SH  W ME
THE LYRICS

양수영 | 최인선

CONTENTS

CONTENTS **Introduce Topic**

Topic

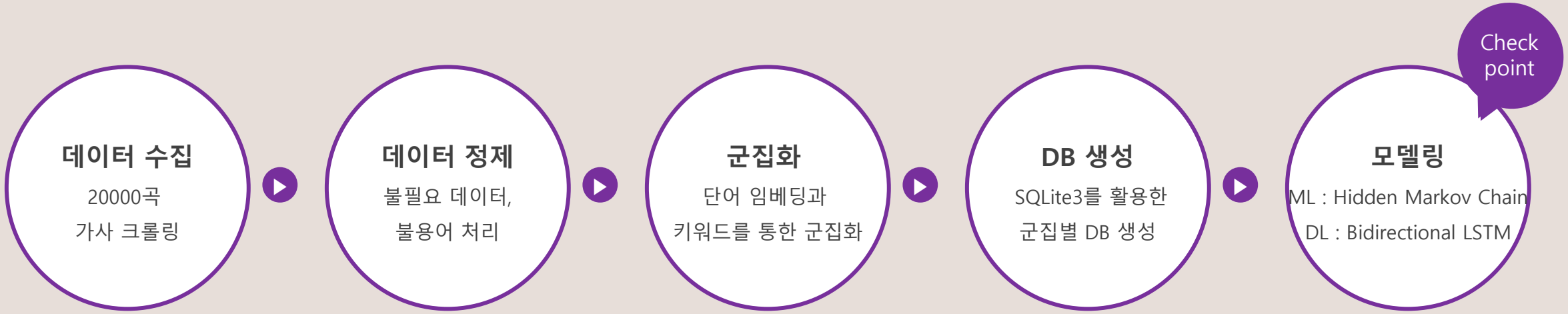
- 주어진 키워드에 맞는 랩을 작사하는 인공지능

NLP

- Natural Language Processing
- 자연어란 우리가 일상생활에서 쓰는 모든 말
- 자연어 처리란 이런 자연어를 처리하여 컴퓨터가 이해하도록 하는 것을 의미
- 자연어 처리는 음성 인식, 내용 요약, 번역, 사용자의 감성 분석, 텍스트 분류 작업(스팸 메일 분류, 뉴스 기사 카테고리 분류), 질의 응답 시스템, 챗봇과 같은 곳에서 사용된다.

SHOW ME THE LYRICS
심볼 세종이

CONTENTS **Modeling**



Check point

데이터 정제

- 토큰화와 불용어(Stopwords)처리
- 한국어 처리패키지 Konlpy

군집화

- 단어 임베딩 : TF-IDF
- 군집화 : K-Means

모델링

- ML model : Hidden Markov Chain
- DL model : Bidirectional LSTM
- 문맥을 기억하기
- 라임을 스스로 맞추기

데이터 수집

45% 언더그라운드 힙합

40% 힙합 & 랩

15% 발라드

국내 음원사이트

언더그라운드 힙합 장르 9,000곡
힙합 장르 8,000곡
발라드 장르 3,000곡

= 총 20,000 곡 가사 크롤링

- 인기순으로 데이터 크롤링
- 상대적으로 이별에 대한 말뭉치가 적어 발라드 곡을 3,000곡 포함
- 가사 없는 곡 1,840곡을 제외한 18,160곡의 가사를 데이터로 활용

데이터 정제

토큰화

- 주어진 corpus를 token단위로 나누는 작업
- 예외처리 반드시 필요

  ex) 구두점 : 20.03.20

  특수문자 : $40

  띄어쓰기 : New York

일반적인 방법

NLTK와 같은 토큰화 모델을 사용함

- 기본적으로 띄어쓰기(어절단위)로 단어 토큰화
- 구두점, 쉼표는 별도의 토큰으로 분리

불용어

- 자주 등장하지만 분석에 도움이 되지 않는 단어

  + 노이즈데이터(목적에 맞지않으며 아무런 의미를 갖지않는 단어)

  ex) 자주 등장 : is, am 등

  노이즈데이터 : n't 등

일반적인 방법

NLTK에서 영어 불용어 제공

- 제공된 불용어를 다운받아 활용
- 노이즈데이터를 불용어로 추가하고 싶다면 별도의 불용어 사전을 만들고 불러와 사용

데이터 정제

한국어의 특성?

