

一、课程介绍

斯坦福大学于2012年3月在Coursera启动了在线自然语言处理课程，由NLP领域大牛Dan Jurafsky 和 Chirs Manning教授授课：
<https://class.coursera.org/nlp/>

以下是本课程的学习笔记，以课程PPT/PDF为主，其他参考资料为辅，融入个人拓展、注解，抛砖引玉，欢迎大家在“[我爱公开课](#)”上一起探讨学习。

课件汇总下载地址：[斯坦福大学自然语言处理公开课课件汇总](#)





二、情感分析 (Sentiment Analysis)

1) What is Sentiment Analysis?

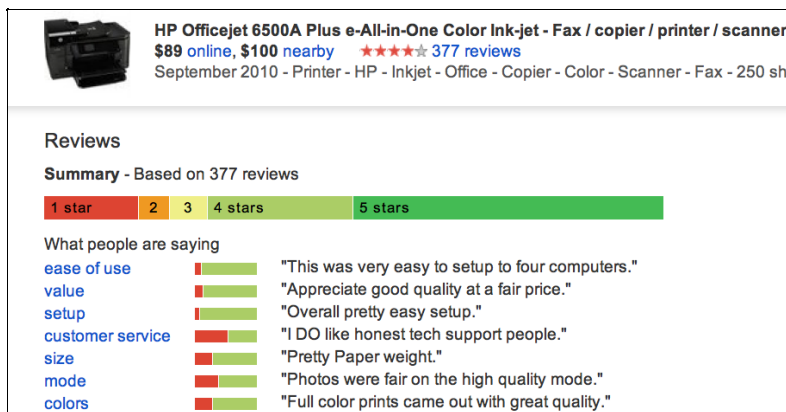
情感分析 (Sentiment analysis)，又称倾向性分析，意见抽取 (Opinion extraction)，意见挖掘 (Opinion mining)，情感挖掘 (Sentiment mining)，主观分析 (Subjectivity analysis)，它是对带有情感色彩的主观性文本进行分析、处理、归纳和推理的过程，如从评论文本中分析用户对“数码相机”的“变焦、价格、大小、重量、闪光、易用性”等属性的情感倾向。

更多例子如下：

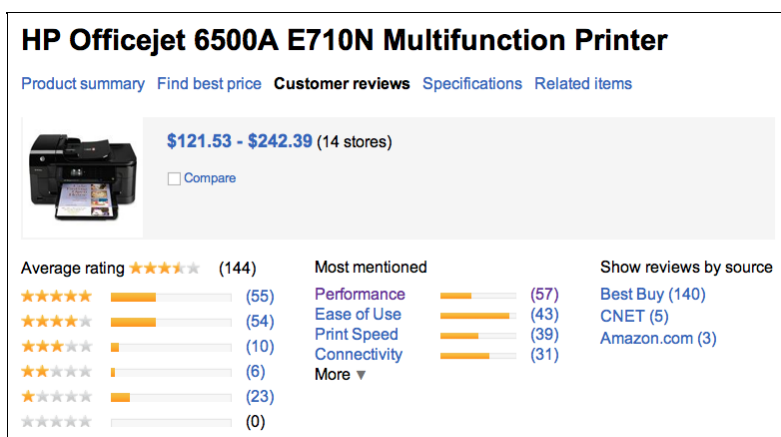
I 从电影评论中识别用户对电影的褒贬评价：

| | |
|---|---|
|  | • unbelievably disappointing |
|  | • Full of zany characters and richly applied satire, and some great plot twists |
|  | • this is the greatest screwball comedy ever filmed |
|  | • It was pathetic. The worst part about it was the boxing scenes. |

