New Media
Data Analytics and Application

Lecture 12: Text Mining and Data Visualization

Ting Wang

Outlines

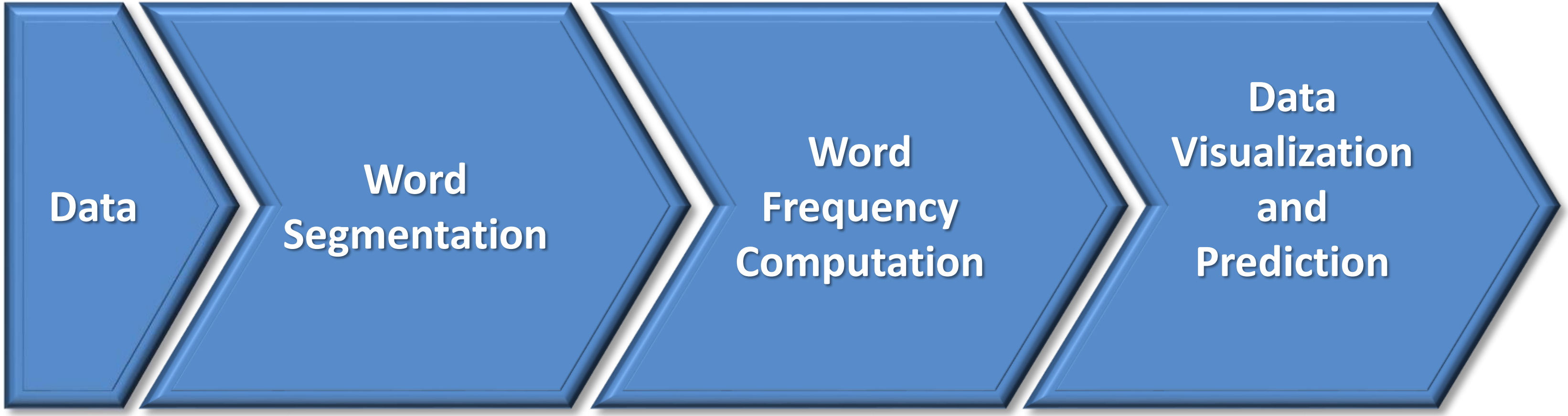
- Text Mining
  - Data Visualization using Python
- Data Mining Essentials

上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

online text data mining based on natural language processing

Text Mining

Now, we have data, how to mining it?



Data

Word Segmentation

Word Frequency Computation

Data Visualization and Prediction

Text Mining

Case Description

Motivations:

- To measure a news objectively
- To obtain new information efficiently

Methodologies:

- Describe a news report by quantitative method
- Technical integration by computer science, statistics and journalism

Text Mining

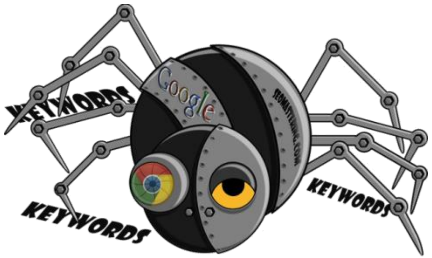
Steps:

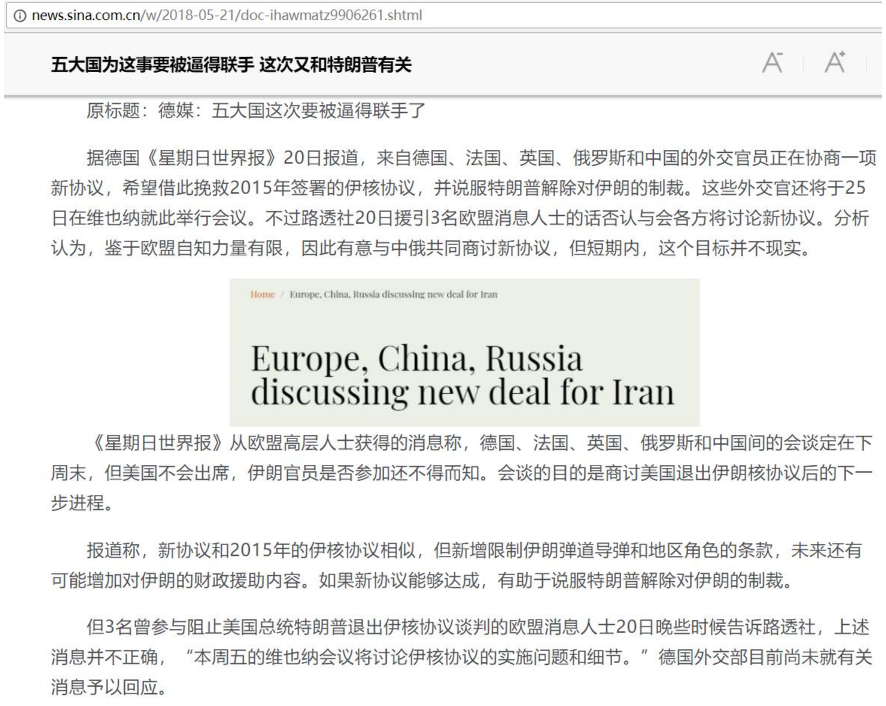
1. Download a news report
2. Word segmentation
3. Word tag extraction and statistical computing
4. Data visualization and news summarization

Step 1: Download a News Report

- Example: http://news.sina.com.cn/w/2018-05-21/doc-ihawmatz9906261.shtml

News also can be obtained by web crawler or databases




news.sina.com.cn/w/2018-05-21/doc-ihawmatz9906261.shtml

五大国为这事要被逼得联手 这次又和特朗普有关

原标题：德媒：五大国这次要被逼得联手了

据德国《星期日世界报》20日报道，来自德国、法国、英国、俄罗斯和中国的外交官员正在协商一项新协议，希望借此挽救2015年签署的伊核协议，并说服特朗普解除对伊朗的制裁。这些外交官还将于25日在维也纳就此举行会议。不过路透社20日援引3名欧盟消息人士的话否认与会各方将讨论新协议。分析认为，鉴于欧盟自知力量有限，因此有意与中俄共同商讨新协议，但短期内，这个目标并不现实。

Home / Europe, China, Russia discussing new deal for Iran

Europe, China, Russia discussing new deal for Iran

《星期日世界报》从欧盟高层人士获得的消息称，德国、法国、英国、俄罗斯和中国间的会谈定在下周末，但美国不会出席，伊朗官员是否参加还不得而知。会谈的目的是商讨美国退出伊朗核协议后的下一步进程。

报道称，新协议和2015年的伊核协议相似，但新增限制伊朗弹道导弹和地区角色的条款，未来还有可能增加对伊朗的财政援助内容。如果新协议能够达成，有助于说服特朗普解除对伊朗的制裁。

但3名曾参与阻止美国总统特朗普退出伊核协议谈判的欧盟消息人士20日晚些时候告诉路透社，上述消息并不正确，“本周五的维也纳会议将讨论伊核协议的实施问题和细节。”德国外交部目前尚未就有关消息予以回应。

Step 2: Word segmentation (1)

Database Preparation

- Word Dictionary (required)
- Stop Word Dictionary (required)
- Dictionaries of Terms (optional)
- Word Chains (required if using N-gram)
- Part of Speech (optional)
- Word Sentiment (optional for Sentiment Analysis)

Step 2: Word segmentation (2)

