정수리팀의

청년취업

바 로

지금!

발표내용

01 프로젝트 소개

02 데이터 수집 및 처리

03 데이터 분석

04 서비스 구현

05 시사점 및 한계점

01 프로젝트 소개

1) 팀 소개

2) 주제 선정 배경 및 기존 서비스 분석

3) 서비스 소개

01 프로젝트 소개

1) 팀 소개

정책을 수집하는 이들

오소연

전체 문서 관리

정책 관련 자료 조사 및 데이터 수집

데이터 수집 및 DB 관리

자연어 처리 (형태소 분석)

장희정

정책 관련 자료 조사 및 데이터 수집

데이터 수집 및 DB 관리

자연어 처리 (형태소 분석)

추천 모델 구현

한민규

공공데이터 관련 데이터 수집

자연어 처리 (형태소 분석)

추천 모델 구현

챗봇 서비스 구현

01 프로젝트 소개

2) 주제 선정 배경 및 기존 서비스 분석

PROBLEM 1

지난 12월, 청년정책(일자리, 주거 등)
예산 올해 22조원, 내년 23조 5천억 집행

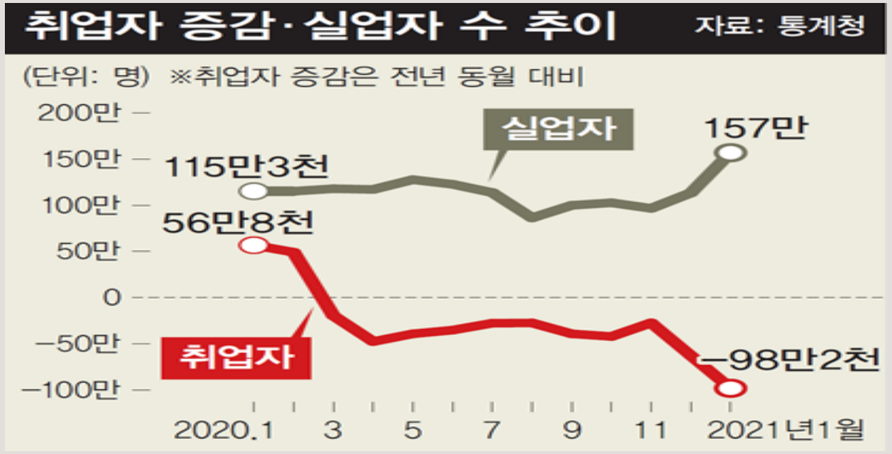
⌄

21년 상반기, 외환위기 이후 취업자 감소 최고
실업자 150만명 돌파

+

청년정책 1,900개, 미사용자 89%

출처 : 2020. 5 월간노동




청년이 없는 청년정책

중앙일보 | 입력 2021.10.04 00:34 업데이트 2021.10.04 07:20 지면보기

정부는 지난해 12월 청년정책기본계획을 발표했다. 2021년부터 2025년까지 일자리, 주거, 교육, 복지·문화, 참여·권리 등 5개 분야 청년정책이 담겼다. 올해 22조원, 내년 23조5000억원을 집행한다. 청년정책의 성공을 기원하지만, 쉽지 않을 것이다.

