UML 시퀀스 다이어그램

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

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 감정의 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 감정을 분석하여 긍정, 중립, 부정 3 가지의 감정으로 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.

1.3 기대 효과

- CNN, LSTM 같은 모델의 연산과정에는 희소 행렬, 밀집 행렬을 포함한 행렬을 통해 연산 과정이 이뤄진다. 이 연산 과정에서 기존에 자주 사용하는 CSR 방식이 아닌, 3X3행렬 크기를 표현하는 4비트와 인덱스를 나타내는 9비트, 총 13비트를 통해 최대 9개의 원소에 접근가능한 NZB 인덱싱 기법을 적용하면 전체 인덱스 사이즈와 메모리 접근 횟수 측면에서 높은 효율을 가질 것이다.

- 트위터 API 유료 버전을 사용할 경우 데이터를 데이터베이스에 저장하고 불러오는 방식이 아닌 사용자가 검색한 특정 인물의 트위터 게시글을 실시간으로 가져와 분석이 가능하다.

1.4 개발 일정

일정	9월			10월				11월				12월
	2	3	4	1	2	3	4	1	2	3	4	1
프로젝트 기획												
요구사항 분석 & 개발 환경 구축												
데이터 수집												
데이터 전처리												
웹 & UI 설계												
워드 임베딩 단계 구현												
Text-CNN 모델 구현												
Text-CNN 모델 평가 및 검증												
NZB 기술 적용												
웹 & UI 구축												
데이터베이스 구축												
NZB 적용 모델 학습												
시각화												
NZB 적용 모델 평가 및 검증												
최종 문서 작성 & 발표												

표에서 회색 색칠한 부분이 계획한 개발 일정이고, 화살표로 나타낸 부분이 실제로 진행된 개발 일정이다.

2. 요구사항 조사

2.1 선행 조사

2.1.1 시장성 조사

- DeepMoji

영어 문장의 감정을 분석해 여러 이모지로 표현한다.

차이점: 한 문장을 대상으로 여러가지 감정으로 표현하지만, 프로젝트에서 만들 웹 애플리케이션은 여러 문장을 분석하여, 긍정, 중립, 부정 문장으로 나누어 통계를 낸다.

- Watson Tone Analyzer

영어, 프랑스어 2 가지 언어의 트위터, 온라인 리뷰, 텍스트를 7 가지의 톤으로 분석한다.

차이점: 하나의 게시글을 가지고 분석을 해주지만, 프로젝트에서 만들 웹 애플리케이션은 게시글을 분석하여, 긍정, 중립, 부정 문장으로 나누어 통계를 낸다.

2.1.2 기술 조사

- Keras

사용자가 손쉽게 딥 러닝을 구현할 수 있도록 도와주는 상위 레벨의 인터페이스다.

① Tokenizer()

토큰화와 정수 인코딩에 사용한다.

② pad_sequence()

모델의 입력을 사용하기 위해 모든 샘플의 길이를 동일하게 맞출 때 사용한다. 정해준 길이보다 길 경우 샘플을 자르고, 길이가 짧은 경우 0으로 채운다.

③ Embedding()

단어를 밀집 벡터로 만드는 역할을 한다. 정수 인코딩이 된 단어들을 입력 받아 임베딩을 수행한다. 총 단어의 개수, 임베딩 벡터의 출력 차원, 입력 시퀀스의 길이를 함수 인자로 받는다.

④ Sequential()

model로 선언한 뒤에 model.add()라는 코드를 통해 층을 단계적으로 추가한다.

층을 입력층, 은닉층, 출력층으로 구성할 수 있다.

⑤ Dense()

전결합층을 추가한다. 출력 뉴런의 수, 입력의 차원, 활성화 함수 종류를 함수의 인자로 받는다.

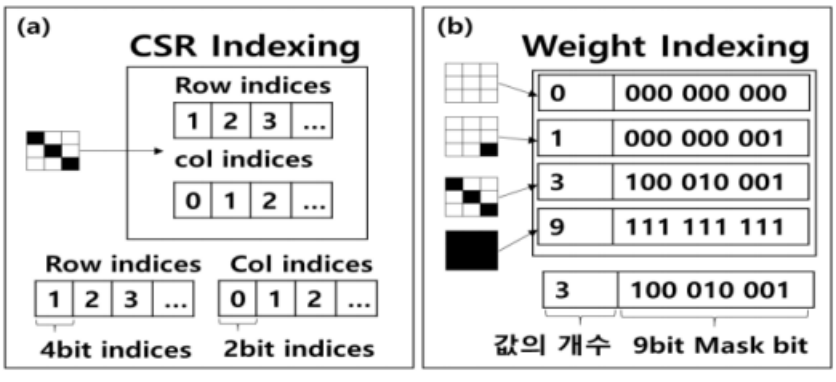
- CSR indexing

희소행렬 처리 기법 중 하나로, 행렬 내의 0이 아닌 값에 행과 열로 표현되는 인덱스를 부여하여 해당 값의 행렬 내 위치정보를 알려주는 방식이다. Convolution 연산에서 주로 파라미터 압축을 수행하는 용도로 사용하고 있다.

