Multimedia Information Retrieval and Technology

Lecture 12 Text classification and Naïve Bayes

By : Laura Liu

Room: EE314

Tel. no. 7756

Xi'an Jiaotong-Liverpool University

西交利物浦大学

Text classification and Naïve Bayes

I. Introduction

II. Naïve Bayes text classification

- ➢ Multinomial Naïve Bayes
- ➢ The Bernoulli model

III. Feature Selection

Standing queries

The path from IR to text classification:

- You have an information need to monitor:
- You want to rerun an appropriate query periodically to find new news items on this topic
- You will be sent new documents that are found

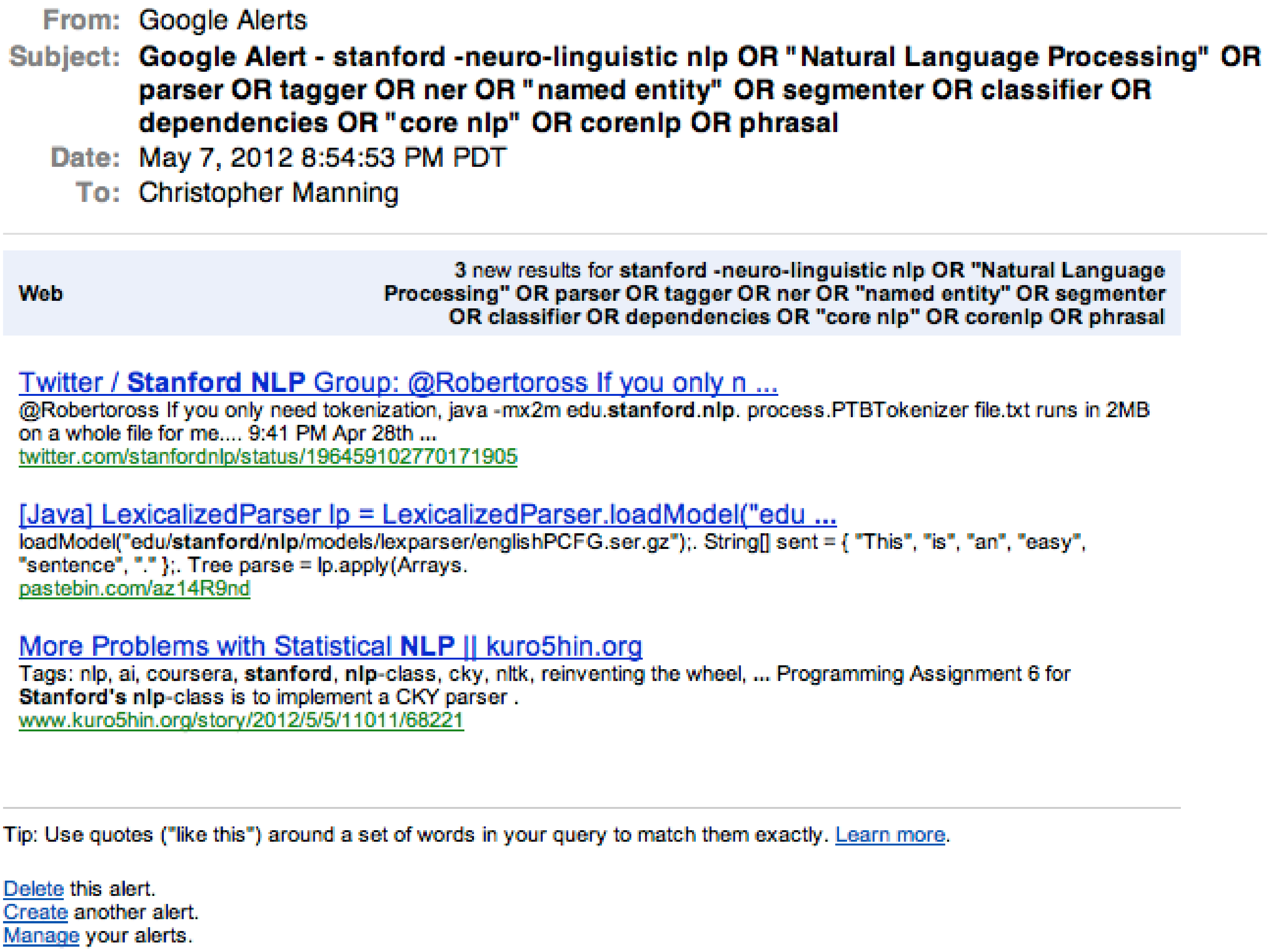
  i.e., it's not ranking but classification (relevant vs. not relevant)