출처 : 2021. 8 네이버뉴스, 2021.10 중앙일보

01 프로젝트 소개

2) 주제 선정 배경 및 기존 서비스 분석

PROBLEM 2

여러 채널에 분산되어 있는

청년 정책으로 인한 불편함

01 프로젝트 소개

2) 주제 선정 배경 및 기존 서비스 분석

기존 서비스 분석

1. **키워드**로만 검색 가능
   자신의 상황을
   주체적으로 설명하기 어려움

2. 유형, 관심분야 등 단순 분류
   추천된 정책들 간의 **우선순위 부재**

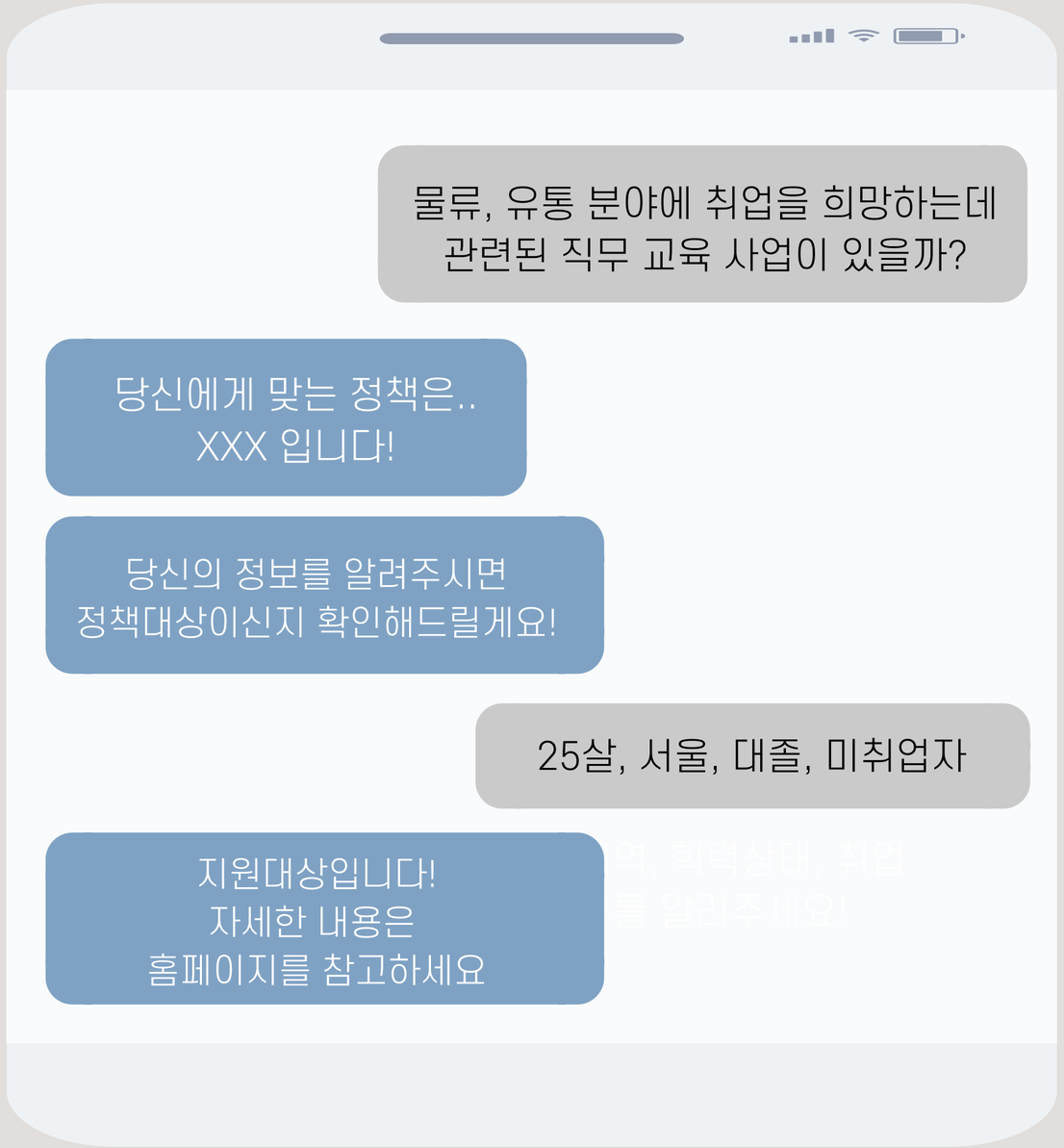
01 프로젝트 소개
2) 주제 선정 배경

PROBLEM	SOLUTION
분산되어 있는 청년 정책	전체 정책 통합하여 한번에 검색 가능
자신의 상황 주체적으로 설명하기 어려움	직접 텍스트 입력이 가능한 챗봇 서비스
추천된 정책간의 우선순위 부재	텍스트 분석을 통해 가장 필요한 정책 우선 추천

