

Contemporary Data Processing Technology (CCOD)

Lab 10 (November 4, 2016)

Savchuk Artem, AS-36

Clustering Russian alphabet: I = {А, Б, В, Г, Д, Е, Ё, Ж, З, И, Й, К, Л, М, Н, О, П, Р, С, Т, У, Ф, Х, Ц, Ч, Ш, Щ, Ъ, Ы, Ъ, Э, Ю, Я};

Table of similarity percentage of letters:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33			
	А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я			
1	А		1	0,1	0,1	0,1	0,8	0,1	0,1	0,1	0,1	0,1	0,95	0,5	0,7	0,1	0,7	0,2	0,1	0,1	0,1	0,2	0,1	0,2	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,7	1	А		
2	Б		0,1		1	0,95	0,8	0,2	0,95	0,9	0,1	0,2	0,1	0,1	0,3	0,1	0,1	0,2	0,4	0,1	0,2	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,8	0,6	0,9	0,3	0,1	0,1	2	Б
3	В		0,1	0,95	1	0,2	0,1	0,95	0,9	0,1	0,9	0,1	0,1	0,4	0,1	0,1	0,1	0,5	0,1	0,3	0,1	0,2	0,1	0,3	0,1	0,1	0,1	0,1	0,7	0,6	0,8	0,4	0,1	0,2	3	В
4	Г		0,1	0,8	0,2	1	0,1	0,5	0,4	0,1	0,1	0,2	0,1	0,3	0,2	0,1	0,2	0,1	0,3	0,4	0,1	0,85	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	4	Г		
5	Д		0,8	0,2	0,1	0,1	1	0,1	0,1	0,1	0,1	0,1	0,9	0,2	0,2	0,1	0,6	0,1	0,1	0,2	0,1	0,1	0,2	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,2	5	Д	
6	Е		0,1	0,95	0,95	0,5	0,1	1	0,9	0,2	0,2	0,1	0,4	0,2	0,2	0,3	0,1	0,3	0,1	0,1	0,2	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,2	6	Е		
7	Ё		0,1	0,9	0,9	0,4	0,1	0,9	1	0,1	0,2	0,1	0,3	0,2	0,2	0,3	0,1	0,2	0,1	0,1	0,3	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,2	7	Ё			
8	Ж		0,1	0,1	0,1	0,1	0,1	0,2	0,1	1	0,1	0,2	0,1	0,6	0,2	0,2	0,1	0,2	0,2	0,1	0,1	0,2	0,1	0,9	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	8	Ж		
9	З		0,1	0,2	0,9	0,1	0,1	0,2	0,2	0,1	1	0,2	0,2	0,2	0,1	0,1	0,1	0,2	0,1	0,3	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,3	0,2	0,4	0,9	0,2	0,4	9	З
10	И		0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,2	2	1	0,95	0,2	0,1	0,4	0,1	0,1	0,2	0,1	0,2	0,1	0,2	0,4	0,2	0,2	0,2	0,1	0,1	0,2	0,1	0,1	10	И		
11	Й		0,1	0,1	0,1	0,1	0,1	0,1	0,2	0,1	0,2	0,95	1	0,2	0,1	0,3	0,1	0,1	0,3	0,2	0,2	0,2	0,2	0,4	0,3	0,3	0,2	0,1	0,1	0,1	0,1	11	Й			
12	К		0,1	0,3	0,4	0,3	0,1	0,4	0,3	0,6	0,2	0,2	1	0,2	0,3	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,6	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	12	К			
13	Л		0,95	0,1	0,1	0,2	0,9	0,2	0,2	0,1	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,2	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	0,1	13	Л			
14	М		0,9	0	0,8	0,8	0	0,8	0	0,8	0	0	0	0	0	0	0	0	0,85	0	0	0	0	0	0	0	0,8	0,8	0,8	0	0	4	Г			
15	Д		0,9	0	0	0	0	0	0	0,7	0	0	0	0	0	0	0	0,9	0,7	0,8	0,7	0,9	0	0	0	0	0	0	0	0,7	0,7	5	Д			
16	Е		0,95	0,95	0,8	0	0	0,9	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0,9	0,9	0	0	0	6	Е			
17	Ё		0,9	0,9	0,8	0	0,9	0	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0,9	0,8	0	0	0	7	Ё			
18	Ж		0,7	0	0	0	0,7	0	0	0	0	0	0,7	0,7	0,7	0,7	0,7	0	0	0	0,8	0,7	0,9	0	0,8	0	0	0	0	0,7	0,7	8	Ж			
19	З		0,9	0,9	0,8	0	0,9	0,9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0,9	0,8	0	0	0	9	З			
20	И		0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	10	И			
21	Й		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	11	Й			
22	К		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	12	К			
23	Л		0,95	0	0	0,9	0	0	0,7	0	0	0	0	0	0	0	0	0	0,7	0,8	0,7	0,9	0	0	0	0	0	0	0	0,7	0,7	13	Л			
24	М		0,7	0	0	0	0,7	0	0	0	0	0	0	0,7	0	0	0	0	0,7	0,7	0	0	0	0	0	0	0	0	0	0	0	14	М			
25	Д		0,8	0	0	0,8	0	0	0,7	0	0	0	0	0,8	0,7	0	0	0	0,7	0,8	0,7	0	0	0	0	0	0	0	0	0	0	15	Д			
26	Е		0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	16	Е			
27	Ё		0,7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	17	Ё		
28	Ж		0,9	0,9	0,8	0	0,9	0,9	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0	0	0	0	0	0	28	Ж		
29	З		0,8	0,8	0,8	0	0,8	0,8	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0	0	0	0	0	0	29	З		
30	И		0,9	0,9	0,8	0	0,9	0,9	0	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0	0	0	0	0	0	30	И		
31	Й		0,9	0,9	0,8	0	0,9	0,9	0	0	0	0	0	0	0	0	0	0	0,8	0,7	0,8	0	0	0	0	0	0	0	0	0	0	0	31	Й		
32	К		0,7	0	0	0	0,7	0	0	0	0	0	0	0,7	0,7	0,7	0,7	0	0	0	0,7	0,8	0	0	0	0	0	0	0	0	0	0	32	К		
33	Л		0,7	0	0	0	0,7	0	0	0	0	0	0	0,7	0,7	0,7	0,7	0	0	0	0,7	0,7	0	0	0	0	0	0	0	0	0	0	33	Л		

