Projeto e questões de pesquisa
Transparência em Tempos de Pandemia

Gabriel Martins Trettel - N. USP 11389471
George Othon Silva Santos - N. USP 10349978
Tiago Lubiana Alves - N. USP 8945857
Wesley Seidel Carvalho - N. USP 6544342

São Paulo, 13 de Setembro de 2020

1 Projeto proposto

O projeto proposto pela *Open Knowledge Foundation Brasil* visa melhorar a visibilidade dos dados disponibilizados em diários oficiais, como parte da operação Querido Diário [1]. A proposta inicial foi colocada de forma aberta como a análise dos textos raspados de portais de diversos municípios brasileiros, buscando facilitar o acesso aos diversos tipos de informação nos textos.

Após conversas do grupo e com o cliente, afunilamos em uma proposta de projeto: encontrar compras suspeitas feitas durante a pandemias. Tal tarefa incluirá etapas de limpeza e normalização dos dados, integração com outras bases de dados, e construção de modelos para detectar anomalias.

2 Questões de pesquisa

Durante o período de pandemia, muitos contratos públicos têm sido feitos em caráter emergencial. Isso é importante para uma resposta rápida, mas abre espaço para irregularidades escaparem verificação.

Nos casos de cidades maiores, onde há mais fiscalização externa do poder público, foram encontrados diversos eventos suspeitos (ex: em Sergipe [2] e no Rio de Janeiro [3]). Contudo, para municípios menores, nem sempre os dados passam por escrutínio, abrindo margem para mal uso.

O objetivo de nosso trabalho é processar os textos obtidos de Diários Oficiais e identificar compras anômalas. Para tal, são necessárias etapas de pré-processamento do *corpus* de textos, etapas que compreendem também uma série de questões de pesquisa.

Em interação com o cliente, convergimos nas seguintes questões:

- A qual *classe* cada texto pertence? A pergunta envolve aplicar abordagens de classificação supervisionada e não-supervisionada para dividir os diários em classes. Exemplos de classes possíveis:
  - Declaração de emergência
  - Divulgação de leis locais
  - Decretos municipais
  - Contratação/licitação
- Quais compras foram feitas? A pergunta envolve extração de padrões do texto, associado a reconhecimento de entidades nomeadas, como:
  - Valor de compra

[1] Mais informações do projeto querido diário no GitHub do projeto
[2] http://www.pf.gov.br/imprensa/noticias/2020/07-noticias-de-julho-de-2020/operacao-serodio-apura-desvios-de-verbas-publicas-
[3] https://www.cnnbrasil.com.br/nacional/2020/05/13/mais-um-acusado-de-desviar-verba-da-saude-publica-e-preso-no-rj

  - CNPJ
  - Nome da empresa
  - Objeto (serviço ou bem) contratado
- Quais das compras feitas são consideradas em caráter emergencial?
- Há padrões de compradores que se repetem?
- Há empresas que ganharam licitações na pandemia que financiaram campanhas para as eleições 2020?

Para responder às questões acima, além do conjunto de dados disponibilizados inicialmente, estamos em contato com a equipe de colaboradores e desenvolvedores da *Open Knowledge Brasil*, para complementarmos os dados com metadados e anotações. Além disso, estamos estabelecendo pontos de contato para integração dos dados com a base de dados de sócios de empresas brasileiras (`https://brasil.io/dataset/socios-brasil/socios/`) e bases de dados do Tribunal Superior Eleitoral (`http://divulgacandcontas.tse.jus.br/divulga/`)

3 Planejamento do projeto

Durante todo o projeto, utilizaremos o GitHub e sua ferramenta de integração contínua, o GitHub Actions. O projeto aberto no GitHub será espelhado no GitLab, possibilitando acesso para os usuários de ambas as ferramentas.

A documentação será levada como um processo integral às etapas descritas adiante. Como o projeto é paralelo a outras abordagens da *Open Knowledge Brasil*, torna-se ainda mais importante a documentação em tempo real, para possibilitar a integração necessária para o sucesso do projeto.

