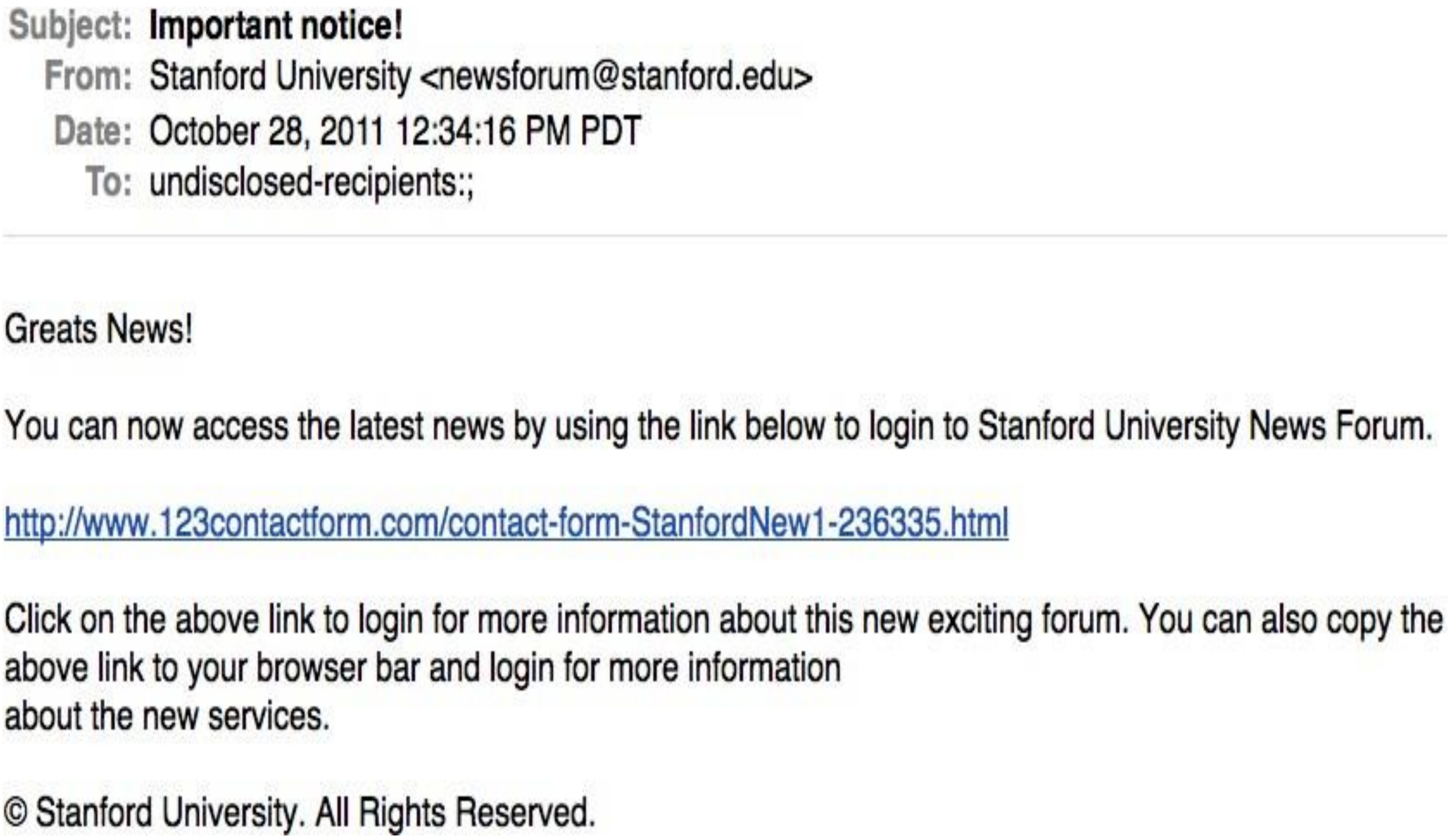
Text Classification

Is this spam?



