上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

New Media
Data Analytics and Application

Lecture 8: Quantitative Analysis for Online Journalism

Ting Wang

Outlines

1. Data Analysis for Online Journalism
2. The Foundation of Statistics
3. Pearson Correlation Coefficient
4. Bayes' Theorem
5. Markov Model

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

some analysis approaches for data journalism

Data Analysis for Online Journalism

Data Analysis for Online Journalism

Now, We have data.

What shall we do in the next step for online journalism analysis?

Data Analysis for Online Journalism

1. How to extract information from data?

Data Clean and Preprocessing

2. How to measure the information of news?

Quantitative Modules

3. How to analysis these information?

Comparison, Classification and Clustering

Data Analysis for Online Journalism

EXAMPLE 1:
papi酱

围脖关键词

围脖关键词利用自然语言处理的关键词抽取技术，分析用户近期发表微博内容，提取代表用户兴趣的关键词，并采用文档可视化技术对关键词进行可视化，便于用户快速了解自己、好友、主题等的关键词。

使用以下账号登录

http://app.thunlp.org/

Data Analysis for Online Journalism

April, 2016

papi酱

November, 2016

Data Analysis for Online Journalism

罗辑思维退出papi酱令人深思 网红经济不行了吗？

3685 我要评论 2016-11-25 15:23 来源：新京报

罗辑思维退出papi酱，网红经济不行了吗？

对个人形象的克制使用和保持健康，才是网红经济不成为一锤子买卖的关键。

据报道，罗辑思维已与著名网红papi酱分手。记者调查发现，早在今年8月29日，papi酱所在的公司春雨听雷在股东一栏里就去掉了罗辑思维的投资经营主体北京思维造物投资管理有限公司。

罗辑思维退出papi酱的背后：互联网风口关闭

2016年11月26日 09:59 创事记 微博 作者： maomaobear 我有话说(5人参与)

+订阅

欢迎关注“创事记”的微信订阅号：sinachuangshiji

Data Analysis for Online Journalism

Technical Approaches

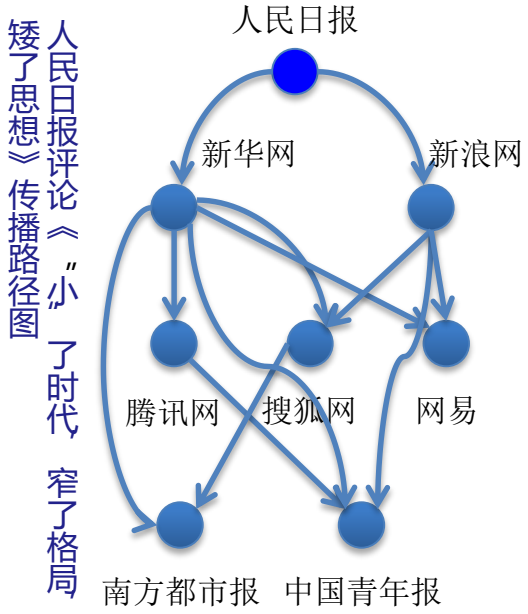
1. Keyword Extraction and Tag Analysis
2. News Tracking
3. News Alignment and Comparison
4. Location-based News Analysis
5. …

