

# 概述

邱锡鹏

复旦大学

<http://nlp.fudan.edu.cn/xpqiu>



# 参考教材

---

- ▶ Chengxiang Zhai, Sean Massung, Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining, ACM and Morgan & Claypool Publishers, July 2016.
  - ▶ <http://dl.acm.org/citation.cfm?id=2915031>
- ▶ Toby Segaran, Programming Collective Intelligence: Building Smart Web 2.0 Applications, O'Reilly Media, 2007.



# 数据

---

- ▶ 结构化数据
- ▶ 非结构化数据
  - ▶ 文本
  - ▶ 声音
  - ▶ 图像
  - ▶ 视频

“Newspapers represent 25 terabytes annually, magazines represent 10 terabytes ... office documents represent 195 terabytes. It is estimated that 610 billion emails are sent each year representing 11,000 terabytes.”

In 2003



# 文本数据

---

## ► 表现形式为自然语言（英文、中文等）

- 网页
- 社交媒体
- 产品评论
- 新闻
- 科学文献
- 电子邮件
- 工作报告
- ...



# 文本数据

---

- ▶ 文本是人们最自然的信息表示
- ▶ 文本是人们最常用的信息表示
- ▶ 文本是最具表达力的信息表示



# 结构化VS非结构化

## 谍影重重5 Jason Bourne (2016)



导演: 保罗·格林格拉斯

编剧: 保罗·格林格拉斯 / 罗伯特·鲁德格姆 / 克里斯多福·劳斯

主演: 马特·达蒙 / 汤米·李·琼斯 / 艾丽西亚·维坎德 / 文森特·卡索 / 朱丽娅·斯蒂尔斯 / 更多...

类型: 动作 / 悬疑 / 惊悚

官方网站: [www.jasonbourne-film.com](http://www.jasonbourne-film.com)

制片国家/地区: 美国

语言: 英语

上映日期: 2016-08-23(中国大陆) / 2016-07-29(美国)

片长: 123分钟 / 124分钟(中国大陆)

又名: 叛谍追击5: 身份重启(港) / 神鬼认证: 杰森伯恩(台)

IMDb链接: [tt4196776](https://www.imdb.com/title/tt4196776)

谍影重重5的影评 ····· (全部501)

我来评论这部电影



画蛇添足: Jason Bourne首映感想

Wenhao 2016-07-28 06:35:06 ★★★★★

(有剧透, 请最好先看完电影以及之前的三部曲) Jason Bourne回来了, 而我的观感却很复杂。技术上来说, 这仍旧是配得上之前三部曲的电影。但是内核却有着重大缺陷。旧的三部曲里, 第一部是逃亡, 第二部是赎罪, 第三部是寻根。B.....

731/883 有用 170回复



疲惫杀手的中年危机, 但该谈情怀的时候还是要谈...

旺财博士 2016-07-29 14:30:51 ★★★★★

不知从什么时候开始, “情怀”在人们的讨论语境中慢慢变得越来越不像个褒义词。每每说起, 仿佛总有半是“将就”半是“卖老”的意思, 仿佛现在的年轻人未曾经历当初的风情万种和惊涛骇浪就没资格对如今的美人白头和英雄迟暮评头.....

816/845 有用 86回复



西方的哪一个国家我没去过

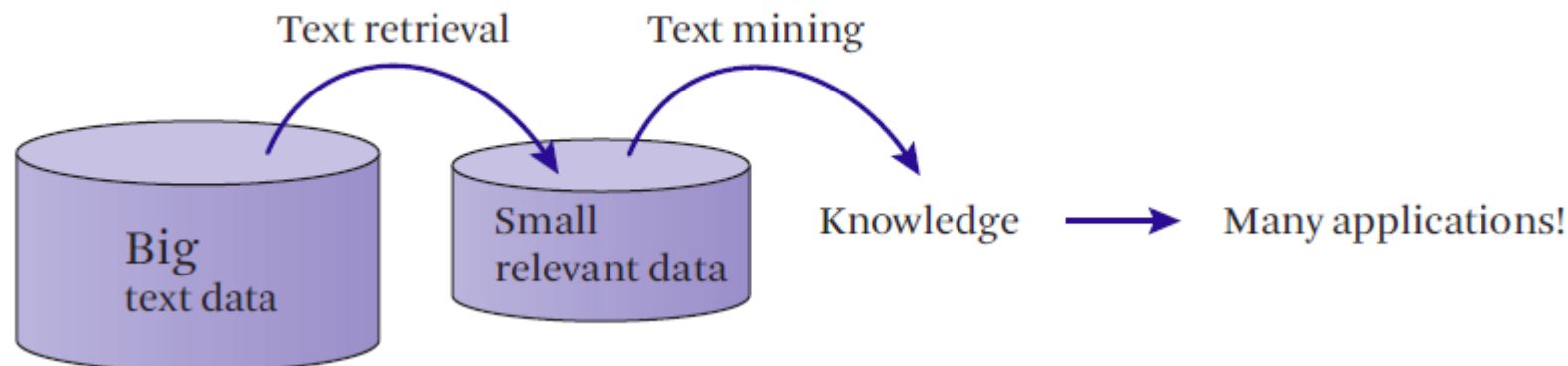
马泽尔法克尔 2016-08-11 01:12:09 ★★★★★

假使一些完全无中生有的东西你再帮他说一遍你等于你也有责任的剧透: 隐居多年的杰森伯恩来到了圣域雅典, 为了成为圣斗士、获得圣衣而苦练天马流星拳。这天正在打拳的他遇到了故人尼基, 后者正遭到CIA的追杀, 原因是她出于个.....

