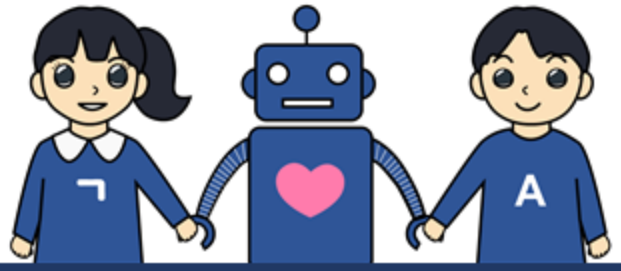
MICROSOFT X 바벨피쉬 2018.01.20

HANDS-ON LAB.

딥러닝 분류 결과 해석하기
: sentiment analysis with grad-CAM and target2vec

김형준



본 발표의 시발점 & 궁극 목표 → 챗봇!?

[목적 지향 검색기 혹은 QA]

Q: 데이터 분석을 공부하고 싶어요

목적(target): 공부하다

대상(object): 데이터 분석

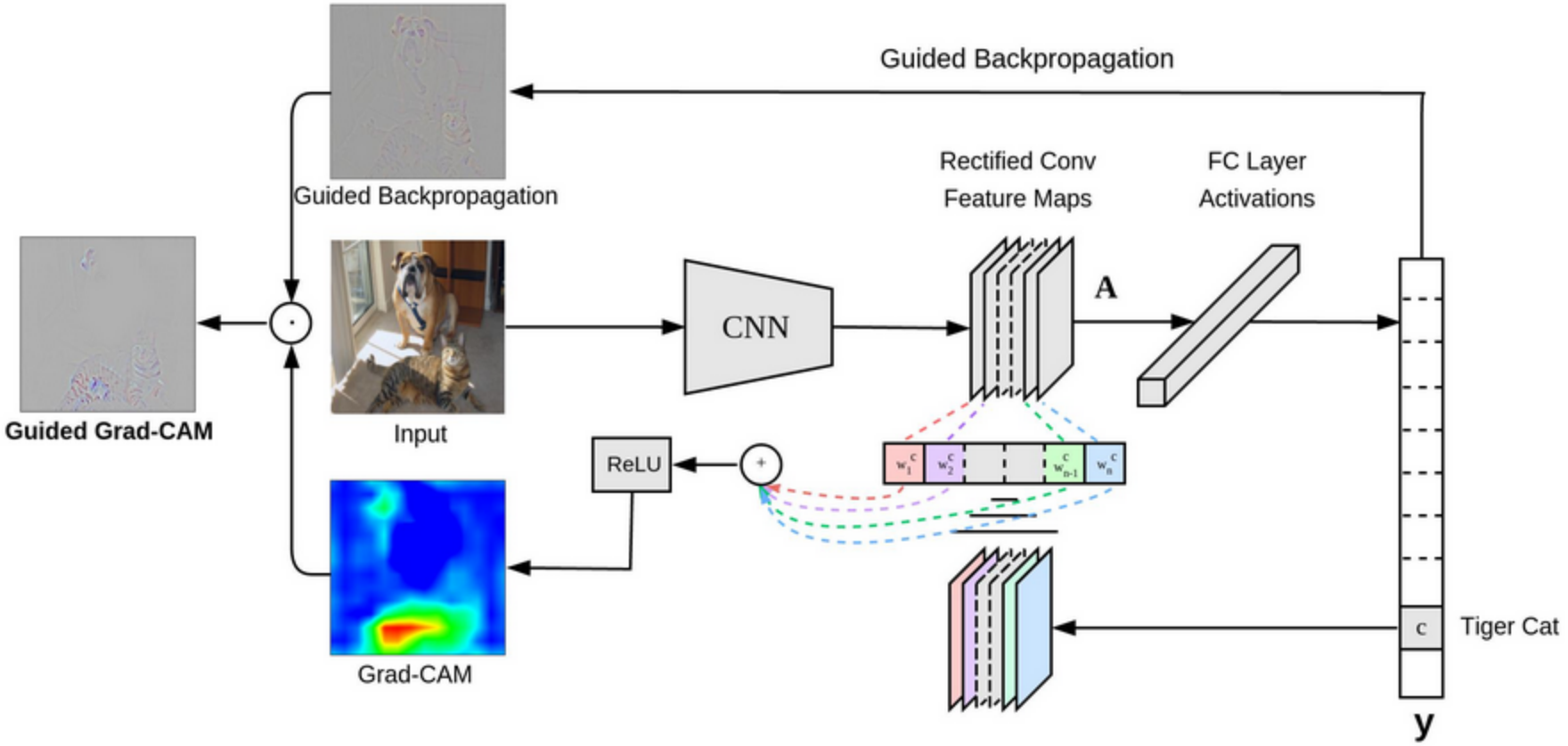
그러나, 질의자는 '데이터 분석'이 무엇인지 잘 모른다.

A: 파이썬? 딥러닝? R?

→ 대상(object)과 관련된 키워드들을 추출하고자 함

다른 예)
Q: 딥러닝을 공부하고 싶어요?

A: 케라스? 파이토치? 파이썬? 이론?

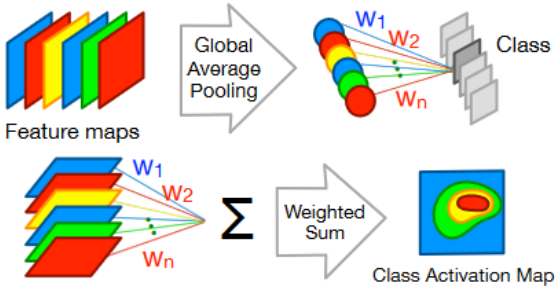


https://github.com/ramprs/grad-cam

1. Class Activation Mapping (CAM)

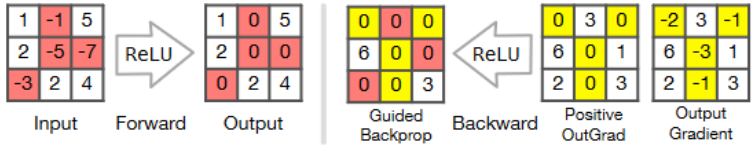
Identify discriminative regions in image classification tasks [Zhou et al., 2016].

