Preparación de Datos

experto en procesamiento del lenguaje natural

Universitat d'Alacant
Universidad de Alicante
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
UJI UNIVERSITAT JAUME I
valgrAI

Contenidos

- Ciclo de Vida Proyecto Aprendizaje Automático
- Obtención de Datos
- Formatos de Archivo
- Codificación de Caracteres
- Preprocesamiento
- Ecosistema de Software

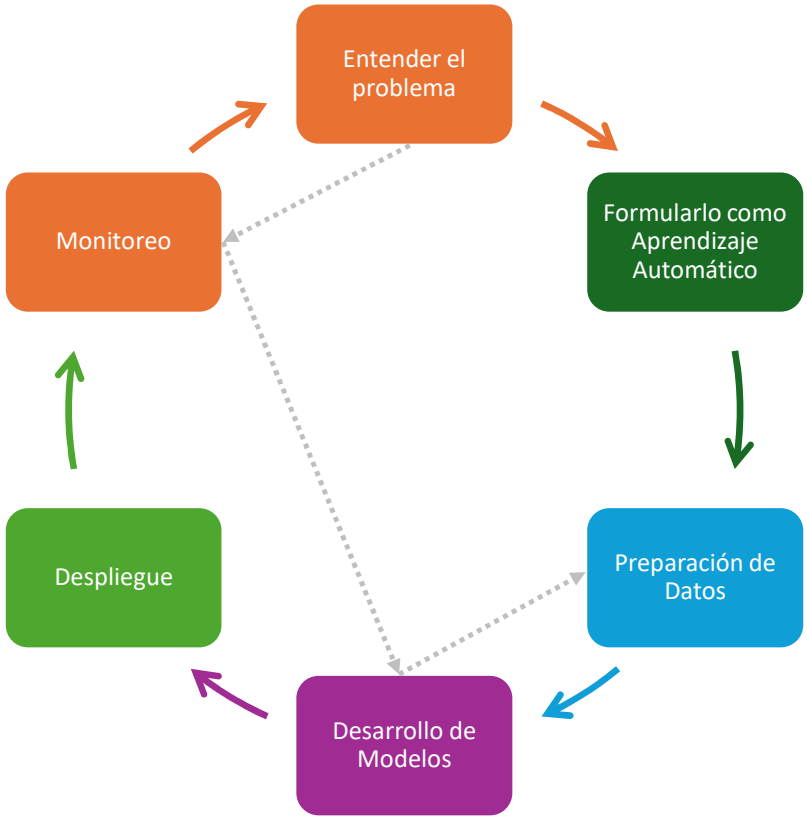
experto en procesamiento del lenguaje natural

Ciclo de Vida Proyecto Aprendizaje Automático

Contextualizando la Preparación de Datos

experto en procesamiento del lenguaje natural

Proyectos Aprendizaje Automático. Ciclo de Vida



experto en procesamiento del lenguaje natural

Entender el problema

- clarificar los objetivos, qué se quiere predecir o clasificar.

- verificar que el problema es apropiado para abordarlo mediante aprendizaje automático. ¿Existe una solución algorítmica? ¿Qué tan efectiva es?

experto en procesamiento del lenguaje natural

Entender el problema

Escenario A

En un fórum de discusión online se recomienda no compartir datos personales como el teléfono o el correo electrónico.

Dado un post o comentario, se desea identificar automáticamente si el usuario ha incluido alguno de estos elementos de modo que se emita una alerta para revisar el contenido.

experto en procesamiento del lenguaje natural

Entender el problema

Escenario B

Para construir un sistema de generación de resúmenes, una etapa puede ser identificar frases que expresan la misma información con diferentes palabras. Eliminando alguna de ellas, se puede condensar el contenido original sin perder su significado.

Se desea implementar una función, que, dadas dos frases, indique su nivel de similitud semántica

experto en procesamiento del lenguaje natural

Entender el problema

¿recomendaría utilizar aprendizaje automático para resolver el problema descritos en el escenario A?

