Tlaquepaque, Jalisco a 9 de enero de 2024

Elia Baltazar
Directora de Comunicación y Análisis Informativo
Coordinación Nacional de Comunicación Social
INE

Estimada Elia

Revisamos con detalle el documento con la propuesta metodológica para el procesamiento de las preguntas que la ciudadanía realizará a través de redes sociodigitales -previa convocatoria del INE-, a las candidaturas a la presidencia durante el tercer debate presidencial.

Es un gusto para mí informarte que el Laboratorio cuenta con la experiencia, el equipo humano, las herramientas de procesamiento de grandes volúmenes de datos, así como de análisis y procesamiento de lenguaje natural a través de distintos algoritmos, para atender los requerimientos que ustedes plantean en su metodología.

Estamos considerando diez días naturales de trabajo intenso a partir de la recepción de la base de datos con las preguntas previamente sistematizadas por el Instituto. Trabajaremos en tres fases o etapas para arribar a los requerimientos claramente expresados en el documento:

I. **Depuración inicial y ponderación** de acuerdo a criterios de elegibilidad y representatividad
II. **Análisis inicial de la rutas semánticas y clusterización**, que sigan las preguntas para arribar a las frecuencias, similitudes y discrepancias. Identificación y agrupación de tópicos comunes a partir de procesamiento de lenguaje natural (PNL), modelos de lenguaje y visualización exploratoria de texto (árboles semánticos, mapas de relaciones semánticas o *embeddings*).
III. **Sistematización y elaboración de selección muestral estratificada de preguntas**, de la población ponderada de preguntas seleccionables, selección de preguntas para el debate, por estratos a partir de los temas previstos para el debate y su frecuencia. Revisión final, a partir de PLN y revisión manual (con criterios como frecuencia por adverbios, por ejemplo, peso de los cómos, los qué, los en dónde, etc.)

El resultado de cada fase será transparentado y claramente documentado.

A continuación, la cotización del trabajo a realizar

Equipo Humano	284 355.00		
Herramientas, algoritmos y procesamiento LN (lenguaje natural)	85 306.50		
Costos indirectos: computadoras, equipo, instalaciones	49 818.00		
Total	**419 480.00 + iva**		

Propuesta adicional
Producción de contenidos interactivos postdebate:
 	
• **Visualizaciones de respuestas a formulario** *preguntar viabilidad de acuerdo a Acuerdo de Confidencialidad
 	
  • Nubes y árboles de palabras de preguntas recibidas.
  • Mapas interactivos de embeddings por categoría de preguntas recibidas.
  • Gráficas con resumen de datos demográficos (top estado, edad, género, tema)
 	
• **Visualizaciones de datos de redes y plataformas:**
 	
  • Nubes y árboles de palabras de comentarios a videos de debates y de transcripción automática de participaciones de candidatos (Whisper)
  • Mapas interactivos de embeddings de
	100 000.00		

comentarios a debates. • Redes de co-comentarios en Youtube. • Nubes y árboles de palabras de titulares de prensa sobre cobertura de debate (Google/GoogleNews) Mapas interactivos de embeddings de palabras de titulares de prensa sobre cobertura de debate (Google/GoogleNews)			

Quedo por supuesto a tus órdenes para resolver cualquier duda o aclaración que fuese necesaria.

