433-352 Data on the Web

Semester 2, 2007

Lecturer: Tim Baldwin

THE UNIVERSITY OF MELBOURNE

Lecture 7

Text Categorisation

What is Text Categorisation?

- Given:
  1. a description of a document $x \in X$
  2. a fixed set of categories $C = \{c_1, c_2, ..., c_n\}$

- Determine:

  the category of $x : c(x) \in C$, where $c(x)$ is a **categorisation function** whose domain is X and whose range is C.

- That is, how do we build categorization functions (**classifiers**) which can operate over an arbitrary description language and within an arbitrary category space?

Jackson and Moulinier (2002:pp119–125); Chakrabarti (2003:pp125–6); McCallum and Nigam (1998)

Is This Really Something I Want to Read?

```
Subject: BUSINESS INVESTMENT
From: COLLINS JAMES <collins55ng@yahoo.co.in>

Attention: President/Director,

I am the chairman of the contract award committee of the Gold and
Natural resources ministry here in Dakar Senegal, for security reasons,
I may not wish to disclose the most important thing for now until I hear
from you.

After due deliberation with my partner, I decided to forward to you this
business proposal, we want you to assist us receive the sum of Twenty
eight million, six hundred thousand united state bills ($28.6M) into
your account.
```

⋮

Example Applications of Text Categorisation

- Assign labels to a document, e.g.:
  - ⋆ Yahoo!-style topic label (e.g. `sport`, `news>world>asia>business`)
  - ⋆ genre (e.g. `job listing`, `news`)
  - ⋆ `spam` vs. `not-spam`
  - ⋆ `contains-adult-language` vs. `conforms-to-Bush-sensibilities`
- Determine the authorship of a paper
- Document routing

Jackson and Moulinier (2002:pp119–125); Chakrabarti (2003:pp125–6); McCallum and Nigam (1998)

- Indexing (digital libraries, etc.)

- Sorting

- Identifying e-scams

⋮

Jackson and Moulinier (2002:pp119–125); Chakrabarti (2003:pp125–6); McCallum and Nigam (1998)

Methods of Categorisation

- Manual categorisation
  - ⋆ very accurate when the job is done by experts
  - ⋆ consistent when the problem size and team are small
  - ⋆ difficult and expensive to scale
- Handcrafted rule-based systems
  - ⋆ e.g., assign a particular category if the document contains a given boolean combination of words
  - ⋆ accuracy is often very high if a rule has been carefully refined over time by a subject expert
  - ⋆ building and maintaining these rules is very expensive

Jackson and Moulinier (2002:pp119–125); Pantel and Lin (1998)

- Automatic classifiers
  - k-Nearest Neighbours (k-NN)
  - Naive Bayes
  - support vector machines
  - presuppose hand-classified seed data to generalise from

- In practise, commercial systems tend to use bits and pieces of all of these
  - Yahoo!, Google Directory (`dmoz.com`), Reuters, ...

Jackson and Moulinier (2002:pp119–125); Pantel and Lin (1998)

What are (Supervised) Classifiers?

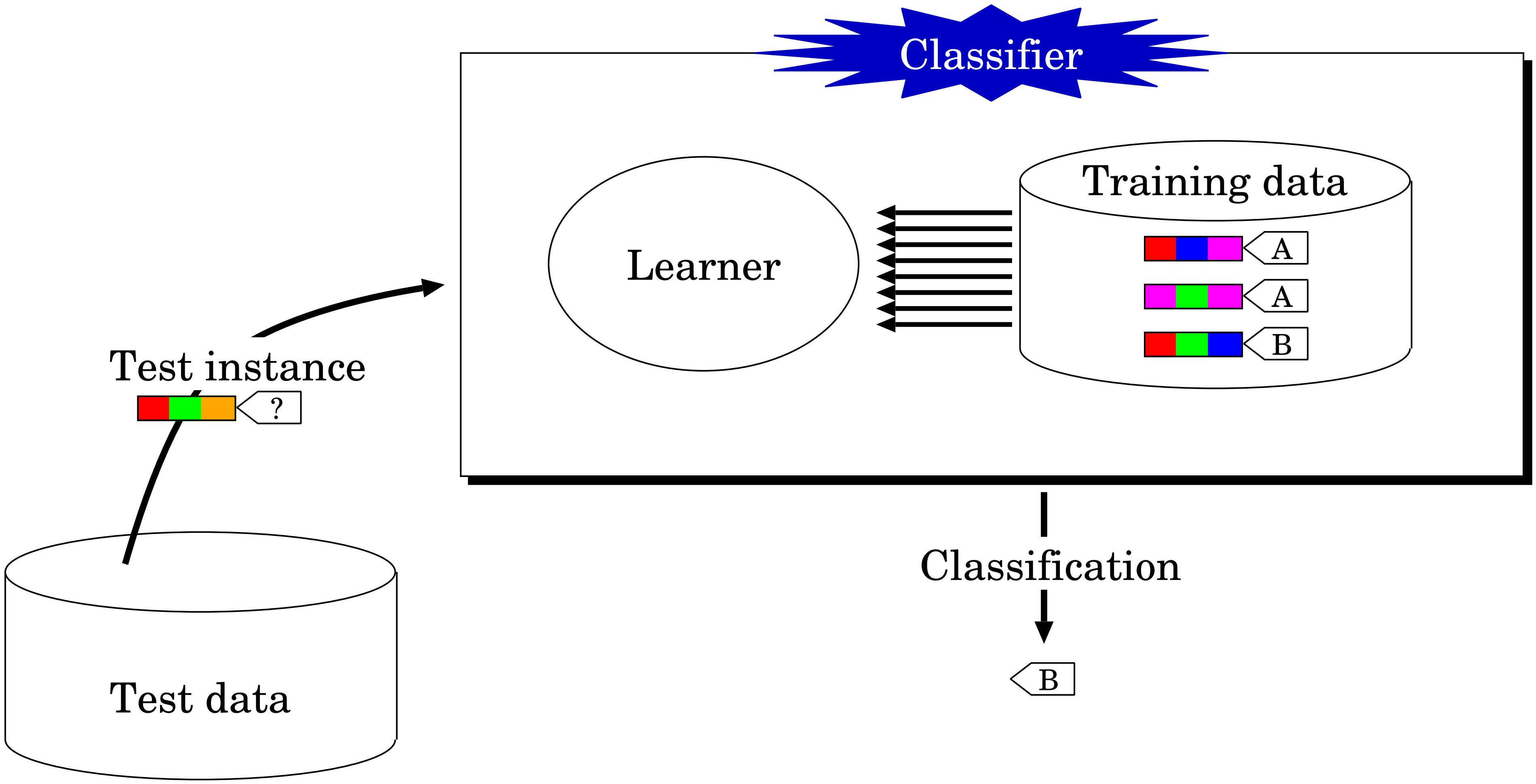
- Given:

