

CIS 530
Computational Linguistics
MW noon-1:30
Towne 217

<http://computational-linguistics-class.org>

Professor Callison-Burch

Professor Callison-Burch

(not Professor Burch)

Bachelors from Stanford

PhD from University of Edinburgh

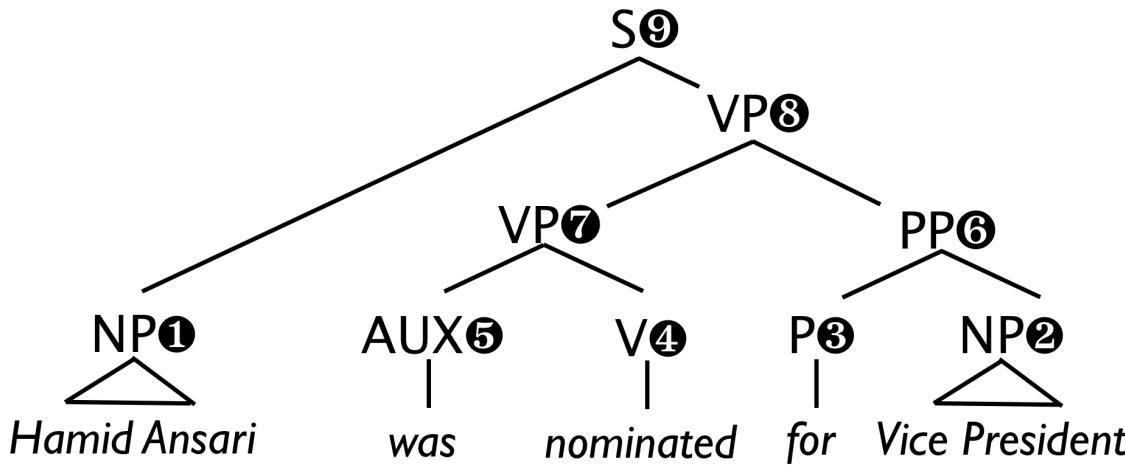
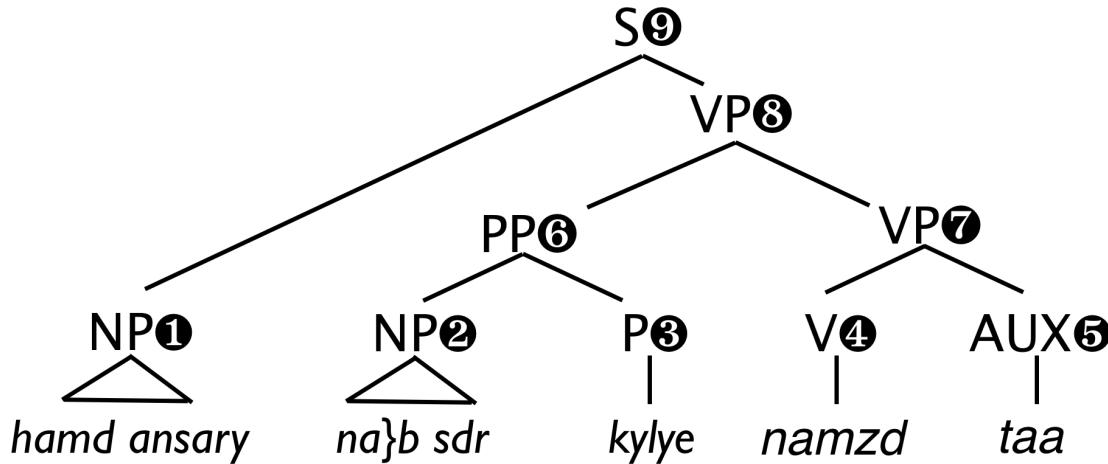
6 years at Johns Hopkins University

5th year here at Penn

I have been working in the field of NLP since 2000. Last year I was the general chair of the 55th meeting of the ACL.



	Urdu	English
$S \rightarrow$	$NP\textcircled{1} \ VP\textcircled{2}$	$NP\textcircled{1} \ VP\textcircled{2}$
$VP \rightarrow$	$PP\textcircled{1} \ VP\textcircled{2}$	$VP\textcircled{2} \ PP\textcircled{1}$
$VP \rightarrow$	$V\textcircled{1} \ AUX\textcircled{2}$	$AUX\textcircled{2} \ V\textcircled{1}$
$PP \rightarrow$	$NP\textcircled{1} \ P\textcircled{2}$	$P\textcircled{2} \ NP\textcircled{1}$
$NP \rightarrow$	<i>hamd ansary</i>	<i>Hamid Ansari</i>
$NP \rightarrow$	<i>na}b sdr</i>	<i>Vice President</i>
$V \rightarrow$	<i>namzd</i>	<i>nominated</i>
$P \rightarrow$	<i>kylye</i>	<i>for</i>
$AUX \rightarrow$	<i>taa</i>	<i>was</i>



