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 pdfLaTeX 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

- $P(\text{event}) \equiv \lim_{n\to\infty} \frac{\#\text{events}}{n}$
  - *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

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

★ product rule

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

★ 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 θ:

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

Prior, likelihood and posterior

- we can think of Bayes' rule as:

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

- 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 $C_1, C_2, \ldots, C_k$
- priors $\pi_1, \pi_2, \ldots, \pi_k$
- models $\mathbb{P}_{C_1}, \mathbb{P}_{C_2}, \ldots, \mathbb{P}_{C_k}$
- 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 $w = w_1 w_2 \ldots w_k$
- reference measure $R_C(w) \equiv p_C(\wedge, w_1) p_C(w_1, w_2) \ldots p_C(w_k, \$)$
  - *this is naïve - it assumes adjacent digrams are statistically independent*
- Dirichlet digram measure $p_C(u, v) = \frac{\#(v|u) + \alpha\mu(v)}{\sum_r \#(r|u) + \alpha}$
- α is a hyperparameter, and the optimum α should be chosen from tests on various corpora

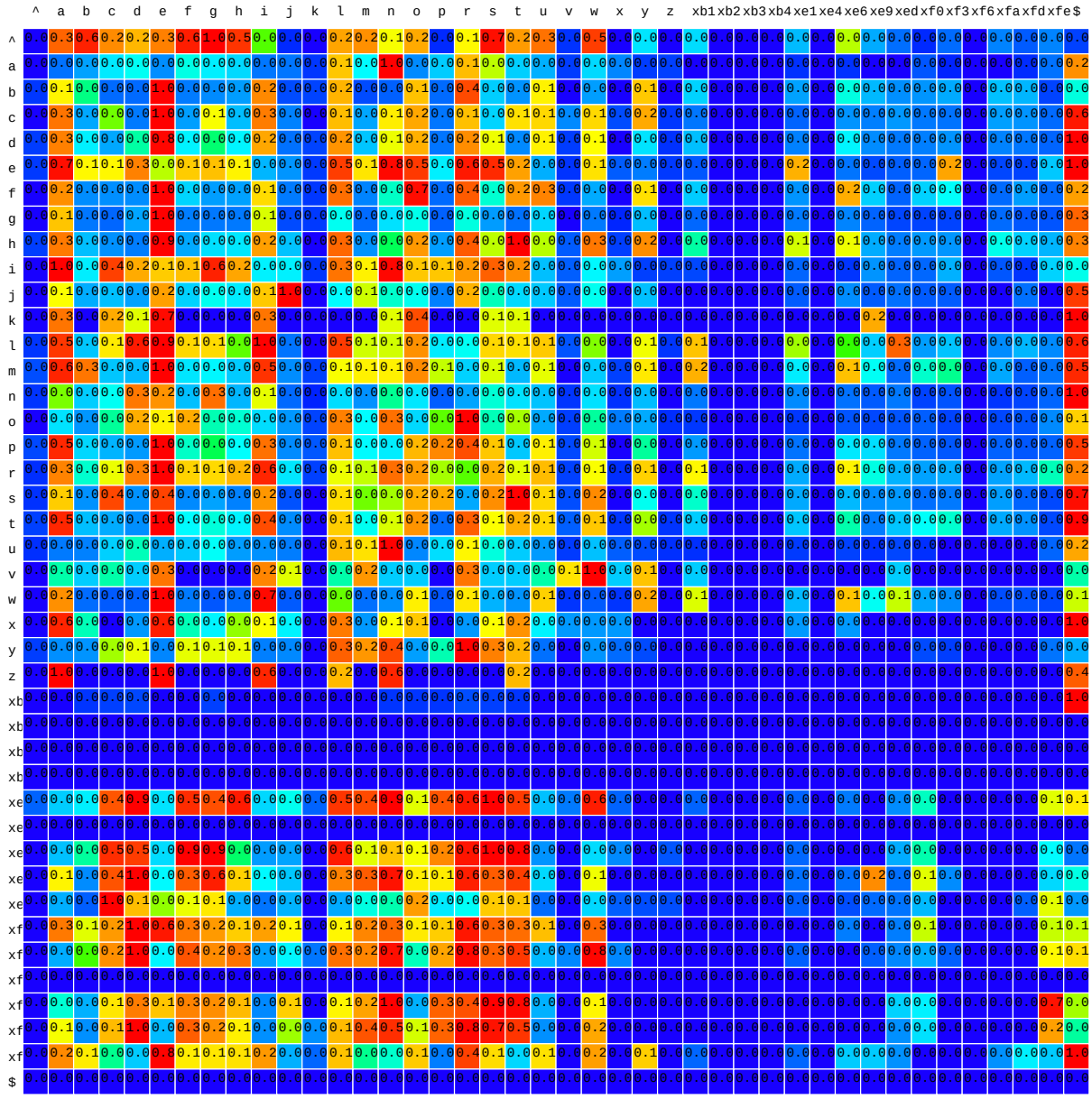
Coding issues

- Only two fixed-width choices - ASCII (1 byte) or Unicode (2 bytes)
- 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* œ

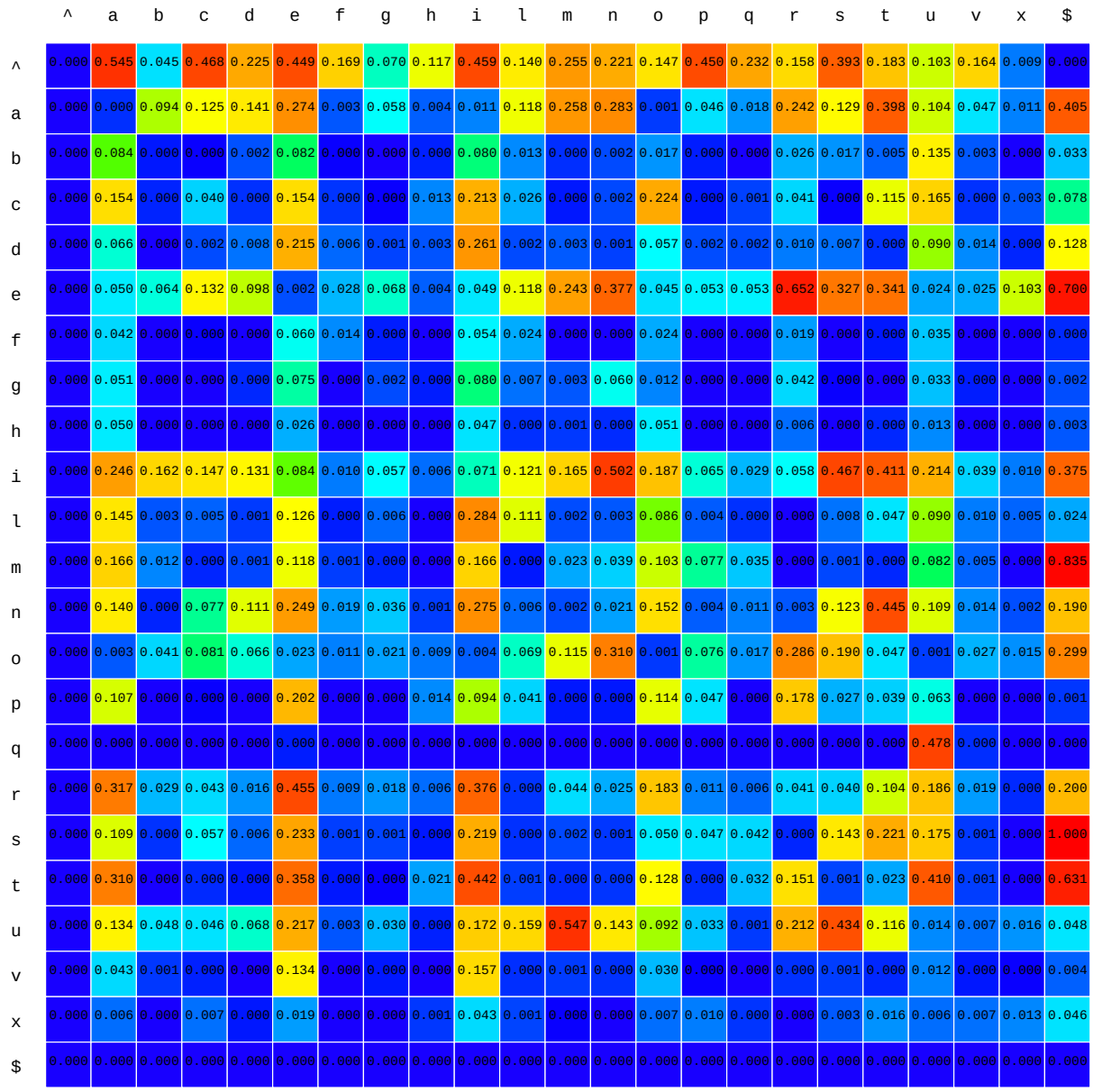
Training

- Collect texts
- split into words; check for obvious errors; fix punctuation and capitalization
- Count trigrams and estimate α

