Data Mining and Analysis

Text Mining: 1

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Resources

https://web.stanford.edu/~jurafsky/slp3/ — slides and readings on Sentiment Analysis

http://www.cs.virginia.edu/~hw5x/Course/TextMining-2019Spring/_site/lectures/ — slides from Hongning Wang's course on Text Mining

http://www.cs.columbia.edu/~blei/topicmodeling.html

Opinion mining and sentiment analysis by Pang and Lee

Probabilistic Topic Models (CACM) by Blei

Latent dirichlet allocation by Blei, Ng, and Jordan

Probabilistic latent semantic analysis by Thomas Hofmann

Text Mining: Wikipedia

Text mining, also referred to as text data mining, roughly equivalent to text analytics, is

> the process of deriving high-quality information from text.

Text Analytics

Zhao and Massung. Text Data Management and Analysis (2016)

Summarization

textsummarization.net/text-summarizer

TextSummarization Text Summarizer Online Text Summarization API

Text Summarization Result

Original URL/Text

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal. Now we are engaged in a great civil war, testing whether that nation, or any nation so conceived and so dedicated, can long endure. We are met on a great battle-field of that war. We have come to dedicate a portion of that field, as a final resting place for those who here gave their lives that that nation might live. It is altogether fitting and proper that we should do this. But, in a larger sense, we can not dedicate -- we can not consecrate -- we can not hallow -- this ground. The brave men, living and dead, who struggled here, have consecrated it, far above our poor power to add or detract. The world will little note, nor long remember what we say here, but it can never forget what they did here. It is for us the living, rather, to be dedicated here to the unfinished work which they who fought here have thus far so nobly advanced. It is rather for us to be here dedicated to the great task remaining before us -- that from these honored dead we take increased devotion to that cause for which they gave the last full measure of devotion -- that we here highly resolve that these dead shall not have died in vain -- that this nation, under God, shall have a new birth of freedom -- and that government of the people, by the people, for the people, shall not perish from the earth.

Summarized Text

Four score and seven years ago our fathers brought forth on this continent, a new nation, conceived in Liberty, and dedicated to the proposition that all men are created equal.

We are met on a great battle-field of that war.

Text Categorization

Google Corporation®
123 Buckingham Palace Road
London SW1W 9SH
United Kingdom

Good day Sir/Madam.

You have successfully been picked as one of our 12 Lucky Winners in this months Lottery Draw, Please see attached file for more details.

Best regards
Sundar Pichai
CEO Google Inc.
sundarpichai@gpromo-team.com

Document Clustering

Suppose you're an analyst for:

GoldmanSachs looking for major shifts in a company (e.g., Theranos)

the NSA looking for evidence of a terror cell operating in a particular country

… ?

Information Extraction

Google

what year was texas founded

All | News | Images | Shopping | Maps | More | Search tools

About 7,490,000 results (0.69 seconds)

Texas / Founded

December 29, 1845

Feedback

People also ask

- What is the history of Texas?
- When was Texas first settled?
- What is the state of Texas?
- Why is the capital of Texas Austin?

Texas - Wikipedia, the free encyclopedia

https://en.wikipedia.org/wiki/**Texas** ▾ Wikipedia ▾

Topic Modeling

AmazonBasics Pre-sharpened Wood Cased #2 HB Pencils, 30 Pack

by AmazonBasics

1,100 customer reviews

| 57 answered questions

Amazon's Choice for "2 pencils"

Price: $4.99 ✓prime

& FREE Returns

Thank you for being a Prime member. Get $70 off instantly: Pay $0.00 upon approval for the Amazon Prime Rewards Visa Card. No annual fee.

Size: **30 Pack**

30 Pack | 150 Pack

1,100 customer reviews

4.5 out of 5 stars

5 star	75%
4 star	9%
3 star	5%
2 star	3%
1 star	8%

By feature

Easy to hold	4.8
Value for money	4.6
Durability	4.5
Softness	4.5
Longevity	4.3

See less

Review this product

Share your thoughts with other customers

Write a customer review

Customer images

See all customer images

Read reviews that mention

already sharpened | presharpened | sharpen well | back to school | good quality | pencil sharpener | waste of money | lead breaks | electric sharpener | pre-sharpened pencils | school supplies | great price

Showing 1-8 of 1,100 reviews

Top Reviews

Elisabeth B

Decent pencil unless you're a pencil snob

April 14, 2018

Size: 150 Pack | Verified Purchase

Great price for order decent pencils although I'm a pencil snob. I purchased these for my children