3.1 Longo prazo

- Etapa 0:
  - Estudar os dados a partir de análises descritivas
  - Lapidar os entregáveis com o cliente
  - Escolher uma amostragem inicial para os arquivos.
- Etapa 1:
  - Pré-processamento dos textos amostrados (conversão em textos processáveis computacionalmente)
  - Segmentação dos textos de diários amostrados em categorias
  - Identificação de contextos associados com compras
- Etapa 2:
  - Enriquecimento semântico (reconhecimento de entidades nomeadas): Nomes de indivíduos; Nomes de empresas; CNPJ e Valores
  - Anotação manual de compras, valores, compradores e empresas associadas
  - Construção de modelos preliminares para extração automática das informações
- Etapa 3:
  - Aplicação das ferramentas das etapas anteriores aos demais dados de diários oficiais
  - Preparação dos dados obtidos para integração com base de dados externas
- Etapa 4:
  - Cruzamento com base de dados de sócios de empresa para identificar CPFs e CNPJs associados a licitações
  - Cruzamento com base de dados de financiamento de campanha
  - Criação de heurísticas para identificar compras anômalas

3.2 Curto prazo

Nos próximos 30 dias, focaremos em terminar a "Etapa 0"e executar a "Etapa 1"do planejamento. Ainda não é possível dimensionar o desafio da limpeza dos dados.

De início, amostraremos 12 textos do *corpus* escolhido. Implementaremos, então, um *framework* para converter os textos do formato bruto (PDF digitalizado) para textos que possamos tratar computacionalmente.

Em paralelo, iremos estudar as categorias adequadas para agrupar os dados oficiais, explorando abordagens supervisionadas (anotando manualmente textos) ou não supervisionadas, agrupando por similaridades.

Nessa etapa também ampliaremos nosso contato com a *Open Knowledge Brasil*. Temos reuniões marcadas com membros da equipe de desenvolvimento, com os quais interagiremos para buscar metadados em enriquecer o conjunto de dados.

Ao fim deste período devemos estar iniciando a etapa 2, com o enriquecimento semântico e anotação do *corpus*.

3.3 Entregáveis por data

3.3.1 Demonstração inicial do software/análise - descrição do código e algoritmos - 5/outubro

Até tal data, esperamos ter encaminhado a limpeza do *corpus* de exemplo, e ter montado e documentado o *framework* mencionado para conversão de um grupo de diários em um formato compatível com processamento de dados.

3.3.2 Planejamento dos ensaios e resultados preliminares - 26/outubro

Para entrega 4, com maior familiaridade com o corpus e com as ferramentas "base"implementadas, pretendemos estar encaminhados na fase 2 do projeto. Teremos, então, possibilidade de definir entregáveis específicos referentes as compras.

Esperamos ter desenvolvido uma parte do enriquecimento semântico também, tendo associado os textos a "Entidades Nomeadas". Isso por si já é mencionado como de interesse pelo cliente, pois tal dado auxilia o desenvolvimento de sistemas de busca mais inteligentes.

3.3.3 Texto jornalístico completo - 16/novembro

Para a entrega 5, pretendemos já termos organizados os dados sobre compras para ao menos alguns municípios. Já teremos exemplos específicos de compras em municípios identificadas pelas heurísticas iniciais para ilustrar o processo para o texto jornalístico.

3.3.4 Esqueleto de artigo científico sobre projeto - 30/novembro

A entrega 6 coincidirá com a etapa de integração as bases de dados externas. O artigo científico sobre o projeto compreenderá dos aspectos técnicos e observações científicas sobre cada uma das etapas de 1-4 e, sendo assim, será encaminhado ao decorrer do projeto.

3.3.5 Demonstração oral final + documentação - 14/dezembro

O período final do projeto, então, será dedicado a lapidar o artigo, a apresentação e a documentação, de forma a possibilitar que as análises feitas e os dados processados sejam reutilizáveis pela comunidade.