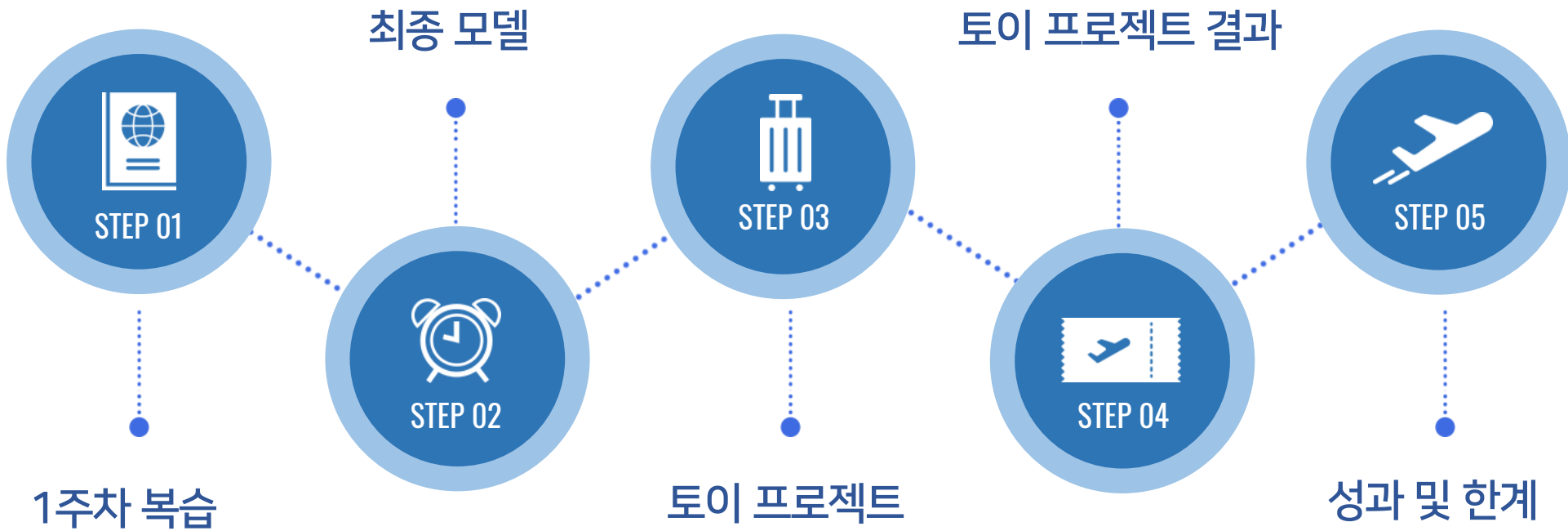




국내 관광 활성화를 위한 카테고리 분류

김예찬 / 박시언 / 박윤아 / 정승민 / 김민

목차



01 1주차 복습





01 1주차 복습



1. 주제



코로나 확산세의 감소로
관광업의 재부흥!



국내 관광 활성화에 도움이 될 방안이 뭐가 있을까?



01 1주차 복습



1. 주제



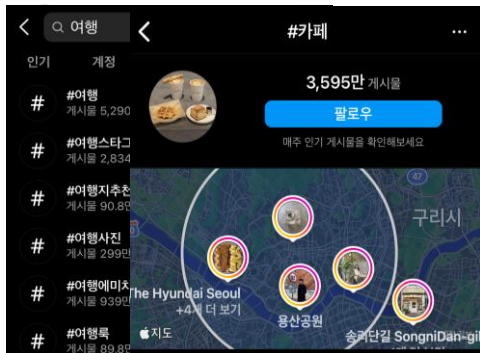
코로나 확산세의 감소로
관광업의 재부흥!



국내 관광 활성화에 도움이 될 방안이 뭐가 있을까?



최근 **해시태그**를 통한 여행, 관광 관련 검색과 홍보가
활발하게 이루어지고 있다는 점을 주목!





01 1주차 복습



1. 주제

기존 해시태그 생성 모델의 단점
리뷰 단어 위주의 단순한 해시태그 구성



이미지와 텍스트를 기반으로 카테고리 대, 중, 소 분류



분류된 데이터를 바탕으로 해시태그 생성



카테고리의 계층적 정보까지 포함한 해시태그로 더 많은 정보 제공 !



01 1주차 복습



2. 텍스트 전처리 _ 불용어 제거



지난 주 설명에 빠졌던 텍스트 전처리 과정을 소개하겠습니다~^^;;



id	img_path	overview	cat1	cat2	cat3
TRAIN_00004	./image/train/...	※ 영업시간 10:30 ~ 20:30WnWn3대에 걸쳐...	음식	음식점	한식
TRAIN_00056	./image/train/...	* 송강 정철의 설화가 있는 곳, 환벽당 ·조대 * ...	인문	역사관광지	유적지/사적지
TRAIN_00057	./image/train/...	※코로나 19 감염 확산방지를 위해...	레포트	육상레포트	야영장, 오토캠핑장
TRAIN_00114	./image/train/...	* 남도 국악의 자긍심, 익산 국악원 * ...	인문	문화시설	문화전수시설
		...			

원본 데이터 overview의 text에 불용어가 존재함



01 1주차 복습



2. 텍스트 전처리 _ 불용어 제거



지난 주 설명에 빠졌던 텍스트 전처리 과정을 소개하겠습니다~^^;;



id	img_path	overview	cat1	cat2	cat3
TRAIN_00004	./image/train/...	※ 영업시간 10:30 ~ 20:30 WnWn 3대에 걸쳐...	음식	음식점	한식
TRAIN_00056	./image/train/...	* 송강 정철의 설화가 있는 곳, 환벽당 ·조대 *
...	인문	역사관광지	유적지/사적지
TRAIN_00057	./image/train/...	 ※코로나 19 감염 확산방지를 위해...	레포트	육상레포트	야영장, 오토캠핑장
TRAIN_00114	./image/train/...	* 남도 국악의 자긍심, 익산 국악원 *

...	인문	문화시설	문화전수시설
		...			

데이터를 크롤링하는 과정에서 함께 저장된 html 구조 문법이기 때문!



01 1주차 복습



2. 텍스트 전처리 _ 불용어 제거



```
# 불용어 제거
import re

for i in tqdm(range(ex.shape[0])):
    text = ex.iloc[i, 2]
    text = text.replace('<br>', '').replace('<br />', '').replace('*', '').replace('<br />', '')
    text = text.replace('&nbsp;', '').replace('<a />', '').replace('<a />', '').replace('₩n', '').replace('₩t', '')
    #text = re.sub('[^a-zA-Z0-9ㄱ-ㅣ가-힣]', ' ', text)
    ex.iloc[i, 2] = text
```

제거한 불용어 리스트

 	<br /	<a />
 	 	₩n
*	<a/>	₩t



01 1주차 복습



2. 텍스트 전처리 _ 불용어 제거



```
# 불용어 제거
import re

for i in tqdm(range(ex.shape[0])):
    text = ex.iloc[i, 2]
    text = text.replace('<br>', '').replace('<br />', '').replace('*', '').replace('<br />', '')
    text = text.replace('&nbsp;', '').replace('<a>', '').replace('<a />', '').replace('¶', '').replace('¶t', '')
    #text = re.sub('[^a-zA-Z0-9ㄱ-힣]', ' ', text)
    ex.iloc[i, 2] = text
```

제거한 불용어 리스트

 <a />

 ¶ ¶t
* <a /> ¶t

단위 등 해석에 유의미한 특수문자는 유지하기 위하여
정규표현식은 생략하고 텍스트 전처리 진행



01 1주차 복습



2. 텍스트 전처리 _ 불용어 제거



id	img_path	overview	cat1	cat2	cat3
TRAIN_00004	./image/train/...	※ 영업시간 10:30 ~ 20:30 3대에 걸쳐 아귀만을...	음식	음식점	한식
TRAIN_00056	./image/train/...	송강 정철의 설화가 있는 곳, 환벽당 · 조대 >>...	인문	역사관광지	유적지/사적지
TRAIN_00057	./image/train/...	※코로나 19 감염 확산방지를 위해 개장 잠정...	레포트	육상레포트	야영장, 오토캠핑장
TRAIN_00114	./image/train/...	남도 국악의 자긍심, 익산 국악원 >>익산 국악원의...	인문	문화시설	문화전수시설
		...			

불용어 제거 이후의 overview를 보면 불용어가 잘 제거 되었음을 확인할 수 있음



01 1주차 복습



3. 텍스트 전처리 _ 토큰화



텍스트 데이터를 처리하는 각 모델의 tokenizer를 활용하여 토큰화 진행

#roberta 토큰화

```
for text in df['overview'].head():  
    rbt = tokenizer.tokenize(text)  
    print(rbt)
```

```
['소', '##안', '##항', '##은', '조용', '##한', '섬', '##으로', '인근',  
['경기도', '이천', '##시', '모', '##가', '##면', '##에', '있', '##는',  
['금오', '##산성', '##숲', '##불', '##갈비', '##는', '한우', '##고기',  
['철판', '위', '##에서', '요리', '##하', '##는', '안동', '##찜', '##닭',  
['영업시간', '10', '30', '20', '30', '##3', '##대', '##에', '걸쳐', '('
```



01 1주차 복습



3. 텍스트 전처리 _ 토큰화



텍스트 데이터를 처리하는 각 모델의 tokenizer를 활용하여 토큰화 진행

```
# kobert 토큰화
for i in range(train_data.shape[0]):
    tokenized = kobert_tokenizer.tokenize(train_data.iloc[i].overview)
    train_data.at[i, 'overview'] = tokenized
train_data.overview
```

0	[_소, 안, 항, 은, _조, 용, 한, _섬, 으로, _인근, 해, 안, 이, ...]
1	[_경기도, _이, 천, 시, _모, 가, 면, 에, _있는, _골프장, 으로, _...
2	[_금, 오, 산, 성, 숲, 불, 갈, 비, 는, _한, 우, 고, 기, 만, 을...
3	[_철, 판, _위, 에서, _요리, 하는, _안, 동, 찜, 닭, 을, _맛, 불...
4	[_영업, 시간, _10, _30, _20, _30, _3, 대, 에, _걸쳐, _...

02 최종 모델





02 최종 모델



1. 라벨링 수정

데이터를 다시 한 번 살펴보던 도중,,,

“ 서울 종로 계동길의 작은 골목에 자리한 ‘멀티스페이스 곳’ 은 80여 년 된 한옥을 개조한 게스트 하우스다. 전통 침구가 깔린 온돌방은 외국인뿐 아니라 우리나라 사람에게도... ”



한옥 스테이? 게스트하우스? 펜션? 고택?

“ ‘장돌 해변’은 바람아래해변에서 10여분 정도 소요되는, 해변의 폭이 그리 크지 않은 아늑하고 조용한 해변이다. 주위가 논경지와 산으로 이루어져 있어 야영하기엔 그리...”



해수욕장? 해안절경? 농.산.어촌 체험?



02 최종 모델



1. 라벨링 수정



“ 서울 종로 계동길의 작은 골목에 자리한 ‘페이스 곳’ 은 80여 년 된 한옥을 개조한 게스트하우스다. 전통 침구가 깔린 온돌방은 외국인뿐 아니라 우리나라 사람에게도... ”

한국관광공사에서 제공된 데이터의 분류 기준은 전반적으로 너무 모호함

= 분류 모델이 학습에 어려움을 겪을 수 있음

한옥 스테이? 게스트하우스? 펜션? 고택?

“ ‘장돌 해변’은 바람아래해변에서 10여분 정도 소요되는, 해변의 폭이 그리 크지 않은 아늑하고 조용한 해변이다. 주위가 논경지와 산으로 이루어져 있어 야영하기엔 그리... ”



해수욕장? 해안절경? 농.산.어촌 체험?



02 최종 모델



1. 라벨링 수정

“ 서울 종로 계동길의 작은 골목에 자리한 ‘**멀딩스페이스** 곳’ 은 80여 년 된 한옥을 개조한 게스트하우스다. 전통 침구가 깔린 온돌방은 **외국인**뿐 아니라 우리나라 사람에게도...”



한국관광공사에서 제공된 데이터의 분류 기준은 전반적으로 너무 모호함

= 분류 모델이 학습에 어려움을 겪을 수 있음

한옥 스테이? 게스트하우스? 펜션? 고택?



“ ‘장돌 해변’은 바람아래해변에서 10여분 정도 소요되는, 해변의 폭이 그리 크지 않은 아늑하고 조용한 해변 **따라서** 구분이 모호한 카테고리에 대해 야영하기엔 그리...”

일관된 기준을 설정하여 라벨 수정을 진행함!



해수욕장? 해안절경? 농.산.어촌 체험?



02 최종 모델



1. 라벨링 수정



Cat3 카테고리 재분류 기준

1. '컨벤션'과 '전시회'를 모두 '전시회'로 통일
2. '해수욕장' 카테고리 중 '섬'에 해당하는 데이터 재분류
3. '상설시장' 카테고리 중 '5일장'에 해당하는 데이터 재분류
4. '호텔' 카테고리 추가
5. '한옥 스테이' 및 '모텔' 카테고리를 '호텔'로 재분류
6. '홈스테이' 카테고리 전부 재분류 후 삭제



02 최종 모델



1. 라벨링 수정



Train set

Id	img_path	overview	cat1	cat2	cat3	수정된 cat3
TRAIN_0	./image/train/...	전남 강진군의 달빛한옥마을은 2013년 7월에 탄생했다. 28가구 77명의 주민이 생활하는 숙박...	숙박	숙박시설	홈스테이	한옥스테이
TRAIN_0	./image/train/...	# 본 업소는 외국인관광 도시민박업으로 외국인만 이용이 가능하며 내국인은 이용할 수 없습니다. 교대게스트하우스는 지하철 2호선과 3호선이 다니는...	숙박	숙박시설	홈스테이	게스트하우스
TRAIN_0	./image/train/...	충남 보령시 이광명 고택은 구한말에 지어진 한옥으로 문화재적 가치가 높다. 대한제국 황제인 고종의 다섯째...	숙박	숙박시설	홈스테이	한옥스테이
TRAIN_0	./image/train/...	전남 보성군 복내면 청염당은 한옥 숙박 체험을 진행한다. 복내면은 수려한 풍경으로 유명한데,...	숙박	숙박시설	홈스테이	한옥스테이
TRAIN_0	./image/train/...	월강고택(최씨고가)는 한국에서 가장 아름다운 마을 제1호인 남사예담촌의 중앙에 있는 경남 문화재 자료 제117호로 지정된 ...	숙박	숙박시설	홈스테이	고택



02 최종 모델



1. 라벨링 수정

Test set

	A	B	C	D
1	id	img_path	overview	cat3
2	TEST_000C	/image/te	신선한 재료로 정성을 다해 만들었다. 늘 변함없는 맛과 서비스로 모실것을 약속한다.	한식
3	TEST_000C	/image/te	청청한 해역 등량만과 울포해수욕장이 한눈에 내려다 보이는 위치에 있으며, 막 잡은 어류로 만든	한식
4	TEST_000C	/image/te	장터설렁탕은 남녀노소 누구나 즐길 수 있는 전통 건강식으로 좋은 재료와 전통 조리방식을 고수	한식
5	TEST_000C	/image/te	다양한 형태의 청소년수련활동을 제공함으로써 청소년들이 민주사회의 주역이 될 수 있도록 건전	수련시설
6	TEST_000C	/image/te	팔공산은 경산시의 북쪽에 위치한 해발 1192.3 m의 높은 산으로 신라시대에는 중악, 부악으로 알	산
7	TEST_000C	/image/te	30여 년의 세월이 느껴지는 실내 분위기가 냉면 맛을 더욱 살린다.	한식
8	TEST_000C	/image/te	코리달리스는 경기도 가평에 위치하고 있는 카페이다. 청명하고 맑은 호수 전경이 아름다운 카페	바/카페
9	TEST_000C	/image/te	신선한 닭갈비를 공급해서 판매하는 곳이다. 대표메뉴는 치즈닭갈비다. 강원도 원주시에 있는 한식	한식
10	TEST_000C	/image/te	정유재란(1597年) 당시 육전에서 패퇴한 왜군선봉장 宇喜多秀家(우끼다히데이)와 藤堂高虎(도도	성
11	TEST_000C	/image/te	약 50여개의 점포가 있는 골목형 시장이다. 시장 내에 개성 있는 인테리어의 카레전문점과 카페가	상설시장

Test set에 대한 정확도 및 f1-score를 계산하기 위해

Test set에 대한 예측 결과를 Train set 재분류 기준에 맞춰 라벨을 수정함



02 최종 모델



1. 라벨링 수정



자세한 내용은 범주팀 3주차 클린업 참고!

F1-Score 란?

