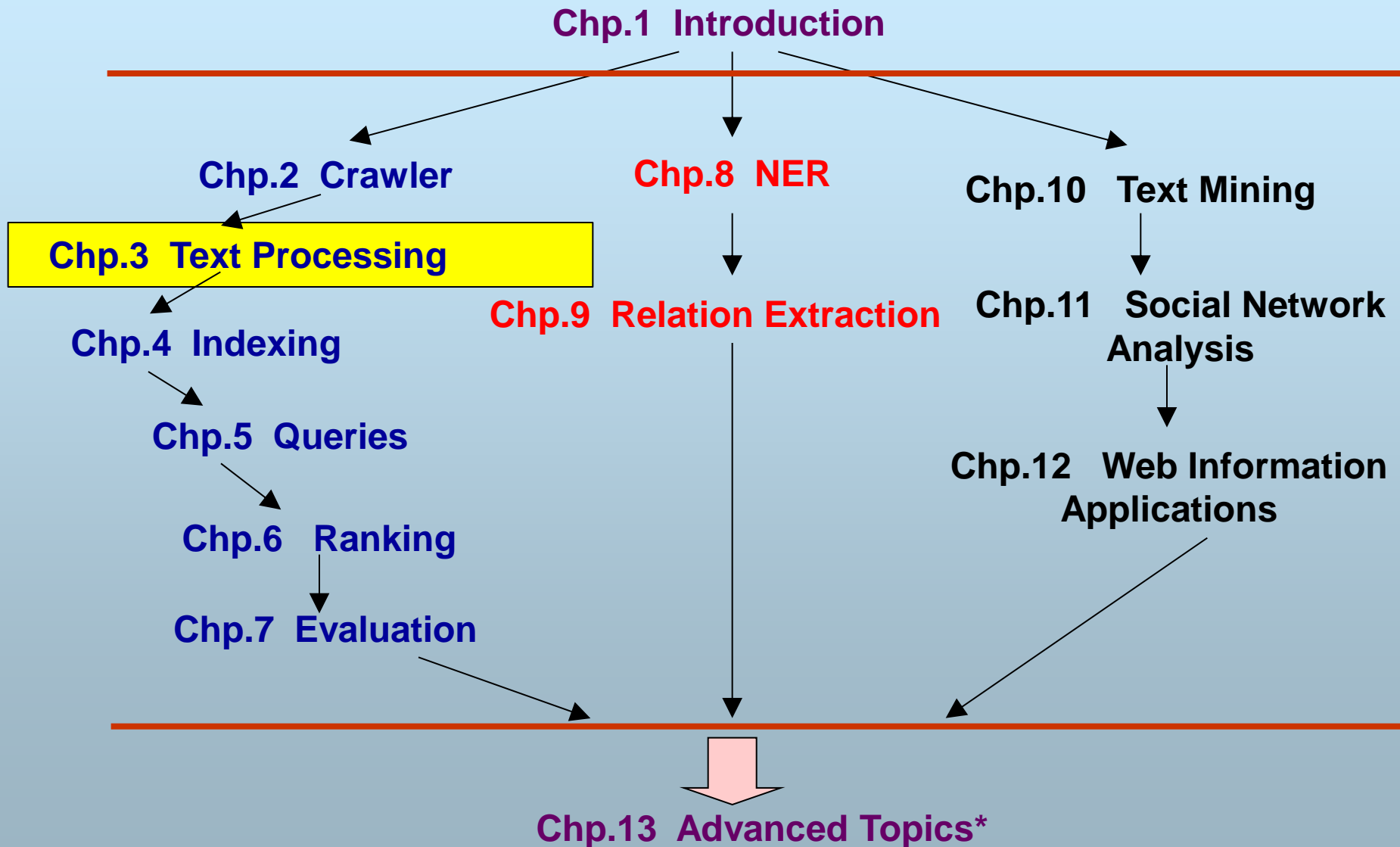


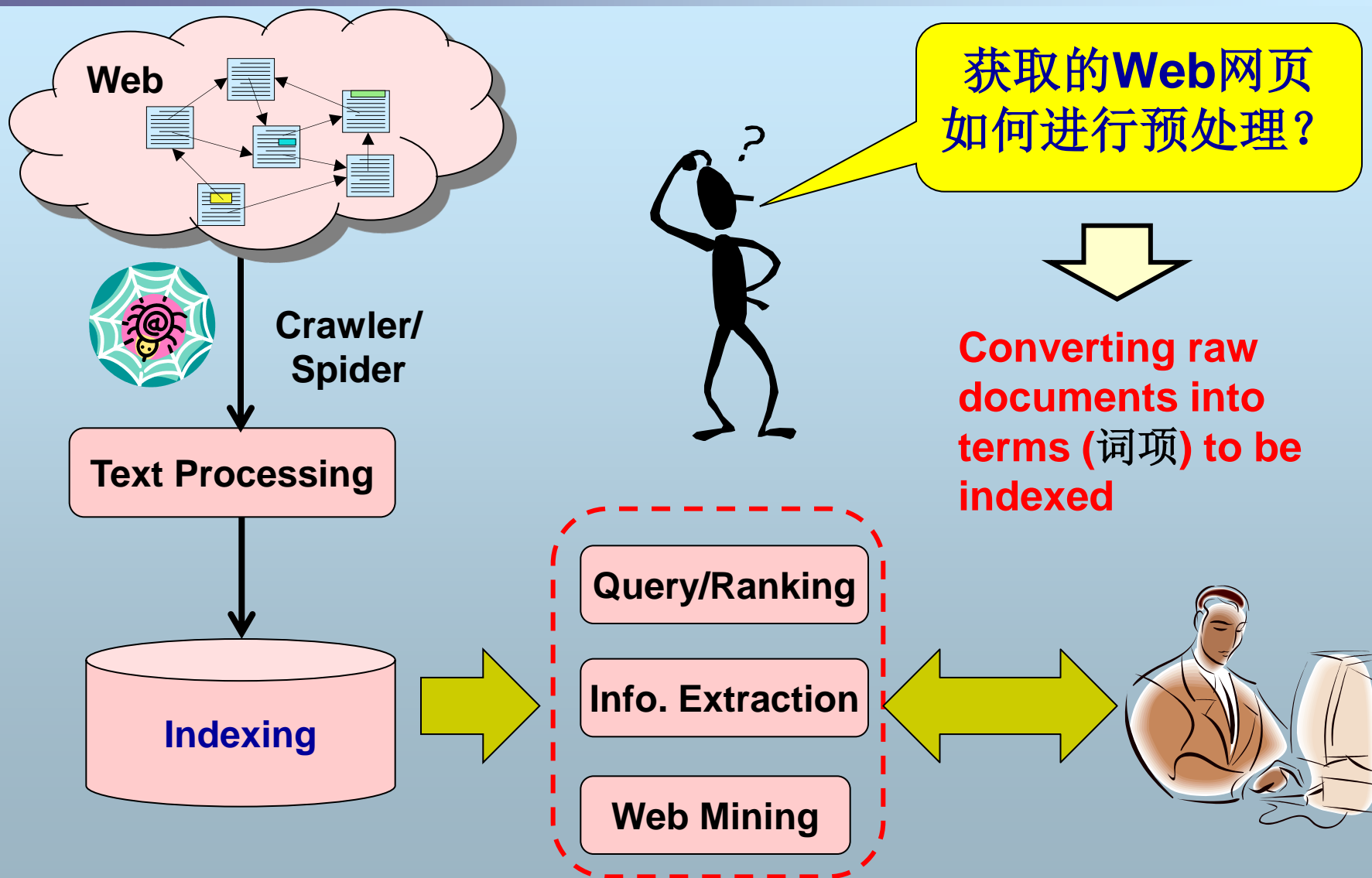
Text Processing



课程知识结构



本章讨论的问题



A Document

中国科学技术大学，^[2]标准简称为中国科大，常用简称科大或USTC，是中国大陆的一所公立研究型大学，^[3]学校主体位于安徽省合肥市。

中国科学技术大学隶属于中国科学院，是全国唯一由中国科学院直属管理的全国重点大学。本科生生源和培养质量一直在全国高校中名列前茅。为中国首批7所“211工程”重点建设的大学和首批9所“985工程”重点建设的大学之一^[4]；是国家“111计划”和“珠峰计划”重点建设的研究型大学；也是“2011计划”中“量子信息与量子科技前沿协同创新中心”的主要协同单位之一。学校在国际上也享有一定声誉，东亚研究型大学协会和环太平洋大学联盟的成员。是九校联盟（C9）和长三角高校合作联盟的重要成员。中国大学校长联谊会成员。中国科学技术大学微尺度物质科学国家实验室入选海外创新人才基地。英国《泰晤士报高等教育副刊》公布该报2010年世界大学排行榜，中国科学技术大学名列全球第49位，中国大陆第二位，同中国内地北京大学，清华大学共有3所高校进入世界百强。英国《泰晤士报高等教育副刊》发布2011~2012世界大学排行榜，中国科学技术大学排名第192位，次于北京大学和清华大学，位居中国大陆第三。^[5]办学目标定位于“质量优异、特色鲜明、规模适度、结构合理的一流研究型大学”。

