Introducción al procesamiento del lenguaje natural (PLN)

Leticia Martín-Fuertes Moreno
Lingüista computacional en Bitext
@nimbusaeta
nimbusaeta@gmail.com

Lingẅars @ Cylicon Valley

9 de noviembre de 2017

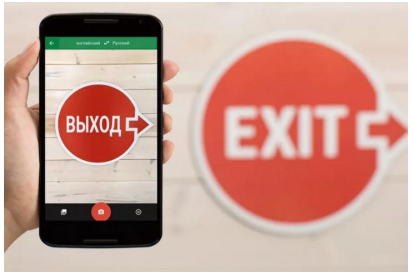
¿QUÉ ES EL PLN?

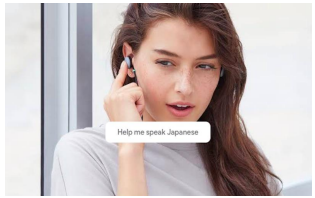
- Lingüística computacional
- Procesamiento del lenguaje natural (PLN o NLProc)
- Text mining

Área multidisciplinar que combina:

- Lingüística: fonética, sintaxis, semántica
- Informática: programación, aprendizaje automático
- Estadística, probabilidad, análisis de datos
- Lógica, formalización y representación del conocimiento

¿QUÉ ES PLN? PLN ERES TÚ





Historia del PLN en un vistazo

- En los 50, al principio de la Guerra Fría, el objetivo del PLN era la **traducción automática**.
- En los 60 aparecen **sistemas expertos** que asistían en la toma de decisiones: sistemas de diálogo que trataban de imitar conversaciones humanas, creación de ontologías para capturar conocimiento del mundo.
- Hasta los 80, la mayor parte de los sistemas de PLN estaban basados en conocimiento y manejaban complejas reglas diseñadas a mano. Influencia de la **lingüística generativa** de Chomsky.
- A partir de esa década, irrumpen las **aproximaciones estadísticas** basadas en sistemas de aprendizaje automático (*machine learning*), que requieren grandes colecciones de datos anotados manualmente. Desarrollo paralelo al aumento de potencia de los ordenadores.
- Actualmente, vivimos un auge de los **sistemas de aprendizaje automático no supervisados** (es decir, no anotados), con especial énfasis en el uso de la Web. Explosión de datos en formato electrónico.
- En la década de 2010 hemos visto el resurgir de los sistemas que utilizan **redes neuronales**.

Tareas propias del PLN

Tokenización

Dividir un texto en palabras.

In [2]:
```
import nltk
from nltk import word_tokenize, pos_tag

tokens = word_tokenize(raw)

print(tokens[:60])
print(len(tokens))
```

```
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME', 'I', 'CHAPTER', 'I', 'Emma', 'Woodhouse',
 ',', 'handsome', ',', 'clever', ',', 'and', 'rich', ',', 'with', 'a', 'comfortable', 'home', 'and', 'happ
y', 'disposition', ',', 'seemed', 'to', 'unite', 'some', 'of', 'the', 'best', 'blessings', 'of', 'existenc
e', ';', 'and', 'had', 'lived', 'nearly', 'twenty-one', 'years', 'in', 'the', 'world', 'with', 'very', 'li
ttle', 'to', 'distress', 'or', 'vex', 'her', '.', 'She', 'was']
191673
```

Tareas propias del PLN

Part-Of-Speech tagging (POS-tagging o etiquetado morfológico)

Asignar a cada palabra su categoría gramatical (o *parte del discurso*). Si puede ser la específica para ese contexto, mejor.

You can't do your work like this. The computer doesn't work!

