New Media
Data Analytics and Application

Lecture 9: Basic Statistics for Natural Language Processing

Ting Wang

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Outlines

- The Foundation of Statistics
- Bayes' Theorem
- Markov Model
- N-gram
- Chinese Word Segmentation

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

introduce some basic statistical metrics to you

The Foundation of Statistics

Average 平均数

Average

- *Mean* 均值
- *Median* 中位数
- *Mode* 众数

The Foundation of Statistics

Mean 均值

Supposing: $X=(x_1, x_2, \ldots, x_n)$

$$\bar{X} = \frac{\sum X}{n}$$

Median 中位数

the value separating the higher half of a data sample, a population, or a probability distribution, from the lower half.

1, 3, 3, **6**, 7, 8, 9

Median = 6

1, 2, 3, **4**, **5**, 6, 8, 9

Median = (4 + 5) ÷ 2

= 4.5

Supposing: $X=(x_1, x_2, \ldots, x_n)$

Sort X from small number to large number,

–if n is an odd number, then the Median of X is the middle one,

–if n is an even number, then the Median of X is the **mean** of the two middle numbers.

Mode 众数

the value that appears most often in a set of data

Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

Type	Description	Example	Result
Arithmetic mean	Sum of values of a data set divided by number of values: $\bar{x}=\frac{1}{n}\sum_{i=1}^{n} x_i$	(1+2+2+3+4+7+9) / 7	**4**
Median	Middle value separating the greater and lesser halves of a data set	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	Most frequent value in a data set	1, **2**, **2**, 3, 4, 7, 9	**2**

Range 极差

the difference between the largest and smallest values

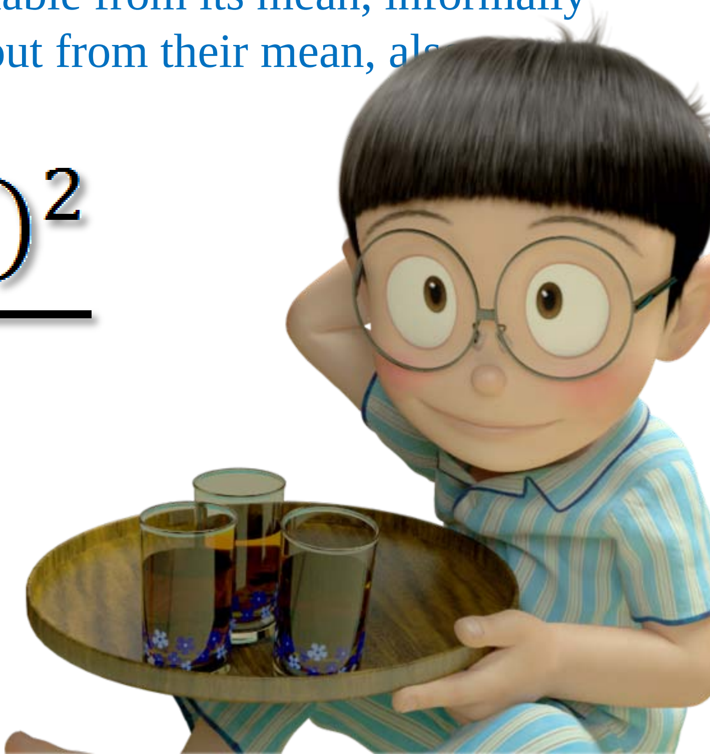
$$r = Max - Min$$

Variance 方差

the expectation of the squared deviation of a random variable from its mean, informally measures how far a set of (random) numbers are spread out from their mean, als known as $D(X)$, $Var(X)$

$$s^2 = \frac{\sum (X - \bar{X})^2}{n-1}$$

Why n-1?



Standard Deviation 标准差

$$s = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

Expected Value 数学期望

$$E[X] = \bar{X} = \sum_{i=1}^{n} x_i P_i$$

Where: $\boldsymbol{P_i}$ is the weight of $\boldsymbol{x}_i$

in Statistics, $\boldsymbol{P}$ is the probability.

