Cosas *que aprendí mirando* 200.000 *anglicismos*

Elena Álvarez Mellado
@lirondos

Prolegómenos


TRABALENGUA
ANGLICISMOS
MACHINE LEARNING
YO
VOSOTROS
imgflip.com

3 motivos por los que el anglicismo merece ser estudiado (y ninguno de ellos es el purismo lingüístico)

3 motivos por los que el anglicismo merece ser estudiado (y ninguno de ellos es el purismo lingüístico)

- El préstamo léxico es una manifestación de cómo las lenguas entran en contacto y cambian con el tiempo

3 motivos por los que el anglicismo merece ser estudiado (y ninguno de ellos es el purismo lingüístico)

- El préstamo léxico es una manifestación de cómo las lenguas entran en contacto y cambian con el tiempo

- El préstamo léxico es una fuente de incorporación de nuevas palabras

El préstamo léxico como fuente de palabras nuevas

- Nuevas realidades, nuevas palabras:

online, software, podcast, streaming…

El préstamo léxico como fuente de palabras nuevas

- Nuevas realidades, nuevas palabras:

  online, software, podcast, streaming…

- Viejas realidades, nuevas palabras:

  nude (color carne), low cost (barato), azeyte (olio)

La sociolingüística del anglicismo en la calle

3 motivos por los que el anglicismo merece ser estudiado (y ninguno de ellos es el purismo lingüístico)

- El préstamo léxico es una manifestación de cómo las lenguas entran en contacto y cambian con el tiempo

- El préstamo es una fuente de incorporación de nuevas palabras

- El proceso de adaptación arroja luz sobre cuáles son las expectativas lingüísticas de los hablantes de una lengua

Existe un interés social sobre la introducción de anglicismos (Fundéu, RAE)

Objetivo del proyecto

Crear un sistema capaz de detectar y monitorizar automáticamente el uso de préstamos léxicos en general y de anglicismos en particular.

Ejemplo: *Recetas de noviembre para el* ***batch cooking***

Las prendas ***bestsellers*** *se estampan con motivos florales, '****animal print****' o a retales tipo* ***patchwork***

¿Cómo lo haríais?

La ambigüedad contraataca: el diccionario no basta

- *Prime time* es un anglicismo:
  - *prime* es una forma del verbo *primar*
  - *time* es una forma del verbo *timar*

La ambigüedad contraataca: el diccionario no basta

- *Prime time* es un anglicismo:
  - *prime* es una forma del verbo *primar*
  - *time* es una forma del verbo *timar*

- *Social media* es un anglicismo:
  - Pero tanto *social* como *media* son también palabras en español
  - *Media social* no es un anglicismo

La ambigüedad contraataca: el diccionario no basta

- *Prime time* es un anglicismo:
  - *prime* es una forma del verbo *primar*
  - *time* es una forma del verbo *timar*

- *Social media* es un anglicismo:
  - Pero tanto *social* como *media* son también palabras en español
  - *Media social* no es un anglicismo

- La palabra *primer*:
  - *Recomendamos usar un 'primer' hidratante*
  - *Ha ganado el primer premio*

La ambigüedad contraataca (II): *nombres propios*

- *La compañía Apple triplica resultados este trimestre*

La ambigüedad contraataca (II): *nombres propios*

- *La compañía Apple triplica resultados este trimestre*

- *Apple triplica resultados este trimestre*

La ambigüedad contraataca (II): *nombres propios*

- *La compañía Apple triplica resultados este trimestre*

- *Apple triplica resultados este trimestre*

- *Ghosting, la tendencia que causa furor en redes*

Cuando la casuística del sistema de reglas es demasiado complejo y no consigue dar cuenta del fenómeno que queremos abordar

→ *Aprendizaje automático*

El aprendizaje automático muy muy muy resumido

- Si le damos a un ordenador una cantidad suficiente de ejemplos del fenómeno que queremos modelar, el ordenador aprenderá a reconocerlo

- Aprendizaje por imitación (¿es verdaderamente aprendizaje? ¿es inteligencia?)

La metáfora del seguro de automóvil

- Cuando compras un seguro de coche, el precio de la póliza dependerá de diversas características: edad, color del coche, experiencia al volante, partes previos:
  - Una mujer de 52 años, 27 años de antigüedad, coche nuevo blanco
  - Un hombre de 23, 1 año de antigüedad, coche viejo rojo

- Imaginad que tuviésemos miles de ejemplos así, y supiésemos además si la persona dio algún parte o no. ¿Podríamos asignar pesos a cada uno de estos rasgos y calcular la probabilidad de que un conductor dé un parte?

¿Cómo lo hacemos?

- En vez de partes del seguro y conductores, pensad en textos.
- Necesitamos una colección de textos/un corpus rico en el fenómeno que queremos modelar
  - Necesitamos un corpus de español rico en anglicismos
- En el corpus deberemos marcar a mano qué cosas son anglicismo y cuáles no (anotación)
- Al ordenador le daremos ese corpus anotado con anglicismos
  - El sistema aprenderá cuáles son las características de los anglicismos de nuestro corpus. Cuando le demos una palabra nueva podrá intentar adivinar si es un anglicismo o no
- Es básicamente reconocimiento de patrones. El ordenador no aprende realmente nada ni tiene inteligencia (tal y como la entendemos los humanos)

