Introducción a Extracción de Información

Ph.D. Cristian Enrique Muñoz Villalobos

Pontifia Universidad Católica del Perú - PUCP

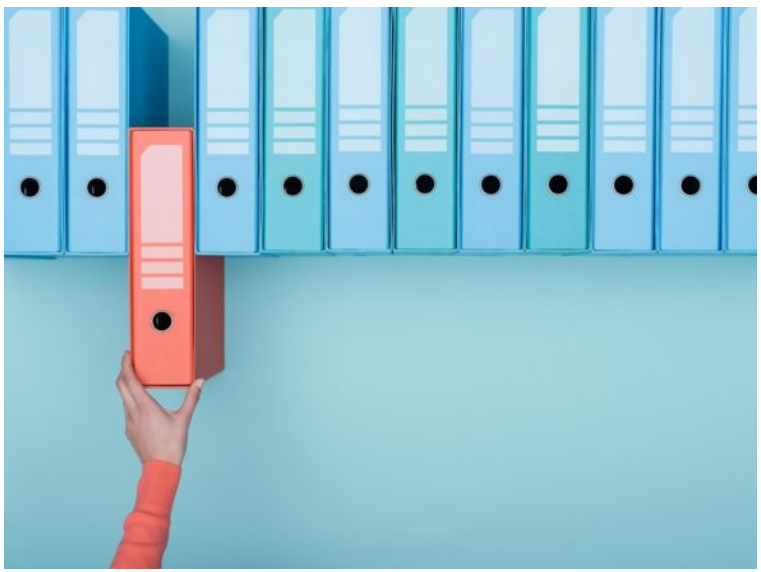
October 4, 2021

Overview

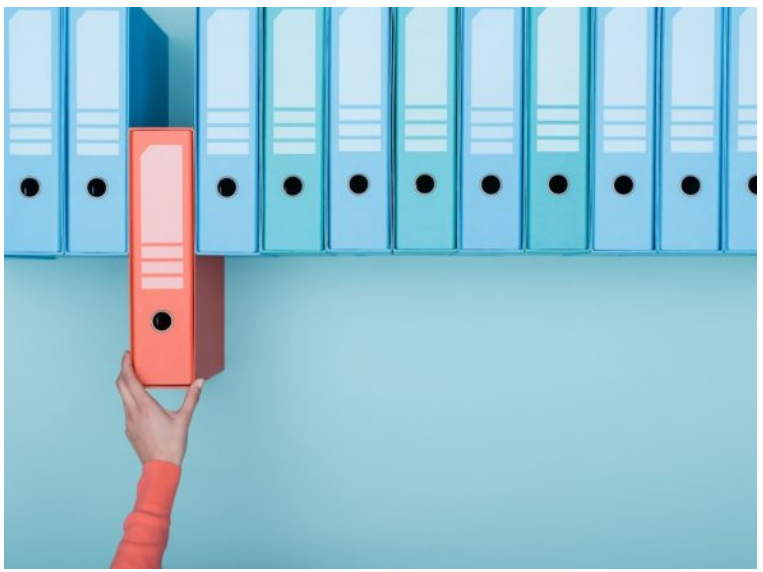
1. Busqueda y Recuperación de la Información
2. Pipeline do Processo de Extración de Información
3. Extracción Sintáctica y Semántica
4. Formatos de dato: JSON y XML
5. Caso práctico: Extracción de documentos word, json, xml, pdf e estructuración en formatos csv.

Busqueda y Recuperación de la Información

La busqueda y recuperación de la información es la ciencia de la búsqueda de información en documentos electrónicos y cualquier tipo de colección documental digital.



(a) Documentos Corporativos



(b) Informaciónes Web

Busqueda y Recuperación de la Información

Componentes:

Web Crawling

Recopilación automatizada de datos y su conversión en información estructurada para su análisis. Ex: datos de paginas web.