¿y en el escenario B?

experto en procesamiento del lenguaje natural

Formular el Problema como Aprendizaje Automático

Determinar:

- objetivos del modelo (problema que resolverá el modelo desde el punto de vista de aprendizaje automático).
- salida del modelo (lo que se quiere predecir, variables dependientes).
- entradas del modelo (las que determinan la predicción, variables independientes).
- seleccionar métricas de desempeño o error que se optimizarán.

experto en procesamiento del lenguaje natural

Formular el Problema como Aprendizaje Automático

frase 1	frase 2	similaridad
the activity of learning or being trained	the gradual process of acquiring knowledge.	4
thin metal strips used in typesetting	thin strip of metal used to separate lines of type in printing.	3.8
falcon: abandon intended game and pursue lesser prey	abandon the intended prey, turn, and pursue an inferior prey.	4.2
the act of designating a role to someone	the act of designating or identifying something.	1.8
take in (mentally or physically)	ke (gas, light or heat) into a solution.	0.8
protect or speak up against criticism or opposition	**protect or fight for as a champion.**	**¿?**

experto en procesamiento del lenguaje natural

Formular el Problema como Aprendizaje Automático

entrada		salida
frase 1	frase 2	similaridad
the activity of learning or being trained	the gradual process of acquiring knowledge.	4
thin metal strips used in typesetting	thin strip of metal used to separate lines of type in printing.	3.8
falcon: abandon intended game and pursue lesser prey	abandon the intended prey, turn, and pursue an inferior prey.	4.2
the act of designating a role to someone	the act of designating or identifying something.	1.8
take in (mentally or physically)	ke (gas, light or heat) into a solution.	0.8
protect or speak up against criticism or opposition	**protect or fight for as a champion.**	**¿?**

experto en procesamiento del lenguaje natural

Entender el problema

Escenario B

Para construir un sistema de generación de resúmenes, una etapa puede ser identificar frases que expresan la misma información con diferentes palabras. Eliminando alguna de ellas, se puede condensar el contenido original sin perder su significado.

Se desea implementar una función, que, dadas dos frases, indique su nivel de similitud semántica

Objetivo: Estimar la similitud semántica entre dos textos en una escala de 0 a 5.

Salida ideal: similitud semántica, valor numérico.

Entrada: par de frases para calcular su similitud.

Métrica: error cuadrático medio entre la similitud calculada por el sistema y la esperada.

experto en procesamiento del lenguaje natural

Preparación de Datos

Entre otras tareas:

- obtener datos relevantes al problema.
- definir estrategia de muestreo.
- crear conjuntos para diferentes tareas (ej. entrenamiento, prueba, validación)
- análisis exploratorio.
- definir y evaluar métricas de calidad de los datos.
- preprocesamiento para normalizar datos, eliminar ruido, etc.
- ingeniería de características (feature engineering): crear, extraer, seleccionar entradas del modelo.

experto en procesamiento del lenguaje natural

Desarrollo de Modelos

Entre otras tareas:

- seleccionar modelos, considerar características de los datos.
- entrenar modelos.
- validar modelos.

- decidir es necesario volver a una etapa previa, ej. obtener más datos.

experto en procesamiento del lenguaje natural

Despliegue

Entre otras tareas:

- preparar modelos para producción (un prototipo lo antes posible).
- estimar recursos necesarios (CPU, RAM, HDD, GPU, etc.)
- seleccionar entorno de despliegue (AWS, Azure, Google Cloud, etc.).
- definir e implementar estrategia de integración continua y despliegue (CI/CD).

