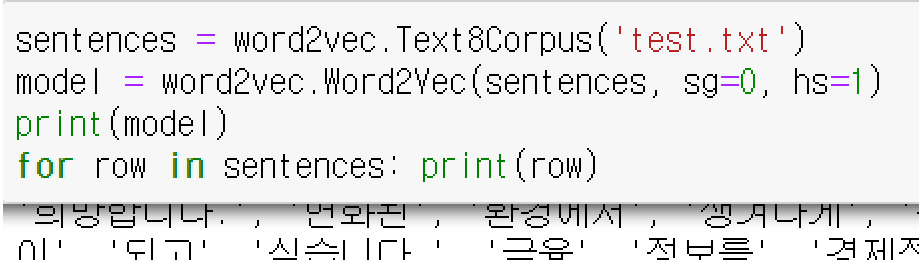
보고서17100239권도한

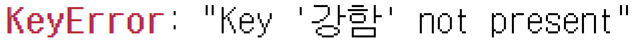
문제와 해결방법

word2vec

이용자의 글을 받아서 글 내부 단어만 벡터화(CBOW)해서 유사한 단어를 볼 경우, 글에 자주 등장하지 않는 단어의 경우 표현이 되지 않는다.



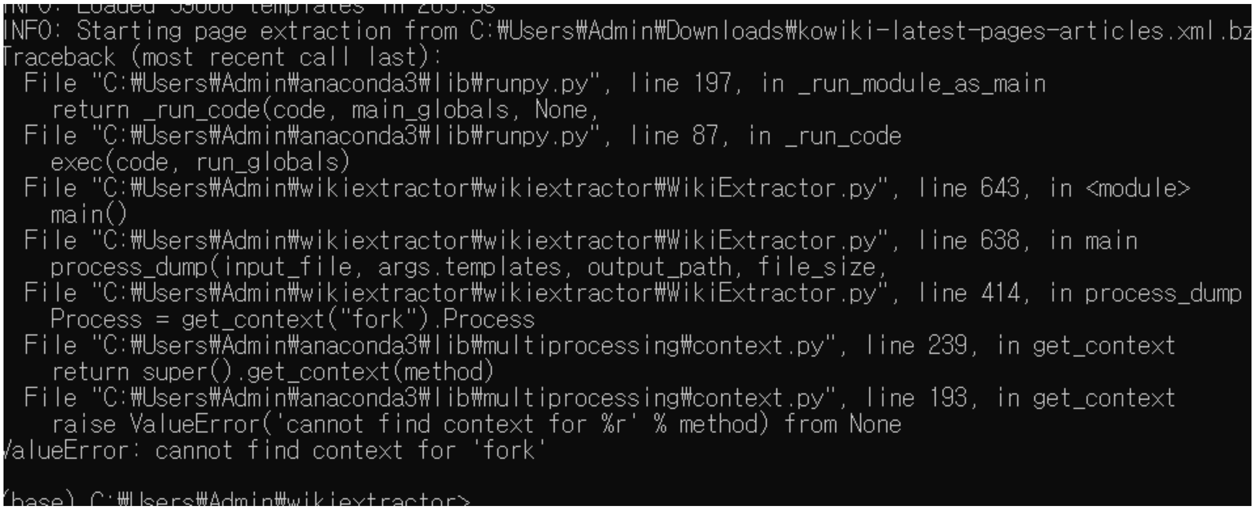
```
sentences = word2vec.Text8Corpus('test.txt')
model = word2vec.Word2Vec(sentences, sg=0, hs=1)
print(model)
for row in sentences: print(row)
'희망합니다.', '변화된', '환경에서', '생겨나게', '
이', '되고', '있습니다.', '금융', '정보를', '경제지
```



```
KeyError: "Key '강함' not present"
```

→ 프로젝트의 목표가 이용자가 이용한 글에서의 단어간의 관계성을 분석하고 밝히는 것이 아닌 효과적인 유의어를 추천하는 것이므로 다른 대용량의 데이터셋을 활용할 필요를 느낌.

