**工作笔记**

目录

[1. 数学基础 3](#_Toc14620577)

[1.1. 三角函数 3](#_Toc14620578)

[直角三角形三角函数定义 4](#_Toc14620579)

[1.1.1. 正弦函数 4](#_Toc14620580)

[1.1.2. 余弦函数 4](#_Toc14620581)

[1.1.3. 正切函数 4](#_Toc14620582)

[1.1.4. 余切函数 4](#_Toc14620583)

[1.1.5. 正割函数 4](#_Toc14620584)

[1.1.6. 余割函数 4](#_Toc14620585)

[1.2. 高等数学基础 5](#_Toc14620586)

[1.2.1. 导数 5](#_Toc14620587)

[1.2.2. 偏导数 5](#_Toc14620588)

[1.2.3. 方向倒数 5](#_Toc14620589)

[1.2.4. 剃度 5](#_Toc14620590)

[1.3. 微积分 5](#_Toc14620591)

[1.3.1. 基本概念 5](#_Toc14620592)

[1.3.2. 定积分 6](#_Toc14620593)

[1.3.3. 定积分性质 6](#_Toc14620594)

[1.3.4. 牛顿-莱布尼茨公式 6](#_Toc14620595)

[1.4. 泰勒公式与拉格朗日 6](#_Toc14620596)

[1.4.1. 基本概念 6](#_Toc14620597)

[1.4.2. 一点一世界 7](#_Toc14620598)

[1.4.3. 阶数的作用 7](#_Toc14620599)

[1.4.4. 阶乘的作用 7](#_Toc14620600)

[1.4.5. 拉格朗日乘子法 7](#_Toc14620601)

[1.4.6. 求解拉格朗日乘子法 7](#_Toc14620602)

[1.5. 线性代数基础 7](#_Toc14620603)

[1.6. 特征值与矩阵分解 8](#_Toc14620604)

[1.6.1. 特征值与特征向量 8](#_Toc14620605)

[1.6.2. 特征空间与应用 8](#_Toc14620606)

[1.6.3. SVD要解决的问题 8](#_Toc14620607)

[1.6.4. 特征值分解 8](#_Toc14620608)

[1.6.5. SVD矩阵分解 8](#_Toc14620609)

[1.7. 随机变量与概率估计 8](#_Toc14620610)

[1.7.1. 离散型随机变量 8](#_Toc14620611)

[1.7.2. 连续型随机变量 8](#_Toc14620612)

[1.7.3. 简单随机抽样 8](#_Toc14620613)

[1.7.4. 似然函数 8](#_Toc14620614)

[1.7.5. 极大似然估计 9](#_Toc14620615)

[1.8. 概率论基础 9](#_Toc14620616)

[1.8.1. 概率与频率 9](#_Toc14620617)

[1.8.2. 古典概型 9](#_Toc14620618)

[1.8.3. 条件概率 9](#_Toc14620619)

[1.8.4. 二维离散随机变量 11](#_Toc14620620)

[1.8.5. 二维连续型随机变量 11](#_Toc14620621)

[1.8.6. 边缘分布 11](#_Toc14620622)

[1.8.7. 期望 11](#_Toc14620623)

[1.8.8. 期望求解 11](#_Toc14620624)

[1.8.9. 马尔科夫不等式 11](#_Toc14620625)

[1.8.10. 切比雪夫不等式 11](#_Toc14620626)

[1.8.11. 后验概率估计 11](#_Toc14620627)

[1.8.12. 贝叶斯拼写纠错实例 12](#_Toc14620628)

[1.8.13. 垃圾邮件过滤实例 12](#_Toc14620629)

[1.9. 数学科学常见分布 12](#_Toc14620630)

[1.9.1. 正太分布 12](#_Toc14620631)

[1.9.2. 二项式分布 14](#_Toc14620632)

[1.9.3. 泊松分布 15](#_Toc14620633)

[1.9.4. 均匀分布 16](#_Toc14620634)

[1.9.5. 卡方分布 16](#_Toc14620635)

[1.9.6. Beta分布 17](#_Toc14620636)

[1.10. 核函数变换 17](#_Toc14620637)

[1.10.1. 核函数的目的 17](#_Toc14620638)

[1.10.2. 线性函数 17](#_Toc14620639)

[1.10.3. 多项式核函数 18](#_Toc14620640)

[1.10.4. 核函数实例 18](#_Toc14620641)

[1.10.5. 高斯核函数 18](#_Toc14620642)

[1.10.6. 参数的影响 19](#_Toc14620643)

[1.11. 熵与激活函数 19](#_Toc14620644)

[1.11.1. 熵的概念 19](#_Toc14620645)

[1.11.2. 熵的大小意义是什么 21](#_Toc14620646)

[1.11.3. 激活函数 21](#_Toc14620647)

[1.11.4. 激活函数的问题 21](#_Toc14620648)

[1.11.5. 回归分析 21](#_Toc14620649)

[1.11.6. 回归分析概述 21](#_Toc14620650)

[1.11.7. 回归方程定义 21](#_Toc14620651)

[1.11.8. 误差项定义 21](#_Toc14620652)

[1.11.9. 最小二乘法推导与求解 21](#_Toc14620653)

[1.11.10. 回归方程求解小例子 21](#_Toc14620654)

[1.11.11. 回归直线拟合优度 21](#_Toc14620655)

[1.11.12. 多元与曲线回归问题 21](#_Toc14620656)

[1.12. 假设检验与相关分析 21](#_Toc14620657)

[1.12.1. 假设检验基本思想 21](#_Toc14620658)

[1.12.2. Z检验基本原理 21](#_Toc14620659)

[1.12.3. T检验基本原理 21](#_Toc14620660)

[1.12.4. 相关分析概念 21](#_Toc14620661)

[1.12.5. 方差基本概念 22](#_Toc14620662)

[2. 人工智能的知识点 22](#_Toc14620663)

[2.1. 基础知识 22](#_Toc14620664)

[2.1.1. Spark 22](#_Toc14620665)

[2.1.2. Graphx 22](#_Toc14620666)

[2.1.3. neo4j 23](#_Toc14620667)

[7.1.1. OrientDB 23](#_Toc14620668)

[7.1.2. Keras 24](#_Toc14620669)

[7.1.3. Tensorflow 24](#_Toc14620670)

[8.1.1. Seq2Seq 27](#_Toc14620671)

[8.1.2. FastText 28](#_Toc14620672)

[8.1.3. word2vector 28](#_Toc14620673)

[8.1.4. LDA 29](#_Toc14620674)

[8.1.5. PyTorch 29](#_Toc14620675)

[8.1.6. Caffe（卷积神经网络框架） 29](#_Toc14620676)

[8.1.7. Theano 30](#_Toc14620677)

[8.1.8. 实体抽取 30](#_Toc14620678)

[8.1.9. 关系抽取 31](#_Toc14620679)

[8.1.10. 实体对齐 32](#_Toc14620680)

[8.1.11. 实体消岐 32](#_Toc14620681)

[8.1.12. 知识挖掘 33](#_Toc14620682)

[8.1.13. 意图识别 33](#_Toc14620683)

[8.1.14. 情感分析 34](#_Toc14620684)

[8.1.15. 语义理解 34](#_Toc14620685)

[8.1.16. 倾向性分析 36](#_Toc14620686)

[8.1.17. 词性标注 36](#_Toc14620687)

[8.1.18. 文本分类聚类 36](#_Toc14620688)

[8.1.19. 热词发现 37](#_Toc14620689)

[8.1.20. 关键词抽取 38](#_Toc14620690)

[8.1.21. 文本纠错 49](#_Toc14620691)

[8.1.22. 语言模型 50](#_Toc14620692)

[8.1.23. 问答系统 50](#_Toc14620693)

[8.1.24. 知识推理 50](#_Toc14620694)

[8.1.25. 自动摘要 50](#_Toc14620695)

[2.2. 算法相关 50](#_Toc14620696)

[2.2.1. CNN（卷积神经网络） 50](#_Toc14620697)

[2.2.2. DNN（深度神经网络算法） 50](#_Toc14620698)

[2.2.3. RNN（循环神经网络） 50](#_Toc14620699)

[2.2.4. LSTM 50](#_Toc14620700)

[2.2.5. CTC 51](#_Toc14620701)

[2.3. 应用相关 51](#_Toc14620702)

[2.3.1. 语音识别 51](#_Toc14620703)

[2.3.2. 图像识别 51](#_Toc14620704)

[2.3.3. 知识图谱 51](#_Toc14620705)

[2.3.4. 智能问答 52](#_Toc14620706)

[2.3.5. 聊天机器人 52](#_Toc14620707)

[3. 人工智能面试题 52](#_Toc14620708)

[3.1. Python基本面试题 52](#_Toc14620709)

[3.1.1. python的基本数据类型有哪些 52](#_Toc14620710)

[3.2. 基本面试题 52](#_Toc14620711)

[3.2.1. 深度学习框架TensorFlow中有哪四种常用交叉熵？ 52](#_Toc14620712)

[3.2.2. 什么叫过拟合，避免过拟合都有哪些措施？ 52](#_Toc14620713)

[3.2.3. 过拟合和欠拟合及解决方法： 53](#_Toc14620714)

[3.2.4. 什么是核函数？ 53](#_Toc14620715)

[3.2.5. 深度学习框架TensorFlow中常见的核函数都有哪些？ 53](#_Toc14620716)

[3.2.6. 朴素贝叶斯方法的优势是什么？ 53](#_Toc14620717)

[3.2.7. 什么是监督学习的标准方法？ 54](#_Toc14620718)

[3.2.8. 在机器学习中，模型的选择是指什么？ 54](#_Toc14620719)

[3.2.9. 图形数据库Neo4J的优劣势？ 54](#_Toc14620720)

[3.2.10. LR和SVM的联系与区别是什么？ 54](#_Toc14620721)

[3.2.11. 什么是聚类，聚类的应用场景？ 54](#_Toc14620722)

[4. 技术预研 55](#_Toc14620723)

[4.1. 模型训练 55](#_Toc14620724)

[4.1.1. CRF模型训练 55](#_Toc14620725)

[4.2. MongoDB 55](#_Toc14620726)

[4.3. Spark 55](#_Toc14620727)

[4.4. OrientDB 55](#_Toc14620728)

[4.5. MySql集群搭建 55](#_Toc14620729)

[4.6. Redis 55](#_Toc14620730)

[4.7. rabbitmq 55](#_Toc14620731)

[5. 整合Acitiviti在线流程设计器(未实现) 55](#_Toc14620732)

[1.1. 月工作日志 56](#_Toc14620733)

[10.1.1. 周工作日志 56](#_Toc14620734)

[10.1.1.1. 2019-06-18 56](#_Toc14620735)

[10.1.1.2. 2019-06-19 60](#_Toc14620736)

[10.1.1.3. 2019-06-20 62](#_Toc14620737)

[10.1.1.4. 2019-06-21 64](#_Toc14620738)

[10.1.1.5. 2019-06-24 65](#_Toc14620739)

[10.1.1.6. 2019-06-25 65](#_Toc14620740)

新版模型设计器

<https://github.com/Activiti/activiti-modeling-app>

<https://github.com/bpmn-io/bpmn-js-examples/tree/master/modeler>

开源bpmn流程设计<http://bpmn.io/>

<https://demo.bpmn.io/>

参考文章

<https://blog.csdn.net/u013488847/article/details/89429643>

开源bpm流程设计器

http://www.hotent.com/h-col-110.html?bd\_vid=10693952536568441774

1.数据模型；

2.表单；

<https://www.cnblogs.com/lyh421/p/6483125.html>

外置表单

<https://blog.csdn.net/hejingyuan6/article/details/46412285>

流程系统参考

<http://sso.hotent.com/login>

http://bpmhome.cn//html/kehu/

http://www.h3bpm.net/Portal/admin/index.html

3.流程图；

npm run serve

# 数学基础

## 三角函数

三角函数是[基本初等函数](https://baike.baidu.com/item/%E5%9F%BA%E6%9C%AC%E5%88%9D%E7%AD%89%E5%87%BD%E6%95%B0/6608669)之一，是以角度（数学上最常用弧度制，下同）为[自变量](https://baike.baidu.com/item/%E8%87%AA%E5%8F%98%E9%87%8F/6895256)，角度对应[任意角](https://baike.baidu.com/item/%E4%BB%BB%E6%84%8F%E8%A7%92/448676)终边与[单位圆](https://baike.baidu.com/item/%E5%8D%95%E4%BD%8D%E5%9C%86/2487023)交点坐标或其比值为[因变量](https://baike.baidu.com/item/%E5%9B%A0%E5%8F%98%E9%87%8F)的函数。也可以等价地用与[单位圆](https://baike.baidu.com/item/%E5%8D%95%E4%BD%8D%E5%9C%86/2487023)有关的各种线段的长度来定义。三角函数在研究三角形和[圆](https://baike.baidu.com/item/%E5%9C%86/54084)等几何形状的性质时有重要作用，也是研究周期性现象的基础数学工具。在[数学分析](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6%E5%88%86%E6%9E%90/3123)中，三角函数也被定义为无穷级数或特定微分方程的解，允许它们的取值扩展到任意实数值，甚至是[复数](https://baike.baidu.com/item/%E5%A4%8D%E6%95%B0)值

### 直角三角形三角函数定义

在直角三角形中，当平面上的三点A、B、C的连线，AB、AC、BC，构成一个[直角三角形](https://baike.baidu.com/item/%E7%9B%B4%E8%A7%92%E4%B8%89%E8%A7%92%E5%BD%A2)，其中∠ACB为[直角](https://baike.baidu.com/item/%E7%9B%B4%E8%A7%92)。对∠BAC而言，[对边](https://baike.baidu.com/item/%E5%AF%B9%E8%BE%B9)（opposite）a=BC、[斜边](https://baike.baidu.com/item/%E6%96%9C%E8%BE%B9)（hypotenuse）c=AB、邻边（adjacent）b=AC，则存在以下关系：

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **基本函数** | **英文** | **缩写** | **表达式** | **语言描述** | [三角形](https://baike.baidu.com/pic/%E4%B8%89%E8%A7%92%E5%87%BD%E6%95%B0/1652457/0/9825bc315c6034a8eb16696fc81349540823766c?fr=lemma&ct=single)三角形 |
| [正弦函数](https://baike.baidu.com/item/%E6%AD%A3%E5%BC%A6%E5%87%BD%E6%95%B0) | ***sine*** | *sin* | *a/c* | *∠A*的对边比斜边 |
| [余弦函数](https://baike.baidu.com/item/%E4%BD%99%E5%BC%A6%E5%87%BD%E6%95%B0) | ***cosine*** | *cos* | *b/c* | *∠A*的邻边比斜边 |
| [正切函数](https://baike.baidu.com/item/%E6%AD%A3%E5%88%87%E5%87%BD%E6%95%B0) | ***tangent*** | *tan* | *a/b* | *∠A*的对边比邻边 |
| [余切函数](https://baike.baidu.com/item/%E4%BD%99%E5%88%87%E5%87%BD%E6%95%B0) | ***cotangent*** | *cot* | *b/a* | *∠A*的邻边比对边 |
| [正割函数](https://baike.baidu.com/item/%E6%AD%A3%E5%89%B2%E5%87%BD%E6%95%B0) | ***secant*** | *sec* | *c/b* | *∠A*的斜边比邻边 |
| [余割函数](https://baike.baidu.com/item/%E4%BD%99%E5%89%B2%E5%87%BD%E6%95%B0) | ***cosecant*** | *csc* | *c/a* | *∠A*的斜边比对边 |

注：正切函数、余切函数曾被写作*tg*、*ctg，*现已不用这种写法*。*

### [正弦函数](https://baike.baidu.com/item/%E6%AD%A3%E5%BC%A6%E5%87%BD%E6%95%B0)

### [余弦函数](https://baike.baidu.com/item/%E4%BD%99%E5%BC%A6%E5%87%BD%E6%95%B0)

### [正切函数](https://baike.baidu.com/item/%E6%AD%A3%E5%88%87%E5%87%BD%E6%95%B0)

### [余切函数](https://baike.baidu.com/item/%E4%BD%99%E5%88%87%E5%87%BD%E6%95%B0)

### [正割函数](https://baike.baidu.com/item/%E6%AD%A3%E5%89%B2%E5%87%BD%E6%95%B0)

### [余割函数](https://baike.baidu.com/item/%E4%BD%99%E5%89%B2%E5%87%BD%E6%95%B0)

## 高等数学基础

### 导数

1. 导数（Derivative），也叫导函数值。又名[微商](https://baike.baidu.com/item/%E5%BE%AE%E5%95%86/11036713)，是[微积分](https://baike.baidu.com/item/%E5%BE%AE%E7%A7%AF%E5%88%86/6065)中的重要基础概念。当函数y=f（x）的[自变量](https://baike.baidu.com/item/%E8%87%AA%E5%8F%98%E9%87%8F/6895256)x在一点x0上产生一个增量Δx时，函数输出值的增量Δy与自变量增量Δx的比值在Δx趋于0时的[极限](https://baike.baidu.com/item/%E6%9E%81%E9%99%90/3564509)a如果存在，a即为在x0处的导数，记作f'（x0）或df（x0）/dx。
2. 导数是函数的局部性质。一个函数在某一点的导数描述了这个函数在这一点附近的变化率。如果函数的自变量和取值都是实数的话，函数在某一点的导数就是该函数所代表的曲线在这一点上的[切线](https://baike.baidu.com/item/%E5%88%87%E7%BA%BF/674562)[斜率](https://baike.baidu.com/item/%E6%96%9C%E7%8E%87/4914111)。导数的本质是通过极限的概念对函数进行局部的线性逼近。例如在[运动学](https://baike.baidu.com/item/%E8%BF%90%E5%8A%A8%E5%AD%A6/650696)中，物体的[位移](https://baike.baidu.com/item/%E4%BD%8D%E7%A7%BB/823868)对于时间的导数就是物体的[瞬时速度](https://baike.baidu.com/item/%E7%9E%AC%E6%97%B6%E9%80%9F%E5%BA%A6/972057)。
3. 不是所有的函数都有导数，一个函数也不一定在所有的点上都有导数。若某函数在某一点导数存在，则称其在这一点[可导](https://baike.baidu.com/item/%E5%8F%AF%E5%AF%BC/11034968)，否则称为不可导。然而，可导的函数一定[连续](https://baike.baidu.com/item/%E8%BF%9E%E7%BB%AD/6532794)；不连续的函数一定不可导。
4. 对于可导的函数f(x)，x↦f'(x)也是一个函数，称作f(x)的[导函数](https://baike.baidu.com/item/%E5%AF%BC%E5%87%BD%E6%95%B0/185399)（简称导数）。寻找已知的函数在某点的导数或其导函数的过程称为[求导](https://baike.baidu.com/item/%E6%B1%82%E5%AF%BC/1063861)。实质上，求导就是一个求极限的过程，导数的四则运算法则也来源于极限的四则运算法则。反之，已知导函数也可以倒过来求原来的函数，即[不定积分](https://baike.baidu.com/item/%E4%B8%8D%E5%AE%9A%E7%A7%AF%E5%88%86/6082893)。[微积分基本定理](https://baike.baidu.com/item/%E5%BE%AE%E7%A7%AF%E5%88%86%E5%9F%BA%E6%9C%AC%E5%AE%9A%E7%90%86/10350012)说明了求原函数与积分是等价的。求导和积分是一对互逆的操作，它们都是微积分学中最为基础的概念。

