

Project Proposal

텍스트マイ닝

NLP 활용 문제 생성을 통한 문해력 향상

5 조

2019204045 윤서환

2019204023 윤성호

2019204094 이건희

2019204013 김승현

Contents

|| 목차 ||

／ 배경 소개 ／ 데이터 수집

／ 문제 생성 ／ 사용자 평가 ／ 최종 결론

1/배경 소개

선정 배경

MZ세대 문해력 저하 심각 수준...책 대신 스마트폰 즐겨보니 글 읽어도 "이게 무슨 뜻?"

"무슨 말인지 모르겠다. 누가 세 줄 요약 좀!"

| OECD 조사, 우리나라 실질 문맹률은 75% 달해
EBS 조사, 중3생 중 27%가 적정 문해력 미달
중3생 중 11%는 초등생 수준의 문해력 보여

1/배경 소개

선정 배경

기업들 절반 "영어잘하는 사람 채용 우선"

배샛별 기자 venus1203@pbj.co.kr

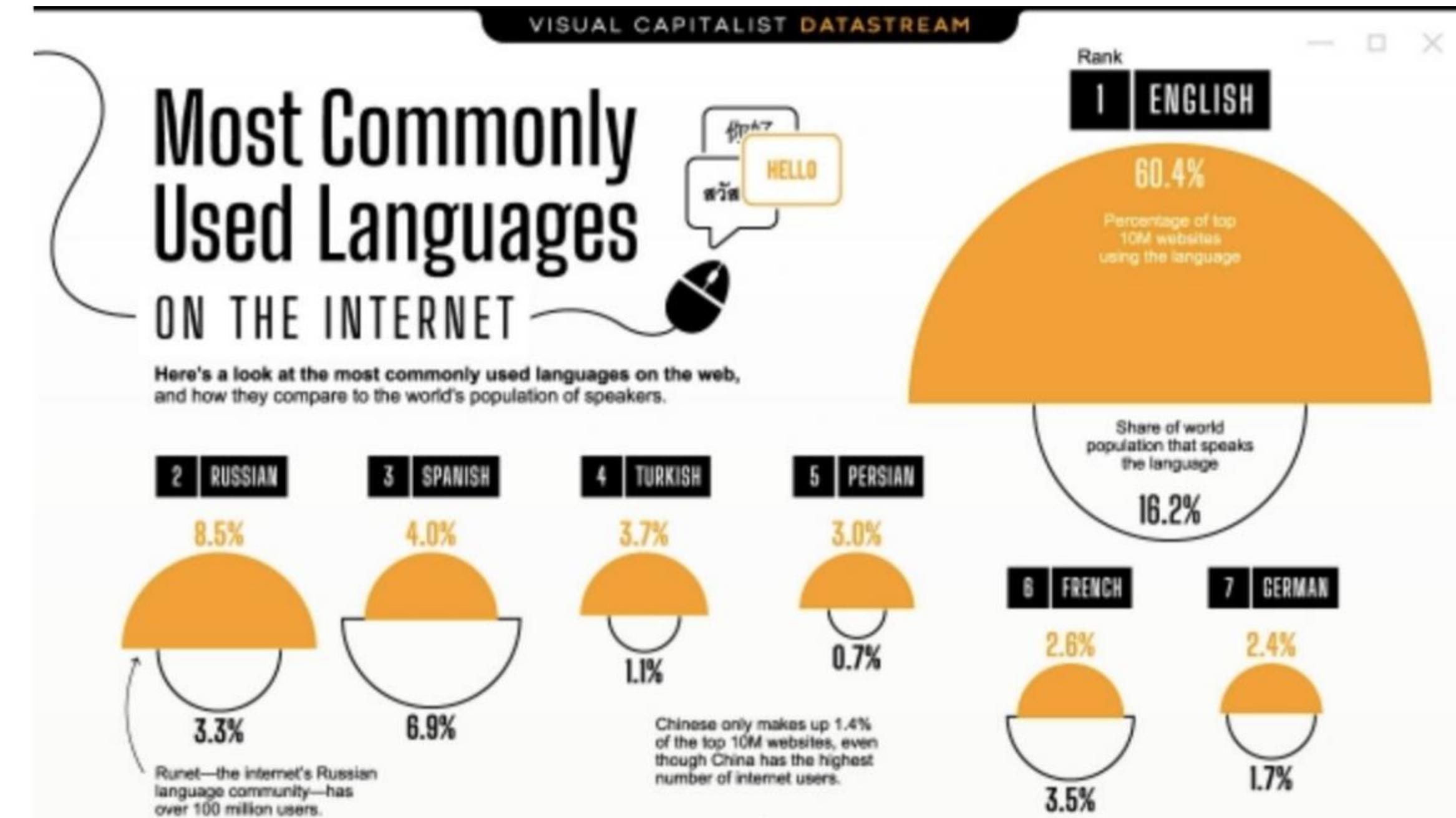
[프라임경제] 대졸 신규직원을 뽑는 기업 2곳 중 1곳은 채용시 영어능력을 중요시하는 것으로 나타났다.

[온라인 취업사이트 “사람인” 조사]

인사담당자 370명을 대상,

34.6%가 ‘우수한 영어 실력으로 연봉을 더 받는 직원이 있다’라고 응답.

46%는 영어실력이 승진 시 유리한 영향을 미친다고 응답했다고 합니다.



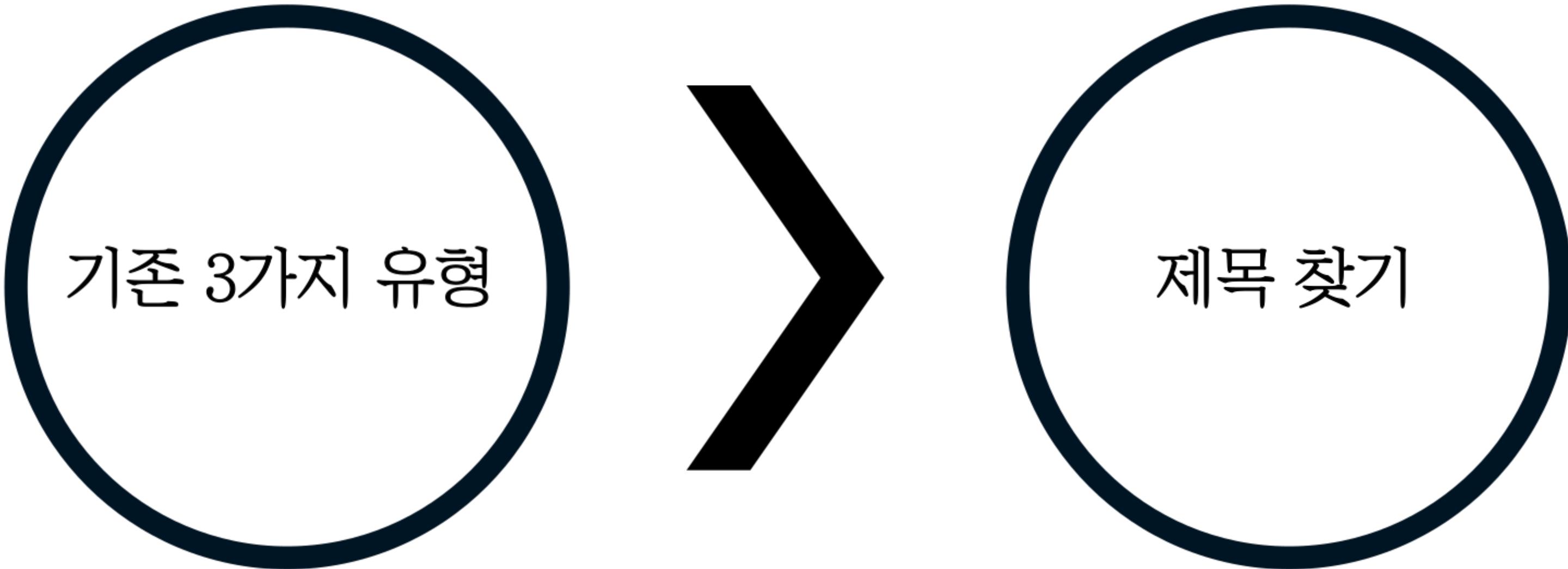
1/배경 소개

【 배경소개 】

- 문해력 저하 및 영어의 중요성 대두
- 최근 동향과 변화 기사를 통한 습득
- 수능 문제 유형 참고 및 활용
- 문해력 및 영어 실력 향상, 사회적 동향 함양 기대

