커뮤니티 상 거래를 위한
게시글 제목 변경 웹 확장 프로그램

팀명 : 제몫
박시현 + 박진영 + 이민희

TABLE OF CONTENTS

프로젝트 개요

프로젝트 설명

팀원 계획

다음주 개발 계획

주차 계획

피드백

피드백 답변

진행 상황

프로젝트 진행 상황

트러블 슈팅

프로젝트 설명

워싱 면 라이더 자켓 | 15000 | 사용감 유 | 2주

필립림 자켓 | 30000 | 새상품 | 0회

존스뉴욕 | 43000 | 최상 | 착용만

지난 주 피드백

Bert 사용 고려?

→ 중고나라 데이터 특성상 문맥의 흐름에 의존하지 않고도
단어의 특징을 잘 반환할 것이라고 생각

프로젝트 진행 상황

박시현	데이터셋 추가 구축 서버에 크롤링 프로그램 등록
박진영	텍스트 대치 테스트 데이터로 구현 웹페이지 데이터 저장
이민희	NER 모델 구축 모델 파라미터 변경

모델 학습 위한 데이터셋 추가 확보

게시글 오류 / 사용자 탈퇴 등 데이터 수집 오류 코드 수정

[노트북, 남성상의]

2개 카테고리 별로 각각 10000개씩 추가 확보

게시물 다양성 부족

최대 목록 페이지 = 1000페이지

1000번째 페이지 게시물들의 작성일은 2019.11.01 (여성 상의 게시판 기준)

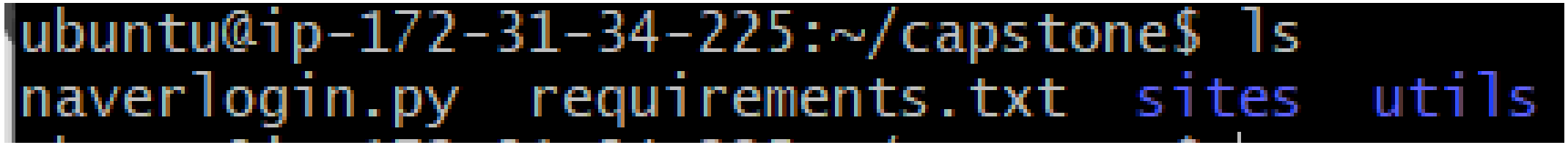
의류의 경우 계절별 의류 종류가 달라 문제

문제점을 인지하고 획득 가능한 데이터로만 학습

우분투 서버 연결

우분투 서버 구축

필요한 모듈 설치 및 환경설정



DOMSubtreeModified

```
var highestScore;

console.log('autoEvent listener!');
document.addEventListener('DOMSubtreeModified', function() {
  console.log('DOMSubtreeModified listener!');
  if(document.body.innerHTML.search('main-area') != -1) {
    console.log('target frame!');

    var table = document.getElementsByClassName('article-board')
    console.log(table.getElementsByTagName('a'));
```

DOMSubtreeModified가 필요한지 깊이 고려 X

DOMSubtreeModified

```
2 ▶Uncaught RangeError: Maximum call stack size exceeded.
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
      at HTMLDocument.<anonymous> (auto.js:26)
3
4
```

DOMSubtreeModified 이벤트 리스너 내부에서 innerText 사용

→ 무한루프 발생

DOMSubtreeModified

반복 횟수 제한 → 호출 스택 순서 실행으로 결과 중복

→ DOMSubtreeModified 이벤트 리스너 제거 (문제 해결)

텍스트 대치 구현 상황

	제목	작성자	작성일	조회
654912086	제목a 판매	자입지	2019.10.19.	17
646926413	제목b 판매	positiveart	2019.10.01.	9
642716194	제목c 판매	한옥부부	2019.09.21.	35
626446408	제목b 판매 [1]	SONGEI0901	2019.08.11.	19
623635707	제목e 판매	궁디애기	2019.08.04.	55
620970984	제목f 판매 안전	jinhee0448	2019.07.28.	85
618993083	제목g 판매	갓요정	2019.07.23.	18
616581663	제목h 판매			
615476511	제목i 판매			
614007547	제목j 판매			
606060821	제목k 판매 안전	p37625	2019.06.20.	22

확장 앱에서 변경 버튼 클릭

→ 다음과 같이 변경

쿠키 vs 웹스토리지

쿠키

작은 데이터 저장
서버와 통신하여 데이터 가져옴

웹스토리지

비교적 큰 데이터 저장
서버 통신 불필요

→ 웹스토리지 사용

utf-8 문제

[Error: ' cp949 ' codec can ' t decode byte 0xec in position 0: illegal multibyte sequence]

```
# -*- coding: utf-8 -*-
from pprint import pprint
from ckonlpy.tag import Twitter #customized twitter
from konlpy.tag import Kkma #문장 분리용
import gensim

twitter = Twitter()

def read_data(filename):
    with open(filename, 'r', encoding='UTF8') as f:
        data = f.read().splitlines()
    return data
```

kkma 문장 분리 라이브러리

```
10: '시 검 형 xl, 검 핑 m 입니다'
11: '직거래 산 본 또는 금정 역 택배거래 시 +5000xl 검
12: 'lx 나이키 빅 스우'
13: '시 후 리스 검 형 ,m 검 핑 팝니다'
14: '[ 판매 ]이 얼 즈 어고 패 커 블 스웻 셔츠 02 오렌…
15: '이 얼 즈 어고 패 커 블 스웻 셔츠
16: '[ 판매 ]이 얼 즈 어고 패 커 블 스웻 셔츠 02 오렌…
17: '코
```

지나친 띄어쓰기 및 문장분리 (토큰화에 불편)