### 偏导数

### 方向倒数

在函数[定义域](https://baike.baidu.com/item/%E5%AE%9A%E4%B9%89%E5%9F%9F)的内点，对某一方向求导得到的[导数](https://baike.baidu.com/item/%E5%AF%BC%E6%95%B0/579188)。一般为[二元函数](https://baike.baidu.com/item/%E4%BA%8C%E5%85%83%E5%87%BD%E6%95%B0/2582023)和三元函数的方向导数，方向导数可分为沿直线方向和沿曲线方向的方向导数。

### 剃度

## 微积分

### 基本概念

微积分（Calculus），数学概念，是[高等数学](https://baike.baidu.com/item/%E9%AB%98%E7%AD%89%E6%95%B0%E5%AD%A6/1182)中研究函数的[微分](https://baike.baidu.com/item/%E5%BE%AE%E5%88%86/317988)(Differentiation)、[积分](https://baike.baidu.com/item/%E7%A7%AF%E5%88%86/5749068)(Integration)以及有关概念和应用的数学分支。它是数学的一个[基础学科](https://baike.baidu.com/item/%E5%9F%BA%E7%A1%80%E5%AD%A6%E7%A7%91/7809018)，内容主要包括[极限](https://baike.baidu.com/item/%E6%9E%81%E9%99%90/3564509)、[微分学](https://baike.baidu.com/item/%E5%BE%AE%E5%88%86%E5%AD%A6/5587083)、积分学及其应用。微分学包括[求导](https://baike.baidu.com/item/%E6%B1%82%E5%AF%BC)数的运算，是一套关于变化率的理论。它使得函数、速度、加速度和曲线的[斜率](https://baike.baidu.com/item/%E6%96%9C%E7%8E%87/4914111)等均可用一套通用的符号进行讨论。[积分学](https://baike.baidu.com/item/%E7%A7%AF%E5%88%86%E5%AD%A6/10350529)，包括求积分的运算，为定义和计算面积、[体积](https://baike.baidu.com/item/%E4%BD%93%E7%A7%AF/4970881)等提供一套通用的方法

### 定积分

**定积分 （definite integral）**

定积分就是求函数f(X)在区间[a,b]中的图像包围的面积。即由 y=0,x=a,x=b,y=f(X)所围成图形的面积。这个图形称为[曲边梯形](https://baike.baidu.com/item/%E6%9B%B2%E8%BE%B9%E6%A2%AF%E5%BD%A2)，特例是[曲边三角形](https://baike.baidu.com/item/%E6%9B%B2%E8%BE%B9%E4%B8%89%E8%A7%92%E5%BD%A2)

å®ç§¯å

定积分是[积分](https://baike.baidu.com/item/%E7%A7%AF%E5%88%86/5749068)的一种，是函数f(x)在区间[a,b]上积分和的[极限](https://baike.baidu.com/item/%E6%9E%81%E9%99%90/3564509)；这里应注意定积分与不定积分之间的关系：若定积分存在，则它是一个具体的数值（曲边梯形的面积），而不定积分是一个函数表达式，它们仅仅在数学上有一个计算关系（[牛顿-莱布尼茨公式](https://baike.baidu.com/item/%E7%89%9B%E9%A1%BF-%E8%8E%B1%E5%B8%83%E5%B0%BC%E8%8C%A8%E5%85%AC%E5%BC%8F)），其它一点关系都没有！一个函数，可以存在不定积分，而不存在定积分；也可以存在定积分，而不存在不定积分。一个连续函数，一定存在定积分和不定积分；若只有有限个间断点，则定积分存在；若有跳跃间断点，则原函数一定不存在，即不定积分一定不存在。

**不定积分（Indefinite integral）**

### 定积分性质

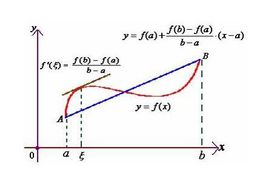
### 牛顿-莱布尼茨公式

1. 牛顿-莱布尼兹公式（Newton-Leibniz formula），通常也被称为微积分基本定理，揭示了[定积分](https://baike.baidu.com/item/%E5%AE%9A%E7%A7%AF%E5%88%86/7128801)与被积函数的原函数或者[不定积分](https://baike.baidu.com/item/%E4%B8%8D%E5%AE%9A%E7%A7%AF%E5%88%86/6082893)之间的联系；牛顿-莱布尼茨公式的内容是一个[连续函数](https://baike.baidu.com/item/%E8%BF%9E%E7%BB%AD%E5%87%BD%E6%95%B0/2716812)在区间 [ a，b ] 上的[定积分](https://baike.baidu.com/item/%E5%AE%9A%E7%A7%AF%E5%88%86/7128801)等于它的任意一个[原函数](https://baike.baidu.com/item/%E5%8E%9F%E5%87%BD%E6%95%B0/2749968)在区间[ a，b ]上的增量。[牛顿](https://baike.baidu.com/item/%E7%89%9B%E9%A1%BF/5463)在1666年写的《[流数](https://baike.baidu.com/item/%E6%B5%81%E6%95%B0)简论》中利用运动学描述了这一公式，[2]  1677年,[莱布尼茨](https://baike.baidu.com/item/%E8%8E%B1%E5%B8%83%E5%B0%BC%E8%8C%A8/417549)在一篇手稿中正式提出了这一公式。[1]  因为二者最早发现了这一公式，于是命名为牛顿-莱布尼茨公式。

## 泰勒公式与拉格朗日

### 基本概念

泰勒公式，应用于[数学](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6/107037)、[物理](https://baike.baidu.com/item/%E7%89%A9%E7%90%86/127205)领域，是一个用[函数](https://baike.baidu.com/item/%E5%87%BD%E6%95%B0/301912)在某点的信息描述其附近取值的公式。如果函数足够[平滑](https://baike.baidu.com/item/%E5%B9%B3%E6%BB%91/560358)的话，在已知函数在某一点的各阶[导数](https://baike.baidu.com/item/%E5%AF%BC%E6%95%B0)值的情况之下，泰勒公式可以用这些导数值做系数构建一个多项式来近似函数在这一点的邻域中的值。泰勒公式还给出了这个多项式和实际的函数值之间的偏差。



[拉格朗日](https://baike.baidu.com/item/%E6%8B%89%E6%A0%BC%E6%9C%97%E6%97%A5)中值定理又称拉氏定理，是微分学中的基本定理之一，它反映了可导函数在闭区间上的整体的平均变化率与区间内某点的局部变化率的关系。[拉格朗日](https://baike.baidu.com/item/%E6%8B%89%E6%A0%BC%E6%9C%97%E6%97%A5)中值定理是[罗尔中值定理](https://baike.baidu.com/item/%E7%BD%97%E5%B0%94%E4%B8%AD%E5%80%BC%E5%AE%9A%E7%90%86/1876399)的推广，同时也是[柯西中值定理](https://baike.baidu.com/item/%E6%9F%AF%E8%A5%BF%E4%B8%AD%E5%80%BC%E5%AE%9A%E7%90%86/7588018)的特殊情形，是[泰勒公式](https://baike.baidu.com/item/%E6%B3%B0%E5%8B%92%E5%85%AC%E5%BC%8F/7681487)的弱形式（一阶展开）

### 一点一世界

### 阶数的作用

阶数只代表[正方形](https://baike.baidu.com/item/%E6%AD%A3%E6%96%B9%E5%BD%A2/1019418)矩阵的大小，并没有太多的意义；

### 阶乘的作用

### 拉格朗日乘子法

### 求解拉格朗日乘子法

## 线性代数基础

行列式概念

矩阵与数据的关系

矩阵基本操作

矩阵的几种变换

矩阵的

内积与正交

## 特征值与矩阵分解

### 特征值与特征向量

### 特征空间与应用

### SVD要解决的问题

### 特征值分解

### SVD矩阵分解

## 随机变量与概率估计

### 离散型随机变量

随机变量分为[离散型随机变量](https://baike.baidu.com/item/%E7%A6%BB%E6%95%A3%E5%9E%8B%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F/9980865)与 [连续型随机变量](https://baike.baidu.com/item/%E8%BF%9E%E7%BB%AD%E5%9E%8B%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F/3318213)两种，随机变量的函数仍为随机变量。有些随机变量,它全部可能取到的不相同的值是有限个或可列无限多个，也可以说概率1以一定的规律分布在各个可能值上。这种随机变量称为"离散型随机变量"

### 连续型随机变量

### 简单随机抽样

### 似然函数

定义：在[数理统计学](https://baike.baidu.com/item/%E6%95%B0%E7%90%86%E7%BB%9F%E8%AE%A1%E5%AD%A6)中，似然函数是一种关于[统计模型](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%A8%A1%E5%9E%8B)中的[参数](https://baike.baidu.com/item/%E5%8F%82%E6%95%B0)的[函数](https://baike.baidu.com/item/%E5%87%BD%E6%95%B0)，表示模型参数中的似然性；给定输出x时，关于参数θ的似然函数L(θ|x)（在数值上）等于给定参数θ后变量X的概率：https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D141/sign=12b0d9d968d0f703e2b291d839f85148/37d3d539b6003af3a22453ae3f2ac65c1138b666.jpg

### 极大似然估计

## 概率论基础

### 概率与频率

**频率**:描述的是事件发生的频繁程度；

**概率:表示某事件出现的可能性大小，最初通过频率稳定性来引出与求得**

### 古典概型

1. 定义

古典概型也叫传统概率、其定义是由[法国](https://baike.baidu.com/item/%E6%B3%95%E5%9B%BD/1173384)数学家[拉普拉斯](https://baike.baidu.com/item/%E6%8B%89%E6%99%AE%E6%8B%89%E6%96%AF/5189) (Laplace ) 提出的。如果一个随机试验所包含的单位事件是有限的，且每个单位事件发生的可能性均相等，则这个[随机试验](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E8%AF%95%E9%AA%8C/124845)叫做拉普拉斯试验，这种条件下的概率模型就叫古典概型;在这个模型下，随机实验所有可能的结果是有限的，并且每个基本结果发生的概率是相同的。古典概型是[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA/829122)中最直观和最简单的模型，概率的许多运算规则，也首先是在这种模型下得到的。

1. 一个试验是否为古典概型，在于这个试验是否具有古典概型的两个特征——有限性和等可能性，只有同时具备这两个特点的概型才是古典概型。
2. **有限性**（所有可能出现的基本事件只有有限个）
3. **等可能性**（每个基本事件出现的可能性相等）
4. P（A）= https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D31/sign=7fc362de7d310a55c024d8f5b6455042/5ab5c9ea15ce36d35dcc7b1d3ef33a87e850b1d7.jpg =A包含的基本事件的个数m/基本事件的总数n

### 条件概率

1. **定义**

条件概率是指事件A在另外一个[事件](https://baike.baidu.com/item/%E4%BA%8B%E4%BB%B6/33582)B已经发生条件下的发生[概率](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87/828845)。条件[概率](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87)表示为：P（A|B），读作“在B的条件下A的概率”。条件概率可以用[决策树](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%91/10377049)进行计算。条件概率的谬论是假设 P(A|B) 大致等于 P(B|A)。数学家John Allen Paulos 在他的《数学盲》一书中指出医生、律师以及其他受过很好教育的非统计学家经常会犯这样的错误。这种错误可以通过用实数而不是概率来描述数据的方法来避免。

1. **概率**[测度](https://baike.baidu.com/item/%E6%B5%8B%E5%BA%A6/12757014)

如果事件 *B* 的概率 *P*(*B*) > 0，那么 *Q*(*A*) = *P*(*A* | *B*) 在所有事件 *A* 上所定义的函数 *Q* 就是概率测度。 如果 *P*(*B*) = 0，*P*(*A* | *B*) 没有定义。 条件概率可以用决策树进行计算。[1]

1. **联合概率**

表示两个事件共同发生的概率。*A*与*B*的[联合概率](https://baike.baidu.com/item/%E8%81%94%E5%90%88%E6%A6%82%E7%8E%87/5901767)表示为 P(AB) 或者*P*(*A***,***B*),或者P(A∩B)。[2]

1. **边缘概率**

是某个事件发生的概率，而与其它事件无关。边缘概率是这样得到的：在联合概率中，把最终结果中不需要的那些事件合并成其事件的全概率而消失（对离散[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)用求和得全概率，对连续随机变量用积分得全概率）。这称为[**边缘化**](https://baike.baidu.com/item/%E8%BE%B9%E7%BC%98%E5%8C%96)（**marginalization**）。*A*的边缘概率表示为*P*(*A*)，*B*的边缘概率表示为*P*(*B*)。

**独立性**

1. 设A，B为随机事件，若同时发生的概率等于各自发生的概率的乘积，则A，B**相互独立**。
2. 一般地，设A1，A2，...，An是n(n≥2) 个事件，如果对于其中任意2个，任意3个，...，任意n个事件的积事件的概率，都等于各事件概率之积，则称A1，A2，...，An**相互独立**。
3. **例题解析**

例1 有两门高射炮独立地射击一架敌机，设甲炮击中敌机的概率为0.8，乙炮击中敌机的概率为0.7，试求敌机被击中的概率。[1] 

解： 设A={甲炮击中敌机}，B={乙炮击中敌机}，则A U B={敌机被击中}，由题意知，P(A)=0.8，P(B)=0.7，由于A，B相互独立。故

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D587/sign=f71da51f083b5bb5bad720f601d2d523/a2cc7cd98d1001e9e628b0bbb30e7bec54e7976a.jpg

例2 有甲、乙两批种子，发芽率分别为0.8和0.7，并假设每批种子发芽与否是相互独立的，从两批种子中各随机地抽取一粒，求：

(1)两粒都能发芽的概率；

(2)至少有一粒种子能发芽的概率；

(3)恰好有一粒种子能发芽的概率。

解: 设A={取自甲批种子中的某粒种子能发芽}，B={取自乙批种子中的某粒种子能发芽}，则所求的概率分别为：

https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D275/sign=0c044cae6b59252da7171a03019a032c/a2cc7cd98d1001e9ecfcbabbb30e7bec55e79746.jpg

。

由于 https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D156/sign=a465a67eedcd7b89ed6c3e8639254291/f9198618367adab43a8fbdf780d4b31c8701e423.jpg ，且 https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D29/sign=334a47ae2334349b7006698cc8eaf6a1/a71ea8d3fd1f4134816da3ca2e1f95cad1c85e2f.jpg 相互独立，故有：

https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D289/sign=1a02350f8cd6277fed12353011381f63/4e4a20a4462309f70d510ca6790e0cf3d7cad64c.jpg

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D428/sign=627e336bc51b9d168ec79b63cbdfb4eb/838ba61ea8d3fd1f558c038e3b4e251f94ca5ff8.jpg

https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D516/sign=e52d3ad5f8deb48fff69a1dfc61e3aef/960a304e251f95cac91f3202c2177f3e6609529f.jpg

### 二维离散随机变量

### 二维连续型随机变量

### 边缘分布

### 期望

### 期望求解

### 马尔科夫不等式

定义

在[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA/829122)中，马尔可夫不等式给出了[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F/828980)的函数大于等于某正数的概率的上界。虽然它以俄国数学家[安德雷·马尔可夫](https://baike.baidu.com/item/%E5%AE%89%E5%BE%B7%E9%9B%B7%C2%B7%E9%A9%AC%E5%B0%94%E5%8F%AF%E5%A4%AB/10445098)命名，但该不等式曾出现在一些更早的文献中，其中包括马尔可夫的老师--巴夫尼提·列波维奇·切比雪夫。

### 切比雪夫不等式

1. 定义

设X是一个随机变数取取区间（0，**∞**）上的值，F(x)是它的分布函数，设Xα（α >0）的数学期望M(Xα )存在，a>0，则不等式成立。这叫做切比雪夫定理，或者切比雪夫不等式。