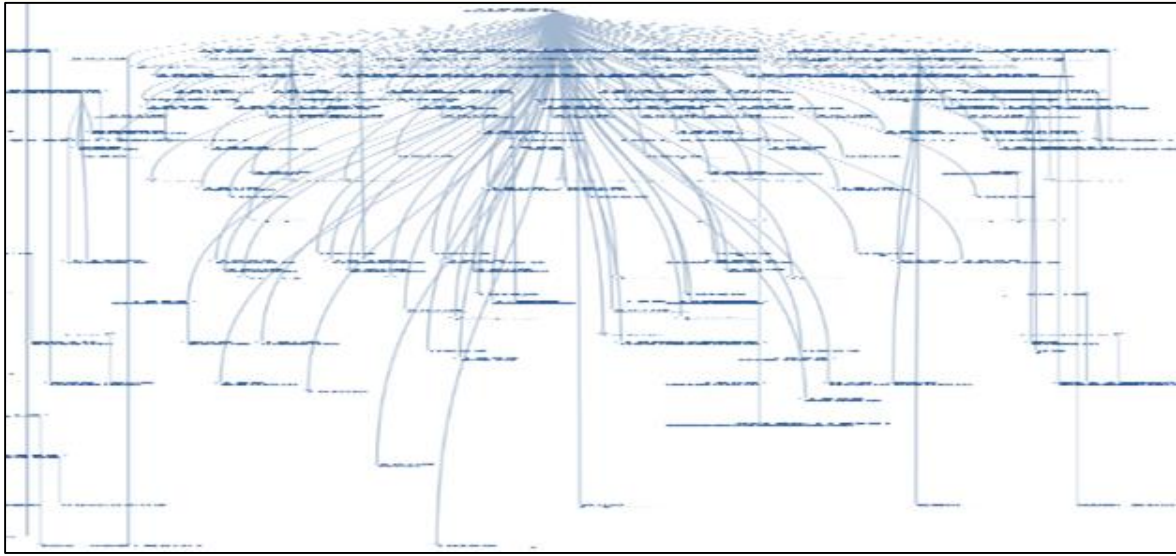
Data Analysis for Online Journalism

1. Keyword Extraction and Tag Analysis

Data Analysis for Online Journalism

2. News Tracking





Communication Efficiency Analysis base on Big Data

媒体	传播效率
新华网	71.67
新浪网	31.0
搜狐网	16.2
中国青年报	9.22
南方都市报	6.41

Data Analysis for Online Journalism

3. News Alignment and Comparison

韩国推迟与日本签署军事协定 被指外交失礼

2012年06月29日 17:37:31
来源： 中国新闻网

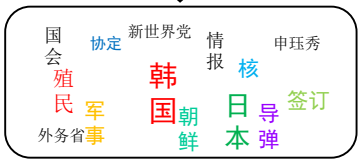
【字号：大 中 小】【打印】
【纠错】

据韩联社报道，韩国政府决定推迟原定在当地时间29日下午4点签署《韩日军事情报综合保护协定》(GSOMIA)。

韩国外交消息人士表示，根据朝野要求，决定在正式签署前先向国会进行说明。今后的日程不得而知。

韩国政府秘密推进签署军事情报协定引发了舆论非议，因此新世界党要求政府推迟或取消签署协定。这虽然属于外交失礼，但韩国政府接受了新世界党的提议。韩国驻日大使申珏秀则向日本外务省转达了韩国政府的立场。

《韩日军事秘密保护协定》将是韩国摆脱日本殖民统治后与日本之间签订的第一个军事协定。如果该协议签订，韩日今后有望互通有关朝鲜军队、朝鲜社会动向、朝核以及导弹问题等方面的情报。（中新网6月29日电）



South Korea to Sign Military Pact With Japan

By CHOE SANG-HUN
Published: June 28, 2012

SEOUL, South Korea — In a significant step toward overcoming lingering historical animosities with its former colonial master, the South Korean government has unexpectedly announced that it will sign a treaty with Japan on Friday to increase the sharing of classified military data on what analysts cite as two major common concerns: North Korea's nuclear and missile threats and China's growing military might.

FACEBOOK
TWITTER
GOOGLE+
EMAIL
SHARE
PRINT
REPRINTS

RUBY SPARKS COMING SOON

Connect With Us on Twitter
Follow @nytimesworld for international breaking news and headlines.
Twitter List: Reporters and Editors

World

The announcement set off a political firestorm in South Korea, where resentment of Japan's early 20th-century colonization remains entrenched and any sign of Japan's growing military role is met with deep suspicion. The opposition accused President Lee Myung-bak of ignoring popular anti-Japanese sentiments in pressing ahead with the treaty, the first military pact between the two nations since the end of colonization in 1945.

Between Different Languages, Websites, Nations, and People

Notes: 1. Words in the same color have the same meaning in translation
2. The size of the word represents the importance of the word, the larger, the more important

4. Location-based News Analysis

(1). Sentiment Analysis
(2). Trend Analysis

大陆：钓鱼岛事件波及日本在华大型企业。9月17日，包括佳能、松下和狮王等在内的企业相继传来工厂生产暂停的消息。另据消息称，9月18日，华堂商场、优衣库、7-11等日资零售企业也在这两天暂时停业。相关企业表示，此举是为了保障员工的安全。记者昨晚从国防部新闻事务局有关负责人获悉，美国国防部长莱昂·帕内塔已于昨晚抵达北京，开始对中国进行为期4天的正式访问。受钓鱼岛事件影响，近日中国多城市出现反日示威游行。

华人：惠誉国际评级周二表示，该机构认为，如果中日钓鱼岛争端升级并持续，日本大型汽车制造商和科技制造公司的评级可能将承压。中日钓鱼岛争议持续升温，中国各地周末爆发反日示威游行。受此影响，日本多家汽车生产企业决定暂停在中国大陆工厂的生产。中国国防部长梁光烈周二上午与到访的美国国防部长帕内塔举行会谈，在中日爆发钓鱼岛争端之际，帕内塔访华备受外界关注。

外媒：The 81st anniversary of a Japanese invasion brought a fresh wave of anti-Japan demonstrations in China on Tuesday, with thousands of protesters venting anger over the colonial past and a current dispute involving contested islands in the East China Sea.China Warns of Further Actions as Anti-Japan Protests Resume.At least two of 11 Chinese ocean surveillance and fishery patrol ships sailing near East China Sea islets claimed by both Tokyo and Beijing have entered what Japan considers its territory.

