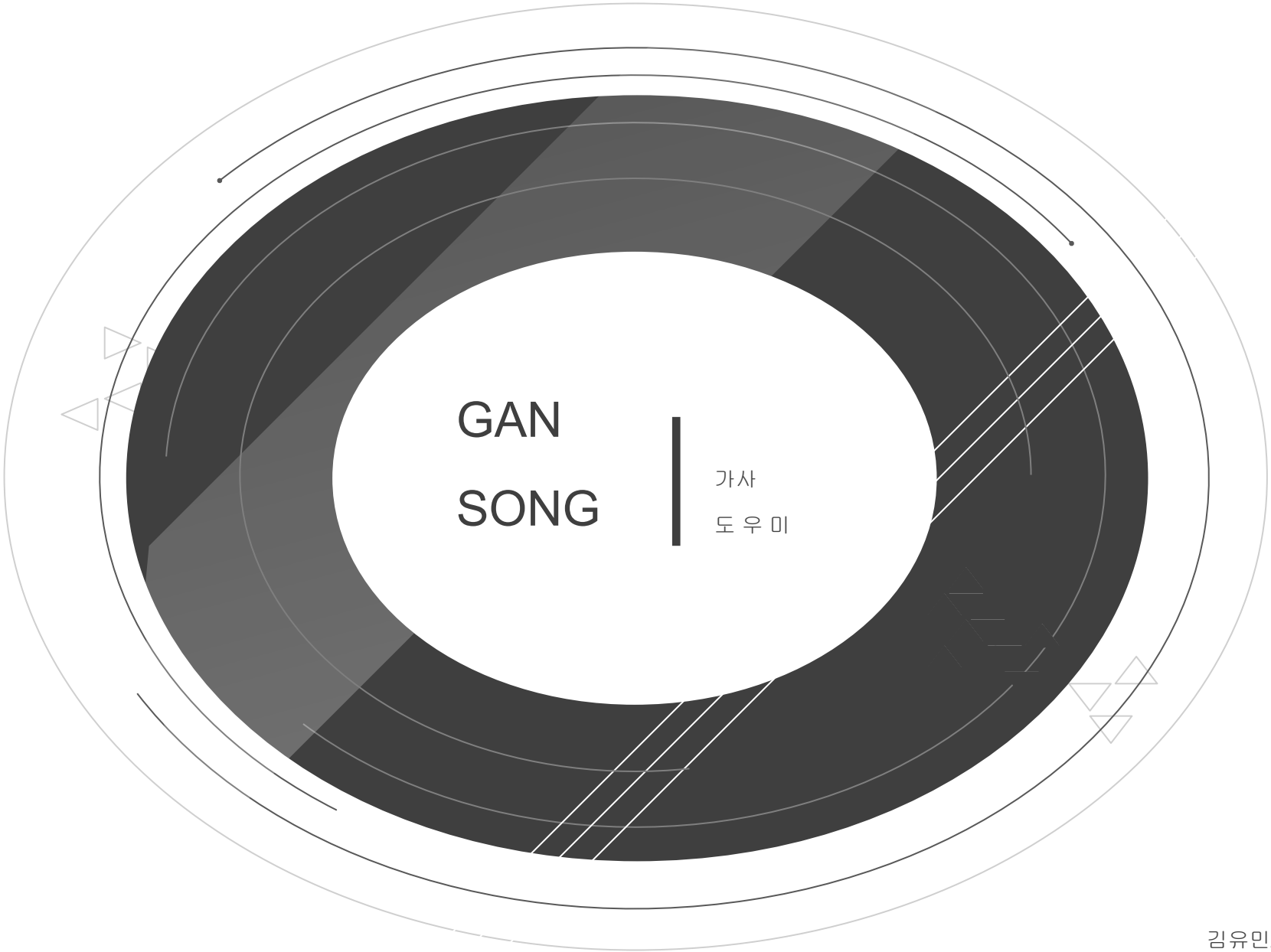


김유민/심은선/이준걸

장청아/정윤호/황이은

목차

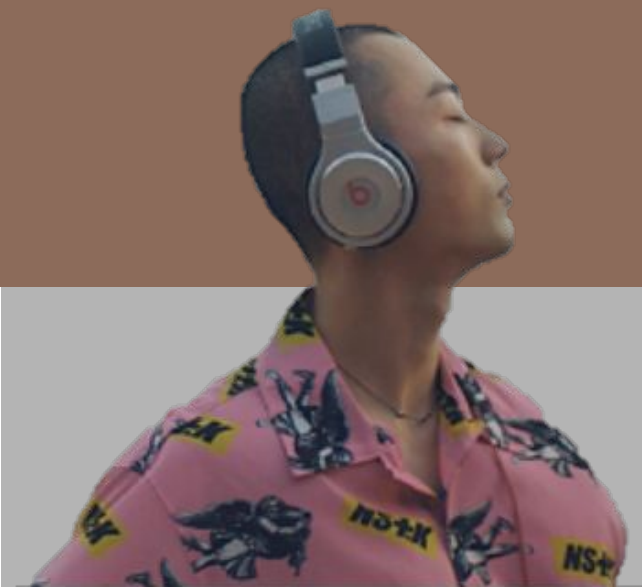
가사도우미

1장 주제 선정 배경

- 우리 모두에게는 각자의 사연이 있다

주제 선정 배경

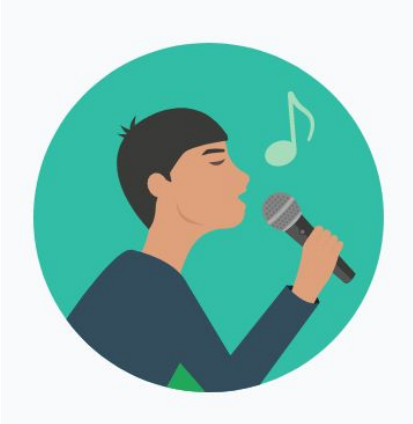
사람들은 음악으로 위로 받고 가사에 공감을 한다

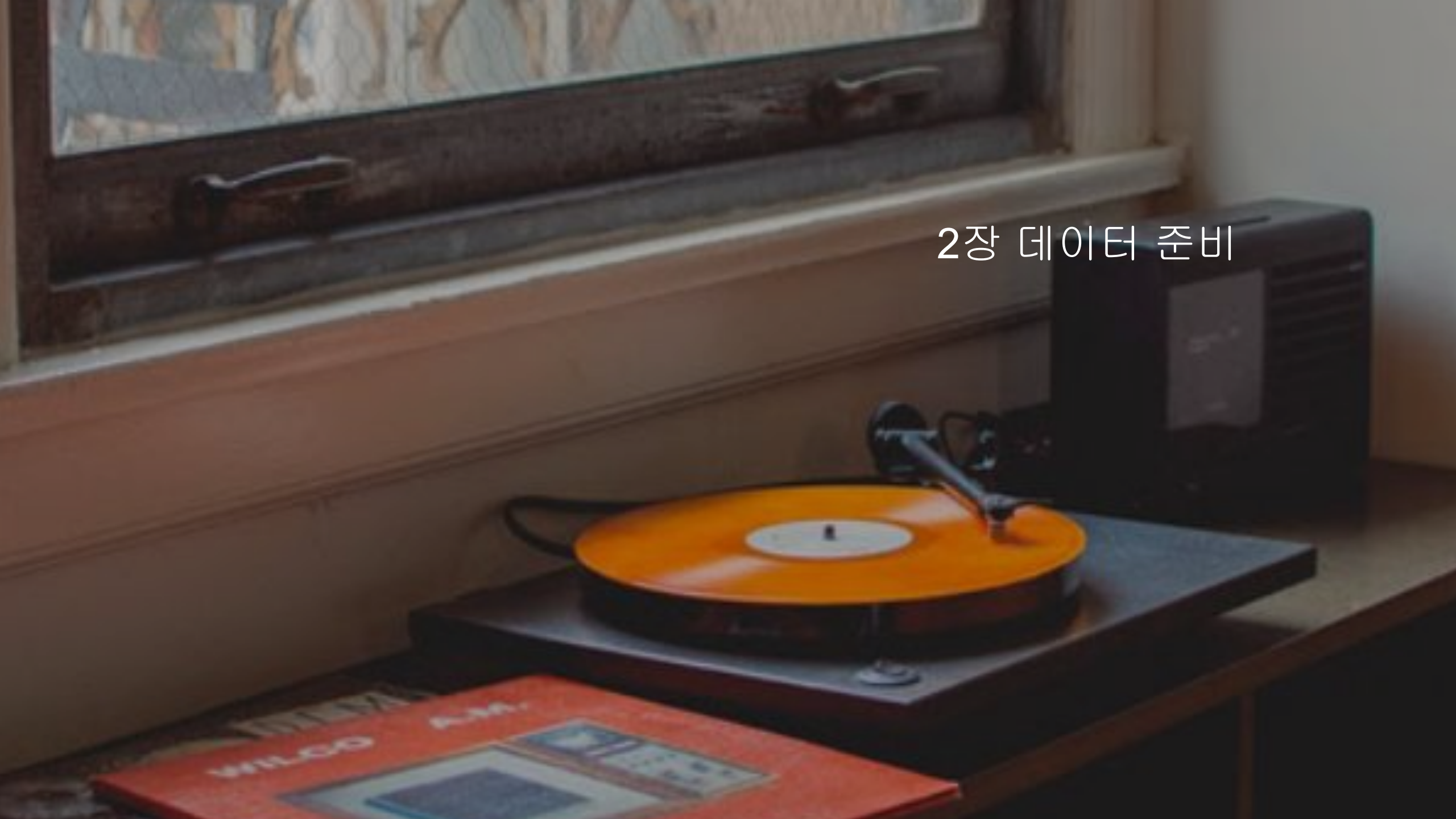


- **사람들에게 맞춤형 노래(customize)를 만들어보자!**
  - 사용자의 사연에 맞추어 옛 아티스트들의 가사를 학습해 감수성을 자극하는 노래 가사를 생성해보자

- 노래 가사가 아름다운 아티스트
-> 김광석, 김현식, 유재하, …

- 다양한 주제의 노래들
-> 그 중에서도 사랑, 이별, 슬픔과 관련된 주제

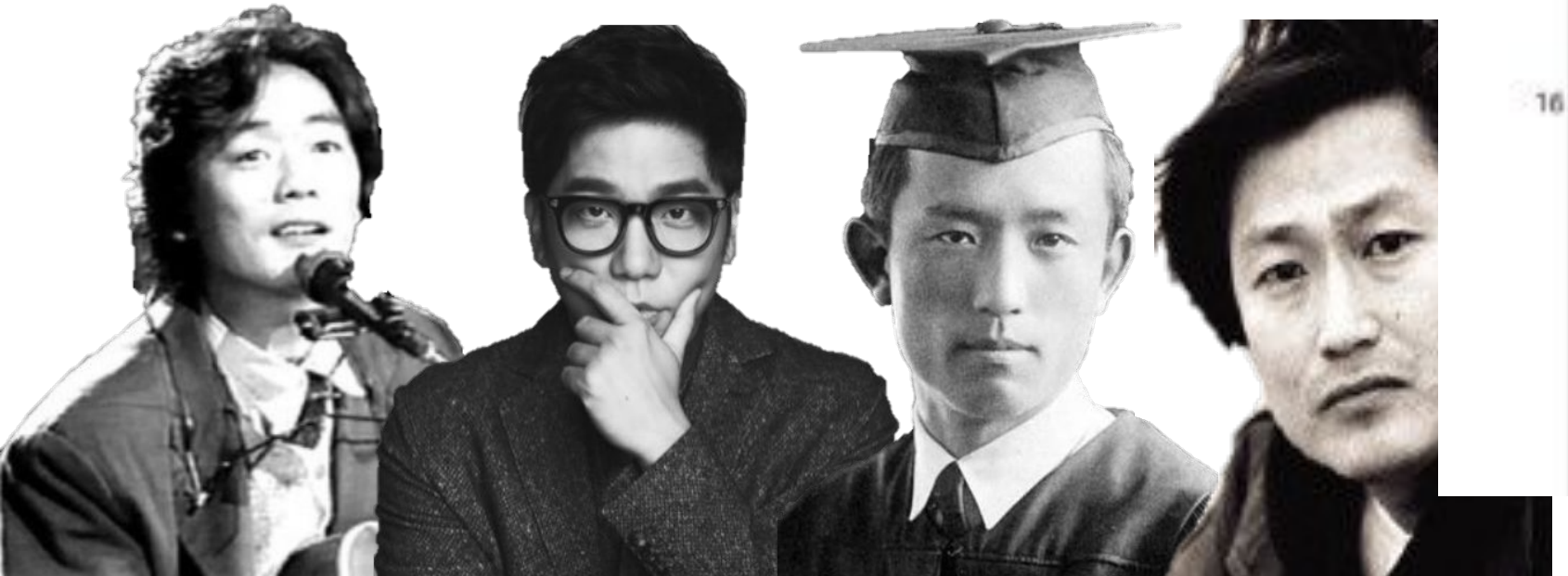




2장 데이터 준비

- **데이터 준비-크롤링**

  - 네이버 뮤직에서 유재하, 김광석 외 6명, 742곡 크롤링
  - 김소월, 윤동주 등 100여개 작품 크롤링



artist_num	title	lyrics
		나의 하늘을 본 적이 있을까
		조각 구름과 빛나는 별들이
		끝없이 펼쳐 있는
		구석진 그 하늘 어디선가
		내 노래는 날 부르고 있을을
		난 알고 있는지
		나의 청춘을 본 적이 있을까
		국화와 장미 예쁜 사루비아가
		끝없이 피어 있는
		언제든 그 곳은 열려 있고
		그 맑기는 날 부르고 있을을
		난 알고 있는지
16	너에게	나의 어릴 적 내 꿈만큼이나
		아름다운 가을 하늘이랑
		내가 그것들과 손잡고
		고요한 달빛으로 내게 오던
		내 여린 마음으로 피워낸 나의 사랑을
		너에게 주어질게
		나의 어릴 적 내 꿈만큼이나
		아름다운 가을 하늘이랑
		내가 그것들과 손잡고

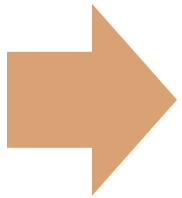
3장 데이터 전처리

- **Tokenization & Regular Expression**

  - 정규표현식으로 특수문자, 숫자, 영어를 제거하고 한글만 추출
