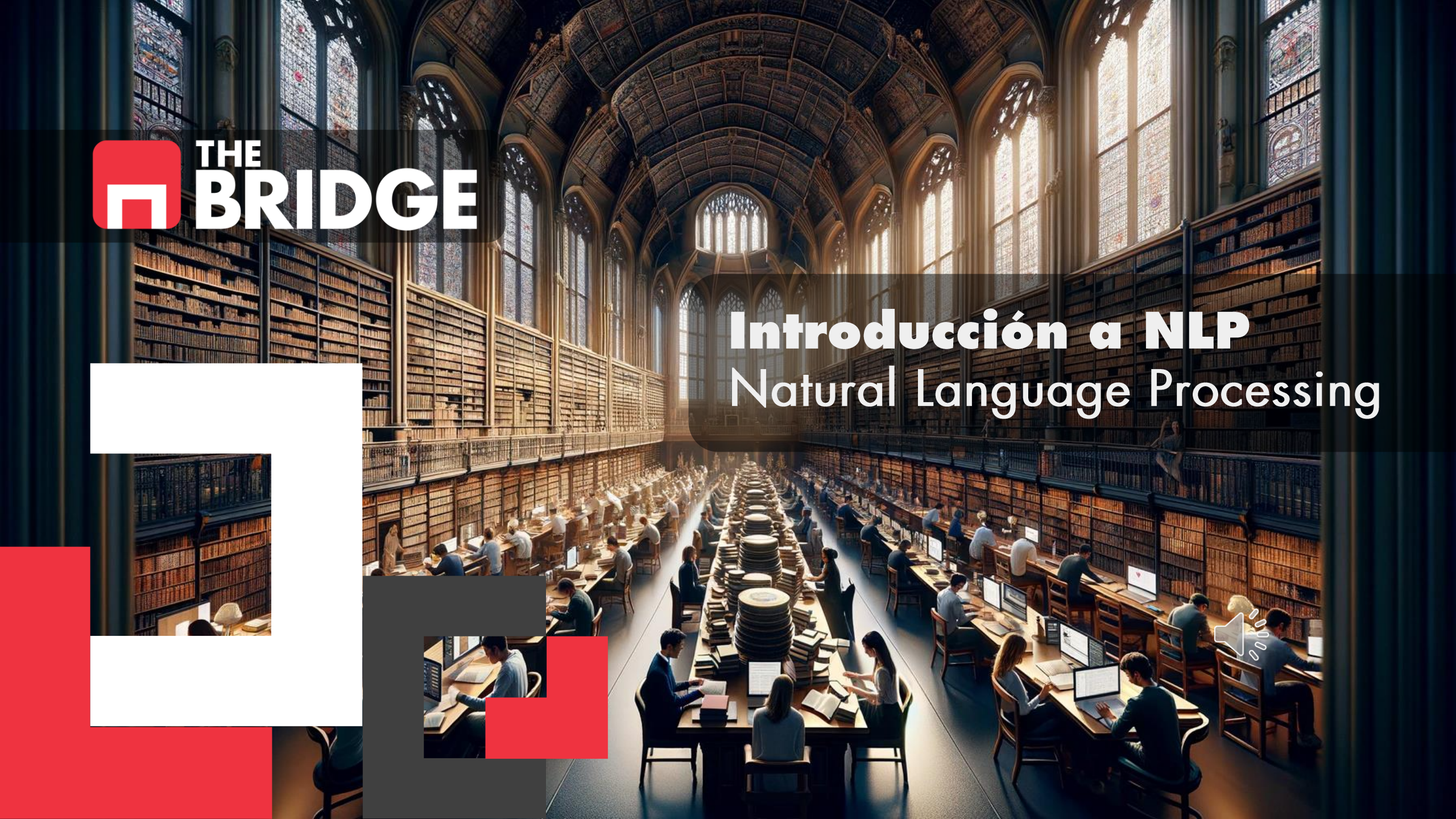
THE BRIDGE

Introducción a NLP

Natural Language Processing



¿Qué es NLP?

Procesamiento de lenguaje (texto, audio) para extraer en features que puedan utilizarse en modelado (machine learning o no) y/o para la resolución de determinados problemas y tareas que tengan ese lenguaje como dato de entrada

NLP como "extracción"/"preprocesamiento"

Hasta ahora hemos manejado datos tabulares con dos tipos de datos básicamente (numéricos y categóricos/etiquetas)

	housing_median_age	total_rooms	total_bedrooms	population	households	median_house_value	ocean_proximity	income_cat
0	41.0	880.0	129.0	322.0	126.0	452600.0	NEAR BAY	5
1	21.0	7099.0	1106.0	2401.0	1138.0	358500.0	NEAR BAY	5
2	52.0	1467.0	190.0	496.0	177.0	352100.0	NEAR BAY	5
3	52.0	1274.0	235.0	558.0	219.0	341300.0	NEAR BAY	4
4	52.0	1627.0	280.0	565.0	259.0	342200.0	NEAR BAY	3

NLP como "extracción"/"preprocesamiento"

