



# 中文情感分类方法简介

Brief Intro to Sentiment Analysis

指导教师： 秦兵教授

主 讲： 李泽魁



HIT-SCIR

## 目录

- 情感分类有什么用
- 情感分类的任务有哪些
- 情感分类的主要方法
- 使用分词&朴素贝叶斯做实验
- 几点思考

哈尔滨工业大学

社会计算与信息检索研究中心



# 句子的情感倾向?

- 胡歌的原音配音好赞，外形声音演技真是得天独厚，偶像派成功转型实力派，赞一个！



- 我从未见过如此厚颜无耻之人~~~



- 中国驻美大使崔天凯接受CNN电话采访，看他如何唇枪舌战、机智对答。我只能说，这个视频非常值得一看。4分56秒处，我大使霸气！



- 刚买的衣服洗了一下线头就开了。



# 情感分析有什么用？

## • 商品口碑分析

 **奥迪A4L 口碑** 全部在售 | 2015款 | 2013款 | 停售年款 ▾

综述 参数配置 图片 视频 报价 降价 口碑 油耗 车型详解 文章 车贷

口碑：★★★★★ 4.3分  
高于同级6.16%(口碑总数499)

优点：

时尚运动	我喜欢	尊贵雅致	好	不错	高速发飙
4561	4398	1542	842	711	371

缺点：

A4烧机油	配置太低	屁股难看	形好质差	满街都是
22692	4709	2628	2560	1481



发布口碑

最新印象：配置丰富 做工精良 个性张扬

添加印象

## 商品评论分析

商品详情
累计评价 607
月成交记录609件
本店同类商品 NEW
扫一扫，手机购买

与描述相符

**4.8**

★★★★★

大家都写到

整体感觉不错 (26)

款式很漂亮 (25)

面料很好 (20)

很修身(14)

手感很舒服 (13)

性价比很高 (11)

不耐洗 (4)

有色差 (4)

质量一般般 (3)

大家认为

尺码：

偏小      适合      偏大

84%

有色差      无色差

色差：      93%

● 全部 ○ 追评 (17) ○ 图片 (0)
☑ 有内容
按时间 ▾

颜色很正。很合身不错！

07.24

颜色：108蓝花灰

尺码：175/96A/M

杨帆gy

这款属于收腰款，我老公人本来就瘦，穿上显得更瘦，不过质量不错，洗过有掉色现象

07.24

解释：亲爱的顾客，由于我们部分产品为成衣染色，最外面会留有一层浮色，掉色属正常现象，请不必过于担心。建议您在第一次洗涤时加入少许盐或醋，可以起到很好的去浮色和护色作用哦。感谢您的支持，欢迎您的再次光临！

颜色：108蓝花灰

尺码：175/96A/M

体重：59kg

身高：181cm

s\*\*\*n (匿名)

