



## TodayNews 중간 결과 보고

Meeting Using DeepLearning

201601364 박주영(T) 201701138 김혜원 201503665 홍승환 201600599 김아연 201700124 이산가 비두샤



## 목차

- 01 서비스 소개
  - 02 시스템 구성도 및 흐름도
    - 03 시스템별 구현 현황
- 04 작품 시연
- 05 향후 계획 및 담당 업무



즉, **뉴스를 봐야겠다고 생각은 하나, 뉴스를 보지 않는 사람들**에게 오늘 하루의 뉴스를 요약하여 제공하는 어플입니다.



# Today News

News summary using machine learning

## 1. TN Services 소개

Today news어플리케이션을 통해 제공받을 수 있는 서비스입니다.



#### 1. 뉴스 요약본 제공

실시간으로 뉴스 요약본 을 제공합니다.



#### 2. 주제별 뉴스

뉴스를 군집화하여 주제 별로 묶어서 보여줍니다.



#### 3. 개인화 추천 기능

사용자별 맞춤 뉴스를 추 천하여 보여줍니다.



#### 4. 북마크 기능

사용자가 원하는 내용을 스크랩하여 후에 다시 볼 수 있습니다.



#### 5. 뉴스 브리핑 기능 군집화된 뉴스의 요약본 을 브리핑 해주는 기능이

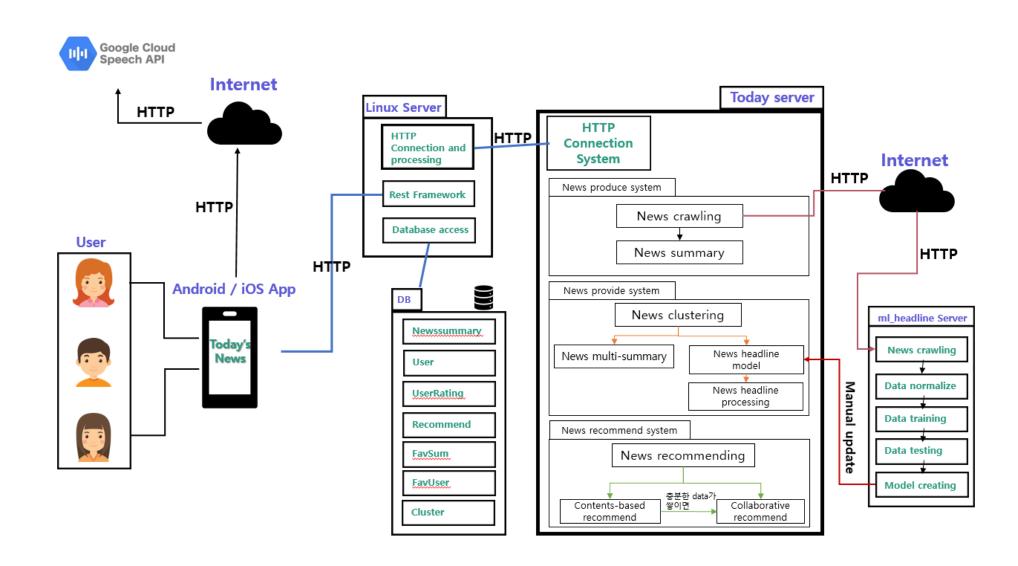


#### 6. 잠금화면

스마트폰의 잠금화면에서 실시간 뉴스를 제공받을 수 있습니다.



## 2. 시스템 구성도 및 흐름도 TN 시스템 구성도



# 2. 시스템 구성도 및 흐름도

TN 시스템 흐름도 실시간 기사 크롤링 TODAY NEW DB <u>Headline1</u> 추상적 요약 content1 각 기사 별 요약 뉴스 기사 요약본 추출적 요약