Then I repeated $R^{(n+1)} = R^{(n)} * R^{(n)}$ till $R^{(n)} = R^{(n+1)}$;

And assuming $\alpha = 0.6$, I got table that shown below:

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33		
	А	Б	В	Г	Д	Е	Ё	Ж	З	И	Й	К	Л	М	Н	О	П	Р	С	Т	У	Ф	Х	Ц	Ч	Ш	Щ	Ъ	Ы	Ь	Э	Ю	Я		
1	А		0	0	0,9	0	0	0,7	0	0	0	0	0,95	0,7	0,8	0,7	0,9	0	0	0	0,7	0,7	0	0,7	0	0	0	0	0	0	0	0,7	0,7	1	А
2	Б		0	0	0,95	0,8	0	0,95	0,9	0	0	0	0	0	0	0	0	0	0,8	0	0	0	0	0	0	0	0,9	0,9	0	0	0	2	Б		
3	В		0	0,95	0	0,8	0	0,95																											

1st cluster:

$I = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33\};$

$C_1 = \{\}$

$a_{2,6} = a_{3,6} = 0.95$ – maximum;

Take $a_{2,6}$, $C_1 = \{2, 6\}$;

$a_{2,3} + a_{6,3} = 1.9$ – maximum;

$C_1 = \{2, 6, 3\}$;

$a_{2,7} + a_{3,7} + a_{6,7} = 2.7$ – maximum;

$C_1 = \{2, 6, 3, 7\}$;

$a_{2,9} + a_{3,9} + a_{6,9} + a_{7,9} = 3.6$ – maximum;

$C_1 = \{2, 6, 3, 7, 9\}$;

Similarly, we include: 28, 30, 31, 4, 20, 29;

Finally we got 1st cluster: $C_1 = \{2, 6, 3, 7, 9, 28, 30, 31, 4, 20, 29\}$;

Letters: $C_1 = \{\text{Б, Е, В, Ѓ, 3, Ѓ, б, Ә, Г, Т, Ы}\}$;

2nd cluster:

$I = \{1, 5, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 32, 33\}$;

$C_2 = \{\}$;

$a_{10,11} = 0.95$ – maximum;

$C_2 = \{10, 11\}$;

There aren't maximum values, because almost all values in 10 and 11 row are 0;

So $C_2 = \{10, 11\}$;

Letters: $C_2 = \{\text{И, Й}\}$;

3rd cluster:

$I = \{1, 5, 8, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 32, 33\};$
 $C_2 = \{\};$
 $a_{1,13} = 0.95 - \text{maximum};$
 $C_3 = \{1, 13\};$

$a_{1,5} + a_{13,5} = 1.8 - \text{maximum};$
 $C_3 = \{1, 13, 5\}$

Similarly, we include: 17, 8, 15, 25, 14, 16, 22, 32, 21, 23, 33;

Finally we got 3rd cluster: $C_3 = \{1, 13, 5, 17, 8, 15, 25, 14, 16, 22, 32, 21, 23, 33\};$
Letters: $C_3 = \{А, Л, Д, П, Ж, Н, Ч, М, О, Ф, Ю, У, Х, Я\};$

4th cluster:

$I = \{12, 18, 19, 24, 26, 27\};$
 $C_4 = \{\};$
 $a_{26,27} = 0.9 - \text{maximum};$
 $C_4 = \{26, 27\};$

$a_{26,24} + a_{27,24} = 1.6 - \text{maximum};$
 $C_4 = \{26, 27, 24\};$

There isn't any maximum, so we stop append new letters into 4th cluster.

$C_4 = \{26, 27, 24\}$
Letters: $C_4 = \{Ш, Щ, Ц\};$

5th, 6th and 7th clusters:

$I = \{12, 18, 19\};$

Now all values in I is zero, so the remaining letters goes into different classes:

$C_5 = \{12\}$, $C_6 = \{18\}$ and $C_7 = \{19\}$;
Letters: $C_5 = \{К\}$, $C_6 = \{Р\}$ and $C_7 = \{С\}$;

Result:

$C = \{\{Б, Е, В, Ё, З, Ъ, ь, Э, Г, Т, ы\}, \{И, Й\}, \{А, Л, Д, П, Ж, Н, Ч, М, О, Ф, Ю, У, Х, Я\}, \{Ш, Щ, Ц\}, \{К\}, \{Р\} \text{ and } \{С\}\}$