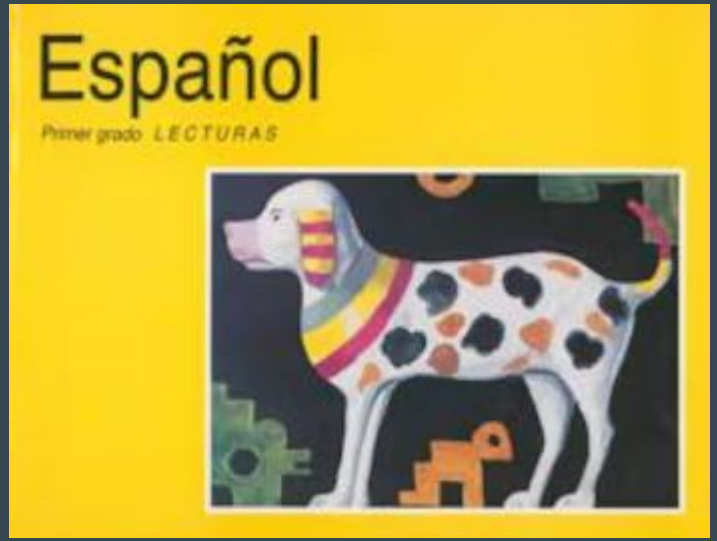
Representación de la Revolución Mexicana en Libros de Texto de la SEP

Elías González Nieto
Seminario de Ciencias de la Computación B
Profesores: Víctor Mireles, Sergio Hernández y Donají Valencia

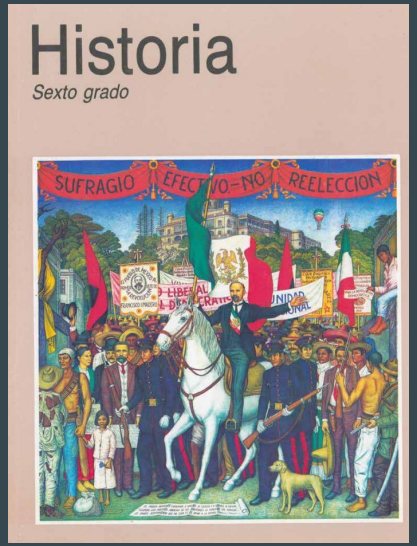
Índice

1. Objetivos
2. Ontología
3. Extracción de Entidades
4. Reporte de Extracción de Entidades
5. Extracción de Relaciones
6. Resultados



Objetivos

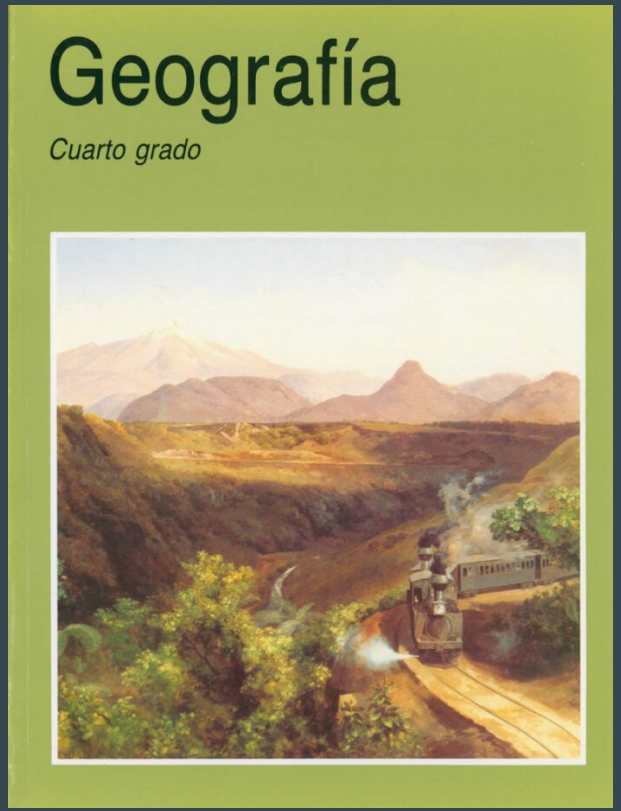
1. Aprender técnicas de procesamiento de lenguaje natural para el análisis de textos.
2. Procesar el texto de las secciones del Porfiriato hasta la Constitución de 1917 de los libros de texto gratuitos de la SEP desde la generación 1960 hasta 2014.
3. Diseñar una ontología para analizar el discurso y los patrones presentes en los libros de texto.
4. Extraer entidades en cada libro y generación.
5. Extraer relaciones en cada libro y generación.
6. * Hacer un grafo de conocimiento de la información conseguida



