Scrapeando el BOE descubrí un tesoro

Leticia Martín-Fuertes
@nimbusaeta
Elen Irazabal
@IrazabalElen

R-Ladies Madrid @ Xantardev
7 de septiembre de 2019

¿Qué es el PLN?

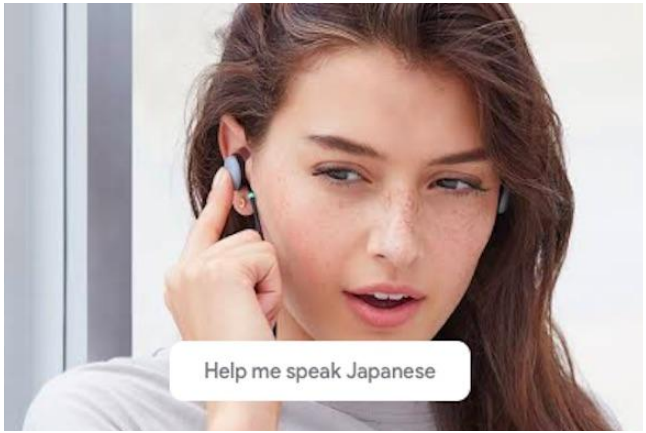
- Procesamiento del lenguaje natural (PLN o NLProc)
- Lingüística computacional
- Tecnologías del lenguaje
- Text mining

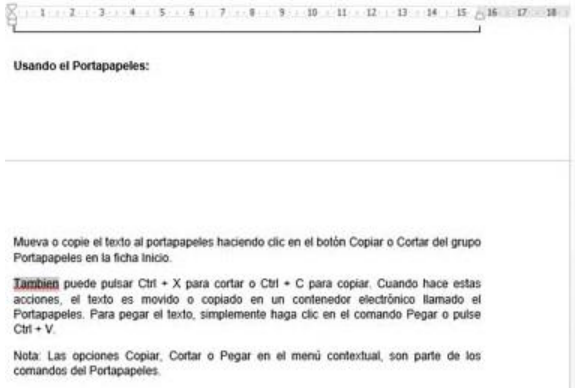
Área multidisciplinar que combina:

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

Área de la IA que trabaja con datos de tipo lingüístico (texto, audio)

Aplicaciones prácticas del PLN



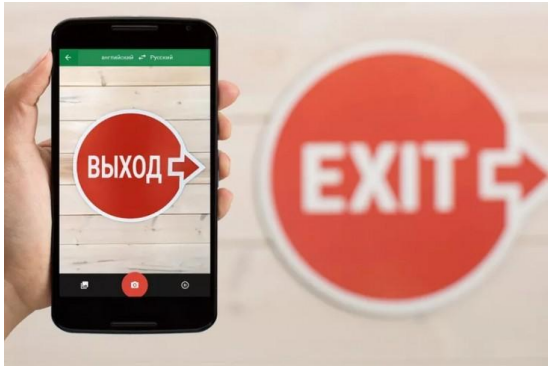


Usando el Portapapeles:

Mueva o copie el texto al portapapeles haciendo clic en el botón Copiar o Cortar del grupo Portapapeles en la ficha Inicio.

También puede pulsar Ctrl + X para cortar o Ctrl + C para copiar. Cuando hace estas acciones, el texto es movido o copiado en un contenedor electrónico llamado el Portapapeles. Para pegar el texto, simplemente haga clic en el comando Pegar o pulse Ctrl + V.

Nota: Las opciones Copiar, Cortar o Pegar en el menú contextual, son parte de los comandos del Portapapeles.



Algunos conceptos

generar el mensaje

Natural Language Generation (NLG)

verbalizar el mensaje

Text-to-speech (TTS)

mensaje (oral o escrito)

percibir el mensaje

Optical Character Recognition (OCR)

Automatic Speech Recognition (ASR)

interpretar el mensaje

Natural Language Understanding (NLU)

La nueva voz de la aplicación de Google (EN, 2016)

Tareas del PLN

Tokenización

Enclíticos:

cuéntamelo = me lo cuentas

Contracciones:

don’t = do not (EN)

al = a el (ES)

