CSCI 3412 – Algorithms
Spring 2013
Problem Set 3 – Programming

A character n-gram is a contiguous sequence of *n* characters from a given sequence of text. In this programming assignment you will extract character n-grams from text files.

All programming assignments require that a written report as well as your source code file(s) be turned in. For multi-part programming assignment, such as this one, you may have one written report covering all of the parts and may even have combined source code file(s) as long as I can easily tell what code goes with each part of the problem – some code may go with multiple parts.

The written report should include, for each part of the programming assignment covered, a description of the problem (basically the information from the assignment itself), an overview of your solution to the problem, a description of the relevant algorithm(s) – including an analysis of the algorithm(s), the results, and an analysis of the results.

The programs may be written in any relevant language. The languages C, C++, Java, Ada, Python, and Racket are acceptable. For other languages, please discuss your language choice with me ahead of time. Generally, you can – are encouraged to – use appropriate libraries for regular expressions, standard data structures, etc.

Please review the Collaboration Policy in the course syllabus. Remember that collaboration is encouraged, but that the work turned in must be your own. Also, please remember to include the names of others you collaborated with in your written report and be sure to cite any sources used.

This programming assignment has multiple parts.

Part 1

The first part of this assignment is to extract sample paragraphs to use to train your language recognizer. Wikipedia has articles in various languages and we will extract sample paragraphs from random articles.

From the main page of Wikipedia (http://www.wikipedia.org), you can select the language you are interested in. For example, selecting Español will take you to http://es.wikipedia.org/wiki/Wikipedia:Portada, which is the main page for Spanish.

From any page, you can select 'Random article', which will be in the selected language – for example, 'Página aleatoria' for Spanish – to get a random article. Select a paragraph from the random article that has about 100 (or more) words. Selected paragraphs should be representative of the language – for example, it should have a lot of foreign words, names, etc. Here is an example from a random Spanish article.

> Victoria era hija del príncipe Eduardo, duque de Kent y Strathearn, cuarto hijo del rey Jorge III. Tanto el duque como el rey murieron en 1820, lo que provocó que Victoria fuera criada bajo la supervisión de su madre, la princesa alemana Victoria de Sajonia-

> Coburgo-Saalfeld. Heredó el trono a los dieciocho años, tras la muerte sin descendencia legítima de tres tíos paternos. El Reino Unido era ya en aquella época una monarquía constitucional establecida, en la que el soberano tenía relativamente pocos poderes políticos directos. En privado, Victoria intentó influir en el gobierno y en el nombramiento de ministros. En público, se convirtió en un icono nacional y en la figura que encarnaba el modelo de valores férreos y de moral personal típico de la época.

Select the paragraph and paste it into a text document. Note that the text editor used must support Unicode (e.g., UTF-8). Add a line feed (i.e., hit the enter key) to act as a separator.

Continue this until you have 100 sample paragraphs. [Note that not all articles will have suitable paragraphs.]

I will provide an example text file for English with 100 sample paragraphs and smaller examples for Spanish and Japanese.

Each student will select (or be assigned) a specific language and all of the text files will be available for all of the students.

Part 2

The second part of this programming assignment is to write a routine to extract the character n-grams. This is similar to the previous programming assignments, but extracts character n-grams instead of word n-grams.

The text files, from Part 1, are formatted as sample paragraphs (with no internal line feeds) separated by line feeds. So, a text file with 100 sample paragraphs will be 100 lines long.

For this programming assignment, separate the text into tokens separated by one or more spaces. Then convert to lowercase – from a Unicode perspective – and remove all of the characters that aren't letters – again, from a Unicode perspective. This will give us a sequence of 'words' from the text file. [Note that 'word' may not be semantically appropriate for some specific language, but works from an abstract perspective.]

Part 3

We want to treat each of the sample paragraphs as a training sample for our language recognizer. We want to determine the mean and standard deviations for each of the character n-grams extracted from the sample paragraphs.

Here are my results from the English, Spanish, and Japanese sample paragraphs.

--- English ---
"a": count = 100, mean = 0.085062, stddev = 0.013927
"b": count = 100, mean = 0.016489, stddev = 0.007065
"c": count = 100, mean = 0.036042, stddev = 0.010339
"d": count = 100, mean = 0.040541, stddev = 0.010548

