

UNIVERSIDADE DE BRASÍLIA

FACULDADE DO GAMA

CURSO:	ENGENHARIAS		
DISCIPLINA:	Estruturas de Dados e Algoritmos	CÓDIGO:	193704
CARGA HORÁRIA:	60 h	CRÉDITOS:	04
PROFESSOR:	Dr. Nilton Correia da Silva / Dr. Fabricio Ataide Braz		

TRABALHO PRÁTICO 01

TEMA: VETORES DINÂMICOS

O Trip Advisor Hotel (<https://www.tripadvisor.com.br/>) é um site de vendas de pacotes de viagens que também coleta avaliações de hotéis a partir das experiências de seus clientes. Parte desta informação foi postada no Kaggle (www.kaggle.com) - um site de competições de Aprendizado de Máquina.

O dataset do Trip Advisor foi estruturado num arquivo texto (CSV) com duas colunas: *Review*: breve texto contendo a avaliação descritiva do cliente e *Rating*: uma nota contendo um valor inteiro: 1,2,3,4 ou 5 (<https://www.kaggle.com/andrewmvd/trip-advisor-hotel-reviews>).

O objetivo deste trabalho é avaliar o teor textual das opiniões escritas pelos clientes da Trip Advisor (coluna *Review*) para cada um dos 5 tipos de notas atribuídas (coluna *Rating*). Para tanto, usaremos um descritor estatístico chamado TF-IDF.

*O valor **tf-idf** (abreviação do inglês term frequency-inverse document frequency, que significa frequência do termo-inverso da frequência nos documentos), é uma medida estatística que tem o intuito de indicar a importância de uma palavra de um documento em relação a uma coleção de documentos ou em um corpus linguístico. Ela é frequentemente utilizada como fator de ponderação na recuperação de informações e na mineração de dados.*

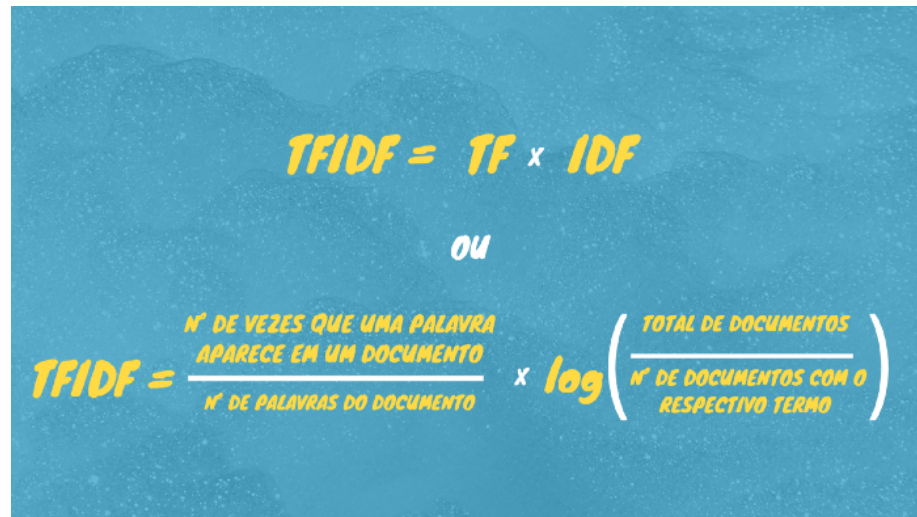
*O valor **tf-idf** de uma palavra aumenta proporcionalmente à medida que aumenta o número de ocorrências dela em um documento, no entanto, esse valor é equilibrado pela frequência da palavra no corpus. Isso auxilia a distinguir o fato de a ocorrência de algumas palavras serem geralmente mais comuns que outras.* (<https://pt.wikipedia.org/wiki/Tf%E2%80%93idf>)

