REPORTE DE AVANCES

MAPEO DE LA INFORMACIÓN ESTADÍSTICA DE COLOMBIA - DANE

Camilo Espitia - 200812115 y Pierre Raimbaud - 837342

REPORTE DE AVANCES **1**
Objetivos del proyecto 2
Objetivo general 2
Objetivos especificos: 2
Análisis por parte del experto 2
Primeros Resultados 3
Actualización del What-Why y How 3
WHAT 3
WHY 4
HOW 5
Fusión de los dos datasets de registros administrativos: 7
Avances en limpieza de datos y procesamiento de texto para encontrar palabras claves 8
Proceso de limpieza de Datos: 8
Resultados del procesamiento preliminar de texto para obtener las palabras claves: 9
Estado del arte y Justificación de la Visualización 10
Avances implementación visualización 13
Archivos Adjuntos 15

1. Objetivos del proyecto

Para contextualizar este informe primero recordaremos cuáles son los objetivos del proyecto.

Objetivo general

Solucionar un problema real del DANE basado en la necesidad de extraer información de un conjunto de datos, mediante el desarrollo de una aplicación de Visual Analytics para responder las preguntas propuestas por los expertos.

Objetivos especificos:

Construir una visualización que permita:

- Conocer qué registros y operaciones estadísticas están relacionados.
- Conocer cuáles registros y operaciones estadísticas se relacionan a una temática clave (ej. línea de política pública y/o indicadores) para agregar valor en la toma de decisiones.
- Descubrir nuevas taxonomías de registros y operaciones estadísticas.

2. Análisis por parte del experto

Realizamos una reunión con el DANE el día viernes 19 de Octubre para mostrar los avances y revisar nuestra propuesta de visualización. Obtuvimos las siguientes conclusiones/tareas:

- El comité de Visual Analytics DANE está de acuerdo con nuestra abstracción del WHY-WHAT y HOW en general. En la reunión discutimos las tareas y les mostramos el mock up para validar que las tareas si pudieran ser realizadas allí.
- Para la tarea principal **T2** (Buscar los registros administrativos y operaciones estadísticas relacionadas a una temática), el comité nos sugirió que en la visualización exista una fuerza que atraiga hacia el centro los RRAA y OOEE más relacionados con el tema clave, mientras los menos relacionados aparecen en la periferia.

- El comité nos entregó un nuevo dataset de Registros Administrativos correspondiente al inventario de 2016 para ampliar la información de registros administrativos del inventario de 2018.
- Como tarea extra nos comprometimos a realizar un informe de compatibilidad entre los dos inventarios de registros administrativos para encontrar diferencias significativas.

3. Primeros Resultados

A. Actualización del What-Why y How

Aunque no hubo muchas actualizaciones en la abstracción del WHAT-WHY y HOW a continuación se encuentran las versiones actualizadas.

WHAT

El DANE nos entregó 3 datasets, 1 inventario de **Operaciones estadísticas** y dos inventarios de dos años diferentes de los **Registros administrativos**. Luego de fusionar los dos datasets de registros administrativos, validamos la información que mostrará la visualización.

Es un **Dataset Type: Network** en donde cada nodo es un registro administrativo o una operación estadística. Los datasets derivados que utilizaremos en la visualización son:

Dataset de nodos: (1059 Filas: 549 RRAA (Registros Administrativos) **+ 510 OOEE** (Operaciones estadísticas)**)**

- **NodeID:** Identificador único del nodo.
- **Tipo:** Registro Administrativo u Operación Estadística.
- **Grupo:** Cluster calculado según palabras clave.

Dataset de links:

- **LinkId**
- **LinkSourceId**
- **LinkTargetId**

2 datasets de atributos de nodos:

Para los nodos de tipo RA.

- NodeId
- Nombre Del Registro Administrativo

- OO.EE Generada A Partir Del R.A
- Objetivo Del Registro Administrativo
- Unidades De Observacion
- Nombre De La Entidad
- Area Tematica
- Tema al Que Pertenece El Registro Administrativo
- Periodicidad De Captura Del Registro
- Cobertura Geografica
- Desagregación Geográfica

Para los nodos de tipo OE.

