



中國人民大學
RENMIN UNIVERSITY OF CHINA

计算传播理论与实务

2019-2020秋季学期

第三讲 文本分析

文本情感分析

授课教师：范举副教授、塔娜讲师

时间：2019年11月4日

上讲回顾：文本分析

- 3.1 文本分类
 - 朴素贝叶斯 Naïve Bayes
 - 分类效果评价
- 3.2 文本数据获取
 - 网页爬虫
 - 网页解析
- 3.3 文本分词与检索
 - 文本分词
 - 文本检索

什么是情感分析？

- 情感分析是让计算机拥有人的情感吗？NO！



什么是情感分析？

- 情感分析是让计算机超越人的情感吗？NO！



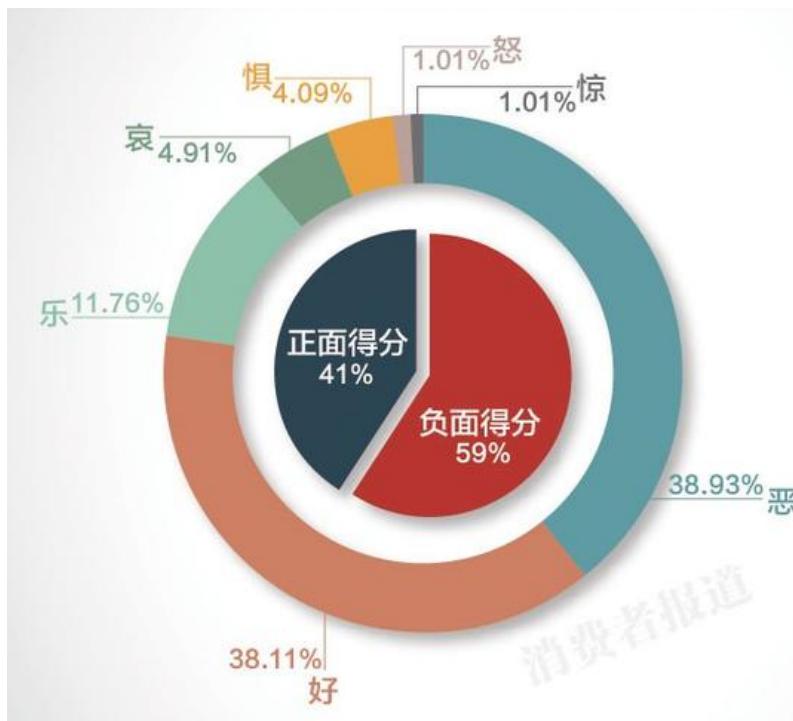
爱，只是一个字而已
但人类千秋和万代
不明白一直到现在
但 A.I.
能克服所有问题

——《A.I.爱》

什么是情感分析？

- 情感分析是让计算机.....

识别人的情感



- Sentiment Analysis
- 对带有情感色彩的主观性文本进行分析、处理、归纳和推理，
- 最后得出文本的二元化或者多元化极性分类。

Scherer Typology of Affective States

- **Emotion:** brief organically synchronized ... evaluation of a major event
 - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood:** diffuse non-caused low-intensity long-duration change in subjective feeling
 - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances:** affective stance toward another person in a specific interaction
 - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes:** enduring, affectively colored beliefs, dispositions towards objects or persons
 - *liking, loving, hating, valuing, desiring*
- **Personality traits:** stable personality dispositions and typical behavior tendencies
 - *nervous, anxious, reckless, morose, hostile, jealous*

Scherer 情感状态类型

- **情绪 (emotion)** : 有一定原因引发的同步反应
 - 例如悲伤 (sadness) 、快乐 (joy)
- **心情 (mood)** : 没有明显原因引发的长期低强度的主观感受变化
 - 例如忧郁 (gloomy) 、倦怠 (listless)
- **人际立场 (interpersonal stance)** : 对他人的特定反应
 - 例如疏远 (distant) 、冷漠 (cold)
- **态度 (attitude)** : 对特定人或事物的带有主观色彩的偏好或倾向
 - 喜欢 (like) 、讨厌 (hate)
- **个性特质 (personal traits)** : 相对稳定的个性倾向和行为趋势
 - 例如焦虑 (nervous) 、渴望 (anxious)

Scherer 情感状态类型

- **情绪 (emotion)** : 有一定原因引发的同步反应
 - 例如悲伤 (sadness) , 快乐 (joy)
- **心情 (mood)** : 没有明显原因引发的长期低强度的主观感受变化
 - 例如忧郁 (gloomy) , 倦怠 (listless)
- **人际立场 (interpersonal stance)** : 对他人的特定反应
 - 例如疏远 (distant) , 冷漠 (cold)
- **态度 (attitude)** : 对特定人或事物的带有主观色彩的偏好或倾向
 - 喜欢 (like) , 讨厌 (hate)
- **个性特质 (personal traits)** : 相对稳定的个性倾向和行为趋势
 - 例如焦虑 (nervous) , 渴望 (anxious)

什么是情感分析

- 情感分析是指从文本（或其它数据）中检测出人的**态度**的技术
 - 态度 (attitude) : 对特定人或事物的带有主观色彩的偏好或倾向
 - *enduring, affectively colored beliefs, dispositions towards objects or persons*

什么是情感分析

- 情感分析有很多别名
 - 意见抽取 (Opinion extraction)
 - 意见挖掘 (Opinion mining)
 - 情感挖掘 (Sentiment mining)
 - 主管倾向分析 (Subjectivity analysis)

案例：商品评论情感分析



HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$89 online, \$100 nearby ★★★★☆ 377 reviews

September 2010 - Printer - HP - Inkjet - Office - Copier - Color - Scanner - Fax - 250 sh