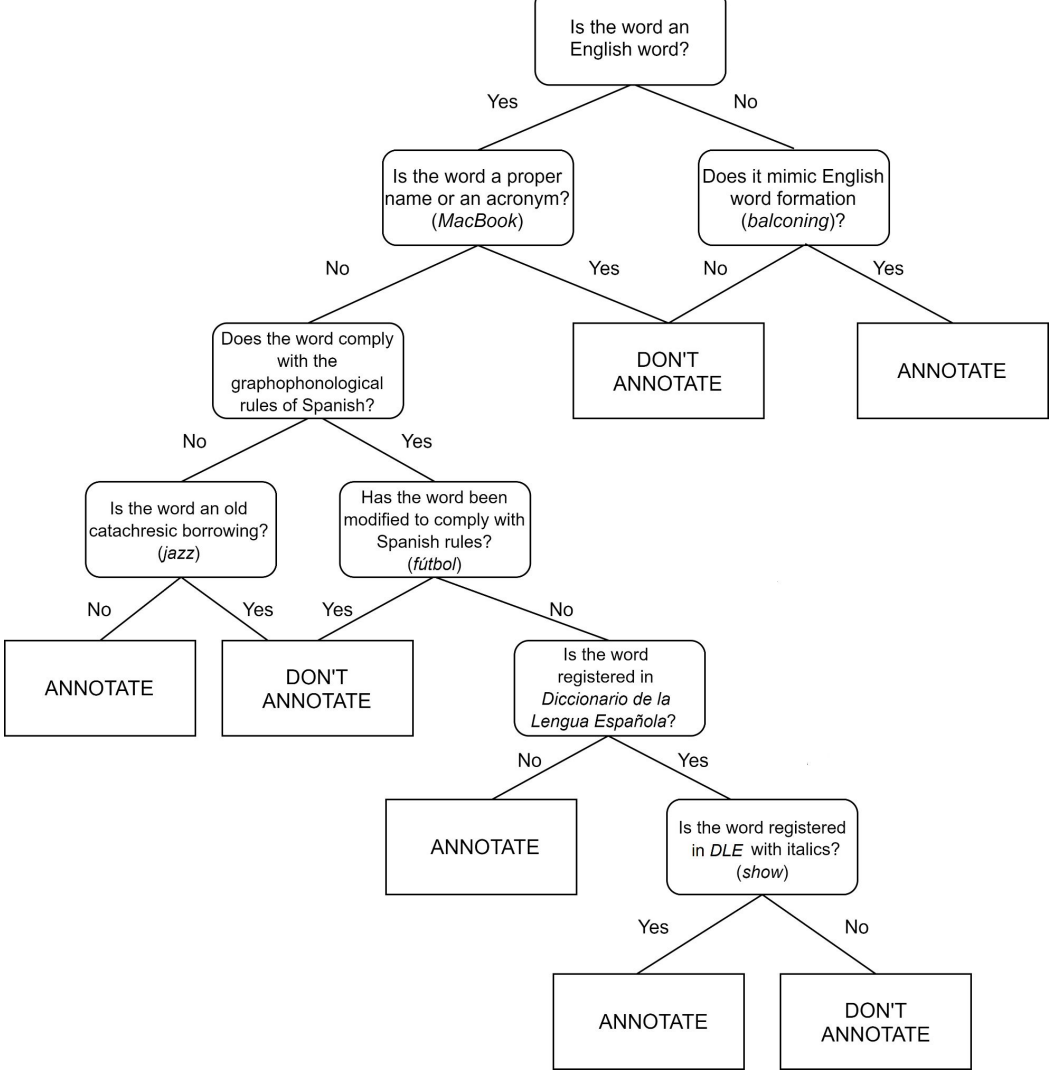
La metáfora del examen de matemáticas

- Le damos al ordenador un montón de problemas resueltos (corpus de entrenamiento)
- El ordenador tiene acceso tanto al problema como a la solución
- El ordenador “aprenderá” a partir de esos problemas resueltos
- Cuando el modelo ya está entrenado (= el ordenador ha terminado de “estudiar”) le damos problemas nuevos (el examen). Esta vez SIN darle la solución (datos de evaluación)
- Evaluamos qué tal ha hecho el examen nuestro modelo

1. *Construir un corpus anotado con anglicismos*

- Recopilar 21,570 titulares de prensa en español de elDiario.es

- Marcar a mano cuáles de esas palabras son anglicismos y cuáles no. Esto es más fácil decirlo que hacerlo → guía de anotación (ponga un lingüista en su vida)


Is the word an English word?
Yes
No
Is the word a proper name or an acronym? (*MacBook*)
Does it mimic English word formation (*balconing*)?
No
Yes
No
Yes
Does the word comply with the graphophonological rules of Spanish?
DON'T ANNOTATE
ANNOTATE
No
Yes
Is the word an old catachresic borrowing? (*jazz*)
Has the word been modified to comply with Spanish rules? (*fútbol*)
No
Yes
Yes
No
ANNOTATE
DON'T ANNOTATE
Is the word registered in *Diccionario de la Lengua Española*?
No
Yes
ANNOTATE
Is the word registered in *DLE* with italics? (*show*)
Yes
No
ANNOTATE
DON'T ANNOTATE

En O

este O

mes O

especialmente O

puede O

ser O

de O

utilidad O

apuntarnos O

al O

batch B-ENG

cooking I-ENG

Benching B-ENG

, O

estar O

en O

el O

banquillo O

de O

tu O

crush B-ENG

mientras O

otro O

juega O

de O

titular O

2. Transformar los datos en algo que el ordenador entienda

La	fiebre	de	los	podcasts	llega	a	España
O	O	O	O	ENG	O	O	O

¿Os acordáis del ejemplo del seguro de coche? Pensad en palabras:

- Cada palabra estará representada por un conjunto de rasgos: qué caracteres que conforman la palabra, si va en mayúscula o no, si lleva signos de puntuación o no, etc.
- La selección de rasgos es una tarea fundamental y muy lingüística (tanto la observación experimental como la intuición lingüística pueden ayudar)

De palabras a rasgos

Cada palabra estará representada por los siguientes rasgos:

- la palabra en sí: podcast
- terminación: ‘ast’
- trigramas de caracteres: ‘pod’, ‘odc’, ‘dca’, ‘cas’, ‘ast’
- mayúscula/minúscula: minúscula
- palabras adyacentes: antecedida por comilla (modelo secuencial)

