

Text Categorization

(actually, methods apply for categorizing anything into fixed categories - tagging, WSD, PP attachment ...)

Why Text Categorization?

- Is it **spam**?
- Is it **Spanish**?
- Is it **interesting to this user**?
 - News filtering
 - Helpdesk routing
- Is it **interesting to this NLP program**?
 - e.g., should my calendar system try to interpret this email as an appointment (using info. extraction)?
- Where should it go **in the directory**?
 - Yahoo! / Open Directory / digital libraries
 - Which mail folder? (work, friends, junk, urgent ...)

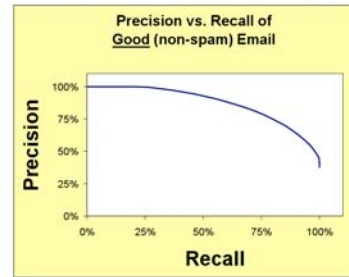
Measuring Performance

- Classification accuracy:** What % of messages were classified correctly?
- Is this what we care about?**

	Overall accuracy	Accuracy on spam	Accuracy on gen
System 1	95%	99.99%	90%
System 2	95%	90%	99.99%

- Which system do you prefer?

Measuring Performance

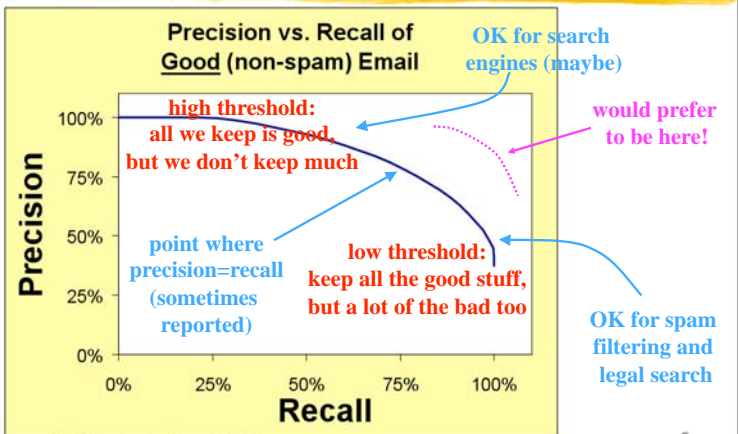


- Precision** = $\frac{\text{good messages kept}}{\text{all messages kept}}$
- Recall** = $\frac{\text{good messages kept}}{\text{all good messages}}$

Trade off precision vs. recall by setting threshold
 Measure the curve on annotated dev data (or test data)
 Choose a threshold where user is comfortable

$$F\text{-measure} = 1 / (\text{average}(1/\text{precision}, 1/\text{recall}))$$

Measuring Performance



More Complicated Cases of Measuring Performance

- For multi-way classifiers:**
 - Average accuracy (or precision or recall) of 2-way distinctions: Sports or not, News or not, etc.
 - Better, estimate the cost of different *kinds* of errors
 - e.g., how bad is each of the following?
 - putting Sports articles in the News section
 - putting Fashion articles in the News section
 - putting News articles in the Fashion section
 - Now tune system to minimize total cost
- For ranking systems:** Which articles / webpages most relevant?
 - Correlate with human rankings?
 - Get active feedback from user?
 - Measure user's wasted time by tracking clicks?

How to Categorize?

Subject: would you like to

. . . drive a new vehicle for free ? ? ? this is not hype or a hoax , there are hundreds of people driving brand new cars , suvs , minivans , trucks , or rvs . it does not matter to us what type of vehicle you choose . if you qualify for our program , it is your choice of vehicle , color , and options . we don ' t care . just by driving the vehicle , you are promoting our program . if you would like to find out more about this exciting opportunity to drive a brand new vehicle for free , please go to this site : http : / / 209 . 134 . 14 . 131 / ntr to watch a short 4 minute audio / video presentation which gives you more information about our exciting new car program . if you do n't want to see the short video , but want us to send you our information package that explains our exciting opportunity for you to drive a new vehicle for free , please go here : http : / / 209 . 134 . 14 . 131 / ntr / form . htm we would like to add you the group of happy people driving a new vehicle for free . happy motoring .

How to Categorize? (supervised)

We've seen lots of options in this course!

1. Build n-gram model of each category
 - Question: How to classify test message?
 - Answer: Bayes' Theorem

How to Categorize? (supervised)

We've seen lots of options in this course!

2. Represent each document as a vector (must choose representation and distance measure; use SVD?)
 - Question: How to classify test message?
 - Answer 1: Category whose centroid is most similar (may not work well if category is diverse)
 - Answer 2: Cluster each category into subcategories (then use answer 1 to pick a subcategory) (return the category that the subcategory is in) (this can also be useful for n-gram models)
 - Answer 3: Just look at labels of nearby training docs (e.g., let the k nearest neighbors vote - flexible!) (maybe the closer ones get a bigger vote)

How to Categorize? (supervised)

We've seen lots of options in this course!

3. Treat it like word-sense disambiguation
 - a) Vector model - use all the features (we just saw this)
 - b) Decision list - use single most indicative feature
 - c) Naive Bayes - use all the features, weighted by how well they discriminate among the categories
 - d) Decision tree - use some of the features in sequence
 - e) Other options from machine learning, like perceptron, Support Vector Machine (SVM), logistic regression, ...

Features matter more than which machine learning method

Review: Vector Model

These two documents are similar:
 After normalizing vector length to 1,
 Close in Euclidean space (similar endpoint)
 High dot product (similar direction)

aardvark abacus abandoned abbot abduct above
 (0, 0, 3, 1, 0, 7, ... 1, 0)

zygote zymurgy
 (0, 0, 1, 0, 0, 3, ... 0, 1)

Can play lots of encoding games when creating vector:
 Remove function words or reduce their weight
 Use features other than unigrams

slide courtesy of D. Yarowsky (modified)

Review: Decision Lists

To disambiguate a token of *lead* :

- Scan down the sorted list
- The first cue that is found gets to make the decision all by itself
- Not as subtle as combining cues, but works well for WSD

Cue's score is its log-likelihood ratio:
 $\log [p(\text{cue} | \text{sense A}) / p(\text{cue} | \text{sense B})]$

Position	Collocation	led	li:d
+1 L	lead level/N	219	0
-1 W	narrow lead	0	70
+1 W	lead in	207	898
..w +1w	of lead in	1167	0

LogL	Evidence	Pronunciation
11.40	follow/V + lead	⇒ li:d
11.20	zinc (in ±k words)	⇒ led
11.10	lead level/N	⇒ led
10.66	of lead in	⇒ led
10.59	the lead in	⇒ li:d
10.51	lead role	⇒ li:d
10.35	copper (in ±k words)	⇒ led
10.28	lead time	⇒ li:d
10.24	lead levels	⇒ led
10.16	lead poisoning	⇒ led
8.55	big lead	⇒ li:d
8.49	narrow lead	⇒ li:d
7.76	take/V + lead	⇒ li:d
5.99	lead, NOUN	⇒ led
1.15	lead in	⇒ li:d
	ooo	

Review: Combining Cues via Naive Bayes

Authorship ID: Who Wrote a Student's Term Paper?

