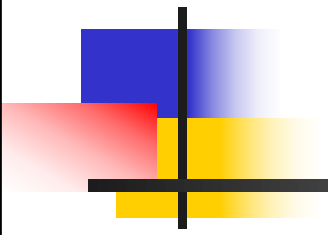


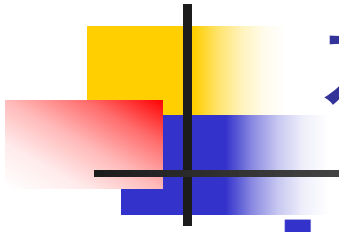
第1章： 自然语言处理基础

木豆

2017-10-17

学习要求

- 不是听故事，记住了、思考了，才算自己的
- 我们这个时代，崇尚短平快、碎片化的时代，学一样东西，很容易浅尝辄止。这个简单，没兴趣学，那个我百度一下就会了不需要再学。结果是，当你学真正想学的东西时发现，完了，怎么基础差这么多？
- 只要我们把准备高考的那种学习精神拿一点点出来，机器学习,NLP，so easy

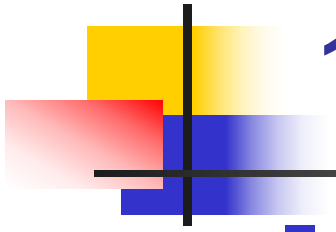


本章提纲

- 1.1 NLP 概述

  概念与发展动力\任务分类\BAT应用举例

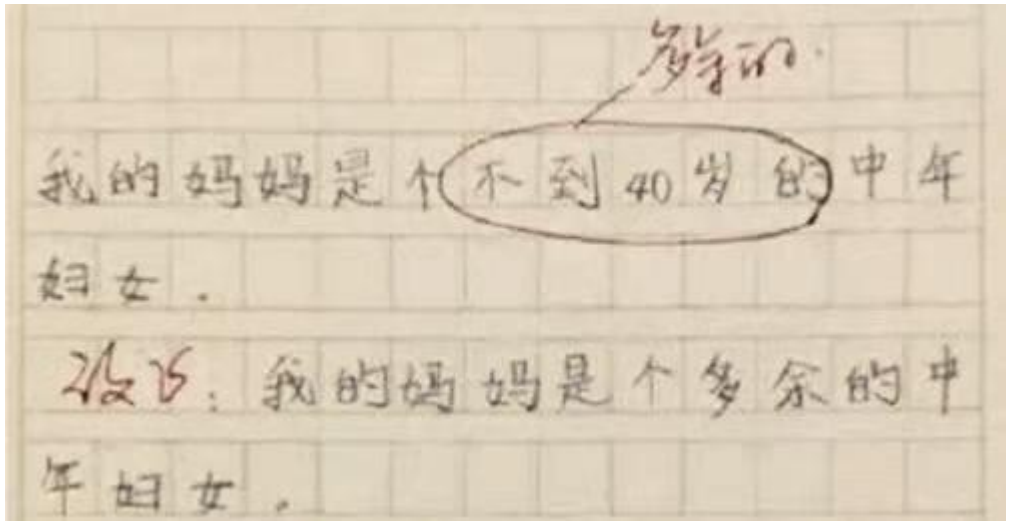
- 1.2 国内外大师
- 1.3 推荐资料



1.1 自然语言处理（NLP）概述

- 1.1.1 NLP的概念与发展动力
- 1.1.2 NLP任务分类
- 1.1.3 BAT应用举例

彻底攻克NLP非常难

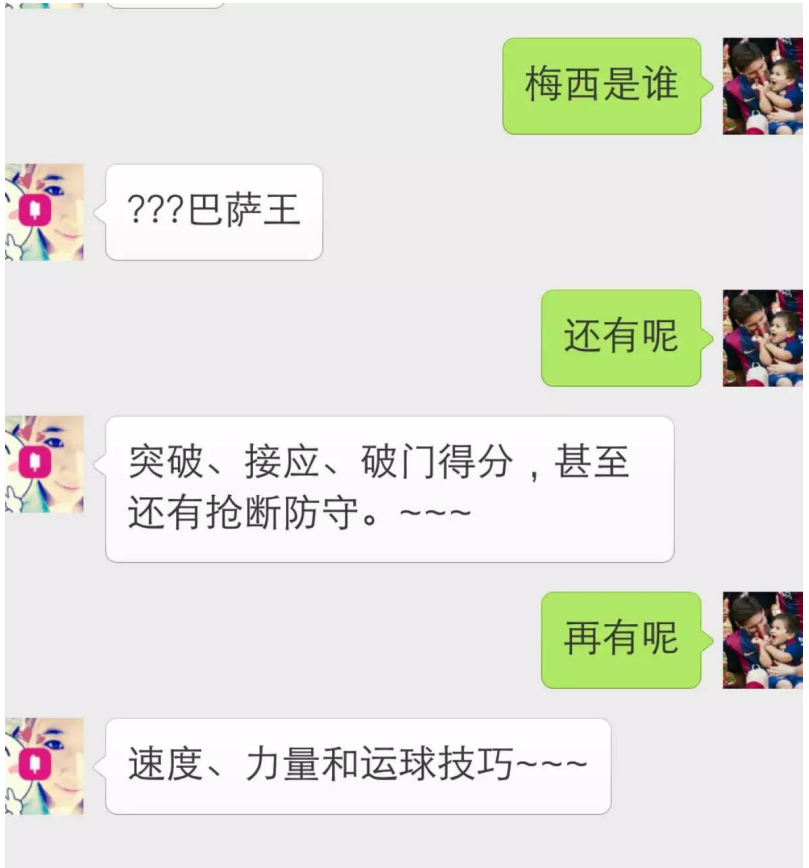


1.1.1自然语言处理的概念和发展动力

- 自然语言处理: NLP,又称计算语言学, 通过建立形式化的计算模型来分析、理解和处理自然语言的一门交叉学科(语言学、计算机科学、数学）
- 人类水平的自然语言处理，是一个人工智能完全问题。
- HIT关毅：支撑NLP发展的三个主要需求，是数据库技术、机器翻译、语音识别
- AI时代NLP的用武之地：speak to everything、analyze all message、mechine speaks and writes

AI时代的NLP：Speak to Everything

1）语音唤醒
DuerOS 唤醒万物
spaek to your phone
speak to your TV
speak to your car
speak to your fridge
speak to your door
....
2) everything can speak
