Nome: _____Matrícula:_____
Data:__/__/____

PROJETO PRÁTICO DE MINERAÇÃO DE DADOS
CLASSIFICAÇÃO

Projeto prático a ser entregue na disciplina Inteligência de Negócios II – 2° Bimestre
Valor: 15 pontos

Objetivo do trabalho:

- Dado um link(URL) de uma determinada notícia da internet, o algoritmo deverá ser capaz de classificar a notícia em uma das seguintes categorias:
  - Agricultura
  - Saúde
  - Economia
  - Educação
  - Política
  - Tecnologia
- A classificação deverá ser feita com base nas **palavras** constantes na notícia.
- O algoritmo deverá ser capaz de processar as **palavras** constantes na notícia e estimar com base em seu aprendizado em qual categoria a notícia melhor se encaixa.

Os dados para treino e testes encontram-se no arquivo CSV: ***urls.csv***

link	**categ**
https://www1.folha.uol.com.br/ilustrada/2019/11/pec-pode-extinguir-fundos-publicos-de-patrocinio-cultural.shtml	politica

ATIVIDADES

1) EXTRAÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

- Ler os dados do CSV e armazenar em um *dataframe* do Pandas.

WEB SCRAPING / WEB CRAWLER

- O programa deverá ler cada URL do *dataframe*, acessar a referida página, capturar a matéria da página e armazená-la em uma nova coluna do *dataframe*.
- O texto da matéria deverá ser armazenado em uma string contínua sem caracteres especiais(quebra de linha, aspas, interrogação, html, js, ...). Para que o algoritmo tenha uma boa eficiência na análise das palavras do texto, deverão ser excluídas as chamadas **STOP WORDS(Palavras irrelevantes na análise de um texto "o, a, de, das, dos, para, com ...")*

PEC dos Fundos Públicos ameaça patrocínio à arte e preocupa setor

Proposta, que ainda precisa ser votada, encaminharia os recursos para pagamento da dívida pública

5

Gustavo Fioratti

SÃO PAULO Nomeado para o comando da Secretaria Especial de Cultura, Roberto Alvim encontrará pela frente um desafio. Ele pode perder dois dos recursos de subsídio mais importantes da subpasta, o Fundo Nacional de Cultura —R$ 1,4 bilhão em 2019—, e o Fundo Setorial do Audiovisual —R$ 724 milhões.

Como parte do Mais Brasil, pacote de medidas do governo lançado na terça-feira (5) visando
cortes de despesas e a flexibilização de orçamentos, a PEC dos Fundos Públicos está causando preocupação entre produtores culturais e artistas.

Exemplo de *String* a ser extraída:

nomeado comando secretaria especial cultura roberto alvim encontrará frente desafio pode perder dois recursos subsídio importantes subpasta fundo nacional cultura bilhão fundo setorial audiovisual milhões parte brasil pacote medidas governo lançado terçafeira visando cortes despesas flexibilização orçamentos pec fundos públicos causando preocupação produtores culturais artistas

Carregamento da *String* em uma nova coluna do dataframe:

link	**categ**	**texto**
https://www1.folha.uol.com.br/ilustrada/2019/11/pec-pode-extinguir-fundos-publicos-de-patrocinio-cultural.shtml	politica	*nomeado comando secretaria especial cultura roberto alvim encontrará frente desafio pode perder dois recursos subsídio importantes subpasta fundo nacional cultura bilhão fundo setorial audiovisual milhões parte brasil pacote medidas governo lançado terçafeira visando cortes despesas flexibilização orçamentos pec fundos públicos causando preocupação produtores culturais artistas*

***Bag of Words* – Processamento de linguagem natural (PLN)**

Os algoritmos de inteligência artificial para análise de textos analisam números para treinamento e aprendizado. Assim, uma técnica de conversão de textos em números se faz necessária. A técnica de Bag of Words traduz um texto em uma lista de palavras não-repetidas e uma outra lista com a frequência de cada palavra no texto. O resultado da *string* acima seria:

Exemplo:

['alvim', 'artistas', 'audiovisual', 'bilhão', 'brasil', 'causando', 'comando', 'cortes', 'cultura', 'culturais', 'desafio', 'despesas', 'dois', 'encontrará', 'especial', 'flexibilização', 'frente', 'fundo', 'fundos', 'governo', 'importantes', 'lançado', 'medidas', 'milhões', 'nacional', 'nomeado', 'orçamentos', 'pacote', 'parte', 'pec', 'perder', 'pode', 'preocupação', 'produtores', 'públicos', 'recursos', 'roberto', 'secretaria', 'setorial', 'subpasta', 'subsídio', 'terçafeira', 'visando']

[[1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]]

→ **Veja que as palavras “cultura” e “fundo” apareceram 2 vezes no texto.**

Os algoritmos de aprendizado da biblioteca SKLearn deverão receber os vetores de palavras para aprendizado.

Dicas:

- Sugestão de bibliotecas Python para a etapa de: “Extração e Pré-Processamento dos Dados”:
  - ***Web Scraping***: Goose3 https://pypi.org/project/goose3/3.1.6/
  - Limpeza do texto:
    - Retirar caracteres especiais, quebra de linha: Unicodedata, normalize
    - **StopWords**: NLTK module - from nltk.corpus import stopwords
  - ****gere um novo CSV com a coluna “texto” preenchida, assim ao rodar o programa novamente não será necessário capturar os textos novamente.***
  - ***Bag of Words:*** CountVectorizer da biblioteca SKLearn
  - **Treino e teste do algoritmo com Bag of Words**: Utilize a classe ***Pipeline*** do SKLearn para facilitar o processo de conversão dos dados X de treino e teste.

2) TREINAMENTO E PREVISÃO

- Utilize a biblioteca ***SKLearn*** para treinar e testar as classificações
  - Para a divisão das bases de treino e teste utilize a função ***train_test_split***
  - Utilize 25% da massa de dados para testes.
  - Utilize um *random_test* padrão com SEED = 20.
  - Estratifique a porcentagem de Y para as amostras de treino e teste.
- Dentre os algoritmos abaixo, verifique qual algoritmo de classificação apresentará melhor taxa de acerto.

3) PREVISÃO PARA NOTÍCIAS DO USUÁRIO

- Implemente uma maneira do usuário passar uma URL de uma notícia e o programa estimar a categoria da notícia.

No exemplo abaixo, o algoritmo apresentou 83% de acerto para os testes e para a URL de exemplo do exercício, ele sugeriu ser uma notícia da área de **“Economia”**.

```
 python3 -W ignore exercicio_04_1.py -p https://www1.folha.uol.com.br/ilustrada/2019/11
/pec-pode-extinguir-fundos-publicos-de-patrocinio-cultural.shtml
0.8342857142857143
['economia']
```

```
from sklearn.dummy import DummyClassifier
from sklearn.svm import LinearSVC
from sklearn.svm import SVC
from sklearn.ensemble import RandomForestClassifier
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import MultinomialNB
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
```