Comunidad MyFuture-AI

Introducción al Procesamiento de Lenguaje Natural (NLP)

By Laura López

Agenda

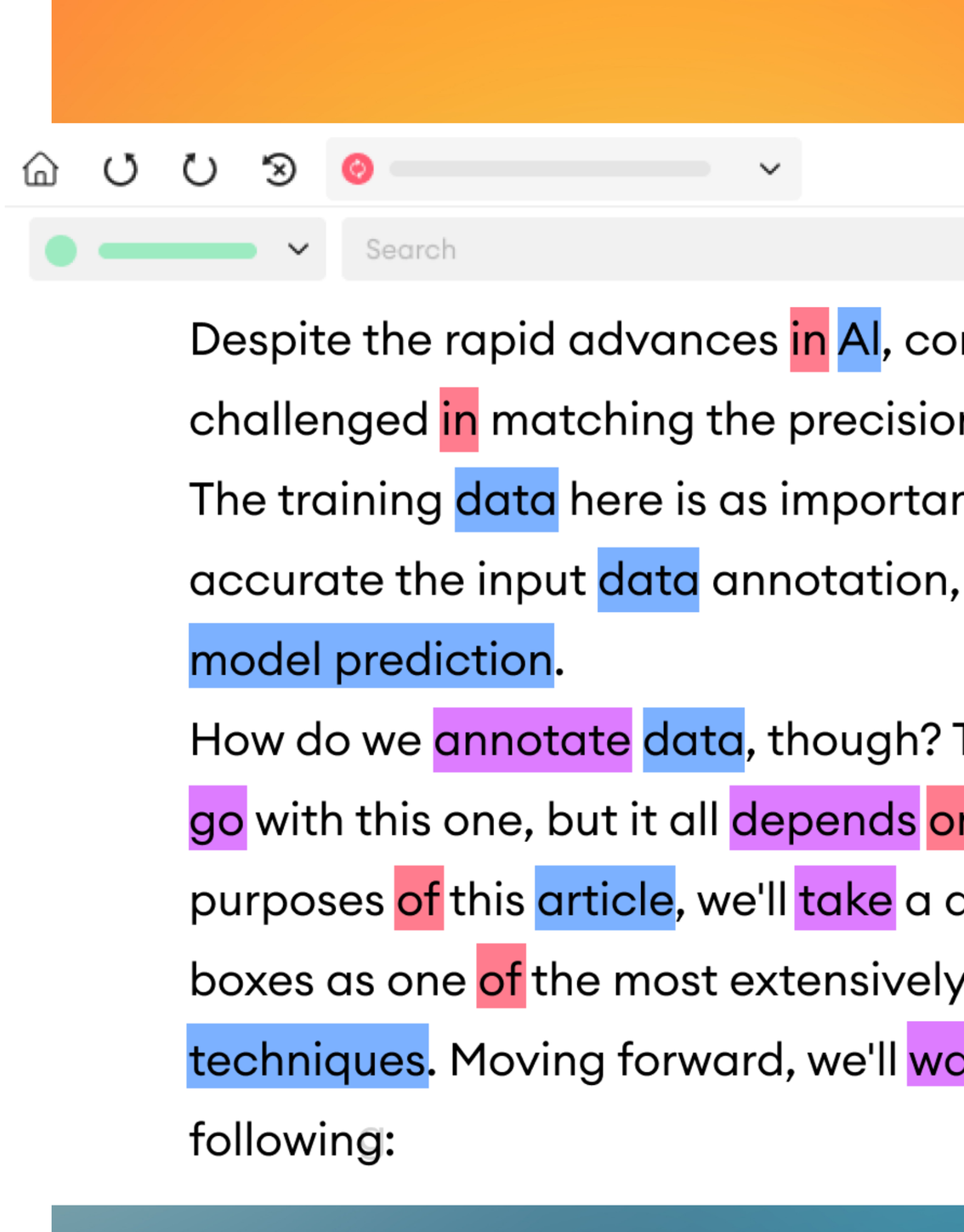
Esto es lo que abarcaremos:

- Embeddings
- Preprocesamiento de texto
- Análisis NLP
- Práctica

Procesamiento de Lenguaje Natural (NLP)

