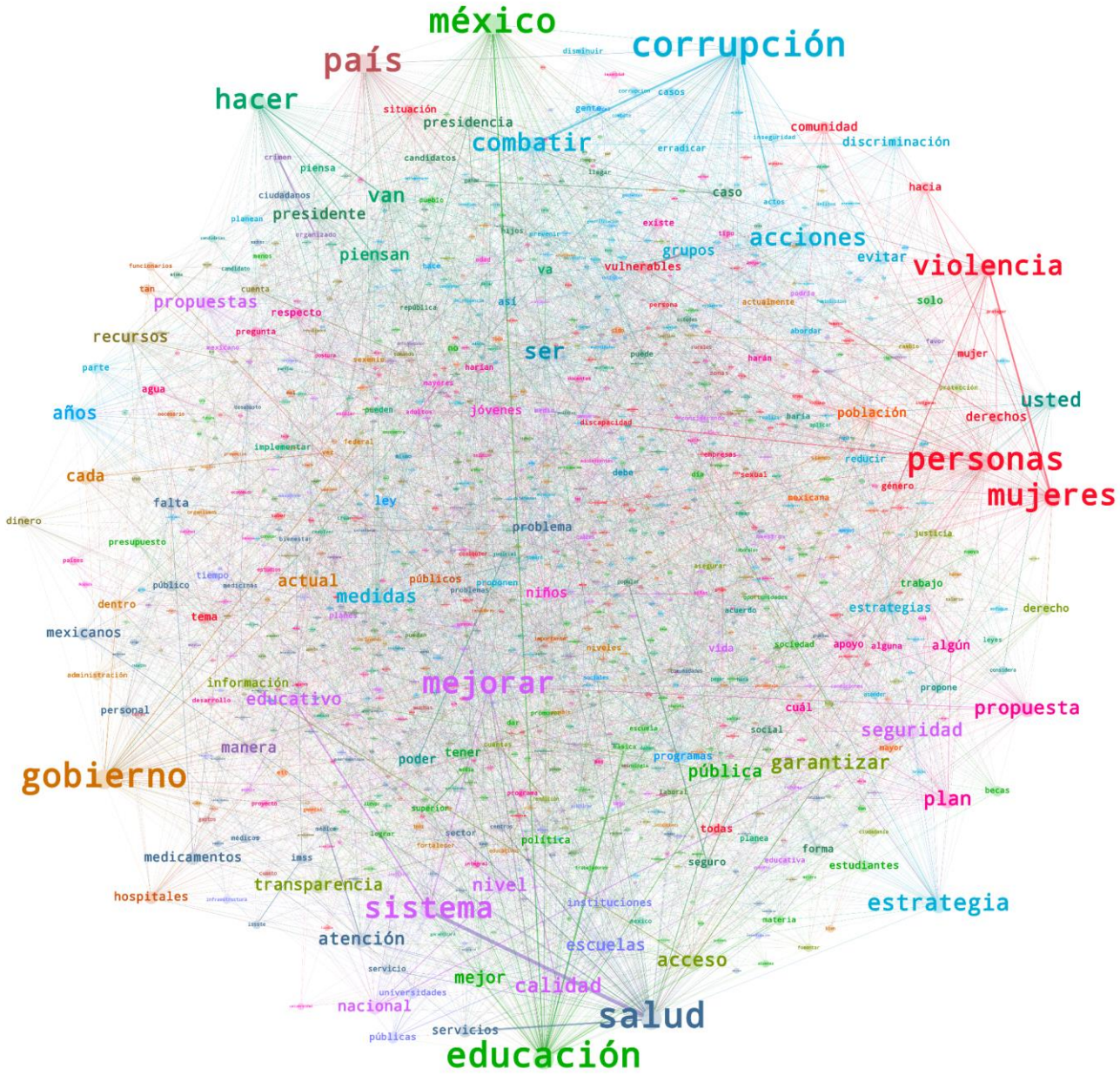
INFORME SOBRE EL PROCESO DE SELECCIÓN DE PREGUNTAS PARA EL 1ER DEBATE PRESIDENCIAL INE 2024 (FORMATO A)



ITESO, Universidad Jesuita de Guadalajara | signa_lab experimentación • creación

Informe sobre el Proceso de Selección de Preguntas de Redes Sociales para el 1er Debate Presidencial INE 2024 (Formato A)

Realizado por el equipo de **Signa_Lab ITESO**

Coordinación del proyecto
Víctor Hugo Ábrego Molina

Supervisión del desarrollo tecnológico, documentación y visualización de datos
Diego Arredondo Ortiz

Coordinación operativa, supervisión de desarrollo de diccionarios y análisis semántico
Paloma López-Portillo Vázquez

Programación de cuadernos de código para limpieza, depuración y análisis exploratorio de datos, análisis semántico y selección aleatoria
Javier de la Torre Silva y José Luis Almendarez González

Diseño muestral y asesoría en estadística aplicada
Radamanto Portilla Tinajero

Asesoría en análisis semántico y exploración de datos
Luz María Sandoval Zavala y Héctor Piña Camacho

Asistencia técnica, exploración de datos y visualización de grafos
Eduardo G. de Quevedo Sánchez

Desarrollo de diccionarios, pruebas de depuración y análisis semántico
Ana Sánchez Muñoz, Daniela Hernández Ramírez, Fernanda Verduzco Hernández y Vanessa Briseño Ramos

Documentación del proceso y realización de bitácoras
Víctor Hugo Ábrego Molina, Diego Arredondo Ortiz, Vanessa Briseño Ramos y Fernanda Verduzco Hernández

Diseño de metodología y supervisión de su cumplimiento (CNCS INE)
Ana Cristina Levy Covarrubias y Zaira Ivonne Medina Gómez

Supervisión desde Oficialía Electoral del INE
Irene Maldonado Cavazos, Adrián Sánchez Sáez y Pedro Alejandro Mendoza Carretero

10 de abril de 2024.
Instituto Tecnológico y de Estudios Superiores de Occidente (ITESO)
Tlaquepaque, Jalisco, México.

ÍNDICE

1. Resumen ejecutivo 04

2. Metodología aplicada por Signa_Lab ITESO 06

3. Resultados, hallazgos y entregables 26

4. Conclusiones 47

5. Lista de Referencias 49

6. Directorio de Anexos 50

1. RESUMEN EJECUTIVO

Para el primer Debate Presidencial del domingo 7 de abril de 2024, organizado por el Instituto Nacional Electoral (INE) y titulado “La Sociedad que Queremos”, se dispuso un formulario digital para que la ciudadanía pudiera realizar preguntas acerca de cada uno de los temas propuestos para el mismo. Para el procesamiento y selección de preguntas a considerar por las personas moderadoras del Debate, el INE diseñó una metodología específica para el Formato A de este primer debate (INE, 2024) y designó a una institución académica externa, Signa_Lab ITESO, para obtener 18 preguntas seleccionadas por frecuencia, representativas de cada tema, y 90 aleatoriamente, distribuidas por cada región del país (Norte, Centro y Sur), que suman un total de 108 preguntas.

Para llevar a cabo este proceso, se definió un plan de trabajo por etapas que se resume a continuación:

Etapa 0. El INE entregó a representantes del ITESO una base de datos, en dos archivos de formato Excel, con los 13,484 registros que contenían las 24 mil preguntas que el instituto recolectó en su sitio web durante el mes que estuvo abierta la convocatoria para que los ciudadanos y ciudadanas pudieran formular sus preguntas. Esta base de datos fue transportada a Guadalajara, donde, bajo la supervisión de expertos en ciberseguridad del ITESO, se aseguró su integridad y se descargó en sistemas aislados de la red, siguiendo las estrictas medidas de seguridad solicitadas por el INE.

Etapa 1. Esta fase estuvo dedicada a que el equipo revisara y depurara la base de datos. Para llevar a cabo este proceso se realizaron dos acciones. La primera consistió en el desarrollo de un diccionario de términos proscritos, elaborado por el equipo de Signa_Lab ITESO en atención a los criterios del INE, que ayudara a identificar preguntas que contenían lenguaje ofensivo o bien sesgos políticos, para cumplir con los criterios de elegibilidad definidos por el INE. El diccionario final estuvo compuesto por 519 de estos términos; como resultado de este proceso, se eliminaron 1,117 preguntas. La segunda fase fue utilizar código informático que permitiera detectar aquellas preguntas cuya redacción fuera idéntica, para catalogarlas como repetidas. A partir de la utilización de este código se descartaron 1,664 preguntas duplicadas. Cabe aclarar que se eliminaron las repeticiones, pero sí se incluyó una instancia de cada pregunta duplicada para ser tomada en cuenta para los siguientes procesos. Mediante estas dos fases, el equipo depuró eficazmente la base de datos, lo cual arrojó un total de 21,219 preguntas libres de términos proscritos y duplicados.

Etapa 2. El propósito de la segunda etapa fue obtener una muestra estratificada para la selección de preguntas. El formulario creado por el INE para la captura de preguntas de los ciudadanos se basó en dos criterios esenciales: temático y territorial. En el

aspecto temático, los participantes seleccionaron uno de los temas sugeridos para el debate, y pudieron formular una pregunta pertinente a esa categoría. Respecto al criterio territorial, debían especificar la entidad del país desde la cual realizaron su consulta, la cual fue categorizada posteriormente por el INE como perteneciente a región Norte, Centro o Sur. El equipo de Signa_Lab ITESO implementó una fórmula estadística que arrojó una muestra estratificada por tema y por región compuesta por 1,701 preguntas, que representan fielmente la distribución por región y por tema de las 21,219 preguntas depuradas en la etapa anterior.

Etapa 3. En esta etapa se llevó a cabo un ejercicio computarizado, a través de herramientas de inteligencia artificial y de lingüística de corpus, para la preselección de preguntas. El proceso incluyó el desarrollo de un algoritmo que permitió la identificación de similitud semántica entre las preguntas de la muestra estratificada. Este algoritmo analizó 1,024 dimensiones dentro de cada pregunta y con ello, las agrupó en clústeres a partir de sus similitudes. Como resultado de este trabajo, se extrajeron 18 preguntas preseleccionadas por frecuencia y 90 preguntas preseleccionadas aleatoriamente.

Etapa 4. En la cuarta etapa, el equipo de Signa_Lab ITESO realizó una revisión manual de las 108 preguntas seleccionadas, con el acompañamiento y supervisión del cumplimiento de criterios de elegibilidad por parte de la representación del INE. Durante la primera ronda de revisión, se identificó que 28 preguntas tenían errores de coherencia argumentativa, de sintaxis, de neutralidad y/o de pertinencia temática. Estas características están claramente señaladas como criterios de invalidación en la metodología del INE, por lo que se procedió a su eliminación y reemplazo por otras de la muestra estratificada por tema y región. Es importante señalar que los motivos de reemplazo de las preguntas no se debieron a errores en el proceso de depuración, sino a fallos y sesgos de origen en el propio registro de la ciudadanía. En la revisión subsiguiente de las nuevas 28 preguntas, 11 aún contaron con alguno de estos criterios para ser invalidadas. Un tercer y cuarto ejercicio de revisión resultaron en la eliminación de dos y una pregunta, respectivamente, por contener alguna de estas cualidades. La tasa total de reemplazo fue del 2.47% en relación con las 1,701 preguntas de la muestra estratificada.

Etapas 5 y 6. Estas etapas contemplan el seguimiento y cumplimiento de cada una de las anteriores en tiempo y forma desde las instalaciones de Signa_Lab ITESO, acompañadas y validadas por la representación del INE durante los días que duró el procesamiento de las preguntas, y la realización y entrega del presente informe, 9 días posteriores a la entrega de las preguntas.

Hallazgos y conclusiones. La y los jóvenes de los 13 a los 27 años fueron el grupo que más participó en este ejercicio, en contraste, los adultos de 68 años en adelante fueron

el grupo con un menor registro de participación. Los temas de salud y educación fueron abordados en términos de diagnóstico deficitario de las condiciones actuales en ambos rubros; la corrupción fue percibida como un problema en todos los niveles de gobierno y el cual no se puede atender sin pensar en estrategias que lidien con el crimen organizado; la violencia en contra de las mujeres resaltó por formas concretas y extremas como los feminicidios y la trata de personas; preocupó también la falta de reconocimiento de las condiciones de desigualdad que viven personas de la diversidad sexual, personas discapacitadas, grupos indígenas y mujeres en el país; las inquietudes acerca de la transparencia giraron alrededor de la rendición de cuentas y el acceso a la información garantizada por parte de los gobiernos.

2. METODOLOGÍA APLICADA POR SIGNA_LAB ITESO[1]

Etapa 0. Entrega de base de datos

El 22 de marzo del 2024, a las 9:00 horas, se llevó a cabo el acto protocolario de entrega de la base de datos en las Oficinas Centrales del INE en la Ciudad de México. En el evento, el INE entregó a Signa_Lab ITESO, vía autoridades del ITESO, las preguntas ciudadanas en una memoria USB. En representación del ITESO, asistieron Humberto Orozco Barba, director de la Oficina de Relaciones Externas; Juan Larrosa Fuentes, Director del Departamento de Estudios Socioculturales[2]; Bernardo Masini Aguilera, Director de Investigación y Posgrado; y Juan Diego Vázquez Valle, Coordinador de Seguridad Informática.

Después del acto protocolario, los funcionarios de ITESO, acompañados por Irene Maldonado Cavazos, Directora de la Oficialía Electoral del INE, y Zaira Ivonne Medina Gómez, Líder de Proyecto de Información en Educación Cívica y Género del Instituto Electoral, volvieron a las instalaciones del ITESO para documentar y dar fe de la entrega de la base de datos con las preguntas recabadas. Signa_Lab ITESO resguardó y respaldó localmente la base de datos entregada por el INE en equipos del laboratorio, con la supervisión de Abraham Carrillo y Guillermo López, del Área de Ciberseguridad de la Oficina de Sistemas de Información del ITESO, quienes revisaron que los códigos de integridad de los archivos fueran los mismos entregados en Ciudad de México.

El equipo de Signa_Lab ITESO se presentó ante las autoridades del INE, quienes hicieron lo propio. Luego de abrir el archivo y validar el número de registros y de

[1] Todas las capturas, gráficas y tablas de esta sección fueron generadas por el equipo de Signa_Lab ITESO, y están documentadas en los Anexos correspondientes.
[2] Departamento al que está adscrito Signa_Lab ITESO.

preguntas entregado por el Instituto, concluyó de manera oficial la etapa de recepción de la base de datos.

Etapa 1. Preparación de la base de datos

Esta etapa consistió en la depuración de las preguntas entregadas por el INE. El equipo de Signa_Lab ITESO desarrolló un diccionario de términos proscritos y utilizó un código informático para la identificación de preguntas repetidas. Las actividades de la metodología contempladas en esta etapa fueron:

- ***Act. 2. Condensar aquellos registros que por su semántica o sintaxis correspondan a una misma pregunta.***
- ***Act. 3. Desechar las preguntas que no cumplan con los criterios de redacción establecidos, como: temática seleccionada, sesgo partidista, lenguaje ofensivo, discurso de odio, discriminación y/o violencia de cualquier tipo.***

Separar registros de preguntas

Previo a la depuración de la base de datos, fue necesario modificar el formato en el que se entregaron las preguntas, ya que dentro de los 13,484 registros (entradas al formulario) estaban contenidas las 24 mil preguntas, por lo que se separaron las preguntas de los registros. Para enlistar el total de preguntas, se generó una nueva columna con identificadores únicos por pregunta (*ID*), con una referencia al registro de origen y el número de pregunta correspondiente en el mismo. Esto permitió obtener un archivo en Excel de 24 mil filas, con el que se trabajó a lo largo de todo el proyecto. El conjunto de datos con la población completa de 24 mil preguntas recibidas se encuentra en los anexos del presente informe.

Depuración de preguntas por uso de términos proscritos

Signa_Lab ITESO desarrolló un diccionario de términos proscritos para la depuración de preguntas para el debate presidencial. Este diccionario integró elementos de un lexicón de términos ofensivos[3] previamente elaborado por el laboratorio (Signa_Lab, 2022), diccionarios externos[4] y términos ofensivos extraídos de conjuntos de datos masivos de redes sociales del repositorio de descargas de Signa_Lab ITESO, como parte del trabajo continuo de monitoreo y análisis del laboratorio.

El diccionario se dividió en términos ofensivos y en términos referidos a sesgos partidistas (menciones a candidaturas, partidos políticos y presidentes de partidos), ideológicos y de religión. Las pruebas llevadas a cabo por el laboratorio antes de recibir las preguntas[5] permitieron afinar la lista y ampliar las variaciones de varios de estos

[3] El *Lexicón de Twitter*, diccionario elaborado por Signa_Lab ITESO mediante técnicas mixtas de exploración de datos de Twitter, permite la identificación y categorización de formas discursivas de violencia digital de género, en distintas variaciones del idioma español en México e Iberoamérica.

[4] Guzmán Falcón, E. (2018). Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones (Tesis de Maestría). Instituto Nacional de Astrofísica, Óptica y Electrónica. Recuperado de https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/1722/1/GuzmanFE.pdf

[5] En las semanas previas a la recepción de la base de datos, Signa_Lab ITESO llevó a cabo pruebas de depuración de preguntas con versiones previas del diccionario de términos proscritos. Para la

términos. La versión final del diccionario utilizada con la base de datos de preguntas para el debate contuvo un total de **519 términos proscritos.**

El refinamiento del diccionario de términos proscritos llevó a que en la base de datos con las preguntas para el debate se identificaran, por ejemplo, 23 variaciones del apellido Sheinbaum, 25 variaciones del nombre Xóchitl y 33 variaciones de las siglas de partidos políticos. La eliminación final de preguntas por utilización de términos proscritos alcanzó un nivel de exhaustividad en el cual se llegaron a ubicar casos concretos de uso de signos de puntuación inmediatamente antes o después de algunas palabras, y fueron integradas como palabras nuevas al diccionario, así como errores ortográficos inusuales (por ejemplo: “Nombre,,no”), pero ya no patrones (es decir, más de un caso similar) en preguntas descartables. La lista completa de palabras del diccionario se encuentra en los anexos de este informe.

Depuración de preguntas por duplicidad en su redacción

Para la detección de preguntas que por su semántica o sintaxis correspondieran a una misma, como solicitaba la metodología definida por el INE para evitar repeticiones (INE, 2024), se diseñó un algoritmo para evaluar las duplicidades tanto por el conteo de palabras repetidas como por el orden de las mismas en la redacción de cada pregunta. De esta manera, se hicieron pruebas con distintos umbrales de similitud (80%, 85%, 90% y 100% de similitud), con el fin de evaluar los resultados de la depuración en cada escenario y de optimizar progresivamente el funcionamiento del algoritmo.

Finalmente, se decidió junto con el personal del INE ajustar el umbral del algoritmo al 100% para identificar y descartar sólo preguntas con coincidencias exactas respecto a otras, para así cumplir con la instrucción explícitamente solicitada en la metodología que indicaba evitar repeticiones. Por cada pregunta identificada con duplicidades, se contabilizó el número de preguntas repetidas y se registró el identificador único (ID) de la pregunta de referencia con la que se detectó la repetición. Con estas categorizaciones se mantuvo una instancia en la población depurada y el resto se descartaron. El código utilizado para dicho proceso y los resultados del mismo se integran en anexos del presente informe.

Resultados de la depuración

Los resultados finales de la etapa de depuración fueron los siguientes:

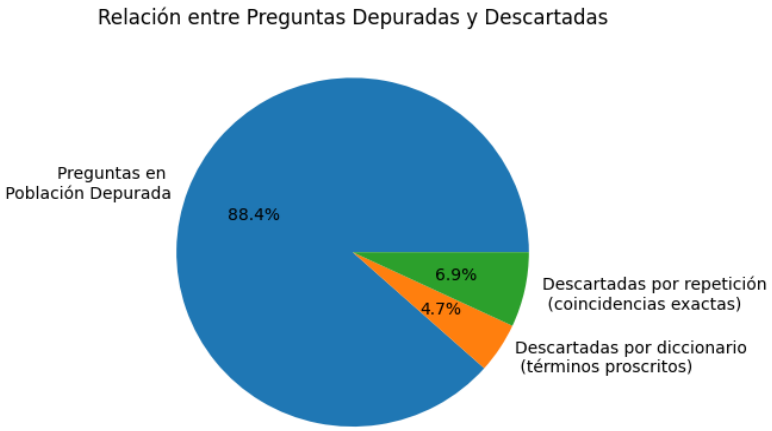
- *1,117 preguntas descartadas por uso de términos proscritos.*
- *1,664 preguntas descartadas por repeticiones.*

identificación de éstos en preguntas reales, el laboratorio utilizó muestras de preguntas del tercer debate presidencial del 2018, con las que el laboratorio ya había publicado un informe disponible en: https://signalab.iteso.mx/informes/informe_3erdebateine.html

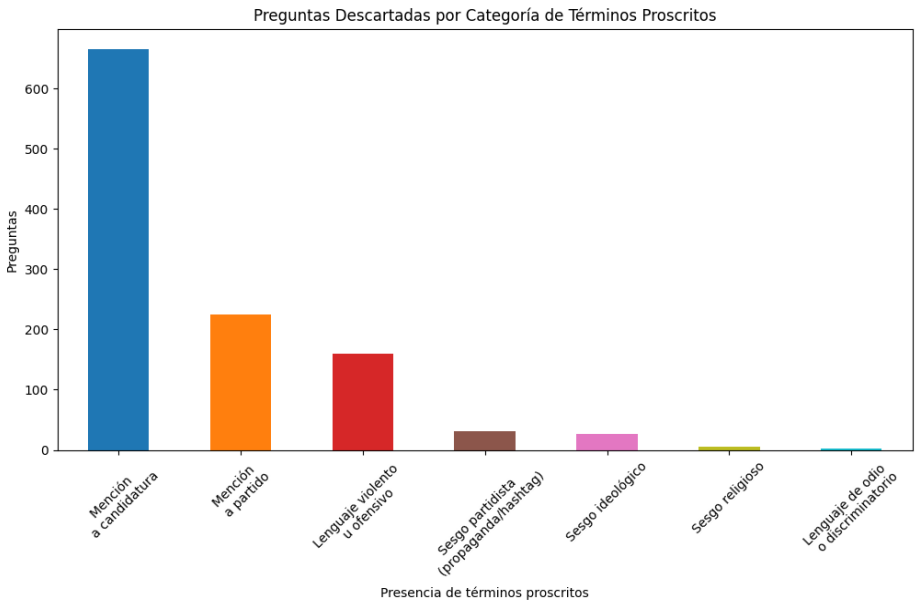
- *21,219 preguntas depuradas finales.*

Este proceso permitió descartar el 11.6% (2,781) de las preguntas iniciales, por incumplir con los criterios de redacción estipulados por el INE, o bien, por tener una redacción idéntica (100% de similitud). Al final, permaneció 88.4% de las preguntas en la población depurada para las etapas posteriores.