# 情感分析有什么用？

## • 网民舆情监控

### 事件详情



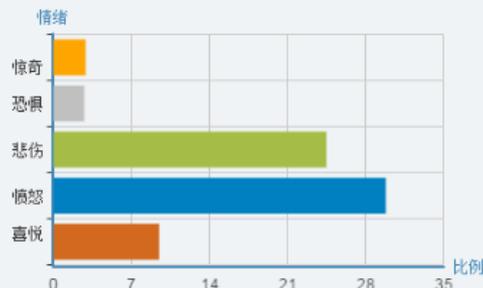
### 宝马撞散马自达

讨论人数：33672

**事件简介：**20日下午2点，江苏南京一陕西牌照宝马车，撞上一正常行驶的马自达轿车，马自达车内一男一女当场身亡。目击者称肇事车加速闯了红灯，后逃离现场。目前肇事者已被抓获，目击者说，在宝马车上查到疑似毒品粉末。

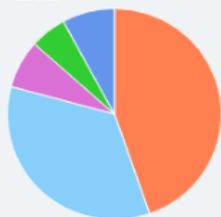
### 情绪分布

#### 事件总情绪分布



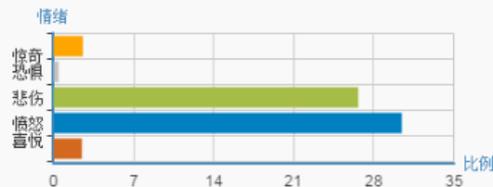
### 事件分布

- 宝马司机闯红灯撞死两人后逃逸
- 南京肇事宝马车系原车主抵债
- 直击第一现场宝马车速真的太可怕
- 出租车车载监控拍下全过程
- 肇事宝马未狂奔



#### 宝马司机闯红灯撞死两人后逃逸

#### 情绪分布



#### 南京肇事宝马车系原车主抵债

#### 直击第一现场宝马车速真的太可怕

# 情感分析有什么用？

## • 根据消费意图做推荐

- ✓ “我想了解”
- ✓ “我要去”
- ✓ “我想要做”
- ✓ “我想买”
- ✓ .....