调戏你的手机 Siri



AI时代的NLP：analyze all message

- 舆情分析
- 精准营销
- 川普竞选
- 从垃圾短信识别到微信群违法言论发现
- 所有信息都会被AI扫描分析

AI时代的NLP：AI speaks and writes

机器写新闻稿只是故事的开始
未来还有哪些文体会被机器写？
Maybe
情书？软文？学生作文？法律文书。。。

AI会写诗
AI会写小说
AI会写模仿赵雷写歌词
AI会主持新闻节目
AI会用张国荣的声音说话
AI会。。。
故事刚刚开始
我们拭目以待

国内写稿机器人成果分析

产品名	Xiaomingbot	Dreamwriter	WritingMaster	度秘（Duer）
所属平台	今日头条	腾讯	第一财经（阿里巴巴）	百度
对接平台	今日头条	腾讯财经 腾讯科技 腾讯体育	第一财经7*24看板、 第一财经电视、 一财网等集团频道	百度
上线时间	2016年8月	2015年9月	2016年5月	2016年8月
覆盖领域	NBA赛事 足球赛事	NBA赛事 足球赛事 财经报道	财经报道	NBA赛事
报道属性	战报	战报、快讯	快讯、长报道、电视新闻	赛事解说
是否配图	是	部分是	少数是	是
日产稿量 （平均值）	20-30篇	30-50篇	数百到上千篇	10-20场

1.1.2 NLP任务分类

- 序列标注问题

  （命名实体 品牌词识别 中文分词（词性标注） 句法分析 新词发现）

- 分类问题

  （情感分析 行业分类 意图识别 ）

- 改写问题

  （query扩展 改写 纠错 翻译）

- 生成问题

  （自动写稿 自动写诗 文本摘要 聊天机器人 自动问答 语音合成 ）

序列标注问题介绍

- 品牌词识别在搜索广告中是一个重要的问题。

  打个比方，你在搜狗广告部门工作，用户搜索”王者荣耀游戏官方网站“，你给出一个”阴阳师游戏官方网站“的广告，这是找死的节奏啊！等着腾讯爸爸来找你谈话了。当然，对于”王者荣耀“这样的热词，这是一个相对好处理的问题。但是搜索广告中有大量的广告主，有时候要去区分它们是很头疼的事情，比如”张爱国足疗店“、”李爱国足疗店“、”王爱国足浴店“，他们的字面相似度是如此之高，但是又是不同的服务商。

