**Weekly report**

날짜 : 2022-03-03

연구원 : 최윤석

* 수행결과

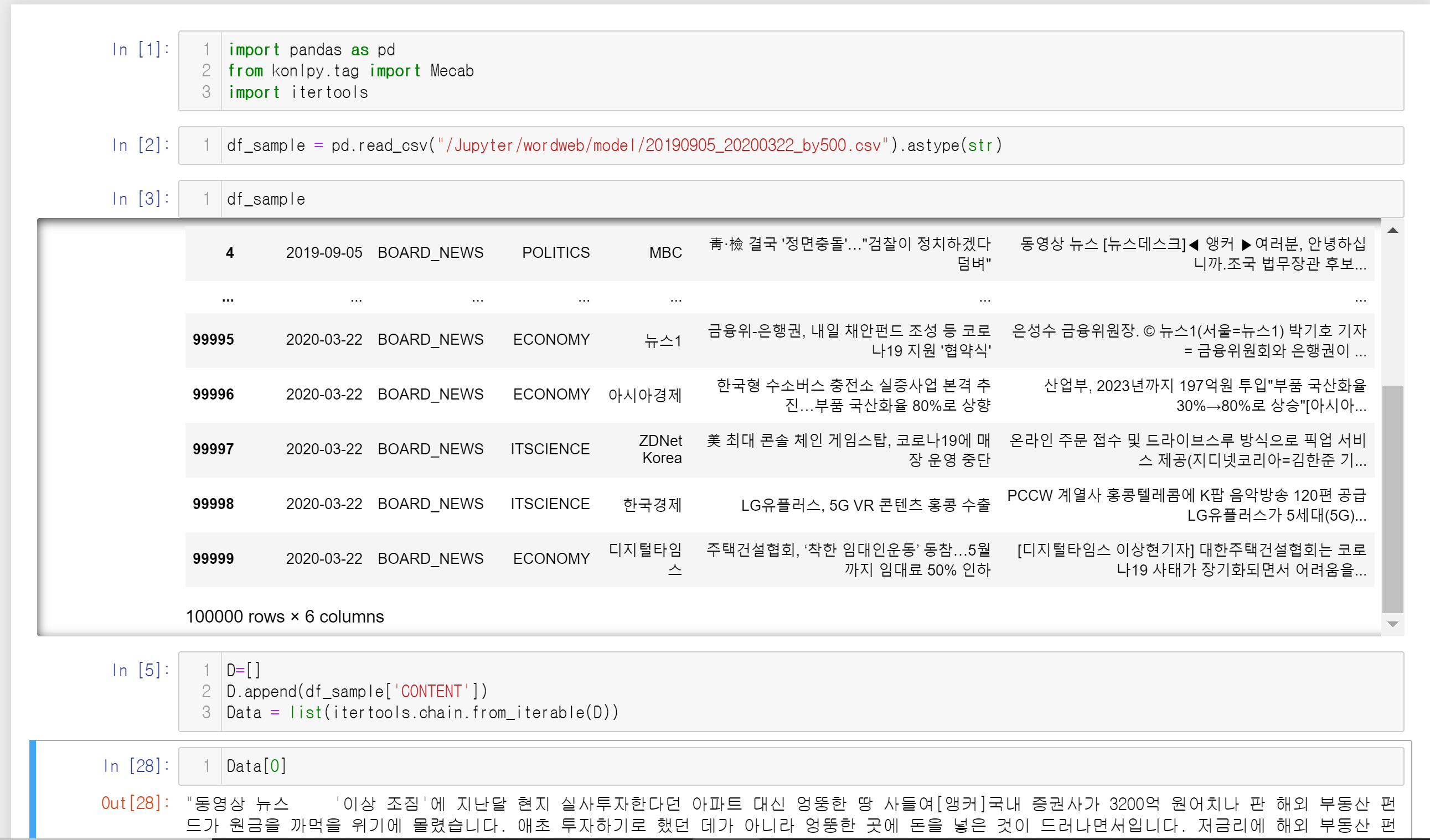
**▶ 연구소개** -> 현재 진행하고 있는 연구는 한국어 임베딩입니다. 임베딩은 자연어를 기계가 이해하는 언어인 숫자의 형태, 즉 Vector로 바꾼 결과 혹은 그 일련의 과정 전체를 의미합니다.

