2019년도 연구용역보고서

빅데이터를 활용한 경제데이터 분석 사례 및 방법론 연구

- 2019. 10. -

이 연구는 국회예산정책처의 연구용역사업으로 수행된 것으로서, 보고서의 내용은 연구용역사업을 수행한 연구자의 개인 의견이며, 국회예산정책처의 공식 견해가 아님을 알려드립니다.

연구책임자

상명대학교 유경원

빅데이터를 활용한 경제데이터 분석 사례 및 방법론 연구

2019. 10.

책 임 연 구 원: 유경원 (상명대학교)

공 동 연 구 원: 최동욱 (상명대학교)

연 구 보 조 원: 정지수 (상명대학교)

이 연구는 국회예산정책처의 연구 용역사업으로
수행된 것으로서, 본 연구에서 제시된 의견이나 대안 등은
국회예산정책처의 공식의견이 아니라 본 연구진의 개인 의견임.

국회예산정책처
National Assembly Budget Office

제 출 문

국회예산정책처장 귀하

본 보고서를 귀 국회예산정책처의 연구과제
「빅데이터를 활용한 경제데이터 분석 사례 및 방법론 연구」의
최종 보고서로 제출합니다.

2019. 10.

상명대학교 유경원 교수

요 약

Ⅰ. 서론

❑ 경제정책 분석에 빅데이터를 활용하는 방안에 대해 최근 많은 논의가 이루어지고 있음

- ㅇ 경제분석과 관련하여 빅데이터는 기존의 정형화된(structured) 데이터를 보완하는 목적으로 활용 가능

❑ 일반적으로 빅데이터는 양적으로 방대한 데이터이며 기존과는 다른 새로운 소스를 통해 생성된 데이터를 지칭

- ㅇ 빅데이터는 만들어진 데이터(made data)가 아닌 발견된 데이터(found data)로서 목적과 설계에 있어서 연구진의 개입이 없다는 특징이 있음(Connelly et al., 2016)

❑ 빅데이터의 장점은 다음과 같은 네 가지로 요약될 수 있음(Stephens-Davidowitz, 2017)

- ㅇ 기존의 정형화된 데이터와 달리 이미지, 텍스트 등 새로운 유형의 데이터를 제공하여 현상에 대한 추가적인 정보와 문제해결이 가능
- ㅇ 이용자가 자발적으로 입력한 포털 검색어는 '디지털 자백약(Truth serum)'이라 불릴 정도로 솔직한 자료를 제공함으로써 조사자료에서 발생할 수 있는 응답편의가 작음[1)]
- ㅇ 작은 집단도 확대해 볼 수 있는 '디지털 확대' 기능 제공
- ㅇ 큰 비용이나 복잡함 없이도 인과관계 파악을 위한 실험의 시행도 가능

❑ 이와 같은 다양한 유형의 유용성이 확인되어짐에 따라 민간을 중심으로 활용되던 빅데이터 분석이 공공 부문에서도 많은 관심의 대상이 되고 있음

- ㅇ 텍스트 데이터의 경우 숫자로 표현하기 힘들었던 새로운 정보들을 제공해줄 수 있으며 경제정책 분석에 대한 활용이 점차 증가하고 있음(이영준 외, 2019)

❑ 본 연구에서는 경제정책 분석에 있어서 국내외적으로 빅데이터가 활용되고 있는 다양한 사례를 살펴보고 앞으로 국회예산정책처의 정책평가와 수립에 참고할 수 있는 시사점을 도출

1) 포털 검색어는 동기 부여된(motivated) 이용자가 스스로 입력한 것이기 때문에 관련된 주제에 대한 솔직한 내용을 담는다는 의미임

Ⅱ. 빅데이터 분석의 국내외 활용 현황

❑ 본 절에서는 기존 연구를 자료의 출처에 따라 크게 텍스트분석, 행정자료, 위치정보, 이미지 정보 및 민간거래정보로 구분하여 정리함

1. 해외사례

❑ (텍스트) 텍스트 자료를 이용하여 경제문제를 분석한 해외의 연구들은 웹 검색어 빈도분석, 소셜미디어 텍스트분석, 뉴스기사분석 등 다양한 방법을 활용

- ○ 검색어를 이용해 실업률 및 실업수당 청구건수, 내구재 수요 등을 예측하거나 FOMC회의록을 통해 금리수준 결정에 대한 논조를 정량화한 사례(McLaren and Shanbogue, 2011; Choi and Varian, 2009, 2012; Hansen and McMahon, 2016)

❑ (행정자료) 해외 연구 중에서 행정데이터를 연계하여 분석한 대표적인 사례는 Chetty et al.(2011), Cellini and Turner(2016), 그리고 Armour and Hung(2017)가 있음

- ○ Chetty et al.(2011)은 과거 성적자료와 국세청의 소득자료를 연계하여 초·중학교 시절 교사들의 역량이 학생의 사회진출 후 소득에 미치는 영향을 분석
- ○ Cellini and Turner(2016)는 설문조사 자료와 사회보장 정보를 연계하여 연금정책이 은퇴계획에 미치는 영향을 확인
- ○ Armour and Hung(2017)는 정부 학자금 지원 자료와 학생의 취업 및 소득 정보를 연계하여 사립대학 교육의 성과를 분석

❑ (위치 & 이미지) 지도, 위치, 이미지 등의 빅데이터를 활용한 사례로는 Dunleavy(2016), Glaeser et al.(2018) 등이 있음

- ○ 특히 Glaeser et al.(2018)은 Google Street View의 이미지를 분석하여 도시의 소득분포 예측 및 개발도상국의 부와 빈곤 수준을 확인

❑ (민간거래정보) 민간거래 빅데이터를 활용하여 분석한 해외 사례로는 Cohen et al.(2016)이 대표적임

○ Uber가 가지고 있는 고객의 이용 정보를 활용하여 Uber서비스에 대한 소비자의 가격탄력성을 추정하고 이를 통해 Uber가 미국 주요 4개 도시에서 창출한 소비자 잉여를 약 29억 달러로 추청

2. 국내사례

❑ (텍스트) 빅데이터 분석을 활용한 국내 사례로는 최동욱(2017), 이영준 외(2019)를 참고할 수 있음

○ 최동욱(2017)은 국회회의록 분석을 통해 정치성향을 측정하고, 이를 이용하여 포털뉴스가 정치적 성향을 가지는 유인이 소비자의 성향에서 비롯됨을 확인

○ 이영준 외(2019)는 금통위 전후 3일간의 뉴스 기사에 대한 텍스트 분석을 통해 통화정책 커뮤니케이션이 경제에 미치는 영향을 확인

❑ (행정자료) 국내 연구 중에서 행정데이터를 연계하여 분석한 대표적인 최근 사례는 이태리 외(2018)를 들 수 있음

○ 개인신용정보와 주택실거래자료를 연계하여 LTV(주택담보대출비율) 및 CoLTV(실효LTV) 지표를 산출하고 임대차주의 상환위험을 분석

❑ (위치정보) 교통망, 거리 등 위치 관련 빅데이터를 활용해 분석한 국내연구 사례

○ 박종수·임금숙(2015)은 교통카드 DB를 기반으로 승객이 이동하는 시간거리를 산출하고 교통망이 제공할 수 있는 목적지와의 접근성을 산출

❑ (이미지) 이미지 데이터를 구축하여 경제 분석에 활용한 사례로는 김규철(2017)이 있음

○ 인공위성 이미지를 통해 북한의 야간 조도를 측정하고 이를 이용하여 북한주민의 후생수준을 파악

❑ (민간거래정보) 신용카드 데이터 등 민간거래 빅데이터를 활용하여 분석한 최근 사례로는 김경근·염병배(2017)가 있음

○ 신용카드 이용자료를 통해 지역별 소비수준을 파악함에 있어서 거주자와 비거주자의 비교를 통해 소비유입률 등을 측정하고 지역에서의 생산 및 고용효과를 분석