### 后验概率估计

1. 定义

在贝叶斯[统计学](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E5%AD%A6)中，“最大后验概率估计”是[后验概率](https://baike.baidu.com/item/%E5%90%8E%E9%AA%8C%E6%A6%82%E7%8E%87)分布的[众数](https://baike.baidu.com/item/%E4%BC%97%E6%95%B0)。利用最大后验概率估计可以获得对实验数据中无法直接观察到的量的[点估计](https://baike.baidu.com/item/%E7%82%B9%E4%BC%B0%E8%AE%A1)。它与[最大似然估计](https://baike.baidu.com/item/%E6%9C%80%E5%A4%A7%E4%BC%BC%E7%84%B6%E4%BC%B0%E8%AE%A1)中的经典方法有密切关系，但是它使用了一个增广的优化目标，进一步考虑了被估计量的[先验概率](https://baike.baidu.com/item/%E5%85%88%E9%AA%8C%E6%A6%82%E7%8E%87)分布。所以最大后验概率估计可以看作是规则化（regularization）的最大似然估计。

### 贝叶斯拼写纠错实例

### 垃圾邮件过滤实例

## 数学科学常见分布

### 正太分布

1. 正态分布（Normal distribution），也称“常态分布”，又名[高斯分布](https://baike.baidu.com/item/%E9%AB%98%E6%96%AF%E5%88%86%E5%B8%83/10145793)（Gaussian distribution），最早由A.棣莫弗在求[二项分布](https://baike.baidu.com/item/%E4%BA%8C%E9%A1%B9%E5%88%86%E5%B8%83)的渐近公式中得到。C.F.高斯在研究测量误差时从另一个角度导出了它。P.S.拉普拉斯和高斯研究了它的性质。是一个在[数学](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6/107037)、物理及工程等领域都非常重要的[概率](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87)分布，在统计学的许多方面有着重大的影响力。
2. 正态曲线呈钟型，两头低，中间高，左右对称因其曲线呈钟形，因此人们又经常称之为[钟形曲线](https://baike.baidu.com/item/%E9%92%9F%E5%BD%A2%E6%9B%B2%E7%BA%BF/7736826)。
3. 若[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F/828980)X服从一个[数学期望](https://baike.baidu.com/item/%E6%95%B0%E5%AD%A6%E6%9C%9F%E6%9C%9B/5362790)为μ、[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE/3108412)为σ^2的正态分布，记为N(μ，σ^2)。其[概率密度函数](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0/5021996)为正态分布的[期望值](https://baike.baidu.com/item/%E6%9C%9F%E6%9C%9B%E5%80%BC/8664642)μ决定了其位置，其[标准差](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E5%B7%AE/1415772)σ决定了分布的幅度。当μ = 0,σ = 1时的正态分布是[标准正态分布](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83)。
4. 一维正态分布

若[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D12/sign=a0c8d1ef241f95caa2f596b4c8171210/267f9e2f0708283828f20885b999a9014d08f1a0.jpg

服从一个位置参数为

https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D13/sign=fc592cb0738b4710ca2ff9cfc1ce63c9/63d9f2d3572c11dfef146bb2622762d0f603c25e.jpg

、尺度参数为

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D10/sign=58e9702ea1ec08fa220017a758ee9a39/9c16fdfaaf51f3de1b656df595eef01f3b297948.jpg

的概率分布，且其[概率密度函数](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0)为[2] 

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D205/sign=2abf505a42166d223c77129473220945/342ac65c1038534384b650b09213b07eca808822.jpg

则这个[随机变量](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)就称为**正态随机变量**，正态随机变量服从的分布就称为**正态分布，记作**

https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D94/sign=6cbb75ab314e251fe6f7e8fca6861361/377adab44aed2e73a65b51bd8601a18b86d6faf5.jpg

，读作

https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D12/sign=293539f5574e9258a23482ec9d8206e4/8b82b9014a90f603539b4dd63812b31bb051ed7c.jpg

服从

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D64/sign=f0de936e3ac79f3d8be1e734bba16935/1e30e924b899a90175eb8da71c950a7b0308f541.jpg

，或

https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D12/sign=f99a29b0738b4710ca2ff9cec2ce638f/63d9f2d3572c11dfead76eb2622762d0f703c219.jpg

服从正态分布。

μ维随机[向量](https://baike.baidu.com/item/%E5%90%91%E9%87%8F)具有类似的概率规律时，称此随机向量遵从多维正态分布。多元正态分布有很好的性质，例如，多元正态分布的边缘分布仍为正态分布，它经任何[线性](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7)变换得到的随机向量仍为多维正态分布，特别它的线性组合为一元正态分布。

本词条的正态分布是一维正态分布，此外多维正态分布参见“[二维正态分布](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83)”。

1. 标准正态分布

当

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D79/sign=96bfd75b838ba61edbeeca2640341961/b21bb051f819861811f03e154bed2e738ad4e6e8.jpg

时，正态分布就成为**标准正态分布**

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D131/sign=aaa1da2e86cb39dbc5c06355e11709a7/728da9773912b31bc73473ed8118367adbb4e19e.jpg

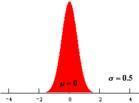
1. 图形特征

**集中性：**正态曲线的高峰位于正中央，即[均数](https://baike.baidu.com/item/%E5%9D%87%E6%95%B0)所在的位置。

**对称性**：正态曲线以均数为中心，左右对称，曲线两端永远不与[横轴](https://baike.baidu.com/item/%E6%A8%AA%E8%BD%B4)相交。

**均匀变动性**：正态曲线由均数所在处开始，分别向左右两侧逐渐均匀下降。

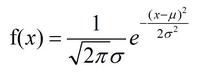
曲线与横轴间的面积总等于1，相当于[概率密度函数](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0)的函数从正无穷到负无穷积分的概率为1。即频率的总和为100%。

[](https://baike.baidu.com/pic/%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83/829892/0/4e83cb6229cf23e7e6113a12?fr=lemma&ct=single)正态分布

关于μ对称，并在μ处取最大值，在正（负）无穷远处取值为0，在μ±σ处有[拐点](https://baike.baidu.com/item/%E6%8B%90%E7%82%B9)，形状呈现中间高两边低，正态分布的概率密度函数[曲线](https://baike.baidu.com/item/%E6%9B%B2%E7%BA%BF/92103)呈钟形，因此人们又经常称之为**钟形曲线。**

1. 参数含义

正态分布有两个参数，即期望（均数）μ和标准差σ，σ2为方差。

[](https://baike.baidu.com/pic/%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83/829892/0/5d212aa89c244b8bca130c13?fr=lemma&ct=single)正态分布公式

正态分布具有两个参数μ和σ^2的[连续型随机变量](https://baike.baidu.com/item/%E8%BF%9E%E7%BB%AD%E5%9E%8B%E9%9A%8F%E6%9C%BA%E5%8F%98%E9%87%8F)的分布，第一参数μ是服从正态分布的随机变量的[均值](https://baike.baidu.com/item/%E5%9D%87%E5%80%BC)，第二个参数σ^2是此随机变量的[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE)，所以正态分布记作N（μ,σ2）。

μ是正态分布的位置参数，描述正态分布的[集中趋势](https://baike.baidu.com/item/%E9%9B%86%E4%B8%AD%E8%B6%8B%E5%8A%BF)位置。概率规律为取与μ邻近的值的概率大，而取离μ越远的值的概率越小。正态分布以X=μ为[对称轴](https://baike.baidu.com/item/%E5%AF%B9%E7%A7%B0%E8%BD%B4)，左右完全对称。正态分布的期望、[均数](https://baike.baidu.com/item/%E5%9D%87%E6%95%B0)、[中位数](https://baike.baidu.com/item/%E4%B8%AD%E4%BD%8D%E6%95%B0)、众数相同，均等于μ。

σ描述正态分布资料数据分布的离散程度，σ越大，数据分布越分散，σ越小，数据分布越集中。也称为是正态分布的形状参数，σ越大，曲线越扁平，反之，σ越小，曲线越瘦高。

**面积分布**

1、实际工作中，正态曲线下横轴上一定区间的面积反映该区间的例数占总例数的百分比，或变量值落在该区间的概率（概率分布）。不同 范围内正态曲线下的面积可用公式计算。

2、正态曲线下，[横轴](https://baike.baidu.com/item/%E6%A8%AA%E8%BD%B4)区间（μ-σ,μ+σ）内的面积为68.268949%。

P{|X-μ|<σ}=2Φ（1）-1=0.6826

横轴区间（μ-1.96σ,μ+1.96σ）内的面积为95.449974%。

P{|X-μ|<2σ}=2Φ（2）-1=0.9544

横轴区间（μ-2.58σ,μ+2.58σ）内的面积为99.730020%。

P{|X-μ|<3σ}=2Φ（3）-1=0.9974

由于**“小概率事件”**和[假设检验](https://baike.baidu.com/item/%E5%81%87%E8%AE%BE%E6%A3%80%E9%AA%8C)的基本思想 “小概率事件”通常指发生的概率小于5%的事件，认为在一次试验中该事件是几乎不可能发生的。由此可见X落在（μ-3σ,μ+3σ）以外的概率小于千分之三，在实际问题中常认为相应的事件是不会发生的，基本上可以把区间（μ-3σ,μ+3σ）看作是随机变量X实际可能的取值区间，这称之为正态分布的“3σ”原则。

1. **曲线应用**

1、估计频数分布 一个服从正态分布的变量只要知道其均数与标准差就可根据公式即可估计任意取值范围内频数比例。[3]

2、制定参考值范围

（1）正态分布法 适用于服从正态（或近似正态）分布指标以及可以通过转换后服从正态分布的指标。

（2）百分位数法 常用于偏态分布的指标。表3-1中两种方法的单双侧界值都应熟练掌握。

3、质量控制：为了控制实验中的测量（或实验）误差，常以 作为上、下警戒值，以 作为上、下控制值。这样做的依据是：正常情况下测量（或实验）误差服从正态分布。

/4、正态分布是许多统计方法的理论基础。检验、[方差分析](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE%E5%88%86%E6%9E%90)、相关和[回归分析](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E5%88%86%E6%9E%90)等多种[统计方法](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%96%B9%E6%B3%95)均要求分析的指标服从正态分布。许多统计方法虽然不要求分析指标服从正态分布，但相应的[统计量](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E9%87%8F)在大样本时近似正态分布，因而大样本时这些[统计推断](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%8E%A8%E6%96%AD)方法也是以正态分布为理论基础的。

### 二项式分布

1. **统计学定义**

在[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA)和统计学中，二项分布是n个独立的是/非试验中成功的次数的离散[概率分布](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83)，其中每次试验的成功概率为p。这样的单次成功/失败试验又称为伯努利试验。实际上，当

https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D35/sign=1b1a826edd2a60595610e71f2934fadc/18d8bc3eb13533fac89efd89a2d3fd1f41345b68.jpg **时，二项分布就是伯努利分布，二项分布是显著性差异的二项试验的基础**

1. **二项分布就是重复n次独立的**[**伯努利试验**](https://baike.baidu.com/item/%E4%BC%AF%E5%8A%AA%E5%88%A9%E8%AF%95%E9%AA%8C/238488)**。在每次试验中只有两种可能的结果，而且两种结果发生与否互相对立，并且相互**[**独立**](https://baike.baidu.com/item/%E7%8B%AC%E7%AB%8B/3415220)**，与其它各次试验结果无关，事件发生与否的概率在每一次**[**独立试验**](https://baike.baidu.com/item/%E7%8B%AC%E7%AB%8B%E8%AF%95%E9%AA%8C/12728918)**中都保持不变，则这一系列试验总称为n重伯努利实验，当试验次数为1时，二项分布服从0-1分布。**
2. **应用实例**

二项分布在心理与教育研究中，主要用于解决含有机遇性质的问题。所谓机遇问题，即指在实验或调查中，实验结果可能是由猜测而造成的。比如，选择题目的回答，划对划错，可能完全由猜测造成。凡此类问题，欲区分由猜测而造成的结果与真实的结果之间的界限，就要应用二项分布来解决

已知有正误题10题，问答题者答对几题才能认为他是真会，或者说答对几题，才能认为不是出于猜测因素?

分析：此题

https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D79/sign=7438a4f1e8cd7b89ed6c388a0f244602/08f790529822720eaccc46c175cb0a46f21fab8f.jpg

，即猜对猜错的概率各为0.5。

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D45/sign=1d2b4783da3f8794d7ff492bd21b7a2f/622762d0f703918f2f8299f15f3d269759eec4be.jpg

，故此二项分布接近正态分布：

https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D147/sign=d7a390cb51b5c9ea66f307e7e238b622/203fb80e7bec54e7d5e9bcdeb3389b504ec26aed.jpg

根据正态分布概率，当Z=1.645时，该点以下包含了全体的95%。如果用原分数表示，则为

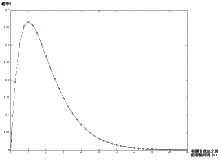
https://gss1.bdstatic.com/-vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D274/sign=5d5d0cd2780e0cf3a4f749fc3e47f23d/64380cd7912397dd6c4b39c45382b2b7d1a287c4.jpg

它的意义是，完全凭猜测，10题中猜对8题以下的可能性为95%，猜对8、9、10题的概率只5%。因此可以推论说，答对8题以上者不是凭猜测，而是会答。但应该明确：作此结论，也仍然有犯错误的可能，即那些完全靠猜测的人也有5%的可能性答对8、9、10道题。

### 泊松分布

1. **定义**

Poisson分布，是一种统计与概率学里常见到的离散[概率分布](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83)，由法国数学家西莫恩·德尼·泊松（Siméon-Denis Poisson）在1838年时发表。



1. **泊松分布与二项分布**

[泊松分布](https://baike.baidu.com/pic/%E6%B3%8A%E6%9D%BE%E5%88%86%E5%B8%83/1442110/0/fd428c45a7399c598694733c?fr=lemma&ct=single)

当二项分布的n很大而p很小时，泊松分布可作为二项分布的近似，其中λ为np。通常当n≧20,p≦0.05时，就可以用泊松公式近似得计算。

事实上，泊松分布正是由二项分布推导而来的，具体推导过程参见本词条相关部分。

1. **应用示例**

泊松分布适合于描述单位时间（或空间）内随机事件发生的次数。如某一服务设施在一定时间内到达的人数，电话交换机接到呼叫的次数，汽车站台的候客人数，机器出现的故障数，自然灾害发生的次数，一块产品上的缺陷数，显微镜下单位分区内的细菌分布数等等。

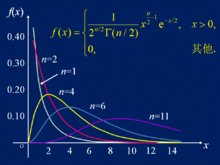
### 均匀分布

1. **定义**

在[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA/829122)和统计学中，均匀分布也叫矩形分布，它是对称概率分布，在相同长度间隔的分布概率是等可能的。 均匀分布由两个参数a和b定义，它们是数轴上的最小值和最大值，通常缩写为U（a，b）。

### 卡方分布

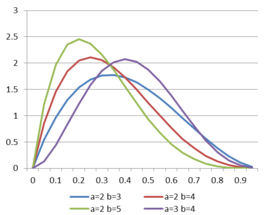
1. **定义**



若n个相互独立的随机变量ξ₁，ξ₂，...,ξn ，均服从[标准正态分布](https://baike.baidu.com/item/%E6%A0%87%E5%87%86%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83/552653)（也称独立同分布于标准[正态分布](https://baike.baidu.com/item/%E6%AD%A3%E6%80%81%E5%88%86%E5%B8%83)），则这n个服从标准正态分布的随机变量的平方和构成一新的随机变量，其分布规律称为卡方分布（chi-square distribution）。

### Beta分布

1. **定义**



贝塔分布（Beta Distribution) 是一个作为[伯努利分布](https://baike.baidu.com/item/%E4%BC%AF%E5%8A%AA%E5%88%A9%E5%88%86%E5%B8%83/7167021)和二项式分布的共轭先验分布的[密度函数](https://baike.baidu.com/item/%E5%AF%86%E5%BA%A6%E5%87%BD%E6%95%B0/12721265)，在机器学习和数理统计学中有重要应用。在[概率论](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E8%AE%BA/829122)中，**贝塔分布**，也称**Β分布，**是指一组定义在(0,1) 区间的连续[概率分布](https://baike.baidu.com/item/%E6%A6%82%E7%8E%87%E5%88%86%E5%B8%83)。

## 核函数变换

### 核函数的目的

### 线性函数

在数学里，线性函数是指那些[线性](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7/5450468)的函数，但也常用作[一次函数](https://baike.baidu.com/item/%E4%B8%80%E6%AC%A1%E5%87%BD%E6%95%B0/1653577)的别称，尽管一次函数不一定是线性的（那些不经过[原点](https://baike.baidu.com/item/%E5%8E%9F%E7%82%B9/8271864)的）。线型函数是一个比较恰当的同义词。

