고려대학교
빅데이터 연구회

KU-BIG

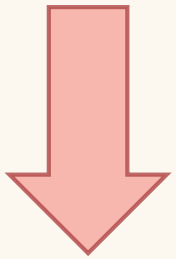
대나무숲 공감 AI

CCP team : 김강우 문선미 박기찬 이세희

Index

주제 선정 배경

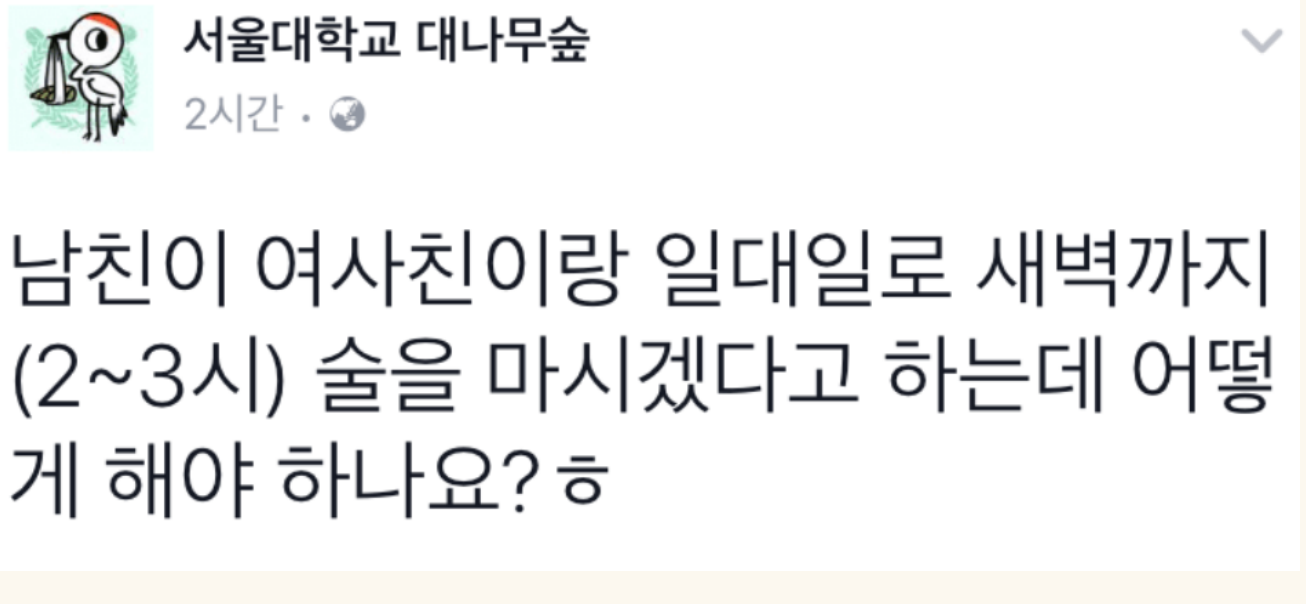
MIT, Generative Model로 논문을 작성해, 학술지에 싣는 데에 성공



generative model로 대나무숲 글을 작성해 대나무 숲에 제보해 보자!