Ⅲ. 빅데이터 분석기반 경제분석 방법론

1. 빅데이터 적용 가능성의 검토

❑ 해결하고자 하는 문제에 가장 적합한 자료를 선정하는 것이 중요함(Stephens-Davidowitz, 2017)

- ㅇ 특히 빅데이터의 경우에는 그 자체로 주요한 결론을 도출하기보다는 기존 데이터를 보완하는 방식으로 활용하는 것이 바람직함
- ㅇ 문제 해결을 중심에 놓고 판단하면 효과적인 데이터를 찾는데 도움이 될 수 있음

❑ 따라서 해결하고자 하는 문제에 대해 구체적으로 정의하는 작업이 우선임

- ㅇ 어떤 문제를 해결하고자 하는지 명확히 할 필요가 있음
- ㅇ 이러한 문제의 성격에 따라 크게 두 가지 종류로 구분 할 수 있음
  - – 속보성과 예측이 중요한 경우로 빅데이터를 활용한 지표를 생성하고 이를 이용하여 현실을 설명(상관관계)
  - – 사후적으로 심층평가가 필요한 경우로 현실의 경제 모형에 빅데이터를 활용한 지표가 포함(인과관계)

❑ 문제가 정의되었다면 문제를 해결하기 위해 필요한 변수를 파악하고 빅데이터의 활용이 필요한지 검토해야 함

- ㅇ 어떤 문제가 빅데이터를 활용해야 하는 문제라면 다음과 같은 성격을 가지고 있기 때문이라고 생각할 수 있음
  - – 기존의 자료로는 풀 수 없는 새로운 문제(예: 청소년기의 학업성취가 성인기의 소득수준에 미치는 영향)
  - – 정성적 정보의 정량화(객관적)가 필요한 경우(예: 금융통화위원회의 금리수준 결정에 대한 논조)

2. 속보성과 예측이 중요한 경우

❑ 속보성이나 예측이 중요한 문제를 해결하기 위해서는 실시간으로 수집되는 데이터를 이용할 수 있음

- ㅇ 포털 검색어나 SNS텍스트 자료 등 실시간으로 생성되는 데이터를 이용하여 현실경제를 예측하는데 활용할 수 있음

❑ 실시간 데이터와 경제변수 간의 상관관계가 성립함을 전제로 실시간 예측 가능

- ㅇ 경제지표의 변동(Choi and Varian, 2012)
  - TV와 냉장고에 대한 검색어가 일반 내구재에 대한 수요와 유사한 추이
- ㅇ 불확실성 지표(Baker, Bloom and Davis, 2013)
  - 언론기사의 불확실성이라는 단어를 언급한 횟수와 현실경제의 관계
- ㅇ GDP나우캐스팅(FRB of New York)
  - 다양한 소스의 자료를 활용하여 매주 GDP추정치를 제공

❑ 시사점

- ㅇ 포털 검색어나 SNS텍스트 등은 실시간으로 생성된다는 점과 동기 부여된 이용자의 자발적인 행위의 결과라는 점에서 신속성이 높고 편의성이 낮은 통계자료로서의 장점을 가지고 있음
- ㅇ 이러한 빅데이터의 특성을 활용하여 실제 경제현상과 높은 상관관계를 가지고 있는 변수를 생성하고 단기 혹은 중장기 예측에 적극적으로 활용할 필요

3. 사후평가 및 심층분석이 필요한 경우

❑ 사후평가나 심층분석은 주로 과거에 도입된 특정한 정책이 경제현상에 미친 인과관계를 사후적으로 엄밀히 검증하는 작업

- ㅇ 단순 상관관계를 파악하는 경우와 달리 경제 이론에 근거한 분석 모형이 제시되어야 함
- ㅇ 분석 모형을 검증할 실증 데이터를 수집하는 과정에서 빅데이터(디지털 기록이나 행정데이터 등)를 이용한 지표를 생성

- ㅇ 이중차분법이나 도구변수추정법 등 인과관계를 파악할 수 있는 계량기법을 적용하여 모형의 파라미터를 추정

❑ 샘플링 기법 등 자료의 대표성이 중요

- ㅇ 행정자료의 연계: 행정데이터는 기존의 조사자료와 달리 샘플링 편의의 문제가 적어 높은 신뢰성을 가짐

❑ 인과관계 분석방법

- ㅇ 가상적인 실험상황을 구성
  - 관측대상을 정책의 영향을 받는 처치군과 영향이 없는 통제군으로 구분하고 정책시행 전후의 변화를 관측
- ㅇ 패널데이터를 이용한 이중차분법
  - 패널자료를 구성할 수 있다면 패널분석 계량모형을 이용하여 정책시행의 효과를 분석할 수 있음

❑ 이처럼 빅데이터를 활용하여 인과관계를 파악하는 심층적 분석이 가능

- ㅇ 행정데이터의 연계를 통해 개인수준의 구체적인 자료를 구축할 수 있으며 이를 이용하여 정책효과나 가격변화의 영향 등을 정확하게 추정할 수 있음
  - 다만 개인정보보호의 문제가 존재하므로 이에 대한 고려가 반드시 필요
- ㅇ 텍스트분석방법을 이용하여 기존에는 알 수 없었던 새로운 정보를 파악할 수 있으며 이를 통해 새로운 통찰력을 제공
- ㅇ 그 밖에도 다양한 소스의 빅데이터를 활용하는 것이 가능하지만 심층분석을 위해서는 일반적인 정형자료들과 연계 및 정치한 경제모형이 필요

❑ 시사점

- ㅇ 단순 상관관계가 아닌 인과관계를 실증적으로 보여줄 수 있는 방법론을 적용하는데 있어서 빅데이터의 활용을 통해 실제 경제현상이나 정책효과에 대한 문제 중 기존에 해결하지 못했던 문제를 해결할 수 있는 가능성이 열림
- ㅇ 경제문제의 해결에 있어서 행정데이터와 연계를 통해 대표성을 확보하고 상당히 구체적인 수준의 개별 정보를 추적할 수 있다는 장점이 있기 때문에 이를 활용하여 정책효과를 분석하는데 기여할 수 있는 잠재력이 존재함

○ 특히 텍스트마이닝 등 비정형데이터를 활용하여 기존에 측정하지 못했던 현상에 대한 지표를 생성할 수 있음

Ⅳ. 경제분야 빅데이터 분석의 적용 사례

1. 소셜 빅데이터 분석: 가계부채 이슈를 중심으로

1) 개인회생 및 개인파산 신청 분석

❑ 본 소절에서는 검색자료의 장점과 한계를 인식하고 보다 정교한 작업을 위해 단계적으로 기술분석과 시계열회귀분석을 수행함

○ 시계열회귀분석에 앞서 주요 변수들의 단위근 검정을 수행하고 McLaren and Shanbhogue (2011)를 원용하여 기본 분석모형을 다음과 같이 설정함

$$R_t = \alpha + \beta_1 R_{t-1} + \beta_2 SR_{t-r} + X_t \Gamma + D_m + \epsilon_t$$

