***2023학년도 2학기 텍스트 정보분석 팀 프로젝트***

**텍스트 마이닝을 활용한 대규모 전세사기 피해사례 분석 :**

**네이버 지식iN을 중심으로**

이진우[[1]](#footnote-1) • 이하은[[2]](#footnote-2) • 최희석[[3]](#footnote-3)

**요 약**

본 프로젝트는 네이버 지식iN을 활용해 '전세사기 피해' 키워드가 포함된 질문 데이터를 추출하고 분석하여 '전세사기 피해'와 관련한 의미 있는 정보를 추출하고, 더 나은 피해 예방, 구제와 관련한 법적, 제도적 장치를 마련하기 위한 기초자료를 만드는데 목적을 둔다. 이를 위해 2019년 1월 1일부터 2023년 10월 31일까지 총 4년 10개월 간의 네이버 지식IN 질문 중 '전세사기 피해'가 중심인 질문들을 수집하였다. 네이버 지식iN 의 질문은 설문조사와 달리 피해자의 보다 자연스럽고 솔직한 목소리를 들을 수 있는 수단이며 익명성이 보장된다는 장점이 있다. 수집된 3,728개의 데이터에 대해 띄어쓰기를 교정하였고 특정 품사의 단어를 추출하는 작업을 진행하였다. 그 후, 단어들의 빈도를 파악하고 N-gram 기법을 활용하여 분석하고 시각화하였다. 데이터 수집은 Python과 lxml, BeautifulSoup을 통해 크롤링하였고 띄어쓰기 교정에는 PyKoSpacing, 형태소 분석에는 Kiwi라이브러리를, 데이터 분석에는 Pandas와 Numpy를, 시각화에는 matplotlib, R언어의 ggraph패키지 그리고 Microsoft Excel을 사용했다. 해당 연구에서는 다음과 같은 결과를 도출하였다. 첫째, ‘전세 사기 피해’ 질문자는 대출을 받은 행동을 가장 많이 했으며, 돈을 못 돌려받은 상황에 처해있다. 둘째, 대출과 개인 회생에 대한 정보가 주로 질문되고 있다. 셋째, 전세 사기에 대한 정보에 쉽게 접근할 수 있는 방안이 마련되어야 한다.

**•주제어 : 전세사기, 텍스트 마이닝, N-gram**

**서 론**

과거 한국전쟁 이후 도시지역으로의 이주자가 급격히 늘어나며 주택 부족이 심각한 상황이 발생했다. 특히 1960~70년대 고도성장기에 따른 도시화로 인해 인구 증가, 지속적인 주택가격 및 제한된 정부의 주택금융 공급정책으로 서울을 중심으로 전세제도가 발전하였고, 전국적으로 확대되었다. 전세제도란 보증금을 맡기고 남의 집에 임차한 뒤 계약기간이 끝나면 보증금을 돌려받는 주택임대차 유형으로[[4]](#footnote-4), 임대인에게는 자본이득을, 임차인에게는 적은 비용으로 양질의 주거환경을 영위할 수 있다. 전세사기란 임대인이 착오에 빠진 상태에서 계약을 체결함으로써 임차인이 전세보증금의 전부 또는 일부를 편취 당하는 경우를 말한다.

연령별로 청년가구는 81.6%의 비율로 임차에 거주하고 있으며, 신혼가구의 경우 53%의 비율로 임차로 거주하고 있다[[5]](#footnote-5). 자가를 마련할 자본적 여유가 충분치 않은 청년가구나 신혼가구처럼 사회초년생들이 많이 이용함을 확인할 수 있다. 이처럼 전세를 특정 연령대에서 전세를 이용하는 비율이 절반을 넘고, 중요한 기능을 함에도 불구하고 전세와 관련된 문제 및 분쟁이 계속해서 발생하고 있다.

따라서 본 프로젝트는 다음과 같은 목적을 지닌다.

목적 1. '전세사기 피해'를 검색어로 한 지식IN 검색결과를 토대로 질문 데이터를 수집한다.