  - KoNLPy의 twitter 형태소 분석을 통해 tokenization을 진행함

<Data>

인생이란 강물 위를 뜻
없이 부초처럼
떠다니다가



<Tokenize>

['인생', '이란', '강물', '위',
'를', '뜻', '없이', '부초',
'처럼', '떠다니다가']

- **Word & Index Dictionary**

  - word 와 word vector의 상호변환을 위해 word에 고유한 index를 부여한 word to index와 index to word을 사전을 구축

인생	이란	강물	위	를	뜻	없이	부초	처럼	떠다니다가
153	94	202	69	7	320	78	3919	9	3920

Word to Index

Index to Word

- **Word Embedding Matrix**

  - 노래 가사 특유의 문맥을 살리기 위해서 pre-trained embedding matrix를 가져다 사용하기 보다는 dataset의 토큰들을 FastText을 통해 word embedding matrix를 구성

"인생이란 강물 위를 뜻 없이 부초처럼 떠다니다가"

["인생", "이란", "강물", "위", "를", "뜻", "없이", "부초", "처럼", "떠다니다가"]

word embedding matrix : 각 행이 특정 word의 word vector로 구성된 array

[
인생(153) : [0.7210467, -0.7817296, ⋯ , -0.8295756]
이란(94) : [-1.0544485, 0.7245336, ⋯ , -0.01476351]
.
.
떠다니다가(20) : [1.431622 , 0.76734734, ⋯ , -0.1439352
]

- **FastText**

입력하십시오 → ‘입’, ‘력’, ‘하’ ‘십’, ‘시’, ‘오’ →

‘<입력’ - 0.0003

‘입력하’ - 0.0021

‘력하십’ - 0.0007

‘하십시’ - 0.0015

‘십시오’ - 0.0069

‘시오>’ - 0.0308

→ ‘입력하십시오’ : 0.0423

단어 Bag-of-Characters	3-gram의 Characters Embedding	최종 단어의 Embedding 값 = 3-gram Embedding의 합

