

# Bayesian text processing

**Keith Briggs**

Keith.Briggs@bt.com

`research.btexact.com/teralab/keithbriggs.html`



CRG meeting 2005 Jan 24 1500

`bayes-2005jan24.tex` TYPESET 2005 JANUARY 24 16:55 IN PDF $\text{\LaTeX}$  ON A LINUX SYSTEM

# Outline

- ★ some problem in text analysis■
- ★ probability theory■
- ★ Bayesian ideas■
- ★ some 'solutions'■

The aim:

to determine how well Bayesian methods work

# Typical problems to be tackled

- ★ anomaly and fraud detection. . .
- ★ and in general to any situation where one has a collection of 'normal' and 'abnormal' documents, log files etc. ■
- ★ The aim is to classify an unknown document as normal or abnormal. ■
- ★ Another use is automatic correction of scanned documents converted to text by optical character recognition. ■
- ★ This is especially challenging when the document contains mixed languages

# Language recognition

★ is amazingly easy:

- ▷ *Zeichen* ■
- ▷ *Teich* ■
- ▷ *étang* ■
- ▷ *raftan* ■
- ▷ *stagnum* ■
- ▷ *piccolo* ■
- ▷ *ddydd* ■
- ▷ *æftercweðan* ■
- ▷ *riðja* ■
- ▷ *négy* ■

★ . . . but what information are we using when we do this? ■

★ and how well can we do it when there are errors?

# Probability vs. degree of belief

★  $P(\text{event}) \equiv \lim_{n \rightarrow \infty} \frac{\#\text{events}}{n}$

- ▷ *objective*
- ▷ *must be able to repeat the experiment indefinitely*
- ▷ *rate of convergence of limit unspecified*
- ▷ *strictly speaking, this rules out using this definition in the real world*

★ 'degree of belief'  $B$  is more or less subjective

- ▷ *meaningful for a single, non-repeatable event*
- ▷ *your  $B$  might not be the same as my  $B$*
- ▷ *chance of rain tomorrow*
- ▷ *chance of horse winning a race*
- ▷ *spamminess of an email*

# Probability theory

## ★ conditional probability

$$P(x = a|y = b) \equiv \frac{P(x = a, y = b)}{P(y = b)} \blacksquare$$

## ★ product rule

$$P(x, y|\mathcal{H}) = P(x|y, \mathcal{H})P(y|\mathcal{H}) = P(y|x, \mathcal{H})P(x|\mathcal{H}) \blacksquare$$

## ★ marginalization

$$\begin{aligned} P(x|\mathcal{H}) &= \sum_y P(x, y|\mathcal{H}) \\ &= \sum_y P(x|y, \mathcal{H})P(y|\mathcal{H}) \end{aligned}$$

# Bayes' theorem

★ Bayes' theorem - is just the product rule:

$$P(y|x, \mathcal{H}) = \frac{P(x|y, \mathcal{H})P(y|\mathcal{H})}{P(x|\mathcal{H})}$$



★ . . . with  $y$  interpreted as the data  $D$ , and  $x$  interpreted as parameters  $\theta$ :

$$P(\theta|D, \mathcal{H}) = \frac{P(D|\theta, \mathcal{H})P(\theta|\mathcal{H})}{P(D|\mathcal{H})}$$

# Prior, likelihood and posterior

- ★ we can think of Bayes' rule as:

$$\text{posterior} \propto \text{likelihood} * \text{prior}$$

- ★ for example, a single bit  $s$  sent twice over a noisy channel, received as  $r_1 r_2$ :

- ▷  $P(s = 1 | r_1 r_2) = P(r_1 r_2 | s = 1) P(s = 1) / P(r_1 r_2)$

- ▷  $P(s = 0 | r_1 r_2) = P(r_1 r_2 | s = 0) P(s = 0) / P(r_1 r_2)$

- ★ that is, your **prior** (degree of belief before you observed that data  $r$ ), is **updated** by the information the data provides about the value of  $s$  (the likelihood), to provide your **posterior** degree of belief



# Text classification theory

- ★ could be based on various choices of *features*: words, or  $n$ -grams ■
- ★ corpora  $C_1, C_2, \dots, C_k$  ■
- ★ priors  $\pi_1, \pi_2, \dots, \pi_k$  ■
- ★ models  $\mathbb{P}_{C_1}, \mathbb{P}_{C_2}, \dots, \mathbb{P}_{C_k}$  ■
- ★ if  $x$  is an unknown document, the posterior probability that  $x$  belongs to  $C_j$  is  $P(C_j|x) \propto \mathbb{P}_{C_j} \pi_j$  ■
- ★ decision rule: choose  $j$  to maximize  $P(C_j|x)$  ■

