Text Mining II - Máster Big Data Analytics (2015 - 2016)
Exploración del dataset HispaTweets

Dr. Paolo Rosso
Casi-casi-Dr. Francisco Rangel ;-)

autoritas

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Objetivos de la exploración de datos

- Entender propiedades de los datos.
- Encontrar patrones.
- Depurar.
- Comunicar resultados.

Descarga del dataset HispaTweets (Fabra et al., 2016)

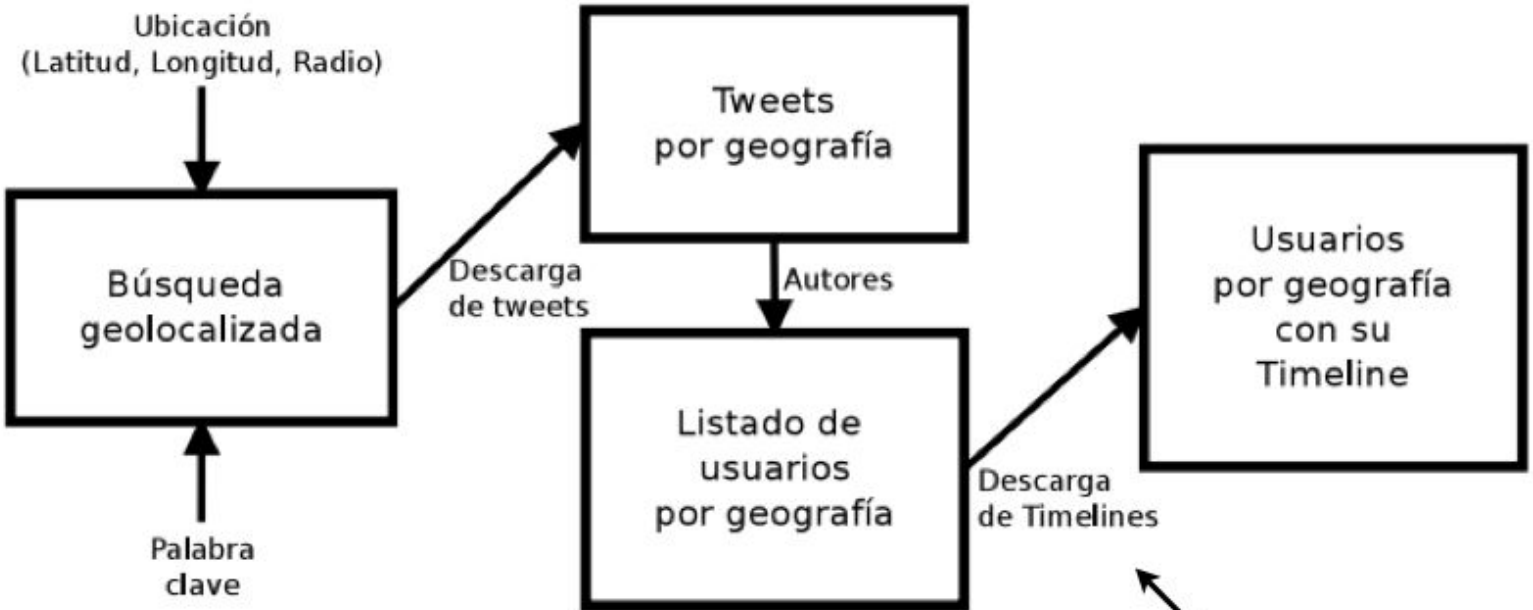
https://s3.amazonaws.com/cosmos.datasets/hispatweets.zip

Características del dataset

- Obtenido de Twitter.
- Colección de decenas/cientos de miles de autores.
- Cientos de tuits por autor.
- Gran variedad de temas.
- ¿Gran tamaño? 700Mb comprimido, 13Gb descomprimido.
- Dificultad para etiquetar la información:
  - Variedad del lenguaje por posición geográfica.
  - Sexo por nombre (mujeres, hombres y ¿perfiles institucionales?)
  - Personas reales vs. robots (chatbots)

¿Big Data?

