# Bayes' Law and Bayesian inference

# Announcements

- Don’t forget to reach out to your team members for A1.
- Walkthrough A0 posted on Canvas

# Conditional probabilities

- Probability that one event occurs, given that another event is known to have occurred
  - Denoted $P(B|A)$ . "Probability of B given A"
  - Defined as:

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$

- Leads directly to the *chain rule*:

$$P(A \cap B) = P(B|A)P(A)$$

  - More generally:

$$P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)...P(A_n|A_1 \cap ... \cap A_{n-1})$$

# Independence of events

- Two events are *independent* if $P(A|B) = P(A)$
  - Or, equivalently, if $P(B|A) = P(B)$
  - Independence denoted $A \perp B$

- The joint probability of independent events A and B both occurring is then simply:

$$P(A \cap B) = P(A)P(B)$$

  - This idea of *factoring* a distribution into a product of two simpler distributions will be a recurring theme!

# Conditional independence

- Sometimes events are both conditioned on the same event, but otherwise are independent
  - Mary and Bob live in same city but independently
  - A denotes event that it's raining, B denotes event that Mary has an umbrella, C denotes event that Bob has an umbrella
  - Events B and C are **not** independent
  - But B and C are **conditionally independent** given A,

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

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

  - Denoted

$$B \perp C|A$$

# Bayes' Law

- For two events *A* and *B*,

Likelihood

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Priors

Posterior

- Useful when you want to know something about A, but all you can directly observe is B
  - This process is called *Bayesian inference*

# Bayes’ Law Example #1

- I have two coins, one fair and one with heads on both sides. I choose a coin at random, flip it, and get heads. What is the probability that it is the fair coin?

# Bayes' Law Example #1

- I have two coins, one fair and one with heads on both sides. I choose a coin at random, flip it, and get heads. What is the probability that it is the fair coin?

# Bayes' Law Example #1

- I have two coins, one fair and one with heads on both sides. I choose a coin at random, flip it, and get heads. What is the probability that it is the fair coin?

```
F = Fair
H = Heads
```

$$P(F|H) = \frac{P(H|F)P(F)}{P(H)} = \frac{\frac{1}{2}\frac{1}{2}}{\frac{3}{4}} = \frac{1}{3}$$

$$P(H|F) = \frac{1}{2}$$

$$P(F) = \frac{1}{2}$$

$$\begin{aligned} P(H) &= P(H \cap F) + P(H \cap \bar{F}) \\ &= P(H|F)P(F) + P(H|\bar{F})P(\bar{F}) \\ &= \frac{1}{2}\frac{1}{2} + 1\frac{1}{2} = \frac{3}{4} \end{aligned}$$

# Bayes’ Law Example #2

- A doctor says you have an illness that afflicts 0.01% of the population. Her diagnoses are right 99% of the time. What’s the probability that you have the illness?

# Bayes' Law Example #2

- A doctor says you have an illness that afflicts 0.01% of the population. Her diagnoses are right 99% of the time. What's the probability that you have the illness?

```
i = Ill
p = Positive
```

$$P(i|p) = \frac{P(p|i)P(i)}{P(p)} = \frac{(0.99)(0.0001)}{0.010098} \approx 0.0098$$

$$P(p|i) = 0.99$$

$$P(i) = 0.0001$$

$$\begin{aligned} P(p) &= P(p \cap i) + P(p \cap \bar{i}) \\ &= P(p|i)P(i) + P(p|\bar{i})P(\bar{i}) \\ &= (0.99)(0.0001) + (0.01)(0.9999) \approx 0.010098 \end{aligned}$$

# Bayes’ Law Example #3a

- Suppose that you have a friend, Mary, who lives in Seattle. She tells the truth 80% of the time and lies the other 20% of the time. The weather in Seattle on any given day is either sunny or cloudy; 30% of days are sunny and 70% are cloudy.

- Mary calls you one day and says that the weather in Seattle is cloudy. What is the probability that the weather is actually cloudy?

# Solution #3a

C: cloudy

S: sunny

Mc: Marry says cloudy

Ms: Marry says sunny

We are interested in finding the probability that it is actually cloudy given that Mary says it is cloudy, P (C|Mc). By Bayes' Law:

$$P(C|Mc) = \frac{P(Mc|C)P(C)}{P(Mc)}$$

We know that P (C) = 0.7 from the problem statement.