Saludos cordiales

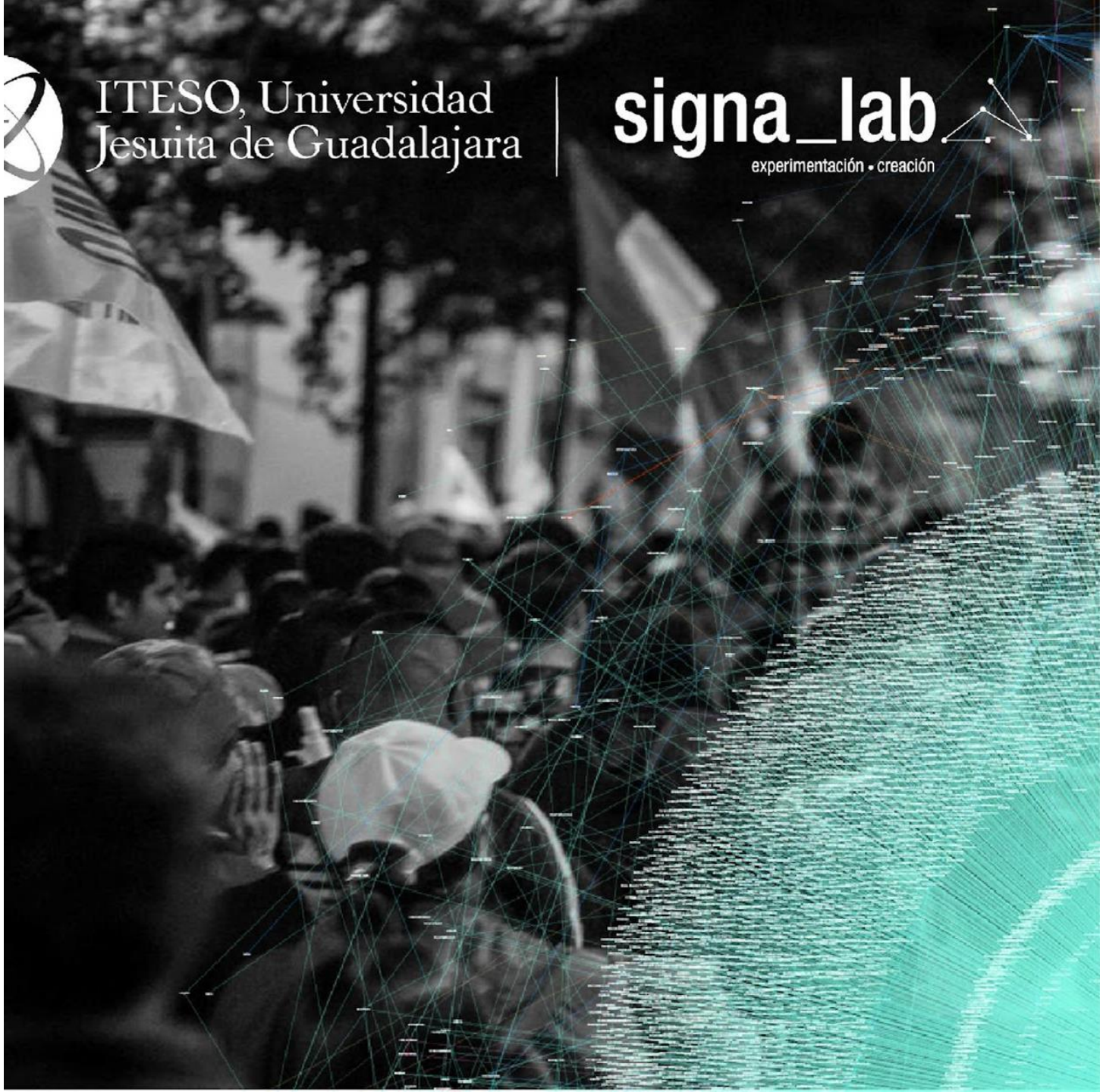
Dra. Rossana Reguillo
Profesora-investigadora Emérita
Coordinadora Signa_Lab
ITESO

INE 2024
propuesta de depuración, análisis y clusterización de preguntas para el 1er debate

@Signa_Lab ITESO

Signa_Lab, es un laboratorio de innovación tecnológica y estudios interdisciplinarios aplicados del ITESO

- Abrió en 2016, pero los trabajos de preparación para el laboratorio empezaron en 2013.
- Conformado por un equipo de profesores-investigadores con formaciones diversas de la antropología a la ingeniería en sistemas y la comunicación y un equipo de becarias y becarios de distintas carreras y posgrados



INVESTIGACIÓN EXPERIMENTACIÓN DESARROLLO TECNOLÓGICO

Somos un espacio interdisciplinario en el que generamos conocimiento, metodologías y herramientas para la comprensión multidimensional del mundo sociodigital.