Reviews

Summary - Based on 377 reviews

1 star 2 3 4 stars 5 stars

What people are saying

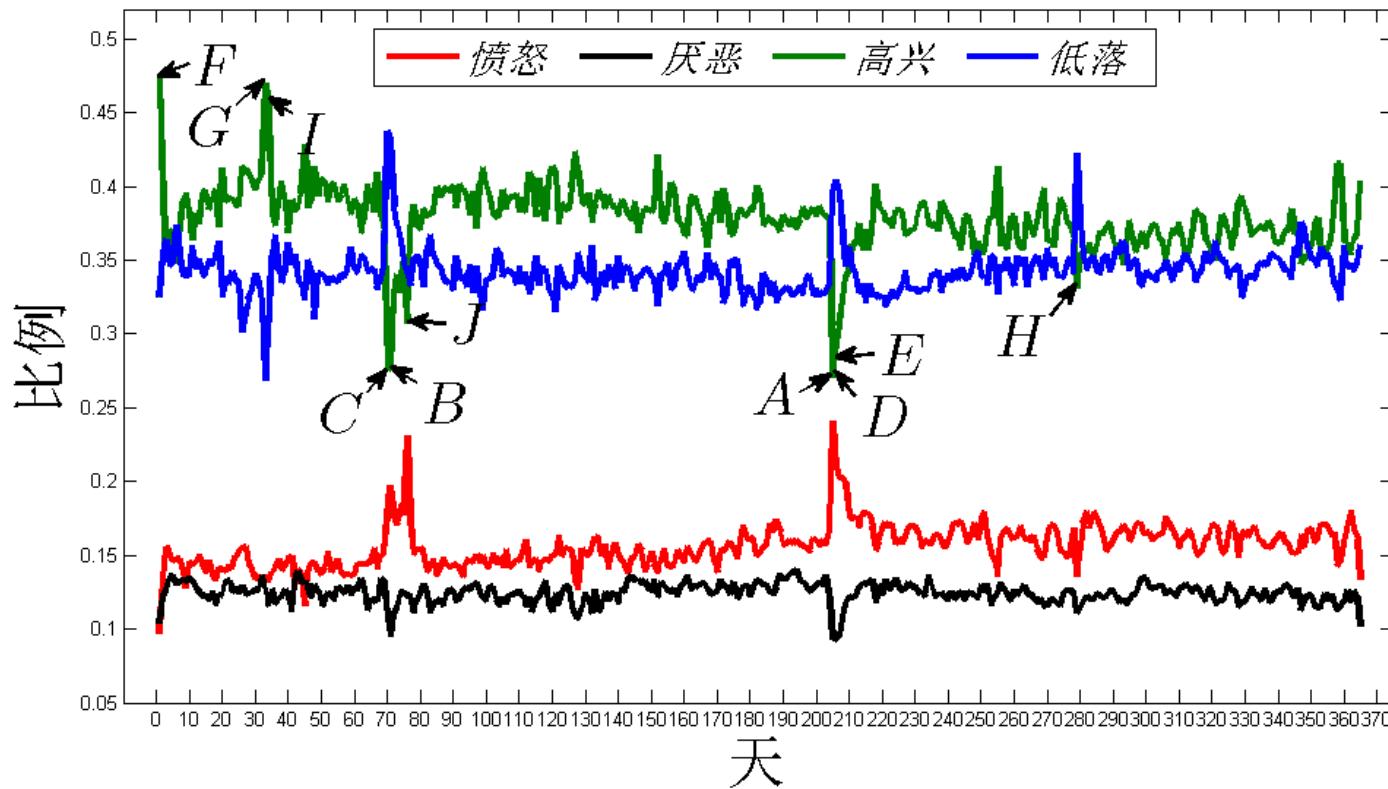
ease of use	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"This was very easy to setup to four computers."
value	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"Appreciate good quality at a fair price."
setup	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"Overall pretty easy setup."
customer service	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"I DO like honest tech support people."
size	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"Pretty Paper weight."
mode	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"Photos were fair on the high quality mode."
colors	<div style="width: 100px; height: 10px; background-color: #ccc; border: 1px solid #ccc; margin-bottom: 5px;"></div>	"Full color prints came out with great quality."



案例：饭店评论情感分析

评论	打分
口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	2
菜品丰富质量好，服务也不错！很喜欢！	4
说真的，不晓得有人排队的理由，香精香精香精，拜拜！	2
菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	5
先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	4
哎，冲着推荐来的，后悔死了，上菜巨慢不说，服务态度还超级差，东西咸的要死，而且普遍的贵，谁来...	1
超级差评????上菜慢??服务态度超级差、先把汤上来喝饱了??完事菜就上了一个，找服务员催半...	1

案例：微博情感分析



- 北航的先进网络分析研究小组(GANA) 对收集到的2011年的近7000万条微博进行情感分析

案例：微博情感分析

- 如F对应新年，以高兴的情绪为主；
- G, I对应春节，也以高兴的情绪为主；
- A、D、E则对应动车事故，明显看到用户以低落悲伤的情感为主，同时愤怒的情绪比例也明显上升至全年最高；
- C、B和J对应日本3月份的地震，以低落的情绪为主，但对点，即2011年03月17日，愤怒的情感比例突然增加，这与当时的盛传碘盐被污染、盐荒等谣言有关；
- H对应苹果前CEO乔布斯逝世，以低落的情绪为主，但同时，愤怒的情绪比例极低，与前面的动车事件和碘盐谣言有明显的区别。

