达观数据
DATA GRAND

文本分析和弹幕审核技术介绍

达观数据 陈运文

- 达观数据 创始人&CEO
- 曾担任盛大文学首席数据官、阅文集团数据中心负责人、腾讯文学高级总监、百度核心技术研发专家
- 复旦大学计算机系博士和杰出毕业生

达观 提供专业的数据技术服务

达观数据成立于2015年，位于上海市张江高科，是上海重点扶持的高科技创新企业，也是著名投资机构真格基金旗下企业

达观数据拥有领先的人工智能、机器学习技术，能自动挖掘数据隐藏的规律，识别文字的语义内容，并进行信息的抓取、搜索、推荐等专业技术服务。

达观核心团队来自腾讯、盛大、阿里、百度等国内一线互联网企业数据部门，具有丰富的研发经验和众多成功应用案例

达观数据
DATA GRAND

ZhenFund
真格基金

直播弹幕：新形式的互动娱乐方式

自然语言处理是文本挖掘的基础

自然语言处理（Natural Language Processing，NLP）是计算机科学领域与人工智能领域中的一个重要方向

它研究能实现人与计算机之间用自然语言进行有效通信的各种理论和方法，能够利用计算机为工具对人类特有的书面形式和口头形式的语言进行各种类型处理和加工的技术。

基础知识

概率，最大似然估计，条件概率，贝叶斯法则，二项式分布，联合概率分布和条件概率分布等知识是nlp研究的基础

信息熵，又称为自信息（self-information），描述一个随机变量的不确定性的数量。一个随机变量的熵越大，它的不确定性越大，正确估计其值的可能性越小，越不确定的随机变量越需要更大的信息量用以确定其值。

$$H(X) = -\sum_{i=1}^{n} P_i \times \log_2 P_i$$

如英语有26个字母，假如每个字母在文章中出现次数平均的话，每个字母的信息量为4.7。而汉字常用的有2500个，假如每个汉字在文章中出现次数平均的话，每个汉字的信息量为11.3。

语言模型

语言模型（language mode）在基于统计模型的语音识别，机器翻译，汉语自动分词和句法分析中有着广泛的应用。

一个语言模型构建字符串的概率分布p(W)，假设p(W)是字符串作为句子的概率

n元语法模型：根据马尔科夫假设，一个词只和他前面n-1个词相关性最高，则概率由下边的公式计算：

$$P(w_1 w_2 \dots w_n) = \prod_i P(w_i \mid w_1 w_2 \dots w_{i-1})$$

中文分词的主要问题

歧义切分：分词后的结果和原来语句所要表达的意思不相符或差别较大，在机械切分中比较常见。

例子：结婚的和尚未结婚的人

正确：结婚/的/和/尚未/结婚/的/人

错误：结婚/的/和尚/未/结婚/的/人

未登录词：指的是词没有在词典中出现，比如一些新的网络词汇：“网红”，“走你”；一些未登录的人名，地名；一些外语音译过来的词等等。简单的case可以通过加词典解决，但是随着字典的增大，可能会引入新的bad case，并且系统的运算复杂度也会增加。

基于词典的机械切分分词方法

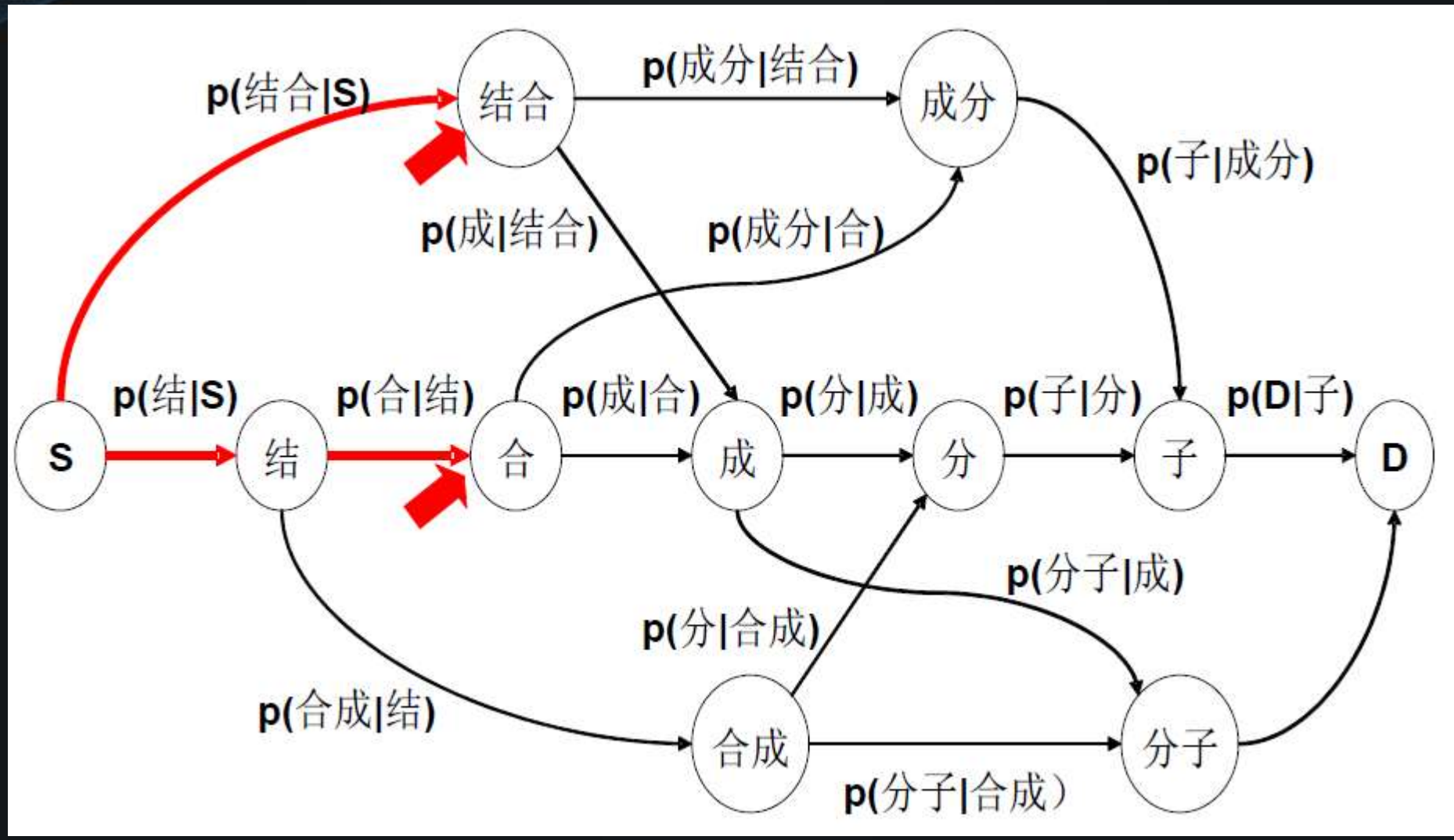
原理：本质上就是字符串匹配的方法，将一串文本中的文字片段和已有的词典进行匹配，如果匹配到，则此文字片段就作为一个分词结果。