1/배경 소개

문제 유형



FeedBack -> 문해력 향상 초점

텍스트 마이닝 기반 보기 선정, 문제 풀이

1/ 배경 소개

파이프라인

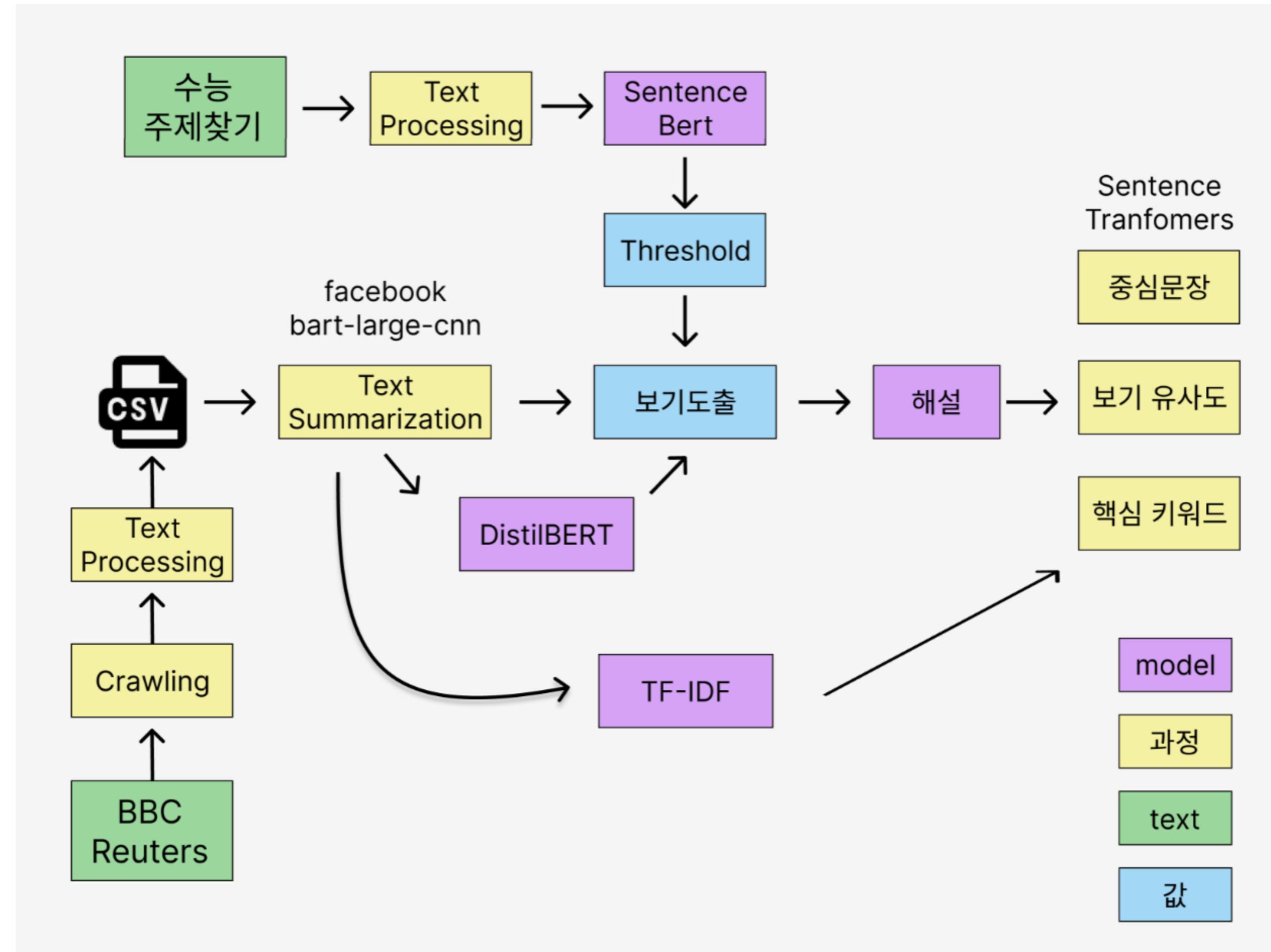
전체 파이프라인

기사 추상요약

수능 오답 유사도 측정

문제 보기 적용

해설



2023학년도 텍스트마이닝 6월 문해력 향상 평가지

문해력 영역

성명	
----	--

수험 번호		-							
-------	--	---	--	--	--	--	--	--	--

- 문제지의 해당란에 성명과 수험 번호를 정확히 쓰시오.
- 답안지의 필적 확인란에 다음의 문구를 정자로 기재하시오.

하버드, 예일, 프린스턴, 광운 Let's go

2/데이터 수집

수집과정



검증된 기사

특정 토픽 &
크롤링

문제 출제
해설 제공

2/데이터 수집

수집과정

BBC 크롤링 결과 예시

	Title	Content	Topic
0	Taurine may extend life and health, scientists...	Taurine - a nutrient found in meat, fish and s...	health
1	Weekend hot weather alert escalated to amber	A weekend heat-health alert has been raised fr...	health
2	Humza Yousaf 'disturbed' by claims NHSGGC spie...	First Minister Humza Yousaf has said he was "d...	health
3	UK heatwave: How does sunscreen work?	A heat-health alert has been issued for parts ...	health
4	Heatwaves: What do they do to the body and who...	A heat-health alert has been issued for parts ...	health
...
213	Jersey-UK reciprocal health agreement updated	A reciprocal health agreement (RHA) between Je...	health
214	'My mum donating her kidney was a gift of life...	An 18-year-old medallist at the World Transpla...	health
215	Savannah Marshall: Boxer admits school struggl...	Former world champion boxer Savannah Marshall ...	health
216	Is the past year's spike in strike action here...	Over the past 12 months, a historic wave of in...	health
217	Health boss says obesity 'bigger health impact...	The long-term impact of obesity could be great...	health

2/데이터 수집

수집 과정

Reuters 크롤링 예시

df_social			
	Title	Content	Topic
0	Analysis: City distances itself from hedge fun...	LONDON, June 10 Big banks and investors quickl...	social_impact
1	'Anti-ESG' funds face slowing deposits, Mornin...	June 9 (This June 8 story has been corrected t...	social_impact
2	US minority-owned banks withstand industry tur...	NEW YORK, June 9 Minority-owned banks in the U...	social_impact
3	US companies face less pressure for climate an...	June 9 Shareholder support for proxy resolutio...	social_impact
4	US DOJ settles probe of illegal dumping in Hou...	WASHINGTON, June 6 The U.S. Department of Just...	social_impact
...
215	Ex-Goldman Sachs banker seeks mercy in 1MDB case	Feb 25 Roger Ng, the former Goldman Sachs Grou...	social_impact
216	Abu Dhabi's IPIC and Aabar to pay Malaysia \$1....	KUALA LUMPUR, Feb 27 Abu Dhabi's International...	social_impact
217	Comment: We can help African countries hold on...	February 24 - Harrison Karnwea is in his late ...	social_impact
218	Majority of global asset managers still not in...	LONDON, Feb 26 Global asset managers controlli...	social_impact
219	Dentsu Group president accepts culpability in ...	TOKYO, Feb 25 The president of Japanese advert...	social_impact

총 4개의 Topic

- Health
- Technology
- Social Impact
- Economy

각 토픽당

200 ± 20 데이터

3/ 문제 생성

제 목 찾기

24. 다음 글의 제목으로 가장 적절한 것은?

