추가 보고서

블로그와 뉴스 데이터의 'token'과 'token_nv' 칼럼을 str 데이터 타입이 아닌 list 데이터 타입으로 변형한 후 한 글자 단어나 추가 불용어는 삭제하는 과정을 거쳐 단어 빈도수 확인을 진행하였다.
이후 추출한 word count를 기반으로 네이버 뉴스, 블로그 각각의 word cloud를 생성하였다.

Table 1. blog_word_count

'사회'	38,630
'건강'	31,189
'정신'	27,733
'지원'	25,309
'복지'	24,184
'출처'	22,593

Table 2. news_word_count

'건강'	9,593
'지원'	8,810
'정신'	8,366
'정책'	7,703
'코로나'	6,058
'사회'	5,929

Fig 1. blog_wordcloud

Fig 2. news_wordcloud

또한, 네이버 뉴스와 블로그의 내용을 KoNLPy를 통해 토큰화를 하고 word embedding 결과와 키워드를 통해 정책 수요를 분석하였다.

Table 3. '정신건강' + '정책'

'분절'	0.76153
'출산장려'	0.74494
'청소년복지정책'	0.72185
'필요할까요민주주'	0.72185
'제안대상'	0.72100

Table 4. **‘**정신건강**’ + ‘**심각**’**

‘정신건강상’	0.76231
‘건강’	0.73761
‘정신적정신과적’	0.73501
‘말한대요외로움’	0.73301
‘높아정신건강’	0.72642

Table 5. **‘**정신건강**’ + ‘**취약**’**

‘정신건강년’	0.74442
‘건강’	0.72937
‘있다정신건강’	0.72614
‘의료서비스체계’	0.72463
‘심리정서’	0.72167

Table 6. **‘**정신건강**’ + ‘**정책**’ + ‘**우선**’**

‘분절’	0.78953
‘출산장려’	0.77102
‘필요할까요민주주’	0.76108
‘시의적절한’	0.75677
‘건강영향평가’	0.75584

마지막으로 word2vec 시각화를 진행하였다.

Fig 3. 정신건강벡터

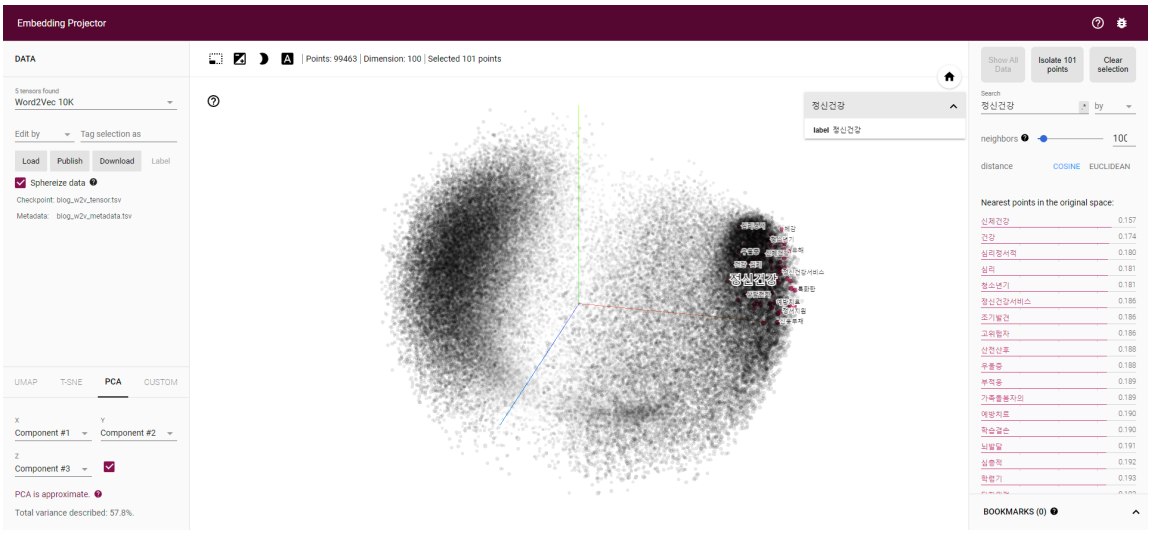

Fig 3은 PCA로 3차원 시각화된 모습이고 우측에 나열된 단어인 '신체건강', '건강', '심리정서적' 등은 검색어인 '정신건강'과 가까운 거리에 있는 단어이다.

