

# 现代信息检索

## Modern Information Retrieval

### 第3讲 词典及容错式检索

Dictionary and tolerant retrieval

授课人：王斌

<http://ir.ict.ac.cn/~wangbin>

# 提纲

- ① 上一讲回顾
- ② 词典
- ③ 通配查询
- ④ 编辑距离
- ⑤ 拼写校正
- ⑥ **Soundex**

# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

# 上一讲内容

---

- 文档
- 词条/词项
- 基于跳表指针的合并
- 短语查询的处理(双词索引和位置索引)

# 文档

---

- 索引的基本单位
  - 与文件不是一回事，严格地说，一篇文档可能包含多个文件，也可能一个文件包含多篇文档
  - 依赖于具体应用
    - 句子级检索：一个句子为一篇文档
    - 段落级检索：一段文本为一篇文档
    - .....

# 词类(type)/词条(token)的区别

- 词条(Token) – 词或者词项在文档中出现的实例，出现多次算多个词条
- 词类(Type) – 多个词条构成的等价类(equivalence class)集合
- *In June, the dog likes to chase the cat in the barn.*
- 12 个词条, 9个词类
- 词类经过一些处理(去除停用词、归一化)之后，最后用于索引的称为为词项

# 词条化中考虑的问题

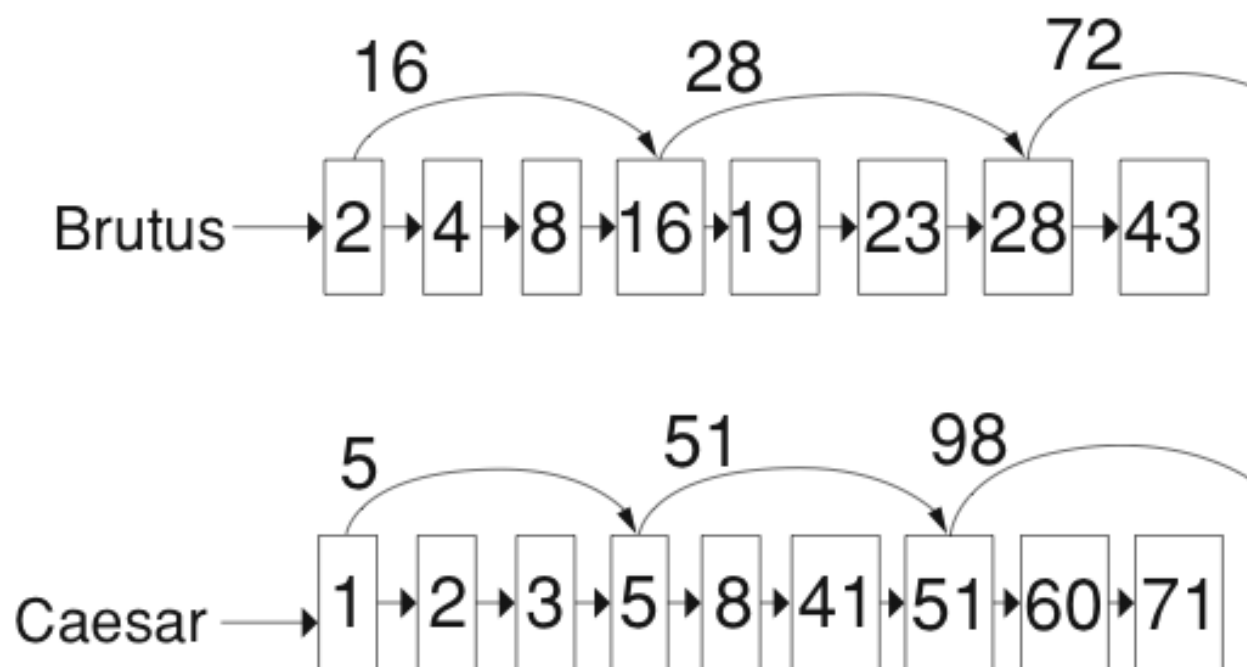
- 词之间的边界是什么？空格？撇号还是连接符？
- 上述边界不一定是真正的边界（比如，中文）
- 另外荷兰语、德语、瑞典语复合词中间没有空格  
(*Lebensversicherungsgesellschaftsangestellter*)

# 词项归一化中的问题

- 词项实际上是一系列词条组成的等价类
- 如何定义等价类？
  - 数字 (3/20/91 vs. 20/3/91)
  - 大小写问题
  - 词干还原，Porter工具
- 形态分析: 屈折 vs. 派生
- 其他语言中词项归一化的问题
  - 比英语中形态更复杂
  - 芬兰语: 单个动词可能有12,000 个不同的形式different forms
  - 重音符号、元音变音问题（umlauts，由于一个音被另一个音词化而导致的变化，尤其是元音的变化）



# 跳表指针



# 位置(信息)索引

- 在无位置信息索引中，每条倒排记录只是一个docID
- 在位置信息索引中，每条倒排记录是一个docID加上一个位置信息表
- 一个查询的例子：“ $to_1 be_2 or_3 not_4 to_5 be_6$ ”

TO, 993427:

1: <7, 18, 33, 72, 86, 231>;

2: <1, 17, 74, 222, 255>;

4: <8, 16, 190, 429, 433>;

5: <363, 367>;

7: <13, 23, 191>; ... >

BE, 178239:

1: <17, 25>;

4: <17, 191, 291, 430, 434>;

5: <14, 19, 101>; ... >

第4篇文档能够与查询匹配！

# 位置信息索引

- 基于位置信息索引，能够处理短语查询(phrase query)，也能处理邻近式查询(proximity query)

# 本讲内容

- 词典的数据结构：访问效率和支持查找的方式
- 容错式检索(Tolerant retrieval): 如果查询词项和文档词项不能精确匹配时如何处理？
  - 通配查询：包含通配符\*的查询
  - 拼写校正：查询中存在错误时的处理

# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

# 倒排索引



# 词典

- 词典是指存储词项词汇表的数据结构
- 词项词汇表(Term vocabulary): 指的是具体数据
- 词典(Dictionary): 指的是数据结构

# 采用定长数组的词典结构

- 对每个词项，需要存储：
  - 文档频率
  - 指向倒排记录表的指针
  - ...
- 暂定每条词项的上述信息均采用定长的方式存储
- 假定所有词项的信息采用数组存储



# 采用定长数组的词典结构

| 词项     | 文档频率    | 指向倒排记录表的指针 |
|--------|---------|------------|
| a      | 656 265 | →          |
| aachen | 65      | →          |
| ...    | ...     | ...        |
| zulu   | 221     | →          |

空间消耗： 20字节   4字节   4字节

# 词项定位(即查词典)

- 输入“信息”，如何在词典中快速找到这个词？18
  - 很多词典应用中的基本问题。19
  - 以下介绍支持快速查找的词典数据结构。20
- 21
- 22



# 用于词项定位的数据结构

- 主要有两种数据结构：哈希表和树
- 有些IR系统用哈希表，有些系统用树结构
- 采用哈希表或树的准则：
  - 词项数目是否固定或者说词项数目是否持续增长？
  - 词项的相对访问频率如何？？
  - 词项的数目有多少？

# 哈希表

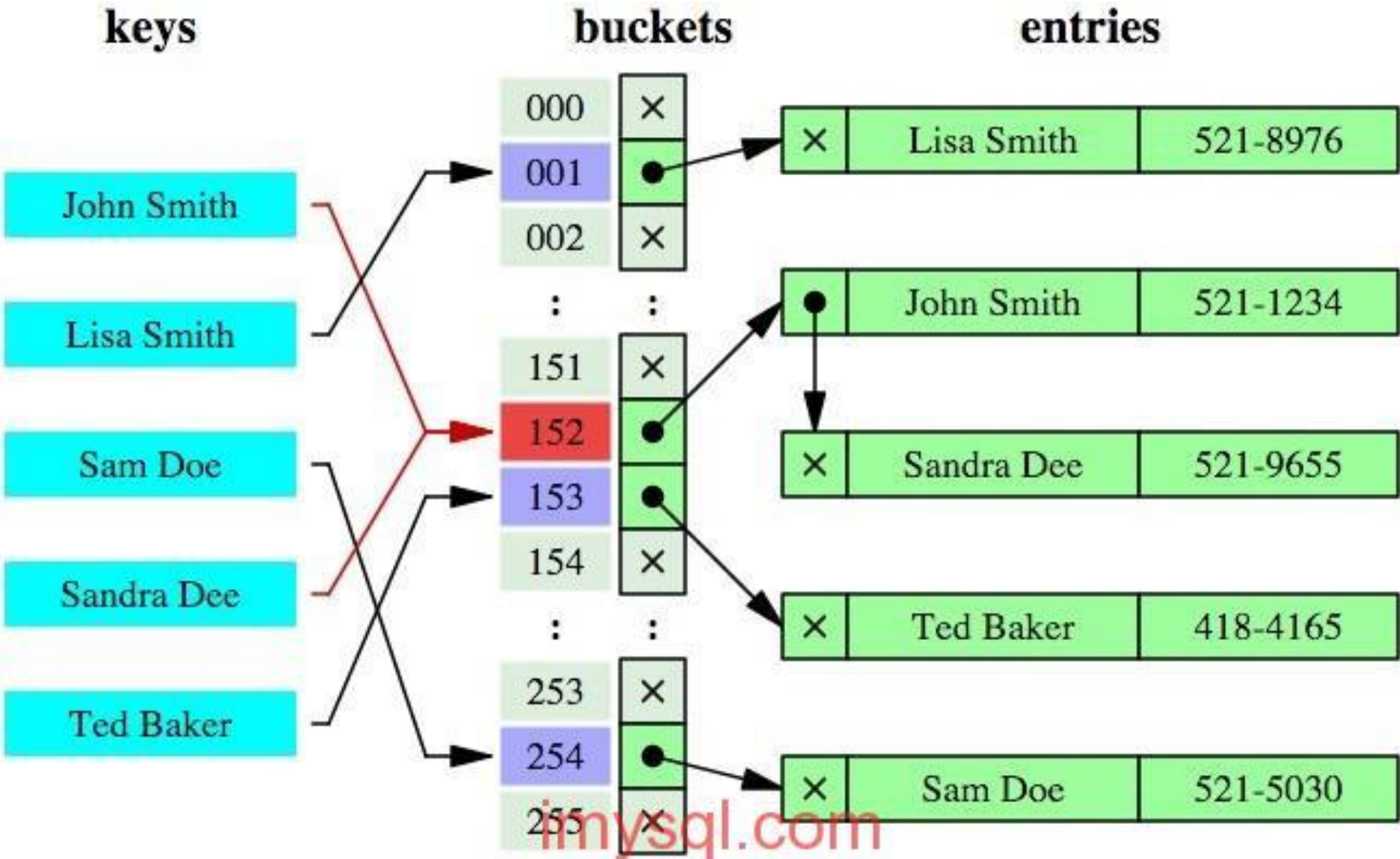
哈希函数，输入词项，输出正整数(通常是地址)