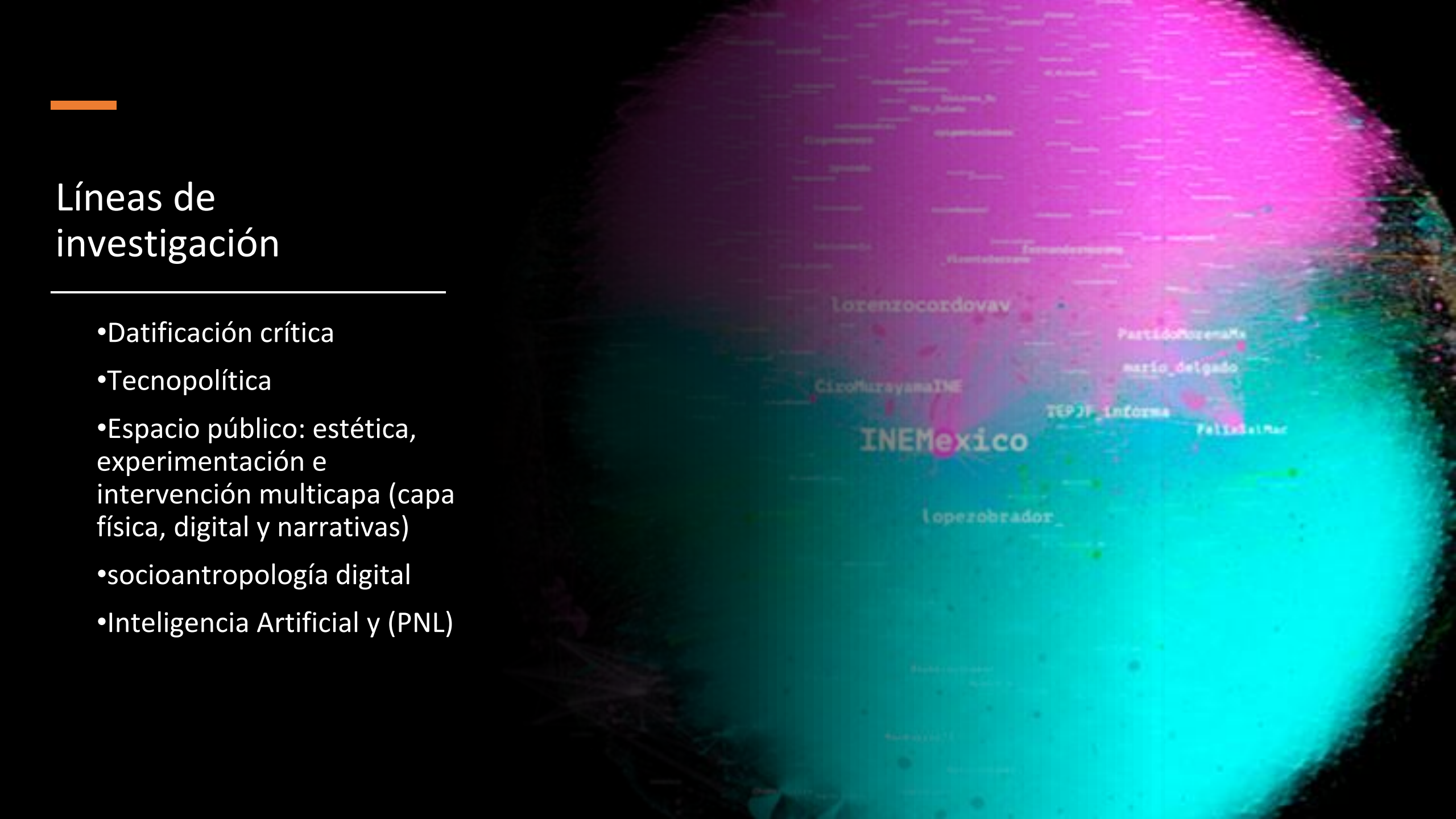
LABORATORIO

Buscamos el conocimiento abierto; la cultura colaborativa y horizontal; propiciamos la acción conectiva a través de la información y la participación; la experimentación como modelo de aprendizaje; fomentamos el trabajo interdisciplinario; procuramos fortalecer la lectura crítica de la realidad y propiciar el desarrollo de tecnologías y metodologías para la innovación social.