Fig 4. 정신건강벡터**2**

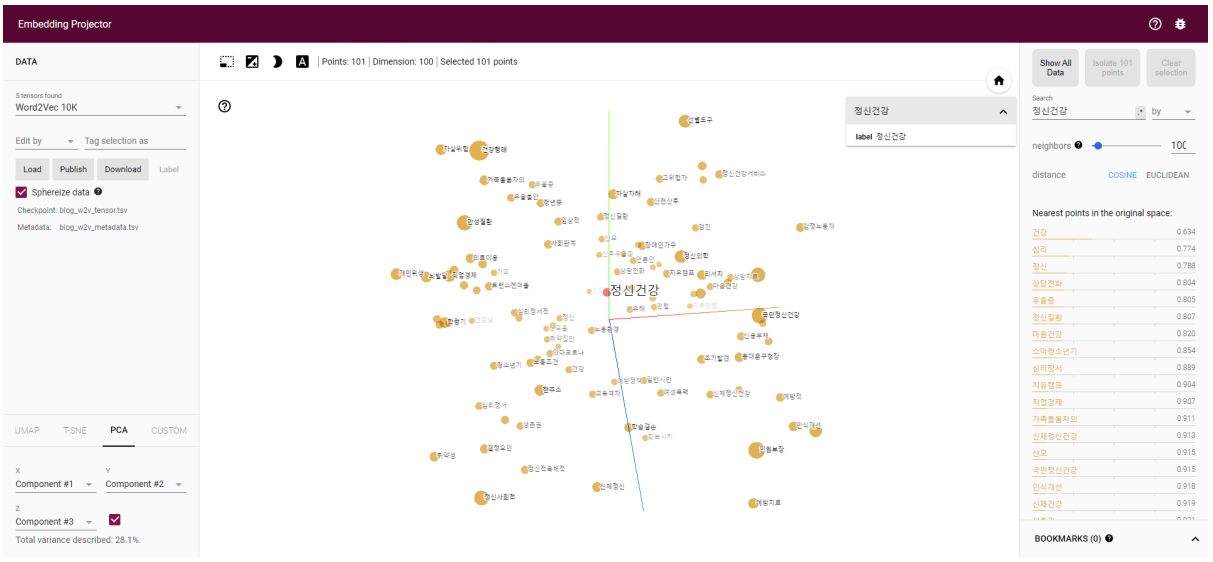

Fig 4를 통해 '정신건강'과 가까운 거리에 있는 단어가 무엇인지 알 수 있다.