The probability that Mary says it is cloudy given that it actually is cloudy is just the probability that she tells the truth, P (Mc|C) = 0.8.

The probability that Mary says cloudy is given by,

$$P(Mc) = P(Mc|C)P(C) + P(Mc|S)P(S) = (0.8)(0.7) + (0.2)(0.3) = 0.62$$

$$P(C|Mc) = \frac{(0.8)(0.7)}{0.62} = 0.903$$

# Making decisions with Bayes' Law

# Applications to AI

- Perhaps the simplest application of probabilistic methods to AI is the Bayesian Classifier
  - Want to make a decision based on some evidence
  - “Train” the classifier by computing prior probability distributions and likelihood functions from training data
  - Then apply Bayes law to compute posterior probabilities

# Bayes' Law: An example

- You're a juror in a murder case
  - You need to decide between guilt (G) and innocence ($\bar{G}$)
  - You have heard some evidence (E)
  - Bayesian approach: Compute *P(G|E),* and vote to convict if

$$P(G|E) > \tau$$

    where $\tau$ is a threshold

  - Or, compute an odds ratio and convict if:

$$\frac{P(G \mid E)}{P(\bar{G} \mid E)} > \tau_2$$

# Bayes' Law: An example

- Say you have to decide before hearing any evidence
  - what is the *prior* probability, $P(G)$?
- How to estimate $P(G)$?
  - Based on population constraints
    - 1 person in Bloomington (~20,000 people) did it
    - $P(G) \approx 1/20000 = 0.00005$
  - Based on historical data
    - U.S. murder conviction rate: 0.06/1000 [BJS96]
    - $P(G) \approx 0.00006$

# Bayesian inference example

- Eyewitness testimony (T) identifies the suspect
  - Now we want to compute $P(G \mid T) = \dfrac{P(T \mid G)P(G)}{P(T)}$
  - P(G) =
  - P(T|G) =

    P(T|$\bar{\text{G}}$) =

$$P(T) = P(T \mid G)P(G) + P(T \mid \bar{G})P(\bar{G})$$
$$= (0.55)(0.00005) + (0.32)(0.99995)$$
$$\approx 0.32001$$
$$P(G \mid T) \approx \frac{(0.55)(0.00005)}{0.32001} \approx 0.0000859$$

- Alternatively, we can compute an *odds ratio:*

$$\frac{P(G \mid T)}{P(\bar{G} \mid T)} = \frac{P(T \mid G)P(G)P(T)}{P(T)P(T \mid \bar{G})P(\bar{G})} = \frac{P(T \mid G)P(G)}{P(T \mid \bar{G})P(\bar{G})} = \frac{(0.55)(0.00005)}{(0.32)(0.99995)} \approx 1:11635$$

# Bayesian inference example (2)

- Now you hear that the murderer had a red car, and that the suspect owns a red car (R)
  - We want to compute $P(G \mid T,R) = \dfrac{P(T,R \mid G)P(G)}{P(T,R)}$. How?
  - Assuming that T and R are independent conditioned on G,

$$P(G \mid T,R) = \frac{P(R \mid G)P(T \mid G)P(G)}{P(R \mid G)P(T \mid G)P(G) + P(R \mid \bar{G})P(T \mid \bar{G})P(\bar{G})}$$

  - Computing odds:

$$\frac{P(G \mid T,R)}{P(\bar{G} \mid T,R)} = \left(\frac{P(G)}{P(\bar{G})}\right)\left(\frac{P(T \mid G)}{P(T \mid \bar{G})}\right)\left(\frac{P(R \mid G)}{P(R \mid \bar{G})}\right)$$

The *posterior* odds

Prior odds

New evidence

# Bayesian inference example (3)

- Given the testimony (T) and red car evidence (R),

$$\frac{P(G \mid T,R)}{P(\bar{G} \mid T,R)} = \left(\frac{P(G)}{P(\bar{G})}\right)\left(\frac{P(T \mid G)}{P(T \mid \bar{G})}\right)\left(\frac{P(R \mid G)}{P(R \mid \bar{G})}\right)$$

  - $P(R|G) = 1$, $P(R|\bar{G}) \approx 0.13$ [DuPont07]

$$\frac{P(G \mid T,R)}{P(\bar{G} \mid T,R)} = \left(\frac{1}{11635}\right)\left(\frac{1}{0.13}\right) \approx 1:1513$$

