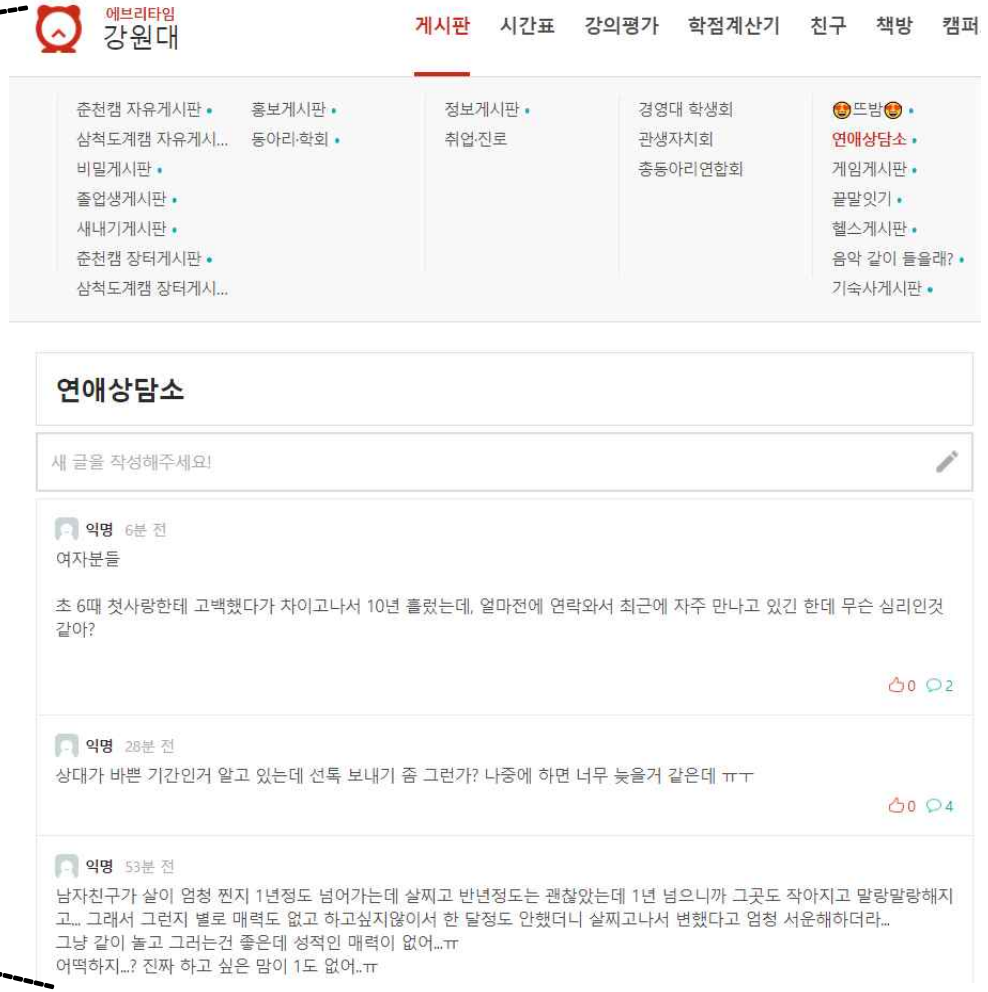
Embedding및 LSTM을 통한 NLP 분석

에브리타임 Sentiment Analysis and Characterization

박성흠

강원대학교 물리학과 제일원리 전자구조계산 연구실