  1. a fixed representation language of **attributes**
  2. a fixed set of pre-classified **training instances**
  3. a fixed set of classes C
  4. a "learner" algorithm which can identify patterns in the training instances

- Estimate:

  the category of a novel input $x : c(x) \in C$

Jackson and Moulinier (2002:pp127–8); Chakrabarti (2003:pp126–8)



Jackson and Moulinier (2002:pp127–8); Chakrabarti (2003:pp126–8)

Example Set-up

- Training data:

Outlook	Temperature	Humidity	Windy	Play
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
overcast	hot	high	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
overcast	cool	normal	TRUE	yes

- Test data:

Outlook	Temperature	Humidity	Windy	Play
sunny	mild	normal	TRUE	???

Supervision

- **Supervised** methods have prior knowledge of a closed set of classes and instances pre-classified according to those classes, and set out to categorise new instances according to those classes

- **Unsupervised** methods dynamically discover the classes in the process of categorising the instances [STRONG DEFINITION]

OR

- **Unsupervised** methods categorise instances without the aid of pre-classified data [WEAK DEFINITION]

Discussion: supervised or unsupervised?

- Given a set of web documents, identify obvious outliers for manual inspection

- From a set of web documents, filter out the spam documents based on a sample of manually-classified documents

- Classify a set of web documents as belonging to SCC, IN, OUT or OTHER

Discussion: supervised or unsupervised?

- Given a set of web documents, identify obvious outliers for manual inspection

  (strongly) unsupervised

- From a set of web documents, filter out the spam documents based on a sample of manually-classified documents

  supervised

- Classify a set of web documents as belonging to SCC, IN, OUT or OTHER

  (weakly) unsupervised

... But First some Basics

- Document representation
- Basics of probability theory
- Basics of entropy

Document Representation

id	*word*	*freq*
1	`a`	233
2	`aardvark`	0
3	`aback`	2
4	`abacus`	0

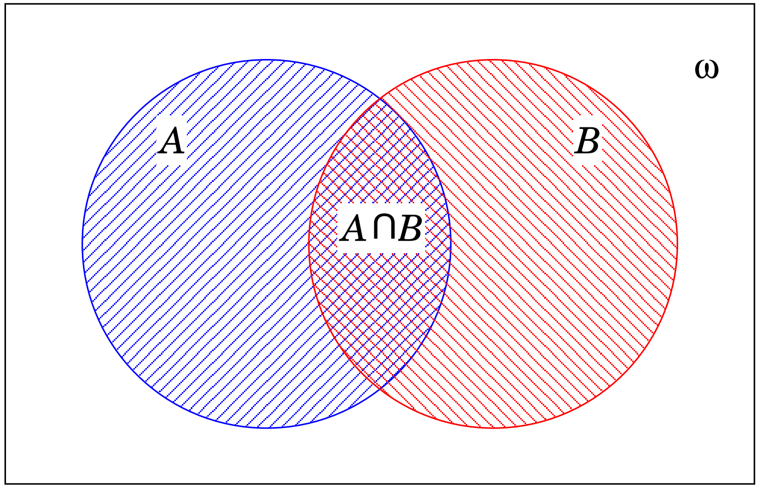
id	*word*	*freq*
5	`abalone`	0
6	`abandon`	1
7	`abandonment`	0
	⋮	

$$\vec{x} = \langle 233, 0, 2, 0, 0, 1, 0, \ldots \rangle$$

- In practice, we tend to weight each term frequency (see later), and also pre-normalise the document vector to unit length

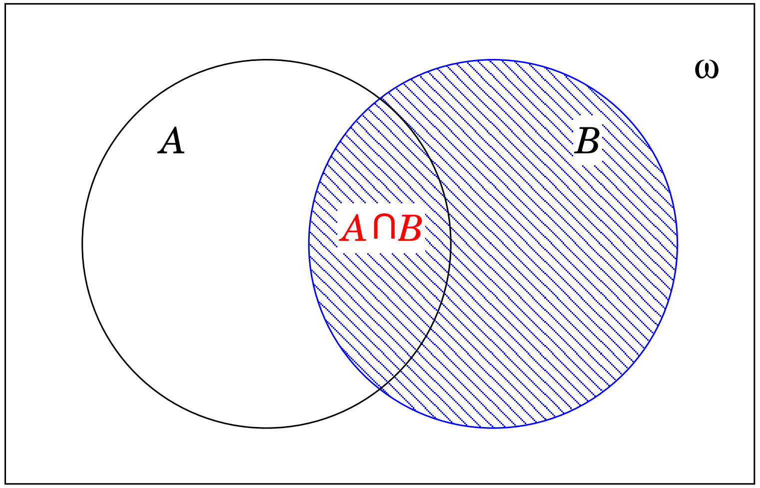
(Very) Basics of Probability Theory

- **Joint probability** $(P(A,B))$: the probability of both A and B occurring $= P(A \cap B)$



$$P(\texttt{ace},\texttt{heart}) = \frac{1}{52},\ P(\texttt{heart},\texttt{red}) = \frac{1}{4}$$

