



Altist

「 함께 」



이승윤



이지현



성민경



장민석

목차

1 데이터 수집

2 Fine-Tuning Strategy

3 Custom Method

4 최종 결과

Part 1.

데이터 수집

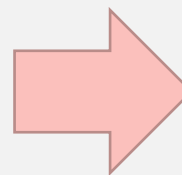


1

데이터 수집



- 수필 관련 콘텐츠를 중심으로 탐색
- Python의 다양한 라이브러리를 이용한 크롤링
- 약 20M 분량의 텍스트 데이터를 수집



원시 데이터

1 데이터 수집

■ 한국 산문작가협회 – 수필공모

[\(주\)한국산문](#)
[월간한국산문](#)
[문학회](#)
[강의실](#)
[누구나참여광장](#)
[자유게시판](#)
[문학상공모안내](#)

로그인

아이디

비밀번호

로그인

회원가입

아이디/비밀번호찾기

누구나참여광장

수필공모

문학정보

Home > 누구나참여광장 > 수필공모

수필공모

비동단자를 위해 열린 공간입니다. 월간 한국산문으로 동단을 원하는 분들의 글을 환영합니다.

전체게시물 391

번호	제 목	글쓴이	날짜	조회
공지	★★수필 응모하는 분들은 꼭 읽어보세요 (4)	웹지기	05-15	61727
391	아홉 살과 예순(수정본2) (3)	고돌선	11-10	1143
390	아홉 살과 예순 (수정본) (2)	고돌선	11-03	1752
389	아홉 살과 예순 (2)	고돌선	10-04	2212
388	구름 (수정본) (2)	고돌선	09-15	2452
387	구름 (2)	고돌선	09-01	2659

■ 글틴 – 중.고등학생 수필 및 공모작

문학광장

[문장 웹진](#)
[문장의 소리](#)
[문학집배원](#)
[문학IN아르코](#)
[글틴](#)
[마로니에백일장](#)
[문장공지](#)
[사이버문학관](#)
[회원마당](#)

글틴 > 쓰면서 뒀글 > 수필

수필

문부일

[공지] 10월 장원 발표+ 함께 보고 싶은 시트콤 영상 2021-11-06

안녕하세요. 11월이 시작되었습니다. 내일은 입동이래고 하네요. 곧 겨울로 접어드는데, 오늘은 포근하고 햇살도 좋습니다. 이렇게 늦가을이 흘러가고 있습니다. 여러분, 잘 지내시나요? 정확히 12일이 지나면 수능이 옵니다. 저는 공부를 잘하겠다는 마음이 1도 없었으니, 당연히 수능 부담이 전혀 없었습니다. 수능 날, 정말 마음이 편했습니다. 수능만 보면 이제 해방이구나! 이제 어른이구나! 이렇게 환호하며 시험장에 갔던 기억이!'' 여러분들도 원하는 점수 얻고 수능에서 해방되시길 바랍니다!'' 저는 최근에 청소년문학 단편집 수정을 끝냈고, 장편 초고 수정도 마쳤습니다. 물론 장편은 또 수정을 해야겠지만 일단 당분간 제 손을 떠나서 정말 출가분하네요. 왜 글을 써야할까, 이런 망상을 하면서 계속하면서!'' 10월에[...]

멘토

[공지] 9월 장원 발표 및 함께 읽고 싶은 책 추천! 2021-10-05

안녕하세요. 어느덧 가을이 왔습니다. 환절기와 감기 걸리기 쉬운 계절인데 다들 건강하신가요? 저는 8월에 썼던 장편 초고를 다시 수정하기 시작했습니다. 출간 날짜를 맞출 수 있을까, 고민하면서! 역시 글쓰기는 너무 어렵네요. 이 세상에는 의미있는 일이 참 많은데 왜 나는 글을 쓰고 있을까? 혼자 망상, 공상, 상상을 하면서 수정하고 있습니다!'' 여러분들은 왜 글을 쓰

데이터 수집

■ 브런치 - 작가별 수집

≡

brunch

시작하기

Q

글이 작품이 되는 공간, 브런치

④

브런치에 담긴 아름다운 작품을 감상해 보세요.