목적2. 수집한 질문 데이터를 기반으로 빈도분석, N-gram 분석을 수행하여 키워드를 파악한다.

목적 3. 키워드를 바탕으로 전세사기 피해예방, 피해구제를 위한 기초자료를 생성한다.

**선행연구**

권경선(2023)은 대규모 전세사기에 대한 공법적 규제방안에 대해 연구하였다. 해당 연구에서 대규모 전세사기 사태의 발생현황을 파악하고 원인을 분석했으며 공법적 규제에 대한 개선 방안을 제시하였다. 위 연구에서 언급한 발생현황 및 규제의 개선 방안은 다음과 같다. 첫째, 최근 3년간의 전세사기 피해자의 수는 2,966명이고 피해금액은 4,599억원이다. 둘째, 전세사기 문제가 커져감에 따라 경찰청에서 2022.07. ~ 2023.05 10개월간 특별단속한 결과 전세사기 사범 2,895명을 검거하였다. 이는 2021년 검거인원 243명에 비교할 때 약 12배 증가한 수치이다. 셋째, 부동산 컨설팅업과 관련한 규정이 전무한 현재, 해당 업종을 관리 감독할 수 있는 법제가 정비되어야 한다. 이와 같이 전세사기는 시간이 지날수록 피의자의 수가 늘어나고 그에 따라 피해액과 피해자의 수도 늘어남을 예상할 수 있다. 뿐만 아니라 전세사기 피해 예방을 위해서는 법적, 제도적 장치 개선이 시급하다는 것을 시사하고 있다.[[6]](#footnote-6)

**연구 방법**

**분석 대상**

본 프로젝트는 전세사기 피해를 입은 당사자의 자연스러운 의견을 분석하기 위하여 네이버 지식iN에 있는 질문 데이터를 활용하였다. 네이버 지식IN 사이트에 '전세사기 피해'를 검색어로 하여 해당 결과들을 크롤링하여 3,728개의 질문 데이터를 수집하고 제목, 내용, 작성일, 링크 열을 만들어서 Pandas의 DataFrame에 정리하였다. 데이터 수집 기간은 2019년 01월 01일부터 2023년 10월 31일까지로 설정하였는데 통칭 '빌라왕'이 전국적으로 출현하기 시작한 시점이 2019년이기 때문이다[[7]](#footnote-7). 월별로 수집된 데이터의 양은 그림 1과 같다.

