Summary

1. 사용한 데이터 셋

- **AI hub의 요약문 및 레포트 생성 데이터에 있는 데이터 셋을 다운로드 받아와 사용함**
- https://www.aihub.or.kr/aihubdata/data/view.do currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582
- 데이터 구축 규모

데이터 종류	원문 규모	어노테이션 규모	결과 규모		비고
			추출요약	생성요약	
뉴스기사	27,000	59,400	14,850	29,700	2~3문장 추출
			14,850		20% 추출
보도자료	20,000	44,000	11,000	22,000	2~3문장 추출
			11,000		20% 추출
역사_문화재	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
보고서	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
회의록	34,000	74,800	18,700	37,400	2~3문장 추출
			18,700		20% 추출
사설	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
간행물	10,000	22,000	5,500	11,000	2~3문장 추출
			5,500		20% 추출
연설문	40,000	88,000	22,000	44,000	2~3문장 추출
			22,000		20% 추출
문학	12,000	26,400	6,600	13,200	2~3문장 추출
			6,600		20% 추출
나레이션	10,000	22,742	5,640	11,371	2~3문장 추출
			5,731		20% 추출
총계	183,000	403,342	201,671	201,671	

※생성요약은 1문장 요약으로 100글자 이하로 구축

- 학습데이터 정보

데이터 종류	학습용(Training)	검증용(Validation)	테스트용(Test)	합계
뉴스	21,600	2,700	2,700	27,000
보도자료	16,000	2,000	2,002	20,002
역사기록물 & 문화재	8,000	1,000	1,002	10,002
보고서	8,000	1,000	1,000	10,000
회의록	27,200	3,400	3,400	34,000
사설	8,000	1,000	1,000	10,000
간행물	8,000	1,000	1,000	10,000
연설물	32,000	4,000	4,000	40,000
문학	9,600	1,200	1,200	12,000
나래이션	8,371	1,000	1,000	10,371
합계	146,771	18,300	18,304	183,375

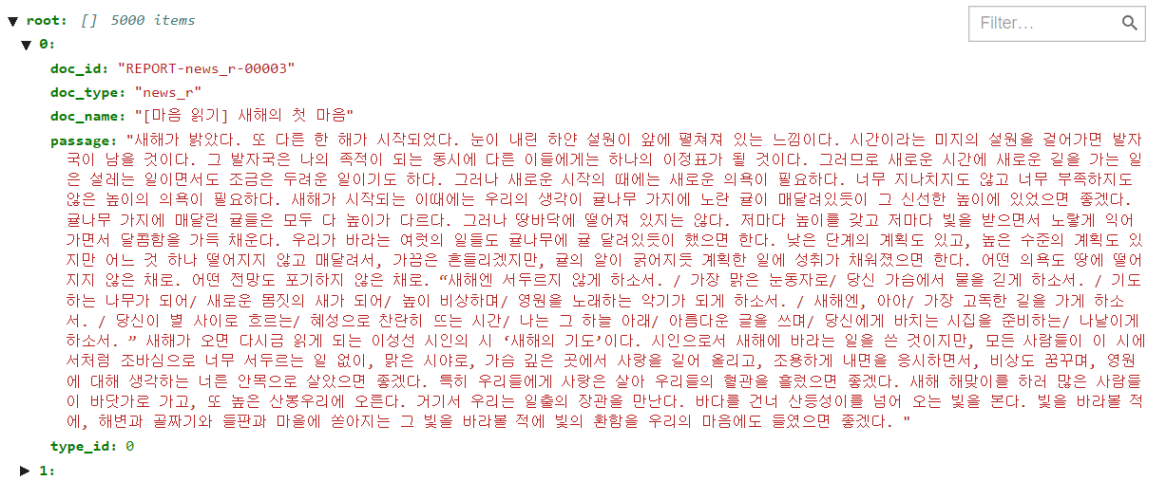
2. 데이터 량

- AI-HUB **요약문 및 레포트 생성 데이터** 활용 : 한국어 원문 문서 범주 10개

typeId	데이터 종류	doc_type
0	뉴스	news_r
1	보도자료	briefing
2	역사기록물&문화재	his_cul
3	보고서	paper
4	회의록	minute
5	사설	edit
6	간행물	public
7	연설물	speech
8	문학	literature
9	나래이션	narration