Standing queries

Such queries are called **standing queries**

- Long used by “information professionals”
- A modern example is **Google Alerts**

Standing queries are text classifiers



From: Google Alerts
Subject: **Google Alert - stanford -neuro-linguistic nlp OR "Natural Language Processing" OR parser OR tagger OR ner OR "named entity" OR segmenter OR classifier OR dependencies OR "core nlp" OR corenlp OR phrasal**
Date: May 7, 2012 8:54:53 PM PDT
To: Christopher Manning

Web — 3 new results for **stanford -neuro-linguistic nlp OR "Natural Language Processing" OR parser OR tagger OR ner OR "named entity" OR segmenter OR classifier OR dependencies OR "core nlp" OR corenlp OR phrasal**

Twitter / **Stanford NLP** Group: @Robertoross If you only n ...
@Robertoross If you only need tokenization, java -mx2m edu.**stanford.nlp**. process.PTBTokenizer file.txt runs in 2MB on a whole file for me.... 9:41 PM Apr 28th ...
twitter.com/stanfordnlp/status/196459102770171905

[Java] LexicalizedParser lp = LexicalizedParser.loadModel("edu ...
loadModel("edu/**stanford**/**nlp**/models/lexparser/englishPCFG.ser.gz");. String[] sent = { "This", "is", "an", "easy", "sentence", "." };. Tree parse = lp.apply(Arrays.
pastebin.com/az14R9nd

More Problems with Statistical **NLP** || kuro5hin.org
Tags: nlp, ai, coursera, **stanford**, **nlp**-class, cky, nltk, reinventing the wheel, ... Programming Assignment 6 for **Stanford's nlp**-class is to implement a CKY parser .
www.kuro5hin.org/story/2012/5/5/11011/68221

Tip: Use quotes ("like this") around a set of words in your query to match them exactly. Learn more.

Delete this alert.
Create another alert.
Manage your alerts.

Spam filtering
Another text classification task

From: "" <takworlld@hotmail.com>
Subject: real estate is the only way... gem oalvgkay

Anyone can buy real estate with no money down

Stop paying rent TODAY !