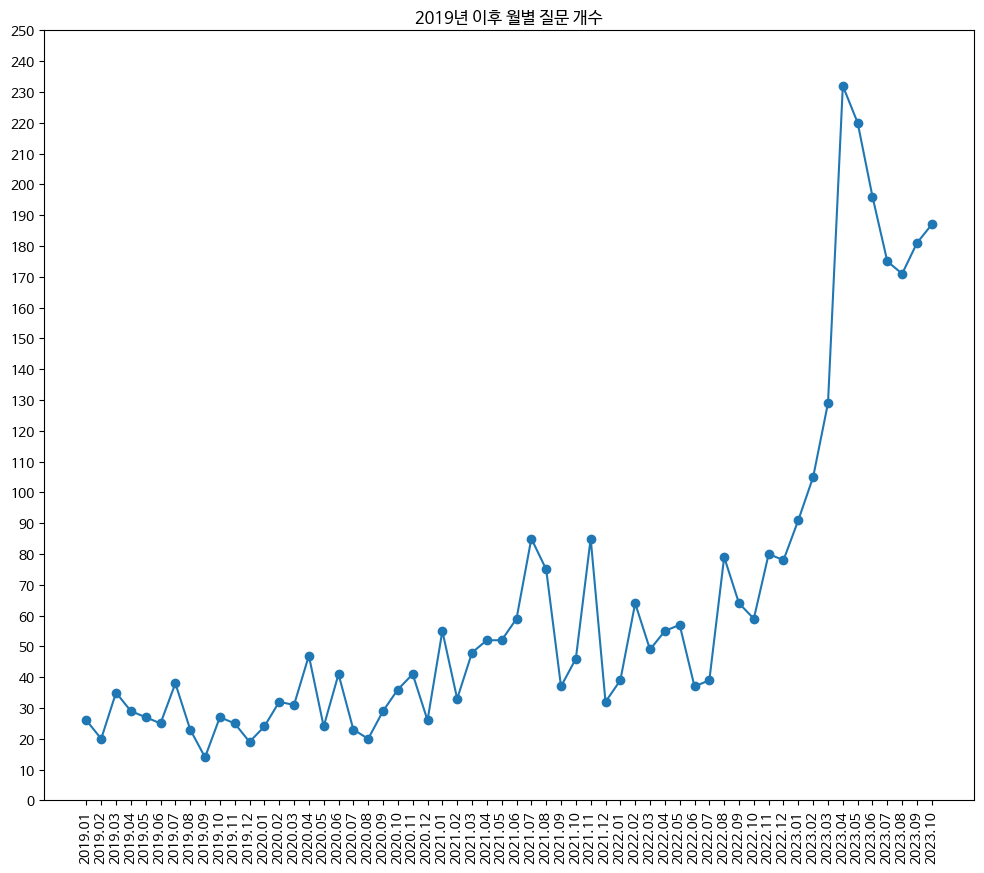


그림 1. 2019년 ~ 2023년 10월 월별 질문 수

**데이터 전처리**

수집한 데이터는 파이썬에서 사용할 수 있는 한국어 띄어쓰기 교정 도구인 PyKoSpacing을 사용하여 띄어쓰기를 교정하였다. PyKoSpacing은 대규모 말뭉치로부터 훈련된 딥러닝 모델에 기반을 두고 있으며, SNS, SMS와 같은 온라인에서 만들어진 텍스트에 높은 성능을 보인다[[8]](#footnote-8). 그 후, 파이썬에서 사용할 수 있는 한국어 형태소 분석기인 Kiwi를 사용하여 동사, 형용사 등이 포함된 용언, 일반 명사, 고유 명사 그리고 부사를 품사로 가지는 단어를 추출했다[[9]](#footnote-9). 추출한 명사 중 1음절짜리 명사는 제외했지만 '집'과 '돈'은 프로젝트 주제와 밀접한 관련을 가지는 단어이므로 제외하지 않았다. 해당 전처리 작업은 제목과 내용에 진행하였다.

**데이터 분석**.

수집한 데이터에서 연도별 대표하는 키워드가 있는지 확인해보기 위해 질문의 제목에서 키워드를 추출하여 분석했다. 분석 결과는 표1과 같다. 5번 나온 단어를 회색으로 표시했고, 3, 4번 나온 단어를 하늘색으로 표시했다. 볼드체로 표시한 단어들은 자료를 수집할 때 사용한 단어인 ‘전세’, ‘사기’, ‘피해’를 제외하고 5번 나온 단어들이다. 분석해 본 결과 5년동안 대체적으로 비슷한 단어가 반복해서 나타났고, 2019년에는 ‘아파트’, 2020년부터 2022년에는 ‘빌라’라는 키워드가 나타났다. 이에 연도별로 전세사기 유형에 차이점이 있는지 확인해보았는데 해당 시기 뉴스들을 검토해 본 결과 특이한 점을 발견하지 못했고, 연도별로 대표하는 키워드가 있지 않다고 판단했다.

텍스트, 스크린샷, 번호, 폰트이(가) 표시된 사진

자동 생성된 설명

표 1. 연도별 단어 빈도수 분석

본 프로젝트의 목표는 자주 사용되는 키워드의 관계를 분석하여 질문을 분석하고 재조합하여 해석하는 것이기 때문에 데이터 분석은 N-gram으로 진행하였다. bi-gram을 사용했을 경우 큰 의미를 형성할 수 없었기 때문에 tri-gram으로 분석을 수행하였다.

tri-gram은 키워드 중심 분석과 서술어 중심 분석으로 나눠 진행하였다. 먼저 주요 키워드를 중심으로 형성한 tri-gram을 알아보겠다. 다음 표2는 tri-gram 결과 중 상위 15개를 나타낸 것이다. 키워드 3개를 순서대로 연결하면 문장 또는 구절이 된다. 이어진 문장 또는 구절만으로 바로 내용을 이해할 수 있는데, 이는 trig-ram이 인간이 말하는 어순대로 연결되었기 때문이다. ‘개인 회생 신청’, ‘주택 담보 대출’, ‘전세 자금 대출’ 이 각각 18.7%, 12.6%, 12.3%로 상위 순위를 차지했다.

