TP3 : word embeddings

Question : 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
```

Réponse :

-train text8 : Spécifie le fichier d'entraînement, dans ce cas, le corpus text8.

-output vectors400.bin : Spécifie le nom du fichier de sortie où les vecteurs word2vec appris seront sauvegardés. Dans cet exemple, les vecteurs sont sauvegardés dans un fichier binaire nommé "vectors400.bin".

-cbow 1 : Indique que l'algorithme de Word2Vec utilisera la méthode Skip-gram pour l'apprentissage (Continuous Bag of Words). Si la valeur est 0, cela signifie que l'algorithme utilisera l'approche CBOW.

-size 400 : Spécifie la dimension des vecteurs d'embedding. Dans cet exemple, chaque mot sera représenté par un vecteur de 400 dimensions.

-window 8 : Définit la taille de la fenêtre contextuelle lors de l'apprentissage. Cela signifie que le modèle prendra en compte les 8 mots environnants pour prédire le mot cible.

-negative 25 : Spécifie le nombre de mots négatifs à échantillonner lors de l'apprentissage. Les exemples négatifs sont utilisés pour améliorer la qualité des représentations apprises.

-hs 0 : Désactive la hiérarchie softmax. Lorsque cette option est à 0, l'algorithme utilise une approche de sampling négatif plutôt que la hiérarchie softmax pour l'entraînement.

-sample 1e-4 : Spécifie le seuil pour la downsampling des mots fréquents. Les mots fréquents peuvent être downsamplés pour éviter de donner trop de poids aux mots très courants.

-threads 20 : Indique le nombre de threads à utiliser pour l'entraînement. Dans cet exemple, 20 threads sont utilisés.

-binary 1 : Spécifie si les vecteurs d'embedding doivent être sauvegardés au format binaire (1) ou texte (0). Dans cet exemple, ils seront sauvegardés au format binaire.
-iter 15 : Indique le nombre d'itérations (epochs) d'entraînement.

Ces options déterminent différents aspects de la manière dont le modèle Word2Vec est entraîné, influençant la qualité et les caractéristiques des embeddings résultants.

Question : Quel est le vecteur du mot hello ?
Question : Ecrivez un programme qui prend en entrée deux mots et qui calcule la similarité cosine entre ces deux mots ?

Vecteur du mot *'hello'*: [-1.008897, -2.308451, -1.194275, -4.362735, -2.566641, 0.314524, 2.26193, 0.054648, 1.564844, -0.460636, 0.792198, 1.368713, 0.78904, 0.023878, -1.936209, -0.780074, -0.206716, 1.281536, 0.136857, -0.289974, 0.343309, 0.07578, 1.463206, 1.09872, 2.726326, -2.116665, 0.761616, 1.469714, 1.25744, 0.550109, 0.109379, -2.392751, 3.322121, 1.719002, 3.541541, 3.793948, 1.122714, -0.523812, -0.055596, 1.201102, 1.378412, -0.55792, -2.854434, -0.524985, -3.309722, -0.389988, 0.938012, -0.784064, -0.417545, -0.082834, 2.830519, -0.120168, 0.725053, 1.539997, 0.262444, -0.274206, -0.680187, 0.994122, -0.016498, 0.76964, 1.805622, -1.43584, 0.346884, 4.835496, -1.867374, -1.036754, 3.820033, -2.141501, -0.833632, 0.315947, -1.105845, 0.848341, 0.776994, 0.357851, -0.861111, 0.366939, -1.105156, 0.93178, -0.60639, -1.727933, 1.057621, 0.53754, -1.28147, -2.091967, 0.352063, -2.116781, -1.415346, 0.168833, -0.666945, 0.59475, 0.29088, 1.932806, 0.87297, 0.789694, -1.182721, 1.373498, -0.151582, 0.971833, 1.092018, 0.838084, -0.420968, -1.018083, 1.815029, -0.51072, 0.266038, -0.968394, -1.331038, -1.31537, 1.583159, 