- **Why FastText?**

<Word2Vec>

✔ 전체 corpus에서 중심 단어와 window를 기준으로 둘러싼 단어들을 한 단어씩 훑어가며 학습함

✔ 중심 단어와 주변 단어 쌍의 관계를 학습(CBOW, Skip-gram 방식)

<FastText>

✔ 같은 어근을 가진 단어들끼리 parameter를 공유하므로 복합 명사를 표현하기 용이함
(ex) ‘disaster’/‘disastrous’

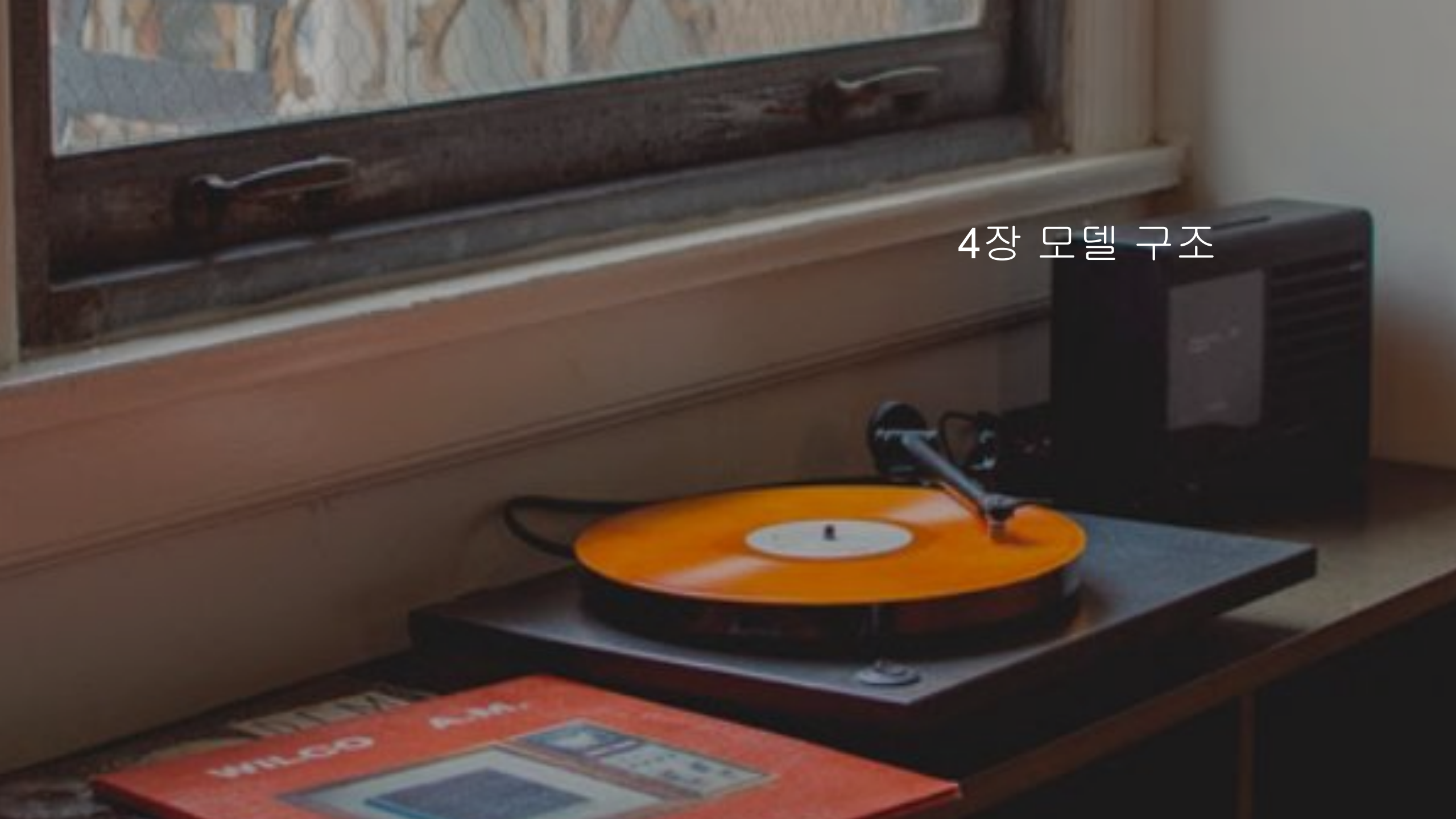
✔ Character n-gram을 통해 Out-of-Vocabulary(OOV) 문제를 해결함

Word2Vec 한계점

- 단어의 형태학적 특성을 반영하지 못함
- 희소한 단어를 Embedding하기 어려움
- Out-of-Vocabulary(OOV)를 처리할 수 없음

FastText

Character n-gram을 통해 희소한 단어에 대해서도
Word2Vec에 비해 dense하게 Embedding 가능



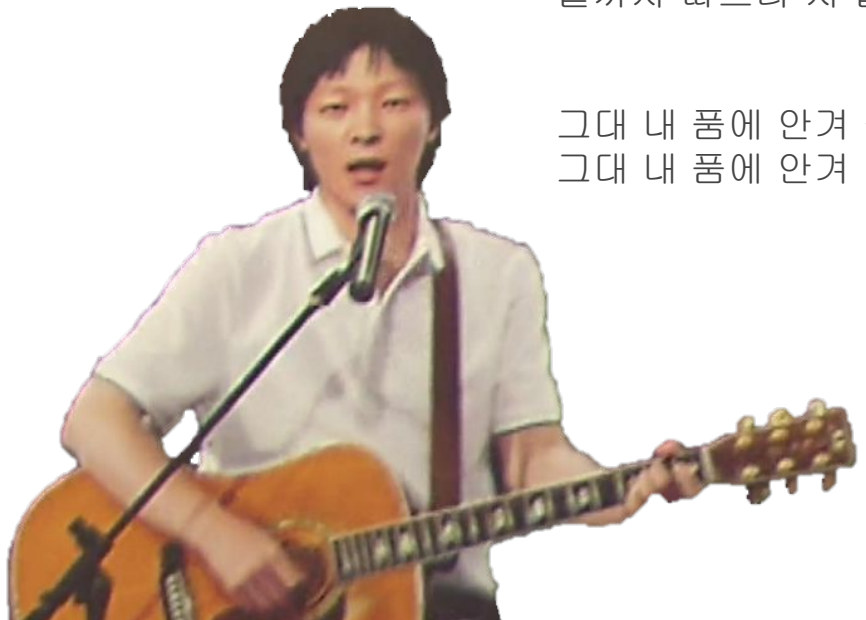
4장 모델 구조

노래 가사 형식

유재하 – 그대 내 품에

별 헤는 밤이면 들려오는 그대의 음성
하얗게 부서지는 꽃가루 되어 그대 꽃 위에 앉고 싶어라
밤하늘 보면서 느껴보는 그대의 숨결
두둥실 떠가는 쪽배를 타고 그대 호수에 머물고 싶어라
만일 그대 내 곁을 떠난다면
끝까지 따르리 저 끝까지 따르리 내 사랑

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요



술잔에 비치는 어여쁜 그대의 미소
사르르 달콤한 와인이 되어 그대 입술에 닿고 싶어라
내 취한 두 눈엔 너무 많은 그대의 모습
살며시 피어나는 아지랑이 되어 그대 곁에서 맴돌고 싶어라
만일 그대 내 곁을 떠난다면
끝까지 따르리 저 끝까지 따르리 내 사랑

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

어둠이 찾아 들어 마음 가득 기댈 곳이
필요할 때

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

노래 가사 형식

유재하 – 그대 내 품에

절(Verse)

후렴구(Hook)

브릿지(Bridge)

별 헤는 밤이면 들려오는 그대의 음성
하얗게 부서지는 꽃가루 되어 그대 꽃 위에 앉고 싶어라
밤하늘 보면서 느껴보는 그대의 숨결
두둥실 떠가는 쪽배를 타고 그대 호수에 머물고 싶어라
만일 그대 내 곁을 떠난다면
끝까지 따르리 저 끝까지 따르리 내 사랑

