


# 9월 회고록

## 전체 돌아보기 :

자연어 처리에 대한 어느 정도의 이해는 모두 갖추게 되었음.  
전처리 및 모델 구현에 집중하여 특별히 강조할 사항은 없음.

## 8월 31일

- **이성준님** → 직접 손으로 코드 필사 (노트에, 컴퓨터 아님...ㄷㄷ)  
전처리 과정에서 original prompt 컬럼, text 컬럼 한개 컬럼으로 합쳐서 모델 돌리는 사람이 있고, 흔히 하는 것처럼 별개로 돌리는 사람들이 있음.  
original prompt를 직접 요약하는 모델 만들어서 비교해 볼 예정
  - **한찬혁님** → 따옴표 처리를 한 것과 안한 것에 상관관계가 크게 있는지 확인 → 크게 상관 없음, 텍스트의 절대적인 양이 많으면 점수가 좋았음.  
오타자와 점수에도 큰 상관관계는 없음
  - **배나영님** → wording 점수는 오타자 여부가 아닌 Summary라는 주제에 맞는 단어 사용 여부일 것 같다.  
(예시: I am → 감점. As mentioned in paragraph 5 → 정상 점수)  
Attention Is All You Need (Transformer 논문) 읽어본 후 MNC 천상진 매니저님과 대화해 봄.  
논문에서 왜 output이 이렇게 나오는지, 왜 잘 나왔는지 설명은 따로 안함.  
논문 저자 : 그냥 이렇게 했더니 결과가 대단하다. 왜 그런진 모른다. 궁금하면 너네가 해 보라  
자세한 내용은 하단 개인노트페이지 참고
-  Transformer
- **임승준님** → 0.482점짜리 코드를 필사해 보았으며, 한국어로 모든 line에 대한 설명과 분석 진행.  
아쉽게도, original 코드의 이슈로 성공적인 제출은 해보지 못함.  
수업시간에 배운 RNN을 사용해 볼 수 있을 것 같다.

- **이희상님** → MobileBERT를 도입하였으나 LB 점수가 5000 ~ 6000점 사이로 나옴.  
동일 한 코드로 타 모델들 돌려보았는데, 정상 범주 사이의 점수가 나옴.  
아마 MobileBERT를 구현하는 방식 자체가 다른 것 같음.  
MobileBERT 논문을 읽어보니 MobileBERT는 타 BERT계열 모델과 다르게 Inverted-Bottleneck BERT (IB-BERT)라는 모델이 Teacher 역할, MobileBERT가 Student 역할을 한다고 함.  
따라서, MobileBERT를 사용하고자하면 IB-BERT도 구현해야 할 듯 함.

- **변웅진님** → 빅콘테스트 관계로 금일 불참

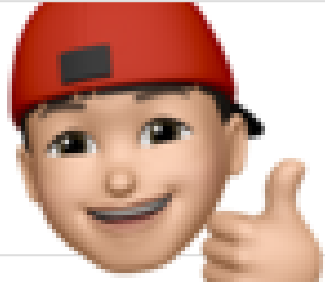
## 8월 31일 ~ 9월 6일

- **이희상님** → EDA 매 학생별로 복붙한 summary가 있으면 해당 text의 복붙 내용을 제거.  
(지지난주 도전과제에서 넘어옴, 이번 주엔 진짜 해보자)  
→ MobileBert를 위한 IB-BERT 구현 및 MobileBERT 모델 자체를 다시 구현  
(사실 이쯤되면 BERT-Tiny도 가벼운데 이미 성능은 잘나와서 이걸 그냥 쓸지도...?)  
(게으름 부릴 생각하지 말고 공부하는 셈치고 만들어보자...!)
- **이성준님** - original prompt를 직접 요약하는 모델을 구현해보고 해당 모델로 생성한 summary의 점수 평가해보기? (흠....)
- **한찬혁님** - EDA 노트북에서 빈도 관련하여 살펴보기 (chi-square 등)

**Commonlit EDA Practice**

Explore and run machine learning code with Kaggle Notebooks | Using data from CommonLit - Evaluate Student Summaries

[k https://www.kaggle.com/code/chanhyukhan/commonlit-eda-practice](https://www.kaggle.com/code/chanhyukhan/commonlit-eda-practice)



- **임승준님** - RNN으로 직접 모델 구현해보기!
- **배나영님** - 1. 개인노트 정리하기 2. 데이터 가지고 놀아보기
- **전체** → 아래 찐코딩 NLP 강좌 들어보기 (9월 3일까지 Available)

## 찐코딩 Jin Coding



누구나 쉽고 재밌게 인공지능의 도움을 받을 수 있는 세상을 꿈꿉니다. 🌟 여러분에게 시간 ⌚ 과 돈 💰 을 가져다줄 수 있는 코드를 만들어 영상으로 공유 할게요. 구독 눌러주시고 매주 만나요. 📺

📺 <https://www.youtube.com/@jincoding>

## 9월 7일

- 4시간 연속 특강으로 스킵...

## 9월 7일 ~ 9월 13일

- 지난 주 할 일에서 각자 추가하여 진행해보기.

## 9월 14일

- 배나영님 → Word Embedding에 관하여 자세히 공부해봄.  
word2vec, GloVe, para\_embedding, fasttext 등의 embedding 라이브러리가 존재함.  
자세한 내용은 하단 개인노트 페이지 참고.

### Word Embedding


- 임승준님 → CommonLit 웹사이트 참고하여 현재 대회로 나온 과제는 이미 해당 업체에서 상용화한 서비스인 것을 알게 되었음. 해당 서비스에서 활용하는 점수 (Lexile 지수)에 관하여도 정리.  
자세한 내용은 하단 개인노트 페이지 참고.

