CIS 530
Computational Linguistics

Mondays and Wednesdays 1:30-3pm

3401 Walnut, room 401B

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

Joined Penn faculy in 2013

I have been working in the field of NLP since 2000. In 2017, I was the general chair of the 55th meeting of the ACL.

Deniz Beser

Diana Marsala

Jasmine Lee

Jishnu Renugopal

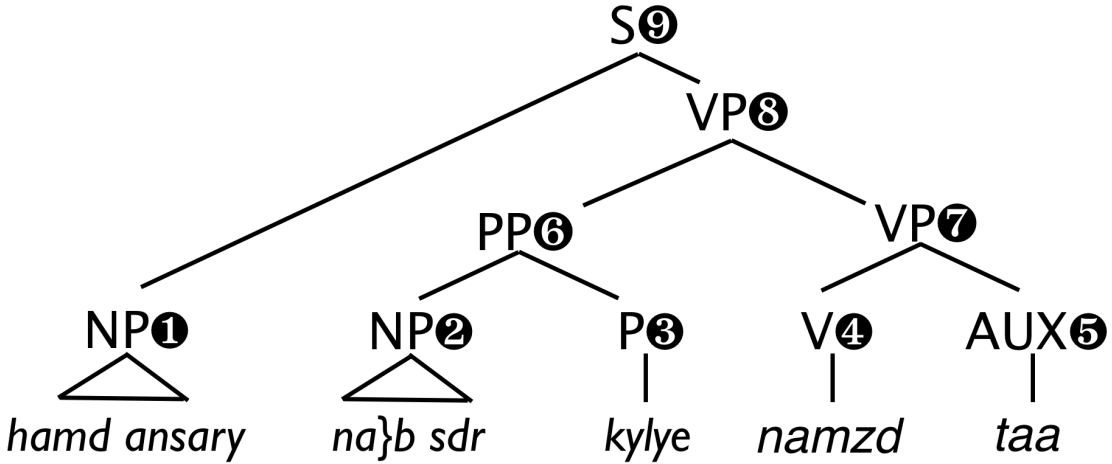
Nidhi Sridhar

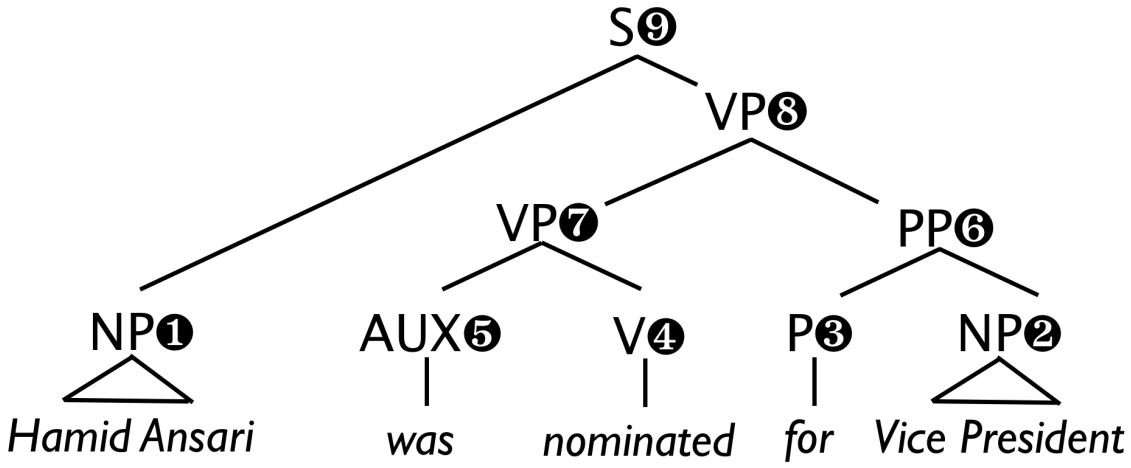
Nina Chang

Maria Kustikova

Shashank Garg

	Urdu	English
S →	NP① VP②	NP① VP②
VP→	PP① VP②	VP② PP①
VP→	V① AUX②	AUX② V①
PP →	NP① P②	P② NP①
NP →	*hamd ansary*	*Hamid Ansari*
NP →	*na}b sdr*	*Vice President*
V →	*namzd*	*nominated*
P →	*kylye*	*for*
AUX →	*taa*	*was*

