**2018년 2학기 데이터마이닝 term project**

**Best Adoption, Best for Each Other**

**: 견종 별 성격 및 특성을 고려한**

**맞춤형 반려동물 매칭 서비스**



201403303 채호연 201100126 고재광

201401186 박서윤 201302510 이우영

201400000 최민준

목차

Ⅰ. 서론

1.1. 문제 의식 ---------------------------------------------------- 3

1.2. 연구 배경 및 목적 -------------------------------------------- 4

1.3. 서비스 제안 -------------------------------------------------- 5

1.4. 기대 효과 및 전제 도출 --------------------------------------- 6

Ⅱ. 데이터 수집

2.1. 데이터 출처 소개 --------------------------------------------- 10

2.2. 데이터 수집 방법 및 이슈 ------------------------------------- 12

2.3. 데이터 전처리 및 이슈 ---------------------------------------- 13

Ⅲ. 서비스 설계

3.1. 1단계 - 자격 요건 판별 프로세스 ------------------------------ 14

3.1.1. 자격 요건 판별 프로세스 설명 --------------------------- 14

3.1.2. 자격 요건 판별 프로세스 알고리즘 및 코드 --------------- 19

3.2. 2단계 - 선호도에 근거한 필터링 프로세스 ---------------------- 20

3.2.1. 견종 성격 선호 테스트 ---------------------------------- 26

3.2.2. 견종 특징 선호 테스트 ---------------------------------- 29

3.3. 3단계 - 결과 출력 -------------------------------------------- 29

Ⅳ. 결론

4.1. 분석 결과 요약 및 인사이트 발굴 ------------------------------- 31

4.2. Lessons Learned ----------------------------------------------- 32

Ⅴ. 프로젝트 관리

5.1. 팀원의 구성 및 역할 ------------------------------------------- 32

5.2. 프로젝트 일정 ------------------------------------------------- 33

5.3. 프로젝트 도중 발생한 이슈 ------------------------------------- 33

5.4. 첨부파일 ------------------------------------------------------ 34

Ⅵ. 참고자료 ----------------------------------------------------------- 35

Ⅰ. 서론

1.1. 문제 의식

2017년 한국농촌경제연구원에 따르면 반려견 및 반려묘의 수가 약 900만 마리에 육박하는 것으로 나타났다.[[1]](#footnote-1) 반려동물을 키우는 가구의 증가는 혼자 사는 가구가 늘고, 고령화가 급격히 진행되면서 동물을 인생의 ‘반려자’로 여기는 분위기가 생겨나고 있기 때문이다. 그러나 반려동물 입양이 늘어나는 만큼 버려지는 유기동물 또한 증가하고 있다. 2017년 농림축산식품부에 따르면 한 해동안 구조된 유실, 유기동물이 10만2593마리로 2015년 8만2천여마리, 2016년 8만9천여마리에 이어 지속적으로 유기되는 동물이 증가하고 있다.[[2]](#footnote-2) 즉, 한 해 동안 수없이 많은 반려동물들이 버려지고 있는 것이다. 이렇게 버려진 동물들은 현재 대한민국 사회에서 큰 문제가 되고있다. 유기동물 관리와 처리에 드는 비용이 만만치 않기 때문이다. 농림축산검역본부에 따르면 유기된 동물을 관리하기 위해 연간 128억원의 비용이 발생하고, 유기동물이 점차 증가함에 따라 이러한 비용 역시 늘어날 것으로 예측된다.[[3]](#footnote-3)

유기동물의 증가로 인한 사회문제를 해결하기 위해서는 왜 사람들이 반려동물을 버리는지 이해할 필요가 있다. 유기동물이 발생하는 이유는 크게 두 가지로 분류할 수 있다. 바로 자격 요건과 선호의 이해 부족이다.

첫째로 자격 요건의 이해 부족에 대해 간략히 설명하면, 앞으로 반려동물을 기르게 될 경제적 여건 및 환경적 여건 등에 대한 이해가 부족하다는 의미이다. 반려동물을 유기하는 가장 큰 이유 중 하나는 바로 경제적 부담이다. 본인이 반려동물을 키울 수 있는 경제적 여건이 된다고 판단하여 반려동물을 입양받지만 실제로 반려동물에게 드는 비용이 많아 유기를 하는 것이다. 이베이코리아에 따르면 반려동물을 키우는데 한달 평균 13만3000원의 비용이 들고 반려동물을 키우는 2명 중 1명은 반려동물을 키우면서 경제적 부담을 느낀 적이 있다고 밝혔다.[[4]](#footnote-4) 위 통계는 간식, 미용용품등 반려동물을 위한 용품만을 고려한 값이라는 것을 고려했을 경우이며, 의료비 등 추가적인 비용을 생각하면 반려동물을 키우는데 드는 한달 평균 비용은 더 많을 것으로 예상된다. 즉, 이러한 경제적인 부담을 고려하지 않거나 과소평가하여 반려동물을 분양받고 이를 감당하지 못하고 반려동물을 유기하는 현상이 발생하는 것이다. 경제적인 여건 외에도 자신이 살고 있는 환경, 여건 등을 고려하지 않고 분양 받는 것, 다시 말해 자신이 반려동물을 키우는데 적합한지에 대한 충분한 고려없이 분양을 받게 되어 유기동물이 증가하고 있다.

둘째, 선호에 대한 이해 부족은 자신이 어떠한 성향의 반려동물을 좋아하고 어떠한 성향의 반려동물을 싫어하는지에 대한 이해없이 분양을 받는 경우를 의미한다. 자신이 반려견의 털 빠짐을 얼마나 용인할 수 있는지, 어떤 성격의 반려견을 선호하는지에 대한 이해 없이 분양한 이후, 자신과 맞지 않을 경우 반려동물을 유기하는 일이 빈번하게 발생한다.

1.2. 연구 배경 및 목적

결국 유기동물의 발생을 줄이기 위해서는 분양자가 자신이 반려동물을 키우는데 적합한지에 대한 판단과 어떠한 종류의 반려동물을 선호하는지에 대해 이해할 필요가 있다. 이러한 문제를 해결하기 위해 현재 다양한 분양 서비스와 유기견 매칭 서비스가 진행되고 있지만 한계가 존재한다. 북미에서 유명한 ‘바크 버디’의 경우 강아지와 고양이의 이미지를 출력하여 사용자가 좋아하는 강아지, 고양이의 이미지를 선택하도록 하고 이를 기반으로 유저의 위치와 가장 가까운 유기동물 보호소의 동물을 추천해주는 서비스를 진행한다. 그러나 바크 버디는 선호 기반 중 외형이라는 측면만 고려하여 강아지와 고양이의 다른 특성을 제시하지 않고, 이용자가 반려동물을 키우는데 적합한지에 대해 고려하지 않는다는 한계를 가진다.

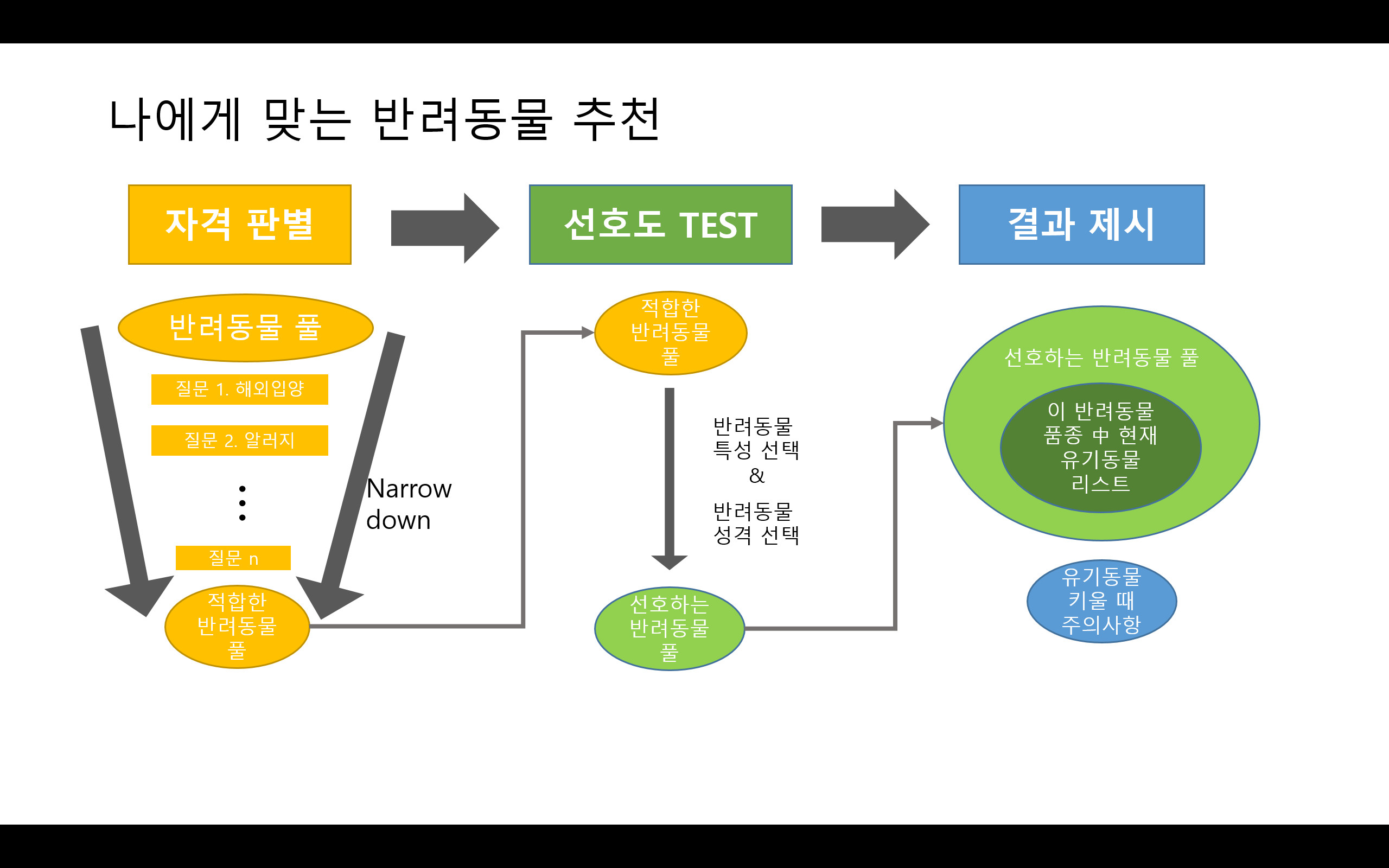
또한 국내의 경우, 한국일보에서 제공하는 ‘나에게 맞는 강아지 찾기’서비스가 있다. 이 서비스의 경우 바크 버디와 달리 강아지와 고양이의 이미지가 아닌 자신이 선호하는 강아지, 고양이의 특성을 선택할 수 있다. 그러나 ‘나에게 맞는 강아지 찾기’ 서비스 역시 개인의 선호만을 고려할 뿐 강아지를 키우는데 적합한지에 대한 질문이 부족하다.

위와 같은 서비스는 결국 이용자가 반려동물을 키우는데 있어 ‘적합한’ 사람인지에 대한 고려를 하지 않아 반려동물을 유기할 가능성을 만든다. 즉, 경제적 여건을 비롯하여 이용자의 환경과 여건을 고려할 필요가 있는 것이다.

이에 따라 우리는 “본인이 선호하는 반려동물을 선택할 때 반려동물과 오래 지낼 수 있다”는 사실에 따라 반려동물에 대한 개인의 선호도를 고려함과 동시에 이용자가 반려동물을 키우는데 있어 경제적으로, 환경적으로 그 외의 여건상 적합한지 또한 파악하여 이에 기반한 강아지 추천 서비스를 제공하고자 한다. 이를 통해 우리는 이용자와 '실질적으로 함께 잘 살 수 있는’ 반려동물을 추천해주고 결론적으로 우리 사회의 문제로 대두되고 있는 유기견 문제를 해소하고자 한다.

1.3. 서비스 제안

이 문서에서는 앞서 언급한 “자격요건”을 판별한 이후, 자격이 있는 사람을 대상으로 “선호도”를 테스트한 후 결과값으로 도출된 견종들을 사용자에게 제공하는 서비스를 제안할 것이다.



< 그림 1. 반려동물 추천 개괄적 프로세스 >

본 서비스에서는 <그림1>과 같은 프로세스로 이용자에게 가장 알맞은 강아지를 추천한다. 이에 따라 우리는 ‘자신과 성격이 비슷한 동물과 오래산다’는 가정을 가지고 응답자와 가장 비슷한 기질의 견종을 추천함으로써 인간과 반려동물이 함께 오래 살 수 있는 환경을 만들고자 한다.