해당 서비스는 Turnitin과 Ecrie와 유사한 방식으로 구동된다.  
자세한 내용은 하단 개인노트 페이지 참고.

또한, discussion을 살펴보면 오타, 문장 길이 등의 정보를 수집.  
자세한 내용은 하단 개인노트 페이지 참고.


이외로도 코딩 실력을 높이기 위해 지속적으로 필사 중.

- 이희상님 → CommonLit 대회보다 현재 집중하고 있는 주제가 있어 해당 내용 중점적으로 공부  
표음문자, 표의문자 처리 관련 차이점에 집중하며 라이브러리 하나하나 뜯어가며 학습 중

 Ideogram-based vs. Phonogram-based Language

Explore and run machine learning code with Kaggle Notebooks | Using data from multiple data sources

<https://www.kaggle.com/code/jasonheesanglee/ideogram-based-vs-phonogram-based-language>



- 이성준님 → 이어드림 본 과정 대회 참여로 인해 새로운 시도를 해보지 못했음
- 한찬혁님 → 공모전 발표로 인한 금주 및 차주 불참

## 9월 14일 ~ 9월 20일

- 이희상님 → MobileBert를 위한 IB-BERT 구현 및 MobileBERT 모델 자체를 다시 구현
- 배나영님 → 단어 수준 등 확인하여, 수준 높은 단어로 구성된 요약본에 대해 Lexile점수 확인 예정
- 이성준님 → BERT에 대해서 조금 더 공부 한 후 모델 구현하여 제출 예정
- 임승준님 → EDA필사 완료하여 debertav3base + LGBM 코드 필사 예정
- 한찬혁님 → 공모전 관계로 차주까지 불참

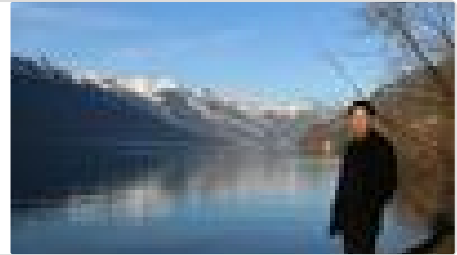
## 9월 21일

- 배나영님 → 저 오늘 약속이 있는걸 깜빡해서 배운 것 노선에 꼭 적어 공유하겠습니당... 암쏘소리...죄송합니다
- 임승준님 → debertav3base + LGBM 코드 등, 대회에 공유된 좋은 코드들 소개 bert-base 모델을 활용한 코드 필사 중
- 이희상님 → MobileBERT를 제대로 구현하기 위해 Teacher모델 구현 완료하였으나, 점수가 제대로 나오지 않아 당황하였음.  
Teacher Model에 deberta를 활용하였으나, 다른 모델로도 해볼 예정.

### [EN/한/中] MobileBERT implementation

Explore and run machine learning code with Kaggle Notebooks | Using data from multiple data sources

<https://www.kaggle.com/code/jasonheesanglee/en-mobilbert-implementation>



- 이성준님 → BERT 모델에 대해 한국어로 설명된 노트북을 공부해 보았고, BERT에 대해 완전히 이해를 함.
- 한찬혁님 → 공모전 발표로 인해 금주까지 불참

## 9월 21일 ~ 9월 26일

- 이희상님 → roberta와 T5를 활용하여 모델 구현 예정. 이제는 점수 상향에 집중 필요.
- 임승준님 → bert-base 베이스라인 코드 필사 및 한국어로 설명함.
- 이성준님 → bert-base 활용하여 제출하여 봄.

✓ [KR code comments] Beginner friendly [BERT]↔ - Version 5	0.568
Succeeded · 12h ago · prompt_id, text	
✓ [KR code comments] Beginner friendly [BERT]↔ - Version 4	1.849
Succeeded · 17h ago · input → prompt_text, text	
✓ [KR code comments] Beginner friendly [BERT]↔ - Version 3	0.572
Succeeded · 20h ago · Notebook [KR code comments] Beginner friendly [BERT]↔   Version 3	
✓ [KR code comments] Beginner friendly [BERT]↔ - Version 2	0.567
Succeeded · 1d ago · Notebook [KR code comments] Beginner friendly [BERT]↔   Version 2	
✓ [KR code comments] Beginner friendly [BERT]↔ - Version 1	0.584
Succeeded · 2d ago · Notebook [KR code comments] Beginner friendly [BERT]↔   Version 1	

version5 = prompt\_id, text

version4 = prompt\_test, text



version3 = prompt\_title, text

version2 = prompt\_question, text





version1 = text 만 사용

그런데 단순히 prompt\_id를 넣은 것도 점수가 좋았다

기준점이 명확하게 구별되는 것이 점수에 영향을 줄 수 있다는 생각을 하게되었다

	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 7</b> Succeeded · 1h ago · epochs = 5   prompt_id, text   Version 7	<b>0.555</b>	<input type="checkbox"/>
	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 6</b> Notebook Timeout · 17h ago · epoch = 5   prompt_id, text		

어제 강의에서 사전학습된 모델의 경우 에폭수가 많이 필요하지 않다는 것을 참고해서 5 에폭만 돌렸음

	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 8</b> Succeeded · 1h ago · epochs = 5   prompt_question, text   Version 8	<b>0.522</b>	<input type="checkbox"/>
	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 7</b> Succeeded · 2h ago · epochs = 5   prompt_id, text   Version 7	<b>0.555</b>	<input type="checkbox"/>
	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 6</b> Notebook Timeout · 18h ago · epoch = 5   prompt_id, text		
	<b>[KR code comments] Beginner friendly [BERT]↔ - Version 5</b> Succeeded · 4d ago · prompt_id, text	<b>0.568</b>	<input type="checkbox"/>

조금 더 수정