Subject: **Important notice!**
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

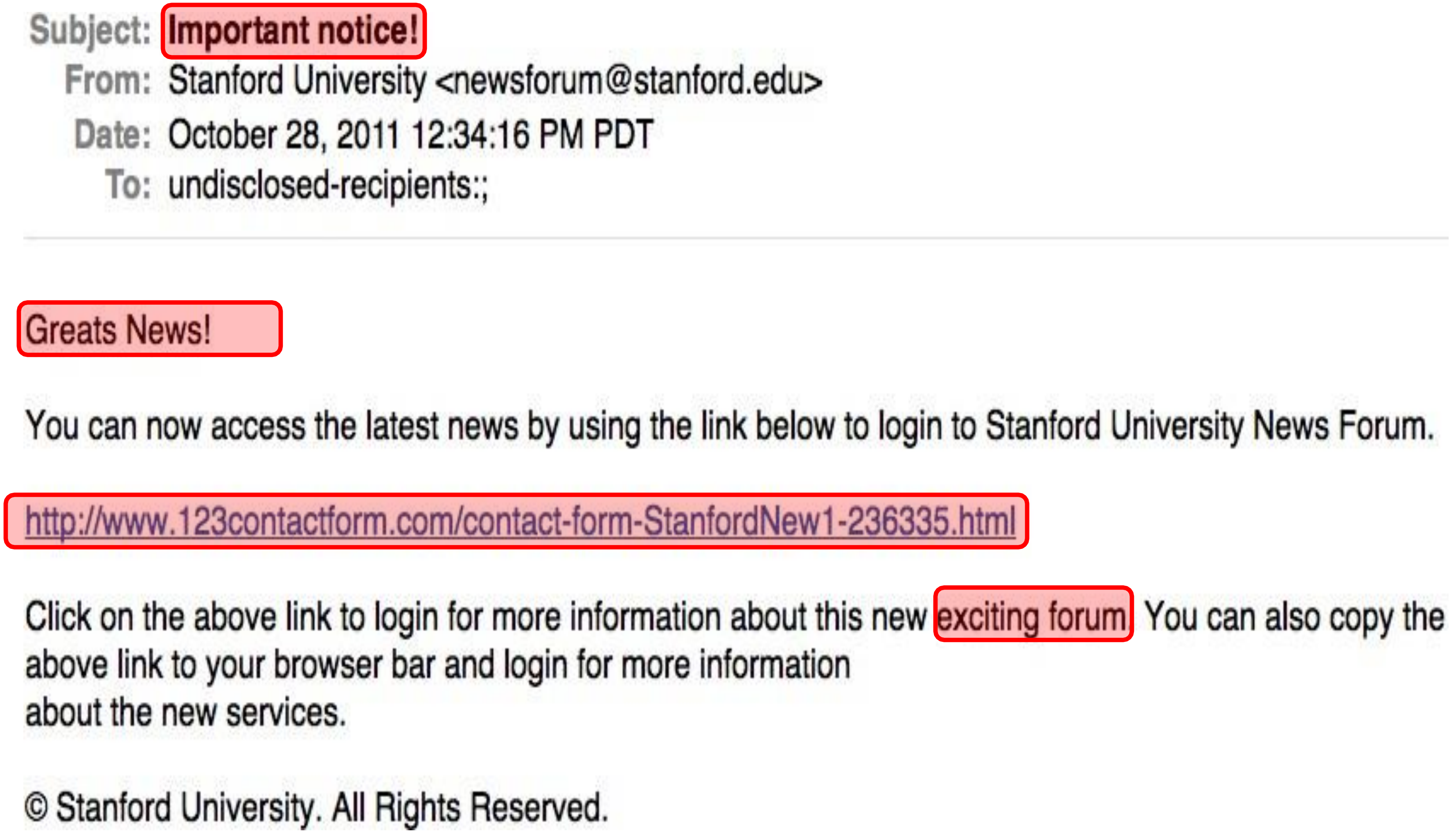
You can now access the latest news by using the link below to login to Stanford University News Forum.

http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

© Stanford University. All Rights Reserved.

Is this spam?



