Relatório - Coleta de Dados

Coleta de dados

Base de dados com informações sobre a fase jurídica da persecução penal à corrupção

Como descrito no projeto inicial de pesquisa, o presente estudo contará com uma lista de processos judiciais, obtidos diretamente dos tribunais e através da leitura automática de DJEs. Este projeto possue como recorte a obtenção de dados em quatro estados, sendo eles, São Paulo, Alagoas, Rio de Janeiro e Distrito Federal.

No que diz respeito a obtenção diretamente dos tribunais, estruturou-se uma estratégia para coletar informações dos Tribunais de Justiça, das Justiças Federais e dos Tribunais Regionais Federais dos respectivos estados citados. Além disso, a coleta teve foco nos processos categorizados dentro da classe **inquérito policial** e nas suas respectivas Jurisprudências e Consultas Processuais, para permitir uma coleta completa dos dados.

Cabe destacar que a coleta manual não seria tão ágil devido a quantidade de processos e consequentemente o tempo atribuido a sua coleta, sendo assim, automatizamos o processo por meio de uma técnica conhecida como *"web scraping"*, que compreende a coleta automatizada, também chamada de raspagem, de informações disponíveis em arquivos HTML das quais todos os tribunais possuem, uma vez que o site de acesso para eles são estruturados neste tipo de arquivo.

Porém, apesar de estruturados em HTML, o sistema para coleta de informações dos processos é diferente entre tribunais e Estados, o que dificultou a estratégia de coleta. A solução, portanto, foi construir "rôbos" especificos para cada site para posteriormente obter as informações sobre os processos.

Após a coleta dos dados, era preciso estruturá-los em tabelas visando uma análise consistente dos dados e assim como os sites era diferentes, a forma como o dado era obtido também. Portanto, foi necessário desenvolver algoritimos diferentes para transformar as informações em tabelas.

A obtenção da lista de processos judiciais através da leitura automática de DJEs, assim como argumentada no projeto inicial é mais difícil, uma vez que muitos destes documentos são obtidos no formato PDF. Logo, para obter estas informações foi necessário contruir "rôbos" para fazer a raspagem dos sites HTML dos diários, estas raspagens então são capazes de obter o download dos diários que posteriormente irão passar por uma

estruturação capaz de permitir análises consistentes dos dados.