miniProject 2 mission : 한글 (뉴스) 자동으로 분류 프로그램 구현

자동 분류할 뉴스 카테고리는 팀별 2개 이상으로 하며, 자유 주제입니다.
예) 부동산 뉴스와 경제 뉴스, 사회 뉴스와 정치 뉴스

* 뉴스 분류 외에 다른 자연어 분류 학습을 주제로 사용해도 괜찮습니다.

[가이드]

1. 문제 정의 : 팀별 뉴스 주제 정하기
2. 일정 계획 및 R&R (Role and Responsibility) 정하기
3. 개발 전 표준 수립
   가. 데이터 포멧 : 수집된 데이터 파일 저장 포멧 결정 – XML or JSON or 기타 / key, tag 명 등 결정
   나. 명명 규칙 : 파일 명, 함수 명 등 결정

[절차]

1. 데이터 수집
Selenium, urllib.request, OPEN API 등을 활용하여 카테고리 별 학습 데이터 수집

* 주의 사항 : 반복 문으로 요청을 여러 번 하는 경우 , 10초에 한번씩만 크롤링하도록
반드시 타임 딜레이(import time ; time.sleep(10) ;) 줍니다.
짧은 시간에 너무 많은 요청을 보내면 수시간에서 하루 이상 서비스요청이 거부될 수 있습니다.
Open api의 경우는 해당 키만 막히지만,
Open api를 사용하지 않고, 요청할 때 IP가 막히면 반 전체가 해당 사이트 요청이 거부되니 더 조심해 주세요.

2. 데이터 전처리

3. 모델 생성 (one hot encoding 또는 embedding 또는 RNN 활용)

4. 학습 : 수집된 데이터로 지도 학습

5. 평가 : 검증 전용 데이터로 평가.

6. 예측 프로그램 작성
성능이 가장 좋은 모델로 예측하는 프로그램을 작성합니다.
뉴스를 입력 받아 어느 카테고리에 속하는 뉴스인지 예측합니다.

7 PPT 작성 및 발표