Memoria del proyecto: Análisis de corpus lingüístico.

Para el análisis exploratorio de datos de los ficheros de frases en inglés con su traducción se pasó por varias etapas. Cada etapa fue algo complicada por el inmenso volumen de datos.

Ficheros .txt o .tsv:

1. Al generar los script para obtener el formato adecuado (source -- target) habían gran cantidad de TU con formato incorrecto.
2. Filas con más de 1 target
3. Varios tab en la misma TU
4. Frases con URL, emoticons, etc.
5. Frases en el target que empezaban con " y entonces a la hora de cargar los datos en un dataframe no reconocía esas TU.

Base de datos

Una vez obtenidos todos los ficheros limpios se concatenaron en un dataframe, obteniendo millones de observaciones. Luego correspondía categorizar todas las frases y obtener todos los valores de las diferentes columnas

Al procesar dichos datos el proceso se hacía muy lento, al punto que no se obtuvo ningún resultado final. A pesar de contar con una buena computadora con buenas características el proceso no se resultó.

Se decide crear una base de datos y guardar toda la información para poder generar los valores de las columnas.

Este proceso duró una 1 semana, utilizando 3 máquinas al mismo tiempo y en cada una de ella varias instancias de "visual code" corriendo el código que actualiza en la base de datos.

No se pudo terminar de actualizar todos los valores de las diferentes columnas, quedando valores nulos.

Pero se logró etiquetar 12 millones filas, quedando solo 4 millones.