- NodeID
- Nombre de la Operación Estadística
- Objetivo
- Unidad De Observación
- Variables
- Resultados Estadísticos (Indicadores, Agregados)
- Usuarios De La Información Estadística
- Área Temática
- Tema
- Entidad Responsable
- Dependencia
- Metodología Estadística
- Desagregación Geográfica
- Periodicidad De Difusión
- Periodicidad Definida

WHY

Tareas principales

T1. Explore / Topology (clusters)

Descubrir nuevas taxonomías de la información estadística en Colombia a través de explorar la topología de los datos. La visualización debe ayudar a establecer relaciones entre Operaciones estadísticas y Registros administrativos, que componen clusters alrededor de diferentes temáticas.

T2. Identify / Features (links)

Identificar cuáles operaciones estadísticas y registros administrativos están conectados con un tema determinado que el usuario filtra para identificar que datasets pueden ser utilizados para crear políticas públicas relacionadas, etc.

Secondary Tasks

T3. Identify / Outliers (nodes)

Identificar las OOEE y RRAA que no pertenecen a ningún cluster y no se relacionan con otros.

T4. Identify / Extremes (nodes)

Identificar qué operaciones estadísticas y registros administrativos tienen el mayor número de conexiones dentro de un cluster para identificar los nodos más relevantes para un tema determinado.

T5. Browse / Features (nodes)

Navegar en los nodos de un cluster para obtener información detallada sobre un nodo en específico.

T6. Enjoy (Explore / Topology)

Entregar una herramienta para el Sistema Estadístico Nacional, en donde usuarios que no son expertos en los datos pueden encontrar un inventario de las OOEE y los RRAA del DANE.

HOW

Recordemos el Mock up (Visualización principal para T1)

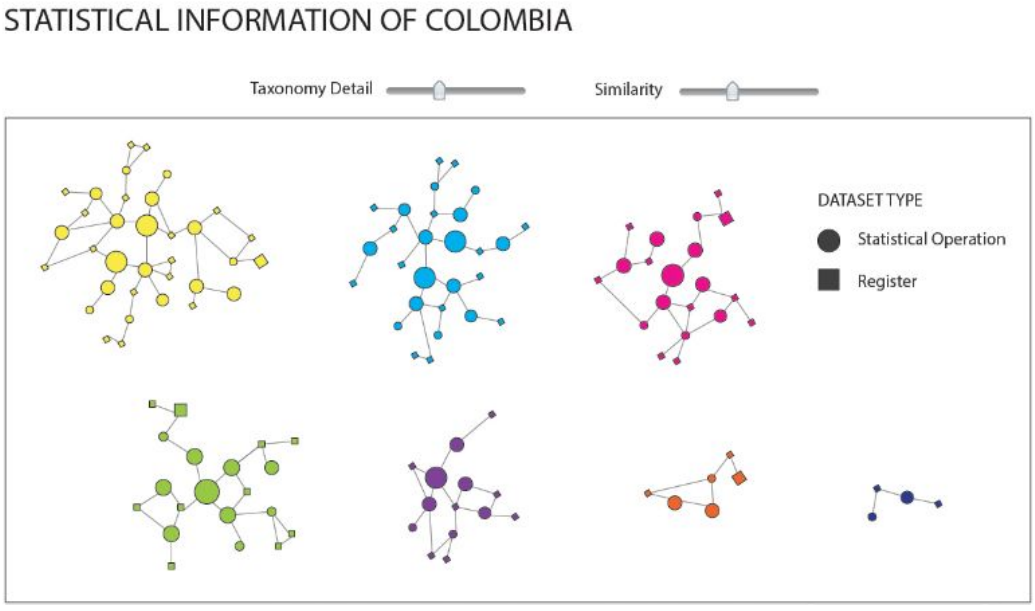

Figura 1: Visualización mock up 1

Recordemos el Mock up (Visualización para T2 - Izquierda y Vista de detalle Derecha)