순서: **추천 시스템** 

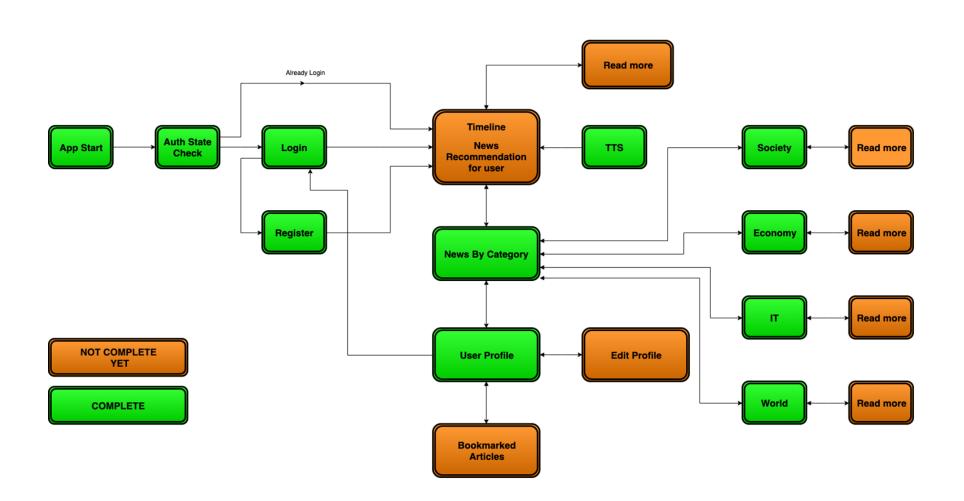
# 3. 시스템 별 구현 현황

#### 주요 기능 목록

유형	주요 기능	설명	진척도(%)
앱	로그인 및 회원가입	사용자가 이메일인증을 통해 회원가입을 하고 로그인 기능을 통해 서비스를 이용할 수 있다.	100%
	화면 제공	뉴스를 볼 수 있는 메인 화면과 카테고리 화면, 회원 정보와 스 크랩 기능이 있는 회원 정보 화면을 제공한다.	90%
	TTS기능	군집화된 뉴스의 요약본을 브리핑 해주는 기능이다.	100%
	스크랩 기능	사용자가 원하는 뉴스를 스크랩하여 뉴스 정보를 DB에 저장하 여 후에 다시 볼 수 있도록 한다.	0%
뉴스 생성 시스템	뉴스 크롤링	실시간으로 정치, 경제, 사회, IT 뉴스를 크롤링 한다.	100%
	뉴스 요약본 생성	Lexrank를 이용하여 뉴스를 3줄로 요약한 후 db에 저장한다.	100%
뉴스 제공 시스템	뉴스 클러스터링	K-means를 이용하여 주제별로 뉴스를 클러스터링한다.	70%
	헤드라인 생성	딥 러닝을 이용하여 생성된 모델을 통해 각 뉴스 군집의 헤드라 인을 추상적 요약 기법으로 생성한다.	70%
	다중 문서 요약	군집화된 뉴스들을 모아 다시 한번 요약하여 각 군집의 대표 요 약본으로 제공	80%
뉴스 추천 시스템	개인화 추천 뉴스 제공	콘텐츠기반, 협업필터링을 이용하여 개인별 맞춤 뉴스를 추천 한다.	90%