Subject: Important notice!
From: Stanford University <newsforum@stanford.edu>
Date: October 28, 2011 12:34:16 PM PDT
To: undisclosed-recipients:;

Greats News!

You can now access the latest news by using the link below to login to Stanford University News Forum.

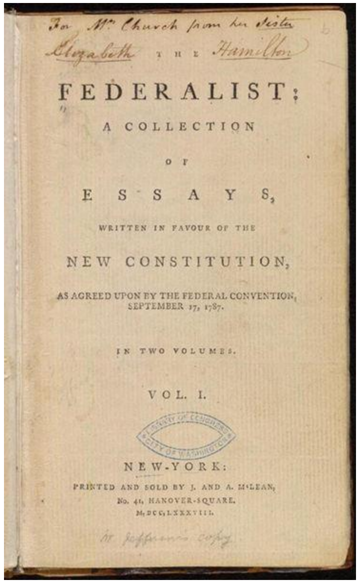
http://www.123contactform.com/contact-form-StanfordNew1-236335.html

Click on the above link to login for more information about this new exciting forum. You can also copy the above link to your browser bar and login for more information
about the new services.

© Stanford University. All Rights Reserved.

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



For Mrs Church from her Sister
Elizabeth THE Hamilton

FEDERALIST:

A COLLECTION

OF

ESSAYS,

WRITTEN IN FAVOUR OF THE

NEW CONSTITUTION,

AS AGREED UPON BY THE FEDERAL CONVENTION,
SEPTEMBER 17, 1787.

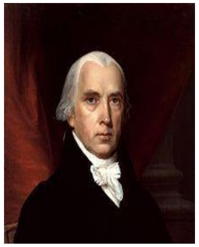
IN TWO VOLUMES.

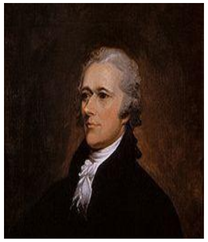
VOL. I.

LIBRARY OF CONGRESS
CITY OF WASHINGTON

NEW-YORK:
PRINTED AND SOLD BY J. AND A. McLEAN,
No. 41, HANOVER-SQUARE.
M,DCC,LXXXVIII.

Mr Jefferson's copy





Male or female author?

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. "Gender, Genre, and Writing Style in Formal Written Texts," Text, volume 23, number 3, pp. 321–346

Male or female author?

More determiners: Male

1. By 1925 present-day Vietnam was divided into three parts under French colonial rule. The southern region embracing Saigon and the Mekong delta was the colony of Cochin-China; the central area with its imperial capital at Hue was the protectorate of Annam…
2. Clara never failed to be astonished by the extraordinary felicity of her own name. She found it hard to trust herself to the mercy of fate, which had managed over the years to convert her greatest shame into one of her greatest assets…

S. Argamon, M. Koppel, J. Fine, A. R. Shimoni, 2003. “Gender, Genre, and Writing Style in Formal Written Texts,” Text, volume 23, number 3, pp. 321–346

More pronouns: Female

Positive or negative movie review?

- 👍 *...zany characters and richly applied satire, and some great plot twists*
- 👎 *It was pathetic. The worst part about it was the boxing scenes...*
- 👍 *...awesome caramel sauce and sweet toasty almonds. I love this place!*
- 👎 *...awful pizza and ridiculously overpriced...*

Text Classification

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

Text Classification: definition

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

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

Classification Methods: Hand-coded rules

- Rules based on combinations of words or other features
  - spam: black-list-address OR (“dollars” AND “have been selected”)
- Accuracy can be high
  - If rules carefully refined by expert
- But building and maintaining these rules is expensive

Classification Methods: Supervised Machine Learning

- *Input:*
  - a document d
  - a fixed set of classes $C = \{c_1, c_2, \ldots, c_J\}$
  - A training set of m hand-labeled documents $(d_1,c_1),\ldots,(d_m,c_m)$
- *Output:*
  - a learned classifier $\gamma: d \rightarrow c$

Classification Methods: Supervised Machine Learning

- Any kind of classifier
  - Naïve Bayes
  - Random forests
  - Logistic regression
  - Perceptrons
- **Generative classifiers** Generative classifiers model the joint distribution of features and labels, estimating both $P(X|Y)$ and $P(Y)$, and then use Bayes' theorem to compute posterior probabilities.
- like Naïve Bayes build a model for each class.
- **Discriminative classifiers** Rather than modeling the joint distribution of features and labels, discriminative classifiers focus on learning the decision boundary between classes based on the observed features.
- They aim to find the function that separates the feature space into regions corresponding to different classes. During training, the classifier learns to distinguish between spam and non-spam emails by finding the optimal decision boundary in the feature space.

