



# 某品牌手机的京东评论数据分析

# 目录

---

1	背景与目标
2	数据预处理
3	可视化分析
4	关键信息分析
5	文本评论倾向分析

# 背景与目标

---

电商成熟发展，引发激烈市场竞争

懂消费者的产品得天下

电商评论文本数据具有分析价值



# 背景与目标

企业可以利用海量评论数据去更好地了解用户喜好，从而提高产品质量，改善服务，获取市场上的竞争优势。

消费者需要评论数据为购物抉择提供参考依据。



## 背景与目标

在诸多外部限制和打压情况下，Apple产品发布iphone13系列手机，一经发布便引发抢购热潮，获得大量关注。

iPhone 13 Pro Max



iPhone 13 Pro




iPhone 13



# 背景与目标

官方旗舰店<https://item.jd.com/100026667928.html>



Apple 产品京东自营旗舰店

Authorized Reseller  
授权经销商

首页MaciPadiPhoneWatch配件教育优惠App Store 充值卡

手机通讯 > 手机 > 手机 > Apple > Apple iPhone 13 Pro ...

自营Apple产品京东自营旗舰店联系客服★关注店铺




< 关注 对比 举报

Apple iPhone 13 Pro Max (A2644) 128GB 远峰蓝色 支持移动联通电信5G 双卡双待手机

该商品已下柜，欢迎挑选其他商品！


相似商品推荐1/2



限时领券立减 200 元

Apple iPhone 13 (A2634) 128GB 星光色 支持移动联通电信5G 双卡


¥4799.00



限时领券立减 3599 元

Apple iPhone 11 (A2223) 128GB 白色 移动联通电信4G手机 双卡双


¥4299.00



限时领券立减 1100 元


Apple iPhone 14 Pro (A2892) 128GB 暗紫色 支持移动联通电信

¥7799.00




限时领券立减 200 元

Apple iPhone 13 promax系列 全新 原装美版有锁 移动联通电信 直播



限时领券立减 200 元

Apple iPhone 14 (A2884) 128GB 黄色 支持移动联通电信5G 双卡双



限时领券立减 200 元

Apple iPhone 14 Plus (A2888) 128GB 蓝色 支持移动联通电信5G

6

# 背景与目标

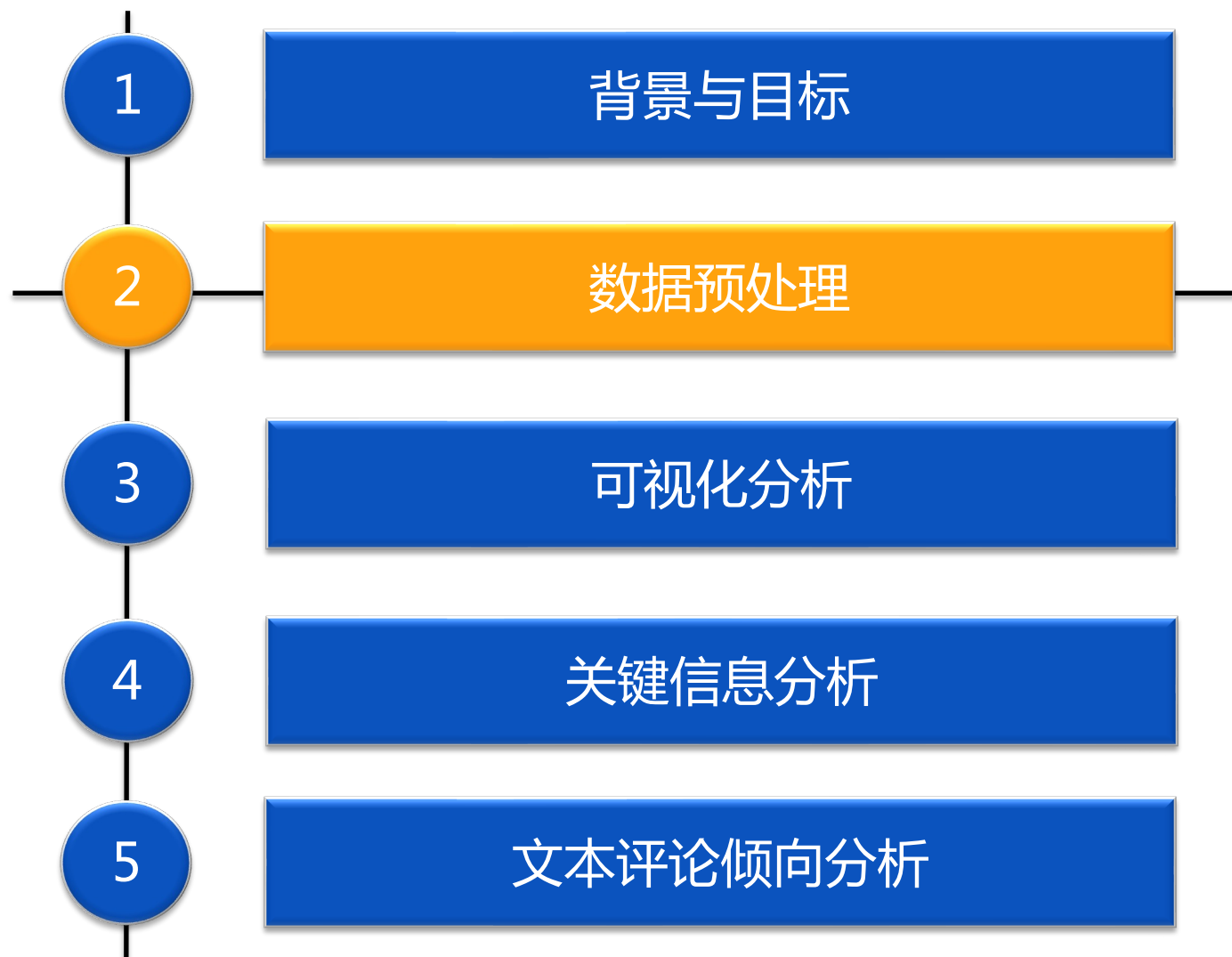
---

本案例主要针对用户在电商平台上留下的评论数据，对其进行预处理及可视化分析，然后提取评论关键信息，了解用户的需求、意见、购买原因以及产品的优缺点，并提出改善产品的建议。最后构建模型判别评论的情感倾向。案例流程：

1. 数据预处理
2. 可视化分析
3. 提取评论关键信息
4. 文本评论倾向分析

# 目录

---





# 数据预处理

## 数据介绍

	content	creationTime	score	userClient	productColor	productSize	referenceTime	nickname	days
0	苹果十三是我用我最好的手机了，节前买了金色，蓝色，黑色，这次又忍不住真的活动买了绿色，还是他...	2023-03-30 16:16:29	5.0	2.0	苍岭绿色	512GB	2023-03-21 07:00:05	超***哈	9.0
1	相比14pm还是觉得13pm顺眼点，不错屏幕很大，很流畅，希望能多用几年。\\n外形外观：酷\\n	2023-03-21 17:59:36	5.0	4.0	石墨色	512GB	2023-03-09 16:58:36	***	12.0
2	外形外观：绿色大气很适合男生\\n屏幕音效：屏幕看着很舒服\\n拍照效果：拍照效果：苹果 拍照不...	2023-03-10 11:50:06	5.0	4.0	苍岭绿色	512GB	2023-03-09 10:38:43	j***k	1.0
3	不愧是13pm，第四次购买了，京东国行正品有保障，还是熟悉的感觉，真的爽的一比，物流也很快， ...	2023-03-04 01:41:13	5.0	4.0	金色	128GB	2023-02-28 13:12:10	黑***6	4.0
4	外形外观：不错看上去很精致\\n屏幕音效：也很不错目前苹果品控还是可以的\\n拍照效果：1200...	2023-03-03 10:24:19	5.0	2.0	银色	1TB	2023-01-27 22:01:20	j***5	35.0

名称	说明
content	评论文本
creationTime	评论时间
score	评分
userClient	设备类型
productColor	商品颜色
productSize	商品规格
referenceTime	购买时间
nickname	用户昵称
days	付款到写评间隔天数

