# **Datos no Estructurados**

# Belizario Mamani, Loja Mamani, Lizárraga Paolo, Kenyi Chino

# August 24, 2020

## Abstract

Before talking about unstructured data, you need to understand what structured data is. When we talk about structured data we refer to the information that is usually found in most databases. They are text type files that are usually displayed in rows and columns with titles. It is data that can be easily ordered and processed by all data mining tools. We could see it as if it were a perfectly organized filing cabinet where everything is identified, labeled and easily accessible.

# **Introduccion**

algunos casos de estos datos o información no están Los datos no estructurados, generalmente son datos binarios que no tienen estructura interna identificable. Es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se identifican y almacenan de manera organizada.

# **RESUMEN**

Antes de hablar de los datos no estructurados, es necesario comprender lo que son datos estructurados. Cuando hablamos de datos estructurados nos referimos a la información que se suele encontrar en la mayoría de bases de datos. Son archivos de tipo texto que se suelen mostrar en filas y columnas con títulos. Son datos que pueden ser ordenados y procesados fácilmente por todas las herramientas de minería de datos. Lo podríamos ver como si fuese un archivador perfectamente organizado donde todo está identificado, etiquetado y es de fácil acceso.

# **TITULO**

Datos no Estructurados

# **AUTORES**

Los autores de este trabajo son:

* Paolo Lizárraga

# Cara de un hombre Descripción generada con confianza alta

* Imagen que contiene persona, interior, hombre, sostener

  Descripción generada con confianza muy altaAnthony Belizario
* Daniel Loja

Una persona sonriendo

Descripción generada con confianza alta

* Kenyi chino

# C:\Users\ASUS\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\532450EC.tmp

# **V.DESARROLLO**

**¿QUÉ SON LOS DATOS NO ESTRUCTURADOS?**

En su definición más básica, simplemente significa cualquier forma de datos que no encaja fácilmente en un modelo relacional o un conjunto de tablas de base de datos desestructurado.

Los datos no estructurados, generalmente son datos binarios que no tienen estructura interna identificable. Es un conglomerado masivo y desorganizado de varios objetos que no tienen valor hasta que se identifican y almacenan de manera organizada.

Una vez que se organizan, los elementos que conforman su contenido pueden ser buscados y categorizados (al menos hasta cierto punto) para obtener información.

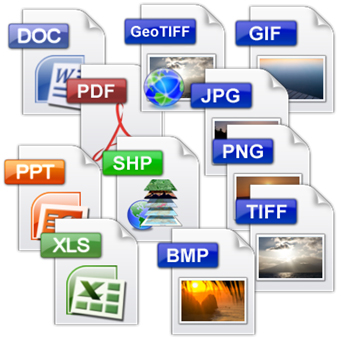
Aunque parezca increíble, la base de datos con información estructurada de una empresa, ni siquiera contiene la mitad de la información que hay disponible en la empresa lista para ser usada. El 80 % de la información relevante para un negocio se origina en forma no estructurada, principalmente en formato texto.

**TIPOS DE DATOS NO-ESTRUCTURADOS:**

Entre los distintos tipos de datos no estructurados tenemos:

* Correos electrónicos
* Archivos de procesador de texto como Word
* Archivos PDF
* Hojas de cálculo como Excel.
* Imágenes digitales como formatos bmp, tiff
* Vídeo como mp4, avi
* Audio como mp3
* Publicaciones en redes sociales.
* Presentaciones como PowerPoint

Mirando esa lista, te podrías preguntar qué tienen en común estos archivos. Se trata de archivos que pueden ser almacenados y administrados sin que el sistema tenga necesidad de entender el formato del archivo. Al no estar organizado el contenido de estos archivos, estos datos suelen ser almacenados en carpetas locales en las redes de las empresas o en la nube como Dropbox, Google drive o SharePoint



**FORMAS DE EXTRACCIÓN DE DATOS NO ESTRUCTURADOS:**

Entre los métodos de extracción de datos no estructurados tenemos:

**Web Scraping (Rascado de datos):**

Se podría definir como la técnica por la que un equipo de desarrolladores es capaz de rascar, escrapear o liberar datos de páginas web de gobiernos, instituciones públicas u organizaciones para acceder a datos privados o públicos que puedan ser publicados o distribuidos en formato abierto. El problema es que la mayoría de los datos de interés están en formatos no reutilizables y poco transparentes como un PDF, por ejemplo.