# Bayesian inference example (4)

- Now suppose a partial fingerprint (F) matches the suspect
  - $P(F|G) = 1$, $P(F|\bar{G})=0.001$

$$\frac{P(G \mid T,R,F)}{P(\bar{G} \mid T,R,F)} = \left(\frac{P(G \mid T,R)}{P(\bar{G} \mid T,R)}\right)\frac{P(F \mid G)}{P(F \mid \bar{G})} \approx \left(\frac{1}{1513}\right)\frac{1}{0.001} = \frac{1}{1.513}$$

# Probabilistic inference

- Notice that we avoided making hard classification decisions until all evidence had been considered
  - Very useful when evidence is weak, noisy, imprecise

# An actual application:

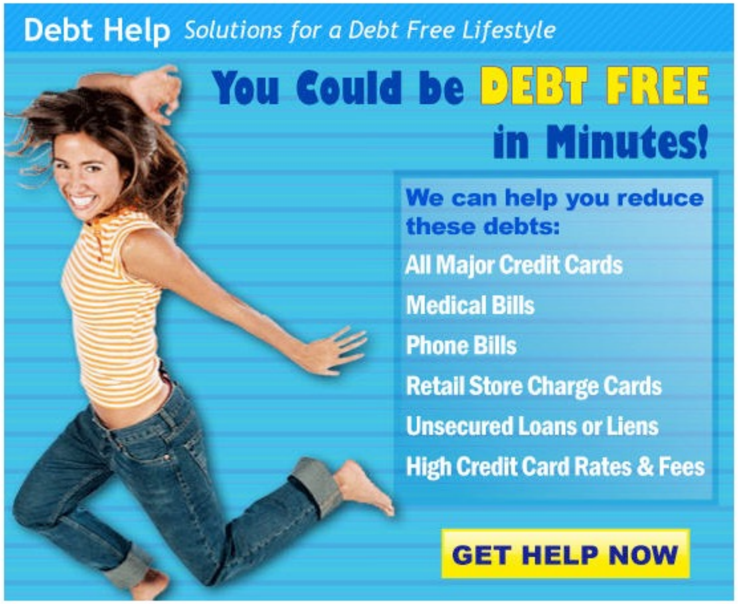
# Spam Filtering

# Spam

- Spam = junk e-mail
- A big problem!
  - ~50% of all email traffic on the Internet
  - ~320 billion junk emails per day
  - >2 petabytes (= 2,000 terabytes = 2,000,000 gigabytes) daily
  - Spreads malware, worms, phishing schemes, etc.
- Possible solutions
  - Block e-mails from blacklisted users and servers
  - Accept e-mails only from whitelisted addresses
  - Cost-based solutions (e.g. micropayments)
  - Filtering rules (ignore mail with “debt”, “viagra”, “stock”)
  - Content-based statistical filtering

**Reduce Debt by up to 60 Percent** Spam | X

from **Financial Assistance** <debtrelief@airving.com> hide details 8:30 PM (3
to david.crandall@gmail.com
date Wed, Sep 10, 2008 at 8:30 PM
subject Reduce Debt by up to 60 Percent




**This is a service of Debt Help. If you no longer wish to receive these announements, Click Here.**
**3409 Executive Center Dr., Suite 110, Austin, TX 78731**

You could be debt free in minutes!

We can help you reduce these debts:
-All major credit cards
-Medical bills
-Phone Bills
-Retail Store Charge Cards
-Unsecured loans or liens
-High credit card rates and fees

Get help from one of our professional consultants today, and get your life back on track:
http://kperducr.airving.com/debt/

This is a service of Debt Help. If you no longer wish to receieve these announements, go here http://kperducr.airving.com/debt/r/ or write to 3409 Executive Center Dr, Suite 110 Austin TX 78731

**Quality watches at 25% discount** Spam | X

from **Exceptional Watches** <info@fliesen-becht.de> hide details 5:58 AM (17 hours ago) Reply
to Great watch Service <david.crandall@gmail.com>
date Wed, Sep 10, 2008 at 5:58 AM
subject Quality watches at 25% discount

