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.)
- ightharpoonup Number of languages \sim 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

0 0 0 0 32 20 40 [space] 64 40 100 @ 66 60 140 1 1 1 1 1 3 32 21 41 ! 65 41 101 A 97 61 141 a 2 2 2 2 34 22 42 ! 66 42 102 B 98 62 142 b 3 3 3 3 3 35 23 43 # 67 43 103 C 99 63 143 c 4 4 4 36 24 44 \$ 66 68 44 101 A D 100 64 144 d 5 6 6 6 6 6 38 26 46 & 70 46 106 F 102 66 145 e 6 6 6 6 6 38 26 46 & 70 46 106 F 102 66 145 e 6 6 6 6 6 38 26 46 & 70 46 106 F 102 66 145 e 7 7 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 49 111 105 69 151 i 11 8 13 43 28 53 + 75 48 113 K 107 69 151 i 12 C 14 44 2C 54 , 76 4C 114 L 108 6C 154 13 D 15 45 20 55 - 77 4D 115 M 109 6C 155 M 14 E 16 46 2E 56 . 78 4E 116 N 110 6E 156 n 15 F 17 47 27 57 7 7 9 47 115 M 109 6C 155 n 16 10 20 48 30 60 0 80 50 120 P 112 70 116 f 157 o 16 10 10 20 48 30 60 0 80 50 120 P 112 70 116 f 157 o 17 11 21 49 31 61 1 1 81 51 121 O 113 67 157 o 18 18 12 22 50 32 66 2 82 52 122 R 114 72 116 67 157 o 18 18 12 22 50 32 66 66 6 86 55 52 122 R 114 72 17 17 17 17 17 17 17 17 17 17 17 17 17		_			. –											
1 1 1 3 33 21 41 1 65 41 101 Â 97 61 141 0 2 2 2 2 2 2 3 43 # 67 43 103 C 99 63 143 C 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 6 5 5 5 5	Dec	Hex	0ct	Char	Dec	Hex	0ct	Char	Dec	Hex	0ct	Char	Dec	Hex	0ct	Char
2 2 2 3 34 22 42 " 66 42 102 B 98 62 142 b 3 3 3 3 3 3 3 3 5 23 43 # 4 67 43 103 C 99 63 143 c 4 4 4 5 68 44 104 D 100 64 144 d 5 6 6 6 6 6 6 38 26 46 8 70 46 106 F 102 66 145 e 6 6 6 6 6 38 26 46 8 70 46 106 F 102 66 145 e 6 6 6 6 7 7 7 7 7 99 9 11 41 29 51) 73 49 111 1 105 67 147 g 8 8 10 40 28 50 (72 48 110 H 104 68 150 h 100 67 147 g 9 9 9 11 41 29 51) 73 49 111 1 105 66 150 h 100 67 147 g 107 6 103 6	0	0	0		32	20	40	[space]	64	40	100	@	96	60	140	`
3 3 3 4 4 4 4 4 4 4 4 104 D 100 64 144 d d d 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 6 68 4 4 1 504 D 100 66 146 f f 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6 6	1	1	1		33	21	41	1	65	41	101		97	61	141	a
4 4 4 5 66 64 44 104 D 100 64 144 d 68 66 6 6 6 6 6 38 26 46 8 70 46 105 E 101 66 145 e 6 6 6 6 6 6 38 26 46 8 70 46 106 F 102 66 146 f 7 7 7 7 7 7 39 27 47 ' 7 11 47 107 6 103 67 147 g 8 8 10 40 28 50 (72 48 110 H 104 68 150 h 9 9 9 11 41 29 51) 73 49 111 105 69 151 i 100 A 12 42 2A 52 * 74 4A 112 J 106 6A 152 J 118 B 13 43 28 53 + 75 48 113 K 107 66 153 k 112 C 144 44 2C 54 , 76 4C 114 L 108 6C 154 I 113 D 15 45 25 2 50 15 5 7 7 4D 115 M 109 6C 155 m 145 6 15 6 1 15 6	2		2		34	22	42		66	42	102	В	98	62	142	b
5 5 5 6 6 6 6 6 8 38 26 46 & 6 70 46 106 F 102 66 146 f 7 7 7 7 7 39 27 47 ' 71 47 107 G 103 67 147 9 8 8 8 10 40 28 50 (72 48 110 H 104 68 150 h 100 A 12 42 2A 52 * 74 47 17 44 112 H 105 69 151 i 10 41 29 51) 73 49 111 I 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 20 51 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	3	3	3		35	23	43	#	67	43	103	С	99	63	143	C
5 5 5 6 6 6 6 6 8 38 26 46 & 6 70 46 106 F 102 66 146 f 7 7 7 7 7 39 27 47 ' 71 47 107 G 103 67 147 9 8 8 8 10 40 28 50 (72 48 110 H 104 68 150 h 100 A 12 42 2A 52 * 74 47 17 44 112 H 105 69 151 i 10 41 29 51) 73 49 111 I 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 29 51) 73 49 111 I 1 105 69 151 i 11 41 20 51 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	4	4	4		36	24	44	\$	68	44	104	D	100	64	144	d
