텍스트마이닝 개인과제 Individual Assignment (Topic Modeling Project)

광운대학교 정보융합학부 2021204051 우지윤

1. 서론

저는 토픽 모델링 대상 주제로 현재 주목받고 있는 분야인 '메타버스(Metaverse)'를 선택하였습니다. 아시아 최대 메타버스 플랫폼이자 증강현실 아바타 서비스인 '제페토(ZEPETO)'가 널리 유행하며 작년 초 기업가치를 상승시켰다는 소식을 접한 적이 있습니다. 그리고 <'메타버스는 새 먹거리'..통신·플랫폼사도 뛰어들어_메타버스의 열풍은 계속된다②>, <메타버스서도 뜨거운 부동산 열풍…가상토지·건물 흥행몰이> 와 같이 메타버스를 다루는 뉴스 기사들도 자주 읽는데, 메타버스가 사회에 미치는 영향력과 중요도에 대해서도 매번 느낍니다. 그리고 작년에 수강한 지능형로봇학과의 '로봇명사와의 만남' 강의에서 오픈형 메타버스 플랫폼 ZEP을 이용하여 비대면으로 메타버스 공간에서 강의와 소통이 진행된 경험으로 인해, 메타버스 분야에 대해 분석하고 이해해보고 싶다는 생각이 생겼습니다.

메타버스는 가상 공간일 뿐만 아니라 우리가 사는 현실 세계와 가상 세계를 연결하는 연결고리이자 교차점이고, 가상 공간과 현실 세계가 융합하며 상호작용하는 공간을 말합니다. 메타버스는 최근 몇 년 동안 빠른 속도로 성장하고 있으며, 가상 현실 기술과 인공지능의 발전으로 더욱더 현실적이고 풍부한 경험을 제공할 수 있게 되었습니다. 메타버스는 현실 세계에서의 제약을 넘어서고, 창의적인 자유를 제공함으로써 사용자들에게 경험과 기회를 제공합니다. 메타버스의 잠재력은 상당히 크며, 향후 더욱 발전할 것으로 예상됩니다. 이에 '메타버스'를 주제로 선정하였습니다.

2. 데이터 수집 및 처리

광운대학교 중앙도서관의 해외 학술 DB 메뉴에서 자연 과학 분야 학술 저널 데이터베이스를 제공하는 Web of Science 사이트에 접속하였습니다. 주제 필드에서 'metaverse'를 검색창에 입력하였습니다. 총 663개의 결과가 나왔으며, 결과 범위를 다음과 같이 일부 재설정하였습니다. 최신 결과를 반영하기 위해 2021/2022/2023년의 연도, 문서 유형으로 Article/Proceeding Paper/Early Access/Editorial Material을 선택하여 총 520개의 최종 검색 결과를 도출하였습니다. 레코드 출처 1~520을 선택하고 excel로 초록까지 포함하여 내보낸 이 엑셀 파일이 과제의 전처리 전 토픽 모델링 대상 데이터입니다.

Jupyter Notebook에서 작업을 진행하였습니다. 총 72개의 열 중에서 작가(Authors)/제목(Article Title)/초록(Abstract)/출판연도(Publication Year)의 4개의 열만 추출하여 데이터를 재구성하였습니다. 데이터를 불러온 결과 (520, 4)의 shape를 가지고 있었고, 결측치 값을 isna().sum()으로 확인하며 행이 초록이나 연도 열의 데이터를 결측치로 가지면 dropna()를 통하여 제거하였습니다. 다음과 같은 전처리 작업을 수행한 후 최종 데이터의 shape는 (425, 4)로 줄어든 형태를 보이는 걸 확인하였고, 데이터 작업 전 결측치 처리의 중요성과 대처에 대한 필요성을 느꼈습니다.

