Лекция 6
Кодировки. Регулярные выражения

23 марта 2017 г.

Работа с кодировками

Кодировки

Кодировка

- Алгоритм *кодирования* E
  Символы $\rightarrow$ **Байты**
- Алгоритм *декодирования* D
  Байты $\rightarrow$ **Символы**

$$D = E^{-1}$$

Кодировки

Кодировка

- Алгоритм *кодирования* E
  Символы $\rightarrow$ **Байты**
- Алгоритм *декодирования* D
  Байты $\rightarrow$ **Символы**

$$D = E^{-1}$$

Обычно просто таблица **Символ** $\leftrightarrow$ **Байты**

ASCII

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

UTF

UTF-32

Символ → 4 байта.

a	0x00 0x00 0x00 0x61
b	0x00 0x00 0x00 0x62
Й	0x00 0x00 0x04 0x19

UTF-8

Символ → 1-4 байта.
Совпадает с ASCII на символах ASCII.

a	0x61
b	0x62
Й	0xD0 0x99

Стандарт Unicode

Стандарт Unicode

- UCS (Universal Character Set). Таблица символов.

a	97
b	98
Й	1049

Стандарт Unicode

- UCS (Universal Character Set).
  Таблица символов.

a	97
b	98
Й	1049

- UTF (Unicode Transformation Format).
  Кодировки UTF-8, UTF-16, UTF-32.

Тип `str`

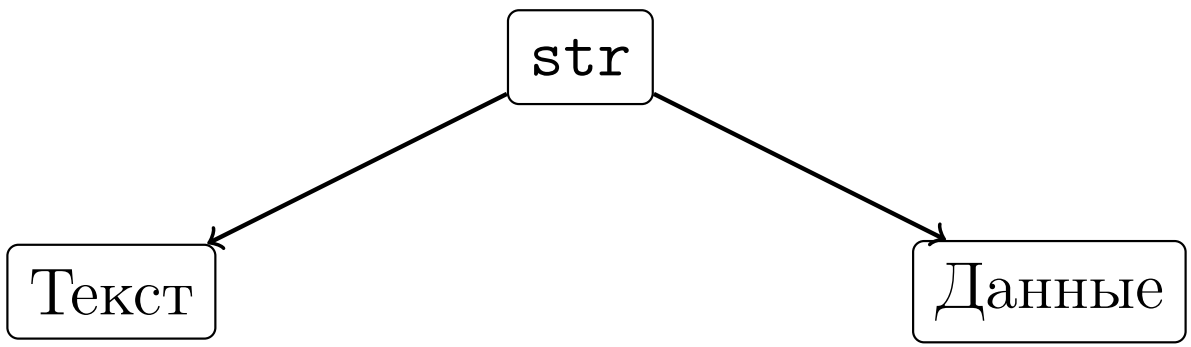
`str` — последовательность байт

`'a'`	0x61
`'\n'`	0x0A
`'\xff'`	0xFF

Тип `str`

`str` — последовательность байт

`'a'`	0x61
`'\n'`	0x0A
`'\xff'`	0xFF



Отображение `str`

```
>>> 'a'
'a'
>>> '\x61'
'a'
>>> '\x0a'
'\n'
>>> '\xff'
'\xff'
```

Отображение `str`

```
>>> 'a'
'a'
>>> '\x61'
'a'
>>> '\x0a'
'\n'
>>> '\xff'
'\xff'
>>> '\xf0\x20\xd5\x63\x33'
'\xf0 \xd5c3'
```

Тип `unicode`

`unicode` — последовательность символов.

`u'Q'`	Q
`u'щ'`	щ
`u'\u0449'`	щ

Тип unicode

`unicode` — последовательность символов.