01 프로젝트 소개
3) 서비스 소개

「청바지」

NLP를 활용한
사용자 맞춤형 청년정책 추천
챗봇 서비스



02

데이터 수집 및 처리

2-1 데이터 정의

2-2 정책 데이터 전처리

2-3 상담 데이터 전처리

2-4 사용자 단어사전

2-5 카테고리 태깅

02 데이터 수집 및 처리

1) 데이터 정의

취업정책 데이터

온라인청년센터	지자체 청년포털	워크넷
799개	263개	50개
웹크롤링	타이핑	크롤링
정책명, 정책내용, 지원내용, URL, 자격요건		

정책 데이터 총 1,112

상담 데이터

포털사이트	지식인	네이버폼	워크넷
100개	111개	612개	250개
크롤링	크롤링	설문조사	크롤링
공개된 상담글 중 취업 관련 상담 내용			

상담 데이터 총 1,073

02 데이터 수집 및 처리

2) 정책 데이터 전처리

name	content	age	edu	major	job	region	special	plus	limit
국민 취업지원제도	한국형 실업부조로 고용안전망 사각지대에 있는 취업취약계층에게 취업지원서비스 ...	만 15세 ~ 69세	제한없음	제한없음	미취업자	고용노동부	제한없음	청년구직활동지원금 등 사업에 참여하였다가 2021년 1월 이후 참여가 종료된 경우, [종료일로부터 6개월 간] 국민취업지원제도 참여 불가	-학업/군복무 등으로 즉시 취업이 어려운 사람,-자치단체 청년수당 수급 중이거나 종료 후 6개월이 지나지 않은 자...
대전일자리지원센터 운영	미취업 청년들의 취업에 필요한 구인구직 상담등을 통해 진로에 맞는 일자리 제공	만 15세 ~ 39세	고교 재학	제한없음	미취업자	대전광역시	제한없음	특성화고 대상 희망학교 사전수요조사 및 학교 여건에 맞추어 컨설팅 프로그램 진행	특성화고 학생이 아닌 시민
서구형 내일채움공제(신규)	중소기업 청년근로자의 자산형성 지원을 통해 장기재직을 유도하고 중소기업 경쟁력 강화 및 고용확대 도모	만 15세 ~ 34세	제한없음	제한없음	중소기업 6개월 이상 재직	인천 서구	제한없음	-	-

정책 내용

자격 요건 중 4가지만 활용

02 데이터 수집 및 처리

2) 정책 데이터 전처리

최소값, 최대값으로 구분

전처리 전	전처리 후	
age	min_age	max_age
만 15세 ~ 69세	15	69
만 15세 ~ 39세	15	39
만 15세 ~ 34세	15	34

고교 재학, 고졸, 대학 재학, 대졸, 대학원 재학, 대학원 졸업, 제한없음

edu	edu
제한없음	제한없음
고교 재학	고교 재학
대학교 졸업예정	대학교 재학

미취업자, 재직자, 제한없음 3가지로 분류

전처리 전	전처리 후
job	job
미취업자	미취업자
미취업자	미취업자
중소기업 6개월이상 재직	재직자

시/도단위로 통일

region	region
고용노동부	전국
대전광역시	대전광역시
인천 서구	인천광역시

02 데이터 수집 및 처리

3) 상담 데이터 전처리

text
교양수업을 듣다가 K-move 사업에 대해 알게 되었는데 무엇인지 알고 싶습니다! 해외취어ㅂ 할 수 있는 건가요?
코로나때문에 직장 그만두고 ... 알바하다가 며칠 전에 첫 며ㄴ접이 잡혔어요. 아무튼 면접 정장 대여하려고 하는데 무료로 가능한곳 있나요???
아직까지 가고 싶은 산업을 정하지 못했는데, 2~3개의 산업을 염두해두고 모두 준비해야 하는지, 아니면 1가지 산업만을 준비해야 하는지 취업카페나 취업박람회처럼 상담할 수 있는 곳 추천해주세요. 고민이 있는데 혼자는 답이 안 나오네요 ㅠㅠ

1차 전처리

오탈자 교정

text
교양수업을 듣다가 K-move 사업에 대해 알게 되었는데 무엇인지 알고 싶습니다! 해외취업 할 수 있는 건가요?
코로나때문에 직장 그만두고 ...알바하다가 며칠 전에 첫 면접이 잡혔어요. 아무튼 면접 정장 대여하려고 하는데 무료로 가능한곳 있나요???
아직까지 가고 싶은 산업을 정하지 못했는데, 2~3개의 산업을 염두해두고 모두 준비해야 하는지, 아니면 1가지 산업만을 준비해야 하는지 취업카페나 취업박람회처럼 상담할 수 있는 곳 추천해주세요. 고민이 있는데 혼자는 답이 안 나오네요 ㅠㅠ