I [Google Product Search](#)识别用户对产品各种属性的评价，并从评论中选择代表性评论展示给用户：

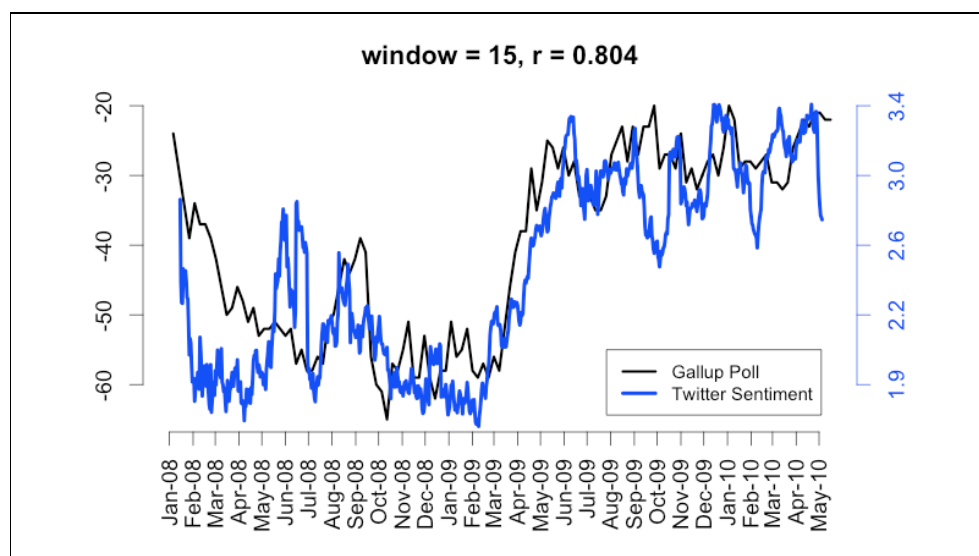


I [Bing Shopping](#)识别用户对产品各种属性的评价：

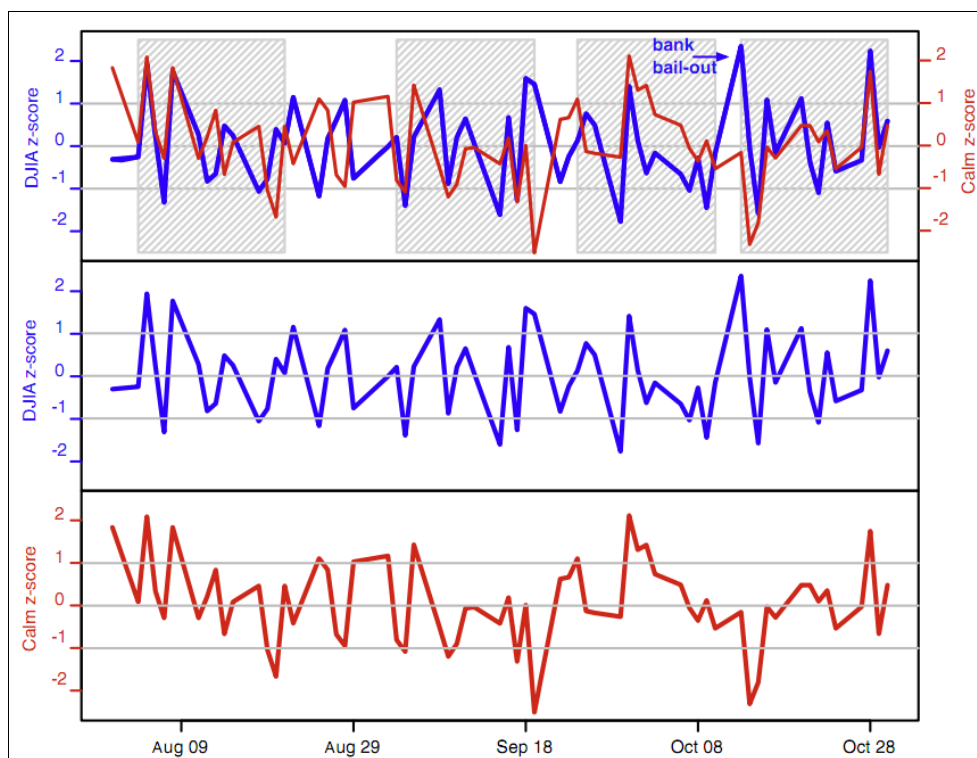
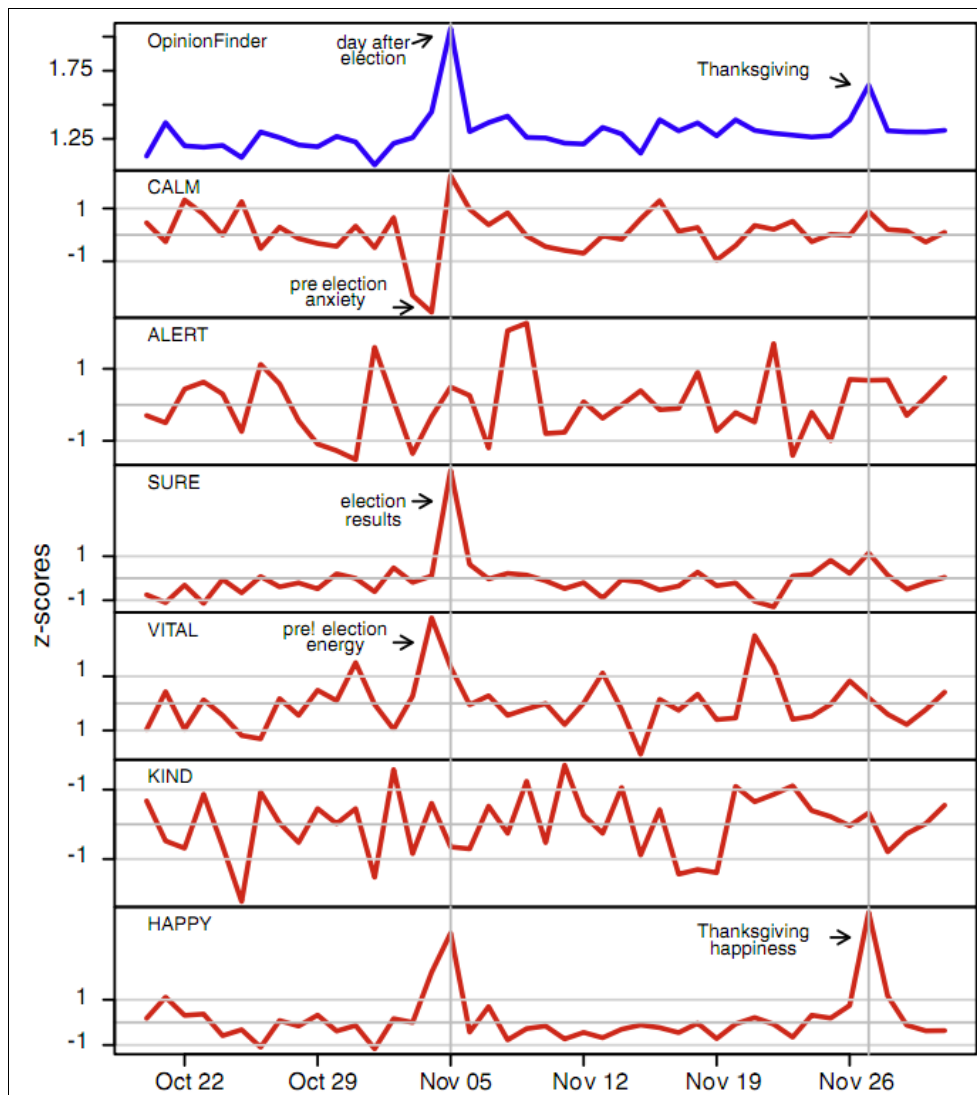


I [Twitter sentiment versus Gallup Poll of Consumer Confidence](#)：挖掘Twitter（中

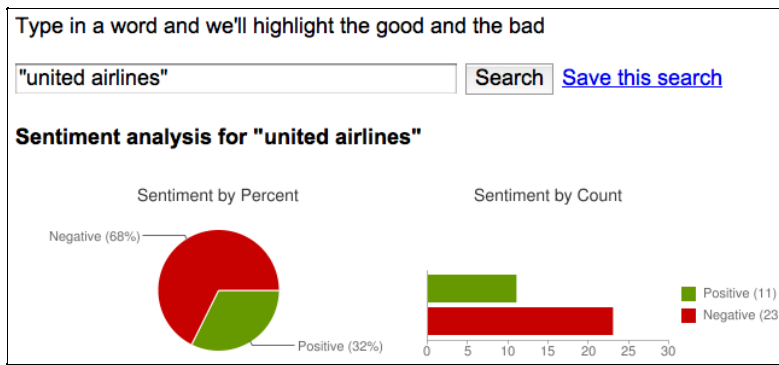
文：微博）中的用户情感发现，其与传统的调查、投票等方法结果有高度的一致性（以消费者信心和政治选举为例，correlation达80%），详细见论文：Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. [From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series](#). In ICWSM-2010。（注：下图中2008年到2009年初，网民情绪低谷是金融危机导致，从2009年5月份开始慢慢恢复）



Twitter sentiment: 通过Twitter用户情感预测股票走势，2012年5月，世界首家基于社交媒体的对冲基金 Derwent Capital Markets 在屡次跳票后终于上线。它会即时关注Twitter 中的公众情绪指导投资。正如基金创始人保罗·赫汀（Paul Hawtin）表示：“长期以来，投资者已经广泛地认可金融市场由恐惧和贪婪驱使，但我们从未拥有一种技术或数据来量化人们的情感。”一直为金融市场非理性举动所困惑的投资者，终于有了一扇可以了解心灵世界的窗户——那便是 Twitter 每天浩如烟海的推文，在一份八月份报道中显示，利用 Twitter 的对冲基金 Derwent Capital Markets 在首月的交易中已经盈利，它以1.85%的收益率，让平均数只有0.76%的其他对冲基金相形见绌。类似的工作还有预测电影票房、选举结果等，均是将公众情绪与社会事件对比，发现一致性，并用于预测，如将“冷静CLAM”情绪指数后移3天后和道琼斯工业平均指数DIJA惊人一致。详细见论文：Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. [Twitter mood predicts the stock market](#), Journal of Computational Science 2:1, 1-8.（注：DIJA，全称Dow Jones Industrial Average）



I **Target Sentiment on Twitter (Twitter Sentiment App)** : 对Twitter中包含给定query的 tweets进行情感分类。对于公司了解用户对公司、产品的喜好，用于指导改善产品和服务，公司还可以据此发现竞争对手的优劣势，用户也可以根据网友甚至亲友评价决定是否购买特定产品。详细见论文：Alec Go, Richa Bhayani, Lei Huang. 2009. [Twitter Sentiment Classification using Distant Supervision](#).



| |
|--|
| jijacobson: 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 |

情感分析的意义何在？下面以实际应用为例进行直观的阐述：

- **Movie:** is this review positive or negative?
- **Products:** what do people think about the new iPhone?
- **Public sentiment:** how is consumer confidence? Is despair increasing?
- **Politics:** what do people think about this candidate or issue?
- **Prediction:** predict election outcomes or market trends from sentiment

情感分析主要目的就是识别用户对事物或人的看法、态度（attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons），参与主体主要包括：

1. **Holder (source)** of attitude: 观点持有者
2. **Target (aspect)** of attitude: 评价对象
3. **Type** of attitude: 评价观点
 - From a set of types: *Like, love, hate, value, desire, etc.*
 - Or (more commonly) simple weighted **polarity**: *positive, negative, neutral*, together with *strength*
4. **Text** containing the attitude: 评价文本，一般是句子或整篇文档

更细更深入的还包括评价属性，情感词/极性词，评价搭配等、

通常，我们面临的情感分析任务包括如下几类：

1. **Simplest task:** Is the attitude of this text positive or negative?
2. **More complex:** Rank the attitude of this text from 1 to 5
3. **Advanced:** Detect the target, source, or complex attitude types

后续章节将以Simplest task为例进行介绍。

2) A Baseline Algorithm

本小节对影评进行情感分析为例，向大家展示一个简单、实用的情感分析系统。详细见论文: Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. [Thumbs up? Sentiment Classification using Machine Learning Techniques](#). EMNLP-2002, 79—86.

