哈爾濱工業大學
社会计算与信息检索研究中心

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

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

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

情感分析有什么用?

- 商品评论分析

情感分析有什么用?

- 网民舆情监控

情感分析有什么用?

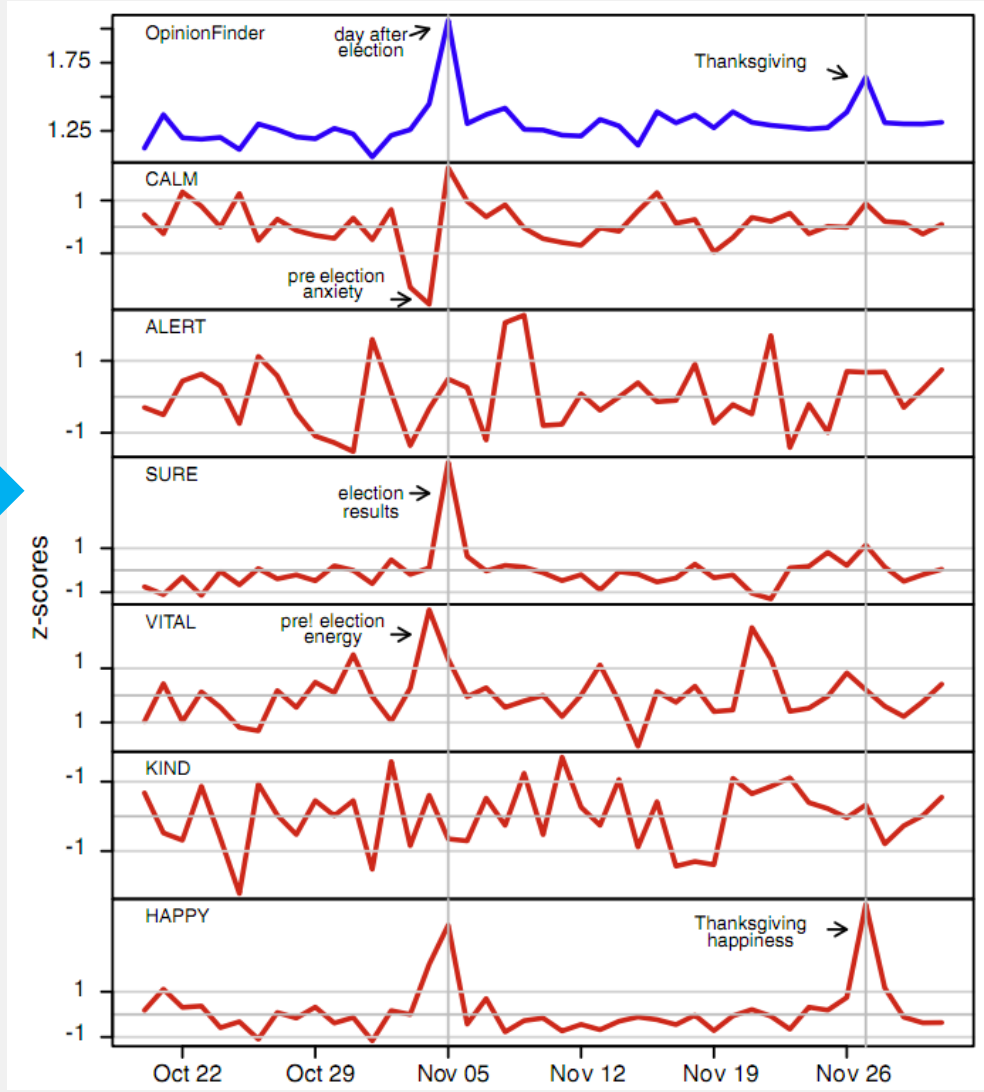
- 根据消费意图做推荐

✓ “我想了解”
✓ “我要去”
✓ “我想要做”
✓ “我想买”
✓ ……

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

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

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

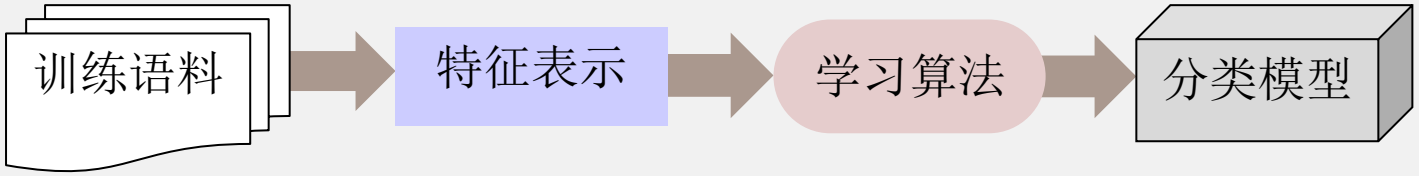
- 思考几个例子:
  - 虽然他是个呆和尚，但是我喜欢帅气和尚爱上我。
  - 尚选不是家服务优良的店。
  - 尔康的特点是鼻孔大、演技好。
  - 《我爱你塞北的雪》是彭麻麻唱的歌。
  - 你们知不知道我当年和他谈笑风生?