From: <http://zh.wikipedia.org>

Text Processing

- Basic component in IR systems (not only for Web search).
- Also known as **Document Processing**
 - Converting raw documents into terms (词项) to be indexed
 - Enabling the matching of terms in the query to those in the documents.
- Document processing and query parsing are connected.

本章主要内容

- 分词
- 去除停用词
- 规范化

一、分词

■ Segmentation / Tokenization

- 将文档的字符串序列变成词序列

■ 英文词语空格区分

- “University of Science and Technology of China”

■ 汉语、日语等无空格区分，分词困难

- “中国科学技术大学”
- “ニューヨーク大学”

相当于：UniversityofScienceandTechnologyofChina

一、分词

- **语素**是最小的语音语义结合体，是最小的语言单位。
 - “字”：简单高效,国家标准**GB2312-80** 中定义的常用汉字为**6763**个。表示能力比较差，不能独立地完整地表达语义信息。
- 词是代表一定的意义，具有固定的语音形式，可以独立运用的最小的语言单位。
 - “词”：表示能力较强，但汉字的词的个数在**10**万个以上，面临复杂的分词问题

一、分词

■ 中文分词的挑战

- 英语——词的集合 vs. 汉语——字的集合
- 汉字之间存在着不同的组词方式
 - ◆ 如“**发展中国家兔的饲养**”一句，现有的汉语词就可能导致有两组分隔结果：**发展中国家/兔/的/饲养**，**发展/中国/家兔/的/饲养**。
- 汉语虚词众多，而且绝大多数汉字当与不同的汉字组词时，其词可能为关键词，也可能为停用词
 - ◆ 如，“非”与“洲”、“常”分别组成不同意义的词“**非洲**”（关键词）、“**非常**”（停用词）。
- 分词歧义
- 新词的频繁出现也给汉语分词增添了难度
 - ◆ 未登录词

分词歧义

■ 交集型歧义（交叉歧义）

- 如果**AB**和**BC**都是词典中的词，那么如果待切分字符串中包含“**ABC**”这个子串，就必然会造成两种可能的切分：“**AB/C/**”和“**A/BC/**”。这种类型的歧义就是交叉歧义。
- 比如“网球场”就可能造成交叉歧义
 - ◆ 网球 / 场 /
 - ◆ 网 / 球场 /

■ 组合型歧义（组合歧义）

- 如果**AB**和**A**、**B**都是词典中的词，那么如果待切分字符串中包含“**AB**”这个子串，就必然会造成两种可能的切分：“**AB/**”和“**A/ B/**”。这种类型的歧义就是组合歧义。
- 比如“个人”就可能造成组合歧义
 - ◆ 我 / 个人
 - ◆ 三 / 个 / 人 /