Como se solicitó por parte del INE, el razonamiento detrás del descarte de cada pregunta en esta etapa fue debidamente registrado de manera individual por cada una de las 2,781 preguntas eliminadas. El conjunto de datos con las preguntas eliminadas y su respectivo razonamiento pueden encontrarse en anexos del presente informe.



Aun cuando en las instrucciones del formulario publicado por el INE se hicieron explícitos los criterios de descarte de una pregunta, la principal causa de depuración por diccionarios fue la mención de candidaturas, seguido de las menciones a alguno de los partidos y, en tercer lugar, por la utilización del lenguaje violento u ofensivo.



A partir de las dos etapas de depuración inicial mencionadas, por términos proscritos y por repeticiones, se consolidó una población depurada de 21,219 preguntas considerables para ser seleccionadas en la muestra estratificada que se detalla en la siguiente sección. Cabe mencionar que, en las siguientes etapas de selección y revisión, no aparecieron términos proscritos ni repeticiones exactas, lo cual da cuenta de la eficacia con la que se realizó la depuración de preguntas en esta primera etapa.

Etapa 2. Obtención de la muestra estratificada y clasificación por tema y por región

Esta etapa consistió en la obtención de una muestra estratificada de 1,701 preguntas, representativas proporcionalmente por tema y por región[6] de las 21,219 preguntas depuradas de la fase anterior. La fórmula utilizada para obtener esta muestra operó con un nivel de confianza de 99% y un margen de error de 3%. Las actividades contempladas en esta etapa fueron:

- ***Act. 4. Determinación del tamaño de la muestra estratificada***
- ***Act. 5. Obtención de la muestra estratificada por región-tema***

Propuesta de una doble estratificación

De acuerdo con la metodología propuesta por el INE, la selección de la muestra debería utilizar un procedimiento de estratificación a partir de los **temas definidos para el debate**, con el objetivo de asegurar la "representatividad en la selección sin importar las diferencias en la cantidad de preguntas que lleguen sobre un tema con respecto a los otros, [por tanto] la muestra resultante sea autoponderada" (INE, 2024). Por otro lado, también se solicitó que "la muestra obtenida se divida o clasifique en 3 grupos, dependiendo de la región" de procedencia de las preguntas. Finalmente, de la muestra estratificada se debía obtener un conjunto de 90 preguntas (cinco para cada tema y región), seleccionadas de manera aleatoria, y 18 preguntas (tres para cada tema) a partir del criterio de prevalencia.

El equipo de Signa_Lab ITESO realizó ejercicios de simulacro para el cálculo y la selección de la muestra considerando los criterios solicitados por la metodología del INE. Para estos simulacros, se utilizaron dos bases de datos con preguntas del tercer debate presidencial del 2018[7] que se enfocaron en temas sugeridos para este debate:

[6] ***Región Norte***: Baja California, Baja California Sur, Chihuahua, Coahuila, Durango, Nayarit, Nuevo León, Sinaloa, Sonora, Tamaulipas y Zacatecas. ***Región Centro*:** Aguascalientes, Ciudad de México, Colima, Estado de México,Guanajuato, Hidalgo, Jalisco, Michoacán, Morelos, Querétaro, San Luis Potosí y Tlaxcala. ***Región Sur*:** Campeche, Chiapas, Guerrero, Oaxaca, Puebla, Quintana Roo, Tabasco, Veracruz y Yucatán.

[7] Signa_Lab ITESO recogió y filtró las preguntas realizadas por la ciudadanía en redes sociales para el tercer debate presidencial en 2018, por lo que utilizó y segmentó una muestra de esas preguntas, con regiones simuladas de manera aleatoria con los dos temas (educación y salud) que se repitieron en dicho

salud y educación, de modo que se contara con un par de muestras con semejanzas al que sería entregado por el INE, uno con 3,734 registros y otro con 7,094 registros. Como resultado se llegó a la siguiente conclusión:

- La selección aleatorizada de las unidades muestrales (las preguntas), estratificadas solamente por la variable "Tema", genera sobre o subrepresentación de la variable "Regiones" e incrementa la probabilidad de que no se obtengan las unidades (preguntas) mínimas requeridas para el conjunto de 90 preguntas (cinco para cada tema y región), seleccionadas aleatoriamente.

Frente a ese escenario, Signa_Lab ITESO propuso que la selección de la muestra se realizara mediante un procedimiento bietápico que considerara dos niveles de estratificación (Región y Tema), de tal manera que se garantizara la representatividad proporcional con respecto a la población depurada (conjunto total de preguntas seleccionables). La propuesta fue aceptada por la representación del INE y se implementó para el cálculo del tamaño de la muestra y la selección de las unidades muestrales. Como resultado, se obtuvo una muestra estratificada y representativa de la población de preguntas, con variaciones máximas de 0.1% en las distribuciones proporcionales de preguntas para cada tema y región.

Previsiones ante un posible escenario donde no se contara con las unidades suficientes para alguna región y tema

Debido a la variabilidad esperada en las distribuciones proporcionales de las preguntas para cada tema y región –es decir, que la mayoría de las preguntas se concentraran en unos cuantos temas y regiones, lo que provocaría que el resto de temas y regiones tuvieran muy poca participación de preguntas–, existía la posibilidad de que el cálculo del tamaño de la muestra estratificada no proporcionara la cantidad mínima necesaria de preguntas para cada estrato (cinco por tema y región), por lo que el equipo de Signa_Lab ITESO preparó cuatro posibles soluciones:

a. La primera era incrementar el tamaño de la muestra, cambiando los parámetros de la fórmula de cálculo (nivel de confianza y/o margen de error), con lo que se mantenía la distribución proporcional por estratos, hasta obtener las unidades muestrales mínimas requeridas para la lista de 90 preguntas seleccionadas de manera aleatoria solicitada por el INE.
b. La segunda era incrementar la representatividad de las regiones-temas que tuvieran menor participación, lo que provocaría una sobrerrepresentación de estos estratos.
c. La tercera consistía en igualar aritméticamente las preguntas de las tres regiones en la muestra estratificada, lo cual implicaría redistribuir 33% la

debate de 2018 y en el presente debate de 2024. El informe acerca del debate de 2018 está disponible en: https://signalab.iteso.mx/informes/informe_3erdebateine.html

representatividad para cada región, independientemente de su porcentaje en la base de datos de preguntas depuradas.

d. La última solución consistía en seleccionar las 90 preguntas (cinco por región y tema) de la base de preguntas depuradas y utilizar la muestra estratificada solamente para la obtención de las 18 preguntas mediante el criterio de prevalencia.

Signa_Lab ITESO puso a consideración de la representación del INE estas cuatro posibles soluciones y se llegó al acuerdo de utilizar la primera, en caso de que fuera necesario.

Implementación de la fórmula para el cálculo del tamaño de la muestra estratificada por tema-región

El cálculo del tamaño de la muestra estratificada se realizó mediante la fórmula de estimación de proporciones, la cual es consistente con las variables "Tema" y "Región" utilizadas para la estratificación. La fórmula de cálculo es:

$$n = \frac{NP(1-P)}{(N-1)\left(\frac{\delta^2}{\left(Z_{1-\frac{\alpha}{2}}\right)^2}\right) + P(1-P)}$$

Donde el límite para el error de estimación (δ) para poblaciones finitas es:

$$\delta = Z_{1-\frac{\alpha}{2}}\sqrt{\left(\frac{N-n}{N}\right)\frac{P(1-P)}{n}}$$

Con un nivel de confianza del 95% y un margen de error de 5%, el tamaño de la muestra fue de 389 preguntas, distribuidas proporcionalmente por cada región y tema. La distribución proporcional de la base de datos depurada permitió obtener la cantidad mínima de preguntas requeridas por el INE (cinco por tema y región). Sin embargo, en el caso de la región y tema que tuvo la menor participación (región Norte, tema *Violencia en contra de las mujeres*), con estos criterios establecidos en la fórmula de cálculo solamente se obtenían cinco preguntas, lo que limitaba la posibilidad de hacer reemplazos, en caso de ser necesario, durante la última fase de revisión de las preguntas obtenidas.

Resultados de primer cálculo de muestra estratificada, con parámetros originales:

tema	Combate a la corrupción	Educación	No discriminación y grupos vulnerables	Salud	Transparencia	Violencia en contra de las mujeres	Total por Región
region							
centro	53	55	29	51	32	25	245
norte	16	15	7	13	8	5	64
sur	19	19	9	16	10	7	80
Total por Tema	88	89	45	80	50	37	389

Con el objetivo de incrementar el número de preguntas en las regiones y temas con menor participación, y así tener un mayor número de elementos para posibles reemplazos, con la autorización de la representante del INE, se decidió incrementar el nivel de confianza al 99% y reducir el margen de error al 3%. Con estos criterios, el tamaño de la muestra fue de 1,701 preguntas, distribuidas proporcionalmente por cada estrato de tema-región. La región y tema con el menor número de unidades muestrales fue Región norte, tema *Violencia en contra de las mujeres*, con un total de 20 preguntas; seguida de la Región Sur y el tema *Violencia en contra de las mujeres*, con 28 preguntas. Los temas *Educación*, *Combate a la corrupción* y *Salud* en la Región Centro, fueron los que contaron un mayor número de unidades muestrales, con 243, 236 y 225 respectivamente.

Captura de la implementación en el código de la fórmula implementada con parámetros ajustados:

```
# Definir parámetros y función para cálculo de tamaño de muestra
Z=2.58 # Nivel de confianza: 99%
p=0.5 # Proporción que presenta el atributo
q=1-p # Complemento de "p"
N=len(df_filtered) # Población. Se debe ajustar según el total de la base depurada
e=0.03 # Error de estimación: 3%

# Fórmula para calcular tamaño de muestra
tam=math.ceil((Z**2*N*p*q)/(e**2*(N-1)+Z**2*p*q))
```

Distribución de preguntas por tema y región en población depurada:

tema	Combate a la corrupción	Educación	No discriminación y grupos vulnerables	Salud	Transparencia	Violencia en contra de las mujeres	Total por región
region							
centro	2939	3030	1574	2811	1789	1346	13489
norte	861	831	380	724	418	249	3463
sur	1052	1011	450	872	529	353	4267
Total por tema	4852	4872	2404	4407	2736	1948	21219

Resultados del cálculo definitivo aplicado a la muestra estratificada, con parámetros:

tema	Combate a la corrupción	Educación	No discriminación y grupos vulnerables	Salud	Transparencia	Violencia en contra de las mujeres	Total por Región
region							
centro	236	243	126	225	143	108	1081
norte	69	67	31	58	34	20	279
sur	84	81	36	70	42	28	341
Total por Tema	389	391	193	353	219	156	1701

Etapa 3. Selección de preguntas

A partir de la utilización de herramientas de Inteligencia Artificial y de lingüística de corpus, en esta etapa se llevó a cabo el procesamiento, identificación y categorización de relaciones semánticas dentro de las 1,701 preguntas de la muestra estratificada por tema y por región, necesarios para obtener la preselección de 18 preguntas por frecuencia y 90 preguntas por aleatoriedad. Las actividades contempladas para esta etapa fueron:

- ***Act. 6. Selección del primer bloque de tres preguntas con mayor frecuencia sobre cada uno de los temas establecidos para el debate.***
- ***Act. 7. Selección aleatoria del segundo bloque de preguntas por región a partir de la muestra estratificada por tema.***

Utilización de herramientas de inteligencia artificial para la identificación de relaciones semánticas

Debido a la pluralidad de expresiones en las preguntas de la ciudadanía, donde incluso algunas con redacciones muy distintas podrían considerarse como preguntas equivalentes, el equipo de Signa_Lab ITESO propuso implementar modelos de lenguaje, basados en inteligencia artificial, para el cálculo de similitud semántica y la identificación de clústeres representativos de las preocupaciones ciudadanas con mayor frecuencia por tema.

Para cumplir con la solicitud del INE de mantener el procesamiento a nivel local, en los equipos del laboratorio, y garantizar la transparencia, trazabilidad y replicabilidad del ejercicio, se propuso trabajar con modelos de lenguaje de software libre y librerías de programación en Python, especializadas para la detección de similitud entre oraciones[8] (Reimers, N. & Gurevych, 2019) y ejecutables localmente.

Después de la implementación de prueba con distintos modelos de lenguaje preentrenados, se optó por el modelo de reciente creación *multilingual-e5-large-instruct*[9] el cual produjo mejores resultados, atribuibles a su mayor densidad semántica (1024 dimensiones), a la presencia considerable de contenido en español en sus datos de entrenamiento (Wang et al, 2024) y a su capacidad de añadir tareas específicas vinculadas a la ejecución de consultas de búsqueda semántica, lo cual permitió una

[8] Se utilizó la librería de software libre *Sentence Transformers*, en su implementación para el lenguaje Python. La documentación completa puede consultarse aquí: https://www.sbert.net/

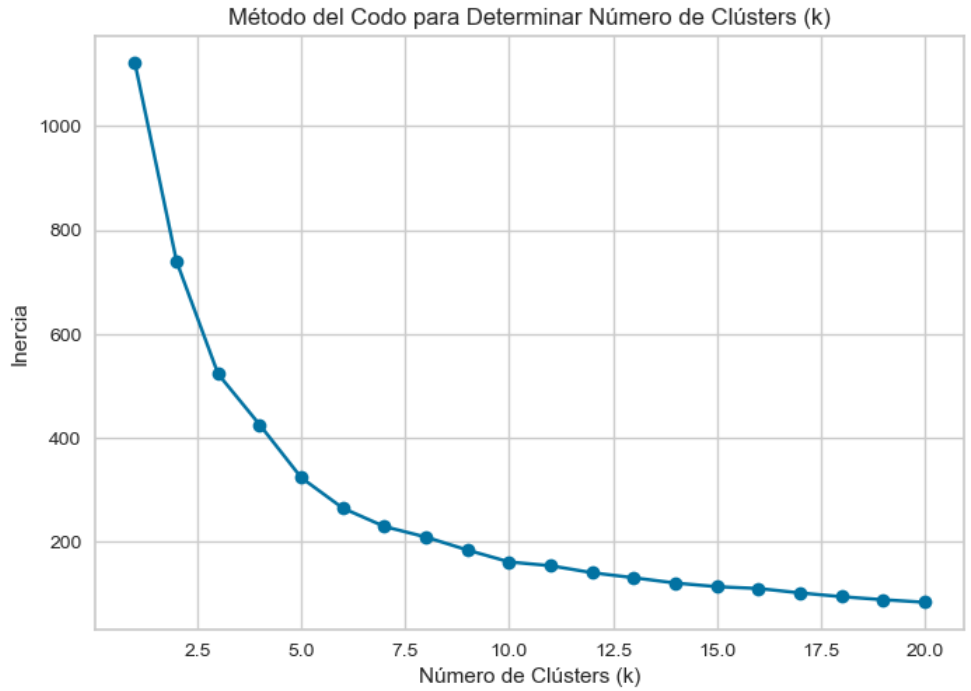
[9] El código y la documentación completa detrás del modelo de lenguaje elegido, *multilingual-e5-large-instruct*, puede consultarse en la ficha dedicada al mismo dentro del portal de *Hugging Face*: https://huggingface.co/intfloat/multilingual-e5-large-instruct

mayor precisión en su orientación para explorar las preguntas a seleccionar como más frecuentes por tema.

Por cada uno de los seis temas del debate, se procesó el cálculo de relaciones semánticas entre las preguntas de la muestra, llamadas incrustaciones o *embeddings*, las cuales permiten codificar las características de cada oración en vectores de 1,024 dimensiones identificables por el modelo de lenguaje implementado.

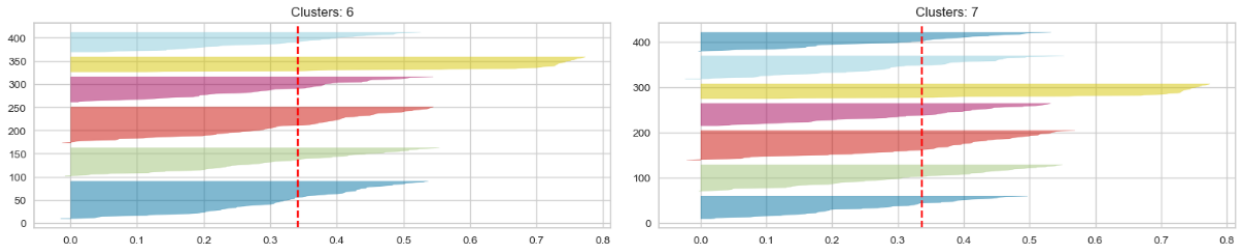
Para la visualización de relaciones semánticas y la identificación de clústeres en las mismas, se aplicaron algoritmos de reducción de dimensionalidades, *UMAP* (McInnes et al, 2018), y de clusterización, *k-means* (Pedregosa et al., 2011). Para elegir el número óptimo de clústeres por muestra de cada tema, se implementó una verificación a doble ciego de dos métodos para identificar el número ideal de clústeres por tema (Método "del Codo" y Método "de Silueta")[10].

Ejemplo de Método del codo (muestra del tema Salud):

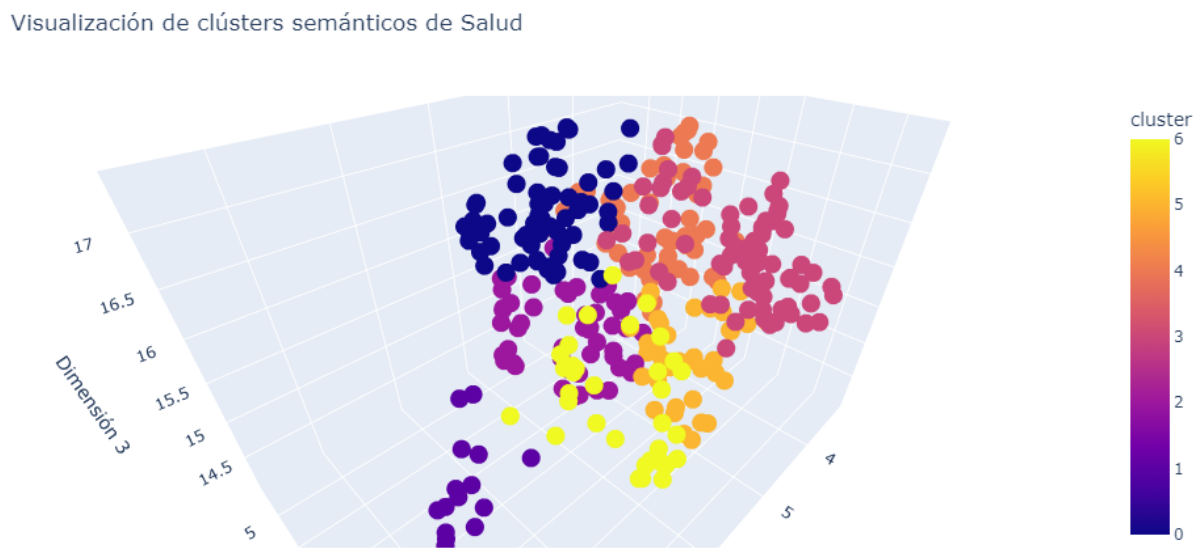


Ejemplo de Método de Silueta (muestra del tema Salud):

[10] El análisis por ambos métodos, del "Codo" (Thorndike, 1953) y de "Silueta" (Rousseeuw, 1987), permitió mitigar la ambigüedad que puede surgir a la hora de optar por un número específico de clústeres. Por cada tema, se aplicaron ambos métodos y se cotejaron sus resultados para elegir un número de clústeres orientado por ambos criterios.



Ejemplo de visualización de relaciones semánticas y clústeres (muestra del tema Salud):



Este procedimiento se realizó de manera segmentada con las preguntas en la muestra estratificada de cada uno de los seis temas, para poder visualizar y explorar las relaciones semánticas y los clústeres reconocibles en ellas.