Data Analysis for Online Journalism

EXAMPLE 2:
Under the Dome

The influence by *Under the Dome*, made by Jing Chai, February 28, 2015

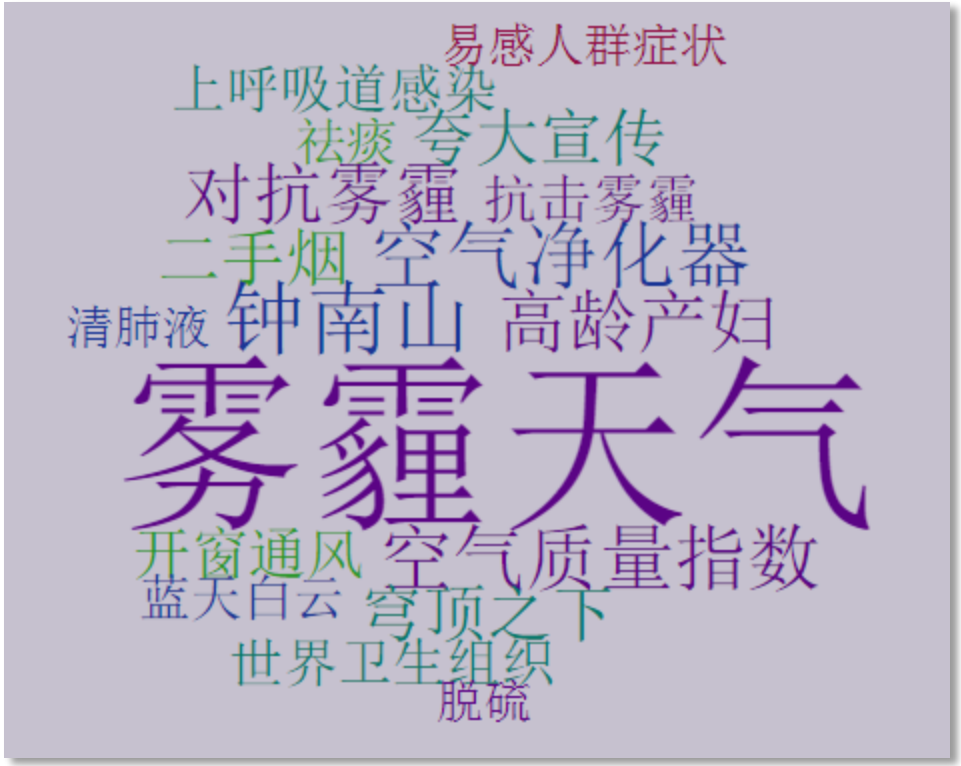
Data Description

Keyword Category	Keywords	Number of Weibo	Number of Weibo without repetition
呼吸系统疾病	呼吸系统疾病支气管炎　哮喘　咳嗽感冒胃肠型感冒、咽炎、支气管肺炎、上呼吸道感染、尘肺、结核病、鼻炎、咽喉炎、鼻窦炎、扁桃体炎	11321	6648
肿瘤	肿瘤新生儿肿瘤肺癌肺肿瘤	10655	7005
汽油	汽油质量	16	14
发电	发电、电力行业	1620	1032
净化	净化器、清新剂	9887	4770
煤	燃煤、煤炭	6587	4613
	总　计	40086	24082

Data Analysis for Online Journalism

Keyword Extraction Based on Weibo

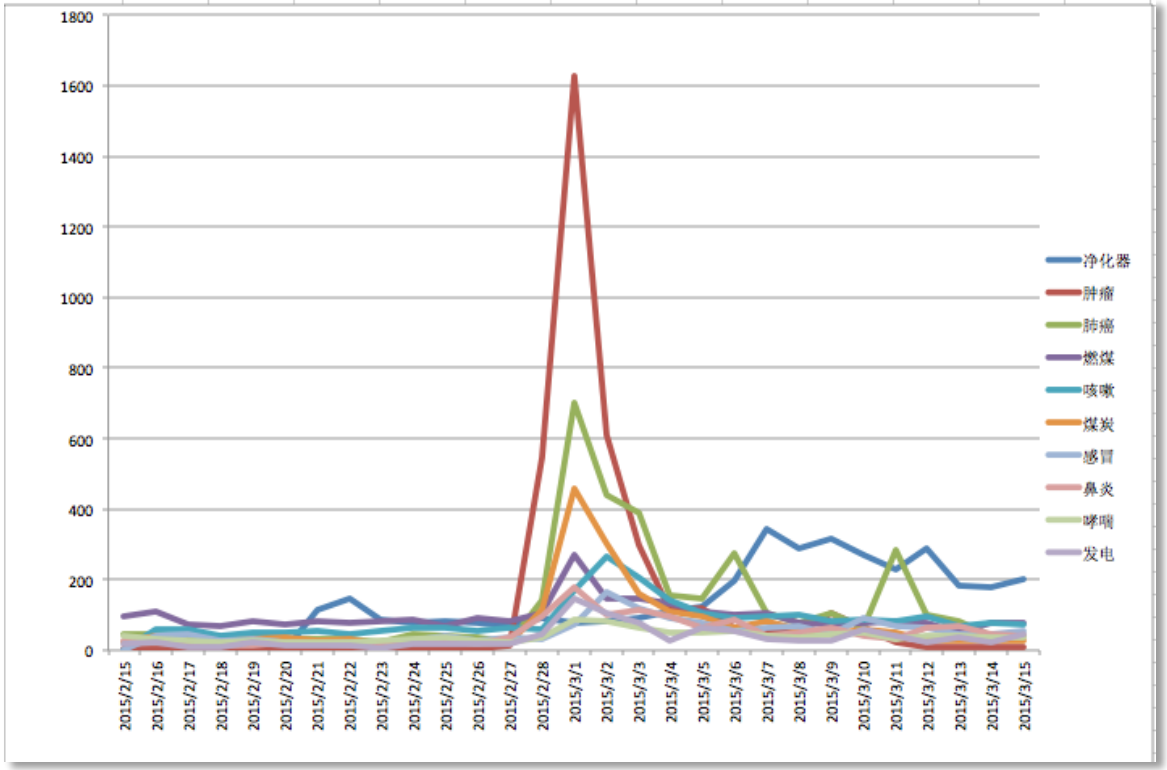
- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Keyword Analysis Based on Weibo