常见方法：匹配法（正向最大匹配方法，逆向最大匹配法，双向最少切分法）；全切分路径选择法（n最短路径方法，n元语法模型法）。

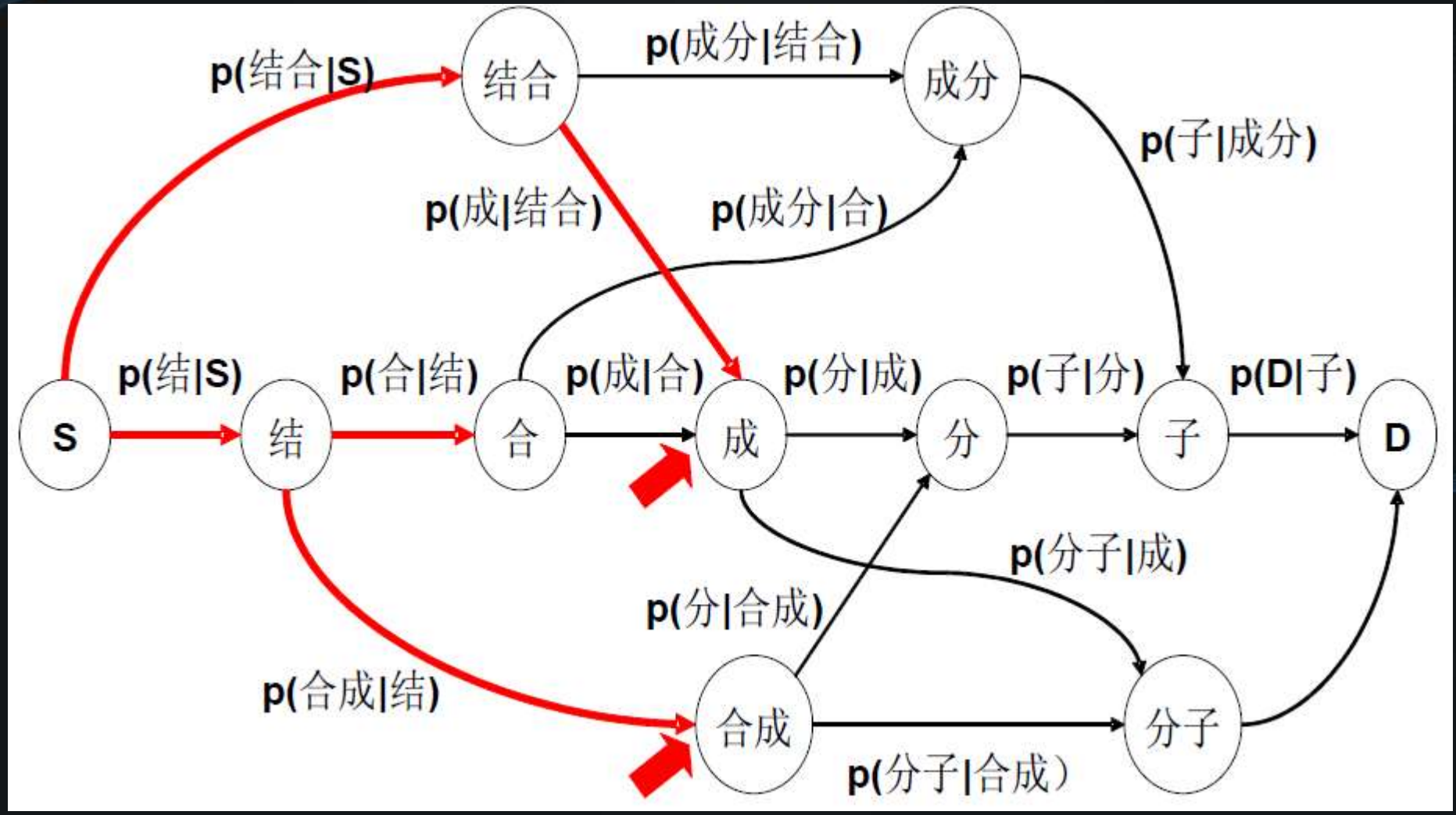
n最短路径方法：将所有的切分结果组成有向无环图，每个切词结果作为一个节点，词之间的边赋予一个权重，最终找到权重和最小的一条路径作为分词结果。

n元语法模型法：根据n元语法模型，路径构成时会考虑词的上下文关系，根据语料库的统计结果，找出构成句子最大模型概率。一般情况下，使用unigram和bigram的n元语法模型的情况较多。

