社交网络与舆情预测

第二讲 自然语言处理基础：情感计算

计算机学院 任昭春

每周课程

- 第1周 课程简介（已讲）
- 第2周 NLP基础：语言模型，情感计算
- 第3周 NLP基础：自动文摘
- 第4周 NLP基础：文档分类与聚类
- 第5周 NLP基础：问答系统与对话
- 第6周 机器学习基础：主题建模
- 第7周 机器学习基础：word2vec与embedding方法
- 第8周 机器学习基础：基础神经网络模型简介
- 第9周 图与网络
- 第10周 社区分析
- 第11周 社交网络中的信息传播1
- 第12周 社交网络中的信息传播2
- 第13周 社交媒体内容的自动摘要
- 第14周 社交媒体的聚类与分类
- 第15周 基于社交媒体的推荐技术与预测应用
- 第16周 前沿讲座 （From an invited speaker）
- 第17周 复习与回顾

课程信息

- 1. 微信群

- 2. 课程内容

社交网络与舆情发现课程微信群

该二维码7天内(9月10日前)有效，重新进入将更新

https://share.weiyun.com/5yBGMQq-g

自然语言处理基础

- 语言建模
- 情感计算

语言建模 （Language Modeling）

Borrowed from Stanford NLP course slides

Probabilistic Language Models

- Today’s goal: assign a probability to a sentence
    - Machine Translation:
        - P(**high** winds tonite) > P(**large** winds tonite)
    - Spell Correction
        - The office is about fifteen **minuets** from my house
            - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
    - Speech Recognition
        - P(I saw a van) >> P(eyes awe of an)
    - + Summarization, question-answering, etc., etc.!!

Probabilistic Language Models

- Goal: compute the probability of a sentence or sequence of words:

  $P(W) = P(w_1,w_2,w_3,w_4,w_5\ldots w_n)$

- Related task: probability of an upcoming word:

  $P(w_5|w_1,w_2,w_3,w_4)$

- A model that computes either of these:

  $P(W)$ or $P(w_n|w_1,w_2\ldots w_{n-1})$ is called a **language model**.

- Better: **the grammar** But **language model** or **LM** is standard

How to compute P(W)

- How to compute this joint probability:
  - P(its, water, is, so, transparent, that)
- Intuition: let’s rely on the Chain Rule of Probability

Reminder: The Chain Rule

- Recall the definition of conditional probabilities

Rewriting:

- More variables:

$$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$$

- The Chain Rule in General

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

The Chain Rule applied to compute joint probability of words in sentence

$$P(w_1 w_2 \ldots w_n) = \prod_i P(w_i \mid w_1 w_2 \ldots w_{i-1})$$

P("its water is so transparent") =

P(its) × P(water|its) × P(is|its water)

× P(so|its water is) × P(transparent|its water is so)

How to estimate these probabilities

- Could we just count and divide?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- No! Too many possible sentences!
- We’ll never see enough data for estimating these

Markov Assumption

Andrei Markov

- Simplifying assumption:

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

- Or maybe

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

Markov Assumption

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i \mid w_{i-k} \ldots w_{i-1})$$

- In other words, we approximate each component in the product

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-k} \ldots w_{i-1})$$

Bigram model

- Condition on the previous word:

$$P(w_i \mid w_1 w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$$

```
texaco, rose, one, in, this, issue, is, pursuing, growth, in,
a, boiler, house, said, mr., gurria, mexico, 's, motion,
control, proposal, without, permission, from, five, hundred,
fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

this, would, be, a, record, november
```

N-gram models

- We can extend to trigrams, 4-grams, 5-grams
- In general this is an insufficient model of language
  - because language has **long-distance dependencies**:

  "The computer which I had just put into the machine room on the fifth floor crashed."

- But we can often get away with N-gram models

Estimating N-gram Probabilities

Estimating bigram probabilities

- The Maximum Likelihood Estimate

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

An example

$$P(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i)}{c(w_{i-1})}$$

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

$$P(\texttt{I} \mid \texttt{<s>}) = \tfrac{2}{3} = .67 \qquad P(\texttt{Sam} \mid \texttt{<s>}) = \tfrac{1}{3} = .33 \qquad P(\texttt{am} \mid \texttt{I}) = \tfrac{2}{3} = .67$$

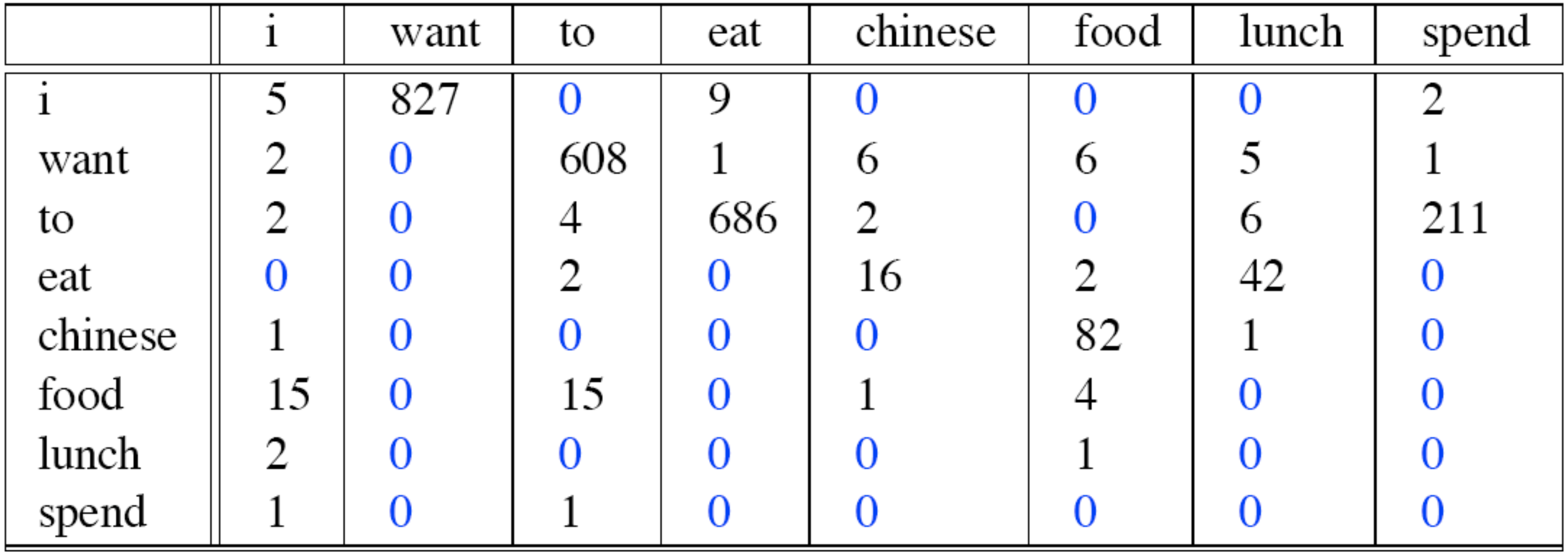
$$P(\texttt{</s>} \mid \texttt{Sam}) = \tfrac{1}{2} = 0.5 \qquad P(\texttt{Sam} \mid \texttt{am}) = \tfrac{1}{2} = .5 \qquad P(\texttt{do} \mid \texttt{I}) = \tfrac{1}{3} = .33$$

More examples:
Berkeley Restaurant Project sentences

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i’m looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i’m looking for a good place to eat breakfast
- when is caffe venezia open during the day

Raw bigram counts

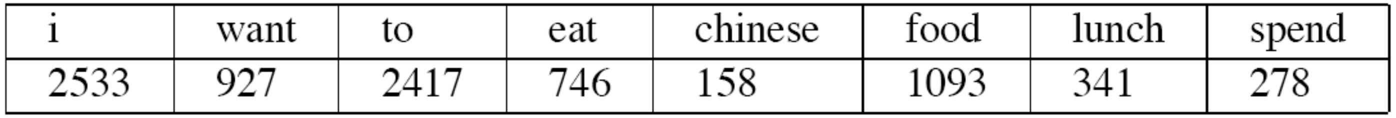
- Out of 9222 sentences



	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

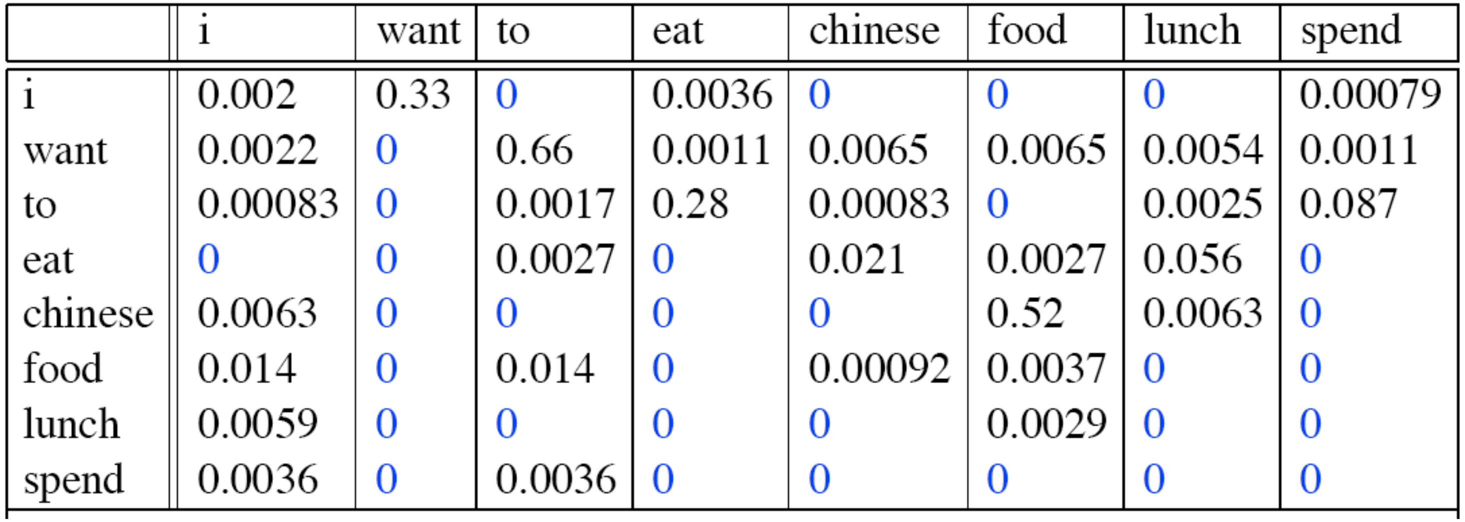
Raw bigram probabilities

- Normalize by unigrams:



i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Result:



	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Bigram estimates of sentence probabilities

P(<s> I want english food </s>) =

P(I|<s>)

× P(want|I)

× P(english|want)

× P(food|english)

× P(</s>|food)

= .000031

What kinds of knowledge?

