



Embedding 및 LSTM을 통한 NLP 분석

에브리타임 Sentiment Analysis and Characterization

박성흠

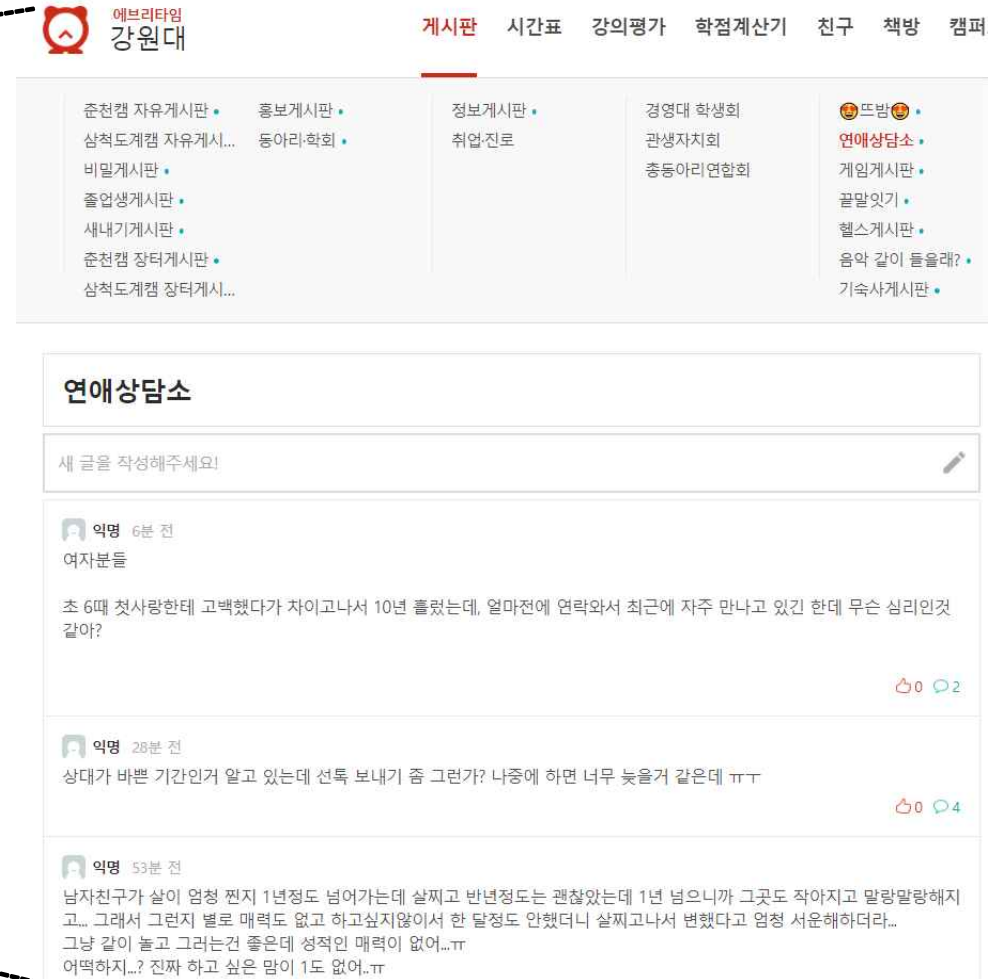
강원대학교 물리학과 제일원리 전자구조계산 연구실

1-1. 웹 크롤링(Web Crawling)