"e": count = 100, mean = 0.118994, stddev = 0.013137
"f": count = 100, mean = 0.023558, stddev = 0.007731
"g": count = 100, mean = 0.018307, stddev = 0.005227
"h": count = 100, mean = 0.048435, stddev = 0.010061
"i": count = 100, mean = 0.073663, stddev = 0.011792
"j": count = 100, mean = 0.00136, stddev = 0.002351
"k": count = 100, mean = 0.006291, stddev = 0.004895
"l": count = 100, mean = 0.043581, stddev = 0.011868
"m": count = 100, mean = 0.025405, stddev = 0.00828
"n": count = 100, mean = 0.073228, stddev = 0.012592
"o": count = 100, mean = 0.076391, stddev = 0.012863
"p": count = 100, mean = 0.02119, stddev = 0.007062
"q": count = 100, mean = 0.000943, stddev = 0.001301
"r": count = 100, mean = 0.064811, stddev = 0.011682
"s": count = 100, mean = 0.06572, stddev = 0.012162
"t": count = 100, mean = 0.089809, stddev = 0.01291
"u": count = 100, mean = 0.025492, stddev = 0.007282
"v": count = 100, mean = 0.009654, stddev = 0.004668
"w": count = 100, mean = 0.015937, stddev = 0.006302
"x": count = 100, mean = 0.001958, stddev = 0.002094
"y": count = 100, mean = 0.016246, stddev = 0.006474
"z": count = 100, mean = 0.000875, stddev = 0.001419
"é": count = 100, mean = 0.000019, stddev = 0.000192

--- Spanish ---
"a": count = 50, mean = 0.121909, stddev = 0.017371
"b": count = 50, mean = 0.011632, stddev = 0.004822
"c": count = 50, mean = 0.048877, stddev = 0.010645
"d": count = 50, mean = 0.052358, stddev = 0.011805
"e": count = 50, mean = 0.128017, stddev = 0.013998
"f": count = 50, mean = 0.008169, stddev = 0.004373
"g": count = 50, mean = 0.013369, stddev = 0.006488
"h": count = 50, mean = 0.006525, stddev = 0.003667
"i": count = 50, mean = 0.062588, stddev = 0.014958
"j": count = 50, mean = 0.003848, stddev = 0.002824
"k": count = 50, mean = 0.000404, stddev = 0.001008
"l": count = 50, mean = 0.060093, stddev = 0.009529
"m": count = 50, mean = 0.027214, stddev = 0.007634
"n": count = 50, mean = 0.069529, stddev = 0.012624
"o": count = 50, mean = 0.083413, stddev = 0.013624
"p": count = 50, mean = 0.026008, stddev = 0.006751
"q": count = 50, mean = 0.007111, stddev = 0.003399
"r": count = 50, mean = 0.067223, stddev = 0.013873
"s": count = 50, mean = 0.071083, stddev = 0.018094
"t": count = 50, mean = 0.045317, stddev = 0.008214
"u": count = 50, mean = 0.040069, stddev = 0.007926
"v": count = 50, mean = 0.009481, stddev = 0.005593

"w": count = 50, mean = 0.000127, stddev = 0.0004
"x": count = 50, mean = 0.001827, stddev = 0.002368
"y": count = 50, mean = 0.00791, stddev = 0.003292
"z": count = 50, mean = 0.004749, stddev = 0.002898
"á": count = 50, mean = 0.003284, stddev = 0.003349
"é": count = 50, mean = 0.002655, stddev = 0.001994
"í": count = 50, mean = 0.004798, stddev = 0.003261
"ñ": count = 50, mean = 0.001691, stddev = 0.001851
"ó": count = 50, mean = 0.007746, stddev = 0.004218
"ú": count = 50, mean = 0.000941, stddev = 0.001099
"ü": count = 50, mean = 0.000035, stddev = 0.000244

--- Japanese ---
"b": count = 6, mean = 0.000995, stddev = 0.002225
"c": count = 6, mean = 0.005776, stddev = 0.008384
"l": count = 6, mean = 0.000697, stddev = 0.001559
"m": count = 6, mean = 0.00994, stddev = 0.014367
"v": count = 6, mean = 0.000697, stddev = 0.001559
"x": count = 6, mean = 0.000995, stddev = 0.002225
"あ": count = 6, mean = 0.00418, stddev = 0.00433
"い": count = 6, mean = 0.010377, stddev = 0.002761
"う": count = 6, mean = 0.002861, stddev = 0.004318
"お": count = 6, mean = 0.002148, stddev = 0.002466
"か": count = 6, mean = 0.012912, stddev = 0.010863
"が": count = 6, mean = 0.017296, stddev = 0.004819
"き": count = 6, mean = 0.0037, stddev = 0.004491
"く": count = 6, mean = 0.003916, stddev = 0.006779
"け": count = 6, mean = 0.003202, stddev = 0.004376
"げ": count = 6, mean = 0.000697, stddev = 0.001559
"こ": count = 6, mean = 0.009041, stddev = 0.007296
"さ": count = 6, mean = 0.013678, stddev = 0.01049
"ざ": count = 6, mean = 0.000995, stddev = 0.002225
"し": count = 6, mean = 0.011829, stddev = 0.004799
"じ": count = 6, mean = 0.000498, stddev = 0.001112
"す": count = 6, mean = 0.001982, stddev = 0.003633
"ず": count = 6, mean = 0.003293, stddev = 0.003508
"せ": count = 6, mean = 0.000324, stddev = 0.000724
"そ": count = 6, mean = 0.005193, stddev = 0.004203
"た": count = 6, mean = 0.022506, stddev = 0.013758
"ち": count = 6, mean = 0.00205, stddev = 0.002386
"っ": count = 6, mean = 0.004643, stddev = 0.006957
"つ": count = 6, mean = 0.005733, stddev = 0.00919
"て": count = 6, mean = 0.011122, stddev = 0.007086
"で": count = 6, mean = 0.017058, stddev = 0.007266
"と": count = 6, mean = 0.014814, stddev = 0.006504
"ど": count = 6, mean = 0.003916, stddev = 0.004581