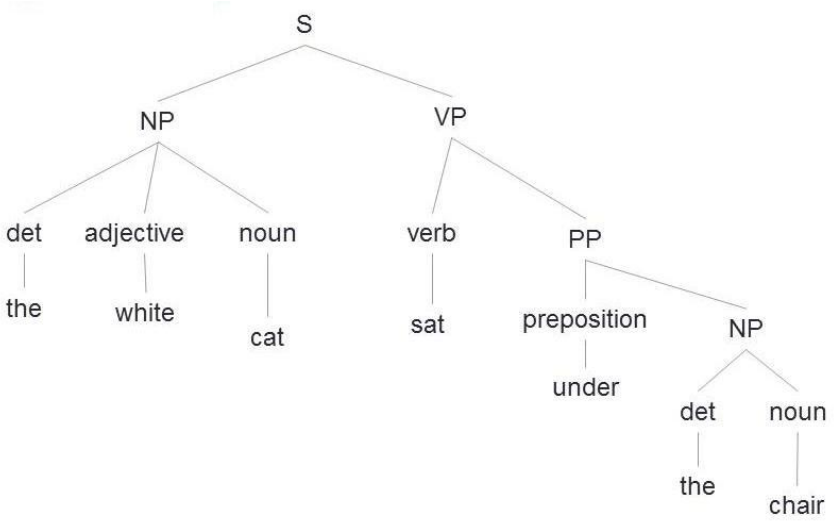
Form	POS
You	pronoun
can	verb
not	adverb
do	verb
your	determiner
work	noun
like	preposition
this	pronoun

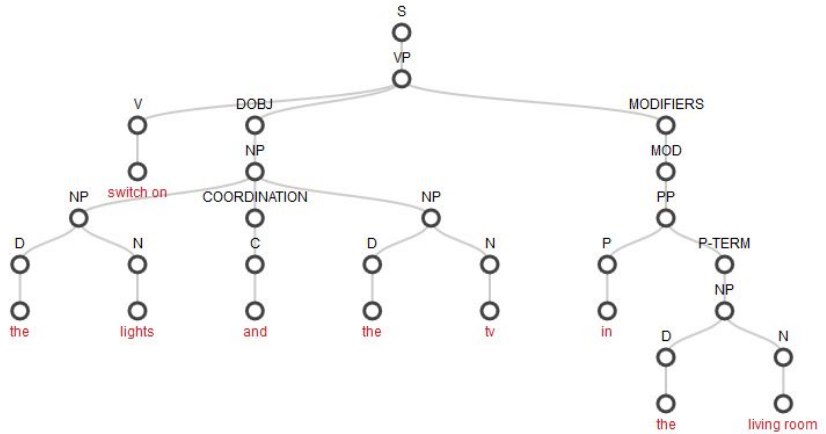
Form	POS
The	determiner
computer	noun
does	verb
not	adverb
work	verb

Tareas propias del PLN

Parsing (o análisis sintáctico)

Analizar una cadena de caracteres según las reglas de una gramática formal. En la práctica, consiste en dividir un texto en constituyentes, obteniendo como resultado un árbol que muestre las relaciones entre estos.





Tareas propias del PLN

Reconocimiento de entidades (o NER, *Named entities recognition*)

En un texto dado, localizar y clasificar entidades con nombre en categorías definidas, como pueden ser nombres de personas, organizaciones, lugares, expresiones de tiempo, cantidades, valores monetarios, porcentajes, etc.