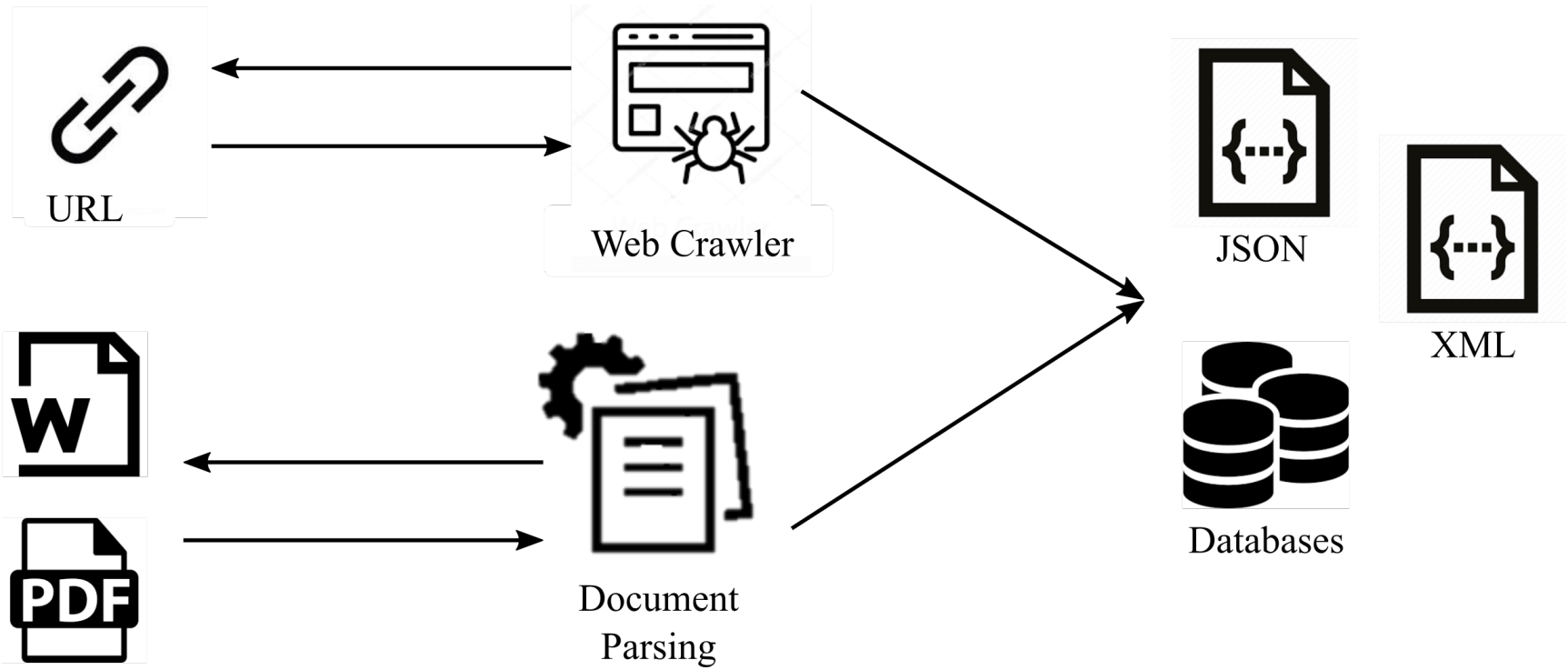
Document Parsing

Implica la examinación de datos presentes en un documento y extraer información útil de él. Ex: datos de PDF, Word, etc.