### 多项式核函数

### 核函数实例

### 高斯核函数

1. 高斯函数具有五个重要的性质，这些性质使得它在早期图像处理中特别有用．这些性质表明，高斯平滑滤波器无论在空间域还是在[频率域](https://baike.baidu.com/item/%E9%A2%91%E7%8E%87%E5%9F%9F)都是十分有效的低通[滤波器](https://baike.baidu.com/item/%E6%BB%A4%E6%B3%A2%E5%99%A8)，且在实际图像处理中得到了工程人员的有效使用．高斯函数具有五个十分重要的性质，它们是：
2. （1）[二维](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4)高斯函数具有旋转对称性，即滤波器在各个方向上的平滑程度是相同的．一般来说，一幅图像的边缘方向是事先不知道的，因此，在滤波前是无法确定一个方向上比另一方向上需要更多的平滑．旋转对称性意味着高斯平滑滤波器在后续[边缘检测](https://baike.baidu.com/item/%E8%BE%B9%E7%BC%98%E6%A3%80%E6%B5%8B)中不会偏向任一方向．
3. （2）高斯函数是[单值函数](https://baike.baidu.com/item/%E5%8D%95%E5%80%BC%E5%87%BD%E6%95%B0)．这表明，高斯滤波器用像素邻域的加权均值来代替该点的像素值，而每一邻域像素点[权值](https://baike.baidu.com/item/%E6%9D%83%E5%80%BC)是随该点与中心点的距离单调增减的．这一性质是很重要的，因为边缘是一种图像局部特征，如果平滑运算对离算子中心很远的像素点仍然有很大作用，则平滑运算会使图像失真．
4. （3）高斯函数的[傅立叶变换](https://baike.baidu.com/item/%E5%82%85%E7%AB%8B%E5%8F%B6%E5%8F%98%E6%8D%A2)频谱是单瓣的．正如下面所示，这一性质是高斯函数傅立叶变换等于高斯函数本身这一事实的直接推论．图像常被不希望的高频信号所污染(噪声和细纹理)．而所希望的图像特征（如边缘），既含有低频分量，又含有高频分量．高斯函数傅里叶变换的单瓣意味着平滑图像不会被不需要的高频信号所污染，同时保留了大部分所需信号．
5. （4）高斯滤波器宽度(决定着平滑程度)是由参数σ表征的，而且σ和平滑程度的关系是非常简单的．σ越大，高斯滤波器的[频带](https://baike.baidu.com/item/%E9%A2%91%E5%B8%A6)就越宽，平滑程度就越好．通过调节平滑程度参数σ，可在图像特征过分模糊(过平滑)与平滑图像中由于噪声和细纹理所引起的过多的不希望突变量(欠平滑)之间取得折中．
6. （5）由于[高斯函数](https://baike.baidu.com/item/%E9%AB%98%E6%96%AF%E5%87%BD%E6%95%B0)的可分离性，大[高斯滤波](https://baike.baidu.com/item/%E9%AB%98%E6%96%AF%E6%BB%A4%E6%B3%A2)器可以得以有效地实现．[二维](https://baike.baidu.com/item/%E4%BA%8C%E7%BB%B4)高斯函数[卷积](https://baike.baidu.com/item/%E5%8D%B7%E7%A7%AF)可以分两步来进行，首先将图像与一维高斯函数进行卷积，然后将卷积结果与方向垂直的相同一维高斯函数卷积．因此，二维高斯滤波的计算量随滤波模板宽度成[线性](https://baike.baidu.com/item/%E7%BA%BF%E6%80%A7)增长而不是成平方增长．

### 参数的影响

## 熵与激活函数

### 熵的概念

1. 信息论解释

1948年，[香农](https://baike.baidu.com/item/%E9%A6%99%E5%86%9C)将统计物理中熵的概念，引申到信道通信的过程中，从而开创了”信息论“这门学科。香农定义的“熵”又被称为“香农熵” 或 “信息熵”, 即

https://gss3.bdstatic.com/-Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D240/sign=8387bc3eac18972ba73a07ced6cc7b9d/8718367adab44aed7de23eceb41c8701a18bfb68.jpg

其中i标记概率空间中所有可能的样本，

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D14/sign=0ff310d67bd98d1072d40835213fd079/50da81cb39dbb6fd7348630c0e24ab18972b378c.jpg

表示该样本的出现几率，K是和单位选取相关的任意常数。可以明显看出“信息熵”的定义和“热力学熵”（玻尔兹曼公式）的定义只相差某个比例常数。数学上，可以证明“香农熵”的定义，具有以下良好性质：

**连续性**

该度量应该是连续的，即，若样本概率值有微小变化，由此引起的熵变化也是微小的。

**对称性**

样本重新排序后，该度量应保持不变，即

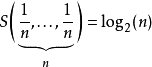
https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D186/sign=e514cb04d51b0ef468e89c56ebc551a1/b219ebc4b74543a9c3f079e619178a82b9011434.jpg

**极值性**

当所有样本等几率出现的情况下，熵达到最大值（所有可能的事件等概率时不确定性最高）

https://gss2.bdstatic.com/9fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D255/sign=10edd53d65d0f703e2b292d93dfb5148/359b033b5bb5c9ea8f3080c9d239b6003bf3b3dd.jpg

对于样本等几率分布而言，样本数越大，熵值越大（可能的事件越多，不确定性越高）



**可加性**

熵的值与过程如何被划分无关。它描述了一个系统与其子系统熵的关系。如果子系统之间的相互作用是已知的，则可以通过子系统的熵来计算一个系统的熵。例如：给定一个有n个样本的均匀分布集合，分为k个箱子（子系统），每个里面有 b1, ..., bk 个样本，合起来的熵应等于系统的熵与各个箱子的熵的和，每个箱子的权重为在该箱中样本的总概率。即，对于正整数bi其中b1 + ... + bk = n来说，

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D370/sign=14aac8134c36acaf5de090fb4cd98d03/0823dd54564e92580e4c0dc49b82d158ccbf4ea4.jpg

其中S的脚标，标记对应概率空间的样本点个数。

事实上，香农证明如果要求度量满足这些性质，则可以完全确定“信息熵”的定义表达式。

### 熵的大小意义是什么

### 激活函数

### 激活函数的问题

### 回归分析

### 回归分析概述

### 回归方程定义

### 误差项定义

### 最小二乘法推导与求解

### 回归方程求解小例子

### 回归直线拟合优度

### 多元与曲线回归问题

## 假设检验与相关分析

### 假设检验基本思想

### Z检验基本原理

### T检验基本原理

### 相关分析概念

### 方差基本概念

# 人工智能的知识点

## 基础知识

### Spark

1. Apache Spark 是专为大规模数据处理而设计的快速通用的计算引擎；
2. Spark，拥有Hadoop MapReduce所具有的优点；但不同于MapReduce的是——Job中间输出结果可以保存在内存中，从而不再需要读写HDFS，因此Spark能更好地适用于数据挖掘与机器学习等需要迭代的MapReduce的算法；
3. 高级 API 剥离了对集群本身的关注，Spark 应用开发者可以专注于应用所要做的计算本身。
4. 其次，Spark 很快，支持交互式计算和复杂算法。
5. 最后，Spark 是一个通用引擎，可用它来完成各种各样的运算，包括 SQL 查询、文本处理、机器学习等，而在 Spark 出现之前，我们一般需要学习各种各样的引擎来分别处理这些需求。
6. 更快的速度：内存计算下，Spark 比 Hadoop 快100倍；
7. 易用性：Spark 提供了80多个高级运算符。
8. 通用性：Spark 提供了大量的库，包括Spark Core、Spark SQL、Spark Streaming、MLlib、GraphX。 开发者可以在同一个应用程序中无缝组合使用这些库。

### Graphx

1. GraphX是Spark中用于图和图计算的组件，GraphX通过扩展Spark RDD引入了一个新的图抽象数据结构，一个将有效信息放入顶点和边的有向多重图。
2. Graphx最大的贡献，也是大多数开发喜欢它的原因是，在Spark之上提供了一站式解决方案，可以方便且高效地完成图计算的一整套流水作业；即在实际开发中，可以使用核心模块来完成海量数据的清洗与与分析阶段，SQL模块来打通与数据仓库的通道，Streaming打造实时流处理通道，基于GraphX图计算算法来对网页中复杂的业务关系进行计算，最后使用MLLib以及SparkR来完成数据挖掘算法处理。

### neo4j

1. Neo4j的特点

* SQL就像简单的查询语言Neo4j CQL
* 它遵循属性图数据模型
* 它通过使用Apache Lucence支持索引
* 它支持UNIQUE约束
* 它它包含一个用于执行CQL命令的UI：Neo4j数据浏览器
* 它支持完整的ACID（原子性，一致性，隔离性和持久性）规则
* 它采用原生图形库与本地GPE（图形处理引擎）
* 它支持查询的数据导出到JSON和XLS格式
* 它提供了REST API，可以被任何编程语言（如Java，Spring，Scala等）访问
* 它提供了可以通过任何UI MVC框架（如Node JS）访问的Java脚本
* 它支持两种Java API：Cypher API和Native Java API来开发Java应用程序

### OrientDB

1. OrientDB是兼具文档数据库的灵活性和图形数据库管理链接能力的可深层次扩展的文档-图形数据库管理系统。可选无模式、全模式或混合模式下。支持许多高级特性，诸如ACID事务、快速索引，原生和SQL查询功能。可以JSON格式导入、导出文档。若不执行昂贵的JOIN操作的话，如同关系数据库可在几毫秒内可检索数以百计的链接文档图。
2. https://www.w3cschool.cn/orientdb/orientdb-mx931xi1.html

### Keras

1. Keras是一个由[Python](https://baike.baidu.com/item/Python/407313)编写的开源[人工神经网络](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/382460)库，可以作为[Tensorflow](https://baike.baidu.com/item/Tensorflow/18828108)、Microsoft-CNTK和Theano的高阶[应用程序接口](https://baike.baidu.com/item/%E5%BA%94%E7%94%A8%E7%A8%8B%E5%BA%8F%E6%8E%A5%E5%8F%A3/10418844)，进行[深度学习](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0/3729729)模型的设计、调试、评估、应用和可视化；
2. Keras在代码结构上完全由[面向对象](https://baike.baidu.com/item/%E9%9D%A2%E5%90%91%E5%AF%B9%E8%B1%A1/2262089)方法编写，高度[模块化](https://baike.baidu.com/item/%E6%A8%A1%E5%9D%97%E5%8C%96/3295536)并具有[可扩展性](https://baike.baidu.com/item/%E5%8F%AF%E6%89%A9%E5%B1%95%E6%80%A7/8669999)，其调用机制和说明文档有将[用户体验](https://baike.baidu.com/item/%E7%94%A8%E6%88%B7%E4%BD%93%E9%AA%8C/1994)和使用难度纳入考虑，简化了很多复杂算法的实现难度[1]  。Keras支持现代[人工智能](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E6%99%BA%E8%83%BD/9180)领域的主流算法，包括前馈结构和递归结构的深度神经网络，也可用和参与构建统计学习模型。在硬件和开发环境方面，Keras支持多[操作系统](https://baike.baidu.com/item/%E6%93%8D%E4%BD%9C%E7%B3%BB%E7%BB%9F/192)下的多GPU[并行计算](https://baike.baidu.com/item/%E5%B9%B6%E8%A1%8C%E8%AE%A1%E7%AE%97/113443)，可以根据后台设置转化为Tensorflow、Microsoft-CNTK等系统下的组件

### Tensorflow

1. TensorFlow是一个深度学习库，他是基于计算图的深度学习库，可以用它多快好省的完成目前绝大部分深度学习任务；
2. 复杂的说：它具有更广泛的应用领域，良好的语言支持，部署性能等优点，同时他是如今最广泛使用的深度学习框架，使用的公司包括Google、Uber等。
3. TensorFlow™是一个基于[数据流编程](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%B5%81%E7%BC%96%E7%A8%8B/22735640)（dataflow programming）的符号数学系统，被广泛应用于各类[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599)（machine learning）算法的编程实现，其前身是[谷歌](https://baike.baidu.com/item/%E8%B0%B7%E6%AD%8C/117920)的神经网络算法库DistBelief
4. 安装 tensorflow

Pip install tensorflow #安装tensorflow cpu稳定版本

pip install tensorflow-gpu#安装ensorflow gpu稳定版本

**import** tensorflow **as** tf  
*#第一个注释*hello=tf.constant(**"Hello TensorFlow"**)  
*#创建一个常量*sess=tf.Session();  
*#执行*print(sess.run(hello));