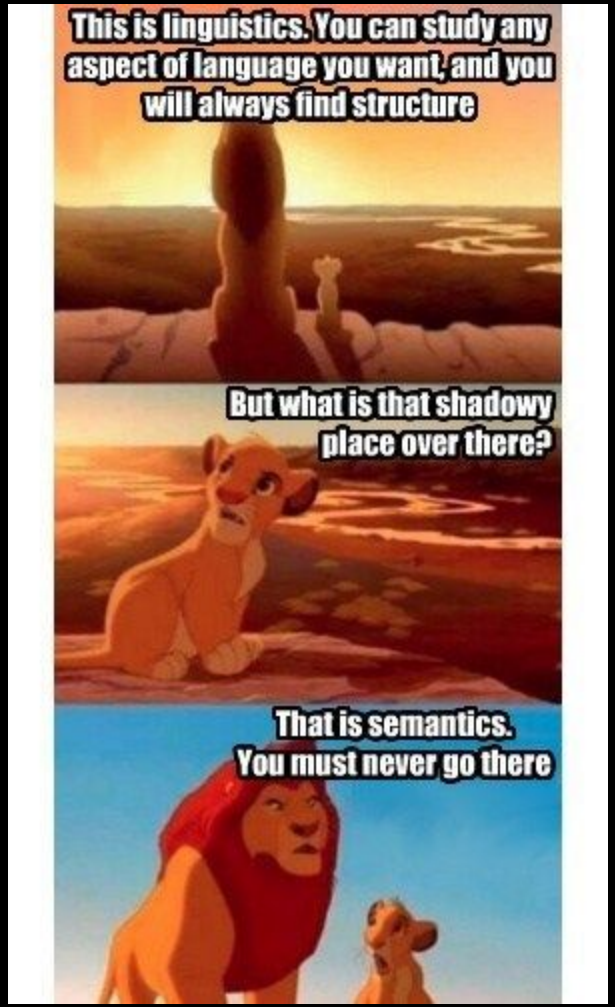
Durante el entrenamiento, el modelo “aprenderá” cuánto de importantes son estos rasgos y cuáles son los indicios de que algo sea o no un anglicismo.

Más allá de la forma

- Nuestros rasgos giran todos en torno a la forma de la palabra

- La forma puede representar aspectos morfológicos, pero no relativos al significado de las palabras

- ¿Es posible representar la semántica de una forma numérica que un ordenador pueda procesar?


This is linguistics. You can study any aspect of language you want, and you will always find structure
But what is that shadowy place over there?
That is semantics. You must never go there

Palabras como vectores (aka word embeddings)

El agua hierve a 100ºC

He pedido un vaso de limonada

Luisa no bebe limonada

Prefiero beber un vaso de agua

Siempre come helado de
chocolate de postre

Palabras como vectores (aka word embeddings)

El agua hierve a 100ºC

He pedido un vaso de limonada

Luisa no bebe limonada

Prefiero beber un vaso de agua

Siempre come helado de chocolate de postre

	hervir	beber	vaso	chocolate
agua				
limonada				
helado				

Palabras como vectores (aka word embeddings)

El agua hierve a 100ºC

He pedido un vaso de limonada

Luisa no bebe limonada

Prefiero beber un vaso de agua

Siempre come helado de chocolate de postre

	hervir	beber	vaso	chocolate
agua	1	1	1	0
limonada	0	1	1	0
helado	0	0	0	1

agua = (1, 1, 1, 0)

limonada = (0, 1, 1, 0)

helado = (0, 0, 0, 1)

Palabras como vectores (aka word embeddings)

El agua hierve a 100ºC

He pedido un vaso de limonada

Luisa no bebe limonada

Prefiero beber un vaso de agua

Siempre come helado de chocolate de postre

	hervir	beber	vaso	chocolate
agua	1	1	1	0
limonada	0	1	1	0
helado	0	0	0	1

agua = (1, 1, 1, 0)

limonada = (0, 1, 1, 0)

helado = (0, 0, 0, 1)

Estos 2 son más parecidos = espacialmente más cerca

3. *Cómo entrenar a tu ~~dragón~~ modelo* 🐉

- Tenemos nuestro corpus anotado

- Hemos transformado los datos en algo que el ordenador pueda beber (rasgos, embeddings)

- Ahora, a entrenar (a estudiarse los exámenes resueltos)
  - El entrenamiento como dardos

4. Hora de evaluar

- Cuando el entrenamiento ha terminado, evaluamos

- Proporcionamos ejemplos nuevos al modelo (sin solución). El modelo analizará cada palabra y a la luz de los rasgos que observe intentará adivinar qué palabras son anglicismos

- Es decir, nuestro modelo está ahora haciendo predicciones

La evaluación

- ¿Cómo de bien le sale el examen?

- Para evaluar un modelo tenemos en cuenta 2 aspectos:
  - Cuántas palabras marcadas por el modelo como “anglicismo” eran verdaderamente anglicismos
  - Cuántos anglicismos se dejó el modelo fuera (eran anglicismos y modelo los marcó como no anglicismo)
  - La combinación de estas dos se llama valor F1

  Para el modelo de Lázaro, F1 = 87 (100 sería un modelo perfecto). No es perfecto, ¡pero no está mal!

5. *Creando Observatorio Lázaro*

- Ya tenemos un modelo capaz de detectar automáticamente anglicismos en prensa española. ¡Ya podemos crear el observatorio!
- El observatorio es un sistema que:
  - Extrae los artículos publicados cada día en 8 medios españoles de prensa escrita: El País, La Vanguardia, ABC, El Mundo, elDiario, 20minutos, El Confidencial, EFE
  - Los artículos se mandan al modelo
  - El modelo extrae los anglicismos que aparecen en los artículos
  - Cada anglicismo se almacena en una base de datos junto a la información del avistamiento (fecha, URL, medio, contexto)

6. *Visualizando, que es gerundio*

Con la información almacenada podemos:

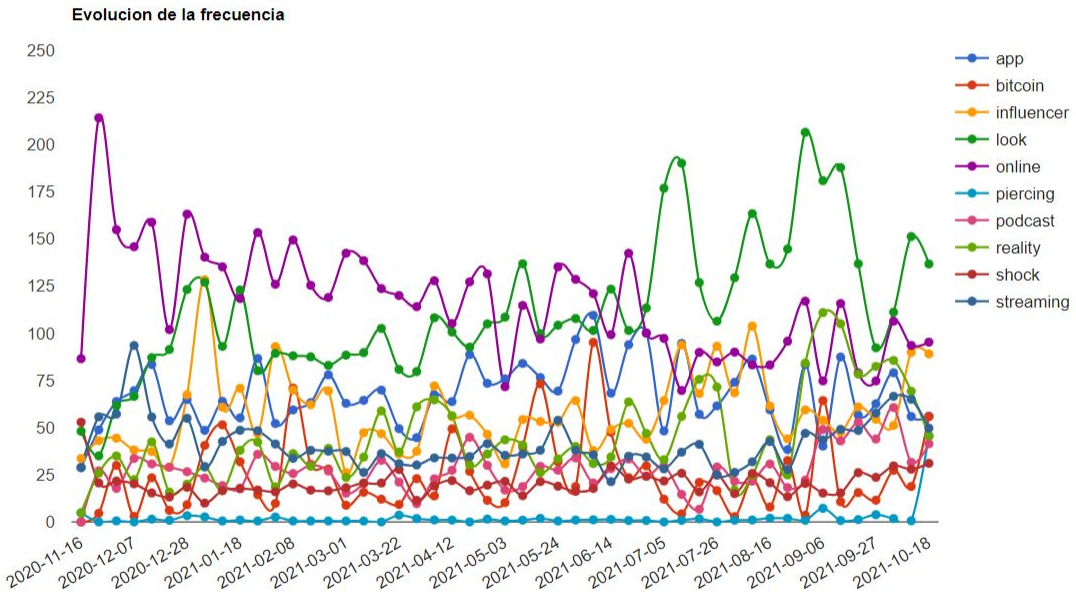
- ver cómo cambia la frecuencia de uso a lo largo del tiempo
- hacer resumen con los anglicismos nuevos del día