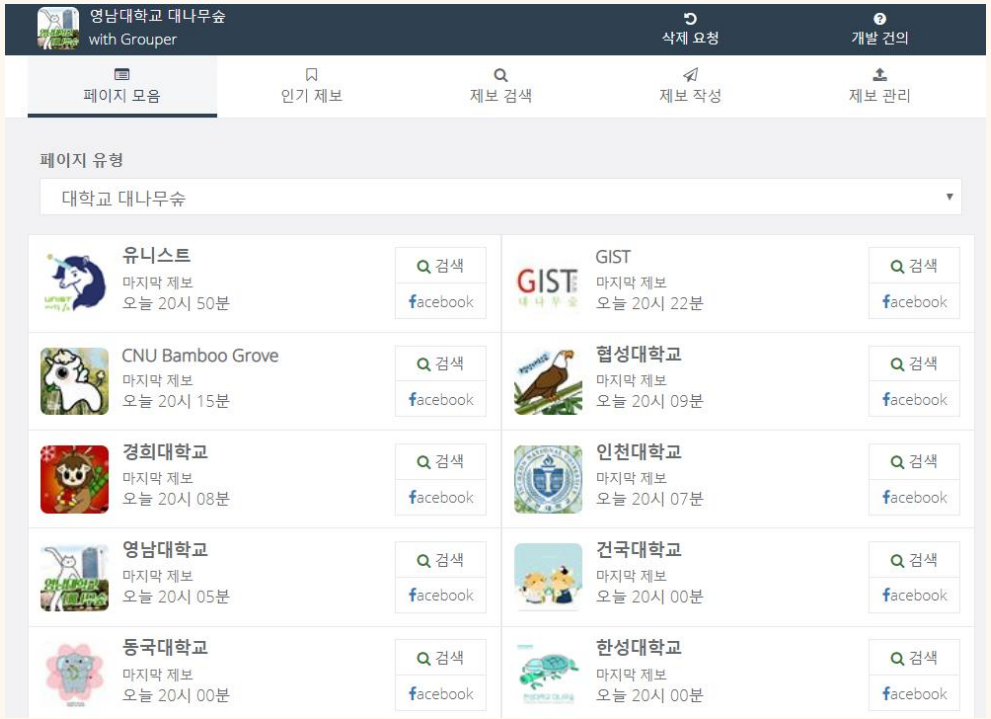
데이터 소개

대학생 공감의 대명사 -> 대나무숲



데이터 수집

From 대나무숲 Grouper, by Selenium



날짜별 추출!

4개의 대학교 고려대, 연세대, 성균관대, 경희대

120000여개의 문서

데이터 전처리

정규 표현식

#고려대숲/r/n/r/n대하!!..

#고대표효/r/n2015.01.01/r/n대숲!!

#연대숲/r/n/r/n2014.01.12/r/n/r/n대숲!..

#경희대체대의 글/n/n # 운동 # 학교/n/n...

형래야 좋아해.

학교마다 다른 패턴을 고려해 불필요한 여백, 기호, 의성어들을 제거하는

“전처리 자동화 시스템” 구축

데이터 전처리

정규 표현식

```
cleaning('연대', '\r\n\r\n',data)
```

1437	제보시간 지워주세요. 기숙사는. 혼자서 펑펑 울 데가 없어요. 방에도 룸메가있구. ...
1438	새내기 여학우에요. 있잖아요 저는 고등학교때 진짜 자타공인 또 이 였어요 완전활발하...
1439	제보시간 지워주세요. 인간적으로 엘레베이터 앞으로 뛰어가는 사람 있으면 엘레베이터 ...
1440	그 사람을 좋아해요. 한 번도 이런 적 없었는데 정말 열렬히. 처음엔 봄 타느라 외...
1441	제보시간 지워주세요. 옷 위로 남자 젖꼭지 튀어나온거 보이면 어떤가요? 신경써서 안...
1442	익명 제보시간 지워주세요. 초등학교 4학년때부터 혼자서 자던 송도러입니다. 어릴때부...
1443	제보시간 지워주세요. 학교 다니는 4년동안 주위 사람들. 특히 여자애들의 이성문제를...
1444	남자친구네 아버지가 알고보니 저희 아파트 옆 아파트에 경비아저씨였어요. 5년이란 시...
1445	제보시간 지워주시면 감사하겠습니다. 여자친구하고 이야기를 할 때 가끔씩 여자친구의 ...