CSR indexing에서는 가중치 행렬 내부에서 0인 원소를 2bit, 0이 아닌 원소를 4bit로 표현한다.

- NZB indexing

가중치 행렬 내부의 0이 아닌 원소의 개수 정보와 행렬 내 원소의 값이 0의 값인지를 나타내는 bitmap 정보로 인덱싱을 수행한다.



2.2 요구 사항 명세서

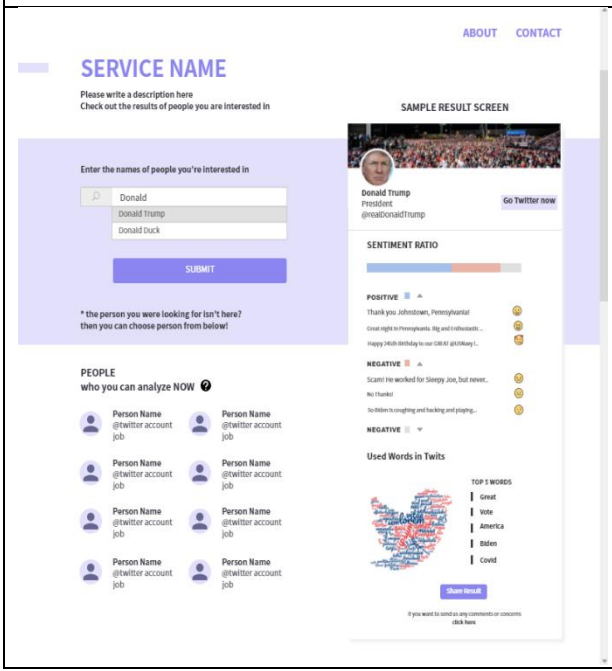
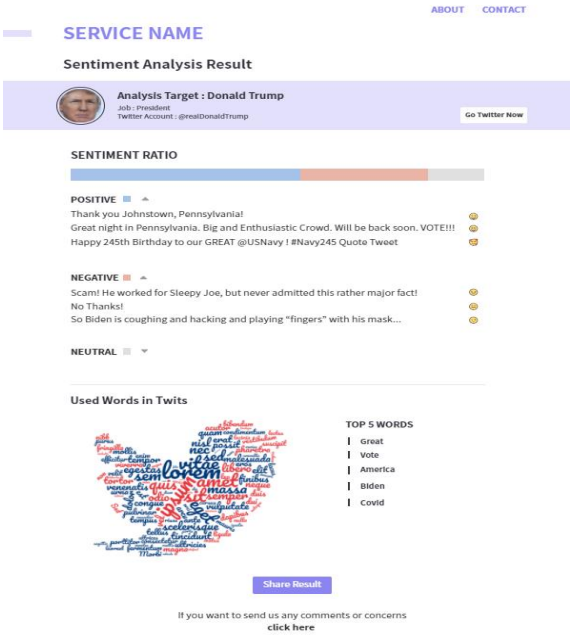
2.2.1 인터페이스 요구사항

번호	요구사항 내용
R-I-01	트위터 API 라이선스 문제로 유명인의 트윗을 데이터 관리자가 매주 화요일마다 데이터를 수집 후, 글자 수 제한으로 잘린 문장을 확인해 txt 파일 형식으로 저장 후, 데이터베이스에 업데이트한다.
R-I-02	트위터 API 라이선스 문제로 데이터베이스에 저장된 50 명의 유명인 트윗만 분석이 가능하다.
R-I-03	데이터 분석 모델 훈련 시 영어 데이터를 사용해, 영어가 아닌 다른 언어의 분석은 불가능하다.
R-I-04	PC/모바일에 따라 별도의 화면을 제공한다. 브라우저에 제한을 받지 않는다.