▶ **연구목적** -> 제 연구의 목적은 어떠한 한국어 단어를 입력으로 집어넣었을 때, 출력으로 입력으로 넣은 단어와 의미적으로 유사한 단어들을 뽑아내는 것입니다. 예를 들어, 김정은을 입력으로 넣었을 때, 북한이나 미사일 같은 출력이 나오게 하는것입니다. 추가적으로 NER(Named Entity Recogniton)을 이용해 인명, 지명, 기관명 등을 분류하고, 어떤 입력이 들어오면 그 입력과 유사한 뜻을 가지거나 관련된 인명, 지명, 기관명 등을 각각 출력하는 모듈 제작을 목표로 하고 있습니다.

이 연구의 목적을 달성하면, 검색사이트나 쇼핑사이트에서의 연관검색어 기능이나, 고객상담센터에서 고객의 문의글에 대해 자동답변 기능을 향상할 수 있다고 생각합니다.

**▶ 진행상황** -> 저는 지금까지 뉴스기사에 나오는 단어들 중 고유명사만을 뽑아, 각각의 단어들끼리의 코사인 유사도를 계산하고, 이 데이터를 T-sne라는 시각화 도구를 이용해 2차원 평면에 나타내는 작업을 진행하였습니다. T-sne는 높은 차원의 복잡한 데이터를 n차원에 차원 축소하는 방법으로, T-sne를 거친후 차원이 축소되면 비슷한 구조끼리 데이터를 정리한 상태이므로, 유사한 의미를 가진 단어들이 비슷한 곳에 분포한 것을 확인할 수 있습니다.

이에 대해 지금까지 짰던 코드에 대해 간략히 설명을 진행하겠습니다.



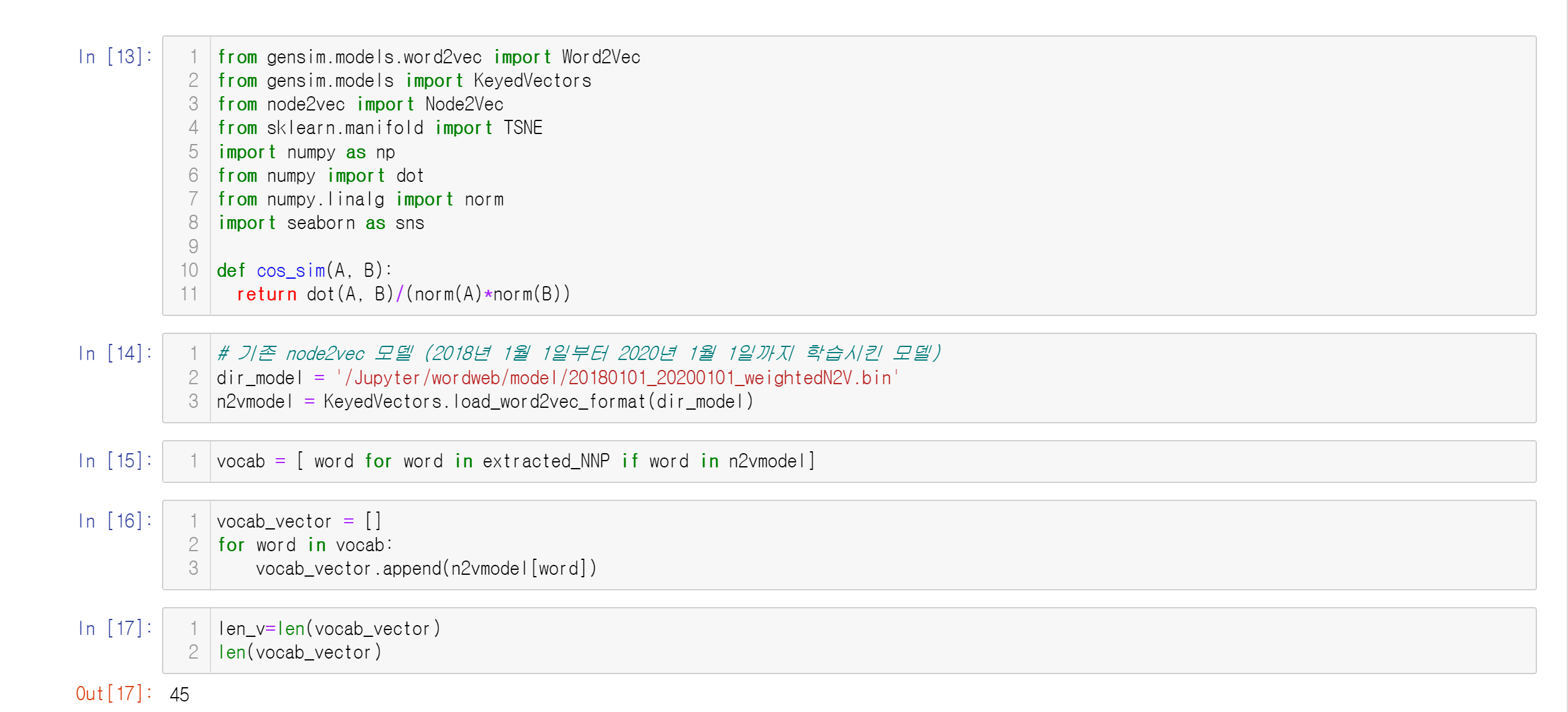
먼저 위의 코드를 통해 2019년 9월 - 2020년 3월까지의 뉴스데이터를 가져옵니다.



