빅데이터 입문 Project Proposal

201901621 공진성

– 프로젝트 주제 및 분석하고자 하는 Target Question

프로젝트 주제는"야생동물의 유해야생동물 지정에 대한 가능성 예측"입니다. Target Question을 위 주제로 생각하게 된 계기는 다음과 같습니다.
최근 유튜브에서 시청했던 영상 중, 늑대거북이 생태계교란생물로 지정되었단 내용의 영상이 있었습니다. 영상을 시청하기 전부터, 늑대거북은 매력적인 등딱지를 가지고 있어 소비자들에게 인기를 많이 끌었단 사실을 알고 있었고, 이런 늑대거북의 생태계교란종 지정 소식은 충격적이었습니다. 지정된 이유는 먹성이 좋은 늑대거북을 반려동물로 키우던 사람들이 유기했기 때문이었습니다. 늑대거북은 물 안과 밖에서 모두 생활할 수 있었고, 물고기뿐 아닌 새, 작은 포유류 동물까지도 모두 먹어 치우는 포식자였습니다. 이런 뭐든 먹어버리는 늑대거북 때문에 먹이사슬이 무너지기 시작하였고, 심지어 생명력이 질기기에, 개체 수가 점점 증가하였습니다. 결국 환경부에서는 생태계 교란, 즉 유해야생동물로 지정하였습니다.

제가 이번 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 선정하였습니다.

– (오픈 데이터 사용의 경우) 사용할 데이터셋 종류, 형태, 및 크기 (volume)

오픈 데이터셋으로 전국적으로 야생동물에 대한 정보가 있는 오픈 데이터는 구하였습니다. 데이터는 csv 파일이였고, 안에는 Phylum, Class, Order등 학명(생물 명)의 데이터가 58,050개 존재하지만, 포유류, 조류, 파충류 등 유해동물로 지정될 수 있는 가능성 높은 종의 데이터는 12,667개 존재합니다.

자세한 정보는 다음과 같습니다.
포유류(125), 조류(550), 파충류(32), 양서류(28), 어류(1339), 미삭동물(123), 무척추동물(10,459)

– (데이터 직접 수집의 경우) Target 서비스, 수집할 데이터의 종류, 양 (가능한 자세하게)

유해 동물을 예측하는 데에 있어, 오픈 데이터셋만 이용할 수 없습니다. 오픈 데이터셋은 현재 학명, 국명, 대표 국명, 명명자, 명명년도등의 데이터만 가지고 있습니다. 유해야생동물을 예측하려면 각각의 동물들 특성들이 필요하므로, 오픈 데이터셋의 학명들을 이용하여 국립생물자원관(https://species.nibr.go.kr/)에서 크롤링함으로써 특성을 수집해 올 예정입니다.

– 대략적인 프로젝트 진행 Timeline

이 프로젝트를 진행하려면, 유해 동물(생태계교란종)의 특성들을 종합하여 데이터를 확보하고, 야생동물들의 특성들을 크롤링해 와야 하며, 유해 동물 특성과 야생동물의 특성을 비교하여 유사도 측정을 진행해야 합니다. 대략적인 Timeline은 다음과 같습니다.

	4주차	5주차	6주차	7주차	8주차	9주차	10주차	11주차	12주차	13주차	14주차	15주차
①												
②												
③												
④												
⑤												

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

– 예상되는 어려운 점 및 해결방법

야생동물 크롤링에서의 걱정이 많습니다. 결국 12,667종의 데이터를 모두 수집해 와야 하는데, 그 과정도 어려우리라 생각되고, 간혹 국립생물자원관에도 학명만 기재되어 있고 특성란이 공백인 동물들도 있기 때문입니다.

특성이 기재되어 있지 않은 동물의 데이터는 직접 손으로 검색하여 가져오는 방법으로 해결하려 합니다. 마땅히 데이터가 뭉쳐있는 공간이 있다면 거기도 크롤링해서 오면 되지만, 현재까지 찾은 바로는 국립생물 자원관 외의 뭉쳐있는 공간은 확인하지 못하였습니다.