2.2.2 기능 요구사항

- **트윗 분석 프로그램**

<웹 실행화면>

웹 처음 화면	웹 분석 결과 화면
	

요구사항명	데이터베이스내 유명인 목록 표시	ID	R-F-01	우선순위	중
요구상황설명	사용자가 분석 가능한 유명인들을 웹에서 확인한다.				
해결방안	분석 가능한 50명의 유명인들 프로필을 표시하고, 클릭하면 분석이 가능하다. 프로필에는 사진, 이름, 트위터 아이디를 표기한다.				
위험요소	사용자가 프로필을 실수로 클릭할 경우를 위해, 선택된 인물의 분석을 진행할지 확인하는 안내창을 보인다.				
설계 시 고려사항	50명의 프로필을 화면에 나타낼 경우, 가독성에 지장이 생길 수 있다.				
관련요구사항	분석 결과 조회				
시나리오	1. 사용자가 웹페이지 하단에 존재하는 분석 가능한 유명인의 프로필을 확인한다. 2. 분석을 하고 싶은 유명인의 프로필을 클릭한다. 3. 해당 유명인의 트윗 분석을 진행할지 확인하는 안내창이 뜬다.				

요구사항명	유명인 이름 검색	ID	R-F-02	우선순위	하
요구상황설명	사용자가 프로그램에서 분석하고 싶은 유명인 이름을 검색한다.				
해결방안	검색창에 유명인의 이름을 입력할 경우, 자동완성 기능으로 검색을 돕는다.				
위험요소	데이터베이스에 존재하지 않는 경우, 분석이 불가능하다고 안내한다.				
설계 시 고려사항	유명인의 이름, 예명을 정확하게 사용해야 한다.				
관련요구사항	데이터베이스내 유명인 목록 표시				
시나리오	1. 사용자가 웹페이지 중앙의 검색창에 유명인의 이름을 입력한다. 2. 해당 유명인이 존재할 경우 자동완성 기능으로 입력을 돕는다.				

요구사항명	트윗 분석 요청	ID	R-F-03	우선순위	상
요구상황설명	사용자가 프로그램에서 유명인의 트윗 분석을 요청한다.				
해결방안					
위험요소	해당 유명인의 트윗 데이터(영어)를 수집한 후 진행한다는 점을 공지한다. (2020.09.22 부터 매주 화요일마다 txt 파일로 저장)				
설계 시 고려사항					
관련요구사항	데이터베이스내 유명인 목록 표시, 유명인 이름 검색				
시나리오	1. 사용자가 '데이터베이스내 유명인 목록 표시', '유명인 이름 검색'을 통해 분석할 인물을 선택한다. 2. 인물을 선택 후, 분석을 진행할지 확인하는 안내창을 통해 분석을 요청한다.				

요구사항명	트윗 분석 결과 조회	ID	R-F-04	우선순위	상
요구상황설명	사용자가 확인하고 싶은 유명인의 트윗 분석 결과를 화면에 표시한다.				
해결방안	데이터 분석 결과를 통계 제출 후, 시각화를 통해 화면에 표시한다. 1. 긍정 / 중립 / 부정 문장의 비율과 해당 감정의 문장을 최대 2개 제공한다. 2. 자주 사용하는 단어를 최대 5개 태그 형식으로 표현한다.				
위험요소	사용자가 긍정 / 중립 / 부정 의 판단 기준을 궁금해 할 수 있다.				
설계 시 고려사항	긍정 / 중립 / 부정 의 기준을 잘 설정해야 한다. 데이터 수집 기간을 정확히 표시해야 한다.				
관련요구사항	트윗 분석 요청				
시나리오	1. 사용자가 트윗 분석을 요청해, 유명인 트윗 분석을 한다. 2. 분석 결과를 화면에 표시한다. 3. 사용자는 유명인의 트윗 게시글 내 문장의 감정 비율과 자주 사용하는 단어를 확인한다.				

요구사항명	트윗 분석 결과 공유	ID	R-F-05	우선순위	하
요구상황설명	유명인의 트윗 분석 결과를 외부로 공유한다.				
해결방안	분석 결과를 조회할 때, 공유하기 버튼을 추가한다. 페이스북, 트위터, 카카오톡 공유를 가능하게 한다.				
위험요소	공유 가능한 범위를 지정한다.				
설계 시 고려사항	많은 사람들이 사용하는 사이트의 공유 기능은 필수로 제공한다.				
관련요구사항	트윗 분석 결과 조회				
시나리오	1. 사용자가 트윗 분석 결과를 조회한다. 2. 사용자가 웹 외부로 분석 결과를 공유한다.				