- $P(\text{english}|\text{want}) = .0011$
- $P(\text{chinese}|\text{want}) = .0065$
- $P(\text{to}|\text{want}) = .66$
- $P(\text{eat} \mid \text{to}) = .28$
- $P(\text{food} \mid \text{to}) = 0$
- $P(\text{want} \mid \text{spend}) = 0$
- $P(\text{i} \mid \text{<s>}) = .25$

Practical Issues

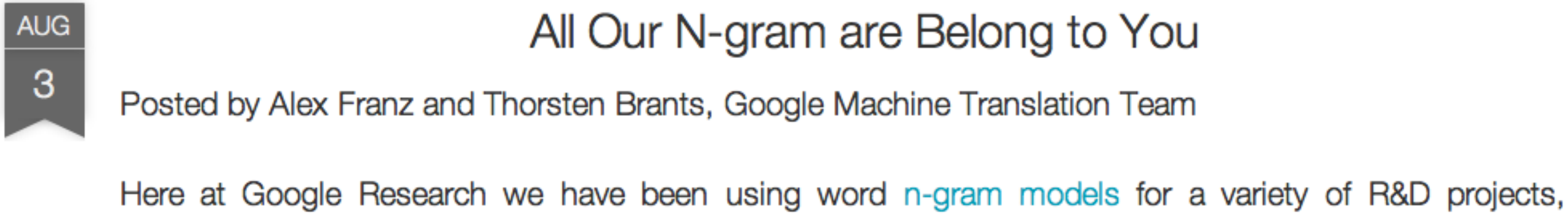
- We do everything in log space
  - Avoid underflow
  - (also adding is faster than multiplying)

$$\log(p_1 \times p_2 \times p_3 \times p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Language Modeling Toolkits

- SRILM
  - http://www.speech.sri.com/projects/srilm/
- IRSTLM
  - (http://hlt.fbk.eu/en/irstlm)
- MITLM
  - (http://code.google.com/p/mitlm/)
- BerkeleyLM
  - http://code.google.com/p/berkeleylm/

Google N-Gram Release, August 2006



AUG

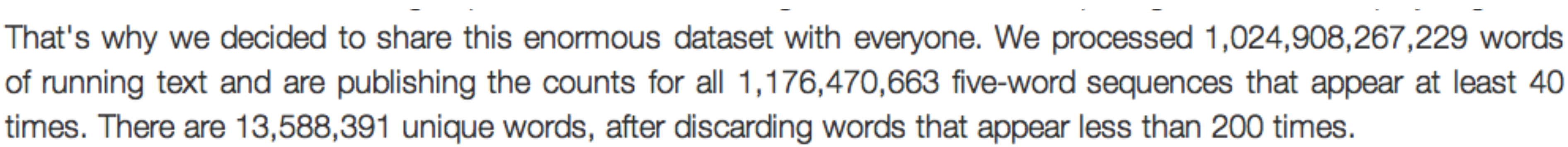
3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

…



That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

Google N-Gram Release

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Google Book N-grams

- https://aws.amazon.com/cn/datasets/google-books-ngrams/

Evaluation and Perplexity

Evaluation: How good is our model?

- Does our language model prefer good sentences to bad ones?
    - Assign higher probability to “real” or “frequently observed” sentences
        - Than “ungrammatical” or “rarely observed” sentences?
- We train parameters of our model on a **training set**.
- We test the model’s performance on data we haven’t seen.
    - A **test set** is an unseen dataset that is different from our training set, totally unused.
    - An **evaluation metric** tells us how well our model does on the test set.

Extrinsic evaluation of N-gram models

- Best evaluation for comparing models A and B
  - Put each model in a task
    - spelling corrector, speech recognizer, MT system
  - Run the task, get an accuracy for A and for B
    - How many misspelled words corrected properly
    - How many words translated correctly
  - Compare accuracy for A and B

Difficulty of extrinsic evaluation of N-gram models

- Extrinsic evaluation
  - Time-consuming; can take days or weeks
- So
  - Sometimes use **intrinsic** evaluation: **perplexity**
  - Bad approximation
    - unless the test data looks **just** like the training data
    - So **generally only useful in pilot experiments**
  - But is helpful to think about.

Intuition of Perplexity

- Let’s suppose a sentence consisting of random digits
- What is the perplexity of this sentence according to a model that assign P=1/10 to each digit?

$$
\begin{aligned}
\mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\
&= (\frac{1}{10}^{N})^{-\frac{1}{N}} \\
&= \frac{1}{10}^{-1} \\
&= 10
\end{aligned}
$$

Lower perplexity = better model

- Training 38 million words, test 1.5 million words, WSJ

N-gram Order	Unigram	Bigram	Trigram
Perplexity	962	170	109

情感分析 （sentiment analysis）

简介

- **两种类型的文本信息**
  - **事实(Facts) 与观点(Opinions)**
- **大多数文本信息处理技术 (e.g., web search, text mining) 面向事实信息.**
- **情感分析/观点挖掘**
  - **对文本中表达的观点与情感进行计算与分析**
- **为什么现在需要情感分析？**
  - **主要因为互联网上有海量的观点文本**

简介

- **观点的重要性**
  - **当我们需要做决定时观点/意见会很重要, 我们通常需要听听其他人的意见.**
  - **过去,**
    - **对于个人: 朋友, 家庭**
    - **对于商务: 调查, 咨询**
  - **现在，互联网上有大量用户生成内容，表达对任意事物的观点**
    - **可以帮助个人或商务进行参考决策**

应用

- **产品比较与推荐**
- **个人与机构声誉分析**
- **电视节目满意度分析**
- **互联网舆情分析**
  - **利用文本情感计算技术深入分析人们对社会现实和现象的群体性情绪、观点、思想、心理、意志和要求；**

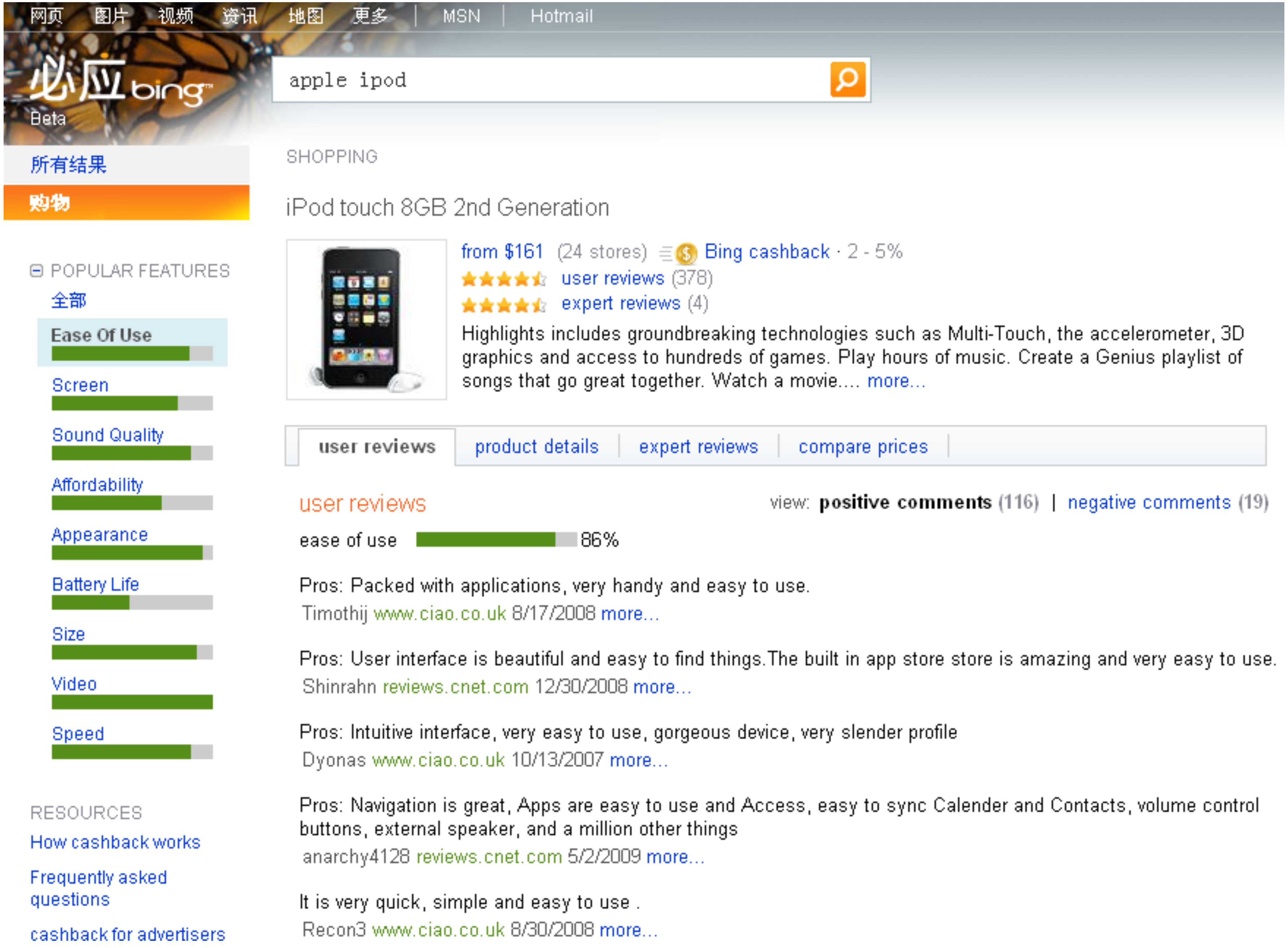
例子

美国花大钱全球找"坏话 “

- ……
- 通过"情绪分析"寻找威胁
- 长久以来，美国官员一直依赖报纸和其他消息来源，追踪美国和海外发生的事 件和舆论。据美国国土安全部官员透露，海外报纸及其他刊物对美国或美国领袖 的负面看法，可能会暗示恐怖分子活动的蛛丝马迹。他们将这种负面信息的收集 分析称作海外"情绪分析"，通过了解"报道词句及用词的情绪有多强烈"，帮助美 国情报人员找出美国可能面临的威胁以及这些威胁的常见类型。
- 为此，国土安全部斥资240万美元作为研究经费，帮助研究机构开发先进的软 件系统来更快、更全面地监视全球媒体。参与研究的包括康奈尔大学、匹兹堡大 学、犹他大学等美国著名学府。开发工作大概需要数年的时间才能完成。
- ……
- 来源：参考消息，CRI国际在线
- http://gb.cri.cn/12764/2006/10/09/2225@1248813.htm
- 国土安全部支持的系统为CERATOPS http://www.cs.pitt.edu/mpqa/ceratops/

原型系统与产品

- **英文：**
  - **OpinionFinder**
  - **RapidMiner**
  - **TextMap**
  - **微软Bing产品搜索**
  - **condensr.com**
  - **tweetfeel.com**
- **中文:**
  - **爱搜车众评：抽取汽车评论中的评价对象和褒贬倾向，对比展示**
  - **雅虎人物搜索**



网页 图片 视频 资讯 地图 更多 | MSN | Hotmail
必应 bing Beta
apple ipod
所有结果
购物

SHOPPING

iPod touch 8GB 2nd Generation

from $161 (24 stores) Bing cashback · 2 - 5%

★★★★½ user reviews (378)

★★★★½ expert reviews (4)

Highlights includes groundbreaking technologies such as Multi-Touch, the accelerometer, 3D graphics and access to hundreds of games. Play hours of music. Create a Genius playlist of songs that go great together. Watch a movie.... more...

user reviews | product details | expert reviews | compare prices

⊟ POPULAR FEATURES

- 全部
- Ease Of Use
- Screen
- Sound Quality
- Affordability
- Appearance
- Battery Life
- Size
- Video
- Speed

user reviews

view: **positive comments** (116) | negative comments (19)

ease of use 86%

Pros: Packed with applications, very handy and easy to use.
Timothij www.ciao.co.uk 8/17/2008 more...

Pros: User interface is beautiful and easy to find things.The built in app store store is amazing and very easy to use.
Shinrahn reviews.cnet.com 12/30/2008 more...

Pros: Intuitive interface, very easy to use, gorgeous device, very slender profile
Dyonas www.ciao.co.uk 10/13/2007 more...

Pros: Navigation is great, Apps are easy to use and Access, easy to sync Calender and Contacts, volume control buttons, external speaker, and a million other things
anarchy4128 reviews.cnet.com 5/2/2009 more...