- 균등 데이터 셋으로 구성

- train data set : 각 분야별 문서 4000개 * 범주 10개 (총 40000개)
- valid data set : 각 분야별 문서 500개 * 범주 10개 (총 5000개)



└─test.json

→ 균형 데이터 사용함으로 불균형 데이터 학습보다 성능이 좋아짐

→ 학습에 사용되는 항목은 ‘passage’ 와 ‘doc_type’ 레이블

3. 학습 모델
   - ratsnlp nlpbook에 소개된 beomi/kcbert-base 모델 선정.
   - BertConfig

     https://ratsgo.github.io/nlpbook/docs/language_model/tutorial/#그림1-pretrained_model_config

```
BertConfig {
  "_name_or_path": "beomi/kcbert-base",
  "architectures": [
    "BertForMaskedLM"
  ],
  "attention_probs_dropout_prob": 0.1,
  "directionality": "bidi",
  "gradient_checkpointing": false,
  "hidden_act": "gelu",
  "hidden_dropout_prob": 0.1,
  "hidden_size": 768,
  "initializer_range": 0.02,
  "intermediate_size": 3072,
  "layer_norm_eps": 1e-12,
  "max_position_embeddings": 300,
```

```
  "model_type": "bert",
  "num_attention_heads": 12,
  "num_hidden_layers": 12,
  "pad_token_id": 0,
  "pooler_fc_size": 768,
  "pooler_num_attention_heads": 12,
  "pooler_num_fc_layers": 3,
  "pooler_size_per_head": 128,
  "pooler_type": "first_token_transform",
  "position_embedding_type": "absolute",
  "transformers_version": "4.2.2",
  "type_vocab_size": 2,
  "use_cache": true,
  "vocab_size": 30000
}
```


4. 학습 환경 설정 (args 설정)

- batch size : 512
  - 에폭 수와 학습량에 따른 조절 수치임
  - 모델과 GPU 환경의 최대 가용 배치 사이즈 수치는 512
- learning rate : 5e-5
- max_seq_length : 256
  - 데이터를 보고 토큰 수 처리 기준은 1024 = 2^10으로 하려함.
  - AI hub에서 가져온 데이터의 글자 수가 대충 700~1000자 사이임
  - 하나의 텐서가 가용할 수 있는 수치 : 300 미만
  - 따라서, 256으로 설정하였음.
- epochs : 79
  - epochs = train data / batch_size = 40000 / 512 ≒ 79

5. 프로젝트 코드 트리 구조

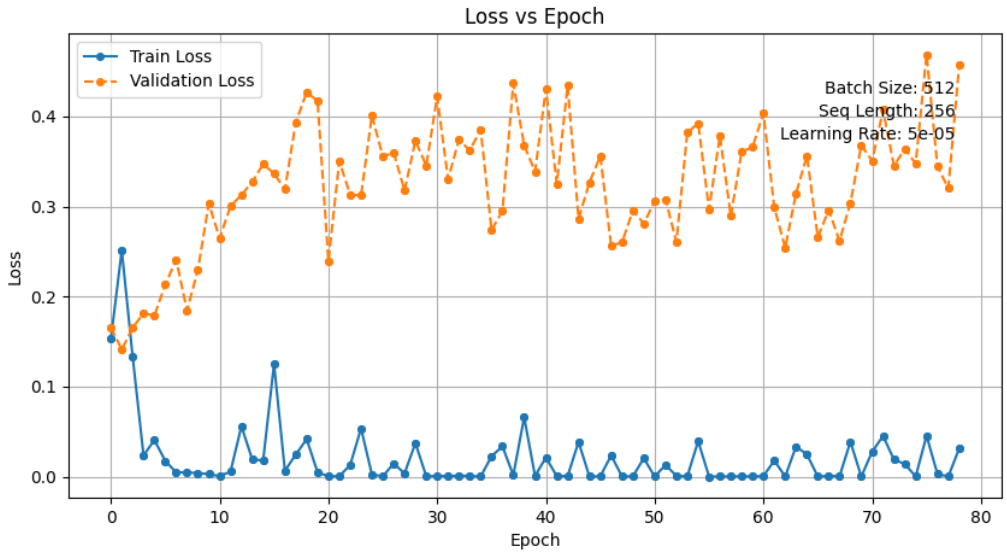
   ratsnlp 실습 코드를 수정해서 활용. [link]