요구사항명	프로그램 이용 안내 확인	ID	R-F-06	우선순위	중
요구상황설명	웹 애플리케이션의 이용 안내를 확인한다.				
해결방안	웹 상단의 about 카테고리를 통해 이용 안내를 확인한다.				
위험요소					
설계 시 고려사항	웹의 어느 페이지에서도 접근 가능하도록 한다. 사용자가 사용하기 쉬운 UI로 구성한다.				
관련요구사항					
시나리오	1. 사용자가 웹 상단의 about을 클릭한다. 2. 사용자는 개발 의도, 서비스 설명(분석 방법), 개발자 소개를 확인한다.				

- **데이터 분석 모델**

요구사항명	문장의 감정분석	ID	R-F-07	우선순위	상
요구상황설명	트윗 문장의 감정분석을 한다.				
해결방안	문장의 감정을 긍정 / 중립 / 부정 3가지로 분류한다. 분류 기준 -> 긍정 : / 중립 : / 부정 :				
위험요소	문장 속 사전에 학습되지 않은 단어가 존재할 경우 분류의 정확도가 낮다.				
설계 시 고려사항	문장 속 이모지의 감정을 나타낸다.				
관련요구사항	트윗 분석 프로그램 : 트윗 분석 요청, 데이터베이스 : 분석할 데이터 제공				
시나리오	1. 데이터베이스에서 분석할 유명인의 트윗 데이터(txt)를 불러온다. 2. 트윗을 문장 단위로 분석해 문장의 긍정, 중립, 부정의 비율을 계산 후, 해당 감정의 문장 2개를 제공한다.				

요구사항명	자주 사용하는 단어 통계	ID	R-F-08	우선순위	중
요구상황설명	트윗 속 자주 사용하는 단어를 통계 낸다.				
해결방안	자주 사용하는 단어 최대 5개까지 태그 형식으로 제공한다.				
위험요소					
설계 시 고려사항	특수 문자, 이모지를 사용하는 트윗이 많아 이모지도 같이 통계를 낸다.				
관련요구사항	트윗 분석 프로그램 : 트윗 분석 요청, 데이터베이스 : 분석할 데이터 제공				
시나리오	1. 데이터베이스에서 분석할 유명인의 트윗 데이터(txt)를 불러온다. 2. 자주 사용하는 단어의 통계를 낸 후, 최대 5개까지 제공한다.				

2.2.3 성능 요구사항

번호	요구사항 내용
R-P-01	기존 모델의 convolution 연산에 NZB 인덱싱 기법을 적용한 모델을 개발한다.
R-P-02	기존 연산 모델과 NZB 인덱싱 기법을 적용한 모델의 성능을 측정하고 비교한다. 비교 항목 : 연산 속도, 메모리 접근 횟수, 정확도
R-P-03	실시간 처리를 통해 신속하고 정확하게 화면에 분석 결과 출력한다. -> 분석 시간 최장 5초 이내 처리한다. -> 10 초 이상 지연 시 오류 메시지 사용자에게 제시한다.