"な": count = 6, mean = 0.017077, stddev = 0.015445
"に": count = 6, mean = 0.033978, stddev = 0.00615
"の": count = 6, mean = 0.04156, stddev = 0.012572
"は": count = 6, mean = 0.02192, stddev = 0.016076
"び": count = 6, mean = 0.001029, stddev = 0.0023
"ま": count = 6, mean = 0.004346, stddev = 0.005359
"み": count = 6, mean = 0.001676, stddev = 0.002461
"む": count = 6, mean = 0.001658, stddev = 0.003708
"め": count = 6, mean = 0.004106, stddev = 0.005998
"も": count = 6, mean = 0.010933, stddev = 0.007413
"や": count = 6, mean = 0.002863, stddev = 0.003258
"よ": count = 6, mean = 0.005615, stddev = 0.00406
"ら": count = 6, mean = 0.00851, stddev = 0.008223
"り": count = 6, mean = 0.00823, stddev = 0.005388
"る": count = 6, mean = 0.015932, stddev = 0.014842
"れ": count = 6, mean = 0.016481, stddev = 0.01419
"ろ": count = 6, mean = 0.000324, stddev = 0.000724
"わ": count = 6, mean = 0.002887, stddev = 0.004087
"を": count = 6, mean = 0.019255, stddev = 0.007981
"ん": count = 6, mean = 0.001021, stddev = 0.001582
"ア": count = 6, mean = 0.00259, stddev = 0.004587
"ィ": count = 6, mean = 0.000697, stddev = 0.001559
"イ": count = 6, mean = 0.003287, stddev = 0.006113
"ウ": count = 6, mean = 0.000697, stddev = 0.001559
"ェ": count = 6, mean = 0.001395, stddev = 0.003119
"エ": count = 6, mean = 0.001692, stddev = 0.002448
"ォ": count = 6, mean = 0.001395, stddev = 0.003119
"カ": count = 6, mean = 0.004281, stddev = 0.006575
"ガ": count = 6, mean = 0.000697, stddev = 0.001559
"ギ": count = 6, mean = 0.001395, stddev = 0.003119
"ク": count = 6, mean = 0.00371, stddev = 0.005311
"グ": count = 6, mean = 0.000697, stddev = 0.001559
"ゲ": count = 6, mean = 0.000324, stddev = 0.000724
"コ": count = 6, mean = 0.000821, stddev = 0.0012
"サ": count = 6, mean = 0.000697, stddev = 0.001559
"シ": count = 6, mean = 0.003361, stddev = 0.00346
"ジ": count = 6, mean = 0.001692, stddev = 0.002448
"ス": count = 6, mean = 0.004382, stddev = 0.004802
"ソ": count = 6, mean = 0.000697, stddev = 0.001559
"タ": count = 6, mean = 0.004755, stddev = 0.006567
"ダ": count = 6, mean = 0.000697, stddev = 0.001559
"チ": count = 6, mean = 0.000697, stddev = 0.001559
"ッ": count = 6, mean = 0.004034, stddev = 0.005711
"ツ": count = 6, mean = 0.003784, stddev = 0.006764
"テ": count = 6, mean = 0.001692, stddev = 0.002448
"デ": count = 6, mean = 0.002689, stddev = 0.003807

