

English and Latin digram and trigram frequencies

Keith Briggs Keith.Briggs@bt.com

research.btexact.com/teralab/keithbriggs.html

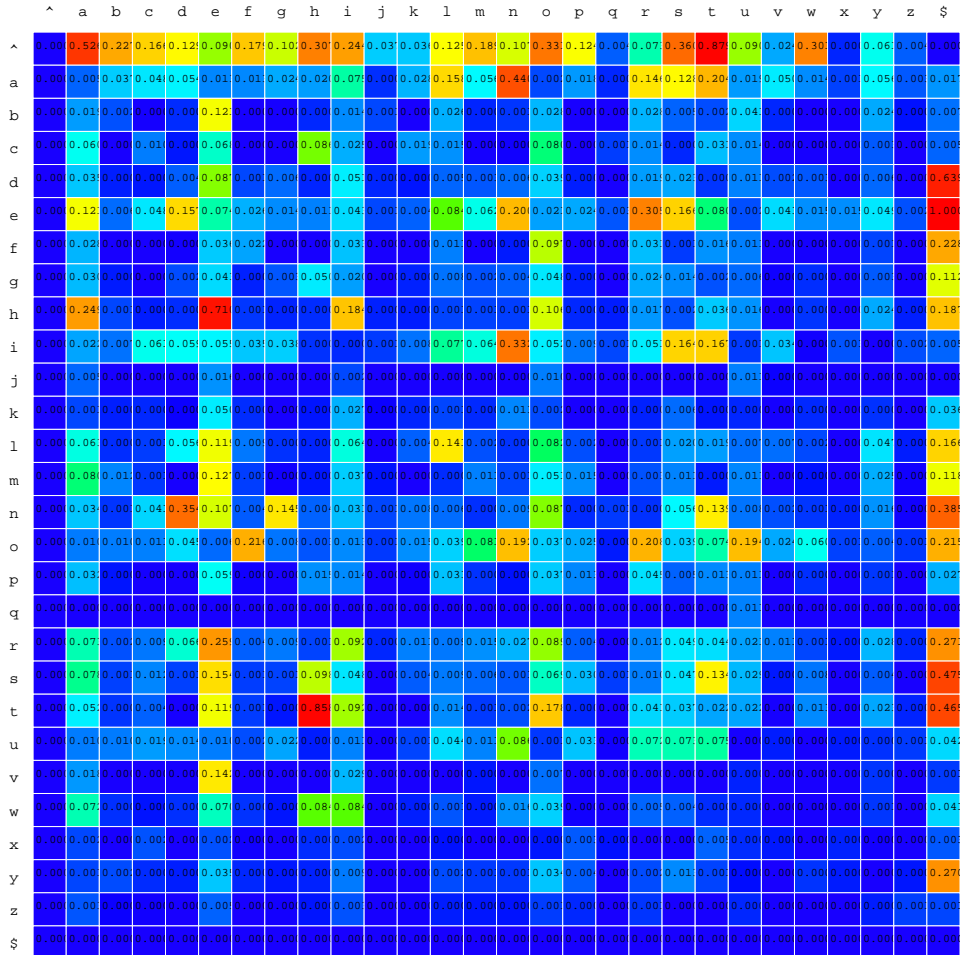
2004 November 11 10:50

TYPESET IN PDF \LaTeX ON A LINUX SYSTEM

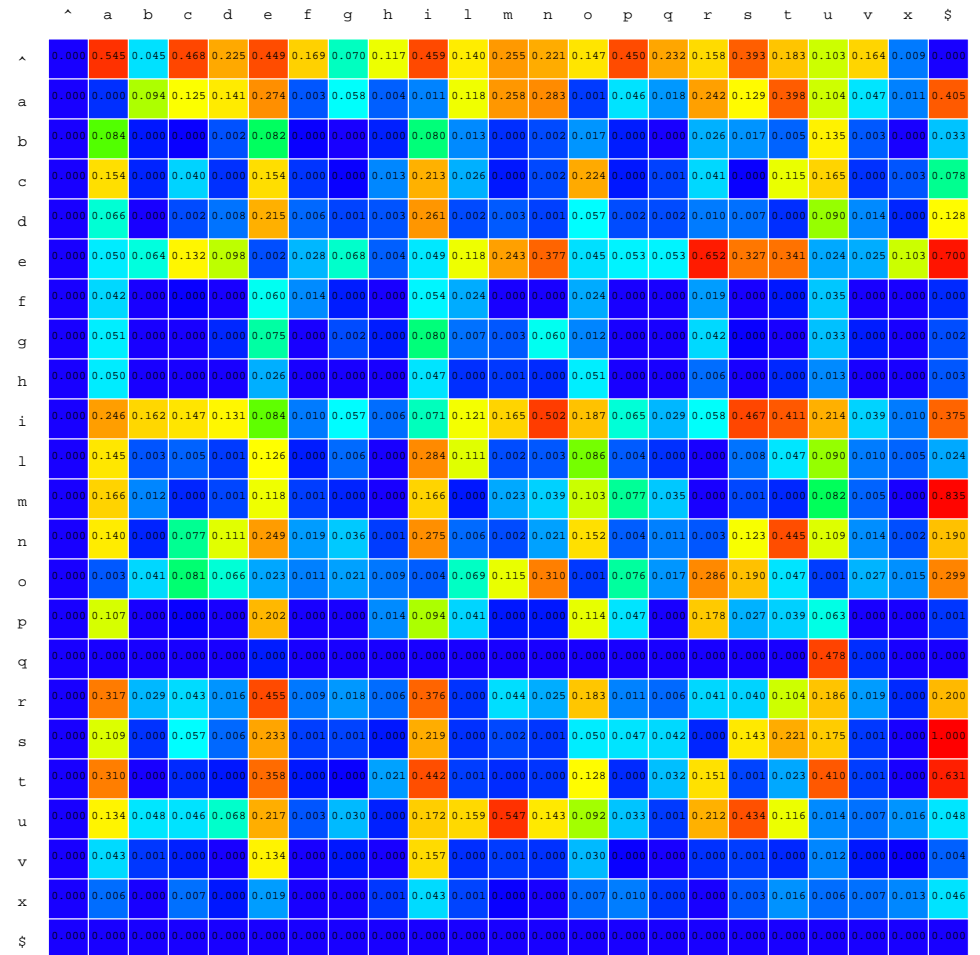
Introduction

- I wanted and look at distributions of letter pairs and triples in English and Latin as part of a project for OCR correction and language recognition
- I wrote a code to count letter pairs, including ^ for start-of-word and \$ for end-of-word
- The rows top-to-bottom correspond to the first character and the columns left-to-right to the second character
- I scaled all cells to a maximum of 1, so the number is proportional the observed frequency of pairs. (The most frequent pair is s\$.)
- The colour is roughly spectral with blue=0, red=1
- For trigram figures I counted occurrences of the 2 characters following the specified one
- in all cases the left figure is for English and the right for Latin