## 3-1. APP개발 세부 진행 사람

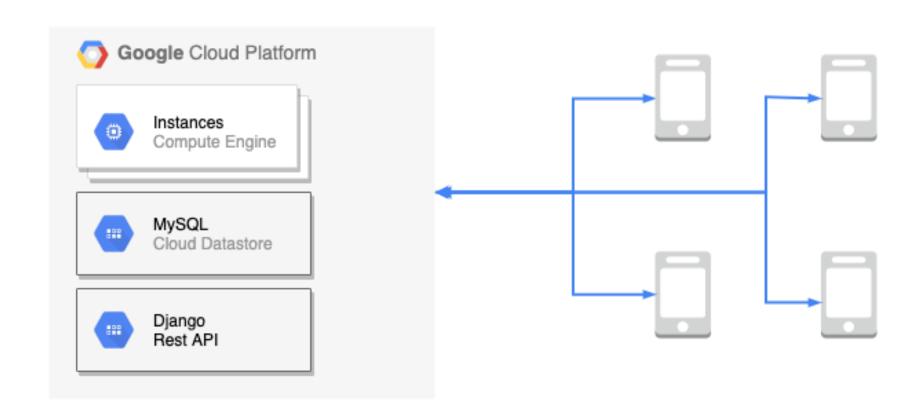
#### **MUD APP Flowchart**



### **Cloud Architecture**

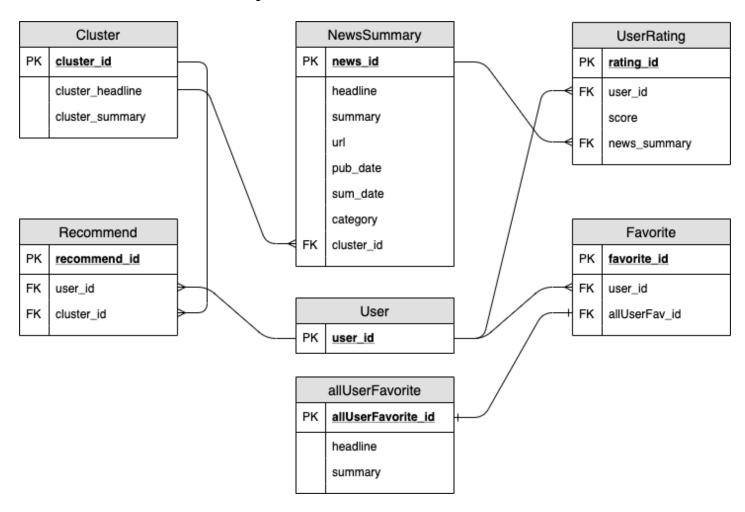


Architecture: Django > Rest API > Content Hosting



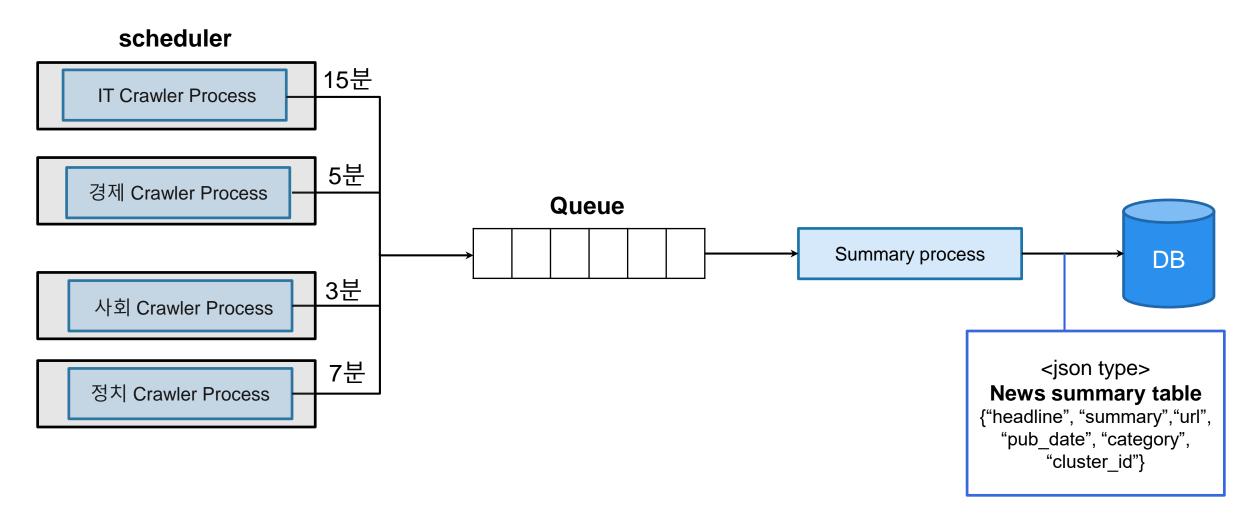
## 데이터 베이스 설계

#### MySQL Database



## 3-2. 뉴스 생성 시스템 구현 형황

#### **Crawling & summary**



## 3-2. 뉴스 생성 시스템 구현

#### Crawling &summary 결과

#### 1. 결과를 리스트로 출력

- 크롤링 후 뉴스 본문을 요약한 데이터를 리스트로 저장한다.
- Data = [오늘날짜,카테고리,언론사,제목,요약본문,ur,뉴스생성시간]으로 저장한다.
- 결과 예시: [ '20191119','society','연합뉴스','프랑스 남부서 현수교 붕괴,,,','스용차 타고 있던 15세 청소년 숨진 채 발견. 트럭 운전사도....','2019.11.19.12.50']

