중간 보고서

희소행렬 처리를 통한 트윗 분석 프로그램

강의 : Capstone 디자인 (001)

담당 교수 : 안용학, 양효식, 박기호

팀장 : 16010945 노수민 컴퓨터공학

팀원 : 16010573 이아현 컴퓨터공학

16011041 홍주희 컴퓨터공학

16011189 양승주 디지털콘텐츠

목 차

중간 보고서	희소행렬 처리를 통한 트윗 분석 프로그램
	작성 날짜: 2020년 11월 04일

1. 개요

1.1 개발 동기

기존의 convolution 연산이 이뤄지는 모델의 연산 부분을 수정해 기존 모델과 성능 비교가 가능한 프로그램을 조사했다. 실시간 반영이 가능한 트위터를 이용해 사용자의 언어습관을 분석하기로 했다. 해당 사용자가 평상시에 어떤 감성의 문장을 사용하는지, 어떤 단어를 자주 사용하는지 분석해주는 프로그램을 개발한다.

1.2 목표

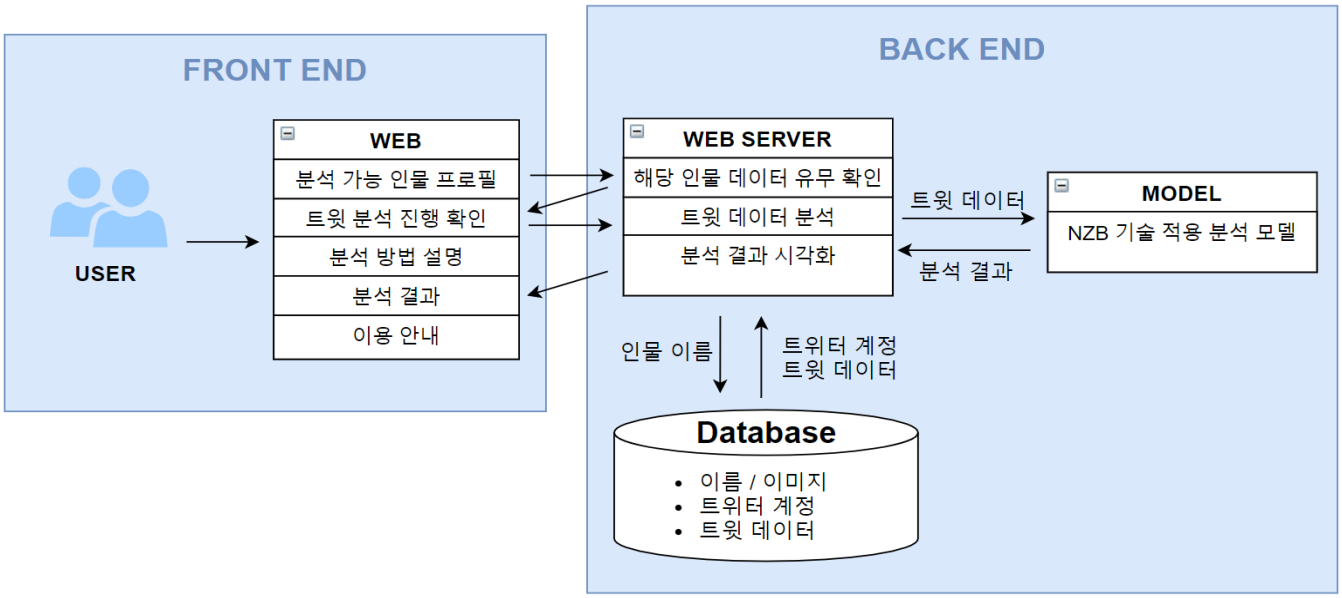
- 기존 모델의 convolution 연산 방식을 NZB 인덱싱 기법을 적용해 메모리 접근효율을 증가시킨다.