Digrams

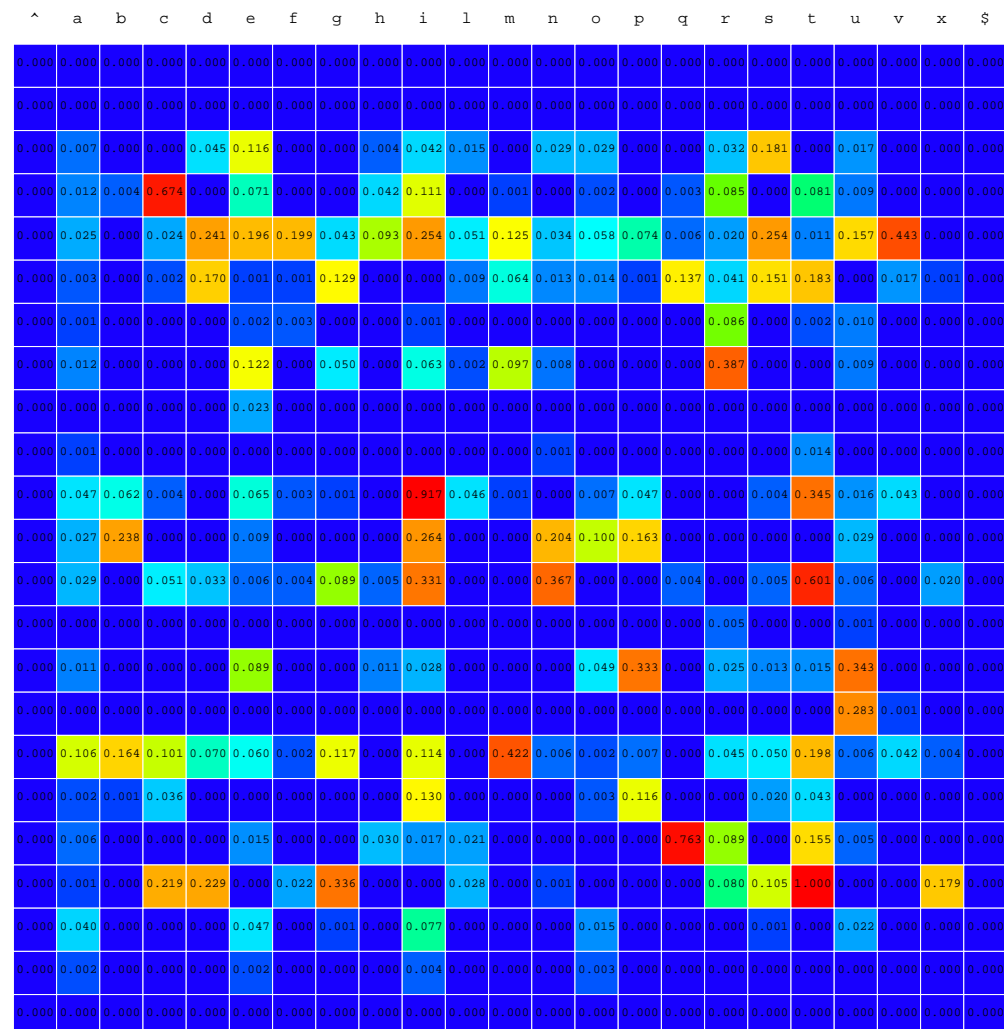
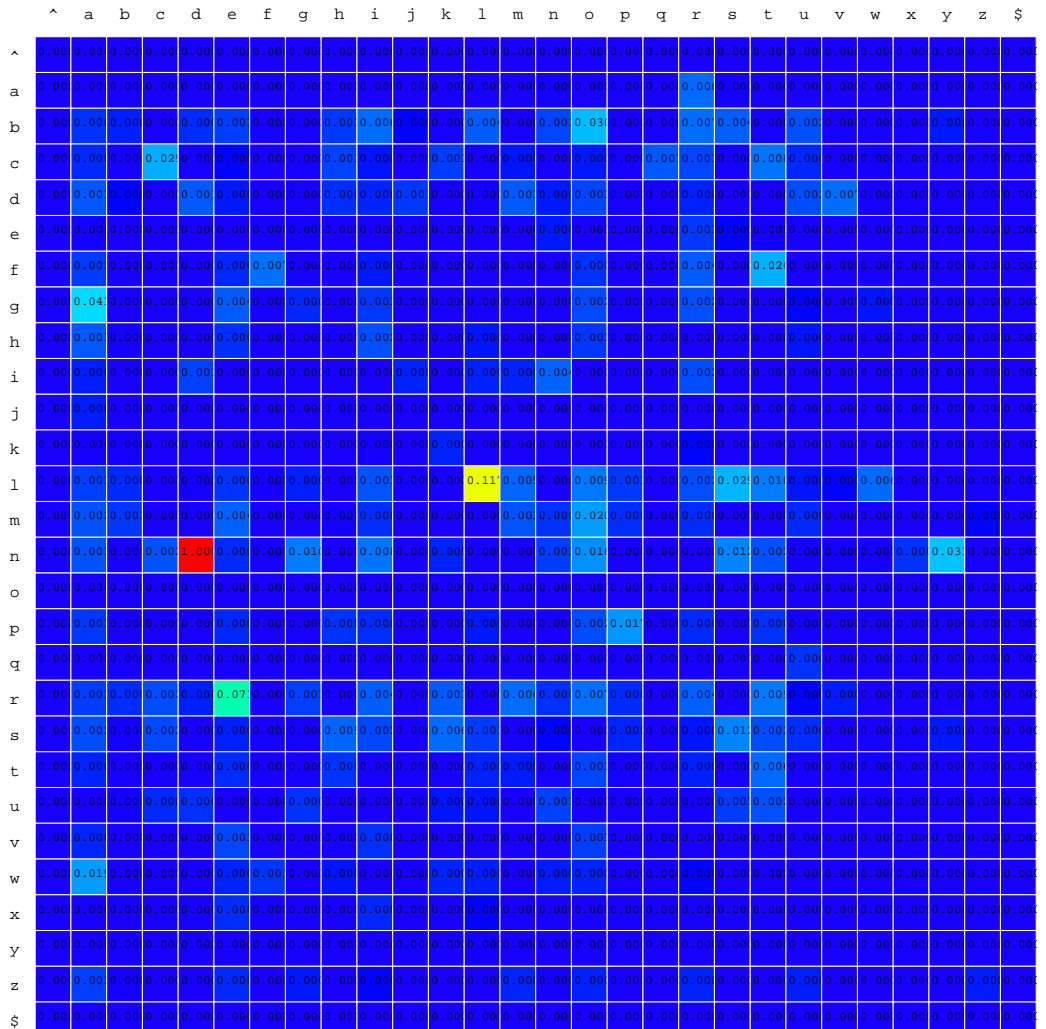