未登录词

- 未登录词即未包括在分词词表中但必须切分出来的词，包括各类专名（人名、地名、企业字号、商标号等）和某些术语、缩略词、新词等等
 - “于大海发明爱尔肤护肤液”需要切分成“于大海/发明/爱尔肤/护肤液”，并需要识别出“于大海”是人名，“爱尔肤”是商标名，“护肤液”是术语名词。
 - 如“斯普林菲尔德是伊利诺州首府”、“丹增嘉措70多岁了”，其中的美国地名、藏族人名都需识别。
 - 机构名和商品品牌名：“希望电脑”、“国际乒联”、“非常可乐”。
 - 专业领域的大量术语：“线性回归”、“A*算法”。
 - 新词语，缩略语：“粉丝”、“E时代”、“坑爹”。

一、分词

■ 常用的分词方法

- 基于字符串匹配的方法
- 基于理解的方法
- 基于统计的方法

1、基于字符串匹配的分词方法

- 又叫**机械分词**方法,它按照一定的策略将待分析的汉字串与一个“充分大的”机器词典中的词条进行匹配,若在词典中找到某个字符串,则匹配成功（识别出一个词）
- 主要特征：**有词典**
- 分类：
 - 按照扫描方向的不同：**正向匹配和逆向匹配**
 - 按照不同长度优先匹配的情况：**最大匹配和最小匹配**

1、基于字符串匹配的分词方法

■ 常用的机械分词方法

- 正向最大匹配分词 (**FMM**)
- 反向最大匹配分词 (**BMM, RMM**)
- 双向最大匹配分词 (**BM: FMM+RMM**)
- 最少切分分词

正向最大匹配分词

- Forward Maximum Matching method, FMM
- 从左至右尽可能查找最长的词，直到当前字符与已经处理的字符串不构成词，输出已经识别的词，并从识别出来的词后面接着查找词。
- 分词速度比较快
- 但分词错误率比较高，错误率约 $1 / 169$

例1: “使用户满意”



使用 / 户 / 满意

例2: “只有在市场中国有企业才能发展”

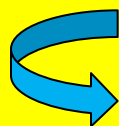


只有 / 在 / 市场 / 中国 / 有 / 企业 / 才能 / 发展

反向最大匹配分词

- Backward Maximum Matching method, BMM
 - 也称Reverse Maximum Matching method, RMM
- 从右至左尽可能查找最长的词
- 实验表明分词效率优于正向最大匹配分词方法
 - 错误率 1 / 245

例：“使用户满意”



使 / 用户 / 满意

例：“只有在中国有企业才能发展”



只有 / 在 / 市场 / 中 / 国有 / 企业 / 才能 / 发展

双向最大匹配分词

- Bi-direction Matching method, BM
- 比较FMM法与BMM法的切分结果，从而决定正确的切分
- 可以识别出分词中的交叉歧义

例：“南京市长江大桥” → 南京市 / 长江大桥 (BMM)

↓
南京市长 / 江 / 大桥 (FMM)

↓
南京市，长江大桥，南京市长，江，大桥 (union)

or

南京市，长江大桥 (minimum)

