뉴스로 살펴본 광진구의 주요 이슈와 핵심 인물-기관-지역:

LDA토픽모델링과 개체명 분석을 통해

팀명(개인): ku-optimizer(이승현)

**목차**

**1. 연구의 배경 및 목적**

**2. 이론적 배경: 지자체의 역할**

**3. 빅데이터 분석**

1) LDA 토픽 모델링(Latent Dirichlet Allocation topic modeling)

2)개체명 분석(Named Entity Recognition, NER)

**4. 연구방법(절차)**

1)웹 스크래핑

2)형태소 분석 및 키워드 추출

3)LDA토픽모델링

**5. 연구결과**

1)추출된 토픽의 해석과 재구성

2) LDA토픽모델링 분석 결과

3) 개체명 분석 결과

**6. 연구 논의**

1) 결과분석

2) 개선방안

**7. 참고문헌**

**1. 연구의 배경 및 목적**

뉴스는 최신성과 객관성, 공공성을 지닌 매체로서, 정치, 경제, 문화 등 다양한 주제와 사회적으로 중요한 이슈들을 다룬다. 이러한 특성으로 인해 뉴스는 우리 사회를 가장 잘 반영하는 지표라고 할 수 있다. 광진구 또한 하나의 작은 사회로서, 구성원들이 상호작용하며 역동적으로 변화하고 발전한다. 이렇게 다양성을 띠는 광진구를 파악하기 위해 가장 적합한 데이터는 뉴스의 빅데이터라고 판단하였다. 방대한 양의 뉴스 빅데이터를 분석함으로써 광진구에 대한 통찰력과 숨어있는 패턴을 발견할 수 있을 것이다. 예를 들어, 어떤 분야가 관심을 끌고 있는지, 또는 어떤 주제가 더 다루어져야 하는지 확인할 수 있다. 나아가 기존 데이터 분석을 통해 향후 발전방향을 예측할 수 있을 것이다.

사회는 정형화된 데이터만으로는 완전히 설명되기 어려우며, 오히려 비정형 텍스트가 보다 나은 이해를 제공한다. 단순한 숫자나 표로는 복합적인 사회 현상을 온전히 파악하기 어렵기 때문이다. 비정형 텍스트에는 문맥 속에서 다양한 의견과 관점이 녹아 있어, 이를 통해 사회현상을 보다 용이하게 이해할 수 있다. 따라서 뉴스 빅데이터를 분석하기 위해서는 텍스트 마이닝이라는 분석기법을 활용한다. 텍스트 마이닝은 대량의 텍스트를 구조화하고 유의미한 정보를 추출하여 패턴과 지식을 발견하는 기법으로, 자연어 처리, 통계분석 등의 기술을 활용한다. 텍스트 마이닝은 다양한 분야에 적용 가능하며, 중요한 정보만을 자동화된 방식으로 처리하므로 효율적이다.

빅데이터는 대규모의 다양한 정보를 제공하기에 사회를 다각도로 조명할 수 있다. 따라서 빅데이터 분석의 요점은 광진구의 특정 문제점에 국한되지 않고, 전반적인 문제들이 얽혀 있는 패턴을 파악하는 데 있다. 본 연구에서는 빅데이터를 활용하여 광진구 내에 숨어 있는 패턴을 발견하고자 한다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

△텍스트 마이닝 시각화 자료

**2. 이론적 배경: 지자체의 역할**

광진구의 다양한 이슈를 고찰하기에 앞서, 지방자치단체가 주민들의 삶의 질 향상을 위해 어떠한 역할을 수행해야 하는지에 대한 선행 이해가 필수적이다. 이에 관련하여 다수의 연구논문에서는 지자체의 기능과 역할에 대한 논의와 제언을 다각도로 제시하고 있다(마지막장 참조문헌 열거). 해당 선행연구들을 통해 지방자치단체의 주요 기능과 역할, 나아가 광진구가 지향해야 할 방향성에 대한 고찰이 가능할 것이다.

상기 내용에 따르면, 지방자치단체는 지역발전과 주민 삶의 질 향상을 위해 다음과 같은 다방면에 걸친 역할과 기능을 수행해야 할 것으로 요약된다.

지역산업 분야에서는 산업클러스터 조성, 중소기업 지원, 창업 촉진 프로그램 운영 등을 통해 지역경제 성장과 발전을 도모해야 한다. 아울러 기업 유치를 통한 일자리 창출과 경제 활성화에 기여해야 한다.