데이터 전처리 - 토큰화

정규 표현식

단어 단위가 아닌 형태소 단위 토큰화 필요

형태소 분석기 **Mecab in Konlpy** 사용

Example)

흔한 연대생들의 연대 -> '흔한',연대','생','들','의','연대' <- 띄어쓰기는 따로 고려

띄어쓰기도 하나의 형태소로 인지시키기 위해 **공백을 특수문자 &로 변경**

->'흔한','&','연대','생','들','의','&','연대'

데이터 전처리 – 포스트 선정

학습 용이성을 고려하여 학습에 사용할 일부 포스트 선정
120000여개 -> 30000개

제외 기준

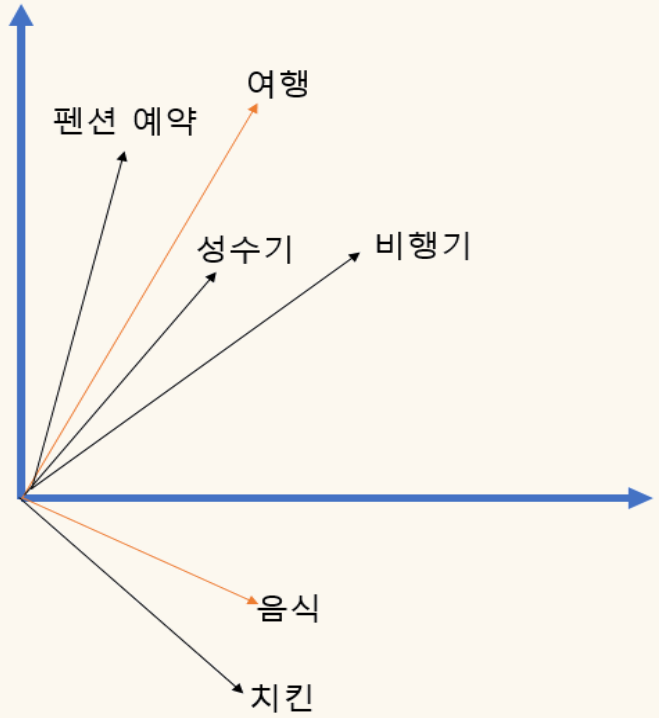
한 포스트에 여러 글이 담긴 경우 제거

전체 문서에서 3번 이하로 나온 형태소를 포함한 문서 제거

좋아요 수 기준 상위 30000개의 포스트를 선정
좋아요가 많은 포스트일수록, 공감할 수 있는 주제를 다루었거나 온전한 문장으로 구성되어 있을 확률이 높다고 판단

데이터 전처리

Word Embedding by Word2Vec



유사한 단어, 관련 있는 단어

⇨ 벡터공간 상 가까운 곳에 위치

⇨ 높은 코사인 유사도

데이터 전처리

Word Embedding by Word2Vec

'여친'입력 시

1.000000	남친
0.890239	여친
0.727867	애인
0.625837	사친
0.608522	썸남
0.604570	뺐기
0.596657	남자
0.592809	친구
0.581338	게이머
0.552695	독설

'남친'입력 시

1.000000	여친
0.890239	남친
0.770095	애인
0.612826	사친
0.592885	나마나
0.581850	썸남
0.565032	게이머
0.560651	뺐기
0.557853	첫사랑
0.554026	자이로

Most.similarity를 통해 유사한,

관련도가 높은 단어 추출

= 코사인 유사도가 높은 단어 추출

= 벡터공간상 근접한 단어 추출

남친 & 여친 벡터 공간상 아주 가까이 위치

-> 유사 단어 리스트가 거의 동일

추출된 리스트를 보았을때, 임베딩이 성공적으로 이루어졌음을 확인 할 수 있다.

데이터 전처리

Word Embedding by Word2Vec

딕셔너리	인덱싱	벡터 리스트
'&'	'&' : 0	[-0.15168379, 0.1794301 ...]
'.'	'.' : 1	[-0.07421526, 0.02408372 ...]
'들'	'들' : 2	[0.02735069, 0.11275656, ...]
'생'	'생' : 3	[-0.10620534, -0.14361914, ...]
'연대'	'연대' : 4	[-0.26902023, 0.27805564, ...]
'의'	'의' : 5	[-0.01524859, 0.12166952, ...]
'흔한'	'흔한' : 6	[0.14256145, 0.00432931, ...]
'10'	'10' : 7	[-0.27286306, -0.09492918, ...]
'100'	'100' : 8	[-0.07621562, 0.08695877, ...]
'2'	'2' : 9	[0.10022658, 0.03568741, ...]
'200'	'200' : 10	[-0.10620534, -0.14361914, ...]