→ kkma 라이브러리 대신 다른 라이브러리 사용

불필요한 띄어쓰기 제거

```
02: '상태 최상입니다'
03: '디자인,핏 다  이뻐요신랑한테 안어울려서 내놓아요'
04: '젊은분취향입니다'
05: '인천직거래)코모도스퀘어 울100% 세미오버핏 싱글코트'
06: '1x나이키 빅스우시 후리스 검형,m 검핑팝니다'
07: '홍대 나이키 새제품 나이키 빅스우시 검형 x1,검핑m입니다'
08: '직거래 산본또는금정역택배거래시+5000x1검형22m핑검19일괄 구매시 390000010 5261 6722'
09: '1x나이키 빅스우시 후리스 검형,m 검핑팝니다'
10: '[판매]이얼즈어고 패커블 스웻셔츠 02 오렌지 M'
11: '이얼즈어고 패커블 스웻셔츠 02 판매착불 13만9천원5회 실착쿨거래하실분 연락주세요'
12: '[판매]이얼즈어고 패커블 스웻셔츠 02 오렌지 M'
13: '코오롱 스노우볼 카키, 블랙색상 상급 팝니다'
14: '그레이쉬 카키, 블랙 색상 105사이즈입니다.'
15: '코트위주로 입다보니 착용횟수가굉장히 적습니다.'
```

```
13
14  def sentenceSlice(phrase_list):
15      sentence_list = []
16      for phrase in phrase_list:
17          sentence_list.extend(kss.split_sentences(phrase))
18      return sentence_list
```

kss 라이브러리 사용

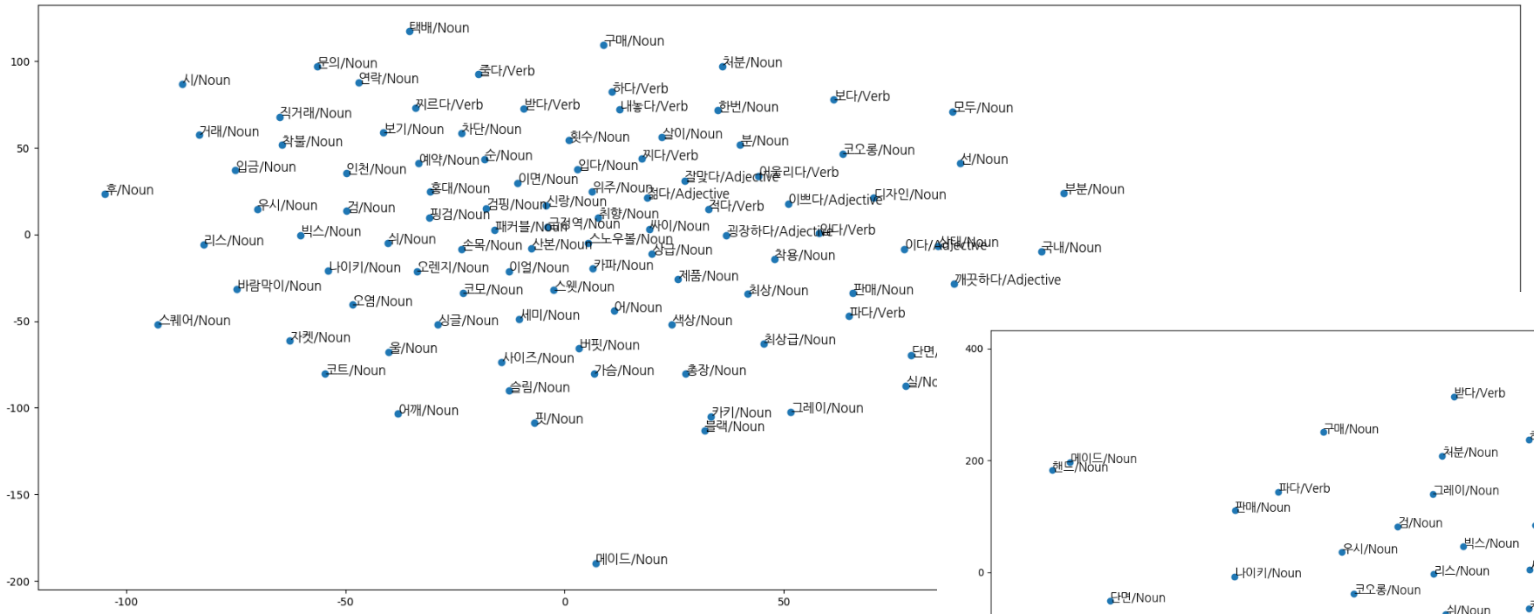
(문장 분리 속도개선 + 불필요한 띄어쓰기 제거)