# Digram measure

★ word  $w = w_1w_2 \dots w_k$  ■

★ reference measure  $R_C(w) \equiv p_C(\wedge, w_1)p_C(w_1, w_2) \dots p_C(w_k, \$)$  ■

▷ *this is naïve - it assumes adjacent digrams are statistically independent*

■

★ Dirichlet digram measure  $p_C(u, v) = \frac{\#(v|u)}{\sum_r \#(r|u)} \frac{+ \alpha \mu(v)}{+ \alpha}$  ■

★  $\alpha$  is a hyperparameter, and the optimum  $\alpha$  should be chosen from tests on various corpora

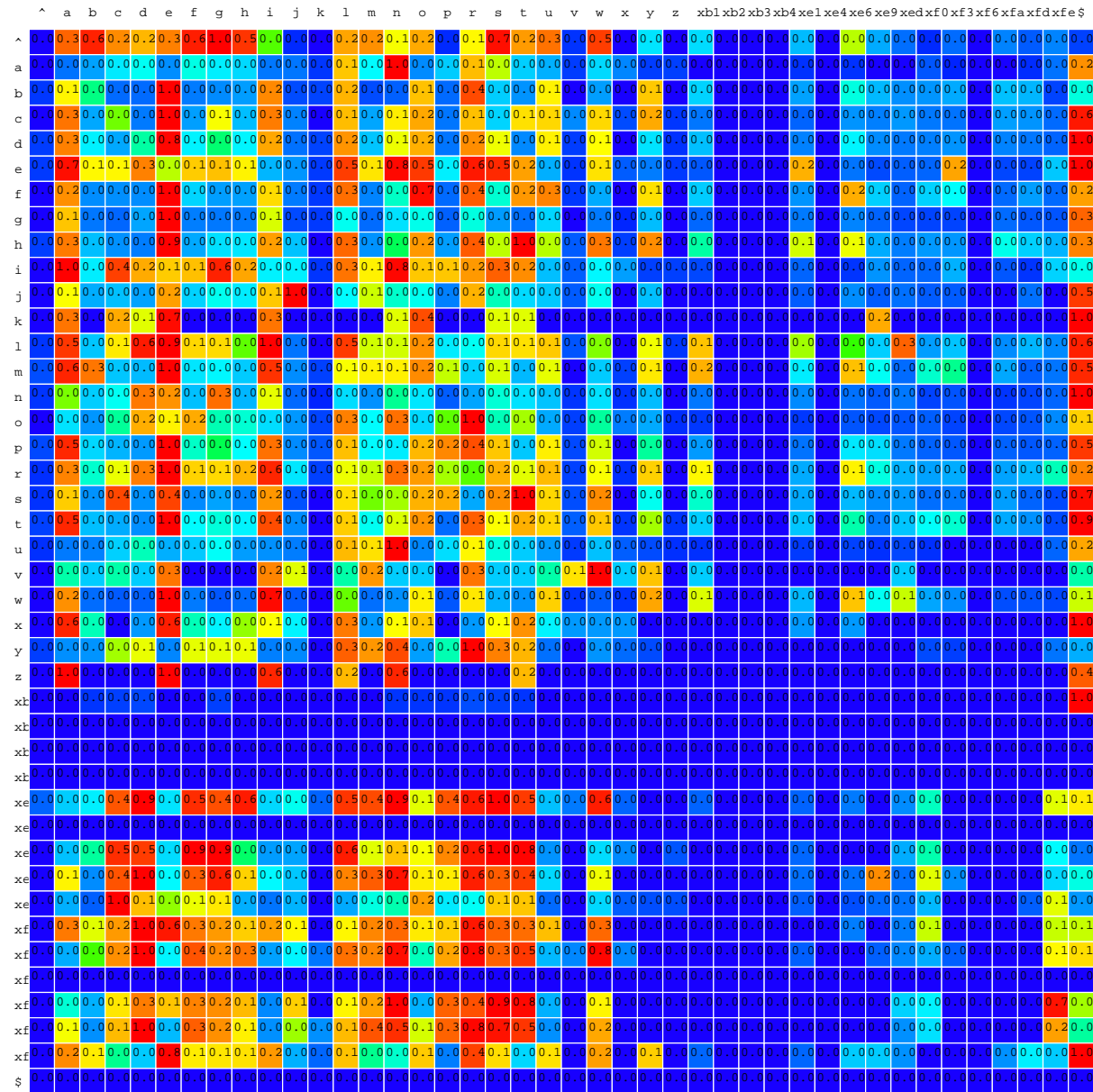
# Coding issues

- ★ Only two fixed-width choices - ASCII (1 byte) or Unicode (2 bytes) ■
- ★  $\text{\TeX}$  or html are possible, but not fixed-width ■
- ★ Unfortunately, ASCII cannot do all characters used in OE or Icelandic ■
- ★ Therefore, I moved some characters to unneeded ascii positions ■
  - ▷ e.g. hex b1 (really the  $\pm$  sign) for  $\bar{æ}$