Text Visualization

people
care
will
health
system
insurance
can
American
doctors
know
want
year
medical
money
cost
small
every
us
Americans
pay
time
savings
costs
tonight
new
get
think
change
quality
coverage
plan
children
must
business
government
employees
today
businesses
doctor
plans
sick
better
also
one
way
achieve
security
just
responsibility
country
believe
work
now
reform
inflation
Congress
ask
employer
nation
never
told
together
large
provide
years
first
principle
right
still
simply
economy
already
hospital
continue
proposal
thousands
spend
third
may
Now
even
paperwork
taken
medicine
benefits
tell
ever
services
information
family
disagree
away
jobs
afford
everybody
elderly
face
higher
able
create
covered
much
look
many
without
workers
premiums
small-business
say
competition
world
give
comprehensive
major
keep
let
cover
arguments
means
else
less
thing
make
budget
companies
understand
hospitals
public
important
someone
America
company
principles
long
got
agree
fast
go
journey
whether
percent
choice
best
someone
understand
hospitals
never
employer
nation
told
together
large
provide
years
first
doctors
American
want
know
year
medical
money
cost
small
every
us
Americans
pay
time
savings
costs
tonight
new
get
think
change
quality
coverage
plan
children
must
business
government
employees
today
businesses
doctor
plans
sick
better
also
one
way
achieve
security
just
responsibility
country
believe
work
now
reform
inflation
Congress
ask
employer
nation
never

Word Tree

Shift-click to make that word the root.

good fun that is funny!” “I know some good games we could play,” Said the cat. “I know some new tricks,” Said the Cat in the Hat. “A lot of good tricks. I will show them to you. Your mother Will not mind at all if I do.” Then Sally and I Did not know what to say. Our mother was out of the house For the day. But the fish said, “No! No! Make that cat go away! Tell that Cat in the Hat you do NOT want to play. He should not be here. He should not be about. He should not be here When your mother is out!” “Now! Now! Have no fear. Have no fear!” said the cat. “My tricks are not bad,” Said the Cat in the Hat. “Why, we can have lots of good fun, if you wish, With a game that I call UP UP UP with a fish!” “Put me down!” said the fish. “This is no fun at all! Put me down!” said the fish. “I do NOT wish to fall!” “Have no fear!” said the cat. “I will not let you

https://www.jasondavies.com/wordtree/

Phrase Nets

collins
charlotte
miss
lady
proud
play
sing
care
anxiety
attentive
kind
grave
anxious
william
maria
mrs
bingley
hurst
bennet
elizabeth
jane
catherine
mr
father
friend
brother
lydia
sisters
friends
mary
sister
kitty
mother
aunt
uncle
air
manner
countenance
address
fortune
consequence
ease
good
sense
disposition
sweetness
humour
kindness
common
ignorance
assurance
understanding
folly
temper
conceit
pride
health
happiness
vanity
follies
affection
wishes
solicitude
vexation
regret
shame
grabitude
esteem
smiles
bears
general
astonished
vexed
amiable
pleasing
character
books
make

TreeMap

https://newsmap-js.herokuapp.com/

Text Over Time

https://books.google.com/ngrams

http://projects.fivethirtyeight.com/reddit-ngram/

Sentiment Analysis / Opinion Mining

Opinions

I'm a tech CEO, and I don't think tech CEOs should be making policy

A person walks past a Google sign in San Francisco on May 1. (Jeff Chiu/AP)

Dan Waylonis 1 day ago

Very interesting take. Tech companies need to lead and. It just react to the bleatings if disgruntled employees.

Like Reply Link Report

Ling Ling Ling 1 day ago *(Edited)*

Disgruntled employees move to work for competitors or launch new companies and become competitors.

Like Reply Link Report

SaneBrain 1 day ago

How did this tech exec escape from the mental programming imposed on the valley? I mean good for him, but he's pretty alone.

Like Reply Link Report

pjhinton 1 day ago

?