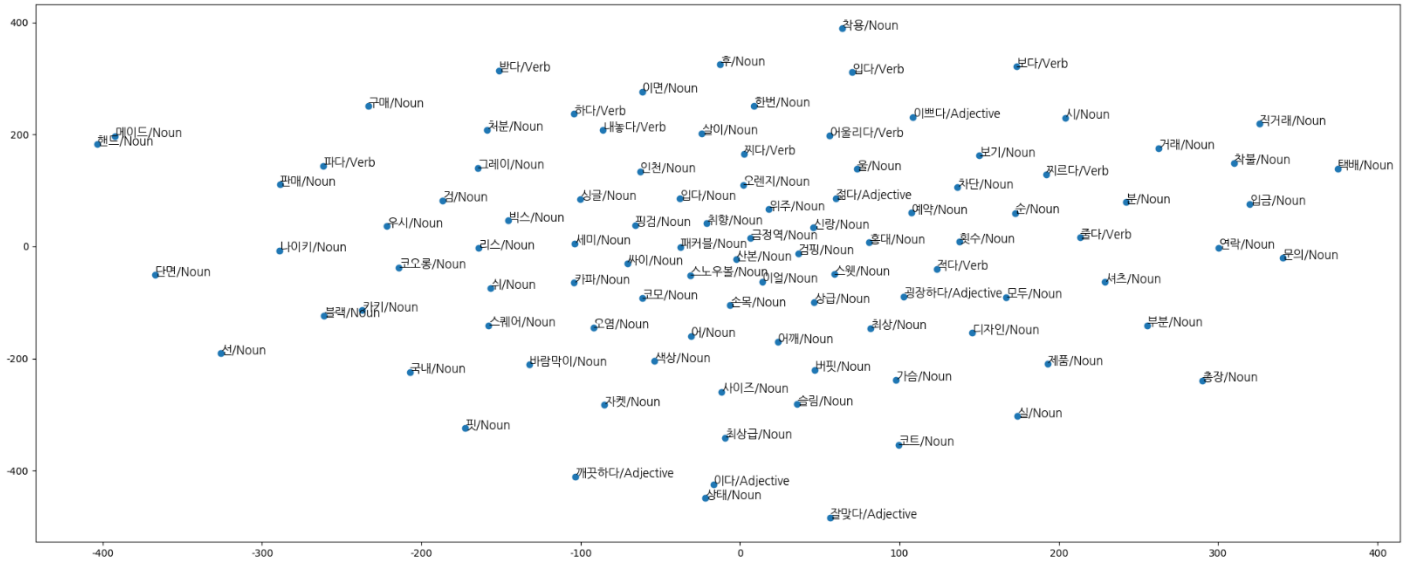
불필요한 띄어쓰기 제거

```
13: '시 후 리스 검 형 ,m 검 핑 팝니다'
14: '[ 판매 ]이 얼 즈 어고 패 커 블 스웻 셔츠 02 오렌…
15: '이 얼 즈 어고 패 커 블 스웻 셔츠 02 판매 착불 13
16: '[ 판매 ]이 얼 즈 어고 패 커 블 스웻 셔츠 02 오렌…
17: '코오롱 스노우 볼 카 키, 블랙 색상 상급 팝니다'
```

```
09: '1x나이키 빅스우시 후리스 검형,m 검핑팝니다'
10: '[판매]이얼즈어고 패커블 스웻셔츠 02 오렌지 M'
11: '이얼즈어고 패커블 스웻셔츠 02 판매착불 13만9천원5회 실착쿨거래하실분 연락주세요'
12: '[판매]이얼즈어고 패커블 스웻셔츠 02 오렌지 M'
13: '코오롱 스노우볼 카키, 블랙색상 상급 팝니다'
14: '그레이쉬 카키, 블랙 색상 105사이즈입니다.'
```

시각화 및 정확도


택배/Noun
구매/Noun
문의/Noun
처분/Noun
시/Noun
연락/Noun
줍다/Verb
하다/Verb
보다/Verb
직거래/Noun
지르다/Verb
받다/Verb
내놓다/Verb
환번/Noun
모두/Noun
거래/Noun
보기/Noun
차단/Noun
횟수/Noun
삼이/Noun
분/Noun
참불/Noun
찌다/Verb
곱오롱/Noun
선/Noun
입금/Noun
인천/Noun
예약/Noun
순/Noun
입다/Noun
어울리다/Verb
이면/Noun
홈대/Noun
위주/Noun
걸맞다/Adjective
부분/Noun
후/Noun
유시/Noun
걸/Noun
핑검/Noun
검핑/Noun
신랑/Noun
적다/Verb
이쁘다/Adjective
디자인/Noun
취향/Noun
리스/Noun
빅스/Noun
쉬/Noun
패커블/Noun
금정역/Noun
싸이/Noun
광장하다/Adjective
입다/Verb
이다/Adjective
상태/Noun
손목/Noun
산본/Noun
스노우볼/Noun
상급/Noun
국내/Noun
나이키/Noun
오렌지/Noun
이열/Noun
카파/Noun
착용/Noun
바람막이/Noun
코모/Noun
스웻/Noun
제품/Noun
최상/Noun
깨끗하다/Adjective
오염/Noun
판매/Noun
스퀘어/Noun
싱글/Noun
세미/Noun
어/Noun
색상/Noun
파다/Verb
자켓/Noun
올/Noun
버핏/Noun
최상급/Noun
사이즈/Noun
가슴/Noun
홍장/Noun
코트/Noun
슬림/Noun
어깨/Noun
핏/Noun
키키/Noun
블랙/Noun
그레이/Noun
메이드/Noun


착용/Noun
받다/Verb
휴/Noun
입다/Verb
보다/Verb
이면/Noun
구매/Noun
하다/Verb
환번/Noun
이쁘다/Adjective
시/Noun
걸맞다/Adjective
메이드/Noun
처분/Noun
내놓다/Verb
삼이/Noun
어울리다/Verb
거래/Noun
파다/Verb
그레이/Noun
찌다/Verb
보기/Noun
참불/Noun
택배/Noun
인천/Noun
홈/Noun
지르다/Verb
판매/Noun
오렌지/Noun
차단/Noun
검/Noun
싱글/Noun
입다/Noun
분/Noun
입금/Noun
유시/Noun
빅스/Noun
핑검/Noun
취향/Noun
신랑/Noun
예약/Noun
순/Noun
나이키/Noun
리스/Noun
세미/Noun
패커블/Noun
금정역/Noun
홈대/Noun
횟수/Noun
줍다/Verb
역락/Noun
문의/Noun
단면/Noun
코오롱/Noun
싸이/Noun
산본/Noun
검핑/Noun
걸다/Verb
버츠/Noun
쉬/Noun
카파/Noun
스노우볼/Noun
이열/Noun
스웻/Noun
모두/Noun
코모/Noun
손목/Noun
상급/Noun
광장하다/Adjective
부분/Noun
스퀘어/Noun
오염/Noun
어/Noun
어깨/Noun
최상/Noun
디자인/Noun
선/Noun
색상/Noun
제품/Noun
국내/Noun
바람막이/Noun
버핏/Noun
가슴/Noun
홍장/Noun
사이즈/Noun
자켓/Noun
슬림/Noun
실/Noun
핏/Noun
최상급/Noun
코트/Noun
깨끗하다/Adjective
이다/Adjective
상태/Noun
걸맞다/Adjective