# 数据预处理

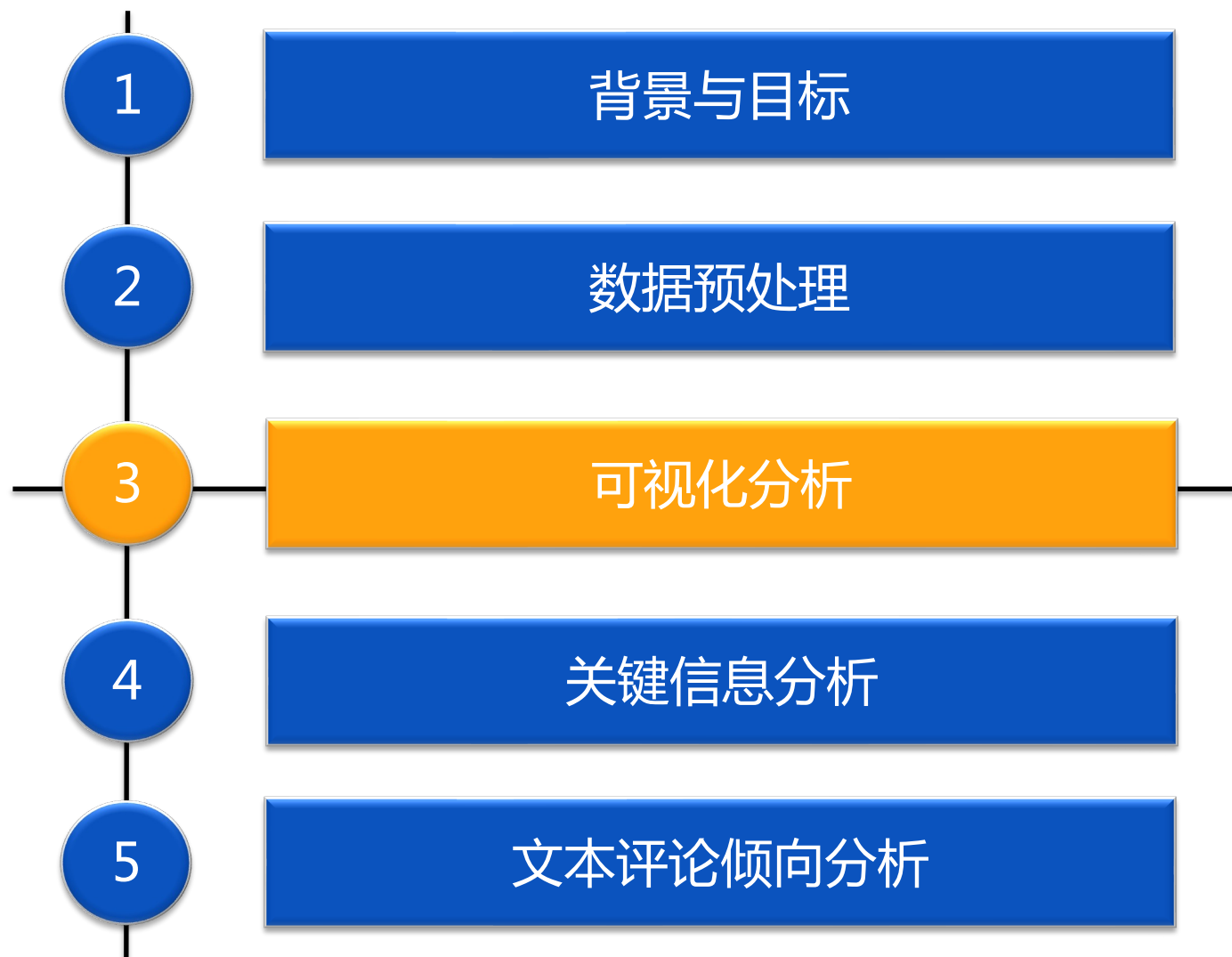
---

## 数据预处理介绍

- 合并数据集
- 换行符处理
- 商品规格数据处理: ['512GB', '128GB', '1TB', '256GB', '256G', '128G', nan]
- 时间格式转换

# 目录

---



# 可视化分析

---

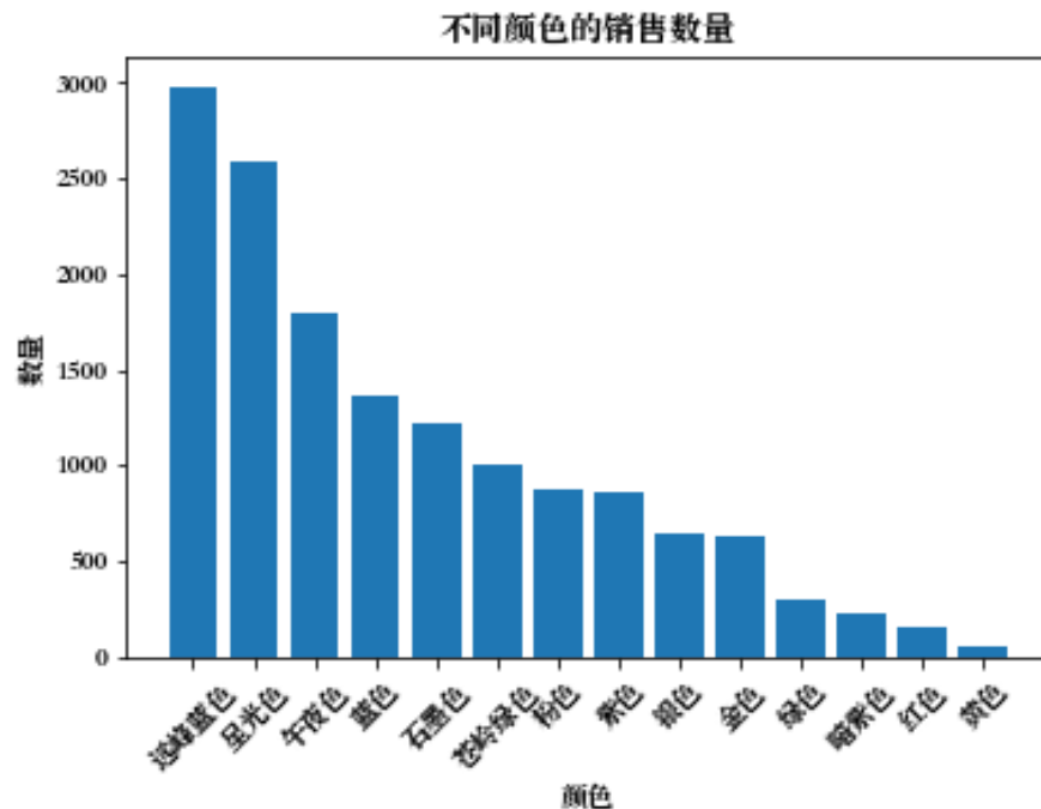
可视化分析介绍：

1. 分析不同颜色商品销售情况
2. 分析购买该商品不同配置（规格）的比例
3. 分析该商品的销售数量和评论数量和时间关系
4. 分析评论时间与购买时间间隔天数

# 可视化分析

## 购买该商品不同颜色的分析

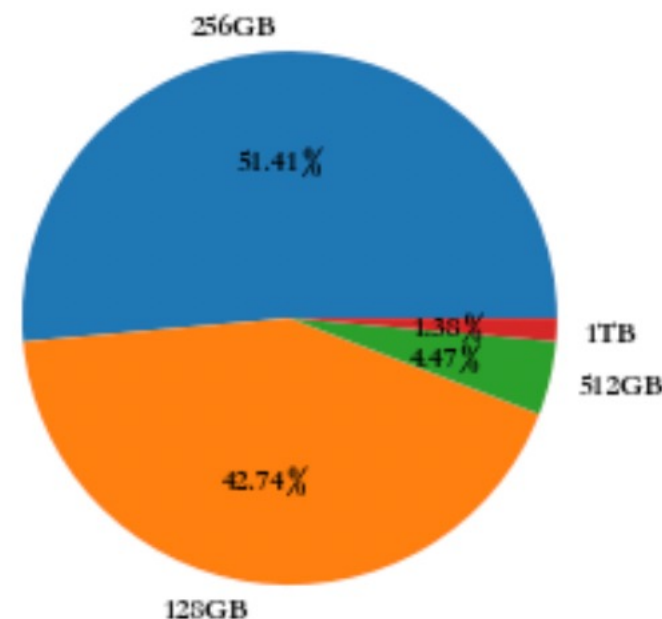
- 购买颜色中偏好不一，但是黄色手机的购买量较少。远峰蓝色的购买量最多，其次是星光色和午夜色。
- 只有iPhone13Pro系列才有远峰蓝色，深受大众喜爱。
- 星光色和午夜色经典耐看，选择的人较多。
- 黄色比较个性化，选择的人不多。



# 可视化分析

## 购买该商品不同配置的比例

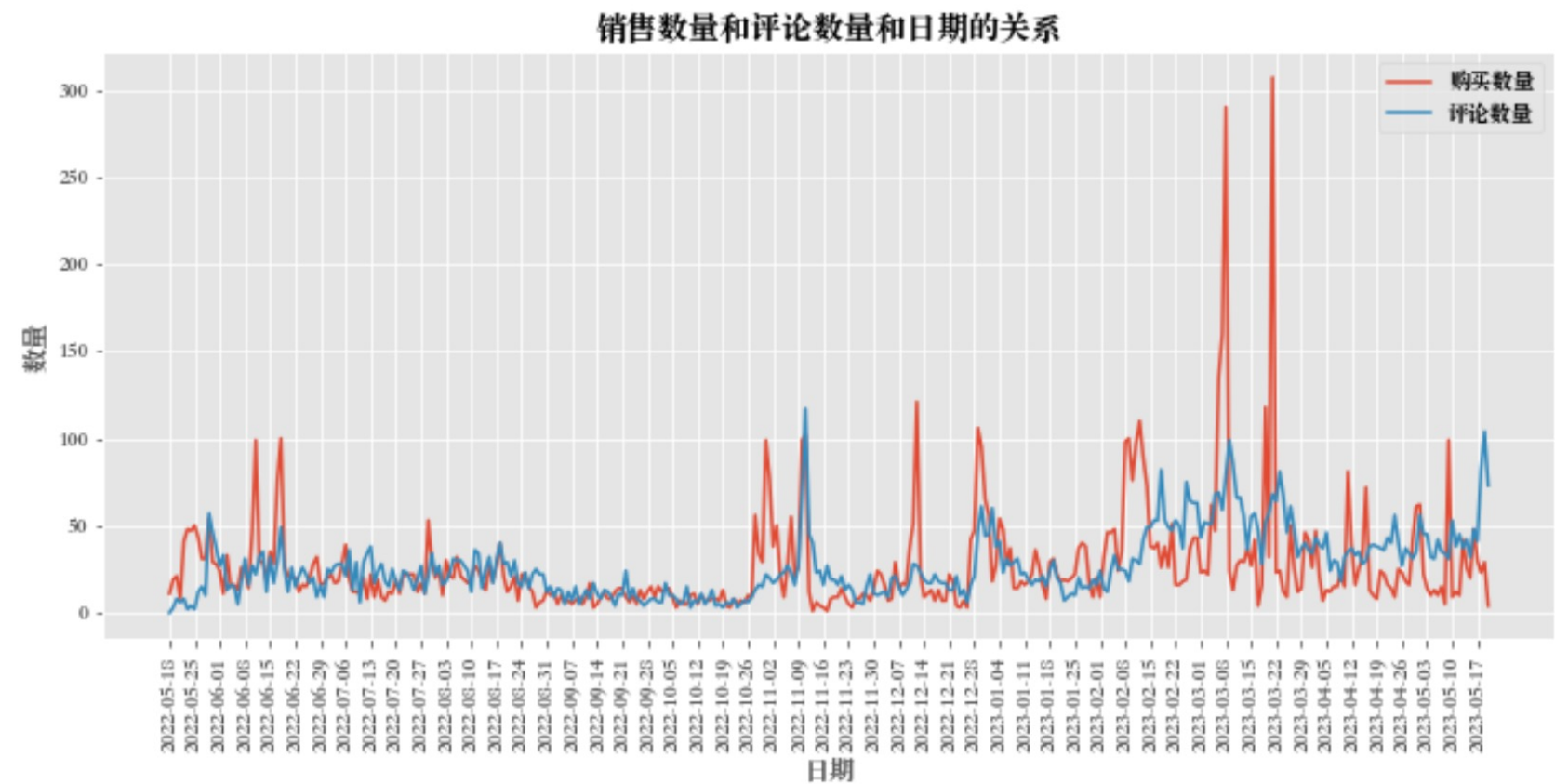
- 256GB内存容量需求最大，其次是128GB版本，最少的是1TB。配置主要权衡使用体验与价格，每个内存版本都有几百的差距。
- 128GB内存也略显不足，考虑长时间使用导致内存积压很容易影响日常用机，故需求变少。
- 256GB大小适中实用，相对来说性价比最高，选择的人数较多。
- 部分高端消费者追求极致，选择购买512GB和1TB版本的手机。



