第12章 文本的分类与聚类

宗 成 庆

中国科学院自动化研究所

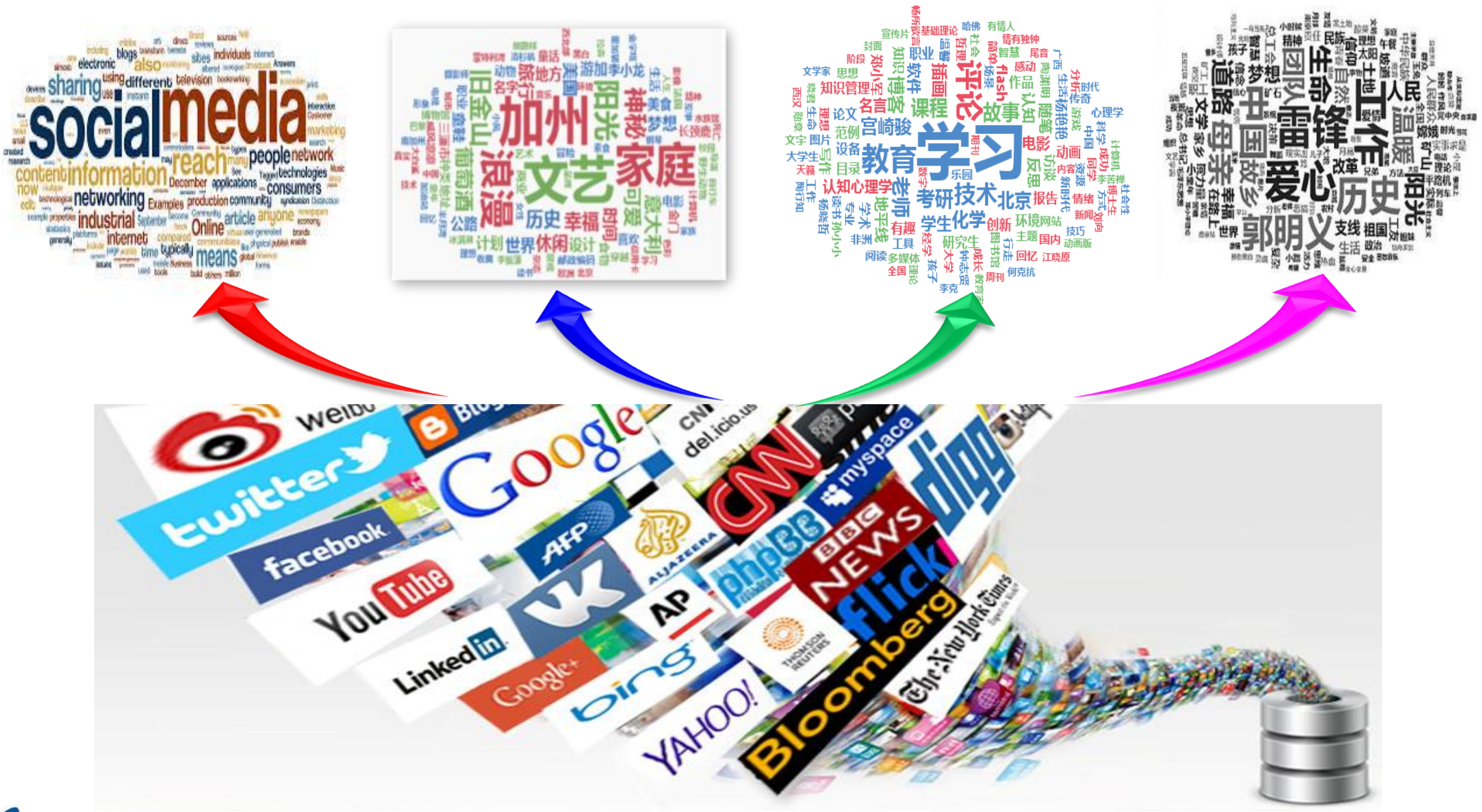
cqzong@nlpr.ia.ac.cn

本章内容

1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

1. 文本分类

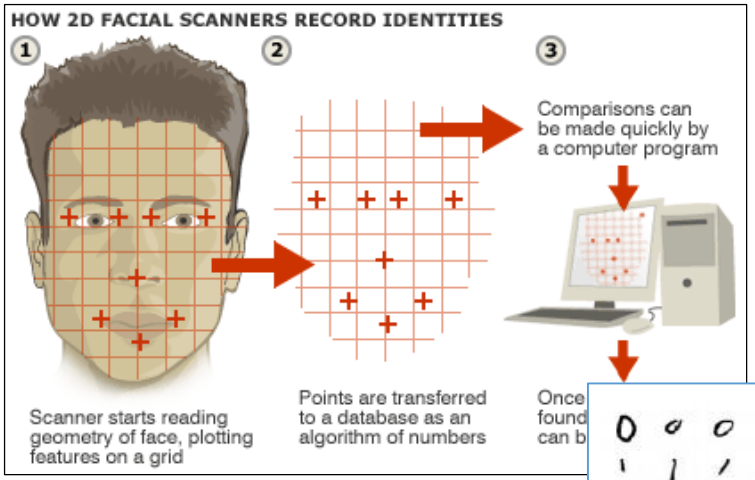
◆ 问题提出

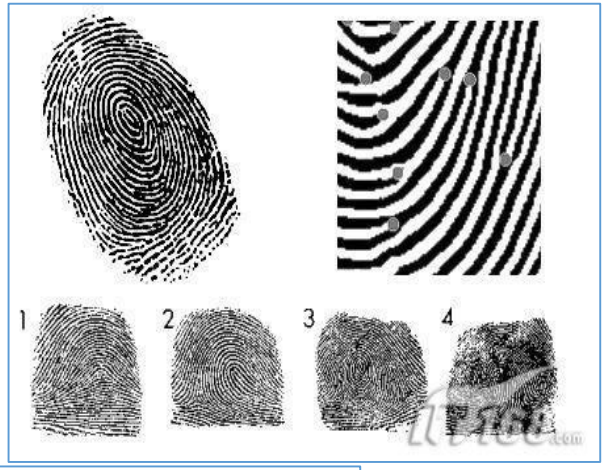


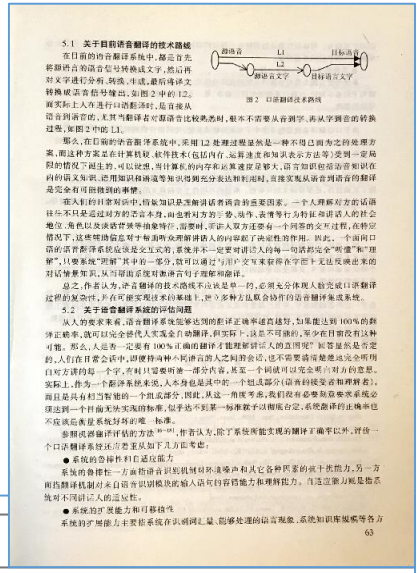
social
media
information
reach
people
network
networking
industrial
internet
means
Online
加州
文艺
家庭
浪漫
阳光
旧金山
葡萄酒
学习
教育
评论
技术
考研
宫崎骏
雷锋
爱心
历史
郭明义
工作
温暖
中国
母亲
weibo
twitter
facebook
YouTube
LinkedIn
Google
Google+
bing
YAHOO!
AFP
AP
CNN
del.icio.us
myspace
digg
BBC NEWS
flickr
Bloomberg
The New York Times

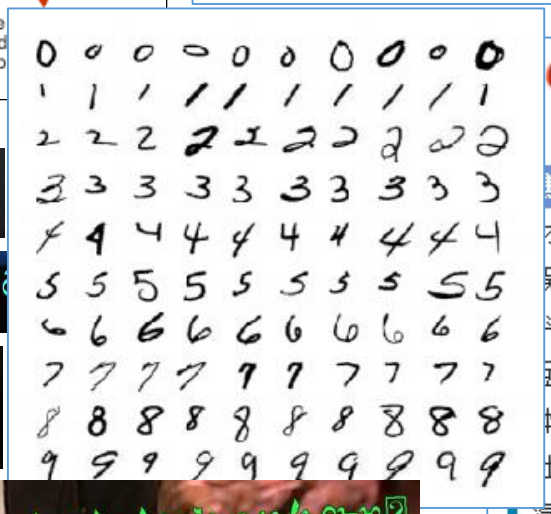
1. 文本分类

◆ 问题提出


HOW 2D FACIAL SCANNERS RECORD IDENTITIES
1
2
3
Comparisons can be made quickly by a computer program
Scanner starts reading geometry of face, plotting features on a grid
Points are transferred to a database as an algorithm of numbers







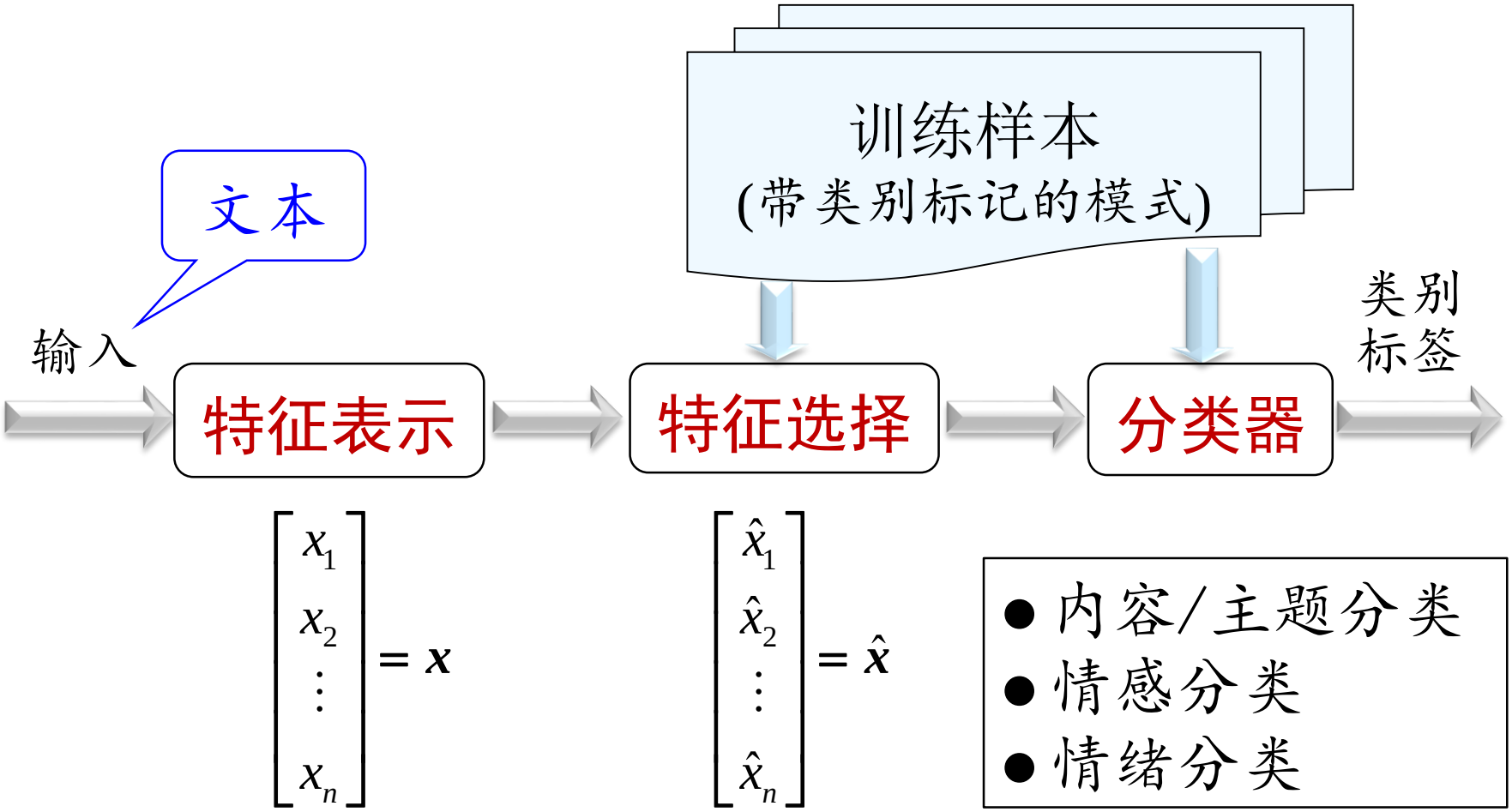

And eventually, I would let it go.


最后，我又会放弃。


oogle 资讯
谷歌 中国版
焦点资讯
财经
娱乐
科技
互联网
体育
社会
汽车
房产
教育
焦点资讯
胡锦涛离开哈萨克斯坦
新华网 - 2小时前
12月12日，中国国家主席胡锦涛
新华社记者兰红光摄 新华网阿斯
坦的工作访问，13日离开阿斯塔
光明网
腾讯网 - 中新网 - 搜狐
此专题所有 965 篇报道 »
男子枪击纵火致12人死引出两点疑问
欧盟坚持将中国减排行动纳入国际协议
下周冷空气活动频繁 气温将持续下降

1. 文本分类

◆ 基于统计学习的模式分类框架



1. 文本分类

◆ 特征表示—文本表示

- 向量空间模型 → 文本被表示成特征向量。特征是离散的符号，词、词性、短语、*n*-grams等。特征项的权重通常采用IDF, TF-IDF等。

  见第7章。

- 分布式表示模型 → 词和句子的分布式向量学习。

  见第7章（词向量）、第11章（预训练语言模型）。

1. 文本分类

◆ 特征选择

(1)文档频率(document frequency, DF)

(2)互信息(mutual information, MI)

(3)信息增益(information gain, IG)

(4)Chi-Square统计(chi-square statistics，CHI)

1. 文本分类

(1)文档频率

用特征词在一个类别中出现的文档数量表示这个特征词与该类别的相关度。出现文档数多的特征词被保留的可能性大。

请看下面的例子：

文本A 关于特征 t_i 与类别 c_j的统计表如下：

特征 \ 类别	c_j	$\overline{c}_j$
t_i	A_{ij}	B_{ij}
$\overline{t}_i$	C_{ij}	D_{ij}

A_{ij}: 类别 c_j 的文档中出现特征 t_i 的文档数；

B_{ij}: 类别 $\overline{c}_j$ 的文档中出现特征 t_i 的文档数；

C_{ij}: 类别c_j 的文档中未出现特征 t_i 的文档数；

D_{ij}: 类别 $\overline{c}_j$ 的文档中未出现特征t_i的文档数；

1. 文本分类

文本A 关于特征 t_i 与类别 c_j的统计表如下：

特征 \ 类别	c_j	$\overline{c}_j$
t_i	A_{ij}	B_{ij}
$\overline{t}_i$	C_{ij}	D_{ij}

A_{ij}: 类别 c_j 的文档中出现特征 t_i 的文档数；

B_{ij}: 类别 $\overline{c}_j$ 的文档中出现特征 t_i 的文档数；

C_{ij}: 类别c_j 的文档中未出现特征 t_i 的文档数；

D_{ij}: 类别 $\overline{c}_j$ 的文档中未出现特征t_i的文档数；

假设 N_{all} 是所有文档的总数；C 是所有的类别数。可以计算出如下概率：

$$P(c_j) \approx (A_{ij} + C_{ij}) / N_{all}$$

$$P(t_i) \approx (A_{ij} + B_{ij}) / N_{all}$$

$$P(\overline{t}_i) \approx (C_{ij} + D_{ij}) / N_{all}$$

$$P(c_j \mid t_i) \approx \frac{A_{ij} + 1}{A_{ij} + B_{ij} + C}$$

$$P(c_j \mid \overline{t}_i) \approx \frac{C_{ij} + 1}{C_{ij} + D_{ij} + C}$$

分子上的“+1”和分母上的“+C”起数据平滑效果。

根据概率$P(c_j|t_i)$等综合确定哪些特征有用。

1. 文本分类

例如：一共有20篇文档(N_{all}=20)，分为两类(C=2)：教育类(c_1)和非教育类$\overline{c}_1$。其中，教育类7篇(c_1=7)，非教育类13篇($\overline{c}_1$=13)。特征词t_2=“计算机”是教育类文档的特征之一。那么，A_{12}表示教育类的7篇文档中出现特征词“计算机”的个数，假设有3篇，即A_{12}=3，那么教育类文档中没出现特征词“计算机”的文档数C_{12}=4。同时，在非教育类的13篇文档中，假设有5篇出现了特征词“计算机”，即B_{12}=5，那么D_{12}=13-5=8。于是，