540/540 有用 109回复

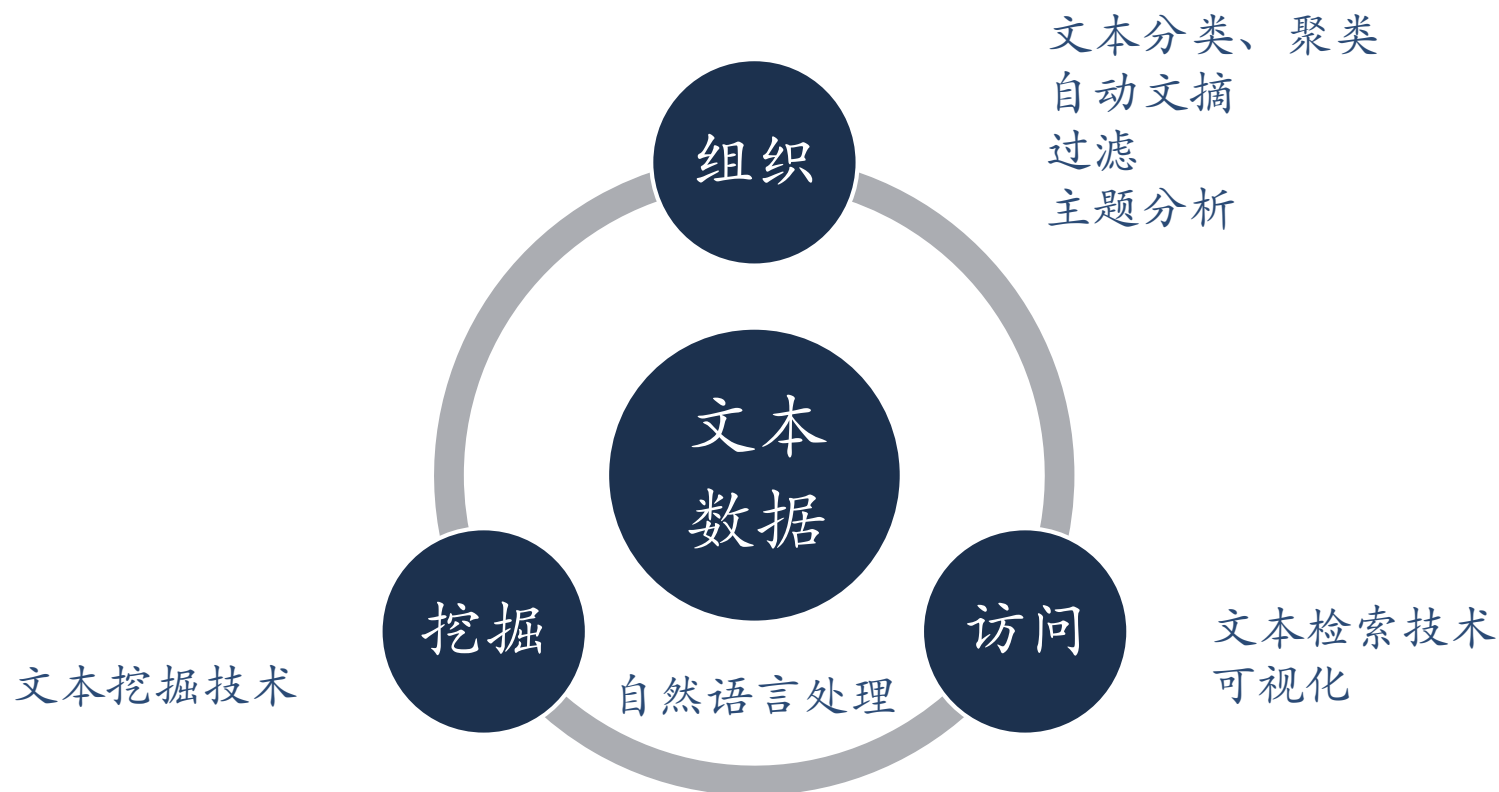
# 文本数据管理与分析

- ▶ 文本挖掘（分析）
- ▶ 信息检索
- ▶ 自然语言处理
- ▶ 机器学习





# 文本数据管理与分析





# 文本挖掘



# 数据挖掘

- ▶ 数据挖掘 (Data Mining) 就是从大量的、不完全的、有噪声的、模糊的、随机的实际应用数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程。





# 数据挖掘的社会需求

- ▶ 著名的“啤酒尿布”案例：美国加州某个超级卖场通过数据挖掘发现，下班后前来购买婴儿尿布的男顾客大都购买啤酒。于是经理当机立断，重新布置货架，把啤酒类商品布置在婴儿尿布货架附近，并在二者之间放置佐酒食品，同时还把男士日常用品就近布置。这样，上述几种商品的销量大增。



# 数据挖掘-分类

---

## ▶ 技术分类

- ▶ 预言 Predication
  - ▶ 用历史预测未来
- ▶ 描述 Description
  - ▶ 了解数据中潜在的规律

## ▶ 数据挖掘技术

- ▶ 关联分析
- ▶ 序列发现
- ▶ 分类（预言）
- ▶ 聚集
- ▶ 异常检测
- ▶ 汇总
- ▶ 回归
- ▶ 时间序列分析



# 文本挖掘

---

- ▶ 从文本中发现有用信息的过程。
- ▶ 挖掘策略
  - ▶ 利用检索技术直接挖掘文档的内容；
  - ▶ 在搜索引擎等工具处理基础上做进一步的处理，以便获得更为精确和有用的信息。
- ▶ 面临的问题
  - ▶ 算法效率
  - ▶ 数据噪声或缺失
  - ▶ 数据隐私
  - ▶ 数据安全问题