- 트위터 API 를 통해 유명인의 트윗 데이터를 수집하고, 게시글 속 문장의 감성을 분석하여 긍정, 중립, 부정 3 가지의 감성으로 분류해 문장의 비율과 해당 문장을 보여주며, 자주 사용하는 단어를 시각화 한다.

- 웹 기반으로 누구나 접근이 가능하며, 쉬운 UI 구성으로 사용하기 편하고, 오락성과 정보성을 동시에 포함하는 프로그램을 개발한다.

2. 시스템 구조

2.1 전체 시스템 구조



2.2 subsystem 설명

2.2.1 웹

- ABOUT 에서 분석 알고리즘 확인, CONTACT 에서 개발자에게 연락 가능하다.
- 분석 가능한 인물의 프로필을 확인 후, 트위터 분석을 요청할 수 있다.

2.2.2 웹 서버

① **node.js 서버**

- 사용자가 웹에서 분석 요청한 인물의 데이터가 존재하는지 데이터베이스에서 확인을 한다.
- 분석 모델에서 제공받은 분석 결과를 시각화 한다.

② **flask 서버**

- predict 엔드 포인트로 요청이 들어온 경우 텍스트 문서를 전달받아 전처리 함수와 단어 카운트함수, text-CNN 모델을 수행하는 함수 실행한다.

2.2.3 모델

- text-CNN 모델에 NZB Indexing 기술을 적용해 convolution 연산을 수행한다.
- 긍정/중립/부정 3 가지 타입의 결과로 문장의 감성분석을 진행 후, 타입 별 비율을 계산 후, 해당하는 3 가지 타입의 문장을 최대 2 개까지의 결과를 제공한다.
- 자주 사용하는 단어를 최대 5 개까지 제공한다.

2.2.4 데이터베이스

- mongoDB 서버 사용
- 50 명의 데이터는 "이름/트위터 계정/이미지/트윗 데이터" 구성으로 저장되어 있다.

3. 팀원 역할 분담

Task 목록	진행자
PM 및 문서 작성	노수민
회의록 작성	홍주희
PPT 제작 및 대본 초안 작성	양승주
분석 가능한 특정 유명인 조사	노수민, 홍주희
데이터 수집 및 전처리 (잘린 문장 복구)	노수민, 홍주희
워드 임베딩 진행	양승주
Text-CNN 모델 구현	이아현
학습 데이터 조사 및 수집	노수민, 양승주, 이아현, 홍주희
이모지 데이터셋 제작	홍주희
모델 학습	양승주, 이아현(SUB)
분석 데이터 전처리 과정 구현	노수민, 홍주희
문장 감성분석 결과 제공 구현	노수민, 홍주희
자주 사용하는 단어 카운트 구현	노수민
웹 설계 및 웹 서버 구현	양승주
NZB 인덱싱 기술 자료 조사	이아현
NZB 인덱싱 기술 적용	노수민, 이아현, 홍주희
데이터베이스 구축	양승주
모델 평가 및 검증	노수민, 이아현, 홍주희
분석 결과 시각화	양승주

4. 제안서 수정 내용

4.1 요구사항 수정 내역

이전 제안서 요구사항	수정된 제안서 요구사항
사용자의 언어습관을 분석한다.	사용자의 트윗 게시글의 감성분석을 한다.
긍정, 부정 단어의 비율 차이를 계산해 5 가지 유형으로 결과를 제공한다.	긍정, 중립, 부정 문장의 비율을 제공 후, 해당 감성의 문장을 예시로 최대 2 개 제공한다.
긍정적, 부정적 단어의 비율을 그래프로 시각화 한다.	긍정, 중립, 부정 문장의 비율을 막대 그래프로 시각화 한다.
word2vec 모델을 사용한다.	랜덤하게 단어 임베딩을 진행해, text-CNN 모델을 사용한다.
pytorch 프레임워크를 사용한다.	Keras 프레임워크를 사용한다.
실시간 트위터 분석으로 진행한다.	트위터 API 라이선스 문제로 유명인 50 명을 선정 후, 데이터를 직접 수집한다.
Frontend 에 React 를 사용해 개발한다.	Frontend 에 bootstrap 을 이용하여 개발한다.