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

술잔에 비치는 어여쁜 그대의 미소
사르르 달콤한 와인이 되어 그대 입술에 닿고 싶어라
내 취한 두 눈엔 너무 많은 그대의 모습
살며시 피어나는 아지랑이 되어 그대 곁에서 맴돌고 싶어라
만일 그대 내 곁을 떠난다면
끝까지 따르리 저 끝까지 따르리 내 사랑

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

어둠이 찾아 들어 마음 가득 기댈 곳이
필요할 때

그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

노래 가사 형식

김광석 - 일어나

절(Verse)

후렴구(Hook)

브릿지(Bridge)

검은 밤의 가운데 서 있어
한 치 앞도 보이질 않아
어디로 가야 하나
어디에 있을까 둘러 봐도
소용없었지

인생이란 강물 위를
끝없이 부초처럼 떠다니다가
어느 고요한 호숫가에 닿으면
물과 함께 썩어가겠지

일어나 일어나
다시 한번 해보는 거야
일어나 일어나 봄의
새싹들처럼

끝이 없는 말들 속에
나와 너는 지쳐가고
또 다른 행동으로 또 다른
말들로 스스로를 안심시키지

인정함이 많을수록
새로움은 점점 더 멀어지고
그저 왔다 갔다 시계추와
같이 매일매일 흔들리겠지

일어나 일어나
다시 한번 해보는 거야
일어나 일어나
봄의 새싹들처럼

가볍게 산다는 건
결국은 스스로를 얽어 매고
세상이 외면해도
나는 어차피 살아 있는 걸

아름다운 꽃일수록
빨리 시들어 가고
햇살이 비치면 투명하던
이슬도 한 순간에 말라 버리지

일어나 일어나
다시 한번 해보는 거야
일어나 일어나
봄의 새싹들처럼

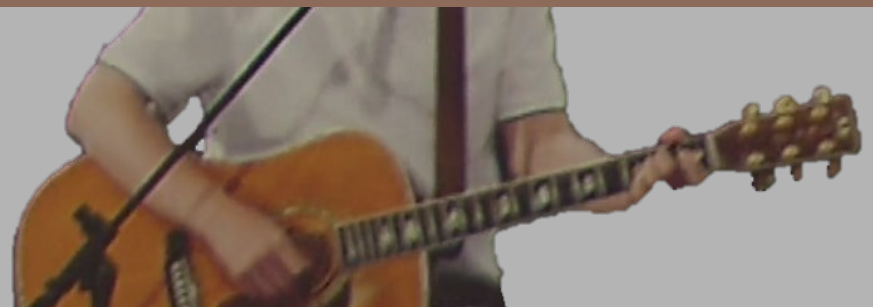
일어나 일어나
다시 한번 해보는 거야
일어나 일어나
봄의 새싹들처럼

노래 가사 형식

노래 가사는 매우 구조적으로 짜여 있음

1. 노래 가사의 형식적 구조
Verse Hook Bridge

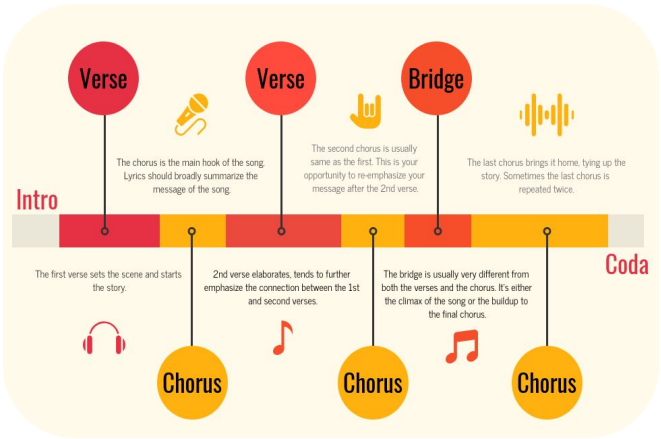
2. Hook에서 나타나는 반복적 구조



그대 내 품에 안겨 눈을 감아요
그대 내 품에 안겨 사랑의 꿈 나눠요

- **우리의 Task**

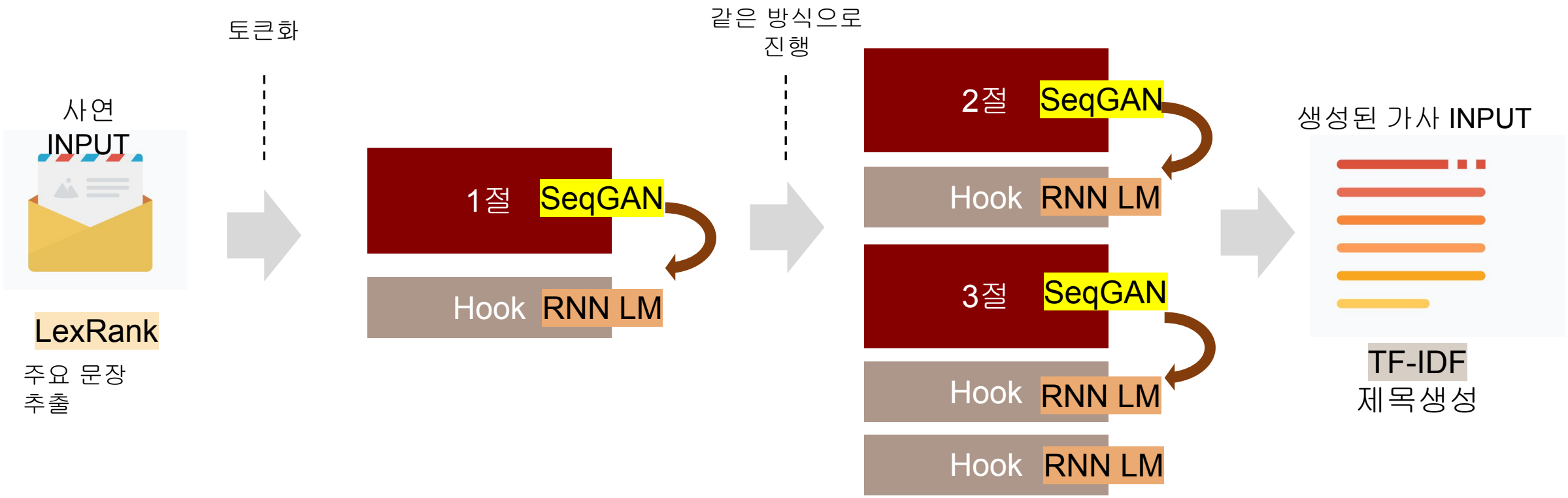
1. 사연 기반



2. 노래 가사 형식 유지(절, 훅, 라임)