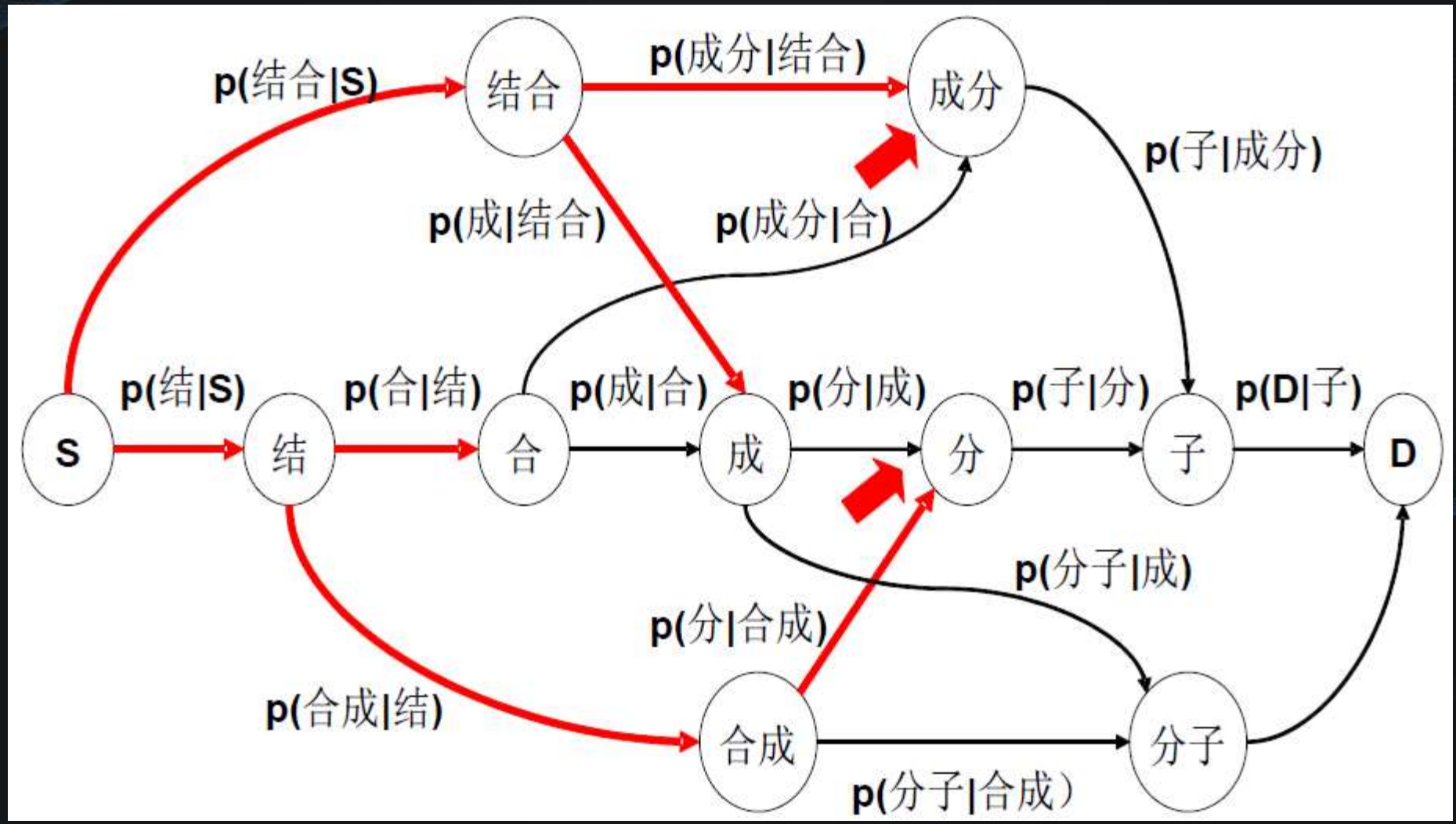
基于n元语法模型的分词方法



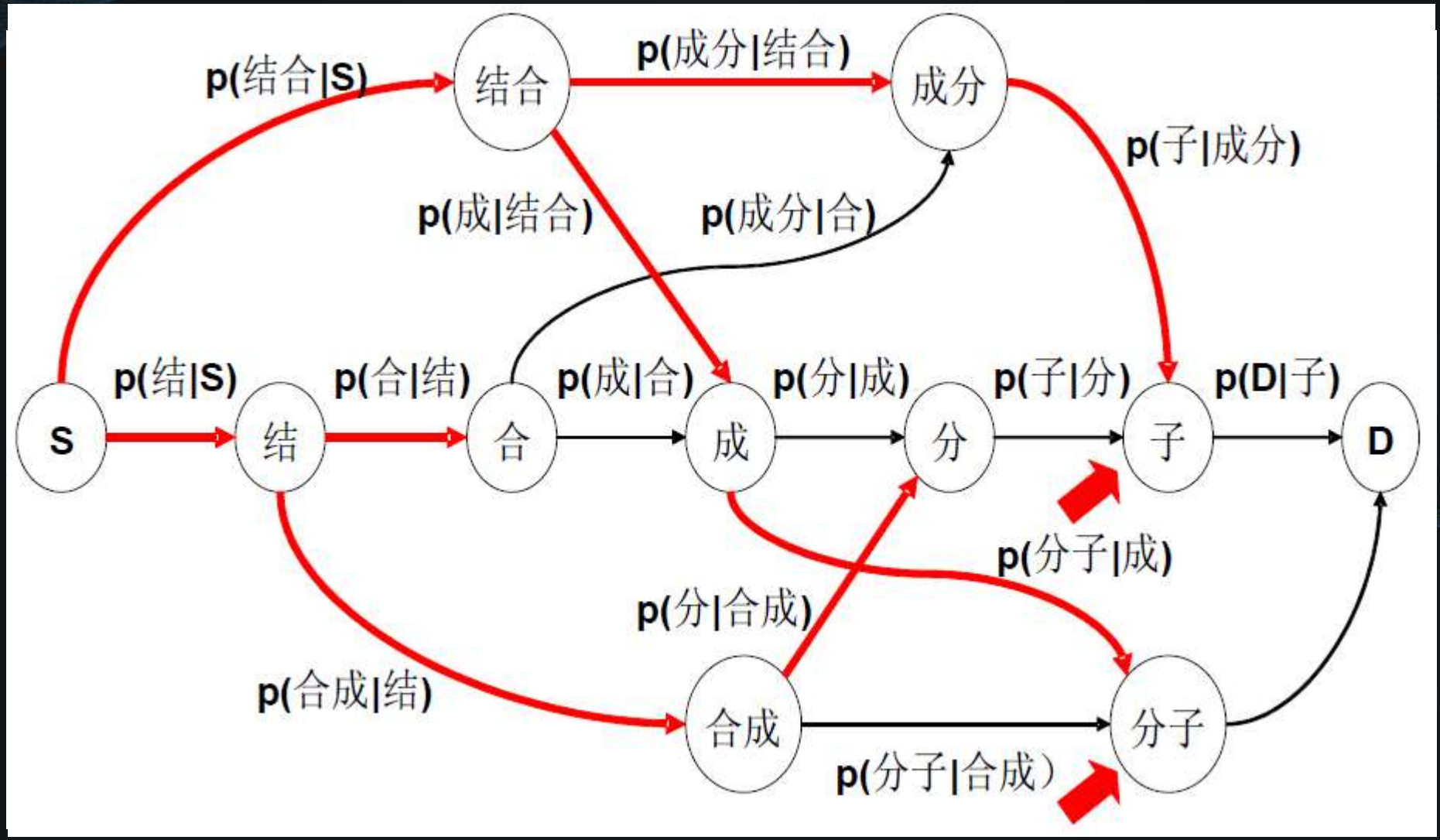
基于n元语法模型的分词方法



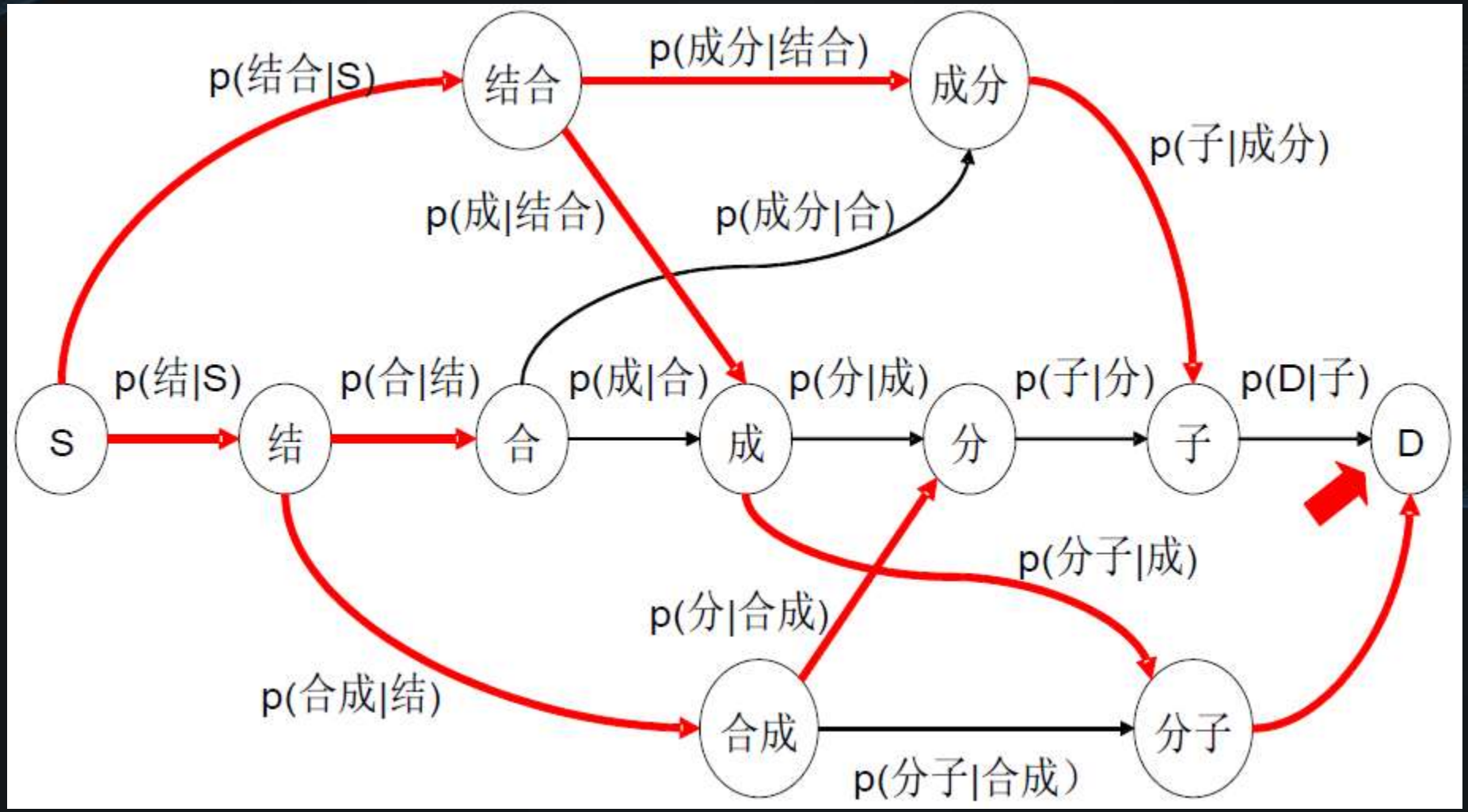
基于n元语法模型的分词方法



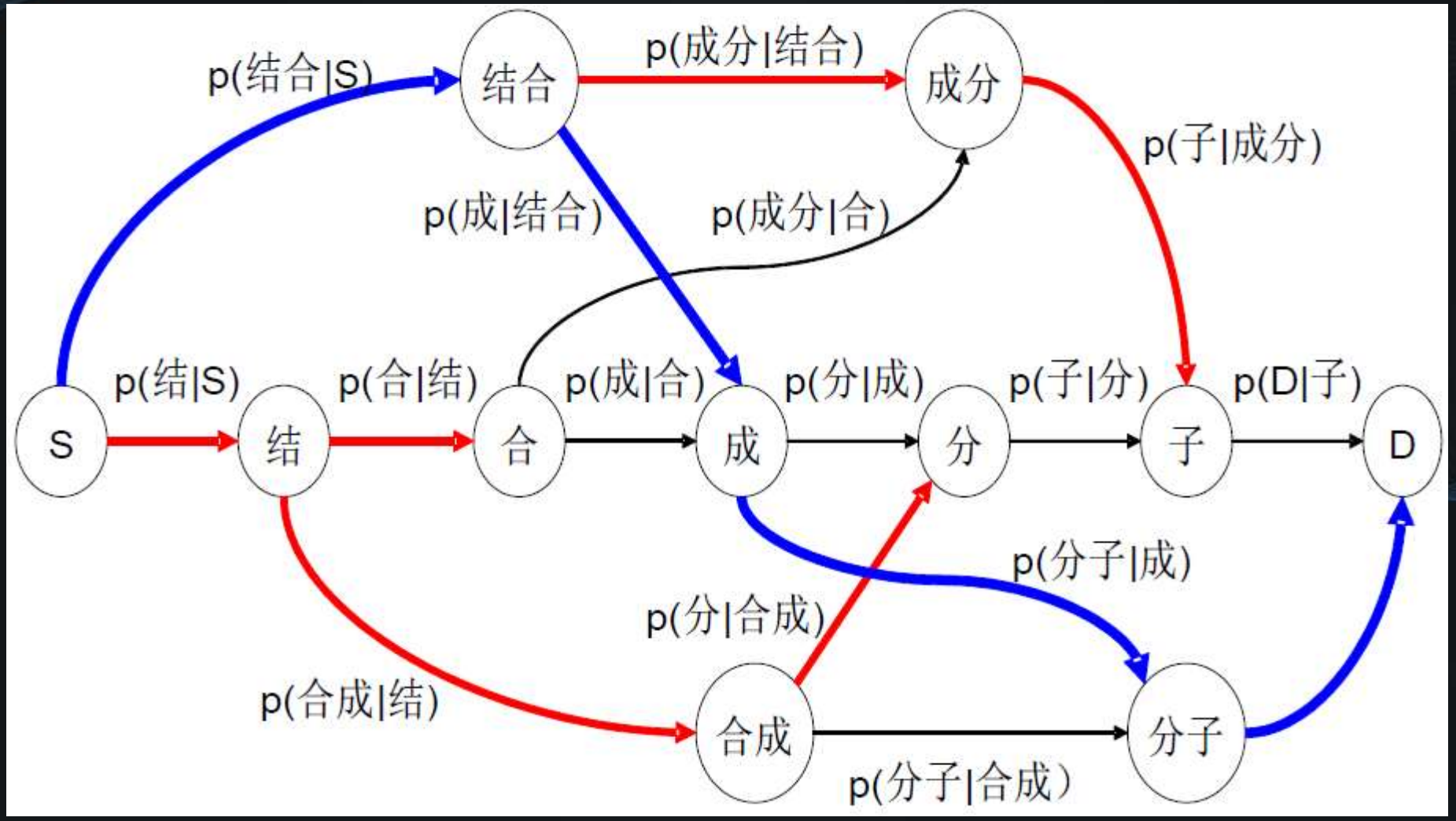
基于n元语法模型的分词方法



