Introducción

El libro **México: del antiguo régimen a la revolución**, profundiza en la historia de la primer Revolución del siglo XX. La obra del historiador Françoise Xavier Guerra, se basó en una investigación prosopográfica y se convirtió en una referencia fundamental para el estudio del movimiento revolucionario en México. Su trabajo revela las relaciones y tensiones en la transición de la sociedad tradicional -un sistema heredado de la colonia- a un Estado moderno proveniente en gran medida de los ideales liberales de la revolución francesa. En el nuevo orden se erige al individuo como único repositorio de la razón y se sitúa como el eje central de la vida política, en detrimento de actores colectivos como la iglesia o los pueblos de indios. En ese sentido, rechaza cualquier vínculo que no surja de la libre asociación.

Sin embargo, la transición dista mucho de su modelo ideal, no solo por el arraigo de las instituciones antiguas, sino porque de manera tácita, incluso entre la elite liberal se conservan las lealtades familiares, y en los nuevos colectivos como las logias o las sociedades de oficios, frecuentemente se recrean las viejas dinámicas. Por ello consideramos que la elección de la estrategia prosopográfica como herramienta de aproximación al tema es sumamente fructífera ya que permite observar vínculos desestimados en el nuevo sistema, que no obstante, podrían contribuir a la comprension de los movimientos en la cúpula política , que se caracterizó por un reducido número de personajes.

Objetivo

La base de datos de la obra fue almacenada originalmente en tres cintas magnéticas, y el único medio en que se encuentra disponible actualmente es impresa en los anexos del libro. El objetivo de este trabajo es la digitalización de la base para permitir la reproducción de los análisis del autor, así como la generación de nuevo conocimiento a partir del cruce de variables.

Organización de la base de datos

El cuerpo biográfico de la investigación está compuesto por más de tres mil actores sociales entre los que figuran individuos y colectividades, con aproximadamente cien mil datos asociados a los movimientos políticos. La información se sistematizó en tres secciones, el anuario, los catálogos y el diario.

El anuario enlista la totalidad de los personajes, su registro consta de tres componentes, las iniciales, un número de identificación que es único para cada personaje y su nombre completo.

Los catálogos sistematizan variables en cuarenta y dos categorías, las cuales pueden referir a ubicación -estado, ciudad o año-, a datos personales -nivel escolar, el grado de riqueza familiar o a la tendencia política- y datos relacionados con la vida política por -los cargos ocupados o el grado militar-.

En el diario, para cada personaje, se conjunta la informacion de los catalógos en módulos , los cuales contextualizan sucesos como nacimiento, batallas o un cargo politico ocupado, con fechas y lugares específicos. Los modulo constan de 10 carácteres, los primeros dos, son letras mayúsculas que identifican el suceso descrito, cada módulo es independiente lo cual permite enriquecer la base de datos con la captura de nuevos módulos para personajes ya establecidos.

Metodología

Se ocupó el framework OpenCV una biblioteca libre para reconocimiento visual artificial para trabajar sobre las hojas escaneadas de la obra, en este paso se procesaron las imagenes, y se detectó la estructura espacial del texto, para que posteriormente se procesarán pedazos de texto organizados

En la digitalización se creó un programa en Python que ocupa Tesseract, y Pytesseract, ambas son bibliotecas de reconocimiento óptico de caracteres (OCR).

El programa es un proceso semi-automatizado, que se detiene al detectar que la lectura OCR es inconsistente con las características de la base, entonces solicita a la persona que corrija la entrada manualmente para continuar con la lectura,

General

a. hoja original

b. se mejoró la cálidad de las imagenes, haciendola binaria

Anuario

c. se identificaron las columnas en el texto mientras el barrido de las letras hacia abajo, lo cual permitio identificar los tres componentes

d. se seccionaron las hojas en columnas para trabajar una a la vez. En cada columna se identificaron renglones haciendo un barrido horizontal.

e. Se ejecuta el OCR con tesseract y en cada renglon de las columnas se identificaron los tres componentes del anuario: iniciales, numero de identificacion y nombre completo

\*. se dieron condiciones excluyentes, por ejemplo que no fuera capaz de reconocer alguna de las partes, o que detectara letras en el componente de números o viceversa. Al detectar errores el programa presenta el renglón específico que esta procesando y escribe el siguiente mensaje:

Diario

Debido a la estructura modular del diario, los renglones, columnas y espacios en blanco son significativos.

c.1) se identificaron renglones haciendo un barrido horizontal. (primer y segundo cuadrito del paso c)

c.2) Aprovechando la indentación de los módulos respecto a los personajes, se distinguieron los personajes de sus módulos (es el tercer cuadrito del paso c)

d) En cada sección cortada se ejecuta el OCR con tesseract; Se incluyen unicamente mayusculas, numeros y punto (sin imagen)

e) Si al buscar al personaje en el Anuario hay discrepancias, el programa solicita rectificar:

f) se seccionan los modulos de cada personaje por renglon, se ejecuta OCR,

g) cada renglon contiene 3 módulos con 10 caracteres, separados por un espacio (máximo 33 caracteres) y los espacios en blanco se leen cómo ‘\_’. Cuando encuentra una excepcion señala el error con el siguiente mensaje

\*) Los módulos refieren a los catálogos en un orden específico reconocidos por el prefijo al inicio de cada módulo. Cuando no encuentra el valor que lee en las variables del catálogo que le corresponde, pide que se complete con el siguiente mensaje

Al finalizar cada hoja 'Thanks ( ♥ ♥ )!'

En lugar del Perfect, El anuario el diario y los catálogos se almacenaron en archivos .csv

Resultados

Hasta el momento, se han completado el anuario y los catálogos. El diario, esta en proceso de digitalización, entre las dificultades para su avance encontramos que, en ocasiones, los valores de los módulos no corresponden con ninguna variable asignada en los catálogos, consideramos que pudieron haber sido resumidos en el proceso de publicación. Otro percance es la falta de información en los módulos que lo hacen carecer de sentido, por ejemplo el prefijo LX de los módulos debe referir a un vínculo, el siguiente espacio en el módulo designa qué tipo de vínculo y por último de coloca el numero de identificación del personaje al que esta relacionado, sin embargo, en múltiples ocasiones falta el dato del otro personaje. En ese sentido, consideramos que quiza sería de utilidad contar con la tesis doctoral del historiador Xavier Guerra, el documento que dio pie a la publicación del libro.

.

Perspectivas

Se planea poner a disposición la base de datos en un sitio web, para que los usuarios puedan realizar queries ineditos e incluso alimentar la base completando o creando módulos. Consideramos que este trabajo ofrecerá la oportunidad de profundizar en el conocimiento sobre los periodos del Porfiriato y la Revolución Mexicana a partir de información sistematizada por el historiador Xavier Guerra.

Contacto

diegomontesinos@ciencias.unammx

[luciagranadosriveros@gmail.com](mailto:luciagranadosriveros@gmailcom)