3. 문맥에 맞게 생성

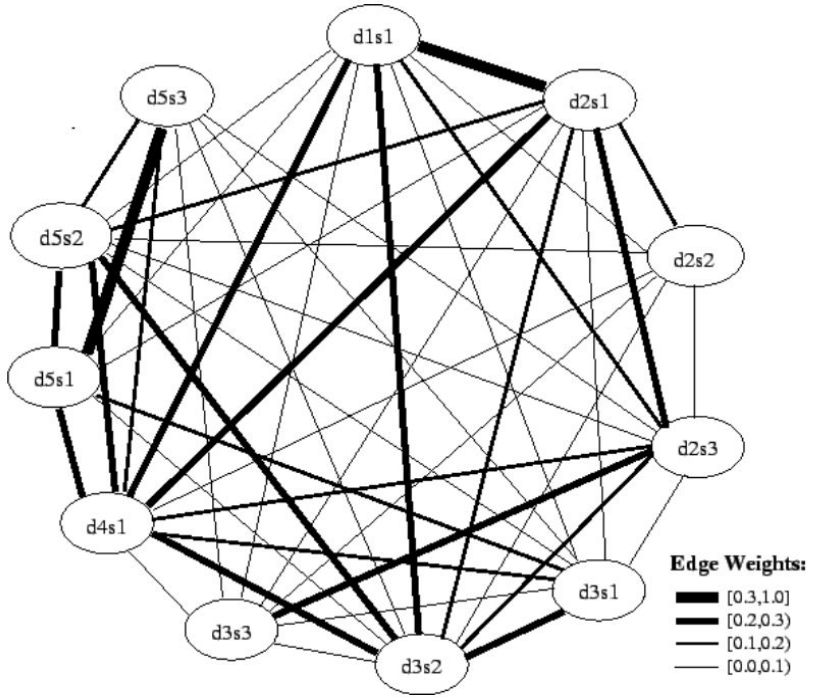
프로젝트 OVERVIEW




사연 요약 모델

1. LexRank

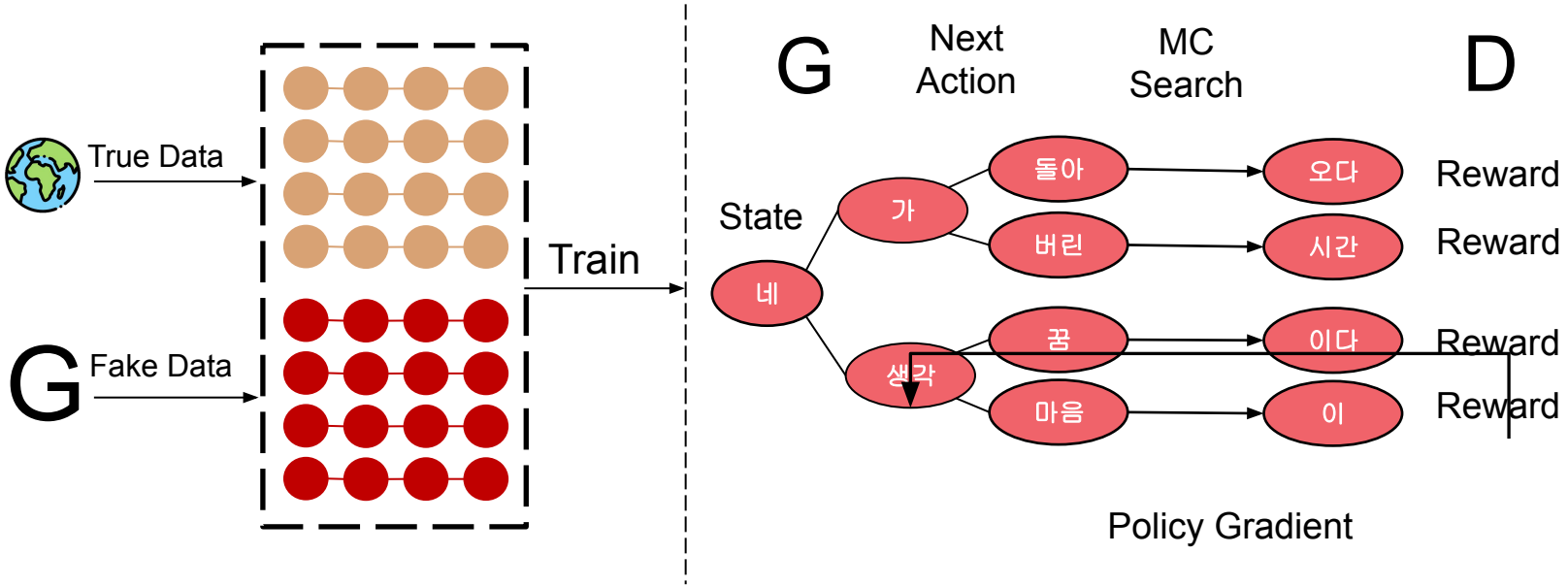
- LexRank 알고리즘은 TextRank와 비슷하게, 문서 내의 각 문장들을 노드로,
  문장들 간 유사도를 선의 값으로 그래프를 만든 후
  PageRank를 적용해서 중요한 문장을 추출해내는 추출 기반 문서 요약 알고리즘




가사생성 모델

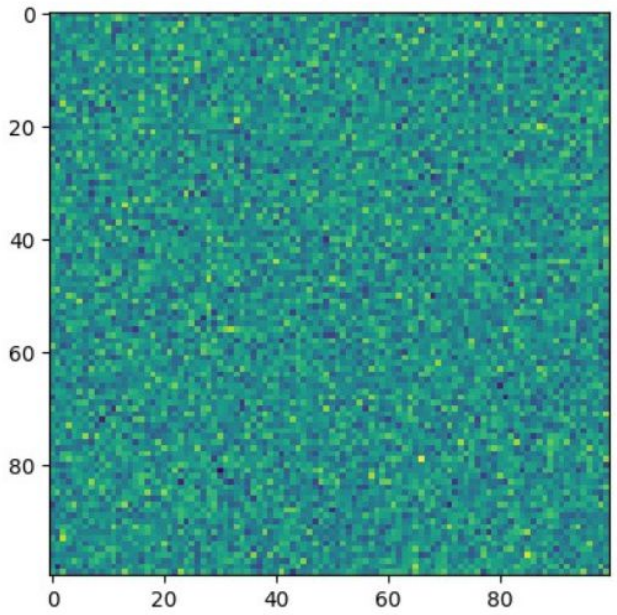
2. SeqGAN

- 강화학습을 적용해 discrete한 text data를 생성하는 GAN
  가짜 데이터를 생성하는 Generator와 가짜 데이터를 구별하는 Discriminator로 구성됨