   data , train/valid/test, model

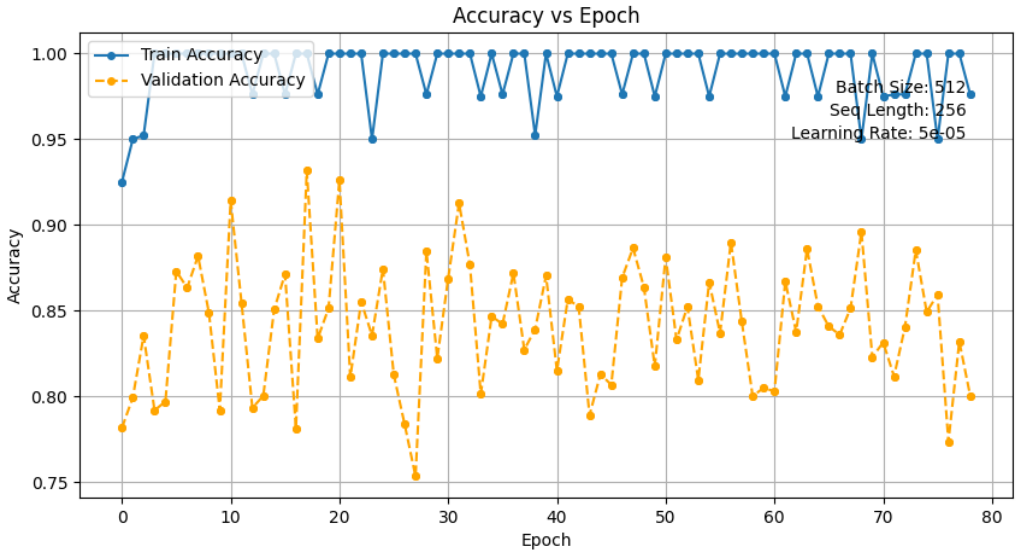
```
📦task12-main
 ┣ 📂documents
 ┃ ┣ 📜briefing.txt
```

```
│ ├ 📜edit.txt
│ ├ 📜history.txt
│ ├ 📜koreabank.txt
│ ├ 📜meeting.txt
│ ├ 📜meeting2.txt
│ ├ 📜meeting3.txt
│ ├ 📜minute1.txt
│ ├ 📜news.txt
│ ├ 📜news2.txt
│ ├ 📜test05.txt
│ ├ 📜ti.txt
│ └ 📜untitled.txt
├ 📂model
│ └ 📜readme.md
├ 📂plots
│ ├ 📜readme.md
│ ├ 📜test07_acc_batch512_seq256_epoch79_lr5e-05.png
│ └ 📜test07_loss_batch512_seq256_epoch79_lr5e-05.png
├ 📂report
│ ├ 📜data_frame.ipynb
│ ├ 📜readme.md
│ ├ 📜test.json
│ └ 📜train.json
├ 📂src
│ ├ 📂ratsnlp
│ │ ├ 📂nlpbook
│ │ │ ├ 📂classification
│ │ │ │ ├ 📜arguments.py
│ │ │ │ ├ 📜corpus.py
│ │ │ │ ├ 📜deploy.py
│ │ │ │ ├ 📜task.py
│ │ │ │ └ 📜__init__.py
│ │ │ ├ 📜data_utils.py
│ │ │ ├ 📜metrics.py
│ │ │ ├ 📜trainer.py
│ │ │ ├ 📜utils.py
│ │ │ └ 📜__init__.py
│ │ └ 📜__init__.py
│ ├ 📜readme.md
│ └ 📜requirements.txt
├ 📂train_result_csv
│ ├ 📜loss_acc_info_batch512_seq256_epoch79_lr5e-05_test07.csv
│ └ 📜readme.md
├ 📜.gitignore
├ 📜classification_finetuning.ipynb
├ 📜doc_cls_deploy_finetuning.ipynb
├ 📜LICENSE
├ 📜README.md
├ 📜requirements.txt
├ 📜setup.py
└ 📜summary.pdf
```

