Embedding 및 LSTM을 통한 댓글분석

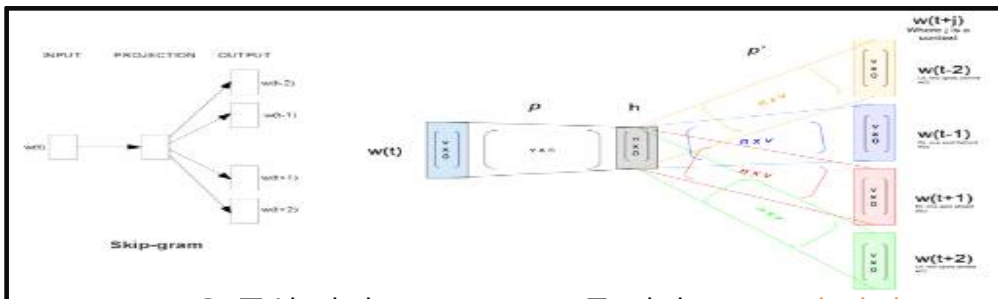
팀명 : 곰두리타 **이름 : 박성흠**

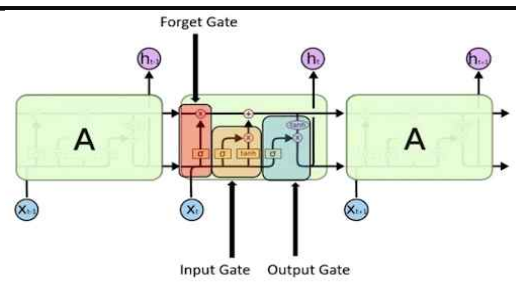
담당교수 : 이수안

1.목적 및 배경

1. 4차산업으로 인한 초개인화시대에 적응하기 위한 개인의 특성, 니즈파악
2. DeepLearning을 통한 감정분석 및, 댓글 오염도 확인
3. Skip-gram을 통해 특성추출하여 개인에 맞춘 시각화 및 캐릭터화

2.내용 및 방법

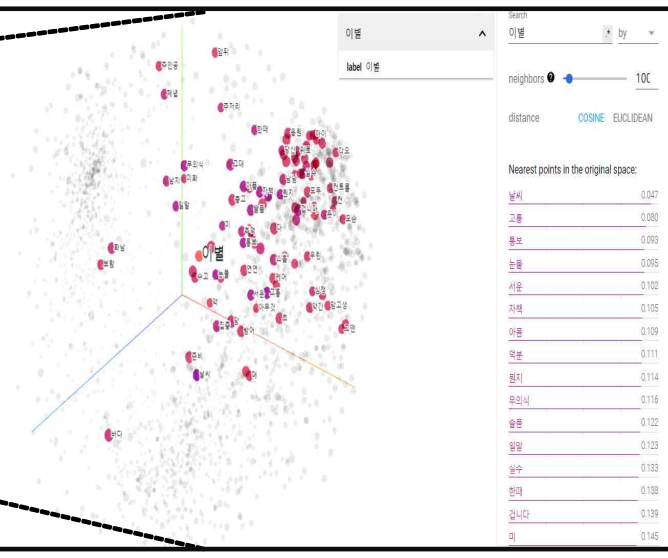




$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] + b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$
$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t] + b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Skip-grams은 중심 단어(Central word)를 기반으로 주변 단어(Neighboring word) 추측하여 유사도 기반의 분산표현으로 밀집 백터(Dense Vector)화 하여 임베딩(Embedding)시킨다. 임베딩된 데이터는 반복되는 순환구조를 통해 과거의 데이터를 반영하고, 정보들이 셀(cell)을 넘어갈 때마다 선택적으로 받아들여 기존 RNN형식의 장기 의존성 문제(Problem of Long-Term Dependencies)을 해결한 LSTM모델을 사용한다.

3.데이터 소개- 강원대학교 커뮤니티



1.단어/문장 간 관련도 계산

단어를 전체 단어들간의 관계에 맞춰 해당 단어의 특성을 갖는 백터로 바꾸면 단어들 사이의 유사도를 계산하는 방법 가능

2.의미적/문법적 정보 함축

단어 백터 간 덧셈/뺄셈을 통해 단어들 사이의 의미적, 문법적 관계를 도출 가능

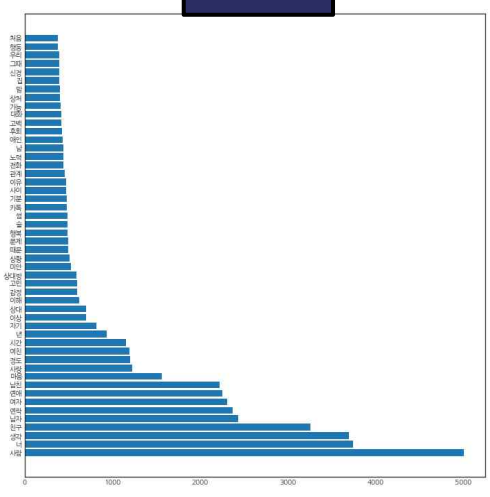
3.전이학습(Transfer learning)