# 文本挖掘

---

## ▶ 主要应用研究内容

- ▶ 文本摘要
- ▶ 文本分类
- ▶ 文本聚类
- ▶ 关联分析
- ▶ 分布分析
- ▶ 趋势预测



# 文本挖掘

---

## ► 文本摘要

- ▶ 从文档中抽取关键信息，用简洁的形式对文档内容进行摘要或解释。这样，用户不需要浏览全文就可以了解文档或文档集合的总体内容。
- ▶ 有篇首截取法、上下文截抽取法、论题句抽取法、仿人法等。



# 文本挖掘

---

## ► 文本分类

- 文本分类是指按照预先定义的主题类别，为文档集中的每个文档确定一个类别。
  - Yahoo!采用人工分类，大大影响了索引的页面数目。
  - 利用自动文本分类技术可以对大量文档进行快速、有效分类，大型搜索引擎都采用自动分类技术。





# 文本挖掘

---

## ► 文本聚类

- 文本聚类是将文档集合分成若干个簇，要求同一簇内文档内容的相似度尽可能地大，而不同簇间的相似度尽可能地小。
- “聚类假设”
  - 与用户查询相关的文档通常会聚类得比较靠近，而远离与用户查询不相关的文档。
- 意义
  - 利用文本聚类技术将搜索引擎的检索结果划分为若干个簇，用户只需要考虑那些相关的簇，大大缩小了所需要浏览的结果数量。



# 文本挖掘

## ► 关联分析

- 从文档集合中找出不同词语之间的关系。
- 实例：
  - 有人提出一种算法，可以从大量文档中发现一对词语同时出现的模式，利用该算法可在Web上寻找作者和书名的出现模式，从而发现了若干本在Amazon网站上找不到的新书籍。
  - 以Web上的电影介绍作为测试文档，通过使用OEM模型从页面中抽取词语，进而得到一些关于电影名称、导演、演员、编剧的出现模式。
  - 从科技论文中挖掘主题词演变模式，发现学科发展趋势。



# 文本挖掘

---

## ► 分布分析

- 指通过对文档的分析，得到特定数据在某个历史时刻的分布情况。
- 实例：
  - Feldman等人使用多种分布模型对路透社的两万多篇新闻进行了挖掘，得到主题、国家、组织、人、股票交易之间的相对分布情况。



# 文本挖掘

---

## ► 趋势预测

- 指通过对文档的分析，得到特定数据将来的取值趋势。
- 实例
  - Wuthrich等人通过分析Web上出版的权威性经济文章，对每天的股票市场指数进行预测，取得了良好的效果。



# 文本挖掘

---

## ► 挖掘对象

- 网站中超级链接结构之间的关系，它体现了文档之间的逻辑关系，与文档所处位置无关。

## ► 目标

- 找到隐藏在一个个页面之后的链接结构模型，可以用这个模型对Web页面重新分类，用于寻找相似的网站，评价网站社会关系及其对应用影响。

# 信息检索



# 信息检索

- ▶ 依照用户的信息需求提供查询的方法以及查询的过程。包括
  - ▶ 用户接口(User Interface)
    - ▶ 输入查询(Query), 返回排序后的结果文档
    - ▶ 可视化(Visualization)
    - ▶ 支持用户进行相关反馈(Feedback)
  - ▶ 两种模式: pull (ad hoc) 和push (filtering)。
    - ▶ Pull: 用户是主动的发起请求, 在一个相对稳定的数据集上进行查询。
    - ▶ Push: 用户事先定义自己的兴趣, 系统在实时数据上进行操作, 将满足用户兴趣的数据推送给用户



# 文本信息检索

---

- ▶ 文本信息检索是针对文本的信息检索技术
- ▶ 对于大规模的语料库，任何检索都可能返回数量众多的结果，因此对检索结果进行排序是必须的。

<https://zh.wikipedia.org/wiki/文本信息检索>





# 文本信息检索

- ▶ 形式上说，信息检索中的相关度是一个函数 $R$ ，输入是查询 $Q$ 、文档 $D$ 和文档集合 $C$ ，返回的是一个实数值  $R=f(Q,D,C)$
- ▶ 信息检索就是给定一个查询 $Q$ ，从文档集合 $C$ 中计算每篇文档 $D$ 与 $Q$ 的相关度并排序(Ranking)。
- ▶ 相关度通常只有相对意义，对一个 $Q$ ，不同文档的相关度可以比较，而对于不同的 $Q$ 的相关度不便比较相关度的输入信息可以更多，比如用户的背景信息、用户的查询历史等等



# 文本信息检索

---

- ▶ 相关(relevant)、相关度(relevance)
  - ▶ 相关取决于用户的判断，是一个主观概念，不同用户做出的判断很难保证一致，即使是同一用户在不同时期、不同环境下做出的判断也不尽相同。



# 文本信息检索模型

---

## ► 常用的信息检索模型：

- 向量空间模型 (Vector Space Model, VSM)
- 概率模型 (Probabilistic Model)