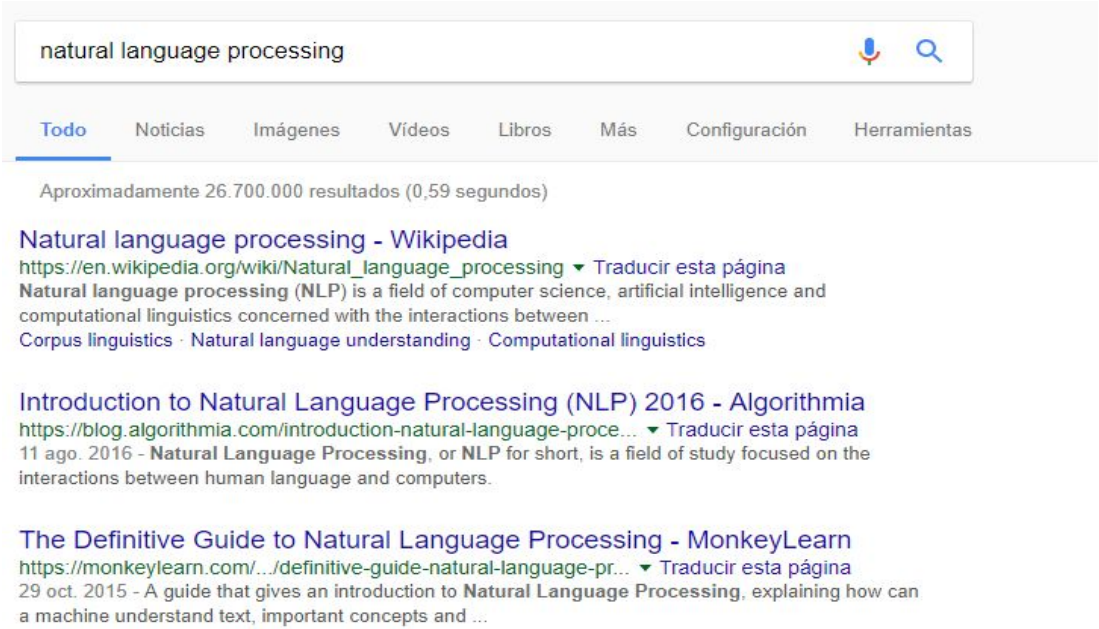
In [15]:
```
Dracula = read_lines("data/corpus_misc/en/literature/Dracula.txt", 200, 500)
sample = Dracula
sample = sample.replace('\n', ' ')
sentences = nltk.sent_tokenize(sample)
tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
tagged_sentences = [nltk.pos_tag(sentence) for sentence in tokenized_sentences]
chunked_sentences = nltk.ne_chunk_sents(tagged_sentences, binary=True)

def extract_entity_names(t):
    _entity_names = []
    if hasattr(t, 'label') and t.label():
        if t.label() == 'NE':
            _entity_names.append(' '.join([child[0] for child in t]))
        else:
            for child in t:
                _entity_names.extend(extract_entity_names(child))
    return _entity_names

entity_names = []
for tree in chunked_sentences:
    entity_names.extend(extract_entity_names(tree))

entity_names = set(entity_names)
print(entity_names)
```

```
{'Hospadars', 'German', 'Dacians', 'Mittel Land', 'Servian', 'Carpathian', 'Castle Dracula', 'Pass', 'Borgo Prund', 'Germany',
 'English Churchman', 'Wallachs', 'Hotel Royale', 'Attila', 'Borgo Pass', 'Jonathan Harker', 'Golden Mediasch', 'Turkish', 'Kla
usenburgh', 'Turk', 'Count Dracula', 'London', 'St. George', 'Europe', 'Slovak', 'France', 'British Museum', 'English', 'Mina',
'Golden Krone Hotel', 'Isten', 'Transylvania', 'Moldavia', 'Herr Englishman', 'Danube', 'Cszeks', 'Bukovina', 'Bistritz', 'Sout
h', 'Vienna', 'East', 'West', 'DRACULA', 'North', 'Huns', 'China', 'Magyars', 'Oriental', '_3 May', 'Slovaks', 'Szekelys', 'Mun
ich'}
```

Tareas propias del PLN

Recuperación de información (*information retrieval*)

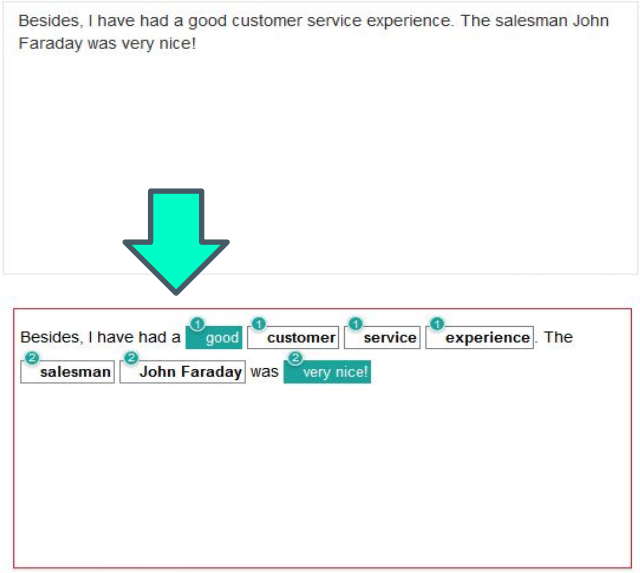
Devolver una lista de documentos relevantes para una búsqueda.



Tareas propias del PLN

Análisis de opinión (*sentiment analysis*)

Determinar de una forma automática la actitud del hablante hacia un tema, o la polaridad contextual o reacción emocional a un determinado producto, documento, interacción o evento.