The CAM compute the **linear combination of the feature maps** of the last convolutional layer using the **weights** corresponding to a given class. However, the modification of architecture requires re-training.



2. Guided Back-Propagation

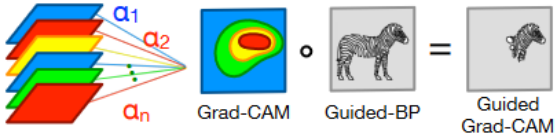
Use the positive gradient of output for back-propagation, for a sharper visualization [Springenberg et al., 2015].



3. Grad-CAM

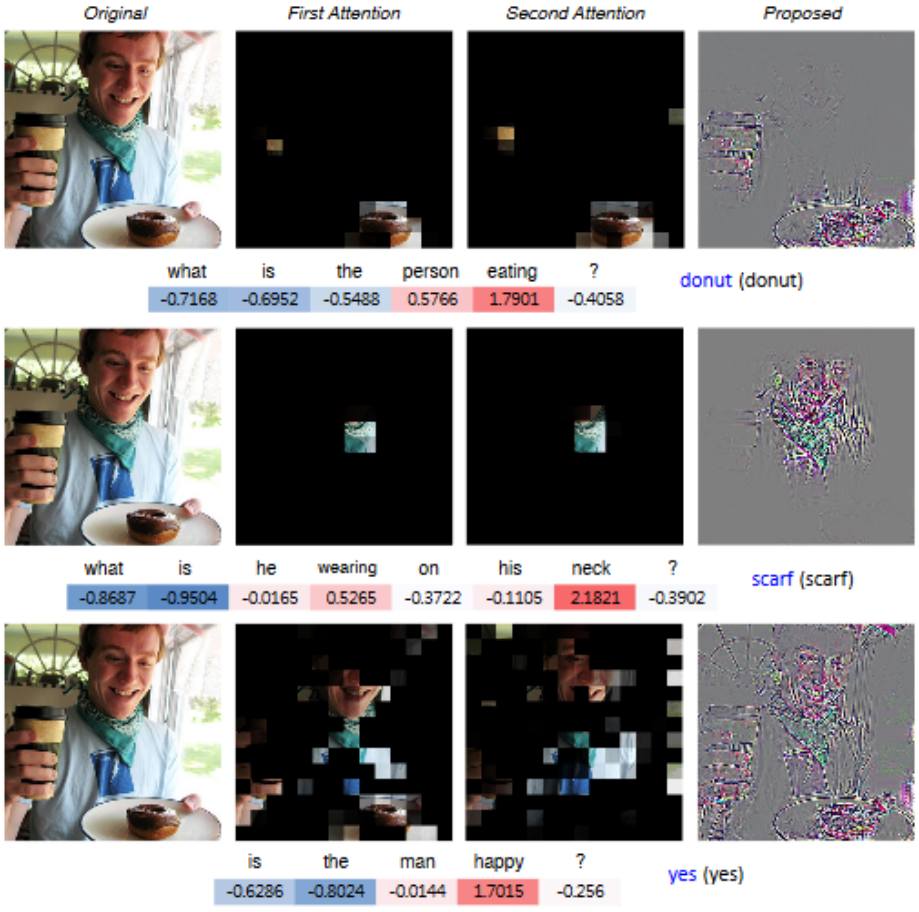
A generalization to CAM, replacing the weights *w* by the summation of the gradient of a class with respect to feature map [Selvaraju et al., 2017]

