CLI Applications: argparse, logging, unicode

Dral Alexey, aadral@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org/
https://www.facebook.com/bigdatateam/

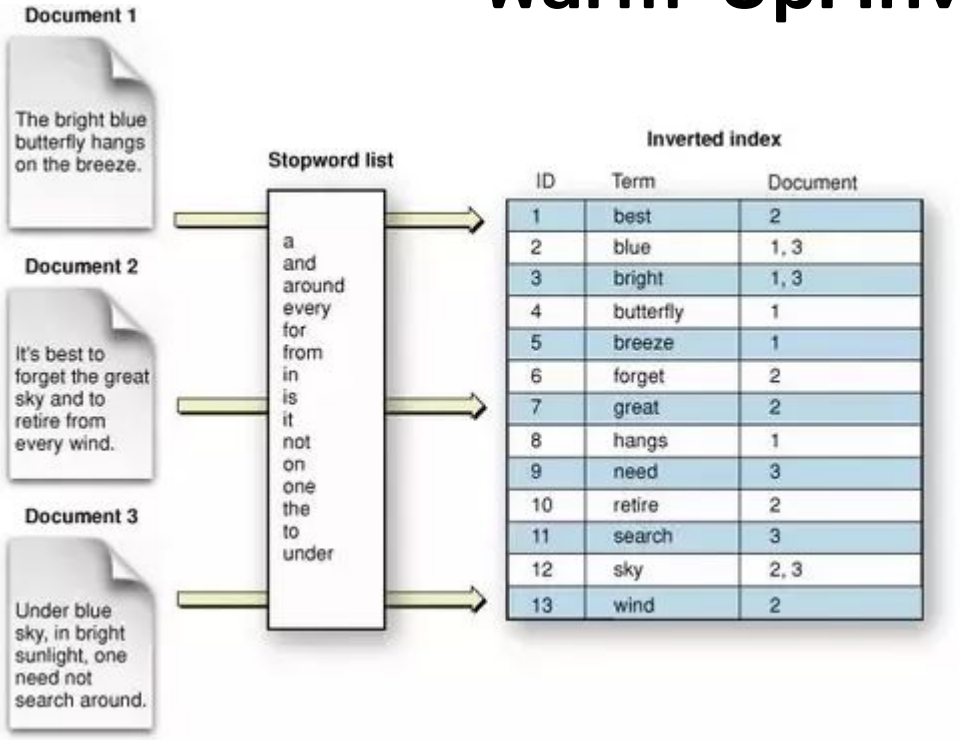
{22,29}.10.2019, Moscow, Russia

Outline

- Using argparse
- Working with Unicode
- Setting up logging

BIGDATA TEAM

Warm-Up: Inverted Index



- input format: "`document_ID <tab> content`"
- index format: "`{term: set<document_ID>}`"

Initial Solution Overview

- https://gitlab.com/snippets/1906309

argparse

The argparse module makes it easy to write user-friendly command-line interfaces. The program defines what arguments it requires, and argparse will figure out how to parse those out of sys.argv. The argparse module also automatically generates help and usage messages and issues errors when users give the program invalid arguments.

© https://docs.python.org/3/library/argparse.html

CLI Workshop

- http://rebrand.ly/mailpy19q4_02_cli_workshop

Unicode

- **Unicode** is a specification that aims to list every character used by human languages and give each character its own unique code.
- A **character** is the smallest possible component of a text (‘A’, ‘B’, ‘C’, ‘È’, ‘Í’, ‘췵‘, ...)
- The Unicode standard describes how **characters** are represented by **code points**.