Word in Text	Frequency as Student A	Frequency as Student B
optimally	97	1
certainly	84	3
typically	46	4
perspicuous	26	0
actually	13	4
whilst	6	0
the	241	229
awesome	0	63
totally	0	40
wonderful	0	26
incredibly	0	13

these stats come from term papers of *known* authorship
(i.e., supervised training)

$$\frac{P(\text{optimally}|\text{Student A})}{P(\text{optimally}|\text{Student B})} = \frac{97}{1} \quad \frac{P(\text{the}|\text{Student A})}{P(\text{the}|\text{Student B})} = \frac{1.1}{1}$$

Review: Combining Cues via Naive Bayes

$$\frac{P(\text{optimally}|\text{Student A})}{P(\text{optimally}|\text{Student B})} = \frac{97}{1} \quad \frac{P(\text{the}|\text{Student A})}{P(\text{the}|\text{Student B})} = \frac{1.1}{1}$$

$$\frac{P(\text{awesome}|\text{Student A})}{P(\text{awesome}|\text{Student B})} = \frac{0}{63}$$

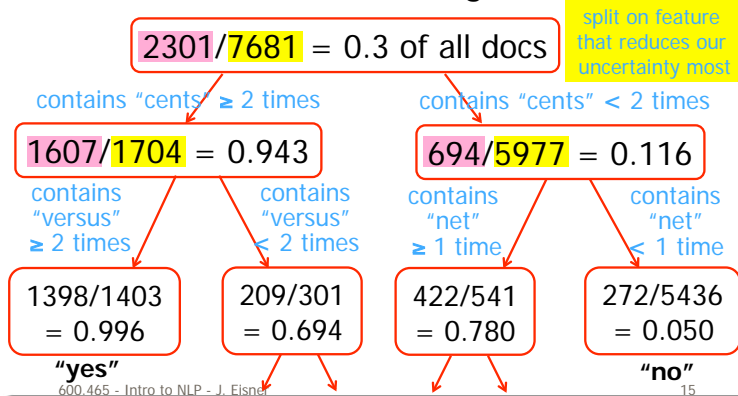
$$\frac{P(\text{Student A})}{P(\text{Student B})} \times \frac{P(w_1|\text{Student A})}{P(w_1|\text{Student B})} \times \frac{P(w_2|\text{Student A})}{P(w_2|\text{Student B})} \times \dots$$

"Naive Bayes" model for classifying text
(Note the naive independence assumptions!)

Would this kind of sentence be more typical of a student A paper or a student B paper?

Decision Trees

Is this Reuters article an Earnings Announcement?



Features Besides Unigrams

- All these approaches (except n-gram model) can use "interesting" features, not just unigrams.
- There's generally a heuristic feature selection problem
 - Use some very large set of features defined by a template
 - Maybe restrict to features that look useful in isolation?
 - Add features greedily, one at a time
 - Measure or guess expected improvement of each feature
 - Make sure to smooth when doing this - why?
 - At the end, remove features that hurt performance on held-out data
- What does SpamAssassin use?

SpamAssassin Features

- 100 From: address is in the user's black-list
- 4.0 Sender is on www.habeas.com Habeas Infringer List
- 3.994 Invalid Date: header (timezone does not exist)
- 3.970 Written in an undesired language
- 3.910 Listed in Razor2, see http://razor.sf.net/
- 3.801 Subject is full of 8-bit characters
- 3.472 Claims compliance with Senate Bill 1618
- 3.437 exists:X-Precedence-Ref
- 3.371 Reverses Aging
- 3.350 Claims you can be removed from the list
- 3.284 'Hidden' assets
- 3.283 Claims to honor removal requests
- 3.261 Contains "Stop Snoring"
- 3.251 Received: contains a name with a faked IP-address
- 3.250 Received via a relay in list.dsbl.org
- 3.200 Character set indicates a foreign language

SpamAssassin Features

- 3.198 Forged eudoraimail.com 'Received:' header found
- 3.193 Free Investment
- 3.180 Received via SBLed relay, seehttp://www.spamhaus.org/sbl/
- 3.140 Character set doesn't exist
- 3.123 Dig up Dirt on Friends
- 3.090 No MX records for the From: domain
- 3.072 X-Mailer contains malformed Outlook Expressversion
- 3.044 Stock Disclaimer Statement
- 3.009 Apparently, NOT Multi Level Marketing
- 3.005 Bulk email software fingerprint (jpfree) found inheaders
- 2.991 exists:Complain-To
- 2.975 Bulk email software fingerprint (VC_IPA) found inheaders
- 2.968 Invalid Date: year begins with zero
- 2.932 Mentions Spam law "H.R. 3113"
- 2.900 Received forged, contains fake AOL relays
- 2.879 Asks for credit card details

SpamAssassin Features

- 2.858 To: username at front of subject
- 2.851 Claims you actually asked for this spam
- 2.842 To header contains 'recipient' marker
- 2.826 Compare Rates
- 2.800 Received: says mail bounced all around the world
- 2.800 Mentions Spam Law "UCE-Mail Act"
- 2.796 Received via buggy SMTP server (MDaemon2.7.4SP4R)
- 2.795 Bulk email software fingerprint (StormPost) found in headers
- 2.786 Broken CGI script message
- 2.784 Message-Id generated by a spam tool
- 2.783 Urges you to call now
- 2.782 Tells you it's an ad
- 2.782 RAND found, spammer forgot to run the random-IDgenerator
- 2.748 Cable Converter
- 2.744 No Age Restrictions
- 2.737 Possible porn - Celebrity Porn