Invasions of natural communities by non-indigenous species are currently rated as one of the most important global-scale environmental problems. The loss of biodiversity has generated concern over the consequences for ecosystem functioning and thus understanding the relationship between both has become a major focus in ecological research during the last two decades. The “biodiversity-invasibility hypothesis” by Elton suggests that high diversity increases the competitive environment of communities and makes them more difficult to invade. Numerous biodiversity experiments have been conducted since Elton’s time and several mechanisms have been proposed to explain the often observed negative relationship between diversity and invasibility. Beside the decreased chance of empty ecological niches but the increased probability of competitors that prevent invasion success, diverse communities are assumed to use resources more completely and, therefore, limit the ability of invaders to establish. Further, more diverse communities are believed to be more stable because they use a broader range of niches than species-poor communities.

* indigenous: 토착의 ** niche: 생태적 지위

- ① Carve Out More Empty Ecological Spaces!
- ② Guardian of Ecology: Diversity Resists Invasion
- ③ Grasp All, Lose All: Necessity of Species-poor Ecology
- ④ Challenges in Testing Biodiversity-Invasibility Hypothesis
- ⑤ Diversity Dilemma: The More Competitive, the Less Secure



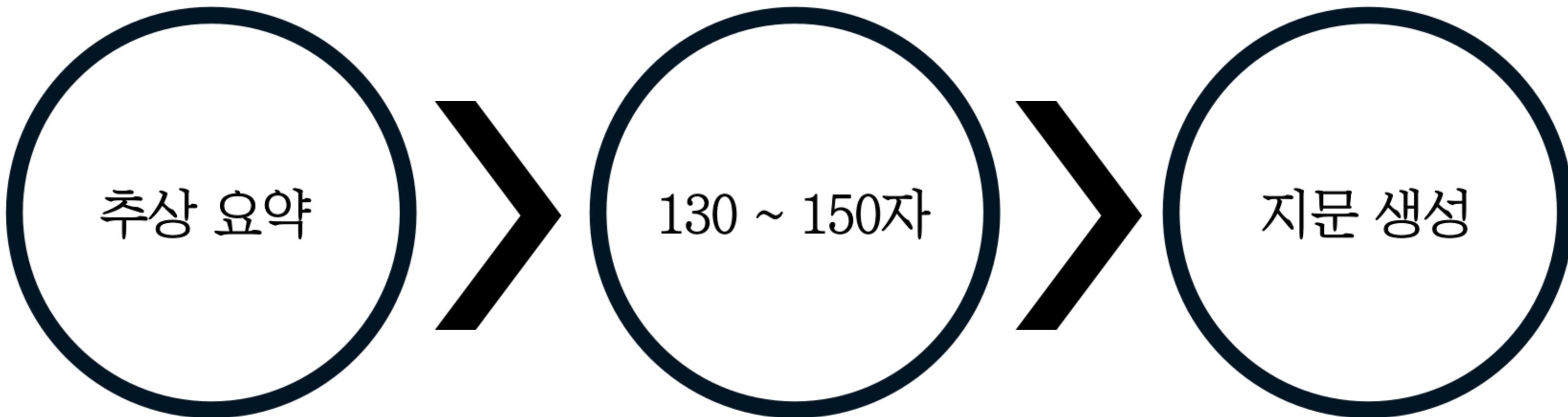
1. 지문 생성

3/ 문제 생성

제 목 찾기



수능 영어지문의 글자 수는 130자, 많아야 150자 이내이다.



3/ 문제 생성

제목 찾기

text=""There is a critical need for a new maternity strategy in Northern Ireland, as midwives deal with a growing number of "complicated pregnancies", the Royal College of Midwives (RCM) has said.
A new report detailed mounting challenges for maternity services as the demographics of pregnant women have "rapidly" changed.
But it added that it was facing a tightening budget and would have to decide which work could be done with its available funding.
"Services are coping because of the incredible efforts of staff who go above and beyond on a daily basis, often to the detriment of their mental and physical health," she added.
"We are seeing far more complicated pregnancies and this is increasing the demands on midwives and their colleagues.
They want to ensure these women get the best possible support, advice and care. That is proving more and more difficult without the resources to back them up."
Speaking to BBC Radio Ulster's Good Morning Ulster programme, Ms Murray said issues which arise during pregnancy can be difficult for expectant mothers to deal with alone.
"These conversations are easier if you have a continuity relationship with a midwife that you know and you trust," she said.
"If you are in a situation where you're meeting a different professional every time you come for an appointment it's much more difficult to address those issues."
The number of births in Northern Ireland has fallen since 2011 from just above 25,000 to 20,500 in 2022.
But the RCM said demand on services could not be measured by that alone and that the demographic profile of women giving birth had changed significantly over the same period.
There was an increasing number of women being cared with additional health needs, the report stated.
It said less than half of pregnant patients in Northern Ireland had a weight in the healthy range.
They are also much more likely to have been diagnosed with diabetes than a decade ago, which can cause problems for mothers and babies.
In 2012-13 the number was 899. It grew to 3,177 in 2020-21.
There were also more births registered to women aged 32 or over in 2021 compared with 2011.
The RCM said older women were more likely to experience higher rates of pregnancy-related conditions and higher rates of assisted birth, leading to longer stays in hospital.
The State of Maternity Services 2023 report said "all of these factors and more have massively increased and complicated the workload of our midwives".
The RCM said Northern Ireland needed to take several steps to deliver high quality, safe, care for women and babies, including the funding and publication of a new maternity and neonatal strategy.
It would help design and plan services based on current trends.
Northern Ireland is the only region in the UK without an active maternity strategy. The last one ended in 2018.
The report said: "Demand for maternity services cannot be managed through waiting lists, so it is vitally important to base workforce planning on evidence rather than assumptions."
It said without a strategy, policies risked "being piecemeal, contradictory and without buy-in from everyone involved".
The RCM said a new strategy had cross-party support but was "held back by the frustrating continuation of the suspension of the assembly and executive" in Northern Ireland.
In response, the Department of Health said an independent review of midwifery services was due to be published in autumn this year.
"It is envisaged that this programme of work will help form the basis of a new maternity strategy in due course," it said.
The RCM report comes amid wide-scale budgetary pressures for the Department of Health, which has seen cuts to some health programmes and services.
The RCM welcomed an increase in student midwives, with more than 70 graduating each year.
But retention of existing staff also needed to be prioritised in order to see the workforce grow. The report pointed to 129 midwife vacancies across Northern Ireland in December 2022.
"We need urgently for politicians to get back to work in the institutions. Added to this we need more midwives, something we can achieve not just through continued investment in more training but by keeping the staff we have," the report added.
"How? We pay them better and treat them better."
In March midwives in Northern Ireland voted to take industrial action over pay.
The Department of Health's permanent secretary recently said it was impossible to give pay rises under the current budget.
Peter May said the health service was facing unfunded spending pressures of £472m of which £375m relates to a pay claim matching that of NHS England.""

```
model_name = "facebook/bart-large-cnn"
tokenizer = BartTokenizer.from_pretrained(model_name)
model = BartForConditionalGeneration.from_pretrained(model_name)

inputs = tokenizer([text], truncation=True, padding='longest', return_tensors='pt')

summary_ids = model.generate(inputs['input_ids'], num_beams=4, min_length=200, max_length=220, early_stopping=True)
summary = [tokenizer.decode(s, skip_special_tokens=True, clean_up_tokenization_spaces=False) for s in summary_ids]

for line in summary.split("."):
    print(line)
```

