

自然语言处理

Natural Language Processing

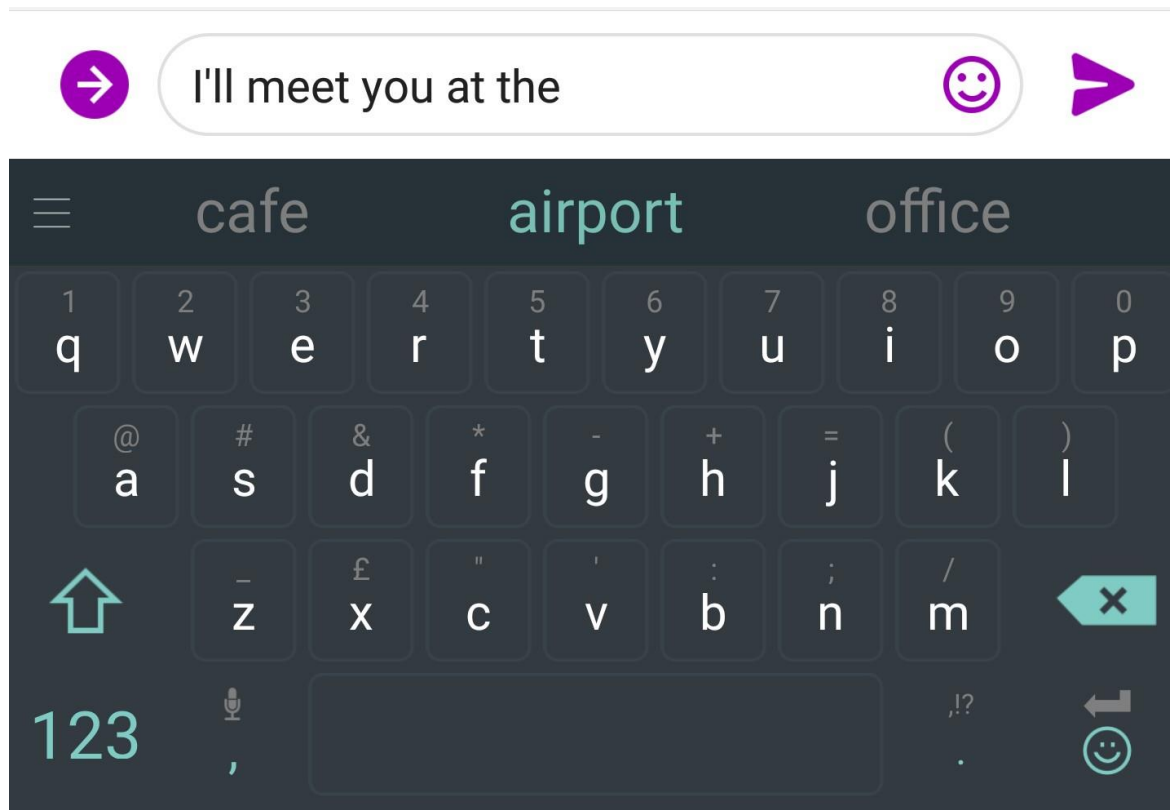
权小军 教授

中山大学数据科学与计算机学院

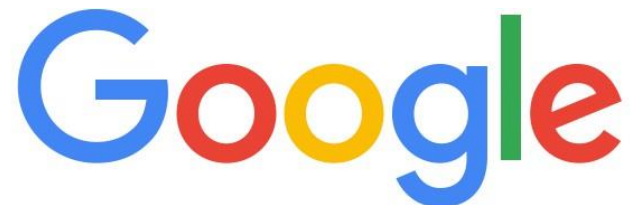
quanxj3@mail.sysu.edu.cn

应用展示

应用展示



应用展示



what is the |



what is the **weather**

what is the **meaning of life**

what is the **dark web**

what is the **xfl**

what is the **doomsday clock**

what is the **weather today**

what is the **keto diet**

what is the **american dream**

what is the **speed of light**

what is the **bill of rights**

Google Search

I'm Feeling Lucky

Lecture 7: 语言模型（上）

语言模型

- 
1. 统计语言模型
 2. 神经语言模型

统计语言模型

1. 基本概念

2. 参数估计

3. 数据平滑

基本概念

- 大规模语料库的出现为自然语言统计处理方法的实现提供了可能，统计方法的成功使用推动了语料库语言学的发展。

基本概念

- 大规模语料库的出现为自然语言统计处理方法的实现提供了可能，统计方法的成功使用推动了语料库语言学的发展。
- 基于大规模语料库和统计方法，可以
 - 发现语言使用的普遍规律