## 3-2. 뉴스 생성 시스템 구현

#### Crawling &summary 결과

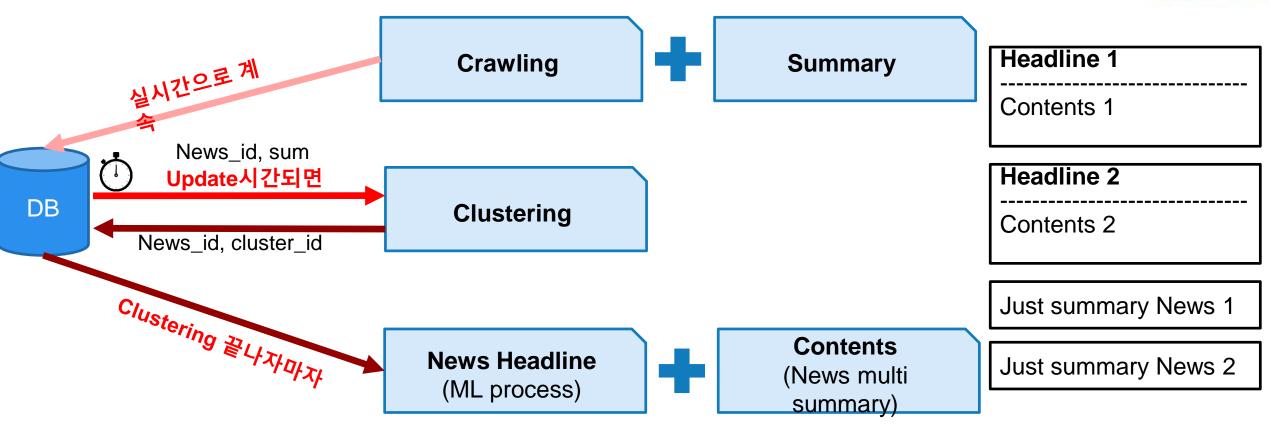
#### 2. DB에 json형식으로 결과를 저장

```
"news_id": "5165a1e2-2c04-48f6-ad67-faf4be51a426",
    "headline": ""선명도 미달" vs "번인 현상" LG·삼성 8K TV 주도권 전쟁",
    "summary": "LG전자는 삼성전자의 양자점발광다이오드 QLED TV를 화질 선명도가 떨어지는 유사 유기발광다이오드 OLED TV라 공격하고 삼성전자는 LG전자 OLED TV의 번인 Burn in·열화 현:
    "url": "https://news.naver.com/main/read.nhn?mode=LSD&mid=sec&sid1=101&oid=353&aid=0000035532",
    "pub_date": "2019-11-16T12:35:00+09:00",
    "sum_date": "2019-11-16T02:46:18.239363+09:00",
    "category": "경제",
    "cluster_id": "01873b9c-244d-4744-afb9-3a2dc34dbfd1"
},
```

- REST API를 이용해 News summary table의 url과 통신을 한다.
- Date = {"headline" : date[3], "summary" : date[4], "url": date[5], "pub\_date":date[6], "category":date[1], "cluster\_id":"..."} 의 json 형식으로 DB에 저장하며 앞서 생성된 data 리스트에서 알맞은 값을 value로 지정하여 준다.
- 요약된 뉴스는 실시간은로 앱에서 확인 가능하다.

## 3-3. 뉴스 제공 시스템 구현 형황

- K\_means 이용한 Clustering



Clustering 도출 결과 : 유사 주제를 가진 뉴스들을 묶는다.

