Projeto de Pesquisa

Nº de Estudantes:	**1**

Indicação do perfil de cada estudante	
Curso Bacharelado em Sistemas de Informação	Período que está Cursando 3

Titulo

O desenvolvimento de um software em Python para baixar, no formato SQL, os dados dos alunos dos cursos do IFG cadastrados no SISTEC.

Palavras-chave

Python, SQL, Download, SISTEC

Resumo

O presente projeto de pesquisa tem por objetivo criar um software para baixar, no formato SQL, os dados de todos os alunos dos cursos do IFG que estão cadastrados no SISTEC. O site do SISTEC retorna os dados dos alunos por Ciclo de Matrícula de cada curso, no formato JSON. Assim sendo, o software deverá também transformar esses dados JSON no formato SQL. Uma vez no formato SQL, esses dados podem ser usados para comparação com os dados dos alunos do IFG cadastrada pelo Q-Acadêmico.

Apresentação/Justificativa

Segundo o próprio site do SISTEC[1], temos a seguinte informação sobre o que é o SISTEC:

> O SISTEC é o Sistema Nacional de Informações da Educação Profissional e Tecnológica. Esse sistema é pioneiro e, portanto, inovador no País por disponibilizar, mensalmente, informações sobre cursos técnicos de nível médio, respectivas escolas e alunos desse nível de ensino. Caso a escola também ofereça cursos de formação inicial e continuada, o SISTEC apresentará ainda dados referentes aos cursos e aos alunos dessa oferta de ensino. Contudo, é importante ressaltar que os cursos de formação inicial e continuada só serão cadastrados se a escola ofertar ensino técnico de nível médio.
>
> Os órgãos competentes de cada sistema de ensino dispõem agora de um importante instrumento para atestar a validade nacional dos diplomas.

O IFG utiliza o sistema Q-Acadêmico[2] para gestão dos seus processos administrativos. O Q-Acadêmico é desenvolvido e mantido pela Qualidata[3]. Segundo a própria Qualidata, temos a seguinte definição do Q-Acadêmico:

> O sistema, batizado de **Q-Acadêmico**, é modularizado de forma em que os sistemas de Controle Acadêmico, Controle de Processo Seletivo, Controle de Acesso e Controle de Biblioteca integram-se totalmente gerando uma única base de informações para toda instituição de ensino, permitindo os mais diversos relatórios gerenciais e estatísticos.

Desta forma é de extrema importância manter os dados do SISTEC atualizados com os dados do IFG. Assim, informações conflitantes entre as duas bases de dados – do SISTEC e do IFG – poderão ser resolvidas mais facilmente.

Portanto, o presente projeto de pesquisa se justifica pelo fato do site do SISTEC não disponibilizar nenhum método de baixar todos os dados automaticamente. Assim, se querermos obter os dados de todos os alunos dos cursos do IFG que estão cadastrados no SISTEC, tem-se que acessar cada Ciclo de Matrícula cadastrado no SISTEC. E, em se tratando de uma instituição como o IFG com 14 campus, com muitos cursos, fica portanto muito tedioso e lento ter que baixar toda a informação manualmente, ciclo por ciclo de matrícula de cada curso.

Vale ressaltar que os dados obtidos do SISTEC serão confrontados com os dados do IFG por outro software denominado VISÃOIFG, o qual, não faz parte do escopo da presente proposta de projeto de pesquisa.

1) SISTEC: http://sitesistec.mec.gov.br/o-sistema-menu-principal-140
2) Q-Acadêmico: http://www2.qualidata.com.br/gestaoacademica.htm
3) Qualidata: http://www2.qualidata.com.br/

Objetivos

- Objetivo Geral

Desenvolver um software em Python para efetuar automaticamente o download de todos os dados dos alunos dos cursos do IFG que estão cadastrados no SISTEC.

Esse software em Python deverá raspar as páginas do site do SISTEC e extrair as informações necessárias para confrontar com a base de dados do IFG.