Naïve Bayes (I)

Naïve Bayes Intuition

- Classification method based on Bayes rule
- A simple (naïve) assumption about how the features interact
- Relies on a 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!

fairy
always
love
to
it
it
whimsical
it
I
and
seen
are
anyone
friend
happy
dialogue
adventure
recommend
who
sweet
of
satirical
it
movie
it
I
but
to
romantic
I
several
yet
again
it
the
humor
the
seen
would
to
scenes
I
the
manages
fun
the
times
I
and
and
about
while
whenever
have
conventions
with

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
…	…

Formalizing the Naïve Bayes Classifier

Bayes' Rule Applied to Documents and Classes

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

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

Naïve Bayes Classifier (I)

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

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

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

Likelihood

Prior

Dropping the denominator

Naïve Bayes Classifier (II)

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

$$= \underset{c \in C}{\operatorname{argmax}} P(f_1, f_2, \ldots, f_n \mid c) P(c)$$

Document d represented as features $f_1..f_n$

Naïve Bayes Classifier (IV)

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(f_1, f_2, \ldots, f_n | c) P(c)$$

- Hard to compute directly
- Every possible set of words and positions
- Could only be estimated if a very, very large number of training examples was available.
- Naïve Bayes makes two simplifying assumptions

Multinomial Naïve Bayes Independence Assumptions

$$P(f_1, f_2, \ldots, f_n \mid c)$$

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

$$P(f_1,\ldots, f_n \mid c) = P(f_1 \mid c) \bullet P(f_2 \mid c) \bullet P(f_3 \mid c) \bullet \ldots \bullet P(f_n \mid c)$$

Multinomial Naïve Bayes Classifier

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(f_1, f_2, \dots, f_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c) \prod_{f \in F} P(f \mid c)$$

Applying Multinomial Naive Bayes Classifiers to Text Classification

positions ← all word positions in test document

$$c_{NB} = \operatorname*{argmax}_{c \in C} P(c) \prod_{i \in positions} P(w_i \mid c)$$

Naïve Bayes: Learning

Learning the Multinomial Naïve Bayes Model

- First attempt: maximum likelihood estimates
  - simply use the frequencies in the data
  - fraction of documents in each class

$$\hat{P}(c) = \frac{N_c}{N_{doc}}$$

  - Assume a feature is just the existence of a word in the document’s bag of words
  - the fraction of times the word w_i appears in all documents of topic *c*

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

  - V consists of the union of all the word types in all classes, not just the words in one class *c*.

Parameter estimation

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

fraction of times word *w*$_i$ appears among all words in documents of topic *c*

- Create mega-document for topic *c* 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*** in the topic **positive**?

