**Weekly report**

날짜 : 2022-05-05

연구원 : 최윤석

* 수행결과

▶ 신문기사 데이터를 가져와서 문장단위로 파싱텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 1. 문장단위로 파싱한 것을 inference해서 품사와 NE가 무엇인지 알려준다.

▶ 단어들 중 품사가 NNP이고 PER/ORG/LOC인 단어들만 추출

▶ 추출한 단어들 중 기존 N2V모델에 존재하는 단어들만 중복없이 한번 더 추출

▶ PER / ORG / LOC를 따로 분류해서 인명, 기관명, 지명 리스트를 생성후 추출한 단어들을 각각의 리스트에 분류

▶ 사용자가 임의의 input을 집어넣으면 인명, 기관명, 지명과 코사인 유사도 계산 후 유사도가 높은 순서대로 5개를 출력하는 프로그램 구현텍스트이(가) 표시된 사진

자동 생성된 설명

Figure 2. input단어 청와대와 유사한 인명, 기관명, 지명을 출력하는 것. (30개 뉴스기사 기준이라 정확한 값은 나오지 않습니다.)

* 수행계획

▶ 뉴스기사가 총 10만개가 있는데, 10만개를 전부다 학습하기 위해서는 많은 시간이 필요함 (30개 - 5분이상, 오래 돌리기 어려운 Colab 환경에서의 어려움) -> Local 환경(jupyter notebook)에서 이부분을 진행할 예정

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

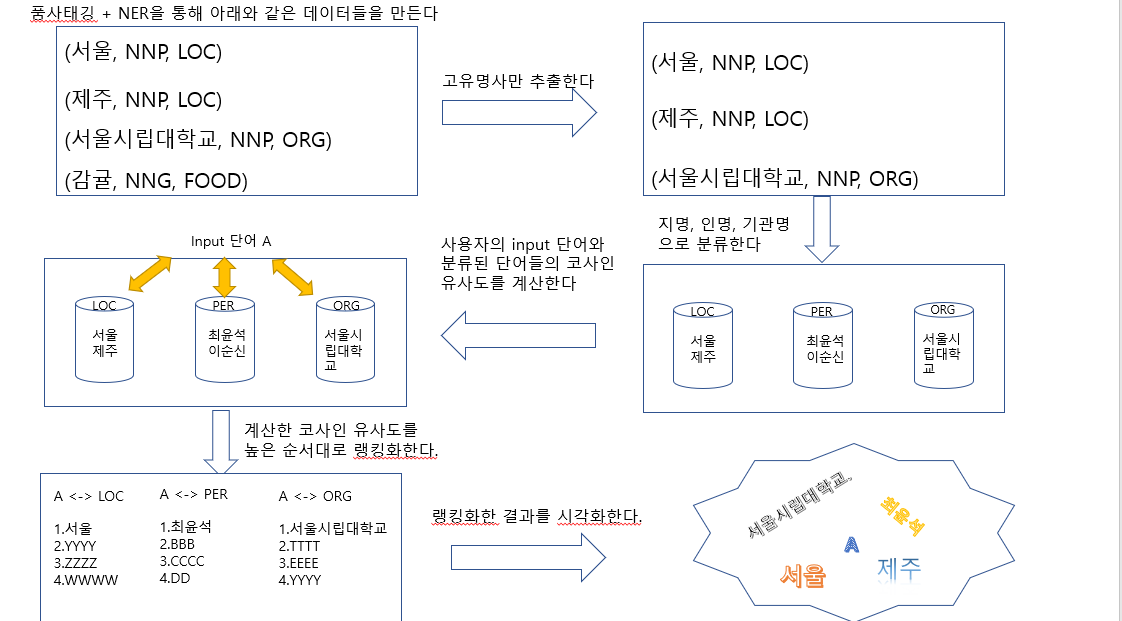


Figure 3. 앞으로의 진행 흐름

* 기타사항

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

▶

▶