Proyecto Semantic Web

Etapa 1 - Grupo 1

Integrantes:

Kevin Cohen Solano – 202011864

Omar Esteban Vargas Salamanca – 201921271

Santiago Pardo Bravo – 202013024

Universidad de los Andes

Marzo 7, 2023

Descripción de las Fuentes

HP Life

HP Life es una página de cursos en línea desarrollada por la empresa HP. Esta plataforma está enfocada principalmente en la oferta de **cursos** del ámbito empresarial y emprendedor, contando con un catálogo decente de cursos y la gran ventaja de estar disponible en 8 idiomas diferentes. Los cursos se encuentran agrupados en 5 **categorías** diferentes, las cuales abarcan todos los aspectos importantes que debe tener un emprendedor. Cualquier persona puede acceder a cualquier curso sin tener que cumplir prerrequisitos de ningún tipo

La plataforma cuenta con una estructura clara que podemos entender de la siguiente manera:

- **Categorías**, siendo estas el grado más alto de la página que contienen una serie de cursos y un nombre.
- **Cursos**, siendo estas el core de la plataforma y el segundo escalón en la plataforma, contando con un nombre, una duración, un ritmo, un tema, un idioma, unos autores, un código y una serie de lecciones.
- **Lecciones**, las cuales componen a los cursos y a nivel superficial las podemos desglosar en tipo de lección y nombre.

La página presenta mucha información al principio, pero desde la vista de todos los cursos tenemos acceso a las categorías, para luego poder avanzar a cada una de ellas y ver sus cursos, por último, pudiendo entrar al preview de un curso. La opción de inscribirse a un curso no la tendremos en cuenta dado que requiere de registro, pero para el proyecto no será necesario.

GCF Global

GCF Global es una organización sin fines de lucro que brinda educación en línea gratuita a través de su sitio web https://edu.gcfglobal.org/. Ofrecen cursos y tutoriales en una amplia variedad de temas, como tecnología, matemáticas, lectura y escritura, habilidades laborales, entre otros está disponible en 3 idiomas en español, inglés y portugués. Su objetivo es ayudar a las personas a adquirir las habilidades necesarias para tener éxito en la vida y en el trabajo. Además, la organización se dedica a mejorar la calidad de vida de las personas mediante la educación y la capacitación en todo el mundo. No hay necesidad de registrarse para acceder a la información y contenido de todos los cursos, pero para guardar el progreso en dichos cursos y obtener el certificado de aprobación si se requiere que el usuario se registre de alguna forma ya sea via Gmail Facebook o creando una cuenta. Dependiendo el Idioma se enseñanza se ofrecen variedad de cursos

GCF para su versión en español ofrece todos sus cursos distribuidos en 25 Categorías(topics) las cuales son: Aritmética, Recomendaciones tecnológicas, Aplicaciones, Redes sociales, Dispositivos móviles, Sistemas operativos, Microsoft

Office, Internet, Ofimática, Informática básica, Conjuntos, Geometría, Aplicaciones y curiosidades de la matemática, Álgebra, Física, Estadística, Vida profesional, Habilidades blandas, Emprendimiento, Inglés, Podcast La Dimensión Conocida, Diseño, Herramientas de educación virtual, Ortografía y gramática, Educación financiera.

Cada uno de los topics tiene uno o más cursos asociados

La organización de esos cursos se da por medio de lecciones Cada lección esta enumerada y tiene su respectivo título y una descripción o pequeña frase relacionada con la lección. Cada lección está constituida en la misma página y tiene recursos como test, videos de youtube, texto(subtítulos y párrafos) hipervínculos a cursos o lecciones similareas. Después de terminada el contenido de la lección se puede continuar con la siguiente lección o dentro de la lección se sugieren lecciones anteriores o cursos relacionados.

Khan Academy

KhanAcademy es una empresa con fines de lucro cuya misión es proporcionar una educación gratuita de clase mundial a cualquier persona, en cualquier lugar. Este objetivo lo logran a través de categorías, cursos y lecciones interactivas que tienen en su plataforma. La información se presenta de manera organizada y separada de forma clara para hacer que el proceso de aprendizaje de los usuarios sea más fácil. Los cursos se pueden encontrar en 36 lenguajes diferentes que varían en diferentes temas, pudiendo cualquier usuario entrar a cualquiera de los cursos y avanzar entre sus lecciones de manera lineal.

