pISSN: 1226-7147, eISSN: 2383-9171
https://doi.org/10.17208/jkpa.2025.06.60.3.84

Journal of Korea Planning Association Vol.60, No.3 (2025)
대한국토·도시계획학회지 「국토계획」 제60권 제3호 pp.84-100

Check for updates

구조적 토픽 모델을 이용한 상업용 부동산 통계의 제공자와 사용자 간 문제의식 비교

Comparative Analysis of Issues Perceived by Providers and Users of Commercial Real Estate Statistics Using Structural Topic Model

민성훈*
Min, Seonghun

Abstract

Commercial real estate statistics remain relatively underdeveloped compared with land or housing statistics, limiting their ability to enhance market efficiency. To address this issue, the present study conducted ten rounds of interviews with 33 experts and analyzed the collected data using the Structural Topic Model. The results were as follows. First, statistics providers emphasized the need to reduce the burden of surveys and increase data reliability by sharing data with relevant institutions rather than relying solely on direct surveys. Meanwhile, statistics users highlighted the need to establish statistical areas, such as office regions and retail districts, with greater precision. Both providers and users shared concerns about the sample design and the need to consider a wider range of factors in the design. Second, the Structural Topic Model proved to be useful in analyzing expert interviews, as metadata provided insights into which topics to prioritize to address real-world issues. Although this study confirms the utility of the Structural Topic Model, it does not suggest that the model alone is sufficient for a comprehensive interpretation of interviews. Instead, it recommends the model as a tool that can help researchers interpret interviews more objectively.

주제어 상업용 부동산, 부동산 통계, 문제의식, 구조적 토픽 모델, 네트워크 분석
Keywords Commercial Real Estate, Real Estate Statistics, Problem Perception, Structural Topic Model, Network Analysis

Ⅰ. 서 론

토픽 모델(Topic Model)은 텍스트 분석(Text Analysis) 또는 텍스트 마이닝(Text Mining)의 한 종류로서, 어떤 말뭉치(Corpus)의 단어와 문장을 분석하여 그 속에 포함된 의미 있는 주제를 추출하는 기법이다. 여기서 단어와 문장의 분석이란 어떤 단어가 자주 출현하는지, 어떤 단어들이 함께 출현하는지, 그러한 단어들이 얼마나 많은 문장에서 출현하는지 등을 확률분포를 이용하여 파악하는 것을 말한다. 토픽 모델의 결과는 의미 있는 단어의 조합을 연구자가 원하는 개수만큼 추출하는 형태로 제시된다. 토픽 모델의 가장 큰 특징은 계량경제 모형에서 사용되는 것과 같은 확률분포를 이용한다는 점이다. 이는 단순히 단어의 빈도수를 세는 것과 다른 의미를 가진다. 토픽 모델은 회귀모형과 유사한 형태를 가지므로 여러 메타변수에 따른 토픽의 차이도 분석할 수 있기 때문이다. 토픽 모델은 텍스트 분석 중에서는 계량경제 모형에 가까운 편에 속한다.

토픽 모델은 단순히 원하는 개수만큼 토픽을 도출하는 LDA(Latent Dirichlet Allocation)에서부터, 각 토픽별로 텍스트의 작성자나 작성시기와 같은 메타데이터(Meta Data)에 따른 차이도 파악하는 DMR(Dirichlet Multinomial Regression), 메타

* Professor, Department of Urban Planning and Real Estate, The University of Suwon (smin@suwon.ac.kr)

데이터에 따른 차이의 유의성까지 검정하는 구조적 토픽 모델(Structural Topic Model, STM)로 꾸준히 발전하고 있다.

이 중 가장 발전된 기법인 STM은 2013년 Roberts, Stewart, Tingley, and Airoldi에 의해 처음 소개되었으며, 이후 여러 학문 분야에서 활용되고 있다. 신문 기사에서 특정 이슈가 정치적 시기에 따라 어떻게 다르게 다뤄지는지 분석하거나, 소셜 미디어 작성자의 지역이나 성별에 따라 주제가 달라지는 양상을 분석하거나, 설문 응답자의 인구 통계적 특성에 따라 응답 내용이 어떻게 달라지는지 분석하는 것이 대표적인 사례다.

STM은 부동산 분야에서도 유용하게 활용될 수 있다. 언론 기사, 소셜 미디어, 그것들에 반응한 댓글, 학술논문, 설문조사, 전문가 인터뷰 등 중요한 의미를 가지는 텍스트 데이터가 부동산 분야에도 풍부하게 존재하기 때문이다. 하지만 부동산 관련 학술연구에서는 STM을 비롯한 텍스트 분석을 활용한 사례를 찾아보기 어렵다. 이는 인문학, 사회학, 교육학과 같이 텍스트 분석이 널리 활용되는 분야와 대조적인 모습이다. 결국 부동산학 분야에서 텍스트 데이터를 정리하고 해석하는 일은 현재로서는 주로 연구자의 주관과 전문성에 의존하여 이루어지고 있는 것이다.

본 연구는 부동산 통계 특히 토지나 주택에 비해 빈약한 상업용 부동산 통계가 가지는 문제점을 추출하기 위해 다수의 전문가와 진행한 인터뷰를 STM을 활용하여 분석한다. 또한 인터뷰 참여자의 포지션(사용자 또는 제공자)과 직업군에 따른 문제의식의 차이를 STM을 활용하여 검정함으로써 입장 차이에 따라 쟁점이 되는 토픽이 무엇인지도 파악한다. 이러한 분석은 내용적, 방법적 측면에서 다음과 같은 두 가지 의의를 가진다.

첫째, 부동산 통계는 국가경제와 국민생활에 직접 연관되어 있어서 사용자의 관심과 요구가 큰 반면 정확성, 신뢰성, 적시성에 대한 비판에 늘 직면해 있다. 부동산은 개별성이 강해서 조사의 범위가 넓고, 조사에 많은 시간과 비용이 투입되며, 조사내용을 토대로 안정적이고 신뢰할 만한 통계를 작성하는 것 또한 쉽지 않기 때문이다. 이러한 문제를 해결하기 위해서는 상업용 부동산 통계와 관련된 다양한 집단의 의견을 폭넓게 수렴하여 해결과제를 객관적으로 도출할 뿐 아니라, 집단 간 문제의식의 차이도 엄밀하게 파악하여 해결방안의 우선순위와 집단 간 균형을 도모해야 한다. 본 연구가 수행하는 토픽 추출과 포지션 및 직업군 간 차이 검정은 이러한 접근에 단초를 제공할 것이다.

둘째, 부동산 분야에서는 전문가 인터뷰가 질적 연구의 한 방법으로 빈번하게 사용되고 있다. 그런데, 인터뷰는 설문과 달리 질문과 답변이 정형적이지 않아서 요약과 해석에 많은 노력이 필요하고, 연구자의 관심사나 선입관에 의해 해석이 달라질 수도 있다. 텍스트 분석은 이러한 문제를 보완하는 데 유용한 수단이 될 수 있을 것이다. 본 연구는 STM이 전문가 인터뷰 해석에 유용하게 활용될 수 있는지 실제 사례를 통해 확인한다.

본 연구는 다음의 순서로 진행된다. 2장에서는 국내 상업용 부동산 통계의 현황과 관련 선행연구를 살펴본다. 선행연구는 상업용 부동산 통계의 전반적인 문제점과 해결방안을 다룬 사례와 본 연구의 분석모형인 STM의 발전과정과 그것을 국내에 적용한 사례 두 가지로 나누어 살펴본다. 3장에서는 분석모형과 분석자료를 설명한다. 분석모형에 대해서는 STM의 함수구성과 본 연구의 투입변수를 소개하고, 분석자료에 대해서는 전문가 인터뷰 개요와 인터뷰 텍스트의 전처리 과정을 서술한다. 4장에서는 분석결과를 제시한다. 분석은 인터뷰 텍스트로부터 주요 토픽을 도출하고, 인터뷰 참여자의 포지션과 직업군에 따라 유의한 차이가 있는지 검정하는 순서로 진행된다. 끝으로 5장에서는 분석결과를 요약하고 시사점을 도출한다.

II. 통계현황 및 선행연구

1. 상업용 부동산 통계현황

국내에서는 상업용 부동산에 대한 통계가 매우 제한적으로 제공되고 있다. 오피스, 리테일, 인더스트리얼, 호스피탈리티 등 상업용 부동산의 하위 섹터 중에서 시장정보가 정기적으로 제공되는 것은 오피스와 리테일 정도이며, 그중 리테일의 경우 체계적인 통계라기보다는 시장동향을 보여주는 리포트가 다수를 차지하고 있다. 결과적으로 국내에서 장기간의 시계열을 가지는 상업용 부동산 통계는 공공부문이 제공하는 상업용 부동산 임대동향조사와 민간부문이 제공하는 오피스 마켓리포트 두 가지가 전부라고 해도 과언이 아니다.

한국부동산원이 수행하는 상업용 부동산 임대동향조사는 전국을 대상으로 오피스와 리테일 두 시장을 조사하고 있다. 조사의 내용은 임대현황뿐 아니라 평가금액, 투자수익률과 같은 자본시장 정보도 포함하고 있다. 조사는 2002년 시작되었으며, 2014년 국가통계 승인을 받았다.

오피스의 경우 표본을 규모와 용도 두 가지 기준으로 선정한다. 규모는 6층 이상일 것, 용도는 건축물대장상 주용도가 업무시설일 것을 충족해야 한다. 리테일의 경우 건축물대장상 주용도가 제1종 및 제2종 근린생활시설, 판매시설, 운동시설, 위락시설인 것을 대상으로 한다. 그리고 규모와 소유형태에 따라 중대형 상가, 소규모 상가, 집합상가 세 가지로 하위 유형을 구분한다. 구체적인 구분 기준은 〈Table 1〉과 같다.

조사권역은 총 328개 설정되어 있으며, 상업용 부동산의 유형에 따라 적절한 것을 선별하여 사용한다. 즉 328개 권역 중에는 네 가지 유형에 모두 사용되는 것도 있고, 일부 유형에만 사용되는 것도 있다. 조사표본은 오피스의 경우 824동(서울 404동)이 초기부터 현재까지 계속 유지된 반면, 상가의 경우 지속적으로 확대되었다. 초기에는 중대형 상가만 조사하였으나, 2014년 집합상가, 2015년 소규모 상가가 추가된 것이다. 유형별 권역 및 표본의

Table 1. Classification criteria for samples in commercial real estate rental trend surveys

Type		Classification criteria (building)		
		Main Use	Size	Own
Office		Business facilities	6 floors or higher	Single
Retail	Mid-large	Neighborhood living facilities, Sales Facilities, Sports facilities, Amusement facilities	3 floors or higher or GFA 330 m^2 or larger	Single
	Small		2 floors or lower and GFA 330 m^2 or smaller	
	Strata		-	Multiple

배분 현황은 〈Table 2〉와 같다. 조사결과는 한국부동산원이 운영하는 통계정보시스템인 R-ONE을 통해 일반에게 제공된다. R-ONE은 조사결과를 임대정보, 수익률정보, 권리금정보 세 가지로 구분하고 있다.

오피스 마켓리포트는 여러 민간기업이 제공하고 있다. 대표적인 업체로는 KB국민은행, 젠스타메이트, 에스원, 교보리얼코, 한화호텔&리조트, 신영에셋, 알스퀘어 등이 있으며, CBRE, Savills, JLL, C&W 등 외국계 업체도 리포트를 발간하고 있다. 이중 C&W는 리테일에 대해서도 마켓리포트를 발간하고 있다.

민간기업의 오피스 표본선정과 등급부여 기준은 대체로 유사하며, 주로 3,300m^2 이상의 중대형 빌딩을 대상으로 한다. 하지만 KB 국민은행과 같은 일부 업체는 소형빌딩에 대한 통계도 다루고 있다. 지역적으로는 대부분 서울로 권역을 한정하고 있다. CBD는 대체로 중구와 종로구로 구성되나, 일부 업체는 용산구 동자동을 포함시키거나 중구 중림동을 제외하는 등 차이를 두고 있다. GBD는 대체로 강남구와 서초구로 구성되나, 한화호텔&리조트와 신영에셋은 잠실도 포함시키고 있다. YBD는 여의도동만으로 구성하는 업체와 마포구 공덕역 일대까지 포함시키는 업체가 양립하고 있다. 이외에 서울 기타지역, 분당, 6대 광역시의 포함 여부는 업체마다 차이가 있다.

업체마다 발표하는 통계지표에는 적지 않은 차이가 있다. 임대료, 공실률, 거래가격 등 핵심지표는 모두 발표하고 있지만, 가격지수나 자본환원율을 발표하는 업체는 KB국민은행, 젠스타메이트, 알스퀘어 정도로 매우 드물며, 투자수익률을 발표하는 기관은 하나도 없다.