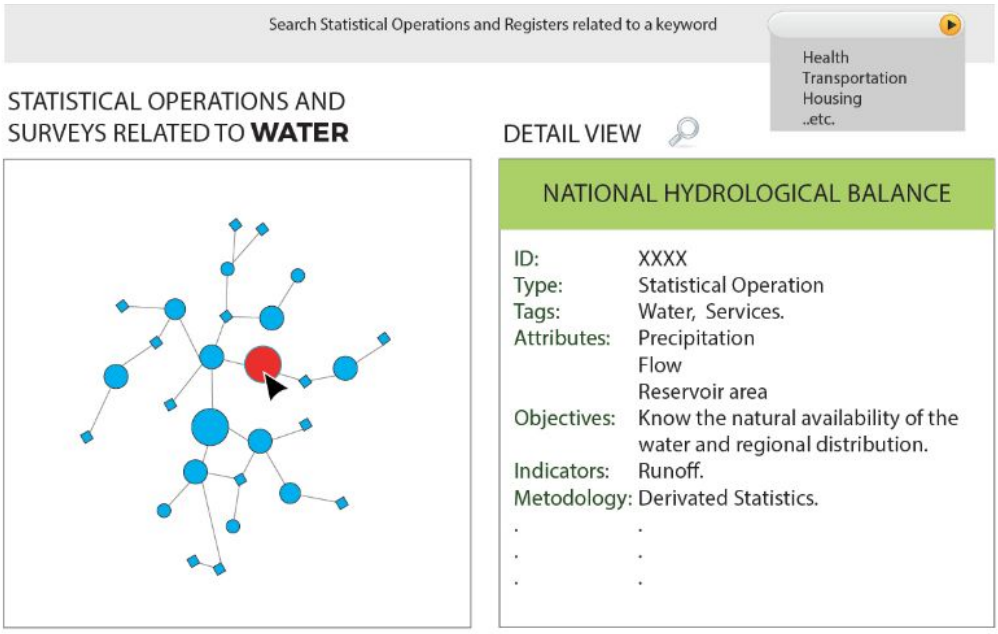

Figura 2: Visualización mock up 2

Marks

Point-Shape: Para representar cada nodo / item (Attribute: Type, Type: Categorical)
Line: Como conector de nodos.

Channels

Spatial Region-Position: To separate clusters. (Attribute: Taxonomy, Type: Categorical) Encoding: Arrange: Separate and Order.

Color Hue: To separate clusters (Attribute: Taxonomy, Type: Categorical)

Shape: To specify the category of each node: Register - Square, Statistical Operation - Circle. (Attribute: Type of Item, Type: Categorical)

Note: Es posible que utilicemos el tamaño del nodo (Area 2D) para expresar el número de conexiones de un nodo (Type: Quantitative).

Expressiveness and Effectivity
Expresividad: Proponemos un doble encoding de posición y Color Hue para separar los diferentes clusters.

Efectividad: como estamos expresando atributos categóricos usamos dos de los canales más efectivos para atributos categóricos, posición y Color Hue.

Idiom: Node-Link Diagram, Force Directed Placement.

Interactivity

Aggregate:
Para agregar y desagregar los clusters y explorar la taxonomía descrita en T1, permitiendo explorar una taxonomía más general o una más detallada.

Filter (by keyword) :
Para lograr T2 es necesario tener un filtro por palabra clave. Es así que proponemos un filtro que sólo permita ver los nodos relacionados con un tema específico.

Filter (by number of nodes):
Para lograr T3 e identificar los outliers proponemos un filtro en la visualización principal para que muestre los nodos que no tienen conexiones.

Manipulate/Select:
Para conocer cuál es el nodo más relacionado con un tema (T4) proponemos una tabla en donde aparezcan los nodos ordenados por número de conexiones.

Manipulate/Select:
Para ver el detalle de los nodos propuesto en T5 proponemos una herramienta de selección para mostrar los atributos de un nodo en la vista de detalle.

B. Fusión de los dos datasets de registros administrativos:

Según las conclusiones de la reunión, nos comprometimos a unificar los dos datasets de registros administrativos para complementar el inventario más actual (de 2018). Llevamos a cabo un análisis entre estos dos inventarios y encontramos las siguientes conclusiones:

- El inventario de 2018 tiene 549 registros mientras que el de 2016 tiene 395. (Diferencia de 154 registros)
- Los registros no tienen un identificador único.
- Comparamos los registros por nombre y sólo 90 coincidían en un principio, después de realizar una unificación manual de los dos registros, sobre todo por diferencias de estilo, tildes, uso de guiones, parentesis,

etc, sólo hubo un registro de los 395 del inventario de 2016 que no se encontraba en el de 2018 ya que estaba dividido en 2. El registro se llama: AFILIADOS O PENSIONADOS DE COLPENSIONES.