Lo anterior permitió identificar los tres clústeres más grandes por tema, dada la cantidad de preguntas que los conformaban, para así orientar la exploración e identificación de núcleos semánticos representativos de los tres subtemas, considerados como las preocupaciones más recurrentes en las preguntas de la ciudadanía.

Ejemplo de identificación de 3 mayores clústeres por tema (muestra del tema Salud):

Identificar clústers más grandes en la muestra del tema (mayor frecuencia de preguntas por similitud semántica):

```
# Identificar y enlistar clusters con mayor cantidad de preguntas (más frecuentes) del tema
for cluster, freq in dfClusterizado['cluster'].value_counts(ascending=False).head(10).items():
  print(f"Cluster {cluster} - {freq} preguntas")

print()
```

```
Cluster 5 - 62 preguntas
Cluster 3 - 62 preguntas
Cluster 6 - 61 preguntas
Cluster 1 - 47 preguntas
Cluster 0 - 45 preguntas
Cluster 2 - 42 preguntas
Cluster 4 - 34 preguntas
```

A su vez, la codificación completa de vectores por cada pregunta (con 1024 dimensiones) se retomó posteriormente para la ejecución de consultas de búsqueda semántica, formuladas a partir de la exploración con métodos mixtos de lingüística de corpus que se describen a continuación.

Utilización de herramientas de lingüística de corpus para la identificación de núcleos semánticos

Dentro de la lingüística, el campo de la lingüística de corpus se refiere "al estudio del lenguaje que incluye recolectar grandes cantidades de lenguaje ocurrido en contextos naturales y el uso de software especializado que trabaja con ese lenguaje para obtener información acerca de frecuencias, co-ocurrencias y significados" (Hunston, 2022). Así, este campo ha desarrollado un amplio repertorio de conceptos y herramientas para identificar y comparar estructuras sintácticas, categorizar usos y orientaciones semánticas y medir frecuencias de distinto tipo en corpus de texto que pueden contener desde algunos cientos a miles de millones de palabras.

La lingüística de corpus brinda herramientas útiles para identificar segmentos de palabras dentro de textos y/o dentro de las oraciones que componen un texto, y categorizarlos de acuerdo con distintos criterios y objetivos, por ejemplo, de acuerdo con la frecuencia con la que aparecen, con el énfasis semántico que tienen dentro de un discurso, o con la intención de identificar variaciones de una misma enunciación o de enunciaciones similares. Así, existen los "paquetes lexicales" (*lexical bundles*) que se refieren a frases coloquiales repetidas en un corpus de palabras (suelen ser saludos, muletillas, regionalismos, etc.), y permiten identificar usos de la lengua de algún grupo social o región; o "plantillas" (Longui, 2021) que sirven para identificar desviaciones u orientaciones semánticas mediante la utilización constante de la suma de ciertos sufijos, prefijos y adjetivos alrededor de ciertas palabras clave dentro de un discurso, lo cual permite señalar los modos de nombrar a ciertos grupos o de asociarlos a ciertos significados en la arena pública; o también existen los enegramas, que se refieren a grupos de palabras (de 1, 2, 3, 4, 5 o más) que aparecen juntas en un corpus textual, y sirven para identificar qué segmentos de oraciones (y sus significados) aparecieron con

más o menos frecuencia en un corpus con respecto a otros de igual o distinto número de palabras.

Como parte de la limpieza y transformación inicial de las preguntas, con técnicas y librerías comunes (Bird et al, 2009) de procesamiento de lenguaje natural (PLN[11]), se generó una columna adicional con una nueva versión de las preguntas sin artículos y sin preposiciones (*stop words*), es decir, preguntas únicamente con palabras lexicales o palabras con un significado independiente de su uso en una oración (como "educación", "propuesta", "escuela" o "ley"). A partir de esta columna, se prepararon listas, por cada tema en la muestra y de sus tres clústeres más grandes respectivamente, para generar los enegramas más relevantes de 1, 2 y 3 palabras de cada uno de los temas del debate, con los que el equipo de Signa_Lab ITESO identificó lo que podemos nombrar como "núcleos semánticos" de cada tema, es decir, los principales significados dentro de varias preguntas de cada tema. Así, la identificación de los principales núcleos semánticos permitió señalar las inquietudes ciudadanas más frecuentes colocadas en las preguntas segmentadas por tema y en los tres clústeres más grandes de cada uno.

Clústeres y enegramas para la identificación de las preguntas más frecuentes

Los enegramas de 1, 2 y 3 palabras generados a través de técnicas de procesamiento del lenguaje natural (PLN), con librerías especializadas de Python, hicieron posible contabilizar la frecuencia ponderada Tf-idf ("frecuencia de término – frecuencia inversa de documento") de cada palabra (Spärck-Jones, 1972) y de cada grupo de 2 y de 3 palabras. Esto dio como resultado una lista de los núcleos semánticos más relevantes de cada tema.

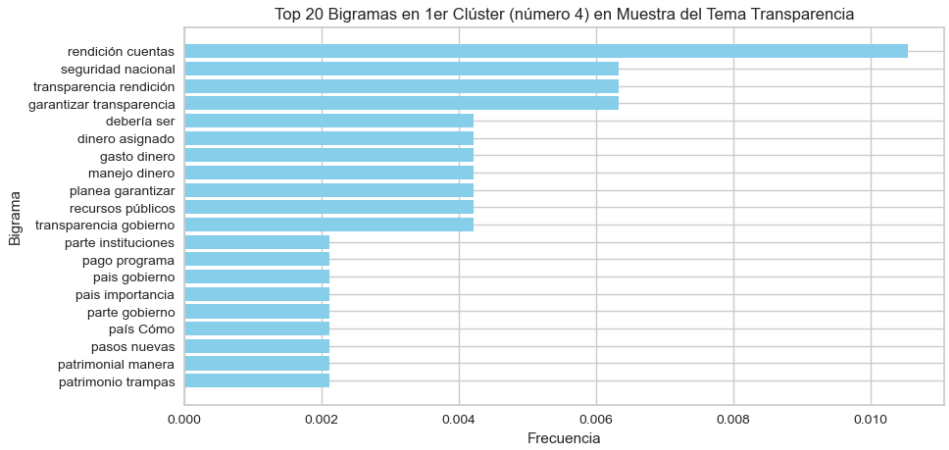
Estas listas de enegramas más importantes por tema fueron obtenidas tanto a partir de la filtración automatizada de los clústeres por tema, como a partir de un proceso manual de filtrado de esos clústeres con herramientas de lingüística de corpus[12]. Ambas versiones de las listas fueron obtenidas a la par por el equipo de Signa_Lab ITESO, para después realizar una lectura a doble ciego que coincidió en todos los casos.

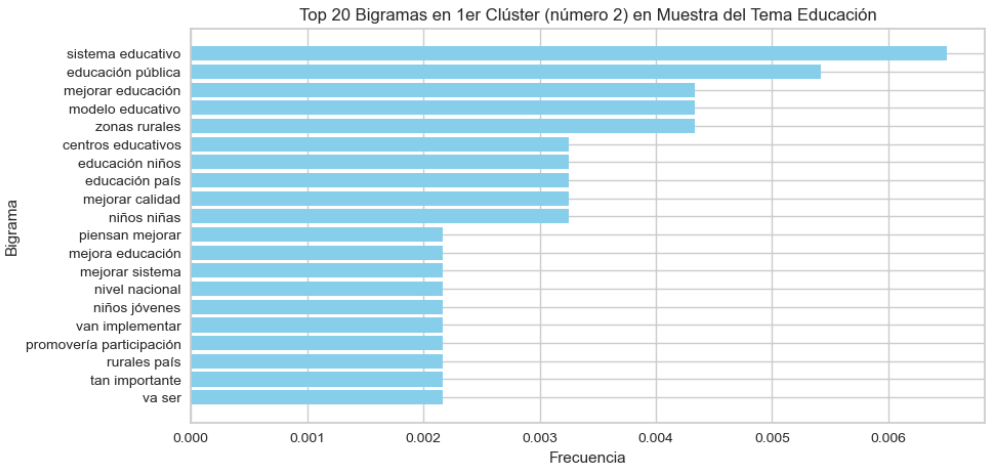
Con esta comprobación del resultado, el equipo desarrolló consultas basadas en técnicas de búsqueda semántica[13] para preseleccionar las 3 preguntas más frecuentes

[11] El Procesamiento de Lenguaje Natural se refiere al conjunto de herramientas computacionales para la identificación, categorización y análisis de las características de un texto que, una vez procesadas, son utilizadas para la generación de insumos útiles para resolver problemas, diseñar estrategias, llevar a cabo diagnósticos, etc., a partir del corpus de texto analizado con estas herramientas.
[12] Para este procesamiento manual se usó Antconc:
https://www.laurenceanthony.net/software/antconc/
[13] El modelo de lenguaje utilizado (multilingual-e5-large-instruct) para la generación de vectores de texto (embeddings) se implementó también para realizar consultas de búsqueda semántica, con los términos formulados a partir de la identificación de los núcleos semánticos más representativos de cada tema y la definición de una tarea para orientar la búsqueda (por ejemplo: "busca propuestas sobre el tema de salud"), habilitadas por la misma naturaleza instruccional (instruct) del modelo.

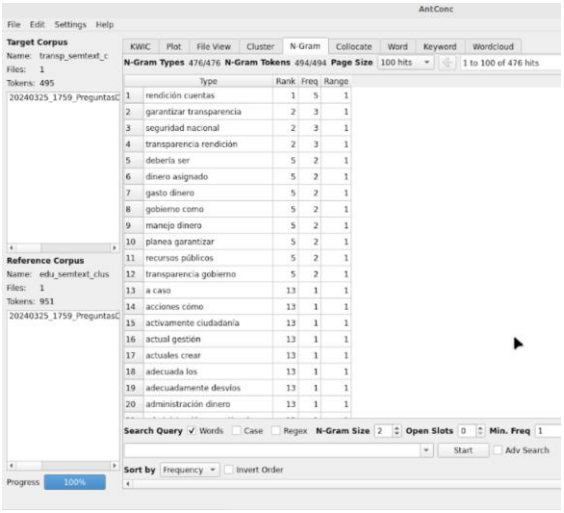
en cada tema. Estas búsquedas arrojaron, por cada uno de los tres clústeres más grandes de cada tema, un listado de preguntas ponderado por similitud semántica respecto a los términos de la consulta. De este listado de preguntas más representativas de cada núcleo semántico identificado, se priorizaron aquellas que cumplían de manera más clara con los criterios de elegibilidad especificados en la metodología del INE, como “coherencia argumentativa, buena sintaxis y neutralidad” (INE, 2024), para asegurar la claridad y pertinencia de las 18 preguntas seleccionadas por el criterio de frecuencia.

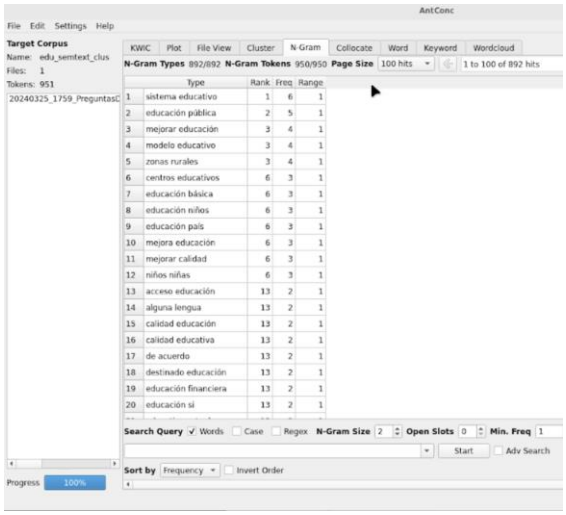
Ejemplos de bigramas de clústeres más relevantes en temas transparencia y educación, generados con librerías de programación especializadas para el Procesamiento de Lenguaje Natural (PLN) en Python:





Ejemplos de bigramas de clústeres más relevantes en temas transparencia y educación filtrados con herramientas de lingüística de corpus:





Ejemplo de búsqueda semántica por clúster con el tema Salud:

```
# Parámetros para ejecutar consulta por búsqueda semántica de 2ndo clúster:
task_2 = 'Preguntas sobre el tema de Salud'

# Términos de consulta definidos tras el análisis cualitativo orientado la cotejación de palabras más frecuentes (método TF-IDF)

query_2 = "falta de medicamentos en hospitales, clínicas y en el sector salud"

columnaTexto = "clean_text"
columnaEmbeddings = "Embedding"
cluster = cluster_2

# Ejecutar consulta por búsqueda semántica:
searchIntFloat(embedder, task_2, query_2, dfClusterizado, columnaTexto, columnaEmbeddings, cluster)
```

```
Búsqueda en clean_text de: falta de medicamentos en hospitales, clínicas y en el sector salud
id: 79071
Similitud: 0.9268273343544456 - 0 - Que propuesta propondrían para el re abastecimiento de medicamentos y la creciente deficie
ncia de las instituciones de salud públicas y su acceso?
--------
id: 91401
Similitud: 0.9199991623958812 - 1 - Es bien conocido que el sector salud desde hace muchos años tiene desabasto en medicación
y en atención a infantes hay un gran resago que estrategias plantean para el mejoramiento en esta área
--------
id: 81611
Similitud: 0.9157401522623755 - 2 - Es urgente que en los centros de salud se cuente con los medicamentos que necesita la pobl
ación en general
--------
id: 27801
Similitud: 0.9149053262385038 - 3 - Cómo solucionarían el desabasto de medicinas que está presente desde hace mucho tiempo en
los servicios de salud del país.
--------
```

En el anterior ejemplo, se retoma la consulta realizada con búsqueda semántica en uno de los clústeres del tema de salud. A partir de la instrucción general, dada al modelo de lenguaje, de priorizar “preguntas sobre el tema de salud”, se orienta la búsqueda a partir de la formulación construida después del análisis de enegramas, que en el caso de este ejemplo se definió “falta de medicamentos en hospitales, clínicas y en el sector salud” como la formulación de términos con mayor frecuencia en las preguntas de este clúster del tema de Salud. Del listado de las preguntas más cercanas a dicha consulta, se seleccionó la que mejor cumplía con los criterios establecidos por la metodología.

Selección aleatoria de preguntas por tema y región

Una vez seleccionadas las 18 preguntas por frecuencia, se utilizaron sus identificadores únicos (IDs) para retirarlas del resto de la muestra estratificada y evitar una pregunta duplicada en la selección por aleatoriedad. Para cumplir con las cantidades y distribución de preguntas indicadas por la metodología del INE, se implementó el código necesario para extraer aleatoriamente 5 preguntas por tema y región. Cabe mencionar que, aun cuando la extracción fue completamente aleatoria, se incorporó a la función de extracción una "semilla" (*seed number*) para fijar los resultados aleatorios y garantizar su trazabilidad y replicabilidad.

Una vez realizado este procedimiento por cada tema, se obtuvo una preselección de 90 preguntas aleatorias para pasar a la etapa de revisión. Los resultados de esta selección preliminar de 18 preguntas por frecuencia y 90 por aleatoriedad se pueden encontrar en los anexos del presente informe.

Etapa 4. Revisión de preguntas

Esta etapa consistió en la revisión de las 108 preguntas preseleccionadas con el objetivo de verificar que cumplieran con los criterios definidos por el INE. Dicha revisión se llevó a cabo mediante la lectura y análisis humano, realizados por el equipo de Signa_Lab ITESO y la representante de Comunicación Social del INE. Estas fueron las actividades correspondientes a esta etapa:

Act. 8. Realizar la revisión de las 108 preguntas

Act. 9. En su caso, sustitución de preguntas por frecuencia y aleatoriedad

Act. 10. Definición de las 108 preguntas finales

Act. 11. Acto de entrega de preguntas seleccionadas

Identificación de preguntas para ser reemplazadas

Con las 108 preguntas preseleccionadas, se llevó a cabo una primera ronda general de revisión en la cual Signa_Lab ITESO, con el acompañamiento y supervisión del cumplimiento de criterios por parte de la representación del INE, identificó 28 que necesitaban ser reemplazadas, mientras que 80 fueron aprobadas. De las reemplazables, solamente una pertenecía a las preguntas preseleccionadas por frecuencia, el resto era parte de las preguntas elegidas de manera aleatoria.

En cuanto a la temática de las preguntas reemplazadas en esta primera ronda, 5 correspondieron a Combate a la corrupción; 4 a Educación; 2 a No discriminación y grupos vulnerables; 7 a Salud; 8 a Transparencia, y 1 a Violencia contra las mujeres.

Los motivos de reemplazo de estas preguntas fueron:

- Carecer de neutralidad.
- No tener suficiente coherencia argumentativa.
- Constituían propaganda gubernamental o mencionaban programas de gobierno.
- Y/o no se adscribían al tema en el que fueron registradas en el sitio habilitado por el INE para capturar las preguntas.

Después de esta primera ronda de reemplazo, se realizó nuevamente el ejercicio de revisión manual únicamente para las preguntas reemplazadas. En esta ocasión, se detectaron 11 preguntas por aleatoriedad que no cumplían con los criterios estipulados para pasar a la selección final: 2 del tema Combate a la corrupción, 1 del tema No discriminación y grupos vulnerables, 2 del tema Salud y 6 del tema Transparencia.

En un tercer ejercicio de revisión, 2 preguntas fueron identificadas como reemplazables, ambas de la categoría de Transparencia. En la cuarta ronda de revisión

sólo 1 pregunta perteneciente a la categoría de Transparencia fue identificada como reemplazable.

Finalmente, en el quinto ejercicio de revisión manual, se determinó que todas las preguntas cumplían con los criterios de selección, lo que dio como resultado las 108 preguntas de la selección final que fueron enviadas a las personas moderadoras del debate. Es importante destacar que no hubo preguntas descartadas por contener términos proscritos o por ser idénticas a otras. Es decir, la implementación del diccionario elaborado por el laboratorio y del código de identificación de preguntas redactadas con 100% de similitud entre sí, funcionó de manera óptima. En todas las rondas de revisión, los motivos de reemplazo de preguntas fueron los arriba mencionados, y se registran, por cada caso de reemplazo, en los anexos del presente informe.

Los criterios del INE para la selección de las preguntas fueron los siguientes:

- Que las preguntas se ciñan al tema definido.
- Que las preguntas sean comprensibles.
- Que las preguntas brinden la posibilidad de plantearse a cualquier tipo de candidatura de forma directa.
- Que las preguntas no contengan algún sesgo partidista.
- Que las preguntas no contengan lenguaje ofensivo y/o expresiones en doble sentido que resten seriedad al ejercicio.
- Que las preguntas no contengan discurso de odio, discriminación y/o violencia de cualquier tipo.

A modo de resumen, los conteos de revisión y reemplazo por cada ronda se detallan a continuación:

Ronda de Revisión	**Preguntas aprobadas**	**Preguntas por reemplazar**
1ra Ronda	80	28 (1 por frecuencia, 27 por aleatoriedad)
2nda Ronda	97	11 (por aleatoriedad)
3ra Ronda	106	2 (por aleatoriedad)
4ta Ronda	107	1 (por aleatoriedad)
5ta Ronda	108	0

Lo anterior dio un total de 42 preguntas reemplazadas, es decir, una **tasa de reemplazo de 2.47%** considerando como base para contabilizar el reemplazo las 1,701 preguntas de la muestra estratificada por tema-región, bajo el razonamiento de que cualquiera de

esas preguntas pudo haber sido incorporada en la selección final de 108, en el caso de haber cumplido con los criterios definidos por la metodología.

Etapa 5. Plazos establecidos

Esta es una etapa transversal en el proyecto y corresponde a la actividad número 12 de la metodología:

- ***Act. 12. Verificar el cumplimiento de los plazos establecidos en la metodología***

Para cumplir con esta etapa, se diseñó un plan de trabajo bajo el acompañamiento de la Coordinación Nacional de Comunicación Social (CNCS) y de la Oficialía Electoral del INE. El plan de trabajo y las bitácoras generadas entre el 22 y el 29 de marzo en las jornadas de procesamiento de la base de datos en Signa_Lab ITESO se pueden consultar en la sección de anexos de este informe. A continuación, se hace el recuento de las actividades realizadas y de la culminación de las etapas 0 a la 4 del plan de trabajo como quedó asentado en las observaciones realizadas por el INE.

0. Entrega de la base de datos

22 de marzo:

CDMX:

- Acto Protocolario Lobby del INE a las 9:00 horas. Presidido por la Consejera Electoral y Presidenta de la Comisión Temporal de Debates del Instituto Nacional Electoral, Carla Astrid Humphrey Jordan y el Director de Relaciones Externas de SIGNA LAB ITESO, Doctor Humberto Orozco Barba. Los consejeros y consejeras Martín Faz Mora, Jorge Montaño Ventura, Dania Paola Ravel Cuevas, Claudia Zavala Pérez; el Director de Investigación y Posgrado de SIGNA LABITESO, Doctor Bernardo Masini Aguilera; el encargado del Despacho de la Coordinación Nacional de Comunicación Social del Instituto Nacional Electoral, Licenciado Iván Flores Ramírez, el Director del Departamento de Estudios Socioculturales de Signa_Lab ITESO, Doctor Juan Larrosa Fuentes, el encargado de Despacho de la Coordinación General de la Unidad de Servicios de Informática, Ingeniero Félix Manuel de Brasdefer Coronel; la Directora del Secretariado del Instituto Nacional Electoral, en su calidad de Coordinadora de la Oficialía Electoral, Maestra Rosa María Bárcena Canuas.