$f(\text{信息})=18$ ,  $f(\text{数据})=19$ ,  
 $f(\text{挖掘})=19$



# 哈希表

- 每个词项通过哈希函数映射成一个整数
- 尽可能避免冲突
- 查询处理时：对查询词项进行哈希，如果有冲突，则解决冲突，最后在定长数组中定位
- 优点：在哈希表中的定位速度快于树中的定位速度
  - 查询时间是常数
- 缺点：
  - 没办法处理词项的微小变形 (*resume* vs. *résumé*)
  - 不支持前缀搜索 (比如所有以*automat*开头的词项)
  - 如果词汇表不断增大，需要定期对所有词项重新哈希



# 字符串Hash入门

字符串Hash可以通俗的理解为，**把一个字符串转换为一个整数**。如果我们通过某种方法，将字符串转换为一个整数，就可以方便的确定的某个字符串是否重复出现过，这是最简单的字符串Hash应用情景了。

当然也不难想到，如果有不同的两个字符串同时Hash到一个整数，这样就比较麻烦了。我们希望这个映射是一个单射，所以问题就是如何构造这个Hash函数，使得他们成为一个单射。不用担心，接下来的内容正要讲解。

• Hash方法

给定一个字符串 $S=s_1s_2s_3..s_n$ ，对字母 $x$ ，我们规定 $idx(x)=x-'a'+1$ 。  
(当然也可以直接用 $si$ 的ASCII值)

• 自然溢出方法

声明： unsigned long long Hash[n]  
公式：

$hash[i]=hash[i-1]*p+id(s[i])$

利用unsigned long long的范围自然溢出，相当于自动对 $2^{64}-1$ 取模

| ASCII 字符代码表 一 |   |               |    |    |    |    |    |    |      |    |    |    |    |    |    |            |    |      |    |      |    |      |    |      |    |      |    |
|---------------|---|---------------|----|----|----|----|----|----|------|----|----|----|----|----|----|------------|----|------|----|------|----|------|----|------|----|------|----|
| 高四位<br>低四位    |   | ASCII 非打印控制字符 |    |    |    |    |    |    |      |    |    |    |    |    |    | ASCII 打印字符 |    |      |    |      |    |      |    |      |    |      |    |
|               |   | 0000          |    |    |    |    |    |    | 0001 |    |    |    |    |    |    | 0010       |    | 0011 |    | 0100 |    | 0101 |    | 0110 |    | 0111 |    |
|               |   | 0             | 1  | 2  | 3  | 4  | 5  | 6  | 7    | 8  | 9  | 10 | 11 | 12 | 13 | 14         | 15 | 16   | 17 | 18   | 19 | 20   | 21 | 22   | 23 | 24   | 25 |
| 0000          | 0 | 0             | 0  | 0  | 0  | 0  | 0  | 0  | 0    | 0  | 0  | 0  | 0  | 0  | 0  | 0          | 0  | 0    | 0  | 0    | 0  | 0    | 0  | 0    | 0  | 0    | 0  |
| 0001          | 1 | 1             | 1  | 1  | 1  | 1  | 1  | 1  | 1    | 1  | 1  | 1  | 1  | 1  | 1  | 1          | 1  | 1    | 1  | 1    | 1  | 1    | 1  | 1    | 1  | 1    | 1  |
| 0010          | 2 | 2             | 2  | 2  | 2  | 2  | 2  | 2  | 2    | 2  | 2  | 2  | 2  | 2  | 2  | 2          | 2  | 2    | 2  | 2    | 2  | 2    | 2  | 2    | 2  | 2    | 2  |
| 0011          | 3 | 3             | 3  | 3  | 3  | 3  | 3  | 3  | 3    | 3  | 3  | 3  | 3  | 3  | 3  | 3          | 3  | 3    | 3  | 3    | 3  | 3    | 3  | 3    | 3  | 3    | 3  |
| 0100          | 4 | 4             | 4  | 4  | 4  | 4  | 4  | 4  | 4    | 4  | 4  | 4  | 4  | 4  | 4  | 4          | 4  | 4    | 4  | 4    | 4  | 4    | 4  | 4    | 4  | 4    | 4  |
| 0101          | 5 | 5             | 5  | 5  | 5  | 5  | 5  | 5  | 5    | 5  | 5  | 5  | 5  | 5  | 5  | 5          | 5  | 5    | 5  | 5    | 5  | 5    | 5  | 5    | 5  | 5    | 5  |
| 0110          | 6 | 6             | 6  | 6  | 6  | 6  | 6  | 6  | 6    | 6  | 6  | 6  | 6  | 6  | 6  | 6          | 6  | 6    | 6  | 6    | 6  | 6    | 6  | 6    | 6  | 6    | 6  |
| 0111          | 7 | 7             | 7  | 7  | 7  | 7  | 7  | 7  | 7    | 7  | 7  | 7  | 7  | 7  | 7  | 7          | 7  | 7    | 7  | 7    | 7  | 7    | 7  | 7    | 7  | 7    | 7  |
| 1000          | 8 | 8             | 8  | 8  | 8  | 8  | 8  | 8  | 8    | 8  | 8  | 8  | 8  | 8  | 8  | 8          | 8  | 8    | 8  | 8    | 8  | 8    | 8  | 8    | 8  | 8    | 8  |
| 1001          | 9 | 9             | 9  | 9  | 9  | 9  | 9  | 9  | 9    | 9  | 9  | 9  | 9  | 9  | 9  | 9          | 9  | 9    | 9  | 9    | 9  | 9    | 9  | 9    | 9  | 9    | 9  |
| 1010          | A | 10            | 10 | 10 | 10 | 10 | 10 | 10 | 10   | 10 | 10 | 10 | 10 | 10 | 10 | 10         | 10 | 10   | 10 | 10   | 10 | 10   | 10 | 10   | 10 | 10   | 10 |
| 1011          | B | 11            | 11 | 11 | 11 | 11 | 11 | 11 | 11   | 11 | 11 | 11 | 11 | 11 | 11 | 11         | 11 | 11   | 11 | 11   | 11 | 11   | 11 | 11   | 11 | 11   | 11 |
| 1100          | C | 12            | 12 | 12 | 12 | 12 | 12 | 12 | 12   | 12 | 12 | 12 | 12 | 12 | 12 | 12         | 12 | 12   | 12 | 12   | 12 | 12   | 12 | 12   | 12 | 12   | 12 |
| 1101          | D | 13            | 13 | 13 | 13 | 13 | 13 | 13 | 13   | 13 | 13 | 13 | 13 | 13 | 13 | 13         | 13 | 13   | 13 | 13   | 13 | 13   | 13 | 13   | 13 | 13   | 13 |
| 1110          | E | 14            | 14 | 14 | 14 | 14 | 14 | 14 | 14   | 14 | 14 | 14 | 14 | 14 | 14 | 14         | 14 | 14   | 14 | 14   | 14 | 14   | 14 | 14   | 14 | 14   | 14 |
| 1111          | F | 15            | 15 | 15 | 15 | 15 | 15 | 15 | 15   | 15 | 15 | 15 | 15 | 15 | 15 | 15         | 15 | 15   | 15 | 15   | 15 | 15   | 15 | 15   | 15 | 15   | 15 |

注：表中的ASCII字符可以用ALT + “小键盘上的数字键”输入



## 单Hash方法

- 公式

$$\text{hash}[i] = (\text{hash}[i-1]) * p + \text{idx}(s[i]) \% \text{mod}$$

其中， $p$ 和 $\text{mod}$ 均为质数，且有 $p < \text{mod}$ 。对于此种Hash方法，将 $p$ 和 $\text{mod}$ 尽量取大即可。这种情况下，冲突的概率很低。