SpamAssassin Features

- 2.782 Tells you it's an ad
- 2.782 RAND found, spammer forgot to run the random-IDgenerator
- 2.748 Cable Converter
- 2.744 No Age Restrictions
- 2.737 Possible porn - Celebrity Porn
- 2.735 Bulk email software fingerprint (JiXing) found in headers
- 2.730 DNSBL: sender is Confirmed Spam Source
- 2.726 Bulk email software fingerprint (MMailer) found in headers
- 2.720 exists:X-Encoding
- 2.720 DNSBL: sender is Confirmed Open Relay
- 2.702 SEC-mandated penny-stock warning -- thanks SEC
- 2.695 Claims you can be removed from the list
- 2.693 Removes Wrinkles
- 2.668 Offers a stock alert
- 2.660 Listed in DCC, see <http://rhyolite.com/anti-spam/dcc/>
- 2.658 Common pyramid scheme phrase (1)

SpamAssassin Features

- 2.654 Offers a free consultation
- 2.645 Bulk email software fingerprint (EVAMAIL) found in headers
- 2.642 Possible porn - Amateur Porn
- 2.640 Listed in Razor1, see <http://razor.sf.net/>
- 2.639 Subject contains lots of white space
- 2.622 exists:X-x
- 2.620 Received via a relay in relays.visi.com
- 2.611 Bulk email software fingerprint (IMktg) found in headers
- 2.566 Compete for your business
- 2.565 Possible porn - Pay Site
- 2.541 Contains "CBIYI"
- 2.516 Spam phrases score is 34 to 55 (high)
- 2.513 Possible porn - Lesbian Site
- 2.510 Contains 'free installation' with capitals
- 2.502 Free Grant Money
- 2.500 Listed in Pyzor, see <http://pyzor.sf.net/>

SpamAssassin Features

- 2.500 TreŃlæ zawiera 'odes³anie z dopiskiem NIE'
- 2.500 TreŃlæ zawiera 'Artykuł 25 ust 2 punkt 2'
- 2.500 Treść zawiera 'przepraszamy za zajęty czas'
- 2.500 Treść zawiera 'Zamów teraz!!!'
- 2.500 Treść zawiera 'Jeżeli (Państwo) nie życzycie(sz)sobie'
- 2.500 Treść zawiera 'Aby usun±æ adres e-mail...'
- 2.496 Spam tool pattern in MIME boundary
- 2.492 'Message-Id' was added by a relay
- 2.488 Bulk email software fingerprint (screwup 1) found in headers
- 2.456 University Diplomas
- 2.450 Character set indicates foreign language body
- 2.445 Claims you can be removed from the list
- 2.443 Headers include 3 consecutive 8-bit characters
- 2.425 Date: is 24 to 48 hours after Received: date
- 2.421 'From' jun0.com does not match 'Received' headers
- 2.398 Meet Singles

SpamAssassin Features

- 2.362 Serious Enquiries Only.
- 2.361 Claims auto-email removal
- 2.357 MIME-Version header (oddly capitalized)
- 2.357 A "microsoft" header was found
- 2.351 X-Mailer contains "OutLook Express 3.14159"
- 2.334 Possible porn - Rape
- 2.331 "Collect Child Support" Scam
- 2.314 Claims spam helps the environment
- 2.292 Free Leads
- 2.290 Fake name used in SMTP HELO command
- 2.280 Received via a relay in ipwhois.rfc-ignorant.org
- 2.276 Possible porn - Cum Shot
- 2.261 Amazing Stuff
- 2.250 Received via a relay in orbs.dorkslayers.com
- 2.242 Possible porn - Mega Porn
- 2.240 Offers pure profit

SpamAssassin Features

- 2.216 Received contains a faked HELO hostname
- 2.210 Tells you it's an ad
- 2.209 Uses control sequences inside a URL's hostname
- 2.206 Claims spam helps the environment
- 2.203 Tells you to 'take action now!'
- 2.203 Cash Bonus
- 2.202 From an address @btamail.net.cn
- 2.180 exists:X-Library
- 2.176 Contains "My wife, Jody" testimonial
- 2.170 Possible porn - Nasty Girls
- 2.145 Promise you ...!
- 2.114 Claims to be in accordance with some Spam law
- 2.109 Uses a numeric IP address in URL
- 2.100 Possible porn - Live Porn
- 2.088 Discusses search engine listings
- 2.083 HTML comments which obfuscate text

SpamAssassin Features

- 2.066 Information on getting a larger penis or breasts (2)
- 2.066 Contains 'free preview' with capitals
- 2.060 A foreign language charset used in headers
- 2.052 Says "We strongly oppose the use of spam email"
- 2.044 trail of Received: headers seems to be forged
- 2.030 Credit Bureaus
- 2.022 Claims compliance with House Bill 4176
- 2.011 No Investment
- 2 Treŋjæ zawiera 'adres e-mail zostalnaleziony/pozyskany'
- 2 Treŋjæ zawiera 'adres (e-mail) pochodzi zogólnodostępných....'
- 2 Treŋjæ zawiera 'Ustawy o ochronie danychosobowych'
- 2 Tresc zawiera 'temat USUN'
- 2 Tresc zawiera 'na podstawie adresow e-mailpublicznie...'
- 2 Tresc zawiera 'kliknij w poniższy link'
- 2 Tresc zawiera 'do nabycia u nas'
- 2 Tresc zawiera 'Wys³aæ pusty mail'

SpamAssassin Features

- 2 Tresc zawiera 'Wiadomoŋjæ nadano na podstawie...'
- 2 Tresc zawiera 'Wiadomoŋjæ nadano jednorazowo...'
- 2 Tresc zawiera 'USUN Z BAZY'
- 2 Tresc zawiera 'Prosimy o przes³anie pustego maila'
- 2 Tresc zawiera 'Jeżeli nie interesuj±...'
- 2 Tresc zawiera 'Jeżeli nie chcesz (otrzymywac)...'
- 2 Tresc zawiera '...prosimy o zwrotny e-mail...'
- 2 Tresc zawiera '...adres z bazy...'
- 2 Dice cumplir con la ley
- 2 Clama cumplir con la normativa SPAM
- 1.995 Serious cash
- 1.984 Viagra and other drugs
- 1.977 If only it were that easy
- 1.952 Nigerian scam key phrase (million dollars)
- 1.910 Drastically Reduced
- 1.904 Contains "Temple Kiff"

