채호연고재광박서윤이우영최민준

We Want Better Life. For Both Of Us.

BABAE

I. 서론

I. 서론

1년에,

100,000 마리

2015년 유기동물관리 비용

12,880,000,000 원

I. 서론

반려 동물을 왜 유기할까요?

이해의 부족

I. 서론

반려동물을 키우는 비용

한달 평균 13만원

반려동물과 함께 하며 경제적 부담을 느낀 적이

“있다”

I. 서론

하지만 이 마저도, 미용 및 사료와 같은 가장 기초적 지출.

의료 비용이 함께 청구된다면,

실제로 사람들의 사전 이해를 한참 웃도는 지출 발생.

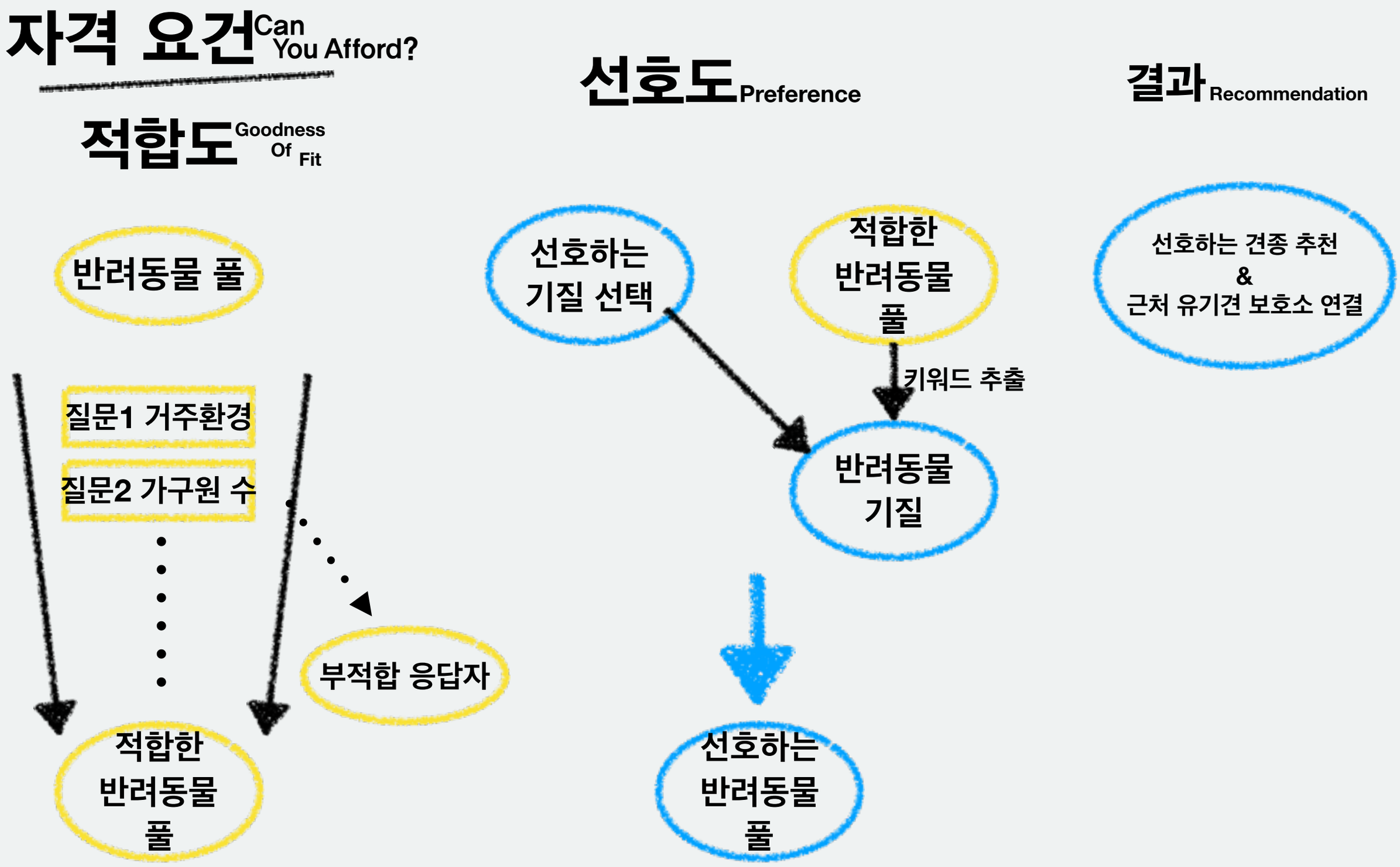
How Can We Make Both HAPPY?