Chinese Word Segmentation

- FMM
- BMM
- N-gram

```
def word_seg_fmm(content): #正向匹配
    MaxLen=10 #最大词长
    Len=MaxLen #动态切割词长
    Seg_Content="" #返回的切割结果

    while len(content)>0:
        if content[0:Len] in WordMap: #词典中有匹配
            Seg_Content=Seg_Content+content[0:Len]+"|"
            content=content[Len:]
            Len=MaxLen
            #print("Seg_Content1:"+Seg_Content)
            continue
        else: #词典中无匹配
            Len=Len-1
            if Len==1:#仅剩一个词还没匹配到
                Seg_Content = Seg_Content + content[0:Len] + "|"
                content = content[Len:]
                Len = MaxLen
                #print("Seg_Content2:" + Seg_Content)
    return Seg_Content[:-1]
```

```
def word_seg_bmm(content): #逆向匹配
    MaxLen=10 #最大词长
    Len=MaxLen #动态切割词长
    Seg_Content="" #返回的切割结果

    while len(content)>0:
        if content[-Len:] in WordMap: #词典中有匹配
            Seg_Content=content[-Len:]+"|"+Seg_Content
            content=content[:-Len]
            Len=MaxLen
            #print("Seg_Content1:"+Seg_Content)
            continue
        else: #词典中无匹配
            Len=Len-1
            if Len==1:#仅剩一个词还没匹配到
                Seg_Content = content[-Len:] + "|" + Seg_Content
                content = content[:-Len]
                Len = MaxLen
                #print("Seg_Content2:" + Seg_Content)
    return Seg_Content[:-1]
```

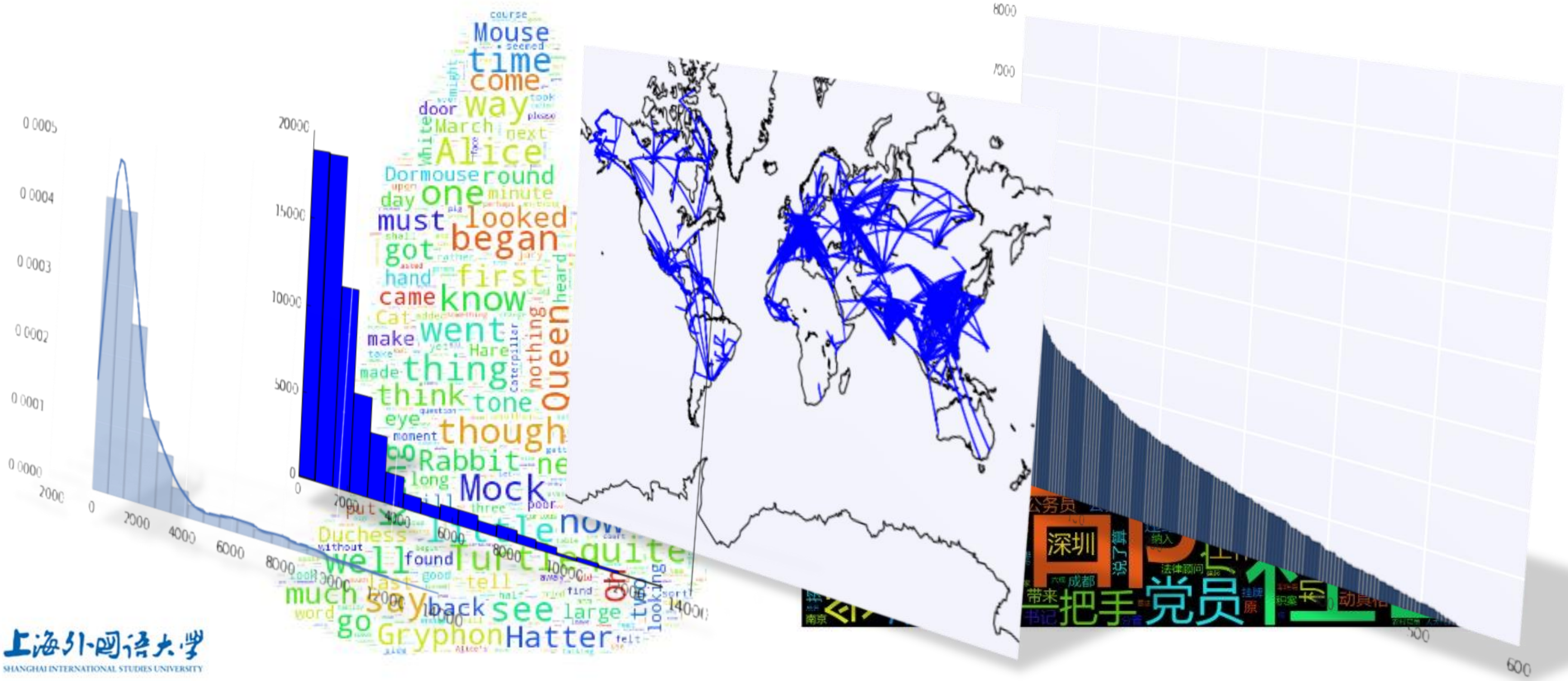
Step 2: Word segmentation (3)

- Tips for Chinese Word Segmentation
  - Initialization is very important
  - Segment in the memory (not hard disk or data bases) to accelerate the segmentation speed
  - Using “set” to store the dictionary, and “dict” for segmented words in Python
  - For Tag Analysis, a precise word segmentation is unnecessary

Step 3:Word Tag Extraction and Statistical Computing

- str.split() for all tags
- Discarding One-Char tags
- Discarding Stop-Word tags
- Select tags whose term frequencies are larger than a threshold (for example >2)
- Other statistical computing

Step 4: Data Visualization and News Summarization



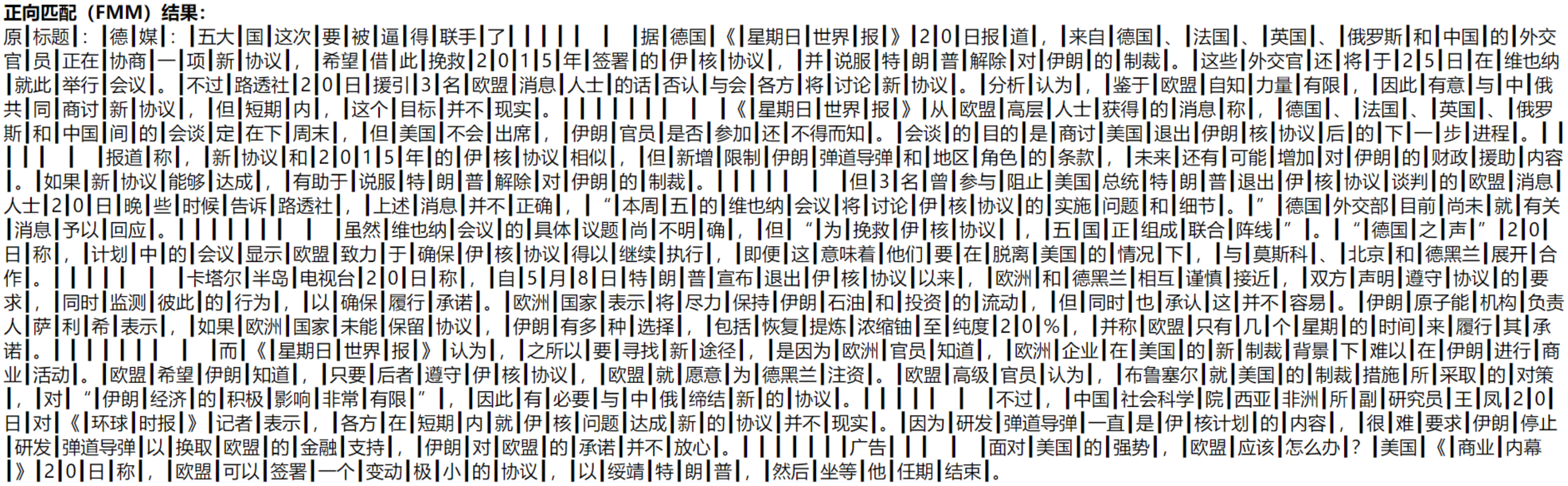
Data Visualization using Python

- Necessity:
  - NumPy (Computing Package)
  - Scipy (Scientific Computing Package)
  - Pillow(Image)
  - Matplotlib (Diagram Package)
  - wordcloud (Word Cloud Package)