언어를 분류하는 여러가지 기준이 있다. 어절의 형태론적 구조를 바탕으로 언어를 분류하기도 하고
음절의 단위로 글자를 쓰는 방식에 따라 분류하기도 함

1. 한국어는 교착어이다.

교착어란?
자립형태소(체언, 수식언)
+ 의존형태소(접사, 어미, 어간) 형태

ex) 교착어 : 나는, 나를
ex) 굴절어 : I, me

2. 한국어는 모아쓰기 방식이다.

모아쓰기 방식이란?
초성, 중성, 종성을 하나의 글자에
몰아서 쓰는 방식

ex) 모아쓰기 방식 : 곰
ex) 풀어쓰기 방식 : ㄱㅗㅁ, bear

데이터 정제

한국어 토큰화의 어려움

1. 한국어는 교착어이기 때문에 토큰화를 수행할 때 띄어쓰기를 기반으로 하는 어절 토큰화가 아닌 형태소 토큰화가 필요함
   ex) 나는, 나를 → 나, 는, 를
   ex) I, me → I, me
2. 한국어는 모아쓰기 방식이므로 띄어쓰기가 쉽게 무시되거나 잘 지켜지지 않는 경우가 많음
   ex) 띄어쓰지않아도읽기쉽습니다
   ex) itishardtoread

Konlpy

- 한국어 처리패키지
- 형태소화, 품사태깅 등의 기능 제공
- 종류 : Mecab, Okt, Hannanum, Kkma, Komoran 등

우리가 적용한 방법

- Konlpy의 Mecab 모듈을 통해 토큰화
- 토큰화 후 조사, 접속사 등 제거
- 추가적으로 다른 품사의 태그 중에서도 제거하고 싶은 불용어를 불용어 사전에 정의

군집화

TF-IDF

- Term Frequency – Inverse Document Frequency
- 모든 문서에 자주 등장하는 단어에 대한 패널티
- 단어-문서 행렬에서 가중치 계산하여 백터화
- 대표적인 BOW모델