Properties of Expected Value

- If C is a constant, $E[C]=C$
- If X and Y are random variables such that $X \leq Y$, then $E[X] \leq E[Y]$
- $E[X+C]=E[X]+C$
- $E[X+Y]=E[X]+E[Y]$
- $E[CX]=CE[X]$
- $D[X]=E[X^2]-(E[X])^2$

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

very useful for natural language processing

Bayes’ Theorem

Probability 概率

$P(x_i)=1/6$

$P(x_i)=1/2$

Sample Space：

{1，2，3，4，5，6}

{H，T}

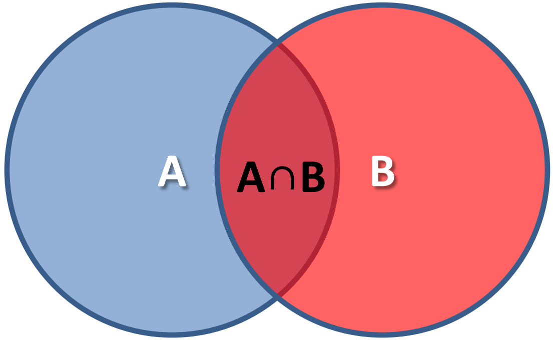
Properties of Probability

$$P(x_i) \geq 0$$

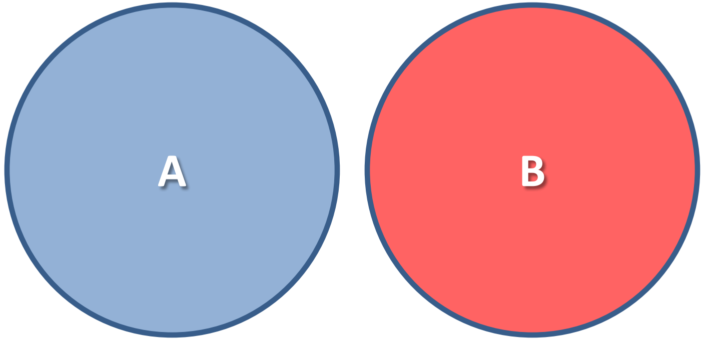
$$P(x_i) \in [0,1]$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

Independence 独立性



Dependent



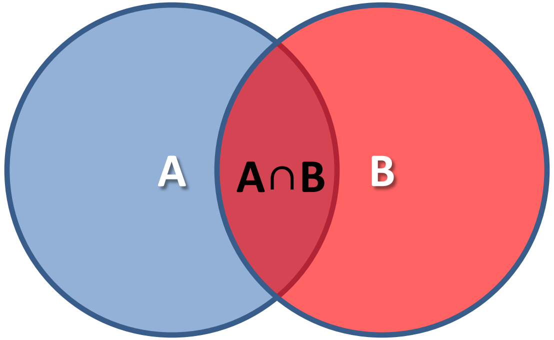
Independent

Bayes' Theorem

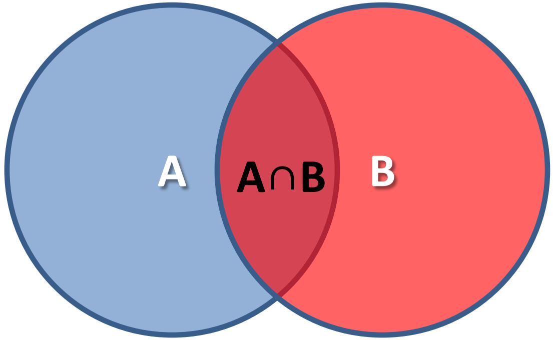
Conditional Probability 条件概率

P(A | B), is the probability of observing event A given that B is true

$$P(A|B) = P(A \cap B)/P(B)$$



Bayes’ Theorem 贝叶斯定理



$$P(A|B) = P(A \cap B)/P(B)$$

$$P(A \cap B) = P(A|B)P(B)$$

$$P(A \cap B) = P(B|A)P(A)$$

$$P(A|B)P(B) = P(B|A)P(A)$$

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes' Theorem

Bayes' Theorem plays an very important role in statistical NLP.

- We can predict what you will say!
  - **Uncle Sam:** How are you?
  - **Chinese student:** Fine, Thank you, and you?
  - **Chinese student's Predictive Answer:** I am fine, too!
  - **Uncle Sam:** Nothing much.
  - **Chinese student:**。。。（不多？？）

Bayes' Theorem

- Because, for Chinese students:

P(Fine, Thank you, and you? | How are you?) $\nearrow$

P(I am fine, too! | Fine, Thank you, and you?) $\nearrow$

P(Nothing much | Fine, Thank you, and you?) $\searrow$

In the corpus of Chinese students,

P(I am fine, too! | Fine, Thank you, and you?)>P(Nothing much | Fine, Thank you, and you?)