Dan Jurafsky

Language Technology

mostly solved

Spam detection

Let's go to Agra! ✓

Buy V1AGRA ... ✗

Part-of-speech (POS) tagging

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Named entity recognition (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

making good progress

Sentiment analysis

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Coreference resolution

Carter told Mubarak he shouldn't run again.

Word sense disambiguation (WSD)

I need new batteries for my ***mouse***.

Parsing

I can see Alcatraz from the window!

Machine translation (MT)

第13届上海国际电影节开幕...

The 13th Shanghai International Film Festival...

Information extraction (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party May 27 add

still really hard

Question answering (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Paraphrase

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Summarization

The Dow Jones is up

The S&P500 jumped

Housing prices rose

→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

Lo último de lo último

Aplicar *machine learning*, *deep learning* y redes neuronales a PLN para desarrollar bots conversacionales (chatbots)

Plataformas PLN

Dialogflow

wit.ai

LUIS

Lex

IBM Watson

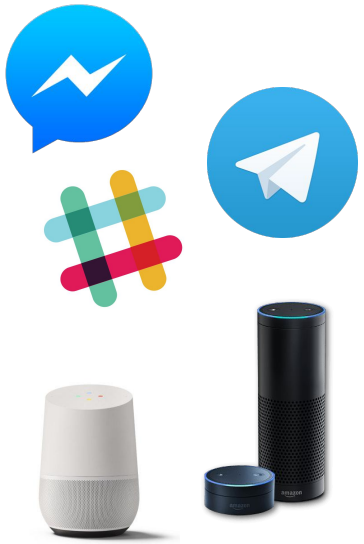
ORACLE

Construcción

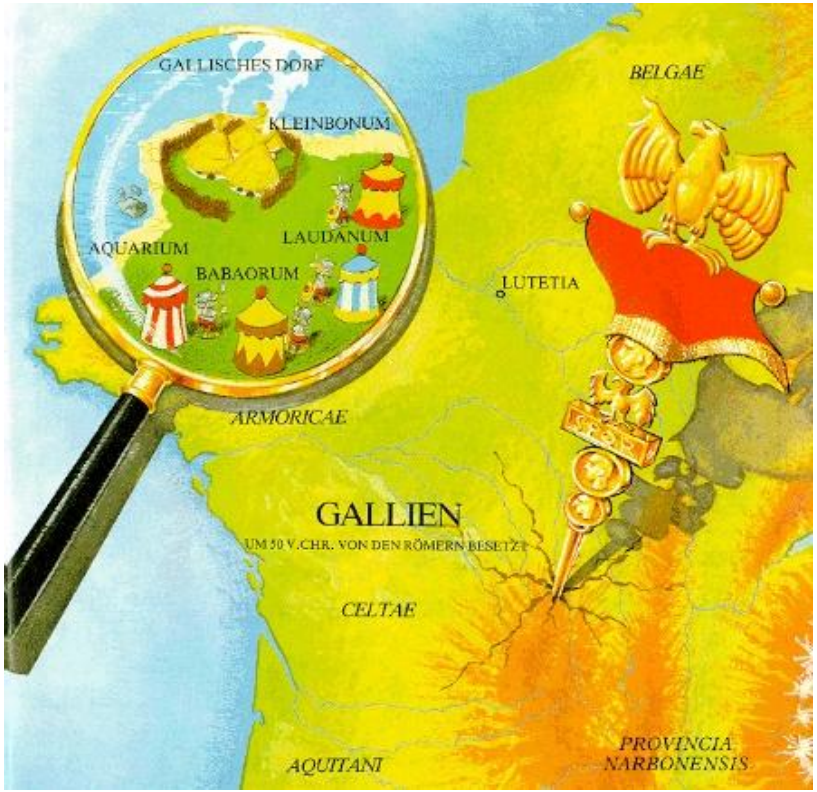
Chatfuel

RECAST.AI

Canales



Y en una irreductible aldea gala...