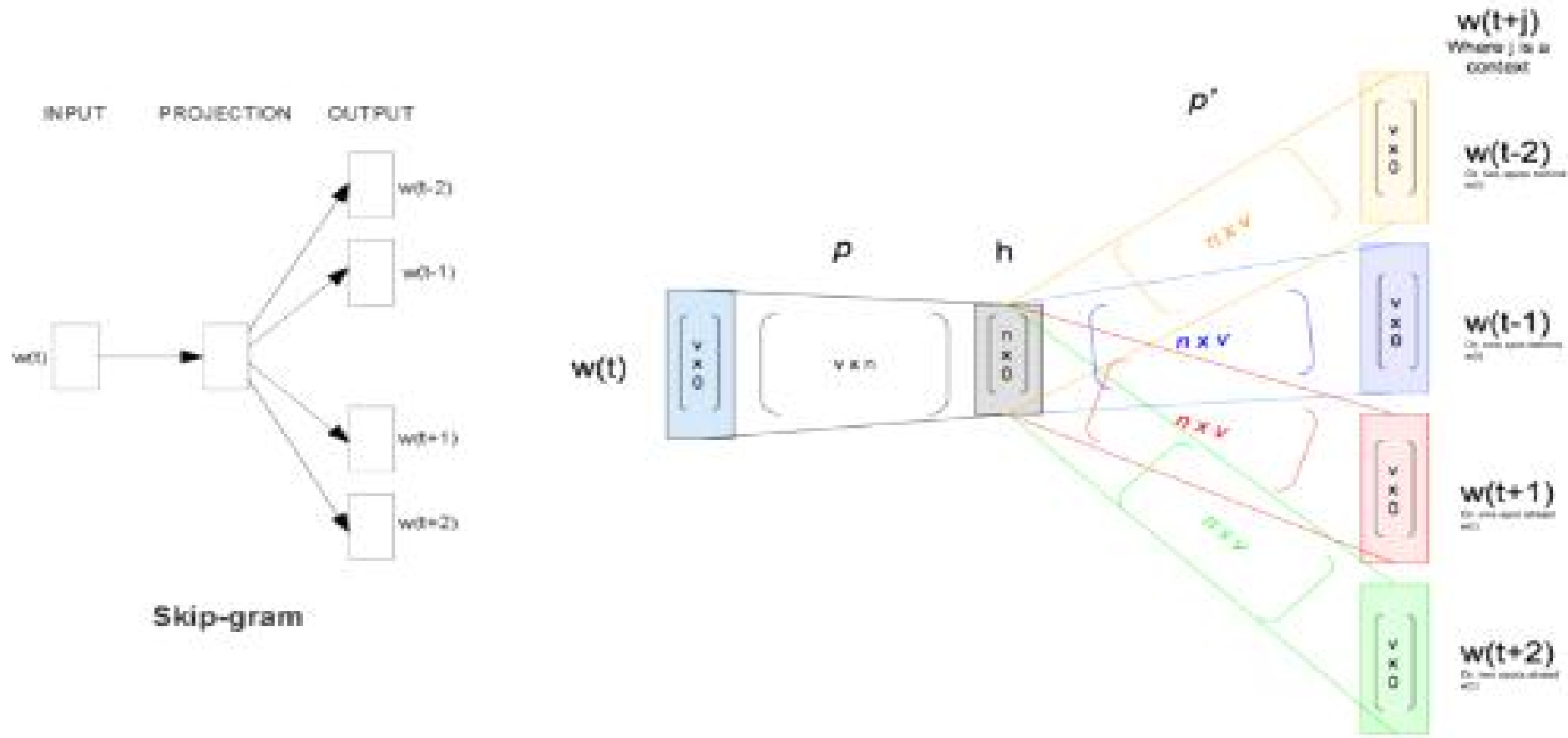


Web Crawling

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

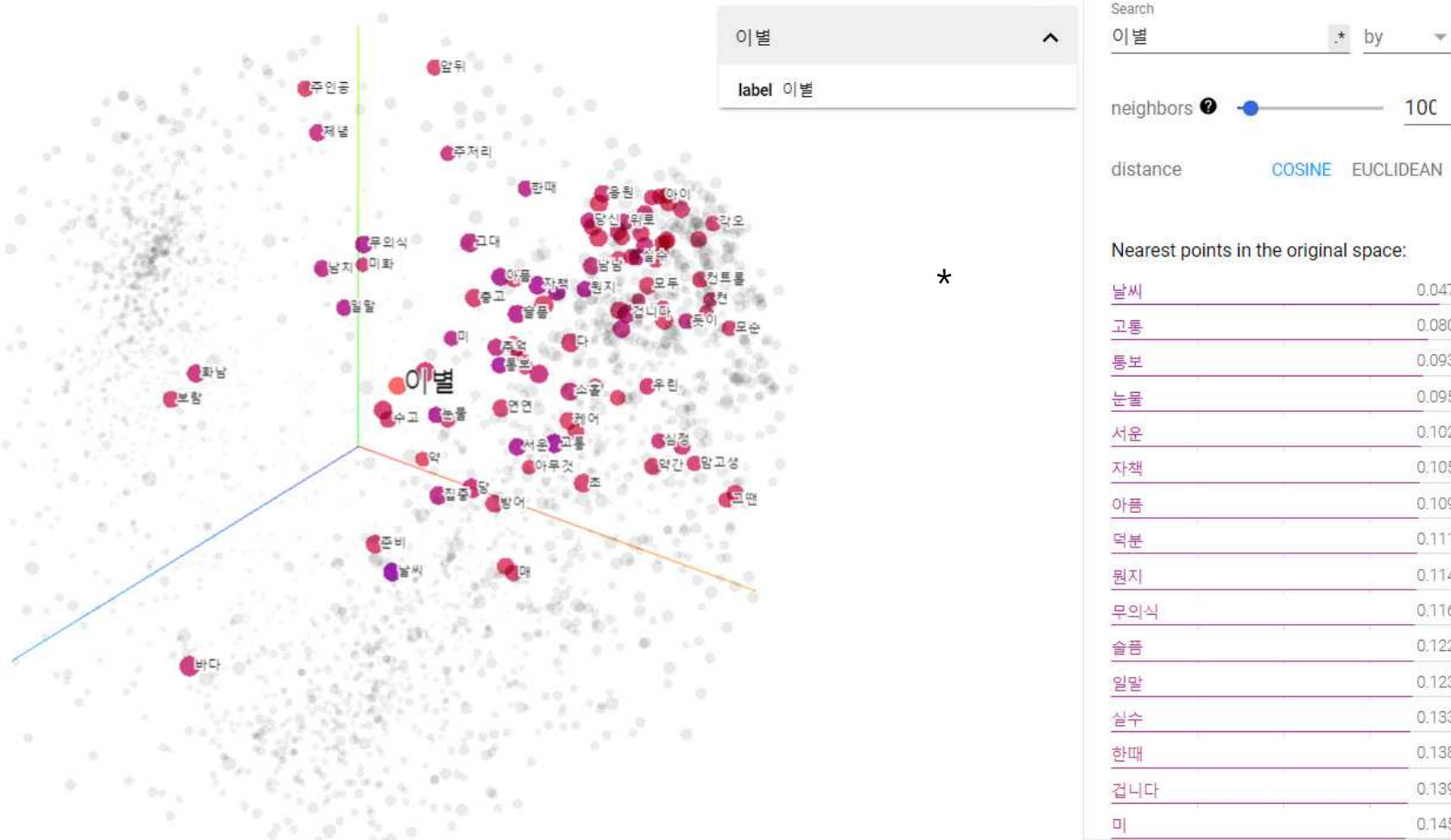


단어 임베딩(Word Embedding) - Skip gram