English

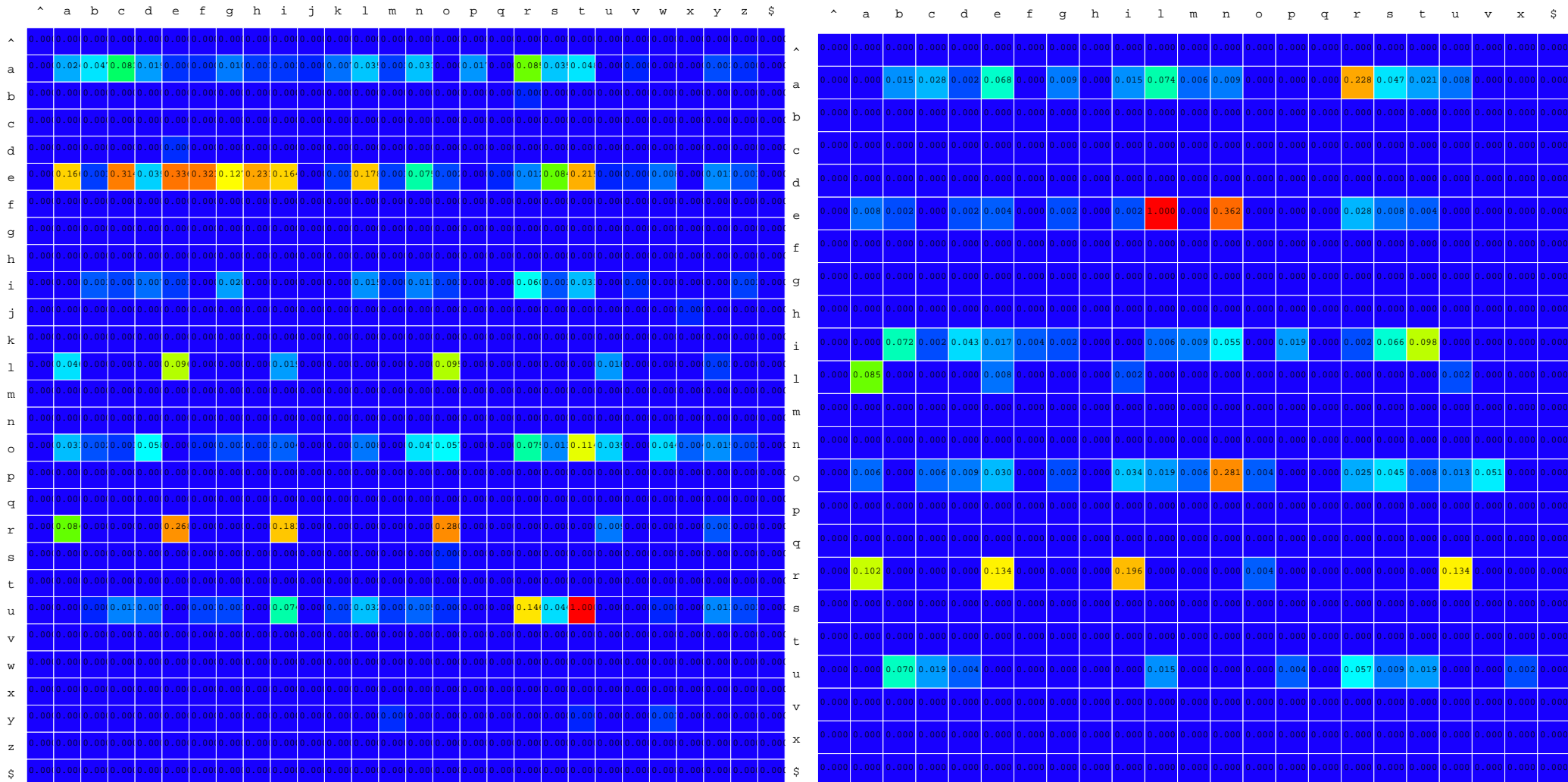


Latin

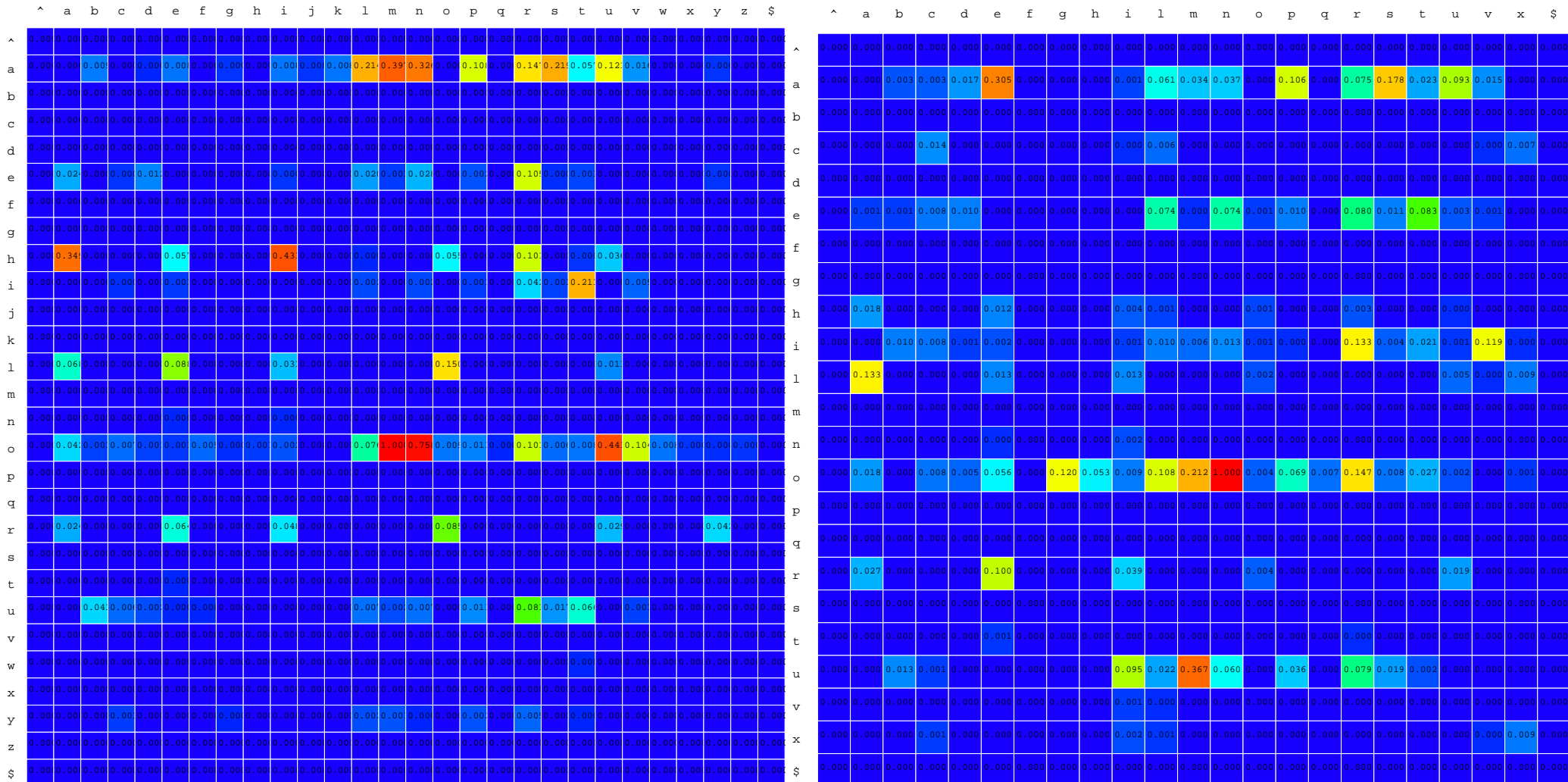
Trigrams - a..



