PROYECTOS MCDI 2025-1

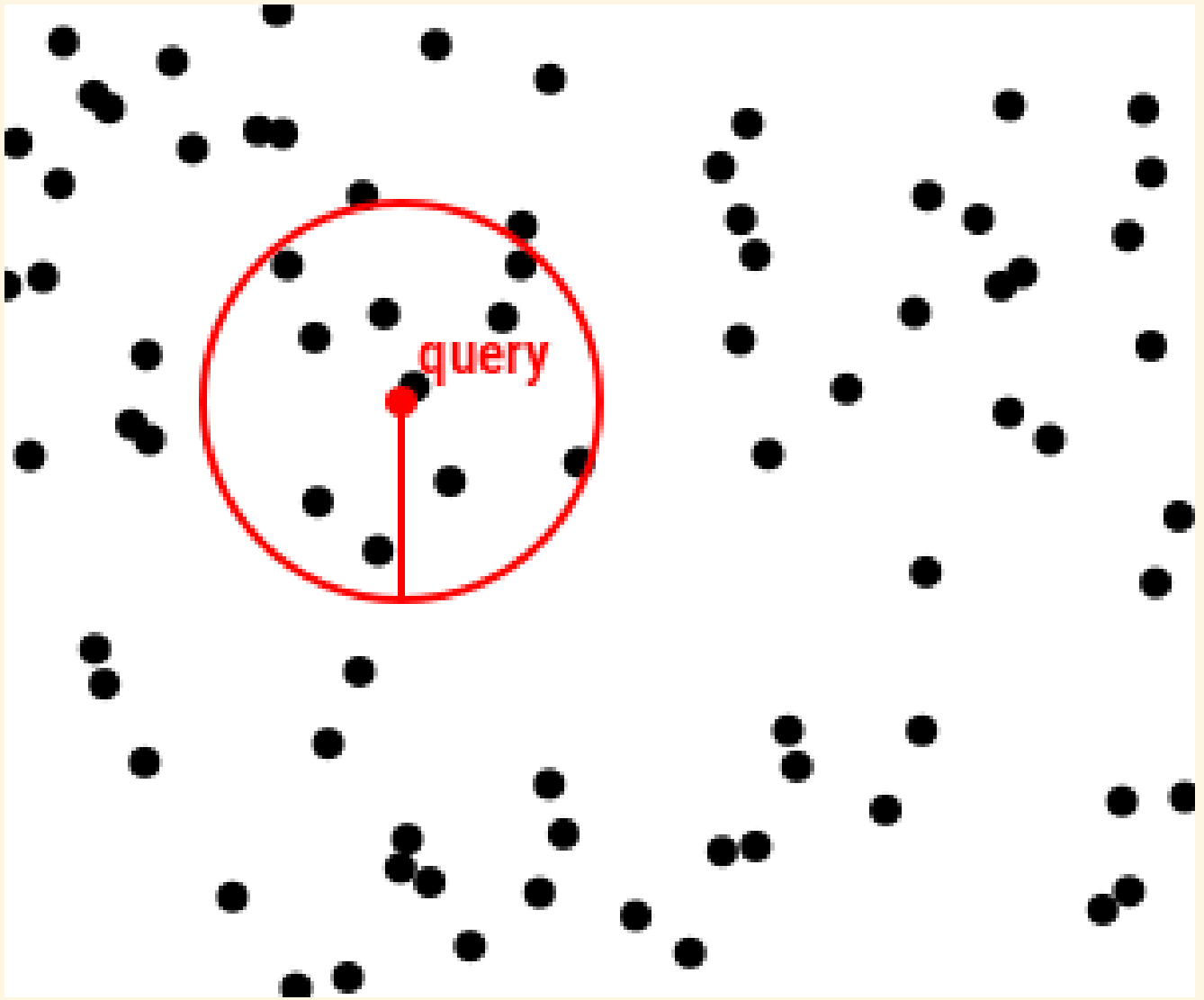
Eric S. Téllez

eric.tellez@infotec.edu.mx

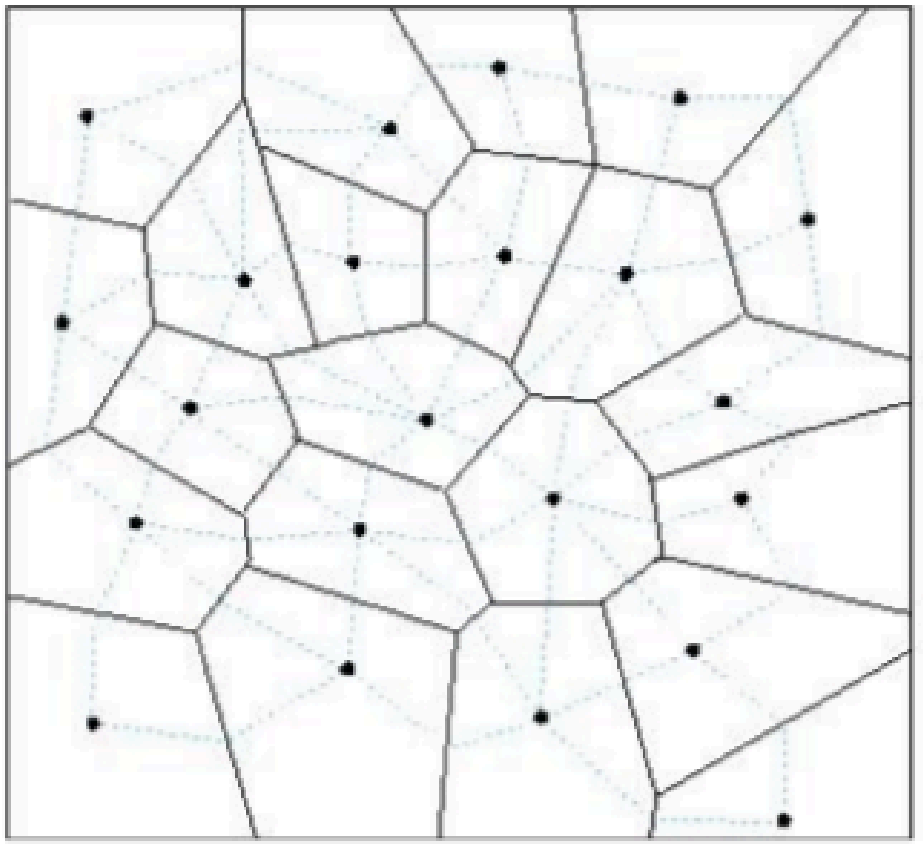
INFOTEC sede Aguascalientes, Ags.

