Web Scraping. Audiencias de programas y televisiones en España

Victor Manuel Vásquez Rivas y Francisco Javier Moreno Hernández
Abril del 2019

1.- Contexto. Explicar en qué contexto se ha recolectado la información. Explique por qué el sitio web elegido proporciona dicha información.

La recolección de información se ha llevado a cabo bajo el contexto de una actividad práctica para el ejercicio en la utilización de técnicas de web scraping, en donde se ha escogido el sitio web https://ecoteuve.eleconomista.es para la obtención de información de canales televisivos, en donde este sitio nos proporciona datos de canales, programas y audiencias, siendo un sitio ideal para el análisis y aplicación de conocimiento práctico en web scraping.

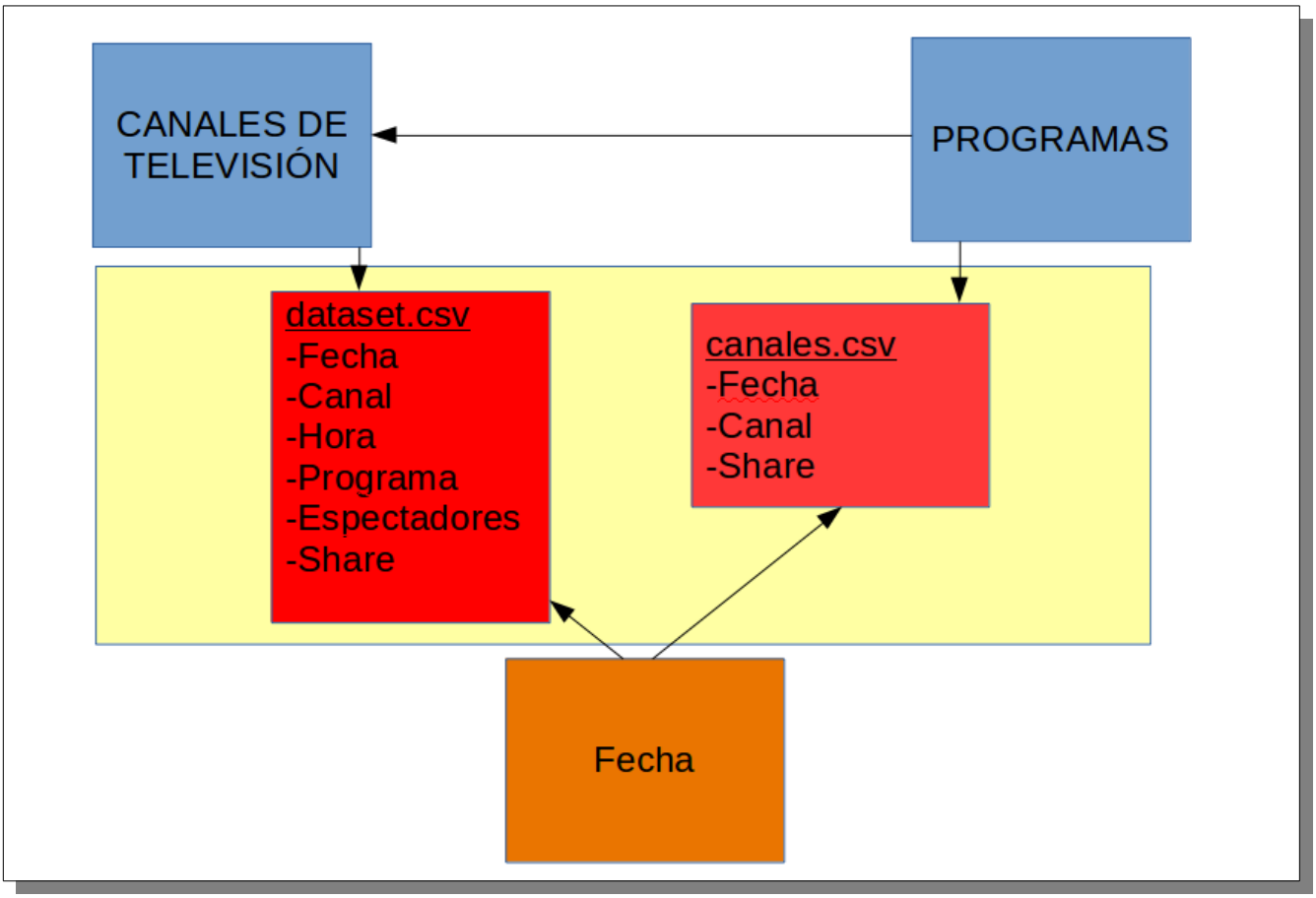
2.-Definir un título para el dataset. Elegir un título que sea descriptivo.