6. 학습 결과 및 성능



Loss 그래프



Accuracy 그래프

loss_acc_info_batch512_seq256_epoch79_lr5e-05_test07.csv

	Batch Size	Max Seq Length	Epochs	Learning Rate
1	512	256	79	5e-05
2	Epoch	Train Loss	Train Accuracy	Validation Loss,Validat...
3	0	0.15314006805419922	0.925000011920929	0.1659211218357086...
4	0	0.15314006805419922	0.925000011920929	0.1659211218357086...
5	1	0.25144848227500916	0.949999988079071	0.1409371942281723,...
6	1	0.25144848227500916	0.949999988079071	0.1409371942281723,...
7	2	0.1332518756389618	0.9523809552192688	0.1654251217842102,...
8	2	0.1332518756389618	0.9523809552192688	0.1654251217842102,...
9	3	0.023548580706119537	1.0	0.1814982891082763...
10	3	0.023548580706119537	1.0	0.1814982891082763...

156	76	0.0029196147806942...	1.0	0.3445680737495422...
157	77	7.699000707361847e-...	1.0	0.3202364146709442,...
158	77	7.699000707361847e-...	1.0	0.3202364146709442,...
159	78	0.031169572845101357	0.976190447807312	0.456879585981369,0....
160	78	0.031169572845101357	0.976190447807312	0.456879585981369,0....

▼ 문서 추론 결과 (**실제값, 추론값**) 추론이 맞았을 경우

- 0~9까지 문서 범주별로 확률적으로 통계를 내어서 결과를 보여줌

▼ 예시1: 뉴스 (인터넷에서 크롤링).

Content of the file:
국책연구기관인 한국개발연구원(KDI)은 최근 우리나라 경제가 서비스업과 제조업을 중심으로 경기 부진이 점진적으로 완화되고 있지만, 글로벌 하방 위험은 여전하다고 진단했다.

KDI가 7일 발표한 '경제동향 8월호'을 통해 "6월 전(全)산업생산은 전월(-1.1%)보다 높은 1.1%의 증가율을 기록했다"며, 이같이 밝혔다.

제조업은 평균가동률(72.8%→71.9%)이 다소 낮은 수준에 머물러 있으나, 재고율(122.7%→111.4%)이 대폭 하락하면서 부진 완화를 시사하고 있다는 것이 KDI의 분석이다. 특히, 반도체 생산은 4월 마이너스(-) 21.6%에서 5월 -18.7%, 6월 -15.9%를 기록했다.

KDI는 "반도체의 생산 감소폭이 축소되는 가운데, 출하와 재고 지표들이 개선되고 수출물량이 크게 증가했다"라며, 경기 부진의 주요인인 반도체 지표가 증가하면서 경기의 부진 완화를 시사하고 있다"라고 내다봤다.

◆…(사진=KDI)

서비스업생산도 완만한 증가세를 지속하고 있다. KDI에 따르면, 서비스업생산은 전월(1.9%)보다 증가폭이 확대된 3.5%로, 서비스소비가 완만한 증가세를 유지하고 있다.

6월 수출의 경우, 전월(-6.0%)보다 낮은 -16.5%의 증가율을 기록했다. 자동차(58.3%→15.0%)의 증가폭이 축소된 가운데, 반도체(-33.6%)와 석유제품(-42.3%)은 수출가격 하락으로 감소세가 지속되고 있다. 조업일수 변동과 기저효과 등 일시적 요인들이 일부 품목에 영향을 미쳤지만, 반도체 지표가 증가하면서 제조업 부진이 완화되고 있다는 것이 KDI의 설명이다.

상품소비를 반영하는 6월 소매판매도, 내구재를 중심으로 전월(-0.6%)보다 높은 1.4%의 증가율을 기록했다. KDI는 개별소비세 인하 종료에 대비한 수입차 구매 증가로 내구재(1.9%→8.2%) 증가세가 확대된 영향으로 풀이했다. 전월 대비로도 1.0% 증가해 부진이 완화되는 분위기다.

7월 소비자심리지수는 지난달 100.7(기준치 100)을 기록한 데 이어, 이번 달도 103.2로 기준치를 상회하며 상승 흐름을 지속하고 있다. KDI는 "소비자심리지수가 상승세를 지속하는 가운데, 승용차 소매판매가 크게 증가하는 등 소비 부진이 일부 완화되고 있다"라고 진단했다.

