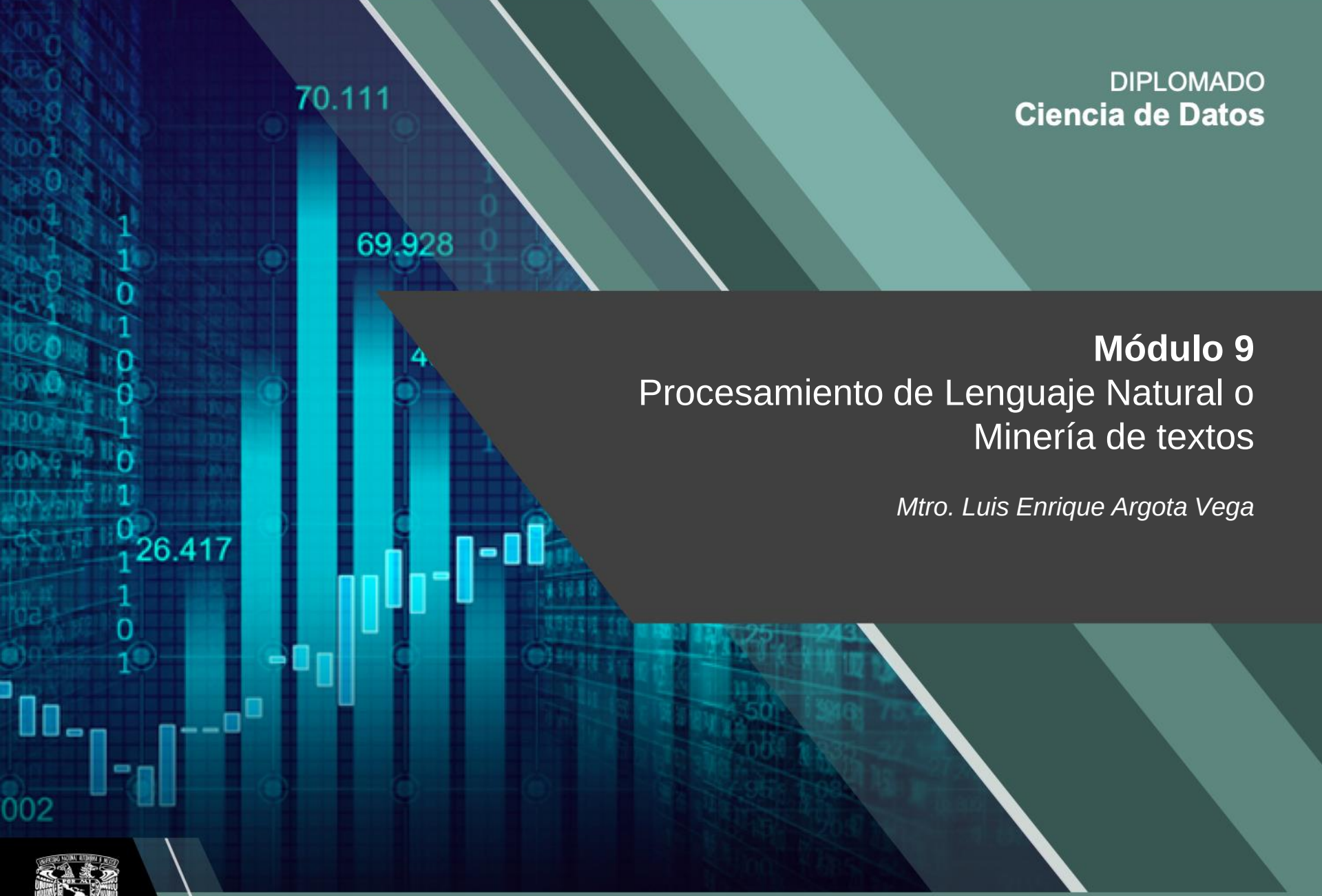


DIPLOMADO
Ciencia de Datos

Módulo 9
Procesamiento de Lenguaje Natural o Minería de textos

Mtro. Luis Enrique Argota Vega

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Tema 3: Procesamiento de Lenguaje Natural

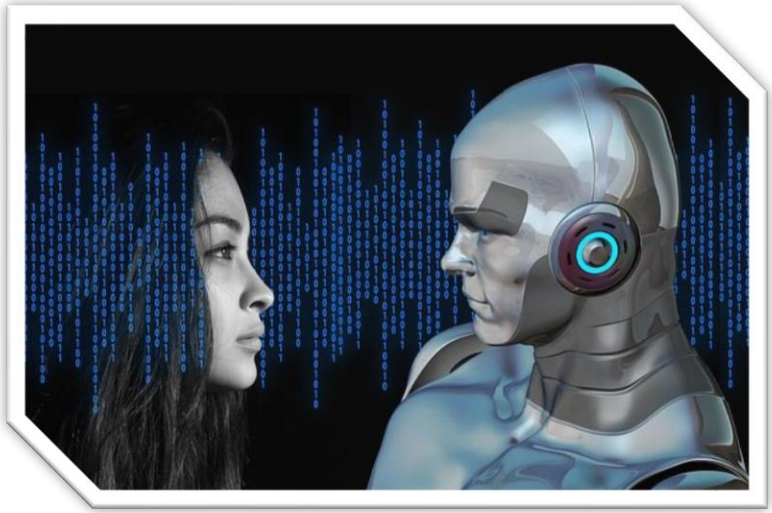
Objetivo

El participante identificará el tipo de operaciones usadas en el preprocesamiento de textos, con la finalidad de descubrir patrones en la colección de textos y darle estructura a los mismos para su posterior tratamiento computacional.