정보통신기술(ICT) 활용을 통해서는 행정서비스의 효율성과 접근성을 제고하고, 스마트시티 솔루션 등을 도입하여 도시문제 해결에 나서야 한다.

복지 분야에서는 사회적 약자 보호와 삶의 질 향상을 위해 노인, 장애인, 저소득층 등을 대상으로 한 맞춤형 복지서비스 제공을 강화해야 한다.

교육 증진을 위해서는 지역주민 교육의 질 제고와 평생교육 기회 확대로 주민 경쟁력을 높이고, 교육 인프라 개선 및 새로운 프로그램 지원이 요구된다.

교통 분야에서는 교통 혼잡 완화와 대중교통 접근성/편리성 향상을 위한 지속가능한 교통체계 구축이 필요하다.

삶의 질 향상을 위해서는 생활편의시설 확충과 보건/안전시설 접근성 제고, 공원/녹지/레크리에이션 공간 조성 및 관리 강화가 요구된다.

주민 상호작용 활성화를 위해서는 커뮤니티 활성화 프로그램, 주민참여 플랫폼 개발 등으로 주민 간 네트워킹과 소통, 협력 기반을 구축해야 한다.

문화 분야에서는 지역 문화의 독창성과 다양성 보존/발전, 주민 문화참여 활성화가 필요하다. 특히 대중문화콘텐츠 개발/보급 지원, 지역예술가 협력 기반 문화행사/축제 개최를 통해 지역문화 다양성과 창조성을 촉진해야 한다.

관광 활성화를 통해 지역경제에 기여하고자 한다면, 지역 자원을 활용한 관광상품 개발과 관광인프라 구축 노력이 수반되어야 한다.

**3. 빅데이터 분석기법**

**1) LDA토픽모델링(Latent Dirichlet Allocation topic modeling)**

LDA 토픽 모델링은 자연어 처리 및 텍스트 분석 분야에서 문서 내 잠재된 토픽을 추론하고, 토픽과 단어 간의 관계를 모델링하는 기법이다. 이는 문서에서 추출된 단어를 기반으로 유사한 단어들을 군집화하여 토픽을 생성한다. 생성된 토픽에 해당하는 문서들을 모아 대량의 문서를 토픽별로 체계적으로 정리할 수 있다. 상기 그림에서 D는 전체 문서의 개수, K는 토픽의 수, N은 d번째 문서의 단어 수를 나타낸다. Wd,n은 d번째 문서의 n번째 단어, ϕk는 k번째 토픽, θd는 d번째 문서의 각 토픽에 대한 비중을 나타내는 벡터, Zd,n은 d번째 문서의 n번째 단어가 어느 토픽에 해당하는지를 지시하는 변수이다. 문서-단어 행렬을 단어-토픽 행렬과 토픽-문서 행렬로 분해한다.

토픽 모델링은 텍스트 자료를 통계적으로 분석하여 객관적인 결과를 도출할 수 있다는 장점이 있다. 숨겨진 관계를 발견할 수 있으며, 대량의 데이터를 소수의 토픽으로 신속하게 추출할 수 있다. 특히 다양한 원인이 복합적으로 작용하는 자치구 문제를 해결할 때, 모든 주제를 포텍스트, 도표, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명괄적으로 다루며 다각적인 해결방안을 모색하는 데 유용하다

△LDA모델의 원리

LDA 토픽 모델링의 일반화 능력을 평가하는 지표로는 perplexity와 coherence가 있다. Perplexity(복잡도)는 확률 모델이 실제 값을 얼마나 정확하게 예측하는지를 나타내는 척도로, 값이 낮을수록 예측 정확도가 높음을 의미한다. Coherence(일관성)는 토픽 내 단어들이 의미론적으로 일관성 있게 구성되었는지를 평가한다. Coherence 값이 높을수록 모델 결과의 의미 있는 정도가 높다는 것을 나타낸다. Coherence를 측정하는 다양한 지표 중 가장 널리 사용되는 'c\_v'를 본 연구에서도 활용하였다.

**2) 개체명 분석(Named Entity Recognition, NER)**

개체명 인식(Named Entity Recognition, NER)은 텍스트 데이터에서 인명, 기관명, 지명, 날짜 등과 같은 특정 유형의 정보를 분류하고 탐지하는 자연어처리 기술이다. 이는 정보추출, 문서요약, 검색엔진 등 다양한 분야에서 활용되고 있다. 뉴스 데이터 분석과 관련해서는 한국언론진흥재단에서 개발한 KPF-BERT 기반 개체명 인식 모델이 높은 성능을 보이는 것으로 알려져 있다.

