

국제/정치 뉴스 요약 서비스 최종 보고서

영어분석을위한 기계학습
202102045 국제학 전공
양유민

1. 프로젝트 개요

1.1 프로젝트 주제

"국제/정치 뉴스 및 공공기관(외교부) 소식 자동 요약 모델"

1.2 문제 정의

- 정보 과잉: 매일 쏟아지는 국제 정세 뉴스와 외교부의 공지사항을 개인이 모두 읽고 파악하기에는 시간이 부족함.
- 접근성 문제: 긴 줄글 형태의 블로그 포스팅이나 보도자료는 핵심 내용을 빠르게 파악하기 어려움.
- 해결 목표: 사용자가 URL만 입력하거나 버튼 한 번만 누르면, AI가 핵심 내용을 3줄로 요약하여 제공함으로써 정보 습득 효율을 극대화함.

2. 진행 과정

2.1 데이터 수집 및 분석 (Assignment 4)

- 초기 시도: 네이버 뉴스 웹 크롤링을 시도했으나 접근 차단 및 데이터 품질 문제 발생.
- 전략 변경: 허깅페이스의 오픈소스 데이터셋(daekeun-ml/naver-news-summarization-ko) 활용.
- EDA 및 전처리:
 - 데이터 22,194건 확보.
 - 분석 결과, 요약문이 본문보다 긴 이상치(Outlier) 1,739건과 초단문 데이터를 발견하여 전량 제거함.

2.2 모델 학습 및 평가 (Assignment 5)

- 모델 선정: 한국어 문서 요약에 특화된 Pre-trained 모델인 KoBART를 선정.
- 학습(Training):
 - Train/Val/Test (8:1:1) 분할.
 - 3 Epoch 학습 수행 (Training Loss의 지속적 감소 확인).

- **평가(Evaluation):**
 - Test Set 기준 **ROUGE-1 58.49%, ROUGE-L 55.39%** 달성.
 - 정성 평가를 통해 문장 끊김 현상과 반복 생성 문제를 발견하고 파라미터 튜닝으로 개선함.

3. 서비스 개발 (Assignment 6)

3.1 서비스 구조 (Architecture)

학습된 모델을 실제 사용 가능한 웹 애플리케이션 (**Web App**) 형태로 구현하였습니다.

<https://assignment6-deploy-fqz3medkju8xgtr8fxbmdu.streamlit.app/>

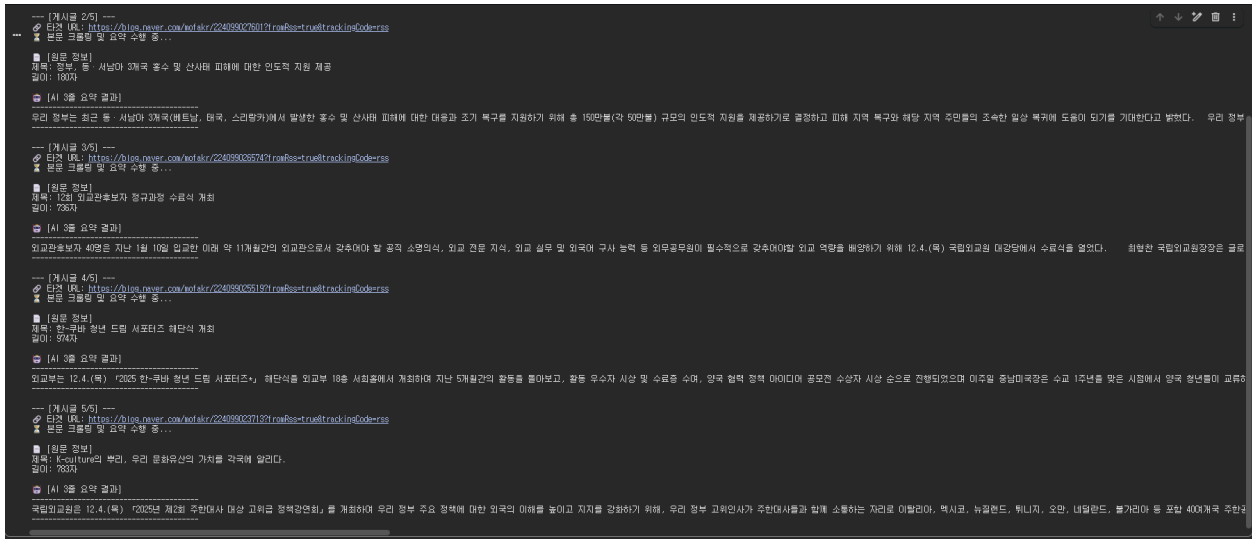
- **Frontend:** Streamlit 프레임워크를 사용하여 직관적인 GUI(Tab 방식) 구현.
- **Backend Pipeline:**
 1. **Input:** 기사 원문 텍스트 입력 또는 네이버 외교부 블로그 RSS 자동 감지.
 2. **Crawler:** BeautifulSoup을 이용해 네이버 블로그 Iframe 내부 본문 추출.
 3. **Inference:** 실제 웹 서비스 구현(Streamlit Cloud 배포) 단계에서는 서버의 저장소 용량 제한 및 메모리 효율성을 고려해야 했음. 이에 따라, 대용량의 커스텀 모델 가중치 파일(`final_model`)을 직접 업로드하는 대신, 동일한 **KoBART** 아키텍처를 기반으로 요약 성능이 검증된 공개 모델([gogamza/kobart-summarization](#))을 API 형태로 호출하여 시스템의 경량화와 안정성을 확보했음.
 4. **Output:** 요약된 텍스트를 화면에 출력.