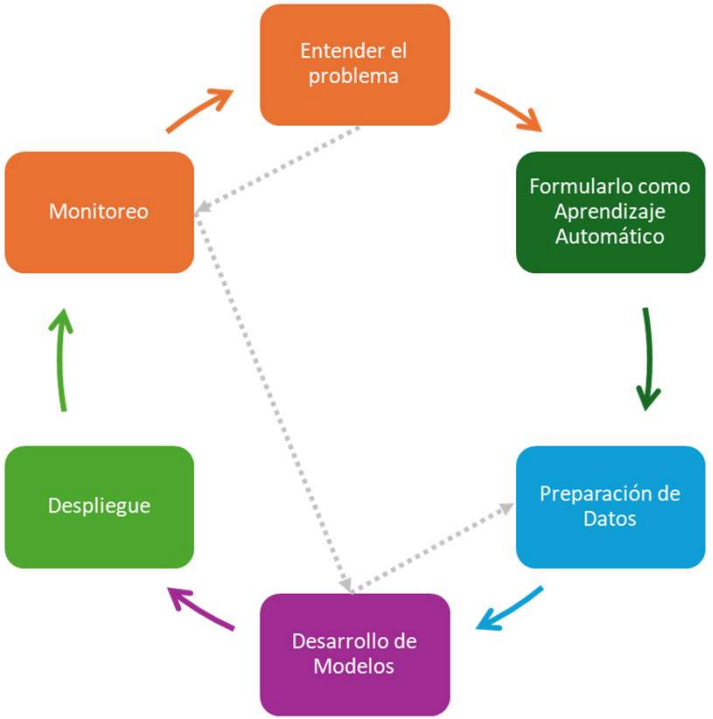
experto en procesamiento del lenguaje natural

Monitoreo

Continuar evaluando:

- métricas de desempeño del modelo.
- métricas de rendimiento de hardware o software.

Realizar "mantenimiento". En casos extremos puede requerir verificar nuevamente todo el ciclo de vida del proyecto.



experto en procesamiento del lenguaje natural

Obteniendo Datos

Consideraciones

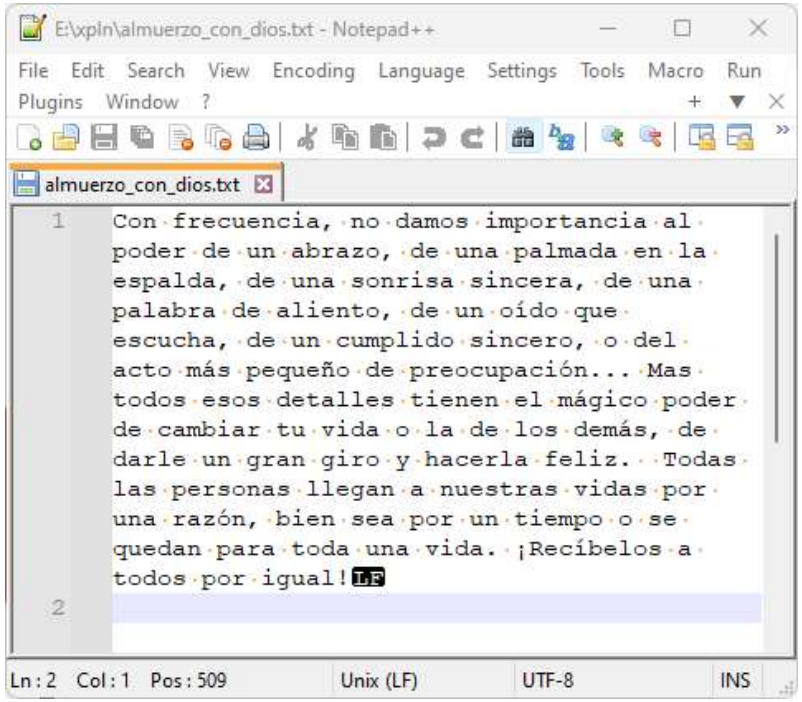
experto en procesamiento del lenguaje natural

Formatos de Archivo (algunos de los más frecuentes)

- Archivo de texto plano, sin información de formato como color de letras, uso de cursiva, etc.

- Fáciles de leer y editar.

