사용자 기반 Text Cloud Visualization

2024년 전기 졸업과제

Department of Computer Science and Engineering, Pusan National University

목차

과제 목표

현대 사회는 빅데이터 시대에 접어들면서 방대한 양의 텍스트 데이터가 생성되고 있다. 이러한 데이터는 소셜 미디어 포스트, 온라인 리뷰, 뉴스 기사, 이메일, 연구 논문 등 다양한 형태로 존재하며, 매일 엄청난 속도로 증가하고 있다. 이러한 방대한 텍스트 데이터를 효과적으로 처리하고 의미 있는 정보를 추출하는 것은 매우 중요한 연구 주제이다.

텍스트 데이터를 처리하고 시각화 하는 효과적인 방법 중 하나로 사용자 기반 텍스트 클라우드 시각화가 있다. 텍스트 클라우드는 단어의 빈도를 시각적으로 표현하여 텍스트 데이터의 구조와 주요 패턴을 쉽게 이해할 수 있게 해준다. 단어의 빈도에 따라 글자 크기를 다르게 표시하여, 중요한 키워드나 주제가 한눈에 들어오도록 한다. 이는 복잡한 데이터셋을 간단하고 직관적으로 요약할 수 있는 효과적인 도구이다.

특히, 사용자 기반 텍스트 클라우드 시각화는 사용자가 특정 관심사나 분석 목표에 따라 텍스트 클라우드를 사용자화(customize)할 수 있는 기능을 제공한다. 예를 들어, 특정 시간대별로 변화하는 트렌드를 시각화 하거나, 특정 주제나 키워드와 관련된 텍스트만을 필터링하여 클라우드를 생성할 수 있다. 이를 통해 사용자는 자신이 원하는 정보를 더 정확하고 빠르게 얻을 수 있다.

또한, 사용자 기반 텍스트 시각화는 교육, 연구 등 다양한 분야에서 활용될 수 있다. 예를 들어, 교육 분야에서는 학생들의 피드백을 시각화 하여 교육 방식의 개선점을 찾을 수 있으며, 연구 분야에서는 학술 논문의 주제를 한눈에 파악하여 연구 동향을 쉽게 파악할 수 있다. 이러한 시각화 도구는 복잡한 정보를 보다 쉽게 이해하고 활용할 수 있도록 돕는다.

대상 문제 및 요구조건

주요 문제점은 다음과 같다.

- 대량의 한국어 텍스트 데이터를 효율적으로 처리하고 분석하는 기술적으로 매우 복잡하다.
- 텍스트 데이터를 효과적으로 시각화 하는 방법이 부족하여, 중요한 정보나 패턴을 쉽게 인식하기 어렵다.
- 텍스트 클라우드 시각화 도구가 사용자들의 특정 목표나 선호도를 충분히 반영하지 못한다.
- 각 소스에서 수집된 데이터의 품질이 다를 수 있으며, 이를 정제하여 정확하고 일관된 데이터를 확보하는 것이 중요한 문제이다.

요구조건은 다음과 같다.

- 대량의 텍스트 데이터를 효율적으로 수집, 처리할 수 있는 시스템이 필요하다.
- 텍스트 데이터의 구조와 패턴을 직관적으로 파악할 수 있는 시각화를 설계하여야 한다.
- 중요한 키워드나 주제를 신속하게 인식할 수 있는 텍스트 클라우드 레이아웃을 개발하여야 한다.
- 자연어 처리 기술을 통합하여 텍스트 데이터의 의미를 더 정확하게 분석할 수 있어야 한다.
- 자연어 처리 기술을 활용하여 텍스트 클라우드의 품질을 개선하고, 사용자 맞춤형 시각화를 제공하여야 한다.

제약사항 분석 결과 및 대책

제약사항은 다음과 같다.

- 한국어 텍스트 데이터는 특유의 언어적 특성으로 인해 자연어 처리와 분석이 기술적으로 매우 복잡하다. 특히, 문맥을 정확히 이해하거나 복잡한 구문을 해석하는 데 어려움이 따른다. 한국어는 조사가 붙거나 어미가 변화함에 따라 단어의 의미가 달라질 수 있으며, 구문 구조도 영어와 크게 다르다. 또한, 띄어쓰기가 명확하지 않거나 잘못된 경우도 많아, 이를 정확하게 처리하는 것이 어렵다.
- 텍스트 클라우드 시각화 도구가 사용자에게 직관적이고 사용하기 쉽게 설계되지 않을 경우, 비전문가나 일반 사용자가 데이터를 이해하는 데 어려움을 겪을 수 있다.

제약사항에 대한 대책은 다음과 같다.

- 최신 자연어 처리 알고리즘과 학습 모델을 사용하여 한국어 텍스트의 문맥을 보다 정확하게 이해하고 분석할 수 있도록 한다. 특히, 한국어에 특화된 BERT 모델이나 GPT-3 등의 언어 모델을 활용하여 문맥 이해를 향상시킨다.
- 관련 논문을 통해 다양한 사용자 그룹의 요구와 선호도를 반영한 직관적인 사용자 인터페이스를 설계한다. 사용자 맞춤형 옵션을 제공하여 개별 사용자 요구에 맞춘 시각화를 제공한다.

설계

사용자 맞춤 텍스트 클라우드를 제공하기 위해서는 다음과 같은 Input 이 요구된다.

- 사용자의 정보
  - 지금의 기분
  - MBTI
  - 착용한 옷 색깔
  - 자기 소개 문장

- 뉴스 내용
  - 언론사에 따른 분류 데이터
  - 뉴스 카테고리에 따른 분류 데이터 (정치, 문화 등)

Input이 모두 갖춰졌으면, 사용자의 정보와 뉴스 내용을 각각 embedding한다. 사용자의 정보 vector을 뉴스 내용 vector에 넣어보거나, 두 vector의 유사도를 측정한다. 유사도에 따라 뉴스 카테고리를 추천해주며, 추천한 뉴스의 내용을 요약하는 모델을 만든다. 사용자 정보의 vector에 따라 어떠한 중요한 단어들이 추출되는지 관찰한다.

Output은 text cloud이며, 사용자에 맞춤형으로 제공되는 것을 목표로 한다. 이때 고려할 점은 사람이 심리적으로 인식하기 편한 형태로 text cloud의 형태가 제공되어야 한다는 점이다.

일정

일정은 다음과 같다.

- 5월
  - 데이터 수집
    - 뉴스 크롤링
    - 감정 언어 데이터 수집
  - 데이터 전처리

- 6월
  - 모델 개발
    - 모델에 embedding vector 넣어보며 결과 관찰
    - NLP 모델 별 테스트 및 성능 측정
    - 성능 스코어링

- 7월
  - 프론트엔드 디자인
    - 논문 참조해, 사람이 심리적으로 편안하다고 느끼는 디자인 구축

- 8월
  - 백엔드 구축
    - 사용자에게 빠른 text cloud visualization을 전달하기 위해 최대한 효율적인 visualization 코드 개발

역할 분담

역할 분담은 다음과 같다.

- 금비 -
  - NLP 모델 구현 및 성능 평가
  - 프론트엔드 디자인
  - 백엔드 디자인
  - 보고서 작성

- 원윤서 -
  - 프론트엔드 디자인
  - 백엔드 디자인
  - 보고서 작성

- 채문석 -
  - 데이터 수집
  - 데이터 전처리
  - 보고서 작성