Contenido

1. Tareas básicas: conteo de palabras, tokenización, normalización, extracción de raíces, lematización, división de oraciones
2. Etiquetado gramatical y Análisis sintáctico
3. Herramientas para usar con Python (NLTK, FreeLing, Spacy, Stanza)

Introducción



Lenguaje

LENGUAJE NATURAL

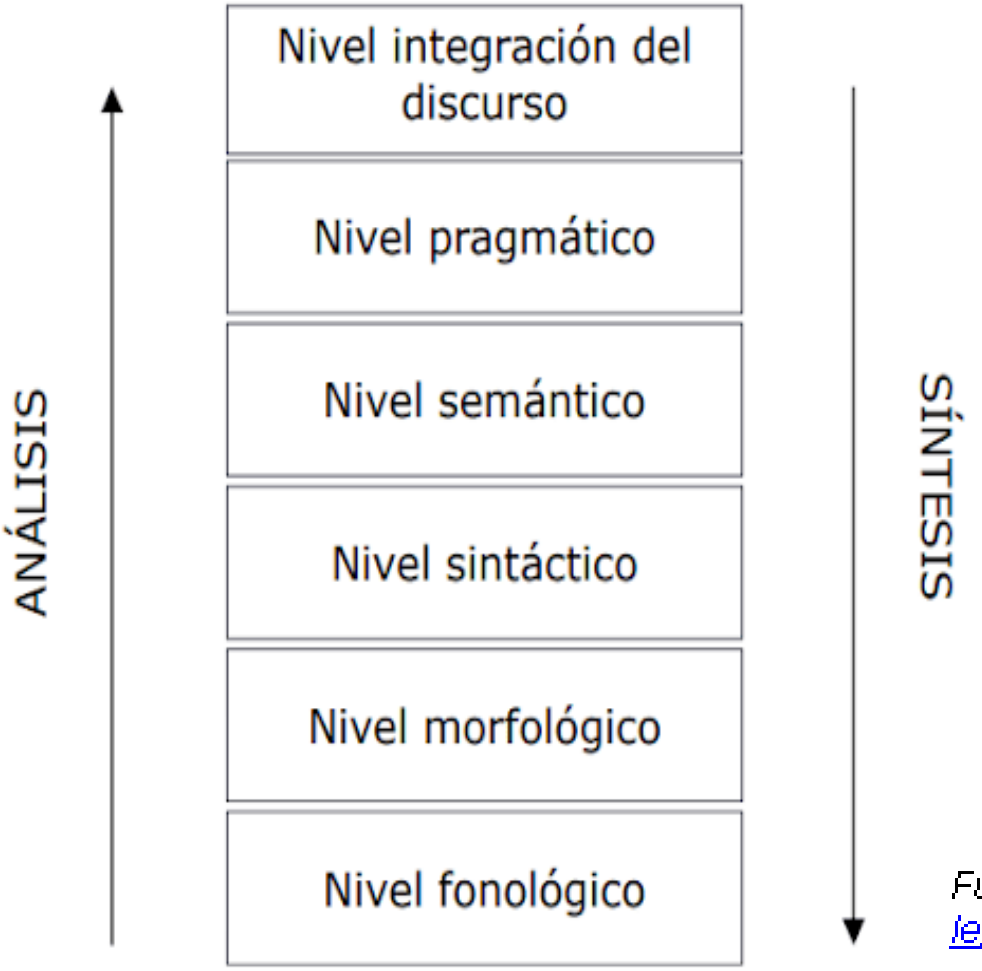
LENGUAJE FORMAL

LENGUAJE ARTIFICIAL

Visualización de video

- **Niveles del Procesamiento de Lenguaje Natural**

Arquitectura de un sistema PLN



Arquitectura de un sistema de PNL

Fuente: https://itelligent.es/es/procesamiento-del-lenguaje-natural-aplicaciones/

Problemas del PLN

- ✓ Ambigüedad
- ✓ Multiplicidad de variantes
- ✓ Evolución y cambio
- ✓ Oscuridad, slang, etc

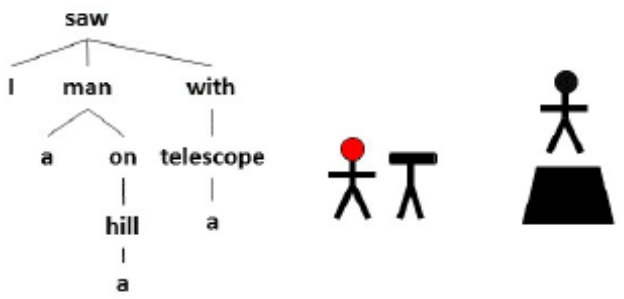
Ambigüedad

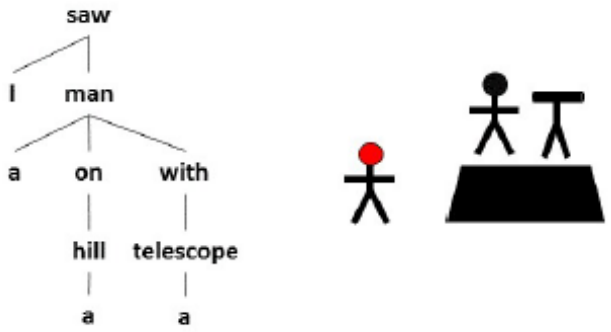
- ✓ Este problema de la ambigüedad se presenta en todos los niveles del lenguaje, sin excepción. Ejemplo:

  “Hay alguien en la puerta, que te quiere hablar”
  “Hay alguien, en la puerta que te quiere hablar”

Ambigüedad sintáctica

- Ambigüedad





I saw a man with a telescope

Ambigüedad léxico-semántica

- ¿Se quedará a dormir?
- Sí.
- Quizá debería saber que la casa está encantada.
- Ah, pues dígale que a mí también me hace ilusión quedarme.

El capitán dijo: "¡Bajen las velas!" Y los de arriba se quedaron sin luz.

- ¡Qué fresca está la mañana!
- Normal, es de hoy.

- ¿Por qué los de Lepe tiran a sus hijos a un pozo?
- Porque saben que en el fondo son buenos.

- ¿Qué pasa si un elefante se queda de pie encima de una pata?
- a- Que se cae.
- b- Que el pato se queda viudo.
- c- Que aplasta a su domador.

Ambigüedad fonológica / morfosintáctica

Dígame su nombre.

- Peter O´ Brian

- Decídase por favor.

- ¡Acusado! ¡Hable ahora o calle para siempre!

- Elijo calle.

Mi marido se ha ido de ca[S]a

1. de casa
2. de caza

Plata no es. Oro tampoco. ¿Qué es?

¿Qué esconde?

¿Que es conde?

Ambigüedad pragmática

- ¡Camarero! ¿Se puede saber qué está haciendo esta mosca en mi sopa?

- Mmm, yo creo que está nadando a braza, señor."

- ¿Cómo estás?
- Han perdido los Pumas

- Golpeó el armario con un palo y lo rompió.

MARÍA SIMARRO VÁZQUEZ, Humor verbal basado en la ambigüedad léxica y competencia léxico-semántica, Pragmalingüística 25 (2017): 618-636

Multiplicidad de variantes

- ✓ Existen alrededor de 7.097 idiomas en el mundo, según la revista “Ethnologue”.
- ✓ Con diferentes palabras, estructuras sintácticas, reglas morfológicas, sistemas fonéticos y escrituras.
- ✓ El intercambio – traducción entre unas y otras no es obvio.

En el país existen **11 familias lingüísticas…**
Se hablan 364 variantes lingüísticas

https://www.europapress.es/sociedad/noticia-idiomas-cifras-cuantas-lenguas-hay-mundo-20190221115202.html

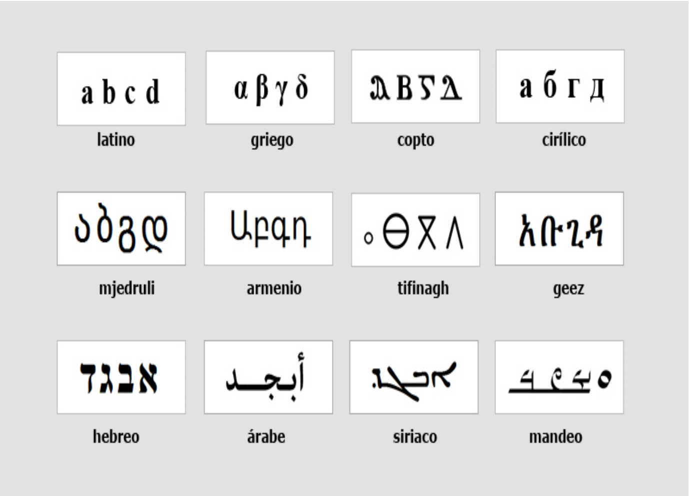
https://wals.info/

- ✓ ¿Sabías que en México hay 68 lenguas indígenas, además del español?

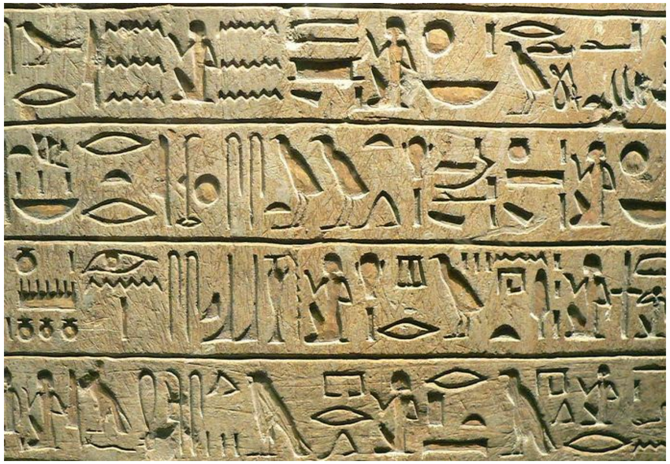
https://www.gob.mx/cultura/es/articulos/lenguas-indigenas?idiom=es