SpamAssassin Features

- 1.889 Forged 'by gw05' 'Received:' header found
- 1.889 Credit Card Offers
- 1.880 Find out Anything
- 1.858 Contains "Gentle Ferocity"
- 1.856 Spam phrases score is 21 to 34 (high)
- 1.844 Possible Porn - Porn membership
- 1.842 Potential Earnings
- 1.839 Bulk email software fingerprint (Group Mail) found in headers
- 1.836 Once in a lifetime, apparently
- 1.831 Offers Free (often stolen) Passwords
- 1.824 Contains 'Dear (something)'
- 1.813 Possible porn - Porn Password
- 1.778 Message is 90-100% HTML tags
- 1.772 Sent using a trial version of CommuniGate
- 1.754 Date: is 48 to 96 hours after Received: date
- 1.744 To: has no local-part before @ sign

SpamAssassin Features

- 1.739 Talk about a check or money order
- 1.721 Contains 'for only pennies a day'
- 1.697 Spam tool pattern in MIME boundary
- 1.690 Form for checking email address
- 1.687 Subject: contains advertising tag
- 1.686 Talks about bulk email
- 1.682 Claims you registered with some kind of partner
- 1.681 Long Distance Phone Offer
- 1.663 Additional Income
- 1.640 Spam phrases score is 05 to 08 (medium)
- 1.640 Contains 'subject to credit approval'
- 1.639 Talks about tracing by SSN
- 1.631 Possible Porn - XXX Photos
- 1.625 Contains 'earn (dollar) something per week'
- 1.598 Message-Id has characters often found in spam
- 1.591 'X-Mailer' line contains gibberish

SpamAssassin Features

- 1.591 Cures Baldness
- 1.578 Subject starts with "Hello"
- 1.552 "Refinance your home"
- 1.548 Doing something with my income
- 1.546 Date: is 96 hours or more before Received: date
- 1.544 To: address contains spaces
- 1.539 Cents on the Dollar
- 1.526 Uses a username in a URL
- 1.523 Secretly Recorded
- 1.518 Invalid Date: header (not RFC 2822)
- 1.506 From and To are same (3)
- 1.505 Valid-looking To "undisclosed-recipients" exists:Date-warning
- 1.500 Temat zawiera 'oferta'
- 1.500 Treŋjæ zawiera 'Zaprosiaæ państwo'
- 1.500 Treŋjæ zawiera 'Szanowni Państwo'

SpamAssassin Features

- 1.500 Tresc zawiera 'publicznie dostępný (email)'
- 1.500 Tresc zawiera 'Upowaznienie do wystawiania fakturVAT...'
- 1.500 Tresc zawiera '...mail z tematem...'
- 1.495 Possible registry spammer
- 1.490 Possible porn - Adult Web Sites
- 1.486 'one time mailing' doesn't mean it isn't spam
- 1.479 Forged hotmail.com 'Received:' header found
- 1.470 Talks about opting in
- 1.466 Possible porn - Barely Legal
- 1.459 Claims compliance with Senate Bill 1618
- 1.435 Direct Marketing
- 1.410 Money back guarantee.
- 1.404 Date: is 48 to 96 hours before Received: date
- 1.404 Instructions on how to increase something
- 1.400 NOS CHILLAN PARA DECIR QUE ES GRATIS
- 1.394 Plugs Viagra

SpamAssassin Features

- 1.385 Spam phrases score is 08 to 13 (medium)
- 1.382 URL uses words and phrases which indicate porn (4)
- 1.373 As seen on national TV!
- 1.370 Message text disguised using base-64 encoding
- 1.368 Date: is 3 to 6 hours after Received: date
- 1.363 Score with babes!
- 1.361 From and To are same (6)
- 1.352 'From' yahoo.com does not match 'Received' headers
- 1.337 Spam phrases score is 13 to 21 (high)
- 1.332 Not intended for residents of XYZ.
- 1.319 Faked To "Undisclosed-Recipients"
- 1.314 From and To are same (5)
- 1.306 Only thing addresses on CD are useful for is spam
- 1.302 Contains "Vjestika Aphrodisia"
- 1.301 Lower Monthly Payment
- 1.293 HTML comment has 3 consecutive 8-bit characters

SpamAssassin Features

- 1.285 From: does not include a real name
- 1.283 Uses a dotted-decimal IP address in URL
- 1.275 Contains link without http:// prefix
- 1.274 'Subject' contains G.a.p.y-T.e.x.t
- 1.273 Marketing Solutions
- 1.270 Spam tool pattern in MIME boundary
- 1.269 'Prestigious Non-Accredited Universities'
- 1.253 Spam tool pattern in MIME boundary
- 1.253 Incorporates a tracking ID number
- 1.247 From and To are same (2)
- 1.246 Contains 'free sample' with capitals
- 1.231 Claims compliance with spam regulations
- 1.226 Online Pharmacy
- 1.224 Received via SMTPD32 server (SMTPD32-n.n)
- 1.218 Includes a form which will send an email
- 1.201 While you Sleep

SpamAssassin Features

- 1.187 Uses non-standard port number for HTTP
- 1.175 Possible porn - in ALL CAPS
- 1.148 Subject contains a unique ID
- 1.146 Bulk email software fingerprint (hash 2) found inheaders
- 1.138 Get Paid
- 1.131 Contains 'URGENT BUSINESS'
- 1.119 Why Pay More?
- 1.118 Requires Initial Investment
- 1.112 Javascript to open a new window
- 1.110 exists:X-List-Unsubscribe
- 1.099 Date: is 6 to 12 hours after Received: date
- 1.098 Subject starts with dollar amount
- 1.092 Increase your ejaculation!
- 1.084 Subject: contains Korean unsolicited email tag
- 1.084 Spam phrases score is 03 to 05 (medium)
- 1.078 Plugs "Herbal Viagra"

SpamAssassin Features

- 1.187 Uses non-standard port number for HTTP
- 1.175 Possible porn - in ALL CAPS
- 1.148 Subject contains a unique ID
- 1.146 Bulk email software fingerprint (hash 2) found inheaders
- 1.138 Get Paid
- 1.131 Contains 'URGENT BUSINESS'
- 1.119 Why Pay More?
- 1.118 Requires Initial Investment
- 1.112 Javascript to open a new window
- 1.110 exists:X-List-Unsubscribe
- 1.099 Date: is 6 to 12 hours after Received: date
- 1.098 Subject starts with dollar amount
- 1.092 Increase your ejaculation!
- 1.084 Subject: contains Korean unsolicited email tag
- 1.084 Spam phrases score is 03 to 05 (medium)
- 1.078 Plugs "Herbal Viagra"