○ 여기서 R은 개인회생 및 개인파산의 월 신청건수 그리고 SR은 이와 같은 개인회생 및 개인파산의 월 검색지표, X는 이와 같은 채무조정을 설명하는 경제변수들의 벡터로 은행연체율, 은행대출금리, 실업률 등을 포함하고 D는 월별(m) 더미변수를 의미

❑ McLaren and Shanbhogue(2011) 방법을 원용하여 검색지수의 추가 설명력을 분석한 결과 개인회생의 경우 어느 정도의 설명력을 가지고 있는 것으로 나타남

❑ 회귀분석에서 추정한 결과를 가지고 모형의 적합도를 살펴본 결과 아래 그림에서와 같이 실제치를 잘 추정하고 있는 것으로 판단됨

[그림] 개인회생 실제치와 모형 추정치 추이

❑ 개인회생에 대한 검색기록은 가계재무 상태의 건전성을 평가할 때 중요한 지표로 활용될 수 있을 것임

- ㅇ 검색기록에 대한 현재와 시차변수 등은 개인회생을 추정오차 5% 내외에서 추정할 수 있으며 속보성 있는 자료의 속성상 현재의 가계재무 건전성에 대한 유용한 판단지표를 제시할 수 있는 것으로 판단됨

❑ 다만 파산에 대한 검색기록은 좀 더 정교한 추정기법과 자료를 활용하여 추가적인 연구를 통해 개선할 필요가 있음

- ㅇ 향후 시계열이 확장된 네이버 검색기록 등 다양한 검색자료와 기법을 활용하여 추가적인 연구를 통해 관계를 추정해 낼 수 있을 것으로 판단됨

2) 전략적 개인파산 분석

❑ 가계부채가 지속적으로 증가함에 따라 채무불이행 등에 따라 향후 사적·공적 채무조정제도의 활용이 늘어나게 되고 채무조정제도를 어떤 식으로 운영하느냐에 따라 가계부채발 경제위기의 영향이 달라질 수 있음

❑ 따라서 본 절에서는 지속적으로 늘어나고 있는 가계부채가 부실화되었을 때 채무조정제도의 기조를 어떻게 운영할지에 대한 시사점을 얻기 위하여 개인들의 전략적 또는 외생적 파산 행태에 대한 분석을 검색자료를 이용하여 살펴봄

❑ 소셜빅데이터의 장점이 보다 '솔직한' 검색기록을 제시하고 있으므로 연관어 분석 등을 통해 전략적인 행동과 외생적인 파산 행동을 식별하고 이들의 추이를 살펴봄

- ㅇ 인터넷 상에서 개인파산 검색 시 이로 인한 편익, 비용 등을 함께 검색한다면 전략적 파산의 의도가 있는 것으로 판정하며, 그렇다면 여기에서의 관건은 '전략적'이라고 판별할 수 있는 연관어의 선택임

❑ 검색기록의 일별 추이를 통해 전략적인 그리고 외생적인 파산 가능성 추이를 살펴볼 수 있으며, 성별, 연령대별 검색기록을 통해 이와 같은 파산 가능성의 인구사회학적 특징을 일부 파악할 수 있었음

❑ 향후 이와 같은 검색기록을 통해 파산신청과 관련된 질적 분석이 가능할 수 있으며 이를 기반으로 개인채무구제정책의 방향성을 설정하는데 있어 참고자료로 활용할 수 있을 것임

- ㅇ 현재 가계부채 취약가구가 지속적으로 늘어나고 있는 가운데 향후 이와 같은 검색기록에 대한 모니터링을 통해 향후 개인채무구제정책 방향을 설정하는데 참고자료로 활용

3) 주택담보대출 수요 분석

❑ 본 절에서는 한국은행 대출태도조사와 관련된 한계를 극복하기 위해 검색결과를 활용하여 대출수요 조사를 대체할 수 있는지를 살펴봄

- ㅇ 본 절에서는 가계부채의 핵심이라고 할 수 있는 주택담보대출과 관련된 태도조사 결과와 검색기록을 활용한 수요조사 결과 그리고 실제 주택담보대출 증가율간의 관계를 분석함

❑ 본 연구에서는 '주택담보대출 금리' 검색기록을 실질적인 가계 주택담보대출 수요로 파악하고 기존의 대출태도조사에서 나타난 대출수요와의 비교를 통해 그 유용성을 파악하고자 할 것임

- 실질적으로 주택담보대출 증가율과 이들 변수들이 어떤 관계가 있으며 검색기록이 이와 같은 대출증가율에 추가적인 정보를 제공할 수 있는지를 살펴봄

❑ 대출태도조사에서 파악되는 가계대출태도 및 가계대출수요 지표에 비해 가계대출금리 검색기록은 가계대출수요를 반영하는 지표로서 가계대출증가율을 보다 잘 설명하는 것으로 판단됨

- 분기별로 발표되는 대출태도조사는 가계대출증가율을 제대로 반영하지 못하는 것으로 나타나고 오히려 대출검색지표가 상대적으로 설명력이 높은 것으로 나타남

❑ 월별자료 분석에 있어서는 대출수요에 직접적인 영향을 미치는 대출금리의 영향이 큰 것으로 나타나고 있으나 대체적으로 대출수요를 속보성 있게 파악할 수 있는 것으로는 이와 같은 검색기록이 유용할 수 있음을 확인

2. 텍스트마이닝 적용사례: 금융통화위원회 회의록 분석

❑ 2019년 금융통화위원회 의사록을 이용하여 공개적으로 발표된 수치 정보가 아닌 새로운 관점으로 금융통화정책을 파악해볼 수 있는 방안을 소개하고자 함

- 금통위 의사록 분석을 통해 파악할 수 있는 한가지 유용한 정보는 금리인상과 인하에 대한 예측으로 소위 매파와 비둘기파의 의견을 구분하고 우세를 비교하는 것임(한국은행, 2017)
- 분석 사례로 제13차 금융통화위원회(2019년 7월 18일)의 회의자료를 이용

❑ 전처리과정(pre-processing)

- 텍스트자료로부터 형태소를 분리 추출하고 명사, 형용사, 동사 등 문장 내에서의 역할을 파악함
- 분석과 관련이 없는 불용어를 삭제하고 길이가 과도하게 짧거나 긴 단어들은 제거함

❑ 추출된 단어들의 빈도를 계산하여 분석하면 관심을 가진 주제를 파악할 수 있음

- 단어를 빈도순으로 나열하고 이를 바그래프로 표현하거나 워드클라우드로 표현가능
- 금통위의 회의에서 가장 자주 등장하는 단어는 ‘전망’, ‘경제’, ‘금리’ 등으로 나타남
- 경기부진과 둔화에 대한 언급이 많다는 것을 확인할 수 있음

❑ 토픽모델은 주어진 문서에서 유사한 용도나 의미를 갖는 단어들의 그룹을 식별하여 분류하는 방법으로 일종의 머신러닝 기법임

- ㅇ 단어들 간의 상관관계 및 유사성을 이용하여 통계적 방법 통해 단어의 그룹을 형성하고 분류하는 기법
- ㅇ 본 연구에서 활용할 방법은 잠재 디리클레 할당(Latent Dirichlet Allocation, 이하 LDA) 기법임
- ㅇ 분석결과(아래 표 참조)
  - – 일반적으로 금통위의 회의에서는 경기 상황 평가 및 금리수준에 대한 판단에 대해 논의
  - – 두 그룹으로 분류한 결과 중, 첫 번째 토픽 그룹은 대체로 경제에 대한 평가(부정적)와 관련된 단어들로 보이고, 두 번째 그룹은 그 이외의 정책관련 단어들로 보임