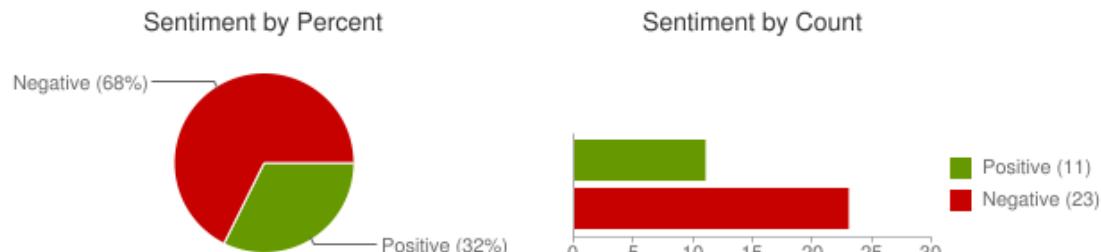
案例：Twitter情感分析

- Twitter Sentiment App

- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision

Type in a word and we'll highlight the good and the bad

Sentiment analysis for "united airlines"



jacobson: OMG... Could @United airlines have worse customer service? W8g now 15 minutes on hold 4 questions about a flight 2DAY that need a human.
Posted 2 hours ago

12345clumsy6789: I hate United Airlines Ceiling!!! Fukn impossible to get my conduit in this damn mess! ?
Posted 2 hours ago

EMLandPRGbelgiu: EML/PRG fly with Q8 united airlines and 24seven to an exotic destination. <http://t.co/Z9QloAjF>
Posted 2 hours ago

CountAdam: FANTASTIC customer service from United Airlines at XNA today. Is tweet more, but cell phones off now!
Posted 4 hours ago

情感分析的主要任务

- 情感分析是指从文本（或其它数据）中检测出人的**态度**的技术
 - 态度由谁持有（source）
 - 态度针对哪些方面（aspect）：饭菜味道、服务
 - 态度的类型（type）
 - 表达态度的文本范围：句子 or 整篇文档
- 态度的类型
 - like, love, hate, value, desire等
 - positive, negative, neutral
 - positive, negative

情感分析的主要任务

- 简单任务
 - 文章的整体感情是积极/消极的?
- 复杂任务
 - 对文章的态度从1-5打分
 - 在文章中识别高兴、愤怒、低落等多种情绪
- 更复杂的任务
 - 检测态度针对的不同侧面
 - 检测态度的持有者

你觉得应如何解决？

好评

菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...

差评

超级差评????上菜慢??服务态度超级差、先把汤上来喝饱了?? 完事菜就上了一个，找服务员催半...

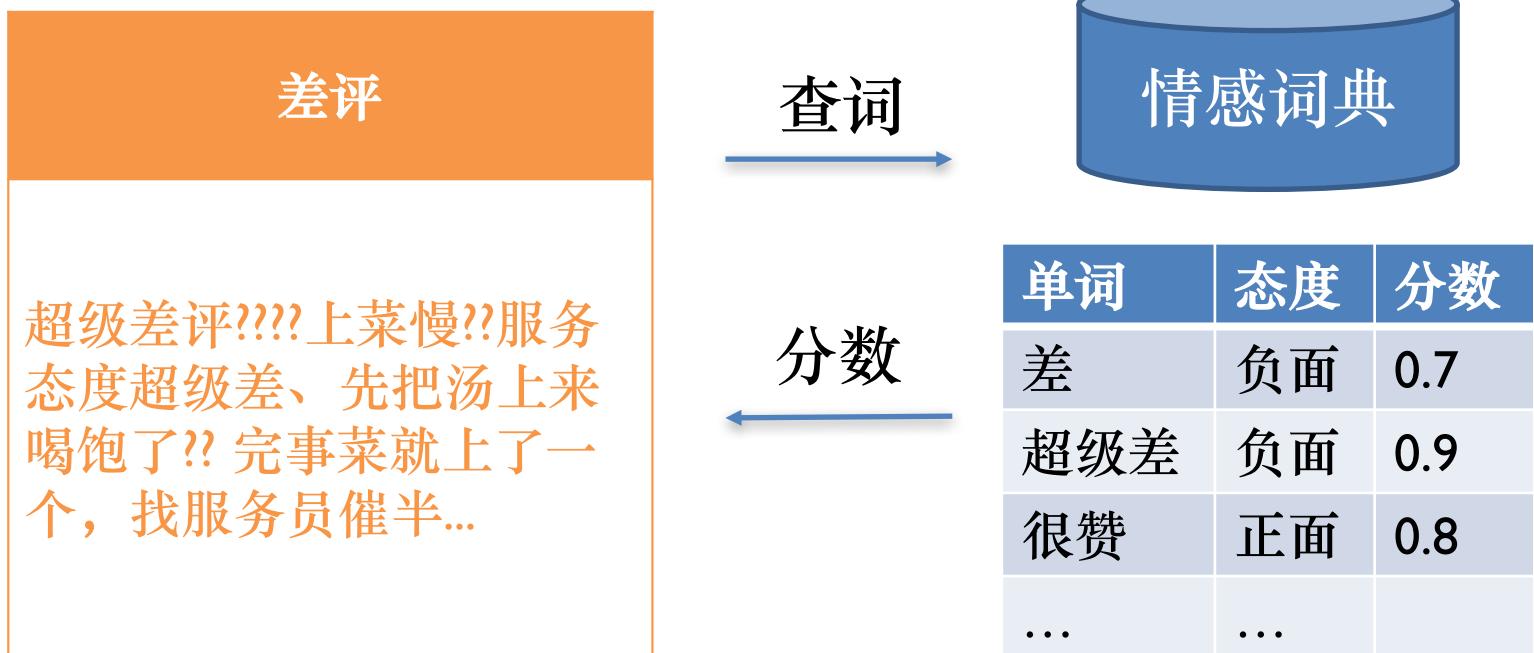
数据来源：大众点评 (www.dianping.com)