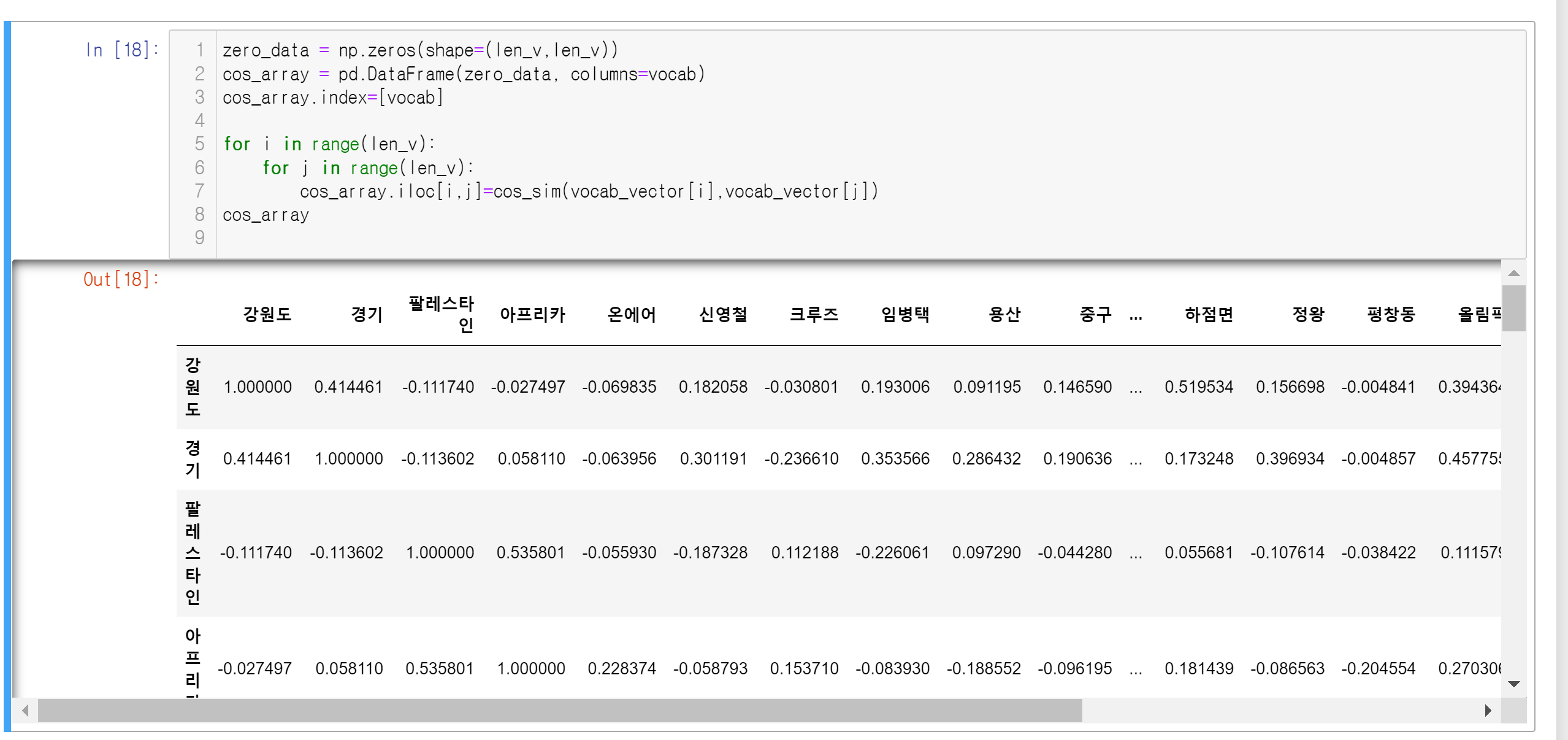
이후 가져온 10만개의 데이터중 랜덤으로 10개를 뽑고, 이를 은전한닢으로 형태소 분석을 해줍니다.