# 可视化分析

## 销售数量、评论数量和日期的关系

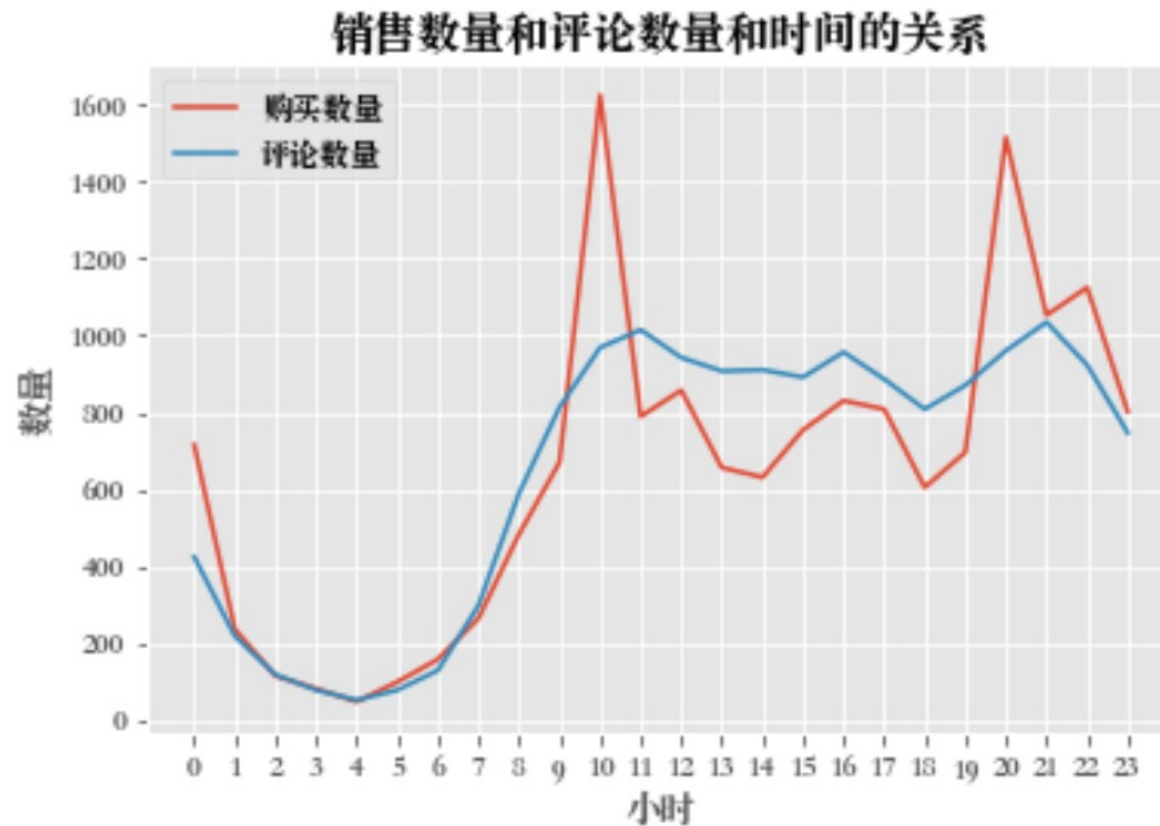
购买手机有波峰，波峰出现主要有两种情况，一种是首次发售刺激的消费，另外一种就是电商平台促销活动刺激的消费，如双十一、双十二。



# 可视化分析

## 购买数量与评论数量随时间的变化情况

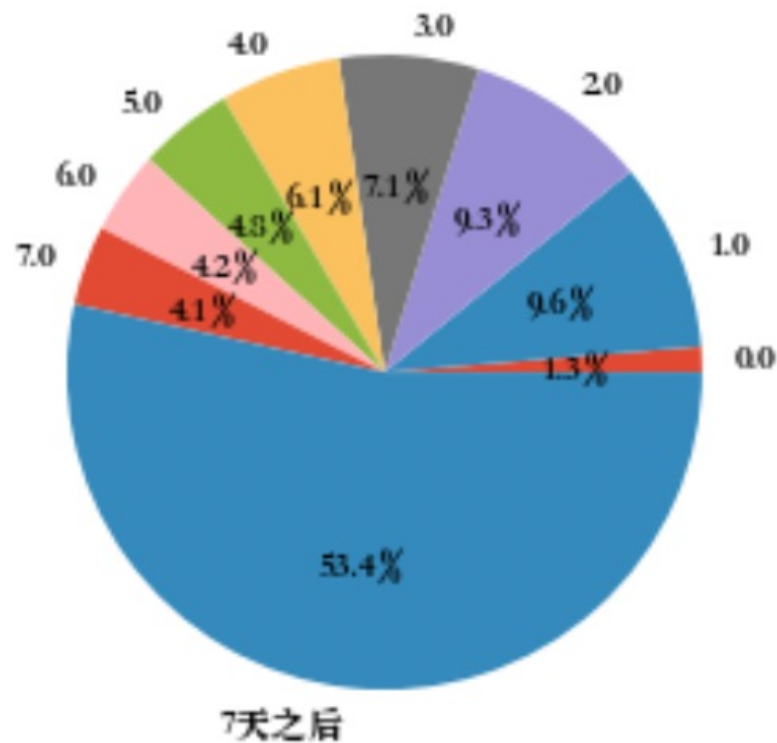
- 购买手机多集中在9点到11点之间，评论时间均匀分布在上午10点至晚11点之间。
- 商家一般会把秒杀，发放优惠券和红包的时间定在早上10点或者晚上8点。
- 建议：更多宣传的成本放到晚上8点之后的时间段





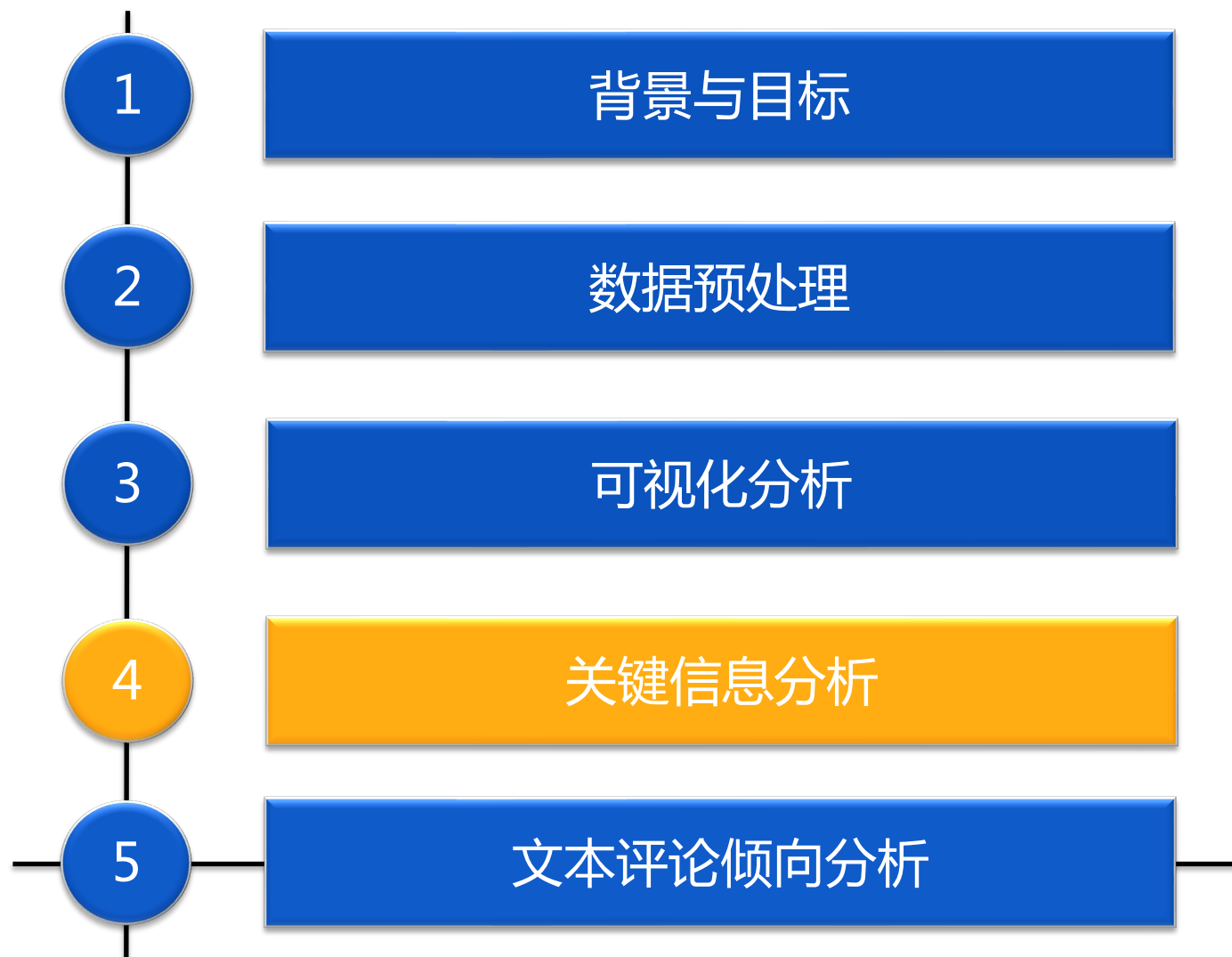
# 可视化分析

- 假设所有人都是在购买并拿到货物后发表评论，有近一半的人是购物后一周内拿到货物的（因没有强制要求拿到货物后马上评论，所以不能看最晚收货时间）。
- 部分用户能够在购买当天收到商品，约1.3%的用户选择购买当天评论。