基于n元语法模型的分词方法



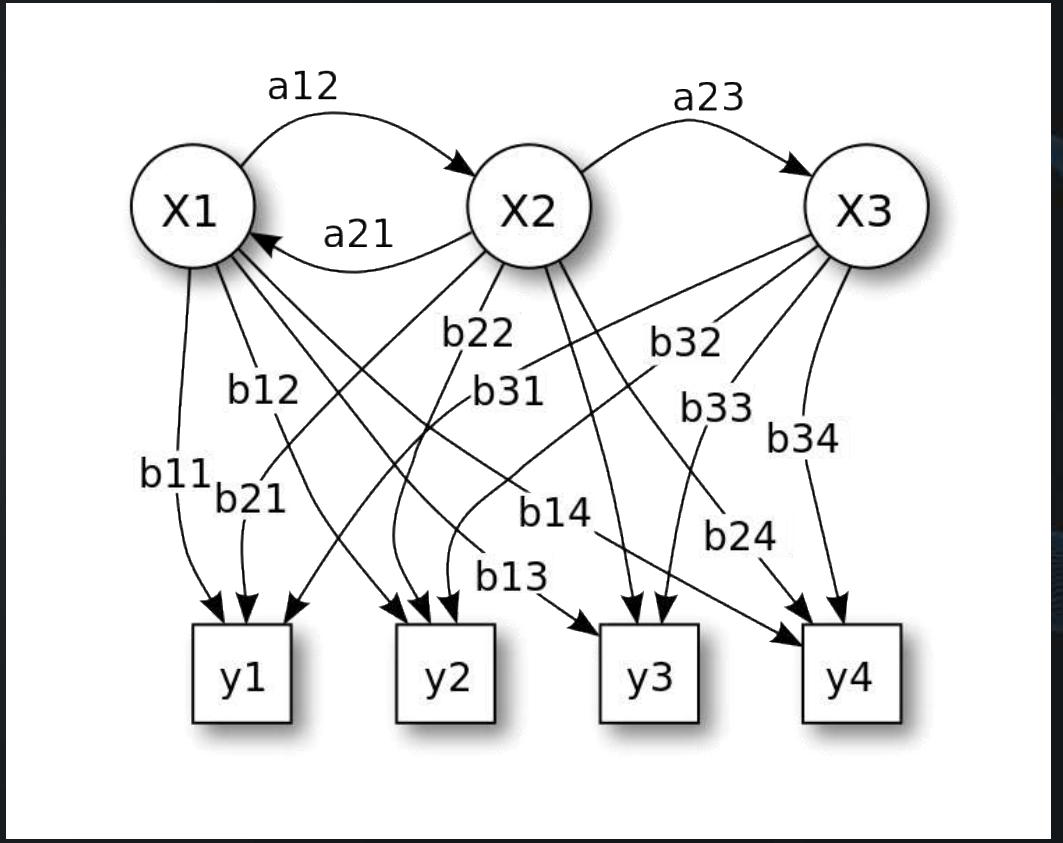
基于n元语法模型的分词方法



序列标注问题的常见模型HMM和CRF

HMM（Hidden Markov Model）

隐马尔科夫模型，基本的思想就是根据观测值序列找到真正的隐藏状态值序列。在中文分词中，一段文字的每个字符可以看作是一个观测值，而这个字符的词位置label（BEMS）可以看作是隐藏的状态。



