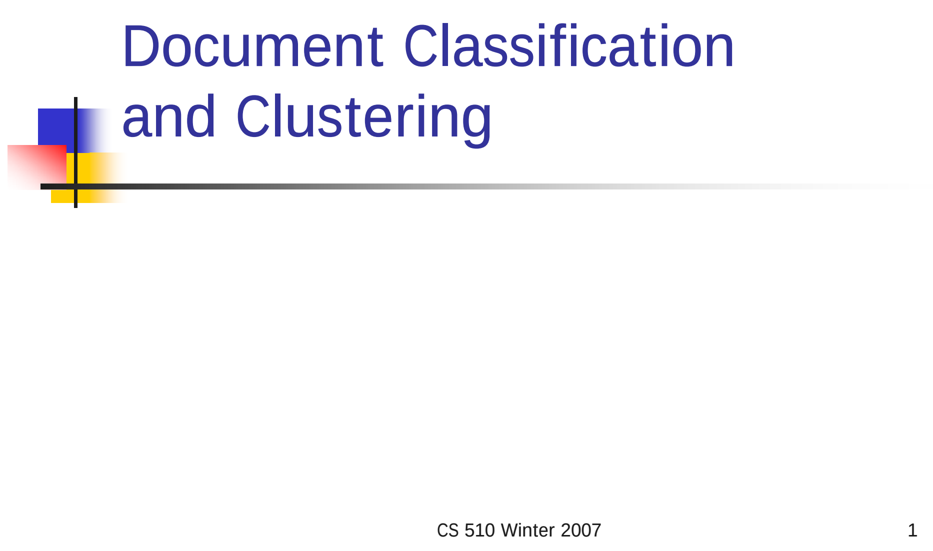
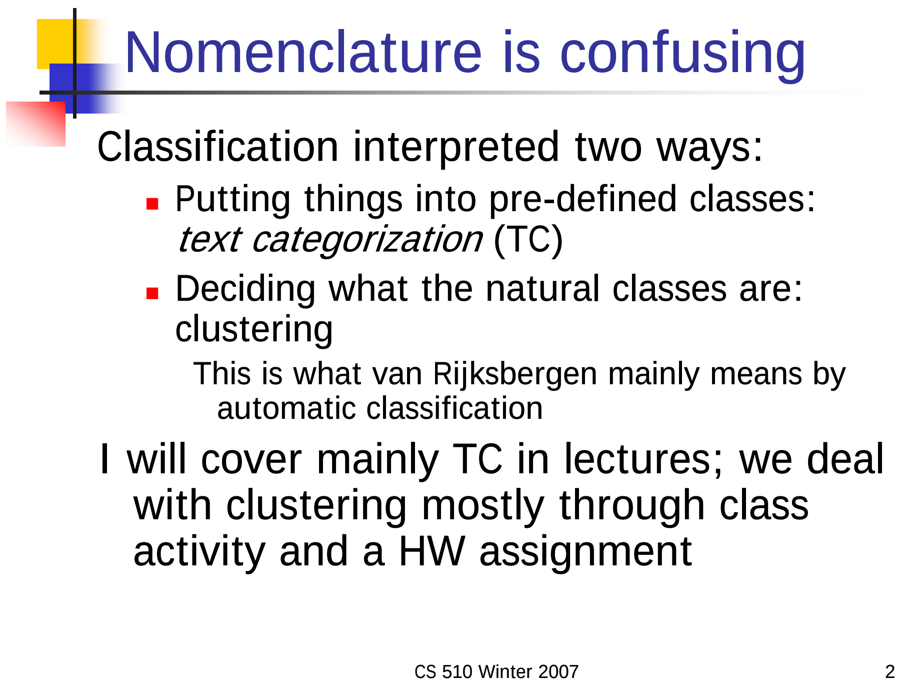


Document Classification and Clustering



Nomenclature is confusing

Classification interpreted two ways:

- Putting things into pre-defined classes: *text categorization* (TC)
- Deciding what the natural classes are: clustering

  This is what van Rijksbergen mainly means by automatic classification

I will cover mainly TC in lectures; we deal with clustering mostly through class activity and a HW assignment

Approaches

I will cover mainly approaches that are related to IR search techniques

There are many other approaches

- Probabilistic
- Regression
- Neural nets

Example: Classified Ads

Boats

16' 2001 ALUMACRAFT with 25HP Yamaha. Tiller. Elect trolling motor. Excellent cond. Many extras. Galv trailer. $4500

18.2 – 1994 ALUMAWELD Intruder Full top, 90hp Mercury with kicker, less than 15 hrs, like new. $14,000/obo.