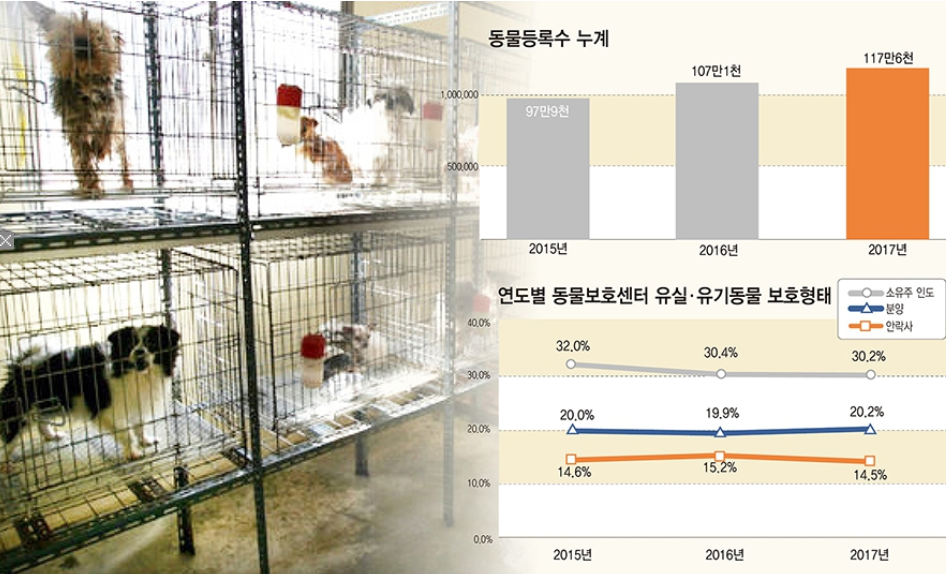
우선 이용자가 반려동물(강아지)을 키우는데 적합한지 파악하는 것부터 프로세스가 시작된다. 이 과정을 통해 우선 이용자가 반려동물을 키울 수 있는 환경인지를 파악하고자 한다. 즉, 응답자가 반려동물을 키울 수 없는 환경의 경우 더 이상 질문을 진행하지 않는 것이다. 이 과정을 통해 우선 응답자가 반려동물을 키우는데 적합한 환경인지 확인을 하고 응답자의 환경에 맞는 반려동물의 풀을 줄일 수 있다. 예를 들어 단독주택에 사는 응답자의 경우 소형견, 중형견, 대형견을 모두 추천 풀에 들어가는 반면 다가구주택에 사는 응답자의 경우에는 대형견을 제외한 소형견, 중형견을 추천해주고 활동량이 많지않은 견종으로 추천 풀을 줄이는 것이다.

이후 응답자의 반려동물 선호도를 파악하는 프로세스로 넘어가게 된다. 이 과정에서 응답자는 적합도 테스트를 통해 나온 적합한 반려동물 풀 내의 반려동물의 성격과 각종 특성을 본인의 선호에 따라 선택을 하게 된다. 특히 선호도 테스트 과정에서는 강아지의 기질과 함께 응답자와 가장 비슷한 모습의 기질을 선택하라는 방향을 제시하여 가장 적합하고, 응답자가 가장 선호하는 견종을 추천하고자 한다.

마지막으로 응답자가 가장 선호하고, 적합한 동물을 추천해줌과 동시에 응답자 주변 유기견 센터에 해당 견종이 존재하는지를 확인하여 응답자에게 유기견을 우선적으로 추천하고자한다.

1.4. 기대 효과 및 전제 도출

반려동물 추천 프로세스를 통해 우리는 크게 두 가지 기대효과가 발생할 것으로 예상한다. 첫째, 앞서 언급한 것처럼 유기동물로 인해 한 해 사용되는 예산이 100억원을 넘고있다. 우리가 제시하는 서비스는 이용자에게 유기견을 우선적으로 추천하기 때문에 결론적으로 현재 유기된 동물을 줄임으로써 유기동물 관리에 필요한 비용을 낮출 수 있다. 둘째, 주인에게 꼭 맞는 반려동물을 매칭함으로써 유기율 자체를 낮출 수 있다. 이용자의 경제적, 환경적 상황을 고려하여 적합하면서도 이용자의 취향과 성격을 반영하여 반려동물을 매칭함으로써 경제적 이유, 환경적 이유, 선호도의 문제로 인해 발생하는 유기동물을 줄일 수 있기 때문이다.



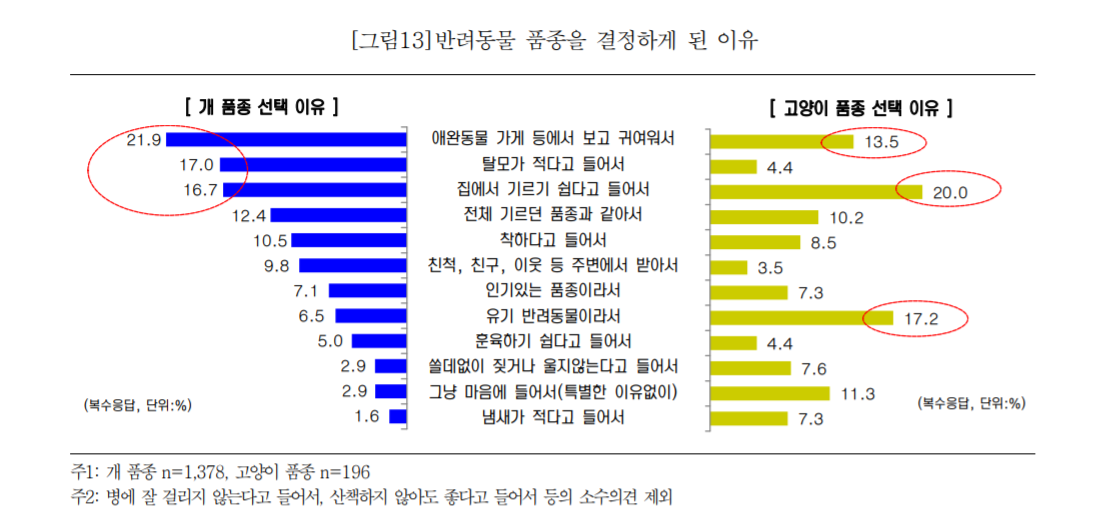
<그림 2. 2017년 동물보호와 복지 관리 실태 조사 결과, 농림축산식품부>

작년 한국에서 한 해 발생한 유기동물은 그 수가 10만 마리에 달했다. 이미 자국민에 의한 유기만해도 수없이 발생하는 상황에서, 최근에는 외국인 유학생·노동자 등 단기체류자에 의한 반려동물 유기가 중요한 사회적 문제로 대두되고 있다. 문제의 원인을 분석하지 않는다면 단기체류자의 수만큼이나 유기되는 반려동물의 수도 점차 증가하게 될 것이다.

“도대체 사람들은 왜 반려동물을 유기하는 것일까?”

처음 이 문제에 대해 팀 내에서 다양한 자료를 조사하던 중에 난항을 겪었는데. 사람에 대한 입양이나 파양, 또는 영아 유기에 대한 국내외 자료는 많았지만 반려동물 유기 사유에 대해 정성적인 방법을 통한 추측을 제외한 자료를 찾기란 쉽지 않았기 때문이다. 정성적 자료도 물론 적극적으로 참고 했지만 조금 더 정량적이고 실증적인 조사로 이루어진 자료를 찾던 중 네덜란드의 와게니겐 대학교 연구소(Wageningen University and Research) 에서 해당 주제와 관련된 행동생태학적 연구논문을 발견했다. 해당 논문의 저자들은 입을 모아 "반려견 주인의 만족도는 주인과 반려견 양측 모두의 삶의 질과 높은 연관성이 있으며 견주의 만족도가 심히 침해될 때 유기로 이어질 확률이 있다. (Dog ownership satisfaction relates to the quality of life of both owner and dog, and when seriously compromised may even lead to dog abandonment.)"다며 “결론적으로 첫째, 견주가 원치 않는 반려견의 행동양식과, 둘째, 관계를 유지하기 위한 비용이 높은 경우는 낙관적인 관계에 대한 조기 경보 신호로서 반려동물의 유기로 이어질 수 있다"(Consequently, unwanted behaviour and high perceived costs of ownership have potential as early warning signals of a less optimal relationship that could, in the end, result in dog abandonment.) 고 주장한다.

위 논문의 연구결과를 바탕으로 팀 BABE는 한 가지 자료를 추가로 검토해보았다. 그 두 번째 자료는 국내외로 활발하게 활동중인 동물권 운동단체 ‘동물권행동 카라(Korea Animal Rights Advocates)의 반려동물 유기사유이다. 카라는 “한국에서는 누구나 반려동물을 쉽게 구매할 수 있습니다. … 외국인이든 한국인이든, 혹은 다른 어떤 조건을 가지고 있든, 한국에서 사람이 가진 조건은 ‘반려동물의 구매자격’에 아무런 영향을 미치지 않습니다. 동물의 ‘가격’을 지급할 수 있는 사람이라면 지역 곳곳에 있는 펫숍에서 누구나 쉽게 동물을 구매할 수 있습니다.”라고 외치며 한국의 반려동물 유기 사유로 “손쉬운 반려동물 구매”를 꼽는다. 사람들이 반려동물을 물건처럼 쉽게 사갔기 때문에 마음에 들지 않을 때 물건을 버리듯 쉽게 버린다는 주장이다. 이는 다양한 자료를 통해 일관된 유기 사유가 ‘부담증가’임을 고려해봤을 때 구매 당시에 예상 부담에 대해 충분히 생각해보지 않고 쉽게 구매한 사람들이 부담이 실제로 닥쳤을 때 구매만큼 쉽게 유기한다는 주장이기에 설득력이 있는 주장이었다. KB연구보고서의 2017년도 반려동물 양육 실태 조사를 보면 실제로 반려동물을 선정하는데 있어서 개의 경우, 흔히 말하는 ‘펫샵’에서 외관만 보고 매력을 느껴 선택하게 된 비율이 가장 많은 것으로 나타났다.



<그림 3. 2017 반려동물 품종을 결정하게 된 이유, KB연구보고서>

이에 대해 의문을 갖고 추가 조사를 하던 중에 해당 문제는 우리나라의 동물보호법에 문제가 있음을 알게 되었다. 사람을 좁은 공간에 가두고 전시하거나 매매하는 것이 불법이듯 독일은 동물보호법으로 반려동물을 전시하여 판매하는 행위를 금지하고 있다. 추가적으로 반려동물 번식업과 판매업에 대한 기준이 한국보다 훨씬 까다롭고 엄격하다. 이에 반려동물을 가족으로 맞이하고자 하는 사람들은 대부분 ‘티어하임(Tierheim)’이라 불리는 독일의 동물보호소를 찾아간다. 티어하임에서는 입양을 희망하는 모든 사람은 가족원들의 동의, 보호소 방문, 입양동물과의 사전 교감 등 반려동물을 제대로 돌볼 수 있는지를 평가하는 입양자격 테스트를 통과해야한다. 이러한 절차는 반려동물을 위한 일종의 안전장치로서 입양희망자에게 자신의 태도, 자원, 환경 등을 사전에 점검해보도록 돕는다. 반면, 대한민국의 동물보호법에 따라 국내 펫샵에서는 반려동물 구매 희망자에게 이러한 사전평가를 시행하지 않고 금액만 지불할 수 있다면 판매하기 때문에 반려동물이 그만큼 쉽게 유기된다.



<그림 4. 독일 민간 동물보호소 티어하임(Tierheim, Berlin)>

위의 두 자료를 종합적으로 분석해보았을 때 팀 BABE는 이와 같은 전제들을 도출할 수 있었다. 첫째, 반려동물의 주인이 반려동물의 행동양식을 좋아하지 않고 관계를 위해 예상보다 많은 비용(high perceived cost)을 지불해야 한다면 유기 확률이 높아진다. 둘째, 국내에서는 반려동물과의 동거에 드는 예상 부담에 대한 충분한 평가나 검토 없이 손쉽게 구매가 가능하기 때문에 유기가 그만큼 쉽다. 두 전제를 근거로 팀 BABE는 이와 같은 근본적인 결론을 도출했다. “독일과 같은 동물보호 선진국들이 이미 시행중인 사례와 같이 반려동물 입양 절차에 있어 희망입양자의 자격요건 판단 절차를 도입하는 것이 반려동물 유기 확률을 낮추는데 기여할 수 있을 것이다.”

Ⅱ. 데이터 수집

2.1. 데이터 출처 소개

견종에 대한 정보는 <https://www.akc.org/dog-breeds> 에서 얻을 수 있었다. 이 웹사이트는 130년의 역사를 자랑하는 American Kennel Club, 세계에서 가장 큰 애견협회의 웹사이트이다. 기본적으로 이 사이트에 가서 페이지 구조를 확인하고, [개발자 도구 – 요소 검사]를 통해 html의 특정 값을 추출했다.

