CEX - CÂMARA DE CIÊNCIAS EXATAS E DOS MATERIAIS (PÔSTER)

NOME: VICENTE CESAR AMORIM SILVA

TÍTULO: DESENVOLVIMENTO DE UMA PLATAFORMA PARA EXTRAÇÃO, INTEGRAÇÃO E ANÁLISE DE GRANDES REPOSITÓRIOS DE DADOS WEB

AUTORES: PATRÍCIA MASCARENHAS DIAS, VICENTE CESAR AMORIM SILVA, VICENTE CESAR AMORIM SILVA

AGÊNCIA FINANCIADORA (se houver): PAPq/UEMG

PALAVRA CHAVE: EXTRAÇÃO DE DADOS, SCRAPY, TESES E DISSERTAÇÕES

RESUMO

Nos últimos anos diversos trabalhos de teses e dissertações estão sendo coletados e reunidos em bancos de dados online, para que possam ser acessados por pessoas em todo mundo. Um exemplo desse tipo de base dados é a plataforma “Catálogo de Teses e Dissertações” da CAPES, onde mais de um milhão de trabalhos coletados ao longo dos anos foram armazenados. Com esse trabalho, está sendo desenvolvida uma plataforma de extração de dados em Phyton, que recolherá da plataforma da CAPES a informação de todos os trabalhos realizados e salvos.

Utilizando da linguagem Phyton juntamente com a biblioteca Scrapy, é possível se conectar a base de dados construída pela CAPES e pesquisar sobre todas as teses contidas nesse repositório, através de requisições consecutivas à página. Os dados serão filtrados retirando algumas informações importantes como autor, área de atuação e instituição de ensino. Depois disso, os dados filtrados serão formatados em um arquivo e passarão por uma análise.

Utilizando dos resultados obtidos pela plataforma de extração, será aplicado futuramente uma análise sobre os dados utilizando o software Gephi. Baseado nas análises irão se relacionar os dados encontrados, para que pessoas com interesse em determinada área científica possam encontrar facilmente trabalhos e pessoas que entendem da área, facilitando o uso das bases de dados existentes. O intuito é criar uma rede de conhecimento onde pessoas possam facilmente encontrar bases teóricas e auxílio para o desenvolvimento de futuros trabalhos.