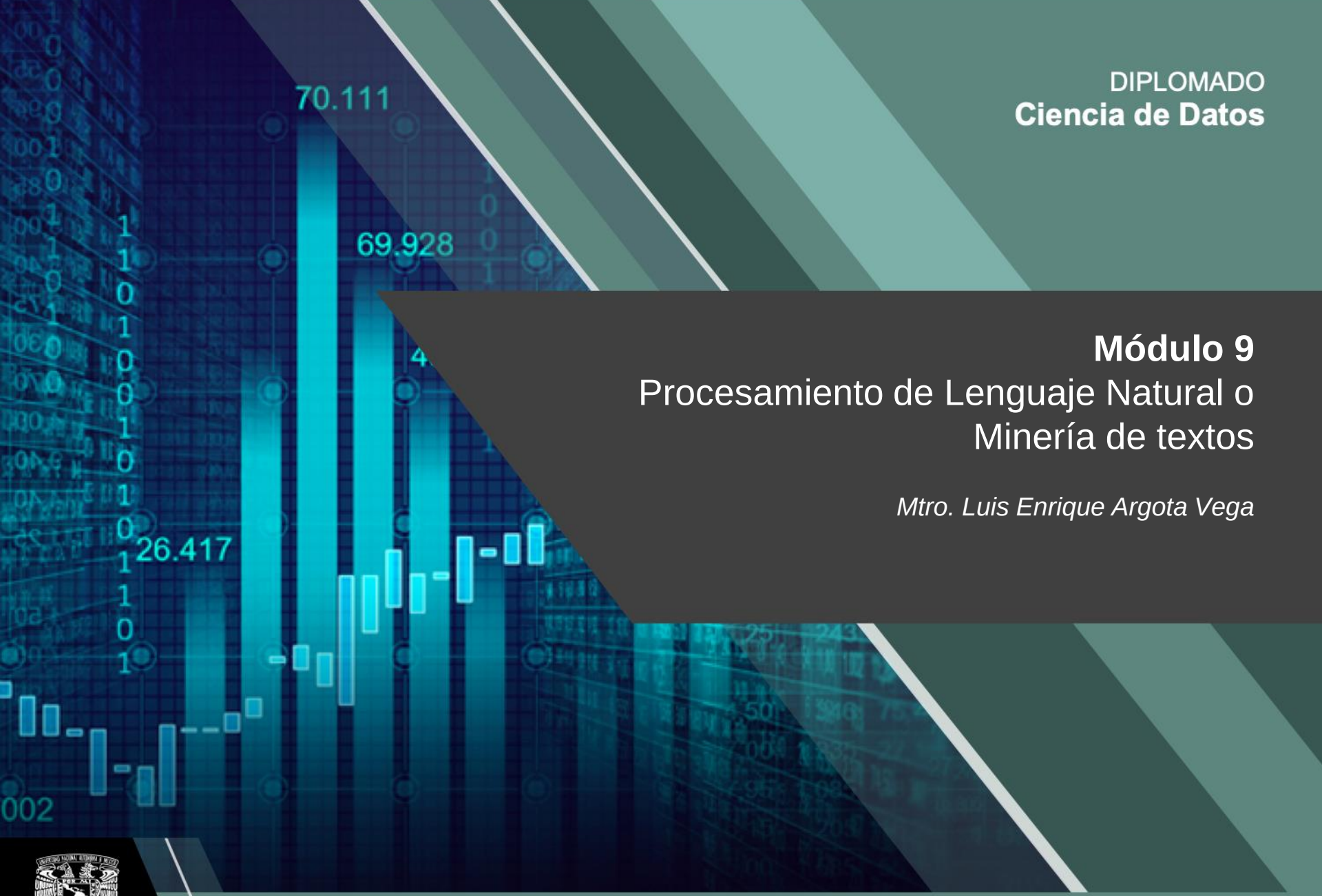


DIPLOMADO
Ciencia de Datos

Módulo 9
Procesamiento de Lenguaje Natural o Minería de textos

Mtro. Luis Enrique Argota Vega

DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Tema 1: Introducción al procesamiento de textos

Objetivo

El participante identificará el procesamiento de textos a partir de los conceptos relacionados con el Procesamiento del Lenguaje Natural y la Minería de Textos, a través del lenguaje de programación Python, para el descubrimiento, extracción y almacenamiento de la información.