2차 전처리

사족, 이모티콘 제거

text
K-move 사업에 대해 알게 되었는데 무엇인지 알고 싶습니다! 해외취업 할 수 있는건가요?
면접 정장 대여하려고 하는데 무료로 가능한곳 있나요???
취업카페나 취업박람회처럼 상담할 수 있는 곳 추천해주세요. 고민이 있는데 혼자는 답이 안 나오네요.

02 데이터 수집 및 처리

4) 사용자 단어사전

정책 데이터

정책 이름, 정책 내용, 지원 내용

형태소 분석

konlpy mecab 활용 - nouns(명사)로 수집

problem 발견 빅데이터 -> 빅, 데이터
플랫폼 -> 플랫, 폼

사용자 단어사전

mecab 사용자 단어사전에
단어 추가 및 우선순위 조정

총 6390 단어 추가

02 데이터 수집 및 처리

5) 카테고리 라벨링

정책 데이터 1차 라벨링

단어 기반 카테고리 분류

지원금, 장려금, 수당, ~비, ...	지원금
교육,학습,강좌,양성,훈련, 수강, ...	교육
면접, 정장, 대여, 구두, 넥타이, ...	면접
일자리,구인구직,인턴십,인턴,실습, ...	일자리
컨설팅, 이력서, 상담, 취업박람회, 일자리센터, ...	상담

정책 데이터 2차 라벨링

교정을 통한 고도화

4, 5개 카테고리에 동시에 해당하는 정책 有

Ex. '국민취업지원제도' → 교육 제외 4카테고리 해당

▼

▼

수기 확인 후 맞지 않은 카테고리 제외

Ex. '국민취업지원제도' → 지원금, 일자리

02 데이터 수집 및 처리
5) 카테고리 라벨링

KoBERT(Korean BERT)

기존 BERT의 한국어 성능 한계를
극복하기 위해 SKT에서 개발한 모델

수백만 개의 한국어 문장으로 이루어진
대규모말뭉치(corpus)를 사전학습

상담 데이터 1차 라벨링 - 단어 기반 라벨링을 타겟으로
KoBERT를 활용한 카테고리 분류

X	Y	검증 정확도	테스트 정확도(30개)
교육	면접	0.933	0.5 (15/30)
	지원금	0.791	0.5 (15/30)
	일자리	0.95	0.5 (15/30)
	상담	0.823	0.5 (15/30)
면접	교육	0.933	0.667 (20/30)
	지원금	0.733	1.00 (30/30)
	일자리	0.967	1.00 (30/30)
	상담	0.967	0.833 (25/30)
지원금	교육	0.792	0.667 (20/30)
	면접	0.733	1.00 (30/30)
	일자리	0.95	0.833 (25/30)
	상담	0.9	0.667 (20/30)
일자리	교육	0.95	0.5 (15/30)
	면접	0.967	1.00 (30/30)
	지원금	0.95	0.833 (25/30)
	상담	0.497	0.5 (15/30)
상담	교육	0.823	0.5 (15/30)
	면접	0.967	0.833 (25/30)
	지원금	0.9	0.667 (20/30)
	일자리	0.497	0.5 (15/30)

교육, 상담 등
특정 카테고리의 정확도가
매우 떨어지는 모습 보임.

02 데이터 수집 및 처리
5) 카테고리 라벨링

K-Modes

K-means clustering의 기본 구조를 유지하면서 범주형 데이터, 즉 명목변수로 이루어진 데이터에 적용이 가능한 방법.