- Difícil estructurar contenido.



experto en procesamiento del lenguaje natural

Formatos de Archivo

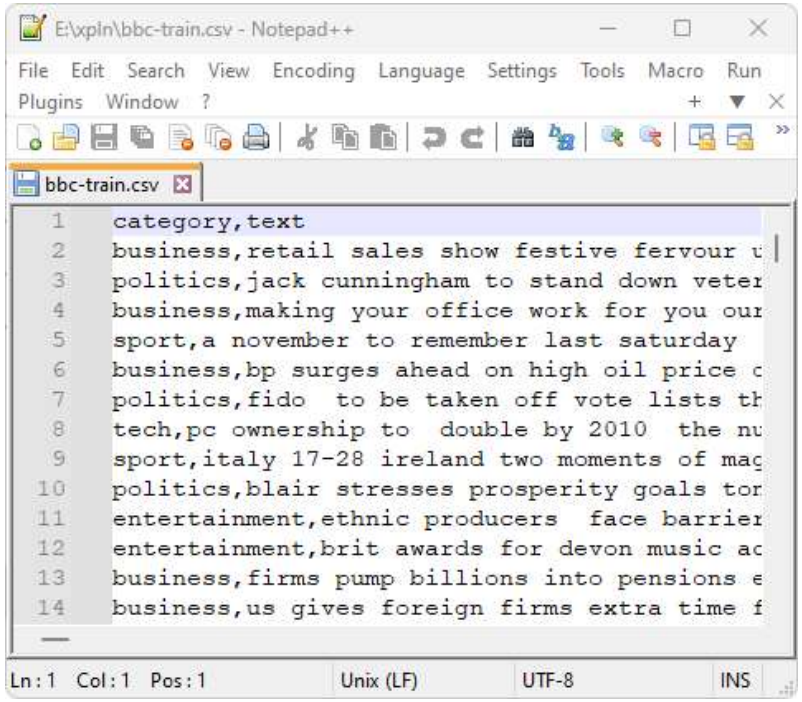
- Formato tabular basado en texto.

- CSV = comma-separated values.

- Fáciles de leer y editar.

- No distingue tipo de datos en cada columna.

- No escalan bie cuando la cantidad de datos aumenta.



experto en procesamiento del lenguaje natural

Formatos de Archivo

- Formato de datos anidado basado en texto.

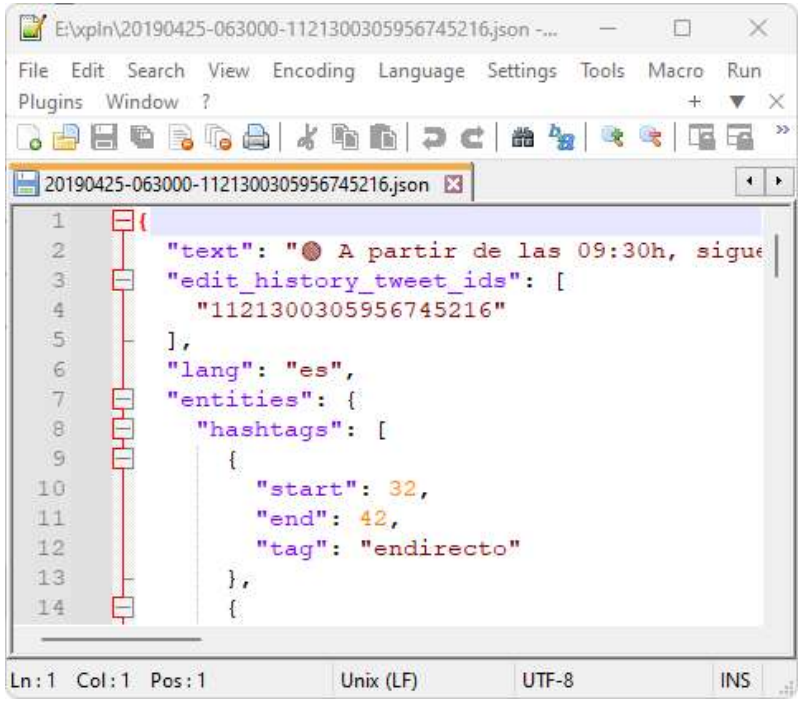
- JSON = **J**ava**S**crip **O**bject **N**otation.