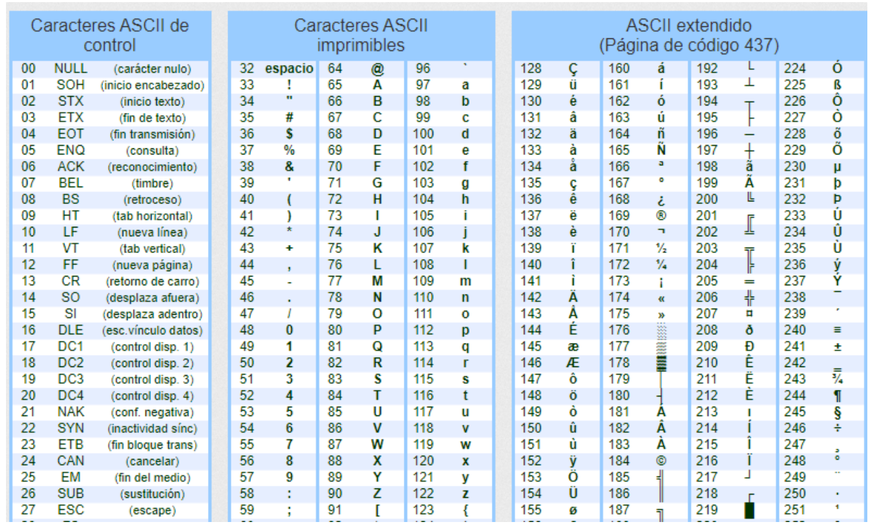
Procesamiento de Lenguaje Natural (NLP)

“Es una tecnología de machine learning que brinda a las computadoras la capacidad de interpretar, manipular y comprender el lenguaje humano.” - Amazon



¿Cómo representas un texto con números?

Asignando a cada carácter un valor



Caracteres ASCII de control

00	NULL	(carácter nulo)
01	SOH	(inicio encabezado)
02	STX	(inicio texto)
03	ETX	(fin de texto)
04	EOT	(fin transmisión)
05	ENQ	(consulta)
06	ACK	(reconocimiento)
07	BEL	(timbre)
08	BS	(retroceso)
09	HT	(tab horizontal)
10	LF	(nueva línea)
11	VT	(tab vertical)
12	FF	(nueva página)
13	CR	(retorno de carro)
14	SO	(desplaza afuera)
15	SI	(desplaza adentro)
16	DLE	(esc.vínculo datos)
17	DC1	(control disp. 1)
18	DC2	(control disp. 2)
19	DC3	(control disp. 3)
20	DC4	(control disp. 4)
21	NAK	(conf. negativa)
22	SYN	(inactividad sínc)
23	ETB	(fin bloque trans)
24	CAN	(cancelar)
25	EM	(fin del medio)
26	SUB	(sustitución)
27	ESC	(escape)

Caracteres ASCII imprimibles

32	espacio	64	@	96	`
33	!	65	A	97	a
34	"	66	B	98	b
35	#	67	C	99	c
36	$	68	D	100	d
37	%	69	E	101	e
38	&	70	F	102	f
39	'	71	G	103	g
40	(	72	H	104	h
41	)	73	I	105	i
42	*	74	J	106	j
43	+	75	K	107	k
44	,	76	L	108	l
45	-	77	M	109	m
46	.	78	N	110	n
47	/	79	O	111	o
48	0	80	P	112	p
49	1	81	Q	113	q
50	2	82	R	114	r
51	3	83	S	115	s
52	4	84	T	116	t
53	5	85	U	117	u
54	6	86	V	118	v
55	7	87	W	119	w
56	8	88	X	120	x
57	9	89	Y	121	y
58	:	90	Z	122	z
59	;	91	[	123	{

ASCII extendido (Página de código 437)

128	Ç	160	á	192	└	224	Ó
129	ü	161	í	193	┴	225	ß
130	é	162	ó	194	┬	226	Ô
131	â	163	ú	195	├	227	Ò
132	ä	164	ñ	196	─	228	õ
133	à	165	Ñ	197	┼	229	Õ
134	å	166	ª	198	ã	230	µ
135	ç	167	º	199	Ã	231	þ
136	ê	168	¿	200	╚	232	Þ
137	ë	169	®	201	╔	233	Ú
138	è	170	¬	202	╩	234	Û
139	ï	171	½	203	╦	235	Ù
140	î	172	¼	204	╠	236	ý
141	ì	173	¡	205	═	237	Ý
142	Ä	174	«	206	╬	238	¯
143	Å	175	»	207	¤	239	´
144	É	176	▒	208	ð	240	≡
145	æ	177	▒	209	Ð	241	±
146	Æ	178	▓	210	Ê	242	‗
147	ô	179	│	211	Ë	243	¾
148	ö	180	┤	212	È	244	¶
149	ò	181	Á	213	ı	245	§
150	û	182	Â	214	Í	246	÷
151	ù	183	À	215	Î	247	¸
152	ÿ	184	©	216	Ï	248	°
153	Ö	185	╣	217	┘	249	¨
154	Ü	186	║	218	┌	250	·
155	ø	187	╗	219	▒	251	¹

Tomada de: https://elcodigoascii.com.ar

Tiene sentido, pero ¿qué pasa si tengo un documento?

