

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ім. Ігоря Сікорського
Кафедра системного проектування

РОЗРАХУНКОВО-ГРАФІЧНА РОБОТА
з дисципліни

"Теорія інформації і кодування"

на тему: Програма автоматичного визначення
кодової таблиці текстового файлу

Студента 2го курсу
групи ДА-61

Кравченко Богдана Євгеновича
Керівник доц., к.т.н. Капшук О.О.

Зміст

- [Короткі відомості](#)
- [ASCII](#)
- [Windows-1251](#)
- [Unicode](#)
- [Версії Юнікод](#)
- [UTF-8](#)
- [UTF-16 і UTF32](#)
- [Розробка програми](#)
- [Інтерфейс програми](#)
- [Існуючі програми для перевірки кодування](#)
- [Тестування](#)
- [Висновок](#)
- [Список літератури](#)



Мета роботи

Розробити програму автоматичного визначення
кодової таблиці текстового файлу

Короткі відомості



- Безліч символів, за допомогою яких записується текст, називається **алфавітом**.
- Число символів в алфавіті - це його **потужність**.
- Формула визначення кількості інформації: $N = 2^b$, де N – потужність алфавіту (кількість символів), b - кількість біт (інформаційна вага символу).
- В алфавіт потужністю 256 символів можна помістити практично всі необхідні символи. Такий алфавіт називається **достатнім**.
- Оскільки $256 = 2^8$, то вага 1 символу - 8 біт.
- Одиниці виміру 8 біт присвоїли назву 1 байт:
- 1 байт = 8 біт.
- **Двійковий код кожного символу в комп'ютерному тексті займає 1 байт пам'яті.**

ASCII

```
!"#$%&'()*+,-./  
0123456789:;<=>?  
@ABCDEFGHIJKLMNO  
PQRSTUVWXYZ[\]^_  
`abcdefghijklmno  
pqrstuvwxyz{|}~
```

ASCII (англ. American Standard Code for Information Interchange) - американський стандартний код для обміну інформацією.

ASCII представляє собою кодування для представлення десяткових цифр, латинської та національного алфавітів, розділових знаків і керуючих символів. Спочатку розроблена як 7-бітна, з широким розповсюдженням 8-бітного байта ASCII стала сприйматися як половина 8-бітної.



Таблиця ASCII

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040	 	Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	\$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	((72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051))	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[[123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174	|	
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135]]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	>	95	5F	137	_	_	127	7F	177		DEL

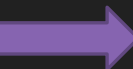
Windows-1251

Windows-1251 (також вживаються назви **Win1251**, **CP1251**) — кодування символів, що є стандартним 8-бітовим кодуванням для всіх локалізованих українських і російських версій Microsoft Windows. Користується досить великою популярністю. Була створена на базі кодувань, що використалися в ранніх «саморобних» русифікаторах Windows в 1990—1991 рр. спільно представниками «Параграфа», «Діалогу» і російського відділення Microsoft. Початковий варіант кодування помітно відрізнявся від сучасного, приведеного нижче в таблиці (зокрема, там було значне число «білих плям»).



Таблиця Windows-1251

	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	.A	.B	.C	.D	.E	.F
В.	Ъ 402	Ѓ 403	, 201A	ѓ 453	„ 201E	… 2026	† 2020	‡ 2021	€ 20AC	% 2030	Љ 409	< 2039	Њ 40A	Ќ 40C	Ѓ 40B	Ц 40F
Г.	ђ 452	‘ 2018	’ 2019	“ 201C	” 201D	• 2022	— 2013	— 2014		™ 2122	љ 459	> 203A	њ 45A	ќ 45C	ѓ 45B	ц 45F
Д.		Ў 40E	ў 45E	Ј 408	Ѡ A4	Ѓ 490	Ї A6	§ A7	Ё 401	© A9	Є 404	« AB	¬ AC		® AD	Ї 407
В.	° B0	± B1	І 406	і 456	г 491	μ B5	¶ B6	· B7	ё 451	№ 2116	є 454	» BB	ј 458	Ѕ 405	ѕ 455	ї 457
С.	А 410	Б 411	В 412	Г 413	Д 414	Е 415	Ж 416	З 417	И 418	Й 419	К 41A	Л 41B	М 41C	Н 41D	О 41E	П 41F
Д.	Р 420	С 421	Т 422	У 423	Ф 424	Х 425	Ц 426	Ч 427	Ш 428	Щ 429	Ъ 42A	Ы 42B	Ь 42C	Э 42D	Ю 42E	Я 42F
Е.	а 430	б 431	в 432	г 433	д 434	е 435	ж 436	з 437	и 438	й 439	к 43A	л 43B	м 43C	н 43D	о 43E	п 43F
Ф.	р 440	с 441	т 442	у 443	ф 444	х 445	ц 446	ч 447	ш 448	щ 449	ъ 44A	ы 44B	ь 44C	э 44D	ю 44E	я 44F