그리고 다시 꺼내 보세요.

서랍 속 간직하고 있는 글과 감성을.


Update

브런치 작가님과 커커오톡의 수많은 독자가 연결되는 기회


간파인



소고기 미역국을 끓이다가




■ 재미수필문학가협회 - 추천 수필



재미수필문학가협회

Korean Essayist Association of America



홈
협회소개
협회게시판
동네방 안내
재미수필집
퓨전수필
문예광장
공부합시다
회원소식
회원서제
커뮤니티
세미나

☐ Keep me signed in.
[Sign Up](#) | Find Account Info
Request for Activation Mail


[▶ 문예광장 > 추천 수필](#)

No.	Subject	Author	Date	Views
Articles 840				
Notice	디카에세이 우산·최창순 수필가	정조연	Jan 06, 2021	307
Notice	수필가 반숙자 초기작품 수필집 <몸으로 우는 사과나무> 80편	admin	Mar 16, 2016	10765
840	건널목 / 정선모	정조연	Nov 13, 2021	37
839	오래된 편지 / 권남희	정조연	Nov 13, 2021	38
838	은이 / 김정섭	정조연	Nov 13, 2021	25
837	슴베 / 배재록 - 제8회 경북일보 청송객주문학대전 금상	정조연	Nov 13, 2021	30


today: 430

yesterday: 682

Total: 769,273



등업신청



공지사항

제23집 재미수필 출판기념회....

2021년 수입&지출 보고서...

Part 2.

Fine-Tuning Strategy



2 데이터 정제 - <1>

```
def preprocess(data, column='contents'):
    data=data.dropna(subset=[column]).reset_index(drop=True)

    for i in range(len(data)):
        data.at[i,column]=re.sub('\\\\d{4}\\\\#.*\\\\d{1,2}\\\\#.*\\\\d{1,2}\\\\#.*',' ',data[column][i]) # 날짜형식 제거
        data.at[i,column]=re.sub('<\\\\\\\\\\\\>|<\\\\\\\\\\\\>|\\\\\\\\\\\\>-\\\\\\\\\\\\>|<*\\\\\\\\\\\\>|<*\\\\\\\\\\\\>', '',data[column][i]) # 끝 제거
        data.at[i,column]=re.sub('[a-z0-9]+@[a-z]+\\\\.[a-z]+[a-z\\\\.]*', '',data[column][i]) # 이메일 양식 제거


        data.at[i,column]=data[column][i].replace('\\\\u200b','').replace('\\\\xa0','').replace('\\\\ufeff','').replace('\\\\u3000','') # 유니코드 제거
        data.at[i,column]=re.sub('\\\\([\\\\\\\\(\\\\\\\\)+\\\\\\\\)\\\\\\\\[. +\\\\\\\\]', ' ',data[column][i]) # (), [] 안의 내용제거
        data.at[i,column]=re.sub('_\\\\\\\\\\\\|\\\\\\\\+|=|\\\\\\\\^|;', ' ',data[column][i]) # 특수기호 제거
        data.at[i,column]=re.sub('--+|--+', '-',data[column][i]) # --- 대체
        data.at[i,column]=re.sub('[ㄱ-ㅎㅏ-ㅣ]+', ' ',data[column][i]) # 자모음 제거
        data.at[i,column]=re.sub('[--龀]', ' ',data[column][i]) # 한자제거
        data.at[i,column]=re.sub('!+', '!',data[column][i]) # ! 중복제거
        data.at[i,column]=re.sub('\\\\?+', '?',data[column][i]) # ? 중복제거
        data.at[i,column]=re.sub('~+', '~',data[column][i]) # ~ 중복제거


        data.at[i,column]=re.sub('(\\\\n)+|(\\\\r)+(\\\\t)+', '\\\\n',data[column][i])
        data.at[i,column]=re.sub('\\\\n', ' ',data[column][i])
        data.at[i,column]=re.sub(r'\\\\\\\\\\\\\\\\', ' ',data[column][i])
        data.at[i,column]=re.sub('.', ' ',data[column][i])
        data.at[i,column]=re.sub('(\\\\.){3,}|(\\\\\\\\.+)|\\\\\\\\\\\\\\\\+|\\\\\\\\\\\\\\\\+|\\\\\\\\\\\\\\\\+|\\\\\\\\\\\\\\\\+|\\\\\\\\\\\\\\\\+', '...',data[column][i]) # ...대체
        data.at[i,column]=re.sub('+', ' ',data[column][i])
        data.at[i,column]=data[column][i].strip()

    return data
```