第3.4.1节 文本分类方法

基于情感词典的方法

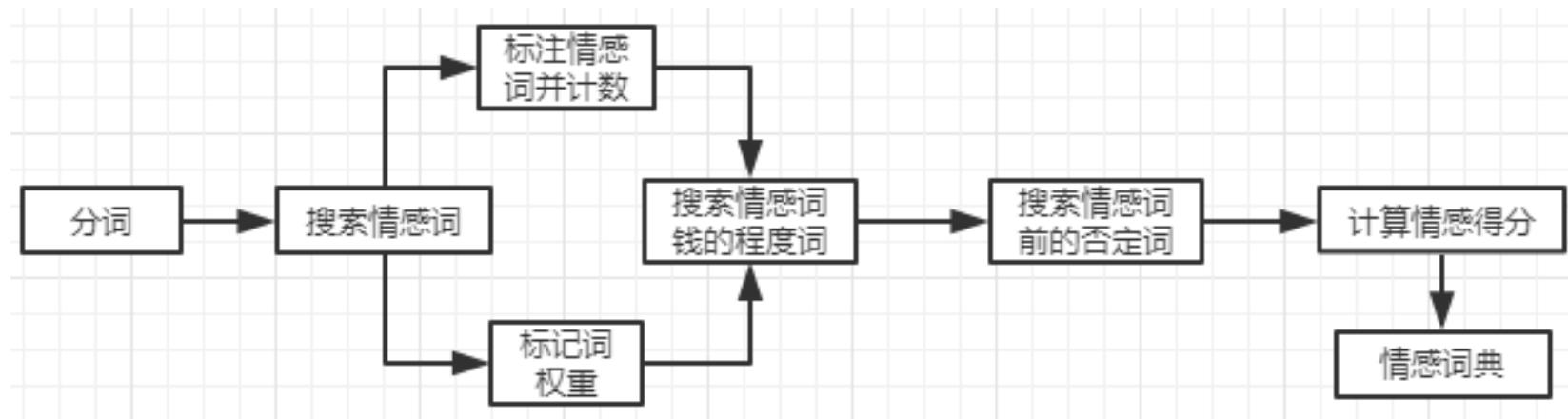
基于情感词典的方法

- 利用构建的文本情感词典，并对情感词典进行极性和强度标注，进而进行文本情感分类



基于情感词典的方法

- 利用构建的文本情感词典，并对情感词典进行极性和强度标注，进而进行文本情感分类



基于情感词典的方法

- 如何构造情感词典
 - 手工构造

情感	情感类
乐	快乐、安心
好	尊敬、赞扬、相信、喜爱
怒	愤怒
哀	悲伤、失望、疚、思
惧	慌、恐惧、羞
恶	烦闷、憎恶、贬责、妒忌、怀疑
惊	惊奇

基于情感词典的方法

- 如何构造情感词典
 - 手工构造：准确性、覆盖率上存在局限性
 - 基础扩展词典

狗血，泡沫剧，伤不起，腹黑.....

基于情感词典的方法

- 如何构造情感词典
 - 手工构造：准确性、覆盖率上存在局限性
 - 基础扩展词典

狗血，泡沫剧，伤不起，腹黑.....

- 表情符号词典



基于情感词典的方法

- 如何构造情感词典
 - 手工构造：准确性、覆盖率上存在局限性
 - 基础扩展词典

狗血，泡沫剧，伤不起，腹黑.....

- 表情符号词典



- 否定词和双重否定词词典

不，不可以，怎么不，几乎不，从来不，没，难以.....
绝非不，并非不，无不，不会不.....

情感词典资源

- 哈佛大学的General Inquirer Lexicon
- 匹兹堡大学提供的Opinion Finder主观情感词典
- 伊利诺伊大学Bing Liu提供的词典资源
- 台湾大学的中文情感极性词典(NTUSD)
- 知网情感词典HowNet
- 普林斯顿大学构建的英文情感词典WordNet