2.581931, 0.521297, 0.95444, -0.137104, 0.160658, 2.170321, 1.767607, 0.676746, 0.713049, -0.82168, 0.914592, -2.497558, 3.07474, 0.178676, 3.856671, -0.154814, 2.022769, 0.244235, -0.663178, 2.118647, 0.157353, 1.95425, -1.544127, -2.111917, 1.159798, 0.201265, -0.885333, -0.910447, 2.66564, 0.514706, 0.55289, -1.140822, 0.440602, -0.608855, -1.937298, 0.536562, -2.935745, -1.704946, 1.441226, -1.967131, 2.579138, -1.645447, -1.120946, 1.101451, 1.109597, 0.419982, 0.825251, -1.625364, 0.442138, -0.728012, -0.41016, 0.142235, -2.428148, 0.139592, -1.615654, 0.144462, -1.407016, 0.399617, -1.278677, -0.615815, 1.96234, 1.425909, -3.031413, 1.193219, -1.138618, 0.736095, 3.09842, 0.436526, 0.70312, 0.314236, 1.333311, -1.566566, 0.920972, 0.175618, -2.399523, -3.665779, 0.868129, 2.06984, 1.566378, -0.307899, -1.528883, 0.995289, 1.738929, -0.536378, -0.082065, 1.008371, 0.17479, -0.586092, 1.509668, 2.335393, -0.948153, 0.484745, -0.287459, 0.02674, 1.923079, -0.833939, -1.582553, 0.171307, 0.903237, 0.600609, -1.395303, 0.853194, 0.531409, 0.276078, -1.10811, -0.316211, -0.785847, -1.827981, 1.327896, 0.847857, 0.807385, -2.556619, -0.69821, 0.815272, 1.034415, 2.137674, 1.608291, -1.046017, 2.023288, -0.186057, -1.202421, -1.454511, -1.067128, -0.369343, 1.796559, 0.390947, -0.916154, 1.022539, 1.273545, -3.337748, -1.343771, 0.709108, -3.916027, -1.296888, 0.319055, 2.343475, -1.91354, -0.192783, 0.506014, 1.506822, 0.477322, -0.097857, 0.176178, -0.505493, -1.291207, -0.447858, -0.208582, -0.212973, -1.202021, -1.050056, -0.024908, -0.128233, 0.364386, 2.419054, 1.029398, 0.463365, 3.264183, 1.320884, 0.388911, 2.808021, -1.085764, -2.876244, -0.451306, -1.010831, -2.971694, 1.091635, -1.888687, -3.036384, 1.145404, -1.292646, 1.006114, 1.126793, -2.813588, 1.247441, -1.225877, -0.982587, -1.219586, -0.852469, -1.678341, -2.309407, -0.885314, 0.15971, -1.358233, -0.244009, 0.74595, -0.446151, -1.490872, -0.322777, 1.89056, 0.53922, 2.864799, -2.672858, 0.316342, -1.322061, -0.150949, 1.851633, 0.499951, -1.398617, -1.050762, -0.532836, -3.746859, 3.780633, -1.158964, -0.203239, -1.829477, 3.483067, -0.549747, 0.997633, -0.279116, -1.696163, 0.987432, 2.013392, -0.42713, -1.831511, -0.68266, 0.795205, -2.774332, -0.753904, -1.854837, 2.391213, -0.80639, -3.14955, 5.203102, -1.366704, 0.848538, -0.471454, 1.19461, -1.710024, 1.31943, 1.194559, -1.186882, -1.157679, -0.348767, 3.868819, -0.575271, -3.38658, -3.120433, -2.885225, -1.682682, 0.958329, 2.4061, -1.091857, -0.461187, 0.807513, -1.143139, 1.268874, -4.077825, 2.000273, -0.577026, 0.040827, -2.797378, -1.556533, -1.408881, -0.733918, -0.014902, 1.020664, -0.209256, -2.343189, 0.143903, 2.937984, -1.786882, -0.221494, 0.376469, 0.426055, -1.863796, -0.34389, 1.077638, -1.661607, 2.57488, 0.554016, -1.21701, -1.149036, -1.215664, 0.856744, 0.643309, -2.601124, 0.401588,

-0.665818, -1.181297, 0.273477, -0.499626, -1.430754, 4.122453, 0.987856, -0.15448, -0.063721, -0.247651, 0.167338, 2.61871, -0.377334, -0.996266]

Similarité cosinus entre 'hello' et 'world': -0.028946194748510966

Question : Quels sont les mots les plus proches des mots suivants : best, football, france,wine?

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

Enter word or sentence (EXIT to break): wine

