Coleta de dados

3 de Maio de 2018

Conteúdo

1 Base de dados com informações sobre a fase jurídica da persecução penal à corrupção **2**

1.1 Listagem dos processos (Tribunais) . . . 3

1.2 Listagem dos processos (DJE) . . . 4

2 Base de dados de perfis dos entrevistados **4**

2.1 Identificação . . . 5

2.2 Perfis . . . 6

3 Anexo **6**

3.1 Lista de sites consultados . . . 6

1 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, permitindo então uma consistência maior para as futuras análises.

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 das informações. Devido ao fato dos sites terem formatos diversos, a transformações dos dados brutos em tabelas também teve que passar por diferentes tipos de estruturação.

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 de informações, por meio do que chamamos de *mineração de texto*.

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.

Tabela 1: Relação de tribunais coletados

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

1.1 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 e estruturado.

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

[1]Os links utilizados para acessar os sites estão no Anexo.

caso de São Paulo especificamente, foi utilizado o pacote `esaj` desenvolvido pela ABJ para uso na linguagem R.

Porém, é preciso destacar alguns problemas que dificultaram a coleta, como por exemplo, `CAPTCHAs` presentes na consulta processual do TRF-3 e da JFAL e problemas na busca por inquérito policial na jurisprudência do TJRJ. Mais especificamente sobre o TJRJ, o site impossibilitava a busca pela classe de inquérito policial e as busca apenas por inquérito retornavam resultados não satisfatórios. Sendo assim, a estratégia utilizada foi realizar a coleta por intervalos menores de tempo e filtrar os casos desejados posteriomente. Apesar dos problemas relatados, cerca de 85% dos códigos para os respectivos tribunais estão implementados de forma integral, isto é, os códigos de coleta e estruturação dos dados estão feitos. Aqueles sinalizados como "PARCIAL" significam que possuem o código de estruturação feito, porém o de obtenção está em andamento.

1.2 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 e estruturá-los em arquivos de texto. A estratégia utilizada para a coleta e estruturação dos dados compreende baixar todos os diários de interesse, neste caso relacionados com a área Judicial, transformá-los em arquivos de texto para que em seguida possamos utilizar expressões regulares[2] de acordo com assuntos específicos, obtendo as pagínas que posteriormente coletaremos os números CNJ.

Com o número CNJ em mãos podemos executar os códigos das Consultas Processuais e assim estruturar as respectivas bases de dados.

2 Base de dados de perfis dos entrevistados

Paralelamente a coleta de dados dos fluxos dos processos penais dos casos de corrupção nos estados previamente selecionados, a pesquisa se propõe a criar um banco de perfis dos agentes chave desse processo, a saber: juízes federais e estaduais, procuradores federais e estaduais e delegados federais e estaduais.

[2] Expressões regulares são sequências específicas utilizadas para encontrar padrões em caracteres.

Para além de variáveis demográficas, como idade, sexo, ocupação, escolaridade, local de nascimento etc., o banco visa combinar características referentes a trajetória de carreira dos entrevistados, se passaram pelo setor privado ou não, se possuem formação secundária, se possuem algum tipo de vinculação a associações ou movimentos sociais. Assim, pode-se cruzar essas variáveis, criando grupos de perfis (clusters) e analisar as relações entre esses grupos e os resultados das decisões dos processos.

2.1 Identificação

A obtenção desse banco de perfis se deu, em primeiro lugar, pela identificação dos atores pertinentes; através dos sites oficiais dos Tribunais de Justiça dos quatro estados selecionados e dos sites dos Tribunais Regionais de Justiça, foi possível obter acesso às listas de juízes, que foram filtradas pelas categorias "criminal", "crim", "vara única" ou "vara cumulativa", ou no próprio site ou em comparação com o banco de dados interativo da Conselho Nacional de Justiça (nesse caso, foi produzida uma nova tabela e realizado o cruzamento entres as comarcas e os nomes). As identificações dos procuradores se deram por meio dos sites da Advocacia Geral da União e do Ministério Público em cada estado, sendo as listas dos procuradores estaduais encontradas nos tópicos "Colégio de Procuradores Estaduais", nos sites do MP, ao passo que para os procuradores federais essa informação constava nas seções institucionais dos site da AGU, geralmente no tópico "Quem é Quem". Por fim, os delegados foram identificados nas páginas da Polícia Federal e da Polícia Civil, sendo possível realizar uma busca por estados a partir da página principal da PF, no caso dos delegados federais. No entanto, em relação os delegados estaduais essa informação só foi encontrada na página da Polícia Civil do Rio de Janeiro.

É importante ressalvar que as listas de identificação não estão todas completas, isso se deve a disparidade da disponibilização de informações entre os sites, às vezes entre os sites de uma mesma instituição, como no caso dos Tribunais de Justiça, que ora apresentam informações agregadas, incluindo diversos servidores em uma mesma lista, ora apresentam documentos separados e ora apresentam buscas com filtros na própria interface do site, ao invés de um documento com uma lista, dificultando a busca e a extração da informação necessária.