	A	B		D
1	id	img_path overview		cat3
2	TEST_000C./image/te	신선한 재료만을 사용해 만든다는 마가사시음료의 맛을 약속한다.		한식
3	TEST_000C./image/te	청청한 해산물과 달고메가 특징인 한산해물탕이며, 막 잡은 어류로 만든		한식
4	TEST_000C./image/te	장터설렁탕은 남녀노소 누구나 즐길 수 있는 전통 건강식으로 좋은 재료와 전통 조리방식을 고수		한식
5	TEST_000C./image/te	다양한 형태의 청소년수련활동을 제공함으로써 청소년들이 민주시민의 주역이 될 수 있도록 건전		수련시설
6	TEST_000C./image/te	한국의 전통산시의 북쪽에 위치한 해발 1163.3m의 높은 산으로 산사시내에는 음악, 부악으로 알		산
7	TEST_000C./image/te	30여 년의 세월이 느껴지는 실내 분위기가 냉면 맛을		한식
8	TEST_000C./image/te	코리달리스는 카페이다. 청명하고 맑은 호수 전경이 아름다운 카페5바/카페		
9	TEST_000C./image/te	신선한 닭갈비를 공급해서 판매하는 곳이다. 대표메뉴는 치즈닭갈비다. 강원도 원주시에 있는 한스		한식
10	TEST_000C./image/te	정유재란(1597年) 당시 육전에서 패퇴한 왜군선봉장 宇喜多秀家(우끼다히데이)와 藤堂高虎(도도 正		
11	TEST_000C./image/te	각 50여개의 점포가 있는 골목명 시장이다. 시장 내에 개장 있는 인테리아의 가래진문점과 카페가		상설시장



1주차에서 언급한 클래스 불균형 문제를 고려하여

각 클래스에 속하는 표본의 개수로 계산한 가중평균을 활용하는

weighted avg f1-score 사용

Test set에 대한 예측 결과를 기준으로 train set에 분류 기준에 맞춰 라벨링



02 최종 모델



2. 모델시도



모델의 성능을 위한 시도들

시도1) 앞뒤 문장 추출

시도2) KeyBERT

시도3) 원문 텍스트 사용



KoBERT

RoBERTa

Multimodal (ViT+RoBERTa)



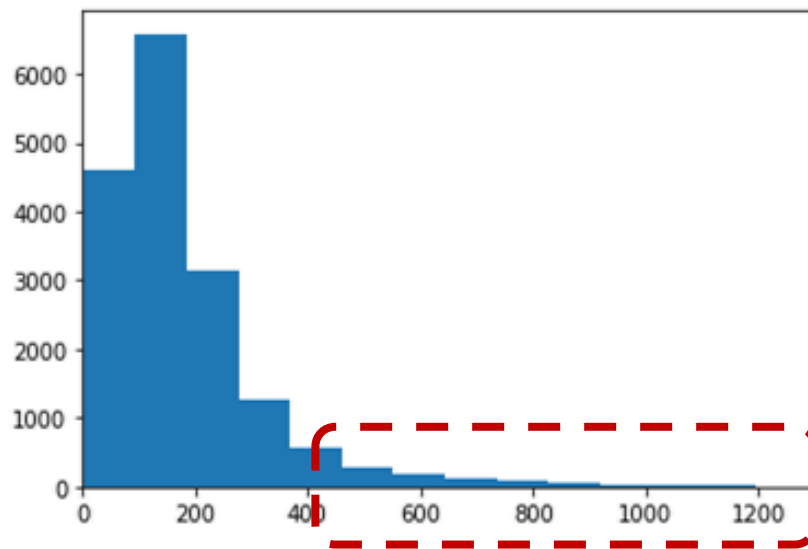
02 최종 모델



2. 모델시도1

정유재란(1597年) 당시 육전에서 패퇴한 왜군선봉장 宇喜多秀家(우끼다히데이)와 藤堂高虎(도도 다카토라)가 호남을 공략하기 위한 전진기지 겸 최후 방어기지로 삼기 위해 3개월간 쌓은 토석성으로 왜장 소서행장(小西行長)이 이끈 1만 4천여 명의 왜병이 주둔하여 조·명연합군과 두차례에 걸쳐 격전을 벌였던 곳으로 남해안 26 왜성 중 유일하게 한 곳만 남아 있다. 순천왜성은 수륙 요충지로서 성곽 규모가 120,595m²(36,480평), 외성 2,502m, 내성 1,342m로 외곽성(토석성)3개, 본성(석성) 3첩, 성문 12개로 축조된 성곽으로 검단산성쪽의 육지부를 파서 바닷물이 차도록 섬처럼 만들고 연결 다리가 물에 뜨게 하여 예교, 왜교성이라 하며 일인들은 순천성이라 부르고 있다.

임진란 패인이 전라도 의병과 수군의 용전에 있었다고 보고 전라도를 철저히 공략키 위해 풍신수길의 야심에 따라 전라도 각처에 진지를 구축해 공세를 강화하였으나 무술년(1598년) 8월 그가 급사 후 왜성에 주둔해 있던 침략 최정에 부대인 소서행장 왜군과 조·명 수륙 연합군 사이에 2개월에 걸친 최후·최대의 격전을 펼친 곳이다. 순천시가지에서 여수쪽으로 6km쯤 가다가 왼쪽으로 6km를 가면 200여호가 사는 신성리 마을과 이충무공을 배향한 충무사가 있고 남쪽 200m 지점 광양만에 접한 나지막한 송림에 위치한 왜성은 유정.권율이 이끄는 육군 3만6천, 진린, 이순신이 이끄는 수군 1만 5천병력이 왜성을 비롯 장도등을 오가며 왜군을 격멸했고 이충무공이 27일간을 머물면서 전사 하루 전 소서행장을 노랑 앞바다로 유인하여 대첩을 거둔 유서 깊은 전적지로서 자라나는 후손들에게 역사의 산교육장이기도 하다.



데이터에 길이가 매우 긴 텍스트 일부 존재



02 최종 모델



2. 모델시도1



정유재란(1597年) 당시 육전에서 패퇴한 왜군선봉장 宇喜多秀家(우끼다히데이)와 藤堂高虎(도도 다카토라)가 호남을 공략하기 위한 전진기지 겸 최후 방어기지로 삼기 위해 3개월간 쌓은 토석성으로 왜장 소서행장(小西行長)이 이끈 1만 4천여 명의 왜병이 주둔하여 조·명연합군과 두차례에 걸쳐 격전을 벌였던 곳으로 남해안 26 왜성 중 유일하게 한 곳만 남아 있다. 순천왜성은 수륙 요충지로서 성곽 규모가 120,595m²(36,480평), 외성 2,502m, 내성 1,342m로 외곽성(토석성)3개, 본성(석성) 3첩, 성문 12개로 축조된 성곽으로 검단산성쪽의 육지부를 파서 바닷물이 차도록 섬처럼 만들고 연결 다리가 물에 뜨게 하여 예교, 왜교성이라 하였다.

임진란 패인이 전라도 의병과 수군의 용전에 있었다고 보고 전라도를 철저히 에 따라 전라도 각처에 진지를 구축해 공세를 강화하였으나 무술년(1598년) 8월 그가 급사 후 왜성에 주둔해 있던 침략 최정예 부대인 소서행장 왜군과 조·명 수륙 연합군 사이에 2개월에 걸친 최후·최대의 격전을 펼친 곳이다. 순천시가지에서 여수쪽으로 6km쯤 가다가 왼쪽으로 6km를 가면 200여호가 사는 신성리 마을과 이충무공을 배향한 충무사가 있고 남쪽 200m 지점 광양만에 접한 나지막한 송림에 위치한 왜성은 유정.권율이 이끄는 육군 3만6천, 진린, 이순신이 이끄는 수군 1만 5천병력이 왜성을 비롯 장도등을 오가며 왜군을 격멸했고 이충무공이 27일간을 머물면서 전사 하루 전 소서행장을 노랑 앞바다로 유인하여 대첩을 거둔 유서 깊은 전적지로서 자라나는 후손들에게 역사의 산교육장이기도 하다.



Padding의 max length를 256으로 설정했기 때문에
뒤쪽에 위치한 텍스트는 학습에 반영이 되지 않음



길이가 매우 긴 텍스트 일부 존재



02 최종 모델



2. 모델시도1



< 가정 1 >

텍스트가 긴 경우 가운데보다 초반과 후반부에 분류에 핵심적인 내용이 많을 것이다.



문장의 개수가 5개 이상일 경우
앞 문장 3개와 뒷 문장 2개만 추출하여 학습시켜보자!



“ 부여안방마님은 충남 부여군의 유일한 **한옥 체험 숙소**다. 안채 상량을 기준으로 1896년 지어진 **한옥을 복원**했다. 안채와 별채, 사랑채, 행랑채 등으로 이뤄진 한옥이다.

...

가족 또는 동호회 모임, 작은 음악회 등의 장소로도 **대관**한다. 한옥 카페에서는 직접 만든 쌍화차, 대추차, 한방차 등을 판매한다. ”



02 최종 모델



2. 모델시도1

절망



"부여안방마님은 충남 부여군의 유일한 한옥 안채 상량을 기준으로 1896년 지어진 한옥을 복원했다. 안채와 별채, 사랑채, 행랑채 등으로 구성되었다."

...

가족 또는 동호회 모임, 작은 음악회 등의 장소로도 대관한다. 한옥 카페에서는 직접 만든 쌍화차, 대추차, 한방차 등을 판매한다."

텍스트 길이 조정 전 데이터로 학습시켰을 때보다 **성능 하락**

< 걱정 >

텍스트가 긴 경우 가운데보다 초반과 후반부에 분포에 해시적인 내용이 많을 것이다.

텍스트 중반부의 내용도 모델의 분류 성능에

유의미한 역할을 했을 것...

문장의 개수가 5개 이상일 경우

앞 문장 3개와 뒷 문장 2개만 추출하여 학습시켜보자!



02 최종 모델



3. 모델시도2



<가정 2>

핵심 내용만을 추출하여 모델을 학습시키면 분류 성능이 좋아질 것이다.



KeyBERT를 활용하여 키워드를 추출한 후
모델을 학습시켜 카테고리 예측을 진행해보자 !





02 최종 모델

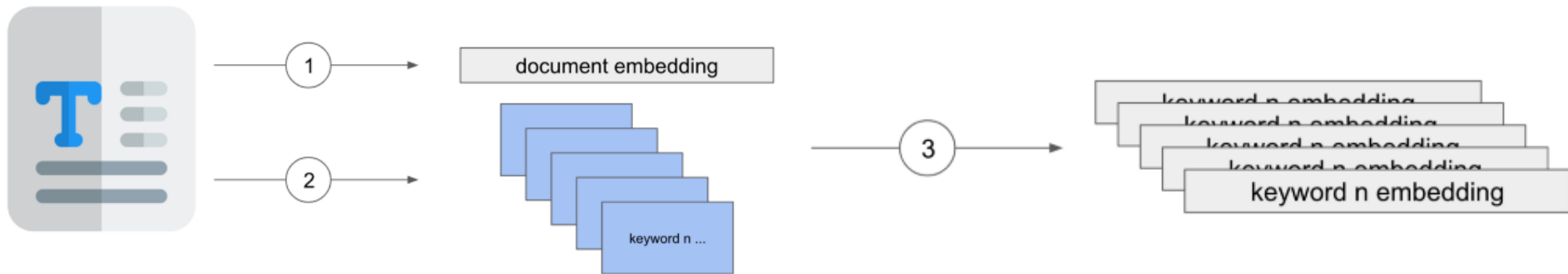


3. 모델시도2



KeyBERT

BERT 기반 키워드 추출 모델로,
BERT를 통해 문서의 주제를 파악하고, bag-of-words 기법으로 n-gram 임베딩
이후 코사인 유사도를 계산하여 키워드 추출



총 10개의 키워드를 추출하고 이를 통해 학습을 진행!!



02 최종 모델



3. 모델시도2

절망



<가정 2>

원본 텍스트 데이터로 학습시켰을 때보다 **성능 하락**
핵심 내용만을 추출하여 모델을 학습시키면 분류 성능이 좋아질 것이다.

문장의 문맥적인 의미가 모델 학습에 반영되지
않은 점이 성능에 부정적으로 작용했을 것 ...
KeyBERT를 활용하여 키워드로 모델을 학습시킨 후
카테고리 예측을 진행해보자!





02 최종 모델



3. 모델시도3



모델 성능을 향상시키기 위해 다양한 시도를 했음에도 불구하고
좋은 성능이 나오지 않았음



텍스트는 원문 그대로를 이용하여 카테고리 예측을 진행하기로 결정





02 최종 모델

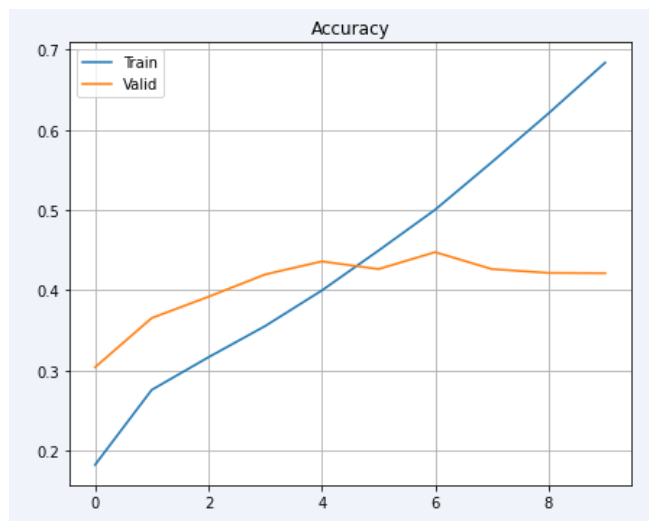


4. 모델 성능 비교

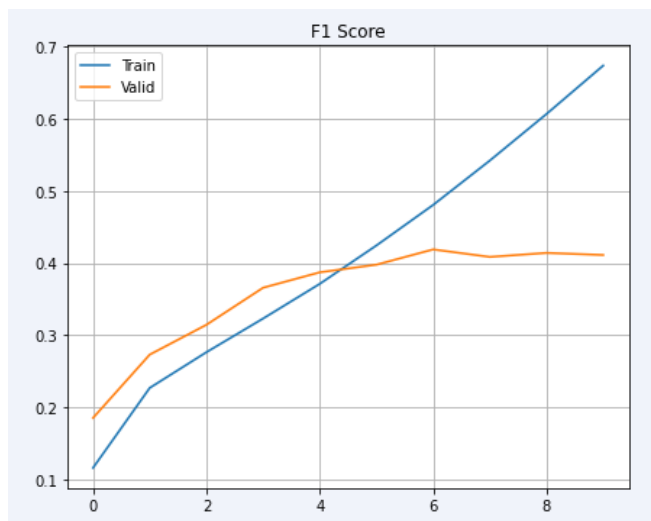
KoBERT

Accuracy = 0.3909

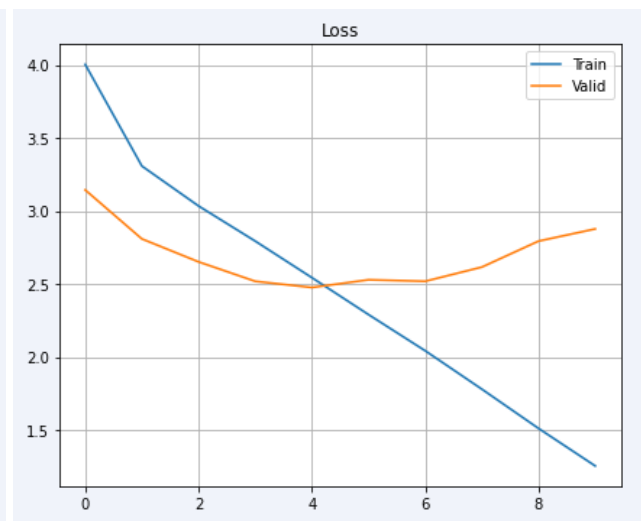
F1-score = 0.3858



Accuracy



F1 - score



Loss



02 최종 모델

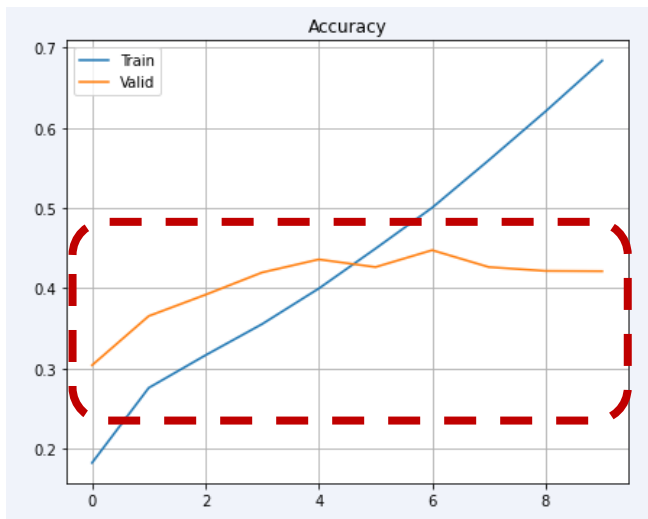


4. 모델 성능 비교

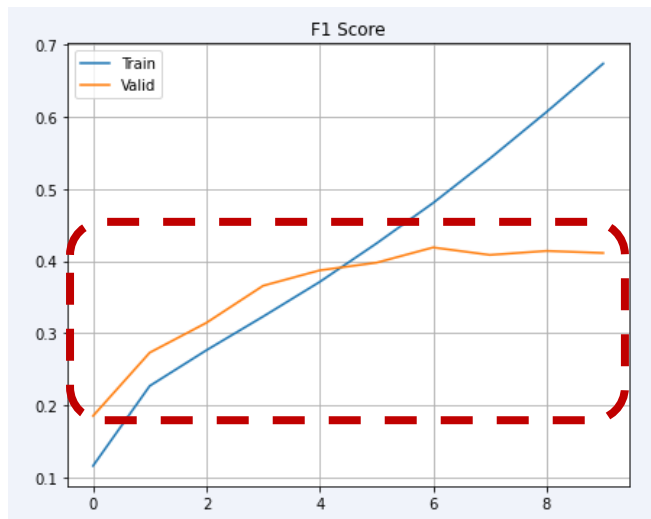


KoBERT

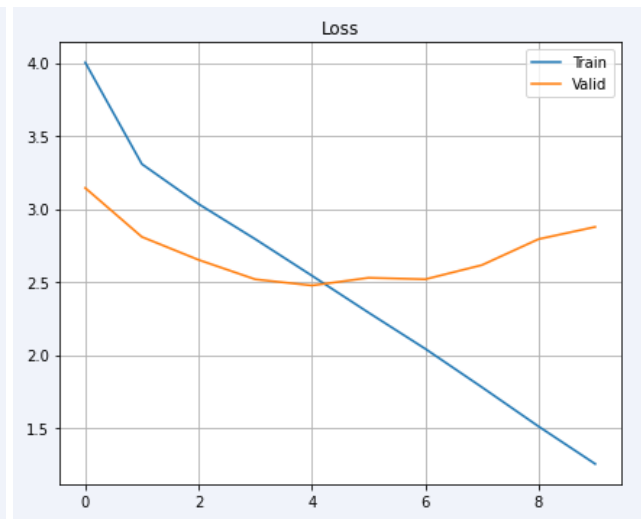
0.5 미만의 낮은 validation score 기록, Test set에 대해서도 가장 낮은 test score 기록