3.2 핵심 기능

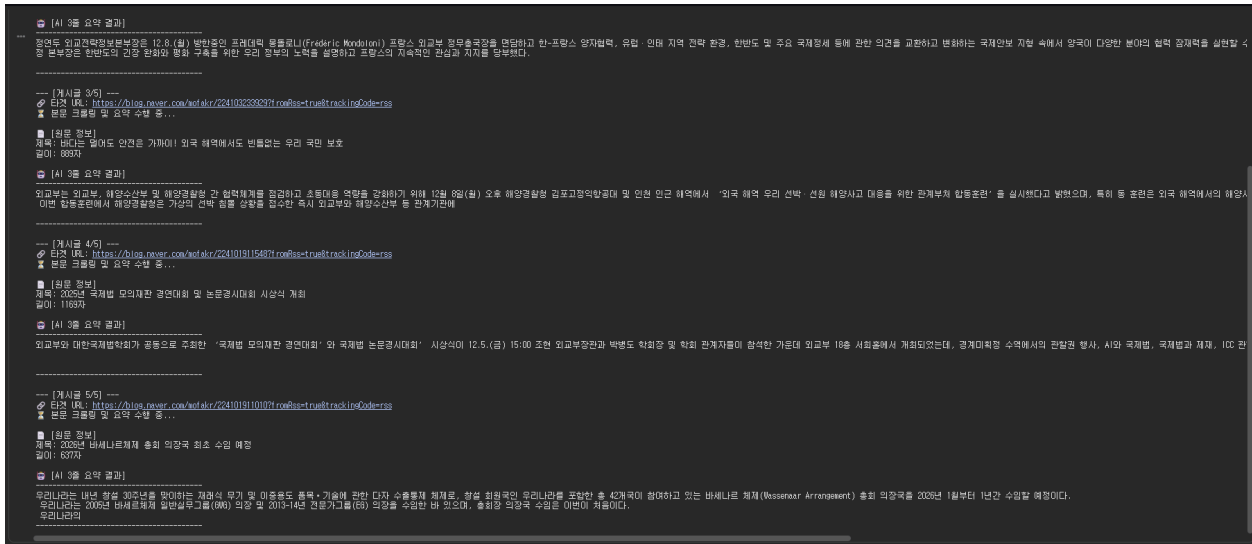
1. 직접 입력: 사용자가 읽고 싶은 기사 원문을 긁어 넣으면 즉시 요약.
2. 외교부 소식통: 버튼 클릭 시 외교부 공식 블로그 RSS를 스캔하여 최신 '외교부 소식' 게시글들을 자동으로 가져와 요약 브리핑 제공.