I. 서론

Right Pet To Right Person,

“사용자의 상황을 고려해 가장 적합한 견 종을 추천해주자.”

I. 서론


자격 요건 Can You Afford?
적합도 Goodness Of Fit
선호도 Preference
결과 Recommendation
반려동물 풀
선호하는 기질 선택
적합한 반려동물 풀
선호하는 견종 추천 & 근처 유기견 보호소 연결
키워드 추출
반려동물 기질
질문1 거주환경
질문2 가구원 수
부적합 응답자
적합한 반려동물 풀
선호하는 반려동물 풀

II. 데이터 수집

II. 데이터 수집

AKC(American Kennel Club)

130년의 역사를 자랑하는 American Kennel Club, 세계에서 가장 큰 애견협회의 웹사이트

II. 데이터 수집

크롬 브라우저 상에서 요소검사를 통해
특정 HTML 값을 추출하는 방식으로 크롤링.

II. 데이터 수집

크롤링 스크립트 플랫폼 : Python

사용된 라이브러리 : Beautifulsoup, csv, requests.

II. 데이터 수집

크롤링 스크립트 플랫폼 : Python

사용된 라이브러리 : Beautifulsoup, csv, requests.

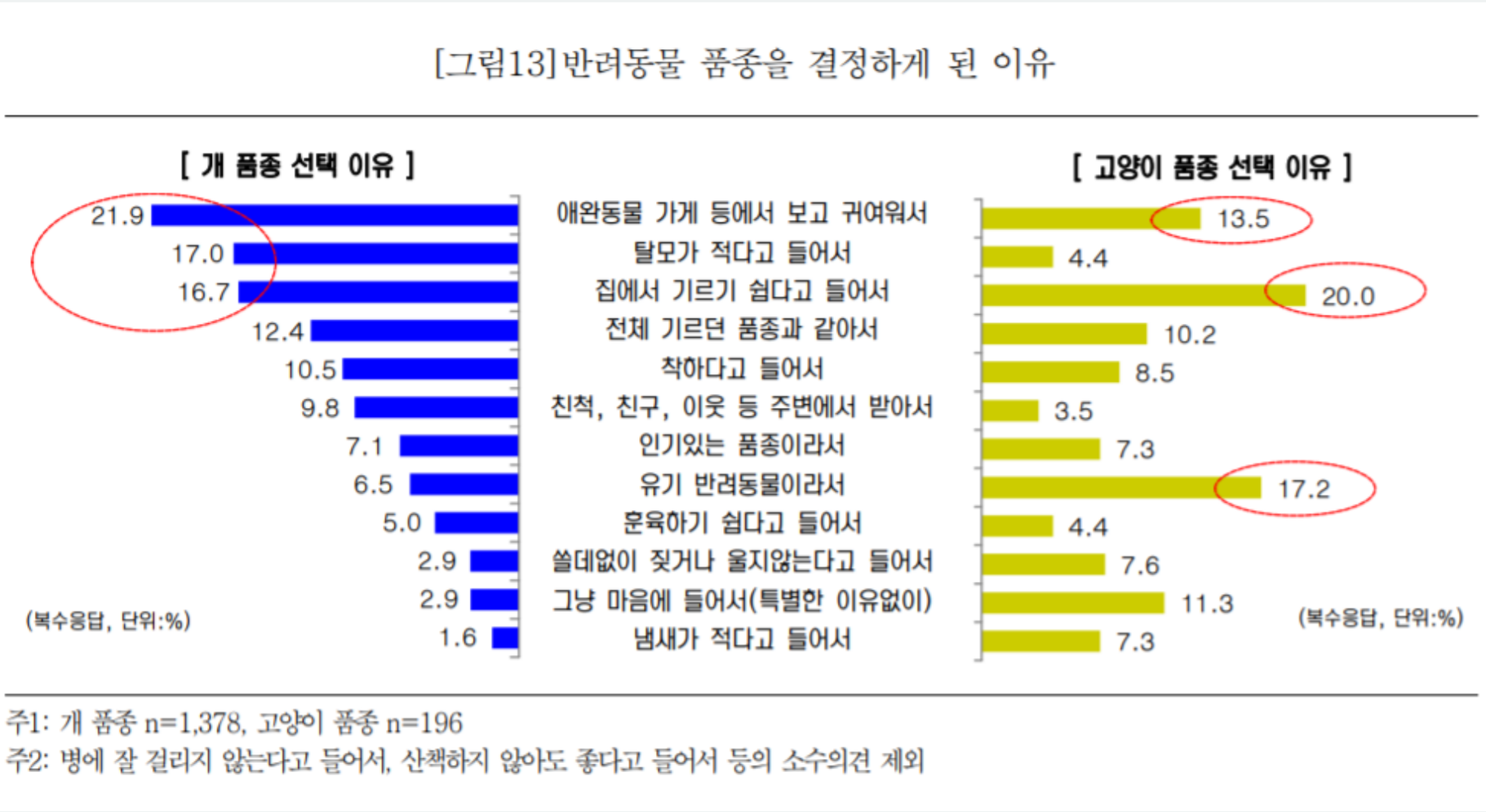
II. 데이터 수집

견 종 별 이름, 특성/성격 키워드 크롤링 후,

Csv 칼럼에 맞추어 가공 후 csv파일의 형태로 저장.

III. 1차 프로세스 설명

III. 1차 프로세스 설명



유기사유: 이해의 부족 (본인&반려동물 둘 다)