基本概念

- 大规模语料库的出现为自然语言统计处理方法的实现提供了可能，统计方法的成功使用推动了语料库语言学的发展。
- 基于大规模语料库和统计方法，可以
 - 发现语言使用的普遍规律
 - 进行机器学习、自动获取语言知识

基本概念

- 大规模语料库的出现为自然语言统计处理方法的实现提供了可能，统计方法的成功使用推动了语料库语言学的发展。
- 基于大规模语料库和统计方法，可以
 - 发现语言使用的普遍规律
 - 进行机器学习、自动获取语言知识
 - 对未知语言现象进行推测

基本概念

如何计算一段文字(句子)的概率?

基本概念

如何计算一段文字(句子)的概率?

阳春三月春意盎然，少先队员脸上荡漾着喜悦的笑容，鲜艳的红领巾在他们的胸前迎风飘扬。

基本概念

如何计算一段文字(句子)的概率?

阳春三月春意盎然，少先队员脸上荡漾着喜悦的笑容，鲜艳的红领巾在他们的胸前迎风飘扬。

- 以一段文字(句子)为单位统计相对频率?
- 根据句子构成单位的概率计算联合概率?

$$p(w_1) \times p(w_2) \times \cdots \times p(w_n)$$

基本概念

如何计算一段文字(句子)的概率?

阳春三月春意盎然，少先队员脸上荡漾着喜悦的笑容，鲜艳的红领巾在他们的胸前迎风飘扬。

- 以一段文字(句子)为单位统计概率?
- 根据句子统计词联合概率?

$$p(w_1, w_2, \dots, w_n)$$

太简单

基本概念

语句 $s = w_1 w_2 \dots w_m$ 的先验概率：

$$\begin{aligned} p(s) &= p(w_1) \times p(w_2/w_1) \times p(w_3/w_1w_2) \times \dots \\ &\quad \times p(w_m/w_1 \dots w_{m-1}) \\ &= \prod_{i=1}^m p(w_i | w_1 \dots w_{i-1}) \end{aligned}$$

基本概念

语句 $s = w_1 w_2 \dots w_m$ 的先验概率：

$$\begin{aligned} p(s) &= p(w_1) \times p(w_2/w_1) \times p(w_3/w_1w_2) \times \dots \\ &\quad \times p(w_m/w_1 \dots w_{m-1}) \\ &= \prod_{i=1}^m p(w_i | w_1 \dots w_{i-1}) \end{aligned}$$

当 $i=1$ 时, $p(w_1|w_0) = p(w_1)$ 。

基本概念

语句 $s = w_1 w_2 \dots w_m$ 的先验概率：

$$\begin{aligned} p(s) &= p(w_1) \times p(w_2/w_1) \times p(w_3/w_1w_2) \times \dots \\ &\quad \times p(w_m/w_1 \dots w_{m-1}) \\ &= \prod_{i=1}^m p(w_i | w_1 \dots w_{i-1}) \end{aligned}$$

当 $i=1$ 时, $p(w_1|w_0) = p(w_1)$ 。

语言模型！！

基本概念

说明：

w_i 可以是字、词、短语或词类等等，称为统计基元。通常以“词”代之。

w_i 的概率由 w_1, \dots, w_{i-1} 决定，由特定的一组 w_1, \dots, w_{i-1} 构成的一个序列，称为 w_i 的 **历史**(history)。

基本概念

问题：随着历史基元数量的增加，不同的“历史”(路径)按指数级增长。对于第 i ($i > 1$) 个统计基元，历史基元的个数为 $i-1$ ，如果共有 L 个不同的基元，如词汇表，理论上每一个单词都有可能出现在1到 $i-1$ 的每一个位置上，那么， i 基元就有 L^{i-1} 种不同的历史情况。我们必须考虑在所有的 L^{i-1} 种不同历史情况下产生第 i 个基元的概率。

基本概念

□ 问题解决方法

设法减少历史基元的个数，将 $w_1 w_2 \dots w_{i-1}$ 映射到等价类 $S(w_1 w_2 \dots w_{i-1})$ ，使等价类的数目远远小于原来不同历史基元的数目。则有：