- En conclusión, los inventarios se corresponden mutuamente, sin embargo, el de 2018 se encuentra mucho más completo.

C. Avances en limpieza de datos y procesamiento de texto para encontrar palabras claves

Proceso de limpieza de Datos:

El proceso de limpieza de datos que realizamos comprende los siguientes pasos:

- Eliminación de espacios al principio y al final de las oraciones.
- Eliminación de espacios dobles.
- Revisión de caracteres “especiales”: paréntesis, guiones, barras oblicuas, tildes, etc.
- Revisión manual uno a uno de las palabras en los títulos de los registros de los dos inventarios para poder unificarlos.
- Es necesario realizar este proceso con el resto de las columnas para no obtener un diccionario de palabras claves con palabras “repetidas”. Por ejemplo: **Hidrología** e **hidrologia** (sin tilde).

Una vez los documentos Excel limpios, otro paso útil es exportarlos a documentos CSV con el fin de poder procesarlos de manera más simple y más “universal” (el CSV es más estándar). Este proceso se hace simplemente exportando el fichero en archivo CSV desde Excel. Pero para poder realizar este proceso sin problema primero hay que tomar en cuenta unas etapas de preparación:

- Eliminación de retornos a la línea
- Eliminación (o cambios para unificación, por ejemplo a espacio) de las comas y/o punto y comas: aquí la idea es poder exportar a CSV con separador coma o punto y comas sin problema

Se puede obtener entonces un archivo CSV donde, abriendolo en Excel o en un simple editor de texto, una línea es una fila, es decir un ítem, lo cual permite un procesamiento

más simple del archivo CSV, ya que en un programa “custom”, como él que hicimos en Java, lo más simple y más típico es leer un archivo línea por línea.

Resultados del procesamiento preliminar de texto para obtener las palabras claves:

Según el primer procesamiento de texto que realizamos con los atributos de los RRAA *(Objetivo Del Registro Administrativo, Unidades De Observación, Nombre De La Entidad)* y de las OOEE *(Objetivo, Unidad De Observación, Variables, Resultados Estadísticos (Indicadores, Agregados), Usuarios De La Información Estadística)*, obtuvimos:

192 - palabras clave repetidas (*en el archivo de Operaciones Estadísticas)*
Por ejemplo:
credito,81
servicio,76
tarjetas,76
transacciones,68
impuesto,66
tasa,58
interes,57
salud,51
bienes,51
acciones,50
transporte,50
educacion,50
renta,49

79 - “frases” palabras clave agrupadas de a 2 (*también en Operaciones Estadísticas)*
Por ejemplo:
tarjetas debito,26
energia electrica,21
seguridad social,13
interes social,12
politicas publicas,12

88 - “frases” palabras clave agrupadas de a 3 (*también en Operaciones Estadísticas)*
Por ejemplo:
tarjetas de credito,38
bienes y servicios,16

transacciones por retiros,12
costos de produccion,10
tarjetas debito emitidas,10
rendicion de cuentas,9
saldo a pagar,9
tasa de cambio,9

El detalle de los resultados se encuentra en el archivo PreliminaryKeywordsResult.txt. Conviene precisar que es un resultado preliminar ; desde la última vez que fue generado el archivo, el procesador de texto ha subido nuevas implementaciones y mejoras, lo cual permitirá generar muy pronto un nuevo archivo con menos palabras erróneas o mejor dicho inútiles (es decir, las que no dan valor añadido para armar unos clusters). El documento no ha sido aún generado de nuevo porque es mejor hacerlo una vez la limpieza de datos totalmente finalizada.

Una vez terminemos de limpiar los datos, agruparemos los RRAA y las OOEE según estos grupos de palabras claves (nuevamente generados) para producir la primera versión de la aplicación de Visual Analytics que validaremos con el DANE en la primera semana de noviembre.

D. Estado del arte y Justificación de la Visualización

En esta parte queremos presentar unas técnicas y unos ejemplos de visualización adaptadas a las tareas que nos piden el DANE: (**Explore / Topology and Identify / Features (links))** con el fin tanto de presentar lo existente como de justificar que la visualización que proponemos corresponde con las tareas que se quieren hacer.