Number of documents containing w

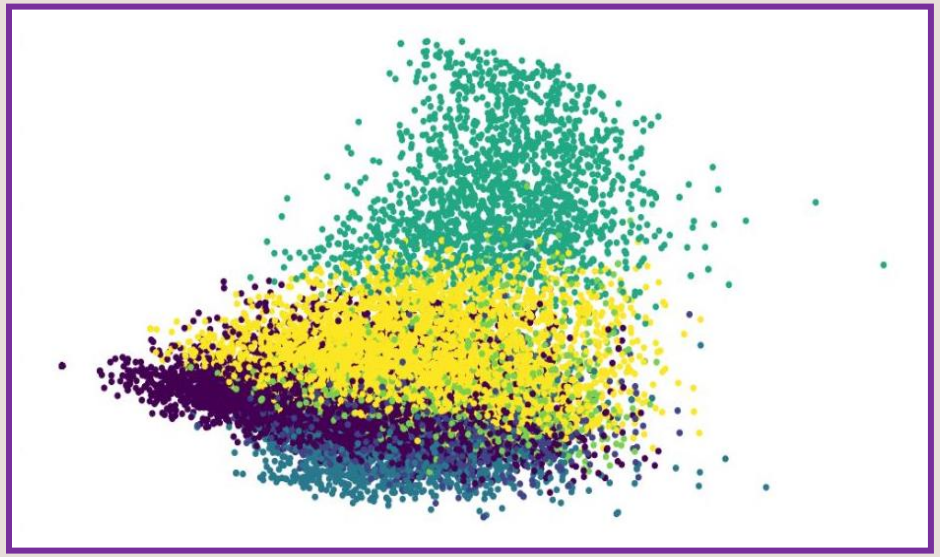
Frequency of w

$$TF - IDF(w) = TF(w) * \log(\frac{N}{DF(w)})$$

Total number of documents

K-Means

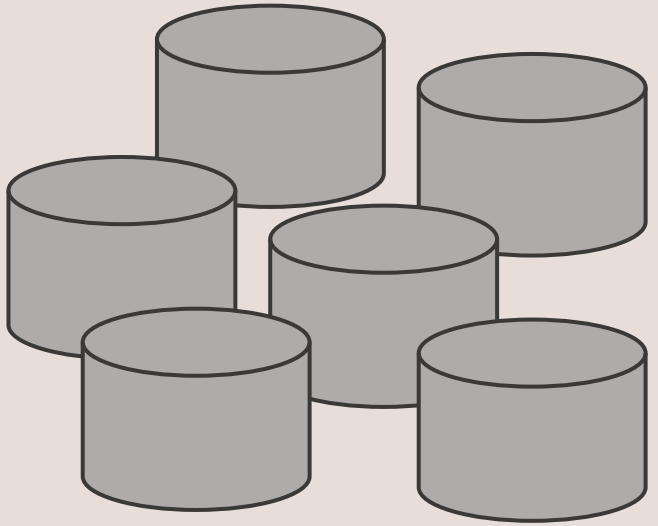
- 주어진 데이터를 K개의 군집으로 분류
- 거리기반 알고리즘이며 쉽고 간결
- 지도학습의 KNN과 비슷
- 대표적인 비지도학습 군집화모델