Accuracy



F1 - score



Loss



02 최종 모델

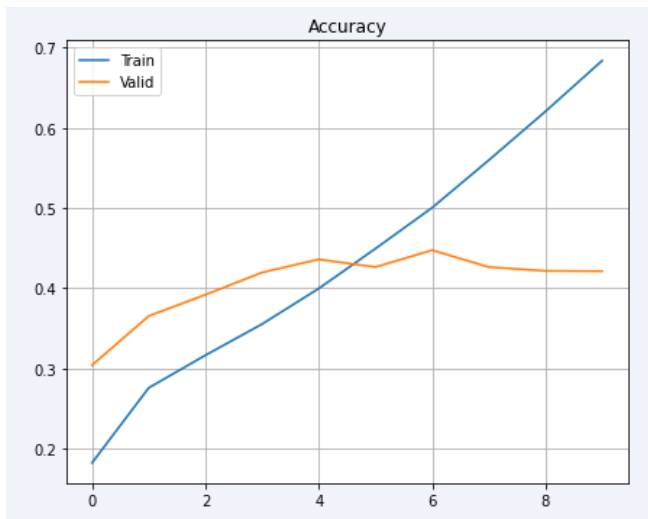


4. 모델 성능 비교

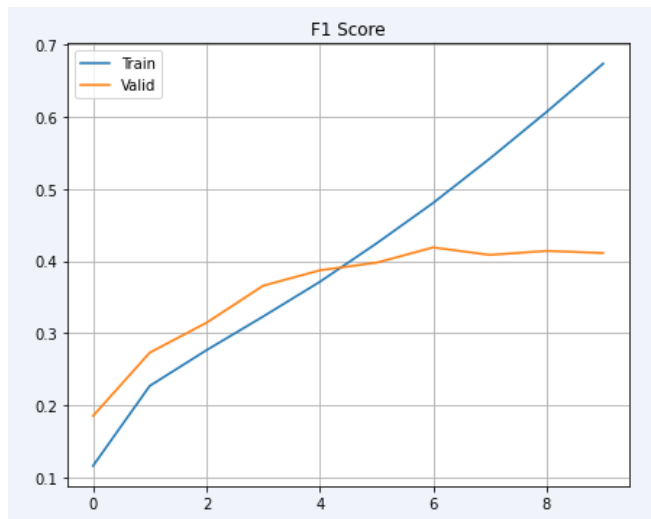


KoBERT

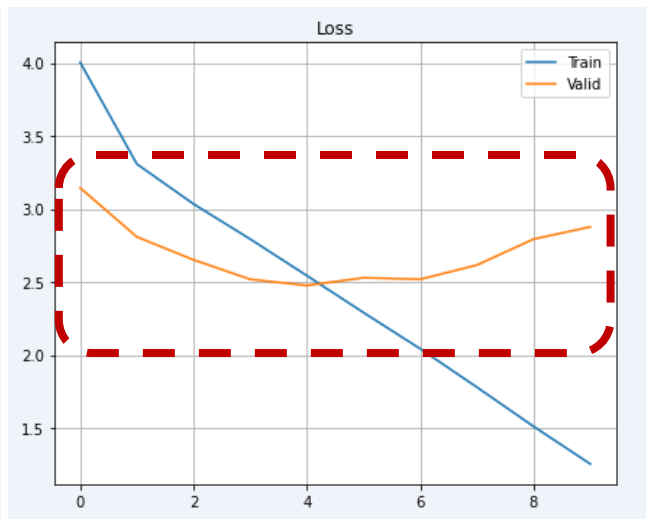
Train 데이터는 Loss가 감소하지만 Valid는 높은 Loss에 수렴



Accuracy



F1 - score



Loss



02 최종 모델

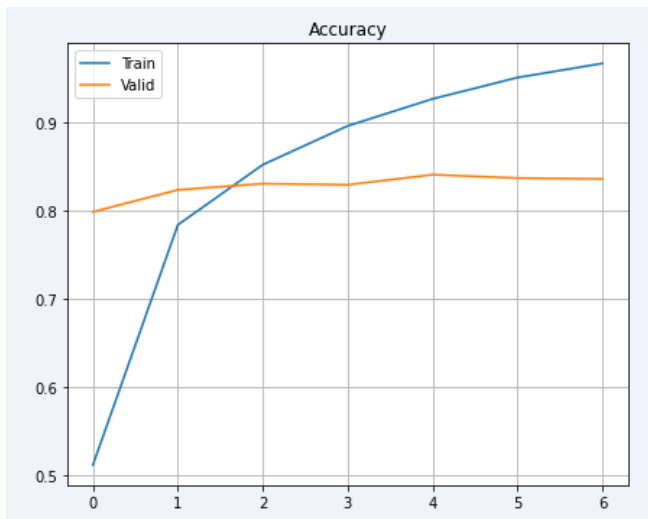


4. 모델 성능 비교

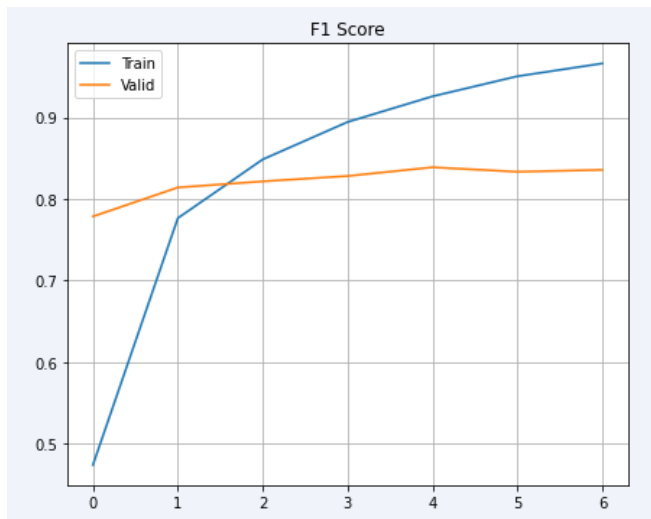
RoBERTa

Accuracy = 0.8733

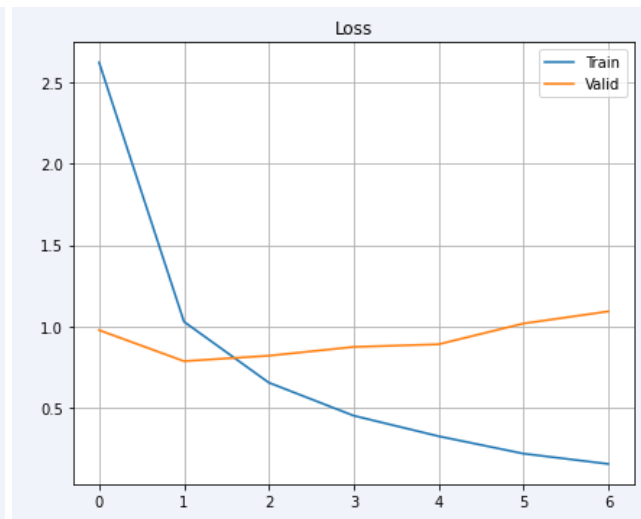
F1-score = 0.8740



Accuracy



F1 - score



Loss



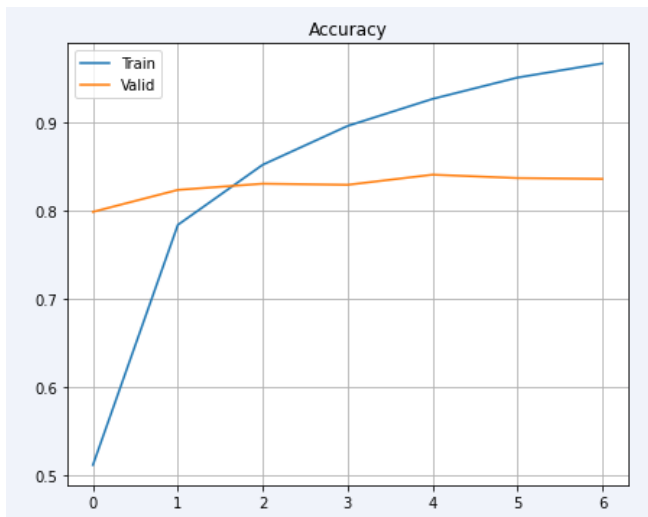
02 최종 모델



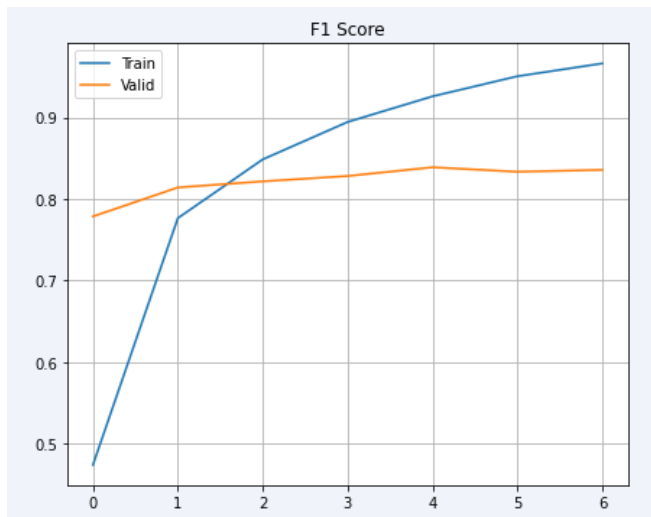
4. 모델 성능 비교

RoBERTa

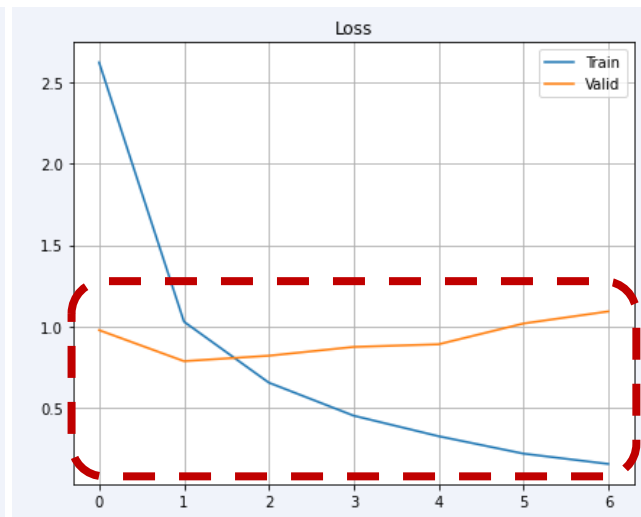
꾸준히 증가하는 validation score, 비교적 낮은 loss 기록



Accuracy



F1 - score



Loss



02 최종 모델



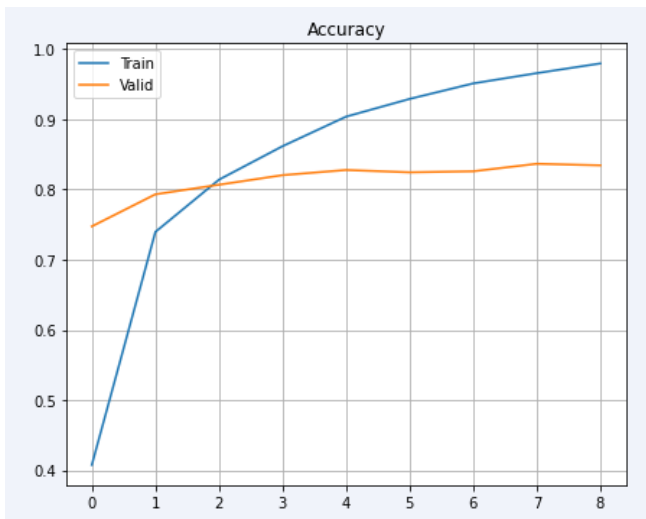
4. 모델 성능 비교



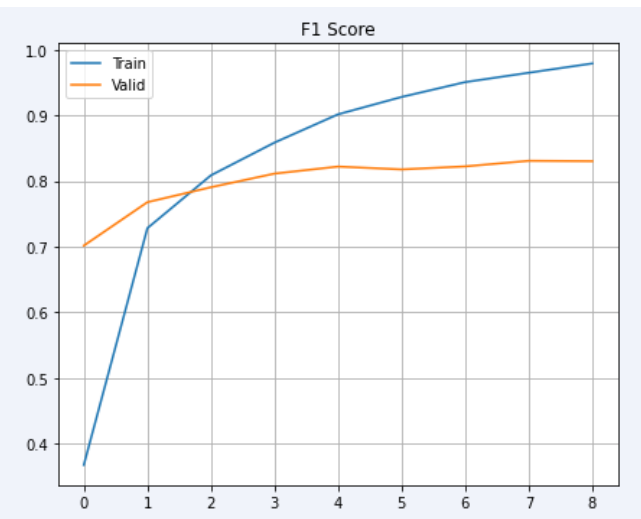
ViT + RoBERTa

Accuracy = 0.8744

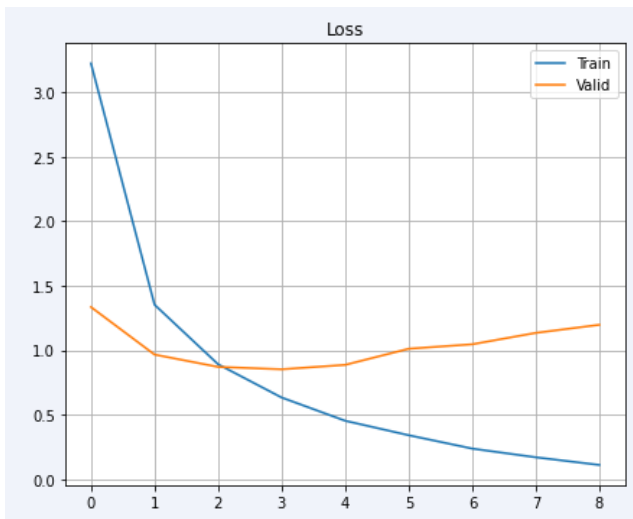
F1-score = 0.8728



Accuracy



F1 - score



Loss



02 최종 모델

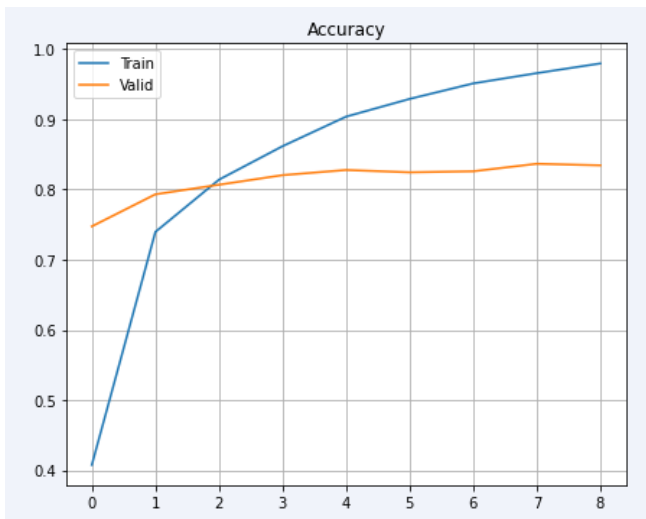


4. 모델 성능 비교

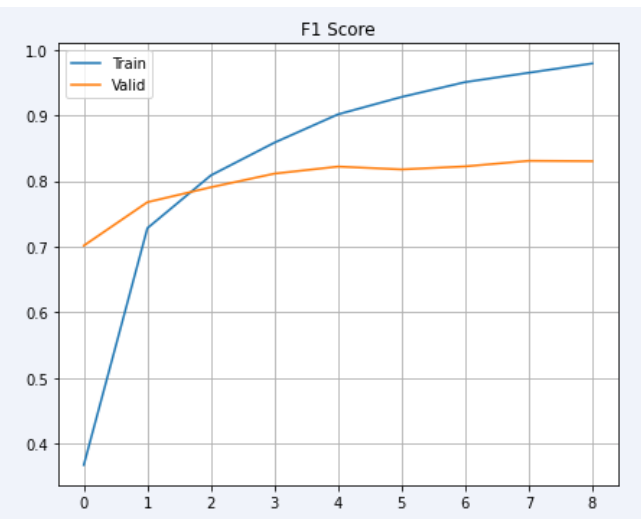


ViT + RoBERTa

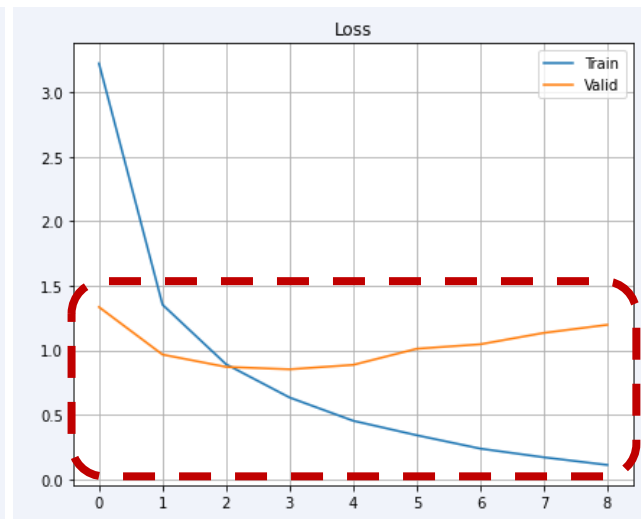
꾸준히 증가하는 validation score, 비교적 낮은 loss 기록