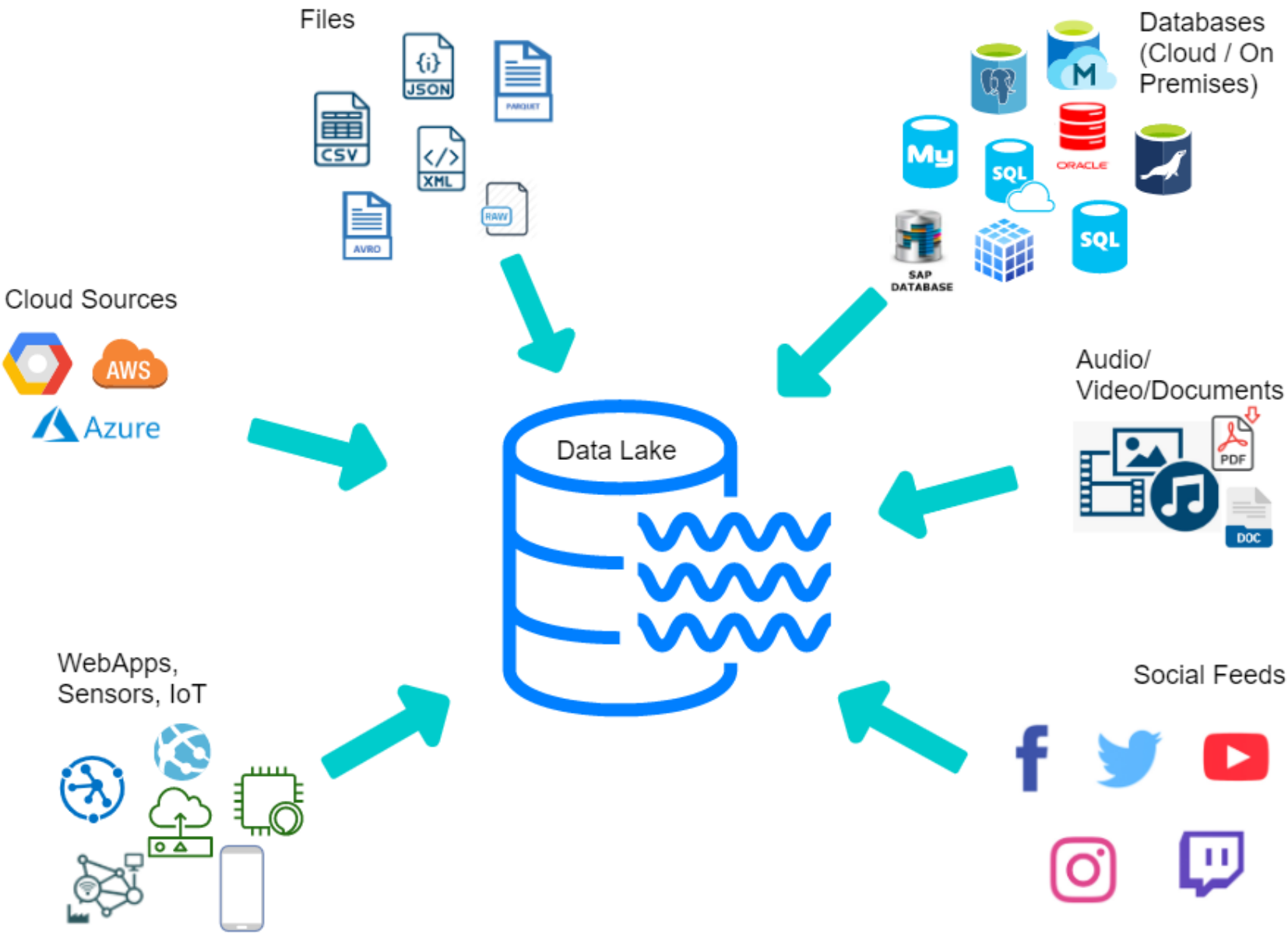
Open Format

Un formato abierto es un formato de archivo para almacenar datos digitales, generalmente mantenida por una organización de estándares, y que puede ser utilizado e implementado por cualquier persona.