En la actualidad, KhanAcademy ha establecido una estructura jerárquica para clasificar y ordenar la información, que se puede resumir en:

- **Categorías** que son el grado más alto de la página, las cuales encierran una gran cantidad de cursos.
- **Cursos** que son el segundo grado de la página pues se componen de Unidades y les ponen un orden lineal a las mismas(prerrequisitos).
- **Unidades** que son el tercer grado de la página y están compuestos por una cantidad variable de lessons, entre ellas tienen relación de prerrequisitos.
- **Lecciones** que son las unidades más específicas de conocimiento, y entre ellas hay relaciones de prerrequisitos.

Al momento de navegar atreves de la página, primero se encuentran las categorías y al hacer click sobre una de esta aparecen los cursos, los cuales tienen por dentro unidades en las cuales se muestra una relación de contenencia.

Estructura Básica de Tablas

FreeCodeCamp

Category

Name

Course

Name	Duration	Pace	Language	Author	Code	FK_CATEGORY	Link

Lesson

Type	Name	FK_COURSE

GCF Global

Category

nombre	Idioma

Course

nombre	Idioma

Lesson

nombre	número	descripción	url	idioma	topico

KhanAcademy

Category

Name	Link

Course

Name	Link	Category	CategoryLink	Language

Unit

Name	Link	Course	CourseLink	Description

Lesson

Name	Link	Unit	UnitLink	Content

Proceso de Crawling y Limpieza de datos

HP Life

Para la recuperación de la información utilizamos las librerías Selenium y BeautifulSoup. El uso de Selenium es necesario pues los HTML de la plataforma se generanm de manera dinámica, así que la librería lanza un browser que lo carga y obtiene la información completa. BeautifulSoup se usa para obtener la información de los tags necesarios.

Para el crawling se empieza obteniendo todas las categorías de la página general, esto pues están separadas con ciertas etiquetas. Luego se obtienen los enlaces que dirigen a las vistas de cada categoría y se ingresa a ellos con la ayuda de Selenium. Luego se obtienen los nombres y enlaces de todos los cursos de cada categoría (de manera muy similar) y se asocian entre ellos. Seguido a esto se ingresa a cada uno de los enlaces de los cursos y allí se recupera la información de cada curso y sus lecciones (estas últimas causaron algunos problemas, por lo que en el código dejamos una aclaración). Cada uno de estos elementos se organiza en las tablas (las mismas del paso anterior) gracias a la librería de pandas.

Por último, los datos son recuperados de la manera más organizada posible, evitando así la necesidad de hacer un proceso de limpieza posterior a los dataframes.

GCF Global

El proceso de crawling empieza identificando los labels, etiquetas y clases de la página web , en este proceso se puede observar que para los 3 idiomas la composición de las páginas es similar lo unico que cambia en la url es la especificación del idioma /es/, /pt o /en/ dependiendo el idioma despues en la página de topics se extraen todos los posibles urls de las categorias y los cursos con sus reespectivas urls y lecciones en la página de cada curso es posible acceder a si título , subtitulo de las lecciones, descripcion de las lecciones y el número de las lecciones.

Aunque en general se observa en el crawling un proceso para el proceso de limpieza de datos se hace uso de herramientas y expersiones regulares en pandas por ejemplo `re.sub('[^0-9]', '', b.text)` para obtener caracteres numericos y eliminar caracteres tipicos en html como (\n,’’,?)

Finalmente se forman los datasets con pandas para ser exportados en un .csv uno por cada idioma.

KhanAcademy

El proceso que crawling para Khan Academy empieza con el idioma a seleccionar en la url. Por ejemplo, si se quiere extraer los curso en ingles la url a iniciar seria en.khanacademy…. . Ya estando en la página principal se extraen todas las categorías con su enlace respectivo, accediendo al href de la etiqueta html.

Seguidamente, ya teniendo todos los links de las categorías, se extrañen todos los cursos de cada una de las categorías. Ya teniendo todos los cursos, se hace el scraping sobre todas las url de cada curso y de esta manera se obtienen las unidades de un curso. De mismo modo, se hace scraping sobre las unidades y se obtienen las lecciones y todo su contenido.

Finalmente, este proceso se hizo utilizando dataframes de pandas, por lo que los datos ya están limpios, debido a que mientras se ingresaban los datos se limpiaba la información. Por ejemplo en el caso de los elementos nulos se cambiaba por un string vacío y se verificaban los tipos de dato de cada columna del dataframe.