© https://docs.python.org/3/howto/unicode.html

BIGDATA TEAM

koi8-r encoding

KOI8-R

	_0	_1	_2	_3	_4	_5	_6	_7	_8	_9	_A	_B	_C	_D	_E	_F
0_ 0																
1_ 16																
2_ 32	SP 0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3_ 48	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4_ 64	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5_ 80	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6_ 96	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7_ 112	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	
8_ 128	─ 2500	│ 2502	┌ 250C	┐ 2510	└ 2514	┘ 2518	├ 251C	┤ 2524	┬ 252C	┴ 2534	┼ 253C	▀ 2580	▄ 2584	█ 2588	▌ 258C	▐ 2590
9_ 144	░ 2591	▒ 2592	▓ 2593	⌠ 2320	■ 25A0	· 2219	√ 221A	≈ 2248	≤ 2264	≥ 2265	NBSP 00A0	⌡ 2321	° 00B0	² 00B2	· 00B7	÷ 00F7
A_ 160	═ 2550	║ 2551	╒ 2552	ё 0451	╓ 2553	╔ 2554	╕ 2555	╖ 2556	╗ 2557	╘ 2558	╙ 2559	╚ 255A	╛ 255B	╜ 255C	╝ 255D	╞ 255E
B_ 176	╟ 255F	╠ 2560	╡ 2561	Ё 0401	╢ 2562	╣ 2563	╤ 2564	╥ 2565	╦ 2566	╧ 2567	╨ 2568	╩ 2569	╪ 256A	╫ 256B	╬ 256C	© 00A9
C_ 192	ю 044E	а 0430	б 0431	ц 0446	д 0434	е 0435	ф 0444	г 0433	х 0445	и 0438	й 0439	к 043A	л 043B	м 043C	н 043D	о 043E
D_ 208	п 043F	я 044F	р 0440	с 0441	т 0442	у 0443	ж 0436	в 0432	ь 044C	ы 044B	з 0437	ш 0448	э 044D	щ 0449	ч 0447	ъ 044A
E_ 224	Ю 042E	А 0410	Б 0411	Ц 0426	Д 0414	Е 0415	Ф 0424	Г 0413	Х 0425	И 0418	Й 0419	К 041A	Л 041B	М 041C	Н 041D	О 041E
F_ 240	П 041F	Я 042F	Р 0420	С 0421	Т 0422	У 0423	Ж 0416	В 0412	Ь 042C	Ы 042B	З 0417	Ш 0428	Э 042D	Щ 0429	Ч 0427	Ъ 042A

5	0438	0439	043A	043B	043C	043D
C	ы 044B	з 0437	ш 0448	э 044D	щ 0449	ч 0447
5	и 0418	й 0419	к 041A	л 041B	м 041C	н 041D

© https://en.wikipedia.org/wiki/KOI8-R

utf-8 encoding

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

© https://en.wikipedia.org/wiki/UTF-8

CLI Workshop

- http://rebrand.ly/mailpy19q4_03_cli_workshop

Logging

Level	When it's used
DEBUG	Detailed information, typically of interest only when diagnosing problems.
INFO	Confirmation that things are working as expected.
WARNING	An indication that something unexpected happened, or indicative of some problem in the near future (e.g. 'disk space low'). The software is still working as expected.
ERROR	Due to a more serious problem, the software has not been able to perform some function.
CRITICAL	A serious error, indicating that the program itself may be unable to continue running.

Logging Workshop

- https://rebrand.ly/mailpy19q4_04_logging_workshop

References

- argparse: https://docs.python.org/3/library/argparse.html
- logging: https://docs.python.org/3/howto/logging.html https://docs.python.org/3/howto/logging-cookbook.html
- unicode: https://docs.python.org/3/howto/unicode.html

Summary

- You can use argparse functionality: ArgumentParser with default values formatter, subparsers, actions (e.g. FileType, nargs, choices)
- You can explain the concept of Unicode and the difference with utf-8. You can encode and decode texts to and from binary formats
- You can setup logging for your application and can explain the usage of different levels

BIGDATA TEAM

Thank you! Questions?

Dral Alexey, aadral@bigdatateam.org
CEO at BigData Team, http://bigdatateam.org/
https://www.linkedin.com/in/alexey-dral
https://www.facebook.com/bigdatateam/