特征 \ 类别	c_1	$\overline{c}_1$
t_2	$A_{12}=3$	$B_{12}=5$
$\overline{t}_2$	$C_{12}=4$	$D_{12}=8$

$$P(c_1) \approx (A_{12}+C_{12})/N_{all} = 7/20 = 0.35$$

$$P(t_2) \approx (A_{12}+B_{12})/N_{all} = 8/20 = 0.40$$

$$P(c_1 \mid t_2) \approx \frac{A_{12}+1}{A_{12}+B_{12}+C} = 4/10 = 0.40$$

……

1. 文本分类

(2)互信息(mutual information, MI)

互信息是关于两个随机变量互相依赖程度的一种度量。

$$I(X,Y)=H(X)-H(X\mid Y)=\sum_{y}\sum_{x}P(x,y)\log\frac{P(x,y)}{P(x)P(y)}$$

$$MI(t_i,c_j)=\log\left(\frac{P(t_i,c_j)}{P(t_i)P(c_j)}\right)\approx\log\left(\frac{A_{ij}N_{all}}{(A_{ij}+C_{ij})(A_{ij}+B_{ij})}\right)$$

点式互信息

$$MI_{\text{avg}}(t_i)=\sum_{j=1}^{C}P(c_j)MI(t_i,c_j)$$

根据特征与类别之间的互信息大小综合确定哪些特征有用。

1. 文本分类

(3)信息增益(information gain, IG)

特征 T_i 对训练数据集 C 的信息增益定义为集合 C 的经验熵 $H(C)$ 与特征 T_i 给定条件下 C 的经验条件熵 $H(C|T_i)$ 之差，即：$IG(C, T_i) = H(C) - H(C|T_i)$。 参阅：李航《统计机器学习》。

$$IG(t_i) = \{-\sum_{j=1}^{C} P(c_j)\log P(c_j)\}$$

$$+\{P(t_i)[\sum_{i=1}^{C} P(c_j \mid t_i)\log P(c_j \mid t_i)]$$

$$+ P(\overline{t_i})[\sum_{j=1}^{C} P(c_j \mid \overline{t_i})\log P(c_j \mid \overline{t_i})]\}$$

在决策树学习中信息增益等价于训练数据集中类与特征的互信息。

1. 文本分类

推导(T_i 表示随机变量 t_i 的出现或者不出现):

$$IG(C,T_i) = H(C) - H(C \mid T_i)$$

$$H(C) = -\sum_{c_j} p(c_j)\log p(c_j)$$

$$H(C \mid T_i) = p(t_i)H(C \mid t_i) + p(\bar{t}_i)H(C \mid \bar{t}_i)$$

$$= -p(t_i)\left(\sum_{c_j} p(c_j \mid t_i)\log p(c_j \mid t_i)\right) - p(\bar{t}_i)\left(\sum_{c_j} p(c_j \mid \bar{t}_i)\log p(c_j \mid \bar{t}_i)\right)$$

$$IG(C,T_i) = H(C) - H(C \mid T_i)$$

$$= -\sum_{c_j} p(c_j)\log p(c_j)$$

$$+ p(t_i)\left(\sum_{c_j} p(c_j)\log p(c_j \mid t_i)\right)$$

$$+ p(\bar{t}_i)\left(\sum_{c_j} p(c_j)\log p(c_j \mid \bar{t}_i)\right)$$

1. 文本分类

- 举例：设有如下训练数据(带类别标签的文档，共4篇)

教育	体育
①北京 理工 大学 计算机 专业 创建 于 1958 年 是 中国 最早 设立 计算机 专业 的 高校 之一	③北京 理工 大学 体育馆 是 2008 年 中国 北京 奥林匹克 运动会 的 排球 预赛 场地
②北京 理工 大学 学子 在 第四 届 中国 计算机 博弈 锦标赛 中 夺冠	④第五 届 东亚 运动会 中国 军团 奖牌 总数 创 新高 男女 排球 双双 夺冠

1. 文本分类

- 词袋表示(含40个词，即词表大小为40)

1958	2008	奥林匹克	北京	博弈	场地	创	创建	大学	
的	第四	第五	东亚	夺冠	高校	计算机	奖牌	届	锦标赛
军团	理工	男女	年	排球	设立	是	双双	体育馆	新高
学子	于	预赛	运动会	在	之一	中	中国	专业	总数	最早

1. 文本分类

特征项“计算机”的信息增益：

特征 \ 类别	教育	体育
计算机	2	0
$\overline{\text{计算机}}$	0	2

$P(\text{计算机})=1/2 \quad P(\overline{\text{计算机}})=1/2$

$P(\text{教育}\mid\text{计算机})=(2+1)/(2+2)=3/4$

$P(\text{体育}\mid\text{计算机})=1/(2+2)=1/4$

$P(\text{教育}\mid\overline{\text{计算机}})=1/(2+2)=1/4$

$P(\text{体育}\mid\overline{\text{计算机}})=(2+1)/(2+2)=3/4$

$$IG(\text{计算机})=-0.5\log 0.5-0.5\log 0.5$$
$$+0.5(0.75\log 0.75+0.25\log 0.25)$$
$$+0.5(0.75\log 0.75+0.25\log 0.25)$$
$$=-\log 0.5+0.75\log 0.75+0.25\log 0.25=0.1308$$

1. 文本分类

特征项“北京”的信息增益：

特征 \ 类别	教育	体育
北京	2	1
$\overline{\text{北京}}$	0	1

$P(\text{北京})=(1+2)/4=3/4$ $P(\overline{\text{北京}})=1/4$

$P(\text{教育}\mid\text{北京})=(2+1)/(3+2)=3/5$

$P(\text{体育}\mid\text{北京})=(1+1)/(3+2)=1/5$

$P(\text{教育}\mid\overline{\text{北京}})=1/(1+2)=1/3$

$P(\text{体育}\mid\overline{\text{北京}})=(1+1)/(1+2)=2/3$

$$IG(\text{北京})=-0.5\log 0.5-0.5\log 0.5$$
$$+0.75(0.6\log 0.6+0.4\log 0.4)$$
$$+0.25(0.667\log 0.667+0.333\log 0.333)$$
$$=0.0293$$

1. 文本分类

根据信息增益的特征排序：

特征	*IG*
计算机 排球 运动会	0.1308
1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 高校 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最早 北京 大学 理工	0.0293
的 夺冠 届 年 是 中国	0.0000

1. 文本分类

- **选择的特征**

计算机 排球 运动会 高校 大学 1958 2008 奥林匹克 博弈 场地 创 创建 第四 第五 东亚 奖牌 锦标赛 军团 男女 设立 双双 体育馆 新高 学子 于 预赛 在 之一 中 专业 总数 最 早 北京 理工

- **精简后的训练数据**

教育	体育
①大学 计算机 高校	③大学 运动会 排球
②大学 计算机	④运动会 排球

1. 文本分类

(4)Chi-Square 统计量 (CHI)

CHI统计量用于检验两个事件之间的独立性，CHI度量了期望计数 E 和实际观察计数 N 之间的相互关系。

$$\chi^2(t,c) = \sum_{It\in\{0,1\}} \sum_{Ic\in\{0,1\}} \frac{(N_{It,Ic} - E_{It,Ic})^2}{E_{It,Ic}}$$

$$\chi^2(t_i,c_j) = \frac{N_{all}\cdot(A_{ij}D_{ij} - C_{ij}B_{ij})^2}{(A_{ij}+C_{ij})\cdot(B_{ij}+D_{ij})\cdot(A_{ij}+B_{ij})\cdot(C_{ij}+D_{ij})}$$

$$CHI_{avg}(t_i) = \sum_{j=1}^{C} P(c_j)\chi^2(t_i,c_j)$$

卡方数值越大，说明“特征 t 和类别 c 越相关”。

1. 文本分类

◆ 分类器

- 关于分类算法的基本概念
  - **模型表示**
    - 用参数进行建模（构建目标函数）
  - **学习算法**
    - 最大似然估计，计算最大后验概率（生成式模型）
    - 梯度下降法、牛顿法（判别式模型）
  - **推断**
    - 决策/预测规则

1. 文本分类

- 分类器
  - 监督学习(supervised learning)
  - 无监督(unsupervised learning)
  - 半监督学习(semi-supervised learning)

1. 文本分类

◆监督学习过程

- **我们需要准备什么？**
  - 训练数据
- **我们的任务是什么？**
  - 利用参数构建模型（确定目标函数）
- **如何进行参数估计？**
  - 根据某个准则从训练数据中学习
  - 学习在训练数据上准则最优的参数

1. 文本分类

- 过程描述
  - 给定有限的、人工标注好的大量数据，假设这些数据符合独立同分布(训练集，training data)；
  - 假设要学习的模型属于某个函数的集合，即存在假设空间(hypothesis space)；
  - 应用某（些）个评价准则(evaluation criterion)，从假设空间中选取最优的模型，使其对已知的训练数据和未知的测试数据(test data)在给定的评价准则下有最优的预测。

1. 文本分类

- 一般步骤

① 获得一个有限的训练数据集合；

② 确定包含所有可能的模型的假设空间，即学习模型的集合；

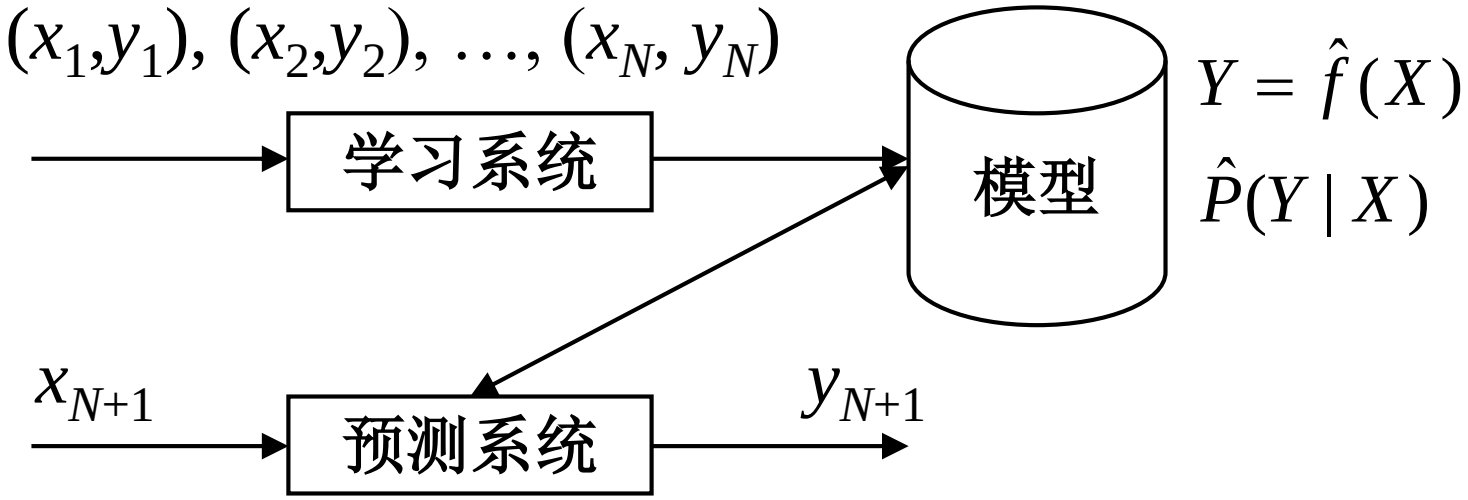
③ 确定模型选择的准则，即学习的策略；

④ 通过学习方法选择最优模型；

⑤ 利用学习到的最优模型对新数据进行预测或分析。

1. 文本分类

- 问题的形式化



给定一个训练数据集：$T=\{(x_1,y_1), (x_2,y_2), \ldots, (x_N, y_N)\}$，其中，$(x_i,y_i)$, $i=1, 2, \ldots, N$，称为样本。x_i是输入的观测值，也称输入或实例；y_i 是输出的观测值，也称输出。

1. 文本分类

◆模型

- 生成式模型
  - 朴素贝叶斯 (na ïve Bayes, NB)
  - 隐马尔可夫模型 (hidden Markov model, HMM)
- 判别式模型
  - 线性判别函数 (linear discriminate function)
  - 支持向量机 (support vector machine, SVM)
  - 最大熵模型 (maximum entropy, ME)
  - 条件随机场 (conditional random fields, CRFs)

1. 文本分类

- 朴素贝叶斯分类器
  - 贝叶斯公式

$$P(B \mid A) = \frac{P(A,B)}{P(A)} = \frac{P(B)P(A \mid B)}{P(A)}$$

  - 贝叶斯决策理论

$$P(c_j \mid \boldsymbol{x}) = \frac{P(c_j, \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(c_j) \times P(\boldsymbol{x} \mid c_j)}{P(\boldsymbol{x})}$$

$$c^* = \arg\max_{j=1,\ldots,C} P(c_j \mid \boldsymbol{x}) \propto \arg\max_{j=1,\ldots,C} P(c_j) \times P(\boldsymbol{x} \mid c_j)$$

1. 文本分类

- 贝叶斯假设

$$P(\boldsymbol{x}\,|\,c_j) \approx P([w_1,\ldots,w_N]\,|\,c_j) \approx \prod_{k=1}^{N} P(w_k\,|\,c_j) = \prod_{i=1}^{M} P(w_i\,|\,c_j)^{N(w_i)}$$

N是 $\boldsymbol{x}$ 中所有的单词数（语料规模）；$N(w_i)$是词w_i出现的次数；M 是不同单词的个数（词汇量）。

例如：$\boldsymbol{x}$ = 我 和 我 的 祖国 。

$N = 6$，N(我)=2，$M = 5$

$$P(\boldsymbol{x}|c_j) \approx P(\text{我}|c_j)\times P(\text{和}|c_j)\times P(\text{我}|c_j)\times P(\text{的}|c_j)\times P(\text{祖国}|c_j)\times P(\text{。}|c_j)$$

$$= [P(\text{我}|c_j)]^2 \times P(\text{和}|c_j)\times P(\text{的}|c_j)\times P(\text{祖国}|c_j)\times P(\text{。}|c_j)$$

1. 文本分类

因此，

$$c^* = \arg\max_{j=1,\ldots,C} P(c_j \mid \boldsymbol{x}) \propto \arg\max_{j=1,\ldots,C} P(c_j) \times P(\boldsymbol{x} \mid c_j)$$

$$= \arg\max_{j=1,\ldots,C} P(c_j) \prod_{i=1}^{M} P(w_i \mid c_j)^{N(w_i)}$$

- 参数估计—最大似然估计

$$P(c_j) \approx \frac{1 + N(c_j)}{C + N_{all}} \qquad P(w_i \mid c_j) \approx \frac{1 + N(w_i, c_j)}{M + \sum_{i'=1}^{M} N(w_{i'}, c_j)}$$

分子加1的目的是用于概率平滑。

贝叶斯模型是理论上最优的分类器!