Dashboards

Product Reviews
Sentiments Analysis
FACTION A
Date
3/21/2011
2/6/2018
Products
All
Rating
All
Sentiments
(Blank)
anticipation
fear
negative
sadness
trust
anger
disgust
joy
positive
surprise
KPIs
Rating
4.3
Rated Products
8
Reviews
511
Average Price
$150.73
Word Frequency by Sentiment
positive 407
trust 305
anticipation 264
joy 259
negative 220
sadness 139
anger 115
fear 107
surprise 90
disgust 78
Reviews Classification
negative 28%
positive 72%
Reviews by Year/Month/day
Sentiment negative positive
500
400
300
200
100
0
2012
2014
2016
2018
Total Score by Word
perfect 660
great 435
top 414
happy 372
love 372
nice 288
excellent 198
easy 165
beautiful 138
0K
1K
Rating vs Product Score
Rating (*)
4.5
4.0
3.5
1.0
1.5
Score
Word Cloud
together
great
easy
looks
back
nice
cabinet
assemble
piece
Appreciation
Sentiments
Reviews
FILTERS

Contrast with lots of our experience with data so far this semester

Text is high-dimensional

Words have meanings and relations

- Correlations: College Station, BCS, Austin
- Order: Monday, Tuesday, …
- Membership: Basketball, football, sports
- Hierarchy, antonyms, synonyms, entities, …

Text Processing Pipeline

Tokenization

- how to separate? what to separate?
- unigrams, bigrams
- stopwords
- numbers, symbols

(Sophisticated?) Linguistic analysis

- stemming
- part-of-speech tagging

Bag-of-words model

- Documents as vectors
- How to weight? By count, by TFIDF?

Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adaped from Schultz [44] page 120)

Variants of term frequency (tf) weight

weighting scheme	tf weight
binary	$0, 1$
raw count	$f_{t,d}$
term frequency	$f_{t,d} \Big/ \sum_{t' \in d} f_{t',d}$
log normalization	$\log(1 + f_{t,d})$
double normalization 0.5	$0.5 + 0.5 \cdot \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$
double normalization K	$K + (1 - K) \frac{f_{t,d}}{\max_{\{t' \in d\}} f_{t',d}}$

Variants of inverse document frequency (idf) weight

weighting scheme	idf weight ($n_t = \lvert\{d \in D : t \in d\}\rvert$)
unary	1
inverse document frequency	$\log \frac{N}{n_t} = -\log \frac{n_t}{N}$
inverse document frequency smooth	$\log\left(\frac{N}{1+n_t}\right)$
inverse document frequency max	$\log\left(\frac{\max_{\{t' \in d\}} n_{t'}}{1+n_t}\right)$
probabilistic inverse document frequency	$\log \frac{N-n_t}{n_t}$

https://scikit-learn.org/stable/tutorial/text_analytics/working_with_text_data.html

scikit-learn

Working With Text Data

The goal of this guide is to explore some of the main `scikit-learn` tools on a single practical task: analyzing a collection of text documents (newsgroups posts) on twenty different topics.

In this section we will see how to:

- load the file contents and the categories
- extract feature vectors suitable for machine learning
- train a linear model to perform categorization
- use a grid search strategy to find a good configuration of both the feature extraction components and the classifier

Tutorial setup

To get started with this tutorial, you must first install *scikit-learn* and all of its required dependencies.

Please refer to the installation instructions page for more information and for system-specific instructions.

The source of this tutorial can be found within your scikit-learn folder:

https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html

scikit-learn :: CountVectorizer

sklearn.feature_extraction.text.CountVectorizer

class `sklearn.feature_extraction.text.` **`CountVectorizer`** (*input='content', encoding='utf-8', decode_error='strict', strip_accents=None, lowercase=True, preprocessor=None, tokenizer=None, stop_words=None, token_pattern='(?u)\b\w\w+\b', ngram_range=(1, 1), analyzer='word', max_df=1.0, min_df=1, max_features=None, vocabulary=None, binary=False, dtype=<class 'numpy.int64'>*) [source]

Convert a collection of text documents to a matrix of token counts

This implementation produces a sparse representation of the counts using scipy.sparse.csr_matrix.

If you do not provide an a-priori dictionary and you do not use an analyzer that does some kind of feature selection then the number of features will be equal to the vocabulary size found by analyzing the data.

Read more in the User Guide.

Parameters: **input :** ***string {'filename', 'file', 'content'}***

If 'filename', the sequence passed as an argument to fit is expected to be a list of filenames that need reading to fetch the raw content to analyze.

If 'file', the sequence items must have a 'read' method (file-like object) that is called to fetch the bytes in memory.

Otherwise the input is expected to be the sequence strings or bytes items are expected to be analyzed directly.

Sentiment Analysis and Opinion Mining

ONCE UPON A TIME IN HOLLYWOOD

Critics Consensus

Thrillingly unrestrained yet solidly crafted, *Once Upon a Time in Hollywood* tempers Tarantino's provocative impulses with the clarity of a mature filmmaker's vision.