Guadalajara:

- Para dar Fe de este acto, la Directora de Oficialía Electoral del Instituto Nacional Electoral, Licenciada Irene Maldonado Cavazos.
- Traslado y entrega de la base con 24,000 preguntas a Signa Lab ITESO.
- 16:15 hrs. Entrega de la base de datos, lectura por parte de Victor Hugo Ábrego, de las etapas y actividades de la Metodología aprobada por el INE.
- 16:45 hrs. revisión del archivo.
- 17:00 hrs. Verificación del conteo de preguntas.

1. Preparación de la base de datos

23 de marzo:

- 10:15 hrs. Presentación del equipo y objetivos del día a cargo de Victor Hugo Ábrego.

- 11:10 hrs. Discusión de posibles escenarios en la redacción de las preguntas: más de una pregunta en el mismo campo de 300 caracteres.
- Discusión sobre los diccionarios.
- a) Revisión inicial de los diccionarios con corte a la semana del 18 de marzo.
- b) De la versión previa de los diccionarios, se discutió incorporar términos por sesgo ideológico y por mención de actores políticos (como presidente, presidenta, que no serán motivo de exclusión, y términos relacionados con la administración actual, como “cuarta transformación”, que sí serán excluidos). En el informe final se especificarán dichos cambios y el porqué de los mismos.
- c) Documentar el proceso de construcción y depuración de los diccionarios (con las personas involucradas y las actividades desarrolladas) dentro del informe final.
- Resultado: Diccionario inicial de palabras
- 11:50 hrs. Ejercicio de prueba
- 12:40 hrs.Primera ejecución de depuración y obtención de resultados
- 13:20 hrs. Redacción de insumos para el informe
- 13:30 hrs. Selección de las gráficas más relevantes para el ejercicio.
- 14:30 hrs. Procesamiento de la base de datos con los datos reales
- 16:10 hrs. Ruta de trabajo posterior al procesamiento de los datos.

24 de marzo:

- 10:20 hrs. Revisión de fase de depuración de la base de datos.
- 12:40 hrs. Actividades de procesamiento
- 14:30 hrs. Revisión de preguntas descartadas por diccionario y de fórmula de estratificación

25 de marzo:

- 11:00 hrs. Identificación de registros extemporáneos en los registros de pregunta
- 12:47 hrs. Obtención de base de datos depurada por diccionario y por criterios de repetición actualizados

2. Obtención de la muestra estratificada por tema-región

25 de marzo:

- 16:05 hrs. Inicio de la Etapa 2: Obtención de la muestra estratificada y clasificación por región

26 y 27 de marzo:

- 10:40 hrs. Puesta a punto del código para la generación de embeddings y clústeres de la muestra estratificada por tema-región y pruebas de rastreo manual de núcleos semánticos con herramientas de lingüística de corpus con la muestra estratificada por tema-región.

3. Selección de las preguntas

27 de marzo:

- 16:01 hrs. Obtención de preguntas
- 23: 43 hrs. Culminación de la obtención de las preguntas
- Se preseleccionaron las 3 preguntas más frecuentes por tema; se verificó y aprobó la fórmula de obtención de las 90 preguntas preseleccionadas aleatoriamente.
- Se procedió a la extracción de las 108 preguntas preseleccionadas por frecuencia y por aleatoriedad.
- 18 preguntas preseleccionadas por frecuencia
- 90 preguntas preseleccionadas aleatoriamente
- La etapa culminó a la 01:45 hrs.

4. Revisión de las preguntas

27 de marzo:

- 16:20 hrs. Presentación del equipo. Lectura de las etapas, actividades y avances.
- 16:50 hrs. Revisión y explicación general de las etapas y actividades cumplidas hasta ahora. Lectura de las preguntas
- 17:25 hrs. Primera lectura general de las 108 preguntas preseleccionadas

28 de marzo:

- 10:40 hrs. Primera revisión y reemplazo de las preguntas preseleccionadas que no cumplieron con los criterios de la metodología
- 15:40 hrs. Ejercicios de reemplazo de preguntas preseleccionadas que con cumplen con los criterios de la metodología
- ENTREGABLES:
- 1.108 preguntas seleccionadas del proceso de revisión de preguntas.
- 2.Tasa de reemplazo por sustitución de preguntas de 2.47%, en relación con las 1,701 unidades de la muestra estratificada por tema-región.

29 de marzo:

- 10:00 hrs. Preparación de la base de datos para entrega al INE.
- 12:30 hrs. Acto protocolario para bajar a dos unidades USB la base de datos con las 108 preguntas seleccionadas y el código de integridad generado para su protección.
- Entrega en dos sobres sellados de las USB al Director del Departamento de Estudios Socioculturales de Signa_Lab ITESO, Doctor Juan Larrosa Fuentes para su resguardo.
- Para dar Fe de este acto, el Subdirector de Oficialía Electoral del Instituto Nacional Electoral, Licenciado Adrián Sánchez Sáez.
- Lunes 1° de abril de 2024 ACTO PROTOCOLARIO DE RECEPCIÓN DE LAS PREGUNTAS SELECCIONADAS A PARTIR DE LA BASE DE DATOS PROCESADA POR ITESO-SIGNA LAB.
- 10:00 hrs., Lobby del Instituto Nacional Electoral con la presencia de la Consejera Carla Humphrey Presidenta de la Comisión Temporal de debates, Dra. Catalina Morfín López Directora General Académica ITESO, Dr. Juan Larrosa, Director del Departamento de Estudios Socioculturales ITESO-Signa Lab e Iván Flores Ramírez Encargado de Despacho de la CNCS.

Etapa 6. Informe final

La última etapa del proyecto consiste en la redacción del presente informe y cumple con la actividad 13 de la metodología del INE:

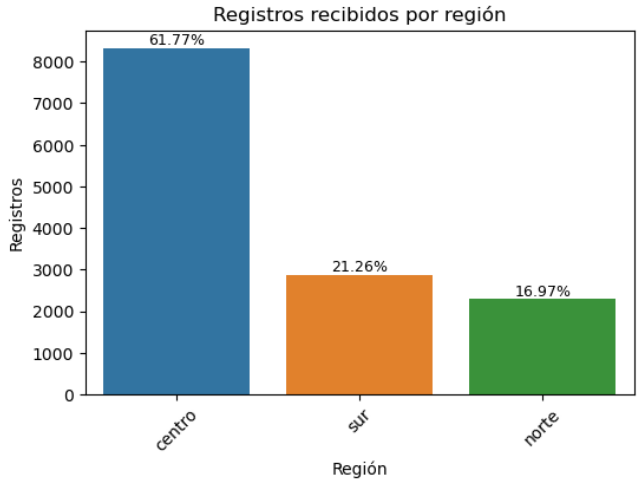
Act. 13. Redacción de informe documental y de informe técnico del proyecto

3. RESULTADOS, HALLAZGOS Y ENTREGABLES[14]

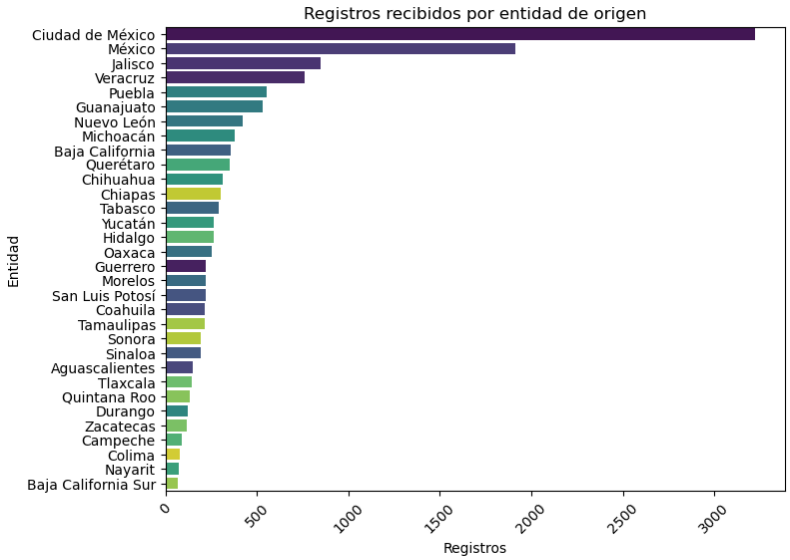
Resultados de registros totales

Participación por entidad y por región

Los 13,484 registros capturados por el INE están distribuidos por región en 8,329 pertenecientes a la región centro, 2,867 pertenecientes a la región sur y 2,288 pertenecientes a la región norte. Esto, traducido a porcentajes del total de los registros obtenidos, arroja un 61.77% correspondiente al centro, 21.26% al sur y 16.97% a la región norte.

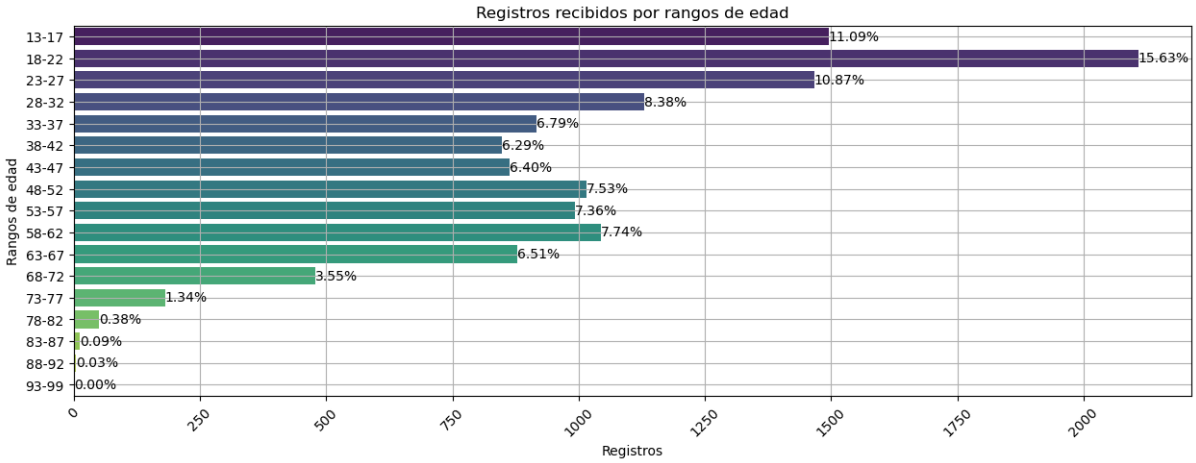


De acuerdo con los registros recibidos por entidad de origen, se identificó que las cinco entidades con mayor participación fueron Ciudad de México, con un total de 3,223 registros (23.90% del total), seguida del Estado de México con 1,911 (14.17%), Jalisco con 850 registros (6.30%), Veracruz con 761 (5.64%) y Puebla con un total de 553 registros (4.10%).



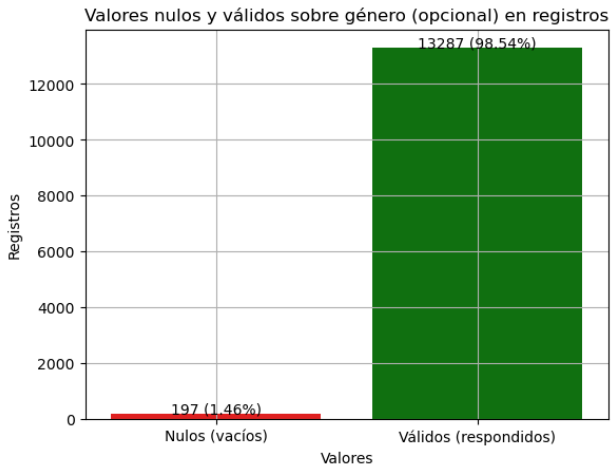
[14] Todas las capturas, gráficas y tablas de esta sección fueron generadas por el equipo de Signa_Lab ITESO, y están documentadas en los Anexos correspondientes.

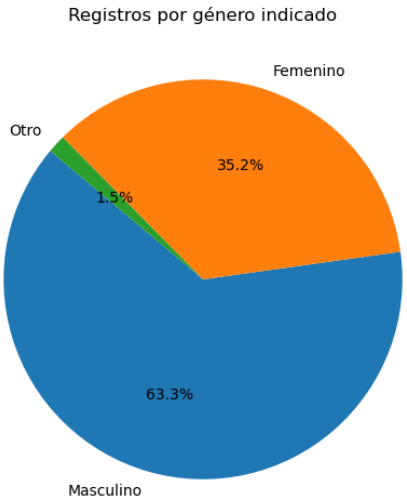
Participación de registros por edad



El grupo de edad que registró la mayor participación fue el de 18 a 22 años, con un 15.63% del total de los registros. El segundo grupo con mayor participación correspondió al rango de edad entre 13 y 17 años, con 11.09%; le siguió el grupo de 23 a 27 años, con un 10.87% de participación. En cuarto lugar, estuvo el grupo de 28 a 32 años, que conformó el 8.38% de los registros. Los rangos con menor participación se hicieron visibles a partir de 68 a 72 años (3.55%) en adelante. En medio de la gráfica hay un pico a la baja que indica una menor participación de los rangos entre 38 a 42 años (6.29%), y de 43 a 47 años (6.40%).

Participación de registros por género

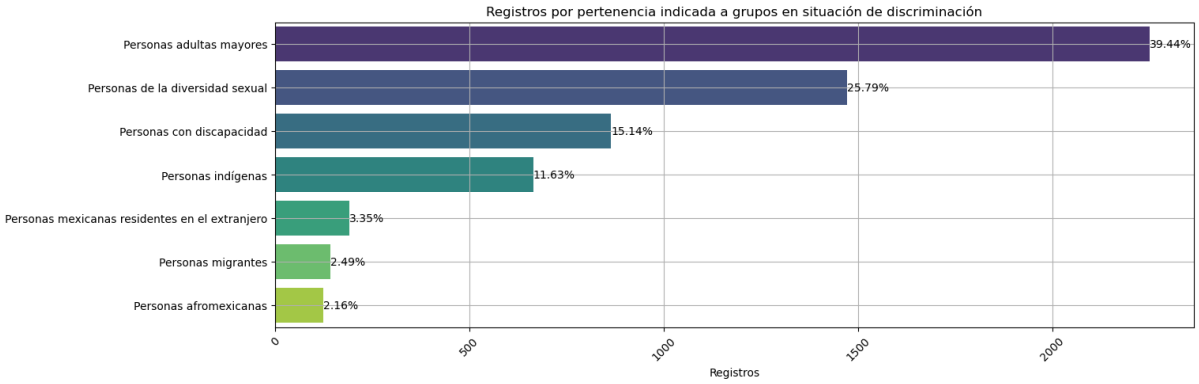




En cuanto al género, de los 13,484 registros totales, 13,287 (98.54%) fueron válidos (respondidos) y 197 (1.46%) fueron nulos, esto último quiere decir que no respondieron esta parte del formulario. De los registros válidos, 8,413 (63.3%) correspondieron al género *masculino* , 4,679 (35.2%) al género *femenino* y 195 (1.5%) a *otro*.

Registros por Grupos en Situación de Discriminación

Se recibieron 5,707 registros que respondieron sí a la pregunta *¿Te identificas con alguno de los siguientes Grupos en Situación de Discriminación?* Estos registros correspondieron al 42.32% de los registros totales, mientras que el 57.68%, conformado por 7,777 registros, no se consideró perteneciente a ningún grupo en situación de discriminación.



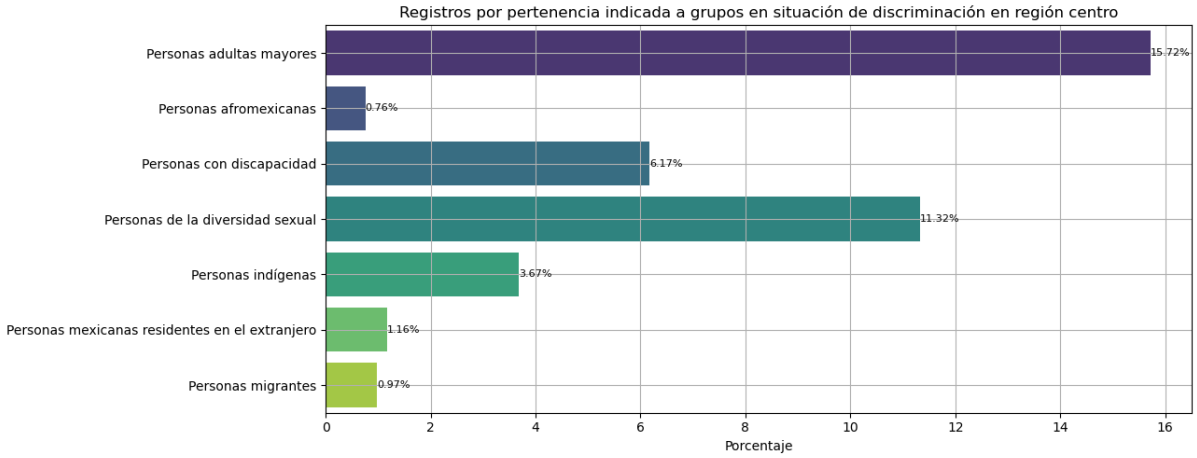
Los grupos en situación de discriminación que indicaba el formulario eran los siguientes: personas adultas mayores, personas de la diversidad sexual, personas con discapacidad, personas indígenas, personas mexicanas residentes en el extranjero, personas migrantes y personas afromexicanas.

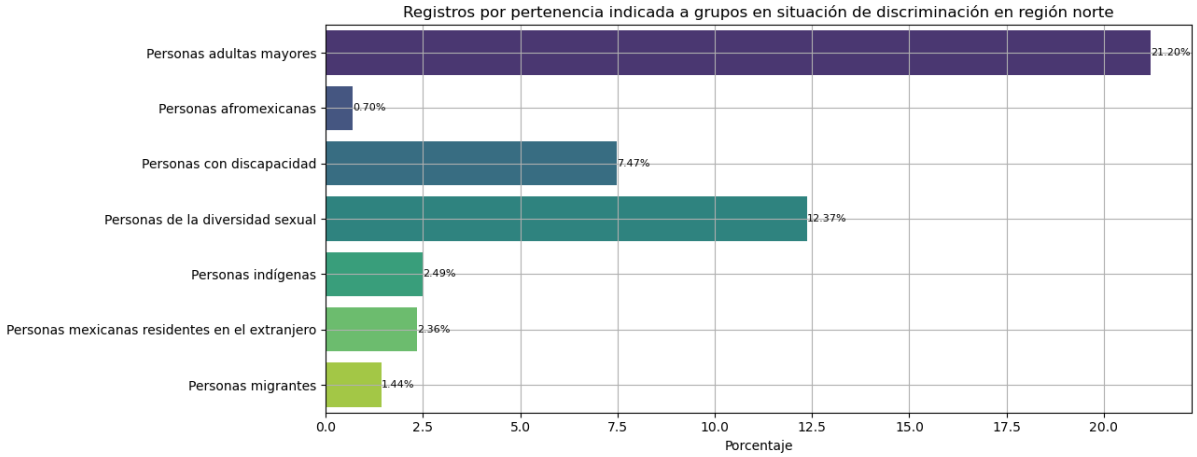
De los 5,707 registros en esta categoría, los grupos que tuvieron mayor número de registros correspondieron, en primer lugar, a *personas adultas mayores*, con 2,251 (39.44%); *personas de la diversidad sexual*, con 1,472 registros (25.79%); *personas con discapacidad*, con 864 registros (15.14%), y en cuarto lugar estuvo el grupo de *personas indígenas*, con 664 registros (11.63%).

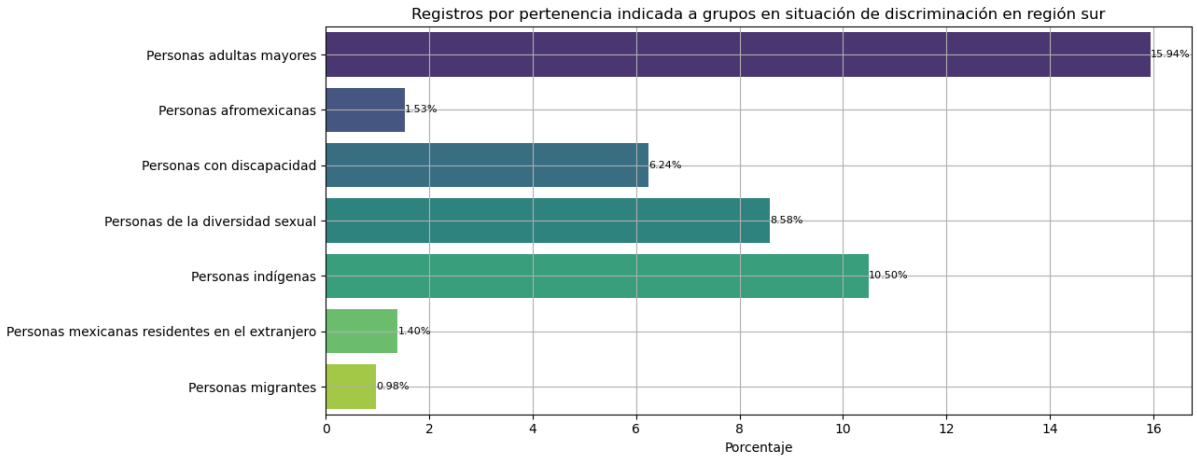
Ahora bien, la distribución de personas pertenecientes a grupos en situación de discriminación por región mostró a *personas adultas mayores* como el grupo de mayor participación en las tres regiones (15.72% en la región centro, 21.20% en la región norte y 15.94% en la región sur). Sin embargo, mientras en las regiones centro y norte el segundo grupo con mayor participación fue el de *personas de la diversidad sexua*l (con 11.32% en la región centro y 12.37% en la región norte); en la región sur el segundo lugar de participación fue el de *personas indígenas* (con 10.5%).