2.2. 데이터 수집 방법 및 이슈

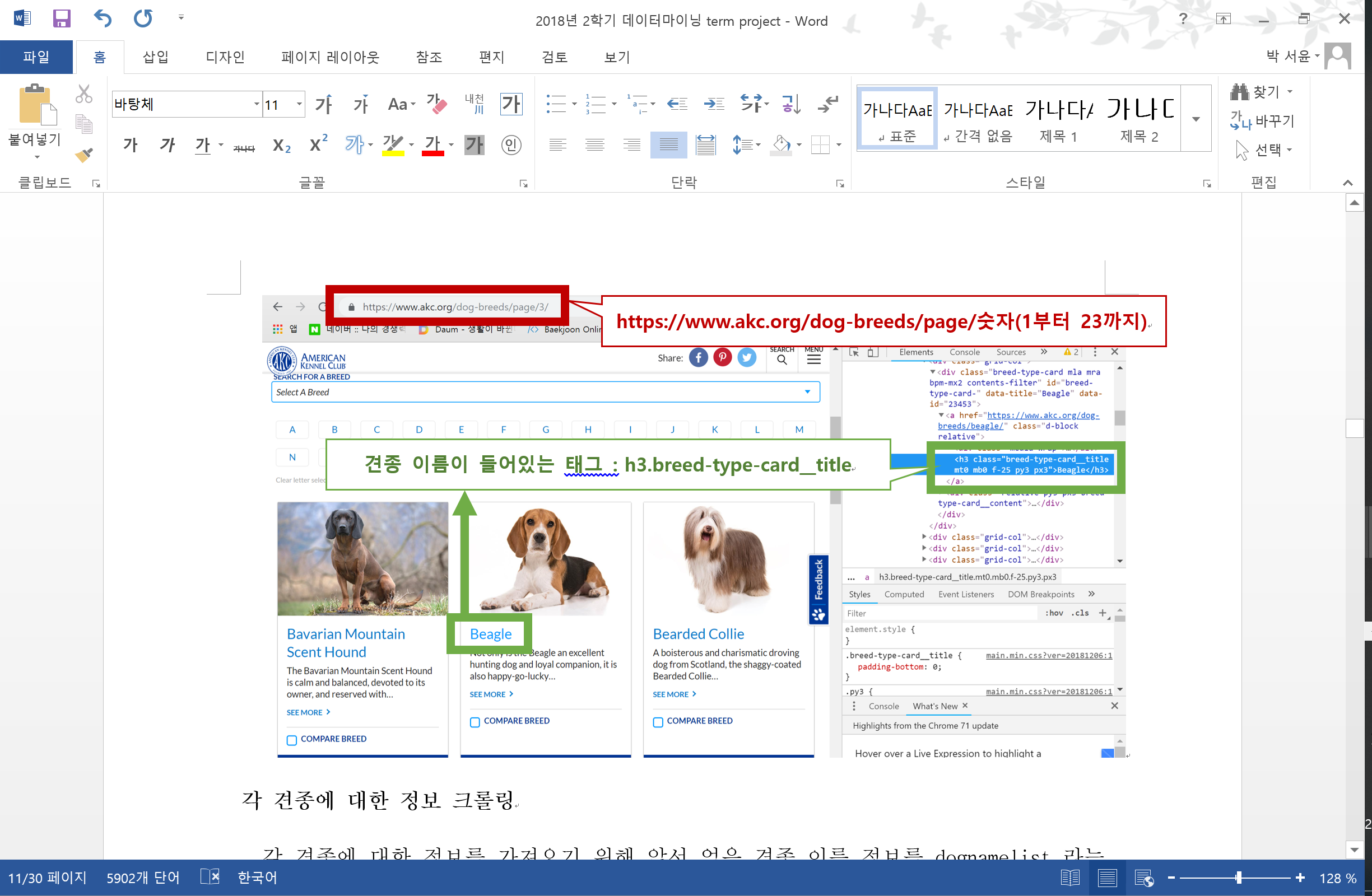
프로그래밍 언어는 Python, 라이브러리는 BeautifulSoup, csv, requests 등을 사용하였다. 데이터는 csv 확장자로 저장하였다.

**1) 모든 견종 이름 크롤링**

모든 견종 이름 데이터를 가져오기 위해 “<https://www.akc.org/dog-breeds/page/>숫자(1부터 23까지)” 페이지의 h3.breed-type-card\_\_title 라는 태그 내용을 불러왔다. 이를 통해 총 272종의 견종 데이터를 얻을 수 있었다.

**2) 각 견종에 대한 정보 크롤링**

각 견종에 대한 정보를 가져오기 위해 앞서 얻은 견종 이름 정보를 dognamelist 라는 리스트에 넣고 이를 그대로 복사한 dogurllist를 만들었다. 그리고 나서 i는 0부터 리스트의 끝까지 반복하며 “'https://www.akc.org/dog-breeds/'+dogurllist[i]+'/'” 주소에 접근하여 특정 태그 내용을 불러왔다. ‘특정 태그’에는 견종을 어떻게 그루밍해야 할지, 얼마나 자주 산책시켜줘야 하는지 등의 정보가 담겨 있다.



<그림 #. AKC 웹사이트와 구조>

**2-1) 이슈 발생**

여기서의 문제 사항은 “<https://www.akc.org/dog-breeds/page/>숫자(1부터 23까지)” 의 h3.breed-type-card\_\_title 태그로부터 얻은 견종 이름을 dogurllist에 넣고 “'https://www.akc.org/dog-breeds/'+dogurllist[i]+'/'” 라는 주소로 접근했을 때, 몇 개의 주소에서 404 not found error 가 발생했다는 점이다. 예를 들어 dogurllist[13] = “Appenzeller Sennenhund”인데, 웹페이지 주소는 “[https://www.akc.org/dog-breeds/Appenzeller Sennenhund/](https://www.akc.org/dog-breeds/Appenzeller%20Sennenhund/)” 가 아니라 “<https://www.akc.org/dog-breeds/appenzeller-sennenhunde/>” 로 단어 사이에 하이픈(-)이 있는 경우, dognamelist[157] = “Löwchen” 처럼 영어 알파벳이 아닌 문자(ö)가 이름에 포함된 경우 등에서 404 not found error가 발생했다. 이러한 경우, 아래와 같이 수작업으로 이름을 변경해주었다.

***# 13 Appenzeller Sennenhund***

**dogurllist[err[0]] = 'appenzeller-sennenhunde'**

***# 17 Australian Stumpy Tail Cattle Dog***

**dogurllist[err[1]] = 'australian-stump-tail-cattle-dog'**

***# 77 Cirneco dell’Etna***

**dogurllist[err[2]] = 'cirneco-delletna'**

***# 120 Grand Basset Griffon Vendéen***

**dogurllist[err[3]] = 'grand-basset-griffon-vendeen'**

***# 157 Löwchen***

**dogurllist[err[4]] = 'lowchen'**

***# 186 Petit Basset Griffon Vendéen***

**dogurllist[err[5]] = 'petit-basset-griffon-vendeen'**

***# 192 Poodle (Miniature / Standard)***

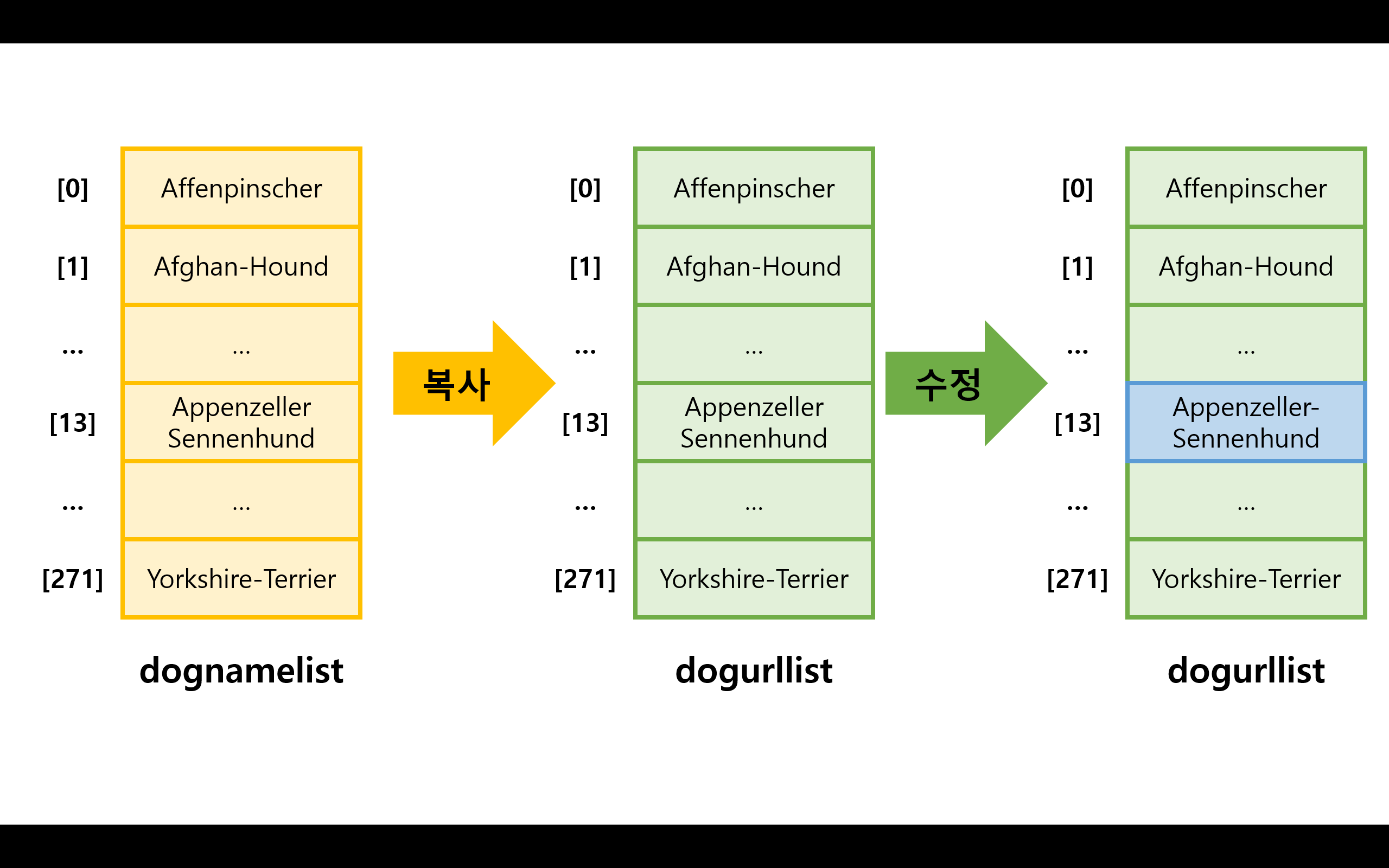
**dogurllist[err[6]] = 'poodle-miniature-standard'**

***# 193 Poodle (Toy)***

**dogurllist[193] = 'poodle-toy'**

***# 215 Saint Bernard***

**dogurllist[err[7]] = 'st-bernard'**



< 그림 #. dognamelist에서 dogurllist로 복사하고, dogurllist를 통해 웹사이트에 접근할 때 404 error 발생하는 이름을 200 정상적 응답 받는 웹사이트에 매치하는 이름으로 수정한 리스트 >

**3) csv 파일로 저장하기**

앞서 크롤링한 데이터들을 컬럼에 맞추어 csv 파일에 저장하였다. Csv 파일을 생성하고, 컬럼 네임을 넣은 뒤, 반복문을 통해 데이터를 각 행에 넣어주었다.

**with open('dog\_cares.csv', 'w', newline='', encoding='UTF8') as outfile:**

**writer = csv.writer(outfile)**

**print(writer)**

**try :**

***writer.writerow(('Breed Name',care\_col[0], care\_col[1], care\_col[2], care\_col[3], care\_col[4]))***

**for j in range(len(dognamelist)) :**

**writer.writerow((dognamelist[j], dog\_cares[j][0], dog\_cares[j][1], dog\_cares[j][2],**

**dog\_cares[j][3], dog\_cares[j][4]))**

**print((dognamelist[j], dog\_cares[j][0], dog\_cares[j][1], dog\_cares[j][2],**

**dog\_cares[j][3], dog\_cares[j][4]))**

**finally :**

**outfile.close()**

**3-1) 이슈 발생**

한 견종에 대하여, NUTRITION, GROOMING, EXERCISE, TRAINING, HEALTH 에 대한 정보를 크롤링하여 넣으려 했지만 어떤 견종은 위 다섯 가지 항목 중 한두개가 누락되어 있는 경우가 발견되었다. 이 경우 위와 같은 반복문 처리에서 index error가 발생하였기 때문에 dog\_cares[j] 의 길이가 5보다 작은 경우를 모두 출력하며 누락된 값이 있으면 해당 컬럼에 빈 string 을 넣어서라도 리스트의 길이를 맞춰주었다.

견종 1의 dog\_cares[1] = [“NUTRITION 관련 정보 … ”, “GROOMING 관련 정보 … ”, “EXERCISE 관련 정보 … ”, “TRAINING 관련 정보 … ”, “HEALTH 관련 정보 … ”] ; 모든 정보가 다 들어가 있으므로 에러 발생하지 않고 정상적으로 처리됨.

