UNIVERSIDADE FEDERAL DO PIAUÍ
CAMPUS SENADOR HELVÍDIO NUNES DE BARROS
DISCIPLINA: SISTEMAS INTELIGENTES
PROFESSORA: DEBORAH MAGALHÃES
MONITORA: ORRANA LHAYNHER

TRABALHO FINAL

1. Descrição do Desafio

O Twitter se tornou um importante canal de comunicação em tempos de emergência. O uso de smartphones e a sua presença em todas as situações do cotidiano permitem que as pessoas anunciem uma emergência que estão observando em tempo real. Por causa disso, mais agências estão interessadas em monitorar programaticamente o Twitter (ou seja, organizações de ajuda a desastres e agências de notícias).

Entretanto, nem sempre é claro se as palavras de uma pessoa estão realmente anunciando um desastre ou não. Algumas expressões, como “o programa hoje estava pegando fogo”, tem seu sentido claro para um humano imediatamente, especialmente com o auxílio visual. Mas é menos claro para uma máquina. Neste trabalho, deverá ser construído um modelo de aprendizado de máquina que preveja quais tweets são sobre desastres reais e quais não são. O dataset contém dados de mais de 10,000 tweets classificados manualmente e foi retirado do link https://www.kaggle.com/competitions/nlp-getting-started/overview.

Descrição dos Arquivos:

- **dataset_treino.csv** - dados de tweets para treino de modelos, incluindo o rótulo de se é um tweet sobre desastre real ou não.
- **dataset_teste.csv** - dados de tweets para testes, excluindo os rótulos.
- **sample_submission** - contém o template para envio dos rótulos preditos.

Formato da entrada:

Os dados de treinamento apresentam o seguinte formato:

1. **id** - um identificador exclusivo para cada tweet

2. **palavra-chave** - uma palavra-chave específica do tweet (pode estar em branco)
3. **location** - o local de onde o tweet foi enviado (pode estar em branco)
4. **texto** - o texto do tweet
5. **target** - apenas em train.csv, isso denota se um tweet é sobre um desastre real (1) ou não (0)

```
1,,,Our Deeds are the Reason of this #earthquake May ALLAH Forgive us all,1

4,,,Forest fire near La Ronge Sask. Canada,1

5,,,All residents asked to 'shelter in place' are being notified by officers. No other evacuation or shelter in place orders are expected,1

6,,,"13,000 people receive #wildfires evacuation orders in California ",1

444,apocalypse,Tokyo,Enjoyed live-action Attack on Titan but every time I see posters I'm reminded how freshly clean and coiffed everyone is in the apocalypse.,0

445,apocalypse,,I liked a @YouTube video http://t.co/ki1yKrs9fi Minecraft: NIGHT LUCKY BLOCK MOD (BOB APOCALYPSE WITHER 2.0 &amp;amp; MORE!) Mod Showcase,0

446,armageddon,"California, United States",#PBBan (Temporary:300) avYsss @'aRmageddon | DO NOT KILL | FLAGS ONLY | Fast XP' for Reason,0
```

Formato da saída:

Para cada amostra, você deve produzir uma saída contendo o identificador e a decisão tomada pelo seu classificador, conforme o exemplo abaixo:

```
1, 1
4, 1
5, 1
6, 1
444, 0
445, 0
446, 0
```

Pontuação

A métrica F1-score será utilizada para avaliar os resultados. Essa métrica corresponde a média harmônica entre precisão (a capacidade do classificador de não rotular

como positiva uma amostra que é negativa) e sensibilidade (a habilidade do classificador em encontrar todas as amostras positivas). O valor mais alto possível do F1-score é 1, indicando precisão e sensibilidade perfeitos, e o valor mais baixo possível é 0, se a precisão ou sensibilidade for zero. O pacote utilizado para o cálculo da métrica F1-score será o sklearn (https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html).

Onde está o dataset?
O dataset está disponível no link: https://www.kaggle.com/competitions/nlp-getting-started/data

1.1. Quais os requisitos da solução esperada?

- Você deve realizar o pré-processamento do texto;
- Você deve realizar a vetorização do texto;
- Você deve escolher um modelo de classificação que corresponda a uma rede neural;
- Você deve utilizar um otimizador para ajustar os parâmetros do modelo de classificação;
- Você deve computar a taxa de acerto conforme especificado no tópico pontuação.

2. Instruções da Entrega

Essa atividade corresponde a nota da terceira avaliação da disciplina, portanto 0-10. Ela deverá ser realizada em equipe e entregue até o **dia 12/05**.

Os seguintes critérios serão considerados na avaliação:

1. Atender ao que foi pedido na descrição deste documento;
2. Código está executando sem erros;
3. Enviar **o link do vídeo no Sigaa** explicando a solução dada, é importante que todos os membros participem (1 por grupo);
4. Enviar o código (jupyter notebook) da solução de classificação juntamente com o arquivo **sample_submission** contendo as labels preditas pelo classificador para o conjunto de teste (1 por grupo);