Accuracy



F1 - score



Loss



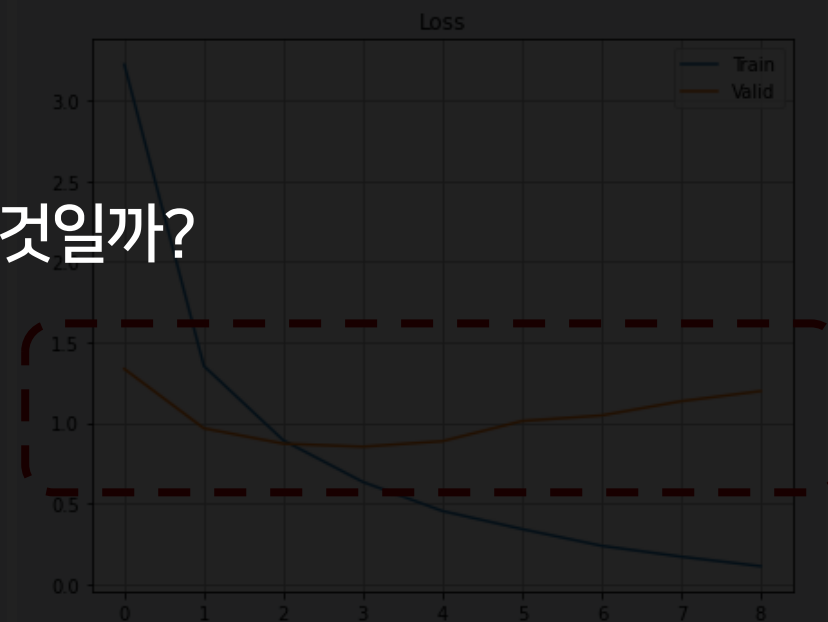
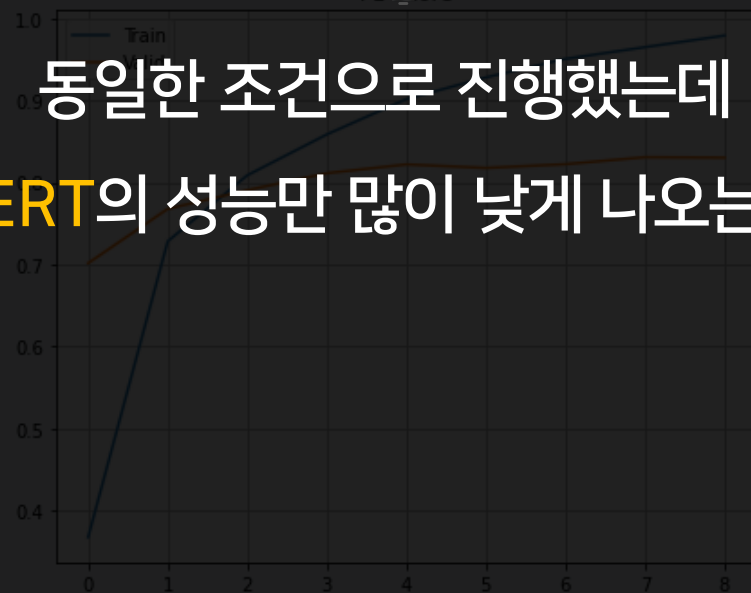
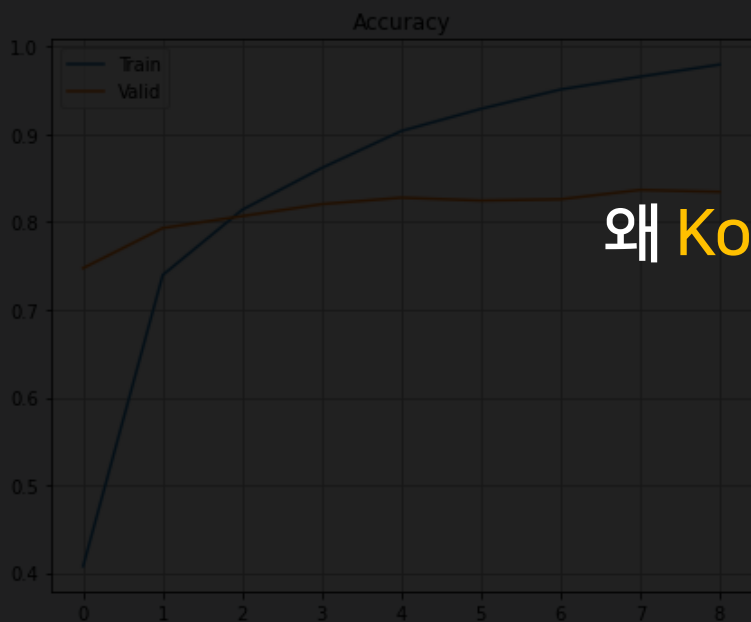
02 최종 모델



4. 모델 성능 비교

ViT + RoBERTa

꾸준히 증가하는 validation score, 비교적 낮은 loss 기록



동일한 조건으로 진행했는데
왜 **KoBERT**의 성능만 많이 낮게 나오는 것일까?

Accuracy

F1 - score

RMSE loss



02 최종 모델



4. 모델 성능 비교



전전 데마 팀장님의 피드백

성능이 잘 나오지 않은 모델에서
어떠한 이유를 딱 찾는 것은 위험!
사전학습 모델을 쓸 때 Fine-tuning을 진행한다고 해도
파라미터가 미세 조정되는 정도이기 때문에
도메인의 차이가 큰 경우에는 성능이 낮을 수 있다
RoBERTa가 학습한 도메인이 KoBERT보다
우리 관광데이터에 잘 맞은 정도로..

본인 사진은 에바참치콩치라 하셔서
젊은 연예인이라도 넣어드립니다!





02 최종 모델



4. 모델 성능 비교

Multimodal

true

한식	1471
야영장,오토캠핑장	650
바/카페	335
유적지/사적지	236
일반축제	222
...	...
터널	2
헬스투어	1
이색체험	1
백화점	1
요트	1

112 rows × 1 columns

RoBERTa

true

한식	1443
야영장,오토캠핑장	649
바/카페	332
일반축제	227
유적지/사적지	222
...	...
문화관광축제	1
헬스투어	1
영화관	1
백화점	1
요트	1

112 rows × 1 columns

?

데이터의 개수가 적은 클래스에 대해서 얼마나 분류가 잘 이뤄졌을까?

Multimodal VS RoBERTa

Test set에 존재하는 cat3: 총 123개

Multimodal과 RoBERTa 모두 112개의
Cat3에 대한 예측 결과가 존재

F1-score는 RoBERTa가 더 높았지만

예측한 소분류의 개수는 동일함



02 최종 모델



4. 모델 성능 비교



Overview	prediction		true
	RoBERTa	Multimodal	
"해양관광 도시 부산에는 광안대교와 부산바다를 구경할 수 있는 요트투어가 다양하다. 대표적으로 다이아몬드베이의 마이다스호가 사람들에게 널리 알려져 있으며 출항시간도 다양하여 선택의 폭이 넓다. ..."	요트	요트	요트
"2017년도에 이어 3회연속 충북 농특산물 판매활성화 최우수축제로 선정된 보은 대추축제는, 임금님께 진상하였던 명품 보은 대추와 보은의 청정한 자연에서 자란 우수한 품질의 농특산물을"	문화관광축제	일반축제	문화관광축제
"팔공산은 경산시의 북쪽에 위치한 해발 1192.3 m의 높은 산으로 신라시대에는 중악, 부악으로 알려진 명산이다. 이곳에는 관봉석조여래좌상(갯바위), 원효사, 천성사, 불굴사 등 신라 고찰과 문화유적이 많다."	군립공원	산	산

개수가 적은 카테고리도 충분히 잘 예측하는 결과를 보임!



02 최종 모델



4. 모델 성능 비교



그렇다면 어떤 모델을 **최종모델**로 선택해야 할까???

Overview	prediction		true
	RoBERTa	Multimodal	
"해양관광 도시 부산에는 광안대교와 부산바다를 구경할 수 있는 요트 표적으로 다이아몬드베이의 마이다스호가 사람들에게 널리 알려져 있 양하여 선택의 폭이 넓다. ..."	<u>요트</u>	<u>요트</u>	<u>요트</u>
"2017년도에 이어 3회연속 충북 농특산물 판매활성화 최우수축제로 선정된 보은대추축제 는, 임금님께 진상하였던 명품 보은 대추와 보은의 청정한 자연에서 자란 우수한 품질의 농 특산물을"	문화관광축제	일반축제	문화관광축제
"팔공산은 경산시의 북쪽에 위치한 해발 1192.3 m의 높은 산으로 신라시대에는 중악, 부 악으로 알려진 명산이다. 이곳에는 관봉석조여래좌상(갯바위), 원효사, 천성사, 불굴사 등 신라 고찰과 문화유적이 많다."	군립공원	산	산



02 최종 모델



4. 모델 성능 비교



Multimodal



ViT (Vision Transformer) + RoBERTa

Accuracy = 0.8744

F1-score = 0.8728

VS

Single Modal



RoBERTa

Accuracy = 0.8733

F1-score = 0.8740

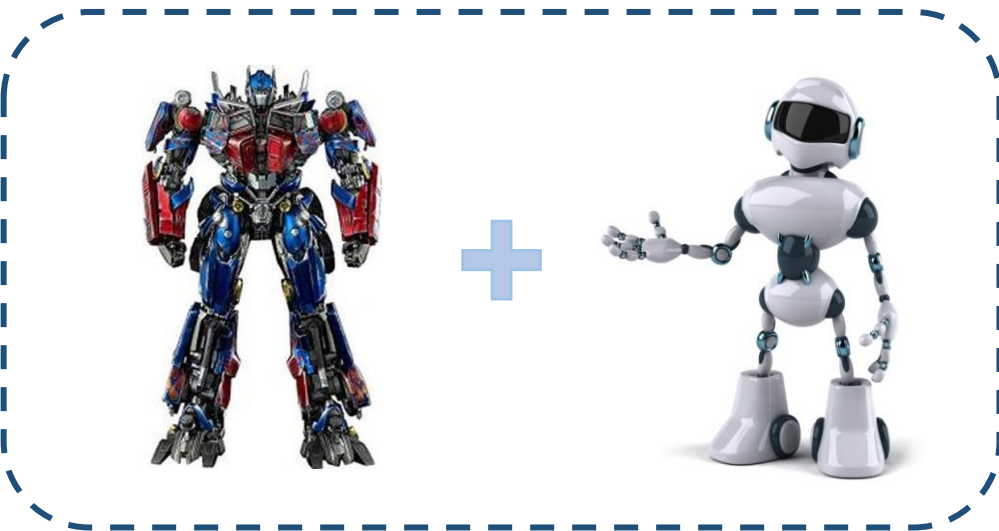


02 최종 모델



4. 모델 성능 비교

Multimodal 



Multimodal을 택한 이유

- ✓ 성능 지표는 RoBERTa가 멀티모달에 비해 약간 좋음

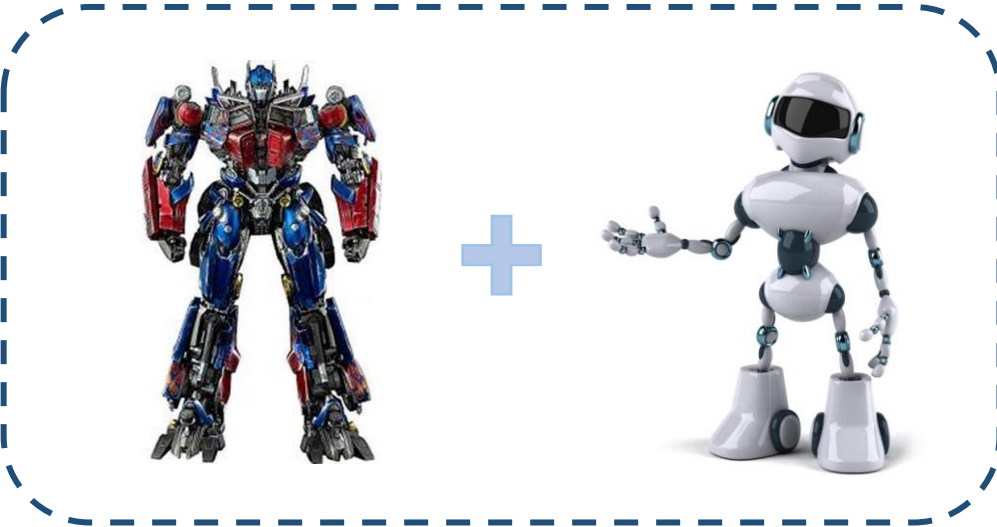


02 최종 모델



4. 모델 성능 비교

Multimodal 

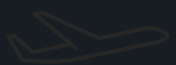


Multimodal을 택한 이유

- ✓ 성능 지표는 RoBERTa가 멀티모달에 비해 약간 좋음

BUT,

- ✓ 이미지 데이터가 일관성이 없다는 점을 고려해야함



02 최종 모델



4. 모델 성능 비교



택한 이유

BERTa 가

약간 좋음

✓ 이미지 데이터가 일관성이 없다는

점을 고려해야함

같은 **한식** 카테고리임에도 불구하고,
하나는 **음식** 사진, 하나는 **간판** 사진으로

일관성이 없음

특징을 추출하기에 **부적합!**





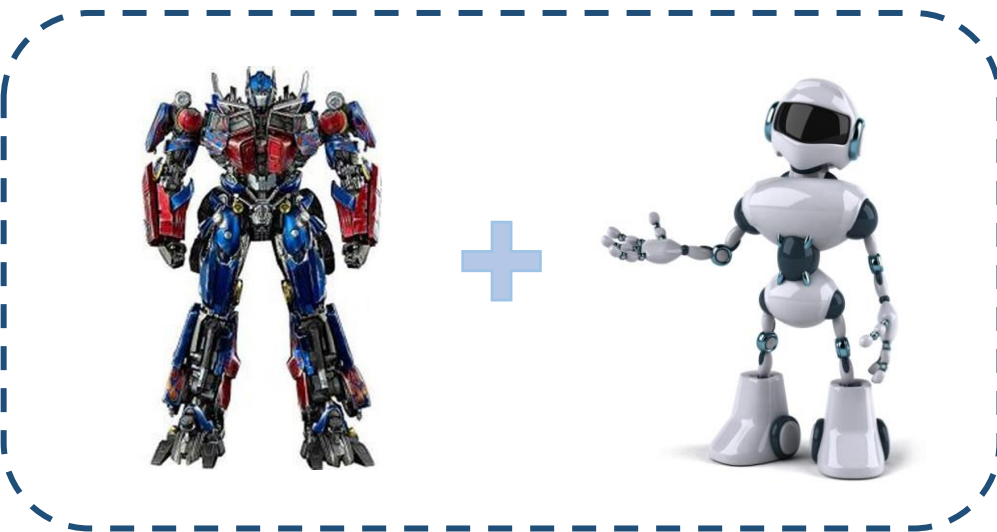
02 최종 모델



4. 모델 성능 비교



Multimodal 



이미지 데이터에 일관성이 생긴다면
유의미한 성능 향상을 기대할 수 있을수도..?

Multimodal을 택한 이유

- ✓ 성능 지표는 RoBERTa가
멀티 모달에 비해 약간 좋음

BUT,

- ✓ 이미지 데이터가 일관성이 없다는
점을 고려해야함
- ✓ 성능 차이가 그렇게 크지 않기에,
SNS에서 활용하기 좋은 멀티모달을
최종적으로 채택!!

03 토이 프로젝트





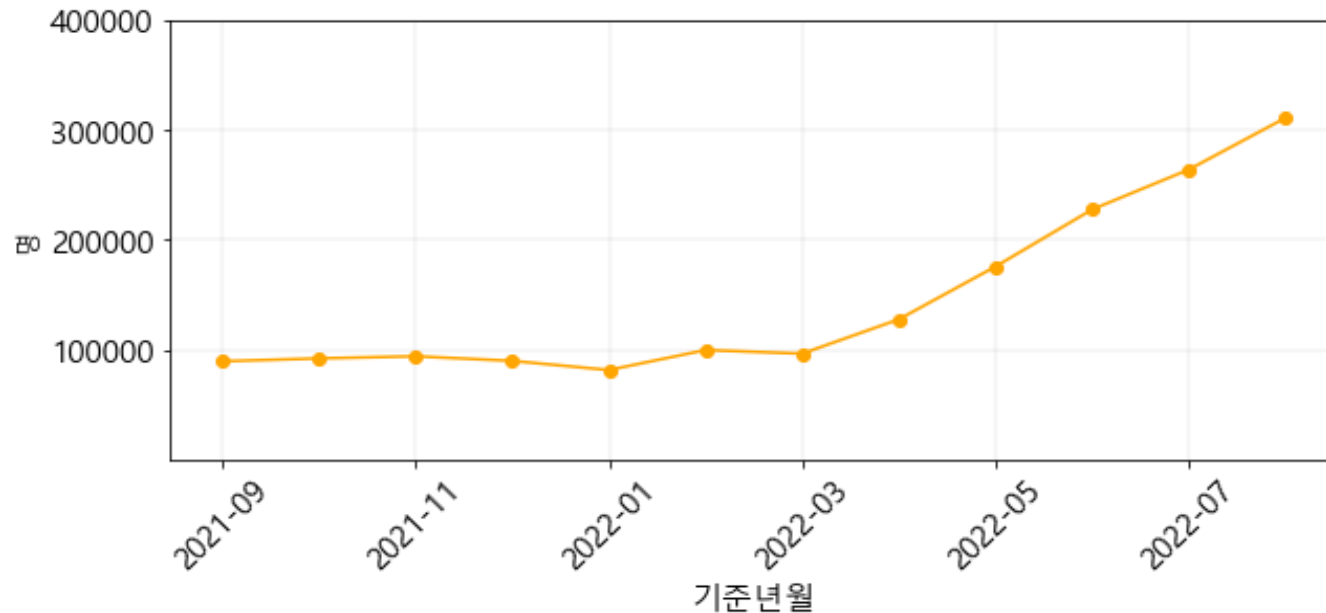
03 토이 프로젝트



1. 토이 프로젝트 선정 배경



방한 외래관광객 추이



외래 관광객 수

276% 증가

작년 9월 대비 올해 9월 방한 외래 관광객의 수는 276% 증가하였으며,
방한 외래 관광객 수는 계속 증가할 것으로 예상됨



03 토이 프로젝트



1. 토이 프로젝트 선정 배경



한국여행 경험 및 의향

구분	2021년
전 생애 한국여행 경험	19.4%
향후 3년 내(~2024년) 한국여행 의향	47.0%
한국 방문 예상 시기	2024년(35.7%)
	2022년(32.0%)
	2023년(28.9%)
	2021년 7~12월(3.4%)

2021년 잠재 방한여행객 조사 결과 향후 3년 내에 한국여행 의향이 47.0%로 나타났으며,
앞으로 많은 외래 관광객이 한국을 방문할 것으로 예상됨