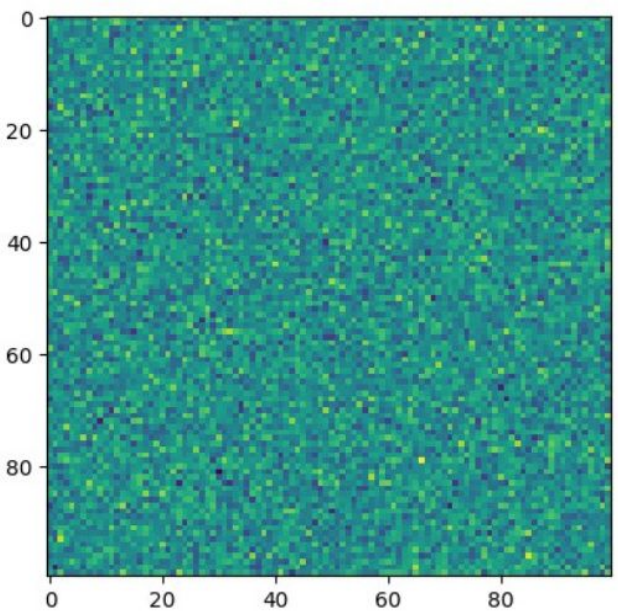


2. SeqGAN

< Image >



M

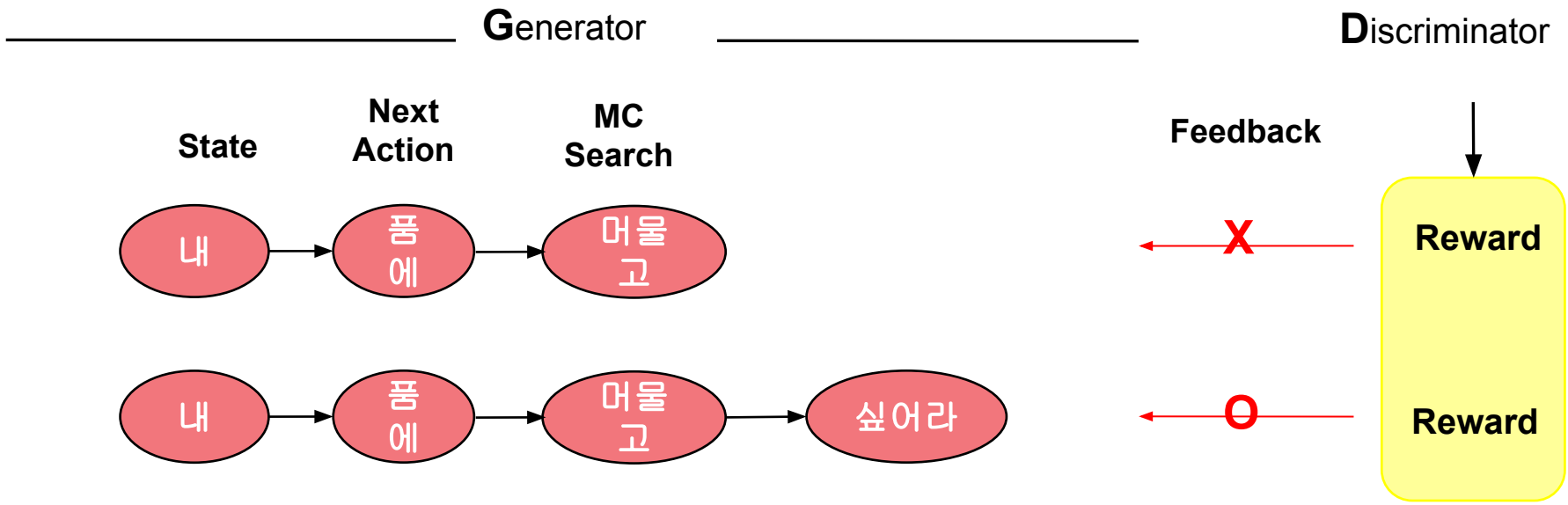


M + 0.08

< Text >

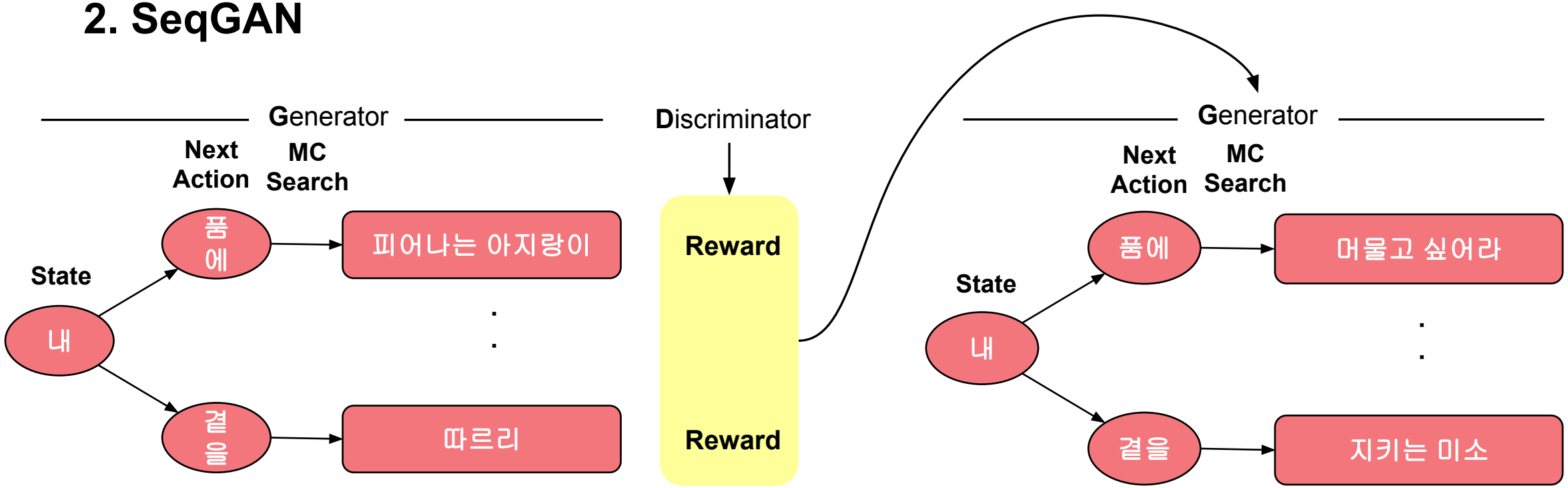
이미지와 달리 텍스트 데이터에서는 token들이 discrete하기 때문에
D model에서 G model까지의 gradient update가 되기 어려움

2. SeqGAN



GAN에서 D model은 오직 entire sequence에 대해서만 feedback을 줄 수 있기 때문에
문장이 partial sequence인 경우에는 어떠한 feedback도 줄 수 없음

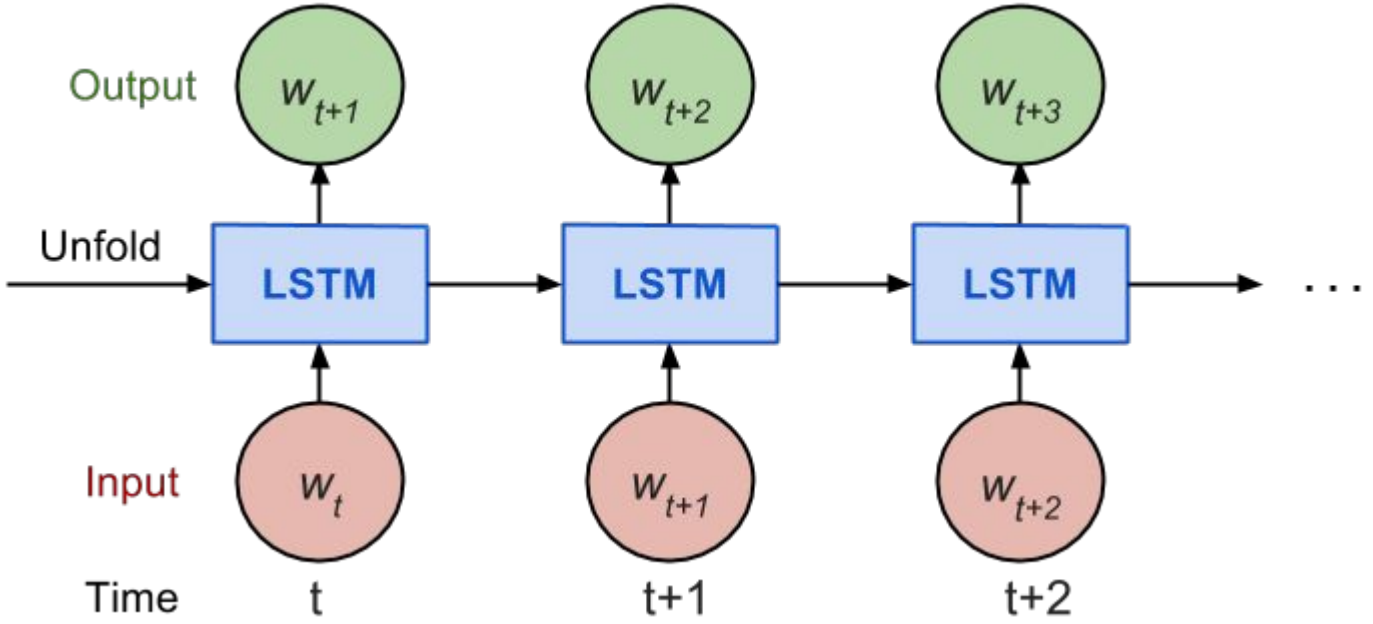
2. SeqGAN



State	현재까지 생성한 문장
Next Action	생성할 다음 토큰
Reward	G가 생성한 한 문장에 대한 Reward

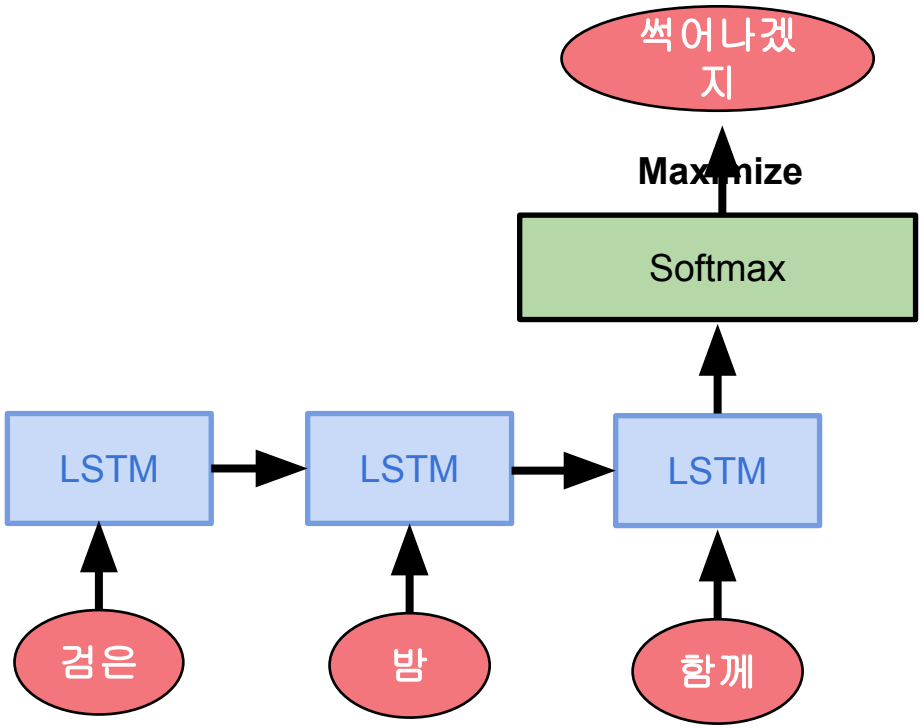
3. RNN-Language Model

- **Language Model** : 주어진 문장에서 이전 단어들을 보고 다음 단어가 나올 확률을 계산해주는 모델
- **생성(generative) 모델**로 적용하면 출력 확률 분포에서 샘플링을 통해 문장의 다음 단어가 무엇이 되면 좋을지 정한다면 기존에 없던 새로운 문장을 생성할 수 있다.