5. 개발 추진 계획

5.1 데이터베이스 관련 유스케이스

일정	9 월			10 월				11 월			
	2	3	4	1	2	3	4	1	2	3	4
분석할 유명인 자료 조사											
트윗 데이터 수집											
트윗 데이터 전처리 (잘린 문장 복구)											
데이터베이스 구축											

5.2 웹 관련 유스케이스

일정	9 월	10 월				11 월			
	4	1	2	3	4	1	2	3	4
웹 & UI 설계									
웹 서버 구현									
분석 가능한 인물 프로필 조회									
분석 요청 화면									
분석 결과 조회 화면									
모델 배포 및 결과 제공									
NZB 기술 적용 모델 배포									
분석 결과 시각화									

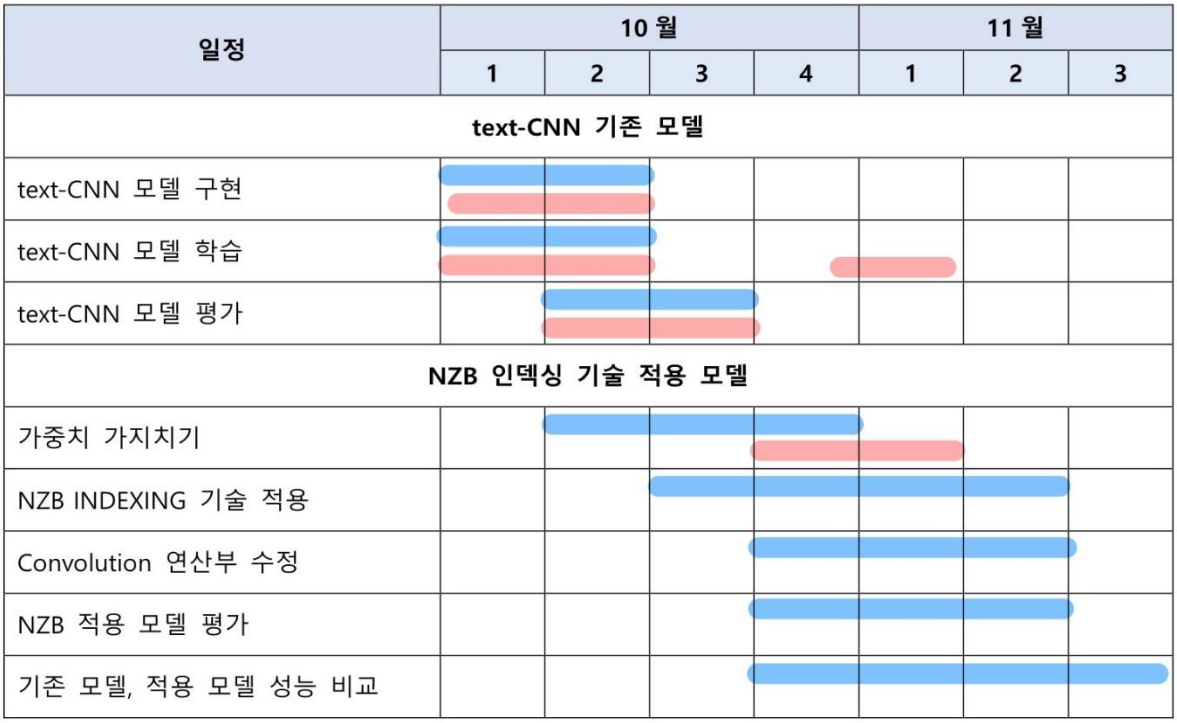
[웹 서버 구현] **(진행 중)**

분석 결과 조회 화면

- 3가지 각 감성 해당 문장 예시 2개 추출 및 각 감성 비율, 단어 5개 전달하여 결과 페이지에 표시
- mongoDB txt파일 & 타겟 정보 컬렉션 합치기 + txt 파일 접근