추상 요약 결과(예시)

Northern Ireland is the only region in the UK without an active maternity strategy
The last one ended in 2018
Department of Health said it was committed to working with the RCM to improve maternity services
But it added that it was facing a tightening budget and would have to decide which work could be done with its available funding
The RCM welcomed an increase in student midwives, with more than 70 each year
But the report pointed out that there were still 129 midwives vacancies across Northern Ireland in December 2022
It said the health service was facing unfunded spending pressures of £472m of permanent claim money, which relates to a pay claim of £375m to midwives
The report said: "Demand for maternity services cannot be managed through waiting lists, so it is vitally important to base workforce planning on evidence rather than assumptions" It said a new strategy had cross-party support but was "held back by the frustrating continuation of the suspension of the

3/ 문제 생성

보기 선정

24. 다음 글의 제목으로 가장 적절한 것은?

Invasions of natural communities by non-indigenous species are currently rated as one of the most important global-scale environmental problems. The loss of biodiversity has generated concern over the consequences for ecosystem functioning and thus understanding the relationship between both has become a major focus in ecological research during the last two decades. The “biodiversity-invasibility hypothesis” by Elton suggests that high diversity increases the competitive environment of communities and makes them more difficult to invade. Numerous biodiversity experiments have been conducted since Elton’s time and several mechanisms have been proposed to explain the often observed negative relationship between diversity and invasibility. Beside the decreased chance of empty ecological niches but the increased probability of competitors that prevent invasion success, diverse communities are assumed to use resources more completely and, therefore, limit the ability of invaders to establish. Further, more diverse communities are believed to be more stable because they use a broader range of niches than species-poor communities.

* indigenous: 토착의 ** niche: 생태적 지위

- ① Carve Out More Empty Ecological Spaces!
- ② Guardian of Ecology: Diversity Resists Invasion
- ③ Grasp All, Lose All: Necessity of Species-poor Ecology
- ④ Challenges in Testing Biodiversity-Invasibility Hypothesis
- ⑤ Diversity Dilemma: The More Competitive, the Less Secure