시각화 및 정확도

다차원의 word2vec을 시각화를 위해 2차원으로 축소

축소할 때 관계를 유지하기 위해 t-SNE 사용

t-SNE : t-distributed Stochastic Neighbor Embedding

고차원 공간에서의 유클리디안 거리측정방법 활용해

데이터 포인트들의 유사성을 표현하는 조건부 확률로 변환하는 방법

시각화 및 정확도

단점 : 조건부 확률의 기준이 정해져 있지 않아
생성할 때마다 모양이 다르다

→ 대략적인 정확도 유추 가능하지만
→ 정확한 정확도 지표로는 사용 불가능

→ 새로운 방법 모색 중

토큰화 시 필요로 하는 품사 변경

알파벳, 숫자

similarity 비교 시 단어 벡터 학습이 잘 안되는 문제점

Similarity 처리 대신에 다른 처리 모색 중

→ 명사, 형용사, 동사만 토큰화 후 사용

NER

```
[('푸마/Noun', 0.507811367511749 3),
 ('소우/Noun', 0.469389945268631),
 ('리복/Noun', 0.462925195693969 7),
 ('뉴발란스/Noun', 0.462383717298507 7),
 ('루나/Noun', 0.459559977054595 95),
 ('줌/Noun', 0.454317212104797 36),
 ('아디다스/Noun', 0.452597200870513 9),
 ('카모/Noun', 0.44942444562911 99),
 ('콜라보/Noun', 0.439646750688552 86),
 ('후디/Noun', 0.43563652038574 22),
 ('데상트/Noun', 0.434751778841018 7),
 ('휠라/Noun', 0.43122792243957 52),
 ('카파/Noun', 0.42955279350280 76),
 ('리액트/Noun', 0.42689082026481 63),
 ('레오/Noun', 0.42623710632324 22),
```

이름 브랜드를 처리해주는
사전을 만들기 위한 모델 생성

→ 나이키, 아디다스와 유사도가
높게 추출된 목록 출력

Logistic Regression

패커블/Noun 16
스노우볼/Noun 10
평창/Noun 13
모노크롬/Noun 10
케이스위스/Noun 33
곰돌이/Noun 24
빨릴레리/Noun 20
빅바/Noun 20
녹/Noun 20
단군/Noun 14
쉐펠/Noun 20
뱅뱅/Noun 39
에르메스/Noun 42
일삼일삼/Noun 16
레알마드리드/Noun 11
볼빅/Noun 46
훼르자/Noun 14

Logistic Regression을 이용,
위에서 추출한 seed words로
positive class 예측

비슷하다고 예측된 단어가 선택

BUT, 브랜드 같지 않은 단어들 출력

정리

seed words → 사전에 추가

Logistic Regression 추출 값 → 튜닝 예정

정리

여성 옷 / 남성 옷 모델을 나눠서 생성 예정
(옷 종류, 구매하는 브랜드의 차이 발생)