5.3 모델 관련 유스케이스

계획 실적



[text-CNN 기존 모델] **(진행 중)**

text-CNN 모델 학습

- Emoji sentiment 학습 데이터 셋 제작 완료 후 학습

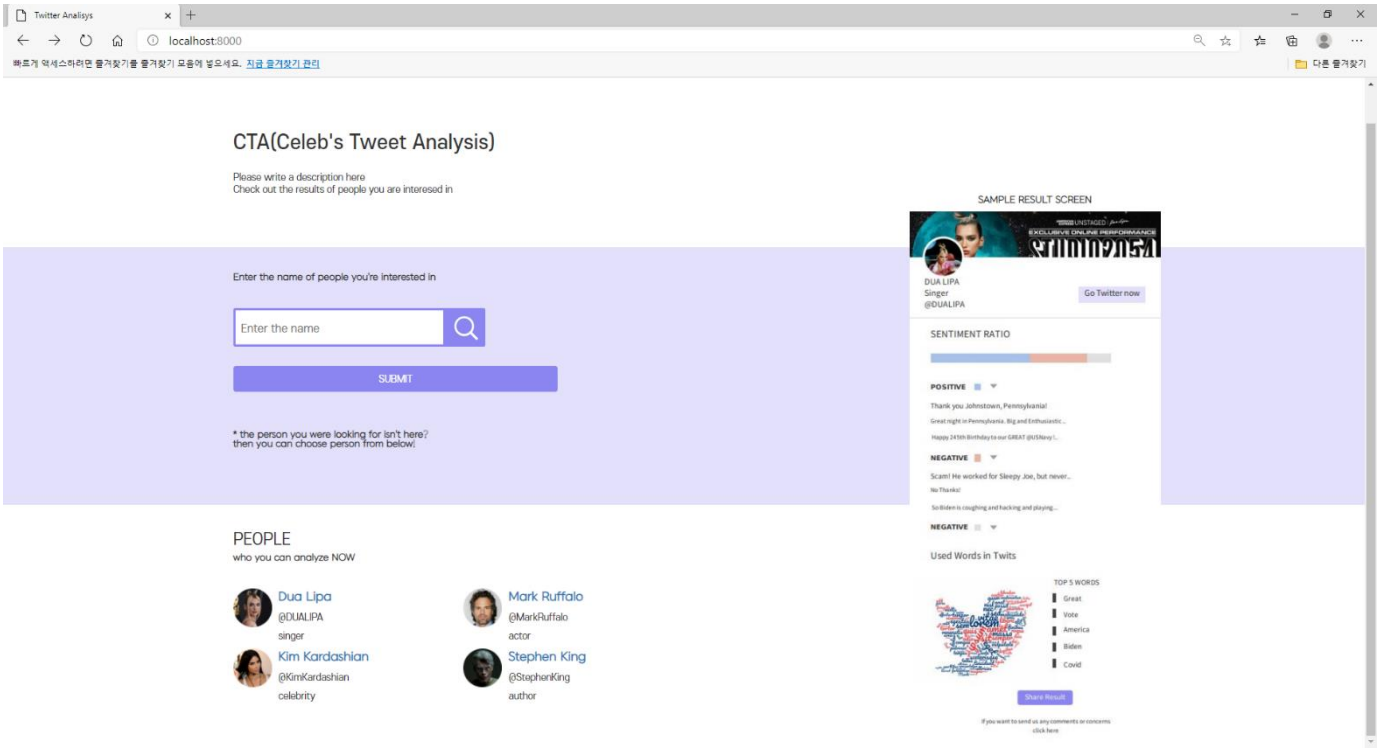
[NZB 인덱싱 기술 적용 모델] **(진행 중)**

가중치 가지치기

6. 현재 구현 내용

6.1 웹

6.1.1 분석 가능 인물 프로필 화면



6.1.2 데이터 값 가져오기


```
{
  "label": [
    "NEGATIVE",
    "NEGATIVE",
    "POSITIVE",
    "POSITIVE",
    "NEGATIVE",
    "POSITIVE",
    "NEGATIVE",
    "NEUTRAL",
    "NEGATIVE",
    "NEUTRAL",
    "NEUTRAL",
    "NEGATIVE",
    "NEUTRAL",
    "NEUTRAL",
    "NEUTRAL",
    "NEUTRAL",
    "NEUTRAL",
    "NEGATIVE",
    "NEGATIVE",
    "NEUTRAL",
    "POSITIVE",
    "NEGATIVE",
    "NEUTRAL",
    "NEGATIVE",
    "POSITIVE",
    "NEUTRAL",
    "POSITIVE",
    "NEGATIVE",
    "POSITIVE",
    "NEUTRAL",
    "POSITIVE",
    "NEUTRAL",
    "NEGATIVE",
    "NEGATIVE",
    "NEUTRAL",
    "NEGATIVE",
    "POSITIVE",
    "NEGATIVE",
    "POSITIVE",
    "NEUTRAL",
    "NEGATIVE",
    "NEUTRAL"
  ],
  "score": [
    0.3224688768386841,
    0.32626259326934814,
    0.8126254677772522,
    0.7995036840438843,
    0.32897132635116577,
    0.7307500839233398,
    0.07198041677474976,
    0.6186363697052002,
    0.2714122235774994,
    0.6173527240753174,
    0.608356237411499,
    0.28538769483566284,
    0.7583063840866089,
    0.23229213058948517,
    0.7931454181671143,
    0.6972705721855164,
    0.18583008646965027,
    0.6186363697052002
  ],
  "text": [
    "You are so great.",
    "I' m gonna stop tweeting coz",
    "If you' re a massive fan of someone and their music, can you
    "Imagine if I was actually like that hahahahahaha",