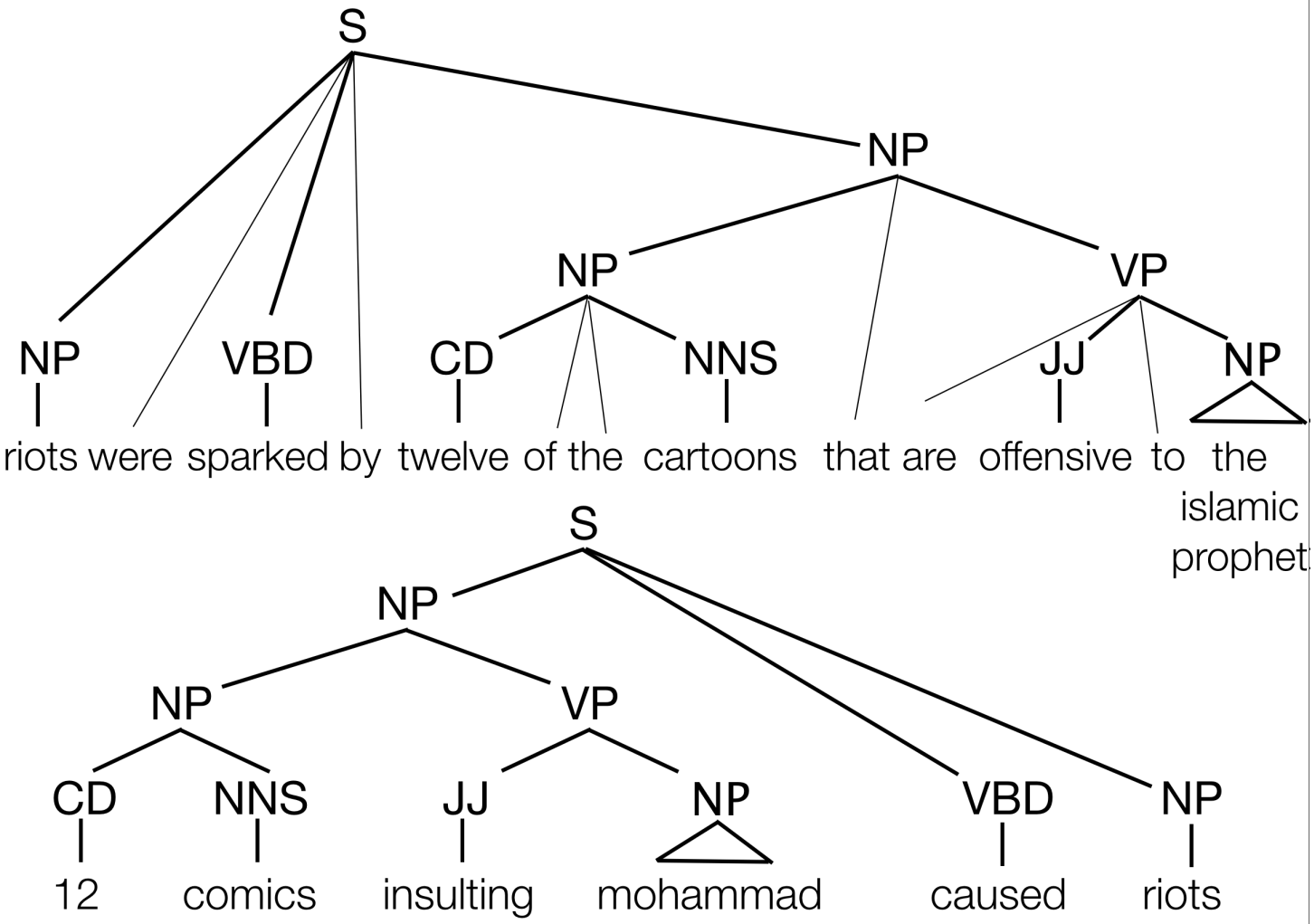

S❾
VP❽
PP❻
VP❼
NP❶
NP❷
P❸
V❹
AUX❺
hamd ansary
na}b sdr
kylye
namzd
taa


S❾
VP❽
VP❼
PP❻
NP❶
AUX❺
V❹
P❸
NP❷
Hamid Ansari
was
nominated
for
Vice President

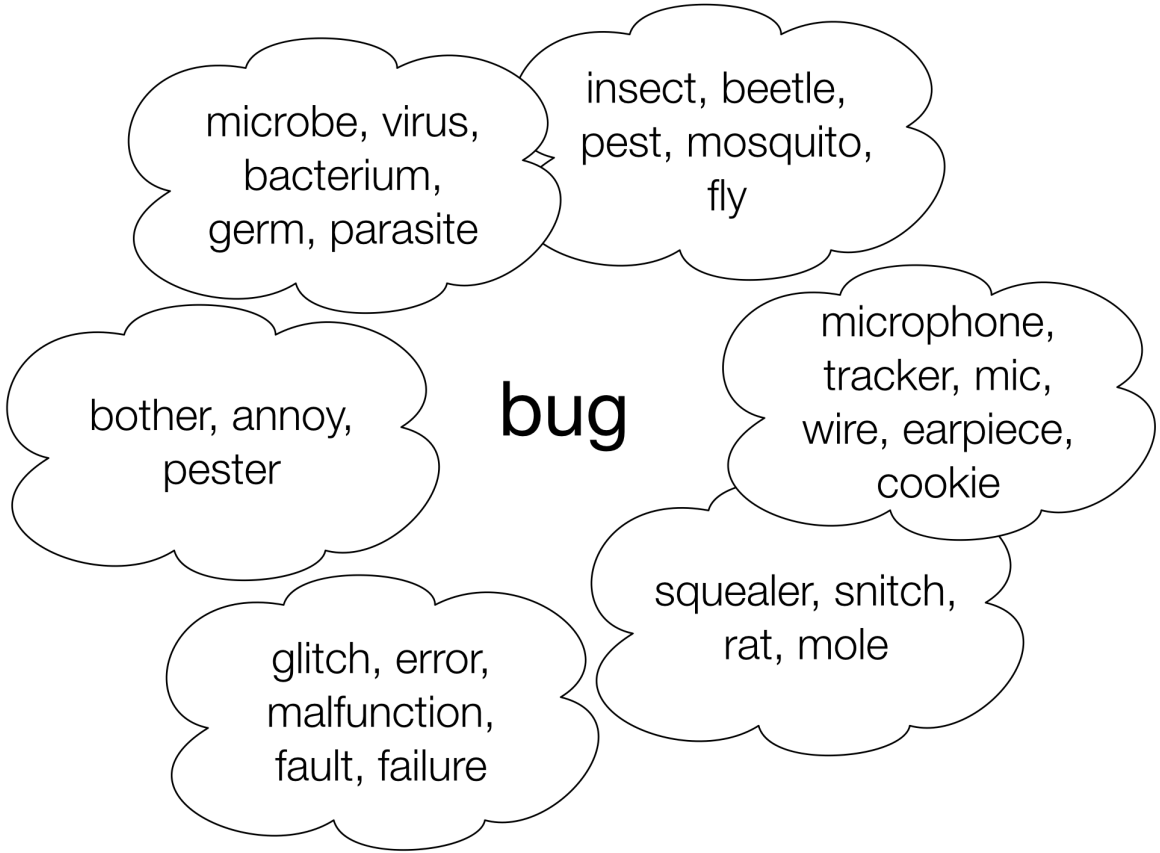
Paraphrases

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


S
NP
NP
VP
NP
VBD
CD
NNS
JJ
NP
riots were sparked by twelve of the cartoons that are offensive to the islamic prophet
S
NP
NP
VP
CD
NNS
JJ
NP
VBD
NP
12
comics
insulting
mohammad
caused
riots