⇒ 1) 주인 본인의 능력/상태/환경 (=1차 자격요건절차) 분석

\+ 2) 강아지의 능력/상태/성향 (=2차 선호도) 분석

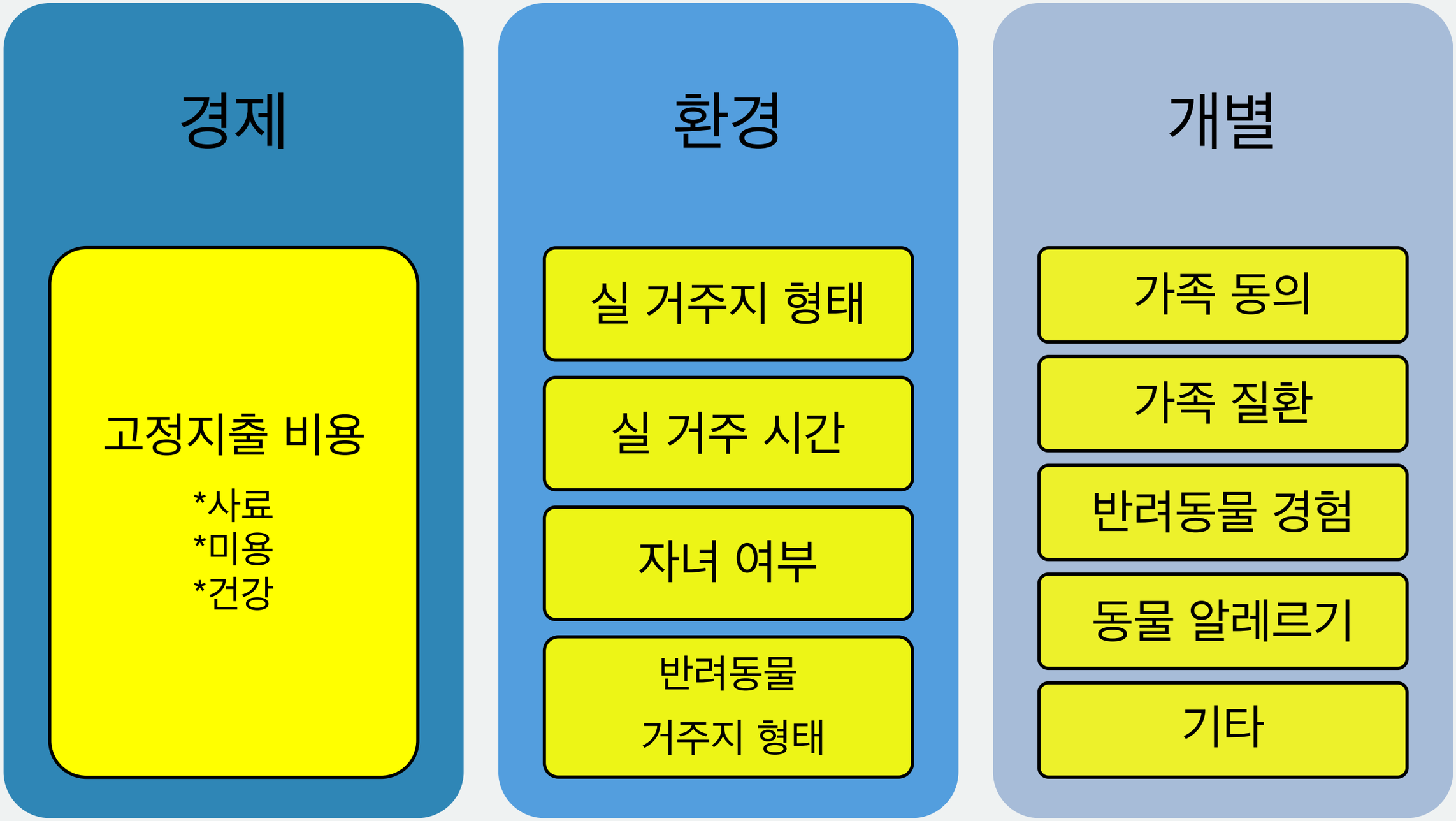
III. 1차 프로세스 설명

"관계를 유지하기 위한 비용이 높은 경우는 반려동물의 유기로 이어질 수 있다"

⇒ 경제적인 역량이 관계유지의 주요한 기준

*이를 중점으로 총 3가지 자격판별절차 구성

III. 1차 프로세스 설명



총 10가지 질문의 답안은 각각 다지선다와 O/X로 구성되어 있다.

III. 1차 프로세스 설명

경제적 요건 질문

1. 반려 동물 생활비

반려동물 생활에 고정적으로 지출 가능한 한달 비용은?

1)5만원 이하 2) 5~13만원 3) 13만원~18만원 4)18만원 이상

III. 1차 프로세스 설명

환경적 요건 질문

2. 3인이상의 가족이 실 평수 10평이하의 공간에 거주하십니까? (O/X)

(2번 질문 'X'의 경우)

1) 단독주택 2) 다가구주택(오피스텔/아파트 등) 3) 기타

3. 강아지가 혼자 있어야 하는 시간이 8시간 이상이에요. (O/X)

4. 3세 미만의 자녀가 있다 (O/X)

5. 현재 거주 중이신 곳이 다음 중 하나에 해당하시나요?

: 공장/회사/군부대 등 사람들의 이동이 많은 곳(환경)(O/X)

: 농장과 식당, 사무실 등 영업장(환경)(O/X)

: 양로원,고아원과 같은 복지 시설(환경)(O/X)

: 위 어느 선지에도 해당하지 않음(O/X)

III. 1차 프로세스 설명

개별적 요건 질문

6. 반려동물과 함께하는 것에 대한 가족 간 의견 합치 여부/
미성년자의 경우 동의현황 (O/X)

7. 동거인 중 우울증 등 정신 질환이 있는 경우(환경)(O/X)

8. 반려 동물을 키우다 중간에 포기한 경우가 2번 이상인가요? (O/X)