Multiplicidad de variantes

Diferentes Alfabetos



Escritura no alfabética



的 不 用 一 新 时 吗 所 被 说 男 事
人 在 之 二 市 也 家 后 只 着 女 得
上 于 要 三 与 还 可 分 都 位 几 真
中 前 好 四 内 出 件 种 做 把 各 对
下 者 了 五 本 去 里 将 已 吧 谁 看
大 会 年 六 地 到 最 很 长 难 找 见
小 号 月 十 这 他 回 而 行 来 子 加
是 我 日 个 此 性 万 数 等 站 字 更
没 和 为 次 建 就 能 天 再 每 那 多
有 你 名 元 全 部 爱 无 以 起 哪 少

Evolución y cambio

- ✓ Hay palabras que se incorporan a la lengua: tuit, celular, selfie.
- ✓ Hay palabras que desaparecen: doncel, jumento, vuestra merced, vosotros
- ✓ Otras cambian de sentido: hasta, celular, ratón
- ✓ Algunas estructuras sintácticas cambian: SOV (latín) -> SVO (español)
- ✓ Algunos fonemas (sonidos) desaparecen y aparecen otros nuevos: caballo [kabalo] > [kabaio] > [kabaʒo] / [kavaʃo]
- ✓ Muchas lenguas desaparecen. Otras se mezclan (pidgin y criollos). Otras resucitan (hebreo), y de otras solo quedan testimonios escritos (etrusco).

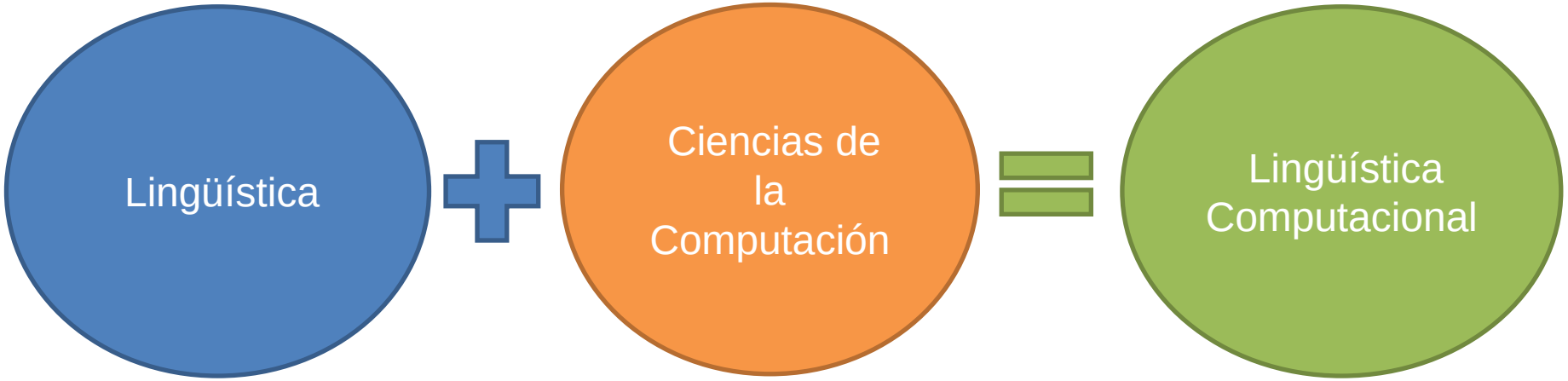
Oscuridad, slang, etc.

En ocasiones los humanos, no son claros ni precisos…

Amor, hagamos cuentas.
A mi edad
no es posible
engañar o engañarnos.
Fui ladrón de caminos,
tal vez,
no me arrepiento.
Un minuto profundo,
una magnolia rota
por mis dientes
y la luz de la luna
celestina.
Muy bien, pero, el balance?

https://www.poemas-del-alma.com/pablo-neruda-oda-al-amor.htm

Lingüística computacional



Tareas básicas para el procesamiento de textos

Prerrequisitos para analizar un texto:

- ✓ Ser capaz de dividir el texto por frases.
- ✓ Ser capaz de encontrar las palabras.
- ✓ DEFINICIÓN DE PALABRA: Todo aquello que se encuentra entre dos espacios en blanco o espacio en blanco y signo de puntuación.

Tokenizador

- ✓ Separa texto en una lista de tokens usando algunos caracteres como referencia para dividir.
- ✓ Los tokens son generalmente palabras y símbolos de puntuación.
- ✓ La tokenización no implica ningún nivel de análisis y se realiza muy rápido.