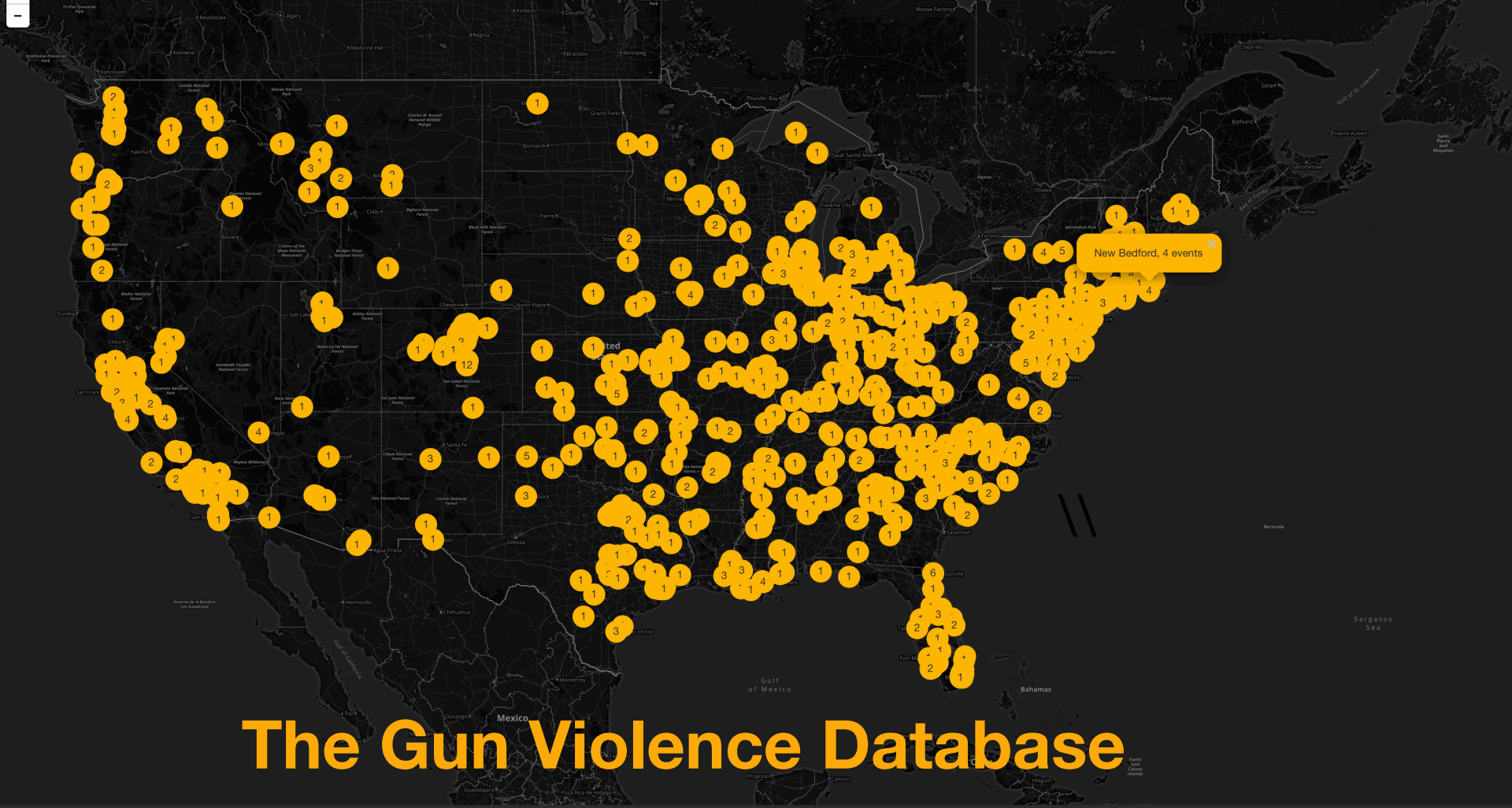
Word Sense


microbe, virus, bacterium, germ, parasite
insect, beetle, pest, mosquito, fly
bother, annoy, pester
bug
microphone, tracker, mic, wire, earpiece, cookie
squealer, snitch, rat, mole
glitch, error, malfunction, fault, failure

Semantic Relationships

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


kucing
★ cat
animal
persian
pet



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 a police 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	

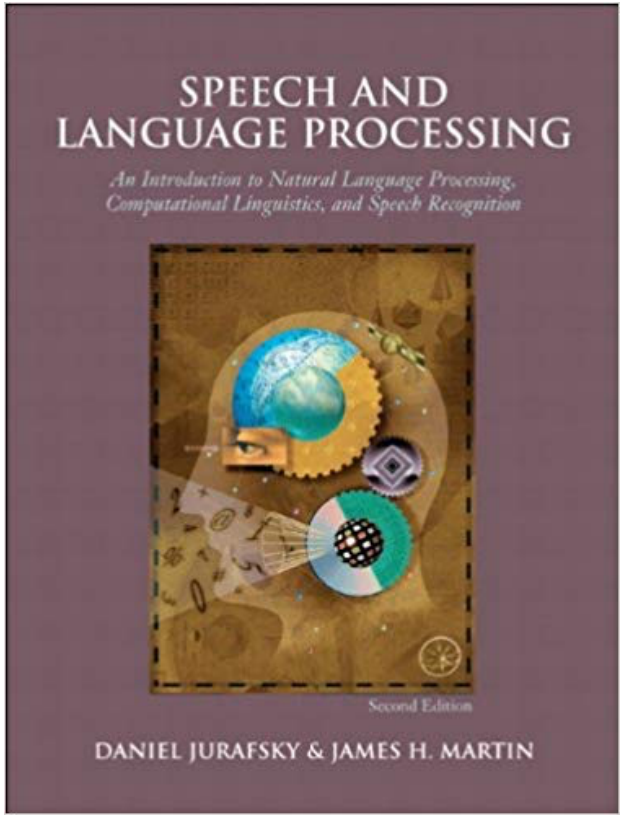
Incident #1053

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
  - Machine translation
  - Information extraction

Course textbook



Don’t buy this book!