→ 모델을 각각 만들었을 때의 정확도 상승 예상

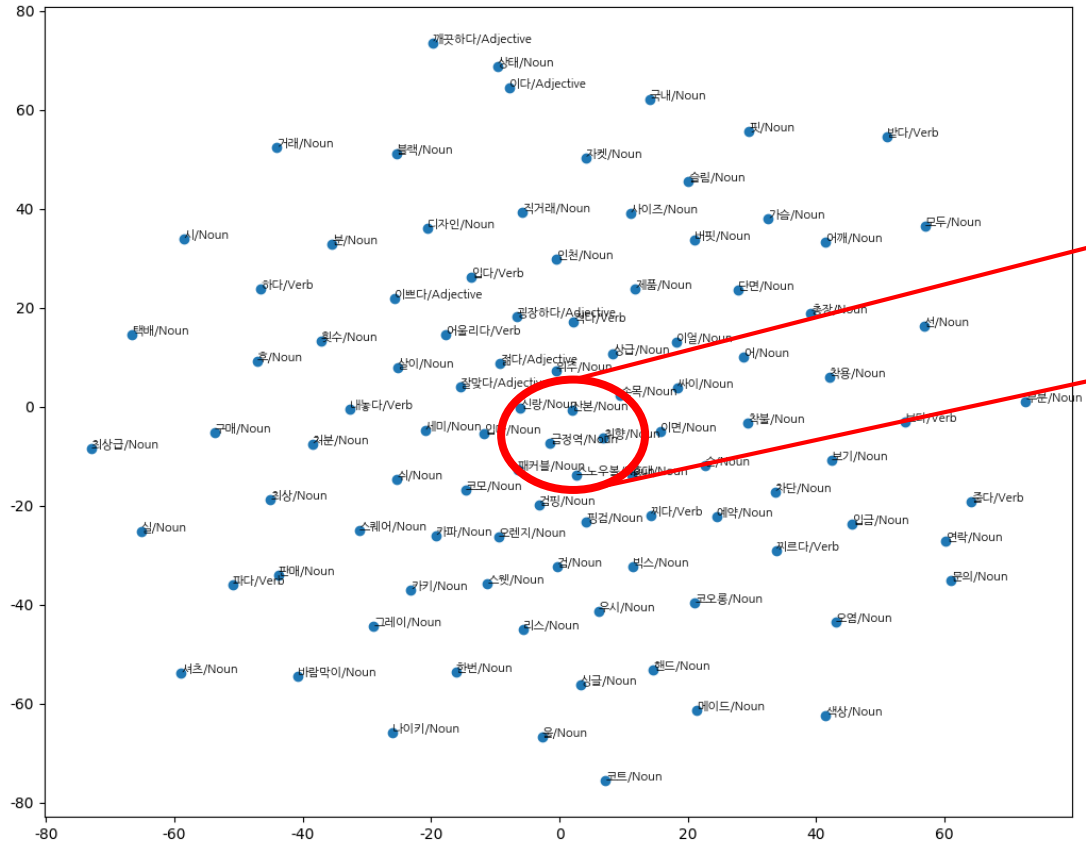
모델 파라미터 변경

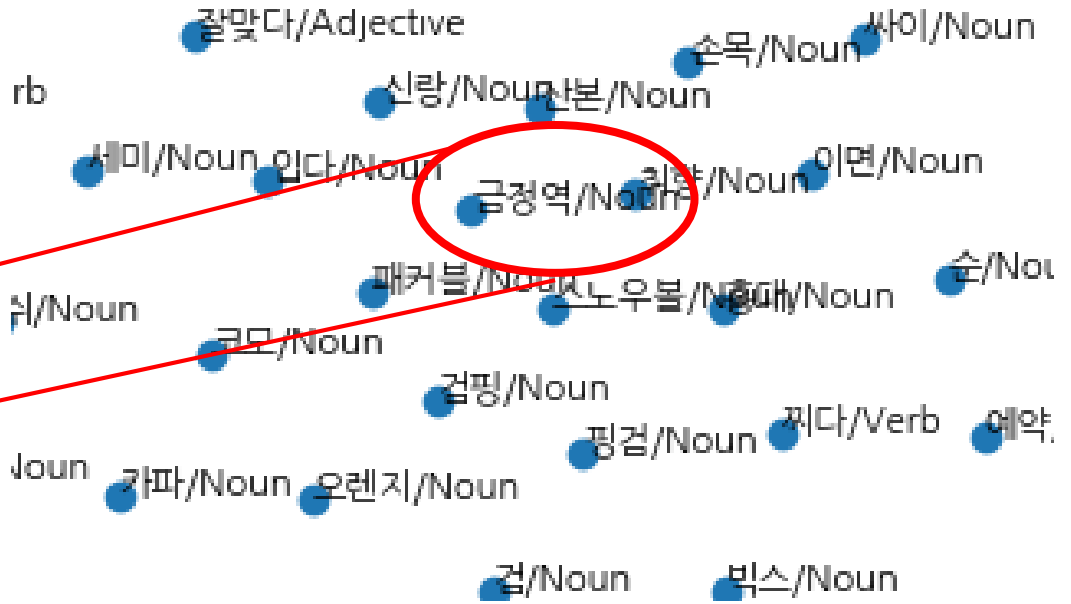
```
def makeM(train_data):
    #하이퍼 파라미터 변경하는 작업을 함
    num_features = 300    #문자 벡터 차원 수(100~300)크면 차원의 저주 최대 6천만 단어의 말뭉치에 300벡터
    min_word_count = 5
    #최소 문자 수(10~100) 최소 문자 빈도가 적으면 학습이 잘 안될 수 있음
    #단어마다 문맥 학습 최소 빈도수가 다름(ex 트와이스는 쉬운데, 아프리카tv, 대륙아프리카가ㅌ이 여러 의미로 사용되는)
    num_workers = 4     #병렬 처리 쓰레드 수(2or4)
    context = 2          #문자열 창 크기 (2~10) 창 크키가 크면 의미적 결과 작으면 구문적 결과
    downsampling = 1e-3  #문자 빈도 수 Downsample

    model = gensim.models.Word2Vec(train_data, workers=num_workers,
                                   size=num_features, min_count = min_word_count,
                                   window = context, sample = downsampling)
```