# Training

- ★ Collect texts ■
- ★ split into words; check for obvious errors; fix punctuation and capitalization ■
- ★ Count trigrams and estimate  $\alpha$  ■

# Example digram measure for Old English



# Example digram measure for Latin

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	§
^	0.000	0.545	0.045	0.468	0.225	0.449	0.169	0.070	0.117	0.459	0.140	0.255	0.221	0.147	0.450	0.232	0.158	0.393	0.183	0.103	0.164	0.009	0.000
a	0.000	0.000	0.094	0.125	0.141	0.274	0.003	0.058	0.004	0.011	0.118	0.258	0.283	0.001	0.046	0.018	0.242	0.129	0.398	0.104	0.047	0.011	0.405
b	0.000	0.084	0.000	0.000	0.002	0.082	0.000	0.000	0.000	0.080	0.013	0.000	0.002	0.017	0.000	0.000	0.026	0.017	0.005	0.135	0.003	0.000	0.033
c	0.000	0.154	0.000	0.040	0.000	0.154	0.000	0.000	0.013	0.213	0.026	0.000	0.002	0.224	0.000	0.001	0.041	0.000	0.115	0.165	0.000	0.003	0.078
d	0.000	0.066	0.000	0.002	0.008	0.215	0.006	0.001	0.003	0.261	0.002	0.003	0.001	0.057	0.002	0.002	0.010	0.007	0.000	0.090	0.014	0.000	0.128
e	0.000	0.050	0.064	0.132	0.098	0.002	0.028	0.068	0.004	0.049	0.118	0.243	0.377	0.045	0.053	0.053	0.652	0.327	0.341	0.024	0.025	0.103	0.700
f	0.000	0.042	0.000	0.000	0.000	0.060	0.014	0.000	0.000	0.054	0.024	0.000	0.000	0.024	0.000	0.000	0.019	0.000	0.000	0.035	0.000	0.000	0.000
g	0.000	0.051	0.000	0.000	0.000	0.075	0.000	0.002	0.000	0.080	0.007	0.003	0.060	0.012	0.000	0.000	0.042	0.000	0.000	0.033	0.000	0.000	0.002
h	0.000	0.050	0.000	0.000	0.000	0.026	0.000	0.000	0.000	0.047	0.000	0.001	0.000	0.051	0.000	0.000	0.006	0.000	0.000	0.013	0.000	0.000	0.003
i	0.000	0.246	0.162	0.147	0.131	0.084	0.010	0.057	0.006	0.071	0.121	0.165	0.502	0.187	0.065	0.029	0.058	0.467	0.411	0.214	0.039	0.010	0.375
l	0.000	0.145	0.003	0.005	0.001	0.126	0.000	0.006	0.000	0.284	0.111	0.002	0.003	0.086	0.004	0.000	0.000	0.008	0.047	0.090	0.010	0.005	0.024
m	0.000	0.166	0.012	0.000	0.001	0.118	0.001	0.000	0.000	0.166	0.000	0.023	0.039	0.103	0.077	0.035	0.000	0.001	0.000	0.082	0.005	0.000	0.835
n	0.000	0.140	0.000	0.077	0.111	0.249	0.019	0.036	0.001	0.275	0.006	0.002	0.021	0.152	0.004	0.011	0.003	0.123	0.445	0.109	0.014	0.002	0.190
o	0.000	0.003	0.041	0.081	0.066	0.023	0.011	0.021	0.009	0.004	0.069	0.115	0.310	0.001	0.076	0.017	0.286	0.190	0.047	0.001	0.027	0.015	0.299
p	0.000	0.107	0.000	0.000	0.000	0.202	0.000	0.000	0.014	0.094	0.041	0.000	0.000	0.114	0.047	0.000	0.178	0.027	0.039	0.063	0.000	0.000	0.001
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.478	0.000	0.000	0.000
r	0.000	0.317	0.029	0.043	0.016	0.455	0.009	0.018	0.006	0.376	0.000	0.044	0.025	0.183	0.011	0.006	0.041	0.040	0.104	0.186	0.019	0.000	0.200
s	0.000	0.109	0.000	0.057	0.006	0.233	0.001	0.001	0.000	0.219	0.000	0.002	0.001	0.050	0.047	0.042	0.000	0.143	0.221	0.175	0.001	0.000	1.000
t	0.000	0.310	0.000	0.000	0.000	0.358	0.000	0.000	0.021	0.442	0.001	0.000	0.000	0.128	0.000	0.032	0.151	0.001	0.023	0.410	0.001	0.000	0.631
u	0.000	0.134	0.048	0.046	0.068	0.217	0.003	0.030	0.000	0.172	0.159	0.547	0.143	0.092	0.033	0.001	0.212	0.434	0.116	0.014	0.007	0.016	0.048
v	0.000	0.043	0.001	0.000	0.000	0.134	0.000	0.000	0.000	0.157	0.000	0.001	0.000	0.030	0.000	0.000	0.000	0.001	0.000	0.012	0.000	0.000	0.004
x	0.000	0.006	0.000	0.007	0.000	0.019	0.000	0.000	0.001	0.043	0.001	0.000	0.000	0.007	0.010	0.000	0.000	0.003	0.016	0.006	0.007	0.013	0.046
§	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