基于序列标注的分词方法

将文本中每个字按在词中的位置进行标注，常用BMES标记

B，Begin，表示这个字是一个词的首字

M，Middle，表示这是一个词中间的字

E，End，表示这是一个词的尾字

S，Single，表示这是单字成词

分词的过程就是将一段字符输入模型，然后得到相应的标记序列，再根据标记序列进行分词。

输入：达观数据是企业大数据服务商

输出：BMMESBEBMEBME

结果：达观数据/是/企业/大数据/服务商

词汇级其他相关工作

词性标注：是指对于句子中的每个词都指派一个合适的词性，也就是要确定每个词是名词、动词、形容词或其他词性的过程，又称词类标注或者简称标注。词性标注的主要问题是标注歧义问题，比如“book”，名词还是动词？

命名实体识别：Named Entity Recognition，简称NER，又称作“专名识别”，是指识别文本中具有特定意义的实体，主要包括人名、地名、机构名、专有名词等。通常包括实体边界识别和确定实体类别。

篇章级应用：文本审核

文本审核：在预定义的审核体系下，根据文本的特征（内容或属性），将给定文本与一个或多个类别相关联的过程。

常见方法：根据分类知识获取方法的不同，文本自动分类系统大致可分为基于知识工程和基于机器学习的分类系统，目前大部分都是基于机器学习的分类系统。

机器学习分类方法：一般包括两个过程：训练阶段和预测阶段

文本语义处理面临的困难

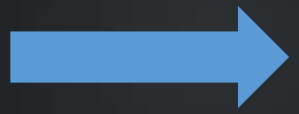
语义层次：喜欢乡下的孩子，关于鲁迅的著作

未知新词：杀马特，高富帅，约炮，微信，欧巴，酱紫，弓虽，毒趴，萌萌哒，么么哒， 美刀，作死，逼格....

未知词义：兰州烧饼，八的平方，小强，炮灰，油菜花，捉鸡，山寨，脱光，上海西南某高校...

直播弹幕审核：异形文本变换

以前我总是买了跌，卖了涨，
入市不久就亏了15万，后来认
识了李老师，就一直跟着他做，
我从10万做到25万，他的微信
《y411580》,qq
《869456317》，需要的可以
加下，也许马上就能改变你在
股市的命运

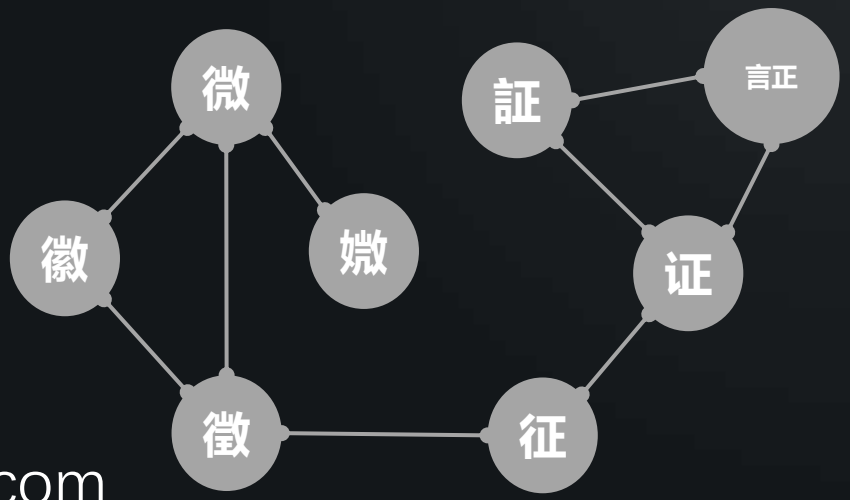


以前我總是買了跌，賣了漲，
入市不久就虧了15萬，后來認
識了李老師，就壹直跟着他做，
我從十萬做到二十五萬，他的
溦《y4.①①㈤8零》,☆釦釦☆
《86九4㈤6Ɔ㈠⑦》，懦要的
可以伽芊，也許馬上就能改變
妳茬股sんi葯命運﹀ㄣ

机器学习的基础：特征抽取

- 自动化生成变形词词库

种子关键词：
{微，证，…}



关键词	变形词
微信	徽亻言 、徵信、嫩亻言
假证	叚証、假言正

关键字	变形词
微	徽、徵、嫩
证	証、言正
…	……

特征提取

文本分类问题中，词语作为主要的特征。但是对于一个分类器来说，并不是特征越多越好。特征越多不仅会导致计算复杂度增加，训练和预测时间加长，还会代入一些噪音，导致模型效果差。因此特征选择是必须也是有效的。

文本特征提取有很多方法，常见的包括基于文档频率（Document Frequency，DF），信息增益（Information Gain，IG），卡方统计，互信息等多种方法。