9. “알러지 약을 복용할 경우 증상을 완화할 수는 있습니다.
하지만 장기적으로 복용하며 함께 지내는 것은 또 다른 문제입니다.
이에 대해 충분히 고려해보셨나요? (O/X)”
(e.g 재채기, 기침, 콧물, 코 막힘, 눈 가려움증, 충혈, 피부 발진,
두드러기, 호흡곤란, 가슴, 답답함, 천명 (호흡 시 쌕쌕거림) 등)

10. 외국으로 입양을 원하거나 한국에 거주하는 외국인인가요? (O/X)

IV. 2차 프로세스 설명

IV. 2차 프로세스 설명

위의 적합도 검사를 통과한 사람들에 한해 선호도 검사를 실시

이 선호도 검사는 다음의 전제에서 시작한다

“삶을 함께 살아가는 파트너들이 더 많은 성격의 유사성을 갖고 있다면
더 오래 결혼 생활을 유지한다.”

Anthony C. Little, D. Michael Burt, David I. Perret

IV. 2차 프로세스 설명


클러스터링
성격 POOL 군집화
선호도조사
최종 추천
견종들의 성격을 클러스터링 방법 구체적 서술
견종의 성격을 기준으로 군집을 만들어 추천 POOL 완성
성격
외형
털 손질 빈도
털빠짐 정도
활동량
총명함
훈련 시 반려견의 태도
최종 견종 추천

IV. 2차 프로세스 설명

클러스터링

어떻게 텍스트의 의미를 반영한 클러스터링을 할 수 있을까?

IV. 2차 프로세스 설명

클러스터링

첫 번 째 방안,

유클리디안 거리에 근거해 유사도 측정 후 이를 근거로 클러스터링 실시.

IV. 2차 프로세스 설명

클러스터링

한계

1. 성격 키워드가 61차원의 데이터였기에 기하적인 거리를 측정하는 것은 가능하였으나, 지나치게 차원이 많았다.

2. 의미를 반영하기 위해 사전에 우리가 직접 의미 기준으로 5가지의 군집으로 나누어 본 후, 이에 근거하여 가중치를 부여하는 방식으로 의미 반영을 시도해봄.

IV. 2차 프로세스 설명

클러스터링

하지만 가중치를 부여하는 방식 마저도,

우리가 개별적으로 부여할 수 있는 수의 데이터였기에 가능한 것이었으며,

만약 데이터의 수가 많아지거나, 차원의 수가 증가하였을 경우에도

일관되게 사용할 수 없는 방식이었으므로

신뢰성이 없다고 판단, 기각하였음.

IV. 2차 프로세스 설명

클러스터링

하지만 가중치를 부여하는 방식 마저도,

우리가 개별적으로 부여할 수 있는 수의 데이터였기에 가능한 것이었으며,

만약 데이터의 수가 많아지거나, 차원의 수가 증가하였을 경우에도

일관되게 사용할 수 없는 방식이었으므로

신뢰성이 없다고 판단, 기각하였음.

IV. 2차 프로세스 설명

클러스터링

Robust Semantic Similarity Measuring 알고리즘에
근거한 클러스터링

IV. 2차 프로세스 설명

클러스터링

다음과 같은 두 가지 근거로 단어 간 유사도를 측정.

1. 해당 단어를 AND연산자로 묶어 검색엔진에 검색한 후
   결과로 출력되는 페이지의 수.
2. 검색 결과로 출력되는 텍스트 스니펫(Snippets) 내의
   비교 대상 단어 간 관계를 정의하는
   어휘 구문론적 유사 패턴(Lexico-Syntactical Patterns)

IV. 2차 프로세스 설명

클러스터링

1. 해당 단어를 AND연산자로 묶어 검색엔진에 검색한 후 결과로 출력되는 페이지의 수.

: “AND 연산자를 이용해 두 단어 간 유사관계가 가능할 수록 검색 결과로 출력되는 페이지의 수가 그렇지 않은 경우보다 많다.”