1. 文本分类

- 应用举例

考虑前面的例子：

$P(c_j)$	P(教育)=0.5	P(体育)=0.5
$P(w_i\|c_j)$	P(计算机|教育)=0.3	P(计算机|体育)=0.1
	P(排球|教育)=0.1	P(排球|体育)=0.3
	P(运动会|教育)=0.1	P(运动会|体育)=0.3
	P(高校|教育)=0.2	P(高校|体育)=0.1
	P(大学|教育)=0.3	P(大学|体育)=0.2

1. 文本分类

给定如下文本：

“北京 理工 大学 是 理工 为主 工理文 协调 发展 的 全国 重点 高校”

特征集= {计算机, 排球, 运动会, 高校, 大学}

$\boldsymbol{x} = [0,0,0,1,1]^{\mathrm{T}}$

$$P(\text{教育}) \times P(\boldsymbol{x} \mid \text{教育}) = 0.5 \times 0.3 \times 0.2 = 0.03$$

$$P(\text{体育}) \times P(\boldsymbol{x} \mid \text{体育}) = 0.5 \times 0.1 \times 0.2 = 0.01$$

$$P(\text{教育} \mid \boldsymbol{x}) = \frac{0.03}{0.03 + 0.01} = 0.75$$

$$P(\text{体育} \mid \boldsymbol{x}) = 0.25$$

$P(\text{教育}|\boldsymbol{x}) > P(\text{体育}|\boldsymbol{x})$

1. 文本分类

给定如下文本：

“复旦 大学 排球 队 获得 本届 大学生 运动会 排球 比赛 冠军”

特征集 = [计算机，排球，运动会，高校，大学]

$$\boldsymbol{x} = [0,1,1,0,1]^{\mathrm{T}}$$

$$P(\text{教育}) \times P(\boldsymbol{x} \mid \text{教育}) = 0.5 \times 0.1 \times 0.1 \times 0.3 = 0.0015$$

$$P(\text{体育}) \times P(\boldsymbol{x} \mid \text{体育}) = 0.5 \times 0.3 \times 0.3 \times 0.2 = 0.0090$$

$$P(\text{教育} \mid \boldsymbol{x}) = \frac{0.0015}{0.0015 + 0.0090} = 0.1429$$

$$P(\text{体育} \mid \boldsymbol{x}) = 0.8571$$

$P(\text{体育}|\boldsymbol{x}) > P(\text{教育}|\boldsymbol{x})$

1. 文本分类

- 组合分类器
  - 多种特征组合
  - 分类器
    - 朴素贝叶斯
    - SVM
    - 最大熵

      ……
  - 分类器组合
    - 投票(voting)/ 加和(sum)/ 最大(max)/ 乘法(product) ……

1. 文本分类

- 开源工具：
  - 贝叶斯分类器：http://www.openpr.org.cn
  - 支持向量机(LibSVM):
    http://www.csie.ntu.edu.tw/~cjlin/libsvm
  - 隐马尔可夫模型：http://htk.eng.cam.ac.uk/
  - 最大熵：
    - OpenNLP：http://incubator.apache.org/opennlp/
    - Malouf： http://tadm.sourceforge.net/
    - Tsujii：http://www-tsujii.is.s.u-tokyo.ac.jp/~tsuruoka/maxent/
    - 张乐：http://homepages.inf.ed.ac.uk/lzhang10/maxent.html
    - 林德康：http://webdocs.cs.ualberta.ca/~lindek/downloads.htm

1. 文本分类

- **条件随机场：**
  - **CRF++** (C++版)：
    http://crfpp.googlecode.com/svn/trunk/doc/index.html
  - **CRFSuite** (C语言版)：
    http://www.chokkan.org/software/crfsuite/
  - **MALLET** (Java版，通用的NLP工具包，包括分类、序列标注等机器学习算法)：http://mallet.cs.umass.edu/
  - **NLTK** (Python版，通用的NLP工具包，很多工具是从MALLET中包装转成的Python接口)：http://nltk.org/

1. 文本分类

◆ 基于神经网络的分类方法

● 词的分布式表示

- ➢ C&W 模型
- ➢ CBOW模型
- ➢ Skip-gram 模型

见第7章

● 句子表示

① 句子中词向量的平均：$\boldsymbol{e}_s = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{e}(w_k)$

其中，n为句子中词的个数；$\boldsymbol{e}(w_k)$ 表示词 w_k 的向量。

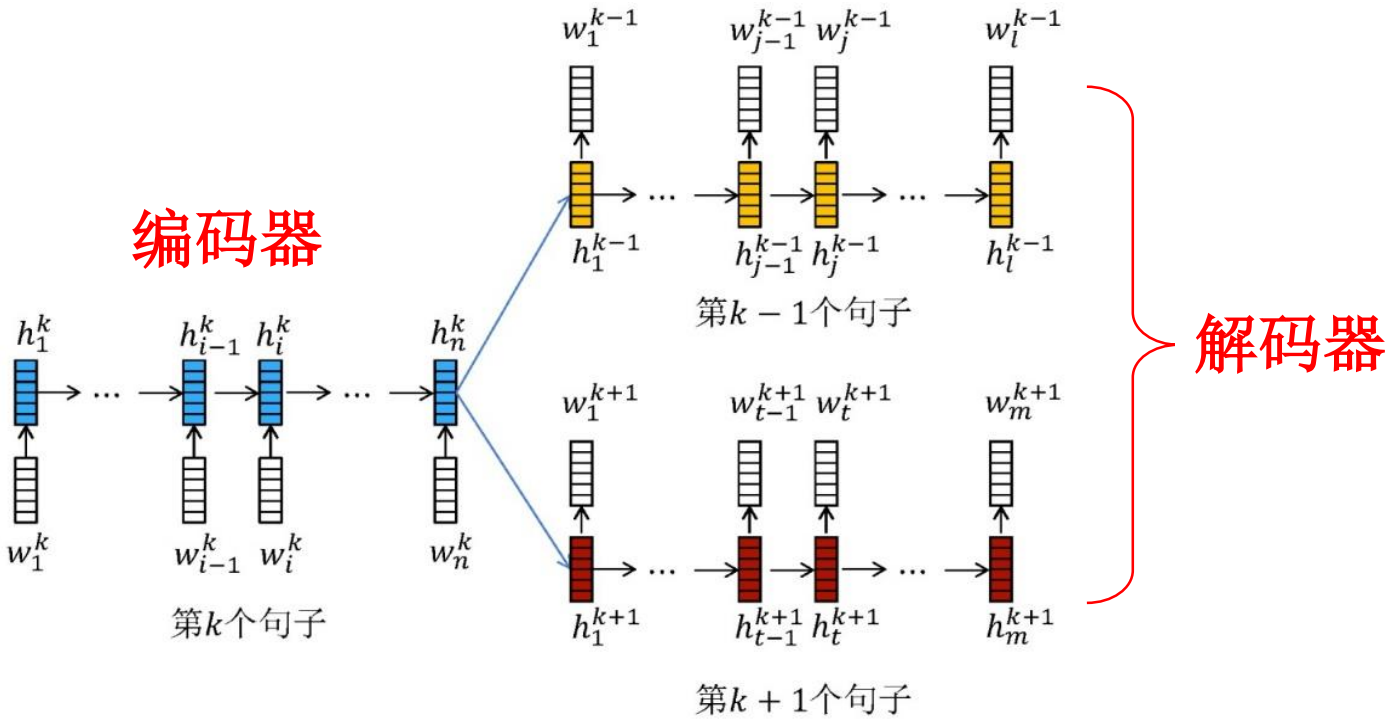
如果考虑句子中不同词的权重：$\boldsymbol{e}_s = \frac{1}{n}\sum_{k=1}^{n} v_k \cdot \boldsymbol{e}(w_k)$

权重v_k可以采用TF-IDF值。

1. 文本分类

② Skip-Thought 模型

借鉴Skip-gram模型，它利用当前句子D_k预测前一个句子D_{k-1}和后一个句子 D_{k+1}。其基本假设是连续出现的句子$D_{k-1}D_kD_{k+1}$ 表达的意思比较接近或相关，因此根据句子D_k可以重构出前后两个句子。



1. 文本分类

➢编码器

可直接采用基于循环神经网络的语言模型，或者循环神经网络的每个神经元采用门控循环单元(gated recurrent unit, GRU)（是LSTM的简化，省去了记忆单元）；或者采用基于LSTM的语言模型。

对于句子$D_k=w_1w_2...w_n$，最后一个位置（第n个词）的隐藏表示($\boldsymbol{h}_n^k$)将作为整个句子的语义编码表示。

那么，如何训练LSTM 或者GRU的参数呢？

—通过解码前后句子。

1. 文本分类

➢ 解码器

以预测前一个句子为例。

每一时刻的输入包括：上一个句子的隐层表示 $\boldsymbol{h}_j^{k-1}$，已经产生的前一个句子的词语序列 $w_1^{k-1}w_2^{k-1}\cdots w_{j-1}^{k-1}$，当前句子$D_k$ 的隐层表示 $\boldsymbol{h}_n^k$，生成前一个句子的下一个词汇的概率为：

$$p\left(w_j^{k-1} \mid w_{<j}^{k-1}, \boldsymbol{h}_n^k\right) \propto \exp(\boldsymbol{e}(w_j^{k-1}), \boldsymbol{h}_j^{k-1})$$

内积

目标函数：

$$\sum_{k=1}^{M}\left\{\sum_{j=1}^{l} p\left(w_j^{k-1} \mid w_{<j}^{k-1}, \boldsymbol{h}_n^k\right) + \sum_{t=1}^{m} p\left(w_t^{k+1} \mid w_{<t}^{k+1}, \boldsymbol{h}_n^k\right)\right\}$$

其中，M为训练集合中句子的数目，l 和 m 分别是前一个句子和后一个句子的长度。

1. 文本分类

- 文档表示（同句子表示）

① 文档中词向量的平均：$\boldsymbol{e}_{\mathrm{s}} = \frac{1}{n}\sum_{k=1}^{n}\boldsymbol{e}(w_k)$

其中，n为文档中词的个数；$\boldsymbol{e}(w_k)$ 表示词 w_k 的向量。

如果考虑文档中不同词的权重：$\boldsymbol{e}_{\mathrm{s}} = \frac{1}{n}\sum_{k=1}^{n} v_k \cdot \boldsymbol{e}(w_k)$

权重v_k可以采用TF-IDF值。

1. 文本分类

②基于句子语义组合的文档表示

假设文档$D = (D_i)_{i=1}^{M}$由M个句子组成，其中第i个句子由n个词汇组成：$S_i = w_{i,1} \cdots w_{i,n}$。利用LSTM可以获得句子$S_i$的分布式表示：

$$\boldsymbol{e}_{s_i} = \text{LSTM}(w_{i,1} \cdots w_{i,n})$$

以句子的分布式表示$\boldsymbol{e}_{s_1} \cdots \boldsymbol{e}_{s_M}$作为输入，用双向LSTM分别学习每个句子$S_i$的正向（从文档第1个句子到当前位置$i$）隐层表示$\overrightarrow{\boldsymbol{h}_i}$和逆向隐层（从最后一个句子到当前位置$i$）表示$\overleftarrow{\boldsymbol{h}_i}$：

$$\overrightarrow{\boldsymbol{h}_i} = \text{LSMT}\left(\boldsymbol{e}_{s_i}, \overrightarrow{\boldsymbol{h}_{i-1}}\right)$$

$$\overleftarrow{\boldsymbol{h}_i} = \text{LSMT}\left(\boldsymbol{e}_{s_i}, \overleftarrow{\boldsymbol{h}_{i+1}}\right)$$

从文档第1个句子到第i-1个句子的隐藏向量。

从文档最后句子到第i+1个句子的隐藏向量。

1. 文本分类

将双向隐层表示拼接，形成句子S_i对应的隐层表示：

$$\boldsymbol{h}_i = \left[\overrightarrow{\boldsymbol{h}}_i; \overleftarrow{\boldsymbol{h}}_i\right]$$

那么，由文档中所有句子的隐层表示可以得到整个文档的表示：

$$\boldsymbol{e}_D = \sum_{i=1}^{M} v_i \boldsymbol{h}_i$$

其中，v_i为权重。如果取平均的话，$v_i = \frac{1}{M}$。

也可以通过注意力机制模型学习权重。

1. 文本分类

● 分类任务

在分类问题中，类别标签 $y \in \{1, 2, \cdots, C\}$可以有$C$个取值，对于给定的文本表示$\boldsymbol{e}_D$，分类层首先采用一个全连接网络将$\boldsymbol{e}_D$转换为维度为类别数目$C$的分值向量 $\boldsymbol{x}=[x_1, x_2, \ldots, x_C]$:

$$\boldsymbol{x} = f(\boldsymbol{W}\boldsymbol{e}_D + \boldsymbol{b})$$

$$\boldsymbol{W} = \begin{bmatrix} w_{c1} \\ w_{c2} \\ \vdots \\ w_{cn} \end{bmatrix} \quad \boldsymbol{b} = \begin{bmatrix} b_{c1} \\ b_{c2} \\ \vdots \\ b_{cn} \end{bmatrix} \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

Softmax 回归预测属于类别c 的条件概率为:

$$p(L(D) = c \mid \boldsymbol{x}) = \text{Softmax}(\boldsymbol{x})$$

$$= \frac{\exp(x_c)}{\sum_{c'=1}^{C} \exp(x_{c'})}$$

文档D的类别标签

关于Softmax(.)，请见第6章。

那如何获得参数 $\boldsymbol{W}$ 和 $\boldsymbol{b}$ 呢？

1. 文本分类

在文本/情感/情绪分类中，有大量的标注数据 $T=\{(D, L)\}$，D为文档/句子，L为文档/句子/属性等分析对象正确的标签。训练过程以最小化交叉熵损失为模型优化的目标：

$$Loss = -\sum_{D \in T} \sum_{k=1}^{C} L_k(D) \log(p_k(D))$$

如果文档 D 属于第 k 类，则 $L_k(D)=1$；否则，$L_k(D)=0$。利用梯度下降法训练参数 $\boldsymbol{W}$ 和 $\boldsymbol{b}$。