**Extracción de datos con Python:**

En este ejemplo con librerías como BeautifulShop que nos sirve para la extracción sencilla de datos concretos de una página web en HTML sin excesiva programación. Es lo que técnicamente recibe el nombre de parsear HTML. Una de las ventajas de esta biblioteca en Python es que todos los documentos salientes de la extracción de datos lo hacen en UTF-8, lo cual es bastante interesante porque el problema típico de las codificaciones queda totalmente resuelto

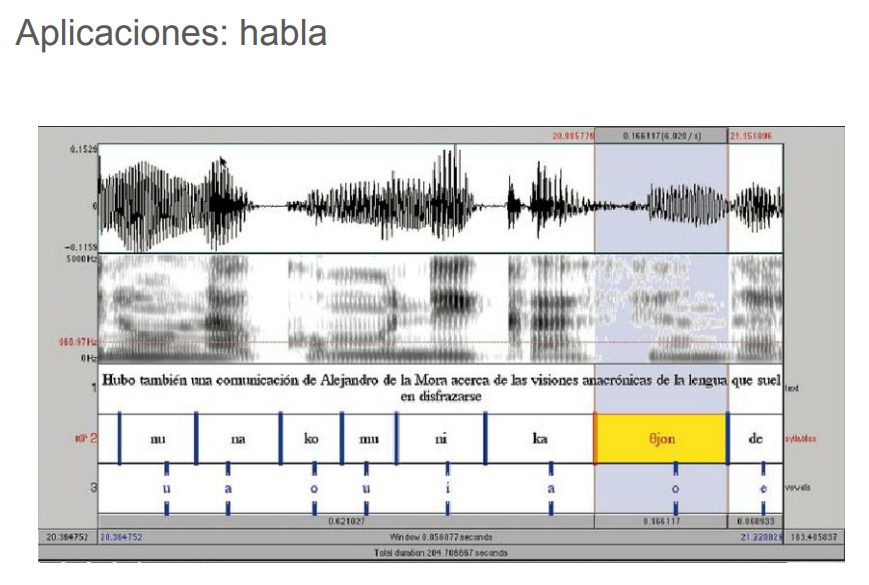
**Document Parsing (Análisis de documentos):**

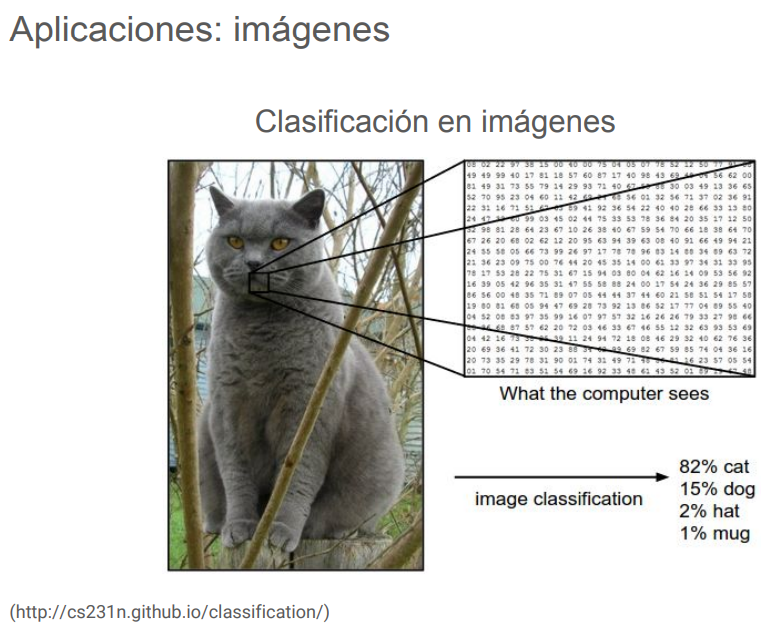
Se utiliza para analizar diferentes tipos de documentos como pdf, html, doc, presentaciones o imágenes. En algún momento es necesario que conserve el formato y la disposición del documento original, por ejemplo, en ocasiones, las estructuras originales de los párrafos, las estructuras de las tablas, los encabezados y subtítulos y el mapeo de las secciones respectivas son importantes para una mejor precisión, por lo que debe conservarlos.