7 7 7 8 8 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 101 64 12 42 2A 52 * 74 4A 112 J 106 6A 152 j 11	5	5	5		37	25	45	%	69	45	105	E	101	65	145	e
8 8 8 10 40 28 50 (72 48 110 H 1 104 68 150 N 1 10	6	6	6		38	26	46	&	70	46	106	F	102	66	146	f
8 8 8 10 40 28 50 (72 48 110 H 104 68 150 ĥ 100 A 12 41 29 51) 73 49 111 105 69 151 i i 100 A 12 42 2A 52 * 74 4A 112 J 106 6A 152 j 110 B 13 43 28 53 + 75 48 113 K 107 66 S 153 k 12 C 14 44 2C 54 , 76 4C 114 L 108 6C 154 113 D 15 45 2 j 114 L 108 6C 154 114 L 108 6C 154 115 F 17 47 2F 57 / 79 4F 117 0 111 6F 157 0 115 F 17 47 2F 57 / 79 4F 117 0 111 6F 157 0 116 10 20 48 30 60 0 80 50 120 P 112 70 160 p 112 70 160 p 18 18 18 18 18 18 18 18 18 18 18 18 18	7	7	7		39	27	47		71	47	107	G	103	67	147	q
100	8	8	10		40	28	50	(72	48	110	Н	104	68	150	
$\begin{array}{cccccccccccccccccccccccccccccccccccc$	9	9	11		41	29	51)	73	49	111	1	105	69	151	i
12 C 14 4 2C 54 , 76 4C 114 L 108 6C 154 1 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 O 113 71 161 q 18 12 22 50 32 62 2 82 52 122 R 114 72 162 r 19 13 23 551 33 63 3 83 53 123 5 115 73 163 s 20 14 24 52 34 64 4 84 54 124 T 116 74 164 t 21 15 25 53 36 66 6 7 88 55 125 U 117 75 165 u 22 1 15 25 53 36 66 6 7 88 55 125 U 117 75 165 u 23 17 27 18 30 56 38 70 8 8 88 88 70 8 70 8 70 8 70 8 70	10	Α	12		42	2A	52	*	74	4A	112	1	106	6A	152	i
133 D 15	11	В	13		43	2B	53	+	75	4B	113	K	107	6B	153	k
144 E 16	12	C	14		44	2C	54	,	76	4C	114	L	108	6C	154	1
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	13		15		45	2D	55	-	77	4D	115	M	109	6D		m
16 10 20 48 30 60 0 80 50 120 P 112 70 160 p 17 11 21 24 31 61 1 81 51 121 0 113 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 33 123 51 157 31 163 s 20 14 24 52 34 64 4 84 84 54 124 T 116 74 164 t 21 15 25 53 35 65 5 85 55 125 U 118 76 166 v 22 16 26 54 36 66 6 86 56 126 V 118 76 166 v 23 17 27 55 37 67 7 87 57 27 W 119 77 167 w <t< td=""><td>14</td><td>E</td><td>16</td><td></td><td>46</td><td>2E</td><td>56</td><td></td><td>78</td><td>4E</td><td>116</td><td>N</td><td>110</td><td>6E</td><td>156</td><td>n</td></t<>	14	E	16		46	2E	56		78	4E	116	N	110	6E	156	n
$ \begin{array}{cccccccccccccccccccccccccccccccccccc$	15	F	17		47	2F	57	/	79	4F	117	0	111	6F	157	0
188 12 22 50 32 62 2 82 52 122 R 114 72 162 r 190 139 23 51 51 33 63 3 83 59 38 73 . 91 58 130 75 192 193 164 32 72 72 78 16 35 74 72 72 73 16 35 74 72 74 75 74 75 74 75 74 75 74 75 74 75 75 74 75 7	16	10	20		48	30	60	0	80	50	120	P	112	70	160	p
199 13 23 51 33 63 3 83 53 123 5 115 73 163 5 20 14 164 t 21 15 25 5 52 34 64 4 8 84 54 124 T 116 74 164 t 22 1 15 25 5 33 55 65 5 85 55 125 U 117 75 165 U 23 17 27 55 37 67 7 87 57 127 W 119 77 167 w 24 18 30 56 38 70 8 8 88 58 130 X 120 78 170 X 25 19 31 57 39 71 9 89 59 131 Y 121 79 171 Y 25 61 14 32 56 14 32 58 34 72 : 90 5A 132 Z 122 7A 172 Z 27 18 33 59 38 73 ; 91 58 133 [123 78 173 (17	11	21		49	31	61	1	81	51	121	Q	113	71	161	q
199 13 23 51 33 63 3 83 53 123 5 115 73 163 5 20 14 164 t 21 15 25 5 52 34 64 4 8 84 54 124 T 116 74 164 t 22 1 15 25 5 33 55 65 5 85 55 125 U 117 75 165 U 23 17 27 55 37 67 7 87 57 127 W 119 77 167 w 24 18 30 56 38 70 8 8 88 58 130 X 120 78 170 X 25 19 31 57 39 71 9 89 59 131 Y 121 79 171 Y 25 61 14 32 56 14 32 58 34 72 : 90 5A 132 Z 122 7A 172 Z 27 18 33 59 38 73 ; 91 58 133 [123 78 173 (18	12	22		50	32	62	2	82	52	122	R	114	72	162	r