[표] 토픽모델 분석 결과

토픽 1						토픽 2					
전망	부진	둔화	지속	금년	투자	기준금리	가계부채	언급	금융시장	가격	
증가세	성장률	수출	가운데	예상	성장세	필요	일부	평가	유의	의견	충격
무역분쟁	미	중	교역	모습	상당	안정	정책	견해	공급	동향	당시
세계	불확실	소비	기업	하방	완화	금리	나라	건전	측면	일반	발생
중심	고용	상승률	제조업	정부	지난해	통화정책	하기	고려	금번	부동산	규제
감소	작용	내년	양호	올해	수요	분석	결과	시장	현상	완화적	불균형
압력	증가율	우려	위축	흐름	약화	견해	주택	장단기	강화	다양	생각
확대	반면	소득	일본의	감소세	예상	당부	부채	때문	중앙은행	상대	중요
소비자물가	완만	개선	민간소비	보호무역주의		이번	결정	거시	상황	제시	초래
달러	상반기	회복	하반기	재정정책	가계	이슈	문제	과거	연구	유동성	점검
부문	무역협상	심화	경로	심리	전환	대내외	만큼	외환	인플레이션		
생산	gdp	당초	하회	국의	작년						
반등	소폭	마이너스	실적	확산	물가상						
승률	수출규제										

❑ 토픽모델은 파라미터 등 일정 조건이 주어진 상황에서 인간의 자의적 판단이 아닌 자료의 성격 즉, 텍스트 안의 단어 구성만을 이용해서 분류 결과를 제시한다는 장점이 있음

- ㅇ 인간의 눈으로는 미처 판단하지 못한 잠재된 맥락과 뉘앙스를 발견할 수 있으며 사람이 직접 눈으로 읽고 판단하기보다는 통계적 기법을 통해 좀 더 객관적으로 측정할 수 있다는 점이 중요

- 물론 파라미터의 설정에 있어서는 인간의 판단이 개입될 필요가 있으며 신뢰성 있는 분류 결과를 얻기 위해서는 다양한 파라미터 조합의 변화에 대해 안정적인(robust) 분류가 이루어지는지 확인할 필요가 있음

Ⅴ. 결론 및 시사점

❑ 최근 주요 정책 분석기관에서 빅데이터 분석에 대한 수요가 커지고 있는 만큼 본 연구를 기반으로 국회예산정책처에서 경제분석 업무에 실제 활용할 수 있는 방안 등을 마련할 수 있을 것임

- 다양한 유형의 자료를 이용, 실제 경제분석에 활용할 수 있는 기반을 마련하고 이를 통해 제도 개선 및 국회 의정활동 지원에 활용될 수 있는 정책 시사점 등을 도출하는데 활용될 수 있을 것임

❑ 국회예산정책처에서도 향후 빅데이터 분석을 적극 활용할 필요가 있음

- 본 연구결과에서 보는 바가 빅데이터 활용이 경제정책 분야에 다양하고 유용하게 활용될 수 있으므로 향후 국회예산정책처에서도 빅데이터 활용을 적극 모색할 필요
- 이미 많은 연구기관이나 정책기관에서 빅데이터 활용을 위해 내부적인 연구와 조직정비가 이루어지고 있는 실정임

❑ 구체적으로 빅데이터 업무활용은 속보성과 예측이 중요한 경우와 사후평가 및 심층분석이 필요한 경우로 나누어 볼 수 있으므로 담당 업무와 과제 성격에 따라 거시적인 분석과 빠른 대응이 필요한 경우 검색어 기반 분석과 소셜미디어 텍스트 분석이 이루어질 필요가 있을 것임

- 검색어 기반 연구를 통해 주요 거시경제변수에 대한 속보성 있는 예측 등이 비교적 손쉽게 이루어질 수 있을 것임
- 거시정책 분석 내지 평가 목적에 있어서는 정도 높은 데이터의 확보가 필요하므로 이를 위해서는 거래자료 기반 민간 빅데이터와 행정데이터와 같은 공공 빅데이터의 활용이 함께 이루어질 필요가 있음

❑ 다음으로 정책에 대한 사후 평가나 심층분석이 필요할 경우 연구 이슈 등에 부합되는 빅데이터를 선정하고 적절한 분석기법을 적용하여 분석할 필요가 있음

- ㅇ 예를 들어 Ⅳ장에서 제시한 텍스트마이닝 기법을 통해 그동안 분석하기 어려웠던 비정형화된 정성적인 정보들의 자료를 활용하여 경제정책에 대한 평가 내지 분석이 이루어질 수 있을 것임

❑ 향후 국회예산정책처의 빅데이터 이용 활성화와 관련하여 다음과 같은 단계별 접근이 필요할 것임

- ㅇ 외부전문가와 공동 연구를 통해 빅데이터 활용과 관련된 다양한 주제를 발굴하고 이를 내부 부서와 공유함으로써 관련 빅데이터 활용을 보다 확산시킬 필요가 있음
- ㅇ 외부 전문가 등과 1~2년간의 빅데이터 활용에 대한 기초연구를 기반으로 빅데이터를 활용한 핵심 연구주제를 발굴하여 자체 연구와 과제를 수행하도록 함
  - – 두 트랙으로 진행될 수 있으며 먼저 정량적인 자료 기반 연구 강화를 위해서는 공공부문 빅데이터를 활용할 수 있도록 다양한 공공기관과의 협업 내지 업무 협조를 진행할 필요가 있음
- ㅇ 4차 산업혁명 시대가 도래함에 따라 중요한 한 축이라고 할 수 있는 빅데이터의 활용 가능성이 높아지고 이에 따라 빅데이터의 접근성이 확대되고 있으나 아직까지도 개인정보보호 이슈 등 법적 제한이 큰 상태이므로 이에 대한 해결 노력이 긴요함
- ㅇ 또한 정성적인 자료 기반 연구는 비정형데이터를 활용하여 속보성 있는 미래 경제 변수 예측이나 심도 깊은 경제분석에 활용될 수 있으므로 내부 전문가 그룹의 전문성과 이를 기반으로 한 통찰력(insight)을 활용할 필요가 있음

- 목 차 -

Ⅰ. 서론 ····· 1

Ⅱ. 빅데이터 분석의 국내외 활용 현황 ····· 5

1. 해외사례 ····· 6

2. 국내사례 ····· 12

Ⅲ. 빅데이터 분석기반 경제분석 방법론 ····· 17

1. 빅데이터 적용 가능성의 검토 ····· 17

2. 속보성과 예측이 중요한 경우 ····· 19

3. 사후평가 및 심층분석이 필요한 경우 ····· 23

Ⅳ. 경제분야 빅데이터 분석의 적용 사례 ····· 29

1. 소셜 빅데이터 분석: 가계부채 이슈를 중심으로 ····· 29

1) 개인회생 및 개인파산 신청 분석 ····· 31

가. 검토배경 ····· 31

나. 방법론 및 분석결과 ····· 35

다. 시사점 ····· 44

2) 전략적 개인파산 분석 ····· 45

가. 검토배경 ····· 45

나. 방법론 및 분석결과 ····· 46

다. 시사점 ····· 53

3) 주택담보대출 수요 분석 ····· 55

가. 검토배경 ····· 55

나. 방법론 및 분석결과 ····· 59
다. 시사점 ····· 66
2. 텍스트마이닝 적용사례: 금융통화위원회 회의록 분석 ····· 67
가. 개요 ····· 67
나. 빈도분석 및 워드클라우드 ····· 69
다. 토픽모델링의 활용 ····· 75
다. 시사점 ····· 77