4. 실제 사용 결과 (Usage Log)

1. 2025년 12월 7일



2. 2025년 12월 9일



3. 2025년 12월 10일

📌 [시 3종 요약 결과]

12월 9 외교부와 해양수산부는 미국 뉴욕 UN 해양수산부에서 열린 UN 총회의 결의로 2028년 6월 제4차 UN해양총회(United Nations Ocean Conference)의 대한민국 개최가 확정되었다고 밝혔는데, 이는 “UN 지속가능발전목표 14번(해양환경, 해양자원의 보전 및 지속가능한 활용)”의 이행을 위해 3년 주기로 개최되는 해양

--- [개시글 3/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410447674271rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 조현 장관, 미국 외교정책위원회(NCAPP) 대표단 집견

길이: 1043자

📌 [시 3종 요약 결과]

조현 외교 장관은 12.9.(화) 방한 중인 미국 외교정책위원회(NCAPP: National Committee on American Foreign Policy) 대표단을 만난 집견하고, 한미동맹, 한반도 및 지역 질서 등에 대해 의견을 교환하고, 금번하는 안보, 경제 환경에 발맞추어 한미동맹을 더 호혜적이고 다원지향적으로 발전시키기 위해 더욱 거

--- [개시글 4/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410447674271rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 『제6차 한-미행정당서국 외교장관회의』 개최 양자회담 개최 결과

길이: 1212자

📌 [시 3종 요약 결과]

조현 장관은 장관은 『피터 샤넬 아고베카 (Peter Shanell Agosaka)』, 솔로몬제도 외교장관과의 면담에서 솔 한 제 솔로몬제도가 태평양도서국포함(PIF)의장국으로서 역할을 해 온 것을 평가하고, 내년 양국 수교 35주년을 앞두고 재반 분야 협력을 더욱 강화하기로 희망한다고 하였으며, 『칼라니 카타리(Kalani

--- [개시글 5/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410447674271rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 김건희 2차관 - 고지 외교부 부장관 면담(12.9.)

길이: 661자

📌 [시 3종 요약 결과]

김건희 외교부 2차관을 12.9.(화) 오후 제6차 한-미행정당서국 외교장관회의 참석차 방한 중인 렌노라 게레레타부아(Lenora Geraereta) 피지 외교부 부장관과 면담을 갖고 양국 관계, 경제협력, 개발협력, 기후변화 등 국제무대 협력 관련 의견을 교환하고 양국을 지난 한 해가 이상 우호협력관계를 발전시켜

4. 2025년 12월 11일

--- [개시글 2/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410570089571rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 신규 『아세안 연계성 전략계획(2026-2035)』 下, 한-아세안 협력 방향 논의

길이: 1801자

📌 [시 3종 요약 결과]

이기가 외교부 아세안국장은 12.10.(수) 한-아세안엑스포가 주최하고 외교부가 후원하는 『제13차 아세 안 연계성 포럼』에 참석해 한국의 역할과 기여 방안에 대해 설명하고, 아세안의 지속가능한 도시 조성을 지원하기 위한 협력기금 사업으로, 역내 통합과 지속 가능한 성장을 목표로 하고 있다고 전했다. 이 국장은 2009년부터 2025년까지 아세안이 회원국 간의 “지속가능한 인프라, 소프트웨어산업 혁신, “디지털 혁신 등 다양한

--- [개시글 3/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410570089571rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 일본 10건 발생 관련 자위대인 보호를 위한 분부-공관 합동 상황점검회의 개최

길이: 431자

📌 [시 3종 요약 결과]

외교부는 최근 일본에서의 자위대 발생과 관련하여 12.10.(수) 을주석 영사안전국장 주재로 일본 내 10개 재외공관과 분부-공관 합동 상황점검회의를 개최하여 일본 동북부 지역에서 규모 7.5 지진이 발생하였으며, 일본 정부는 12.9.(화) ‘후발지진 주의 정보’를 발표하면서 향후 1주간 여진이나 대규모 후4 이후 관원에 이번 회의에서는 현지 우리 국민 안전 상황을 점검하고 합동 안전 대책을 논의했다.

한편 주전도이출형

--- [개시글 4/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410569954971rceBss+true&trackinCode=rs>

🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 외교 시 외부 전문가 자문단 발대식 개최

길이: 7193자

📌 [시 3종 요약 결과]

외교부는 조현 외교부장관 주재로 2025년 12월 10일(수), 『외교 시 외부 전문가 자문단 발대식』을 개최하고 국제 전문기공을 구성한 민간 협력체계를 본격 가동하여 올해부터 3개년 계획으로 추진 중인 외교 시 구축사업의 성공적인 이행을 위하여 전략적인 조언과 지원을 수행할 예정이다.