- 举例

如取 $p=13, \text{mod}=101$ ，对字符串 $\text{abc}$ 进行哈希。

$$\text{hash}[0] = 1$$

$$\text{hash}[1] = (\text{hash}[0] * 13 + 2) \% 101 = 15$$

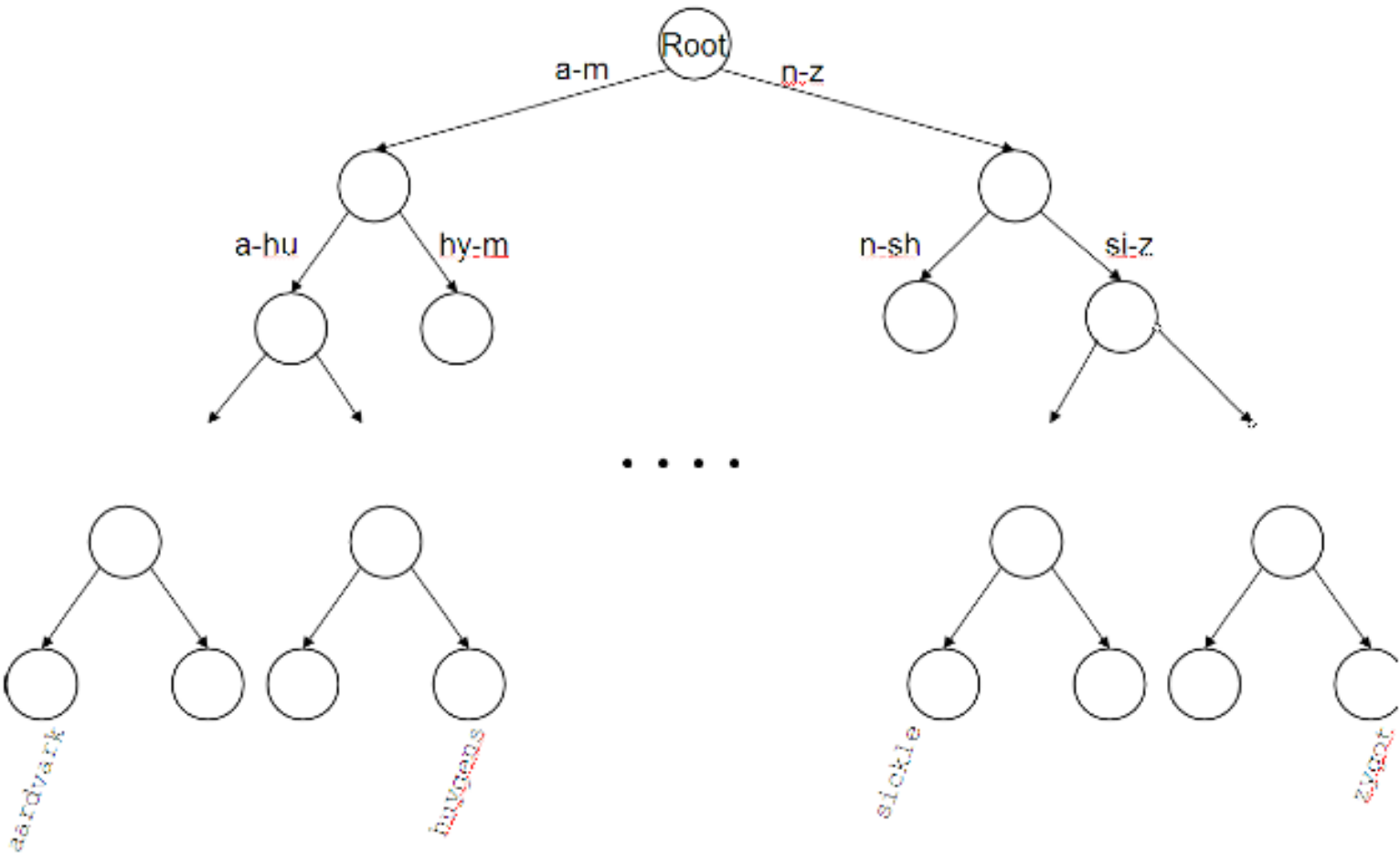
$$\text{hash}[2] = (\text{hash}[1] * 13 + 3) \% 101 = 97$$

这样，我们就认为字符串 $\text{abc}$ 当做97，即97就是 $\text{abc}$ 的hash值。

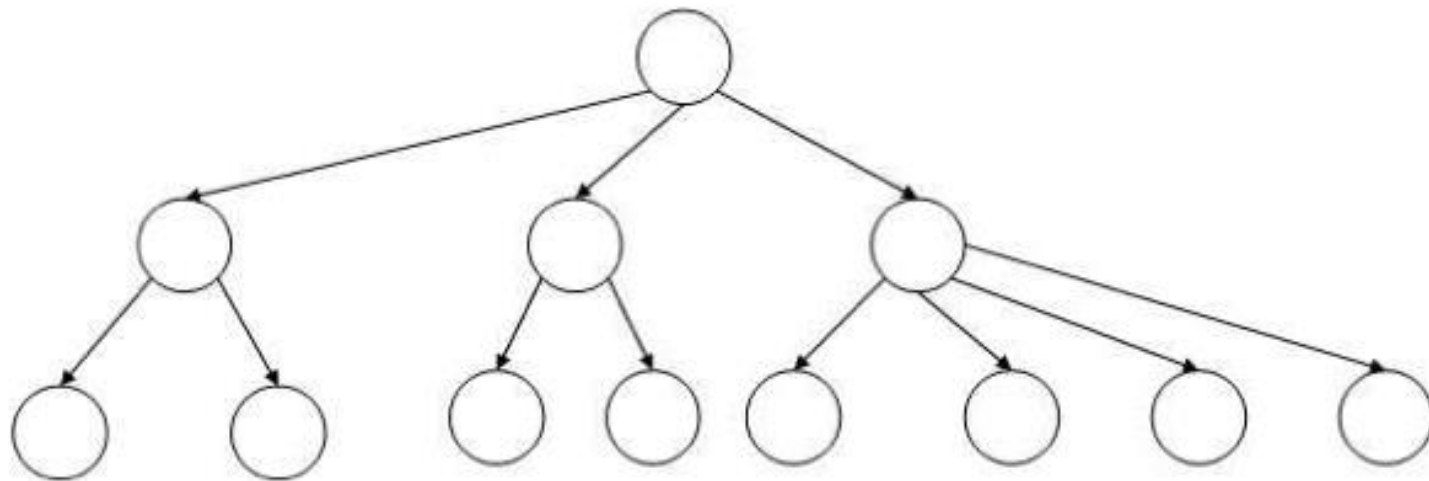
# 树

- 树可以支持前缀查找
- 最简单的树结构：二叉树
- 搜索速度略低于哈希表方式：  $O(\log M)$ , 其中  $M$  是词汇表大小，即所有词项的数目
- $O(\log M)$  仅仅对平衡树成立
- 使二叉树重新保持平衡开销很大
- B-树 能够减轻上述问题
- B-树定义：每个内部节点的子节点数目在  $[a, b]$  之间，其中  $a, b$  为合适的正整数, e.g.,  $[2, 4]$ .

# 二叉树



# B-树



# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

# 通配查询的处理

- $mon^*$ : 找出所有包含以 *mon* 开头的词项的文档
- 如果采用B-树词典结构，那么实现起来非常容易，只需要返回区间  $mon \leq t < moo$  上的词项  $t$
- $*mon$ : 找出所有包含以 *mon* 结尾的词项的文档
  - 将所有的词项倒转过来，然后基于它们建一棵附加的树
  - 返回区间  $nom \leq t < non$  上的词项  $t$
- 也就是说，通过上述数据结构，可能得到满足通配查询的一系列词项，然后返回任一词项的文档

# 词项中间的 \*号处理

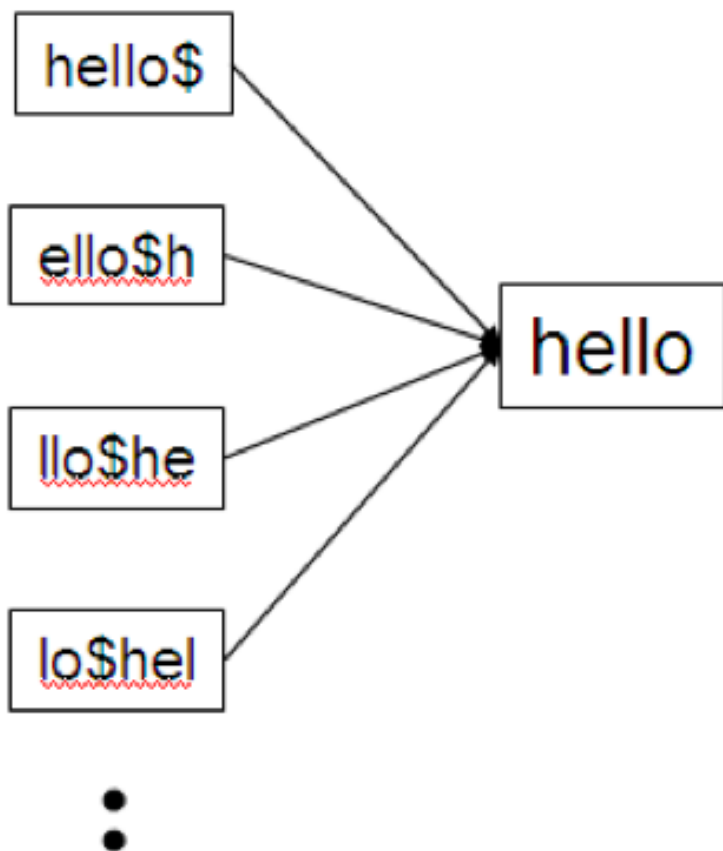
- 例子: m\*nchen
- 在B-树中分别查找满足m\*和 \*nchen的词项集合，然后求交集
- 这种做法开销很大
- 另外一种方法: 轮排([permuterm](#)) 索引
- 基本思想: 将每个通配查询旋转，使\*出现在末尾
- 将每个每个旋转后的结果存放在词典中，即B-树中