Ⅴ. 결론 및 시사점 ····· 78

참 고 문 헌 ····· 83

<부록 1> 국내 연구기관 빅데이터 관련 연구 현황 ····· 87
<부록 2> 빅데이터 분석도구 소개 ····· 90
<부록 3> 국내 부동산 및 가계부채 관련 정책 일지 ····· 108
<부표> ····· 113
<부도> ····· 120

- 표 목 차 -

[표 Ⅱ-1] 빅데이터의 분류 ····· 5
[표 Ⅱ-2] 미국 행정자료 이용 연구사례 ····· 10
[표 Ⅲ-1] 실험상황의 구성 ····· 24
[표 Ⅳ-1] 인터넷 사용 추이(통계청 사회지표) ····· 31
[표 Ⅳ-2] 네이버와 구글 점유율 평균 추이(2010~2019) ····· 36
[표 Ⅳ-3] 채무조정제도 실적과 검색기록과의 상관관계 ····· 38
[표 Ⅳ-4] 개인회생과 개인파산의 단순회귀분석 결과 ····· 39
[표 Ⅳ-5] 시차변수 추가한 회귀분석 결과 ····· 39
[표 Ⅳ-6] 개인회생 추정 결과(1) : 검색기록 설명변수 ····· 40
[표 Ⅳ-7] 개인회생 추정 결과(2) : 경제변수 추가 ····· 42
[표 Ⅳ-8] 모형별 예측정확도 결과 ····· 44
[표 Ⅳ-9] 개인파산 연관 검색어(Google Trends) ····· 47
[표 Ⅳ-10] 개인파산 주요 검색 지표 및 지수(일별) 요약 통계량 ····· 49
[표 Ⅳ-11] 성별 및 연령대별 전략파산지수 요약 통계량 ····· 50
[표 Ⅳ-12] 성별 및 연령대별 외생파산지수 요약 통계량 ····· 51
[표 Ⅳ-13] 성별 및 연령대별 외생파산지수 대비 전략파산지수 요약 통계량 ····· 51
[표 Ⅳ-14] 주요 지수와 연령 관계 회귀분석 결과 ····· 52
[표 Ⅳ-15] 대출수요지수와 대출수요 검색지수를 이용한 대출증가율 회귀분석 결과 ····· 61
[표 Ⅳ-16] 월별 검색자료를 활용한 가계주택담보대출 증가율 회귀분석 결과 ····· 64
[표 Ⅳ-17] 토픽모델 분석 결과 ····· 77

- 그 림 목 차 -

[그림 Ⅰ-1] 조사자료와 행정자료 연계를 이용한 연구 증가 추세 ····· 3
[그림 Ⅱ-1] 영국의 실업수당 검색 빈도수 활용사례 ····· 6
[그림 Ⅱ-2] 미국과 유럽의 뉴스기사 활용사례 ····· 7
[그림 Ⅱ-3] Chetty et al.(2011)의 자료 연계 개념도 ····· 8
[그림 Ⅱ-4] Armour and Hung(2017)의 자료 연계 개념도 ····· 9
[그림 Ⅱ-5] 2015년 국내의 주요 정치적 사건과 포털뉴스의 정치성향 변화추이 ····· 13
[그림 Ⅲ-1] 뉴욕 FRB의 GDP나우캐스팅 페이지 ····· 20
[그림 Ⅲ-2] 부동산 관련 키워드를 통한 연령대별 관심정도의 비교 ····· 21
[그림 Ⅲ-3] 부동산 보유세 인상에 대한 연관검색어(이창근 외, 2017) ····· 22
[그림 Ⅲ-4] 실험상황의 구성 ····· 24
[그림 Ⅲ-5] 자료연계 개념도 ····· 26
[그림 Ⅳ-1] 가계부채(household debt)에 대한 언급 회수 ····· 29
[그림 Ⅳ-2] 가계신용 추이 ····· 32
[그림 Ⅳ-3] 처분가능소득 대비 가계부채 비율 ····· 32
[그림 Ⅳ-4] 금융자산 대비 금융부채 비율 ····· 32
[그림 Ⅳ-5] 은행 및 비은행금융기관 가계대출 연체율 ····· 33
[그림 Ⅳ-6] 개인파산, 개인회생 월별 신청 추이 ····· 34
[그림 Ⅳ-7] 인터넷 검색엔진 최근 점유율(2019.8.12.~2019.9.2.) ····· 36
[그림 Ⅳ-8] 개인파산과 개인회생 실적과 검색 추이 ····· 38
[그림 Ⅳ-9] 개인회생 실제치와 모형 추정치 추이 ····· 44
[그림 Ⅳ-10] 네이버 데이터랩 상에서의 개인파산 관련 검색 조건 ····· 48
[그림 Ⅳ-11] 전략파산과 외생파산 추이 ····· 50
[그림 Ⅳ-12] 대출태도조사와 주택담보대출 증가율 관계 ····· 56
[그림 Ⅳ-13] 주택대출 관련 대출태도 지수와 대출수요 관계 추이 ····· 57
[그림 Ⅳ-14] 주택담보대출 증가율과 검색기록간 관계 추이 ····· 60

[그림 Ⅳ-15] 월별 가계대출금리 검색과 대출증가율 관계 추이 ····· 63
[그림 Ⅳ-16] 빈도분석 결과 그래프 ····· 71
[그림 Ⅳ-17] 워드클라우드 결과 그래프 ····· 72

Ⅰ. 서론

❑ 경제정책 분석에 빅데이터를 활용하는 방안에 대해 최근 많은 논의가 이루어지고 있음

- ㅇ 경제분석과 관련하여 빅데이터는 기존의 정형화된(structured) 데이터를 보완하는 목적으로 활용 가능
  - ICT기술의 발전에 따라 텍스트데이터, 이미지데이터, 위치데이터 등 비정형(unstructured)데이터를 수집・처리 가능해짐
  - 통계청의 승인 데이터처럼 공식적인 수집과정을 거쳐 구성된 데이터는 정형화된 데이터로서 공신력을 가지고 있지만 그만큼 전달할 수 있는 정보량과 범위에 한계가 존재
  - 빅데이터를 적절히 이용한다면 기존의 정형 데이터만으로는 표현할 수 없었던 추가적인 정보를 얻을 수 있음
  - 다만, 빅데이터가 전통적인 데이터를 완전히 대체할 수는 없으며 기존의 데이터를 보완하는 의미로 활용하는 것이 적절
- ㅇ 빅데이터의 단점과 한계에도 불구하고 향후 빅데이터가 경제통계 작성 및 분석에 유용하게 활용될 가능성이 높기에 한국은행 등 경제정책기관에서 물가지표, 심리지표, GDP 관련 지표 등 다양한 분야에서 빅데이터 활용 방안에 대한 연구를 진행 중임(한국은행, 2017; 이영준 외, 2019)
  - 통계청, 한국은행 등 주요 경제·통계 기관에서는 빅데이터의 중요성을 인식하여 2~3년 전부터 관련 연구용역과제를 외부에 공모하고 공동연구를 통해 구성원들의 역량을 증진시키는 노력을 해오고 있으며 아울러 조직적인 측면에서도 빅데이터센터 및 빅데이터연구팀 등을 신설·운영하고 있음[2)]