Faça um programa para avaliar as características nos textos (*Review*) das notas (*Rating*) 1,2,3,4 e 5. Para tanto, deve-se calcular o TF-IDF de cada uma das 5 notas. Dado o dataset do TripAdvisor que está em um arquivo CSV, você deve seguir os seguintes primeiros passos:

1. Aglutinar todos os textos por nota. Desta forma você vai gerar um **corpus textual** contendo os seguintes 5 arquivos:
 - a. Nota1.txt: Concatenação de todos os textos de nota 1 (coluna *Review* igual a 1)
 - b. Nota2.txt: Concatenação de todos os textos de nota 2 (coluna *Review* igual a 2)
 - c. Nota3.txt: Concatenação de todos os textos de nota 3 (coluna *Review* igual a 3)
 - d. Nota4.txt: Concatenação de todos os textos de nota 4 (coluna *Review* igual a 4)
 - e. Nota5.txt: Concatenação de todos os textos de nota 5 (coluna *Review* igual a 5)

EDA – TRABALHO PRÁTICO - TEMA: VETORES DINÂMICOS

2. Gerar vocabulário: vetor com as diferentes palavras que ocorrem no **corpus textual**. *Sugestão: elimine palavras com menos de 3 caracteres. Elimine pontuações.*
3. Calcular o vetor de IDF para cada termo do vocabulário. Obs: *O valor idf de um termo não varia conforme o documento.*
4. Gerar os 5 vetores de TF-IDF (conforme os 5 documentos listados no item 1).


$$TFIDF = TF \times IDF$$

OU

$$TFIDF = \frac{\text{Nº DE VEZES QUE UMA PALAVRA APARECE EM UM DOCUMENTO}}{\text{Nº DE PALAVRAS DO DOCUMENTO}} \times \log\left(\frac{\text{TOTAL DE DOCUMENTOS}}{\text{Nº DE DOCUMENTOS COM O RESPECTIVO TERMO}}\right)$$

Figura 1. TF-IDF. Fonte: <https://medium.com/turing-talks/introdu%C3%A7%C3%A3o-a-bag-of-words-e-tf-idf-43a128151ce9>

- a. Veja detalhes do cálculo de TF-IDF em:
 - <https://towardsdatascience.com/tf-idf-for-document-ranking-from-scratch-in-python-on-real-world-dataset-796d339a4089>
 - http://www.facom.ufu.br/~wendelmelo/ori201802/4_ponderacao_de_termos.pdf

Seu programa deve possuir um menu com as seguintes opções mínimas:

- Ler o dataset do Trip Advisor. Usuário deve informar o nome do arquivo.
 - Os 5 arquivos do corpus textual devem ser gerados neste momento (item 1 acima).
- Gerar vocabulário.
- Exibir TF-IDFs. Vide formato na Tabela 1.

Vocabulário	Nota 1	Nota 2	Nota 3	Nota 4	Nota 5
Bom	0.9	0.01	0.5	2.8	0.77
Casa	2.8	15.8	3.4	1.5	0.45
Desleixado	5.9	1.8	3.7	6.8	4.2

Tabela 1. Exemplo de visualização dos 5 vetores de TF-IDFs

- Exibir TF-IDF de uma Nota. Em ordem decrescente de valor de TF-IDF. Usuário deve informar qual TF-IDF a ser visualizado (da nota 1,2,3,4 ou 5). Vide formato na Tabela 2.
- Sair
 - Não esqueça de desalocar os vetores todos os vetores/matrizes que foram alocados.

EDA – TRABALHO PRÁTICO - TEMA: VETORES DINÂMICOS

Vocabulário	Nota 1
Desleixado	5.9
Casa	2.8
Bom	0.9

Tabela 2. Exemplo de visualização o TF-IDF para Nota 1.

Condições de contorno:

1. Você deve implementar as funcionalidades específicas em bibliotecas (crie os arquivos header (.h) e de código (.c)).

Ótimo trabalho!