개인 맞춤형 뉴스레터 추천 프로그램 구현

김원종, 강규창

국립군산대학교

01055130143a@gmail.com, kc.kang@kunsan.ac.kr

Implementation of Personalized Newsletter Recommendation Program

Wonjong Kim, Kyuchang Kang

Kunsan National University

요 약

본 구현은 이용자가 원하는 카테고리의 뉴스레터를 간단하게 확인할 수 있는 프로그램 구현에 관한 것으로 정치, 경제, 사회, 문화, 과학 5개 분야의 최신 뉴스 제목을 실시간으로 제공하여 사용자가 관심 있는 정보를 빠르고 효율적으로 확인할 수 있도록 개발된 개인 맞춤형 뉴스레터 추천 프로그램이다. 이를 위해 Python 언어를 중심으로 C++과 Java를 함께 활용해 웹 크롤링, 데이터 정제, 출력까지 멀티 언어 통합 환경을 활용하여 구현하였다. 구현 프로그램은 정보 접근성을 높이고 사용자 편의성을 강화하는 것을 핵심 취지로 하며, 그래픽 사용자 인터페이스와 명령줄 인터페이스를 모두 지원해 다양한 환경에서 쉽게 사용할 수 있는 장점이 있다.

Ⅰ. 서 론

현대 사회에서 뉴스는 사회·경제·정치 등 다양한 분야의 최신 정보를 실시간으로 제공하는 핵심 매체로 자리 잡고 있다. 그러나 사용자가 관심 있는 분야의 뉴스를 신속하고 간편하게 확인하기 위해서는 포털 사이트에서 원하는 카테고리를 일일이 탐색하고 여러 기사를 거쳐야 하는 번거로움이 여전히 존재한다. 특히 모바일 환경에서는 뉴스 페이지를 여러 번 클릭해야 원하는 정보를 찾을 수 있어 사용자 편의성이 떨어지는 문제가 있다. 본 구현은 이러한 불편함을 해소하고 뉴스 접근성을 향상시키기 위해, 정치, 경제, 사회, 문화, 과학 등 5개 주요 카테고리별 최신 기사 제목을 한 번의 클릭만으로 확인할 수 있는 개인 맞춤형 뉴스레터 추천 프로그램을 개발하는 것을 목표로 하였다. 본 구현의 차별점은 단순히 웹 크롤링을 통한 데이터 수집에 그치지 않고, Python·C++·Java 세 가지 언어를 통합적으로 활용하여 데이터 수집, 처리, 출력까지의 전 과정을 멀티 언어 환경에서 구현했다는 점에 있다. 이를 통해 편리한 뉴스 탐색 경험을 얻을 수 있다.

Ⅱ. 본론

1. 프로그램 구성

구현된 프로그램은 Python, C++, Java 세 언어가 유기적으로 결합 된 멀티 언어통합 구조를 이루고 있다. Python은 전체 애플리케이션의 중심 역할을 하며, 웹 크롤링 및 기사 제목 추출을 담당한다. 이를 위해 urllib, 정규 표현식, 선택적으로 BeautifulSoup 라이브러리[1]를 활용하여 그림 1과 같은 네이버 뉴스의 HTML 문서에서 기사 제목을 수집한다. 또한 Python은 GUI (Graphic User Interface) 및 CLI (Command Line Interface) 환경을 모두 지원하여 사용자가 버튼 클릭이나 콘솔 입력만으로 원하는 카테고리의 뉴스를 확인할 수 있도록 한다. 데이터 정제와 출력 단계에서는 C++과 Java 모듈이 활용된다. Python은 서브프로세스를 통해 C++과 Java 프로그램을 호출하고, 이 과정에서 기사 제목의 중복 제거 및 번호 매기기와 같은 포맷팅 작업이 수행된다. 이러한 구조는 단일 언어 프로젝트보다 복잡하지만, 언어 간 상호 운용성과 프로세스 간 통신에 대한 이해를 심화할 수 있다는 학습적 의의가 크다. 또한 윈도우즈와 유닉스 계열 환경 모두에서 실행 가능하도록 exe 및 ELF (Executable and Linkable Format) 바이너리, Java 바이트코드로 자동 변환되어 크로스 플랫폼 호환성을 제공한다.

그림 1. 크롤링의 바탕이 되는 웹 사이트 예

2. 웹 크롤링 과정

본 프로그램의 핵심 기능은 네이버 뉴스의 각 카테고리(정치, 경제, 사회, 문화, 과학)에서 최신 기사 제목을 자동으로 수집하는 것이다. 이를 위해 Python의 urllib 모듈을 이용하여 지정된 URL에 HTTP 요청을 보내고, 응답으로 받은 HTML 문서를 UTF-8로 디코딩하였다. 디코딩된 HTML은 BeautifulSoup 또는 정규 표현식을 활용하여 <a> 태그 중 실제 기사 링크를 포함하는 요소를 선별하였다. 이후 extract_titles_from_html() 함

수에서 제목 문자열을 추출하고, 불필요한 태그나 공백을 제거하였다. 기사 제목 중 중복되거나 불완전한 항목은 필터링되며, 추가로 모바일 뉴스 페이지를 폴백 소스로 활용하여 부족한 데이터(5개 미만)를 보완하였다. 마지막으로 추출된 제목의 특수문자(", ' 등 HTML 엔티티)는 html.unescape() 함수[2]를 이용해 사람이 읽을 수 있는 문자로 복원하였다. 이러한 크롤링 절차를 통해 프로그램은 실시간으로 최신 뉴스를 안정적으로 수집할 수 있으며, HTML 구조 변경이나 인코딩 문제에 대응하기 위한 예외 처리가 포함되어 있다.

