

인공지능기초응용 II

10 주차 과제

인공지능응용

K2025029 금동환

목차

1. [실습내용](#)

1. 실습내용

A. 필요 라이브러리 설치

```

(.venv) PS E:\Project\kmu-ai> pip install nltk gensim konlp konlpy
Collecting nltk
  Obtaining dependency information for nltk from https://files.pythonhosted.org/packages/4d/66/7d9e26593edda06e8cb5318746337fc2372279c3b0f4623539fe546df8b/nltk-3.9.1-py3-none-any.whl.metadata
  Downloading nltk-3.9.1-py3-none-any.whl.metadata (2.9 kB)
Collecting gensim
  Obtaining dependency information for gensim from https://files.pythonhosted.org/packages/79/7b/747fcb06280764cf28353361162eff68c6b0a3be34c43ead5ae393d3b19e/gensim-4.3.3-cp312-cp312-win_amd64.whl.metadata
  Downloading gensim-4.3.3-cp312-cp312-win_amd64.whl.metadata (8.2 kB)
Requirement already satisfied: konlp in e:\project\kmu-ai\.venv\lib\site-packages (0.0.58)
Requirement already satisfied: konlpy in e:\project\kmu-ai\.venv\lib\site-packages (0.6.0)
Collecting click (from nltk)
  Obtaining dependency information for click from https://files.pythonhosted.org/packages/7e/d4/7ebdbd03970677812aac39c869717059abb71a4cfc033ca6e5221787892c/click-8.1.8-py3-none-any.whl.metadata
  Downloading click-8.1.8-py3-none-any.whl.metadata (2.3 kB)
Requirement already satisfied: joblib in e:\project\kmu-ai\.venv\lib\site-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in e:\project\kmu-ai\.venv\lib\site-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in e:\project\kmu-ai\.venv\lib\site-packages (from nltk) (4.67.1)
Collecting numpy<2.0,>=1.18.5 (from gensim)
  Obtaining dependency information for numpy<2.0,>=1.18.5 from https://files.pythonhosted.org/packages/16/2e/86f24451c2d530c88daf997cb8d6ac622c1d46d19f5a031ed68a4b73a374/numpy-1.26.4-cp312-cp312-win_amd64.whl.metadata
  Downloading numpy-1.26.4-cp312-cp312-win_amd64.whl.metadata (2.5 kB)

```

B. 실습 내용

i. nltk의 'movie_review' 데이터

```

1: from nltk.corpus import movie_reviews
2:
3: sentences = []
4: for s in movie_reviews.sents():
5:     sentences.append(s)
6:
7: from gensim.models.word2vec import Word2Vec
8:
9: model = Word2Vec(sentences)
10: model.save('mytest.model')
11: model.init_sims(replace=True)
12: model.wv.similarity('actor', 'actress')
13: model.wv.most_similar("accident")
14: model.wv.most_similar(positive=['she', 'actor'], negative='actress', topn=5)
15: model.wv.get_vector('actor') # a vector for 'actor'
16: model.wv.get_vector('actress') # a vector for 'actress'

```

```

E:\Project\kmu-ai\venv\Scripts\python.exe E:\Project\kmu-ai\102\movie_review.py
[nltk_data] Downloading package movie_reviews to
[nltk_data]   C:\Users\dhkeu\AppData\Roaming\nltk_data...
[nltk_data]   Unzipping corpora\movie_reviews.zip.
E:\Project\kmu-ai\102\movie_review.py:11: DeprecationWarning: Call to deprecated 'init_sims' (Gensim 4.0.0 implemented internal optimizations that make calls to init_sims() unnecessary. init_sims() will be removed in a future version.)
  model.init_sims(replace=True)

```

ii. 벡터 유사도 계산

```

1: return sumxy / math.sqrt(sumxx + sumyy)
2:
3: v1, v2 = [3, 45, 7, 2], [2, 54, 13, 15]
4: print(v1, v2, cosine_similarity(v1, v2))
5:
6: print('-----')
7:
8: from gensim.models import Word2Vec
9: from sklearn.metrics.pairwise import cosine_similarity
10: import numpy as np
11: from gensim.models.word2vec import Word2Vec
12: model = Word2Vec.load("mytest.model")
13:
14: va = model.wv.get_vector('actor').reshape(1, -1) # a vector for 'actor'
15: vb = model.wv.get_vector('actress').reshape(1, -1) # a vector for 'actress'

```

```

E:\Project\kmu-ai\venv\Scripts\python.exe E:\Project\kmu-ai\102\vector_similarity.py
[3, 45, 7, 2] [2, 54, 13, 15] 0.97228425171235
-----
[[ 0.43530163 -0.1481481  0.47547887 -0.8363712  1.0041441 -0.11163238
 -0.2889088  -0.10212886  0.26232174 -0.30038667 -0.08797348 -0.8711544
 -0.15492052  0.5808744  -0.15711911 -0.16956213 -1.1743414  0.61737573
  1.216472  -0.95696163 -0.3227595  0.9622122  0.19120777 -0.64779186
 -0.06530997 -0.07187551  0.08704498  0.62366563  0.09052198 -0.1942127
 -0.42828846 -1.2525979  0.4557532  0.52569443  1.0936611  0.60951525]]

```

iii. 네이버 영화평 데이터

```

1  import codecs
2
3
4  def load_pre_tokenized(filename):
5      sentences = []
6      with open(filename, 'r', encoding='utf-8') as f:
7          for line in f:
8              line = line.strip()
9              if not line:
10                 continue
11             # 토큰은 "타이/PoS" 토큰으로, 토큰과 토큰을 split('/')로 쪼개서 사용
12             tokens = [ tok.split('/')[0] for tok in line.split() ]
13             sentences.append(tokens)
14         return sentences
15
16 def read_data(filename):
17     with open(filename, mode='r', encoding='utf-8') as f:

```

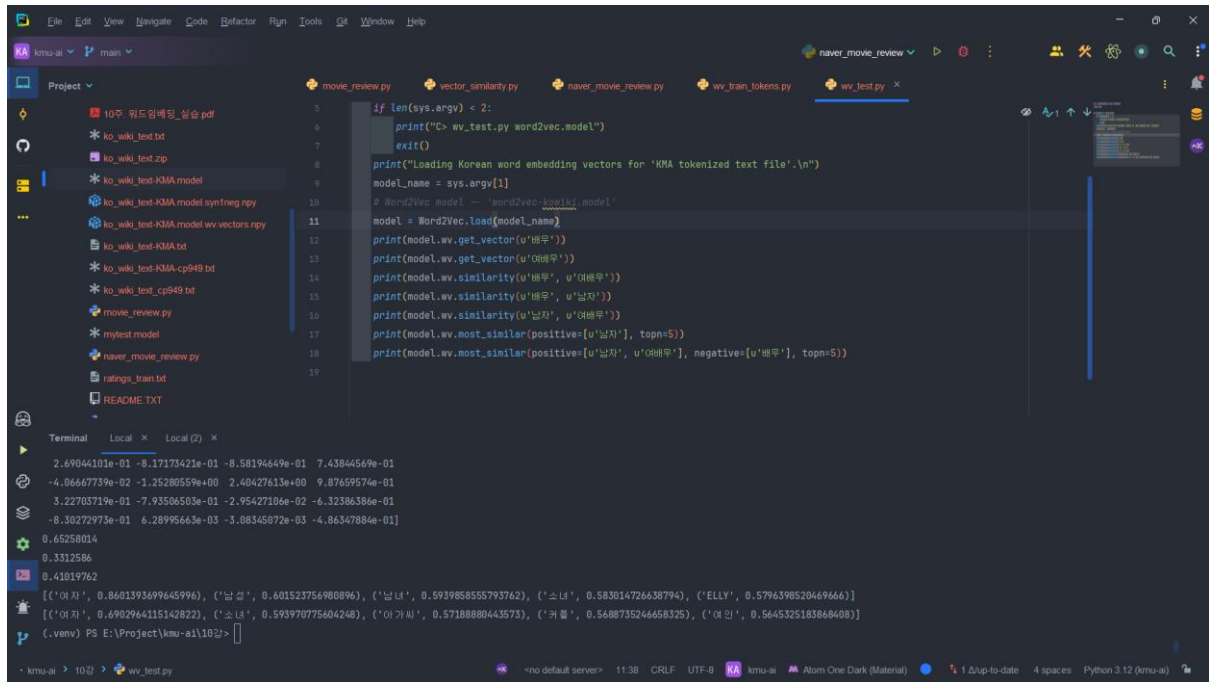
Run: E:\Project\kmu-ai\venv\Scripts\python.exe E:\Project\kmu-ai\102\naver_movie_review.py
KLT2000 -- morph analysis for 200K Naver movie reviews. Waiting several minutes...

```

17  if __name__ == "__main__":
18      if len(sys.argv) < 2:
19          print("C: vw_train_tokens.py token-list.txt")
20          exit()
21
22      tokenized_file = sys.argv[1]
23      model_file = tokenized_file[:-4] + '.model'
24      model = vw.train_tokens(tokenized_file)
25      # 'KMA tokenized text file'
26      model.save(model_file)
27      print(f"--> Model file <{model_file}> was created!\n")

```

Terminal: (.venv) PS E:\Project\kmu-ai\102> python .\vw_train_tokens.py ko_wiki_text-KMA.txt
Training word embedding vectors for <ko_wiki_text-KMA.txt>.
--> Model file <ko_wiki_text-KMA.model> was created!
(.venv) PS E:\Project\kmu-ai\102>



```
File Edit View Navigate Code Refactor Run Tools Window Help
kmu-ai main
Project
  10주 워드임베딩_실습.pdf
  * ko_wiki_test.txt
  ko_wiki_test.zip
  * ko_wiki_test-KMA.model
  ko_wiki_test-KMA.model.synfreg.npy
  ko_wiki_test-KMA.model.wv.vectors.npy
  ko_wiki_test-KMA.txt
  * ko_wiki_test-KMA.cp949.txt
  * movie_review.py
  * mytest.model
  naver_movie_review.py
  ratings_train.txt
  README.TXT
Terminal Local Local (2)
2.69044101e-01 -8.17173421e-01 -8.58194449e-01 7.43944569e-01
-4.06667739e-02 -1.25280559e+00 2.40427613e+00 9.87659574e-01
3.22703719e-01 -7.93586503e-01 -2.95427106e-02 -6.32386386e-01
-8.3927273e-01 6.28995663e-03 -3.08345072e-03 -4.86347884e-01]
0.65258014
0.3312586
0.41019762
[('여자', 0.8401393699645996), ('남성', 0.401523756980896), ('남녀', 0.593985855793762), ('소년', 0.583014726638794), ('ELLY', 0.5796398520469666)]
[('여자', 0.69029644115142822), ('소년', 0.593970775604248), ('여가씨', 0.57188880443573), ('커피', 0.5680735246658325), ('여인', 0.5645325103869408)]
(.venv) PS E:\Project\kmu-ai\10강>
```