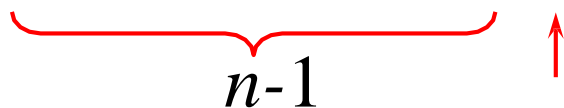
$$p(w_i | w_1, \dots, w_{i-1}) = p(w_i | S(w_1, \dots, w_{i-1}))$$

基本概念

□ 如何划分等价类

- 将两个历史映射到同一个等价类，当且仅当这两个历史中的最近 $n-1$ 个基元相同，即：

$$H_1: w_1 w_2 \dots \dots w_{i-n+1} w_{i-n+2} \dots w_{i-1} w_i \dots \dots$$


 $n-1$

$$H_2: v_1 v_2 \dots \dots v_{k-n+1} v_{k-n+2} \dots v_{k-1} v_k \dots \dots$$

$$S(w_1, w_2, \dots, w_i) = S(v_1, v_2, \dots, v_k)$$

$$\text{iff } H_1: (w_{i-n+1}, \dots, w_i) = H_2: (v_{k-n+1}, \dots, v_k)$$

基本概念

- 这种情况下的语言模型称为 n 元文法(n -gram)模型
- 通常地,
 - 当 $n=1$ 时, 即出现在第 i 位上的基元 w_i 独立于历史。
一元文法也被写为 uni-gram 或 monogram;

基本概念

- 这种情况下的语言模型称为 n 元文法(n -gram)模型
- 通常地,
 - 当 $n=1$ 时, 即出现在第 i 位上的基元 w_i 独立于历史。
一元文法也被写为 uni-gram 或 monogram;
 - 当 $n=2$ 时, 2-gram (bi-gram) 被称为1阶马尔可夫链;

基本概念

□ 这种情况下的语言模型称为 n 元文法(n -gram)模型

□ 通常地,

- 当 $n=1$ 时, 即出现在第 i 位上的基元 w_i 独立于历史。
一元文法也被写为 uni-gram 或 monogram;
- 当 $n=2$ 时, 2-gram (bi-gram) 被称为1阶马尔可夫链;
- 当 $n=3$ 时, 3-gram(tri-gram)被称为2阶马尔可夫链,
依次类推。

基本概念

为了保证条件概率在 $i=1$ 时有意义，同时为了保证句子内所有字符串的概率和为 1，即 $\sum_s p(s)=1$ ，可以在句子首尾两端增加两个标志：**<BOS>** $w_1 w_2 \dots w_m$ **<EOS>**。不失一般性，对于 $n>2$ 的 n -gram, $p(s)$ 可以分解为：

$$p(s) = \prod_{i=1}^{m+1} p(w_i | w_{i-n+1}^{i-1})$$

其中， w_i^j 表示词序列 $w_i \dots w_j$ ， w_{i-n+1} 从 w_0 开始， w_0 为 **<BOS>**， w_{m+1} 为 **<EOS>**。

基本概念

□ 举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

基本概念

□ 举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

Unigram: <BOS>, John, read, a, book, <EOS>

基本概念

□ 举例：

给定句子：John read a book

增加标记：<BOS> John read a book <EOS>

Unigram: <BOS>, John, read, a, book, <EOS>

Bigram: (<BOS>John), (John read), (read a), (a book), (book <EOS>)

基本概念

□ 举例：

给定句子： John read a book

增加标记： <BOS> John read a book <EOS>

Unigram: <BOS>, John, read, a, book, <EOS>

Bigram: (<BOS>John), (John read), (read a), (a book), (book <EOS>)

Trigram: (<BOS>John read), (John read a), (read a book), (a book <EOS>)

基本概念

<BOS> John read a book <EOS>

基于2元文法的概率为：

基本概念

<BOS> John read a book <EOS>

基于2元文法的概率为：

$$\begin{aligned} p(\text{John read a book}) &= p(\text{John}|\text{<BOS>}) \times \\ &\quad p(\text{read}|\text{John}) \times p(\text{a}|\text{read}) \times \\ &\quad p(\text{book}|\text{a}) \times p(\text{<EOS>}|\text{book}) \end{aligned}$$

基本概念

□ 应用-1: 音字转换问题

给定拼音串: ta shi yan jiu sheng wu de

基本概念

□ 应用-1: 音字转换问题

给定拼音串: ta shi yan jiu sheng wu de

可能的汉字串: 踏实研究生物的
他实验救生物的
他使烟酒生物的
他是研究生物的
... ..