**Why would you want to purchase a replica watch from King-replicas?**

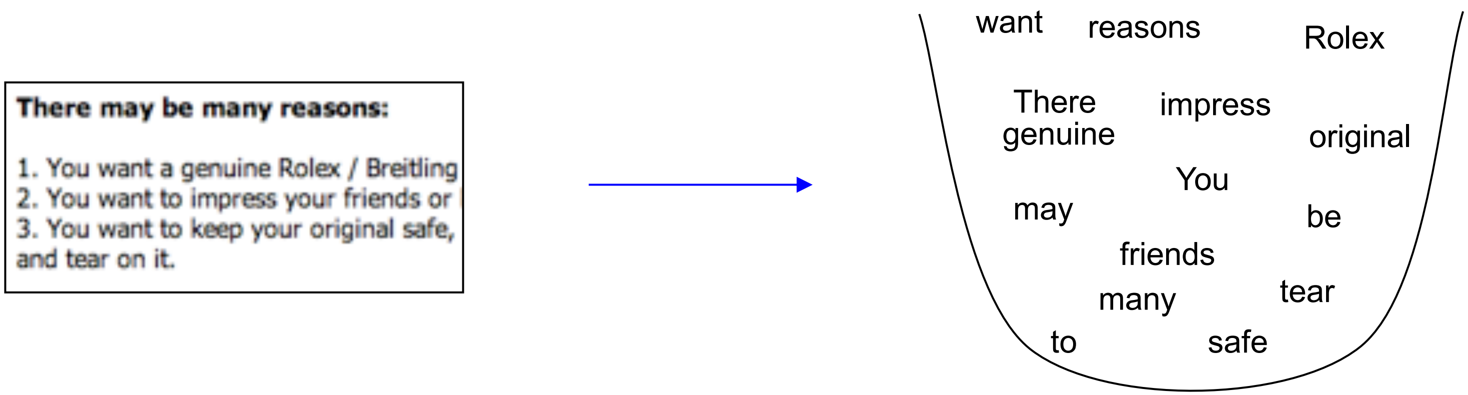
**There may be many reasons:**

1. You want a genuine Rolex / Breitling watch, but the price is too ridiculous
2. You want to impress your friends or business clients
3. You want to keep your original safe, while using the replica for daily wear and tear on it.