observatoriolazaro.es

observatoriolazaro.es

Anglicismos más frecuentes

Evolución de los 10 anglicismos más frecuentes (frecuencia por cada millón de palabras)



podcast

Anglicism: yes

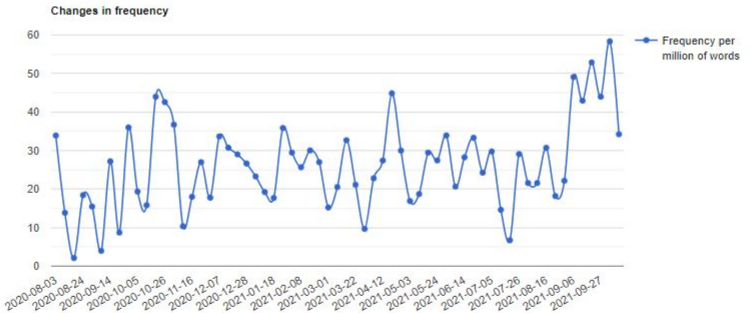
Forms: podcast podcasts

Average frequency*: 26.313

Frequency during the last 30 days*: 47.818

Newspaper sections: Portada Televisión Cultura Gente Economía

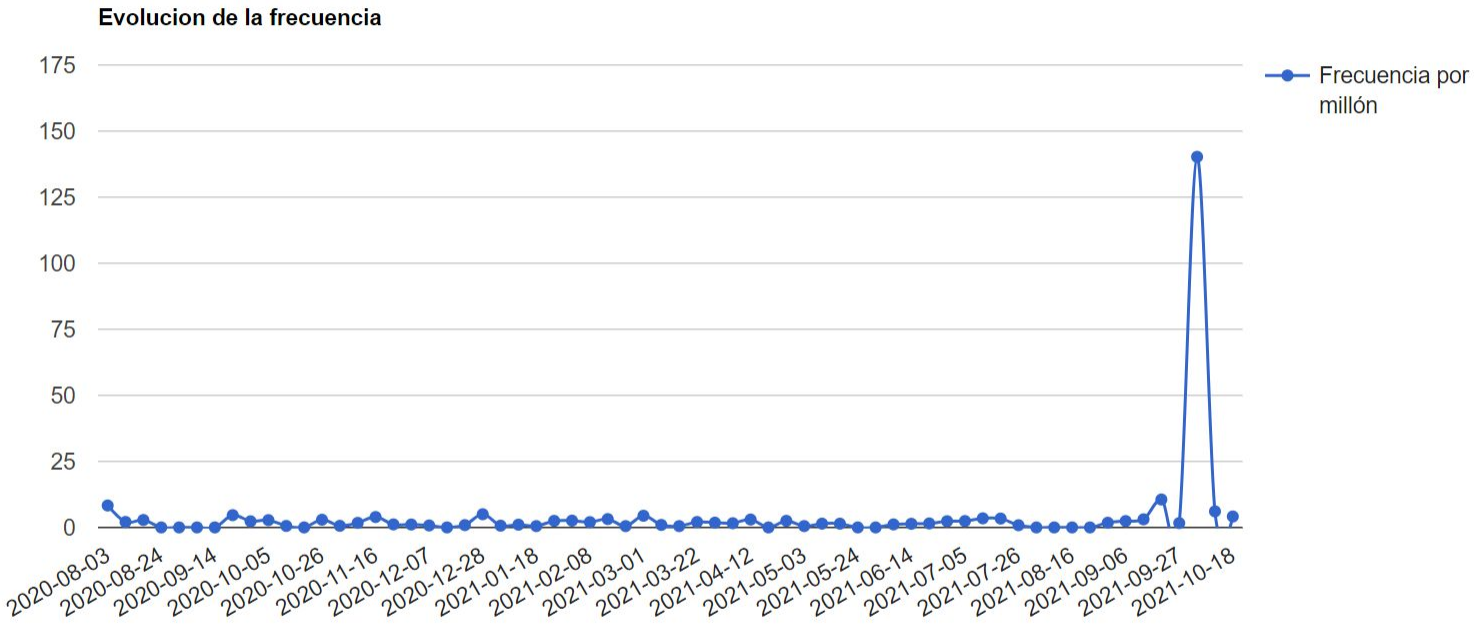
* Frequency per million of words (measured from August 2020 onwards).