机械分词方法一般模型

- 对于机械分词方法，可以建立一个一般的模型，形式地表示为：

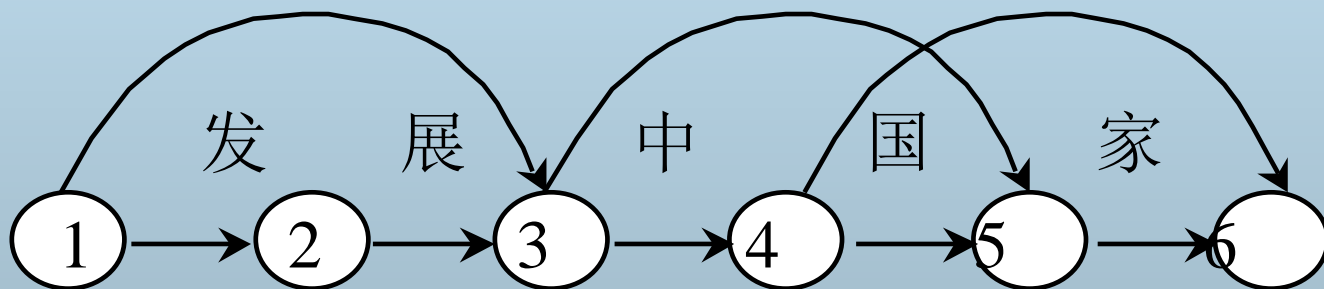
- **ASM(d, a, m)**，即**Automatic Segmentation Model**。其中，
 - ◆ **d**：匹配方向，**+1**表示正向，**-1**表示逆向；
 - ◆ **a**：每次匹配失败后增加/减少字符串长度（字符数），**+1**为增字，**-1**为减字；
 - ◆ **m**：最大/最小匹配标识，**+1**为最大匹配，**-1**为最小匹配。

- 例如：

- **ASM (+, -, +)** 就是正向减字最大匹配（即**FMM**方法）。
- **ASM (-, -, +)** 就是逆向减字最大匹配法（即**BMM**方法）。
- 对于现代汉语来说，**m=+1**是实用的方法。

最少切分分词方法

- 使句子中切出的词数目最少
- 等价于在有向图中搜索最短路径问题



2、基于理解的分词方法

- 这种分词方法是通过让计算机模拟人对句子的理解，达到识别词的效果；
- 基本思想
 - 在分词的同时进行句法、语义分析，利用句法信息和语义信息来处理歧义现象。
- 需要使用大量的语言知识和信息。由于汉语语言知识的复杂性，难以将各种语言信息组织成机器可直接读取的形式，因此目前基于理解的分词系统还处在试验阶段

3、基于统计的分词方法

- 字与字相邻共现的频率或概率能够较好的反映成词的可信度。
- 如果某两个词的组合，在概率统计上出现的几率非常大，那么我们就认为分词正确。
- 例如，“南京市长江大桥”
 - 统计结果表明，“南京市 / 长江大桥”同时出现的概率大于“南京市长 / 江 / 大桥”的概率
 - 则可以认为“南京市 / 长江大桥”是正确分词结果的可能性更大

3、基于统计的分词方法

■ 问题形式化描述

- 令 $c = c_1 \dots c_n$, c 是待分词的句子（字串）
 $w = w_1 \dots w_n$ 是切分的结果。
- 设 $P(w|c)$ 是 c 切分为 w 的某种估计概率。
- w_a, w_b, \dots, w_k 是 c 的所有可能的分词方案。
- 那么，基于统计的分词模型就是找到目的词串 w ，使得 w 满足：

$$P(w|c) = \max\{P(w_a|c), P(w_b|c), \dots, P(w_k|c)\}$$

◆ 即估计概率为最大之词串。

3、基于统计的分词方法

■ 一般过程

- (1) 建立统计语言模型
- (2) 对句子进行分词
- (3) 计算概率最大的分词结果

■ 理论上可以不需要词典，但实际应用中第(2)步可以采用机械分词方法进行分词，以获得候选的分词集合

- 既发挥匹配分词切分速度快、效率高的特点，又利用了无词典分词结合上下文识别生词、自动消除歧义的优点。

3、基于统计的分词方法

■ 常用的统计语言模型

- N元语法模型 (N-gram)
- 隐马尔可夫模型 (HMM)

■ 马尔可夫假设

- 当前状态出现的概率仅同过去有限的历史状态有关，而与其他状态无关。具体到分词任务，就是文本中第N个词出现的概率仅仅依赖于它前面的N-1个词，而与其他词无关
 - ◆ 当前状态只跟(N-1)阶马尔可夫状态有关
 - ◆ N-gram = (N-1)阶马尔可夫模型
 - ◆ N=1: 一元语法模型，最大概率模型
 - ◆ N=2: bigram
 - ◆ N=3: trigram