"ト": count = 6, mean = 0.002987, stddev = 0.003161
"ド": count = 6, mean = 0.001395, stddev = 0.003119
"ハ": count = 6, mean = 0.001021, stddev = 0.001582
"バ": count = 6, mean = 0.001021, stddev = 0.001582
"パ": count = 6, mean = 0.000498, stddev = 0.001112
"ビ": count = 6, mean = 0.002092, stddev = 0.004678
"フ": count = 6, mean = 0.002789, stddev = 0.006237
"プ": count = 6, mean = 0.001021, stddev = 0.001582
"へ": count = 6, mean = 0.000697, stddev = 0.001559
"べ": count = 6, mean = 0.000647, stddev = 0.001447
"ぺ": count = 6, mean = 0.000324, stddev = 0.000724
"ボ": count = 6, mean = 0.000971, stddev = 0.002171
"マ": count = 6, mean = 0.001145, stddev = 0.00164
"ミ": count = 6, mean = 0.000697, stddev = 0.001559
"ム": count = 6, mean = 0.004779, stddev = 0.006571
"メ": count = 6, mean = 0.000995, stddev = 0.002225
"ヤ": count = 6, mean = 0.000697, stddev = 0.001559
"ョ": count = 6, mean = 0.000697, stddev = 0.001559
"ラ": count = 6, mean = 0.002092, stddev = 0.004678
"リ": count = 6, mean = 0.006, stddev = 0.007038
"ル": count = 6, mean = 0.002116, stddev = 0.003566
"レ": count = 6, mean = 0.001021, stddev = 0.001582
"ン": count = 6, mean = 0.009858, stddev = 0.012372
"ー": count = 6, mean = 0.01016, stddev = 0.018124
"一": count = 6, mean = 0.002182, stddev = 0.002519
"万": count = 6, mean = 0.001244, stddev = 0.002781
"三": count = 6, mean = 0.001244, stddev = 0.002781
"上": count = 6, mean = 0.002821, stddev = 0.003023
"下": count = 6, mean = 0.000324, stddev = 0.000724
"不": count = 6, mean = 0.000829, stddev = 0.001854
"世": count = 6, mean = 0.000697, stddev = 0.001559
"両": count = 6, mean = 0.000324, stddev = 0.000724
"並": count = 6, mean = 0.000324, stddev = 0.000724
"中": count = 6, mean = 0.002795, stddev = 0.002203
"丸": count = 6, mean = 0.001029, stddev = 0.0023
"主": count = 6, mean = 0.001842, stddev = 0.001877
"乗": count = 6, mean = 0.000498, stddev = 0.001112
"乾": count = 6, mean = 0.001244, stddev = 0.002781
"事": count = 6, mean = 0.002306, stddev = 0.003701
"交": count = 6, mean = 0.000697, stddev = 0.001559
"京": count = 6, mean = 0.000622, stddev = 0.001391
"仁": count = 6, mean = 0.001244, stddev = 0.002781
"他": count = 6, mean = 0.001195, stddev = 0.001725
"付": count = 6, mean = 0.000697, stddev = 0.001559
"代": count = 6, mean = 0.001941, stddev = 0.002904
"以": count = 6, mean = 0.000498, stddev = 0.001112

"伏": count = 6, mean = 0.001244, stddev = 0.002781
"休": count = 6, mean = 0.000498, stddev = 0.001112
"似": count = 6, mean = 0.000324, stddev = 0.000724
"佑": count = 6, mean = 0.001866, stddev = 0.004172
"体": count = 6, mean = 0.003666, stddev = 0.00438
"作": count = 6, mean = 0.000946, stddev = 0.001433
"使": count = 6, mean = 0.004448, stddev = 0.006588
"供": count = 6, mean = 0.000829, stddev = 0.001854
"便": count = 6, mean = 0.000697, stddev = 0.001559
"係": count = 6, mean = 0.000829, stddev = 0.001854
"修": count = 6, mean = 0.001195, stddev = 0.001725
"備": count = 6, mean = 0.001493, stddev = 0.003337
"先": count = 6, mean = 0.001029, stddev = 0.0023
"光": count = 6, mean = 0.001244, stddev = 0.002781
"八": count = 6, mean = 0.000324, stddev = 0.000724
"公": count = 6, mean = 0.000622, stddev = 0.001391
"内": count = 6, mean = 0.000821, stddev = 0.0012
"円": count = 6, mean = 0.002381, stddev = 0.004512
"出": count = 6, mean = 0.001029, stddev = 0.0023
"分": count = 6, mean = 0.000995, stddev = 0.002225
"列": count = 6, mean = 0.000324, stddev = 0.000724
"初": count = 6, mean = 0.001692, stddev = 0.002448
"制": count = 6, mean = 0.000498, stddev = 0.001112
"前": count = 6, mean = 0.003352, stddev = 0.004921
"力": count = 6, mean = 0.000995, stddev = 0.002225
"加": count = 6, mean = 0.000324, stddev = 0.000724
"助": count = 6, mean = 0.000498, stddev = 0.001112
"勇": count = 6, mean = 0.001244, stddev = 0.002781
"動": count = 6, mean = 0.003411, stddev = 0.004604
"化": count = 6, mean = 0.000995, stddev = 0.002225
"北": count = 6, mean = 0.001866, stddev = 0.004172
"単": count = 6, mean = 0.001294, stddev = 0.002895
"厚": count = 6, mean = 0.001468, stddev = 0.002233
"参": count = 6, mean = 0.000829, stddev = 0.001854
"反": count = 6, mean = 0.001029, stddev = 0.0023
"取": count = 6, mean = 0.001195, stddev = 0.001725
"口": count = 6, mean = 0.001676, stddev = 0.002461
"可": count = 6, mean = 0.001327, stddev = 0.001962
"台": count = 6, mean = 0.00165, stddev = 0.001875
"右": count = 6, mean = 0.000971, stddev = 0.002171
"号": count = 6, mean = 0.002488, stddev = 0.005562
"合": count = 6, mean = 0.001658, stddev = 0.003708
"名": count = 6, mean = 0.000622, stddev = 0.001391
"向": count = 6, mean = 0.001294, stddev = 0.002895
"君": count = 6, mean = 0.001244, stddev = 0.002781
"否": count = 6, mean = 0.002488, stddev = 0.005562

