Laboratorio de Analítica Computacional de Grandes Cúmulos de Información

CONACYT
Consejo Nacional de Ciencia y Tecnología

INFOTEC

Laboratorio Big Data (LaBD)

Aguascalientes, México

¿Qué es el laboratorio de Big Data?

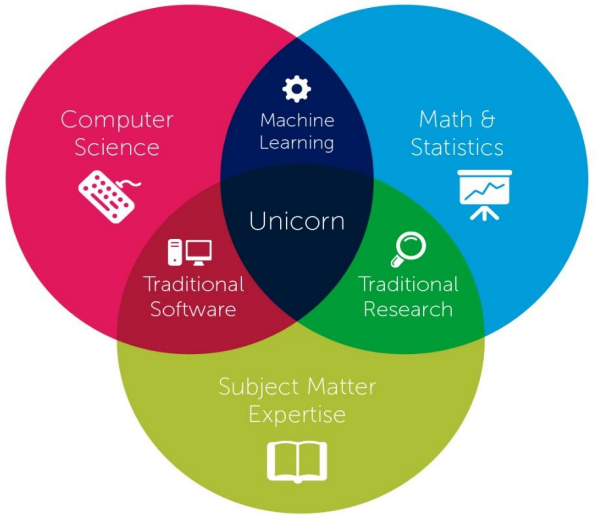
El Laboratorio de Big Data es un espacio de experimentación científico-computacional en ciencia de datos donde se busca

- Desarrollar investigación
- Desarrollar tecnologías
- Generar servicios

Laboratorio de Big Data

Involucra el descubrimiento de conocimiento a través del procesamiento de grandes colecciones de datos, provenientes de fuentes heterogéneas, mediante la aplicación de métodos de inteligencia artificial.

Data Science



Investigadores asociados

- Dr. Eric Sadit Téllez Avila
- Dr. Mario Graff Guerrero
- Dr. Dagoberto Armenta Medina
- Dr. Sabino Miranda Jiménez
- Dra. Daniela Moctezuma (CentroGeo)
- Dr. Luis Guillermo Ruiz (CentroGeo)
- Dra. Tania Anglaé Ramírez del Real (CentroGeo)
- Dr. Elio Villaseñor García (INEGI)
- Dr. José Luis Manzanares (Colegio de la Frontera Norte)

CONACYT
Consejo Nacional de Ciencia y Tecnología

INFOTEC

Estudiantes de doctorado

- M. C. Claudia Sánchez (UP)
- M. C. José Ortíz (UMSNH)
- M. C. Abel Coronado (INEGI)
- M. C. Sergio Nava (CIMAT)
- M. C. José Luis Jiménez (FC-UNAM)

LaBD - Participación - Posgrado

- Maestría de Sistemas Embebidos
  - MSE - Profesionalizante
- Maestría en Ciencia de Datos e Información
  - MCDI - Profesionalizante - En línea
- Maestría en Ciencias en Ciencia de Datos
  - MCCD - Investigación
- Doctorado en Ciencias en Ciencia de Datos
  - DCCD - investigación