Another Example:

I ate a red _________ .

A. telephone   B. light   C. swim   D. tomato

No Grammar! But the Frequency of use!

- The most successful Chinglish:

Long time no see!

- Chinglish Future Star:

Good Good Study, Day Day UP!

your future is decided by now, not the past

Markov Model

Stochastic Process 随机过程

Markov Chain 马尔科夫链

$$X=(x_1, x_2, \ldots, x_n)$$

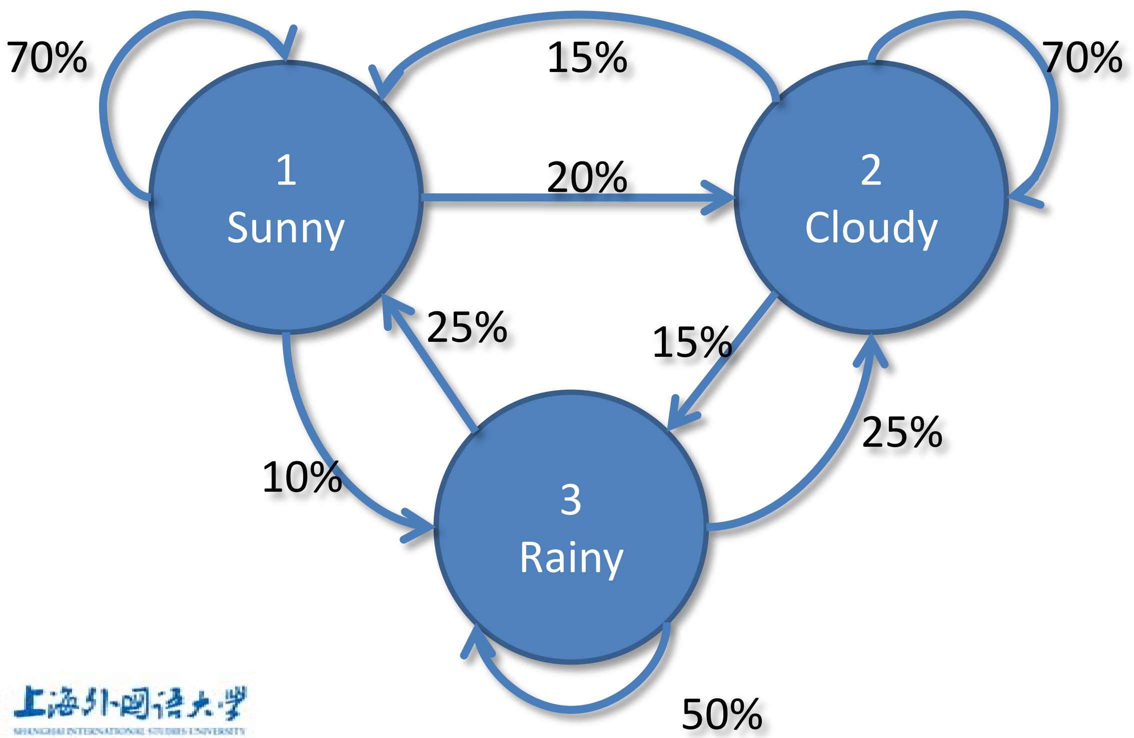
x_i is a Stochastic Process

1,3,5,2,1,4,2,6,3,……

X is a Markov Chain

Markov Model

Transition Probability 转移概率



$$\begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix} =$$

$$\begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.15 & 0.7 & 0.15 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