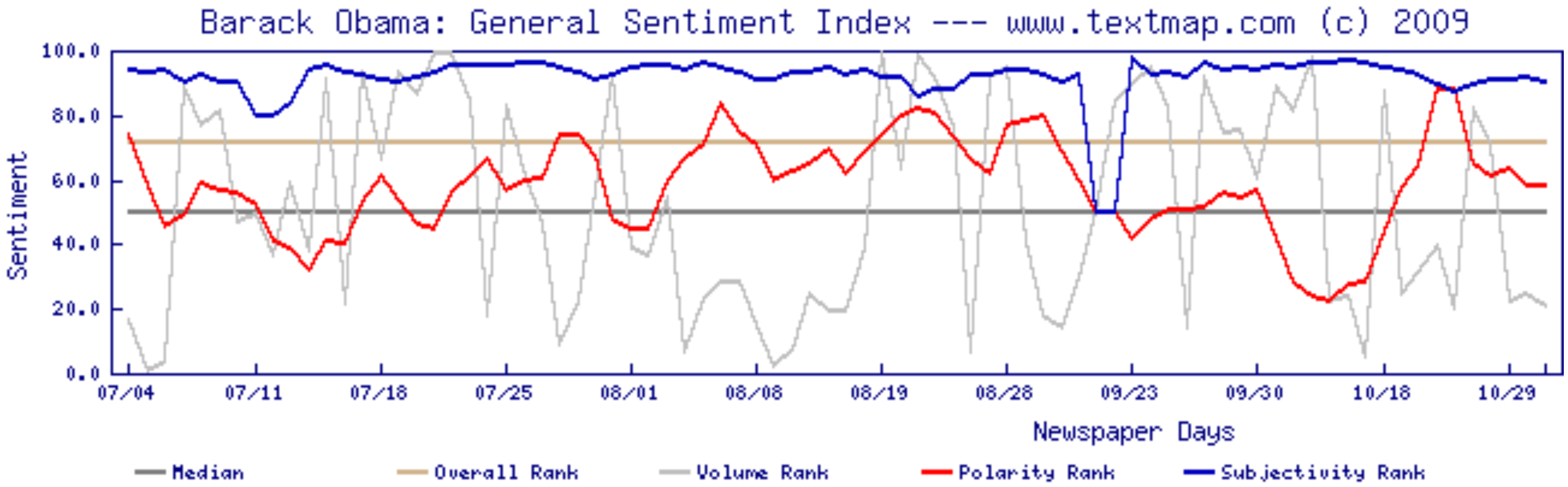
It is very quick, simple and easy to use .
Recon3 www.ciao.co.uk 8/30/2008 more...

RESOURCES

How cashback works

Frequently asked questions

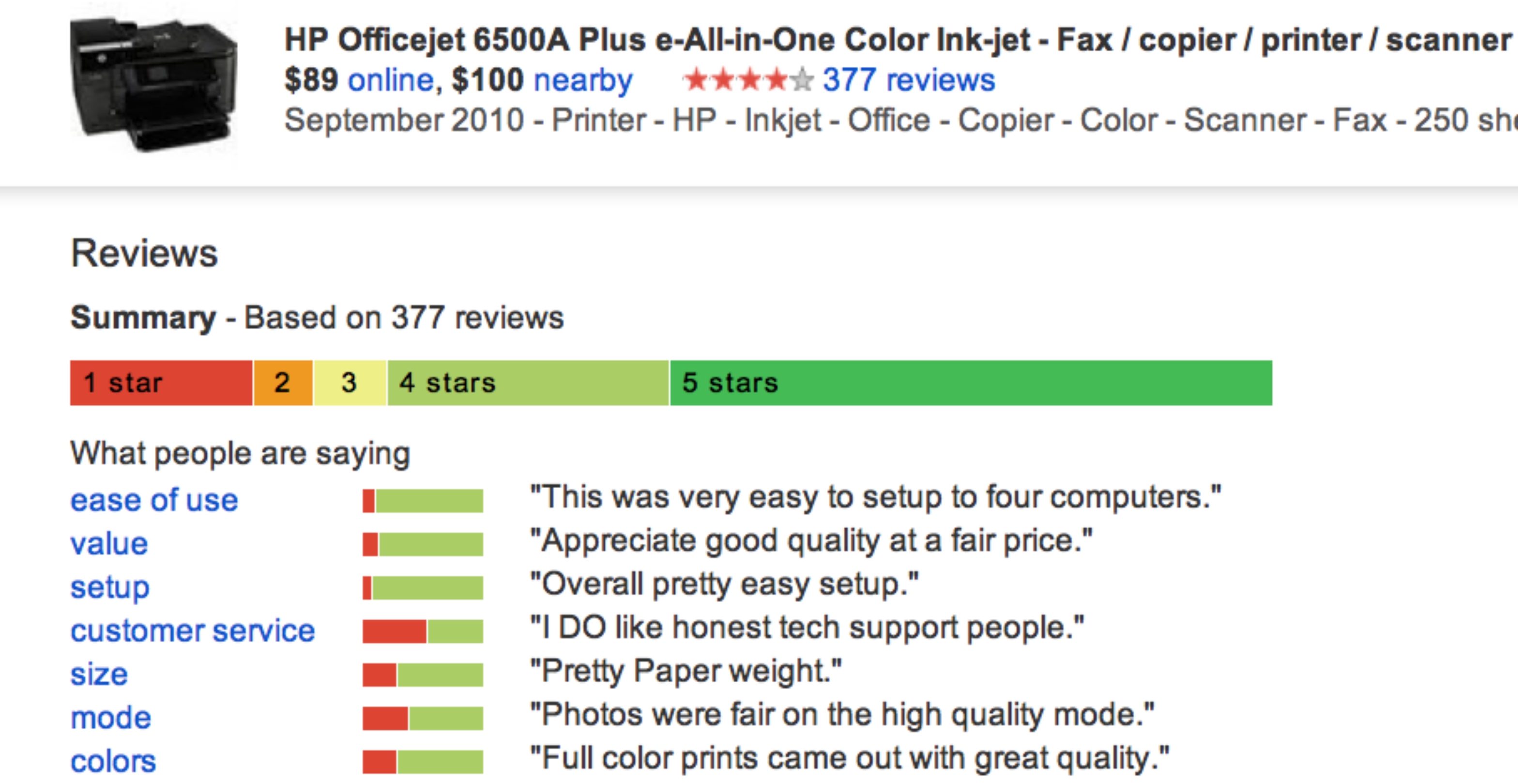
cashback for advertisers


Barack Obama: General Sentiment Index --- www.textmap.com (c) 2009
Sentiment
100.0
80.0
60.0
40.0
20.0
0.0
07/04
07/11
07/18
07/25
08/01
08/08
08/19
08/28
09/23
09/30
10/18
10/29
Newspaper Days
Median
Overall Rank
Volume Rank
Polarity Rank
Subjectivity Rank

Positive or negative movie review?

- 👎 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.

Google Product Search



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

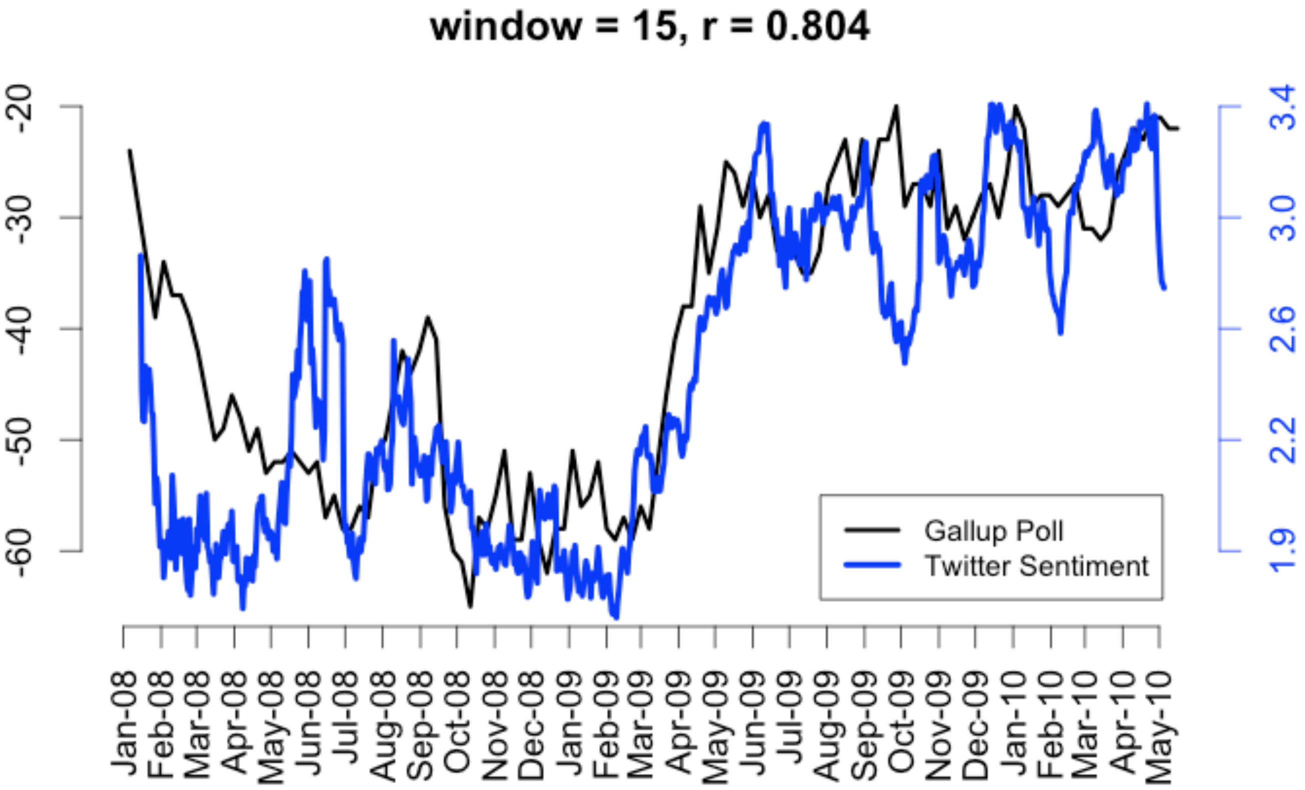
Summary - Based on 377 reviews

What people are saying

ease of use	"This was very easy to setup to four computers."
value	"Appreciate good quality at a fair price."
setup	"Overall pretty easy setup."
customer service	"I DO like honest tech support people."
size	"Pretty Paper weight."
mode	"Photos were fair on the high quality mode."
colors	"Full color prints came out with great quality."

Twitter sentiment versus Gallup Poll of Consumer Confidence

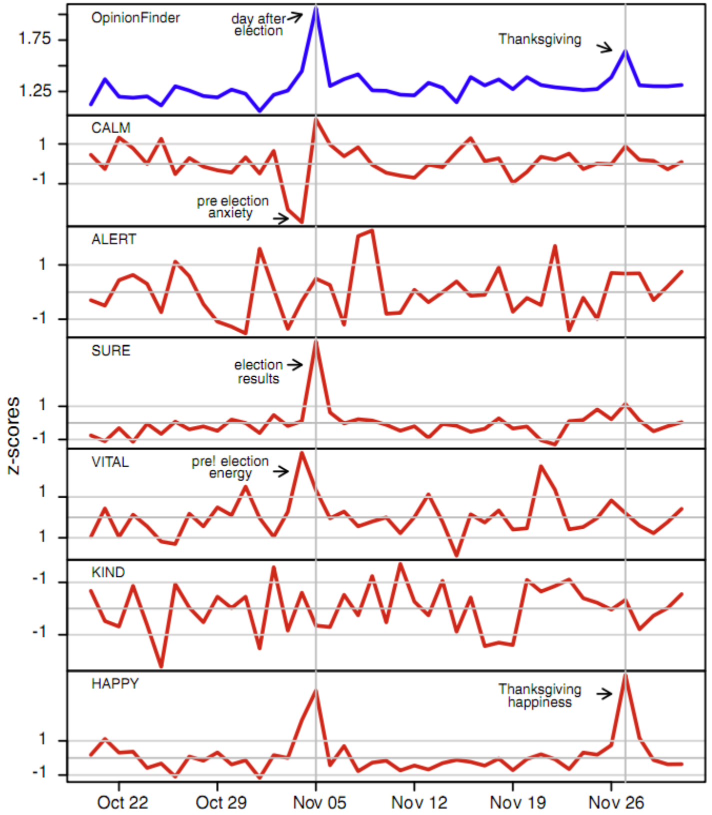
Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In ICWSM-2010



Twitter sentiment:

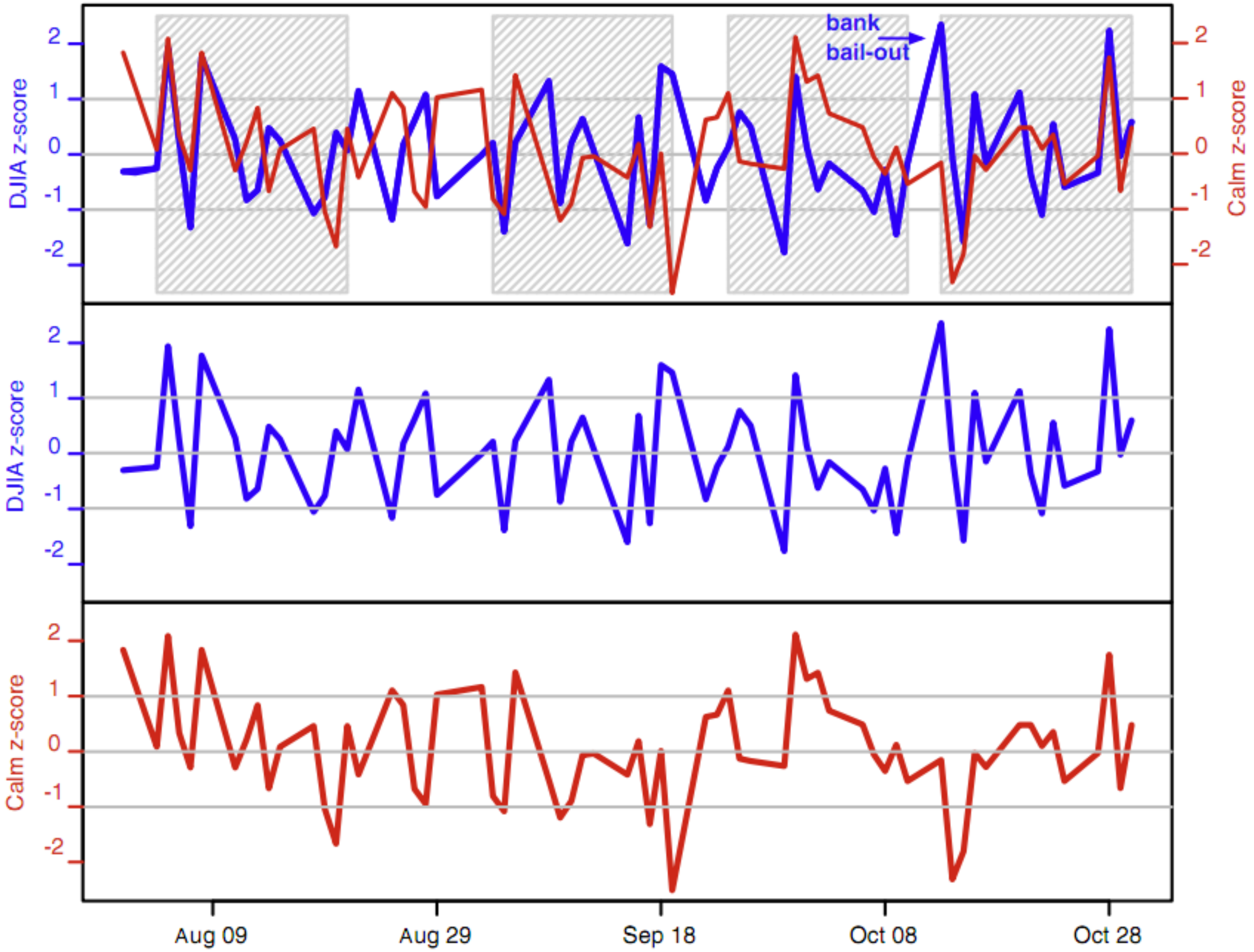
Johan Bollen, Huina Mao, Xiaojun Zeng. 2011. Twitter mood predicts the stock market,

Journal of Computational Science 2:1, 1-8. 10.1016/j.jocs.2010.12.007.



Bollen et al. (2011)

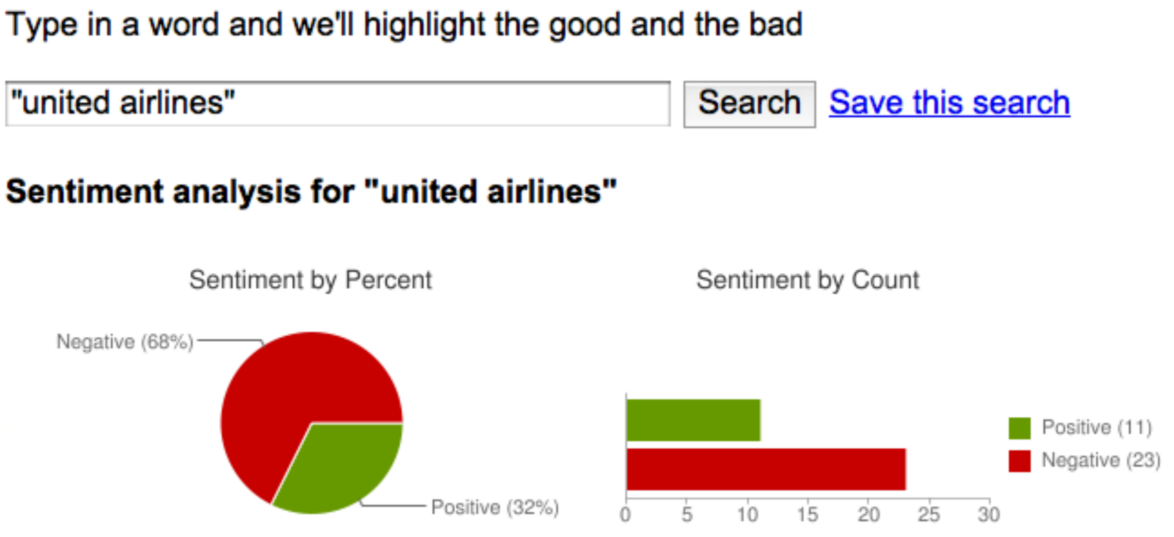
- CALM predicts DJIA 3 days later
- At least one current hedge fund uses this algorithm



Target Sentiment on Twitter

- Twitter Sentiment App

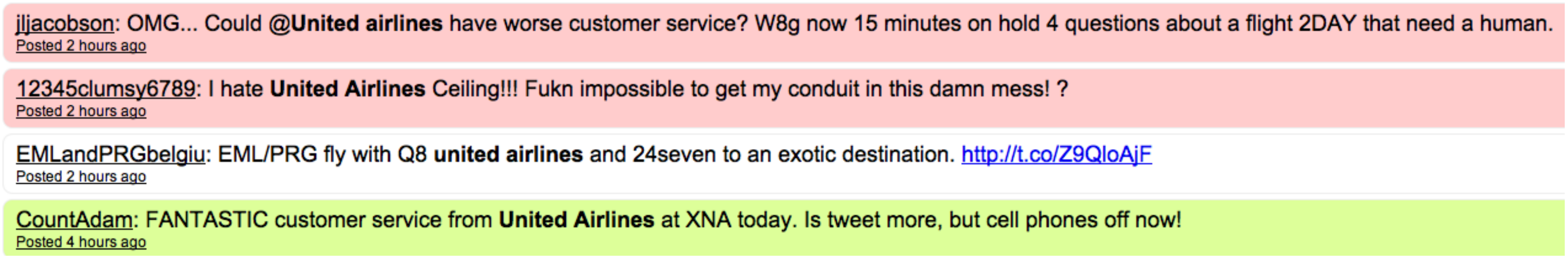
- Alec Go, Richa Bhayani, Lei Huang. 2009. Twitter Sentiment Classification using Distant Supervision



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

"united airlines" Search Save this search

Sentiment analysis for "united airlines"



jljacobson: 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

Why sentiment analysis?

- *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

Sentiment Analysis

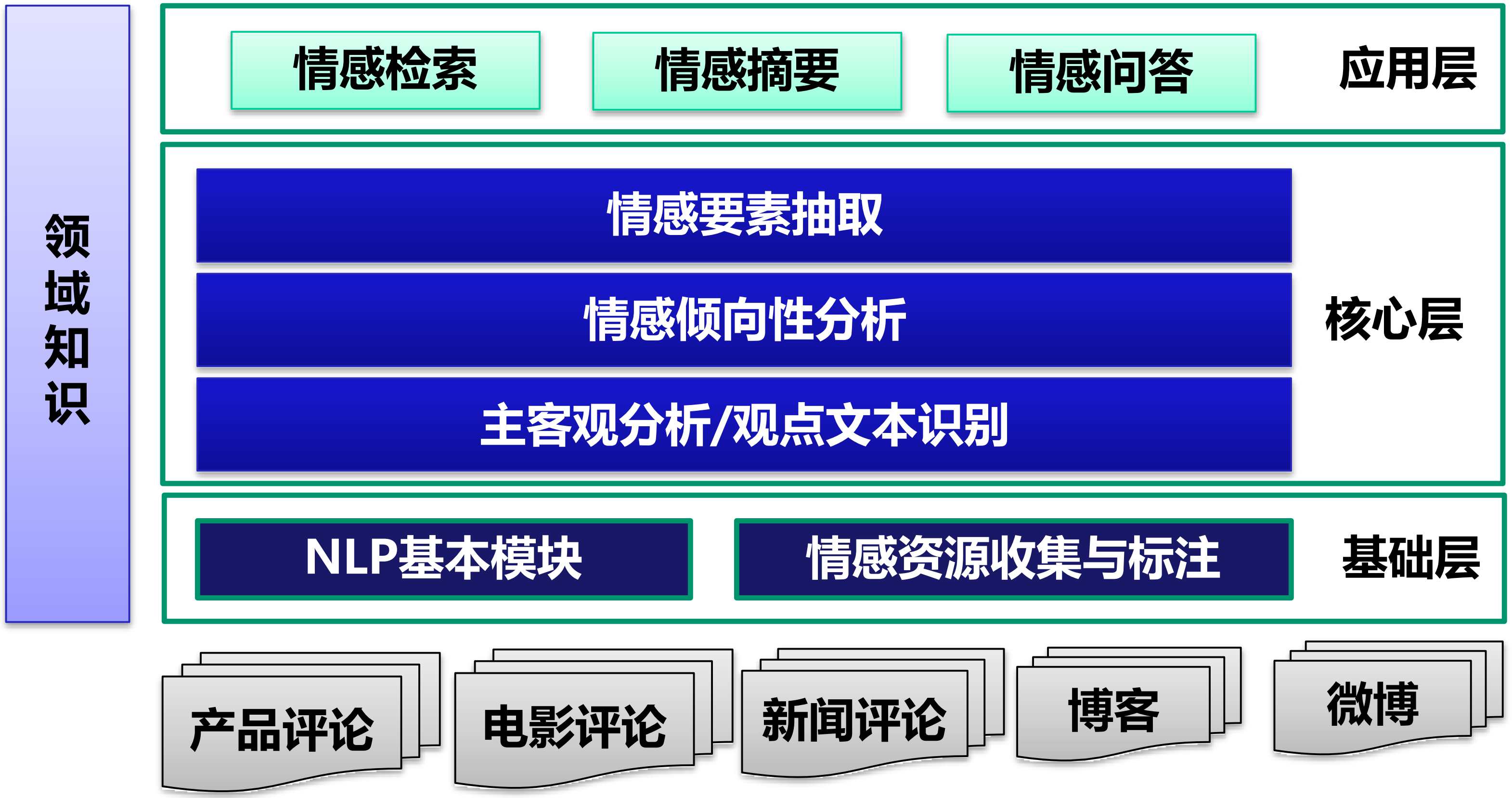
- Sentiment analysis is the detection of **attitudes**

  1. **Holder (source)** of attitude
  2. **Target (aspect)** of attitude
  3. **Type** of attitude
     - 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
     - Sentence or entire document

Sentiment Analysis

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

研究框架



领域知识
情感检索
情感摘要
情感问答
应用层
情感要素抽取
情感倾向性分析
主客观分析/观点文本识别
核心层
NLP基本模块
情感资源收集与标注
基础层
产品评论
电影评论
新闻评论
博客
微博

情感分类

- **将文本按照所表达的总体情感进行分类**
  - **例如：正面(Positive), 负面(negative), (possibly) 中性 (neutral)**

- **与基于话题的文本分类相似又不同**
  - **对于基于话题的文本分类, 话题词汇很重要**
  - **情感分类中，情感词汇更加重要，例如 great, excellent, horrible, bad, worst, etc.**

情感分类任务

- **主客观分析/观点文本识别**
  - 客观：反映关于世界的事实信息，“北京是中国的首都”
  - 主观：反映个人情感、信念等，“我爱北京天安门”
- **倾向性分析(可看作主客观分析的细粒度处理)**
  - 对包含观点的文本进行倾向性判断
  - 一般为以下三类
    - 褒义： “外观不错”
    - 贬义： “软件目前不丰富”
    - 中性： “我认为中国需要治理环境”
      - 在一些问题中不考虑中性
- **情绪分析**
  - 用户情绪识别：愤怒、高兴、喜好、悲哀、吃惊，等
- **粒度**
  - 词、句子、文档

A base algorithm

Sentiment Classification in Movie Reviews

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

- Polarity detection:
  - Is an IMDB movie review positive or negative?
- Data: *Polarity Data 2.0:*
  - http://www.cs.cornell.edu/people/pabo/movie-review-data

IMDB data in the Pang and Lee database

✓

when _star wars_ came out some twenty years ago , the image of traveling throughout the stars has become a commonplace image . […]

when han solo goes light speed , the stars change to bright lines , going towards the viewer in lines that converge at an invisible point .

cool .

october sky offers a much simpler image–that of a single white dot , traveling horizontally across the night sky . [. . .]

