

8월 회고록

전체 돌아보기 :

처음 스터디를 구성하였음.

대회 취지, 데이터, 평가 Metric 등 대회에 대한 전반적인 이해를 함.

Code, Discussion 살펴보기를 통해 다른 실력자들은 어떤 방식으로 접근을 하는지, 그들이 고민하는 사항은 무엇인지 확인하였음.

NLP관련 블로그 및 논문을 살펴보며 자신이 사용할 전처리 방식, 모델에 대해서 공부해 보았음

8월 1일 첫 모임

• NLP 대회를 도전해 봐야겠다고 생각한 이유

이희상님

- 원래 언어에 관심이 많았고, 여러 대회를 진행하다 보니 자연스레 자연어 처리 관련 대회를 위주로 진행을 하게 되었다.

이성준님

- 챗봇에 관심이 있어 자연어를 접해보고 싶었다.

한찬혁님

- 심리 상담 쪽 창업을 준비함에 있어 추후 자연어 관련 스킬셋이 필요하여 미리 경험해 보고 싶었다.

임승준님

- IoT에 관심이 있고, 그것을 위해 자연어를 다루어 보고 싶었다.

배나영님

- 어떤 분야이던 한 번씩은 해보고 싶었다.
나중에 배울 분야이지만 미리 도전해보고 싶었다.

• 진행 내용

- 대회 데이터 셋에 대한 간략한 소개
- Evaluation Metric에 대한 고찰 (MCRMSE)

- PyTorch, TensorFlow에 대해 잠깐 이야기 함 (차주에 다시 공유)
- NLP Terminology (용어) 간략한 소개

8월 1일 ~ 8월 11일

1. 다른 사람이 적은 코드, 디스커션 읽어보기

Most Voted (가장 많은 좋아요 수)

Best Score (리더보드 (LB) 최고 점수)


Hotness(실시간 조회수 급상승)로 정렬 가능

- 코드를 읽으면서 원본 코드가 있는 코드일 경우 (Copied from ~~~ 이 있는 경우) 원본 코드와 어떠한 점에서 차이가 있는지 확인해 보기
- 가능하다면** 변경된 부분을 직접 바꾸어 보며 제출해 보기
- 이해가 안되는 코드나 디스커션이 있다면 슬랙이나 노션에 미리 공유 후 다음 회의 때 함께 논의해 보기

2. 하단의 유튜브 Playlist를 통해 NLP에서 사용하는 용어 정리해 보기

Natural Language Processing (NLP) Zero to Hero

Welcome to Zero to Hero for Natural Language Processing using TensorFlow! If you're not an expert on AI or ML, don't worry -- we're taking the concepts of NL...

 <https://www.youtube.com/playlist?list=PLQY2H8rRoyvzDbLUZKbudP-MFQZwNmU4S>



3. 가능하다면 타 EDA 코드 참고하면서 직접 데이터 가지고 놀아보기

4. 이번 주 공부 참고 키워드

- 패딩 (Padding)
- 임베딩 (Embedding)
- 마스킹 (Masking)
- 토큰나이징 (Tokenizing)

5. 참고자료

- NLP 기초 용어

1. [https://seokii.tistory.com/27#자연어처리\(NLP\)_기초_용어](https://seokii.tistory.com/27#자연어처리(NLP)_기초_용어)
2. <https://cold-soup.tistory.com/244>

[밑바닥부터 시작하는 데이터과학 - 자연어처리.pdf](#)

- NLP 전처리 관련

1. <https://velog.io/@cateto/NLP-텍스트전처리Text-preprocessing>

8월 11일

각자 지난 한 주 동안 어떤 것을 공부했는지 나누어 보았다.

배나영님, 한찬혁님, 이성준님은 하단의 캐글 노트를 참고하여 공부하였고, 임승준님은 이희상님의 EDA 코드를 참고하여 공부하였다.

배나영님 → GPT와 BERT의 차이에 대해서 특히 더 자세히 살펴보았다.

한찬혁님 → 여러가지 모델을 활용한 토큰나이징 방식에 대해 코드를 리뷰 해보았다.

이성준님 → 어떠한 모델들이 있는지 살펴보았고, 이에 더불어 텍스트 전처리 (토큰화, 패딩) 등에 대해 공부해 보았으며, 단어표현 (Word Embedding), 텍스트 유사도 등을 살펴보았다.

임승준님 → 지난 주 공유받았던 TensorFlow의 자연어 입문 영상을 바탕으로 상세히 공부해 보았고, 평가지표인 MCRMSE를 분석해보았다.