그중에서 고유명사인 단어들만 따로 추출해서 리스트로 만들어줍니다.



이후 기존의 만들어진 n2v 모델을 사용해서 앞서 추출한 고유명사중 n2v 모델에 존재하는 단어에 대해 추가적인 추출을 하고, 단어에 해당하는 벡터값을 넣어줍니다. 결과를 보면 10개의 무작위로 추출된 신문기사들에서 나온 고유명사(총 70개)중 n2v모델에 있는 고유명사는 45개입니다.



45\*45의 데이터프레임을 만들고, 각각 단어들의 코사인 유사도를 모두 구해줍니다.



추가적으로 특정 단어를 입력하였을때, 코사인 유사도가 높은 순서대로 내림차순 정리를 해줍니다.

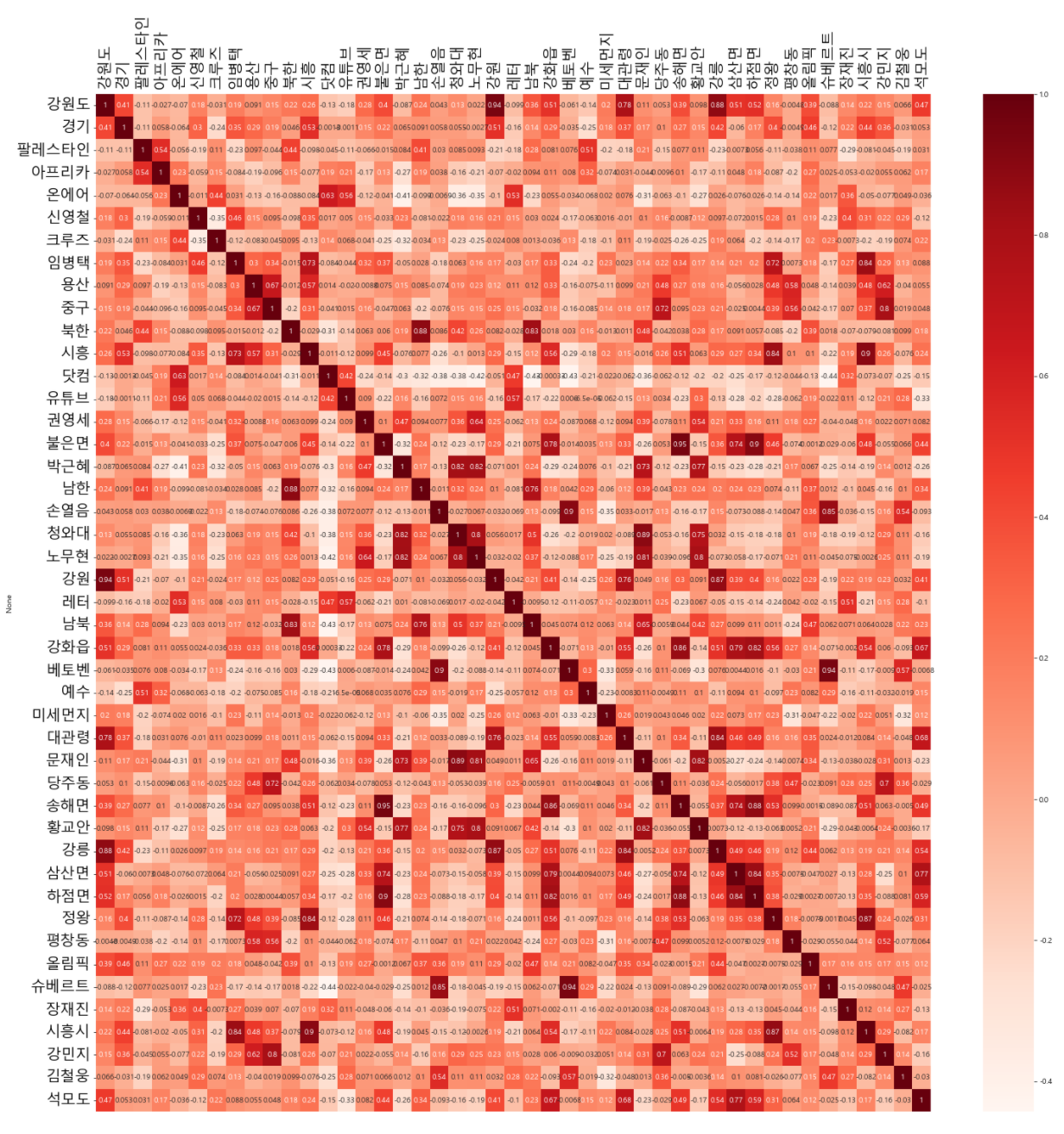


이후 벡터들을 t-sne를 이용해 2차원 공간으로 축소시키는 코드입니다. 여기서 T-sne의 파라미터를 설명하자면,   
-n\_components는 차원 축소 결과 임베딩되는 차원

-learning\_rate 는 일반적으로 10 - 1000 정도의 값을 가지며, 너무 높은 값이면 데이터가 가장 가까운 이웃과 너무 겹쳐져 같은거리에 있는 것처럼 보이고, 너무 낮은 값이면 대부분의 포인트가 특이치가 거의 없이 흩어져 있는것처럼 보입니다. 저는 변수를 써서 고유명사 수를 learning rate로 사용하였더니, 대부분의 경우 포인트가 있게 적당히 뭉쳐있는 좋은결과가 나옴을 확인했습니다.