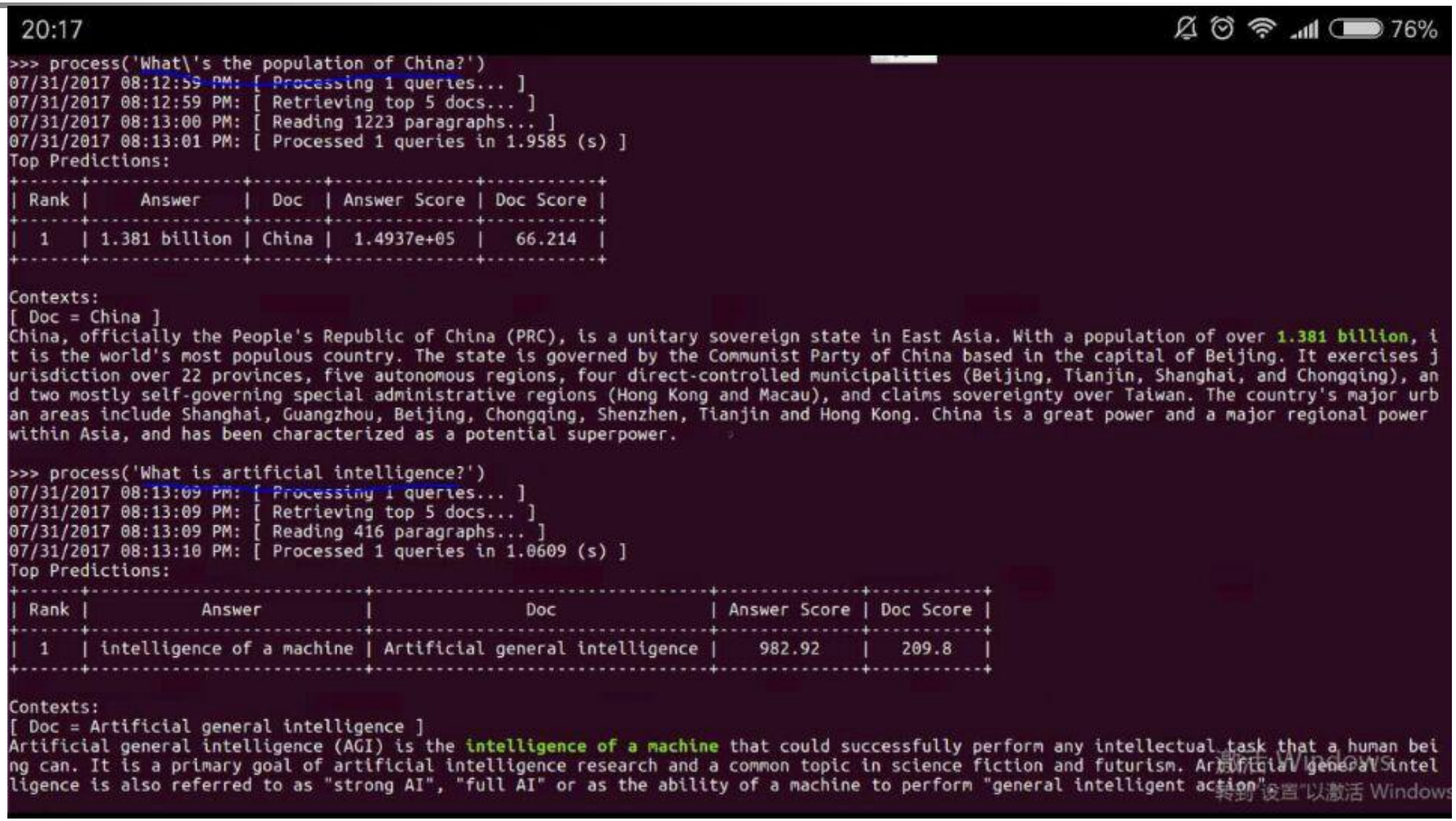
- 序列标注问题的经典方法是 CRF。我们后续在《进阶分享》中会介绍

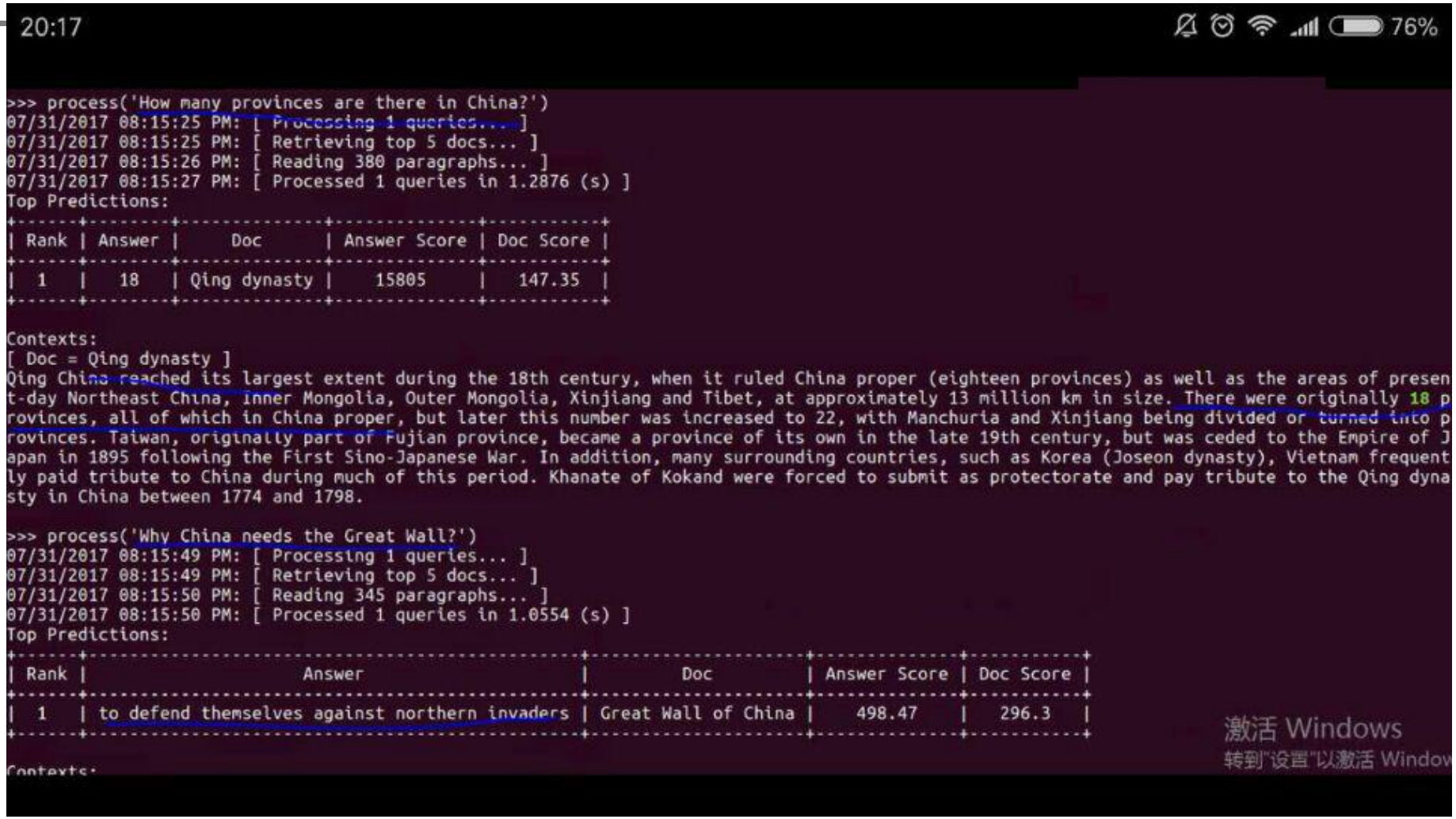
1.1.2 NLP任务分类：RNN写诗

娼男离身还自惭。空梁几日近晓真，人生一望月归时。
细低三年前柳家，我吟曾恨见人家。庆天天绕烟罩尽，曙作长安人旧来
临思将照事，游处映春来。出故至何早，月冠琼虮风。
独坐此山下，除看闻住香。好掌从夜树，不事若轻低。
张掉飞幕来情何，不在忽知不必分。如今欢客上烟酒，阋待归青是寂寥。
雨滤富下鸳惊春，绿叶初黄不见规。欲问好人新作地，终岁西南似坟花。
山泉流前梦入破，旺们一树尔真后，怜君幽士洛阳秋。
越北源如生，征帆今净洁。东陵道史冷，榜路不言楼。
蜀尉送恩春展眉，和啼似思对心情。一宵不咒二终鸟，明月将军紫暗销。
天下眼前瞿闭名，野花谢断入中花。半春中鸟鞭襄星，时有窗中霞满书。
楚天庆阁开斋长，霜阴灯和白柳台。还将帝城轻须缩，一吟风雨更惆怅
关门同丛语，却窗悠随行。年来青桥说，江冲易郸漫。
秋雨言露此流霞，风岁落浦月鹃斜。正知月频如霜雨，嫫上大空满北光。
但臣访舸掌，兰少夜为猫。风桥见西围，回代向高黯。湖碧莫比马蹄落，为君落尽到青山。
一声池水竹山醒，圣历喧人长满尘。一回遗看随不得，半涩无限官夜迎。
落日悲寒竹，君英遥榴阳。秋风流满千，斜方多古寺。
九华清阴溪，何人拜王骨。天下何时阻，雪头上万年。堪君垂露常，只有抱庭寒。
行行鸟事不成身，无夜平间去暮寒。晋今却吟千多夜，鹪鸪飞病素人同。