2.2.4 데이터베이스 요구사항

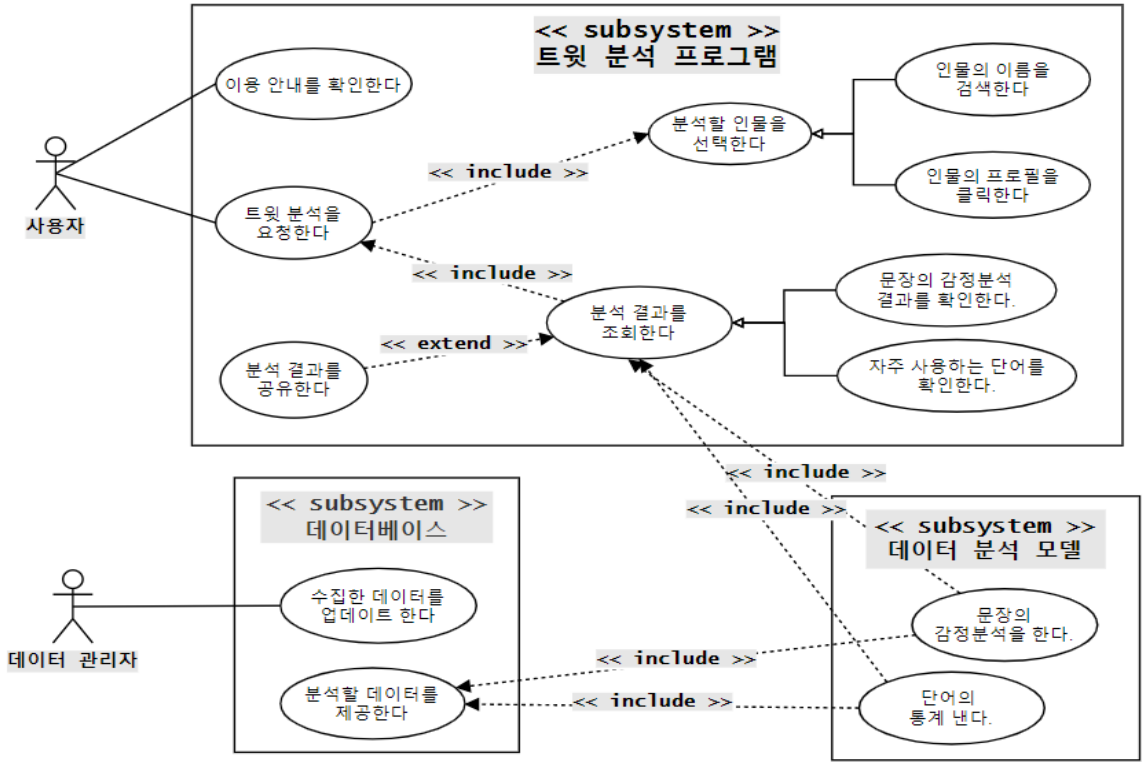
요구사항명	수집한 데이터 업데이트	ID	R-F-09	우선순위	상
요구상황설명	트위터 API 라이선스 문제로 데이터 관리자가 수집한 데이터를 업데이트 한다.				
해결방안	트위터 API 라이선스 문제로 트윗 데이터를 실시간으로 가져올 수 없다. 데이터 관리자가 매주 화요일마다 50명의 트윗 데이터를 수집해 txt파일로 저장한다.				
위험요소	많은 시간이 소요된다.				
설계 시 고려사항	문장을 단위로 데이터를 수집한다.				
관련요구사항					
시나리오	1. 데이터 관리자는 50명의 트윗을 트위터 API 무료 버전을 통해 수집한다. 2. 글자 수 제한으로 잘린 문장을 확인 후 실제 문장으로 수정한다.				

요구사항명	분석할 데이터 제공	ID	R-F-10	우선순위	중
요구상황설명	트윗 분석을 진행할 데이터를 데이터 분석 모델에 제공한다.				
해결방안	저장된 트윗 데이터 파일(txt)을 모델에 제공한다.				
위험요소					
설계 시 고려사항	유명인의 트위터 아이디와 트윗 데이터(txt파일)를 저장한다.				
관련요구사항	데이터 분석 모델 문장의 감정분석, 자주 사용하는 단어 통계				
시나리오	1. 데이터 분석 모델에서 분석을 위해 데이터를 요청한다. 2. 데이터베이스에 미리 저장해둔 데이터를 제공한다.				

2.2.5 제약조건

소프트웨어 환경		하드웨어 환경		네트워크 환경	
Language	Javascript, python, C++	CPU	Intel core i5	Cloud computing service	AWS
DBMS	Mongodb	RAM	8GB	VM	EC2
WAS	Node.js	O/S	Windows10	AMI	Ubuntu Server
Editor	Jupyter notebook, VS Code			Server region	Seoul
Tech. Used	ejs				

2.3 유스케이스 다이어그램



2.3.1 유스케이스 기술서

유스케이스명	이용 안내 사항을 확인한다
액터명	사용자, 트윗 분석 프로그램
개요	사용자는 프로그램에 대한 이용 안내를 확인한다.
사전 조건	사용자가 웹페이지에 접속해야 한다.
정상 흐름	1. 사용자가 웹페이지에 접속한다. 2. 사용자가 웹페이지 상단의 ABOUT 버튼을 누른다. 3. 사용자는 트윗 분석 프로그램의 이용 안내 사항을 확인한다.
사후 조건	사용자는 프로그램에 대한 안내 사항을 확인했다.

유스케이스명	인물의 프로필을 클릭한다
액터명	사용자, 트윗 분석 프로그램
개요	사용자가 분석하고 싶은 인물을 선택한다.
사전 조건	사용자가 웹페이지에 접속해야 한다. 사용자가 하단의 분석 가능한 인물의 프로필을 확인한다.
정상 흐름	1. 사용자가 하단의 분석 가능한 인물의 프로필을 확인한다. 2. 사용자가 분석을 원하는 인물의 프로필을 클릭한다.
사후 조건	사용자가 선택한 인물의 분석을 진행할지 확인하는 안내창이 나타난다.