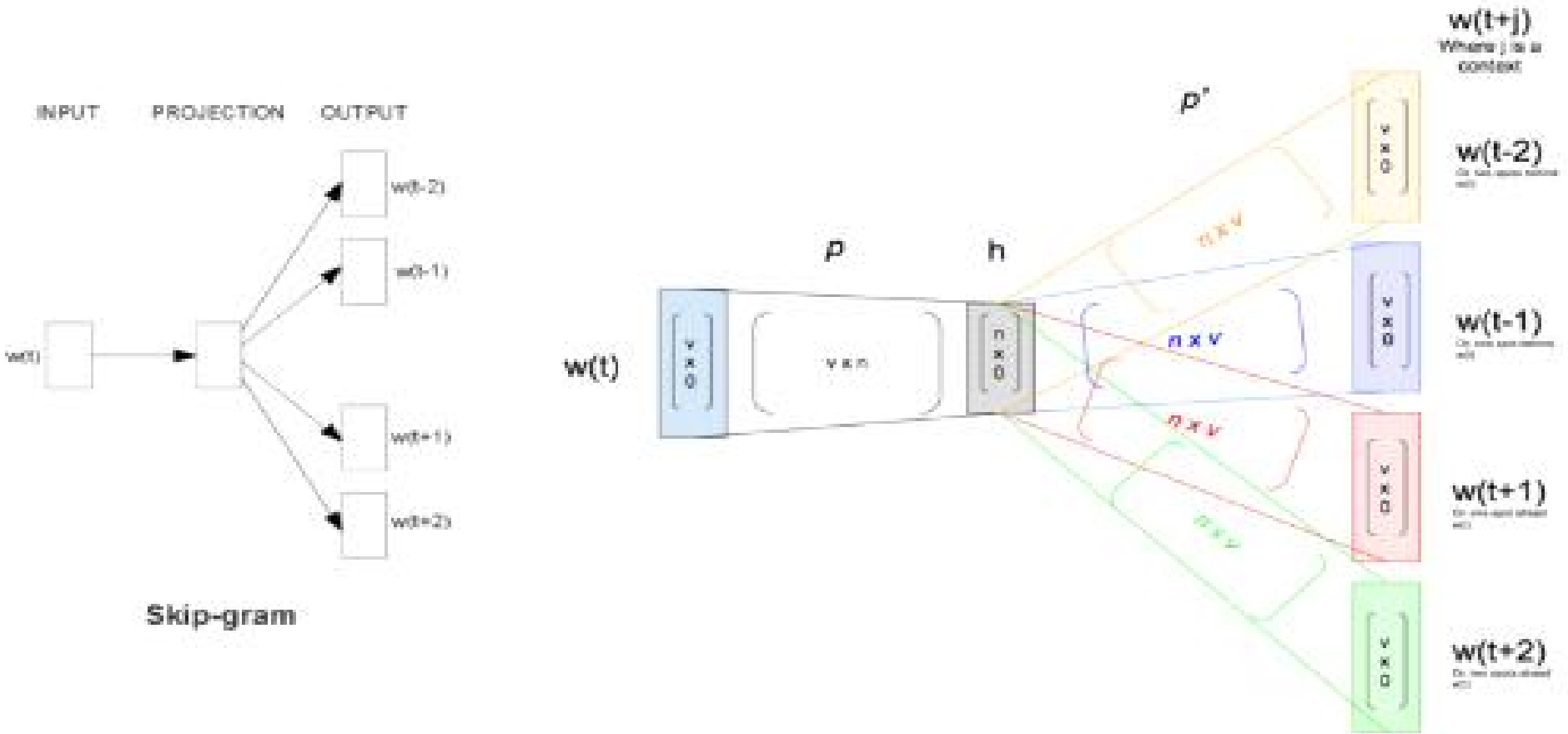
1-1. 웹 크롤링(Web Crawling)



Web Crawling

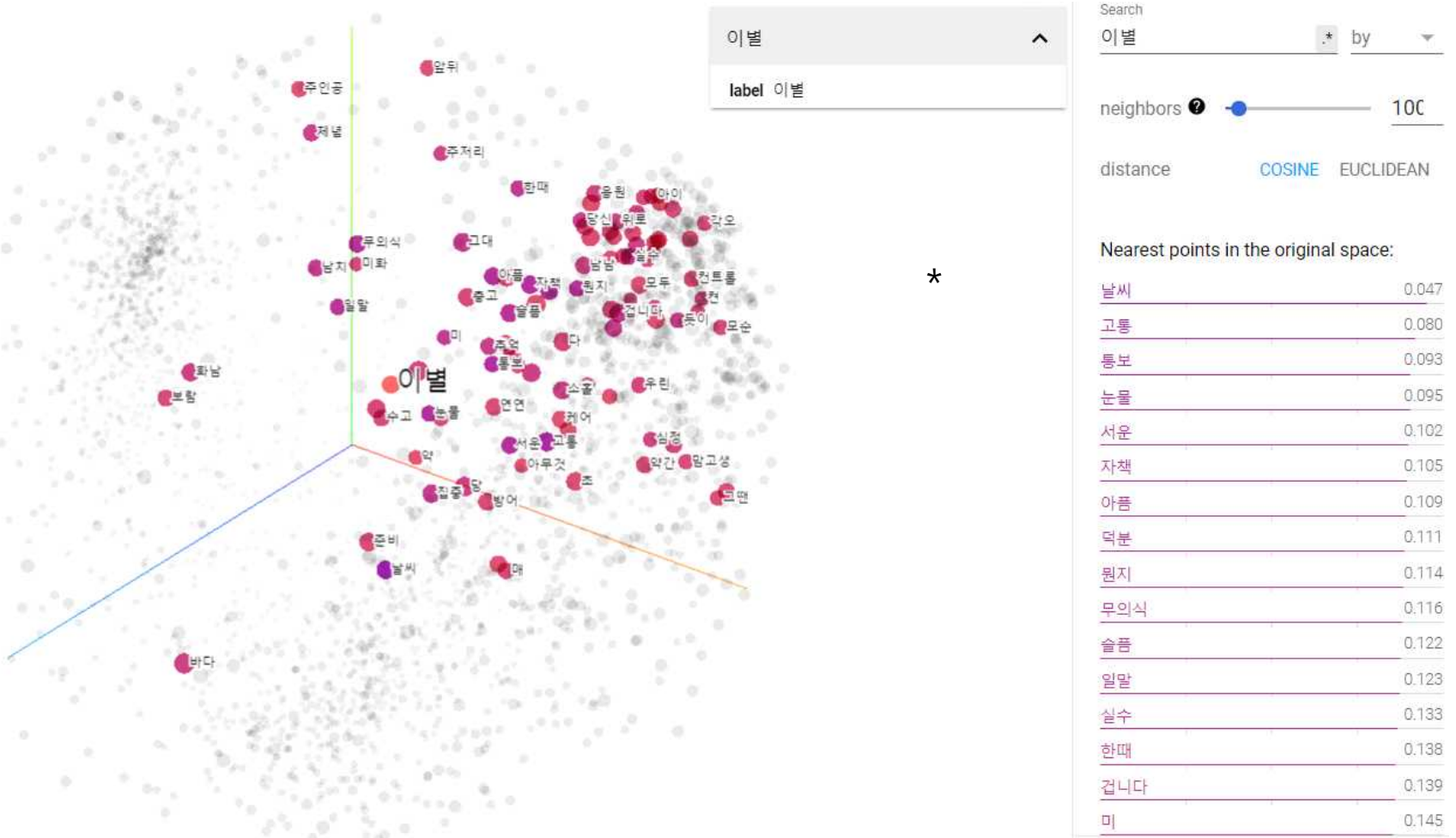
인터넷에서 존재하는 데이터를 컴퓨터 프로그램을 통해
자동화된 방법으로 웹에서 데이터를 수집하는 작업