LaBD - Áreas de Investigación

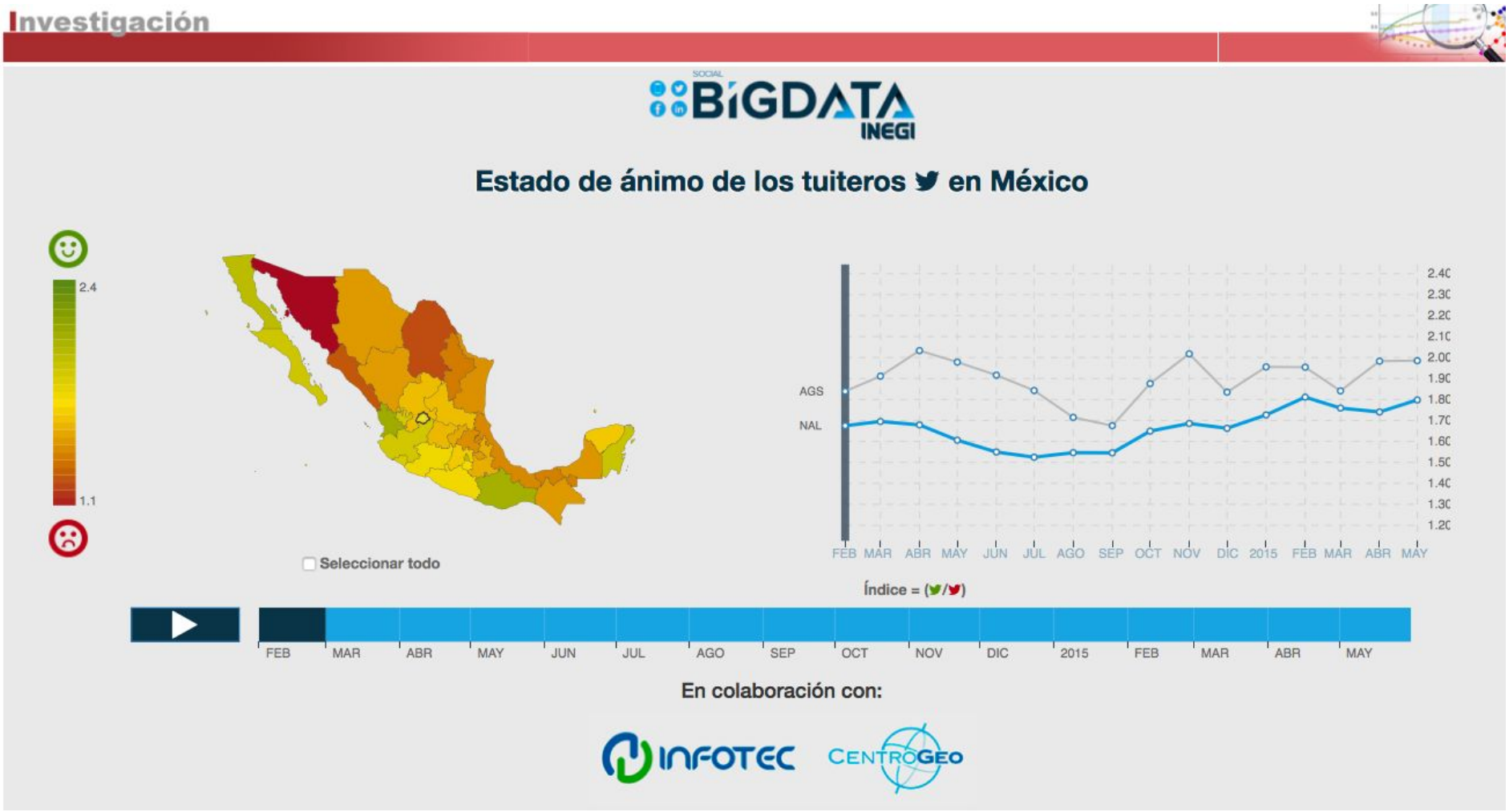
- Aprendizaje computacional
- Cómputo evolutivo
- Procesamiento de lenguaje natural
- Clasificación de textos
- Visión artificial
- Visualización de información y datos
- Análisis de datos biológicos