# 排序

---

## ▶ 排序算法

### ▶ PageRank

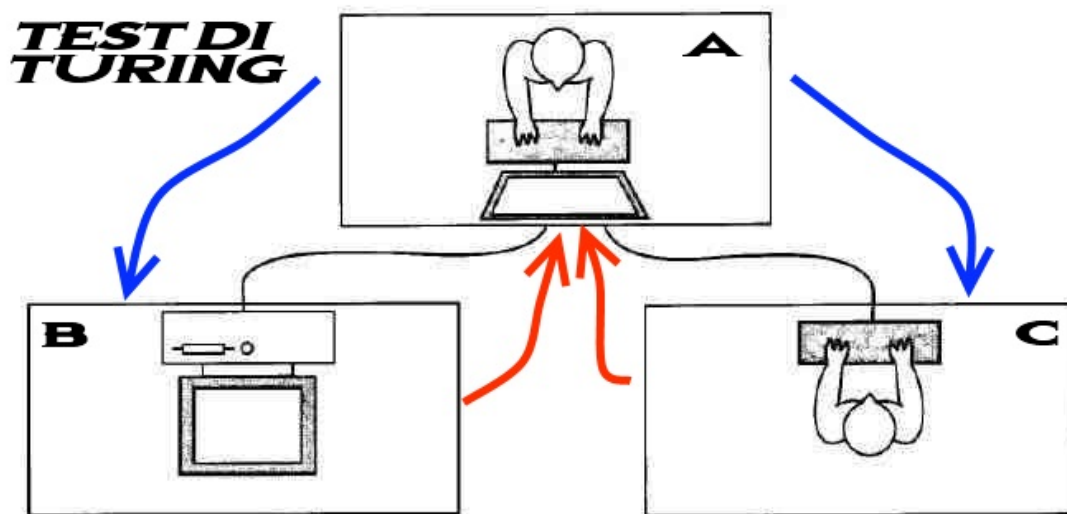
- ▶ 一个页面尽管没有被多次引用，但被一个重要页面引用，则这个页面很可能是重要的；一个页面的重要性被均分并被传递到它所引用的页面

### ▶ HITS

- ▶ 衡量网页重要性有两个要素 (ranking)：权威级别（依赖于指向它的页面）、中心级别（依赖于它指向别人的页面）

# 自然语言处理

# 从人工智能开始



Alan Turing

## 自然语言处理：理解和生成

# 自然语言处理是人工智能的瓶颈

- ▶ 图灵测试涉及的众多技术中就包括自然语言处理技术。
  - ▶ 自然语言理解系统
    - ▶ 自然语言转化为计算机程序更易于处理的形式
  - ▶ 自然语言生成系统
    - ▶ 把计算机数据转化为自然语言

自然语言处理不等于研究语言学（计算语言学）、文学。

“我每开除一名语言学家，我的语音识别系统错误率就降低一个百分点。” -- Frederick Jelinek



# 语言

- ▶ 语言是指在一个有限的字符集上，产生的符合一定规则的字符串集合。
- ▶ 自然语言 VS 人工语言
  - ▶ 形式语言 (Chomsky, 1950)
  - ▶ 区别
    - ▶ 自然语言：歧义性
    - ▶ 人工语言：确定性







# 歧义：中文分词

▶ 不同的语言环境中的同形异构现象，按照具体语言环境的语义进行切法。

▶ 交叉歧义

▶ 他/说/的/确实/在理

▶ 组合歧义

▶ 两个/人/一起/过去、个人/问题

▶ 从马/上/下来、马上/就/来

▶ 句子级歧义

▶ 白天鹅在水里游泳

▶ 该研究所获得的成果

} 伪歧义



# 歧义：指代消解

- ▶ 指代是自然语言表达中的常见现象。
- ▶ 在语言学中，
  - ▶ 避免已经出现的字词重复出现在文章的句子上，导致语句结构过于赘述和语意不够清晰
  - ▶ 使用代名词或是普通名词来代替已经出现过的字词称之为共指。
- ▶ 我们把香蕉给猴子，因为它们饿了
- ▶ 我们把香蕉给猴子，因为它们熟透了



# 中文分词难点：新词

---

## ▶ 也称为：未登录词

- ▶ 在字典或训练语料中都没有出现过的词。

## ▶ 常见新词

- ▶ 人名、机构名、地名、产品名、商标名、简称、省略语等
- ▶ 专业术语
- ▶ 网络新词



# 时间短语识别

- ▶ 时间相关信息的处理是自然语言理解过程中一个非常重要的部分。

Input:

08 年北京举行奥运会，8 月 8 号开幕。四年后的七月二十七日，伦敦奥运开幕。

The Beijing Olympic Games took place from August 8, 2008. Four years later, the London Olympic Games took place from July 21.

今天我很忙，晚上 9 点才能下班。周日也要加班。

I'm busy today, and have to come off duty after 9:00 PM. And I also have to work this Sunday.

08 年 (2008)	2008
8 月 8 号 (August 8)	2008-8-8
七月二十七日 (July 21)	2012-7-27
今天 (today)	2012-2-22 <sup>1</sup>
晚上 9 点 (9:00 PM)	2012-2-22 21:00
周日 (this Sunday)	2012-2-26

<sup>1</sup> The base time is 2012-02-22 10:00AM.