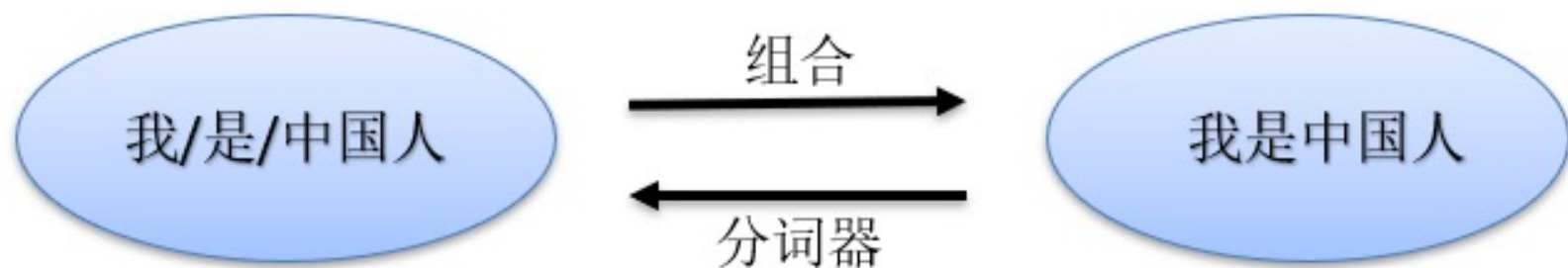
# 目录

---



## 文本评论分词

在英文文本中，词和词之间有明显的分隔符，通过空格可以划分出每一个单词，而对于中文文本中的“词”和“词组”来说，它们的边界模糊，没有一个形式上的分界符。因此，进行中文文本挖掘时，首先应对文本分词，即将连续的字序列按照一定的规范重新组合成词序列。



分词结果的准确性对后续文本挖掘算法有着不可忽视的影响，如果分词效果不佳，即使后续算法优秀也无法实现理想的效果。例如在特征选择的过程中，不同的分词效果，将直接影响词语在文本中的重要性，从而影响特征的选择。

# 分词的基本方法

---

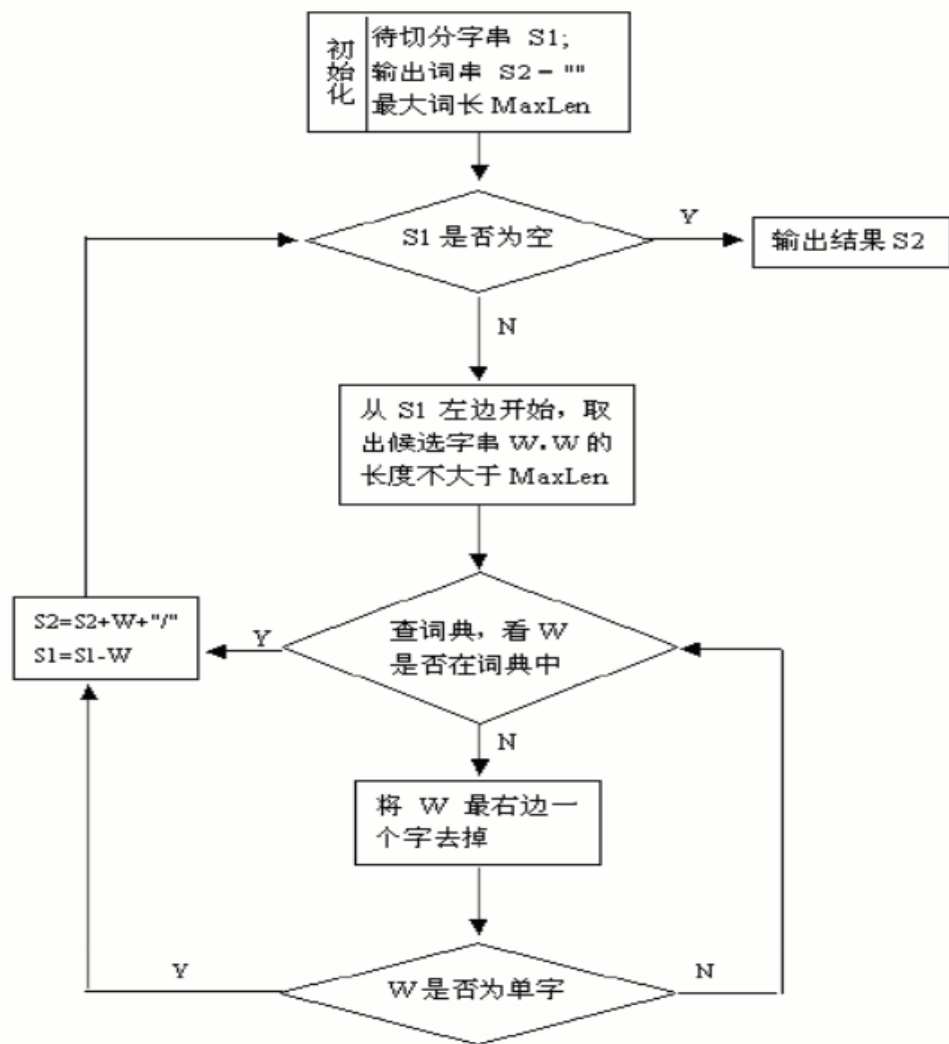
- 基于字符串匹配
  - ✓ 最大匹配法和逆向最大匹配法
  - ✓ 双向最大匹配法

基于统计模型

- ✓ 最大概率法
- ✓ 隐马尔科夫
- ✓ Jieba分词

# 分词的基本方法

- ✓ 最大匹配法和逆向最大匹配法
- ✓ 双向最大匹配法



# 分词的基本方法

## ✓ 最大匹配法和逆向最大匹配法

S1=“今天我们要学习文本挖掘案例”为输入字符串。设定最大词长 MaxLen=5，S2=“”为初始输出字符串。

分词步骤：

- (1) S2=“”；S1 不为空，从 S1 左边取出候选子串 W=“今天我们要”；
- (2) 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W=“今天我们”；
- (3) 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W=“今天我”；
- (4) 查词表，“今天”在词表中，将 W 加入到 S2 中，S2=“今天”，并将 W 从 S1 中去掉，此时 S1=“我们要学习文本挖掘案例”；
- (5) S1 不为空，于是从 S1 左边取出候选子串 W=“我们要学习”；
- (6) 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W=“我们要学”；
- (7) 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W=“我们要”；
- (8) 查词表，W 不在词表中，将 W 最右边一个字去掉，得到 W=“我们”；
- (9) 查词表，“我们”在词表中，将 W 加入到 S2 中，S2=“今天/我们”，并将 W 从 S1 中去掉，此时 S1=“要学习文本挖掘案例”；
- .....
- (n-1) S2=“/今天/我们/要/学习 /文本/挖掘/案例/”，此时 S1=“”。
- (n) S1 为空，输出 S2 作为分词结果，分词过程结束。

# 分词的基本方法

---

## 双向最大匹配法

- ✓ 双向最大匹配法是将正向最大匹配法得到的分词结果和逆向最大匹配法得到的结果进行比较，然后按照最大匹配原则，选取词数切分最少的作为结果。
- ✓ 双向最大匹配的规则是：
  - (1) 如果正反向分词结果词数不同，则取分词数量少的那个。
  - (2) 如果分词结果词数相同：
    - a) 分词结果相同，没有歧义，返回任意一个。
    - b) 分词结果不同，返回其中单字数量较少的那个。
- ✓ 例如：“独立自主和平等互利的原则”

最大匹配法的划分：“独立自主 / 和平 / 等 / 互利 / 的 / 原则”一共有 6 个词；

双向最大匹配法的划分“独立自主 / 和 / 平等互利 / 的 / 原则”，一共只有 5 个词。

## 分词的基本方法

---

但是，很多句子也有不止一个词数最少的分词方案，双向最大匹配法并不能从中选出一个最佳答案。

- ✓ “为人民办公益”的最大匹配划分和双向最大匹配法划分都是“为人 / 民办 / 公益”
- ✓ 正确的划分则是“为 / 人民 / 办 / 公益”



# 分词的基本方法

---

## 最大概率法

假设一个句子S有多种分词方法，为了简单起见，假定有三种：

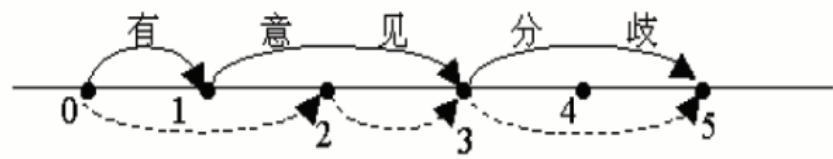
- $A_1, A_2, A_3, \dots, A_k$
- $B_1, B_2, B_3, \dots, B_m$
- $C_1, C_2, C_3, \dots, C_n$

我们用k,m,n三个不同的下标表示这个句子在采用不同分词结果时词的数目。最好的一种分词方法应该保证分词后这个句子出现的概率最大：

$P(A_1, A_2, A_3, \dots, A_k) > P(B_1, B_2, B_3, \dots, B_m)$  并且  $P(A_1, A_2, A_3, \dots, A_k) > P(C_1, C_2, C_3, \dots, C_n)$

# 分词的基本方法

一个待切分的汉字串可能包含多种分词结果， 将其中概率最大的那个作为该字串的分词结果。



路径 1: 0-1-3-5

路径 2: 0-2-3-5

该走哪一条路呢？

S: 有意见分歧

W1: 有/ 意见/ 分歧/

W2: 有意/ 见/ 分歧/

$Max(P(W1 | S), P(W2 | S))$  ?

$P(W | S) = \frac{P(S | W)P(W)}{P(S)} \approx P(W)$

$P(W) = P(w_1, w_2, ..., w_i) \approx P(w_1) \times P(w_2) \times ... \times P(w_i)$

$P(w_i) = \frac{w_i \text{在语料库中的出现次数}n}{\text{语料库中总词数}N}$

词语	概率
...	...
有	0.0180
有意	0.0005
意见	0.0010
见	0.0002
分歧	0.0001
...	...

