|  |
| --- |
|  |
| 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** |

[1. 개요 3](#_Toc55169200)

[1.1 개발 동기 3](#_Toc55169201)

[1.2 목표 3](#_Toc55169202)

[2. 시스템 구조 4](#_Toc55169203)

[2.1 전체 시스템 구조 4](#_Toc55169204)

[2.2 subsystem 설명 4](#_Toc55169205)

[2.2.1 웹 브라우저 4](#_Toc55169206)

[2.2.2 웹 서버 (모델 API) 4](#_Toc55169207)

[2.2.3 데이터베이스 4](#_Toc55169208)

[3. 팀원 역할 분담 5](#_Toc55169209)

[4. 제안서 수정 내용 6](#_Toc55169210)

[4.1 요구사항 수정 내역 6](#_Toc55169211)

[5. 개발 추진 계획 7](#_Toc55169212)

[5.1 데이터베이스 관련 유스케이스 7](#_Toc55169213)

[5.2 분석 모델 관련 유스케이스 8](#_Toc55169214)

[5.3 웹 관련 유스케이스 8](#_Toc55169215)

[6. 현재 구현 내용 9](#_Toc55169216)

[6.1 웹 9](#_Toc55169217)

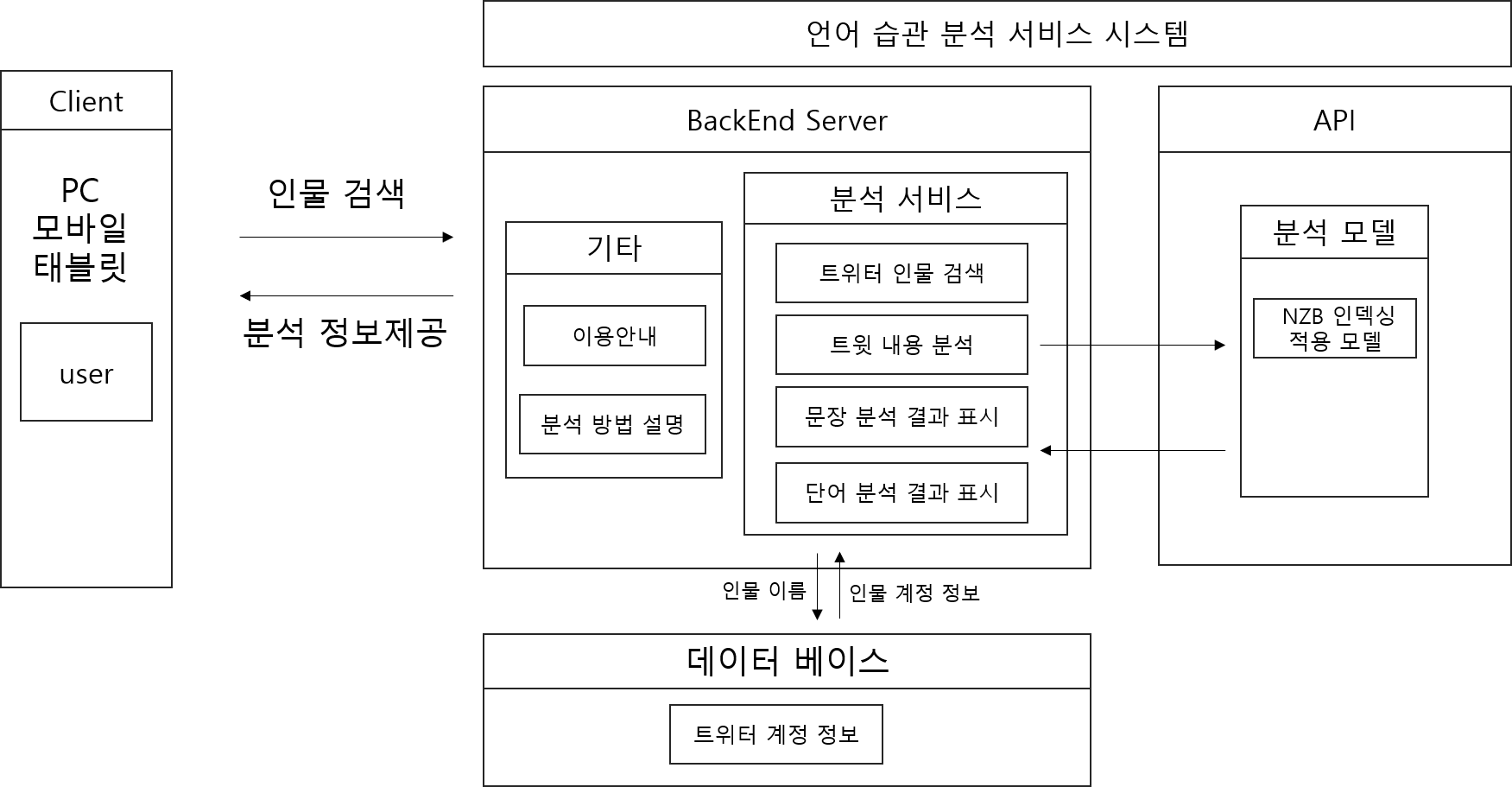
[6.2 분석 모델 10](#_Toc55169218)