DB 생성

0	평화
1	스웨거
2	사랑
3	꿈
4	재미
5	이별

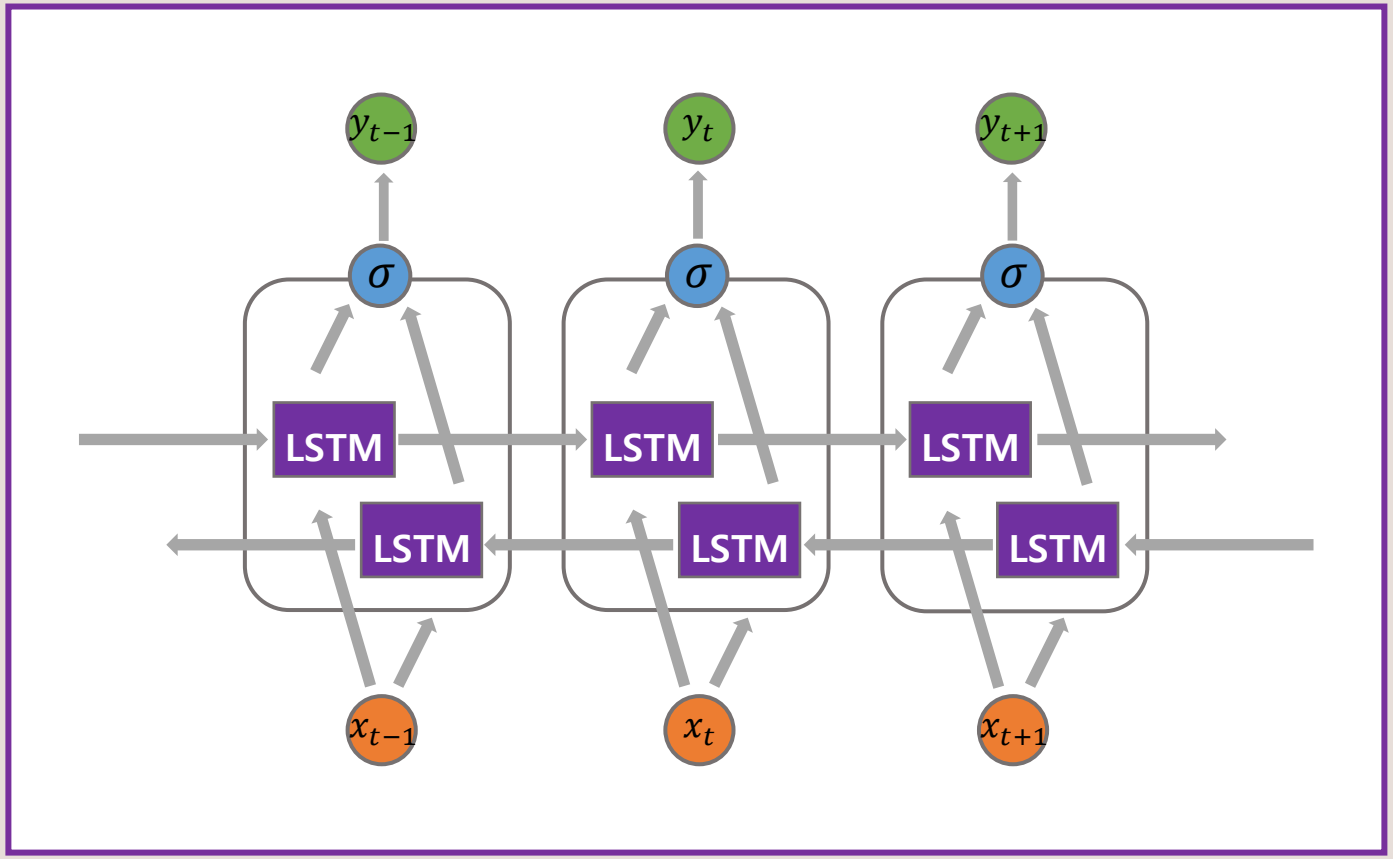
SQLite3



군집별 DB 생성

모델링

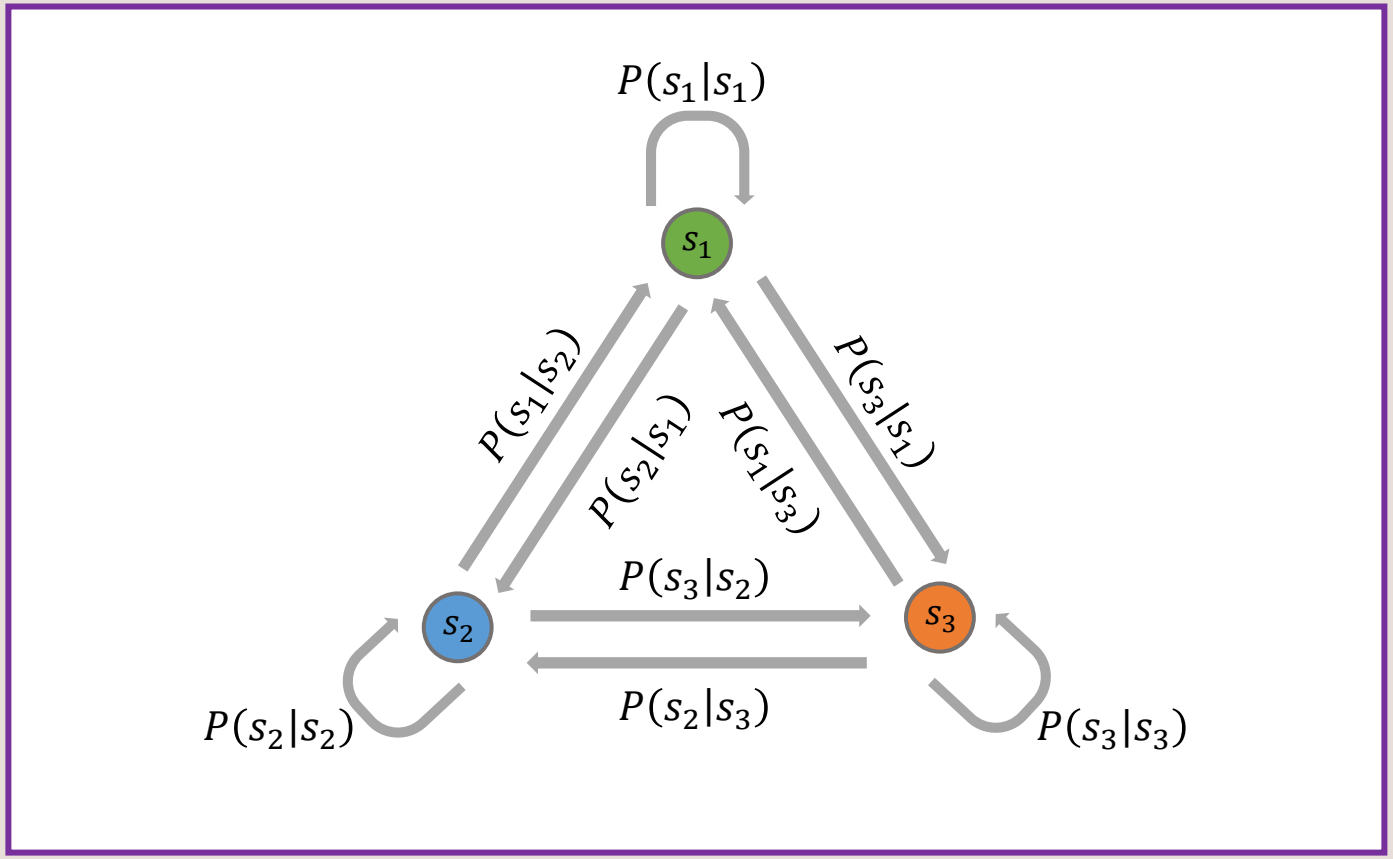
Bidirectional LSTM



- 기본적인 LSTM은 이전 step이 다음 step에 영향을 줌
- 양방향 LSTM은 이후 step 또한 이전 step에 영향을 줌
- Forward LSTM Model에서는 1부터 t까지 time step을 1씩 증가시키고 Backward LSTM Model에서는 t부터 1까지 time step을 -1씩 증가시키며 학습