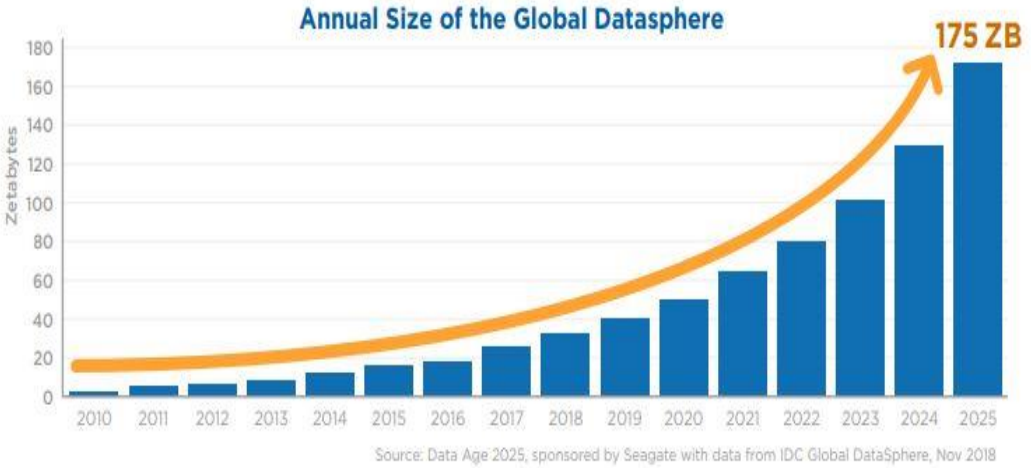
Contenido

1. ¿Qué es PLN?
2. Problemas de ambigüedad
3. Construcciones primitivas en texto
4. Funciones de cadenas y de comparación de textos
5. Manejo de archivos de texto
6. Internacionalización y problemas con caracteres no ASCII

Introducción

Incremento de datos textuales

- El volumen de datos llegará a 175 zettabytes en 2025[1], según un informe de la consultora IDC[2], lo que significa el equivalente a 175 veces la información generada en 2011.



Fuente: The Digitization of the World From Edge to Core

- *Más de la mitad de los datos permanecerá guardado en la nube*[3]. Aproximadamente, el 80% de los datos de una organización se encuentra en **formato no estructurado**[4].

[1] https://www.fundacionbankinter.org/blog/noticia/en-2025-el-volumen-de-datos-en-el-mundo-sera-175-veces-mas-que-en-2011
[2] https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf
[3] https://www.fundacionbankinter.org/blog/noticia/en-2025-el-volumen-de-datos-en-el-mundo-sera-175-veces-mas-que-en-2011
[4] https://support.sas.com/resources/papers/proceedings14/1288-2014.pdf

Datos escondidos a plena vista

Datos escondidos a plena vista

- Tópicos
- Sentimiento

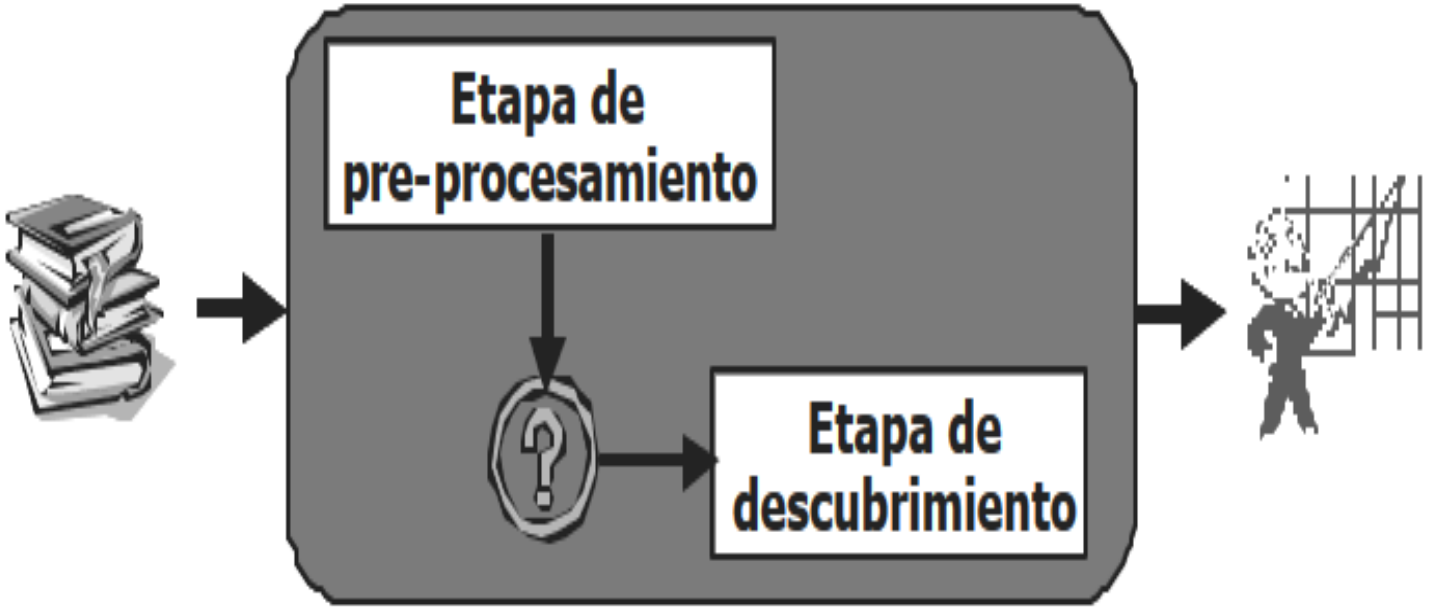
Procesamiento del lenguaje natural

Es un área encargada del desarrollo eficiente de algoritmos para procesar textos y hacer la información accesible a las aplicaciones informáticas.