基于词典规则的无监督分类算法

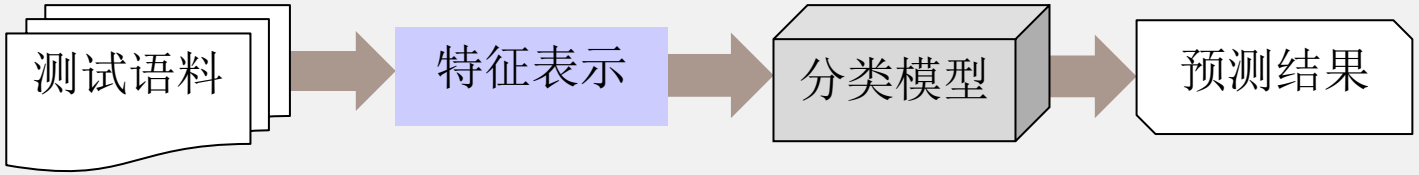
- 换个任务：垃圾邮件分类任务
  - 按照”Hand-coded Rules”方法来判别
  - 例如邮件中同时出现”低价”、”秒杀”、”办证”等词汇，那么将其判定为垃圾邮件
- 点评：
  - 这种方法往往准确率非常高召回率很低
  - 规则集需要人工精心撰写
  - 建立和维护规则集的过程比较费事费力
  - 能否让机器自动构建与维护规则?

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

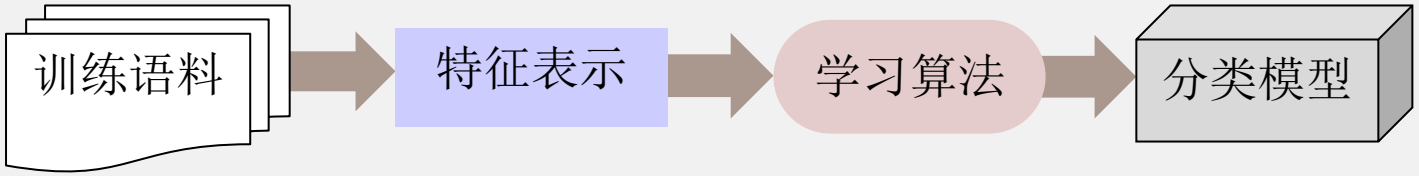


  - 预测过程



分类模型的学习

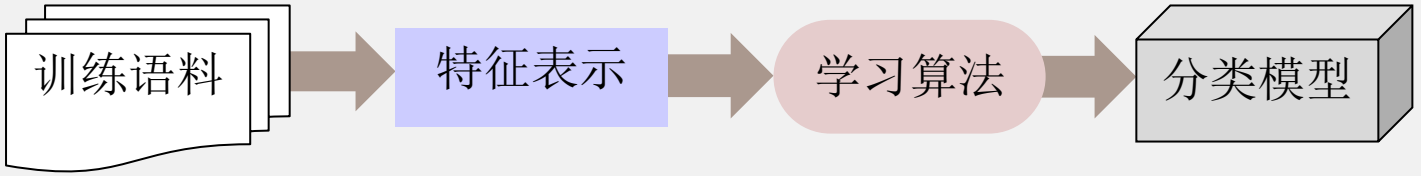
- 模型的学习(Model Learning/Training)



  - 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

情感分类的特征抽取

- 简单的特征的抽取(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

- 简单的特征的抽取(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

- 简单的特征的抽取(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

- 简单的特征的抽取(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
  - 数据格式：Label + \t + Sentence

```
-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。。。
1→蒙牛很牛
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
```