¿Cómo representas un texto con números?

Asignando a cada palabra un valor o posición de un vector

My Future IA

11 10 23

Pero ¿**My** es mayor que **Future**?

¿Cómo representas un texto con números?

Vectores para representar las palabras y sentido semántico

One-Hot encoding

[My, Future, IA]
My: [1,0,0]
Future:[0,1,0]
IA:[0,0,1]

Todas las palabras tendrían la misma distancia, ¿qué pasa si tengo palabras con un contexto en común?

¿Cómo representas un texto con números?

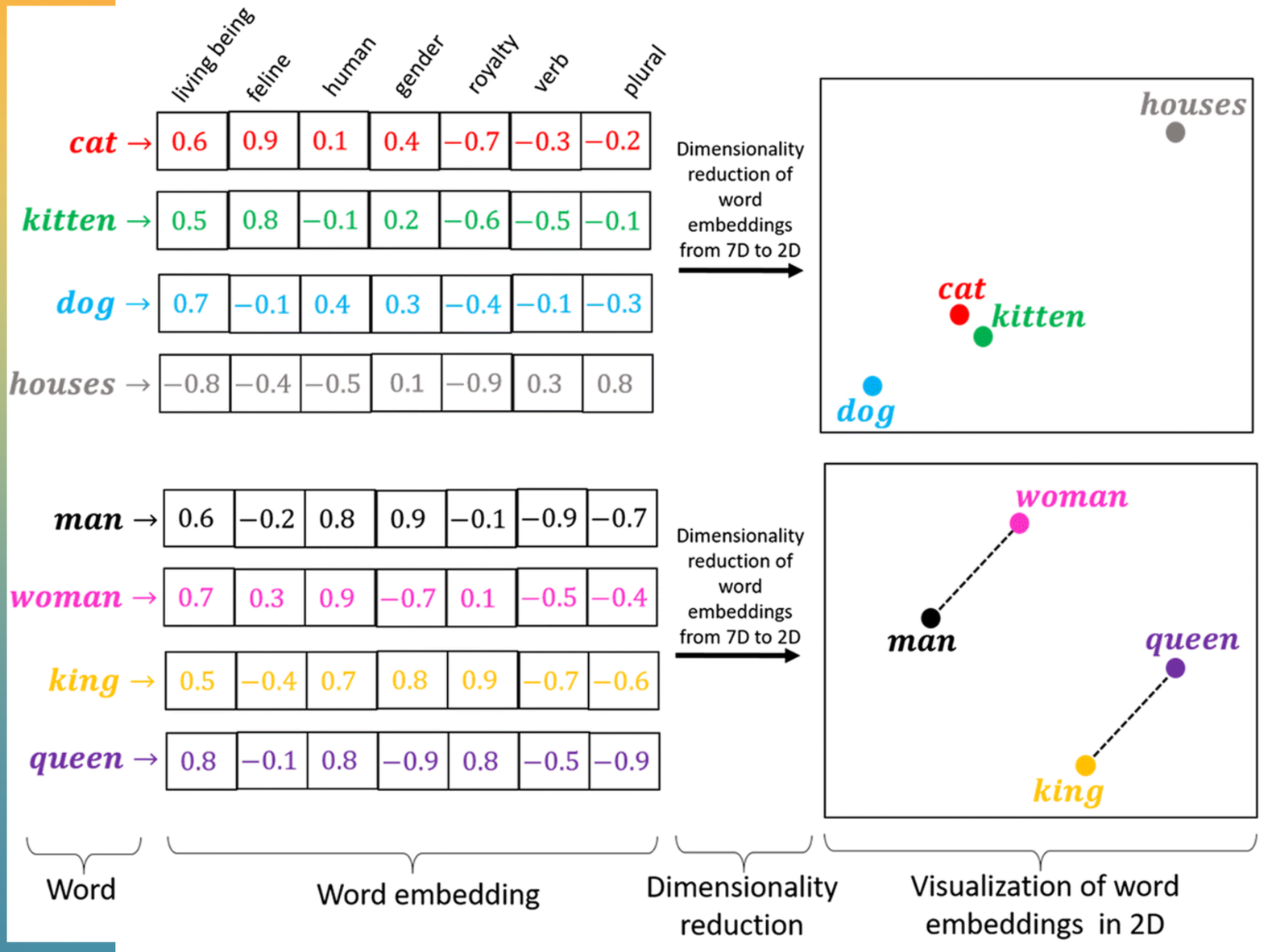