묶음 속에서 중심 주제와 가까운 것은

Cluster\_id 부여 / 이 외 default

이때의 Cluster\_id는 UUID Cluster 당 포함 뉴스기사 수 == Cluster 크기

## 3-3. 뉴스 제공 시스템 구현 현황

- K\_means 이용한 Clustering



#### ②불용어 제거 및 tokenizer

토큰화 시 참조 API: Countvectorizer vs Tfidfvectorizer vs HashingVectorizer

## 3-3. 뉴스 제공 시스템 구현 형황

- K\_means 이용한 Clustering

#### <u>④최적의 K찾기</u>

-최적의 K값 찾기 방법 1) elbow 기법

2) 실루엣 기법



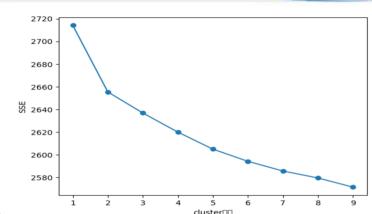
두 방법은 K값을 하나하나 다 넣어보고 나온 성능평가 결과를 의미한다. 나는 나는 그를 되고로 한 카테고리 당 24시간에 (사회 1만개, 경제 5천개, IT 2천개, 정치 4천개) crawling 시,

- 시간이 오래 걸린다. 1만 dataset을 range 6 내에 Kmeans □ 20분 소요
- 분산이 작을 시
- 문자 데이터에 대한 내용이므로 시각적 판단 기준이 없음 (분산도 그래프 작성의 어려움)
   골바른 clustering 여부

-해결책: K값에 최대 최소 범위를 지정 (10< K < 20): 추천, 보여지는 Headline의 편리성과 관련 \* 유동 K

: multiprocessing 이용

\* 고정 K



## 3-3. 뉴스 제공 시스템 구현 형황

- K\_means 이용한 Clustering

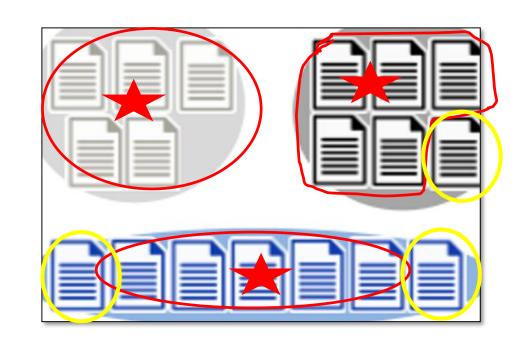
#### ⑥Centroid와 근접 기사 찾기

== MORE 선택 시 보여지는 기사들

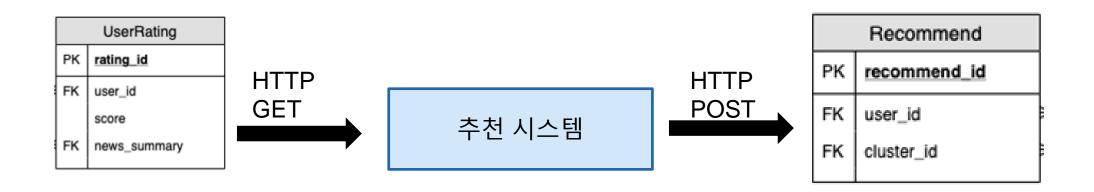


(전처리 된 data + 최적의 K) + Kmeans의 각 중심 centroid, Cn과의 vector data의 거리를 측정하여 가까운 순으로 list화

근접 기사로 측정된 것만 Cluster\_id 부여 / 나머지는 default



## 3-4. 추천 시스템 구현 진행 형황



#### 진행현황

UserRating 테이블에서 user\_id, score 와 news\_summary를 가져와 cluster\_id와 user\_id를 recommend 테이블로 업데이트 완료

## 3-4. 추천 시스템 구현 진행 현황

#### Rating table

# { "rating\_id": "055c8196-9b47-4289-8592-a586a7625696", "score": 5, "user\_id": "pFbDe7m55DYtXMgcarlBFgEFhYr1", "news\_summary": "b47b1456-d816-4f30-8d5c-9d5846a79bb9" }, { "rating\_id": "4f71ee52-e260-4382-8b30-5e9847765b6b", "score": 2, "user\_id": "R7TdA4s1Di05QwN7iTJL1ad6e2V2", "news\_summary": "6fd0f744-5758-45c1-94c5-84a69b893429"