1.1.2 NLP任务分类：自动问答

- https://github.com/facebookresearch/DrQA




20:17
76%
>>> process('How many provinces are there in China?')
07/31/2017 08:15:25 PM: [Processing 1 queries...]
07/31/2017 08:15:25 PM: [Retrieving top 5 docs...]
07/31/2017 08:15:26 PM: [Reading 380 paragraphs...]
07/31/2017 08:15:27 PM: [Processed 1 queries in 1.2876 (s)]
Top Predictions:
| Rank | Answer | Doc | Answer Score | Doc Score |
| 1 | 18 | Qing dynasty | 15805 | 147.35 |
Contexts:
[Doc = Qing dynasty]
Qing China reached its largest extent during the 18th century, when it ruled China proper (eighteen provinces) as well as the areas of present-day Northeast China, inner Mongolia, Outer Mongolia, Xinjiang and Tibet, at approximately 13 million km in size. There were originally 18 provinces, all of which in China proper, but later this number was increased to 22, with Manchuria and Xinjiang being divided or turned into provinces. Taiwan, originally part of Fujian province, became a province of its own in the late 19th century, but was ceded to the Empire of Japan in 1895 following the First Sino-Japanese War. In addition, many surrounding countries, such as Korea (Joseon dynasty), Vietnam frequently paid tribute to China during much of this period. Khanate of Kokand were forced to submit as protectorate and pay tribute to the Qing dynasty in China between 1774 and 1798.
>>> process('Why China needs the Great Wall?')
07/31/2017 08:15:49 PM: [Processing 1 queries...]
07/31/2017 08:15:49 PM: [Retrieving top 5 docs...]
07/31/2017 08:15:50 PM: [Reading 345 paragraphs...]
07/31/2017 08:15:50 PM: [Processed 1 queries in 1.0554 (s)]
Top Predictions:
| Rank | Answer | Doc | Answer Score | Doc Score |
| 1 | to defend themselves against northern invaders | Great Wall of China | 498.47 | 296.3 |
Contexts:
激活 Windows
转到"设置"以激活 Windows

1.1.2 NLP任务分类:几个高级任务

- 语言模型

kenlm语言模型：http://kheafield.com/code/kenlm/

rnn语言模型 ：https://github.com/wpm/tfrnnlm

语言模型可以对 句子通顺度建模；可以用于纠错

- 句法分析

Stanford parser 最经典

syntaxNet最快、准确率最高

- 机器翻译

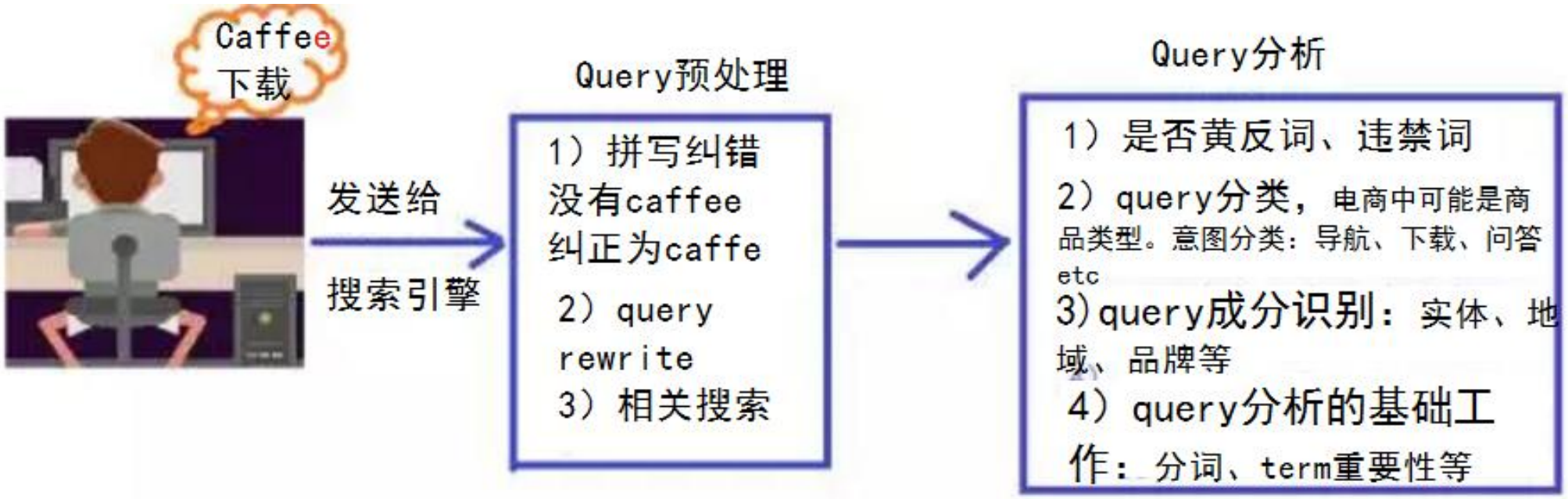
统计机器翻译 vs 神经机器翻译

引入seq2seq模型的神经机器翻译，不仅仅可以用于语言的翻译

1.1.3 BAT应用举例：NLP对BAT有多重要

- 在像百度、搜狗、360、淘宝、京东等这样的互联网公司里，核心数据仍然是文本数据，**基于文本**可以做大量的工作，来为公司的核心业务目标服务。例如搜索业务（包括网页搜索，以及电商搜索、地图搜索等都是类似原理），需要对用户的搜索query做大量的分析，分析用户的搜索意图，进而给用户呈现他需要的搜索结果，并且顺带捎上搜索广告。在这个看似简单的过程中，其实衍生出非常多的技术问题.