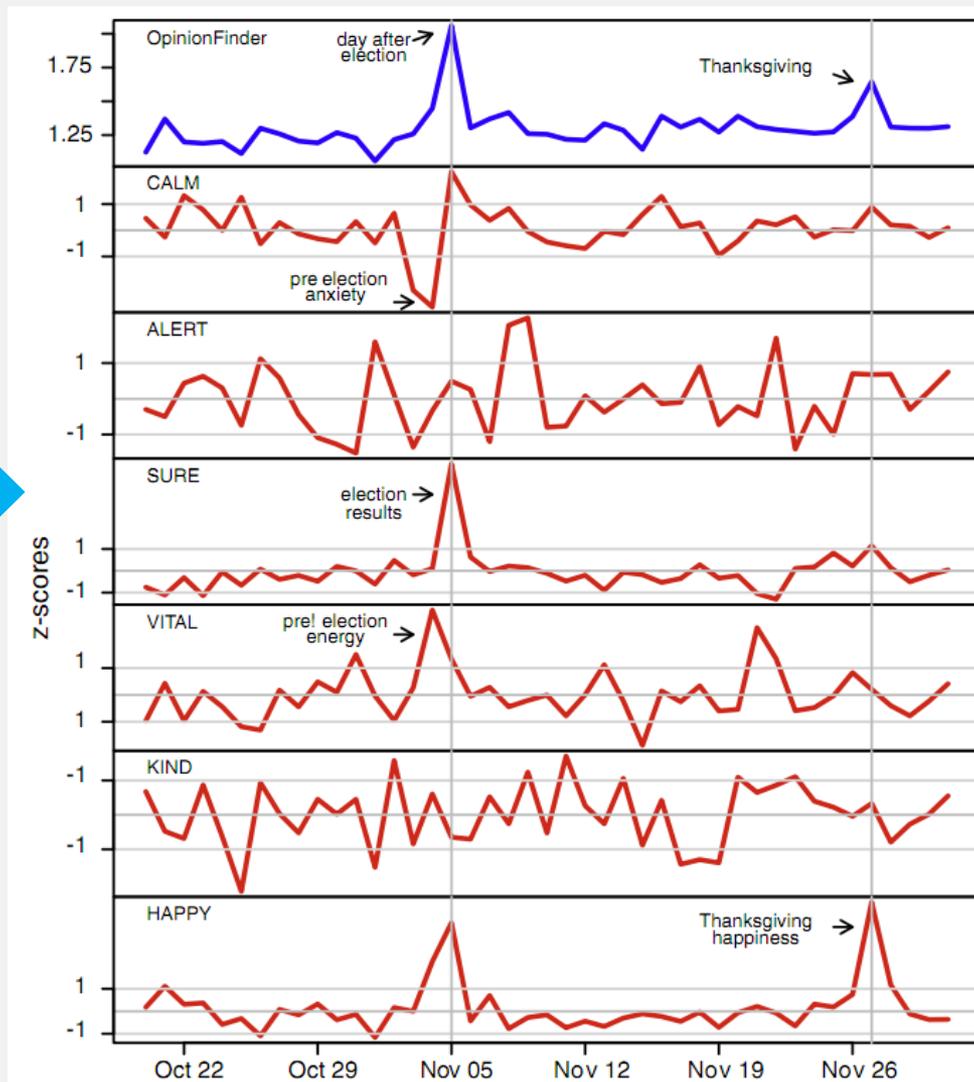


The image shows a Weibo post from BMW China and a recommendation from Rejoice. The BMW post features a large Chinese character '悦' (Yue) formed by car parts, with the text '宝马中国 越是期待已久，悦是如期而至。' and '查看详情'. The Rejoice recommendation includes a video thumbnail and text: '飘柔Rejoice 你是大王的#空气感柔顺女孩#吗？看大王手里拿的是什么？要「发根清爽、发尾柔顺」才能通过大王的纸梳挑战哦~女王快来，证明你就是大王寻找的她？' and '争当男神de空气...'. At the bottom, there are statistics: '收藏', '转发 860', '评论 511', and '3584'.

# 情感分析有什么用？

## • 股票预测

- ✓ 政策变动
- ✓ 经营现状
- ✓ 网民情绪
- ✓ .....



# 情感分析有什么用？

## • 搜索引擎中的应用

[饭后吃水果好吗 \(共61条网友回答\)](#)



不好



好

认为“不好”的网友回答：

- [饭后立刻吃水果好吗?\\_百度知道](#)

不好，饭后吃水果会给胃带来沉重的负担，不愿望消化会导致腹胀，便秘，从而引起胃肠疾病的发生，专家认为，饭后最好在1个小时以后在吃水果，有些人认为吃水果减肥，而饭后吃水果往往相反，现在的年轻女性一般都想... [显示全部](#) ▾

来自百度知道 | 2006-10-21

- [饭后立即吃水果好吗?\\_百度知道](#)

不好，水果会冲淡胃液，会引起消化不良！饭后一小时是最好的时间！如果想减肥，可以饭前吃水果！吃苹果最有效哦！

来自百度知道 | 2005-08-26

[查看53条认为“不好”回答>>](#)

# 情感分类的课本定义

- 情感分析，是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程
- 别名
  - Sentiment analysis
  - Opinion extraction
  - Opinion mining
  - Sentiment mining
  - Subjectivity analysis
- 按照处理文本的粒度不同可以分为词语级、短语级、句子级和篇章级等





# 情感分类的研究任务

- 褒贬(中)分类:
  - 一句话是褒义还是贬义
- 细粒度分类:
  - 喜怒悲恐惊(微博情绪指数系统)
  - 将情感极性打分(例如1-5颗星)
  - 褒贬中更细化(强褒义、褒义、些许褒义等)
- 进阶分类:
  - 评价词(Opinion)、评价对象(Target)抽取
  - 复杂观点抽取 等



# 本节课介绍的情感分类任务

- 褒贬(中)分类:
  - **一句话是褒义还是贬义**
- 细粒度分类:
  - 喜怒悲恐惊(微博情绪指数系统)
  - 将情感极性打分(例如1-5颗星)
  - 褒贬中更细化(强褒义、褒义、些许褒义等)
- 进阶分类:
  - 评价词(Opinion)、评价对象(Target)抽取
  - 复杂观点抽取 等



# 情感分类相关分类方法

- 无监督的分类算法(unsupervised)
  - 基于情感词典及规则
  - 优点：??
  - 缺点：??
  
- 有监督的分类算法(supervised)
  - 基于机器学习(Machine Learning)
  - 优点：??
  - 缺点：??

# 基于词典规则的无监督分类算法

- 直观的思路

$$Polarity = \begin{cases} positive, & \text{if } positive_{count} > negative_{count} \\ negative, & \text{if } positive_{count} < negative_{count} \end{cases}$$

- 思考几个例子:

- 虽然他是个呆和尚，但是我喜欢帅气和尚爱上我。
- 尚选不是家服务优良的店。
- 尔康的特点是鼻孔大、演技好。
- 《我爱你塞北的雪》是彭麻麻唱的歌。
- 你们知不知道我当年和他谈笑风生？