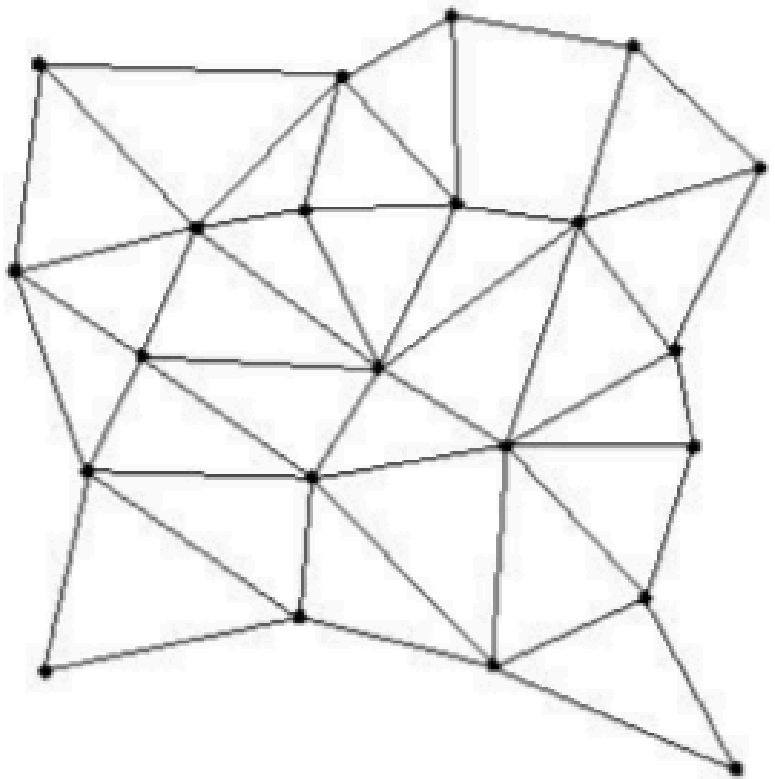
BÚSQUEDA POR SIMILITUD

PROBLEMA



Base de datos y consulta





Preprocesamiento

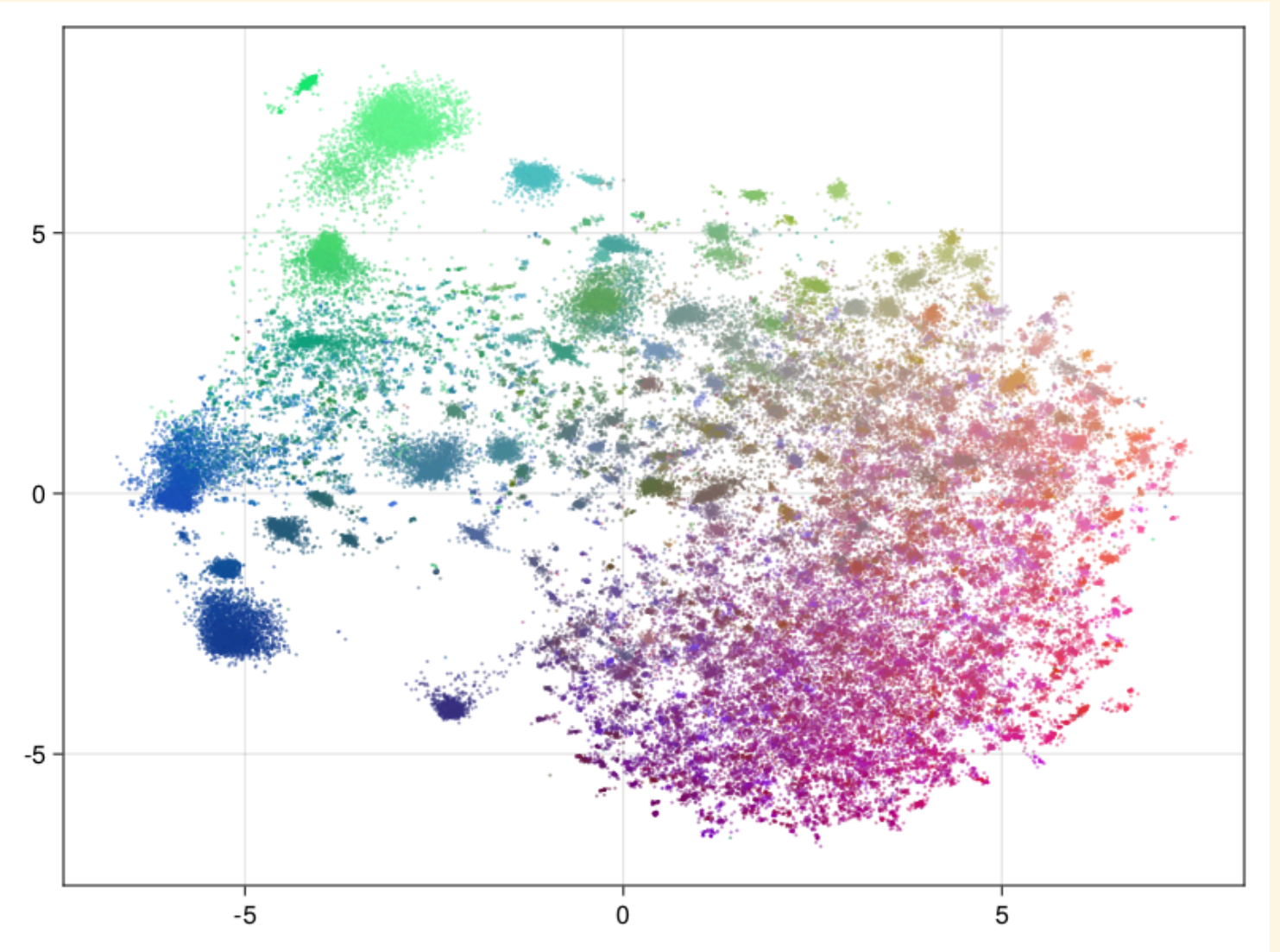
- Indexamiento y búsqueda para k vecinos cercanos.
- Determinación de los pares más cercanos y los k centros más alejados.
- Construcción de grafos de k vecinos.