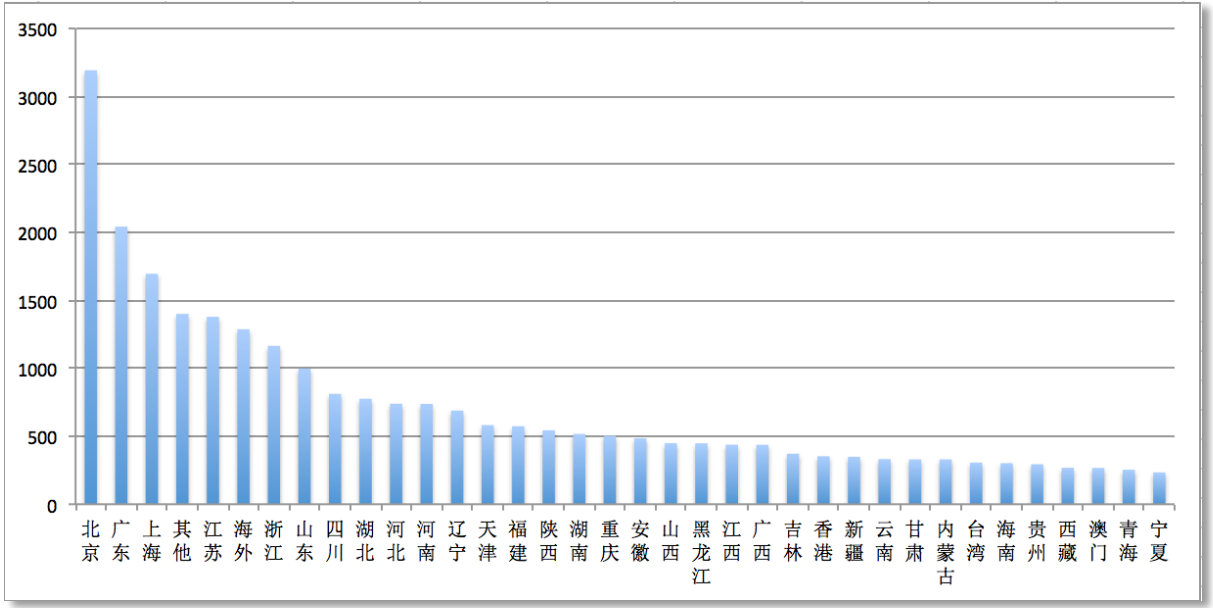
- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Keyword Analysis Based on Weibo

- What can you find in this graph?
- What will you do for your group?



Data Analysis for Online Journalism

Conclusions:

- Keyword is an abstract of online media
- The frequency of using keywords is important

Data Analysis for Online Journalism

Correlative Scientific Technologies

Natural Language Processing

Statistics ***Machine Learning***

ARTIFICIAL INTELLIGENCE

Machine Translation **Psychology**

Computer Science **Linguistics**

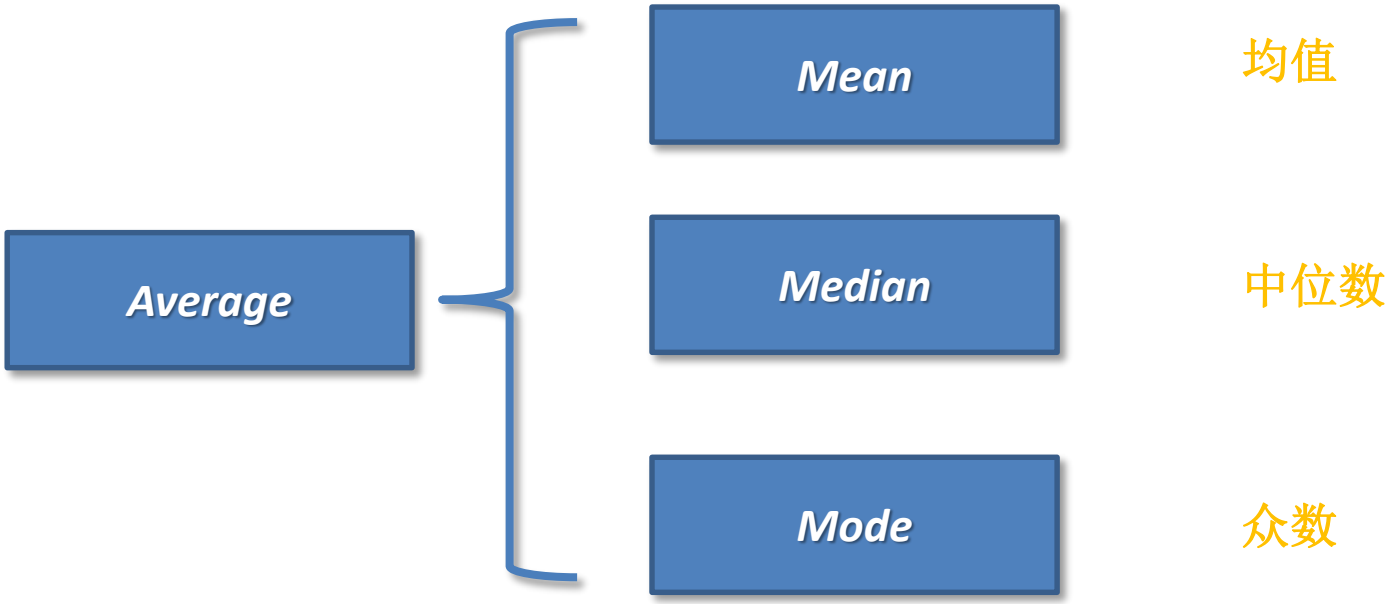
上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