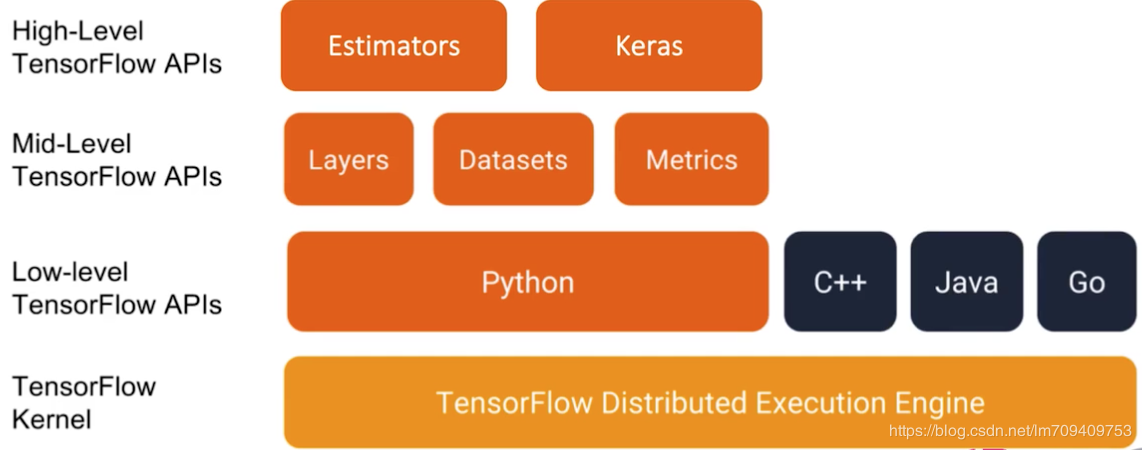
1. 安装 jupyter

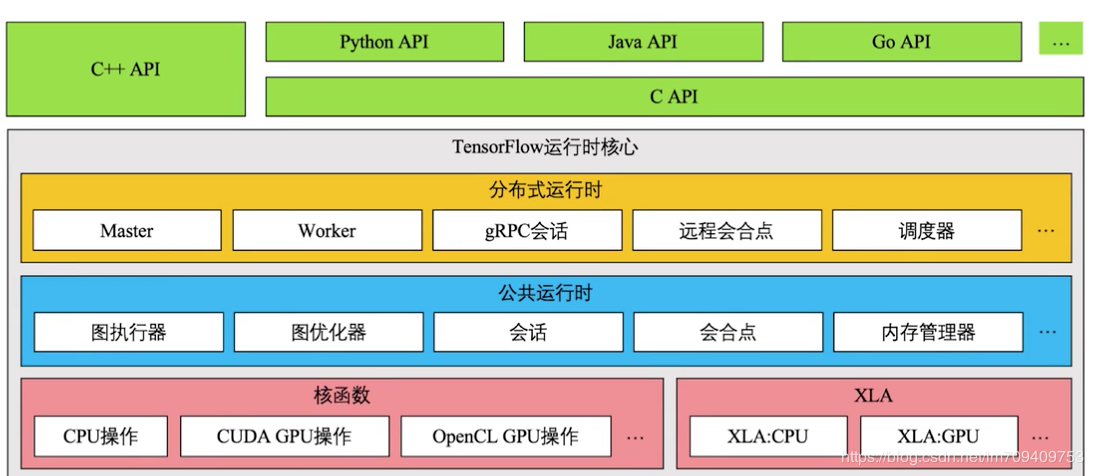
pip install jupyter

1. 使用jupyter

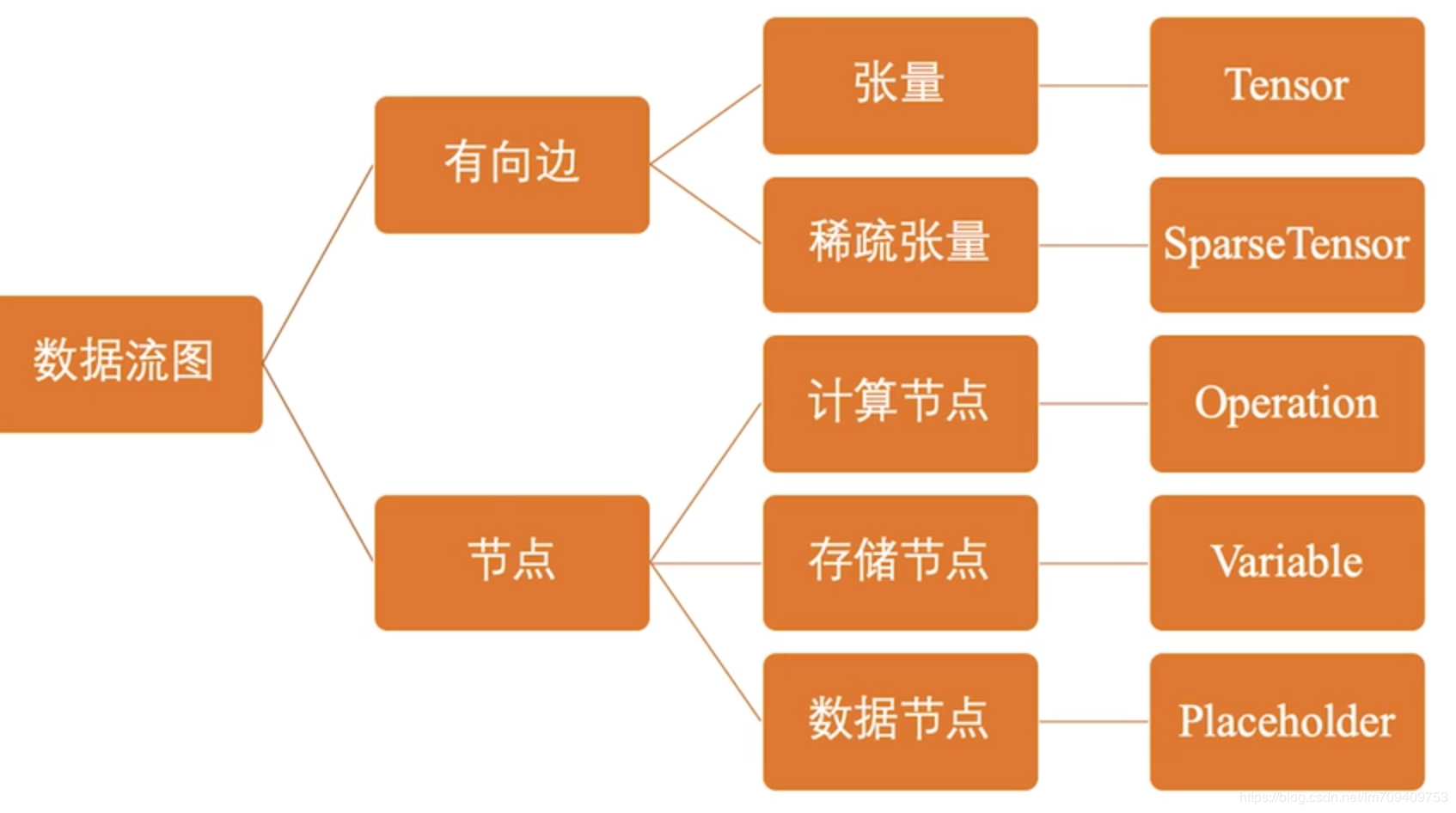
jupyter notebook

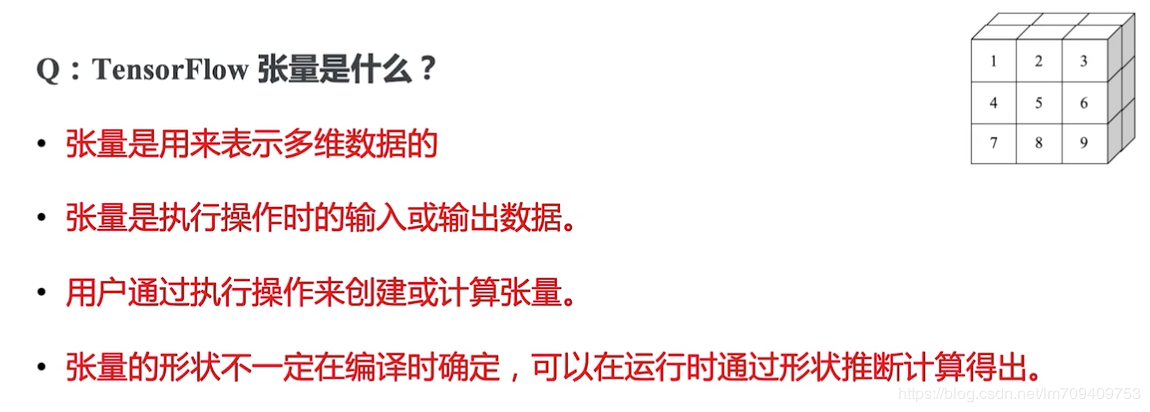
1. TensorFlow模块与架构介绍

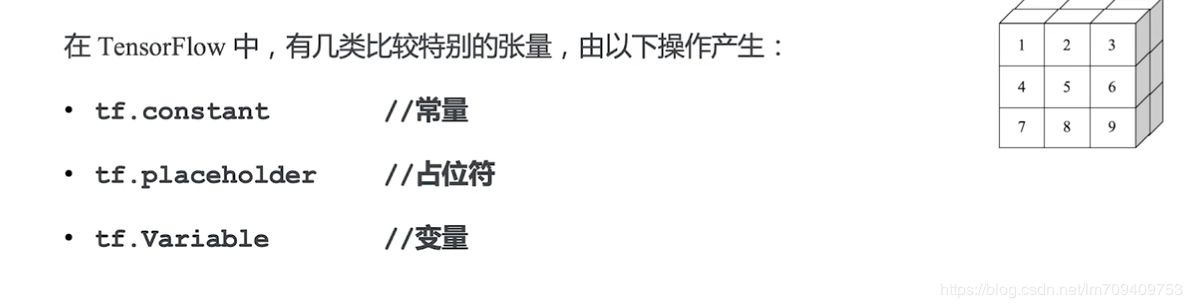














### Seq2Seq

1. Seq2Seq模型是RNN最重要的一个变种：N vs M（输入与输出序列长度不同）。
2. 这种结构又叫Encoder-Decoder模型。
3. 原始的N vs N RNN要求序列等长，然而我们遇到的大部分问题序列都是不等长的，如机器翻译中，源语言和目标语言的句子往往并没有相同的长度。
4. 为此，Encoder-Decoder结构先将输入数据编码成一个上下文向量c：
5. 得到c有多种方式，最简单的方法就是把Encoder的最后一个隐状态赋值给c，还可以对最后的隐状态做一个变换得到c，也可以对所有的隐状态做变换。
6. 拿到c之后，就用另一个RNN网络对其进行解码，这部分RNN网络被称为Decoder。具体做法就是将c当做之前的初始状态h0输入到Decoder中：
7. 还有一种做法是将c当做每一步的输入：
8. 由于这种Encoder-Decoder结构不限制输入和输出的序列长度，因此应用的范围非常广泛，比如：
9. 机器翻译。Encoder-Decoder的最经典应用，事实上这一结构就是在机器翻译领域最先提出的
10. 文本摘要。输入是一段文本序列，输出是这段文本序列的摘要序列。
11. 阅读理解。将输入的文章和问题分别编码，再对其进行解码得到问题的答案。
12. 语音识别。输入是语音信号序列，输出是文字序列。

### FastText

1. **fasttext是facebook开源的一个词向量与文本分类工具，在2016年开源，典型应用场景是“带监督的文本分类问题”**。提供简单而高效的文本分类和表征学习的方法，性能比肩深度学习而且速度更快。
2. fastText结合了自然语言处理和机器学习中最成功的理念。这些包括了使用**词袋以及n-gram袋**表征语句，还有使用子词(subword)信息，并通过隐藏表征在类别间共享信息。我们另外**采用了一个softmax层级**(利用了类别不均衡分布的优势)来加速运算过程。
3. FastText原理：fastText方法包含三部分，**模型架构，层次SoftMax和N-gram子词特征；**
4. fastText的架构和word2vec中的CBOW的架构类似，因为它们的作者都是Facebook的科学家Tomas Mikolov，而且确实fastText也算是word2vec所衍生出来的；
5. **fastText和word2vec的区别**

* 相似处：

1. 图模型结构很像，都是采用embedding向量的形式，得到word的隐向量表达。
2. 都采用很多相似的优化方法，比如使用Hierarchical softmax优化训练和预测中的打分速度。

* 不同处：

1. 模型的输出层：word2vec的输出层，对应的是每一个term，计算某term的概率最大；而fasttext的输出层对应的是分类的label。不过不管输出层对应的是什么内容，起对应的vector都不会被保留和使用。
2. 模型的输入层：word2vec的输出层，是 context window 内的term；而fasttext 对应的整个sentence的内容，包括term，也包括 n-gram的内容。

* 两者本质的不同，体现在 h-softmax的使用：

1. Word2vec的目的是得到词向量，该词向量 最终是在输入层得到，输出层对应的 h-softmax  
   也会生成一系列的向量，但最终都被抛弃，不会使用。
2. fastText则充分利用了h-softmax的分类功能，遍历分类树的所有叶节点，找到概率最大的label（一个或者N个）

### word2vector

1. **word2vector，顾名思义，就是将语料库中的词转化成向量，以便后续在词向量的基础上进行各种计算。最常见的表示方法是counting 编码**

### LDA

1. **LDA（Latent Dirichlet Allocation）是一种文档主题生成模型，也称为一个三层**[**贝叶斯**](https://baike.baidu.com/item/%E8%B4%9D%E5%8F%B6%E6%96%AF)**概率模型，包含词、主题和文档三层结构。所谓生成模型，就是说，我们认为一篇文章的每个词都是通过“以一定概率选择了某个主题，并从这个主题中以一定概率选择某个词语”这样一个过程得到。文档到主题服从多项式分布，主题到词服从多项式分布。**
2. LDA是一种非监督机器学习技术，可以用来识别大规模文档集（document collection）或语料库（corpus）中潜藏的主题信息。它采用了词袋（bag of words）的方法，这种方法将每一篇文档视为一个词频向量，从而将文本信息转化为了易于建模的数字信息。但是词袋方法没有考虑词与词之间的顺序，这简化了问题的复杂性，同时也为模型的改进提供了契机。每一篇文档代表了一些主题所构成的一个概率分布，而每一个主题又代表了很多单词所构成的一个概率分布。

### PyTorch

1. PyTorch 是一个基于 Python 的科学计算包

### Caffe（卷积神经网络框架）

1. Caffe是一个[深度学习框架](https://baike.baidu.com/item/%E6%B7%B1%E5%BA%A6%E5%AD%A6%E4%B9%A0%E6%A1%86%E6%9E%B6/22718084)，最初开发于加利福尼亚大学伯克利分校；
2. Caffe，全称Convolutional Architecture for Fast Feature Embedding，是一个兼具表达性、速度和思维模块化的深度学习框架。由[伯克利](https://baike.baidu.com/item/%E4%BC%AF%E5%85%8B%E5%88%A9/25375)人工智能研究小组和伯克利视觉和学习中心开发。虽然其内核是用C++编写的，但Caffe有Python和Matlab 相关接口。Caffe支持多种类型的深度学习架构，面向图像分类和图像分割，还支持CNN、RCNN、LSTM和全连接神经网络设计。Caffe支持基于GPU和CPU的加速计算内核库，如NVIDIA cuDNN和Intel MKL。
3. Caffe 完全开源，并且在有多个活跃社区沟通解答问题，同时提供了一个用于训练、测试等完整工具包，可以帮助使用者快速上手。此外 Caffe 还具有以下特点：
4. 模块性：Caffe 以[模块化](https://baike.baidu.com/item/%E6%A8%A1%E5%9D%97%E5%8C%96/3295536)原则设计，实现了对新的数据格式，网络层和损失函数轻松扩展。
5. 表示和实现分离：Caffe 已经用谷歌的 Protocl Buffer定义模型文件。使用特殊的文本文件 prototxt 表示网络结构，以有向非循环图形式的网络构建。
6. Python和MATLAB结合： Caffe 提供了 Python 和 MATLAB 接口，供使用者选择熟悉的语言调用部署算法应用。
7. GPU 加速：利用了 MKL、Open BLAS、cu BLAS 等计算库，利用GPU实现计算加速。

### Theano

Theano不太好，主要原因是：

1. theano 定义 function 时缺乏灵活的多态机制。

2. 困难的调试方法

### 实体抽取

1. 命名实体识别的主要任务：

要识别出文本中出现的专有名称和有意义的数量短语并加以归类

1. 命名实体识别的主要研究内容：

就整个的命名实体识别的研究结果而言，时间表达式和数字表达式的识别相对简单，其规则的设计、数据的统计训练等也比较容易。而对于实体中的组织名、人名、地名，因为其具有开放性和发展性的特点，而且构成规律有很大的随意性，所以其识别就可能会有较多的错选或漏选。现在大多数的命名实体识别的研究都集中于对这三种实体的识别技术的研究。

1. 3.命名实体识别的发展历程：
2. 基于规则的方法->基于统计的方法->混合方法
3. 4.汉语命名实体识别中的特殊难点：
4. （1）分词：边界模糊不仅存在于非实体词之间，也出现于实体词和非实体词之间。
5. （2）汉语命名实体的生成规律以及结构更加复杂，尤其是缩略语的表示形式具有多样性，很难提取构成规则，因此不可能用一种识别模型应用于所有的命名实体。
6. （3）与西方语言比较，汉语缺少在命名实体识别中起重要作用的词形变换特征。
7. （4）汉语中除比较特殊的字词外，命名实体也可包含普通字词。
8. （5）能用于汉语命名实体识别的开放型语料还很少，因此一方面需要开发大型命名实体标注语料库，另一方面研究不依赖大型命名实体标注文本库的算法也具有重要意义。
9. 5.命名实体识别的结果：
10. （1）正确（correct） ：系统识别结果和标准结果相同。
11. （2）丢失（missing）：系统未识别而标准结果中有。
12. （3）虚假（spurious）：系统识别但标准结果中没有。
13. 6.衡量命名实体识别系统性能主要的两个评价指标：
14. 查全率：正确/（正确+丢失）
15. 查准率：正确/（正确+虚假）
16. 有时为了综合评价系统的性能，通常还计算查全率和查准率的加权几何平均值即F指数。
17. 7.命名实体识别方法：
18. （1）基于规则：
19. 如：NTU系统、FACILE系统、OKI系统。
20. 缺点：缺乏鲁棒性和可移植性，对于每个新领域的文本都需要更新规则来保持最优性能，而这需要大量的专门知识和人力，代价往往非常大。
21. （2）基于统计：
22. 如：n元模型、隐马尔科夫模型（HMM）、最大熵模型（ME）、决策树、基于转换的学习方法、推进方法、表决感知器方法、条件马尔科夫模型等。评价性能最好的是HMM。而ME因其自身的特点仍是当前主要的研究方向。
23. 缺点：性能较基于规则的方法而言偏低，因为基于统计的方法获取的概率知识总赶不上人类专家的专业知识的可靠性，而且有些知识获取必需专家的经验。
24. （3）混合方法：
25. 借助规则知识及早剪枝，再用统计模型是比较好的方法。

### 关系抽取

1. 关系抽取: 从一个句子中判断两个entity是否有关系，一般是一个二分类问题，指定某种关系
2. 关系分类: 一般是判断一个句子中 两个entity是哪种关系，属于多分类问题。

优点:

* 使用Pooling 操作，提取了每一个卷积核对应的最优特征
* 引入Position Feature: 相当于在一定程度上引入了位置信息，因为CNN更多是局部N-Gram特征; PF突出了entity的影响，包括后续的PieceWise Max Pooling，同样也是对entity与其他词分开考虑。

缺点:

* 虽然使用了多个卷积核，但是仅仅使用同一个window-size, 提取特征比较单一。
* 结构还是相对简单，更多的是类似于CNN 文本分类的结构。 而且还是加入了人为构造的Lexical Feature。

### 实体对齐

1. 知识库的实体对齐(entity alignment)工作是近年来的研究热点问题.知识库实体对齐的目标是能够高质量链接多个现有知识库，并从顶层创建一个大规模的统一的知识库，从而帮助机器理解底层数据.然而，知识库实体对齐在数据质量、匹配效率等多个方面存在很多问题与挑战有待解决.从这些挑战出发，对十几年来的可用于知识库实体对齐的技术和算法进行综述，通过分类和总结现有技术，为进一步的研究工作提供可选方案.首先形式化定义了知识库实体对齐问题；然后对知识库的实体对齐工作进行总体概述，并从对齐算法、特征匹配技术和分区索引技术3个方面详细总结了各种可用方法和研究进展，重点从局部和全局2个角度对主流的集体对齐算法进行详细阐述，并介绍了常用的评测数据集；最后对未来重点的研究内容和发展方向进行了探讨和展望.

### 实体消岐

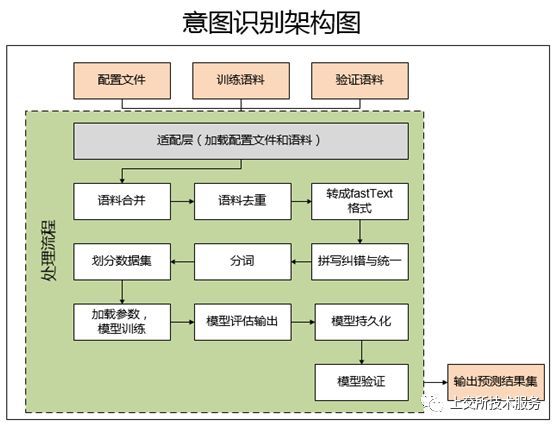
1. 实体消歧的本质在于一个单词很可能有多个意思，也就是在不同的上下文中所表达的含义可能不太一样

### 知识挖掘

1. [知识挖掘](https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E6%8C%96%E6%8E%98/2272520)源于全球范围内数据库中存储的数据量急剧增加，人们的需求已经不只是简单的查询和维护，而是希望能够对这些数据进行较高层次的处理和分析以得到关于数据总体特征和对发展趋势的预测。知识挖掘最新的描述性定义是由UsamaM.Fayyyad等给出的：[知识挖掘](https://baike.baidu.com/item/%E7%9F%A5%E8%AF%86%E6%8C%96%E6%8E%98/2272520)是从数据集中识别出有效的、新颖的、潜在有用的，以及最终可理解的模式的非平凡过程
2. 知识挖掘已被越来越多的领域所采用，并取得了较好效果。这些领域有科学研究、[市场营销](https://baike.baidu.com/item/%E5%B8%82%E5%9C%BA%E8%90%A5%E9%94%80)、金融投资、欺诈甄别、产品制造、通信网络管理等。由加州理工学院喷气推进实验室与天文科学家合作开发的SKICAT(SkyImageCatalogingandAnalysisTool)是第一个获得相当成功的知识挖掘应用，已经帮助科学家发现了16颗极其遥远的类星体

### 意图识别

1. **意图识别：**意图识别是通过分类的办法将句子或者我们常说的query分到相应的意图种类,这在搜索引擎以及智能问答中都起很重要的作用。简单来说，就是当用户输入一句话或者一段文本时，意图识别可以准确识别出它是哪个领域的问题，然后分配给相应的领域机器人进行二次处理，这在问题分类很多的情况，可以明显提升问题匹配的准确度。
2. 自动文摘--通过NLP技术对金融领域的新闻、研究报告、上市公司的公告进行分析，生成自动文摘和机器自动写作;
3. 事件分析--基于热点事件，然后结合行业产业链、知识图谱和推理决策系统的推理体系，针对个股和行业，生成研究报告和及时报告;
4. 热点挖掘--基于最新的新闻资讯，聚类出当前的热点资讯，提取摘要;
5. 智能客服--通过NLP的意图识别，问答相似度分析，语义理解等，可以用机器代替人工客服.