텍스트, 번호, 폰트, 낱말맞추기 퍼즐이(가) 표시된 사진

자동 생성된 설명

표 2키워드 중심 tri-gram

키워드 중심으로 분석한 tri-gram을 보면 세 개의 키워드를 그대로 조합했을 때 바로 그 의미를 파악할 수 있다는 것을 알 수 있다. 이는 tri-gram이 어순으로 분석되었기 때문이다. 한글은 주어, 목적어, 서술어 형식으로 문장이 구성된다. 따라서 세 번째 키워드 자리에 동사를 고정하여 서술어 중심 tri-gram을 다시 생성해보았다. 다음 표3은 tri-gram 결과 중 상위 15개를 나타낸 것이다. ‘전세 사기 당하다’, ‘전세 대출 받다’, ‘집 경매 넘어가다’가 순서대로 22%, 12%, 10%를 차지했다.

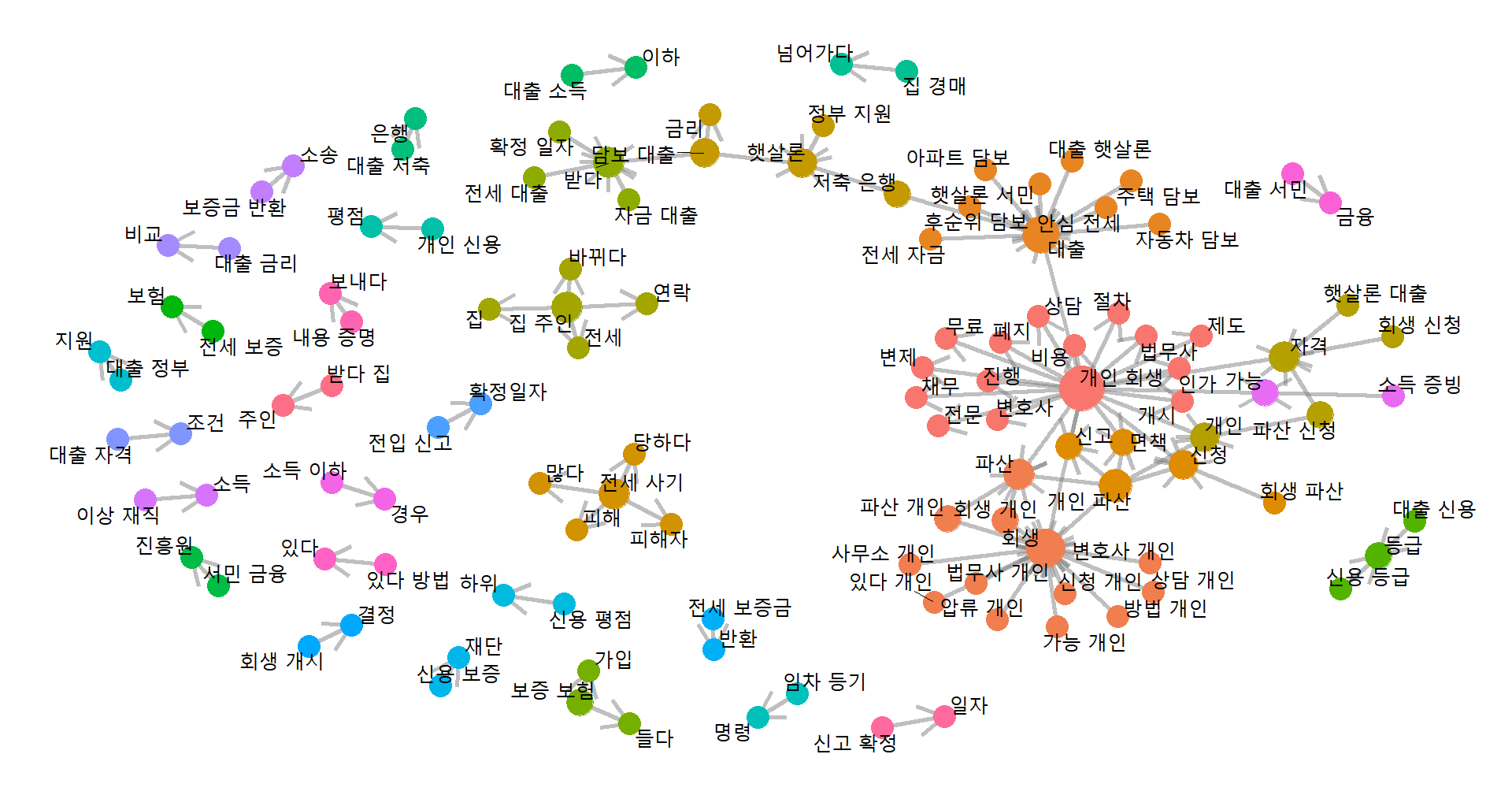
****

그림 2. 키워드 중심 tri-gram word network

그림 2는 키워드 1과 2를 합쳐 한 문자열로 만들고 해당 열을 키워드 3과 연결하여 만든 워드 네트워크이다. 노드의 크기는 연결중심성을 나타내고 노드의 색깔을 커뮤니티를 나타낸다. 즉, 노드가 크면 클수록 다른 노드들과 밀접하게 연결되어 있다는 의미이고 같은 색깔의 노드는 빈번하게 연결되어 있다는 의미이다. 해당 그림에서 눈에 띄게 확인할 수 있는 키워드는 ‘개인 회생’이다. 이 그림을 확인하고 전세사기 피해자들이 개인 회생을 진행중이거나 개인 회생을 준비를 위한 상담 또는 답변을 원한다는 사실을 원 데이터에서 다시 한번 확인 할 수 있었다.