特征权重计算方法

布尔权重：出现为1，否则为0

绝对词频：tf，特征项在文本中出现的频度

倒排文档频度：idf，稀有特征项比常用特征项含有更多的信息

TF-IDF：tf*idf，特征与在文档中出现的频度成正比，与整个语料中出现的该特征项的文档数成反比

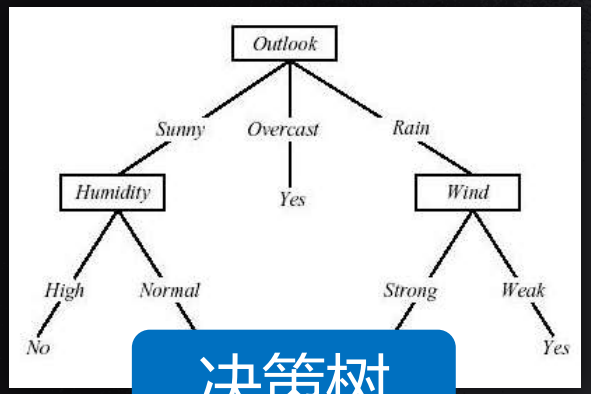
TFC：对文本长度进行归一化处理后的TF-IDF

ITC：在TFC基础上，对tf的对数值取代tf

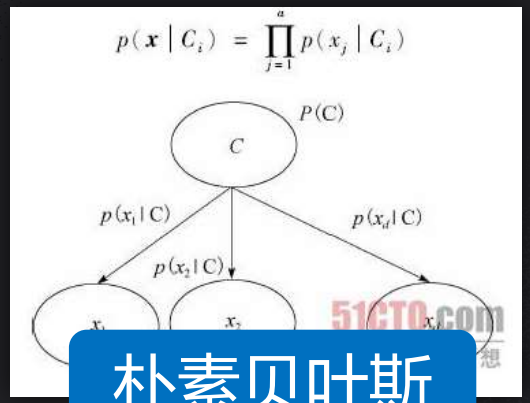
熵权重：建立在信息论基础上

TF-IWF：在TF-IDF基础上，用特征项频率倒数的对数值IWF代替IDF，并用IWF的平方平衡权重值对于特征项频度的倚重

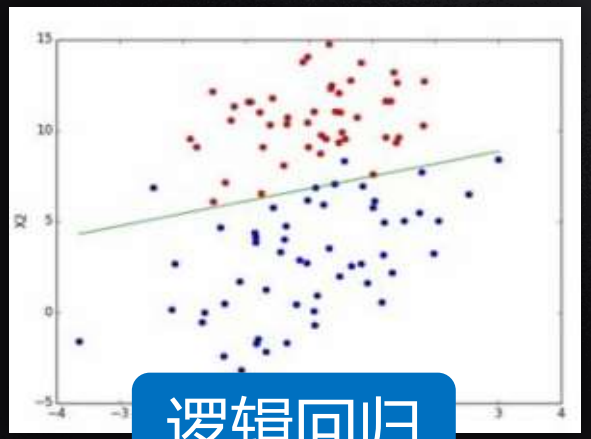
常用的文本分类模型



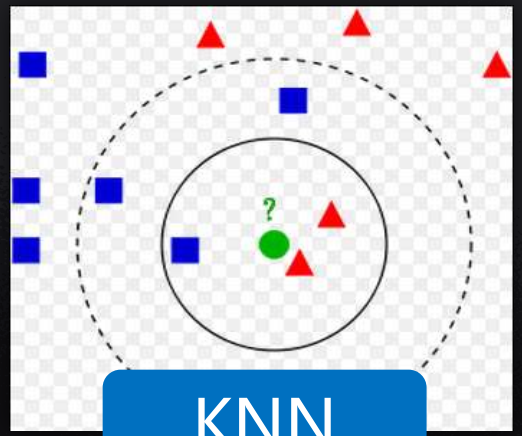
决策树



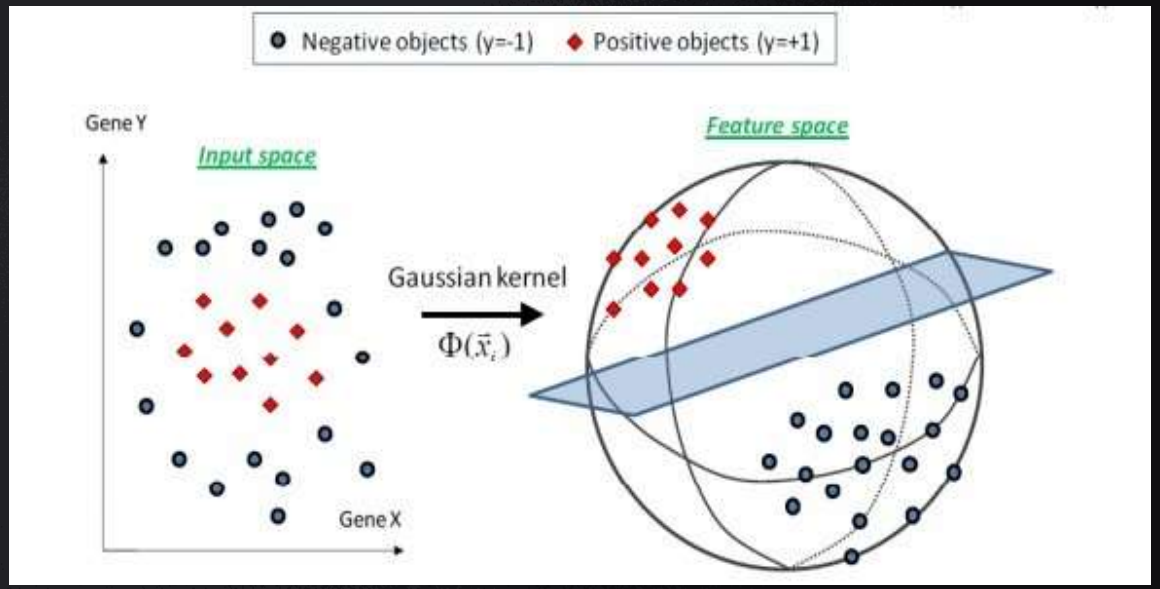
朴素贝叶斯

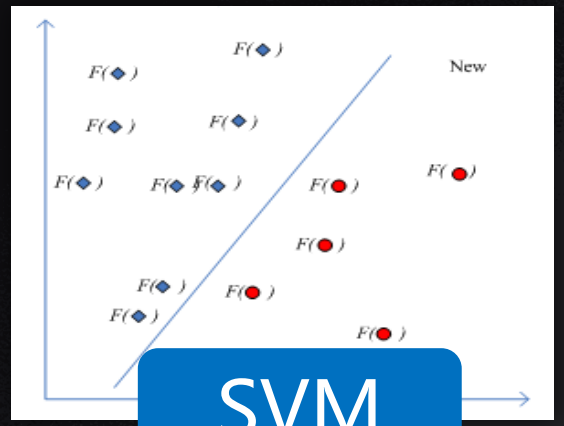


逻辑回归

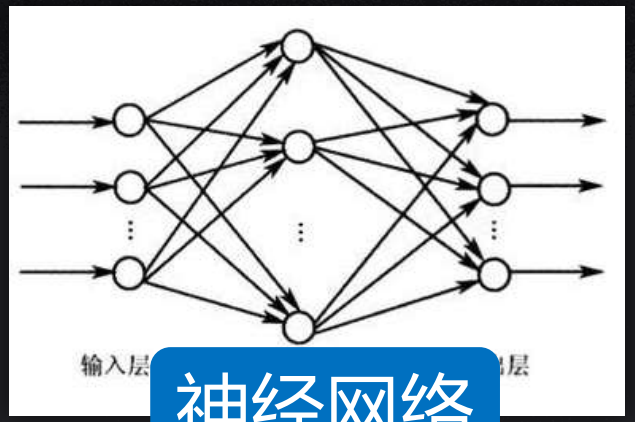


KNN



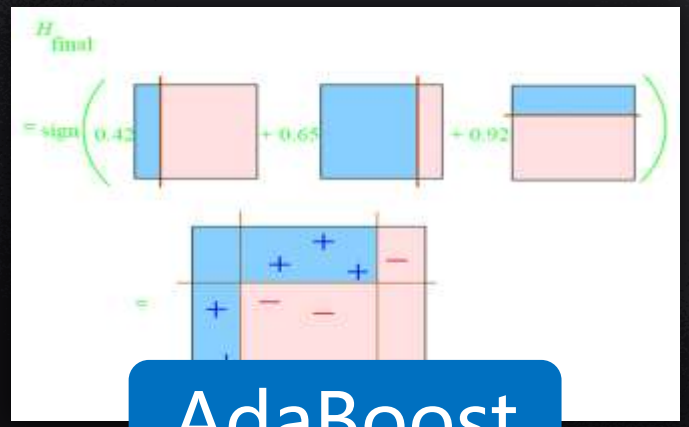


SVM

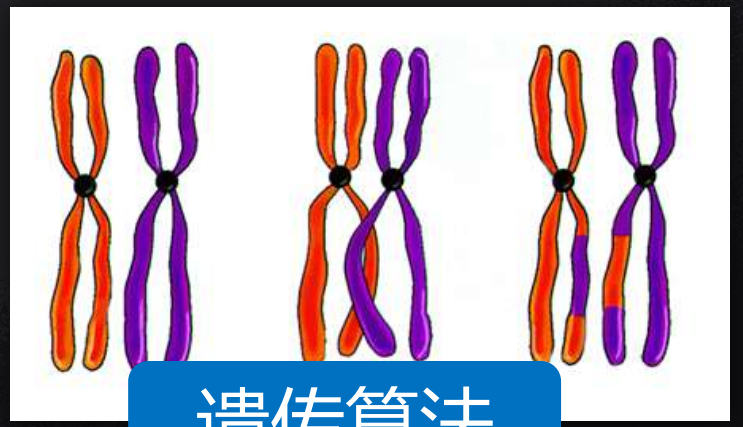