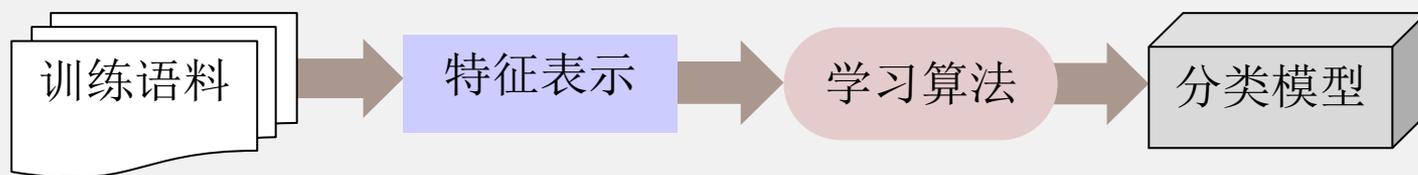
# 基于词典规则的无监督分类算法

- 换个任务：垃圾邮件分类任务
  - 按照“Hand-coded Rules”方法来判别
  - 例如邮件中同时出现“低价”、“秒杀”、“办证”等词汇，那么将其判定为垃圾邮件
- 点评：
  - 这种方法往往准确率非常高召回率很低
  - 规则集需要人工精心撰写
  - 建立和维护规则集的过程比较费事费力
  - 能否让机器自动构建与维护规则？

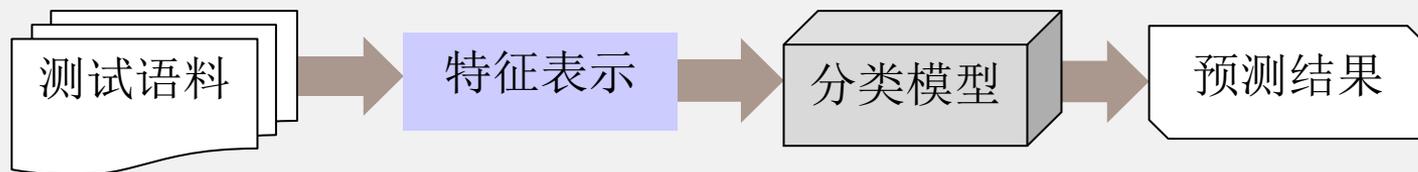
# 基于机器学习的有监督分类算法

- 有监督的机器学习算法

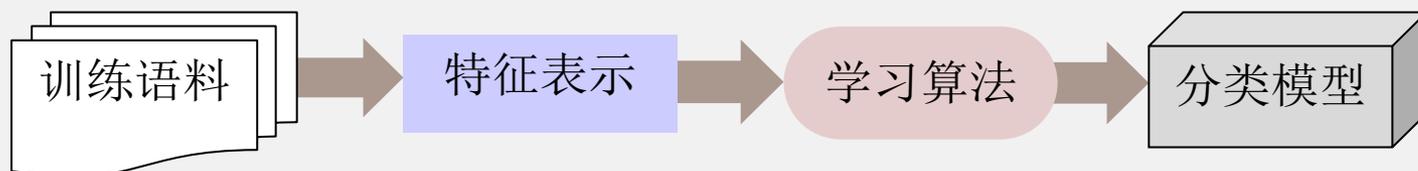
- 训练过程



- 预测过程



## • 模型的学习(Model Learning/Training)



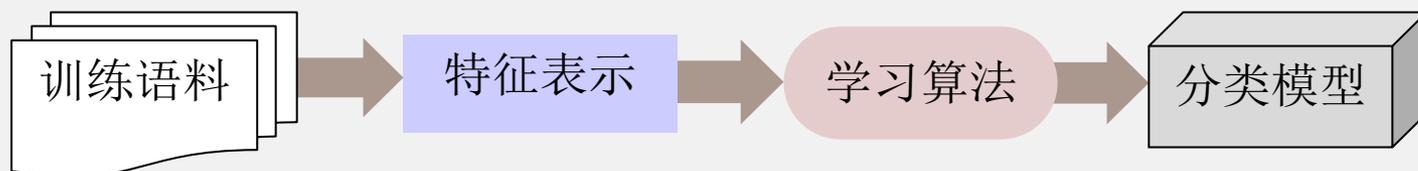
### – 特征表示:

- 对文本进行特征的抽取，转化为机器可理解的向量的表达形式

### – 学习算法:

- 朴素贝叶斯(Naïve Bayes)、最大熵(MaxEnt)、支持向量机(SVM)等

- 简单的特征抽取(Feature Extraction)