Show 10 entries

Search:

Borrowing	Context	Newspaper	Date
podcast	*Grito Sordo y que guioniza el contenido que propone Ignatius Farray, como el 'podcast' 'Payasos y fuego'. En*	elpais	18-10-2021


Evolucion de la frecuencia
175
150
125
100
75
50
25
0
2020-08-03
2020-08-24
2020-09-14
2020-10-05
2020-10-26
2020-11-16
2020-12-07
2020-12-28
2021-01-18
2021-02-08
2021-03-01
2021-03-22
2021-04-12
2021-05-03
2021-05-24
2021-06-14
2021-07-05
2021-07-26
2021-08-16
2021-09-06
2021-09-27
2021-10-18
Frecuencia por millón

offshore

Anglicismo: sí

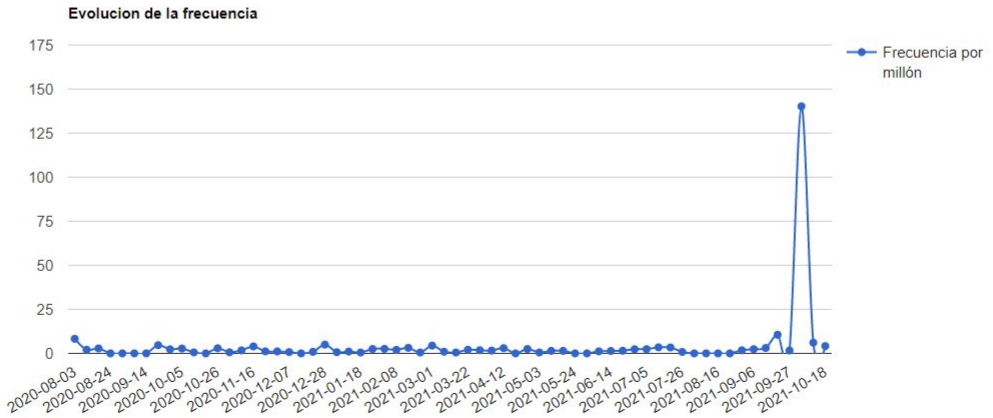
Formas: offshore offshores

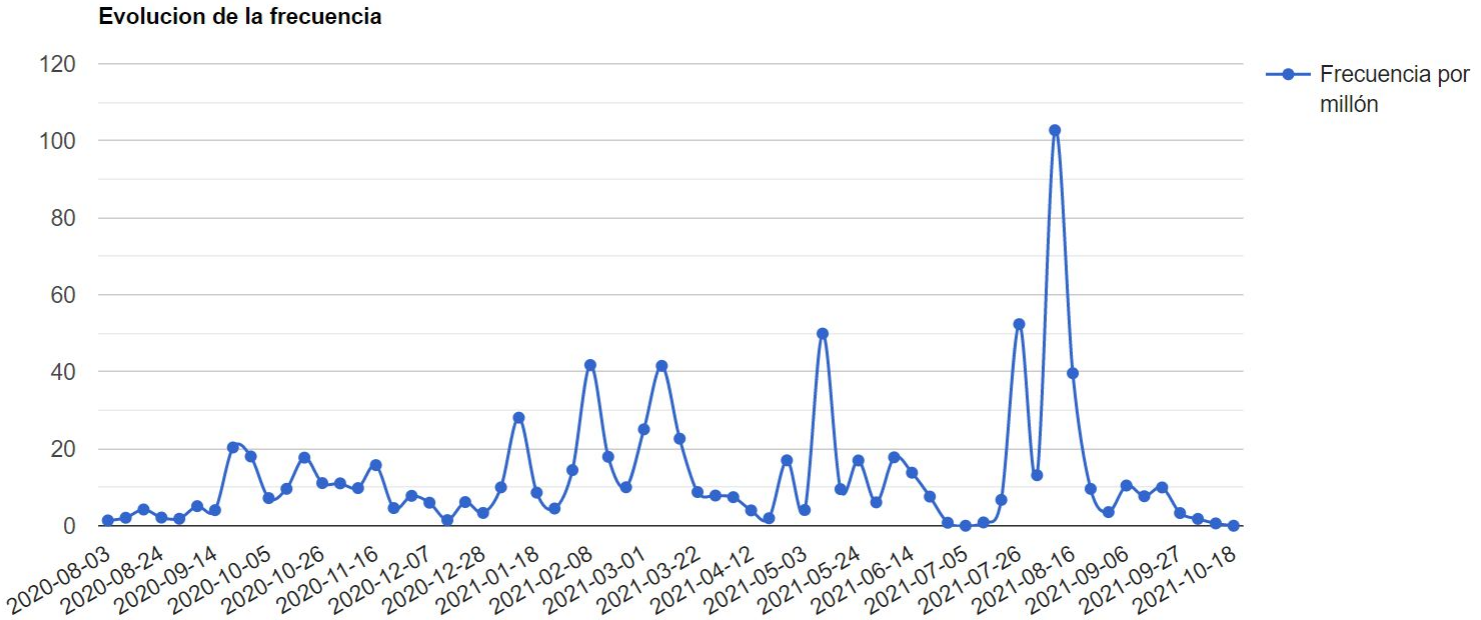
Frecuencia media de aparición*: 4.143

Frecuencia de aparición en el último mes*: 35.18

Secciones habituales: Economía Portada Internacional Política España

* Frecuencia por cada millón de palabras medida desde agosto de 2020.




Evolucion de la frecuencia
120
100
80
60
40
20
0
2020-08-03
2020-08-24
2020-09-14
2020-10-05
2020-10-26
2020-11-16
2020-12-07
2020-12-28
2021-01-18
2021-02-08
2021-03-01
2021-03-22
2021-04-12
2021-05-03
2021-05-24
2021-06-14
2021-07-05
2021-07-26
2021-08-16
2021-09-06
2021-09-27
2021-10-18
Frecuencia por millón