❑ 일반적으로 빅데이터는 양적으로 방대한 데이터이며 기존과는 다른 새로운 소스를 통해 생성된 데이터를 지칭

- ㅇ 초창기의 논의는 소위 3V를 빅데이터의 특성으로 제시(Mayer-Schonberger and Cukier, 2013)

2) 보다 자세한 논의는 부록의 국내 연구기관 빅데이터 관련 연구 활동 현황을 참조

- Volume: 양적으로 규모가 큰 데이터
- Variety: 다양한 소스를 통해 얻어지는 데이터(텍스트, 이미지, 위치정보 등)
- Velocity: 실시간으로 현황을 파악할 수 있는 데이터

○ 방대한 데이터 규모에서 얻을 수 있는 장점은 데이터 샘플링의 필요성이 감소했다는 점이라 할 수 있음
  - Mayer-Schonberger and Cukier(2013)는 빅데이터 분석이 미시적 차원의 정확성을 포기하되 거시적 차원의 통찰을 얻는 과정이라고 주장

○ 빅데이터는 만들어진 데이터(made data)가 아닌 발견된 데이터(found data)로서 목적과 설계에 있어서 연구진의 개입이 없다는 특징이 있음(Connelly et al., 2016)

○ Stephens-Davidowitz(2017)는 빅데이터에 대한 논의에 있어서 가장 중요한 것은 '문제를 해결할 수 있어야 한다'는 점임을 강조
  - 새로운 데이터 뿐만 아니라 새로운 분석방식이 적용되어야 함
  - 분석의 목적은 문제를 더욱 효과적으로 해결하는 것이라는 의미

❑ 한편, 경제적 분석이나 정책 결정에 활용하기 위한 빅데이터의 특성에 대해서는 Dunleavy(2016)와 Einav and Levin(2014)의 논의가 대표적이라 할 수 있음

○ 빅데이터는 기존의 조사자료에 비해 훨씬 구체적이고(detailed), 시의적절하고(timely), 방대하다는(large) 특성을 가짐(Dunleavy, 2016)
  - 다양한 변수를 활용하여 더욱 세밀한 분석이 가능하다는 장점이 있음
  - 조사자료와 달리 응답편의(response bias)가 발생할 확률이 적다는 장점이 있음

○ 경제분석에서 빅데이터를 활용할 때의 가장 중요한 장점 중 하나는 Ceteris Paribus, 즉 'All other things equal'이라는 강한 가정을 완화하고 이질성을 분석에 포함시킬 수 있다는 점(Einav and Levin, 2014)이라 할 수 있음
  - 사회과학 분야에서는 이전 수행하기 어려웠던 인과관계 분석을 위한 실험 등 새로운 연구 설계가 가능하고 경제 정책 수립과 운용의 효율성을 높일 수 있음
  - 이와 같은 과정에서 빅데이터와 경제 이론의 결합 중요성이 강조됨

❑ 빅데이터의 장점은 다음과 같은 네 가지로 요약될 수 있음(Stephens-Davidowitz, 2017)

○ 기존의 정형화된 데이터와 달리 이미지, 텍스트 등 새로운 유형의 데이터를 제공하여 현상에 대한 추가적인 정보와 문제해결이 가능

○ 포털 검색어는 ‘디지털 자백약(Truth serum)’이라 불릴 정도로 솔직한 자료를 제공함으로써 조사자료에서 발생할 수 있는 응답편의가 작음[3)]

○ 작은 집단도 확대해 볼 수 있는 ‘디지털 확대’ 기능 제공

○ 큰 비용이나 복잡함 없이도 인과관계 파악을 위한 실험의 시행도 가능

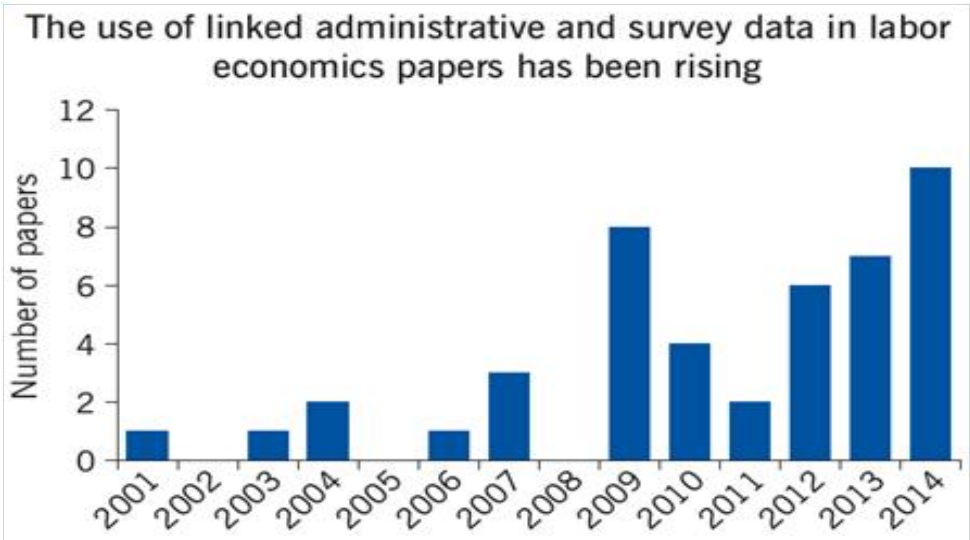
❑ 이와 같은 다양한 유형의 유용성이 확인되어짐에 따라 민간을 중심으로 활용되던 빅데이터 분석이 공공 부문에서도 많은 관심의 대상이 되고 있음

○ 민간부문에서는 주로 기업의 매출을 높이기 위한 마케팅 전략수립에 활용

○ 행정데이터의 활용 가능성을 중심으로 공공부문에서의 빅데이터 활용에 대해 관심이 증대

– 독일의 노동경제연구소인 IZA의 경우, 아래 그림에서 볼 수 있듯이 행정자료와 조사자료의 연계를 활용한 연구가 급격히 증가하는 추세를 확인할 수 있음

[그림 Ⅰ-1] 조사자료와 행정자료 연계를 이용한 연구 증가 추세



출처: Kunn(2015)

○ 텍스트 데이터의 경우 숫자로 표현하기 힘들었던 새로운 정보들을 제공해줄 수 있으며 경제정책 분석에 대한 활용이 점차 증가하고 있음

– 정부 회의록의 논조, 금융당국의 정책방향, 국회의원의 정치성향 등 정량화하기 힘들었던 정보들을 문서자료로부터 추출가능

3) 포털 검색어는 동기 부여된(motivated) 이용자가 스스로 입력한 것이기 때문에 관련된 주제에 대한 솔직한 내용을 담는다는 의미이며, 일반적인 조사자료는 응답자가 사회적으로 바람직한 내용으로 응답해야 한다는 생각으로 인해 편의가 발생할 수 있음

- 이러한 데이터는 향후 경제정책과 관련한 정책방향을 예측하고 분석하는데 다양하게 활용될 수 있음(이영준 외, 2019)

❑ 본 연구에서는 경제정책 분석에 있어서 국내외적으로 빅데이터가 활용되고 있는 다양한 사례를 살펴보고 앞으로 국회예산정책처에서 정책평가와 수립에 참고할 수 있는 시사점을 도출해보고자 함