18' Alumaweld 200 Merc w/fresh pump, $15,000 OBO. 9.9 Kicker, F. Finder, VHF Radio. New: Seats, Top, Rogue Galv. Trailer.

21' 03 TRITON TR21DC, Merc 225 Optimax. 36V troll motor. Boat loaded. Low hours on motor. Like new. $28,000.

21' ALUMAWELD, 200HP Mercury Jet & 15HP kicker, top & side covers, FF, trailer, lots of extras. $9,500 obo.

Furniture

$100/SET > Full – Queen – King – Mattress & box spring, in plastic. Used, but in good condition.

$125 QUEEN SET Dbl Pillowtop. New in plastic, factory warranty. Can deliver

$185 KING Double Pillowtop NEW! Mattress Set. W/Warranty. Can Deliver.

DINING Table Solid Maple, excellent condition, W/6 Chairs. $275 OBO.

NEW LEATHER SOFA & LOVE Lifetime warranty. Still in crates. Retail $1850. Sell $699. Can deliver.

SOFA: 3 piece dark green Sectional, pull-out bed, 2 recliners & phones $350

Dogs

ENGLISH BULLDOG, AKC, F, brindle, house broke, inside dog only, all shots, very friendly like kids $500 cash only

ENGLISH BULLDOG PUPS, born Thanksgiving day, sweet natured & beautiful blood lines, AKC, $1800-$2000

CHIHUAHUAS $300 each. Purebred, males & females available.

CHIHUAHUA, AKC with ped. Pups $600, 6 wks, 1 girl, 4 boys.

COCKER SPANIEL, AKC reg., $300 each; Black F 3-yr, Buff F 18 mo, Buff M 2-yr.

GOLDEN Retriever Pups, bred for hlth & beauty, 4 F $750 ea, 7 M $700 ea, born 12/7.

GOLDEN Retriever puppies, male & female, ready now. $300 with shots.

Classified ad scenarios

Mentioned in Sebastiani paper

Want to categorize ads under headings; possibilities

- Help a person placing an ad find an appropriate category
- Redo categorization when ad is reused in another venue

`Dogs → Pets`

`Furniture → Tables, Sofas, Beds`

Issue here

Note that these ads tend not to include the category name explicitly

Could be a problem for boats, dogs

Furniture, not so much

Definitions for TC

D: documents

C: categories, with labels

$\langle d_j, c_i \rangle \rightarrow \{T,F\}$

Whether document d_j in categoryc_i

Have a classifier function that attempts to determine this relationship

$\Phi: \mathbf{D} \times \mathbf{C} \rightarrow \{T,F\}$

Single label vs. multi-label

Might want classified ads classified to single best label

Might want news stories about presidential candidates labeled by multiple candidates

Multi-label

- Arbitrary number
- Exactly k
- At most k

Binary TC

Given category c, classify a document as c or $\neg c$

Note that multi-label TC can be viewed as multiple binary TC tasks for categories c_1, c_2, …, c_n

Two perspectives

- Document-pivoted categorization (DPC)

  Given document d, find the category (or categories)

- Category-pivoted categorization (CPC)