Lingwarriors
de **Lingwars**

Charlas y talleres

GAPLEN

Objetivos

- Lo que nos une es el interés por el PLN. No hay un lenguaje de programación preferido.
- Quién: gente «de letras» + gente «de ciencias»
- Se valoran todos los conocimientos y todos los niveles
- Enfoque lingüístico sobre el enfoque estadístico

Trayectoria

22/07/2015 - Primera quedada. Hasta mediados de noviembre: quedadas semanales los miércoles o jueves

03/10/2015 - Jornada entera: **presentación de proyectos** desarrollados en Lingwars + Introducción a Python

14/11/2015 - Jornada entera, edición especial corpus: AntConc + **presentación de proyectos** (Neutrón, Aracne, enclitizador) + Recuperación de información

Trayectoria

09-23/04/2016 - Taller de **tratamiento de datos digitales** (scraping, análisis de corpus, visualización de datos) por Javi G. Sogo + Tania Karaseva

07/05/2016 - Taller de la **API de Twitter** por Javi G. Sogo

09/07/2016 - Charla sobre **semántica estructural** por Mónica González Manzano

16/07/2016 - Taller de **Jupyter para lingüistas** por Javi G. Sogo

12/12/2016 - Charla sobre la **sílaba y Prolog** por Elena Álvarez Mellado

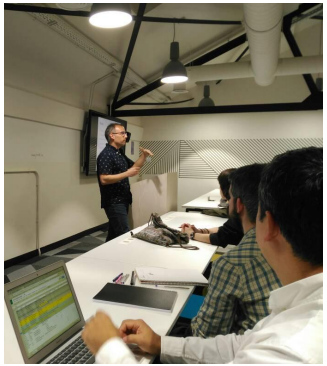
Trayectoria

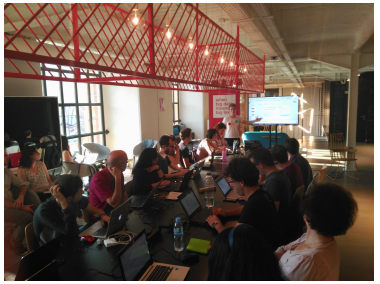
22/02/2017 - Charla sobre **cómo hacerse una carrera en PLN** por Víctor Peinado

08/07/2017 - Taller de **NLTK** por Leticia Martín-Fuertes, Tania Karaseva y Javi G. Sogo

05/10/2017 - GAPLEN #1 - Charla de **introducción a PLN** por Luis Anke Espinosa

22/10/2017 - Charlas sobre **ontologías + PLN en R** por Lorena Giusio y Claudia Guirao @ R-Ladies Madrid





Trayectoria

02/11/2017 - GAPLEN #2 - Charla **Rapidminer** por Enrique Puertas

09/11/2017 - Charlas **Corpus + Ontologías** por Leticia Martín-Fuertes y Javi G. Sogo @ UVa

09/11/2017 - Charlas de **introducción a PLN + ontologías** por Leticia Martín-Fuertes y Javi G. Sogo @ Cylicon Valley

22/11/2017 - Charla sobre **resumen automático** por Luis Anke Espinosa

???

Nos inspiran

El granjero que creó un tractor robot aprendió programación con un curso online gratuito

- Matthew Reimer es un granjero canadiense que usó software libre para programar un tractor que funciona de forma autónoma
- No tenía idea de programación hasta que tomó un curso gratuito del MIT y modificó software diseñado para drones para adaptarlo a su tractor

El procesador de textos que gustó a Apple

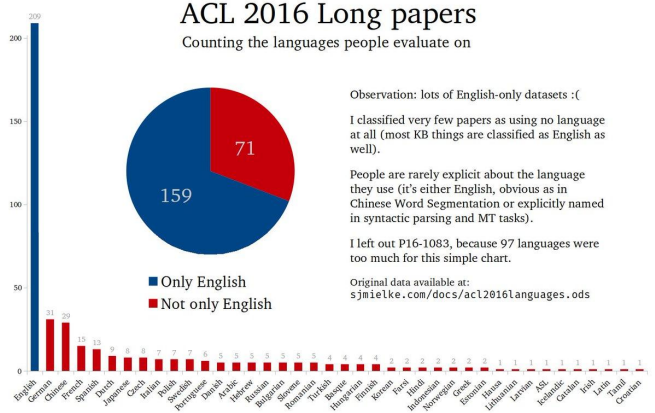
Paul Lutus empezó a escribir programas por diversión que dibujaban imágenes en la pantalla o realizaban acciones "elegantes y matemáticas" con su Apple II. Envió algunos a la compañía, que decidió comprárselos, así que en poco tiempo había recuperado el coste de aquel novedoso ordenador lanzado en 1977.