Stochastic Matrix

概率转移矩阵

出度之和100%

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

Markov Model 马尔科夫模型

$$P(x_{t+1}|x_1, x_2, \cdots, x_t) = P(x_{t+1}|x_t)$$

First-Order Markov Model

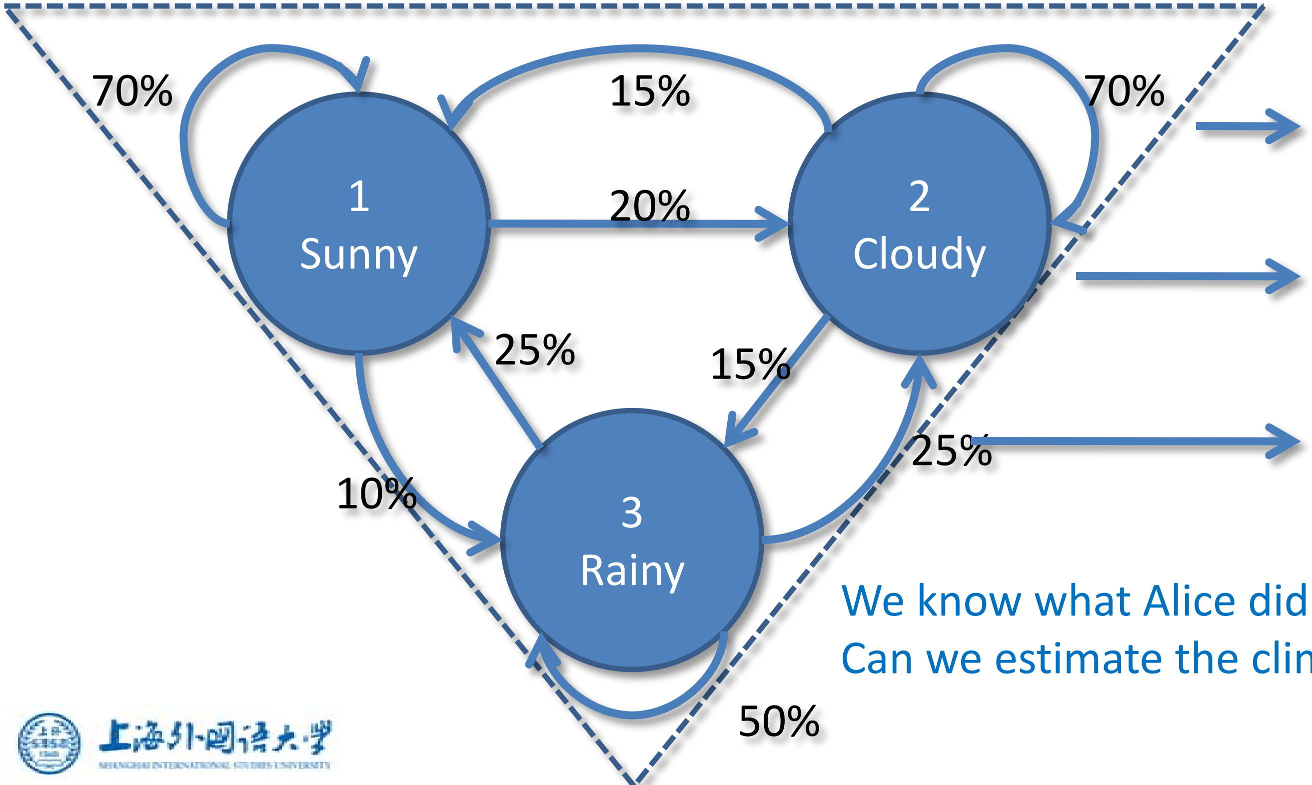
Your future is not decided by your past, but now!

Second-Order Markov Model

$$P(x_{t+1}|x_1, x_2, \cdots, x_t) = P(x_{t+1}|x_t x_{t-1})$$

Markov Model

Hidden Markov Model 隐马尔科夫模型



Day 1 : Shopping

Day 2 : Cleaning

Day 3 : Camping

We know what Alice did in the past three days.
Can we estimate the climate of the place Alice lived in?

The Applications of Markov Model in NLP

- Machine Translation
- Word Segmentation
- Speech Recognition
- Part-of-speech Tagging
- Natural Language Generation
- …

one of the most important statistical computational linguistic models

N-gram

Definition of N-gram N元文法

An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a (n − 1)-order Markov model.

N	N-gram	(N – 1)-order Markov model	Example
1	1-gram(unigram)	Independent from history	One Word
2	2-gram(bigram)	1-order (HMM-1)	Two Words
3	3-gram(trigram)	2-order (HMM-2)	Three Words
…	…	…	…

Unigram *上下文无关文法*

- Only consider the probability of the word itself
- Hypothesis: Every word is independent.

$$P(X) = P(x_1, x_2, \cdots, x_N) = \prod_{i=1}^{N} P(x_i)$$

$$P(x_i) = \frac{Number\ of\ x_i\ in\ the\ artical}{Number\ of\ all\ words\ in\ the\ artical}$$