03 토이 프로젝트

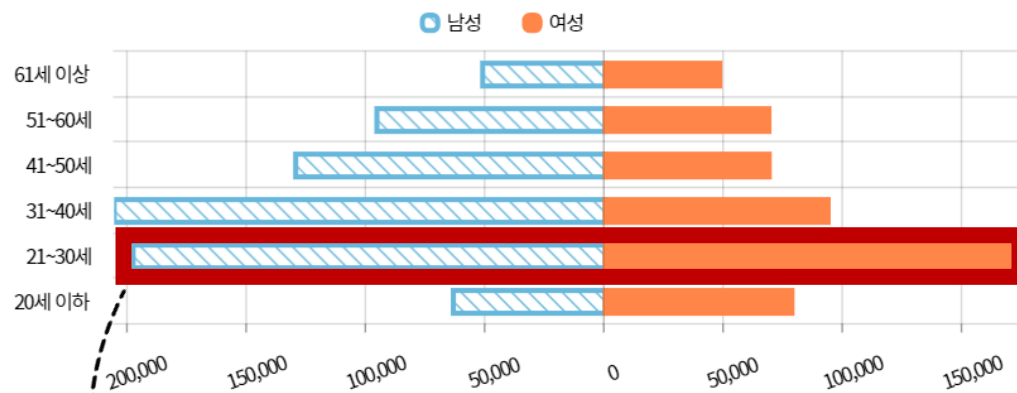


1. 토이 프로젝트 선정 배경



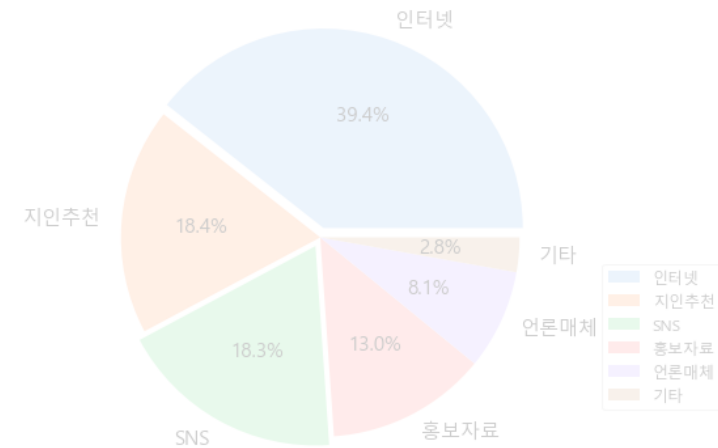
출처: 한국여행협회(KATA)

방한 외래관광객 성·연령별



특히, 방한 외래관광객은 21-30세가 가장 많음

한국여행 정보 획득경로



방한 외래 관광객의 20-30대는 인터넷, SNS에서 정보를 획득하는 비중이 높은 것으로 나타남



03 토이 프로젝트

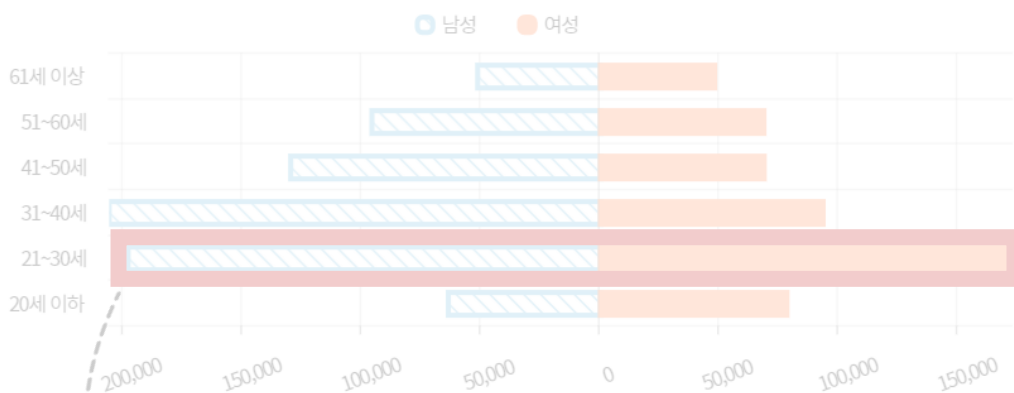


1. 토이 프로젝트 선정 배경



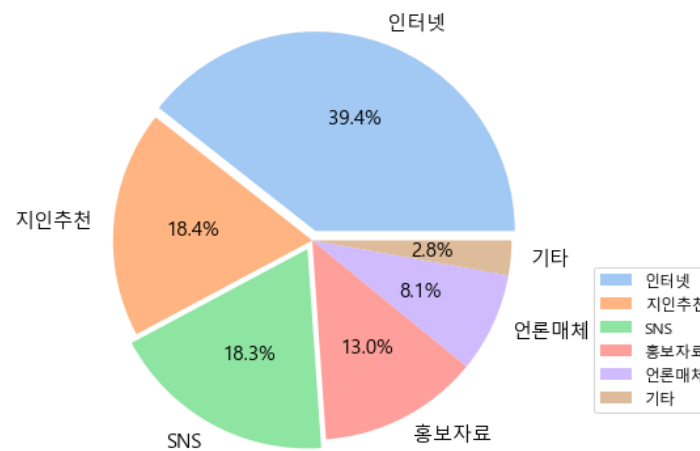
출처: 한국여행협회(KATA)

방한 외래관광객 성·연령별



특히, 방한 외래관광객은 21-30세가 가장 많음

한국여행 정보 획득경로



방한 외래 관광객의 20-30대는 인터넷, SNS에서 정보를 획득하는 비중이 높은 것으로 나타남



03 토이 프로젝트



1. 토이 프로젝트 선정 배경



현재 SNS상에서 외국인 관광객을 위한 **영문 관광자료**의 수요가 높음

구분	2021년
전 생애 한국여행 경험	19.4%
향후 3년 내(~2024년) 한국여행 의향	47.0%
한국 방문 예상 시기	2024년(35.7%)
	2022년(32.0%)
	2023년(28.9%)
	2021년 7~12월(3.4%)

2021년 잠재 방한여행객 조사 결과 향후 3년 내에 한국여행 의향이 47.0%로 나타났으며,
앞으로 많은 외래 관광객이 한국을 방문할 것으로 예상됨



03 토이 프로젝트



1. 토이 프로젝트 선정 배경



현재 SNS상에서 외국인 관광객을 위한 **영문 관광자료**의 수요가 높음

구분	2021년
전 생애 한국여행 경험	19.4%
향후 3년 내(~2024년) 한국여행 의향	47.0%
한국 방문 예상 시기	2024년(35.7%)
	2022년(32.0%)
	2023년(28.9%)
	2021년 7~12월(3.4%)

외국인 관광객을 위한
해시태그 생성 모델을 추가로 만들어 보자!



기계번역



카테고리 분류



해시태그 생성 모델

2021년 잠재 방한여행객 조사 결과 향후 3년 내에 한국여행 의향이 47.0%로 나타났으며

앞으로 많은 외래 관광객이 한국을 방문할 것으로 예상됨



03 토이 프로젝트



1. 토이 프로젝트 선정 배경



?

단순히 이미 분류한 데이터를 가지고 번역을 진행하면 되지 않나요?

1. 단순히 한국 관광자료에서 해시태그를 추출하여 번역한 자료는
영어 문화권의 표현의 차이를 반영하지 못함



영문 데이터 모델을 구축해 놓으면 추가적인 데이터 수집을 통해 활용 가능할 것

2. 한국 데이터 분류 모델과 기계번역 이후 분류 모델이
똑같이 유의미한 성능이 나올지 알아보고자 함



03 토이 프로젝트



2. 기계번역



인공 지능을 사용하여 사람의 개입 없이
한 언어에서 다른 언어로 텍스트를 자동으로 번역하는 프로세스



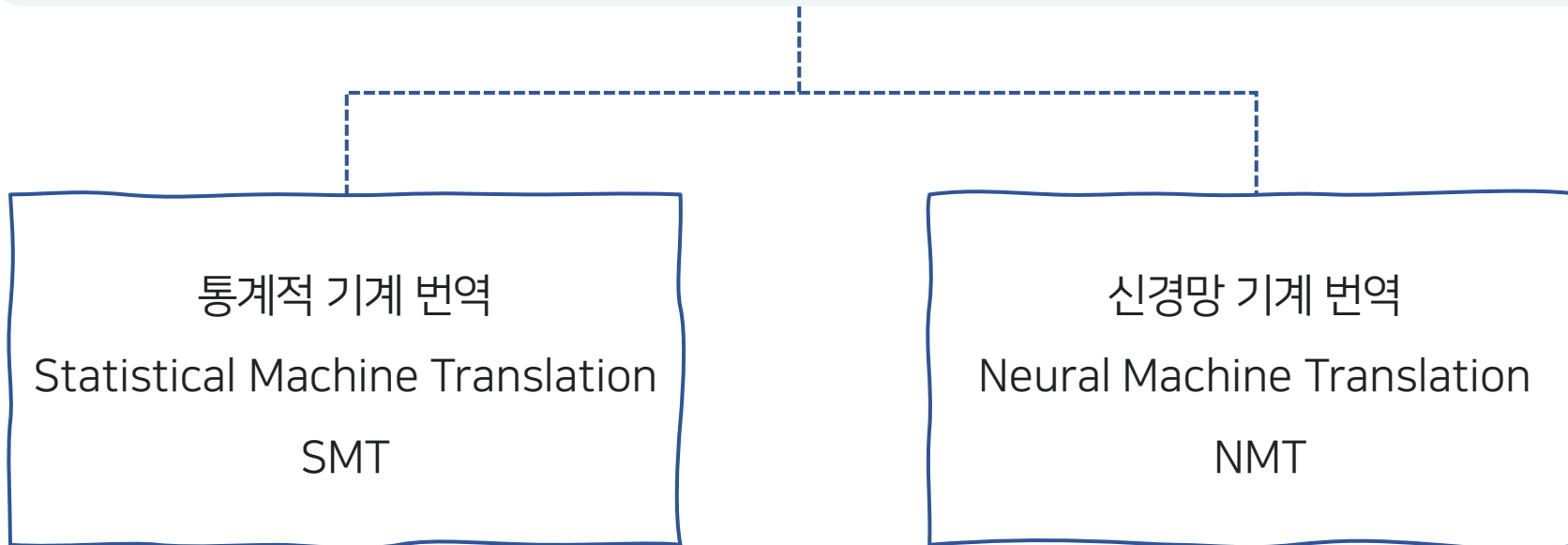
03 토이 프로젝트



2. 기계번역



인공 지능을 사용하여 사람의 개입 없이
한 언어에서 다른 언어로 텍스트를 자동으로 번역하는 프로세스





03 토이 프로젝트

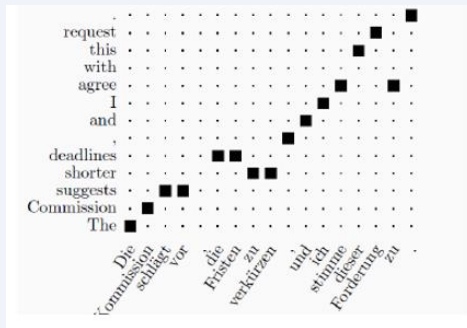


2. 기계번역



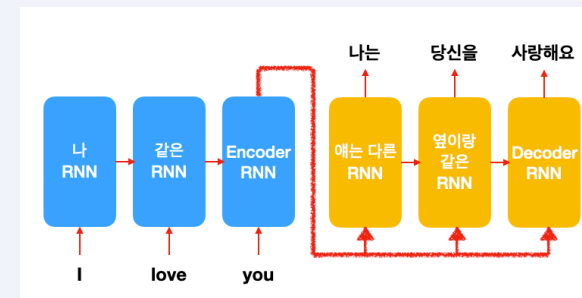
SMT

- ✓ 통계기반 기계번역
- ✓ 특정 단어가 나왔을 때 다음 단어가 나올 확률을 구하는 통계적 기법



NMT

- ✓ 신경망기반 기계번역
- ✓ 두 언어의 말뭉치를 학습시켜 번역하는 모델을 구현





03 토이 프로젝트

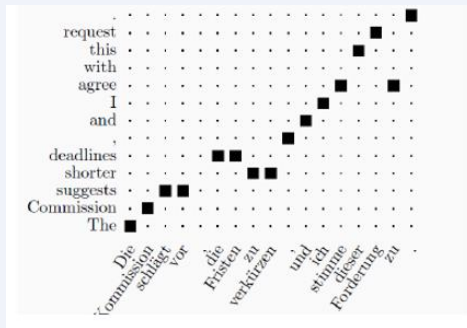


2. 기계번역

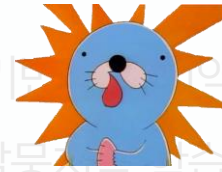


SMT

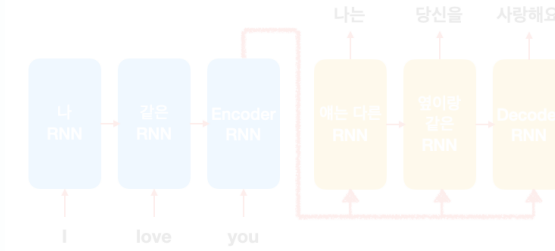
- ✓ 통계기반 기계번역
- ✓ 특정 단어가 나왔을 때 다음 단어가 나올 확률을 구하는 통계적 기법



NMT



- ✓ 신경망기반 기계번역
- ✓ 두 언어의 말뭉치를 학습시켜 통계적 기법을 활용했기 때문에, 동음이의어나 전체 문장이 흘러가는 맥락을 이해하지 못했음






03 토이 프로젝트



2. 기계번역



SMT



역

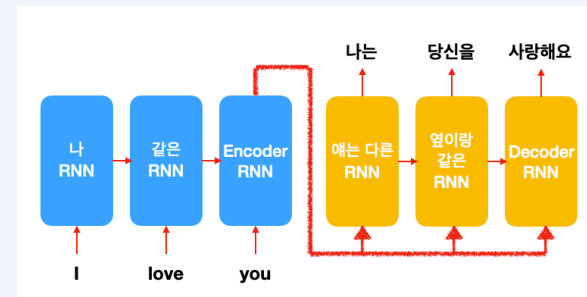
✓ 특...는 그거 곳이라고 생각해요

신경망 모델을 활용하여 단어의 맥락을 부여,
SMT와 달리 양질의 데이터만
있다면 좋은 번역 성능을 얻을 수 있음



NMT

- ✓ 신경망기반 기계번역
- ✓ 두 언어의 말뭉치를 학습시켜
번역하는 모델을 구현

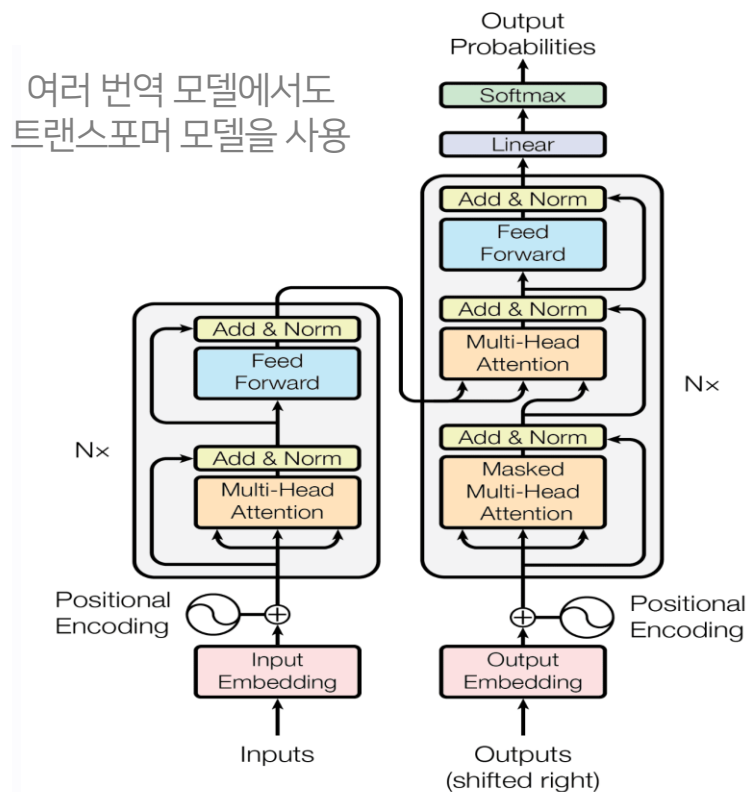




03 토이 프로젝트



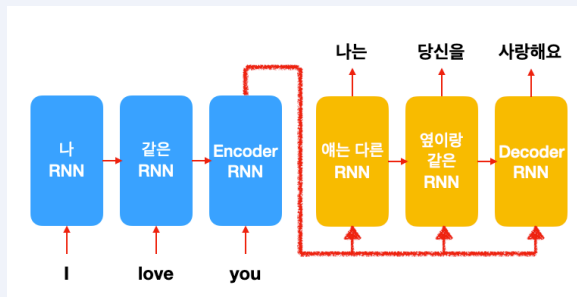
2. 기계번역



자세한 내용은 딥러닝 1주차 주제분석 참고

NMT

- ✓ 신경망기반 기계번역
- ✓ 두 언어의 말뭉치를 학습시켜 번역하는 모델을 구현





03 토이 프로젝트



2. 기계번역



Step 1. 말뭉치 토큰나이징

Source: 도깨비에 대한 현대적 활용은 앞으로도 과제가 될 것이다.

Target: The modern use of goblin will be a challenge in the future.