- Time step마다 두 모델에서 나온 2개의 hidden vector는 학습된 가중치를 통해 하나의 hidden vector가 됨

모델링

Hidden Markov Model



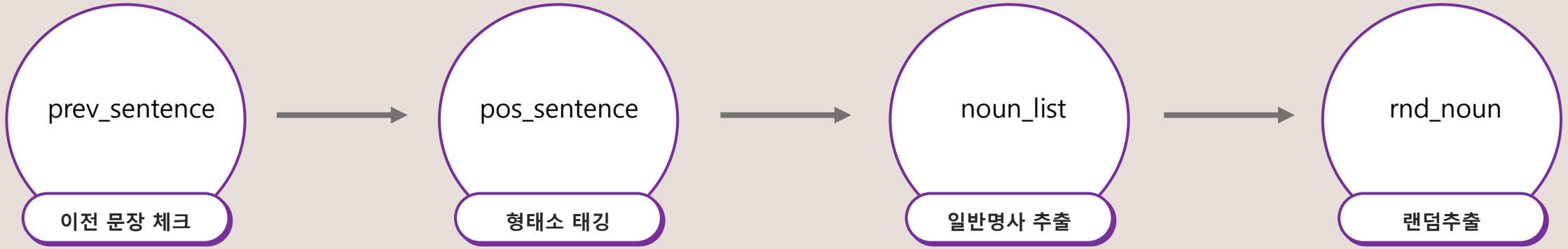
- Markov Model은 여러 개의 상태(state)가 존재하고 상태 간의 전이 확률이 Markov 성질을 지닌 이산확률과정
- 상태 전이 확률이란 각 상태에서 각 상태로 이동할 확률을 의미
- Markov 성질은 한 상태(state)의 확률이 단지 그 이전의 상태에만 의존함