"含": count = 6, mean = 0.001658, stddev = 0.003708
"命": count = 6, mean = 0.000622, stddev = 0.001391
"員": count = 6, mean = 0.000498, stddev = 0.001112
"哲": count = 6, mean = 0.000622, stddev = 0.001391
"唐": count = 6, mean = 0.000622, stddev = 0.001391
"商": count = 6, mean = 0.000697, stddev = 0.001559
"問": count = 6, mean = 0.000498, stddev = 0.001112
"啓": count = 6, mean = 0.001244, stddev = 0.002781
"嘉": count = 6, mean = 0.001244, stddev = 0.002781
"回": count = 6, mean = 0.001395, stddev = 0.003119
"国": count = 6, mean = 0.00228, stddev = 0.003691
"地": count = 6, mean = 0.000622, stddev = 0.001391
"型": count = 6, mean = 0.001021, stddev = 0.001582
"基": count = 6, mean = 0.003618, stddev = 0.005786
"場": count = 6, mean = 0.001658, stddev = 0.003708
"塔": count = 6, mean = 0.000647, stddev = 0.001447
"填": count = 6, mean = 0.000324, stddev = 0.000724
"売": count = 6, mean = 0.000697, stddev = 0.001559
"変": count = 6, mean = 0.000697, stddev = 0.001559
"外": count = 6, mean = 0.001651, stddev = 0.002439
"多": count = 6, mean = 0.001029, stddev = 0.0023
"大": count = 6, mean = 0.007755, stddev = 0.007772
"天": count = 6, mean = 0.002488, stddev = 0.005562
"太": count = 6, mean = 0.000622, stddev = 0.001391
"央": count = 6, mean = 0.000647, stddev = 0.001447
"威": count = 6, mean = 0.001866, stddev = 0.004172
"子": count = 6, mean = 0.000622, stddev = 0.001391
"孔": count = 6, mean = 0.000622, stddev = 0.001391
"守": count = 6, mean = 0.000324, stddev = 0.000724
"安": count = 6, mean = 0.000622, stddev = 0.001391
"宋": count = 6, mean = 0.000622, stddev = 0.001391
"宗": count = 6, mean = 0.000622, stddev = 0.001391
"定": count = 6, mean = 0.000829, stddev = 0.001854
"実": count = 6, mean = 0.005307, stddev = 0.007849
"寝": count = 6, mean = 0.000498, stddev = 0.001112
"察": count = 6, mean = 0.000697, stddev = 0.001559
"対": count = 6, mean = 0.004406, stddev = 0.006556
"封": count = 6, mean = 0.000622, stddev = 0.001391
"尊": count = 6, mean = 0.001244, stddev = 0.002781
"小": count = 6, mean = 0.001021, stddev = 0.001582
"尖": count = 6, mean = 0.001029, stddev = 0.0023
"尾": count = 6, mean = 0.000324, stddev = 0.000724
"局": count = 6, mean = 0.000498, stddev = 0.001112
"属": count = 6, mean = 0.000697, stddev = 0.001559
"州": count = 6, mean = 0.000622, stddev = 0.001391
"工": count = 6, mean = 0.000324, stddev = 0.000724

