10.16 NLP 실습 : NLP 모델로 위키레이스 프로그램 구현하기

수강생 이름 : 곽태경

구현한 프로그램 : selenium, kcbert 모델을 이용한 자동 위키레이스 프로그램

위키레이스

위키백과, 나무위키 등의 위키에서, 특정 문서에서 출발하여 링크 텍스트만 타고 넘어가는 식으로 최대한 빠르게 목표한 문서까지 도달하는 게임을 말합니다. 이 게임을 프로그램으로 구현한다면 특정 문서로 이동하는 과정을 자동화할 수 있습니다.

구현한 프로그램 설명

한국어 위키백과에서 시작 문서 → 목표 문서까지 링크 텍스트만으로 자동으로 도달하는 프로그램입니다.

selenium 라이브러리를 이용하여 자동으로 진행되도록 하였고, 사전 학습된 kcbert 모델을 사용하여 목표 문서의 제목과 링크 텍스트들의 코사인 유사도를 체크하여 유사도가 가장 높은 링크 텍스트를 타고 다음 문서로 넘어갑니다. 이 과정을 목표 문서에 도달할 때까지 반복합니다.

처음에는 본문 div 태그를 가져와서, 그 속에 있는 a 태그들 중에서 목표 문서의 제목과 가장 유사도가 높은 태그를 클릭하도록 했습니다.

(이 과정에서, 최소 유사도 기준 없이 가장 높은 것을 선택합니다.)

여기서, a 태그의 텍스트가 비었거나, 연결될 문서 주소가 비어있는 경우는 제외했습니다.

그리고 페이지 내에서 유효한 링크만을 선택하도록 여러 필터링 조건을 추가했습니다. 예를 들어, "위키미디어 공용"과 같은 한국 위키백과가 아닌 사이트 링크나, 각주 링크, 편집 링크는 제외하고, 방문한 적이 있거나 없는 문서로 표시된 링크도 제외했습니다.

그리고 이미 방문한 페이지는 다시는 방문하지 않도록 했고, 만약 정상적으로 링크 텍스트를 클릭해 들어갔지만 내용의 길이가 100자가 넘지 않는 경우에는 뒤로가기를 하고, 다음으로 유사도가 높은 태그를 클릭하도록 했습니다.

그리고 검색 깊이 제한을 50으로 두었습니다. 페이지를 이동하거나, 적절한 페이지를 찾지 못하여 다시 시도하는 과정에서 깊이가 1씩 증가합니다. 깊이가 50이 될때까지 계속 실행되고, 50이 되면 자동으로 도달 실패 처리됩니다.

링크를 이동할 때마다 이동한 링크, 이동한 링크가 담긴 텍스트의 유사도를 출력하고, 만약 짧은 문서를 무시하고 뒤로가기 했을 때에도 무시한 링크가 출력됩니다.

만약 유사도가 1.0이 되어 도달에 성공했다면, 목표 페이지에 도달했다는 출력이 나오고, 프로세스가 종료됩니다.

테스트

목표한 문서로 도달하는 것 자체가 어렵다고 판단하여, 성공 여부를 체크했습니다.

시작 문서는 '파이썬' 문서로 했습니다.

테스트한 목표 문서들입니다.

1. 특수_상대성이론 : 40번의 이동 후 성공
2. 커피 : 7번의 이동 후 성공
3. 거짓말 : 실패(깊이 제한 도달)
4. 이재용 : 9번의 이동 후 성공
5. 안젤리나_졸리 : 31번의 이동 후 성공
6. 제4천년기(4000년) : 실패(깊이 제한 도달)
7. 독도 : 24번의 이동 후 성공

성공률 : 71.43%

성공 시 평균 이동 횟수 : 22.2번

생각보다 높은 성공률을 보여, 제4천년기(4000년) 같은 연속적인 숫자 문서들을 제외하면 도달 성공을 기대할 수 있을 것으로 보입니다.

결론

이번 프로젝트를 통해 NLP 모델을 활용한 위키레이스 프로그램을 구현해보았습니다. 자연어 처리 기술을 활용해 링크 텍스트와 목표 문서 제목의 유사도를 계산하고, 이를 바탕으로 목표 페이지까지 경로를 찾아가는 기능을 성공적으로 구현할 수 있었습니다. 또한, 여러가지 필터링을 통해 원활한 페이지 이동이 가능하도록 했습니다.

이번 실습을 통해 자연어 처리 기술의 활용 가능성을 다시 한 번 확인할 수 있었고, 프로그램의 성능을 더욱 향상시키기 위한 여러 가지 최적화 방안을 생각해 볼 수 있었습니다.

개선 방안

1. 내용 전체를 살펴보기 때문에 링크 텍스트 목록을 만들 때 중복된 링크, 텍스트는 다르지만 똑같은 문서로 가는 링크들은 제외하는 과정을 추가하면 더 빠른 결과가 나올 것으로 예상됩니다.
2. 숫자와 관련된 문서에서는 숫자 관련 문서에서 벗어나기 어려워하는 모습을 보였습니다. 숫자 관련 문서를 잘 처리할 수 있다면 더 빠른 결과를 얻을 수 있을 것으로 보입니다.