基本概念

$$\begin{aligned}\hat{CString} &= \arg \max_{CString} p(CString | Pinyin) \\ &= \arg \max_{CString} \frac{p(Pinyin | CString) \times p(CString)}{p(Pinyin)} \\ &= \arg \max_{CString} p(Pinyin | CString) \times p(CString) \\ &= \arg \max_{CString} p(CString)\end{aligned}$$

基本概念

$CString = \{ \text{踏实研究生物的, 他实验救生物的, 他是研究生物的, 他使烟酒生雾的, ... } \}$

基本概念

$CString = \{\text{踏实研究生物的, 他实验救生物的, 他是研究生物的, 他使烟酒生雾的,}\}$

如果使用 2-gram:

$$\begin{aligned} p(CString_1) &= p(\text{踏实} | \langle BOS \rangle) \times p(\text{研究} | \text{踏实}) \times \\ &\quad p(\text{生物} | \text{研究}) \times p(\text{的} | \text{生物}) \times p(\langle EOS \rangle | \text{的}) \\ p(CString_2) &= p(\text{他} | \langle BOS \rangle) \times p(\text{实验} | \text{他}) \times p(\text{救} | \text{实验}) \times \\ &\quad p(\text{生物} | \text{救}) \times p(\text{的} | \text{生物}) \times p(\langle EOS \rangle | \text{的}) \\ &\dots\dots \end{aligned}$$

基本概念

如果汉字的总数为： N

- 一元语法：1) 样本空间为 N
- 2元语法：1) 样本空间为 N^2
2) 效果比一元语法明显提高
- 估计对汉字而言四元语法效果会好一些
- 智能狂拼、微软拼音输入法基于 n -gram.

基本概念

□ 应用-2：汉语分词问题

给定汉字串：他是研究生物的。

基本概念

□ 应用-2：汉语分词问题

给定汉字串：他是研究生物的。

可能的汉字串：

- 1) 他|是|研究生|物|的
- 2) 他|是|研究|生物|的

基本概念

$$\begin{aligned}\hat{Seg} &= \arg \max_{Seg} p(Seg | Text) \\ &= \arg \max_{Seg} \frac{p(Text | Seg) \times p(Seg)}{p(Text)} \\ &= \arg \max_{Seg} p(Text | Seg) \times p(Seg) \\ &= \arg \max_{Seg} p(Seg)\end{aligned}$$

基本概念

如果采用2元文法:

$$p(\text{Seg1}) = p(\text{他} | \langle \text{BOS} \rangle) \times p(\text{是} | \text{他}) \times p(\text{研究生} | \text{是}) \times \\ p(\text{物} | \text{研究生}) \times p(\text{的} | \text{物}) \times p(\text{的} | \langle \text{EOS} \rangle)$$

$$p(\text{Seg2}) = p(\text{他} | \langle \text{BOS} \rangle) \times p(\text{是} | \text{他}) \times p(\text{研究} | \text{是}) \times \\ p(\text{生物} | \text{研究}) \times p(\text{的} | \text{生物}) \times p(\text{的} | \langle \text{EOS} \rangle)$$

基本概念

如果采用2元文法:

$$p(\text{Seg1}) = p(\text{他} | \langle \text{BOS} \rangle) \times p(\text{是} | \text{他}) \times p(\text{研究生} | \text{是}) \times \\ p(\text{物} | \text{研究生}) \times p(\text{的} | \text{物}) \times p(\text{的} | \langle \text{EOS} \rangle)$$

$$p(\text{Seg2}) = p(\text{他} | \langle \text{BOS} \rangle) \times p(\text{是} | \text{他}) \times p(\text{研究} | \text{是}) \times \\ p(\text{生物} | \text{研究}) \times p(\text{的} | \text{生物}) \times p(\text{的} | \langle \text{EOS} \rangle)$$

问题：如何获得 n 元文法模型？

统计语言模型

1. 基本概念

2. 参数估计

3. 数据平滑

参数估计

□ 两个重要概念:

参数估计

□ 两个重要概念:

- 训练语料(training data)

用于建立模型确定模型参数的已知语料

参数估计

□ 两个重要概念:

- 训练语料(training data)

用于建立模型确定模型参数的已知语料

- 极大似然估计(MLE)

