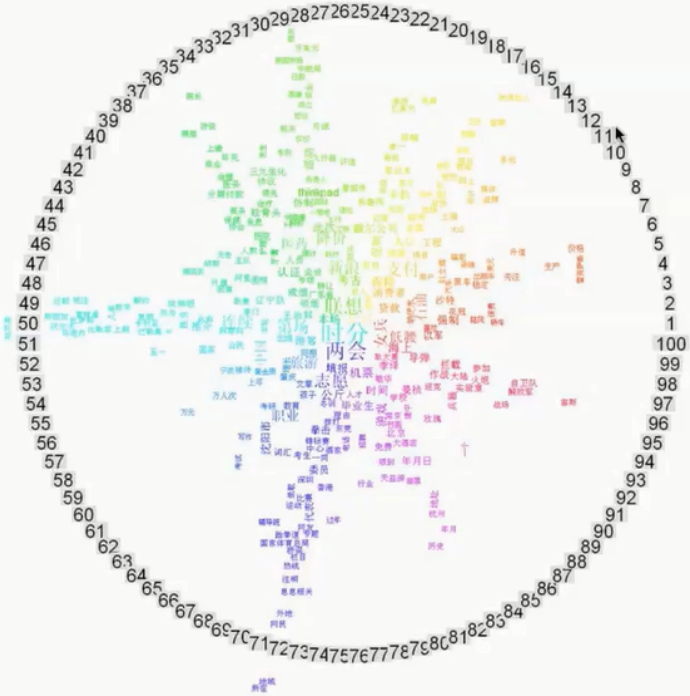
1、自然语言处理概述

NLP的研究内容

- 信息检索
- 机器翻译
- 文档分类
- 问答系统
- 信息过滤
- 自动文摘
- 信息抽取
- 文本挖掘
- 舆情分析
- 机器写作
- 文稿机器校对
- OCR或语音识别



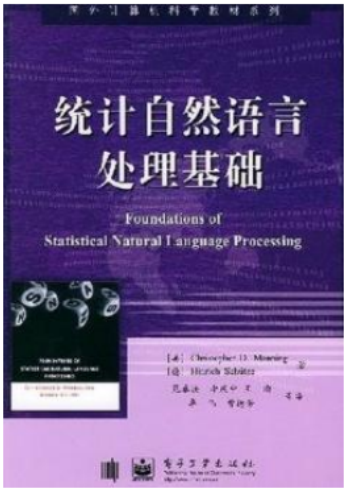
DATAGURU专业数据分析社区

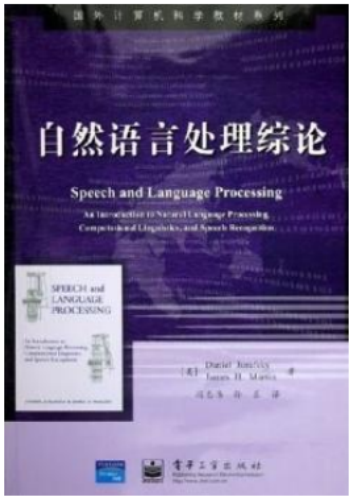
机器读心术之自然语言处理 讲师 黄志洪

（学习）NLP的困难

- 场景的困难：语言的多样性，多变性，歧义性
- 学习的困难：艰难的数学模型（概率图模型：隐马尔科夫过程HMM，最大熵模型，条件随机场CRF等），有人戏称深度学习和NLP是数据科学家的标配（都很难理解）
- 语料的困难：什么是语料？语料有什么作用？如何获得语料？

参考资料：





重点1：形式语言

一般地，描述一种语言可以有三种途径：

（1）穷举法：

（2）文法描述（形式语言）：语言中的每个句子用严格定义的规则来构造，利用规则生成语言中合法的句子。

（3）自动机法：通过对输入的句子进行合法检验，区别哪些是语言中的句子，利用规则生成语言中的句子。

形式语法是一个四元组$G=(N,\Sigma,P,S)$，N是非终结符的有限集合；$\sum$是终结符的有限集合；$V=N\cup\Sigma$总词汇表；P是重写规则(或称之为语法范畴)的有限集合：$P=\{a\to b\}$,其中，a，b是由V中元素构成的串，但是，a中至少应含有一个非终结符号：S称为句子符或初始符。

在不特别强调推导的直接性时，“直接推导”可以简称为**推导**，有时也称推导为**派生**。与之相对应，在不特别强调归约的直接性时，“直接归纳”可以简称为**归约**。

读者不难看出，对任意的$x,y\in\sum^{+}$，要使用一个语法范围D代表的集合为$\{x^{n}y^{n},n\geq 0\}$，可用产生式组$\{D\to xy|xDy\}$

进而，对任意的$x,y\in\sum^{+}$，要使一个语法范畴D代表的集合为$\{x^{n}|n\geq 0\}$，可用产生式组$\{D\to\varepsilon|xD\}$

句子ω是从S开始，在G中可以推导出来的终极符号行，它不含语法变量。

句型α是从S开始，在G中可心推导出来的符号行，它可能含有语法变量。所以，句子一定是句型，但句型不一定是句子。

阅读作业 1 宗成庆《统计自然语言处理》第2章“预备知识”，《形式语言与自动机》第1章1.2节，打好知识基础，后面会用到 2 宗成庆《统计自然语言处理》第4章“语料库与语言知识库”，按照所给的资源链接观摩一下资源的状况

书面作业 《形式语言与自动机》一书第2章习题：第5题，第7(1)题，第8(3)(6)(8)题

5、设方法G的产生式集如下，试给出句子abeebbeeba的推导。你能给出句子abeebbeeb的归约吗？如果能，请给出它的一个归约；如果不能，请说明为什么？

S->aAa|bAb|e

A->SS

bB->bAb

bC->bc

B->bAbS

答：

$$\begin{aligned}S&\Rightarrow aAa\\&\Rightarrow aSSa\\&\Rightarrow abAbSa\\&\Rightarrow abSSbSa\\&\Rightarrow abeebSa\\&\Rightarrow abeebbAba\\&\Rightarrow abeebbSSba\\&\Rightarrow abeebbeeba\end{aligned}$$