- 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 벡터로 바꾸면
단어들 사이의 유사도를 계산하는 일이 가능

2) 의미적/문법적 정보 함축

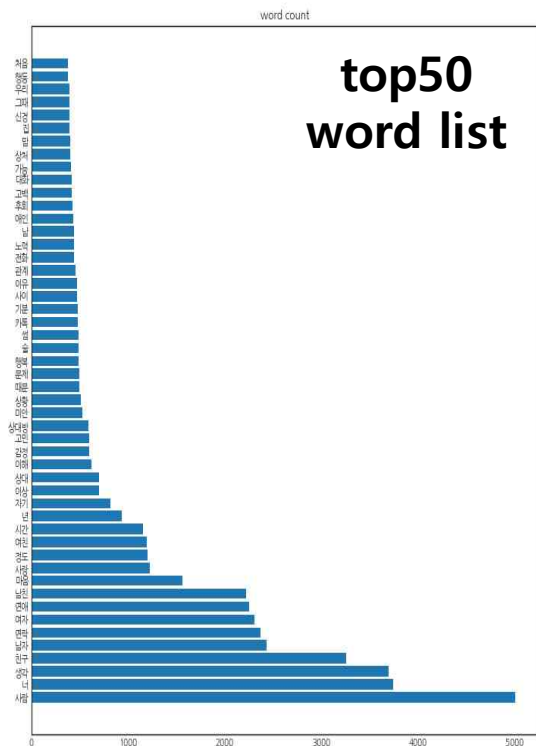
단어 벡터간 덧셈/뺄셈을 통해 단어 들 사이의 의미적, 문법적 관계를 효율적으로 도출 가능