2. 보기 생성 및 선정



3/ 문제 생성

보기선정

94년 ~ 22년 수능 제목/주제 지문, 보기 추출



수능 지문 중심 문장과
오답 보기와의 유사도 파악

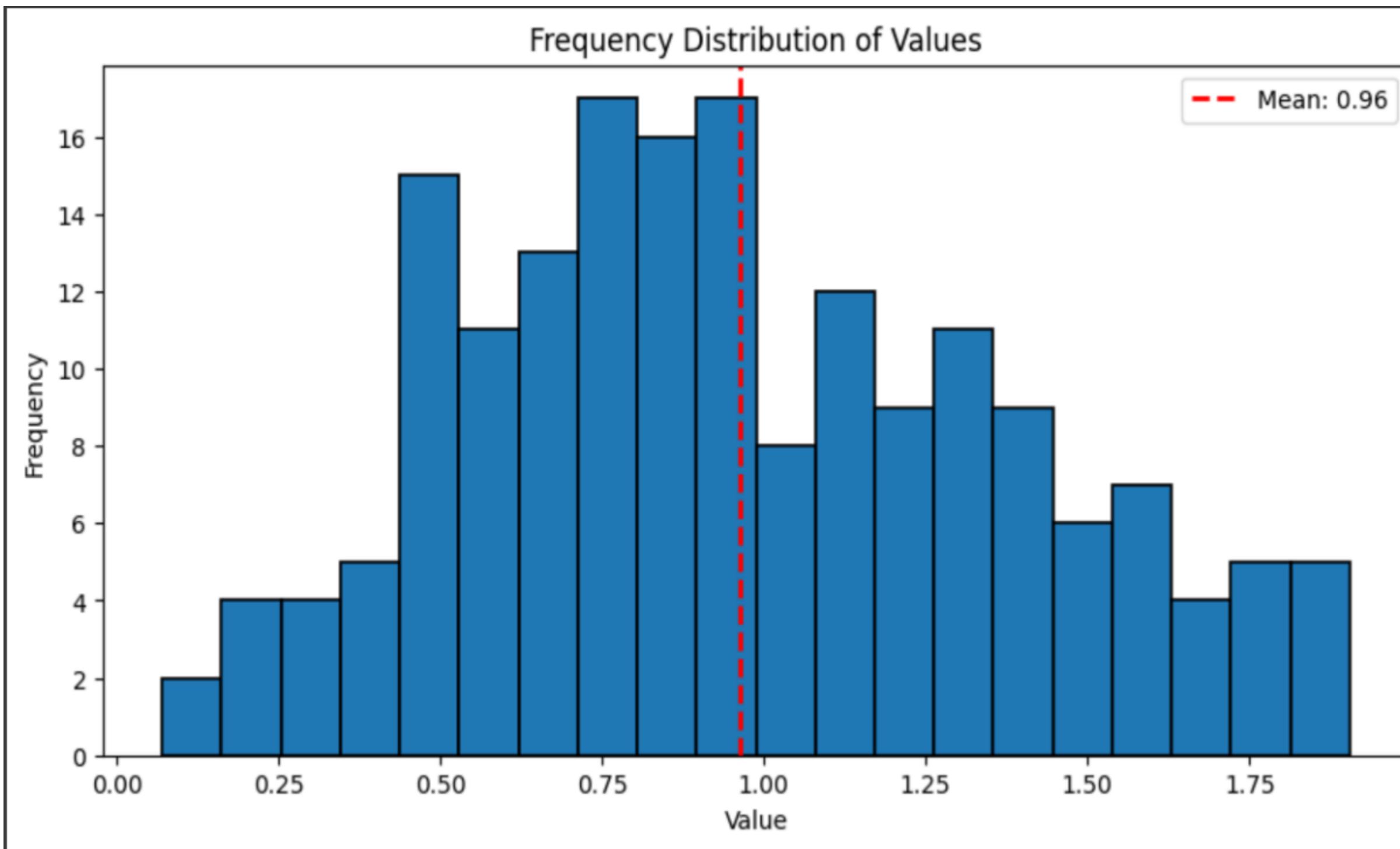


정답과 오답 간 차이(유사도) 도출

3/ 문제 생성

보기선정

지문 핵심 문장 4개 | 오답 보기 유사도합 비교



분포를 참고
다양한 Threshold
설정하여 난이도 조절

오답 선정 접목

3/ 문제 생성

보기선정

토픽 별 제목

200개

핵심 문장 4개

보기 유사도 계산

상위 3개

하위 1개 보기

3/ 문제 생성

보기선정

- ('Causeway Hospital: Mums worried about births moving from Coleraine', 1.2799387574195862)
- ('Airedale hospital maternity services require improvement, say CQC inspectors', 1.2110414057970047)
- ('Stockton health hub to be built on former shopping site', 1.2050038576126099)
- ("East Kent maternity deaths: Staff 'committed' to improvements", 1.164552167057991)
- ("NI maternity strategy 'critical' as complicated pregnancies rise", 1.1576507240533829)
- ('Daisy Hill Hospital: Staff notified of further changes', 1.1530259251594543)
- ('Propped-up hospitals in the East of England to be rebuilt by 2030', 1.141053944826126)
- ('Coleraine: No more births to take place at Causeway Hospital', 1.1331289559602737)
- ('Ageing hospital to be rebuilt by 2030', 1.112069457769394)
- ('Mental health services that fail to improve could be shut, says watchdog', 1.1031081974506378)
- ('Watchdog says Devon NHS trust requires improvement', 1.0777377486228943)
- ("Daisy Hill Hospital's inpatient care 'at risk' in general medicine", 1.0767159909009933)
- ("Menopause and periods guidance is a 'magnificent step forward'", 1.0726054310798645)
- ('Mental health unit to open two beds for Welsh mums', 1.0623609721660614)
- ('University duty of care campaign to be debated after suicides', 1.0622076392173767)

유사도 기준
오름차순 정렬
상위 3개
하위 1개
보기 출제

3/ 문제 생성

해설

정답

해설

유사도 높은
3개 오답

유사도 낮은
1개 오답

공통 해설



지문의 중심 문장 제공

보기와 지문 중심 문장 간 유사도 제공

5/ 최종 결론

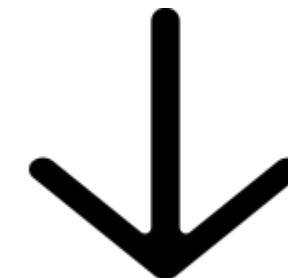
최종 결론

향후 계획

다양한 언론사로부터 더욱 방대한 데이터 확보

요약 문장의 길이 조절

핵심문장 - 보기 간 유사도 엄밀히 분석 후, 나이도에 따른 유사도 도출



문제풀이로 문해력 향상 및 사회적 동향 함양 기대

향후 App, Web 등 실용화

3/ 문제 생성

해설

유사도 높은 3개 오답시

상위 5개 중요 단어: board, health, added, sarwar, leadership

```
def preprocess_text(text):
    text = re.sub(r'₩+', ' ', text) # 특수문자를 공백으로 대체
    text = re.sub(r's+', ' ', text) # 연속된 공백 문자를 하나의 공백으로 대체
    text = text.lower().strip() # 소문자로 변환하고 양쪽 공백 제거
    return text

def remove_stopwords(text, lang="english"):
    stop_words = set(stopwords.words(lang))
    word_tokens = word_tokenize(text)
    filtered_text = [word for word in word_tokens if word not in stop_words]
    return " ".join(filtered_text)

def remove_verbs(text):
    word_tokens = word_tokenize(text)
    pos_tags = pos_tag(word_tokens) # 품사태깅 수행

    nouns = [word for word, pos in pos_tags if pos in ['NN', 'NNP', 'NNS', 'NNPS']]
    return " ".join(nouns)

def get_top_n_words(sentence, n=3):
    sentence = preprocess_text(sentence)
    sentence = remove_stopwords(sentence)
    sentence = remove_verbs(sentence)

    vectorizer = TfidfVectorizer()
    tfidf_matrix = vectorizer.fit_transform([sentence])
    feature_array = np.array(vectorizer.get_feature_names_out())
    tfidf_sorting = np.argsort(tfidf_matrix.toarray()).flatten()[:-1]

    top_n_words = feature_array[tfidf_sorting][:n]
    return top_n_words
```

품사 태깅 > 명사 추출

TF-IDF 이용

+ 주요 5개 단어 제공

3/ 문제 생성

해설

유사도 낮은

1개 오답시

이전 3가지 풀이 + 추가 문제 제공

3/ 문제 생성

시행 예시

시행 예시

다음글의 주제를 고르시오

Yoshikazu Higashitani, known on YouTube as GaaSyy, is famous for his celebrity gossip videos. He is accused of threatening to defame an actor, an entrepreneur and a designer between February and August last year. Tokyo police had been sending investigators to the UAE since May, and urging authorities there to extradite him to Japan. Earlier this year, he became the first lawmaker in Japanese history to be kicked out of parliament without ever entering it. During his seven months in office, he refused to leave his home in Dubai and did not attend a single legislative session. His absence angered the Senate – whose members unanimously voted for his expulsion. Instead, he announced on his YouTube channel that he was going to Turkey – and that he planned to donate his salary to the Turkish earthquake relief. His passport expired the following month, but he said he'd lost it. Japan's foreign minister ordered him to return his passport after Tokyo police obtained his arrest warrant in March.

- 1, 'GaaSyy: Japan YouTuber arrested over celebrity threats'
- 2, "Vodafone to cut 11,000 jobs as new boss says firm 'not good enough'"
- 3, 'Twitter engineering boss Foad Dabiri quits day after DeSantis launch glitches'
- 4, 'Tucker Carlson: Former Fox anchor to launch new show on Twitter'
- 5, 'TikTok ban: What could it mean for creators?'

정답을 고르세요

3/ 문제 생성

풀이 예시

풀이 예시

오답입니다.

해당 문제에 대한 핵심문장입니다.

['He is accused of threatening to defame an actor, an entrepreneur and a designer between February and August last year', "Japan's foreign minister ordered him to return his passport after Tokyo police obtained his arrest warrant in March.", 'Earlier this year, he became the first lawmaker in Japanese history to be kicked out of parliament without ever entering it', 'Tokyo police had been sending investigators to the UAE since May, and urging authorities there to extradite him to Japan']

각 보기별 중심문장과의 유사도를 보여드리겠습니다.

['GaaSyy: Japan YouTuber arrested over celebrity threats', 2.5407528281211853]

["Vodafone to cut 11,000 jobs as new boss says firm 'not good enough'", 2.4321629405021667]

['Twitter engineering boss Foad Dabiri quits day after DeSantis launch glitches', 2.429233133792877]

['Tucker Carlson: Former Fox anchor to launch new show on Twitter', 2.4179360270500183]

['TikTok ban: What could it mean for creators?', 0.4533626139163971]

추가적으로 해당 문제의 핵심 키워드를 보여드리겠습니다.

['year' 'police' 'passport' 'arrest' 'august']

추가적인 문제 풀이가 필요합니다. 추가 문제를 제공하겠습니다.

