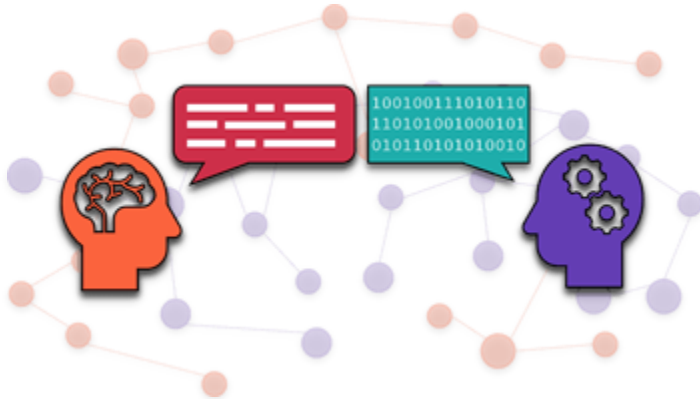
Modelos de Inteligencia Artificial

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

:: U3 ::

Procesamiento del Lenguaje Natural



2. Análisis de texto

Curso 2023-24

Tabla de contenidos

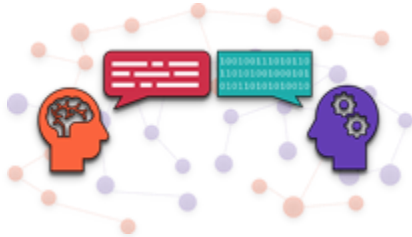
CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

1. Introducción

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- El **análisis de texto** permite:
  - identificar términos y patrones
  - descubrir dónde y con qué frecuencia tales términos se usan un texto
  - clasificar el tipo de términos que un texto contiene
  - ...

- Tanto las **expresiones regulares** como el **reconocimiento de entidades** (NER) son herramientas útiles para tal fin.



2. Expresiones regulares

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Una **expresión regular** es un patrón que puede reconocerse en un fragmento de texto
- Usos principales:
  - Buscar elementos particulares dentro de un texto extenso
  - Reemplazar términos o partes de ese texto por otros
  - Reformatear el texto
  - Validar entrada de datos
  - ...

2. Expresiones regulares

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo**:
  - formato de un email: **username@domain.extension**

2. Expresiones regulares

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejemplo**:

  1 2 3 4 5

  - formato de un email: **username@domain.extension**
    1. nombre usuario (combinación letras y números)
    2. + símbolo "**@**"
    3. + combinación letras y números
    4. + símbolo "**.**"
    5. + combinación letras y números

2. Expresiones regulares

- **Juego de símbolos**:
  - **“.”** → cualquier carácter o símbolo disponible
  - **“[]”** → un grupo de posibles caracteres
  - **“^”** → negación del carácter al que acompaña
  - **“\w”** → coincide con cualquier carácter (igual a [a-zA-Z0-9])
  - **“\d”** → coincide con cualquier dígito decimal (equivalente a [0-9])
  - ...