Bo Pang and Lillian Lee. 2004. [A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts](#). ACL, 271-278

我们面临的任务是“Polarity detection: Is an **IMDB** movie review positive or negative?”，数据集为“**Polrity Data 2.0**”: <http://www.cs.cornell.edu/people/pabo/movie-review-data>。作者将情感分析当作分类任务，拆分成如下子任务：

- a. **Tokenization:** 正文提取，过滤时间、电话号码等，保留大写字母开头的字符串，保留表情符号，切词；
- b. **Feature Extraction:** 直观上，我们会认为形容词直接决定文本的情感，而Pang和Lee的

实验表明，采用所有词（unigram）作为特征，可以达到更好的情感分类效果。

其中，需要对否定句进行特别的处理，如句子“I **didn't** like this movie”vs “I really like this movie”，unigram只差一个词，但是有着截然不同的含义。为了有效处理这种情况，Das and Chen (2001)提出了“Add NOT_ to every word between negation and following punctuation”，根据此规则可以将句子“didn't like this movie , but I”转换为“didn't NOT_like NOT_this NOT_movie, but I”。

另外，在抽取特征时，直观的感觉“Word occurrence may matter more than word frequency”，这是因为最相关的情感词在一些文本片段中仅仅出现一次，词频模型起得作用有限，甚至是负作用，则使用多重伯努利模型事件空间代替多项式事件空间，实验也的确证明了这一点。所以，论文最终选择二值特征，即词的出现与否，代替传统的频率特征。 $\log(\text{freq}(w))$ 也是一种值得尝试的降低频率干扰的方法。

α. **Classification using different classifiers:**如Naïve Bayes、MaxEnt、SVM，以朴素贝叶斯分类器为例，训练过程如下：

- From training corpus, extract *Vocabulary*
- Calculate $P(c_j)$ terms
 - For each c_j in C do
$$docs_j \leftarrow \text{all docs with class} = c_j$$
$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$
- Calculate $P(w_k | c_j)$ terms
 - Remove duplicates in each doc:
 - For each word type w in doc_j
 - Retain only a single instance of w
 - $Text_j \leftarrow$ single doc containing all $docs_j$
 - For each word w_k in *Vocabulary*
$$n_k \leftarrow \text{\# of occurrences of } w_k \text{ in } Text_j$$
$$P(w_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

预测过程如下：

- First remove all duplicate words from d
- Then compute NB using the same equation:
$$c_{NB} = \operatorname{argmax}_{c_j \in C} P(c_j) \prod_{i \in \text{positions}} P(w_i | c_j)$$

实验表明，MaxEnt和SVM相比Naïve Bayes可以得到更好的效果。

最后，通过case review可以总结下，影评情感分类的难点是什么？

- 语言表达的含蓄微妙：“If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”，“She runs the gamut of emotions from A to B”。
- 挫败感表达方式：先描述开始的期待（不吝赞美之词），后表达最后失望感受，如“This film should be **brilliant**. It sounds like a **great plot**, the actors are **first grade**, and the supporting cast is **good** as well, and Stallone is attempting to deliver a **good performance**. However, it **can't hold up**.”，“Well as usual Keanu Reeves is nothing special, but surprisingly, the **very talented** Laurence Fishbourne is **not so good** either, I was surprised.”。

3) Sentiment Lexicons

情感分析模型非常依赖于情感词典抽取特征或规则，以下罗列了较为流行且成熟的开放情感词典资源：

- GI (The General Inquirer)：该词典给出了每个词条非常全面的信息，如词性，反义词，褒贬，等，组织结构如下：

| | Entry | Positiv | Negativ | Hostile | ...184 classes ... | Othtags | Defined |
|-------|-------------|---------|---------|---------|--------------------|---------|---------|
| 1 | A | | | | | DET ART | ... |
| 2 | ABANDON | | Negativ | | | SUPV | |
| 3 | ABANDONMENT | | Negativ | | | Noun | |
| 4 | ABATE | | Negativ | | | SUPV | |
| 5 | ABATEMENT | | | | | Noun | |
| ... | | | | | | | |
| 35 | ABSENT#1 | | Negativ | | | Modif | |
| 36 | ABSENT#2 | | | | | SUPV | |
| ... | | | | | | | |
| 11788 | ZONE | | | | | Noun | |

详细见论文：Philip J. Stone, Dexter C Dunphy, Marshall S. Smith, Daniel M. Ogilvie. 1966. [The General Inquirer: A Computer Approach to Content Analysis](#). MIT Press

- [LIWC \(Linguistic Inquiry and Word Count\)](#): 该词典通过大量正则表达式描述不同类别的情感词规律，其类别体系与GI ([The General Inquirer](#)) 基本一致，组织结构如下：

| Category | Examples |
|----------|--|
| Negate | aint, ain't, arent, aren't, cannot, cant, can't, couldnt, ... |
| Swear | arse, arsehole*, arses, ass, asses, asshole*, bastard*, ... |
| Social | acquainta*, admit, admits, admitted, admitting, adult, adults, advice, advis* |
| Affect | abandon*, abuse*, abusi*, accept, accepta*, accepted, accepting, accepts, ache* |
| Posemo | accept, accepta*, accepted, accepting, accepts, active*, admir*, ador*, advantag* |
| Negemo | abandon*, abuse*, abusi*, ache*, aching, advers*, afraid, aggravat*, aggress*, |
| Anx | afraid, alarm*, anguish*, anxi*, apprehens*, asham*, aversi*, avoid*, awkward* |
| Anger | jealous*, jerk, jerked, jerks, kill*, liar*, lied, lies, lous*, ludicrous*, lying, mad |

详细见论文：Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). [Linguistic Inquiry and Word Count: LIWC 2007](#). Austin, TX

- [MPQA Subjectivity Cues Lexicon](#): 其中包含Positive words: 2718, Negative words: 4912, 组织结构如下图所示：

| | Strength | Length | Word | Part-of-speech | Stemmed | Polarity |
|-------|-----------------|--------|-------------------|----------------|------------|------------------------|
| 1. | type=weaksubj | len=1 | word1=abandoned | pos1=adj | stemmed1=n | priorpolarity=negative |
| 2. | type=weaksubj | len=1 | word1=abandonment | pos1=noun | stemmed1=n | priorpolarity=negative |
| 3. | type=weaksubj | len=1 | word1=abandon | pos1=verb | stemmed1=y | priorpolarity=negative |
| 4. | type=strongsubj | len=1 | word1=abase | pos1=verb | stemmed1=y | priorpolarity=negative |
| 5. | type=strongsubj | len=1 | word1=abasement | pos1=anypos | stemmed1=y | priorpolarity=negative |
| 6. | type=strongsubj | len=1 | word1=abash | pos1=verb | stemmed1=y | priorpolarity=negative |
| 7. | type=weaksubj | len=1 | word1=abate | pos1=verb | stemmed1=y | priorpolarity=negative |
| 8. | type=weaksubj | len=1 | word1=abdicate | pos1=verb | stemmed1=y | priorpolarity=negative |
| 9. | type=strongsubj | len=1 | word1=aberration | pos1=adj | stemmed1=n | priorpolarity=negative |
| 10. | type=strongsubj | len=1 | word1=aberration | pos1=noun | stemmed1=n | priorpolarity=negative |
| ... | | | | | | |
| 8221. | type=strongsubj | len=1 | word1=zest | pos1=noun | stemmed1=n | priorpolarity=positive |

详细见论文：Theresa Wilson, Janyce Wiebe, and Paul Hoffmann (2005). [Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis](#). Proc. of HLT-EMNLP-2005.