Paraphrases

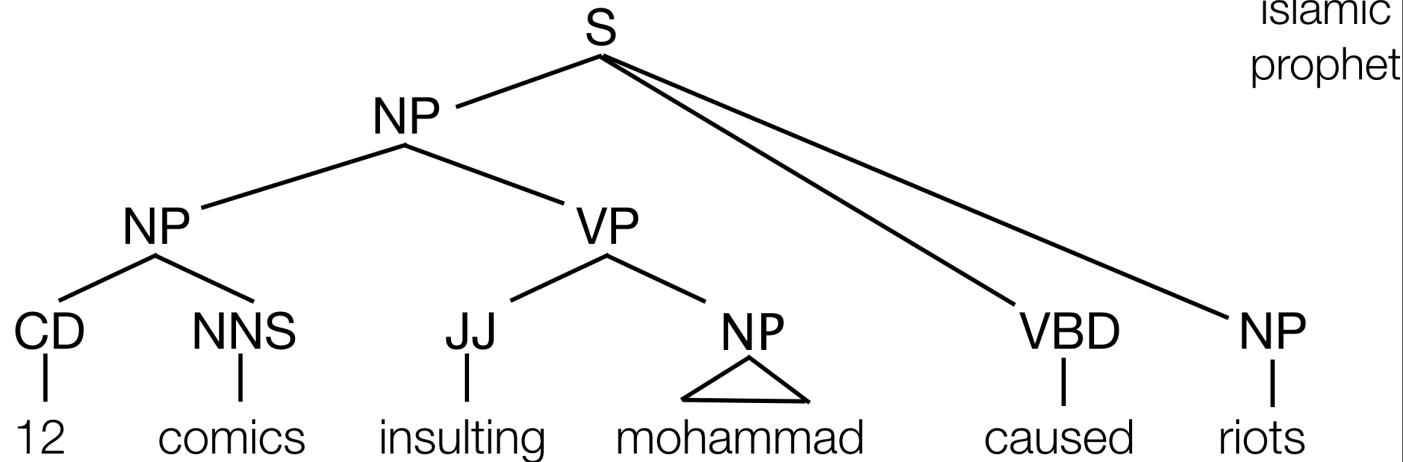
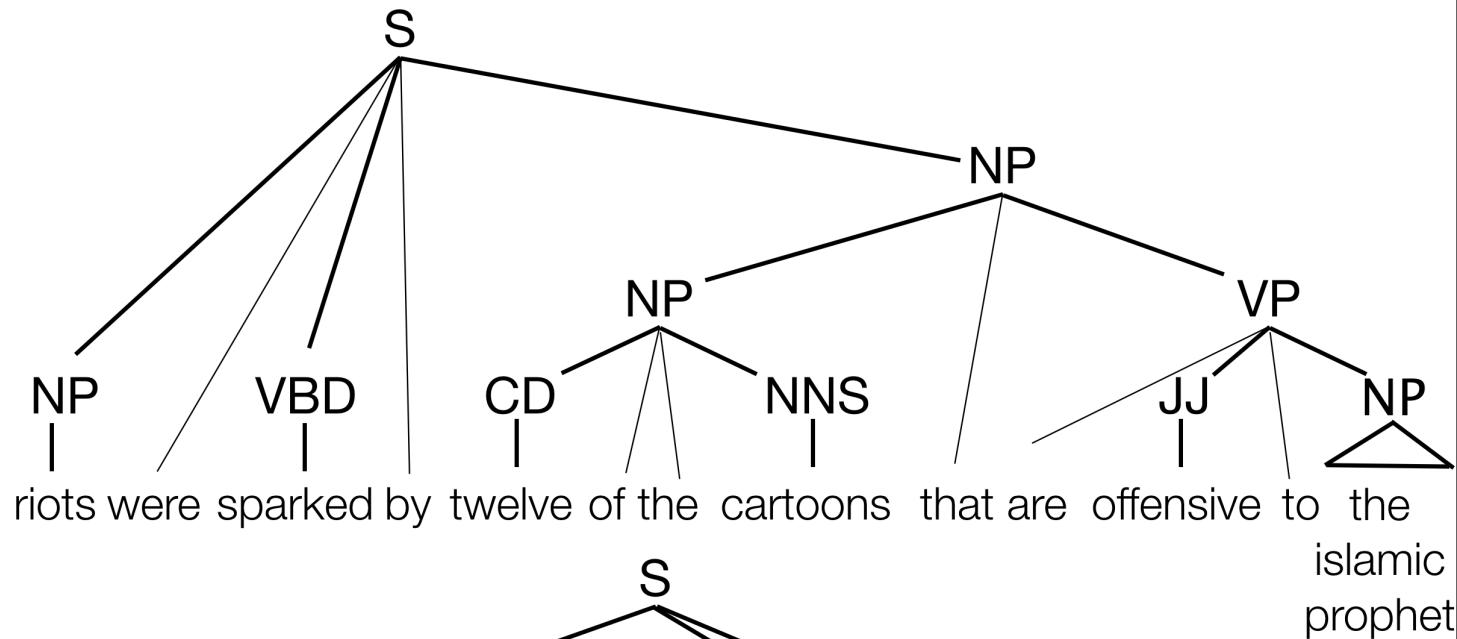
Differing **textual** expressions of the same meaning:

cup \leftrightarrow mug

the king's speech \leftrightarrow His Majesty's address

X_1 devours X_2 \leftrightarrow X_2 is eaten by X_1

one JJ instance of NP \leftrightarrow a JJ case of NP



Word Sense

bug

microbe, virus,
bacterium,
germ, parasite

insect, beetle,
pest, mosquito,
fly

bother, annoy,
pester

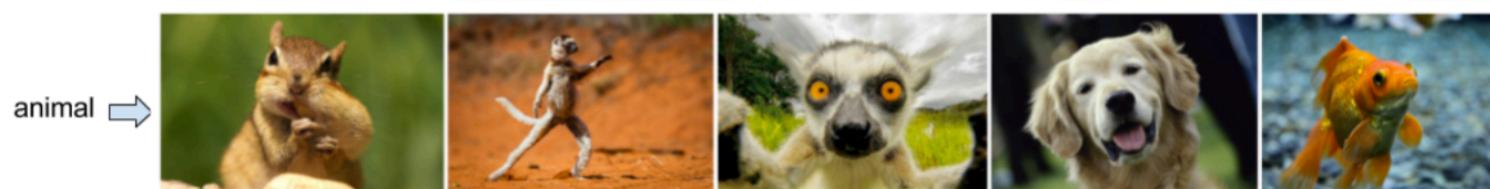
microphone,
tracker, mic,
wire, earpiece,
cookie

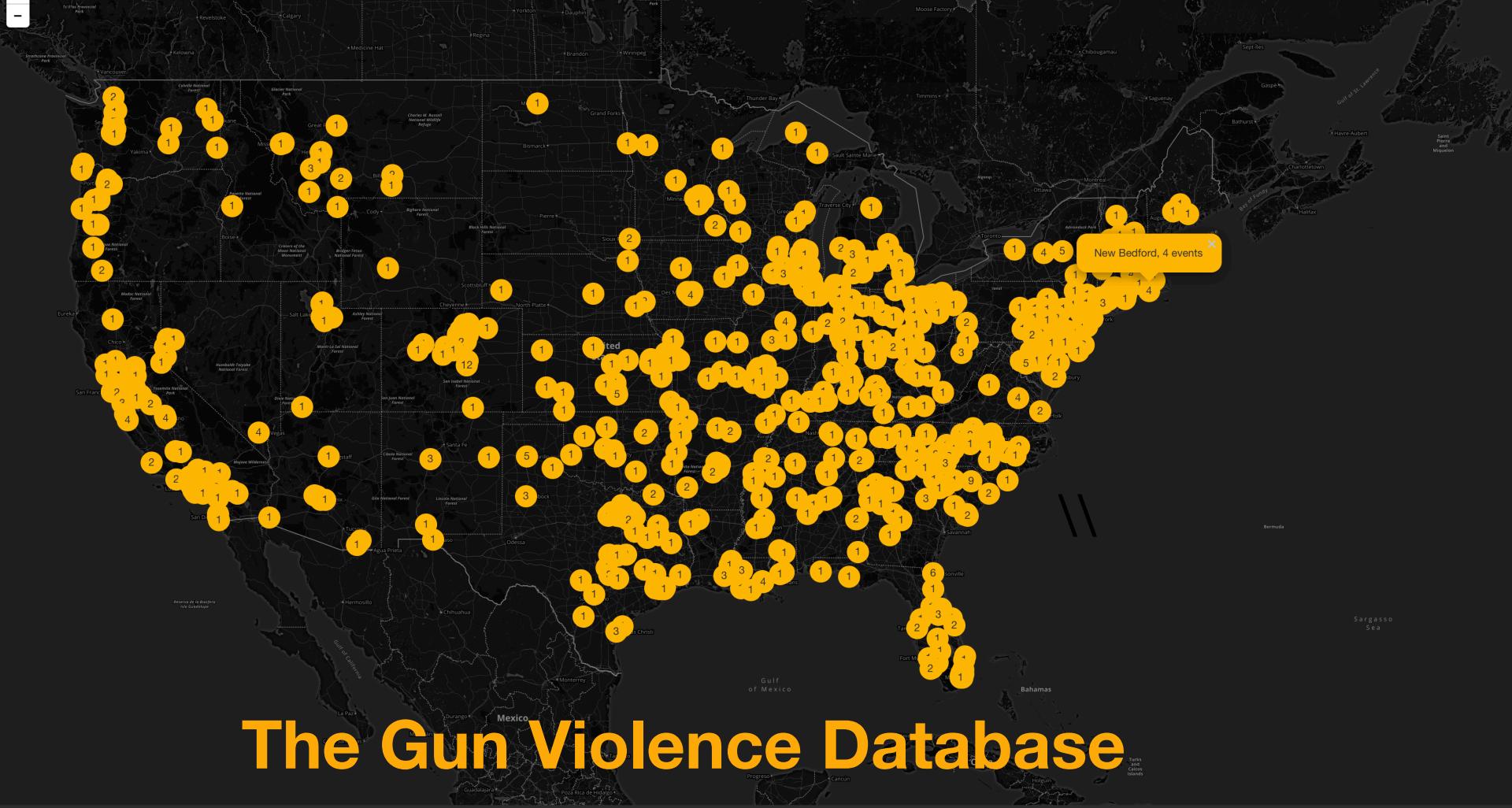
glitch, error,
malfunction,
fault, failure