견종 2의 dog\_cares[2] = [[“NUTRITION 관련 정보 … ”, “GROOMING 관련 정보 … ”, “EXERCISE 관련 정보 … ”, “TRAINING 관련 정보 … ”] ; HEALTH 관련 정보가 누락되었으므로 error 발생 -> 리스트의 맨 마지막에 “” 이라는 빈 string을 삽입하여 리스트 element의 수를 5개로 맞춰줘야 한다. 그 결과는 아래와 같다.

견종 2의 dog\_cares[2] = [[“NUTRITION 관련 정보 … ”, “GROOMING 관련 정보 … ”, “EXERCISE 관련 정보 … ”, “TRAINING 관련 정보 … ”, “” ]

위와 같은 원리로 크롤링한 스크립트와 데이터에 대한 파일명은 아래와 같으며 별도로 첨부하였다.

|  |  |
| --- | --- |
| 크롤링 스크립트 | 데이터 파일 |
| dog\_cares.html  dog\_cares.ipynb | dog\_cares.csv |
| dog\_features\_general.html  dog\_features\_general.ipynb | dog\_features\_edited\_size.csv |
| dog\_features\_percent.html  dog\_features\_percent.ipynb | dog\_features\_percent.csv |

2.3. 데이터 전처리 및 이슈

1) 인코딩 이슈

|  |  |
| --- | --- |
| Affenpinscher | The혻Affenpinscher should do well on a high-quality혻dog food, whether commercially manufactured or home-prepared with your veterinarian?셲 supervision and approval. Any diet should be appropriate to the dog?셲 age (puppy, adult, or senior). Some dogs are prone to getting혻overweight, so watch your dog?셲 calorie consumption and weight level.혻Treats혻can be an important aid in training, but giving too many can cause obesity. Learn about which혻human foods혻are safe for dogs, and which are not. Check with your vet if you have any concerns about your dog?셲 weight or diet.혻Clean, fresh water should be available at all times. |

위 표는 크롤링한 값을 csv에 저장하였을 때 몇몇 글자가 깨지는 현상이 발생한 내용이다. 따라서 AKC 웹사이트로부터 원래 값을 찾아서, 깨진 값을 엑셀에서 찾기 및 바꾸기 기능을 통해 원래 값으로 바꿔주었다.

2) 표기된 데이터들의 형식

|  |
| --- |
| weight |
| 7-10 pounds |
| 50-60 pounds |
| 50-70 pounds |
| 100-130 pounds (male), 70-100 pounds (female) |

< dog\_features\_edited\_size.csv의 weight 컬럼 >

위와 같이 무게 데이터는 범위로 이루어져 있다. 22파운드 미만 견종은 소형견, 22파운드이상 55파운드 미만 견종은 중형견, 55파운드 이상의 견종은 대형견으로 직접 분류하여 size 컬럼을 만들어 값을 넣었다. Weight 값이 없는 경우, 위키피디아에 검색하여 직접 넣고 분류하였다.

또한 수명(Life\_expectancy) 또한 범위의 형태로 주어졌기 때문에 차후에 이를 프로세스 해야할 상황이 발생할 경우를 염두에 두고 이를 minimum과 maximum으로 열을 나누어 최솟값과 최대값의 형태로 기술하였다. 이 외에 띄어쓰기를 ‘\_’로 대체하거나 대문자들을 소문자로 바꾸는 마이너한 데이터 마사지들을 시행하였다.

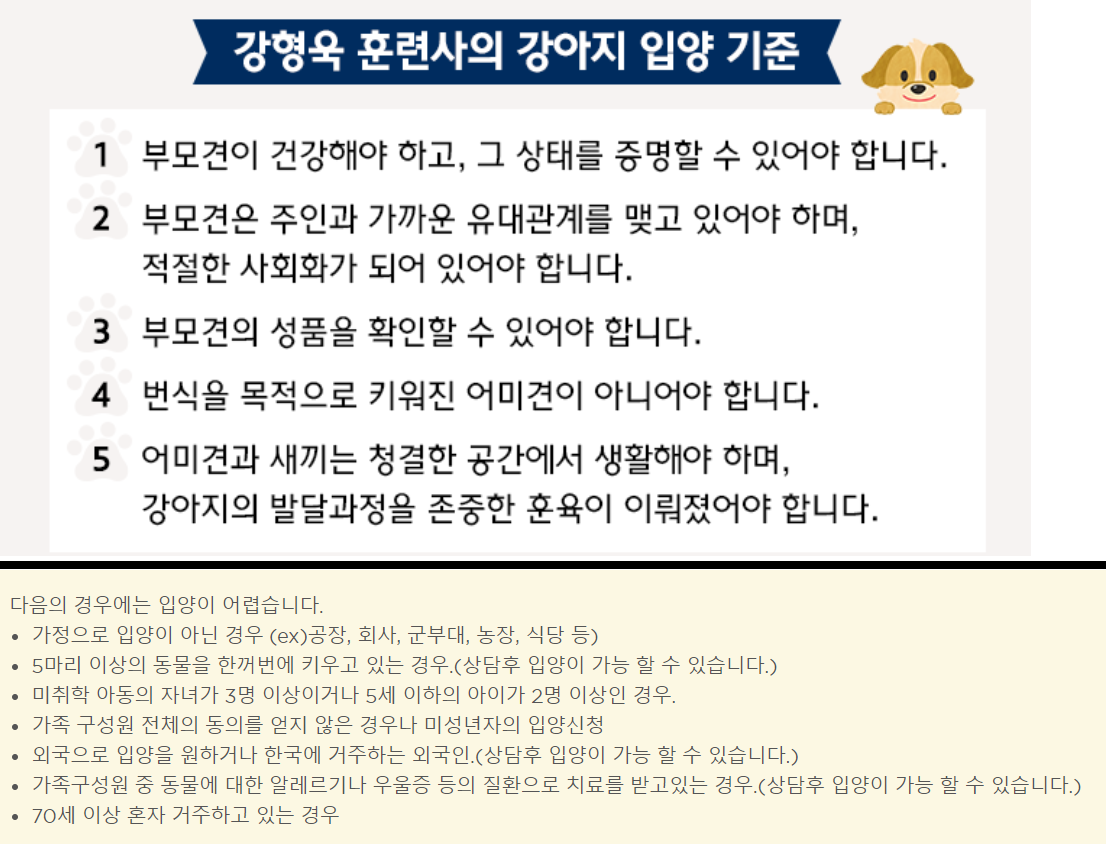
Ⅲ. 서비스 설계

팀 BABE가 제시하는 서비스는 두 단계로 구성되어 있다. 1단계는 반려동물 입양을 하기 위한 최소한의 자격 요건을 판별하는 프로세스, 2단계는 서비스 사용자가 선호하는 반려동물의 특성 및 성격을 선택하는 프로세스이다. 이 프로세스를 거친 후에 사용자에게 적합한 반려견의 견종을 추천하는 시스템을 설계하였다.

3.1. 1단계 - 자격 요건 판별 프로세스

3.1.1. 자격 요건 판별 프로세스 설명

반려동물 입양을 희망할 시에 요구되는 기준은 동물 보호단체나 기관마다 상이하다. 앞에서 언급했던 독일의 민간 동물보호소 ‘티어하임(Tierheim)’의 경우 가족원들의 동의, 보호소 방문, 입양동물과의 사전 교감 등을 자격판별요건으로 꼽는다. 국내 단체들을 보면 최근 대한민국 ‘개통령’으로 화제가 되고 있는 보듬 컴퍼니 대표 강형욱 훈련사는 부모견의 건강 상태와 사회화 여부, 부모견과 새끼견의 발달과정 등의 자체 기준을 가지고 입양을 권하는 것을 확인해볼 수 있다. 또 다른 국내 단체를 살펴보면 앞서 언급한 동물보호단체 카라(KARA)를 들 수 있다. 카라는 “입양을 결정하시기 까지 얼마나 오랜 시간을 고민하셨나요?”, “입양을 원하시는 가장 큰 이유는 무엇인가요?”등의 질문을 통해 입양에 대한 충분한 고민과 적절한 이유에 대해 생각하게 만든다. 추가적으로 “정원이 있다면 담이나 울타리가 되어있나요? 되어있다면 높이는 얼마나 되나요?”와 같은 질문을 통해 주거 환경에 대한 점검을 하고 “입양 후 모니터링을 위한 전화연락이나 방문에 동의하십니까?”등의 질문을 통해 입양 후 관리절차에 대해서도 강조하는 모습을 볼 수 있다. 이와 더불어 입양 신청서의 가장 상단에 입양 결격 사유를 명시함으로 자격요건을 한층 더 엄격하게 제정한 것을 확인할 수 있다.



< 그림 #. (위) 강형욱 훈련사의 입양 기준

(아래) 동물보호단체 카라(KARA) 입양결격사유 >

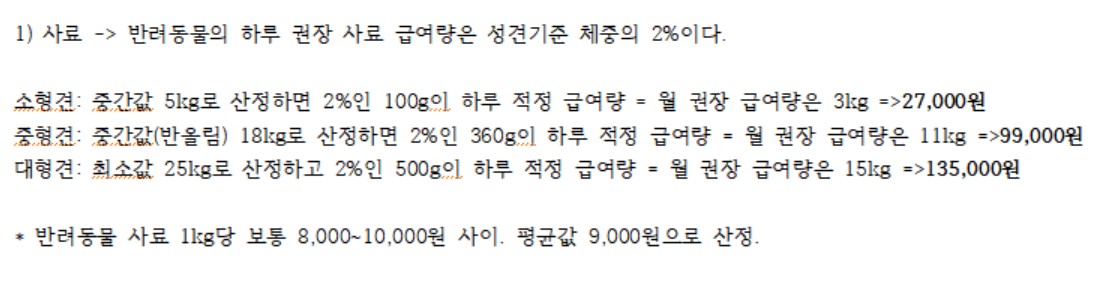
이처럼 단체에 따라 입양 절차가 다양하다는 말은 즉 각 단체마다 중요하게 생각하는 입양 기준이 다르다는 말이다. 이 말을 다시 한번 뒤집어서 해석해보면 단체마다 가정하는 유기 사유가 상이하다는 뜻으로도 볼 수 있을 것이다. 그러나 다양한 기준 속에서도 공통적으로 명시되어있는 부분이 존재하는 것은 사실이다. 이에 팀 BABE는 지금까지 분석한 유기 사유들을 종합하고 각종 유기동물 입양단체 기준을 참고하여 자체적인 자격 판별 절차를 작성하기로 결정했다. 팀 BABE는 경제적 요건, 환경적 요건, 개별적 요건, 총 3가지 부문으로 자격 판별 절차를 나눠 각 영역에서 필수 질문을 함으로 입양을 희망하는 사람의 입양 적합여부를 판별해보고자 한다.

첫째로 경제적 요건을 살펴보자. 경제적 요건에 관해 팀 BABE의 자격요건판별 질문은 하나의 질문으로 압축되어 있다.

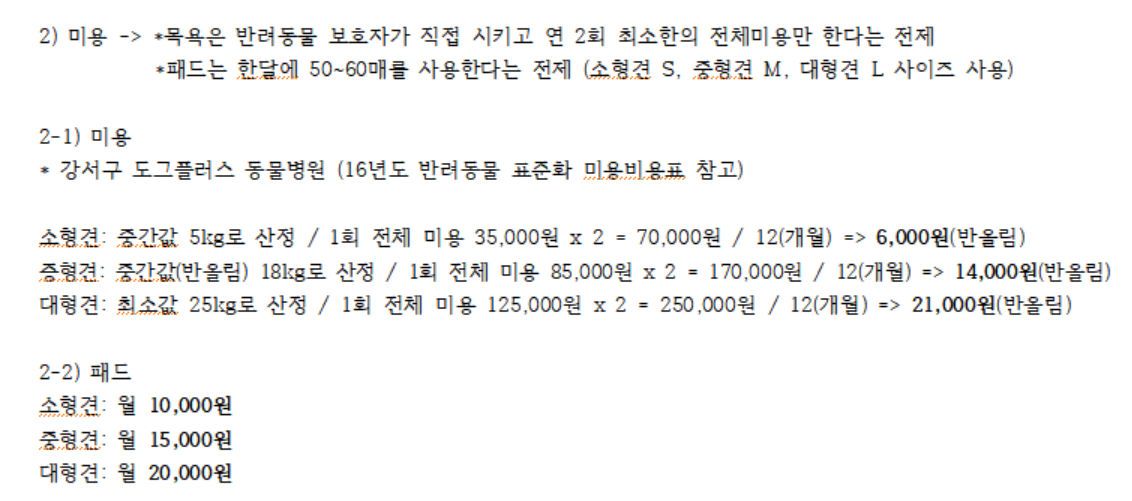
**1. 반려동물 생활비에 고정적으로 지출 가능한 한달 비용은?**

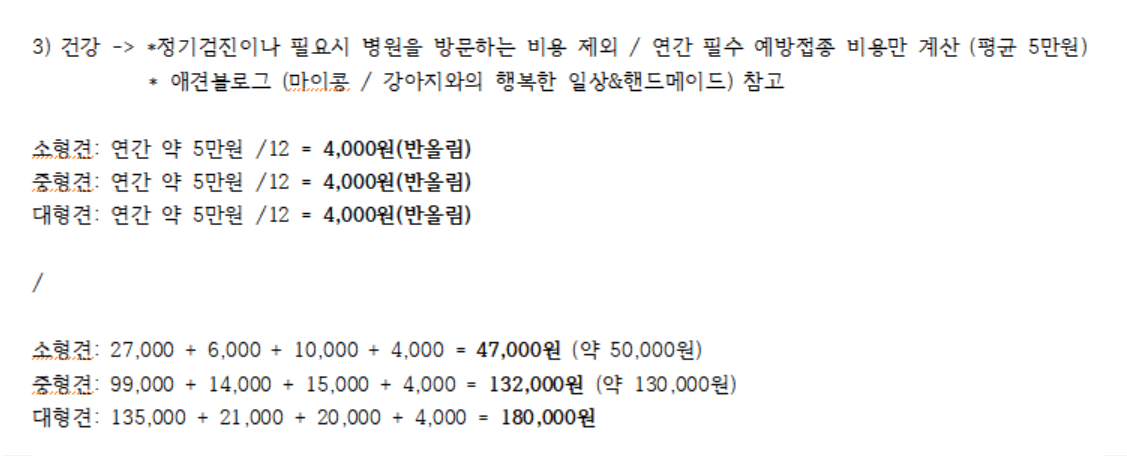
1. **5만원 이하 2) 5~13만원 3) 13만원~18만원 4)18만원 이상**