Riloff and Wiebe (2003). [Learning extraction patterns for subjective expressions](#). EMNLP-2003.

- [Bing Liu Opinion Lexicon](#): 其中包含Positive words: 2006, Negative words: 4783, 需要特别说明的是，词典不但包含正常的用词，还包含了拼写错误、语法变形，俚语以及社交媒体标记等，详细见论文：Minqing Hu and Bing Liu. [Mining and Summarizing Customer Reviews](#). ACM SIGKDD-2004.
- [SentiWordNet](#): 其通过对WordNet中的词条进行情感分类，并标注出每个词条属于positive和negative类别的权重大小，组织结构如下：

| POS | ID | PosScore | NegScore | SynsetTerms | Gloss |
|-----|----------|----------|----------|------------------------------------|---|
| a | 00001740 | 0.125 | 0 | able#1 | (usually followed by 'to') having the necessary means or [...] |
| a | 00002098 | 0 | 0.75 | unable#1 | (usually followed by 'to') not having the necessary means or [...] |
| a | 00002312 | 0 | 0 | dorsal#2 abaxial#1 | facing away from the axis of an organ or organism; [...] |
| a | 00002527 | 0 | 0 | ventral#2 adaxial#1 | nearest to or facing toward the axis of an organ or organism; [...] |
| a | 00002730 | 0 | 0 | acroscopic#1 | facing or on the side toward the apex |
| a | 00002843 | 0 | 0 | basiscopic#1 | facing or on the side toward the base |
| a | 00002956 | 0 | 0 | abducting#1 abducent#1 | especially of muscles; [...] |
| a | 00003131 | 0 | 0 | adductive#1 adducting#1 adducent#1 | especially of muscles; [...] |
| a | 00003356 | 0 | 0 | nascent#1 | being born or beginning; [...] |
| a | 00003553 | 0 | 0 | emerging#2 emergent#2 | coming into existence; [...] |

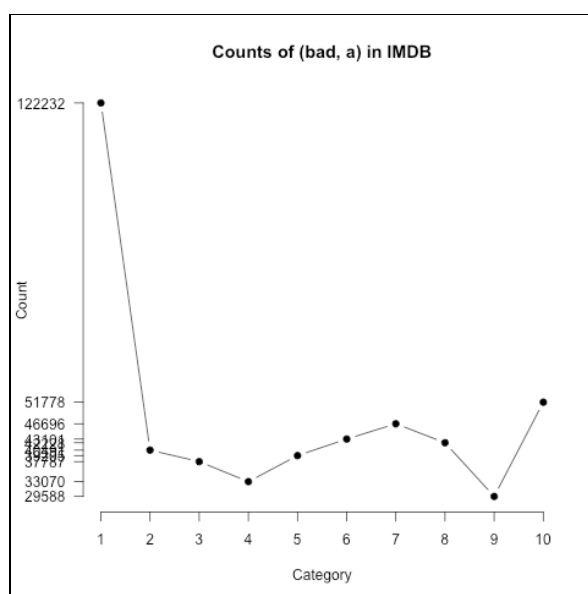
详细见论文：Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani.
2010SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and
Opinion Mining. LREC-2010

以上给出了一系列可用的情感词典资源，但是，如何选择一个合适的为我所用呢？这里，通过对比同一词条在不同词典之间的分类，衡量词典资源的不一致程度，如下：

| | Opinion Lexicon | General Inquirer | SentiWordNet | LIWC |
|------------------|--------------------|---------------------|-----------------|---------------|
| MPQA | 33/5402 (0.6%) | 49/2867 (2%) | 1127/4214 (27%) | 12/363 (3%) |
| Opinion Lexicon | | 32/2411 (1%) | 1004/3994 (25%) | 9/403 (2%) |
| General Inquirer | | | 520/2306 (23%) | 1/204 (0.5%) |
| SentiWordNet | | | | 174/694 (25%) |
| LIWC | | | | |

对于在不同词典中表现不一致的词条，我们至少可以做两件事情。第一，review这些词条，通过少量人工加以纠正；第二，可以得到一些存在褒贬歧义的词条。

给定一个词，如何确定其以多大概率出现在某种情感类别文本中呢？以IMDB下不同打分下影评为例，最简单的方法就是计算每个分数（星的个数）对应的文本中词条出现的频率，如下图所示为Count(“bad”)分布情况：