#### Recommend table

```
"recommend_id": "862c007f-d49d-4265-8bce-814f7d2e18df",
    "user_id": "pFbDe7m55DYtXMgcar1BFgEFhYr1",
    "cluster_id": "58d0aa77-1664-4bed-9f45-0e28e8b11171"
},
```

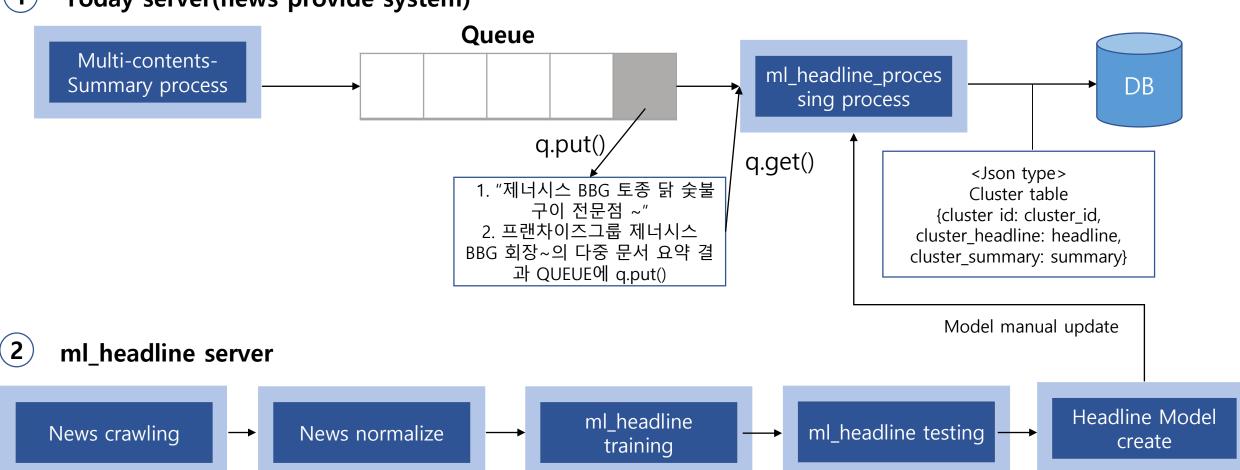
user\_id가 pFbDe7m55DYtXMgcarlBFgEFhYr1인 유저에 대한 새로운 뉴스 군집을 Recommend table에 등록 했다.