本章内容

1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

2. 文本聚类

- 目的

  在某种度量指标或维度上，相似度大的文本聚集在一起。

- 假设

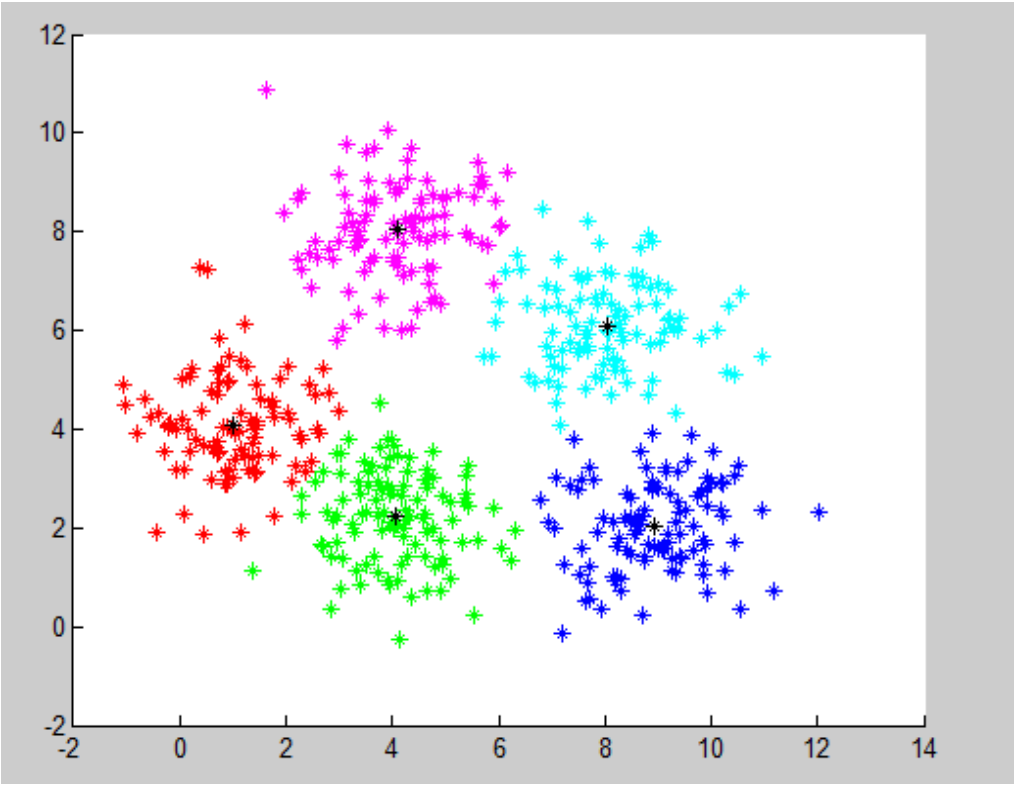
  –同类的文本相似度较大

  –不同类的文本相似度较小

- 与文本分类的区别

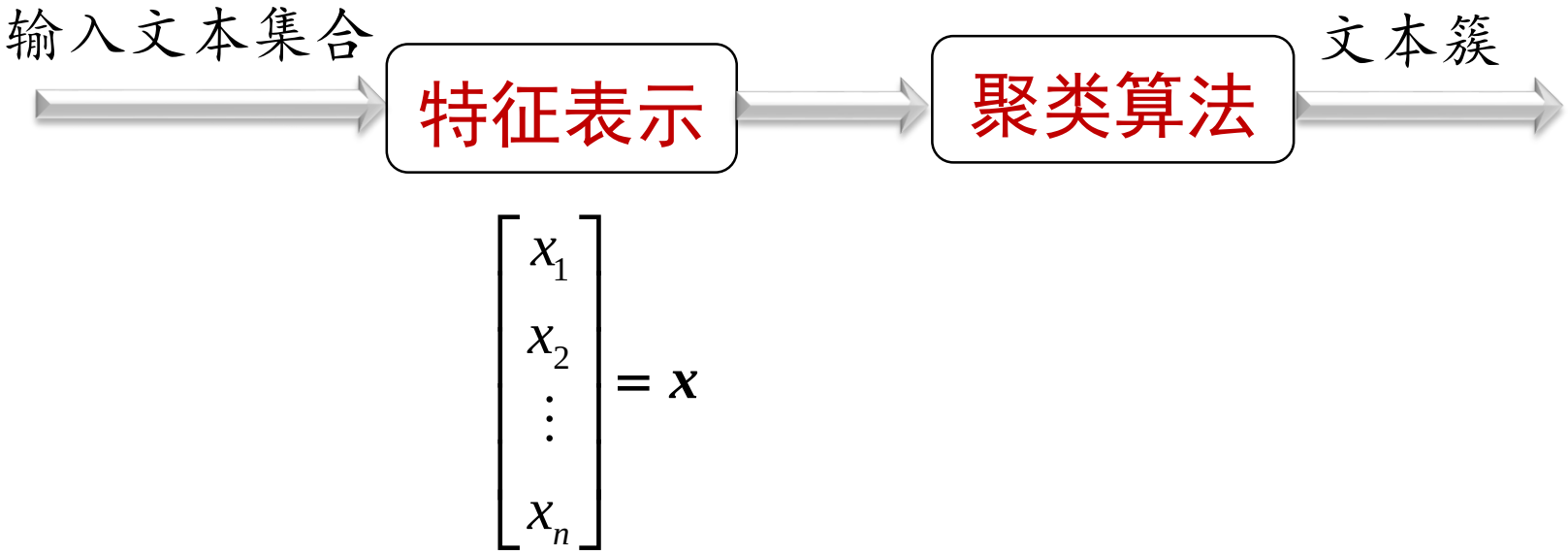
  –没有带标签的训练数据；

  –分类任务中类别是确定的，而聚类任务中类别是不确定的。



2. 文本聚类

◆ 基于统计学习的模式聚类框架



2. 文本聚类

- 聚类算法
  - 分割法
    - K-means 算法
    - K-medoids (质心)算法
    - CLARANS 算法
  - 层次法
    - CURE 算法
    - BIRCH 算法
  - 基于密度的方法
  - 基于网格的方法

2. 文本聚类

◆K-means 算法

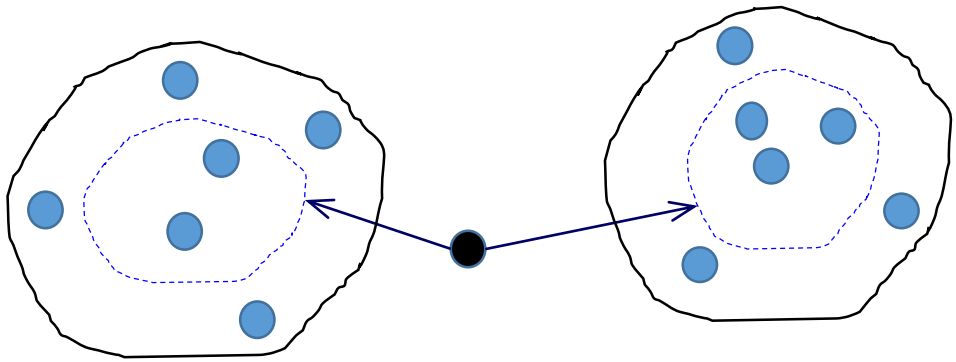
● 基本思路

① 随机选取k个文本作为初始的聚类种子;

② 根据聚类种子的值，将每个文本重新赋给最相似的簇;

③ 重新计算每个簇中**所有文本的平均值**，用此平均值作为新的聚类种子;

④ 重复执行②③步，直到各个簇不再发生变化。



2. 文本聚类

◆K-medoids (质心)算法

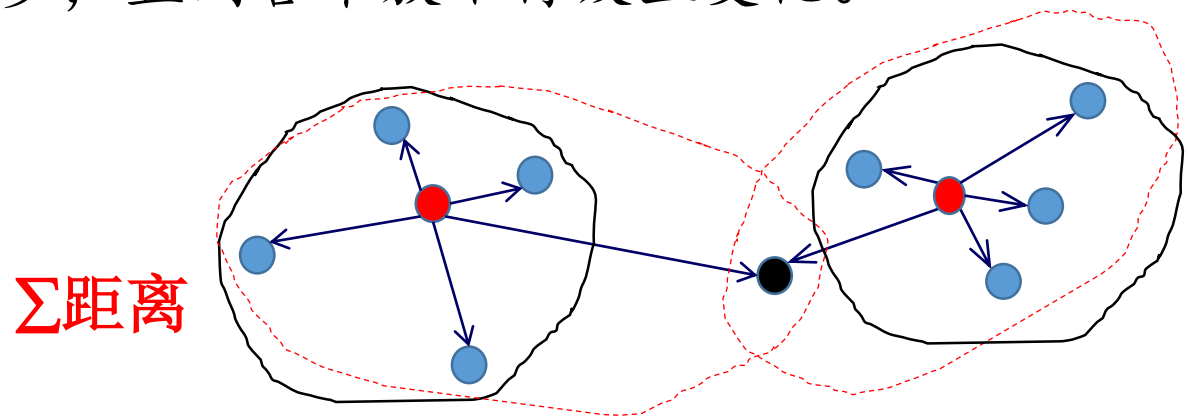
● 基本思路

① 随机选取 k 个文本作为初始的聚类种子;

② 根据聚类种子的值，将每个文本重新赋给最相似的簇;

③ 重新计算每个簇的**中心文本**，要求该文本到簇中其它所有文本的距离之和最小，用此文本作为新的聚类种子;

④ 重复执行②③步，直到各个簇不再发生变化。



2. 文本聚类

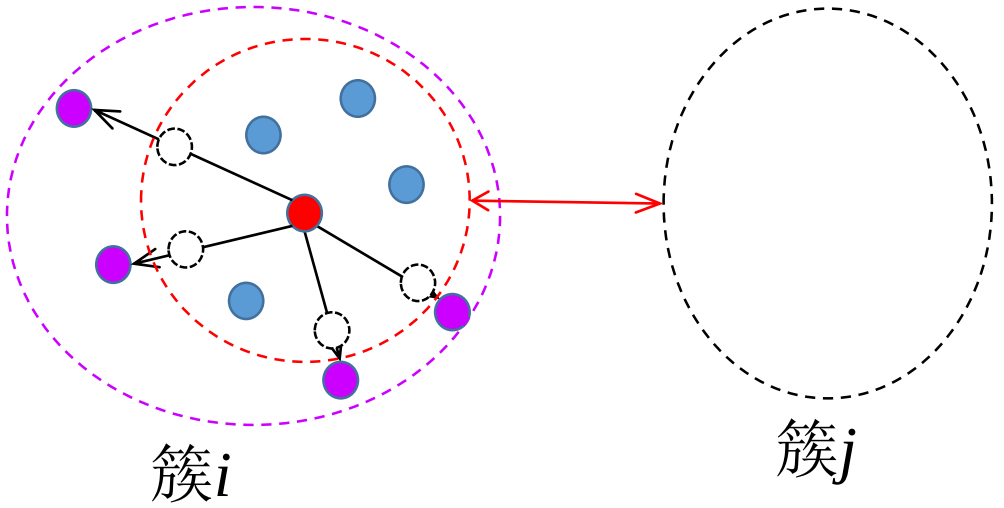
◆CURE 算法

— Clustering Using Representatives

● 基本思路

与用质心算法和所有点（文本）代表一个簇相比，CURE采用折中的方法，通过选取多个“代表性的样本(representatives)”表示一个簇，这种方法更加鲁棒。

$c = 4$
(实际上通常$c \geq 10$)



2. 文本聚类

- 实现算法

假设最终要聚成k个簇，或者根据簇之间的距离决定是否进一步合并，每个簇用c个“代表(representatives)”表示(通常$c \geq 10$)。

① 将各个样本单独看成一个簇;

② 合并最近的簇，直到每个簇中至少包含c个样本;

③ 对于每个簇，执行如下循环:

(a)选取簇中距离质心最远的一个样本作为第一个点，然后依次选取离已选的点最远的样本，直到选定c个样本为止(这些点捕获了簇的形状和大小，作为该簇的代表(representatives));

(b)利用选取到的这c个样本与质心点共同确定簇的代表点：根据参数α ($0<\alpha<1$, 通常取值区间为0.2~0.7可获得较好的效果)向质心收缩, 收缩后得到的样本即为该簇的“代表”:

representatives = 选择样本c_i+ α(质心点 – 样本c_i)

④ 如果多于k个簇，或者有的簇之间的距离小于某个阈值(利用“代表(representatives)”计算距离)，则合并距离最近的两个簇，返回到上面的(a)步，重复上述过程，直到聚成k个簇，或者无可合并的簇为止。

本章内容

1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

3. 情感与情绪分析

◆文本信息划分

- 客观性事实(Facts)和内容(contents) ⟹ 文本分类
- 主观性观点(Opinions)
  - ➢ 情感(sentiment)倾向性(orientation)
  - ➢ 情绪(emotion)

观点挖掘
(opinion mining)

3. 情感与情绪分析

◆基本概念

情感(sentiment)是人对客观事物是否满足自己的需要而产生的态度体验。在自然语言处理中所说的**情感分析**(sentiment analysis)是指对文本作者在文中所表达的对特定事件、物品、观点等目标的喜好倾向进行分析。通常用喜欢/赞同/支持(positive)、厌恶/反对/排斥(negative)和中立(neutral) 三种状态表示。

情绪(emotion)是对一系列主观认知经验的通称，是人对客观事物的态度体验以及相应的行为反应。一般认为，情绪是以个体愿望和需要为中介的一种心理活动。在自然语言处理中所说的**情绪分析**(emotion analysis)是指对文本作者在文中所表达的对某个特定事件、物品、观点等目标的情绪类别进行分析。常见的情绪类别包括：喜、怒、忧、思、悲、恐、惊7种。

3. 情感与情绪分析

◆研究意义

- 随着Web2.0飞速发展和Web3.0的兴起，互联网中出现大量的UGC数据，其中包含了大量的观点信息，如博客、微博、商品评论、论坛 ……
- 企业对观点挖掘和倾向性分析的需求，自动发现用户情感与观点（市场智能化）；感知社会发展趋势；获取商业机会；在线名誉管理；目标导向地广告…
- 普通用户对观点挖掘和倾向性分析的需求，有助于购买产品；有利于发现针对政治话题的观点…
- 政府对观点挖掘和倾向性分析的需求, 实时了解社会舆情，掌握公众整体情绪，预警或检测公共事件…

3. 情感与情绪分析

◆问题与挑战

- 从海量非结构化文本中挖掘分析情感倾向性或情绪
- 情感/情绪表达的隐蔽性(委婉、含蓄、隐喻)、多样性(针对不同的方面)和复杂性(不是简单的喜欢和排斥)
- 不同语言(汉语、英语、阿拉伯语等)、不同文体(评论、对话等)、不同领域(图书、电影、点钟产品等)的差异性

……

3. 情感与情绪分析

◆举例

“我今年入手诺基亚5800，把玩不到24小时，目前感觉5800屏幕很好，操作也很方便，通话质量也不错，但是外形有些偏女性化，不适合男生。这些都是小问题，最主要的问题是电池不耐用，只能坚持一天，反正我觉得对不起这个价格。”

- 外形
- 电池

- 屏幕
- 操作
- 通话质量

3. 情感与情绪分析

◆任务划分

- 文本情感分析

(1)(整体)情感识别(sentiment identification)

(2)性级情感分析(aspect-based sentiment analysis)

(3)观点要素抽取

–观点属性抽取(opinion attribute extraction)

–观点摘要(opinion summarization)

(3)观点检索

3. 情感与情绪分析

- 文本情绪分析
  - 情绪分类
  - 情绪溯因
  - 情绪-原因对抽取
  - 情绪强度检测

3. 情感与情绪分析

◆任务举例

(1)情感识别

- **极性分类** (Positive/ Negative/ Neutral)
  - 这家餐厅总体来说还可以。(Neutral)
  - 但是价格偏贵，人均消费100块。(Negative)
  - 抛开价格的因素还是很不错的。(Positive)
- **强度识别**
  - iPhone 6s的价格太贵了，两个肾没了。(强烈反对)
  - iPhone 6s的价格有点贵。(有点差)

3. 情感与情绪分析

- 层次类别

✧**词级别：**识别一个词的倾向性

✧**特征/要素级别**(aspect level)：识别一个要素(aspect)的倾向性

例如：“这家餐厅价格偏贵，人均消费100块”→ 价格

✧**句子级别：**识别一个句子的观点倾向性

✧**文档级别：**识别一篇文本(包含多个句子)整体的倾向性。

3. 情感与情绪分析

(2)观点要素识别

●观点持有者抽取

- 中方发言人对彭佩奥的“甩锅”言论表示强烈愤慨。

在新闻文本中大量出现，通常为命名实体、名词性短语或者术语，但在商品评论文本中很少出现。

●观点目标抽取

- 中方发言人痛斥彭佩奥的“甩锅”言论。
- 这款手机的屏幕太小，分辨率不足。

通常为术语、事件、实体等。

3. 情感与情绪分析

● 观点摘要

例如：

“I bought an iPhone a few days ago. It was such a nice phone. The touch screen was really cool. The voice quality was clear too. Although the battery life was not long, that is ok for me. However, my mother was mad with me as I did not tell her before I bought the phone. She also thought the phone was too expensive, and wanted me to return it to the shop. …”

观点摘要：

特征 1: touch screen

Positive: 212(文档中该特征的正向评论数)

- *The touch screen was really cool.*
- *The touch screen was so easy to use and can do amazing things.*

… …

Negative: 6(文档中该特征的负向评论数)

- The screen is easily scratched.
- I have a lot of difficulty in removing finger marks from the touch screen.