Ontología

Clases:

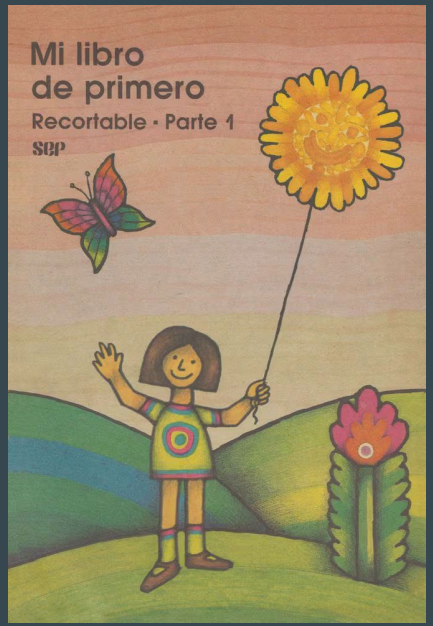
- Actor
- Adjetivo
- Ideología
- Institución
- Libro (contiene Generación y Número de Libro)
- Lugar
- Sentimiento
- Tema (contiene Evento)
- Verbo



Relaciones

Las relaciones que propuse fueron diferentes a las extraídas, pero las que puse en Protegé fueron:

- fue publicado en
- funda
- lidera
- lucha en
- menciona a
- pertenece a
- se describe con
- se identifica con
- tiene sesgo político hacia
- viaja a
- es contemporáneo de
- se alía con



Extracción de Entidades

A continuación se da una explicación de los pasos a seguir en el código, el cual puede ser consultado en aquí. Además, se usó el modelo de HuggingFace mrm8488/bert-spanish-cased-finetuned-ner para hacer el NER inicial.

Extracción de Entidades

1. Por medio del modelo de HuggingFace, extrajimos las entidades y las guardamos en DataFrames junto con su contexto (que también nos arroja este modelo).. Se fue haciendo de 100 en 100 y haciendo una función para regresar diccionarios con entidades como llaves y contextos como valores.
2. Convertimos todo a DataFrames con dos columnas: Entidad y Contexto.
3. Después pasamos nuestros datos por un LLM para hacer deduplicación con el siguiente prompt:

```python
# Prompt del sistema
sysprompt = "Eres un experto en NLP. Deduplicas entidades con nombres variantes y devuelves diccionarios JSON limpios."

# Función para crear el mensaje del usuario
def crear_prompt(diccionario):
    return f'''Aquí hay una lista de entidades, cada una con una lista de contextos donde se menciona.
Por favor, deduplica las entidades y devuelve un diccionario JSON que tenga como llave el nombre más común
de la entidad, y como valor la lista de otras formas en que aparece (sin repetir el principal).

Aquí tienes los datos:
{json.dumps(diccionario, indent=2, ensure_ascii=False)}
'''
```

Extracción de Entidades

Para este punto, el objetivo era tener un nombre homologado para todas las entidades que estaban repetidas o tenían errores. Queríamos conseguir un diccionario de esta forma: {

'Madero' : ['Madero', 'Francisco I Madero', 'Francisco I Mader0', ‘Don Madero’]

}

donde la llave fuera el nombre normalizado y tuviera como valor una lista con los nombres utilizados

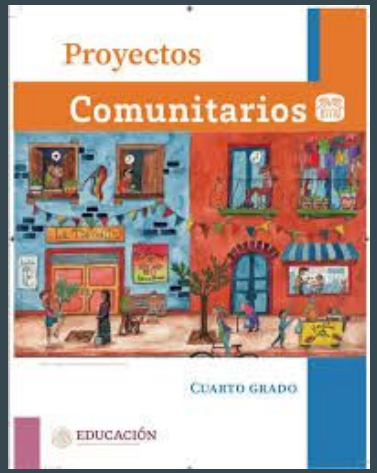
4. Quitamos de las entidades errores del OCR como valores como ## o errores del NER, que hayan identificado entidades que realmente no lo eran.