- Compuesto por pares clave-valor. Las claves son cadenas. Los valores pueden ser cadenas, números, arreglos u objetos

- Los elementos se separan por “,”

- [] denotan un arreglo

- {} denotan un objeto



experto en procesamiento del lenguaje natural

Formatos de Archivo

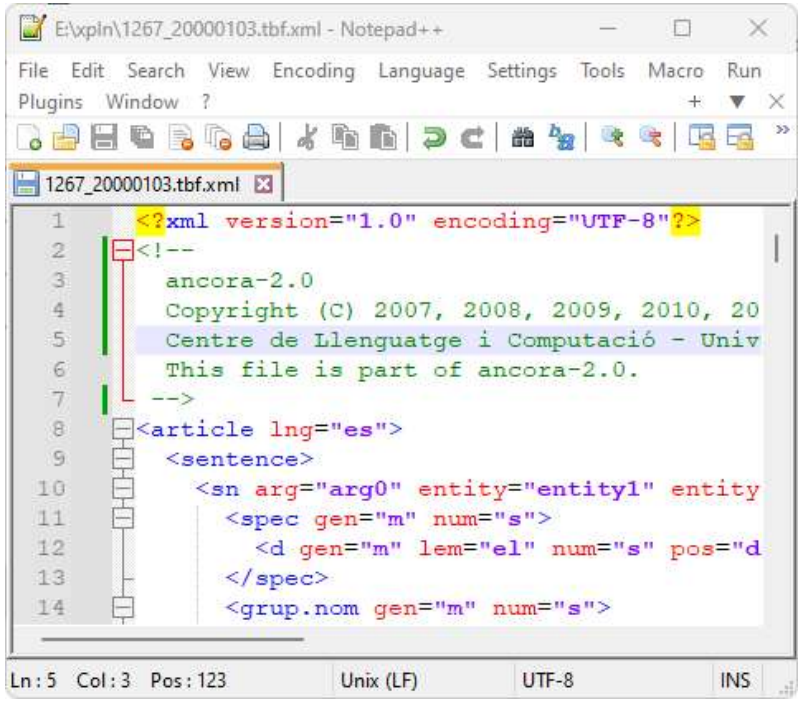
- Formato de datos anidado basado en texto.

- XML = e**X**tensible **M**arkup **L**anguage

- Los documentos siguen una estructura jerárquica.

- Compuesto elementos, que pueden contener texto, otros elementos o estar vacíos.

- Los elementos se identifican por etiquetas de comienzo y fin, ej. <article> </article>



experto en procesamiento del lenguaje natural

Codificación de Caracteres. ASCII

- ASCII = **A**merican **E**stándar **C**ode for **I**nformation **I**nterchange
- 7 bits por símbolo = 128 símbolos

Ejemplo,

b_7	b_6	b_5	b_4	b_3	b_2	b_1
1	0	0	0	0	0	0

1000000 binario = 64 decimal = @

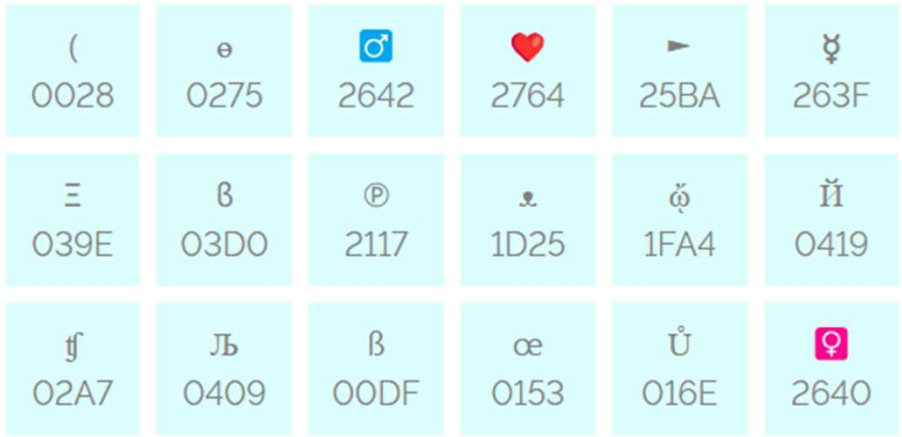
Bits b_4	b_3	b_2	b_1	Row \ Column	0 (000)	1 (001)	2 (010)	3 (011)	4 (100)	5 (101)	6 (110)	7 (111)
0	0	0	0	0	NUL	DLE	SP	0	@	P	`	p
0	0	0	1	1	SOH	DC1	!	1	A	Q	a	q
0	0	1	0	2	STX	DC2	"	2	B	R	b	r
0	0	1	1	3	ETX	DC3	#	3	C	S	c	s
0	1	0	0	4	EOT	DC4	$	4	D	T	d	t
0	1	0	1	5	ENQ	NAK	%	5	E	U	e	u
0	1	1	0	6	ACK	SYN	&	6	F	V	f	v
0	1	1	1	7	BEL	ETB	´	7	G	W	g	w
1	0	0	0	8	BS	CAN	(	8	H	X	h	x
1	0	0	1	9	HT	EM	)	9	I	Y	i	y
1	0	1	0	10	LF	SUB	*	:	J	Z	j	z
1	0	1	1	11	VT	ESC	+	;	K	[	k	{
1	1	0	0	12	FF	FS	,	<	L	\	l	¦
1	1	0	1	13	CR	GS	—	=	M	]	m	}
1	1	1	0	14	SO	RS	.	>	N	^	n	~
1	1	1	1	15	SI	US	/	?	O	_	o	DEL