CERTIFIED FRESH Rotten Tomatoes **85%**

TOMATOMETER
Total Count: 492

70%

AUDIENCE SCORE
Verified Ratings: 25,234

NEW MORE INFO

There seems little doubt that every detail of the tableau has been selected to give [Tarantino] pleasure, and we are being permitted to witness a private ritual of self-gratification, on more than one level.

Aug 14, 2019 | Rating: 4/5 | Full Review...

Jake Wilson
The Age (Australia)
★ Top Critic

The plot is a tapestry template that drops in characters like shining beads and watches the story weave slowly around them.

Aug 8, 2019 | Rating: 5/5 | Full Review...

Kevin Maher
Times (UK)
★ Top Critic

A bittersweet, complex, conversation-starting look back at the film business at the close of a tumultuous decade, Once Upon a Time in Hollywood also works as a strong argument for why it's too soon for Tarantino to pack it in.

Aug 5, 2019 | Full Review...

Keith Phipps
Slate
★ Top Critic

Don't mistake me: This movie is good. It all depends on how hard you're willing to work to justify its pleasures.

Aug 2, 2019 | Full Review...

Josephine Livingstone
The New Republic
★ Top Critic

It feels like Tarantino is so wrapped up in indulging in his own filmmaking style that he loses sight of what made his best movies work.

Jul 30, 2019 | Rating: 2.5/4 | Full Review...

Matthew Rozsa

It's dispiriting to suspect that part of why Tarantino wanted to stage a Manson-adjacent story was because the accoutrements-the period cars and costumes and neon signs, the glowering barefoot hippie girls, the acid-laced cigarettes-were just so cool.

From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

Twitter mood predicts the stock market.

Johan Bollen[1,⋆],Huina Mao[1,⋆],Xiao-Jun Zeng[2].

⋆: authors made equal contributions.

3003v1 [cs.CE] 14 Oct 2010

***Abstract*—Behavioral economics tells us that emotions can profoundly affect individual behavior and decision-making. Does this also apply to societies at large, i.e. can societies experience mood states that affect their collective decision making? By extension is the public mood correlated or even predictive of economic indicators? Here we investigate whether measurements of collective mood states derived from large-scale Twitter feeds are correlated to the value of the Dow Jones Industrial Average (DJIA) over time. We analyze the text content of daily Twitter feeds by two mood tracking tools, namely OpinionFinder that measures positive vs. negative mood and Google-Profile of Mood States (GPOMS) that measures mood in terms of 6 dimensions (Calm, Alert, Sure, Vital, Kind, and Happy). We cross-validate the resulting mood time series by comparing their ability to detect the public's response to the presidential election and Thanksgiving day in 2008. A Granger causality analysis and a Self-Organizing Fuzzy Neural Network are then used to investigate the hypothesis that public mood states, as measured by the OpinionFinder and GPOMS mood time series, are predictive of changes in DJIA closing values. Our results indicate that the accuracy of DJIA predictions can be significantly improved by the inclusion of specific public mood dimensions but not others. We find an accuracy of 87.6% in predicting the daily up and down changes in the closing values of the DJIA and a reduction of the Mean Average Percentage Error by more than 6%.**

***Index Terms*—stock market prediction — twitter — mood analysis.**

sentiment from blogs. In addition, Google search queries have been shown to provide early indicators of disease infection rates and consumer spending [14]. [9] investigates the relations between breaking financial news and stock price changes. Most recently [13] provide a ground-breaking demonstration of how public sentiment related to movies, as expressed on Twitter, can actually predict box office receipts.

Although news most certainly influences stock market prices, public mood states or sentiment may play an equally important role. We know from psychological research that emotions, in addition to information, play an significant role in human decision-making [16], [18], [39]. Behavioral finance has provided further proof that financial decisions are significantly driven by emotion and mood [19]. It is therefore reasonable to assume that the public mood and sentiment can drive stock market values as much as news. This is supported by recent research by [10] who extract an indicator of public anxiety from LiveJournal posts and investigate whether its variations can predict S&P500 values.

However, if it is our goal to study how public mood influences the stock markets, we need reliable, scalable and early assessments of the public mood at a time-scale and resolution appropriate for practical stock market prediction. Large surveys of public mood over representative samples of

Fig. 3. A panel of three graphs. The top graph shows the overlap of the day-to-day difference of DJIA values (blue: $\mathbb{Z}_{D_t}$) with the GPOMS' Calm time series (red: $\mathbb{Z}_{X_t}$) that has been lagged by 3 days. Where the two graphs overlap the Calm time series predict changes in the DJIA closing values that occur 3 days later. Areas of significant congruence are marked by gray areas. The middle and bottom graphs show the separate DJIA and GPOMS' Calm time series.