ㅇ 빅데이터 활용과 관련하여 데이터의 유형, 방법론, 정책기관의 활용사례를 중심으로 하여 정리

❑ 이를 위해 본 연구에서는 우선 빅데이터를 이용한 경제분석 동향 및 활용사례를 조사하고 방법론을 정리한 뒤 경제이슈에 대한 분석 사례를 제시하고자 함

ㅇ 경제정책에 대한 빅데이터 분석 현황과 관련하여 해외의 연구사례와 국내의 연구사례를 조사하고 시사점을 도출

ㅇ 실제 우리 경제의 주요 이슈인 가계부채 문제를 빅데이터 분석을 통해 사례분석을 수행

❑ 선행연구와 방법론을 정리한 다음으로 실제 비정형 데이터를 수집 및 활용하여 주요 경제문제에 적용해보고자 함

ㅇ 소셜 빅데이터를 활용하여 주요 경제정책 이슈 중 하나인 가계부채 이슈, 특히 개인파산, 개인회생 등 채무자구제제도 그리고 대출 수요와 관련된 이슈 들을 분석함

ㅇ 대체적으로 가계부채 분석이 속보성이 떨어지는 부채 총량자료나 정확성이 상대적으로 떨어지는 가계 서베이 등 미시자료를 이용하는 경우가 많아 분석에 한계가 있어 입수가능한 시계열 자료와 결합하여 빅데이터 분석을 보완적으로 수행하여 속보성 있고 정확도 높은 관련 이슈 분석을 수행할 수 있는 방안을 모색함

ㅇ 또한 금융통화운영위원회 회의록을 분석한 텍스트 마이닝 적용 사례를 살펴봄으로써 이를 기반으로 향후 비정형 데이터를 활용한 다양한 경제정책에 대한 사후 평가 및 심층분석이 가능할 수 있도록 함

❑ 마지막으로 이와 같은 연구결과를 기반으로 향후 국회예산정책처에서 빅데이터의 정책 분석 활용과 관련된 시사점 등을 제시함

Ⅱ. 빅데이터 분석의 국내외 활용 현황

❑ 빅데이터 분석의 국내외 활용 현황은 관련된 기존 연구의 분석을 통해 수행함

○ 빅데이터는 다음과 같이 출처에 따라 사회관계망 자료, 거래내역자료, 사물 인터넷 자료로 분류될 수 있으며 자료 성격에 따라 연구도 다양한 분야에서 진행되고 있음

[표 Ⅱ-1] 빅데이터의 분류

사회관계망 자료	거래내역 자료	사물인터넷 자료
- 소셜미디어(트위터 등) - 블로그, 코멘트 - 개인문서 - 사진, 동영상(유튜브 등) - 인터넷 검색 - 모바일데이터 - 사용자 생성지도 - 이메일	- 공공기관생성자료 의료기록 등 - 기업생성자료 상업적 거래 은행/증권 기록 전자상거래 신용카드	- 센서 데이터 홈자동화 기후/오염센서 교통센서/웹캠 과학센서 보안/감시용 비디오 사진 휴대폰 위치 자동차 위성사진 - 컴퓨터 시스템 데이터 로그/웹로그

출처: UNECE(2013)
재인용: 한국은행(2017)

❑ 본 절에서는 기존 연구를 자료의 출처에 따라 크게 텍스트분석, 행정자료, 위치정보, 이미지정보 및 민간거래정보로 구분하여 정리함

○ 텍스트

– 포털검색어, 소셜미디어와 인터넷뉴스(연관어 및 감성분석), 웹사이트 로그 정보, 문서자료분석(회의록)

○ 행정자료

– 인구센서스, 국세청, 건강보험, 국민연금, 고용보험, 산재보험 자료 등

○ 위치정보

- 교통정보(교통카드, 우버이용), 지역정보(공간계량분석), 부동산, 상권분석

○ 이미지

- 위성이미지, 소셜이미지(인스타그램 등)

○ 민간 거래 정보

- 신용카드, 스캐너데이터(마트, 수퍼마켓 등), 신용정보 등

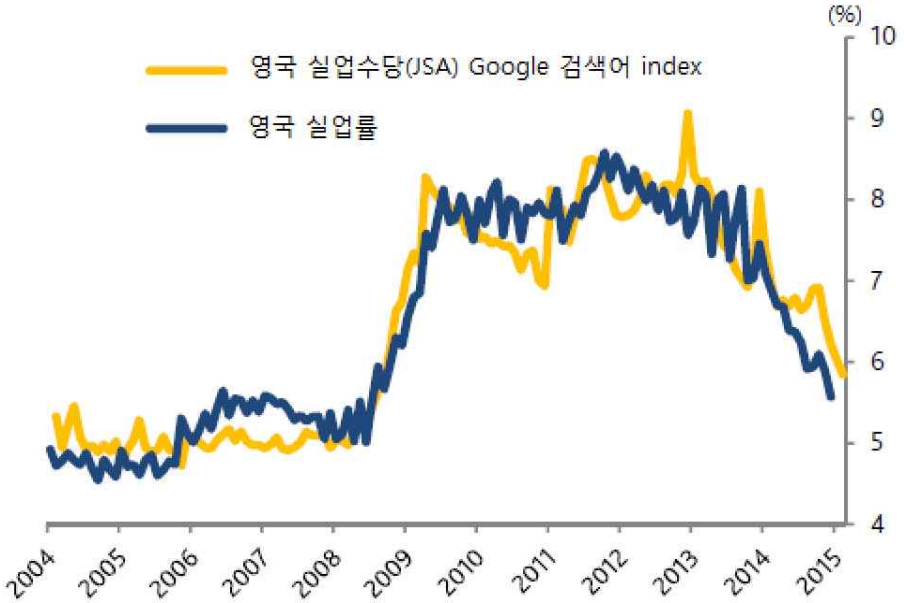
1. 해외사례

❑ (텍스트) 텍스트 자료를 이용하여 경제문제를 분석한 해외의 연구들은 웹 검색어 빈도분석, 소셜미디어 텍스트분석, 뉴스기사분석 등 다양한 방법을 활용

○ 검색엔진의 검색어 데이터베이스를 활용한 연구로 McLaren and Shanbogue(2011)과 Choi and Varian(2009, 2012) 등을 들 수 있음

- McLaren and Shanbogue(2011)은 영국의 실업수당 검색 빈도수를 이용하여 실업률을 예측할 수 있음을 보임

[그림 Ⅱ-1] 영국의 실업수당 검색 빈도수 활용사례



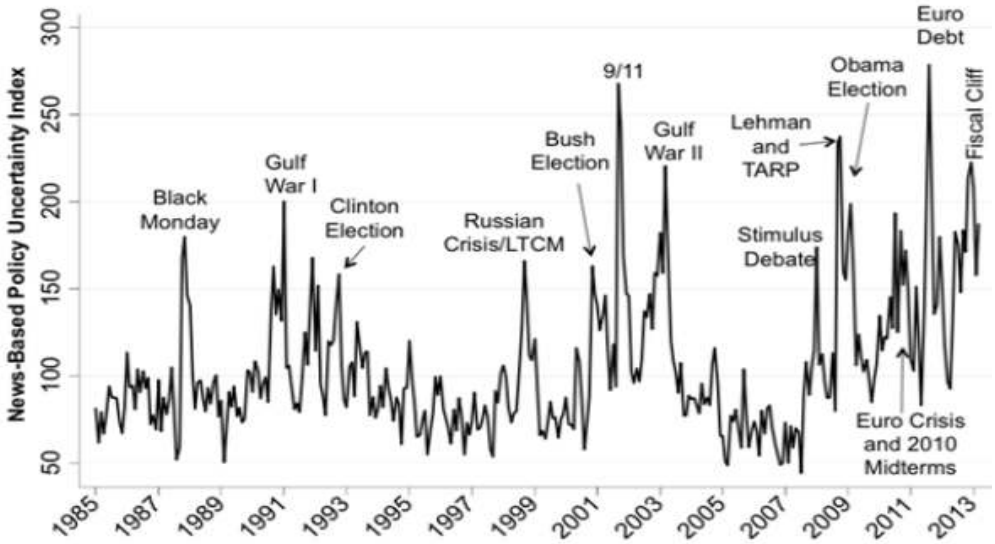
출처: 김수현 외(2019)에서 재인용

- Choi and Varian(2009, 2012)는 TV와 냉장고 등의 구글 검색어가 내구재 수요를 보여주고, 일자리와 복지 관련 검색어가 실업수당 청구건수에 대한 높은 설명력을 가짐을 보임