- Investigación (interdisciplinaria y multicapa)
- Diseño y producción de narrativas a través de visualización de datos y desarrollo de interfaces interactivas.
- Uso y desarrollo de herramientas para la gestión y el análisis de datos.

Líneas de investigación

- Datificación crítica
- Tecnopolítica
- Espacio público: estética, experimentación e intervención multicapa (capa física, digital y narrativas)
- socioantropología digital
- Inteligencia Artificial y (PNL)



Espacialidades
la red digital como

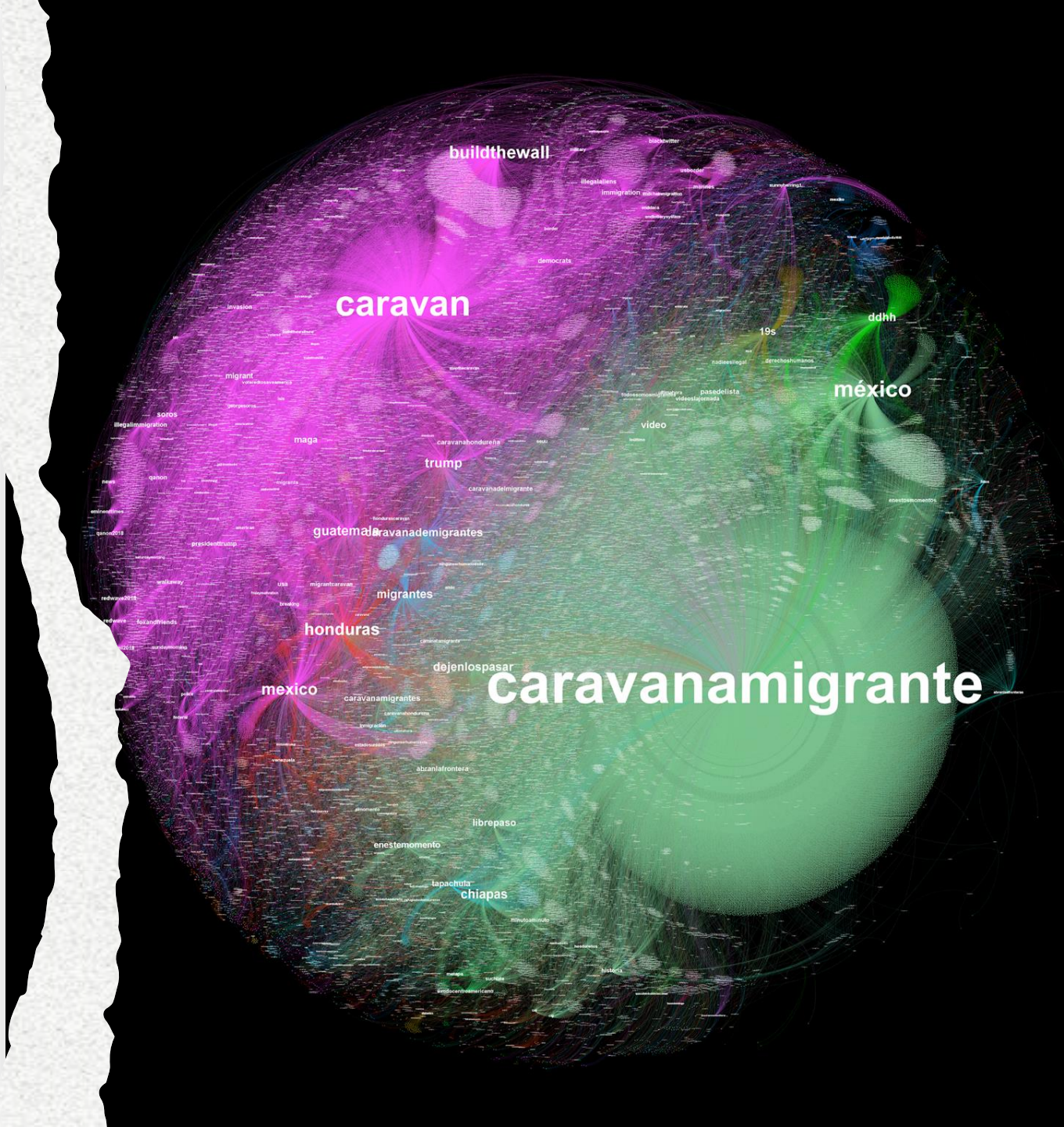
(y no como un dominio alterno o diferente)