Bigram 二元文法

The current word is influenced by the previous one word

$$P(X) = P(x_1, x_2, \cdots, x_N) = P(x_1)P(x_2|x_1)P(x_3|x_2)\cdots P(x_N|x_{N-1})$$

$$= P(x_1)\prod_{i=2}^{N} P(x_i|x_{i-1})$$

$$P(x_i|x_{i-1}) = \frac{Number\ of\ (x_{i-1}x_i)\ in\ the\ artical}{Number\ of\ all\ x_{i-1}\ in\ the\ artical}$$

Trigram 三元文法

The current word is influenced by the previous two words

$$P(X) = P(x_1, x_2, \cdots, x_N) = P(x_1)P(x_2|x_1)P(x_3|x_2x_1)P(x_4|x_3x_2)\cdots P(x_N|x_{N-1}x_{N-2})$$

$$= P(x_1)P(x_2|x_1)\prod_{i=3}^{N} P(x_i|x_{i-1}x_{i-2})$$

$$P(x_i|x_{i-1}x_{i-2}) = \frac{Number\ of\ (x_{i-2}x_{i-1}x_i)\ in\ the\ artical}{Number\ of\ all\ (x_{i-2}x_{i-1})\ in\ the\ artical}$$

Tips

1. Previous studies showed that trigram and four-gram often have better performance

2. The larger of *N*, the more complex of the computation

3. N-gram needs training data set, while it is impossible for a training data set to contain all the matches of a word

Smoothing 平滑

- Zero Probability 零概率
- Small Probability 小概率
- Laplace Smoothing 拉普拉斯平滑

$$P(x_i|x_1,x_2,\cdots,x_{i-1}) = \frac{Number\ of\ (x_1 \ldots x_i) in\ the\ artical + 1}{Number\ of\ all (x_1 \ldots x_{i-1}) in\ the\ artical + Number\ of\ words\ in\ dictionary}$$

Commonly used Smoothing Approaches

- Linear interpolation (e.g., taking the weighted mean of the unigram, bigram, and trigram)
- Good–Turing discounting
- Witten–Bell discounting
- Lidstone's smoothing
- Katz's back-off model (trigram)
- Kneser–Ney smoothing

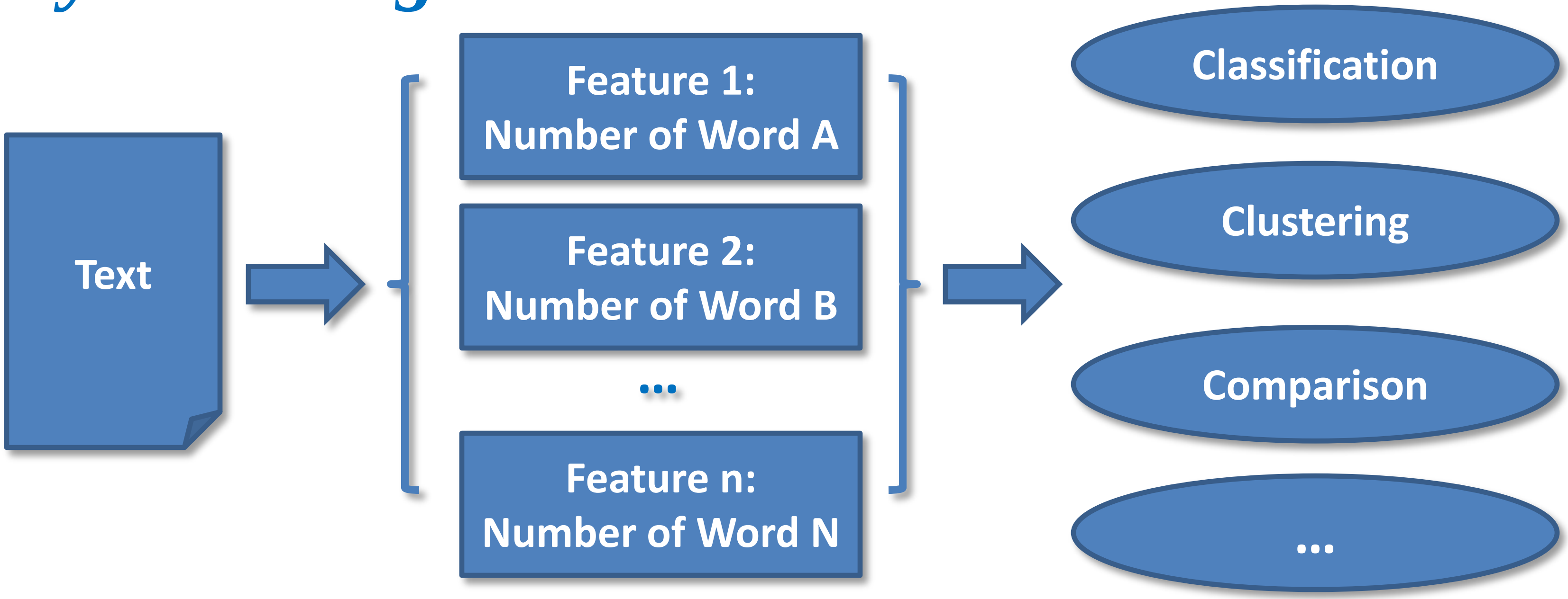
Ref. https://en.wikipedia.org/wiki/N-gram