단어 임베딩(Word Embedding) - Skip gram



- 중심 단어(Central word) → 주변 단어(Neighboring word) 추측
- 유사도 기반의 분산 표현

1-2. 임베딩 공간 시각화(Dense vector visualization)



임베딩의 역할

1) 단어/문장 간 관련도 계산

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

2) 의미적/문법적 정보 함축

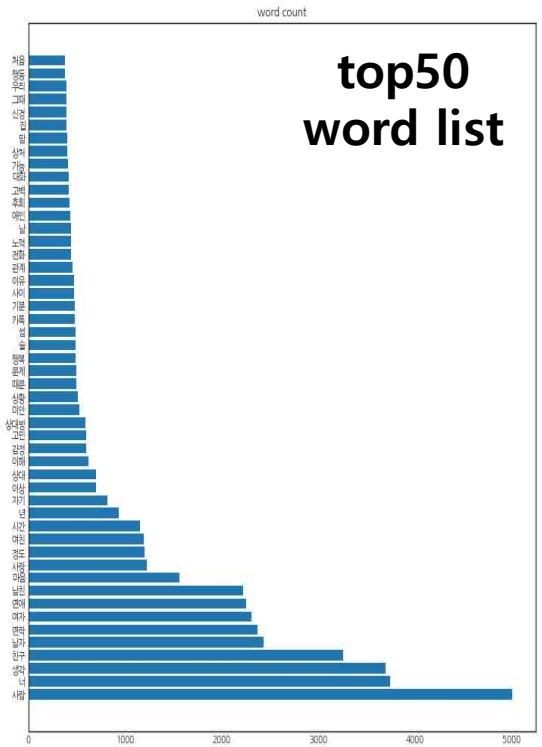
단어 백터간 덧셈/뺄셈을 통해 단어들 사이의 의미적, 문법적 관계를 효율적으로 도출 가능

