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 [SE O PROJETO SE RESTRINGE SOMENTE A SEGUNDA INSTANCIA JF SAI] 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,

Table 1: Relação de tribunais coletados

Tipo do tribunal	Tipo do scraper	Implementado
JFAL	Consulta Processual	SIM
JFDF		SIM
JFRJ		SIM
JFSP		SIM
TJAL	Jurisprudência	SIM
	Consulta Processual	SIM
TJDFT	Jurisprudência	SIM
	Consulta Processual	SIM
TJRJ	Jurisprudência	SIM
	Consulta Processual	SIM
TJSP	Jurisprudência	SIM
	Consulta Processual	SIM
TRF-1	Jurisprudência	SIM
	Consulta Processual	SIM
TRF-2	Jurisprudência	SIM
	Consulta Processual	SIM
TRF-3	Jurisprudência	SIM
	Consulta Processual	SIM
TRF-5	Jurisprudência	SIM
	Consulta Processual	SIM

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.

Nas seções seguintes apresentaremos mais detalhadamente sobre como foram realizadas as coletas, bem como os sites visitados para obtenção das bases. As seções serão divididas pelo tipo de coleta, isto é, a listagem dos processos através dos tribunais e pelos DJEs.

Listagem dos processos (Tribunais)

A tabela abaixo apresenta quais tribunais foram foco de coleta e o andamento delas até então. A coluna `Tipo do tribunal` apresenta o tribunal foco da coleta, a `Tipo do scraper` é para indicar qual foi a coleta realizada, a `Implementado` é para sinalizar se o algoritimo para obter os dados foi realizado.

A coleta foi realizada nos sites oficiais dos tribunais[1]. Sites como o de São Paulo e Alagoas, por possuirem sistemas semelhantes (e-SAJ), tiveram rotina parecida de coleta. Além disso, para o caso de São Paulo especificamente, foi utilizado o pacote `esaj` desenvolvido pela ABJ para uso na linguagem R.

[1]Os links utilizados para acessar os sites estão no Anexo. [TALVEZ ISSO MUDE PARA SEÇÃO 2.2]

Table 2: Relação de DJEs coletados e estruturados

Tipo do tribunal	Implementado	Estruturação do PDF
TJAL	SIM	NÃO
TJSP	SIM	NÃO
TJRJ	SIM	NÃO
TJDFT	SIM	NÃO
JFAL	NÃO	NÃO
JFSP	SIM	NÃO
JFRJ	SIM	NÃO
JFDF	SIM	NÃO

Listagem dos processos (DJE)

A coleta de dados dos Diários de Justiça Estaduais segue os passos da listagem anterior, isto é, é preciso utilizar a técnica de *web scraping* para coletar informações sobre os diários, que normalmente são disponibilizados em PDF, após a coleta destas informações e download dos arquivos, é preciso também transformá-las em arquivos que permitem uma análise consistente.

A tarefa relacionada com a coleta dos dados está na sua etapa final, uma vez que a maioria dos diários já possui algoritmos para obter os arquivos em PDF. Sendo assim, ao terminar a última coleta poderemos dar início a estruturação dos arquivos. A tabela abaixo mostra os trabalhos realizados até então.