<sos>도깨비/에/대한/현대적/활용은/앞으로도/과제가/될/것/이다.<eos>

<sos>The/modern/use/of/goblin/will/be/a/challenge/in/the/future.<eos>



Step 2. 임베딩 벡터

[0, 234, 34, ..., 34, 451, 203]

[0, 1234, 3, ..., 34, 456, 234]



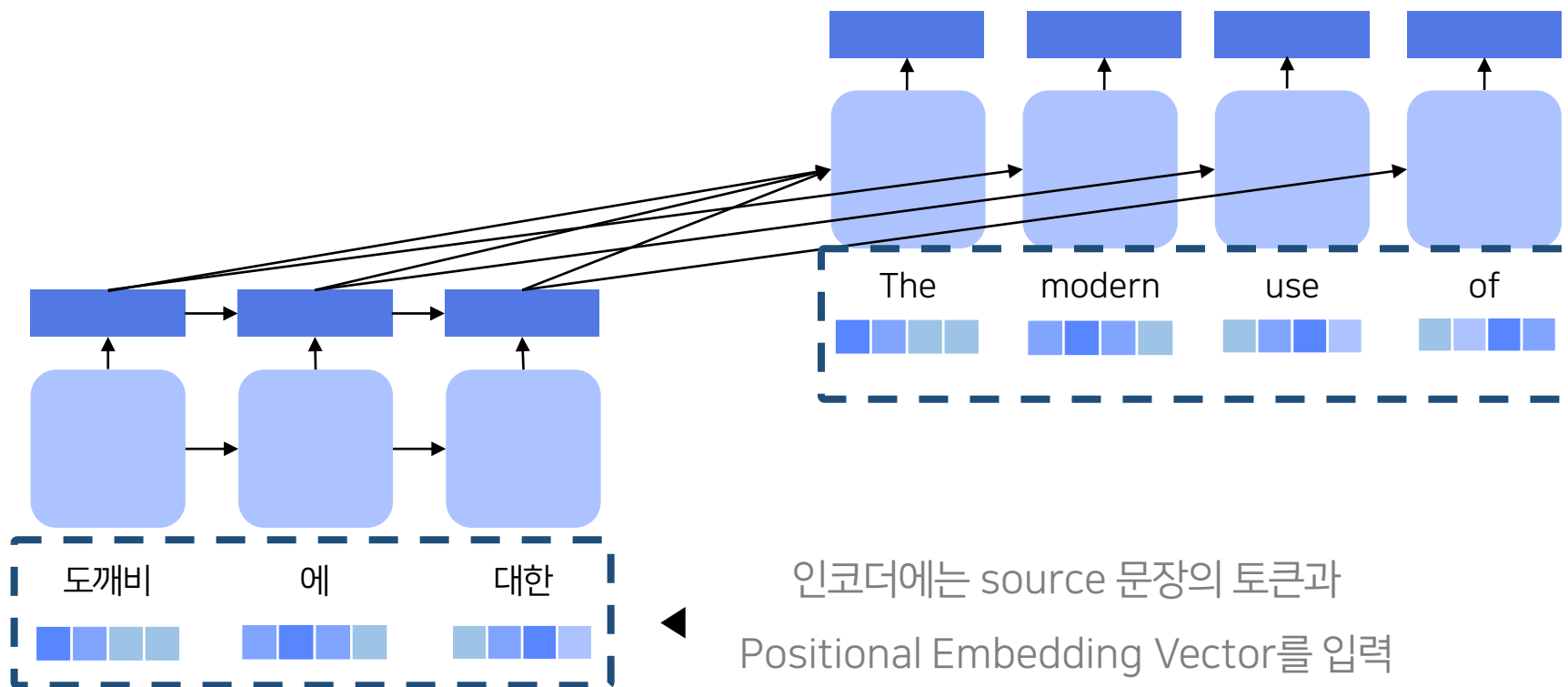
03 토이 프로젝트



2. 기계번역



Step 3. Transformer 모델을 이용한 학습





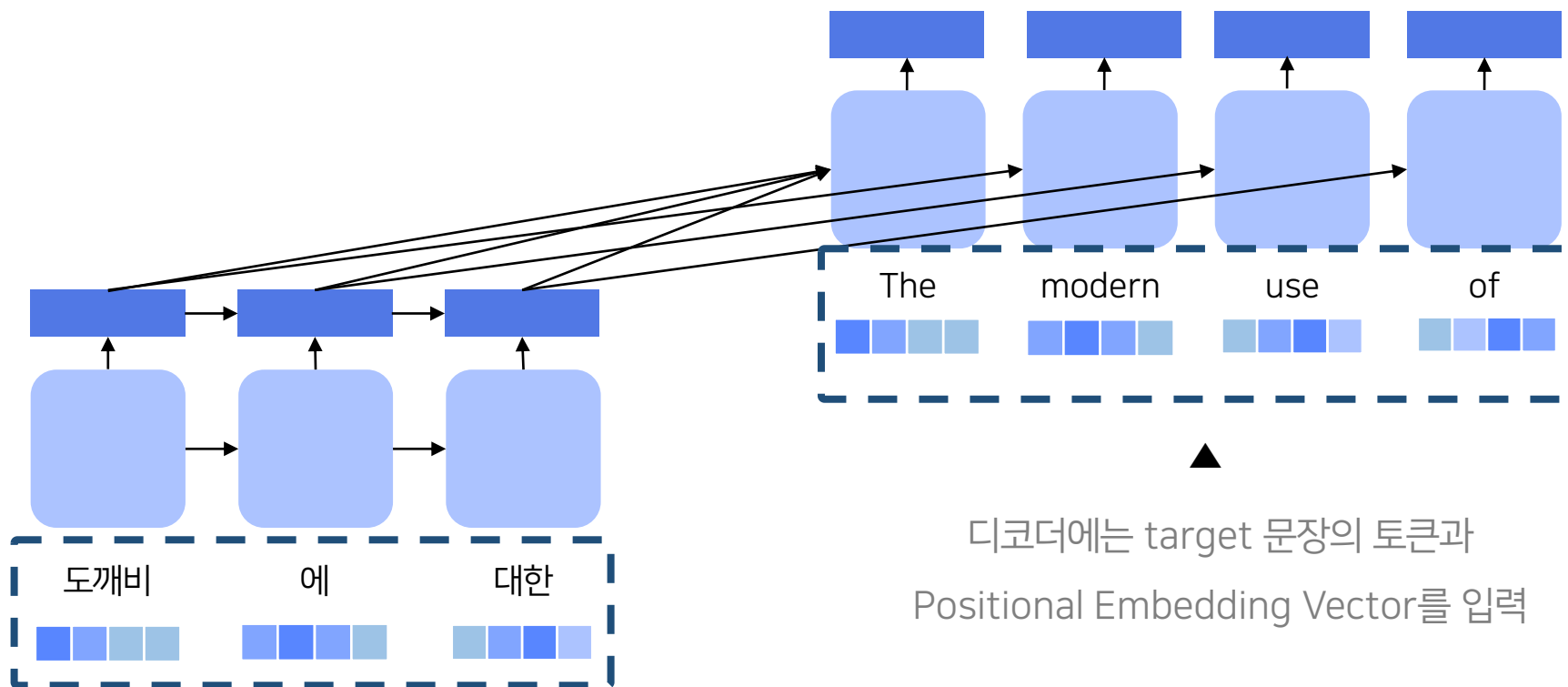
03 토이 프로젝트



2. 기계번역



Step 3. Transformer 모델을 이용한 학습





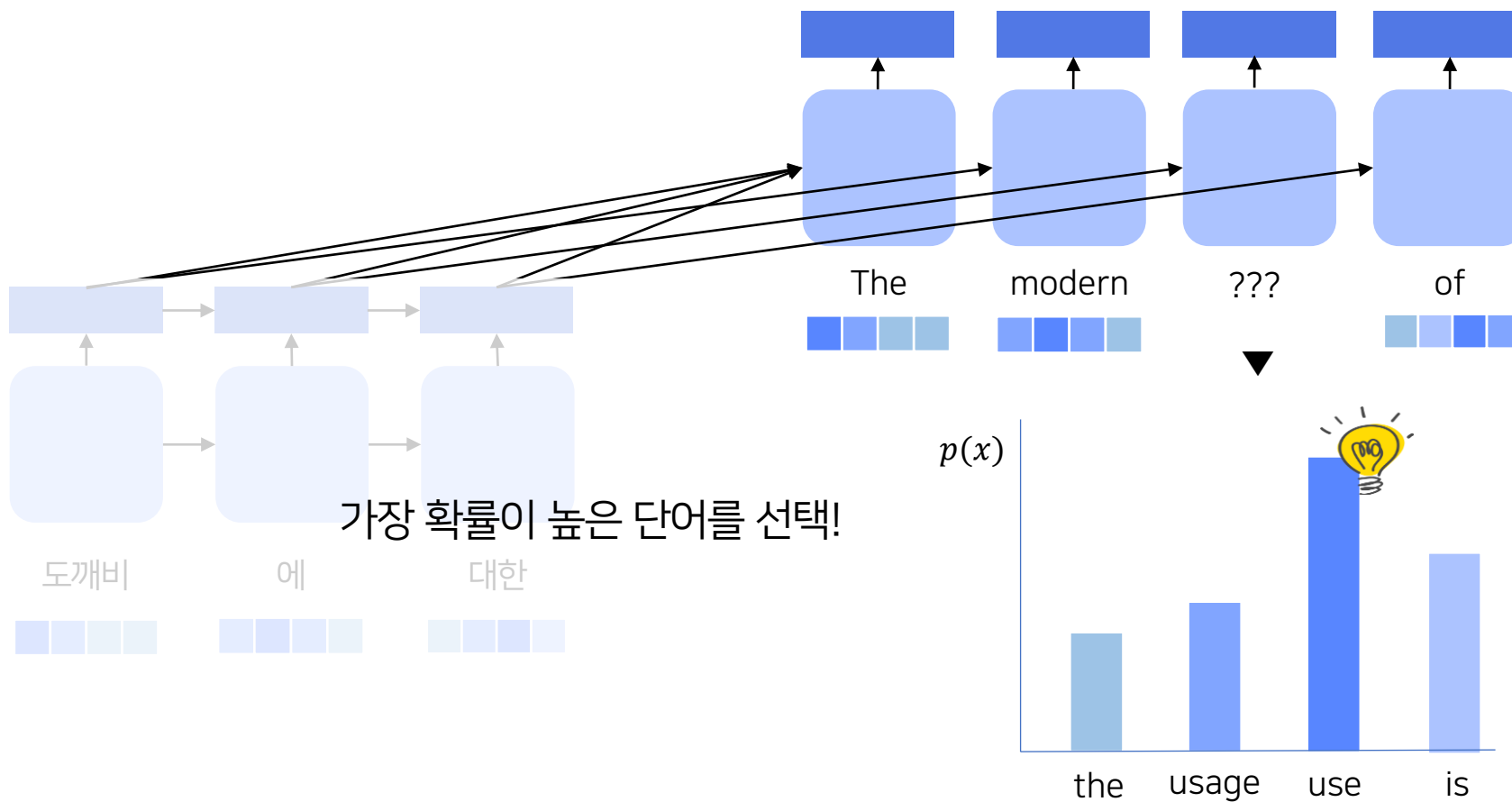
03 토이 프로젝트



2. 기계번역



Step 4. 학습된 모델을 통해 단어를 한 단어 씩 예측





03 토이 프로젝트



2. 기계번역



OOV(Out-Of-Vocabulary) 문제

OOV에 대한 자세한 설명은 딥러닝팀 3주차 클린업 참고!

Source: 사창시장은 충북 청주시 서원구에 자리 잡고 있다.

Target: The <unk> market is established in <unk> <unk>



'사창', '충북', '청주시', '서원구'와 같은 고유 명사들이
단어 사전에 존재하지 않아 <unk>가 반환되는 문제 발생



국립 국어원에서 웹페이지 크롤링을 통해 고유 명사 단어들을 추가적으로 학습!



03 토이 프로젝트



2. 기계번역



OOV(Out-Of-Vocabulary) 단어

Kor	Eng	Kor	Eng
가경동	Gagyeon-dong	가계	Gagye
가곡	Gagok	가곡동	Gagok-dong
가곡면	Gagok-myeon	가나안	Ganaan
가남읍	Ganam-eup	가남정	Ganamjeong



문화체육관광부
국립국어원

국립국어원에서 웹페이지 크롤링을 통해 고유명사 단어들을 추가적으로 학습!



03 토이 프로젝트



2. 기계번역



Source: 경기도 이천시 모가면에 있는 골프장으로 대중제 18홀이다.

Target: it is a golf course that is closed due to golf courses in Gyeonggi-do, which were preserved for Gyeonggi-do province.

경기도 -> gyeonggi-do, Gyeonggi-do province로
고유명사가 학습된 것을 확인할 수 있음



03 토이 프로젝트



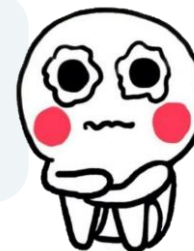
2. 기계번역



Source: 경기도 이천시 모가면에 있는 골프장으로 대중제 18홀이다.

Target: it is a golf course that is closed due to golf courses in Gyeonggi-do,
which were preserved for Gyeonggi-do province.

하지만 여전히 '이천시', '모가면', '대중제'와 같은 고유명사는 학습이 안 되었고,
문맥이 맞지 않는 번역과 반복되는 구가 늘었음





03 토이 프로젝트



2. 기계번역



Source: 경기도 이천시 모가면에 있는 골프장으로 대중제 18홀이다.

Target: it is a golf course that is closed due to golf courses in Gyeonggi-do, which were preserved for Gyeonggi-do province.

실제 BLEU 점수 계산을 통해서
기계번역 모델의 성능을 평가해보자!

하지만 여전히 '이천시', '모가면', '대중제'와 같은 고유명사는 학습이 안 되었고,
문맥이 맞지 않는 번역과 반복되는 구가 늘었다.



3. BLEU(Bilingual Evaluation Understudy)

BLEU란? 

기계 번역의 성능이 얼마나 뛰어난가를 측정하기 위해 사용되는 대표적인 방법

측정 기준은 n-gram에 기반



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



BLEU란? 

기계 번역의 성능이 얼마나 뛰어난가를 측정하기 위해 사용되는 대표적인 방법
측정 기준은 n-gram에 기반

 n- gram 이란? — — — — —

“N개의 연속적인 단어 나열”

n-gram 언어 모델은 카운트에 기반한 통계적 접근을 사용

이전에 등장한 모든 단어를 고려하는 것이 아니라 일부 단어만 고려

*이때 일부단어의 개수가 n을 의미함



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



Human evaluation

사람이 직접 평가
정확하지만 평가하려는 언어에 대한
제한이 발생하며 오랜 시간이 걸림



BLEU

언어에 구애 받지 않고 사용 가능
계산 속도가 빠름



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)

Human evaluation

사람이 직접 평가
정확하지만 평가하려는 언어에 대한
제한이 발생하며 오랜 시간이 걸림



BLEU

언어에 구애 받지 않고 사용 가능
계산 속도가 빠름



3. BLEU(Bilingual Evaluation Understudy)



단어 개수 카운트로 측정하기



목표: 한국어- 영어 번역기의 성능을 측정해보자

두 기계 번역기에 같은 한국어 문장을 입력하여 번역된 영어 문장의 성능을 측정하고자 함

Example 1

- Candidate1 : It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate2 : It is to insure the troops forever hearing the activity guidebook that party direct.
- Reference1 : It is a guide to action that ensures that the military will forever heed Party commands.
- Reference2 : It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference3 : It is the practical guide for the army always to heed the directions of the party.



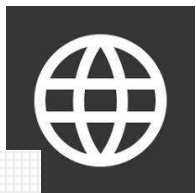
03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



단어 개수 카운트로 측정하기



목표: 한국어- 영어 번역기의 성능을 측정해보자

두 기계 번역기에 같은 한국어 문장을 입력하여 번역된 영어 문장의 성능을 측정하고자 함

번역된 문장

Example 1

- Candidate1 : It is a guide to action which ensures that the military always obeys the commands of the party.
- Candidate2 : It is to insure the troops forever hearing the activity guidebook that party direct.
- Reference1 : It is a guide to action that ensures that the military will forever heed Party commands.
- Reference2 : It is the guiding principle which guarantees the military forces always being under the command of the Party.
- Reference3 : It is the practical guide for the army always to heed the directions of the party.

정답으로 비교되는 문장



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



단어 개수 카운트로 측정하기

<유니그램 정밀도>

$$\text{Unigram Precision} = \frac{\text{ref들 중에서 존재하는 } ca \text{의 단어의 수}}{ca \text{의 총 단어 수}}$$

번역된 문장을 정답문장(사람이 번역한 문장)들과 비교하여 성능을 측정!

*직관적인 성능 평가 방법

: 정답문장들 중 한 번이라도 등장한 단어의 개수를
번역된 문장에서 세고, 이를 번역된 문장의 총 단어 수로 나눠줌



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



단어 개수 카운트로 측정하기

Example 1

- Candidate1 : *It is a guide to action which ensures that the military always obeys the commands of the party.*
- Candidate2 : *It is to insure the troops forever hearing the activity guidebook that party direct.*
- Reference1 : *It is a guide to action that ensures that the military will forever heed Party commands.*
- Reference2 : *It is the guiding principle which guarantees the military forces always being under the command of the Party.*
- Reference3 : *It is the practical guide for the army always to heed the directions of the party.*

Ca1의 단어들은 Ref1, Ref2, Ref3에서 전반적으로 등장
반면, Ca2는 그렇지 않음
➡ Ca1이 Ca2보다 더 좋은 번역 문장



Candidate1 정확도 = $\frac{17}{18}$

Candidate2 정확도 = $\frac{8}{14}$



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



중복을 제거하여 보정하기

Example 2

- Candidate : *the the the the the the the*
- Reference1 : *the cat is on the mat*
- Reference2 : *there is a cat on the mat*

Candidate는 *the*만 7개가 등장한 *터무니 없는* 번역

하지만 이 번역은 앞서 배운 유니그램 정밀도에 따르면 $7/7=1$ 이라는 최고의 성능 평가를 받게 됨



유니그램 정밀도를 다소 **보정할 필요성 존재**



03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



중복을 제거하여 보정하기

Example 2

- Candidate : *the the the the the the the*
- Reference1 : *the cat is on the mat* → "the"가 최대 **두번** 등장
- Reference2 : *there is a cat on the mat*

$$\text{Modified Unigram Precision} = \frac{\text{ca의 각 유니그램에 대해 count를 수행한 값의 총 합}}{\text{ca의 총 단어 수}}$$

정밀도의 분자를 계산하기 위한 각 유니그램의 카운트는

"유니그램이 하나의 Ref에서 최대 몇 번 등장했는지"를 카운트 하는것"으로 수정

Ca의 기존 유니그램 정밀도는 $7/7=1$ 이었으나 보정된 유니그램 정밀도는 $2/7$ 와 같이 변경



3. BLEU(Bilingual Evaluation Understudy)



문장 간결성 패널티(Sentence brevity penalty)

Example 3

- Candidate : *the cat*
- Reference1 : *the cat is on the mat*
- Reference2 : *there is a cat on the mat*

번역된 문장이 간결하면 유니그램 정밀도의 분모가 작아지므로
번역 문장의 정확도와 상관없이 정밀도가 올라감



문장 간결성에 대해 패널티를 부여



3. BLEU(Bilingual Evaluation Understudy)



문장 간결성 패널티(Sentence brevity penalty)

Example 3

- Candidate : *the cat*
- Reference1 : *the cat is on the mat*
- Reference2 : *there is a cat on the mat*

Reference **문장의 길이**를 고려하는데, candidate 문장과 가장 길이가 비슷한

Reference 문장의 길이를 **best match length(r)**라고 함

▼ Brevity penalty (c: candidate 문장의 길이)

