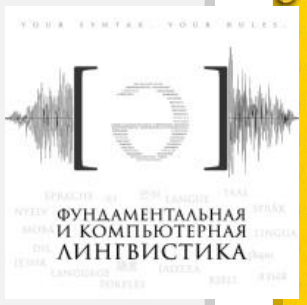
Ольга
Ляшевская

Распаковываем текст

Структурированное
представление
текстовых данных



Курс
«Лингвистические данные»
НИУ ВШЭ, ФиКЛ, 1 курс бакалавриата

Что сегодня?

- Форматы представления данных
- Кодировки файлов
- Поиск по (не)структурированному тексту

Форматы представления данных

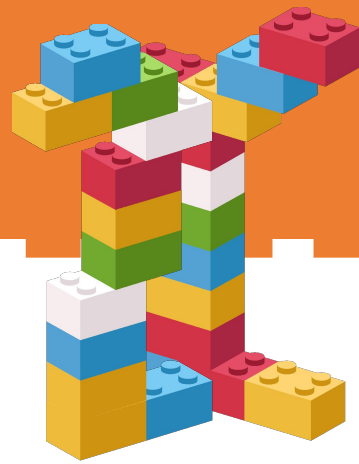
- Текст:
  - содержание
  - структура - логическая и визуальная
  - шрифтовое оформление - выделение элементов внутри структурных блоков
    (также бывает не только визуальным, но и логическим)
- А кто читает текст?

Форматы представления данных

- Текст:
  - содержание
  - структура - логическая и визуальная
  - шрифтовое оформление - выделение элементов внутри структурных блок
    (также бывает не только визуальным, но и логическим)
- Текст для человека и текст *для машины*
  - кодировка текста
  - расширение файла - для графических редакторов
  - метаданные в начале файла
  - формат конца строки

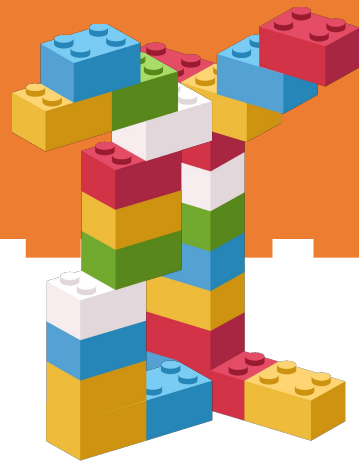
```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN" "http://www.w3.org/TR/html4
```

Расширения файлов



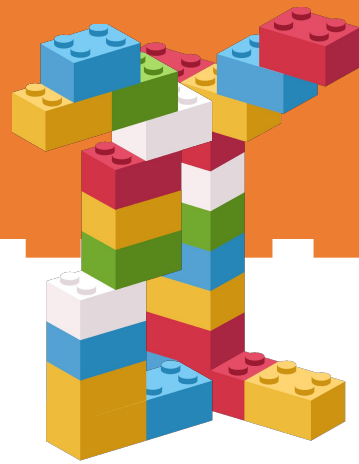
- … и ассоциированные типы кодировок
  - txt - чистый текст (bare text)
  - doc (docx, odt) - форматы для офисных документов
  - xls (xlsx, ods) - табличные форматы
  - …
  - pdf - Portable Document Format
- переходные форматы (кросс-форматы):
  - csv, tsv, rtf

Форматы для разметки



- … и ассоциированные типы расширений
  - HTML-формат
  - Вики-формат
  - Markdown
  - XML
  - JSON
  - YAML
  - лингвистических редакторов (*.eaf* для ELAN, *.TextGrid* и *.Pitch* для Praat, *.exb* для EXMARaLDA…)

Пример слоев лингвистической разметки



Salut, Tom!

	0	1	2	3
TOM [c]	*waving*			
TOM [v]	Hello,	Tim!		
TOM [a]	Salut, Tim!			
TIM [c]		*waving*		
TIM [v]		Hello,	Tom.	
TIM [a]		Salut, Tom!		

Кодировки текста

- Несколько названий:
  - ASCII
  - Cyrillic Windows (1251)
  - KOI8-R
  - Unicode (UTF-8)

Кодировки текста: ASCII

Bits b_4	b_3	b_2	b_1	Row ↓ / Column →	0 (000)	1 (001)	2 (010)	3 (011)	4 (100)	5 (101)	6 (110)	7 (111)
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	'	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	|
1	1	0	1	13	CR	GS	—	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

Кодировки текста: KOI-8

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
80	2500 ─	2502 │	250C ┌	2510 ┐	2514 └	2518 ┘	251C ├	2524 ┤	252C ┬	2534 ┴	253C ┼	2580 ▀	2584 ▄	2588 █	258C ▌	2590 ▐
90	2591 ░	2592 ▒	2593 ▓	2320 ⌠	25A0 ■	2219 ∙	221A √	2248 ≈	2264 ≤	2265 ≥	A0	2321 ⌡	B0 °	B2 ²	B7 ·	F7 ÷
A0	2550 ═	2551 ║	2552 ╒	451 ё	2553 ╓	2554 ╔	2555 ╕	2556 ╖	2557 ╗	2558 ╘	2559 ╙	255A ╚	255B ╛	255C ╜	255D ╝	255E ╞
B0	255F ╟	2560 ╠	2561 ╡	401 Ё	2562 ╢	2563 ╣	2564 ╤	2565 ╥	2566 ╦	2567 ╧	2568 ╨	2569 ╩	256A ╪	256B ╫	256C ╬	A9 ©
C0	44E ю	430 а	431 б	446 ц	434 д	435 е	444 ф	433 г	445 х	438 и	439 й	43A к	43B л	43C м	43D н	43E о
D0	43F п	44F я	440 р	441 с	442 т	443 у	436 ж	432 в	44C ь	44B ы	437 з	448 ш	44D э	449 щ	447 ч	44A ъ
E0	42E Ю	410 А	411 Б	426 Ц	414 Д	415 Е	424 Ф	413 Г	425 Х	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О
F0	41F П	42F Я	420 Р	421 С	422 Т	423 У	416 Ж	412 В	42C Ь	42B Ы	417 З	428 Ш	42D Э	429 Щ	427 Ч	42A Ъ

Предлагаемые решения

- Эксплицитное распределение ролей в группе, индивидуальные роли по Р. Белбину (1993)
- Ощущение успешности себя как профессионала, социальной успешности через успешность всей команды. “Successful collectives demonstrate not just individual but collective intelligence”(Alberola et al. 2013, Gupta & Dardaman 2023)
- Выделение роли тим-лида, который готов разобраться в задачах, коммуницировать задачи, заниматься тайм-менеджментом
- Роль тим-лида получают студенты, которые успешно представили проект, его задачи и критерии, и за которых проголосовали потенциальные участники
- Техническое задание и стадии определяются командой, а не куратором
- Опрос об оценке своего вклада, оценке группы и группового взаимодействия как вспомогательный инструмент оценивания - большая объективность?
- Куратор-преподаватель - скорее фасилитатор и эксперт, нежели менеджер

Кодировки текста: Юникод

https://symbl.cc/ru/