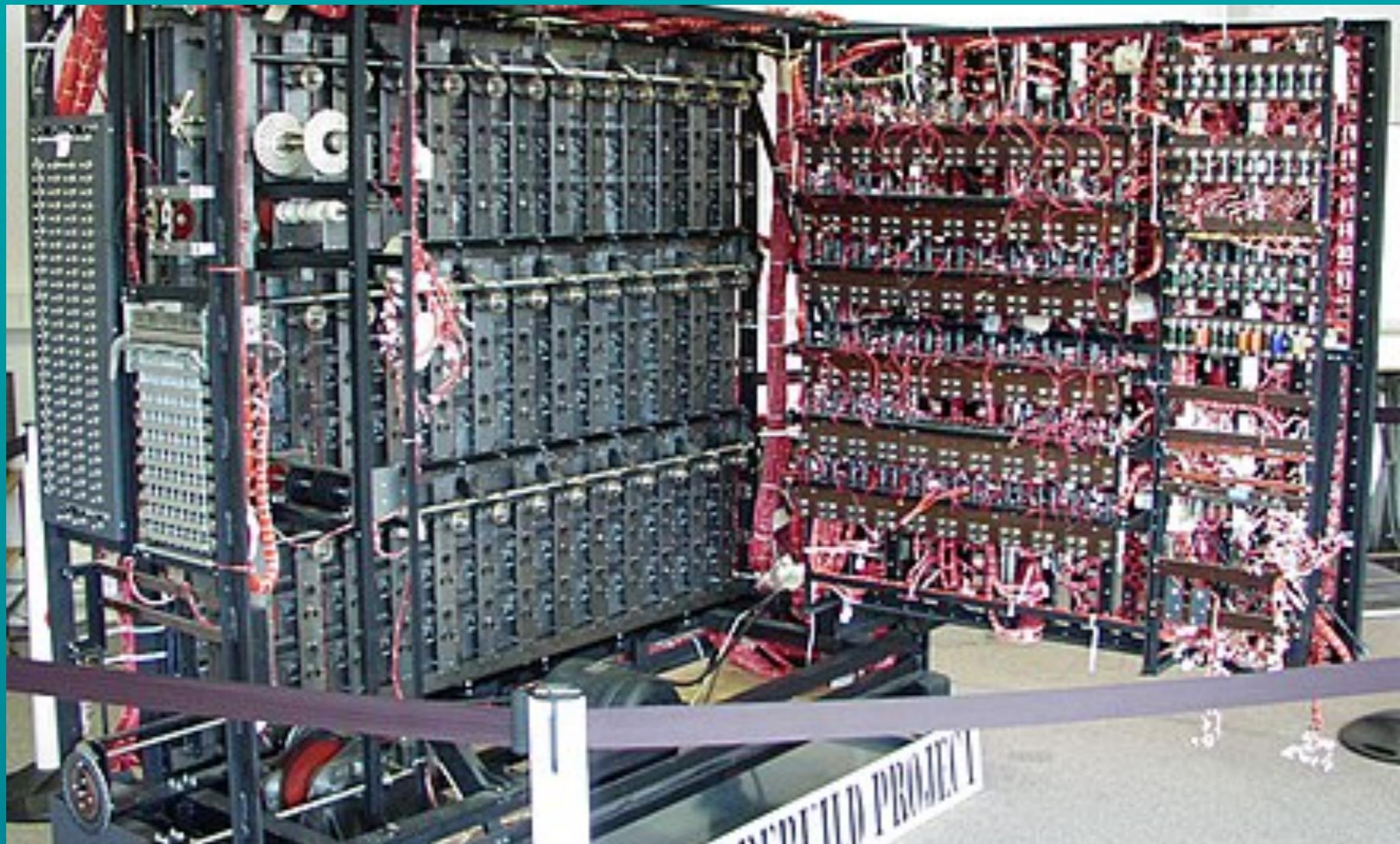


Alan Mathison Turing

阿兰图灵

- 《论可计算数及其在判定问题上的应用》 On Computable Numbers, with an Application to the Entscheidungsproblem, “这样的机器有能力解决任何可想像的数学难题”
- 《计算机械和智能》 , Computing Machinery and Intelligence, 图灵测试
- 混沌理论《形态发生的化学基础》 (英语: The Chemical Basis of Morphogenesis)





Bombe

Dartmouth workshop

达特茅斯会议

1956, Dartmouth

- 1) [Dr. Marvin Minsky](#)
- 2) [Dr. Julian Bigelow](#)
- 3) [Professor D.M. Mackay](#)
- 4) [Mr. Ray Solomonoff](#)
- 5) [Mr. John Holland](#)
- 6) [Mr. John McCarthy.](#)

For four weeks:

- 7) [Dr. Claude Shannon](#)
- 8) [Mr. Nathaniel Rochester](#)
- 9) [Mr. Oliver Selfridge](#)

For the first two weeks:

- 10) [Mr. Allen Newell](#)



Machine Learning

- Supervised Learning 监督学习
- Unsupervised Learning 非监督学习
- Reinforcement Learning 强化学习
- Representation (表征) + Policy (决策)