- **Conditional probability** $(P(A|B))$: the probability of A occurring given the occurrence of $B = \frac{P(A \cap B)}{P(B)}$



$P(\texttt{ace}|\texttt{heart}) = \frac{1}{13}$, $P(\texttt{heart}|\texttt{red}) = \frac{1}{2}$

- **Multiplication rule**: $P(A \cap B) = P(A|B)P(B) = P(B|A)P(A)$
- **Chain rule**: $P(A_1 \cap ... \cap A_n) = P(A_1)P(A_2|A_1)P(A_3|A_2 \cap A_1) ... P(A_n| \cap_{i=1}^{n-1} A_i)$
- **Prior probability** ($P(A)$): the probability of A occurring, given no additional knowledge about A
- **Posterior probability** ($P(A|B)$): the probability of A occurring, given background knowledge about event(s) B leading up to A
- **Independence:** A and B are independent iff $P(A \cap B) = P(A)P(B)$

Binomial Distributions

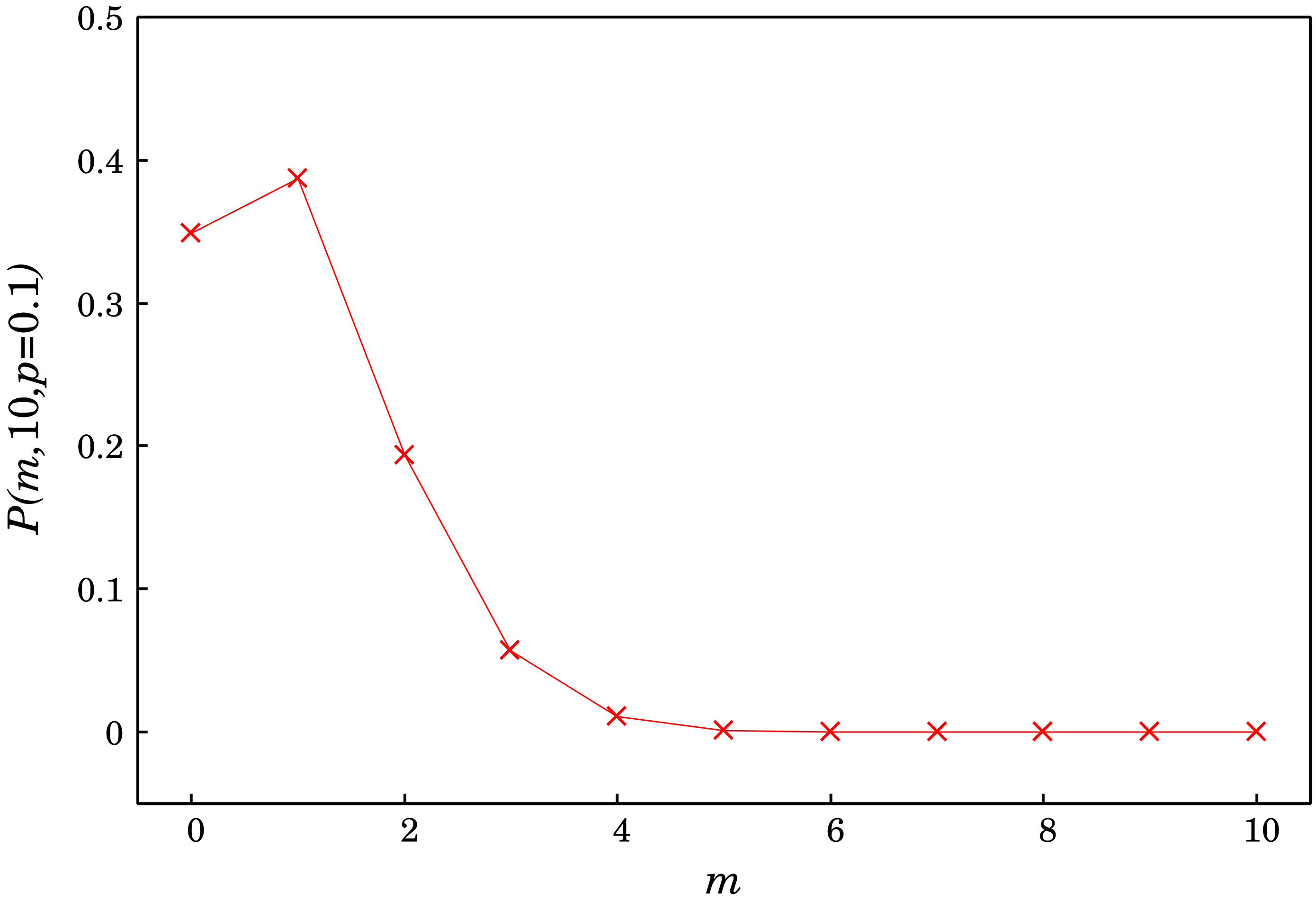
- A **binomial distribution** results from a series of independent trials with only two outcomes (i.e. **Bernoulli trials**)

  e.g. multiple coin tosses ($\langle H, T, H, H, ..., T \rangle$)

- The probability of an event with probability p occurring exactly m out of n times is given by

$$P(m, n, p) \;=\; \frac{n!}{m!(n-m)!} p^m (1-p)^{n-m}$$

Binomial Example: $P(m, 10, p = 0.1)$



Multinomial Distributions

- A **multinomial distribution** results from a series of independent trials with more than two outcomes

  e.g. balls in cricket ($\langle \cdot, \cdot, 1, \mathrm{out}_{\mathrm{LBW}}, ..., 4 \rangle$)

- The probability of events $X_1, X_2, ..., X_n$ with probabilities $p_1, p_2, ..., p_n$ occurring exactly $x_1, x_2, ..., x_n$ times, respectively, is given by

$$P(X_1 = x_1, X_2 = x_2, ..., X_n = x_n) = (\sum_i x_i)! \prod_i \frac{p_i^{x_i}}{x_i!}$$