Fig 5. 정책벡터

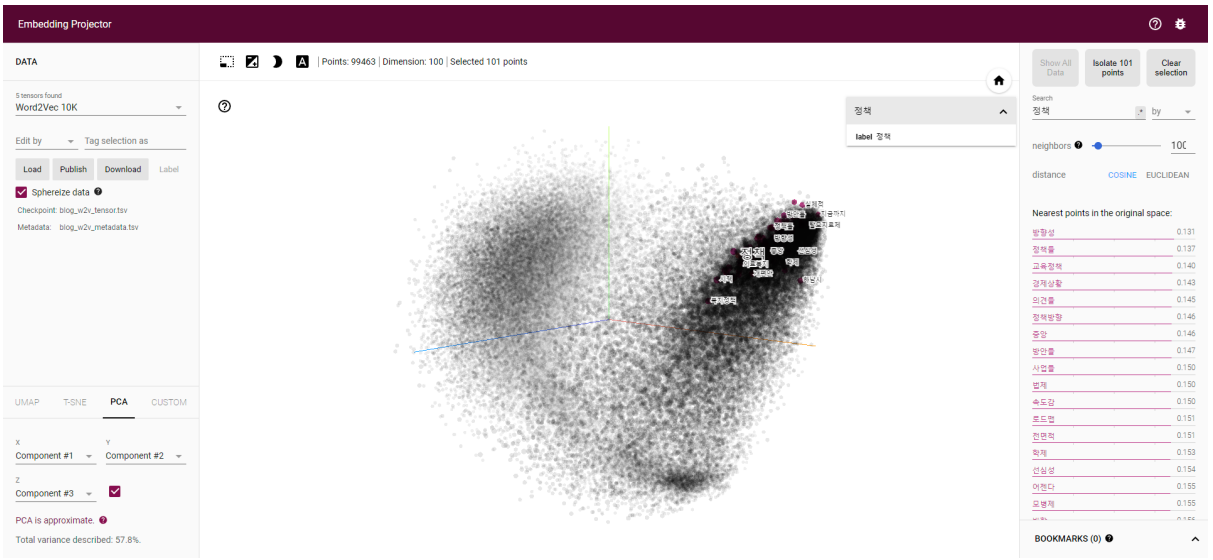

Fig 3과 같은 방식으로, 검색어는 '정책'이다. 우측에 있는 '방향성', '정책들', '교육정책' 등은 '정책'과 가깝다는 것을 알 수 있다.

Fig 6. 정책벡터**2**

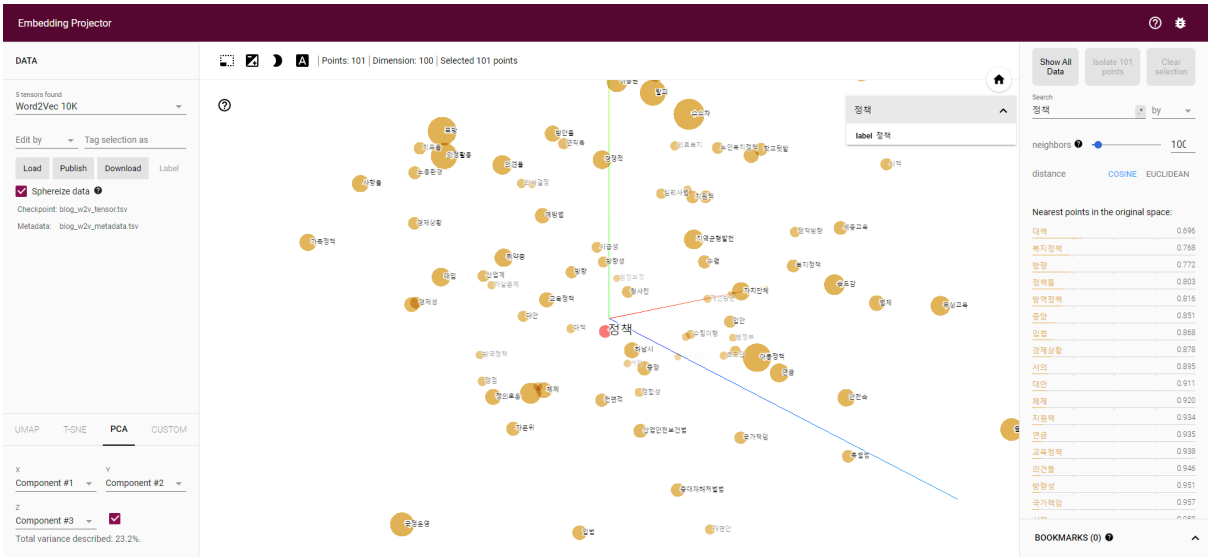

Fig 4와 같은 방식으로, 검색어인 '정책'과 가까운 거리에 있는 단어들을 확인할 수 있다.