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text)
```

RUN

```
Veo al hombre con el telescopio.
Veo
al
hombre
con
el
telescopio
.
```

https://spacy.io/models/es

Análisis de textos

- ✓ **Niveles**:

- **Fonética-fonología (sonidos) – Corresponde al análisis de habla**
- Morfología (clases de palabras y segmentación)
- Sintaxis (oraciones, sintagmas y orden de las palabras)
- Semántica (significados)
- Pragmática (interacciones, uso y contexto)
- Discurso (expresiones correferenciales, estructura retórica)

Morfología

- ¿Cuáles son las clases de palabras y por qué importa saberlo?
- ¿Qué partes tienen las palabras? ¿Cómo se segmentan?

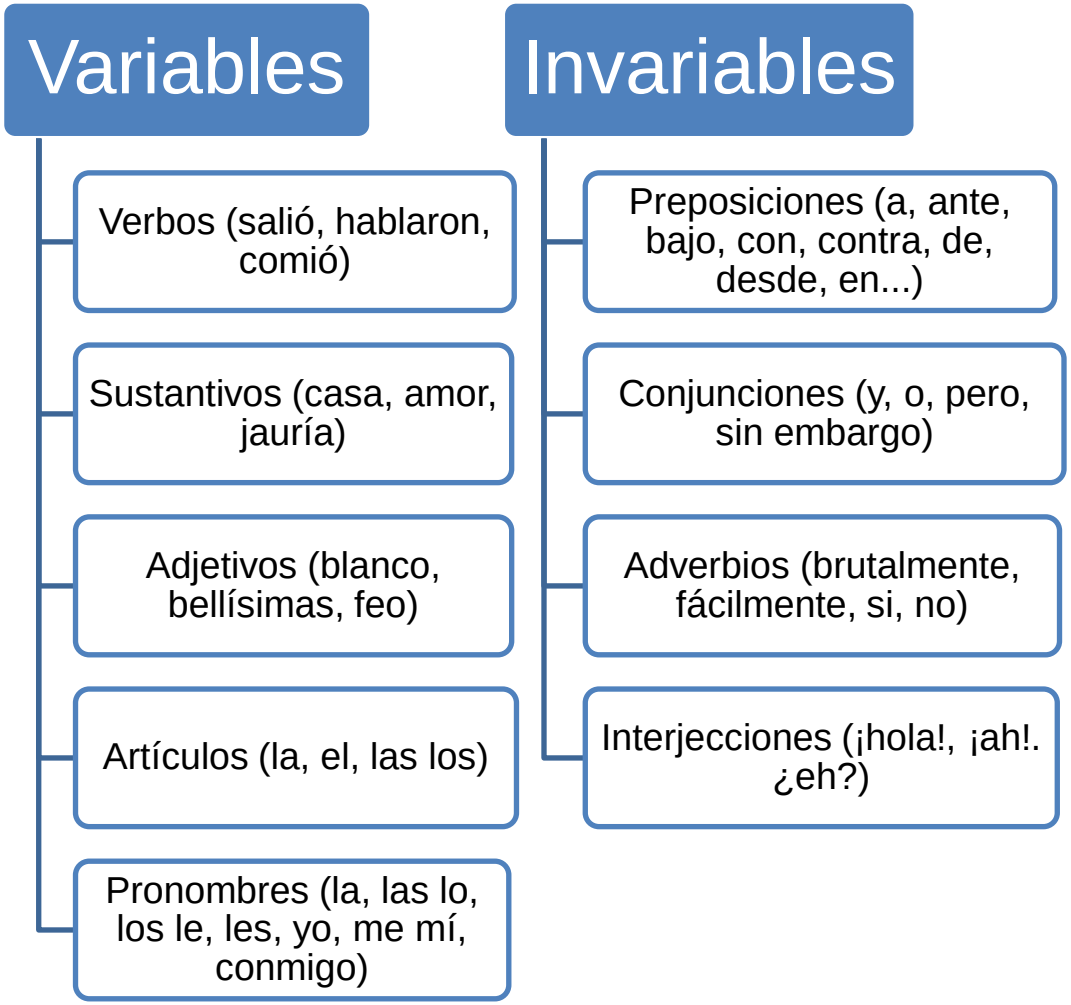
¿Cuáles son las palabras?

Pago por $475 347.50 M.N. (cuatrocientos setenta y cinco mil trescientos cuarenta y siete pesos 50/100 M.N.) del Restaurant Pomme de terre D'Opera por haberlo traspasado a Mary-Carmen da Cunha en San Luis Potosí.

7/02/2017

Clases de palabras

Part-of-speech (POS, etiquetado de parte del discurso), clases morfológicas, categorías gramaticales




CLASES DE PALABRAS
SUSTANTIVO
Se refiere a
seres
objetos
ideas
acontecimientos
Clases
Común
o
Concreto
Propio
o
Individual
Abstracto
o
Contable
Colectivo
o
No contable
ADJETIVO
Expresa
Cualidad o propiedad de un sustantivo
en grado
positivo
superlativo
comparativo
DETERMINANTES
Clases
Artículos
Demostrativos
Posesivos
Numerales
Indefinidos
Interrogativos
Exclamativos
clase de palabras que van **junto al nombre** para concretarlo o identificarlo
PRONOMBRES
Clases
Personales
Demostrativos
Posesivos
Numerales
Indefinidos
Interrogativos
Exclamativos
Relativos
Son
clase de palabras que **sustituyen** al nombre
PREPOSICIÓN
Sirven para
relacionar palabras, o grupos de palabras, entre sí en la oración
palabras invariales
No pueden aparecer solas
a, ante, bajo, cabe, con, contra,de, desde, en, entre, hacia, hasta, para, por, según, sin, so, sobre, tras
CONJUNCIÓN
Sirven para
unir dos elementos de una misma categoría gramatical (dos nombres, dos adjetivos o dos oraciones)
y, e, ni, pero, sino, aunque, sin embargo, no obstante, porque, cuando, donde,...
ADVERBIO
modifica el significado de un verbo, de un adjetivo o de otro adverbio
palabra invariable
VERBO
Expresa
acción, proceso o estado que ha sucedido, sucede o sucederá
Accidentes gramaticales
son
persona
tiempo
modo
número
aspecto
voz

Etiquetado Penn TreeBank

Penn TreeBank Tags está basado en el corpus Brown, pionero en etiquetado POS para inglés.

Number	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential *there*
