Recomendando Arquivos

tempo limite de execução: `1s`

Neste problema N arquivos de textos foram coletados ao longo do tempo e, para cada arquivo, técnicas de remover *stop words* e de *stemming* foram aplicadas.

Dado um arquivo de texto, $N+1$ (arquivo de consulta), seu programa deverá apresentar, em ordem decrescente, a similaridade de todos os N arquivos com o arquivo de consulta. Para isso seu programa deverá:

1 - ler todo o conteúdo dos N arquivos pré-processados e aplicar a técnica de *tfidf* para fazer a representação veterial de cada um deles. Para o cálculo do *tfidf* levar em consideração os conteúdos dos N arquivos;

2 - em seguida o programa deverá ler o conteúdo do arquivo $N+1$ e montar a representação vetorial do mesmo. Para o cálculo do *tfidf* levar em consideração o conteúdo deste único arquivo, perceba que, neste caso, o *idf* terá valor igual a 1 e, portanto, o valor de *tfidf* se resume no cálculo de *tf* (verifique você mesmo);

3 - por útlimo, o programa deverá determinar a similaridade dos N arquivos com o arquivo $N+1$, usando cosseno como métrica:

$$sim_{ij}(\mathbf{v}_i, \mathbf{v}_j) = \frac{\sum_{t=1}^{T} v_{it} \times v_{jt}}{\sqrt{\sum_{t=1}^{T} (v_{it})^2} \times \sqrt{\sum_{t=1}^{T} (v_{jt})^2}}, \quad (1)$$

onde sim_{ij} é a similaridade existente entre os documentos i e j, cuja representação vetorial é dada por $\mathbf{v}_i$ e $\mathbf{v_j}$, respectivamente. Cada termo v_{it} e v_{jt} corresponde ao valor *tfidf* do termo t sobre os documentos i e j, respectivamente.

Entrada:

Entrada:

A entrada é composta por um conjunto de linhas onde:

- a primeira linha contém um inteiro $N+1$, onde $1 \leq N \leq 1000$, indicando o número de documentos de texto;
- a segunda linha começa com o conteúdo do primeiro documento coletado ao longo de um tempo. O conteúdo de um documento, neste exercício, é simplificado por caracteres. Cada caractere, por sua vez, representa um palavra pré-processada e pode ser qualquer caractere no intervalo $[A-Z]$, conforme a tabela ASCII. Um documento possui $1 \leq nC \leq 50$ caracteres, $1 \leq nL \leq 100$ linhas e é finalizado com o inteiro 0 (zero);
- O último documento da entrada corresponde ao arquivo de texto $N+1$;
- Os documentos aparecem conforme coletados ao longo do tempo.

Saída:

A saída deverá imprimir em ordem decrescente os N documentos e suas respectivas similaridade com o documento $N+1$, conforme ilutra o exemplo abaixo. O valor de similaridade deverá ser calculado e computado considerando apenas as 4 casas decimais após a vírgula e deverá ser apresentado como um grau de porcentagem. Se houverem documentos com sim-

ilaridades iguais, então o documento coletado por último, ao longo do tempo, deve aparecer antes.

Entrada:	Saída:
5 A R T Y U H N S D F D V N M K E R T G V B 0 A B C N F D J K L K H N A B N M G 0 A B C N F D J K L K K J A H N A B N M G 0 A R T Y U H N S D F D V N M K E R T G V B 0 A R T Y U H N S D F D V N M K E R T G V B W 0	D4:93.47 D1:93.47 D2:51.61 D3:45.64

Considerações práticas:

- Pode ser que um ou mais termos presentes no documento $N+1$ não estejam presentes nos demais N documentos. Neste caso, deve-se tomar um cuidado para não determinar a similaridade sobre vetores cujo elementos e/ou dimensões não sejam equivalentes. Por exemplo: considere que os termos t_1, t_2, t_3 aparecem nos documentos 1 e 2, permitindo as seguintes representações:
  - $\mathbf{v}_1 :< v_{11}, v_{12}, v_{13} >$
  - $\mathbf{v}_2 :< v_{21}, v_{22}, v_{23} >$

  Já o documento 3 apresenta como termos t_1, t_4, t_5 e, consequentemente, ao processar o $tfidf$ obtém-se:
  - $\mathbf{v}_3 :< v_{31}, v_{34}, v_{35} >$

  Embora a dimensão dos vetores $\mathbf{v}_1, \mathbf{v}_2$ e $\mathbf{v}_3$ sejam as iguais, seus elementos não são correspondentes e, portanto, o cálculo da similaridade resultaria em um valor inválido. O correto seria efetuar o cálculo da similaridade levando em consideração os seguintes vetores:

- $\mathbf{v}_1 :< v_{11}, v_{12}, v_{13}, 0, 0 >$
- $\mathbf{v}_2 :< v_{21}, v_{22}, v_{23}, 0, 0 >$
- $\mathbf{v}_3 :< v_{31}, 0, 0, v_{34}, v_{35} >$