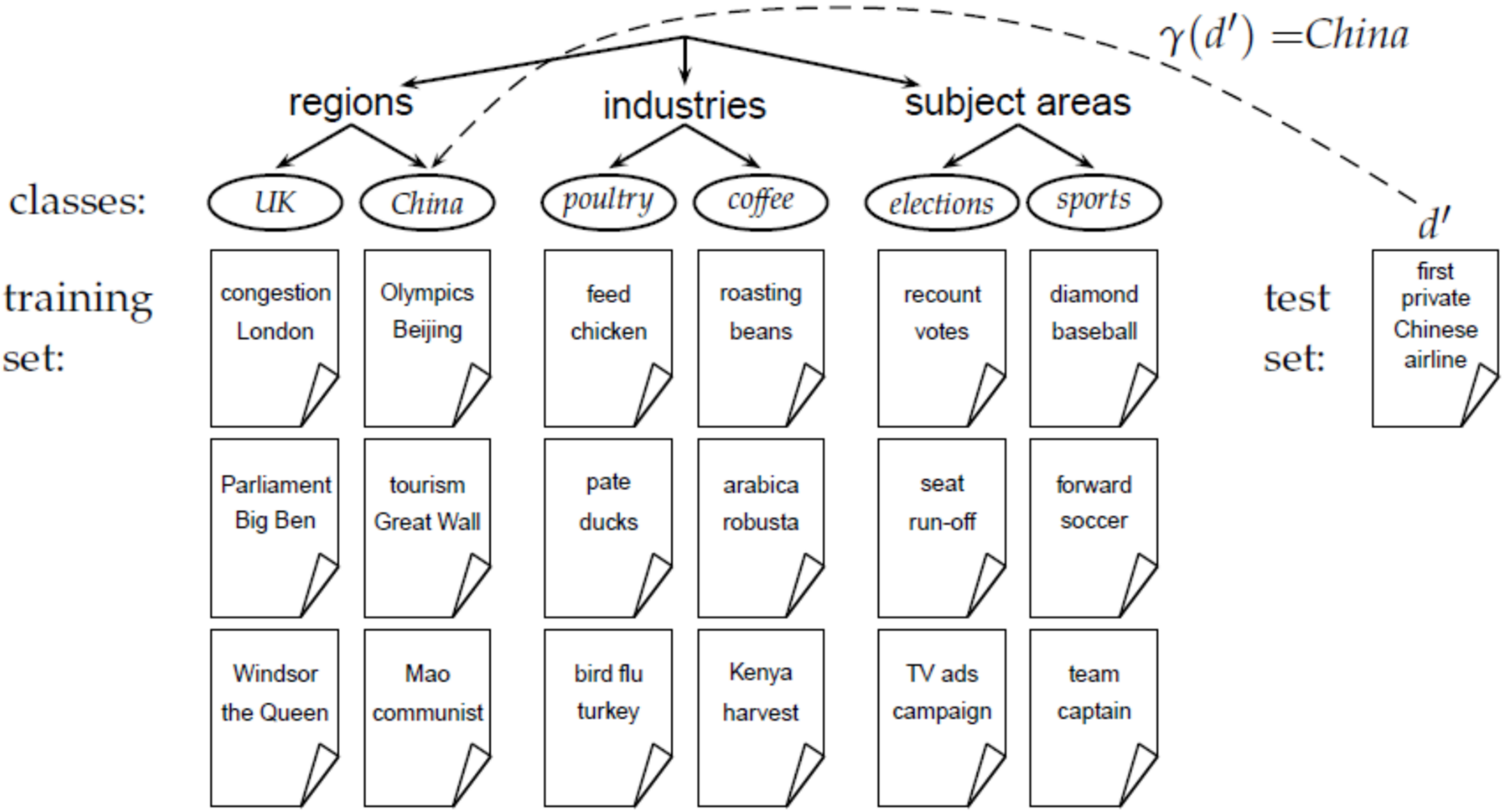
There is no need to spend hundreds or even thousands for similar courses

I am 22 years old and I have already purchased 6 properties using the methods outlined in this truly INCREDIBLE ebook.

Change your life NOW !

===
Click Below to order:
http://www.wholesaledaily.com/sales/nmd.htm
===

Example



- **Figure 13.1** Classes, training set, and test set in text classification .

Example

There are six classes (*UK*, *China*,. . . , *sports*), each with three training documents.

The training set provides some typical examples for each class, so that we can learn the classification function γ.

TEST SET: Once we have learned γ, we can apply it to the *test set* (or *test data*), for example, the new document *first private Chinese airline* whose class is unknown.

Supervised learning

Given:

A document d

A fixed set of classes:

$C = \{c_1, c_2, \ldots, c_J\}$

A <u>training set</u> D of documents each with a label in C

Determine:

A learning method or algorithm which will enable us to learn **a classifier γ**

For a test document x, we assign it the class $\gamma(x) \in C$

Supervised learning

A *training set* $\mathbb{D}$ is consist of labeled documents $< d, c >$

E.g.:

$< d, c >$=<Beijing joins the World Trade Organization, *China*>

- Document: ***Beijing joins the World Trade Organization***
- The class (or label) ***China***.

Classification Methods

Supervised learning

- Naive Bayes (simple, common)
- k-Nearest Neighbors (simple, powerful)
- Support-vector machines (newer, generally more powerful)

… plus many other methods

No free lunch: requires hand-classified training data

Many commercial systems use a mixture of methods

The bag of words representation

$\gamma($

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.

$)=c$

The bag of words representation

$\gamma($

great	2
love	2
recommend	1
laugh	1
happy	1
...	...

$)=c$

Features

- Supervised learning classifiers can use any sort of feature
  - URL, email address, punctuation, capitalization, dictionaries, network features

- In the simplest bag of words view of documents
  - We use **only** word features
  - we use **all** of the words in the text (not a subset)

I. Introduction

II. Naïve Bayes text classification

- ➢ Multinomial Naïve Bayes
- ➢ The Bernoulli model

III. Feature Selection

Naïve Bayes text classification

- The Naive Bayes classifier is a probabilistic classifier.
- We compute the probability of a document d being in a class c as follows:

$$P(c|d) \propto P(c) \prod_{1 \leq k \leq n_d} P(t_k|c)$$

- $P(t_k|c)$ is the conditional probability of term t_k occurring in a document of class c
- $P(t_k|c)$ as a measure of how much evidence t_k contributes that c is the correct class.
- $P(c)$ is the prior probability of c.