CONACYT
Consejo Nacional de Ciencia y Tecnología

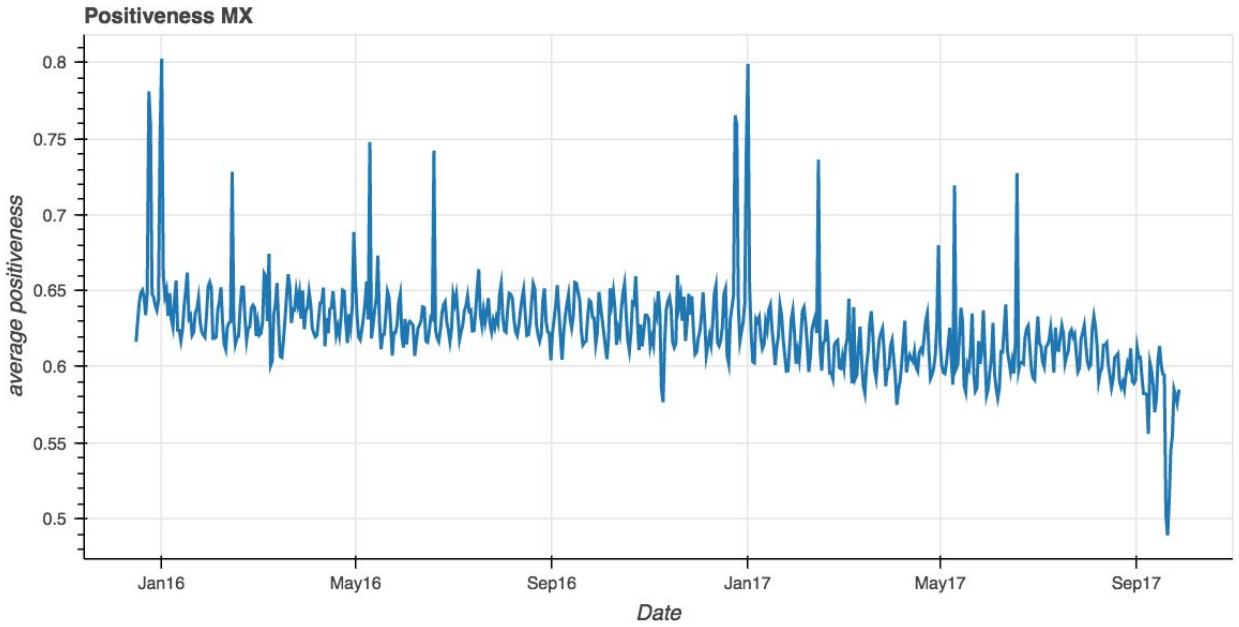
INFOTEC

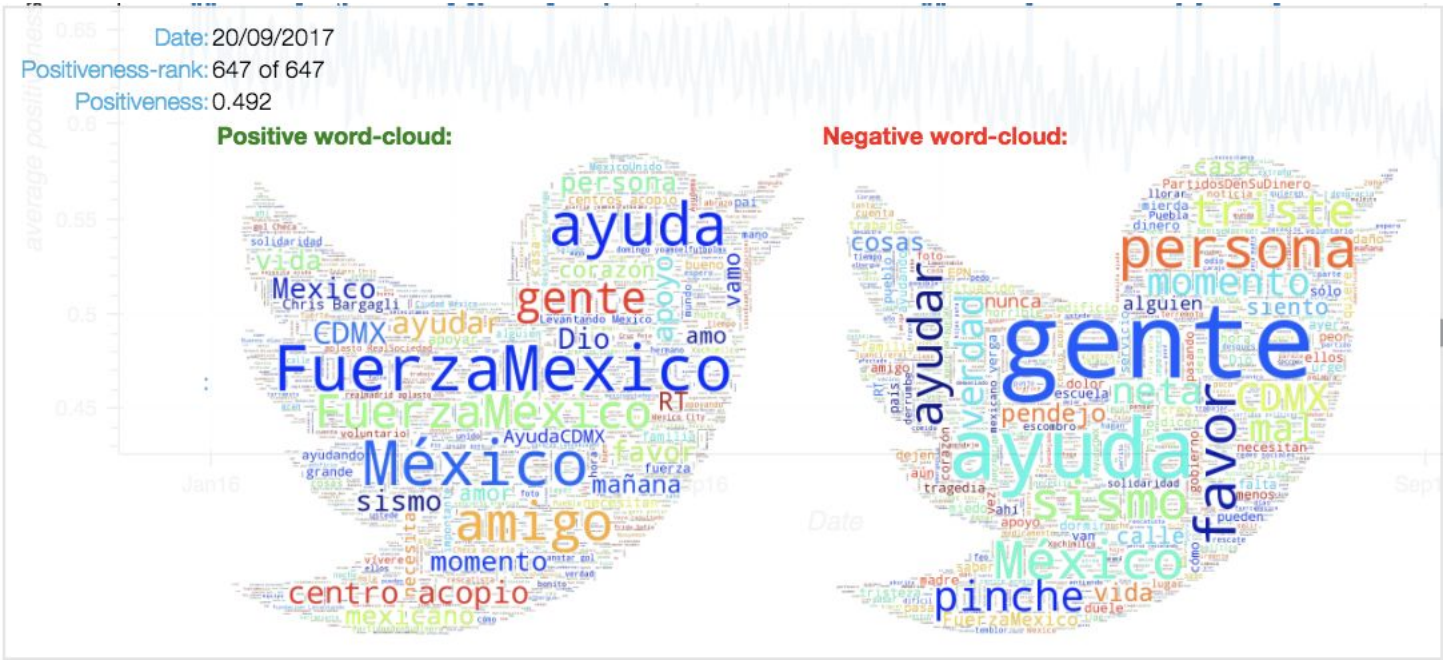
Desarrollos tecnológicos

Estado del ánimo tuitero en México



Positividad de México





CONACYT
Consejo Nacional de Ciencia y Tecnología

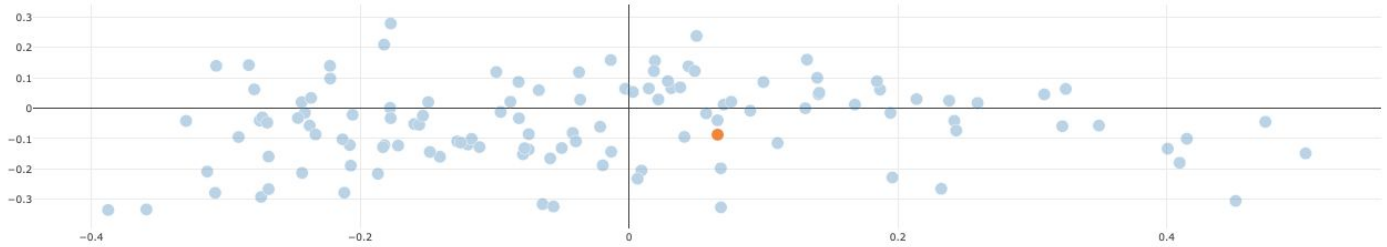
INFOTEC

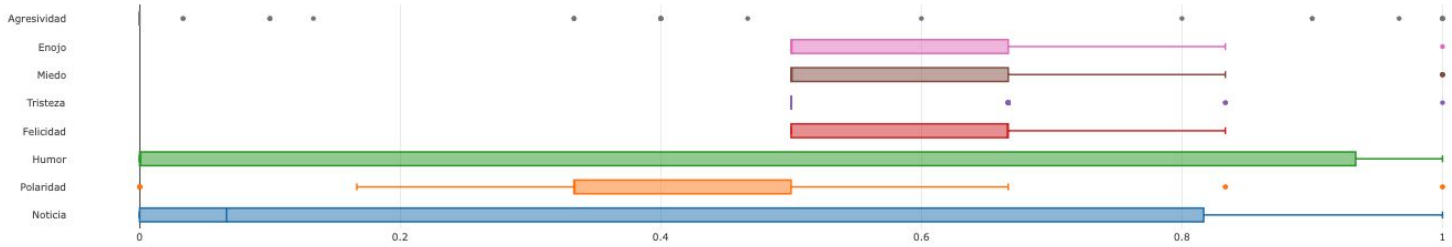
Modelado Automático de Usuarios

Basado en las Emociones Expresadas en sus Textos

Conacyt_MX

😉👍😀✨😱😎👏😊💙😘





Noticia
Polaridad
Humor
Felicidad
Tristeza
Miedo
Enojo
Agresividad

Opiniones

ubícam

Técnicas de IA utilizadas:

- B4MSA - detalles
- microTC - detalles
- EvoMSA - detalles

Áreas relacionadas:

- Minería de Opinión
- Análisis de Sentimientos
- Aprendizaje Comutacional
- Categorización Automática de Texto

Problemas relacionados:

- Ciberacoso
- Misoginia
- Noticias Falsas
- Monitoreo de Eventos
- entre otras

En 2013, la tesis de doctorado de Omar Felipe Giraldo obtuvo el Premio Arturo Fregoso Urbina de la Universidad Autónoma Chapingo a la mejor tesis, y mención honorífica del Premio Cátedra Jorge Alonso a la mejor tesis doctoral en ciencias sociales. https://t.co/kaQkEo7MFJ https://t.co/mRl6DstpVd 😊😎👍

El Conacyt felicita al @INMEGEN por un año más de contribuir al cuidado de la salud de los mexicanos, a través del desarrollo de proyectos de investigación científica con tecnología de vanguardia. #UnDíaComoHoy https://t.co/nt1Uw8XmaJ 😄🙏😊

El [illegible] cyt y la @amciencias entregaron esta semana el Premio de Investigación José Antonio Alzate, como reconocimiento los principales promotores de la cooperación entre México y Alemania. https://t.co/TZbNAqRVpJ https://t.co/FcFrtdFSIS 👏😊