**Browse our King-replica watches shop!**

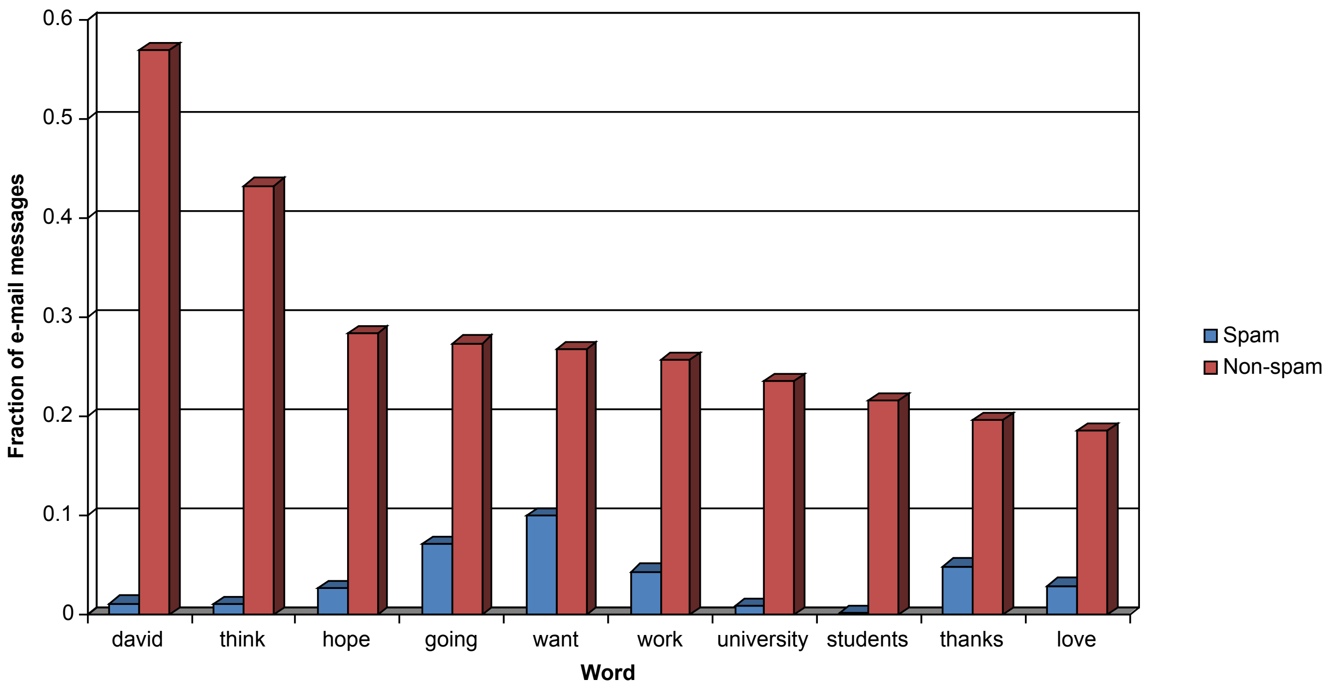
# Modeling a document

– Represent a document as an unordered collection of words (a *bag of words* model)



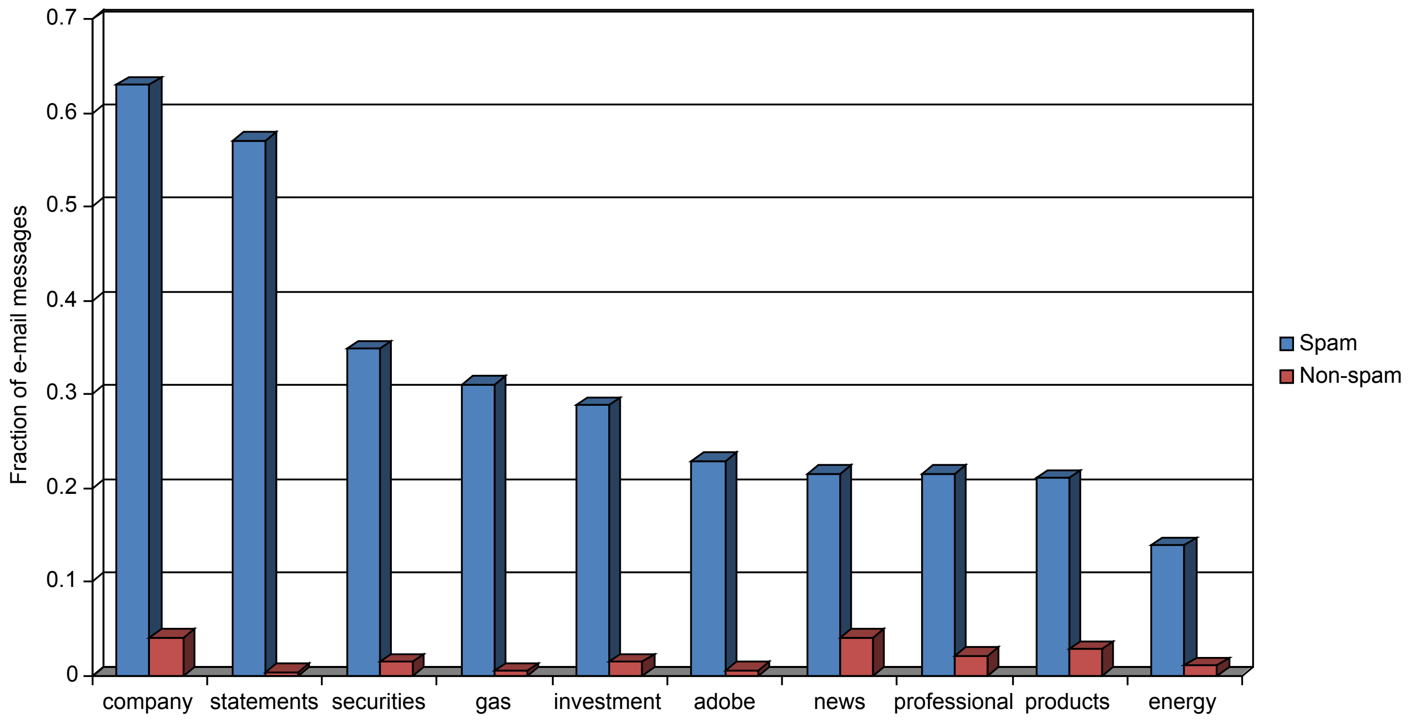
# Statistical motivation

- Spam and (my) non-spam are statistically very different



# Statistical motivation

- Spam and (my) non-spam are statistically very different



# Bayesian spam filtering

- Suppose we get an email containing the word "debt"
  - What is the probability it is spam (S), P(S|debt)?