Entropy

- Given a probability distribution, the information (in bits) required to predict an event is the distribution's **entropy** or **information value**

- The entropy of a discrete random event x with possible states $1, ..n$ is:

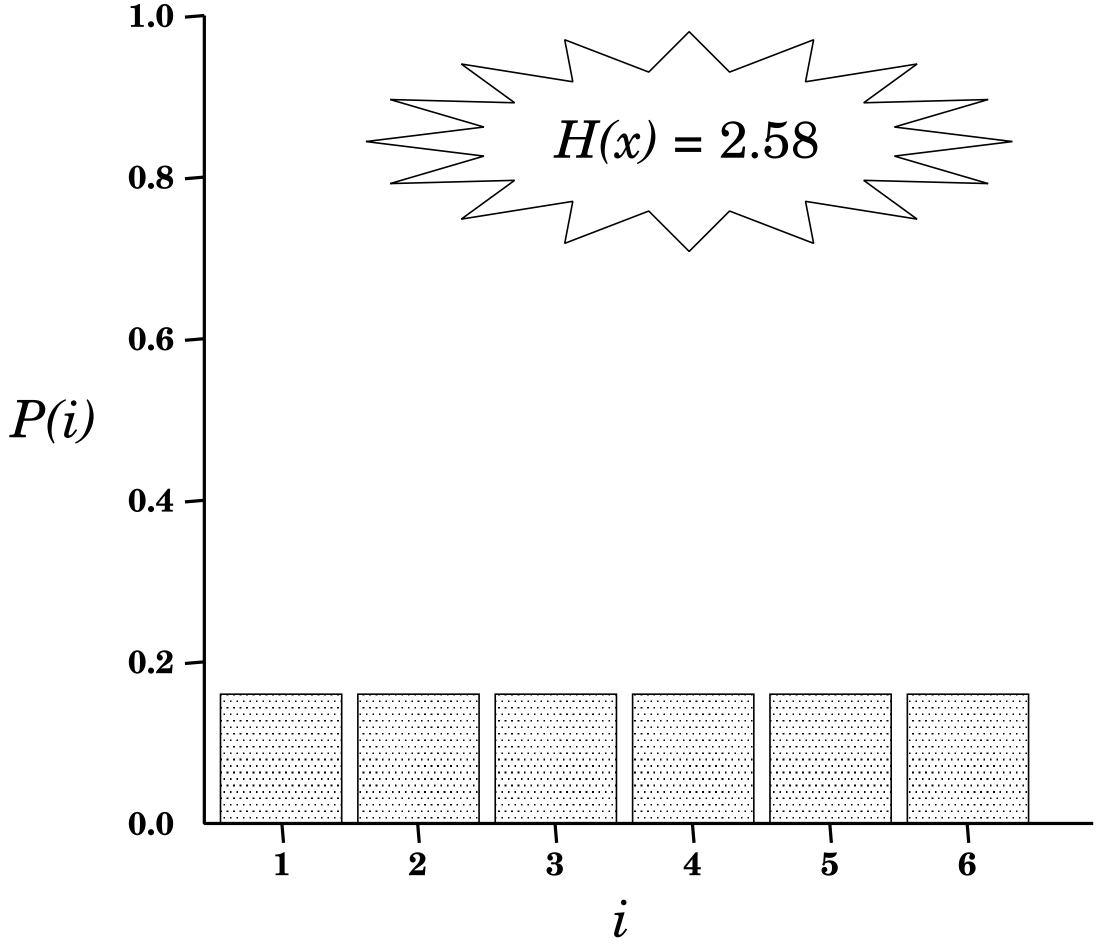
$$
\begin{aligned}
H(x) &= -\sum_{i=1}^{n} P(i) \log_2 P(i) \\
&= \frac{freq(*) \log_2(freq(*)) - \sum_{i=1}^{n} freq(i) \log_2(freq(i))}{freq(*)}
\end{aligned}
$$