$$\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$

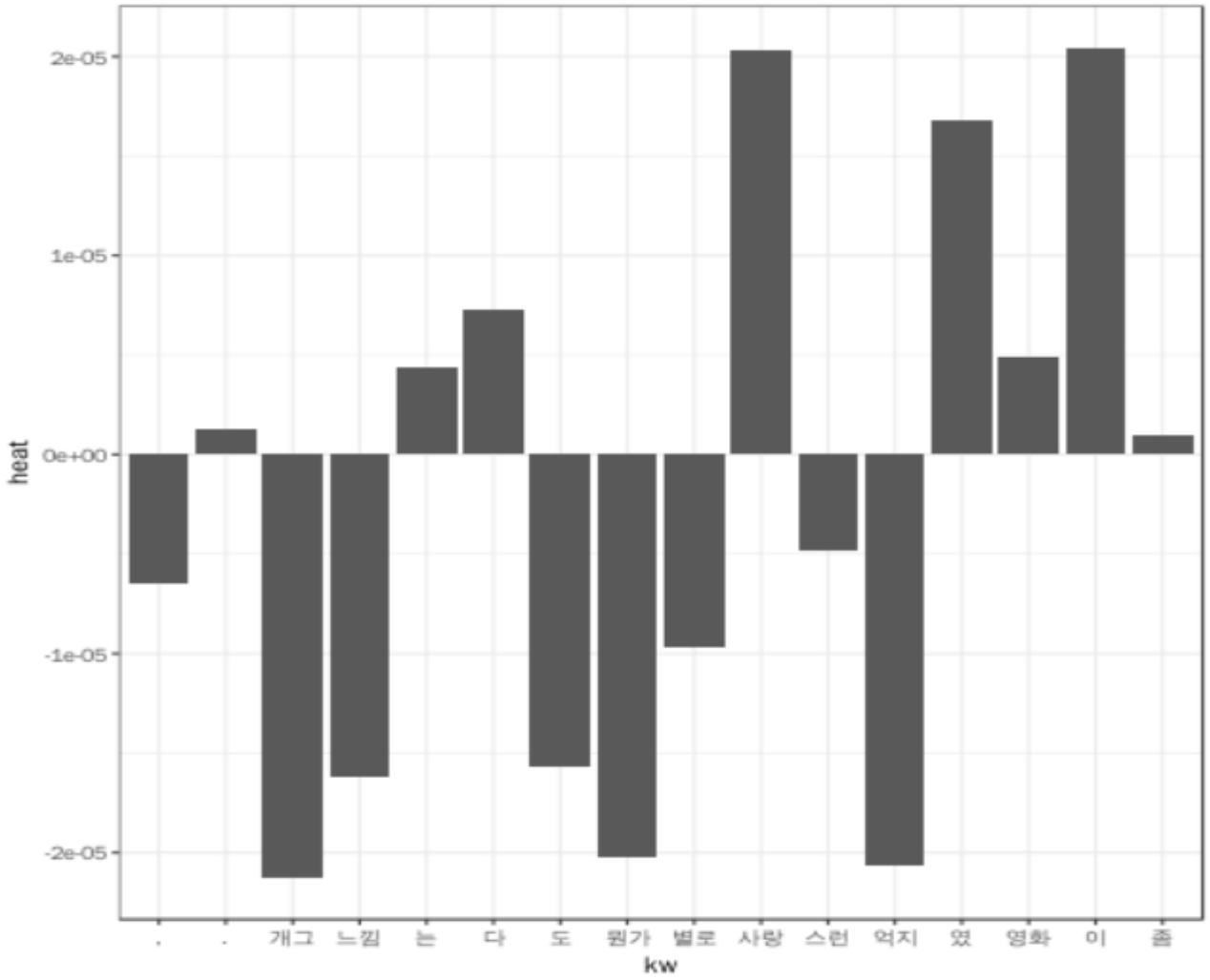


https://bi.snu.ac.kr/~jhkim/posters/vis-nips17ws-poster.pdf

Visualizations



https://bi.snu.ac.kr/~jhkim/posters/vis-nips17ws-poster.pdf



http://freesearch.pe.kr/archives/4685

[목적 지향 검색기 혹은 QA]

Q: 데이터 분석을 공부하고 싶어요

목적(target): 공부하다

대상(object): 데이터 분석

A: 파이썬? 딥러닝? R?

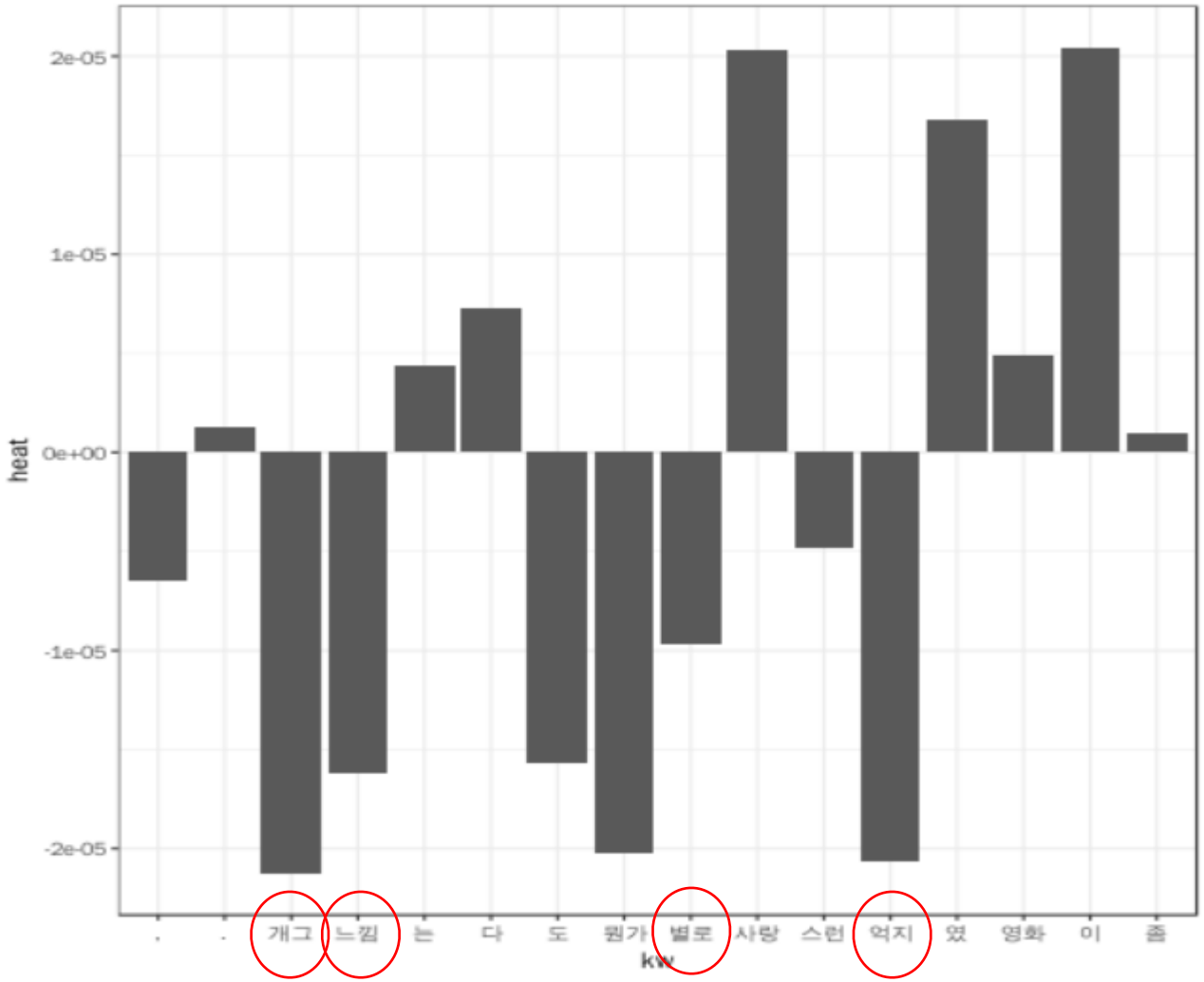
[딥러닝의 분류 과제]

목적(target): 분류 (긍정이냐 부정이냐)

