uc3m | Universidad Carlos III de Madrid

Máster en Ciencia Tecnología Informática

2 - Tareas Básicas PLN: NLTK y Spacy

Tecnologías Emergentes en la Sociedad de la Información

AUTOR: ISABEL SEGURA BEDMAR

NLTK

- Es una librería de Python para realizar tareas de PLN.
  - División de oraciones y tokenización.
  - Análisis Morfosintáctico.
  - Lematización y stemming.
  - Reconocimiento de Entidades
  - Expresiones regulares.

Spacy

- Es una librería de Python para realizar tareas de PLN.
  - Tokenización
  - Análisis Morfosintáctico
  - Reconocimiento de Entidades
  - Análisis de Dependencias
  - Similitud Semántica
  - Word Embeddings
- https://nlpforhackers.io/complete-guide-to-spacy/

uc3m | Universidad Carlos III de Madrid

Máster en Ciencia Tecnología Informática

2 - Tareas Básicas PLN: Expresiones Regulares

Tecnologías Emergentes en la Sociedad de la Información

AUTOR: ISABEL SEGURA BEDMAR

Expresiones Regulares

- Son patrones que se utilizan para buscar determinadas combinaciones de caracteres en un texto.
- Ej: Encontrar todas las posibles variaciones de la palabra niño (niño, niña, niños, niñas, Niños, Niñas) en un texto.

Expresiones Regulares

- [] el patrón mapea cualquiera de los caracteres contenidos entre corchetes:

Patrón	Qué detecta
[012345]	Cualquier dígito entre 0 y 5
[Aa]	La letra A o a
[nN]iñ[oa]	niño,niña,Niño,Niña

https://www.regexpal.com/

Expresiones Regulares

- Caracteres dentro de [] significa que el patrón detectará cualquier texto que contenga algunos de los caracteres entre los corchetes:

Patrón	Qué detecta
[012345]	Cualquier dígito entre 0 y 5
[Aa]	La letra A o a
[nN]iñ[oa]	niño,niña,Niño,Niña

https://www.regexpal.com/

Expresiones Regulares

- Rangos [A-Z]

https://www.regexpal.com/

Patrón	Qué detecta
[0-9]	Cualquier dígito entre 0 y 9
[A-Z]	Cualquier letra mayúscula
[a-z]	Cualquier letra minúscula
[A-Za-z]	Cualquier letra mayúscula o minúscula

Expresiones Regulares

- | : significa o

Patrón	Qué detecta
-|/	- o /
a|b|c = [abc]	letra a b o c

https://www.regexpal.com/

Expresiones Regulares

- | : significa o

Patrón	Qué detecta
-|/	- o /
a|b|c = [abc]	letra a b o c

https://www.regexpal.com/

Expresiones Regulares

Patrón	Significado	Ejemplo
?	el carácter anterior es opcional	*colou?r -> color, colour*
*	0 o más ocurrencias	*ja* -> j, ja, jaa, jaa, etc*
+	1 o más ocurrencias	*ja+ -> j, ja, jaa, jaa, etc*
.	cualquier carácter	*l.s -> los, las, l-s, ..*

https://www.regexpal.com/

Expresiones Regulares

- ^representa el principio de la línea.
- $ el final de línea
- \ permite representar caracteres especiales como ., *, +

Patrón	Significado
^[0-9]	Identifica las líneas que empiezan con un dígito
;$	identifica el texto que termina con un ;
\.$	indentifica el texto que termina con un .

https://www.regexpal.com/

Expresiones Regulares

Patrón	Significado
\w	mapea cualquier letra, digito o underscore
\W	cualquier carácter que no sea \w
\d	cualquier dígito
\s	un espacio en blanco
\t	tabular
\n	salto de línea

https://www.regexpal.com/

Ejemplos

- El documento nacional (cédula) de Venezuela tiene el siguiente formato: la letra V (para indicar que es venezolano) o la letra E, un guión, seguido de un número de 8 dígitos.
- Escribe el patrón que permita identificar este tipo de expresiones.

Ejemplo

Regular Expression

Javascript flags

```
/[VE]-[\d]{8}/g
```

2 matches

Test String

```
En Venezuela, al documento nacional de identidad (DNI) se le llama cédula de identidad o simplemente cédula, y generalmente consiste en: una letra (V o E) que indica si se es venezolano o extranjero y un número entre 1 y 99 999 999, separados por un guión, como por ejemplo: V-15000123 ó E-82055055. Sin embargo, el patrón no va a identificar un NIF español como el 08927350-S.
```

Ejemplo

Escribe la expresión resultar para identificar DNI españoles.