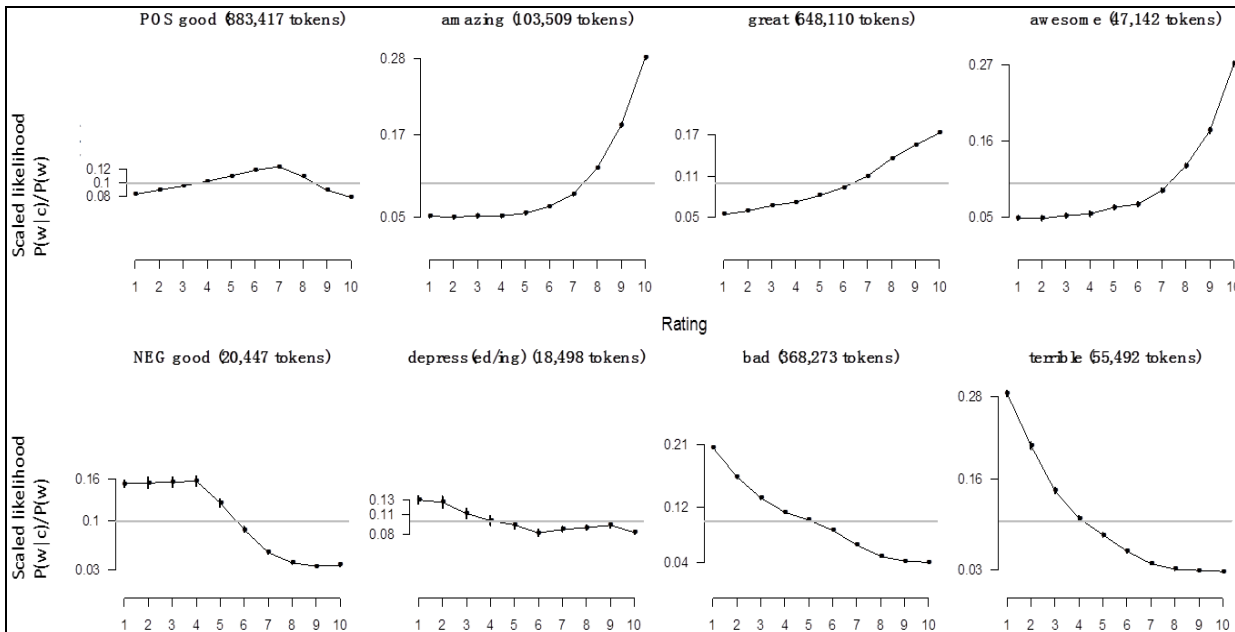
使用更多的是likelihood公式：

$$P(w|c) = \frac{f(w, c)}{\sum_{w \in c} f(w, c)}$$

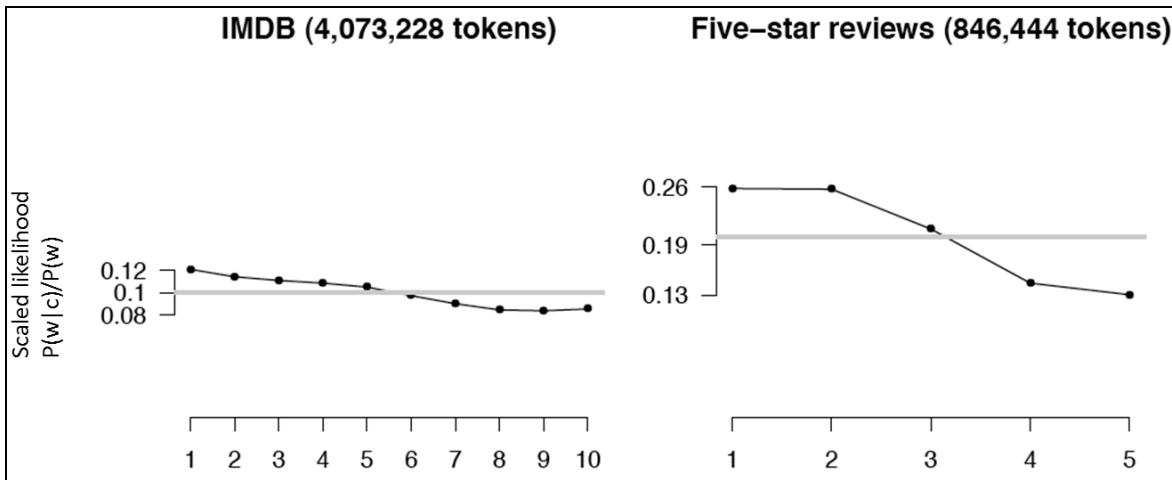
为了使得不同词条在不同类别下的概率可比，通常使用Scaled likelihood公式代替，如下：

$$\frac{P(w|c)}{P(w)}$$

如下图所示，列出了部分词条在不同类别下的Scaled likelihood，据此可以判断每个词条的倾向性。



另外，我们通常会有这么一个疑问：否定词（如*not*, *n't*, *no*, *never*）是否更容易出现在negative情感文本中？[Potts, Christopher \(2011\)](#) 等通过实验给出了答案：More negation in negative sentiment，如下图所示：



4) Learning Sentiment Lexicons

我们在庆幸和赞扬众多公开情感词典为我所用的同时，我们不免还想了解构建情感词典的方法，正所谓知其然知其所以然。一方面在面临新的情感分析问题，解决新的情感分析任务时，难免会需要结合实际需求构建或完善情感词典，另一方面，可以将成熟的词典构建方法应用于其他领域，知识无边界，许多方法都是相通的。

常见的情感词典构建方法是基于半指导的bootstrapping学习方法，主要包括两步：

1. Use a small amount of information (Seed)
 - a. A few labeled examples
 - b. A few hand-built patterns
2. To bootstrap a lexicon

接下来，通过相关的几篇论文，详细阐述下构建情感词典的方法。具体如下：

1. Hatzivassiloglou & McKeown: 论文见Vasileios Hatzivassiloglou and Kathleen R. McKeown. 1997. [Predicting the Semantic Orientation of Adjectives](#). ACL, 174 – 181，基于这样的一种语言现象：“Adjectives conjoined by ‘and’ ’ have same polarity; Adjectives conjoined by ‘but’ ‘do not’ ”，如下示例：

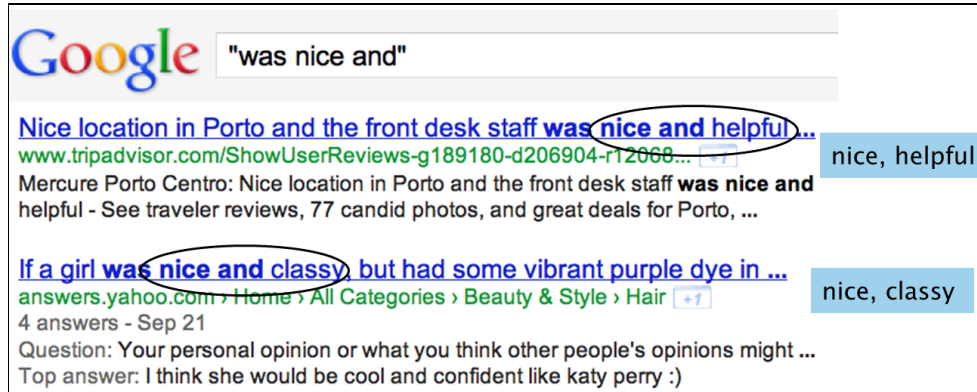
- Fair and legitimate, corrupt and brutal
- *fair and brutal, *corrupt and legitimate
- fair but brutal

Hatzivassiloglou & McKeown (1997) 提出了基于bootstrapping的学习方法，主要包括四步：

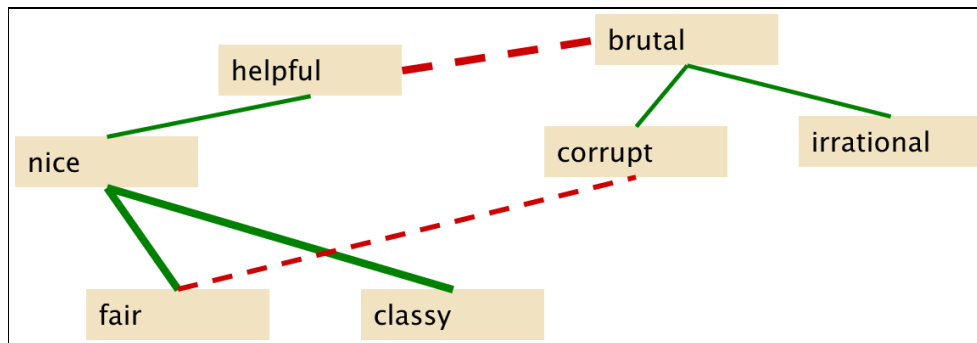
- Step 1: Label seed set of 1336 adjectives (all >20 in 21 million word WSJ corpus)

初始种子集包括657个 positive words (如adequate central clever famous intelligent remarkable reputed sensitive slender thriving...) 和679个 negative words (如contagious drunken ignorant lanky listless primitive strident troublesome unresolved unsuspecting...)

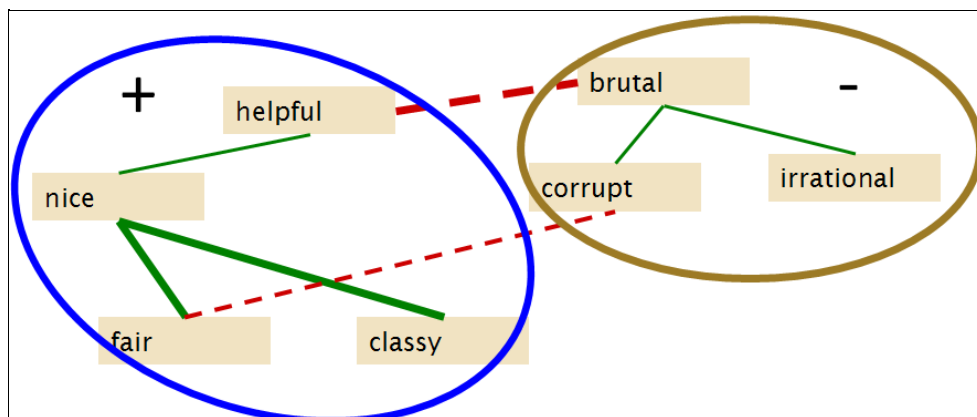
- Step 2: Expand seed set to conjoined adjectives, 如下图所示：



- Step 3: Supervised classifier assigns “polarity similarity” to each word pair, resulting in graph, 如下图所示：



- Step 4: Clustering for partitioning the graph into two



最终，输出新的情感词典，如下（加粗词条为自动挖掘出的词条）：

- Positive: **bold** **decisive** **disturbing** **generous** **good** **honest** **important** **large** **mature** **patient** **peaceful** **positive** **proud** **sound** **stimulating** **straightforward** **strange** **talented** **vigorous** **witty**...
- Negative: **ambiguous** **cautious** **cynical** **evasive** **harmful** **hypocritical** **inefficient** **insecure** **irrational** **irresponsible** **minor** **outspoken** **pleasant** **reckless** **risky** **selfish** **tedious** **unsupported** **vulnerable** **wasteful**...