SpamAssassin Features

- 1.077 Apparently, you'll be amazed
- 1.057 People just leave money laying around
- 1.045 Bulk email software fingerprint (eGroups) found inheaders
- 1.042 Date: is 24 to 48 hours before Received: date
- 1.039 Talks about direct email
- 1.038 Unneeded encoding of HTML tags
- 1.023 Javascript to move windows around
- 1.021 No such thing as a free lunch (3)
- 1.009 Save big money
- 1 Frequent SPAM content
- 1 Frequent SPAM content
- 1 Frequent SPAM content
- 1 Frequent SPAM content
- 1 Frequent SPAM content
- 1 Frequent SPAM content
- 1 Frequent SPAM content

SpamAssassin Features

- 1 Filename is just a '\#'; probably a JS trick
- 1 Old Murkowski disclaimer
- 1 Obfuscated action attribute in HTML form
- 1 Mentions monsterhut.com
- 1 Form for verifying email address
- 1 Contains signature of unregistered spam tool
- 1 Publicidad por e-mail
- 1 Contiene la palabra gratis en las cabeceras
- 1 exists:X-Fix
- 1 To: non-existent 'Investors' address
- 1 Subject contains 'Your Membership Exchange'
- 1 Spam tool pattern in MIME boundary
- 1 Reply-To: is empty
- 1 Received via a relay in bl.spamcop.net
- 1 Received via RSSed relay, see<http://www.mail-abuse.org/rss/>
- 1 Received via RBLed relay, see<http://www.mail-abuse.org/rbl/>

SpamAssassin Features

- 1 Received from first hop dialup, see <http://www.mail-abuse.org/dul/>
- 1 Received from dialup, see <http://www.mail-abuse.org/dul/>
- 1 Received contains fake 'Post.cz' hostname
- 1 From an address @email-publisher.com
- 1 Bulk email software fingerprint (xmailer tag) found in headers
- 1 Bulk email software fingerprint (pascual) found in headers
- 1 Bulk email software fingerprint (eBizmailer) found in headers
- 1 Bulk email software fingerprint (charset) found in headers
- 1 Bulk email software fingerprint (Yam) found in headers
- 1 Bulk email software fingerprint (V3161) found in headers
- 1 Bulk email software fingerprint (Uproar) found in headers
- 1 Bulk email software fingerprint (Seednet) found in headers
- 1 Bulk email software fingerprint (PowerCampaign) found in headers
- 1 Bulk email software fingerprint (Opt-In Lightning) found in headers
- 1 Bulk email software fingerprint (Matchmaker) found in headers
- 1 Bulk email software fingerprint (Mail Bomber) found in headers

SpamAssassin Features

- 1 Bulk email software fingerprint (Henry Su) found in headers
- 1 Bulk email software fingerprint (GRMessageQueue) found in headers
- 1 Bulk email software fingerprint (EPaper) found in headers
- 1 Bulk email software fingerprint (DiffondiCool) found in headers
- 1 Bulk email software fingerprint (CurrentMailer) found in headers
- 1 Bulk email software fingerprint (Caretop) found in headers
- 1 Bulk email software fingerprint (Campaign Blaster) found in headers
- 1 Bulk email software fingerprint ("outlook") found in headers
- 1 'Received:' contains huge hostname
- 1 'From' contains more than one address
- 1 Treŕlæ jest od wydawnictwa Verlag Dashofer (spamerzy)
- 1 Tresc zawiera 'Za zaliczeniem pocztowym...'
- 1 /zam.wieni/i
- 1 /zainteresowan.{0,50}wsp..prac/
- 1 /www\..adresy\.org/i
- 1 /specjaln.{0,50}ofert/i

SpamAssassin Features

- 1 Presentación de un nuevo producto.
- 1 Porno gratis.
- 1 Para dejar de fumar
- 1 Pago contra reembolso.
- 1 Nos animan a contestar si estamos interesados
- 1 No se puede considerar spam
- 1 Mensaje enviado por error
- 1 Mas informacion.
- 1 Los regalos no existen, salvo de nuestros amigos.
- 1 Inmigración legal (?) a los Estados Unidos
- 1 Informacion y reserva
- 1 If you want to subscribe...
- 1 If you send an email you will be OptOut
- 1 IMPERATIVOS EN MAYUSCULAS.
- 1 Haga click aqui.
- 1 Ha sido ganador.

SpamAssassin Features

- 1 Ha sido ganador.
- 1 El correo como alternativa comercial
- 1 Conviertete en Spammer.
- 1 Claims you can opt-out
- 1 Claims you can be removed in Spanish
- 1 Claims not to be spam in Spanish
- 1 Alta en buscadores hispanos.
- 1 spam software: PopLaunch
- 1 mentions Cyber FirePower!, a spam-tool
- 1 Will not Belive your Eyes!
- 1 Well known spam senders
- 1 Wants you to do business online
- 1 Things incredible
- 1 They keep your money -- No Refund!
- 1 Terms and conditions
- 1 Suspect you might have received the message by mistake

SpamAssassin Features

- 1 Slashed Price
- 1 SSPL found, spammer forgot to run the random-IDgenerator
- 1 Psychics Scam
- 1 Prices won't Last
- 1 Possible porn - Galleries of Pictures
- 1 Plugs "Natural Viagra"
- 1 Outstanding Values
- 1 Orders shipped by priority mail
- 1 No Middleman
- 1 No Medical Exams
- 1 No Gimmick
- 1 Nigerian scam, cf <http://www.snopes2.com/inboxer/scams/nigeria.htm>
- 1 New Customers Only
- 1 More Internet Traffic
- 1 Luxury Car
- 1 List removal information

SpamAssassin Features

- 1 Get Started Now
- 1 Cyber FirePower! rant about losing dropboxes
- 1 Confidentially on all orders
- 1 Claims you were on a list
- 1 Claims to listen to some removal request list
- 1 Claims not to be spam
- 1 Claims not to be selling anything
- 1 Claims compliance with spam regulations
- 1 Claims compliance with spam regulations
- 1 Claims "This is not junk email"
- 1 Cell Phone Cancer Scam
- 1 Buying judgements
- 1 Achieve Wealth
- 0.982 Talks about future mailings
- 0.977 Excessive quoted-printable encoding in body
- 0.975 Multi Level Marketing mentioned

SpamAssassin Features

0.968 Possible porn - Hardcore Porn
0.959 Missing To: header
0.954 From: has no local-part before @ sign
0.952 Targeted Traffic / Email Addresses
0.948 Information on getting a larger penis or breasts
0.947 Message is 70-90% HTML tags
0.935 Free Membership
0.931 To: and Cc: contain similar domains at least 8 times
0.910 Received contains a (dollar) variable reference
0.908 Claims compliance with spam regulations
0.906 'From' ebay.com does not match 'Received' headers
0.904 Unlimited in caps
0.900 Accept Credit Cards
0.893 From: ends in numbers
0.885 'Message-Id' was added by a relay (3)
0.882 Gives information about an opportunity