… …

特征 2: battery life

… …

3. 情感与情绪分析

(3)观点检索

根据用户的查询找出对于主题信息发表了观点的文档。

- 主题相关并且具有主观倾向性
- Blog Search, Twitter, Forum……

3. 情感与情绪分析

◆基本识别方法

以情感识别为例：

- 词汇级情感识别
- 句子级情感识别
- 文档级情感识别

3. 情感与情绪分析

- **词汇级情感识别方法**
  - ✧**任务：**识别词汇的情感倾向性。
  - ✧**思路：**利用词汇之间的相似度进行扩展。
  - ✧**方法：**
    - 基于词典的方法
    - 基于语料资源的方法

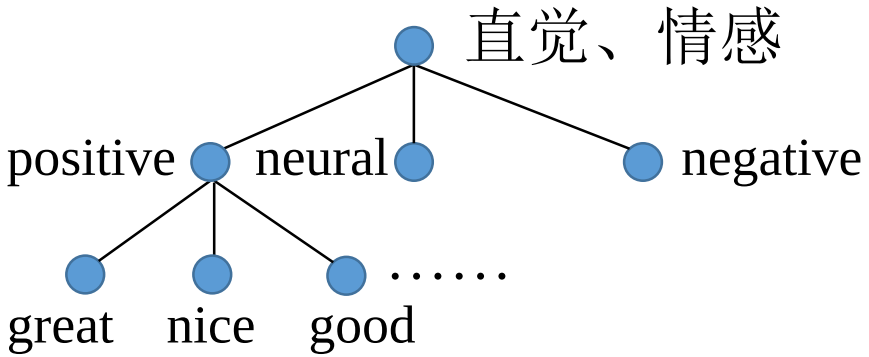
3. 情感与情绪分析

①基于词典的方法

利用词汇之间在WordNet、HowNet 等中的同义、反义关系对情感词典(词汇+极性)进行扩展。

- **positive adjectives:** great, fantastic, nice, cool …
- **negative adjectives:** bad, dull …

WordNet:



3. 情感与情绪分析

②基于语料资源的方法

利用网络资源计算两个词之间的相关度（点式互信息）。

$$\mathrm{PMI}(word_1, word_2) = \log_2 \left[\frac{\mathrm{p}(word_1 \,\&\, word_2)}{\mathrm{p}(word_1)\,\mathrm{p}(word_2)} \right]$$

利用相关度识别词语的情感倾向性:

SO(*word*) = PMI(*word*, “excellent”) – PMI(*word*, “poor”)

3. 情感与情绪分析

✧方法评价

- **优点：**模型直观，易于计算。
- **缺点：**
  - 利用词典或大规模语料计算词之间的相似性易产生噪音；
  - 部分词语的倾向性与上下文相关，与主题相关，如：屏幕大，体积太大。
  - 大部分方法只计算了形容词的倾向性，忽略了动词、副词、名词以及网络用语的情感倾向性。

3. 情感与情绪分析

- 句子级情感识别方法
  - ✧ **任务：**识别句子的情感倾向性。

    如：“美国新冠肺炎感染者人数的持续增长是特朗普政府不作为的有力证据。”

    
  - ✧ **思路：**通过特征表示和选择构建分类器；神经网络。
  - ✧ **方法：**
    - 基于词典的方法
    - 基于语料的统计方法
    - 神经网络方法

3. 情感与情绪分析

①基于词典的句子情感识别方法

利用句子中词的倾向性确定句子的倾向性。

Step1: 词性标注，利用模板选择带有情感的表达短语，如(JJ NN)

Step2: 利用点式互信息计算短语的情感极性:

PMI(*word*, “excellent”) – PMI(*word*, “poor”)

Step3: 计算句子中所有情感短语的平均情感极性。

3. 情感与情绪分析

②基于统计学习的句子情感识别方法

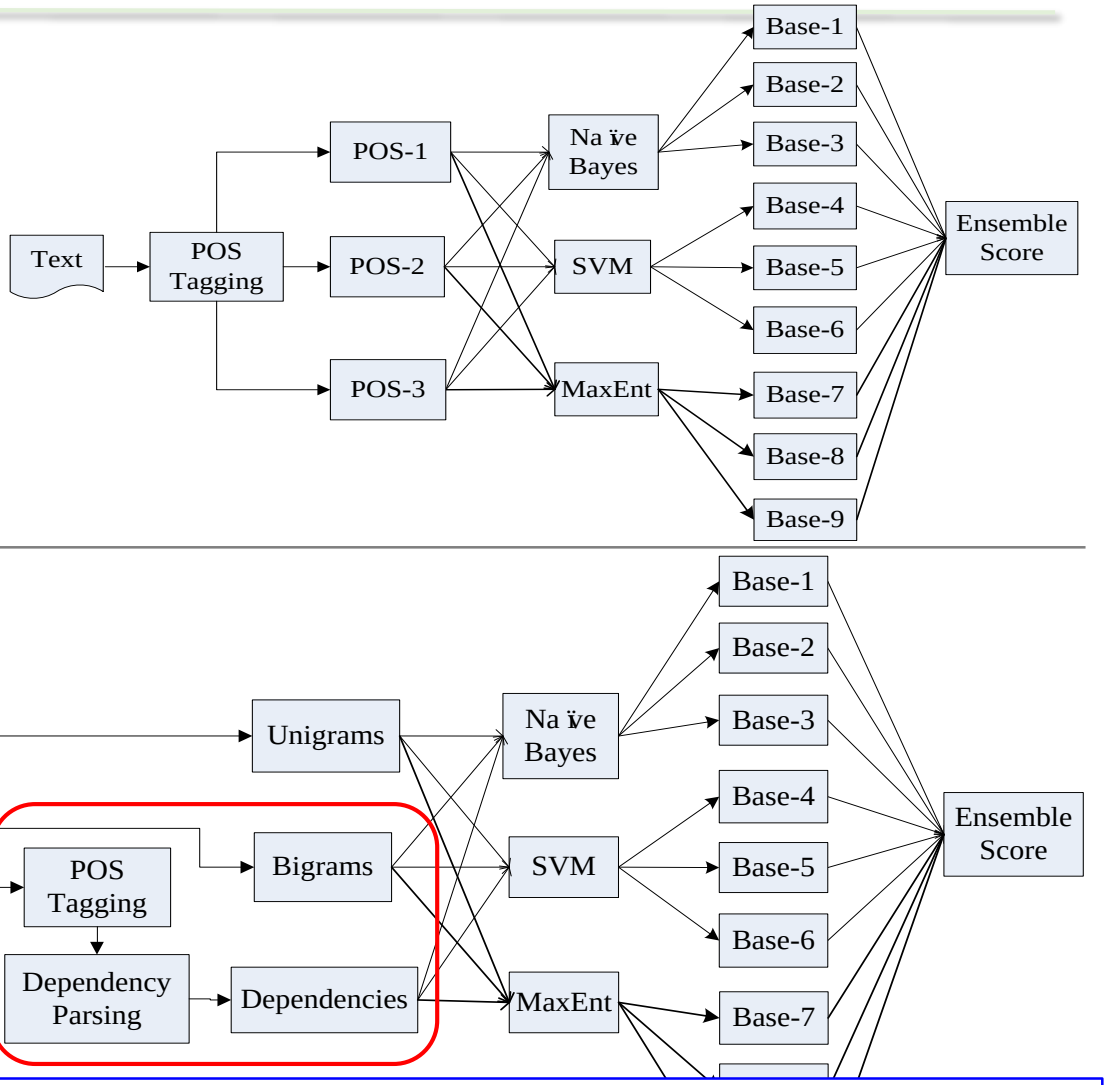
基本思路同基于内容的文本分类方法：

- 常用的特征：uni-gram, bi-gram, POS, Adj, position etc.
- 常用的分类器：NB, SVM, ME …...
- 组合分类器

3. 情感与情绪分析

- 组合学习方法
  - 多种特征组合
  - 分类器
    - 朴素贝叶斯
    - SVM
    - 最大熵

组合决策！

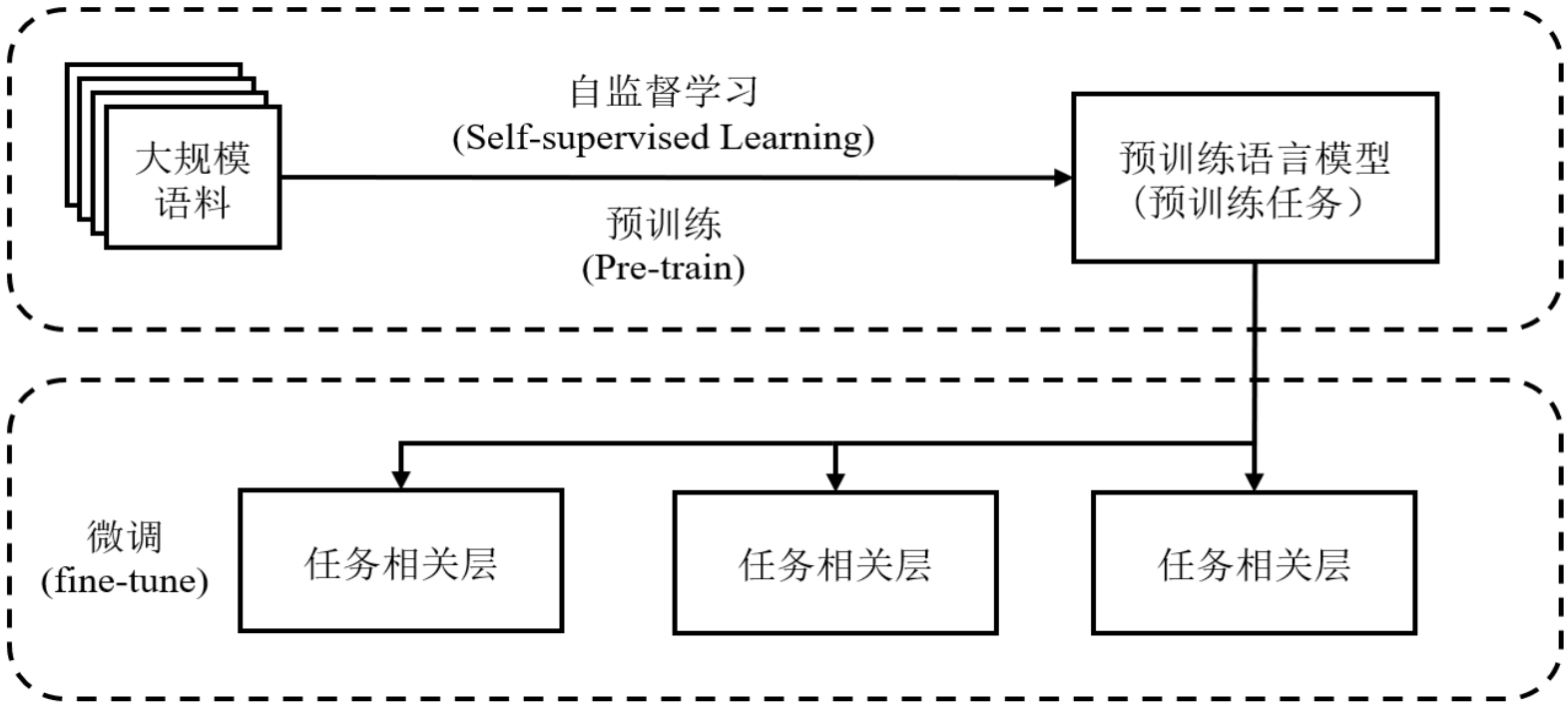


R. Xia, C. Zong, and S. Li. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(2011): 1138–1152

3. 情感与情绪分析

③ 神经网络方法

- 分布式表示 + Softmax(.)：同基于内容的文本分类
- 基于预训练语言模型的分类方法

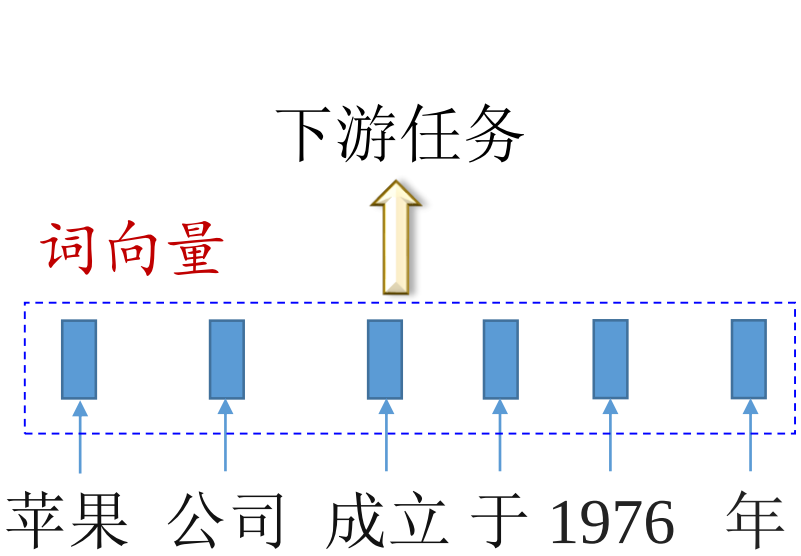


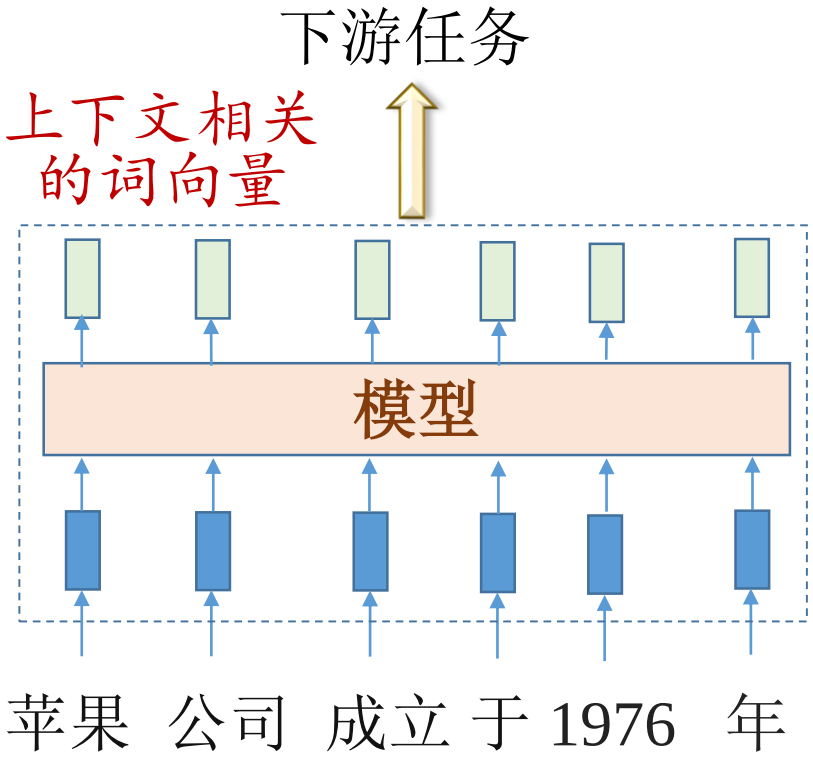
3. 情感与情绪分析

- 预训练语言模型
  - **动机：**
    - 先利用大量无标注数据学习通用的上下文相关词向量
    - 再利用少量标注数据对参数进行调优(fine-tuning)，以提升对不同任务的效果
  - **回顾：词向量**(word embeddings)（见第7章文本表示）
    每个词被表示为一个向量，词向量用于下游任务。问题是：同一单词在不同上下文的向量相同。
  - **上下文相关的词向量**(contextual word embeddings)
    目标：模型既学习词向量，也学习句子中词之间的编码方式。
    方法：通过预训练语言模型得到上下文相关的动态词向量。

