Herramientas utilizadas:

Dataiku y Excel para la manipulación, transformación y tratamiento de los datos, así como las visualizaciones de estadística básica

RStudio, específicamente la libería igraph para la el análisis y visualización de las redes de coautoría

Metodología utilizada:

1. Extracción de los datos.

Aquí podrías hacer una breve explicación de las características de la información que guarda el repositorio. Mencionas las colecciones que seleccionaste para ser analizadas y por qué. Explicas los indicadores de interés y el sistema de metadatos que utiliza el repositorio que permite extraer estos datos de forma estructurada.

2. Preprocesamiento de datos:

- se concatenaron los 3 extraídos de las colecciones anteriormente mencionadas utilizando Excel, para poder contar con un juego de datos unificado.

-Se unieron columnas que contenían información redundante respecto a las dimensiones autor, fecha, descripción, identificadores ISBN, ISSN y URI, lenguaje, editorial de publicación, tema, título, y tipo de publicación.

- Se normalizó la dimensión fecha, convirtiendo todos los registros en solo el año

- las dimensiones tema y título se unificaron en una sola dimensión "tema"

Una vez obtenido un marco de datos en bruto se realizó un análisis exploratorio y extracción de indicadores de interés utilizando Dataiku

3. Análisis preliminares utilizando Dataiku

- Para la visualización de frecuencias de publicación, por año y por tipo. (Fig y tabla “1.Productividad por año”)

- Para el conteo de frecuencias totales por tipos de publicación o actividad. (Tablas “2.Conteo de Eventos”, “3.Conteo de Revistas” y la figura “2.Publicaciones por tipo de actividad”)

En algunos casos se detectó ausencia de información que pudo completarse a partir de información relacionada con las publicaciones disponible en Internet u otras dimensiones dentro del mismo marco de datos en bruto. (Por ejemplo, las editoriales de los capítulos de libro)

4. Transformaciones utilizando Dataiku para el análisis estadístico descriptivo

-Separación de los autores por cada publicación. O sea, para cada publicación cuyo registro aparecieron todos los autores unidos, se creó un registro individual, para poder realizar el análisis por cada autor por separado. (Tabla “4.Lista de coautores.xlsx”)

- Se pudo identificar y solucionar problemas de la normalización en los nombres de algunos autores.

- Se realizó un conteo de frecuencias de autoría, independendientemente de su importancia dentro de la publicación. (Tabla “5. Rank de autores”)

-Para el análisis cualitativo se realizó un conteo de autores por cada publicación, por orden de aparición. El dataframe resultante se visualizó de forma transpuesta, con la posición de cada autor como dimensión y se esa forma se logra obtener una matriz de cada autor, con un conteo de cada posición de autoría y se puede evaluar cualitativamente su producción. (Tabla “6.Conteo de autores por lugar.xlsx”)

En general hay una correspondencia entre los autores más prolíficos y su prevalencia como autores principales. Aunque hay autores como por ejemplo, U. Puentes que de ser la segunda con más actividad absoluta, pasa al 7mo lugar en prevalencia en los trabajos en los que participa, queda claro que los 10 primeros autores más prolíficos, no dejan de formar parte de los 10 más destacados cuando se analiza su trabajo en colectivo.

De este mismo análisis es posible identificar 49 trabajos con un solo autor. (Tabla “7. Autores solitarios”)

5. Transformaciones para la elaboración del juego de datos necesario para la visualización de las redes de coautoría

* Transformaciones utilizando Dataiku:
  1. Identificación de la posición de cada autor por cada artículo, con una dimensión para cada autor (primero, segundo, etc)
  2. Creación de un data set que contiene solo un identificador de registro y las dimensiones para cada autor según su posición. (Tabla Todos\_los\_autores\_lugar.xlsx)
  3. Eliminación de los autores que no tienen ninguna colaboración con otros autores
* Transformación utilizando Python
  + Creación de una lista de relaciones a partir de todas las combinaciones existentes de dos autores. (por ejemplo para un listado de autores: A1, A2, A3 se obtendrían las parejas A1+A2; A2+A3 y A1+A3). Para ello se utiliza la el paquete “combinations” de la librería “itertools”, como se muestra a continuación:

pseudo código:

from itertools import combinations

parejas\_autores = registro\_de\_coautores\_df.apply(lambda x: list(combinations(x, 2)), axis=1)

*#es necesario convertir la lista en un data frame*

relaciones = pd.DataFrame(parejas\_autores)

* + “relaciones” es una tabla que contiene 170 colaboraciones por parejas de autores, consta de dos columnas o dimensiones, una para cada autor. (Tabla “8. UAPA\_edgelist”)
  + La tabla “relaciones” es modificada para normalizar los identificadores de los autores para que quede con el formato “Apellido, N.”, para facilitar la interpretación del etiquetado de los grafos. (Tabla 9. “UAPA\_edgelist\_normalizado”)

6. Análisis de coautoría. Herramienta utilizada: Rstudio, lenguaje R, librería: igraph

* Ya que las relaciones que se establecen en la colaboración de autores es recíproca y no existe direccionalidad, el grafo se define como indirecto. Contiene 90 nodos o vértices que representan a los autores y 170 relaciones.
* Figura: “1.Visualización básica” (.png y .html). En este primer acercamiento se puede observar una gran comunidad interconectada, 3 islas de entre 5 y 11 autores y 7 parejas.
* Vértices más lejanos y diámetro de la red: Los nodos o autores más lejanos dentro de la comunidad principal son  L . Alcántara y D.M. Mora, con una distancia de 5 contactos.
* En la tabla “10.Conteo de autores por cantidad de conexiones” puede observarse una tendencia habitual en los ámbitos académicos. 52 (57%) de autores tienen menos de 3 conexiones, de ellos 25 (27%) tienen solo una, mientras que 4 autores concentran 79 (46%) de las conexiones posibles.

