

OpenCourseWare
**Procesamiento de Lenguaje Natural con
Aprendizaje Profundo,**
Máster en Ciencia y Tecnología Informática

Tema 6 Data Augmentation

Objetivos

- Comprender el concepto de Data Augmentation (DA) y conocer sus principales técnicas.
- Conocer las principales librerías que implementan las técnicas de DA para PLN y saber utilizarlas para generar datos sintéticos que puedan ser utilizados en tareas de PLN.

Índice

- ¿Qué es Data Augmentation (DA)?
- DA en visión artificial
- DA en PLN
- Principales técnicas / librerías de DA para PLN

Data Augmentation (DA)

- **DA** son técnicas dirigidas a **generar nuevos datos sintéticos** a partir de los datos disponibles.
- Los algoritmos de **aprendizaje automático** (supervisado) requieren **grandes colecciones de datos anotados** para la tarea a realizar (por ejemplo, clasificación de imágenes, clasificación de textos, generación de resúmenes, etc), para **obtener buenos resultados**.
- El proceso de **anotación** de estos datos es muy **costoso** (requiere conocimiento experto) y lento.
- DA es un proceso automático que puede reducir el coste en el proceso de creación de nuevos datos.

Data Augmentation (DA)

- **Generan nuevas instancias sintéticos** aplicando **modificaciones** sobre los **datos** del conjunto **original**.
- Además de aumentar la cantidad de datos, también **añaden diversidad** al conjunto de datos.
- DA actúa como una técnica de **regularización**, evitando el **sobre-entrenamiento** en los modelos.

Índice

- ¿Qué es Data Augmentation (DA)?
- **DA en visión artificial**
- DA en PLN
- Principales técnicas / librerías de DA para PLN

DA en visión artificial

- En este campo, las modificaciones más habituales son cambios en tamaño, color, contraste, brillo, rotaciones, zoom, etc.
- [albumentations](#) es una librería de Python para generar nuevas imágenes.

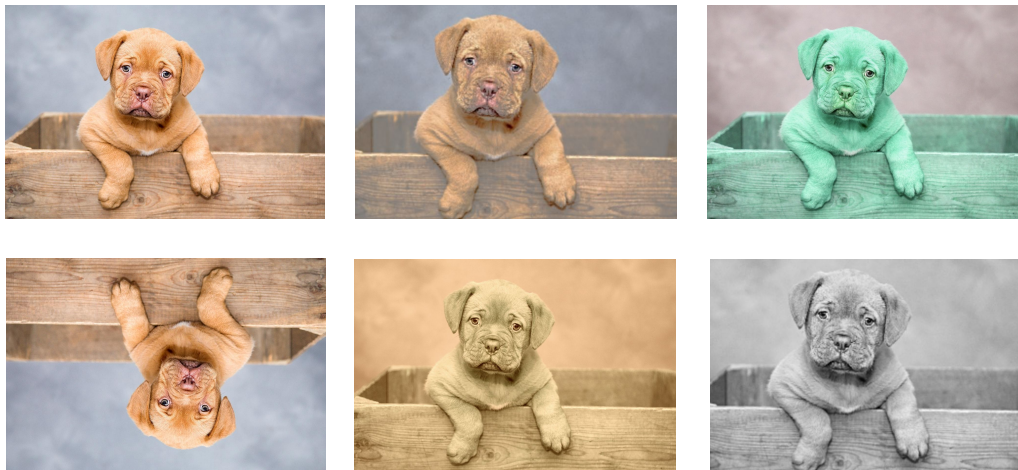


Imagen de [Ilna Krijgsman](#) en [Pixabay](#)

