**Weekly report**

날짜 : 2022-05-12

연구원 : 최윤석

* 수행결과

▶ 신문기사 데이터를 파싱해 (단어, 품사, NE)셋을 만들어주는 모델(inference)을 로컬환경에서 사용해 10만개의 뉴스데이터를 학습 -> 오류발생

▶ Colab 환경에서 1000개의 뉴스데이터를 이용해 학습 후 학습한 데이터를 파일로 저장해 추후 재사용 가능

▶ 1000개의 뉴스데이터 학습에 5시간 소요, 대부분의 시간이 inference에서 사용되는 것으로 확인

텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 1. 1000개의 뉴스데이터를 학습해 돌린 연관단어 검색기의 결과

* 수행계획

▶ inference 모델이 로컬환경에서 돌아갈 때 발생하는 오류 해결

▶ 오류를 해결하더라도, 10만개의 뉴스데이터를 돌리려면 500시간 이상이 걸릴 것으로 예상, 1만개의 뉴스데이터도 50시간 이상으로 예상됨. 따라서 수행시간의 대부분을 차지하는 inference 코드를 빠르게 돌릴 방법이 없는지 확인

▶ 고유명사와 개체명 인식할 때, 단어에 조사가 붙는 경우가 많은데 (ex: 최윤석은 – NNP – PER) 이를 해결할 방안 찾기

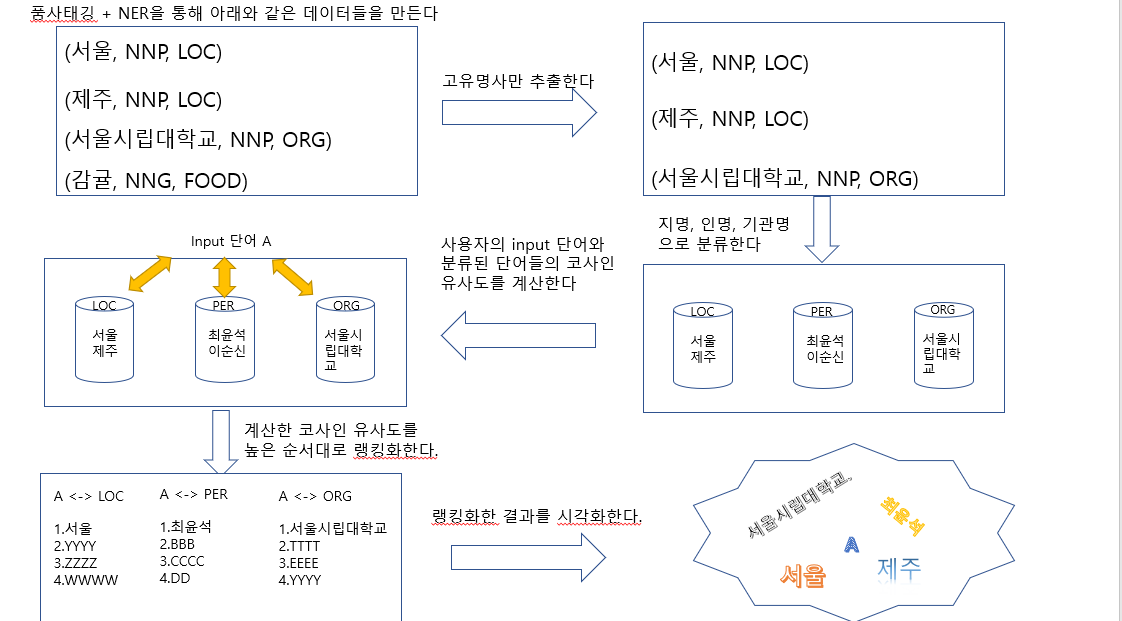


Figure 2. 앞으로의 진행 흐름

* 기타사항

▶ 카테고리별로 고유명사가 잘 나오는지 UI를 먼저 완벽하진 않지만 먼저 테스팅해보기 – 조사제거는 UI가 먼저 어느정도 완성된 후에 진행

▶ 한국 전자 거래 학회에 제출할 논문 작성, 웹에서 UI 작성

▶