3. 情感与情绪分析

- 词向量与上下文相关词向量的对比

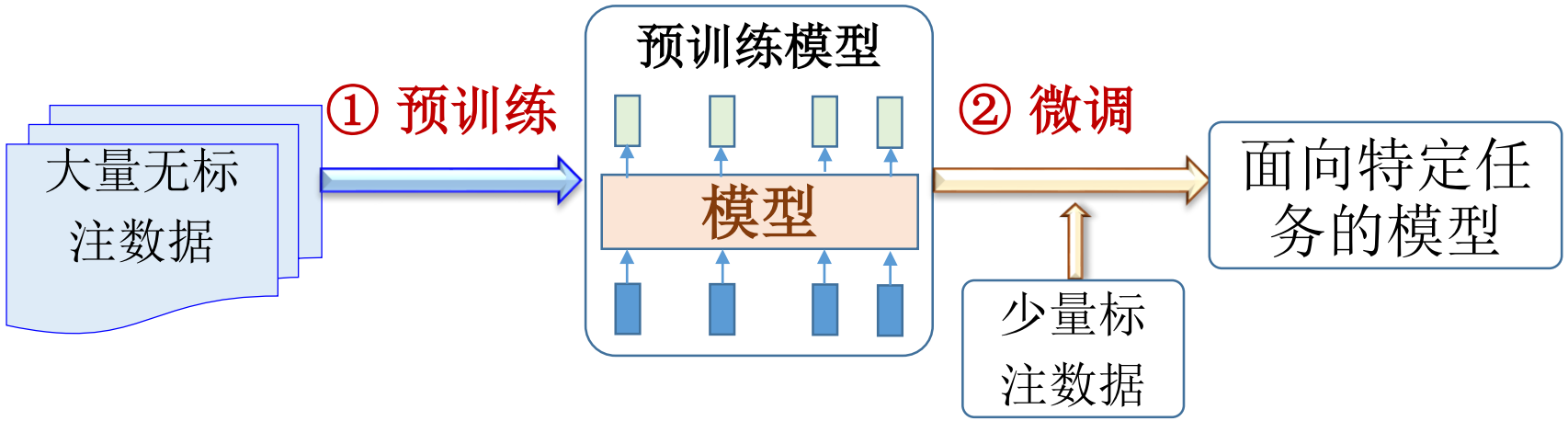




预训练语言模型不仅包含词向量，还包含一个模型(结构和参数)

3. 情感与情绪分析

● 使用预训练语言模型的一般过程



➢ **预训练**

- 收集大量无标注数据
- 设计各种自监督语言模型任务
- 各种模型结构
  - LSTM（ELMo [1]）
  - Transformer编码器 (BERT[2])

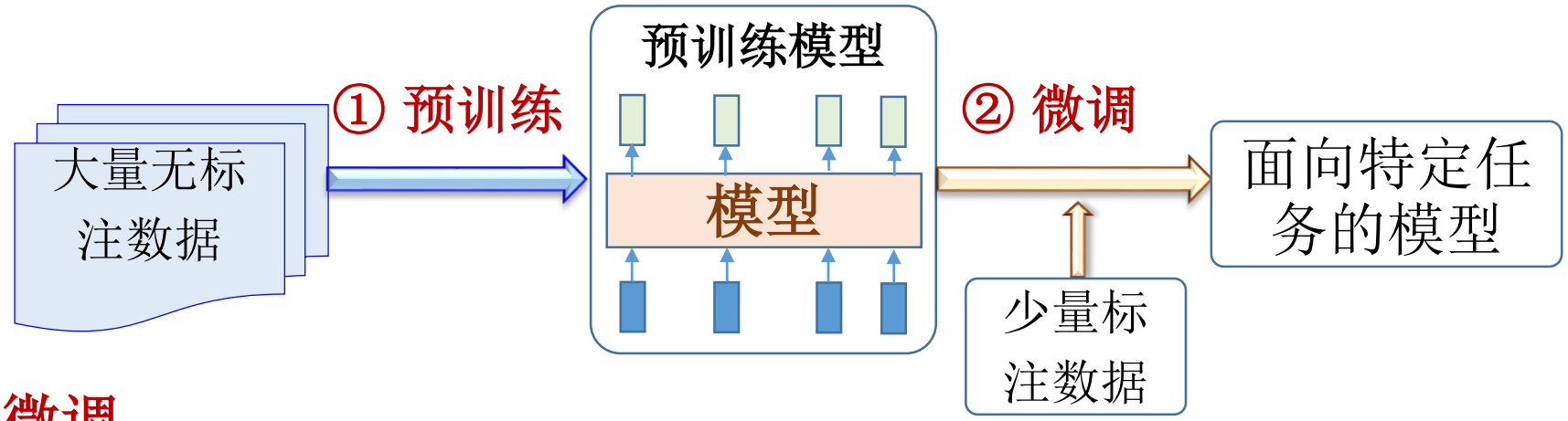
➢ **预训练语言模型的特点：**

- 数据量大
- 层数深
- 参数多

需要足够的算力和财力支持。

3. 情感与情绪分析

- 使用预训练语言模型的一般过程



➢ **微调**

- 确定特定任务，并收集相关数据
- 根据特定任务，选择一个预训练语言模型（如BERT[2]）作为初始模型
- 利用小规模任务数据对预训练模型的参数进行微调：将一个通用的预训练模型优化为一个面向特定任务的模型

3. 情感与情绪分析

参考文献：

[1]**ELMo (E**mbeddings from **L**anguage **Mo**dels**)**

Matthew E. Peters et al. Deep contextualized word representations. *Proc.* NAACL-HLT 2018

[2]**BERT(B**idirectional **E**ncoder **R**epresentations from **T**ransformers**)**

Jacob Devlin et al. BERT: pre-training of deep bidirectional transformers for language understanding. *Proc.* NAACL-HLT 2019

[3]**GPT (G**enerative **P**re-trained **T**ransformer**)**

Alec Radford et al. Improving language understanding by generative pre-training. *Technical report, OpenAI* 2018

[4]**GPT-2**

Alec Radford et al. Language models are unsupervised multitask learners. *OpenAI Blog*, 2019

[5]**GPT-3**

Tom B. Brown et al. Language models are fewshot learners. In *Proc. NeurIPS* 2020

见第6章

3. 情感与情绪分析

- 基于BERT的分类方法（分别以内容分类和情感分类为例）

➢ 收集数据

(1)大规模无标注文本语料（用于预训练）

文档1： 国科大 的 研究生 教育 发端 于 中科院 。 1950 年，中科院 启动 研究 实习员 的 培养 工作 。（上述两句话在文档中是连续的）……

(2)少量基于内容的文本分类标注数据（用于微调内容分类模型）

句子1： 理工 大学 学子 在 第四 届 中国 计算机 博弈 锦标赛 中 夺冠 。

类别： 教育

……

(3)少量情感分析标注数据（用于微调情感分类模型）

句子1： 这 家 餐厅 总体 来 说 不行　　**情感标签：** 负

……

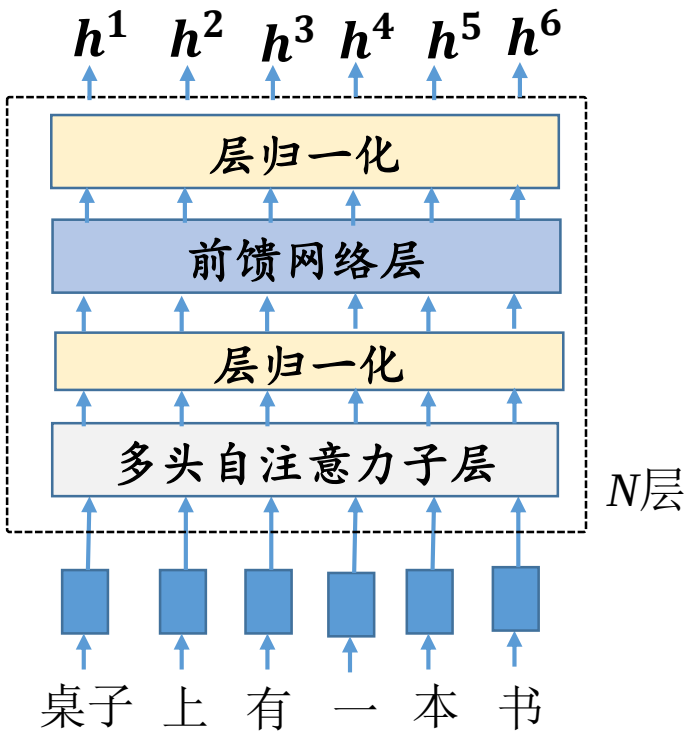
3. 情感与情绪分析

Step1 预训练

➢BERT的结构:

- 回顾Transformer (**见第11章机器翻译**)
  编码-解码框架
  编码器和解码器采用基于自注意力机制的神经网络
- BERT仅使用Transformer的**编码器**（见右图）
- BERT的预训练包含两个任务：
  (a)掩码语言模型(Masked language model)
  (b)句子衔接预测(Next sentence prediction)

为更多的下游任务做准备。

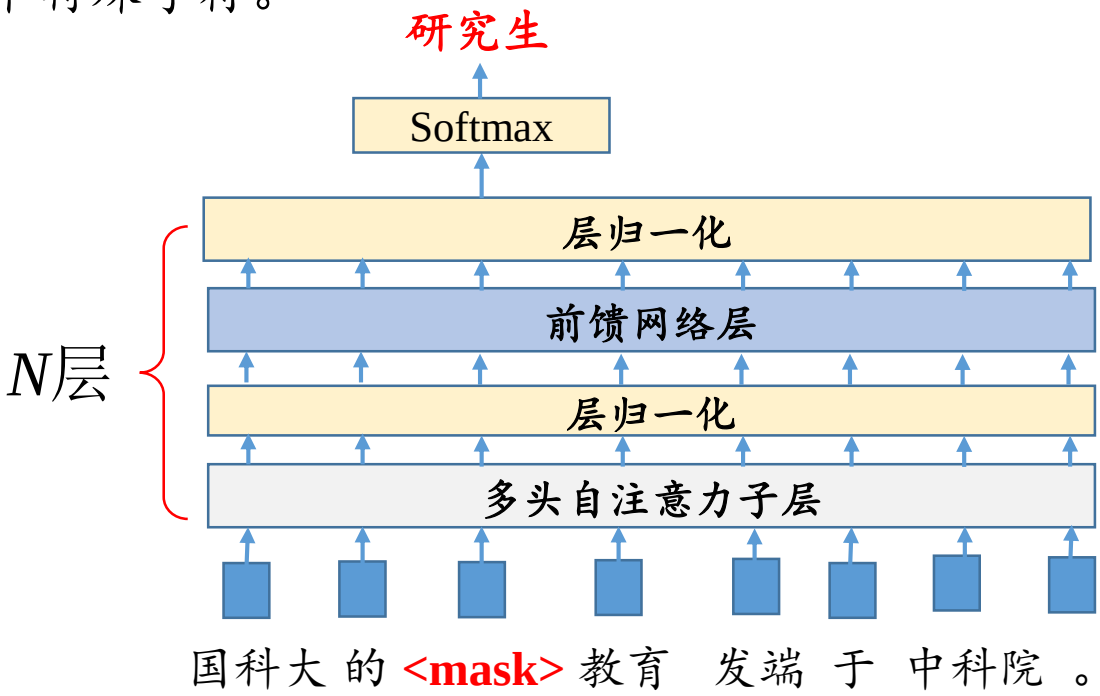


Transformer的编码器

3. 情感与情绪分析

(a)任务1：掩码语言模型（masked language model）

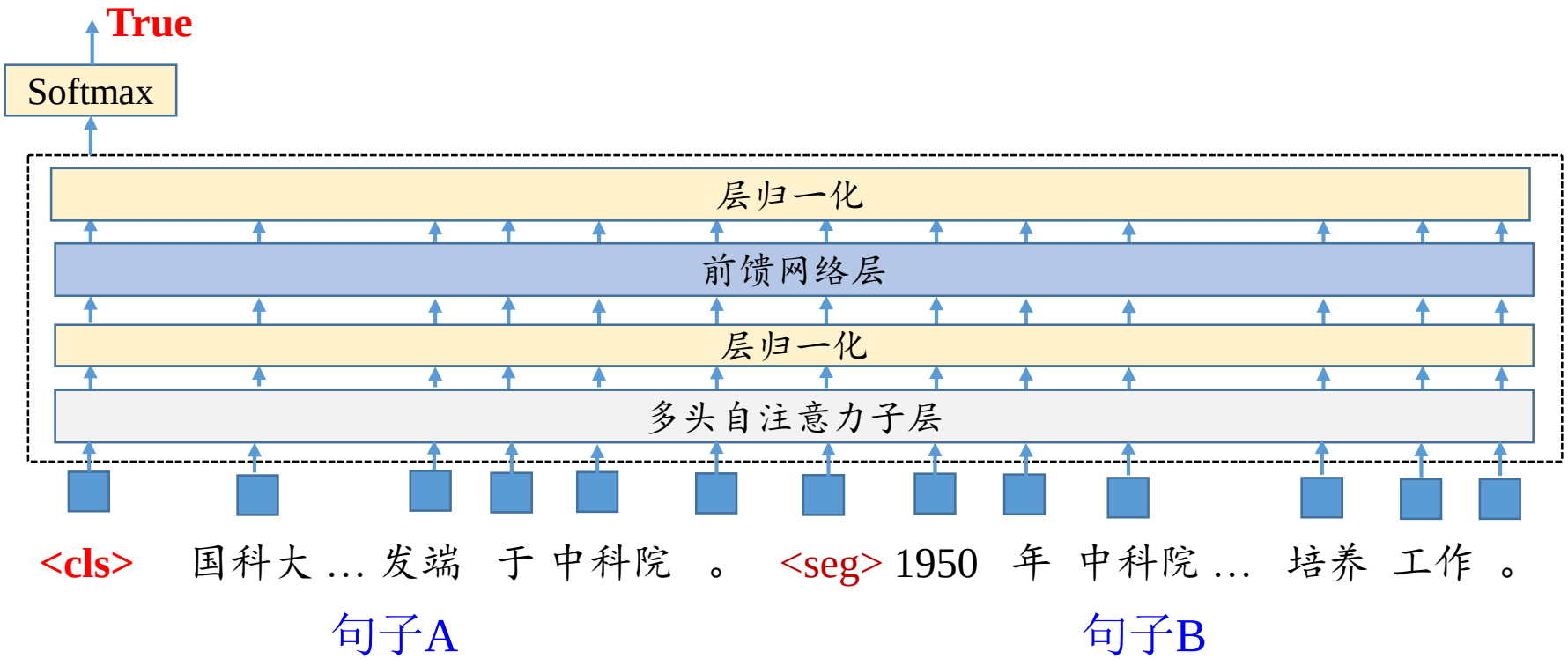
- 随机地将一句话中某个单词替换成<mask>，输入*N*层Transformer编码器；
- 将<mask>的隐层状态输入 Softmax 进行预测，正确预测结果为该单词；
- 根据预测结果和正确结果的交叉熵，更新参数；
- <mask>是词表中的一个特殊字符。