품질 좋은 임베딩은 모형의 성능과 수렴속도가 빨라 좋은 품질의 임베딩을 다른 딥러닝 모델의 입력값으로 사용 가능

4.프로젝트 결과

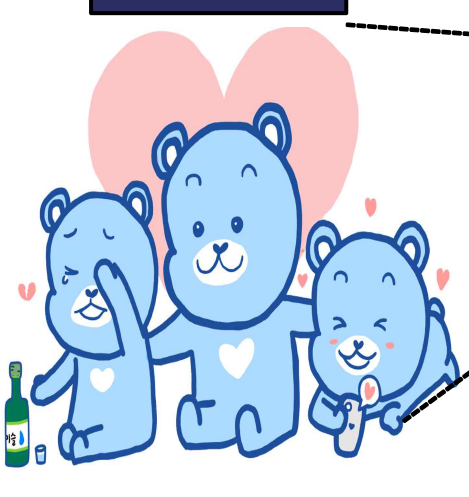
I.특성파악 및 캐릭터화

분석

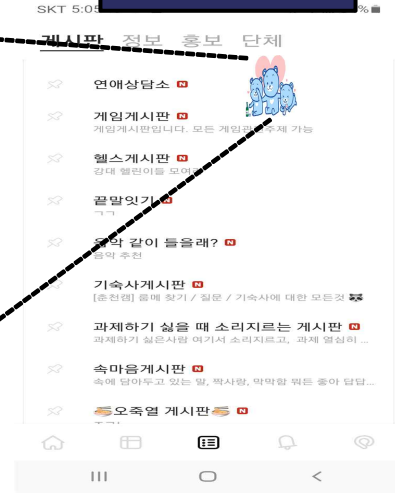


특성 시각화

캐릭터화

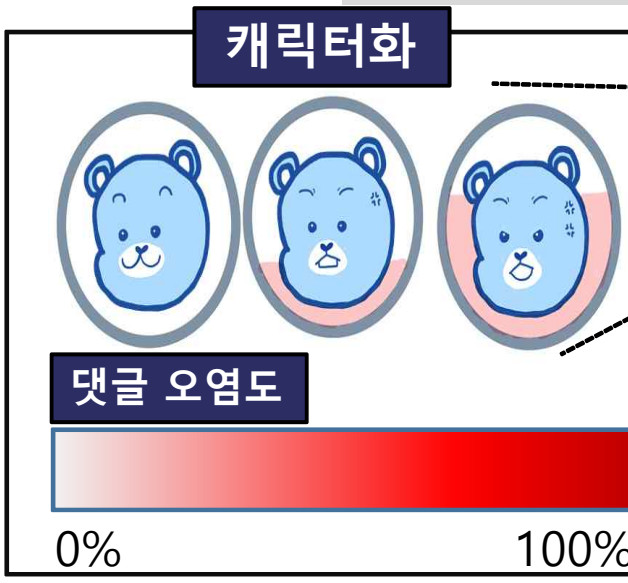


적용

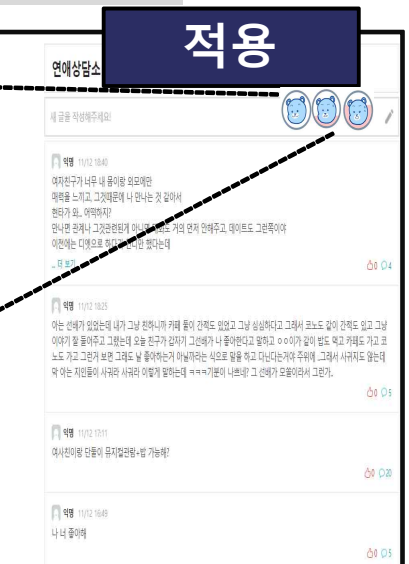


II.댓글 오염도 측정

캐릭터화



적용



5. 기대효과

1. 막연한 상식, 편견, 주관적 기준이 아닌 분석을 통해 객관적인 정보 기준으로 데이터를 시각화 하여 **사실기반의 분석,이해가능**
2. 데이터 기반으로 특성추출하여 가장 많이 언급된 단어를 통 **니즈에 알맞은 특성으로 캐릭터화 가능**
3. 전체 댓글 대비 악성 댓글 기준으로 비율이 높아질수록 게시판의 우측 상단에 곰두리의 표정과 색이 변하여 즉각적인 피드백이 가능
4. 악성 댓글에 대한 경각심을 일으키고, 시각적으로 상황을 인지하여 악성,혐오조성댓글을 경계하고 간접적인 방식으로 **자연스러운 배려문화 조성**