Naïve Bayes text classification

- $\langle t_1, t_2, \ldots, t_{n_d} \rangle$ are the tokens in document *d* that are part of the vocabulary we use for classification
- n_d is the number of such tokens in document *d* .

Eg: *Beijing and Taipei join the WTO*

$\langle t_1, t_2, \ldots, t_{n_d} \rangle$=<Beijing, Taipei, join,WTO>, with $n_d = 4$

treat the terms *and* and *the* as stop words.

Naïve Bayes text classification

- Our goal is to find the "best" class.
- The best class in Naive Bayes classification is the most likely or maximum a posteriori (MAP) class c_{map}:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} \hat{P}(c|d) = \arg\max_{c \in \mathbb{C}} \ \hat{P}(c) \prod_{1 \leq k \leq n_d} \hat{P}(t_k|c)$$

- We write $\hat{P}$ for P since these values are estimates from the training set.

In mathematics, the arguments of the maxima (abbreviated **arg max** or **argmax**) are the points of the domain of some function at which the function values are maximized.

Multinomial NB (Naïve Bayes) model

$$\hat{P}(\mathrm{c}|\mathrm{d}) = \hat{P}(c) * \prod_{k=1}^{n_d} \hat{P}(t_k|c)$$

How do we estimate the parameters $\hat{P}(c)$ and $\hat{P}(t_k|c)$?

Multinomial NB (Naïve Bayes) model

The maximum likelihood estimate (MLE), which is simply the relative frequency and corresponds to the most likely value of each parameter given the training data.

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

N_c is the number of documents in class *c* and N is the total number of documents.

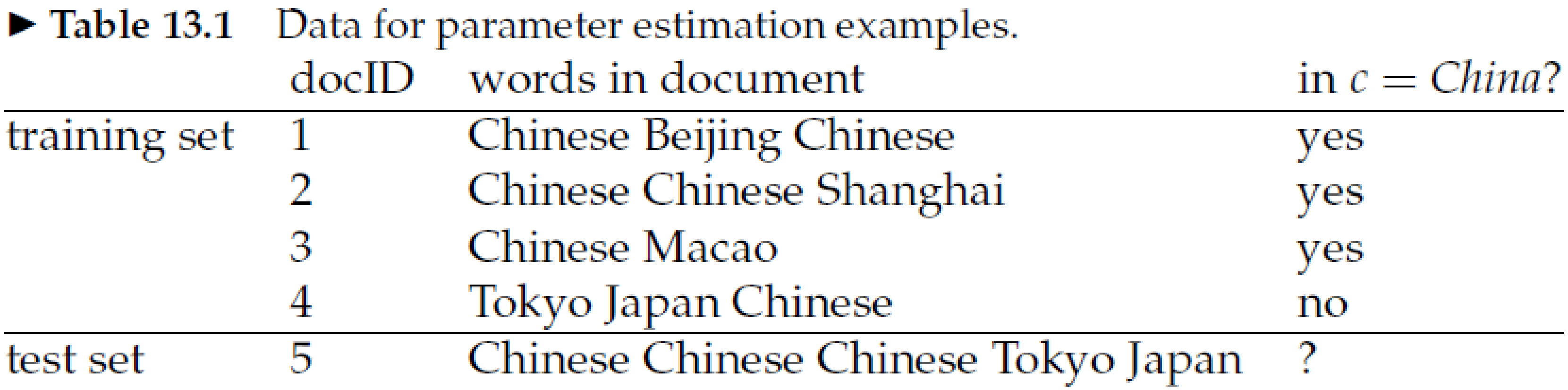
Maximum Likelihood Estimate (MLE)-- recap

For trials with categorical outcomes (such as noting the presence or absence of a term), one way to estimate the probability of an event from data is simply to count the number of times an event occurred divided by the total number of trials.

This is referred to as the ***relative frequency*** of the event.

Estimating the probability as the relative frequency is the ***maximum likelihood estimate* (or *MLE*)**, because this value makes the observed data maximally likely.