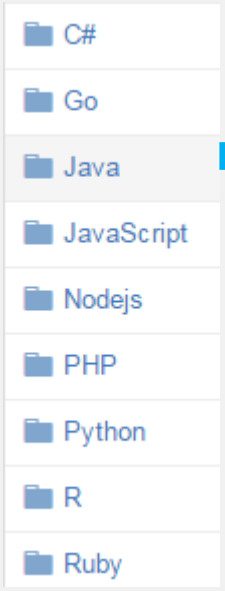
  - 文本分词 (Java、Python、C等)
    - LTP Cloud

      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



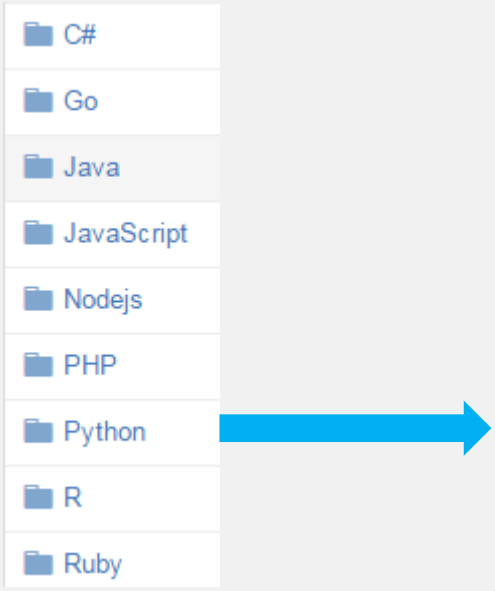

```
String api_key = "YourApiKey";
String pattern = "all";
String format  = args[0];
String text    = "我爱北京天安门。";
text = URLEncoder.encode(text, "utf-8");

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 innet = new BufferedReader(new InputStreamReader(
                        conn.getInputStream(),
                        "utf-8"));
String line;
while ((line = innet.readLine())!= null) {
    System.out.println(line);
}
innet.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




```python
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]UgGEL5ZvIhWuAaxtz
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

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

  - 基于机器学习的情感分类 (例如NB分类器)

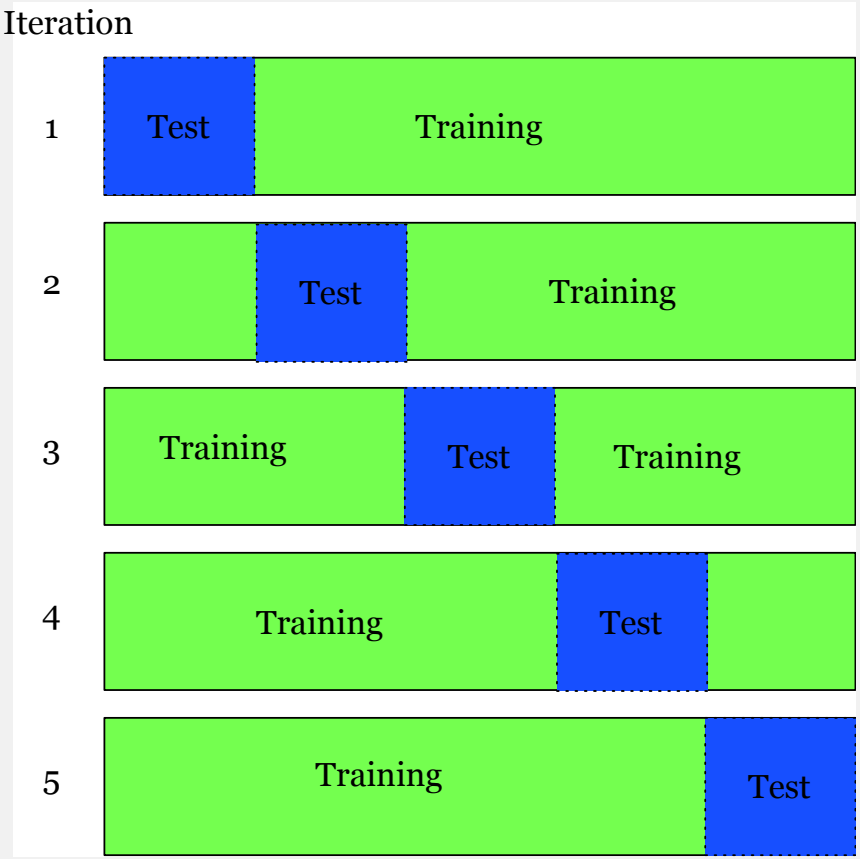
$$c_{NB} = \underset{c_j \in C}{\operatorname{argmax}} P(c_j \mid w1, w2 \ldots wi) = \underset{c_j \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

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

情感分类 in Action —— 效果评估

- 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