유스케이스명	인물의 이름을 검색한다
액터명	사용자, 트윗 분석 프로그램
개요	사용자가 분석하고 싶은 인물을 선택한다.
사전 조건	사용자가 웹페이지에 접속해야 한다
정상 흐름	1. 사용자가 하단의 감정분석 가능한 인물의 프로필을 확인한다. 2. 사용자가 감정분석 하려는 인물의 이름을 화면 중앙 검색창에 입력한다.
사후 조건	사용자가 선택한 인물의 분석을 진행할지 확인하는 안내창이 나타난다.

유스케이스명	트윗 분석을 요청한다
액터명	사용자, 트윗 분석 프로그램, 데이터 분석 모델
개요	사용자가 인물의 트윗 분석을 요청한다.
사전 조건	사용자가 "분석할 인물을 선택한다".
정상 흐름	1. 사용자가 "분석할 인물을 선택한다". 2. 사용자가 "트윗 분석을 요청한다". 3. 트윗 분석 프로그램은 분석할 인물의 트위터 아이디(@userid : str형)를 데이터 분석 모델에 전송한다.
사후 조건	데이터 분석 모델이 "문장의 감정분석을 한다", "단어의 통계를 낸다"가 실행된다.

유스케이스명	문장의 감정분석을 한다
액터명	트윗 감정분석 프로그램, 데이터 분석 모델, 데이터베이스
개요	감정분석 모델이 트윗 문장의 감정분석을 한다.
사전 조건	사용자가 "트윗 감정분석을 요청한다".
정상 흐름	1. 사용자가 "트윗 감정분석을 요청한다". 2. 데이터베이스가 "분석할 데이터를 제공한다". 데이터는 txt 파일 형태로 제공된다. 3. 데이터 분석 모델은 제공받은 데이터로 "트윗 문장의 감정분석을 한다".
사후 조건	감정분석 결과 3가지의 타입 긍정 / 중립 / 부정 으로 분류된다. 3가지 감정의 비율을 확인 가능하다. 각 타입의 실제 트윗 문장을 최대 2개씩 확인 가능하다.

