Meu Querido Diário

Andre Assumpcao

1 Introdução

Neste arquivo, eu computo o fluxo dos dados e da geração de produtos do projeto Meu Querido Diário.

2 Geração

Os raspadores serão executados na cloud do Scraping Hub e enviados ao nosso servidor da Digital Ocean. Em princípio, nós apenas baixaremos os arquivos mas não faremos nenhum tratamento do texto (ainda).

Nós mapeamos alguns dos problemas dos diários oficiais que podem dificultar a extração do texto ou a coleção de diários disponíveis:

1. Nem todos os municípios têm "diários" oficiais. Alguns têm boletins, outros semanários, outros publicam seus diários junto com outros municípios, etc.
2. Muitos diários têm duas ou mais colunas de informação. Em alguns casos, as assinaturas e headings são centralizados na página do diário.
3. Há muitos casos com tabelas e gráficos nos diários oficiais, o que exige ainda mais atenção no parseamento dos dados.
4. Em alguns casos, temos textos escritos à mão, como atas de reuniões, anexadas aos diários.
5. Há constante quebra de conteúdo, como a transição de uma para duas para três colunas.
6. Em muitos casos, temos uma série de assinaturas e títulos que acompanham algumas páginas do diário.
7. Alguns diários estão em Word, outros em JPEG, alguns em PDF.
8. Alguns caracteres de unicode são escapados pelos spiders.

3 Armazenamento

1. Infraestrutura HDFS no servidor da Digital Ocean. Conexão via Spark.
2. Dados NoSQL (json) com dados e metadados no mesmo lugar.
3.

4 Catalogação

5 Análise

6 Produtos