"左": count = 6, mean = 0.000971, stddev = 0.002171
"市": count = 6, mean = 0.001658, stddev = 0.003708
"帝": count = 6, mean = 0.008706, stddev = 0.019468
"帯": count = 6, mean = 0.001029, stddev = 0.0023
"帳": count = 6, mean = 0.000829, stddev = 0.001854
"年": count = 6, mean = 0.006916, stddev = 0.01368
"床": count = 6, mean = 0.000324, stddev = 0.000724
"店": count = 6, mean = 0.000697, stddev = 0.001559
"府": count = 6, mean = 0.001658, stddev = 0.003708
"度": count = 6, mean = 0.000498, stddev = 0.001112
"庫": count = 6, mean = 0.000324, stddev = 0.000724
"廃": count = 6, mean = 0.001658, stddev = 0.003708
"廟": count = 6, mean = 0.003731, stddev = 0.008344
"建": count = 6, mean = 0.000324, stddev = 0.000724
"式": count = 6, mean = 0.001468, stddev = 0.002233
"弾": count = 6, mean = 0.000324, stddev = 0.000724
"当": count = 6, mean = 0.001319, stddev = 0.002198
"形": count = 6, mean = 0.003734, stddev = 0.006762
"影": count = 6, mean = 0.000697, stddev = 0.001559
"径": count = 6, mean = 0.001676, stddev = 0.002461
"後": count = 6, mean = 0.006471, stddev = 0.007062
"得": count = 6, mean = 0.000829, stddev = 0.001854
"御": count = 6, mean = 0.000324, stddev = 0.000724
"忠": count = 6, mean = 0.002488, stddev = 0.005562
"恐": count = 6, mean = 0.000697, stddev = 0.001559
"態": count = 6, mean = 0.003317, stddev = 0.007416
"慌": count = 6, mean = 0.000697, stddev = 0.001559
"慶": count = 6, mean = 0.001244, stddev = 0.002781
"憩": count = 6, mean = 0.000498, stddev = 0.001112
"成": count = 6, mean = 0.001119, stddev = 0.001598
"戦": count = 6, mean = 0.000324, stddev = 0.000724
"所": count = 6, mean = 0.000324, stddev = 0.000724
"技": count = 6, mean = 0.000498, stddev = 0.001112
"持": count = 6, mean = 0.003352, stddev = 0.004921
"指": count = 6, mean = 0.000324, stddev = 0.000724
"採": count = 6, mean = 0.000498, stddev = 0.001112
"接": count = 6, mean = 0.000324, stddev = 0.000724
"搭": count = 6, mean = 0.000498, stddev = 0.001112
"撮": count = 6, mean = 0.000697, stddev = 0.001559
"操": count = 6, mean = 0.000498, stddev = 0.001112
"改": count = 6, mean = 0.000498, stddev = 0.001112
"故": count = 6, mean = 0.000697, stddev = 0.001559
"数": count = 6, mean = 0.000498, stddev = 0.001112
"整": count = 6, mean = 0.000498, stddev = 0.001112
"文": count = 6, mean = 0.000622, stddev = 0.001391
"斜": count = 6, mean = 0.000647, stddev = 0.001447

"方": count = 6, mean = 0.002971, stddev = 0.004489
"施": count = 6, mean = 0.000498, stddev = 0.001112
"日": count = 6, mean = 0.000498, stddev = 0.001112
"明": count = 6, mean = 0.000622, stddev = 0.001391
"映": count = 6, mean = 0.000697, stddev = 0.001559
"時": count = 6, mean = 0.001643, stddev = 0.00178
"暦": count = 6, mean = 0.001244, stddev = 0.002781
"更": count = 6, mean = 0.000697, stddev = 0.001559
"最": count = 6, mean = 0.001294, stddev = 0.002895
"望": count = 6, mean = 0.000622, stddev = 0.001391
"朝": count = 6, mean = 0.000622, stddev = 0.001391
"期": count = 6, mean = 0.000697, stddev = 0.001559
"本": count = 6, mean = 0.003592, stddev = 0.004288
"村": count = 6, mean = 0.001658, stddev = 0.003708
"来": count = 6, mean = 0.000622, stddev = 0.001391
"板": count = 6, mean = 0.001294, stddev = 0.002895
"架": count = 6, mean = 0.000324, stddev = 0.000724
"構": count = 6, mean = 0.000324, stddev = 0.000724
"様": count = 6, mean = 0.000324, stddev = 0.000724
"標": count = 6, mean = 0.000995, stddev = 0.002225
"機": count = 6, mean = 0.003309, stddev = 0.006568
"止": count = 6, mean = 0.000829, stddev = 0.001854
"武": count = 6, mean = 0.004055, stddev = 0.008229
"歴": count = 6, mean = 0.000622, stddev = 0.001391
"残": count = 6, mean = 0.000498, stddev = 0.001112
"殻": count = 6, mean = 0.005144, stddev = 0.011502
"水": count = 6, mean = 0.000324, stddev = 0.000724
"氷": count = 6, mean = 0.000498, stddev = 0.001112
"決": count = 6, mean = 0.000829, stddev = 0.001854
"治": count = 6, mean = 0.001244, stddev = 0.002781
"泉": count = 6, mean = 0.000622, stddev = 0.001391
"法": count = 6, mean = 0.000829, stddev = 0.001854
"海": count = 6, mean = 0.000324, stddev = 0.000724
"消": count = 6, mean = 0.000697, stddev = 0.001559
"清": count = 6, mean = 0.000622, stddev = 0.001391
"点": count = 6, mean = 0.000829, stddev = 0.001854
"烈": count = 6, mean = 0.000622, stddev = 0.001391
"煙": count = 6, mean = 0.000324, stddev = 0.000724
"照": count = 6, mean = 0.000829, stddev = 0.001854
"爵": count = 6, mean = 0.000622, stddev = 0.001391
"片": count = 6, mean = 0.000647, stddev = 0.001447
"状": count = 6, mean = 0.00384, stddev = 0.005419
"玉": count = 6, mean = 0.000622, stddev = 0.001391
"王": count = 6, mean = 0.001244, stddev = 0.002781
"現": count = 6, mean = 0.000829, stddev = 0.001854
"理": count = 6, mean = 0.002853, stddev = 0.003573