Має три недоліки:

- мала (рядкова) буква «я» має код 0xFF (255 в 10-овій системі). Вона є «винуватицею» ряду несподіваних проблем в програмах без підтримки чистого 8-го біту.
- відсутні символи псевдографіки.
- при сортуванні в алфавітному порядку літери не йдуть підряд, оскільки між літерами ŷŸіієЄііґґёё і основним блоком літер йдуть спецсимволи.

Unicode



Юнікод (англ. Unicode) - стандарт кодування символів, що включає в себе знаки майже всіх письмових мов світу. В даний час стандарт є домінуючим в Інтернеті. Стандарт запропонований в 1991 році некомерційною організацією «Консорціум Юнікоду» (англ. Unicode Consortium, Unicode Inc.). Застосування цього стандарту дозволяє закодувати дуже велике число символів з різних систем писемності: в документах, закодованих за стандартом Юнікод, можуть бути сусідами китайські ієрогліфи, математичні символи, букви грецького алфавіту, латиниці і кирилиці, символи музичної нотної нотації, при цьому стає непотрібним переключення кодових сторінок.



Стандарт складається з двох основних частин: універсального набору символів (англ. Universal character set, UCS) і сімейства кодувань (англ. Unicode transformation format, UTF). Універсальний набір символів перераховує допустимі за стандартом Юнікод символи і привласнює кожному символу код у вигляді невід'ємного цілого числа, що записується зазвичай в шістнадцятковій формі з префіксом U +, наприклад, U + 040F. Сімейство кодувань визначає способи перетворення кодів символів для передачі в потоці або в файлі. Коди в стандарті Юнікод розділені на кілька областей. Область з кодами від U + 0000 до U + 007F містить символи набору ASCII, і коди цих символів збігаються з їх кодами в ASCII. Далі розташовані області символів інших систем писемності, знаки пунктуації та технічні символи. Частина кодів зарезервована для використання в майбутньому. Під символи кирилиці виділені області знаків з кодами від U + 0400 до U + 052F, від U + 2DE0 до U + 2DFF, від U + A640 до U + A69F (див. Кирилиця в Юнікоде).



Способи представлення

Юнікод має кілька форм представлення (англ. Unicode transformation format, UTF): UTF-8, UTF-16 (UTF-16BE, UTF-16LE) і UTF-32 (UTF-32BE, UTF-32LE). Була розроблена також форма подання UTF-7 для передачі по семибітним каналах, але через несумісність з ASCII вона не набула поширення і не включена в стандарт.



Версії Юнікода



UTF-8

UTF-8 - уявлення Юнікода, що забезпечує найбільшу компактність і зворотну сумісність з 7-бітної системою ASCII; текст, що складається тільки з символів з номерами менше 128, при записі в UTF-8 перетворюється в звичайний текст ASCII і може бути відображений будь-якою програмою, що працює з ASCII; і навпаки, текст, закодований 7-бітної ASCII може бути відображений програмою, призначеної для роботи з UTF-8. Решта символів Юнікоду зображуються послідовностями довжиною від 2 до 4 байт, в яких перший байт завжди має маску 11xxxxxx, а решта - 10xxxxxx. В UTF-8 не використовуються сурогатні пари.