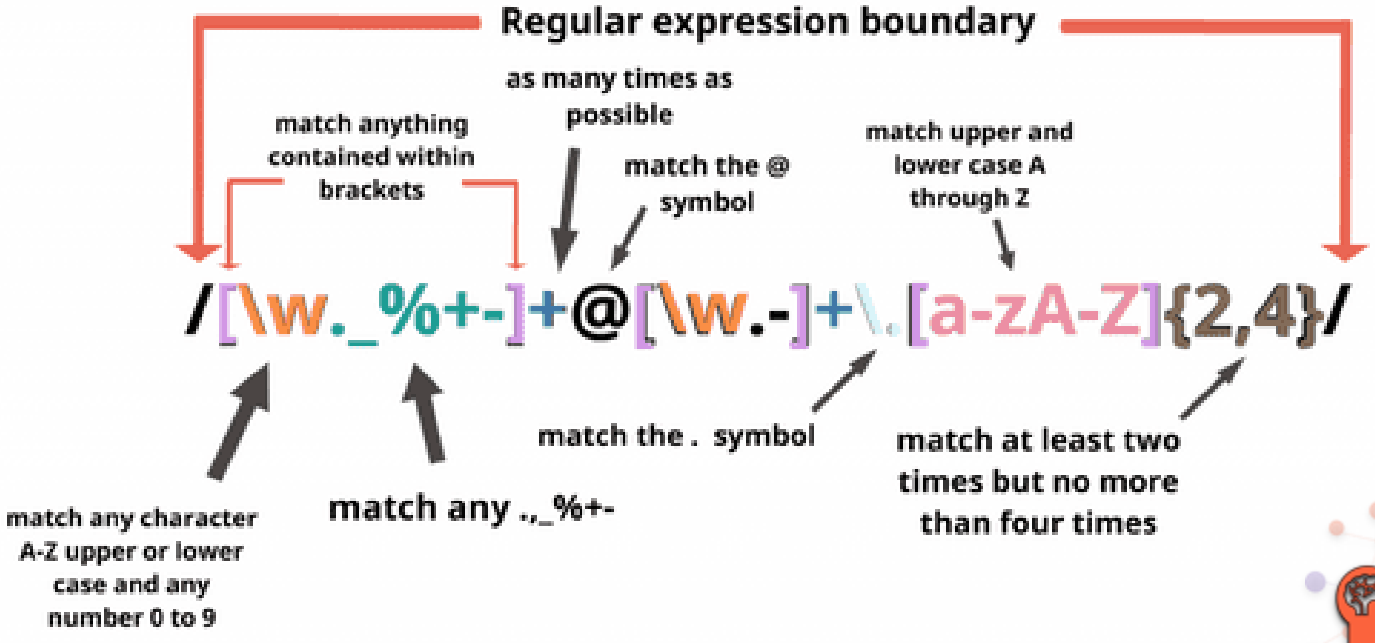
2. Expresiones regulares

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Cuantificadores**:
  - **“*”** → el carácter ocurre cero o más veces.
  - **“+”** → el carácter ocurre una o más veces.
  - **“?”** → el carácter ocurre cero o una vez.
  - **“{5}”** → el carácter ocurre cinco veces.
  - **“{3,7}”** → el carácter ocurre entre 3 y 7 veces.
  - **“{2,}”** → el carácter aparece al menos 2 veces.

2. Expresiones regulares

- **Ejemplo**:
  - formato de un email: **username@domain.extension**
  - Expresión regular: **/[\w._%+-]+@[\w.-]+\.[a-zA-Z]{2,4}/**



2. Expresiones regulares

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Ejercicio 1**:

  Construye las expresiones regulares que se proponen en el listado derecho de esta página: https://regexone.com/

3. **Operaciones** (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Python dispone de la librería **‘re’** para la definición y aplicación de expresiones regulares en tareas de análisis de texto.
- Métodos más útiles:
  - **compile**
  - **match**
  - **search**
  - **findall**

3. **Operaciones** (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Python dispone de la librería **‘re’** para la definición y aplicación de expresiones regulares en tareas de análisis de texto.
- Métodos más útiles:
  - **compile:** define un patrón (expresión regular)
  - **match:** busca un patrón al principio del texto
  - **search:** busca la primera ocurrencia del patrón en un texto
  - **findall:** busca todas las ocurrencias

3. Operaciones (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Interpretación y emparejamiento**:
  - uso los métodos **compile()** y **match()**

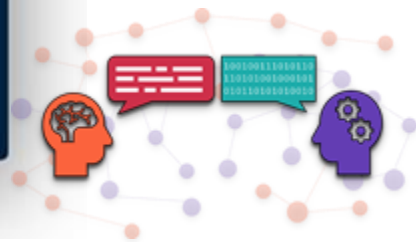
```python
import re
# dos nombres
persona_1 = "Luisa"
persona_2 = "Javier"

# se define un patrón
exp_reg = re.compile("[a-zA-Z]{5}")

# se rastrean posibles emparejamientos
resultado_1 = re.match(exp_reg, persona_1)
print(resultado_1)

# guarda e imprime los emparejamientos
emparejamiento_1 = resultado_1.group(0)
print(emparejamiento_1)

# compile a regular expression to match a 7 character string of word characters and check for a match
to character_2 here
emparejamiento_2 = re.match("[a-zA-Z]{6}", persona_2)
print(emparejamiento_2)

# <re.Match object; span=(0, 5), match='Luisa'>
# Luisa
# <re.Match object; span=(0, 6), match='Javier'>
```

3. **Operaciones** (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- **Búsqueda**:
  - con **search()** se busca alguna ocurrencia en el texto
  - con **findall()** se buscan varias ocurrencias en el texto

```python
# EJEMPLOS DE USO DE 'SEARCH' Y 'FINDALL'

import re
from google.colab import files

turing_text = "Alan Mathison Turing fue un matemático, lógico, informático teórico, criptógrafo,
filósofo y biólogo teórico. Además, fue corredor de ultradistancia británico."

# buscamos el término 'corredor' en todo el texto
corredor = re.search("corredor", turing_text)
res = corredor.group(0)
print(res)

# buscamos todas las ocurrencias de "fue" en el texto
todos_fue = re.findall("fue", turing_text)
print(todos_fue)

# store and print the length of all_lions here
num_fue = len(todos_fue)
print(num_fue)

# corredor
# ['fue', 'fue']
# 2
```