Example



▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}(\mathrm{c}|\mathrm{d}oc5) = \hat{P}(c) * \prod_{k=1}^{n_d} \hat{P}(t_k|c) \quad ?$$

$$\hat{P}_{MLE}(c) = ? \;\; \hat{P}_{MLE}(\bar{c}) = ?$$

Multinomial NB (Naïve Bayes) model

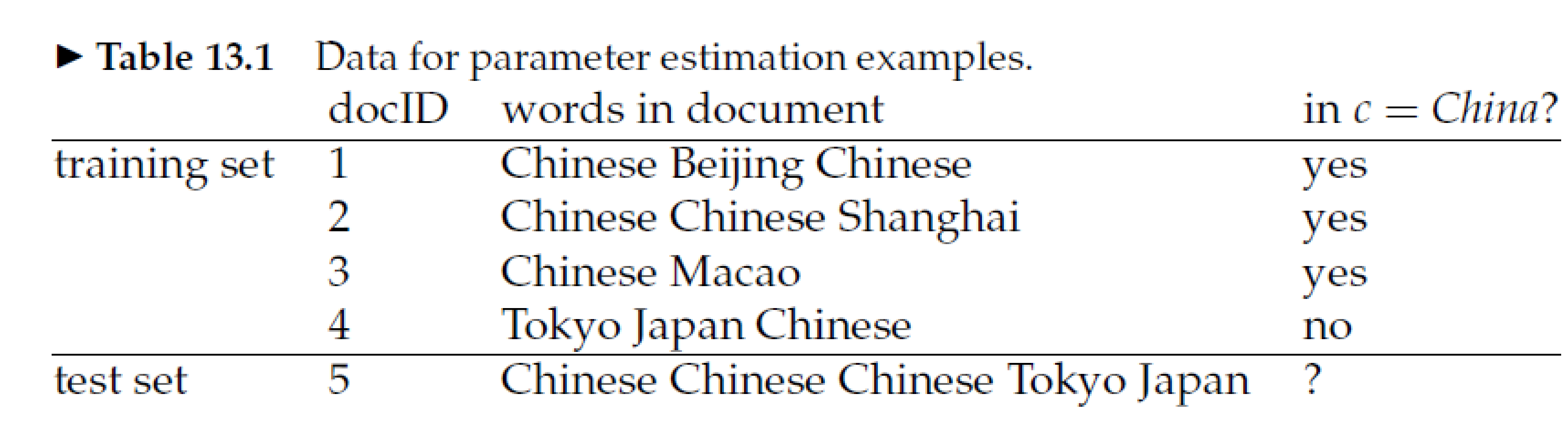
We estimate the conditional probability $\hat{P}(t|c)$ as the relative frequency of term t in documents belonging to class c:

$$\hat{P}_{MLE}(t|c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

Multinomial NB (Naïve Bayes) model:

T_{ct} is the term frequency of term *t* in training documents of class *c* (includes multiple occurrences);