Installation Sequence

- Some packages also need some other required packages

Result

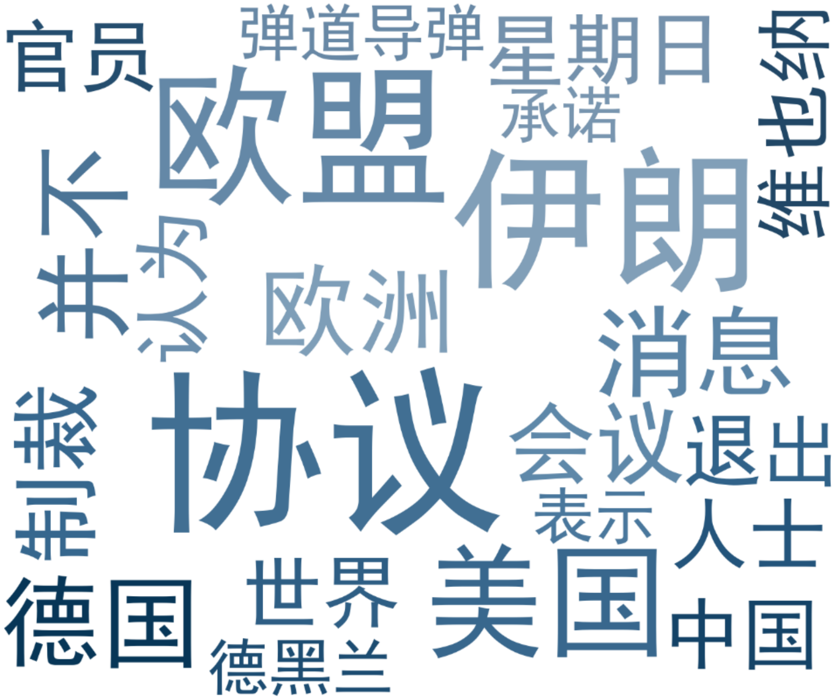


正向匹配（FMM）结果：

```
原|标题|：|德|媒|：|五大|国|这次|要|被|逼|得|联手|了| | | | |  |  |据|德国|《|星期日|世界|报|》|2|0|日报|道|，|来自|德国|、|法国|、|英国|、|俄罗斯|和|中国|的|外交
官|员|正在|协商|一|项|新|协议|，|希望|借|此|挽救|2|0|1|5|年|签署|的|伊|核|协议|，|并|说服|特|朗|普|解除|对|伊朗|的|制裁|。|这些|外交官|还|将|于|2|5|日|在|维也纳
|就此|举行|会议|。|不过|路透社|2|0|日|援引|3|名|欧盟|消息|人士|的话|否认|与会|各方|将|讨论|新|协议|。|分析|认为|，|鉴于|欧盟|自知|力量|有限|，|因此|有意|与|中|俄
共|同|商讨|新|协议|，|但|短期|内|，|这个|目标|并不|现实|。| | | | | | |  |  |《|星期日|世界|报|》|从|欧盟|高层|人士|获得|的|消息|称|，|德国|、|法国|、|英国|、|俄罗
斯|和|中国|间|的|会谈|定|在下|周末|，|但|美国|不会|出席|，|伊朗|官员|是否|参加|还|不得而知|。|会谈|的|目的|是|商讨|美国|退出|伊朗|核|协议|后|的|下|一|步|进程|。| |
| | |  |  |报道|称|，|新|协议|和|2|0|1|5|年|的|伊|核|协议|相似|，|但|新增|限制|伊朗|弹道导弹|和|地区|角色|的|条款|，|未来|还有|可能|增加|对|伊朗|的|财政|援助|内容
|。|如果|新|协议|能够|达成|，|有助于|说服|特|朗|普|解除|对|伊朗|的|制裁|。| | | | |  |  |但|3|名|曾|参与|阻止|美国|总统|特|朗|普|退出|伊|核|协议|谈判|的|欧盟|消息|
人士|2|0|日|晚|些|时候|告诉|路透社|，|上述|消息|并不|正确|，|“|本周|五|的|维也纳|会议|将|讨论|伊|核|协议|的|实施|问题|和|细节|。|”|德国|外交部|目前|尚未|就|有关
|消息|予以|回应|。| | | | | | |  |  |虽然|维也纳|会议|的|具体|议题|尚|不明|确|，|但|“|为|挽救|伊|核|协议| |，|五|国|正|组成|联合|阵线|”|。|“|德国|之|声|”|2|0|
日|称|，|计划|中|的|会议|显示|欧盟|致力|于|确保|伊|核|协议|得以|继续|执行|，|即便|这|意味着|他们|要|在|脱离|美国|的|情况|下|，|与|莫斯科|、|北京|和|德黑兰|展开|合
作|。| | | | |  |  |卡塔尔|半岛|电视台|2|0|日|称|，|自|5|月|8|日|特|朗|普|宣布|退出|伊|核|协议|以来|，|欧洲|和|德黑兰|相互|谨慎|接近|，|双方|声明|遵守|协议|的|要
求|，|同时|监测|彼此|的|行为|，|以|确保|履行|承诺|。|欧洲|国家|表示|将|尽力|保持|伊朗|石油|和|投资|的|流动|，|但|同时|也|承认|这|并不|容易|。|伊朗|原子能|机构|负责
人|萨|利|希|表示|，|如果|欧洲|国家|未能|保留|协议|，|伊朗|有多|种|选择|，|包括|恢复|提炼|浓缩铀|至|纯度|2|0|%|，|并称|欧盟|只有|几|个|星期|的|时间|来|履行|其|承
诺|。| | | | | | |  |  |而|《|星期日|世界|报|》|认为|，|之所以|要|寻找|新|途径|，|是因为|欧洲|官员|知道|，|欧洲|企业|在|美国|的|新|制裁|背景|下|难以|在|伊朗|进行|商
业|活动|。|欧盟|希望|伊朗|知道|，|只要|后者|遵守|伊|核|协议|，|欧盟|就|愿意|为|德黑兰|注资|。|欧盟|高级|官员|认为|，|布鲁塞尔|就|美国|的|制裁|措施|所|采取|的|对策
|，|对|“|伊朗|经济|的|积极|影响|非常|有限|”|，|因此|有|必要|与|中|俄|缔结|新|的|协议|。| | | | |  |  |不过|，|中国|社会科学|院|西亚|非洲|所|副|研究员|王|凤|2|0|
日|对|《|环球|时报|》|记者|表示|，|各方|在|短期|内|就|伊|核|问题|达成|新|的|协议|并不|现实|。|因为|研发|弹道导弹|一直|是|伊|核计划|的|内容|，|很|难|要求|伊朗|停止
|研发|弹道导弹|以|换取|欧盟|的|金融|支持|，|伊朗|对|欧盟|的|承诺|并不|放心|。| | | | | | |广告| | |  |  |面对|美国|的|强势|，|欧盟|应该|怎么办|？|美国|《|商业|内幕
|》|2|0|日|称|，|欧盟|可以|签署|一个|变动|极|小|的|协议|，|以|绥靖|特|朗|普|，|然后|坐等|他|任期|结束|。
```

Conclusions

本文与伊朗问题有关，可能跟武器和制裁有关，起决定力量的应该是美国、德国、中国和伊朗。欧洲与此消息关系较大。



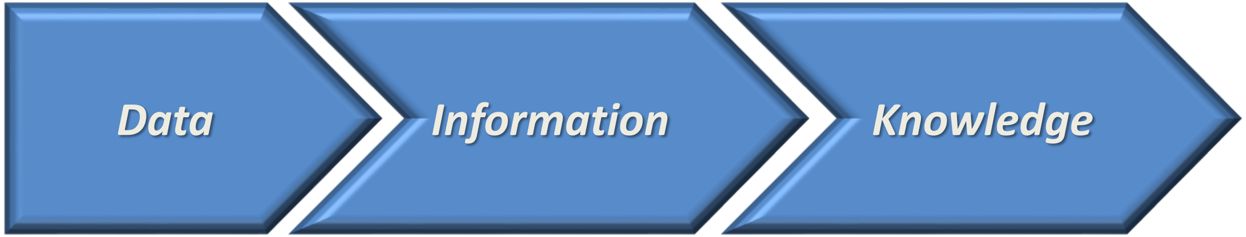
上海外国语大学
SHANGHAI INTERNATIONAL STUDIES UNIVERSITY

machine learning approaches for data mining

Data Mining Essentials

Data Mining ***数据挖掘***

- – Data Mining is the power for producing high-quality journalism.
- – Data Mining is an interdisciplinary subfield of computer science, and statistics.