神经网络

随机森林



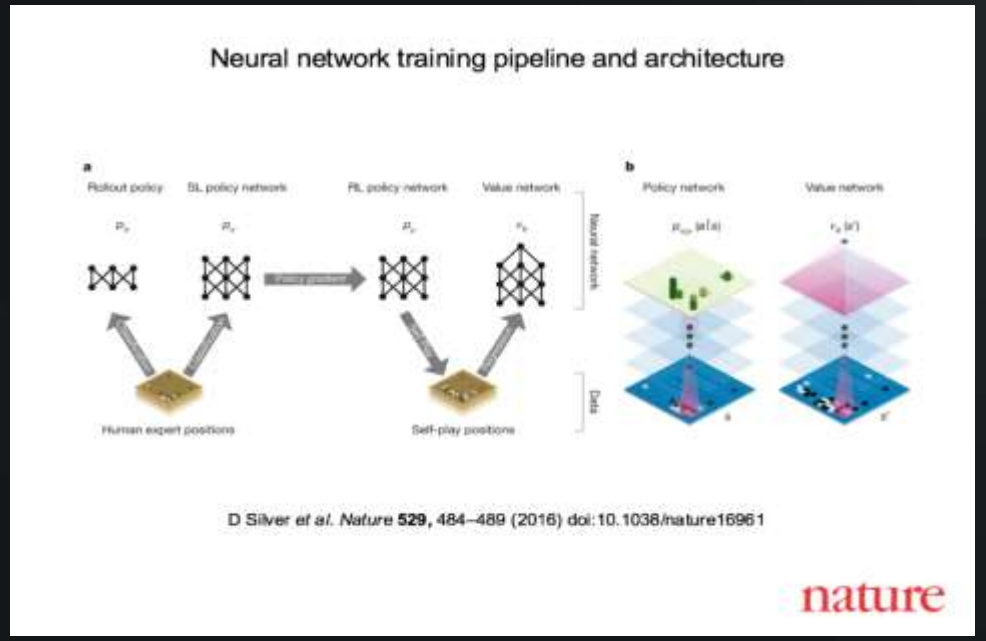
AdaBoost



遗传算法

深度学习开始影响NLP领域

Deep Learning（深度学习）的NLP运用是近年来的研究热点

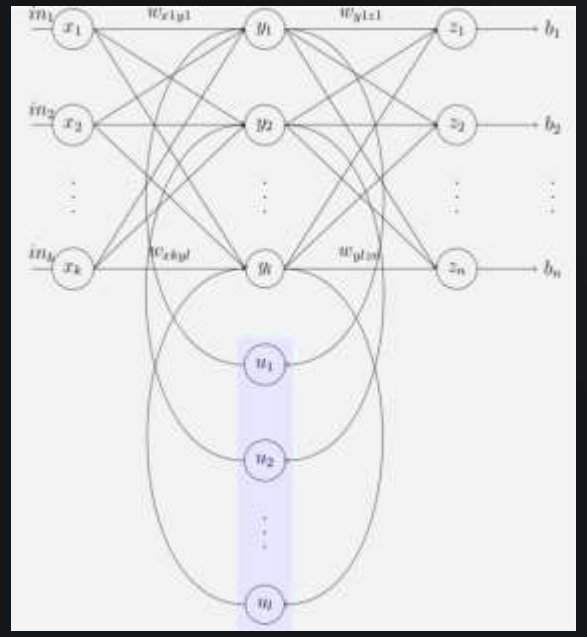


Word2Vec由三层神经网络构成，将词映射为向量，方便的找到同义词或联系紧密的词，扩展词汇特征

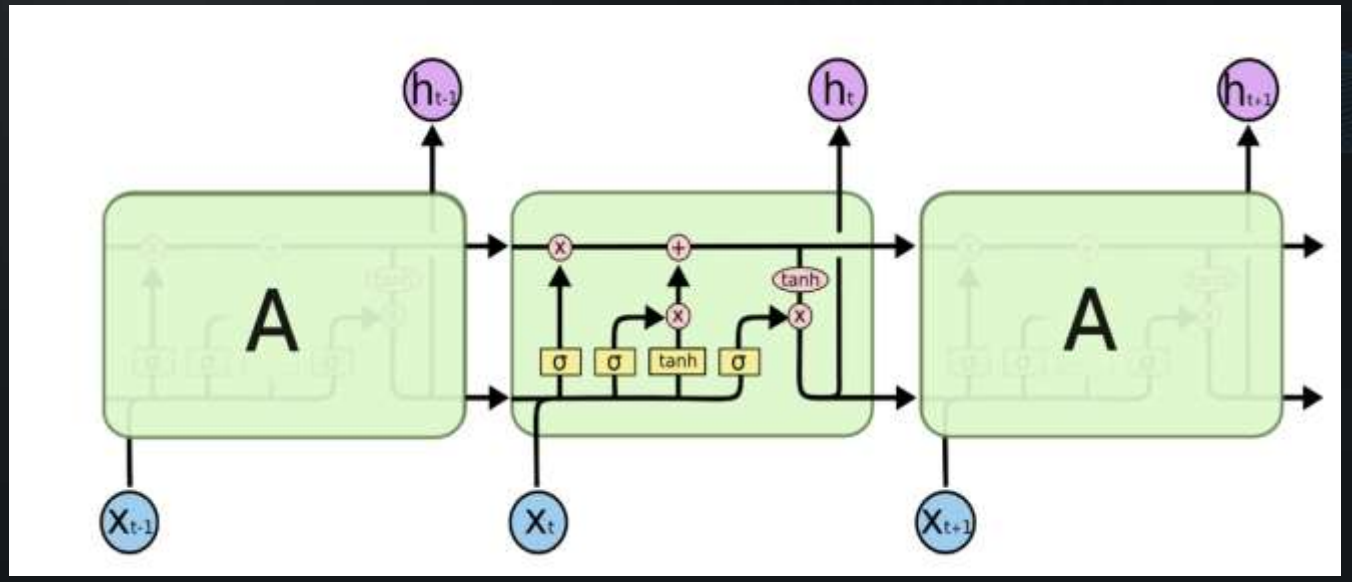
```
2. gaoxiang@iZ2547mxdm8Z:~/datagrand/siterec/code/rpc_services/...
(env_dl) [gaoxiang@iZ2547mxdm8Z word2vec]$ python client.py 编程
java 0.771900594234
程序开发 0.762071967125
编程语言 0.752881407738
python 0.752307057381
程序设计 0.743839502335
程序语言 0.735327124596
c++ 0.713602125645
rubyonrails 0.710829496384
javascript 0.7074457407
计算机 0.69321423769
(env_dl) [gaoxiang@iZ2547mxdm8Z word2vec]$
```