팀 BABE는 해당 질문을 통해 입양희망자가 반려견과의 동거에 드는 최저 예상 비용을 지불할 능력이 되는지를 물어보고자 한다. 반려견의 크기 별로 필요로 하는 최소 비용이 다르기 때문에 희망하는 반려견을 키우는데 적합한 경제적 능력을 갖추고 있는지에 대한 문항이다. 반려동물 생활비용 3대 주요 항목은 사료, 미용(패드+전체미용), 건강(필수 예방접종)이다. 소형견은 성견 기준 몸무게 10kg 미만, 중형견은 성견 기준 몸무게 25kg 미만, 대형견은 성견 기준 몸무게 25kg 이상으로 국내 표준이 설정 되어있다. 팀 BABE는 이를 바탕으로 반려동물 월 평균 생활비용을 산출해보았다. 모든 계산에 사용한 몸무게값은 성견 기준 몸무게 중간값으로 하였고 산출된 월 평균 생활비용은 적정 비용이 아니라 반려동물 양육에 필요한 최소 생활비용임을 강조하는 바이다. 질문의 각 답안을 계산한 기준과 방식은 아래 사진에 요약해놓았다.



<그림 #. 반려동물 월 평균 사료비용 계산 방식 및 기준>



<그림 #. 반려동물 월 평균 미용 비용 계산 방식 및 기준>

<그림 #. 반려동물 월 평균 건강 비용 계산 방식 및 기준>

둘째로는 환경적 요건이 있겠다. 총 4가지 질문으로 구성된 환경적 영역은 대한민국 농림수산식품부 소관의 사단법인인 비영리단체 ‘동물자유연대’의 입양기준의 환경부문을 참고하였다. 4가지 질문은 아래와 같다.

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

**(2번 질문에 ‘X’라고 답했다면)**

**1) 단독주택 2) 다가구주택(오피스텔/아파트 등) 3) 기타**

**3. 강아지가 혼자 있어야 하는 시간이 8시간 이상이다. (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**

자격요건판별 질문 영역의 마지막 부문으로 개별적 요건 확인이 있겠다. 개별적 요건은 카라(KARA)와 동물자유연대의 입양절차기준을 참고하여 재구성해보았다. 개별적 요건에 대한 질문은 총 5가지로 아래와 같다.

**6. 반려동물 입양에 대한 가족 및 동거인 간 의견 합치 여부 (O/X)**

**7. 본인 또는 동거인이 우울증 등 정신 질환이 있나요? (환경)(O/X)**

**8. 반려 동물을 키우다 중간에 포기한 경우가 2번 이상인가요? (경험)(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)**

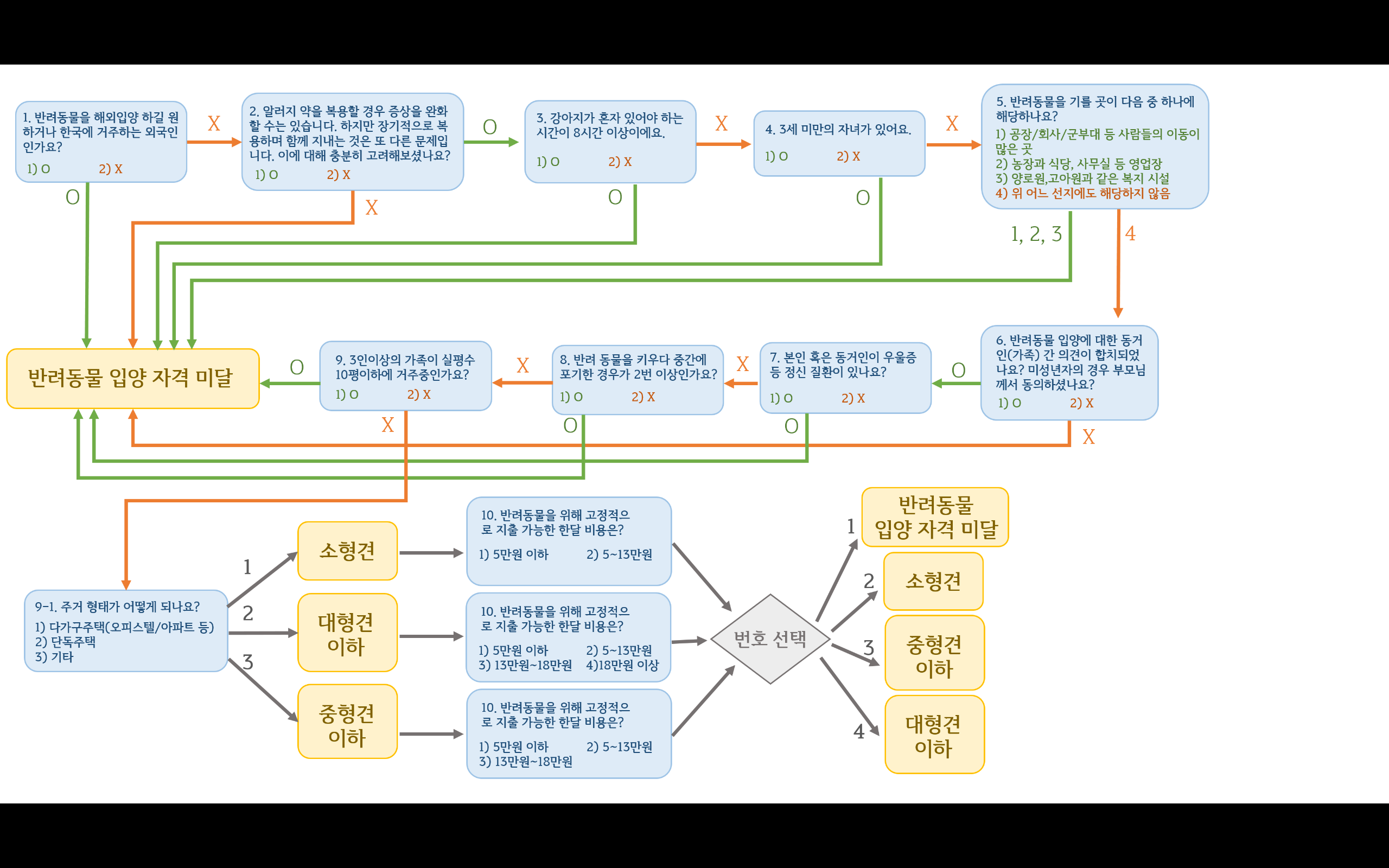
**(e.g 재채기, 기침, 콧물, 코막힘, 눈 가려움증, 충혈, 피부 발진, 두드러기, 호흡곤란, 가슴, 답답함, 천명 (호흡 시 쌕쌕거림) 등)**

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

총 10가지 질문으로 구성된 자격요건판별 시스템은 팀 BABE의 솔루션의 1차적인 필터로서 작용한다. 각 답항은 결격사유 판단 필터로 작용하기도 하고 동시에 최종 품종 추천 알고리즘의 분류 필터로 작용하기도 한다. 예를 들어, 반려동물 생활에 고정적으로 지출 가능한 한달 비용을 묻는 1번 문항에 2번 (5~13만원)을 체크했다면 알고리즘의 필터링에 의해 자동적으로 추천 범위가 소형견으로 축소된다. 다른 예를 또 들어보자면 입양 부적격 환경에 관한 5번 문항에 하나라도 체크를 한다면 최종 결과로 “반려견을 입양하기에 부적합”이라는 판정을 받게 되는 것이다. 자세한 사항은 알고리즘과 필터에 대한 구체 설명 부문에 설명 되어있다.

3.1.2. 자격 요건 판별 프로세스 알고리즘 및 코드

위 자격 요건을 판별하는 알고리즘을 아래와 같이 도식화하였다. 구현 상의 편의를 위하여 개별적 요건, 환경적 요건, 경제적 요건 순으로 질문지를 구성하였다.



<그림 #. 자격요건판별 알고리즘>

1번부터 9번까지의 대답을 통해 반려동물 입양 자격이 있는지, 자격 미달인지를 판단한 후 반려동물 입양 자격이 있는 경우 주거 형태 질문과 지출 가능 비용 질문을 통해 입양 가능한 반려견 풀을 소형견, 중형견 이하, 대형견 이하로 좁혔다. 이에 대한 코드는 아래와 같다. Python 3.7 을 통해 구현하였다. 이 함수의 call이 완료되면 size를 리턴하는데, 이를 통해 적합한 반려동물의 사이즈를 알 수 있다. Size가 ‘S’이면 소형견, ‘M’이면 중형견 이하, ‘L’이면 대형견 이하이며 ‘ ’이면 자격 미달이다.

코드는 따로 첨부한 IsQualified.py 파일의 is\_qualified() 함수 안에 구현되어 있다.

3.2. 2단계 - 선호도에 근거한 필터링 프로세스

3.2.1. 견종 성격 선호 테스트

**1. 이론적 배경 및 방법론**

위의 적합도 검사를 통과한 사람들에 한해 선호도 검사를 시행한다. 자격요건이 갖춰진 예비 입양자를 대상으로 진행되는 이 선호도 검사는 다음과 같은 전제에서 시작한다.

*There was, however, a significant tendency for partners to appear more similar in personality traits the longer they had been married.[[5]](#footnote-5)*

앤써니 C. 리틀 등 3명의 연구에 따르면 사람은 본인과 유사한 성격을 가진 사람과 더 오래 행복하게 살 가능성이 높았다. 결론적으로 본인을 잘 아는 것은 타인이 아닌 본인이라는 점에서, 본인이 직접 자신의 성격과 가까운 반려견을 직접 선택하고 이에 기반하여 추천한다면 입양 후 만족도는 높아질 것이다.

견 종 추천을 위해 성격, 외형, 털 손질 빈도, 털 빠짐 정도, 활동량, 총명함, 훈련 시 반려견의 태도를 묻는다. 먼저 예비 입양자가 선호하는 반려견의 성격을 알기 위해 다음의 질문을 제시한다.

*“다음은 강아지들의 기질을 나타내는 단어들입니다.*

*이들 중 당신과 가장 가까운 모습이라고 생각되는 단어들을 골라주세요.*

견종의 성격 특성을 소개하고 있는 American Kennel Club(이하 AKC)에서 성격 키워드들을 출력한 다음, 예비 입양자로 하여금 선택하도록 한다.

이를 위해 해당 웹사이트를 파이썬 언어를 사용해 작성한 스크립트를 이용해 크롤링하였고 이를 다시 정제하여 각 견 종 별 성격 키워드 세 가지/크기/무게/털갈이 정도/필요로 되는 운동량/훈련 가능성 과 같은 특성들을 추출하여 데이터셋으로 가공하였다. 견종별 특성 데이터는 성격을 규정짓는 세 가지 키워드 외에 모두 수치 데이터였기에 데이터에 대해 별도의 변환 작업을 실시하지 않고 곧바로 분석을 시행할 수 있었다. 하지만 성격을 규정짓는 세 가지 키워드는 텍스트 데이터였기에 이를 Rapid Miner와 같은 툴 상에서 데이터 마이닝 프로세스를 수행하기 이전에 컴퓨터가 인지할 수 있는 체계 혹은 차원의 데이터로의 변환이 필요했다. 이를 위해 우선 성격 키워드에 대한 EDA를 실시하였고 그 결과 몇 가지의 키워드들이 반복되는 것을 관측했다. 이를 통해 특정 성격적 특성에 대한 정의가 일관성을 갖추고 있음을 확인하였으므로 추가적인 분석을 위해 우선 반복되는 것을 제외하고 전체 데이터에서 관측된 키워드들을 추려내 목록으로 만들어보았다. 그 결과 견종별 성격을 규정 짓는 성격 키워드는 총 61가지의 텍스트 데이터로 구성됨을 알 수 있었다. 상술한 서비스를 실체화 하기 위해선 견 종 별 성격 키워드를 기준으로 견 종을 구분해 군집화(Clustering)할 필요가 있었으므로 이를 위해 해당 텍스트 데이터를 분석할 수 있는 알고리즘에 대한 고민을 시작하였다.