APLICACIONES

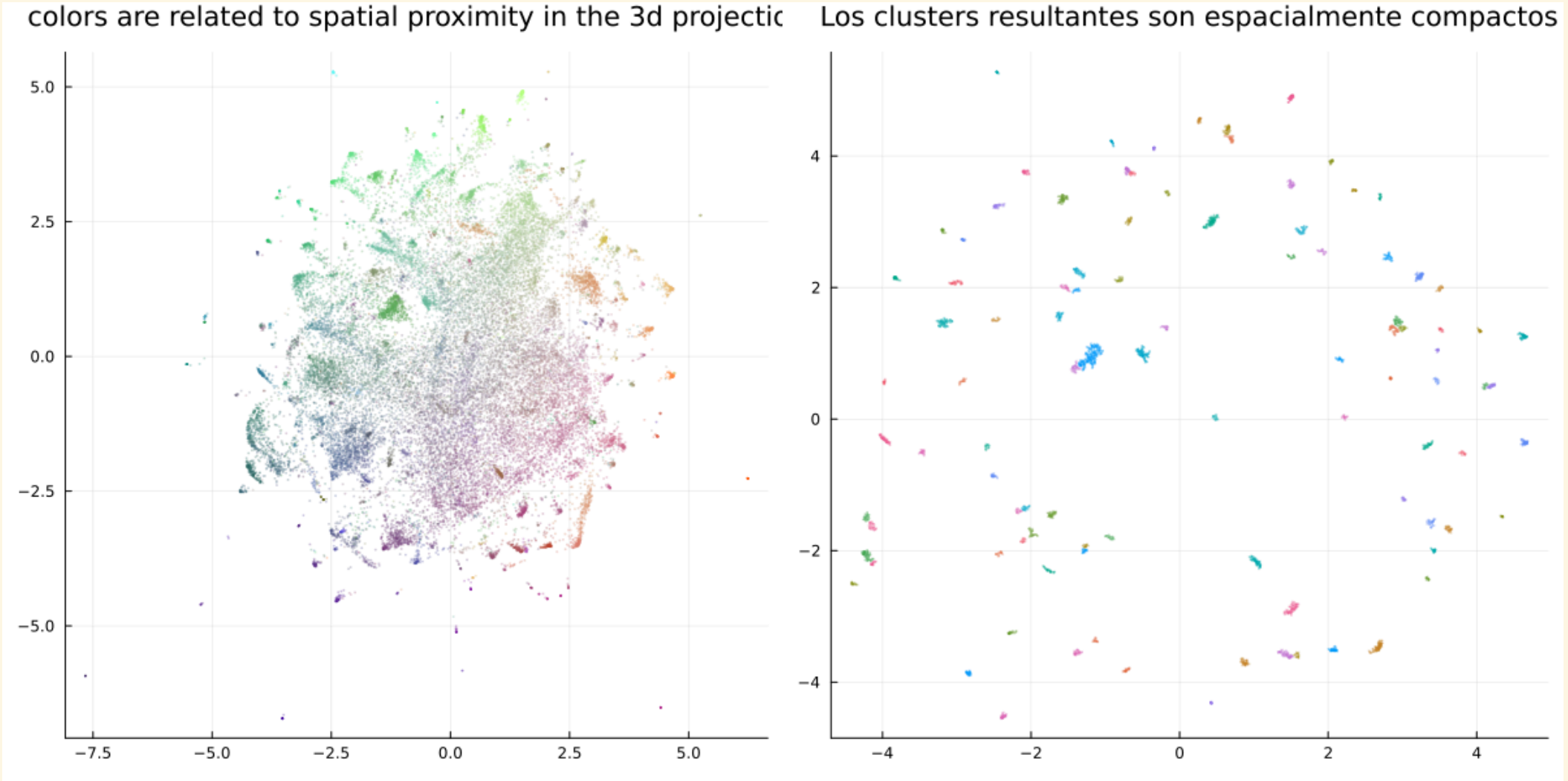
- **Recuperación densa (dense retrieval)**: Búsqueda de documentos a partir de consultas que no necesariamente se parecen léxicamente pero si semánticamente.
- **Búsqueda multimodal**: Búsqueda de video o imágenes por medio de texto usando descripciones del contenido.

- **Generación aumentada por recuperación**: Conocida como *Retrieval Augmented Generation* o (RAG), la idea es reducir alucionaciones de LLM generativos a partir de recuperación factual; también funciona para que los LLM tengan conocimiento fuera de su entrenamiento.
- **Acelerador de algoritmos** de agrupamiento y visualización.

PROYECCIONES A BAJA DIMENSIÓN



UMAP de primos



Ejemplo Clustering/tópicos – Unidad 6 Recuperación de Información

PROBLEMAS

- Sketches binarios sobre la distancia de Hamming.
- Cuantización basada en:
  - grafo kn.
  - grafo HSP.