Table 2. Districts and samples in commercial real estate rental trend surveys

Type		Number of districts		Number of samples	
		Seoul	Korea	Seoul	Korea
Office		29	52	404	824
Retail	Mid-large	61	248	1,445	5,761
	Small	54	232	911	5,526
	Strata	40	222	5,592	29,500

한편 리테일의 경우 소상공인시장진흥공단, 서울신용보증재단 등 소상공인 또는 자영업자를 지원하는 공공기관에 의해 매우 풍부한 시장정보가 제공되고 있다. 또한, 이들과 유사한 서비스를 제공하는 민간기업도 빠르게 성장하고 있는데, 오픈업, 네모, 마이프차 등이 창업자를 대상으로 다양한 정보서비스를 제공하고 있다. 다만, 이러한 리테일 시장정보는 임대료와 같은 부동산 시장정보보다는 매출액과 같은 소매업 시장정보를 다루고 있다.

상업용 부동산 임대동향조사와 오피스 마켓리포트는 특정한 표본을 선정한 후 직접 조사하는 방식으로 데이터를 수집하고 있다. 이에 반해 리테일 관련 통계를 제공하는 공공기관이나 민간기업은 통신회사, 카드회사, 국세청 등 리테일 시장 소비자와 사업자의 활동내역을 축적하고 있는 타 기관의 데이터를 입수하여 통계를 작성하고 있다. 두 가지 방식은 각각의 장단점을 가진다. 직접 조사의 경우 표본이 일정하여 일관되고 안정된 통계를 생산할 수 있는 반면, 비용이 많이 들고 새로운 시장변화를 민감하기 포착하기 어렵다는 단점을 가지고 있다.

2. 상업용 부동산 통계 선행연구

상업용 부동산이 기관투자자의 포트폴리오에서 적지 않은 비중을 차지하게 되면서, 이에 대한 연구도 다양하게 이루어지고 있다. 이러한 연구는 앞에서 살펴본 통계들을 주로 활용하는데, 통계가 제한적이다 보니 연구범위 또한 오피스와 서울에 집중되어 있다. 그럼에도 연구의 기초가 되는 통계 자체에 대해 현황을 분석하고, 문제점과 개선 방안을 도출한 연구는 흔하지 않다.

상업용 부동산 통계의 현황을 전반적으로 다룬 사례는 이태리 외(2017)에서 찾을 수 있다. 그들은 통계의 내용뿐 아니라 통계 작성을 위한 정보체계를 전반적으로 검토하기 위해 미국 NCREIF (National Council of Real Estate Investment Fiduciaries)와 싱가포르 URA(Urban Redevelopment Authority) 사례를 분석하였다. 그 결과 임대 및 매매 관련 데이터 수집 표준화, 체계적인 모니터링 시스템 구축, 정보에 대한 접근성 강화의 필요성을 역설하였다. 방보람 외(2017)는 AHP(Analytic Hierarchy Process)를 통해 상업용 부동산 정보의 우선순위를 평가하였다. 투자자와 정책 입안자를 대상으로 한 분석결과 상업용 부동산의 경우 주거용 부동산에 비해 거래량, 매매가, 임대료, 접근성 등이 중요하게 취급된다는 것을 밝혔다.

이들 연구는 문헌조사와 사례조사를 주된 연구방법으로 사용하고 있으며, 전문가의 의견을 청취하기 위해 FGI(Focus Group Interview)와 AHP를 보조적인 연구방법으로 채택하였다. FGI

는 특정한 이슈를 심도 있게 파악하는 데 유용한 반면 인터뷰 내용의 해석에 연구자의 주관이나 선입관이 개입될 여지가 있다는 단점을 가진다. 또한 AHP는 여러 변수 간 우선순위를 계량적으로 측정하는데 유용한 반면 연구자가 사전적으로 정한 질문과 항목으로 의견이 제한된다는 단점을 가진다. 본 연구는 이러한 방법들과 함께 활용하여 앞에서 지적한 문제를 최소화할 수 있는 수단으로 STM을 적용한다는 점에서 선행연구와 차별성을 가진다.

3. 토픽 모델 선행연구

STM의 상위 개념인 토픽 모델에 관한 선행연구부터 하나씩 살펴보면 다음과 같다. 토픽 모델의 가장 고전적인 방법은 Blei et al.(2003)이 제시한 LDA(Latent Dirichlet Allocation)다. LDA는 하나의 텍스트가 여러 토픽의 혼합물이고, 각 토픽은 특정 단어의 분포로 표현된다고 가정한다. 그리고 단어-텍스트 행렬을 기반으로 토픽을 학습하여 각 텍스트가 어느 정도의 비율로 각 토픽을 포함하는지, 각 토픽이 어떤 단어들로 구성되는지 추론한다. 이 과정에서 디리클레(Dirichlet) 분포를 사용하여 텍스트별 토픽의 분포와 토픽별 단어의 분포를 확률적으로 추정한다.

LDA의 발전된 형태인 DMR(Dirichlet Multinomial Regression)은 회귀모형의 형태를 가진다. 즉 저자, 연도, 범주와 같은 텍스트의 메타데이터를 함께 고려하여 각 텍스트의 토픽 분포를 예측한다. LDA가 각 텍스트의 토픽분포가 일정한 디리클레 분포를 따른다고 가정하는 반면, DMR은 토픽분포의 초매개변수(Hyperparameter)가 텍스트의 메타데이터에 의해 달라지도록 설정하여 더 정교한 예측을 한다. DMR은 Mimno and McCallum(2008)에 의해 처음 제시되었다.

STM은 Roberts et al.(2013)에 의해 제시되었다. 그들은 기존 LDA나 DMR과 달리, 토픽빈도(Topic Prevalence)와 토픽내용(Topic Content)이라는 두 가지 요소를 도입하여 토픽이 메타데이터로부터 얼마나 유의하게 영향을 받는지 분석했다. Roberts et al.(2014)은 STM을 개방형 설문조사 데이터에 적용하여 유용성을 입증했다. 그리고, Roberts et al.(2016)은 텍스트 빅데이터에서부터 경향(Mode)을 추출하는 데 STM을 활용하였다. 그 과정에서 STM의 학습 방법을 더욱 효율화하였다.

국내에서는 문안나·이신행(2020)이 사회서비스원 정책 보도의 프레임 변화를 분석하는 데 STM을 활용했다. 이들은 선행연구를 통해 주요 프레임을 도출한 후 2010년 9월 1일부터 2019년 12월 10일까지 사회서비스원 정책 관련 기사를 분석한 결과 언론사 성향에 따라 특정 주제에 대한 보도의 양상이 다르다는 것을 발견했다.

조성배·하성호(2020)는 공공데이터에 대한 수요를 파악하는 데 STM을 활용했다. 이들은 2017년부터 2019년까지 3년간 공공데이터포털(data.go.kr)에 접수된 공공데이터 제공신청 12,814건을 분석하였다. 그 결과 시기, 유형, 자료형태, 제공방식 등 여러 요소별로 데이터 수요에 대한 차이가 있다는 것을 발견하였다.

강현희·백영민(2022)은 사회적기업에 대한 사회적 인식 형성 과정에서 언론 보도가 가지는 영향을 분석하기 위해 STM을 활용했다. 이들은 1997년부터 2021년까지 25년간 종합일간지와 경제지에 보도된 사회적기업 관련 기사를 분석하여 공공 조직이 등장한 기사에는 생태계 조성과 공공 지원 중심 담론이, 민간 조직이 언급된 기사에는 민간 지원 중심 담론이 주로 나타난 것을 발견하였다.

지금까지 살펴본 바와 같이 STM은 LDA와 DMR이 가지는 유의성 검정의 부재라는 문제를 해결하면서 다양한 분야로 확산되고 있으며, 국내에서도 정책, 언론 등의 분야에서 주요하게 논의되는 토픽이 무엇이고, 그것이 시기나 주체에 따라 어떤 차이를 보이는지 밝히는 데 활용되고 있다. 본 연구는 부동산 분야 특히 상업용 부동산 통계와 관련된 전문가 문제의식을 분석하는 데 STM을 활용한다는 점에서 선행연구와 차별성을 가진다.

Ⅲ. 분석모형 및 분석자료

1. 전문가 인터뷰

본 연구는 총 33인의 전문가와 인터뷰를 진행하였다. 상업용 부동산 통계 사용자는 학계와 업계로 구분하고, 제공자는 공공부문과 민간부문으로 구분하였다. 여기서 학계 사용자는 교육 및 연구 분야 종사자 중 상업용 부동산 통계를 활용하여 연구를 수행한 경험이 있는 자, 업계 사용자는 금융투자 및 컨설팅 분야 종사자 중 상업용 부동산에 대한 실무 경험이 있는 자, 공공부문 제공자는 국토교통부, 중소벤처기업부, 지방자치단체 산하기관의 오피스와 리테일 관련 통계 작성자, 민간부문 제공자는 부동산 정보제공업체의 마켓리포트 작성자로 각각 구성하였다. 결과적으로 모든 인터뷰 참여자는 국내외에서 발표되고 있는 상업용 부동산 통계에 대해 충분한 지식과 경험을 가지고 있는 전문가라고 할 수 있다.

인터뷰 참여자는 전문성 확보를 위해 10~20년 경력을 보유한 전문가로 섭외하였으며, 공공통계 종사자 6인, 민간통계 종사자 12인, 교육연구 종사자 6인, 금융투자 종사자 9인으로 구성되었다. 이들의 평균경력과 인터뷰에서 발언한 문장 수는 〈Table 3〉과 같다. 1인당 평균 문장 수는 134개이며, 제공자와 사용자가 각각 133개와 135개로 거의 동일하였다. 직업군에 따라서는 공공부문 제공자가 165개로 가장 많고, 민간부문 제공자가 117개로 가장 적었으나, 그 차이가 크지는 않았다. 포지션별, 직업군별로 평균경력과 한 사람당 발언 문장 수가 고르게 분포되어 인터뷰가 균형 있게 진행된 것을 확인할 수 있다.

인터뷰는 2024년 7~8월 중 총 10회에 걸쳐 각 2~3시간 동안

Table 3. Summary of interviewee

Type	Number	Career (year)	Sentences	Sentences / Number
Provider	18	14.03	2,393	133
Public	6	15.25	987	165
Private	12	13.17	1,406	117
User	15	17.45	2,025	135
Research	6	18.90	886	148
Finance	9	16.31	1,139	127
Sum	33	15.60	4,418	134

진행하였다. 한 회에 3~4인의 전문가를 대상으로 하였으며, 대부분 참여자가 한 차례 회의에 참석하였다. 〈Table 4〉에서 보는 바와 같이 각 회의는 특정한 주제를 상정하여 진행하였다. 이 중 1회차 상업용 부동산 통계에 대한 전반적 평가, 2회차 부동산 통계 연구의 최근 발전은 학계 사용자, 3회차 해외 상업용 부동산 통계 사례, 8회차 향후 개발이 필요한 통계지표, 9회차 가격지수 개발에 대한 의견은 업계 사용자, 나머지 4, 5, 6, 7회차 표본선정과 권역설정에 대한 의견, 10회차 데이터 공유에 대한 의견은 공공부문 및 민간부문 제공자 모두를 대상으로 하였다. 그러나 회의마다 해당 주제에 대한 논의가 끝나면 지체없이 다른 주제로 넘어갔기 때문에 대부분 회의에서 전체 주제가 두루 다루어졌다. 또한 참여자의 발언을 10개 주제로 제한하지 않았으며, 참여자 간 개방적 토론도 독려하는 반구조적 인터뷰(Semi-Structured Interview) 형식을 취하였다.

Table 4. Summary of interview sessions

No	Date	Topic
1	2024.07.01. (Mon.)	Overall opinion about commercial real estate statistics
2	2024.07.04. (Thu.)	Current academic advancements in real estate statistics
3	2024.07.08. (Mon.)	Global examples of commercial real estate statistics
4	2024.07.12. (Fri.)	Sampling design of office property market
5	2024.07.15. (Mon.)	Sampling design of retail property market
6	2024.07.19. (Fri.)	District establishment of office property market
7	2024.07.23. (Tue.)	District establishment of retail property market
8	2024.07.26. (Fri)	Statistics should be developed
9	2024.07.29. (Mon.)	Opinion about development of price indices
10	2024.08.08. (Thur.)	Data sharing with other statistical institutions

인터뷰 내용은 참여자의 동의를 얻어 녹음하였다. 참여자에게 인터뷰 내용이 학술연구에 사용된다는 점과 텍스트 분석을 시행하는 연구방법까지 명확하게 설명하였다. 모든 참여자가 이러한 취지에 동의하였으나, 일부 참여자는 본인의 이름과 소속이 공개되는 것을 원하지 않았다. 따라서 본 연구는 전체 참여자의 신원을 밝히지 않는다.

2. 분석모형

본 연구의 인터뷰 분석은 크게 세 단계로 이루어진다. 첫째, 텍스트에서 키워드를 추출하고, 네트워크 분석을 시행하여 키워드의 중요도와 연관성을 살펴본다. 둘째, 문장과 단어의 연관성을 분석하여 최적의 토픽 수와 각 토픽에서 중요한 역할을 하는 단어를 도출한다. 그리고 주요 단어를 참고하여 토픽명을 정한다. 셋째, 통계의 제공자와 사용자 두 개의 값을 가지는 명목변수가 토픽의 출현에 미치는 영향의 유의성을 검정한다. 그리고, 보다 세부적으로 두 포지션을 구성하는 네 가지 직업군에 따른 차이의 유의성도 검정한다.