BERT(Bidirectional Encoder Representations from Transformers)는 구글에서 개발한 대규모 사전훈련 언어모델로, 양방향 컨텍스트를 효과적으로 반영할 수 있다. KPF-BERT는 한국어 뉴스와 문서를 대량으로 학습한 모델로, 한국어 기사 분석에 활발히 활용되고 있다. 이 모델은 한국어의 미묘한 문맥과 언어적 특성을 잘 포착하여, 한국어 텍스트에 대한 개체명 인식 작업에서도 우수한 성능을 보인다고 평가받고 있다. 본 연구에서는 KPF-BERT 기반 개체명 인식기를 활용하여 인명, 기관명, 지명 등의 탐지를 수행하였다.

**4. 연구방법(절차)**

**1) 웹 스크래핑**

웹 스크래핑(Web Scraping)이란 월드 와이드 웹에 저장된 대량의 데이터를 자동으로 추출하여 가공 및 활용 가능한 데이터로 변환하는 과정을 일컫는다. 이를 위해 urllib, BeautifulSoup, pandas 등의 파이썬 패키지가 활용된다. 수집 대상 URL은 웹사이트의 구조에 따라 원하는 정보를 함유하고 있다. '&query=' 뒤에는 검색어의 URL 인코딩 값이 위치하며, '%sort=' 뒤의 숫자에 따라 뉴스의 정렬 방식(0: 관련도 순, 1: 최신 순, 2: 오래된 순)이 결정된다. '&photo=' 뒤의 숫자는 수집할 데이터 유형(0: 전체, 1: 포토, 2: 동영상, 3: 지면기사, 4: 보도자료, 5: 자동생성기사)을 지정한다. 언론사 정보는 고유 ID 번호로 저장되며, 'from'과 'to' 사이에는 수집 기간의 시작일과 종료일이 입력된다. 본 연구에서는 이러한 URL 구조를 활용하여 특정 기간 내 작성된 중심 키워드를 포함한 뉴스의 전체 내용을 자동으로 추출하고자 하였다.

검색어를 '광진구'로 설정하고, 뉴스 작성 기간을 2023년 4월 1일부터 2024년 3월 31일까지로 한정하였다. 이후 언론사, 기사제목, 작성날짜, URL, 기사내용 순으로 데이터를 데이터프레임에 수집하여 엑셀 파일로 정리하였다. 다양한 언론사에서 수집한 초기 데이터 수는 총 6,058개였으며, 중복 제목 기사를 제거한 뒤 최종적으로 4,561개의 데이터를 확보하였다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

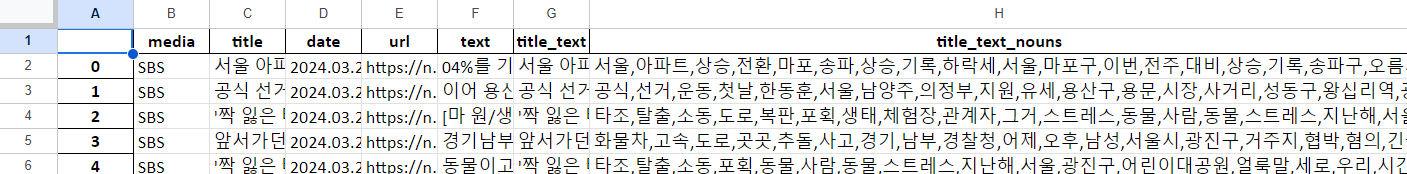
텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

△언론사에서 수집한 광진구 관련 기사

**2) 형태소 분석 및 키워드 추출**

광진구 뉴스 데이터를 수집한 이후에는 기사 내용 속에 포함된 명사를 추출하는 전처리 과정이 필요하다. 이를 위해 기사 내용을 형태소 단위로 분석하여 문장의 주요 구성 요소인 명사만을 추출한다. Konlpy 라이브러리와 Kiwipier 패키지를 활용하여, 명사인 경우에는 리스트에 저장하고 명사가 아닌 경우에는 제거하는 조건문을 적용한다. 다량의 뉴스 기사에서 추출된 명사들 중에는 자주 등장하는 고빈도 명사들이 존재한다. 이러한 고빈도 명사는 광진구에서 비중 있게 다루어지는 주요 키워드일 가능성이 높다. 따라서 출현 빈도가 높은 상위 100개의 단어를 추출하여 수집한다. 나아가 전체 기사 내용에서 단어를 추출하는 것뿐만 아니라, 단어를 특정 범주로 묶은 후 각 범주 내에서 출현 빈도가 높은 순서대로 단어를 추출할 수 있다. 본 연구에서는 '사람', '장소', '지역' 등 3개의 범주를 설정하였다. 추가적으로 단어 출현 빈도를 기반으로 하는 워드 클라우드(Word Cloud) 기법을 활용하여 주요 키워드를 한눈에 시각화하였다.