$$P(S \mid \text{debt}) = \frac{P(\text{debt} \mid S)P(\text{S})}{P(\text{debt})}$$

  - P(debt | S) = 0.309 , P(debt | $\bar{S}$) = 0.00447
  - P(S) = 0.5
  - P(debt) = 0.157

$$P(S \mid \text{debt}) = \frac{P(\text{debt} \mid S)P(\text{S})}{P(\text{debt})} \approx 0.986$$

0.986/0.014 ≈ 70:1 odds that message is spam

$$P(\bar{S} \mid \text{debt}) = 1 - P(S \mid \text{debt}) \approx 0.014$$

# More examples

– Assuming a uniform prior, *P(S)*=0.5

| Word | P(word\|spam) | P(word\|not spam) | P(word) | P(spam\|word) |
|---|---|---|---|---|
| debt | 0.309 | 0.00447 | 0.157 | 0.986 |
| news | 0.215 | 0.0395 | 0.127 | 0.845 |
| investment | 0.288 | 0.0137 | 0.151 | 0.955 |
| david | 0.012 | 0.575 | 0.294 | 0.020 |
| want | 0.101 | 0.268 | 0.185 | 0.274 |
| thanks | 0.0491 | 0.196 | 0.123 | 0.200 |

Computed from Bayes' Law

# Bayesian spam filtering

- A new email has the words “debt” and “price”
  - What is the probability it is spam (S), P(S|debt, price)?

$$P(S \mid \text{debt}, \text{price}) \quad \propto \quad P(\text{debt}, \text{price} \mid S)P(S)$$

  - If we assume that the occurrence of the words “debt” and “price” are independent events *conditioned on S,*

$$P(S \mid \text{debt}, \text{price}) \quad \propto \quad P(\text{debt} \mid S)P(\text{price} \mid S)P(S)$$

  - This is called the *naïve Bayes* assumption.

# Bayesian spam filtering

- Generalize to an arbitrary number of words,

$$P(S \mid \mathrm{W}_1, \mathrm{W}_2, \mathrm{W}_3, \ldots, \mathrm{W}_n) \propto P(W_1 \mid S) P(W_2 \mid S) \ldots P(W_n \mid S) P(S)$$

which is equivalent to,

$$P(S \mid \bigcap_{i=1}^{n} W_i) \propto P(S) \prod_{i=1}^{n} P(W_i \mid S)$$

- For example,

$$P(S \mid \text{debt, free, credit}) \propto P(S) P(\text{debt} \mid S) P(\text{free} \mid S) P(\text{credit} \mid S)$$

# A practical spam filter [Graham02]

- Break a message into *tokens* of words, numbers, etc.
- Look for the 15 “most interesting words”
  - I.e. words for which P(S|W) is farthest from 0.5
  - Then compute $P(S|W_1, W_2, \ldots, W_{15})$

```
Dear Sir or Madam:

Please reply to
Receiver: China Enterprise Management Co., Ltd. (CMC)
E-mail: unido@chinatop.net

As one technical organization supported by China Investment and
Technical Promotion Office of United Nation Industry Development
Organization (UNIDO), we cooperate closely with the relevant Chinese
Quality Supervision and Standardization Information Organization. We
provide the most valuable consulting services to help you to open
Chinese market within the shortest time:

1. Consulting Service on Mandatory National Standards of The People's
Republic of China.

2. Consulting Service on Inspection and Quarantine Standards of The
People's Republic of China.

3. Consulting Service for Permission to Enter Chinese Market

We are very sorry to disturb you!

More information, please check our World Wide Web:
http://www.chinatop.net

Sincerely yours
```

| | |
|---|---|
| madam | 0.99 |
| promotion | 0.99 |
| republic | 0.99 |
| shortest | 0.047225013 |
| mandatory | 0.047225013 |
| standardization | 0.07347802 |
| sorry | 0.08221981 |
| supported | 0.09019077 |
| people's | 0.09019077 |
| enter | 0.9075001 |
| quality | 0.8921298 |
| organization | 0.12454646 |
| investment | 0.8568143 |
| very | 0.14758544 |
| valuable | 0.82347786 |