이러한 일련의 과정에서 분석도구로 활용되는 STM을 좀 더 자세히 설명하면 다음과 같다. STM은 LDA나 DMR과 마찬가지로 각 문서가 하나 이상의 토픽으로 구성되고, 각 토픽이 고유한 단어 분포를 가진다고 가정한다. 이를 위해 두 가지 주요 변수, 즉 토픽빈도(Topic Prevalence)와 토픽내용(Topic Content)을 도입한다.

각 문서 d의 토픽분포를 나타내는 토픽빈도는 토픽 k의 빈도가 문서의 메타데이터와 관련이 있다고 가정한다. 여기서 θ_d는 문서 d에 대한 토픽분포를 나타내는 벡터, X_d는 문서 d의 메타데이터로 구성된 벡터, β는 토픽빈도에 대한 회귀계수를 각각 의미한다. f(·)는 연결함수로서 일반적으로 로지스틱 함수가 사용된다.

$$\theta_d = f(X_d\beta) \tag{1}$$

토픽내용은 토픽을 논의하는 방식이 달라지는 것을 모형화한다. 예를 들어, 동일한 토픽이라도 시간이나 저자에 따라 특정 단어의 사용 빈도가 달라질 수 있다는 것이다. 여기서 β_k는 토픽 k의 단어분포 벡터, Z_d는 문서 d의 내용에 관한 메타데이터, κ는 토픽내용에 대한 회귀계수를 각각 의미한다. g(·)는 연결함수로서 주로 로그선형모형이 사용된다.

$$\beta_k = g(Z_d k)\ \beta_k = g(Z_d \kappa) \tag{2}$$

STM은 사전확률로서 각 문서의 토픽분포(θ_d)와 각 토픽의 단어분포(β_k)를 정의한다. 수식에서 보는 바와 같이 토픽분포(θ_d)는

메타데이터(X_d)에 따라 사전분포가 결정되고, 각 토픽의 단어분포(β_k)는 그 내용에 대한 메타데이터(Z_d)에 따라 사전분포가 결정된다. STM은 잠재변수를 추정하기 위해 변이형 추론(Variational Inference)을 사용한다. STM에서 사용하는 Semi-Collapsed Variational EM은 다음과 같은 절차로 진행된다. 먼저 변이형 E-스텝에서는 토픽빈도(θ_d)와 토픽내용(β_k)에 대한 변이형 분포를 업데이트한다. 그리고 M-스텝에서는 사전확률에 대한 매개변수 β와 κ를 최적화한다. 이 과정은 토픽빈도(θ_d)와 토픽내용(β_k)이 수렴할 때까지 반복된다.

본 연구에서는 유의성 검정을 위한 반응변수로 토픽빈도(θ_d)를 사용한다. 설명변수로는 분석의 관심대상인 제공자와 사용자 두 포지션을 나타내는 명목변수 또는 네 개의 직업군(금융투자, 교육연구, 공공통계, 민간통계)을 나타내는 명목변수 메타데이터를 투입한다. 이때 경력에 따른 인식 차이를 통제하기 위해 해당 직업군 종사연수를 함께 투입한다. 연속형 변수인 종사연수에는 스플라인 함수를 적용하여 분포를 유연화 한다. 모델의 초기화(Initialization)는 무작위성을 포함하지 않아 일관된 결과를 생산하는 스펙트럴(Spectral) 기법을 적용한다.

3. 데이터 전처리

33인의 전문가와 진행한 10회의 인터뷰 결과 방대한 텍스트 데이터가 생성되었다. 하지만, 전문가를 대상으로 특정한 주제를 논의하는 본 연구의 특성상 대체로 필요한 문장과 불필요한 문장이 명확히 구분되고, 필요한 문장의 경우 문장 구조가 완전하고 의미 전달이 명확한 편이었다.

첫 번째 단계는 녹음 데이터를 문자 데이터로 전환하고, 인사, 환담 등 불필요한 문장을 삭제한 후, 필요한 문장에서 오류를 수정하는 전사(Transcript) 작업이다. 과거에는 녹음 내용을 들으면서 일일이 문자를 입력하였으나, 최근에는 음성-텍스트 변환 프로그램이 발달하여 전사 작업이 수월해졌다. 본 연구는 이러한 프로그램을 이용하여 음성을 텍스트로 변환한 후 다시 녹음 내용을 들으면서 오류를 수정하였으며, 그 결과 총 4,418개 문장을 얻을 수 있었다.

두 번째 단계는 전체 문장을 의미를 가지는 최소 언어 단위인 형태소로 분해하는 작업이다. 형태소는 어간, 어미 등으로 구성된 단어보다 더 작은 개념으로서 더 이상 분해하면 의미를 상실하는 언어 단위를 말한다. 한국어는 형태소마다 띄어쓰기를 하는 영어와 달리 한 어절 내에 여러 개의 형태소가 포함되는 특성을 가진다. 이 작업 역시 과거에는 사람이 수행하였으나, 최근에는 형태소 분석기라 불리는 프로그램이 발달하여 수월해졌다. 본 연구는 최근 많이 사용되는 Kiwi 형태소 분석기를 사용하여 형태소 분석을 시행하되, 명사만을 추출하였다. 동사, 형용사, 부사 등은 엄밀하게 형태소 분석을 하기 어렵고, 명사만 추출해도 토픽을 도출하는 데 큰 문제가 없기 때문에 많은 연구가 명사만을 대상으로 형태소 분석을 하고 있다. 흔히 형태소 분석은 토큰화(Tokenization), 토큰화를 통해 추출된 형태소를 토큰(Token)이라고 부른다.

형태소 분석은 분석기에 내장된 사전을 이용하여 시행된다. 따라서 분석기의 사전에는 없지만, 인터뷰에서 자주 거론된 전문용어를 추가해 줄 필요가 있다. 본 연구는 사용자 사전에 보통명사 "상업용 부동산", "오피스", "리테일", "인더스트리얼", "호스피탈리티", "데이터센터", "재생에너지", "지식산업센터", "레지덴셜", "데이터", "상업용 부동산통계", "오피스통계", "리테일통계", "마켓리포트"와 고유명사 "임대동향조사", "한국부동산원", "국토교통부"를 등록하였다. 만약 이들 용어를 사용자 사전에 등록하지 않으면 "상업용 부동산"이 "상업", "용", "부동산"으로 분해되어 화자의 의도와 다른 분석결과가 도출될 수도 있다.

세 번째 단계는 형태소 분석을 통해 추출한 토큰 중 불필요한 것을 제거하는 작업이다. 일반적으로 한 음절의 형태소는 큰 의미를 가지지 않기 때문에 분석대상에서 제거하는데, 본 연구는 "층", "평", "시", "도", "군", "구", "읍", "면", "동", "리"와 같이 부동산 분야에서 중요한 형태소는 남기고 제거하였다. 또한, 문장에 일상적으로 출현하지만 토픽과의 관련성이 떨어지는 불용어도 제거해야 한다. 본 연구는 빈도 기준으로 상위 30위에 포함되는 토큰 중 불용어가 없을 때까지 제거를 반복하였는데, 결과적으로 "생각", "말씀", "때문", "사실", "경우", "필요", "사람", "얘기", "다음", "자체", "의미" 등이 삭제되었다.

네 번째 단계는 동의어를 하나의 대표 토큰으로 통일시키는 작업이다. 실제로 같은 의미를 가지는 한국어와 외래어, 동일한 대상을 지칭하지만, 띄어쓰기가 다른 복합어 등이 그 대상이다. 본 연구는 〈Table 5〉와 같이 여러 동의어를 기준단어로 통일하였다.

다섯 번째 단계는 한 문장 내에 토큰의 개수가 너무 적은 문장들을 제거하는 작업이다. 〈Table 6〉에서 보는 바와 같이 토큰화를 시행하고 나면 의미 있는 토큰이 하나도 존재하지 않는 문장도 다수 존재한다. 본 연구는 토큰을 하나씩 분석할 뿐 아니라 앞뒤 토큰과의 관련성도 분석하므로, 최소 3개 이상의 토큰을 보유한 문장만을 남기고 나머지는 제거하였다. 그 결과 3,474개 문장이 남았다.

Ⅳ. 분석결과

1. 키워드 추출 및 네트워크 분석

STM을 시행하기 전에 키워드 즉 가장 빈번하게 출현한 토큰을 살펴보면 〈Table 7〉과 같다. 키워드는 최소 20개의 문서에서 등장하는 토큰만을 대상으로 하였으며, 세 개까지 연결된 토큰조합(Trigram)을 고려하였다. 빈도는 문장 내에서 얼마나 자주 출현하는가를 측정하는 토큰빈도(Token Frequency)와 너무 많은

Table 5. Synonym processing

Representative	Synonyms
상업용 부동산 Commercial real estate	상업용 부동산, 커머셜프라퍼티, 커머셜 프라퍼티
오피스 Office	사무실, 사무용부동산, 사무용 부동산, 업무용부동산, 업무용 부동산
리테일 Retail	상가, 점포, 매장, 업장, 가게, 판매점, 상가용부동산, 점포용부동산, 매장용부동산, 판매용부동산
인더스트리얼 Industrial	물류, 창고, 공장, 물류센터, 물류 센터, 물류창고, 물류 창고, 물류부동산, 산업용 부동산
호스피탈리티 Hospitality	호텔, 숙박시설, 숙박용부동산, 숙박용 부동산, 리조트
레지덴셜 Residential	주택, 아파트, 공동주택, 주거, 주거용부동산, 주거용 부동산
지식산업센터 Knowledge industrial center	지식산업 센터, 지식 산업센터, 지식 산업 센터, 지산, 아파트형공장, 아파트형 공장
데이터 Data	자료, 정보, 데이터
임대동향조사 Rental trend surveys	임대동향 조사, 임대 동향조사, 상업용부동산 임대동향조사, 상업용부동산 임대 동향 조사, 상업용부동산 임대동향 조사, 상업용부동산 임대동향조사
가격지수 Price index	가격 지수, 매매가격지수, 매매가격 지수, 매매 가격지수, 거래가격지수, 거래가격 지수, 거래 가격지수, 매매가지수, 매매가 지수
임대료지수 Rent index	임대료 지수, 임대가 지수, 임대가지수, 월세 지수, 월세지수
수익률지수 Return index	수익률 지수
상업용 부동산통계 Commercial real estate statistics	상업용부동산 통계, 부동산시장통계, 부동산시장 통계, 부동산 시장통계, 부동산 시장 통계, 부동산관련통계, 부동산관련 통계, 부동산 관련통계, 부동산 관련 통계
오피스통계 Office statistics	오피스 통계, 오피스시장통계, 오피스시장 통계, 오피스 시장통계, 오피스관련통계, 오피스관련 통계, 오피스 관련통계, 오피스 관련 통계
리테일통계 Retail statictics	리테일 통계, 리테일시장통계, 리테일시장 통계, 리테일 시장통계, 리테일관련통계, 리테일관련 통계, 리테일 관련통계, 리테일 관련 통계

* Since the process was applied to Korean text, the synonyms are written in the original language.

Table 6. Number of tokens by sentence

Sentences	Mean	SD0	Min	25%	50%	75%	Max
4,418	5.59	3.74	0	3	5	8	27

문장에 출현하는 토큰에 페널티를 부과하는 역문서 토큰빈도(Token Frequency - Inverse Document Frequency) 두 가지로 산출하였다. 다수의 문장에 흔히 출현하는 토큰은 오히려 분석에 큰 의미를 가지지 않는 경우가 많기 때문이다.

두 가지 빈도 지표는 큰 차이를 보이지 않았다. 가장 빈도가 많은 키워드는 "데이터"와 "조사"로서 통계의 기본이 되는 자료수집에 대한 언급이 많았음을 반영하였다. 그 뒤로는 "리테일", "상권", "부분"이 자주 언급되었는데, 이는 특히 리테일에 대한 통계를 작성하는 것이 상권 설정이나 부분(또는 집합) 건물의 처리 측면에서 쉽지 않다는 논의가 많았던 것을 반영하는 것으로 보인다. 5위권 밖의 키워드는 표를 참고하기 바란다.

Table 7. Keywords

Rank	Word (TF)	TF	Word (TF-IDF)	TF-IDF
1	Data	863	Data	239
2	Survey	478	Survey	152
3	Retail	434	Retail	130
4	District	337	Part	116
5	Part	336	District	111
6	Office	300	Degree	105
7	Statistics	289	Statistics	101
8	Rent	287	Office	101
9	Degree	284	Sample	99
10	Sample	262	Rent	98
11	Real estate	195	Index	68
12	Index	189	Real estate	67
13	Building	167	Region	63
14	Region	167	Criteria	63
15	Market	164	Market	61
16	Criteria	158	Building	59
17	Lease	144	Problem	54
18	Price	129	Usage	50
19	Problem	127	Lease	50
20	Institution	123	Price	48

키워드 간의 관계는 네트워크 분석을 통해 보다 명확히 파악할 수 있다. 네트워크 분석은 위에서 추출된 키워드를 대상으로 문장 내에서 함께 등장하는 것들을 노드(Node)로 연결하되, 빈도에 따라 가중치를 부여하는 가중 동시출현 네트워크(Weighted Co-occurrence Network) 방식으로 시행하였다. 두 단어가 함께 등장하는 빈도가 30 이상일 때만 네트워크 연결을 인정하였고(Edge Threshold Value 30), 노드의 중요도를 나타내는 DC(Degree Centrality)값이 하위 20%인 것을 제외하여 시각화를 간결하게 하였다.