Se marcó entonces su siguiente objetivo: desarrollar un procesador de textos que dejara obsoleta su vieja máquina de escribir. Con él podría escribir más cómodamente un artículo sobre la teoría de la relatividad que **le había encargado una revista**. **A Apple también le interesó** y le pagó 7.500 dólares (unos 26.000 euros al cambio actual, teniendo en cuenta la **inflación**) por ese 'software'.

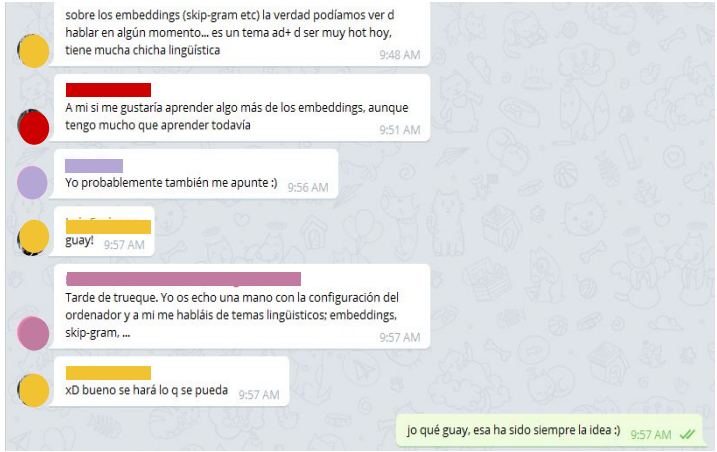
'Obijuan' te quiere enseñar a hacer un chip

El doctor en Informática y Telecomunicaciones Juan González, "Obijuan" en la comunidad 'online', está enfrascado en el "gran reto" de enseñar a hacer chips a cualquier persona, incluso a quien es "no técnico".

EFEFUTURO BILBAO | SÁBADO 28.10.2017



Espíritu



juliverx
Google Campus Ma...
Following

juliverx Aquí estamos: lingüinis y techies con @nimbusaeta y @lirondos en @lingwars

playmobiles Mola. Me apunto a la próxima. Sea lo que sea

elizazuben Love your pictures!

nimbusaeta @juliverx ¡Gracias! @playmobiles Somos una especie de grupo de apoyo para aprender procesamiento de lenguaje natural, resumidamente :)

LingWars @lingwars
Following

«Yo estoy emocionado solo con el hecho de que estemos aquí reunidas unas cuantas personas hablando de estas cosas» se ha dicho hoy aquí <3

Elena A. Mellado @lirondos · 19 ago. 2015
En respuesta a @lirondos
Representación gráfica de lingufriendly.

LingWars @lingwars · 8 May 2016
El currazo que se ha pegado @jgsogo últimamente nos ha dejado 2 magníficos #tutoriales: cómo scrapear la Web lingwars.github.io/blog/scrape-xp...
Translate from Spanish
1 2 4

LingWars
@lingwars
Following

Y cómo conectarse a la API de Twitter lingwars.github.io/blog/twitter-s... (sí, contiene información sobre qué leches es una API) #tutoriales

Elena A. Mellado @lirondos · 6 Dec 2016
"Dime, oh Azrael, en qué idioma está escrito este texto".

Hoy ha sido un día provechoso :_)
Translate from Spanish

ɬta aszetika deritzon a: (1643) euskal prosa
ɔo Agerre Azpilkueta (A.33333333333333
ɔarraldeko Saran emai.22222222222222
53.333333333333336
, geroago lapurtera kl.66666666666667
azpideak bilatzen ditu.7.77777777777778
enetako sortzailea, Le
4 1 15

Elena A. Mellado
@lirondos
Following

De esta criatura y del cómo se hizo hablaremos entre otras cosas el lunes. Es gratis, es chachi, somos majos

Leticia MFM
@nimbusaeta

Bieeeen 👏👏 ya hemos llegado al momento típico en que alguien dice que las regex son un quebradero de cabeza xD @claudiaguirao

¿Qué nos depara el futuro?

THAT'S ALL, FOLKS!

lingwars@gmail.com
@lingwars