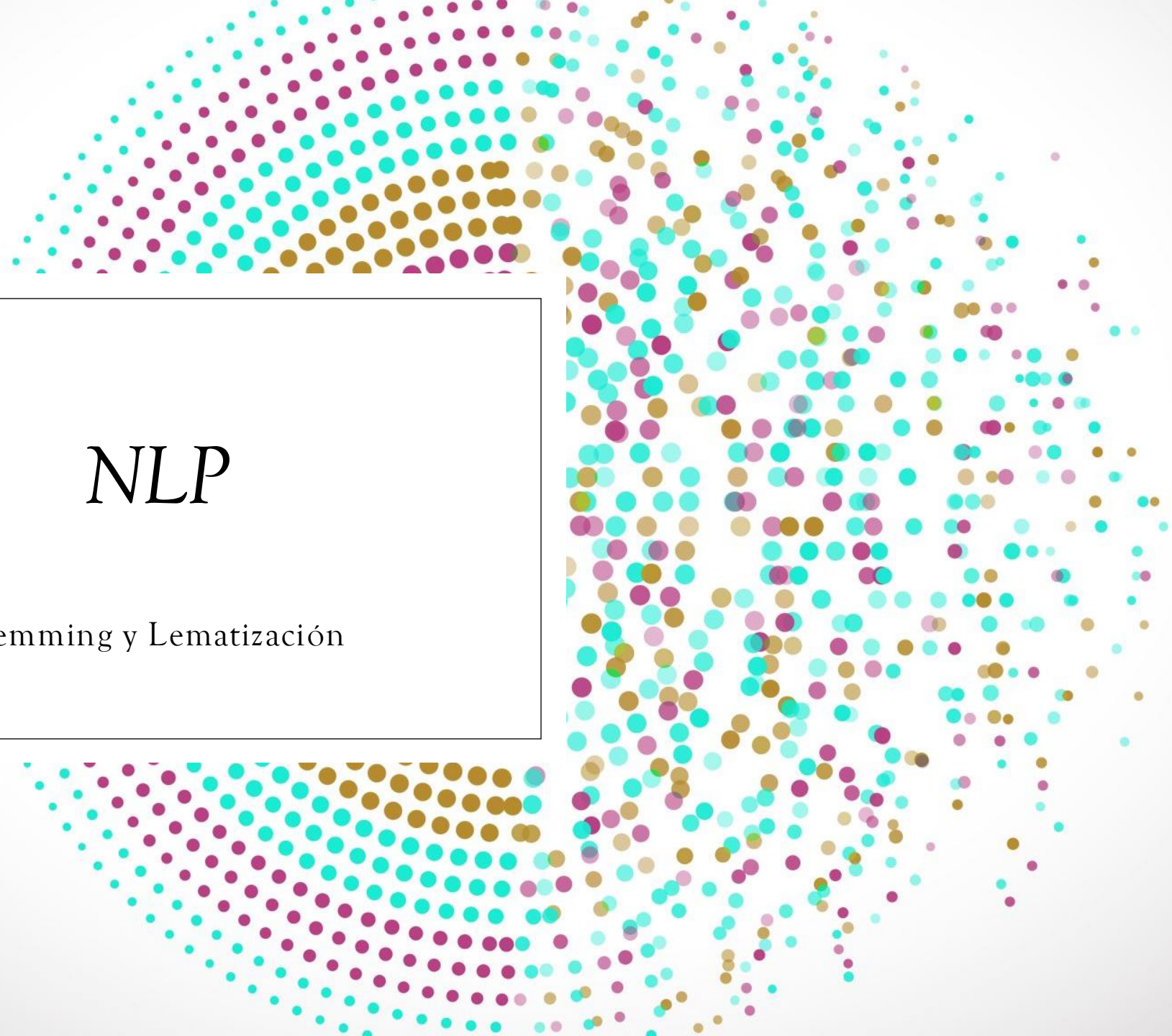
NLP

Stemming y Lematización



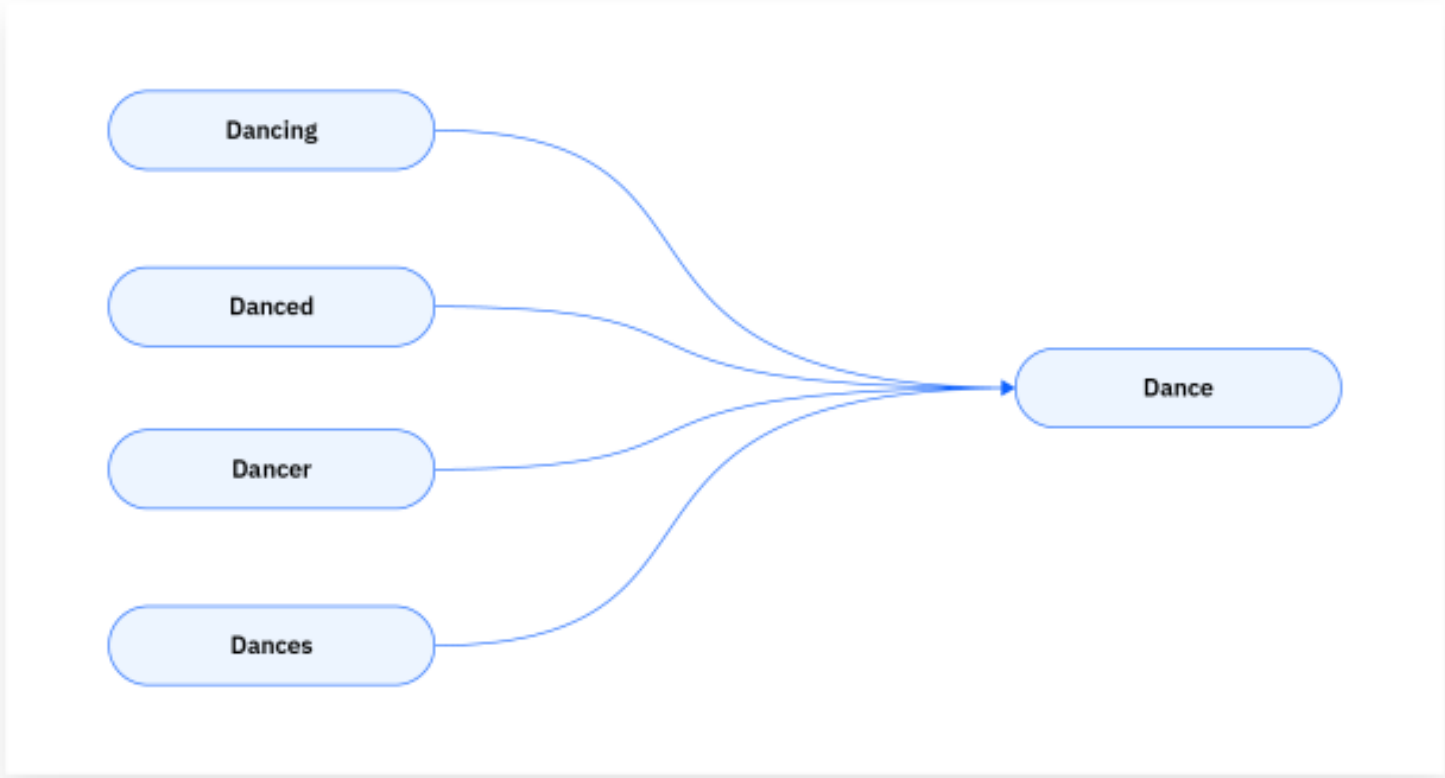
Nota

Faltan las diapositivas con el contenido de Stemming[1], serán agregadas más adelante.

[1]: https://www.ibm.com/topics/stemming

Lematización

Igual que en en Stemming, pero la representación intermedia/lema tiene significado.



Stemming vs. Lematización

El stemming y la lematización funcionan como una etapa en las etapas de minería de texto que convierten los datos de texto sin procesar en un formato estructurado para el procesamiento automático. Tanto la lematización como el stemming eliminan los afijos de las formas de palabras flexionadas, dejando solo una forma raíz. Estos procesos equivalen a eliminar caracteres del principio y el final de los tokens de palabra. Las raíces resultantes, o palabras base, se transmiten para su posterior procesamiento. Más allá de esta similitud básica, la lematización y el stemming tienen diferencias clave en la forma en que reducen las diferentes formas de una palabra a una forma base común.

¿Cómo *funciona la lematización?*

La literatura generalmente define la lematización como el proceso de quitar los afijos de las palabras para obtener cadenas de palabras derivadas, y la lematización como la técnica más ampliamente utilizada para reducir las variantes morfológicas a una forma base de diccionario. La distinción práctica entre lematización y lematización es que, mientras la lematización simplemente elimina los sufijos comunes del final de los tokens de palabras, la lematización garantiza que la palabra de salida sea una forma normalizada existente de la palabra que se puede encontrar en el diccionario.

¿Cómo *funciona la lematización?*

Debido a que la lematización tiene como objetivo generar formas base de diccionario, requiere un análisis morfológico más sólido que el stemming. El etiquetado de partes de voz (Part of speech, POS) es un paso crucial en la lematización. POS asigna esencialmente a cada etiqueta de palabra su función sintáctica en la oración. El módulo NLTK de Python tiene la implementación del algoritmo de lematización Word Net.

Stemming vs. Lematización

Stemming

La representación de palabras puede no tener ningún significado.

Lleva menos tiempo.

Usa stemming cuando el significado de las palabras no es importante para el análisis. Ejemplo: Detección de spam.

Lematización

La representación de palabras tiene significado.

Lleva más tiempo que Stemming.

Usa lematización cuando el significado de las palabras sea importante para el análisis. Ejemplo: aplicación para respuesta de preguntas.