군집화는 유사한 것들이 하나의 공통된 성질을 갖는다고 전제하고 이들을 인지하기 편리하도록 각각의 군집으로 묶는 분석을 말한다. 이 때 유사한 것들의 기준엔 다양한 것들이 이용되는데 가장 대표적으로 사용되는 것이 “거리”이다. 유클리디안 거리가 가장 대표적으로 사용되는 거리 측정 방식인데 결국 ‘거리’라는 것은 공간의 개념 속에서 존재할 수 밖에 없는 것이기에 우리는 성격 데이터들을 어떻게 공간에 위치시킬 수 있을 지에 관해 고찰하기 시작했다. 하지만 성격 데이터는 1)텍스트이며, 2) 61차원의 데이터이기 때문에 데이터 그 자체로는 지각 혹은 관측이 불가능한 차원 상의 데이터였고 그렇기에 인지의 차원에서만 다루어 볼 수 있었다. 우선 처음 시도해본 방식은 각각 다른 61가지의 성격 데이터들을 각각의 차원으로 상정한 후 좌표의 형태로 데이터를 표현해보는 것이었다. 하지만 이 방식은 결정적으로 각각의 성격 키워드들이 갖는 ‘의미(Semantics)’를 전혀 반영하지 못한다는 점에서 한계를 가졌다. 이를 위해 우리가 직접 61가지의 키워드들을 의미 기준으로 다섯 가지의 군집으로 군집화한 후 이에 근거하여 단어 간 거리에 가중치를 부여하는 방식을 고려해보았으나 이러한 방식은 우리가 셀 수 있는 차원의 데이터이기에 적용 가능할 뿐 데이터의 개수가 우리가 직접 손으로 군집화할 수 없을 만큼 만을 경우엔 전혀 사용할 수 없는 방식이기에 신뢰성이 떨어진다고 판단, 기각하였다.

이후 다양한 논문과 이론을 탐색한 결과, 우리의 목적에 적합한 이론을 발견할 수 있었다. 의미에 기반한 유사도(Semantic Similarity)를 측정하기 위해 측정 대상이 되는 단어를 함께 검색엔진에 검색한 결과로 출력되는 페이지의 수와 텍스트 스니펫(Text snippets) 속 두 단어의 구조적 연관관계를 사용한 연구가 존재했다. 연구 결과, 두 단어가 의미적으로 유사 관계에 있을 경우 검색 엔진 상에서 두 단어가 함께 검색된 결과의 페이지 수가 그렇지 않을 때보다 많다는 것이었다. 실제로 apple/computer 두 단어가 함께 관측된 페이지의 수가 apple/banana보다 80배 더 많은 288, 000, 000개 였다고 한다. 이를 귀무가설로 설정한 후 Correlation Coefficient 테스트를 시행한 결과 0.834의 상관관계가 있는 것으로 확인되었다고 한다. 이는 충분히 유의미한 수치로 유사도와 검색 결과 페이지 수 간의 상관관계가 있다고 할 수 있는 수치이다. 또한 이에 대한 F-Measure를 시행한 결과 또한 0.78로 귀무가설이 성립함을 뒷받침해주고 있다. 하지만 검색 결과 페이지 수에만 근거한 유사도 측정은 한계를 갖는데 첫 번째로 두 단어가 페이지 내에서 어떤 맥락과 관계로 함께 다루어 지고 있는 지에 관한 것은 반영하지 못한다는 것이다. 두 번 째로 대상이 되는 단어가 다양한 맥락에서의 정의를 함께 갖고 있을 경우 이에 대한 결과를 중복으로 포함할 가능성이 존재한다는 점이다. 가령 Apple이라는 키워드를 검색한 결과 내에 기업 ‘Apple’과 과일 ‘Apple’이 함께 존재할 가능성이 다분하다는 것이다. 우리가 원하는 Apple의 의미 외에 다른 의미를 갖는 결과가 반영되기에 데이터에 노이즈가 발생하는 것이다. 따라서 검색 결과 페이지 수 만으로는 충분한 신뢰성을 갖춘 의미 유사도를 측정할 수 없다.

이를 보완하기 위해 검색 결과 중 스니펫(Snippets)을 분석하여 두 단어의 직접적인 의미관계를 반영하는 방식을 더하여 분석을 진행한다. 스니펫(Snippet)이란 검색 결과로 출력되는 항목 옆에 링크로 연결되는 페이지의 내용이 요약되어 출력되는 본 내용의 토막을 의미한다. 대상이 되는 두 단어를 함께 검색한 결과 중 이 두 단어가 함께 검출된 스니펫의 리스트를 추출한 후 해당 스니펫 내의 대상 단어들을 미지수 X, Y로 대체한다. 이를 대체하는 이유는 언어에서 문장이라는 것이 단어와 단어의 연결로 이루어지고 이러한 단어와 단어가 연결되는 방식이 일정한 양식을 갖추고 있음에 기인한 것이고 이렇게 미지수로 대체함으로써 언어 간 유사관계를 나타내는 구조적인 양식을 도출해내기 위해서이다. 가령 Cricket 과 Sport를 함께 검색한 결과 스니펫 중 “**Cricket** is a **sport** played between two teams, each with eleven players.”라는 것이 존재했다고 가정할 수 있다. 이때 이 두 단어 사이에 있는 “is a” 혹은 “is known as”와 같은 것들이 바로 앞서 말한 공통적으로 나타나는 두 어휘 간 의미 유사를 나타내는 구조적 양식인 것이다. 그렇기에 이를 X, Y라는 미지수로 치환함으로써 어휘 간 의미 유사를 나타내는 구조적 양식, 어휘 구문론적 패턴(Lexico-Syntactic Patterns)을 추출해낼 수 있는 것이다. 이를 도출할 수 있다면 후에 크롤링 할 스니펫 속에서 의미 유사도를 측정하고 싶은 단어들이 유의관계를 나타내는 어휘 구문론적 패턴과 함께 출현할 경우 이 두 가지를 유사한 관계로 정의할 수 있을 것이다. 하지만 위에서 서술한 “is a” 혹은 “is known as”처럼 두 단어 사이에 어휘 구문론적 패턴이 위치하는 것은 아주 일반적이고 직접적인 구문이기에 파악하기가 편리하지만 모든 문장들이 이처럼 직접적이라고 할 수는 없을 것이다. “Toyota and Nissan is a car-manufacturing company.”과 같이 두 단어 사이가 아닌 바깥에 어휘 간 관계를 설명해주는 구문론적인 패턴이 위치할 경우도 분명히 다수 존재할 것이기 때문이다. 이처럼 일반적이지 않은 패턴을 정교하게 분석하기 위해서는 보다 더 정교한 분석이 필요한데 웹 상의 일반적인 패턴 혹은 일반적이지 않더라도 알고리즘으로 대응할 수 있는 범위까지의 구문론적 패턴을 갖는 스니펫의 갯수만으로도 분석 시행이 필요로 하는 최소한의 신뢰성을 갖추기에 충분하다고 할 수 있다. 연구 논문의 저자는 그렇기에 알고리즘을 고안하며 지나치게 복잡한 패턴 보다는 일정 수준 이상의 분석을 필요로 하지 않는 패턴을 상정하였다고 한다. 앞서 언급하였듯 이와 같이 두 단어의 관계를 알려주는 어휘 구문론적 패턴을 추출하는 이유는 유사도를 측정하고 싶은 두 단어가 사전과 유사어 말뭉치에 등재 되어있지 않은 비공식적인 유사관계일 경우 이를 확인하기 위해서 필요하기 때문이다. 이를 위해서 WordNet과 같은 미리 만들어진 말뭉치 중 유사어 뭉치들 사이에서 나타나는 어휘 구문론적 패턴과 유사관계로 정의 되지 않은 단어들 사이에서 나타나는 패턴들을 아래 기술할 알고리즘을 이용해 각각 추출한다. 그 후 대상이 되는 두 단어를 함께 검색한 결과 스니펫 속에서 추출되는 어휘 구문론적 패턴 중에서 어느 쪽이 더 많이 관측되었는지를 비교함으로써 의미 유사도 측정 대상이 된 두 단어의 관계를 유추하는 것이다. 이때 어휘 구문론적 패턴을 추출하는 알고리즘은 다음과 같다.

**Algorithm 3.1: ExtractPatterns(S)**

for each word-pair (A,B) S

do D <- GetSnippets(“A B")

N <- null

for each snippet d D

do N <- N + GetNgrams(d, A, B)

Pats <- CountFreq(N)

return (Pats)

1. **어휘 구문론적 패턴 추출 알고리즘**

이때 S는 WordNet과 같은 말뭉치에서 유사어로 정의된 단어들의 조합 집합을 의미한다. GetSnippets(“A B”)는 웹 상에서 A와 B를 검색한 결과로 출력된 Snippets들을 크롤링하는 행위를 의미하며, 이렇게 추출된 모든 스니펫들의 집합을 D로 정의한 것이다. 이후 수행되는 프로세스를 계속해서 이해하기 위해선 N-gram의 개념에 대해 파악할 필요가 있다. N-gram이란, 문자열에서 연속되는 N개의 구성요소를 추출하여 만든 문자열을 의미한다. “I am your father”와 같은 문장에서 2-gram을 구해보면 다음과 같다.

**2-gram**

I am

Am your

Your father

N을 Null로 초기값을 설정해준 후엔 for이라는 구문에서 알 수 있 듯, 특정 스니펫 d로부터 N-gram을 시행하는 것을 반복한다. 이후 N-gram을 시행했을 당시의 N-gram으로 구해진 문자열의 갯수를 패턴으로 상정하고 이 패턴의 갯수를 새어 Pats라는 객체로 선언 후 이를 출력한다. 이렇게 추출한 패턴들을 상술하였듯, WordNet과 같은 말뭉치에서 사전에 정의된 유사관계의 단어들과 그렇지 않은 단어들 간의 스니펫 검색 결과를 추출한 후 이들로부터 패턴을 추출하여 의미 유사 관계에서의 어휘 구문론적 패턴/그렇지 않은 경우의 어휘 구문론적 패턴으로 구분한다. 그리고 대상이 되는 단어들의 스니펫 검색 결과들로부터 패턴을 추출한 후 이들 속에서 양 쪽의 어휘 구문론적 패턴 빈출을 확인하고 어떤 쪽의 패턴이 통계적으로 유의미하게 검출되었는지 분석해봄으로써 두 단어 사이의 관계를 유사어와 그렇지 않은 것, 둘 중 한 쪽으로 정의할 수 있는 것이다.

이와 같은 두 가지 관점에서의 의미 유사도 측정을 통해 두 가지 관점에서의 유사도를 수치 형태의 결과로 데이터를 출력받을 수 있다. 이렇게 도출된 두 가지의 수치 데이터를 하나로 통합하여 의미 유사도를 벡터데이터로 변환하는 프로세스를 거치면 우리는 두 단어 간 의미의 유사도를 수치 데이터의 형태로 인지할 수 있게 된다. 이렇게 가공된 데이터를 클러스터링하면 결과적으로 단어들의 의미에 기반한 유사도에 근거해 이들은 군집화할 수 있게 되는 것이다. 우리의 견 종 성격 키워드에도 충분히 적용 가능한 알고리즘이지만 61가지 단어들 간의 유사도를 모두 체크해야하기에 위의 연산을 61\*61의 횟수만큼 반복해야한다는 점에서 작은 한계를 가진다. 또한 해당 알고리즘을 직접 진행해보려고 시도하였으나 필요로 되는 프로세스들이 웹 크롤링부터 직접 각각의 프로세스들을 진행하기 위한 기능(Function)들을 개개로 정의하고 사용할 필요가 있다는 부분을 확인하였다. 하지만 안타깝게도 이러한 부분들을 진행할 수 있을 만큼 숙련된 개발 능력을 갖고 있지 못했기에 모든 프로세스들을 직접 진행해볼 수 없었다는 점에서 우리의 프로젝트는 한계를 갖는다. 하지만 이러한 과정을 통해 프로젝트 진행을 통틀어 가장 큰 문제가 되었던 “의미 기반의 단어 유사도 측정”이에 기반한 클러스터링 방안에 대해 정확한 응답이 가능하게 되었으며 결과적으로 자연어 처리 프로세스(NLP)와 클러스터링과 같은 일반적인 데이터 마이닝 알고리즘을 적용하는 것 사이의 큰 간극을 해결하는 방법을 알게 되었다는 점이 고무적이었다고 할 수 있다. 이 후의 프로젝트 프로세스들은 이러한 알고리즘을 적용하여 클러스터링이 완료되었다는 전제 하에 진행될 방식들을 서술하도록 해보겠다.

