Social Networks
Course Project
KML (Knowledge Markup Language)

Amit Kumar Verma (mentor)
Paras Kumar
Nitin Gandhi

Contents

What is KML?

KML stands for Knowledge Markup Language, a standard format for storing the data of all the Knowledge Building Portals. Knowledge Building portals like Wikipedia, Stack Exchange, GitHub, e.t.c provides their data dump in their own formats. We are trying to propose a new standard format for all these types of portals such that the analysis is easy.

Need for KML

All these Knowledge Building portals provide their data dumps in their own format. For example, Wikipedia provides its data in an XML format with their own schema definition. Similarly, Stack Exchange provides its data dump in an XML format with different schema definition. The KML will be a new standard format for these kinds of Knowledge Building portals with a standard schema definition. The idea is to make KML flexible enough such that it can store the data of any kind of Knowledge Building portals.

Components

- KML-Compressor-Decompressor
- KML-for-Github
- Github-Spider
- User-Agent-Spider
- Wiki-Satck-KML-Downloader

Source Code: `https://github.com/csl-622/KML`

KML-Compressor-Decompressor

These programs compresses and decompresses a KML file using diff algorithm inspired from git. It can compress the KML file by 30 percent or more depending upon the edits made on an article. The compressed KML is also very easy to read, since the number of lines is reduced from 1 lakh to 9k and it retains the original structure of the KML so that it remains redable even when compressed.

KML-for-Github

We had to make KML for github and we had KML for wikipedia earlier. The attempt of this program is to retain the structure of KML for wkipedia while representing all the data in a git workflow in this newly generated KML file.

Github-Spider

This multithreaded web-spider follows the rules of /robots.txt file and is useful for fetching small sized GitHub repositories. It outputs the results in JSON format.

Github-Spider-Proxy-Rotated

Github-Spider-Proxy-Rotated is an advanced version of the previous spider; it contains pipelines, middlewares, and throttling parameters, which is in the middleware and settings file. This spider is for advanced scraping. It has a bottleneck mechanism for limiting the number of HTTP requests per second, and also a technique for HTTP headers rotation. It outputs the results in JSON format.

User-Agent-Spider

This spider gets the latest user-agent headers from an online forum, output of this file is directly put into use in other two spiders namely Github-Spider and Github-Spider-Proxy-Rotated. The result of this spider is a text file containing a list of HTTP headers which can is useful for spoofing browser's HTTP request activity.

Wiki-Satck-KML-Downloader

Dowloads data from Wikipedia and Stackoverflow and then coverts it to KML.