Wiki 데이터를 training에 활용



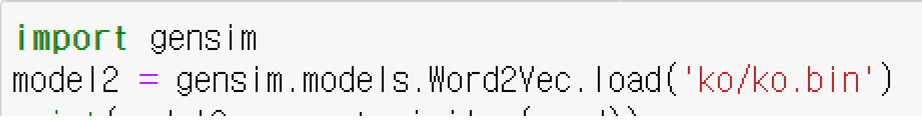
```
INFO: Loaded 35000 templates in 203.3s
INFO: Starting page extraction from C:₩Users₩Admin₩Downloads₩kowiki-latest-pages-articles.xml.bz
Traceback (most recent call last):
  File "C:₩Users₩Admin₩anaconda3₩lib₩runpy.py", line 197, in _run_module_as_main
    return _run_code(code, main_globals, None,
  File "C:₩Users₩Admin₩anaconda3₩lib₩runpy.py", line 87, in _run_code
    exec(code, run_globals)
  File "C:₩Users₩Admin₩wikiextractor₩wikiextractor₩WikiExtractor.py", line 643, in <module>
    main()
  File "C:₩Users₩Admin₩wikiextractor₩wikiextractor₩WikiExtractor.py", line 638, in main
    process_dump(input_file, args.templates, output_path, file_size,
  File "C:₩Users₩Admin₩wikiextractor₩wikiextractor₩WikiExtractor.py", line 414, in process_dump
    Process = get_context("fork").Process
  File "C:₩Users₩Admin₩anaconda3₩lib₩multiprocessing₩context.py", line 239, in get_context
    return super().get_context(method)
  File "C:₩Users₩Admin₩anaconda3₩lib₩multiprocessing₩context.py", line 193, in get_context
    raise ValueError('cannot find context for %r' % method) from None
ValueError: cannot find context for 'fork'

(base) C:₩Users₩Admin₩wikiextractor>
```

os 및 환경의 문제로 제대로 토크나이저가 작동하지 않고, 사전 훈련된 언어 모델이 효율이 좋다는 것을 알게되어서 이쪽으로 넘어가게 된다.

사전 훈련된 언어 모델(Pre-trained language model)

https://stackoverflow.com/questions/66952438/attributeerror-cant-get-attribute-vocab-on-module-gensim-models-word2vec한국어 pretrain된 모델을 구해서 넣는데, gensim에서 버전이 지원하지 않아서 3.8.3버전으로 내려줬다.



```
import gensim
model2 = gensim.models.Word2Vec.load('ko/ko.bin')
```

프로젝트 수행 중 새롭게 배운 점 및 느낀점

사전에서 제공하는 유의어를 사용자가 자신의 글의 문맥에 맞게 활용하도록 제공하고 싶다고 생각했는데, 이미 좋은 모델이 다양한 언어로 제공되고 있어서 활용만 잘해도 정말 괜찮은 결과가 나와서 놀라웠다.

또한 이러한 임베딩 방식으로 키워드확장 및 의미찾기 정보검색에 많이 활용되고 있음을 알게되어서 흥미가 생겼다.

예상되는 프로젝트 점수

기본점수 20

완성도 19

아이디어6

난이도6

외부 오픈소스 사용/참고

word2vec

import gensim

API references: https://radimrehurek.com/gensim/models/word2vec.html

korean pretrain word2vec:

https://github.com/Kyubyong/wordvectors

BERT

from transformers import TFBertForMaskedLM
from transformers import AutoTokenizer

bertbase:

https://github.com/google-research/bert/blob/master/multilingual.md

klue-bert:

https://github.com/KLUE-benchmark/KLUE

학습자료:

https://riverkangg.github.io/nlp/nlp-bertWordEmbedding/

https://wikidocs.net/152922

pytorch

tensorflow