- 词袋模型(Bag of Words)
- 否定特征(Negation Features)
- 情感词频率特征(Lexicon Features)

# 情感分类的特征抽取

- 简单的特征抽取(Feature Extraction)
  - 词袋模型(Bag of Words)
    - “吃葡萄不吐葡萄皮”
    - “吃 葡萄 不 吐 葡萄 皮”
    - Word frequency: “吃:1葡萄:2 不:1 吐:1 皮:1”
    - Word occurrence: “吃:1葡萄:1 不:1 吐:1 皮:1”
  - 否定特征(Negation Features)
  - 情感词频率特征(Lexicon Features)

# 情感分类的特征抽取

- 简单的特征抽取(Feature Extraction)
  - 词袋模型(Bag of Words)
  - 否定特征(Negation Features)
    - “我不喜欢这件衣服” vs “我喜欢这件衣服”
    - “我不喜欢这件衣服” →  
“我不喜欢\_NEG 这件\_NEG 衣服\_NEG”
  - 情感词频率特征(Lexicon Features)



# 情感分类的特征抽取

- 简单的特征抽取(Feature Extraction)
  - 词袋模型(Bag of Words)
  - 否定特征(Negation Features)
  - 情感词频率特征(Lexicon Features)
    - “我就会升职加薪 当上总经理 出任CEO 迎娶白富美 走向人生巅峰 想想还有点小激动”
    - 褒义词数目: 3
    - 贬义词数目: 0



# 情感分类的特征抽取

- 特征的抽取还有那些?
  - 词性特征?
  - N-gram特征?
  - 强度词词典特征?
  - 句法依存特征?
  - 词向量特征?



# 情感分类的特征抽取

- 特征的抽取还有那些？

- 词性特征？

- 某些可以影响情感的词性，例如形容词、副词

- N-gram特征？

- 针对词组表达，例如“给力”、“哔了狗了”

- 强度词词典特征？

- 很 非常 十分

- 句法依存特征？

- 主谓结构 动宾结构

- 词向量特征？

- 与深度学习结合，词表达成另一向量空间唯一表示



# 情感分类相关分类方法 —— 总结

- 无监督的分类算法(unsupervised)
  - 基于情感词典及规则
  - 优点：无需标注数据
  - 缺点：构建词典和规则耗时耗力，准确率不高
- 有监督的分类算法(supervised)
  - 基于机器学习(Machine Learning)
  - 优点：分类效果提升
  - 缺点：依赖标注语料和特征选择



# 情感分类 in Action —— Overview

- 下载数据
- 数据预处理
- 文本情感分类
- 分类效果评估



# 情感分类 in Action —— Overview

- 下载数据
  - 3000句褒贬中数据 (已标注)
- 数据预处理
  - 数据清洗 (@USER、URL等)
  - 文本分词 (Java、Python、C等)
- 文本情感分类
  - 基于词典规则的情感分类
  - 基于机器学习的情感分类
- 分类效果评估
  - 交叉验证和准确率



# 情感分类 in Action —— 数据格式

- 下载数据
  - 3000句褒贬中数据 (已标注)
  - 下载链接: [baiduyun/exp/sentiment-data](http://baiduyun/exp/sentiment-data)
  - 数据格式: Label + \t + Sentence

1 -1 → 我不是蒙牛、没你想象那么纯。--要不要这么讽刺啊？蒙牛好尴尬。。。  
2 1 → 蒙牛很牛  
3 0 → 据说古时人们曾用公道杯对付贪酒者，斟酒如超过高度，则会全部漏光。

## – 数据标签:

- 褒义: 1
- 贬义: -1
- 中性: 0



# 情感分类 in Action —— 数据预处理

- 数据预处理

- 数据清洗 (@USER、URL等, 已完成)

3 @帅气的追风少年:我分享了<http://t.cn/RLCcvnC>

- 文本分词 (Java、Python、C等)

- LTP Cloud

- <https://github.com/HIT-SCIR/ltp-cloud-api-tutorial>

- Stanford Parser

- <http://nlp.stanford.edu/software/lex-parser.shtml#Download>

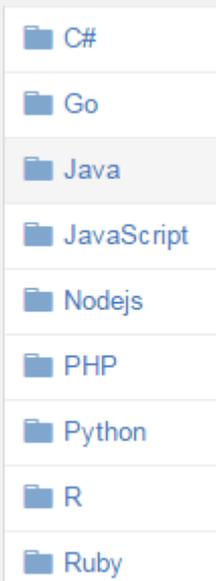
- 中科院分词系统ICTCLAS、腾讯文智平台

- ansj分词、jieba分词、PaodingAnalyzer、IKAnalyzer

- .....