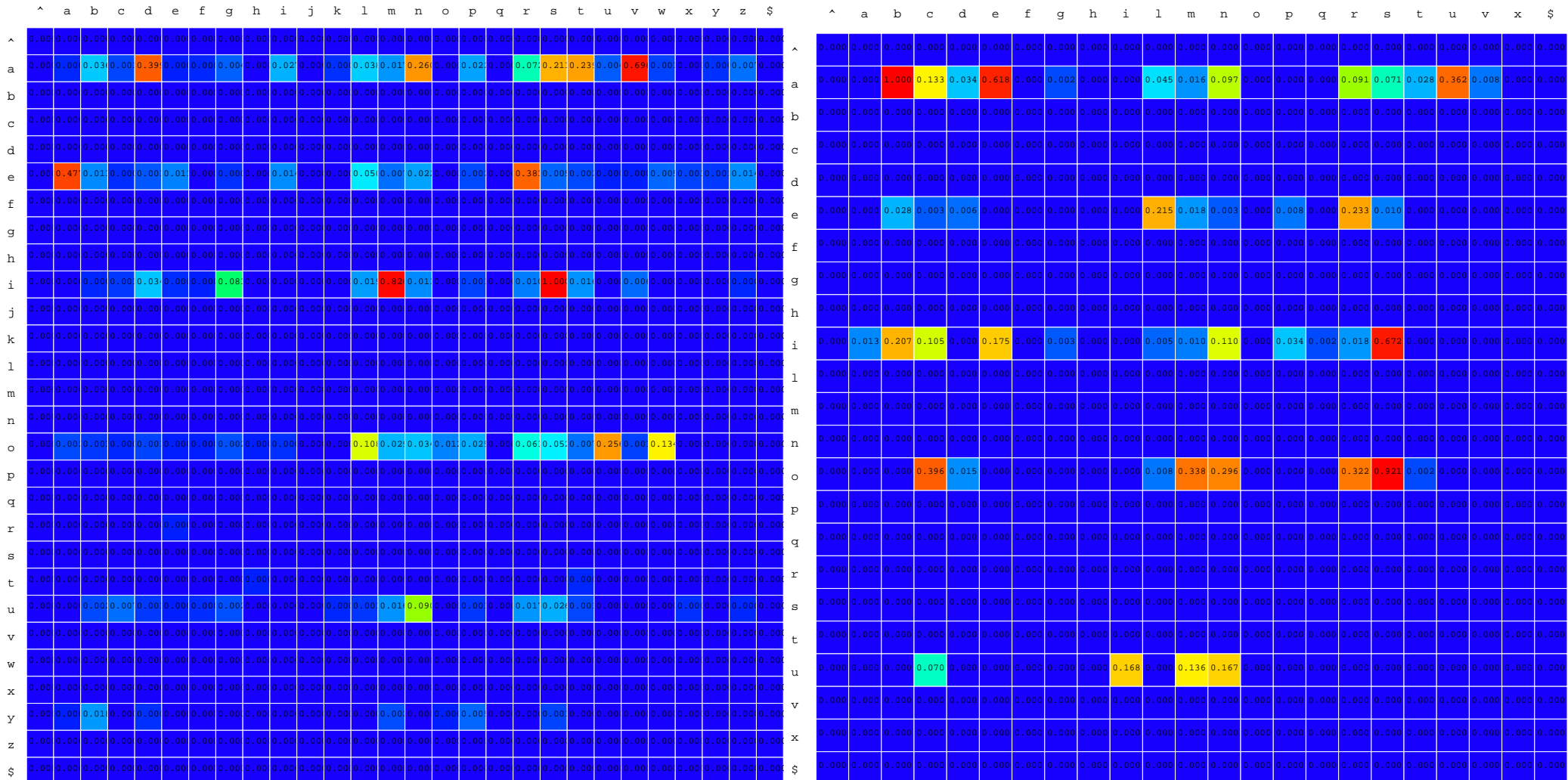
Trigrams - b..



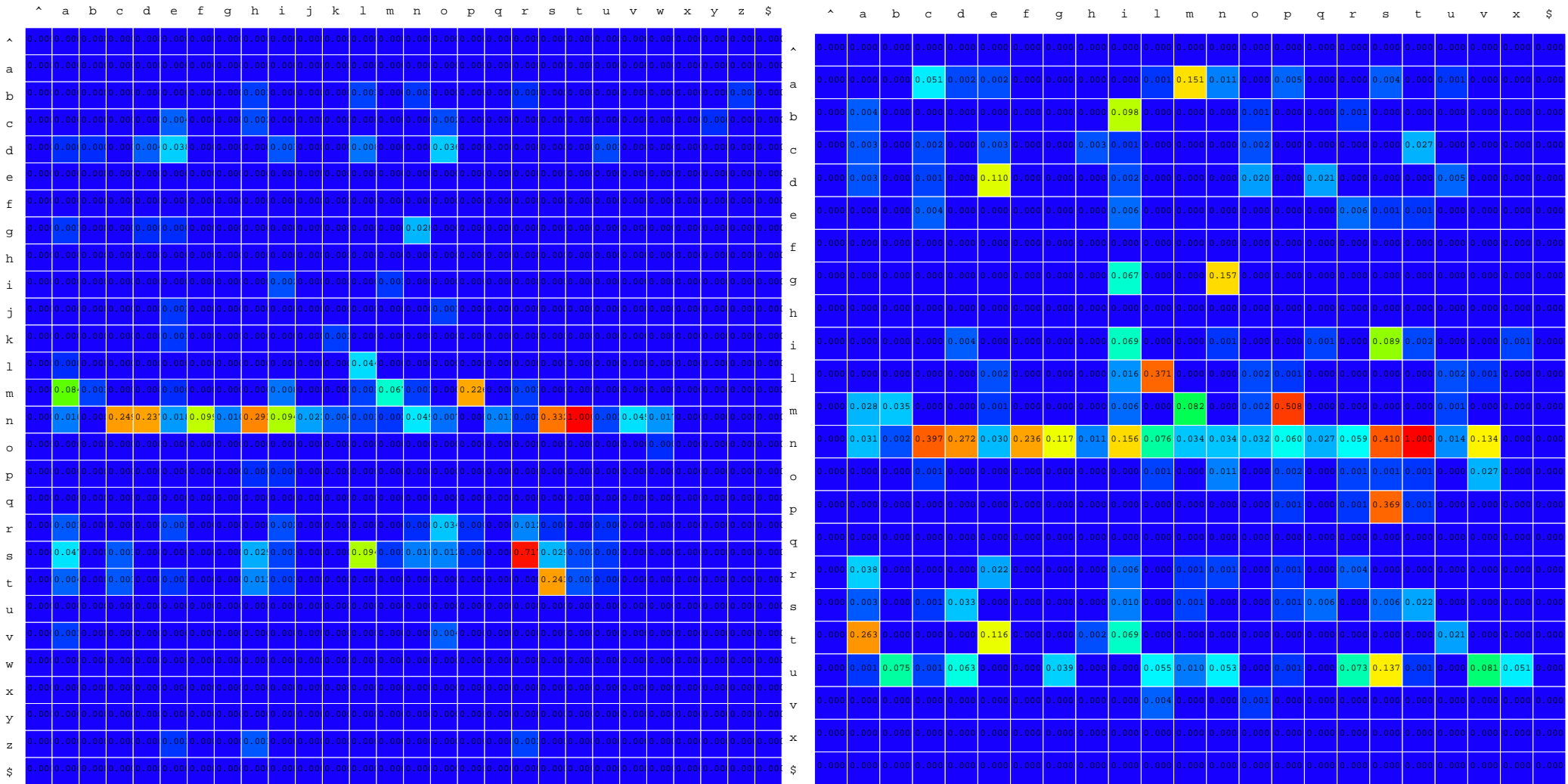
Trigrams - C..