[6.2.1 기존 분석 모델 10](#_Toc55169219)

1. 개요
   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. 목표
* 기존 모델의 convolution 연산 방식을 NZB 인덱싱 기법을 적용해 메모리 접근효율을 증가시킨다.
* 트위터 API를 통해 유명인의 트윗 데이터를 수집하고, 게시글 속 문장의 감정을 분석하여 긍정, 중립, 부정 3가지의 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. 시스템 구조
   1. 전체 시스템 구조
   2. subsystem 설명
      1. 웹 브라우저

* 분석 가능한 인물 프로필 안내 화면
* ㅇ
  + 1. 웹 서버 (모델 API)
* 긍정/중립/부정 3가지 타입의 결과로 문장의 감정분석을 진행 후, 타입 별 비율을 계산 후, 해당하는 3가지 타입의 문장을 최대 2개까지의 결과를 제공한다.
* 자주 사용하는 단어를 최대 5개까지 제공한다.
  + 1. 데이터베이스
* 50명의 데이터는 “이름/트위터 계정/이미지/직업 정보/트윗 데이터”를 포함해 저장되어 있다.

1. 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 감정분석 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 | 이아현, 노수민(SUB), 홍주희(SUB) |
| 데이터베이스 구축 | 양승주 |
| 모델 평가 및 검증 | 노수민, 이아현, 홍주희 |
| 분석 결과 시각화 | 양승주 |

1. 제안서 수정 내용
   1. 요구사항 수정 내역

|  |  |
| --- | --- |
| 이전 제안서 요구사항 | 수정된 제안서 요구사항 |
| 긍정, 부정 단어의 비율 차이를 계산해 5가지 유형으로 결과를 제공한다. | 긍정, 중립, 부정 문장의 비율을 제공 후, 해당 감정의 문장을 예시로 최대 2개 제공한다. |
| 긍정적, 부정적 단어의 비율을 그래프로 시각화 한다. | 긍정, 중립, 부정 문장의 비율을 막대 그래프로 시각화 한다. |
| word2vec 모델을 사용한다. | text-CNN 모델을 사용한다. |
|  |  |

1. 개발 추진 계획
   1. 데이터베이스 관련 유스케이스

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **일정** | **9월** | | | **10월** | | | | **11월** | | | |
| **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 |  |  |  |  |  |  |  |  |  |  |  |

* 1. 분석 모델 관련 유스케이스

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 일정 | 10월 | | | | 11월 | | |
| 1 | 2 | 3 | 4 | 1 | 2 | 3 |
| text-CNN 기존 모델 | | | | | | | |
| text-CNN 모델 구현 |  |  |  |  |  |  |  |
| text-CNN 모델 학습 |  |  |  |  |  |  |  |
| text-CNN 모델 평가 |  |  |  |  |  |  |  |
| NZB 인덱싱 기술 적용 모델 | | | | | | | |
| NZB 인덱싱 기술 적용 |  |  |  |  |  |  |  |
| NZB 적용 모델 평가 |  |  |  |  |  |  |  |
| 기존 모델, 적용 모델 성능 비교 |  |  |  |  |  |  |  |