squealer, snitch,
rat, mole

Semantic Relationships

twelve	12	equivalence
cartoons	illustrations	forward entailment
ε	in Denmark	reverse entailment
caused	prevented	negation
Europe	the middle East	alternation





Information Extraction

Chicago Police release Laquan McDonald shooting video | National News

Three seconds. On a dashcam video clock, that's the amount of time between the moment when two officers have their guns drawn and the point when Laquan McDonald falls to the ground. The video, released to the public for the first time late Tuesday, is a key piece of evidence in a case that's sparked protests in Chicago and has landed one officer behind bars. The 17-year-old McDonald was shot 16 times on that day the video shows in October 2014. Chicago police Officer Jason Van Dyke was charged Tuesday with first-degree murder....

Person #1014

Name	Laquan McDonald
Gender	
Age	
Race	

Incident #1053

City	
Date	
Shooter	
Victim	McDonald
Victim Killed	

What will you learn?

- This will be a survey class in natural language processing
- Focus will be programming assignments for hands-on learning
- Topics will include things like
 - Sentiment analysis
 - Vector space semantics
 - Parsing
 - Machine translation
 - Information extraction

Active Learning Class

- We are going to divide the course roughly in half
- Mondays will be traditional lectures like this one
- Wednesdays will be a “flipped classroom” with hands-on in-class group activities
- For Wednesdays you will be required to advanced preparation by reading the textbook and assigned research papers beforehand
- In-class quizzes on the assigned materials

Course Grading

- Weekly programming assignments (due Wednesdays by 11am)
 - In-class quizzes on the assigned readings
 - Self-designed final project
 - No final exam or midterm
-
- All homework assignments can be done in pairs, except for HW1
 - Final project will be teams of ~4-5 (details TBA)
 - 5 free late days for the term (1 minute - 24 hours = 1 day late)



Anne Cocos



Daphne Ippolito



John Hewitt



Nitish Gupta



Reno Kriz



Stephen Mayhew

Text Classification and Sentiment Analysis

Slides from **Speech and Language Processing** (3rd ed. draft) by
Dan Jurafsky and James H. Martin

Positive or negative movie review?



- unbelievably disappointing
- Full of zany characters and richly applied satire, and some great plot twists

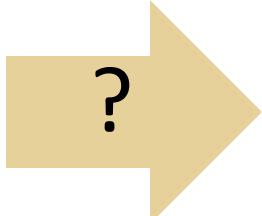


- this is the greatest screwball comedy ever filmed
- It was pathetic. The worst part about it was the boxing scenes.



What is the subject of this article?

MEDLINE Article



MeSH Subject Category Hierarchy

- Antagonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- ...

Classify User Attributes Using Their Tweets



Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping



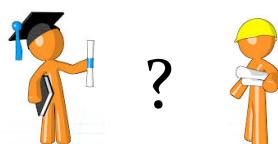
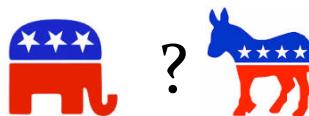
Yesterday's look-my new obsession is this Givenchy fur coat! Wolford sheer turtleneck, Proenza skirt & Givenchy boots



We've already tripled wind energy in America, but there's more we can do.



Two giant planets may cruise unseen beyond Pluto - space - June 2014 - New Scientist: newscientist.com/article/dn2571