where $0 \log_2 0 =^{def} 0$

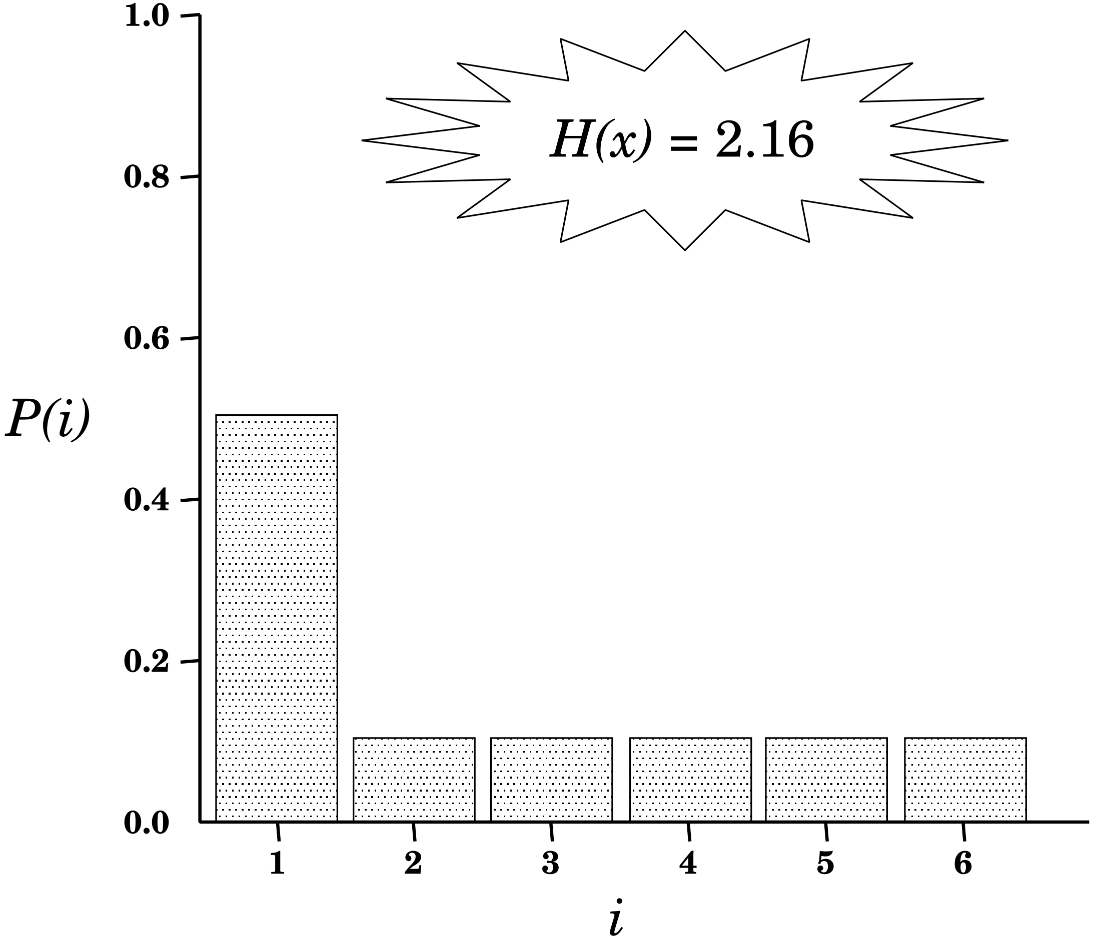
Interpreting Entropy Values

- A high entropy value means x is boring (uniform/flat)
- A low entropy value means x is varied ("peaky")

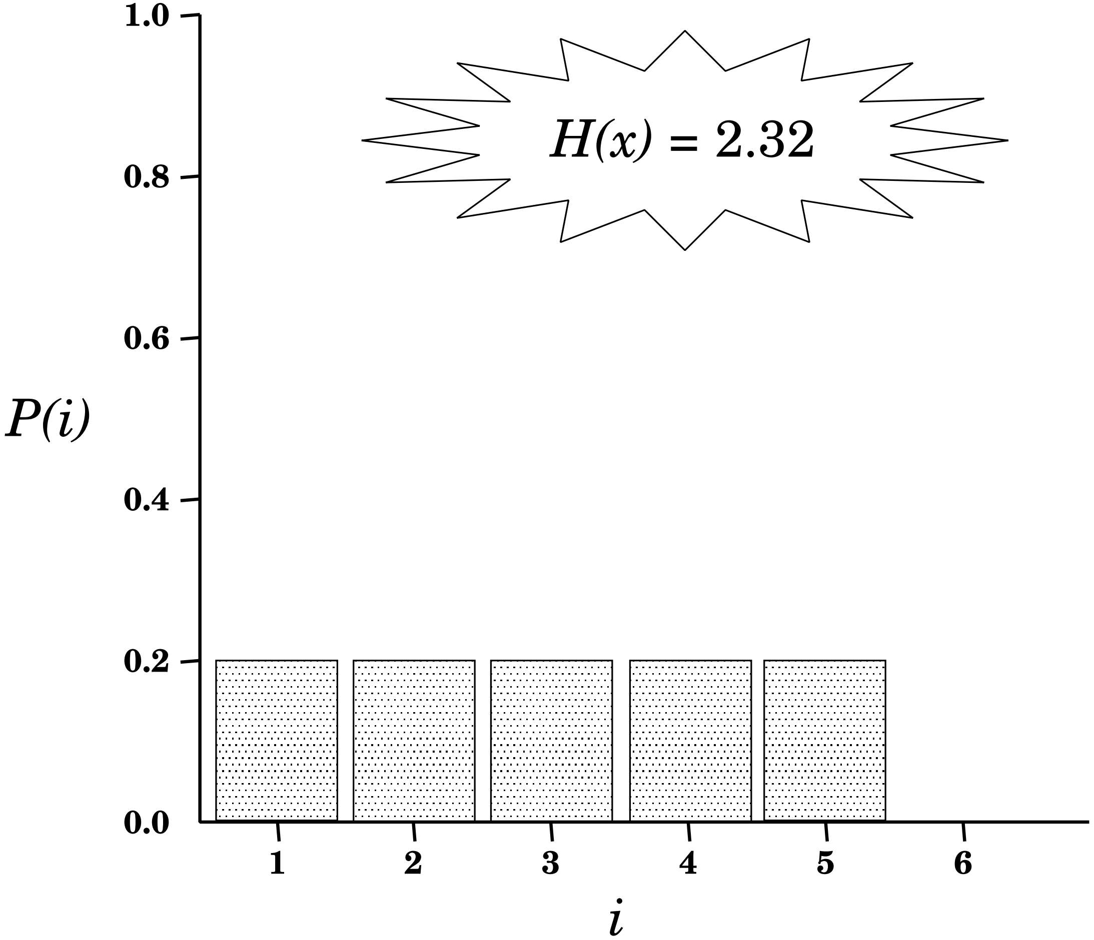
Entropy of Loaded Dice (1)