introduce some basic statistical metrics to you

The Foundation of Statistics

The Foundation of Statistics

Average 平均数



The Foundation of Statistics

Mean 均值

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

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

The Foundation of Statistics

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

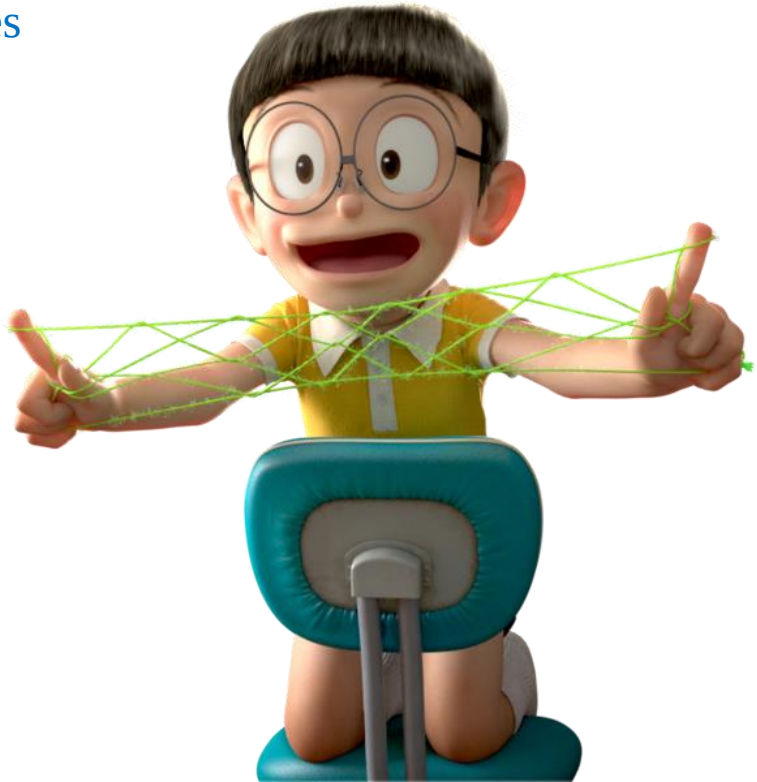
Comparison of common averages of values { 1, 2, 2, 3, 4, 7, 9 }

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

$$r = Max - Min$$



The Foundation of Statistics

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, also 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: P_i is the weight of x_i

in Statistics, 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 $E[X] \leq E[Y]$
- $E[X+C]=E[X]+C$
- $E[X+Y]=E[X]+E[Y]$
- $E[CX]=CE[X]$
- $D[X]=E[X^2]-(E[X])^2$

Covariance 协方差

a measure of the joint variability of two random variables

$$\begin{aligned} Cov(X,Y) &= E[(X-E[X])(Y-E[Y])] \\ &= E[XY]-2E[Y]E[X]+E[X]E[Y] \\ &= E[XY]-E[X]E[Y] \end{aligned}$$

Cov(X，X)=D(X)，Cov(Y，Y)=D(Y)

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

Properties of Covariance

$$D(X+Y)=D(X)+D(Y)+2Cov(X，Y)$$
$$D(X-Y)=D(X)+D(Y)-2Cov(X，Y)$$

$$Cov(X，Y)=E(XY)-E(X)E(Y)$$

$$Cov(X，Y)=Cov(Y，X)$$
$$Cov(aX，bY)=abCov(X，Y)$$
$$Cov(X_1+X_2，Y)=Cov(X_1，Y)+Cov(X_2，Y)$$

Uncorrelatedness and independence

- If X and Y are independent, then their covariance is 0.

$$E[XY]=E[X]E[Y]$$

- The converse, however, is not generally true.

The Foundation of Statistics

When the covariance is ***normalized***, one obtains the ***Pearson correlation coefficient***, which gives the goodness of the fit for the best possible linear function describing the relation between the variables. In this sense covariance is ***a linear gauge of dependence***.

https://www.zhihu.com/question/20852004

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

a measurement of correlation

Pearson Correlation Coefficient

Pearson Correlation Coefficient

Pearson Correlation Coefficient is a measure of the linear correlation between two variables X and Y.