Example

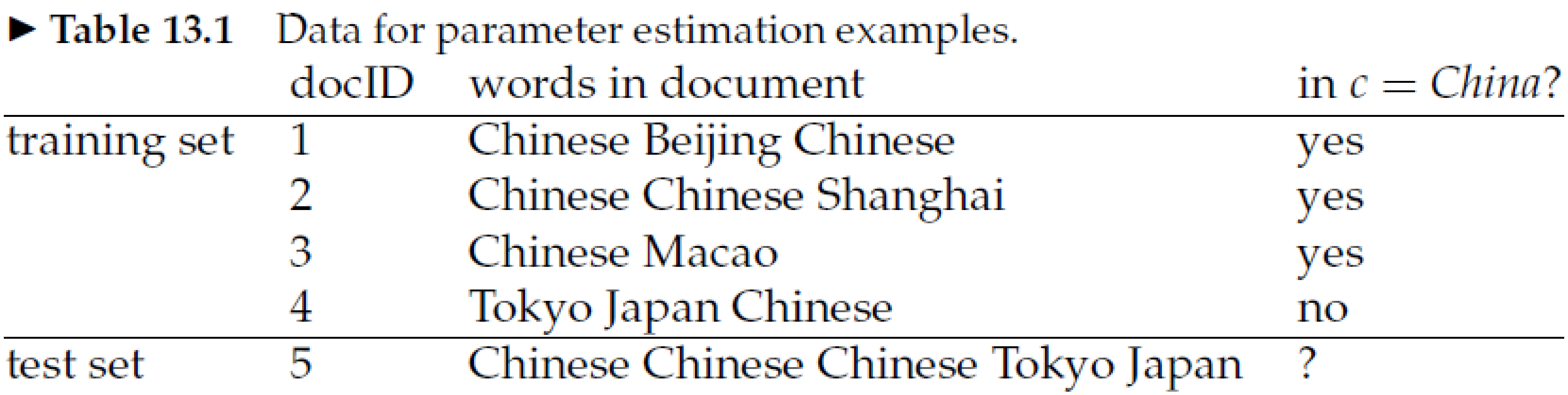


► **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}_{MLE}(chinese|c) = ?$$

$$T_{ct}? \quad T_{ct'}?$$



▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$T_{c,chinese} = ?$$

$$T_{c,Beijing} = ?\ T_{c,shanghai} = ?\ T_{c,macau} = ?$$

$$T_{c,\mathrm{Tokyo}} = T_{c,Japan} = ?$$

The problem with MLE: zeros

$$\hat{P}_{MLE}(Japan|c) = 0$$

The probability of a document *d* being in a class *c:*

$$\hat{P}(\mathrm{c|d}) = \hat{P}(c) * \prod_{k=1}^{n_d} \hat{P}(t_k|c) \;\; \hat{P}(\mathrm{c|d}oc5)$$
$$= \hat{P}(c) * \hat{P}_{MLE}(chinese|c) \; * \hat{P}_{MLE}(chinese|c)$$
$$* \hat{P}_{MLE}(chinese|c) * \hat{P}_{MLE}(\mathrm{Tokyo}|c) * \hat{P}_{MLE}(Japan|c)$$

If “Japan” is not occurred in the class c=China in training data set, we ***get 0 estimate*** for any test document that with a term “Japan”.

Xi’an Jiaotong-Liverpool University 西交利物浦大学

To avoid zeros: add one smoothing.

Add one to each count to avoid zeros

$$\hat{P}_{MLE}(t|c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\sum_{t' \in V} T_{ct'} + B}$$

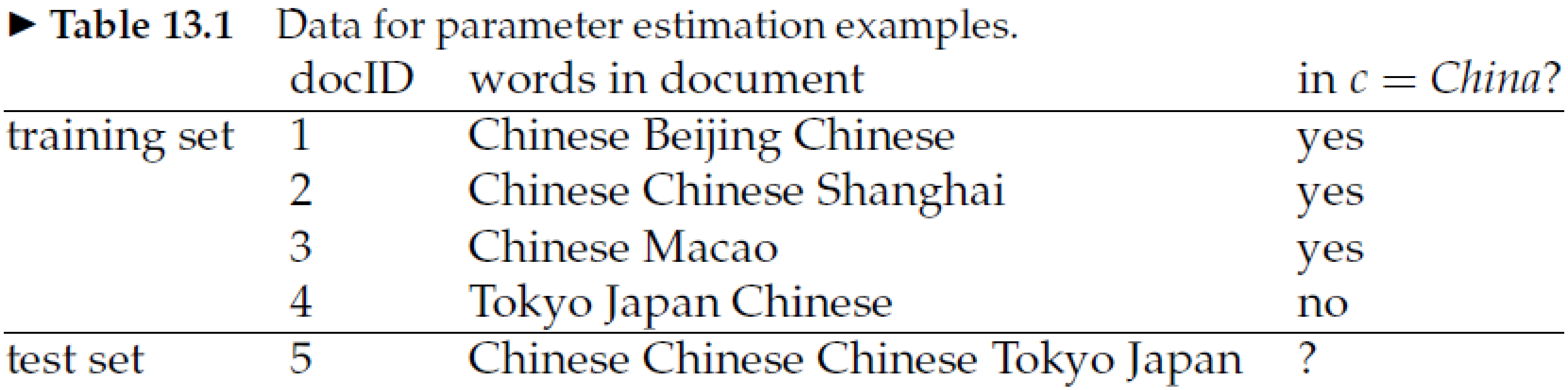
Equation 13.7

B: the number of terms in the vocabulary training set.

