DESAFIO - PARTE 1 (ENTREGA INDIVIDUAL)

Módulo Ibiapaba - Ciência de Dados

Prof. Wellington Franco

VÍDEO EXPLICATIVO

- A solução para as questões abaixo deve ser feita no formato .ipynb;
- A solução para as questões abaixo deve ser enviado até no dia 22/11 para o seguinte e-mail: wellington@crateus.ufc.br. Utilize o seguinte texto no assunto do e-mail "mandacaru.dev - Desafio parte 1";
- As **questões 1 à 5** utilizarão os textos disponibilizados na biblioteca NLTK;
- Link da pasta com dados de apoio para a **questão 6**;
- Link da pasta com dados de apoio para as **questões 7 e 8**.

1) Calcule os seguintes itens:
    a) Número total de palavras
    a) Número total de sentenças
    b) Número total de palavras não repetidas
    c) Número total de palavras repetidas
    d) Média de palavras por sentenças

Esse procedimento deve ser feito para os seguintes textos disponíveis na biblioteca NLTK:

- shakespeare-caesar.txt,
- shakespeare-hamlet.txt,
- shakespeare-macbeth.txt

2) Em relação ao corpus “gutenberg” implemente os algoritmos para responder às seguintes questões:
    a) Total de palavras em cada documento do corpus "gutenberg"
    b) Quem é o maior documento do corpus?

c) Quem é o menor documento do corpus?
d) Calcular a média da quantidade sentenças por palavras do corpus “gutenberg”.
e) Calcule a distribuição de frequência das palavras do livro "shakespeare-macbeth.txt".
f) Calcule 5 palavras mais frequentes nesse corpus.
g) Mostre a diferença entre de palavras entre dois livros. (shakespeare-caesar.txt, shakespeare-hamlet.txt,)

3) Para o corpus “shakespeare-caesar.txt”, usando expressões regulares responda os seguintes itens:
   a) Quantidades de palavras que terminam com “r”;
   b) Quantidade de palavras com 5 letras;
   c) Quantidade de vezes que “err” ocorre no corpus;
   d) Quantidade de vezes que “are” ocorre no corpus para as palavras com 5 ou mais caracteres.

4) Em relação do corpus “shakespeare-hamlet.txt”, faça as seguintes atividades:
   a) Normalize o corpus (Retire os números e deixe todas as palavras minúsculas);
   b) Aplique o lematizador em todas as palavras do corpus;
   c) Aplique o tokenizador em todas as sentenças do corpus;
   d) Aplique os pos tagger e responda a quantidade de adjetivos existem no corpus; Dica a tag de adjetivo é **“JJ”**

5) Retire as stopwords dos seguintes textos:
   a) shakespeare-caesar.txt
   b) shakespeare-hamlet.txt
   c) Qual é a quantidade de palavras restantes em cada texto?

6) Encontre as entidades nomeada presentes para o seguintes textos: apoloxi.txt e french-revolution.txt
   a) Qual é a quantidade de entidades "GPE" presentes em cada um dos texto?
   b) Qual é a quantidade de entidades "LOCATION" presentes em cada um dos texto?

c) Qual é a quantidade de entidades "PERSON" presentes em cada um dos texto?

7) Utililizando as técnicas aprendidas sobre análise de sentimentos defina a polaridade do arquivo ‘reviews”. Utiliza os arquivos positive_words.txt e negative_words.txt presentes na pasta.

8) Utililizando as técnicas aprendidas sobre ***Term Frequency, Inverse Document Frequency***, liste as 5 palavras mais relevantes de cada texto contido no corpus Gutemberg. A sua analise deve ser em cima dos seguintes textos : 'austen-emma.txt', 'bible-kjv.txt', 'carroll-alice.txt', 'melville-moby_dick.txt', 'shakespeare-caesar.txt' e 'shakespeare-hamlet.txt'. Lembre-se que antes de aplicar o as técnicas de TF IDF o texto deve está normalizado, com as stopwords retiradas e lematizado.