캡스톤 주제 제안서

2020-09-24

강좌 : 캡스톤 디자인 (001)

지도 교수 : 양효식, 안용학

16010573 이아현 컴퓨터공학

16010945 노수민 컴퓨터공학

16011041 홍주희 컴퓨터공학

16011189 양승주 디지털콘텐츠

2 조 팀원 소개

			
16010573	16010945	16011041	16011189
이아현	노수민 (팀장)	홍주희	양승주
컴퓨터공학	컴퓨터공학	컴퓨터공학	디지털콘텐츠
ahlee273@gmail.com	chzoss@sju.ac.kr	hongjh2004@gmail.com	uknows j@nate.com

I. 개발 배경 및 중요성

1. 개발 배경

기존에 존재하는 모델을 이용해 실시간 처리가 가능하며, 두개의 모델을 적용했을 때 성능 비교가 가능한 프로그램을 조사를 했다. 두개의 모델 중 하나는 새로운 모델은 기존에 존재하는 모델을 분석 및 수정을 할 것이다. 실시간 반영이 가능한 트위터를 이용해 사용자의 언어습관을 분석하기로 했다. 해당 사용자가 어떤 단어를 자주 사용하며, 긍정적 언어, 부정적 언어의 사용 비율을 보여주는 프로그램을 개발할 예정이다.

2. 중요성

시중에 많이 알려진 AI 관련 알고리즘들은 대개 높은 정확도를 위해 복잡하게 만들어져 있다. 따라서 모델을 훈련시키거나 사용하는 데에 시간이 많이 소모되고, 과다한 연산량으로 인한 메모리 부족의 문제들과 직면하기도 한다. 반면에 모델의 크기를 줄인다면 연산량은 줄어들더라도 성능이 떨어질 것이다. 이러한 연산량과 정확도의 반비례적인 문제점을 해결할 수 있는 기법으로 제시된 NZB 인덱싱을 사용할 것이다. 기존의 밀집 행렬 연산이 이루어지는 부분을 희소 행렬로 수행하도록 바꾸고, 이를 통해 모델의 용량과 연산량을 감소시킬 것이다. 희소행렬을 처리하는 방식인 NZB 인덱싱 기법으로 모델의 연산 과정을 분석 및 수정을 통해 구현한다면 어떤 부분에서 좋은 성능의 결과를 내는지 증명할 수 있다고 예상된다.

II. 개발 목표

1. 데이터를 기존 모델로 연산했을 때와 모델의 밀집 행렬 연산 처리 과정에서 NZB 인덱싱 기술을 적용했을 때, 기존 모델과 성능적으로 어떤 차이가 존재하는지 정량적으로 결과를 보인다.

2. 특정 사용자의 이름을 입력 받아 해당 인물의 트위터 게시글을 실시간으로 분석해 저희 프로그램에서 나누는 5 가지 유형으로 결과를 도출한다.

3. 긍정적 단어와 부정적 단어의 그래프를 통해 시각화 후, 자주 사용하는 단어를 태그 형식으로 나타낸다.

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

III. 차별성

1. 시장성

   A. AI 분야 시장 조사

      ① 인공지능 분야 특허 출원이 2010 년부터 최근 10 년간 16 배 급증하며 4 차 산업혁명을 주도한 것을 알 수 있다.

      ② 세계 인공 지능 소프트웨어 시장은 2019 년 164 억 달러에서 6 배 증가해 2025 년에는 988 억 달러 규모로 확대될 전망이다.

      ③ 정부에서 "데이터·AI 경제 활성화 계획"을 발표를 했고, AI 를 사용하는 분야가 점점 확대되고 있다.

   B. 기존에 존재하는 프로그램

      ① Watson Tone Analyzer

         - 트위터, 온라인 리뷰, 이메일, 텍스트를 7 가지의 톤으로 분석
         - 영어, 프랑스어 2 가지 언어를 제공

      ② DeepMoji

         - 영어 문장의 감정을 분석해 여러 이모지로 표현

      ③ 네이버 요약봇

         - 텍스트에서 문장의 중요도를 분석해 중요한 문장을 추출하는 요약 기술

      ④ CLOVA Speech

         - 한국어 음성 파일을 NEST 엔진을 사용해 텍스트로 변환

2. 기능 구성

   A. 밀집 행렬에 NZB 인덱싱 기법 사용

   기존에 존재하는 word2vec 모델을 사용해 텍스트 분석을 진행할 예정이다. Pytorch 코드에서 Model 을 선언하는 부분, 직접 연산을 하는 부분을 분석 후 기존의 밀집 행렬 연산이 이루어지는 부분을 희소 행렬로 수행하도록 바꿀 예정이다. 기존의 모델과 NZB 인덱싱 기법을 적용한 모델에서 어떤 성능적 차이가 존재하는지 보여줄 것이다. 기존 모델과 기법을 적용한 모델의 처리 속도를 비교할 수 있다.