21 15 25 53 35 65 5 85 55 125 U 117 75 165 U 222 16 26 54 36 66 6 6 86 55 6 126 V 118 76 166 V 233 17 27 55 37 67 7 87 57 127 W 119 77 167 w 24 18 30 56 38 70 8 88 85 81 30 X 120 78 170 x 25 19 31 57 39 71 9 89 59 131 Y 121 79 171 y 26 14 18 32 58 130 X 2 122 74 171 y 27 18 18 33 59 38 73 ; 91 58 138 Z 12 122 74 172 Z 174 174 175 18 175	19	13	23			33	63	3	83	53	123	S	115	73	163	S
222 16 26 54 36 66 6 886 56 126 V 118 76 166 V 23 172 7 55 37 67 7 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 14 32 58 3A 72 : 90 5A 132 Z 122 7A 172 Z 27 18 33 59 38 73 : 91 58 133 [123 78 173 { 22 88 1C 34 60 3C 74 < 92 5C 134 \ 124 7C 174 [175 4] 18 30 18 2 36 61 30 75 = 93 5D 135] 125 7D 175 }	20		24		52	34	64		84	54	124	Т	116	74	164	t
23 17 27 55 37 67 7 87 57 127 W 119 77 167 w 24 18 30 56 38 70 8 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 78 173 (28 1C 34 66 3C 74 < 92 5C 134 124 7C 174 29 1D 35 61 30 75 = 93 5D 135 125 7D 175 } 30 1E 36 62 3E 76 > 94 5E 136 6 126 7E 176 ~	21	15	25		53	35	65	5	85	55	125	U	117	75	165	u
24 18 30 56 38 70 8 88 58 130 X 120 78 170 X 25 199 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 38 73 ; 91 58 133 [123 78 173 { 28 16 34 60 3C 74 < 92 5C 134 \ 124 7C 174 29 1D 35 61 36 62 3E 76 > 93 5D 135 125 7D 175 }	22	16	26		54	36	66	6	86	56	126	V	118	76	166	v
25 19 31 57 39 71 9 89 59 131 Y 121 79 171 Y 26 16 1A 32 58 3A 72 : 90 5A 132 Z 122 7A 172 Z 27 1B 33 59 3B 73 ; 91 5B 133 [123 78 173 (23											W	119	77		W
26 1A 32 58 3A 72 : 90 5A 132 Z 122 7A 172 Z 77 1B 33 59 3B 73 ; 91 5B 133 [123 7B 173 { 22 12 2 7A 172 2 2 13 2 14 1 124 7C 174 129 1D 35 61 3D 75 = 93 5D 135] 125 7D 175 } 93 5D 135] 125 7D 175 } 30 1E 36 62 3E 76 > 94 5E 136 ^ 126 7E 176 ~	24								88	58			120	78		X
26 1A 32 58 3A 72 : 90 5A 132 Z 122 7A 172 Z 77 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 ~	25	19			57	39		9	89	59		Υ	121	79		у
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 ~	26				58			:	90	5A		Z	122	7A		
29 1D 35 61 3D 75 = 93 5D 135] 125 7D 175 } 30 1E 36 62 3E 76 > 94 5E 136 ^ 126 7E 176 ~	27							;]				{
30 1E 36 62 3E 76 > 94 5E 136 ^ 126 7E 176 ~	28							<				1				
30 12 30 02 32 70 > 34 32 130 120 72 170 -	29							-			135]	125			}
31 1F 37 63 3F 77 ? 95 5F 137 _ 127 7F 177	30											^				~
	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 → 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+7FFFFFF	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)

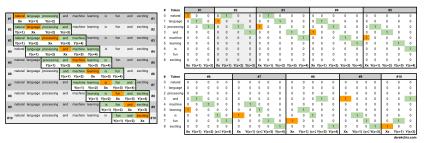
alkalom/NOUN[NEG ATTRIB]/ADJ<PLUR>

```
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>
```

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

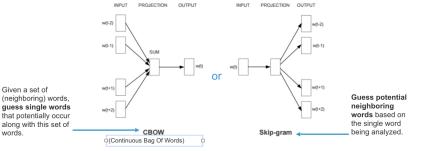


Words prediction

Given a set of

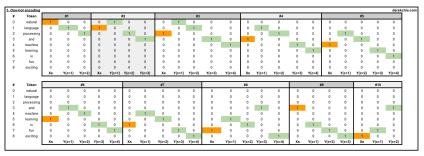
words.

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



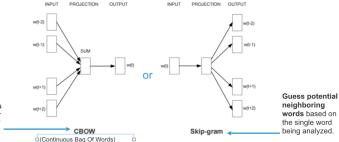
Encoding

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



Word similarity

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



Given a set of (neighboring) words, guess single words that potentially occur along with this set of words.

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		