Also called **Laplace smoothing**.

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Example



▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

$$\hat{P}_{MLE}(chinese|c) = ?$$

$$\hat{P}_{MLE}(Tokyo|c) = ?$$

Naïve Bayes text classification

- Multiplying lots of small probabilities can result in floating point underflow.
- Since $\log(xy) = \log(x) + \log(y)$, we can sum log probabilities instead of multiplying probabilities.
- Since log is a monotonic function, the class with the highest score does not change.
- So what we usually compute in practice is:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

Naïve Bayes text classification

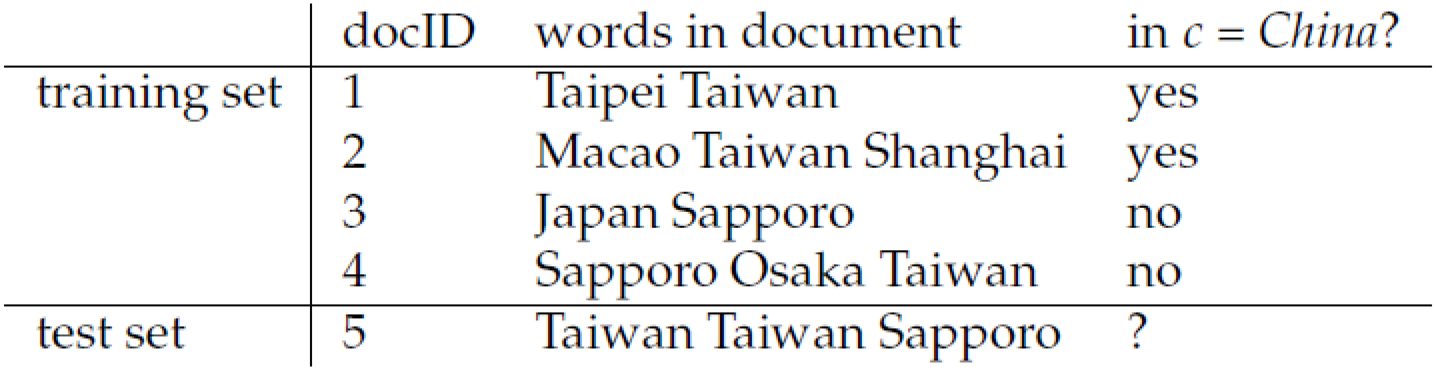
- Classification rule:

$$c_{\text{map}} = \arg\max_{c \in \mathbb{C}} [\log \hat{P}(c) + \sum_{1 \leq k \leq n_d} \log \hat{P}(t_k|c)]$$

- Simple interpretation:
  - Each conditional parameter $\log \hat{P}(t_k|c)$ is a weight that indicates how good an indicator t_k is for c.
  - The prior $\log \hat{P}(c)$ is a weight that indicates the relative frequency of c.
  - The sum of log prior and term weights is then a measure of how much evidence there is for the document being in the class.
  - We select the class with the most evidence.

Exercise

Based on the data in Table 13.10, (i) estimate a multinomial Naive Bayes classifier, (ii) apply the classifier to the test document. You need not estimate parameters that you don't need for classifying the test document.



	docID	words in document	in $c = China$?
training set	1	Taipei Taiwan	yes
	2	Macao Taiwan Shanghai	yes
	3	Japan Sapporo	no
	4	Sapporo Osaka Taiwan	no
test set	5	Taiwan Taiwan Sapporo	?

▶ **Table 5** Data for parameter estimation exercise.

I. Introduction

II. Naïve Bayes text classification

- ➢ Multinomial Naïve Bayes
- ➢ The Bernoulli model

III. Feature Selection

The Bernoulli model

$$\hat{P}_{MLE}(t|c) = \frac{df_{ct} + 1}{N_c + No\ of\ classes}$$

df_{ct} is the document frequency of term t in class c ;
N_c is the number of documents in class c.

The Bernoulli model

The multivariate Bernoulli model:
The conditional probability $\hat{P}(t|c)$ is estimated as fraction of ***documents of class*** c that contain term t.

The multinomial NB model: fraction ***of tokens in documents*** from class *c* that contain term *t*.

The Bernoulli model

The probability of nonoccurrence is factored in when computing $P(c|d)$:

Models absence of terms explicitly.

