**텍스트 마이닝 기법 선정에 대한 고찰**

**1. 초기 선정 기법 (BoW – TF-IDF 방식 사용)**

|  |
| --- |
| 파이썬을 이용한 웹 크롤링을 통해 유튜브 영상의 댓글을 수집하여 유의미한 데이터를 추출하는 과정에서, 댓글 문장을 벡터화하는 단계가 있습니다. 이 단계에서는 BoW(Bag of Words) 기법을 사용하여 문장에 나타난 단어들의 빈도수를 세어 문장 벡터값을 도출합니다. BoW 기법은 각 문장에서 단어의 출현 빈도만을 고려하여 벡터를 생성합니다. 이를 통해 댓글을 통해 유튜버에 대한 팔로워, 즉 지지도를 산출할 수 있으며, 유사한 문장들에 대한 클러스터링도 가능해집니다.  BoW 기법과 함께 TF-IDF(Term Frequency-Inverse Document Frequency) 방식을 사용하여 문장에서 단어의 중요도를 측정할 수도 있습니다. TF-IDF 방식은 특정 단어가 문서 내에서 자주 등장할수록(TF값이 높을수록) 그리고 전체 문서 중에서 그 단어를 포함한 문서가 적을수록(IDF값이 높을수록) 높은 값을 갖습니다. 이는 여러 댓글에서 자주 등장하지만, 전체 댓글들 중에서는 드물게 등장하는 단어의 중요도를 더 높게 평가합니다.  단어 빈도수만을 고려하는 BoW 기법의 단점은 문장 내 단어 순서를 고려하지 않아 단어 순서에 따른 의미 변화를 반영하지 못한다는 점입니다. TF-IDF 역시 단어의 빈도수와 문서 빈도수를 기반으로 하기 때문에, 단어 순서가 주는 의미 변화는 반영하기 어렵습니다. |

**2. 대체 텍스트 마이닝 기법 도안**

|  |
| --- |
| BoW 기법은 간단하고 빠르게 텍스트를 백터화할 수 있다는 장점이 있습니다. 하지만 문장의 맥락이나 단어 간의 관계를 반영하지 못하기 때문에, 텍스트의 의미를 정확하게 분석하는 데에는 한계가 있습니다. 이를 보완하기 위해 TF-IDF를 사용하여 단어의 중요도를 더 정확하게 평가할 수 있지만, 여전히 단어 순서에 따른 의미 변화를 반영하지 못한다는 한계가 존재합니다. 이러한 한계를 극복하기 위해 단어를 의미적 유사도를 반영하여 n차원 좌표공간에 배치하는 Word2Vec, Doc2Vec이나 각 단어를 순환신경망에 순차적으로 넣고 내부적으로 이 문장의 컨텍스트를 은닉층에 가지고 있게 되는 Seq2Seq방식 등 딥러닝 기반의 임베딩 기법들을 사용할 수 있습니다.  Word2Vec – (Word-to-Vector)  주변 단어들의 벡터를 가지고 빈칸에 있는 단어의 벡터를 도출시키는 방법인 CBOW방식과 중심 단어의 백터를 가지고 주변 단어들의 벡터를 예측하는 방법인 Skip-Gram 방법 총 두가지의 분류법이 있습니다. 이는 박지성-축구+야구 = 류현진 과 같은 유의미한 결과값을 도출해 낼 수 있습니다.  Seq2Seq – (Sequence-to-Sequence)  번역기에 대표적으로 사용되는 방식으로서 시퀀스를 전환하는 즉 문장에서 컨텍스트 벡터를 만드는 인코더 부분과 이로부터 새로운 문장을 생성하는 디코더 부분으로 나뉘어 있는걸 볼 수 있습니다. 이 방식은 주로 질의응답 혹은 새로운 텍스트를 생성해야 할 경우에 사용합니다. 문장의 길이가 길지 않고 적어도 사용할 수 있다는 장점이 있습니다. |

**3. 텍스트 마이닝 기법 비교 분석 후 최종 선정**

|  |
| --- |
| BoW방식을 제외한 다른 순서를 고려하는 텍스트 마이닝 기법을 살펴보았으나 Bow를 대체할만한 이점이 크게 존재하지 않다고 판단되었습니다. BoW방식은 문장 그대로 빈도수로 환산해 TF-IDF방식을 통해 정확도를 올려 문장 벡터값을 환산할 수 있지만, 다른 두 모델의 경우 딥러닝 기반의 입베드 기법으로서 사전에 학습된 데이터가 있어야 하며 주로 예측하거나 번역하는 즉 검색엔진 및 챗봇형 AI모델에 사용되는 방식이라 볼 수 있습니다.  또한 본 프로젝트에서 도출해내고 싶은 결과값은 댓글을 통한 지지도이며 예상 댓글 및 예측 댓글, 번역 댓글 등이 아니므로 목표성이 모호하다는 결론에 도달하였습니다. 지지도를 계산하기위한 데이터의 사전 전처리 과정에 의한 비정형데이터의 수치화과정이므로 딥러닝에 기초한 시간을 할애하기에는 무리하다고 판단되었습니다.  최종적으로 댓글에 대한 지지도 파악에 있어서 긍정적인 단어와 해당 컨텐츠에 부정적 및 유투버에 대해 지지하지 않는 단어 등을 확인함에 있어서 비교적 짧은 댓글들에 순서보다는 빈도수에 의한 긍정적인 댓글을 파악하는것에 중요도를 기여한다면 BoW기법을 사용하는게 더 큰 이점이 있다고 판단되어 최종적으로 BoW기법을 선정하게 되었습니다. |