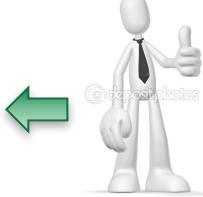


Lexical Markers for Age

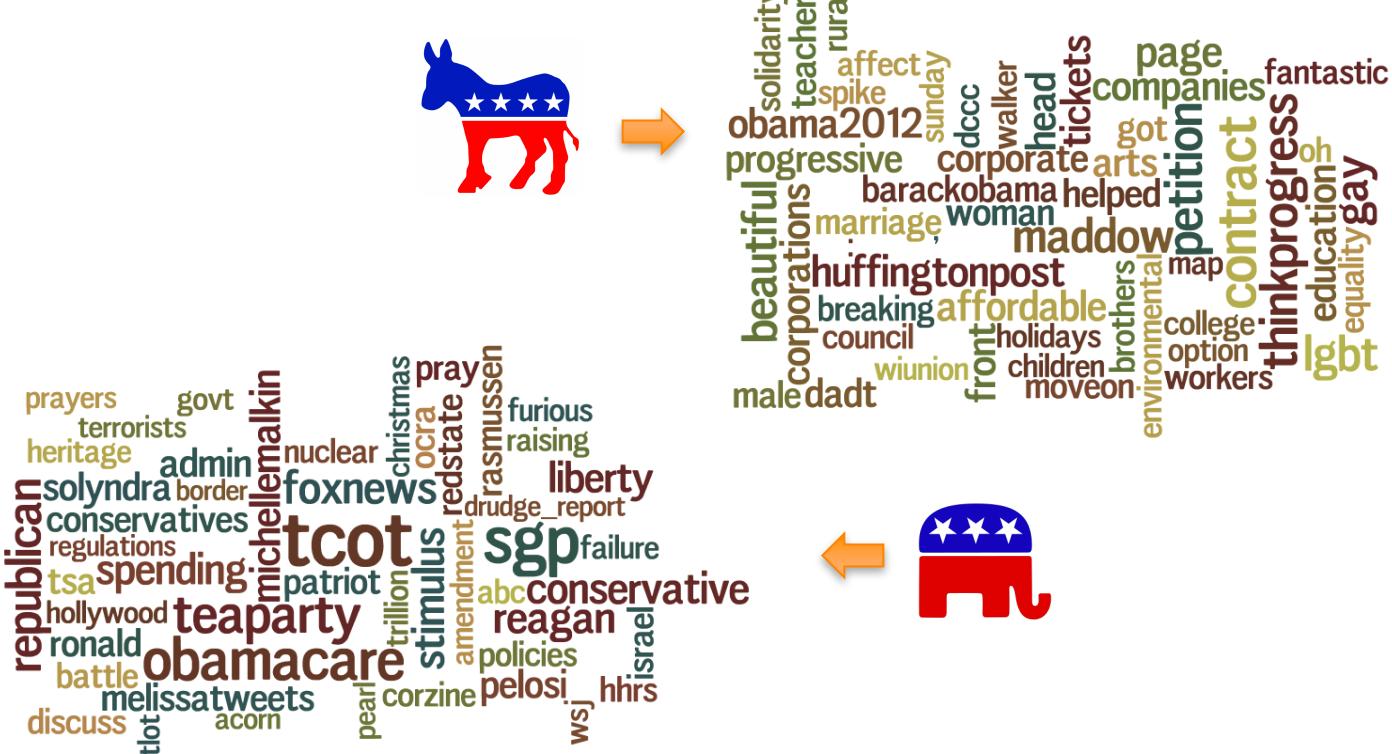


die cute mall
parents trending
obviously classes
18th exam
finals stop late
yeah followed
english sitting
sitting
merry forever me
dad actually college
probably dream
teacher xd
idk
studying
fight hahaha
perfect because
met light 20th
went
21st school
song asked
cant either
either
justin
xd
22nd

the men need
in love good and million
loved eye
sign that matter work
ladies lots getting
interested ticket thought
info quiet simply enjoy
peeps on wedding helps
30th bit with 27th ok
think wine thru of
my 25th 28th your
boo pic n't



Lexical Markers for Political Preferences



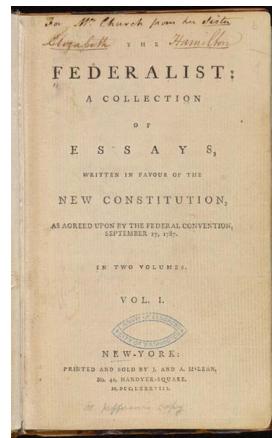
Lexical Markers for Gender



dude
xbox broo
money actor
bro writer
homies my-girl
place pretty
artist life
linux homie gangsta
i-hate-gays news+source
time homies world+series
chelsea homies
truck news
ps3 source
boyz news
bedjobs source
work android
dadcall+of+duty car
beer Carnigga
smoke game
people swearing
power guy
beer politics
smoke
news
cars



love
mother miss+you make-up
wonder his+hugs baby i+want+him my+tit
party dance heels women smile my+boy
ready lovely hair breast+cancer
ladies makeup dreams
layaway teenage+dream please ex+boyfriend
makeup laugh dream beats
cat+mommy simple pink sexy phone
females fashion dress romance people justin+bieber
Woman shopping singing daughter funny
slut beyonce
slut food crying
bitch cry
daddy cute
brunch my+sis
luv wedding jaja
mom peace chill dancing doll
husband chapstick young jealous
my+sis dancing love+him rainbow
best my+babies young
best fresh dancer
best best
luv my+boyfriend awww
wedding my+husband kiss
jaja clothes cuddle
brunch best-wishes
luv love+him girls life
luv my+babies better
brunch my+boyfriend bieber
luv dancing sleep



Who wrote which Federalist papers?

- 1787-8: anonymous essays try to convince New York to ratify U.S Constitution: Jay, Madison, Hamilton.
- Authorship of 12 of the letters in dispute
- 1963: solved by Mosteller and Wallace using Bayesian methods



James Madison



Alexander Hamilton

Your Orders

Q. Search all orders

Search Orders

Orders Open Orders Digital Orders Cancelled Orders

57 orders placed in 2016

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Carolyn Welsh ▾ ORDER # 105-5657220-9369068 Order Details Invoice

This is a gift order



The Federalist Papers (Signet Classics)
Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Carol Sidi ▾

This is a gift order



The Federalist Papers (Signet Classics)
Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Ash Khan ▾

This is a gift order



The Federalist Papers (Signet Classics)
Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Andrew Reilly ▾

This is a gift order



The Federalist Papers (Signet Classics)
Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

ORDER PLACED December 3, 2016 TOTAL \$18.99 SHIP TO Chris Callison-Burch ▾ ORDER # 116-4465007-6265007 Order Details Invoice

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Michael Downing ▾

ORDER # 105-9309957-2631421 Order Details Invoice

This is a gift order

Refund issued

A refund will appear on your original payment method in 2-4 business days. When will I get my refund?



The Federalist Papers (Signet Classics)

Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

Write a product review

Archive order

Write a product review

Archive order

The Federalist Papers (Signet Classics)

Hamilton, Alexander

Sold by: Amazon.com Services, Inc.

\$6.23

Condition: New

Gift options: Gift Receipt and Gift Message selected

"Federalist Paper 68 by Alexander Hamilton gives the rationale for your role as a presidential elector. I implore you to read it and consider it when casting your vote. Sincerely, Professor Chris Callison-Burch, University of Pennsylvania"

Buy it again



Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

Archive order

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Joyce Haas ▾

ORDER # 105-9893256-8192252 Order Details Invoice

This is a gift order

The Federalist Papers (Signet Classics)

Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

Write a product review

Archive order

16 TOTAL \$6.60 SHIP TO Tina Pickett ▾

order

The Federalist Papers (Signet Classics)

Hamilton, Alexander
Sold by: Amazon.com Services, Inc.
\$6.23

Buy it again

16 TOTAL \$6.60 SHIP TO Ted Christian ▾

16 TOTAL \$6.60 SHIP TO Robert Gleason ▾

16 TOTAL \$6.60 SHIP TO Robert Bozzuto ▾

order

16 TOTAL \$6.60 SHIP TO Robert Bozzuto ▾

order

57 orders placed in 2016

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Carolyn Welsh

 This is a gift order Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Carol Sidi

 This is a gift order Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Ash Khan

 This is a gift order Buy it again

ORDER PLACED December 7, 2016 TOTAL \$6.60 SHIP TO Andrew Reilly

 This is a gift order Buy it again

ORDER PLACED December 3, 2016 TOTAL \$18.99 SHIP TO Chris Callison-Bi

THIS MODERN WORLD

by TOM TOMORROW

PHILADELPHIA, 1787.
HEY, FOUNDING FATHERS!
I SEE YOU'RE WORKING
ON THE **CONSTITUTION!**
A TREMENDOUS DOCUMENT,
REALLY ONE OF THE BEST!

AND WHO MIGHT
YOU BE, GOOD
SIR?



I'M DONALD TRUMP--
ENORMOUSLY SUCCESSFUL
DEVELOPER OF FABULOUS
PROPERTIES ALL AROUND
THE WORLD! OH AND I'M
ALSO PRESIDENT-ELECT
IN THE YEAR 2016.