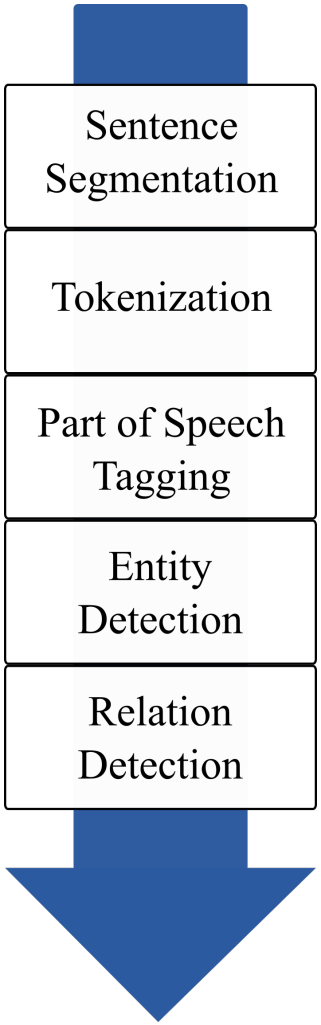
Busqueda y Recuperación de la Información



Almacenamiento: Data Lakes



Extracción Sintáctica y Semántica



Segmentación de Sentencia

- Segmentación de sentencias basicamente é o problema de dividir un texto en una composición de sentencias.
- En varios lenguajes la puntuación es un aproximador razonable. Sin embargo, por ejemplo en español, existe muchas abreviaturas que dificultan la tarea de segmentación.
- La complejidad de esta tarea aumenta con la existencia de tablas, formulas, diferentes formatos, etc.

Que es Tokenización?

- Realiza la segmentación de sentencias en palabras ou unidades atomicas.
- La tokenización es una etapa fundamental usado tanto en métodos tradicionales y avanzados de NLP.

- Tokenización simple (por espacio):

  Dada a frase:

  Se descarto que exista persecución política contra el

  a tokenización seria:

  "Se" , "descarto" , "que" , "exista" , "persecución" , "política", "contra" , "el"

Tokenizadores

Existen varios métodos de tokenización. Cada metodo crea un vocabulario de tokens a partir de un conjunto de reglas o es aprendido mediante un entrenamiento.

- Las tokenizaciones basadas en reglas son utilizadas para conservar la interpretación sintactica del texto. Esto facilita revision y lectura por un ser humano.
- Por otro lado, el entrenamiento de un tokenizadores trae ventajas para la aplicación de modelos de aprendizado automatico. Sin embargo muchas veces la tokenización divide el texto de forma que dificulta la lectura por parte del ser humano.
- Actualmente, los tokenizadores mas comunes mas utilizados para aprendizaje automatico son Byte-Pair Encoding y WordPiece.

Tokenizadores basados en palabras y caracteres

- Tokenizadores basada en palabras o caracteres presenta dificultades en su aplicaciones con modelos de aprendizaje automático.
- Tokenización basada en palabras genera grandes vocabularios de tokens, grande número de tokens no considerados en el vocabulario (out of vobabulary - OOV) y diferentes significados en palabras muy similares. Ex: "persona" e "personas"
- Tokenización basada en caracteres genera sequencias muy largas asi como tambien los tokens individuales son menos significativos. Ex: "persona" $\rightarrow$ "p","e","r","s","o","n","a".

Tokenización basada en sub-palabras

- Tokenización basada en sub-palabras presenta una solución entre los dos mundos. Este abordaje divide palabras poco frecuentes en menos subpalabras com maior frecuencia y significancia.

- Ejemplo: "persona" se mantiene sin dividir y "personas" puede ser dividia en "persona" y "s". Esto ayuda al modelo aprender que la palabra "personas" es formada usando la palabra "persona" con una pequeña diferencia pero la palabra base es la misma.