3. 情感与情绪分析

(b)任务2：句子衔接预测（Next sentence prediction）

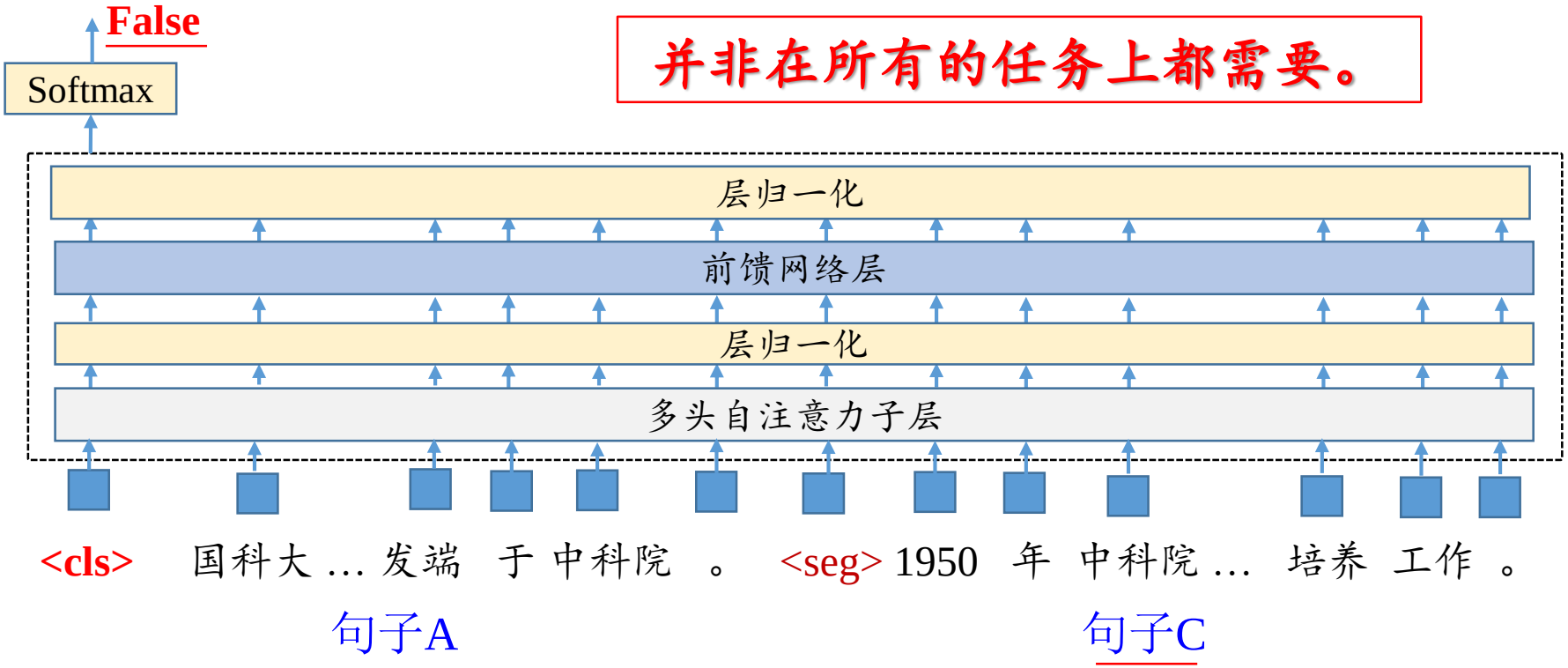
- 将大规模语料库中任意的连续两个句子(A和B)进行拼接；
- 句子前端加入<cls>，两个句子之间加入 <seg> ， <cls>和<seg>均为特殊字符；
- 将<cls>的隐层状态输入 Softmax 进行预测，正确标签为True (即B为A的下一句)；
- 根据预测结果与正确结果的交叉熵，更新参数。



3. 情感与情绪分析

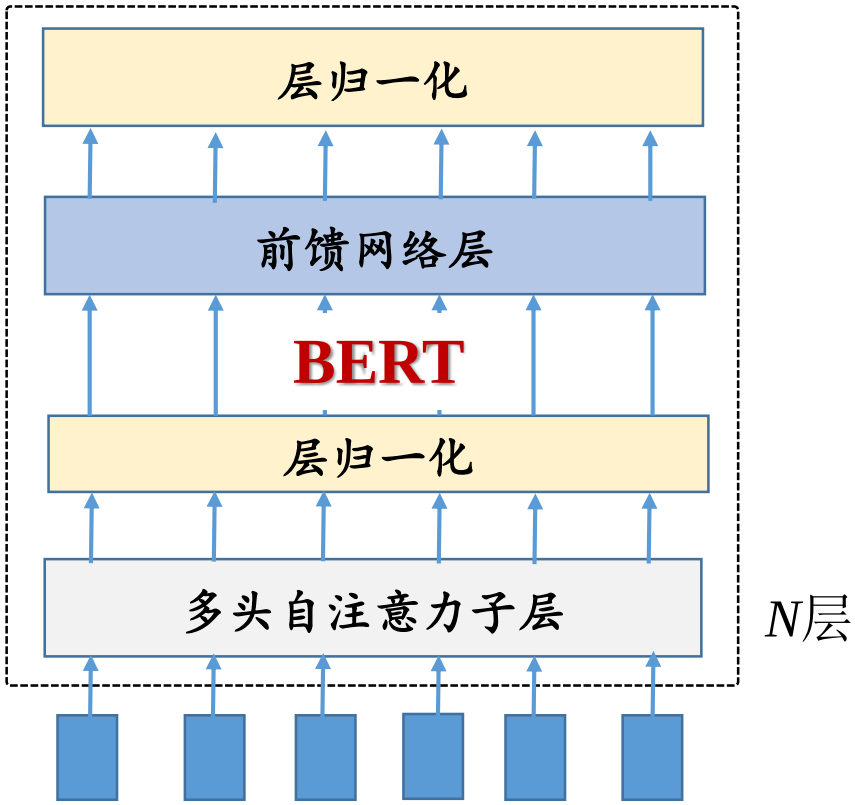
(b)任务2：句子衔接预测（Next sentence prediction）

- 将大规模语料库中任意的连续两个句子(A和B)进行拼接；
- 句子前端加入<cls>，两个句子之间加入 <seg> ， <cls>和<seg>均为特殊字符；
- 将<cls>的隐层状态输入 Softmax 进行预测，正确标签为True (即B为A的下一句)；
- 根据预测结果与正确结果的交叉熵，更新参数。



3. 情感与情绪分析

- 利用大规模语料，反复迭代**掩码语言模型任务**和**句子衔接预测任务**，优化模型参数，当模型收敛或达到最大迭代次数时，即得到最终的BERT。
- 句子经过 BERT 得到的表示为通用表示。**通用表示不能直接用于分类**。

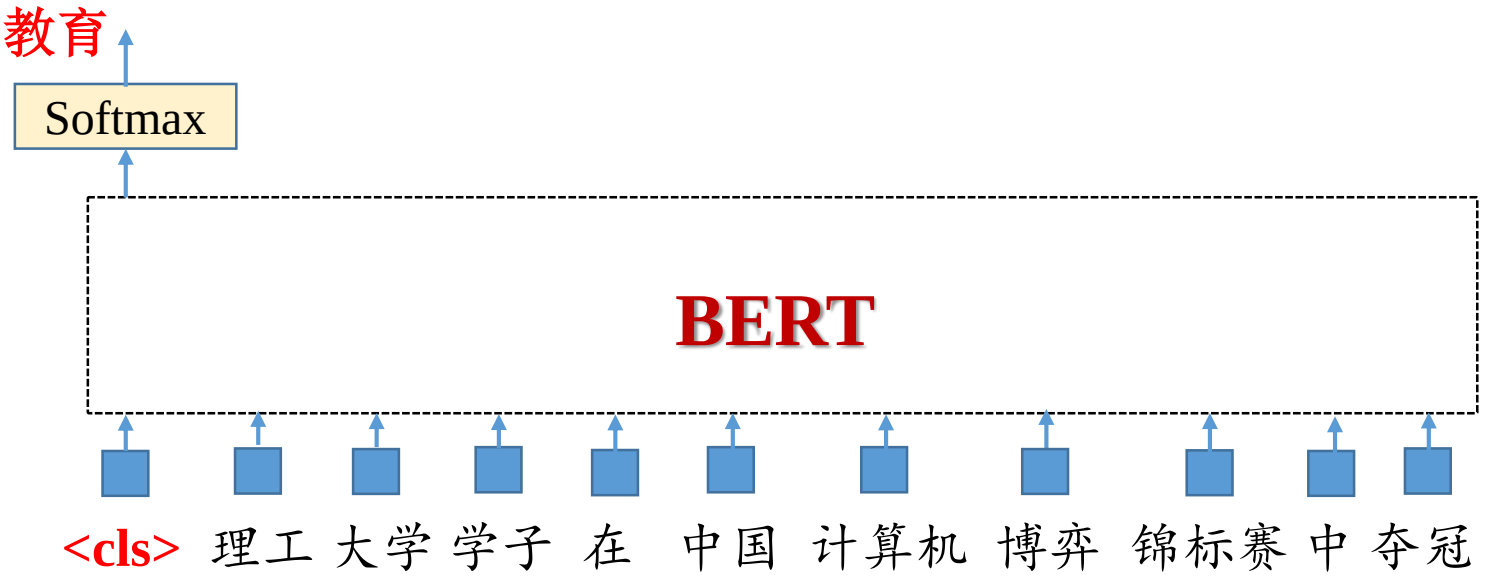


3. 情感与情绪分析

Step2 微调(fine-tuning) — **基于内容的文本分类**

“**句子1：** 理工 大学 学子 在 中国 计算机 博弈 锦标赛 中 夺冠 **类别：** 教育”

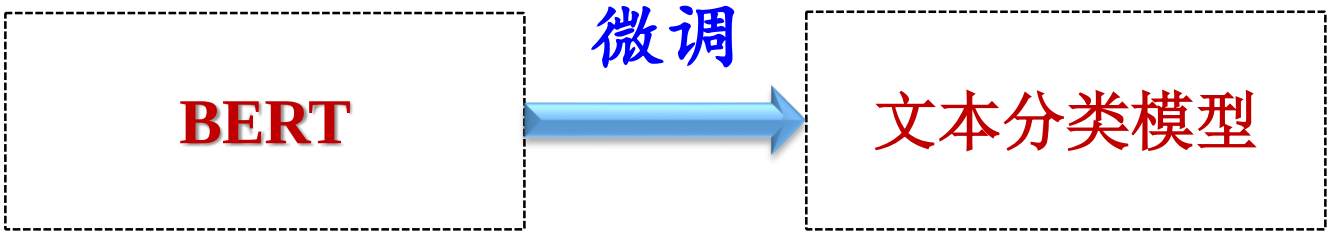
- 将该句子前面加入<cls>后，输入 BERT；
- 将<cls>的隐层状态输入 Softmax 进行预测，正确标签为教育；
- 根据预测标签和正确标签的交叉熵，更新BERT中的所有参数。



3. 情感与情绪分析

Step2 微调(fine-tuning) — 基于内容的文本分类

- 利用少量的基于内容的文本分类标注数据进行少量次数的迭代，即可以将BERT模型微调为一个文本分类模型；



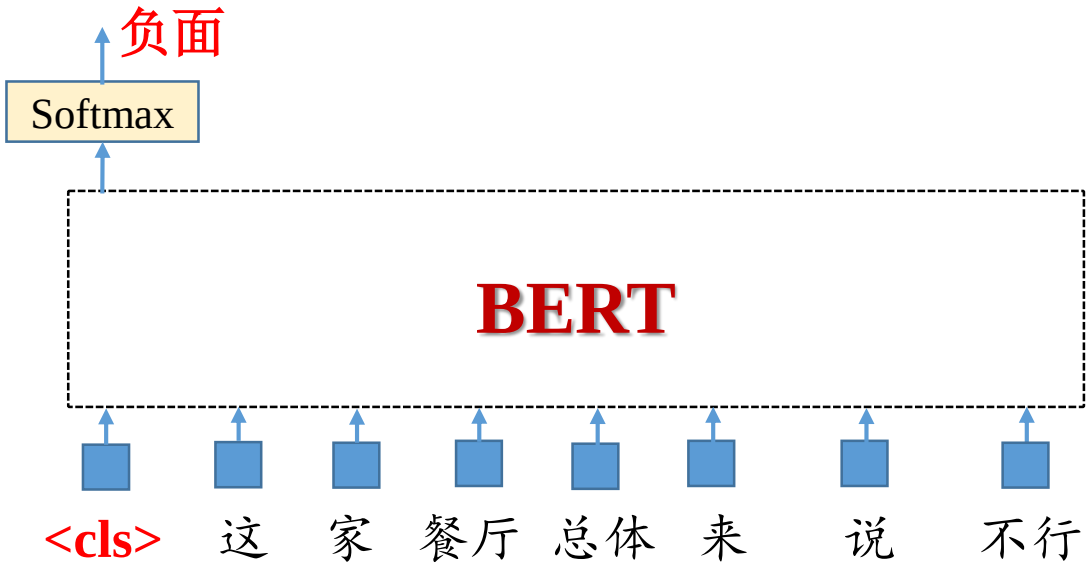
3. 情感与情绪分析

Step2 微调(fine-tuning) — 情感文本分类

“**句子1：** 这 家 餐厅 总体 来 说 不行　**情感类别：** 负面”

执行与文本内容分类同样的微调过程：

- 将该句子前面加入<cls>后，输入BERT。
- 将<cls>的隐层状态输入 Softmax 进行预测，正确标签为负面；
- 根据预测标签和正确标签的交叉熵，更新BERT中的所有参数。



Refer to:

J. Devlin et al. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proc.* NAACL-HLT, pp. 4171–4186

3. 情感与情绪分析

◆方法对比

●不同方法各有利弊

- 基于词典的方法具有领域独立性，但缺乏领域词典，因此效果不如有监督的统计学习方法；
- 基于语料资源的有监督统计学习方法和神经网络方法都受到领域和语料本身及标注质量的限制，尤其对于多层次属性而言，其划分的粒度对性能影响至关重要
- 多种方法结合使用在一定程度上可以实现优势互补

●对于句子级情感识别，仍面临很多挑战

- 如何处理比较句，如：诺基亚5800比5230更超值
- 如何处理否定词，尤其多重否定
- 对于较短的、缺乏足够上下文的句子情感分析困难
- 隐式表达的情感极性常常具有判断陷阱

……

本章内容

1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

4. 资源与评测

◆情感词典：英文

- General Inquier (http://www.wjh.harvard.edu/~inquirer/)
  - Manually labeled terms (positive, negative)
- SentiWordnet (http://sentiwordnet.isti.cnr.it/)
  - Extend from WordNet
  - Each synset is automatically labeled as P, N, O
- OpinionFinder's Subjectivety Lexicon(http://www.cs.pitt.edu/mpqa/)
  - Subjective words provided by Opinion Finder
- Taboada and Grieve's Turney adjective list
  - Available through Yahoo Sentiment AI group. 1700 words
- IBM Lexicon
  - 1,267 positive words and 1,701 negative words (Melville 2009)