〈Figure 1〉에서 보는 바와 같이 가장 빈도가 높은 키워드인 "데이터"와 "리테일"이 중심 노드가 되어 다른 노드들과 연결된 것을 알 수 있다. "오피스", "임대료", "부분", "정도" 등의 키워드는 "데이터"와 "리테일" 모두와 밀접하고, "데이터"는 기관투자자, 국가승인통계, 소상공인 정보제공 시스템과 관련된 키워드와 밀접하며, "리테일"은 집합건물, 구분건물, 평이나 층과 같은 규

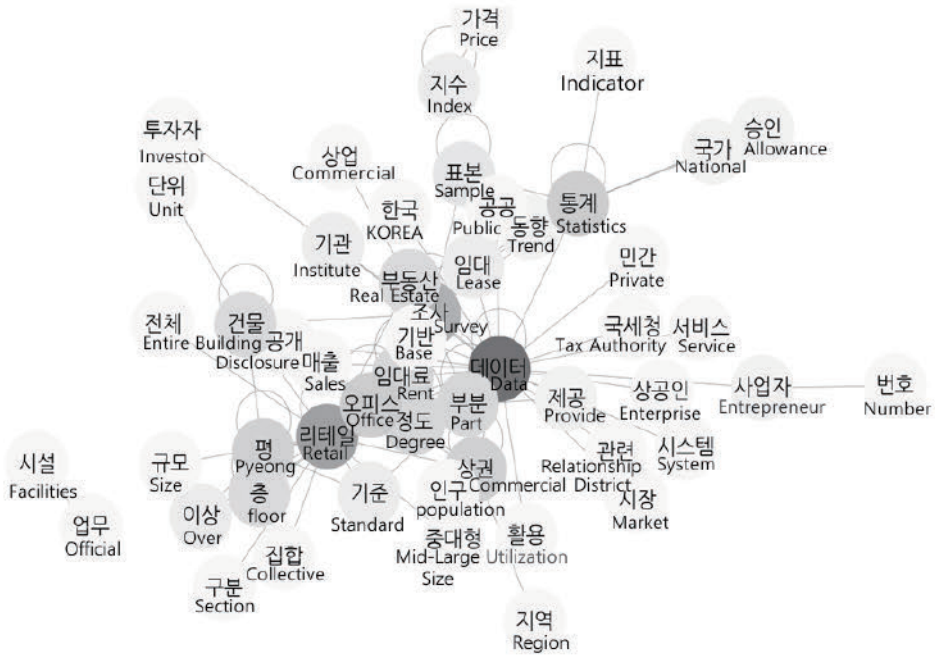

Figure 1. Network analysis

모 기준과 관련된 키워드와 밀접하게 연결되어 있다. 이러한 결과는 이후 진행하는 STM에서 토픽을 정의 또는 명명하는 데 도움을 준다.

2. 제공자와 사용자 간 차이

STM을 이용하여 3,474개 문장의 메타데이터 중 하나인 화자의 포지션에 따른 문제의식 차이를 검정하는 작업은 크게 두 단계로 이루어진다. 첫째, 전체 텍스트 데이터로부터 의미 있는 토픽을 도출한다. 이 토픽이 문제의식의 내용이라고 볼 수 있다. 이를 위해서는 먼저 토픽의 개수부터 결정해야 한다. 둘째, 도출된 토픽빈도를 반응변수로 하여 설명변수 즉 화자의 포지션을 나타내는 명목변수가 미치는 영향의 방향, 크기, 유의성 등을 검정한다. 앞에서 언급한 바와 같이 화자가 해당 업무를 수행한 종사연수를 함께 투입하여 경력에 따른 문제의식 차이를 통제한다.

최적의 토픽 수를 결정하는 데는 네 가지 지표가 주로 사용된다. Held-Out Likelihood는 모델의 일반화 성능을 나타내는데, 값이 클수록 새로운 데이터를 잘 예측한다는 것을 의미한다. Residuals는 모델의 설명력을 나타내는데, 값이 작을수록 문서 내 단어 분포를 잘 설명한다는 것을 의미한다. Semantic Coherence는 각 토픽을 구성하는 단어들의 의미적 일관성을 평가하는데, 값이 클수록 해석 가능성이 높다는 것을 의미한다. Lower Bound는 모델의 토픽 적합성을 나타내는데, 값이 클수록 모델이 데이터를 잘 설명한다는 것을 나타낸다.

〈Figure 2〉는 토픽 수에 따른 네 가지 지표의 크기를 보여준다. 여기서 그래프의 상승 또는 하락 추세가 크게 꺾이는 점이 최적 토픽 수에 해당하는데, Held-Out Likelihood와 Lower Bound는 대체로 일정하게 상승하는 반면 Residuals와 Semantic Coherence는 토픽 수 10개 내외에서 하락 추세가 약해지는 것을 관찰할 수 있다. 본 연구를 이를 참고하여 최적 토픽 수를 10개로 설정한다.

각 토픽별로 주요 토큰을 할당하는 기준에는 크게 네 가지가 있다. Highest Probability는 각 토픽에 등장할 확률이 높은 토큰을 말한다. FREX(FRequency and EXclusivity)는 빈도(Frequency)와 배타성(Exclusivity)을 결합했을 때 순위가 높은 토큰을 말한다. 확률이나 빈도만을 고려했을 때 다수의 문장에 출현하는 흔한 토큰이 중요하게 취급되는 문제를 보완한 것이다. Lift는 어느 토큰이 특정 토픽에서 나타날 확률을 전체 텍스트 데이터에서 나타날 확률로 나눈 값이 큰 토큰을 말한다. Score는 확률, 빈도, 배타성 등을 종합적으로 고려하여 자동으로 선택된 토큰을 말한다. 이는 앞의 세 가지 기준을 종합한 것인 반면, 어떤 기준이 큰 역할을 했는지 알 수 없어 해석이 어렵다는 문제를 가진다.

이 중 가장 널리 사용되는 기준은 FREX이며, 본 연구도 이를 기준으로 주요 토큰을 할당한다. 10개의 토픽에 대해 주요 토큰을 할당한 결과는 〈Table 8〉과 같다. FREX 외의 기준에 따른 주요 토큰 할당 결과도 이와 크게 다르지 않은데, 참고를 위해 〈Appendix 1〉에 수록하였다.

STM은 10개의 토픽에 대해 주요 토큰을 할당할 뿐 토픽의 이름까지 명명하지는 않는다. 각 토픽을 무엇이라 부를지는 인터뷰를 진행하고, 텍스트 데이터를 처리한 연구자가 해당 토큰의 조합을 참고하여 직접 결정해야 한다. 가장 일반적인 방법은 해당 토큰의 조합이 출현하는 대표적인 문장들을 찾아서 그 내용을 표현하는 것이다.

토픽 1은 '신축건물과 같이 일시적으로 공실이 큰 표본을 어떻

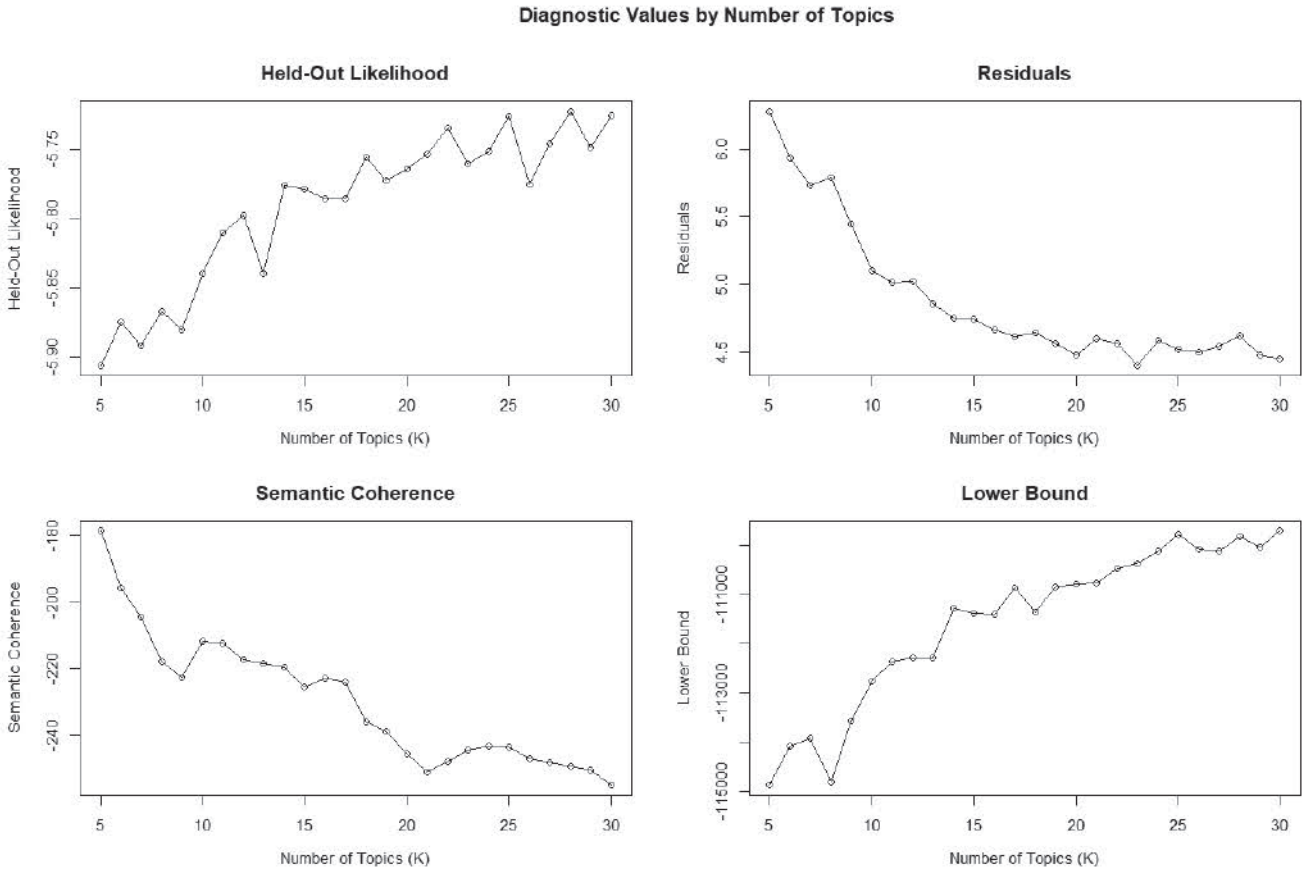

Figure 2. Optimal number of topics (providers and users)

Table 8. 10 Topics (providers and users)

Topic 1: 공실, 부분, 선택, 반영, 만약, 렌트, 젠스타 Vacancy, Part, Selection, Reflection, If, Rent, GenStar
Topic 2: 지수, 임대, 가격, 레지덴셜, 상업, 거래, 동향 Index, Lease, Price, Residential, Commercial, Transaction, Trend
Topic 3: 상공인, 관련, 국세청, 검증, 데이터, 영역, 유동 Enterprise, Relationship, National Tax Service, Validation, Data, Area, Footprint
Topic 4: 결국, 분석, 동, 펀드, 위치, 인덱스, 유저 Finally, Analysis, Dong, Fund, Location, Index, User
Topic 5: 리테일, 평, 구분, 서울, 규모, 집합, 중대형 Retail, Pyeong, Section, Seoul, Size, Collective, Mid-large size
Topic 6: 국가, 분기, 승인, 가치, 확인, 오픈, 상황 National, Quarter, Approval, Value, Confirmation, Open, Situation
Topic 7: 조사, 표본, 추가, 설계, 인더스트리얼, 과거, 이슈 Survey, Sample, Addition, Design, Industrial, Past, Issue
Topic 8: 오피스, 임대료, 지방, 얼마, 계약, 매출액, 단위 Office, Rent, Non-Seoul area, Amount, Contract, Sales, Unit
Topic 9: 상권, 지역, 업무, 업종, 권역, 설정, 용도 Commercial district, Region, Official, Type of business, District, Establishment, Use
Topic 10: 투자, 수익, 투자자, 방식, 빌딩, 임차인, 비용 Investment, Revenue, Investor, Method, Building, Tenant, Cost

게 통계에 반영할 것인가?', '젠스타의 경우 이러한 문제에 대응하기 위해 신축건물을 포함한 통계와 포함하지 않은 통계를 모두 발표하고 있다.', '표본이 집합건물의 부분을 구성하는 경우 임대료(렌트)와 공실률을 어떻게 조사할 것인가?', 등과 같이 조사의 기준이 엄밀하지 않을 때 발생할 수 있는 통계의 편의에 대한 논의를 대변하고 있다. 본 연구는 이 주제를 "조사의 엄밀성"이라고 명명한다.