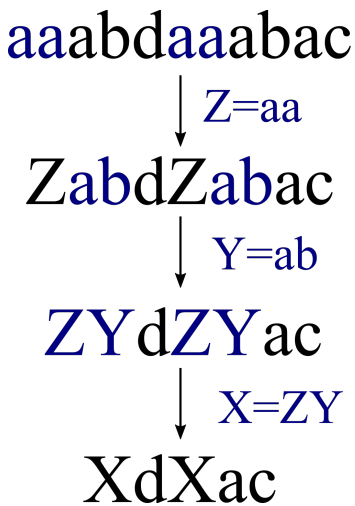
- WordPiece y Byte-Pair Encoding (BPE) son algoritmos de tokenización populares y basado en sub-palabras.

Byte-Pair Encoding (BPE)

Idea: "A new Algorithm for Data Compression", 1994

Exemplo: Comprimir el dato "aaabdaaabac" .

Solución:



Part of Speech Tagging

Part of Speech explica como uma palavra é usada em uma sentencia. A anotación puede usar diferentes formatos. Para comenzar son listadas algunas de las mas importantes: sustantivo (NOUN), pronombre (PRO), adjetivo (ADJ), verbos (VERB), adverbios (ADV), adposiciones (ADP), conjunciones (CONJ) puntuaciones (.).

Ejemplo: Part of Speech de "El mandatario nacio en Cajamarca."

Solución:

```
word: El          upos: DET   xpos: None  feats: Definite=Def|Gender=Masc|Number=Sing|PronType=Art
word: mandatario  upos: NOUN  xpos: None  feats: Gender=Masc|Number=Sing
word: nacio       upos: VERB  xpos: None  feats: Mood=Ind|Number=Sing|Person=1|Tense=Pres|VerbForm=Fin
word: en          upos: ADP   xpos: None  feats: _
word: Cajamarca   upos: PROPN xpos: None  feats: _
word: .           upos: PUNCT xpos: None  feats: PunctType=Peri
```

Deteción de Entidades

La deteción de entidades es la tarea de identificar y categorizar información clave en el texto. La entidad puede ser cualquier tipo de palabra o una serie de palabras.

Ejemplo: Reconocer entidades genericas en el texto "Carlos escucho que la Pontifica Universidad Católica del Perú está Lima."

Solución:

```
entity: Carlos                                      type: PER
entity: Pontifica Universidad Católica del Perú     type: ORG
entity: Lima                                        type: LOC
```

Deteción de Relaciones

La deteción de relaciones semanticas o eventos ocurre entre dos o más entidades.

Ejemplo:

En la frase "Paris está en Francia", la parte "está en" indica la relación entre Paris y Francia. Esta relación es representada por la tripleta (Paris, is in, France).

Formato JSON (JavaScript Object Notation)

JSON es un modelo de almacenamiento y transmición de informaciones en formato de texto.

- Formato simple.
- Estructura compacta (permite una lectura más rápida de la información).

A pesar de ser bien simple, es bastante utilizado por aplicaciones Web. Este formato es utilizado por Google y Yahoo en aplicaciones que necesitan transmitir grandes volumenes de datos.

Formato JSON (JavaScript Object Notation)

Sintaxis Básica:

```
{
    "ano": 2021,
    "altura":1.72,
    "site":"www.mysite.com",
    "temperatura":-2,
    "casado":true
}
```

Formato JSON (JavaScript Object Notation)

Array y Matrices:

```
{
    "estados":["RJ","SP","MG","ES"],
    "matrix":[
    [1,5],
    [-1,9],
    [1000,0]
    ]
}
```

Formato JSON (JavaScript Object Notation)

Objetos:

```
[
    {
      "titulo": "JSON x XML",
      "resumen": "el duelo de dos modelos de representación de informaciones",
      "ano": 2012,
      "genero": ["aventura", "acción", "ficción"]
    },
    {
      "titulo": "JSON James",
      "resumo": "la história de una leyenda del viejo oeste",
      "ano": 2012,
      "genero": ["western"]
    }
]
```

Caso práctico

Extracción de documentos word, json, xml, pdf e estructuración en formatos csv.

Vamos a comenzar!