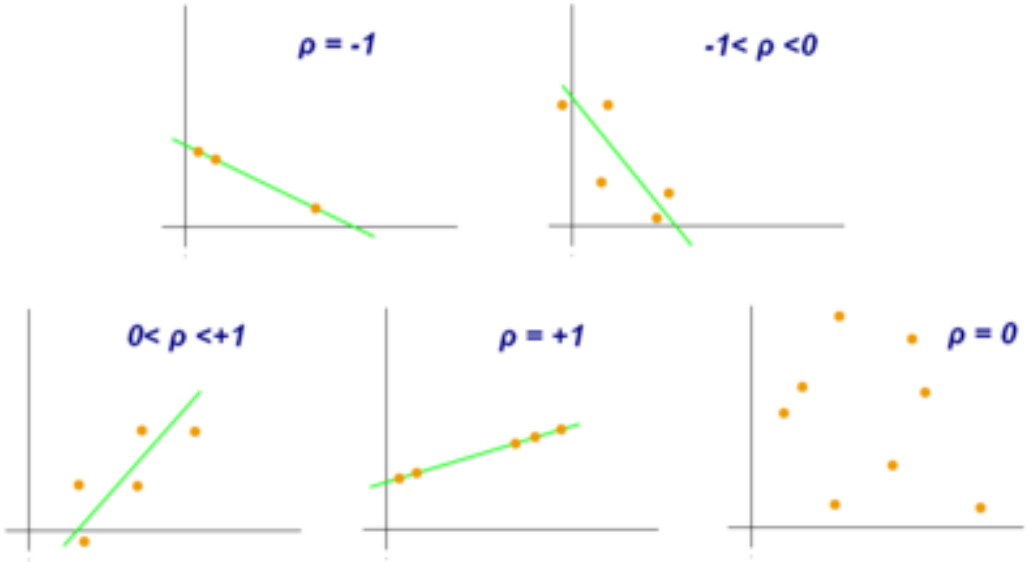
$$r(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var[X]\,Var[Y]}}$$

- It has a value between +1 and −1, where 1 is total positive linear correlation, 0 is no linear correlation, and −1 is total negative linear correlation.

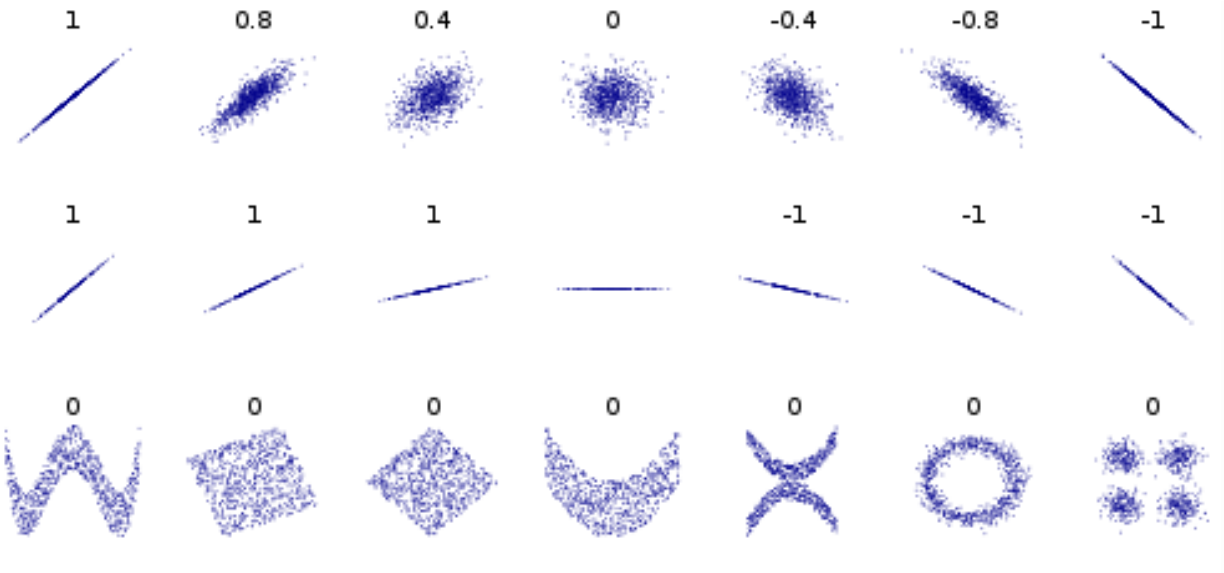
- It is widely used in the sciences. It was developed by Karl Pearson from a related idea introduced by Francis Galton in the 1880s.

Pearson Correlation Coefficient

Examples of scatter diagrams with different values of correlation coefficient (ρ)



Pearson Correlation Coefficient



Pearson Correlation Coefficient

Calculate PCC in Python

```
import numpy as np

a=np.array([1,2,3,4])
b=np.array([8,7,6,5])

print(np.corrcoef(a,b))
```