# 情感分类 in Action —— 文本分词

- 文本分词 —— LTP Cloud
  - <https://github.com/HIT-SCIR/ltp-cloud-api-tutorial>



```
String api_key = "YourApiKey";
String pattern = "all";
String format = args[0];
String text = "我爱北京天安门。";
text = URLEncoder.encode(text, "utf-8");

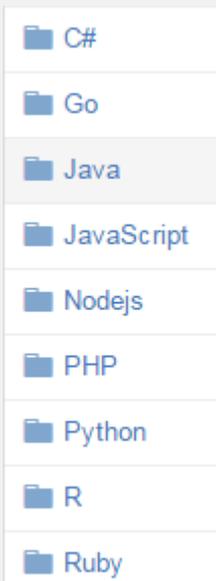
URL url = new URL("http://ltpapi.voicecloud.cn/analysis/?"
    + "api_key=" + api_key + "&"
    + "text=" + text + "&"
    + "format=" + format + "&"
    + "pattern=" + pattern);
URLConnection conn = url.openConnection();
conn.connect();

BufferedReader innnet = new BufferedReader(new InputStreamReader(
    conn.getInputStream(),
    "utf-8"));

String line;
while ((line = innnet.readLine()) != null) {
    System.out.println(line);
}
innnet.close();
```

# 情感分类 in Action —— 文本分词

- 文本分词 —— LTP Cloud
  - <https://github.com/HIT-SCIR/ltp-cloud-api-tutorial>



```
if __name__ == '__main__':
    if len(sys.argv) < 2 or sys.argv[1] not in ["xml", "json", "conll"]:
        print >> sys.stderr, "usage: %s [xml/json/conll]" % sys.argv[0]
        sys.exit(1)

    uri_base = "http://ltpapi.voicecloud.cn/analysis/?"
    api_key = "YourApiKey"
    text = "我爱北京天安门"
    # Note that if your text contain special characters such as linefeed or '&',
    # you need to use urlencode to encode your data
    text = urllib.quote(text)
    format = sys.argv[1]
    pattern = "all"

    url = (uri_base
           + "api_key=" + api_key + "&"
           + "text=" + text + "&"
           + "format=" + format + "&"
           + "pattern=" + "all")

    try:
        response = urllib2.urlopen(url)
        content = response.read().strip()
        print content
    except urllib2.HTTPError, e:
        print >> sys.stderr, e.reason
```



# 情感分类 in Action —— 文本分词

- 文本分词 —— LTP Cloud + Java
  - 安装Java 1.X、Eclipse 3.X (exp/enviroment-java/\*)
  - 下载示例代码 (exp/code-java-cws/LTP4Java.rar)
  - 导入代码到Eclipse中
  - 注册API Key
  - Run Code

## 基本信息

Email:	lzkhit@163.com
Api_key:	31i8x0cEC [redacted] UggEL5ZvlhWuAaxtz
本月流量使用:	312 bytes
本月剩余流量:	18.6 GB

```
//在语言云网站上完成注册后，点击控制面板，就可以看到apikey
String api_key = "HERE IS YOUR KEY";
//用以指定分析模式，可选值包括ws(分词)，pos(词性标注)，ner(
String pattern = "ws";
//用以指定结果格式类型，可选值包括xml(XML格式)，json(JSON
String format = "plain";
//待分析的文本
String text = "我爱北京天安门。";
```

Cut Word Result: 我 爱 北京 天安门。

Index : 0, Word : 我  
Index : 1, Word : 爱  
Index : 2, Word : 北京  
Index : 3, Word : 天安门  
Index : 4, Word : 。

