Rapport TP3

Abdou NIANG

Expliquez les différentes options du programme word2vec ?

```
./word2vec -train text8 -output vectors400.bin -cbow 1 -size 400 -window 8 -negative
25 -hs 0 -sample 1e-4 -threads 20 -binary 1 -iter 15
```

-train: fichier texte qui contient les données d'apprentissage

-output: fichier dans lequel les mots formés seront enregistrés

-cbow: est utilisé pour spécifier s'il faut utiliser le modèle de sac de mots continu

-size: dimension des mots embedding

-window: permet de déterminer le contexte des mots pour chaque mot cible

-negative: indique le nombre d'échantillons négatifs utilisés pour entraîner le modèle.

-hs: indique s'il faut utiliser le softmax hiérarchique au lieu de l'échantillonnage négatif par défaut.

-sample: Le seuil de sous-échantillonnage utilisé pour réduire l'impact des mots fréquents

-threads: Il est utilisé pour spécifier le nombre de threads qui doivent être utilisés pendant la formation

-binary: indique si la sortie doit être sauvegardée au format binaire.

-iter: Il est utilisé pour spécifier le nombre d'itérations sur les données d'entrée qui doivent être utilisées pendant l'entraînement.

Quetion : Même question, quels sont les mots les plus proches des mots suivants : apple, mouse, macron? Que constatez-vous ?

```
Entrée [8]: model.most_similar(["apple","mousse","macron"])

    Out[8]: [('pancakes', 0.44406259059906006),
             ('garnish', 0.4310971796512604),
             ('diaeresis', 0.4090222418308258),
             ('candied', 0.40708065032958984),
             ('puree', 0.40515878796577454),
             ('pentyl', 0.40278443694114685),
             ('cedilla', 0.4003411531448364),
             ('desserts', 0.4003259539604187),
             ('savoury', 0.39990460872650146),
             ('tamarind', 0.3925629258155823)]
```

On voit nettement il y’a une similarité entre ces mots.