- Espacio
- Lugar
- Territorio


Violencia digital
Lexicón para Twitter

Efectos

- Tiempo de viralización
- Intensidad
- Zonas de polarización

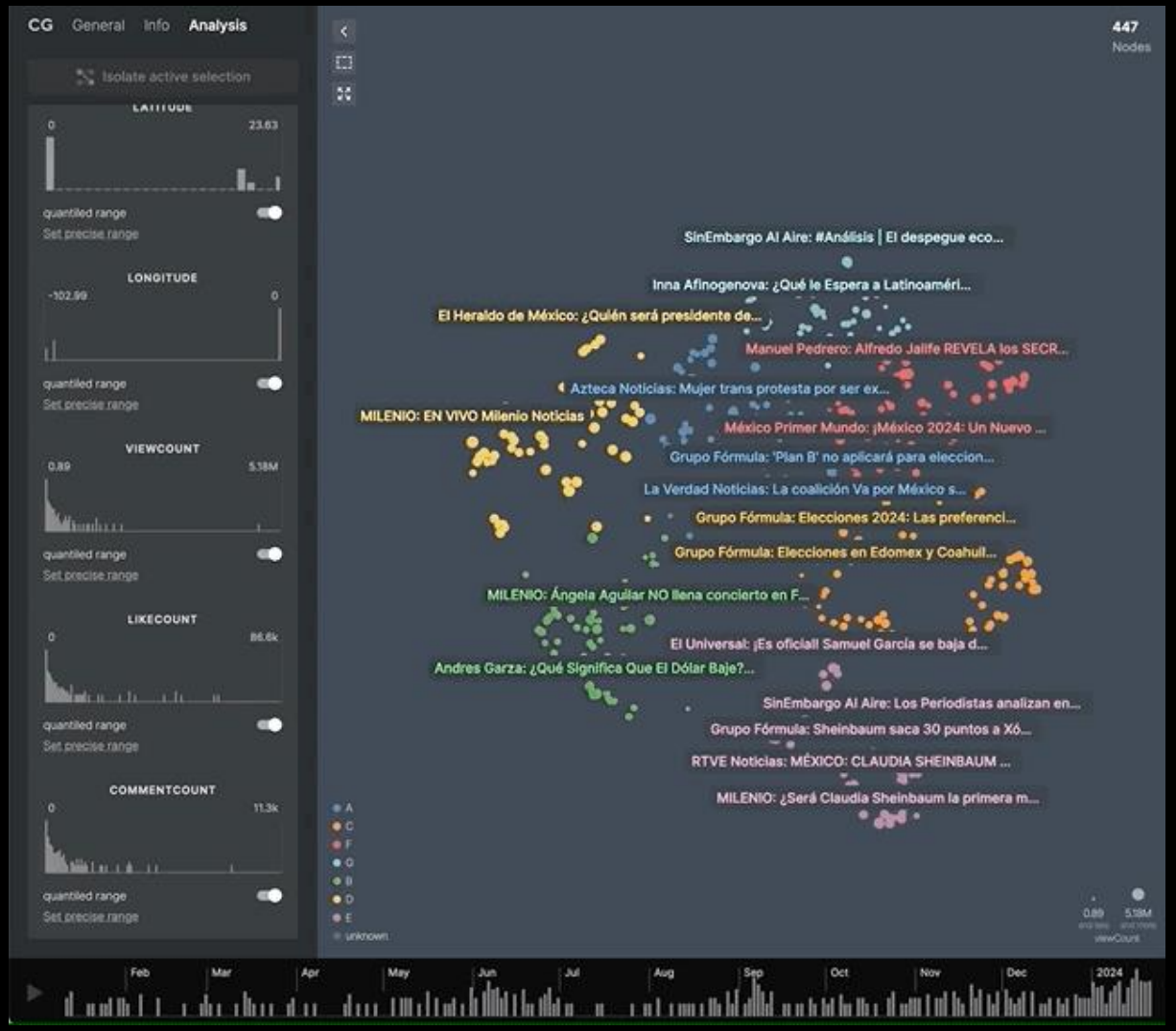


Fases de la propuesta

1.Depuración inicial y ponderación de acuerdo a criterios de elegibilidad y representatividad

2.Análisis inicial de la rutas semánticas y clusterización, que sigan las preguntas para arribar a las frecuencias, similitudes y discrepancias. Identificación y agrupación de tópicos comunes a partir de procesamiento de lenguaje natural (PLN), modelos de lenguaje y visualización exploratoria de texto (árboles semánticos, mapas de relaciones semánticas o *embeddings*).

3.Sistematización y elaboración de selección muestral estratificada de preguntas. A partir de la participación de la ciudadanía y de la ponderación de preguntas seleccionables de acuerdo con los criterios del INE; selección de preguntas para el debate, por estratos a partir de los temas previstos y de su frecuencia. Revisión final mixta, a partir de PLN y procedimiento manual (con criterios como frecuencia por adjetivos, sustantivos, adverbios, y palabras clave, por ejemplo, el peso de los cómos, los qué, los en dónde, etc.)



Signa_Lab ITESO

Especificaciones técnicas sobre equipo técnico y humano

Proyecto: Selección de Preguntas de Redes Sociales para Debate Presidencial INE 2024 (Formato A)

1. Especificaciones técnicas sobre equipos del laboratorio:

1.1 Equipos e infraestructura