조 장관은 임명자간 전략적 시대를 맞아 참신기공을 적용한 외교 인프라의 혁신을 강조하며, 금번하는 국제 환경 속에서 전략적 외교 자문 확대를 위한 자문단의 적극적인 참여와 지원을 당부했다.

이후러, 외교부의 외교 통합 시 구축 추진

--- [개시글 5/5] ---

🔗 타겟 URL: <https://blog.naver.com/notakr/22410569867471rceBss+true&trackinCode=rs>

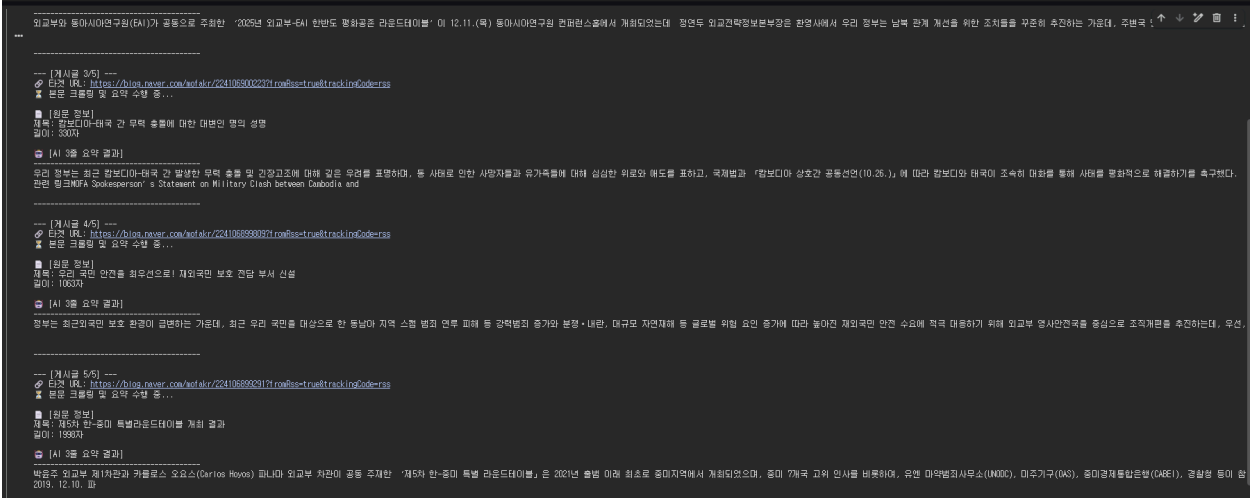
🔖 본문 크롤링 및 요약 수량 중...

■ [원문 정보]

제목: 태국-캄보디아 접경 지역 마약밀반 상황 점검

길이: 894자

5. 2025년 12월 12일



5. 배운 점 및 개선 방향

5.1 한계 및 개선 방향

- 생성 속도: CPU 환경에서 추론 시 약 3~5초 소요됨. GPU 서빙 또는 모델 경량화(Distillation) 필요.
- 문장 완결성: max_length를 늘렸으나 간혹 문장이 매끄럽지 않은 경우가 있어, 후처리 로직(Post-processing)을 더 정교하게 다듬을 필요가 있음.
- 크롤링 문제: Google Colab에서는 크롤링 문제가 발생하지 않았으나, Streamlit에서 서비스 구현 후 보안상의 문제로 크롤링이 되지 않는 문제가 발생하였음. 현재는 개선되었으나 온전한 서비스 구현에는 한계가 있음.



위의 사진에서 파악할 수 있듯이, 각기 다른 기사에 동일한 요약이 제공되고 있는 것을 확인할 수 있음.

5.2 배운 점

- 데이터의 중요성: 단순 크롤링보다 잘 정제된 데이터셋이 모델 성능에 훨씬 큰 영향을 미침을 확인함.
- 서비스화의 가치: 모델 성능이 아무리 좋아도(ROUGE 점수), 실제 사용자가 쓰기 편한 UI(Streamlit)와 자동화 기능(RSS)이 결합되어야 비로소 가치가 생김을 깨달음.