以搜索为例，NLP之于BAT



1.1.3 BAT应用举例

- 百度：

  基于NMT的智能标题生成项目

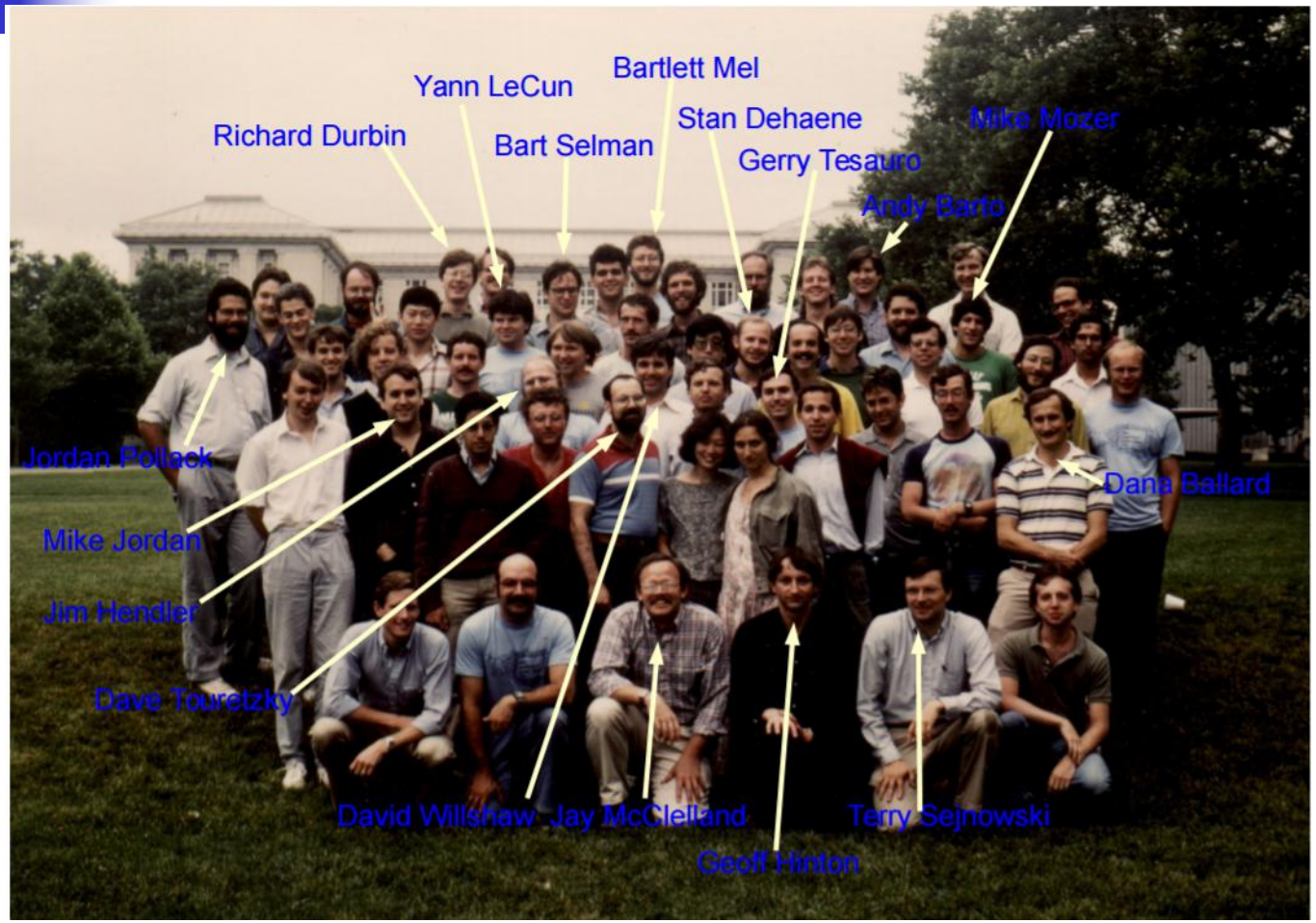
- 其他举例

  - 关键词推荐
  - 生成式触发模型
  - 品牌词保护

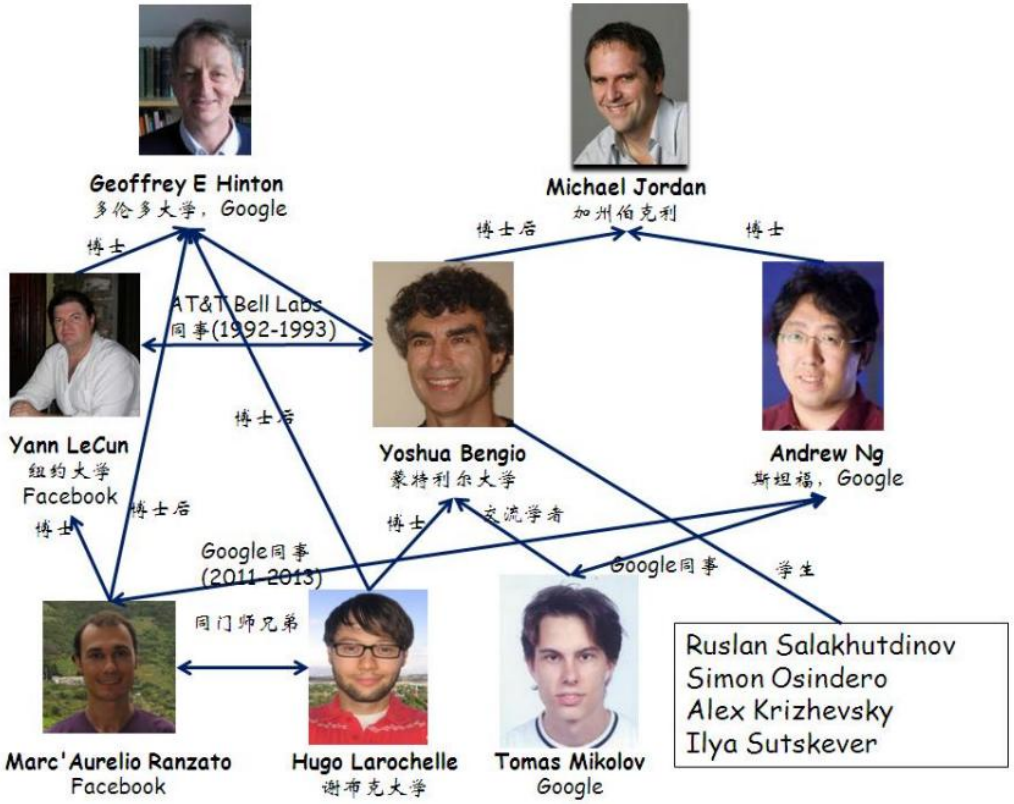
Query	Old	New
初中毕业	初中毕业?学历+技能打造全..	初中毕业**学什么技术有前途?**
独轮电动车代理	独轮电动车代理十大品牌电动车火爆加盟!	独轮电动车代理-开电动车加盟店，**厂家免费铺货!**
电脑编程	电脑编程课程学费介绍 2016年招生中	电脑编程**零基础学习**，电脑编程**入门到精通**

跳转

八卦时间：国际大师


Yann LeCun
Bartlett Mel
Richard Durbin
Bart Selman
Stan Dehaene
Mike Mozer
Gerry Tesauro
Andy Barto
Jordan Pollack
Dana Ballard
Mike Jordan
Jim Hendler
Dave Touretzky
David Willshaw
Jay McClelland
Terry Sejnowski
Geoff Hinton

八卦时间：国际大师



三大巨头，两大门派

三大巨头：
Hinton提出深度学习概念激活了整个领域、LeCun发表了卷积神经网络（CNN）这样的阶段性突破成果的前提下，Bengio对自然语音处理难题的贡献

两大门派：
LeCun是hinton的博士，属于Hinton派