정규표현식

- 특수 기호, 한자, 과도한 중복 제거
- 괄호 및 내부 내용 제거
- 이메일 등의 특수 양식 제거
- 연속 구두점들의 규칙 통일

2

데이터 정제 - <2>

	AB_C Column1	AB_C Column2	AB_C Column3	AB_C Column4
1	refer	types	title	contents
2	https://brunch.co.kr/@seanpyo/176	브런치 에세이 수필	입장료 4만 5천원. 승봉도 이름 ...	허겁지겁 대체할 곳을 검색해 보았으나 알려진 근교 캠핑시설은 ...
3	https://brunch.co.kr/@seanpyo/175	브런치 에세이 수필	여름, 에어컨을 대신하는 것들	우리에게는 그저 복잡한 도심과 일상, 더위를 피해 잠시 떠난 도피...
4	https://brunch.co.kr/@seanpyo/171	브런치 에세이 수필	하늘로 오르다' 제주 오름의 정석...	잠시 쉬어가는 선글러 아담한 카페, 주안에게 근처 오름을 하나 후...
5	https://brunch.co.kr/@seanpyo/159	브런치 에세이 수필	꿈을 꾸듯 시탕	'딩그러니' 우리는 꿈의 한 조각을 더듬는 듯, 잠결에 멍하게 서 있...
6	https://brunch.co.kr/@seanpyo/158	브런치 에세이 수필	특별한 여행을 꿈꾸는 당신에게	몽골 여행은 도시여행보다 안전하고 패키지여행보다 편하다. 언...
7	https://brunch.co.kr/@seanpyo/157	브런치 에세이 수필	몽골 최고의 풍경은 고비도 흡스...	차 한 대 지날 정도의 좁은 길은 높은 언덕으로 이어져 있었다. 경...
8	https://brunch.co.kr/@seanpyo/149	브런치 에세이 수필	몽골 초원에서 차가 고장 나면 생...	바퀴가 진흙에 빠지거나 펑크가 나기도 하고 낡은 자동차로 비포...
9	https://brunch.co.kr/@seanpyo/134	브런치 에세이 수필	피렌체 두오모 보다 미켈란젤로 ...	좁고 긴 나선형 계단을 올라 만나는 피렌체의 풍경은 더없이 아름...
10	https://brunch.co.kr/@seanpyo/125	브런치 에세이 수필	날씨 좋은 5월 아이와 서울 거닐기	엄마와는 충분한 시간을 보내지만 아빠와의 시간은 늘 부족한 아...
11	https://brunch.co.kr/@seanpyo/124	브런치 에세이 수필	No News, No Shoes 몰디브의 매력	갤러리, 공원의 티켓을 미리 예약하고 끼니마다 식당을 정한다. 대...
12	https://brunch.co.kr/@seanpyo/121	브런치 에세이 수필	섬진강을 따라 두근두근 봄이 옵...	하지만 우리가 일상을 보내는 도시에서는 계절의 변화에 두근거림...
13	https://brunch.co.kr/@seanpyo/110	브런치 에세이 수필	도시에서 태어난 당신이 지구인...	점점 더 편해지고, 편리하고, 아무것도 하지 않아도 되는 삶을 무의...
14	https://brunch.co.kr/@seanpyo/87	브런치 에세이 수필	겨울 캠핑, 1살 먹으러 떠납니다.	특히 이번에는 한해의 마지막 일몰과 새해의 첫 일출을 맞이한...
15	https://brunch.co.kr/@seanpyo/80	브런치 에세이 수필	몰디브의 매력, '워터빌라'가 특...	여행의 시작, 이후로 펼쳐질 것을 누릴 시간이 충분하다는 것만으...
16	https://brunch.co.kr/@seanpyo/78	브런치 에세이 수필	몰디브 다가서기, 인천에서 제티...	우기의 끝자락 10월, 하늘은 희뿌옇고 이따금 빗방울이 떨어질 때...
17	https://brunch.co.kr/@seanpyo/77	브런치 에세이 수필	자연으로의 동행 두근두근 몽골...	두 번째는 휴양이다 일상에 지친 몸과 마음을 재충전할 수 있다. 세...
18	https://brunch.co.kr/@seanpyo/66	브런치 에세이 수필	M50 모간산루의 숨은 매력 찾기	실은 모간산루는 이름난 관광지가 아니라 예술가들이 모여서 작업...
19	https://brunch.co.kr/@seanpyo/43	브런치 에세이 수필	6살+아빠의 여행준비	여행의 목적지를 정하는 것부터 여행을 별 탈 없이 보내기 위한 준...
20	https://brunch.co.kr/@seanpyo/38	브런치 에세이 수필	프롤로그	그런 아버지와는 아주 거리가 멀다. 하지만 지금보다 더 좋은 아버...
21	https://brunch.co.kr/@seanpyo/25	브런치 에세이 수필	사라져가는 여행의 전리품들	도시를, 골목을 지날 때마다 갖고 싶은 모든 것을 주워 담고 싶지만...