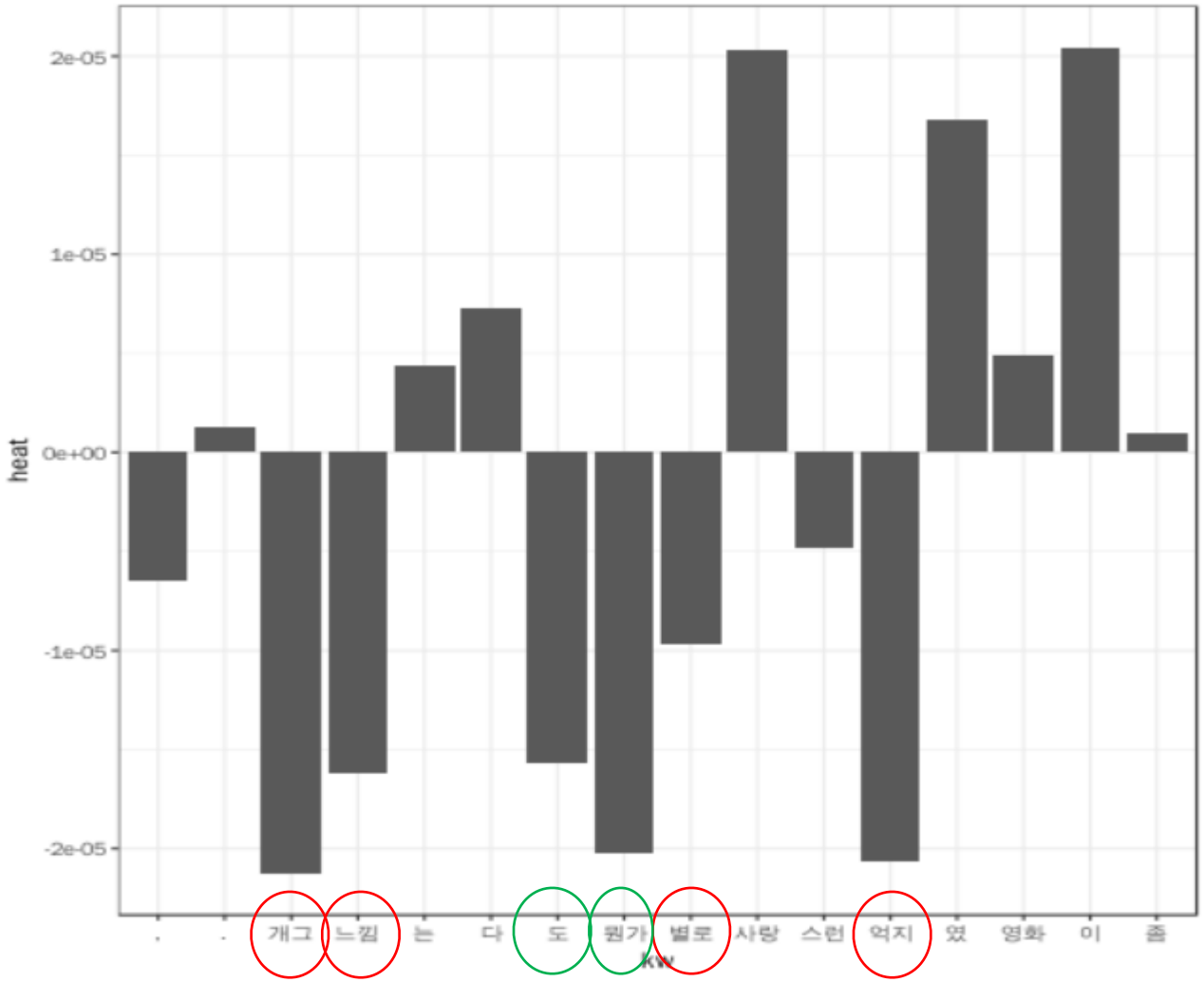
대상(objects): 무엇을 가지고??

긍정에 영향을 끼치는 단어

부정에 영향을 끼치는 단어



http://freesearch.pe.kr/archives/4685



http://freesearch.pe.kr/archives/4685

[Objective]

더 나아가 …

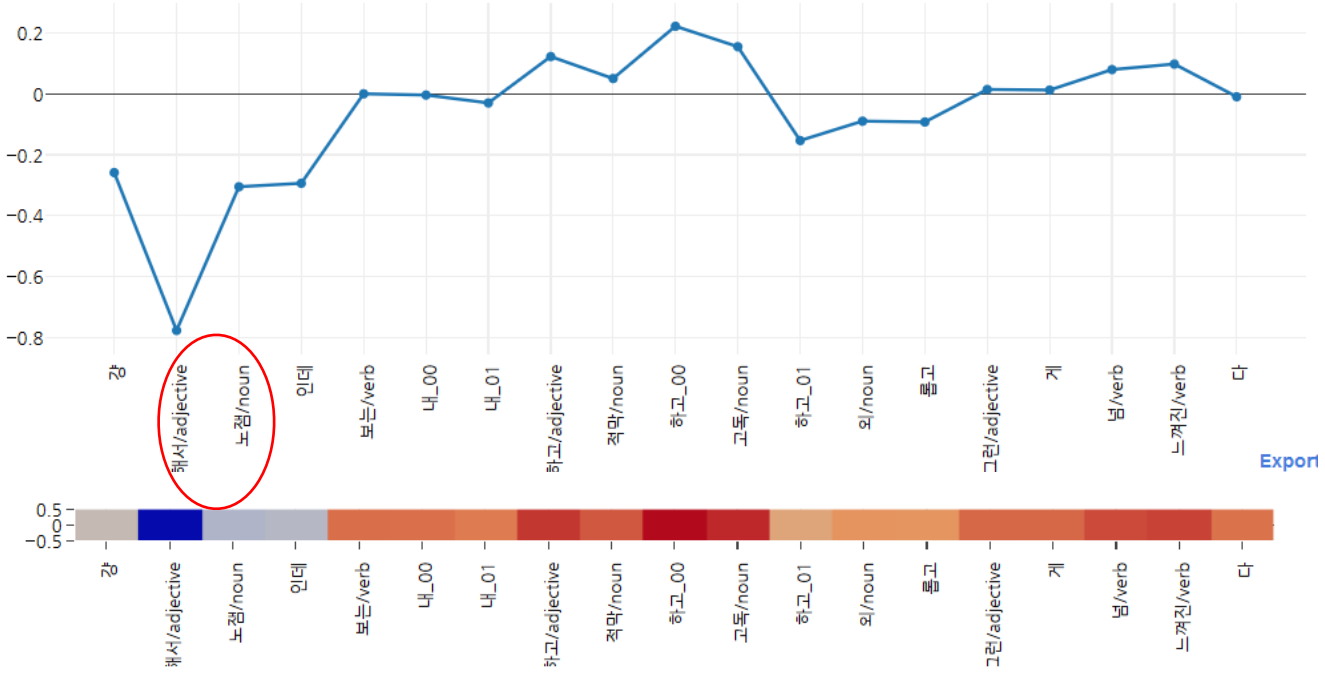
스토리가 부자연스럽고 억지스러움 → 지루함 → 부정적!