    "I just can' t get over how good this song is, I really outd
    "NEXT YEARS CALENDAR ? https://t.co/eW8jPaP8C1",
    "Let go of your story line. Try something new.",
    "WAHOOOOOOOOOOOOO https://t.co/JvkXq5aYUF",
    "OI OIIIIIIIIIII @BBCR1 https://t.co/5gvIFGyOUm",
    "YESSSSSSSS HAHHAHAHA https://t.co/zqwDrCBTkN",
    "Brain farts. I like that 🤣🤣 https://t.co/mSXwF6l1Z1",
    "Do you wanna know if we listen alike? Check out if we do her
    "❤️ @charliemackesy https://t.co/iLqJOGdtU8",
    "LISTEN! https://t.co/RDi2srZYIY",
    "A local barber in my neighbourhood was arrested for selling
    "Unbelievable. I' d been his customer for years and had no i
    "❤️❤️❤️❤️❤️❤️❤️ https://t.co/saSp789XkO",
    "I got fired from my job at the bank today.",
    "An old lady asked me to check her balance - so I pushed her
    "Want more jokes ?",
    "I asked my wife if I was the only one she' d been with. ",
    "She said yes, all the others had been nines and tens.",
    "Next joke in 10 minutes",
    "The man who invented Velcro died.",
    "RIP",
    "WHAT AN INCREDIBLE WOMAN! https://t.co/yfkUytMViT",
    "I went to a psychic and knocked on her front door. ",
    "She yelled. "who is it?" . ",
    "So I left.",
    "Working out is like a drug to me.",
    "I don' t do drugs.",
    "YEAAHAHHHAHAHHAHAHAHAHAHA https://t.co/DsmOEOeZXU",
    "My boss hates it when I shorten his name to Dick. ",
    "Especially when his name is Steve.",
    "You' ll have another one in half hour I gotta spread this ou
    "I met a girl the other night at a club who said she' d show
```

6.2 모델

6.2.1 기존 분석 모델

(@alicia_keys)

