

Project Intermediate Report

201901621 공진성

- 전체 프로젝트 주제

프로젝트 주제는“야생동물의 유해야생동물 지정에 대한 가능성 예측”입니다.
프로젝트를 진행하면서, 야생동물들의 특성을 파악하여, 유해야생동물의 가능성이 높은 동물을 먼저 예측할 수 있다면, 그로 인해 해당 동물을 반려동물로 키우는 사람들에게 주의를 줄 수 있다면, 이로 인해 먹이사슬의 붕괴를 막을 수 있지 않을까 생각되어 위 주제를 선정 하였습니다.

- 데이터셋 설명

두 가지의 데이터셋이 존재합니다.
첫 번째로, 이미 선정된 유해 동물 데이터셋(정답 데이터셋)으로 국문 / 영문 / 학명 / 관리현황 / 분류군 / 학명-수정(excel 내장 함수로 학명을 검색하기 좋게 재명명) / 특성의 칼럼값을 가지는 612종의 유해 동식물 데이터가 있습니다. 프로젝트는 유해 동물 예측이지만, 정답 데이터를 유해 동식물로 지정한 것은, 유해 식물의 특성에서도 충분히 유해 동물을 예측할 만한 특성이 있다고 생각해, 이 데이터셋에서는 식물을 포함하였습니다. 위의 식물에서의 가설을 세운 근거는 황금 아카시아와 유럽독미나리의 특성에서 있었습니다. 아래는 위 식물들의 특성 중 일부입니다.

황금 아카시아(*Acacia saligna*) : At the base of each phyllode is a nectary gland, which secretes a sugary fluid. This attracts ants, which are believed to reduce the numbers of leaf-eating insects.

유럽독미나리(*Aethusa cynapium*): Poisoning from fool's parsley results in symptoms of heat in the mouth and throat and a post-mortem examination has shown redness of the lining membrane of the gullet and windpipe and slight congestion of the duodenum and stomach. Some toxins are destroyed by drying, and indeed, hay containing the plant is not poisonous.

이와 같이, 식물이라도 충분히 유해동물의 특성을 지닐 수 있다고 생각해 데이터셋에 포함했습니다.

두 번째로, 한국 야생동물 데이터셋(예측 데이터셋)으로 관리분류군/ Species / Class 등 종 분류 / 학명 / 대표국명 / 명명자 / 명명년도 / 특성의 칼럼을 가지고 있는 12,667종의 동물들이 있습니다. 12,667종 생물들의 자세한 정보는 다음과 같습니다.


포유류	조류	파충류	양서류	어류	미삭동물	무척추동물
125	550	32	28	1,339	134	10,459

하지만 12,667종에 대해 Wikipedia(ko, en) 데이터 크롤링 결과 8,406의 종들의 데이터가 없다는 결괏값이 도출되었습니다.

CSV 파일에서 무척추동물 10,459종 중, 7,882종의 데이터가 없다고 확인하였고, 결국 유해야생동물을 예측하는데 무척추동물을 빼는 게 더 확실한 결과가 나올 거 같다고 판단하였습니다.

또한, 무척추동물을 제외한 2,208의 종 데이터 중에서도, 524종의 특성에서 Null 값이 추출되었는데, 결국 Wikipedia에 데이터가 없는 생물이라면, 그 생물은 관심도가 떨어진다고 판단하였고, “그렇다면 데이터가 없는(정보가 없는) 동물이 유해 동물이 될 가능성도 현저히 낮지 않을까?”라는 가설을 세워 1,684종의 동물에 대해서만 예측을 진행하고자 합니다.

- Feedback

채점자	 최대진
피드백	<p>— 일단, 유해가능성을 예측하려면 유해가능성에 대한 확률, 유해종인자의 여부 등의 "정답데이터"가 필요한데, 작성된 내용으로는 정답데이터가 있는지 여부를 알 수 없음. 정답 데이터가 없다면 애초에 연구를 수행하기 어려울 것으로 보임. "유해동물로 지정될 수 있는 가능성 높은 종의 데이터는 12,667개 존재합니다." 라고 되어있는데, 어떤 기준으로 가능성이 나와있으며, 확률을 표현한 것인지 등에 대한 자세한 조사를 해야할 것으로 보임. - 동물들의 어떤 특성을 수집해야 유해가능성이 예측될 것인지에 대한 고찰을 더 해서 가설을 세우는 것이 중요함. 단순히 특성을 가져오겠다고 되어있는데, 아무 가설없이 모든 특성을 다 가져오는 것은 노력이 많이 들어갈 뿐더러, 결과가 제대로 나오지 않거나 의미가 없는 결과가 도출되는 경우로 귀결되는 경우가 많음.</p>