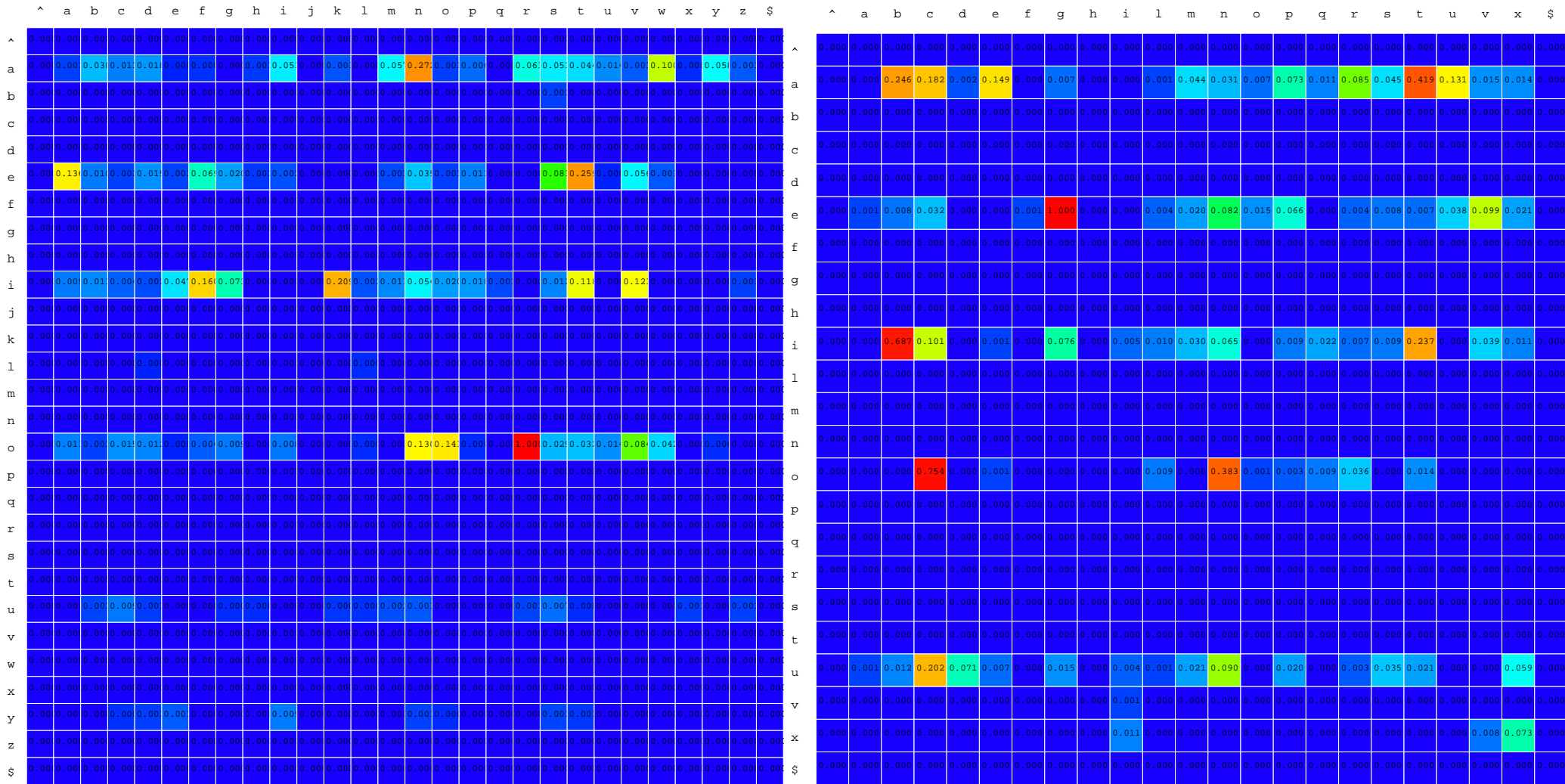
Trigrams - h..



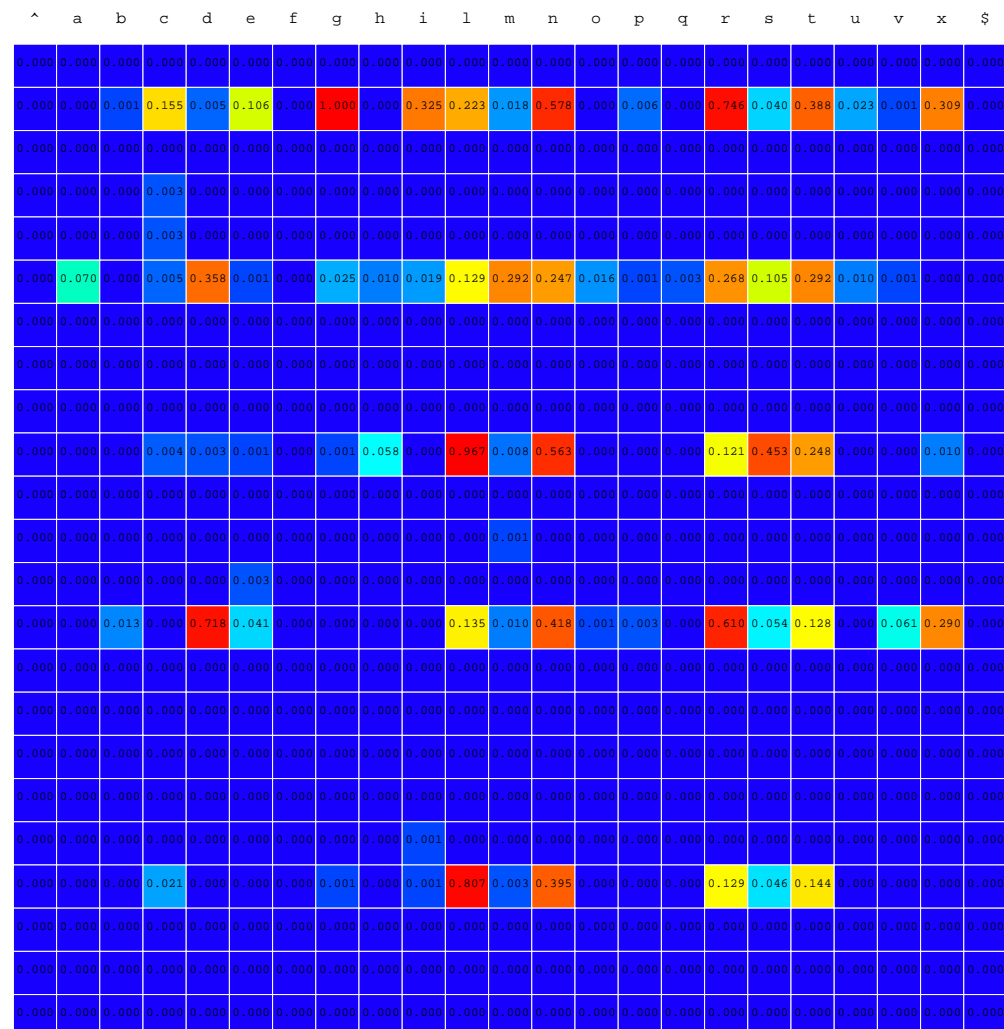
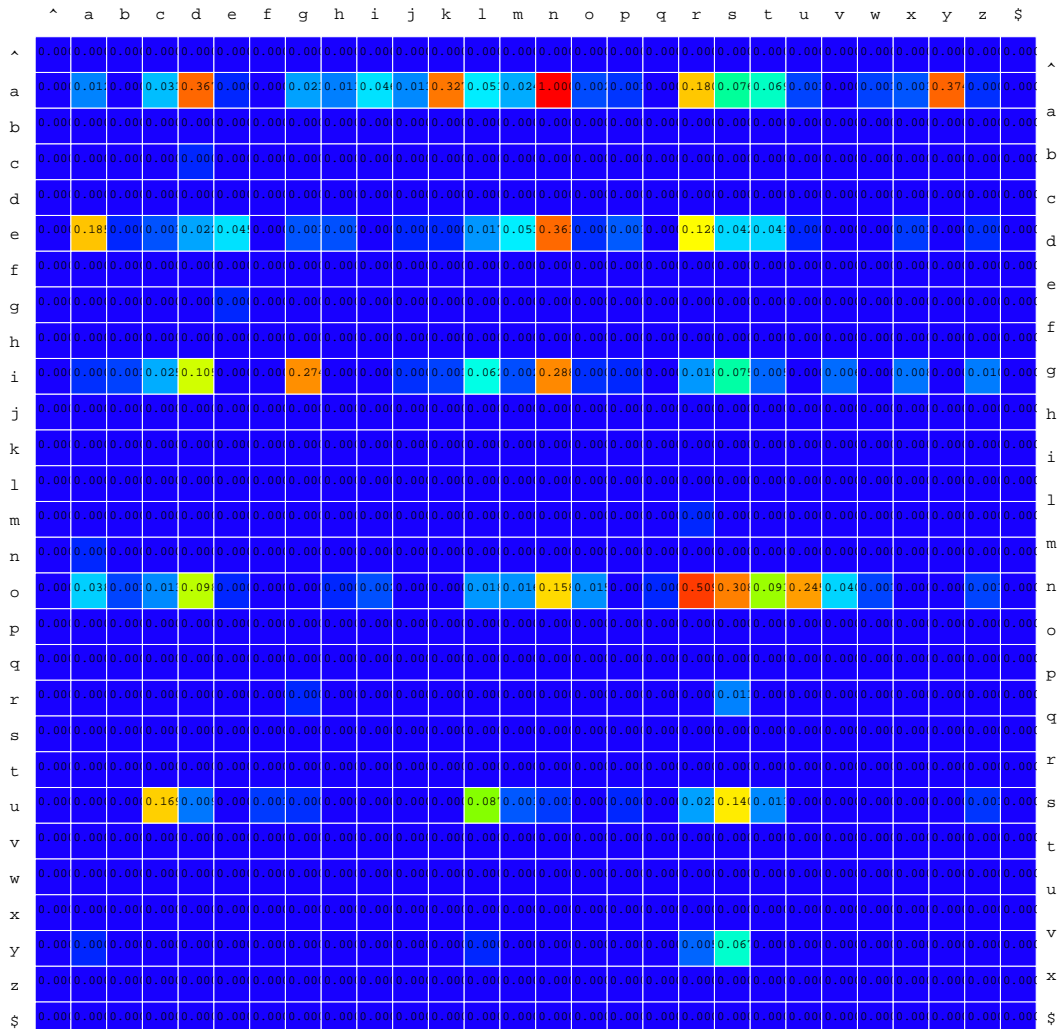
Trigrams - i..