The Authors are releasing free draft chapters of their updated 3rd edition.

https://web.stanford.edu/~jurafsky/slp3/

We will use the draft 3rd edition as our course textbook, along with required reading of research papers.

Course Grading

- Weekly programming assignments
- Short 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
- 5 free late days for the term (1 minute - 24 hours = 1 day late)

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

Syntactic frame and verb bias in aphasia: Plausibility judgments of undergoer-subject sentences

?

MeSH Subject Category Hierarchy

- Antogonists and Inhibitors
- Blood Supply
- Chemistry
- Drug Therapy
- Embryology
- Epidemiology
- …

Classify User Attributes Using Their Tweets

Follow

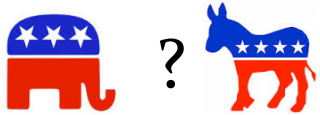
Delighted I kept my Xmas vouchers - Happy Friday to me 😊 #shopping

Follow

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

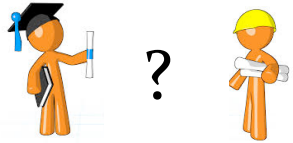
Follow

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



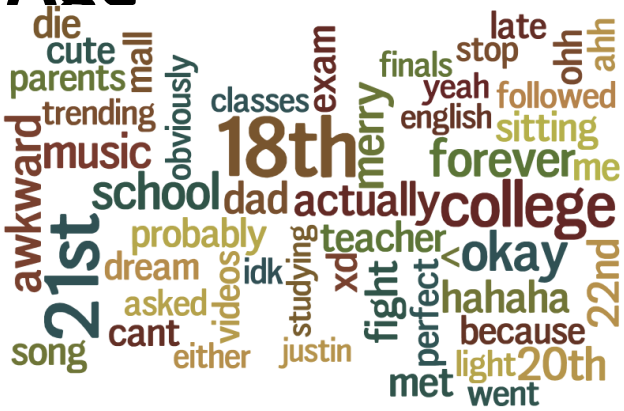
Follow

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



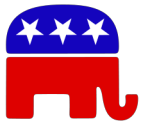
Slide from Svitlana Volkova

Lexical Markers for Age



Slide from Svitlana Volkova

Lexical Markers for Political Preferences

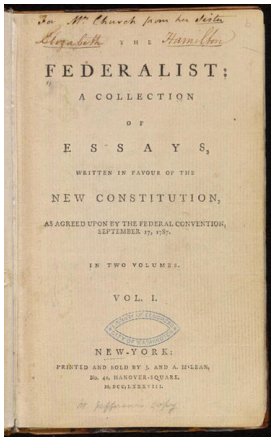


Slide from Svitlana Volkova

Lexical Markers for Gender

Slide from Svitlana Volkova

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

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

Write a product review
Archive order

ORDER PLACED December 7, 2016 | TOTAL $6.60 | SHIP TO Carol Sid