견종의 성격 키워드들을 상술한 알고리즘을 이용해 의미를 반영한 유사도를 측정하고 이에 근거해 클러스터링을 한 결과로 약 5 가지의 군집으로 분류해낼 수 있다. 이때 클러스터링 알고리즘으로는 k-means 알고리즘이 적절할 것으로 판단된다. 이 군집들은 강아지를 추천을 위한 POOL로 사용될 것이다. 한 성격 군집이 예비 입양자에게 선택이 된다면 견종 특징 선호도 조사를 통해서 견종 추천을 더욱 구체화한다.

3.2.2. 견종 특징 선호 테스트

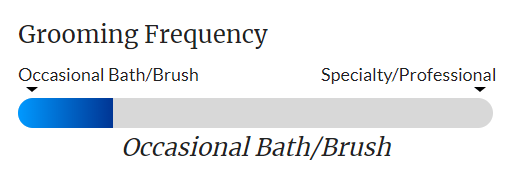
견종 특징 선호도 조사에서는 예비 입양자들에게 외형, 털 손질 빈도, 털빠짐 정도, 활동량, 총명함, 훈련 시 반려견의 태도를 묻는다. 이 문항들은 AKC 홈페이지에 있는 자료에 근거하여 척도화한다. 외형은 소형, 중형, 대형으로 구분하고 나머지 문항들은 5점 척도(예비)로 다음과 같이 구성한다.

**1. 당신은 어느 크기의 견종을 선호합니까?**

1) 소형/10KG 미만 2) 중형/25KG 미만 3) 대형/25KG 이상

**2. 반려견의 털 손질은 몇 번이 적당하다고 생각하십니까?**

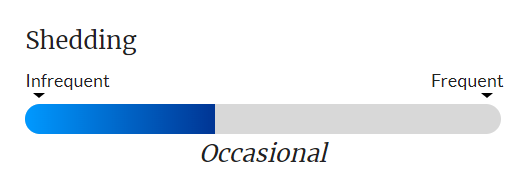
1) 가끔 목욕할 때 2) 매주 한 번 브러싱 3) 주 2-3회 브러싱 4) 매일 브러싱 5) 아주 전문적으로 매일



<그림 #. 샘플 견종에 대한 Grooming Frequency, 출처 AKC>

**3. 미래의 반려견의 털갈이에 대하여, 최대 얼마만큼까지 수용 가능합니까?**

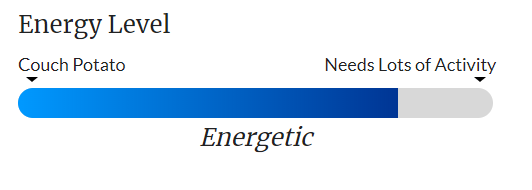
1) 드물게 2) 때때로 3) 계절마다 4) 정기적으로 5) 자주



<그림 #. 샘플 견종에 대한 Shedding, 출처 AKC>

**4. 당신은 산책의 강도와 빈도에 있어서 최대 얼마나 활기찬 반려견을 감당할 수 있으십니까?**

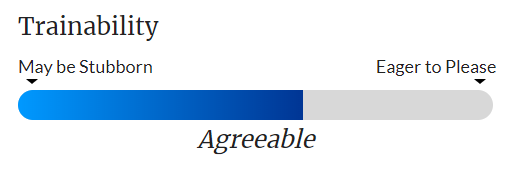
1) Couch Potato 2) 차분한 3) 보통 활동량 4) 활기찬 5) 엄청난 활동량을 자랑하는



<그림 #. 샘플 견종에 대한 Energy Level, 출처 AKC>

**5. 당신이 반려견을 트레이닝 할 때 최대한으로 다룰 수 있는 반려견의 성격은?**

1) 고집이 있는 반려견 2) 독립적인 반려견 3) 순응하는 반려견 4) 트레이닝이 쉬운 반려견 5) 주인을 위해 무엇이든 할 준비가 된 반려견



<그림 #. 샘플 견종에 대한 Trainability, 출처 AKC>

견종 특징 선호도 조사 알고리즘은 어떤 견종의 특성값과 사용자가 선택한 특성값과 큰 차이가 없을수록, 그리고 사용자가 인내할 수 없는 특성일수록 높은 점수가 나오는 방식으로 진행된다. 가장 높은 점수를 가진 견종들이 사용자가 선호하는 견종이 된다. 아래 예시를 통해 견종 특징 선호도 테스트 알고리즘을 설명할 것이다.

예를 들어, 사용자 A는 털날림은 다소 싫어하지만 반려견 산책의 강도나 빈도는 높아도 상관 없는 사람이다. (이하 ‘털’은 털갈이 수준을, ‘활’은 활동량 수준을 나타낸다. ‘털’은 5점 척도이며, 5에 가까울수록 털갈이 빈도가 높음을 의미한다.‘활’또한 5점 척도이며, 5에 가까울수록 활동량이 높아 산책을 자주 시켜줘야 함을 의미한다.)

A의 선택: (털 2, 활 5) = 털날림은 싫지만, 반려견 산책의 강도는 높아도 괜찮은 사람

우선,‘용인할 수 있는 정도’에 따라 가중치를 부여한다. 어떤 특성에 대한 수용도가 낮을수록 중요한 특성이므로 큰 수 가중치로 부여하고, 어떤 특성에 대해 수용도가 높을수록 어떻든 상관 없다는 뜻이므로 작은 수를 가중치로 부여한다.

사용자 A의 경우, 털 날리는 것을 다소 싫어하므로 털날림 값에 다소 큰 가중치를 부여한다. 반면, 강아지의 활동량은 높아도 수용 가능하므로 활동량 값에 작은 가중치를 부여한다. 털날림의 경우 5점 척도의 역수를 가중치로 부여한다. 사용자가 1을 선택한 경우, 털날림을 극도로 싫어하기 때문에 1점의 역수인 5점이 가중치가 된다. 사용자가 5를 선택한 경우, 털이 날려도 상관 없기 때문에 5의 역수인 1이 가중치가 된다. 같은 원리로 활동량의 가중치는 1이 된다.

가중치를 구한 다음, 각 견종 특성과 사용자가 선택한 값 사이의 오차를 구한다. 이 오차가 작을수록 사용자의 선호와 특정 견종이 잘 맞는다는 것을 의미한다.

강아지 1 (털4, 활2) --> 사용자 A의 선호도 (털 2, 활 5)와의 오차 (2, 3) = 오차 합 5

강아지 2 (털5, 활3) --> 사용자 A의 선호도 (털 2, 활 5)와의 오차 (3, 2) = 오차 합 5

두 강아지와 사용자 A의 선호도 사이의 오차합은 같다. 그러나 상식적으로 생각해보면 사용자 A는 털이 날리는건 싫고 훈련 강도가 높은 것은 괜찮으므로 강아지 1이 강아지 2보다 낫다. 따라서 이 점을 반영하기 위해, 앞서 구한 가중치를 오차에 곱할 것이다. 그러나 가중치는 클수록 중요하고 오차는 작을수록 좋다. 따라서 두 값을 곱함에 따라 상쇄 작용이 일어나는 것을 방지하기 위하여 오차를 변환한다. 털날림과 활동량의 경우, 5점 척도의 역수로 변환한다. 4는 1, 3은 2, 2는 3, 1은 4, 0은 5로 변환된다. 그 결과는 아래와 같다.

강아지 1: 사용자 A의 선호도와의 오차 (2, 3) --> 변환 후 (3, 2)

강아지 2: 사용자 A의 선호도와의 오차 (3, 2) --> 변환 후 (2, 3)

오차를 변환하고 가중치도 구했으면, 오차를 변환한 값에 가중치를 곱한 합을 구한다.

강아지 1 오차변환 후 (3,2) \* 가중치(4(다소 싫음), 1(상관 없음)) = 12 + 2 = 14

강아지 2 오차변환 후 (2,3) \* 가중치(4(다소 싫음), 1(상관 없음)) = 8 + 3 = 11

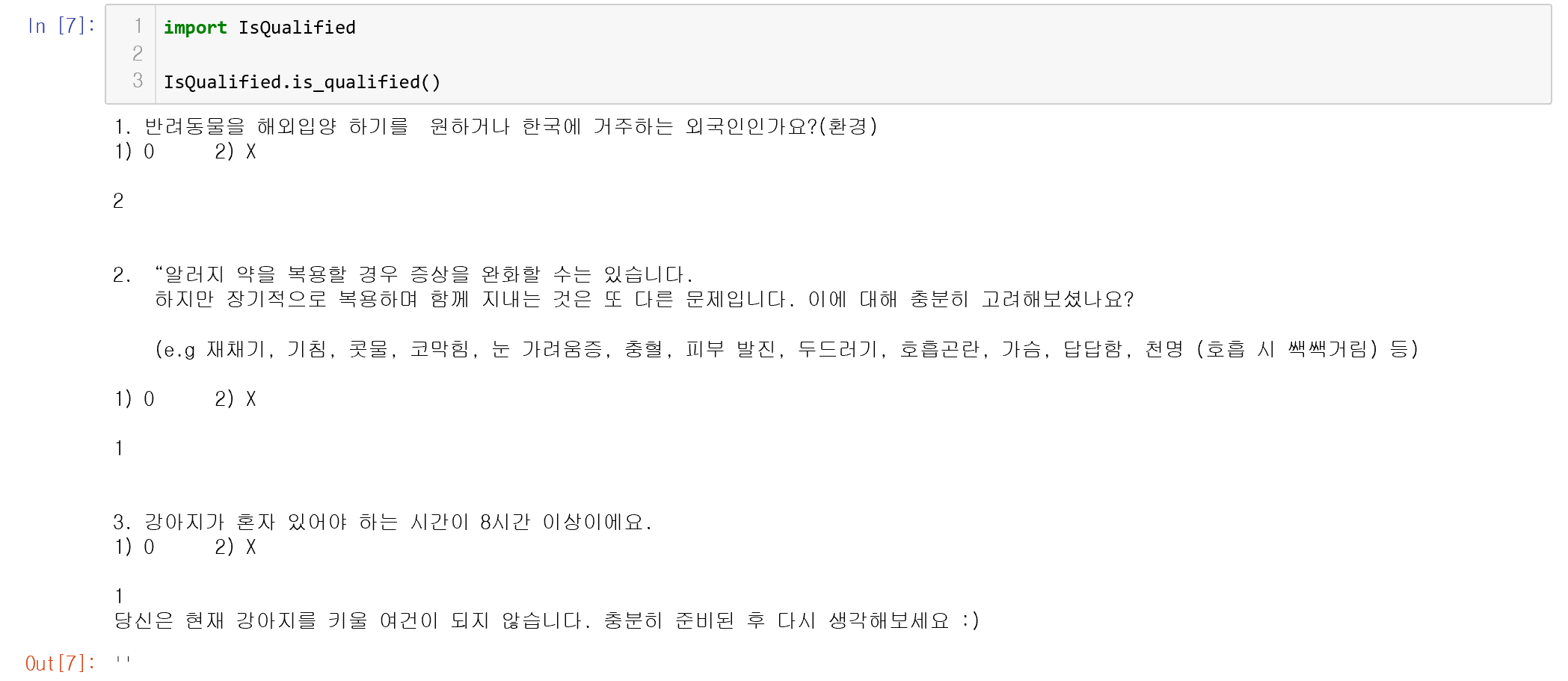
따라서 사용자 A가 선호하는 강아지는 14점을 기록한 강아지 1이다.

즉 (오차변환\*가중치) 합이 클수록 내가 선호하는 견종이다.

본 과정을 통해 도출해낸 점수로 예비 입양자와 가장 잘 맞는 견종들을 도출하고 최종적으로 추천한다.