$P(S|W_1, W_2, \ldots, W_{15})=0.9$

# A true negative

```
Hi,

Do you have any examples online of that continuation style
web programming that you describe?

For example, you mention that you needed the user to go
to a color picker screen and then return to the
same spot. I'm interested on what was required to
achieve that. Did you have to use real continuations
to achieve that?


Dru Nelson
San Carlos, California
```

| | |
|---|---|
| continuation | 0.01 |
| describe | 0.01 |
| continuations | 0.01 |
| example | 0.033600237 |
| programming | 0.05214485 |
| i'm | 0.055427782 |
| examples | 0.07972858 |
| color | 0.9189189 |
| localhost | 0.09883721 |
| hi | 0.116539136 |
| california | 0.84421706 |
| same | 0.15981844 |
| spot | 0.1654587 |
| us-ascii | 0.16804294 |
| what | 0.19212411 |

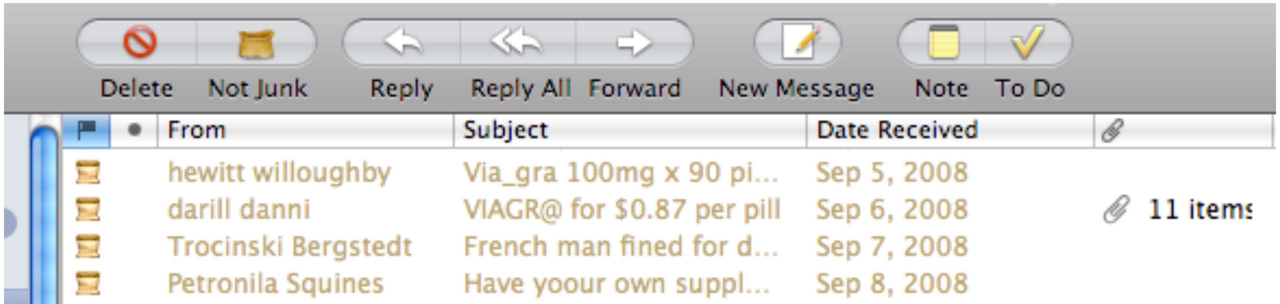
# A false negative

```
Dear Paul Graham,

As a person involved in website development you recognize that finding a
good web host offering reasonably priced quality services can be quite
difficult. We believe that offering a combination of quality services,
timely and responsive customer support and good prices is where our company
excels.

HOSTEX ( http://www.hostex.com ) is a flexible company providing quality and
cost effective web hosting solutions with focus on small and medium businesses.

Recently we introduced new service plans offering flexibility, performance and value:

Our plans are:

Basic    - $6.95 / month
         - 40 Mb of diskspace
         - 10 POP3 mailboxes, unlimited aliases, autoresponders
         - unlimited ftp accounts, URL protection, shell access and more
         - no setup fee

Advanced - $14.95 / month
         - 80 Mb of diskspace
         - 15 POP3 mailboxes, unlimited aliases, autoresponders
         - unlimited ftp accounts, URL protection, shell access
         - CGI/SSI scripting (Perl, Python, Tcl)
         - PHP4 scripting
         - Frontpage 2002 extensions
         - Graphical web statistics
         - various optional services and more
         - no setup fee

Professional    - customizable, make-your-own plan
                - select only those services that you need

All plans include full 30-day money back guarantee and there is no setup fee.
```

| | |
|---|---|
| perl | 0.01 |
| python | 0.01 |
| tcl | 0.01 |
| scripting | 0.01 |
| morris | 0.01 |
| graham | 0.01491078 |
| guarantee | 0.9762507 |
| cgi | 0.9734398 |
| paul | 0.027040077 |
| quite | 0.030676773 |
| pop3 | 0.042199217 |
| various | 0.06080265 |
| prices | 0.9359873 |
| managed | 0.06451222 |
| difficult | 0.071706355 |

# Learning

- The advantage of a Bayesian classifier is that it can learn optimal values for its parameters
  - Given a set of training data
  - No need for hand-crafted rules. More accurate, less work.
  - But a good set of training data is critical
- The classifier can be continue to learn with time
  - User corrects the classifier's errors, classifier adjusts probabilities accordingly



# Naïve Bayes

- Our framework assumes that evidence is independent, given a class label, e.g.

$$P(\text{debt}, \text{loan}, \text{stock}|\text{S}) = P(S)P(\text{debt}|\text{S})P(\text{loan}|\text{S})P(\text{stock}|\text{S})$$

  – This is not generally true
  – It is very popular anyway and often works surprisingly well

# Next class

- **More on Inference and Bayes Nets**