#	Modelo	Almacenamiento	CPU	GPU	Memoria RAM
1	HP-ProDesk-600 G1 TWR	238.47 GB	Intel Core i5-4570 3.20 Ghz x 4	Intel Xeon E3-1200 v3/4th Integrated Graphics	7.7 GB
2	HP-ProDesk-600 G1 TWR	238.47 GB	Intel Core i5-4570 3.20 Ghz x 4	Intel Xeon E3-1200 v3/4th Integrated Graphics	7.7 GB
3	HP-ProDesk-600-G1_TWR	256.1 GB	Intel Core i5-4570 3.20 Ghz x 4	Intel Xeon E3-1200 v3/4th Integrated Graphics	7.7 GB
4	HP-ProDesk-600-G1_TWR	256.1 GB	Intel Core i5-4570 3.20 Ghz x 4	IIntel Xeon E3-1200 v3/4th Integrated Graphics	7.7 GB
5	HP-ProDesk-600 G1 TWR	256.1 GB	Intel Core i5-4570 3.20 Ghz x4	Intel Xeon E3-1200 v3/4th Integrated Graphics	7.7 GB
6	**AX370-Gaming K5-CF**	**1.79 TB**	**AMD Ryzen 7 2700 Eight-Core Processor 3.20 GHz**	**GeForce GTX 1050 TI 20 GB VRAM**	**32 GB**
7	iMac	1 TB	Quad-Core Intel Core i5 3.4GHz	NVIDIA GeForce GTX 775M 2 GB VRAM	16 GB
8	HP-Zbook Firefly G8	512 GB	11th Gen Intel(R) Core(TM) i5-1135G7 @ 2.40GHz 2.42 GHz	Intel Iris Xe Graphics 1.30 GHz	16 GB

***Consideraciones sobre ciberseguridad:**

Todos los equipos del ITESO, como los antes enlistados, cuentan con infraestructura, software y protocolos de seguridad informática gestionados y garantizados por la Oficina Servicios de Información (OSI) de la universidad. En observación a los protocolos de ciberseguridad establecidos, no se pueden dar mayores detalles sobre las medidas y sistemas instalados. Sin embargo, el personal de la OSI está dispuesto a dialogar y buscar la viabilidad de seguimiento a solicitudes específicas.