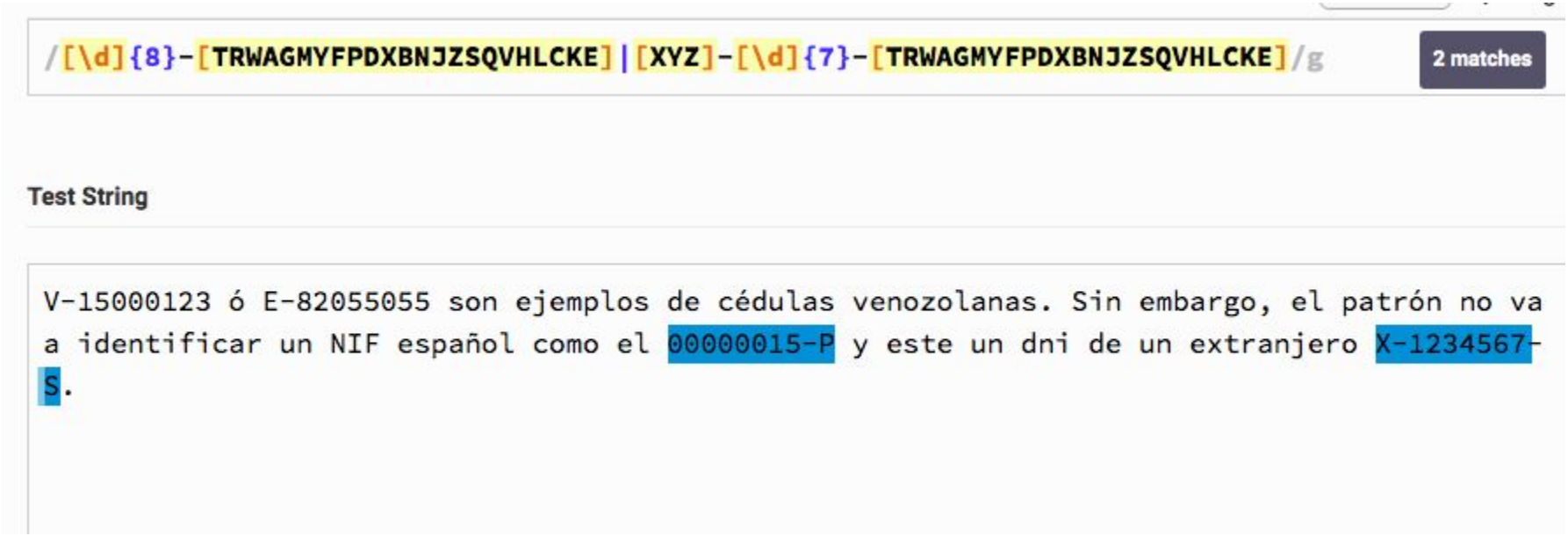
Estos pueden ser:

- NIF (Número de Identificación Fiscal) - 8 numeros y una letra[1]
- NIE (Número de Identificación de Extranjeros) - 1 letra[2], 7 numeros y 1 letra[1]

1 - Una de las siguientes: TRWAGMYFPDXBNJZSQVHLCKE

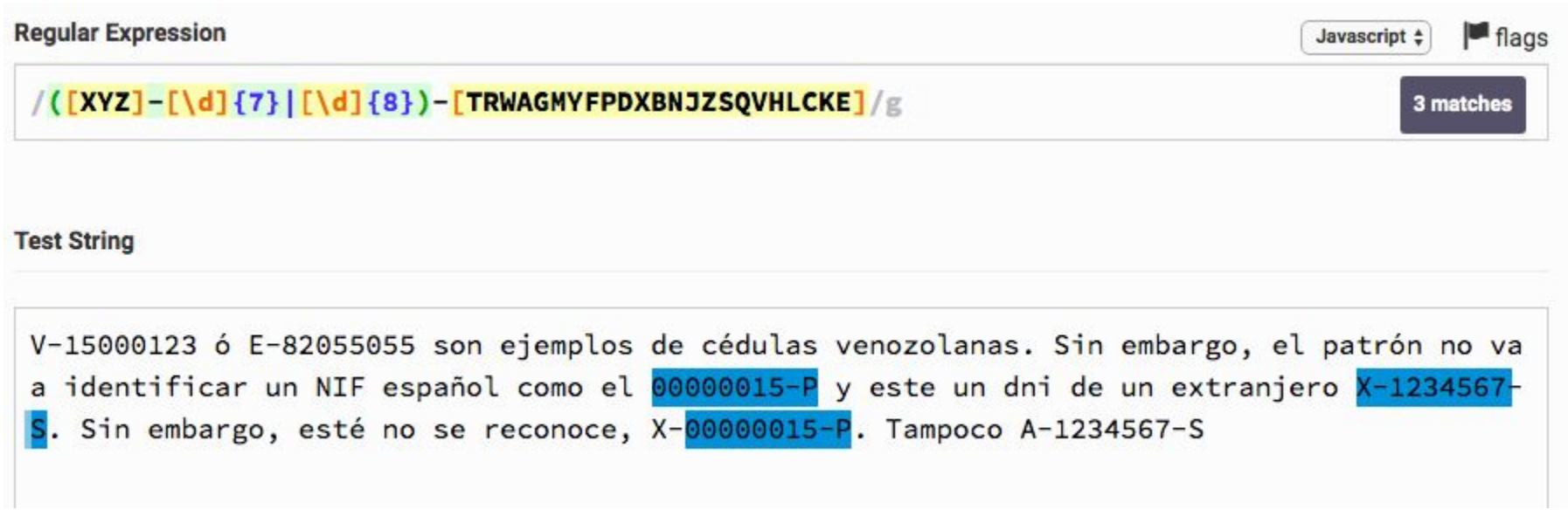
2 - Una de las siguientes: XYZ

Ejemplo



[\d]{8}-[TRWAGMYFPDXBNJZSQVHLCKE]|[XYZ]-[\d]{7}-[TRWAGMYFPDXBNJZSQVHLCKE]

Ejemplo



([XYZ]-[\d]{7}|[\d]{8})-[TRWAGMYFPDXBNJZSQVHLCKE]

Conclusiones

- Expresiones regulares suelen utilizarse como enfoques iniciales para abordar tareas de PLN.
- También pueden ser utilizadas como features (características) dentro de los algoritmos de aprendizaje automático (por ejemplo, si un texto contiene una determinada expresión regular o no).