#### 앞으로의 계획

스케쥴러를 사용하여 주기적으로 추천 시스템 동작

## 3-5. 헤드라인 제공 기능 구현 진행 현황

1 Today server(news provide system)



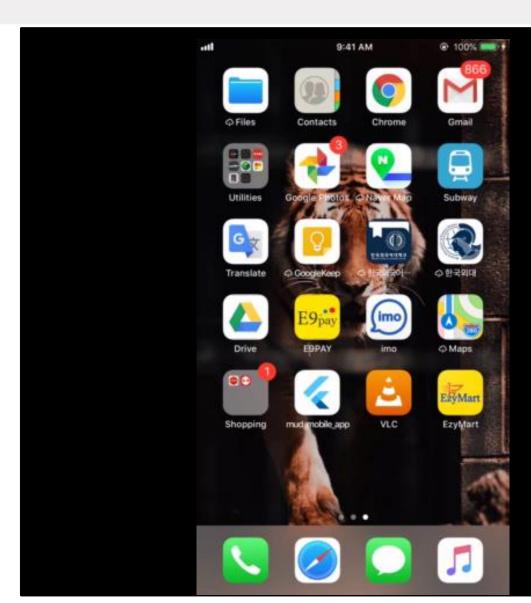
## 3-5. 헤드라인 제공 기능 구현 진행 현황

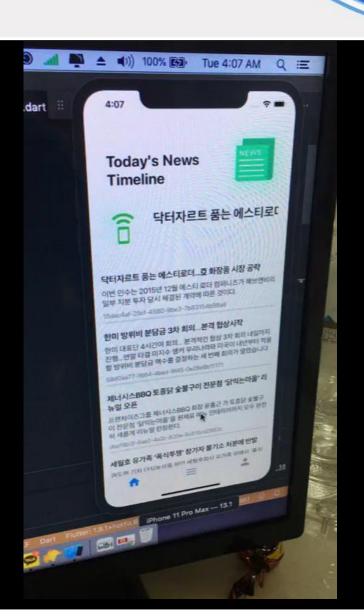
#### 1. 학습 결과

```
Step 수 .
                결과(loss,predition, edit result, actual result).
76450
                 step : 76450
                 time: 14166781839.02959
                LOSS : 36.35411338806152
prediction : 최 신 재 재 조 는 는 는 는
                 edit result : 최 신 책 물 을 것 오 마 자 비 < < < <
                 actual result : 최승호 신동호 책임 물 을 것 오후 마다 자리 비
                Prediction: 최승호 신동호 재귀 재귀 조사장 는 는 는 는
83860.
                 step : 83860
                 time: 14166868565.5033
                 LOSS: 35.155114364624026
                prediction : 롯 연 연 연 연 앞 앞 앞 앞 앞 앞 앞 앞 앞 일 실 실 실 실 < < edit result : 롯 연 앞 고 와 수 늘 어 < < < < <
                 actual result : 롯데마트 연말연시 앞두 고 와인 수요 늘
                94470
                   step : 94470
                   time: 14166991342.8276
                   LOSS: 36.3416446685791
                   edit result : 애 국 특 침 조 핀 펫
                   actual result : 애플 국내 특허 침해 조사 핀 펫 반도체 기술 훔쳤
                Prediction: 무 반도체 반도체 반도체 반도체 반도체 조사 조사 조사 조사 조사,
100640.
                 time: 14167064035.018682
                 LOSS: 35.92008285522461
                 prediction : 프 카 카 전 전 전 전 전 전 조 < < < < < < < <
                 edit result : 프 마 모 에 녹 카 의 캐 < < < < < < < < < < < < < < <
                 actual result : 프렌즈 마블 모노폴리 에 녹인 카카오 의 캐주얼
                Prediction:프렌즈 카카오 카카오 전대 전대 전대 전대 전대 전대,
```



## 4. 작품 시연 동영상





## 5. 향후 계획 및 담당

#### 주차 별 계획

프로젝트 기간	2019.9.9~2019.12.9												
개발 내용	프로젝트 기간												
	1주	2주	3주	4주	5주	6주	7주	8주	9주	10주	11주	12주	13주
아이디어 회의 및 계획				•									
프로젝트 관련 자료 조사					•								
역할 분담 및 핵심 기술 조 사							•						
사업 제안서 작성						•							
요구사항 및 상세 설계 작성							•						
서버 및 데이터 베이스 구현								•					
앱 구현									•				
뉴스 크롤링								•					
클러스터링 시스템구현										<u> </u>	<b>&gt;&gt;</b>		
요약 시스템 구성(추상적)										-	<b>&gt;</b>		
요약 시스템 구성(추출적)										•			
추천 시스템 구현										•			
개별 테스트 및 보안											•		
통합 테스트 및 보안												•	
최종 발표 및 시연													

● 완료

🛑 미완료

● 예정

## 5. 향후 계획 및 담당 <sup>향후 계획</sup>

개발 내용	11주	12주	13주	담당자
클러스터링 시스템 구현				김혜원
헤드라인 추상적 요약 구현				박주영
스크랩 기능 구현				이산가 비두샤
다중요약				홍승환, 김아연
뉴스 제공 시스템 병렬처리				홍승환, 김아연
개별 테스트 및 보안				전 팀원
통합 테스트 및 보안				전 팀원
최종 문서 수정				전 팀원
최종 발표 및 시연				전 팀원



## Thank You

감사합니다.