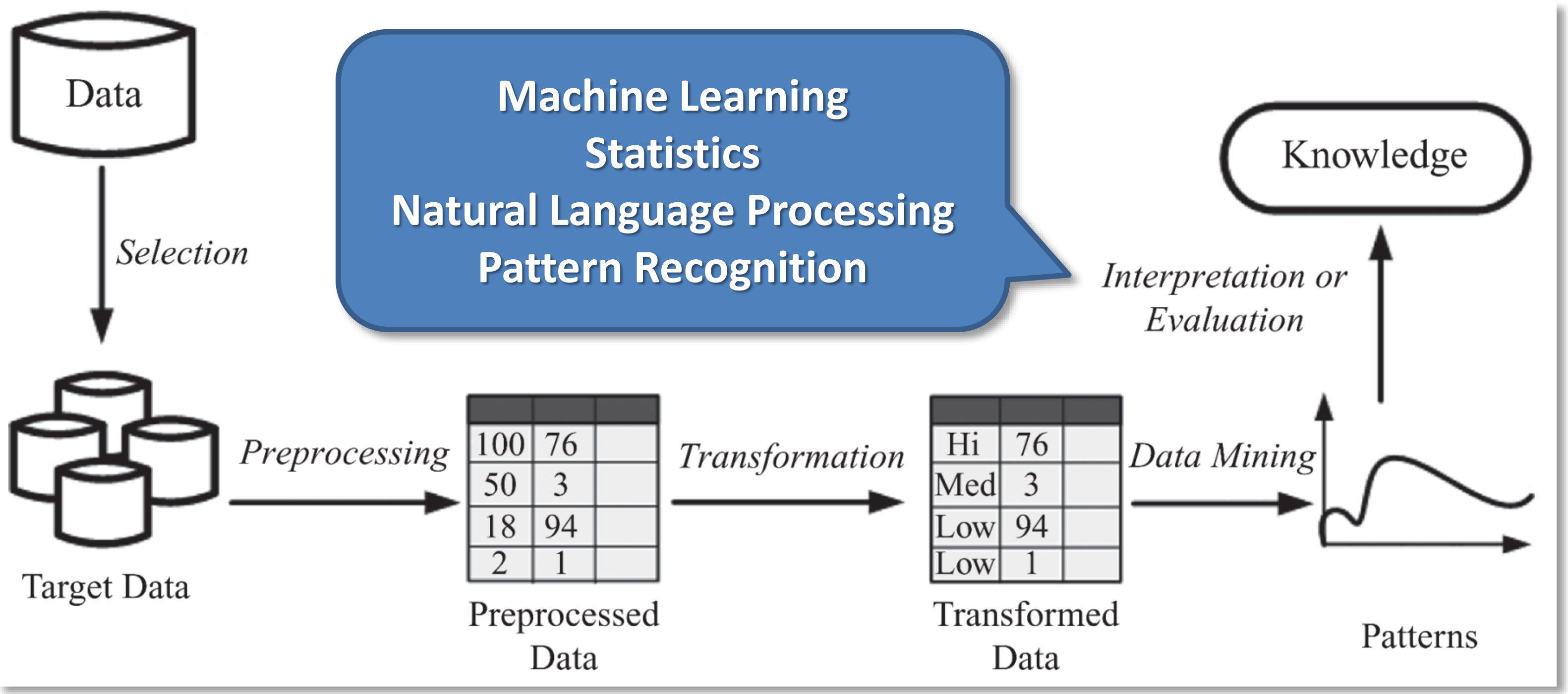


Social Demands

- Data production rate has increased dramatically (**Big Data**) and we are able store much more data
  - E.g., purchase data, social media data, cell phone data
- Businesses and customers need <u>useful</u> or <u>actionable</u> knowledge to gain insight from raw data for various purposes
  - It’s not just searching data or databases

The process of extracting useful patterns from raw data is known as Knowledge Discovery in Databases (KDD)

KDD from Data Bases



100	76	
50	3	
18	94	
2	1	

Hi	76	
Med	3	
Low	94	
Low	1	

Machine Learning
Statistics
Natural Language Processing
Pattern Recognition

Data 数据

- Continuous Data 连续型数据
  - – Regression
- Discrete Data 离散型数据
  - – Classification

Data Feature (1) 数字特征

Feature also called as Measurement, Attribute

- **Nominal 名词性**
  - **Operations**:
    - Mode (most common feature value), Equality Comparison
  - E.g., {male, female}
- **Ordinal 序数性**
  - Feature values have an intrinsic order to them, but the difference is not defined
  - **Operations**:
    - same as nominal, feature value rank
  - E.g., {Low, medium, high}

Data Feature (2) *数字特征*

- **Interval 间隔性**
  - **Operations:**
    - Addition and subtractions are allowed whereas divisions and multiplications are not
  - E.g., 3:08 PM, calendar dates
- **Ratio 比例性**
  - **Operations:**
    - divisions and multiplications are allowed
  - E.g., Height, weight, money quantities

Data Quality 数据质量

- **Noise 噪声数据**
  - Noise is the distortion of the data
- **Outliers 异常值**
  - Outliers are data points that are considerably different from other data points in the dataset
- **Missing Values 缺失值**
  - Missing feature values in data instances
  - **Solution:**
    - Remove instances that have missing values
    - Estimate missing values, and
    - Ignore missing values when running data mining algorithm
- **Duplicate data 重复数据**

Data Preprocessing (1) *数据预处理*

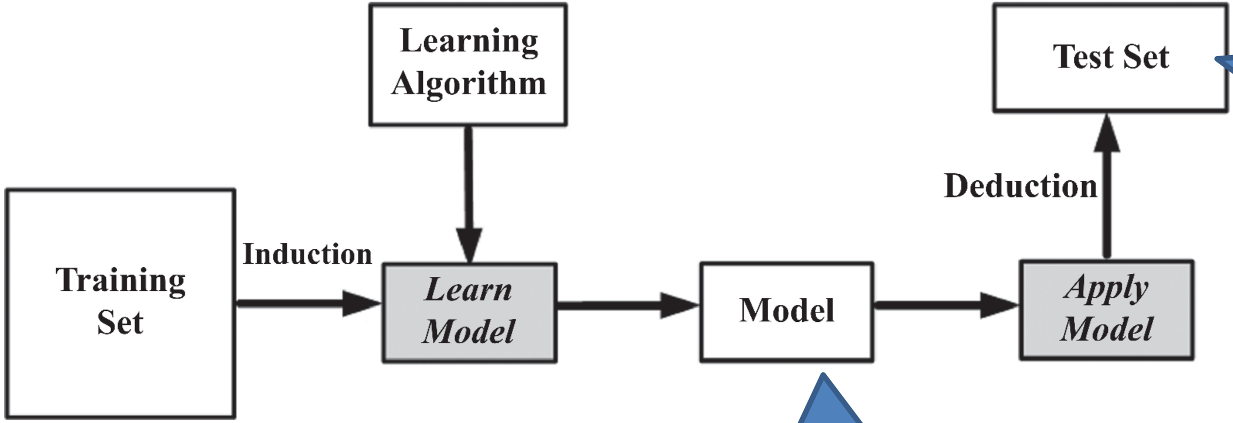
- **Aggregation 聚合**
  - It is performed when multiple features need to be combined into a single one or when the scale of the features change
  - Example: image width , image height -> image area (width x height)
- **Discretization 离散化**
  - From continues values to discrete values
  - Example: money spent -> {low, normal, high}