WHEN THEY SHOWED ME
THE TOP-SECRET **TIME
MACHINE**, I JUST KNEW
YOU GUYS WOULD LOVE
TO MEET ME!



FASCINATING! AND WHAT
HAVE THE PEOPLE, IN
THEIR INFINITE WISDOM,
ELECTED YOU TO **ACHIEVE?**

TO MAKE AMERICA
GREAT AGAIN!
WE'LL WIN SO
MUCH, YOU DEAD
GUYS WILL BE
SPINNING IN YOUR
GRAVES. HERE,
HAVE A HAT!



OF COURSE, I ALSO IN-
TEND TO USE EVERY
POWER AT MY DISPOSAL
TO **CRUSH** MY ENEMIES
BIGLY! WE'LL SEE HOW
FUNNY THEY THINK MY
HAIR IS THEN! AND LOOK
AT THESE HANDS! PER-
FECTLY NORMAL-SIZED,
AM I RIGHT?



WELL, I'D LOVE TO STAY
AND TELL YOU **MORE**
ABOUT MYSELF, BUT I'VE
GOT RALLIES TO LEAD--
AND BIASED, FAILING NEWS-
PAPERS TO **COMPLAIN**
ABOUT! THEY'RE **SO UN-**
FAIR TO **ME**! YOU GUYS
SHOULD **REALLY** RECON-
SIDER THAT FIRST AMEND-
MENT THING!

ZAP!



GENTLEMEN, I DO NOT
FULLY COMPREHEND WHAT
WE HAVE JUST WITNESSED,
BUT ONE THING IS **EX-**
CEEDINGLY CLEAR--



--WE MUST **ABANDON**
ANY INCLINATION WE MAY
HAVE HELD TOWARD A
SYSTEM OF DIRECT DEMO-
CRACY--AND INSTITUTE A
SAFEGUARD TO ENSURE
THAT SUCH A DANGEROUS
NARCISSIST WILL **NEVER**
BECOME PRESIDENT OF
THIS NATION!

IT SHALL BE KNOWN AS--



--**THE ELECTORAL**
COLLEGE!

OUR POSTERITY WILL NEVER
EVEN KNOW HOW CLOSE
THEY CAME TO UTTER
RUINATION!



TOM TOMORROW © 2016

TOTAL \$6.60 SHIP TO Tina Pickett

order

The Federalist Papers (Signet Classics)

Hamilton, Alexander

Sold by: Amazon.com Services, Inc.

\$6.23

 Buy it again

TOTAL \$6.60 SHIP TO Ted Christian

classics

SHIP TO Robert Gleason

classics

SHIP TO Robert Bozzuto

classics

Sold by: Amazon.com Services, Inc.

\$6.23

 Buy it again

TOTAL \$6.60 SHIP TO Robert Asher

order

2d

bear on your original payment method in 2-4 business days. When will I get

Text Classification

- Assigning subject categories, topics, or genres
- Spam detection
- Authorship identification
- Age/gender identification
- Language Identification
- Sentiment analysis
- ...

Sentiment Analysis

What is Sentiment
Analysis?

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



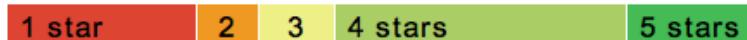
HP Officejet 6500A Plus e-All-in-One Color Ink-jet - Fax / copier / printer / scanner

\$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."

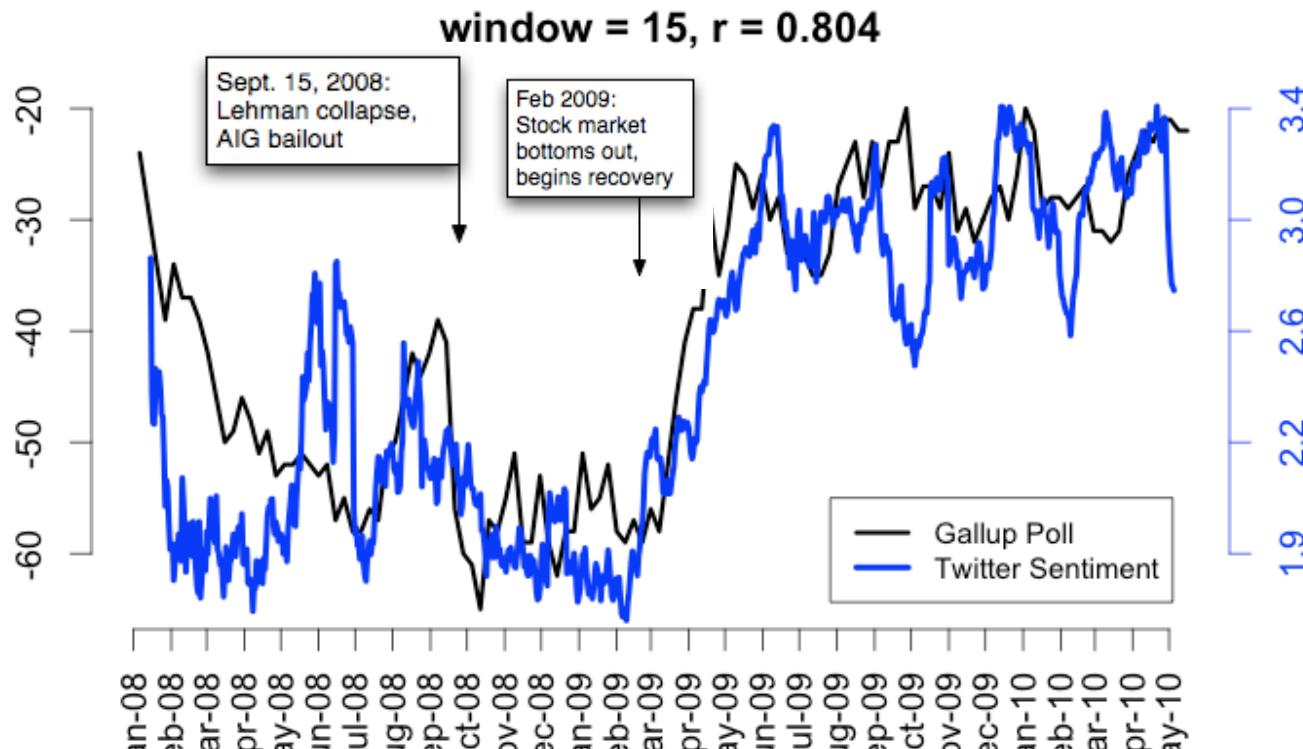
colors



"Full color prints came out with great quality."