✗

“ snake eyes ” is the most aggravating kind of movie : the kind that shows so much potential then becomes unbelievably disappointing .

it’s not just because this is a brian depalma film , and since he’s a great director and one who’s films are always greeted with at least some fanfare .

and it’s not even because this was a film starring nicolas cage and since he gives a brauvara performance , this film is hardly worth his talents .

Baseline Algorithm (adapted from Pang and Lee)

- Tokenization
- Feature Extraction
- Classification using different classifiers
    - Naïve Bayes
    - MaxEnt
    - SVM

Sentiment Tokenization Issues

- Deal with HTML and XML markup
- Twitter mark-up (names, hash tags)
- Capitalization (preserve for words in all caps)
- Phone numbers, dates
- Emoticons
- Useful code:
  - Christopher Potts sentiment tokenizer
  - Brendan O'Connor twitter tokenizer

Potts emoticons

```
[<>]?                       # optional hat/brow
[:;=8]                      # eyes
[\-o\*\']?                  # optional nose
[\)\]\(\[dDpP/\:\}\{@\|\\]  # mouth
|                           #### reverse orientation
[\)\]\(\[dDpP/\:\}\{@\|\\]  # mouth
[\-o\*\']?                  # optional nose
[:;=8]                      # eyes
[<>]?                       # optional hat/brow
```

Extracting Features for Sentiment Classification

- How to handle negation
  - `I` **`didn’t`** `like this movie`

  vs

  - `I really like this movie`
- Which words to use?
  - Only adjectives
  - All words
    - All words turns out to work better, at least on this data

Negation

Das, Sanjiv and Mike Chen. 2001. Yahoo! for Amazon: Extracting market sentiment from stock message boards. In Proceedings of the Asia Pacific Finance Association Annual Conference (APFA).

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment Classification using Machine Learning Techniques. EMNLP-2002, 79—86.

Add NOT_ to every word between negation and following punctuation:

```
didn’t like this movie , but I
```

↓

```
didn’t NOT_like NOT_this NOT_movie but I
```

Reminder: Naïve Bayes

$$c_{NB} = \operatorname*{argmax}_{c_{\mathrm{j}} \in C} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

$$\hat{P}(w \mid c) = \frac{count(w,c) + 1}{count(c) + |V|}$$

Binarized (Boolean feature) Multinomial Naïve Bayes

- Intuition:
  - For sentiment (and probably for other text classification domains)
  - Word occurrence may matter more than word frequency
    - The occurrence of the word *fantastic* tells us a lot
    - The fact that it occurs 5 times may not tell us much more.
  - Boolean Multinomial Naïve Bayes
    - Clips all the word counts in each document at 1

Boolean Multinomial Naïve Bayes: Learning

- From training corpus, extract *Vocabulary*

- Calculate $P(c_j)$ terms
  - For each c_j in C do

    $docs_j \leftarrow$ all docs with class $=c_j$

$$P(c_j) \leftarrow \frac{|docs_j|}{|\text{total \# documents}|}$$

- Calculate $P(w_k \mid c_j)$ terms

- $Text_j \leftarrow$ single doc containing all $docs_j$
- For each word w_k in *Vocabulary*

  $n_k \leftarrow$ # of occurrences of w_k in $Text_j$

$$P(w_k \mid c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

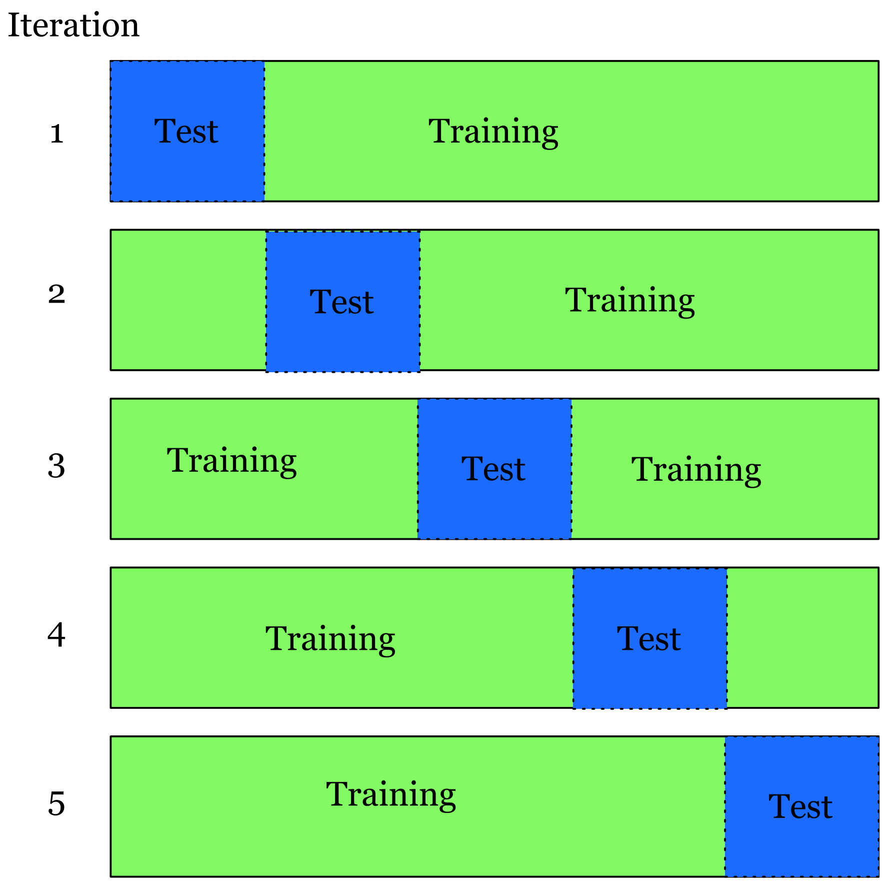
Boolean Multinomial Naïve Bayes on a test document *d*

- First remove all duplicate words from *d*
- Then compute NB using the same equation:

$$c_{NB} = \underset{c_{\mathrm{j}} \in C}{\operatorname{argmax}} P(c_j) \prod_{i \in positions} P(w_i \mid c_j)$$

Cross-Validation

- Break up data into 10 folds
  - (Equal positive and negative inside each fold?)
- For each fold
  - Choose the fold as a temporary test set
  - Train on 9 folds, compute performance on the test fold
- Report average performance of the 10 runs



Other issues in classification

- MaxEnt and SVM tend to do better than Naïve Bayes

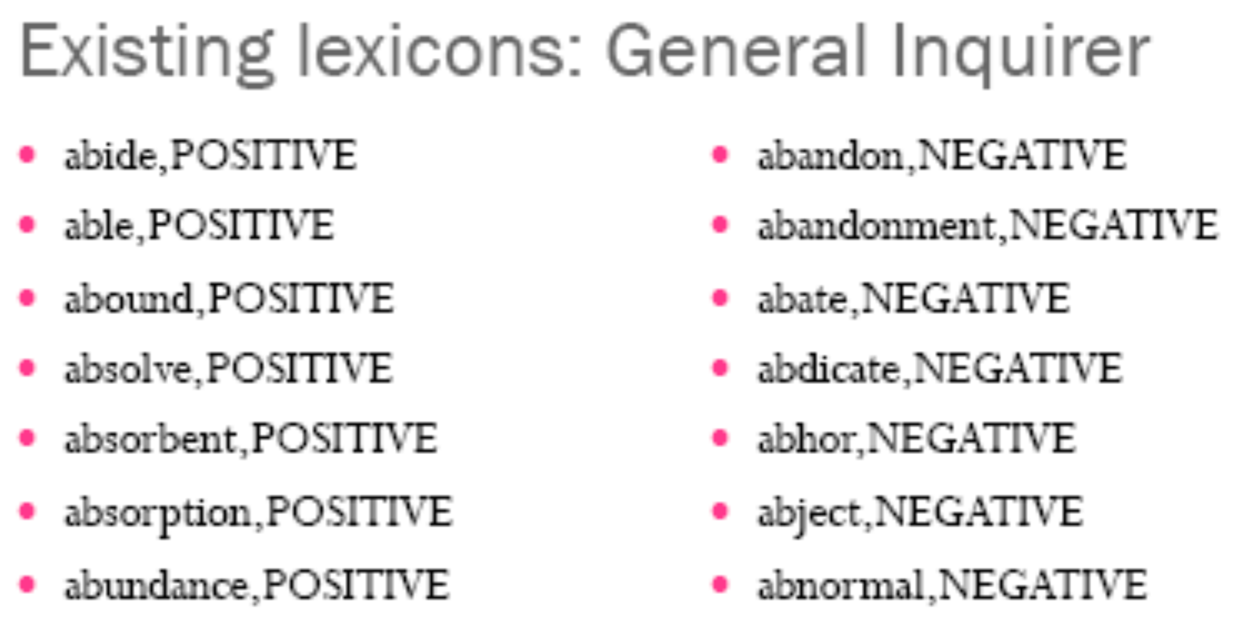
情感资源

- 情感分析的基础
- 英文资源较多
  - 情感词典：SentiWordNet, Inquirer等
    - 包含词语、短语等
    - 倾向性词语，主观性词语
  - 已标注语料库数量较多
  - 提供开源情感分析工具: OpinionFinder

情感资源

- 中文资源较少，逐年增多
  - 知网Hownet提供了部分情感词汇，部分高校也提供了情感词汇，但质量参差不齐
  - 近两年的评测提供了中文标注文本
    - NTCIR， COAE、NLP&CC等
- 情感资源基本上跟领域、语言有关
- 主客观分析与倾向性分析的资源也不一样