3. 프로그램 구현

사용자 인터페이스는 Python의 표준 GUI 라이브러리인 Tkinter를 활용하여 구현하였다. 그림 2와 같이 본 시스템은 직관적인 그래픽 사용자 인터페이스(GUI) 모드를 제공하며, 사용자는 화면에 표시된 다섯 개의 버튼(정치, 경제, 사회, 문화, 과학) 중 하나를 클릭하여 원하는 뉴스 카테고리를 선택할 수 있다. 버튼 클릭 시 해당 카테고리에 대응하는 네이버 뉴스 섹션에서 HTML 데이터를 수집하고, 상위 5개의 기사 제목을 추출하여 출력창에 실시간으로 표시한다. 그 예시는 그림 3과 같다. 결과는 스크롤 가능한 텍스트 영역을 통해 제공되며, 제목 앞에는 자동으로 번호가 부여되고 중복 기사나 불필요한 문구는 제거된다. 이러한 구조를 통해 사용자는 명령어 입력 없이 클릭 한 번으로 최신 뉴스를 확인할 수 있으며, 단순하고 응답성이 빠른 인터페이스를 구현하였다. 또한 GUI 환경이 지원되지 않는 시스템에서도 동일한 기능을 수행할 수 있도록 CLI(Command Line Interface) 모드를 병행 제공하였다. 두 인터페이스는 모두 동일한 백엔드 로직을 공유하여 코드의 일관성을 유지한다. 이를 통해 유지보수성과 확장성을 높였으며, 실제 실행 환경에 따라 자동으로 GUI 또는 CLI 모드가 선택되도록 설계하였다. 결과적으로, 본 구현은 단순한 크롤링 기능을 넘어 사용자 중심의 접근성과 실용성을 강화한 프로그램의 형태로 발전하였다.

그림 2. 구현 프로그램 메인화면

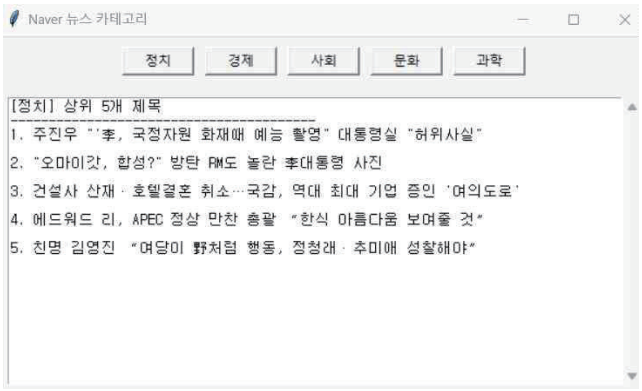

그림 3. 카테고리별 크롤링 결과

Ⅲ. 결론

본 구현에서는 사용자가 관심 있는 뉴스 카테고리를 빠르고 편리하게 탐색할 수 있는 개인 맞춤형 뉴스레터 추천 프로그램을 개발하였다. 본 시스템은 네이버 뉴스의 정치, 경제, 사회, 문화, 과학 5개 주요 카테고리를 대상으로 최신 기사 제목을 실시간으로 수집하고, Python을 중심으로 C++과 Java를 통합하여 데이터의 수집, 정제, 출력까지 하나의 파이프라인으로 처리할 수 있도록 설계되었다. 개발 과정에서 HTML 구조 변화로 인한 파싱 오류와 특수문자 출력 문제를 해결하면서, 프로그램의 데이터 수집 및 처리 안정성을 크게 향상시켰다. 특히 html.unescape() 함수를 통한 이중 디코딩 처리와 예외 제어 로직의 추가로, 다양한 인코딩 환경에서도 깨지지 않는 텍스트 출력을 구현할 수 있었다. 이를 통해 웹 크롤링 환경의 비표준 데이터 문제를 실질적으로 해결하며, 실무 수준의 데이터 정제 기술을 습득할 수 있었다. 또한 Python의 유연한 네트워크 처리 기능과 C++의 빠른 연산 성능, Java의 안정적인 입출력 구조를 유기적으로 결합함으로써 멀티 언어 통합 개발 환경을 성공적으로 구축하였다. 이러한 개발 경험은 단일 언어 기반의 프로젝트를 넘어, 실제 산업 현장에서 필요한 프론트 엔드와 백 엔드 간의 협력 구조 이해, 그리고 이기종 언어 간 데이터 연동 및 최적화 능력을 강화하는 데 중요한 밑거름이 되었다. 본 연구의 결과물은 단순한 뉴스 크롤러를 넘어, 사용자 맞춤형 정보 접근성과 서비스 편의성을 높이는 실질적인 응용 가능성을 지닌다. 향후 연구에서는 기사 요약, 감성 분석, 키워드 기반 추천 기능을 추가하여 보다 고도화된 개인화 서비스로 확장할 계획이다. 이를 통해 정보의 양적 소비를 넘어 질적 탐색이 가능한 뉴스 추천 시스템으로 발전시키는 것이 궁극적인 목표이다.

참 고 문 헌

[1] 나철원, 온병원, 최신 웹 크롤링 알고리즘 분석 및 선제적인 크롤링 기법 제안 (Journal of Internet Computing and Services), 인터넷정보학회논문지, pp. 43-59, 2019

[2] 이종화, Python을 이용한 SNS 크롤링 시스템 구축(Building an SNS Crawling System Using Python), 한국산업정보학회, 한국산업정보학회논문지, pp. 61-76, 2018