# 更困难的例子

---

- ▶ 冬天，能穿多少穿多少；夏天，能穿多少穿多少。
- ▶ 剩女的原因：一是谁都看不上，二是谁都看不上。
- ▶ 单身的来由：原来是喜欢一个人，现在是喜欢一个人。
- ▶ 女致电男友：地铁站见。如果你到了我还没到，你就等着吧。如果我到了你还没到，你就等着吧！！

# 日常生活的自然语言处理应用

## ► 搜索引擎



## ► 输入法



## ► 机器翻译



# 最新进展 ( 续 )

## ▶ 2011年 2月

- ▶ IBM Watson在美国热门的电视智力问答节目《Jeopardy! 》中战胜了两位人类冠军选手。



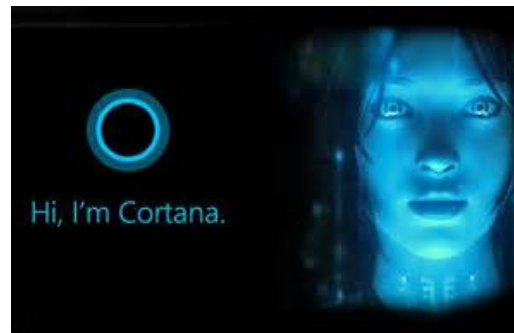
## ▶ 2011年

- ▶ 苹果发布了SIRI



## ▶ 2014年4月

- ▶ 微软首度展示了Cortana



# 最新进展（续）

► 2014年6月8日

- 英国皇家学会举办的图灵测试中，一台由俄罗斯工程师设计的机器（更准确的说是软件）成功通过测试，成为有史以来第一台通过图灵测试的机器。





# 最新进展（续）

## ► 2014年11月

### ► Skype官方近期宣布推出实时语言翻译的预览版

- 在线的语音实时翻译功能支持英语和西班牙语的实时翻译;
- 同时还宣布可以支持40多种语言的文本实时翻译功能。



# 更多应用场景

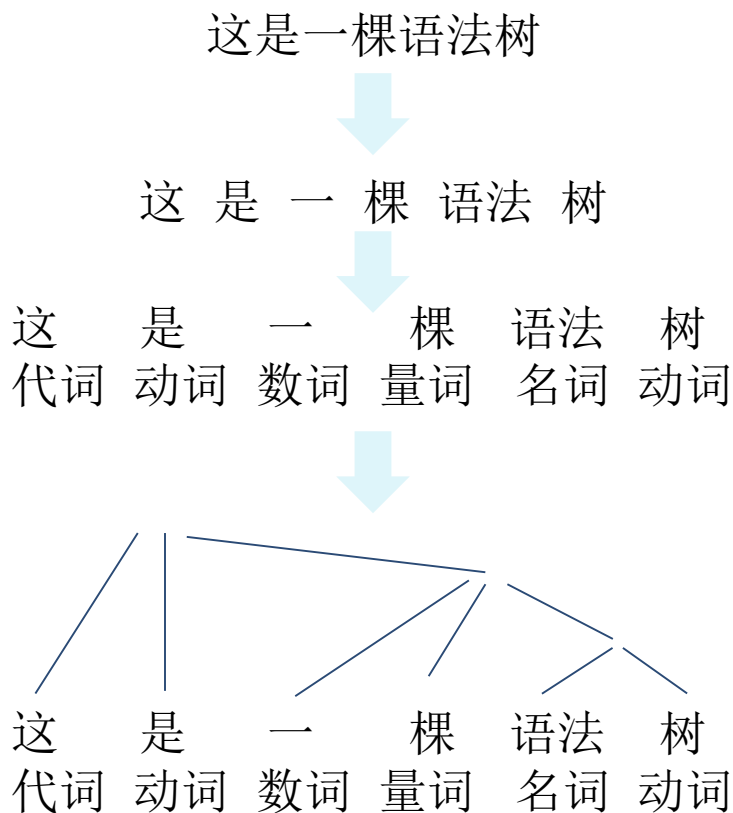


# 当出错时...

民主 ->  江泽民 主席  
(注：早期google的搜索结果)



# 理想中的自然语言处理流程



分词

词性标注

句法分析

语义分析

应用

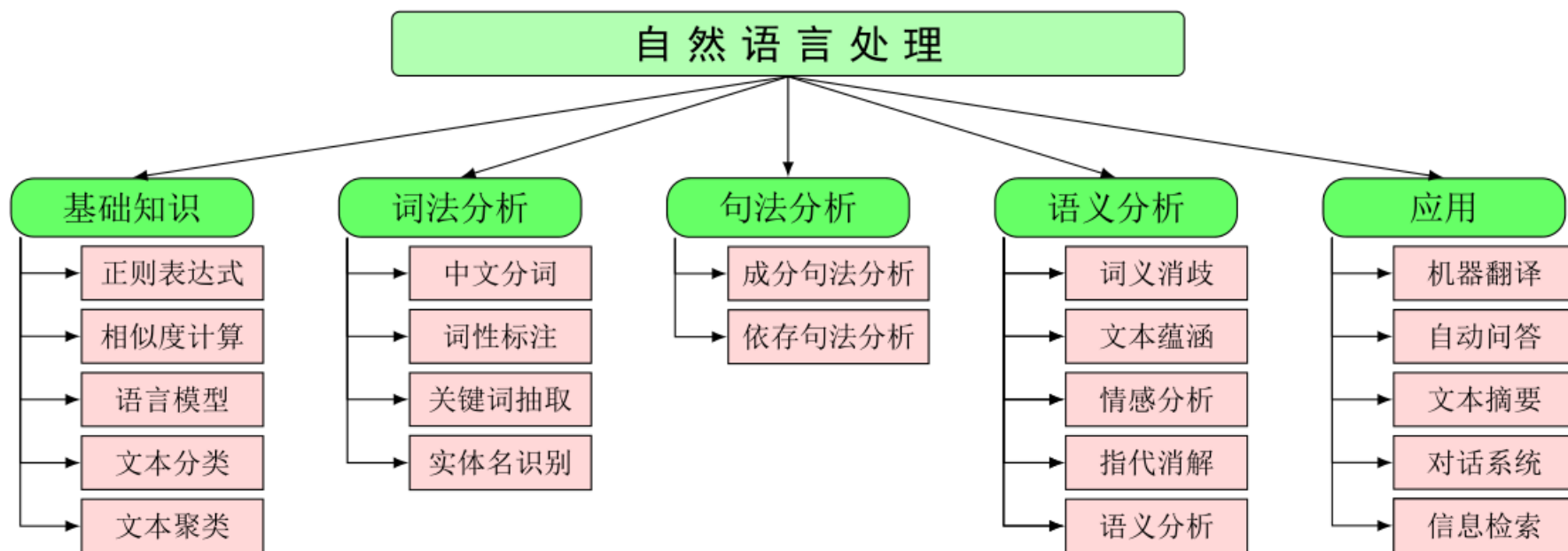
语义分析  
机器翻译  
自动问答  
情感分析  
... ..