  Given category c, find all the documents that belong to it

Not used equally

DPC is more common

- Have categories and a new document arrives, want to categorize it

CPC might happen if a new category shows up from time-to-time

- New presidential candidate declares

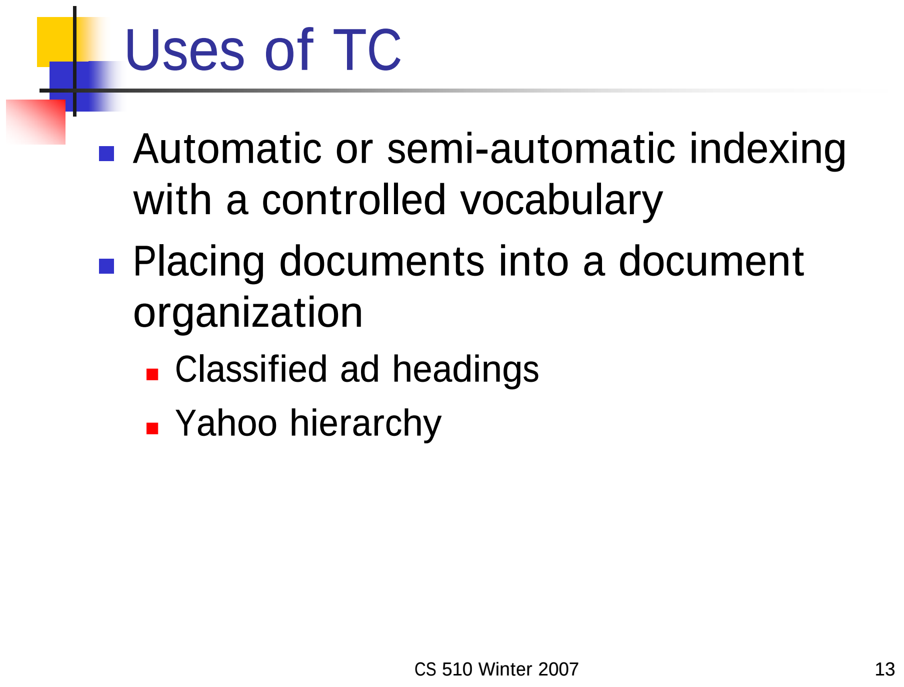
Hard vs. ranked categorization

Rather than T/F, might have a ranking from either perspective

- Given a document d, rank categories in **C** by degree to which d is appropriate to the category
- Given a category c, rank documents in **D** by which are most appropriate to c.

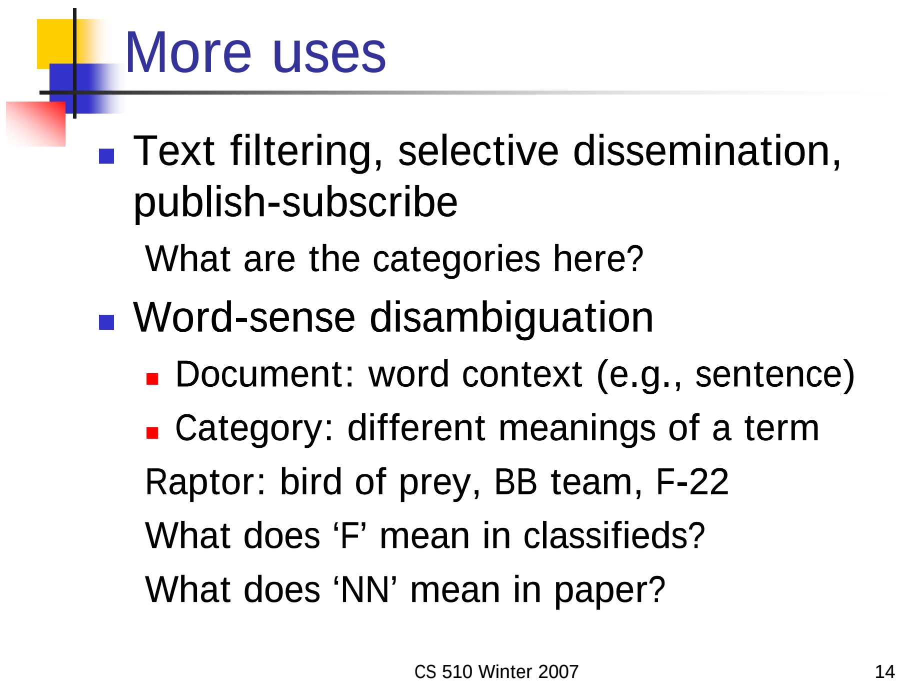
Can view as

$$\Phi:\mathbf{D}\times\mathbf{C} \rightarrow [0,1]$$



Uses of TC

- Automatic or semi-automatic indexing with a controlled vocabulary
- Placing documents into a document organization
  - Classified ad headings
  - Yahoo hierarchy



More uses

- Text filtering, selective dissemination, publish-subscribe

  What are the categories here?
- Word-sense disambiguation
  - Document: word context (e.g., sentence)
  - Category: different meanings of a term

  Raptor: bird of prey, BB team, F-22

  What does ‘F’ mean in classifieds?