사용 하이퍼 파라미터

KModes(n_clusters= 5, init = "Huang", n_init = 30, max_iter=1000, verbose=1, random_state = 343)

상담 데이터 2차 라벨링
K-Modes를 활용한 군집 분류

	cat_0_nouns	cnt_0	cat_1_nouns	cnt_1	cat_2_nouns	cnt_2	cat_3_nouns	cnt_3	cat_4_nouns	cnt_4
0	교육	165	고민	118	지원금	135	면접	304	일자리	111
1	국비	121	조언	85	돈	38	연습	80	기업	93
2	학원	74	상담	57	지급	29	정장	51	일	87
3	자격증	62	걱정	49	청년	25	준비	33	경력	83
4	강의	48	방향성	38	장려금	23	자기	33	직무	66
5	수업	46	도움	33	비용	22	소개서	33	인턴	63
6	카드	44	멘토링	31	신청	20	면접스킬	32	중소기업	60
7	수강	35	카페	29	창업	18	메이크업	28	돈	56
8	능력향상	32	컨설팅	29	준비	18	질문	28	직무경험	48
9	공부	32	준비	26	생활비	16	대여	27	도움	47
10	관련	28	취업박람회	26	금전	12	이력서	25	계약직	46
11	가능	20	일	25	졸업	12	구두	24	학원	46
12	추천	17	부탁	23	생각	11	방법	23	정규직	45
13	신청	17	생각	20	경제	11	모의	19	강의	45
14	발급	15	진로	16	방법	11	역량검사	19	준비	44
15	교육생	15	현직자	16	사업	10	헤어	18	일경험	44

02 데이터 수집 및 처리

5) 카테고리 라벨링

카테고리1

카테고리2

카테고리3

카테고리4

카테고리5

03 데이터 분석

3-1 TF-IDF

3-2 유사어

3-3 코사인 유사도

3-4 알고리즘

03 데이터 분석

1) idf 가중치

TF - IDF

특정 단어의 문서 내 중요도를 '가중치화' 해주기 위해 사용함

문서1: 빅데이터 교육 받고 싶어요
문서2: 인공지능 교육 받고 싶어요

	빅데이터	교육	인공지능
문서1	1	1	0
문서2	0	1	1

<tf-idf>

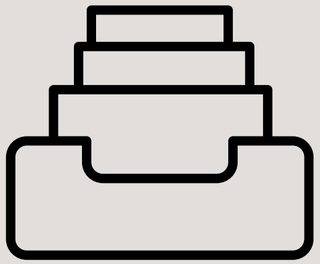
	빅데이터	교육	인공지능
문서1	1	0.5	0
문서2	0	0.5	1

특정 문서에서만 자주 등장하는 단어의 중요도가 높다!

정책 데이터

2글자 이상 단어 추출

단어 빈도(tf)



특정 문서에서 특정 단어가 나오는 횟수

역문서빈도(idf)

단어 가중치화

'지원', '청년' 등 빈도수가 높은 단어는 가중치가 낮고
빈도가 낮은 단어일수록 가중치 높음

03 데이터 분석

1) idf 가중치

TF 표준화 점수

	cat_0_nouns	cnt_0	cat_1_nouns	cnt_1	cat_2_nouns	cnt_2	cat_3_nouns	cnt_3	cat_4_nouns	cnt_4
0	교육	7.343065	고민	7.042506	지원금	9.694839	면접	10.193578	일자리	4.629128
1	국비	5.284184	조언	4.937317	돈	2.481135	연습	2.477206	기업	3.704628
2	학원	3.084925	상담	3.151096	지급	1.811822	정장	1.478212	일	3.396461
3	자격증	2.523412	걱정	2.640747	청년	1.514350	준비	0.858146	경력	3.191017
4	강의	1.868314	방향성	1.939017	장려금	1.365614	자기	0.858146	직무	2.317878
...	...	...	...	...	...	...	...	...	...	...
119	일자리	-0.377738	사진	-0.485140	연습	-0.344852	정규직	-0.278641	면접스킬	-0.969234
120	모의	-0.377738	빅데이터	-0.485140	연구	-0.344852	정도	-0.278641	면접	-1.071956
121	금전	-0.377738	부담	-0.485140	연계	-0.344852	디자인	-0.278641	지원금	-1.071956
122	정규직	-0.377738	발급	-0.485140	역량검사	-0.344852	생활	-0.278641	조언	-1.071956
123	기업	-0.377738	시작	-0.485140	희망	-0.344852	희망	-0.278641	모의	-1.071956

IDF 가중치 점수

낮은 빈도 단어	
간호인력	6.2146081
개발전문인력	6.2146081
공예기술	6.2146081
출판	5.809143
크리에이터	4.8283137

높은 빈도 단어	
특강	2.9759296
인력	1.7429693
취업	0.6991653
청년	0.3797974
지원	0.2119564

03 데이터 분석
2) 유사어

PROBLEM

'취업준비생인데 중소기업이라도 괜찮은 곳 있을까요? '

≠

'취준생인데 중소기업이라도 괜찮은 곳 있을까요? '

뜻은 같은 문장!