6월 설비투자는 -0.6%의 증가율을 기록했다. 설비투자 관련 선행지수도 전달과 동일한 수준으로 부진한 상황이다. 한국은행에 따르면, 8월 설비투자 기업경기실사지수(BSI) 전망(90)은 전월과 동일하게 낮은 수준을 유지하고 있다. 또, 운송장비를 중심으로 감소폭(-4.5%→-0.6%)이 일시적으로 축소했고, 선행지표의 부진으로 설비투자 수요가 제한적인 상황이다.

한편, KDI는 글로벌 경기 하방 위험을 경계해야 한다고 제언했다.

KDI는 "원자재 가격 상승과 중국의 경기회복 지연 등 글로벌 경기 하방 위험이 높게 유지된다"라며, "최근 유가가 상승한 가운데 러시아·우크라이나 전쟁 등의 지정학적 요인과 기상여건 악화로 곡물가격 급등에 대한 우려가 증대되고 있다"라고 전했다.

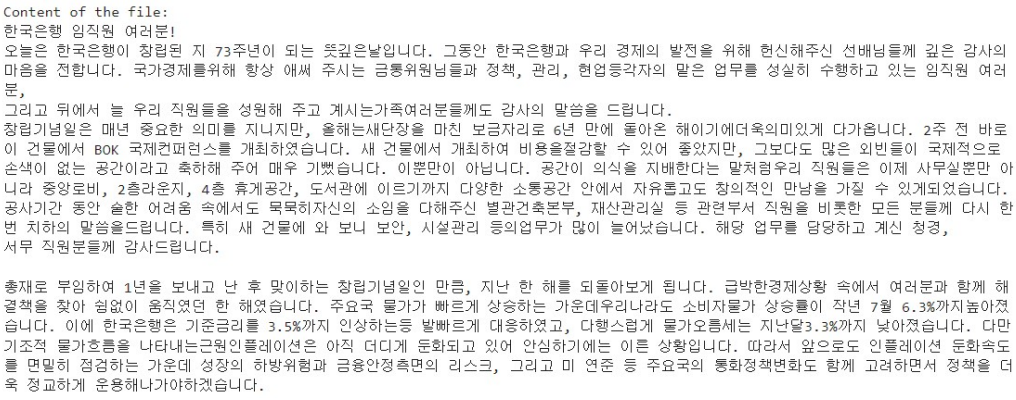
또, "2023년 세계 경제성장률 전망이 소폭 상향 조정됐으나, 인플레이션과 그에 따른 통화긴축, 중국의 경기회복 지연 등 경기 하방 위험은 여전히 높은 상황"이라고 진단했다.

└뉴스 원문.

{'prediction': '뉴스', 'probs': {'뉴스': 0.976, '보도자료': 0.0127, '역사기록물': 0.0005, '보고서': 0.0029, '회의록': 0.001, '사설': 0.002, '간행물': 0.0019, '연설문': 0.0022, '문학': 0.0003, '나래이션': 0.0004}}

└추론 결과 : '뉴스' 0.976의 확률로 추론.

▼ 예시2: 연설문 (인터넷에서 크롤링).



연설문 원문.

{'prediction': '연설문', 'probs': {'뉴스': 0.0004, '보도자료': 0.0409, '역사기록물': 0.0003, '보고서': 0.0014, '회의록': 0.0001, '사설': 0.0005, '간행물': 0.0013, '연설문': 0.9541, '문학': 0.0004, '나래이션': 0.0006}}

추론 결과: ‘연설문’ 0.9541의 확률로 추론.