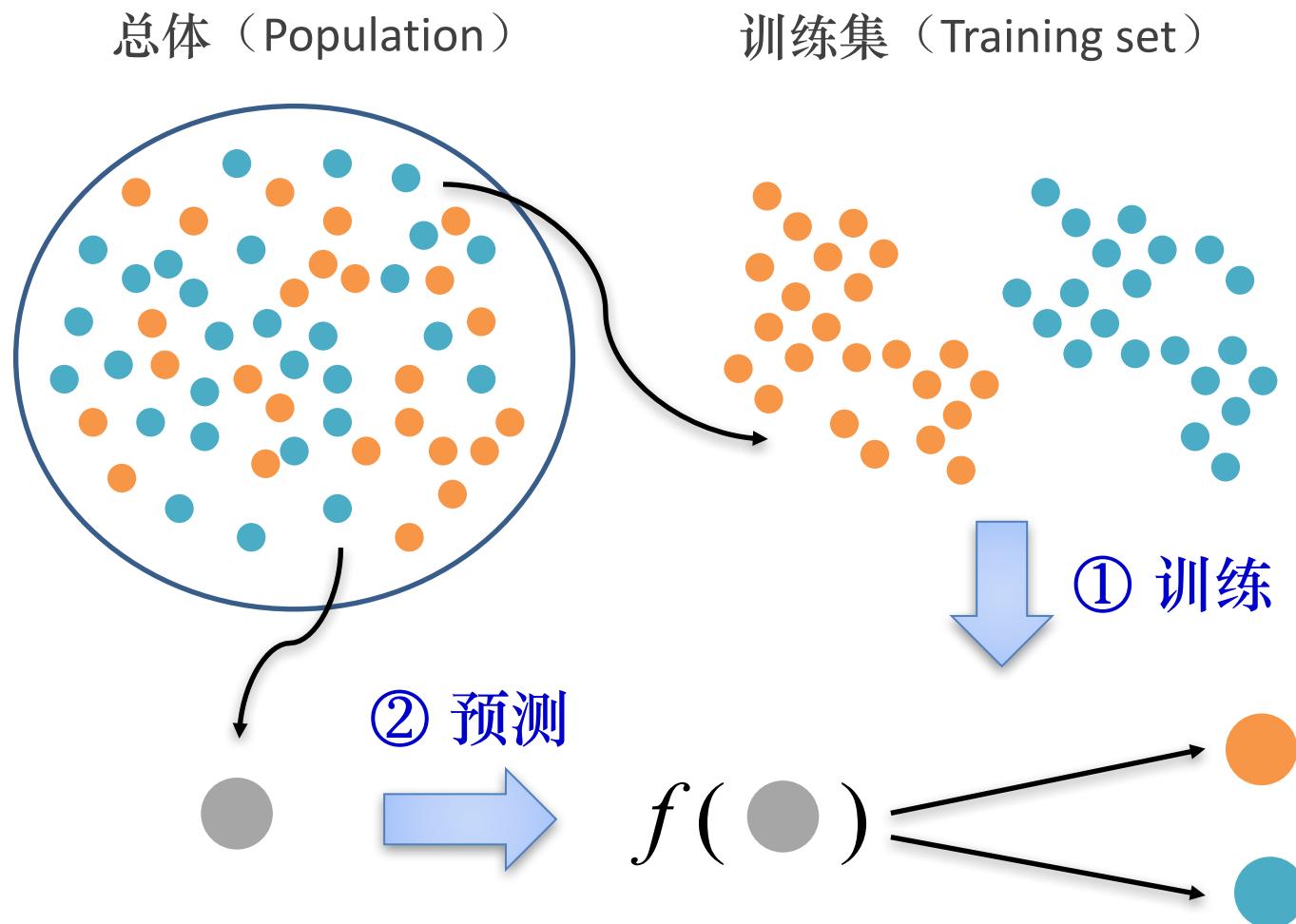
基于情感词典方法的局限性

- 情感词典中的情感词有较大程度的领域依赖性、时间依赖性和语言依赖性，同一词汇在不同的领域、时间和语言环境中可能会表达完全不同的情绪。
- 传统的情感词典方法还存在词典中情感词固定，难以及时捕捉新词、变形词的缺陷。
- 可以加入一些规则来提高情绪分类的准确率，但在数据量较大时，词典与规则的维护比较复杂，且不易扩展。

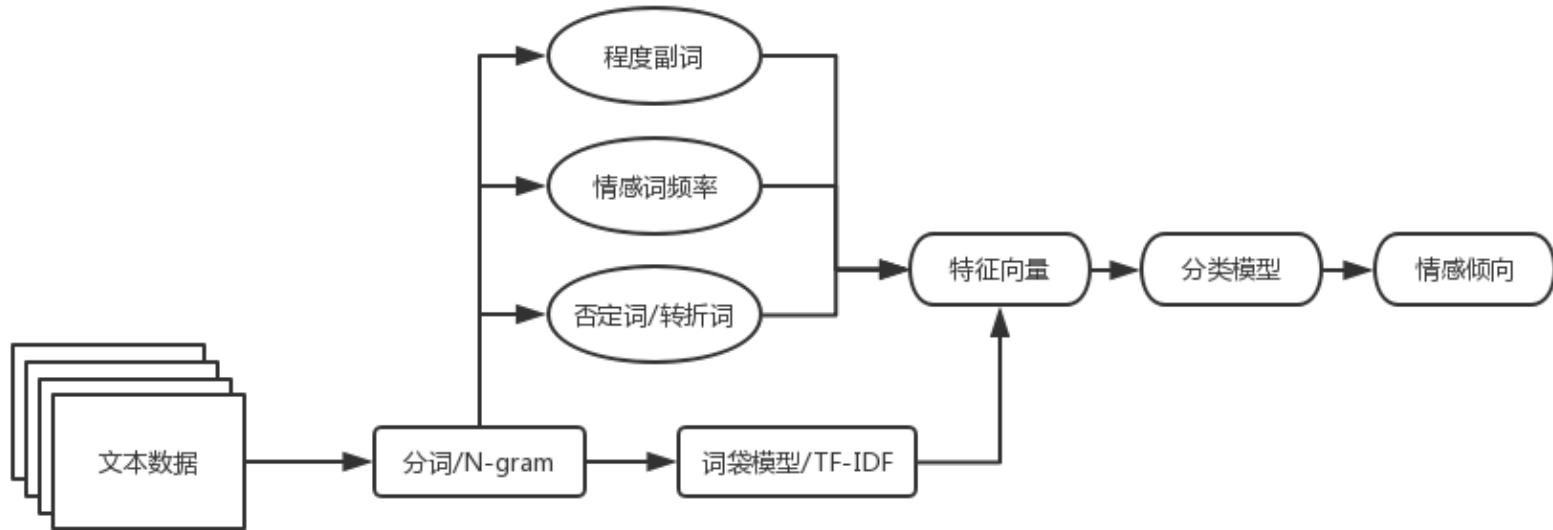
第3.4.2节 文本分类方法

基于机器学习的方法

回顾：监督学习的基本思路



基于机器学习的方法



词袋模型

为单词创建唯一标记

将文本以数值特征向量

形式表示

词频 \times 逆文档频率

特征提取

词级、句子级、篇章级

词袋模型

消极、积极情感词个数

转折词

分类算法

朴素贝叶斯算法

最大熵模型

支持向量机算法

半监督方法

词袋模型 Word Bags

- 基本目标：通过单词集合来表示一篇文档

ID	评论
0	饭店 味道 好 好 好
1	饭店 味道 差
2	饭店 位置 好



词典 Vocabulary
{饭店, 味道, 位置, 好, 差}



ID	表示
0	(1, 1, 0, 1, 0)
1	(1, 1, 0, 0, 1)
2	(1, 0, 1, 1, 0)

