Introducción al Big Data

Ejercicio U11_E02

MasterD

1. EXPLORANDO FUENTES DE DATOS

Pasemos a otro ejercicio sencillo. No te descubrimos nada nuevo a estas alturas al decir que el interés y crecimiento de las tecnologías Big Data es producto del desbordante crecimiento de datos disponibles para analizar.

Hoy en día muchas empresas y organizaciones generan y manejan enormes cantidades de información en su día a día.

Pero para acceder a grandes conjuntos de datos y poder explorar y hacer nuestros pinitos con las diferentes tecnologías Big Data no necesitamos ser una de estas empresas.

Afortunadamente, existen gran cantidad de organismos y compañías que ofrecen conjuntos de datos de mucho valor y utilidad, de manera pública y gratuita a través de páginas y servicios web. Es lo que llamamos *open data*. Las principales fuentes de datos abiertos son las administraciones públicas y los centros de investigación, pero cada vez hay más empresas que publican parte de los datos que generan.

Cualquiera puede acceder y descargar estos datos, y utilizarlos simplemente para curiosear y examinarlos, o realizar estudios, o incluso desarrollar aplicaciones o servicios a partir de ellos.

El ejercicio que te proponemos aquí es que hagas un pequeño trabajo de búsqueda para encontrar algunos de estas fuentes de datos abiertos en internet.

Puedes clasificarlos en las siguientes categorías.

1.1. CATÁLOGOS O LISTADOS DE FUENTES DE DATOS ABIERTOS

Awesome Public Datasets

https://github.com/awesomedata/awesome-public-datasets

Awesome Data es una comunidad de usuarios (alojada en Slack) interesados en los datos abiertos. En su sitio web en GitHub ofrecen un listado de fuentes de datos abiertos de gran calidad, organizadas por temáticas.

Open Data Inception

https://opendatainception.io/

Agregador / catálogo de portales de datos abiertos de todo el mundo. Te permite explorar fuentes de datos de distintos países y regiones.

1.2. CONJUNTOS DE DATOS PARA USO CIENTÍFICO, Y PARA MACHINE LEARNING Y CIENCIA DE DATOS

Kaggle

https://www.kaggle.com/datasets

Kaggle es una comunidad online dirigida a científicos de datos. Los usuarios pueden publicar y acceder de forma abierta a multitud de conjuntos de datos, y proponer problemas de aprendizaje estadístico y *machine learning* sobre dichos datos, para resolverlos de forma colaborativa.

AWS open data

https://registry.opendata.aws/

Listado de conjuntos de datos abiertos alojados en los servicios de almacenamiento en la nube de Amazon AWS, especialmente útiles para análisis científico.

Scientific Data Repository

http://mlvis.com

Repositorio de datos abiertos de diferentes áreas de la ciencia. Además de buscar y acceder a multitud de conjuntos de datos científicos, nos ofrece visualizaciones y análisis estadísticos básicos.

1.3. DATOS ABIERTOS DE ADMINISTRACIONES Y ORGANISMOS PÚBLICOS

Portal de datos abiertos de la Unión Europea

https://data.europa.eu/euodp/en/data/

Portal de datos abiertos de España

http://datos.gob.es/es

Portal de datos abiertos de los Estados Unidos de América

https://www.data.gov/

World Bank Open Data

https://data.worldbank.org/

Repositorio de datos abiertos del Banco Mundial. Ofrece especialmente datos de actividad y desarrollo económico y social de todos los países y regiones.

1.4. DATOS ABIERTOS DE REDES SOCIALES

Twitter - API

https://developer.twitter.com/en/docs

A través de la API de Twitter podemos acceder a datos de los mensajes, temas y usuarios en tiempo real.

Facebook - API

https://developers.facebook.com/

La API de Facebook también nos permite acceder a datos y metadatos del contenido publicado por sus usuarios (siempre que estos hayan autorizado el acceso público).