2023.02.28 미팅 자료

keyBERT, sentiBERT 테스트 결과 및 질문

팀원 : 배한성, 김은서, 조현아, 조유진

CONTENTS

1. keyBERT 테스트 결과
2. sentiBERT 테스트 결과
3. 질문
4. 캡스톤 일정 및 지원금

keyBERT 테스트 결과

테스트1

- BertForSequenceClassification 모델 fine-tuning 진행
  - BertForSequenceClassification.from_pretrained(“bert-base-multilingual-cased”, num_labels = 2)
  - 정확도 = 89%
  - 하지만, 긍/부정 예측 결과가 만족스럽지 않았음
    - train data의 문제라고 예상
- BertModel fine-tuning 진행
  - BertModel.from_pretrained(“monologg/kobert”)
  - 정확도 = 64%
  - 정확도를 올리기 위한 테스트 필요

```
======== Epoch 4 / 4 ========
Training...
  Batch   500  of  2,771.    Elapsed: 0:02:25.
  Batch 1,000  of  2,771.    Elapsed: 0:04:49.
  Batch 1,500  of  2,771.    Elapsed: 0:07:14.
  Batch 2,000  of  2,771.    Elapsed: 0:09:39.
  Batch 2,500  of  2,771.    Elapsed: 0:12:03.

  Average training loss: 0.18
  Training epoch took: 0:13:22

Running Validation...
  Accuracy: 0.89
  Validation took: 0:00:25
```

```
Training...
Epoch 10/13, Train Loss: 1.2171, Val Loss: 1.2048, Val Acc: 0.6457
```

테스트2

- keyBERT는 원작자에 따르면 Hugging Face API 중 sentence_transformers 모듈로 호출할 수 있는 BERT 모델을 추천하였습니다.
  - sentence_transformers 모듈로 생성할 수 있는 모델은 SBERT

- SBERT는 document embedding을 기존 BERT보다 효과적으로 추출할 수 있으며 evaluation 시 코사인 유사도 등을 구할 수 있는 모델
  - keyBERT는 키워드 임베딩과 문서 임베딩 사이의 코사인 유사성을 계산하여 키워드를 추출

- SBERT를 fine-tuning하기 위해서 BERT 모델에 NLI와 STS 데이터를 전이 학습하는 방식을 사용해야 한다.

테스트2

- 현재 github나 Hugging Face에서 KLUE 혹은 Kor-NLU 데이터 셋을 통해 전이 학습한 SBERT 모델을 제공

- 전이 학습을 시도해 보았으나 GPU 메모리 문제로 인해 batch_size를 줄여야 하는 문제 발생
  - 낮은 batch_size로 인해 튜토리얼에서 보여주는 값과 다른 결과가 발생하였다.
  - fine-tuning한 모델을 평가했을 때 상당히 낮은 값을 보여줌

- 결론적으로 keyBERT를 위한 SBERT를 local에서 fine-tuning하지 못하였다.

sentiBERT 테스트 결과

진행 상황

sst3 + bert-base-uncased (대소문자 구별 x) , sst3 + bert-base-multilingual-uncased

sst2 + bert-base-uncased , sst2 + bert-base-multilingual-uncased,

sst2 + kobert , sst3 + kobert 모델을 만들어서 가장 정확도가 높은 모델을 채용할 계획입니다.

현재 완성된 모델은 sst3 + bert-base-uncased 모델 (CPU 학습)

CUDA out of memory 이슈로 finetuning 단계에서 어려움을 겪고 있습니다.

