Universidad Técnica Particular de Loja
Ingeniería en Sistemas Informáticos y Computación

Sistemas Basados en Conocimientos

Proyecto Bimestral - Análisis y recolección de datos

- Fuertes Alejandro
- Román Andrés

Documentación de la API de Wikipedia

La API de MediaWiki es una interfaz madura y estable que se apoya y mejora activamente. En el siguiente enlace se puede ingresar a la portada de la API, en donde se puede observar de manera general acciones de MediaWiki.

https://www.mediawiki.org/wiki/MediaWiki

La API de acción de MediaWiki es un servicio web que permite el acceso a algunas funciones de la wiki como autenticación, operaciones de página y búsqueda. Puede proporcionar meta información sobre la wiki y el usuario que inició sesión.

Entrada

La API toma su entrada a través de parámetros proporcionados por la solicitud HTTP en formato application / x-www-form-urlencoded o multipart / form-data. Cada módulo y submódulo tiene su propio conjunto de parámetros, que se enumeran en la documentación y en action = help. También se pueden recuperar mediante action = paraminfo.

Codificación

Todas las entradas deben ser UTF-8 válidas, en formato NFC. MediaWiki intentará convertir otros formatos, lo que puede resultar en un error.

Salida

El formato de salida estándar y predeterminado en MediaWiki es JSON. Se desaconsejan todos los demás formatos. El formato de salida siempre debe especificarse usando format = yourformat siendo yourformat uno de los siguientes:

1. json
2. php
3. xml
4. txt
5. dbg
6. yaml
7. wdds

8. dump
9. none

Los metadatos seleccionados de la API son: pageid, revid, title, links y secciones.

Obtención de datos

Para la creación de un corpus de información de Wikipedia se debe definir un concepto raíz, en este caso se tomará como concepto el recurso que se aloja bajo la categoria de Realidad Virtual (VR) que se aloja en DBPedia (https://dbpedia.org/page/Category:Virtual_reality). El primer paso es la creación de una base de datos local, para lo cual se utilizó MySQL como servidor y en él se define el esquema con el que se va a trabajar. El siguiente paso es la obtención de subconceptos existentes que tiene el nodo raíz bajo la relación de *skos:broader*, en este caso se llegó hasta el tercer nivel de recorrido en DBPedia, obteniendo un total de 772 subconceptos. Para obtener estos subconceptos se utilizó una consulta SPARQL que recorre la red de conceptos cercanos al nodo raíz que se van insertando en una tabla de datos SQL. El siguiente paso es limpiar los subconceptos encontrados y obtener una lista única, paso que se explica más detalladamente en el siguiente punto.

El segundo paso es obtener los links y los metadatos de las páginas de Wikipedia que están relacionados con los subconceptos encontrados anteriormente para ello se crea una tabla donde se almacenarán estos datos y se hace una consulta a la API de Wikipedia con los datos ya limpiados, obteniendo el link de DBPedia y de Wikipedia de cada uno de los conceptos y sus metadatos. En este paso también se realizó una limpieza de datos de las páginas obtenidas, explicado también en el siguiente punto.

El último paso fue la creación de una nueva tabla donde se almacenarán las URLs de Wikipedia de todos los subconceptos que tienen relación con el concepto raíz, haciendo uso de las redirecciones obtenidas desde la DBPedia.

Limpieza de datos

La primera limpieza de datos que se realizó fue sobre los subconceptos obtenidos en los 3 niveles de recorrido en DBPedia, para ello se recorrió la tabla de subconceptos obtenida en el primer paso y mediante sentencias SQL se realiza selecciones únicas de los subconceptos para que no sean repetidos

La segunda limpieza de datos se realizó sobre los links de Wikipedia obtenidos, para ello se crea una tabla temporal con los IDs y una tabla para las redirecciones de esos mismos links. En este punto se hizo uso de la librería BeautifulSoup que analiza los documentos HTML y crea un árbol con todos sus documentos para extraer la información pertinente, utilizada sobre todo para temas de scrapy, en este caso para el scrapping de la API de Wikipedia.

1. ESPECIFICACION DE FUENTES DE DATOS

1.1 Identificación de Fuentes

Nombre de la Fuente de Datos:	DBPedia
Proveedor:	DBPedia
Sitio Web:	http://dbpedia.org/resource/Category:Virtual_reality
Institución:	Universidad de Leipzig y Universidad de Mannheim
Descripción de la fuente de datos:	DBpedia es un proyecto para la extracción de datos de Wikipedia para proponer una versión Web semántica. Este proyecto es realizado por la Universidad de Leipzig, Universidad Libre de Berlín y la compañía OpenLink Software.
Tamaño del archivo de datos (MB)	2
Licencia	Creative Commons
Fomato del archivo:	Base de datos relacional
Incluye información de los metadatos?	Inglés

1.2 Volumetría de datos

Entidad[atributos]	**Cantidad de registros**	**Descripción**
DBCSubconceptsVR [date, r, c1, c2, c3, level]	772	Subconceptos del nodo raíz existentes en DBPedia hasta el nivel 3
DBCSubconceptsVRIDs [id, c]	651	Lista única de subconceptos del nodo raíz después de la limpieza de datos
DBCSubconceptsVRList [level, c1]	731	Lista única de subconceptos del nodo raíz
DBSubconcepts_WikiPages [date, concept, dbrResource, wikiPage, wikipageModified, max_outDegree]	6728	Páginas de Wikipedia de los resources de cada uno de los subconceptos con sus fechas de modificación
WikiPagesIDs [id, wikipageUrl]	3284	URL de las páginas de Wikipedia después de la limpieza de datos
WikiPagesRedirects [date, id, dbpUrl, returnedUrl]	3410	Redirecciones de la página de Wikipedia con límite de 3410

1.3 Obtención de datos

Método de extracción de datos	Consumo de API
En caso de requerir construir un script o app de extración de datos, indicar detalles como lenguaje de prog., descripción de lo que realizó el programa	Lenguajes: python y SQL Se obtuvo los subconceptos de un nodo ríaz predeterminado y se realizó la limpieza de datos para obtener listas únicas, a partir de los subconceptos se obtuvo los links y los metadatos de DBPedia para después obtener los links de Wikipedia relacionados con el nodo raíz
Base de datos en la que guardó los resultados	MySQL

DIAGRAMA DE INTEGRACIÓN DE FUENTES DE DATOS