1.2 Herramientas y librerías a utilizar para el procesamiento de datos

- **Aplicaciones y entornos de trabajo (locales y de software libre)**
  - Jupyter Notebook v7.1 con Python 3.1 (local)
  - OpenRefine v3.7 (local)
  - LibreOffice v7.6.4 (local)
  - Gephi 0.92 (local)
  - TextAnalysis v0.2 (local)

- **Librerías y dependencias (locales y de software libre)**

○ pytorch ○ pandas ○ sentence_transformers ○ cohere ○ numpy ○ operator ○ nltk	○ matplotlib ○ scikit-learn ○ plotly ○ umap-learn ○ scatter-gl ○ http-server (local)

- **Modelos de software libre para el análisis semántico con aprendizaje automático:**
  - **Generación de *embeddings* (similitud entre oraciones)**

Modelo	**Dimensiones** (densidad de espacio vectorial)	**Licencia**	**Documentación**
paraphrase-distilroberta-base-v1 (local)	768	Apache 2.0 (software libre)	https://huggingface.co/sentence-transformers/paraphrase-distilroberta-base-v1
all-mpnet-base-v2	768	Apache 2.0 (software libre)	https://huggingface.co/sentence-transformers/all-mpnet-base-v2
all-MiniLM-L12-v2	512	Apache 2.0 (software libre)	https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2
distiluse-base-multilingual-cased-v2	384	Apache 2.0 (software libre)	https://huggingface.co/sentence-transformers/distiluse-base-multilingual-cased-v2

- **Clusterización:**
  - **Técnica:** k-means
  - **Librería:** sklearn.cluster.KMeans
  - **Documentación:** https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- **Reducción de dimensionalidades**
  - **Técnica:** UMAP
  - **Librería:** umap-learn
  - **Documentación:** https://pypi.org/project/umap-learn/

2. Composición y perfiles del equipo de Signa_Lab ITESO

2.1 Equipo Signa_Lab

Equipo de Coordinación • Dra. Rossana Reguillo, Coordinadora General de Signa_Lab ITESO • Mtro. Víctor Hugo Ábrego, Coordinador Ejecutivo de Signa_Lab • Mtra. Paloma López Portillo, Profesora, Analista y curadora de contenidos de Signa_Lab ITESO • Lic. Diego Arredondo, Profesor, Coordinador de Tecnologías e Interfaces de Signa_Lab ITESO • Lic. Eduardo G. De Quevedo, Psicólogo, Analista y Responsable Técnico de Signa_Lab ITESO **Consultores** • Asesor Doctorante en Estudios Científico Sociales (ITESO) • Asesor Estudiante de Ingeniería en Sistemas Computacionales (ITESO)	**Becarixs** • 5 estudiantes de la Licenciatura en Ciencias de la Comunicación (ITESO) • 2 estudiantes de Ingeniería y Ciencia de Datos (ITESO)

2.2 Resumen de capacidades técnicas del equipo

- Programación con lenguajes para la aplicación de técnicas de Ciencia de Datos (Python) y desarrollo de interfaces interactivas (JavaScript, HTML y CSS).
- Procesamiento de lenguaje natural (PNL) y análisis semántico basadas en aprendizaje automático (*embeddings de texto*, clusterización y reducción de dimensionalidades.
- Procesamiento, limpieza, transformación análisis de datos masivos con librerías especializadas, entornos de programación (Jupyter Notebook), software libre de ejecución local (OpenRefine, LibreOffice y TextAnalysis).
- Visualización exploratoria e interactiva de datos con librerías especializadas (scatter-gl, plotly y matplotlib) y software libre de ejecución local (Gephi).
- Análisis crítico, contextual y multicapa de grandes volúmenes de datos.