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

- 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

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

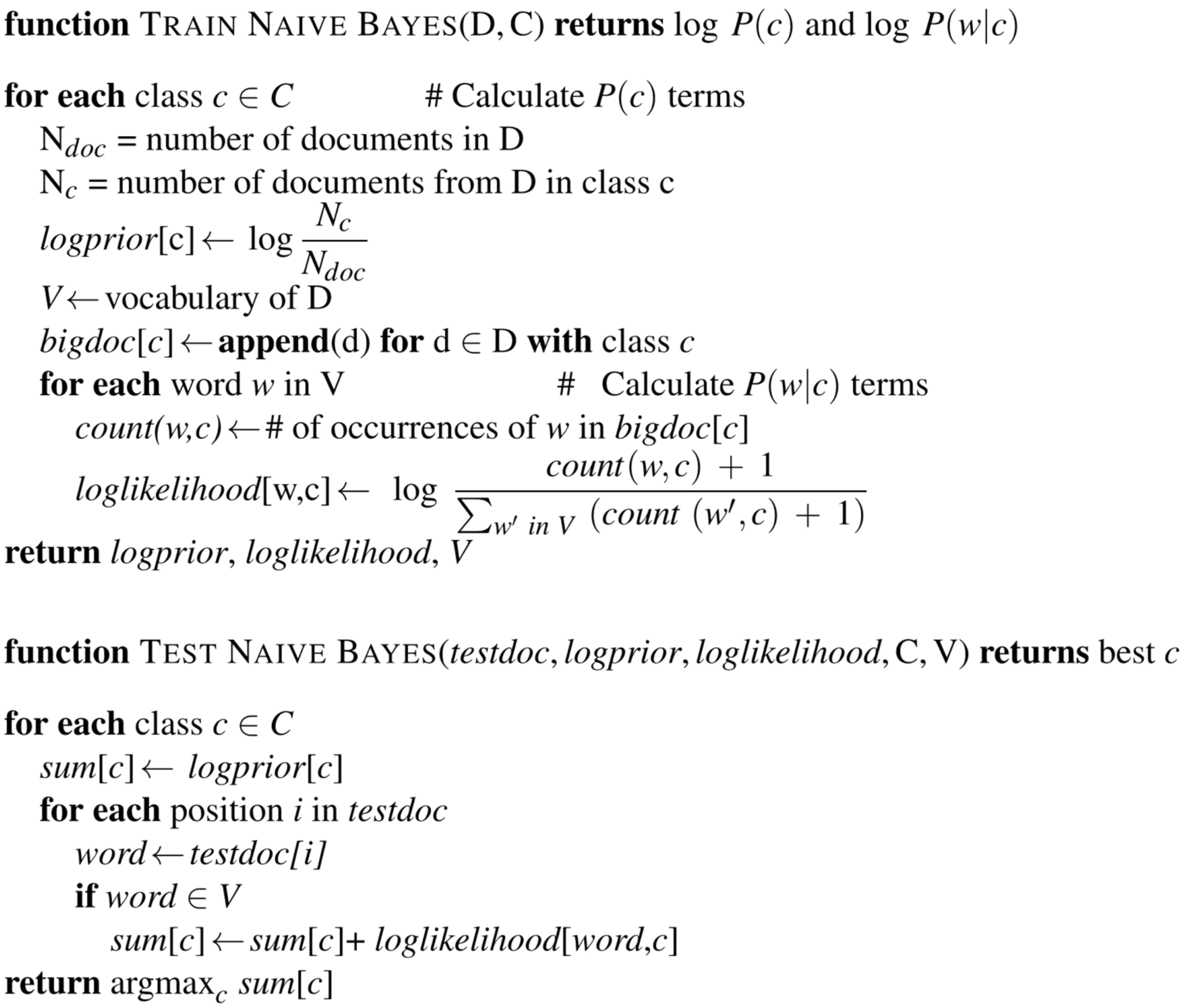
$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} \left(count(w, c) + 1\right)}$$

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

Unknown words

- Words that occur in our test data but not in any training document in any class
- Ignore such words i.e. remove them from the test document and not include any probability for them at all.
- Some systems also ignore **stop words**
  - very frequent words like *the* and *a*.
  - sort the vocabulary by frequency in the training set, and define the top 10–100 entries as stop words
  - or, use a pre-defined stop word list
  - every instance of these stop words are simply removed from both training and test documents as if they had never occurred
- However, using a stop word list doesn’t improve performance