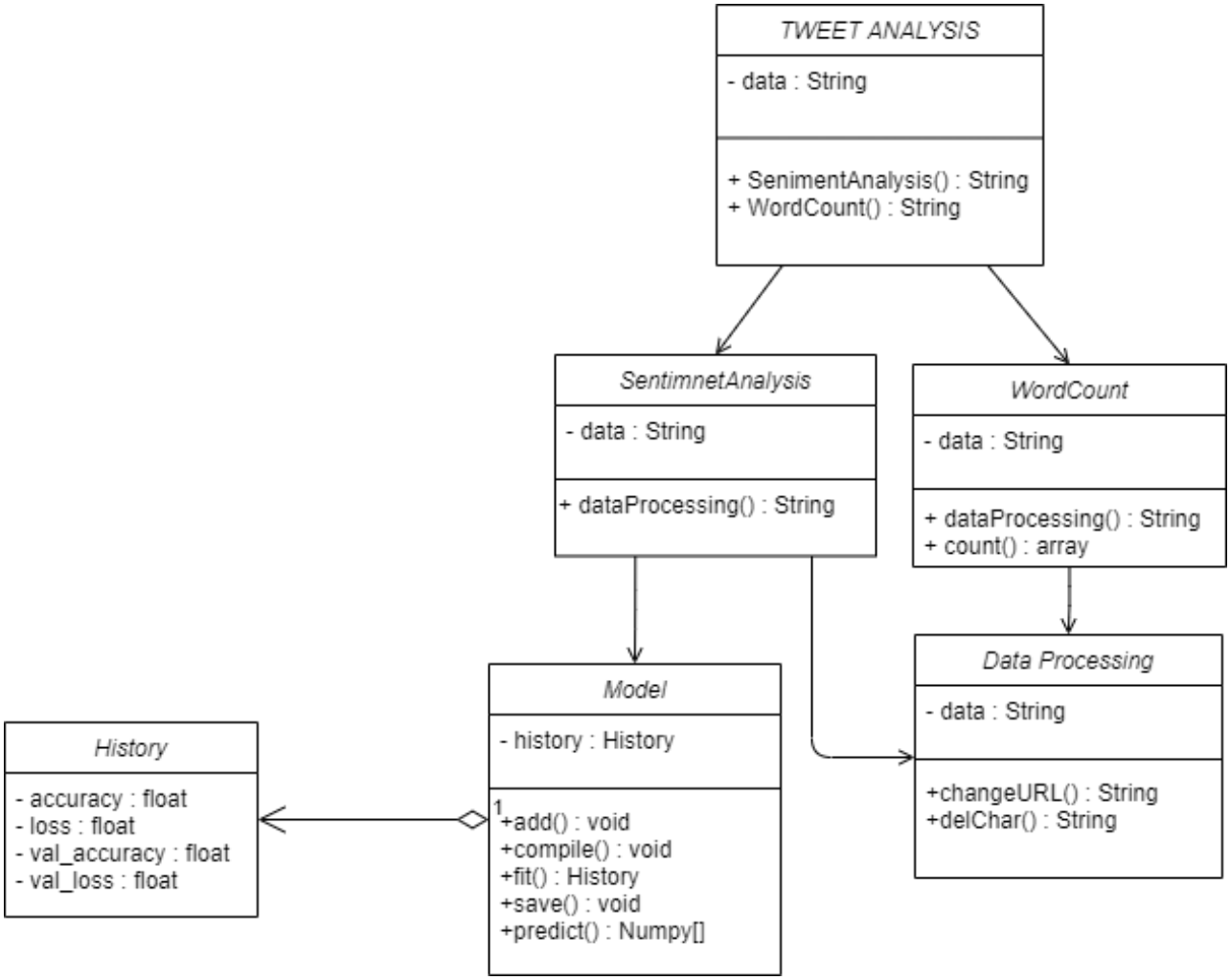
유스케이스명	단어의 통계를 낸다
액터명	트윗 감정분석 프로그램, 데이터 분석 모델, 데이터베이스
개요	감정분석 모델이 트윗 속 단어의 통계를 낸다.
사전 조건	사용자가 "트윗 감정분석을 요청한다".
정상 흐름	1. 사용자가 "트윗 감정분석을 요청한다". 2. 데이터베이스가 "분석할 데이터를 제공한다". 데이터는 txt 파일 형태로 제공된다. 3. 데이터 분석 모델은 제공받은 데이터로 "단어의 통계를 낸다".
사후 조건	이모지를 포함해 자주 사용하는 단어를 최대 5개까지 확인 가능하다.

유스케이스명	분석 결과를 조회한다
액터명	사용자, 트윗 감정분석 프로그램, 데이터 분석 모델,
개요	사용자가 요청한 유명인의 트윗 분석 결과를 조회한다.
사전 조건	데이터 분석 모델은 "트윗 문장의 감정분석을 한다". 데이터 분석 모델은 "단어의 통계를 낸다".
정상 흐름	1. 사용자가 "트윗 분석을 요청한다". 2. 데이터 분석 모델은 "트윗 문장의 감정분석을 한다", "단어의 통계를 낸다". 분석 결과는 json 파일 형태로 제공된다. 3. 사용자는 "분석 결과를 조회한다". 분석 결과는 긍정/중립/부정 문장의 비율과 예시, 자주 사용하는 단어를 확인 가능하다.
사후 조건	사용자는 트윗 분석 결과를 확인했다.