用相对频率计算概率的方法

参数估计

对于 n -gram, $p(w_i | w_{i-n+1}^{i-1})$ 可由最大参数似然估计求得:

$$p(w_i | w_{i-n+1}^{i-1}) = f(w_i | w_{i-n+1}^{i-1}) = \frac{c(w_{i-n+1}^i)}{\sum_{w_i} c(w_{i-n+1}^i)}$$

其中, $\sum_{w_i} c(w_{i-n+1}^i)$ 是历史串 w_{i-n+1}^{i-1} 在给定语料中出现的次数, 即 $c(w_{i-n+1}^{i-1})$, 不管 w_i 是什么。

$f(w_i | w_{i-n+1}^{i-1})$ 是在给定 w_{i-n+1}^{i-1} 的条件下 w_i 出现的相对频度, 分子为 w_{i-n+1}^{i-1} 与 w_i 同现的次数。

参数估计

例如，给定训练语料：

- a) “*John read Moby Dick*”,
- b) “*Mary read a different book*”,
- c) “*She read a book by Cher*”

根据 2 元文法求句子的概率？

参数估计

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

$$p(\text{John} | \langle \text{BOS} \rangle) = \frac{c(\langle \text{BOS} \rangle \text{John})}{\sum_w c(\langle \text{BOS} \rangle w)} = \frac{1}{3}$$

$$p(a | \text{read}) = \frac{c(\text{read } a)}{\sum_w c(\text{read } w)} = \frac{2}{3}$$

$$p(\text{read} | \text{John}) = \frac{c(\text{John } \text{read})}{\sum_w c(\text{John } w)} = \frac{1}{1}$$

$$p(\text{book} | a) = \frac{c(a \text{ book})}{\sum_w c(a w)} = \frac{1}{2}$$

$$p(\langle \text{EOS} \rangle | \text{book}) = \frac{c(\text{book } \langle \text{EOS} \rangle)}{\sum_w c(\text{book } w)} = \frac{1}{2}$$

$$p(\text{John read a book}) = \frac{1}{3} \times 1 \times \frac{2}{3} \times \frac{1}{2} \times \frac{1}{2} \approx 0.06$$

参数估计

$$p(\textit{Cher read a book}) = ?$$

参数估计

$$p(\textit{Cher read a book}) = ?$$

$$= p(\textit{Cher} | \langle \textit{BOS} \rangle) \times p(\textit{read} | \textit{Cher}) \times p(\textit{a} | \textit{read}) \times \\ p(\textit{book} | \textit{a}) \times p(\langle \textit{EOS} \rangle | \textit{book})$$

$$p(\textit{Cher} | \langle \textit{BOS} \rangle) = \frac{c(\langle \textit{BOS} \rangle \textit{Cher})}{\sum_w c(\langle \textit{BOS} \rangle w)} = \frac{0}{3}$$

$$p(\textit{read} | \textit{Cher}) = \frac{c(\textit{Cher} \textit{read})}{\sum_w c(\textit{Cher} w)} = \frac{0}{1}$$

于是, $p(\textit{Cher read a book}) = 0$



参数估计

问题：

数据匮乏(稀疏) (*Sparse Data*) 引起零概率问题，如何解决？

参数估计

问题：

数据匮乏(稀疏) (*Sparse Data*) 引起零概率问题，如何解决？

数据平滑(data smoothing)

统计语言模型

1. 基本概念

2. 参数估计

3. 数据平滑

数据平滑

□ 数据平滑的基本思想：

调整最大似然估计的概率值,使零概率增值,使非零概率下调,消除零概率,改进模型的整体正确率

□ 基本目标：

测试样本的语言模型 困惑度(Perplexity)越小越好

□ 基本约束： $$\sum_{w_i} p(w_i | w_1, w_2, \dots, w_{i-1}) = 1$$

困惑度

➤ 困惑度的定义：

对于一个平滑的 n -gram，其概率为 $p(w_i | w_{i-n+1}^{i-1})$ ，

可以计算句子的概率：
$$p(s) = \prod_{i=1}^{m+1} p(w_i | w_{i-n+1}^{i-1})$$

假定测试语料 T 由 L 个句子构成 (t_1, \dots, t_L) ，则整个测试集的概率为：

$$p(T) = \prod_{i=1}^L p(t_i)$$

困惑度

➤ 困惑度的定义:

$$\begin{aligned} PP(S) &= P(w_1 w_2 \dots w_N)^{-\frac{1}{N}} \\ &= \sqrt[N]{\frac{1}{p(w_1 w_2 \dots w_N)}} \\ &= \sqrt[N]{\prod_{i=1}^N \frac{1}{p(w_i | w_1 w_2 \dots w_{i-1})}} \end{aligned}$$

数据平滑

□ 数据平滑方法

1) 加1法(Additive smoothing)

基本思想: 每一种情况出现的次数加1。

例如, 对于 *uni-gram*, 设 w_1, w_2, w_3 三个词, 概率分别为: $1/3, 0, 2/3$, 加1后情况?