Algorithm



```
function TRAIN NAIVE BAYES(D, C) returns log $P(c)$ and log $P(w|c)$

for each class $c \in C$              # Calculate $P(c)$ terms
    N$_{doc}$ = number of documents in D
    N$_c$ = number of documents from D in class c
    logprior[c] ← log $\frac{N_c}{N_{doc}}$
    $V$ ← vocabulary of D
    bigdoc[$c$] ← append(d) for d ∈ D with class $c$
    for each word $w$ in V                # Calculate $P(w|c)$ terms
        count(w,c) ← # of occurrences of $w$ in bigdoc[$c$]
        loglikelihood[w,c] ← log $\frac{count(w,c) + 1}{\sum_{w' \text{ in } V} (count(w',c) + 1)}$
return logprior, loglikelihood, $V$

function TEST NAIVE BAYES(testdoc, logprior, loglikelihood, C, V) returns best $c$

for each class $c \in C$
    sum[$c$] ← logprior[$c$]
    for each position $i$ in testdoc
        word ← testdoc[i]
        if word ∈ $V$
            sum[$c$] ← sum[$c$]+ loglikelihood[word,$c$]
return argmax$_c$ sum[$c$]
```

Example: Sentiment analysis with add-1 smoothing

	Cat	Documents
Training	-	just plain boring
	-	entirely predictable and lacks energy
	-	no surprises and very few laughs
	+	very powerful
	+	the most fun film of the summer
Test	?	predictable with no fun

Prior: N_c / N_{doc}

$$P(-) = \frac{3}{5} \qquad P(+) = \frac{2}{5}$$

Dropping "with" as an unknown word

$$P(\text{“predictable”}|-) = \frac{1+1}{14+20} \qquad P(\text{“predictable”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“no”}|-) = \frac{1+1}{14+20} \qquad P(\text{“no”}|+) = \frac{0+1}{9+20}$$

$$P(\text{“fun”}|-) = \frac{0+1}{14+20} \qquad P(\text{“fun”}|+) = \frac{1+1}{9+20}$$

$$\boxed{\begin{aligned} P(-)P(S|-) &= \frac{3}{5} \times \frac{2 \times 2 \times 1}{34^3} = 6.1 \times 10^{-5} \\ P(+)P(S|+) &= \frac{2}{5} \times \frac{1 \times 1 \times 2}{29^3} = 3.2 \times 10^{-5} \end{aligned}}$$

The test sentence belongs to class *negative*.

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

P(c|d5) $\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14
≈ 0.0003

P(j|d5) $\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9
≈ 0.0001

Sentiment Analysis: Optimization

- Whether a word occurs or not seems to matter more than its frequency
- Improves performance by clipping word counts in **each document** at 1
  - called **binary multinomial naive Bayes** or **binary NB**
  - for each document remove all duplicate words before concatenating them into the single big document
  - the word *great* has a count of 2 even for Binary NB, because it appears in multiple documents.

Sentiment Analysis: Optimization

Four original documents:

– it was pathetic the worst part was the boxing scenes

– no plot twists or great scenes

+ and satire and great plot twists

+ great scenes great film

After per-document binarization:

– it was pathetic the worst part boxing scenes

– no plot twists or great scenes

+ and satire great plot twists

+ great scenes film

	NB Counts		Binary Counts	
	+	−	+	−
and	2	0	1	0
boxing	0	1	0	1
film	1	0	1	0
great	3	1	2	1
it	0	1	0	1
no	0	1	0	1
or	0	1	0	1
part	0	1	0	1
pathetic	0	1	0	1
plot	1	1	1	1
satire	1	0	1	0
scenes	1	2	1	2
the	0	2	0	1
twists	1	1	1	1
was	0	2	0	1
worst	0	1	0	1

Sentiment Analysis: Optimization

- when a negation is present, the sentiment of the subsequent words may be reversed or altered.
- **Negation**
  - *I really like this movie* (positive)
  - *I didn't like this movie* (negative)
- Prepend the prefix *NOT* to every word after a token of logical negation (*n't, not, no, never*) until the next punctuation mark
  - *didn't like this movie , but I*
  - *didn't NOT_like NOT_this NOT_movie , but I*
- 'words' like *NOT_like*, *NOT_recommend* will occur more often in negative documents, while words like *NOT_bored*, *NOT_dismiss* will acquire positive associations

Sentiment Analysis: Insufficient data

- Insufficient labeled training data to train accurate naive Bayes classifiers
- **Sentiment lexicons**
  - Extract positive and negative word features from **sentiment lexicons**
    - lists of words that are pre-annotated with positive or negative sentiment