N-gram模型

■ Word sequences

$$w = w_1 \dots w_n$$

■ Chain rule of probability

$$P(w) = p(w_1)p(w_2 | w_1)p(w_3 | w_1w_2)\dots p(w_n | w_1w_2\dots w_{n-1})$$

■ Bigram approximation

$$\approx p(w_1)p(w_2 | w_1)p(w_3 | w_2)\dots p(w_n | w_{n-1})$$

■ Trigram approximation

$$\approx p(w_1)p(w_2 | w_1)p(w_3 | w_1w_2)\dots p(w_n | w_{n-2}w_{n-1})$$

N-gram模型

■ 概率估计

● 最大似然估计 Maximum Likelihood Estimate

Bigram:
$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

- ◆ $C(w_{n-1}w_n)$ 是词序列 $w_{n-1}w_n$ 在语料库中出现的次数
- ◆ $C(w_{n-1})$ 是词 w_{n-1} 在语料库中出现的次数

N-gram模型

■ 分词过程，以bigram模型为例

- 构造训练语料库，计算 $C(w_1), C(w_2), \dots, C(w_n)$ 以及 $C(w_1w_2), \dots, C(w_{n-1}w_n)$

- 对于每一个可能的分词序列 w ，计算

$$p(w) = p(w_1)p(w_2 | w_1)p(w_3 | w_2) \dots p(w_n | w_{n-1})$$

其中
$$P(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

- 返回 $p(w)$ 最大的分词序列作为结果

举例

假设语料库总词数为**13,748**词

$$C(w_{n-1})$$

我	3437
想	1215
晚上	3256
去	938
吃	213
意大利	1506
菜	459

举例

$$C(w_{n-1}w_n)$$

	我	想	晚上	去	吃	意大利	菜
我	8	1087	0	13	0	0	0
想	3	0	786	0	6	8	6
晚上	3	0	10	860	3	0	12
去	0	0	2	0	19	2	52
吃	2	0	0	0	0	120	1
意大利	19	0	17	0	0	0	0
菜	4	0	0	0	0	1	0

举例

■ $P(\text{“我|想|晚上|去|吃|菜”})$

$= P(\text{我}) * P(\text{想|我}) * P(\text{晚上|想}) * P(\text{去|晚上}) * P(\text{吃|去}) * P(\text{菜|吃})$

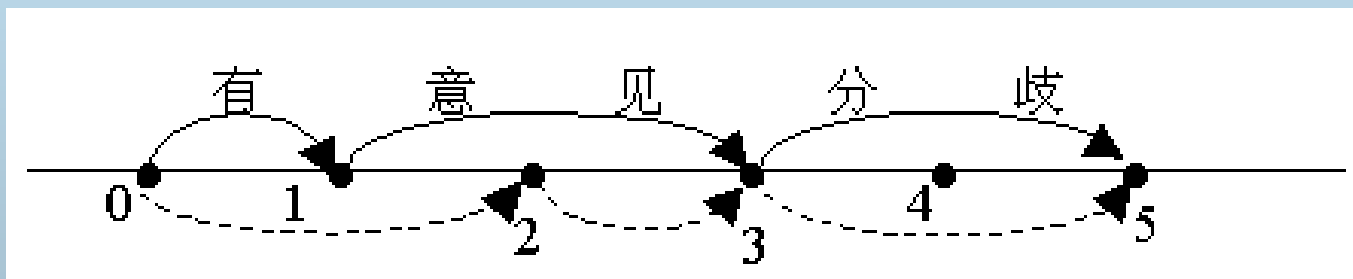
$= 0.25 * 1087/3437 * 786/1215 * 860/3256 * 19/938 * 1/213$

$= 0.00000128$

N-gram

■ N=1时：一元语法模型（最大概率分词）

- 词与词之间是独立的
- 一个待切分的汉字串可能包含多种分词结果
- 将其中概率最大的那个作为分词结果



路径1: 0-1-3-5

路径2: 0-2-3-5

该走哪条路呢？

最大概率分词

- “有意见分歧”
 - W1: 有/ 意见/ 分歧/
 - W2: 有意/ 见/ 分歧/

$$P(W) = P(w_1, w_2, \dots, w_i) \approx P(w_1) \times P(w_2) \times \dots \times P(w_i)$$