Word: wine Position in vocabulary: 3103

Word	Cosine distance
wines	0.529830
bread	0.499896
grape	0.496092
cidre	0.476362
unleavened	0.452229
chianti	0.441797
sparkling	0.438808
vinegar	0.434627
claret	0.428625
brandy	0.426827
vermouth	0.425767
liquors	0.425212
cider	0.420550
drink	0.410988
apfelwein	0.410715
loaf	0.405290
ciders	0.401532
liqueur	0.399411
fermented	0.397799
drank	0.397448
libations	0.397444
flavouring	0.392011
juice	0.391949
stewed	0.390888
liqueurs	0.390839
cherries	0.390575
dessert	0.387234
pickle	0.386277
cognac	0.386240
casks	0.381544
raspberries	0.381359
vodka	0.378089
kumquats	0.377755

almonds	0.376143
currants	0.374187
beverage	0.370944
flavored	0.369633
leche	0.369375
whisky	0.368388
bitters	0.367255

Enter word or sentence (EXIT to break): football

Word: football Position in vocabulary: 623

Word	Cosine distance
soccer	0.561850
basketball	0.556496
rugby	0.527656
baseball	0.499982
hockey	0.499283
interuniversity	0.450845
korfball	0.445676
sports	0.435907
liga	0.433487
snooker	0.432859
midfielder	0.422272
lacrosse	0.417922
coached	0.410455
bulldogs	0.409033
athletic	0.402810
teams	0.400559
ymca	0.400533
jongg	0.399090
fulham	0.398985
volleyball	0.391488
cfl	0.390506
team	0.389559
wafl	0.389150
knute	0.388786
fiorentina	0.388099
sport	0.385548
rockne	0.385478
chievo	0.385009
leagues	0.383837
midfield	0.382000
everton	0.377341
handball	0.377079
defensemen	0.376460
meazza	0.376229
nfl	0.376060
tennis	0.375422
deportivo	0.374899
shinty	0.373387
goalkeeper	0.373011
softball	0.372690

Word: **france** Position in vocabulary: 303

Word	Cosine distance
french	0.559701
spain	0.541337
italy	0.505466
provence	0.495340
netherlands	0.486763
commune	0.477533
belgium	0.467802
germany	0.456381
portugal	0.453496
ferrand	0.451681
alsace	0.448190
paris	0.446346
loire	0.446145
russia	0.445998
mulhouse	0.444192
brittany	0.440423
calais	0.437420
baudouin	0.434878
nantes	0.432362
partement	0.429104
marseille	0.424050
aquitaine	0.423582
toulouse	0.421673
universite	0.419581
picardy	0.419028
denmark	0.418759
britain	0.415788
switzerland	0.414077
vres	0.408530
philippe	0.406850
belgians	0.406513
huguenots	0.404874
elba	0.401733
burgundy	0.401587
albret	0.401348
bordeaux	0.400378
corse	0.400099
austria	0.398334
marne	0.398159
vichy	0.398013

Word: best Position in vocabulary: 299

Word	Cosine distance
finest	0.440608
oscars	0.416204
better	0.392773
worst	0.388984
fondly	0.381800
greatest	0.376055

razzie	0.365956
biggest	0.338933
bafta	0.337587
awards	0.331020
nominations	0.322448
grammy	0.319362
outstanding	0.316899
telemark	0.312753
well	0.312518
filmfare	0.311801
award	0.306036
eastwood	0.292366
gingold	0.287363
mvp	0.281451
earliest	0.275153
favorite	0.274746
hermione	0.273202
clint	0.272508
emmy	0.269486
pagal	0.267734
popularly	0.267573
favourite	0.264448
cate	0.263733
krush	0.262886
famous	0.257959
sharpest	0.257944
funniest	0.257063
showcase	0.252040
gabby	0.249609
artist	0.249501
underrated	0.248733
houdini	0.248115
bittersweet	0.247243
nominated	0.247022

Word: apple Position in vocabulary: 1221

Word	Cosine distance
macintosh	0.556345
imac	0.499862
appleworks	0.467271
performa	0.459667
iigs	0.459127
quickdraw	0.442507
wozniak	0.442321
trs	0.429233
ibook	0.428612
macs	0.425125
raskin	0.422047
microsoft	0.418906
hypercard	0.414907
intel	0.413199
