

# 5.1 Recursos para mi aprendizaje |

## Bolsa de palabras y vectores embebidos



### Contextualización

El uso de la “bolsa de palabras” (bag of words en inglés y sus iniciales BOW) es un primer modelo de representación de los documentos en formato numérico, a través de la tokenización y generación de un “diccionario”.

El diccionario obtenido a través de los documentos es una representación de los tokens en su formato “clave:valor” (“key:value” en inglés), que nos ayudará a transformar los enunciados numéricamente. Dicha representación a su vez nos llevará a obtener representaciones matriciales de los documentos, los cuales podrán ser utilizados como datos de entrada para técnicas de aprendizaje supervisado y no supervisado.

Ya vimos que la representación visual de la “nube de palabras” (“word cloud”) nos permite obtener una representación rápida de la manera en que se distribuyen las palabras o tokens en un documento.

En esta quinta semana aprenderás a trabajar con los llamados vectores embebidos o vectores continuos, una técnica mucho más poderosa que permitirá generar mejores modelos de aprendizaje automático o machine learning.

Revisa a continuación el siguiente material de estudio, el cual te ayudará a realizar las actividades del curso y lograr los objetivos de aprendizaje 2.1 a 2.4.



#### Lectura

Lee las siguientes **secciones** del **capítulo 3, Text Representation**:

- **Basic Vectorization Approaches**


- o **One-Hot Encoding**
- o **Bag of Words**
- o **Embedding vectors: word-to-vector**

Del libro del texto:

Vajjala, S., Majumder, B., Gupta, A., y Surana, H. (2020). *Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems*. O'Reilly.

<https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/>

 [\(https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/\)](https://learning.oreilly.com/library/view/practical-natural-language/9781492054047/)

- Nota: El URL de los ebooks que están en la base de datos de O'Reilly solo va a funcionar si se tiene una sesión activa en la plataforma. Es decir, primero deben entrar este link: <https://biblioteca.tec.mx/oreilly> [\(https://biblioteca.tec.mx/oreilly\)](https://biblioteca.tec.mx/oreilly), autenticarse y posteriormente abrir los links directos a las lecturas.
- Licencia: Copyright © 2020 Anuj Gupta, Bodhisattwa Prasad Majumder, Sowmya Vajjala, y Harshit Surana. All rights reserved. Permisos del editor: se permite copiar/pegar.
- Accesibilidad: En la página de la obra dentro de la base de datos, se puede ajustar el tamaño de la letra, el color de fondo y el ancho del párrafo. Para mayor información, se puede consultar la página [Accessibility Guide](https://www.oreilly.com/online-learning/accessibility-guide.html)  [\(https://www.oreilly.com/online-learning/accessibility-guide.html\)](https://www.oreilly.com/online-learning/accessibility-guide.html), de O'Reilly (en inglés).

Lee el siguiente **artículo**:

Khurana, D., Koli, A., Khatter, K., y Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl* 82, 3713–3744.

<https://link.springer.com/article/10.1007/s11042-022-13428-4> 

[\(https://link.springer.com/article/10.1007/s11042-022-13428-4\)](https://link.springer.com/article/10.1007/s11042-022-13428-4)

- Licencia: © 2023 Springer Nature Switzerland AG. © The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022, corrected publication 2022.
- Accesibilidad: En el documento en formato HTML se puede ajustar el tamaño del texto con la herramienta zoom del navegador web. En el documento en formato PDF se puede ajustar el tamaño de la página con el visor de archivos PDF.

Puedes revisar la teoría también en la siguiente **presentación**:

Falcón Morales, L. E. (2023). *Bolsa de palabras: BOW* [PDF]. Maestría en Inteligencia Artificial Aplicada. ITESM. [Acceso al material](#)

(<https://experiencia21.tec.mx/courses/575069/files/226235211?wrap=1>)\_ ↓

([https://experiencia21.tec.mx/courses/575069/files/226235211/download?download\\_frd=1](https://experiencia21.tec.mx/courses/575069/files/226235211/download?download_frd=1))

- Licencia: D. R. Tecnológico de Monterrey, México, 2023. Prohibida la reproducción total o parcial de esta obra sin expresa autorización del Tecnológico de Monterrey.
- Accesibilidad: En el documento en PDF, se puede ajustar el tamaño de la página con el visor de archivos PDF.

Revisa el siguiente documento para estudiar las principales técnicas de vectorización con los métodos basados en CBOW y skip-gram:, a saber, el de Google, Stanford y Facebook:

- [Embeddings\\_Word2Vec\\_GloVe\\_FastText-clase.html](#)

(<https://experiencia21.tec.mx/courses/575069/files/226236572?wrap=1>)\_ ↓

([https://experiencia21.tec.mx/courses/575069/files/226236572/download?download\\_frd=1](https://experiencia21.tec.mx/courses/575069/files/226236572/download?download_frd=1))

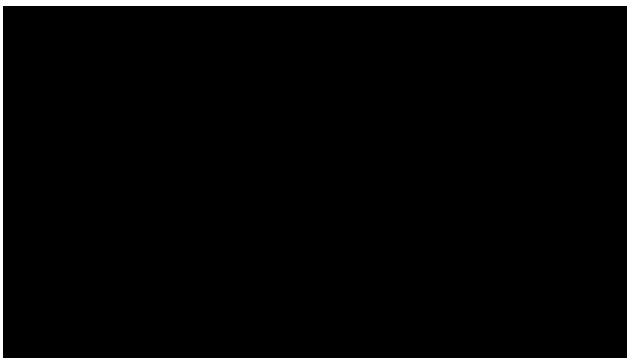
## Video 1/4: Vectorización de palabras - una introducción:



## Video 2/4: Modelos CBOW & Skip-gram:



## Video 3/4: Modelos secuenciales/recurrentes RNN:



## Video 4/4: Ejemplos en JupyterNb:



### Video 5 - complementario:

Descargando los vectores embebidos de fast-text. En mi caso lo hice con Windows10:

