1. 프로젝트 소개

- 발표: NLP 감정분석을 이용한 영화 리뷰 데이터 프로젝트.

- 목표: 영화 리뷰 텍스트 데이터를 분석하여 리뷰가 긍정적인지 부정적인지 감정 상태를 분류.

2. NLP 정의

- 자연어 처리(Natural Language Processing)란 컴퓨터가 인간의 언어를 이해하고 해석하는 기술과 알고리즘.

- 언어 이해: 컴퓨터가 텍스트나 음성 데이터의 의미와 의도를 분석.

- 언어 생성: 컴퓨터가 자연스러운 인간 언어를 생성하는 기능 개발.

3. 사용된 기술과 도구

- NLTK 패키지 사용: 문장 구문 분석, 단어 분할, 어간 추출 및 표제어 추출, 토큰화, 의미론적 추론 지원.

4. 데이터 전처리 과정

- HTML 태그 제거: BeautifulSoup를 이용해 텍스트에서 HTML 태그 제거.

- 정규표현식 사용: 특수문자 제거.

- 토큰화: 문장이나 긴 텍스트를 작은 단위로 분리.

- 불용어 제거: 문장에서 많이 사용되지만 분석에 큰 의미를 가지지 않는 단어들 제거.

- 어간 추출 및 표제어 추출: 단어를 기본형으로 변환.

5. 모델 설명

- 랜덤 포레스트 모델 사용: 여러 의사결정 트리의 출력을 결합.

- 과적합 줄임: 여러 개의 의사 결정 트리를 무작위로 생성하여 모델의 다양성 확보.

- 파라미터 설정: 의사결정 트리의 개수를 100개로 설정, cross validation은 10으로 설정.

6. 결과 및 해석

- 감정 분석 모델이 영화 리뷰를 처리하고 결과를 도출하는 방식 설명.

- 리뷰의 감정을 '0'(부정적인 감정) 혹은 '1'(긍정적인 감정)로 분류.

- 모델은 단어의 의미와 문맥을 고려하여 각 리뷰의 감정을 정확히 분류할 수 있음.

7. 보완점

- Word2Vec을 사용하여 단어의 분산 표현을 학습: 모델의 해석력 향상.

- CountVectorizer 문제 해결: 단어의 순서나 문맥을 고려하지 않는 한계를 Word2Vec으로 해결.

단어 사이의 상관관계와 문맥적 의미 포착 개선.

8. 로컬 표현 대 분산 표현

- 로컬 표현(원-핫 인코딩): 각 단어를 벡터의 하나의 차원과 대응, 단어의 존재 여부를 0과 1로 표현.

- 분산 표현: 단어를 벡터의 여러 차원에 걸쳐 표현, 단어 사이의 의미적 유사성과 문맥적 뉘앙스 반영.

9. Word2Vec 소개

- 단어의 분산 표현을 학습하는 신경망 구현체.

- 단어 간의 유사성을 기반으로 의미를 클러스터링하여 언어 이해를 개선.