¿posibles problemas?

experto en procesamiento del lenguaje natural

Codificación de Caracteres. Unicode

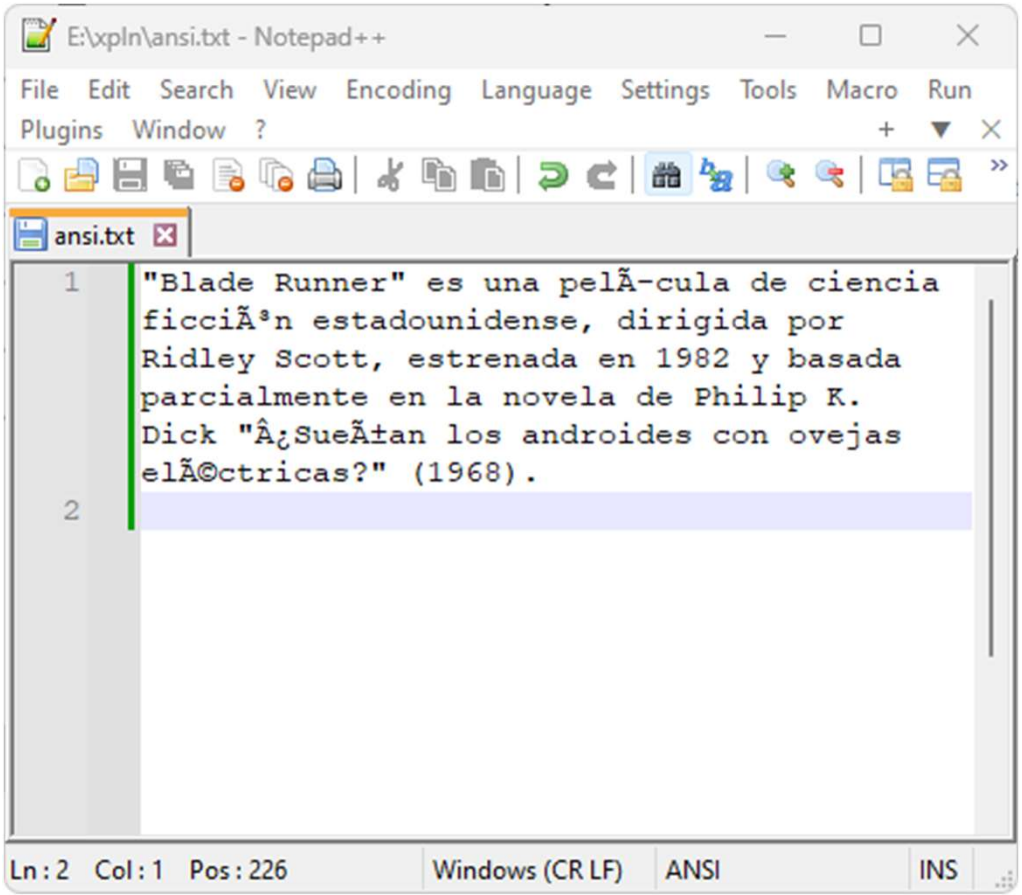
- Asigna un único número a cada carácter, independiente de la aplicación, la plataforma o el idioma.
- Desarrollado por Unicode Consortium
- Más de 150 000 caracteres.



experto en procesamiento del lenguaje natural

Codificación de Caracteres

¿por qué tenerlo en cuenta?