gpu ram 문제

```
Traceback (most recent call last):
  File "finetune_on_pregenerated_sstphrase.py", line 391, in <module>
    main()
  File "finetune_on_pregenerated_sstphrase.py", line 352, in main
    loss = model(input_ids=input_ids, token_type_ids=segment_ids, attention_mask=input_mask, phrase_mask=phrase_mask, masked_lm_labels=lm_label_ids, next_sentence_label=is_next, graph_label
=graph_label, span=span, span_3=span_3)
  File "C:₩anaconda₩envs₩sentibert₩lib₩site-packages₩torch₩nn₩modules₩module.py", line 1130, in _call_impl
    return forward_call(*input, **kwargs)
  File "C:₩SentiBERT₩pytorch_pretrained_bert₩modeling_new.py", line 1088, in forward
    prediction_scores, seq_relationship_score, graph_score = self.cls(sequence_output, pooled_output, graph_output)
  File "C:₩anaconda₩envs₩sentibert₩lib₩site-packages₩torch₩nn₩modules₩module.py", line 1130, in _call_impl
    return forward_call(*input, **kwargs)
  File "C:₩SentiBERT₩pytorch_pretrained_bert₩modeling_new.py", line 614, in forward
    prediction_scores = self.predictions(sequence_output)
  File "C:₩anaconda₩envs₩sentibert₩lib₩site-packages₩torch₩nn₩modules₩module.py", line 1130, in _call_impl
    return forward_call(*input, **kwargs)
  File "C:₩SentiBERT₩pytorch_pretrained_bert₩modeling_new.py", line 571, in forward
    hidden_states = self.decoder(hidden_states) + self.bias
  File "C:₩anaconda₩envs₩sentibert₩lib₩site-packages₩torch₩nn₩modules₩module.py", line 1130, in _call_impl
    return forward_call(*input, **kwargs)
  File "C:₩anaconda₩envs₩sentibert₩lib₩site-packages₩torch₩nn₩modules₩linear.py", line 114, in forward
    return F.linear(input, self.weight, self.bias)
RuntimeError: CUDA out of memory. Tried to allocate 60.00 MiB (GPU 0; 4.00 GiB total capacity; 2.54 GiB already allocated; 36.50 MiB free; 2.56 GiB reserved in total by PyTorch) If reserved
 memory is >> allocated memory try setting max_split_size_mb to avoid fragmentation.  See documentation for Memory Management and PYTORCH_CUDA_ALLOC_CONF

(sentibert) C:₩SentiBERT₩examples₩lm_finetuning>
```

질문

질문

Chat Bot

- 구현 방향성이 궁금합니다.
- chatbot 학습을 위한 데이터를 수작업으로 만들어야 할까요?
    - 수작업으로 만들어야 할 경우, 데이터 개수, 데이터 내용은 어떻게 해야하나요?

설명 가능한 ai를 통한 추천

- 구현 방향성이 궁금합니다.
- 긍부정 bert와 문장의 주제를 판단하는 모델을 사용하여 구현하면 되는 것일까요?

질문

문장의 주제

- 추천과 속성에 대한 질문(색깔, 제품 완성도에 대한 질문)을 할 경우, 리뷰가 어떤 주제(속성)인지 모델이나 수작업으로 구현을 해야할 것으로 생각합니다.

  방법1. 수작업

  - data를 가공할 때 시간이 오래 걸릴 것으로 생각합니다. 왜냐하면 하나의 제품군(ex.노트북)에 대해서 80만개의 문장이 있기 때문입니다. 따라서 저희는 방법2로 하려고 합니다.

  방법2. 모델 이용

  - 시연 시 실시간으로 리뷰를 넣어 주제를 파악하도록 할 것인지, 아니면 db에 리뷰가 어떤 주제인지 미리 담아둔 상태에서 이후 과정을 진행할 것인지 결정해야 할 것 같습니다.
  - 따라서 model을 학습해서 해야할 것으로 저희는 생각합니다.

질문

속성에 대한 자연스러운 문장 구현

- 하나의 속성에 대한 여러 리뷰들 중 비슷한 문장은 하나의 문장으로 요약하는 등 문장들 간의 유사성을 판단하여 자연스러운 문장을 제공하는 모델이 필요할 것이라고 생각하고 있습니다.