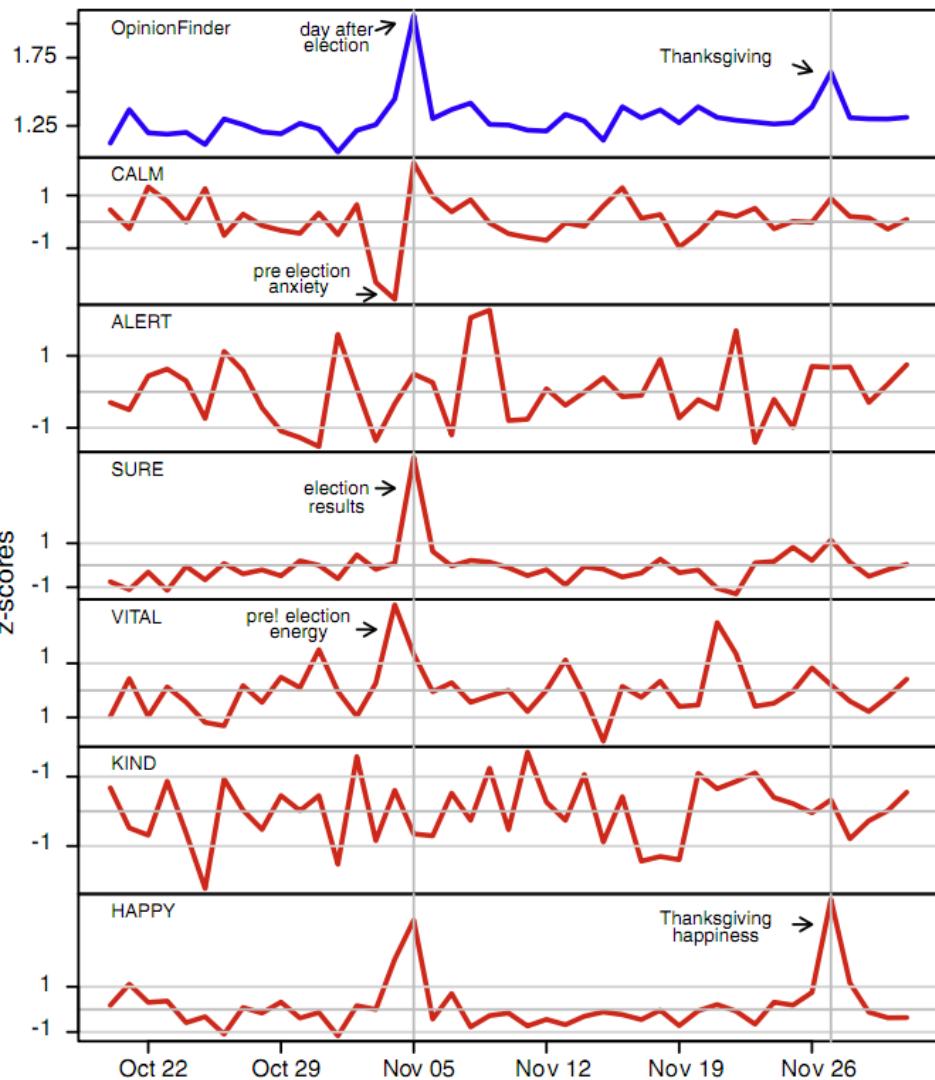
Twitter sentiment versus Gallup Poll of Consumer Confidence

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



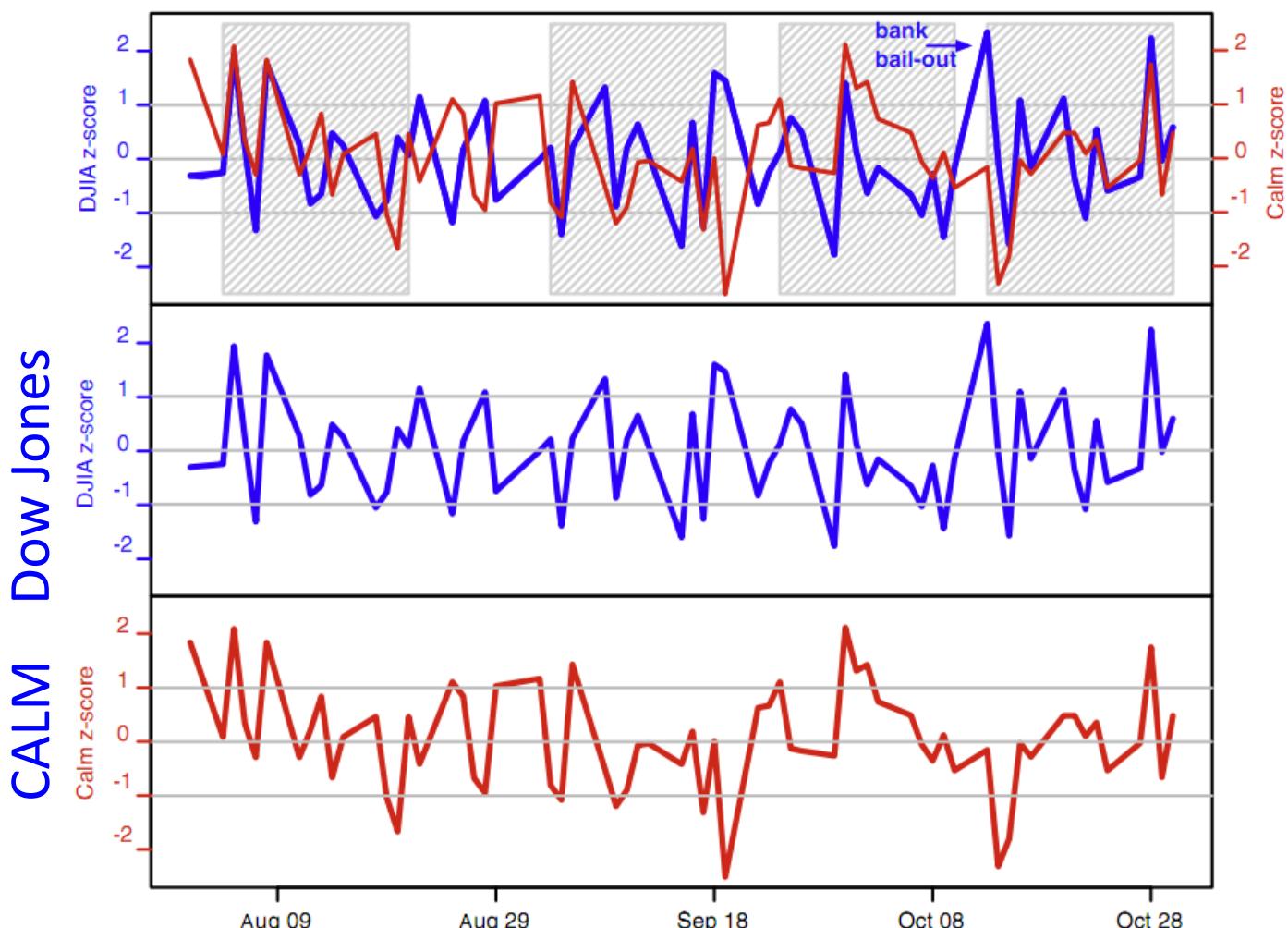
Twitter sentiment:

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



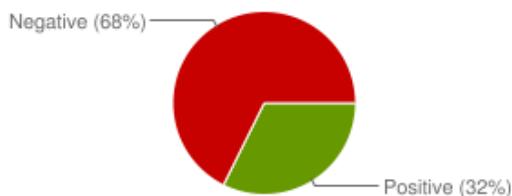
Target Sentiment on Twitter

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

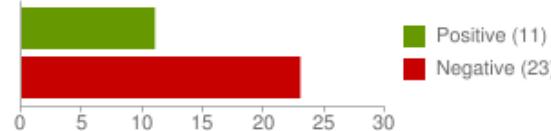
[Save this search](#)

Sentiment analysis for "united airlines"

Sentiment by Percent



Sentiment by Count



jacobson: 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

Sentiment analysis has many other names

- Opinion extraction
- Opinion mining
- Sentiment mining
- Subjectivity analysis

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

Scherer Typology of Affective States

- **Emotion:** brief organically synchronized ... evaluation of a major event
 - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood:** diffuse non-caused low-intensity long-duration change in subjective feeling
 - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances:** affective stance toward another person in a specific interaction
 - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes:** enduring, affectively colored beliefs, dispositions towards objects or persons
 - *liking, loving, hating, valuing, desiring*
- **Personality traits:** stable personality dispositions and typical behavior tendencies
 - *nervous, anxious, reckless, morose, hostile, jealous*

Scherer Typology of Affective States

- **Emotion:** brief organically synchronized ... evaluation of a major event
 - *angry, sad, joyful, fearful, ashamed, proud, elated*
- **Mood:** diffuse non-caused low-intensity long-duration change in subjective feeling
 - *cheerful, gloomy, irritable, listless, depressed, buoyant*
- **Interpersonal stances:** affective stance toward another person in a specific interaction
 - *friendly, flirtatious, distant, cold, warm, supportive, contemptuous*
- **Attitudes: enduring, affectively colored beliefs, dispositions towards objects or persons**
 - *liking, loving, hating, valuing, desiring*
- **Personality traits:** stable personality dispositions and typical behavior tendencies
 - *nervous, anxious, reckless, morose, hostile, jealous*

Sentiment Analysis

- 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
 - 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

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

Sentiment Analysis

A Baseline
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
 - CRF
 - Neural net

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](#)

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

Text Classification with Naïve Bayes

**The Task of Text
Classification**

Text Classification: definition

- *Input:*
 - a document d
 - a fixed set of classes $C = \{c_1, c_2, \dots, c_J\}$
- *Output:* a predicted class $c \in C$

Naïve Bayes Intuition

- Simple (“naïve”) classification method based on Bayes rule
- Relies on very simple representation of document
 - Bag of words

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!



it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

The bag of words representation

seen	2
sweet	1
whimsical	1
recommend	1
happy	1
...	...

Y
(

) = C



Bayes' Rule Applied to Documents and Classes

- For a document d and a class c

$$P(c|d) = \frac{P(d|c)P(c)}{P(d)}$$

Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(c | d)$$

MAP is “maximum a posteriori” = most likely class

$$= \operatorname{argmax}_{c \in C} \frac{P(d | c)P(c)}{P(d)}$$

Bayes Rule

$$= \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

Dropping the denominator

Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(d | c)P(c)$$

$$= \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

Document d
represented as
features
x1..xn

Multinomial Naïve Bayes Independence Assumptions

$$P(x_1, x_2, \dots, x_n | c)$$

- **Bag of Words assumption:** Assume position doesn't matter
- **Conditional Independence:** Assume the feature probabilities $P(x_i | c_j)$ are independent given the class c .

$$P(x_1, \dots, x_n | c) = P(x_1 | c) \cdot P(x_2 | c) \cdot P(x_3 | c) \cdot \dots \cdot P(x_n | c)$$

Multinomial Naïve Bayes Classifier

$$c_{MAP} = \operatorname{argmax}_{c \in C} P(x_1, x_2, \dots, x_n | c)P(c)$$

$$c_{NB} = \operatorname{argmax}_{c \in C} P(c_j) \prod_{x \in X} P(x | c)$$

Problems:

What makes reviews hard to classify?

- Subtlety:
 - Perfume review in *Perfumes: the Guide*:
 - “If you are reading this because it is your darling fragrance, please wear it at home exclusively, and tape the windows shut.”
 - Dorothy Parker on Katherine Hepburn
 - “She runs the gamut of emotions from A to B”

Thwarted Expectations and Ordering Effects

- “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.

Text Classification and Naïve Bayes

Parameter
Estimation and
Smoothing

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
 - simply use the frequencies in the data

$$\hat{P}(c_j) = \frac{\text{doccount}(C = c_j)}{N_{\text{doc}}}$$

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

Parameter estimation

$$\hat{P}(w_i | c_j) = \frac{\text{count}(w_i, c_j)}{\sum_{w \in V} \text{count}(w, c_j)}$$

fraction of times word w_i appears
among all words in documents of topic c_j

- Create mega-document for topic j by concatenating all docs in this topic
 - Use frequency of w in mega-document

Problem with Maximum Likelihood

- What if we have seen no training documents with the word ***fantastic*** and classified in the topic **positive (*thumbs-up*)**?

$$\hat{P}(\text{"fantastic"} \mid \text{positive}) = \frac{\text{count}(\text{"fantastic"}, \text{positive})}{\sum_{w \in V} \text{count}(w, \text{positive})} = 0$$

- Zero probabilities cannot be conditioned away, no matter the other evidence!

$$c_{MAP} = \operatorname{argmax}_c \hat{P}(c) \prod_i \hat{P}(x_i \mid c)$$

Laplace (add-1) smoothing for Naïve Bayes

$$\begin{aligned}\hat{P}(w_i | c) &= \frac{\text{count}(w_i, c) + 1}{\sum_{w \in V} (\text{count}(w, c) + 1)} \\ &= \frac{\text{count}(w_i, c) + 1}{\left(\sum_{w \in V} \text{count}(w, c) \right) + |V|}\end{aligned}$$

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
 - Calculate $P(w_k | c_j)$ 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 $Text_j$
- $$P(c_j) \leftarrow \frac{|docs_j|}{\text{total \# documents}}$$
- $$P(w_k | c_j) \leftarrow \frac{n_k + \alpha}{n + \alpha |Vocabulary|}$$

Text Classification and Naïve Bayes

Precision, Recall, and
the F measure

The 2-by-2 contingency table

	correct	not correct
selected	tp	fp
not selected	fn	tn

Precision and recall

- **Precision:** % of selected items that are correct
Recall: % of correct items that are selected

	correct	not correct
selected	tp	fp
not selected	fn	tn

A combined measure: F

- A combined measure that assesses the P/R tradeoff is F measure (weighted harmonic mean):

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}} = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

- The harmonic mean is a very conservative average; see *IIR* § 8.3
- People usually use balanced F1 measure
 - i.e., with $\beta = 1$ (that is, $\alpha = \frac{1}{2}$):