2 Data collector

```
def sent_tokenizer(data,max_len):
    final_list=[]
    for contents in tqdm(data['contents']):
        sents = split_text_into_sentences(contents)
        sent_len = len(sents)
        if sent_len > 3:
            len_i=[0]
            while sent_len-(sum(len_i)+1)!=0:
                imsi=[]
                sent=[]
                for i in sents[sum(len_i):]:
                    if sum(imsi)<max_len:
                        sent.append(i)
                        imsi.append(len(mecab.morphs(i)))
                    else:
                        pass
                imsi.pop()
                sent.pop()

            final_list.append(' '.join(sent))
            len_i.append(len(imsi))
    return final_list
```

1

문장 단위로 분리

2

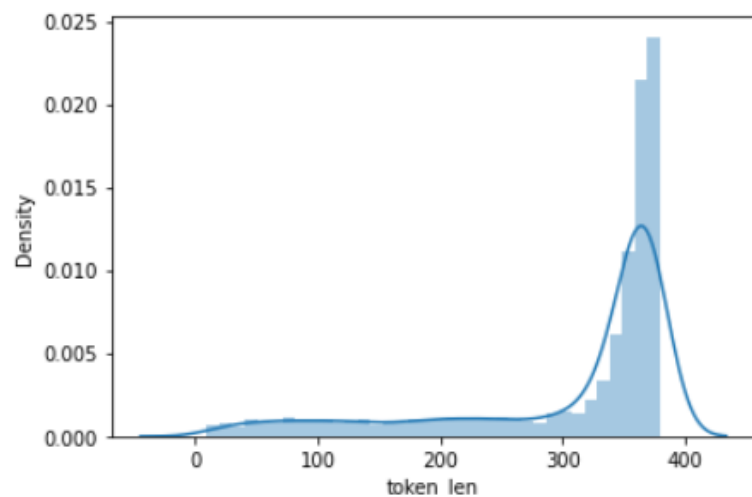
max_len 보다 작을때,
포함 가능한 최대
문장까지 추출

3

Join 으로 나뉜 문장들을 연결

최대한 문맥을 보존하며,
각 데이터 파일이
유사한 문장 길이를 갖도록
데이터를 효율적으로 분할

=> 통일된 학습 SIGNAL주입



2

모델 학습 전략 - <1>

모델 V1

필터링 전 데이터(약 17MB)

