Ejercicios de curación y limpieza de datos

Ejercicio 4:

a. Carga el fichero SAFI_ejercicio4_openrefine.csv para realizar el ejercicio 4.
b. Usa clustering para corregir los errores en el campo “province”. Prueba con diferentes técnicas y observa las diferencias. ¿Cuáles son más útiles?
c. Repite la misma operación que en el punto b para el campo “district”. ¿Se pueden corregir todos los errores? ¿Cuáles no? ¿Por qué crees que sucede esto?
d. Prueba a usar clustering para corregir los errores del campo “other_buildings”. ¿Es útil alguno de los métodos?
e. Prueba a utilizar clustering, usando facets de tipo texto, para corregir errores en el campo “interview_date”. ¿Es factible? ¿Qué es lo que recomiendan los diferentes métodos?
f. Utiliza ahora clustering para tratar de corregir errores en el campo “months_lack_food”. ¿Es útil en este caso?
g. Lee con calma el siguiente enlace, y trata de entender las diferencias de funcionamiento entre los diferentes algoritmos de clustering que ofrece OpenRefine:https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth

M1968 - Ciclo de vida de los datos
Máster en Data Science – Universidad de Cantabria Curso 2018-2019
Diego García Saiz