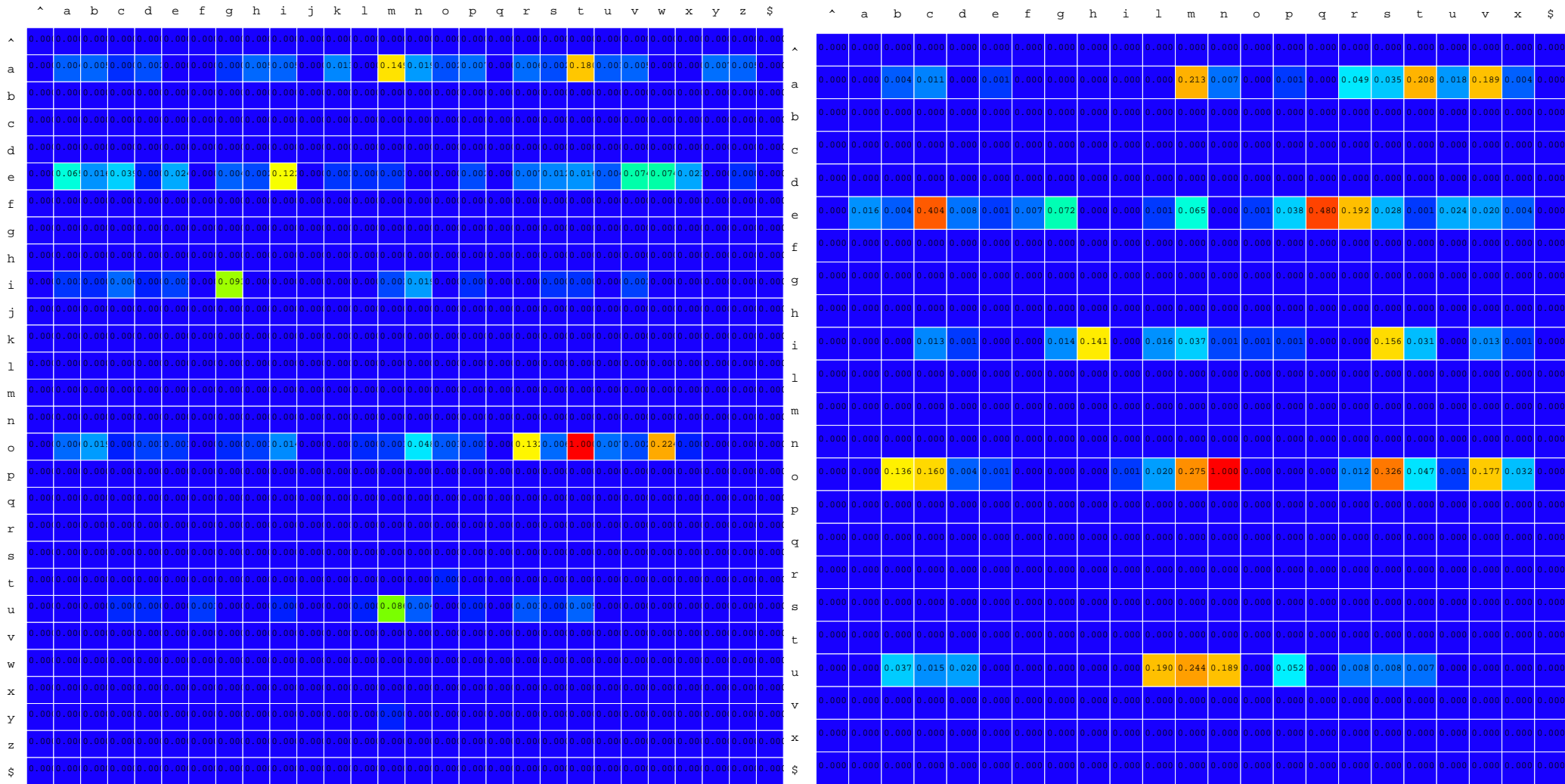
Trigrams - 1..



