Algoritmos de relação de dados

Grupo

Gustavo Murayama

Lenin Cristi

Objetivo

Implementar algoritmos de análise de dados que consigam identificar relações em sequências de dados distintas que variam em função do tempo

Escopo

Temos um projeto de análise de dados tocado por alunos e gostaríamos de usar a oportunidade do curso para tocar uma parte desse projeto, mas como os dados dele dependem de um prazo externo, elegemos um estudo alternativo com dados que já temos à mão para o caso de não termos os primeiros dados a tempo.

Caso 1: Criar ferramentas que detectem a existência ou não de uma relação entre o índice pluviométrico e de temperatura com dados de notificações de casos de dengue no município de São Paulo. Conseguimos o contato com a prefeitura no sentido de recuperar estes dados, e já tivemos uma primeira reunião onde nos orientamos como conseguir dados do SUS via lei de acesso a informação, entramos em contato também com o CGE e recebemos dados históricos de pluviometria e temperatura, infelizmente com o contratempo deles terem sido enviados em formato que não conseguiremos trabalhar nos forçando a pedir novamente em formato específico. Não sabemos se estes dados chegarão a tempo de criar os algoritmos, apesar de ser pedido via lei de acesso existe um prazo legal de 20 dias para esta resposta.

Caso 2: No caso dois os conjuntos de dados utilizados serão os e-mails vazados da candidata à presidência dos Estados Unidos, Hillary Clinton. Pretendemos cruzar tais dados com a popularidade de Hillary, conjuntos de palavras das principais notícias e postagens de redes sociais no período, relacionando-os temporalmente para verificar se eles possuem influência direta na queda ou subida da intenção de votos da candidata em questão.

Componentes

Os componentes das duas alternativas de estudo são similares, inicialmente acreditamos que precisaremos de processos distintos, mas isso vai ficar mais claro durante o desenvolvimento:

Cabeças de leitura
Processos destinados a ler conjuntos de dados e carregá-los na memória, se necessário particionado em diferentes blocos

Sequenciadores
Processos destinados a linearizar dados em função do tempo

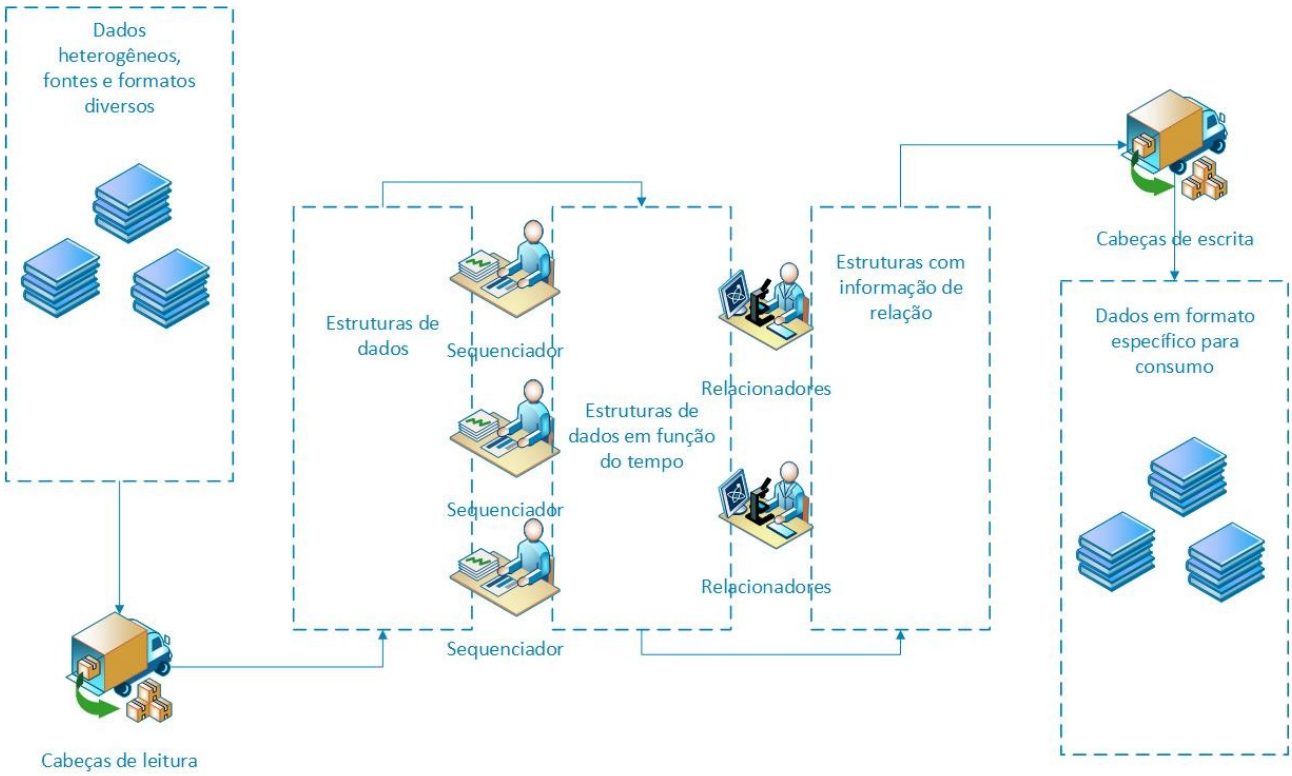
- Exemplo Caso 1: Uma função que distribui casos de notificações, indicadores pluviométricos e de temperatura por regiões da cidade em função do tempo;
- Exemplo Caso 2: Uma função que cria “tag clouds” com base num conjunto de documentos, notícias ou postagens e as distribui em função do tempo;

Relacionadores
Algoritmos que buscam uma relação entre duas funções numa determinada fatia de tempo

- Caso 1: Seria uma elevação do número de notificações numa região específica depois de um período determinado de dias após um período de precipitação;
- Caso 2: Um maior uso de palavras ou verbos nas redes sociais relacionados a determinadas notícias e sua “tag cloud” correspondente nos e-mails vazados;

Cabeças de escrita
Processos que descarregam os dados já computados em disco.

Diagrama



Tipos de dados

Os dados consistem em: planilhas do excel, conjunto de e-mails, dados sobre a variação de temperatura e pluviosidade em arquivos textos, informações retiradas nas redes sociais (twitter, instagram, facebook etc.) e textos-notícias retiradas de sites.

Aplicações similares

Análise de dados para retirada de uma inteligência, tais como: sugestão de filmes baseados na filmografia vista pelo usuário ou até mesmo o processamento de dados para estimar valor de algum produto no mercado (preço do barril de petróleo, por exemplo, utilizando dados como taxa mensal

de refino, preço do barril cru, preço da gasolina na bomba etc.).