@[illegible]zm Buenos días, ¿qué día escribiste el correo? 💕😊💜

http://ingeotec.mx/IA/usuarios/

CONACYT
Consejo Nacional de Ciencia y Tecnología

INFOTEC

Desarrollos de software

Software

- **MicroTC.** Librería para la creación de clasificadores de texto; es tanto independiente del lenguaje como del dominio.
  - https://github.com/INGEOTEC/microtc
- **B4MSA.** Librería multilenguaje para desarrollar analizadores de sentimiento.
  - https://github.com/INGEOTEC/b4msa
- **EvoMSA.** Sistema para generar analizadores de sentimiento basado en programación genética; permite la combinación de diversas fuentes de conocimiento.
  - https://github.com/INGEOTEC/EvoMSA
- **EvoDAG.** Sistema para abordar problemas de aprendizaje supervisado basados en programación genética.
  - https://github.com/mgraffg/EvoDAG
- **SimilaritySearch.jl.** Conjunto de librerías para abordar problemas de búsqueda por similitud y aprendizaje computacional tanto en espacios métricos como en texto.
  - https://github.com/sadit/SimilaritySearch.jl
  - ***TextSearch.jl*** *https://github.com/sadit/TextSearch.jl*
  - ***KernelMethods.jl*** *https://github.com/sadit/KernelMethods.jl*
  - ***KCenters.jl*** *https://github.com/sadit/KCenters.jl*

Plataforma para procesar grandes cúmulos de datos

CONACYT
Consejo Nacional de Ciencia y Tecnología

INFOTEC

DASK

Dask natively scales Python

- **Scikit-learn** — Machine learning in Python
- **Scikit-image** — A collection of algorithms for image processing in Python
- **TPOT** — A Python Automated Machine Learning tool that optimizes machine learning pipelines using genetic programming
- **XGBoost** — Gradient boosted trees for machine learning. XGboost can use Dask to bootstrap itself for distributed training
- **XArray** — Brings the labeled data power of pandas to the physical sciences, by providing N-dimensional variants of the core pandas data structures
- **Iris** — A Python library for analysing and visualising Earth science data
- **Pangeo** — A community effort for big data geoscience in the cloud
- **RAPIDS** — GPU Accelerated libraries for data science
- **Datashader** — Visualization packages for large data
- **Intake** — A lightweight package for finding, investigating, loading and disseminating data
- **Prefect** — A workflow management system, designed for modern infrastructure
- **MDAnalysis** — A Python toolkit to analyze molecular dynamics trajectories generated by a wide range of popular simulation packages
- **Stumpy** — A Python library that can be used for a variety of time series data mining tasks
- **Featuretools** — A Python framework for automated feature engineering
- **Cesium-ML** — Open-Source machine learning for time series analysis
- **SkyPortal** — An astronomical data platform
- **SatPy** — Library for reading and manipulating meteorological remote sensing data and writing it to various image and data file formats
- **Streamz** — A package to help build pipelines to manage continuous streams of data
- **Scikit-allel** — Provides utilities for exploratory analysis of large scale genetic variation data
- **tsfresh** — Automatic extraction of relevant features from time series

APACHE Spark™

APACHE Spark

hadoop

cassandra

MESOS

APACHE HBASE

kubernetes

APACHE hadoop

CONACYT
Consejo Nacional de Ciencia y Tecnología

INFOTEC

SchedMD

Simple Linux utility for resource management

https://slurm.schedmd.com/quickstart.html

`sacct` `salloc` `sattach` **`sbatch`** `sbcast` `scancel` **`scontrol`** **`sinfo`** `squeue` **`srun`** `strigger` `smaps` `sview`

- Pros:
  - Gratuito y de código abierto
  - No requiere ajustarse a un flujo de programa particular o a un entorno de programación específico
  - Puede utilizar cualquier lenguaje de programación
  - No es necesario modificar programas si los datos de entrada pueden particionarse
  - Soporte nativo para MPI
  - Puede correr Hadoop, Spark, Dask, etc.
- Contras:
  - Funciona sobre linux (no necesariamente una contra)
  - Requiere administración
  - Uso de NFS o GlusterFS para el manejo de archivos distribuidos