Bengio师从Mike Jordan，属于Mike Jordan派

Hinton派

- **Hinton**深度学习当之无愧的宗师

  1986年，提出的反向传播算法（BP），神经网络中最重要的理论基础

  提出Distributed Representation，克服 one-hot representation 的缺点

  后来，提出对比散度（contrastive divergence）算法用于训练RBM

  神经网络的低潮期，他不抛弃，不放弃

  扛出“deep learning”的大旗，带领学生参加ImageNet一战成名，神经网络华丽转身

- 卷积神经网络（CNN）的发明人Lecun是Hinton的博士
- AlexNet ： 作者 Alex Krizhevsky，属于多伦多大学Hinton组。

Michael Irwin Jordan派：低调的豪门

伯克利大学教授，门下的牛人辈出,https://en.wikipedia.org/wiki/Michael_I._Jordan

Yoshua Bengio:

青出于蓝而胜于蓝的典范，深度学习三巨头（Lecun Bengio Hinton）之一

加拿大计算机科学家。加拿大这个国家很有意思，深度学习三巨头，独占两席。

David Blei：

LDA主题模型提出人　Latent Dirichlet Allocation

http://www.cs.columbia.edu/~blei/

Andrew Ng：

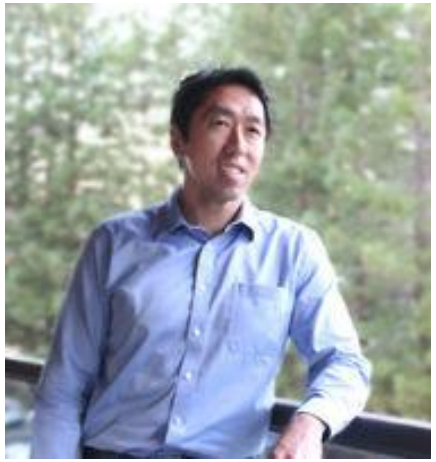
1）Coursera联合创始人，斯坦福大学人工智能实验室主任，百度首席科学家，“Drive.ai创始人的老公”