### 情感分析

1. 情感分析--对新闻、评论等文本中的主观性信息进行分析，来挖掘其态度和情绪，实现情感倾向性分析和观点挖掘;

### 语义理解

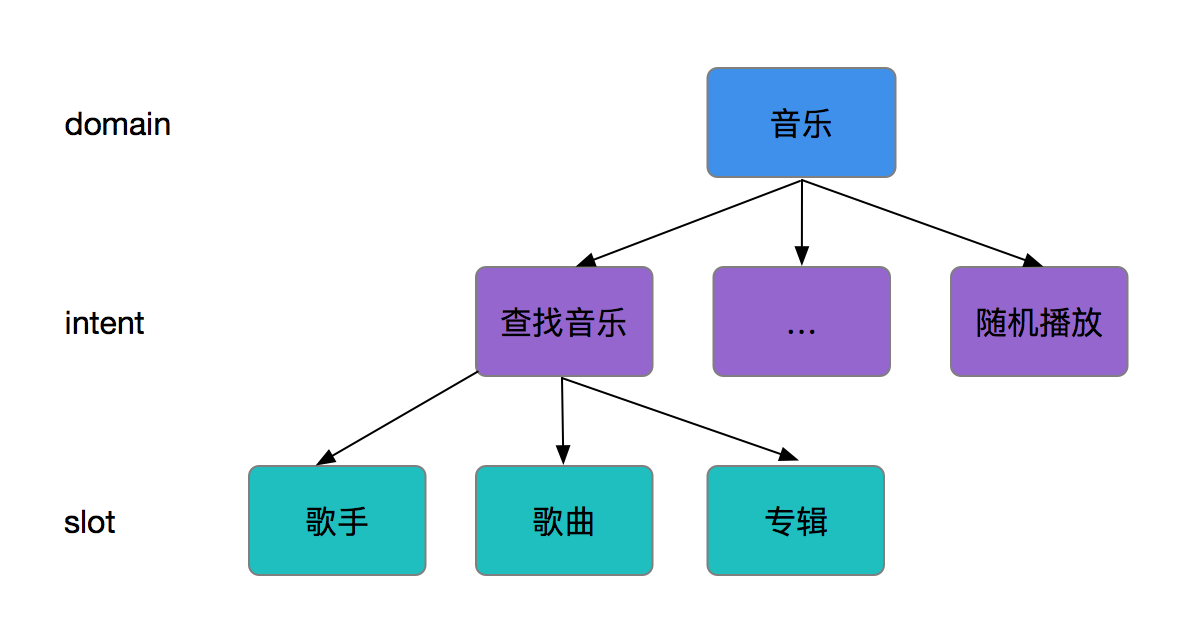
对话的第一步就是理解用户所说的话，也就是自然语言。自然语言理解（Natural Language Understanding，NLU）以语言学为基础，融合逻辑学、心理学和计算机科学等学科，试图解决以下问题：语言究竟是怎样组织起来传输信息的？人又是怎样从一连串的语言符号中获取信息的？换种表达就是，通过语法、语义、语用的分析，获取自然语言的语义表示。

1. 语义表示

自然语言语义的表示主要有三种：分布语义，框架语义，模型论语义。智能对话平台采用模型语义的一个变形：领域（domain）、意图（intent）、词槽（slot）来表示语义。

* 领域（domain）：领域是指同一类型的数据或者资源，以及围绕这些数据或资源提供的服务，比如天气、音乐等。
* 意图（intent）：意图是指对于领域数据的操作，一般以动宾短语来命名，比如询问天气、查找音乐。
* 词槽（slot）：词槽用来存放领域的属性，比如天气领域的日期、天气，音乐领域的歌手、歌曲名等。

对于音乐领域，我们的语义结构可以表示为



### 倾向性分析

### 词性标注

1. 词性标注（Part-Of-Speech tagging, POS tagging）也被称为语法标注（grammatical tagging）或词类消疑（word-category disambiguation），是[语料库语言学](https://baike.baidu.com/item/%E8%AF%AD%E6%96%99%E5%BA%93%E8%AF%AD%E8%A8%80%E5%AD%A6/1046088)（corpus linguistics）中将[语料库](https://baike.baidu.com/item/%E8%AF%AD%E6%96%99%E5%BA%93/11029908)内单词的词性按其含义和上下文内容进行标记的文本数据处理技术；
2. 词性标注可以由人工或特定算法完成，使用[机器学习](https://baike.baidu.com/item/%E6%9C%BA%E5%99%A8%E5%AD%A6%E4%B9%A0/217599)（machine learning）方法实现词性标注是[自然语言处理](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/365730)（Natural Language Processing, NLP）的研究内容。常见的词性标注算法包括[隐马尔可夫模型](https://baike.baidu.com/item/%E9%9A%90%E9%A9%AC%E5%B0%94%E5%8F%AF%E5%A4%AB%E6%A8%A1%E5%9E%8B/7932524)（Hidden Markov Model, HMM）、[条件随机场](https://baike.baidu.com/item/%E6%9D%A1%E4%BB%B6%E9%9A%8F%E6%9C%BA%E5%9C%BA/10804560)（Conditional random fields, CRFs）等。
3. 词性标注主要被应用于[文本挖掘](https://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E6%8C%96%E6%8E%98/4401689)（text mining）和NLP领域，是各类基于文本的机器学习任务，例如[语义分析](https://baike.baidu.com/item/%E8%AF%AD%E4%B9%89%E5%88%86%E6%9E%90/8853372)（semantic analysis）和[指代消解](https://baike.baidu.com/item/%E6%8C%87%E4%BB%A3%E6%B6%88%E8%A7%A3/7605552)（coreference resolution）的预处理步骤。
4. 词性标注在本质上是分类问题，将语料库中的单词按词性分类。一个词的词性由其在所属语言的含义、形态和语法功能决定。以汉语为例，汉语的词类系统有18个子类，包括7类体词，4类谓词、5类虚词、代词和感叹词。词类不是闭合集，而是有兼词现象，例如“制服”在作为“服装”和作为“动作”时会被归入不同的词类，因此词性标注与上下文有关 [3]  。对词类的理论研究可以得到基于人工规则的词性标注方法，这类方法对句子的形态进行分析并按预先给定的规则赋予词类。
5. 词性标注是文本数据的预处理环节之一，原始文本在NLP或文本挖掘应用中，首先通过字符分割（word segmentation）和字符嵌入（word embedding）被向量化，随后通过词性标注得到高阶层特征，并输入[语法分析器](https://baike.baidu.com/item/%E8%AF%AD%E6%B3%95%E5%88%86%E6%9E%90%E5%99%A8/10598664)执行[语义分析](https://baike.baidu.com/item/%E8%AF%AD%E4%B9%89%E5%88%86%E6%9E%90/8853372)（sentiment analysis）、[指代消解](https://baike.baidu.com/item/%E6%8C%87%E4%BB%A3%E6%B6%88%E8%A7%A3/7605552)（coreference resolution）等任务。

### 文本分类聚类

1. 文本聚类（Text clustering）文档[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)主要是依据著名的[聚类假设](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%81%87%E8%AE%BE/10751611)：同类的文档相似度较大，而不同类的[文档](https://baike.baidu.com/item/%E6%96%87%E6%A1%A3/1009768)相似度较小。作为一种无监督的机器学习方法，聚类由于不需要训练过程，以及不需要预先对文档手工标注类别，因此具有一定的灵活性和较高的自动化处理能力，已经成为对文本信息进行有效地组织、摘要和导航的重要手段，为越来越多的研究人员所关注。
2. ①文档[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)可以作为多文档[自动文摘](https://baike.baidu.com/item/%E8%87%AA%E5%8A%A8%E6%96%87%E6%91%98)等[自然语言处理](https://baike.baidu.com/item/%E8%87%AA%E7%84%B6%E8%AF%AD%E8%A8%80%E5%A4%84%E7%90%86/365730)应用的预处理步骤，比较典型的例子是[哥伦比亚大学](https://baike.baidu.com/item/%E5%93%A5%E4%BC%A6%E6%AF%94%E4%BA%9A%E5%A4%A7%E5%AD%A6/513683)开发的多文档文摘系统Newsblaster。Newsblaster将每天发生的重要新闻文本进行聚类处理，并对同主题文档进行冗余消除、[信息融合](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E8%9E%8D%E5%90%88/2272960)、文本生成等处理，从而生成一篇简明扼要的摘要文档；
3. ②对搜索引擎返回的结果进行聚类，使用户迅速定位到所需要的信息。Hua-Jun Zeng等人提出了对搜索引擎返回的结果进行聚类的学习算法。比较典型的系统则有[vivisimo](https://baike.baidu.com/item/vivisimo/9546714)和infonetware等。系统允许用户输入检索关键词，而后对检索到的文档进行聚类处理，并输出各个不同类别的简要描述，从而可以缩小检索的范围，用户只需关注比较有希望的主题。另外这种方法也可以为用户二次检索提供线索；
4. ③对用户感兴趣的文档（如用户浏览器cache中的网页）[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)，从而发现用户的兴趣模式并用于[信息过滤](https://baike.baidu.com/item/%E4%BF%A1%E6%81%AF%E8%BF%87%E6%BB%A4)和信息主动推荐等服务。
5. ④聚类技术还可以用来改善[文本分类](https://baike.baidu.com/item/%E6%96%87%E6%9C%AC%E5%88%86%E7%B1%BB)的结果，如[俄亥俄州立大学](https://baike.baidu.com/item/%E4%BF%84%E4%BA%A5%E4%BF%84%E5%B7%9E%E7%AB%8B%E5%A4%A7%E5%AD%A6/6404922)的Y.C. Fang, S. Parthasarathy和F. Schwartz等人的工作。
6. ⑤[数字图书馆](https://baike.baidu.com/item/%E6%95%B0%E5%AD%97%E5%9B%BE%E4%B9%A6%E9%A6%86/221396)服务。通过SOM神经网络等方法，可以将[高维空间](https://baike.baidu.com/item/%E9%AB%98%E7%BB%B4%E7%A9%BA%E9%97%B4/692559)的文档[拓扑](https://baike.baidu.com/item/%E6%8B%93%E6%89%91/573536)保序地映射到二维空间，使得聚类结果可视化和便于理解，如SOMlib[ ]系统；
7. ⑥[文档集合](https://baike.baidu.com/item/%E6%96%87%E6%A1%A3%E9%9B%86%E5%90%88)的自动整理。如Scatter/Gather[ ]是一个基于[聚类](https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB/593695)的文档浏览系统。而微软的Ji-Rong Wen等人则利用聚类技术对用户提出的查询记录进行聚类，并利用结果更新搜索引擎网站的[FAQ](https://baike.baidu.com/item/FAQ/99550)。

### 热词发现

1. 热词，作为一种词汇现象，反映了一个国家、一个地区在一个时期人们普遍关注的问题和事物。具有时代特征，反映一个时期的热点话题及民生问题。其主要表达形式有语言、文字以及网络图片。
2. 网络热词有两层含义：一是作为一种语言首先是流行于虚拟网络空间并且为网民自己创造或借用的；二是能够记载历史、传递民情。既然我们把一个语言单位称作“网络热词”，那么它应该同时具有“网络”和“热”两个基本要素。“网络”要素，说的是它起源于网络，由网民自己所创；“热”要素，说的是它被网民所熟知并广泛使用，产生巨大社会反响。
3. 热词创作具有和新闻创作一样的精髓和理念：紧扣社会热点，吸引受众眼球，博得舆论关注，常具有“语不惊人死不休”的特点，因此也创造出众多令人感到新奇的词语。每一个热词的产生和热传，往往跟随社会重大新闻事件，具有很强的时效性。
4. 热词不仅仅是一种语言文化现象：热词不仅仅是一种语言文化现，而且已经成为都市人宣泄情绪、自我解压的一种方式方法。人们产生精神危机、对社会缺乏信心时，某种情感就会成为社会的共同心理，当带有这种情感色彩的新闻事件发生时，聪明的人们就会针对一些敏感事件发表异议，用“隐晦”的方式创造出“令人叫绝”的网络热词，借此来调侃稀释心中的无奈，从而迅速引发共鸣。
5. 热词折射社会热点的民意特征：语言符号的意义是人们赋予的，人们通过语言符号来表达自己的看法。在互联移动技术发明之前，面对重大新闻事件，民众只能通过人际传播来交流和表达自己的观点和看法[1]  。在自媒体这个异常开放的空间和平台上，带有公民评判意识的各种网络热词此起彼伏，源源不断。例如在贵州瓮安事件中，有官员说出“做俯卧撑”之后，原本受网友们宠爱的“打酱油”一词立即被“俯卧撑”取代了，用来表示对事情漠不关心、不发表自己的意见，只干与自己有关的事情。2009年2月，云南24岁的李荞明在被看守所关押11天后死亡，警方对于这件事情的说法是他和狱友们玩“躲猫猫”撞墙所致，“躲猫猫”一词迅速蹿红网络。

### 关键词抽取

1. TextRank算法提取关键词的Java实现；
2. **在自然语言处理领域，处理海量的文本文件最关键的是要把用户最关心的问题提取出来。**而无论是对于长文本还是短文本，往往可以通过几个关键词窥探整个文本的主题思想。与此同时，不管是基于文本的推荐还是基于文本的搜索，对于文本关键词的依赖也很大，关键词提取的准确程度直接关系到推荐系统或者搜索系统的最终效果。因此，关键词提取在文本挖掘领域是一个很重要的部分。
3. 关于文本的关键词提取方法分为有监督、半监督和无监督三种；
4. 有监督的关键词抽取算法：它是建关键词抽取算法看作是二分类问题，判断文档中的词或者短语是或者不是关键词。既然是分类问题，就需要提供已经标注好的训练预料，利用训练语料训练关键词提取模型，根据模型对需要抽取关键词的文档进行关键词抽取；
5. 半监督的关键词提取算法：只需要少量的训练数据，利用这些训练数据构建关键词抽取模型，然后使用模型对新的文本进行关键词提取，对于这些关键词进行人工过滤，将过滤得到的关键词加入训练集，重新训练模型。
6. 无监督的方法：不需要人工标注的语料，利用某些方法发现文本中比较重要的词作为关键词，进行关键词抽取。

**有监督的文本关键词提取算法需要高昂的人工成本，因此现有的文本关键词提取主要采用适用性较强的无监督关键词抽取。**其文本关键词抽取流程如下：

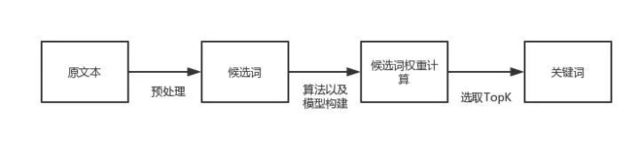


图 1 无监督文本关键词抽取流程图

无监督关键词抽取算法可以分为三大类，基于统计特征的关键词抽取、基于词图模型的关键词抽取和基于主题模型的关键词抽取

基于统计特征的关键词提取算法

基于于统计特征的关键词抽取算法的思想是**利用文档中词语的统计信息抽取文档的关键词。**通常将文本经过预处理得到候选词语的集合，然后采用特征值量化的方式从候选集合中得到关键词。基于统计特征的关键词抽取方法的关键是采用什么样的特征值量化指标的方式，目前常用的有三类：

**1**

**基于词权重的特征量化**

基于词权重的特征量化主要包括词性、词频、逆向文档频率、相对词频、词长等。

**2**

**基于词的文档位置的特征量化**

这种特征量化方式是根据文章不同位置的句子对文档的重要性不同的假设来进行的。通常，文章的前N个词、后N个词、段首、段尾、标题、引言等位置的词具有代表性，这些词作为关键词可以表达整个的主题。

**3**

**基于词的关联信息的特征量化**

词的关联信息是指词与词、词与文档的关联程度信息，包括互信息、hits值、贡献度、依存度、TF-IDF值等。

下面介绍几种常用的特征值量化指标。

词性

词性时通过分词、语法分析后得到的结果。现有的关键词中，绝大多数关键词为名词或者动名词。一般情况下，名词与其他词性相比更能表达一篇文章的主要思想。但是，词性作为特征量化的指标，一般与其他指标结合使用。

词频

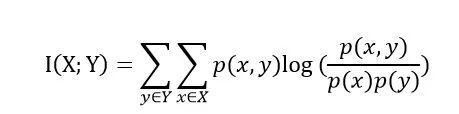
词频表示一个词在文本中出现的频率。一般我们认为，如果一个词在文本中出现的越是频繁，那么这个词就越有可能作为文章的核心词。词频简单地统计了词在文本中出现的次数，但是，只依靠词频所得到的关键词有很大的不确定性，对于长度比较长的文本，这个方法会有很大的噪音。

位置信息

一般情况下，词出现的位置对于词来说有着很大的价值。例如，标题、摘要本身就是作者概括出的文章的中心思想，因此出现在这些地方的词具有一定的代表性，更可能成为关键词。但是，因为每个作者的习惯不同，写作方式不同，关键句子的位置也会有所不同，所以这也是一种很宽泛的得到关键词的方法，一般情况下不会单独使用。

互信息

互信息是信息论中概念，是变量之间相互依赖的度量。互信息并不局限于实值随机变量，它更加一般且决定着联合分布 p(X,Y) 和分解的边缘分布的乘积 p(X)p(Y) 的相似程度。互信息的计算公式如下：

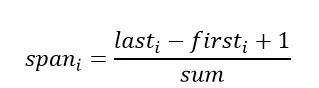


其中，p(x,y)是X和Y的联合概率分布函数，p(x)和p(y)分别为X和Y的边缘概率分布函数。

当使用互信息作为关键词提取的特征量化时，应用文本的正文和标题构造PAT树，然后计算字符串左右的互信息。

词跨度

词跨度是指一个词或者短语字文中首次出现和末次出现之间的距离，词跨度越大说明这个词对文本越重要，可以反映文本的主题。一个词的跨度计算公式如下：



其中，表示词i在文本中最后出现的位置， 表示词 i 在文本中第一次出现的位置，sum表示文本中词的总数。

词跨度被作为提取关键词的方法是因为在现实中，文本中总是有很多噪声（指不是关键词的那些词），使用词跨度可以减少这些噪声。

TF-IDF值

一个词的TF是指这个词在文档中出现的频率，假设一个词w在文本中出现了m次，而文本中词的总数为n，那么。一个词的IDF是根据语料库得出的，表示这个词在整个语料库中出现的频率。假设整个语料库中，包含词w的文本一共有M篇，语料库中的文本一共有N篇，则