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

Simple? Just look for “great” or “terrible”

This laptop is a great deal.

A great deal of media attention surrounded the release of the new laptop.

This laptop is a great deal … and I’ve got a nice bridge you might be interested in.

Simple?

This film should be brilliant. It sounds like a great plot, the actors are first grade, and the supporting case is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up.

Simple?

“She ran the gamut of emotions from A to B.” [Dorothy Parker on Katharine Hepburn]

“Read the book.” [Bob Bland]

Sentiment Analysis is the Detection of Attitudes

"enduring, affectively colored beliefs, dispositions towards objects or persons"

1. Holder (source) of attitude

2. Target (aspect) of attitude

3. Type of attitude (like love, hate, value, desire, etc.) or simple weighted polarity (positive, negative, neutral, together with strength)

4. Text containing the attitude (sentence or entire document)

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

Predicting the Semantic Orientation of Adjectives by Hatzivassiloglou & McKeown (1997)

- Adjectives conjoined by “and” have same polarity
  - Fair and legitimate, corrupt and brutal
  - *fair and brutal, *corrupt and legitimate
- Adjectives conjoined by “but” do not
  - fair but brutal

Predicting the Semantic Orientation of Adjectives by Hatzivassiloglou & McKeown (1997)

Build training set: label all adjectives with frequency > 20; test agreement with human annotators

Extract all conjoined adjectives

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 **...**

nice and comfortable

nice and scenic

Predicting the Semantic Orientation of Adjectives by Hatzivassiloglou & McKeown (1997)

Build a graph of adjectives linked by the same or different semantic orientation

Predicting the Semantic Orientation of Adjectives by Hatzivassiloglou & McKeown (1997)

Partition the adjectives into two subsets

Predicting the Semantic Orientation of Adjectives by Hatzivassiloglou & McKeown (1997)

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…

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews by Turney (2002)

Goal: classify reviews as “positive” or “negative”.

Epinions “[not] recommended” as given by authors.

Method:

- Find (possibly) meaningful phrases from review (e.g., “bright display”, “inspiring lecture”, …)
- (based on POS patterns, like ADJ NOUN)
- Estimate “semantic orientation” of each candidate phrase
- Assign overall orientation of review by averaging orientation of the phrases in the review

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews by Turney (2002)

Semantic Orientation? Using the pointwise-mutual information

$$\mathrm{PMI}(word_1, word_2) = \log_2 \left[\frac{\mathrm{p}(word_1 \ \& \ word_2)}{\mathrm{p}(word_1)\ \mathrm{p}(word_2)} \right]$$

$$\mathrm{SO}(phrase) = \mathrm{PMI}(phrase, \text{“excellent”}) - \mathrm{PMI}(phrase, \text{“poor”})$$

Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews by Turney (2002)

Pointwise mutual information:

How much more do events x and y co-occur than if they were independent?

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

https://en.wikipedia.org/wiki/Pointwise_mutual_information

PMI between two words:

How much more do two words co-occur than if they were independent?

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

How to estimate PMI?

Query search engine (Altavista)

P(word) estimated by hits(word)/N

P(word1,word2) by hits(word1 NEAR word2)/N

(More correctly the bigram denominator should be kN, because there are a total of N consecutive bigrams (word1,word2), but kN bigrams that are k words apart, but we just use N on the rest of this slide and the next.)

$$\mathrm{PMI}(word_1, word_2) = \log_2 \frac{\frac{1}{N} hits(word_1 \text{ NEAR } word_2)}{\frac{1}{N} hits(word_1) \frac{1}{N} hits(word_2)}$$

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

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

$$= \log_2 \frac{\frac{1}{N}\text{hits}(\textit{phrase}\ \text{NEAR "excellent"})}{\frac{1}{N}\text{hits}(\textit{phrase})\frac{1}{N}\text{hits}(\text{"excellent"})} - \log_2 \frac{\frac{1}{N}\text{hits}(\textit{phrase}\ \text{NEAR "poor"})}{\frac{1}{N}\text{hits}(\textit{phrase})\frac{1}{N}\text{hits}(\text{"poor"})}$$

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

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

Table 2. An example of the processing of a review that the author has classified as *recommended*.[6]

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
online experience	JJ NN	2.253
low fees	JJ NNS	0.333
local branch	JJ NN	0.421
small part	JJ NN	0.053
online service	JJ NN	2.780
printable version	JJ NN	-0.705
direct deposit	JJ NN	1.288
well other	RB JJ	0.237
inconveniently located	RB VBN	-1.541
other bank	JJ NN	-0.850
true service	JJ NN	-0.732
Average Semantic Orientation		0.322

Table 3. An example of the processing of a review that the author has classified as *not recommended*.