1. briefing(보도자료)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0205, **'보도자료': 0.3573**, '역사기록물': 0.0079, '보고서': 0.0063, '회의록': 0.0016, '사설': 0.0019, '간행물': 0.0048, **'연설문': 0.5981**, '문학': 0.001, '나래이션': 0.0006}}
2. edit(사설)
   {'prediction': '뉴스', 'probs': {**'뉴스': 0.959**, '보도자료': 0.001, '역사기록물': 0.0004, '보고서': 0.0135, '회의록': 0.0015, **'사설': 0.0071**, '간행물': 0.0162, '연설문': 0.0006, '문학': 0.0005, '나래이션': 0.0003}}
3. history(역사 기록물)
   {'prediction': '역사기록물', 'probs': {'뉴스': 0.0174, '보도자료': 0.0566, **'역사기록물': 0.6604**, '보고서': 0.1267, '회의록': 0.0029, '사설': 0.0062, '간행물': 0.1183, '연설문': 0.0028, '문학': 0.0023, '나래이션': 0.0064}}
4. koreabank(연설문)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0003, '보도자료': 0.0115, '역사기록물': 0.0002, '보고서': 0.0007, '회의록': 0.0001, '사설': 0.0002, '간행물': 0.0013, **'연설문': 0.985**, '문학': 0.0004, '나래이션': 0.0002}}
5. meeting(회의록)
   {'prediction': '보도자료', 'probs': {'뉴스': 0.0014, '**보도자료': 0.9946**, '역사기록물': 0.001, '보고서': 0.0004, **'회의록': 0.0003**, '사설': 0.0001, '간행물': 0.0012, '연설문': 0.0005, '문학': 0.0003, '나래이션': 0.0002}}
6. meeting2(회의록)
   {'prediction': '연설문', 'probs': {'뉴스': 0.0099, '보도자료': 0.0567, '역사기록물':

0.0011, '보고서': 0.0691, **'회의록': 0.0031**, '사설': 0.0182, '간행물': 0.214, **'연설문': 0.6171**, '문학': 0.0076, '나래이션': 0.0031}}

7. meeting3(보고서) - 있던 데이터 원본 txt
   {'prediction': '보고서', 'probs': {'뉴스': 0.0013, '보도자료': 0.0024, '역사기록물': 0.0261, **'보고서': 0.6062**, '회의록': 0.0074, '사설': 0.0183, '간행물': 0.0215, '연설문': 0.0495, '문학': 0.0049, '나래이션': 0.2624}}
8. minute1(회의록) - 있던 데이터 원본 txt
   {'prediction': '회의록', 'probs': {'뉴스': 0.0001, '보도자료': 0.0001, '역사기록물': 0.0002, '보고서': 0.0002, **'회의록': 0.9988**, '사설': 0.0001, '간행물': 0.0001, '연설문': 0.0001, '문학': 0.0002, '나래이션': 0.0001}}
9. news(뉴스)
   {'prediction': '뉴스', 'probs': {**'뉴스': 0.9841**, '보도자료': 0.0004, '역사기록물': 0.0002, '보고서': 0.0104, '회의록': 0.0003, '사설': 0.0031, '간행물': 0.0005, '연설문': 0.0002, '문학': 0.0002, '나래이션': 0.0004}}
10. news2(사설) - 있던 데이터 원본 txt
    {'prediction': '간행물', 'probs': {'뉴스': 0.0036, '보도자료': 0.0005, '역사기록물': 0.0102, '보고서': 0.1301, '회의록': 0.0019, **'사설': 0.2032**, **'간행물': 0.6316**, '연설문': 0.0004, '문학': 0.0134, '나래이션': 0.0052}}
11. test05(사설) - 있던 데이터 원본 txt
    {'prediction': '보고서', 'probs': {'뉴스': 0.0136, '보도자료': 0.0008, '역사기록물': 0.0023, **'보고서': 0.8138**, '회의록': 0.0009, **'사설': 0.0934,** '간행물': 0.0718, '연설문': 0.0011, '문학': 0.0005, '나래이션': 0.0018}}
12. ti(문학작품)
    {'prediction': '보고서', 'probs': {'뉴스': 0.0061, '보도자료': 0.001, '역사기록물': 0.0016, **'보고서': 0.5984**, '회의록': 0.001, '사설': 0.0875, '간행물': 0.0464, '연설문': 0.0013, **'문학': 0.2553**, '나래이션': 0.0014}}
13. untitled(뉴스)
    {'prediction': '뉴스', 'probs': {**'뉴스': 0.9105**, '보도자료': 0.0033, '역사기록물': 0.0007, '보고서': 0.0578, '회의록': 0.0006, '사설': 0.0125, '간행물': 0.0104, '연설문': 0.0036, '문학': 0.0003, '나래이션': 0.0003}}