적은 step (최대 20000 step)

최대 입력 토큰 : 512

Learning rate decay 미적용

데이터 양

주제, 문체에 기반한
직접 필터링



명확한 문장구조를
위해 많은 step 학습

VS

모델 V2

필터링 후 데이터(약 7MB)

많은 step (최대 35000 step)

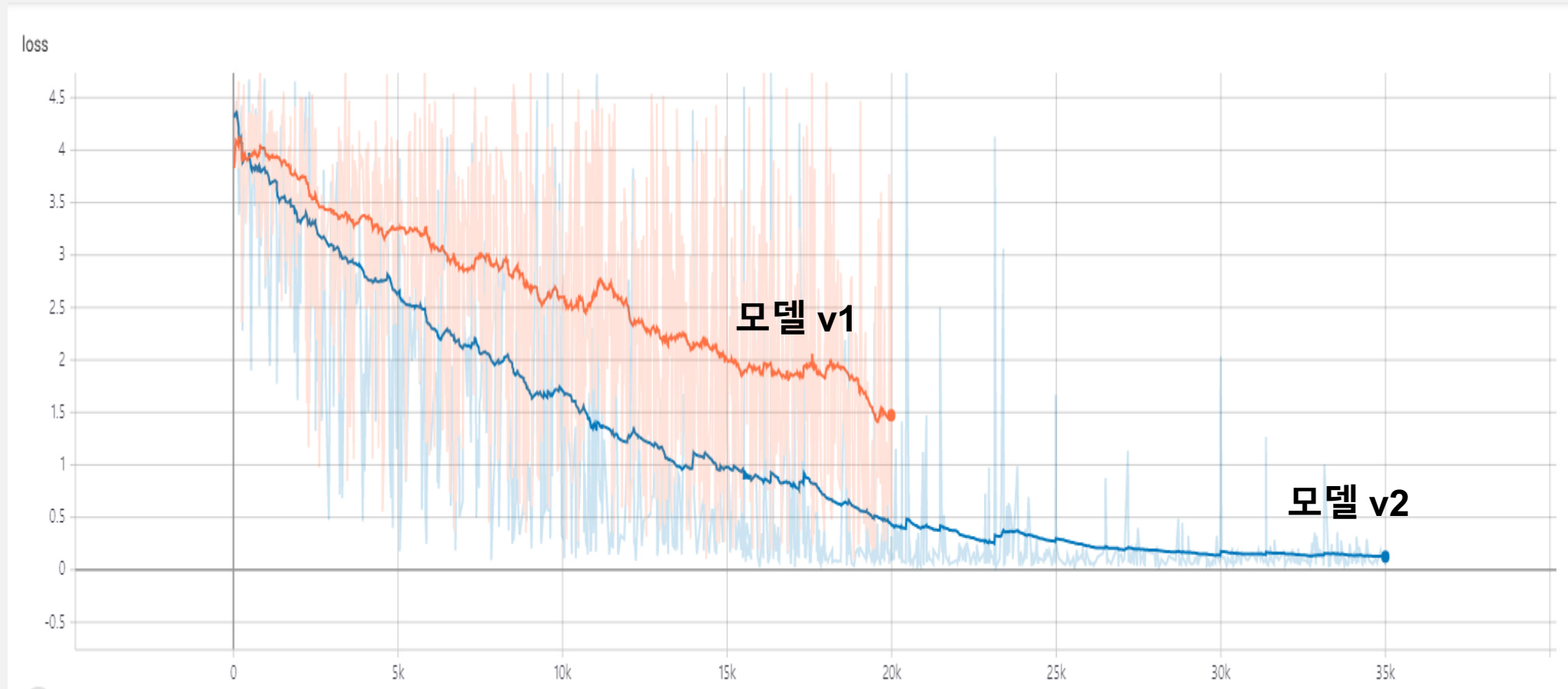
최대 입력 토큰 : 384

Learning rate decay 적용

데이터 질

2

모델 학습 전략 - <2>



Part 3.