Preprocesamiento

Consideraciones

experto en procesamiento del lenguaje natural

Preprocesamiento de Datos

Tokenización

- Divide el texto en componentes pequeños – palabras, sub-palabras o caracteres – que permiten un análisis más refinado. Permite pasar del espacio del texto a espacios numérico.

Eliminar URLs, números, signos de puntuación, emoticonos, etc.

- Hacer el texto más simple eliminando elementos irrelevantes el problema o que puede ocasionar ruido.

Eliminar Stop-words

- Palabras que aparecen con frecuencia y que no aportan mucho significado.

experto en procesamiento del lenguaje natural

Preprocesamiento de Datos

Stemming o Lematización

- Normaliza el texto. Reduce la variabilidad y el vocabulario. Permite eliminar información que quizás no es relevante.

Normalización

- Mejora la consistencia y facilita el análisis, ej. no diferenciando entre mayúsculas y minúsculas, homogenizando formato de fechas, números, corrigiendo errores ortográficos, etc.

experto en procesamiento del lenguaje natural

Ecosistema de Software

(no exhaustivo)

experto en procesamiento del lenguaje natural

Herramientas

spaCy

- Es una biblioteca de procesamiento del lenguaje natural (NLP, por sus siglas en inglés) de código abierto desarrollada principalmente en Python. Es utilizada para realizar tareas de procesamiento de texto, como tokenización, etiquetado de partes del discurso, reconocimiento de entidades nombradas, análisis de dependencias sintácticas y más.

- Incluye modelos pre-entrenados para varios idiomas, lo que facilita su uso en diferentes contextos lingüísticos. Además, ofrece la capacidad de entrenar modelos personalizados para adaptarse a necesidades específicas.

experto en procesamiento del lenguaje natural

Herramientas

- Es una herramienta de Procesamiento del Lenguaje Natural en Python.
- Provee interfaces a más de 50 corpus y recursos léxicos como WordNet.
- Orientado a público variado como lingüistas computaciones, ingenieros, educadores, etc.

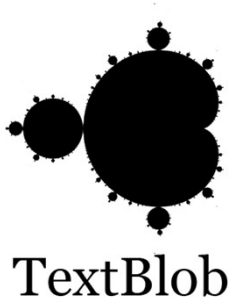
experto en procesamiento del lenguaje natural

Herramientas

- Herramienta de código abierto para la representación de documentos como vectores semánticos.
- Implementa una extensa colección de algoritmos o recursos como word2vect, FastText, LDA, etc.
- Prioriza la facilidad de uso y el rendimiento.

experto en procesamiento del lenguaje natural

Herramientas



- Librería Python para el procesamiento de textos.
- Implementa tareas comunes en NLP como:
  - etiquetado PoS.
  - tokenización.
  - cálculo de frecuencias.
  - Inflexión de palabras.
  - Lematización.
  - análisis de sentimientos.
  - etc.

experto en procesamiento del lenguaje natural

Herramientas

- Librería Java para el procesamiento de textos.
- Implementa tareas comunes en NLP como:
  - etiquetado PoS.
  - lematización.
  - reconocimiento de entidades.
  - correferencia.
  - etc.
- 8 idiomas
- Incluye interfaces de acceso a CoreNLP.

experto en procesamiento del lenguaje natural

Herramientas

- Librería Python para el procesamiento de textos.
- Implementa tareas comunes en NLP como:
  - etiquetado PoS.
  - lematización.
  - reconocimiento de entidades.
  - etc.
- 70 idiomas
- Incluye interfaces de acceso a CoreNLP.

experto en procesamiento del lenguaje natural

Preparación de Datos

Fin

experto en procesamiento del lenguaje natural