El grupo de *personas con discapacidad* ocupó el tercer escaño de estos registros en las regiones centro (6.17%) y norte (7.47%); mientras que este lugar en la región sur correspondió a los registros de *personas de la diversidad sexual* (8.58%).

Registros por Grupos en Situación de Discriminación y por Región



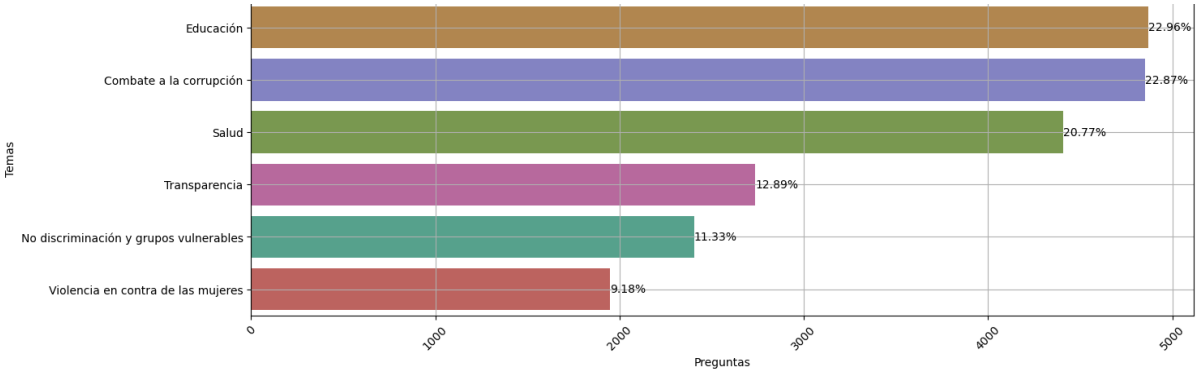




Resultados de población depurada de preguntas

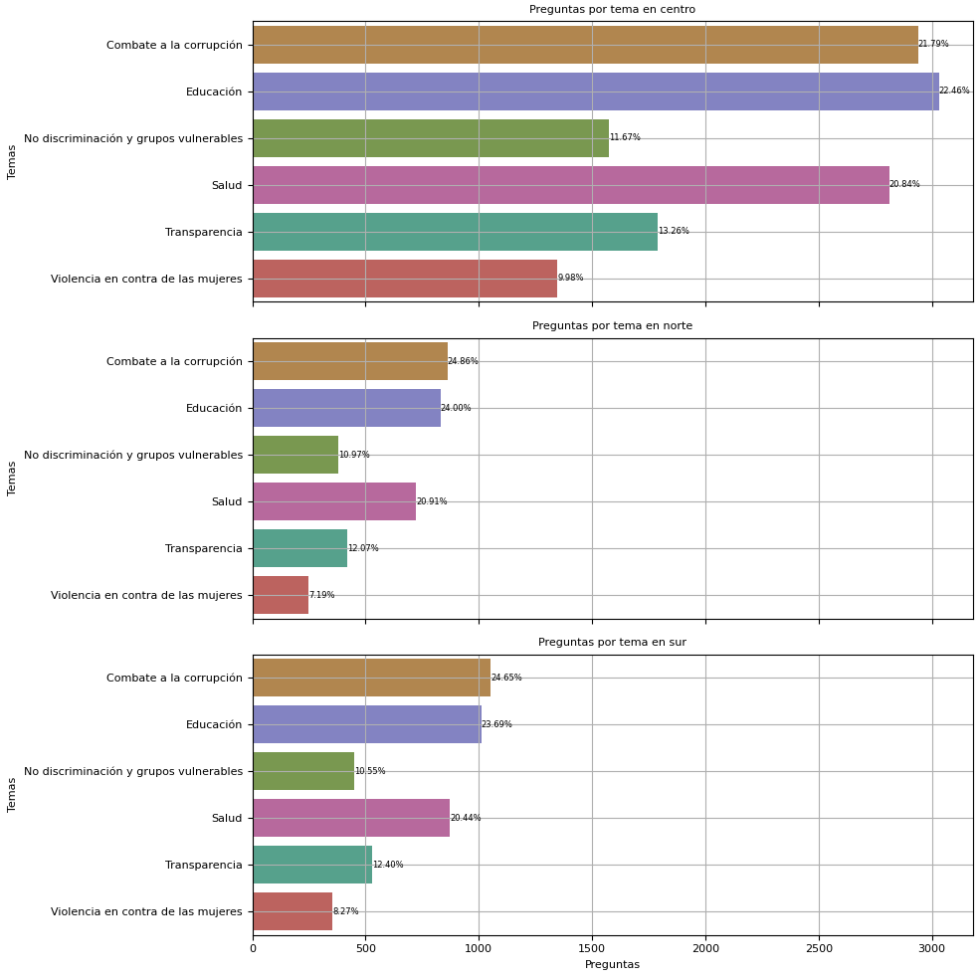
Población Depurada por Tema y Región

La población de preguntas depuradas, es decir, aquellas consideradas para obtener la muestra estratificada por tema y región, fue de 21,219 preguntas. De éstas, los temas con mayor participación registrada fueron *Educación*, *Combate a la corrupción* y *Salud*, con 22.96%, 22.87% y 20.77% de las preguntas respectivamente. Le siguieron *Transparencia*, con 12.89%, *No discriminación y grupos vulnerables*, con 11.33% y *Violencia en contra de las mujeres*, con 9.18% de participación registrada.



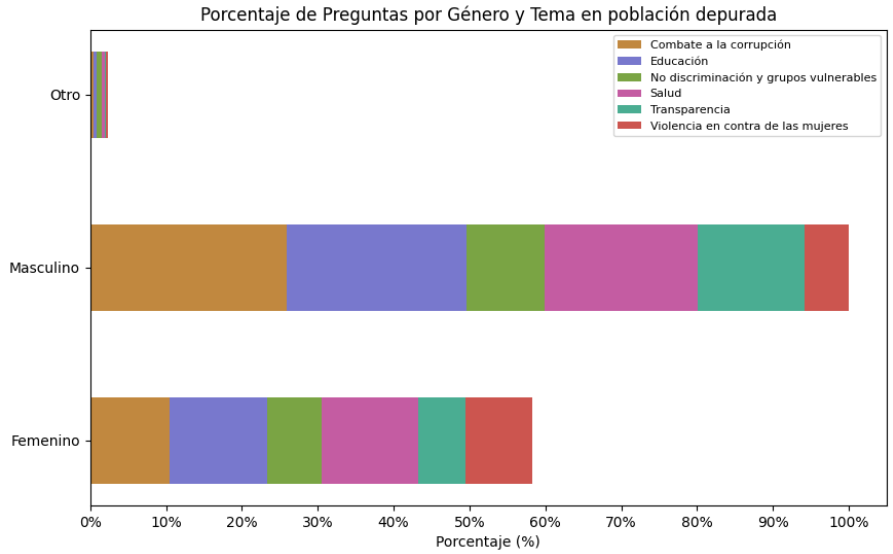
En cuanto a la distribución por región y por tema de estas 21,219 preguntas depuradas, se encontró que en la región centro, el tema *Educación* fue el que tuvo mayor participación, con 22.46%, mientras que en las regiones norte y sur fue *Combate a la corrupción*, con 24.86% y 24.65%. El segundo lugar de participación en las regiones norte y sur fue para *Educación*, con 24% y 23.69% respectivamente, mientras que en la región centro en este escaño estuvo *Combate a la corrupción*, con 21.79%. El tercer tema con mayor participación en las tres regiones fue *Salud*, con 20.84% en la región centro, 20.91% en la región norte, y 20.44% en la región sur.

Destaca la baja participación en *Violencia en contra de las mujeres* en las regiones norte y sur, con 7.19% y 8.27% respectivamente, mientras que en la región centro, este tema alcanzó 9.98% de participación. Por otro lado, el tema *Transparencia* registró un 13.26% de participación en la región centro, 12.07% en la región norte y 12.40% en la región sur. Finalmente, el tema *No discriminación y grupos vulnerables* alcanzó un 11.67% de participación en la región centro, mientras que 10.97% en la región norte y 10.55% en la región sur.



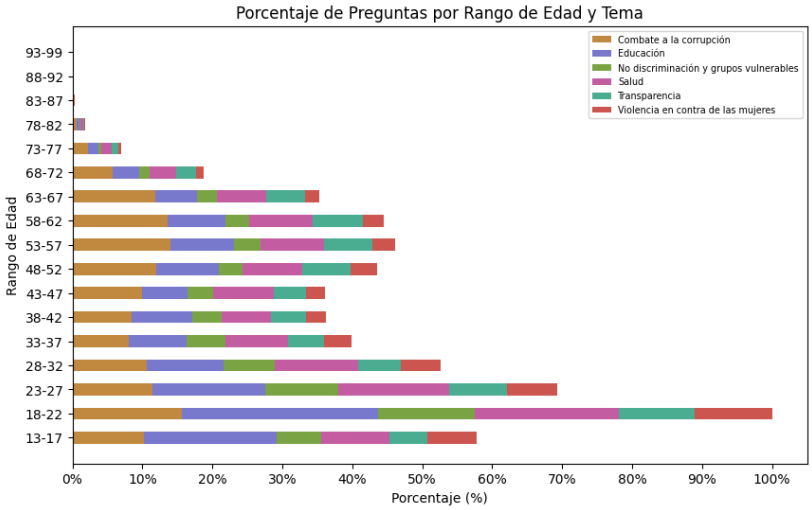
El cruce entre las variables de género y participación por tema de las 21,219 preguntas de la población depurada mostró que para hombres y mujeres el tema de mayor interés fue el de *Combate a la corrupción*, mientras que, para personas no binarias, fue el de *No discriminación y grupos vulnerables*.

Preguntas por Tema y Edad

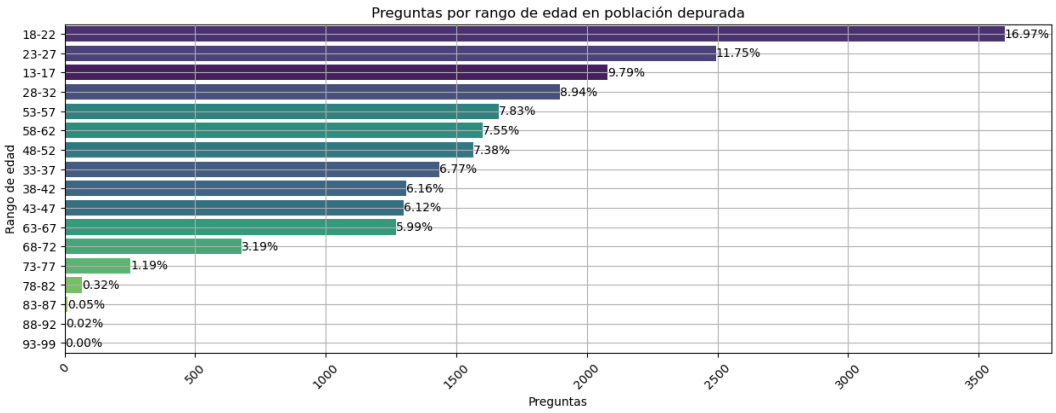


En cuanto a la participación de rangos etarios por tema, la población depurada de 21,219 preguntas

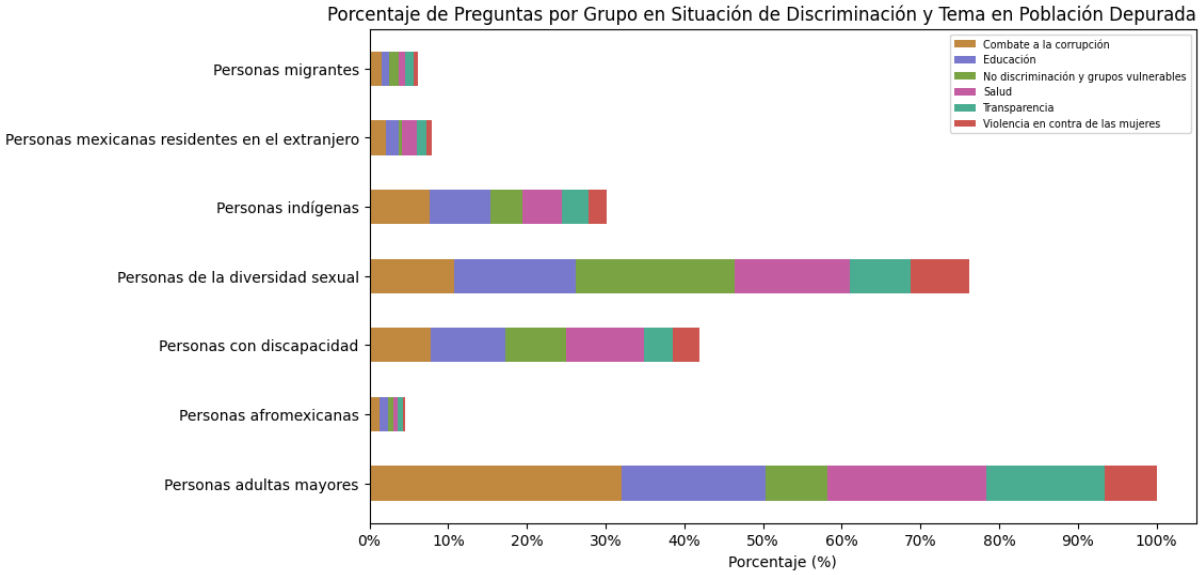
arrojó que el rango de edad de 18 a 22 años, que tuvo la mayor participación en el envío de preguntas con 16.97%, tuvo un interés predominante en el tema de *educación* (32.67%), seguido de *salud* (16.89%) y *combate a la corrupción* (17.80%). Mientras los rangos etarios con menor participación en las preguntas fueron de los 68 a los 92 años. Estos rangos, sumados, acumularon un total del 4.77% de las preguntas, y mostraron mayor preocupación en *combate a la corrupción*, *salud* y *educación*.



Los únicos dos rangos de edad que no abordaron todos los temas en sus preguntas fueron el de 83 a 87 años (0.05%), el cual no envió preguntas sobre el tema *no discriminación y grupos vulnerables*, y el de 88 a 92 años (0.02%), que solamente envió preguntas relacionadas con *combate a la corrupción* y *salud*.



Preguntas por Tema y Grupo en Situación de Discriminación:



De la población depurada de preguntas, el grupo en situación de discriminación con mayor porcentaje de registros fue el de *personas adultas mayores* con 3,273 preguntas (15.42%), seguido por *personas de la diversidad sexual* con 2,496 preguntas (11.76%). En tercer lugar, *personas con discapacidad*, con 1,371 preguntas (6.46%), y en cuarto, *personas indígenas*, con 986 preguntas (4.65%).

Por otro lado, los grupos en situación de discriminación con menor participación en la población depurada de preguntas fueron *personas mexicanas residentes en el extranjero* con 260 preguntas (1.23%), *personas migrantes* con 202 preguntas (0.95%) y *personas afromexicanas* con 150 preguntas (0.71%).

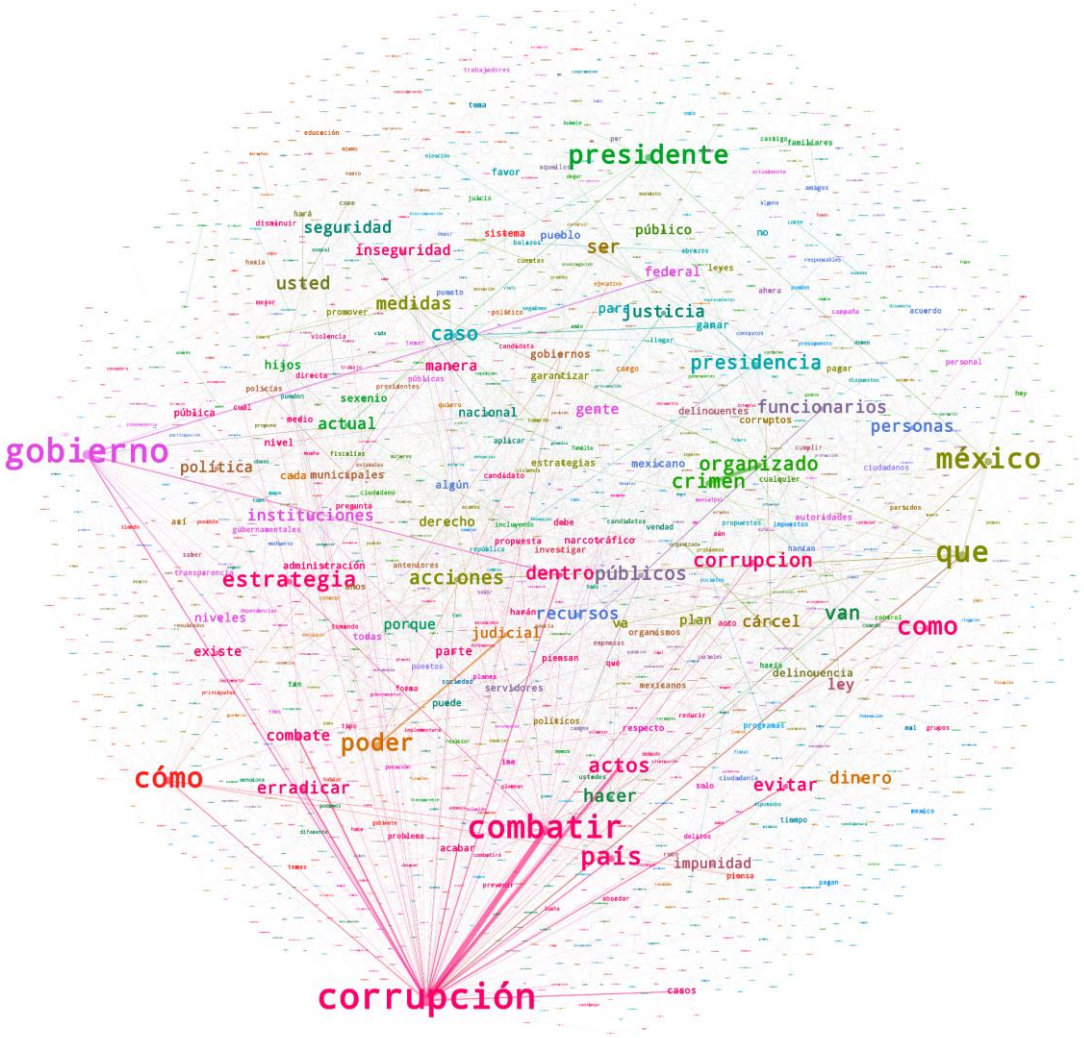
El grupo de *personas adultas mayores* mostró mayor interés en el tema de combate *a la corrupción* (31.98%), seguido por el tema de *salud* (20.13%). Por su parte, el grupo de *personas de la diversidad sexual* tuvo mayor presencia en el tema de *no discriminación y grupos vulnerables* (26.52%), seguido por los temas de *educación* (20.11%) y *salud* (19.27%), mientras que las *personas con discapacidad* centraron sus preguntas en *salud* (23.77%), *educación* (22.68%) y *combate a la corrupción* (18.59%).

Inquietudes ciudadanas. Hallazgos a partir del análisis semántico por tema

El siguiente es un listado de enegramas de 1, 2 y 3 palabras para cada uno de los temas del debate. El ejercicio consiste en la identificación y descripción concisa de campos

semánticos (palabras alrededor de un mismo significado) y de núcleos semánticos (segmentos lexicales de oraciones) con mayor peso en las preguntas de cada tema. Estas listas de palabras permiten un acercamiento a las inquietudes ciudadanas con mayor presencia en cada uno de los temas para este debate[15]. Cada tema es presentado con un grafo elaborado a partir de sus preguntas[16].