# 轮排索引

- 对于词项hello: 将 *hello\$*, *ello\$h*, *llo\$he*, *lo\$hel*, 和 *o\$hell* 加入到 B-树中, 其中 \$ 是一个特殊符号
- 即在词项前面再加一层索引



# 轮排结果 → 词项的映射示意图



# 轮排索引

- 对于hello, 已经存储了 *hello\$, ello\$h, llo\$he, lo\$hel, 和 o\$hell*
- 查询
  - 对于 x, 查询 x\$
  - 对于 x\*, 查询 x\*\$
  - 对于 \*x, 查询 x\$\*
  - 对于 \*x\*, 查询 x\*
  - 对于 x\*y, 查询 y\$x\*
  - 例子: 假定通配查询为 hel\*o, 那么相当于要查询o\$hel\*
- 轮排索引称为轮排树更恰当
- 但是轮排索引已经使用非常普遍

# 使用轮排索引的查找过程

- 将查询进行旋转，将通配符旋转到右部
- 同以往一样查找B-树
- 问题：相对于通常的B-树，轮排树的空间要大4倍以上 (经验值)

# k-gram 索引

- 比轮排索引空间开销要小
- 枚举一个词项中所有连读的k个字符构成的k-gram 。
- 2-gram称为二元组(**bigram**)
- 例子: from *April is the cruelest month* we get the bigrams: \$a  
ap pr ri il l\$ \$i is s\$ \$t th he e\$ \$c cr ru ue el le es st t\$ \$m mo  
on nt h\$
- 同前面一样，\$ 是一个特殊字符
- 构建一个倒排索引，此时词典部分是所有的2-gram，倒排记录表部分是包含某个2-gram的所有词项
- 相当于对词项再构建一个倒排索引(二级索引)

## 3-gram(trigram)索引的例子



## $k$ -gram (bigram, trigram, ...) 索引

- 需要注意的是，这里有两个倒排索引
- 词典-文档的倒排索引基于词项返回文档
- 而 $k$ -gram索引用于查找词项，基于查询包含的 $k$ -gram查找词项

# 利用2-gram索引处理通配符查询

- 查询`mon*` 可以先执行布尔查询: `$m AND mo AND on`
- 该布尔查询会返回所有以前缀`mon`开始的词项 ...
- ... 当然也可能返回许多伪正例, 比如`MOON`.
- 因此, 必须要做后续的过滤处理
- 余下的词项将在词项-文档倒排索引中查找文档
- `k`-gram索引 vs. 轮排索引
  - `k`-gram索引的空间消耗小
  - 轮排索引不需要进行后过滤

# 课堂练习

- Google对通配符查询的支持极其有限
- 比如：在 Google中查询 [gen\* universit\*]
  - 意图：想查 University of Geneva, 但是不知道如何拼写，特别是法语中的拼写
- 按照Google自己的说法, 2010-04-29: “\* 操作符只能作为一个整体单词使用，而不能作为单词的一部分使用”
- 但是这点并不完全对，尝试一下 [pythag\*] 和 [m\*nchen]
- 问题: 为什么Google对通配查询并不充分支持？



# 原因

- 问题 1: 一条通配符查询往往相当于执行非常多的布尔查询
- 对于 [gen\* universit\*]: geneva university OR geneva université OR genève university OR genève université OR general universities OR ...
- 开销非常大
- 问题 2: 用户不愿意敲击更多的键盘
- 如果允许[pyth\* theo\*]代替 [pythagoras' theorem]的话, 用户会倾向于使用前者
- 这样会大大加重搜索引擎的负担
- Google Suggest是一种减轻用户输入负担的好方法

# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

# 拼写校正

- 两个主要用途
  - 纠正待索引文档
  - 纠正用户的查询
- 两种拼写校正的方法
- 词独立(Isolated word)法
  - 只检查每个单词本身的拼写错误
  - 如果某个单词拼写错误后变成另外一个单词，则无法查出，  
e.g., *an asteroid that fell **form** the sky*
- 上下文敏感(Context-sensitive)法
  - 纠错时要考虑周围的单词
  - 能纠正上例中的错误 *form/from*

# 关于文档校正

- 本课当中我们不关心文档的拼写校正问题 (e.g., MS Word)
- 在IR领域, 我们主要对OCR处理后的文档进行拼写校正处理. (OCR = optical character recognition, 光学字符识别)
- IR领域的一般做法是: 不改变文档

# 查询校正

- 第一种方法: 词独立(isolated word)法
- 假设1: 对需要纠错的词存在一系列“正确单词形式”
- 假设2: 需要提供存在错误拼写的单词和正确单词之间的距离计算方式
- 简单的拼写校正算法: 返回与错误单词具有最小距离的“正确”单词
- 例子: *informaton* → *information*
- 可以将词汇表中所有的单词都作为候选的“正确”单词
- 这种方式为什么有问题?

# 使用词汇表的几种其他方式

- 采用标准词典 (韦伯词典, 牛津词典等等)
- 采用领域词典 (面向特定领域的IR系统)
- 采用文档集上的词项词汇表, 但是每个词项均带有权重

# 单词间距离的计算

- 以下将介绍几种计算方法
- 编辑距离(Edit distance或者Levenshtein distance)
- 带权重的编辑距离
- $k$ -gram 重叠率

# 编辑距离

- 两个字符串  $s_1$  和  $s_2$  编辑距离是指从  $s_1$  转换成  $s_2$  所需要的最短的基本操作数目
- Levenshtein距离: 采用的基本操作是插入(insert)、删除(delete)和替换(replace)
- Levenshtein距离 *dog-do*: 1
- Levenshtein距离 *cat-cart*: 1
- Levenshtein距离 *cat-cut*: 1
- Levenshtein距离 *cat-act*: 2
- Damerau-Levenshtein距离 *cat-act*: 1
- Damerau-Levenshtein还包括两个字符之间的交换 (transposition) 操作



# Vladimir Iosifovich Levenshtein

---

- 俄罗斯科学家(1935-)
- 研究信息论、纠错理论
- 毕业于莫斯科国立大学
- 1965年提出Levenshtein距离
- 2006年获得IEEE Richard W. Hamming Medal



# Levenshtein距离: 计算

|   |   |   |   |   |   |
|---|---|---|---|---|---|
|   |   | f | a | s | t |
|   | 0 | 1 | 2 | 3 | 4 |
| c | 1 | 1 | 2 | 3 | 4 |
| a | 2 | 2 | 1 | 2 | 3 |
| t | 3 | 3 | 2 | 2 | 2 |
| s | 4 | 4 | 3 | 2 | 3 |

fast中的f、s、t分别用c、t、s替换，即可得到cats，所以代价是3，即距离是3。

# Levenshtein距离: 算法

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```

1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 

```

|           |         |
|-----------|---------|
| (i-1,j-1) | (i-1,j) |
| (i,j-1)   | (i,j)   |

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

# Levenshtein距离: 算法

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```

1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 

```

|                |              |
|----------------|--------------|
| ( $i-1, j-1$ ) | ( $i-1, j$ ) |
| ( $i, j-1$ )   | ( $i, j$ )   |

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

左邻居

# Levenshtein距离: 算法

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```

1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 

```

|           |         |
|-----------|---------|
| (i-1,j-1) | (i-1,j) |
| (i,j-1)   | (i,j)   |

Operations: insert (cost 1), delete (cost 1), replace (cost 1), copy (cost 0)

上邻居

# Levenshtein距离: 算法

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```

1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 

```

|           |         |
|-----------|---------|
| (i-1,j-1) | (i-1,j) |
| (i,j-1)   | (i,j)   |

Operations: insert (cost 1), delete (cost 1), **replace (cost 1)**, copy (cost 0)

左上邻居

# Levenshtein距离: 算法

LEVENSHTEINDISTANCE( $s_1, s_2$ )

```

1  for  $i \leftarrow 0$  to  $|s_1|$ 
2  do  $m[i, 0] = i$ 
3  for  $j \leftarrow 0$  to  $|s_2|$ 
4  do  $m[0, j] = j$ 
5  for  $i \leftarrow 1$  to  $|s_1|$ 
6  do for  $j \leftarrow 1$  to  $|s_2|$ 
7      do if  $s_1[i] = s_2[j]$ 
8          then  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]\}$ 
9          else  $m[i, j] = \min\{m[i-1, j]+1, m[i, j-1]+1, m[i-1, j-1]+1\}$ 
10 return  $m[|s_1|, |s_2|]$ 

```

|           |         |
|-----------|---------|
| (i-1,j-1) | (i-1,j) |
| (i,j-1)   | (i,j)   |

Operations: insert (cost 1), delete (cost 1), replace (cost 1), **copy**  
(cost 0)