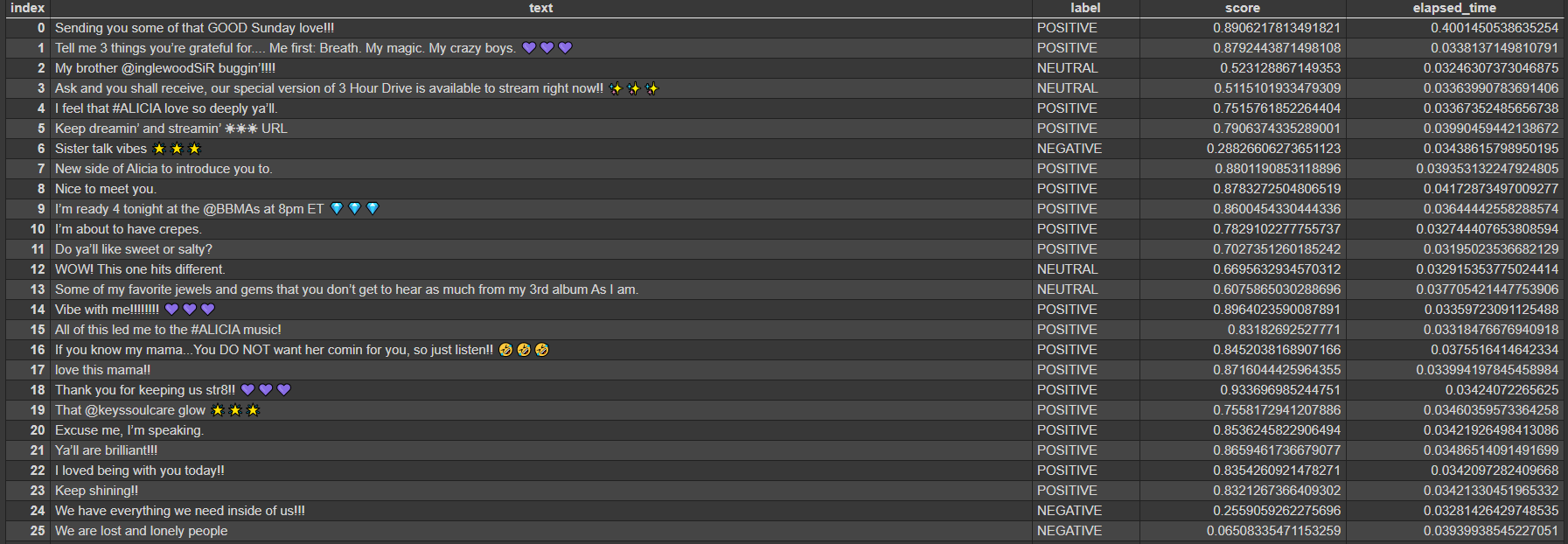
* 1. 웹 관련 유스케이스

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **일정** | **9월** | **10월** | | | | **11월** | | | |
| **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 |  |  |  |  |  |  |  |  |  |

1. 현재 구현 내용
   1. 웹
   2. 분석 모델
      1. 기존 분석 모델

(@alicia\_keys)

* 텍스트이(가) 표시된 사진

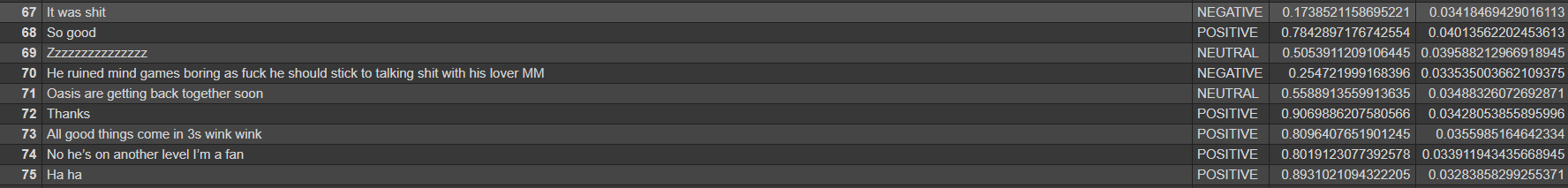
  자동 생성된 설명문장의 감정 비율, 자주 사용하는 단어 결과
* 문장의 감정 분석 결과

(@liamgallagher)

* 문장의 감정 비율, 자주 사용하는 단어 결과

텍스트이(가) 표시된 사진

자동 생성된 설명

* 문장의 감정 분석 결과