1. **Combate a la corrupción**

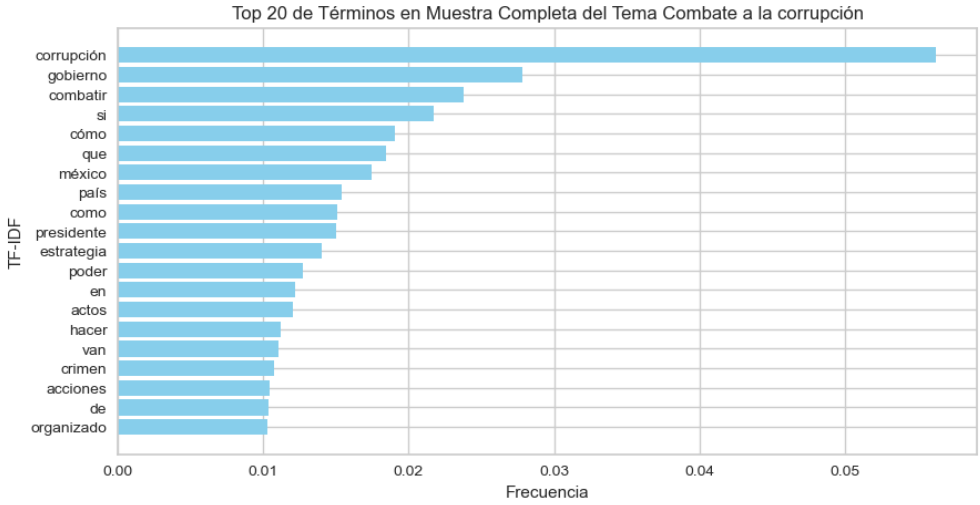


Bigrama de preguntas en muestra del tema: Combate a la Corrupción
Nodos: 2,092, Aristas: 4,050

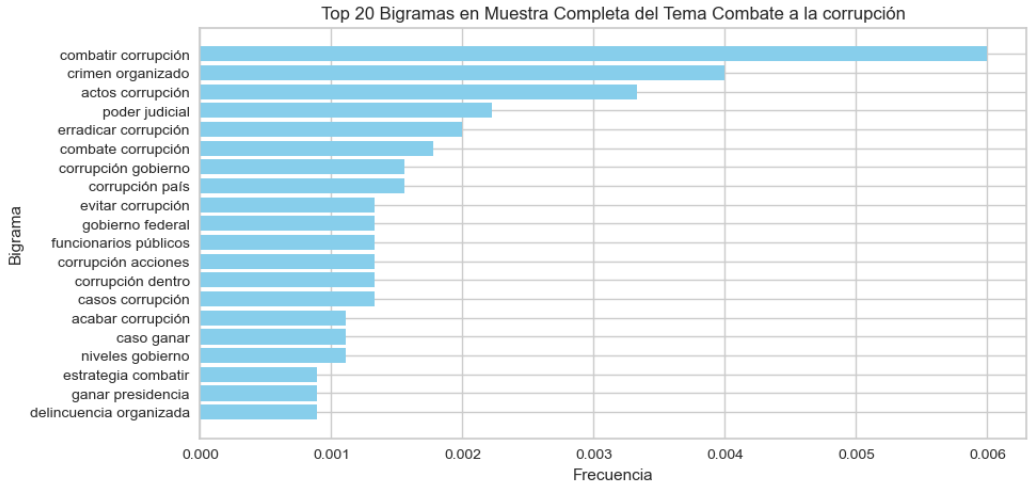
Las palabras más relevantes por número de ocurrencias en *Combate a la corrupción* invitan a las candidaturas a nombrar qué estrategias propondrían desde el gobierno,

[15] El eje horizontal, de "Frecuencia", en todas las listas de enegramas se refiere a "Frecuencia ponderada", no a frecuencia absoluta.
[16] El tamaño de los nodos y de las aristas está en función del número de relaciones entre las palabras y de la frecuencia con que estuvieron unas junto a otras en las preguntas de cada tema.

como indica el nombre del tema, para el combate a la corrupción. En la lista aparecen palabras como “acciones” o “actos”, que dan cuenta de la búsqueda de precisiones en esas estrategias y propuestas, mientras “crimen” y “organizado” cierran la lista e indican a estos grupos como elemento clave a ser tomado en cuenta para el diseño de estrategias contra la corrupción en el país.



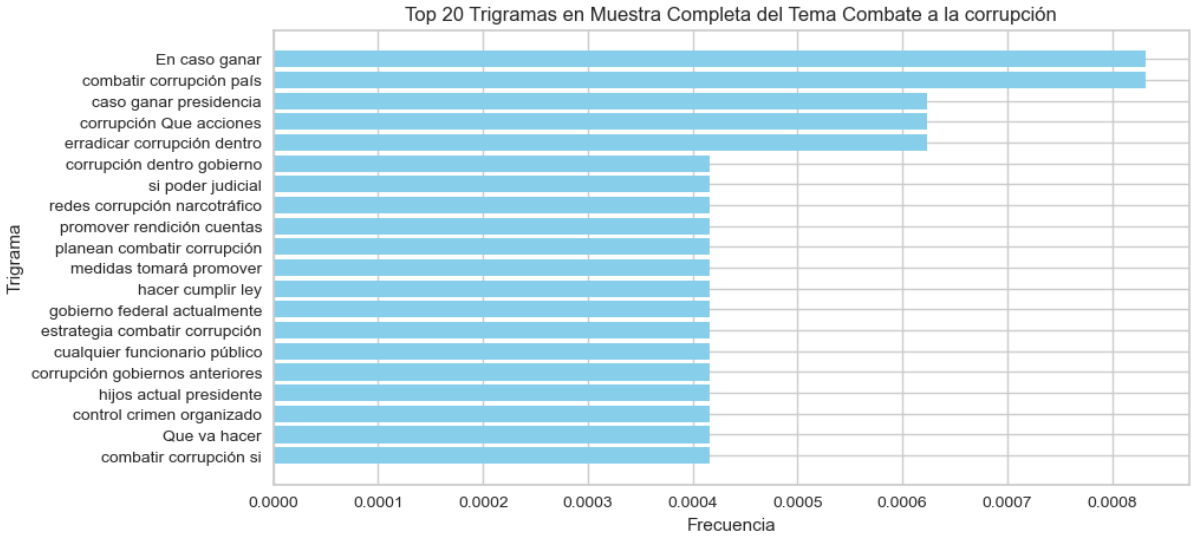
En los enegramas de dos palabras o bigramas se despliegan “poder judicial”, “gobierno federal” y “funcionarios públicos” a lo largo de la lista, lo cual indica preocupaciones acerca de la corrupción en todos los niveles de gobierno. El énfasis en palabras como “erradicar”, “acabar” y “evitar” es claro en cuanto al margen de tolerancia que se espera frente al tema. “Crimen organizado” y “delincuencia organizada” aparecen al inicio y al final de la lista, como parte de un mismo campo semántico presente con distintas palabras.

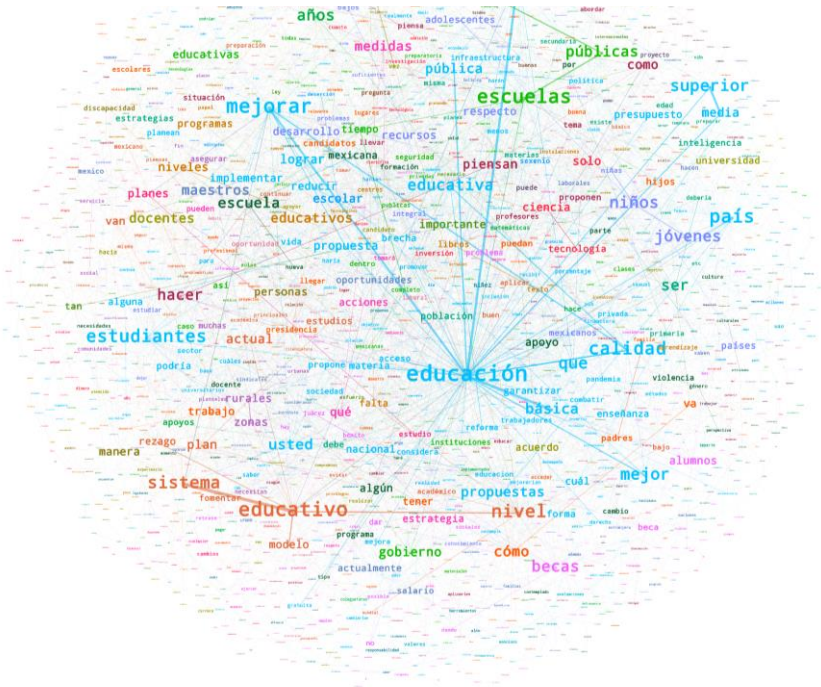


Finalmente, los enegramas de tres palabras enfatizan las inquietudes de las listas previas en tres campos semánticos generales: propuestas para el combate a la corrupción, con grupos de palabras como “combatir corrupción país”, “corrupción Que acciones”, “medidas tomará promover”; delincuencia organizada y corrupción, con

grupos como “control crimen organizado”, “redes corrupción narcotráfico”; y diagnóstico del problema en todos los niveles de gobierno, con núcleos como “cualquier funcionario público”.

2. Educación

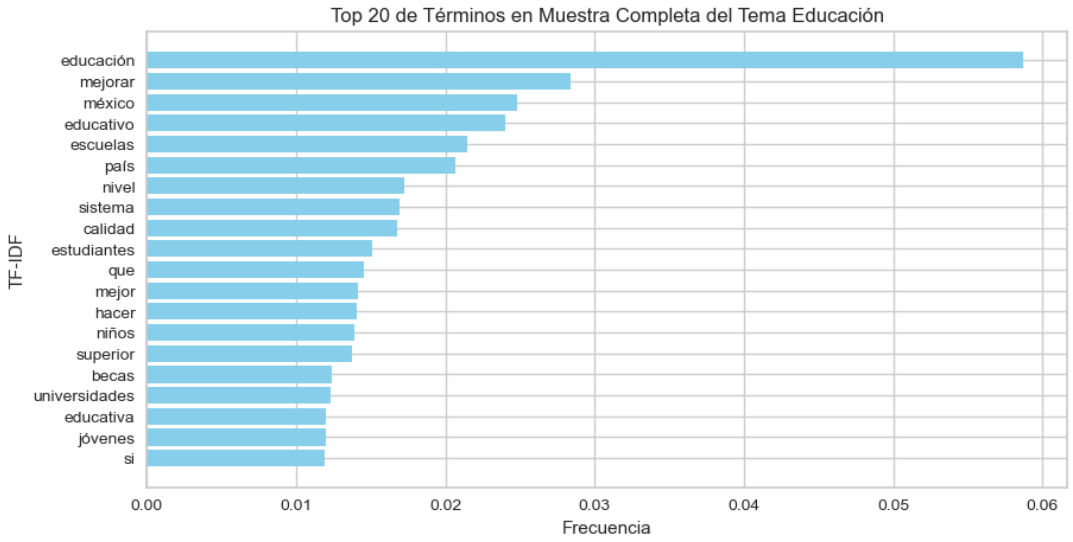




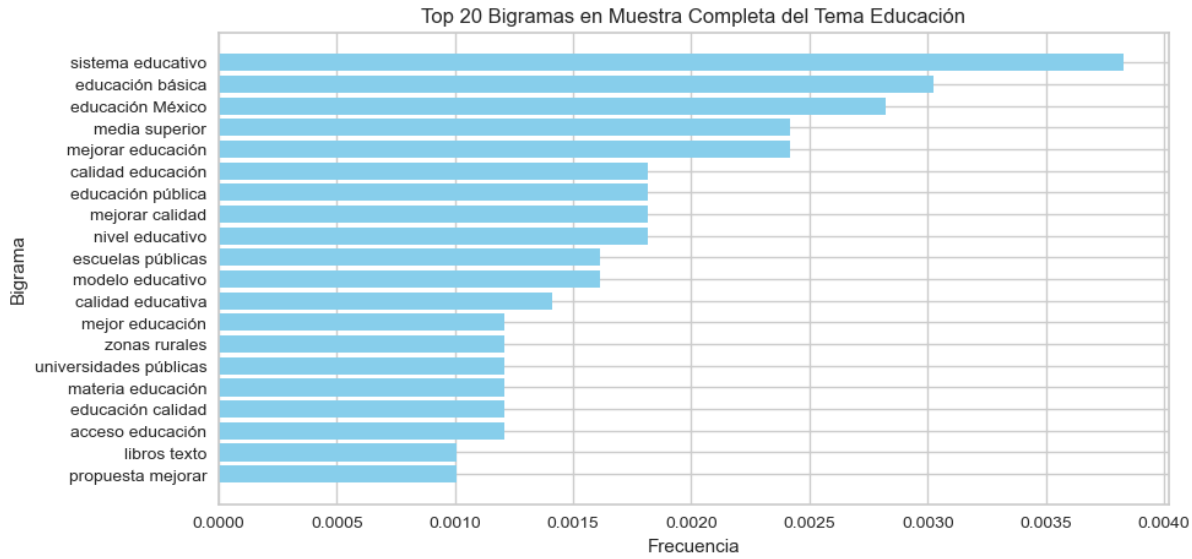
Bigrama de preguntas en muestra del tema: Educación
Nodos: 1,973, Aristas: 4,405

En el tema de educación, las palabras que destacan por frecuencia ponderada señalan inquietudes acerca de la mejora en el sistema educativo a nivel nacional. Las menciones

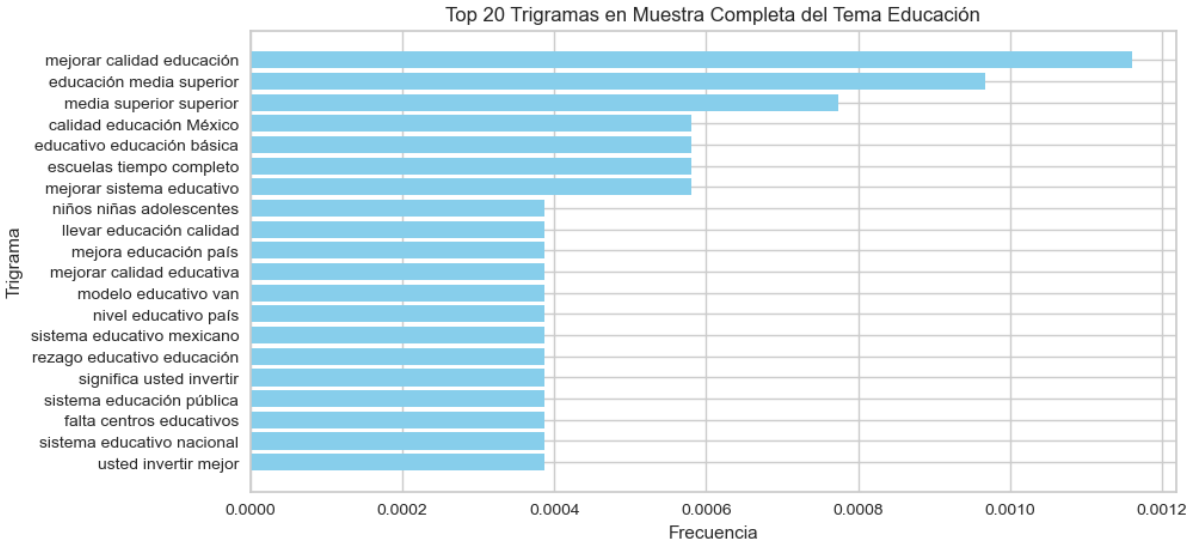
a “jóvenes” y “niños” muestran el rango etario, intergeneracional, que supone la mayor preocupación de la ciudadanía acerca del tema.



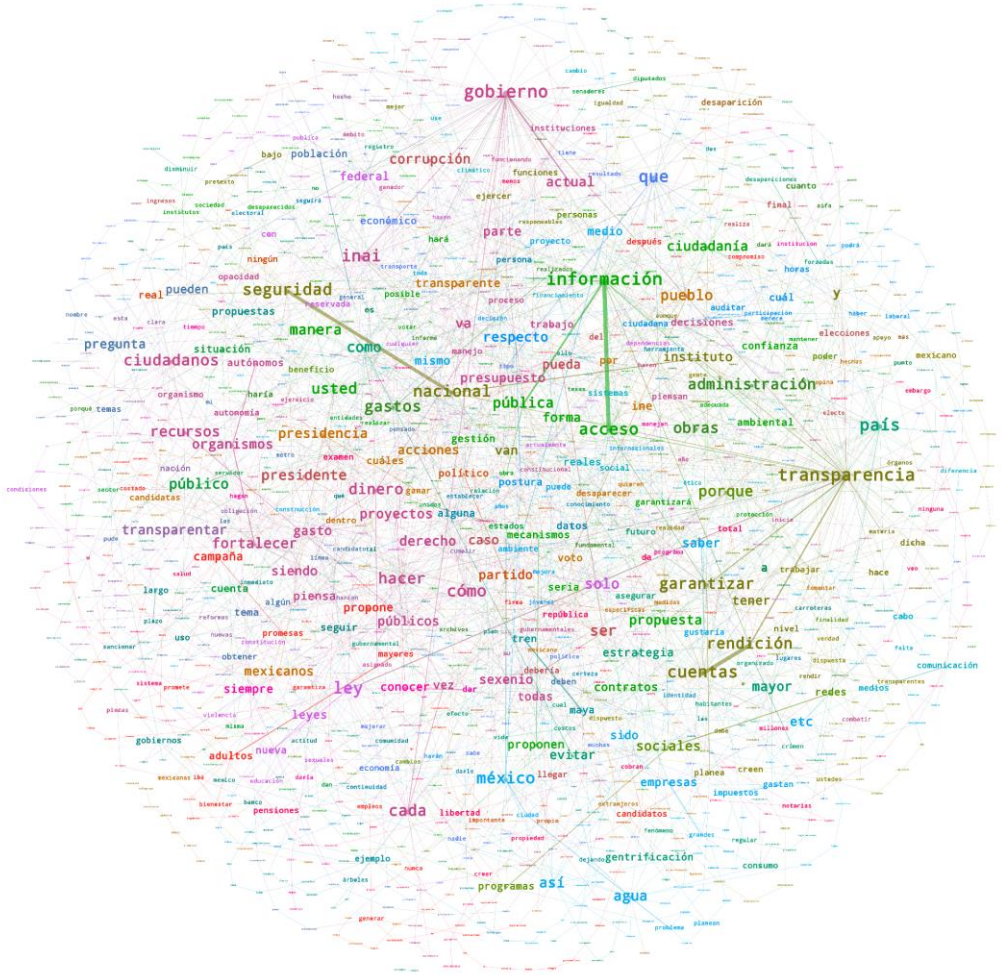
En los bigramas “mejorar educación”, “mejor educación” y “mejorar calidad” enfatizan las expectativas de un cambio en las condiciones educativas actuales. Mientras que “zonas rurales”, “escuelas públicas”, “educación básica” y “universidades públicas” dan cuenta de escenarios específicos donde se esperan las propuestas en este debate acerca de la mejora en la educación.



Con los trigramas se agrega un núcleo semántico desde un diagnóstico deficitario del momento actual en la educación nacional, con conjuntos en el top de menciones como “rezago educativo educación” y “falta centros educativos”. La mención a “sistema” en 4 distintas formulaciones con más de una ocurrencia, amplía la mirada de diagnóstico general de la situación actual, desde la que se elaboraron las preguntas acerca de este tema.



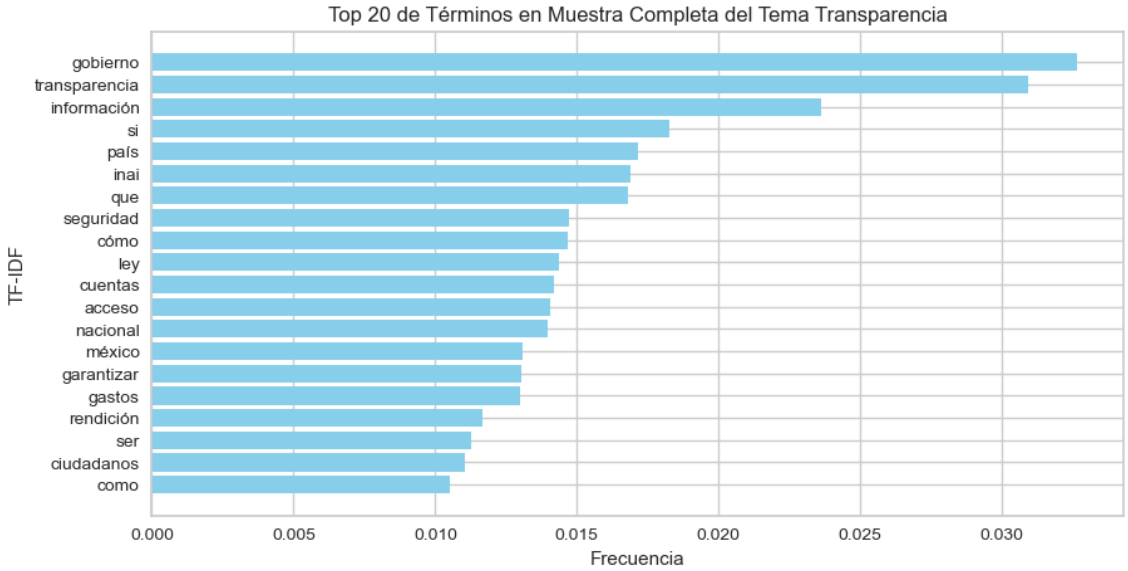
3. Transparencia



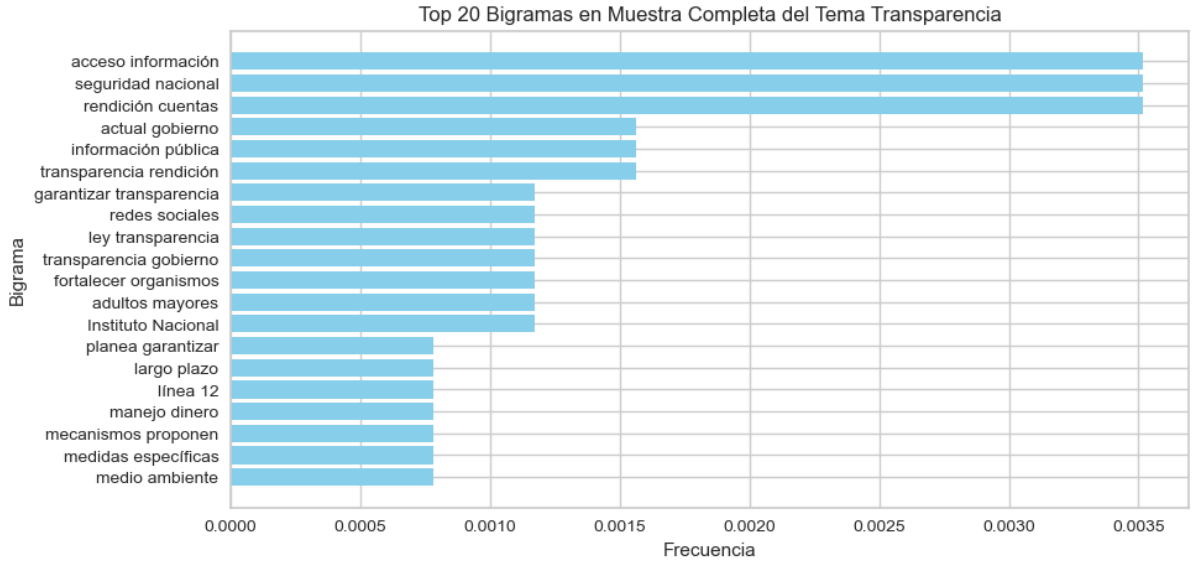
Bigrama de preguntas en muestra del tema: Transparencia
Nodos: 1,455, Aristas: 2,445

