Artificial intelligence in data science

Text prediction

Janos Török

Department of Theoretical Physics

November 9, 2022

Working with text

- Real nightmare
- Lot of data (e.g. books, chats, tweets, etc.)
- Number of languages ~ 6500
- Number of really spoken languages?
  - According to Wikipedia 100th language has 7.5million native speakers
  - Wikipedia with at least 100 pages: 282 languages
- Writing: left to right, right to left, symbols (Chinese)

Encoding text

- ASCII table: American Standard Code for Information Interchange
- 8 bit: 256 different possibilities

ASCII Table

Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char	Dec	Hex	Oct	Char
0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
1	1	1		33	21	41	!	65	41	101	A	97	61	141	a
2	2	2		34	22	42	"	66	42	102	B	98	62	142	b
3	3	3		35	23	43	#	67	43	103	C	99	63	143	c
4	4	4		36	24	44	$	68	44	104	D	100	64	144	d
5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
7	7	7		39	27	47	'	71	47	107	G	103	67	147	g
8	8	10		40	28	50	(	72	48	110	H	104	68	150	h
9	9	11		41	29	51	)	73	49	111	I	105	69	151	i
10	A	12		42	2A	52	*	74	4A	112	J	106	6A	152	j
11	B	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	l
13	D	15		45	2D	55	-	77	4D	115	M	109	6D	155	m
14	E	16		46	2E	56	.	78	4E	116	N	110	6E	156	n
15	F	17		47	2F	57	/	79	4F	117	O	111	6F	157	o
16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
19	13	23		51	33	63	3	83	53	123	S	115	73	163	s
20	14	24		52	34	64	4	84	54	124	T	116	74	164	t
21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
23	17	27		55	37	67	7	87	57	127	W	119	77	167	w
24	18	30		56	38	70	8	88	58	130	X	120	78	170	x
25	19	31		57	39	71	9	89	59	131	Y	121	79	171	y
26	1A	32		58	3A	72	:	90	5A	132	Z	122	7A	172	z
27	1B	33		59	3B	73	;	91	5B	133	[	123	7B	173	{
28	1C	34		60	3C	74	<	92	5C	134	\	124	7C	174	|
29	1D	35		61	3D	75	=	93	5D	135	]	125	7D	175	}
30	1E	36		62	3E	76	>	94	5E	136	^	126	7E	176	~
31	1F	37		63	3F	77	?	95	5F	137	_	127	7F	177	

Encoding text

- ASCII table: American Standard Code for Information Interchange
- 8 bit: 256 different possibilities
- Latin-1: ä,ö,ü,û,à
- Latin-2: á,ő,Ű,í
- Unicode: 16 bit characters $\rightarrow$ died before it could live, but still exists!
- Encoding: utf-8: Special characters:

Bits of code point	First code point	Last code point	Bytes in sequence	Byte 1	Byte 2	Byte 3	Byte 4	Byte 5	Byte 6
7	U+0000	U+007F	1	0xxxxxxx					
11	U+0080	U+07FF	2	110xxxxx	10xxxxxx				
16	U+0800	U+FFFF	3	1110xxxx	10xxxxxx	10xxxxxx			
21	U+10000	U+1FFFFF	4	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx		
26	U+200000	U+3FFFFFF	5	111110xx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	
31	U+4000000	U+7FFFFFFF	6	1111110x	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx	10xxxxxx

Lucky world

- English is just the perfect choice
- Short words
- No fusion or hardly any conjugation
- Very few letters, and all are available as simple ascii

Make the computer understand the text

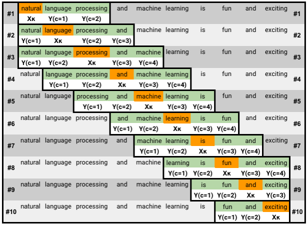
- Analyze the word (problems with same form) e.g. leaves (what trees have and what someone does at the end of the class)
- Get meaning → stem
- Always use purpose made tool on you own language (hunmorph for Hungarian)

```
echo "alkalmatlanok" | ./src/wrappers/ocamorph/ocamorph
--aff ../morphdb.hu/morphdb_hu.aff \
--dic ../morphdb.hu/morphdb_hu.dic
> alkalmatlanok
alkalmatlan/NOUN<PLUR>
alkalmatlan/ADJ<PLUR>
alkalom/NOUN[NEG_ATTRIB]/ADJ<PLUR>
alkalom/NOUN[NEG_ATTRIB]/ADJ<PLUR>
```