- 0 - 相应位置的词没出现
1 - 相应位置的词出现

TF-IDF模型

- 如何度量单词对文档的**重要性**
 - 一篇文档中重复出现的单词**很重要**
 - 大多数文档中都出现的单词**不重要**
- TF: term frequency
 - $tf(t, d) = \frac{Frquence(t, d)}{|d|}$
 - $Frquence(t, d)$: 单词t在文档d中出现了几次
 - $|d|$: 文档d总共出现了多少个单词（包含重复）
- IDF: inverse document frequency
 - $idf(t) = \log \frac{|D|}{1+df(t,f)}$
 - $df(t, f)$: 文档库D中有多少篇文档包含单词t
 - $|D|$: 文档库中总共有多少篇文档

TF-IDF模型

- TF: term frequency
 - $tf(t, d) = \frac{Frequency(t, d)}{|d|}$
 - $Frquence(t, d)$: 单词t在文档d中出现了几次
 - $|d|$: 文档d总共出现了多少个单词 (包含重复)
- IDF: inverse document frequency
 - $idf(t) = \log \frac{|D|}{1+df(t,f)}$
 - $df(t, f)$: 文档库D中有多少篇文档包含单词t
 - $|D|$: 文档库中总共有多少篇文档
- TF-IDF模型: $tf(t, d) \times idf(t)$

练习题

- 计算下面例子中的TF-IDF分值

ID	评论			
0	饭店	味道	好	好
1	饭店	味道	差	
2	饭店	位置	好	



词典 Vocabulary
{饭店, 味道, 位置, 好, 差}



ID	TF-IDF表示
0	
1	
2	

基于机器学习方法实战

- 最基础的情感分析算法 – 掌握原理 & 编程
- 步骤1：分词
 - 中文分词
- 步骤2：特征提取
 - 词袋模型 (word bags)
- 步骤3：应用分类模型
 - 朴素贝叶斯分类模型

基于机器学习方法实战

- 真实数据集dianping.csv – 来自大众点评

A	B
1 comment	star
2 口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢说菜都是提前做好的	2
3 菜品丰富质量好，服务也不错！很喜欢！	4
4 说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	2
5 菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口感绵软停不下来，	5
6 先说我是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐总是不如天津的某些家	4
7 菜很好吃??环境也不错?????值得推荐！！！！！！！！！！！！！！	5
8 很喜欢的餐厅，东北风味儿，以后还会再去！	4
9 口味属于家常菜一类，价格有点小贵，逛街的时候吃吃还可以。	4
10 最近太忙了，估计要暂时告别大众点评一段时间了。这条应该是这个月的最后一条点评了，全五星好评。在天津连续下雨的	5
11 我去了一次感觉喜欢上了这，无论环境还有菜品在自助餐里已经是很不错了	5
12 10点半的餐，12点还没下单，这速度也是醉了。味道还可以，德国香肠没有太好吃，焗土豆和柠檬鸡还是推荐的。	2
13 两个人吃没点太多，要了一份手抓饼当主食，没仔细看菜单，以为手抓饼是咸口的了??结果太甜了，油还有点大。当时特别	4
14 很一般?????地点好真的很重?????不是很推荐	2
15 我们两个人去的，等个100多个号吧，还行因为不着急，就特地为了去拔草，点了一直烤鸭，味道不错，就是上的特别慢，	2
16 哎，冲着推荐来的，后悔死了，上菜巨慢不说，服务态度还超级差，东西咸的要死，而且普遍的贵，谁来谁犯二，这辈子不	1
17 超级差评????上菜慢??服务态度超级差、先把汤上来喝饱了??完事菜就上了一个，找服务员催半天也上不来最后结账态度超	1
18 很好吃，中餐有西餐味。用餐人很多，我们排队了15分钟呢，爆	5
19 平日晚餐去的时间不算晚人也依旧那么火爆??等了一个小时??安排了一张超小的桌子??只适合下午茶用??很无语??点了奶酪	5
20 好吃好吃好吃，就是人不少！烤肉速度人多时候稍有些慢！	5
21 总体还不错??要的都是推荐菜??手撕饼有点腻	5
22 没有菜单??用微信点菜??还是先付款??什么都没吃着就把钱付了??忙活点餐付款就十几二十分钟??还真得不饿的时候来吃饭	1
23 上周末去的娜娜家，一直没来得及点评，今天趁着工作不忙赶紧来补一个??作为一个爱吃甜食的女生，真的是对这种有很多	5
24 服务员跟**似的，摇摇手，没位置，话都不说一句。	1
25 自从这家出了就一直想去，终于赶个闺蜜也有时间的日子早早的去了。不到五点半到那就已经排到10几号了，差不多等了半	5
26 东西什么的都挺好，味道也不错，就是人多点，环境很好	5
27 早就耳闻卤肉饭，可是不太喜欢吃，是甜口的，米饭太硬太硬了应该是隔夜的，肉也不是很烂！！烟花意面挺辣的，就是意	2

程序参考

sklearn-sa.ipynb

读取数据

- 如何从CSV文件中读取数据?
 - 使用pandas !

```
In [2]: data = pd.read_csv('datasets/dianping.csv')
data.head()
```

Out[2]:

		comment	star
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...		2
1	菜品丰富质量好，服务也不错！很喜欢！		4
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！		2
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...		5
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...		4

数据预处理

- 根据star属性设置sentiment属性
 - 三星以上 → positive；三星及以下 → negative

```
In [3]: def make_label(star):
    if star > 3:
        return 1
    else:
        return 0

data['sentiment'] = data.star.apply(make_label)
data = data[['comment', 'sentiment']]
data.head()
```

Out[3]:

	comment	sentiment
--	---------	-----------

0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	0
1	菜品丰富质量好，服务也不错！很喜欢！	1
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	1
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	1

小白版情感分析：snownlp库

```
In [4]: from snownlp import SnowNLP

text1 = '这个东西不错'
text2 = '这个东西很垃圾'

s1 = SnowNLP(text1)
s2 = SnowNLP(text2)

print(s1.sentiments,s2.sentiments)

def snow_result(comemnt):
    s = SnowNLP(comemnt)
    if s.sentiments >= 0.6:
        return 1
    else:
        return 0

data['snlp_result'] = data.comment.apply(snow_result)
data.head()
```