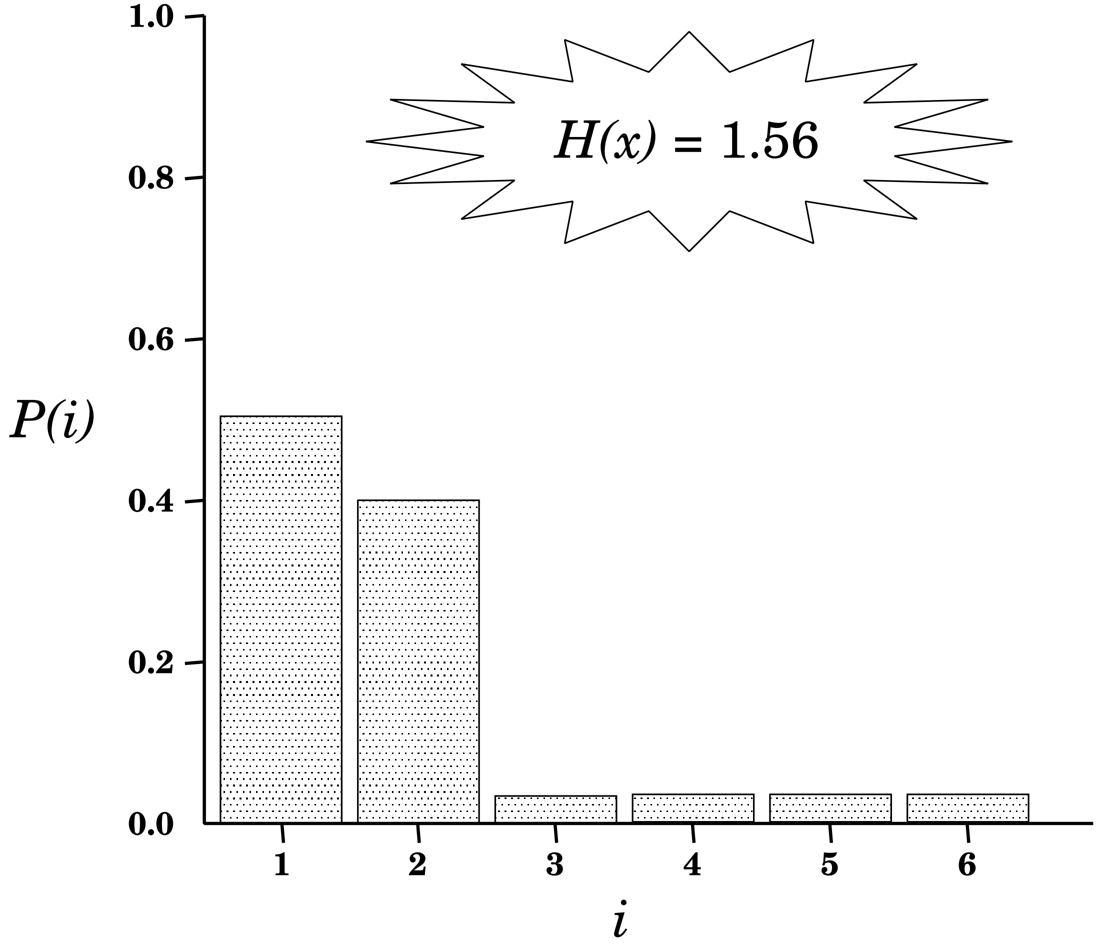
Entropy of Loaded Dice (2)



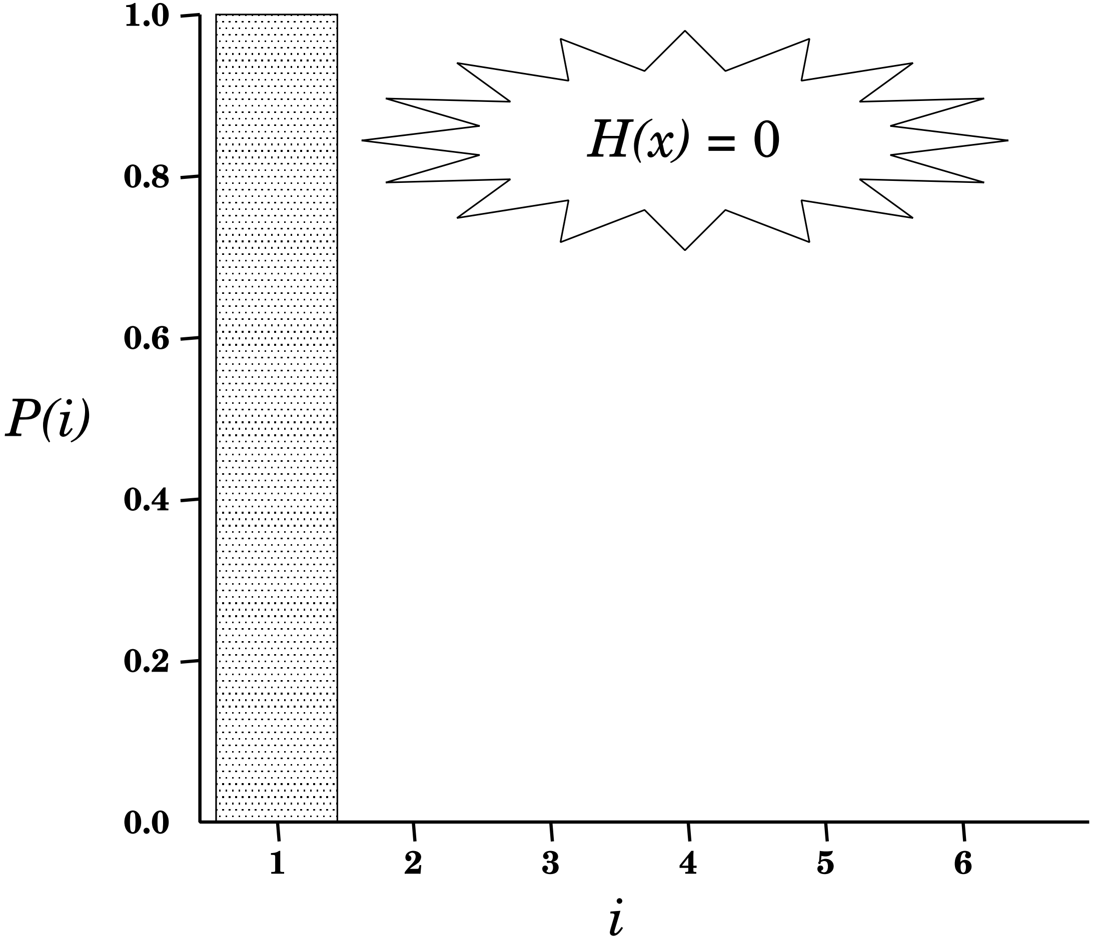
Entropy of Loaded Dice (3)



Entropy of Loaded Dice (4)



Entropy of Loaded Dice (5)