独立性假设，一元文法

$$P(w_i) = \frac{w_i \text{在语料库中的出现次数} n}{\text{语料库中的总词数} N}$$

最大概率分词

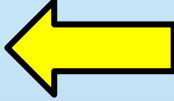
词语	概率
...	...
有	0.0180
有意	0.0005
意见	0.0010
见	0.0002
分歧	0.0001
...	...

$$\begin{aligned}P(W1) &= P(\text{有}) * P(\text{意见}) * P(\text{分歧}) \\ &= 1.8 \times 10^{-9}\end{aligned}$$

$$\begin{aligned}P(W2) &= P(\text{有意}) * P(\text{见}) * P(\text{分歧}) \\ &= 1 \times 10^{-11}\end{aligned}$$

$$P(W1) > P(W2)$$

本章主要内容

- 分词
- 去除停用词 
- 规范化

二、去除停用词

■ 停用词—stopwords

- 在文档中频繁出现的词语
- 与语料库特性有关
 - ◆ 例如，wikipedia语料库中“wiki”是停用词

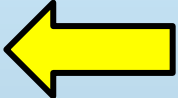
■ 为什么要去除stopwords?

- 重复率很高，会造成索引中的posting list很长，影响查询性能
- 对最后结果的排序没什么贡献

■ 一般可通过维护一个stopwords表来实现

- 如：<http://www.lextek.com/manuals/onix/stopwords2.html>

本章主要内容

- 分词
- 去除停用词
- 规范化 

三、规范化

- 文本处理的主要目标
 - 实现查询词与索引词之间的匹配
- 但用户的查询输入可能出现多种情况
 - 大小写: “New York University” 《=》 “new york university”
 - 缩写: “U.S.A. population” 《=》 “usa population”
 - 标点符号: “What’s the highest mountain?” 《=》 “what’s the highest mountain”
- 规范化的目的就是尽量保证索引的词项符合用户可能的查询输入

三、规范化

■ 规范化需要考虑的因素

● 大小写

- ◆ 全部转换成小写？ JOURNAL->journal

● 标点符号

- ◆ 首尾标点去掉？

● 缩写

- ◆ 转换为统一缩写，例如U.S.A.—>USA

● 词根化

● 拼写错误处理

● 同义词/相关词

1、词根化—Stemming

■ Stemming

- The process of finding the semantic root of a word
- 例如
 - ◆ ran, running → run
 - ◆ universities → university

Porter Stemming

- Most widely used stemming algorithm for English
- A suffix-stripping algorithm
 - 使用一系列后缀变换规则对单词进行变换
- Open source version of the algorithm readily available in many programming languages on the Web.
 - E.g., <http://tartarus.org/~martin/PorterStemmer/>

An algorithm for suffix stripping, by Martin Porter, *Program*, 14(3), 1980.

Porter Stemming

- Step 1a: removes plurals
- Step 1b: removes –ed(ly) or –ing(ly) suffixes
- Step 1c: turns –y to –i
- Step 2: handles double suffixes such as –ization
- Step 3: handles –full, –ness, etc.
- Step 4: handles –ant, –ence, etc.
- Step 5: removes final –e and –ll

2、拼写错误处理

- 查询输入容易出现拼写错误：10%+
- 通常采用拼写错误词典来处理拼写错误
 - 词典
 - Edit distance (编辑距离)

2、拼写错误处理

■ Google中的拼写错误自动纠正

编辑距离

- Given two strings, s and t , the edit distance, or **Levenshtein Distance**, between them is the minimum number of ***edit operations*** required to transform s into t

- 例如

- **Edit-Distance(“kitten” , “sitting”)=3**

- ◆ kitten → sitten (substitution of "s" for "k")
 - ◆ sitten → sittin (substitution of "i" for "e")
 - ◆ sittin → sitting (insertion of "g" at the end).