이희상님 → 지난 한 주간 오타 처리에 집중하였고, 다양한 오타처리 모델 중 symspell을 채택하여 사용하였다. 다만, running time 이슈로 인해 오타 detect에는 pyspellchecker를 사용하였고, 오타 → 정타 conversion에만 symspell을 사용하였다.

캐글 링크

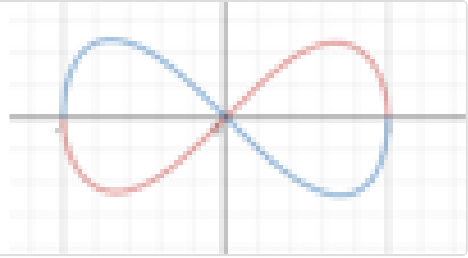
(베이스라인이 될 만한 노트북 참고해보기)

Understanding the Competition | CommonLit    

Understanding the Competition | CommonLit 

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

[k https://www.kaggle.com/code/ayushs9020/understanding-the-competition-commonlit](https://www.kaggle.com/code/ayushs9020/understanding-the-competition-commonlit)



참고 자료

<https://heekangpark.github.io/nlp/huggingface-bert>

알아두기

- nlp에 사용되는 정의 잘 알아두기
 - 예) corpus의 뜻은?

8월 11일 ~ 8월 17일

1. EDA와 전처리 해보기

EDA를 통해 본 대회 데이터에 어떠한 특성이 있는지 확인하여 보고,
확인한 사실에 기반하여 전처리 해보기!

2. BERT, DeBERTa 등 이 코드 참고하여 각 모델 별 특성 알아보기 (가능하면)

3. 이 블로그 읽어보기

추가 :: 참고할만한 블로그 모음

8월 17일

배나영님 → 아래 노트북을 살펴보고, 유사하게 접근해 보고자 한다.

<https://www.kaggle.com/code/sungeun1028/simple-eda-insights/notebook>

이성준님 → 데이터를 점검하였고, 텍스트 유사도 측정을 시도하였다.

이희상님 → 다양한 전처리 방식을 추가해 보았고, 전처리 코드를 간편화시켰다.

임승준님 → 여러 코드를 살펴보고, Baseline을 잡고 해당 코드를 정리해 보았다.

<https://www.kaggle.com/code/seungjunlim/notebookc187d58ee5>

한찬혁님 → 여러 사람의 EDA를 검토해 보았다.

8월 17일 ~ 8월 22일

이희상님 → EDA → 매 학생별로 복불한 summary가 있으면 해당 text의 중복을 제거와 도입하지 않았던 모델인 MobileBert 시도

이성준님 → EDA (이상치 점수 확인) 및 다양한 시각화 도전

임승준님 → <https://www.kaggle.com/code/abhishek123maurya/rsna-eda-understanding-the-data-first> 코드필사

배나영님 → 미팅 중 이야기했던 노트북 뜯어보고 더욱 세밀하게 접근하기

한찬혁님 → EDA와 모델 사용 방식 살펴보기

8월 22일

이성준님 → 추가 그래프를 그려보았으며, 이상치 확인을 해보았으나 크게 의미있는 결과는 없음

→ 서머리 길이 별 점수를 확인해 보았음

이희상님 → 모델 경량화를 위해 MobileBert를 도입하였으나 GPU Quota이슈로 LB 점수를 내보진 못하였음

한찬혁님 → 다른 사람들의 EDA를 꾸준히 살펴보았으나 타 공모전 진행으로 인해 자세히 살펴보진 못하였음

변웅진님 환영

배나영님, 임승준님 → 병가

8월 22일 ~ 8월 31일

이희상님 → EDA → 매 학생별로 복불한 summary가 있으면 해당 text의 중복을 제거. (지난주 도전과제에서 넘어옴)

- MobileBert 시도
- 오타처리한 Train / Test 데이터를 Content 점수 뽑는 것에 활용, 오타처리 안한 Train/Test 데이터를 Wording 점수 뽑는 것에 활용 예정

이성준님 → 어느정도 전처리한 데이터를 딥러닝 코드 필사하여 넣어보기

한찬혁님 → EDA 직접 해보기

변웅진님 → 대회 과제와 데이터 살펴보고 EDA를 통해 데이터의 특징 살펴보기

아래 찐코딩 NLP 강좌 들어보기

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. 🍷 - 4년차 현

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

8월 31일 저녁 미팅 예정