2）google cat 项目负责人；参与谷歌第一代深度学习框架 DistBelief，也就是tensorflow的前身

八卦时间：华人大师









推荐书籍

- 自然语言处理

  Christopher Manning（斯坦福大学教授），统计自然语言处理

  Steven Bird等, Python自然语言处理

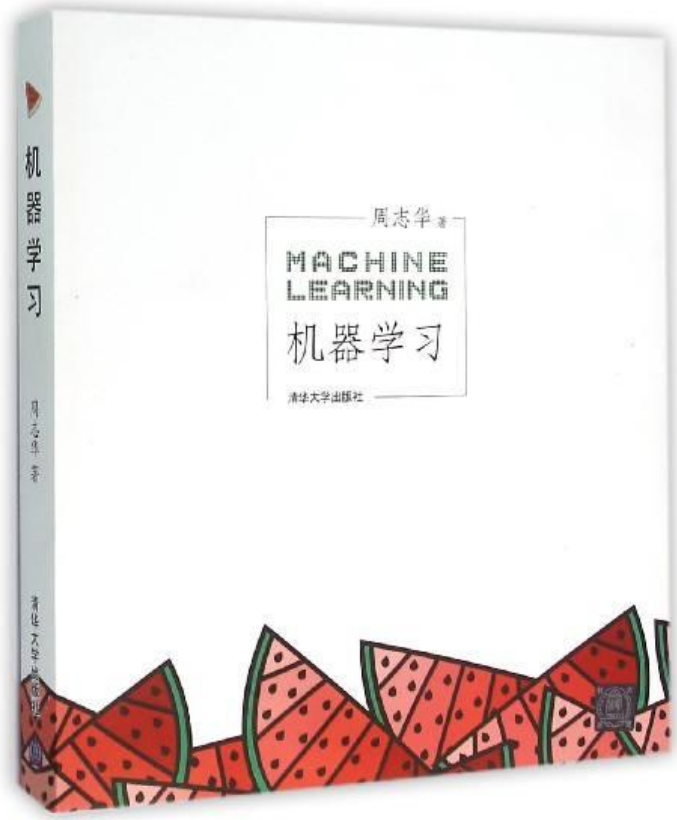
- 机器学习

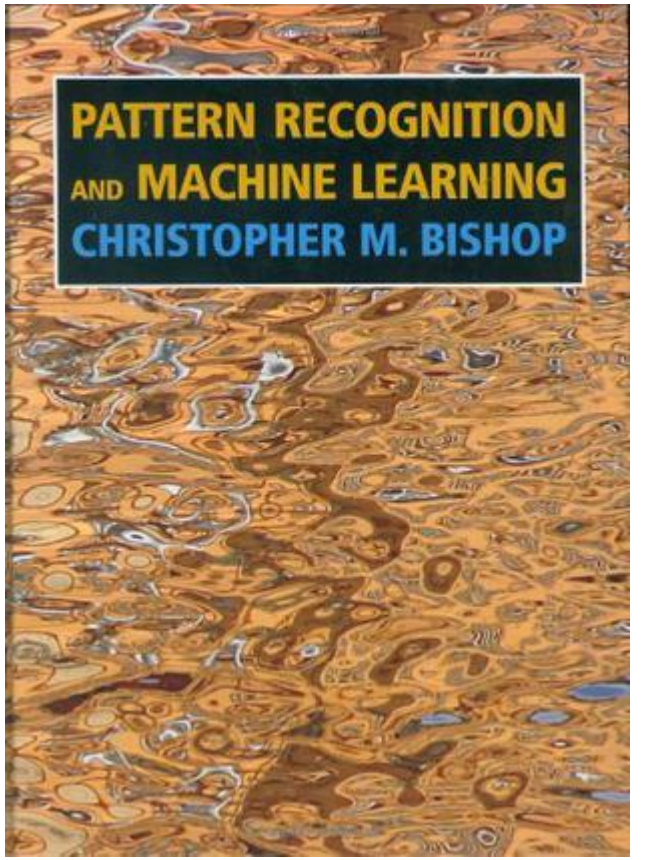
  李航（博士），统计学习方法

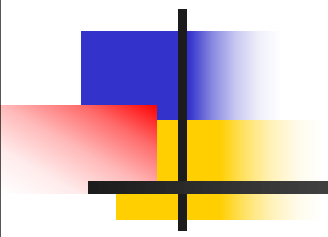
  周志华（南京大学），机器学习

  PRML，Christopher Michael Bishop

推荐书籍


机器学习
周志华 著
MACHINE LEARNING
机器学习
清华大学出版社


PATTERN RECOGNITION
AND MACHINE LEARNING
CHRISTOPHER M. BISHOP



第2章：文本的特征表示与语言模型

这部分内容见纸质打印PPT

更多分享&&关注我

微信扫码