DRAEX - Extractor del Diccionario de la Real Academia Española

DRAEX elimina las limitaciones de la página web del DRAE permitiendo la descarga en formato JSON todas las definiciones asociadas a una palabra (o a un conjunto de ellas) para su posterior consulta.

Página oficial del DRAE: **Diccionario Online de la Real Academia Española**

Instalación

DRAEX necesita Python 3.6 o superior para funcionar correctamente.

Instalación rápida

Con diferencia, la manera más simple y rápida de instalar *DRAEX* es utilizando *pip*:

```
pip3 install draex
```

Instalación manual

Si eso no funciona, una solución podría ser instalar las dependencias de forma manual.

Las siguientes librerías (**externas**) son necesarias para ejecutar *DRAEX*:

1. WebBot
2. BeautifulSoup

Para instalarlas, simplemente utiliza pip3:

```
pip3 install -r requirements.txt
```

Si por algún motivo eso no funcionase, instálalas individualmente con:

```
pip3 install webbot
pip3 install BeautifulSoup
```

Clona este repositorio:

```
git clone http://github.com/rmon-vfer/draex.git
```

Una vez clonado, muevete al directorio del repositorio:

```
cd ./draex/
```

Para ejecutar el programa, lee la siguiente sección

Uso

Por el momento, DRAEX solo cuenta con una interfaz interactiva, existen planes para extender la interfaz actual e implementar un modo *avanzado*, si quieres colaborar, haz un *fork*, haz los cambios que creas convenientes y despues haz un *pull request* a este repositorio.

Para usar el modo interactivo de DRAEX, simplemente escribe:

```
python3 draex.py
```

Implementación

La implementetación de DRAEX es bastante simple si atendemos al problema que la originó, la página web del Diccionario Online de la Real Academia Española , que impide mediante un *challenge* (reto) de JavaScript acceder a la página si no es mediante un navegador con un motor capaz de resolverlo, (*por ejemplo, V8 de Google*)

En la mayoría de los casos es bastante tedioso hacerle ingeniería inversa a un JS Challenge, así que en lugar de hacer eso, decidí renderizar la página web con el WebDriver de Chromium en modo Headless para así resolverlo y obtener el código fuente de la página. Una vez obtenido es bastante sencillo emplear el HTML para obtener las definiciones y demás datos.

Resultados

Logs

El programa genera un archivo log que puede servir posteriormente para su depuración:

main.log

```
[20190225-125655] | Iniciando...
[20190225-125655] | No he podido crear un directorio especial para
almacenar las palabras dumpeadas, ejecútame como administrador.
[20190225-125655] | Almacenaré las definiciones en este mismo directorio
[20190225-130540] | KeyboardInterrupt se ha lanzado, interrumpiendo
aplicacion...
```

Definiciones

Las definiciones de almacenan por duplicado, una copia se almacena en un archivo JSON `general.json` y otra copia en un archivo cuyo nombre es la palabra definida.

El archivo de definición individual, está estructurado en dos secciones, de acuerdo al uso que recibe cada significado de la palabra:

1. Uso Normal (*denominado* `[UNORM]` en el log)
2. Uso Especial (*denominado* `[USPEC]` en el log)

Las diferentes definiciones presentan el mismo orden que en la página web original.

Un ejemplo de `<palabra>.json` es el siguiente

diccionario.json

```
{
    "usoNormal" : {
        1 : "Definicion1",
        2 : "Definición2",
        3 : "Definición3
    },
    "usoEspecial" : {
        1 : "Definicion1",
        2 : "Definición2",
        3 : "Definición3
    }
}
```

Notas

Aviso **legal** (muy importante)

Ni yo ni ninguna de las personas que han colaborado conmigo en el desarrollo de esta herramienta deseamos dañar en modo alguno a la *Real Academia Española*, el presente software es liberado bajo la licencia MIT, y su única finalidad es servir de ejemplo para el aprendizaje de las tecnologías y la seguridad web.

> Si utilizas este programa asumes las condiciones de la licencia MIT y estás de acuerdo en que **no lo utilizarás para obtener ningún beneficio** más allá del puro conocimiento.

Aviso de desarrollo

Este software está en desarrollo, lo que implica que las cosas pueden cambiar de un día para otro sin previo aviso, pueden dejar de funcionar o directamente desaparecer entre una versión y otra.

Colaboración

Si quieres colaborar, **haz un fork** y completa alguno de los (muchos) `TODO`'s que pueblan el código, intenta solucionar algún bug o implementa alguna característica interesante.

Una vez que tengas la nueva característica/bug..., haz una *pull request* a la *main branch* de este repo ~~(quizá estaría bien abrir una *development branch*)~~.