ÉCOLE NATIONALE SUPÉRIEURE D'ARTS ET MÉTIERS
UNIVERSITÉ HASSAN II DE CASABLANCA

Année Universitaire: 2022-2023

TD N° 4 Machine Learning & Text Mining
Techniques de représentation des données textuelles et applications

Considérons le programme suivant :

```
1. from sklearn.feature_extraction.text import CountVectorizer
2. texte = ["La vie est douce","La vie est tranquille, est belle, est douce"]
3. vect = CountVectorizer()
4. T= vect.fit_transform(texte)
5. dictionnaire_des_mots=vect.vocabulary_
6. print("dictionnaire_des_mots :", dictionnaire_des_mots)
7. liste_des_mots=list(dictionnaire_des_mots.keys())
8. print("liste_des_mots :", liste_des_mots)
9. Matrice_sparse_correspondante=T.toarray()
10. print("Matrice_sparse_correspondante:\n",Matrice_sparse_correspondante)
```

Travail à faire :

1. Exécuter le programme précédent
2. Identifier le rôle de la fonction

   fit_transform() : ……
   ……
3. Identifier le rôle de l'attribut

   vocabulary_: …… …
   …….
4. Identifier le rôle de la fonction ***keys***

   () : ……
   …
5. Quel est le résultat final de ce programme : (changer le texte en cas de besoin) ……
   ……
   ……
6. Appeler la fonction CountVectorizer avec l'argument **(binary=True)** dans la ligne 3, analyser le résultat du programme et identifier le rôle du paramètre binary