Extracted Phrase	Part-of-Speech Tags	Semantic Orientation
little difference	JJ NN	-1.615
clever tricks	JJ NNS	-0.040
programs such	NNS JJ	0.117
possible moment	JJ NN	-0.668
unethical practices	JJ NNS	-8.484
low funds	JJ NNS	-6.843
old man	JJ NN	-2.566
other problems	JJ NNS	-2.748
probably wondering	RB VBG	-1.830
virtual monopoly	JJ NN	-2.050
other bank	JJ NN	-0.850
extra day	JJ NN	-0.286
direct deposits	JJ NNS	5.771
online web	JJ NN	1.936
cool thing	JJ NN	0.395
very handy	RB JJ	1.349
lesser evil	RBR JJ	-2.288
Average Semantic Orientation		-1.218

Thumbs up? Sentiment Classification using Machine Learning Techniques by Pang, Lee, and Vaithyanathan (2002)

Movie review classification as positive or negative

✓

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 .

Thumbs up? Sentiment Classification using Machine Learning Techniques by Pang, Lee, and Vaithyanathan (2002)

Method one: count human-provided polar words (sort of like Turney):

E.g., “love, wonderful, best, great, superb, still, beautiful” vs “bad, worst, stupid, waste, boring, ?, !” gives 69% accuracy on 700+/700- movie reviews

Method two: text classification

E.g., Naïve Bayes bag of words: 78.7%; SVM-lite “set of words”: 82.9% was best result

Adding bigrams and/or POS tags doesn’t change things much.

	Features	# of features	frequency or presence?	NB	ME	SVM
(1)	unigrams	16165	freq.	**78.7**	N/A	72.8
(2)	unigrams	"	pres.	81.0	80.4	**82.9**
(3)	unigrams+bigrams	32330	pres.	80.6	80.8	**82.7**
(4)	bigrams	16165	pres.	77.3	**77.4**	77.1
(5)	unigrams+POS	16695	pres.	81.5	80.4	**81.9**
(6)	adjectives	2633	pres.	77.0	**77.7**	75.1
(7)	top 2633 unigrams	2633	pres.	80.3	81.0	**81.4**
(8)	unigrams+position	22430	pres.	81.0	80.1	**81.6**

Mining Opinion Features in Customer Reviews, by Hu and Liu (2004)

Here: explicit product features only, expressed as nouns or compound nouns

Use association rule mining technique rather than symbolic or statistical approach to terminology

Extract associated items (item-sets) based on support (>1%)

Yahoo! for Amazon: Extracting market sentiment from stock message boards, Das and Chen (2001)

The sentence "It is not a bullish market" actually means the opposite of a bull market. Words such as "not", "never", "no", etc., serve to reverse meaning.

We handle negation by detecting these words and then tagging the rest of the words in the sentence after the negation word with markers, so as to reverse inference.

Yahoo! for Amazon: Extracting market sentiment from stock message boards, Das and Chen (2001)

```
didn’t like this movie ,
but I

—>

didn’t NOT_like NOT_this
NOT_movie but I
```

LIWC (Linguistic Inquiry and Word Count)

Pennebaker, J.W., Booth, R.J., & Francis, M.E. (2007). Linguistic Inquiry and Word Count: LIWC 2007. Austin, TX

Home page: http://www.liwc.net/

2300 words, >70 classes

Affective Processes

- negative emotion (bad, weird, hate, problem, tough)
- positive emotion (love, nice, sweet)

Cognitive Processes

- Tentative (maybe, perhaps, guess), Inhibition (block, constraint)

Pronouns, Negation (no, never), Quantifiers (few, many)

$30 or $90 fee

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

Nice Survey of Lexicons

http://sentiment.christopherpotts.net/lexicons.html

Disagreement levels for the sentiment lexicons reviewed above.

	MPQA	Opinion Lexicon	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%)
Inquirer			–	520/2306 (23%)	1/204 (0.5%)
SentiWordNet				–	174/694 (25%)
LIWC					–