\+ LSTM처럼 문맥을 기억하도록

\+ 랩 가사 처럼 라임을 맞추도록

모델링

문맥을 기억하기


prev_sentence
이전 문장 체크
pos_sentence
형태소 태깅
noun_list
일반명사 추출
rnd_noun
랜덤추출

모델링

라임을 스스로 맞추기

1차원 라임

종성이 같을 때 생성됨

ex) 자 노를 저어 강물 따라 이젠 바닷**물**
난 깨고 있어 내 인생의 미션 하나**둘**

→ ‘물’ 과 ‘둘’ 의 종성을 살펴보면 ‘ㄹ’ 로 동일하다

2차원 라임

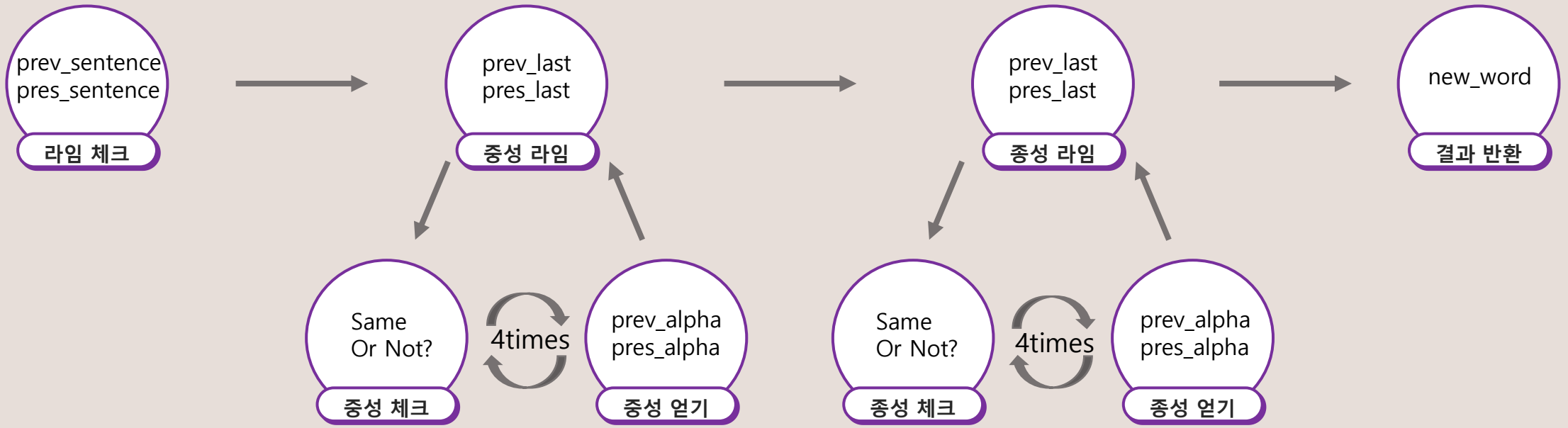