3. 토픽 모델링 결과

저는 토픽 모델링 기법 중 잠재 디리클레 할당(LDA, Latent Dirichlet Allocation) 기법을 적용하였습니다. LDA 기법은 주로 텍스트 마이닝, 정보 검색, 추천 시스템 등 다양한 자연어 처리 태스크에서 활용됩니다. 토픽 모델링은 문서의 구조를 파악하고 의미 있는 정보를 추출하는 데 도움이 되며, LDA는 그중에서도 대표적인 알고리즘입니다. LDA는 각 문서가 여러 개의 토픽으로 구성되어 있다고 가정하고, 문서 내의 단어들이 이러한 토픽들에 따라 분포되어 있다고 가정합니다. LDA를 통해 추출된 토픽은 단어의 확률 분포로 표현되므로, 각 토픽은 해당 토픽에 가장 적합한 단어들을 가지고 있습니다. 이를 통해 토픽의 의미를 이해하고, 새로운 문서의 토픽 분포를 예측할 수 있습니다.

본격적인 LDA 토픽 모델링을 진행하기 전, 최적의 토픽 수를 선택하기 위하여 주제에 대한 일관성 점수(Coherence Score)를 분석하였습니다. 일관성 점수가 높은 모델은 주제 간의 일관성이 높아 의미 있는 토픽을 잘 나타내는 경향이 있습니다. CoherenceModel을 이용하여 일관성 점수를 구한 결과, 토픽의 수가 5개일 때 점수가 약 0.276으로 가장 높아 5개로 토픽의 수를 선정하였습니다. 그리고 크게 총 두 가지 작업을 진행하였습니다. 전체 데이터에 LDA를 활용한 토픽모델링, 연도별 (2021/2022/29023)로 데이터에 LDA를 활용한 토픽모델링을 진행하였습니다.

먼저, 전체 데이터에 대한 LDA 토픽모델링 작업을 수행하였습니다. 불용어를 제거하고, 사전을 구축하고 문서-단어 행렬을 생성한 후 LDA 모델을 생성하여 토픽을 출력한 결과는 다음과 같습니다. 첫번째 토픽은 'metaverse, virtual, based, model, research, technology, new, social, digital, results', 두번째 토픽은 'metaverse, virtual, reality, based, digital, research, vr, using, also, new', 세번째 토픽은 'metaverse, virtual, digital, reality, research, new, data, users, technologies, using', 네번째 토픽은 'metaverse, research, virtual, digital, reality, using, learning, 3d, future, real', 다섯 번째 토픽은 'virtual, metaverse, using, reality, research, digital, vr, new, based, design'으로 구성되어 있었습니다.

LDA 토픽 모델링 적용에 따른 토픽 도출 결과, '새로운 디지털 모델의 기술에 기초한 가상의 메타버스에 관한 연구 결과', '디지털이나 vr을 기반으로 하는 새로운 가상현실 메타버스에 관한 조사 결과', '가상현실 메타버스를 사용하는 사용자들의 데이터에 관한 새로운 디지털 기술 조사 결과', '디지털과 3d를 활용하여 가상현실 메타버스에 대한 연구를 통해 현실과 미래를 배우는 결과', '가상현실 메타버스를 기반으로 한 새로운 디지털 디자인을 활용한 연구 결과'라는 주제를 추론해보았습니다. 더 정확한 주제를 결정하기 위해서는 데이터의 문맥을 자세히 살펴보아야 합니다.

다음으로, 데이터로부터 5개의 주제를 5개의 단어로 개수를 지정하여 끌어내고 그래프/표/인터랙티브 시각화를 진행하였습니다. 토픽별 상위 5개의 단어를 추출하고 데이터프레임으로 변환한 과정을 추가로 진행한 것 외에는 앞과 동일한 과정으로 진행되었습니다. 불용어를 제거함으로써 텍스트를 정제하고, 각 문서를 단어 리스트로 변환하였으며, 사전을 구축하였습니다. 그리고 문서-

