Plan Overview

A Data Management Plan created using DMPTool

Title: Ciência de Dados: Análise de sentimentos

Creator: Tharsis Novais

Affiliation: Universidade de São Paulo (www5.usp.br)

Principal Investigator: Tharsis Novais

Data Manager: Tharsis Novais

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: 30498

Template: NSF-GEN: Generic

Last modified: 12-16-2017

Copyright information:

Ciência de Dados: Análise de sentimentos

Types of data produced

Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project.

Question not answered.

Data and metadata standards

Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies).

São dados de notícias do portal britânico The Guardian (https://www.theguardian.com) coletadas do ano de 2000 à 2010. As notícias foram recuperadas por meio de requisições para uma API aberta disponibilizada pelo próprio portal (http://open-platform.theguardian.com/access/).

As notícias recuperadas vem em formato HTML, por isso utilizamos um código em python para fazer a limpeza dos dados, para retirar as tags de html e ficar apenas com o título e conteúdo da notícia.

Os metadados persistidos serão basicamente data, título, autor e conteúdo da notícia, pois são as variáveis necessárias para realizar a análise.

O objetivo de armazenar estes dados é para que a partir de um código em python possam ser carregados para criar um modelo de análise de sentimentos onde seja possível indentificar o risco de uma empresa se tornar inadimplente, observando se o que as notícias estão dizendo sobre determinada empresa são boas ou ruins.

Policies for access and sharing

Policies for access and sharing; Provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements.

Os dados coletados são de propriedade do portal britânico The Guardian (https://www.theguardian.com), mas o acesso é público para consumo.

Não é permitido a alteração do conteúdo e atribuição ao The Guardian.

Todos os dados gerados por esta pesquisa podem também ser consumidos para implementação em outros trabalhos.

Policies for re-use, redistribution

Policies and provisions for re-use, re-distribution, and the production of derivatives.

Todos os dados gerados por esta pesquisa podem ser utilizados para fins de contribuir com outros trabalhos, desde que os autores originais sejam citados.

Plans for archiving and preservation

Plans for archiving data, samples, and other research products, and for preservation of access to them.

Os dados gerados serão armazenados em um repositório de dados científicos e disponibilizados com um DOI para validação.

Também será disponibilizado o código de um README para facilitar a repodução da pesquisa.