▼ 최종적인 BLEU 점수

$$BP = \begin{cases} 1 & , \text{ if } c < r \\ e^{(1-\frac{r}{c})} & , \text{ if } c \geq r \end{cases}$$

$$BLEU = BP * \exp(\sum_{n=1}^N W_n \log P_n)$$



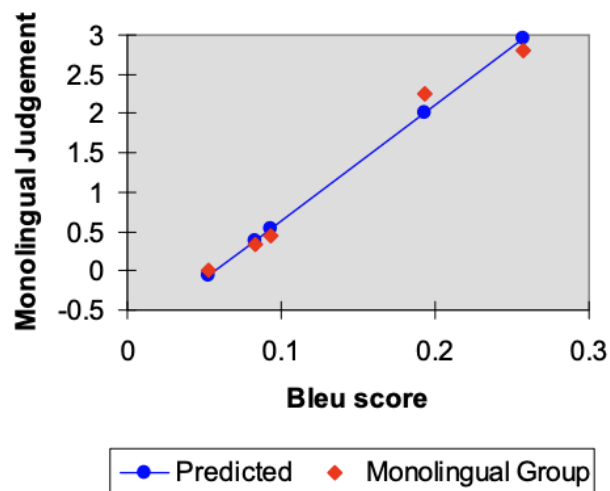
03 토이 프로젝트



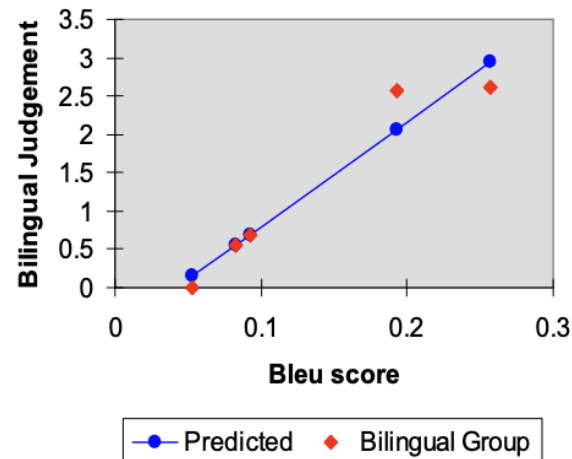
3. BLEU(Bilingual Evaluation Understudy)



사람이 직접 번역한 평가한 결과 VS BLEU score 비교



▲ monolingual인 사람의 평가 결과



▲ bilingual인 사람의 평가 결과

실제 사람이 평가한 결과와 BLEU SCORE가 거의 유사



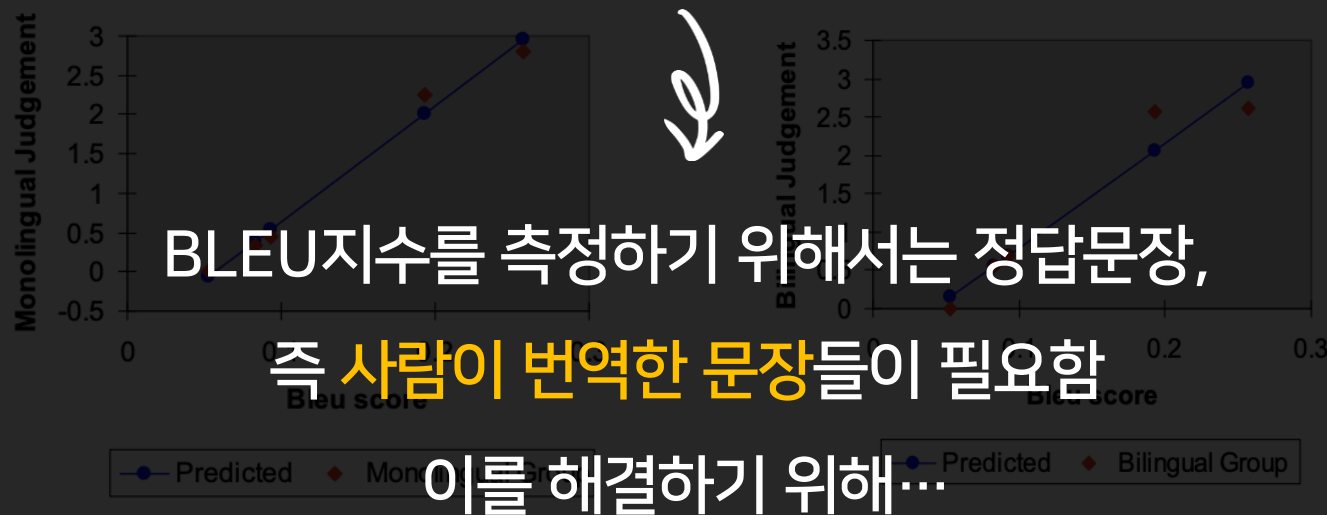
03 토이 프로젝트



3. BLEU(Bilingual Evaluation Understudy)



사람이 직접 번역한 평가한 결과 VS BLEU score 비교



▲ monolingual인 사람의 평가 결과

▲ bilingual인 사람의 평가 결과

실제 사람이 평가한 결과와 BLEU SCORE가 거의 유사



03 토이 프로젝트



4. 기계번역의 성능 평가

P-SAT의 영어 능력자분들께 도움을 받았습니다

고분회귀 징글팀장



선대 Fox



열정데마 썬샤인



딤러닝 다니엘



시계열 금수저



범주 실세





03 토이 프로젝트



4. 기계번역의 성능 평가

다음과 같은 문장들을...

1. 젊음으로 꽉 들어찬 로데오거리는 부평문화의거리와 함께 인천을 대표하는 변화가다.
2. 선사시대를 테마로 한 박물관으로 다양한 선사 시대 모형 유물들이 전시되어 있다.
3. 40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물로 끓인 해물탕을 맛볼 수 있다.



03 토이 프로젝트

4. 기계번역의 성능 평가

다음과 같이 번역해주셨습니다!



Street, which is filled with many young people, is a
town area of Incheon on a par with Bupyeong
Street.

2. This place is the prehistoric themed museum, displayed
with various kinds of prehistoric relics.
3. In Bupyeong Seafood Soup street formed over the 40
years, you can try Seafood Soup made with fresh seafoods
caught off the coast of Incheon.



Street full of youths is the representative
of Incheon along with Bupyeong culture street.

2. It is a prehistoric ages themed museum, where a lot of
prehistoric relic models are displayed.
3. At Bupyeong Haemultang Street, formed over 40 years,
you can enjoy spicy seafood stew made of fresh seafood
from the coast of Incheon.



Street, a street full of youth, is the most famous
as well as Bupyeong Cultural Street in Incheon.
Prehistoric-Age-themed museum exhibits numerous

- kinds of contemporary model remains.
3. You can try a seafood soup made from in-season fresh
seafoods taken from Incheon-nearby sea at 'Bupyeong
Seafood Soup Street' where it has over 40 years history.



Street, full of youth, represents Incheon's main
along with Bupyeong Munhwa Street.

2. It is prehistory museum where various prehistoric relics
are exhibited.
3. At Bupyeong Haemultang Street which has 40 years of
history, you can try Haemultang which is cooked with fresh
seasonal seafood from Incheon offshore.



Rodeo Road and Bupyeong Culture Street
represents Incheon.

- Prehistoric figures and statues are displayed
through a prehistoric-themed museum.
3. Seafood soup made with fresh seafood caught from the
Incheon sea can be tasted at the 40-year-old Bupyeong
Seafood Street.



Street that filled with youth and Bupyeong Culture
Street are a representative mainstreet of Incheon.

2. It is a prehistoric museum which has various prehistoric
model artifacts.
3. At Bupyeong Seafood Stew Street, which has been
formed over 40 years, you can taste seafood stew made of
fresh seafood from the coast of Incheon.



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 젊음으로 꽉 들어찬 로데오거리는 부평문화의거리와 함께 인천을 대표하는 변화가다.

Reference
Feat.(P-SAT) Rodeo Street full of youth is a representative downtown in Incheon along with Bupyeong Culture Street...

Transformer In addition, the gyeongin country is a street that specializes in culture and romance of the majority.

BLEU score -> 7.5656e-155

Cumulative 1-gram: 0.529412

Cumulative 2-gram: 0.257248

Cumulative 3-gram: 0.000000

Cumulative 4-gram: 0.000000

1-gram과 2-gram에서는 어느정도 점수가 나왔지만,
3-gram부터는 점수가 0점이 나와
최종적인 BLEU score가 매우 낮음



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 선사시대를 테마로 한 박물관으로 다양한 선사 시대 모형 유물들이 전시되어 있다.

**Reference
Feat.(P-SAT)** It is a prehistoric museum which has various prehistoric model artifacts...

Transformer various models of the bronze age are held in the morning using the training museum.



BLEU score -> 1.4488e-231

Cumulative 1-gram: 0.400000

Cumulative 2-gram: 0.000000

Cumulative 3-gram: 0.000000

Cumulative 4-gram: 0.000000

1-gram을 제외하고는
모든 n-gram 점수가 0점인 것을 확인 가능



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물...

Reference
Feat.(P-SAT) Formed over the course of 40 years Bupyeong Haemultang Street offers
Haemultang cooked with fresh seasonal seafood ...

Transformer the fukujuen was a full service country in which was added over 60 years old and
it can be said that the instant street in the middle.

BLEU score -> 5.5026e-155

Cumulative 1-gram: 0.481481

Cumulative 2-gram: 0.136083

Cumulative 3-gram: 0.000000

Cumulative 4-gram: 0.000000

1-gram과 2-gram에서는 어느정도 점수가 나왔지만,
3-gram부터는 점수가 0점이 나와
최종적인 BLEU score가 매우 낮음



03 토이 프로젝트



4. 기계번역의 성능 평가

문맥에 맞지 않은 단어,
고유명사의 부족,
적은 데이터 세트의 개수



낮은 N-Gram으로 인한
낮은 BLEU Score



03 토이 프로젝트



4. 기계번역의 성능 평가



문맥에 맞지 않은 단어,
고유명사의 부족,
적은 데이터 세트의 개수



낮은 N-Gram으로 인한
낮은 BLEU Score



낮은 기계번역 성능 때문에
파파고를 이용하여
overview 번역





03 토이 프로젝트



4. 기계번역의 성능 평가



파파고 크롤링을 위한 셀레니움 활용

#셀레니움: 웹 애플리케이션 자동화 및
테스트를 위한 포터블 프레임워크

한국어	영어
점음으 는 부평 을 대표	is a use rehis n dis
한국어 선사시 로 다양 들이 전	영어 In Bupyeong Beach, formed over 40 years, it can taste seafood soup with fresh seafood soup in front sea of Incheon, ...
한국어 40년에 걸쳐 형성된 부평해물탕 거리에서는 인천 앞바다에서 공 수해온 제철의 싱싱한 해산물로 끓인 해물탕을 맛볼 수 있다.	



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 젊음으로 꽉 들어찬 로데오거리는 부평문화의거리와 함께 인천을 대표하는 변화가다.

Reference Rodeo Street full of youth is a representative downtown in Incheon along with
Feat.(P-SAT) Bupyeong Culture Street...



Papago

Rodeo Street filled with youth is a representative downtown of Incheon along with Bupyeong Culture Street.

BLEU score -> 0.819618

Cumulative 1-gram: 1.000000

Cumulative 2-gram: 0.966092

Cumulative 3-gram: 0.902713

Cumulative 4-gram: 0.819619

Reference와 비교해도 큰 차이를 못 느낄 번역

BLEU score가 매우 높음



03 토이 프로젝트



4. 기계번역의 성능 평가

Source 선사시대를 테마로 한 박물관으로 다양한 선사 시대 모형 유물들이 전시되어 있다.

**Reference
Feat.(P-SAT)** It is a prehistoric museum which has various prehistoric model artifacts...



It is a museum with the theme of prehistoric times and exhibits various prehistoric model artifacts.

BLEU score -> 0.352578

Cumulative 1-gram: 0.937500

Cumulative 2-gram: 0.750000

Cumulative 3-gram: 0.588814

Cumulative 4-gram: 0.352578

1-gram에서는 높은 값을 보여주나,
n이 커질수록 점점 줄어들어
최종 BLEU score는 높지 않음



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물...

Reference Formed over the course of 40 years Bupyeong Haemultang Street offers
Feat.(P-SAT) Haemultang cooked with fresh seasonal seafood ...



Papago

At Bupyeong Seafood Soup Street which has been formed over 40 years you can taste seafood soup boiled with seasonal fresh seafood ...

BLEU score -> 0.764851

Cumulative 1-gram: 0.962963

Cumulative 2-gram: 0.902671

Cumulative 3-gram: 0.838628

Cumulative 4-gram: 0.764851

1-gram에서 높은 값을 보여주고,
n이 커질수록 점점 줄어들이지만
감소폭이 크지 않아 최종 BLEU score는 준수함



03 토이 프로젝트



4. 기계번역의 성능 평가



Source 40년에 걸쳐 형성된 부평해물탕거리에서는 인천 앞바다에서 공수해온 제철의 싱싱한 해산물...

Reference Formed over the course of 40 years Bupyeong Haemultang Street offers
Feat.(P-SAT) Haemultang cooked with fresh seasonal seafood ...



최종적으로 파파고로 번역한

데이터셋 사용!

BLEU score -> 0.764851

Cumulative 1-gram: 0.962963

Cumulative 2-gram: 0.902671

Cumulative 3-gram: 0.838628

Cumulative 4-gram: 0.764851

1-gram에서 높은 값을 보여주고,

n이 커질수록 점점 줄어들이지만

감소폭이 크지 않아 최종 BLEU score는 준수함

04 토이 프로젝트 결과





04 토이 프로젝트 결과



1. 한국어 데이터와 한-영 번역데이터 성능 비교



최종선택모델을 통한
한국어 데이터 분류

VS

번역이후
영어 데이터 분류

한글 text 데이터를 영어로 번역한 이후
성능은 어떨까?



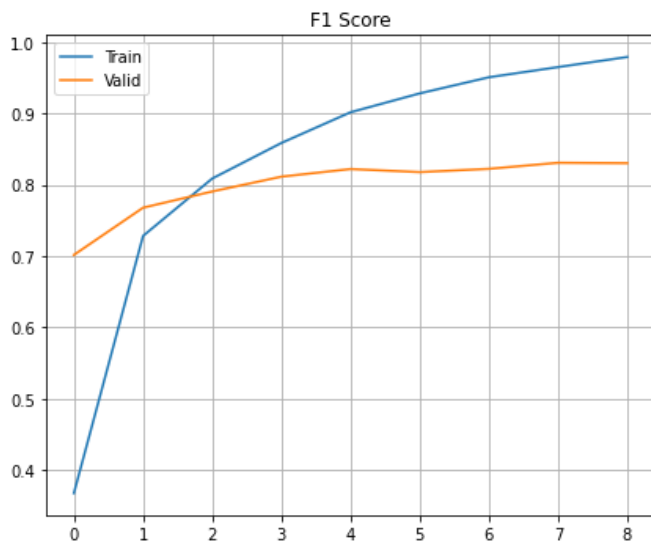
04 토이 프로젝트 결과



1. 한국어 데이터와 한-영 번역데이터 성능 비교

1. F1- SCORE 비교

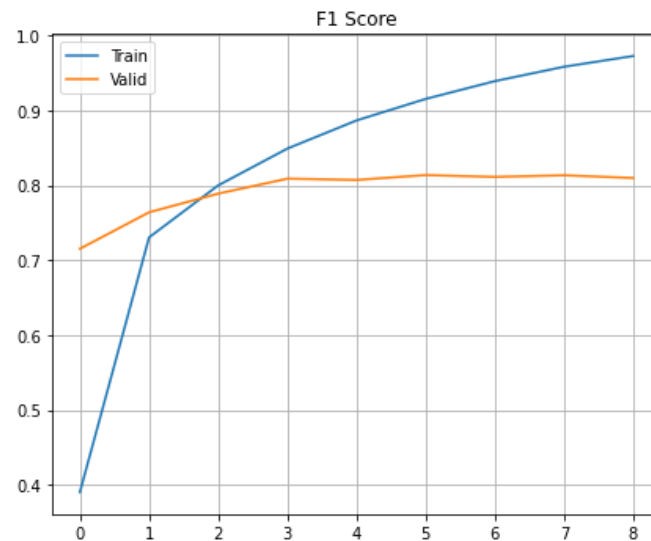
한국어 데이터 분류한 모델



▲ Multi Modal

Test F1 : 0.8728

영어 데이터 분류한 모델



▲ 한영 번역 모델

Test F1 : 0.8459



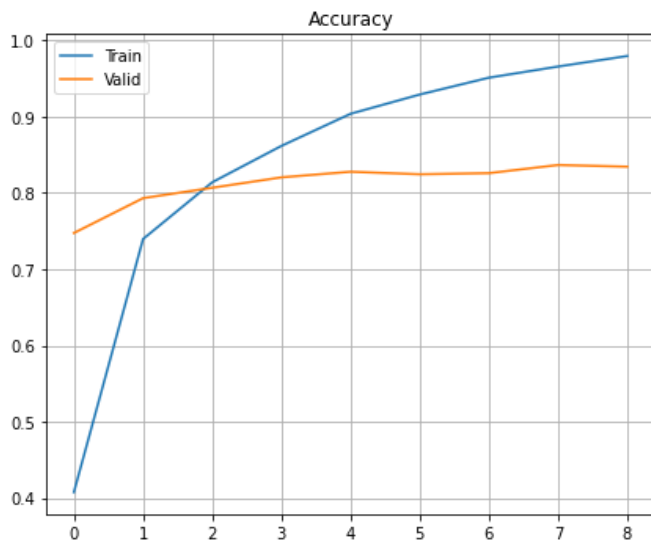
04 토이 프로젝트 결과



1. 한국어 데이터와 한-영 번역데이터 성능 비교

2. Accuracy 비교

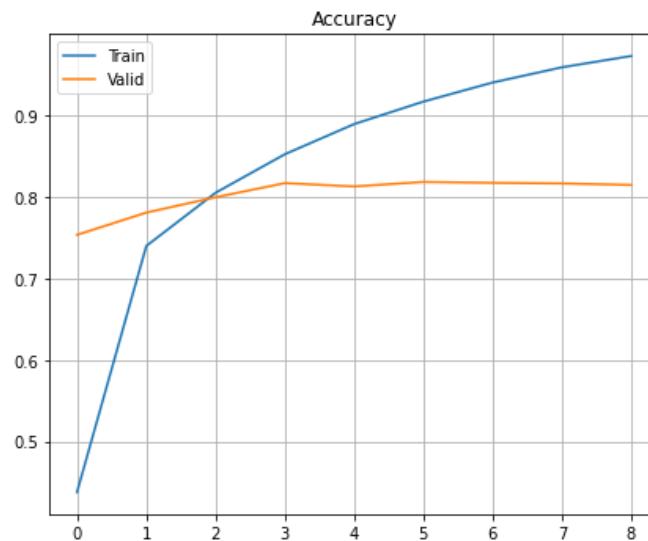
한국어 데이터 분류한 모델



▲ Multi Modal

Test acc : 0.8744

영어 데이터 분류한 모델



▲ 한영 번역 모델

Test acc : 0.8464



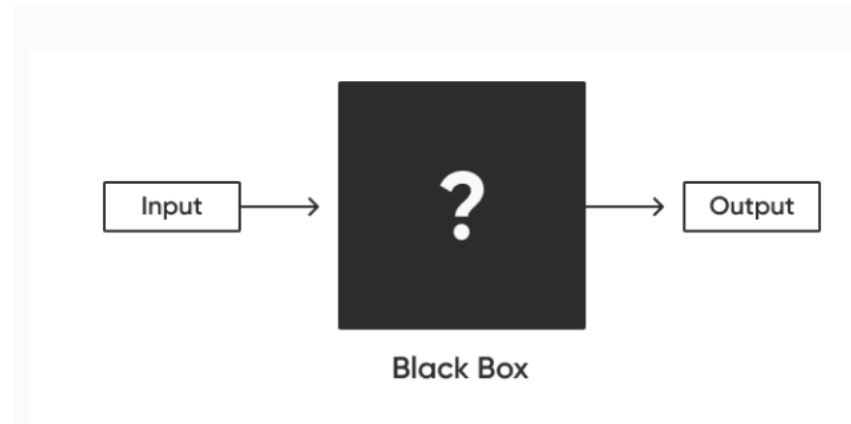
04 토이 프로젝트 결과



1. 한국어 데이터와 한-영 번역데이터 성능 비교



한-영 번역 모델이 성능이 낮은 이유는?