3、同义词/相关词处理

- 比词根化和拼写错误更难处理
- 通常借助人工维护的知识库
- 例如, **WordNet 3.0**
 - <http://wordnet.princeton.edu/wordnet/>
 - 155K words organized into 117K synsets
 - George Miller, a psychologist, started the project in late 1980's
 - A comprehensive lexical for English
 - Relationships between words are labeled

3、同义词/相关词处理

■ 词之间的关系

- **Synonymy**

- ◆ college ~= university

- **ISA relationship**

- ◆ dreamliner ISA plane

- **Is-Part-Of relationship**

- ◆ Nokia Is-Part-Of Microsoft

- **Antonymy**

- ◆ young vs old

WordNet示例


```
dog, domestic dog, Canis familiaris
=> canine, canid
=> carnivore
=> placental, placental mammal, eutherian, eutherian mammal
=> mammal
=> vertebrate, craniate
=> chordate
=> animal, animate being, beast, brute, creature, fauna
=> ...
```


WordNet应用

■ 文献搜索

● 相关词搜索 Microblog ⇔ Twitter

DBLP FILTER Sign in

 Sort by ☐ relevance ☒ importance ☐ date

Scholar About 28 results (5.57sec)  (1998~2013)

Since Time

Since 2013

Since 2012

Since 2009

Custom range...

Sort By

Sort By Relevance

Sort By Importance

Sort By Date

A	EE	Scholar	A Data Model and Data Structures for Moving Objects Databases. (Luca Forlizzi and Ralf Hartmut G and ü) <i>ACM Conference on Management of Data (sigmod) [2000]</i> Cited by 353
A	EE	Scholar	Scientific Data Repositories: Designing for a Moving Target. (Etzard Stolte and Christoph von Praun and Gustavo Alonso) <i>ACM Conference on Management of Data (sigmod) [2003]</i> Cited by 43
A	EE	Scholar	A Data Model for Moving Objects Supporting Aggregation. (Bart Kuijpers and Alejandro A. Vaisman) <i>IEEE International Conference on Data Engineering (ICDE) [2007]</i> Cited by 20
B	EE	Scholar	Spatio-Temporal Data Types: An Approach to Modeling and Querying Moving Objects in Databases. (Martin Erwig and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [1999]</i> Cited by 367
B	EE	Scholar	A generic data model for moving objects. (Jianqiu Xu and Ralf Hartmut G and ü) <i>GeoInformatica (GeoInformatica) [2013]</i>
B	EE	Scholar	An Object-Field Perspective Data Model for Moving Geographic Phenomena. (Kyoung-Sook Kim and Yasushi Kiyoki) <i>Database Systems for Advanced Applications (DASFAA) [2010]</i>
C	EE	Scholar	Place: A Distributed Spatio-Temporal Data Stream Management System for Moving Objects. (Xiaopeng Xiong and Hicham G. Elmongui and Xiaoyong Chai) <i>International Conference on Mobile Data Management (MDM) [2007]</i> Cited by 18
C	EE	Scholar	An analytic solution to the alibi query in the space-time prisms model for moving object data. (Bart Kuijpers and Rafael Grimson and Walied Othman) <i>International Journal of Geographical Information Science (IJGIS) [2011]</i> Cited by 3
C	EE	Scholar	A Scaleless Data Model for Direct and Progressive Spatial Query Processing. (Sai Sun and Sham Prasher and Xiaofang Zhou) <i>International Conference on Conceptual Modeling (ER) [2004]</i> Cited by 2
C	EE	Scholar	Efficient Strip-Mode SAR Raw-Data Simulation of Fixed and Moving Targets. (Ozan Dogan and Mesut Kartal) <i>IEEE Geoscience and Remote Sensing Letters (LGRS) [2011]</i>
			Computational data modeling for network-constrained moving objects (I aurvnas Sneicvs and Christian S. Jensen and Augustas Kliavs) <i>GIS</i>

Jiang Du, Peiquan Jin, et al. DBLP-Filter: Effective Search on the DBLP Bibliography, WWW'14

本章小结

■ Text Processing

- 分词
- 去除停用词
- 规范化