텍스트, 번호, 폰트, 스크린샷이(가) 표시된 사진

자동 생성된 설명

표 3. 서술어 중심 tro-gram

서술어 중심의 tri-gram을 생성하면서 동사로 ‘받다’가 가장 높은 빈도로 사용되었다는 점에 주목하여, ‘받다’를 두번째 키워드 자리에 고정하여 bi-gram을 다시 생성해보았다. 그 결과 ‘대출 받다’와 ‘못 받다’의 조합이 높은 빈도로 도출되었다. 이를 해석하자면, ‘전세 사기 피해’에 대하여 질문한 사람 중 가장 많이 한 행동이 ‘대출을 받다’이고, 질문자가 처한 상황이 ‘돈을 못 돌려 받다’인 것으로 해석할 수 있다.

키워드 중심의 tri-gram과 서술어 중심의 tri-gram은 유의하게 다른 결과가 도출되었다. 키워드 중심의 tri-gram은 질문자가 해당 질문을 통해 얻고 싶어하는 정보 중심으로 결과가 도출되었다. 반면 서술어 중심 tri-gram은 질문자가 마주하고 있는 현재 문제 상황 중심으로 결과가 도출되었다. 또한 키워드 중심 tir-gram 기법은 ‘개인 회생’과 ‘대출’에 관한 질문이 주를 이루었다. 서술어 중심 tri-gram에서 ‘개인 회생’과 ‘대출’에 대한 내용이 포함되어 있었지만 그 이외 다양한 상황이 함께 도출되었다.