토픽 2는 '시장동향을 파악하기 위해서는 임대료나 매매가(가격)에 대한 통계가 꼭 필요하다.', '임대료나 매매가에 대해서는 주택(레지덴셜) 부문과 같이 지수를 개발해서 공표해야 한다.', '임대료와 매매가에 대한 통계는 실제 거래에 기반할 필요가 있다.' 등과 같이 거래 관련 지표의 부족에 대한 문제의식을 대변하고 있다. 본 연구는 이 주제를 "거래관련 통계지표의 필요성"이라고 명명한다.

토픽 3은 '국세청, 소상공인진흥공단 등 상업용 부동산의 운영과 관련된 정보를 보유하고 있는 기관들이 자료를 공유해야 한다.', '민간부문이 조사한 유동인구와 같은 일차적인 정보로는 이용에 한계가 있다.', '영역 간 정보공유는 각 기관이 보유한 정보를 검증하는 역할도 하게 된다.' 등과 같이 양질의 자료를 보유한 기관 간의 협력을 통해 통계의 질을 높여야 한다는 주장을 대변하고 있다. 본 연구는 이 주제를 "유관기관 자료공유의 필요성"이라고 명명한다.

토픽 4는 비록 하나의 토픽을 구성하기는 했으니, 주요 토큰으로부터 다른 토픽들과 차별화되는 명백한 주제를 파악하기 힘들다. 본 연구는 이를 "기타"로 명명한다. 이러한 토픽이 존재하는 것으로부터 10개라는 토픽 수가 중요한 주제를 대부분 포함하는 것임을 알 수 있다.

토픽 5는 '리테일의 면적(평)은 전용면적을 기준으로 해야 한다.', '서울과 지방에서 중대형, 소형과 같은 규모 구분의 기준이 달라야 한다.', '리테일은 집합건물(구분소유권)인가 일반건물인가에 따라 성질이 매우 다르다.' 등과 같이 리테일의 특성에 기반하는 어려움을 대변하고 있다. 본 연구는 이 주제를 "리테일 통계의 어려움"이라고 명명한다.

토픽 6은 '민간부문에서 국가 전체를 대상으로 조사를 하기는 어렵다.', '분기와 같이 잦은 주기로 조사를 하는 데는 많은 시간과 비용이 소요된다.', '가격수준을 파악하기 위해 가치를 평가하는 데에도 많은 비용이 소요된다.', '이렇게 민간부문이 감당하기 어려운 통계는 국가승인통계가 담당하고 자료를 공개(오픈)해야 한다.' 등과 같이 공공부문이 더 큰 역할을 해야 한다는 문제의식을 대변하고 있다. 본 연구는 이를 "공공부문의 역할"이라고 명명한다.

토픽 7은 표본설계와 관련해서 고려해야 하는 사항을 다양하게 언급하고 있다. '표본을 설계하고 추가할 때 시장현황을 잘 고려해야 한다.', '혁신도시와 같이 특별한 이슈가 있는 지역을 별도로 표본 구성할 필요가 있다.', '공장, 물류센터 등 인더스트리얼 섹터에 대한 표본을 추가할 필요가 있다.', '표본을 변경할 때 과거와의 연속성을 잘 고려해야 한다.' 등이 대표적인 사례다. 본 연구는 이 주제를 "표본설계의 고려사항"이라고 명명한다.

토픽 8은 '오피스 임대료 조사에 있어서 층과 같은 물리적 기준도 고려해야 한다.', '호가임대료, 표면임대료, 계약임대료 등 다양한 형식 중에서 계약임대료를 조사해야 한다.', '지방마다 오피스 조사기준을 차별화해야 한다.', 등과 같은 오피스 통계에 대한 문제의식을 대변하고 있다. 본 연구는 이 주제를 "오피스 조사기준의 복잡함"이라고 명명한다.

토픽 9는 '상권이나 지역의 설정에 따라 통계가 달라지므로 신중히 경계를 정해야 한다.', '조사범위를 설정할 때 건물의 용도, 임차인의 업종 등을 고려해야 한다.' 등과 같이 통계조사 및 지표생산의 공간적 단위에 대한 문제의식을 대변하고 있다. 본 연구는 이 주제를 "통계구 설정의 엄밀성"이라고 명명한다.

토픽 10은 '상업용 부동산에 대한 투자자 특히 기관투자자의 관심 대상인 수익과 위험에 관한 통계가 필요하다.', '주거용 부동산과 달리 상업용 부동산에 대해서는 사용자뿐 아니라 투자자의 입장도 잘 반영해야 한다.' 등과 같이 투자지표의 필요성에 대한 주장을 대변하고 있다. 본 연구는 이 주제를 "투자자 관점의 반영"이라고 명명한다.

여기서 토픽의 번호는 무작위로 설정되기 때문에 의미를 가지지 않는다. 10개의 토픽 중 어느 것이 중요한지는 기대토픽빈도(Expected Topic Proportions)를 통해 판단한다. 이는 각 문장에 특정 토픽이 포함된 평균적인 비율을 나타낸다. 〈Figure 3〉은 10개의 토픽을 기대토픽빈도 크기순으로 보여주고 있다. 1위부터 토픽명을 나열하면 다음과 같다. 1위: 유관기관 자료공유의 필요성 〉 2위: 통계구 설정의 엄밀성 〉 3위: 표본설계의 고려사항 〉 4위: 오피스 조사기준의 복잡함 〉 5위: 표본조사의 엄밀성 〉 6위: 리테일 통계의 어려움 〉 7위: 거래관련 통계지표의 필요성 〉 8위: 공공부문의 역할 〉 9위: 기타 〉 10위: 투자자 관점의 반영.

10개의 토픽이, 정확히는 10개의 토픽빈도가 통계의 제공자와 사용자를 나타내는 메타데이터에 따라 어떻게 차이 나는지 검정

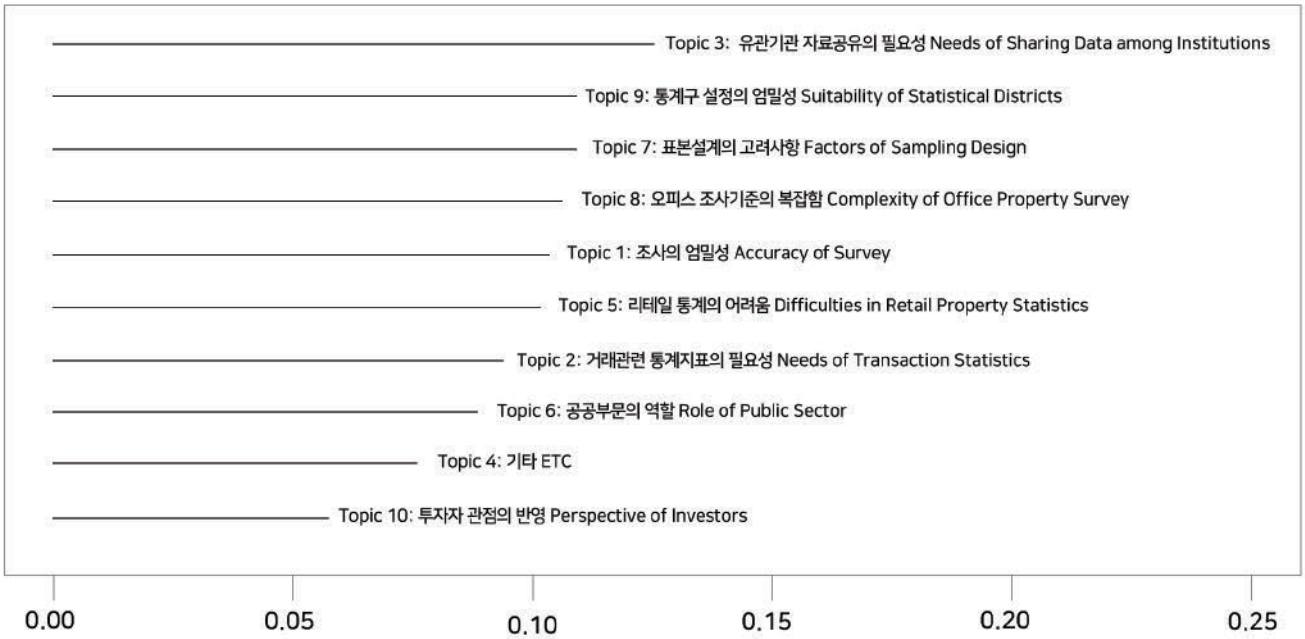

Figure 3. Rank of topic frequencies (providers and users)

한 결과는 〈Table 9〉와 같다. 실제로는 각 토픽 별로 분석이 이루어지며, 종사연수의 유의성도 함께 검정된다. 하지만, 본 연구에서는 관심 대상인 제공자와 사용자 여부만 모아서 하나의 표로 정리했다. 표에서 보는 바와 같이 두 포지션 중에서 제공자를 기준값으로 설정하였다. 〈Figure 4〉는 동일한 결과를 직관적으로 보기 편하게 그래프로 나타낸 것이다.

첫째, 통계 제공자의 문제의식이 강하게 나타난 토픽은 "유관기관 자료공유의 필요성"과 "기타" 두 가지였다. 앞에서 살펴본 바와 같이 "유관기관 자료공유의 필요성"은 10개 토픽 중 가장 토픽빈도가 높았는데, 이것이 주로 통계 제공자에 의해 제기되었다는 것을 알 수 있다. 이러한 결과는 상업용 부동산 통계가 대부분 직접 조사에 의해 작성되기 때문인 것으로 보인다. 상업용 부동산 시장이 성장함에 따라 조사의 부담은 커지는 반면, 이를 수행하기 위한 시간과 비용을 확보하기는 쉽지 않기 때문이다. 이를 해결하기 위한 방안으로 통계 제공자는 필요한 정보를 보유한 기관과의 상호 정보교류를 원하고 있다. 하지만, 그러기 위해서는 법령이나 정보체계를 개편해야 하므로 해결하기 매우 어려운 문제로 인식하고 있다. 통계 제공자는 그 외에도 여러 가지 실무적인 어려움을 제기하였다. 하지만 그 내용이 다양해서 하나의 주제로 명명하기는 어려웠다.

Table 9. Significance of metadata (providers and users)

Topic	Metadata	Coef.	S. Err.	t	Pr(>|t|)
Topic 1	User	0.0196	0.0053	3.642	0.0002 ***
Topic 2	User	0.0173	0.0063	2.734	0.0062 **
Topic 3	User	-0.0560	0.0073	-7.668	2.26e-14 ***
Topic 4	User	-0.0460	0.0056	-8.194	3.53e-16 ***
Topic 5	User	0.0113	0.0061	1.847	0.0648
Topic 6	User	0.0029	0.0053	0.554	0.5794
Topic 7	User	-0.0069	0.0059	-1.174	0.2405
Topic 8	User	0.0027	0.0053	0.518	0.6043
Topic 9	User	0.0315	0.0060	5.171	2.46e-07 ***
Topic 10	User	0.0234	0.0047	4.914	9.35e-07 ***

Signif. codes: ***p<0.001; **p<0.01; *p<0.05; .p<0.1

둘째, 통계 사용자의 문제의식이 강하게 나타난 토픽은 "조사의 엄밀성", "거래관련 통계지표의 필요성", "통계구 설정의 엄밀성", "투자자 관점의 반영" 네 가지였다. 이중 "조사의 엄밀성"과 "통계구 설정의 엄밀성"은 현재 발표되고 있는 통계의 신뢰성이나 활용성과 관련된 토픽이며, "거래관련 통계지표의 필요성"과 "투자자 관점의 반영"은 향후 발표되기를 희망하는 통계와 관련된 토픽이라고 할 수 있다. 이러한 결과는 향후 수요자 중심으로 통계를 개선한다면 이들 문제를 우선 해결해야 하는 것을 알려준다.