- MPQA subjectivity lexicon
  - 6885 words, 2718 positive and 4912 negative
  - + : *admirable, beautiful, confident, dazzling, ecstatic, favor, glee, great*

    − : *awful, bad, bias, catastrophe, cheat, deny, envious, foul, harsh, hate*

- If we do not have a lot of training data, add features:
  - **'this word occurs in the positive lexicon'**
  - **'this word occurs in the negative lexicon'**
  - And treat all instances of words in the lexicon as counts for that one feature, instead of counting each word separately
  - If we have lots of training data, and if the test data matches the training data, using just two features won't work as well as using all the words

SPAM vs HAM: Optimization

- Rather than using all the words as individual features:
  - predefine likely sets of words or phrases as features
  - including features that are not purely linguistic
- E.g. the open-source SpamAssassin has features like:
  - the phrase “one hundred percent guaranteed”
  - the feature *mentions “millions of dollars”* (as a regex)
  - Not purely linguistic features: *HTML has a low ratio oftext to image area*
  - Non-linguistic features: “the path that the email took to arrive”
  - Other features:
    - Email subject line is all capital letters
    - Contains phrases of urgency like “urgent reply”
    - Email subject line contains “online pharmaceutical”
    - HTML has unbalanced ”head” tags
    - Claims you can be removed from the list

SPAM vs HAM: Optimization

Rather than using all the words as individual features:

1.Predefined Sets of Likely Features:

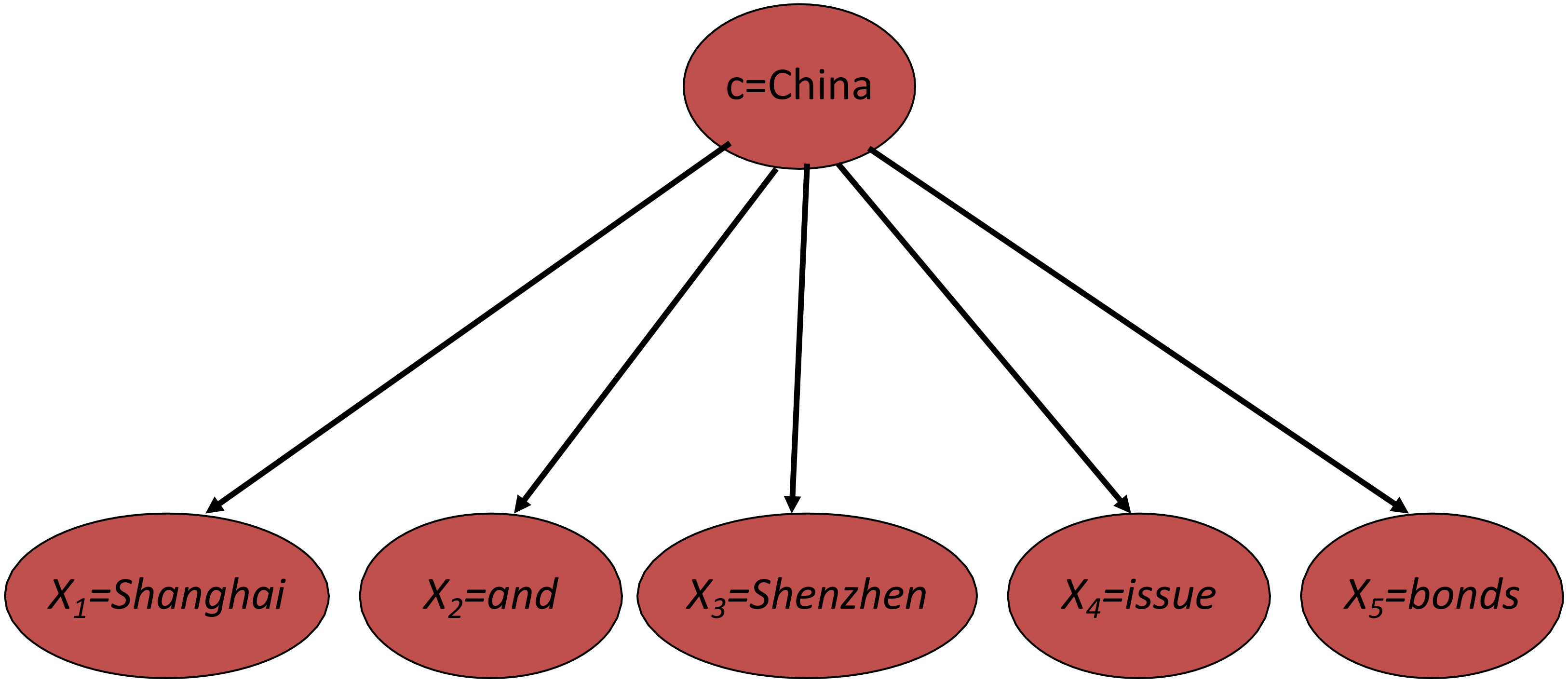
- Instead of using all words as individual features, predefined sets of words or phrases are identified as features.
- These sets are likely to appear frequently in either spam or ham emails. This approach reduces the dimensionality of the feature space and focuses on relevant linguistic patterns.