  What does ‘NN’ mean in paper?

Constructing a categorization function

Could be manual – writing rules for example

Could be via learning

- Want to generalize from manually labeled sample data
- Supervised learning from labeled examples
  - Could have only positive examples, labeled c_i
  - Could have positive and negative examples, labeled c_i and $\neg c_i$

Using labeled examples

Split the example corpus 2 or 3 ways

- Training set: input for learning algorithm
- [Validation set: tuning parameters or thresholds]
- Test set: see how good the classifier is

Generality of a category c

$$\frac{\#\text{docs classified as c}}{\#\text{docs}}$$

Will make a difference in some of the evaluation measures

Document representation

Usually a vector of term weights

What’s a term?

- Usually a word or a stem, maybe a phrase
- Generally stop out function words: prepositions, conjunctions articles

Weights can be 0,1, or tf-idf style

Use of stemming: might improve efficiency, but can reduce effectiveness

Darmstadt Indexing Approach

Used in the AIR/X system

Considers wide range of properties

- term properties (frequency, location)
- document properties (format, length)
- category properties (generality)

Builds from a relevance description r(d,c) for each document-category pair

Class exercise

Food features

Do not open the containers, please

Come up with as many features as you can that might be used to categorize these different items

Issues with term vectors

They can be large

- Can make learning, categorization expensive

  Often doing document-document similarity comparison; inverted index of limited use

- Use of low-frequency terms can cause over-fitting

  Cf ‘Thanksgiving’ in Dog ads

Dimensionality reduction

Term-space reduction: reduce number of terms considered

- Globally
- Per category

Trim vectors individually for documents

Example strategies

Document frequency

- Top 10%
- All terms with frequency > 3

  Low-frequency terms can be misspellings: **`hte`**

Probabilistic notions

Try to find words that are good discriminators between a category and its complement

Odds ratio

```
(TC/NTC)/(TNC/NTNC) =
(TC*NTNC)/(NTC*TNC)
```

	t	¬t
c	TC	NTC
¬c	TNC	NTNC

Examples

	t	¬t
c	80	20
¬c	350	350

	t	¬t
c	12	8
¬c	200	300

	t	¬t
c	12	2
¬c	400	100

Term Vector Database

The Term Vector Database: Fast access to indexing terms for web pages, R. Stata, K. Bharat, F. Maghoul, *Computer Networks,* June 2000

Uses whole term vector

- Topic distillation: Highly connected pages in relevant topic should rank high in search results
- Classify search results into 12 top-level Yahoo categories

Both term-space and vector reduction

- Porter stemming algorithm
- Middle third of AltaVista index, minus stop list
- Select 50 terms per document with greatest tf-idf

  Might drop a few more, so that encoding fits in 128 bits

Term clustering

Treat groups of synonyms or highly related words as a single term

Building classifiers

For each category c_i, have a categorization status value CSV_i

CSV_i:**D** → [0,1]