`u'Q'`	Q
`u'щ'`	щ
`u'\u0449'`	щ

Отображение

```
>>> u'привет'
u'\u043f\u0440\u0438\u0432\u0435\u0442'
>>> u'йQ ?ф'
u'\u0439Q ?\u0444'
```

str и unicode

str и unicode

unicode → str

```
>>> u = u'эюя'
>>> s = u.encode('utf-8')
>>> s
'\xd1\x8d\xd1\x8e\xd1\x8f'
```

str и unicode

unicode → str

```
>>> u = u'эюя'
>>> s = u.encode('utf-8')
>>> s
'\xd1\x8d\xd1\x8e\xd1\x8f'
```

str → unicode

```
>>> u2 = s.decode('utf-8')
>>> u2
u'\u044d\u044e\u044f'
```

Кодирование данных

```
with open('in.txt') as fin:
    data_in = fin.read()
text_in = data_in.decode('utf-8')
```

Кодирование данных

```
with open('in.txt') as fin:
    data_in = fin.read()
text_in = data_in.decode('utf-8')


... = text_in
...
text_out = ...
```

Кодирование данных

```
with open('in.txt') as fin:
    data_in = fin.read()
text_in = data_in.decode('utf-8')


... = text_in
...
text_out = ...

data_out = text_out.encode('utf-8')
with open('out.txt', 'w') as fout:
    fout.write(data_out)
```

Кодирование данных и codecs

```
import codecs

with codecs.open('int.txt',
                 encoding='utf-8') as fin:
    text_in = fin.read()

... = text_in
...
text_out = ...

with codecs.open('out.txt', 'w',
                 encoding='utf-8') as fout:
    fout.write(text_out)
```

Кодирование исходного кода

```
# -*- coding: utf-8 -*-

x = u'эюя'
# Комментарий
```

Строки в Python 3

Тип	Запись	Аналог в Python 2
`str`	`'Привет'`	`unicode`
`bytes`	`b'abc\xff'`	`str`

Регулярные выражения

Регулярные языки

Алфавит $\mathcal{A} = \{0, 1\}$

Выражение	**Формальный язык**
0	$\{0\}$
$(0\|1)1$	$\{01, 11\}$
001^+	$\{001, 0011, 00111, \ldots\}$
0^*10^*	$\{1, 01, 10, 010, 001, \ldots\}$

Простые регулярные выражения

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
import re
```

Простые регулярные выражения

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
import re

m = re.search('i', text)
print m
```

```
<_sre.SRE_Match object at 0x3693648>
```

Простые регулярные выражения

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
import re

m = re.search('i', text)
print m
```

```
<_sre.SRE_Match object at 0x3693648>
```

```
print m.start(), m.end()
```

```
48 49
```

Простые регулярные выражения

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
def search_results(pattern, text):
    for m in re.finditer(pattern, text):
        print "'{0}': {1}-{2}".format(
            m.group(), m.start(), m.end())
```

Простые регулярные выражения

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
def search_results(pattern, text):
    for m in re.finditer(pattern, text):
        print "'{0}': {1}-{2}".format(
            m.group(), m.start(), m.end())

search_results('i', text)
```

```
'i': 48-49
'i': 73-74
'i': 75-76
'i': 79-80
```

Множества символов

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('(i|I)n', text)
```

Множества символов

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('(i|I)n', text)
```

```
'In': 0-2
'in': 79-81
```

Множества символов

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('(i|I)n', text)
```

```
'In': 0-2
'in': 79-81
```

```
search_results('[0-9] ', text)
```

Множества символов

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('(i|I)n', text)
```

```
'In': 0-2
'in': 79-81
```

```
search_results('[0-9] ', text)
```

```
'9 ': 16-18
'7 ': 84-86
```

Экранирование символов

```
search_results('\\', '\\')
```

Экранирование символов

```
search_results('\\', '\\')
```

```
error: bogus escape (end of line)
```

Экранирование символов

```
search_results('\\', '\\')
```

```
error: bogus escape (end of line)
```

```
search_results('\\\\', '\\')
```

```
'\': 0-1
```

Экранирование символов

```
search_results('\\', '\\')
```

```
error: bogus escape (end of line)
```

```
search_results('\\\\', '\\')
```

```
'\': 0-1
```

```
search_results(r'\\', '\\')
```

```
'\': 0-1
```

Любой символ

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""

search_results('s.', text)
```

