팀 미 팅 회 의 록

팀명	4726	차수	3 차
일 시	2019 년 3월 7일 목요일 17시 0분 – 20시 0분 (3시간 00분)		
장 소	국민대학교 7호관 424호실		
참석자	고현경, 김희주, 김혜인, 이선홍, 이수민		
불참자			
안 건	자연어처리 단계별 구체화된 절차는 무엇인가?		
회의내용	1. KoInPy를 활용한 데이터의 정제. KoInPy 데이터 정제를 통해 각 문장을 분절하는 토큰화를 진행한다. 토큰화를 통해 품사별 분석을 마치고 명사를 추출하여 어휘목록을 정리한다. 2. 추출한 데이터를 벡터로 구조화하는 방법: 1) Word2vec에는 알고리즘이 두 가지 방식이 있다. Skip-gram, CBOW. 2개의 알고리즘은 단어를 문장 주변 요소를 통해 분석, 예측 하는 알고리즘이다. 2) Word2vec을 통해 벡터화를 하는 방법도 있지만 TF-IDF(단어의 빈도, 문서의 단어 출현빈도)를 정리하여 벡터화 하는 방법이 있다. 추출한 데이터를 벡터화하는 방법 1)과 2)중에서 어떤 방식을 선택할 것인지 토의한 결과 2)의 TF-IDF 방식을 활용하는 방식으로 결정. 3. 벡터화된 데이터를 비교 분석하는 방식에 대하여. 1) k-means 군집 알고리즘을 통해 군집하는 방법. 2) Bi-LSTM을 적용하여 텍스트를 분류하여 인식하는 방법. 3) 데이터를 피쳐벡터로 만들고 피쳐벡터간의 유사도를 비교하여 결정하는 방법. 자소서의 분석, 학습을 위해서 벡터화된 데이터를 비교하는 단계가 필요하다. 벡터화된 데이터를 비교 분석하는 1), 2), 3)의 방식중에서 3) 피쳐벡터를 생성 후 코사인유사도 분석을 통해 유사도를 연산해내는 방식으로 진행하도록 한다.		