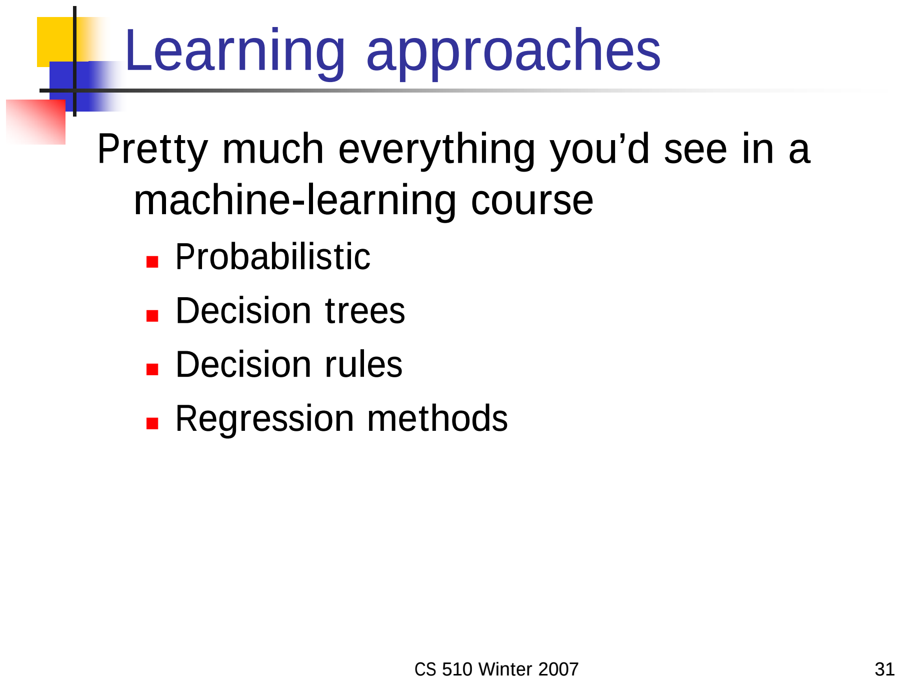
$CSV_i(d)$ is the strength of membership of d in c_i.

Using CSV's

Can use CSV to rank (documents, categories), or can threshold it to get a hard classification: $CSV_i(d) \geq \tau_i$

Setting thresholds

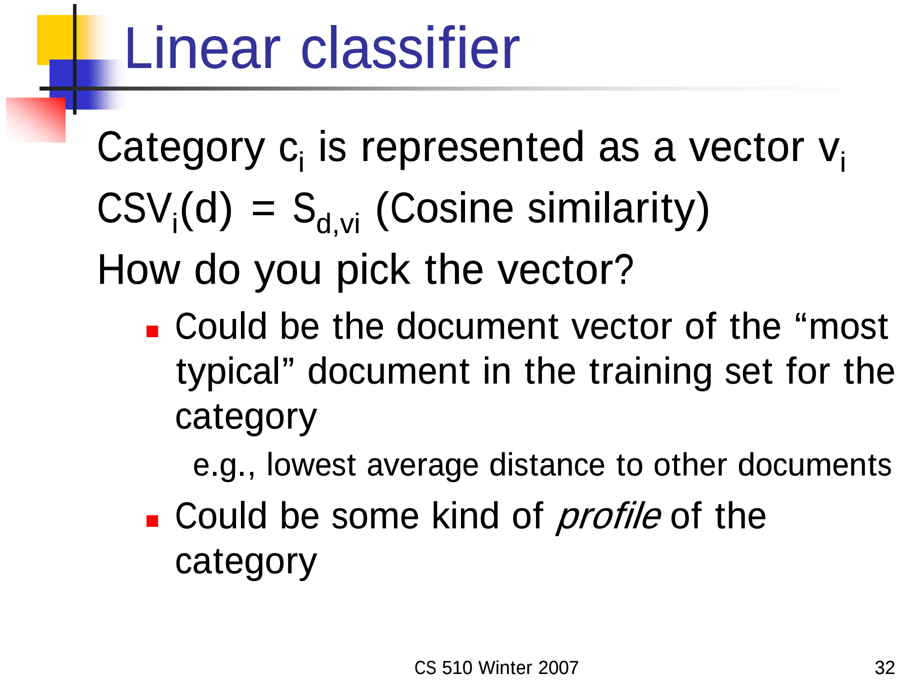
- Try to get the same generality for categories in both training and validation sets
- Tune for a particular effectiveness (precision vs. recall, for example)
- Top k per document (not really a threshold)



Learning approaches

Pretty much everything you'd see in a machine-learning course

- Probabilistic
- Decision trees
- Decision rules
- Regression methods



Linear classifier

Category c_i is represented as a vector v_i

$CSV_i(d) = S_{d,vi}$ (Cosine similarity)

How do you pick the vector?

- Could be the document vector of the "most typical" document in the training set for the category
  - e.g., lowest average distance to other documents
- Could be some kind of *profile* of the category