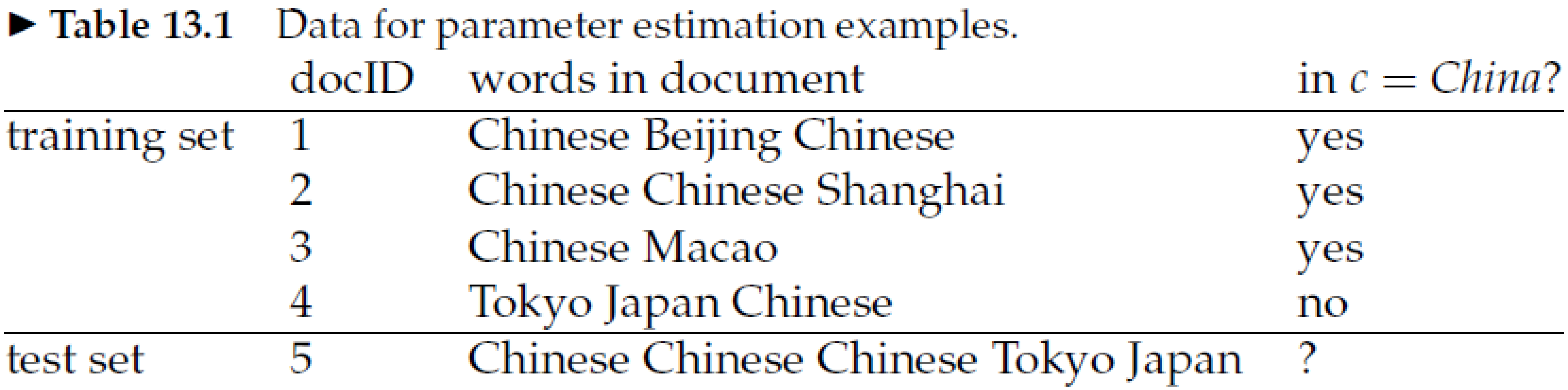
$$c_{map} = \arg\max \hat{P}(c) \prod_{t_k \in q} \hat{P}(t_k|c)) \prod_{t_k \notin q} [1 - \hat{P}(t_k|c)]$$

The Bernoulli model

The Bernoulli model ignore the number of occurrences.

Make mistakes when classifying long documents.

Example



▶ **Table 13.1** Data for parameter estimation examples.

	docID	words in document	in $c = China$?
training set	1	Chinese Beijing Chinese	yes
	2	Chinese Chinese Shanghai	yes
	3	Chinese Macao	yes
	4	Tokyo Japan Chinese	no
test set	5	Chinese Chinese Chinese Tokyo Japan	?

Applying the Bernoulli model to the above example again.

I. Introduction

II. Naïve Bayes text classification

   - Multinomial Naïve Bayes
   - The Bernoulli model

III. Feature Selection

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Feature Selection:

The process of selecting a subset of the terms occurring in the training set and using only this subset as features in text classification.

A ***noise feature*** is one that, when added to the document representation, increases the classification error on new data.

Feature Selection:

Such an incorrect generalization from an accidental property of the training set is called ***overfitting***.

The basic feature selection algorithm : For a given class *c*, we

1. compute a utility measure *A*(*t*, *c*) for each term
2. select the *k* terms that have the highest values of *A*(*t*, *c*).
3. All other terms are discarded and not used in classification.

Feature Selection:

A(*t*, *c*):

1. ***Mutual information***: how much information the presence/absence of a term contributes to making the classification decision.
2. χ^2 ***feature selection***: measure the independence of two events.
3. ***Frequency-based*** feature selection

Xi'an Jiaotong-Liverpool University
西交利物浦大学

Feature Selection: Why?

Text collections have a large number of features

- 10,000 – 1,000,000 unique words … and more

Selection may make a particular classifier feasible

- Some classifiers can't deal with 1,000,000 features

Reduces training time

- Training time for some methods is quadratic or worse in the number of features

Makes runtime models smaller and faster

Can improve generalization (performance)

- Eliminates noise features
- Avoids overfitting

Feature Selection: Frequency

The simplest feature selection method:

- Just use the commonest terms
- No particular foundation
- But it make sense why this works

They' re the words that can be well-estimated and are most often available as evidence

In practice, this is often 90% as good as better methods