$P(W1) = P(\text{有}) \times P(\text{意见}) \times P(\text{分歧}) = 1.8 \times 10^{-9}$

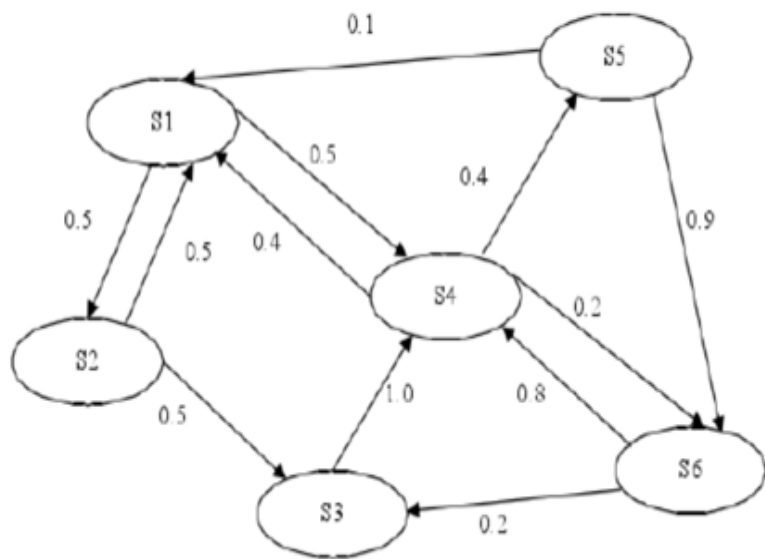
$P(W2) = P(\text{有意}) \times P(\text{见}) \times P(\text{分歧}) = 1 \times 10^{-11}$

$P(W1) > P(W2)$

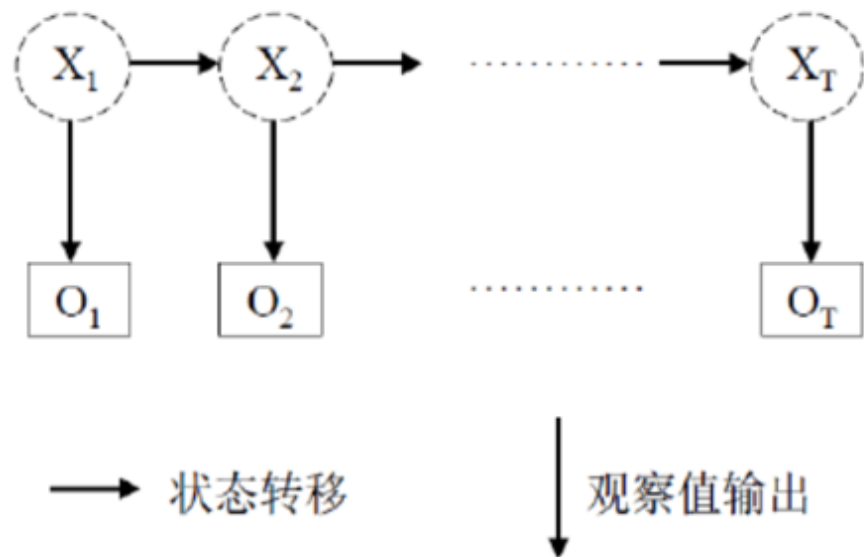
# 隐马尔科夫分词

隐马尔可夫模型（Hidden Markov Model, HMM）是统计模型，它用来描述一个含有隐含未知参数的马尔可夫过程。

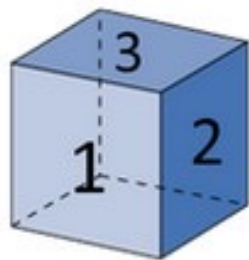
马尔科夫过程



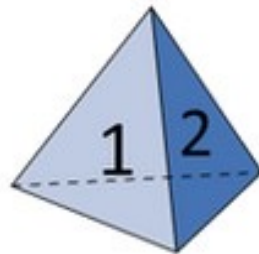
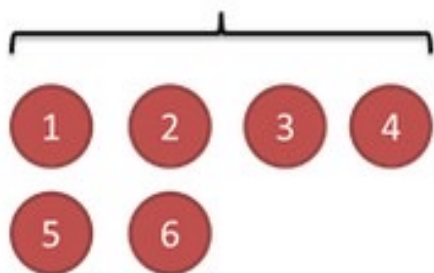
隐马尔科夫过程



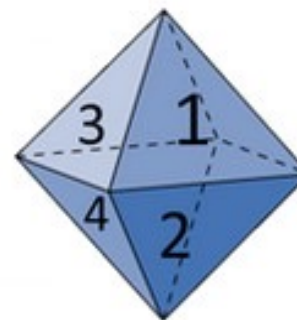
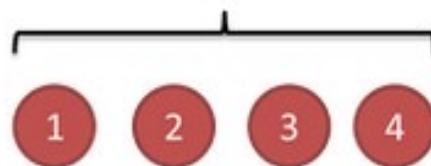
# 隐马尔科夫分词



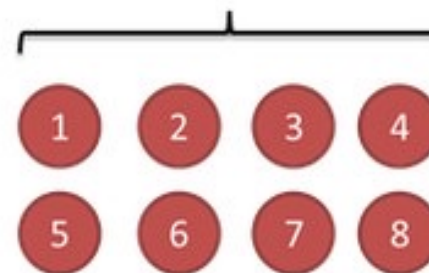
D6



D4



D8

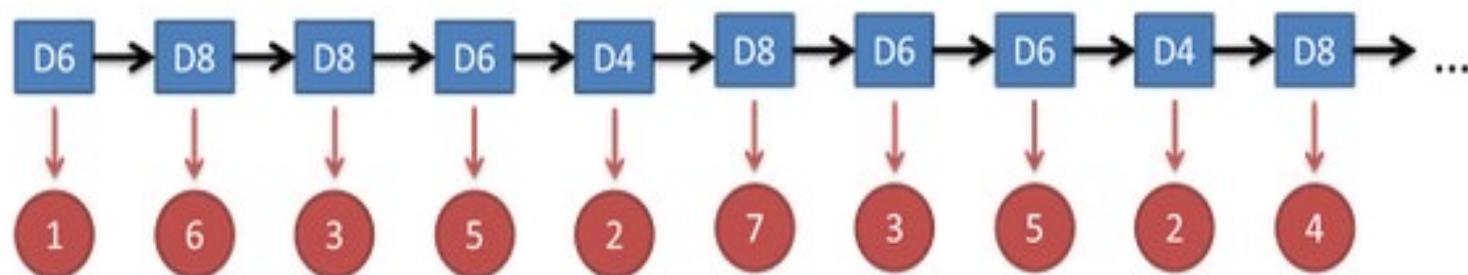


# 隐马尔科夫分词

可见状态：1 6 3 5 2 7 3 5 2 4

隐藏状态：D6 D8 D8 D6 D4 D8 D6 D6 D4 D8……

隐马尔可夫模型示意图

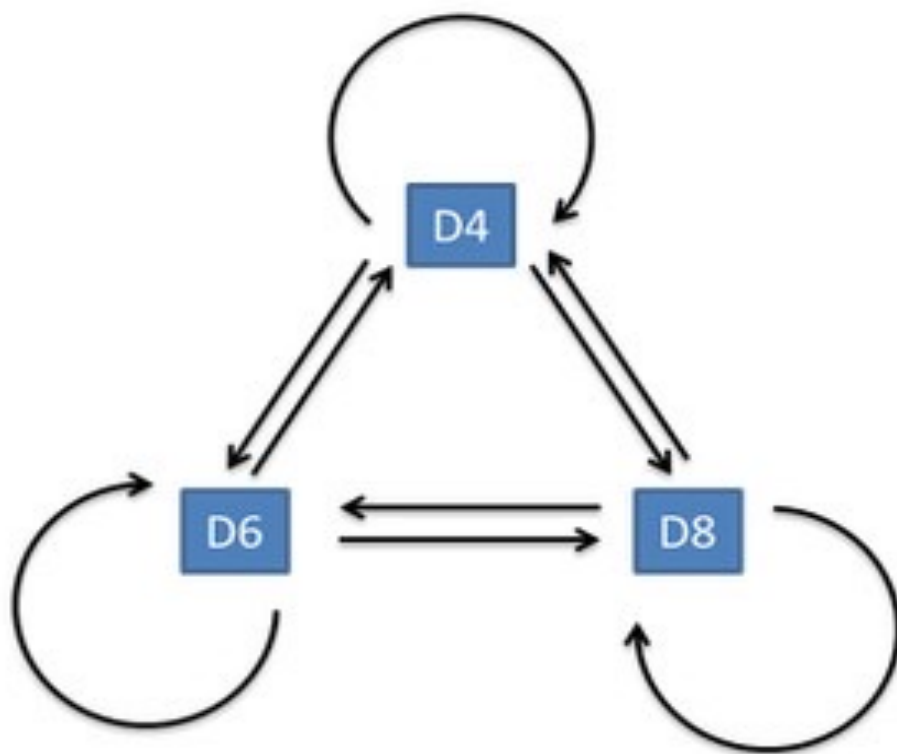


图例说明：



# 隐马尔科夫分词

隐含状态转换关系示意图



# 隐马尔科夫分词

## 模型表示

- 五元组( $S, V, p, A, B$ )

- 符号表

$V$ : 观察值集合,  $\{v_1, \dots, v_M\}$

–  $S$ : 输出状态集合,  $\{s_1, \dots, s_N\}$ 。

- 模型参数

–  $p$ : 初始状态概率。  $p = \{p_i\}; \quad i \in S$

–  $A$ : 状态转移概率。  $A = \{a_{ij}\}; \quad i, j \in S$

–  $B$ : 符号输出概率。  $B = \{b_{jk}\}; \quad j \in S, k \in V$

- 序列

• 输入观察序列:  $A = A_1, A_2 \dots A_T$

输出状态序列:  $O = O_1, O_2 \dots O_T$

# 隐马尔科夫分词

---

HMM模型作的两个基本假设：

1. 齐次马尔科夫性假设，即假设隐藏的马尔科夫链在任意时刻 $t$ 的状态只依赖于其前一时刻的状态，与其它时刻的状态及观测无关，也与时刻 $t$ 无关；
2. 观测独立性假设，即假设任意时刻的观测只依赖于该时刻的马尔科夫链的状态，与其它观测和状态无关。