2. Turney Algorithm: 论文见Turney (2002): [Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews](#), 具体步骤如下:

- Step 1: Extract a *phrasal lexicon* from reviews, 通过规则抽取的phrasal如下图所示:

| First Word | Second Word | Third Word (not extracted) |
|-----------------|-------------------|----------------------------|
| JJ | NN or NNS | anything |
| RB, RBR, RBS | JJ | Not NN nor NNS |
| JJ | JJ | Not NN or NNS |
| NN or NNS | JJ | Nor NN nor NNS |
| RB, RBR, or RBS | VB, VBD, VBN, VBG | anything |

- Step 2: Learn polarity of each phrase, 那么, 如何评价phrase的polarity呢? 直观上, 有这样的结论: “Positive phrases co-occur more with ‘*excellent*’, Negative phrases co-occur more with ‘*poor*’”, 这时, 将问题转换成如何衡量词条之间的共现关系? 于是, 学者们引入了点互信息 (Pointwise mutual information, PMI), 它经常被用于度量两个具体事件的相关程度, 公式为:

$$PMI(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)}$$

两个词条的PMI公式为:

$$PMI(word_1, word_2) = \log_2 \frac{P(word_1, word_2)}{P(word_1)P(word_2)}$$

常用的计算PMI(word1, word2)方法是分别以” word1”, ” word2” 和” word1 NEAR word2” 为query, 根据搜索引擎检索结果, 得到P(word)和P(word1, word2), 如下:

$$P(word) = hits(word)/N$$

$$P(word_1, word_2) = hits(word1 NEAR word2)/N^2$$

则有:

$$PMI(word_1, word_2) = \log_2 \frac{hits(word_1 NEAR word_2)}{hits(word_1)hits(word_2)}$$

那么, 计算一个phrase的polarity公式为 (excellent和poor也可以使用其它已知极性词代替):

$$\begin{aligned}
 Polarity(phrase) &= PMI(phrase, "excellent") - PMI(phrase, "poor") \\
 &= \log_2 \frac{hits(phrase NEAR "excellent")}{hits(phrase)hits("excellent")} - \log_2 \frac{hits(phrase NEAR "poor")}{hits(phrase)hits("poor")} \\
 &= \log_2 \frac{hits(phrase NEAR "excellent")}{hits(phrase)hits("excellent")} \frac{hits(phrase)hits("poor")}{hits(phrase NEAR "poor")} \\
 &= \log_2 \left(\frac{hits(phrase NEAR "excellent")hits("poor")}{hits(phrase NEAR "poor")hits("excellent")} \right)
 \end{aligned}$$

Turney Algorithm在410 reviews (from Epinions) 的数据集上, 其中170 (41%) negative, 240 (59%) positive, 取得了74%的准确率 (baseline为59%, 均标注为positive)。

- Step 3: Rate a review by the average polarity of its phrases

3. Using WordNet to learn polarity: 论文见S.M. Kim and E. Hovy. 2004. [Determining the sentiment of opinions](#). COLING 2004, M. Hu and B. Liu. [Mining and summarizing customer reviews](#). In Proceedings of KDD, 2004.该方法步骤如下:

- Create positive (“good”) and negative seed-words (“terrible”)
- Find Synonyms and Antonyms

Positive Set: Add synonyms of positive words (“well”) and antonyms of

negative words

Negative Set: Add synonyms of negative words (“awful”) and antonyms of positive words (“evil”)

- Repeat, following chains of synonyms
- Filter

以上几个方法都有较好的领域适应性和鲁棒性，基本思想可以概括为 “Use seeds and semi-supervised learning to induce lexicons” ，即：

- Start with a seed set of words (‘good’ , ‘poor’)
- Find other words that have similar polarity:
 - Using “and” and “but”
 - Using words that occur nearby in the same document
 - Using WordNet synonyms and antonyms
 - Use seeds and semi-supervised learning to induce lexicons

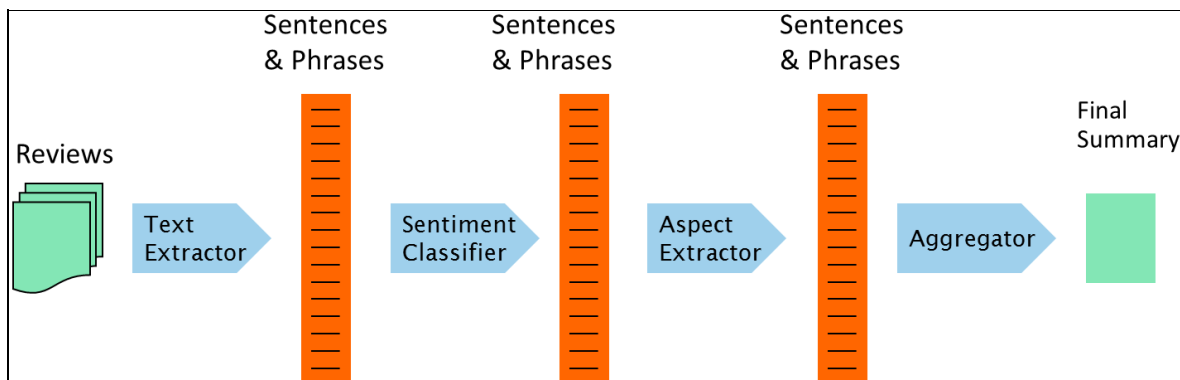
5) Other Sentiment Tasks

上面介绍了文档级或句子级情感分析，但是，实际中，一篇文档（评论）中往往会提及不同的方面/属性/对象（以下统称属性），且可能对不同的属性持有不同的倾向性，如“The **food** was **great** but the **service** was **awful**”。一般通过Frequent phrases + rules的方法抽取评价属性，如下：

- Find all highly frequent phrases across reviews (“fish tacos”)
- Filter by rules like “occurs right after sentiment word” : “...great fish tacos” means fish tacos a likely aspect

通常，我们还会面临一种问题：评价属性缺失，准确的讲，评价属性不在句子中。这是很常见的现象，此时就需要结合上下文环境，如来自某电影的评论缺失的评价属性基本上就是电影名或演员，可以基于已知评价属性的句子训练分类器，然后对评价属性缺失的句子进行属性预测。

Blair-Goldensohn et al.提出了一套通用的aspect-based summarization models，如下图所示：



详细见论文：S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar. 2008. [Building a Sentiment Summarizer for Local Service Reviews](#). WWW Workshop

另外，其他的一些情感分析的相关任务有：

- Emotion: 个人情绪
 - Detecting annoyed callers to dialogue system
 - Detecting confused/frustrated versus confident students
- Mood: 个人情绪
 - Finding traumatized or depressed writers
- Interpersonal stances: 人际关系中的谈话方式
 - Detection of flirtation or friendliness in conversations
- Personality traits: 性格
 - Detection of extroverts

三、参考资料

1. Lecture Slides: [Sentiment Analysis](#)
2. [Sentiment tutorial](#)
3. 赵妍研, [文本情感分析综述](#)



时间: 2012年 6月 24日 分类:[自然语言处理](#) 作者: [fandywang](#) (2,170 基本)
 编辑 2012年 7月 2日 作者:[fandywang](#)

[举报](#)
[回答](#)
[评论](#)

写的很赞，基本上涵盖了SA当前的研究方向。早点看到这篇文章我会少走好多弯路(泪奔)