{ 这, 是, 语法树 }

知识库



# 主要任务



# 实际流程：End-to-End

我喜欢读书。

我讨厌读书。



分类模型



模型表示

特征抽取

参数学习

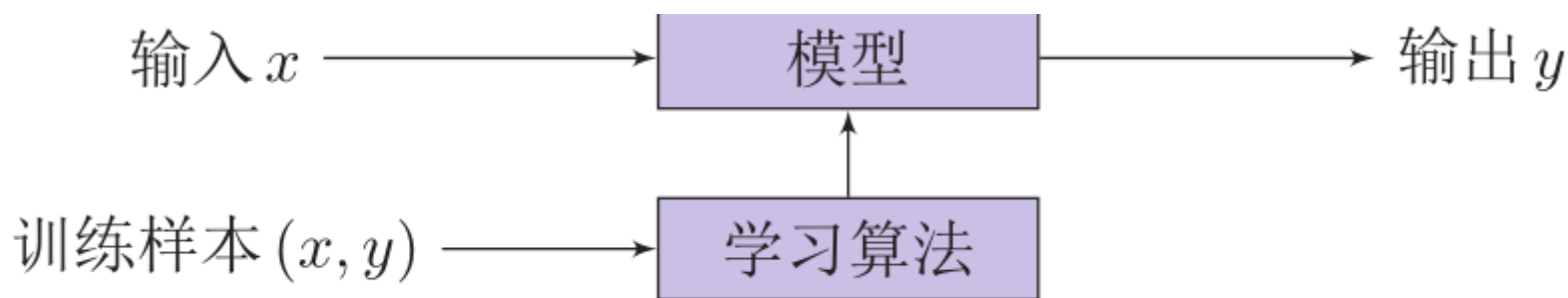
解码算法

情感分析



# 机器学习的基本概念

- ▶ 机器学习主要是研究如何使计算机从给定的数据中学习规律，即从观测数据（样本）中寻找规律，并利用学习到的规律（模型）对未知或无法观测的数据进行预测。
- ▶ 目前，主流的机器学习算法是基于统计的方法，也叫统计机器学习。





# 机器学习类型

---

## ▶ 有监督学习

- ▶ 有监督学习 利用一组已知输入 $x$ 和输出 $y$ 的数据来学习模型的参数，使得模型预测的输出标记和真实标记尽可能的一致。
  - ▶ 回归 如果输出 $y$ 是连续值（实数或连续整数）， $f(x)$ 的输出也是连续值。
  - ▶ 分类 如果输出 $y$ 是离散的类别标记（符号），就是分类问题。





# 机器学习类型

---

## ▶ 无监督学习

- ▶ 用来学习的数据不包含输出目标，需要学习算法自动学习到一些有价值的信息。
- ▶ 一个典型的无监督学习问题就是聚类。



# 机器学习类型

---

## ► 增强学习

- 增强学习也叫强化学习，强调如何基于环境做出一系列的动作，以取得最大化的累积收益。每做出一个动作，并不一定立刻得到收益。
- 增强学习和有监督学习的不同在于增强学习不需要显式地以输入/输出对的方式给出训练样本，是一种在线的学习机制。



# 机器学习

- 训练数据:  $(x_i, y_i), 1 \leq i \leq m$
- 模型:
  - 线性方法:  $y = f(x) = w^T x + b$
  - 广义线性方法:  $y = f(x) = w^T \phi(x) + b$
  - 非线性方法: 神经网络
- 优化
  - 损失函数:  
 $L(y, f(x)) \rightarrow \text{最小化}$
  - 经验风险最小化  
 $Q(\theta) = \frac{1}{m} \cdot \sum_{i=1}^m L(y_i, f(x_i, \theta)) \rightarrow \text{最小化}$
  - 正则化:  $\|\theta\|^2$
- 优化目标函数:  $Q(\theta) + \lambda \|\theta\|^2$