<역으로 계산 >

부정적 영화평 ← 지루함 ← 스토리, 부자연, 억지

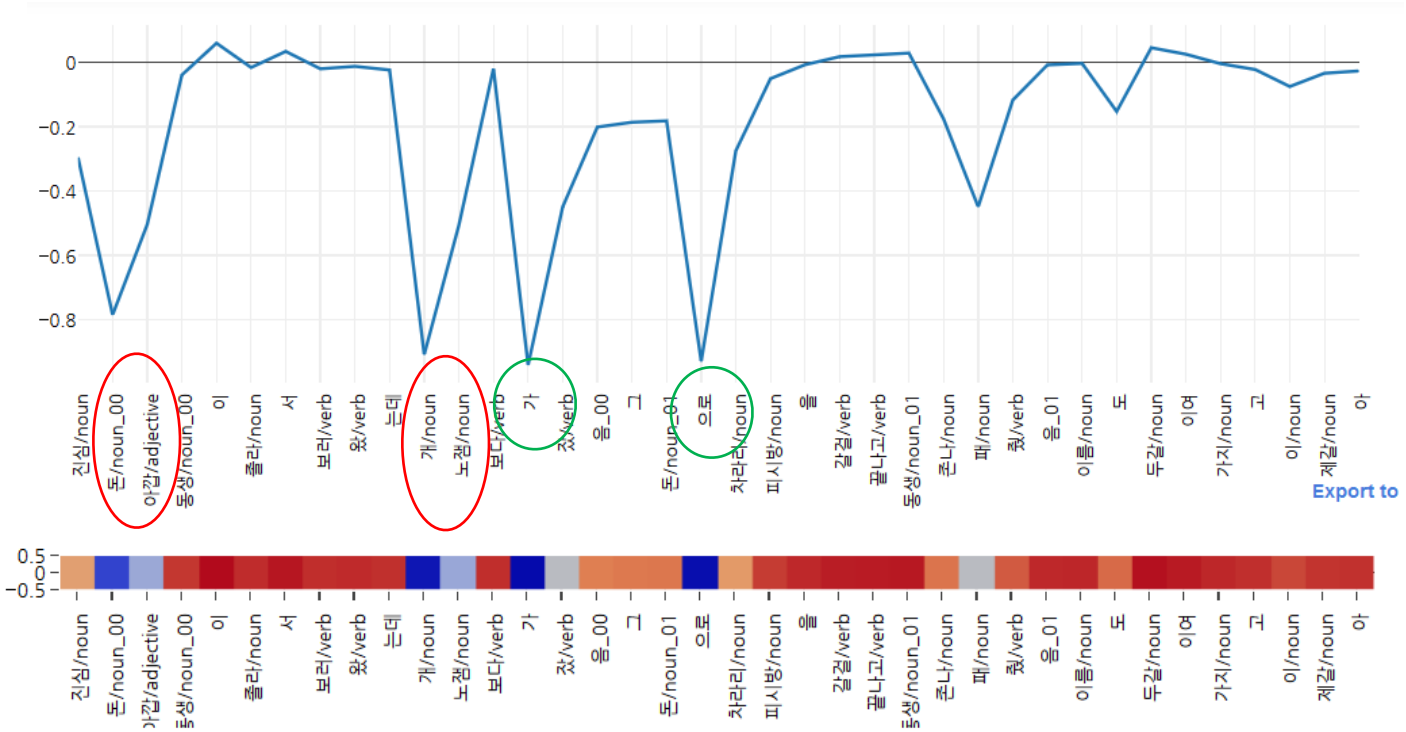
인접한 단어를 추출하기 어려움 → (대안) 관련 있는 단어 추출

All Words:
강 건조해서/adjective 노잼/noun 인데 보는/verb 내_00 내_01 쓸쓸하고/adjective 적막/noun 하고_00 고독/noun 하고_01 외/noun 롭고 그런/adjective 게 넘/verb 느껴진/verb 다