4. 资源与评测

◆情感词典：中文

- HowNet (http://www.keenage.com/html/e_index.html)

  正面情感、负面情感、正面评价、负面评价、程度级别、主张词语6个子集

- 大连理工大学：情感词汇本体库

4. 资源与评测

◆情感分析评测

- **TREC Blog Track** (starting from 2006) (http://trec.nist.gov/)
  - Task: Opinion Retrieval and Polarity Identification
  - Corpus: 3,000,000 English webpages

- **NTCIR** (http://research.nii.ac.jp/ntcir/)
  - **Tasks:**
    ① Topic Relevance
    ② Opinion identification
    ③ Polarity Identification
    ④ Opinion Holder extraction
    ⑤ Opinion Target extraction
  - **Corpus:** news articles (English, Chinese, Japanese, Korea)

4. 资源与评测

◆情感分析评测

- **Chinese** (COAE 2008-2015)

  - Tasks:

  ① Words level (sub/obj, positive/negative)
  ② Documents level (sub/obj, positive/negative)
  ③ Opinion target extraction
  ④ Opinion retrieval

  - Corpus: Chinese (商品属性语料)

  口碑网，it168，494 documents, 5 domains

4. 资源与评测

◆ 部分公开的数据集

数据集	任务	简介	语言	链接
IMDb	句子级情感分类	包括来自Internet Movie Database（IMDb）的50000条电影评论，标注了正面和负面两类情感。每部电影的评论不超过30条。	EN	http://www.cs.cornell.edu/people/pabo/movie-review-data/ https://huggingface.co/datasets/imdb https://datasets.imdbws.com
SST	句子级情感分类（SST-2二分类、SST-5五分类）	斯坦福情感树库数据主要来自电影评论，每个句子分析树的节点均有细粒度的情感标注。总数据量为11855个电影评论句子，句子的解析树中包含215154个带有精细情感标签的短语。可以用于情感二分类和细粒度的情感五分类。	EN	https://nlp.stanford.edu/sentiment/index.html
Yelp	句子级情感分类	Yelp评论数据集包含超过50万条Yelp评论。数据集有二分类版本和细粒度五分类版本。	EN	https://www.yelp.com/dataset
中文微博情感分析测评数据	句子级情感分类	数据来自腾讯微博，包括 20 个话题，每个话题大约1000条微博，共约20000条微博。	ZH	https://pan.baidu.com/s/1psjysSXpKOEb1ciem7DsRw 密码：7hb4　原址： http://tcci.ccf.org.cn/conference/2012/pages/page10_dl.html http://tcci.ccf.org.cn/conference/2013/pages/page04_tdata.html

4. 资源与评测

数据集	任务	简介	语言	链接
SemEval-14	属性级情感分析 任务 1:属性词抽取 任务 2: 属性词情感分类 任务 3: 属性类别检测 任务 4: 属性类别情感分类	SemEval-14包含餐馆和笔记本电脑两个领域超过6000条标有细粒度属性及其对应情感极性的句子。其中餐馆领域标注了属性词及其对应情感、属性类别及属性类别情感，可用于四个子任务；笔记本电脑领域仅标注了属性词及其相应的情感，只能用于子任务1和2。	EN	任务介绍：https://aclanthology.org/S14-2004.pdf 官方网站：https://alt.qcri.org/semeval2014/task4/ 数据下载地址：https://github.com/songyouwei/ABSA-PyTorch/tree/master/datasets/semeval14
Sentihood	属性级情感分析	该数据集用于基于目标属性的情感分析的数据集，旨在识别特定属性的细粒度情感极性。数据集包含5,215个句子，其中3,862个句子包含一个目标其余句子包含多个目标。	EN	数据集描述：https://aclanthology.org/C16-1146.pdf 数据下载地址：https://github.com/uclmr/jack/tree/master/data/sentihood
ASAP	属性级情感分析	包括多达46,730条电子商务平台餐馆领域的评论。预先定义了18个属性类别，并标注了属性类别对应的五类情感极性，同时还标注了5分制的评论整体评分。	ZH	数据集相关论文：https://arxiv.org/abs/2103.06605 数据下载地址：https://github.com/Meituan-Dianping/asap

4. 资源与评测

数据集	任务	简介	语言	链接
AffectiveText	情绪分类	基于新闻标题的Ekman基本情绪的分类和情感效价的预测。	EN	http://web.eecs.umich.edu/~mihalcea/affectivetext/
CBET	情绪分类	基于推特语料的文本情绪分类。	EN	http://www.cs.ualberta.ca/~zaiane/CBET/
CEmo	情绪分类	面向健康领域的情绪分类。	EN	https://github.com/tsosea2/CancerEmo
SemEval-18 Affect in Tweets	① 情绪强度回归； ② 情绪强度分类； ③ 情绪效价回归； ④ 情绪效价分类； ⑤ 情绪分类。	基于阿拉伯语、英语和西班牙语的推特语料的情感分析系列任务。	AR, EN, EN	https://aclanthology.org/S18-1001/
GoEmotions	细粒度情绪分类	基于Reddit英语评论的细粒度情绪分类任务。	EN	https://github.com/google-research/google-research/tree/master/goemotions
EmoBank	情感VAD属性预测	基于维度模型的文本语料。	EN	https://github.com/JULIELab/EmoBank

本章内容

1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

5. 习题

1. 请从新浪或其他门户网上收集一定规模的不同类别文档，进行分类整理，利用不同的特征和不同的分类器实现文本内容分类，并对不同的方法进行对比实验。
2. 请从Twitter或新浪微博等平台上收集一定规模的短文本，进行分类整理，并利用这些整理后的语料进行分类方法实验。分析比较在短文本和长文本上文本分类任务面临的问题及分类方法的性能差异。
3. 请尝试从京东或淘宝网上收集商品的评论语料，并利用这些数据进行句子或文档级的情感类别标注，抽取情感词汇。
4. 尝试将句子中词汇依存关系分析的结果与文本情感分析相结合，在上述语料或其它公开的评测语料上进行方法性能对比分析。
5. 利用公开的情感或情绪分析评测语料，对比分析统计学习方法和神经网络方法的性能差异，并进行错误分析。

5. 习题

6. 搜狗实验室在2006年公布的一批文本分类的数据集，包括汽车、财经、IT、健康、体育、旅游、教育、招聘、文化、军事等9个领域，各1990个文件（见SEP平台本课程资源的附录）。请利用这些数据实验对比不同分类器的性能，并进行特征、分类器组合分析。
7. 中国科学院计算技术研究所谭松波老师整理了一批篇章级情感分类语料，包含图书评论、酒店评论和笔记本评论三大领域（见SEP平台本课程资源的附录）。请利用这些数据进行情感分析方法对比实验。

项目作业：从第1~5题中任选一题，完成技术报告。

本章小结

- 文本内容分类
- 情感文本分析
  - 任务细分：情感识别；观点要素抽取；观点检索；
    情绪分析；情绪-原因对抽取
  - 词汇级、句子级、文档级情感识别方法
- 基本方法
  - 基于词典；统计方法(特征+分类器)；深度学习(表示/预训练+Softmax)
- 相关评测和资源
- 文本聚类

  基本聚类算法原理
- 延伸阅读

本章内容

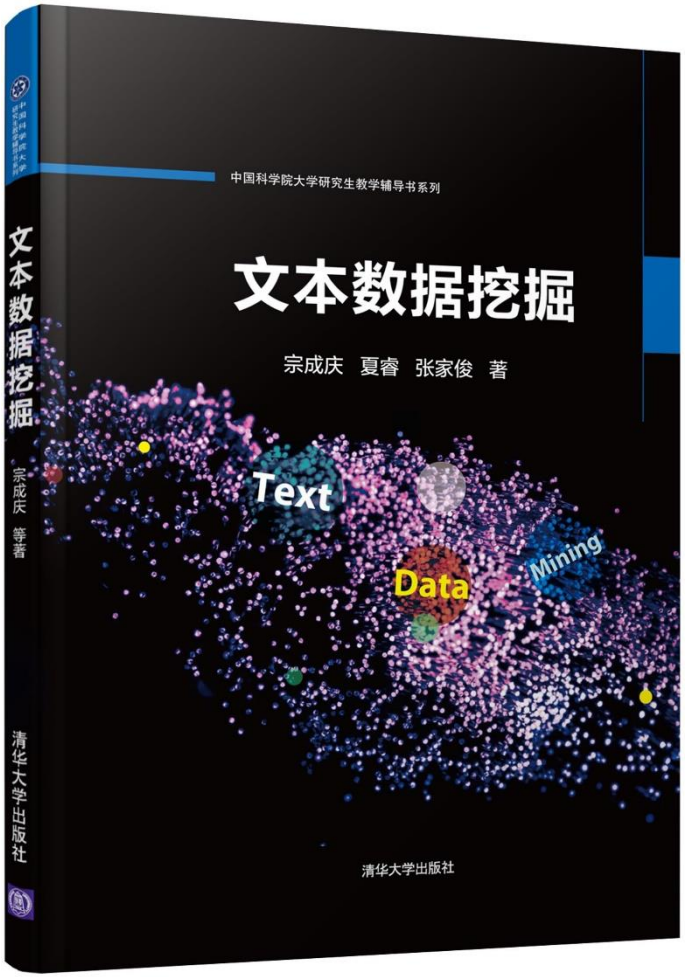
1. 文本分类
2. 文本聚类
3. 情感与情绪分析
4. 资源与评测
5. 习题
6. 附录：延伸阅读

6. 附录：延伸阅读

◆ 关于文本内容或情感分类的专著及综述

1. Bing Liu. 2012. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.
2. Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135.
3. Kumar Ravi and Vadlamani Ravi. 2015. A survey on opinion mining and sentiment analysis:
   Tasks, approaches and applications. *Knowledge-Based Systems*, 89(C):14–46.
4. Mika Viking Mantyla, Daniel Graziotin, and Miikka Kuutila. 2016. The evolution of sentiment analysis - a review of research topics, venues, and top cited papers. *arXiv preprint arXiv:1612.01556*
5. Fabrizio Sebastiani. 2002. Machine Learning in Automated Text Categorization. ACM *computing surveys*.
6. 宗成庆等，《文本数据挖掘》，清华大学出版社，2019(1)；2021(2)

6. 附录：延伸阅读


中国科学院大学研究生教学辅导书系列
文本数据挖掘
宗成庆 夏睿 张家俊 著
Text
Data
Mining
清华大学出版社
文本数据挖掘
宗成庆 等著
清华大学出版社


中国科学院大学研究生教学辅导书系列
文本数据挖掘
（第2版）

6. 附录：延伸阅读

◆ 关于文本分类的论文

1. George Forman. An extensive empirical study of feature selection metrics for text classification. 2003. *JMLR*, 3(2):1289-1305
2. Genkin A, Lewis D D, Madigan D. 2007. Large-Scale Bayesian Logistic Regression for Text Categorization. *Technometrics*, 49(3):291-304
3. Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, Chris Watkins. 2002. Text classification using string kernels. *Journal of Machine Learning Research*, 2(3):419-444
4. Andrew McCallum, Kamal Nigam. 1998. A Comparison of Event Models for Naive Bayes Text Classification. *Proc. AAAI*
5. Andrew McCallum, Nigam K. 1998. Employing EM and Pool-Based Active Learning for Text Classification. *Proc. ICML*
6. Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, Michell Tom. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2):103-134

6. 附录：延伸阅读

7. Joachims T. 1998. Text categorization with support vector machines: Learning with many relevant features. *Proc. European conference on machine learning*
8. Schapire R E, Singer Y. 2000. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning*, 39(2):135-168
9. William B. Cavnar and John M. Trenkle. 1994. N-Gram-Based Text Categorization. *Ann Arbor MI 48113.2*: 161-175
10. Yiming Yang, Jan O Pedersen. 1997. A comparative study on feature selection in text categorization. *Proc. ICML*
11. Yiming Yang, Xin Liu. 1999. A re-examination of text categorization methods. *Proc. SIGIR*
12. Yiming Yang. 1999. An evaluation of statistical approaches to text categorization. *Information Retrieval*, 1999, 1(1-2): 69-90

6. 附录：延伸阅读

◆ 关于情感/情绪分析的论文

1. John Blitzer et al. 2007. Biographies, bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. *Proc. ACL*, pp. 440–447
2. Michael Gamon. 2004. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. *Proc.* COLING
3. Xavier Glorot et al. 2011. Domain Adaptation for Large-Scale Sentiment Classification: A Deep Learning Approach. *Proc. ICML*
4. Alec Go et al. 2009. Twitter Sentiment Classification using Distant Supervision. CS224N Project Report, Stanford University, 2009, 1(12)
5. Alistair Kennedy, Diana Inkpen. 2006. Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2): 110-125
6. Shoushan Li et al. 2011. Semi-supervised Learning for Imbalanced Sentiment Classification. *Proc. IJCAI*

6. 附录：延伸阅读

7. Sinno Jialin Pan et al. 2010. Cross-Domain Sentiment Classification via Spectral Feature Alignment. *Proc. WWW*
8. Bo Pang et al. 2002. Thumbs up?: sentiment classification using machine learning techniques. *Proc. EMNLP*, pp. 79–86
9. Richard Socher et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. *Proc. EMNLP*, pp. 1631–1642
10. Richard Socher et al. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. *Proc. EMNLP*, pp. 151–161
11. Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proc. ACL*, pp. 417–424
12. Xiaojun Wan. 2009. Co-training for cross-lingual sentiment classification. *Proc. ACL*
13. Sida Wang, Christopher Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. *Proc. ACL*

6. 附录：延伸阅读

14. Rui Xia et al. 2011. Ensemble of feature sets and classification algorithms for sentiment classification. *Information Sciences*, 181(6): 1138-1152
15. Richard Socher et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank *Proc. EMNLP*, pp 1631-1642
16. Li Dong et al. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. *Proc. ACL*, pp 49-54
17. Yequan Wang et al. 2016. Attention-based LSTM for aspect-level sentiment classification. *Proc. EMNLP*, pp 606-615
18. Kai Sheng Tai et al. 2015. Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. *Proc. ACL*, pp 1556-1566
19. Zhiyang Teng et al. 2016. Context-sensitive lexicon features for neural sentiment analysis. *Proc. EMNLP*, pp 1629-1638
20. Wang X, Zong C. 2021. Distributed representations of emotion categories in emotion space. Proc. ACL-IJCNLP, pp. 2364-2375

6. 附录：延伸阅读

21. Xia R, Ding Z. 2019. Emotion-cause pair extraction: a new task to emotion analysis in texts. arXiv preprint arXiv:1906.01267
22. Kusuma R M I et al. 2019. Using deep learning neural networks and candlestick chart representation to predict stock market. arXiv preprint arXiv:1903.12258
23. Sun C et al. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. *Proc.* NAACL-HLT, pp. 380-385
24. Xu H et al. 2019. BERT Post-Training for Review Reading Comprehension and Aspect-based Sentiment Analysis. *Proc.* NAACL-HLT
25. Demszky D, Movshovitz-Attias D, Ko J, et al. 2020. GoEmotions: A Dataset of Fine-Grained Emotions. arXiv preprint arXiv:2005.00547
26. Seoh R et al. 2021. Open Aspect Target Sentiment Classification with Natural Language Prompts. *Proc.* EMNLP, pp.6311-6322
27. Liu J et al. 2021. Solving Aspect Category Sentiment Analysis as a Text Generation Task. *Proc.* EMNLP, pp. 4406-4416
28. Li Z et al. 2021. Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training. *Proc.* EMNLP, pp. 246-256

谢谢！

Thanks!