***Assignment week 1 (Johan, Laurence, Jarik)***

***Exercise 1:***

1A: All character types, alphabetically ordered:

['\n', '\r', ' ', '!', '"', '&', "'", '(', ')', ',', '-', '.', '/', '0', '1', '2', '3', '4', '5', '6', '7', '8', '9', ':', ';', '?', 'A', 'B', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z', 'a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', '\xa0', '\xa2', '\xa8', '\xa9', '\xc3']

There are 83 character types in total.

1B: There are 8550 word types, the first 20 (to print 8550 words is a bit much) for example are: ['!', '!"', "!'", "!'--", '!--', '"', '"\'', '",', '"--', '&', "'", "'-", "'--", "':", '(', '),', ')."', '):', ',', ',"']

1C:

Top 20 unigrams, ordered by descending frequency:

[((' ',), 94683), (('e',), 52863), (('t',), 37842), (('a',), 34374), (('o',), 33200), (('n',), 28378), (('h',), 27803), (('s',), 26432), (('i',), 26366), (('r',), 24304), (('d',), 18349), (('l',), 16854), (('u',), 13070), (('\n',), 12649), (('\r',), 12649), (('m',), 11055), (('w',), 10508), (('c',), 10172), (('y',), 8976), (('f',), 8785)])

Top 20 Bigrams, ordered by descending frequency:

[(('e', ' '), 16536), ((' ', 't'), 12954), (('\r', '\n'), 12649), (('t', 'h'), 11765), (('h', 'e'), 11472), ((' ', 'a'), 9948), (('d', ' '), 9539), (('t', ' '), 9149), (('s', ' '), 8237), (('i', 'n'), 7611), (('e', 'r'), 7096), ((' ', 'h'), 6767), ((' ', 'w'), 6757), ((',', ' '), 6581), (('a', 'n'), 6448), ((' ', 's'), 6233), (('n', ' '), 6172), (('h', 'a'), 5730), (('o', 'u'), 5641), (('r', 'e'), 5590)])

Top 20 Trigrams, ordered by descending frequency:

[((' ', 't', 'h'), 8712), (('t', 'h', 'e'), 7037), (('h', 'e', ' '), 6338), (('n', 'd', ' '), 3384), (('a', 'n', 'd'), 3369), ((' ', 'a', 'n'), 3283), (('e', 'd', ' '), 3073), ((' ', 't', 'o'), 2940), (('a', 't', ' '), 2811), (('i', 'n', 'g'), 2762), (('t', 'o', ' '), 2705), ((' ', 'o', 'f'), 2639), (('\n', '\r', '\n'), 2574), (('\r', '\n', '\r'), 2574), ((' ', 'h', 'a'), 2468), (('i', 's', ' '), 2428), ((' ', 'i', 'n'), 2420), (('o', 'f', ' '), 2411), (('a', 's', ' '), 2264), ((' ', 'I', ' '), 2244)])

1D: (a) the most likely 20 collocations using PMI

[('1', '/'), ('31', 'Lyon'), ('ARTHUR', 'CONAN'), ('Alicia', 'Whittington'), ('Amateur', 'Mendicant'), ('Amoy', 'River'), ('Anderson', '",'), ('Apache', 'Indians'), ('Arabian', 'Nights'), ('Arnsworth', 'Castle'), ('Atkinson', 'brothers'), ('BERYL', 'CORONET'), ('BLUE', 'CARBUNCLE'), ('BOSCOMBE', 'VALLEY'), ('Bengal', 'Artillery'), ('Bermuda', 'Dockyard'), ('Blue', 'Carbuncle'), ('Botany', 'variable'), ('CONAN', 'DOYLE'), ('COPPER', 'BEECHES')]

(b) the most likely 20 collocations using χ2[('1', '/'), ('31', 'Lyon'), ('ARTHUR', 'CONAN'), ('Alicia', 'Whittington'), ('Amateur', 'Mendicant'), ('Amoy', 'River'), ('Anderson', '",'), ('Apache', 'Indians'), ('Arabian', 'Nights'), ('Arnsworth', 'Castle'), ('Atkinson', 'brothers'), ('BERYL', 'CORONET'), ('BLUE', 'CARBUNCLE'), ('BOSCOMBE', 'VALLEY'), ('Bengal', 'Artillery'), ('Bermuda', 'Dockyard'), ('Beryl', 'Coronet'), ('Blue', 'Carbuncle'), ('Botany', 'variable'), ('Briony', 'Lodge')]

(c) just looking at them, there are surely differences in the rankings. Can you provide an explanation?

The first 16 items are equal, the last 4 differ from each other. The square-method doesn’t work well with bigrams not occuring very often, so maybe that’s the case why at the lower end of the top 20 differences between the PMI and square-method are being shown.