3) one-hot-encoding의 한계 극복

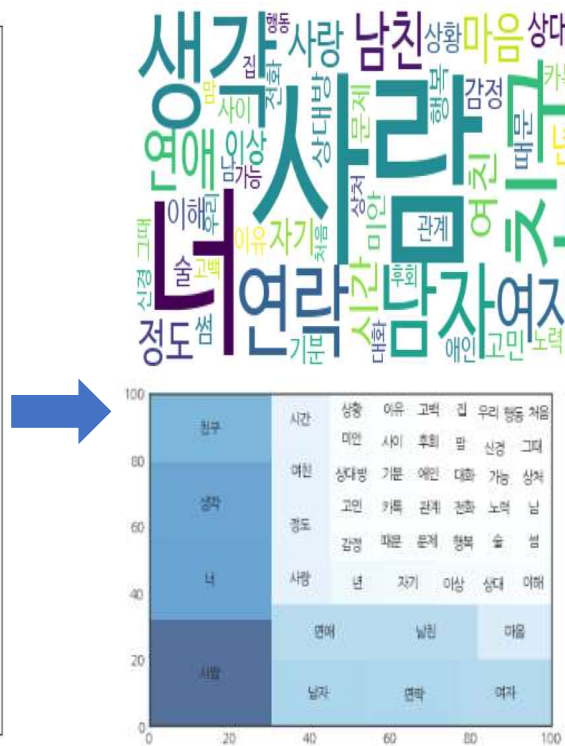
고차원으로 표현된 희소벡터를 저 차원으로 표현하여 효율적인 방식으로 저장

