```
#####################################


import gensim

import pprint

from gensim import corpora

from gensim.utils import simple_preprocess

doc_list = [

  "Hello, how are you?", "How do you do?",

  "Hey what are you doing? yes you What are you doing?"

]

doc_tokenized = [simple_preprocess(doc) for doc in doc_list]

dictionary = corpora.Dictionary()

BoW_corpus = [dictionary.doc2bow(doc, allow_update=True) for doc in doc_tokenized]

print(BoW_corpus)

id_words = [[(dictionary[id], count) for id, count in line] for line in BoW_corpus]

print(id_words)


# OUTPUT


# [[(0, 1), (1, 1), (2, 1), (3, 1)], [(2, 1), (3, 1), (4, 2)], [(0, 2), (3, 3), (5, 2), (6, 1), (7, 2), (8, 1)]]

# [[('are', 1), ('hello', 1), ('how', 1), ('you', 1)], [('how', 1), ('you', 1), ('do', 2)], [('are', 2), ('you', 3), ('doing', 2), ('hey', 1), ('what', 2), ('yes', 1)]]


#Assignment no : 2

#Name : Ishwar wagh

#Batch : B4

#Roll no : 6565

#Title : Natural Language Processing (NLP) using Gensim

#TFID


import gensim

import pprint

from gensim import corpora,models

from gensim.utils import simple_preprocess

import numpy as np

from nltk.tokenize import sent_tokenize, word_tokenize

import gensim

from gensim.models import Word2Vec



doc_list = [

  "Natural language processing (NLP) is a subfield of Artificial Intelligence (AI).",

  "This technology works on the speech provided by the user breaks it down for proper understanding and processes it accordingly.",

  "This is a very recent and effective approach due to which it has a really high demand in today's market. "

  "Natural Language Processing (NLP) is a subfield of artificial intelligence that deals with the interaction between computers and humans in natural language."


tokens1 = [[item for item in line.split()] for line in doc_list]

g_dict1 = corpora.Dictionary(tokens1)


print("The dictionary has: " +str(len(g_dict1)) + " tokens\n")

print(g_dict1.token2id)


g_bow =[g_dict1.doc2bow(token, allow_update = True) for token in tokens1]

print("Bag of Words : ", g_bow)


print("\n---------------------------------------------------------- ----------------------------------------\n")

g_dict = corpora.Dictionary([simple_preprocess(line) for line in doc_list])

g_bow = [g_dict.doc2bow(simple_preprocess(line)) for line in doc_list]


print("Dictionary : ")

for item in g_bow:

    print([[g_dict[id], freq] for id, freq in item])


g_tfidf = models.TfidfModel(g_bow, smartirs='ntc')


print("\nTF-IDF Vector:")
```

```
for item in g_tfidf[g_bow]:
    print([[g_dict[id], np.around(freq, decimals=2)] for id, freq in item])



print("\n---------------------------------------------------------------------------------------------- \n")