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

the first step for Chinese information processing

Chinese Word Segmentation

Why Word Segmentation?



However, it is difficult to extract words from Chinese text.

Difficulties: Disambiguation

乒乓球拍卖完了

乒乓|球拍|卖完了

乒乓球|拍卖|完了

Forward Max. matching method, FMM

正向最大匹配

准备工作：需要分词词典D

设MaxLen表示最大词长度

算法：

1. 从生语料N中取长度为MaxLen的字串str，令Len= MaxLen
2. 把str与D中的词相匹配
3. 若匹配成功，则认为str为词，N中去掉str(指针前移Len个单位)，返回1
4. 若匹配不成功，

◆ 若Len>1则Len--，从生语料N中取长度为Len的字串str返回2；

◆ 否则，得到单字词， N中去掉str(指针前移1个单位)，返回1

若4中得到的单字不是词，则要进行未登录词处理

若待切分的语料字串长度小于MaxLen，则取str为待切分语料

Chinese Word Segmentation

Backward Max. matching method, BMM

逆向最大匹配

1. Similar to FMM, but the text is scanned from the right side
2. Often jointly use with FMM

Chinese Word Segmentation

- **Statistical Matching Method**

```
FMM and BMM
Begin initialize Path←{},AmbiguousString,SubString←{}
   While (AmbiguousString.Length>0)
   {
       //只考虑以当前HMM第一个状态开始的匹配序列
       SubString←以AmbiguousString中的第一个字为基准，取出所有可能的匹配字符串
       Foreach SubString
       {
          //提供当前情况下所有的概率，为判断歧义作参考
              计算当前每一种可能情况的概率P(SubString)  //unigram, bigram, trigram with smoothing
       }
       //选择概率最大的SubString添加到Path
       将argmax(P(SubString))添加到Path
       //准备考察除去最大概率的SubString后的AmbiguousString，从HMM序列首部开始，除去所有的匹配状态
       AmbiguousString.Remove(0, argmax(P(SubString)).Length)
    }
    Return Path
End
```

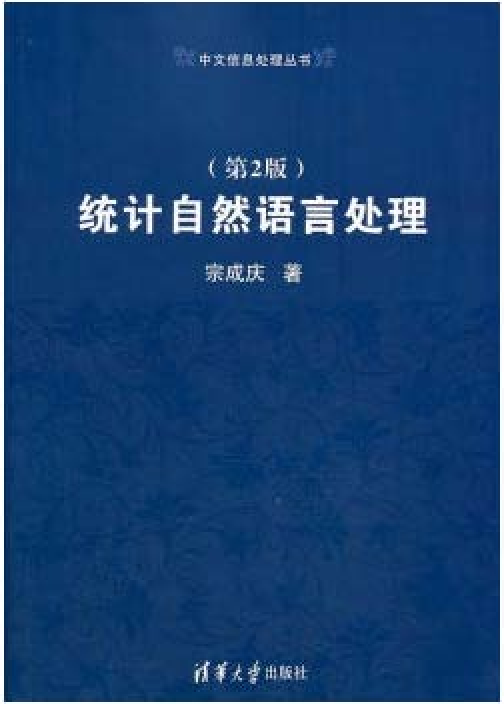
Reference

Reference

- https://item.jd.com/11701113.html

Reference

- https://item.jd.com/1040675628.html



Homework

Homework

- Data Collection for your group.
- Try your best to write a Chinese word segmentation algorithm and run it.
- How work will be presented group by group on Dec. 21 and report should be handed before Jan. 6.

The End of Lecture 9

Thank You

http://www.wangting.ac.cn