# 隐马尔科夫分词

在隐马尔科夫中 $S$ ：输出状态集合， $\{s_1, \dots, s_N\}$ 共有四种情况：

词首（B），词尾（E），词中（M），单字词（S）

✓ 人民收入和生活水平进一步提高。

✓ 人/B民/E 收/B入E 和/S 生/B活/E 水/B平/E 进/B一/M步/E 提/B高/E。

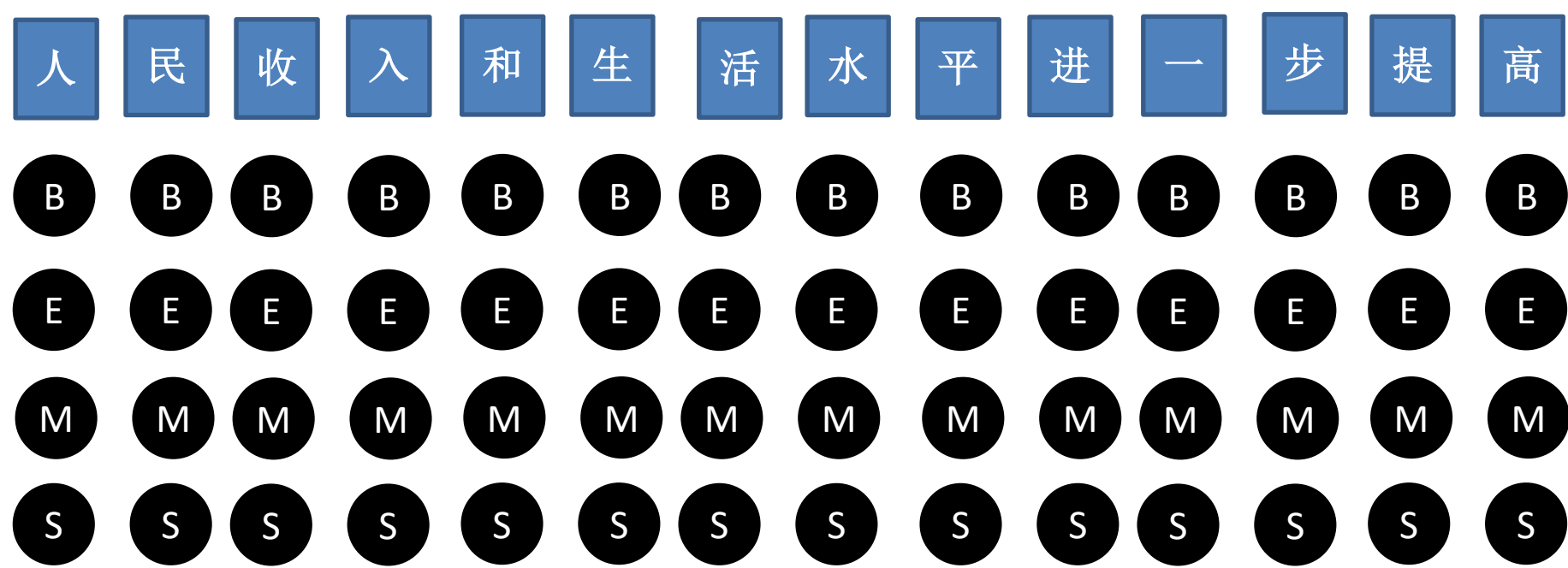
输入：人民收入和生活水平进一步提高。

输出：BEBESBEBEBMEBE .....

根据结果进行切分：人民/收入/和/生活/水平/进一步/提高/。

# 隐马尔科夫分词

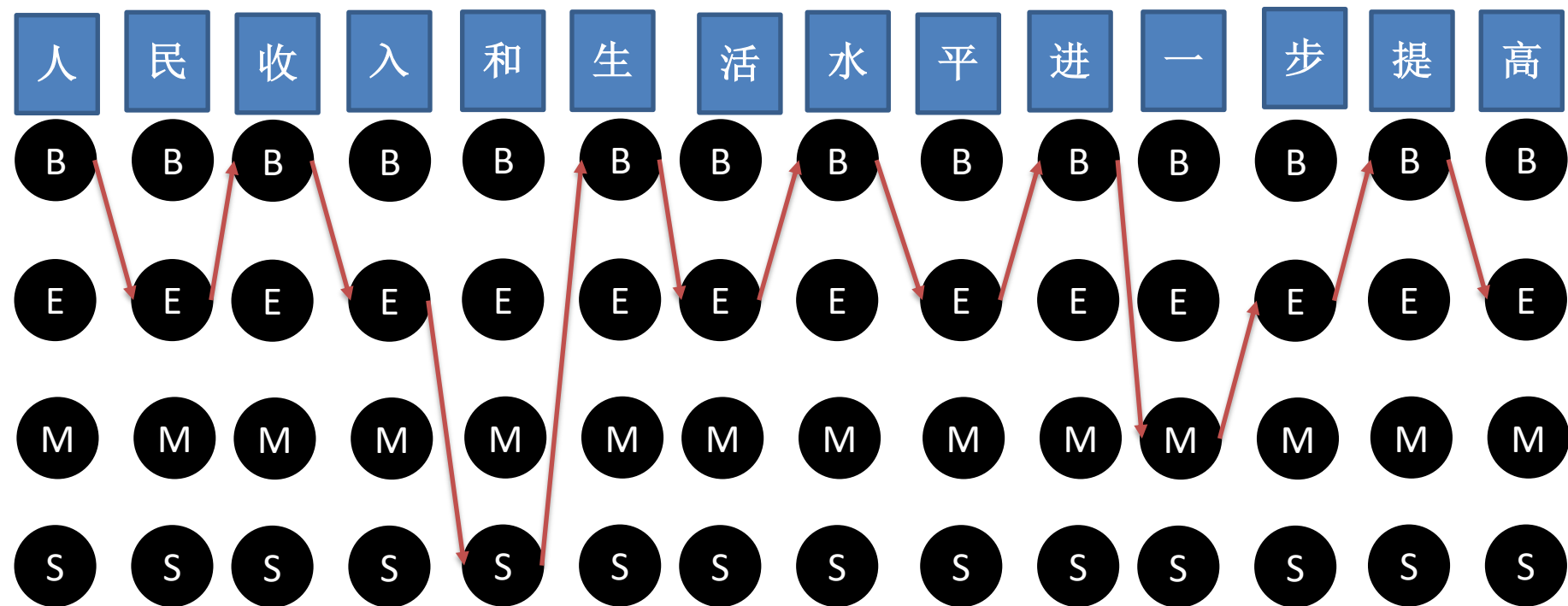
事实上，对于输入的词序列，根据语法词典，列出每个词可能的词性候选，构成词网格，即状态空间。



注：B后面只可能接(M or E)，不可能接(B or S)。而M后面也只可能接(M or E)，不可能接(B, S)

# 隐马尔科夫分词

采用Viterbi算法搜索词网格，搜索最佳路径（词性序列、状态序列）。根据动态规划原理，最优路径具有这样的特性：如果最优路径从结点  $i\{t\}$  到终点  $i\{T\}$ ，那么这两点之间的所有可能的部分路径必须是最优的。依据这一原理，我们只需从时刻  $t=1$  开始，递推地计算在时刻  $t$  状态为  $i$  的各部分路径的最大概率，直至得到时刻  $t=T$  状态  $i$  的各条路径的最大概率  $P$ ，最优路径的终结点  $i\{T\}$  也同时得到，这就是维特比算法。



## 分词器当前存在的问题

### ✓ 切分歧义

✓ 例如：“结婚的和尚未结婚的”

- “结婚/的/和/尚未/结婚/的”
- “结婚/的/和尚/未/结婚/的”

### ✓ 新词识别

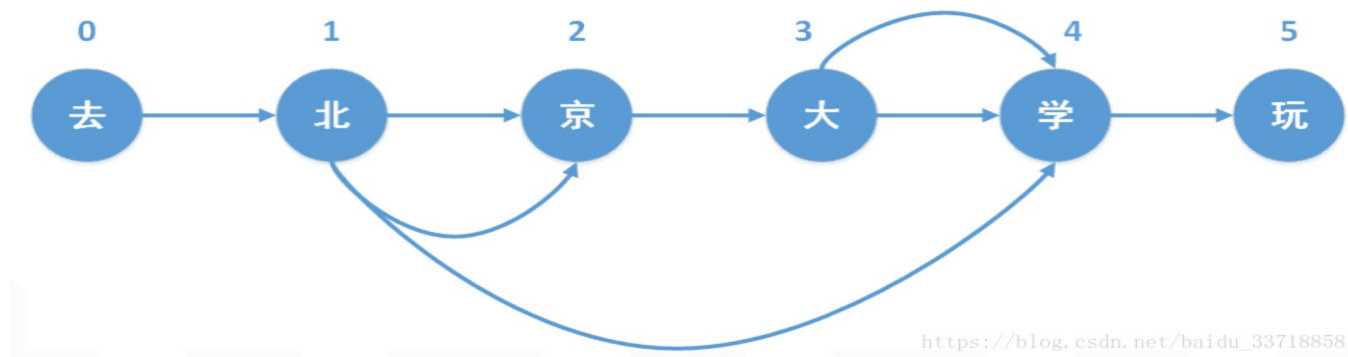
- 新词，专业术语称为未登录词。也就是那些在字典中都没有收录过，但又确实能称为词的那些词。

如：人名、地名、机构名、专业术语；

# jieba分词

➤ jieba——“结巴”中文分词：做最好的 Python 中文分词组件

- 对于切分歧义，基于前缀词典实现高效的词图扫描。生成某个句子中汉字所有可能成词情况所构成的有向无环图（路径）；然后基于前缀词典里面的词语词频，采用动态规划查找最大概率路径，找出最佳的切分组合（Viterbi算法）。



- 对于未登录词，采用了基于汉字成词能力的 HMM 模型。首先通过语料训练出HMM相关的模型，然后利用Viterbi算法进行求解，得到最优的隐藏状态序列，最后再根据隐藏状态序列，输出分词结果。

# 分析方法与过程

## ➤ 去除停用词

- ✓ 停用词(Stop Words)，词典译为“电脑检索中的虚字、非检索用字”。
- ✓ 停用词一定程度上相当于过滤词(Filter Words)，不过过滤词的范围更大一些，包含黄色、政治等敏感信息的关键词都会被视做过滤词加以处理，停用词本身则没有这个限制。通常意义上，停用词(Stop Words)大致可分为如下两类：
  - ① 使用十分广泛，甚至是过于频繁的一些单词。比如英文的“i”、“is”、“what”，中文的“我”、“就”之类词几乎在每个文档上均会出现。
  - ② 文本中出现频率很高，但实际意义又不大的词。这一类主要包括了语气助词、副词、介词、连词等，通常自身并无明确意义，只有将其放入一个完整的句子中才有一定作用的词语。如常见的“的”、“在”、“和”、“接着”之类。
    - # \$ % &
    - 啊啊哎呀哟唉俺俺们
    - ℃&\*一一~~~~，『. 一. /--』