- ***Data Preprocessing (2)*** **数据预处理**
- **Feature Selection 特征选择**
  - Choose relevant features
- **Feature Extraction 特征提取**
  - Creating new features from original features
  - Often, more complicated than aggregation
- **Sampling 取样**
  - Random Sampling
  - Sampling with or without replacement
  - Stratified Sampling: useful when having class imbalance
  - Social Network Sampling

Machine Learning 机器学习

- Supervised Learning
  - Classification
  - Regression
- Unsupervised Learning
  - Clustering
  - Dimensional Reduction

Supervised Machine Learning 有监督学习



Unlabeled

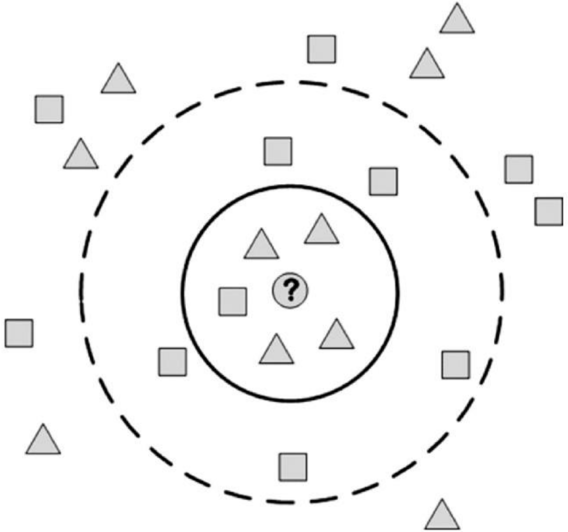
Labeled

Parameter Setting

Classification 分类

Prediction Result with Labeled Discrete Value

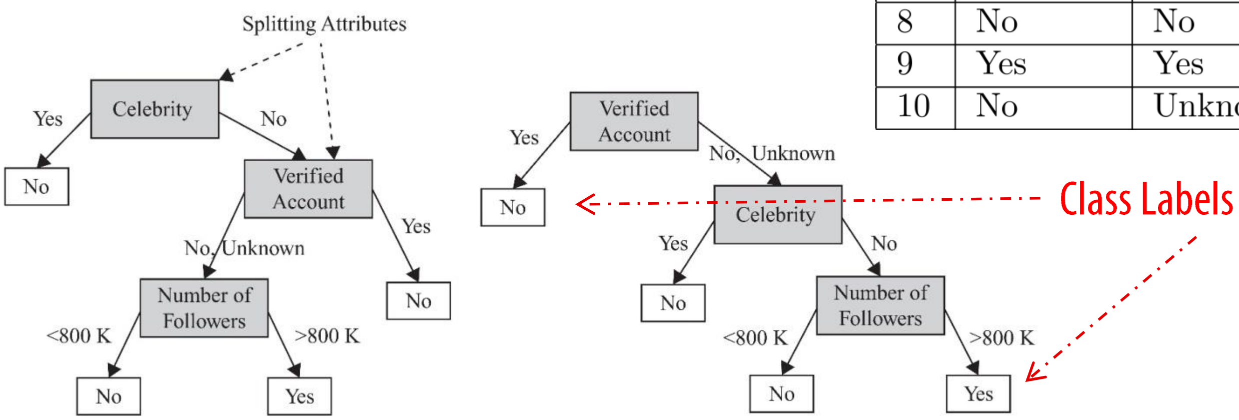
- KNN(K-Nearest Neighbors) K临近原则
- Linear Classifier 线性分类器
- Neural Networks 神经网络
- Support Vector Machine 支撑向量机
- Decision Tree 决策树



Data Mining Essentials

Multiple **decision trees** can be learned from the same dataset

ID	Celebrity	Verified Account	# Followers	Influential?
1	Yes	No	1.25M	No
2	No	Yes	1M	No
3	No	Yes	600K	No
4	Yes	Unknown	2.2M	No
5	No	No	850K	Yes
6	No	Yes	750K	No
7	No	No	900K	Yes
8	No	No	700K	No
9	Yes	Yes	1.2M	No
10	No	Unknown	950K	Yes



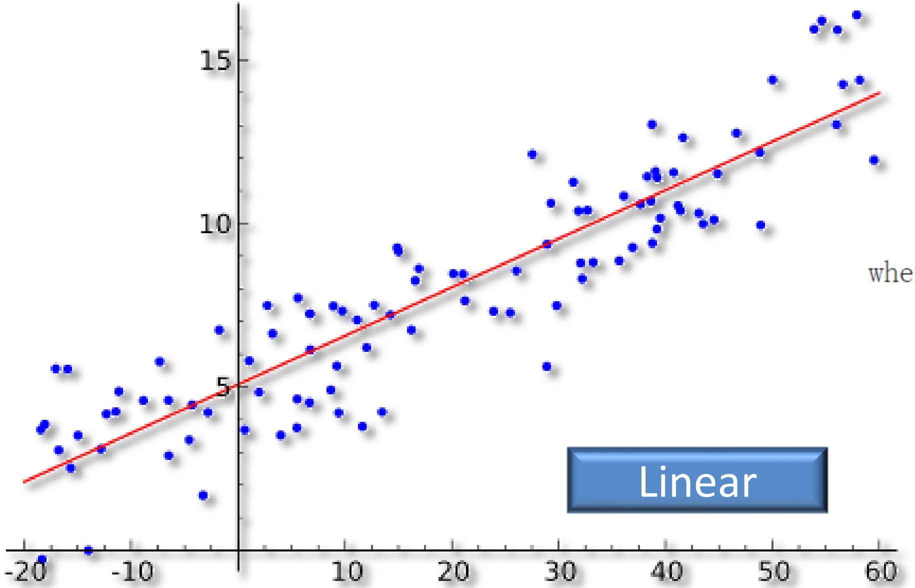
Class Labels

(a) Learned Decision Tree 1

(b) Learned Decision Tree 2

Regression (1) *回归*

Prediction Result with Unlabeled Continuous Value



Eg. Linear least squares 线性最小二乘法

$$\mathbf{X}\boldsymbol{\beta} = \mathbf{y},$$

where

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1n} \\ X_{21} & X_{22} & \cdots & X_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ X_{m1} & X_{m2} & \cdots & X_{mn} \end{bmatrix}, \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}.$$

Regression (2)*回归*

Nonlinear Regression 非线性回归计算

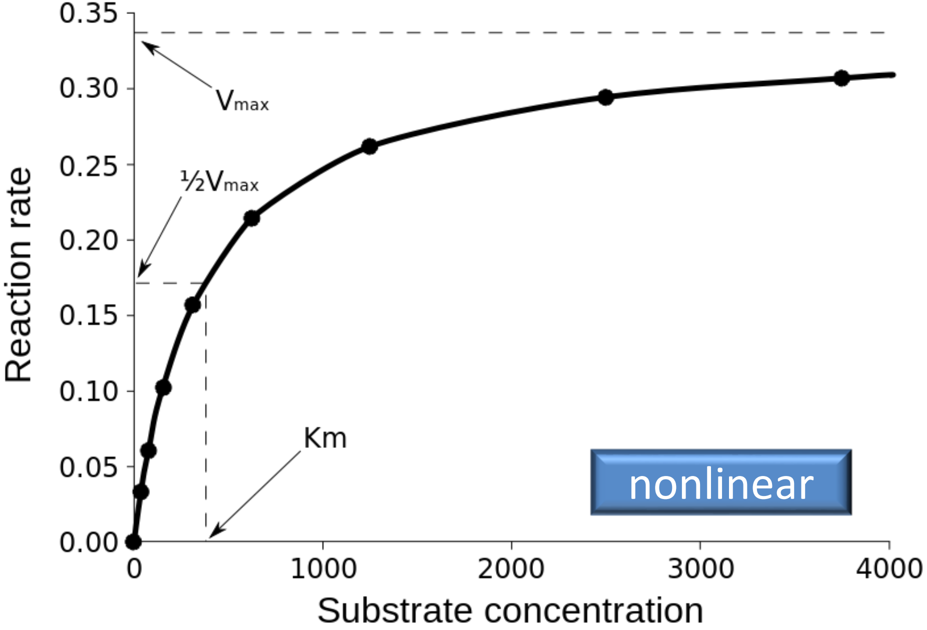
- Linearization 线性化方法

    1. Transformation 变形法

$$y = ae^{bx}U \Rightarrow \ln(y) = \ln(a) + bx + u$$

    2. Segmentation 分割法

split up into classes or segments and *linear regression* can be performed per segment