IV. 2차 프로세스 설명

클러스터링

실제로,

Apple/Computer를 함께 검색한 결과가 Apple/Banana 보다

“80배”많은 288,000,000개 였다고 한다.

해당 가설에 대한 Correlation Test : 0.834

F-Measures : 0.78

따라서,

H0 :“두 단어의 유사관계가 높을 수록,

AND연산자로 함께 검색한 결과로 출력되는

페이지의 수가 그렇지 않은 경우보다 많다.”

귀무가설 H0이 성립한다고 할 수 있음.

IV. 2차 프로세스 설명

클러스터링

하지만 이것만으로는 신뢰성이 충분한 결론을 짓기 힘들다.

따라서, 두 번째, 텍스트 스니펫과 말뭉치를 활용한 방법을 추가하여

보다 정교하게 의미 유사도를 측정한다.

IV. 2차 프로세스 설명

클러스터링

두 번 째,

검색 결과로 출력되는 텍스트 스니펫(Snippets) 내의

비교 대상 단어 간 관계를 정의하는

어휘 구문론적 유사 패턴(Lexico-Syntactical Patterns)

IV. 2차 프로세스 설명

클러스터링

Snippets : 토막

검색엔진에서 특정 단어의 검색 결과 하단에,

해당 페이지의 내용이 요약되어 출력되어있는 본문의 토막.

IV. 2차 프로세스 설명

클러스터링

이는 이미 검색엔진에 의해

본문을 가장 잘 대표하는 것으로 판단된 텍스트가 출력된 것이므로,

스니펫이 본문을 잘 대변한다는 것은 검색엔진 자체의 신뢰성으로 설명이 가능하다.

IV. 2차 프로세스 설명

클러스터링

어휘 구문론적 패턴(Lexico-Syntactics Patterns)

“단어를 포함하는 구문의 패턴을 파악함으로써

두 단어의 관계를 유추하는 것이 가능하다.”

IV. 2차 프로세스 설명

클러스터링

(Cricket, Sport)

“Cricket is a sport played between two teams, each with eleven players.”

‘is a” 와 같은 구문이

두 단어를 같은 성질의 것임을 알려주고 있음.

이것이 바로

어휘 구문론적 패턴

IV. 2차 프로세스 설명

클러스터링

(Cricket, Sport)

“Cricket is a sport played between two teams, each with eleven players.”

‘is a” 와 같은 구문이

두 단어를 같은 성질의 것임을 알려주고 있음.

이것이 바로

어휘 구문론적 패턴

IV. 2차 프로세스 설명

클러스터링

WordNet과 같이 말뭉치에 사전에 정의된

1)유사어로 묶인 단어 짝(Pair)

2) 그렇지 않은 단어

두 가지에 대해 검색 결과로 출력된 스니펫으로부터

어휘 구문론적 패턴을 추출한 후 각각의 패턴들을

유사어를 나타내는 패턴/그렇지 않은 패턴으로 구분하여 상정.

IV. 2차 프로세스 설명

클러스터링

그리고 유사도를 측정하고 싶은 단어들을 검색한 결과로 출력되는

스니펫들로부터 어휘 구문론적 패턴을 추출한 후 ,

유사한 어휘의 구문론적 패턴과/그렇지 않은 패턴 중

어으 것이 더 많이 관측되는 지에 근거해

두 단어의 관계를 유사어/그렇지 않은 단어로 판단 하는 알고리즘.

IV. 2차 프로세스 설명

클러스터링

WordNet과 같이 말뭉치에 사전에 정의된

1)유사어로 묶인 단어 짝(Pair)

2) 그렇지 않은 단어

두 가지에 대해 검색 결과로 출력된 스니펫으로부터

어휘 구문론적 패턴을 추출한 후 각각의 패턴들을

유사어를 나타내는 패턴/그렇지 않은 패턴으로 구분하여 상정.

IV. 2차 프로세스 설명

클러스터링

두 가지 관점을 통해

유사도를 수치 데이터로 도출한 후

이 두 가지를 통합해 벡터데이터로 변환.