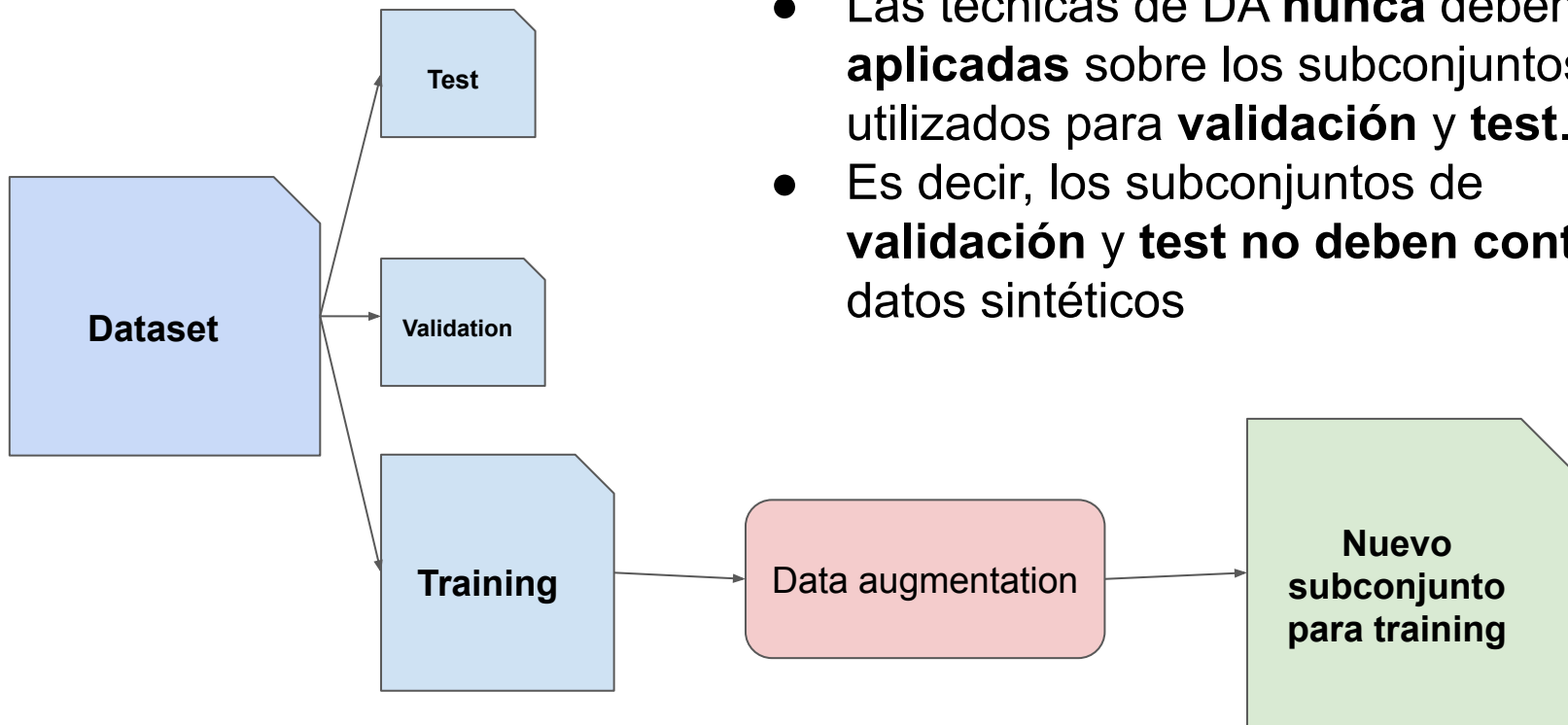
Índice

- ¿Qué es Data Augmentation (DA)?
- DA en visión artificial
- **DA en PLN**
- Principales técnicas / librerías de DA para PLN