# OUTPUT:

#     The dictionary has: 56 tokens


# {'(AI).': 0, '(NLP)': 1, 'Artificial': 2, 'Intelligence': 3, 'Natural': 4, 'a': 5, 'is': 6, 'language': 7, 'of': 8, 'processing': 9, 'subfield': 10, 'This': 11, 'accordingly.': 12, 'and': 13, 'breaks': 14, 'by': 15, 'down': 16, 'for': 17, 'it': 18, 'on': 19, 'processes': 20, 'proper': 21, 'provided': 22, 'speech': 23, 'technology': 24, 'the': 25, 'understanding': 26, 'user': 27, 'works': 28, 'Language': 29, 'Processing': 30, 'approach': 31, 'artificial': 32, 'between': 33, 'computers': 34, 'deals': 35, 'demand': 36, 'due': 37, 'effective': 38, 'has': 39, 'high': 40, 'humans': 41, 'in': 42, 'intelligence': 43, 'interaction': 44, 'language.': 45, 'market.': 46, 'natural': 47, 'really': 48, 'recent': 49, 'that': 50, 'to': 51, 'today’s': 52, 'very': 53, 'which': 54, 'with': 55}

# Bag of Words :  [[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1), (9, 1), (10, 1)], [(11, 1), (12, 1), (13, 1), (14, 1), (15, 1), (16, 1), (17, 1), (18, 2), (19, 1), (20, 1), (21, 1), (22, 1), (23, 1), (24, 1), (25, 2), (26, 1), (27, 1), (28, 1)], [(1, 1), (4, 1), (5, 3), (6, 2), (8, 1), (10, 1), (11, 1), (13, 2), (18, 1), (25, 1), (29, 1), (30, 1), (31, 1), (32, 1), (33, 1), (34, 1), (35, 1), (36, 1), (37, 1), (38, 1), (39, 1), (40, 1), (41, 1), (42, 2), (43, 1), (44, 1), (45, 1), (46, 1), (47, 1), (48, 1), (49, 1), (50, 1), (51, 1), (52, 1), (53, 1), (54, 1), (55, 1)]]

#  ------------------------------------------------------------------------------------


# Dictionary :

# [['ai', 1], ['artificial', 1], ['intelligence', 1], ['is', 1], ['language', 1], ['natural', 1], ['nlp', 1], ['of', 1], ['processing', 1], ['subfield', 1]]

# [['accordingly', 1], ['and', 1], ['breaks', 1], ['by', 1], ['down', 1], ['for', 1], ['it', 2], ['on', 1], ['processes', 1], ['proper', 1], ['provided', 1], ['speech', 1], ['technology', 1], ['the', 2], ['this', 1], ['understanding', 1], ['user', 1], ['works', 1]]

# [['artificial', 1], ['intelligence', 1], ['is', 2], ['language', 2], ['natural', 2], ['nlp', 1], ['of', 1], ['processing', 1], ['subfield', 1], ['and', 2], ['it', 1], ['the', 1], ['this', 1], ['approach', 1], ['between', 1], ['computers', 1], ['deals', 1], ['demand', 1], ['due', 1], ['effective', 1], ['has', 1], ['high', 1], ['humans', 1], ['in', 2], ['interaction', 1], ['market', 1], ['really', 1], ['recent', 1], ['that', 1], ['to', 1], ['today', 1], ['very', 1], ['which', 1], ['with', 1]]


# TF-IDF Vector:

# [['ai', 0.55], ['artificial', 0.28], ['intelligence', 0.28], ['is', 0.28], ['language', 0.28], ['natural', 0.28], ['nlp', 0.28], ['of', 0.28], ['processing', 0.28], ['subfield', 0.28]]

# [['accordingly', 0.25], ['and', 0.12], ['breaks', 0.25], ['by', 0.25], ['down', 0.25], ['for', 0.25], ['it', 0.25], ['on', 0.25], ['processes', 0.25], ['proper', 0.25], ['provided', 0.25], ['speech', 0.25], ['technology', 0.25], ['the', 0.25], ['this', 0.12], ['understanding', 0.25], ['user', 0.25], ['works', 0.25]]

# [['artificial', 0.09], ['intelligence', 0.09], ['is', 0.18], ['language', 0.18], ['natural', 0.18], ['nlp', 0.09], ['of', 0.09], ['processing', 0.09], ['subfield', 0.09], ['and', 0.18], ['it', 0.09], ['the', 0.09], ['this', 0.09], ['approach', 0.18], ['between', 0.18], ['computers', 0.18], ['deals', 0.18], ['demand', 0.18], ['due', 0.18], ['effective', 0.18], ['has', 0.18], ['high', 0.18], ['humans', 0.18], ['in', 0.36], ['interaction', 0.18], ['market', 0.18], ['really', 0.18], ['recent', 0.18], ['that', 0.18], ['to', 0.18], ['today', 0.18], ['very', 0.18], ['which', 0.18], ['with', 0.18]]

# ------------------------------------------------------------------------------------
```