Pearson Correlation Coefficient

EXAMPLE 3:
INOHERB

Pearson Correlation Coefficient

very useful for natural language processing

Bayes' Theorem

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}

Bayes' Theorem

Properties of Probability

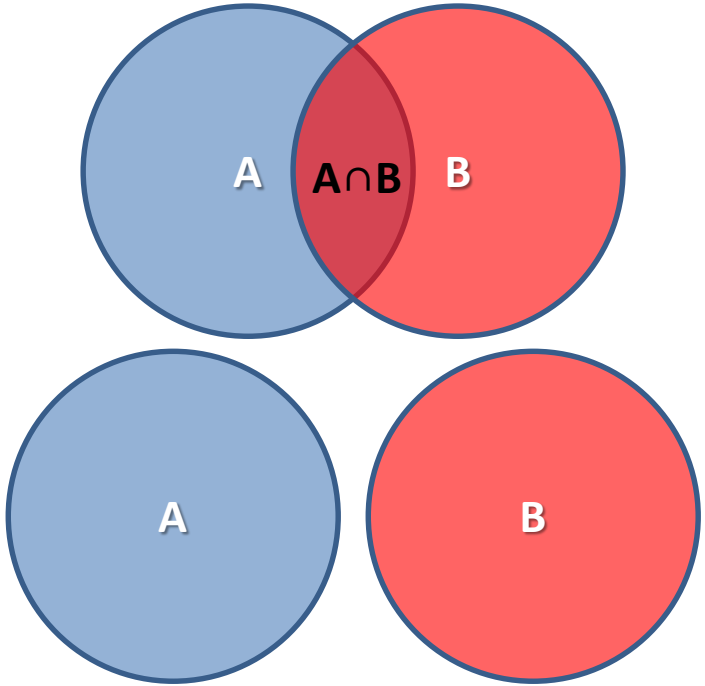
$$P(x_i) \geq 0$$

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

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

Bayes' Theorem

Independence 独立性



Dependent

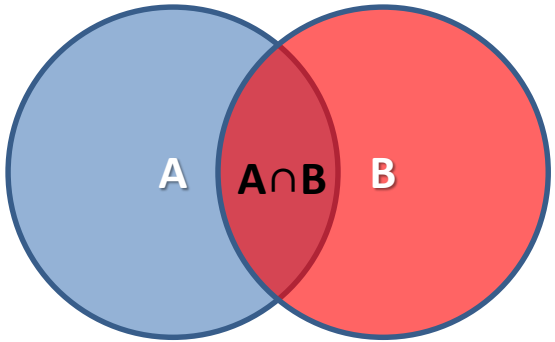
Independent

Bayes' Theorem

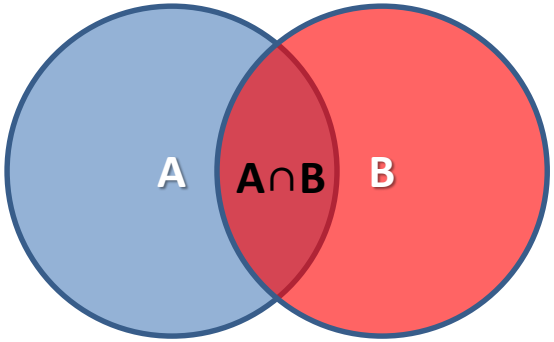
Conditional Probability 条件概率

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?) ↗

P(I am fine, too! | Fine, Thank you, and you?) ↗

P(Nothing much | Fine, Thank you, and you?) ↘

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

Bayes' Theorem

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

Markov Model