# 关键信息分析

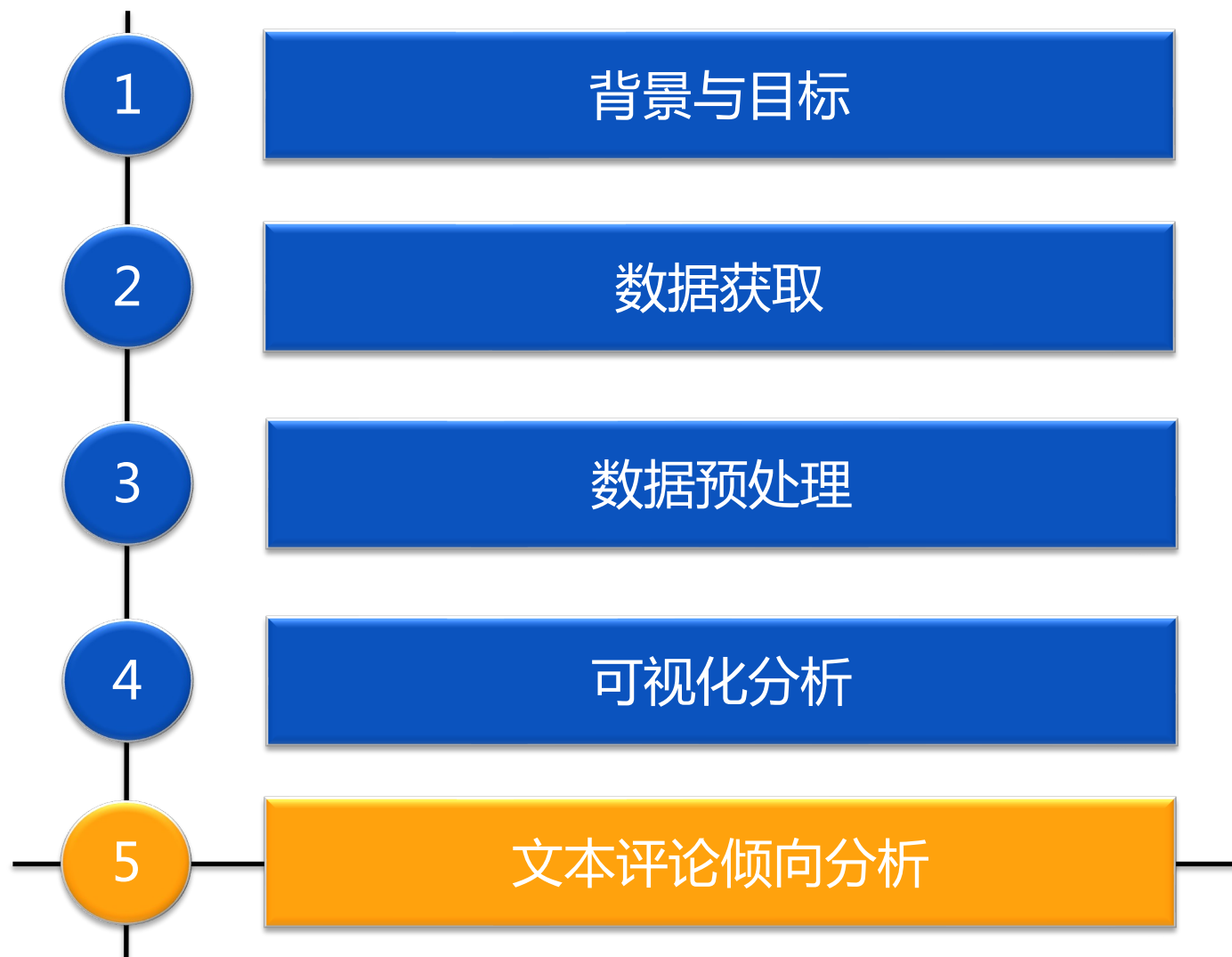
## 评论数据词云图

- “音效”、“拍照”、“运行速度”等功能体验良好
- “外观”、“外形”颜值认同
- “苹果”的品牌效应
- “屏幕”质量的肯定
- “速度”认可，运行速度，快递速度



# 目录

---





# 文本评论倾向分析

---

## 情感分类

- 评论的倾向附带情感，如何构建一个自动识别情感或评分的模型，是京东系统优化的一个方向。
- 目标：基于评论数据，构建一个文本评论分类模型。
  - 文本的向量表示
  - 构建分类模型

# 文本的向量表示

---

1. One-Hot表达：将每个词表示为一个长长的二元向量

- 文本1 : My dog ate my homework.
- 文本2 : My cat ate the sandwich.
- 文本3 : A dolphin ate the homework.

① 词袋：所有词的不重复构成




[a, ate, cat, dolphin, dog, homework, my, sandwich,  
the]

a: [1 0 0 0 0 0 0 0 0]

ate: [0 1 0 0 0 0 0 0 0]

.....

# 文本的向量表示

- 文本转化为词向量矩阵
- [a, ate, cat, dolphin, dog, homework, my, sandwich, the]
- 文本1: [0 1 0 0 1 1 1 0 0]  • 文本1 : My dog ate my homework.
- 文本2: [0 1 1 0 0 0 1 1 1]  • 文本2 : My cat ate the sandwich.
- 文本3: [1 1 0 1 0 1 0 0 1]  • 文本3 : A dolphin ate the homework.
- 缺陷: 忽略了句子词频信息

# 文本的向量表示

## 2. TF-IDF:

TF-IDF的主要思想是：如果某个词或短语在一篇文章中出现的频率TF高，并且在其他文章中很少出现，则为关键词，即此词或者短语具有很好的类别区分能力，适合用来分类。

### 中华蜜蜂养殖的方法

#### 1、选地建场

中华蜜蜂的饲养场适合建在远离公路、工厂、吵闹的地方，并且周围要有充足的水源，此外，中华蜜蜂对蜜源的要求不高，但也需要找到一个持续有蜜可以采集的地方。

#### 2、蜂群距离

中华蜜蜂的野性强，对方向的辨别能力较弱，容易迷巢，因此在摆放蜂群时，要将蜂群之间的距离拉远一些，尤其是交尾群，蜂群之间的距离要拉开得更大，以免处女王进错蜂群被杀害。

#### 3、喂养工具

中华蜂群的嗅觉灵敏，喂养时最好使用没有异味的木箱或者木桶，如果是新的木桶，要在里面涂抹少量的蜂蜡，将木质气味消除，有助于蜜蜂接受新巢的环境，避免逃走。

#### 4、越夏管理

在夏季天气炎热时，要将蜂箱转移到树荫下或者屋檐下，中午朝蜂箱的周围喷洒水分，增加空气湿度，还可以扩大巢门，降低温度，当温度超过35℃时，要给中华蜜蜂喂水。

"词频" (Term Frequency, 缩写为TF)

"蜜蜂"、"养殖" → 关键词

的”、“是”、“在”，“时” → ~~常见词~~ → 关键词

"逆文档频率" (Inverse Document Frequency, 缩写为IDF)

# 文本的向量表示

## ➤ 第一步：计算词频

$$TF = N / M$$

$N$ : 单词在某文档中的频次  
 $M$ : 该文档的单词数

## ➤ 第二步，计算逆文档频率。

$$IDF = \log\left(\frac{D}{D_w}\right)$$

$D$ : 总文档数  
 $D_w$ : 出现了该单词的文档数

如果一个词越常见，那么分母就越大，逆文档频率就越小越接近0。

# 文本的向量表示

- 为了避免分母为0（即所有文档都不包含该词），通常在分母加1。

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

- 第三步，计算TF-IDF

TF-IDF与一个词在文档中的出现次数成正比，与该词在整个语言中的出现次数成反比。计算出文档的每个词的TF-IDF值，然后按降序排列，取排在最前面的几个词。

$$TF - IDF = TF \times IDF$$

# 文本的向量表示

---

## 1. 计算TF

- 文本1:  $[0 \ 1 \ 0 \ 0 \ 1 \ 1 \ 2 \ 0 \ 0]$  “my” 在句子中出现了2次
- 文本2:  $[0 \ 1 \ 1 \ 0 \ 0 \ 0 \ 1 \ 1 \ 1]$
- 文本3:  $[1 \ 1 \ 0 \ 1 \ 0 \ 1 \ 0 \ 0 \ 1]$

- 文本1 : My dog ate my homework.
- 文本2 : My cat ate the sandwich.
- 文本3 : A dolphin ate the homework.

文档TF信息：相当于归一化，避免了句子长度不一致问题。

- 文本1:  $[0 \ 1/5 \ 0 \ 0 \ 1/5 \ 1/5 \ 2/5 \ 0 \ 0]$  “my” 在句子中出现了2次
- 文本2:  $[0 \ 1/5 \ 1/5 \ 0 \ 0 \ 0 \ 1/5 \ 1/5 \ 1/5]$
- 文本3:  $[1/5 \ 1/5 \ 0 \ 1/5 \ 0 \ 1/5 \ 0 \ 0 \ 1/5]$

# 文本的向量表示

## 2. 计算IDF

- 文本1 : My dog ate my homework.
- 文本2 : My cat ate the sandwich.
- 文本3 : A dolphin ate the homework.

a(1), ate(3), cat(1), dolphin(1), dog(1), homework(2), my(3), sandwich(1), the(2)

$$\text{逆文档频率(IDF)} = \log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数} + 1}\right)$$

a:  $\log(3/2)$ ,    ate:  $\log(3/4)$ ,    cat:  $\log(3/2)$ ,    dolphin:  $\log(3/2)$ ,    dog:  $\log(3/2)$ ,

homework:  $\log(3/3)$ ,    my:  $\log(3/3)$ ,    sandwich:  $\log(3/3)$ ,    the:  $\log(3/3)$