0.8623218777387431 0.21406279508712744

Out[4]:

	comment	sentiment	snlp_result
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	0	0
1	菜品丰富质量好，服务也不错！很喜欢！	1	1
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0	1
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	1	0
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	1	0

小白版情感分析：snownlp库

Out[4]:

		comment	sentiment	snlp_result
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...		0	0
1		菜品丰富质量好，服务也不错！很喜欢！	1	1
2		说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0	1
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...		1	0
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...		1	0

- 使用snownlp库的优缺点
 - 优点：简单方便
 - 缺点：准确率低

```
from snownlp import SnowNLP
text1 = '这个东西不错'
text2 = '这个东西很垃圾'
s1 = SnowNLP(text1)
s2 = SnowNLP(text2)
print(s1.sentiments,s2.sentiments)
```

分词：准备词袋模型

- 使用jieba分词进行简单分词

```
def simple_word_cut (texts):
    return ' '.join(jieba.cut(texts, cut_all=False))

data['simple_cut_comment'] = data.comment.apply(simple_word_cut)
data.head()
```

```
Building prefix dict from the default dictionary ...
Loading model from cache /var/folders/8d/62lh93xj51v9v1r3f8qg6m6r0000gn/T/jieba.cache
Loading model cost 0.975 seconds.
Prefix dict has been built successfully.
```

	comment	sentiment	simple_cut_comment
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	0	口味： 不知道 是我口高了， 还是这家真不怎么样。 ?? 我感觉 口...
1	菜品丰富质量好，服务也不错！很喜欢！	1	菜品 丰富 质量 好， 服务 也 不错！ 很 喜欢！
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0	说真的， 不晓得 有人 排队 的 理由， 香精 香精 香精 香精， 拜拜！
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	1	菜量 实惠， 上菜 还 算 比较 快， 疙瘩汤 喝 出 了 秋 日 的 暖 意， 烧 茄子 吃...
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	1	先说 我 算是 娜娜 家风 荷园 开业 就 一直 在 这里 吃 ?? 每次 出去 回 来 总...

去除停用词 (stopwords)

In [9]:

```
def word_cut (texts):
    words_list = []
    word_generator = jieba.cut(texts, cut_all=False) # 返回的是一个迭代器
    with open('hit_stopwords.txt') as f:
        str_text = f.read()
    for word in word_generator:
        if word.strip() not in str_text:
            words_list.append(word)
    #print ('1')
    return ' '.join(words_list) # 注意是空格

data['cut_comment'] = data.comment.apply(word_cut)

data.head()
```

Out[9]:

	comment	sentiment	cut_comment
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	0	口味 知道 我口 高 这家 不怎么样 感觉 口味 确实 很 很 上菜 相当 快 我敢 菜 都...
1	菜品丰富质量好，服务也不错！很喜欢！	1	菜品 丰富 质量 服务 不错 很 喜欢
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0	说真的 晓得 有人 排队 理由 香精 香精 香精 香精 拜拜
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	1	菜量 实惠 上菜 算 比较 快 疙瘩汤 喝出 秋日 暖意 烧茄子 吃 出 大阪 烧 味道 想...
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	1	先说 算是 娜娜 家风 荷园 开业 吃 每次 出去 回来 总想 吃 一回 有时 觉得 外面 ...

数据查看：绘制词云

- 绘制所有positive文档包含单词的词云

```
In [6]: def draw_wordcloud (words, color = 'white'):  
    wordcloud = WordCloud(stopwords = STOPWORDS,  
                          font_path="HYQiHei-105.ttf",  
                          background_color=color,  
                          width=2500,  
                          height=2000  
                      ).generate(words)  
    plt.figure(1,figsize=(13, 13))  
    plt.imshow(wordcloud)  
    plt.axis('off')  
    plt.show()  
  
    print("Positive words")  
  
data_pos = data[ data['sentiment'] == 1 ]  
words_pos = ' '.join(data_pos['cut_comment'])  
draw_wordcloud (words_pos)
```

数据查看：绘制词云

- 绘制所有positive文档包含单词的词云



数据查看：绘制词云

- 绘制所有negative文档包含单词的词云



提取特征

```
: X = data['cut_comment']
y = data['sentiment']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

: from sklearn.feature_extraction.text import TfidfVectorizer

vect = TfidfVectorizer(max_df = 0.8,
                      min_df = 3,
                      token_pattern=u'(?u)\\b[^\\d\\w]\\w+\\b'
                     )

features = pd.DataFrame(vect.fit_transform(X_train).toarray(), columns=vect.get_feature_names()
                        .head()
```

	app	ipad	ok	ps	wifi	一下	一个个	一个半	一个多	一人	...	麻将	麻烦	麻辣	麻酱	黄瓜	黄盖	黄色	黑椒	默默	齐全
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

5 rows × 1910 columns

使用朴素贝叶斯分类器

```
In [13]: from sklearn.naive_bayes import MultinomialNB  
  
nb = MultinomialNB()  
  
X_train_vect = vect.fit_transform(X_train)  
nb.fit(X_train_vect, y_train)  
  
train_accuracy = nb.score(X_train_vect, y_train)  
print (train_accuracy)  
  
0.914375
```

```
In [14]: X_test_vect = vect.transform(X_test)  
test_accuracy = nb.score(X_test_vect, y_test)  
  
y_predict = nb.predict(X_test_vect)  
  
print('测试准确率', test_accuracy)  
  
from sklearn.metrics import classification_report  
print("测试集上其他指标: \n",classification_report(y_test, y_predict))
```

测试准确率 0.82

测试集上其他指标:

	precision	recall	f1-score	support
0	0.85	0.77	0.81	197
1	0.79	0.87	0.83	203
micro avg	0.82	0.82	0.82	400
macro avg	0.82	0.82	0.82	400
weighted avg	0.82	0.82	0.82	400