左上邻居

# Levenshtein距离: 例子

|                 |        |
|-----------------|--------|
| Copy<br>Replace | delete |
| insert          |        |

|   |                      | f                                      | a                                      | s                                      | t                                      |
|---|----------------------|--|--|--|--|
|   | <u>0</u>             | <u>1</u> <u>1</u>                      | <u>2</u> <u>2</u>                      | <u>3</u> <u>3</u>                      | <u>4</u> <u>4</u>                      |
| c | <u>1</u><br><u>1</u> | <u>1</u> <u>2</u><br><u>2</u> <u>1</u> | <u>2</u> <u>3</u><br><u>2</u> <u>2</u> | <u>3</u> <u>4</u><br><u>3</u> <u>3</u> | <u>4</u> <u>5</u><br><u>4</u> <u>4</u> |
| a | <u>2</u><br><u>2</u> | <u>2</u> <u>2</u><br><u>3</u> <u>2</u> | <u>1</u> <u>3</u><br><u>3</u> <u>1</u> | <u>3</u> <u>4</u><br><u>2</u> <u>2</u> | <u>4</u> <u>5</u><br><u>3</u> <u>3</u> |
| t | <u>3</u><br><u>3</u> | <u>3</u> <u>3</u><br><u>4</u> <u>3</u> | <u>3</u> <u>2</u><br><u>4</u> <u>2</u> | <u>2</u> <u>3</u><br><u>3</u> <u>2</u> | <u>2</u> <u>4</u><br><u>3</u> <u>2</u> |
| s | <u>4</u><br><u>4</u> | <u>4</u> <u>4</u><br><u>5</u> <u>4</u> | <u>4</u> <u>3</u><br><u>5</u> <u>3</u> | <u>2</u> <u>3</u><br><u>4</u> <u>2</u> | <u>3</u> <u>3</u><br><u>3</u> <u>3</u> |



# Levenshtein矩阵中每个单元包含4个元素

|                                 |                        |
|---------------------------------|------------------------|
| 从左上角邻居到来的开销<br>(copy 或 replace) | 从上方邻居到来的代价<br>(delete) |
| 从左方邻居到来的代价<br>(insert)          | 上述三者之中最低的代价            |

# Levenshtein距离: 例子

|   |                                     | f   | a   | s   | t   |
|---|-------------------------------------|---|---|---|---|
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| c | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> | <div><div>4</div><div>5</div><div>4</div><div>4</div></div> |
| a | <div><div>2</div><div>2</div></div> | <div><div>2</div><div>2</div><div>3</div><div>2</div></div> | <div><div>1</div><div>3</div><div>3</div><div>1</div></div> | <div><div>3</div><div>4</div><div>2</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| t | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>3</div><div>4</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |
| s | <div><div>4</div><div>4</div></div> | <div><div>4</div><div>4</div><div>5</div><div>4</div></div> | <div><div>4</div><div>3</div><div>5</div><div>3</div></div> | <div><div>2</div><div>3</div><div>4</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> |

# 动态规划算法(Cormen et al.)

- 最优子结构: 最优的问题解决方案中包括子解决方案, 及子问题的最优解决方案 (两点最短路径问题最典型的解法就是动态规划算法)
- 重叠的子解决方案Overlapping subsolutions: 子解决方案中有重叠, 如果采用暴力计算方法(穷举法), 子解决方案将会被反复计算, 从而使得计算开销很大.
- 编辑距离计算中的子问题: 两个前缀之间的编辑距离计算

# 带权重的编辑距离

- 思路：对不同的字符进行操作时权重不同
- 希望能更敏锐地捕捉到键盘输入的错误, e.g.,  $m$  更可能被输成  $n$  而不是  $q$

QWERTY KEYBOARD

|        |        |        |        |         |        |        |        |        |        |        |        |        |        |
|--------|--------|--------|--------|---------|--------|--------|--------|--------|--------|--------|--------|--------|--------|
| ~<br>` | !<br>1 | @<br>2 | #<br>3 | \$<br>4 | %<br>5 | ^<br>6 | &<br>7 | *<br>8 | (<br>9 | )<br>0 | -<br>_ | =<br>+ | Delete |
| Tab    | Q      | W      | E      | R       | T      | Y      | U      | I      | O      | P      | {<br>[ | }<br>] | <br>_  |
| Caps   | A      | S      | D      | F       | G      | H      | J      | K      | L      | ;<br>, | "<br>' | Enter  |        |
| Shift  | Z      | X      | C      | V       | B      | N      | M      | <<br>, | ><br>. | ?<br>/ | Shift  |        |        |
| Ctrl   |        | Alt    |        |         |        |        |        |        |        |        | Alt    |        | Ctrl   |

<http://www.computerhope.com>

- 因此，将  $m$  替换为  $n$  的编辑距离将高于替换为  $q$  的距离
- 也就是输入的操作代价矩阵是一个带权重的矩阵
- 对上述动态规划算法进行修改便可以处理权重计算

# 利用编辑距离进行拼写校正

- 给定查询词，穷举词汇表中和该查询的编辑距离(或带权重的编辑聚类)低于某个预定值的所有单词
- 求上述结果和给定的某个“正确”词表之间的交集
- 将交集结果推荐给用户
- 代价很大，实际当中往往通过启发式策略提高查找效率(如：保证两者之间具有较长公共子串)

# 课堂练习

- ① 给出计算 *OSLO* – *SNOW* 之间Levenshtein距离的距离矩阵
- ② 将 *cat* 转换成 *catcat* 需要哪几步Levenshtein编辑操作？

|   |                   | s             | n             | o             | w             |
|---|-------------------|---------------|---------------|---------------|---------------|
|   | $\frac{\quad}{0}$ | $\frac{1}{1}$ | $\frac{2}{2}$ | $\frac{3}{3}$ | $\frac{4}{4}$ |
| o | $\frac{1}{1}$     |               |               |               |               |
| s | $\frac{2}{2}$     |               |               |               |               |
| l | $\frac{3}{3}$     |               |               |               |               |
| o | $\frac{4}{4}$     |               |               |               |               |

|   |  |    |          |    |    |    |
|---|--|----|----------|----|----|----|
|   |  |    | s        | n  | o  | w  |
|   |  |    |          |    |    |    |
|   |  | 0  | 11       | 22 | 33 | 44 |
| o |  | 11 | 12<br>2? |    |    |    |
| s |  | 22 |          |    |    |    |
| l |  | 33 |          |    |    |    |
| o |  | 44 |          |    |    |    |



|   |                                     |   |                                     |                                     |                                     |
|---|-------------------------------------|---|-------------------------------------|-------------------------------------|-------------------------------------|
|   |                                     | s   | n                                   | o                                   | w                                   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div></div> |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> |                                     |                                     |                                     |
| s | <div><div>2</div><div>2</div></div> |   |                                     |                                     |                                     |
| l | <div><div>3</div><div>3</div></div> |   |                                     |                                     |                                     |
| o | <div><div>4</div><div>4</div></div> |   |                                     |                                     |                                     |

|   |                                     |   |   |                                     |                                     |
|---|-------------------------------------|---|---|-------------------------------------|-------------------------------------|
|   |                                     | s   | n   | o                                   | w                                   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div></div> |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>?</div></div> |                                     |                                     |
| s | <div><div>2</div><div>2</div></div> |   |   |                                     |                                     |
| l | <div><div>3</div><div>3</div></div> |   |   |                                     |                                     |
| o | <div><div>4</div><div>4</div></div> |   |   |                                     |                                     |

|   |                                     |   |   |                                     |                                     |
|---|-------------------------------------|---|---|-------------------------------------|-------------------------------------|
|   |                                     | s   | n   | o                                   | w                                   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div></div> |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |                                     |                                     |
| s | <div><div>2</div><div>2</div></div> |   |   |                                     |                                     |
| l | <div><div>3</div><div>3</div></div> |   |   |                                     |                                     |
| o | <div><div>4</div><div>4</div></div> |   |   |                                     |                                     |

|   |                                     | s   | n   | o   | w                                   |
|---|-------------------------------------|---|---|---|-------------------------------------|
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div> |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>?</div></div> |                                     |
| s | <div><div>2</div><div>2</div></div> |   |   |   |                                     |
| l | <div><div>3</div><div>3</div></div> |   |   |   |                                     |
| o | <div><div>4</div><div>4</div></div> |   |   |   |                                     |

|   |  |                                     |   |   |   |                                     |   |  |   |
|---|--|-------------------------------------|---|---|---|-------------------------------------|---|--|---|
|   |  |                                     | s   |   | n   |                                     | o |  | w |
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div> |   |  |   |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |                                     |   |  |   |
| s |  | <div><div>2</div><div>2</div></div> |   |   |   |                                     |   |  |   |
| l |  | <div><div>3</div><div>3</div></div> |   |   |   |                                     |   |  |   |
| o |  | <div><div>4</div><div>4</div></div> |   |   |   |                                     |   |  |   |

|   |  |                                     |   |  |   |  |   |  |   |  |
|---|--|-------------------------------------|---|--|---|--|---|--|---|--|
|   |  |                                     | s   |  | n   |  | o   |  | w   |  |
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         |  | <div><div>2</div><div>2</div></div>                         |  | <div><div>3</div><div>3</div></div>                         |  | <div><div>4</div><div>4</div></div>                         |  |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |  | <div><div>4</div><div>5</div><div>3</div><div>?</div></div> |  |
| s |  | <div><div>2</div><div>2</div></div> |   |  |   |  |   |  |   |  |
| l |  | <div><div>3</div><div>3</div></div> |   |  |   |  |   |  |   |  |
| o |  | <div><div>4</div><div>4</div></div> |   |  |   |  |   |  |   |  |

|   |                                     |                                       |                                       |                                       |                                       |
|---|-------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                     | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div>  | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>            | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>2</div><div>2</div></div> |                                       |                                       |                                       |                                       |
| l | <div><div>3</div><div>3</div></div> |                                       |                                       |                                       |                                       |
| o | <div><div>4</div><div>4</div></div> |                                       |                                       |                                       |                                       |