SpamAssassin Features

0.874 Don't delete me! Noooooo!!!!
0.863 Fast Viagra Delivery
0.853 Frequent SPAM content
0.849 exists:X-Stormpost-To
0.849 Missing Date: header
0.849 List removal information
0.838 Consolidate Debt and Credit
0.820 Financial Freedom
0.817 Lots and lots of Cc: headers
0.810 Received via a relay in multihop.dsbl.org
0.796 Contains word 'guarantee' in all-caps
0.795 Claims you can be removed from the list
0.781 Spam phrases score is 00 to 01 (low)
0.781 HTML message is a saved web page
0.781 Claims compliance with Senate Bill 1618
0.779 exists:X-PMFLAGS

SpamAssassin Features

0.676 See for yourself
0.673 You'd better read all of this spam!
0.670 Easy Terms
0.666 Contains "Toner Cartridge"
0.665 Human Growth Hormone
0.658 Trying to sell insurance online
0.653 No experience needed!
0.646 Claims to be legitimate email
0.643 Subject: starts with advertising tag
0.630 Frequent SPAM content
0.628 illegal Nigerian transactions (2)
0.622 Subject GUARANTEED
0.620 DNSBL: sender ip address in in a dialup block
0.614 Possible porn - Must be 18
0.612 Tells you to click on a URL (in caps)
0.612 Free Quote

SpamAssassin Features

0.611 Refinance Home
0.610 Received via a relay in relays.ordb.org
0.608 Contains 'free access' with capitals
0.606 Uses a long numeric IP address in URL
0.605 Have you been turned down?
0.601 Includes a URL link to send an email with the subject'remove'
0.601 No Credit Check
0.600 No Inventory
0.594 To: has a malformed address
0.573 Be your own boss
0.563 Information on how to work at home (2)
0.560 Contains mail-in order form
0.556 One hundred percent guaranteed
0.553 Guaranteed Stuff
0.552 Information on mortgage rates
0.549 Frequent SPAM content

SpamAssassin Features

0.544 From and To the same (1)
0.542 Bulk email software fingerprint (screwup 2) found inheaders
0.542 Gives an excuse for why message was sent
0.541 Avoid Bankruptcy
0.539 Includes a link for AOL users to click
0.536 Form for changing email address
0.531 Apply online (with capital O)
0.525 List removal information
0.521 Date: is 12 to 24 hours after Received: date
0.518 Asks you for your signature on a form
0.514 Subject talks about losing pounds
0.513 Lower Interest Rates
0.511 Do it Today
0.506 Unsecured Credit/Debt
0.506 The best Rates
0.505 From: starts with nums

SpamAssassin Features

0.505 Spam phrases score 55 or higher (high)
0.505 Impotence cure
0.503 Vacation Offers
0.503 Spam is 100% natural?!
0.501 Possible porn - Free Porn
0.501 Possible porn - Best, Largest Porn Collections
0.500 Spam phrases score is 01 to 02 (low)
0.496 Can not be combined with any other offer
0.489 Message contains disclaimer
0.488 Claims to be Legal
0.483 Subject is all capitals
0.466 MS-Outlook-style To "<Undisclosed-Recipient:;>"
0.466 Date: is 96 hours or more after Received: date
0.459 Spam tool pattern in MIME boundary
0.448 Date: is 6 to 12 hours before Received: date
0.448 Says: "to be removed, reply via email" or similar

SpamAssassin Features

- 0.448 Possible porn - Porn Fest
- 0.446 Sent with 'X-Priority' set to high
- 0.443 Local part containing a "4u" variant
- 0.443 HTML font color is magenta
- 0.435 Join Millions of Americans
- 0.434 Asks for a billing address
- 0.431 Nigerian scam key phrase ((dollar) NNN.Nm/USDNNN.N m/US(dollar) |
- 0.431 Claims "This is not spam"
- 0.429 Sent with 'X-Msmail-Priority' set to high
- 0.428 Subject contains "FREE" in CAPS
- 0.426 exists:X-MailingID
- 0.424 MIME section missing boundary
- 0.424 Asks you to fill out a form
- 0.422 HTML font color is unknown to us
- 0.422 Domain name containing a "4u" variant
- 0.421 HTML font color is yellow

SpamAssassin Features

- 0.419 Includes a link to send a mail with a subject
- 0.419 Standard investment opportunity spam
- 0.418 Javascript to hide URLs in browser
- 0.417 Offers Extra Cash
- 0.416 Eliminate Bad Credit
- 0.415 Lose Weight Spam
- 0.414 Subject talks about savings
- 0.414 Subject ends with lots of white space
- 0.414 Offers a full refund
- 0.414 Gives instructions for removal from list
- 0.413 Free Cell Phone
- 0.412 Frontpage used to create the message
- 0.411 Offers a limited time offer
- 0.410 Claims you can be removed from the list
- 0.408 Attempt at obfuscating the word "mortgage"
- 0.407 Opportunity - What a deal!

SpamAssassin Features

- 0.407 Nobody's perfect
- 0.406 Tells you about a strong buy
- 0.406 HTML table has thick border
- 0.406 Buy Direct
- 0.405 Instant Access button
- 0.405 HTML font color is green
- 0.405 HTML font color is cyan
- 0.405 Discusses money making
- 0.405 Asks you to click below (in caps)
- 0.404 Uses open redirection service
- 0.404 exists:X-ServerHost
- 0.404 Claims you can be removed from the list
- 0.403 List removal information
- 0.402 Message with extraneous Content-type:...type=header
- 0.402 There is no obligation.
- 0.402 Talks about lots of money

SpamAssassin Features

- 0.402 Contains 'Get it now' with capitals
- 0.401 Supplies are Limited
- 0.401 No such thing as a free lunch (2)
- 0.400 You won't be dissapointed.
- 0.400 Possible porn - Offers Instant Access
- 0.400 Nigerian scam key phrase ((dollar)NN,NNN,NNN.NN)
- 0.400 How dear can you be if you don't know my name?
- 0.386 No Strings Attached
- 0.382 HTML with embedded plugin object
- 0.380 Received via a relay in relays.osirusoft.com
- 0.369 Off Shore Scams
- 0.365 Information on how to work at home (1)
- 0.364 Possible porn - Hot, Nasty, Wild, Young
- 0.364 Contains word 'amazing' in all-caps
- 0.362 exists:X-SMTPExp-Version
- 0.362 There is no catch.