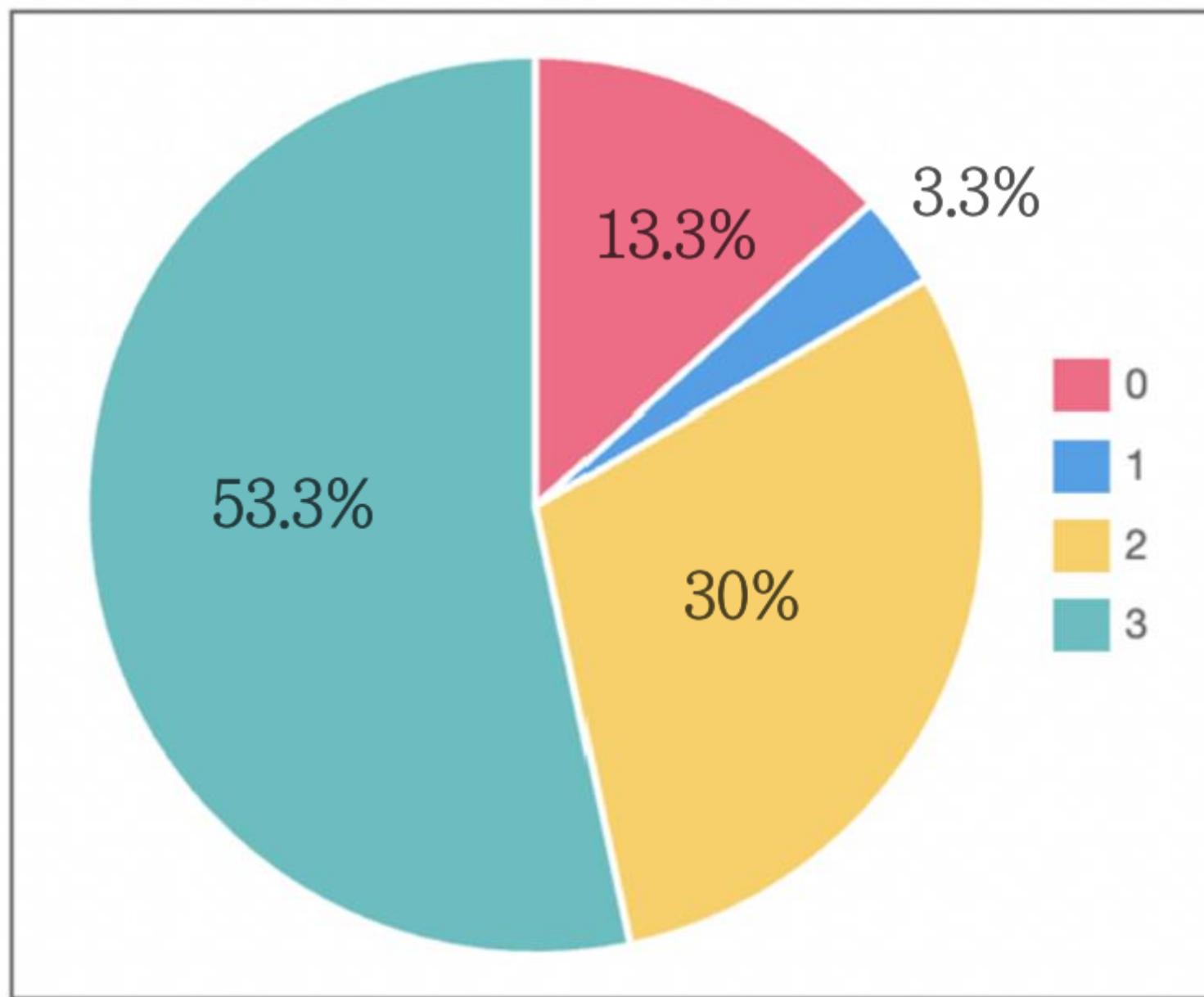
진행하겠습니까? [Y/N]