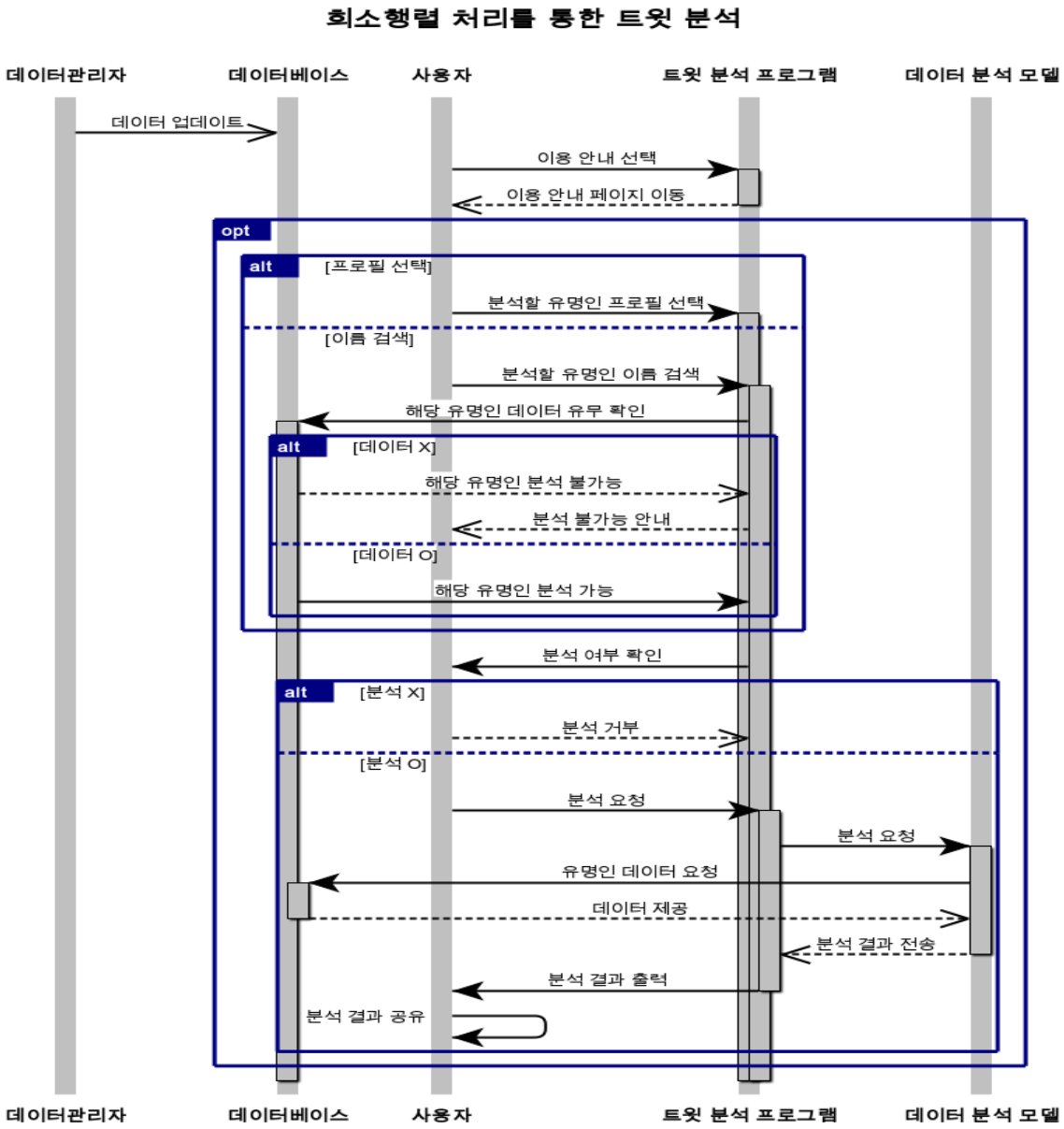
유스케이스명	분석할 데이터를 업데이트 한다
액터명	데이터 관리자, 데이터베이스
개요	데이터 관리자는 직접 수집한 데이터를 업데이트 한다.
사전 조건	데이터베이스에 50명의 인물의 트윗 데이터가 텍스트 파일로 저장되어 있다.
정상 흐름	1. 트위터 API 무료 버전을 사용해 데이터 관리자가 매주 화요일마다 50명의 데이터를 텍스트 파일 형식으로 수집한다. 2. 무료 버전의 글자 수 제한으로 잘린 문장의 데이터를 트위터와 비교해서 수정한다. 3. 데이터 관리자는 "분석할 데이터를 업데이트 한다".
사후 조건	데이터베이스에 매주 월요일까지의 트윗 데이터가 업데이트 되었다.

3. 시스템 분석

3.1 클래스 다이어그램



3.2 시퀀스 다이어그램


회소행렬 처리를 통한 트윗 분석
데이터관리자
데이터베이스
사용자
트윗 분석 프로그램
데이터 분석 모델
데이터 업데이트
이용 안내 선택
이용 안내 페이지 이동
opt
alt
[프로필 선택]
분석할 유명인 프로필 선택
[이름 검색]
분석할 유명인 이름 검색
해당 유명인 데이터 유무 확인
alt
[데이터 X]
해당 유명인 분석 불가능
분석 불가능 안내
[데이터 O]
해당 유명인 분석 가능
분석 여부 확인
alt
[분석 X]
분석 거부
[분석 O]
분석 요청
분석 요청
유명인 데이터 요청
데이터 제공
분석 결과 전송
분석 결과 출력
분석 결과 공유
데이터관리자
데이터베이스
사용자
트윗 분석 프로그램
데이터 분석 모델