Unsupervised Machine Learning

无监督学习

machine learning task of inferring a function to describe hidden structure from unlabeled data

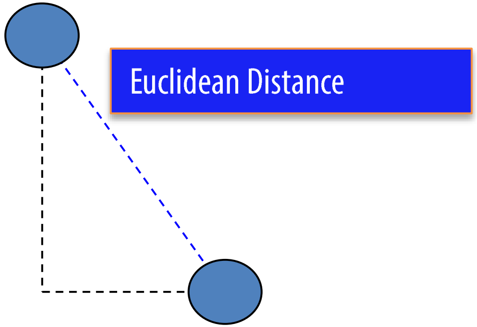
Clustering 聚类

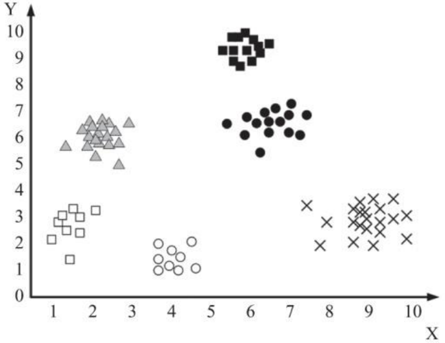
- **Clustering Goal:** Group together similar items
- Clustering algorithms group together **similar items**

  –The algorithm does not have examples showing how the samples should be grouped together (unlabeled data)

Similarity Computing (1) *相似度计算*

–The most popular (dis)similarity measure for continuous features are **Euclidean Distance** and **Pearson Linear Correlation**





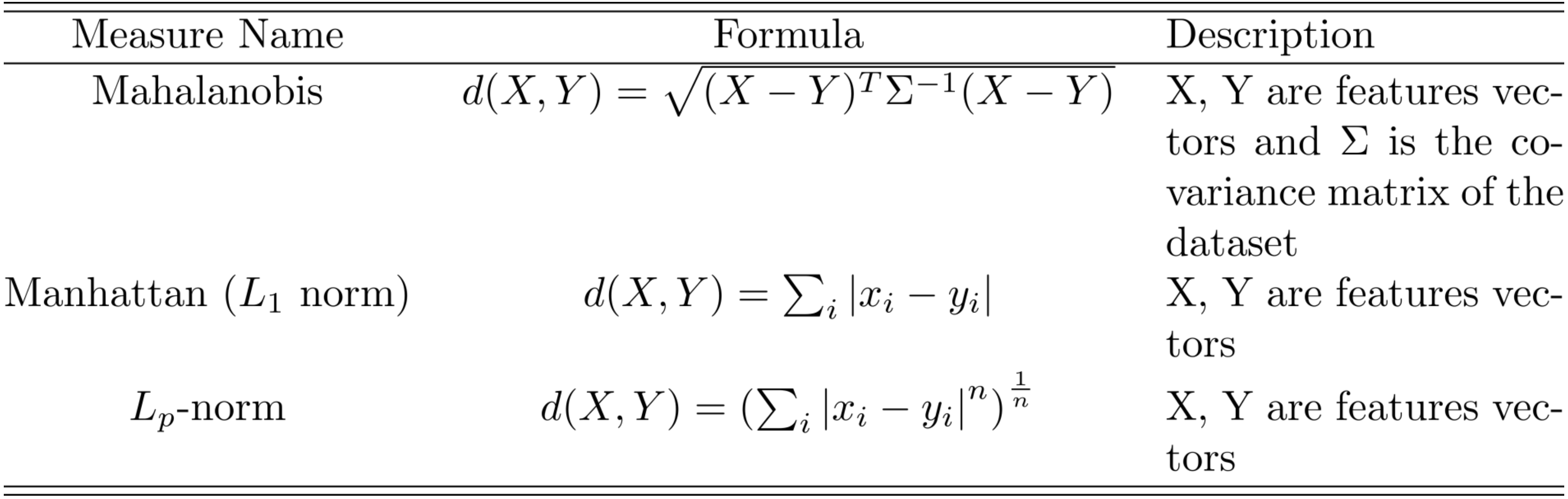
$$d(X,Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

Similarity Computing (2)相似度计算

X and Y are n Dimensional Vectors

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

$$Y = (y_1, y_2, \ldots, y_n)$$



Measure Name	Formula	Description
Mahalanobis	$d(X,Y) = \sqrt{(X-Y)^T\Sigma^{-1}(X-Y)}$	X, Y are features vectors and Σ is the covariance matrix of the dataset
Manhattan (L_1 norm)	$d(X,Y) = \sum_i \lvert x_i - y_i \rvert$	X, Y are features vectors
L_p-norm	$d(X,Y) = (\sum_i \lvert x_i - y_i \rvert^n)^{\frac{1}{n}}$	X, Y are features vectors

Once a distance measure is selected, instances are grouped using it.

Pearson Linear Correlation 皮尔逊线性相关

Correlation Coefficient 相关系数

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Where, cov **is the covariance**

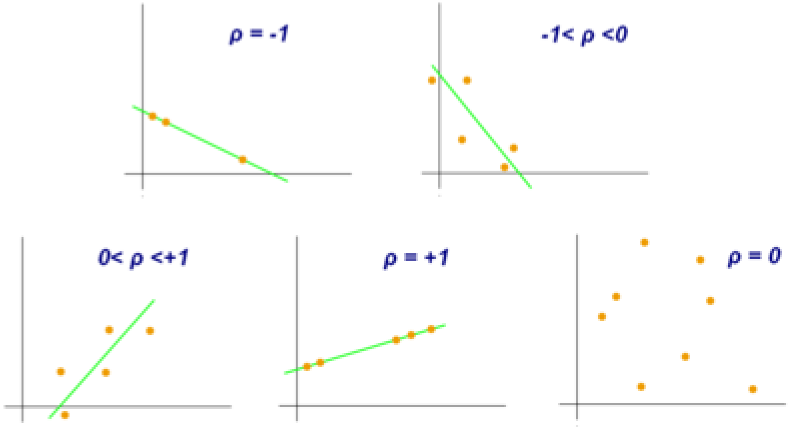
σ **is the standard deviation**

$$\operatorname{cov}(X,Y) = \mathrm{E}[(X-\mu_X)(Y-\mu_Y)]$$

$$\rho_{X,Y} = \frac{\mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]}{\sqrt{\mathrm{E}[X^2] - [\,E[X]]^2}\;\sqrt{\mathrm{E}[Y^2] - [\,E[Y]]^2}}.$$