Asignando a cada carácter un valor

Asignando a cada palabra un valor o posición de un vector

Vectores para representar las palabras y sentido semántico

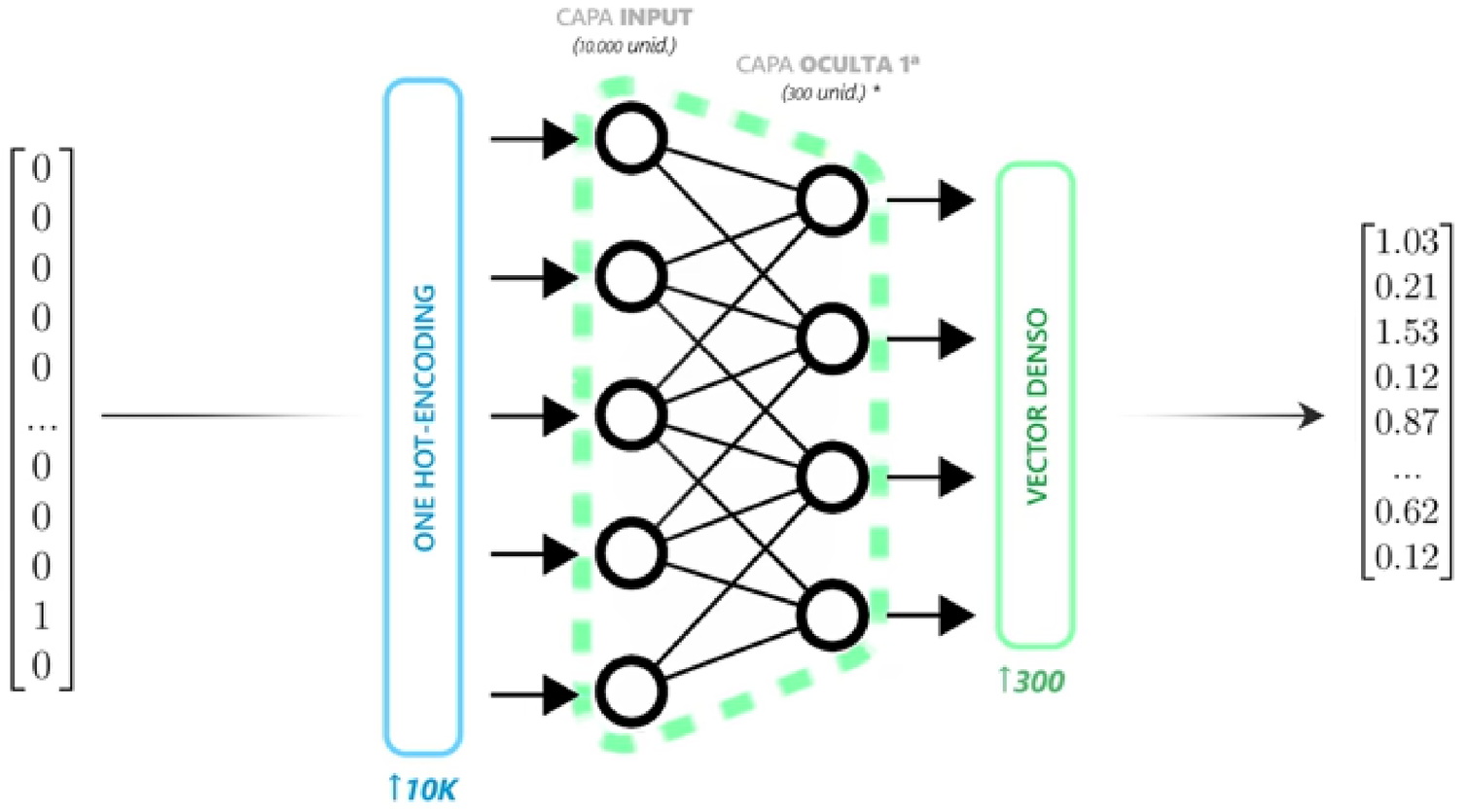
Embeddings

Reducir la dimensionalidad



Embeddings

Reducir la dimensionalidad



Tomado de: https://www.youtube.com/watch?v=RkYuH_K7Fx4

Embeddings

Reducir la dimensionalidad

Word2Vec

Google - 2013

Embedding projector - visualization of high-dimensional data

Visualize high dimensional data.