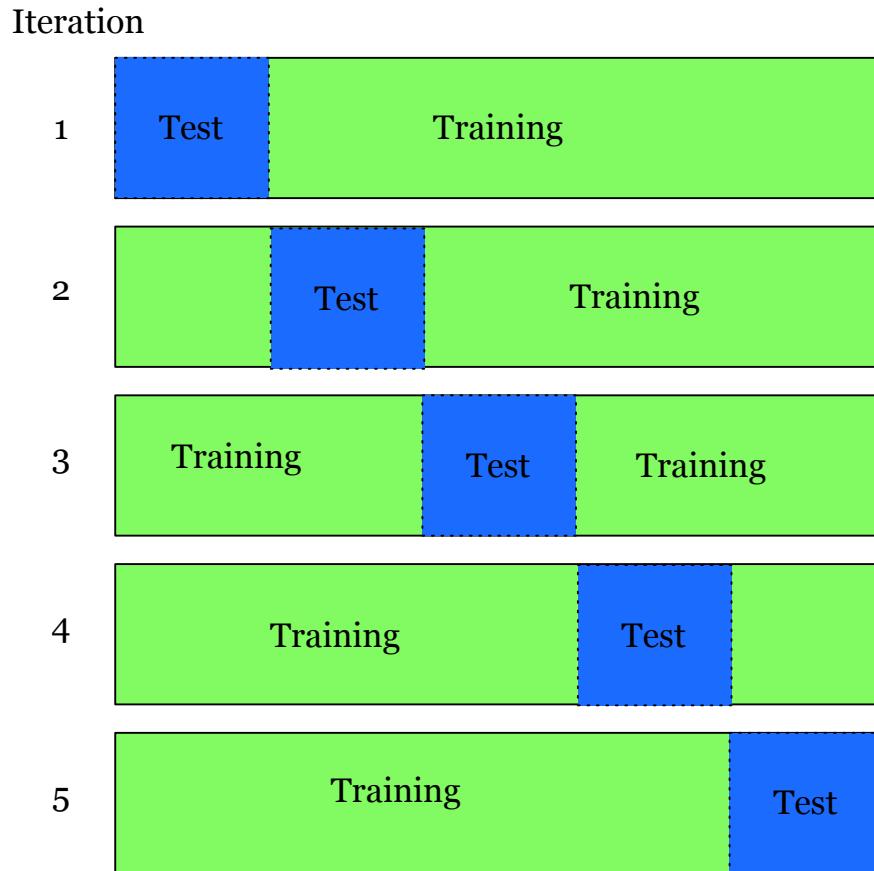
$$F = 2PR/(P+R)$$

Text Classification and Naïve Bayes

**Text Classification:
Evaluation**

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



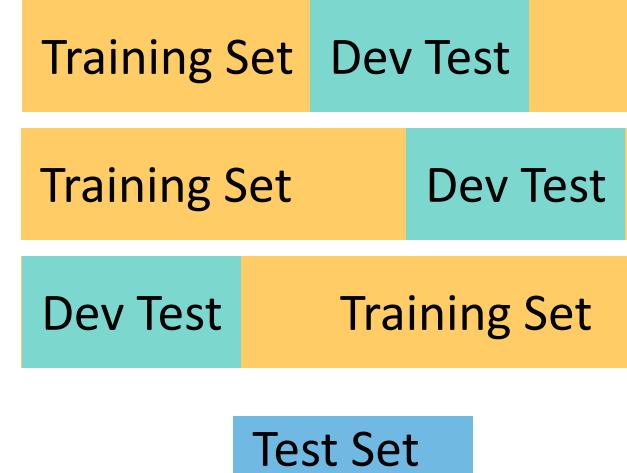
Development Test Sets and Cross-validation

Training set

Development Test Set

Test Set

- Metric: P/R/F1 or Accuracy
- Unseen test set
 - avoid overfitting ('tuning to the test set')
 - more conservative estimate of performance
- Cross-validation over multiple splits
 - Handle sampling errors from different datasets
 - Pool results over each split
 - Compute pooled dev set performance



Text Classification and Naïve Bayes

**Text Classification:
Practical Issues**

The Real World

- Gee, I'm building a text classifier for real, now!
- What should I do?

No training data? Manually written rules

If (wheat or grain) and not (whole or bread) then
 Categorize as grain

- Need careful crafting
 - Human tuning on development data
 - Time-consuming: 2 days per class

Very little data?

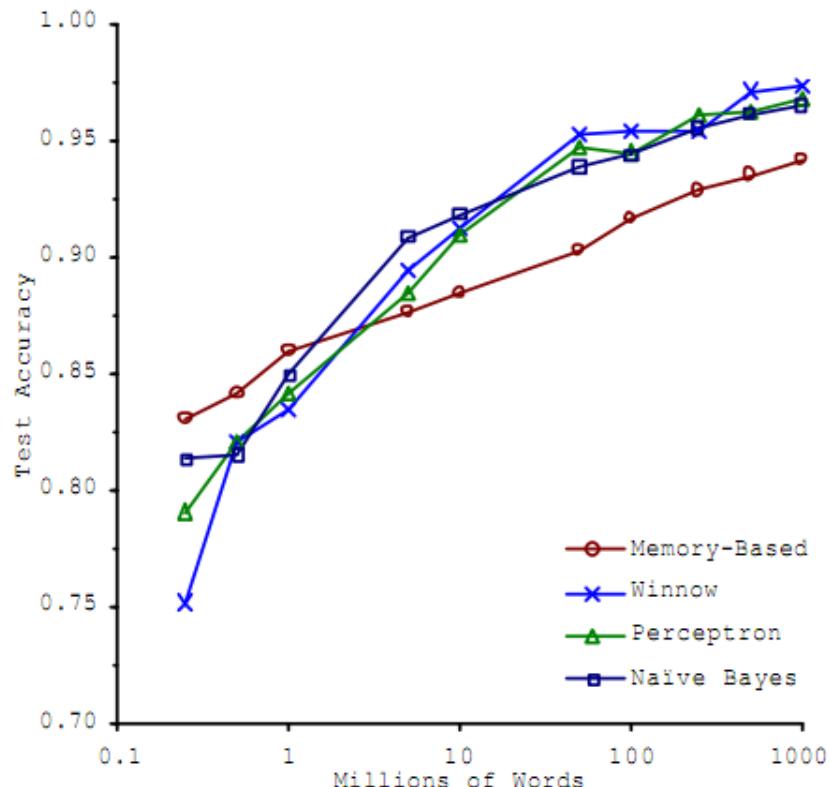
- Use Naïve Bayes
 - Naïve Bayes is a “high-bias” algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
 - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
 - Bootstrapping, EM over unlabeled documents, ...

A reasonable amount of data?

- Perfect for all the clever classifiers
 - SVM
 - Regularized Logistic Regression
- You can even use user-interpretable decision trees
 - Users like to hack
 - Management likes quick fixes

Accuracy as a function of data size

- With enough data
 - Classifier may not matter



Brill and Banko on spelling correction

Homework 1
Bash and Python
Skills
is available on the
web site

For next
Wednesday,
please read J&M v3
Chapters 2 and 6