△뉴스기사 내용 속 명사를 추출하여 만든 엑셀표

3) **LDA토픽모델링**

상기 Perplexity(복잡도) 분포도에서 22번째 지점은 일시적으로 하락하였다가 다시 상승하는 양상을 보였다. 마찬가지로 Coherence(일관성) 지표인 c\_v 분포도에서도 22번째 지점이 상승 후 하락하는 패턴을 나타냈다. 이는 22개의 토픽 수가 Perplexity를 최소화하고 Coherence를 최대화하는 최적점임을 시사한다. 따라서 본 연구에서는 토픽의 개수를 22개로 설정하였다.

추가 분석을 통해 토픽 분산(Topic Dispersion) 그래프를 도출하였다. 상단 그래프의 원 분포는 의미론적으로 유사한 토픽들이 공간상 인접하게 위치하도록 한 것이다. 하단 그래프는 22개 토픽 각각과 가장 연관성이 높은 상위 30개 단어를 추출한 결과이다. 각 토픽의 명칭은 연결된 30개 단어들의 의미를 포괄할 수 있도록 대표성 있게 선정하였다.

|  |
| --- |
| 그래프, 라인, 텍스트, 도표이(가) 표시된 사진  자동 생성된 설명 |
| 텍스트, 그래프, 라인, 도표이(가) 표시된 사진  자동 생성된 설명 |

△perplexity와 coherence를 나타낸 그래프

|  |  |
| --- | --- |
| 텍스트, 스크린샷, 도표이(가) 표시된 사진  자동 생성된 설명 | 텍스트, 스크린샷, 도표이(가) 표시된 사진  자동 생성된 설명 |

* 토픽 분산 그래프

**5. 연구 결과**

**1) 추출된 토픽의 해석과 재구성**

주요 토픽들을 요약하면 다음과 같다:

Topic 0: 보건의료(병원, 환자, 건강, 코로나, 걷기, 장애, 의료 등)

Topic 1: 부동산(상승, 거래, 성동구, 아파트, 강남구 등)

Topic 2, 3: 영화산업(영화, 롯데시네마, 건대, 간담회, 배우, 인사말, 보고회, 강하늘 등)

Topic 4: 공공편의시설(설치, 안전, 광진구청, 서울시, 터미널, 어린이집 등)

Topic 5: 교통(도로, 보드, 사고, 차로, 오토바이, 횡단보도 등)

Topic 6, 11, 12: 사건사고(소방, 화재, 대피, 신고, 발생, 경찰, 살해, 추적, 남성, 혐의, 체포, 흉기, 학대, 교사, 폭행 등)

Topic 7: 공연/전시(공연, 콘서트, 개최, 음악, 라이브, 밴드 등)

Topic 8: 정치(위원장, 공약, 인파, 의원, 대책, 민주당 등)

Topic 9: 노동(폭염, 노동자, 노동, 생수, 청춘, 강력 등)

Topic 13: 복지(출산, 취약, 전세, 상공인, 소득 등)

Topic 14: 대중가요(쇼케이스, 앨범, 미니, 발매, 그룹, 포즈, 걸그룹 등)

Topic 15: 교육(폐교, 학교, 교통, 인구, 운영, 대학, 학습 등)

Topic 16: 관광(뚝섬, 한강공원, 어린이대공원, 물놀이 등)

Topic 18: 행사/모임(회장, 워커힐, 개최, 경영, 호텔, 박람회)

Topic 20, 21: 부동산(분양, 가구, 아파트, 단지, 경쟁, 명의, 법인, 부동산, 최고 등)

한편, Topic 10, 17, 19의 경우 상위 30개 단어만으로는 토픽을 대표할 수 있는 명확한 주제를 도출하기 어려워 기타로 표기했다.

분석 결과, 광진구 뉴스 데이터에서 부동산, 영화산업, 사건사고, 공연/전시, 정치, 교육, 관광 등 다양한 분야의 토픽이 추출되었음을 확인할 수 있다. 이를 통해 지역사회에서 관심을 받고 있는 주요 이슈들을 파악할 수 있었다.