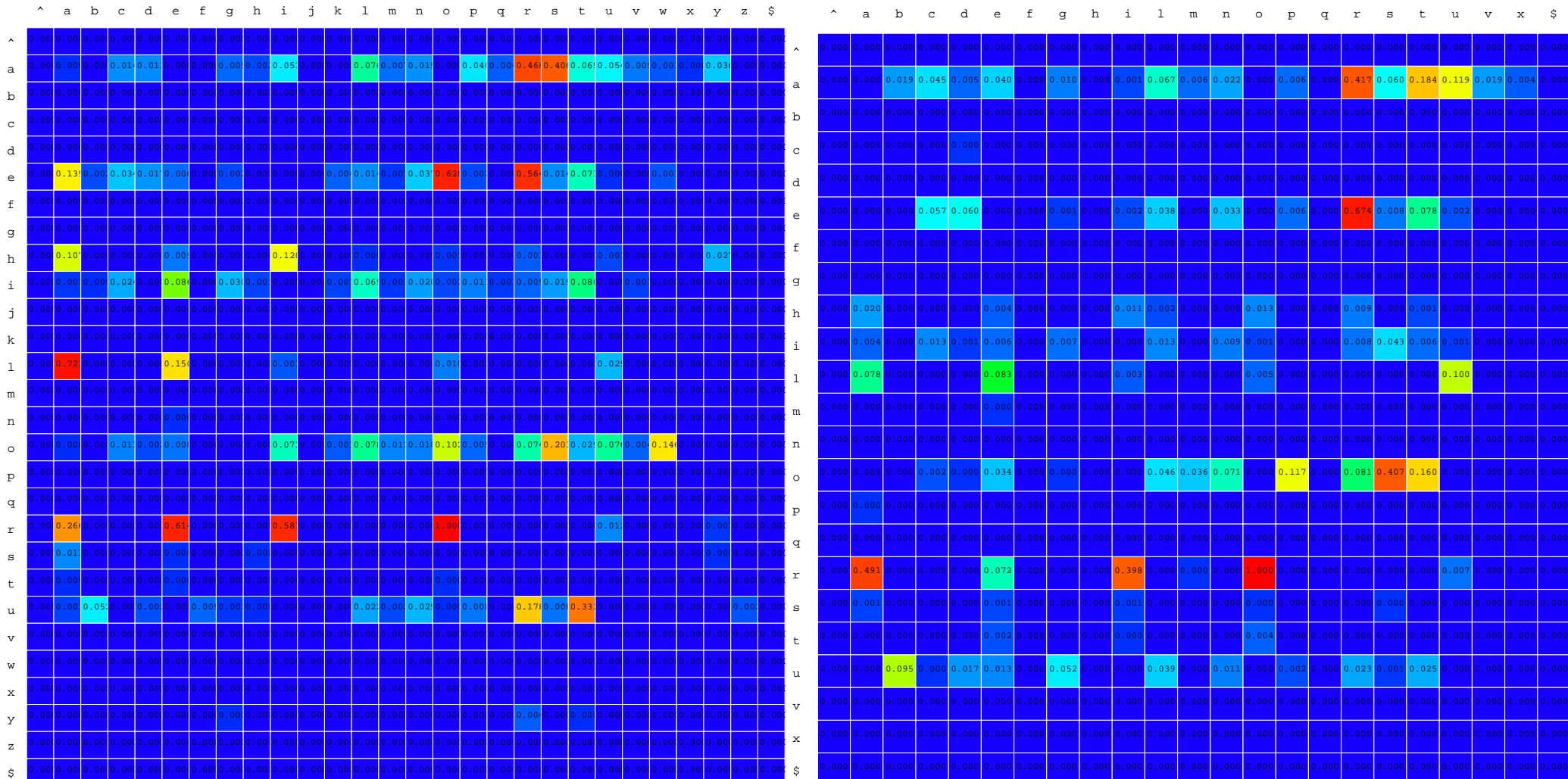
Trigrams - m..



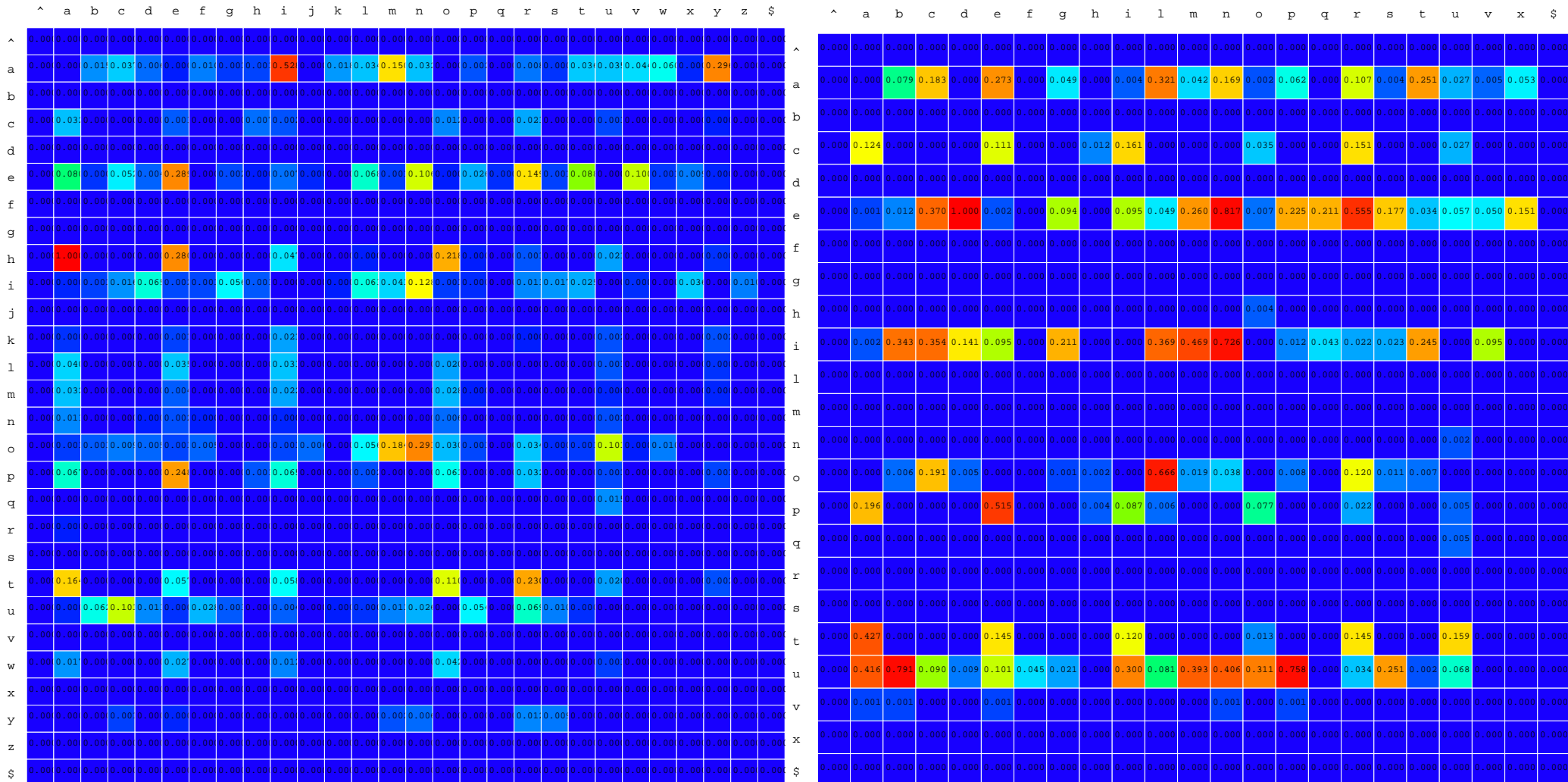
Trigrams - n..