Forces and force in a box:

Cuando se usa una visualización en network para resolver estas tareas, primero, una técnica recurrente y casi imprescindible es el uso de fuerzas para organizar los nodos (siempre usado para visualización en network: fuerza de colisión, fuerza en x o y etc.). Segundo, otra implementación posible para completar esta agrupación en clúster, es obligar a que los nodos estén en un mismo lugar o mismo zona metiéndolos todos en una misma caja (force in a box). Gracias a esto, con la visualización se puede fácilmente hacer las tareas *explore topology (T1)* y *identify / features (links) (T2).*

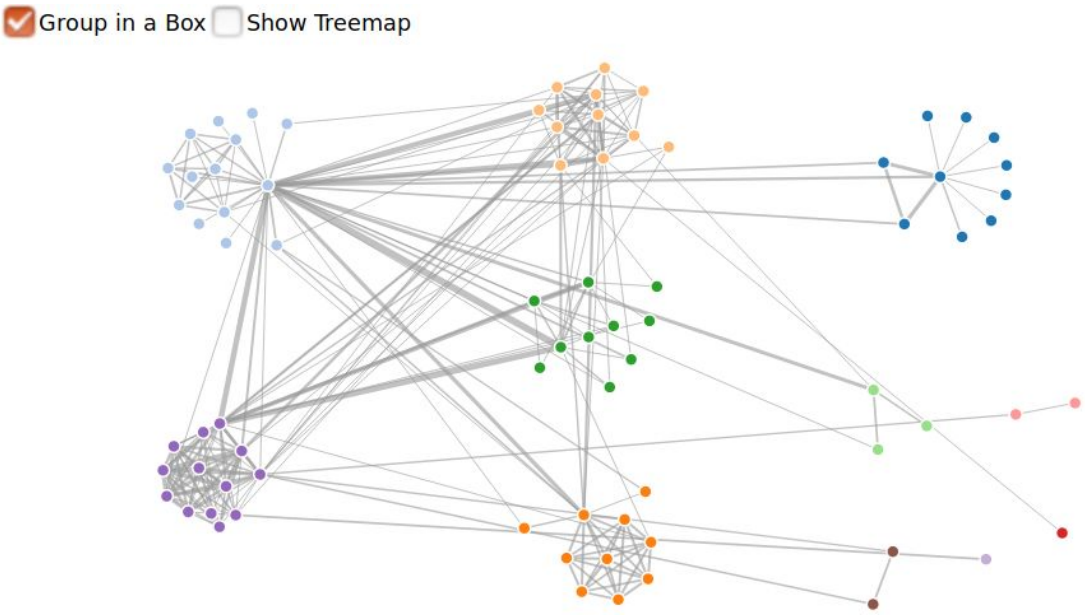

Figura 3: Visualización en network que muestra los clusters de los personajes de Les Misérables, basandose en las co-occurrencias por capitulo
(https://bl.ocks.org/john-guerra/14c943d8f198d9f3fef2/669304a9a47d0fb6dfff4060ceba13d35037fffb)

Otro ejemplo de visualización en network :