数据平滑

□ 数据平滑方法

1) 加1法(Additive smoothing)

基本思想: 每一种情况出现的次数加1。

例如, 对于 *uni-gram*, 设 w_1, w_2, w_3 三个词, 概率分别为: $1/3, 0, 2/3$, 加1后情况?

$2/6, 1/6, 3/6$

数据平滑

对于2-gram 有：

$$\begin{aligned} p(w_i | w_{i-1}) &= \frac{1 + c(w_{i-1}w_i)}{\sum_{w_i} [1 + c(w_{i-1}w_i)]} \\ &= \frac{1 + c(w_{i-1}w_i)}{|V| + \sum_{w_i} c(w_{i-1}w_i)} \end{aligned}$$

其中， V 为被考虑语料的词汇量(全部可能的基元数)。

数据平滑

在前面 3 个句子的例子中,

$$\begin{aligned} p(\textit{Cher read a book}) &= p(\textit{Cher}|\langle\textit{BOS}\rangle) \times \\ &\quad p(\textit{read}|\textit{Cher}) \times p(\textit{a}|\textit{read}) \times p(\textit{book}|\textit{a}) \times \\ &\quad p(\langle\textit{EOS}\rangle|\textit{book}) \end{aligned}$$

$\langle\textit{BOS}\rangle\textit{John read Moby Dick}\langle\textit{EOS}\rangle$

$\langle\textit{BOS}\rangle\textit{Mary read a different book}\langle\textit{EOS}\rangle$

$\langle\textit{BOS}\rangle\textit{She read a book by Cher}\langle\textit{EOS}\rangle$

数据平滑

原来:

$$p(\textit{Cher}|\textit{<BOS>}) = 0/3$$

$$p(\textit{read}|\textit{Cher}) = 0/1$$

$$p(a|\textit{read}) = 2/3$$

$$p(\textit{book}|a) = 1/2$$

$$p(\textit{<EOS>}|\textit{book}) = 1/2$$

数据平滑

词汇量： $|V|=11$

<BOS>John read Moby Dick<EOS>

<BOS>Mary read a different book<EOS>

<BOS>She read a book by Cher<EOS>

平滑以后：

$$p(\text{Cher}|\text{<BOS>}) = (0+1)/(11+3) = 1/14$$

$$p(\text{read}|\text{Cher}) = (0+1)/(11+1) = 1/12$$

$$p(a|\text{read}) = (1+2)/(11+3) = 3/14$$

$$p(\text{book}|a) = (1+1)/(11+2) = 2/13$$

$$p(\text{<EOS>}|\text{book}) = (1+1)/(11+2) = 2/13$$

$$p(\text{Cher read a book}) = \frac{1}{14} \times \frac{1}{12} \times \frac{3}{14} \times \frac{2}{13} \times \frac{2}{13} \approx 0.00003$$

2) 减值法/折扣法(Discounting)

基本思想：修改训练样本中事件的实际计数，使样本中(实际出现的)不同事件的概率之和小于1，剩余的概率量分配给未见概率。

数据平滑

a) 绝对减值法 (Absolute discounting)

- Hermann Ney 和 U. Essen 1993年提出。
- 基本思想：从每个计数 r 中减去同样的量，剩余的概
率量由未见事件均分。
- 设 R 为所有可能事件的数目(当事件为 n -gram 时，如
果统计基元为词，且词汇集的大小为 L ，则 $R=L^n$)。

数据平滑

那么，样本出现了 r 次的事件的概率可以由如下公式估计：

$$p_r = \begin{cases} \frac{r-b}{N} & \text{当 } r > 0 \\ \frac{b(R-n_0)}{Nn_0} & \text{当 } r = 0 \end{cases}$$

其中， n_0 为样本中未出现的事件的数目。 b 为减去的常量， $b \leq 1$ 。 $b(R - n_0)/N$ 是由于减值而产生的剩余概率量。

Thank you!

权小军 中山大学数据科学与计算机学院