여러 예제들 참고 후 값 설정

모델 비교

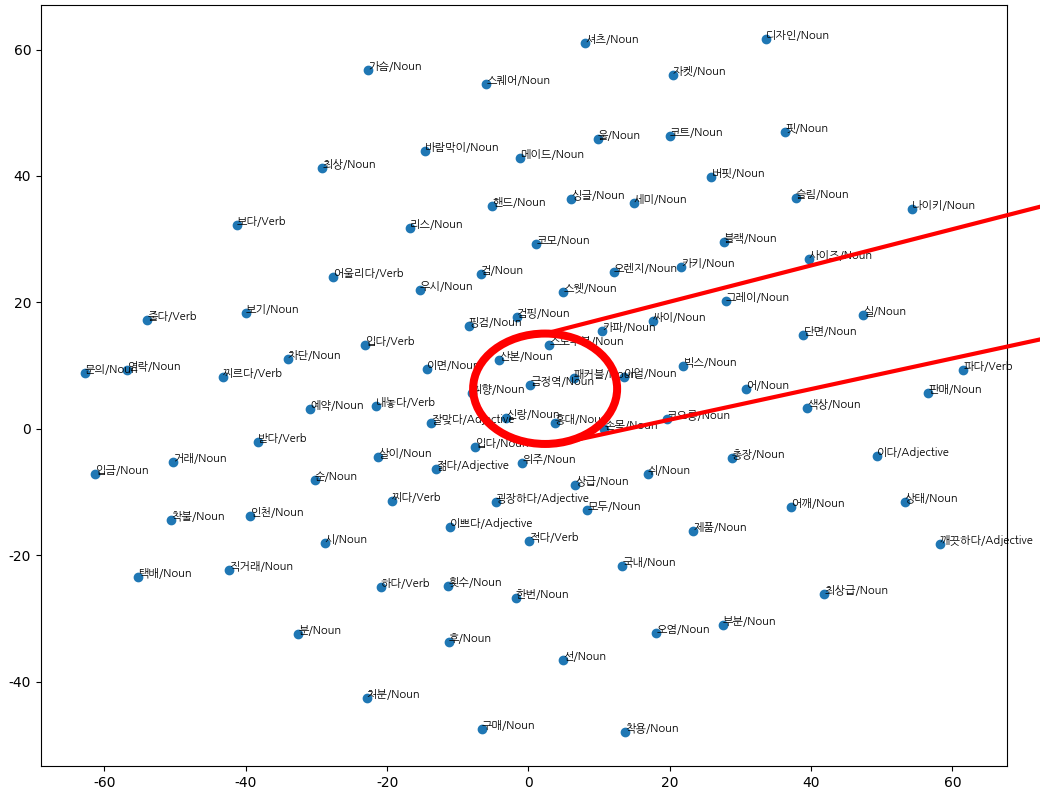


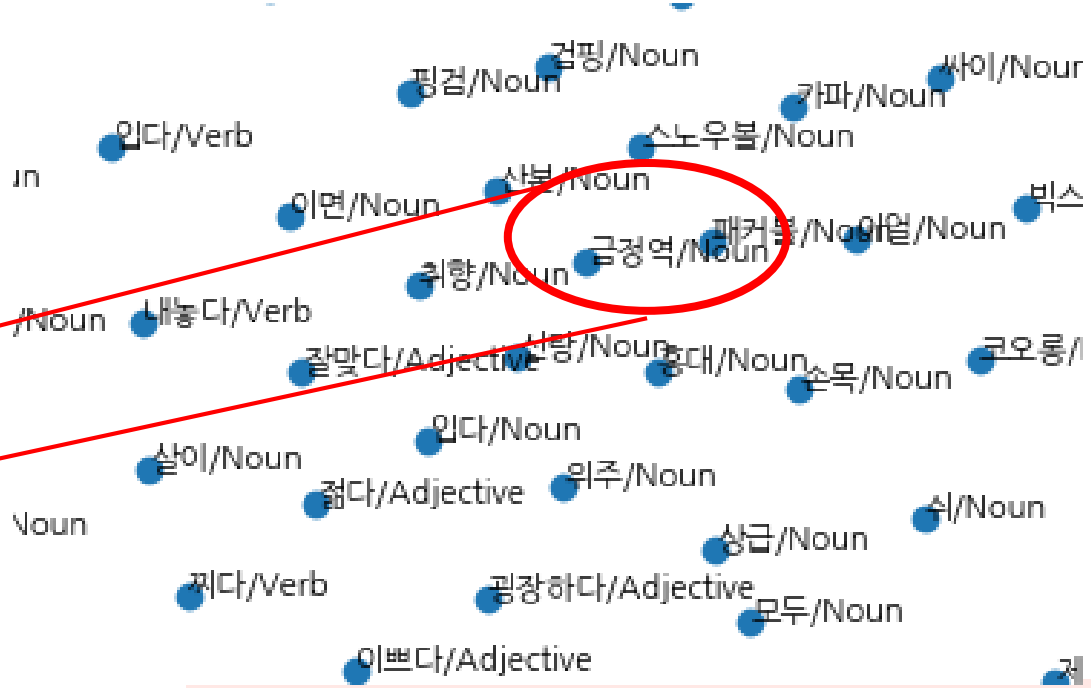


최소단어 수 / 윈도우 크기

5 / 2

모델 비교

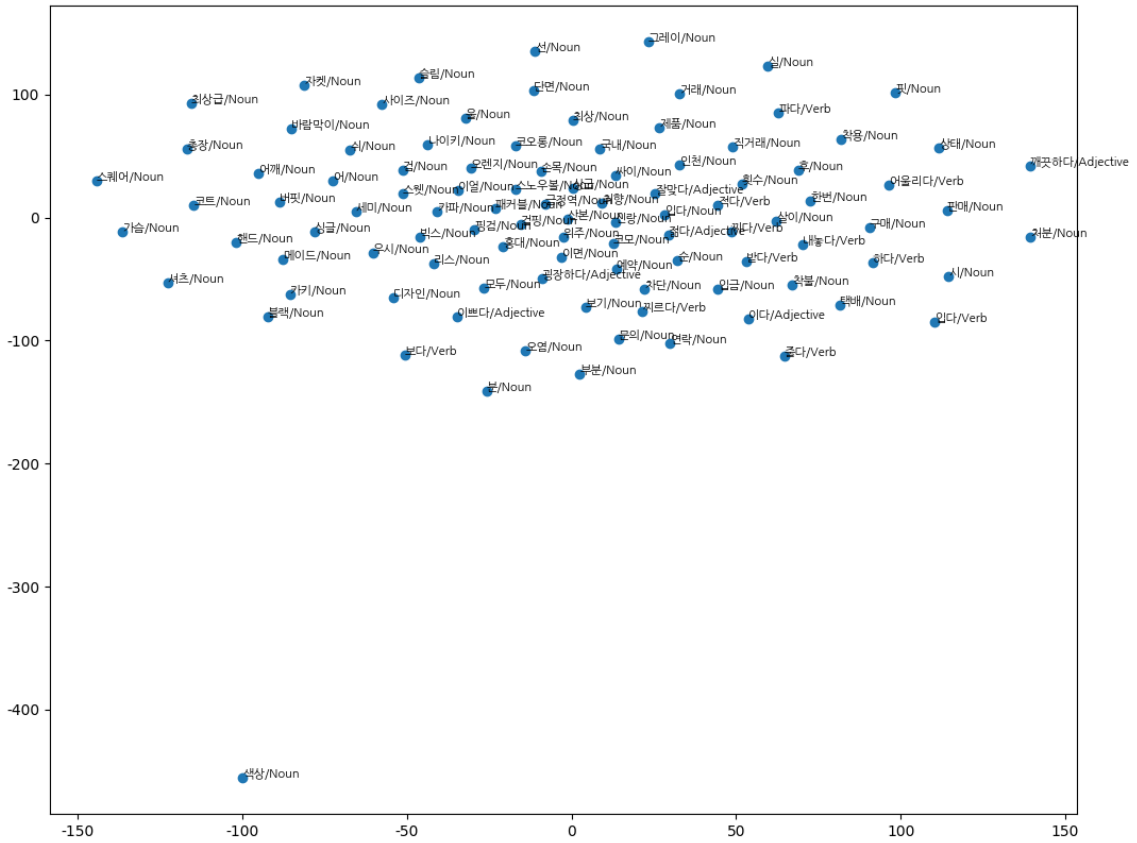




최소단어 수 / 윈도우 크기

10 / 2

모델 비교



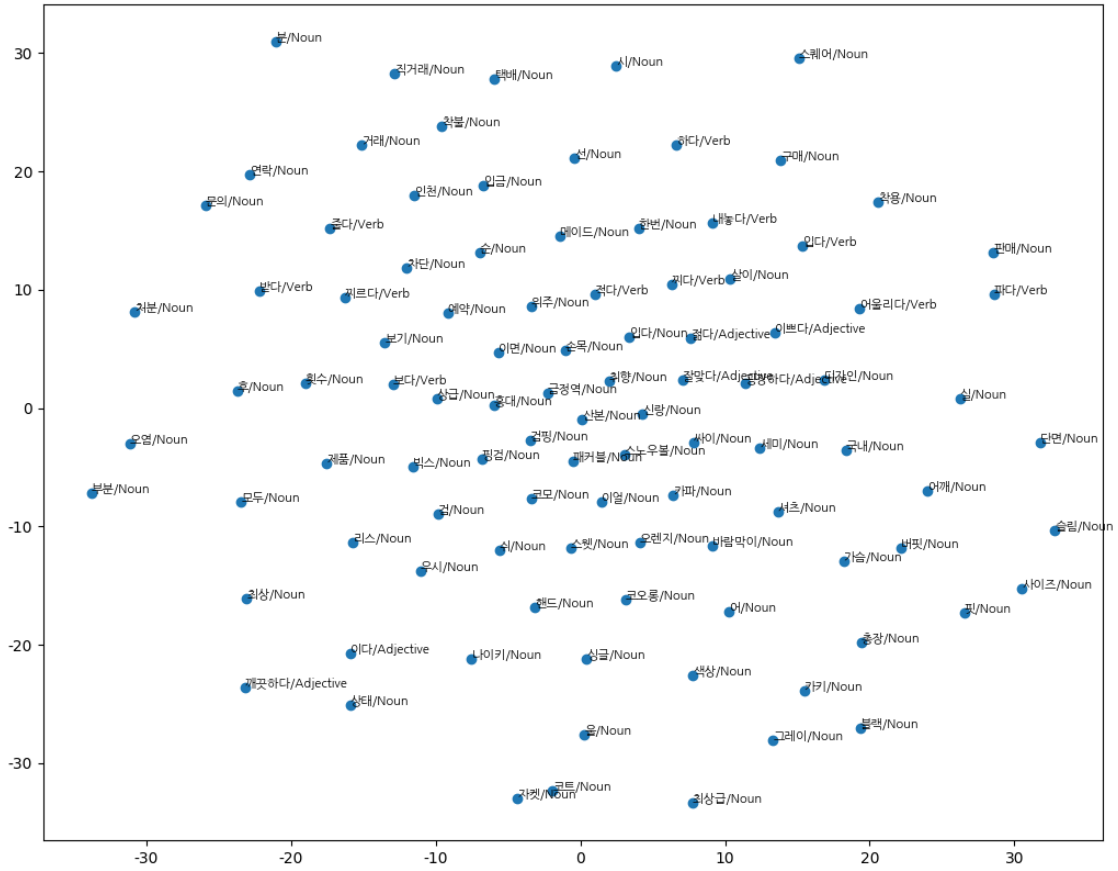
최소단어 수 / 윈도우 크기

10 / 6

하나의 예외값

토큰이 잘 안 나눠져 있음

모델 비교

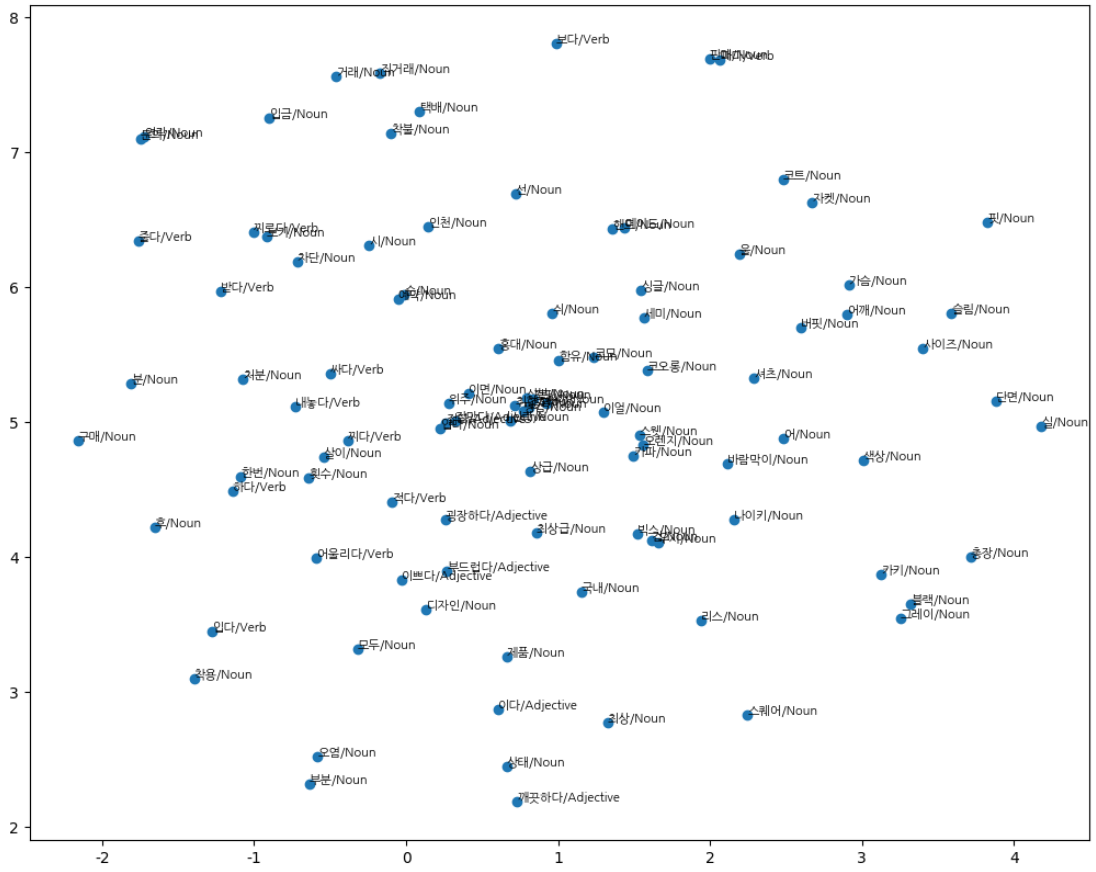


최소단어 수 / 윈도우 크기

10 / 4

잘 출력됨

모델 비교

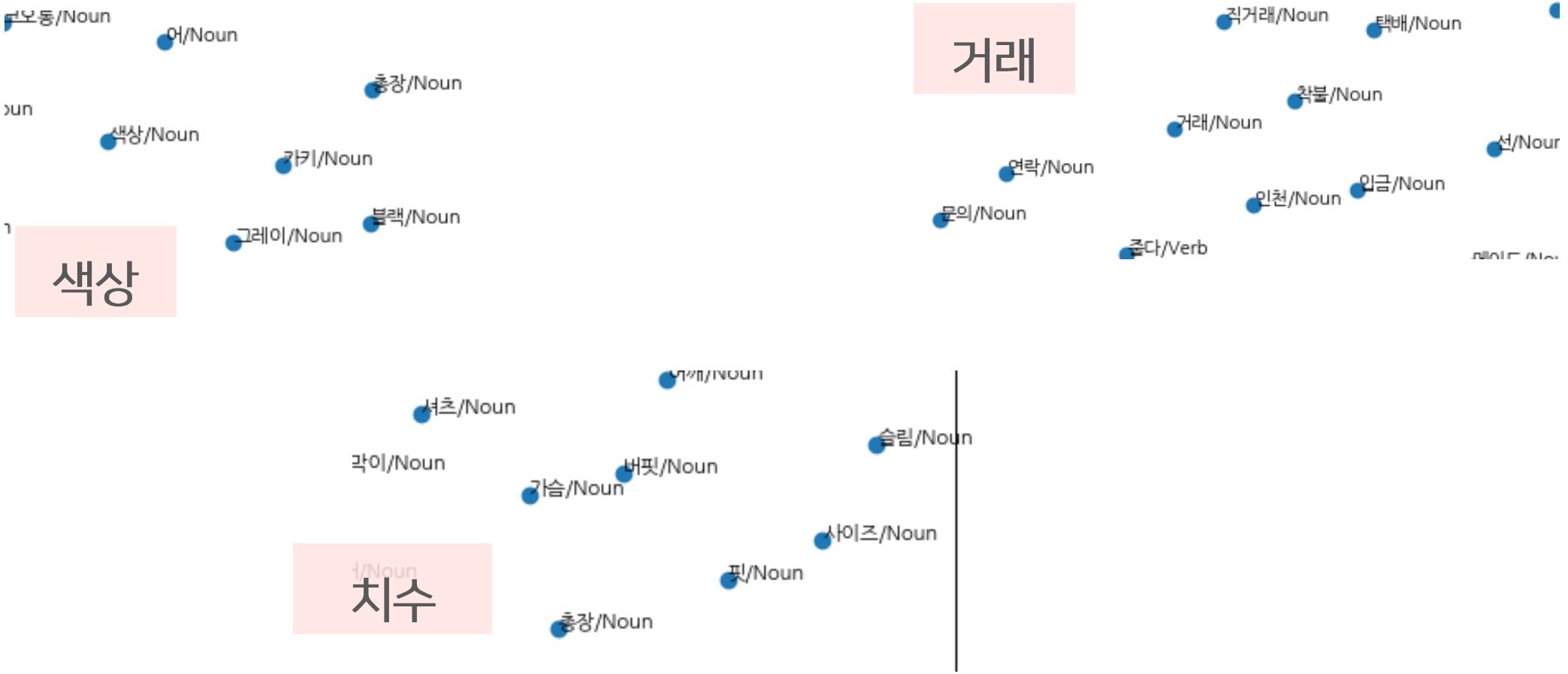


최소단어 수 / 윈도우 크기

20 / 4

잘 출력됨

모델 결과


어/Noun
색상/Noun
그레이/Noun
블랙/Noun
색상
직거래/Noun
택배/Noun
착불/Noun
거래/Noun
연락/Noun
입금/Noun
인천/Noun
문의/Noun
거래
슬림/Noun
버핏/Noun
가슴/Noun
사이즈/Noun
핏/Noun
총장/Noun
치수

다음주 개발 계획

박시현	우분투 서버와 크롤링 프로그램 연결 설정 통신 모듈 구현