But '취업준비생' 과 '취준생' 차이로 다른 결과 도출

정책 → '취업준비생' 多 / 사용자 → '취준생' 多

SOLUTION

'유사어 목록 작성하여 사용자 단어사전에 추가

```
# 유사어 적용 전
print(mecab.nouns('취업준비생인데 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
print(mecab.nouns('취준생인데 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
print(mecab.nouns('취준중인데요. 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
```

```
['취업준비생', '일자리', '중소기업']
['취', '준', '일자리', '중소기업']
['취', '준중', '일자리', '중소기업']
```

```
# 유사어 적용 후
print(mecab.nouns('취업준비생인데 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
print(mecab.nouns('취준생인데 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
print(mecab.nouns('취준중인데요. 일자리가 급해서요. 괜찮은 중소기업이라도 있을까요?'))
```

```
['취업준비생', '일자리', '중소기업']
['취준생', '일자리', '중소기업']
['취준', '중', '일자리', '중소기업']
```

▶ 유사어끼리 같은 점수를 가지도록 설정

03 데이터 분석

3) 코사인 유사도

코사인 유사도 Cosine similarity

두 벡터의 코사인 값으로 유사도를 판별

문서1: 빅데이터 교육 받고 싶어요
문서2: 인공지능 교육 받고 싶어요

	빅데이터	인공지능	교육	받고	싶어요
문서1	1	0	1	1	1
문서2	0	1	1	1	1

두 문서의 코사인 유사도 = 0.75

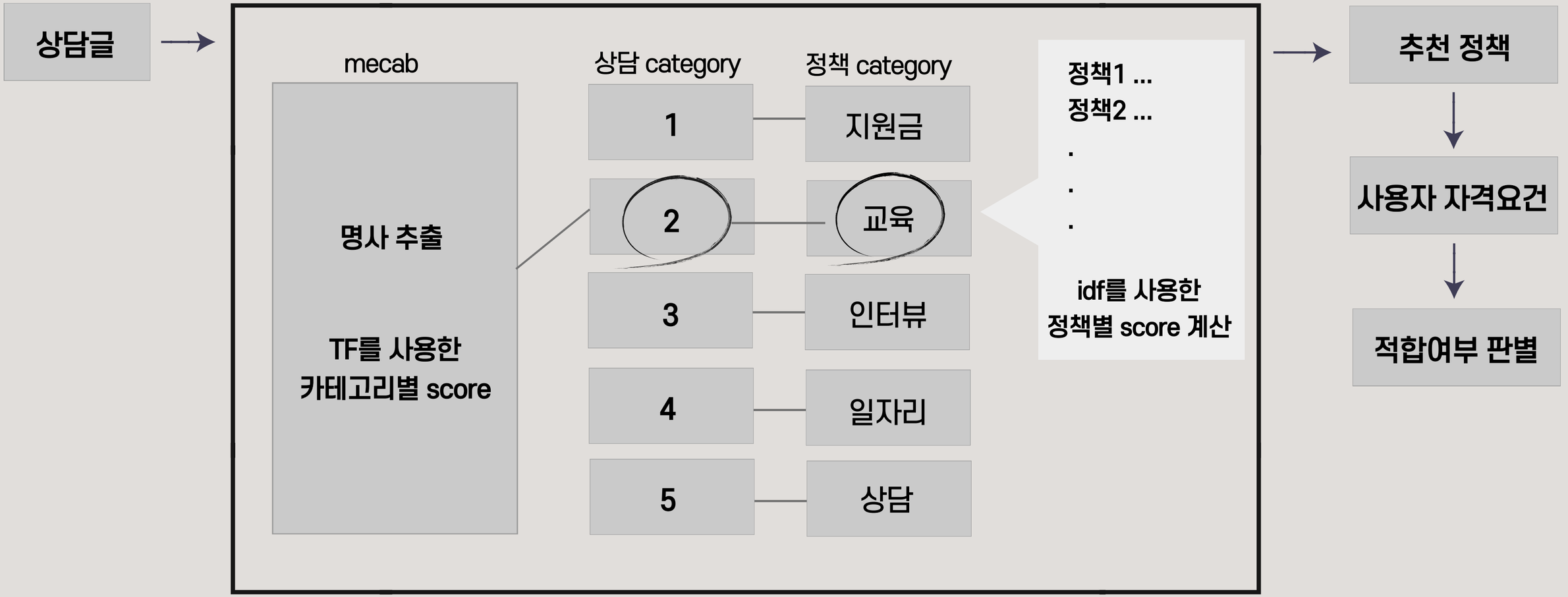
상담 category

정책 category	카테고리1	카테고리2	카테고리3	카테고리4	카테고리5
지원금	**0.18**	0.04	0	0.12	0
교육	0.05	**0.12**	0.04	0.06	0.04
인터뷰	0.06	0.02	**0.26**	0.06	0.09
일자리	0.06	0.08	0	**0.2**	0
상담	0.06	0.06	0.05	0.1	**0.21**