Outra razão para incompletude das listas é a própria falta de informação nos sites oficiais, como no caso dos sites das Policias Civis do Alagoas, do Distrito Federal e de São Paulo que sequer apresentam alguma seção informando quais são os responsáveis pelas delegacias.

A busca pela identificação dos possíveis entrevistados prosseguirá com telefonemas às corregedorias dos órgãos em que não obtivemos informações pelos sites ou que as informações estavam

incompletas, com o intuito de cobrir os *missing values* das tabelas com o fornecimento da informação diretamente dos órgãos.

2.2 Perfis

Em decorrência dos problemas na obtenção das identificações dos possíveis entrevistados, a fase de levantamento dos perfis dos agentes encontra-se em defasagem até o presente momento.

Algumas tentativas e esforços já foram realizados nesse sentido considerando as identificações que já possuímos; ao se procurar individualmente os nomes de delgados, procuradores e juízes em uma busca simples na internet, os resultados foram variados, podendo aparecer informações em plataformas profissionais como Lattes, LinkedIn, Escavador etc. ou até mesmo mais pessoais como Facebook, por exemplo. No entanto, esses experimentos mostraram a inconstância que essas informações podem ser obtidas, como se tratam de plataformas em que os usuários se cadastram voluntariamente e colocam as informações que desejam, era comum não encontrar o entrevistado cadastrado em nenhuma plataforma ou em apenas alguma delas, não havendo uma base comum em que se possa extrair informações para todos eles, dificultando comparações e aproximações.

Outra dificuldade encontrada foi a ausência de dados biográficos dos investigados tanto nas plataformas oficiais, via de regra, quanto em buscas livres na internet, algo que era esperado, por se tratarem de funcionários públicos concursados ou de carreira.

Na tentativa de automatizarmos as buscas, foram testados dois pacotes do software R com o intuito de criar loops de procura e raspagem de dados nas plataformas Lattes e Facebook, porém não obtivemos sucesso. Uma provável dificuldade nessa tentativa talvez seja a de lidar com o léxico da procura, uma vez que resultados idênticos ou semelhantes aos nomes dos procurados pode atrapalhar a busca e conduzir a informações desviantes.

Por fim, novas tentativas serão feitas nesse sentido e considera-se caso essa opção de busca on-line não mostrar-se viável, anexar perguntas de perfil aos questionários enviados aos entrevistados.

3 Anexo

3.1 Lista de sites consultados

Jurisprudência

- TJSP: https://esaj.tjsp.jus.br/cjsg/consultaCompleta.do

- TJAL: https://www2.tjal.jus.br/cjsg/resultadoCompleta.do e http://www.jurisprudencia.tjal.jus.br/Default.aspx
- TJRJ: http://www.tjrj.jus.br/search?site=juris&client=juris&output=xml_no_dtd&proxystylesheet=juris
- TJDFT: https://pesquisajuris.tjdft.jus.br/IndexadorAcordaos-web/sistj?visaoId=tjdf.sistj.acordaoeletronico.buscaindexada.apresentacao.VisaoBuscaAcordao
- TRF-1: https://jurisprudencia.trf1.jus.br/busca/resultado.jsp
- TRF-2: http://www10.trf2.jus.br/consultas/
- TRF-3: http://web.trf3.jus.br/base-textual
- TRF-5: http://www.cjf.jus.br/juris/unificada/Resposta

Consulta Processual

- TJSP: http://esaj.tjsp.jus.br/esaj/portal.do?servico=190090
- TJAL: https://www2.tjal.jus.br/cposg5/search.do
- TJRJ: http://www4.tjrj.jus.br/ejud/WS/ConsultaEjud.asmx/DadosProcesso_1
- TJDFT: https://cache-internet.tjdft.jus.br/cgi-bin/tjcgi1?NXTPGM=plhtml01&MGWLPN=SERVIDOR1&submit=ok&SELECAO=1&CHAVE=&ORIGEM=INTER
- TRF-1: https://processual.trf1.jus.br/consultaProcessual/processo.php
- TRF-2: http://www.trf2.gov.br/cgi-bin/pingres-allen
- TRF-3: http://web.trf3.jus.br/consultas/Internet/ConsultaProcessual
- TRF-5: http://www.trf5.jus.br/cp/cp.do
- JFSP: http://csp.jfsp.jus.br/csp/consulta/consinternet.csp
- JFAL: https://jef.jfal.jus.br/cretainternetal/consulta/processo/pesquisar.wsp
- JFRJ: http://procweb.jfrj.jus.br/portal/consulta/cons_procs.asp
- JFDF: https://processual.trf1.jus.br/consultaProcessual/processo.php