# 情感分类 in Action —— 情感分类

- 文本情感分类
  - 基于词典规则的情感分类

$$Polarity = \begin{cases} positive, & \text{if } positive_{count} > negative_{count} \\ negative, & \text{if } positive_{count} < negative_{count} \end{cases}$$

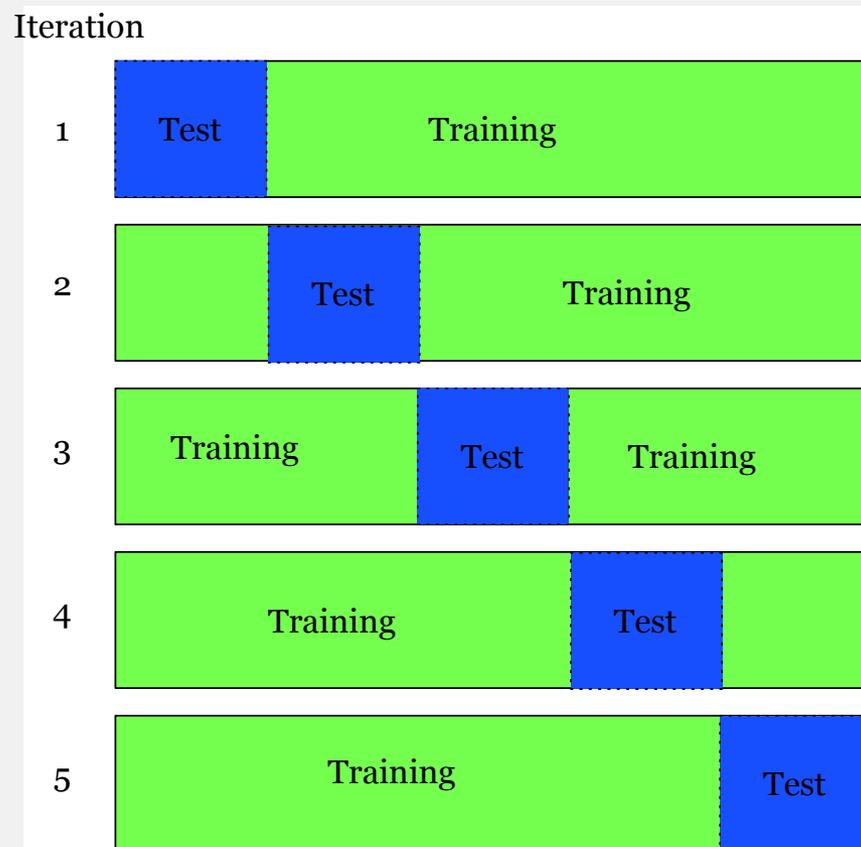
- 基于机器学习的情感分类 (例如NB分类器)

$$c_{NB} = \operatorname{argmax}_{c_j \in \mathcal{C}} P(c_j \mid w_1, w_2 \dots w_i) = \operatorname{argmax}_{c_j \in \mathcal{C}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

$$P(c_j) = \frac{|docs_j|}{|\text{total \# documents}|} \quad \hat{P}(w \mid c) = \frac{count(w, c) + 1}{count(c) + |V|}$$

- 分类效果评估
  - 交叉验证 (cross validation)
    - 例如五折交叉 (5-fold)
    - 4/5的数据作为Train
    - 1/5的数据作为Test
    - 分类效果取平均
  - 准确率 (accuracy)

$$\text{Accuracy} = \frac{|\text{分类正确的样本数目}|}{|\text{总样本数目}|}$$





# 情感分类 —— 思考

- 情感词典怎么来的
  - 纯手工标注？ 机器自动标注？ 半自动？
- 分类模型的思考
  - 决策树、最大熵、支持向量机
  - 同样训练数据为什么分类效果有差异
  - 生成模型与判别模型？
- 没有出现情感词怎么办
  - “今天股票大涨！”，“小明考试又不及格。”
  - 深度学习？
- 其他评价指标
  - 精确率、召回率、F值



哈爾濱工業大學  
社会计算与信息检索研究中心



# Thanks Q&A

答疑邮箱: [zkli \(AT\) ir.hit.edu.cn](mailto:zkli(AT)ir.hit.edu.cn)