- 문장의 감성 비율, 자주 사용하는 단어 결과

```
트윗 문장 감정 비율
POSITIVE    59.15%
NEUTRAL     23.24%
NEGATIVE    17.61%
Name: label, dtype: object


자주 사용하는 단어 TOP5
[('love', 18), ('alicia', 15), ('album', 12), ('im', 11), ('yall', 9)]
```

- 문장의 감성 분석 결과

index	text	label	score	elapsed_time
0	Sending you some of that GOOD Sunday love!!!	POSITIVE	0.8906217813491821	0.4001450538635254
1	Tell me 3 things you're grateful for.... Me first: Breath. My magic. My crazy boys. 💜💜💜	POSITIVE	0.8792443871498108	0.033813714981079
2	My brother @inglewoodSiR buggin'!!!!	NEUTRAL	0.523128867149353	0.03246307373046875
3	Ask and you shall receive, our special version of 3 Hour Drive is available to stream right now!! ✨✨✨	NEUTRAL	0.5115101933479309	0.03363990783691406
4	I feel that #ALICIA love so deeply ya'll.	POSITIVE	0.7515761852264404	0.03367352485656738
5	Keep dreamin' and streamin' ✷✷✷ URL	POSITIVE	0.7906374335289001	0.03990459442138672
6	Sister talk vibes ✨✨✨	NEGATIVE	0.2882660627365112	0.03438615798950195
7	New side of Alicia to introduce you to.	POSITIVE	0.8801190853118896	0.03935313224792480
8	Nice to meet you.	POSITIVE	0.8783272504806519	0.04172873497009277
9	I'm ready 4 tonight at the @BBMAs at 8pm ET 💎💎💎	POSITIVE	0.8600454330444336	0.03644442558288574
10	I'm about to have crepes.	POSITIVE	0.7829102277755737	0.03274440765380859
11	Do ya'll like sweet or salty?	POSITIVE	0.7027351260185242	0.03195023536682129
12	WOW!! This one hits different.	NEUTRAL	0.6695632934570312	0.03291535377502441
13	Some of my favorite jewels and gems that you don't get to hear as much from my 3rd album As I am.	NEUTRAL	0.6075865030288696	0.03770542144775390
14	Vibe with me!!!!!!!! 💜💜💜	POSITIVE	0.8964023590087891	0.03359723091125488
15	All of this led me to the #ALICIA music!	POSITIVE	0.83182692527771	0.03318476676940918
16	If you know my mama...You DO NOT want her comin for you, so just listen!! 🧐🧐🧐	POSITIVE	0.8452038168907166	0.0375516414642334
17	love this mama!!	POSITIVE	0.8716044425964355	0.03399419784545898
18	Thank you for keeping us str8!! 💜💜💜	POSITIVE	0.933696985244751	0.0342407226562

(@liamgallagher)

- 문장의 감성 비율, 자주 사용하는 단어 결과

```
트윗 문장 감정 비율
NEUTRAL     38.16%
POSITIVE    37.63%
NEGATIVE    24.21%
Name: label, dtype: object


자주 사용하는 단어 TOP5
[('yeah', 24), ('ha', 24), ('x', 21), ('know', 21), ('fuck', 21)]
```

- 문장의 감성 분석 결과

67	It was shit	NEGATIVE	0.1738521158695221	0.03418469429016113
68	So good	POSITIVE	0.7842897176742554	0.04013562202453613
69	Zzzzzzzzzzzzzz	NEUTRAL	0.5053911209106445	0.03958821296691895
70	He ruined mind games boring as fuck he should stick to talking shit with his lover MM	NEGATIVE	0.254721999168396	0.03353500366210937
71	Oasis are getting back together soon	NEUTRAL	0.5588913559913635	0.03488326072692871
72	Thanks	POSITIVE	0.9069886207580566	0.03428053855895996
73	All good things come in 3s wink wink	POSITIVE	0.8096407651901245	0.0355985164642334
74	No he's on another level I'm a fan	POSITIVE	0.8019123077392578	0.03391943435668945
75	Ha ha	POSITIVE	0.8931021094322205	0.03283858299255371