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 Khar

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

Write a product review
Archive order

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

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

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

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 | SHIP TO Ted Christian

SHIP TO Robert Gleason

SHIP TO Robert Bozzuto

Sold by: Amazon.com Services, Inc.
$6.23
Buy it again

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

Write a product review
Archive order

TOTAL $6.60 | SHIP TO Robert Asher

order

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

Your Orders

Orders | Open Orders | Digital Orders | Cancelled Ord

57 orders placed in 2016

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

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 Sid

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 Khar

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

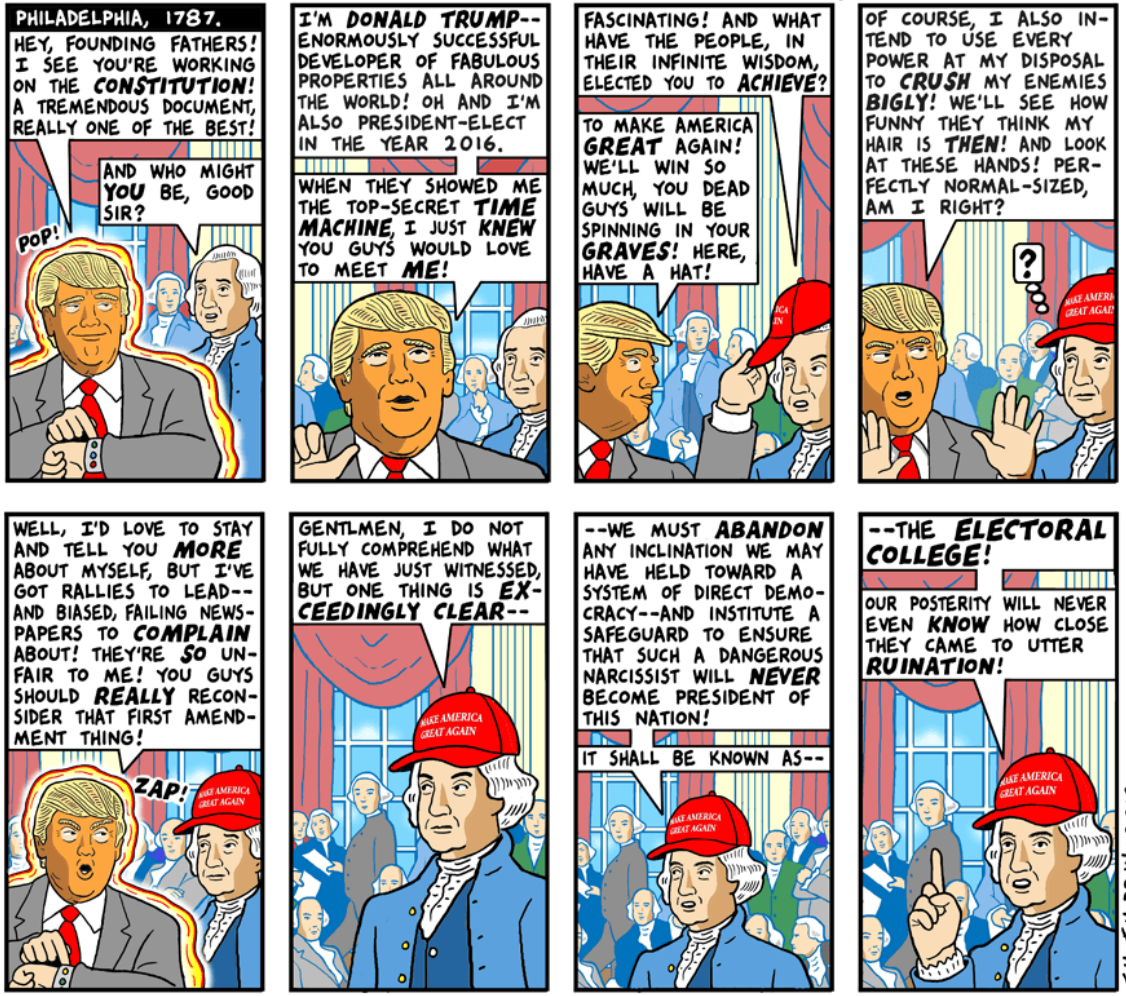
This is a gift order

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-B

THIS MODERN WORLD

by TOM TOMORROW



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 | SHIP TO Ted Christian

SHIP TO Robert Gleason

SHIP TO Robert Bozzuto

Sold by: Amazon.com Services, Inc.
$6.23
Buy it again

TOTAL $6.60 | SHIP TO Robert Asher

order

ear 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

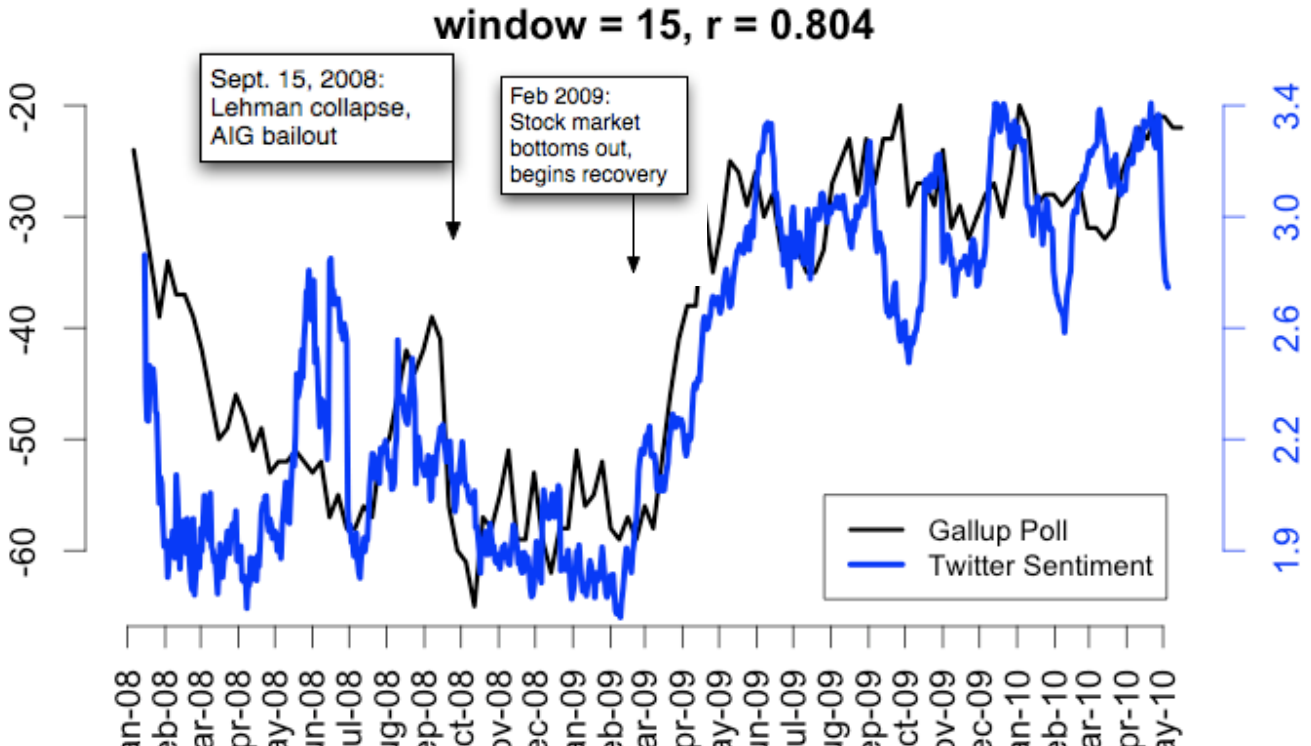
1 star	2	3	4 stars	5 stars

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."