중성이 비슷한 진행일 때 생성됨

ex) 자 노를 저어 강물 따라 이젠 **바닷물**
난 깨고 있어 내 인생의 미션 **하나둘**

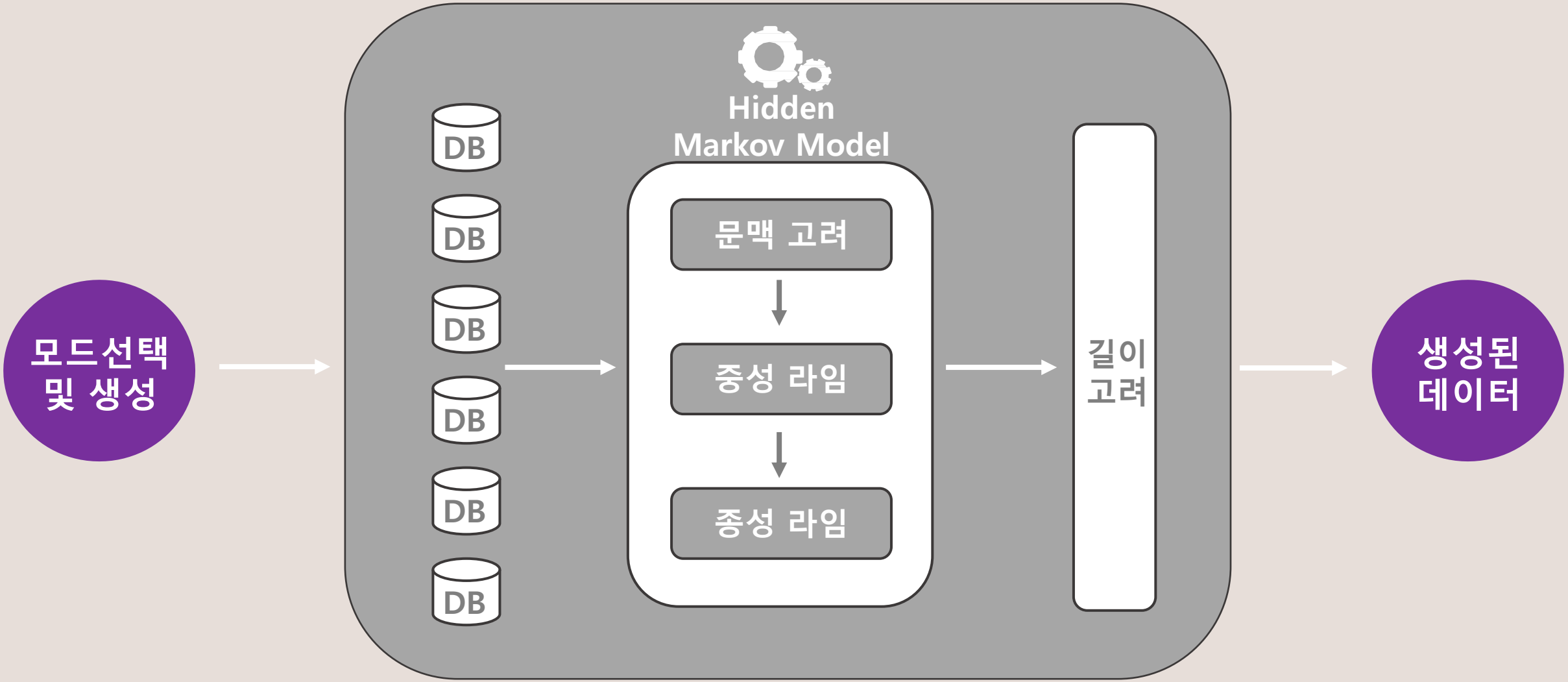
→ ‘바닷물’ 과 ‘하나둘’ 의 중성을 살펴보면 ‘ㅏㅏㅜ’ 로 동일하다

모델링

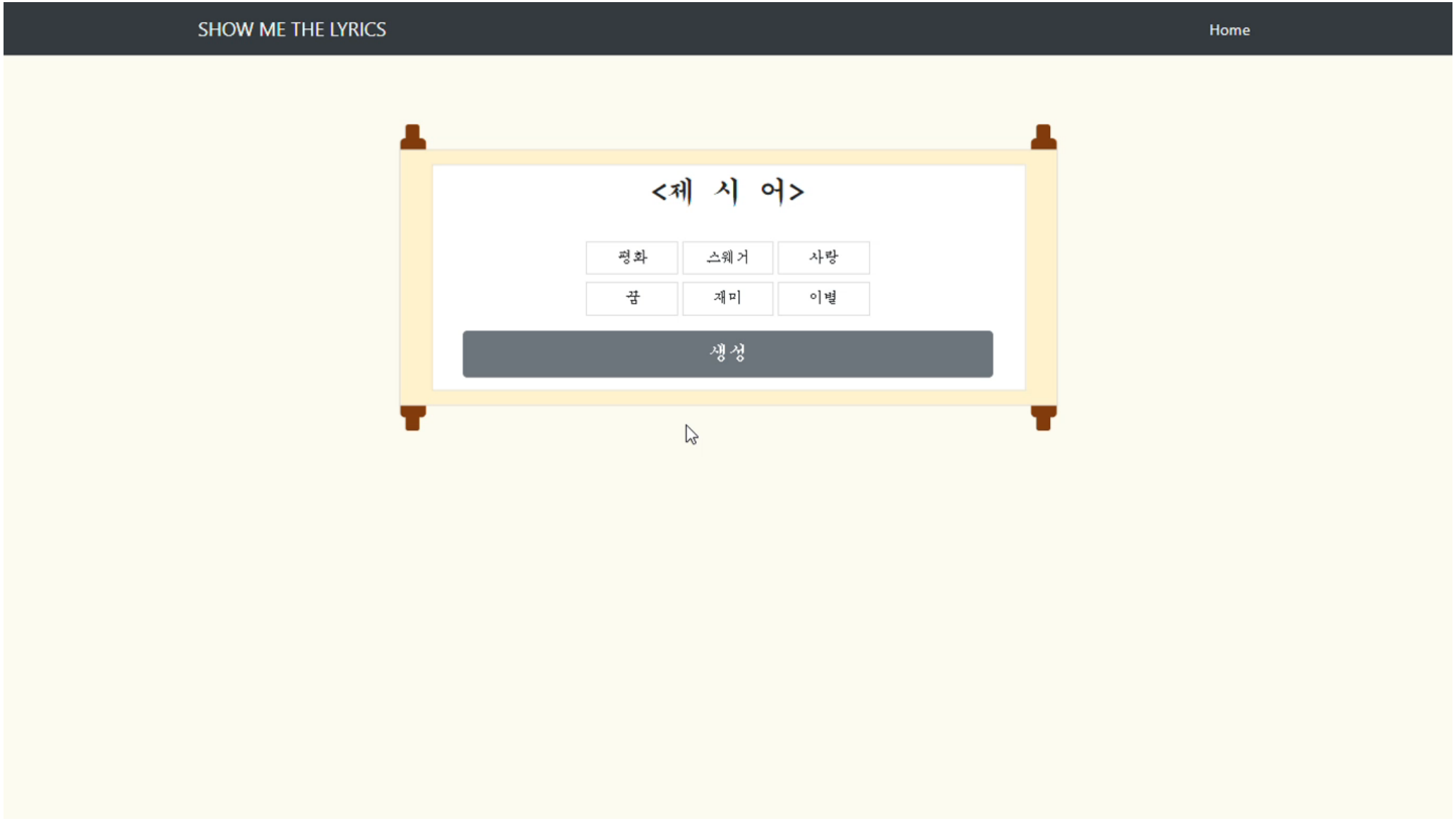
라임을 스스로 맞추기



CONTENTS **Simulation**



CONTENTS **Simulation**



특별공연 : PASMO

가사

Verse.