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	*to*
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Fuente: https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html

Etiquetas EAGLES

Las etiquetas EAGLES codifican todas las características morfológicas existentes para la mayoría de los idiomas europeos, incluido el español. Estas etiquetas consisten en un conjunto de caracteres de longitud variable, donde cada uno corresponde a una característica morfológica.

1. Adjetivos
2. Adverbios
3. Artículos
4. Determinantes
5. Nombres
6. Verbos
7. Pronombres
8. Conjunciones
9. Numerales
10. Interjecciones
11. Abreviaturas
12. Preposiciones
13. Signos de Puntuación

NOMBRES			
Pos.	**Atributo**	**Valor**	**Código**
1	Categoría	Nombre	N
2	Tipo	Común	C
		Propio	P
3	Género	Masculino	M
		Femenino	F
		Común	C
4	Número	Singular	S
		Plural	P
		Invariable	N
5	Caso	-	0
6	Género Semántico	-	0
7	Grado	Apreciativo	A

Forma	Lema	Etiqueta
chico	chico	NCMS000
chicos	chico	NCMP000
chica	chica	NCFS000
chicas	chica	NCFP000
oyente	oyente	NCCS000
oyentes	oyente	NCCP000
cortapapeles	cortapapeles	NCMN000
tesis	tesis	NCFN000
Antonio	antonio	NP00000

https://www.cs.upc.edu/~nlp/tools/parole-sp.html

Problema del POS Tagging

- ✓ Las palabras, tomadas en forma aislada, son ambiguas respecto a su categoría.
- ✓ Pero… La categoría de la mayoría de las palabras no es ambigua dentro de un contexto.

Yo	bajo	con	el	hombre	bajo	a
PP	VM VM AQ NC SP	SP	TD	NC	VM VM AQ NC SP	NC SP

tocar	el	bajo	bajo	la	escalera	.
VM VM	TD	VM VM AQ NC SP	VM VM AQ NC SP	TD NC PP	NC	FP

Yo	bajo	con	el	hombre	bajo	a
PP	VM VM AQ NC SP	SP	TD	NC	VM VM AQ NC SP	NC SP

tocar	el	bajo	bajo	la	escalera	.
VM VM	TD	VM VM AQ NC SP	VM VM AQ NC SP	TD NC PP	NC	FP

- ✓ ***Solución:*** **Desambiguador Morfosintáctico (Pos tagger)**

Desambiguador morfosintáctico

Herramientas en línea:

- ✓ Linguakit: https://linguakit.com/es/etiquetador-morfosintactico
- ✓ Stanford Parser: http://nlp.stanford.edu:8080/parser/index.jsp
- ✓ Desambiguador morfosintáctico: http://protos.dis.ulpgc.es/investigacion/desambigua/morfosintactico.htm
- ✓ Stilus: https://www.mystilus.com/herramientas/analizador-morfosintactico

Etiquetador gramatical (POS)

- ✓ El etiquetado de parte del discurso (POS, Part-of-speech) es el proceso de marcar una palabra en un texto con una parte particular del discurso, en función de su definición y contexto.
- ✓ Requieren un corpus marcado manualmente.
- ✓ Es una forma de análisis morfo-sintáctico.

Try out the model