amigas	0.410028
compaq	0.403991
macintoshes	0.403952

ibm	0.403935
microcomputer	0.403127
amiga	0.402310
iic	0.401004
atari	0.400911
ecs	0.399866
os	0.392552
laptop	0.392457
jef	0.386823
visicalc	0.386124
commodore	0.384758
claris	0.384240
ipod	0.382358
iie	0.381218
prodos	0.380294
pc	0.379028
multiplan	0.375702
macbook	0.374962
microcomputers	0.370556
powerpc	0.369053
truetype	0.367554
tramiel	0.361021
xerox	0.360098

Word: mouse Position in vocabulary: 2800

Word	Cosine distance
mice	0.462794
trackball	0.437051
joystick	0.406525
cursor	0.388708
joysticks	0.386537
touchscreen	0.380922
buttons	0.379336
mickey	0.365166
keystrokes	0.347479
engelbart	0.345172
mousepad	0.341425
keyboard	0.332658
chorded	0.329170
cutouts	0.324654
logitech	0.323490
widgets	0.309987
keyer	0.308482
chording	0.306588
microcebus	0.304270
pad	0.302311
button	0.301446
menus	0.298427
toolbar	0.297240
paddles	0.289704
controllers	0.288574
clicking	0.286677
moth	0.286080
guis	0.284028

keypad	0.283947
pinky	0.283234
pda	0.282275
amstrad	0.281096
arabidopsis	0.279791
cursors	0.276445
cheirogaleidae	0.275020
keyboards	0.271198
intellimouse	0.270800
paddle	0.270799
messagepad	0.269608
graphical	0.269436

Word: macron Position in vocabulary: 21657

Word	Cosine distance
diacritic	0.630436
circumflex	0.608093
diaeresis	0.587922
diacritics	0.560388
cedilla	0.551746
diacritical	0.545063
macrons	0.527692
handakuten	0.506562
breve	0.505365
dakuten	0.498815
dotless	0.495805
ligatures	0.492124
digraphs	0.485479
kahak	0.484778
umlaut	0.483103
digraph	0.472638
buailte	0.464106
alif	0.460524
transliterations	0.459042
ayin	0.450875
semivowel	0.446824
ligature	0.446817
tilde	0.445969
umlauts	0.443205
okina	0.440630
apostrophe	0.435255
vowel	0.434372
uppercase	0.430606
palatalized	0.427483
accent	0.424146
palatalisation	0.423413
yod	0.420437
kana	0.418676
diphthong	0.417356
vowels	0.417056
pronunciations	0.417006
glottal	0.416264
bilabial	0.414842
punctuation	0.413992
palatalization	0.413314

Observation :

Les mots les plus proches pour chaque mot donné sont affichés avec leurs distances cosinus. Voici un résumé des résultats :

Pour les mots "best", "football", "france" et "wine" :

"Best" : ["finest", "oscars", "better", "worst", "fondly"]
"Football" : ["soccer", "basketball", "rugby", "baseball", "hockey"]
"France" : ["french", "spain", "italy", "provence", "netherlands"]
"Wine" : ["wines", "bread", "grape", "cidre", "unleavened"]
Pour les mots "apple", "mouse" et "macron" :

"Apple" : ["macintosh", "imac", "appleworks", "performa", "iigs"]
"Mouse" : ["mice", "trackball", "joystick", "cursor", "joysticks"]
"Macron" : ["diacritic", "circumflex", "diaeresis", "diacritics", "cedilla"]

Observations :

Les mots les plus proches pour chaque terme semblent être sémantiquement cohérents. Par exemple, pour "best", on obtient des mots tels que "finest" et "better", ce qui est conforme aux attentes.
Pour "mouse", les mots proches sont liés aux dispositifs d'entrée, tels que "trackball" et "joystick".
Pour "macron", les mots proches sont des termes linguistiques liés aux diacritiques.

Ces résultats indiquent que les vecteurs de mots captent des relations sémantiques et linguistiques intéressantes entre les mots dans l'espace vectoriel.

Question : Vérifier que les analogies suivantes fonctionnent ?