Words to vector

Mikolov et al. 2013

- Try to predict parts of text
- Take sentences
- consider 5 word grams
- encode them using one hot encoding



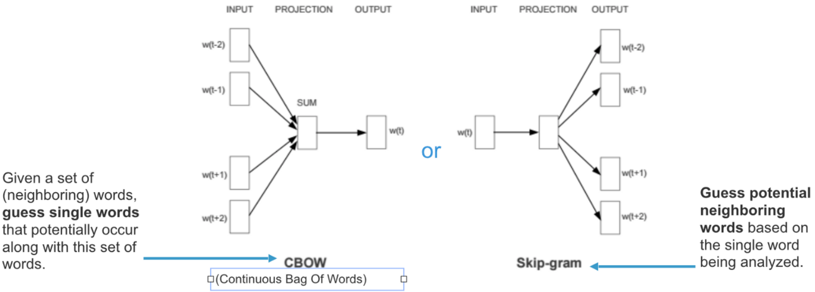
#	Token	#1			#2				#3					#4					#5				
		Xк	Y(c=1)	Y(c=2)	Xк	Y(c=1)	Y(c=2)	Y(c=3)	Xк	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	Xк	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	Xк	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)
0	natural	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2	processing	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
3	and	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
4	machine	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
5	learning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
6	is	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	fun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	exciting	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

#	Token	#6					#7					#8					#9				#10		
		Xк	Y(c=1)	Y(c=2)	(c=:	Y(c=4)	Xк	Y(c=1)	(c=:	Y(c=3)	Y(c=4)	Xк	Y(c=1)	(c=:	Y(c=3)	Y(c=4)	Xк	Y(c=1)	(c=:	Y(c=3)	Xк	Y(c=1)	Y(c=2)
0	natural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	processing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	and	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
4	machine	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	learning	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	is	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
7	fun	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
8	exciting	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0

derekchia.com

Words prediction

- Word is determined by neighboring word and of course context.
- Two way of guessing



Encoding

- Set of words
- Extra words at end of sentence extra encoding

3. One-hot encoding

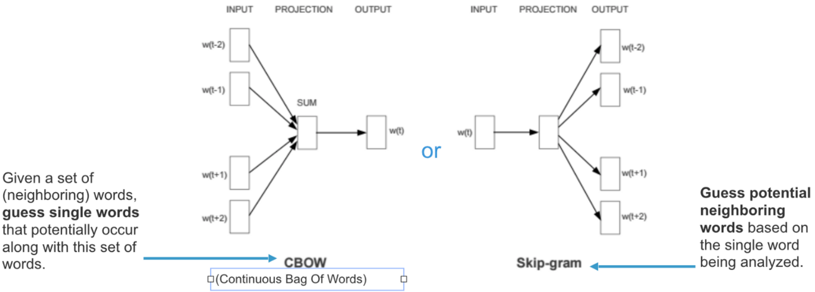
derekchia.com

#	Token	#1			#2				#3					#4					#5				
0	natural	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0
2	processing	0	0	1	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
3	and	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0
4	machine	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
5	learning	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
6	is	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
7	fun	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	exciting	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		X_K	Y(c=1)	Y(c=2)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)

#	Token	#6					#7					#8					#9				#10		
0	natural	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	language	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	processing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	and	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1
4	machine	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	learning	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0
6	is	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0
7	fun	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	1	0
8	exciting	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	1	0	0
		X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	Y(c=4)	X_K	Y(c=1)	Y(c=2)	Y(c=3)	X_K	Y(c=1)	Y(c=2)

Word similarity

- If we have only single layer of neurons
- We can find similar word which have the most similar weights



Word similarity

- If we have only single layer of neurons
- We can find similar word which have the most similar weights

Type of relationship	Word Pair 1		Word Pair 2	
Common capital city	Athens	Greece	Oslo	Norway
All capital cities	Astana	Kazakhstan	Harare	Zimbabwe
Currency	Angola	kwanza	Iran	rial
City-in-state	Chicago	Illinois	Stockton	California
Man-Woman	brother	sister	grandson	granddaughter
Adjective to adverb	apparent	apparently	rapid	rapidly
Opposite	possibly	impossibly	ethical	unethical
Comparative	great	greater	tough	tougher
Superlative	easy	easiest	lucky	luckiest
Present Participle	think	thinking	read	reading
Nationality adjective	Switzerland	Swiss	Cambodia	Cambodian
Past tense	walking	walked	swimming	swam
Plural nouns	mouse	mice	dollar	dollars
Plural verbs	work	works	speak	speaks