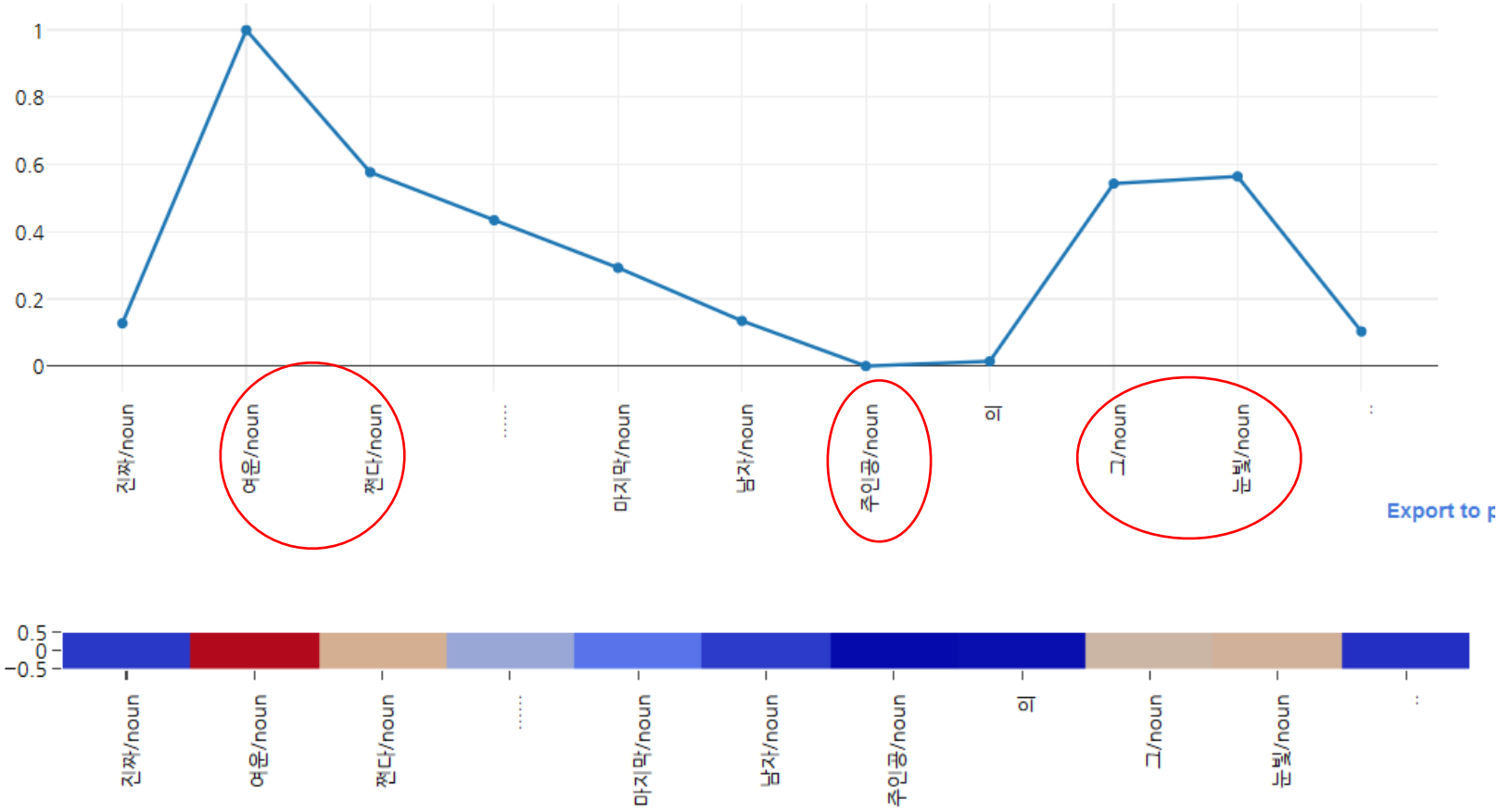
인접한 단어를 추출하기 어려움 → (대안) 관련 있는 단어 추출

All Words:
진심/noun 돈/noun_00 아깝/adjective 동생/noun_00 이 쫄라/noun 서 보러/verb 왔/verb 는데 개/noun 노잼/noun 보다/verb 가 잤/verb 음_00 그 돈/noun_01 으로 차라리/noun 피시방/noun 을 갈걸/verb 끝나고/verb 동생/noun_01 존나/noun 패/noun 줬/verb 음_01 이름/noun 도 두갈/noun 이여 가지/noun 고 이/noun 제갈/noun 아



인접한 단어를 추출하기 어려움 → (대안) 관련 있는 단어 추출

All Words:
진짜/noun 여운/noun 쩐다/noun ..... 마지막/noun 남자/noun 주인공/noun 의 그/noun 눈빛/noun ..



그 밖의 어려운 점

대부분은 이유가 명시되어 있지 않음

233: 정말/noun 노잼/noun 이고 죽여/verb 버리/verb 고 싶 다

246: 노잼/noun 노잼/noun 노잼/noun 노잼/noun 노잼/noun 노잼/noun

406: 존/noun 노잼/noun 존/noun 노잼/noun 존/noun 노잼/noun .

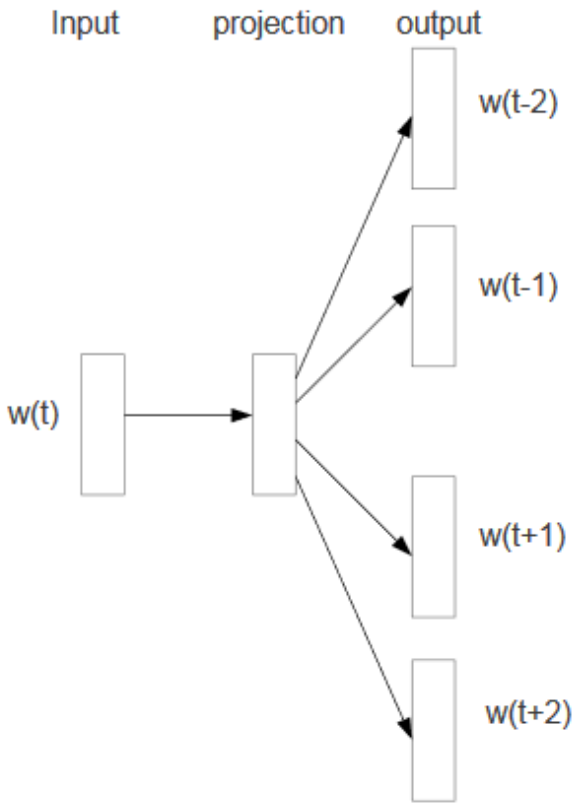
613: 노잼/noun 테드/noun 요 번년/noun 에 본영/noun 화중/noun 젤/noun 노잼/noun

[긍정단어 / 부정단어 사전 만들기]

Tfldf	Grad-Cam
영화/noun	재밋어/noun
정말/noun	쩐다/noun
너무/noun	으리/noun
최고/noun	최고/noun
감동/noun	순수/noun
진짜/noun	굿굿굿/noun
연기/noun	졸라/noun
드라마/noun	up
다시/noun	허니/noun
보고/noun	추강/noun
평점/noun	best
마지막/noun	짱임/noun
…	…