# 机器学习的基本概念

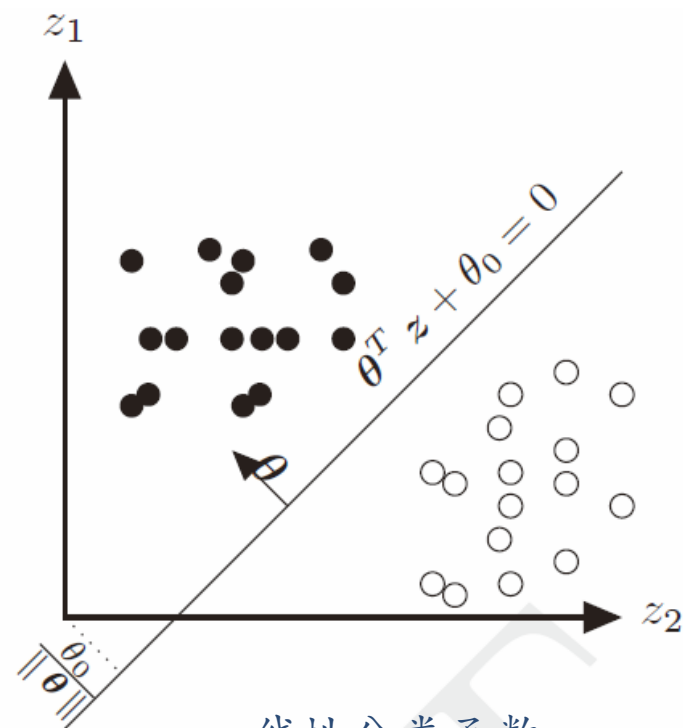
## ► 分类问题

### ► 两类

$$\hat{y} = \begin{cases} +1 & \text{if } f(\mathbf{z}) > 0 \\ -1 & \text{if } f(\mathbf{z}) < 0 \end{cases}$$

### ► 多类

$$\hat{y} = \arg \max_{c=1}^C f_c(\mathbf{z})$$



线性分类函数

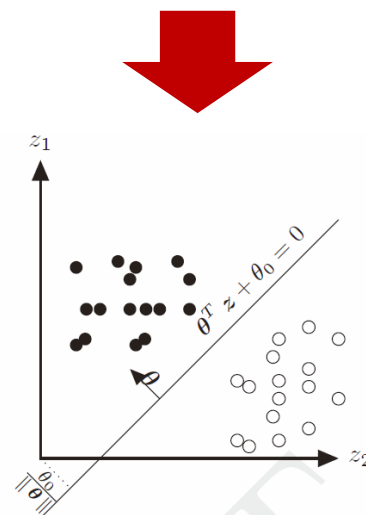
# 文本分类

根据文本内容来判断文本的相应类别

$D_1$ : “我喜欢读书”

$D_2$ : “我讨厌读书”

	我	喜欢	讨厌	读书
$D_1$	1	1	0	1
$D_2$	1	0	1	1



+

-

# 换个角度看中文分词

自	:	然	:	语	:	言	:	处	:	理
0		1		0		1		0		





窗口大小	样本 $x$	类别标签 $y$
2	“自:然”	0
	“然:语”	1
	“语:言”	0
4	“自然:语言”	1
	“然语:言处”	0
	“语言:处理”	1



单字符特征	$x_{-2}y_0, x_{-1}y_0, x_0y_0, x_1y_0, x_2y_0^a$
双字符特征	$x_{-1}x_0y_0, x_0x_1y_0, x_{-1}x_1y_0,$
三字符特征	$x_{-1}x_0x_1y_0$
马氏链特征	$y_{-1}y_0$

1/0

[000010001000100011001]