补充一点，文章最后说道解决情感分析问题都转化为分类或者regression问题

文中没提到的rating prediction/inference问题就是转化为regression求解的，Bo pang 2005年的一篇论文《seeing stats: exploiting class relationships for sentiment categorization with respect to rating scales》，不仅仅满足于对电影评论的正负面分类，考虑更细粒度的分类问题，将评论文本分类为多个类别。比如电影评论文本分为1~5星，1星和2星之间比1星和5星更为相似，所以这种多分类问题可以看做是ordinal regression问题求解。

已评论 2012年 7月 1日 作者: [bitwjg](#)

[举报](#)
[回复](#)

谢谢，看来你是专业研究这个的哈，欢迎多提建议，在这里做些分享，最后的部分内容的确略过去了，后续我会补充上！再次感谢！



已评论 2012年 7月 1日 作者: [fandywang](#)

[举报](#)
[回复](#)

客气了

我现在的研究方向是情感分析，也是刚入门不久

刚开始的时候看了不少论文 走了不少弯路 所以看到你的文章特别有共鸣

希望以后多交流

已评论 2012年 7月 1日 作者: [bitwjg](#)

[举报](#)
[回复](#)

这个对情感分析和观点挖掘讲解的还很粗浅，要全面了解情感分析和观点挖掘的内容，建议看 Bing Liu 的新书《Sentiment Analysis and Opinion Mining》，网上找不到的话可以连系我。353718947@qq.com。有空多交流。



已评论 2012年 8月 17日 作者: [立东](#)

[举报](#)
[回复](#)

4个回答

+

+1

-

投票

通常，情感分析任务都被转化成分类或回归任务，而其中最为关键的就是特征的抽取，其中，需要特别注意几点：

1. 否定词的处理；
2. 仅使用形容词、副词并不一定可以得到最好的结果；
3. 情感词典举足轻重，且不同领域的情感词可能存在差异，一般采用bootstrapping的方法构建情感词典



已回复 2012年 7月 1日 作者: [fandywang](#) (2,170 基本)
 编辑 2012年 7月 2日 作者:[fandywang](#)

[举报](#)
[追问](#)
[评论](#)

+

-1

-

投票

这里简要介绍下[哈工大社会计算与信息检索研究中心](#)做的情感分析系统，如下：

- 爱搜车众评（[zp.isoche.com](#)）：通过挖掘汽车垂直网站网友评论，分析用户对不同属性的情感倾向，可惜现在已经下线无法访问，从[jnwang](#)[百度网盘](#)找到两张珍贵的图片，如下所示：

众评首页 > 车型浏览 > 福克斯



点击进入

满意度 [点击查看详情](#)

统计各大论坛网友的发言，自动计算获得，仅作参考

| | | |
|-----|----|-------------------|
| 油耗 | 47 | 满意: 58 不满: 82 |
| 安全性 | 95 | 满意: 60 不满: 2 |
| 空间 | 67 | 满意: 62 不满: 40 |
| 动力 | 83 | 满意: 126 不满: 33 |
| 操控 | 89 | 满意: 184 不满: 36 |
| 外观 | 92 | 满意: 283 不满: 21 |
| 内饰 | 40 | 满意: 67 不满: 93 |
| 性价比 | 68 | 满意: 73 不满: 39 |
| 配置 | 54 | 满意: 28 不满: 27 |

概览 价格 外形 油耗 内饰 安全 配置 操控 品质

以下评论来自各大汽车论坛，由爱搜车每天自动选取更新，详情可分类查看

转 刹车失灵的几种处理办法 大家学习一下 0天前
造成刹车失灵的原因很多，一是对刹车系统缺乏必要的保养，刹车总泵里杂质太多、密封不严、真空助力泵失效、刹车油过脏或几种刹车油混合使用受热后出现气阻、刹车总泵或分泵漏油、储气罐或管路接口漏气；二 ...
[巴渝车友论坛](#) 发布日期:2008-01-21 浏览:5 回复:0

优惠近万元价格欲破13万 福克斯详请 0天前
优惠，详细价格信息请见下表。**福克斯**最新价格变动报价表 **福克斯**指导价(万元)现价(万元)降幅(万元)1.8 AT 时尚 ...比较大，可以优惠8800元，而且这个价格如果到店详谈还能优惠 ... 当日价格
[汽车之家](#) 发布日期:2008-01-21 浏览:25 回复:1

有没有人喜欢这个颜色福克斯两厢? 0天前
从台湾福特六和的网站上的图，这个蓝色个人感觉真是漂亮，福特六和称之为“极光淡蓝”！我 ... 觉得，长安福特应该在运动版增加这个颜色，红黄偏女性多一些，偏男性的只有钛晶灰，而灰色又过于沉稳，与**福克斯** ...
[汽车之家](#) 发布日期:2008-01-21 浏览:78 回复:4

奇怪的长途油耗 0天前
100公里去湛江三保，晚上晚饭 ... 后保养回来了。去程我开，国道，油耗7.1升/100公里，平均速度54公里 ... 小时；回程朋友开（第一次开**focus**），高速，油耗8.0升/100公里，平均速度45公里
[爱卡汽车网](#) 发布日期:2008-01-21 浏览:40 回复:5

颐达，新宝来，福克斯，卡罗拉之我见！ 1天前
颐达：后排空间大（这是本人最喜欢的一点 ... 均衡原则。四款车中，最看中颐达，其次是宝来，**福克斯**和卡罗拉。而且后二款车比颐达要贵1W多吧，各款车都是指的最低配置。不知大家的感觉怎么样？另：现在最低配 ...
[汽车之家](#) 发布日期:2008-01-20 浏览:81 回复:1

★【新人必读】福克斯俱乐部常用帖速查★ 1天前
... 获取汽车之家车标等信息**福克斯**俱乐部QQ群：24090793——名称：**福克斯**车友俱乐部群51343164——名称：**福克斯**车友俱乐部群 ... 名称：**福克斯**车友俱乐部群351343017——名称：**福克斯**
[汽车之家](#) 发布日期:2008-01-20 浏览:377 回复:11

三厢福克斯运动版还能买到吗？请问？ 1天前
...
[汽车之家](#) 发布日期:2008-01-20 浏览:10 回复:1

车友们在对比查看...

| | |
|---------------------------------------|---------------------------------------|
| 标致307 | 蒙迪欧 |
| | |
| 看看 比比 | 看看 比比 |
| 速腾 | 马自达3 |
| | |
| 看看 比比 | 看看 比比 |
| 新宝来 | 思域 |
| | |
| 看看 比比 | 看看 比比 |

[对比更多车型](#)

爱搜车·众评 众评 经验 问答 热帖

车型搜索 [isoe有什么不一样?](#)



| | |
|-----|----|
| 油耗 | 47 |
| 安全性 | 95 |
| 空间 | 67 |
| 动力 | 83 |
| 操控 | 89 |
| 外观 | 92 |
| 内饰 | 40 |
| 性价比 | 68 |
| 配置 | 54 |

满意

一般不发言：汽车之家 2007-12-15 15:53
福克斯 外形好 [【查看原帖】](#)

cameron168：汽车之家 2007-12-13 15:58
FKS 2厢外观饱满 [【查看原帖】](#)

roeder：东莞车友网 2007-12-04 12:46
福克斯的车漆真的不错 [【查看原帖】](#)

血色浪漫：长安汽车车友会 2007-12-16 10:15
令福特福克斯的外型更符合他"动感设计"的概念 [【查看原帖】](#)

心猿意马：汽车之家 2007-09-07 12:26
今天去试驾两厢福克斯1.8MT给我个人感觉如下外型相当漂亮 [【查看原帖】](#)

roppipe：新浪汽车论坛 2006-04-26 11:13:51
福特福克斯不仅融合了先进的欧洲现代设计风尚 [【查看原帖】](#)

wufajiesdm：Tom汽车论坛 2007-05-14 20:27
两厢福克斯外形极具动感 [【查看原帖】](#)

berpou：Tom汽车论坛 2007-02-07 08:41
造型养眼型：福克斯两厢福特福克斯两厢的造型显得非常时尚动感 [【查看原帖】](#)

berpou：Tom汽车论坛 2007-02-07 08:41
时尚造型是福克斯两厢获得成功的关键因素 [【查看原帖】](#)

berpou：Tom汽车论坛 2007-02-07 08:41
福克斯两厢还一举夺得了2006CCTV年度车评选中的“最佳造型车”大奖 [【查看原帖】](#)