情感资源

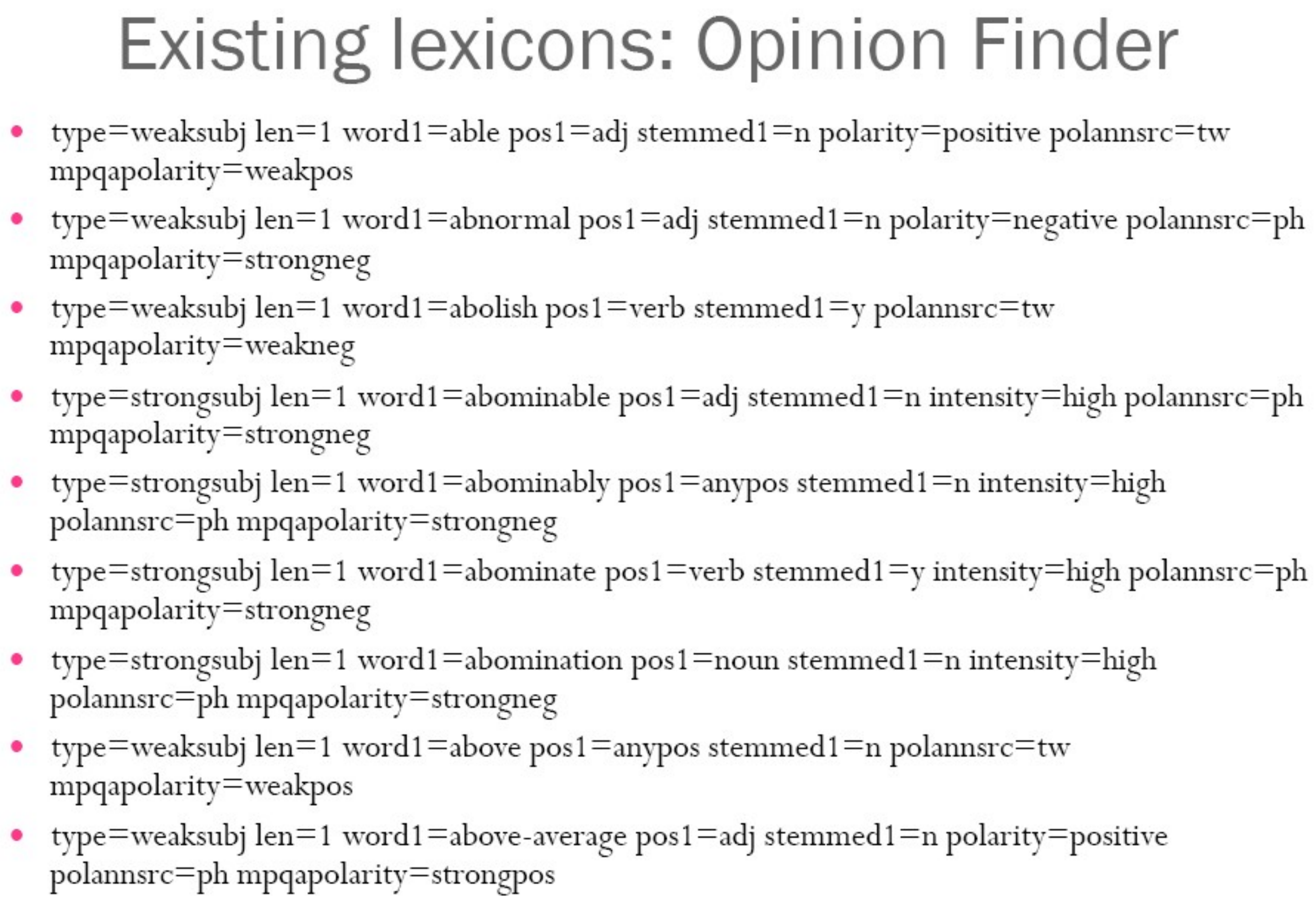


Existing lexicons: General Inquirer

- abide,POSITIVE
- able,POSITIVE
- abound,POSITIVE
- absolve,POSITIVE
- absorbent,POSITIVE
- absorption,POSITIVE
- abundance,POSITIVE

- abandon,NEGATIVE
- abandonment,NEGATIVE
- abate,NEGATIVE
- abdicate,NEGATIVE
- abhor,NEGATIVE
- abject,NEGATIVE
- abnormal,NEGATIVE

情感资源



Existing lexicons: Opinion Finder

- type=weaksubj len=1 word1=able pos1=adj stemmed1=n polarity=positive polannsrc=tw mpqapolarity=weakpos
- type=weaksubj len=1 word1=abnormal pos1=adj stemmed1=n polarity=negative polannsrc=ph mpqapolarity=strongneg
- type=weaksubj len=1 word1=abolish pos1=verb stemmed1=y polannsrc=tw mpqapolarity=weakneg
- type=strongsubj len=1 word1=abominable pos1=adj stemmed1=n intensity=high polannsrc=ph mpqapolarity=strongneg
- type=strongsubj len=1 word1=abominably pos1=anypos stemmed1=n intensity=high polannsrc=ph mpqapolarity=strongneg
- type=strongsubj len=1 word1=abominate pos1=verb stemmed1=y intensity=high polannsrc=ph mpqapolarity=strongneg
- type=strongsubj len=1 word1=abomination pos1=noun stemmed1=n intensity=high polannsrc=ph mpqapolarity=strongneg
- type=weaksubj len=1 word1=above pos1=anypos stemmed1=n polannsrc=tw mpqapolarity=weakpos
- type=weaksubj len=1 word1=above-average pos1=adj stemmed1=n polarity=positive polannsrc=ph mpqapolarity=strongpos

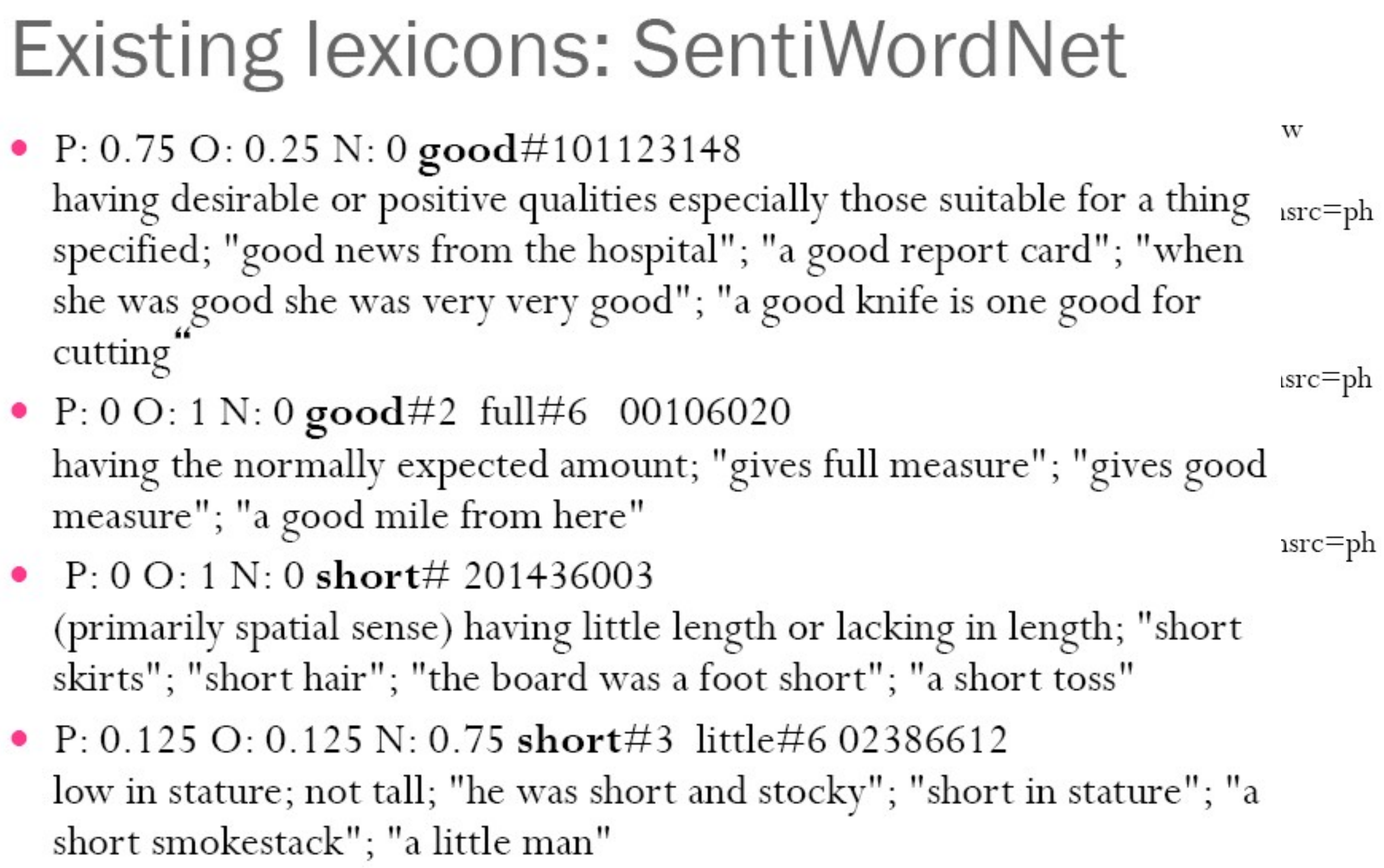
Bing Liu Opinion Lexicon

Minqing Hu and Bing Liu. Mining and Summarizing Customer Reviews. ACM SIGKDD-2004.

- Bing Liu's Page on Opinion Mining
- http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar

- 6786 words
  - 2006 positive
  - 4783 negative

情感资源



Existing lexicons: SentiWordNet

w

- P: 0.75 O: 0.25 N: 0 **good**#101123148
  having desirable or positive qualities especially those suitable for a thing specified; "good news from the hospital"; "a good report card"; "when she was good she was very very good"; "a good knife is one good for cutting“

ısrc=ph

ısrc=ph

- P: 0 O: 1 N: 0 **good**#2 full#6 00106020
  having the normally expected amount; "gives full measure"; "gives good measure"; "a good mile from here"

ısrc=ph

- P: 0 O: 1 N: 0 **short**# 201436003
  (primarily spatial sense) having little length or lacking in length; "short skirts"; "short hair"; "the board was a foot short"; "a short toss"
- P: 0.125 O: 0.125 N: 0.75 **short**#3 little#6 02386612
  low in stature; not tall; "he was short and stocky"; "short in stature"; "a short smokestack"; "a little man"

SentiWordNet

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

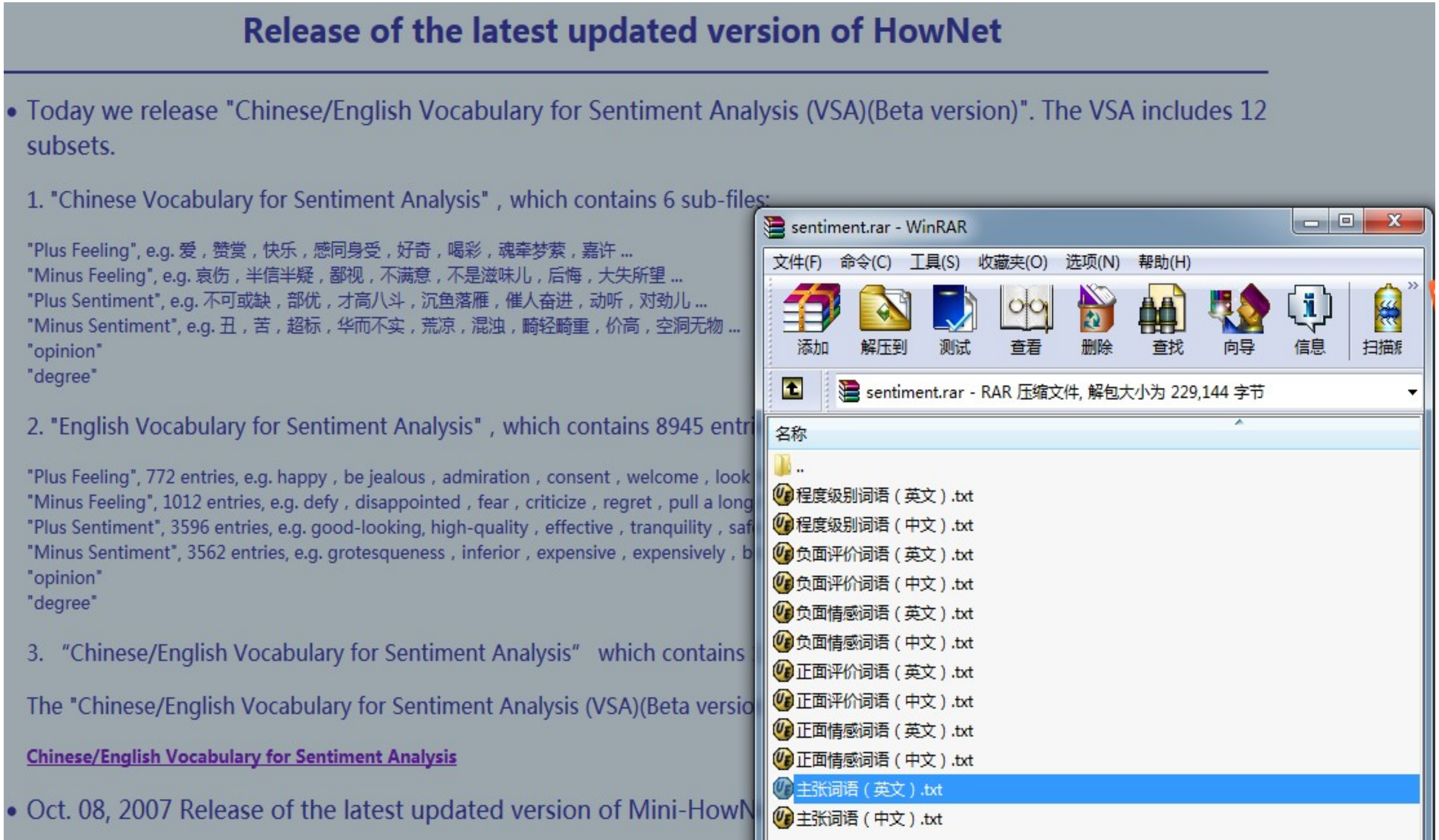
- Home page: http://sentiwordnet.isti.cnr.it/
- All WordNet synsets automatically annotated for degrees of positivity, negativity, and neutrality/objectiveness
- [estimable(J,3)] “may be computed or estimated”

```
Pos   0    Neg 0    Obj 1
```