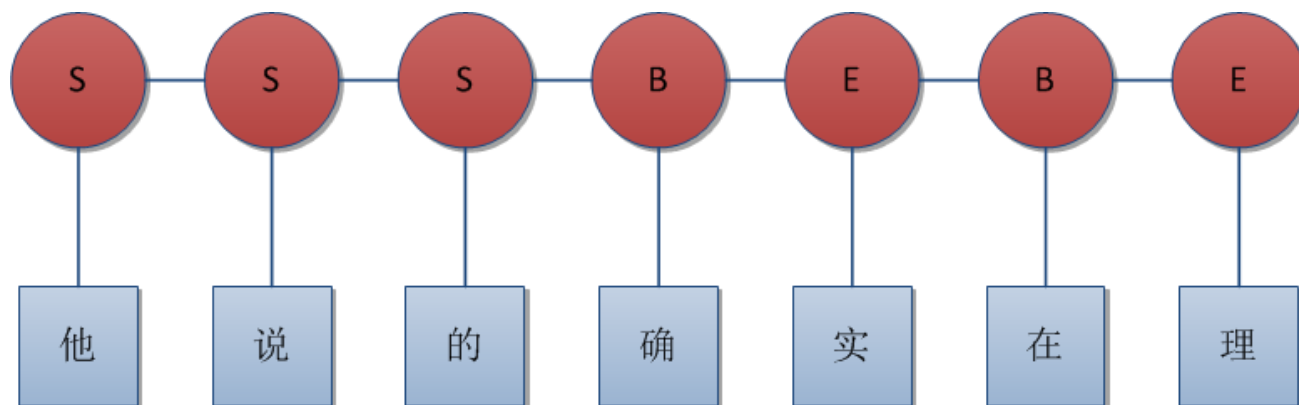



# 换个角度看中文分词

他 / 说 / 的 / 确实 / 在理



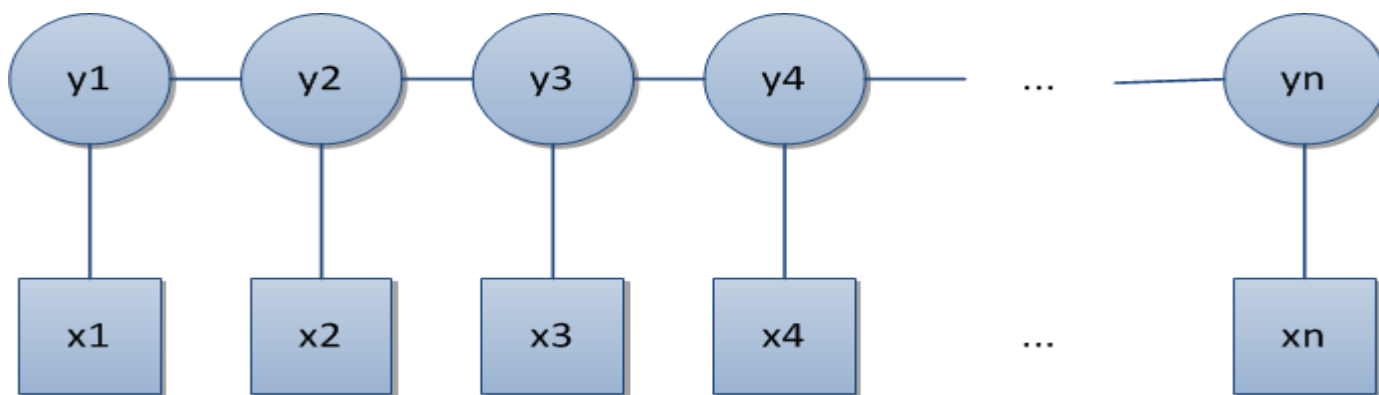
序列标注：结构化学习问题



B: 词的开始字符  
M: 词的中间字符  
E: 词的结尾字符  
S: 单字符词

# 结构化学习

- 在结构化学习中，预测不再局限在一个数，而可以是复杂的结构化对象，比如一个标签序列，或是分析树等。



$$\bar{Y} = \operatorname{argmax}_Y f(\varphi(X, Y), W)$$

C.M. Bishop, Pattern recognition and machine learning, Springer New York, 2006.





# 结构化学习的三个基本问题

---

## ► 模型表示

- 定义一个模型，将自然语言处理任务转换为结构化学习问题。

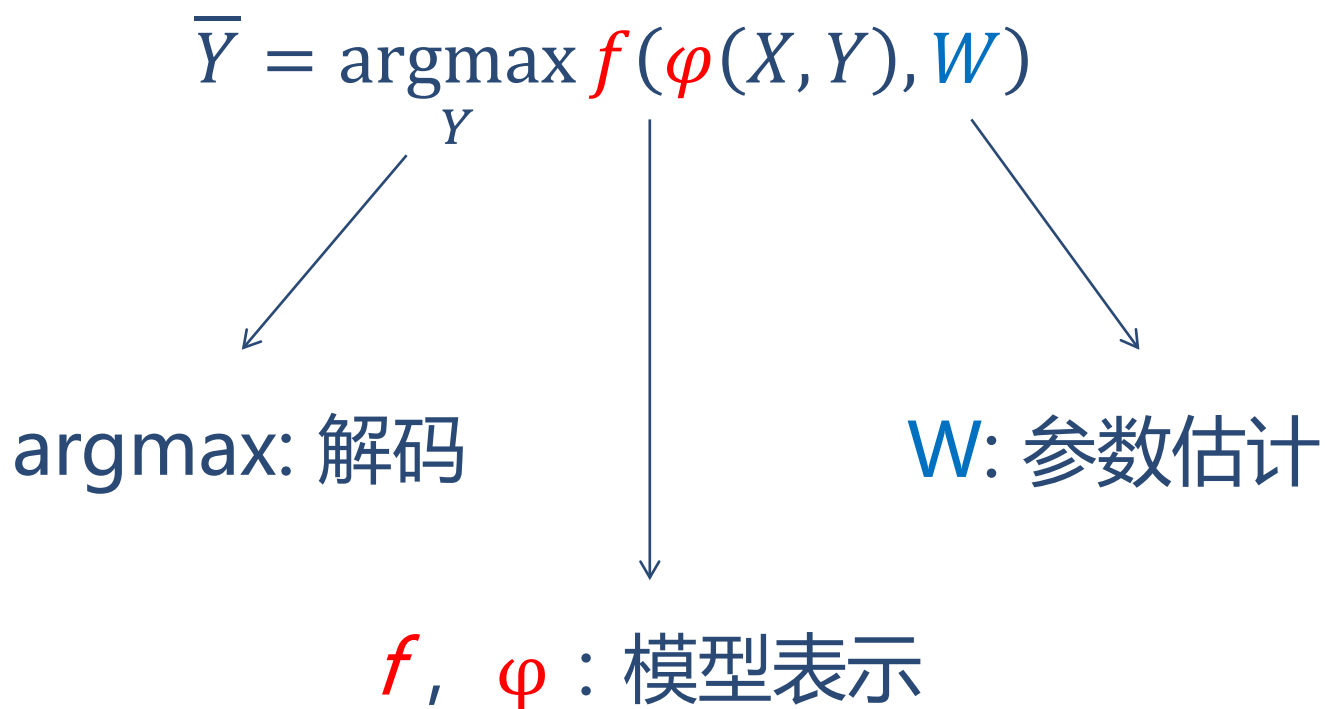
## ► 解码问题

- 给定一个模型，计算最可能的解。

## ► 参数估计

- 给定训练语料，学习模型参数。

# 结构化学习形式化表示





# 基准测试（以FNLP为例）

## ▶ 测试环境

- ▶ CPU: Intel(R) Xeon(R) CPU E5430 @ 2.66GHz
- ▶ 内存: 32G
- ▶ 操作系统: Debian-6.0.1a-amd64
- ▶ java环境: java 1.6.0\_18, OpenJDK

版本	测试集	准确率	万字数/秒
分词	CTB 7.0	95.67%	9.6
词性标注	CTB 7.0	89.29%	0.8
依存句法分析	CTB 7.0	85.9% (UAS)	9840.3 (词/秒)

<https://github.com/FudanNLP/fnlp>



# 预备知识

---

## ► 理论

- 数据结构
- 概率论
- 线性代数
- 机器学习

## ► 实践

- <https://github.com/FudanNLP/fnlp>
  - Java
  - GitHub
  - Maven
  - Eclipse



# 谢 谢