3) one-hot-encoding의 한계 극복

고차원으로 표현된 희소백터를 저차원으로 표현하여 효율적인 방식으로 저장

2. 분석된 데이터를 통한 캐릭터화

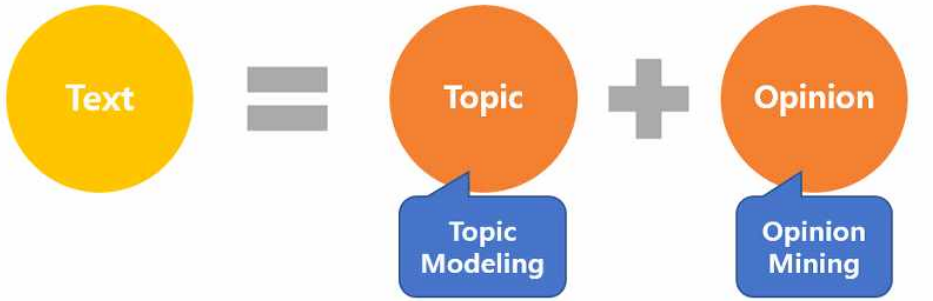
분석 → **시각화** → **캐릭터화** → **적용**



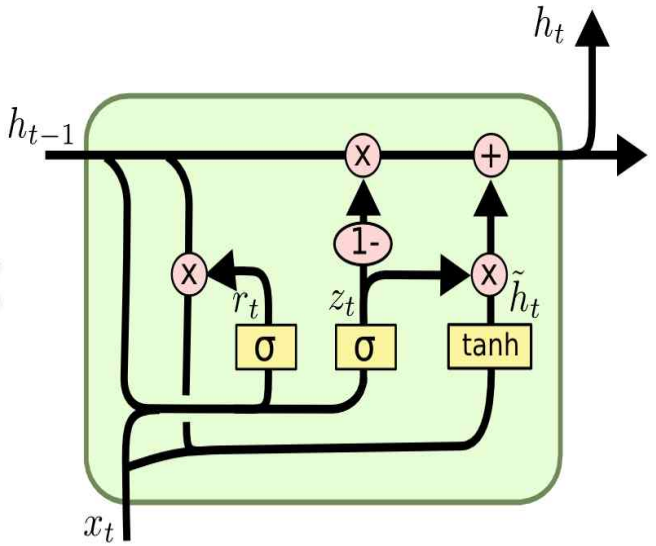
I. 막연한 상식, 편견, 주관적 생각이 아닌 **사실기반의 분석,이해**

II. 각각 커뮤니티의 분위기를 데이터기반으로 이해하고 **분석된 특징을 캐릭터화**

3-1. 감정 분석(Sentiment Analysis)



LSTM의 수학적 기법



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

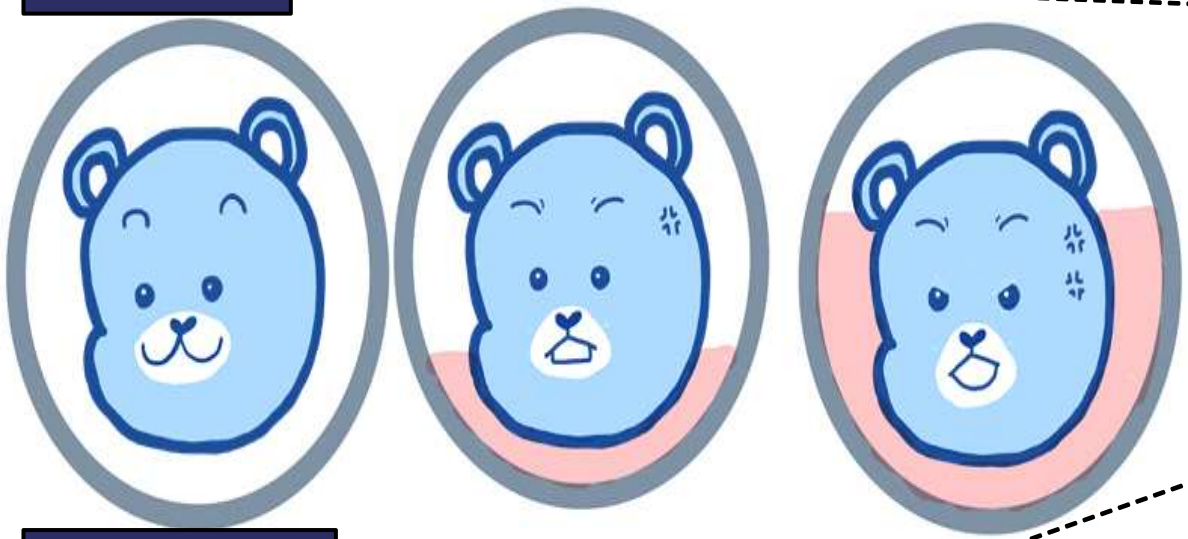
I. 수집한 데이터 기반으로 LSTM 자연어 처리 모델 학습

II. 사용자들의 댓글을 구분하여 긍정/부정 카테고리로 분류

III. 댓글의 긍정 부정의 정도를 확률적으로 판정

3-2. 감정 분석을 통한 시각화

캐릭터화



댓글 오염도

0%

100%

- 전체 댓글 대비 악성댓글 기준으로 비율이 높아질수록 각 게시판마다 우측 상단에 곰두리의 표정과 색을 변화
- 곰두리의 상태를 파악하여 실시간으로 게시판의 분위기,상황 파악
- 악성 댓글에 대한 경각심을 일으키고 배려 문화 조성

적용