# Latin trigrams - a..

	^	a	b	c	d	e	f	g	h	i	l	m	n	o	p	q	r	s	t	u	v	x	
^	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
a	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
b	0.000	0.007	0.000	0.000	0.045	0.116	0.000	0.000	0.004	0.042	0.015	0.000	0.029	0.029	0.000	0.000	0.032	0.181	0.000	0.017	0.000	0.000	0.000
c	0.000	0.012	0.004	0.674	0.000	0.071	0.000	0.000	0.042	0.111	0.000	0.001	0.000	0.002	0.000	0.003	0.085	0.000	0.081	0.009	0.000	0.000	0.000
d	0.000	0.025	0.000	0.024	0.241	0.196	0.199	0.043	0.093	0.254	0.051	0.125	0.034	0.058	0.074	0.006	0.020	0.254	0.011	0.157	0.443	0.000	0.000
e	0.000	0.003	0.000	0.002	0.170	0.001	0.001	0.129	0.000	0.000	0.009	0.064	0.013	0.014	0.001	0.137	0.041	0.151	0.183	0.000	0.017	0.001	0.000
f	0.000	0.001	0.000	0.000	0.000	0.002	0.003	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.086	0.000	0.002	0.010	0.000	0.000	0.000
g	0.000	0.012	0.000	0.000	0.000	0.122	0.000	0.050	0.000	0.063	0.002	0.097	0.008	0.000	0.000	0.000	0.387	0.000	0.000	0.009	0.000	0.000	0.000
h	0.000	0.000	0.000	0.000	0.000	0.023	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
i	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.014	0.000	0.000	0.000	0.000
l	0.000	0.047	0.062	0.004	0.000	0.065	0.003	0.001	0.000	0.917	0.046	0.001	0.000	0.007	0.047	0.000	0.000	0.004	0.345	0.016	0.043	0.000	0.000
m	0.000	0.027	0.238	0.000	0.000	0.009	0.000	0.000	0.000	0.264	0.000	0.000	0.204	0.100	0.163	0.000	0.000	0.000	0.000	0.029	0.000	0.000	0.000
n	0.000	0.029	0.000	0.051	0.033	0.006	0.004	0.089	0.005	0.331	0.000	0.000	0.367	0.000	0.000	0.004	0.000	0.005	0.601	0.006	0.000	0.020	0.000
o	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.005	0.000	0.000	0.001	0.000	0.000	0.000
p	0.000	0.011	0.000	0.000	0.000	0.089	0.000	0.000	0.011	0.028	0.000	0.000	0.000	0.049	0.333	0.000	0.025	0.013	0.015	0.343	0.000	0.000	0.000
q	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.283	0.001	0.000	0.000
r	0.000	0.106	0.164	0.101	0.070	0.060	0.002	0.117	0.000	0.114	0.000	0.422	0.006	0.002	0.007	0.000	0.045	0.050	0.198	0.006	0.042	0.004	0.000
s	0.000	0.002	0.001	0.036	0.000	0.000	0.000	0.000	0.000	0.130	0.000	0.000	0.000	0.003	0.116	0.000	0.000	0.020	0.043	0.000	0.000	0.000	0.000
t	0.000	0.006	0.000	0.000	0.000	0.015	0.000	0.000	0.030	0.017	0.021	0.000	0.000	0.000	0.000	0.763	0.089	0.000	0.155	0.005	0.000	0.000	0.000
u	0.000	0.001	0.000	0.219	0.229	0.000	0.022	0.336	0.000	0.000	0.028	0.000	0.001	0.000	0.000	0.000	0.080	0.105	1.000	0.000	0.000	0.179	0.000
v	0.000	0.040	0.000	0.000	0.000	0.047	0.000	0.001	0.000	0.077	0.000	0.000	0.000	0.015	0.000	0.000	0.000	0.001	0.000	0.022	0.000	0.000	0.000
x	0.000	0.002	0.000	0.000	0.000	0.002	0.000	0.000	0.000	0.004	0.000	0.000	0.000	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