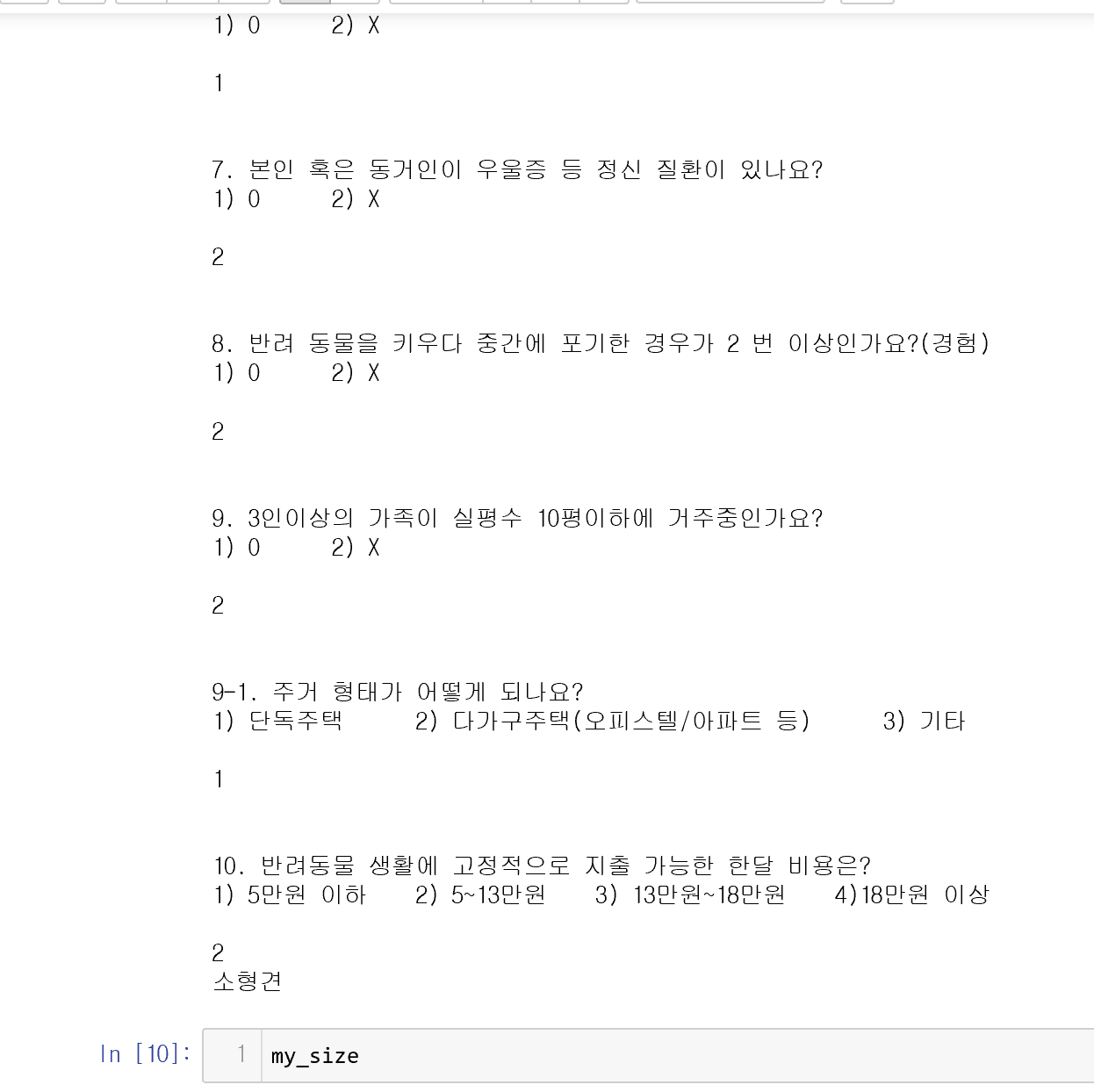
시간이 부족하여 위 코드를 충분히 구현해내지 못하였고, 미완성인 상태로 IsQualifiedWithData.ipynb 코드를 첨부하여 아쉬움이 남는다.

3.3. 3단계 - 결과 출력



< 그림 #. 자격판별요건 프로세스에서 자격 미달 결과가 나온 경우 >

자격판별요건 프로세스를 진행한 결과 자격 미달 결과가 나온 경우에 대한 그림이다. 보기(선지) 아래에 있는 숫자가 사용자가 입력한 숫자이다. 3번 문항에서 1번을 선택하여 강아지가 집에 혼자 있어야 하는 시간이 8시간 이상이라고 답했기 때문에 자격 미달이 된 것이다.



< 그림 #. 자격판별요건 프로세스 결과로 소형견이 나온 경우 >

1번부터 6번까지의 문항과 답에 대한 결과화면은 생략되었다. 9-1번 주거 형태는 단독주택이라 소형, 중형, 대형 모든 견종을 입양할 수 있지만 10번 반려동물 한달 양육비용으로 5~13만원을 지출 가능하다고 답했으므로 소형견이 도출되었다.

Ⅳ. 결론

4.1. 분석 결과 요약 및 인사이트 발굴

위의 프로세스를 위해 현재 접근 가능한 견종들에 대한 데이터를 크롤링 하였고 이를 csv와 같이 데이터 마이닝 툴에서 사용 가능한 방식으로 정제하고 이를 다각도의 관점에서 분석해봄으로써 중요한 의미를 도출하기 위해 노력해보았다. 견 종 별 성격 키워드라는 텍스트 데이터를 가공해봄으로써 이들이 데이터 속에서 중복이 관측되는 것을 확인, 이로부터 견 종 별 성격이라는 추상적인 개념에 대해 데이터가 일관된 정의를 유지하고 있음을 파악하였다. 이에 근거해 견 종 별 성격 키워드를 이용해 ‘성격’이라는 관념을 기준으로 견종들을 클러스터링 할 수 있음을 결론 지을 수 있었다. 하지만 ‘견 종 별 특성’을 이용해 사용자에게 추천할 견 종의 풀을 줄이고 해당 견종 중 가장 적합한 것으로 사료되는 견 종과 유사한 성격을 지닌 견종을 추천해준다는 당초의 목적을 위해서는 ‘성격’을 정의하는 키워드들의 의미를 반영한 클러스터링을 구현할 수 있는 알고리즘이 필요로 되어졌다. 하지만 의미(Semantics)를 반영한 클러스터링 알고리즘에는 현재 해당 분야 내에서도 어려운 것으로 사료되어졌고 우리는 우리가 실현할 수 있는 알고리즘을 찾기 위해 많은 노력을 하였다. 그 결과로 Robust Semantic Similarity Measuring 알고리즘을 찾을 수 있었으나 이 프로젝트를 진행하는 우리가 이를 시행하기 위해 필요로 되는 수준을 만족하지 못하였기에 아쉽게도 완전한 결말을 보지 못했다. 그럼에도 불구하고 성격 데이터를 비롯해 견 종 별 특성 데이터를 가공하고 분석하는 것을 통해 견 종 별로 성격이라는 키워드를 통해 개개별 견종으로 구분되고 또 이들이 군집화 되어질 수 있는 부류의 데이터임을 확인했다는 것이 해당 분석 프로세스를 통해 발굴한 인사이트이고 또 결과라고 할 수 있다.

4.2. Lessons Learned

본 문서를 통해 BABE(Best Adoption, Best for Each Other)라는 서비스에 대한 개괄적인 보고를 완료하였다. 처음 시작이 반려동물과 사람이 모두 함께 할 수 있는 좋은 사회를 다 같이 진심으로 그린 것으로부터 였기에 부족한 과정에도 불구하고 다섯 사람 모두 즐겁고 열정적으로 임할 수 있었다고 생각한다. 우리는 파양과 유기와 같이 반려동물 뿐만 아니라 함께 살아가는 많은 사람들을 실망시키는 현상이 발생하는 이유로 “반려동물에 대한 사람들의 몰이해”를 지목하였고 이를 해결하기 위해 많은 시간을 투자해 고려해야할 사항들을 고르고, 동시에 또 사람들이 더 잘 체계화된 사전 정보를 바탕으로 반려동물을 만나 함께 오래 행복하게 살 수 있도록 하기 위해 프로젝트를 진행해보았다. 이를 위해 다양한 견종들의 특성 중 사람과 함께 살 게 될 때 고려되어야 할 특성들을 구분하기 위해 많은 노력을 하였고 이렇게 구분된 특성들을 데이터화한 것을 바탕으로 비단 직감 혹은 감정적인 판단이 아닌 데이터마이닝이라는 프로세스를 통해 최대한 정확한 결과를 도출해내기 위해 노력하였다. 클러스터링 과정에서 데이터 마이닝 툴을 이용한 실질적인 입증과 관련한 부분에서 당초 계획했던 것에 미치지 못했다는 것이 애석한 부분이나 시간이 조금 더 주어진다면 구현할 수 있는 방안을 찾을 수 있다고 자신할 수 있는 수준만큼 구체적인 알고리즘을 찾아 이해하고 이를 적용할 방안을 생각해낼 수 있었기에 한 학기에 걸친 이번 프로젝트가 충분히 가치있고 유의미했다고 말할 수 있다.

Ⅴ. 프로젝트 관리

5.1. 팀원의 구성 및 역할

|  |  |
| --- | --- |
| 이름(가나다순) | 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 |

5.2. 프로젝트 일정

프로젝트 일정은 중간 발표 당시에 정하였던 일정과 크게 차이가 나지 않도록 진행하였다. 하지만 마지막 발표 일정이 수정되며 보고서 분석 및 피피티에 할애한 일정을 시험이 끝난 후로 미루었다. 수정되어 실시된 프로젝트 일정은 다음과 같다.

|  |  |  |  |
| --- | --- | --- | --- |
|  | 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 |

5.3. 프로젝트 도중 발생한 이슈

반려 동물의 유기 이유에 대한 자료 조사에서 어려움을 겼었다. 사람에 대한 입양이나 파양, 또는 영아 유기에 대한 국내외 자료는 많았지만 반려동물에 대한 유기 사유를 정리한 자료를 찾기가 쉽지 않았기 때문이다. 특히 유기동물 수와 같이 정량적 데이터는 찾을 수 있었지만, 유기 사유, 즉 정성적 데이터는 수집하기 어려웠다.

< 3.2.1. 견종 성격 선호 테스트 >에서 언급한 알고리즘 구현 상의 어려움이 있었다.

< 3.2.2. 견종 특징 선호 테스트 >에서 가중치를 몇으로 부여해야 할지에 대한 고민이 많았다. 사용자가 민감해하는 특성에 가중치를 부여하여 오차합이 같은 경우에도 사용자가 더 선호하는 견종을 추천할 수 있으면서도, 너무 큰 가중치 때문에 사용자와 견종 특성의 오차가 작은 경우가 무시되지 않도록 해야 했기 때문이다. 여러 번의 실험을 거쳐 5점 척도를 역순으로 한 수를 가중치로 두는 것이 적절하다고 판단하였고 이 가중치를 알고리즘에 반영하였다.

5.4. 첨부파일

아래 표와 같이 파일을 첨부하였다.

|  |  |
| --- | --- |
| 코드 | 데이터 파일 |
| dog\_cares.html  dog\_cares.ipynb | dog\_cares.csv |
| dog\_features\_general.html  dog\_features\_general.ipynb | dog\_features\_edited\_size.csv |
| dog\_features\_percent.html  dog\_features\_percent.ipynb | dog\_features\_percent.csv |
| IsQualified.py |  |
| IsQualifiedWithData.ipynb |  |

IsQualifiedWithData.ipynb 코드에서는 IsQualified.py 를 import하여 그 안에 정의된 함수를 사용한다. IsQualified.py 코드에는 자격요건판별, 선호도 test에 대한 질문지가 있다.

Ⅵ. 참고자료

* Herwijnen IRv, van der Borg JAM, Naguib M, Beerda B (2018) Dog ownership satisfaction determinants in the owner-dog relationship and the dog’s behaviour. PLoS ONE 13(9)
* 2018 (사)동물권행동 카라 정기총회 보고서
* (사)동물권행동 카라 입양신청서
* 보듬컴퍼니 강형욱 칼럼 “[분리불안] 애견숍 강아지가 분리불안증이 더 심한 이유”
* 웹진 ‘나의 반려동물(https://myanimals.co.kr/)’
* 2017 반려동물 양육 실태 조사, KB연구보고서
* ‘동물자유연대’ 입양신청기준
* 강서구 도그플러스 동물병원 16년도 반려동물 표준화 미용 비용표
* Danushka Bollegala, Yutaka Matsuo, Mitsuru Ishizuka (2007), “Measuring Semantic Similarity between Words Using Web Search Engines”, WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.

1. 국내 반려동물 사육 인구? ‘574만 가구 1481만명’, 데일리벳, 2018년3월28일, <http://www.dailyvet.co.kr/news/industry/92406> [↑](#footnote-ref-1)
2. 유기.유실 동물 10만 넘었다. 한겨레, 2018년7월13일, <http://www.hani.co.kr/arti/animalpeople/human_animal/851031.html> [↑](#footnote-ref-2)
3. 한해 반려동물 8만 마리 버려져...관리비용만 100억원, 조선일보, 2016년5월15일, <http://news.chosun.com/site/data/html_dir/2016/05/15/2016051501868.html> [↑](#footnote-ref-3)
4. 반려동물 양육비, 한달 평균 ‘13만3000원’, 매일경제, 2018년4월5일, <http://news.mk.co.kr/newsRead.php?no=217103&year=2018> [↑](#footnote-ref-4)
5. ## Anthony C. Little, D. Michael Burt, David I. Perrett, Assortative mating for perceived facial personality traits, [Personality and Individual Differences](https://www.sciencedirect.com/science/journal/01918869) [Volume 40, Issue 5](https://www.sciencedirect.com/science/journal/01918869/40/5), April 2006, Pages 973-984

   [↑](#footnote-ref-5)