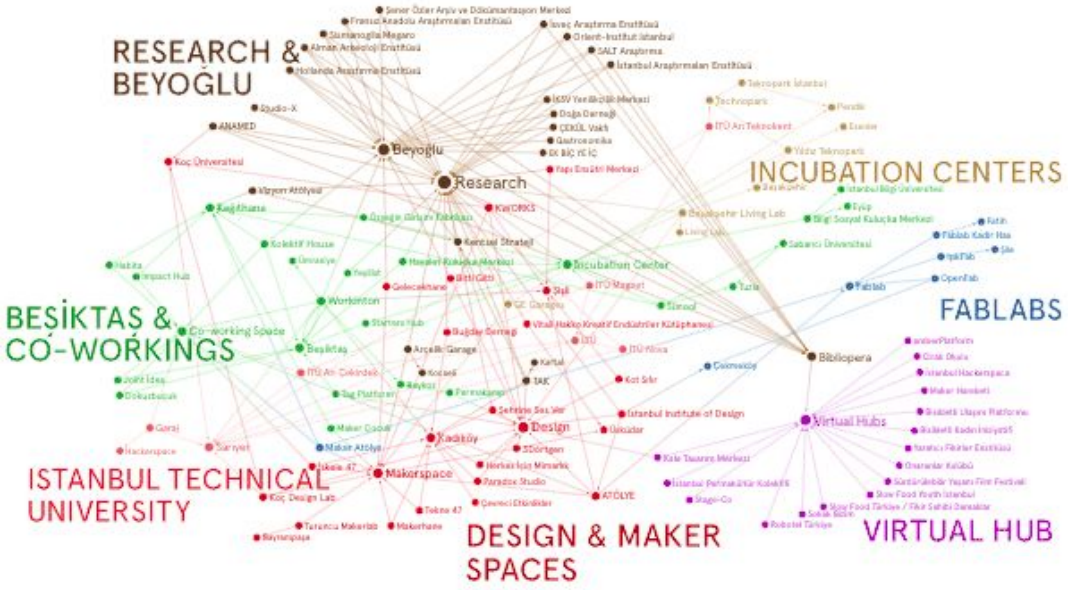

Figura 4: Visualización en network que muestra los clusters de sitios de investigacion en Turquia
https://medium.com/graph-commons/finding-organic-clusters-in-your-complex-data-networks-5c27e1d4645d

A propósito de esta visualización, un punto interesante es el uso de palabras que se pueden ver directamente en el grafo, lo cual permite, a parte de hacer las tareas principales T1 y T2 previamente descritas, de tener una parte de la información

contenida en los atributos de los nodos, antes de realizar la tarea T5 (Browse Features (nodes)).

Matriz de adyacencia

Otra técnica para representar los clusters es el uso de una matriz de adyacencia. Este tipo de visualización es muy útil para las tareas T1 (explore topology) pero no es la mejor para la tarea T2 (identify / features (links)) porque no muestra de manera explícita las relaciones entre los ítems.

Abajo presentamos un ejemplo de este tipo de visualización, que muestra los mismos resultados que en la Figura 3 con el mismo objetivo (explore the topology + identify the clusters of the characters).