Para el cómputo del indicadores de importacia para la red

* Grado de centralidad(“degree centrality”): Corresponde al número de enlaces que posee un nodo con los demás.

Observando este indicador es posible identificar quiénes efectivamente son los autores con mayor cantidad de conexiones dentro de la red.

autor centralidad

1 Puentes P; U . 27

2 Cruz ; M . 22

3 Rodríguez C; J M. 19

4 Moyares N; Y . 11

5 Pérez T; K . 8

6 Mena R; M . 8

7 Leyva V; M . 7

8 Fernández ; J M. 6

9 Guerrero ; C . 6

10 Hernández ; A . 6

* Coeficiente de intermediación (“betweeness”): Cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos.
  + (Nota de Wikipedia sobre este indicador: Los nodos que poseen una posición de intermediarios de alguna manera son también controladores o reguladores del flujo de información. Así, en un proceso de difusión, si el valor de intermediación de un nodo es alto entonces puede actuar como un *broker*; y si es suficientemente alto como para controlar el flujo de información, entonces puede actuar como un *guardián*. )
    - Este indicador muestra cuán importante es un autor para el establecimiento de conexiones entre otros autores. A continuación el resultado de este cómputo. Este resultado confirma la importacia y el peso que tienen las autoras M.Cruz y U. Puentes dentro de la red de investigadores pues no solo son las que tienen la mayor cantidad de contactos, sino que la mayor parte de la red se construye alrededor de ellas.
    - También resulta interesante cuando se visualiza la red de influencia de estas dos autoras principales, que en el caso de U. Puentes, aunque tiene mayor cantidad de colaboraciones, 9 de ellas son con el autor JM Rodríguez. M Cruz tiene una mayor diversidad de autores entre sus contactos. Todo esto se puede observar en las figuras: “Intermediación de M. Cruz” y “Intermediación de U. Puentes”

autor Coef.Intermediación

1 Cruz ; M . 0.19383904

2 Puentes P; U . 0.14101124

3 Martínez ; S . 0.04570991

4 Rodríguez C; J M. 0.03930703

5 Fernández ; J M. 0.02349336

6 Jiminián ; Y Y. 0.01360799

* Densidad : es la relación entre el número el número de aristas o relaciones (edges) del **grafo** y el número de aristas máximo posible a alcanzar. O sea, para este caso, la relación del número de vínculos existentes entre los autores de la red y el potencial de relaciones que pudieran existir si todos los autores colaboraran entre sí. El cómputo de la densidad arroja un valor de 0.04244694, lo que indica que estamos frente a una red dispersa, donde solo 4% de las conexiones potenciales está presente.
* Puntaje de autoridad (centralidad de vector propio o autovector. En inglés “Eigenvector centrality”) se utiliza para medir el nivel de influencia de un nodo dentro de una red. Tiene carácter global. Calcula la centralidad de un nodo como una función de las centralidades de sus vecinos. Cuanto más alta sea el coeficiente, mayor será el nivel de influencia dentro de la red. Esto quiere decir que un autor será más incluyente de forma global dentro de una red en función de cuán bien conectado está en términos de cuántas conexiones tienen los autores con los que trabaja. Según esta medida los autores más influyentes dentro de la red son los siguientes:

autor autoridad

1 Puentes P; U . 1.0000000

2 Rodríguez C; J M. 0.9757066

3 Cruz ; M . 0.3043543

4 Mena R; M . 0.2361688

5 Edwin R; E . 0.2020765

6 Hiraldo T; R . 0.1792240

7 Martínez ; O L.. 0.1751023

8 Martínez ; O L. 0.1751023

9 Wendy Y; F . 0.1751023

10 Jiminián ; Y Y. 0.1465243

Acá resulta interesante el coeficiente obtenido por el autor Rodríguez C; JM, quien aunque ha estado presente entre los primeros 10 autores para cada una de las métricas analizadas aquí asciende hasta el segundo puesto, probablemente por su cercanía en términos de coautorías con la autora Puentes P; U.

Este indicador puede ser visualizado en las figuras “4.Puntaje de autoridad”

plana → .jpg

3D → .html

7. Detección de comunidades con igraph

Dentro de las redes pueden encontrarse sub-redes que están densamente conectadas interiormente, estas subredes se denominan comunidades. Pueden estar muy relacionadas entre sí o escasamente conectadas a otras. Utilizando la función edge.betweenness.community() se identifican 13 comunidades que cuentan con la siguiente estructura:

Comunidad A B C D E F G-M

Membresía 23 20 13 8 6 5 2

En tabla “5. Comunidades” se encuentra el listado de autores por comunidad.

En las figuras “6.Comunidades” se pueden visualizar las estructuras de las comunidades.