Custom Method



3 모델 Customizing

Data Feeder

문단 연결성의 손실을 막는 Feeder 함수 구현

```
def _sent_custom_append(self, text, length):
    sentences = split_text_into_sentences(text)
    sentences_tokenized = [self.tokenizer.tokenize(sent) for sent in sentences if sent is not '']

    begin_index = 0

    selected_sents = sentences_tokenized[begin_index:]

    token_ids = self.tokenizer.convert_tokens_to_ids([token for sent in selected_sents for token in sent])
    while len(token_ids) < length:
        appending_text = random.choice(self.text_files).read_text(encoding='utf-8')
        token_ids += self.tokenizer.convert_tokens_to_ids(self.tokenizer.tokenize(appending_text))
    return token_ids[:length]
```

3 모델 Customizing

Data Sampler

Top-p 방식과 Top-k 방식을 동시에 적용하는 Sampler 함수 구현

```
def top_p_top_k_logits(logits, p, k):  
    with tf.variable_scope('top_p_top_k_logits'):  
        logits_sort = tf.sort(logits, direction='DESCENDING')  
        probs_sort = tf.nn.softmax(logits_sort)  
        indices = tf.constant(np.tile(np.arange(logits.shape[1].value), (logits.shape[0].value, 1)))  
        probs_sums = tf.cumsum(probs_sort, axis=1, exclusive=True)  
        logits_masked = tf.where((probs_sums < p) & (indices < k), logits_sort, tf.ones_like(logits_sort)*1000) # [batchsize, vocab]  
        min_logits = tf.reduce_min(logits_masked, axis=1, keepdims=True) # [batchsize, 1]  
  
        return tf.where(  
            logits < min_logits,  
            tf.ones_like(logits, dtype=logits.dtype) * -1e10,  
            logits,  
        )
```


Part 4.

최종 결과





A photograph of a paved path in a park, lined with cherry blossom trees in full bloom. The path leads into the distance, and the trees are covered in pink blossoms. The image is slightly dimmed to make the text stand out.

인연(人然)

4 소설 흐름

“자연과 함께 성장하는 인간”

기

자연과의 교감

승

봄과 여름의
아름다움

전

가을에 마주한
고독, 고통

결

타인과의 교감으로
극복과 성장

4

소설 흐름

봄과 여름의 아름다움

여름은 희망의 불씨가 되어 세상을 향해
뻗어나가는 것 같아 참으로 아름답다.

타인과의 교감으로 극복과 성장

마치 겨울이 지나가는 것처럼. 봄별이 흑독
해도 단 한 번도 눈이 오지 않을 봄도 있다.
그러나 그럴 때마다 설렘으로 서툰 마음을
조금이나마 달래주는 것은 봄이다.

승

기

자연과 교감

창을 통해 밖을 보니 오후의 햇살이
저만큼 투명하다. 창을 통해 비추는
햇살이 내 손을 잡았다.

결

전

가을에 마주한 고독, 고통

추운 가을에 홀로 지내는 외로움은 나의 우울함이다.
이 가을, 홀로 떠난 여행은 슬픔으로 다가온다.

4

Inference 문장

기승전결 문장은 아니지만, 문학적으로 예쁜 문장

동행의 언어에는 삶의 이야기와 한 장의 문장이 담겨 있다.

창을 통해 비추는 햇살이 내 손을 잡았다. 싹이 돋은 솜털 같았다.

봄의 산은 우리를 두근거리게 하고 그에 스며든 신록의 초록빛은 희망과 온기를 안겨 준다.

창을 통해 비친 푸른 바다와 눈부시게 비치는 햇살은 우리를 새로운 환상의 세계로 스며들게 하는 듯하다.

감사합니다