

Streamlit 대시보드 프로젝트 기획서

프로젝트명

네이버 뉴스와 수치형 데이터 시각화를 활용한 대시보드

1. 프로젝트 개요

본 프로젝트는 사용자가 네이버 뉴스를 크롤링하여 특정 카테고리의 토픽을 분석하고, 이를 워드 클라우드 형태로 시각화하는 동시에, 수치형 데이터를 활용한 시각화를 통해 직관적이고 효율적인 데이터 분석 경험을 제공하는 Streamlit 기반 대시보드 애플리케이션을 제작하는 것을 목표로 합니다.

2. 주요 기능

2.1 네이버 뉴스 크롤링 및 토픽 모델링

- 사용자가 정치, 경제, 사회, 생활/문화, IT/과학, 세계 카테고리 중 하나를 선택하면 해당 카테고리의 최신 기사를 크롤링합니다. (Selenium 은 사용하지 않음)
- 텍스트 데이터를 정제하고 불용어를 처리하여 텍스트 클렌징을 수행합니다.
- Latent Dirichlet Allocation (LDA) 모델을 활용하여 토픽 모델링을 수행한 후, 결과를 워드 클라우드로 시각화 합니다.

2.2 수치형 데이터 시각화

- 수치형 데이터(예: 주식, 부동산, 국제 유가 등)를 1 개 이상 크롤링하여 시각화를 진행하고 유의미한 정보를 제공합니다.
- 시각화는 최소 3 가지 이상을 포함합니다. (다음의 내용은 예시입니다.)
 - 시간에 따른 변화 추세 그래프 (Line Chart)
 - 카테고리별 비교를 위한 막대 그래프 (Bar Chart)
 - 비율 데이터를 시각화하는 파이 차트 (Pie Chart)

2.3 모듈화된 파이썬 코드

- 프로젝트의 각 기능은 모듈화하여 유지보수성과 재사용성을 높입니다.
 - 모든 모듈은 재사용 가능한 함수와 클래스를 기반으로 설계됩니다.
 - 설정값과 공통 상수는 별도의 설정 파일에서 관리하여 일관성을 유지합니다.
 - 예외 처리를 각 모듈에 포함하여 안정적인 실행을 보장합니다
-

3. 기술 스택

- **Frontend:** Streamlit
 - **Backend:** Python (requests, BeautifulSoup, pandas, matplotlib, seaborn 등)
 - **데이터 분석:** scikit-learn, Gensim, nltk, konlpy 등
 - **데이터 시각화:** WordCloud, Matplotlib, Plotly
-

4. 기대 결과물

1. 사용자가 네이버 뉴스 카테고리를 선택하여 최신 기사를 기반으로 한 토픽 모델링 결과와 워드 클라우드를 확인할 수 있음.
 2. 수치형 데이터를 시각화하여 정보를 효과적으로 전달할 수 있음.
 3. Streamlit 기반으로 직관적이고 반응형 UI를 제공함.
-

5. 프로젝트 구조 예시

```
project/
|
|—— app.py                # Streamlit 애플리케이션 실행 파일
|—— modules/
|   |—— crawler.py         # 네이버 뉴스 및 수치형 데이터 크롤링
|   |—— text_processing.py # 텍스트 정제 및 불용어 제거
|   |—— topic_modeling.py  # LDA 기반 토픽 모델링
|   |—— visualization.py  # 워드 클라우드 및 데이터 시각화
|   |—— config.py         # 공통 설정 (API 키, 상수 등)
|
|—— data/                 # 크롤링된 데이터 저장
|—— assets/               # 기타 리소스
|—— requirements.txt      # 의존성 패키지 목록
|—— README.md            # 프로젝트 설명 파일
```