… ahora vamos a ver técnicas para poder tratar features que son "textuales" y en general casos en los que la feature única o principal es un texto (el contenido de un tweet, una review, un capítulo de un libro, etc)

CASO 1: Utilizar las features marcadas

urlDrugName	rating	effectiveness	sideEffects	condition	benefitsReview	sideEffectsReview
enalapril	4	Highly Effective	Mild Side Effects	management of congestive heart failure	slowed the progression of left ventricular dys...	cough, hypotension , proteinuria, impotence , ...
ortho-tri-cyclen	1	Highly Effective	Severe Side Effects	birth prevention	Although this type of birth control has more c...	Heavy Cycle, Cramps, Hot Flashes, Fatigue, Lon...
ponstel	10	Highly Effective	No Side Effects	menstrual cramps	I was used to having cramps so badly that they...	Heavier bleeding and clotting than normal.
prilosec	3	Marginally Effective	Mild Side Effects	acid reflux	The acid reflux went away for a few months aft...	Constipation, dry mouth and some mild dizzines...
lyrica	2	Marginally Effective	Severe Side Effects	fibromyalgia	I think that the Lyrica was starting to help w...	I felt extremely drugged and dopey. Could not...

NLP como "extracción"/"preprocesamiento"

… ahora vamos a ver técnicas para poder tratar features que son "textuales" y en general casos en los que la feature única o principal es un texto (el contenido de un tweet, una review, un capítulo de un libro, etc)

CASO 2: El "texto" es la FEATURE

	User	Content	Date	Lang
0	ccifuentes	Salgo de #VeoTV , que día más largoooooo...	2011-12-02T00:47:55	es
1	CarmendelRiego	@PauladeLasHeras No te libraras de ayudar me/n...	2011-12-02T00:49:40	es
2	CarmendelRiego	@marodriguezb Gracias MAR	2011-12-02T00:57:40	es
3	mgilguerrero	Off pensando en el regalito Sinde, la que se v...	2011-12-02T02:33:37	es
4	paurubio	Conozco a alguien q es adicto al drama! Ja ja ...	2011-12-02T02:59:03	es

Vamos a convertir esos textos en nuevas features numéricas que puedan entender nuestros modelos (por ejemplo podríamos convertir los textos en una feature que sea "num_palabras",…)

NLP como resolución de problemas

Además de procesar texto, también incluímos en NLP el conjunto de problemas que ayuda a tratar

Clasficación de Textos (supervisada y no supervisada) -> Sentimental Analysis, Topic Classification, Triage (*) , Spam detection, Languge detection,…

Resumen de textos

Extracción de información (por ejemplo procesado de CVs, formularios, etc)

Traducción

Generación de "textos" (cuentos, código, etc)

Interpretación de Textos (Semántica, intención))

(*) Analizar el texto y decidir siguiente paso a realizar

NLP como resolución de problemas

Además de procesar texto, también incluímos en NLP el conjunto de problemas que ayuda a tratar

Clasficación de Textos (supervisada y no supervisada) -> Sentimental Analysis, Topic Classification, Triage (*) , Spam detection, Languge detection,…

Resumen de textos

Q&A

Extracción de información
(por ejemplo procesado de CVs, formularios, etc)

Traducción

Asistentes Virtuales

Interpretación de Textos (Semántica, intención))

Generación de "textos"
(cuentos, código, etc)

ChatBots

NLP en el Bootcamp (I)

Preprocesado Tradicional

Clasificación de textos: Predicción de Reviews, Analisis Sentimental

NLP flujo de preprocesado: Objetivo

	Texto
0	El misterio del Banco Central: quién estuvo de...
1	El 23 de mayo de 1981, sólo tres meses depué...
2	Cuatro décadas más tarde, un bootcamp bucea ...
3	Es uno de los grandes misterios de la Transición

Pasar de una representación a otra (ojo: no necesariamente la mostrada

	el	misterio	del	banco	central	mayor	la	mayo	meses	barcelona	más	bootcamp	bucea	las	motivaciones	reales	misterios	w_count	num_count
0	0	0	1	0	1	1	0	0	1	0	0	0	0	1	0	1	0	18	0
1	1	2	0	1	0	0	1	0	1	0	0	1	0	0	0	1	0	22	4
2	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	14	1
3	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	1	9	0

NLP flujo de preprocesado: General

No difiere en el tipo de pasos de un procesado al que estamos acostumbrados

1. Entendimiento del problema
2. Obtención de datos y primer contacto
3. Train y Test (para supervisados)
4. **Extracción de Features: Vectorización de textos**
5. MiniEDA: Análisis del target, análisis bivariante, entendimiento de las features, selección de las mismas
6. **Reducción dimensionalidad (caso de ser necesario)**
7. Preparación del dataset de Train: Conversión de categóricas, tratamiento de numéricas
8. Selección e instanciación de modelos. Baseline.
9. Comparación de modelos
10 Selección de modelo:
11. Evaluación contra test.
12. Análisis de errores, posibles acciones futuras.

.


THE
BRIDGE