Любой символ

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('s.', text)
```

```
'sh': 18-20
'st': 50-52
'ss': 64-66
'si': 74-76
's.': 90-92
```

Любой символ

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('s.', text)
```

```
'sh': 18-20
'st': 50-52
'ss': 64-66
'si': 74-76
's.': 90-92
```

```
search_results(r's\.', text)
```

Любой символ

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('s.', text)
```

```
'sh': 18-20
'st': 50-52
'ss': 64-66
'si': 74-76
's.': 90-92
```

```
search_results(r's\.', text)
```

```
's.': 90-92
```

Звезда и плюс Клини

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""

 search_results('sh*e', text)
```

Звезда и плюс Клини

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('sh*e', text)
```

```
'she': 18-21
'se': 65-67
```

Звезда и плюс Клини

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('sh*e', text)
```

```
'she': 18-21
'se': 65-67
```

```
search_results('.s+.', text)
```

Звезда и плюс Клини

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('sh*e', text)
```

```
'she': 18-21
'se': 65-67
```

```
search_results('.s+.', text)
```

```
' sh': 17-20
'rst': 49-52
'esse': 63-67
'isi': 73-76
'rs.': 89-92
```

Другие диапазоны

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""


search_results('i?n', text)
```

Другие диапазоны

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('i?n', text)
```

```
'n': 1-2
'n': 39-40
'n': 60-61
'in': 79-81
```

Другие диапазоны

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('i?n', text)
```

```
'n': 1-2
'n': 39-40
'n': 60-61
'in': 79-81
```

```
search_results('[0-9]{4,6}', text)
```

Другие диапазоны

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('i?n', text)
```

```
'n': 1-2
'n': 39-40
'n': 60-61
'in': 79-81
```

```
search_results('[0-9]{4,6}', text)
```

```
'1769': 13-17
```

Жадность поиска

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""

search_results('e.*e', text)
```

Жадность поиска

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('e.*e', text)
```

```
'eptember 1769 she reached Ne': 4-32
'ealand, the first European vesse': 35-67
```

Жадность поиска

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('e.*e', text)
```

```
'eptember 1769 she reached Ne': 4-32
'ealand, the first European vesse': 35-67
```

```
search_results('e.*?e', text)
```

Жадность поиска

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results('e.*e', text)
```

```
'eptember 1769 she reached Ne': 4-32
'ealand, the first European vesse': 35-67
```

```
search_results('e.*?e', text)
```

```
'epte': 4-8
'er 1769 she': 10-21
'eache': 23-28
'ealand, the': 35-46
'ean ve': 58-64
```

Другие спецсимволы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""

search_results(r'n\s+', text)
```

Другие спецсимволы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results(r'n\s+', text)
```

```
'n ': 1-3
'n ': 60-62
'n ': 80-82
```

Другие спецсимволы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results(r'n\s+', text)
```

```
'n ': 1-3
'n ': 60-62
'n ': 80-82
```

```
search_results(r't\S', text)
```

Другие спецсимволы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
search_results(r'n\s+', text)
```

```
'n ': 1-3
'n ': 60-62
'n ': 80-82
```

```
search_results(r't\S', text)
```

```
'te': 6-8
'th': 43-45
'to': 69-71
```

Группы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""


m = re.search('(\w+)\s*([0-9]+)', text)
print m.group(0)
```

```
September 1769
```

Группы

```
text = """In September 1769 she reached New
Zealand, the first European vessel to visit
in 127 years."""
```

```
m = re.search('(\w+)\s*([0-9]+)', text)
print m.group(0)
```

```
September 1769
```

```
print m.group(1)
print m.group(2)
```

```
September
1769
```

Другие возможности

Функции

- `match` — поиск только в первой позиции.
- `split` — разбиение текста на части.
- `sub` — замена в тексте.
- `compile` — компиляция выражения для многократного использования.

Другие возможности

Функции

- `match` — поиск только в первой позиции.
- `split` — разбиение текста на части.
- `sub` — замена в тексте.
- `compile` — компиляция выражения для многократного использования.

Флаги

- `IGNORECASE`
- `UNICODE`
- `VERBOSE`
- `MULTILINE`