DA en PLN

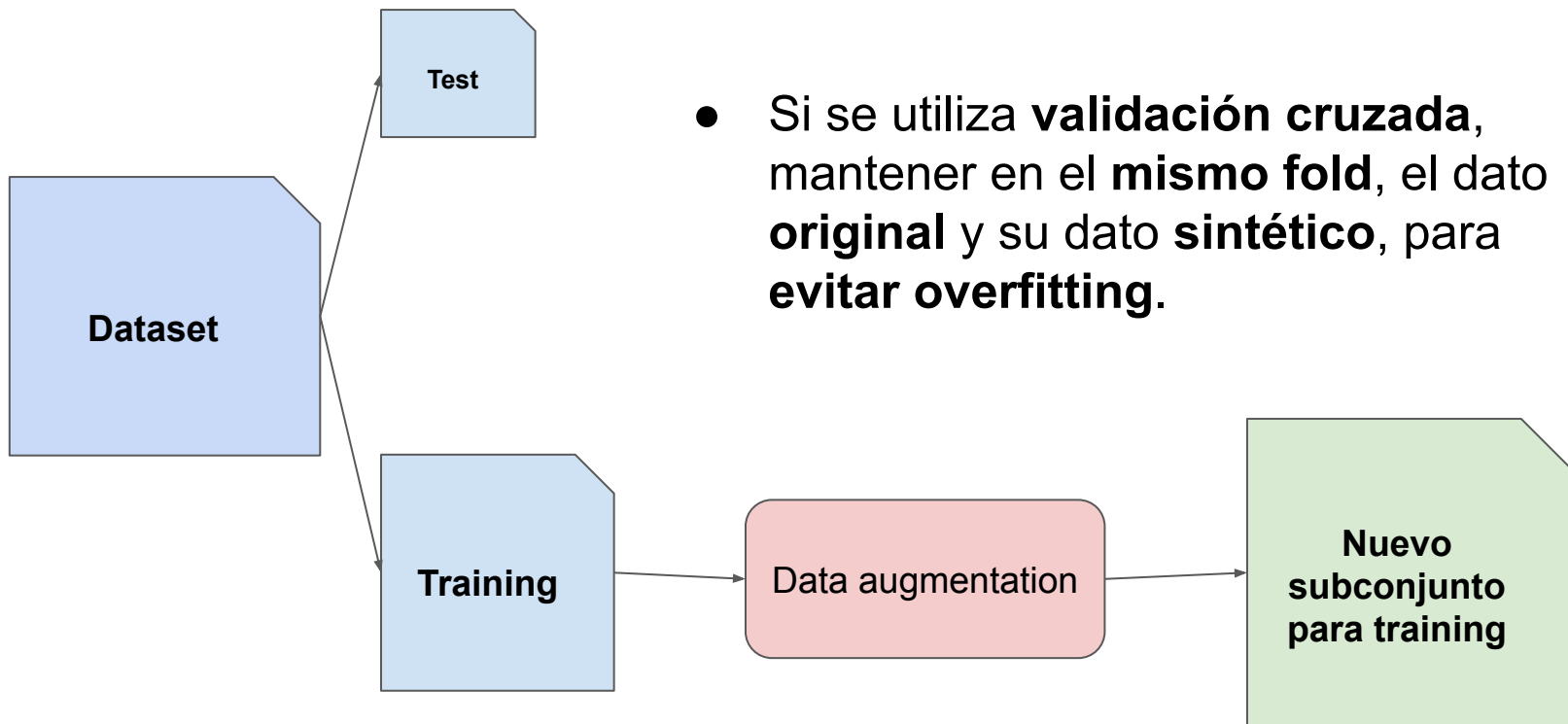
- Operaciones que consisten en **añadir, eliminar o intercambiar caracteres y/o tokens**.
- En PLN, DA es una **tarea** muy **compleja** ya que las modificaciones realizadas sobre el texto original podrían generar un **nuevo texto** con **errores** gramaticales o que haya perdido su significado.
- Además, mientras que en visión artificial, los datos sintéticos pueden generarse durante el entrenamiento, en el caso de PLN, estos nuevos datos sintéticos deben haber sido generados antes de comenzar el entrenamiento.

DA en PLN



- Las técnicas de DA **nunca** deben ser **aplicadas** sobre los subconjuntos utilizados para **validación** y **test**.
- Es decir, los subconjuntos de **validación** y **test** **no deben contener** datos sintéticos

DA en PLN



Índice

- ¿Qué es Data Augmentation (DA)?
- DA en visión artificial
- DA en PLN
- **Principales técnicas / librerías de DA para PLN**

Principales técnicas (librerías) de DA para PLN

- Back Translation
- Easy Data Augmentation (EDA).
- NLP Aug.

Back translation

- Traducir el texto original a un idioma, y volver a traducirlo al idioma original.