UTF-16 і UTF-32

- UTF-16 - кодування, що дозволяє записувати символи Юнікоду в діапазонах $U + 0000 \dots U + D7FF$ і $U + E000 \dots U + 10FFFF$ (загальною кількістю 1 112 064). При цьому кожен символ записується одним або двома словами (сурогатна пара). Кодування UTF-16 описана в додатку Q до міжнародного стандарту ISO / IEC 10646, а також їй присвячений документ IETF RFC 2781 під назвою «UTF-16, an encoding of ISO 10646».
- UTF-32 - спосіб представлення Юнікоду, при якому кожен символ займає рівно 4 байта. Головна перевага UTF-32 перед кодуваннями змінної довжини полягає в тому, що символи Юнікод в ній безпосередньо індексованих, тому знайти символ за номером його позиції в файлі можна надзвичайно швидко, і отримання будь-якого символу n -ї позиції при цьому є операцією, що займає завжди однакове час. Це також робить заміну символів в рядках UTF-32 дуже простий. Навпаки, кодування зі змінною довжиною вимагають послідовного доступу до символу n -ї позиції, що може бути дуже витратною за часом операцією. Головний недолік UTF-32 - це неефективне використання простору, так як для зберігання будь-якого символу використовується чотири байти. Символи, що лежать за межами нульовий (базової) площині кодового простору, рідко використовуються в більшості текстів. Тому подвоєння, в порівнянні з UTF-16, займаного рядками в UTF-32 простору, часто не виправдано.

Розробка програми

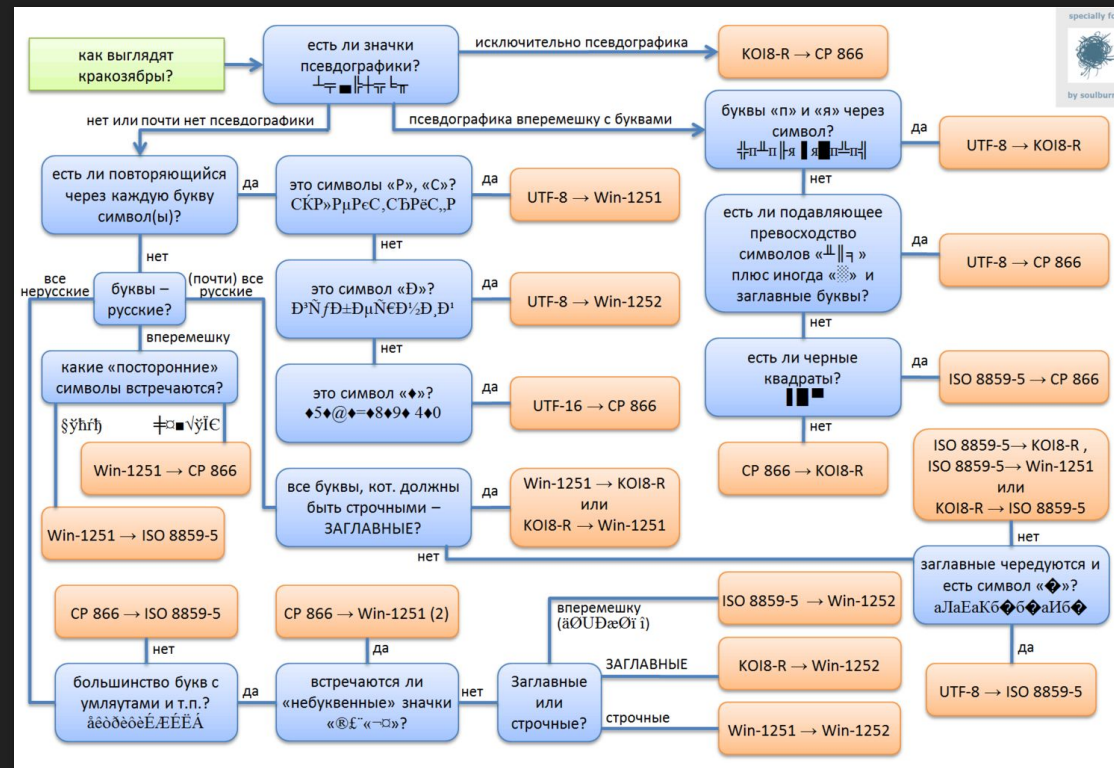
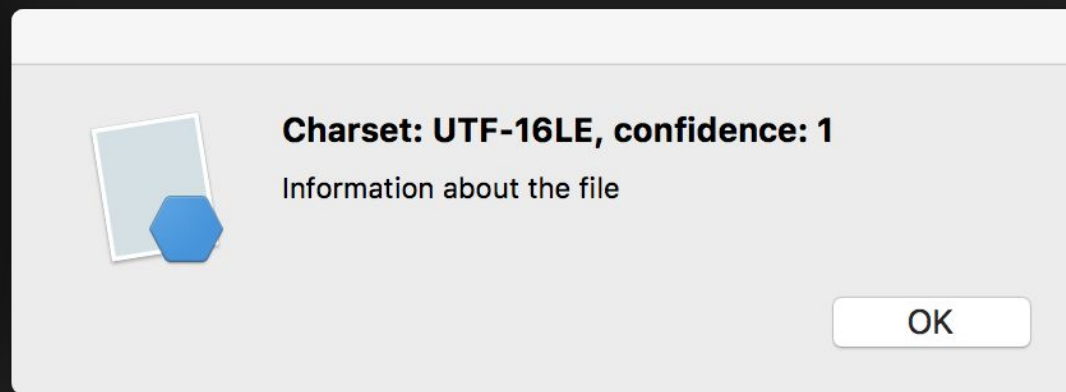
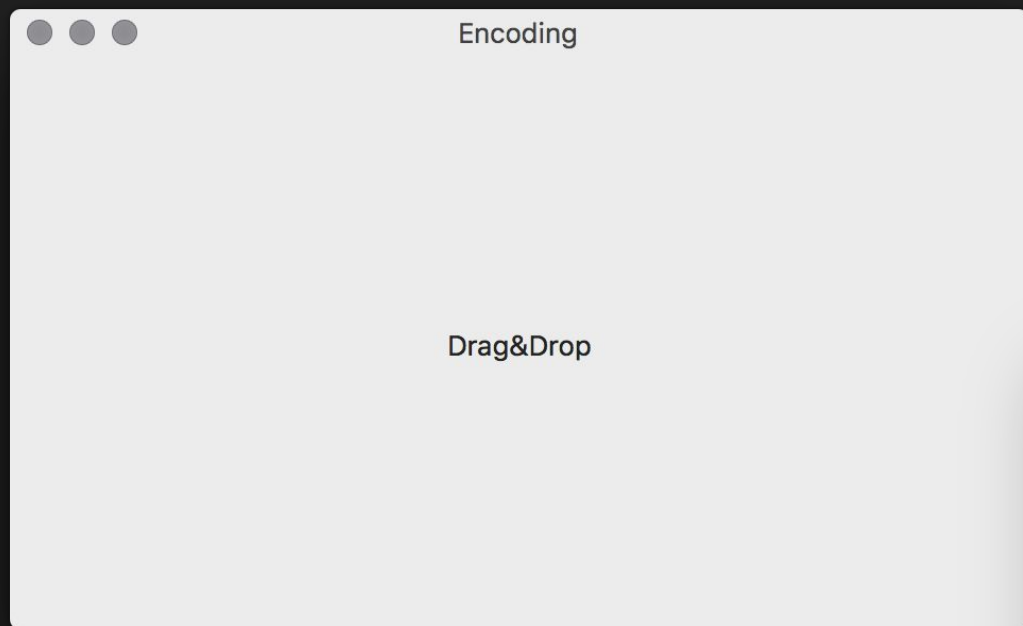