德国一马克：Tom汽车论坛 2007-07-11 14:02
福克斯外形还不错 [【查看原帖】](#)

不满

xiaode555：Tom汽车论坛 2007-07-22 10:22
FKS这边说307的屁股难看 [【查看原帖】](#)

休闲人：汽车之家 2007-09-09 19:46
也是福克斯设计最差又容易撞车之处 [【查看原帖】](#)

linhanheman：网易汽车论坛 2006-12-03 22:07
过小的车内外后视镜影响了福克斯对外信息的反馈 [【查看原帖】](#)

荒荒天使：爱卡汽车网 2006-12-08 23:01
觉得两厢福克斯的屁股远没有VW的好看 [【查看原帖】](#)

飞知：汽车之家 2008-01-08 14:51
造成福克斯的左前轮眉和我的S右侧前门来了个接触 [【查看原帖】](#)

huanggangb：中国汽车网 2007-11-21 11:04
福克斯的手套箱设计的很不好 [【查看原帖】](#)

linhanheman：网易汽车论坛 2007-05-30 11:26
我个人感觉福克斯三厢的尾灯不好看 [【查看原帖】](#)

zjx781027：汽车之家 2007-11-02 10:48
而且老婆认为307的外形比FKS好看 [【查看原帖】](#)

geshijw：车友社区 2007-09-23 08:37
但凯悦的造型实在是非常的漂亮（至今我仍然认为凯悦比福克斯还好看 [【查看原帖】](#)

ar8633：汽车之家 2007-07-30 13:59
如果说景程的储物点设计比福克斯周到的话 [【查看原帖】](#)

219.142.178.*：百度汽车吧 2006-12-22 13:25
ST的外形比FKS更时尚运动 [【查看原帖】](#)

梦想福克斯：汽车之家 2007-11-16 16:46
...
[汽车之家](#) 发布日期:2007-11-16 16:46

- 八维音乐挑歌 (<http://yue.8wss.com/screener/>)：通过挖掘豆瓣音乐评论，分析用户对不同歌曲、专辑、歌手的情感倾向，并基于此提供挑歌功能，这完全不同于Google音乐搜索的基于人物、年代、音频的挑歌功能，如下图所示：



请尝试输入关键词: 欢快的 空灵的 厚重的 高亢的 ...

八维搜索

音乐首页

音乐新闻

音乐人物

歌曲专辑

挑歌

预告

<< 隐藏左侧面板

- ★ 欢快的
- ★ 空灵的
- ★ 厚重的
- ★ 高亢的
- ★ 低沉的
- ★ 舒缓的

欢快的音乐:

1. 牛仔很忙
2. 礼物
3. EVERYTHING
4. I'm Yours
5. 冰菊物语

- 欢快的 旋律
- 欢快的 节奏
- 欢快的 氛围
- 欢快的 歌声
- 欢快的 歌曲

更多>>

空灵的音乐:

1. 暗涌
2. 天空
3. 想哭
4. 琴麻岛的海
5. I Do

- 空灵的 感觉
- 空灵的 声音
- 空灵的声音
- 空灵的 编曲
- 空灵的和声

更多>>

厚重的音乐:

1. LOVE SONG
2. 简单
3. Meds
4. Final Straw
5. 是我

- 厚重的 全曲
- 厚重的 音墙
- 厚重的 层次感
- 厚重的 专辑
- 厚重的 Bass

更多>>

高亢的音乐:

1. 消失
2. Soul
3. Goodbye
4. 我就是这样
5. 杀破狼

- 高亢的 声线
- 高亢的 声线
- 高亢的 噪音
- 高亢的 副歌
- 高亢的 歌曲

更多>>

低沉的音乐:

1. Nebelung

- 低沉的 吉他

更多>>



欢快

八维搜索

音乐首页

音乐新闻

音乐人物

歌曲专辑

挑歌

预告

<< 隐藏左侧面板

- ★ 欢快的
- ★ 空灵的
- ★ 厚重的
- ★ 高亢的
- ★ 低沉的
- ★ 舒缓的

欢快的音乐有:

1. 牛仔很忙
2. 礼物
3. EVERYTHING
4. I'm Yours
5. 冰菊物语
6. 小快乐
7. 泥娃娃
8. 约定
9. 我们的歌
10. 威廉古堡
11. Fly with you
12. 第一次
13. My Love
14. Good Morning
15. 轻微
16. 不由自主
17. 星象仪
18. Driving
19. 择偶条件
20. 一个像夏天一个像秋天

- 欢快的 旋律
- 欢快的 节奏
- 欢快的 氛围
- 欢快的 歌声
- 欢快的 歌曲
- 欢快的小调
- 欢快的 童谣
- 欢快的 词曲
- 欢快的 演绎
- 欢快的 歌曲
- 欢快的 鼓点节奏
- 欢快的 歌曲
- 欢快的 歌曲
- 欢快的 歌曲
- 欢快的 歌曲
- 欢快的 歌曲
- 欢快的情歌
- 欢快的 歌曲
- 欢快的 和弦
- 欢快的 歌曲

1 2 3 4 5 6 7 8 9 10 下一页

八维音乐

欢快

八维搜索

音乐首页 | 音乐新闻 | 音乐评论

<< 隐藏左侧面板

★ 欢快的

★ 空灵的

★ 厚重的

★ 高亢的

★ 低沉的

★ 舒缓的

评论来源

• ... 唱片的五首歌，分别采用了五种不同的音乐风格，不管是Indiepop也好或者Chillout、Folkpop、Air电子，我们体会到的是《塞宁》中拥有淡淡失意的歌手塞宁（小说人物），《玛格丽特》里抑郁自闭的失忆女孩，《蕾丝边》中永远充满活力与希望的私奔女王蕾丝边，《**轻微**》里**欢快**、天真烂漫的轻微姑娘。...

来源：豆瓣 作者：朴九月 发表日期：2006-11-15

<http://www.douban.com/review/1090747>

• ... 唱片的五首歌，分别采用了五种不同的音乐风格，不管是Indiepop也好或者Chillout、Folkpop、Air电子，我们体会到的是《塞宁》中拥有淡淡失意的歌手塞宁（小说人物），《玛格丽特》里抑郁自闭的失忆女孩，《蕾丝边》中永远充满活力与希望的私奔女王蕾丝边，《**轻微**》里**欢快**、天真烂漫的轻微姑娘。...

来源：豆瓣 作者：朴九月 发表日期：2006-11-15

<http://www.douban.com/review/1090747>

10. 飘像风一样

11. Fly with you

12. 第一次

13. My Love

14. Good Morning

15. 轻微

16. 不由自主

17. 星象仪

18. Driving

19. 择偶条件

20. 一个像夏天一个像秋天

欢快的 旋律

欢快的 节奏

欢快的 氛围

欢快的 歌声

欢快的 歌曲

欢快的小调

欢快的 童谣

欢快的 词曲

欢快的 演绎

欢快的 歌曲

欢快的 鼓点节奏

欢快的 歌曲

欢快的 歌曲

欢快的 歌曲

欢快的 歌曲

欢快的情歌

欢快的 歌曲

欢快的 和弦

欢快的 歌曲

1

2

3

4

5

6

7

8

9

10

下一页



已回复 2012年 7月 1日 作者: **fandywang** (2,170 基本)
编辑 2012年 7月 2日 作者:**fandywang**

举报 追问 评论