4/사용자 평가

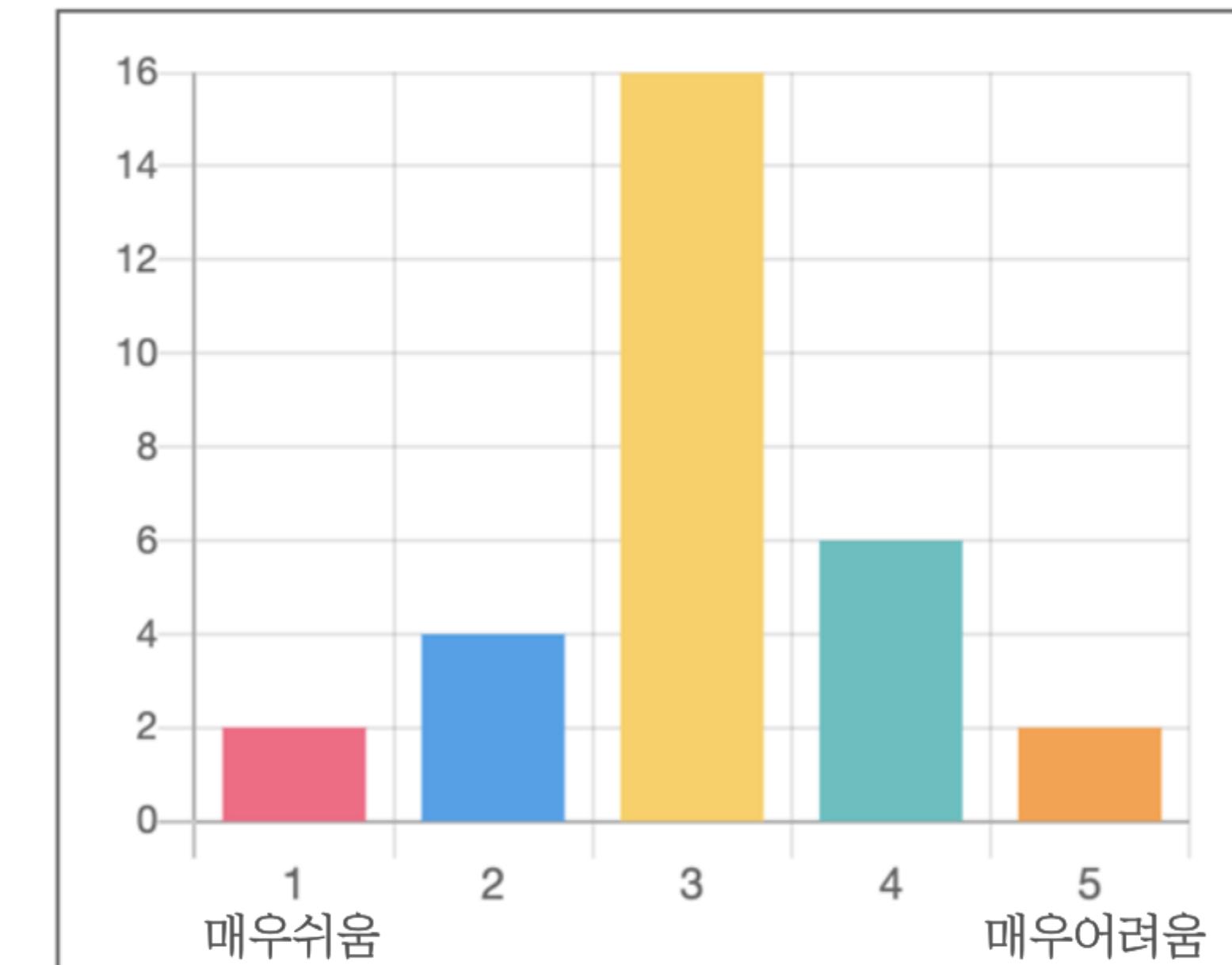
■■■유저평가■■■

고1 ~ 20대 30명 대상 생성 문제 3개에 대한 평가

1. 맞은 문제의 개수를 알려주세요



2. 문제의 난이도를 평가해주세요

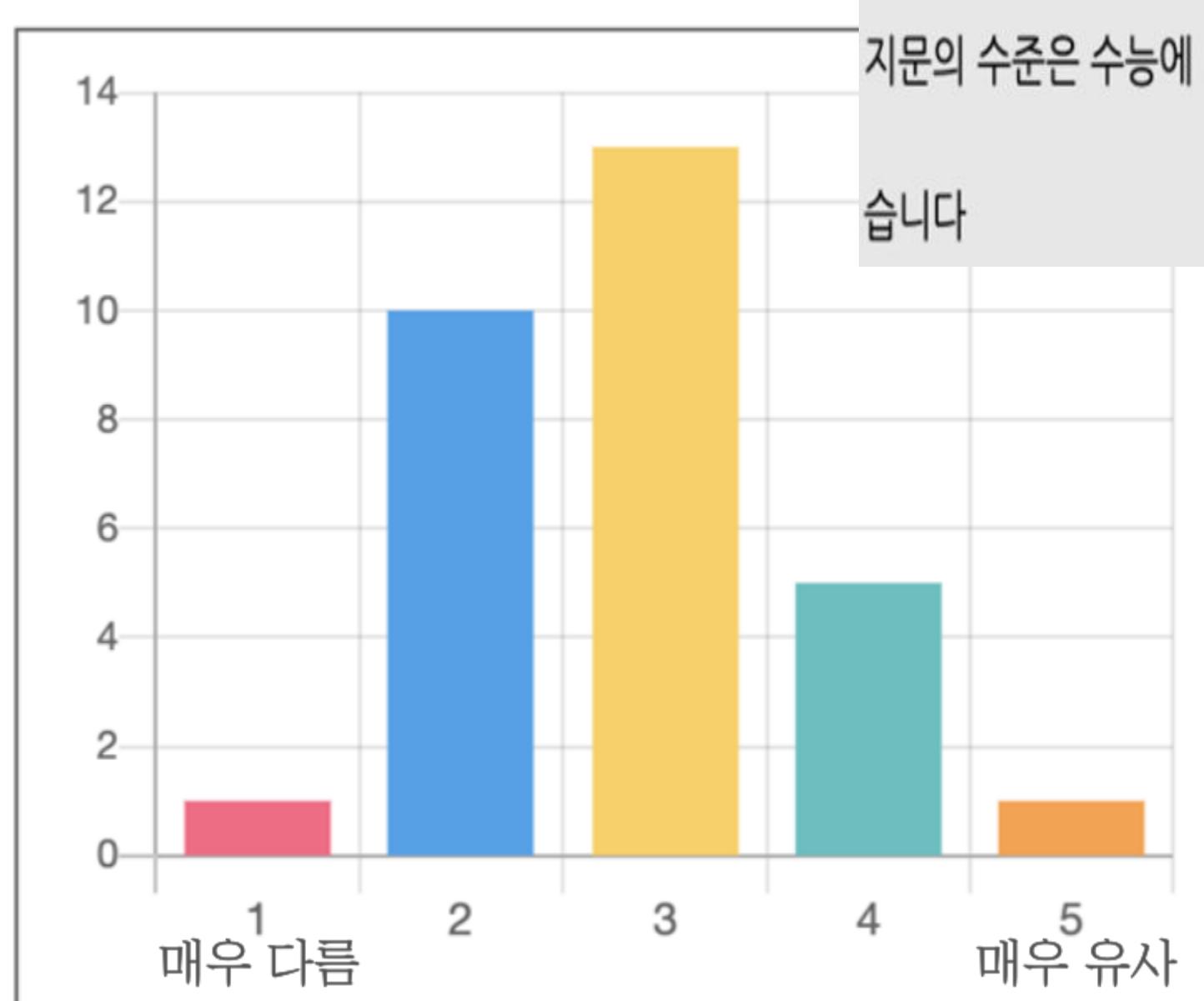


4/사용자 평가

■■■유저평가■■■

고1 ~ 20대 30명 대상 생성 문제 3개에 대한 평가

3. 해당 지문이 실제 수능과 비슷하게 느껴졌나요?



지문의 수준은 수능에 나오는 중하~중상 정도의 난이도였지만 수능 답지 보기처럼 헷갈리게 하는 요소들이 거의 없어서 정답 찍기가 수월했습니다

- 1번은 주제가 왔다갔다 하는 느낌이 있었다
- 2번은 답안 구분이 너무 쉬웠음
- 3번은 답안 구분이 까다로웠다

고유명사가 많아서 개헷갈려요ㅠㅠ

5/ 최종 결론

최종결론

한계점

지문에 고유명사가 있으면 정답을 찾기 쉽다

정답 유사도 보다 높은 유사도 오답 존재

수능 문제 길이에 맞게 추상요약 -> 가끔 해석의 어색함이 존재

유사도와 난이도가 정비례하지 않음

5/최종 결론

질의 응답

질의 응답

Q&A

5/최종 결론

질의응답

Q.

수능의 경우 실제 사용하는 언어와 차이가 심해, 원어민들이 읽는것도 어렵다고 하는데, 이런 부분에 대해서 어려운 것은 없었을까요?

A.

실제 수능 지문과 저희가 요약한 지문은 다릅니다. 저희는 실제 기사를 추상요약 하였기 때문에, 실제 사용하는 언어로 구성되어 있습니다.

Q.

학습하는 모델을 사용한다면 모델이 만드는 오답들이 너무 꼬게 되는 문제가 되어 문해력이 상승한다는 보장이 있는지, 아니면 그 정도가 아닐 정도로 오답이 nice하게 나오는지 궁금합니다!!

A.

저희는 오답들을 만들지 않습니다. 실제 기사의 제목을 모두 한 집합에 넣은 다음, 유사도 분석을 하여 설정한 기준으로 오답을 선정합니다. 실제 사용자 평가 시, 대체적으로 수능 난이도와 유사한 오답들이 선정되었다는 평가를 받았습니다.

문해력 적인 측면에서는 기사(요약 지문)를 읽고, 기사의 내용을 이해하여야 문제를 풀 수 있기 때문에, 문해력 향상 측면에서는 효과가 있다고 생각합니다.

5/최종 결론

질의응답

Q.

훌륭한 모델이라고 생각합니다. 그와 별개로, 영어 지문의 경우 이미 지문 제작자들이 많이 있고 쌓인 데이터도 많은데 해당 문제 생성 모델이 만드는 문제가 기존의 것보다 뛰어날 수 있을까요?

A.

저희는 실제 최신 기사를 요약해서 지문을 생성하기 때문에, 다른 지문들과 최신 동향을 파악 할 수 있다는 점에서 차별점이 있다고 생각합니다. 물론 사람이 직접 지문을 만드는 것보다 뛰어나다고 장담할 순 없지만, 시간적, 효율적인 측면에서는 뛰어나다고 생각합니다.

Q.

수능 형식의 문제가 정말 독해력을 높여주는게 맞을까요...?

A.

네. 저희는 문해력이 상승된다고 생각합니다. 지문을 읽고 해당 지문의 내용을 이해하여야 문제를 풀 수 있으므로, 자연스럽게 문제를 풀며 문해력이 상승됩니다.

5/최종 결론

질의응답

Q.

결국 영어 독해를 키우는 프로젝트처럼 느껴지는데, 관련 어플들이 많이 존재하는 것으로 알고 있습니다. 관련 어플의 경우 보통 생활영어를 키워드로 잡아서 독해 및 스피킹 실력향상에 초점을 두는데, 수능영어라는 데이터 셋을 선택하신 이유가 궁금합니다!