Lematización

hablan → hablar

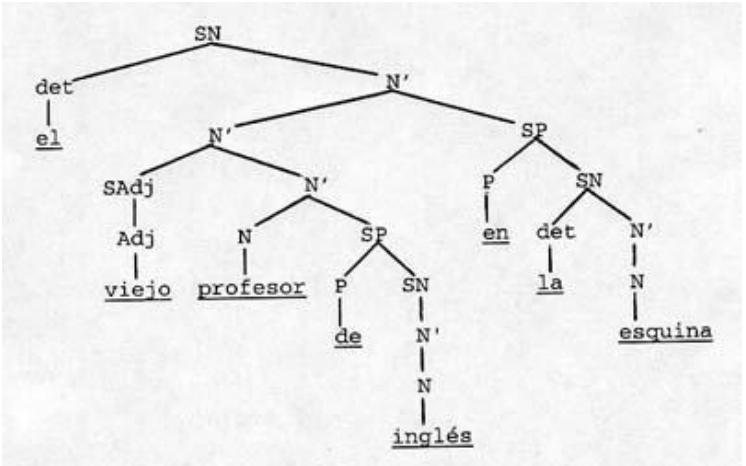
curiosas → curioso

POS-tagging

hablan<verb>

curiosas<adj>

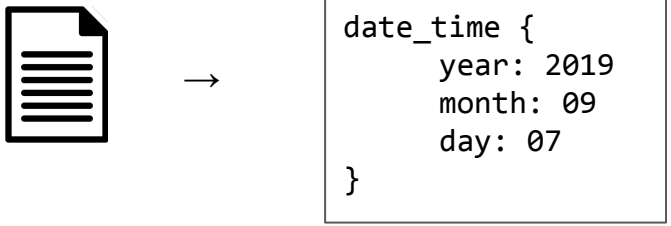
Parsing



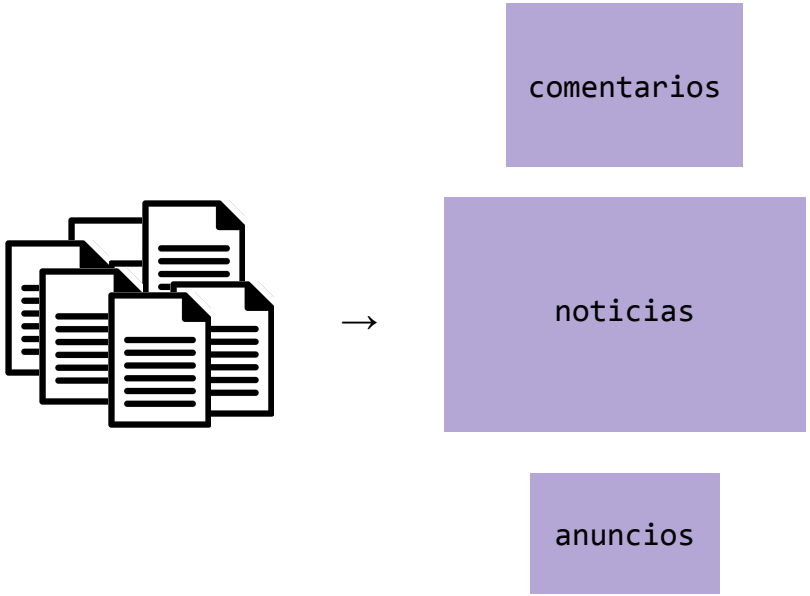
Tareas del PLN

Information retrieval (IR) o recuperación de información

Information extraction (IE) o extracción de información



Topic modeling



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

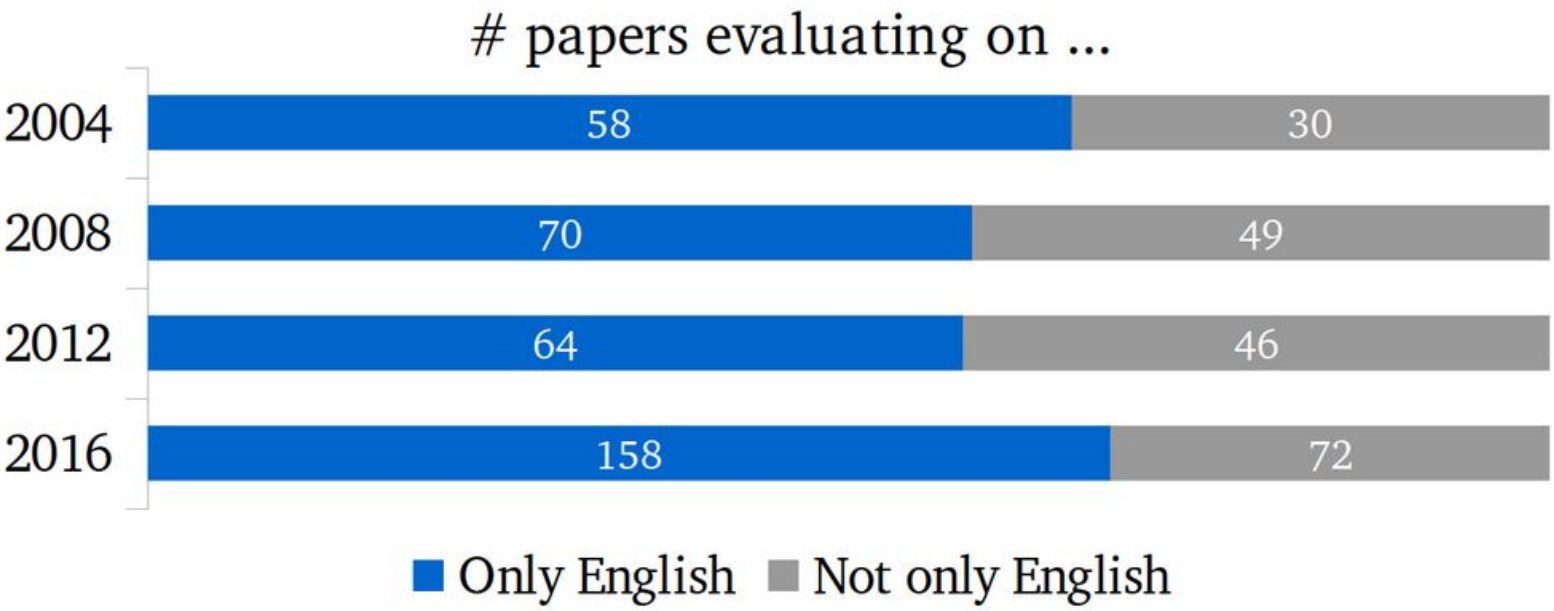
→ Economy is good

Dialog

Where is Citizen Kane playing in SF?

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

Procesamiento de... ¿la lengua inglesa?



[Language diversity in ACL 2004 - 2016]

Recursos y curiosidades

Newsletter de Sebastian Ruder: últimos papers, noticias, avances… tanto del mundo académico como del industrial - http://newsletter.ruder.io

Newsletter de Mariya Yao: más enfocada a chatbots, sistemas de diálogo, interfaces conversacionales, management… - https://mariyayao.com

NLTK book (Python) - https://www.nltk.org/book

Curso gratuito online de Spacy - https://course.spacy.io

OpenAI GPT2 demo (NLG) - https://gpt2.ai-demo.xyz

@Lingwars - https://twitter.com/lingwars