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명

△웹 스크롤링한 문서들을 소속토픽과 최종토픽명으로 정렬한 엑셀표

텍스트, 스크린샷, 폰트, 라인이(가) 표시된 사진

자동 생성된 설명**2) LDA토픽모델링 분석 결과**

△월별 기사 수

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명2023년 7월에 가장 많은 기사 수가 작성되었고 2023년 4월에 가장 적은 기사 수 가 작성되었다.

△토픽별 기사 수

부동산, 사건사고, 영화산업 및 대중가요에 관한 기사 수가 많이 작성되었다. 노동, 복지, 교통, 보건의료에 관한 기사 수는 적게 작성되었다.

텍스트, 스크린샷, 폰트, 그래프이(가) 표시된 사진

자동 생성된 설명

△언론사별 기사 수

연합뉴스에서 보도된 기사 수가 가장 많았으며 지상파 3사(SBS, MBC, KBS)의 기사 수가 적게 확인된다.

텍스트, 스크린샷, 번호이(가) 표시된 사진

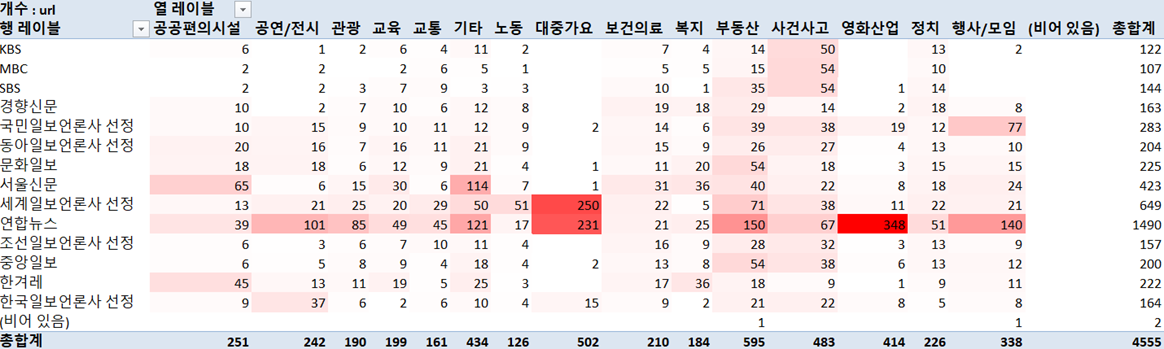
자동 생성된 설명

△월별 언론사의 기사 수

텍스트, 스크린샷, 번호, 달력이(가) 표시된 사진

자동 생성된 설명

△월별 토픽의 기사 수



△언론사별 토픽의 기사 수

**3) 개체명 분석 결과**

|  |
| --- |
| 텍스트, 폰트, 그래픽, 그래픽 디자인이(가) 표시된 사진  자동 생성된 설명 |
|  |
| △광진구 뉴스에서 등장한 주요 인물 |
| 텍스트, 폰트, 그래픽, 그래픽 디자인이(가) 표시된 사진  자동 생성된 설명 |
|  |
| △광진구 뉴스에서 등장한 주요 기관(조직) |
| C:\Users\82104\AppData\Local\Temp\BNZ.6634279e73243983\all_ner_LCP.png |
|  |
| △광진구 뉴스에서 등장한 주요 장소(지역) |
|  |

**6. 연구 논의**

**1) 결과 분석**

토픽별 기사 수 분포를 살펴보면, 광진구의 주요 산업이 대중문화나 부동산 등의 특정 영역에 집중되어 있음을 알 수 있다. 2023년 7월에는 다수의 가수 앨범 발매와 영화 개봉 등 대중문화 콘텐츠가 증가하면서 관련 기사 수가 급증한 것으로 해석된다. 또한, 인물 출현 빈도 그래프에서 연예인들이 주로 상위권에 분포해 있는 것으로 보아, 광진구가 대중문화의 영향을 많이 받고 있음을 시사한다. 반면에, 교육, 노동, 보건의료 등의 토픽 기사 수가 상대적으로 적어, 주민들에게 필수적인 사회복지 공급이 충분치 않을 수 있음을 유추할 수 있다.

