Similarité cosinus

1 - Définir une liste "docs" qui contient les trois strings
"The sky is blue"
"The sun is dark dark red"
"The moon is gray"

2 - Définir une fonction "unique_words" qui prend comme argument une list de strings. Cette fonction doit retourner un dict où chaque mot est associé à un numéro séquentiel commençant 0

3 - Définir une fonction "term_freq" qui prend comme argument une list de strings et retourne une matrice n x m d'où n est le nombre de strings et m est le nombre de mots uniques. Utilisez "unique_words" pour obtenir la position de chaque mot dans la matrice. Chaque element de la matrice correspond au nombre de fois où chaque mot est dans chaque chaîne.

4 - Définir une fonction "inv_doc_freq" qui prend la matrice renvoyée par term_freq et calcule, pour chaque mot

$$\log\left[\frac{n\text{docs}}{\text{word\_docs}+1}\right]$$

où ndocs est le nombre de strings et word_docs le nombre de strings avec ce mot.

5 - Définir une fonction "norm_rows" qui divise chaque ligne d'une matrice par la norme de cette ligne

6- Utilisez toutes les fonctions définies ci-dessus pour calculer la matrice tf*idf. Cette matrice est définie comme la matrice normalisée obtenue par le produit scalaire de la matrice calculé avec "term-freq" et le vecteur calculé par "inv-doc_freq" représenté comme une matrice diagonale.

7- Chaque ligne de cette matrice représente l'une des strings d'origine. Nous pouvons maintenant mesurer la similarité de chaque paire de strings est ~~pas~~ de calculer le cosinus de les deux vecteurs lignes correspondantes. Comme les lignes sont déjà normalisés, ce n'est que le produit scalaire de chaque paire de vecteurs. Calculer les similitudes pour les paires suivantes:

0, 1

0, 2

1, 2