단어 행렬(corpus)을 생성하고 LDA 모델도 생성하여 결과를 출력하는 과정을 거쳤습니다. 토픽별 단어 빈도를 그래프와 표로 시각화하여 나타낸 결과, 첫번째 토픽은 단어 'metaverse, virtual, digital, research, proposed'로 구성되어 있었고 두번째 토픽은 'metaverse, virtual, based, digital, new'로, 세번째 토픽은 'metaverse, virtual, reality, research, using'으로, 네번째 토픽은 'metaverse, virtual, research, digital, study'로, 다섯 번째 토픽은 'virtual, metaverse, reality, digital, vr'로 구성되어 있었습니다. 토픽별 주어진 5개의 단어로부터 '가상의 메타버스와 디지털에 제안된 조사', '가상의 메타버스와 디지털에 기반한 새로운 것', '메타버스의 가상현실을 이용한 조사, '가상의 메타버스와 디지털에 관한 조사와 연구', '가상현실 메타버스와 디지털과 vr'라는 주제를 생각해보았습니다.

Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
metaverse	metaverse	metaverse	metaverse	virtual
virtual	virtual	virtual	virtual	metaverse
digital	based	reality	research	reality
research	digital	research	digital	digital
proposed	new	using	study	vr

더불어 pyLDAvis 모델을 이용하여 LDA 토픽모델링 결과를 인터랙티브 시각화를 진행하였습니다. 우측 상단 슬라이딩 바를 통해 관련성을 조절할 수 있고, 관련성 값은 0부터 1사이의 값을 갖습니다. 값이 0에 가까울수록 전체 문서에서 출현 횟수는 적더라도, 해당 토픽을 다른 토픽과 차별성 있게 구분할 수 있는 단어인지에 집중한다는 의미입니다. 좌측 Intertopic Distance Map에서 원이 모두 각각의 토픽이며, 서로 가까이 붙어있는 원일수록 유사한 토픽입니다. 원이 클수록 토픽에 해당되는 단어의 개수가 많다는 것도 알 수 있습니다. 원에 마우스를 가져다 대면 우측에 해당 토픽을 구성하는 단어가 전체 문서 데이터 대비 현재 토픽의 키워드로 구성되었는지 비율을 알려줍니다. 또한, 전체 문서 데이터의 단어 대비 해당 토픽이 구성하는 단어들의 비율을 제공합니다. 이처럼 토픽별로 어떤 단어들이 어떤 비율로 구성되어 있는지 파악함으로써 주제를 유추할 수 있으며, 나아가 전체 문서 데이터에 어떤 토픽이 어떤 비율로 구성되어 있는지 중요도에 대해 파악할 수 있습니다. Relevance metric이 1인 경우에서 이 문서의 토픽들을 살펴본 결과, 1번과 2번 원이 가까이 붙어있는 걸로 보아 유사한 토픽임을 알 수 있습니다. 그리고, 1번 원의 크기가 가장 큰 것으로 보아 토픽에 해당하는 단어의 수가 가장 많음을 알 수 있습니다. 그리고 각 토픽에서 가장 많이 등장한 상위 5개의 단어를 추출하여 주요 토픽을 추출하고, 각 주요 토픽의 등장 횟수도 확인하였습니다. 두 번째 토픽의 등장 횟수가 130번으로 가장 빈도수가 높았으며, 각 문서가 어떤 토픽에 속하는지도 확인하였습니다. 대다수의 문서가 첫 번째 혹은 세 번째 토픽의 구성 비율이 0.99로 가장 높은 것을 알 수 있었습니다.

