**Unidad 3 – Procesamiento del lenguaje natural**

**Uso de modelos lingüísticos**

**Introducción**

Los modelos lingüísticos son modelos probabilísticos del lenguaje que se basan en representaciones numéricas del lenguaje

**Uso de modelos lingüísticos**

El modelo de **saco de palabras** es una técnica de procesamiento de lenguaje natural utilizada para representar un documento de texto. Su objetivo es tratar un documento como un conjunto de palabras, sin tener en cuenta la estructura gramatical ni el orden de las palabras, simplemente cuenta la frecuencia de cada palabra en el documento. A través de este modelo se pueden utilizar técnicas estadísticas como Naïve Bayes. El modelo BoW se construye a partir de 3 elementos

* Tokenización
* Diccionario de características: crear un diccionario donde las claves son las palabras y los valores su orden de aparición.
* Vectorización: construir un vector con la frecuencia de aparición de cada una de los términos

Una forma de implementar el BoW es utilizando el CountVectorizer.

**Ventajas**:

* Simplicidad: BoW es fácil de entender e implementar
* Eficiencia computacional: computacionalmente eficiente y rápido de calcular:
* Versatilidad: se puede utilizar en diversas tareas de PLN
* Robustez ante ruido: capaz de procesar correctamente texto con errores o ruido
* Interpretabilidad: fácil de interpretar, identificar fácilmente las palabras que contribuyen más a la representación del documento.

**Inconvenietes**:

* Baja capacidad predictiva: no es muy preciso, ya que depende de las palabras utilizadas más frecuentemente
* Descontextualización: los términos de un BoW no tienen contexto, lo que dificulta la comprensión de su significado.
* Tamaño del vocabulario: puede generar vectores enormes de difícil almacenamiento.