# W2V : Hierarchical Softmax & Negative Sampling & Subsampling

# 2013250422 보건정책관리학부 김도향

● 실험 보고서

본 실험은 다음과 같은 조건에 따라 분류되어 진행되었다.

1. Skip Gram or CBOW
2. Subsampling or Not
3. Hierarchical Softmax or Negative Sampling

이와 함께, 다음과 같은 Hyperparameter를 조정하여 유의미한 변화를 도출하고자 하였다.

* Subsampling의 threshold
* 학습에 사용할 Dataset의 수
* Weight의 Dimension
* Negative Sampling의 Sampling 수

결론적으로, 실험은 실패하였다.

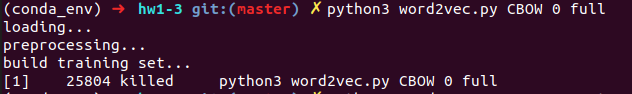
다양한 파라미터 조건 설정을 시도하였고, 위에서 언급한 샘플링 조건 등 또한 여러 케이스를 시도해보았으나, Analogical Reasoning 결과 정답률이 0에 수렴하였다.

예상되는 원인으로는

1. 학습에 사용한 데이터셋이 Full이 아닌 Part를 사용하였다.

첫 실험에서 Full 설정을 통해 전체 데이터셋을 학습시키도록 시도했다.

그러나 해당 실험을 진행하는 과정에서, 기기가 연산과정을 버티지 못해 프로세스가 강제로 종료되는 케이스가 발생하였다. 기기 자체의 성능 한계로 예상되며, 이를 해결하기 위해서는 더 좋은 기기를 사용하거나 Goole colab 등 클라우드 서비스를 사용하는 것이 바람직할 것으로 판단된다.



1. 잘못된 Skip Gram과 CBOW 코드를 작성했다.

학습 데이터를 Part로 재설정하고 실험을 계속 진행해보았다.

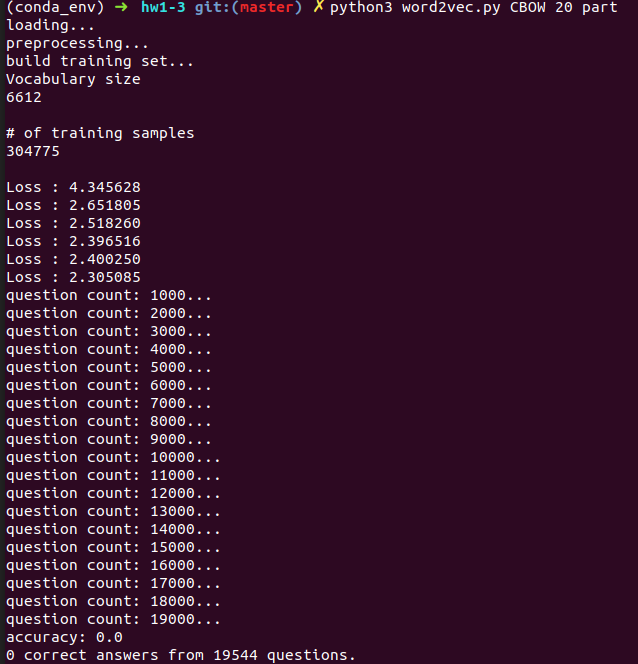
그 결과, 학습에 따른 Loss가 2 미만으로 떨어지지 않는 상황이 이어졌다.

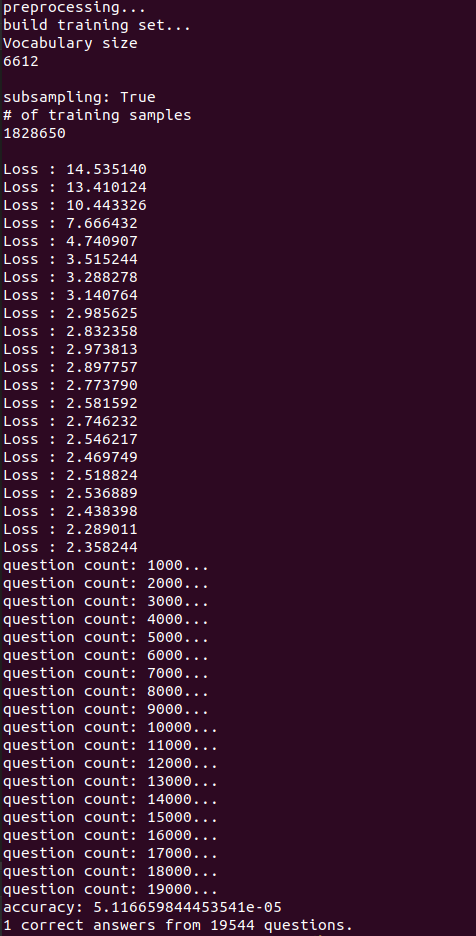
이를 해결하기 위해서,

* 학습데이터를 늘리는 방식
  + 기본 설정은 100000 라인의 텍스트 데이터를 읽어들이나, 이를 2~5배까지 늘려서 실행하였다.
* Dimension을 줄이는 방식(100 → 64)
* Negative Sampling의 샘플링 수 변화(5, 10, 15, 20)
* Subsampling Threshold 변화(0.001, 0.005, 0.0001)

이를 Skip Gram, CBOW를 바꿔가며 다양한 조건의 실험을 계속했으나,

다음과 같은 결과가 일관되게 나타났다.





학습 데이터의 수를 늘렸을 때 성공한 경우에도 단 1개의 정답에 그쳤기 때문에, 사실상 실패라고 볼 수 있다.

학습 데이터의 수를 더욱 늘리는 방식을 사용할 수도 있으나, 연산 과정에 소요되는 시간이 너무 오래걸려, 실험을 진행하기에 무리가 있다고 판단하여 중지하였다.

이외에도 window size, learning rate 또한 조정을 해보았으나, 눈에 띄는 차이가 발생하지 않았다.

1. Question Text의 대소문자 처리 문제

Analogical Reasoning 과정에 소문자 전환을 위한 코드를 추가하였으나,

해당 코드가 제대로 작동하지 않았을 경우 또한 존재한다. 따라서 이에 대한

추가적 검토가 필요하다.

이와 같은 Loss 학습 개선 정도의 한계 문제, 지나치게 큰 시간 복잡도 문제 등은 내부적으로 코드가 비효율적으로 작성되었을 것이라고 예상된다. 따라서 이후 다시 한 번 코드를 재작성하고, 더욱 성능을 향상시킨 기기 혹은 클라우드 환경에서 코드를 실행하여 새롭게 제출할 계획이다.