由此可得词w的TF-IDF值为：

**TF-IDF的优点是实现简单，相对容易理解。**但是，TFIDF算法提取关键词的缺点也很明显，严重依赖语料库，需要选取质量较高且和所处理文本相符的语料库进行训练。另外，对于IDF来说，它本身是一种试图抑制噪声的加权，本身倾向于文本中频率小的词，这使得TF-IDF算法的精度不高。TF-IDF算法还有一个缺点就是不能反应词的位置信息，**在对关键词进行提取的时候，词的位置信息，例如文本的标题、文本的首句和尾句等含有较重要的信息，应该赋予较高的权重。**

基于统计特征的关键词提取算法通过上面的一些特征量化指标将关键词进行排序，获取TopK个词作为关键词。

基于统计特征的关键词的重点在于特征量化指标的计算，不同的量化指标得到的记过也不尽相同。同时，不同的量化指标作为也有其各自的优缺点，在实际应用中，通常是采用不同的量化指标相结合的方式得到Topk个词作为关键词。

NO.2

文本关键词提取算法

基于词图模型的关键词抽取算法

基于词图模型的关键词抽取首先要构建文档的语言网络图，然后对语言进行网络图分析，在这个图上寻找具有重要作用的词或者短语，这些短语就是文档的关键词。语言网络图中节点基本上都是词，根据词的链接方式不同，语言网络的主要形式分为四种：**共现网络图、语法网络图、语义网络图和其他网络图。**

在语言网络图的构建过程中，都是以预处理过后的词作为节点，词与词之间的关系作为边。语言网络图中，边与边之间的权重一般用词之间的关联度来表示。在使用语言网络图获得关键词的时候，需要评估各个节点的重要性，然后根据重要性将节点进行排序，选取TopK个节点所代表的词作为关键词。节点的重要性计算方法有以下几种方法。

**1**

**综合特征法**

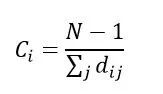
综合特征法也叫社会网络中心性分析方法，这种方法的核心思想是节点中重要性等于节点的显著性，以不破坏网络的整体性为基础。此方法就是从网络的局部属性和全局属性角度去定量分析网络结构的拓扑性质，常用的定量计算方法如下。

度

节点的度是指与该节点直接向量的节点数目，表示的是节点的局部影响力，对于非加权网络，节点的度为：对于加权网络，节点的度又称为节点的强度，计算公式为：

接近性

节点的接近性是指节点到其他节点的最短路径之和的倒数，表示的是信息传播的紧密程度，其计算公式为：



特征向量

特征向量的思想是节点的中心化测试值由周围所有连接的节点决定，即一个节点的中心化指标应该等于其相邻节点的中心化指标之线性叠加，表示的是通过与具有高度值的相邻节点所获得的间接影响力。特征向量的计算公式如下：

集聚系数

节点的集聚系数是它的相邻的节点之间的连接数与他们所有可能存在来链接的数量的比值，用来描述图的顶点之间阶级成团的程度的系数，计算公式如下：

平均最短路径

节点的平局最短路径也叫紧密中心性，是节点的所有最短路径之和的平均值，表示的是一个节点传播信息时对其他节点的依赖程度。如果一个节点离其他节点越近，那么他传播信息的时候也就越不需要依赖其他人。一个节点到网络中各点的距离都很短，那么这个点就不会受制于其他节点。计算公式如下：

因为每个算法的侧重方向的不同，在实际的问题中所选取的定量分析方法也会不一样。同时，对于关键词提取来说，也可以和上一节所提出的统计法得到的词的权重，例如词性等相结合构建词搭配网络，然后利用上述方法得到关键词。

**2**

**系统科学法**

系统科学法进行中心性分析的思想是节点重要性等于这个节点被删除后对于整个语言网络图的破坏程度。重要的节点被删除后会对网络的呃连通性等产生变化。如果我们在网络图中删除某一个节点，图的某些指定特性产生了改变，可以根据特性改变的大小获得节点的重要性，从而对节点进行筛选。

**3**

**随机游走法**

随机游走算法时网络图中一个非常著名的算法，它从给定图和出发点，随机地选择邻居节点移动到邻居节点上，然后再把现在的节点作为出发点，迭代上述过程。

**随机游走算法一个很出名的应用是大名鼎鼎的PageRank算法，PageRank算法是整个google搜索的核心算法，是一种通过网页之间的超链接来计算网页重要性的技术，其关键的思想是重要性传递。**在关键词提取领域， Mihalcea 等人所提出的TextRank算法就是在文本关键词提取领域借鉴了这种思想。

PageRank算法将整个互联网看作一张有向图，网页是图中的节点，而网页之间的链接就是图中的边。根据重要性传递的思想，如果一个大型网站A含有一个超链接指向了网页B，那么网页B的重要性排名会根据A的重要性来提升。网页重要性的传递思想如下图所示：

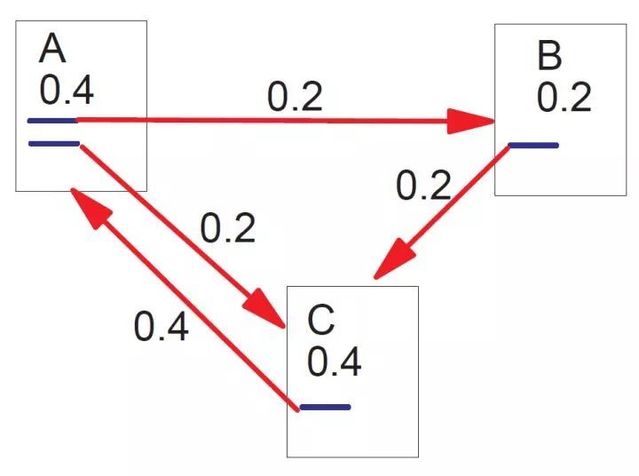
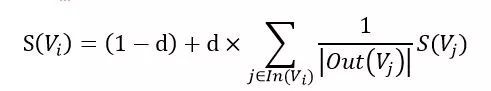


图 2 PageRank简单描述（来自PageRank论文）

在PageRank算法中，最主要的是对于初始网页重要性（PR值）的计算，因为对于上图中的网页A的重要性我们是无法预知的。但是，在原始论文中给出了一种迭代方法求出这个重要性，论文中指出，幂法求矩阵特征值与矩阵的初始值无关。那么，就可以为每个网页随机给一个初始值，然后迭代得到收敛值，并且收敛值与初始值无关。

PageRank求网页i的PR值计算如下：



其中，d为阻尼系数，通常为0.85。是指向网页 i 的网页集合。是指网页j中的链接指向的集合，是指集合中元素的个数。

TextRank在构建图的时候将节点由网页改成了句子，并为节点之间的边引入了权值，其中权值表示两个句子的相似程度。其计算公式如下：

公式中的为图中节点和的边的权重。其他符号与PageRank公式相同。

TextRank算法除了做文本关键词提取，还可以做文本摘要提取，效果不错。但是TextRank的计算复杂度很高，应用不广。

NO.3

文本关键词提取算法

基于主题模型的关键词抽取

**基于主题关键词提取算法主要利用的是主题模型中关于主题的分布的性质进行关键词提取。算法步骤如下：**

1

获取候选关键词

从文章中获取候选关键词。即将文本分词，也可以再根据词性选取候选关键词。

2

语料学习

根据大规模预料学习得到主题模型。

3

计算文章主题分部

根据得到的隐含主题模型，计算文章的主题分布和候选关键词分布。

4

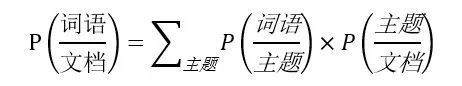
排序

计算文档和候选关键词的主题相似度并排序，选取前n个词作为关键词。

**算法的关键在于主题模型的构建。**主题模型是一种文档生成模型，对于一篇文章，我们的构思思路是先确定几个主题，然后根据主题想好描述主题的词汇，将词汇按照语法规则组成句子，段落，最后生成一篇文章。

主题模型也是基于这个思想，它认为文档是一些主题的混合分布，主题又是词语的概率分布，pLSA模型就是第一个根据这个想法构建的模型。同样地，我们反过来想，我们找到了文档的主题，然后主题中有代表性的词就能表示这篇文档的核心意思，就是文档的关键词。

pLSA模型认为，一篇文档中的每一个词都是通过一定概率选取某个主题，然后再按照一定的概率从主题中选取得到这个词语，这个词语的计算公式为：



一些贝叶斯学派的研究者对于pLSA模型进行了改进，他们认为，文章对应主题的概率以及主题对应词语的概率不是一定的，也服从一定的概率，于是就有了现阶段常用的主题模型--LDA主题模型。

LDA是D.M.Blei在2003年提出的。LDA采用了词袋模型的方法简化了问题的复杂性。在LDA模型中，每一篇文档是一些主题的构成的概率分布，而每一个主题又是很多单词构成的一个概率分布。同时，无论是主题构成的概率分布还是单词构成的概率分布也不是一定的，这些分布也服从Dirichlet 先验分布。

文档的生成模型可以用如下图模型表示：

其中和为先验分布的超参数，为第k个主题下的所有单词的分布，为文档的主题分布，w为文档的词，z为w所对应的主题。

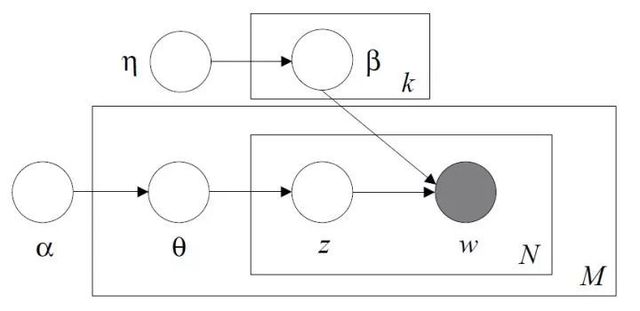


图 3 Blei在论文中的图模型

DA挖掘了文本的深层语义即文本的主题，用文本的主题来表示文本的也从一定程度上降低了文本向量的维度，很多人用这种方式对文本做分类，取得了不错的效果。**具体LDA的算法在请参考《一文详解LDA主题模型》。**

LDA关键词提取算法利用文档的隐含语义信息来提取关键词，但是主题模型提取的关键词比较宽泛，不能很好的反应文档主题。另外，对于LDA模型的时间复杂度较高，需要大量的实践训练。

NO.4

文本关键词提取算法

应用

现阶段，文本的关键词提取在基于文本的搜索、推荐以及数据挖掘领域有着很广泛的应用。同时在实际应用中，因为应用环境的复杂性，对于不同类型的文本，例如长文本和短文本，用同一种文本关键词提取方法得到的效果并相同。因此，在实际应用中针对不同的条件环境所采用的算法会有所不同，没有某一类算法在所有的环境下都有很好的效果。

相对于上文中所提到的算法，一些组合算法在工程上被大量应用以弥补单算法的不足，例如将TF-IDF算法与TextRank算法相结合，或者综合TF-IDF与词性得到关键词等。同时，工程上对于文本的预处理以及文本分词的准确性也有很大的依赖。对于文本的错别字，变形词等信息，需要在预处理阶段予以解决，分词算法的选择，未登录词以及歧义词的识别在一定程度上对于关键词突提取会又很大的影响。

**关键词提取是一个看似简单，在实际应用中却十分棘手的任务，从现有的算法的基础上进行工程优化，达观数据在这方面做了很大的努力并且取得了不错的效果。**

NO.5

文本关键词提取算法

总结

本文介绍了三种常用的无监督的关键词提取算法，并介绍了其优缺点。关键词提取在文本挖掘领域具有很广阔的应用，现有的方法也存在一定的问题，我们依然会在关键词提取的问题上继续努力研究，也欢迎大家积极交流。

### 文本纠错

1. 【**文本纠错**】可快速对文本进行错误识别与纠正，帮助提升文案准确性；
2. 【**对话情绪识别**】可识别客服聊天文本情绪，帮助优化企业客服的产品体验

* **文本纠错**

支持短文本、长文本、语音识别结果等多种文本内容，可以为您快速、准确找到文本中的错误并给出对应正确的文本以供参考。

文本纠错能力，广泛应用在搜索引擎、人机对话、语音识别、内容审核等方面，显著提高这些场景下的语义准确性和用户体验。

* **对话情绪识别**

你想知道用户在客服聊天里的打出的这段话蕴含什么样的情绪？对话情绪识别可以实现啦！

这个功能主要针对对话文本，自动检测用户日常对话文本中蕴含的情绪特征，输出情绪类别及置信度。

目前可检测的情绪包含强烈负向、非强烈负向两种情绪分类。这项功能能够帮助企业更全面的把握产品体验、监控客户服务质量。后续我们还会开放更多情绪类别。

### 语言模型

### 问答系统

### 知识推理

### 自动摘要

## 算法相关

### CNN（卷积神经网络）

### DNN（深度神经网络算法）

DNN是指深度神经网络算法，是近几年在工业界和学术界新型的一个机器学习领域的流行话题。DNN算法成功的将以往的识别率提高了一个显著的档次

### RNN（[循环神经网络](https://baike.baidu.com/item/%E5%BE%AA%E7%8E%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/23199490)）

1. [循环神经网络](https://baike.baidu.com/item/%E5%BE%AA%E7%8E%AF%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/23199490)，Recurrent Neural Network。神经网络是一种节点定向连接成环的[人工神经网络](https://baike.baidu.com/item/%E4%BA%BA%E5%B7%A5%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C)。这种网络的内部状态可以展示动态时序行为。不同于[前馈神经网络](https://baike.baidu.com/item/%E5%89%8D%E9%A6%88%E7%A5%9E%E7%BB%8F%E7%BD%91%E7%BB%9C/7580523)的是，RNN可以利用它内部的记忆来处理任意时序的输入序列，这让它可以更容易处理如不分段的[手写识别](https://baike.baidu.com/item/%E6%89%8B%E5%86%99%E8%AF%86%E5%88%AB/2724189)、语音识别等。

### LSTM

1. long short term memory，即我们所称呼的LSTM，**是为了解决长期以来问题而专门设计出来的**，所有的RNN都具有一种重复神经网络模块的链式形式。

### CTC

1. CTC，Connectionist Temporal Classification，用来解决输入序列和输出序列难以一一对应的问题。
2. 举例来说，在语音识别中，我们希望音频中的音素和翻译后的字符可以一一对应，这是训练时一个很天然的想法。但是要对齐是一件很困难的事，如下图所示（图源见参考资料[1]），有人说话块，有人说话慢，每个人说话快慢不同，不可能手动地对音素和字符对齐，这样太耗时。

## 应用相关

### 语音识别

### 图像识别

### 知识图谱

1. 知识图谱完整案例演示
2. 知识图谱典型应用
3. 知识图谱与人工智能
4. 知识图谱的系统架构
5. 知识图谱设计-知识设计方法
6. 知识图谱设计-语义类型设计
7. 语义关系设计
8. 本地对象设计
9. 知识图谱存储-知识存储模型
10. 知识图谱存储-知识存储系统选型

### 智能问答

### 聊天机器人

# 人工智能面试题

## Python基本面试题

### python的基本数据类型有哪些

python的数据类型有int整型，long长整型，float浮点数，complex复数，布尔值（0和1），字符串str，列表list，**元组**，**字典**，**集合**等十种。  
拓展：为了对比记忆，我们再来看看java的基本数据类型。

java的有8大基本数据类型，包括四种整数类型，byte(8位),short（16位）,int(32位),long（64位）,2种浮点数类型，float(32位),double（64位）,一种字符类型char（16位），一种布尔（boolean）类型。

## 基本面试题

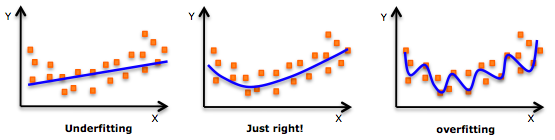
### 深度学习框架TensorFlow中有哪四种常用交叉熵？

答： tf.nn.weighted\_cross\_entropy\_with\_logits  
　　tf.nn.sigmoid\_cross\_entropy\_with\_logits  
　　tf.nn.softmax\_cross\_entropy\_with\_logits  
　　tf.nn.sparse\_softmax\_cross\_entropy\_with\_logits

### 什么叫过拟合，避免过拟合都有哪些措施？

答：过拟合：就是在机器学习中，我么测试模型的时候，提高了在训练数据集的表现力时候，  
　　但是在训练集上的表现力反而下降了。  
　　解决方案：  
　　1.正则化 ；  
　　2.在训练模型过程中，调节参数。学习率不要太大；  
　　3.对数据进行交叉验证；  
　　4.选择适合训练集合测试集数据的百分比，选取合适的停止训练标准，使对机器的训练在合适；  
　　5.在神经网络模型中，我们可以减小权重；

### 过拟合和欠拟合及解决方法：



如上图所示，左一和右一分别表示欠拟合和过拟合，中间表示的则是恰好拟合。过拟合的特点是模型的参数过多，提取的特征过多，模型的泛化能力差，即模型在训练集上表现很好，在测试集上则表示很差。欠拟合的特点则是模型的参数过少，提取的特征太少，以至于模型不能很好地完成分类或者预测的任务。

对于过拟合，可以通过使用正则项、pooling、dropout、增大训练样本规模等方法来解决。对于欠拟合，则可以通过增加模型的参数及复杂度来解决。

### 什么是核函数？

核函数是将线性不可分的特征隐射到高位特征空间，从而让支持向量机在这个高维空间线性可分，也就是使用核函数可以向高维空间映射并解决非线性的分类问题。包括线性核函数，多项式核函数，高斯核函数等，其中高斯核函数最为常用。