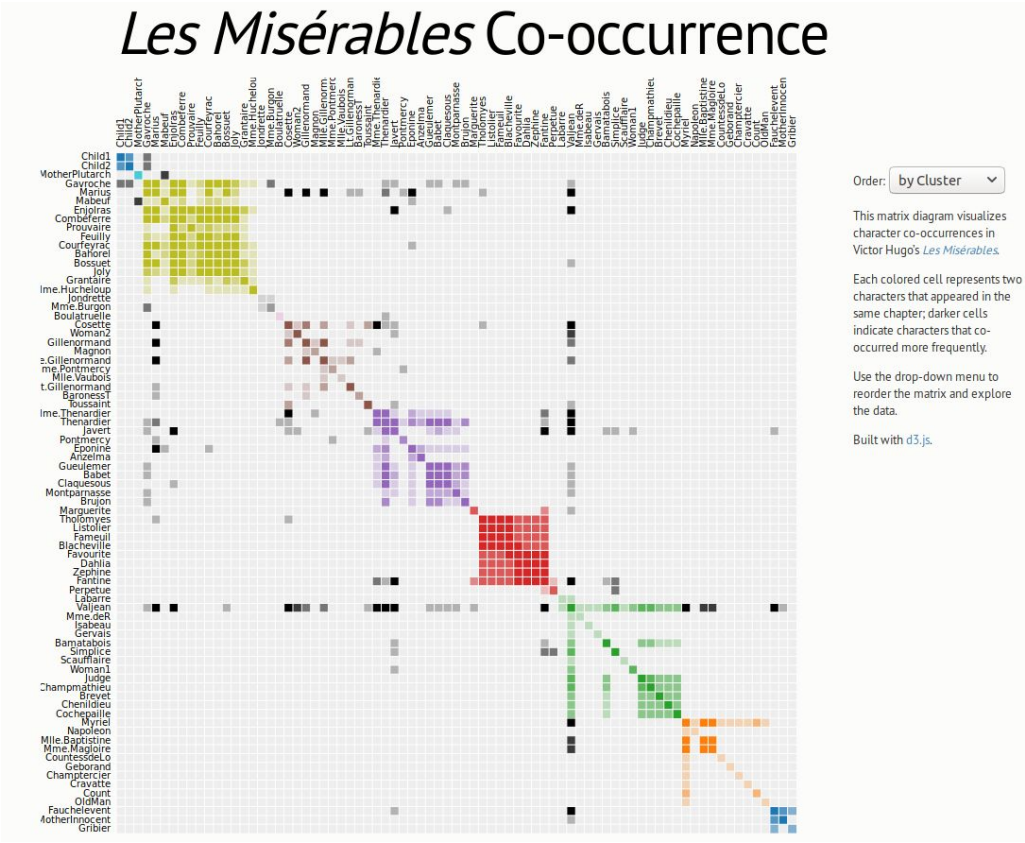

Figura 5: Matriz de adyacencia que muestra los clusters de los personajes de Les Misérables
https://bost.ocks.org/mike/miserables/

Este estado del arte nos indica que estos tipos de visualizaciones están adecuadas y adaptadas para poder realizar estos tipos de tareas: permite justificar la elección de visualización en network que hicimos, cumpliendo con las tareas T1 y T2 del proyecto.

E. Avances implementación visualización

Con el fin de poder hacer una primera implementación de la visualización, una vez la limpieza de datos hecha y los archivos del DANE exportados a CSV, el paso siguiente es procesar estos archivos para crear un nuevo archivo JSON cuya estructura debe ser la siguiente:

```
{
        "nodes": [
                {"nodeID": "XX", "type": "XXX", "group": "XXX"},
                …
                 ],
        "links": [
                {"linkID": "XX", "linkSource": "XX", "linkTarget": "XX"},
                 …
                ]
}.
```

Para poder hacer esto de manera fiable y automática, creamos un programa Java que lee los archivos CSV y cree a partir de estos dos ficheros un archivo txt de palabras clave y un archivo JSON que tiene todos los nodos y enlaces de la visualización. Este último archivo se cree usando tanto el archivo de palabras claves previamente creado como los archivos CSV.

Las etapas del programa Java son:

1. Leer cada línea del archivo CSV y construir un Map <palabra, número de ocurrencias>
2. Escribir los elementos con número de ocurrencias más alto en un archivo txt.
3. Crear una lista con las líneas *{"nodeID": "XX", "type": "XXX", "group": "XXX"}* para cada item
4. Crear una lista con las líneas *{"nodeID": "XX", "type": "XXX", "group": "XXX"}* para cada “cluster node”, es decir un nodo que tiene solamente como información el “cluster name” (nodos centrales de cluster, ver abajo la Figura 6)
5. Crear una lista con las líneas *{"linkID": "XX", "linkSource": "XX", "linkTarget": "XX"},* para cada enlace
6. Leer estas 3 listas y escribir sus líneas en un archivo JSON final.

Una vez creado el archivo JSON, se puede usar para crear una visualización en network. Por ejemplo, en d3, se puede usar una combinación de fuerzas como force charge, force link y force center, lo cual permite organizar los nodos espacialmente (channel Spatial position). Al nivel del channel Color hue se puede usar una escala ordinal de colores como schemeCategory10. Y por fin hay que notar que también se puede mostrar los nombres de clúster encima de los nodos, los cuales son puntos o formas de círculos (marca) y líneas para los enlaces (marca). A continuación pueden ver ejemplos de visualización, usando archivos JSON creados con el programa Java.

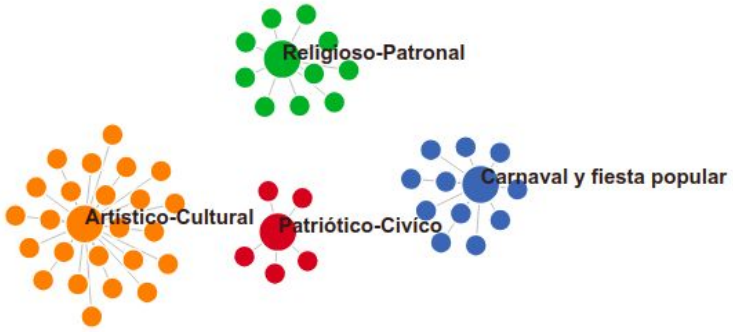

Figura 6 Ejemplo de visualización en network con archivo JSON nodes/links, por grupo temático