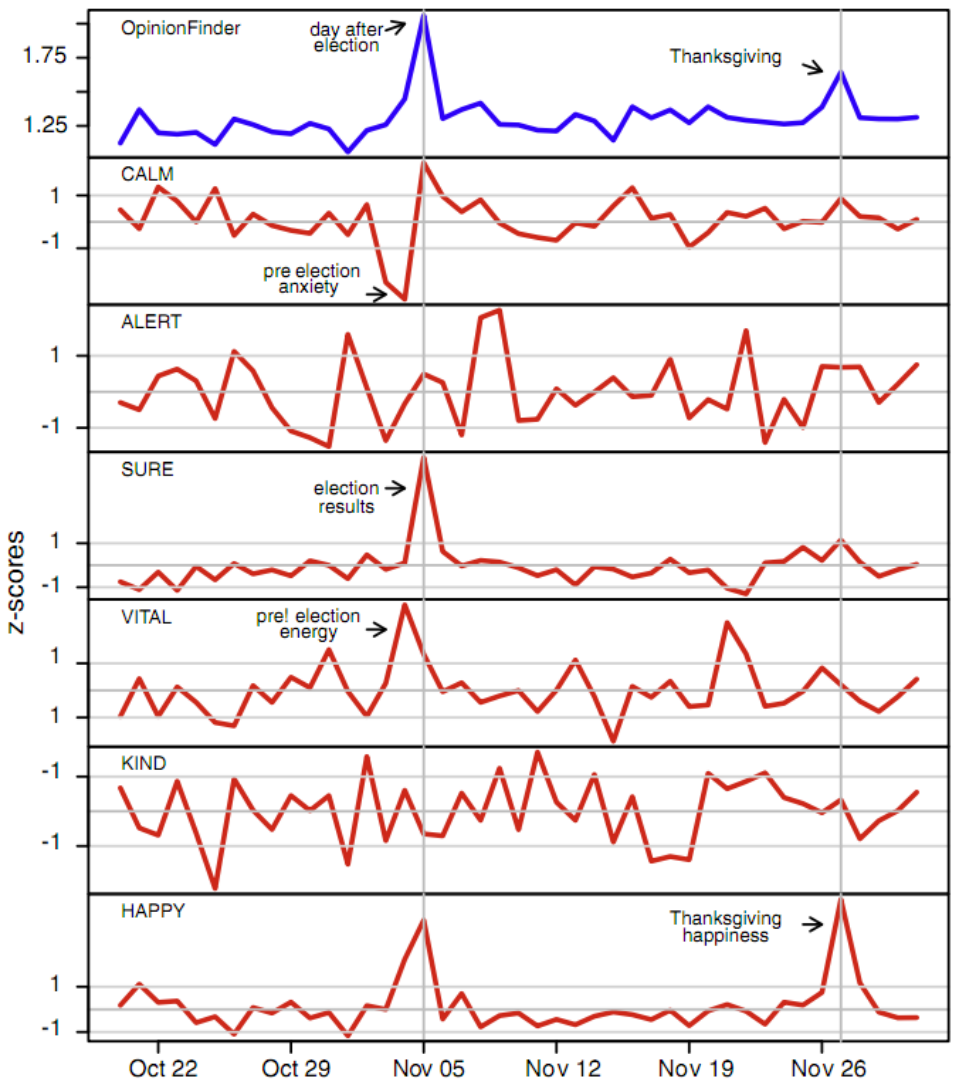
Twitter sentiment versus Gallup Poll of Consumer Confidence

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



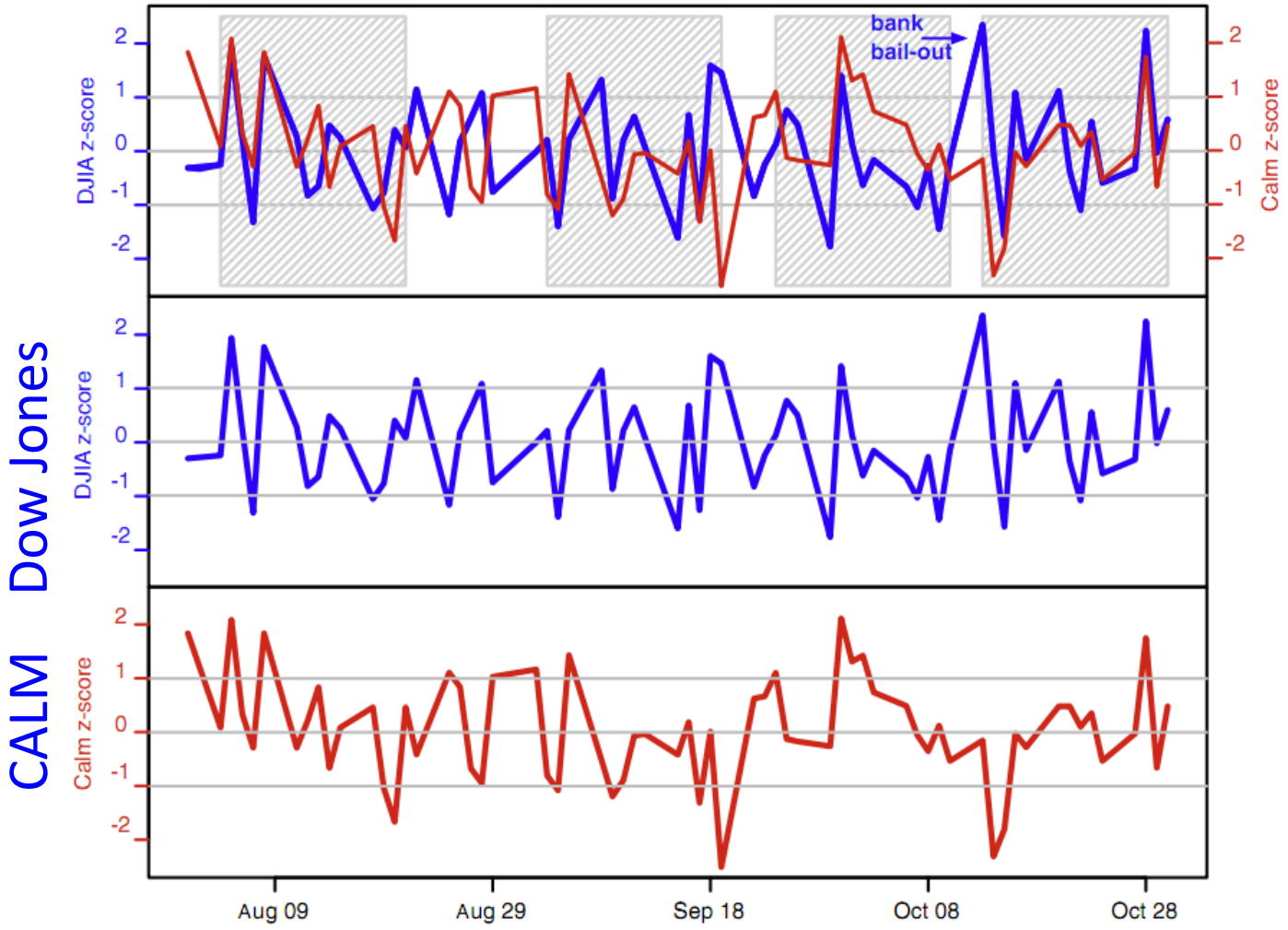
Twitter sentiment:

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

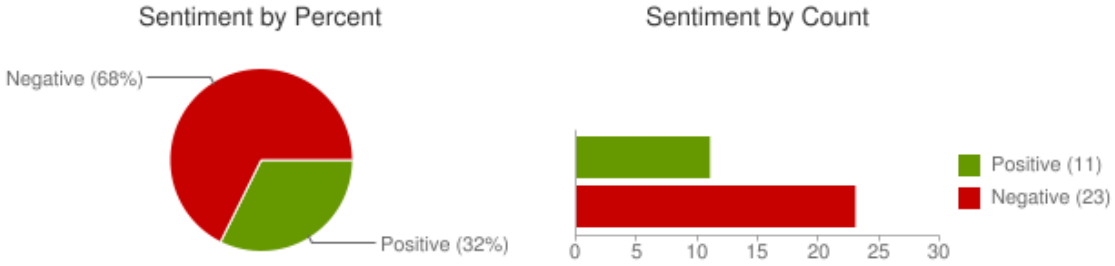


Target Sentiment on Twitter

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

"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

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, Klaus R. 1984. Emotion as a Multicomponent Process: A model and some cross-cultural data. In *Review of Personality and Social Psych* 5: 37-63.

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

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
```

Text Classification with Naïve Bayes

The Task of Text Classification

Text Classification: definition

- *Input*:
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$

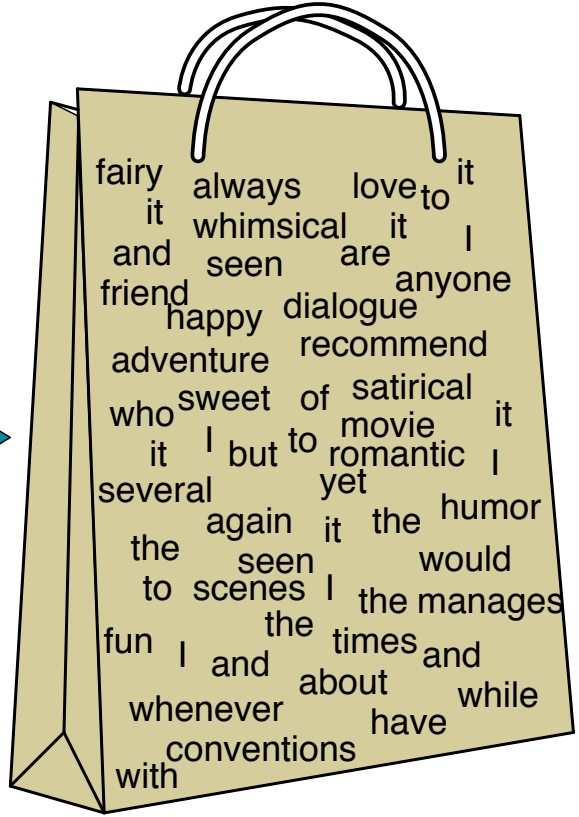
- *Output*: a predicted class $c \in C$

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

$\gamma($

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

$) = c$

Bayes' Rule Applied to Documents and Classes

- For a document *d* and a class *c*

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

Naïve Bayes Classifier

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

MAP is "maximum a posteriori" = most likely class

$$= \underset{c \in C}{\operatorname{argmax}} \frac{P(d \mid c)P(c)}{P(d)}$$

Bayes Rule

$$= \underset{c \in C}{\operatorname{argmax}} P(d \mid c)P(c)$$

Dropping the denominator

Naïve Bayes Classifier

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

$$= \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

Document d represented as features x1..xn

Multinomial Naïve Bayes Independence Assumptions

$$P(x_1, x_2, \ldots, x_n \mid c)$$

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

$$P(x_1, \ldots, x_n \mid c) = P(x_1 \mid c) \bullet P(x_2 \mid c) \bullet P(x_3 \mid c) \bullet \ldots \bullet P(x_n \mid c)$$

Multinomial Naïve Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