SpamAssassin Features

- 0.361 sent to you@you.com or similar
- 0.360 Received from first hop dialup listed inrelays.osirusoft.com
- 0.344 HTML font color is same as background
- 0.336 Subject: is empty or missing
- 0.335 FONT Size +2 and up or 3 and up
- 0.334 Lowest Price
- 0.333 HTML font color has unusual name
- 0.333 Contains word 'profits' in all-caps
- 0.330 HTML font color is gray
- 0.329 What are you waiting for
- 0.329 One Time Rip Off
- 0.327 Talks about prizes
- 0.327 Free Website
- 0.326 To: and Cc: contain similar usernames at least 5 times
- 0.325 HTML font face is not a commonly used face
- 0.324 Quoted-printable line longer than 76 characters

SpamAssassin Features

- 0.324 From: has a malformed address
- 0.323 exists:X-SMTPExp-Registration
- 0.323 Message-Id has no @ sign
- 0.323 No such thing as a free lunch (1)
- 0.321 URL of CGI script called "unsubscribe" or "remove"
- 0.321 Satisfaction Guaranteed
- 0.321 "if you do not wish to receive any more"
- 0.320 Message contains a lot of ^M characters
- 0.320 exists:x-esmtp
- 0.320 Claims you are a winner
- 0.319 From: contains numbers mixed in with letters
- 0.318 Can't live without?
- 0.317 HTML mail with non-white background
- 0.315 Talks about email marketing
- 0.315 Save big money
- 0.315 HTML font color is red

SpamAssassin Features

- 0.315 3 WHOLE LINES OF YELLING DETECTED
- 0.313 Save Up To
- 0.313 Domain registration spam body
- 0.312 Tells you to click on a URL
- 0.312 Subject: domain registration spam subject
- 0.308 URL contains spamhaus signature: numbered servers
- 0.308 Name Brand
- 0.307 Asks you to click below
- 0.307 Act Now! Don't Hesitate!
- 0.306 Talks about Hidden Charges
- 0.305 Message is 50-70% HTML tags
- 0.304 While Supplies Last
- 0.302 Easily-executed JavaScript code
- 0.302 Subject starts with "Free"
- 0.302 HTML font color not within safe 6x6x6 palette
- 0.301 No Purchase Necessary

SpamAssassin Features

- 0.301 Auto-executing JavaScript code
- 0.300 DNSBL: sender is a Spamware site or vendor
- 0.300 Significant Savings
- 0.300 No Fees
- 0.300 Click-to-remove with PHP/ASP action found
- 0.299 X-Mailer header indicates a non-spam MUA (TheBat!)
- 0.296 'remove' URL contains an email address
- 0.294 Being a Member
- 0.281 Investment Decision
- 0.279 Date: is 3 to 6 hours before Received: date
- 0.245 Contains a Privacy Statement
- 0.242 Tells you how to stop further spam
- 0.239 Month Trial Offer
- 0.229 Save (dollar) (dollar) (dollar)
- 0.224 Sign up Free Today
- 0.222 To: repeats address as real name

SpamAssassin Features

- 0.218 Congratulations - you've been scammed?
- 0.217 2 WHOLE LINES OF YELLING DETECTED
- 0.216 Weekend Getaway
- 0.214 Trying to offer you something
- 0.212 Member Stuff
- 0.212 HTML font color is missing hash (
- 0.212 Doesn't ask any questions
- 0.212 Contains 'Special Promotion'
- 0.212 A WHOLE LINE OF YELLING DETECTED
- 0.211 To: is empty
- 0.211 Winning in Caps
- 0.211 Stuff on Sale
- 0.211 Only (dollar) (dollar) (dollar)
- 0.211 Encourages you to waste no time in ordering
- 0.210 Who really wins?
- 0.210 HTML font face has excess capital characters

SpamAssassin Features

- 0.209 Free DVD
- 0.207 Date: is 12 to 24 hours before Received: date
- 0.207 JavaScript code
- 0.206 Header with all capitals found
- 0.205 HTML font color is blue
- 0.204 Winner in Caps
- 0.204 HTML font face is not a word
- 0.204 Fantastic Deal
- 0.203 Includes a 'remove' email address
- 0.203 Includes a URL link to send an email
- 0.203 Possible porn - Large Number of movies, pics
- 0.203 Free Offer
- 0.202 Contains a tollfree number
- 0.201 illegal Nigerian transactions (1)
- 0.201 Image tag with an ID code to identify you
- 0.201 Frame wanted to load outside URL

SpamAssassin Features

- 0.201 Contains 'for only' some amount of cash
- 0.181 X-Mailer header indicates a non-spam MUA(Outlook Express)
- 0.150 Spam tool pattern in MIME boundary
- 0.146 Cancel at any time!
- 0.144 Talks about social security numbers
- 0.137 Click to perform an action on an account
- 0.134 Gives an excuse about why you were sent this spam
- 0.127 Nigerian scam key phrase ((dollar) NNN.Nm/USDNNN.N m/US(dollar))
- 0.123 Contains a comment with nothing but unique ID
- 0.117 No Claim Forms
- 0.114 'Message-Id' was added by a relay (2)
- 0.114 Free Trial
- 0.111 They're just giving it away!
- 0.108 Message-Id has characters indicating spam
- 0.107 Dear you@you.com?
- 0.106 Free Hosting

SpamAssassin Features

- 0.105 Contains an ASCII-formatted form
- 0.104 I wonder how many emails they sent in error...
- 0.103 URL of page called "unsubscribe"
- 0.102 Subject has exclamation mark and question mark
- 0.101 Offer Expires
- 0.101 Contains 'Dear Somebody'
- 0.100 Javascript protocol in a URI
- 0.100 Message includes Microsoft executable program
- 0.100 MIME filename does not match content
- 0.100 Spam tool pattern in MIME boundary
- 0.038 'Received:' has 'may be forged' warning
- 0.032 Message-Id is not valid, according to RFC 2822
- 0.031 Offers Coupon
- 0.028 Please read this! Please oh please oh please!
- 0.014 Shopping Spree
- 0.009 Contains a line ≥ 199 characters long