B. 언어습관을 파악 후 5 가지 유형으로 분류

실시간으로 분석하고자 하는 검색 대상의 이름을 검색 후, 사용자가 특정 계정을 선택한다. 해당 계정의 공개된 트위터 게시글에서 텍스트를 가져온다. 가져온 텍스트들을 분석을 통해 2 가지의 결과 긍정, 부정이 아니라 단어 사용의 비율을 측정해 5 가지의 유형 'A master of positive speech / A sprout of positive speech / A master of various speech / A sprout of hate speech / A master of hate speech ' 으로 분류할 예정이다. 긍정적 단어 사용 비율과 부정적 단어 사용 비율의 편차를 이용해 위의 유형을 판단할 예정이다. 또한, 긍정적 단어와 부정적 단어의 사용 비율을 그래프를 통해 시각화하고, 자주 사용하는 단어들을 최대 4 개까지 태그형식으로 화면에 표현함으로써 해당 인물의 언어습관에 대해 간단하게 파악이 가능해 더 실용성 있을 것이다.

3. 활용성

A. NZB 인덱싱 기법의 성능 향상

① 두 모델의 성능 비교를 통해 성능 비교를 통해 더 큰 데이터를 다루는 프로그램에 적용할 경우 얼마나 효과적일지 예상할 수 있다.

B. 정보 제공

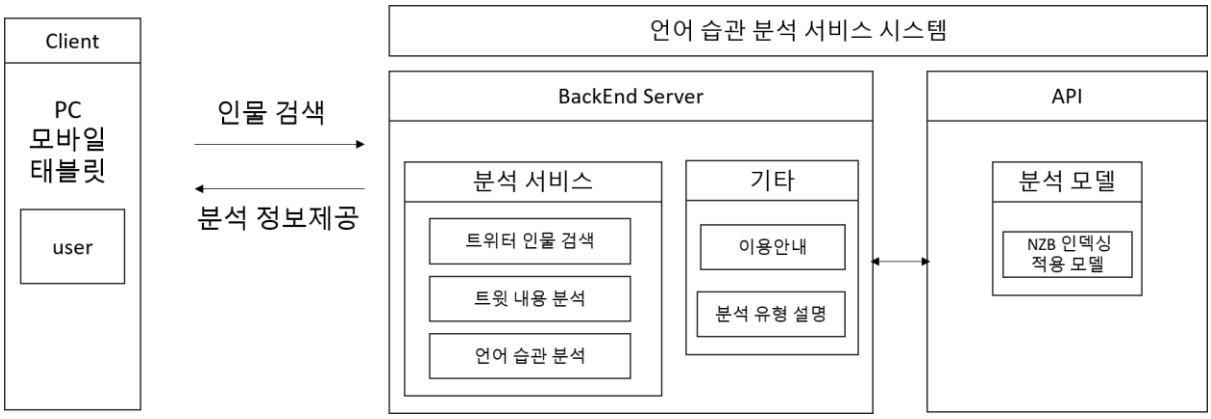
① 미국 대선에 관심있는 사람은 Donald Trump, Joe Biden 을 검색해 후보들의 평상시 언어습관을 확인할 수 있을 것이다.

C. 엔터테인먼트

① 좋아하는 가수, 배우 등에 대해 검색을 해 팬들끼리 공유를 할 수 있다.