2.Including Non-Purely Linguistic Features:

- Features are not limited to linguistic elements alone.
- They may include non-linguistic aspects of emails, such as their structure, formatting, or metadata.
- For instance, features like HTML text-to-image ratio or the email's routing path can be considered.

Naïve Bayes: Relationship to Language Modeling

Generative Model for Multinomial Naïve Bayes



Naïve Bayes and Language Modeling

- Naïve bayes classifiers can use any sort of feature
  - URL, email address, dictionaries, network features
- But if, as in the previous slides
  - We use **only** word features
  - we use **all** of the words in the text (not a subset)
- Then
  - Naïve bayes has an important similarity to language modeling.

Each class = a unigram language model

- Assigning each word: P(word | c)
- Assigning each sentence: $P(s|c) = \prod_{i \in positions} P(w_i|c)$

Class	*pos*
0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film
…	

I	love	this	fun	film
0.1	0.1	.05	0.01	0.1

P(s | pos) = 0.0000005

Naïve Bayes as a Language Model

- Which class assigns the higher probability to s?

Model pos

0.1	I
0.1	love
0.01	this
0.05	fun
0.1	film

Model neg

0.2	I
0.001	love
0.01	this
0.005	fun
0.1	film

I	love	this	fun	film
0.1	0.1	0.01	0.05	0.1
0.2	0.001	0.01	0.005	0.1

$$P(s|pos) > P(s|neg)$$

Multinomial Naïve Bayes: Another Worked Example

$$\hat{P}(c) = \frac{N_c}{N}$$

$$\hat{P}(w \mid c) = \frac{count(w,c)+1}{count(c)+|V|}$$

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

Priors:

$P(c) = \frac{3}{4}$

$P(j) = \frac{1}{4}$

Conditional Probabilities:

P(Chinese|*c*) = (5+1) / (8+6) = 6/14 = 3/7

P(Tokyo|*c*) = (0+1) / (8+6) = 1/14

P(Japan|*c*) = (0+1) / (8+6) = 1/14

P(Chinese|*j*) = (1+1) / (3+6) = 2/9

P(Tokyo|*j*) = (1+1) / (3+6) = 2/9

P(Japan|*j*) = (1+1) / (3+6) = 2/9

Choosing a class:

P(c|d5)

$\propto$ 3/4 * $(3/7)^3$ * 1/14 * 1/14

≈ 0.0003

P(j|d5)

$\propto$ 1/4 * $(2/9)^3$ * 2/9 * 2/9

≈ 0.0001

Summary: Naive Bayes is Not So Naive

- Very good in domains with many equally important features: It excels in scenarios where multiple features are equally relevant, as it doesn't prioritize one feature over another.
- Optimal if the independence assumptions hold:
  - If assumed independence is correct, then it is the Bayes Optimal Classifier for problem
- A good dependable baseline for text classification
  - **But we will see other classifiers that give better accuracy**