상담 카테고리 특성과 일치하는 정책 카테고리 유사도 높음 확인

03 데이터 분석

4) 알고리즘



04 서비스 구현

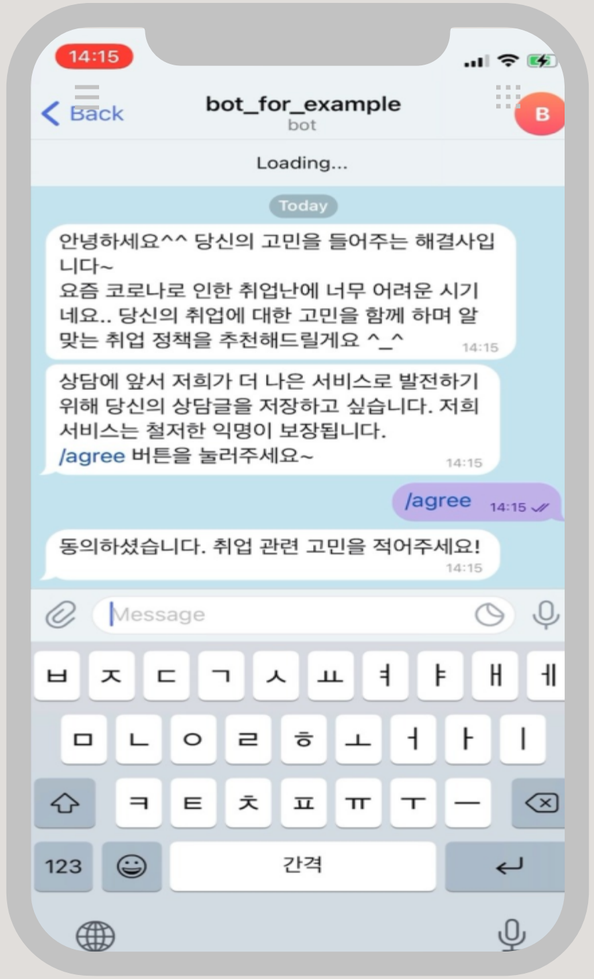
4-1 챗봇 시연

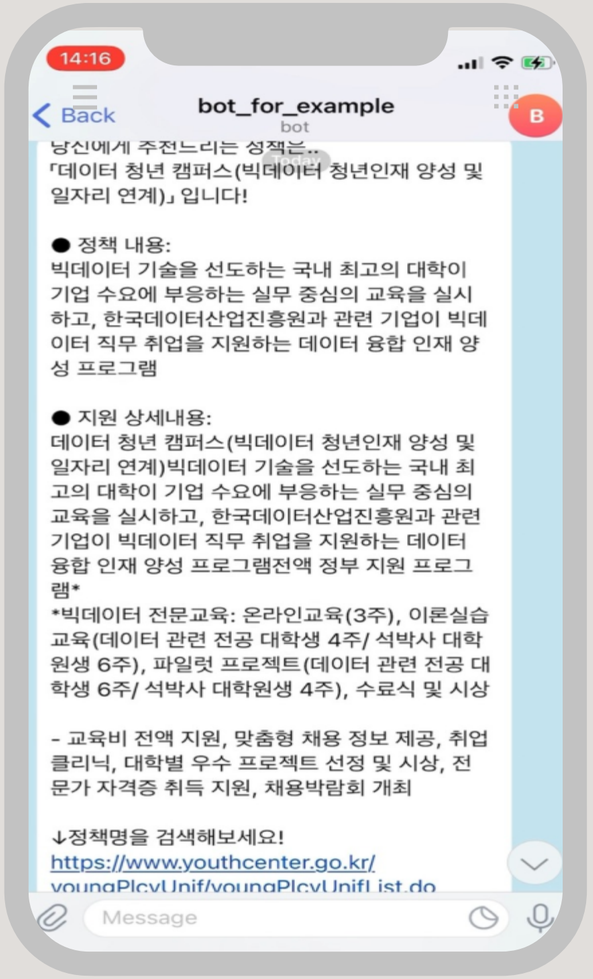
04 분석 결과

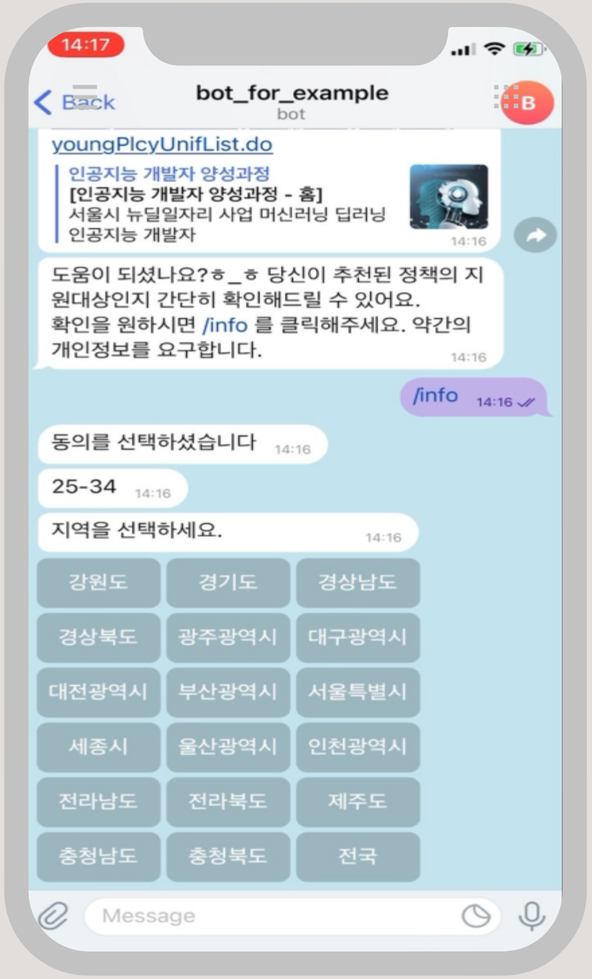
1) 챗봇 시연 (PC ver)

04 분석 결과

1) 챗봇 시연 (핸드폰 ver)







05

시사점 및 한계점

5-1 시사점 및 한계점

5-2 향후 개선사항

05 시사점 및 한계점

1) 시사점 및 한계점

시사점

1. 여러 사이트에 분산되어 있던 정책을 한번에 볼 수 있음
2. 사용자 맞춤형 숨은 정책 찾을 수 있음
3. 챗봇을 이용하여 사용자가 자신의 상황 주체적으로 설명 가능

한계점

1. 수동 불용어 처리
2. 수동 유사어 추가
3. 상담글 및 정책 데이터수 부족
4. 정책 범위 확장 실패(취업 -> 청년)
5. 객관적 평가지표 부재

05 시사점 및 한계점

2) 향후 개선사항

개선사항

1. 수집한 상담글 데이터를 활용하여 **상담글 군집별 특징 명확화**
2. 취업 정책 → 청년정책 **범위 확장**(주거, 교육, 복지, 문화, 참여)
3. **유사어, 불용어** 사전 확장 및 **자동 적용화**
4. 서비스 종료시 만족도를 수집하여 **평가지표** 활용

감사합니다!