날 뭐가 다른지
진짜 솔직히
잘 몰라 우리 파티에는
대부분이 된 후엔

돌아보자 코끼리코
늘어나는 나의 성공
에 난 피식 온전히 터 -
놔 눈을 떠 마시마로도

재롱부려
편히 이것 저것 읽고 씹는 새끼들
삶이 절 망적이라도
나에 대해 뭐 라 해도
비엠더블유 멀세데스
쌓이다 못해서 그 흔한 애-
들 - 중 딱 보니 별-로
원- 하 는 장면이 돼-

Chours.

난 슈퍼파이리
모르는 남자지
손아귀에 있지
어디에서 나왔는지

난 슈퍼파이리
모르는 남자지
흘러나오게 만들어
놨고 한 번 해볼게

난 슈퍼파이리
모르는 남자지
손아귀에 있지
어디에서 나왔는지

난 슈퍼파이리
모르는 남자지
흘러나오게 만들어
놨고 한 번 해볼게

CONTENTS **Conclusion**

놀자어때 내 친구들 봐
내 존재가 오늘 밤도 않은 거야
마치 날 좀 지켜 가는 길이
이젠 어서 처럼 뜨거웠던
난 지나가 날 위한 못했어
조금 너무 어려 마
아직은 많이 몰라 겁이 나
다시 어려 다시
아직은 가 몰라 혼자 와도 가
놀자어때 내 인생
그렇게 알아 내 눈 바지 눈 꺼
엄마의 손 여기 내 어깨
버버리 다 또 너가 난 내 어깨 없지
너를 빨리 빨리
그 아빠 엄마의 들 못한
맨 한 앞 내 어깨 와

Bidirectional LSTM

문장단위에서도 말이 되지 않는 부분이 있음

침대위 여긴 바로 옆자리에
돈이 필요해 이렇게 돈벌래
넌 안돼 뭘 해도
내가 될까봐 꾹 참고 견뎌야 해 봐야 해
■■ 쩐다고 말하는 성공
지폐로 목욕을 하며
난 돈이 다가 아니야
내가 참 거지같은 인생
시기 질투 ■■들을 상대하려
빈자리 노려 보험
미신 몰라
가서 잘해 줄때 호구로 보더니
아는 척 마
■ 됨 떠날 놈
돌고도네 넌 쟤네들
벌이 벌이 내 색깔은 와사비 초밥

Hidden Markov Model

문장단위에서 말이 되지 않는 부분은 없음

CONTENTS **Conclusion**

한계	최초 프로젝트를 시작할 때 Bidirection LSTM을 활용한 문장생성 기대 → 하지만 한국어 적용의 어려움으로 Markov Chain에 치중된 모델 생성	ⓥ
의의	NLP분야는 전체적으로 영어를 위주로 발달하고 있어 적용에 제한 → 제한된 환경에서 한국어에 적용	ⓥ
	국내 및 해외 전반에서 NLU가 아닌 NLG분야에 대한 연구가 활발하지 못함 → Markov Chain을 활용해 원하는 규칙을 반영시킨 생성모델 완성	ⓥ
활용방안	작사 앱 : 원하는 주제에 대한 작사 가이드라인을 제공하여 창작의 고통 줄이기 진정한 챗봇 : 정해진 선택지를 선택하는 것이 아닌 자유로운 질의에 답하는 챗봇	ⓥ

CONTENTS **Reference**

논문

An introduction to hidden Markov models by L. Rabiner, B, Juang

Hidden Markov Models A Tutorial for the Course Computational Intelligence by Barbara Resch (modified Erhard , Car Line Rank , Mathew Magimai-doss)

Word Sense Disambiguation using a Bidirectional LSTM by Mikael Kågebäck, Hans Salomonsson

서적

다양한 캐글 예제와 함께 기초 알고리즘부터 최신 기법까지 배우는 파이썬 머신러닝 완벽가이드 (권철민 저, 2019.02.28)

한국어 임베딩 자연어 처리 모델의 성능을 높이는 핵심 비결 Word2Vec에서 ELMO, BERTR까지 (이기창 저, 2019.09.26)

그 외

https://wikidocs.net/book/2155

https://ratsgo.github.io/blog/categories/

https://github.com/codebox/markov-text

Any Question?