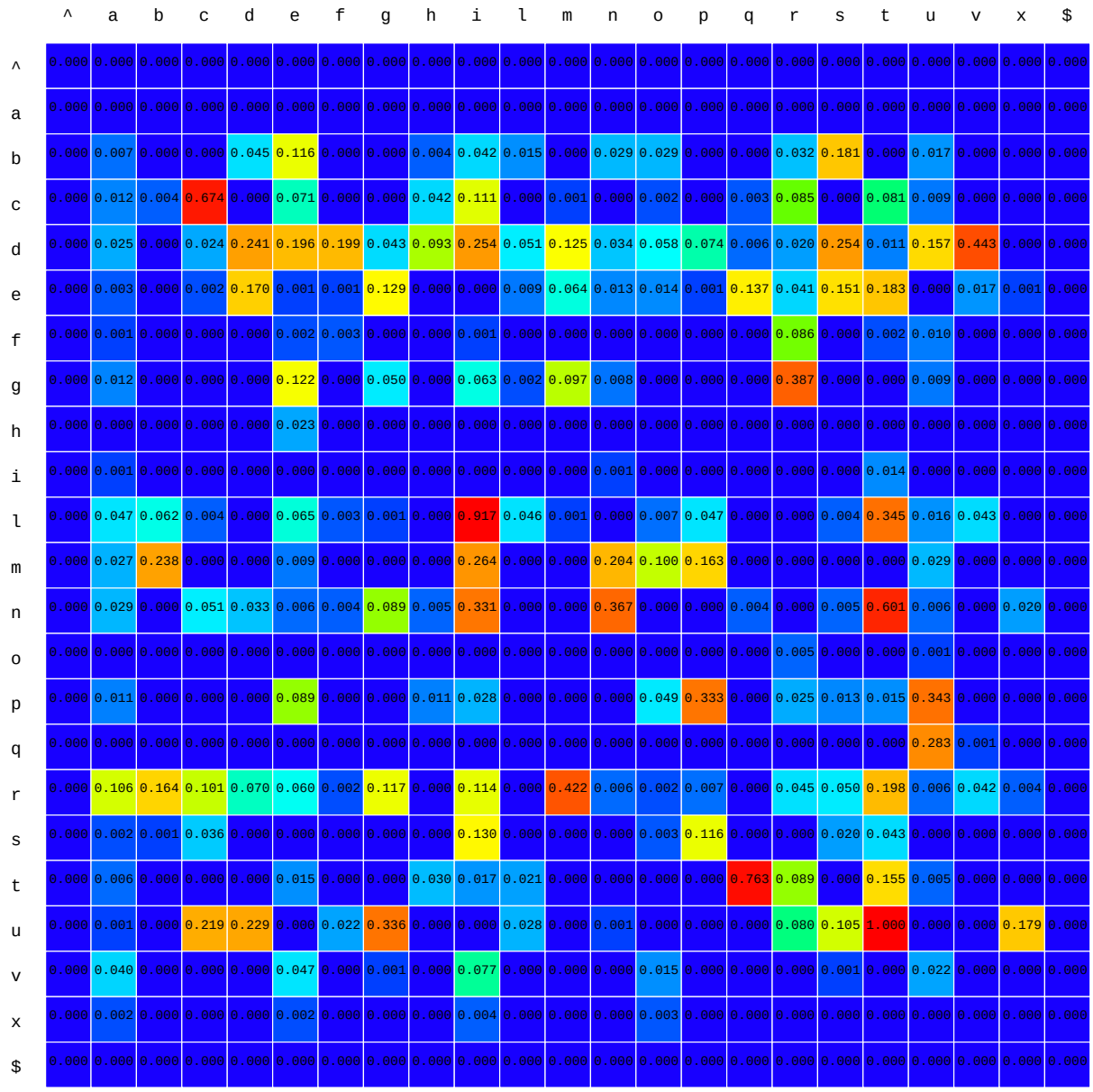
Example digram measure for Old English



Example digram measure for Latin



Latin trigrams - a..