```
man woman king : queen
athens greece paris : france
berlin germany madrid : spain
man woman son : daughter
```

Question : Vérifier que les analogies suivantes fonctionnent ?

```
write writes decrease : decreases
man woman husband : wife
us italy hamburger : bologna
us australia hamburger : hotdog
```

```
uapv2403399@pedago01c:~/Donnees_itinerantes_depuis_serveur_pedagogique/Mes D
Enter three words (EXIT to break): man woman king

Word: man  Position in vocabulary: 243

Word: woman  Position in vocabulary: 1013

Word: king  Position in vocabulary: 187

                                             Word              Distance
------------------------------------------------------------------------
                                            queen              0.446631
                                            anjou              0.393614
```

```
Enter three words (EXIT to break): athens greece paris

Word: athens  Position in vocabulary: 3066

Word: greece  Position in vocabulary: 1248

Word: paris  Position in vocabulary: 1055

                                             Word              Distance
------------------------------------------------------------------------
                                           france              0.496559
                                            xmas               0.409582
```

```
Enter three words (EXIT to break): berlin germany madrid

Word: berlin  Position in vocabulary: 1360

Word: germany  Position in vocabulary: 324

Word: madrid  Position in vocabulary: 4732

                                             Word              Distance
------------------------------------------------------------------------
                                           spain               0.491485
```

```
Enter three words (EXIT to break): man woman son

Word: man  Position in vocabulary: 243

Word: woman  Position in vocabulary: 1013

Word: son  Position in vocabulary: 388

                                             Word              Distance
------------------------------------------------------------------------
                                         daughter              0.575615
```

Reponse 2 :

```
Enter three words (EXIT to break): write writes decrease

Word: write  Position in vocabulary: 1214

Word: writes  Position in vocabulary: 4309

Word: decrease  Position in vocabulary: 5368

                                                  Word              Distance
--------------------------------------------------------------------------------
                                             increases              0.336839
                                              increase              0.331326
                                             decreases              0.313760
```

```
Enter three words (EXIT to break): man woman husband

Word: man  Position in vocabulary: 243

Word: woman  Position in vocabulary: 1013

Word: husband  Position in vocabulary: 2471

                                                  Word              Distance
--------------------------------------------------------------------------------
                                                   her              0.427702
                                                  wife              0.421972
```

```
Enter three words (EXIT to break): us italy hamburger

Word: us  Position in vocabulary: 251

Word: italy  Position in vocabulary: 843

Word: hamburger  Position in vocabulary: 24840

                                                  Word              Distance
--------------------------------------------------------------------------------
                                                friuli              0.435056
                                              cagliari              0.433817
```

```
Enter three words (EXIT to break): us australia hamburger

Word: us  Position in vocabulary: 251

Word: australia  Position in vocabulary: 565

Word: hamburger  Position in vocabulary: 24840

                                                  Word              Distance
--------------------------------------------------------------------------------
                                             tasmania               0.397141
                                               monash               0.375199
                                              burgers               0.356754
```

Commentaire sur l'évaluation des analogies :

L'évaluation des analogies à l'aide des vecteurs de mots est une tâche intéressante mais délicate. Lors de l'exécution des analogies fournies, le modèle a donné des résultats qui peuvent sembler inattendus. Il est important de noter que les modèles de vecteurs de mots ont des limites et peuvent ne pas toujours saisir toutes les nuances des relations sémantiques.

Dans le cas spécifique des analogies proposées, il semble y avoir une divergence entre les résultats attendus et ceux générés par le modèle. Cela pourrait être dû à plusieurs facteurs, notamment la complexité des relations entre les mots, la taille et la qualité du corpus utilisé pour entraîner le modèle, ainsi que les choix de paramètres spécifiques.