复习：朴素贝叶斯

	docID	words in document	in $c = China?$
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{(\sum_{t' \in V} T_{ct'}) + B}$$

$$c_{\text{map}} = \arg \max_{c \in \mathcal{C}} \hat{P}(c|d) = \arg \max_{c \in \mathcal{C}} \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

训练过程 - I

- 步骤1：提取词典（Vocabulary）
 - $V = \{\text{Chinese}, \text{Beijing}, \text{Shanghai}, \text{Macao}, \text{Tokyo}, \text{Japan}\}$
- 步骤2：计算训练文档个数
 - $N = 4$
- 步骤3：针对正例文档，表示为 c
 - 计算正例文档个数 $N_c = 3$
 - 计算先验概率 $P(c) = \frac{3}{4} = 0.75$
 - 将正例文档进行合并形成文本 text_c

训练过程 - 2

- 步骤3：针对正例文档，即 $c=China$
 - 针对词典 V 中的每个单词，进行如下计算

$$\bullet P(Chinese|c) = \frac{5+1}{8+6} = \frac{3}{7}$$

$$\bullet P(Beijing|c) = \frac{1+1}{8+6} = \frac{1}{7}$$

$$\bullet P(Shanghai|c) = \frac{1+1}{8+6} = \frac{1}{7}$$

$$\bullet P(Macao|c) = \frac{1+1}{8+6} = \frac{1}{7}$$

$$\bullet P(Tokyo|c) = \frac{0+1}{8+6} = \frac{1}{14}$$

$$\bullet P(Japan|c) = \frac{0+1}{8+6} = \frac{1}{14}$$

训练过程 - 3

- 步骤4：针对负例文档，表示为 \bar{c}

- 计算先验概率 $P(\bar{c}) = \frac{1}{4} = 0.25$

- 将负例文档进行合并形成文本text $_{\bar{c}}$

- $P(Chinese|\bar{c}) = \frac{1+1}{3+6} = \frac{2}{9}$

- $P(Beijing|\bar{c}) = \frac{0+1}{3+6} = \frac{1}{9}$

- $P(Shanghai|\bar{c}) = \frac{0+1}{3+6} = \frac{1}{9}$

- $P(Macao|\bar{c}) = \frac{0+1}{3+6} = \frac{1}{9}$

- $P(Tokyo|\bar{c}) = \frac{1+1}{3+6} = \frac{2}{9}$

- $P(Japan|\bar{c}) = \frac{1+1}{3+6} = \frac{2}{9}$

测试（预测）过程

- 给定测试文本
 - Chinese Chinese Chinese Tokyo Japan
- 计算正例的后验概率
 - $P(c|d) = \frac{3}{4} \cdot \left(\frac{3}{7}\right)^3 \cdot \frac{1}{14} \cdot \frac{1}{14} \approx 0.0003$
- 计算负例的后验概率
 - $P(\bar{c}|d) = \frac{1}{4} \cdot \left(\frac{2}{9}\right)^3 \cdot \frac{2}{9} \cdot \frac{2}{9} \approx 0.0001$
- 由于 $P(c|d) > P(\bar{c}|d)$, 预测测试文本的类别为正例
- 请你计算以下文本
 - Beijing Tokyo Japan

练习

- 使用朴素贝叶斯分类器完成以下情感分析

数据集	ID	评论	态度
训练集	0	饭店 味道 好 好 好	1
	1	饭店 味道 差	0
	2	饭店 位置 好	1
测试集		饭店 好 差	?

记录预测结果

```
In [15]: X_vec = vect.transform(X)
nb_result = nb.predict(X_vec)
data['nb_result'] = nb_result

data.head()
```

Out[15]:

	comment	sentiment	simple_cut_comment	cut_comment	nb_result
0	口味：不知道是我口高了，还是这家真不怎么样。??我感觉口味确实很一般很一般。上菜相当快，我敢...	0	口味： 不知道 是我口 高 了， 还是 这家 真 不怎么样 。 ?? 我 感觉 口...	口味 知道 我 口 高 这家 不 怎么样 感觉 口味 确实 很 很 上菜 相当 快 我 敢 菜 都...	0
1	菜品丰富质量好，服务也不错！很喜欢！	1	菜品 丰富 质量 好， 服务 也 不错！ 很 喜欢！	菜品 丰富 质量 服务 不错 很 喜欢	1
2	说真的，不晓得有人排队的理由，香精香精香精香精，拜拜！	0	说真的， 不 晓得 有 人 排队 的 理由， 香 精 香 精 香 精 香 精， 拜拜！	说真的 晓得 有 人 排队 理由 香精 香精 香精 香精 拜拜	0
3	菜量实惠，上菜还算比较快，疙瘩汤喝出了秋日的暖意，烧茄子吃出了大阪烧的味道，想吃土豆片也是口...	1	菜量 实惠， 上菜 还 算 比较 快， 疙瘩汤 喝出 了 秋日 的 暖意， 烧茄子 吃...	菜量 实惠 上菜 算 比较 快 疙瘩汤 喝出 秋日 暖意 烧茄子 吃 出 大阪 烧 味道 想...	1
4	先说我算是娜娜家风荷园开业就一直在这里吃??每次出去回来总想吃一回??有时觉得外面的西式简餐...	1	先说 我 算是 娜娜 家风 荷园 开业 就 一直 在 这里 吃 ?? 每次 出去 回来 总...	先说 算是 娜娜 家风 荷园 开业 吃 每次 出去 回来 总想 吃 一回 有时 觉得 外面 ...	0

总结

- 简单任务
 - 文章的整体感情是积极/消极的?
- 复杂任务
 - 对文章的态度从1-5打分
 - 在文章中识别高兴、愤怒、低落等多种情绪
- 更复杂的任务
 - 检测态度针对的不同侧面
 - 检测态度的持有者