Extracción de Entidades

5. Guardamos los diccionarios hechos para homologar entidades y los pasamos por los df de cada generación y cada libro para obtener entidades y contextos normalizados.

6. Guardamos esto en la carpeta de drive con la sintáxis n{gen}l{libro}.

Entonces ya teníamos nuestras listas de entidades deduplicadas y homologadas por cada generación y por cada libro



Ideologías

Para las ideologías hicimos un enfoque diferente, el cual se puede consultar en el repositorio del proyecto.

Usamos el modelo ‘paraphrase-multilingual-MiniLM-L12-v2’ de SentenceTransformers, creamos un diccionario cuya llave era el nombre de una ideología y el valor eran ejemplos de oraciones donde estaba presente. Después, hicimos embedding a nuestras oraciones y las pasamos por un regresor logístico con algunos datos de entrenamiento y otros de prueba para identificar la ideología presente en ciertas partes del texto. El resultado fue que obtuvimos una ideología asociada con cada página de cada libro de cada generación, los resultados se pueden ver aquí.

Ideologías

	generacion	libro	pagina	texto	prediccion
0	2008	1	1	El Porf\n\ndel pueblo y con alegorias de la me...	Nacionalismo
1	2008	1	2	115 de julio de 1867,\n\nla capital de México\...	Comunismo
2	2008	1	3	Sin embargo, consiguieron promulgar\nleyes que...	Comunismo
3	2008	1	4	Ferrocarril porfiriano con la\npirdmide de Cho...	Comunismo
4	2008	1	5	1s Eee ona Ses ramiiones Visca\nel cine, las b...	Comunismo

Reporte de Extracción de Entidades

Algunos datos importantes son:

- En total se detectaron: 3748 entidades, contando los duplicados. Además, fueron 1784 entidades únicas.
- Las entidades con más de 50 apariciones fueron:

	Entidad	Apariciones
0	México	245
7	Francisco I. Madero	135
5	Porfirio Diaz	110
124	Francisco Villa	110
174	Venustiano Carranza	74
115	Madero	67
161	Victoriano Huerta	63
48	Veracruz	60
11	Revolución	59
422	Villa	56

Duplicados en cada libro y generación

Estos resultados se pueden consultar aquí

```
n1960l2: total=62, duplicados=14, únicos=48, duplicados repecto al total: 22.58%
n1960l3: total=527, duplicados=66, únicos=461, duplicados repecto al total: 12.52%
n1962l1: total=76, duplicados=12, únicos=64, duplicados repecto al total: 15.79%
n1962l2: total=550, duplicados=77, únicos=473, duplicados repecto al total: 14.00%
n1972l1: total=32, duplicados=6, únicos=26, duplicados repecto al total: 18.75%
n1982l1: total=39, duplicados=12, únicos=27, duplicados repecto al total: 30.77%
n1993l1: total=444, duplicados=49, únicos=395, duplicados repecto al total: 11.04%
n1993l2: total=563, duplicados=65, únicos=498, duplicados repecto al total: 11.55%
n2008l1: total=443, duplicados=53, únicos=390, duplicados repecto al total: 11.96%
n2008l2: total=320, duplicados=27, únicos=293, duplicados repecto al total: 8.44%
n2011l1: total=363, duplicados=33, únicos=330, duplicados repecto al total: 9.09%
n2014l1: total=448, duplicados=33, únicos=415, duplicados repecto al total: 7.37%
```

Extracción de Relaciones

Algunas reflexiones en torno a la extracción de relaciones después de haberlo hecho son las siguientes:

- Fue complicado debido a la cantidad de entidades
- Fue diferente para las ideologías que para las relaciones en general
- Se probaron diversos métodos, pero al final usar LLM's fue lo más conveniente
- El LLM tardaba horas en compilar toda la información

En las siguientes diapositivas se explica los pasos que se llevaron a cabo. El código de extracción de relaciones se puede consultar aquí.