NLP新模型：RNN网络和LSTM模型

考虑到上下文的RNN（Recurrent Neural Networks，循环神经网络）将隐藏层内部的节点也连接起来，增加了网络前一时刻的输出对当前输出的影响



LSTM（Long Short-Term Memory，长短时记忆）在RNN基础上，增加了远距离的上下文依赖，能够存储较远距离上下文对当前时间节点的影响。



弹幕文本调用API和统计分析

自动识别垃圾广告、色情政治、暴力辱骂等文本，降低运营风险

相应时间<0.1秒，支持超高并发

示例

CURL调用示例：

```
curl -i -X POST -d 'appid=12345&text=3m每天百分之1利息，60元起步，有需要可以联系我，Q49663537，或者关注百度贴吧，老马平台吧！&title=顶顶顶&textid=435386945382932&user=巧儿' 'http://commentaggreapi.datagrand.com/commentaggre/datagrand'
```

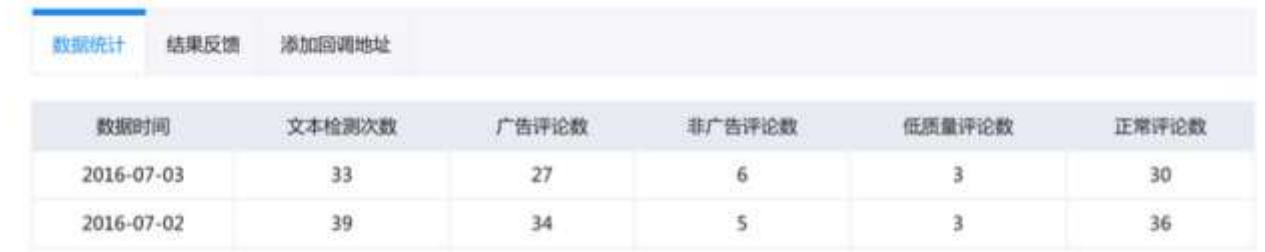
成功返回示例:

```
{
    "status": "OK",
    "result": {
        "weight_ad": 0.696227989117567,
        "reaction": 0.0096,
        "porn": 0.1469,
        "is_ad": 1,
        "politic": 0.0096
    },
    "request_id": "1470899110107282"
}
```

数据统计和结果反馈

数据统计和结果反馈目前仅限于文章自动归类、文章自动审核和垃圾评论自动过滤三种服务中。

在上述操作正确执行，服务运行正常情况下，可以登录到系统后台查看统计数据。



数据时间	文本检测次数	广告评论数	非广告评论数	低质量评论数	正常评论数
2016-07-03	33	27	6	3	30
2016-07-02	39	34	5	3	36

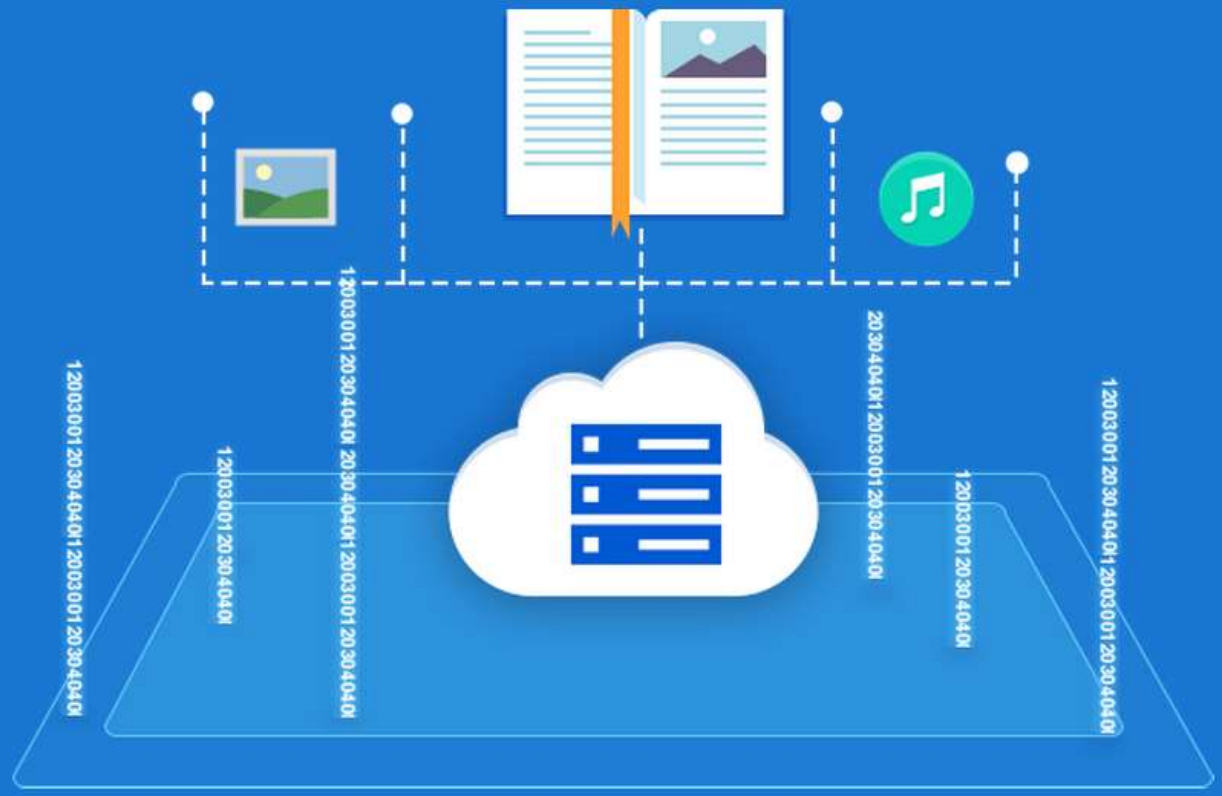
同时，用户可在系统后台查看到并且进行结果反馈操作。

文本挖掘

从海量文本中精炼高质量信息

高精确度的挖掘结果，毫秒级的处理速度，高度稳定的云服务，简单至极的使用方法。达观自动化的文本挖掘服务为企业节省大量人力物力，提升经济效益。

立即体验



文本分类应用：文本情感分析

准确分析文本中的情感倾向，帮助应用方把握用户好恶，及时进行调整优化

文本情感分析

请输入一段需要分析的文字

刚失业，钱包就被偷，又要交房租了，没有比我更倒霉悲催的人了吧。

换一换　　提交文本

分析结果

情感倾向	权重
正面	0.0096
负面	0.9904

达观文本情感倾向分析功能可针对一段文本分析出其表达的是正面情绪还是负面情绪，以及情感倾向的程度。

情感微弱（0 - 0.5），情感一般（0.5 – 0.75），情感强烈（0.75 - 1）

文本分类应用：文本自动归类

依据预设的分类体系对文本进行自动归类，帮助高效管理和使用海量文本数据

文本自动归类

请输入一段文字

今天是世界卫生日，今年世界卫生日的关注重点为“应对糖尿病”。世卫组织报告指出，全球糖尿病患者人数已超过4亿人，而中国糖尿病的发病率正在呈现“爆炸式”增长，专家呼吁加强预防管理，控制糖尿病发病率逐年升高的趋势。

换一换　　提交文本

归类结果

类别	权重
社会资讯	0.605
健康	0.5
科技	0.203

达观文本自动归类服务可对文本内容进行分析，给出文本所属的类别和置信度。

非常相关（0.85 – 1.0），一般相关（0.5 – 0.85），少量相关（0 – 0.5）

该demo依照新榜公众号类别进行归类，若您有定制化需求，请联系我们。


达观数据
DATA GRAND
QINIU

达观数据
DATA GRAND

联系方式：400-175-9889

Thank you