Рис. Алгоритм

Інтерфейс програми



Існуючі програми для перевірки кодування

NotePad++

```
1 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec condimentum molestie arcu, posuere scelerisque nisi semper vitae.
2 Pellentesque ut sem diam. Nullam finibus sagittis mi nec interdum. Nunc ac pulvinar dui, quis iaculis dui. Sed non mauris tortor.
3 Praesent imperdiet egestas congue. Curabitur ipsum lectus, consectetur ac accumsan non, blandit et erat.
4
5 Maecenas efficitur, lacus eget pharetra congue, nisl tellus egestas sapien, id condimentum sapien lectus ac urna.
6 Pellentesque sagittis suscipit dolor nec maximus. Donec varius hendrerit feugiat. Nullam vulputate dolor ut tempus faucibus.
7 Curabitur pretium, libero id euismod efficitur, dolor elit cursus est, at aliquet felis risus non metus. Pellentesque habitant
8 morbi tristique senectus et netus et malesuada fames ac turpis egestas. Quisque non odio ante. Proin dictum, elit id egestas
9 mattis, tortor felis faucibus neque, eget bibendum nulla mi sit amet ante. Lorem ipsum dolor sit amet, consectetur adipiscing
10 elit. Phasellus ipsum metus, aliquam eu placerat in, porta ut lectus. Nam purus dui, scelerisque nec felis at, cursus efficitur
11 lorem. Morbi lacinia interdum varius. Morbi orci velit, laoreet a tempus et, venenatis at dui. Sed tristique, sapien nec varius
12 rhoncus, lacus elit tincidunt nisi, eu tincidunt sem est at odio. Nam vel ex nunc. Donec pulvinar purus eget orci molestie, eget
13 blandit odio malesuada.
14
15 Vestibulum mattis egestas ultricies. Vivamus mollis auctor condimentum. Nullam risus ligula, ultrices a urna a, pellentesque
16 pretium justo. Pellentesque commodo venenatis dolor, ac lacinia nibh sollicitudin nec. Sed consequat tortor quis tempus sagittis.
17 Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Integer imperdiet volutpat tempor. Praesent
18 ultrices enim ac consectetur porta. Integer nec orci sit amet odio cursus imperdiet in ac quam. Curabitur a condimentum urna.
19 Etiam mi felis, congue non massa accumsan, viverra mattis ex. Proin pulvinar nunc sed sodales pulvinar.
20
21 Lorem ipsum dolor sit amet, consectetur adipiscing elit. Mauris lacinia enim orci, vitae vestibulum urna ornare sed.
22 Praesent cursus, mauris quis ullamcorper tempor, dolor mauris fermentum ipsum, ut ullamcorper ligula diam eu ligula.
23 In dui ex, mattis sed dui quis, placerat ultricies arcu. Phasellus sed odio luctus, volutpat mi vel, lacinia urna.
24 Aliquam a dolor quis diam convallis elementum. Proin ornare viverra orci, ac sollicitudin nisi iaculis sit amet.
25 Donec blandit dui nec egestas luctus. Nunc convallis ante in metus porta gravida. Integer malesuada massa a turpis malesuada
26 hendrerit et vestibulum nibh. Proin tincidunt, lectus eu pellentesque egestas, mauris ipsum ultricies dui, quis sagittis purus
27 eros eget purus. Duis finibus nibh porttitor, malesuada libero ac, porta arcu. Aenean dolor risus, posuere sed enim at, cursus
28 consequat ante. Nullam in dui malesuada, elementum magna eu, aliquet nunc.
29
30 Morbi ut tellus a odio viverra venenatis. Integer mattis, metus eget aliquet pretium, nibh nibh blandit nisi, a accumsan
31 ligula urna a nulla. Aliquam ut maximus risus. Vestibulum sit amet iaculis lectus, quis euismod erat. Cras eget tincidunt est.
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
```

<http://foxtools.ru/Text>

Универсальный декодер текста

Укажите текст, кодировку которого необходимо определить или изменить

-- выберите читаемый вариант из списка --

Исходная кодировка:

Конечная кодировка:

Автоматически определять:

Отправить

Очистить форму

Тестування програми

File name	Створена програма	Онлайн тест
euc.txt	EUC-JP	ISO-2022-JP
iso2022jp.txt	ISO-2022-JP	Shift-JIS
shiftjis.txt	Shift-JIS	EUC-JP
utf8.txt	UTF-8	UTF-8
utf8n.txt	UTF-8	UTF-8
utf16le.txt	UTF-16LE	UTF-16LE

ВИСНОВОК

При виконанні розрахунково-графічної роботи було розглянуто кодування файлів. Було описано основні кодування текстових файлів.

У процесі виконання було розроблено програму, що реалізує автоматичне визначення кодової таблиці текстового файлу. Було виконано ряд тестів, які підтвердили правильність роботи програми(правильність визначення кодування текстового файлу)



Список літератури

- Вернер.М. Основы кодирования. Учебник для ВУЗов. Москва: Техносфера. 2004. – 288с.
- Dave Tomas, Endi Hat — The Pragmatic Programmer, 1999
- https://ru.wikibooks.org/wiki/Кодирование_текста
- <http://school497.ru/download/u/02/les10/les.html>
- <https://uk.wikipedia.org/wiki/Windows-1251>
- <https://uk.wikipedia.org/wiki/KOI-8>
- <https://uk.wikipedia.org/wiki/CP866>
- <https://ru.wikipedia.org/wiki/MacCyrillic>
- <https://uk.wikipedia.org/wiki/UTF-8>
- <https://uk.wikipedia.org/wiki/UTF-16>



Дякую за увагу!