Procedimento metodológico

## Introdução

A metodologia fará uso de meios para identificar as influências recíprocas entre a imprensa jornalística impressa e digital, de um lado, e o sistema de justiça, de outro, a fim de verificar a existência de papéis compartilhados na superlotação e na superpopulação carcerária no Brasil.

Para tanto, os pesquisadores irão coletar, organizar, anotar e analisar notícias jornalísticas que façam referência ao crime e ao sistema de justiça, bem como, julgados de primeiro e de segundo grau dos tribunais listados no anexo@.

A pesquisa tem caráter eminentemente exploratório, de modo que, a princípio, não há hipóteses a serem testadas, mas sim questões a serem respondidas durante a análise do material coletado. Dito isso, caso alguma hipótese seja levantada durante a fase exploratória, ela poderá ser incluída no desenho da pesquisa a fim de ser testada.

Todo o processo de coleta, organização e análise dos dados será feito no ambiente de programação *R*, um software livre voltado à pesquisa científica. O trabalho de anotação será realizado por meio do software igualmente livre inception [[1]](#footnote-21)

## Coleta

As notícias da imprensa digital e os julgados dos tribunais de justiça, do Supremo Tribunal Federal e do Superior Tribunal de Justiça serão obtidos via webscraping, isto é, por meio de técnicas de raspagem de dados. Os scripts para coleta serão convertidos em pacotes do **R**, os quais poderão ser acessados pela equipe de supervisão do projeto.

A propósito, já iniciamos a construção dos pacotes. O pacote para coleta e organização das notícias pode ser visualizado por meio [desta página](https://jjesusfilho.github.io/jornais). O pacote para baixar e organizar os dados do Tribunal de Justiça de São Paulo se encontra [nesta página](https://jjesusfilho.github.io/tjsp).

Os pacotes dos demais tribunais estaduais poderão acessados por meio da alteração do sufixo da url. Assim, bastará substituir tjsp por tjms, para acessar o código de coleta e organização dos dados do Tribunal de Justiça do Mato Grosso do Sul. Igualmente, os códigos do Supremo Tribunal Federal e do Superior Tribunal de Justiça já se encontram disponíveis, bastando substituir tjsp por stf e stj respectivamente.

Os dados coletados serão “parseados”, ou seja, lidos e estruturados em tabelas. Uma vez estruturados, esses dados serão armazenados em uma base SQL, preferivelmente em uma base do PostgreSQL. Os dados brutos, pdf e html, após “parseados” serão mantidos em uma armazenamento na núvel, Dropbox, Googledrive ou ou outro serviço.

## Transformação

Uma vez estruturados em tabelas, os dados passarão por transformação, a fim de se extrair informações relevantes para a pesquisa, as quais serão organizadas em novas tabelas e armazenadas na base SQL. Para as notícias jornalísticas, esseas tabelas conterão data do evento, título da notícia, assunto principal, nome do jornalista, unidade federativa, município, local do evento e outras informações que a equipe julgar pertinentes.

Quanto aos textos, esses serão submetidos ao processamento de linguagem natural, isto é, tokenização, stemming, exclusão de diacríticos e acentos, exclusão de “stopwords” e outras transformações, e convertidos em um index invertido ou índice remissivo. Tais transformações visam otimizar as buscas via motor de busca que será construído ao longo do projeto.

## Anotação

Numa fase posterior, a equipe irá criar “tags”, os quais utilizados para anotar os textos com a ajuda dos estagiários. Os textos anotados serão então submetidos a aprendizado de máquina (machine learning) e aprendizado profundo (deep learning). Possivelmente, utilizaremos LSTM para treinar partes dos textos e classificar novos textos com base nas anotações.

## Estatísticas descritivas

Uma vez transformados e classificados, os dados serão serão agregados em novas tabelas contendo frequências e outras estatísticas descritivas, as quais servirão para a produção de gráficos.

1. <https://inception-project.github.io/> [↑](#footnote-ref-21)