La lista de palabras por frecuencia ponderada más relevantes en el tema de transparencia abarca dos campos semánticos principales, uno alrededor de cuentas públicas, con palabras como “rendición”, “gastos”, “acceso”, “cuentas” e “información”,

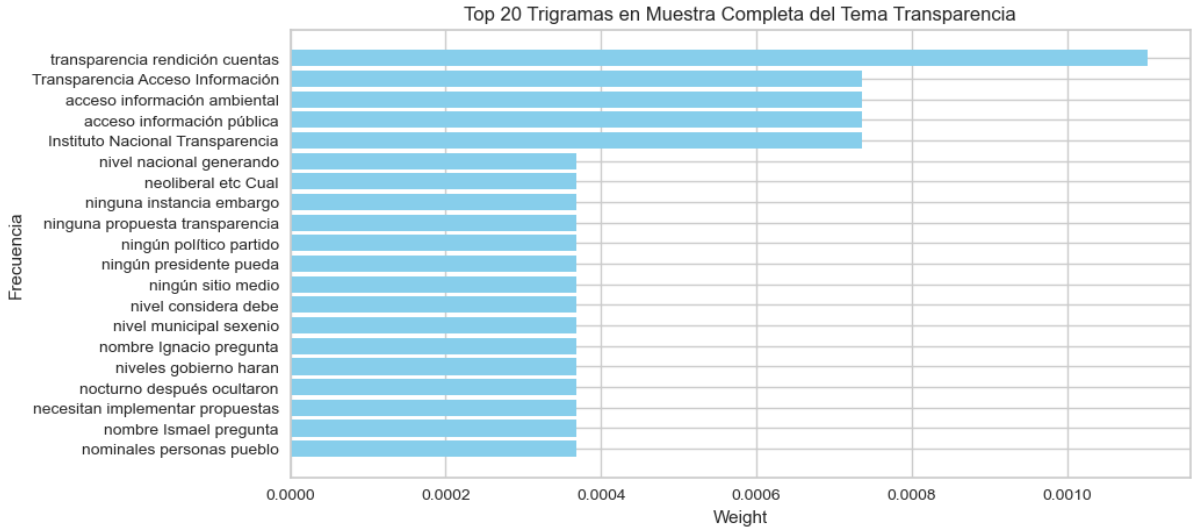
y otro alrededor de las normas y del órgano encargado del tema en México: “inai”, “ley”, “país” y “nacional”.



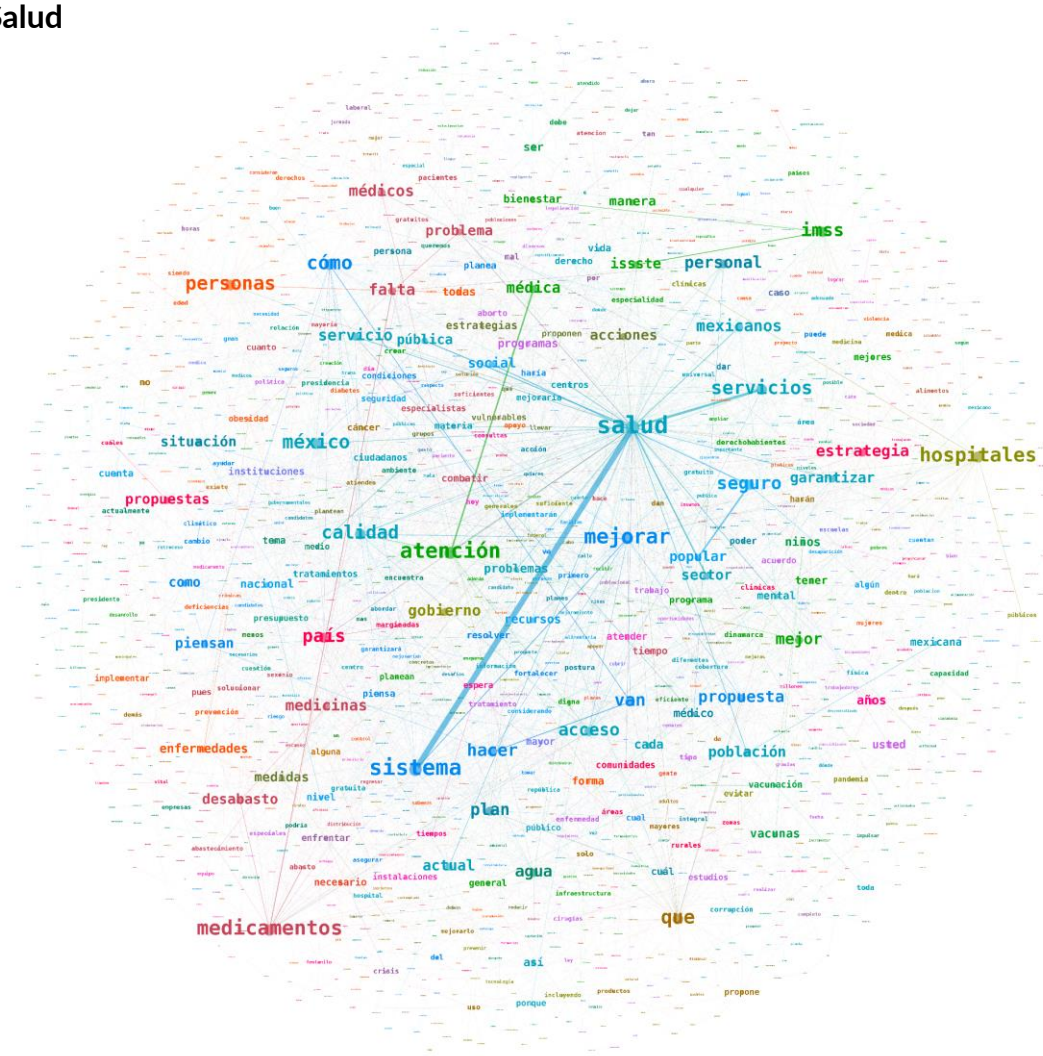
Los bigramas más relevantes dan cuenta de la dispersión en los significados alrededor de este tema. Aparecen pares de palabras como “seguridad nacional”, “redes sociales”, “ley transparencia” o “medio ambiente”, lo cual es indicativo de que fue en este tema en donde se tuvieron que hacer más reemplazos de preguntas en la fase de revisión, pues desde el propio registro no necesariamente estaban ceñidas a la transparencia.



En los trigramas acerca de transparencia se reafirma la rendición de cuentas así como las menciones a “Instituto Nacional Transparencia” como preocupaciones sobre el tema, y al mismo tiempo se amplía la diversidad de los campos semánticos en este tema, pues la mayoría de estos conjuntos tienen muy pocas repeticiones como para distinguirse del resto y poder señalar intereses más específicos.



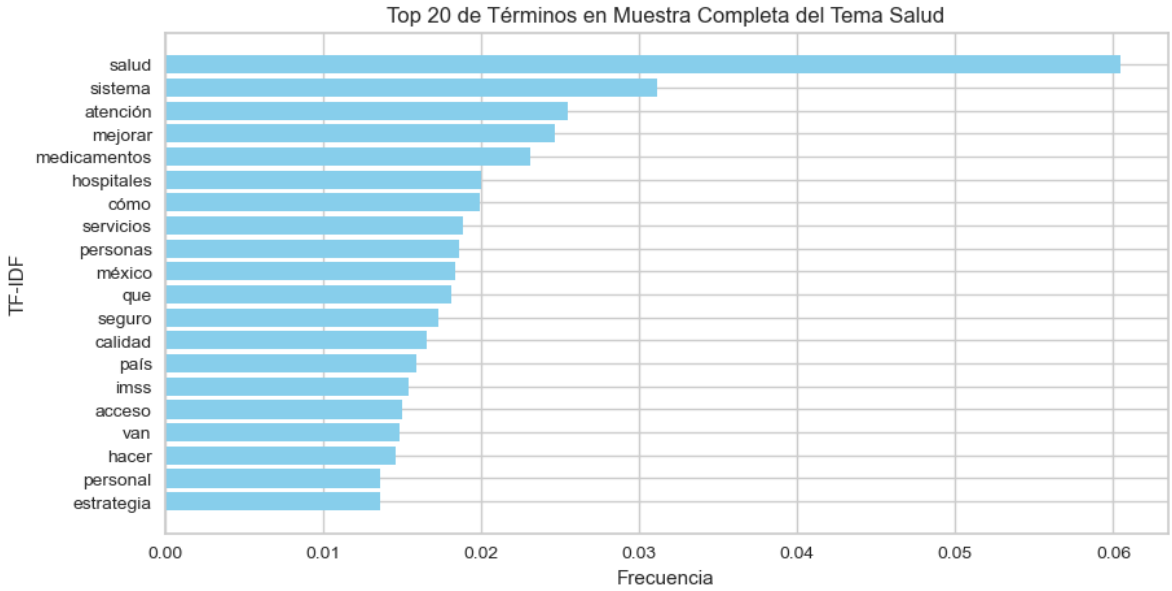
4. Salud



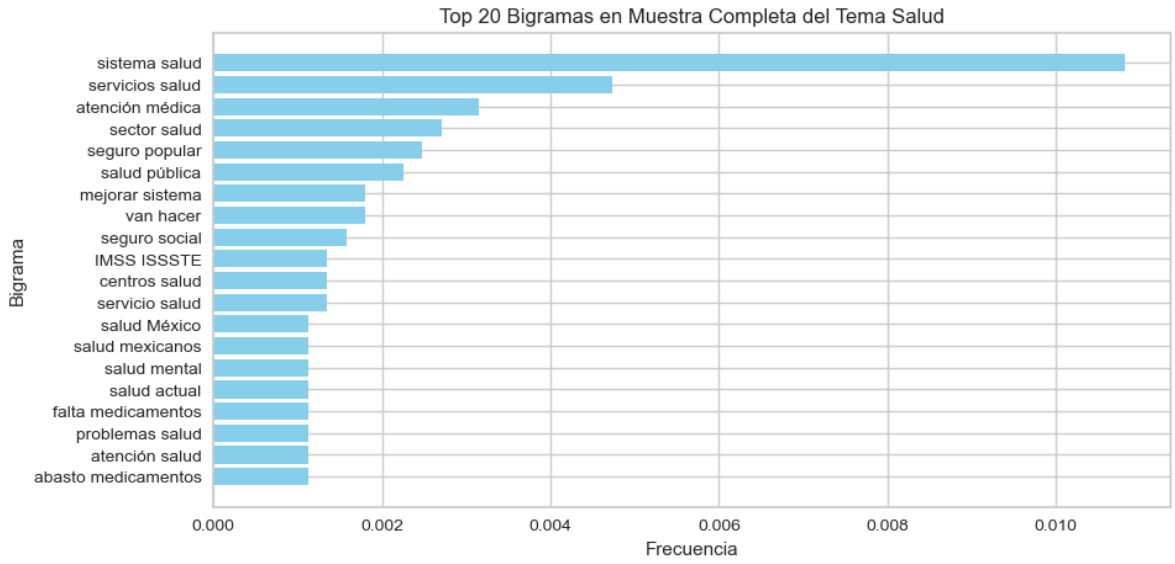
Bigrama de preguntas en muestra del tema: Salud
Nodos: 1,800, Aristas: 3,873

Las palabras más relevantes relacionadas con el tema de salud se centran en la búsqueda de propuestas para la mejora del sistema público de salud. Las palabras

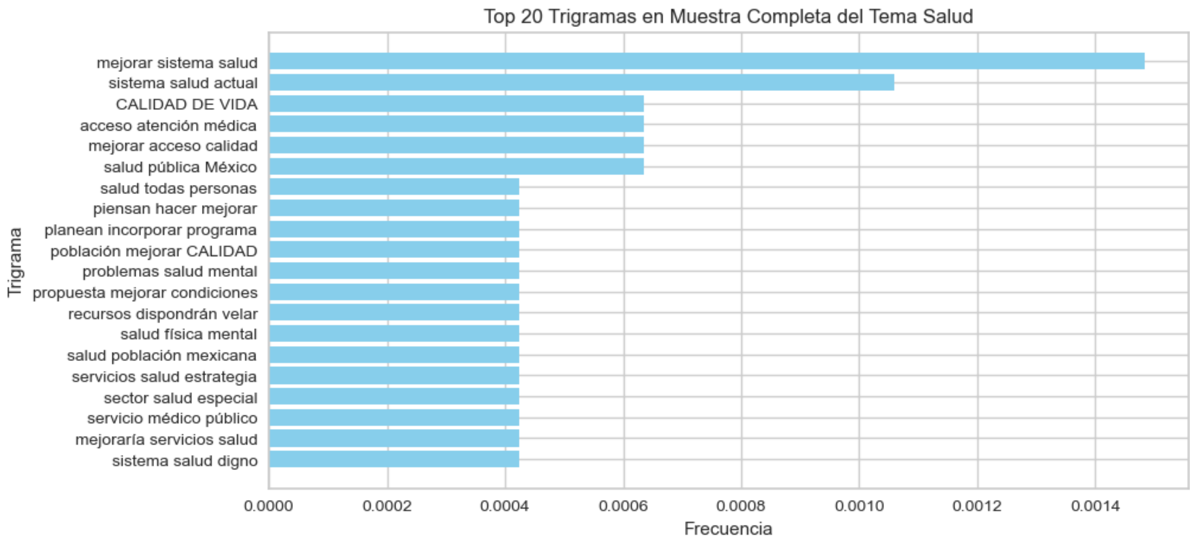
“atención”, “medicamentos” y “hospitales” apuntan a la atención médica actual en los hospitales y la falta de medicamentos.



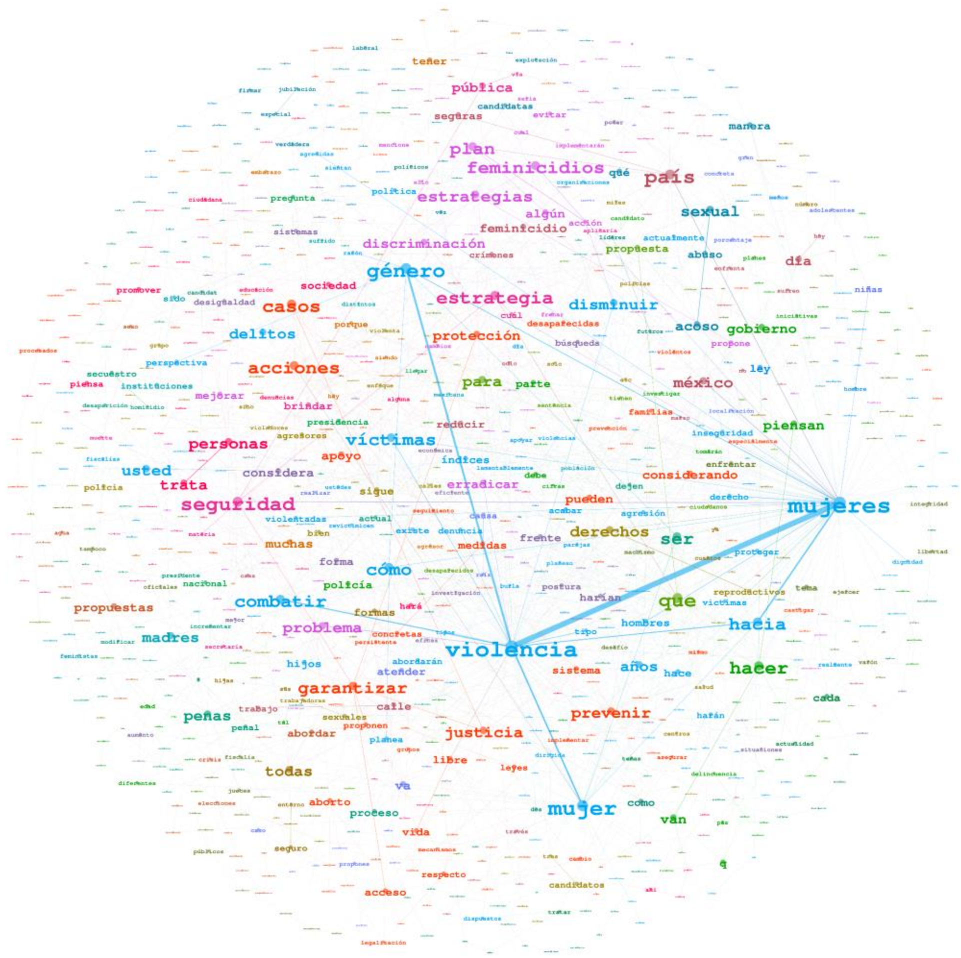
Los bigramas o enegramas de dos palabras muestran que “sistema salud”, “servicios salud”, “seguro popular”, “salud pública”, “mejorar sistema” y “seguro social” son diferentes maneras de referirse a la preocupación por mejorar el sistema de salud pública actual y a programas sociales que garanticen su acceso. Además, se retoma el tema de los medicamentos con “falta medicamentos” y “abasto medicamentos”.



Por último, los enegramas de tres palabras subrayan, con distintas formulaciones, la inquietud por mejorar el sistema de salud actual, como “mejorar sistema salud”, “sistema salud actual”, “mejorar acceso calidad”, “propuesta mejorar condiciones”, “mejoraría servicios salud”, entre otros. Cabe señalar que también menciona la salud mental, con palabras como “problemas salud mental” y “salud física mental”.



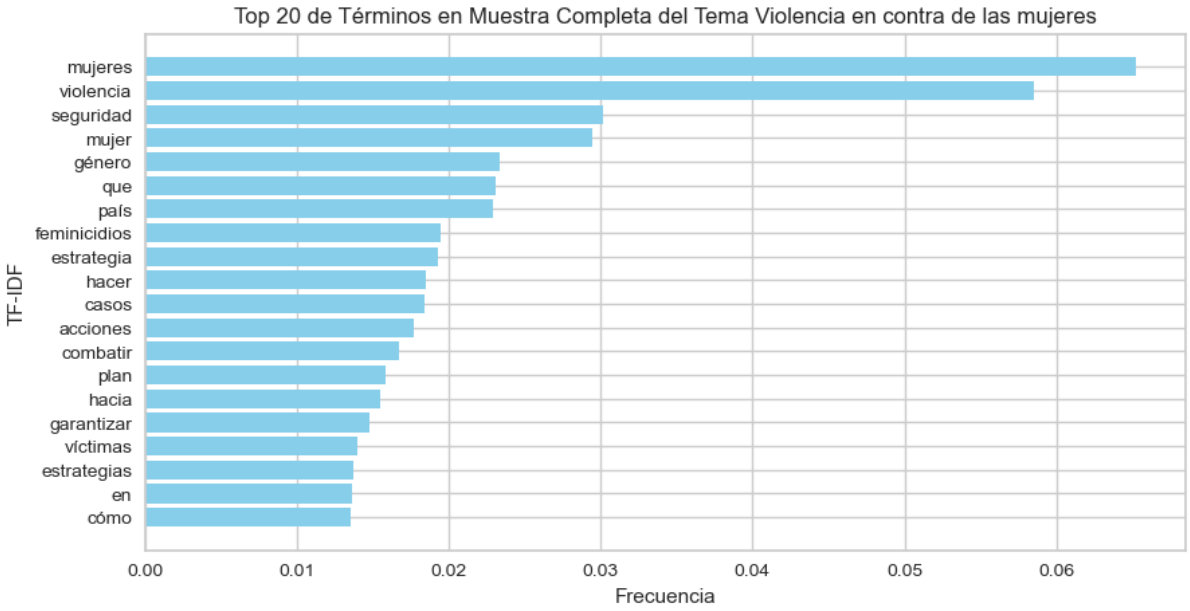
5. Violencia en Contra de las Mujeres



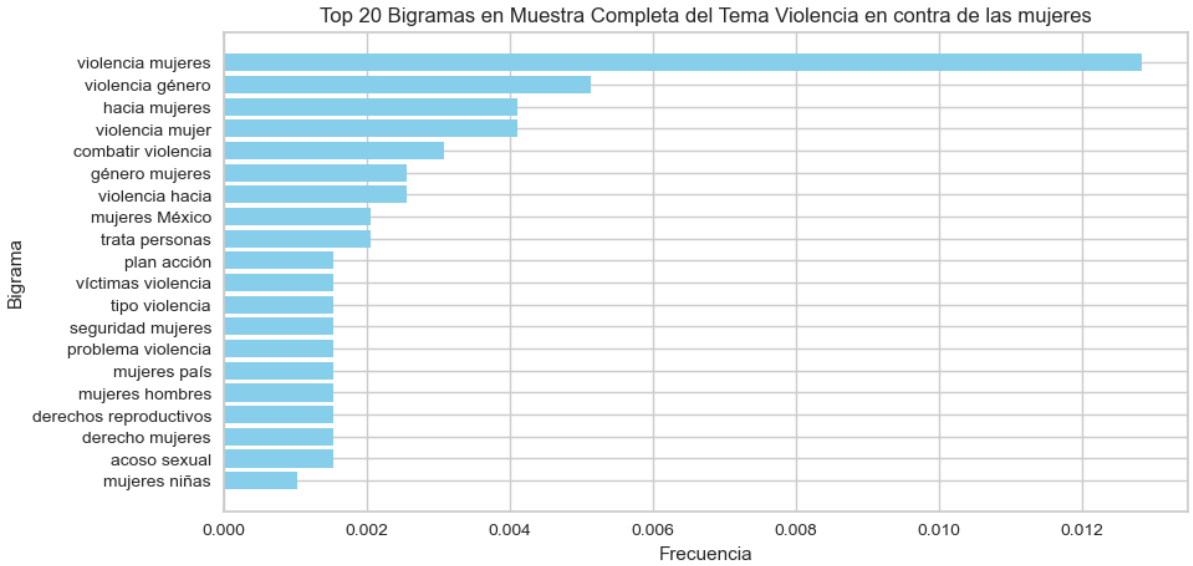
Bigrama de preguntas en muestra del tema: Violencia Contra las Mujeres
Nodos: 1,010, Aristas: 1,766

Las palabras más relevantes por frecuencia ponderada en el tema de violencia contra las mujeres se preocupan, en términos generales, por conocer acciones y estrategias en el combate a este tipo de violencia en México, y por garantizar la seguridad de las

mujeres, aparecen las palabras “violencia”, “mujeres”, “seguridad”, “feminicidios”, “estrategia”, “acciones”, “combatir” y “garantizar”.