tensorflow.org /

Preprocesamiento de texto

Preprocesamiento

“Hey Amazon - my package never arrived https://www.amazon.com/gp/css/order-history?ref_=nav_orders_first PLEASE FIX ASAP! @amazonhelp”

Normalizar

“hey amazon my package never arrived please fix asap”

Eliminar stopwords

“amazon package never arrived fix asap”

Tokenizar

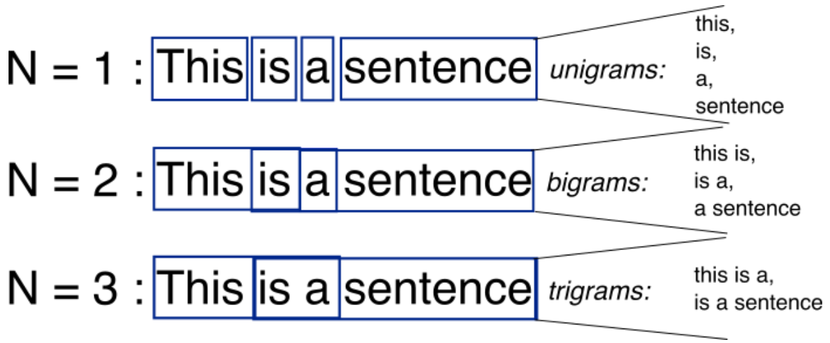
["amazon", "package", "never", "arrived", "fix", "asap”]

Preprocesamiento

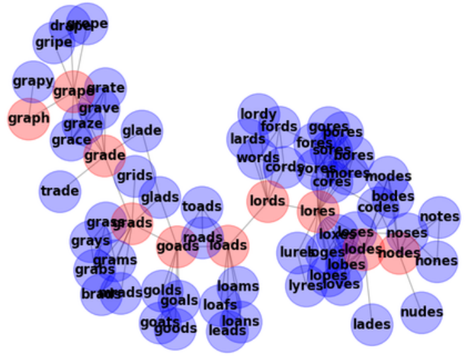
Normalizar	Unificar el uso de acentos, mayúsculas, abreviaciones
Stopwords	Eliminar palabras que no aportan al contexto de una oración
Símbolos y caracteres especiales	Eliminar signos de puntuación, símbolos, caracteres especiales
Tokenizar	Dividir un texto en entidades más pequñas llamadas tokens
Stemming	Reducir una palabra a su raíz
Lemmatization	A una forma flexionada de una palabra pasarla a su lema

Aplicaciones

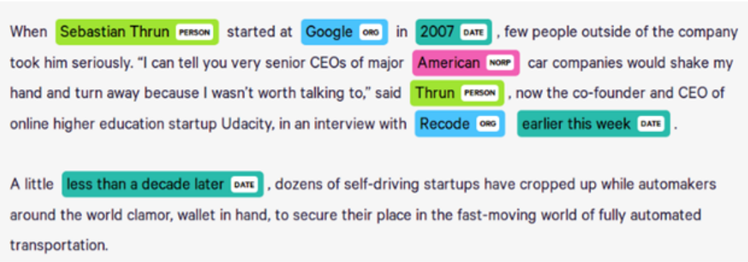
¿Qué podemos hacer con NLP?



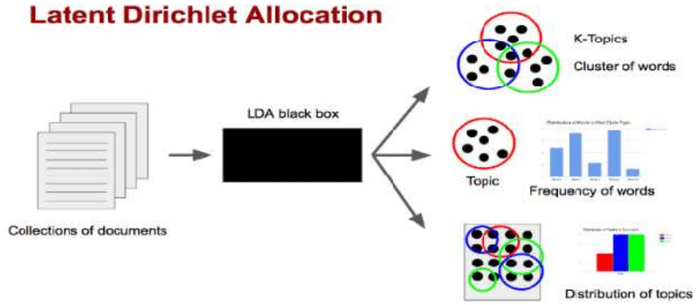
N-Grams



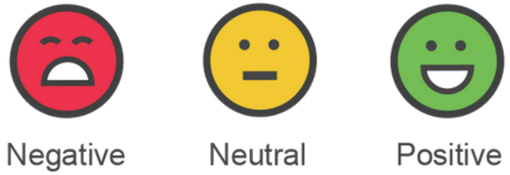
¿Qué podemos hacer con NLP?



Named Entity Recognition (NER) y Part Of Speech Tagging (POS)



Latent Dirichlet Allocation (LDA)



Análisis de sentimientos y emociones

Vamos al código con algo de PySpark

¿Tienes alguna pregunta?

github.com/lauralpezb/NLP_MyFutureAI

@lauralpezb