소셜 미디어 뉴스 제목의 잠재 주제
- LDA 모델을 기반으로 한 분석 -

WU YUHANG(부경대학교 미디어커뮤니케이션학과)
LI WENQI(부경대학교 미디어커뮤니케이션학과)

요약

소셜 미디어 시대의 뉴스 헤드라인 중요성이 증가하고 있다. 본 연구는 이러한 헤드라인의 잠재적 주제를 Latent Dirichlet Allocation (LDA) 모델로 분석하였다. 37,288개 뉴스 헤드라인 데이터를 통해 주제 분포의 차이와 주요 주제들을 식별하였다. 이 결과는 소셜 미디어 뉴스의 주제 이해에 중요한 인사이트를 제공한다.

서론

소셜 미디어의 빅데이터는 인간 행동에 대한 중요한 통찰력의 핵심이라고 칭송되었으며, 학자들, 기업, 정치인들, 기자들 및 정부들에 의해 광범위하게 분석되었습니다 (Boyd and Crawford 2012; Lazer et al, 2009).

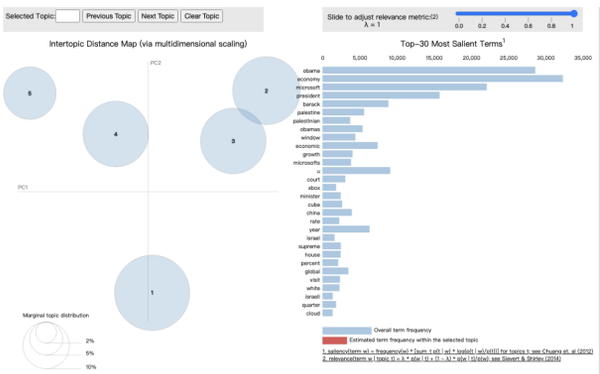
빅데이터는 다양한 질문에 대한 매력적인 통찰력을 제공하며, 이전에는 상상할 수 없었던 수준에서 사회 현상을 관찰할 수 있게 해줍니다. 예를 들어, 84개 국가의 수백만명의 사람들의 감정의 진동과 같은 것들을 (Golder et al., 2011).

LDA 토픽모델링은 하나의 문서에 복수의 토픽이 내재된 상황을 전제로 하여 문서 내 여러 단어의 조합을 토대로 문서의 토픽들을 확률적으로 파악하는 통계적 분석 방법이다(Blei, D. M.,2012). 토픽모델링은 단순히 단어의 출현 빈도수를 기반으로 하는 텍스트 분석과는 다르며, 토픽이라는 변수를 통한 확률 기반의 분석을 함으로써 과적합 문제가 적고, 새로운 데이터가 투입되어도 분석 가능하다는 장점이 있다(박영욱, 정규엽, 2021).

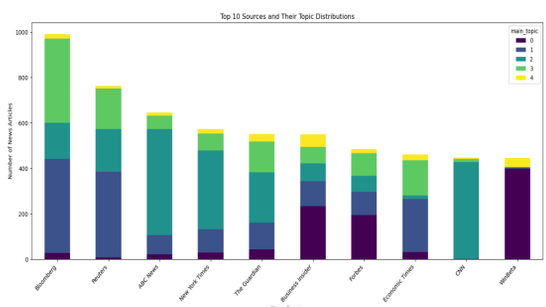
연구방법

본 연구는 여러 소셜 미디어 플랫폼에서 가져온 37,288개의 뉴스 제목을 포함하는 공개 데이터셋을 사용하였다. 데이터를 사전 처리하여 텍스트를 정리하고 불필요한 문자와 불용어를 삭제한 후, Gensim 라이브러리의 LDA 모델을 활용하여 주제 모델링을 수행하였다.이 모델은 텍스트 내의 단어의 동시 발생 패턴을 통해 잠재적 주제를 식별하며, 실험을 통해 최적의 주제 수를 결정하였다.모델의 출력을 직관적으로 표현하기 위해 pyLDAvis 라이브러리를 사용하여 상호 작용식 주제 시각화를 진행하였고,matplotlib와 seaborn 라이브러리를 활용하여 주요 뉴스 출처와 그들의 주제 간의 관계를 시각화하였다. 이러한 방법을 통해, 본 연구는 현재의 뉴스 트렌드와 주요 사안에 대한 깊은 이해를 제공하였다.

연구결과



1)마이크로소프트와 그 제품
2) 세계 경제, 특히 중국 경제
3) 오바마의 정치 활동 및 쿠바와 대법원과 관련된 주제
4) 국제 경제 및 정책
5) 마이크로소프트 기술과



중동 이슈
행은 다양한 뉴스 출처를 나타낸다
열은 LDA로 식별된 주제(0, 1, 2, 3, 4)를 나타낸다
표 내의 값은 해당 출처에서 특정 주제를 중점으로 다룬 뉴스의 수를 나타낸다

결론

본 연구는 소셜 미디어 뉴스 제목의 주제 분포에 대한 초기 탐색을 수행하였으며, LDA 모델과 관련 시각화 도구를 사용하여 초기적인 관점을 제공하였다. 분석을 통해, 소셜 미디어에서 다양한 뉴스 주제의 관심도와 뉴스 기관이 더 많은 독자의 관심을 끌기 위해 어떻게 뉴스 제목을 선택하는지에 대한 선도적인 역할을 볼 수 있다. 본 연구는 단순히 시작점에 불과하지만, 디지털 시대의 뉴스 전파 추세에 대한 일부 통찰력을 제공한다. 앞으로의 연구에서 이 분야에 대한 더 깊은 이해와 확장을 기대한다.