No me esperes para el almuerzo



Don't wait me for lunch



No me esperes para almorzar



Imagen de [Raphael Silva](#) en [Pixabay](#)



Imagen de [OpenClipart-Vectors](#) en [Pixabay](#)



Imagen de [Raphael Silva](#) en [Pixabay](#)

Easy Data Augmentation (EDA)

- Operaciones muy básicas para generar nuevos datos sintéticos para tareas de NLP, implementadas en la librería de Python [TextAugment](#)*:
 - sustitución de sinónimos.
 - inserción de sinónimos.
 - intercambio aleatorio.
 - borrado aleatorio.
- **TextAugment** utiliza la librería **nlk** para **obtener** la lista de **stopwords** y los **sinónimos** de [WordNet](#) (base de datos léxica con más de 150.000 palabras en inglés: nombres, verbos, adjetivos y adverbios), o de su versión multilingüe [OMW](#).

EDA: sustitución de sinónimos

- En esta operación, se seleccionan **n palabras** (que no sean stopwords) en un texto para ser **reemplazadas** por **sinónimos** para generar un nuevo texto.

*This **article** will focus on summarizing data augmentation **techniques** in NLP*



*This **write-up** will focus on summarizing data augmentation **methods** in NLP*

EDA: inserción aleatoria

- En esta operación, se selecciona una **palabra** (que no sea una stopword) de forma **aleatoria**, e **inserta** un **sinónimo** de dicha palabra en una **posición aleatoria** en la oración. Pueden producir errores gramaticales y semánticos
- Este proceso se realiza n veces.
- Las palabras originales no son eliminadas o reemplazadas. Por ejemplo:

*This **article** will focus on summarizing data augmentation techniques in NLP*



*This article will focus on **write-up** summarizing data augmentation **techniques** in NLP*



*This article will focus on **write-up** summarizing data augmentation techniques in NLP **methods***

EDA: intercambio aleatoria

- En esta operación, se seleccionan **aleatoriamente dos palabras**, y se **intercambian** sus posiciones.
- Puede producir errores gramaticales y semánticos

*This **article** will focus on summarizing data augmentation **techniques** in NLP*



*This **techniques** will focus on summarizing data augmentation **article** in NLP*

EDA: borrado aleatorio

- Se selecciona una palabra aleatoriamente y se elimina del texto. Puede ejecutarse varias veces.
- Puede producir errores gramaticales y semánticos

*This **article** will focus on summarizing data augmentation techniques in NLP*



*This will focus on summarizing data augmentation **techniques** in NLP*



This will focus on summarizing data augmentation in NLP

NLP Aug

- [Librería](#) de Python que proporciona una **implementación eficiente** de las técnicas de DA para texto.
- En tres **niveles: carácter, palabra u oración**.
- Para cada nivel:
 - eliminación aleatoria,
 - inserción aleatoria,
 - alteración del orden,
 - sustitución de sinónimos.

Resumen

- Los algoritmos de **aprendizaje automático**, y en particular, las redes neuronales profundas **requieren grande cantidades de datos** para poder **entrenar modelos** que puedan obtener **buenos resultados**.
- Las técnicas de **DA** permiten **generar nuevos datos sintéticos** a partir de los datos del **conjunto de entrenamiento**.
- Los subconjuntos de test y validación no deben incluir datos sintéticos.
- **DA en PLN** consiste en **operaciones** para **añadir, borrar o intercambiar caracteres, tokens u oraciones** en el texto original. También se utilizan algoritmos de **traducción automática** (back translation: español -> inglés -> español).

Resumen

- **DA en PLN** es una **tarea compleja** porque el **nuevo texto** generado podría contener **errores** y haber **perdido** la **coherencia** con el significado del texto original.
- Es necesario un **estudio empírico** para determinar la **mejor combinación** de **técnicas** de DA y la cantidad de nuevos datos sintéticas, que permitan mejorar los resultados del algoritmo de aprendizaje automático.
- No siempre ayudan a mejorar los resultados en PLN.

OpenCourseWare
Procesamiento de Lenguaje Natural con
Aprendizaje Profundo,

Gracias!!!

<https://github.com/iseaura>