Extracción de Relaciones

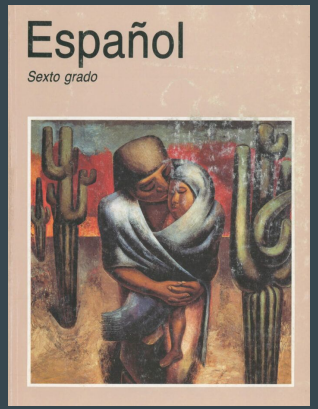
1. Primero unimos todos los DataFrames y e hicimos uno nuevo filtrando por contexto.
2. Después, generamos un prompt para el LLM con el objetivo de obtener las relaciones entre pares de entidades de cada concepto. Pedimos que lo devolviera en un JSON con la siguiente estructura:

{{"sujeto": "entidadA", "relacion": "acción o vínculo", "objeto": "entidadB"}},

...

]

A continuación el prompt



```
"""Eres un experto en procesamiento de lenguaje natural e historia de México.
Tu tarea es identificar relaciones relevantes entre entidades mencionadas en un
contexto histórico. Las opciones que
más nos interesan son: fuepublicadoen, funda, lidera, lucha en, menciona a, no
menciona a,pertenece a, se describe con, se identifica con, tiene sesgo político
hacia, viaja a, es aliado de, es contemporáneo de.

Devuelve únicamente una lista en formato JSON con esta estructura:
[
  {{"sujeto": "entidadA", "relacion": "acción o vínculo", "objeto":
"entidadB"}},
  ...
]
No agregues explicaciones ni texto adicional. No uses valores genéricos como
"entidad1" o "relación".
Un ejemplo es:
```

```
{{
  "sujeto": "Fransisco I. Madero",
  "relacion": "lucha en",
  "objeto": "Revolución Mexicana"
}}

para cada una de las relaciones encontradas en cada contexto.

Solo incluye relaciones que sean explícitas o muy probables.
No agregues explicaciones ni ningún texto adicional."""
```

Extracción de Relaciones

3. Lo pasamos por el DataFrame con los contextos y los prompts y fuimos guardando todas las diferentes relaciones. Algunas son:

Eulalio Gutiérrez	es aliado de	Villa
Eulalio Gutiérrez	avanza sobre	Ciudad de México
Carranza	es aliado de	Obregón
Carranza	es aliado de	Pablo Gonzalez
Carranza	se retira a	Puerto de Veracruz
Carranza	instala	Primera Jefatura
Eulalio Gutiérrez	es nombrado	Presidente de la República
Mi patria	es	México
México	necesita	trabajo material e intelectual de sus hijos
México	necesita	moralidad de todos ellos
México	merece	trabajo material e intelectual de sus hijos

Estas se pueden consultar en el repositorio del proyecto.

Conclusiones

A lo largo de este proyecto se logró aplicar técnicas de procesamiento de lenguaje natural para el análisis de textos. Se procesaron sistemáticamente las secciones referentes al Porfiriato y a la Constitución de 1917 en los libros de texto gratuitos de la SEP, abarcando desde la generación de 1960 hasta la de 2014.

Mediante el uso de herramientas de NER (como HuggingFace y LLMs), se extrajeron entidades clave y se identificaron relaciones relevantes entre actores, lugares y conceptos históricos. Esto nos permitió tener un panorama más completo y sistematizado acerca de la forma de la narrativa en los libros.

Finalmente, el proyecto fue de gran aprendizaje en el tratamiento de textos, en el uso extensivo de LLMs así como de otros modelos como Transformers o HuggingFace.