CONACYT
Consejo Nacional de Ciencia y Tecnología

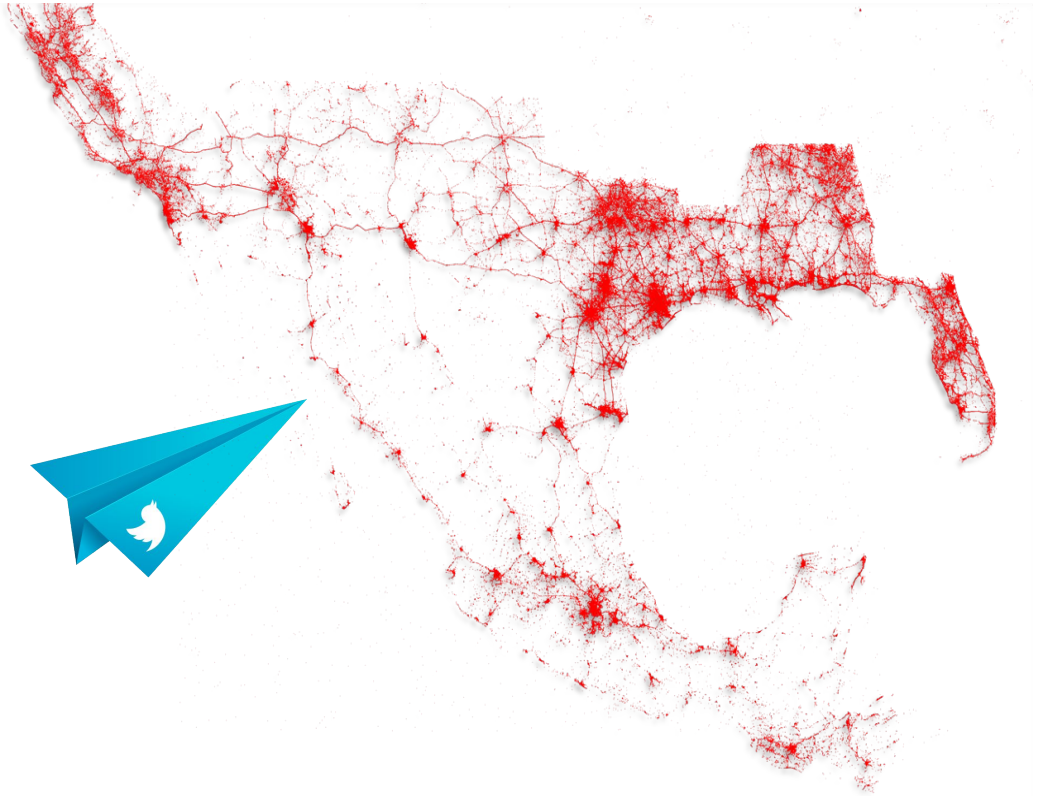
INFOTEC

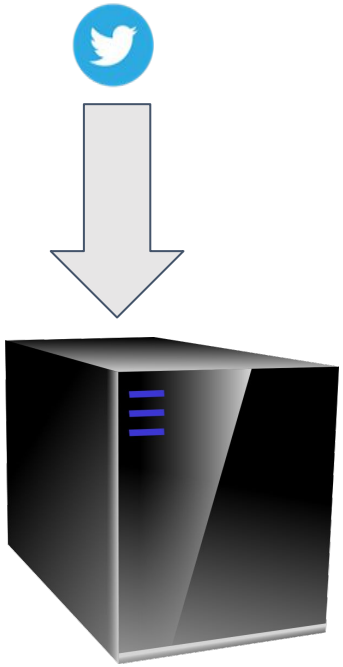
Infraestructura

- Penguin Computing (investigación)
  - Cluster SLURM con 7 nodos con 16 cores@2.6GHz con hyperthreading; en total 224 threads, 923 GB RAM conectadas mediante ethernet de 10-gigabit.
  - Dos tarjetas Xeon Phi con 420 threads
- Dell (Prácticas alumnos)
  - Cluster SLURM con 1 nodos Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz; 28 total threads, 256GB RAM conectadas mediante ethernet 1-gigabit
  - Cinco GPUs Titan Xp cada una con 3840 NVIDIA CUDA® Cores y 12 GB de memoria

Recolectamos el stream público de Twitter desde diciembre 2015 a la fecha varios millones de tweets en español, inglés, árabe, ruso, portugués, francés, entre otros.

Esto nos representa cerca de 30TB comprimidos (~200TB sin compresión)





¡Gracias por su atención!

Laboratorio Big Data (LaBD)

https://www.infotec.mx/es_mx/infotec/Laboratorio_de_Analitica_Big_Data

Aguascalientes, México