딥러닝은 블랙박스 모델이기에 정확한 해석은 힘들지만
그래도 예상해보자면...



04 토이 프로젝트 결과



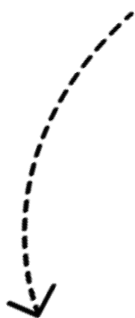
1. 한국어 데이터와 한-영 번역데이터 성능 비교



한-영 번역 모델이 성능이 낮은 이유는?

다음의 이유일 것으로 추측

- ① 번역 성능이 충분하지 않음
- ② 고유명사 번역의 문제



“혜화수산”이라는 상호명을

“Hyehwa Seafood Restaurant”가 아니라 “Hyehwa SUSAN”으로 반영



04 최종 결과

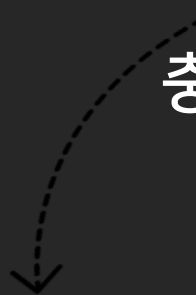


1. Multi Modal 과 영-한 번역 모델 비교

한-영 번역 모델의 성능이 낮은 이유는?



다음의 이유인 것으로 추측
하지만 **한-영 번역 모델의**
① 번역 성능의 불완전성
평가 지표 점수도 꽤 높기 때문에
② 문맥을 충분히 반영하지 못함
충분히 활용 가능하다고 판단!



“혜화 수산”이라는 상호명을

“Hyehwa **Seafood Restaurant**”가 아니라 “Hyehwa **SUSAN**”으로 반영

이제 KeyBERT를 통한
해시태그 출력으로 가보자구~





04 토이 프로젝트 결과



2. KeyBERT 결과 비교



한국어 기반

수도권에서 가까운 위치, 문산천을 따라 걷는 산책코스, 한여름 더위를 날려버릴 시원한 물놀이장 등 가족이 함께 즐기기 좋은 캠핑장이다...



위치 # 눈썰매장 # 캠핑장
한여름 # 문산천

번역 모델 기반

Located close to the metropolitan area, it is a good camping site for families to enjoy together, including a walking course along Munsancheon Stream, and a cool water playground to blow away the midsummer heat...



camping # outdoor # summer
playground # pool



04 토이 프로젝트 결과



2. KeyBERT 결과 비교



한국어 KeyBERT 방식

형태소 분석기를 통해 명사 추출



추출한 명사들을 SBERT를 통해 수치화



문맥과의 코사인 유사도를 계산해 키워드 추출

영어 KeyBERT 방식

```
# kw_model = KeyBERT()  
# keywords = kw_model.extract_keywords(df.iloc[0,6])
```



영어는 교착어가 아니기 때문에
바로 코사인 유사도를 계산해
키워드 추출이 가능함



04 토이 프로젝트 결과



2. KeyBERT 결과 비교



TEST_07277에 대한
image_data

한국어 기반

위치 # 눈썰매장 # 캠핑장
한여름 # 문산천

번역 모델 기반

camping # outdoor # summer
playground # pool

한국어 기반, 번역 모델 기반 각각에서
다른 스타일의 해시태그 출력 가능



04 토이 프로젝트 결과



3. 최종 결과 출력

카테고리 분류와
키워드 추출을 통한
최종 해시태그 출력!



Psat_deep
Hyewha, Seoul



♥ 610 Likes

Psat_deep

Strawberry cream cake is a famous bakery.
The representative menu is bread. It is a cafe
located in Songpa-gu, Seoul.
[#food](#) [#restaurant](#) [#bar/café](#) [#bakery](#)
[#strawberry](#) [#cake](#) [#cafe](#) [#seoul](#)



Psat_deep
Hyewha, Seoul



♥ 610 Likes

Psat_deep

딸기생크림 케이크가 유명한 베이커리이다.
대표메뉴는 빵이다. 서울특별시 송파구에 있는 카페다.
[#음식](#) [#음식점](#) [#바/카페](#) [#카페](#) [#생크림](#)
[#송파구](#) [#베이커리](#) [#딸기](#)

05 성과 및 한계





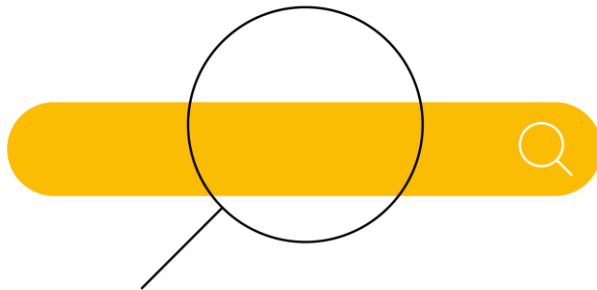
05 성과 및 한계



1. 성과



검색의 접근성 향상



#음식 #음식점 #바/카페 #카페

#생크림 #송파구 #베이커리 #딸기

카테고리도 해시태그로 활용함으로써



검색의 접근성이 높아짐



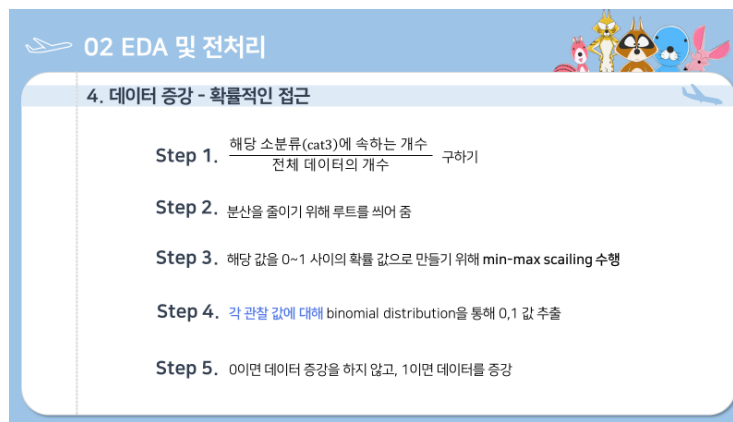
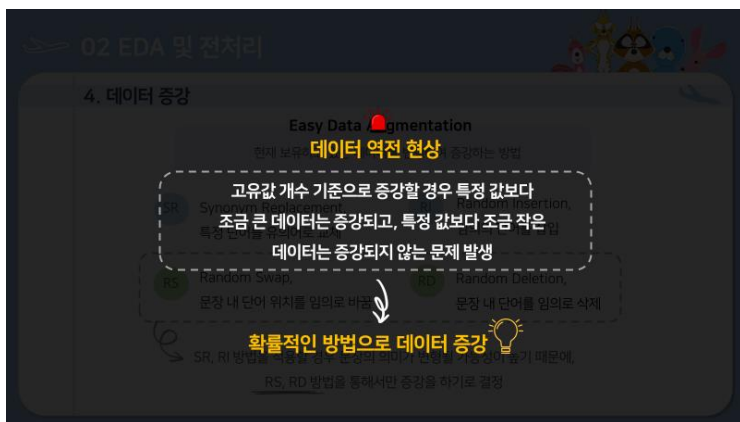
05 성과 및 한계



1. 성과



라벨 수 적은 데이터 예측에 성공



1주차 주제분석 PPT를 참고해주세요~

확률적인 방법으로 데이터를 증강하는 방법을 고안💡

➔ 불균형한 데이터임에도 라벨 수가 적은 데이터 예측에 성공



05 성과 및 한계



2. 한계



데이터셋의 문제

id	Img_path	Overview	Cat1	Cat2	cat3
TRAIN_09562	./image/train/TRAIN_09562.jpg	구매탄 시장은 경기도 수원시 영통구의 유일한 재래시장이다. Wn50년 전통의 역사를...	쇼핑	쇼핑	상설시장
TRAIN_09572	./image/train/TRAIN_09572.jpg	영산포풍물시장은 일제강점기에 형성된 장으로, 영산포 포구가 번성하였을 때에는 서남해...	쇼핑	쇼핑	5일장
...

▲ 텍스트 데이터 셋



▲ 이미지 데이터

라벨링이 애매하고, 이미지 데이터의 일관성이 부족해 명확한 분류에 한계가 존재



05 성과 및 한계



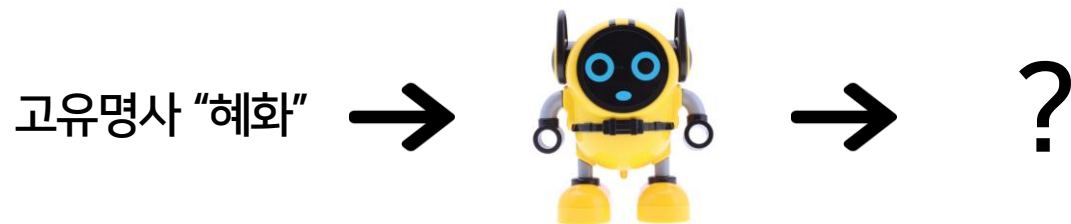
2. 한계



기계번역 성능의 한계

기계번역모델을 구현하는 과정에서
Tokenizer의 선택과 적은 데이터셋으로 인해
성능이 잘 나오지 않았음

예시) 기계번역모델에서 '고유명사' 를 Unknown 토큰으로 반환





05 성과 및 한계



3. 제언



데이터 분류 관련 제언

한국관광공사가 관광지점 정보 (POI: point of interest) 데이터를 구축할 때

① 카테고리를 더 세분화하고

② 개념이 겹치지 않게 분류할 것

③ 일관된 이미지 데이터 구축



[라벨 16211]경기도 남양주시의 천마산 산자락에 자리잡고 있는 몽골문화촌은 1998년 남양주시와...

[라벨 150] 충남 북부 해안에 접해있는 단호박올리고 마을은 서해안 고속도로 충남의 관문에...

각각 "이색거리", "농, 산, 어촌체험"으로 분류되어 있음

각 분류보다는 "이색문화마을"이라는 카테고리를 새로 생성



05 성과 및 한계



3. 제언



데이터 분류 관련 제언

한국관광공사가 관광지점 정보 (POI: point of interest) 데이터를 구축할 때

① 카테고리를 더 세분화하고

② 개념이 겹치지 않게 분류할 것

③ 일관된 이미지 데이터 구축

[라벨 97] 서울시 강남구 신사동에 자리한 근린**공원**이다. 1970년 3월 10일 박정희 전 대통령은 도산 안창호 선생이 이 나라 자주와 독립을 위해...

[라벨 158] 도심에 있는 큰 규모의 **공원**으로 가족들과 나들이하기에 좋은 곳이다. 나무 데크 길과 운동기구, 쉼터 등이 잘 구성되어 있으며, 공원을...

각각 "자연생태관광지", "공원"으로 분류되어 있음

"공원"이라는 개념이 겹침



05 성과 및 한계



3. 제언



데이터 분류 관련 제언

한국관광공사가 관광지점 정보 (POI: point of interest) 데이터를 구축할 때

- ① 카테고리를 더 세분화하고
- ② 개념이 겹치지 않게 분류할 것
- ③ 일관된 이미지 데이터 구축



Ex) 음식점에 관련된 카테고리는
음식사진이든 가게사진이든 하나로 통일



05 성과 및 한계



4. 의의, 기대효과



의의, 기대효과



① 관광정보의 생산을 인공지능의 힘으로 자동화
더 적은 공공의 예산으로 더 많은 POI 데이터 만들 수 있음

② 이를 해시태그로 활용하여
SNS상에서 국내 관광정보의 접근성을 향상할 수 있음



국내 관광 활성화에 도움

06 후기





06 후기

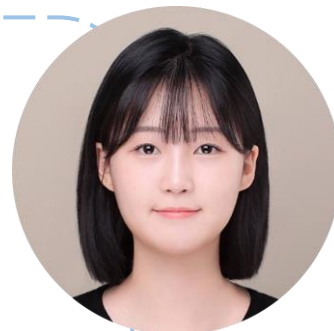


딥러닝 보나(우주소녀)



한 학기 동안 예찬 오빠와 윤아, 민이, 승민이랑 함께하는 딥러닝 팀원으로 활동할 수 있어서 너무 행복했습니다,,,,♡ 능력자 팀장님과 팀원들을 만나서 정말 많이 배우고 성장할 수 있는 기회였던 것 같습니다. 벌써 피셋에 들어온 지 1년이 다 되어 가네요,,,, 뿌듯하기도 하고 아쉽기도 합니다ㅎㅎ 이번 학기 다들 너무 고생 많았고 우리 팀 두 민이들은 한학기 더 파이팅이야!!

딥러닝 임현주



정말 배워보고 싶었던 분야였던 딥러닝팀의 일원으로서 한학기를 보낼 수 있어서 행복했습니다. 어렵고 복잡한 내용일수도 있는데, 팀장님께서 교안을 정말 정성스럽게 써주신 덕분에, 그리고 궁금한 점이나 모르는 점이 있으면 잘 알려주는 우리 팀원분들 덕분에 덜 힘들게 배우며 지금까지 올 수 있었던 것 같습니다. 부족한 팀원과 함께 하느라 다들 고생했어요 아직 부족하지만 여기서 배운걸 시작으로 저는 멋진 사람이 될 거랍니다! 한 학기동안 감사했습니다!



06 후기



딤러닝 다니엘



한 학기 동안 딤러닝 팀으로 함께하게 되어서 정말 영광이었습니다! 클린업 기간에도, 주분 기간에도 정말 배울 게 많고 부족한 점이 많다는 걸 느꼈습니다ㅠㅠ 킹왕짱 딤예찬님과 피셋의 자랑 시언언니, 윤아언니, 승민이와 함께 하면서 많이 배우고 성장한 거 같습니다! 다들 고맙습니다! 딤러닝팀 사랑해요!!!!!!

딤러닝 강남

어느새 벌써 한 학기가 지나간 주제분석 3주차 입니다. 통계에서 생소한 딤러닝이라는 분야를 처음 배웠던 게 엇그제 같은데 벌써 끝이라니 감개무량하네요. 인생의 그 어느 때보다도 가파른 기울기를 가지고 성장했고 피셋을 통해 이것보다 더 빠르게 성장할 것이라는 기대도 생기네요. 그러면서 만나게 된 딤러닝 팀원, 예찬, 시언, 윤아, 민 외의 다른 피셋 구성원들도 더 나은 나를 위한 좋은 자양분이었던 것 같습니다. 다시 생각해봐도 쉬운 일은 없었고 아쉬운 일도 많은 시원섭섭한 한 학기였습니다. 29기 및 학회장팀 분들은 다시 만났을 때 더 높은 곳에서 뵙기를, 같은 30기 동기들은 다음 학기에 더 나은 피셋을 위해 잘 부탁드립니다. 한 학기 동안 고생 많으셨습니다.





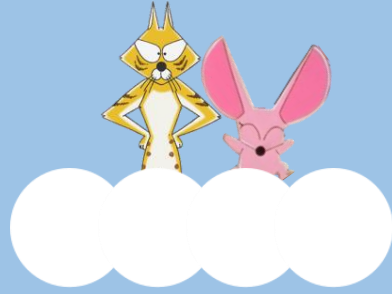
06 후기



딥러닝 래원



로피탈!! 을 외치던게 어제같은데 벌써 피셋이 끝나다니 시간 정말 빠르군요. 맨 처음 피셋에 들어와서 딥러닝팀에서 지멘과 함께 딥러닝을 처음 배웠었는데 부족한 제가 참 뭐라고 이렇게 팀장까지 하게되어 주제분석을 이렇게 마무리짓게 되었습니다.. 그래도 우리 딥러닝 이사님 시언이, 딥러닝 국장님 윤아, 딥러닝 다니엘 민이, 그리고 차기 딥러닝 팀장을 맡을 승민이까지 모두랑 같이 잘 헤쳐나가서 잘 마무리할 수 있었던 것 같네요.. 피셋 안하면 큰일난다고 해서 들어왔었는데, 정말 안했으면 큰일났을 뻔 했네요. 피셋을 통해 정말 많이 성장할 수 있는 기회였던 것 같습니다! 다른 팀 분들도 정말 고생많으셨구, 신입분들은 남은 한 학기동안 열심히 성장하셨으면 좋겠습니다. 저희는 먼저 갈게요! 하하하하하하



감사합니다