Asignatura	M2.851 Tipología y ciclo de vida de los datos
Semestre	2017-18-s2
Práctica	1
Fecha	15-04-2018
Autor	Alberto Gómez

Dataset: Las obras de J.S. Bach

Recopilación de datos sobre las obras de Johan Sebastian Bach y su familia a partir de la información de **bach-digital.de.**

Contexto

En este *dataset* se recogen datos de las obras musicales de Johan Sebastian Bach y algunos miembros de su familia.

La web **bach-digital.de** recopila información sobre las obras de Johan Sebastian Bach y algunos miembros de su familia, junto con información sobre las fuentes bibliográficas donde se pueden encontrar estas obras.

Para recopilar la información de este *dataset* se ha hecho *web scraping* sobre la parte de la web relacionada con las obras, sin tener en cuenta las fuentes bibliográficas.

Se ha realizado la búsqueda sobre la información en inglés (la web presenta información en varios idiomas).

Contenido

Para cada obra recopilada en este *dataset* se ha recogido la siguiente información:

- **Autor**: identificado por su nombre completo junto con el año de nacimiento y muerte.
- **Título**: cadena de texto con el nombre completo de la obra.
- **Catálogo**: identificador dentro del catálogo de obras más conocido de J.S. Bach (catalogo BWV, Bach-Werke-Verzeichnis) o de otros catálogos del resto de autores.
- **Descripción**: texto con una descripción del tipo de obra (cantata, coral, música de cámara, etc) y alguna información adicional en algunos casos.
- **Instrumentación**: texto con información sobre los instrumentos para los que está escrita la obra. No se sigue siempre el mismo código para identificar los instrumentos.
- **Fecha**: fecha de publicación de la obra. No siempre está completa ni se sigue la misma codificación.
- **Letras**: nombre del autor de la letra de la obra o enlace donde se puede consultar.
- **Comentarios**: texto con información adicional sobre la obra.
- **Edición**: información sobre la edición de la obra.
- **Editor**: información sobre el editor de la información.
- **Estreno**: fecha del estreno público de la obra.
- **URL**: URL de la página de donde se ha extraído la información.

Muchos de estos campos pueden estar vacíos.
Además, la información no está siempre representada de la misma forma en todas las obras, lo que disminuye la calidad del *dataset* y hace necesario una tarea de procesamiento adicional.

El *dataset* se ha recogido el 16 de abril de 2018 mediante técnicas de *web scraping*.

La información en la web se sigue actualizando, según se puede ver en la fecha de actualización de algunas obras. Probablemente los campos que más se actualicen sean los relacionados con las fuentes de datos, que no se han incluido en el data set.

Agradecimientos

Agradezco el trabajo realizado por los autores, gestores y patrocinadores de la web www.bach-digital.de
(https://www.bach-digital.de/content/contact.xml).

En la siguiente página se pueden consultar los objetivos del proyecto del portal web: https://www.bach-digital.de/content/project.xml.

Inspiración

Me gusta la música clásica y estuve buscando web con información sobre autores, obras, etc.

Esta fue la web que encontré con información más completa sobre obras de un autor y con un formato en el que, en principio, se podía extraer la información mediante *web scraping*.

Este conjunto de datos se podría emplear, sobre todo, para visualizaciones sobre la obra de J.S. Bach y su familia. Por ejemplo, se podrían intentar hacer visualizaciones para responder a las siguientes preguntas:

- ¿Cómo se relaciona el tipo de obra con la instrumentación?
- ¿Se fue modificando la instrumentación de las obras de J.S.Bach con el paso del tiempo?
- Presentación de las obras de J.S.Bach según su fecha de composición (el catálogo BWV no sigue un orden cronológico como en otros autores)
- Tipos de obras más habituales (relacionándolas con la época del año, ya que J.S.Bach debía componer corales para determinadas fiestas religiosas).

Es poco probable que se puedan desarrollar tareas de clasificación o de predicción con la información que contiene el *dataset*.

Licencia

La web de donde se ha sacado la información tiene la licencia Creative Commons Attribution-NonCommercial 4.0 International License (https://www.bach-digital.de/content/license.xml).

Por tanto, el *dataset* se debe publicar con la misma licencia.

Comentarios

El código para el proceso de *web scraping* se ha implementado en Python 2.

Se han utilizado las librerías **lxml** y **cssselect** para buscar la información en las páginas web.

Se han intentado hacer funciones generales que funcionen para los distintos casos que se han ido encontrando. A medida que se han ido haciendo pruebas se ha aprendido mejor cómo funcionaban las librerías y cómo se podían hacer búsquedas más complejas.

Probablemente se podría simplificar algo el código de la implementación, sin usar tantos tipos de búsqueda (de hecho, el tipo "cssselect-text" ya no se usa), pero se ha preferido dejar todo el código para mostrar también la evolución.

Aunque se ha intentado revisar la información del fichero robots.txt del sitio para ver si se podía hacer *web scraping*, no estaba disponible.

Mejoras

El *dataset* se puede mejorar en muchos aspectos:

- Codificar los tipos de obras que aparecen en la descripción.
- Construir otro *dataset* donde se ponga por separado la instrumentación de cada obra. Ahora mismo es una cadena de texto con diferentes codificaciones. Para poder hacer visualizaciones o clasificaciones se deberían codificar los tipos de instrumentos y separar la información en diferentes datos.
- Codificar y normalizar las fechas de creación y estreno.
- Obtener información adicional de las obras de otras fuentes (a partir, por ejemplo, de su identificador dentro del catálogo BWV).

El programa para genera el *dataset* también se puede mejorar:

- Permitir el paso de parámetros al programa para, por ejemplo, establecer el rango de páginas que se van a procesar o el nombre del *dataset*.
- Hacer varios intentos si una página no es accesible. No se ha considerado adecuado hacerlo porque no se quería incrementar la carga del servidor. Cuando no se ha podido acceder a una página ha sido porque no existía.
- Mejorar la eficiencia del código.

Recursos

A continuación se presentan algunas páginas consultadas con información sobre las librerías:

- Web de donde se ha extraído el *dataset*:
  www.bach-digital.de
- Sobre CSS Selectors:
  https://www.w3.org/TR/selectors-3/
- Sobre LXML:
  http://lxml.de/xpathxslt.html
  http://lxml.de/api/lxml.etree._Element-class.html
  http://lxml.de/tutorial.html
  http://blog.datahut.co/beginners-guide-to-web-scraping-with-python-lxml/