# 文本的向量表示

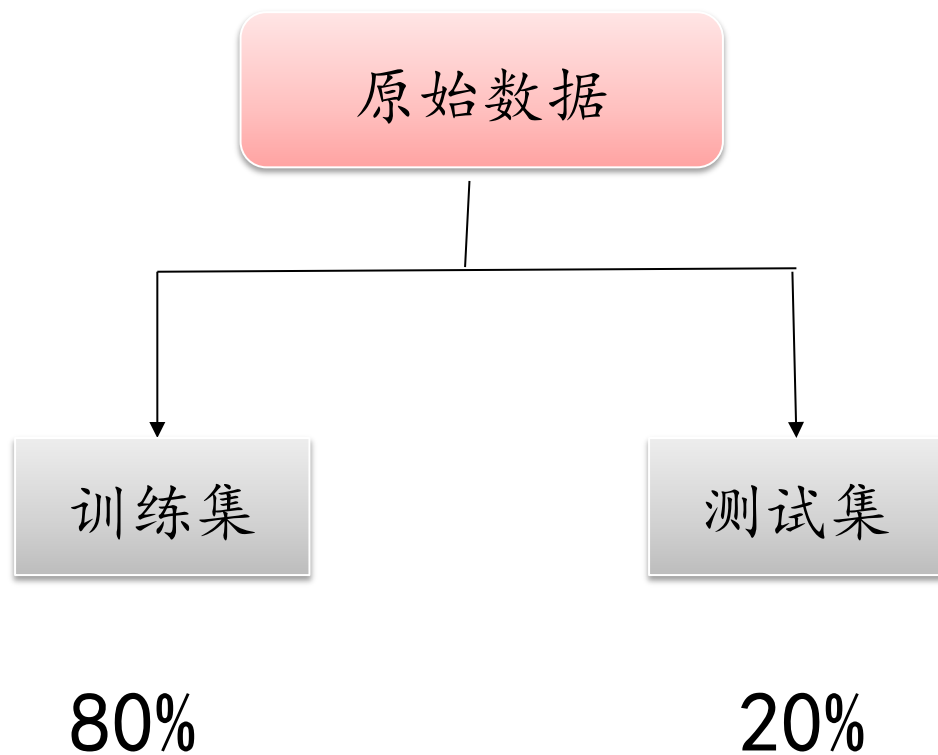
## TF-IDF权值向量

1. 'My dog has flea problems, help please.' .....> [0.,0.,0.,0.,0.,0.,0.27912828,0.40318254,0.,...]
  2. 'Maybe not take him to dog park is stupid.' .....> [0.,0.,0.,0.,0.,0.,0.25617597,0.,0.,0.,0.,0.,...]
  3. 'My dalmation is so cute. I love him my.' .....> [0.,0.3240.,0.57964,0.,0.,0.,0.27912828,0.,...]
  4. 'Stop posting stupid worthless garbage.' .....> [0.,0.,0.,0.,0.,0.,0.25617597,0.,0.,0.,0.,0.,...]
  5. 'Mr licks ate my steak, what can I do?.' .....> [0.,0.,0.,0.,0.,0.,0.27912828,0.40318254,0.,...]
  6. 'Quit buying worthless dog food stupid' .....> [0.,0.,0.,0.,0.,0.,0.25617597,0.,0.,0.,0.,0.,...]
- labels = [0, 0, 1, 1, 0, 1]      #文档标签: 是否是消极情感

# 建模准备

---

训练集、测试集划分



# 朴素贝叶斯

引例: 如果一对男女朋友, 男生想女生求婚, 男生的四个特点分别是不帅, 性格不好, 身高矮, 不上进, 请你判断一下女生是嫁还是不嫁?

帅?	性格好?	身高?	上进?	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

分类  
问题

$p(\text{嫁} | (\text{不帅}, \text{性格不好}, \text{身高矮}, \text{不上进})) > p(\text{不嫁} | (\text{不帅}, \text{性格不好}, \text{身高矮}, \text{不上进}))$

$p(\text{嫁} | (\text{不帅}, \text{性格不好}, \text{身高矮}, \text{不上进})) \leq p(\text{不嫁} | (\text{不帅}, \text{性格不好}, \text{身高矮}, \text{不上进}))$

# 朴素贝叶斯

贝叶斯定理

嫁不嫁?

贝叶斯公式:

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

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

$$\xrightarrow{\text{red dashed arrow}} P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

当A与B相互独立时:

$$P(AB) = P(A)P(B)$$



$$p(\text{嫁} | \text{不帅、性格不好、身高矮、不上进}) = \frac{p(\text{不帅、性格不好、身高矮、不上进} | \text{嫁}) * p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

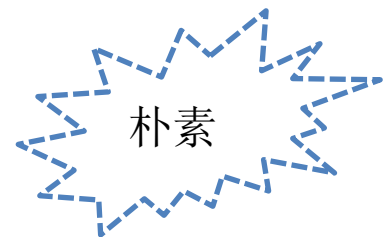
$$p(\text{不嫁} | \text{不帅、性格不好、身高矮、不上进}) = \frac{p(\text{不帅、性格不好、身高矮、不上进} | \text{不嫁}) * p(\text{不嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

# 朴素贝叶斯

$$p(\text{嫁} | \text{不帅、性格不好、身高矮、不上进}) = \frac{p(\text{不帅、性格不好、身高矮、不上进} | \text{嫁}) * p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})}$$

基于独立性假设：

$$P(\text{不帅、性格不好、身高矮、不上进} | \text{嫁}) = P(\text{不帅} | \text{嫁}) * p(\text{性格不好} | \text{嫁}) * P(\text{身高矮} | \text{嫁}) * P(\text{不上进} | \text{嫁})$$



# 朴素贝叶斯

$$\frac{p(\text{嫁}|\text{帅, 性格好, 身高矮, 不上进})}{p(\text{嫁}|\text{帅, 性格好, 身高矮, 不上进}) + p(\text{不嫁}|\text{帅, 性格好, 身高矮, 不上进})}$$
  
$$p(\text{嫁}|\text{帅, 性格好, 身高矮, 不上进}) = p(\text{帅}) * p(\text{性格好}) * p(\text{身高矮}) * p(\text{不上进}) * p(\text{嫁}|\text{帅, 性格好, 身高矮, 不上进})$$

帅？	性格好？	身高？	上进？	嫁与否
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

$$P(\text{嫁}) = P(\text{不嫁}) = 1/2$$

$$p(\text{不帅}|\text{嫁}) = 3/6 = 1/2$$

$$p(\text{性格不好}|\text{嫁}) = 1/6$$

$$p(\text{矮}|\text{嫁}) = 1/6$$

$$p(\text{不上进}|\text{嫁}) = 1/6$$

$$P(\text{不帅}|\text{不嫁}) = 1/6$$

$$P(\text{性格不好}|\text{不嫁}) = 3/6 = 1/2$$

$$P(\text{矮}|\text{不嫁}) = 6/6 = 1$$

$$P(\text{不上进}|\text{不嫁}) = 3/6 = 1/2$$

# 朴素贝叶斯

## 拉普拉斯平滑处理

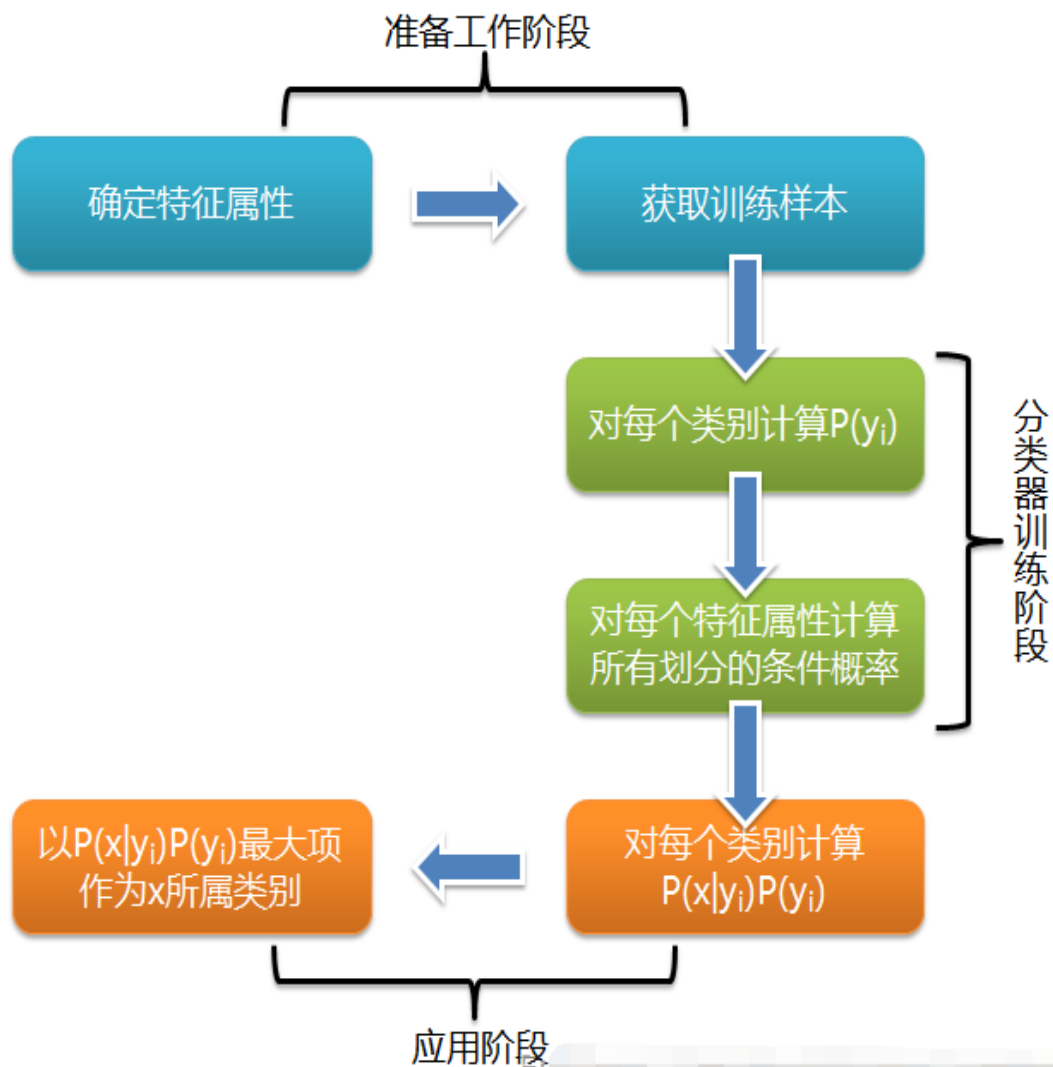
- 缺陷：受样本个数限制，若某个属性值在训练集中没有与某个同类同时出现过
- 如  $p(\text{不帅}) * p(\text{性格不好}) * p(\text{身高矮}) * p(\text{不上进}) = 0$ ，则出现分母为零。
- 修正方法：拉普拉斯平滑处理

$$\begin{array}{ccc} P(y) = \frac{|D_y|}{|D|} & \xrightarrow{\text{red dashed arrow}} & \hat{P}(y) = \frac{|D_y| + 1}{|D| + N} \\ P(x|y) = \frac{|D_{y,x}|}{|D_c|} & & \hat{P}(x|y) = \frac{|D_{y,x}| + 1}{|D_c| + N_i} \end{array}$$

- $N$  表示训练集样本的类别数， $N_i$  表示训练集样本在第  $i$  个属性上的取值个数

# 朴素贝叶斯

## 算法处理流程







# Thank you!