"生": count = 6, mean = 0.000697, stddev = 0.001559
"産": count = 6, mean = 0.000498, stddev = 0.001112
"用": count = 6, mean = 0.007357, stddev = 0.008642
"甲": count = 6, mean = 0.002589, stddev = 0.005789
"町": count = 6, mean = 0.001658, stddev = 0.003708
"画": count = 6, mean = 0.000697, stddev = 0.001559
"界": count = 6, mean = 0.001941, stddev = 0.002904
"異": count = 6, mean = 0.001858, stddev = 0.00265
"発": count = 6, mean = 0.000498, stddev = 0.001112
"的": count = 6, mean = 0.002024, stddev = 0.002863
"目": count = 6, mean = 0.001824, stddev = 0.002596
"県": count = 6, mean = 0.001658, stddev = 0.003708
"眼": count = 6, mean = 0.002058, stddev = 0.004601
"砲": count = 6, mean = 0.005502, stddev = 0.012302
"示": count = 6, mean = 0.000829, stddev = 0.001854
"神": count = 6, mean = 0.004975, stddev = 0.011125
"祠": count = 6, mean = 0.000622, stddev = 0.001391
"祭": count = 6, mean = 0.000622, stddev = 0.001391
"私": count = 6, mean = 0.001658, stddev = 0.003708
"称": count = 6, mean = 0.001119, stddev = 0.001598
"移": count = 6, mean = 0.000324, stddev = 0.000724
"種": count = 6, mean = 0.001029, stddev = 0.0023
"空": count = 6, mean = 0.000498, stddev = 0.001112
"突": count = 6, mean = 0.000324, stddev = 0.000724
"立": count = 6, mean = 0.000324, stddev = 0.000724
"竣": count = 6, mean = 0.000324, stddev = 0.000724
"端": count = 6, mean = 0.001029, stddev = 0.0023
"筒": count = 6, mean = 0.000324, stddev = 0.000724
"管": count = 6, mean = 0.001658, stddev = 0.003708
"箱": count = 6, mean = 0.000324, stddev = 0.000724
"納": count = 6, mean = 0.000324, stddev = 0.000724
"紹": count = 6, mean = 0.000622, stddev = 0.001391
"組": count = 6, mean = 0.000647, stddev = 0.001447
"結": count = 6, mean = 0.000498, stddev = 0.001112
"線": count = 6, mean = 0.000829, stddev = 0.001854
"締": count = 6, mean = 0.000697, stddev = 0.001559
"縁": count = 6, mean = 0.001029, stddev = 0.0023
"縦": count = 6, mean = 0.000498, stddev = 0.001112
"置": count = 6, mean = 0.002439, stddev = 0.004261
"義": count = 6, mean = 0.002488, stddev = 0.005562
"翼": count = 6, mean = 0.000995, stddev = 0.002225
"者": count = 6, mean = 0.001658, stddev = 0.003708
"聖": count = 6, mean = 0.004353, stddev = 0.009734
"能": count = 6, mean = 0.000498, stddev = 0.001112
"自": count = 6, mean = 0.001892, stddev = 0.003094
"致": count = 6, mean = 0.000829, stddev = 0.001854

