NLP 자료조사

1. 한국어 맞춤법 검사 및 띄어쓰기 패키지

Py-Hanspell

1. 키워드 추출 모델

Keybert

konlpy or soynlp로 명사 추출 ->sbert를 이용한 키워드 추출

1. 데이터

뉴스 대본 및 앵커 음성 데이터 셋(<https://kdx.kr/data/view/32567>)

한국전자통신연구원\_한국어 및 영어 음향모델 훈련용 음성 데이터(<https://www.data.go.kr/data/15073486/fileData.do>)

AI HUB 요약 문데이터셋(https://aihub.or.kr/aihubdata/data/view.do?currMenu=115&topMenu=100&aihubDataSe=realm&dataSetSn=582)

1. 요약

4-1. STT model: kospeech학습 or EspNet중 성능시험

4-2. summarization: KoBert or gpt

4-3. keyword extraction model:

4-3-1. key-bert

4-3-2. 명사 추출(soybert, konlpy) -> sbert를 사용하여 키워드 추출