- [estimable(J,1)] “deserving of respect or high regard”

```
Pos .75   Neg 0    Obj .25
```

知网情感词典



Release of the latest updated version of HowNet

- Today we release "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta version)". The VSA includes 12 subsets.

1. "Chinese Vocabulary for Sentiment Analysis" , which contains 6 sub-files:

"Plus Feeling", e.g. 爱，赞赏，快乐，感同身受，好奇，喝彩，魂牵梦萦，嘉许 ...
"Minus Feeling", e.g. 哀伤，半信半疑，鄙视，不满意，不是滋味儿，后悔，大失所望 ...
"Plus Sentiment", e.g. 不可或缺，部优，才高八斗，沉鱼落雁，催人奋进，动听，对劲儿 ...
"Minus Sentiment", e.g. 丑，苦，超标，华而不实，荒凉，混浊，畸轻畸重，价高，空洞无物 ...
"opinion"
"degree"

2. "English Vocabulary for Sentiment Analysis" , which contains 8945 entri

"Plus Feeling", 772 entries, e.g. happy , be jealous , admiration , consent , welcome , look
"Minus Feeling", 1012 entries, e.g. defy , disappointed , fear , criticize , regret , pull a long
"Plus Sentiment", 3596 entries, e.g. good-looking, high-quality , effective , tranquility , saf
"Minus Sentiment", 3562 entries, e.g. grotesqueness , inferior , expensive , expensively , b
"opinion"
"degree"

3. “Chinese/English Vocabulary for Sentiment Analysis” which contains

The "Chinese/English Vocabulary for Sentiment Analysis (VSA)(Beta versio

Chinese/English Vocabulary for Sentiment Analysis

- Oct. 08, 2007 Release of the latest updated version of Mini-HowN

sentiment.rar - RAR 压缩文件, 解包大小为 229,144 字节

..
程度级别词语（英文）.txt
程度级别词语（中文）.txt
负面评价词语（英文）.txt
负面评价词语（中文）.txt
负面情感词语（英文）.txt
负面情感词语（中文）.txt
正面评价词语（英文）.txt
正面评价词语（中文）.txt
正面情感词语（英文）.txt
正面情感词语（中文）.txt
主张词语（英文）.txt
主张词语（中文）.txt

情感词汇资源构建

- **基于心理学的评价理论**
- **任务**
  - **确定词语的主观性(subjectivity)**
  - **确定词语的倾向(orientation)**
  - **确定词语态度的强度(strength)**
- **例子**
  - **Objective: vertical, yellow, liquid**
  - **Subjective**
    - **Positive: good < excellent**
    - **Negative: bad < terrible**

情感词汇资源构建

- **连接词方法(Conjunction Method)**
- **PMI方法**
  - Orientation
  - Subjectivity
- **WordNet扩展方法**
- **释义方法(Gloss Use Method)**
  - Orientation
  - Subjectivity
  - SentiWordNet

How to measure polarity of a phrase?

- Positive phrases co-occur more with *“excellent”*
- Negative phrases co-occur more with *“poor”*
- But how to measure co-occurrence?

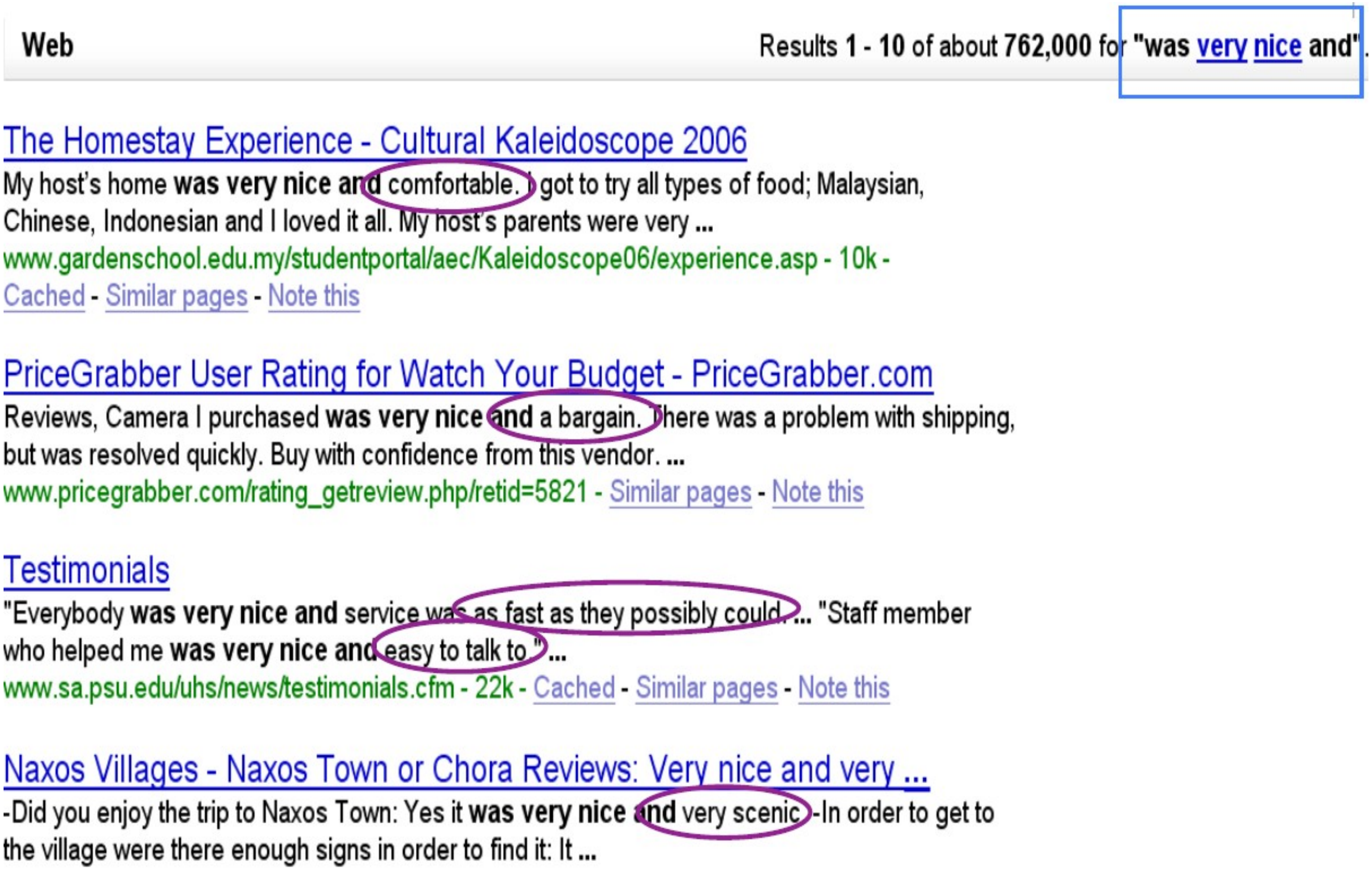
连接词方法

- **假设**
  - **用‘and’ 相连的形容词通常具有相同的倾向，而用 ‘but’ 相连的形容词通常具有相反的倾向**

“beautiful and clever”，

“beautiful but stupid”

- **对形容词按照不同倾向聚类**



Web

Results **1** - **10** of about **762,000** for **"was very nice and"**.

The Homestay Experience - Cultural Kaleidoscope 2006

My host's home **was very nice and** comfortable. I got to try all types of food; Malaysian, Chinese, Indonesian and I loved it all. My host's parents were very **...**

www.gardenschool.edu.my/studentportal/aec/Kaleidoscope06/experience.asp - 10k - Cached - Similar pages - Note this

PriceGrabber User Rating for Watch Your Budget - PriceGrabber.com

Reviews, Camera I purchased **was very nice and** a bargain. There was a problem with shipping, but was resolved quickly. Buy with confidence from this vendor. **...**

www.pricegrabber.com/rating_getreview.php/retid=5821 - Similar pages - Note this

Testimonials

"Everybody **was very nice and** service was as fast as they possibly could. **...** "Staff member who helped me **was very nice and** easy to talk to." **...**

www.sa.psu.edu/uhs/news/testimonials.cfm - 22k - Cached - Similar pages - Note this

Naxos Villages - Naxos Town or Chora Reviews: Very nice and very ...

-Did you enjoy the trip to Naxos Town: Yes it **was very nice and** very scenic -In order to get to the village were there enough signs in order to find it: It **...**

Pointwise Mutual Information

- **Mutual information** between 2 random variables X and Y

$$\mathrm{I}(X,Y) = \sum_{x}\sum_{y} P(x,y)\log_2 \frac{P(x,y)}{P(x)P(y)}$$

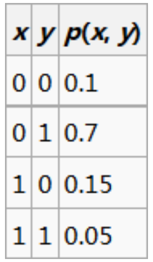
- **Pointwise mutual information**:
  - How much more do events x and y co-occur than if they were independent?