rider

Anglicismo: sí

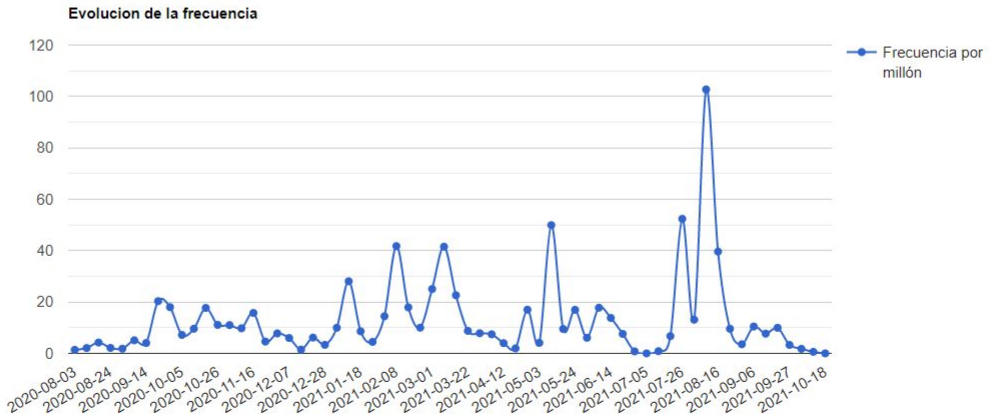
Formas: rider riders

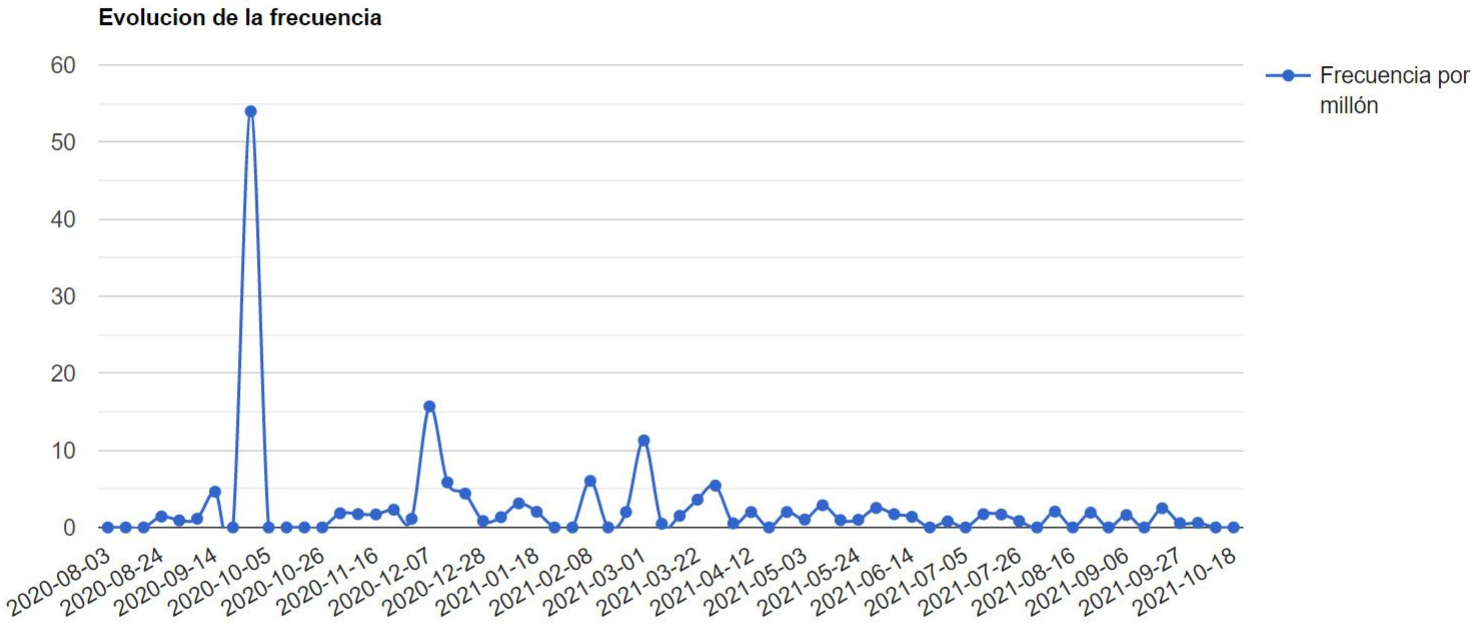
Frecuencia media de aparición*: 12.965

Frecuencia de aparición en el último mes*: 2.39

Secciones habituales: Economía Portada Tecnología España Medio Ambiente

* Frecuencia por cada millón de palabras medida desde agosto de 2020.




Evolucion de la frecuencia
60
50
40
30
20
10
0
2020-08-03
2020-08-24
2020-09-14
2020-10-05
2020-10-26
2020-11-16
2020-12-07
2020-12-28
2021-01-18
2021-02-08
2021-03-01
2021-03-22
2021-04-12
2021-05-03
2021-05-24
2021-06-14
2021-07-05
2021-07-26
2021-08-16
2021-09-06
2021-09-27
2021-10-18
Frecuencia por millón