IV. 개발 방법 및 체계

1. 시스템 구성도

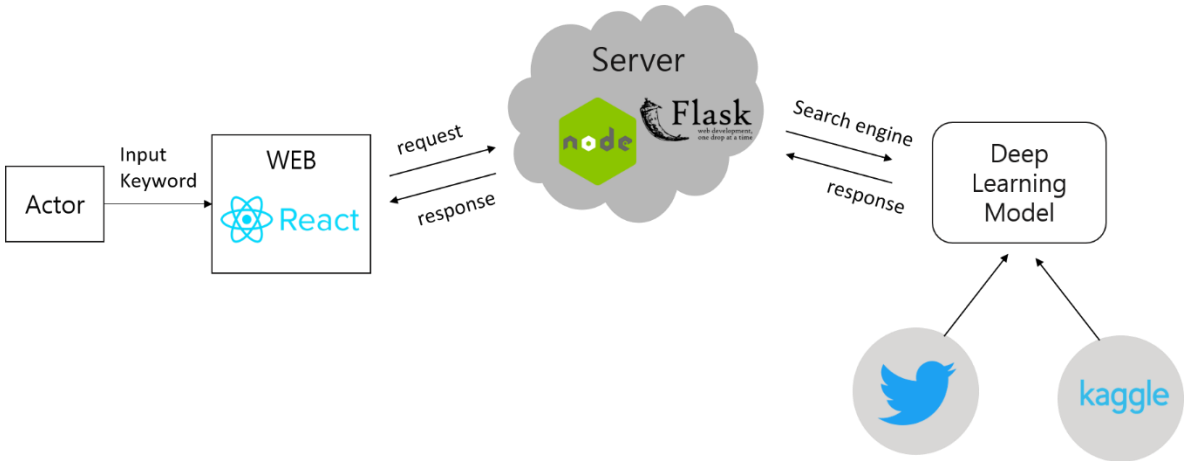


A. 사용자가 웹에서 특정 인물의 이름을 검색해, 계정을 선택

B. 해당 계정의 트윗 내용들을 추출 후 NZB 인덱싱 적용 모델로 분석

C. 분석 후, 웹에서 언어습관 유형, 유형 설명, 긍정/부정 단어 비율, 자주 사용하는 단어, 트위터 계정 출처를 확인 가능

2. 개발 환경



A. 언어
    ① Python
    ② C
    ③ JavaScript

B. 프레임워크
    ① Flask

C. 협업 툴
    ① GitHub

D. IDE
    ① Pycharm
    ② Visual Studio Code
    ③ Jupyter Notebook

E. 플랫폼
    ① Anaconda
    ② Kaggle
    ③ Node.js

F. 라이브러리
    ① React.js
    ② PyTorch
    ③ Scikit learn
    ④ TensorFlow

G. 모델
    ① Word2vec

3. 팀원 역할 분담

PM 및 문서 작성	노수민
회의록 작성	홍주희
PPT 제작 및 대본 작성	양승주
데이터 수집	노수민, 홍주희
데이터 전처리	노수민, 홍주희
텍스트 분석 모델	노수민, 홍주희, 이아현(SUB)
NZB 인덱싱 기법 적용	노수민, 이아현, 홍주희
웹 설계 및 구축	양승주
UI 설계 및 구현	양승주
모델 학습	이아현, 홍주희
모델 검증 및 통계 제출	노수민, 이아현
시각화	양승주

고정 역할은 위의 표와 같이 분담하였으며, 프로젝트를 진행하면서 개발 일정과 다르게 먼저 진행을 끝낸 팀원은 다른 팀원을 도와서 진행한다.

V. 개발 추진 계획

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 및 전처리												
텍스트 분석												
기존 모델 검증												
모델 NZB 기술 적용												
웹 설계												
모델 학습												
웹 UI 구현 및 구축												
모델 검증 및 통계 제출												
시각화												
최종 문서 작성												
최종 발표												

1. 모델 학습

   A. 모델의 정확도를 위해 twitter sentiment analysis dataset (Kaggle), SentiWordNet, SentimentIntensityAnalyzer 등 이미 존재하는 데이터셋을 활용할 예정이다.

VI. 기대 효과

1. 정성적 기대효과

   A. CNN, RNN, LSTM 같은 모델의 연산과정에는 희소 행렬, 밀집 행렬을 포함한 행렬을 통해 연산 과정이 이뤄진다. 희소행렬을 처리하기 위해 CSR 방식이 자주 사용한다. CSR 방식은 행렬에서 값이 존재하는 부분을 추출할 때, 해당 데이터의 열 정보, 행 정보, 데이터 값 순서대로 접근이 필요해 밀집 행렬에서 성능이 떨어진다. 감정을 분류하는 학습에는 행렬 연산이 많다. 이 연산 과정에서 3X3 행렬에 크기를 표현하는 4 비트와 인덱스를 나타내는 9 비트로 총 13 비트를 통해 최대 9 개의 원소에 접근가능한 NZB 인덱싱 방식을 적용하면 전체 인덱스 사이즈와 메모리 접근 횟수 측면에서 높은 효율을 가질 것이다.

   B. 미리 분석한 결과를 DB에 저장하고 불러오는 방식이 아닌 사용자가 검색한 특정 인물의 트위터 게시글을 실시간으로 가져와 분석을 진행을 한다. 기존에 존재하는 데이터가 아닌, 새로운 데이터가 반영이 되어서 특정 인물이 언어습관이 계속해서 반영될 것이다.

2. 정량적 기대효과

   A. 모델의 정확성을 위해 KAGGLE에 존재하는 twitter sentiment dataset을 모델 학습에 사용할 예정이다. 기존 모델과 NZB 기법을 적용한 모델의 수행 속도 시간을 비교했을 때, NZB 기법을 적용한 모델이 연산 처리 부분에서 능력이 빨라 수행 속도가 빠를 것이라고 예상된다. 수행 속도가 향상되었다는 걸 증명하면, 방대한 데이터를 처리할 때 사용하면 더 효율적일 것이라고 기대한다.