Audiencias de programas y televisiones en España.

3.- Descripción del dataset. Desarrollar una descripción breve del conjunto de datos que se ha extraído (es necesario que esta descripción tenga sentido con el título elegido).

Los datos corresponden a datos televisivos de canales de España, en donde se presenta información de sus programadas, espectadores y porcentaje de audiencias por programa y canal (share).
Los datos generados se dividen en dos dataset. En el primero llamado dataset.csv, en éste se busca el listado de canales existentes en https://ecoteuve.eleconomista.es/canales/, luego por cada canal se va recorriendo por día, búscando de los datos de audiencias de cada programa por canal de televisión obtenido anteriormente, por ejemplo https://ecoteuve.eleconomista.es/cadena/La2/audiencias-programas/2019-03-22 . El segundo conjunto de datos llamado canales.csv, corresponde a los datos de audiencia diaria de aquellos canales obtenidos que presentan esta información, en donde el dato disponible es el porcentaje/ share.

4.-Representación gráfica. Presentar una imagen o esquema que identifique el dataset visualmente



5.- Contenido. Explicar los campos que incluye el dataset, el periodo de tiempo de los datos y cómo se ha recogido.

dataset.csv. Para la obtención del dataset se ha recorrido un listado de canales guardándolos en un array, luego por cada uno de ellos se recorre 3 días a partir del día anterior al día de hoy, por ejemplo si hoy es 25-03-2019, se buscará información para TELECINCO los 22, 23 y 24 de marzo. Los datos obtenidos son los siguientes:

- Fecha: corresponde a la fecha de la información obtenida de la audiencia.
- Canal: corresponde al canal de televisión en que se obtiene información.
- Hora: horario en que el programa se televisa.
- Programa: nombre del programa de televisión.
- Espectadores: cantidad de espectadores que obtuvo el programa el día especificado.
- Share(%): porcentaje de audiencia del programa respecto a otros programas en la misma hora.

canales.csv. Para la obtención del dataset se ha recorrido el listado de canales obtenidos en el dataset anterior, para luego ir a buscar la información de audiencia general del canal. Los datos son los siguientes:

- Fecha: corresponde a la fecha de la información obtenida de la audiencia.
- Canal: corresponde al canal de televisión en que se obtiene información.
- Share(%): porcentaje de audiencia del programa respecto a otros programas en la misma hora.

En este caso no se obtiene el número de espectadores porque no se dispone de esa información en esta web.

6.-Agradecimientos. Presentar al propietario del conjunto de datos. Es necesario incluir citas de investigación o análisis anteriores (si los hay).

Agradecemos al portal https://ecoteuve.eleconomista.es por contener los datos que hemos querido extraer. También presentamos a la empresa Kantar Media (https://www.kantarmedia.com/es) como empresa propietaria de los datos. En el caso de uso comercial sería necesario contactar con la empresa para asegurar la licencia de los datos.

7.- Inspiración. Explique por qué es interesante este conjunto de datos y qué preguntas se pretenden responder.

El conjunto de datos pretende entregar la información necesaria para poder mostrar visualmente aquellos cambios que a través del tiempo han tenido los programas y canales de televisión española, en donde se puede utilizar para evidenciar horarios y mejores programaciones.
Por lo que con este conjunto de datos se puede realizar un estudio de la evoluciones de los programas televisivos, en donde se podrían buscar ciertas tendencias, predecir que programas elevarán su número de espectadores, entregar alternativas de horarios para publicidad en los que va en aumento de espectadores y a la vez predecir aquellos que van en baja, entre otras predicciones que se pueden generar en el área de minería de datos.

8-. Licencia. Seleccione una de estas licencias para su dataset y explique el motivo de su selección:

Hemos seleccionado de la lista la “Unknown License”. Hemos investigado si la empresa que ofrece los datos informaba de la licencia en la web, pero no hemos encontrado nada al respecto. En el caso de uso comercial sería necesario contactar con la empresa para asegurar la licencia escogida.

Contribuciones	Firma
Investigación Previa	V.M.V.R,F.J.M.H
Redacción de las Respuestas	V.M.V.R,F.J.M.H
Desarrollo de Código	V.M.V.R,F.J.M.H