Estimating the Probabilities

- The most obvious way of generating the probabilities is via **maximum likelihood estimation** (MLE), using the frequency counts in the training data:

$$\hat{P}(c_j) = \frac{freq(c_j)}{\sum_k freq(c_k)}$$

$$\hat{P}(x_i|c_j) = \frac{freq(x_i, c_j)}{freq(c_j)}$$

- Based on this, our document representation would look something like:

$$\vec{x} = \langle 0.1, 0, 0.0002, 0, 0, 0.0001, 0, \ldots \rangle$$

Modelling Document Similarity: Cosine Similarity

- Given two documents x and y, and their corresponding feature vectors $\vec{x}$ and $\vec{y}$, respectively, we can calculate their similarity via their **vector cosine**:

$$sim(x,y) = \frac{\vec{x} \cdot \vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\sqrt{\sum_i y_i^2}}$$

Jackson and Moulinier (2002:pp32–6); Chakrabarti (2003:pp133–6)

Cosine Similarity Example

- Calculate the cosine similarity of the following documents:

A = | aardvark back abandon abandon abandon |

B = | aardvark abandonment back back back |

$$\vec{A} = \langle 1, 3, 0, 1 \rangle \equiv \langle \frac{1}{\sqrt{11}}, \frac{3}{\sqrt{11}}, 0, \frac{1}{\sqrt{11}} \rangle$$

$$\vec{B} = \langle 1, 0, 1, 3 \rangle \equiv \langle \frac{1}{\sqrt{11}}, 0, \frac{1}{\sqrt{11}}, \frac{3}{\sqrt{11}} \rangle$$

$$\vec{A} \cdot \vec{B} = \frac{\frac{1}{\sqrt{11}} \times \frac{1}{\sqrt{11}} + \frac{3}{\sqrt{11}} \times 0 + 0 \times \frac{1}{\sqrt{11}} + \frac{1}{\sqrt{11}} \times \frac{3}{\sqrt{11}}}{1 \times 1} = \frac{4}{11}$$

Jackson and Moulinier (2002:pp32–6); Chakrabarti (2003:pp133–6)

Modelling Document Distance: Relative Entropy

- Given two documents x and y, and their corresponding feature **unit-length** vectors $\vec{x}$ and $\vec{y}$, respectively, we can interpret the feature vector as a probability distribution and calculate the **relative entropy** (or KL divergence):

$$D(x \mid\mid y) = \sum_i x_i(\log_2 x_i - \log_2 y_i)$$

or alternatively **skew divergence**:

$$s_\alpha(x, y) = D(x \mid\mid \alpha y + (1 - \alpha)x)$$

- This causes considerable grief for our MLE-based probabilities: why?

- A simplistic way of getting around this is via **Laplacian smoothing**:

$$\hat{P}(c_j) = \frac{freq(c_j) + 1}{k + \sum_k freq(c_k)}$$

$$\hat{P}(x_i|c_j) = \frac{freq(x_i, c_j) + 1}{freq(c_j) + l}$$

Relative Entropy Example

- Calculate the relative entropy and skew divergence of the following documents:

aardvark back abandon abandon abandon	aardvark abandonment back back back

$$\mathbf{A} = \langle \frac{2}{9}, \frac{4}{9}, \frac{1}{9}, \frac{2}{9} \rangle \qquad \mathbf{B} = \langle \frac{2}{9}, \frac{1}{9}, \frac{2}{9}, \frac{4}{9} \rangle$$

$$D(\mathbf{A}||\mathbf{B}) = \sum_i a_i (\log_2 a_i - \log_2 b_i)$$

$$
\begin{aligned}
= \;& \frac{2}{9}(\log\frac{2}{9} - \log\frac{2}{9}) + \frac{4}{9}(\log\frac{4}{9} - \log\frac{1}{9}) + \\
& \frac{1}{9}(\log\frac{1}{9} - \log\frac{2}{9}) + \frac{2}{9}(\log\frac{2}{9} - \log\frac{4}{9}) \\
\approx \;& 0.56
\end{aligned}
$$

Skew Divergence Example

- Calculate the relative entropy and skew divergence of the following documents:

aardvark back abandon abandon abandon	aardvark abandonment back back back

$$\mathbf{A} = \langle 0.2, 0.6, 0.0, 0.2 \rangle \qquad \mathbf{B} = \langle 0.2, 0.0, 0.2, 0.6 \rangle$$

$$\begin{aligned} s_{0.99}(\mathbf{A}, \mathbf{B}) &= D(\mathbf{A} \mid\mid 0.99\mathbf{A} + 0.01\mathbf{B}) \\ &= \sum_i a_i(\log a_i - \log(0.99b_i + 0.01a_i)) \end{aligned}$$

$$
\begin{aligned}
&= 0.2(\log 0.2 - \log(0.99 \times 0.2 + 0.01 \times 0.2)) + \\
&\quad 0.6(\log 0.6 - \log(0.99 \times 0.0 + 0.01 \times 0.6)) + \\
&\quad 0.0(\log 0.0 - \log(0.99 \times 0.2 + 0.01 \times 0.0)) + \\
&\quad 0.2(\log 0.2 - \log(0.99 \times 0.6 + 0.01 \times 0.2)) \\
&\approx 3.67
\end{aligned}
$$

Nearest Neighbour Classifiers

- There are various ways to combine these document–document scores to form an overall categorisation function, e.g.:

- **Method 1:** index all training documents, and query the training document set with each test document; classify the test document according to the class of the top-ranked training document **[1-NN]**

- **Method 2:** index all training documents, and query the training document set with each test document; classify the test document according to the **majority class** within the k top-ranked training documents **[k-NN]**

Jackson and Moulinier (2002:pp148–9); Chakrabarti (2003:pp133–6)

- **Method 3:** index all training documents, and query the training document set with each test document; classify the test document according to the class with the best accumulative score **[weighted k-NN]**