Latin verbs

Fourth conjugation, indicative voice, active mood

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

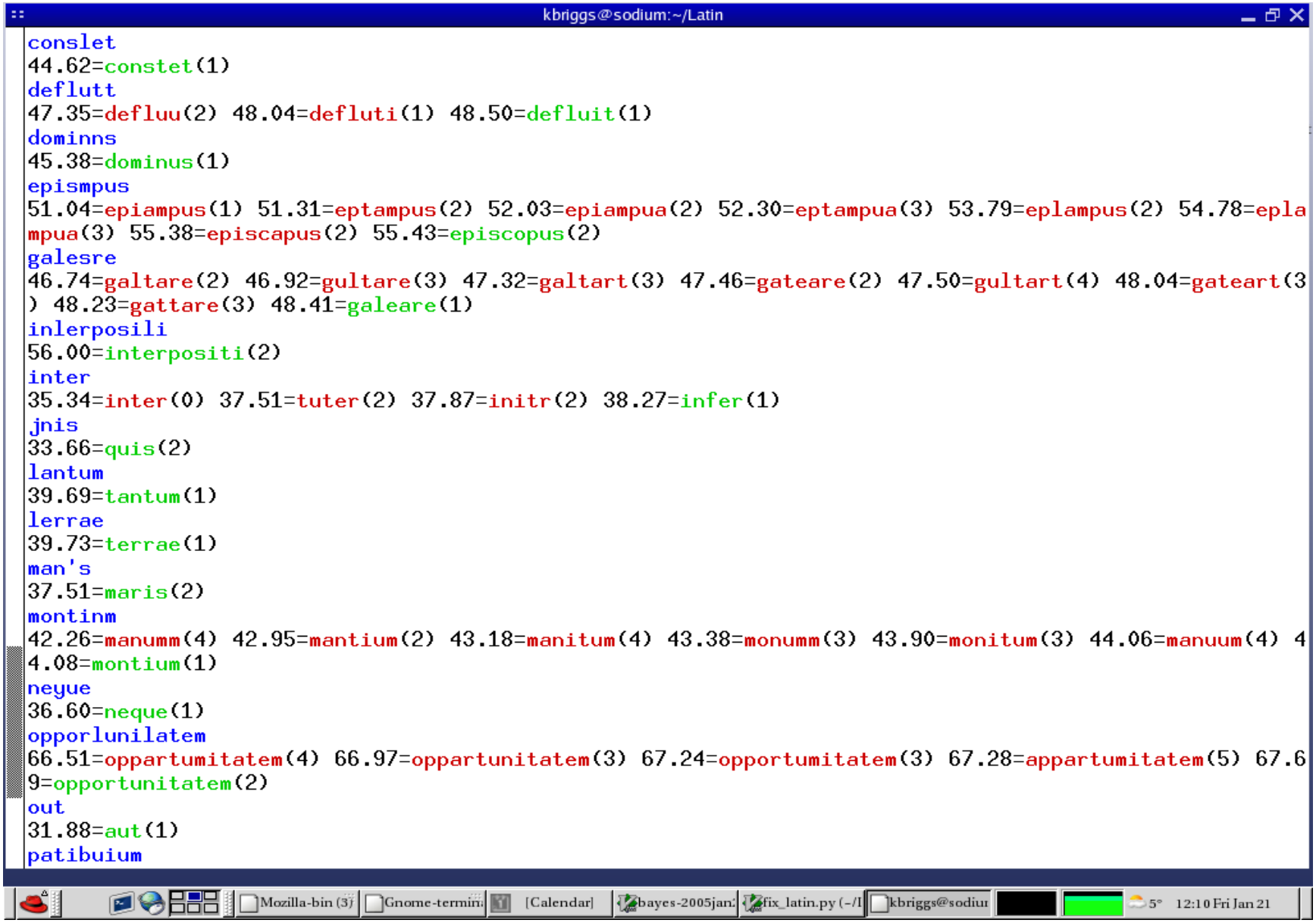
subjunctive

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

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=epla
mpua(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)
jnis
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)
neyue
36.60=neque(1)
opporlunilatem
66.51=oppartumitatem(4) 66.97=oppartunitatem(3) 67.24=opportumitatem(3) 67.28=appartumitatem(5) 67.6
9=opportunitatem(2)
out
31.88=aut(1)
patibuium
Mozilla-bin (3)
Gnome-termini
[Calendar]
bayes-2005jan:
fix_latin.py (~/l
kbriggs@sodiu
5°
12:10 Fri Jan 21