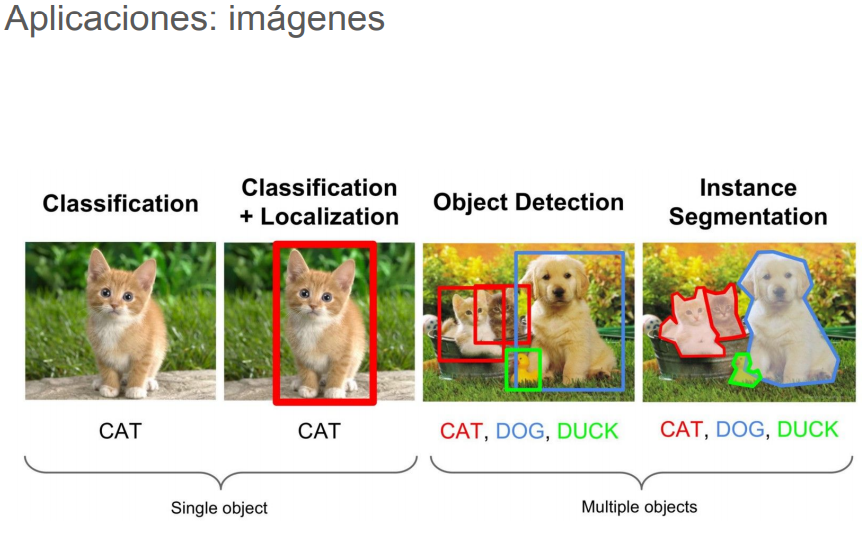
**Tokenización:**

Se trata de dividir el texto en varias oraciones, ya que ciertos procesos solo toman una oración por vez. De manera similar, las oraciones deben convertirse en una secuencia de fichas para ciertos pasos.

**APLICACIÓN DE LOS DATOS NO ESTRUCTURADOS, EJEMPLOS:**

****

****

. ****



**VI. CONCLUCIONES**

•No sólo los datos estructurados pueden ser analizado.

•Hay una gran disponibilidad de datos no estructurados.

•Existen múltiples técnicas para analizar distintos tipos de datos no estructurados.

•Hoy en día se están haciendo grandes avances en esta línea.

# **RECOMENDACIONES**

Para una correcta extracción de datos no estructurados recomendamos seguir 10 pautas:

•Decide tu fuente de datos

•Administra tu búsqueda de datos no estructurados

•Elimina datos inútiles

•Prepara los datos para su almacenamiento

•Decide la tecnología para la pila y el almacenamiento de datos

•Conserva los datos hasta que se almacenen

•Recupera la información útil

•Realizar la evaluación de ontología

•Mantener un registro de estadísticas

•Analizar los datos

# **VIII. BIBLIOGRAFIA**

• https://sommet.mx/blog/que-son-los-datos-no-estructurados

• https://www.kyoceradocumentsolutions.es/es/smarter-workspaces/insights-hub/articles/diferencia-entre-datos-estructurados-y-no-estructurados.html#:~:text=Los%20datos%20no%20estructurados%2C%20generalmente,y%20almacenan%20de%20manera%20organizada.

• https://bbvaopen4u.com/es/actualidad/herramientas-de-extraccion-de-datos-para-principiantes-y-profesionales

• http://mbda.es/que-son-los-datos-no-estructurados/

• http://www.eco.unc.edu.ar/files/ief/workshops/2018/Galvez\_Extraccin\_y\_Anlisis\_de\_Datos\_No\_Estructurados\_\_Aplicaciones\_usando\_texto\_audio\_imgenes\_y\_video.pdf

• https://www.astera.com/es/topic/automated-data-extraction/automated-data-extraction-tools-for-faster-insights/

• Ontología: Se ocupa de determinar dos aspectos importantes de las ontologías: calidad y corrección. ... Esto también ha presentado el desafío de decidir la idoneidad de una ontología dada para los propósitos de uno en comparación con otra ontología en un dominio similar.