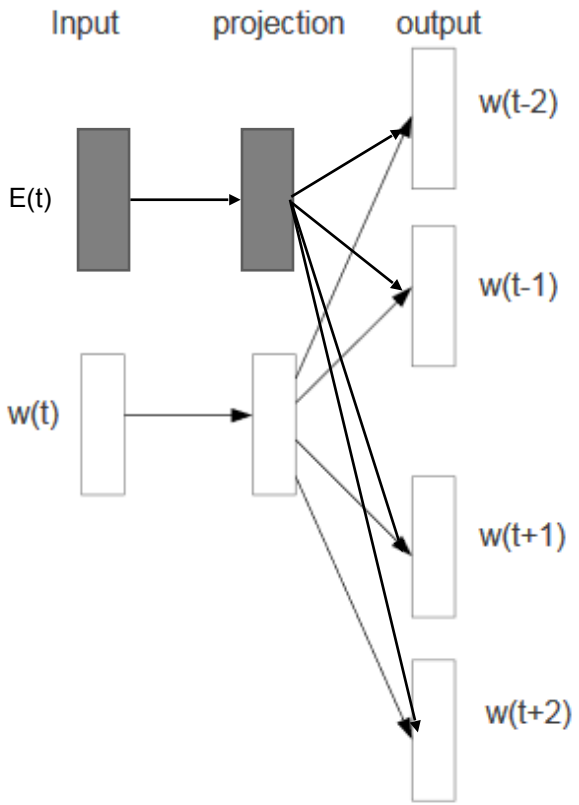
Tfldf	Grad-Cam
영화/noun	별로/noun
너무/noun	지루/noun
진짜/noun	ooo
평점/noun	노잼/noun
정말/noun	최악/noun
스토리/noun	쓰레기/noun
그냥/noun	어의/noun
쓰레기/noun	글쎄요/noun
감독/noun	밋밋하/noun
시간/noun	심해/noun
재미/noun	졸라/noun
내용/noun	증말/noun
…	…

Word2Vec 중 Skip-gram



$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0}\log p(w_{t+j}|w_t)$$

Target2Vec



E(t) = {0, 1, 2}

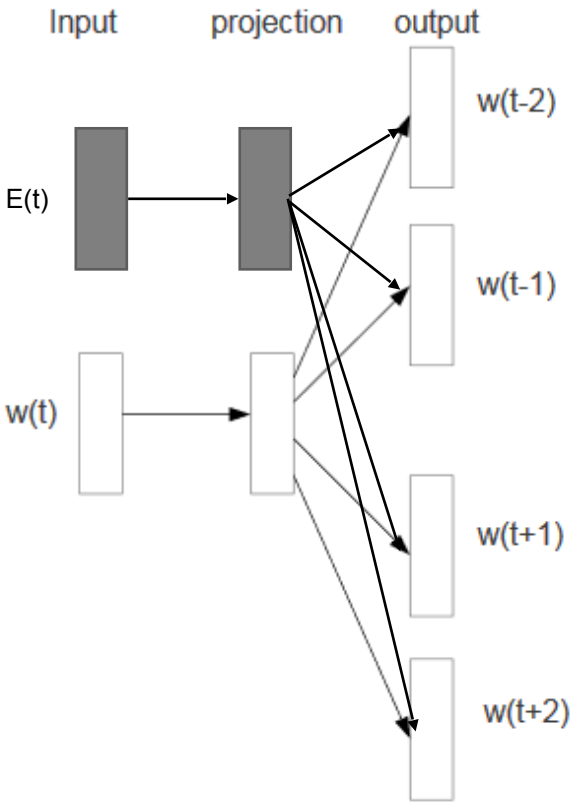
0: 부정
1: 긍정
2: 중립

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\left[\log p(w_{t+j}|w_t) + \log p(w_{t+j}|E_t)\right]$$

Target2Vec

- 긍정은 긍정끼리 잘 모이도록 부정은 부정끼리 잘 모이도록 제약 조건 하 임베딩
- King + Man - Woman = ? Queen → 최악 + 부정 – 긍정 = ? 최고
  (워드와 워드라벨간 벡터 연산)

- 부정적 문서들과 관련 있는 단어 = 네이버 영화평에서 부정적 표현



E(t) = {0, 1, 2}

0: 부정
1: 긍정
2: 중립

$$\frac{1}{T}\sum_{t=1}^{T}\sum_{-c\le j\le c, j\neq 0}\left(\log p(w_{t+j}|w_t)_+ \quad \log p(w_{t+j}|\ \epsilon_t)\right)$$

[결과 비교]

1. 일반적인 doc2vec
2. 부정문서만 대상으로 doc2vec (긍정문서만 대상으로 doc2vec)
3. Target2Vec (doc 버전)

[결과 요약]

1의 경우: 부정단어나 긍정단어의 유사 맥락(단어)이 잘 나타나지 않음

2의 경우: 부정 혹은 긍정 단어들은 각각 잘 나타남, 중립단어들은 풀 수 없음

3의 경우: '스토리'라는 중립 단어가 부정적 맥락에서 어떻게 사용되는지 알 수 있음

'최악/noun'과 유사단어

Doc2vec	Doc2Vec(N)	Target2Vec
'최악/noun '최고/noun '별로/noun '베스트/nou '내생/noun '노잼/noun '여신/noun '삼류/noun '최대/noun '최강/noun '연속/noun '구식/noun '생애/noun '극장판/nou '천원/noun '향연/noun '아오', '실수/noun '노답/noun	'최악/noun', '최고/noun', '내생/noun', '생애/noun', '역대/noun', '인생/noun', '극치/noun', '구식/noun', '최대/noun', '노답/noun', '망작/noun', '실수/noun', '극장판/noun' '은걸', '관람/noun', '오랜만/noun' '후회/noun', '제일/noun', '취향/noun',	'최악/noun', '생애/noun', '내생/noun', '최대/noun', '여태/noun', '졸작/noun', '역대/noun', '제일/noun', '가히', '가장/noun', '인생/noun', '지구/noun', '시리즈/noun' '최고/noun', '실수/noun', '이후/noun', '극장판/noun' 'cf', '망작/noun',

'최악/noun'의 맥락 단어들

Doc2vec	Doc2Vec(N)	Target2Vec
'임/noun', '였/verb', '중', '역대/noun', '내생/noun', '라고 '생애/noun', '베스트/noun' '사상/noun', '의 '완전/noun', '중/noun', '좀비/noun', '무비/noun', '작/noun', '노잼/noun', '쵝/noun', '입니/adject	'의', '중', '임/noun' '생애/noun '내생/noun '인생/noun '드라마/nou '내', '베스트/nou '가장/noun '최고/noun '입니/adje '역대/noun '내/noun' '정말/noun '이다', 0. '최대/noun '망작/noun	'중', '역대/noun' '생애/noun' '내생/noun' '중/noun', '가본/verb' '제일/noun' '가장/noun' '시리즈/noun '내/noun', '조선/noun' '젤/noun', '최악/noun' '여태/noun' '꼽/noun', '화중/noun' '근래/noun' '태어나서/ver

Target2Vec으로 할 수 있는 것?!