특히 "통계구 설정의 엄밀성"은 10개 토픽 중 두 번째로 토픽빈도가 높았다. 실제로 인터뷰에서 CBD, YBD, GBD로 대표되는 서울 오피스 권역의 경계를 어떻게 나눌 것인가에 대한 논의도 많았고, 서울뿐 아니라 전국을 대상으로 리테일 상권을 어떻게 설

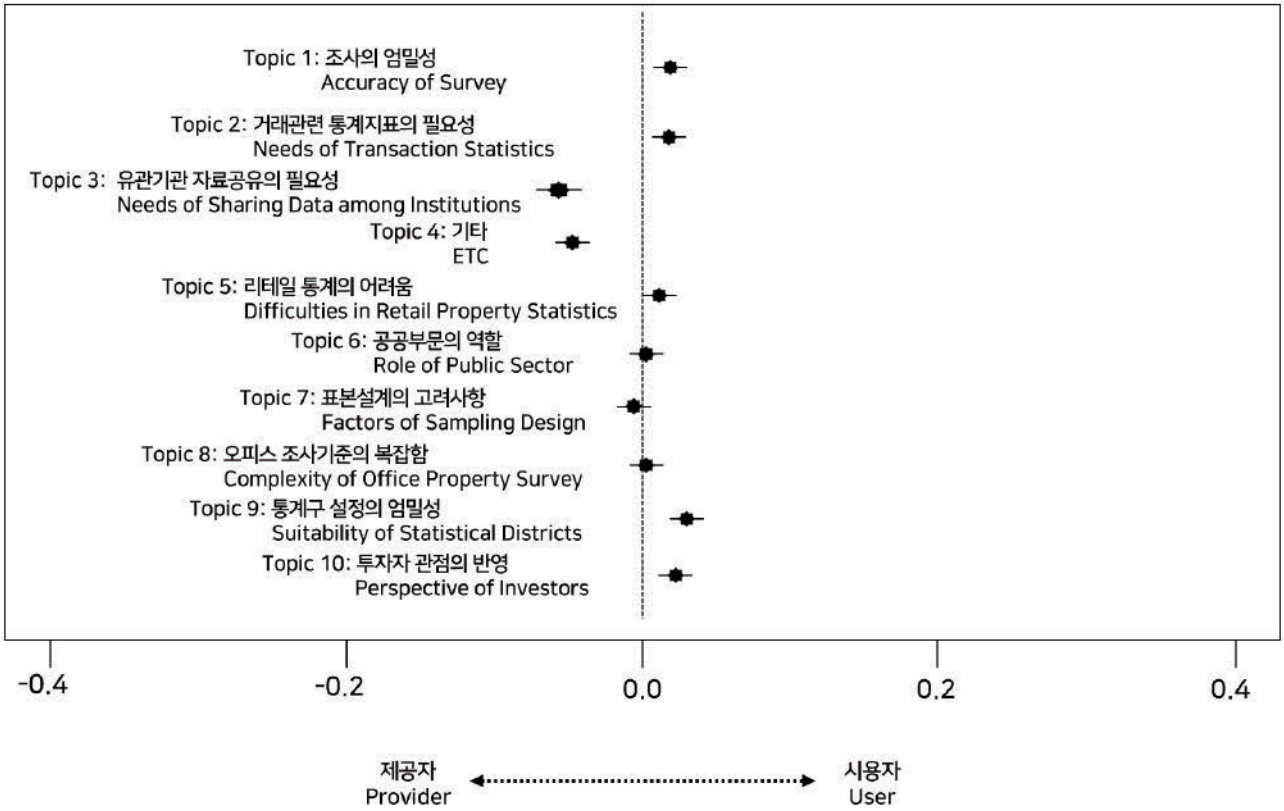

Figure 4. Significance of metadata (providers and users)

정할지에 대해서는 참여자 간 이견도 많았다. 통계구를 작게 설정할수록 특정 지역에 대해 보다 세밀한 정보를 제공할 수 있는 반면 표본수가 적어서 통계의 안정성이 낮아질 수 있기 때문이다. 이에 대한 해결방안은 보다 심도 있는 연구를 통해 찾아야 하겠지만, 현재의 통계구가 사용자의 니즈를 충족하지 못한다는 사실은 분명하게 확인할 수 있다.

셋째, 통계 제공자와 사용자 간에 유의한 차이가 발견되지 않은 토픽은 "리테일 통계의 어려움", "공공부문의 역할", "표본설계의 고려사항", "오피스 조사기준의 복잡함" 네 가지였다. 이중 "표본설계의 고려사항", "오피스 조사기준의 복잡함", "리테일 통계의 어려움"은 상업용 부동산의 용도별로 적절한 표본을 선정하고 조사하는 것이 쉽지 않다는 문제의식을 보여주고, "공공부문의 역할"은 상업용 부동산 통계가 가진 문제를 해결하기 위해서는 정부가 나서서 역할을 해야 한다는 주장을 보여준다. 이중 "표본설계의 고려사항"은 10개의 토픽 중 세 번째로 토픽빈도가 높아서 통계 제공자와 사용자 모두가 심각한 문제의식을 가지고 있다는 것을 알 수 있다.

3. 직업군에 따른 차이

지금까지 통계 제공자와 사용자의 문제의식 차이를 살펴보았다. 하지만, 제공자와 사용자 내에서도 문제의식에 차이가 있을 수 있으며, 이는 해결방안을 모색하는 데 중요한 정보가 될 수도 있다. 이를 확인하기 위해 제공자를 공공통계와 민간통계, 사용자를 교육연구와 금융투자 네 가지 직업군으로 나누어 앞에서와 동일한 절차로 STM을 시행하였다.

〈Figure 5〉, 〈Table 10〉, 〈Figure 6〉은 토픽 수를 결정하고, 토픽을 도출하여 명명한 절차를 보여준다. 여기서도 토픽별로 주요 토큰을 할당하는 작업은 FREX를 기준으로 하였다. 그 외의 기준에 따른 주요 토큰 할당 결과는 〈Appendix 2〉에 수록하였다. 비록 명목변수인 메타데이터의 값을 2개에서 4개로 증가시켰지만, 최적 토픽 수, 토픽별 주요 토큰, 토픽명에는 거의 차이가 없었다. 또한 토픽 순위 역시 "통계구 설정의 엄밀성"이 두 계단 하락한 것 외에는 한 계단의 상승이나 하락과 같이 경미한 차이밖에 없었다.

직업군의 유의성을 검정한 결과는 〈Table 11〉과 같다. 참고로 의미 있는 토픽이라고 보기 어려운 Topic 4는 표에서 제외하였다. 여기서는 통계 제공자 중 하나인 공공통계를 기준값으로 설정하였는데, 다음과 같은 의미 있는 사실들을 발견할 수 있다.

첫째, 통계 제공자의 문제의식을 강하게 나타낸 "유관기관 자료공유의 필요성"은 교육연구, 금융투자, 민간통계 세 직업군이 모두 음(-)의 값을 가지며, 매우 유의하였다. 특히 계수값의 크기

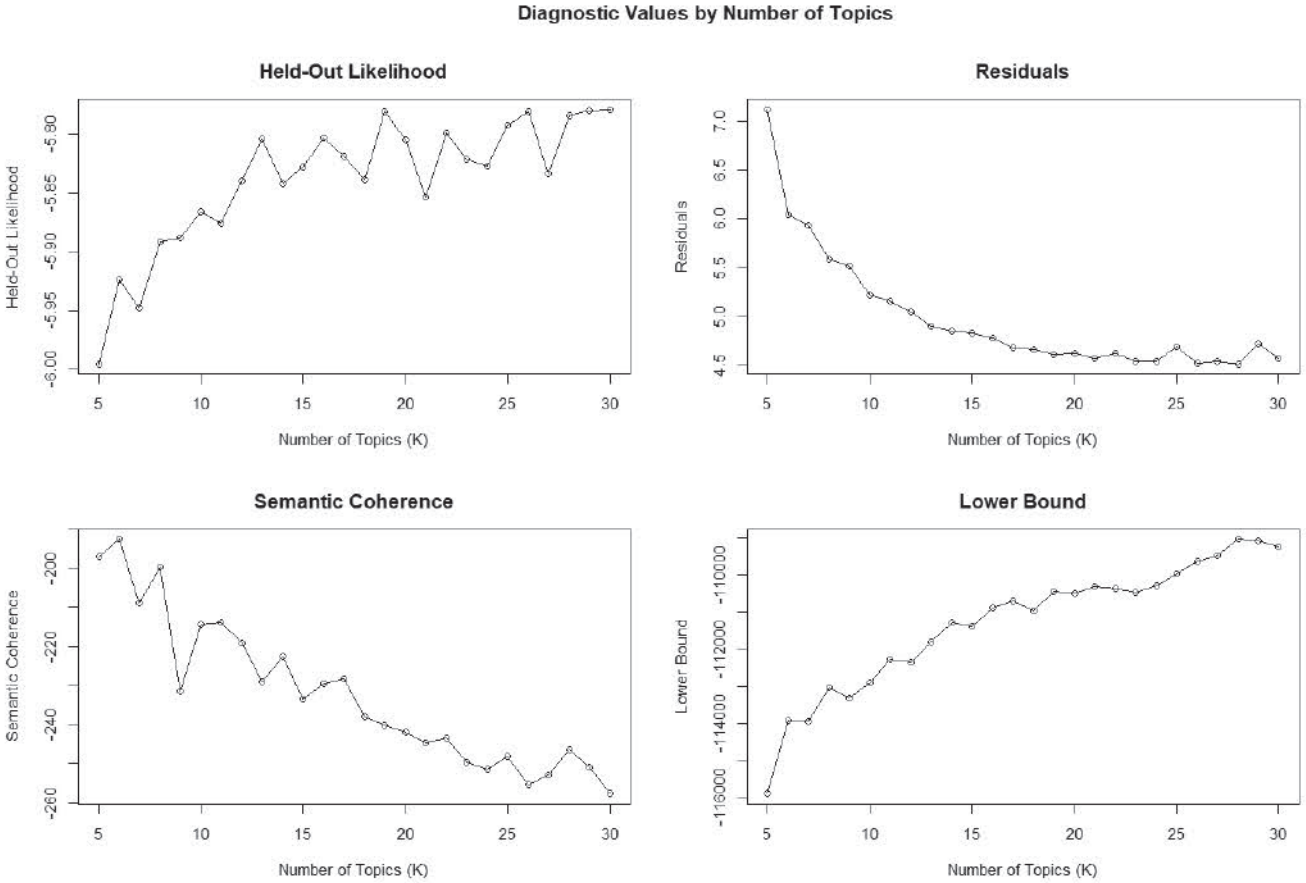

Figure 5. Optimal number of topics (by profession)

와 P값이 거의 동일하여 공공통계 외에는 모두 비슷한 문제의식을 가지는 것으로 나타났다. 이는 통계 제공자의 문제의식을 강하게 나타낸 "기타"에서도 마찬가지였다. 다만, "유관기관 자료공유의 필요성"에 비해 공공통계와 민간통계의 문제의식 차이는 다소 작았다.

Table 10. 10 Topics (by profession)

Topic 1: 부분, 통계, 공실, 반영, 만약, 실제, 과정 Part, Statistics, Vacancy, Reflection, If, Actual, Process
Topic 2: 지수, 임대, 가격, 레지덴셜, 상업, 거래, 동향 Index, Lease, Price, Residential, Commercial, Transaction, Trend
Topic 3: 데이터, 제공, 회사, 기반, 상공인, 관련, 국세청 Data, Provide, Company, Base, Enterprise, Relationship, National Tax Service
Topic 4: 결국, 중요, 분석, 사업, 동, 펀드, 리츠 Finally, Important, Analysis, Business, Dong, Fund, REITs
Topic 5: 리테일, 정도, 평, 구분, 서울, 규모, 집합 Retail, Degree, Pyeong, Section, Seoul, Size, Collective
Topic 6: 시장, 국가, 분기, 상황, 평가, 자산, 대표 Market, National, Quarter, Situation, Evaluation, Asset, Representative
Topic 7: 조사, 표본, 지금, 대상, 교수, 이슈, 예전 Survey, Sample, Now, Target, Professor, Issue, Previous
Topic 8: 오피스, 임대료, 지방, 얼마, 계약, 매출액, 단위 Office, Rent, Non-Seoul area, Amount, Contract, Sales, Unit
Topic 9: 상권, 지역, 업무, 업종, 권역, 설정, 용도 Commercial district, Region, Official, Type of business, District, Establishment, Use
Topic 10: 투자, 수익, 투자자, 방식, 빌딩, 최근, 임차인 Investment, Revenue, Investor, Method, Building, Recent, Tenant

Table 11. Significance of metadata (by profession)

Topic	Metadata	Coef,	S. Err.	t	Pr(>|t|)
Topic 1	Edu & Res	0.0050	0.0079	0.634	0.5260
	Fin & Inv	0.0025	0.0064	0.391	0.6960
	Priv. Stat.	-0.0141	0.0066	-2.113	0.0346 *
Topic 2	Edu & Res	0.0623	0.0089	6.957	4.15e-12 ***
	Fin & Inv	0.0165	0.0082	2.005	0.0451 *
	Priv, Stat	0.0337	0.0083	4.067	4.87e-05 ***
Topic 3	Edu & Res	-0.1274	0.0102	-12.447	<2e-16 ***
	Fin & Inv	-0.1235	0.0098	-12.595	< 2e-16 ***
	Priv, Stat	-0.1251	0.0101	-12.300	< 2e-16 ***
Topic 5	Edu & Res	0.0097	0.0087	1.112	0.2660
	Fin & Inv	0.0174	0.0079	2.201	0.0277 *
	Priv, Stat	0.0043	0.0081	0.529	0.5964
Topic 6	Edu & Res	0.0059	0.0071	0.831	0.4060
	Fin & Inv	0.0365	0.0078	4.634	3.72e-06 ***
	Priv, Stat	0.0445	0.0067	6.624	4.04e-11 ***
Topic 7	Edu & Res	-0.0144	0.0082	-1.753	0.0797 .
	Fin & Inv	-0.0129	0.0074	-1.747	0.0807 .
	Priv, Stat	-0.0117	0.0074	-1.569	0.1168
Topic 8	Edu & Res	-0.0003	0.0082	-0.045	0.9645
	Fin & Inv	0.0174	0.0081	2.142	0.0323 *
	Priv, Stat	0.0113	0.0087	1.296	0.1952
Topic 9	Edu & Res	0.0852	0.0085	9.949	< 2e-16 ***
	Fin & Inv	0.0165	0.0080	2.058	0.0396 *
	Priv, Stat	0.0188	0.0083	2.251	0.0244 *
Topic 10	Edu & Res	0.0337	0.0074	4.553	5.47e-06 ***
	Fin & Inv	0.0863	0.0073	11.797	< 2e-16 ***
	Priv, Stat	0.0559	0.0072	7.677	2.11e-14 ***