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

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

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

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

Parameter estimation

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

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)***?

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

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

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

Laplace (add-1) smoothing for Naïve Bayes

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) \right) + 1)}$$

$$= \frac{count(w_i, c) + 1}{\left(\sum_{w \in V} count(w, c) \right) + |V|}$$

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 $Text_j$

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

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):

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

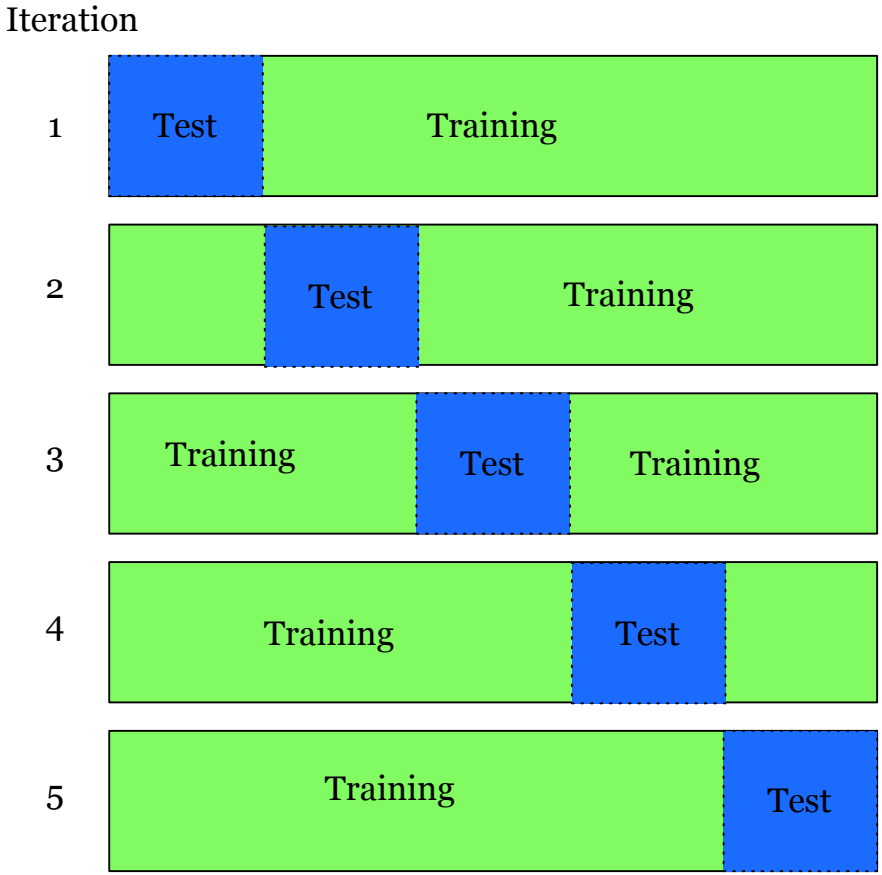
- The harmonic mean is a very conservative average; see *IIR* § 8.3
- People usually use balanced F1 measure
  - i.e., with $\beta = 1$ (that is, $\alpha = ½$): $F = 2PR/(P+R)$

Text Classification and Naïve Bayes

Text Classification: Evaluation

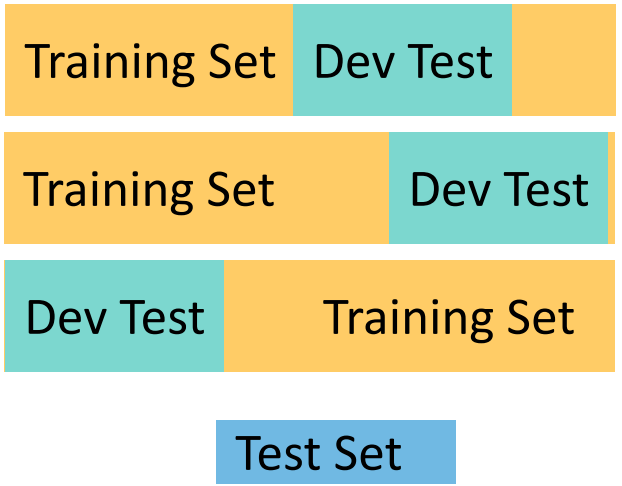
Cross-Validation

- 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

- 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?

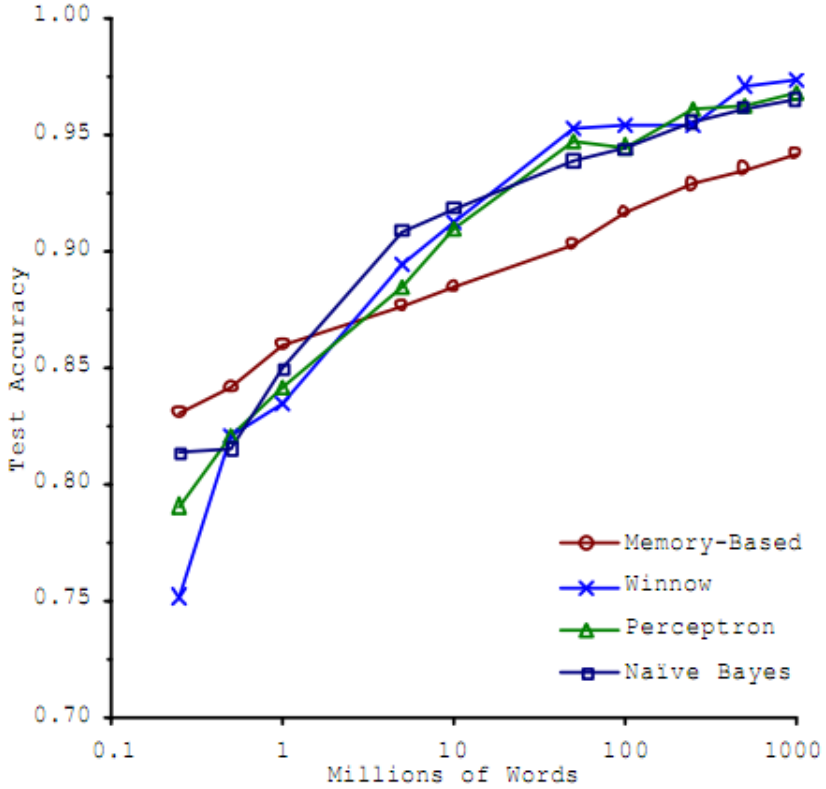
- Use Naïve Bayes
  - Naïve Bayes is a “high-bias” algorithm (Ng and Jordan 2002 NIPS)
- Get more labeled data
  - Find clever ways to get humans to label data for you
- Try semi-supervised training methods:
  - Bootstrapping, EM over unlabeled documents, …

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

For next Wednesday,

please read Jurafsky and Martin Chapter 4, and Thumbs up? Sentiment Classification using Machine Learning Techniques