‘뻔한/adj’의 맥락 단어들

Doc2vec	Doc2Vec(N)	Target2Vec
'스토리/noun' '억지/noun', '결말/noun', '스런', '반전/noun', '연기력/noun' '대사/noun', '내용/noun', '실망/noun', '어설픈/noun' '코미디/noun' '지루하고/adje '스럽/adject '스러운', '장면/noun', '상황/noun', '줄거리/noun'	'억지/noun', '결말/noun', '어설픈/noun' '스럽/adject '너무나/noun' '반전/noun', '감동/noun', '스토리/noun' '실망/noun', '스러운', '지루함/noun' '눈물/noun', '극치/noun', '지루하고/adje '이야기/noun' '연출/noun', '대사/noun'	'전개/noun', '빈약/noun', '엉', 0.039 '라인/noun', '억지/noun', '성/noun', '구성/noun', '부실/noun', '플롯/noun', '마무리/noun "단순하고/adj '조잡/noun', '진행/noun', '진부/noun', '유치/noun’, '설정/noun’, '스토리/noun’

'이야기/noun’의 맥락 단어들

Doc2vec	Doc2Vec (N)	Target2Vec (N)	Target2Vec (G)
'와', '사랑/noun' '인', '가족/noun' '대한/noun' '독특한/adj '따뜻한/adj '아름다운/ adj '처럼', '뻔한/ad '않은/verb' '할/verb', '로서/noun' '요즘/noun' '다른/noun' '흔한/adjec '풋풋/noun'	'사랑/noun' '풀어/verb' '를', '꼴/noun', '복수/noun' '메세지/ noun '뻔한/adjec '감동/noun' '와', '방식/noun' '인', '나', '로', '흥미/noun' '미친/adjec '빈약/noun' '기',	'전개/*noun* '뻔한/*adj* '빈약/*noun* '진부/*noun* '탄탄/*noun* '유치/*noun* '단순하고/ *ad* '구성/*noun* '억지/*noun* '조잡/*noun* '뒤죽박죽 '부실/*noun* '라인/*noun* '플롯/*noun* "없는/*adj* '빠른/*adj* '어설픈*nou*	'따뜻해/adj '따뜻한/adj '아프/adj '아름다운/ adj '찡', '훈훈한/adj '면서도', '슬프/adj '잔잔/noun' '아름답/adj '믿음/noun' '순수한/adje '따뜻하고/ adj '풋풋/noun'

'부정단어들'과 유사단어 =
네이버 영화평에서 부정적 표현

Target2Vec
'반성/noun', '누가/noun', '——', 0.39 '알바생/noun' '베스트/noun' '동안/noun', '댓글/noun', '이모/noun', '우와', '낚시/noun', '저리/noun', '반개/noun', '별하나/noun' '독립영화/noun '최강/noun', '아무리', '알바/noun', '냄새/noun', '열자/noun', '로는',

' 긍정단어들'과 유사단어 =
네이버 영화평에서 긍정적 표현

Target2Vec
'남은/verb' '깊게/adjec '깊다/adjec '소중함/adje '지나도/verb '슬픈/adjec '완벽하다/adj '흘린/verb' '굉장한/adje '생/noun', '적/verb', '갑/noun', '안타까운/adj '감명/noun' '울컥', '마음/noun' '세라/noun' '함께', '방황/noun'

Quiz 1

최악 : 부정 = ? : 긍정

Quiz 1

최악 : 부정 = 베스트, 최고 : 긍정

Quiz 2

노잼 : 부정 = ? : 긍정

Quiz 2

노잼 : 부정 = 쩐다, 레전드 : 긍정

Quiz 3

뻔한 : 부정 = ? : 긍정

Quiz 3

뻔한 : 부정 = 반전 : 긍정

[오분류]

```
('0_1025', 0.5986143350601196),
('0_4527', 0.5860131978988647),
('1_3507', 0.5740727186203003),
('0_3997', 0.5421779155731201),
('0_1375', 0.5139888525009155),
('0_3835', 0.48261043429374695),
('0_2091', 0.4777252559661865),
('0_5647', 0.4672212600708008),
('0_4946', 0.466841459274292),
('0_123', 0.46322327852249146),
('0_4617', 0.45621258020401),
('0_3474', 0.45521730184555054),
('0_1266', 0.44517603516578674),
('0_2256', 0.4439569115638733),
('0_5652', 0.4407670795917511),
('0_1936', 0.43370863795280457),
('0_4551', 0.4332246780395508),
('0_2746', 0.4312360882759094),
('0_4668', 0.4311937093734741),
('0_3909', 0.42830002307891846),
('0_3306', 0.4279792308807373),
('0_5409', 0.42482253909111023),
('0_1400', 0.4200928807258606),
('0_2804', 0.41977816820144653),
('0_990', 0.4177834391593933),
```

<문서 차원>

1_3507: ['평점/noun 이 왜/noun 이리 낮/adjective 음 ——
베스트/noun 오브/noun 베스트/noun 이구만']

<단어 차원>

['재밌다', '재밌네'] 긍정 부정 둘 다 출현하여

→ 중립단어로 분류

그래서 챗봇은?

[용가리] 영화 [재미 없었어]

- [킬링타임, 그래픽, 아쉬움] 어떤 부분이?

- [그래픽]

- 그랬구나. [시나리오, 연출] 호평이 많은 AA는 어때?

[용가리 + 노잼] → [킬링타임, 그래픽, 아쉬움] ? → ? [시나리오, 연출] → [AA]

언젠가…

[코멘트]

- 영화평 당 관련 구문 추출 (합산이 아니라)

- Poincare Embedding

- Acoustic Embedding