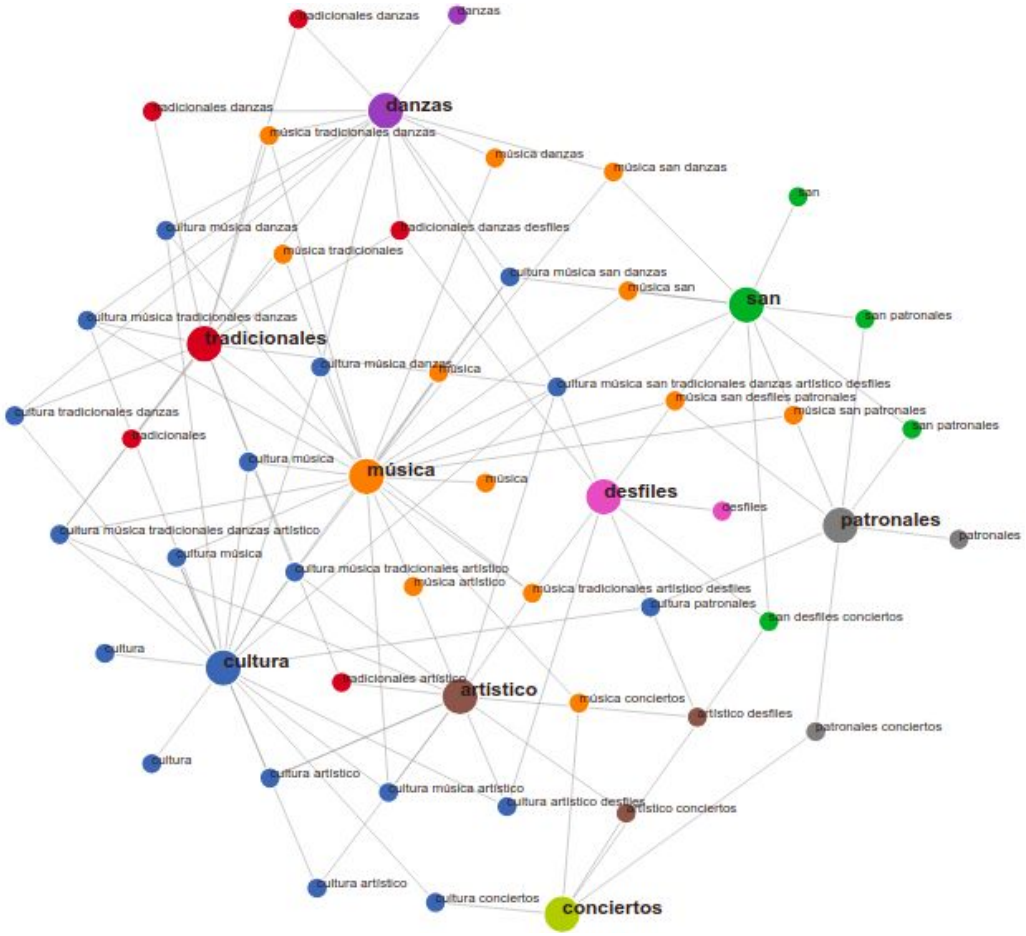

Figura 7 Otro ejemplo de visualización en network con archivo JSON nodes/links, por grupo temático

Estos ejemplos, como lo pueden ver, utilizan otros datos pero entonces los pasos seguidos aquí son los mismos que los que están descrito arriba. De hecho, para hacer la primera visualización con los datos del DANE, para este proyecto, los pasos que quedan por terminar o por hacer en adelante son: terminar la limpieza y preparación de datos, realizar la fusión final de los archivos Excel y pasarlo a CSV, crear el archivo JSON gracias al programa Java (el programa mismo necesita limpieza de código, documentación y refactoring) y después crear unas visualizaciones de pruebas con estos datos adaptados a las visualizaciones en network.

4. Archivos Adjuntos

En el archivo zip mandado y en el repositorio Github del proyecto (https://github.com/pierreraimbaud/DANEColombiaStatisticOperationsAdminRegisters) se encuentran los siguientes archivos con los avances:

A. Reporte de avance (este mismo documento)
B. Resultados preliminares del procesamiento de texto para obtener las palabras clave (PreliminaryKeywordsResult.txt)
C. Código fuente del procesador de texto en Java (carpeta KeywordsProcessor)