### 深度学习框架TensorFlow中常见的核函数都有哪些？

SVM高斯核函数，应为如果想要分割非线性数据集，改变线性分类器隐射到数据集，就要改变SVM损失函数中的核函数  
　　线性核函数  
　　多项式核函数

### 朴素贝叶斯方法的优势是什么？

朴素贝叶斯有稳定的分类效率，对于小规模的数据表现很好，能处理多分类问题，可以再数据超出内存时，去增量训练  
　　对缺失数据不太敏感，算法比较简单，常用于文本分类。

### 什么是监督学习的标准方法？

有的回归算法和分类算法都属于监督学习并且明确的给给出初始值在训练集中有特征和标签，并且通过训练获得一个模型，在面对只有特征而没有标签的数据时，能进行预测。

### 在机器学习中，模型的选择是指什么？

根据一组不同复杂度的模型表现，从某个模型中挑选最好的模型。选择一个最好模型后，在新的数据上来评价其预测误差等评价和指标。

### 图形数据库Neo4J的优劣势？

　优势：1.更快的数据库操作，前提是数据量足够大。  
　　2. 数据更加直观，相应的SQL语句更加好写。  
　　3.更灵活，不管有什么新的数据需要储存，都是一律的节点，只需要考虑节点属性和边属性。  
　　4.数据库的操作不会随着数据库的增大有明显的降低。  
　　劣势：1.极慢的插入速度。  
　　2.超大的节点。当一个节点的边非常多，  
　　有关这个节点的操作速度就会大大下降

### LR和SVM的联系与区别是什么？

都是分类算法  
　　如果不考虑核函数，LR和SVM都是线性分类算法，也就是说他们的分类决策面都是线性的。  
　　LR和SVM都是监督学习算法  
　　LR和SVM的损失函数不同  
　　SVM只考虑局部的边界线附近的点 ，LR考虑全局，远离的点对边界线的确定也起作用。

### 什么是聚类，聚类的应用场景？

聚类是指根据一定的准则，把一份事物按照这个准则归纳成互不重合的几份，机器学习中，聚类指按照一个标准，这个标准通常是相似性，把样本分成几份，是得相似程度高的聚在一起，相似程度低的互相分开。  
　　聚类的应用场景，求职信息完善（有大约10万份优质简历，其中部分简历包含完整的字段，部分简历在学历，公司规模，薪水，等字段有些置空顶。希望对数据进行学习，编码与测试，挖掘出职位路径的走向与规律，形成算法模型，在对数据中置空的信息进行预测。）

# 技术预研

## 模型训练

### CRF模型训练

CRF模型广泛应用于命令实体识别。本平台提供命名实体识别服务，目前，本版本的命名实体识别服务主要采用CRF模型。

模型训练的难点，第一，用户提供的训练语料规模可能比较大，并且标注项可能比较多，则生成的特征会非常庞大。在这种情况下，训练任务对算法和机器内存都提出了更高的要求；第二，多个用户同时训练模型的情况下，如何做到对不同的用户的命令进行隔离，分配硬件和资源，对系统算法设计提出了更高的要求。

## [MongoDB](https://www.baidu.com/link?url=1J8unZor153sqOZQEekQOauUwAbTutx0qGi1FRE47IsZ72rI5cwzbV5GsB6EiCZMvKZoIObjcI9bkwC4S2ykBK&wd=&eqid=c22df1bf00086d83000000035d3198cd)

## Spark

## OrientDB

## MySql集群搭建

## Redis

## rabbitmq

# 整合Acitiviti在线流程设计器(未实现)

## 月工作日志

### 周工作日志

### 2019-06-18

1. **产品基本介绍**
2. 业务级；

产品-PPT;

NLP;

OCR;

1. 知识图谱；
2. 智慧语义；
3. 工作流；
4. **Activiti技术研究应用**
5. 启动事件（Start Event）的类型

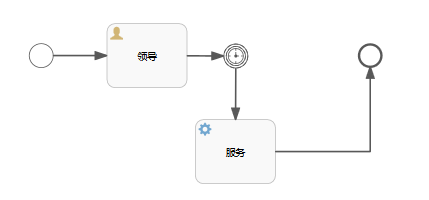
* **分类：**空启动事件（None Satrt Event）、定时启动事件（Timer Satrt Event）、信号事件(Signal Start Event)、消息事件(Message Start Event)、异常事件(Error Start Event)
* 1．空启动事件就是一个圆圈。在工作流里面使用的非常多。但是空启动事件必须要人工去启动一个流程。空启动事件是建立在部署完成的流程上面的。每启动一次，就会产生一个新的流程实例。
* 2．定时器启动事件用于在给定的时间点创建流程实例。它可以用在只启动一次的流程中，也可以用在特定时间间隔下启动。如果是有重复执行的,流程在周期内都有效，可以重复利用的流程；子流程中不能使用定时器启动事件。 定时器是从流程部署开始计时，不需要去启动流程。
* 3．信号启动事件，使用具名信号启动流程实例；信号启动事件的两种启动方式：信号可以是流程实例中抛出的信号事件、API触发；信号事件的执行方式：异步执行、同步执行。
* 4．消息事件，启动方式：流程实例中的信号抛出事件，API触发（*runtimeService.signalEventReceivedXXX*方法）触发，
* 注意事项：
* 1）流程的消息名称必须是唯一的，一个流程定义不得包含多个同名的启动消息。否则部署流程的时候就会抛异常。
* 2）消息启动事件，在所有部署的流程里面必须要唯一，否则也会抛异常。
* 3）直接启动消息定义事件，会当作一个普通启动事件执行。
* 4）新版本发布，会取消上一版本的消息订阅。
* 5）启动流程实例的三种方法 ProcessInstance startProcessInstanceByMessage
* 6）只有顶层流程（top­level process）才支持消息启动事件 ，嵌入流程（子流程）不支持消息事件。
* 7）如果一个流程定义多个消息事件无法使用runtimeService.startProcessInstanceByKey(…) 与runtimeService.startProcessInstanceById(…) 启动消息事件，但是一个流程定义单个消息事件可以启动。如果同一个流程定义同时具有多个消息事件和空启动事件，就会启动空启动事件，同一个流程定义多个消息事件的时候，使用这个方法启动，就会抛异常。
* 5．BPMN2.0规定了错误开始事件只能使用在事件子流程（Event Sub-Process）中，该该事件不能使用在其他流程中，包括最高级流程（Top-Level Process）、嵌套子流程（Sub-Process）和调用子流程（Call Activity）。BPMN错误与Java异常不是一回事。事实上，这两者毫无共同点。BPMN错误事件是建模业务异常（business exceptions）的方式。

1. **用户任务（UserTask）**

* 单人签收、用户组待签收/待办理
* 代理人,完成签收这个时候就任务Task的Assignee属性就有了值；
* 推荐使用候选人或者指定委托人，已经签收过的任务列表，某种意义上我理解为真正的办理人
* 查询候选人或者指定委托人
* 已完成/已办结;任务的已完成（调用过complete方法）是指当前待办的任务状态，而非整个流程实例的完成状态；

1. **服务任务（ServiceTask）**

* **应用场景：**当客户有这么一个需求,下一个任务我需要自动执行一些操作，并且这个节点不需要任何的人工干涉，也就是说这个节点是自动化的。那么，这个当前面一个经办人员把任务发送下去的时候，自然而然的下一个节点就会开始马上执行。这个时候。我们就需要使用Activiti工作流的ServiceTask任务。
* **升级：**假如客户说，我并不想马上去执行这个任务，我只是想在某一个时间之后才去执行这个任务。那么上面的要求显然又不能满足需求了。我们又要怎么去升级改进呢？其实也很简单。我们前面已经实现了定时边界任务。我们只需要增加一个定时边界来处理就可以了。



1. 脚本任务（ScriptTask）

* **应用场景：**脚本任务一般是用在当前的监听器或者监听服务类都不能满足的情形下面，或者说后期系统维护，突然在不想改动系统的情况下需要对流程做一些适当的改变。仅仅是几个变量或者仅仅是一个计算公式等等。这个时候可以使用脚本任务。

1. [**排他网关（ExclusiveGateWay）**](https://www.cnblogs.com/dengjiahai/p/8542929.html)

* 排他网关exclusive gateway 经常使用流程变量决定流程下一步要选择的路径
* 排他网关（也叫异或网关 XOR gateway，或者更专业的，基于数据的排他网关 exclusive data¬based gateway），用于为流程中的决策建模。
* 当执行到达这个网关时，所有出口顺序流会按照它们定义的顺序进行计算。条件计算为true的顺序流（当没有设置条件时，认为顺序流定义为true）会被选择用于继续流程。
* 用排他网关时，只会选择一条顺序流。当多条顺序 流的条件都计算为true时，其中在XML中定义的第一条（也只有这条）会被选择，用于继续流程。如果没有可选的顺序流，会抛出异常。
* 排他网关，用内部带有’X’图标的标准网关（菱形）表示，'X’图标代表异或（XOR）的含义。请注意内部没有图标的网关默认为排他网关。BPMN 2.0规范不允许在同一个流程中，混合使用带有及没有X的菱形标志
* 一个排他网关对应一个以上的顺序流
* 1、语法结构 ${expression}

1. [**并行网关（ParallelGateway）**](https://www.cnblogs.com/dengjiahai/p/8542929.html)

* **什么是并行网关：**它可以将分支（fork）为多个路径，也可以合并（join）多个入口路径。
* **并行网关的前提条件：**基于出口顺序流和入口顺序流。
* 并行网关两个重要特性：
* 分支（fork）： 并行后的所有外出顺序流，为每个顺序流都创建一个并发分支。
* 合并（join）： 所有到达并行网关，在此等待的进入分支， 直到所有进入顺序流的分支都到达以后， 流程就会通过汇聚网关。
* **注意：**
* 1．如果并行网关同时具有入口顺序流和出口顺序流，并行网关会先执行入口顺序流，然后执行再分裂成多条可以执行的路径。
* 2．并行网关不执行计算条件。并行网关上面的计算条件会被忽略。
* 3．并行网关分支和合并是同时存在的，就是说，一个至少有一个分支和一个合并。但是，网关是可以不平衡的，分支和合并的数量可以不一致

1. **事件监听**

* 工作流程事件监听可用于任务提醒、超时提醒等的模块的设计

1. **表单的介绍以及formService使用**

* 在Activiti中总共有三种表单，动态表单，普通表单和外置表单.
* **动态表单:** 用activiti:formProperty属性定义，可以在开始事件(startEvent)和任务(Task)上设置表单的动态内容，表单的内容都是以key和value的形式数据保存在引擎表中 ;因此，动态表单的布局在顺序的显示在页面上，基本上毫无布局可言，但是同时这种方式对于不需要布局的流程，在开发和实施的过程中，是最为方便的 .
* **普通表单:** 其中activiti:formKey就是来定义我们启动界面或者某个任务环节处理界面的对应的表单标识的key ;所以，普通表单的启动界面和任务处理界面都是需要我们自己来定义维护的，因此布局肯定比动态表单要优美很多。
* **外置表单：**这种方式常用于基于工作流平台开发的方式，代码写的很少，开发人员只要把表单内容写好保存到.form文件中即可，然后配置每个节点需要的表单名称（form key），实际运行时通过引擎提供的API读取Task对应的form内容输出到页面。

### 2019-06-19

1. **国特色的流程审批服务**
2. **事件监听**
3. **流程串行、并行审批**
4. **子流程的审批**
5. **流程版本变更**
6. **流程的自由流程**
7. **流程的人员变更处理**
8. **流程任务的分发与汇总**
9. **流程定义的会签**
10. **流程的催办**
11. **流程超时跳过**
12. **流程的工作日历**
13. **流程表单的在线配置**
14. **流程分支的配置处理**
15. **流程的组架构整合**
16. **流程表单的权限配置控制**
17. **流程消息通知**
18. **流程的回退与追回**
19. **流程的抄送与阅读控制**
20. **流程表单的模板打印等**
21. **多种流程开发工具集的支持**
22. **流程在线定义、**
23. **表单自定义配置、**
24. **查询设计、**
25. **列表设计、**
26. **表单方案设计、**
27. **流程方案设计、**
28. **组织架构设计、**
29. **数据字典、**
30. **选择对话框设计、**
31. **数据源设计、**
32. **流水号设计、**
33. **流程授权设计等，**
34. **可以满足各种流程的业务扩展需求。**
35. **流程方案配置**
36. **提供组装流程业务的解决方案的管理，把流程定义、审批人员、流程表单、审批时的事件及交互脚本调用等组装起来，实现真正意义上的BPMN的流程业务规则。如审批时，执行写入其他数据库的操作。支持各种事件的脚本交互处理；同时可让平台扮演流程管理中心，支持第三方平台的流程应用调用。**



1. **提供流程级别与节点大量不同的属性配置，以增加流程在执行过程中与外部的交互能力：**



1. **待办任务处理**
2. **在线可实现任务的办理、转办、抄送、加签、作废、回退上一步、回退发起人、自由跳转等。**



1. **工作流系统业务功能梳理**

### 2019-06-20

1. **人员工作量梳理**

**Java高级工程师 1名、承担工作：**

基于Spring Boot/Cloud、MyBatis、Redis、MongoDB、RabbitMQ、mysql、HBase、swagger、security、Oauth2、Elasticsearch、bootstrap、Angular等开源框架和数据库整体架构搭建微服平台；

1. 负责平台架构及优化，对技术细节负责，负责设计和开发平台核心代码；
2. 指导开发人员开发，负责评审开发人员的开发结果，指导解决各类技术问题，并跟踪执行；
3. 负责完成重要业务模块及核心框架的搭建和持续优化；
4. 负责平台的性能评测、容量评估，线上技术问题跟踪和处理；
5. 负责关键技术研究及技术预研，并结合实际应用需求运用于平台系统架构中；
6. 通过开发工具或开发方法的改进，提高开发效率，参与制定设计及实现各项研发规范，指导设计、实现及部署工作。
7. 带领java工程师团队和前端团队按照项目计划，与项目组其他成员协同工作，在保证质量的前提下，按时完成开发任务；
8. 负责技术团队的管理，制定开发规范，撰写相关技术文档指导、培训工程师；
9. 负责解决核心技术问题，对技术方案进行决策；
10. 工作流系统设计及业务整合设计；
11. 改进并优化activiti6工作流引擎开源框架；

**Java中级级工程师 1名**

1. 实现系统工作流java源码修改，并应用到实际系统；
2. 协助前端调试并修改activiti6相关接口；
3. 具体开发大的模块有流程管理、表单管理、用户组织管理、系统配置、个人配置、流程应用；

**Java中级级工程师 1名**

1. 编写NLP自然语言平台后端业务代码编写；
2. 编写前端所需业务应用接口及业务功能实现代码；
3. 编写与Python算法应用接口调用业务功能实现的代码；
4. 具体开发大的模块有

**Java中级级工程师 1名**

1. 编写NLP自然语言平台后端业务代码编写；
2. 编写前端所需业务应用接口及业务功能实现代码；
3. 编写与Python算法应用接口调用业务功能实现的代码；
4. 具体开发大的模块有；

**前端工程师 1名、承担工作：**

1. 实现前端总体框架搭建及各种应用插件基础选型初步验证；
2. 实现工作流业务系统前端功能代码编写；
3. 具体开发大的模块有流程管理、表单管理、用户组织管理、系统配置、个人；配置、流程应用；

**前端工程师 1名、承担工作：**

1. 实现NLP智慧语义分析平台前端业务功能代码编写；
2. 具体开发大的模块；

**前端工程师 1名、承担工作：**

1. 实现NLP智慧语义分析平台前端业务功能代码编写；
2. 具体开发大的模块；

**知识图谱，**

**关键字搜索，知识图谱搜索；**

**智能推荐；**

**智能编写；**

**所以浏览器要兼容；**

### 2019-06-21

1. **人员工作量梳理**

**Elasticsearch**

1. ElasticSearch是一个基于Lucene的搜索服务器；
2. 它提供了一个分布式多用户能力的全文搜索引擎，基于RESTful web接口；
3. 设计用于[云计算](https://baike.baidu.com/item/%E4%BA%91%E8%AE%A1%E7%AE%97/9969353)中，能够达到实时搜索，稳定，可靠，快速，安装使用方便；
4. 分布式的实时文件存储，每个字段都被索引并可被搜索
5. 分布式的实时分析搜索引擎
6. 可以扩展到上百台服务器，处理PB级结构化或非结构化数据
7. 学习资料https://es.xiaoleilu.com/index.html

基于Oauth2 实现SSO单点登录，

[**https://blog.csdn.net/qq\_36807862/article/details/81284990**](https://blog.csdn.net/qq_36807862/article/details/81284990)

[**https://github.com/chenfromsz/spring-boot-sso**](https://github.com/chenfromsz/spring-boot-sso)

Mysql数据库读写分离，

1. **what 读写分离**   
   读写分离，基本的原理是让主数据库处理事务性增、改、删操作（INSERT、UPDATE、DELETE），而从数据库处理SELECT查询操作。数据库复制被用来把事务性操作导致的变更同步到集群中的从数据库。

分类体系（目录结构）-》（语料）标注——》机器训练-----》训练模型🡪

智能填写----》实体抽取（规则方式（创建规则库）；学习方式抽取（模型训练））-------》

四大架构：业务（用户业务逻辑）、数据（数据单元，数据流转）、技术（技术实现）、应用（应用实现）

架构三级：初级设计（费用，做什么），概要设计（怎么实现，实现方法、有节点、有方法），详细设计（实现步骤，（元素，属性））；

架构设计方法论：

角色（用户，管理员、合作伙伴），场景（每个业务流程，对象（属性），做什么），流程，节点，元素，属性

### 2019-06-24

### 2019-06-25

1. **数据库设计**
2. **基础环境搭建**