현재 정답으로써 사용될 선정된 유해 동물들(유입주의 생물, 외래생물, 생태계 교란 생물, 생태계위해우려생물)의 데이터는 한국 외래생물 정보 시스템, 환경부에서 xlsx 파일로 작성된 유해 동물들의 국명, 학명 등을 받아왔습니다. 그 후, Wikipedia를 통한 크롤링으로 현재 유해 동물 612종 중 49종을 제외한 유해 동물 특성까지 CSV 파일에 입력된 상태입니다. 그러므로 정답 데이터는 문제없이 구했다고 생각합니다. 49종의 데이터에 대해서는 검색하여 직접 채워 넣었습니다.

또한, "유해 동물로 지정될 수 있는 가능성 높은 종의 데이터는 12,667개 존재합니다."라는 문장은 단어 선택이 잘못되어 교수님께 잘못된 전달해 드린 것 같습니다. 위처럼 말했던 근거는, 주제 자체가 “유해 동물”이었기에 50,000여 가지의 우리나라 생물 중 식물류 / 미생물 등 동물로 분류하기 어려운 생물을 제외한 개수를 도출했을 때, 12,667종이었습니다. 그렇기에 유해 동물로 지정될 가능성 있는 종은 12,667종이라 말씀드렸습니다.

마지막으로, 모든 특성을 다 가져오는 것에 대해서는 다음과 같이 생각합니다.

차원의 저주로 인해 정확한 특성의 데이터들을 가져와야 정확한 학습이 가능한 것을 인지하고 있지만, 크롤링 환경에서 12,667종의 동물 세부 특성을 분류해 데이터를 받아오는 것이 불가능하다고 판단했습니다.

하지만, 데이터양을 최대한 늘리는 것도 차원의 저주를 해결하는 방법으로 알고 있어, 특성의 세부 사항은 아니지만 특성 데이터양 늘려 종마다 많은 글자를 받아와서 분류한다면, 의미 있는 결과를 도출할 수 있을 것이라 예상됩니다.

- Timeline

[illegible]

- ① : 유해동물(생태계 교란종) 특성 종합 및 전처리
- ② : 야생동물의 특성 크롤링
- ③ : 야생동물의 특성 토큰화 및 정제
- ④ : 야생동물의 특성 정규화 및 태깅
- ⑤ : 야생동물과 유해동물간의 유사도 측정 및 표현

- 현재까지 분석된 내용

위에 쓰여 있는 Timeline과 같이 ①, ②항목을 계획대로 완료하였습니다.

Proposal에서는 국립생물자원관에서 크롤링을 진행한다고 하였지만, 국립생물자원관에서 크롤링을 진행하기엔 Wikipedia보다 값이 많다고 생각하지 않았기에 Wikipedia에서 크롤링을 진행하였습니다.

유해 동물 종류는 생각했던 것보다 매우 많아서 다행이었다고 생각합니다. 600종으로 학습 시킨다고 학습이 잘 될지는 모르겠지만, 유해 동물들의 특성 중에 α -Latrotoxin(신경독의 한 종류)과 같은 여러 상세한 특성들까지 포함되어 있어, 어느 정도는 학습이 되리라고 생각합니다.

그리고 우리나라 야생동물을 크롤링하며 1,684종의 특성을 이용하기 쉽게 CSV 파일에 넣는 것에 성공했습니다.

- 앞으로 수행할 내용

현재 우리나라 야생동물들을 크롤링한 결과는 영문 Wikipedia에서 먼저 크롤링해 왔고, 만약 영문 wiki에서 찾지 못한 값이라면 한국 wiki에서 찾도록 설계하여 특성값을 부여했습니다. 하지만 유해 동물 데이터 셋의 특성도 영어로 표기되어 있어, 한국 wiki에서 찾은 야생동물의 특성값을 영문으로 바꾸는 데이터 전처리 작업이 우선일 거 같습니다.

그 후, 야생동물의 특성을 정제 및 토큰화, 정규화를 진행하여 여러 문서 분류 방식을 사용해 가장 효율적인 문서분류 방법을 찾아보려고 합니다. 그 후, 학습시킨 유해 동물 데이터 셋과 비교하여 유사도를 측정해 예측한 결과값을 프로젝트 결과물로 제출하려 합니다.

해당 프로젝트 Github(<https://github.com/KKardd/prediction-of-hazardous-animals>)

한국 외래정보 시스템(<https://kias.nie.re.kr/home/main/main.do>)

환경부(<https://www.me.go.kr/home/web/main.do>)