벡터데이터화 되었으므로 k-Means와 같은

일반적인 클러스터링 알고리즘이 사용가능하다.

IV. 2차 프로세스 설명

성격 POOL 군집화

군집화된 결과치 제시

예) 알고리즘을 돌려본 결과 이런 POOL이 나온다

IV. 2차 프로세스 설명

선호도조사

1. 성격
2. 견종크기
3. 털 손질 빈도
4. 털빠짐 정도
5. 활동량
6. 총명함
7. 훈련 시 반려견의 태도

최종 추천

5점 척도를 사용
개인마다 최종 점수를 도출(가중치 부여)

성격으로 선택된 POOL 안에서
예비 보호자와 가장 잘 맞는 견종들 추천

V. 결론

V. 결론

견 종 별로 공통적으로 갖고있는 속성이지만 구분되는 특징들을 크롤링한 후 다각도에서 접근.

V. 결론

견 종 별 성격 데이터는 텍스트 데이터,

이를 분석한 결과 중복이 관측되었다.

따라서 견 종 별 성격을 정의한 키워드들의

특정 성격에 대해 일정한 정의를 공유하고 있음을 확인.

V. 결론

특정 성격에 대한 일정한 정의를 공유하고 있다는 것은,

성격 키워드를 통해 유의미한 군집으로 군집화하는 것이 가능하다는 것.

V. 결론

Robust Semantic Similarity Measuring 알고리즘을 발견하였으나,

이를 실제로 적용하기엔 우리에게 기술적 한계가 있었음.

V. 결론

그럼에도 불구하고,

1. 성격 키워드로 출력된 데이터들로부터 중복을 관측하고,
2. 이로부터 클러스터링할 수 있는 성격의 데이터임을 발견하였으며,
3. 현재에는 힘들지만 조금 시간이 더 주어진다면 충분히 가능할 만큼 이해를 하였다고 할 수 있음.

V. 결론

이로부터 프로젝트의 실현가능성을 보았으므로,
충분한 인사이트를 도출하였다고 조심스럽게 말할 수 있다.

VI. 역할 분담 및 프로젝트 일정

VI. 역할 분담 및 프로젝트 일정

이름(가나다 순)	역할
고재광	자격 요건 판별 프로세스 전담 기술 및 체계화, 프로세스 전반적 논리 체계화, 선호도 알고리즘 체계화 참여, 자격 요건 판별 프로세스 파트 보고서 기술 및 PPT 작성
박서윤	데이터 수집(스크래핑) 및 전처리, 자격 요건 판별 알고리즘 구현, 선호도 알고리즘 고안, 선호도 알고리즘 구현, 최종 보고서 문서 형식 만들기 및 문서 마무리 작업 등
이우영	데이터 전처리, 자격요건, 선호도 문항 구조 및 질문 작성, 최종보고서와 PPT 선호도 파트 작성, 팀플 장소 예약
채호연(팀장)	데이터 전처리, 전체적인 일정/테스크 조율 및 할당, Robust Semantic Similarity Measuring Algorithm 분석 및 관련한 기술/서술작업 일체, 보고서 취합
최민준	프로젝트 연구 목적/배경/기대효과/프로세스 소개 기술 위한 구조 체계화, PPT 취합, 보고서 취합, 자격 요건 판별 알고리즘 체계화, 프로세스 전반적 논리 체계화, 서론 PPT/보고서 작성

역할 분담.

VI. 역할 분담 및 프로젝트 일정

	작업 시작	기간	작업 종료
브레인스토밍	2018.10.10	7일	2018.10.17
주제선정	2018.10.17	1일	2018.10.17
문헌 및 통계 자료 조사	2018.10.21	9일	2018.10.30
중간발표	2018.10.31	1일	2018.10.31
데이터 수집	2018.10.31	20일	2018.11.20
데이터 전처리 및 분석	2018.11.20	20일	2018.12.10
문서화	2018.12.21	4	2018.12.24
최종발표	2018.12.24	2일	2018.12.26

프로젝트 일정