"航": count = 6, mean = 0.000498, stddev = 0.001112
"般": count = 6, mean = 0.001029, stddev = 0.0023
"舷": count = 6, mean = 0.000971, stddev = 0.002171
"船": count = 6, mean = 0.000971, stddev = 0.002171
"艦": count = 6, mean = 0.002589, stddev = 0.005789
"良": count = 6, mean = 0.000697, stddev = 0.001559
"荊": count = 6, mean = 0.000622, stddev = 0.001391
"落": count = 6, mean = 0.000697, stddev = 0.001559
"蓋": count = 6, mean = 0.003086, stddev = 0.006901
"薄": count = 6, mean = 0.001029, stddev = 0.0023
"薬": count = 6, mean = 0.000324, stddev = 0.000724
"行": count = 6, mean = 0.00199, stddev = 0.00445
"術": count = 6, mean = 0.000498, stddev = 0.001112
"衝": count = 6, mean = 0.000324, stddev = 0.000724
"装": count = 6, mean = 0.005376, stddev = 0.008658
"補": count = 6, mean = 0.000498, stddev = 0.001112
"製": count = 6, mean = 0.001345, stddev = 0.001904
"見": count = 6, mean = 0.000324, stddev = 0.000724
"角": count = 6, mean = 0.002705, stddev = 0.004539
"触": count = 6, mean = 0.002058, stddev = 0.004601
"計": count = 6, mean = 0.000647, stddev = 0.001447
"記": count = 6, mean = 0.000829, stddev = 0.001854
"許": count = 6, mean = 0.000829, stddev = 0.001854
"試": count = 6, mean = 0.000995, stddev = 0.002225
"認": count = 6, mean = 0.000829, stddev = 0.001854
"誕": count = 6, mean = 0.000697, stddev = 0.001559
"語": count = 6, mean = 0.001451, stddev = 0.002083
"諡": count = 6, mean = 0.000622, stddev = 0.001391
"警": count = 6, mean = 0.000697, stddev = 0.001559
"象": count = 6, mean = 0.000697, stddev = 0.001559
"貼": count = 6, mean = 0.000324, stddev = 0.000724
"贈": count = 6, mean = 0.001866, stddev = 0.004172
"路": count = 6, mean = 0.003814, stddev = 0.007276
"身": count = 6, mean = 0.001029, stddev = 0.0023
"車": count = 6, mean = 0.002921, stddev = 0.004675
"軍": count = 6, mean = 0.000324, stddev = 0.000724
"載": count = 6, mean = 0.000829, stddev = 0.001854
"輸": count = 6, mean = 0.000498, stddev = 0.001112
"込": count = 6, mean = 0.001021, stddev = 0.001582
"近": count = 6, mean = 0.000324, stddev = 0.000724
"返": count = 6, mean = 0.001029, stddev = 0.0023
"追": count = 6, mean = 0.001567, stddev = 0.00273
"送": count = 6, mean = 0.000498, stddev = 0.001112
"途": count = 6, mean = 0.000829, stddev = 0.001854
"通": count = 6, mean = 0.001195, stddev = 0.001725
"速": count = 6, mean = 0.001195, stddev = 0.001725

"造": count = 6, mean = 0.001668, stddev = 0.002406
"道": count = 6, mean = 0.01534, stddev = 0.031083
"達": count = 6, mean = 0.001195, stddev = 0.001725
"違": count = 6, mean = 0.001029, stddev = 0.0023
"遠": count = 6, mean = 0.001244, stddev = 0.002781
"部": count = 6, mean = 0.003144, stddev = 0.003696
"郵": count = 6, mean = 0.000697, stddev = 0.001559
"都": count = 6, mean = 0.001658, stddev = 0.003708
"配": count = 6, mean = 0.002639, stddev = 0.00431
"里": count = 6, mean = 0.000829, stddev = 0.001854
"量": count = 6, mean = 0.000498, stddev = 0.001112
"鉄": count = 6, mean = 0.001153, stddev = 0.001851
"錐": count = 6, mean = 0.001029, stddev = 0.0023
"鎮": count = 6, mean = 0.000622, stddev = 0.001391
"長": count = 6, mean = 0.000324, stddev = 0.000724
"門": count = 6, mean = 0.00448, stddev = 0.007215
"開": count = 6, mean = 0.00185, stddev = 0.002245
"間": count = 6, mean = 0.002488, stddev = 0.005562
"関": count = 6, mean = 0.004884, stddev = 0.008027
"闘": count = 6, mean = 0.000324, stddev = 0.000724
"防": count = 6, mean = 0.001021, stddev = 0.001582
"限": count = 6, mean = 0.001658, stddev = 0.003708
"除": count = 6, mean = 0.000498, stddev = 0.001112
"隆": count = 6, mean = 0.001244, stddev = 0.002781
"際": count = 6, mean = 0.001029, stddev = 0.0023
"障": count = 6, mean = 0.000697, stddev = 0.001559
"雄": count = 6, mean = 0.001029, stddev = 0.0023
"雌": count = 6, mean = 0.001029, stddev = 0.0023
"震": count = 6, mean = 0.000622, stddev = 0.001391
"霊": count = 6, mean = 0.001866, stddev = 0.004172
"面": count = 6, mean = 0.000647, stddev = 0.001447
"頂": count = 6, mean = 0.001029, stddev = 0.0023
"項": count = 6, mean = 0.000829, stddev = 0.001854
"順": count = 6, mean = 0.001244, stddev = 0.002781
"題": count = 6, mean = 0.000498, stddev = 0.001112
"顕": count = 6, mean = 0.001244, stddev = 0.002781
"類": count = 6, mean = 0.001029, stddev = 0.0023
"飛": count = 6, mean = 0.000995, stddev = 0.002225
"首": count = 6, mean = 0.000647, stddev = 0.001447
"馬": count = 6, mean = 0.000498, stddev = 0.001112
"駆": count = 6, mean = 0.000697, stddev = 0.001559
"験": count = 6, mean = 0.001493, stddev = 0.003337
"高": count = 6, mean = 0.001029, stddev = 0.0023
"魔": count = 6, mean = 0.001244, stddev = 0.002781