Minería de textos

Se define como el proceso de descubrimiento de patrones interesantes y nuevos conocimientos en una colección de textos, es decir, es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos (Hearst, 1999) (Kodratoff, 1999).

Metodología para minería de textos



Proceso de minería de textos

Fuente: Montes y Gómez, M. (2001). Minería de texto: Un nuevo reto computacional. Obtenido de https://ccc.inaoep.mx/~mmontesg/publicaciones/2001/MineriaTexto-md01.pdf

¿Qué se puede hacer con datos textuales?

- Encontrar, identificar y extraer información relevante
- Clasificar documentos
- Búsqueda de documentos relevantes
- Análisis de sentimientos, Clasificación de Opiniones
- Agrupamiento de documentos
- Identificación de tópicos
- Resúmenes

Python para minería de textos

Python 3

Entorno de trabajo interactivo:

- **Jupyter Notebooks**
- **Colaboratory**

Construcciones primitivas en texto

- Oraciones / cadenas de entrada
- Palabras o tokens
- Caracteres
- Documentos, archivos más grandes

Ejercicio1(es)-Introducción al procesamiento de textos.ipynb

Tarea

A) Crear una función en Python que:

- Permita leer un archivo (por ejemplo: trabalenguas.txt y/o frases_famosas.txt)
- Extraiga del archivo cada texto en una sola línea.

Debe de tener en cuenta lo siguiente:

- En el archivo *trabalenguas.txt*, los trabalenguas están denotados entre comillas
- En el archivo *frases_famosas.txt*, las frases famosas están denotadas entre guiones

- Guarde los textos extraídos en un nuevo archivo.

Internacionalización

- La codificación de caracteres es el método que permite convertir un carácter de un lenguaje natural (como el de un alfabeto o silabario) en un símbolo de otro sistema de representación, como un número o una secuencia de pulsos electrónicos en un sistema electrónico aplicando normas o reglas de codificación.

Ejemplos:

- ASCII
- IBM EBCDIC
- Latin-1
- JIS: Estándar industrial Japonés
- CCCII: Código de caracteres chinos para el intercambio de información
- EUC: Código extendido de Unix
- Otros estándares nacionales
- Unicode y UTF-8

```
# -*- coding: utf-8 -*-
```

Tarea

B) Indagar en las diferencias de implementación entre Python2 y Python3. Profundice con respecto a la codificación de caracteres.

Conclusiones

- Debido al **incremento de los datos** en formato texto, toma vital importancia el **PLN** a partir del **reconocimiento de patrones** y de la **interpretación de cadenas de textos** para analizar de forma efectiva éstos grandes volúmenes de datos.
- Cuando se enfrenta al texto con la idea de **descubrir conocimiento**, se encuentra con el problema de la **falta de estructura** de este. Esta falta de estructura es solo aparente, porque, realmente, **el texto presenta una estructura demasiado compleja y difícil de tratar computacionalmente**.
- Dependiendo del tipo de operaciones usadas en este **pre – procesamiento de datos**, será el **tipo de patrones** a descubrir en esta colección.

Referencias

- AMPLN. (2019). CICLing: International Conference on Computational Linguistics and Intelligent Text Processing. Obtenido de AMPLN Asociación Mexicana para el Procesamiento del Lenguaje: https://www.cicling.org/ampln/
- Justicia de la T., M. d. (2017). Nuevas Técnicas de Minería de Textos: Aplicaciones. Universidad de Granada. Tesis Doctorales. Obtenido de http://hdl.handle.net/10481/46975
- Gomez-Adorno, H., Bel-Enguix, G., Sierra, G., Sánchez, O., & Quezada, D. (2018). A machine learning approach for detecting aggressive tweets in spanish. In Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), CEUR WS Proceeding.
- Sidorov, G., Markov, I., Kolesnikova, O., & Chanona-Hernández, L. (2019). Human interaction with shopping assistant robot in natural language. Journal of Intelligent & Fuzzy Systems, 36(5), 4889-4899.

Contacto

Luis Enrique Argota Vega

Máster en Ciencia e Ingeniería de la Computación

luiso91@gmx.com

Tels: 5578050838

Redes sociales:

LinkedIn https://cutt.ly/ifPyTEH

Google Scholar https://cutt.ly/WfPtYZz