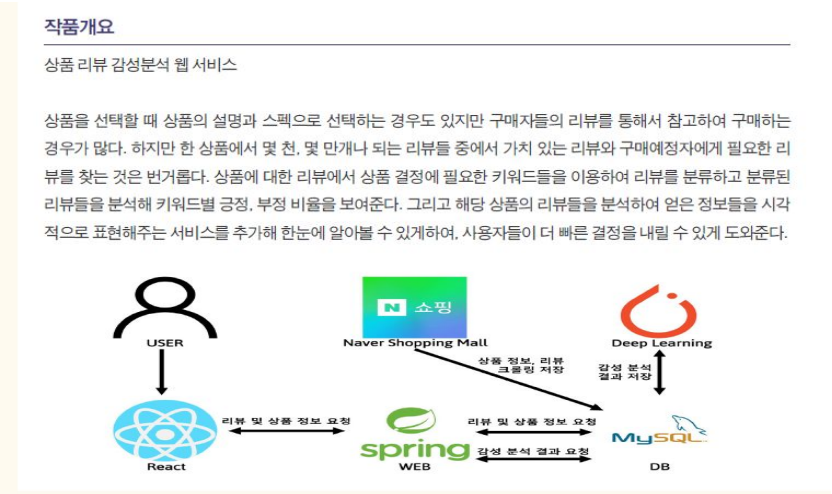
DB에 넣을 데이터 양의 수준

- DB에 데이터를 넣을 때, git에 있는 json 파일까지의 범위만을 넣을 것인지, 혹은 json 파일의 데이터 외에 리뷰의 주제나 다른 다양한 정보도 DB에 넣어야 할지 궁금합니다.
- DB에 어느 범위까지 넣는지에 따라 캡스톤에서 보여지는 기술 완성도 또한 다를 것 같기 때문에 사전에 어느정도 정해두는게 좋을 것 같다고 판단했습니다.

질문

현재 진행 중인 프로젝트와 거의 유사한 출품작 존재

- 기존 출품작과의 차별점
  - 설명 가능한 ai를 통한 추천 시스템 도입
  - 긍/부정 비율뿐만 아니라 요약본까지 제공
  - 챗봇 이용



- 2021 캡스톤에서의 출품작
  - https://github.com/cse-hansung/capstone2021/blob/main/bigdata.md
  - [캡스톤디자인] 엄빌리버블 - 텍스트마이닝 기반 가구 온라인 고객 리뷰 분석 및 추천 웹사이트 최종 발표 영상

캡스톤 일정 및 지원금

캡스톤 일정

주차	날짜	내용	강의방식
1주차	3월3일	강의 소개 및 팀 구성	온라인 동영상
2-13주차	지도교수와 상의	지도교수와 미팅	지도교수와 상의
14주차	6월2일	최종 발표/전시회	
15주차	6월9일	최종 결과물, 팀원 평가지 제출	지도교수와 상의

캡스톤 일정

- 팀 구성/지도교수 정보 제출, 3월 10일까지 e-class에 제출
- 프로젝트 제안서 (e-class로 제출) 3월 31일
- 최종 발표회 자료 제출 (제출 링크 e-class에 공지 예정)
  - github 저장소 주소 5월 26일
  - 발표 영상 주소 5월 26일 (2분 이내, 최대한 짧은 시간에 작품의 핵심을 강조하여 설명)
    - 시연 영상을 추가로 제출해도 됨(시간 제한 없음)
  - 팜플렛 제작을 위한 정보, 판넬 제작을 위한 파일 (제출 기한 추후 공지)
- 최종 결과물 (e-class로 제출): [양식][팀명]최종결과물.zip 참고
  - 프로젝트 제안서, 최종보고서, 최종발표자료
  - 주간보고서(2-13주차)
  - 소스코드(github 주소와 branch/tag를 명시하는 것으로 대체 가능)
  - 제출 기한: 6월 9일

캡스톤 지원금

10. 프로젝트 지원금

- 창의융합교육원(https://capstone.hansung.ac.kr/)에서 캡스톤디자인 기본지원금 지급 예정
  - 영수증 처리 방법 미리 확인 후 사용할 것
  - 예전에는 학교 사업자 번호로 발행된 현금 영수증만 가능
  - 올해(2023년)는 학생 개인 카드 사용 가능하다고 함
- 기본지원금 이외에 팀별로 추가적인 비용이 불가피할 경우 지도교수와 상의하여 청구할 수 있음
  - (학과 실험실습비로 추가 지원, 절차는 담당 조교에게 별도로 문의할 것)
  - 되도록 교내(컴퓨터공학부)에서 조달 가능한 장비를 사용하고, 부득이 새로운 장비가 필요한 경우에는 지도교수님과 상의하여 조교를 통해 구매추진
  - 지도교수님과 상의하였어도 예산 문제나 실험실습비로 구매 불가한 장비일 경우 거절될 수 있음