Atividade 02

Introdução

Folha de S.Paulo, também conhecida como Folha de São Paulo ou simplesmente Folha, é um jornal brasileiro editado na cidade de São Paulo e é atualmente o segundo maior jornal do Brasil em circulação, com 366.087 exemplares (incluindo assinantes digitais), segundo o Instituto Verificador de Comunicação (IVC), em dezembro de 2021, ficando atrás apenas do carioca O Globo.

Fundada por um grupo de jornalistas liderado por Olival Costa e Pedro Cunha em 19 de fevereiro de 1921, a Folha foi criada em oposição ao principal jornal da cidade, O Estado de S. Paulo, que representava as elites rurais e assumia uma posição mais conservadora, tradicional e rígida. Em 1950, todas as Folhas passaram a ser impressas num prédio na Alameda Barão de Campinas, ampliado no final dos anos 1960 com a construção de um segundo prédio na alameda Barão de Limeira, no bairro dos Campos Elísios. Hoje, esse prédio é a entrada principal da empresa.

Em 1986, a Folha tornou-se o jornal de maior circulação em todo o país, liderança que manteve até 2021 (atualmente, o jornal de maior circulação no Brasil é O Globo, segundo dados do Instituto Verificador de Comunicação - IVC). Em 1995, um ano depois de ultrapassar a marca de 1 milhão de exemplares aos domingos, a Folha inaugurou seu novo parque gráfico, considerado o maior e mais atualizado tecnologicamente na América Latina. O recorde de tiragem e de vendas do jornal foi alcançado em 1994, na época do lançamento do Atlas Folha/The New York Times (1.117.802 exemplares no domingo).

Base de Dados

Fonte:
https://www.kaggle.com/datasets/marlesson/news-of-the-site-folhauol

Notícias do Jornal Brasileiro
167.053 notícias do site Folha de São Paulo

O conjunto de dados consiste em 167.053 exemplos e contém títulos, URL do artigo, artigo completo e categoria.

Foram reunidas as notícias resumidas e apenas coletadas as reportagens da Folha de São Paulo:
http://www.folha.uol.com.br/

O intervalo de tempo é entre janeiro de 2015 e setembro de 2017.

Repositório do WebScrapy do autor da base de dados:
https://github.com/marlesson/scrapy_folha

Etapas da Atividade

- (03/05/2024) Realizar o procedimento do notebook:
  - https://www.kaggle.com/code/marlesson/vocabulary-analysis-word2vec
- (10/05/2024) Adaptar o notebook:
  - https://www.kaggle.com/code/madz2000/sarcasm-detection-with-glove-word2vec-83-accuracy
  - para classificação das categorias via a tokenização das notícias
- (17/05/2024) Utilizando um modelo do Huggingfaces de Text Classification:
  - https://huggingface.co/models?pipeline_tag=text-classification
  - Obter um resultado melhor que o notebook da etapa 2 (GloVe)

Entrega da Atividade

- Formato de Entrega:
  - Jupiter Notebook
  - Arquivo ".IPYNB"
  - Qualquer outro formato anulará a nota
- Sistema de entrega:
  - Tarefa SIGAA
- Cuidados:
  - Não permite envio com atrasos
  - Pode substituir o arquivo quando quiser