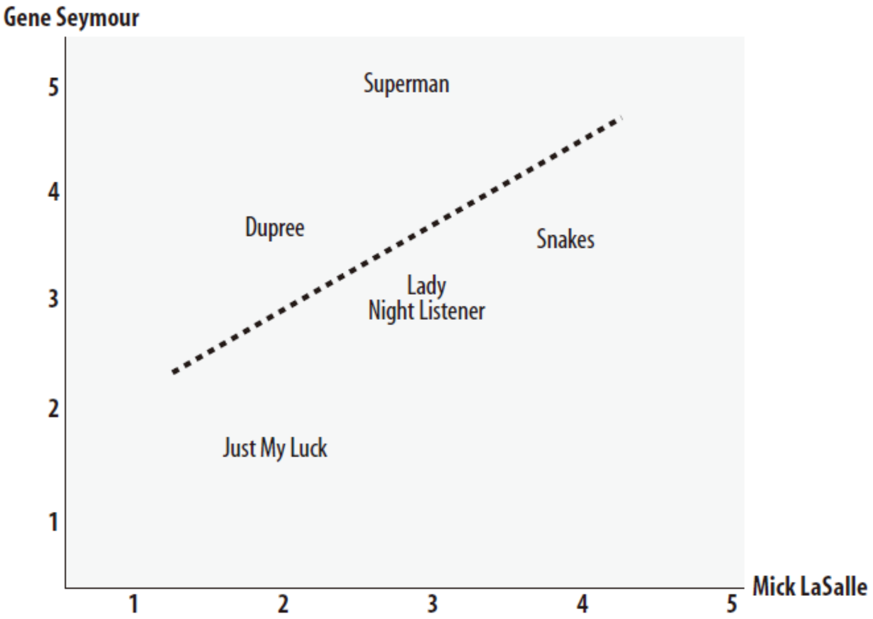
Relations between Variance and Covariance

$$\mu_X = \mathrm{E}[X]$$
$$\mu_Y = \mathrm{E}[Y]$$
$$\sigma_X^2 = \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - [\,E[X]]^2$$
$$\sigma_Y^2 = \mathrm{E}[(Y - \mathrm{E}[Y])^2] = \mathrm{E}[Y^2] - [\,E[Y]]^2$$
$$\mathrm{E}[(X-\mu_X)(Y-\mu_Y)] = \mathrm{E}[(X-\mathrm{E}[X])(Y-\mathrm{E}[Y])] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y]$$

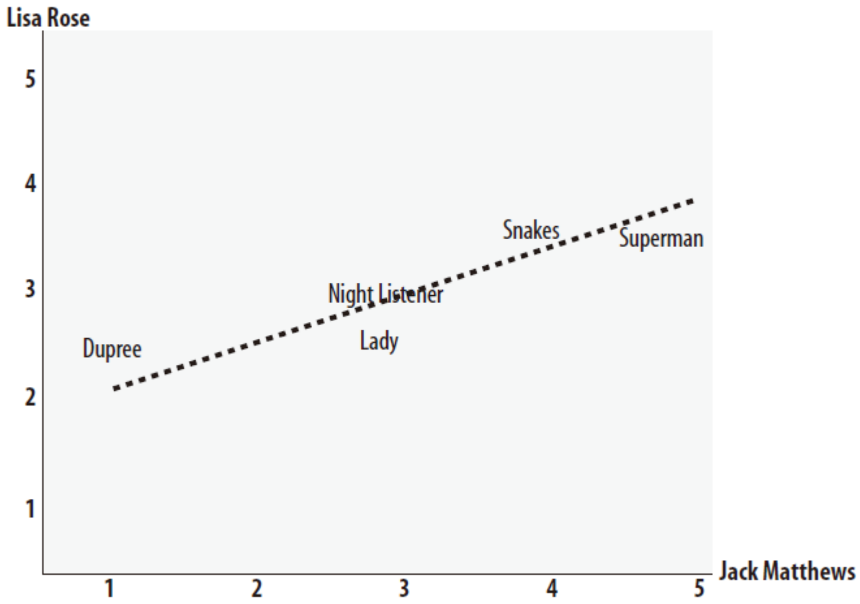


Film Ranking Correlation

Superman *was rated 3 by Mick LaSalle and 5 by Gene Seymour,* so it is placed at (3,5) on the chart.



$$\rho_{X,Y} = 0.4$$



$$\rho_{X,Y} = 0.75$$

Conclusion: Films recommended to Lisa, also can be recommended to Jack.

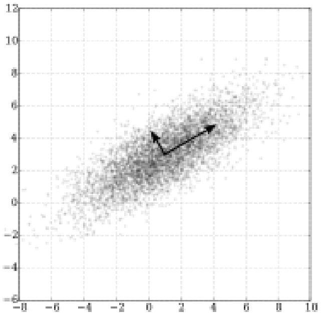
Dimensional Reduction 降维

Principal Component Analysis (PCA) 主成份分析

1. PCA is a statistical procedure ***converts*** a set of observations of possibly **correlated variables *into*** a set of values of linearly **uncorrelated variables** called principal components.

2. The number of principal components is **less than or equal to** the number of original variables.

3. This transformation is defined in such a way that **the first principal component** has ***the largest possible variance***, and each **succeeding component** in turn has **the highest variance possible under the constraint** that it is orthogonal to the preceding components.



Ref. http://www.cnblogs.com/SCUJIN/p/5965946.html

Reference

Reference

Books and Chapters (1)

https://item.jd.com/11983227.html

Chapter 1-2

Machine Learning Package Installation

Machine Learning Theory Foundations

Reference

Books and Chapters (2)

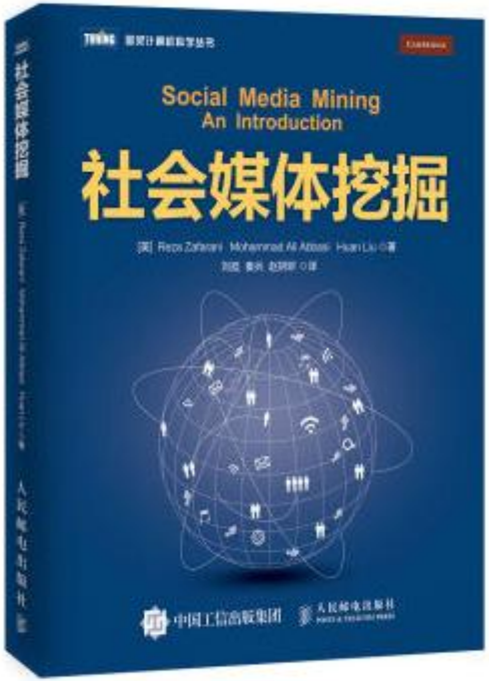
https://item.jd.com/11803260.html

Chapter 5

Data Mining Essentials

Online Reference:

http://www.public.asu.edu/~huanliu/

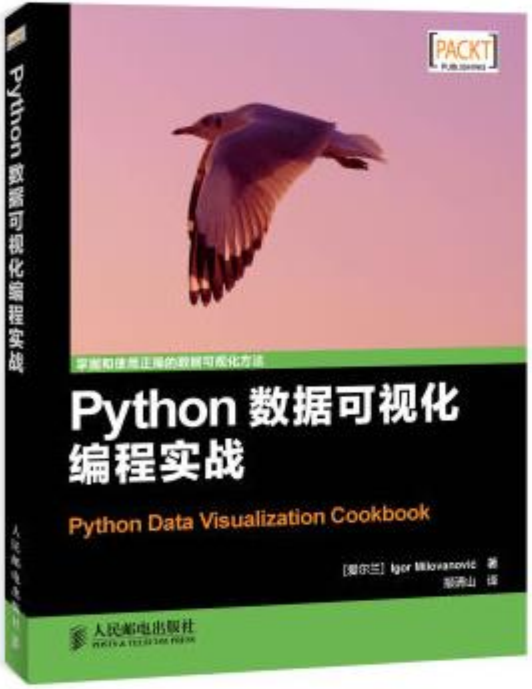


Reference

Books and Chapters (3)

https://item.jd.com/11676691.html

Python Data Visualization

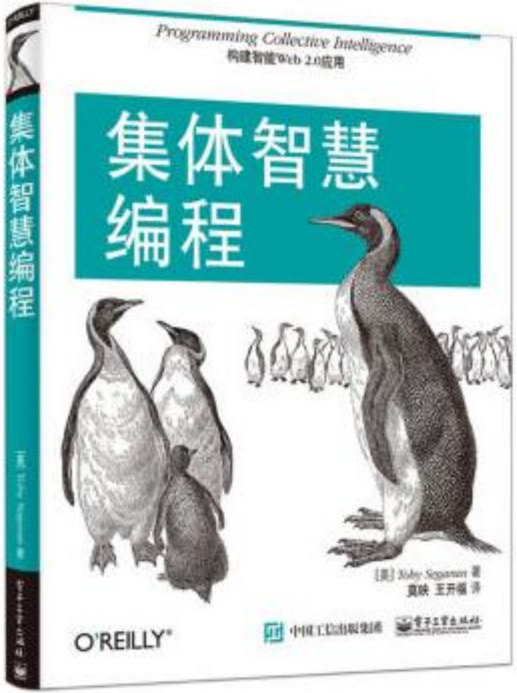

PACKT
Python 数据可视化编程实战
Python 数据可视化编程实战
Python Data Visualization Cookbook
Igor Milovanović 著
人民邮电出版社