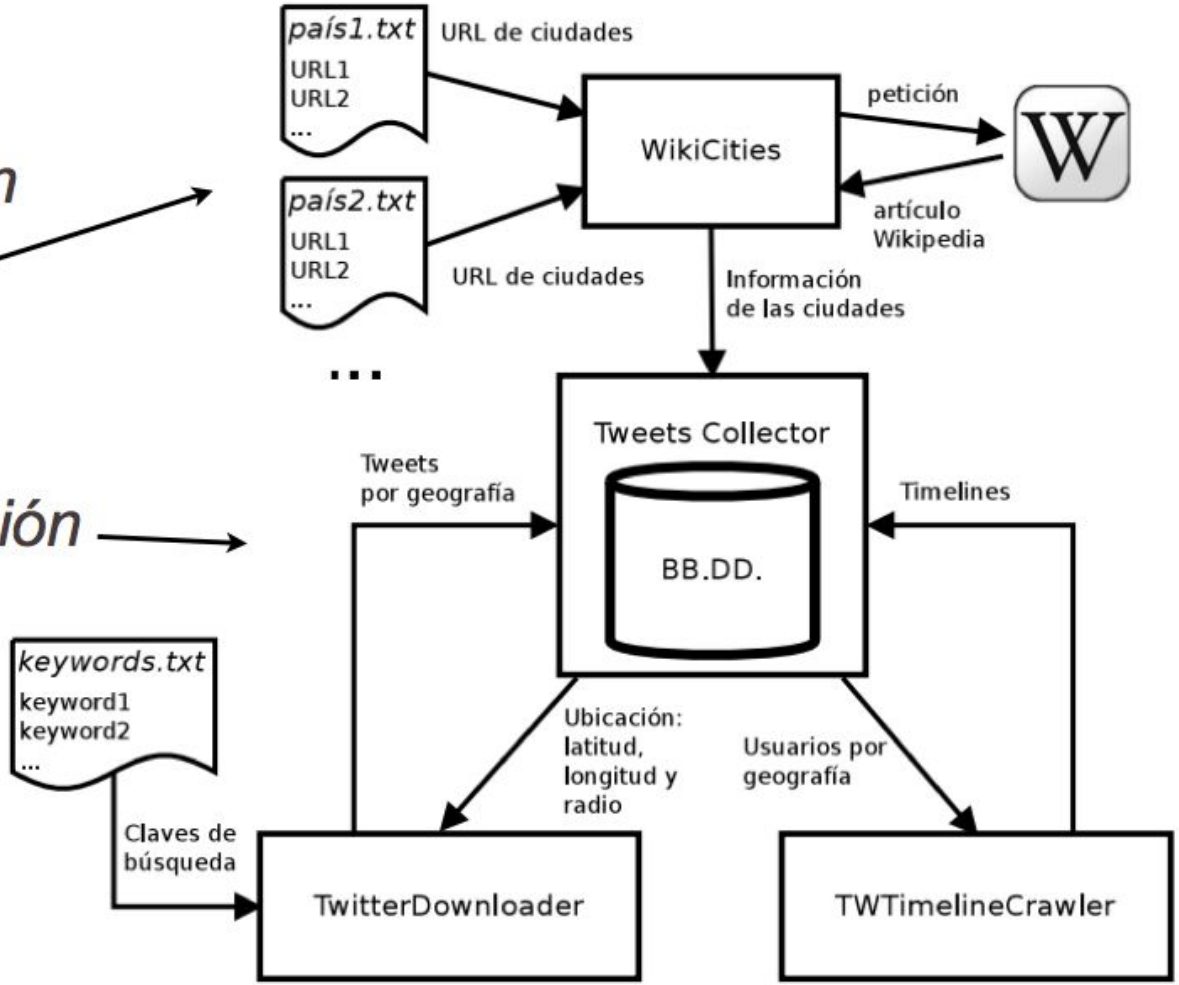
Proceso de construcción



hasta los últimos 1000 tweets de sus respectivas *timelines*.

Proceso de obtención de coordenadas geográficas

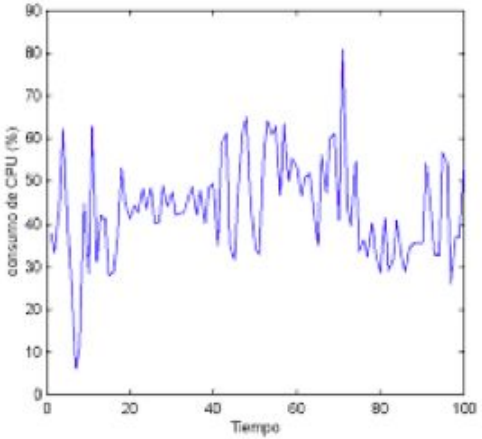
Proceso de construcción del corpus



Criterio de selección

Depuración, refinamiento y filtrado del corpus:

Geográfico



Temporal

Usuarios

Etiquetado del sexo

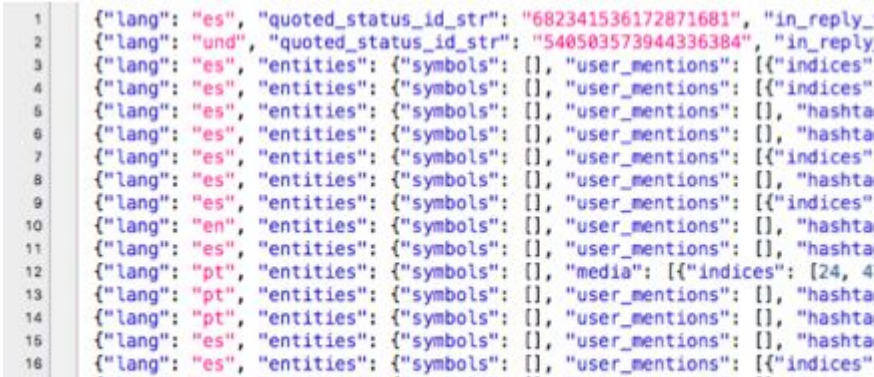
Basado en diccionario de nombres propios:

Mujeres	24.429
Hombres	25.949
Total	*50.378*

Formato de los ficheros

Una vez descomprimido el dataset, la estructura es la siguiente:

- Un par de ficheros de verdad: training.txt y test.txt. El formato es:
  - id:::variedad:::sexo
- Una carpeta por variedad, un fichero .json por autor:
  - Cada línea del fichero un tuit en formato json.
  - A la derecha un ejemplo ->
  - Se recomienda el uso de jsonlint.com



¿Qué nos interesa explorar?

- Número de autores por clase (sexo y variedad del lenguaje)
- Número de tuits por autor.
- Número de tuits por clase.
- Número de palabras por documento / autor / clase.
- Distribución de palabras/documentos/autores por documento/autor/clase…
- Longitud media de tuits, palabras, documentos...por clase.
- Distribución temporal de los tuits, tuit más antiguo, más nuevo, media, desviación…
- Palabras extrañas, frecuentes, comunes…
- … cualquier información que nos describa el dataset y aporte conocimiento nuevo y valuable.