BÚSQUEDA SIN INDEXAMIENTO

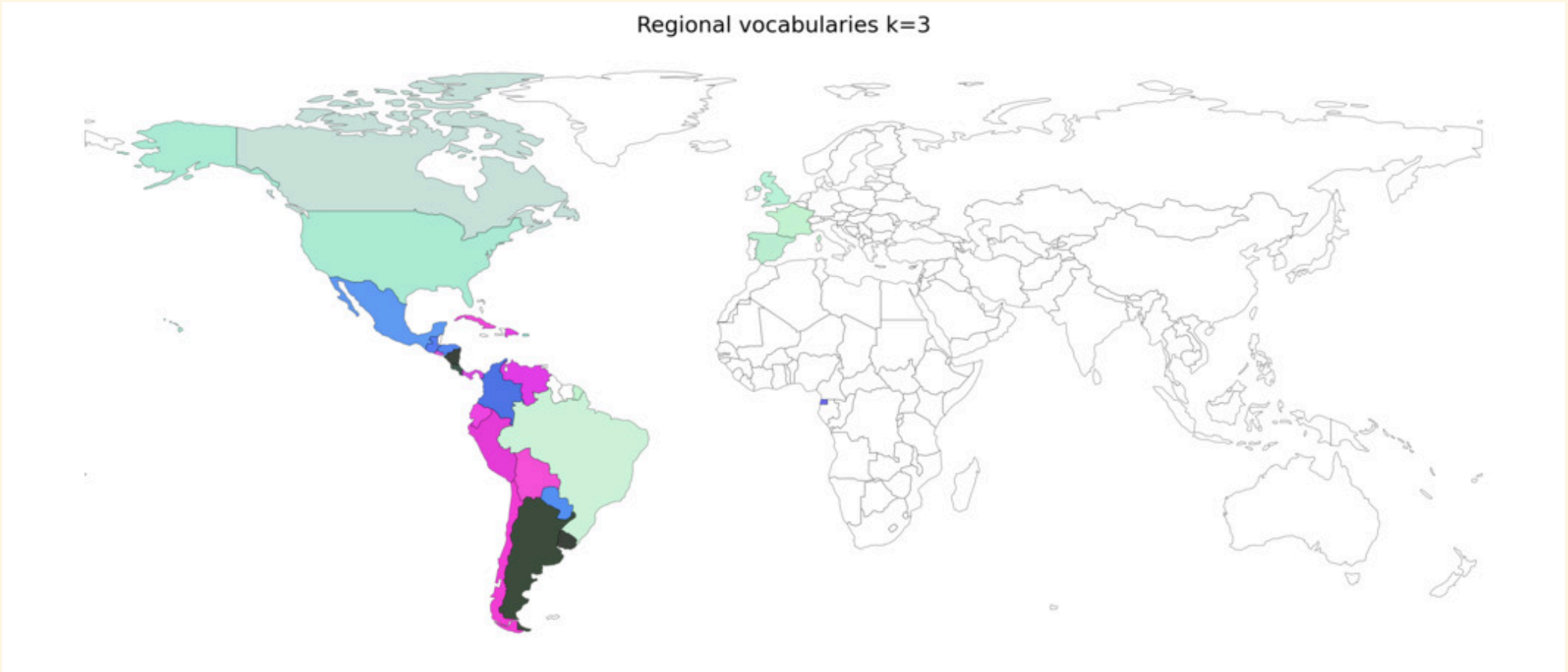
Entre las aplicaciones posibles, no siempre se necesita un índice (preprocesamiento)

- Archivado de datos.
- Análisis de datos históricos.
- Bases de datos que raramente son actualizadas.
- Agrupamiento.
- Visualización.

PROCESAMIENTO DE LENGUAJE NATURAL

RECURSOS REGIONALIZADOS PARA EL ESPAÑOL

ESPAÑOL



Similitud léxica entre regiones

MÉXICO

Similitud semática entre regiones de México

PROBLEMAS DE CLASIFICACIÓN

Entender el lenguaje y los mensajes escritos en redes sociales.

- **Minería de opinión** (análisis de sentimiento): determinar sí algo es positivo :), neutro :), o negativo :(
- **Análisis de tópicos**: ¿Qué temas hay en un corpus?
- **Carga emotiva de un mensaje**: *enojo*, *anticipación*, *disgusto*, *miedo*, *gozo*, *tristeza*, *sorpresa*, *confianza*.
- Identificación de **humor**, **odio**, o **esperanza**, …y un largo etcétera.

PERFILADO

- Predicción indicadores socio-demográficos de los usuarios.
- Identificación de autoría.
- Entender como se comportan usuarios.
- Medición de violencia en redes sociales.
- Identificación de posibles trastornos mentales.

COMPETENCIAS PLN

- IberLEF
- PAN
- FIRE
- SemEval

CLUSTERING Y TÓPICOS

- Clustering de documentos utilizando sentence BERT o BoW.
- Identificación de tópicos, e.g., LDA, BERT Topic.

BÚSQUEDA

- Búsqueda de texto completo con modelo léxico:
  - TFIDF
  - BM25
- Búsqueda densa:
  - Sentence BERT
  - ColBERT

GRACIAS

Dr. Eric S. Téllez
Investigador SECIHTI-INFOTEC
eric.tellez@infotec.edu.mx

Aguascalientes, Aguascalientes