한편, 토픽 추출 과정에서 IT 기술 관련 주제가 수집되지 않았는데, 이는 해당 분야에 대한 관심 부족이 현지 경제에 부정적 영향을 미칠 수 있음을 시사한다. 특히 첨단 기술의 부재는 미래 산업 경쟁력에 지대한 영향을 끼칠 수 있다. 경기도 분당구의 '판교테크노밸리'는 국내외 IT 기업과 인재가 모여 4차 산업혁명을 선도하는 거점 지역으로 자리매김했다. 광진구 또한 IT 관련 산업 육성을 통해 미래 경쟁력 확보에 주력할 필요가 있어 보인다.

지역명 출현 빈도 분석 결과, 광진구 내 동보다는 강남구, 마포구 등 타 지역이 더 많이 언급되었다. 이는 자양동, 광장동 등 광진구 동의 특색이 부족함을 보여준다. 반면에, 성동구의 성수동은 개성 있는 소품샵, 이국적 맛집, 다양한 전시회 등으로 젊은 소비자층을 불러모으며 활기찬 상권을 이루고 있다. 광진구 또한 동별 특색 있는 상권 육성이 필요해 보인다.

마지막으로, 지상파 3사(KBS, MBC, SBS)의 광진구 관련 보도 건수가 서초구에 비해 현저히 적은 것으로 나타나, 광진구의 방송 노출도가 낮음을 확인할 수 있었다. 광진구의 지상파 3사의 기사 수는 약 100개정도이다. 서초구의 지상파 3사의 기사 수를 웹 스크롤링을 통해 구한 결과 약 700개의 기사 수가 보도됐다.

**2) 개선방안**

광진구에는 예스24 라이브홀, CGV 강변점(1호점), KU시네마테크(건국대학교 내 독립영화 상영관) 등 대중문화를 선도하는 다양한 문화시설들이 위치해 있다. 이러한 차별화된 문화인프라를 활용하여 대중문화 관련 산업을 고부가가치화할 필요가 있다. 예를 들어, 대한민국 영화역사 전시회 기획, 독립영화 지원 확대를 통한 예술인 유치 노력 등을 통해 광진구를 영화와 대중문화의 중심지로 인식시키는 것이 중요하다. 혜화동이 연극의 중심지로 유명한 것처럼, 광진구가 문화산업을 전폭적으로 육성한다면 지역 이미지 제고와 경쟁력 있는 문화산업으로의 재탄생이 가능할 것이다.

한편, IT 산업 육성을 위해서는 인재양성과 더불어 지속적인 관심이 필수적이다. 광진구는 학구열이 높으며 건국대학교, 세종대학교 등 다수의 교육기관이 소재하고 있다. 대학의 이공계 지원 확대, 공모전 및 박람회 개최 등을 통해 IT 산업 인재를 적극 육성할 필요가 있다. 건국대학교의 영화학과 명성이 영화 관련 산업 발전을 이끈 것처럼, 젊은 이공계 인재에 대한 지원이 IT 산업 발전으로 이어질 수 있을 것이다. 아울러 주민 IT 특강 프로그램 등을 활발히 운영하여 주민들의 지속적인 IT 기술 관심을 유도해야 한다.

**7. 참고문헌**

김종, & 조성식. (2007). 지방자치단체 체험형 스포츠 테마파크의 지역사회 발전 및 지역경제 기여방안. 체육과학연구, 18(4), 185-196.

강현철, & 최재용. (2019). 문화예술기반 도시재생사업의 쟁점과 함의-2019 년 매니페스토 선정 지방자치단체를 중심으로. GRI 연구논총, 21(4), 107-138.

이완복, & 정진도. (2009). 지방자치단체 이미지제고를 위한 문화마케팅 전략에 관한 연구. 인문콘텐츠, (15), 29-53.

심익섭, & 정동일. (2009). 지방자치단체의 세계화: 서울특별시 중구를 중심으로. 사회과학연구, 16(1), 39-67.

서라벌, & 성욱준. (2019). 지방자치단체의 공공서비스 질이 이주 의향에 미친 영향 연구: 경기도 삶의 질 조사를 활용하여: 경기도 삶의 질 조사를 활용하여. 한국사회와 행정연구, 29(4), 199-223.

최명진, 동재욱, & 이화룡. (2020). 지방자치단체 공공시설의 효율적 관리방안에 관한 연구-한국과 일본의 정책을 중심으로. 교육시설 논문지, 27(6), 3-12.

김근호. (2013). 기초지방자치단체 사회복지예산 결정요인에 관한 연구. 보건과 복지, 15, 7-29.

오자왕, 방신웅, & 인상우. (2010). 지방자치단체의 공공체육시설 활성화방안: 성남시 종합운동장 사례를 중심으로. 한국스포츠산업경영학회지, 15(1), 75-90.