black

Anglicismo: sí

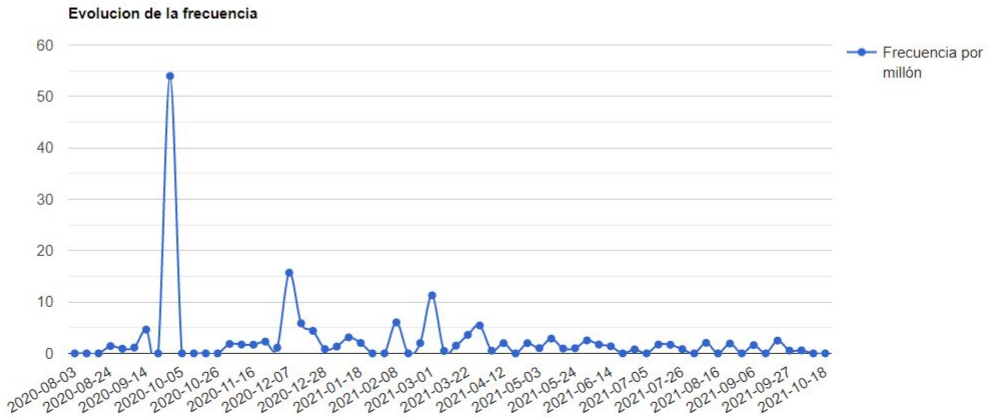
Formas: black

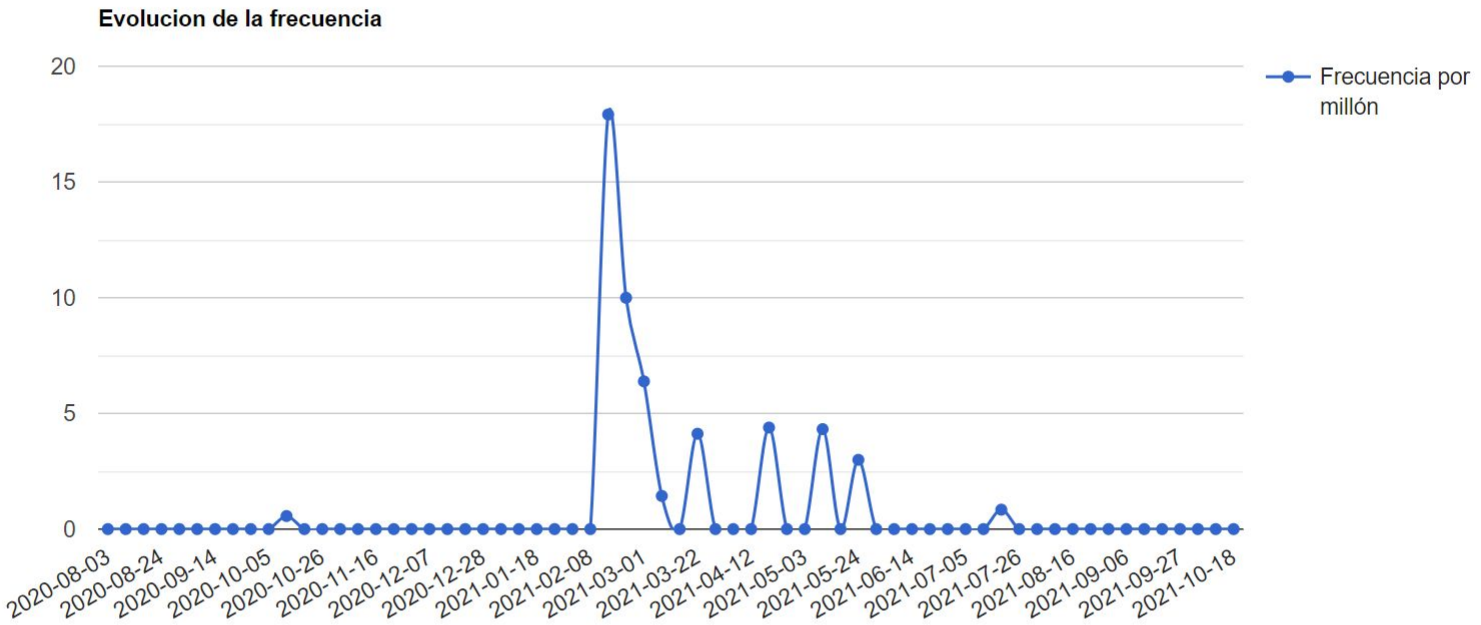
Frecuencia media de aparición*: 2.647

Frecuencia de aparición en el último mes*: 0.531

Secciones habituales: Portada Economía España Política Cultura

* Frecuencia por cada millón de palabras medida desde agosto de 2020.




Evolucion de la frecuencia
20
15
10
5
0
Frecuencia por millón
2020-08-03
2020-08-24
2020-09-14
2020-10-05
2020-10-26
2020-11-16
2020-12-07
2020-12-28
2021-01-18
2021-02-08
2021-03-01
2021-03-22
2021-04-12
2021-05-03
2021-05-24
2021-06-14
2021-07-05
2021-07-26
2021-08-16
2021-09-06
2021-09-27
2021-10-18

foam

Anglicismo: sí

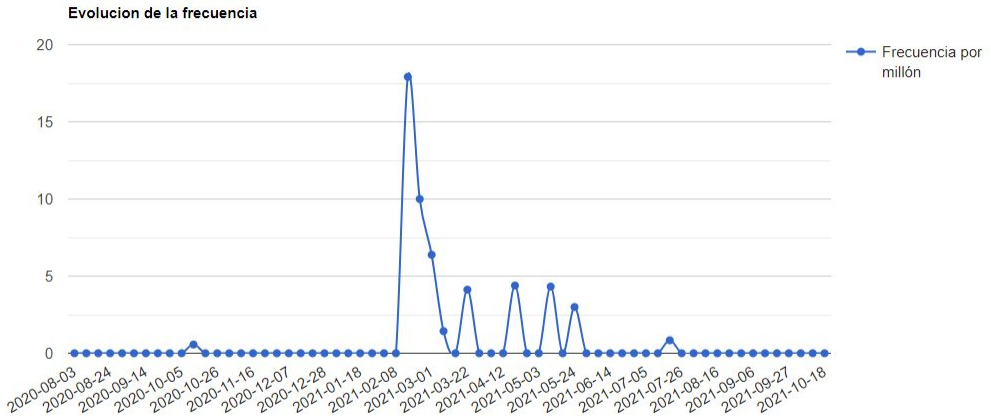
Formas: foam

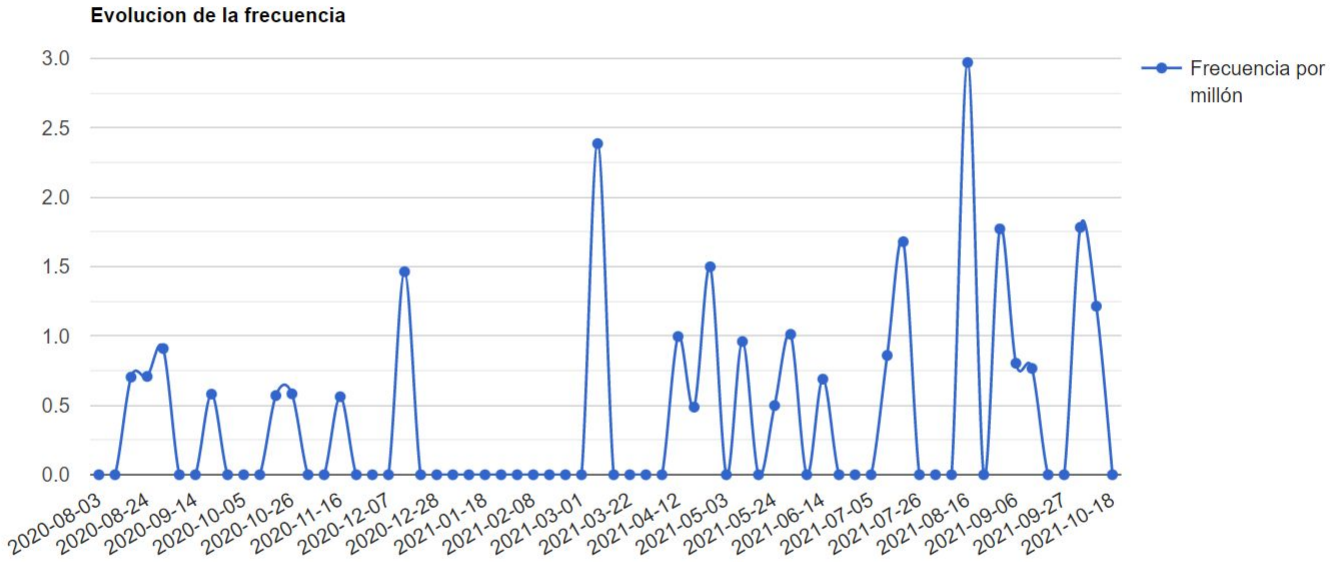
Frecuencia media de aparición*: 0.997

Frecuencia de aparición en el último mes*: 0

Secciones habituales: Portada Política España Opinión Moda

* Frecuencia por cada millón de palabras medida desde agosto de 2020.