|   |                                     |   |  |   |  |   |  |   |  |
|---|-------------------------------------|---|--|---|--|---|--|---|--|
|   |                                     | s   |  | n   |  | o   |  | w   |  |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                                     |  | <div><div>2</div><div>2</div></div>                                     |  | <div><div>3</div><div>3</div></div>                                     |  | <div><div>4</div><div>4</div></div>                                     |  |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div></div> <div><div>2</div><div>1</div></div> |  | <div><div>2</div><div>3</div></div> <div><div>2</div><div>2</div></div> |  | <div><div>2</div><div>4</div></div> <div><div>3</div><div>2</div></div> |  | <div><div>4</div><div>5</div></div> <div><div>3</div><div>3</div></div> |  |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div></div> <div><div>3</div><div>?</div></div> |  |   |  |   |  |   |  |
| l | <div><div>3</div><div>3</div></div> |   |  |   |  |   |  |   |  |
| o | <div><div>4</div><div>4</div></div> |   |  |   |  |   |  |   |  |



|   |                                     |   |   |   |   |
|---|-------------------------------------|---|---|---|---|
|   |                                     | s   | n   | o   | w   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> |   |   |   |
| l | <div><div>3</div><div>3</div></div> |   |   |   |   |
| o | <div><div>4</div><div>4</div></div> |   |   |   |   |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>2?</div></div> |                                       |                                       |
| l | <div><div>33</div></div>           |                                       |                                       |                                       |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |

|   |                                      |                                       |                                       |                                       |                                       |
|---|--------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                      | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div>   | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>             | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div><div>2</div></div> | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> |                                       |                                       |
| l | <div><div>3</div><div>3</div></div>  |                                       |                                       |                                       |                                       |
| o | <div><div>4</div><div>4</div></div>  |                                       |                                       |                                       |                                       |

|   |  |                                     |   |   |   |   |   |  |   |
|---|--|-------------------------------------|---|---|---|---|---|--|---|
|   |  |                                     | s   |   | n   |   | o |  | w |
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                                     | <div><div>2</div><div>2</div></div>                                     | <div><div>3</div><div>3</div></div>                                     | <div><div>4</div><div>4</div></div>                                     |   |  |   |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div></div> <div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div></div> <div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div></div> <div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div></div> <div><div>3</div><div>3</div></div> |   |  |   |
| s |  | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div></div> <div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div></div> <div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div></div> <div><div>3</div><div>?</div></div> |   |   |  |   |
| l |  | <div><div>3</div><div>3</div></div> |   |   |   |   |   |  |   |
| o |  | <div><div>4</div><div>4</div></div> |   |   |   |   |   |  |   |

|   |  |   |   |   |   |   |   |   |   |   |
|---|--|---|---|---|---|---|---|---|---|---|
|   |  |   | s |   | n |   | o |   | w |   |
|   |  |   |   |   |   |   |   |   |   |   |
|   |  | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| o |  | 1 | 1 | 2 | 2 | 3 | 2 | 4 | 4 | 5 |
|   |  | 1 | 2 | 1 | 2 | 2 | 3 | 2 | 3 | 3 |
| s |  | 2 | 1 | 2 | 2 | 3 | 3 | 3 |   |   |
|   |  | 2 | 3 | 1 | 2 | 2 | 3 | 3 |   |   |
| l |  | 3 |   |   |   |   |   |   |   |   |
|   |  | 3 |   |   |   |   |   |   |   |   |
| o |  | 4 |   |   |   |   |   |   |   |   |
|   |  | 4 |   |   |   |   |   |   |   |   |

|   |  |                                     |   |  |   |  |   |  |   |
|---|--|-------------------------------------|---|--|---|--|---|--|---|
|   |  |                                     | s   |  | n   |  | o   |  | w   |
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         |  | <div><div>2</div><div>2</div></div>                         |  | <div><div>3</div><div>3</div></div>                         |  | <div><div>4</div><div>4</div></div>                         |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |  | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s |  | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> |  | <div><div>3</div><div>4</div><div>4</div><div>?</div></div> |
| l |  | <div><div>3</div><div>3</div></div> |   |  |   |  |   |  |   |
| o |  | <div><div>4</div><div>4</div></div> |   |  |   |  |   |  |   |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           |                                       |                                       |                                       |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>4?</div></div> |                                       |                                       |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |



|   |                                     |   |  |   |  |   |  |   |  |
|---|-------------------------------------|---|--|---|--|---|--|---|--|
|   |                                     | s   |  | n   |  | o   |  | w   |  |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         |  | <div><div>2</div><div>2</div></div>                         |  | <div><div>3</div><div>3</div></div>                         |  | <div><div>4</div><div>4</div></div>                         |  |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |  | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |  |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> |  | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |  |
| l | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> |  |   |  |   |  |   |  |
| o | <div><div>4</div><div>4</div></div> |   |  |   |  |   |  |   |  |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>3?</div></div> |                                       |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> |                                       |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>3?</div></div> |                                       |
| o | <div><div>44</div></div>           |                                       |                                       |                                       |                                       |

|   |                                     |   |   |   |   |
|---|-------------------------------------|---|---|---|---|
|   |                                     | s   | n   | o   | w   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> |   |
| o | <div><div>4</div><div>4</div></div> |   |   |   |   |

|   |                                     |   |   |   |   |
|---|-------------------------------------|---|---|---|---|
|   |                                     | s   | n   | o   | w   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div><div>4</div><div>?</div></div> |
| o | <div><div>4</div><div>4</div></div> |   |   |   |   |

|   |  |                                     |   |  |   |  |   |  |   |
|---|--|-------------------------------------|---|--|---|--|---|--|---|
|   |  |                                     | s   |  | n   |  | o   |  | w   |
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         |  | <div><div>2</div><div>2</div></div>                         |  | <div><div>3</div><div>3</div></div>                         |  | <div><div>4</div><div>4</div></div>                         |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> |  | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s |  | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> |  | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> |  | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> |  | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |
| l |  | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> |  | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> |  | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> |  | <div><div>4</div><div>4</div><div>4</div><div>4</div></div> |
| o |  | <div><div>4</div><div>4</div></div> |   |  |   |  |   |  |   |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>5?</div></div> |                                       |                                       |                                       |



|   |                                     | s                                     | n                                     | o                                     | w                                     |
|---|-------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div>  | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>            | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>4</div><div>4</div></div> | <div><div>43</div><div>53</div></div> |                                       |                                       |                                       |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>4?</div></div> |                                       |                                       |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> |                                       |                                       |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>4?</div></div> |                                       |

|   |                                     |   |   |   |   |
|---|-------------------------------------|---|---|---|---|
|   |                                     | s   | n   | o   | w   |
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div><div>4</div><div>4</div></div> |
| o | <div><div>4</div><div>4</div></div> | <div><div>4</div><div>3</div><div>5</div><div>3</div></div> | <div><div>3</div><div>3</div><div>4</div><div>3</div></div> | <div><div>2</div><div>4</div><div>4</div><div>2</div></div> |   |

|   |                                     | s   | n   | o   | w   |
|---|-------------------------------------|---|---|---|---|
|   | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |
| o | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div><div>4</div><div>4</div></div> |
| o | <div><div>4</div><div>4</div></div> | <div><div>4</div><div>3</div><div>5</div><div>3</div></div> | <div><div>3</div><div>3</div><div>4</div><div>3</div></div> | <div><div>2</div><div>4</div><div>4</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>?</div></div> |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>42</div></div> | <div><div>45</div><div>33</div></div> |

|   |                                    | s                                     | n                                     | o                                     | w                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>42</div></div> | <div><div>45</div><div>33</div></div> |



|   |  | s                                  |                                     | n                                   |                                     | o                                   |  | w |  |
|---|--|------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|--|---|--|
|   |  | <div><div></div><div>0</div></div> | <div><div>1</div><div>1</div></div> | <div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div></div> |  |   |  |
| o |  | <div>1</div>                       | <div><div>1</div><div>2</div></div> | <div><div>2</div><div>3</div></div> | <div><div>2</div><div>4</div></div> | <div><div>4</div><div>5</div></div> |  |   |  |
|   |  | <div>1</div>                       | <div><div>2</div><div>1</div></div> | <div><div>2</div><div>2</div></div> | <div><div>3</div><div>2</div></div> | <div><div>3</div><div>3</div></div> |  |   |  |
| s |  | <div>2</div>                       | <div><div>1</div><div>2</div></div> | <div><div>2</div><div>3</div></div> | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div></div> |  |   |  |
|   |  | <div>2</div>                       | <div><div>3</div><div>1</div></div> | <div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>3</div></div> |  |   |  |
| l |  | <div>3</div>                       | <div><div>3</div><div>2</div></div> | <div><div>2</div><div>3</div></div> | <div><div>3</div><div>4</div></div> | <div><div>4</div><div>4</div></div> |  |   |  |
|   |  | <div>3</div>                       | <div><div>4</div><div>2</div></div> | <div><div>3</div><div>2</div></div> | <div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div></div> |  |   |  |
| o |  | <div>4</div>                       | <div><div>4</div><div>3</div></div> | <div><div>3</div><div>3</div></div> | <div><div>2</div><div>4</div></div> | <div><div>4</div><div>5</div></div> |  |   |  |
|   |  | <div>4</div>                       | <div><div>5</div><div>3</div></div> | <div><div>4</div><div>3</div></div> | <div><div>4</div><div>2</div></div> | <div><div>3</div><div>3</div></div> |  |   |  |

