캡스톤 주제 제안서

2020-09-17

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

1. 개발 배경 및 중요성

문서 요약 프로그램 중 생소한 주제인 유명인들의 SNS 와 연설을 통해 언어습관을 분석하기를 선택하게 된 계기는 유명인이 주로 사용하는 언어습관에서 그 사람의 특징을 파악하기 위해서다. 사람의 성격을 확인할 수 있는 좋은 방법 중 하나로, 그가 평소에 어떤 단어를 쓰며 어떤 말투로 이야기를 하는지 분석하는 것이다. 연예인 혹은 정치인이 방송이나 연설, 또는 SNS 공간에서 사람들이 불쾌하게 느낄 수 있는 혐오성 발언을 했는지, 긍정적인 이미지를 보여줄 수 있는 적절한 언어 선택을 하는지를 파악할 수 있다는 점은 큰 장점이 될 것이다. 특히, 정치인의 평소 언어습관을 파악할 수 있다면 이 사람의 순간적인 모습으로 판단을 하지 않을 수 있을 것이다.

시중에 많이 알려진 딥러닝 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 기법이 NZB 인덱싱이다. 희소행렬을 처리하는 방식인 NZB 인덱싱 기법으로 모델을 만든다면 좋은 성능의 결과를 보여줄 수 있을 것이다.

2. 개발 목표

1. 특정 인물의 발화 패턴을 분석하는 딥러닝 모델의 연산에 NZB 인덱싱 기법을 적용해서 데이터 처리 속도를 향상시킨다. 이렇게 개발된 딥러닝 모델은 웹서비스를 통해 불특정 다수의 사용자가 사용할 수 있도록 한다.
2. 모델을 학습시키는 데이터는 유저가 검색한 유명인의 유튜브 영상과 트위터에서 추출한 텍스트로 한다.
3. 유저가 검색한 결과는 시각화 하여 표현해서 한 눈에 해당 인물의 특징을 파악할 수 있도록 한다.

3. 차별성

1. 행렬 연산에서 NZB 인덱싱 기법 사용

최근 논문으로 발표된 연산 성능 향상 기법인 NZB 인덱싱 기법을 기존에 존재하는 딥러닝 모델에 연산 과정에 적용해 효율적으로 수행할 수 있다. 이전 시장에 나와있는 텍스트 분석 프로그램들보다 더 적은 메모리 사용이 가능해짐으로써 개발 비용이 절감된다. 동시에 처리 속도도 훨씬 빨라진다는 점에서 다른 프로그램보다 사용성이 좋아 경쟁성을 가진다.

2. 텍스트 감정 분석이 아닌 인물의 언어습관을 파악 가능

기존의 텍스트 분석 프로그램들은 크게 2 가지로 분류할 수 있다. 영화, 드라마 속의 대사 또는 트위터 같은 SNS 의 문장들을 통해 감정을 분석을 진행한 후 감정 표현의 비율이 어떻게 되는지 분석하는 프로그램.[1] 뉴스와 논문 같은 지식 전달 목적의 글을 요약해주는 프로그램.[2] 이런 프로그램들과 다르게 본 프로젝트에서 개발 예정인 프로그램은 실제 인물의 영상과 SNS 에서 나타나는 언어습관, 자주 사용하는 추임새 같은 특징을 추출한다. 이 특징들을 가지고 단순한 2 가지의 결과 긍정, 부정이 아니라 단어 사용의 비율을 측정해 크게 5 가지의 유형으로 분류할 예정이다. 이러한 통계적 수치를 사용해 시각적으로 표현함으로써 해당 인물의 언어습관에 대해 간단하게 파악이 가능해 더 실용성 있을 것이다.

[1] 예) DeepMoji : 사용자가 문장을 입력하면 문장 속 비율이 높은 감정 표현을 이모티콘으로 보여준다.
[2] 예) 네이버 요약봇 : 뉴스 기사 내용을 간략하게 요약해서 보여준다.

4. 개발 방법 및 체계

1. 개발 방법

A. Github

- 팀 회의, 교수님 미팅 회의록 및 공지사항을 공유한다.
- 과제 제출물을 공유한다.
- 개발 진행 시 프로젝트 code를 공유한다.

B. Pytorch로 진행

- python-tiwtter github open source를 통해 트위터 자료를 수집 예정이다.
- youtube 영상 인물의 음성을 텍스트로 추출할 예정이다.
    ▷ 사용할 open source를 조사 중이다.
- word2vec 모델을 활용해 텍스트 속 단어를 분석할 예정이다.
- 연산 처리 과정에 NZB 인덱싱 기법을 적용할 예정이다.

C. 웹 진행

- 프론트 엔드는 react.js를 사용할 예정이다.
- 백 엔드는 node.js를 사용할 예정이다.
- 모델 API는 flask를 사용할 예정이다.
- 데이터베이스는 mongoDB를 사용할 예정이다.

2. 팀원 역할

- 노수민 (팀장)
  - PM 및 문서 작성
  - 데이터 전처리 (SUB)
  - 텍스트 분석 모델 설정
  - NZB 인덱싱 기법 적용
  - 통계 제출 및 시각화
- 양승주
  - PPT 제작 및 대본 구성
  - 웹 설계 및 구축
  - 데이터베이스 구축
  - UI 설계 및 구현
  - 시각화
- 이아현
  - 영상 텍스트 추출
  - NZB 인덱싱 기법 적용
  - 모델 테스트
  - 통계 제출
- 홍주희
  - 회의록 작성
  - 데이터 수집 및 전처리
  - 영상 텍스트 추출 (SUB)
  - NZB 인덱싱 기법 적용

5. 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												
웹 디자인 및 설계												
영상에서 텍스트 추출												
텍스트 분석												
데이터 전처리												
모델 NZB 기술 적용												
웹 UI 구현 및 구축												
검증 및 테스트												
단어 사용 통계												
시각화												
최종 문서 작성												
최종 발표												

6. 기대 효과

1. 성능적 측면

기존에 존재하는 모델들은 희소행렬을 CSR, DOK 방식으로 처리했다. 이 방식들을 쓸 경우, 데이터의 행, 열, 값의 정보를 다 확인해야 접근이 가능해 그만큼 연산량이 많다. 이런 과정에서 시간을 줄일 수 있는 NZB 인덱싱 기법을 적용한다면 메모리 접근 횟수도 3 배가 줄어들 것이다.

2. 사회적 측면

SNS가 일상인 요즘 시기에 단순한 문장인 텍스트로 사람을 판단하는 경우가 늘어나고 있다. 이런 상황에서 특정 인물에 대한 자료를 통해 이 사람의 평소 언어습관을 보여주면 오해의 소지가 줄어들 수 있을 것이다.