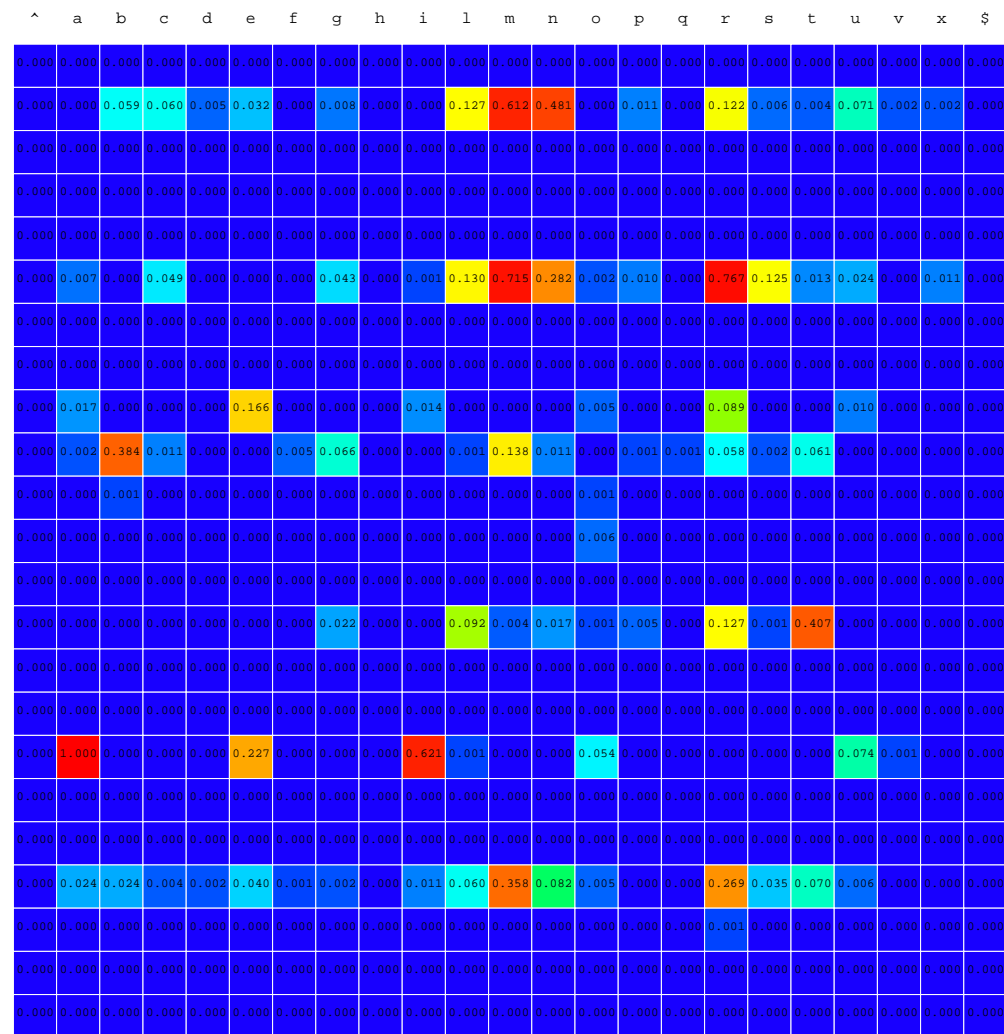
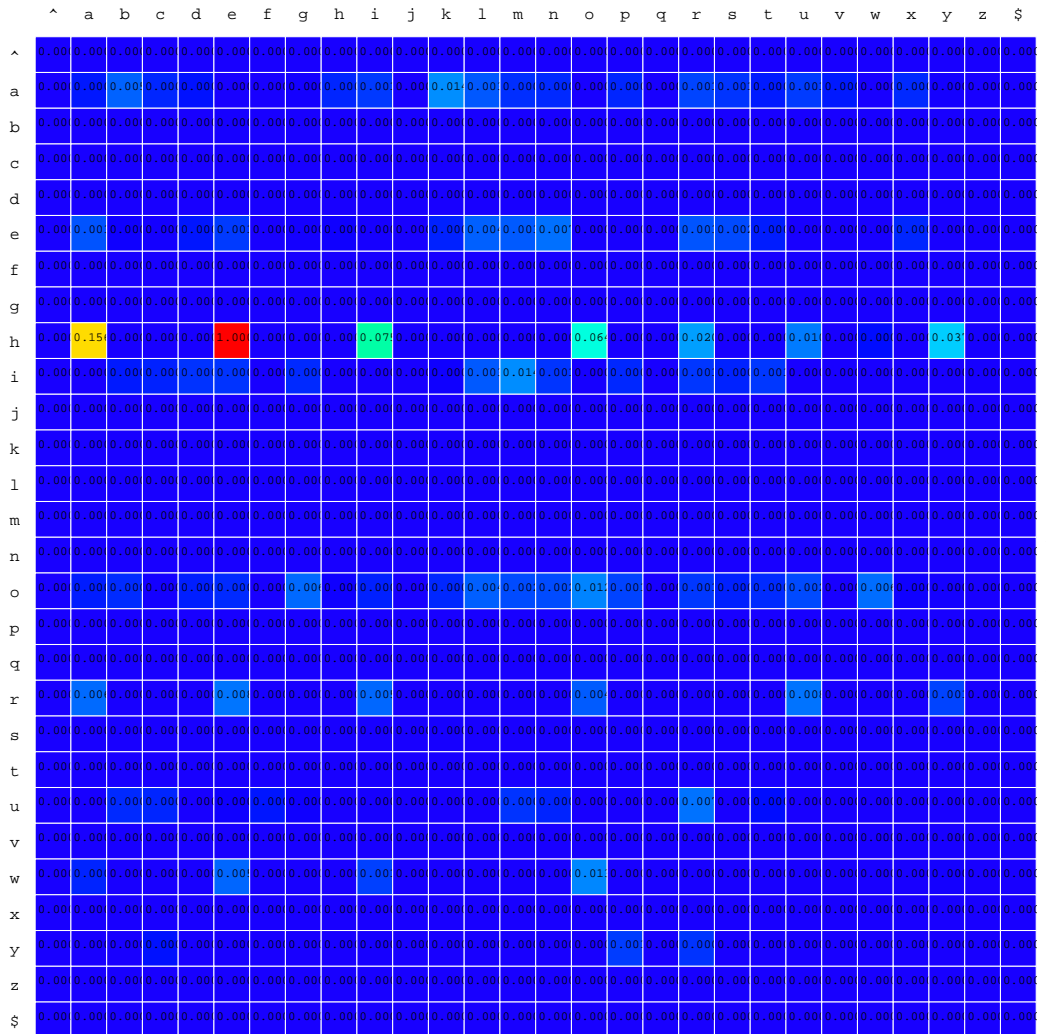
Trigrams - p..



Trigrams - s..



Trigrams - t..



Source files used

- English:

- ▷ *From* <http://www.ota.ahds.ac.uk/> *cayley.txt franklin.txt huck.txt origin-of-species.txt silas.txt*
- ▷ *From* <http://promo.net/pg/> *kjv10.txt (King James Bible)*

- Latin:

- ▷ *From* <http://www.thelatinlibrary.com> *ammianus14.txt gall2.txt gall5.txt gall8.txt pliny.nh3.txt tac.ann11.txt tac.ann14.txt tac.ann1.txt tac.ann4.txt cato.agri.txt gall3.txt gall6.txt l.txt pliny.nh4.txt tac.ann12.txt tac.ann15.txt tac.ann2.txt tac.ann5.txt gall1.txt gall4.txt gall7.txt pliny.nh5.txt tac.ann13.txt tac.ann16.txt tac.ann3.txt tac.ann6.txt*
- ▷ *From* http://penelope.uchicago.edu/Thayer/E/Roman/Texts/Pliny_the_Elder/home.html *pliny2.txt*