Deep Learning

- Complicated Structure 结构复杂
- Advanced Structure 先进的结构
 - RNN
 - CNN
 - ...
- Computing and Data consuming 需要大量计算力和数据

- Q1: 深度学习和机器学习的关系是什么?
- A. 深度学习是一种机器学习方法
- B. 机器学习是以前老的方法， 深度学习是新的方法
- C. 深度学习是用来取代机器学习的东西

- Q2: 为什么深度学习最近几年忽然崛起?
- A. 因为人们需要新的概念
- B. 因为数据量增多
- C. 因为硬件的提升
- D. 因为深度学习的专家年龄大了

自然语言处理的问题

- 1. 分词 (新词发现)
- 2. 实体发现
- 3. 词汇, 句子相似度的判断
- 4. 词汇, 句子含义的表示
- 5. 自动对话机器人
- 6. 阅读理解
- 7. 文本生成
- 表征 (Representation) + 决策 (Policy)
 - 表征: 词向量, 句子向量, 知识图谱
 - 决策: 主题模型, 自动对话机器人

自然语言处理的特点

- 1. Information Chaos 信息杂乱
- 2. Discrete 离散值
- 3. Unconventional 不规律
- 4. OOV 缺失值多

- Q3：关于自然语言处理正确的是
- A. 主要应用语言学知识
- B. 不同的语言需要的算法大不相同
- C. 数据数量多， 高质量数据少

Talk

- Assignments:
 - 1. 选取一个日常事件将其抽象为机器学习模型
 - 2. 设想一个自然语言的应用场景

推荐读物

- SICP, *Structure and Interpretation of Computer Programming*. 《计算机程序设计的解释和构造》
- *Introduction to Algorithms* 《算法导论》
- *Artificial Intelligence A Modern Approach (3rd Edition)* 《人工智能：一种现代方法》
- *Paradigms-of-Artificial-Intelligence-Programming*
- *Code Complete 2* 《代码大全》
- *Programming Pearls* 《编程珠玑》
- *Deep Learning*, 《深度学习》
- 《黑客与画家》
- 《数学之美》， 吴军
- *Fluent Python*
- *Hands on Tensorflow*
- **Conference: NIPS, ICML, ICLR, ACL, AAAI, ICJAI**
- <https://github.com/Deep-Learning-for-NLP-Chinese/References>

文本相似度的判断

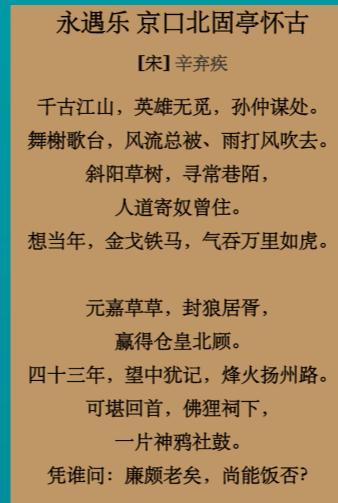
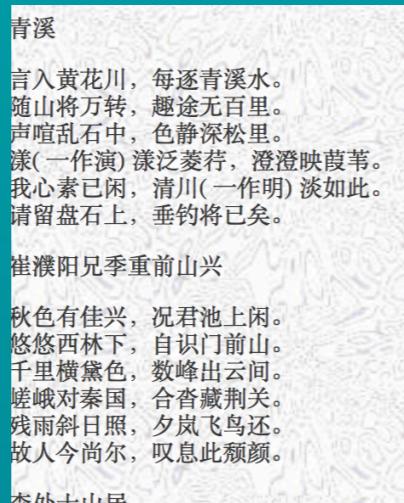
- 如何表征一段文字？如何表征一个单词？
- 我们先从最原始的开始：
 - Edit Distance 编辑距离
 - Insertion
 - Deletion
 - Substitution
 - 今天是个好日子 → 今天真是个好日子

Talk

- Edit Distance 的优点与缺点：
 - 语义层面的把握
 - 可解释性
 - 精准定位
 - 实现的难易程度

keywords

- 两篇文章，判断作者？



- 搜索时候网页与查询的相关性

- 例如：输入：机器学习的应用；机器学习，的，应用
- 哪些网页包含的这些词汇多；
- 篇幅长短？
- “的”停用词
- “机器学习”与“应用”哪个更重要？

tf-idf

- 1. 对于关键词而言， 其具有较强的标示性；
- 2. 对于关键词而言， 停用词应该自动排除；
- tf: term frequency
- idf: inverse document frequency
 - $\log(D/D_w)$; D_w is all the related documents

文本句子的表示

- tfidf不仅可以查询关键词；
- tfidf其实表示的是每个单词的重要性；
- 依据这种新型，可以为每个文章，句子建模；

Project-1

- 使用新闻数据库， 判断一篇文章是否为新华社发布的；
- 准确率达到一定程度之后， 判读一篇文章抄袭与修改的地方；
- hints:
 - 1. 先构建每个文本的tf-idf向量；
 - 2. 使用tf-idf进行向量搜索；
 - 3. 建立机器学习模型；
 - 4. 使用 edit distance 定位错误位置