SpamAssassin Features

- 0.009 Spam tool pattern in MIME boundary
- 0.009 Risk free. Suuurreeee....
- 0.008 Reserves the right
- 0.008 Expect to earn
- 0.005 Contains 'G.a.p.p.y-T.e.x.t'
- 0.004 Gift Certificate
- 0.003 Big Bucks
- 0.006 X-Mailer header indicates a non-spam MUA(Outlook)
- 0.019 From Majordomo
- 0.026 Missing From: header
- 0.069 Free money!
- 0.075 Forwarded email (Outlook style)
- 0.102 Email came from some known mailing list software
- 0.118 Mailer daemon failure notice (1)
- 0.123 Message text is over 40K in size
- 0.133 Came via Internet Mail Service plugin

SpamAssassin Features

- 0.137 Correct for MIME 'null block'
- 0.143 X-Mailer header indicates a non-spam MUA(Netscape)
- 0.196 Mailing list headers are suspicious
- 0.200 exists:Resent-To
- 0.207 exists:X-Authentication-Warning
- 0.211 Where are you working at?
- 0.215 exists:X-Accept-Language
- 0.217 Subject contains newsletter header (list)
- 0.231 'Message-Id' was added by yahoo.com, that's OK
- 0.233 exists:X-Loop
- 0.240 X-Mailer header indicates a non-spam MUA (AOL)
- 0.298 To: repeats local-part as real name
- 0.300 User-Agent header indicates a non-spam MUA(Entourage)
- 0.301 Short signature present (no empty lines)
- 0.302 exists:X-Mailing-List
- 0.304 Long signature present (empty lines)

SpamAssassin Features

- 0.484 Subject contains a month name - probable newsletter(2)
- 0.484 Subject contains a month name - probable newsletter
- 0.489 Common footer for Hotmail
- 0.506 Contains a PGP-signed message
- 0.506 Appears to be from yahoo groups
- 0.506 Yahoo! Groups message
- 0.518 exists:User-Agent
- 0.522 Has a valid-looking References header
- 0.558 Forwarded email
- 0.601 User-Agent header indicates a non-spam MUA(Mozilla)
- 0.605 User-Agent header indicates a non-spam MUA(Outlook Express)
- 0.616 Subject contains newsletter header (news)
- 0.641 Message-Id indicates a non-spam MUA (Pine)
- 0.695 Contains what looks like an 'E-Mail Disclaimer'
- 0.708 Contains a PGP-signed message (signature attached)
- 0.708 Message text is over 20K in size

SpamAssassin Features

- 0.725 Subject contains a frequency - probable newsletter
- 0.754 X-Mailer header indicates a non-spam MUA(T-Offline)
- 0.832 Contains what looks like a quoted email text
- 0.847 exists:In-Reply-To
- 0.864 Has an Approved-By moderated list header
- 0.897 User-Agent header indicates a non-spam MUA(IMP)
- 0.949 Contains what looks like a patch from diff -u
- 0.986 Mailer daemon failure notice (2)
- 1 X-Mailer header indicates a non-spam MUA (Gnus)
- 1 User-Agent header indicates a non-spam MUA(Gnus)
- 1 Subject contains newsletter header (in review)
- 1 From: looks like US Telephone Number
- 1 recommended page from MailBits.com
- 1 Talks about tracking numbers
- 1 Common footer for MSN
- 1 A MailMan confirm-your-address message

SpamAssassin Features

- 1.118 Common footer for MSN
- 1.128 Contains a password retrieval system
- 1.152 Something about registration
- 1.176 User-Agent header indicates a non-spam MUA(Mutt)
- 1.301 Came from MSN Communities
- 1.334 exists:X-Cron-Env
- 1.433 Subject looks like order info
- 1.451 From the Mailer-Daemon
- 1.596 Subject contains a date
- 1.628 Contains what looks like an email attribution
- 1.696 Common footer for Hotmail
- 1.780 X-Mailer header indicates a non-spam MUA (AppleMail)
- 1.801 Common footer for Hotmail
- 1.898 Sent through Microsoft's ListBuilder service
- 2.092 Short signature present (empty lines)
- 2.170 Common footer for Hotmail

SpamAssassin Features

- 2.174 Message from eBay
- 2.442 Contains what looks like a patch from diff -c
- 2.473 Looks like a Debian BTS bug
- 2.475 Common footer for Hotmail
- 2.550 Subject is an eBay question
- 2.699 Looks like a Bugzilla bug
- 2.863 User-Agent header indicates a non-spam MUA(KMail)
- 3.052 non-spam Yahoo! Groups banner found
- 3.127 Long signature present (no empty lines)
- 4.0 Uses the Habeas warrant mark(<http://www.habeas.com/>)
- 6 User is listed in 'whitelist_to'
- 10 Not Matt's Scripts formmail.pl
- 10 Bonded sender, see<http://www.bondedsender.org/referred.html>
- 20 User is listed in 'more_spam_to'
- 100 User is listed in 'all_spam_to'
- 100 From: address is in the user's white-list

How to Categorize? (unsupervised)

What if we don't have supervised training data?

Might try an iterative approach as usual:

1. Cluster the messages
2. Train n-gram, Naive Bayes, or decision list model to discriminate among the clusters
3. Use the model to reassign messages to clusters (most will stay put but some will move)
4. Return to step 2 until convergence

How to Categorize? (semisupervised)

What if we have only a little supervised data?

Could try bootstrapping like Yarowsky's WSD:

1. Start with very small, rather accurate classes
2. Train n-gram, Naive Bayes, or decision list model to discriminate among the classes
3. Augment each class with new messages that the model **confidently** classifies there (maybe also move or remove some existing messages)
4. Return to step 2 until convergence

How to Categorize? (adaptive)

What if we gradually get more new data over time?

- User feedback (active or passive) on our classifications
- News / email systems that categorize, or judge relevance
 - Add new articles / messages to training data
 - If they're unlabeled (no supervision), label them automatically
 - Add them only if we're confident? Add them fractionally, like EM?

So model adjusts over time:

- E.g., change the cluster centroids or n-gram parameters
- May want to **weight the more recent data more heavily**, since the future is more like the present than the past
 - E.g., message from k days ago has weight 0.9^k ($k=0,1,2, \dots$)
 - So today's model = today's data + $0.9 * \text{yesterday's model}$

How to Categorize? (hierarchical)

What if we are putting document in a Yahoo! category?

- There are thousands of categories (at least) – too hard!
- Choose one of the 14 top-level categories, e.g., Science
- Then use a Science-specific classifier to choose one of the 54 second-level categories within Science (14 are symlinks)
- Continue working your way down the tree ...
- When you can't classify with high confidence, ask a human (then use the human's answer as more training data)