A.

저희는 최근 기사들을 지문으로 영어 문해력을 향상시키기 위한 주제로서 저희에게 친숙한 수능형식을 바탕으로 문제를 생성하고자 하였습니다. 답을 풀려면 지문을 다 읽고 중요한 문장을 생각해야되기 때문에 이러한 과정 문해력 향상에 도움이 될거라 생각해서 결정하게 되었습니다

Q.

문제풀이는 유사도로 진행을 하셨는데 절대적인 수치인가요?

A.

4개의 핵심문장에 대하여 유사도의 합을 구했기 때문에 절대적인 수치는 아니나 정답값이 2.0에 오차율 ± 0.2 사이에 해당되어서 스케일링은 따로 진행하지 않았습니다.

5/최종 결론

질의응답

Q.

MZ세대의 문해력 저하를 문제점 및 주제 선정 배경으로 삼으셨는데, 정작 결과물은 영어라서 주제와 살짝 동떨어진 느낌이 듭니다. '수능 문제 생성기'에 더 가깝다는 생각이 드는데, 이에 대해 어떻게 생각하시는지 궁금합니다.

A.

저희는 한글에 대하여 문해력을 키우자는것이 아닌 전세계인들이 가장 많이쓰는 '영어'에 대해서 문해력을 향상시키는 주제를 삼았습니다. 또한 수능 문제 형식을 가져와서 최근 해외 토픽들에 대하여 요약을 하고 문제를 생성하면, 문제를 풀어봄으로써 문해력에 향상이 될 수 있을거라 생각을 하였습니다.

Q.

어떤 사용자가 이걸 사용하고, 어떤 서비스 제공자가 이런 서비스를 제공하게 될 지 궁금합니다.

A.

수능 형식을 본 떠 기사를 잘 읽을수 있게 만들기 때문에 문해력이 부족한 사람이라면 나이 불문하고 사용자가 될 수 있고, 만약 상용화가 될 시, 학원 업계, 공공 기관 등에서 학술적 보조 수단으로 사용 될 수도 있을 것 같습니다.

5/최종 결론

질의응답

Q.

문해력의 근본은 결국 직접 문장들을 읽어나가면서 해석해나가는 것인데, 위 모델을 통해 해결할 수 있을지 궁금합니다

A.

지문의 문장을 읽어 나가면서 해석 해야지 문제를 풀 수 있으므로 해결할 수 있다고 생각합니다.

Q.

핵심단어가 내용적으로 중요한 핵심단어인지 빈도수로 추출된 핵심단어인지 구분없이 추출된 것인지 궁금합니다. 프로젝트의 의도에 따르면 지문 내용에 핵심적인 단어이어야 될 것 같아서요.

A.

발표 때도 말씀 드렸듯이 TF-IDF를 이용하여 품사 태깅 > 명사 추출 상위 5개 단어를 추출하였습니다. 저희 프로젝트에서는 지문 하나에 TF-IDF를 적용해, 빈도수 기반으로 추출된 핵심단어라고 생각합니다. 많이 사용된 단어들이 결국엔 핵심 단어라고 생각합니다. 말씀대로, 핵심 단어 추출로 추후에 TextRank를 활용하는 방안도 고려중입니다.

5/최종 결론

질의응답

Q.

문제 생성하면 그 문제에 대한 저작권은 누가 갖나요?

A.

아직 AI가 생성한 창작물 저작권에 대한 제도적인 방안이 마련 되어 있지 않습니다. 따라서 향후 법안 제도를 수용해 저작권을 부여 할 것입니다.

Q.

이미 시중에 있는 영어 문제가 많은데 차이점이 뭔지 궁금합니다

A.

저희의 의도는 최신 동향의 해외 기사를 요약해 국제 정세를 파악함과 동시에 문해력 향상을 목표로 하고 있습니다. 다른 문제집들의 문제 제작 방법을 알 수는 없으나, 실제 최신 해외 기사를 요약하기 때문에 최신 동향 파악적인 측면에서 차별점이 있다고 생각합니다.

5/최종 결론

질의응답

Q.

문해력을 높이기 위해 다양한 모델을 사용하여 나타낸 것은 좋지만, 더 다양하고 유용한 토픽을 보여주었으면 어땠을까하는 생각이 듭니다. 그리고 보여주신 분석결과보다 얼마나 더 유용하게 높일 수 있는지 궁금합니다.

A.

TF-IDF 기반과 더불어 TextRank를 통한 단어 추출로 해당 문제를 향후 보완할 계획입니다. 분석결과에 대해서는 사람별 자주 틀리는 주제와 유형에 맞춰 조금 더 다양한 풀이를 제공할 계획입니다. 그래서 개개인의 취약부분을 보완하여 문해력 향상을 기대할 수 있습니다.

Q.

지문에 고유명사가 있으면 정답을 맞추기 쉽다는 한계점이 있었는데, 이를 극복하기 위한 대안으로는 무엇이 있나요?

A.

해당 문제는 고유명사의 수가 보기 개수만큼 존재하지 않기 때문에 발생하는 문제로 ppt설명에서 말씀드렸듯이, 더 방대한 데이터를 통해, 부족한 고유명사에 대한 보완을 할 계획입니다.

5/최종 결론

질의응답

Q.

신문은 사실 전달을 위한 글이라고 생각합니다. 따라서 사실 전달에 대한 글과 문해력은 상관성이 떨어진다고 생각합니다. 실제 생성된 지문에 대한 답이 추론하기 쉬울 수도 있어 문해력 향상에 진정 도움이 되는지 견해가 궁금합니다.

A.

문해력은 글을 읽고 이해하는 능력을 뜻합니다. 사실 전달에 대한 글을 이해하여야 지식 전달이 되므로 문해력과 높은 상관성이 있다고 생각합니다. 생성된 지문의 보기는 해당 지문의 중심 문장을 기반으로 보기를 선정했기 때문에, 단순히 답을 추론하는 과정에 비해서는 난이도가 있습니다. 또 threshold를 조정하여, 문제의 보기 난이도를 조절할 수 있습니다.

Q.

모델을 통해 문제를 생성하면 결국 수능도 문제은행식이 되지 않나 싶습니다.

A.

저희는 수능문제를 만든 것이 아닌, 수능의 유형을 참고하여 기사를 기반으로 문제를 생성했습니다. 따라서 수능이 문제 은행이 된다는 것과의 상관성은 떨어진다고 생각합니다.

5/최종 결론

질의응답

Q.

보기에 대한 유사도 계산이 지문과의 유사도도 봐야하는 것 아닌가?

A.

실제 전체 지문과의 유사도 비교를 해보았으나, 실제 정답과의 유사도보다 오답의 유사도가 대부분의 경우에서 높게 나왔고, 그래서 중심 문장을 추출하여 비교를 진행했을 때, 더 높은 유사도 결과값이 나왔습니다.

Q.

실용 가능성에 대해 설명해주셨으면 합니다

A.

수능형식을 본떠서 기사를 잘 읽을 수 있게 만드는 것이기 때문에 문해력이 부족한 사람이라면 나이 불문하고 사용자가 될 수 있다고 생각합니다. 이를 토대로 문해력과 관련된 기업(ex 시원스쿨, 책을 보여주는 교보문고)에 서비스를 실행한다면 충분히 실용가능성이 있다고 생각합니다.



발	표	를		마	침	니	다
---	---	---	--	---	---	---	---

감사합니다