

## Plan Overview

---

*A Data Management Plan created using DMPTool-Stage*

**Title:** Ciência de Dados: Análise de sentimentos

**Creator:** Tharsis Novais

**Affiliation:** Universidade de São Paulo ([www5.usp.br](http://www5.usp.br))

**Principal Investigator:** Tharsis Novais

**Data Manager:** Tharsis Novais

**Funder:** National Science Foundation ([nsf.gov](http://nsf.gov))

**Funding opportunity number:** 30498

**Template:** NSF-GEN: Generic

**Last modified:** 12-16-2017

**Copyright information:**

The above plan creator(s) have agreed that others may use as much of the text of this plan as they would like in their own plans, and customize it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

---

## Ciência de Dados: Análise de sentimentos

Question not answered.

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.

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.

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

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 reprodução da pesquisa.

---