```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text, token.pos_)
```

RUN

```
Veo al hombre con el telescopio.
Veo VERB
al ADP
hombre NOUN
con ADP
el DET
telescopio NOUN
. PUNCT
```

https://spacy.io/models/es

Comparación de herramientas

Herramienta	Código	Segmentación y tokenización	Etiquetado POS	Lematización
NLTK	Nativo Python	Si	Eagles	No
Freeling	API	Si	Eagles	Si
Pattern.es	Nativo Python	Si	Penn TreeBank	Si
Spacy	Nativo Python	Si	Penn TreeBank	Si
Stanford NLP	API	Si	Eagles	Si

Fuente: Talamé, L., Cardoso, A., & Amor, M. (2019). Comparación de herramientas de procesamiento de textos en español extraídos de una red social para Python. In XX Simposio Argentino de Inteligencia Artificial (ASAI 2019)-JAIIO 48 (Salta). Consultado en: http://170.210.201.137/pdfs/asai/ASAI-06.pdf

Segmentación

- Algunas palabras son indivisibles: que, no, ya, y…
- Pero en general las palabras tienen partes:
  - Raíz o raíces
  - Afijos
    - Flexivos: singular/plural, femenino/masculino, tiempos verbales.
    - Derivativos: prefijos, interfijos, sufijos

CANT-O
CANT-ABA-MOS
CANT-O-S

ABRE-LATA-S
CORRE-CAMINO-S

GAT-A
GAT-IT-O
GAT-IT-OS

AMOR
EN-AMOR-AR
DES-EN-AMOR-AR

CONSTITU-IR
CONSTITU-CIÓN
CONSTITU-CION-AL
CONSTITU-CION-AL-IZ-AR
CONSTITU-CION-AL-IZ-A-CIÓN

Segmentación

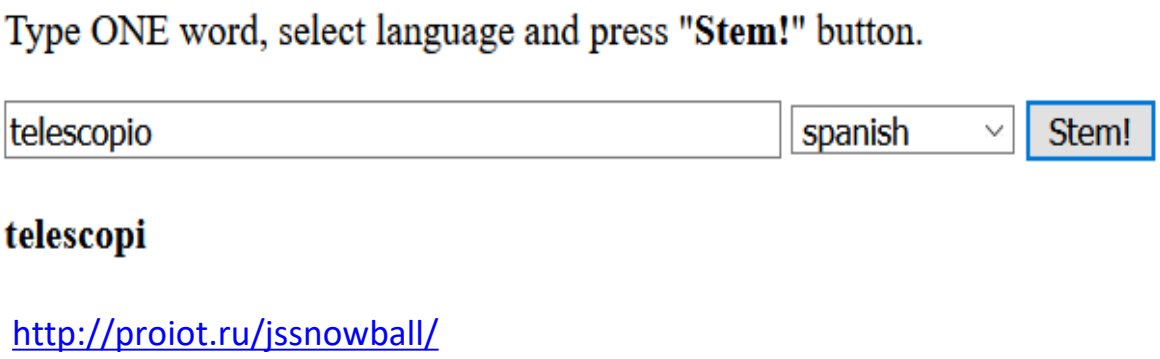
“A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS”

- ¿Cuántas palabras hay en este texto?
- ¿Cuántas palabras diferentes?

Hay 2 tareas diferentes que ayudan a buscar las palabras y sus raíces: ***stematización y lematización***.

Stemmer

- ✓ Los algoritmos de stemming intentan reducir las palabras flexionadas y derivadas en su forma raíz, es decir, extraer la raíz de una palabra, la raíz lingüística a la que pertenece. Generalmente son algoritmos muy rápidos.
- ✓ El stemming es una forma de análisis morfológico.
- ✓ Este proceso se realiza porque la raíz de una palabra puede aparecer más veces en un texto.



Stemmer

- ✓ El algoritmo más común para stemming es el algoritmo de Porter. Existen además métodos basados en análisis lexicográfico y otros algoritmos similares (KSTEM, stemming con cuerpo, métodos lingüísticos, entre otros).

```
from nltk import word_tokenize
from nltk.stem.snowball import SnowballStemmer

stemmer = SnowballStemmer("spanish")
doc = "A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS"
' '.join( [stemmer.stem(word) for word in word_tokenize(doc)] )
```

```
'a la gat le gust tra gat que tra a mas gat y a sus gatit'
```

Lematizador

- ✓ Lematización de los términos, es una parte del procesamiento lingüístico que trata de determinar el lema de cada palabra que aparece en un texto.
- ✓ Su objetivo es reducir una palabra a su raíz, de modo que las palabras clave de una consulta o documento se representen por sus raíces en lugar de por las palabras originales.
- ✓ El lema de una palabra comprende su forma básica, más sus formas declinadas.
- ✓ La lematización requiere etiquetado POS.

Lematizador

- ✓ La lematización requiere un diccionario como WordNet o Wikipedia.
- ✓ La lematización es un proceso computacionalmente costoso, en comparación con el stemming.

https://spacy.io/models/es

Try out the model

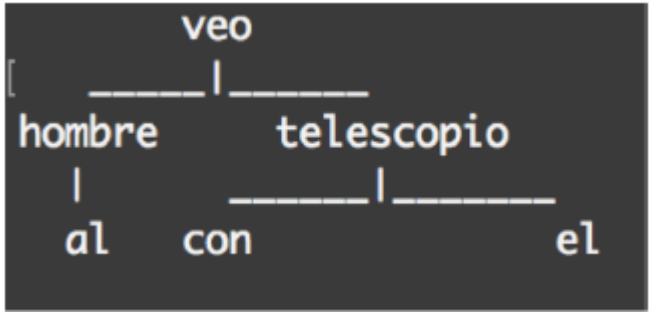
```
import spacy

nlp = spacy.load("es_core_news_sm")
doc = nlp("A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS")
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.lemma_)
```

RUN