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

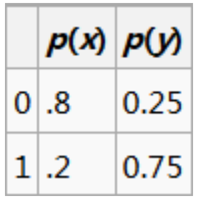
PMI方法

- **PMI: 点互信息(Pointwise Mutual Information)**

$$\operatorname{pmi}(x;y) \equiv \log\frac{p(x,y)}{p(x)p(y)} = \log\frac{p(x|y)}{p(x)} = \log\frac{p(y|x)}{p(y)}.$$



x	*y*	*p*(*x*, *y*)
0	0	0.1
0	1	0.7
1	0	0.15
1	1	0.05



	p(*x*)	*p*(*y*)
0	.8	0.25
1	.2	0.75

⇨

pmi(x=0;y=0) −1

pmi(x=0;y=1) 0.222392421

pmi(x=1;y=0) 1.584962501

pmi(x=1;y=1) −1.584962501

- **判别倾向性**
  - **具有相似倾向性的词语倾向于在文档中共同出现**
- **判别主观性**
  - **主观性形容词倾向于出现在其他主观性形容词周围**

How to Estimate Pointwise Mutual Information

- Query search engine (Altavista)
  - P(word) estimated by `hits(word)/N`
  - P(word$_1$,word$_2$) by `hits(word1 NEAR word2)/N`^{2}

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

Does phrase appear more with “poor” or “excellent”?

$$\text{Polarity}(\textit{phrase}) = \text{PMI}(\textit{phrase},\text{"excellent"}) - \text{PMI}(\textit{phrase},\text{"poor"})$$

$$= \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"excellent"})} - \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"poor"})}$$

$$= \log_2 \frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})}{\text{hits}(\textit{phrase})\text{hits}(\text{"excellent"})} \frac{\text{hits}(\textit{phrase})\text{hits}(\text{"poor"})}{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})}$$

$$= \log_2 \left(\frac{\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})\text{hits}(\text{"poor"})}{\text{hits}(\textit{phrase}\ \text{NEAR "poor"})\text{hits}(\text{"excellent"})} \right)$$

Results of Turney algorithm

- 410 reviews from Epinions
  - 170 (41%) negative
  - 240 (59%) positive
- Majority class baseline: 59%
- Turney algorithm: 74%

- Phrases rather than words
- Learns domain-specific information

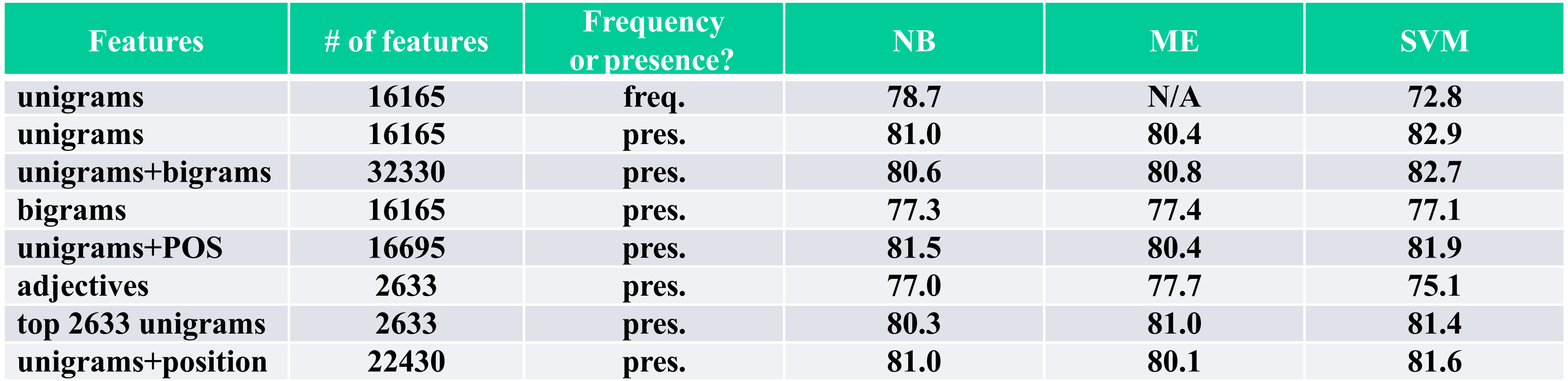
基于PMI的情感分类

- **步骤**
  - **只抽取包含形容词或副词的两个词构成的短语**
  - **短语phrase的语义倾向**
    - **SO(phrase) = PMI(phrase, “excellent”) – PMI(phrase, “poor”)**
  - **文档的语义倾向为所有短语语义倾向的平均值**
- **实验**
  - **410 reviews from Epinions (epinion.com): 170 positive, 240 negative**

Domain of review	Accuracy	Domain of review	Accuracy
Automobiles	**84.00%**	Movies	**65.83%**
- Honda Accord	83.78%	- The Matrix	66.67%
- Volkswagen Jetta	84.21%	- Pearl Harbor	65.00%
Banks	80.00%	Travel Destination	70.53%
- Bank of America	78.33%	- Cancun	64.41%
- Washington Mutual	81.67%	- Puerto Vallarta	80.56%

基于分类学习的情感分类

- **看作是特殊的文本分类任务**
- **文档采用标准的特征向量表示**
  - **特征包括：unigram, bigram, POS, sentiment lexicon, etc.**
- **实验**
  - **Data：movie reviews (Internet Movie Database), rating -> negative, neutral, positive**
  - **Naïve Bayes, Maximum Entropy, Support Vector Machine**



Features	# of features	Frequency or presence?	NB	ME	SVM
unigrams	16165	freq.	78.7	N/A	72.8
unigrams	16165	pres.	81.0	80.4	82.9
unigrams+bigrams	32330	pres.	80.6	80.8	82.7
bigrams	16165	pres.	77.3	77.4	77.1
unigrams+POS	16695	pres.	81.5	80.4	81.9
adjectives	2633	pres.	77.0	77.7	75.1
top 2633 unigrams	2633	pres.	80.3	81.0	81.4
unigrams+position	22430	pres.	81.0	80.1	81.6

情感分类现状与难点

- 产品评论的情感分类效果较好，可实用
- 社交媒体情感分析相当困难
  - 写作自由，不规范
    - @JeremyOnMarz just be tryna hate on ma‘ 2nd babydaddy Kirko Bangz. （negative ）
  - 反讽
    - 余教授真是为北大争光！
  - 情感倾向需要跟对象关联才有意义
    - 我反对站中 Vs. 我支持警察清场

观点抽取

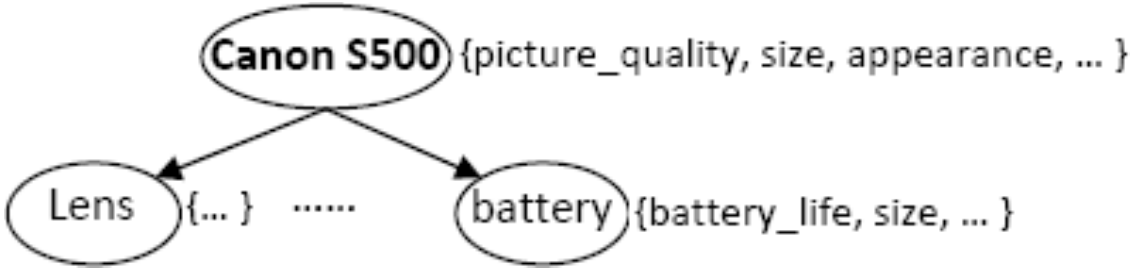
观点抽取

- **观点的组成**
  - **观点持有者(Opinion holder): 持有/表达观点的人或机构**
  - **目标对象(Object): 观点的表达对象，所指向的对象**
  - **观点表达(Opinion): 持有者对目标对象的态度、评价**

- **一般针对产品评论进行分析**

目标对象

- **一个对象o 是一个产品、人物、事件、机构、或话题，可表示为**
  - **一个部件/子部件构成的层次结构**
  - **每个部件都有一系列属性(attributes)**


Canon S500
{picture_quality, size, appearance, ... }
Lens
{... }
......
battery
{battery_life, size, ... }

- **观点可表达于任一节点或属性**
- **部件或属性也称为特征(aspect，features)**

观点表示

- **一个观点表示为五元组**

$$(o_j, a_{jk}, so_{ijkl}, h_i, t_l),$$

其中

- o_j 为目标对象.
- a_{jk} 是对象o_j的特征
- so_{ijkl} 为观点所表达的情感值（如倾向性分类）
- h_i为观点持有者
- t_l 为观点表达的时间

观点抽取的目标

- **给定观点文本**
  - **抽取所有的五元组(o_j, a_k, so_{ijkl}, h_i, t_l)**
- **基于五元组，可将无结构化文本结构化**
  - **可利用传统数据挖掘与可视化技术进行挖掘与呈现**
  - **可以定量与定性分析**

观点抽取任务很困难

- ***“This past Saturday, I bought a Nokia phone and my girlfriend bought a Motorola phone with Bluetooth. We called each other when we got home. The voice on my phone was not so clear, worse than my previous phone. The battery life was long. My girlfriend was quite happy with her phone. I wanted a phone with good sound quality. So my purchase was a real disappointment. I returned the phone yesterday.”***

观点抽取子任务

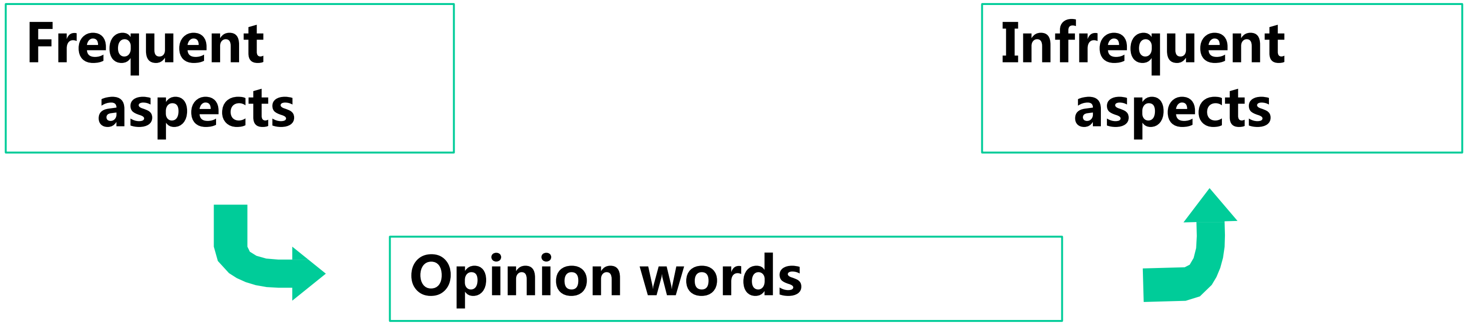
- **重点关注两个子任务**
  - **特征抽取与聚类 (aspect extraction and grouping)**
    - **抽取对象的所有特征表达，并将同义特征表达聚类。每个特征**
    - **类表示了关于该对象的独一无二的某个特征**
  - **特征情感分类(aspect sentiment classification)**
    - **确定观点针对每个特征的情感倾向：正面、负面、中性.**

对象特征抽取

- **频繁特征: 被许多评论提及的特征**
  - **找到评论中频繁出现的名词短语**
    - **描述产品特征的词语比较有限**
    - **主要特征通常出现比较频繁.**
    - **手机： 大小、电池、价格、屏幕像素、内存, 等等**

非频繁特征抽取

- **基于: 同一情感词被用来描述不同特征与对象**
  - **“The pictures are absolutely amazing.”**
  - **“The software that comes with it is amazing.”**



Frequent aspects

Opinion words

Infrequent aspects

特征聚类

- **基于相似度对特征词进行聚类**
  - **将同一特征的不同表达聚到一起**
    - **例如：价格、价钱、售价**
  - **可考虑字符串相似度，同义词、其他基于WordNet的距离**

特征情感分类

- 对于每个特征，判别观点持有者所表达的情感倾向性

- 基于句子
  - 一个句子可包含多个特征
  - 针对不同的特征可以有不同的观点
  - E.g., The battery life and picture quality are great (+), but the view founder is small (-).

特征情感分类

- **输入: (f, s)，f 为产品特征，s 为包含f的一个句子**
- **输出: s中针对 f 的观点倾向**
- **两个步骤:**
  - **基于转折连词切分句子(but, except that, etc).**
  - **针对包含f的句子分段s_f ，将其所有情感词的倾向性值(1, -1, 0)进行求和.**
- **对于上下文相关的特征情感分类，需要充分利用句子之间和句子内部的连词、否定词等词汇，加以约束**
  - **Negation Neg → Positive**
  - **Negation Pos → Negative**
  - **Desired value range → Positive**
  - **Below or above the desired value range → Negative ; …**

基于特征的产品观点摘要

- *“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. …”*

- Feature Based Summary:
- Feature1: Touch screen
- Positive: 212
  - *The touch screen was really cool.*
  - *The touch screen was so easy to use and can do amazing things.*
- …
- Negative: 6
  - The screen is easily scratched.
  - I have a lot of difficulty in removing finger marks from the touch screen.
- …
- Feature2: battery life
- …

- *Note: We omit opinion holders*

其他相关研究工作

- **多模态情感分析**
  - **图像**
  - **音乐**
  - **语音**
  - **视频**
- **机器人情感合成 (视频)**

总结

- 语言建模
  - 概率语言模型
  - 马尔科夫链假设
  - Uni-gram 到 bi-gram
  - 语言建模工具
- 情感分析
  - 简介
  - 框架
  - 情感分类：朴素贝叶斯模型
  - 情感词构建：连接词与PMI
  - 观点抽取

下周：文档摘要 ：D