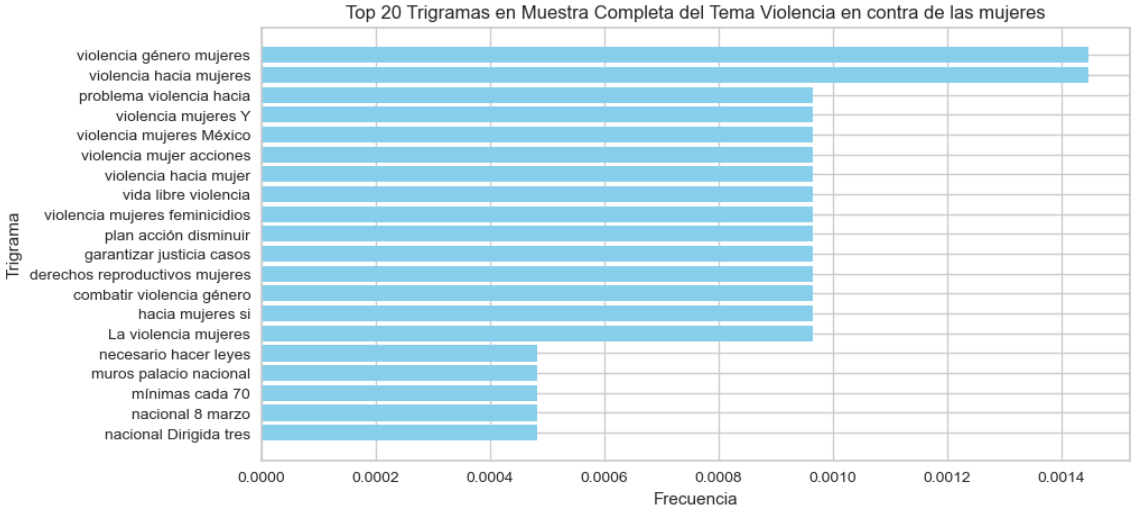


Los bigramas reflejan una manera alternativa para referirse al mismo campo semántico, con palabras como “violencia mujeres”, “violencia género” o “violencia mujer”, pero también dan cuenta de problemas y luchas más específicas, como “trata personas”, “acoso sexual” y “derechos reproductivos”.



En los trigramas acerca de la violencia en contra de las mujeres se precisan formas de esta violencia en México, con núcleos semánticos como “violencia género mujeres”, “violencia mujeres México” o “violencia mujeres feminicidios”. Adicionalmente, se muestran conjuntos de palabras que aluden a la búsqueda de propuestas y leyes para

disminuir la violencia, como “plan acción disminuir”, “garantizar justicia casos”, “combatir violencia género” y “necesario hacer leyes”.

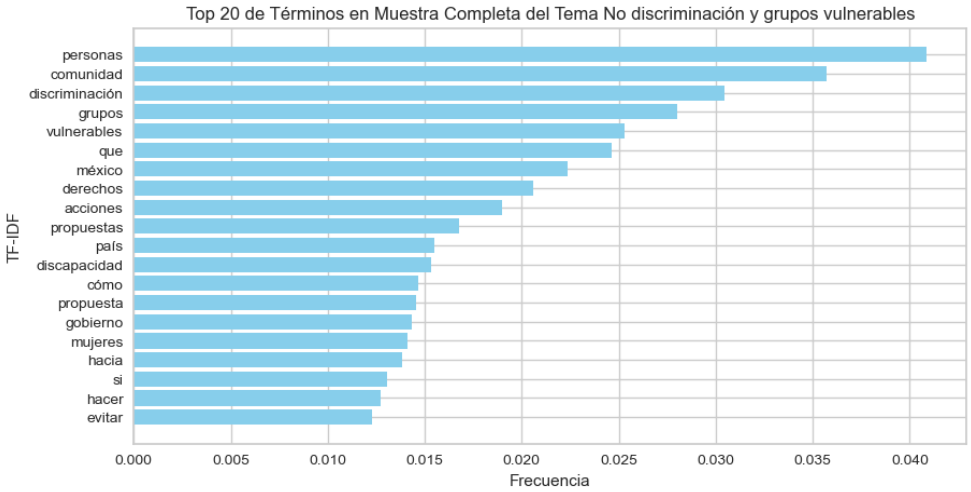


6. No Discriminación y Grupos Vulnerables

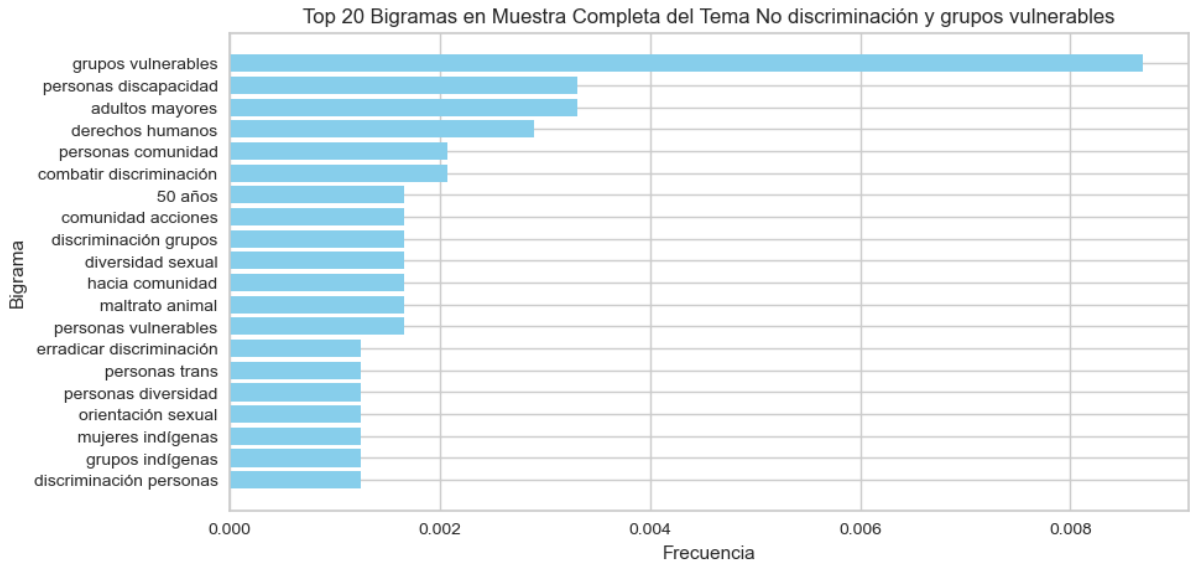
Bigrama de preguntas en muestra del tema: No Discriminación y Grupos Vulnerables
Nodos: 1,264, Aristas: 2,238

La lista de palabras con mayor peso por frecuencia ponderada en este tema nombra a los grupos más mencionados en las preguntas, en las que sobresalen las personas

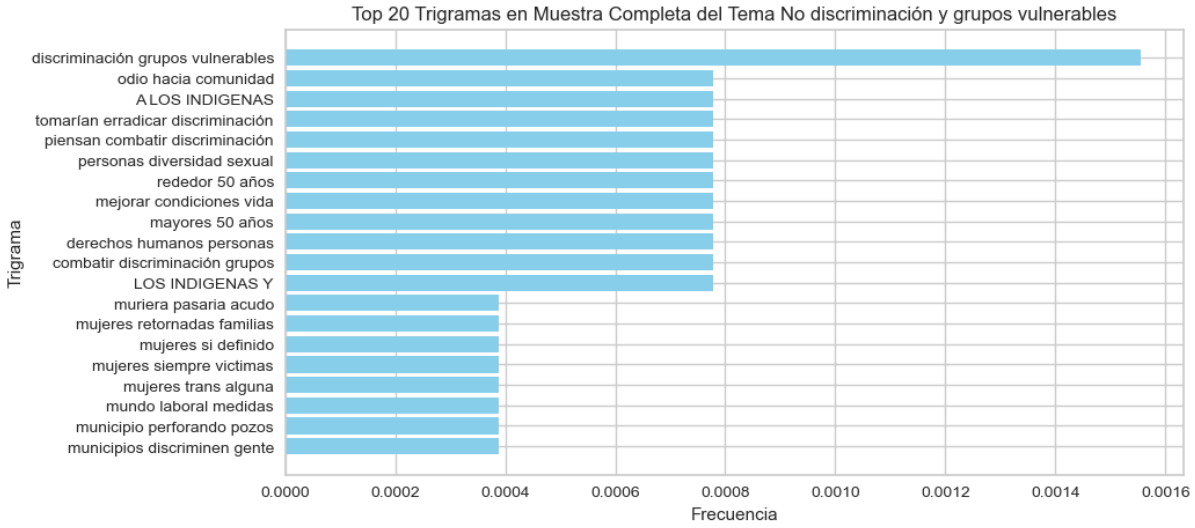
“discapacitadas” y las “mujeres”. A la lista de palabras más relevantes se suman “propuestas” y “derechos” “hacia” estos grupos.



En los bigramas alrededor de este tema aparecen dos campos semánticos principales, uno acerca de la diversidad sexogenérica, con menciones a “personas diversidad”, “orientación sexual”, “diversidad sexual”, “personas trans” ; y otro acerca de comunidades indígenas, específicamente “mujeres indígenas” y “grupos indígenas”. Verbos como “erradicar” y “combatir” dan cuenta del tono de exigencia en las preguntas acerca del tema.



En los trigramas “discriminación grupos vulnerables” aparece en primer lugar, seguido de una listade núcleos semánticos donde destacan “odio hacia la comunidad”, “A LOS INDIGENAS”, “personas diversidad sexual” y “mayores 50 años”. Es decir, alrededor de la no discriminación y grupos vulnerables lo que más aparece son las menciones a los propios grupos vulnerables. Síntoma de una probable sensación de falta de reconocimiento de las desigualdades que padecen quienes pertenecen a estos grupos.



4. CONCLUSIONES

El abanico de la participación

La participación de personas identificadas como pertenecientes a grupos en situación de discriminación fue del 42% del total de los registros. Queda este como un antecedente para futuros ejercicios de este tipo, que intenten medir y dar cuenta de la participación de personas sistemática e históricamente excluidas en México. Los adultos mayores, las personas de la diversidad sexual y los grupos indígenas encabezan la lista de estos registros.

Los jóvenes de los 13 a los 27 años fueron los que más participaron en este ejercicio. Sus intereses estuvieron colocados sobre todo en *Educación*, *Salud* y *Combate a la corrupción*. Destaca la alta participación del rango etario entre 18 y 22 años, pues serán las y los jóvenes que votarán por primera vez en elecciones federales. Los adultos mayores a partir de los 68 años fueron los que menos participaron.

Por otro lado, los adultos de entre 38 y 47 años tuvieron un pico a la baja en su participación en este ejercicio, en comparación con los rangos etarios inmediatamente superiores e inferiores, es decir, las personas con mayor margen de representatividad institucional, de mayor acceso a la toma de decisiones y que además entran en el rango más reconocido de población económicamente activa, no pusieron tanta atención a este ejercicio.

Con respecto a la participación por regiones, la región centro tuvo 61% de los registros totales, mientras que las regiones sur y norte tuvieron el 21% y el 16% respectivamente.

Las entidades con mayor participación fueron la capital (3,223 registros), el Estado de México (1,911 registros), Jalisco (850 registros), Veracruz (761 registros) y Puebla (553 registros).

Los adultos mayores fueron el grupo en situación de discriminación con mayor número de registros en las tres regiones (15.72% en la región centro, 21.20% en la región norte y 15.94% en la región sur), mientras que para las regiones centro y norte el segundo grupo en situación de discriminación con mayor participación fue el de *personas de la diversidad sexua*l (con 11.32% y 12.37% respectivamente), y en la región sur el segundo lugar de participación fue el de *personas indígenas* (con 10.5%).

Diagnósticos y preocupaciones

Los temas de *Salud* y *Educación* compartieron una mirada en tono de diagnóstico deficitario acerca de las condiciones actuales a nivel nacional en ambos rubros. Las listas de campos y núcleos semánticos de estos temas mostraron intereses claros en cuanto a propuestas de mejora en las infraestructuras de los sistemas educativo y de salud a nivel nacional.

En cuanto a *Salud*, el énfasis estuvo en la falta de abastecimiento de medicamentos y la mejora en los hospitales, y acerca de la *Educación*, el énfasis estuvo en el mejoramiento del nivel educativo en todos los niveles, particularmente en escuelas y universidades públicas.

La atención a la salud mental es una preocupación creciente a nivel nacional, entre las preguntas se presentaron incertidumbres acerca de la capacitación a maestros para atender alumnos con capacidades diferentes y el acceso a la atención de especialistas.

Acerca del *Combate a la corrupción*, las palabras más visibles enfatizaron un tono de tolerancia cero frente a ésta de parte de la ciudadanía, y señalaron la necesidad de combatirla en todos los niveles de gobierno. También quedó claro que para quienes participaron en el ejercicio no se puede pensar en este tema sin propuestas claras de cómo lidiar con el crimen organizado.

La *Violencia en contra de las mujeres* fue considerada como un problema de seguridad que requiere estrategias concretas. Feminicidios y trata de personas fueron señaladas como formas de violencia de género extremas y que requieren de mayor atención. Sobresalieron inquietudes alrededor de propuestas de ley y formas de garantizar vías institucionales para hacerle frente al abanico de problemas que incluye la violencia de género.

Acerca de la *No discriminación y grupos vulnerables*, los grupos más mencionados fueron las personas de la diversidad sexual, personas discapacitadas, grupos indígenas y mujeres. El tono principal de los núcleos y campos semánticos en este tema dio cuenta

de una necesidad de mayor reconocimiento institucional de las desigualdades que enfrentan estos grupos, y que sería el punto de inicio para un cambio en estas condiciones.

Finalmente, acerca de la *Transparencia* hay mucha dispersión y se entiende de distintas formas por distintos grupos. Sin embargo, pese a esta dispersión es posible identificar inquietudes claras acerca de la rendición de cuentas y del acceso a la información pública, garantizada por parte de los gobiernos y las instancias nacionales correspondientes.

5. LISTA DE REFERENCIAS

Bird, S., Loper, E. & Klein, E. (2009). *Natural Language Processing with Python*. O'Reilly Media Inc.

Guzmán Falcón, E. (2018). *Detección de lenguaje ofensivo en Twitter basada en expansión automática de lexicones*. Instituto Nacional de Astrofísica, Óptica y Electrónica. https://inaoe.repositorioinstitucional.mx/jspui/bitstream/1009/1722/1/GuzmanFE.pdf

Hunston, S. (2022). *Corpora in Applied Linguistics* (2a ed.). Cambridge University Press. https://doi.org/10.1017/9781108616218

Instituto Nacional Electoral. (2024). *Acuerdo INE/CG95/2024. Acuerdo del Consejo General del Instituto Nacional Electoral por el que se define la metodología, así como la instancia que seleccionará las preguntas provenientes de redes sociales relativas al Formato Tipo A que se utilizará en el Primer Debate Presidencial en el Proceso Electoral Federal 2023-2024*. Repositorio Documental INE. https://repositoriodocumental.ine.mx/xmlui/bitstream/handle/123456789/164296/CGex202402-08-ap-3.pdf

Instituto Nacional Electoral. (2024). *Anexo I. Metodología Selección de Preguntas para Debate Formato A*. Repositorio Documental INE. https://repositoriodocumental.ine.mx/xmlui/bitstream/handle/123456789/164296/CGex202402-08-ap-3-a1.pdf

Kiss, T., & Strunk, J. (2006). *Unsupervised Multilingual Sentence Boundary Detection.* Computational Linguistics, 32(4), 485-525. https://doi.org/10.1162/coli.2006.32.4.485

Longhi, J. (2021). *Mapping information and identifying disinformation based on digital humanities methods: From accuracy to plasticity*. *Digital Scholarship in the Humanities*, 36 (4), 980–998. https://doi.org/10.1093/llc/fqab005

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426. https://arxiv.org/abs/1802.03426

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., & others. (2011). *Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research*, 12, 2825–2830.

Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (arXiv:1908.10084). arXiv. http://arxiv.org/abs/1908.10084

Rousseeuw, P. (1987). *Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis*. Computational and Applied Mathematics. 20: 53–65. doi:10.1016/0377-0427(87)90125-7.

Signa_Lab ITESO. (2022). *Lexicon para identificar violencia con razón de género*. Signa_Lab ITESO.

Spärck-Jones, K. (1972). *A statistical interpretation of term specificity and its application in retrieval*. Journal of Documentation, 28(1), 11-21. https://www.staff.city.ac.uk/~sbrp622/idfpapers/ksj_orig.pdf

Thorndike, R. (1953). *Who Belongs in the Family?*. Psychometrika. 18 (4): 267–276. doi:10.1007/BF02289263. S2CID 120467216.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). *Multilingual E5 Text Embeddings: A Technical Report* (arXiv:2402.05672). arXiv. http://arxiv.org/abs/2402.05672

7. DIRECTORIO DE ANEXOS

Anexo 1. Bitácoras y documento de seguimiento del INE

Anexo 2. Compilación de gráficas de los 13,484 registros recibidos

Anexo 3. Compilación de gráficas de la población depurada de 21,219 preguntas

Anexo 4. Compilación de gráficas de análisis semántico por tema en la muestra

Anexo 5. Archivo de datos tabulares (CSV) con los 519 términos proscritos del diccionario de depuración desarrollado por Signa_Lab ITESO

Anexo 6. Archivo de datos tabulares (CSV) con los 13,484 registros recibidos

Anexo 7. Archivo de datos tabulares (CSV) con las 24,000 preguntas y su ID generado

Anexo 8. Archivo de datos tabulares (CSV) con las 21,219 preguntas de la población depurada

Anexo 9. Archivo de datos tabulares (CSV) con las 2,781 preguntas descartadas

Anexo 10. Archivo de datos tabulares (CSV) con las 1,701 preguntas de la muestra estratificada

Anexo 11. Archivos de datos tabulares (CSV) con preguntas preseleccionadas y sus respectivas rondas de revisión

- **11.1** 1ra Preselección 108 preguntas
- **11.2** 1ra Ronda de revisión 108 preguntas
- **11.3** 2nda Preselección 108 preguntas
- **11.4** 2nda Ronda de revisión 108 preguntas
- **11.5** 3ra Preselección 108 preguntas
- **11.6** 3ra Ronda de revisión 108 preguntas
- **11.7** 4ta Preselección 108 preguntas
- **11.8** 4ta Ronda de revisión 108 preguntas
- **11.9** 5ta Preselección 108 preguntas
- **11.10** 5ta Ronda de revisión 108 preguntas

Anexo 12. Archivo de datos tabulares (CSV) con las 108 preguntas en selección final

Anexo 13. Documentación de código desarrollado por Signa_Lab ITESO para la aplicación de la metodología en cuadernos de programación (*Python*)

- **13.1** Cuaderno 01. Depuración, A.E.D. y Elaboración de Muestra Estratificada
- **13.2** Cuaderno 02. Generación de Relaciones Semánticas (*Embeddings*)
- **13.3** Cuaderno 03 (Combate a la Corrupción). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad
- **13.4** Cuaderno 03 (Educación). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad
- **13.5** Cuaderno 03 (Salud). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad
- **13.6** Cuaderno 03 (Transparencia). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad
- **13.7** Cuaderno 03 (Violencia en Contra de las Mujeres). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad

13.8 Cuaderno 03 (No Discriminación y Grupos Vulnerables). Análisis Semántico, Selección por Frecuencia y por Aleatoriedad

13.9 Cuaderno 04. Revisión de preguntas.

13.10 Cuaderno 05. Análisis exploratorio de resultados