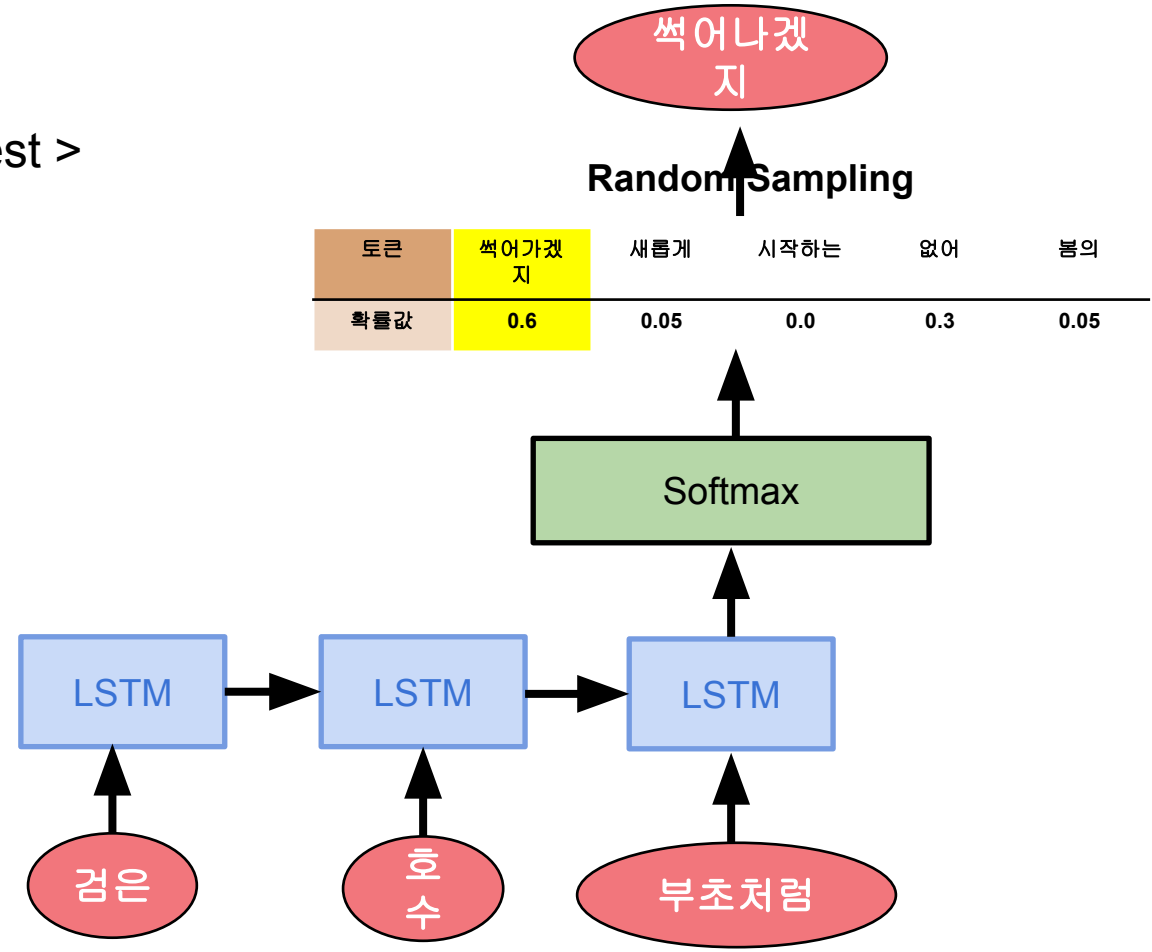
3. RNN-Language Model

< Train >



다음으로 Token으로 예측할 Softmax 확률을 Maximize하는 방식으로 학학습이 진행

< Test >



새로운 Input를 받아 계산된 Softmax 확률 분포에서 샘플링을 통해 문장의 다음 단어를 샘플링하여 기존에 없던 새로운 문장을 생성할 수 있다.

제목 생성 및 키워드 추출
생성 모델

4. TF-IDF

- 문서 내의 빈도 TF와 역문서 빈도 IDF의 곱으로 단어의 빈도를 나타냄
  특정 문서에서 자주 나타날 수록, 다른 문서에서 적게 나타날 수록 높음
- TF-IDF가 높은 단어(구)가 제목

5. 자카드 유사도

- 두 집합의 교집합의 크기를 합집합의 크기로 나눈 값으로 두 문서(집합)의 유사도를 측정
- 0에서 1사이의 값을 가지며 두 집합 사이에 교집합이 없으면 0, 두 집합이 동일하면 1의 값을 가짐

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

문서 A: 그대 내품에 안겨 눈을 감아요
문서 B: 그대 내품에 안겨 사랑의 꿈 나눠요

	그대	내품에	안겨	눈을	감아요	사랑의	꿈	나눠요
문서 A	O	O	O	O	O	X	X	X
문서 B	O	O	O	X	X	O	O	O

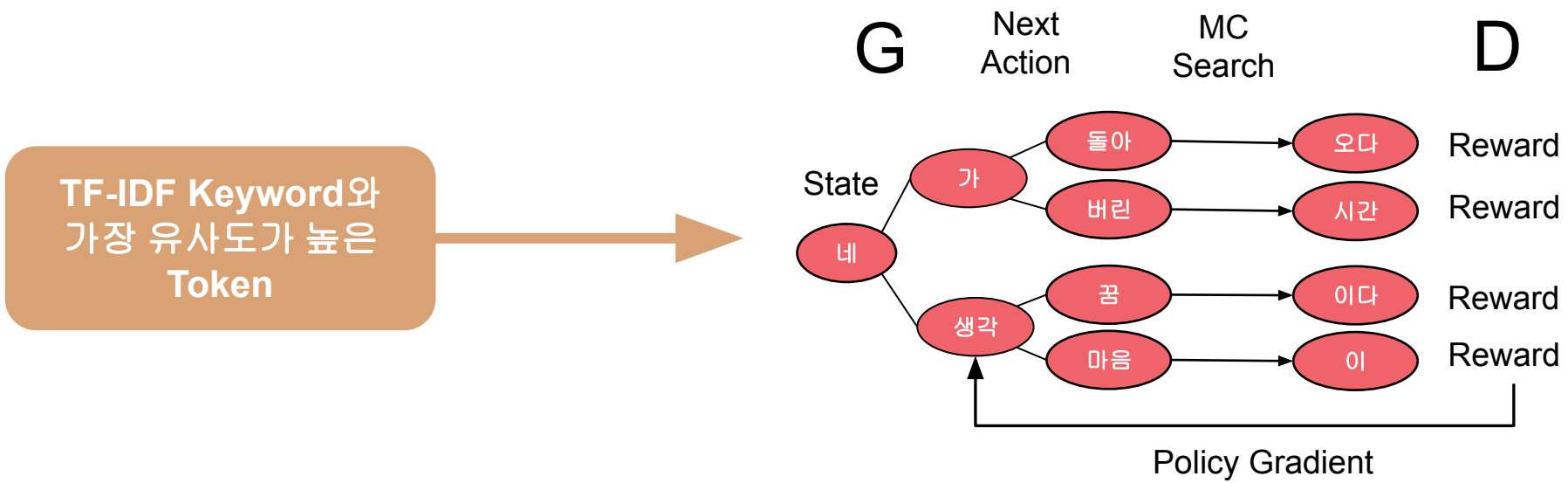
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{3}{8}$$