2. 분석된 데이터를 통한 캐릭터화

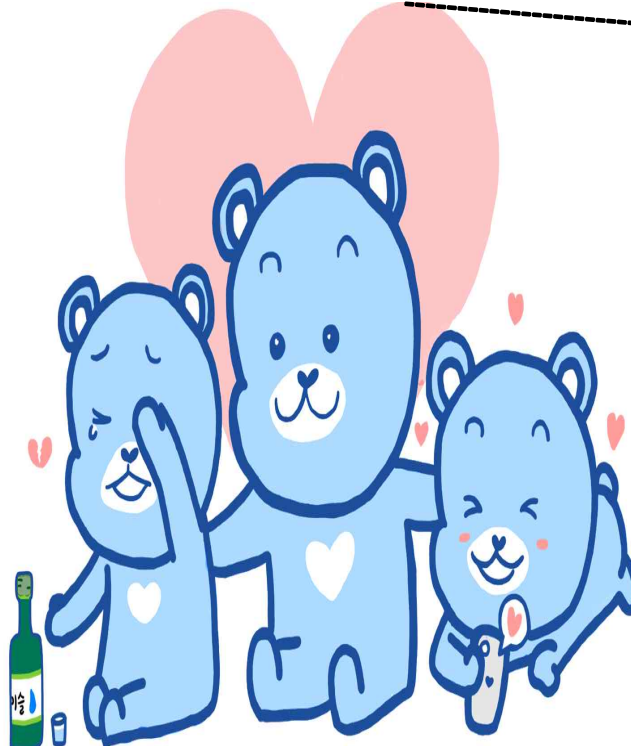
분석



시각화



캐릭터화

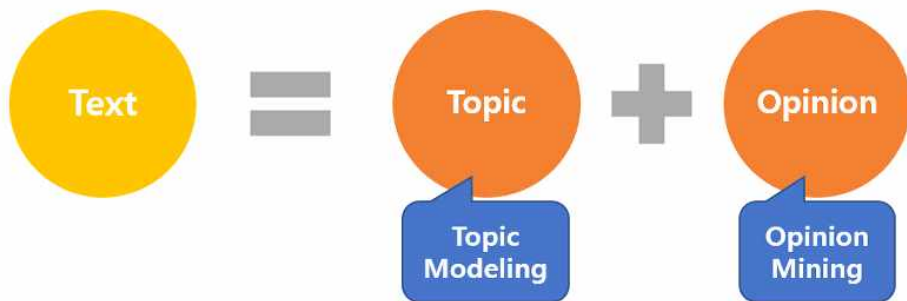


적용

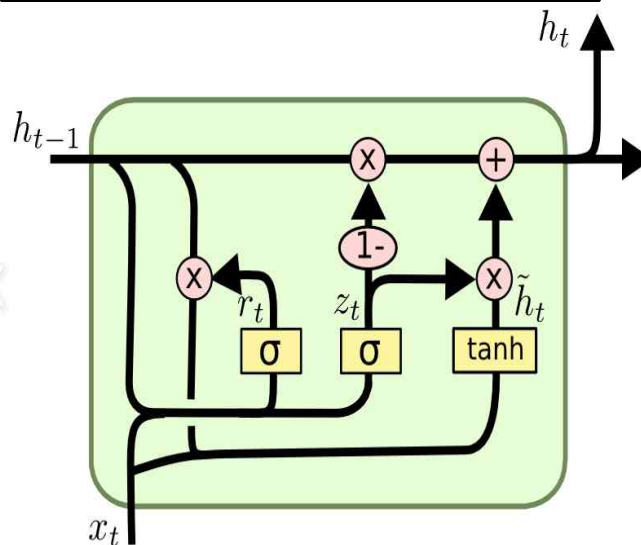


- I. 막연한 상식, 편견, 주관적 생각이 아닌 **사실기반의 분석,이해**
- II. 각각 커뮤니티의 분위기를 데이터기반으로 이해하고 **분석된 특징을 캐릭터화**

3-1. 감정 분석(Sentiment Analysis)



LSTM의 수학적 기법



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

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

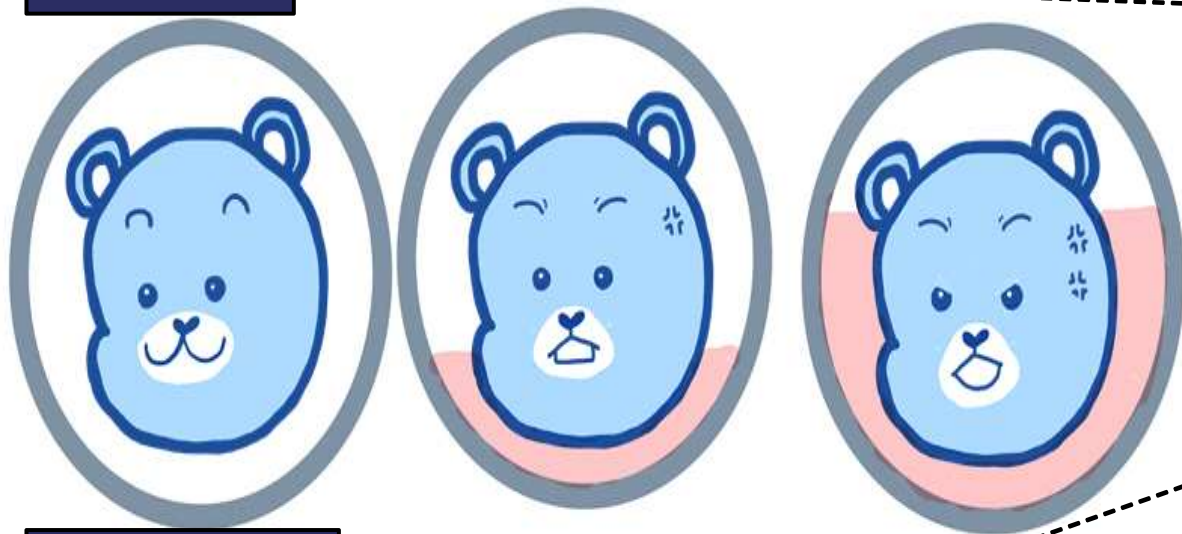
$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t])$$

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

- 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

연애상담소



새 글을 작성해주세요!

익명 11/12 18:40

여자친구가 너무 내 몸이랑 외모에만 매력을 느끼고, 그것때문에 나 만나는 것 같아서 현타가 와... 어떡하지?
만나면 관계나 그것관련된게 아니면 대화도 거의 먼저 안해주고, 데이트도 그런쪽이야 이전에는 디오텔로 하다가 연디만 했다는데

... 더 보기

0 4

익명 11/12 18:25

아는 선배가 있었는데 내가 그냥 친하니까 카페 둘이 간적도 있었고 그냥 심심하다고 그래서 코노도 같이 간적도 있고 그냥 이야기 잘 들어주고 그랬는데 오늘 친구가 갑자기 그선배가 나 좋아한다고 말하고 ㅇㅇ이가 같이 밥도 먹고 카페도 가고 코노도 가고 그런거 보면 그래도 날 좋아하는거 아닐까라는 식으로 말을 하고 다닌다는거야 주위에 ..그래서 사귀지도 않는데 막 아는 지인들이 사귀라 사귀라 이렇게 말하는데 ㅋㅋㅋ기분이 나쁘네? 그 선배가 모쪼니까서 그런가..

0 5

익명 11/12 17:11

여자친구랑 단둘이 뮤지컬관람+밥 가능해?

0 20

익명 11/12 16:49

나 너 좋아해

0 5