```
A LA GATA LE GUSTA TRAER GATAS QUE TRAEN A MÁS GATOS Y A SUS GATITOS
A ADP a
LA DET el
GATA NOUN gata
LE PRON él
GUSTA ADP GUSTA
TRAER PROPN TRAER
GATAS NOUN gata
QUE PRON que
TRAEN PROPN TRAEN
A ADP A
MÁS ADV más
GATOS NOUN gatos
Y CCONJ Y
A ADP A
SUS PROPN SUS
GATITOS NOUN gatito
```

Analizador de dependencias

- ✓ El análisis de dependencia extrae un árbol sintáctico de un texto dado.
- ✓ Los árboles representan las relaciones sintácticas entre las palabras de una oración.
- ✓ El análisis utiliza un corpus marcado manualmente.
- ✓ El análisis requiere más recursos computacionales que los anteriores.



Try out the model

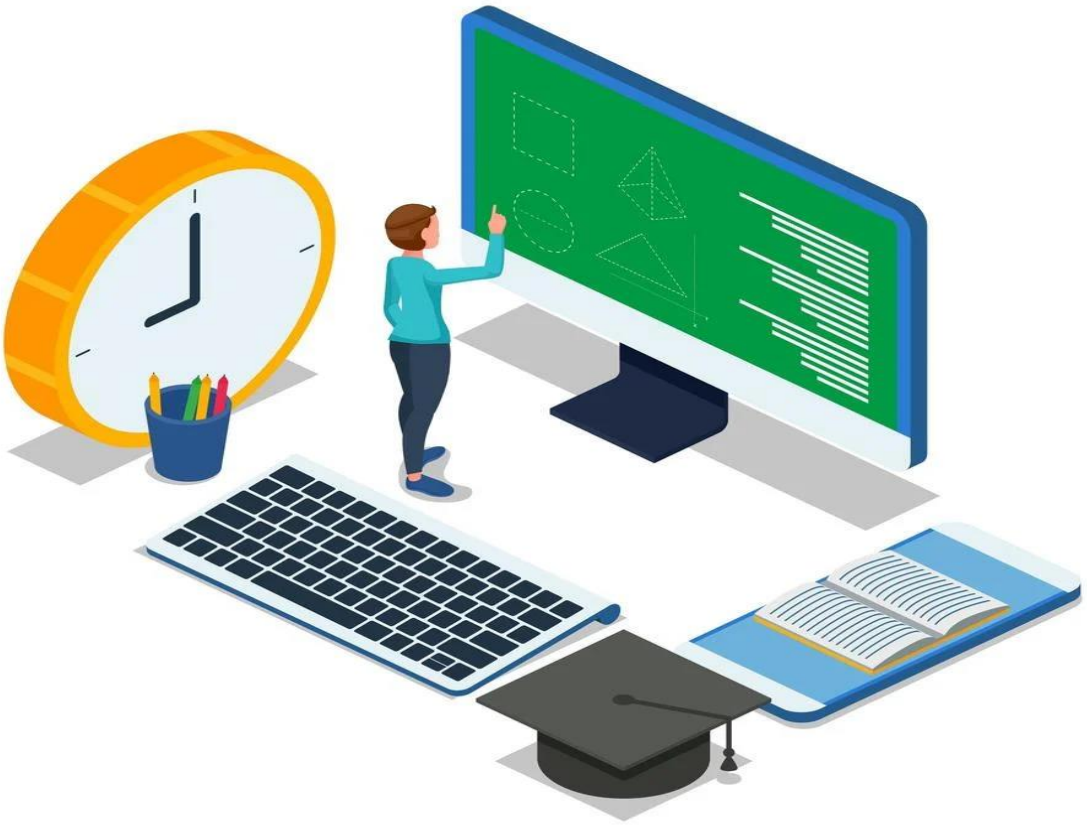
```
import spacy
from spacy.lang.es.examples import sentences

nlp = spacy.load("es_core_news_sm")
doc = nlp(sentences[5])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

RUN

```
Veo al hombre con el telescopio.
Veo VERB ROOT
al ADP case
hombre NOUN obj
con ADP case
el DET det
telescopio NOUN obl
. PUNCT punct
```

https://spacy.io/models/es

Práctica



Ejercicio3(es)-PLN.ipynb

Tarea

1. A partir del texto “El Ramo Azul”, de Octavio Paz, publicado en el libro español “Arenas movedizas” en 1949.

“El ramo azul” trata de un viajero que pasa una noche en un pueblo inquietante. La trama se centra en el diálogo que ocurre cuando un hombre débil se acerca al narrador para intentar sacarle los ojos. Lo que expone Paz, con los ojos como un símbolo, es que hay límites para la percepción.

A) ¿Cuántas palabras hay en el texto?

B) ¿Cuántas palabras diferentes existen?

C) ¿Qué cantidad de sustantivos, adjetivos y verbos posee el texto?

• Para el inciso C) puede tomarse en cuenta la ayuda de otras herramientas implementadas. Ejemplo: el servicio FreeLing: http://www.corpus.unam.mx/servicio-freeling/

Conclusiones

- El PLN es fácil de entender, posible y tiene gran importancia en nuestra época de información.

- La comprensión total del lenguaje natural es un objetivo aún distante. Pero existen herramientas y sistemas útiles para resolver varios problemas prácticos de PLN. El reto consiste en encontrar la correcta adecuación entre los diferentes tipos de problemas y las herramientas disponibles para resolverlos.

Referencias

- Applied Text Analysis with Python / by Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda : O'Reilly Media, Inc. [2018] 1 recurso en línea (xii, 334 páginas) : ilustraciones https://www.oreilly.com/library/view/applied-text-analysis/9781491963036/
- Natural language processing recipes : unlocking text data with machine learning and deep learning using Python / Akshay Kulkarni, Adarsha Shivananda -- [Berkeley, California] : Apress, [2019].-- xxv, 234 páginas : ilustraciones
- Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit 1st Edition / by Steven Bird, Ewan Klein, Edward Loper : O'Reilly Media, Inc. [2009] 1 recurso en línea (xi, 512 páginas) : ilustraciones https://itbook.store/books/9780596516499
- Vásquez, A. C., Quispe, J. P., & Huayna, A. M. (2009). Procesamiento de lenguaje natural. Revista de investigación de Sistemas e Informática, 6(2), 45-54.

Contacto

Luis Enrique Argota Vega

Máster en Ciencia e Ingeniería de la Computación

luiso91@gmx.com

Tels: 5578050838

Redes sociales:

LinkedIn https://cutt.ly/ifPyTEH

Google Scholar https://cutt.ly/WfPtYZz