텍스트, 번호, 스크린샷, 폰트이(가) 표시된 사진

자동 생성된 설명

표 3 <키워드 중심과 서술어 중심 비교>

**예방법 또는 해결책 제시**

예방법과 해결책을 제시하기 위해서 전세 사기 피해자들의 현재 상황을 파악하는 것도 중요하지만 질문자가 직접적으로 물어보는 것을 제시하는 것이 더 적절하다고 생각했기 때문에 서술어 중심 tri-gram이 아닌 키워드 중심 tri-gram을 활용하여 방안을 제시하였다. 질문자가 필요한 정보에 바로 접근할 수 있도록 Linktree를 이용하였다. 키워드 분석에서 빈도수가 높았던 순서대로 정렬하였으며, 키워드 분석을 토대로 해석한 결과에 맞는 전세 사기 예방법 또는 전세 사기 피해를 입은 후 구제 방법을 확인할 수 있도록 바로가기를 만들었다.[[10]](#footnote-10)

**결론**

본 프로젝트는 네이버 지식iN을 활용해 '전세사기 피해' 키워드가 포함된 질문 데이터에서 의미 있는 정보를 추출하여 전세 사기 피해자들의 상황을 파악하고 더 나은 피해 예방, 구제와 관련한 법적, 제도적 장치를 마련하기 위한 기초자료를 만드는데 목적으로 수행되었다. N-gram 기법을 활용하여 도출한 주요 시사점은 다음과 같다.

첫째, ‘전세 사기 피해’ 질문자는 대출을 받은 행동을 가장 많이 했으며, 돈을 못 돌려받은 상황에 처해있다. 이를 연결하여 상황을 파악해본다면 두 가지로 해석해볼 수 있다. 첫번째는 질문자가 돈을 못 돌려받았기 때문에 대출을 받는 상황이다. 두번째는 질문자가 대출을 받아 전세 계약을 했는데 사기를 당해 전세 계약금을 못 돌려받는 상황이다. 두 상황 모두 임대인에게 임차권등기명령 신청을 하는 등 돈을 되돌려 받을 수 있도록 해야 하며, 현재 상황에 맞는 대출을 받을 수 있는 대출 비교 정보가 필요하다. 더불어 사전에 보증보험을 들어 돈을 못 돌려받는 상황에 피해를 줄여야 한다.

둘째, 대출과 개인 회생에 대한 정보가 주로 질문되고 있다. 전세 계약은 큰 단위의 금전 거래이므로 전세 사기를 당하는 경우 금전적으로 입는 피해가 크다. 또한 대출은 받은 금액을 사기당하는 경우, 상환을 할 능력이 없어 개인 회생까지 신청하기에 이를 수 있다. 전세 계약을 하는 국민이 사기의 위험을 걱정하지 않고 거래를 할 수 있도록 법적 영역에서 이를 예방하는 방안을 고려해야 한다.

셋째, 전세 사기에 대한 정보에 쉽게 접근할 수 있는 방안이 마련되어야 한다. 전세 사기 예방법과 해결 방안이 이미 법으로 규정되어 있는 경우가 있지만, 법적 용어가 많아 내용이 어렵고 필요한 정보에 접근하기 또한 어렵다. 이를 해결하기 위해 일반인이 이해하기 쉽도록 전세 사기에 대한 구체적이고 종합적인 정보가 제공되는 사이트가 필요해 보인다.

1. 문헌정보학과 [↑](#footnote-ref-1)
2. 문헌정보학과 [↑](#footnote-ref-2)
3. 프랑스•아프리카학과 [↑](#footnote-ref-3)
4. 한국민족문화대백과사전 [↑](#footnote-ref-4)
5. 국토교통부, 2021년도 주거실태조사 결과 [↑](#footnote-ref-5)
6. 대규모 전세사기(빌라왕)에 대한 공법적 규제방안 [↑](#footnote-ref-6)
7. https://premium.sbs.co.kr/article/MvIsIIRo5r [↑](#footnote-ref-7)
8. https://github.com/haven-jeon/PyKoSpacing [↑](#footnote-ref-8)
9. https://github.com/bab2min/kiwipiepy [↑](#footnote-ref-9)
10. https://linktr.ee/text\_mining\_project [↑](#footnote-ref-10)