# Latin verbs

Fourth conjugation, indicative voice, active mood

PRESENT	<b>audio</b>	<b>audis</b>	<b>audit</b>	<b>audimus</b>	<b>auditis</b>	<b>audiunt</b>
PREFECT	<b>audivi</b>	<b>audivisti</b>	<b>audivit</b>	<b>audivimus</b>	<b>audivistis</b>	<b>audiverunt</b>
IMPERFECT	<b>audiebam</b>	<b>audiebas</b>	<b>audiebat</b>	<b>audiebamus</b>	<b>audiebatis</b>	<b>audiebant</b>
PLUPERFECT	<b>audiveram</b>	<b>audiveras</b>	<b>audiverat</b>	<b>audiveramus</b>	<b>audiveratis</b>	<b>audiverant</b>
FUTURE	<b>audiam</b>	<b>audies</b>	<b>audiet</b>	<b>audiemus</b>	<b>audietis</b>	<b>audient</b>
FUTURE PERFECT	<b>audivero</b>	<b>audiveris</b>	<b>audiverit</b>	<b>audiverimus</b>	<b>audiveritis</b>	<b>audiverint</b>

subjunctive

PRESENT	<b>audiam</b>	<b>audias</b>	<b>audiat</b>	<b>audiamus</b>	<b>audiatis</b>	<b>audiant</b>
PERFECT	<b>audiverim</b>	<b>audiveris</b>	<b>audiverit</b>	<b>audiverimus</b>	<b>audiveritis</b>	<b>audiverint</b>
IMPERFECT	<b>audirem</b>	<b>audires</b>	<b>audiret</b>	<b>audiremus</b>	<b>audiretis</b>	<b>audirent</b>
PLUPERFECT	<b>audivissem</b>	<b>audivisses</b>	<b>audivisset</b>	<b>audivissemus</b>	<b>audivissetis</b>	<b>audivissent</b>



# Spelling correction

- ★ Idea: keep a list of common errors (perhaps with priors) ■
- ★ Try all corrections and sort them by likelihood ■
- ★ Give the users a list of the few most likely to select from ■
- ★ Could use heuristics: likelihood 'jumps' ■

# Screenshot

```
kbriggs@sodium:~/Latin
conslet
44.62=constet(1)
deflutt
47.35=defluu(2) 48.04=defluti(1) 48.50=defluit(1)
dominns
45.38=dominus(1)
epismpus
51.04=epiampus(1) 51.31=eptampus(2) 52.03=epiampua(2) 52.30=eptampua(3) 53.79=eplampus(2) 54.78=eplam
pua(3) 55.38=episcapus(2) 55.43=episcopus(2)
galesre
46.74=galtare(2) 46.92=gultare(3) 47.32=galtart(3) 47.46=gateare(2) 47.50=gultart(4) 48.04=gateart(3
) 48.23=gattare(3) 48.41=galeare(1)
inlerposili
56.00=interpositi(2)
inter
35.34=inter(0) 37.51=tuter(2) 37.87=initr(2) 38.27=infer(1)
jniss
33.66=quis(2)
lantum
39.69=tantum(1)
lerrae
39.73=terrae(1)
man's
37.51=maris(2)
montinm
42.26=manumm(4) 42.95=mantium(2) 43.18=manitum(4) 43.38=monumm(3) 43.90=monitum(3) 44.06=manuum(4) 4
4.08=montium(1)
neque
36.60=neque(1)
opporlunilatam
66.51=oppartumitatem(4) 66.97=oppartunitatem(3) 67.24=opportunitatem(3) 67.28=appartumitatem(5) 67.6
9=opportunitatem(2)
out
31.88=aut(1)
patibuium
```