Reference

Books and Chapters (4)

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

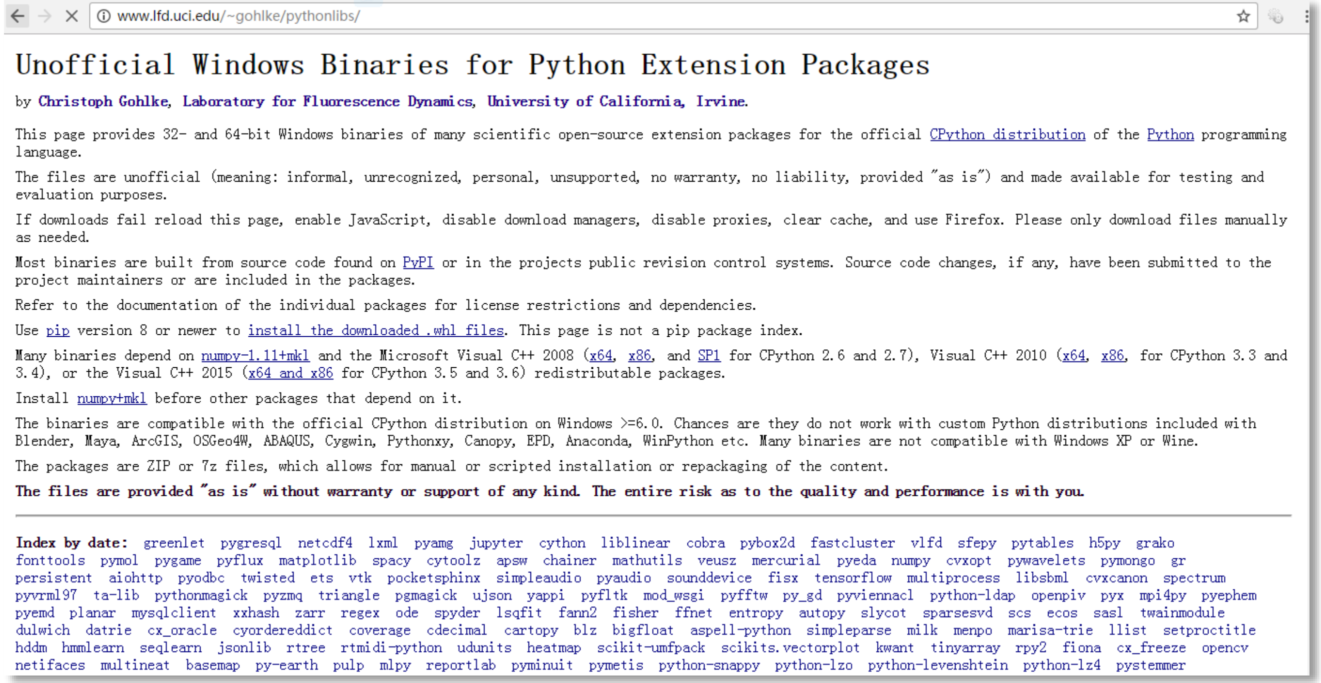
Programming Collective Intelligence



Reference

Python Extension Packages

http://www.lfd.uci.edu/~gohlke/pythonlibs/



Unofficial Windows Binaries for Python Extension Packages

by **Christoph Gohlke, Laboratory for Fluorescence Dynamics, University of California, Irvine.**

This page provides 32- and 64-bit Windows binaries of many scientific open-source extension packages for the official CPython distribution of the Python programming language.

The files are unofficial (meaning: informal, unrecognized, personal, unsupported, no warranty, no liability, provided "as is") and made available for testing and evaluation purposes.

If downloads fail reload this page, enable JavaScript, disable download managers, disable proxies, clear cache, and use Firefox. Please only download files manually as needed.

Most binaries are built from source code found on PyPI or in the projects public revision control systems. Source code changes, if any, have been submitted to the project maintainers or are included in the packages.

Refer to the documentation of the individual packages for license restrictions and dependencies.

Use pip version 8 or newer to install the downloaded .whl files. This page is not a pip package index.

Many binaries depend on numpy-1.11+mkl and the Microsoft Visual C++ 2008 (x64, x86, and SP1 for CPython 2.6 and 2.7), Visual C++ 2010 (x64, x86, for CPython 3.3 and 3.4), or the Visual C++ 2015 (x64 and x86 for CPython 3.5 and 3.6) redistributable packages.

Install numpy+mkl before other packages that depend on it.

The binaries are compatible with the official CPython distribution on Windows >=6.0. Chances are they do not work with custom Python distributions included with Blender, Maya, ArcGIS, OSGeo4W, ABAQUS, Cygwin, Pythonxy, Canopy, EPD, Anaconda, WinPython etc. Many binaries are not compatible with Windows XP or Wine.

The packages are ZIP or 7z files, which allows for manual or scripted installation or repackaging of the content.

The files are provided "as is" without warranty or support of any kind. The entire risk as to the quality and performance is with you.

Index by date: greenlet pygresql netcdf4 lxml pyamg jupyter cython liblinear cobra pybox2d fastcluster vlfd sfepy pytables h5py grako fonttools pymol pygame pyflux matplotlib spacy cytoolz apsw chainer mathutils veusz mercurial pyeda numpy cvxopt pywavelets pymongo gr persistent aiohttp pyodbc twisted ets vtk pocketsphinx simpleaudio pyaudio sounddevice fisx tensorflow multiprocess libsbml cvxcanon spectrum pyvrml97 ta-lib pythonmagick pyzmq triangle pgmagick ujson yappi pyfltk mod_wsgi pyfftw py_gd pyviennacl python-ldap openpiv pyx mpi4py pyephem pyemd planar mysqlclient xxhash zarr regex ode spyder lsqfit fann2 fisher ffnet entropy autopy slycot sparsesvd scs ecos sasl twainmodule dulwich datrie cx_oracle cyordereddict coverage cdecimal cartopy blz bigfloat aspell-python simpleparse milk menpo marisa-trie llist setproctitle hddm hmmlearn seqlearn jsonlib rtree rtmidi-python udunits heatmap scikit-umfpack scikits.vectorplot kwant tinyarray rpy2 fiona cx_freeze opencv netifaces multineat basemap py-earth pulp mlpy reportlab pyminuit pymetis python-snappy python-lzo python-levenshtein python-lz4 pystemmer

Reference

Data Visualization in Python

- http://it.sohu.com/20151119/n427117609.shtml
- http://www.oschina.net/translate/python-data-visualization-libraries

Reference

Using WordCloud

- – http://blog.csdn.net/tanzuozhev/article/details/50789226
- – https://www.oschina.net/code/snippet_2294527_56155

Chinese Display

- – http://blog.csdn.net/u012705410/article/details/47379957

Reference

Provided Repositories for Social Mining

- http://socialcomputing.asu.edu
- http://snap.Stanford.edu
- https://github.com/caesar0301/awesome-public-datasets

The End of Lecture 12

Thank You

http://www.wangting.ac.cn