다음으로, 연도별 데이터에 대한 토픽모델링 작업을 수행하기 위해 임의로 토픽의 수를 2개로 정하였습니다. 연도별로 각각 2개의 토픽을 3개의 단어를 통해 끌어내고 이에 대해 히트맵으로 토픽별 단어 빈도 시각화를 진행하였습니다. 2021년~2023년 사이에 토픽의 변화 양상을 살펴보고 싶었습니다. 데이터의 Publication Year 열의 데이터가 각각 2021, 2022, 2023인 경우를 나누어

분석을 진행하였습니다. 2021년의 첫 번째 토픽은 단어 ‘virtual’, ’metaverse’, ’world’, 두 번째 토픽은 단어 ‘virtual’, ‘metaverse’, ‘reality’ 과 같은 토픽별 단어 빈도가 나타났습니다. 이에 따라 제가 예측한 주제는 ‘가상 세계 메타버스’와 ‘가상현실 메타버스’입니다. 2022년의 첫 번째와 두 번째 토픽은 단어 ‘metaverse’, ‘virtual’, ‘reality’와 같은 토픽별 단어 빈도가 나타나 ‘가상현실 메타버스’로 생각해보았습니다. 마지막으로 2023년의 토픽들은 단어 ‘metaverse’, ‘research’, ‘virtual’이 나타났습니다. 제가 예측한 주제는 ‘가상의 메타버스에 관한 조사’ 입니다. 이를 통하여, 3년간 메타버스 분야에 관한 주제와 토픽-단어 분포의 큰 변동은 없었으며 유사한 양상임을 알 수 있었습니다.

4. ChatGPT 적용 및 비교 결과

ChatGPT에게 공통된 형식으로 프롬프트(명령문)를 입력하여 메타버스 분야에 관련된 주요 토픽 5개를 도출하고자 하였습니다. “메타버스 분야와 관련된 주요 토픽 5개를 알려줘”를 프롬프트에 적용하였습니다. 이에 ChatGPT가 제공한 메타버스 분야와 관련해 도출한 주요 토픽 5가지는 ‘가상현실(VR)’, ‘증강현실(AR)’, ‘가상경험 경제’, ‘가상 소셜 네트워킹’, ‘가상현실 교육’입니다.

LDA 토픽 모델링을 기반으로 해당 분야 내 5개의 주요 토픽을 도출한 결과는 ‘새로운 디지털 모델의 기술에 기초한 가상의 메타버스에 관한 연구 결과’, ‘디지털이나 VR을 기반으로 하는 새로운 가상현실 메타버스에 관한 조사 결과’, ‘가상현실 메타버스를 사용하는 사용자들의 데이터에 관한 새로운 디지털 기술 조사 결과’, ‘디지털과 3D를 활용하여 가상현실 메타버스에 대한 연구를 통해 현실과 미래를 배우는 결과’, ‘가상현실 메타버스를 기반으로 한 새로운 디지털 디자인을 활용한 연구 결과’입니다.

두 기법에 따른 메타버스 분야 관련 주요 토픽 도출을 본 결과, 모두 주어진 텍스트 데이터로부터 텍스트와 관련된 주제를 예측 및 추정한다는 점을 알 수 있었습니다. 그리고 텍스트의 문맥을 이해하며, 텍스트 데이터를 분석하고 해석한다는 공통점이 있었습니다. 이러한 공통점은 두 기법이 텍스트 데이터를 분석하고 처리함에 있어서 일부 유사성을 가지고 있다는 점을 나타냅니다.

LDA 토픽 모델링 기법은 통계적 모델링을 사용하며 메타버스 분야를 다룬 문서 집합에서 단어의 출현 패턴과 분포를 통해 토픽을 추론하므로, 토픽은 주로 단어의 연관성을 기반으로 합니다. 그러나, ChatGPT는 사용자가 입력한 문장의 문맥을 이해하고 학습된 모델을 기반으로 주제를 예측합니다. 이에 ChatGPT는 메타버스와 관련된 더 넓은 의미와 문맥을 파악할 수 있으며, 도메인 지식이나 문장 구조 등 다양한 특징을 활용하여 결과를 도출한다는 특징이 있습니다. 이처럼 토픽 모델링과 ChatGPT는 주제 추론에 사용되는 방식과 목적과 기능이 다릅니다. 그러므로 각 각의 특징과 장점, 사용 사례를 더 자세히 살펴보아야 하며 주어진 상황에 맞게 적절한 모델을 선택하는 것이 중요합니다.