- **Method 4:** index all training documents, and query the training document set with each test document; classify the test document according to the class with the best accumulative score based on scores, factoring in an offset to indicate the prior expectation of a test document being classified as being a member of that class **[offset weighted k-NN]**

Jackson and Moulinier (2002:pp148–9); Chakrabarti (2003:pp133–6)

- Overall advantages of the nearest neighbour approach:
  - ⋆ simple
- Overall disadvantages of the nearest neighbour approach:
  - ⋆ expensive (in terms of index accesses)
  - ⋆ everything is done at run time (**lazy learner**)
  - ⋆ prone to bias
  - ⋆ arbitrary k value

Jackson and Moulinier (2002:pp148–9); Chakrabarti (2003:pp133–6)

Feature Selection

Jackson and Moulinier (2002:pp148–9); Chakrabarti (2003:pp133–6)

Feature Selection

- Classes will tend to have medium-frequency membership, suggesting that we need some extra mechanism of identifying the terms which best discriminate the classes

  $\rightarrow$ enter **feature selection**

- We will focus on **greedy inclusion algorithms** for feature selection:

  rank terms in descending order of class discrimination, and select the top N features

Chakrabarti (2003:pp136–47)

Feature Selection via MI

- **Mutual information** (MI) provides an information-based estimate of the (in)dependence of two discrete random variables T (term) and C (class):

$$MI(T,C) = \sum_{t \in \{0,1\}} \sum_{c} P(t,c) \log \frac{P(t,c)}{P(t)P(c)}$$

- If T and C are independent, $MI(T,C) = 0$
- If T and C are positively correlated, $MI(T,C) > 0$
- If T and C are negatively correlated, $MI(T,C) < 0$

Chakrabarti (2003:pp139–40)

- We select our N "best" features by taking the terms T with the highest MI value
- The method is greedy in that it doesn't take term inter-dependence into consideration
- Bias towards rare uninformative terms
- Example features (on 20 Newsgroups):

  `sci.electronics`: circuit, voltage, amp, ground, copy, battery, electronics, cooling, ...
  `rec.autos`: car, cars, engine, ford, dealer, mustang, oil, collision, autos, tires, toyota, ...

Chakrabarti (2003:pp139–40)

Mutual Information Example

- Perform feature selection over a document collection with the following characteristics:

Term	*Class A*	*Class B*	*Total*
a	1,000	1,000	2,000
aardvark	30	50	80
aback	10	20	30
abacus	100	0	10
TOTAL	1,000	1,000	2,000

$$
\begin{array}{ll}
P(A) = \frac{1000}{2000} & P(B) = \frac{1000}{2000} \\
P(aback) = \frac{30}{2000} & P(\overline{aback}) = \frac{1970}{2000} \\
P(aback, A) = \frac{10}{2000} & P(\overline{aback}, A) = \frac{990}{2000} \\
P(aback, B) = \frac{20}{2000} & P(\overline{aback}, B) = \frac{980}{2000}
\end{array}
$$

$$
\begin{aligned}
MI(aback) &= \frac{990}{2000} \log \frac{\frac{990}{2000}}{\frac{1970}{2000}\frac{1000}{2000}} + \frac{980}{2000} \log \frac{\frac{980}{2000}}{\frac{1970}{2000}\frac{1000}{2000}} \\
&\quad + \frac{10}{2000} \log \frac{\frac{10}{2000}}{\frac{30}{2000}\frac{1000}{2000}} + \frac{20}{2000} \log \frac{\frac{20}{2000}}{\frac{30}{2000}\frac{1000}{2000}} \\
&= 0.0009
\end{aligned}
$$

Feature Selection via χ^2

- χ^2 ("kai-square") provides an estimate of the level of statistical significance of the correlation between two discrete random variables T (term) and C (class):

	Term present	***Term absent***
$\boldsymbol{class} = c_i$	W	X
$\boldsymbol{class} \neq c_i$	Y	Z

$$\chi^2 = \frac{N(WZ - XY)^2}{(W+X)(W+Y)(X+Z)(Y+Z)}$$

Chakrabarti (2003:pp138–9)

- The higher the value of χ^2, the less confident we are of T and c_i being independent

- We select our N "best" features by taking the terms T with the highest χ^2 value

- Bias towards frequent uninformative terms

χ^2 Example

- Perform feature selection over a document collection with the following characteristics:

$Term$	$Class\ A$	$Class\ B$	$Total$
a	1,000	1,000	2,000
$aardvark$	30	50	80
$aback$	10	20	30
$abacus$	100	0	10
TOTAL	1,000	1,000	2,000

Chakrabarti (2003:pp138–9)

	$aback$	$\overline{aback}$
A	10	990
$\bar{A}$	20	980

$$\chi^2 = \frac{2000(10 \times 980 - 990 \times 20)^2}{(10 + 990)(10 + 20)(990 + 980)(20 + 980)}$$
$$= 3.38$$

Chakrabarti (2003:pp138–9)

Summary

- What are the basic methods of text categorisation?
- What is the different between supervised and unsupervised classifiers?
- How does the k-nearest neighbour method operate, and what are some of the variants on the original algorithm?
- What different methods are used to calculate document similarity?
- How can we perform feature selection?

References

CHAKRABARTI, SOUMEN. 2003. *Mining the Web: Discovering Knowledge from Hypertext Data*. San Francisco, USA: Morgan Kaufmann.

JACKSON, PETER, and ISABELLE MOULINIER. 2002. *Natural Language Processing for Online Applications: Text Retrieval, Extraction and Categorization*. Amsterdam, Netherlands: John Benjamins.

MCCALLUM, ANDREW, and KAMAL NIGAM. 1998. A comparison of event models for Naive Bayes text classification. In *Proc. of the AAAI-98 Workshop on Learning for Text Categorization*, Madison, USA.

PANTEL, PATRICK, and DEKANG LIN. 1998. SpamCop: A spam classification & organization program. In *Proc. of the 1998 AAAI Workshop on Learning for Text Categorization*, Madison, USA.