Assim, ficará mais fácil visualizar e corrigir as diferenças entre as bases de dados do SISTEC e do IFG.

- Objetivos Específicos

O software a ser desenvolvido deverá extrair os dados das páginas do SISTEC – que se encontram no formato JSON – e inseri-los num banco de dados no formato SQL, para então outro software – VISÃOIFG – confrontar os dados obtidos do SISTEC com os dados do IFG.
Primeiramente o software efetuará o login do usuário no SISTEC, para somente depois baixar os dados no formato JSON e convertê-los em um arquivo SQL que poderá ser inserido em qualquer banco de dados padrão SQL ANSI.

Material e métodos

O software a ser desenvolvido usará a biblioteca gráfica WxPython, por se tratar de uma biblioteca de código Python multi plataforma, isto é, o mesmo software python poderá ser executado em diferentes plataformas de software como: Mac OSX, Windows, Linux, FreeBSD, e demais Unix-like.

Usaremos também a biblioteca PycURL para fazer as requisições ao site do SISTEC por se tratar de uma biblioteca com mais recursos do que a biblioteca padrão do Python. Usaremos os recursos da PycURL para efetuar o login no SISTEC, coletar os cookies da sessão do usuário no SISTEC para então acessar os dados no SISTEC e transformá-los, passo a passo, em dados num arquivo SQL local.

Assim, o software a ser desenvolvido executará o loop tradicional de raspar a web segundo Etheriel (2015):

1. Fazer uma requisição para uma URL;
2. Processar a resposta (HTML, XML ou JSON);
3. Extrair os dados;
4. Deduzir as próximas URLs a visitar;
5. Repetir o loop.

A Lista dos campus do IFG no SISTEC é obtido por meio de uma resposta HTML do site do SISTEC.

Para extrair as informações dos campus do IFG do HTML obtido do SISTEC, usaremos a tecnologia de expressões regulares por meio da biblioteca padrão 're' do Python.

Segundo Jargas (2012) a tecnologia de expressões regulares é uma forma prática de economizar muito tempo de serviço ao lidar com textos padronizados como HTML.

De acordo com a lista de campus obtida, faremos requisições para o site do SISTEC a fim de obtermos a lista de todos os Ciclos de Matrícula por curso e por campus. Obteremos essas informações no formato JSON. E, por meio da biblioteca JsonPare[1] faremos a extração desses dados obtidos, para então podermos requisitar os dados dos alunos de cada Ciclo de Matrícula do IFG cadastrado no SISTEC, por curso e por campus. Assim, obteremos um novo arquivo JSON como resposta. Então, extraímos os dados obtidos desse novo arquivo JSON novamente com a biblioteca JsonPare, e, transformamos esta informação em uma linha do arquivo SQL com o devido comando 'insert into table'.

Desta maneira, o arquivo SQL gerado, poderá ser importado em qualquer gerenciador de banco de dados que seja compatível com o formato SQL ANSI, como MySQL, PostgreSQL, SQLite, etc.

Resultados esperados

Espera-se produzir um meio mais fácil para efetuar o download dos dados referentes a todos os Ciclos de Matrícula do IFG cadastrados no SISTEC.

Também esperamos aprofundar e difundir o conhecimento e a técnica de raspagem de dados da Web por meio da linguagem de programação Python, apresentando o presente trabalho na PythonBrasil 2015 que será realizada no *Parque Tecnológico de São José dos Campos – São Paulo*, de 13 a 17 de outubro de 2015.

E, por fim, esperamos difundir esta prática para outras instituições de ensino que usam o SISTEC.

1) JsonPare: o código desta biblioteca Python pode ser obtido a partir de https://github.com/dsoprea/JsonPare

Referências bibliográficas

Etheriel, Capi. **Raspando a Web com Python: Introdução.** Disponível em: http://pythonclub.com.br/raspando-a-web-com-python-parte-1.html. Acessado em: 13/07/2015.

Jargas, Aurelio Marinho. **Expressões Regulares - Uma abordagem divertida.** São Paulo: Novatec, 2012.