3. Operaciones (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

```python
import re

# Patrón con grupos de captura para buscar nombres y apellidos
pattern = re.compile(r'Nombre: (\w+) Apellido: (\w+)')

# Cadena de ejemplo
texto = "Nombre: John Apellido: Doe"

# Buscar la coincidencia en la cadena
result = pattern.search(texto)

# Imprimir el texto coincidente y las partes capturadas
if result:
    print("Texto coincidente completo:", result.group())
    print("Primer grupo de captura (Nombre):", result.group(1))
    print("Segundo grupo de captura (Apellido):", result.group(2))
else:
    print("No se encontró ninguna coincidencia.")

----

Texto coincidente completo: Nombre: John Apellido: Doe
Primer grupo de captura (Nombre): John
Segundo grupo de captura (Apellido): Doe
```

3. Operaciones (con ‘re’)

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Práctica 2.1

- Realiza la práctica “**PLN - P2.1 :: Procesamiento de información médica”**, cuyo enunciado encontrarás en el moodle

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Qué es NER?

- NER, que significa "**Named Entity Recognition**" en inglés, se refiere al proceso de **identificar** y **clasificar entidades** nombradas en un texto
- Las entidades nombradas son **sustantivos** que se refieren a personas, lugares, organizaciones, fechas, cantidades, valores monetarios y otros tipos de entidades con nombres específicos.

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

¿Dónde se usa NER?

- Análisis de redes sociales
- Análisis de documentos legales
- Servicios de atención al cliente (chatbots)
- Detección del fraude
- Clasificación temática de textos
- …

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

- Librerías como **NLTK**, **Stanza** o **SpaCy** permiten hacer el reconocimiento de términos como entidades.
  - En inglés todas ellas arrojan buenos resultados.
  - En español se recomienda el uso **SpaCy** o **Stanza**.

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Ejemplo en la librería **SpaCy** …

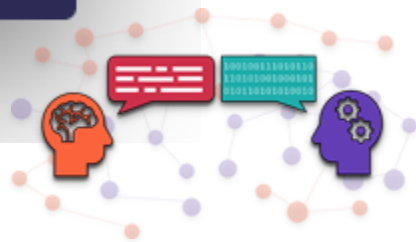
But Google ORG is starting from behind. The company made a late push into hardware, and Apple ORG 's Siri PRODUCT , available on iPhones PRODUCT , and Amazon ORG 's Alexa PRODUCT software, which runs on its Echo PRODUCT and Dot PRODUCT devices, have clear leads in consumer adoption.

PERSON: People, including fictional.
NORP: Nationalities or religious or political groups.
FAC: Buildings, airports, highways, bridges, etc.
ORG: Companies, agencies, institutions, etc.
GPE: Countries, cities, states.
LOC: Non-GPE locations, mountain ranges, bodies of water.
PRODUCT: Objects, vehicles, foods, etc. (Not services.)
EVENT: Named hurricanes, battles, wars, sports events, etc.
WORK_OF_ART: Titles of books, songs, etc.
LAW: Named documents made into laws.
LANGUAGE: Any named language.
DATE: Absolute or relative dates or periods.
TIME: Times smaller than a day.
PERCENT: Percentage, including "%".
MONEY: Monetary values, including unit.
QUANTITY: Measurements, as of weight or distance.
ORDINAL: "first", "second", etc.
CARDINAL: Numerals that do not fall under another type.

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

```python
import spacy

# Cargar el modelo preentrenado de spaCy para NER en español
nlp = spacy.load("es_core_news_sm")

# Texto de ejemplo
texto = "El presidente de Estados Unidos, Joe Biden, \
se reunió con la canciller alemana Angela Merkel en Berlín."

# Procesar el texto con spaCy
doc = nlp(texto)

# Iterar sobre las entidades reconocidas e imprimir la etiqueta
# y el texto de cada entidad
for entidad in doc.ents:
    print(f"Entidad: {entidad.text:<15} Etiqueta: {entidad.label_}")
```



4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

```
Entidad: Estados Unidos  Etiqueta: LOC
Entidad: Joe Biden       Etiqueta: PER
Entidad: Angela Merkel   Etiqueta: PER
Entidad: Berlín          Etiqueta: LOC
```

4. Reconocimiento de entidades

CIPFP Mislata
Centre Integrat Públic
Formació Professional Superior

Práctica 2.2

- Realiza la práctica "**PLN - P2.2 :: Agenda personal con recordatorios"**, cuyo enunciado encontrarás en el moodle