5장 결론 및 개선

- **결과 및 특성 – Verse(SeqGAN)**

IDEA1) Token을 Input을 받아 문장을 생성하는 SeqGAN의 특징을 이용해 이전 문장의 키워드(TF-IDF)로 생성을 시켜 다음 문장을 생성하여 문맥을 반영하고자 함

IDEA2) TF-IDF로 Token을 선정 후 이 Token과 가장 유사도가 높은 다른 Token으로 대체하기도 함



- **결과 및 특성 – Verse(SeqGAN)**

사연

LexRank로

문장 3개 추출

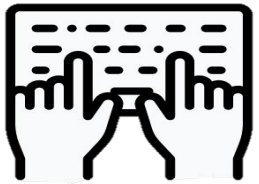
나는 항상 너를 사랑하는데 넌 가끔 날 사랑하더라
내가 유달리 몸이 아프던 날, 평소와 같이 넌 아무 연락이 없었어
그 때 말 못했지만 내가 바란건 네 사랑 하나 뿐이었어

전체 corpus에 추가

TF-IDF가 높은 명사 선택

```
('평소', 'noun')
('연락', 'noun')
('유달리', 'noun')
```

Start token



SeqGAN

사연과의 자카도 유사도를 통해

32개 문장 중 최종 문장 선택

```
["('연락','Noun')",
 "('이','Josa')",
 "('없이','Adverb')",
 "('내','Determiner')",
 "('마음','Noun')",
 "('이','Josa')",
 "('아려','Noun')",
 "('와','Josa')"]
```

- **결과 및 특성 – Verse(SeqGAN)**

User Token | **Similarity Token** | **TF-IDF Token**

이별이란 말 할 수 없는지

내 마음 아파

너와 함께한 시간의 끝을

말 할 수 없던 거예요

- **결과 및 특성 – Verse(SeqGAN)**

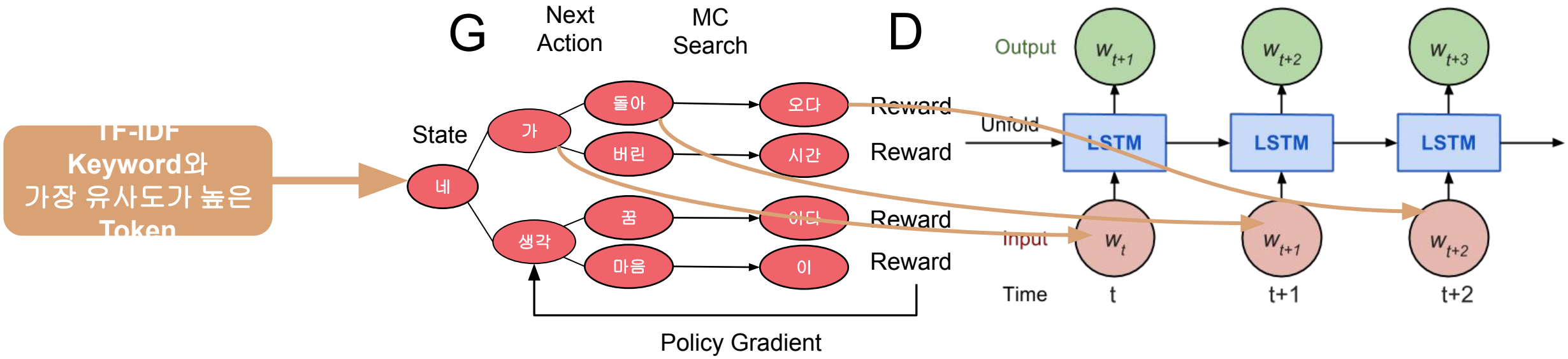
이별이란 말 할 수 없는지
내 마음 아파
너와 함께한 시간의 끝을
말 할 수 없던 거예요

이별 내게 남아
추억도 기억 속으로
기억 속으로

- **결과 및 특성 – Hook(RNN-LM)**

IDEA1) SeqGAN의 결과를 RNN-LM 모델의 Input으로 하여 Hook을 만들어 냄

IDEA2) 라임을 만들기 위해 RNN-LM에서 뱉은 토큰의 조합을 그대로 사용하여 Softmax의 확률 바꿔 생성



- **결과 및 특성 – Hook(RNN-LM)**

그대 내 품에 안겨
눈을 감아요
그대 내 품에 안겨
사랑의 꿈 나눠요

<Input>

품에 | 안겨

원래 모델

토큰	그대	눈을	바라보고	싶어	눈물
확률값	0.4	0.3	0.1	0.13	0.07

고안한 모델

토큰	그대	눈을	바라보고	싶어	눈물
확률값	0	0.3	0.1	0.13	0.07

Hook 을 만들기 위해 ‘그대’ 의 확률값을 0으로 주어 새로운 문장을 생성함

- **결과 및 특성 – Verse(SeqGAN) & Hook(RNN-LM)**

추억도 기억 속으로
기억 속으로

SeqGAN Output → **User Token**

터지는 두 눈 함께
감고 돌아오지

터지는 두 눈에

가득 찬 나 너

RNN-LM Output → **RNN의 Output을 그대로 사용한 Token**

RNN-LM Output → **Softmax 확률값을 변형하여 생성한 Token**

- **Best Example – Verse(SeqGAN) & Hook(RNN-LM)**

이별이란 말 할 수 없는지
내 마음 아파
너와 함께한 시간의 끝을
말 할 수 없던 거예요
이별 내게 남아
추억도 기억 속으로
기억 속으로

터지는 두 눈 함께
감고 돌아오지
터지는 두 눈에
가득 찬 나 너

- **Non-Best Example– Verse(SeqGAN)**

이별을 못하고
너의 두 손을 가득 하는 마음에
나에게 나는 아무 말 없이
무서운 것이 **다것을 때**
언제부턴 너와 나의 생각만
내 맘에 **음이 끝에** 인생 네게
가끔은 내 모습이 날을 때면
사랑해 내 맘은 **다시면 내가**

결론 및 개선

SeqGAN에서 생성한 32개 문장 중 자카드 유사도를 통해
문장은 선택하는 방법이 비문이 선택되는 경우가 있음

따라서 좋은 문장을 선택하는 것에 대한 Rule이 한계가 있어
사람이 읽고 선택하는 방식으로 최종 문장을 선택

- **최종결과 – 대나무숲 사연**

“내 인생에서 가장 두근거렸던 날은 확실히 그 날,
너가 나를 좋아한다고 내 눈을 보며 떨리지만 확신에 찬 그 예쁜 목소리로 말해주던 날.
너는 내가 네 고백을 거절할 거라 생각하는 듯 보였어.
사실 처음엔 널 이성으로 생각하지 않았어.

(중략)

난 그 때에 멈춰 있어. 내 **첫사랑**, 내 탈출구, 내 빛, 내 천사, 너.
사랑했어 정말 진심으로 너가 상상하지도 못 할 만큼
나는 너를 그 무엇보다도 그 누구보다도 사랑했어.
그리고 지금도 사랑하고 앞으로도 사랑할거야.
다시 말 할게. 나는 그 때에 멈춰 있어.
항상 그 곳에 있을 거야. 그러니까 다시 한 번만 더 나를 사랑해줘.”

“ 첫사랑”

작사 : **GANSONG**

첫사랑 내 아픈 가슴 속에 남아
외로운 새벽별처럼 빛나고
가슴이 일렁거렸지

하지만 그대여 그대여
있어 우리가 함께 했던
모든 것이
그대로 우리 손에 흔적 있어

그 세상엔 내겐 너
그 세상엔 우리들 처럼

첫사랑 내 아픈 가슴 속에 남아
난 추억이 있었죠
피어나네요 바람이 눈가에
눈물의 슬픈 눈
하지만 그대여 그대여
있어 우리가 함께 했던
모든 것이
추억만 남았네

그 세상엔 내겐 너
그 세상엔 우리들 처럼

그 세상엔 내겐 너
그 세상엔 우리들 처럼

Q & A

THANK YOU