박진영	웹 스토리지 구현 서버 통신 구현 시작
이민희	서버 모듈 연결 / 모델 파라미터 변경 상세 구조 정리

팀원 별 각 주간 개발 계획

박시현	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주제 확정/제안서 작성								중간 고사 & 중간 데모 준비						최종 데모	최종 리포트 제출
크롤링 · 전처리															
데이터셋 구축															
서버 구축															
통신 모듈 작업															
서버 통신 구현															
데모 준비 · 서비스 테스트															

팀원 별 각 주간 개발 계획

박진영	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주제 확정/제안서작성								중간 고사 & 중간 데모 준비						최종 데모	최종 리포트 제출
프론트엔드 화면 구현															
프론트엔드 동작 구현															
텍스트 대치 구현															
확장 앱 통신 구현															
웹 페이지 쿠키 구현															
테스트															

팀원 별 각 주간 개발 계획

이민희	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주제 확정/제안서작성								중간 고사 & 중간 데모 준비						최종 데모	최종 리포트 제출
자연어 처리 공부															
라이브러리 사용 전처리															
알고리즘 구체화															
문장분리															
주변 토큰 처리															
키워드 사전 변경															
모델 구축, 파라미터 변경															
서버 모듈 연결															
성능 및 정확도 향상															

전체 각 주간 개발 계획

공통: 연두색 / 박시현:핑크색 / 박진영:하늘색 / 이민희:주황색

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
주제확정/제안서 작성								중간 고사 & 중간 데모 준비						최종 데모	최종 리포트 제출
크롤링/데이터 확보															
데이터 전처리															
프론트 엔드															
서버 구축															
텍스트 대치 구현															
자연어 처리 모델 구축															
웹 페이지 쿠키 구현															
모듈 연결															
유닛 · 전체 테스트															

감사합니다! QnA