如何从上述矩阵中找到编辑操作的路径？

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>42</div></div> | <div><div>45</div><div>33</div></div> |

|      |           |       |        |
|------|-----------|-------|--------|
| cost | operation | input | output |
| 1    | insert    | *     | w      |

|   |                                    |                                       |                                       |                                       |                                       |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                    | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>           | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>22</div></div>           | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>44</div></div>           | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>42</div></div> | <div><div>45</div><div>33</div></div> |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 0    | (copy)    | o     | o      |
| 1    | insert    | *     | w      |

|   |  | s                                   |   | n   |   | o   |  | w |  |
|---|--|-------------------------------------|---|---|---|---|--|---|--|
|   |  | <div><div></div><div>0</div></div>  | <div><div>1</div><div>1</div></div>                         | <div><div>2</div><div>2</div></div>                         | <div><div>3</div><div>3</div></div>                         | <div><div>4</div><div>4</div></div>                         |  |   |  |
| o |  | <div><div>1</div><div>1</div></div> | <div><div>1</div><div>2</div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>2</div><div>4</div><div>3</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |  |   |  |
| s |  | <div><div>2</div><div>2</div></div> | <div><div>1</div><div>2</div><div>3</div><div>1</div></div> | <div><div>2</div><div>3</div><div>2</div><div>2</div></div> | <div><div>3</div><div>3</div><div>3</div><div>3</div></div> | <div><div>3</div><div>4</div><div>4</div><div>3</div></div> |  |   |  |
| l |  | <div><div>3</div><div>3</div></div> | <div><div>3</div><div>2</div><div>4</div><div>2</div></div> | <div><div>2</div><div>3</div><div>3</div><div>2</div></div> | <div><div>3</div><div>4</div><div>3</div><div>3</div></div> | <div><div>4</div><div>4</div><div>4</div><div>4</div></div> |  |   |  |
| o |  | <div><div>4</div><div>4</div></div> | <div><div>4</div><div>3</div><div>5</div><div>3</div></div> | <div><div>3</div><div>3</div><div>4</div><div>3</div></div> | <div><div>2</div><div>4</div><div>4</div><div>2</div></div> | <div><div>4</div><div>5</div><div>3</div><div>3</div></div> |  |   |  |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 1    | replace   | l     | n      |
| 0    | (copy)    | o     | o      |
| 1    | insert    | •     | w      |

|   |  | s  |      | n    |      | o    |  | w |  |
|---|--|----|------|------|------|------|--|---|--|
|   |  | 0  | 11   | 22   | 33   | 44   |  |   |  |
| o |  | 11 | 1221 | 2322 | 2432 | 4533 |  |   |  |
| s |  | 22 | 1231 | 2322 | 3333 | 3443 |  |   |  |
| l |  | 33 | 3242 | 2332 | 3433 | 4444 |  |   |  |
| o |  | 44 | 4353 | 3343 | 2442 | 4533 |  |   |  |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 0    | (copy)    | s     | s      |
| 1    | replace   | l     | n      |
| 0    | (copy)    | o     | o      |
| 1    | insert    | *     | w      |

|   |                                     |                                       |                                       |                                       |                                       |
|---|-------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   |                                     | s                                     | n                                     | o                                     | w                                     |
|   | <div><div></div><div>0</div></div>  | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              |
| o | <div><div>11</div></div>            | <div><div>12</div><div>21</div></div> | <div><div>23</div><div>22</div></div> | <div><div>24</div><div>32</div></div> | <div><div>45</div><div>33</div></div> |
| s | <div><div>2</div><div>2</div></div> | <div><div>12</div><div>31</div></div> | <div><div>23</div><div>22</div></div> | <div><div>33</div><div>33</div></div> | <div><div>34</div><div>43</div></div> |
| l | <div><div>3</div><div>3</div></div> | <div><div>32</div><div>42</div></div> | <div><div>23</div><div>32</div></div> | <div><div>34</div><div>33</div></div> | <div><div>44</div><div>44</div></div> |
| o | <div><div>4</div><div>4</div></div> | <div><div>43</div><div>53</div></div> | <div><div>33</div><div>43</div></div> | <div><div>24</div><div>42</div></div> | <div><div>45</div><div>33</div></div> |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 1    | delete    | o     | *      |
| 0    | (copy)    | s     | s      |
| 1    | replace   | l     | n      |
| 0    | (copy)    | o     | o      |
| 1    | insert    | *     | w      |

从*cat*到*catcat*

|   |  | c |   | a |   | t |   | c |   | a |   | t |   |   |   |
|---|--|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|   |  |   | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 |
| c |  | 1 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 | 5 | 6 | 6 | 7 |
|   |  | 1 | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| a |  | 2 | 2 | 2 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 | 5 | 6 |
|   |  | 2 | 2 | 3 | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| t |  | 3 | 3 | 3 | 2 | 2 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 |
|   |  | 3 | 3 | 4 | 2 | 3 | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |

|   |                                    | c                                     | a                                     | t                                     | c                                     | a                                     | t                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              | <div><div>55</div></div>              | <div><div>66</div></div>              |
| c | <div><div>11</div></div>           | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> | <div><div>56</div><div>44</div></div> | <div><div>67</div><div>55</div></div> |
| a | <div><div>22</div></div>           | <div><div>21</div><div>31</div></div> | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> | <div><div>56</div><div>44</div></div> |
| t | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>21</div><div>31</div></div> | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 1    | insert    | *     | c      |
| 1    | insert    | *     | a      |
| 1    | insert    | *     | t      |
| 0    | (copy)    | c     | c      |
| 0    | (copy)    | a     | a      |
| 0    | (copy)    | t     | t      |



|   |                                    | c                                     | a                                     | t                                     | c                                     | a                                     | t                                     |
|---|------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|---------------------------------------|
|   | <div><div></div><div>0</div></div> | <div><div>11</div></div>              | <div><div>22</div></div>              | <div><div>33</div></div>              | <div><div>44</div></div>              | <div><div>55</div></div>              | <div><div>66</div></div>              |
| c | <div><div>11</div></div>           | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> | <div><div>56</div><div>44</div></div> | <div><div>67</div><div>55</div></div> |
| a | <div><div>22</div></div>           | <div><div>21</div><div>31</div></div> | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> | <div><div>56</div><div>44</div></div> |
| t | <div><div>33</div></div>           | <div><div>32</div><div>42</div></div> | <div><div>21</div><div>31</div></div> | <div><div>02</div><div>20</div></div> | <div><div>23</div><div>11</div></div> | <div><div>34</div><div>22</div></div> | <div><div>35</div><div>33</div></div> |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 0    | (copy)    | c     | c      |
| 1    | insert    | *     | a      |
| 1    | insert    | *     | t      |
| 1    | insert    | *     | c      |
| 0    | (copy)    | a     | a      |
| 0    | (copy)    | t     | t      |

|   |                                     | c                                   |                                     | a                                   |                                     | t                                   |                                     | c                                   |                                     | a                                   |                                     | t                                   |                                     |
|---|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|-------------------------------------|
|   | <div><div></div><div>0</div></div>  | <div>1</div>                        | <div>1</div>                        | <div>2</div>                        | <div>2</div>                        | <div>3</div>                        | <div>3</div>                        | <div>4</div>                        | <div>4</div>                        | <div>5</div>                        | <div>5</div>                        | <div>6</div>                        | <div>6</div>                        |
| c | <div><div>1</div><div>1</div></div> | <div><div>0</div><div>2</div></div> | <div><div>2</div><div>0</div></div> | <div>2</div> <div>3</div>           | <div>1</div> <div>1</div>           | <div>3</div> <div>4</div>           | <div>2</div> <div>2</div>           | <div>3</div> <div>5</div>           | <div>3</div> <div>3</div>           | <div>5</div> <div>6</div>           | <div>4</div> <div>4</div>           | <div>6</div> <div>7</div>           | <div>5</div> <div>5</div>           |
| a | <div><div>2</div><div>2</div></div> | <div>2</div> <div>3</div>           | <div><div>1</div><div>1</div></div> | <div><div>0</div><div>2</div></div> | <div><div>2</div><div>1</div></div> | <div><div>2</div><div>3</div></div> | <div><div>1</div><div>1</div></div> | <div><div>3</div><div>4</div></div> | <div><div>2</div><div>2</div></div> | <div><div>3</div><div>5</div></div> | <div><div>3</div><div>3</div></div> | <div>5</div> <div>6</div>           | <div>4</div> <div>4</div>           |
| t | <div><div>3</div><div>3</div></div> | <div>3</div> <div>4</div>           | <div>2</div> <div>2</div>           | <div>2</div> <div>3</div>           | <div>1</div> <div>1</div>           | <div>0</div> <div>2</div>           | <div>2</div> <div>0</div>           | <div>2</div> <div>3</div>           | <div>1</div> <div>1</div>           | <div>3</div> <div>4</div>           | <div>2</div> <div>2</div>           | <div><div>3</div><div>3</div></div> | <div><div>5</div><div>3</div></div> |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 0    | (copy)    | c     | c      |
| 0    | (copy)    | a     | a      |
| 1    | insert    | *     | t      |
| 1    | insert    | *     | c      |
| 1    | insert    | *     | a      |
| 0    | (copy)    | t     | t      |

|   |  |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|--|---|---|---|---|---|---|---|---|---|---|---|---|---|
|   |  | c |   | a |   | t |   | c |   | a |   | t |   |   |
|   |  | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 |
| c |  | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 | 5 | 6 | 6 | 7 |
|   |  | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 |
| a |  | 2 | 2 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 | 5 | 6 |
|   |  | 2 | 3 | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 | 4 | 4 |
| t |  | 3 | 3 | 2 | 2 | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 3 | 5 |
|   |  | 3 | 4 | 2 | 3 | 1 | 2 | 0 | 1 | 1 | 2 | 2 | 3 | 3 |

| cost | operation | input | output |
|------|-----------|-------|--------|
| 0    | (copy)    | c     | c      |
| 0    | (copy)    | a     | a      |
| 0    | (copy)    | t     | t      |
| 1    | insert    | *     | c      |
| 1    | insert    | *     | a      |
| 1    | insert    | *     | t      |

# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ Soundex

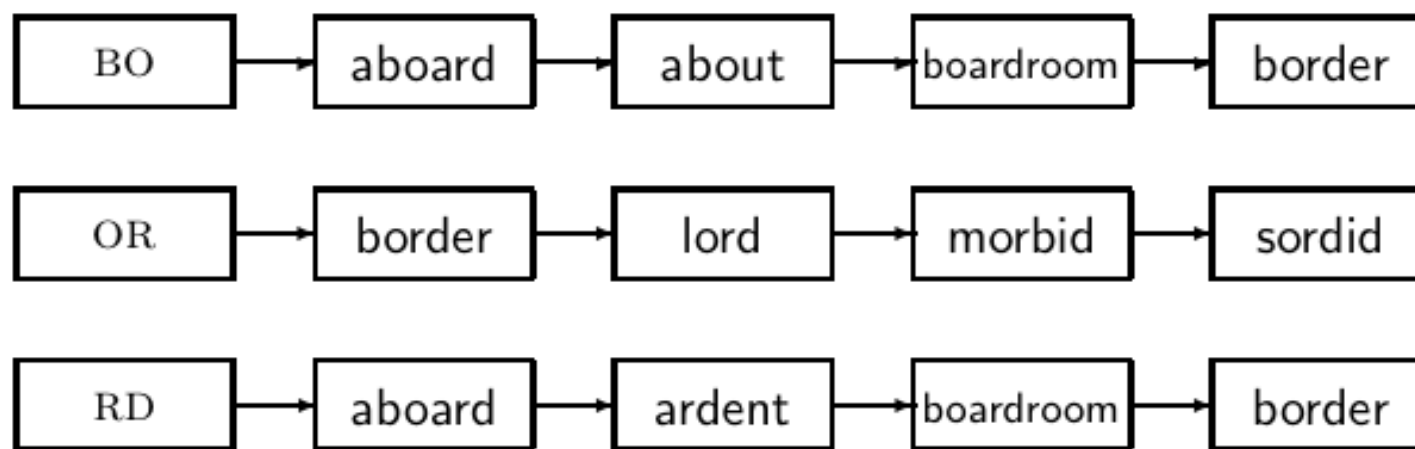
# 拼写校正

- 刚才已经介绍如何利用编辑距离进行词独立方式下的拼写校正
- 另一种方法： $k$ -gram索引
- 上下文敏感的拼写校正
- 拼写校正中的一般问题

# 基于 $k$ -gram索引的拼写校正

- 列举查询词项中的所有 $k$ -gram
  - 例子：采用2-gram索引, 错误拼写的单词为bordroom
  - 2-gram: *bo, or, rd, dr, ro, oo, om*
- 利用  $k$ -gram索引返回和能够匹配很多查询 $k$ -gram的正确单词
- 匹配程度(数目或者指标)上可以事先设定阈值
- E.g., 比如最多只有 3 个 $k$ -gram不同

## 2-gram索引示意图



# 上下文敏感的拼写校正

- 例子: *an asteroid that fell **form** the sky*
- 如何对*form*纠错?
- 一种方法: 基于命中数(**hit-based**)的拼写校正
  - 对于每个查询词项返回 相近的“正确” 词项
  - *flew form munich: flea ->flew, from -> form, munch ->munich*
  - 组合所有可能
  - 搜索 “*flea form munich*”
  - 搜索 “*flew from munich*”
  - 搜索 “*flew form munch*”
  - 正确查询 “*flew from munich*” 会有最高的结果命中数
- 假定 *flew*有7个可能的候选词, *form* 有20个, *munich* 有3个, 那么需要穷举出多少个查询?



# 上下文敏感的拼写校正

- 刚才提到的基于命中数的算法效率不高The “hit-based” algorithm we just 提纲d is not very efficient.
- 一种更高效的做法是： 从查询库(比如历史查询)中搜索而不是从文档库中搜索

# 拼写校正中的一般问题

- 用户交互界面问题
  - 全自动 vs. 推荐式校正方法(Did you mean...?)
  - 推荐式校正方法通常只给出一个建议
  - 如果有多个可能的正确拼写怎么办?
  - 平衡: 交互界面的简洁性 vs. 强大性
- 开销问题
  - 拼写校正的开销很大
  - 避免对所有查询都运行拼写校正模块
  - 只对返回结果很少的查询运行拼写校正模块
  - 猜测: 主流搜索引擎的拼写校正模块非常高效, 有能力对每个查询进行拼写校正

# 课堂练习: Peter Norvig拼写校正工具的理解

```
import re, collections
def words(text): return re.findall('[a-z]+', text.lower())
def train(features):
    model = collections.defaultdict(lambda: 1)
    for f in features:
        model[f] += 1
    return model
NWORDS = train(words(file('big.txt').read()))
alphabet = 'abcdefghijklmnopqrstuvwxyz'
def edits1(word):
    splits      = [(word[:i], word[i:]) for i in range(len(word) + 1)]
    deletes     = [a + b[1:] for a, b in splits if b]
    transposes  = [a + b[1] + b[0] + b[2:] for a, b in splits if len(b)>1]
    replaces    = [a + c + b[1:] for a, b in splits for c in alphabet if b]
    inserts     = [a + c + b      for a, b in splits for c in alphabet]
    return set(deletes + transposes + replaces + inserts)
def known_edits2(word):
    return set(e2 for e1 in edits1(word) for e2 in
               edits1(e1) if e2 in NWORDS)
def known(words): return set(w for w in words if w in NWORDS)
def correct(word):
    candidates = known([word]) or known(edits1(word)) or
                 known_edits2(word) or [word]
    return max(candidates, key=NWORDS.get)
```

# 提纲

① 上一讲回顾

② 词典

③ 通配查询

④ 编辑距离

⑤ 拼写校正

⑥ **Soundex**

# Soundex

- Soundex是寻找发音相似的单词的方法
- 比如: *chebyshev / tchebyscheff*
- 算法:
  - 将词典中每个词项转换成一个4字符缩减形式
  - 对查询词项做同样的处理
  - 基于4-字符缩减形式进行索引和搜索

# Soundex 算法

- ① 保留词项的首字母
- ② 将后续所有的A、E、I、O、U、H、W及Y等字母转换为0。
- ③ 按照如下方式将字母转换成数字：
  - B, F, P, V → 1
  - C, G, J, K, Q, S, X, Z → 2
  - D, T → 3
  - L → 4
  - M, N → 5
  - R → 6
- ④ 将连续出现的两个同一字符转换为一个字符直至再没有这种现象出现。
- ⑤ 在结果字符串中剔除0，并在结果字符串尾部补足0，然后返回前四个字符，该字符由1个字母加上3个数字组成。

# 例子: 采用Soundex算法处理 *HERMAN*

- 保留 H
- *ERMAN* → *ORMON*
- *ORMON* → 06505
- 06505 → 06505
- 06505 → 655
- 返回 *H655*
- 注意: *HERMANN* 会产生同样的编码

# Soundex的应用情况

- 在IR中并不非常普遍
- 适用于“高召回率”任务 (e.g., 国际刑警组织Interpol在全球范围内追查罪犯)
- Zobel and Dart (1996)提出了一个更好的发音匹配方法



# 课堂练习

- 计算你的姓的拼音的Soundex编码

# 本讲小结

- 词典的数据结构：访问效率和支持查找的方式
  - 哈希表 vs. 树结构
- 容错式检索(Tolerant retrieval):
  - 通配查询：包含通配符\*的查询
    - 轮排索引 vs. k-gram索引
  - 拼写校正：
    - 编辑距离 vs. k-gram相似度
    - 词独立校正法 vs. 上下文敏感校正法
    - Soundex算法

# 前一段小结

---

- 目前的索引方式和查询方式
  - 布尔查询：普通倒排索引、+跳表
  - 短语查询：双词索引、位置索引
  - 通配查询：轮排索引、k-gram索引
  - 拼写错误的查询：k-gram索引
  - 发音错误的查询：Soundex索引
- 包含各种格式索引的搜索引擎甚至可以处理如下查询
- (SPELL(moriset) /3 toron\*to) OR  
SOUNDEX(chaikofski)

# 参考资料

- 《信息检索导论》第3章、MG4.2
- 高效拼写校正方法:
  - K. Kukich. Techniques for automatically correcting words in text. ACM Computing Surveys 24(4), Dec 1992.
  - J. Zobel and P. Dart. Finding approximate matches in large lexicons. Software - practice and experience 25(3), March 1995.  
<http://citeseer.ist.psu.edu/zobel95finding.html>
  - Mikael Tillenius: Efficient Generation and Ranking of Spelling Error Corrections. Master's thesis at Sweden's Royal Institute of Technology.  
<http://citeseer.ist.psu.edu/179155.html>
  - Peter Norvig: How to write a spelling corrector  
<http://norvig.com/spell-correct.html>
- <http://ifnlp.org/ir>
  - Soundex演示
  - Levenshtein距离的演示
  - Peter Norvig的拼写校正工具

# 课后练习

---

- 习题3-6
- 习题3-7
- 习题3-10
- 习题3-12