Generator Model (SeqGAN)

세희는 오늘 학교에 갔다.

토큰화

'세희','는','&', '오늘','& ','학교','에','&', '갔다.'

Indexing

& : 1
세희 : 30
는 : 2
오늘 : 8
학교: 11
에: 76
갔다: 5

SeqGAN Input

(30,2,1,8,1,11,76,5)

SeqGAN Output

(30,2,1,8,1,11,76,5)

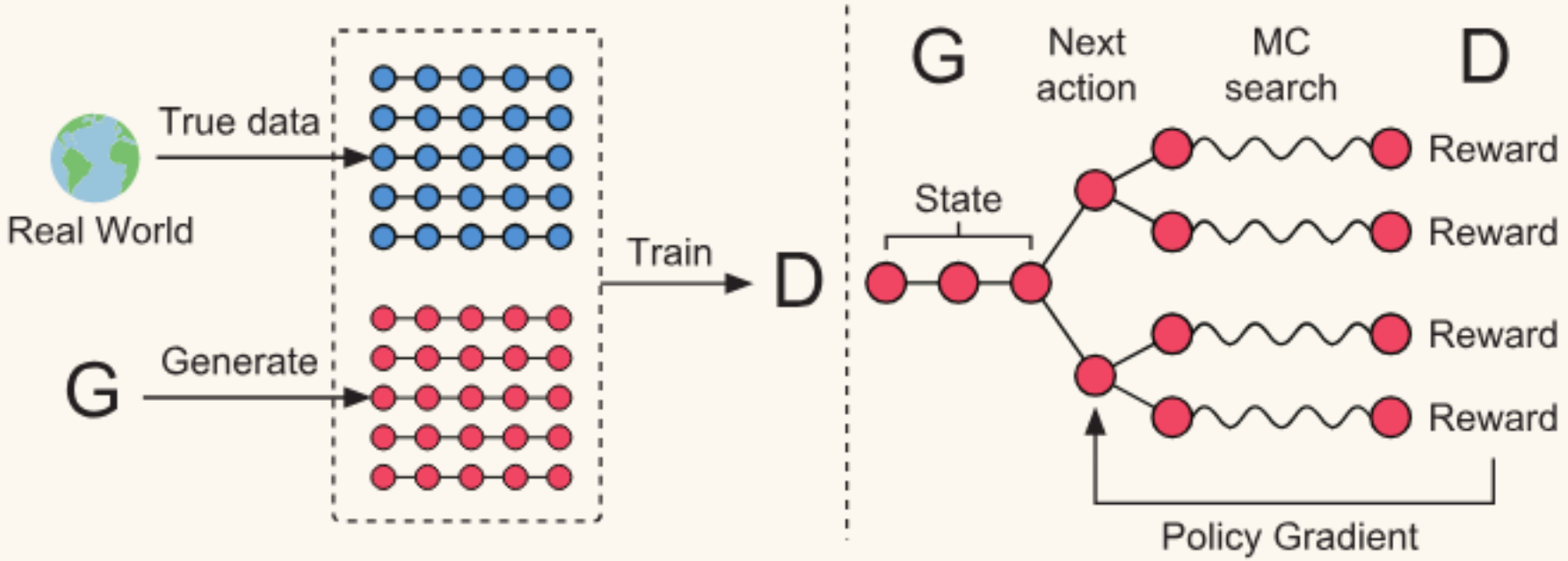
& : 1
세희 : 30
는 : 2
오늘 : 8
학교: 11
에: 76
갔다: 5

'세희','는','&', '오늘','& ','학교','에','&', '갔다.'

Generator Model (SeqGAN)

[Generator]
EMB_DIM = 100
SEQ_LENGTH = 20

[Discriminator]
dis_embedding_dim = 64
dis_filter_sizes = [1, 2, 3, 4, 5, 6, 7, 8, 9]
dis_num_filters = [100, 200, 200, 200, 200, 100,100,100,100]



Thank You