Signif. codes: ***p<0.001; **p<0.01; *p<0.05; .p<0.1
Edu: Education, Res: Research, Fin: Finance, Inv: Investment, Priv Stat: Private Statistics Provider

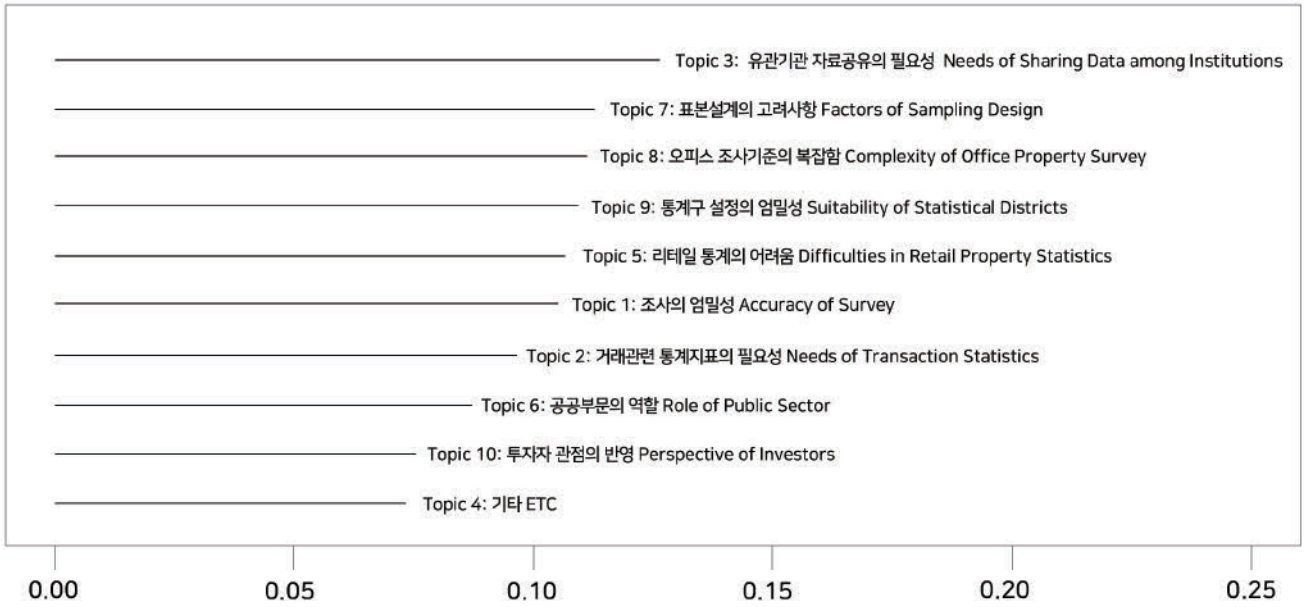

Figure 6. Rank of topic frequencies (by profession)

둘째, 통계 사용자의 문제의식을 강하게 나타낸 "조사의 엄밀성"은 민간통계만이 음(-)의 값으로 유의하였다. 이는 공공통계가 사용자와 다른 문제의식을 가지고 있지 않다는 것을 의미한다. 민간통계 종사자가 높은 전문성을 보유한 반면 그에 상승하는 엄밀한 조사를 하기 힘든 입장에 처해 있기 때문인 것으로 보인다. 그 외에 "거래관련 통계지표의 필요성", "통계구 설정의 엄밀성", "투자자 관점의 반영"은 반대로 교육연구, 금융투자, 민간통계 모두 양(+)의 값으로 유의하게 나타나 민간통계가 사용자와 다른 문제의식을 가지고 있지 않은 것으로 나타났다. 세 토픽이 대체로 자본시장과 연관되어 있어서 민간통계도 강한 문제의식을 가지고 있기 때문인 것으로 보인다.

셋째, 통계 제공자와 사용자 간에 유의한 차이가 발견되지 않은 토픽에 대해서는 예상대로 세 가지 직업군이 대체로 유의하지 않게 나타났다. 하지만 "공공부문의 역할"에 대해서는 금융투자와 민간통계 두 직업군이 양(+)의 값으로 매우 유의하였다. 앞에서와 마찬가지로 두 직업군은 자본시장과 밀접한 연관을 가지고 있다. 하지만, 민간부문에서는 수익률과 같은 지표를 생산하기 위한 정보를 구하기 어렵기 때문에 공공부문의 적극적인 역할을 기대하는 것으로 보인다. 통계 제공자와 사용자에 대한 분석에서는 이들 직업군이 두 포지션으로 분산되어 유의성을 발견하지 못한 것으로 판단된다.

V. 결 론

상업용 부동산은 연기금이나 공제회와 같은 기관투자자의 투자대상일 뿐 아니라 기업의 경제활동을 위한 생산요소라는 점에서 국민경제에 미치는 영향이 작지 않다. 그럼에도 상업용 부동산 통계는 토지나 주택 통계에 비해 상대적으로 빈약하여 시장의 효율성을 제고하는 데 충분히 기여하지 못하고 있다.

본 연구는 상업용 부동산 통계의 문제점을 구체적으로 포착하기 위해 33인의 전문가와 10회에 걸쳐 인터뷰를 진행하고, 그 내용을 STM으로 분석하였다. 인터뷰의 해석에는 연구자의 주관과 선임관이 개입될 여지가 있는데, 텍스트 분석을 통해 해석의 객관성을 높일 수 있다고 기대했기 때문이다. 연구의 결과와 시사점을 내용과 방법 두 가지 측면에서 요약하면 다음과 같다.

먼저 내용적 측면에서, 통계의 제공자는 직접 조사에만 의존하기보다 상업용 부동산 유관기관과 자료를 공유하여 조사의 부담을 줄임과 동시에 데이터의 신뢰성을 높일 필요가 있다는 점을, 통계의 사용자는 오피스 권역이나 리테일 상권과 같은 통계구 설정을 보다 엄밀하게 할 필요가 있다는 점을 가장 중요한 문제로 생각하였다. 또한, 표본선정 시보다 다양한 요소를 고려할 필요가 있다는 점에 대해서는 양자가 문제의식을 공유하였다. 특히 통계의 제공자와 사용자를 불문하고 자본시장과 연관이 큰 금융투자와 민간통계 종사자는 상업용 부동산 통계의 개선을 위해 공공부문이 더 많은 역할을 해야 한다고 주장하였다.

이상의 결과는 다음과 같은 네 가지 시사점을 제공한다. 첫째, 상업용 부동산 통계에 대한 다양한 비판 중에서 권역설정, 표본선택, 자료조사, 공공부문의 역할이 가장 중요한 것으로 지목되었다. 통계의 개선을 위해서는 이 네 가지 문제에 주목할 필요가 있다. 둘째, 상업용 부동산 통계의 활용성을 개선하기 위해서는 적절한 권역을 대상으로 유의미한 표본을 조사해야 한다. 그러나 사용자의 입장에서 볼 때, 현재의 통계들은 이러한 요구를 충족하지 못하고 있다. 권역설정과 표본선정에 있어서 실제 사용자의 의견을 더욱 진지하게 수렴할 필요가 있다. 셋째, 사용자의 요구에 적합한 통계 작성에는 상당한 조사의 부담을 감내해야 한다. 하지만, 통계의 제공자들은 그러한 여력을 가지고 있지 못하다. 이것이 적절한 권역설정과 표본선정을 어렵게 하는 원인일 수도 있다. 조사의 부담을 줄이기 위해서는 상업용 부동산 정보를 보유하고 있는 유관기관 간 정보공유가 절실하다. 통계의 제공자들은 조사에 더 많은 시간과 비용을 투입하기 전에 정보공유를 먼저 폭넓게 시도할 필요가 있다. 단, 공공부문이 보유한 정보의 경우 법적 제한에 의해, 민간부문이 보유한 정보의 경우 상업적 이유로 공유가 상당히 제한되어 있는 것이 현실이다. 이에 대해 제도의 개선, 상호 호혜적 합의 등을 적극적으로 모색해야 한다. 넷째, 상업용 부동산 통계의 작성에 더 많은 자원을 투입하고, 유관기관 간 공신력 있는 협조를 도모하기 위해서는 정부의 역할이 요구된다. 민간부문의 자발적 노력으로는 한계가 있다는 것을 경험했기 때문에 중요한 이슈로 추출되었을 것이기 때문이다. 공공부문의 역할은 국가승인통계인 상업용 부동산 임대동향조사의 확대와 같은 제도적 뒷받침이 전제되어야 한다. 하지만, 이를 위해서는 통계청과의 협의와 같은 엄격한 절차를 거쳐야 한다.

한편, 방법적 측면에서 STM은 전문가를 대상으로 한 인터뷰 분석에 유용하게 활용될 수 있다는 것을 알 수 있었다. STM은 방대한 텍스트 데이터에서 주요 토픽의 개수와 내용을 추출하고, 참여자 특성과 같은 메타데이터의 유용성을 검정하는데 좋은 성능을 보였다. 특히 메타데이터의 유의성은 현실 문제를 해결함에 있어서 어떤 토픽에 주목할 것인가를 결정하는 데 의미 있는 시사점을 제공하였다. 이러한 STM의 유용성은 본 연구의 인터뷰가 전문가를 대상으로 특정한 주제를 토론하는 것이어서 더욱 강화된 것으로 판단된다.

비록 STM이 전문가 인터뷰 해석에 유용하다는 것을 확인하였지만, 이것이 STM으로 충분한 해석이 가능하다는 것을 의미하지는 않는다. 참여자에 따라 발언한 문장과 단어의 양이 다르고, 화법에도 차이가 있어서 기계적인 텍스트 분석이 잠재된 의미를 포착하는 데 한계가 있기 때문이다. 따라서 STM은 연구자가 인터뷰를 보다 객관적으로 해석할 수 있게 돕는 하나의 도구라고 보는 것이 바람직하다. 또한, STM은 주요 토픽을 추출하고 메타데이터의 유의성을 검정할 뿐 그러한 토픽이 도출된 논리의 흐름은

보여주지 못한다. 이에 대해서는 근거이론(Grounded Theory)과 같은 질적 연구를 활용하는 방안이 강구되어야 할 것이다.

인용문헌
References

1. 강현희·백영민, 2022. "구조적 토픽 모형을 활용한 언론의 사회적기업 관련 담론 분석", 「사회적가치와 기업연구」, 15(2): 169-206.
   Kang, H.H. and Baek, Y.M., 2022. "An Analysis of Media Discourse on Social Enterprises Using Structural Topic Modeling", *Journal of Social Value and Enterprise,* 15(2): 169-206.
2. 문안나·이신행, 2020. "사회서비스원 정책 보도의 프레임 분석: 구조적 주제모형(Structural Topic Modeling)과 내용분석(Content Analysis)의 보완적 적용", 「한국광고홍보학보」, 22(4): 100-134.
   Mun, A.N. and Lee, S.H., 2020. "News Frames in the Coverage of Social Service Agency Policy: Complementary Application of Structural Topic Modeling and Content Analysis", *The Korean Journal of Advertising and Public Relations,* 22(4): 100-134.
3. 방보람·이태리·조정희, 2017. "상업용 부동산 정보의 중요도 평가 연구", 「부동산학연구」, 23(3): 29-39.
   Bang, B.R., Lee, T.R., and Cho, J.H., 2017. "A Study on the Importance Evaluation of Commercial Real Estate Information", *Journal of the Korea Real Estate Analysts Association*, 23(3): 29-39.
4. 이태리·조정희·최진·권건우, 2017. "미국, 싱가포르 사례를 통한 한국의 상업용 부동산 정보체계 구축 방안 연구", 「정보화정책」, 24(4): 44-67.
   Lee, T.R., Cho, J.H., Choi, J., and Kwon, G.W., 2017. "A Study on the Establishment of Commercial Real Estate Information Framework in Korea Compared with the Case of USA and Singapore", *Informatization Policy*, 24(4): 44-67.
5. 조성배·하성호, 2020. "구조적 토픽 모델링을 활용한 공공데이터 수요 분석", 「정보화연구」, 17(2): 103-118.
   Cho, S.B. and Ha, S.H., 2020. "Analysis of Open Government Data Demand Using Structural Topic Modeling", *Journal of Information Technology and Architecture*, 17(2): 103-118.
6. Blei, D.M., Ng, A.Y., and Jordan, M.I., 2003. "Latent Dirichlet Allocation", *Journal of Machine Learning Research*, 3: 993-1022.
7. Mimno, D. and McCallum, A., 2008. "Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression", Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI), 411-418.
8. Roberts, M.E., Stewart, B.M., Tingley, D., and Airoldi, E. M., 2013. "The Structural Topic Model and Applied Social Science", Proceedings of Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.
9. Roberts, M.E., Stewart, B.M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., and Rand, D.G., 2014. "Structural Topic Models for Open Ended Survey Responses", *American Journal of Political Science*, 58(4): 1064-1082.
10. Roberts, M., Stewart, B., and Tingley, D., 2016. "Navigating the Local Modes of Big Data: The Case of Topic Models", In *Computational Social Science: Discovery and Prediction*, New York: Cambridge University Press.

