Tipología y ciclo de vida de los datos I

El objetivo de esta actividad será la creación de un dataset a partir de los datos contenidos en la web. Tenéis que indicar las siguientes características del dataset general:

1. Título del dataset. Poned un título que sea descriptivo.
   Directorio de empresas
2. Subtítulo del dataset. Agregad una descripción ágil de vuestro conjunto de datos por vuestro subtítulo.
   Categorización de empresas gallegas por localidad y actividad.
3. Imagen. Agregad una imagen que identifique vuestro dataset visualmente

4. Contexto. ¿Cuál es la materia del conjunto de datos?

Existe una página web https://www.paxinasgalegas.es/ donde se publicitan las empresas situadas en Galicia. Esta página cataloga la información por localidad, por actividad y sub-actividad.

Para crear el conjunto de datos se ha cogido toda la información de las empresas de la localidad de Santiago de Compostela.

Se ha generado el código para recoger todas las empresas de Santiago de Compostela catalogadas por actividad y subactividad. El dataset creado es una lista de todas las empresas de Santiago de Compolesta dedicadas a la administración.

5. Contenido. ¿Qué campos incluye? ¿Cuál es el periodo de tiempo de los datos y cómo se ha recogido?

El conjunto de datos generado incluye los siguientes campos:

- Nombre: nombre de la empresa

- Telefono: teléfono de contacto de la empresa
- Correo: dirección de correo de la empresa
- Pagina Web: dirección de la página web
- Dirección: dirección física de la situación de la empresa
- Latitud: coordenada gps de la latitud
- Longitud: coordenada gps de la longitud
- Descripción: pequeña descripción de la actividad de la empresa.
- Actividad: clasifica la empresa en una actividad (ej: servicios, telecomunicaciones, administración, alimentación…)
- Subactividad: indica la subactividad a la que pertenece (ej: administración bancaria, alimentación animal...)

En cuanto al período de datos recogidos no depende del tiempo. Salvo que se den de baja empresas o se den de alta nuevas empresas, que es algo que ocurre con muy poca frecuencia, los datos pueden ser los mismos durante años.

Los datos se han recogido de la página web https://www.paxinasgalegas.es/empresas-santiago-de-compostela-79ay.html. En esta página se muestra la lista de las diferentes actividades. Se trata de una lista de enlaces que al seleccionarla te re direcciona a una nueva página que muestra otra lista con las sub-actividades. Cada una de estas sub-actividades es un enlace a una página web con las empresas que componen esas características.

Por lo tanto lo que se hizo para capturar la información, es recoger los enlaces web de cada actividad, y a su vez los enlaces web de las sub-actividades. Una vez cargadas las páginas web correspondientes, en función de actividad y sub-actividad, obtenemos la lista de empresas, donde extraemos los valores como el nombre, dirección, teléfono, email, página web, latitud, longitud y descripción.

6. Agradecimientos. ¿Quién es propietario del conjunto de datos? Incluid citas de investigación o análisis anteriores.

Se extrae la información de la página https://www.paxinasgalegas.es/ el propietario es Páginas Telefónicas S.L

No he encontrado ni investigación ni análisis previo de esta información. Las empresas se ponen en contacto con Páginas Telefónicas para publicitarse y destacarse frente a las demás. Ellos generan sus propios datos, pero no he encontrado hasta el momento ninguna investigación.

7. Inspiración. ¿Por qué es interesante este conjunto de datos? ¿Qué preguntas le gustaría responder la comunidad?

Porque puede responder a muchas preguntas a las diferentes comunidades.

Aquellas comunidades que están interesadas en ofrecer sus servicios, por ejemplo empresas de diseño web pueden ver si las empresas disponen de páginas web y si nos así ofrecer sus servicios. También puede interesar a personas que quieran crear una nueva empresa y ver cuál es la competencia que existe en esta localidad. O para realizar estudios y ver cuál es la mayor actividad en esa localidad, si servicios, administración, etc. O incluso realizar estudios en que zona se alojan la mayoría de empresas,etc.

8. Licencia. Seleccionad una de estas licencias y decid porqué la habéis seleccionado:

Released Under CC0: Licencia de Dominio Publico

He seleccionado esta licencia porque así se puede copiar, modificar, distribuir los datos y hacer comunicación pública y así permitir fines comerciales. Como se indica en el apartado anterior, este conjunto de datos tiene mayor interés para obtener beneficios comerciales.

9. Código: Hay que adjuntar el código con el que habéis generado el dataset, preferiblemente con R o Python, que os ha ayudado a generar el dataset

10. Dataset: Dataset en formato CSV