***Stochastic Process* 随机过程**

***Markov Chain* 马尔科夫链**

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

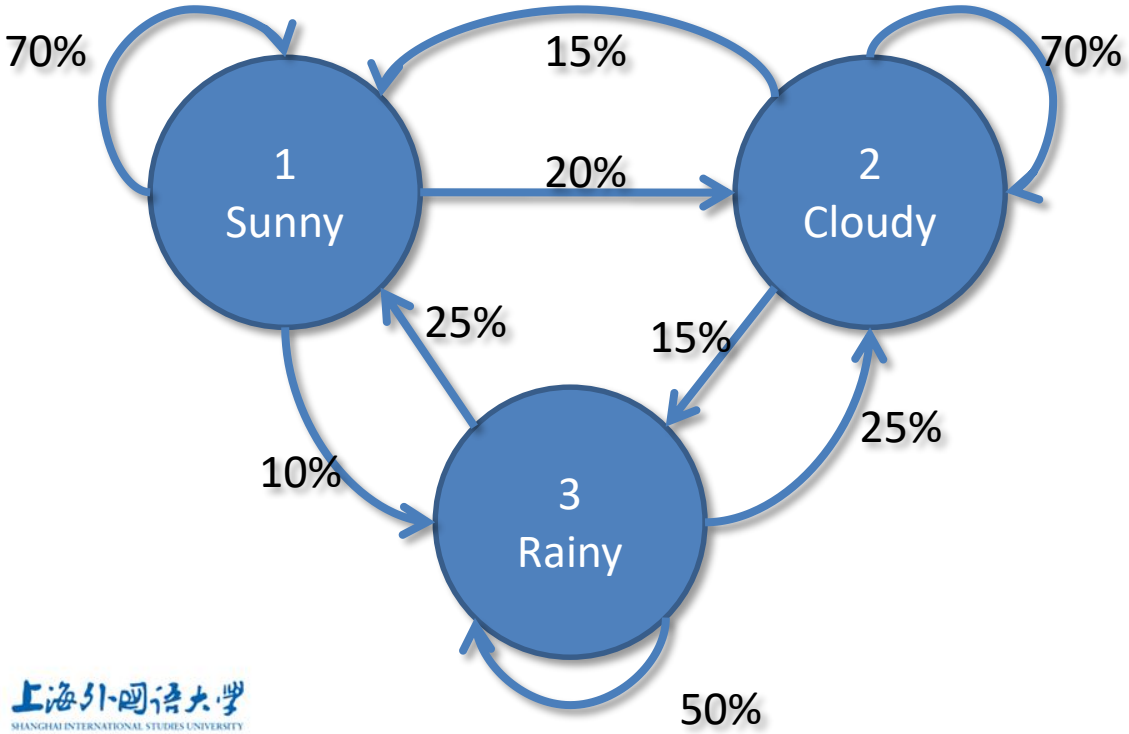
x_i ***is a Stochastic Process***

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%

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

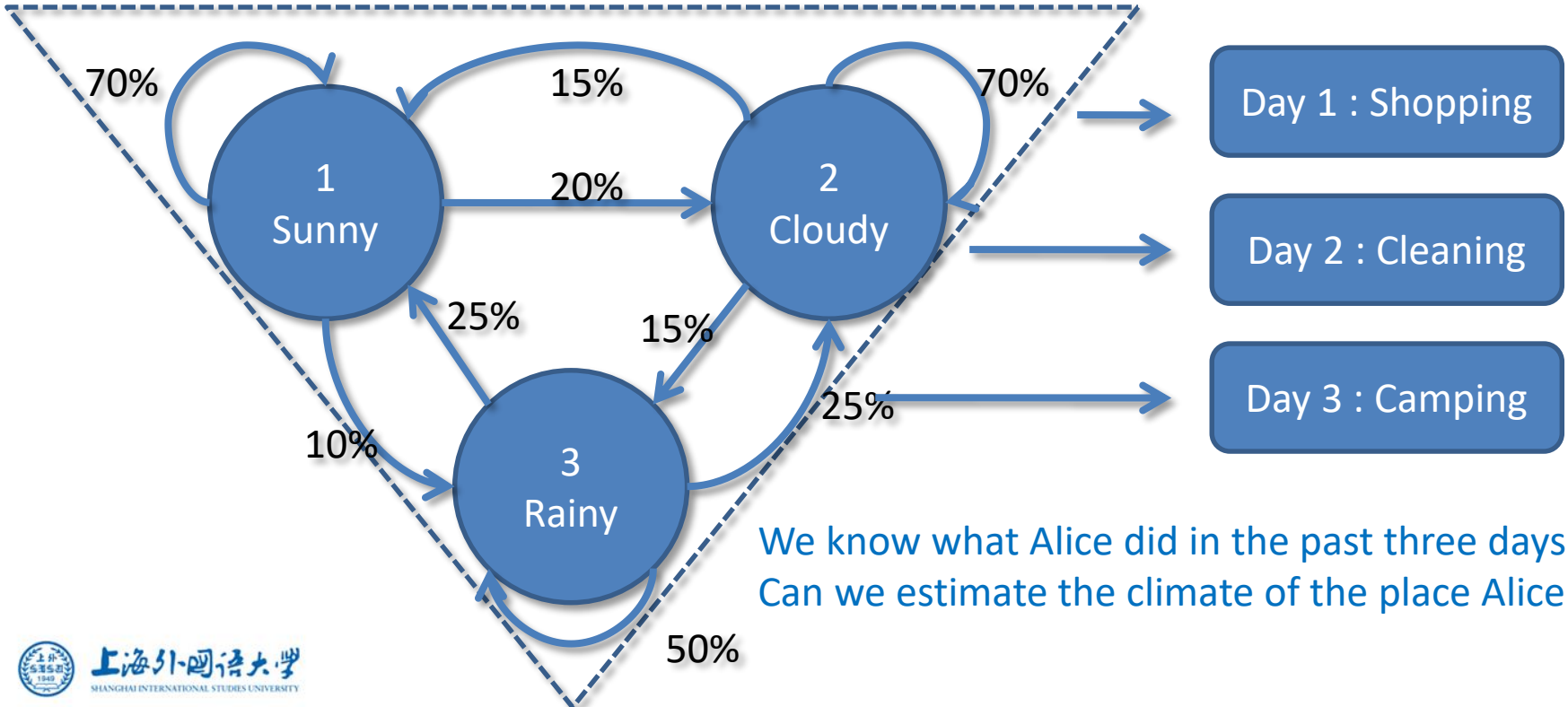
Your future is not decided by your past, but now!

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



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

Markov Model

The Applications of Markov Model in NLP

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

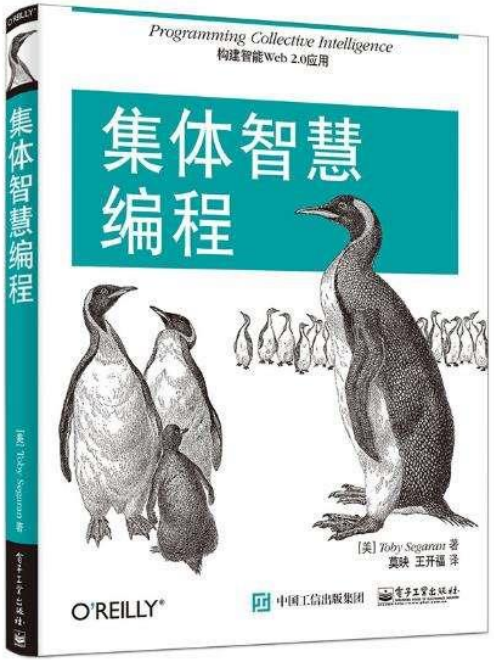
Reference

Reference

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

Reference

- https://item.jd.com/11667512.html



上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

The End of Lecture 8

Thank You

http://www.wangting.ac.cn