-perplexity는 다른 manifold learning의 nearest neighbors 갯수에 사용되는 값을 뜻합니다. 일반적으로 큰 데이터 셋은 큰 perplexity 값을 필요로합니다. 대게 5-50정도의 값을 가지는게 좋습니다. 이 부분도 코드에서 변수로 처리하였는데, 수식은 제가 임의로 지정한 것이며, 10개의 추출에서 대부분 좋은 결과를 보였습니다.

early\_exaggerattion은 기존 공간에서 데이터의 클러스터 간 거리가 타겟 공간에서 얼마나 조밀하거나 먼 지를 나타내는 파라미터입니다. 이 값이 커지면 기존 공간에서 클러스터 사이의 공간이 타겟 공간에서 더 커지도록 학습됩니다. 기본값은 12이지만 저는 20을 사용하였습니다.



마지막으로 각각 단어간의 유사도를 쉽게 알 수 있게 히트맵으로 나타내었습니다.

* 수행계획

▶ 현재 임베딩에 사용하고 있는 모델은 조새롬 연구원이 만드신 것을 그대로 사용하고 있었는데, 이를 이해하고 추가적으로 모델을 더 좋게 발전시키는 작업을 진행하겠습니다. (매주)

▶ 현재까지 찾은 한국어 NER 코드는 라이브러리가 충돌해서 실행이 안되고 있었고, 여러 연구원분들께서 조언해 주셨지만, 아직 해결되지 않아서, 한국어 NER에 대해 사용할 수 있는 코드가 있는지 추가적으로 더 찾아보고 실행해 보겠습니다. (3/10까지)

▶ 만약 기존의 한국어 NER 코드가 도저히 사용이 어렵다면 관련 논문이나 인터넷을 참고하여 직접 NER 코드를 만드는 방식으로 진행할 예정입니다.

* 기타사항