Date Received	2024-11-20
Reviewed(1st)	2025-01-23
Date Revised	2025-02-14
Reviewed(2nd)	2025-03-04
Date Accepted	2025-03-04
Final Received	2025-04-22

부 록
Appendix

Appendix 1. Detail indices of 10 topics (providers and users)

Topic 1	Highest prob: 부분, 통계, 활용, 지표, 공실, 차이, 반영 Part, Statistics, Utilization, Indicator, Vacancy, Difference, Reflection
	FREX: 공실, 부분, 선택, 반영, 만약, 렌트, 젠스타 Vacancy, Part, Selection, Reflection, If, Rent, GenStar
	Lift: 공실, 논의, 만족, 번째, 보조, 연관, 연속 Vacancy, Discussion, Satisfaction, th, Auxiliary, Related, Continuation
	Score: 부분, 통계, 공실, 지표, 활용, 반영, 측면 Part, Statistics, Vacancy, Indicator, Utilization, Reflection, Aspect
Topic 2	Highest prob: 부동산, 지수, 임대, 가격, 레지덴셜, 상업, 거래 Real estate, Index, Lease, Price, Residential, Commercial, Transaction
	FREX: 지수, 임대, 가격, 레지덴셜, 상업, 거래, 동향 Index, Lease, Price, Residential, Commercial, Transaction, Trend
	Lift: 거래량, 반복, 보도, 브랜드, 블룸버그, 임차, 지속 Volume, Repeated, Reporting, Brand, Bloomberg, Lease, Continue
	Score: 지수, 가격, 임대, 부동산, 상업, 레지덴셜, 동향 Index, Price, Lease, Real estate, Commercial, Residential, Trend
Topic 3	Highest prob: 데이터, 기관, 제공, 회사, 기반, 상공인, 관리 Data, Agency, Provide, Company, Base, Enterprise, Management
	FREX: 상공인, 관련, 국세청, 검증, 데이터, 영역, 유동 Enterprise, Relationship, National Tax Service, Validation, Data, Area, Footprint
	Lift: 감사, 경영, 국세청, 내년, 로우, 마이, 신보 Audit, Management, National Tax Service, Next year, Low, My, ShinBo
	Score: 데이터, 기반, 제공, 서비스, 기관, 상공인, 국세청 Data, Base, Provide, Service, Agency, Enterprise, National Tax Service
Topic 4	Highest prob: 결국, 입장, 중요, 사업자, 분석, 시스템, 사업 Finally, Position, Important, Entrepreneur, Analysis, System, Business
	FREX: 결국, 분석, 동, 펀드, 위치, 인덱스, 유저 Finally, Analysis, Dong, Fund, Location, Index, User
	Lift: 거시, 경제, 라인, 무시, 보호, 상장, 차원 Macro, Economy, Line, Disregard, Protect, Listing, Dimension
	Score: 사업자, 결국, 입장, 번호, 중요, 인덱스, 시스템 Entrepreneur, Finally, Position, Number, Important, Index, System
Topic 5	Highest prob: 리테일, 정도, 기준, 평, 전체, 구분, 서울 Retail, Degree, Criteria, Pyeong, Total, Division, Seoul
	FREX: 리테일, 평, 구분, 서울, 규모, 집합, 중대형 Retail, Pyeong, Division, Seoul, Size, Collective, Mid-Large Size
	Lift: 전면, 구분, 규모, 중대형, 권리금, 남대문, 도로 Front, Division, Size, Mid-Large Size, Premium, Namdaemun, Road
	Score: 리테일, 평, 정도, 기준, 커피, 규모, 집합 Retail, Pyeong, Degree, Criteria, Coffee, Size, Collective
Topic 6	Highest prob: 시장, 국가, 분기, 상황, 평가, 자산, 한국 Market, National, Quarter, Situation, Evaluation, Asset, Korea
	FREX: 국가, 분기, 승인, 가치, 확인, 오픈, 상황 National, Quarter, Approval, Value, Confirmation, Open, Situation
	Lift: 걱정, 분자, 운용, 의사, 자리, 정량, 지적 Worry, Molecule, Operation, Intention, Place, Quantity, Point
	Score: 시장, 국가, 평가, 운용, 분기, 자산, 상황 Market, National, Evaluation, Operation, Quarter, Asset, Situation
Topic 7	Highest prob: 조사, 표본, 민간, 지금, 가지, 대상, 방법 Investigation, Sample, Private, Now, Things, Target, Method
	FREX: 조사, 표본, 추가, 설계, 인더스트리얼, 과거, 이슈 Survey, Sample, Addition, Design, Industrial, Past, Issue
	Lift: 물량, 설계, 세빌스, 임의, 협조, 과거, 기사 Volume, Design, Savills, Random, Cooperate, Past, Article
	Score: 조사, 표본, 민간, 지금, 지출, 설계, 이슈 Survey, Sample, Private, Now, Expense, Design, Issue
Topic 8	Highest prob: 오피스, 임대료, 건물, 층, 단위, 이상, 지방 Office, Rent, Building, Floor, Unit, More, Non-Seoul area
	FREX: 오피스, 임대료, 지방, 얼마, 계약, 매출액, 단위 Office, Rent, Non-Seoul area, Amount, Contract, Sales, Unit.
	Lift: 보통, 인상, 학원, 강남대로, 갱신, 건축물, 군 Normal, Increase, Academy, Gangnam-daero, Renewal, Building, Country
	Score: 임대료, 층, 오피스, 건물, 이상, 단위, 지방 Rent, Floor, Office, Building, More, Unit, Non-Seoul area
Topic 9	Highest prob: 상권, 지역, 문제, 업무, 가능, 시설, 업종 Commercial District, Region, Problem, Official, Availability, Facilities, Industry
	FREX: 상권, 지역, 업무, 업종, 권역, 설정, 용도 Commercial District, Region, Official, Type of business district, Establishment, Use
	Lift: 그림, 상업시설, 업데이트, 용지, 지가, 크기, 권역 Painting, Commercial facility, Update, Plot, Land price, Size, Type of business district
	Score: 상권, 지역, 시설, 권역, 문제, 설정, 업무 Commercial District, Region, Facility, Type of business district, Problem, Setting, Official
Topic 10	Highest prob: 투자, 수익, 투자자, 방식, 빌딩, 임차인, 비용 Investment, Revenue, Investor, Method, Building, Tenant, Cost
	FREX: 투자, 수익, 투자자, 방식, 빌딩, 임차인, 비용 Investment, Revenue, Investor, Method, Building, Tenant, Cost
	Lift: 감정평가, 경험, 리스크, 방식, 밸류, 외국, 이사 Appraisal, Experience, Risk, Method, Value, Foreign, Director
	Score: 수익, 투자, 빌딩, 방식, 임차인, 비용, 경험 Revenue, Investment, Building, Method, Tenant, Cost, Experience

Appendix 2. Detail indices of 10 topics (by profession)

Topic 1	Highest prob: 부분, 통계, 활용, 지표, 공실, 차이, 반영 Part, Statistics, Utilization, Indicator, Vacancy, Difference, Reflection
	FREX: 부분, 통계, 공실, 반영, 만약, 실제, 과정 Part, Statistics, Vacancy, Reflection, If, Actual, Process
	Lift: 공실, 과정, 논의, 번째, 보조, 연속, 예측 Vacancy, Process, Discuss, th, Assist, Sequence, Predict
	Score: 부분, 통계, 공실, 지표, 활용, 반영, 만약 Part, Statistics, Vacancy, Indicator, Utilization, Reflection, If
Topic 2	Highest prob: 부동산, 지수, 임대, 가격, 레지덴셜, 상업, 거래 Real Estate, Index, Lease, Price, Residential, Commercial, Transaction
	FREX: 지수, 임대, 가격, 레지덴셜, 상업, 거래, 동향 Index, Price, Lease, Real Estate, Commercial, Residential, Trend
	Lift: 개념, 거래량, 금액, 대안, 레지덴셜, 매매, 반복 Concept, Transaction volume, Amount, Alternative, Residential, Sale, Repeat
	Score: 지수, 가격, 임대, 부동산, 상업, 레지덴셜, 동향 Index, Price, Lease, Real estate, Commercial, Residential, Trend
Topic 3	Highest prob: 데이터, 제공, 회사, 기반, 상공인, 관리, 공공 Data, Provide, Company, Base, Enterprise, Management, Public
	FREX: 데이터, 제공, 회사, 기반, 상공인, 관련, 국세청 Data, Provide, Company, Base, Enterprise, Relationship, National Tax Service
	Lift: 감사, 경영, 내년, 로우, 마이, 신보, 영역 Audit, Management, Next year, Low, My, ShinBo, Area
	Score: 데이터, 기반, 제공, 서비스, 상공인, 국세청, 공공 Data, Base, Provide, Service, Enterprise, National Tax Service, Public
Topic 4	Highest prob: 결국, 중요, 사업자, 분석, 시스템, 사업, 동 Finally, Important, Businessman, Analysis, System, Business, Dong
	FREX: 결국, 중요, 분석, 사업, 동, 펀드, 리츠 Finally, Important, Analysis, Business, Dong, Fund, REITs
	Lift: 거시, 경제, 기본, 대국민, 라인, 마스터, 무시 Macro, Economic, Fundamental, To the nation, Line, Master, Disregard
	Score: 사업자, 결국, 번호, 중요, 시스템, 인덱스, 펀드 Businessman, Finally, Number, Important, System, Index, Fund
Topic 5	Highest prob: 리테일, 정도, 기준, 평, 전체, 구분, 서울 Retail, Degree, Criteria, Pyeong, Entire, Section, Seoul
	FREX: 리테일, 정도, 평, 구분, 서울, 규모, 집합 Retail, Degree, Pyeong, Section, Seoul, Size, Collective
	Lift: 도입, 스퀘어, 오피스텔, 전면, 중대형, 퍼센트, 구분 Introduction, Square, Efficiency apartment, Front, Mid-large size, Percentage, Section
	Score: 리테일, 평, 정도, 기준, 규모, 집합, 구분 Retail, Pyeong, Degree, Criteria, Size, Collective, Section
Topic 6	Highest prob: 시장, 국가, 분기, 상황, 평가, 자산, 한국 Market, National, Quarter, Situation, Evaluation, Asset, Korea
	FREX: 국가, 분기, 승인, 가치, 확인, 오픈, 상황 National, Quarter, Approval, Value, Confirmation, Open, Situation
	Lift: 걱정, 분자, 운용, 의사, 자리, 정량, 지적 Worry, Molecule, Operation, Intention, Place, Quantity, Point
	Score: 시장, 국가, 평가, 운용, 분기, 자산, 상황 Market, National, Evaluation, Operation, Quarter, Asset, Situation
Topic 7	Highest prob: 조사, 표본, 민간, 지금, 가지, 대상, 방법 Investigation, Sample, Private, Now, Things, Target, Method
	FREX: 조사, 표본, 추가, 설계, 인더스트리얼, 과거, 이슈 Survey, Sample, Addition, Design, Industrial, Past, Issue
	Lift: 물량, 설계, 세빌스, 임의, 협조, 과거, 기사 Volume, Design, Savills, Random, Cooperate, Past, Article
	Score: 조사, 표본, 민간, 지금, 지출, 설계, 이슈 Survey, Sample, Private, Now, Expense, Design, Issue
Topic 8	Highest prob: 오피스, 임대료, 건물, 층, 단위, 이상, 지방 Office, Rent, Building, Floor, Unit, More, Non-Seoul area
	FREX: 오피스, 임대료, 지방, 얼마, 계약, 매출액, 단위 Office, Rent, Non-Seoul area, Amount, Contract, Sales, Unit
	Lift: 보통, 인상, 학원, 강남대로, 갱신, 건축물, 군 Normal, Increase, Academy, Gangnam-daero, Renewal, Building, Country
	Score: 임대료, 층, 오피스, 건물, 이상, 단위, 지방 Rent, Floor, Office, Building, More, Unit, Non-Seoul area
Topic 9	Highest prob: 상권, 지역, 문제, 업무, 가능, 시설, 업종 Commercial district, Region, Problem, Official, Availability, Facilities, Industry
	FREX: 상권, 지역, 업무, 업종, 권역, 설정, 용도 Commercial district, Region, Official, Type of business district, Establishment, Use
	Lift: 그림, 상업시설, 업데이트, 용지, 지가, 크기, 권역 Painting, Commercial facility, Update, Plot, Land price, Size, Type of business district
	Score: 상권, 지역, 시설, 권역, 문제, 설정, 업무 Commercial district, Region, Facility, Type of business district, Problem, Setting, Official
Topic 10	Highest prob: 투자, 수익, 투자자, 방식, 빌딩, 임차인, 비용 Investment, Revenue, Investor, Method, Building, Tenant, Cost
	FREX: 투자, 수익, 투자자, 방식, 빌딩, 임차인, 비용 Investment, Revenue, Investor, Method, Building, Tenant, Cost
	Lift: 감정평가, 경험, 리스크, 방식, 밸류, 외국, 이사 Appraisal, Experience, Risk, Method, Value, Foreign, Director
	Score: 수익, 투자, 빌딩, 방식, 임차인, 비용, 경험 Revenue, Investment, Building, Method, Tenant, Cost, Experience