Evolucion de la frecuencia
3.0
2.5
2.0
1.5
1.0
0.5
0.0
2020-08-03
2020-08-24
2020-09-14
2020-10-05
2020-10-26
2020-11-16
2020-12-07
2020-12-28
2021-01-18
2021-02-08
2021-03-01
2021-03-22
2021-04-12
2021-05-03
2021-05-24
2021-06-14
2021-07-05
2021-07-26
2021-08-16
2021-09-06
2021-09-27
2021-10-18
Frecuencia por millón

booster

Anglicismo: sí

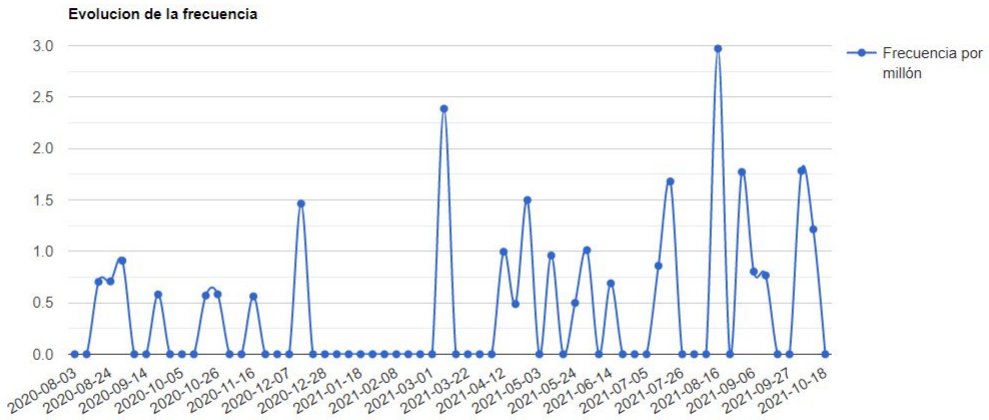
Formas: booster boosters

Frecuencia media de aparición*: 0.393

Frecuencia de aparición en el último mes*: 0.664

Secciones habituales: Moda Salud Sociedad Portada España

* Frecuencia por cada millón de palabras medida desde agosto de 2020.



lazarobot
@lazarobot

anxiety baking

"...milénicos son especialmente dados al anxiety baking, la práctica de preparar..."

Translate Tweet

Horneamos por encima de nuestras posibilidades
El frenesí repostero de la cuarentena se explica por la necesidad de llenar horas muertas, las ansias de aplausos en las redes sociales y la búsqueda de un ...
elpais.com

6:35 PM · May 3, 2020 · lazarobot

3 Retweets **9** Likes

lazarobot
@lazarobot

old school

"...en su diseño básico muy old school, pero también decisivo en..."

Translate Tweet

EL PAÍS

La función no puede continuar

El dramaturgo alemán Roland Schimmelpfennig llora el cierre de los teatros en este artículo escrito durante el confinamiento por el coronavirus

elpais.com

10:43 AM · May 14, 2020 · lazarobot

Observatorio Lázaro
@lazarobot

male tears

"...y a formatos clásicos de meme) para tratar algunos temas, como las male tears vertidas tras el estreno del videojuego Last of us 2, cuya..."

Translate Tweet

verne.elpais.com
Videojuegos y el machismo que siempre pica
Hablamos con la creadora de Feminismoen8bits, una cuenta de Instagram en la que se habla de videojuegos desde una ...

4:12 PM · Nov 23, 2020 · lazarobot

Observatorio Lázaro
@lazarobot

red carpet

"...Laguna solo hace vestidos largos de red carpet, este look de Clara Courel..."

Translate Tweet

vanitatis.elconfidencial.com
Señoras y señores, comienza la MBFWM: la crónica de las ...
Repasamos los momentos más señalados de la pasarela madrileña

9:53 PM · Sep 12, 2020 · lazarobot

lazarobot
@lazarobot

Crying

"...:. . ORIGINAL:. . Cryings not for me. . .."

Esta es la playlist que necesitas para salir a la calle y disfrutar del aire fresco
Si buscas temazos para practicar ejercicio al aire libre o, simplemente, quieres dar un paseo y mantener el ánimo arriba, aquí tienes lo que necesitas.
20minutos.es

12:18 PM · May 14, 2020 · lazarobot

lazarobot
@lazarobot

brilli-brilli

"...glitter o brilli-brilli de la artista..."

Translate Tweet

Tres españoles recopilan más de 300 obras de 'Arte Covid' de todo el mundo. ¿H...
Cuando esto acabe será fundamental recoger un testimonio emocional y artístico de cómo el virus nos ha afectado , dice uno de los impulsores de Covid Art ...
elpais.com

7:24 PM · Apr 23, 2020 · lazarobot

Observatorio Lázaro
@lazarobot

shosho

"...eso sería como el shosho de la Bernarda...."

Translate Tweet

elpais.com
El ‘divorcio’ obligado de los Estopa
David y José Muñoz, 21 años después de su primer éxito, llevan una vida corriente y gozan de un sólido patrimonio qu...

8:26 PM · May 5, 2020 · lazarobot

1 Quote Tweet **2** Likes

Observatorio Lázaro
@lazarobot

date prisa

"...Si te decides por estas, date prisa y fíchalas por el mismo precio..."
vanitatis.elconfidencial.com/estilo/moda/20...

Translate Tweet

4:46 PM · May 4, 2020 · lazarobot

¿Preguntas?

Elena Álvarez Mellado
http://observatoriolazaro.es/
@lirondos @lazarobot