○ 거시경제변수에 대한 예측이나 분석에 활용되는 경제정책 불확실성지수(economic policy uncertainty, EPU)를 도출하기 위해 뉴스기사에서 나타난 불확실성 관련 단어의 빈도를 활용

- Baker, Bloom and Davis(2013)는 미국과 유럽의 뉴스기사에서 불확실성 관련 단어가 등장한 기사 수를 측정하고, 이 지표가 주가변동성과 정(+)의 관계를 가지며 투자 및 고용과 부(-)의 관계를 갖는다는 것을 보임

[그림 Ⅱ-2] 미국과 유럽의 뉴스기사 활용사례



출처: Baker, Bloom and Davis(2013)

- 그 밖에 트위터와 같은 소셜미디어에 특정 사건과 관련한 심리를 나타내는 단어의 검색 빈도수를 통해 시장의 심리를 예측하기도 함

○ Hansen and McMahon(2016)은 연준 위원회 회의록에 텍스트 분석을 적용하여 논조를 분석하고 이것이 실물경제에 영향을 주는 것을 확인

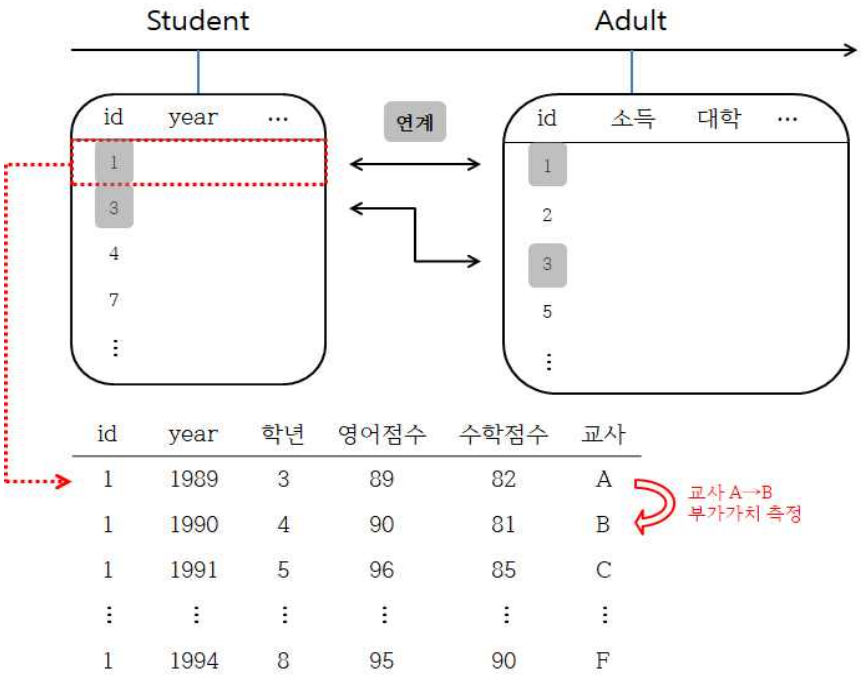
- LDA(Latent Dirichlet Allocation)의 방법론을 활용하여 자동적으로 회의록 표현들의 주제를 분류하고 논조를 정량화하여 분석

❑ (행정자료) 해외 연구 중에서 행정데이터를 연계하여 분석한 대표적인 사례는 Chetty et al.(2011), Cellini and Turner(2016), 그리고 Armour and Hung(2017)가 있음

○ Chetty et al.(2011)은 학생의 능력과 교사의 능력이 장기적으로 가져오는 성과를 측정

- 기존의 전통적인 데이터를 사용한 분석으로는 교육의 성과를 단기적으로 추정할 수밖에 없다는 문제의식이 제기됨
- 학생들의 학업 성취 데이터와 국세청의 세금 데이터를 연계하여 분석
- 1989~2009년 미국 대도시의 3~8학년 학생의 영어 및 수학 시험성적과 1996년~2010년 동일한 인물의 소득, 대학진학 정보를 연계함
- 분석결과 학창시절에 교사의 분포 중간값보다 1표준편차(sigma) 높은 교사의 학습지도는 학생의 평생 소득을 약 25,000달러 높인다는 결과를 보고

[그림 Ⅱ-3] Chetty et al.(2011)의 자료 연계 개념도



출처: 이창근 외(2017)

○ Armour and Hung(2017)은 은퇴 연령의 증가로 연금 수급액의 가치가 감소할 때 인구 그룹별 반응을 비교하여 연금 정책이 은퇴 계획에 미치는 영향을 확인

- 개인별로 노후대비 자산의 구조가 다르고 이에 따라 연금의 변화에 대한 반응이 다를 것이라는 점에 착안

- 개인별로 가지고 있는 다양한 자산을 측정하기 위해 조사자료(Health and Retirement Survey)와 Social security 행정기록을 연계
- 설문조사 데이터와 사회보장 기록데이터를 사회보장번호(SSN)로 연계해서 분석함으로써 조사자료의 심도있는 정보와 행정자료의 대표성을 모두 확보
- 분석 결과 은퇴 자산과 연금 수령의 보완적 관계가 존재함을 확인

[그림 Ⅱ-4] Armour and Hung(2017)의 자료 연계 개념도

SSN	Non-Social Security			인구학적 특성	
	Income	Asset	Pension	Age	Married
1	30K	500K	650K	57	1
2	50K	630K	500K	60	0
-	10K	610K	440K	59	1
4	60K	550K	460K	61	1

SSN	Social Security		
	수령 연령	수령 종류	수령 액수
4	63	A	400K
3	67	B	350K
2	69	B	340K
1	62	A	210K

SSN	Non-Social Security			Social Security	
	Income	Asset	Pension	수령 연령	수령 액수
1	30K	500K	650K	62	210K
2	50K	630K	500K	69	340K
4	60K	550K	460K	63	400K

출처: 이창근 외(2017)

○ Cellini and Turner(2016)는 행정데이터를 활용하여 사립대학 교육의 성과를 분석
- 기존의 연구들이 쉽게 자료를 얻을 수 있는 공립교육의 효과만을 분석했던 점에 대한 문제의식
- 정부 학자금 지원 자료와 학생의 취업과 소득 정보를 연계
- 분석 결과 4년제 대학은 소득이 오히려 감소한다는 결과를 확인했으며, 석사의 경우에는 소득이 증가한다는 결과를 보고
- 행정데이터는 기존 설문조사 데이터에 비해 더 정확하고 신뢰성 있는 정보를 제

공해주고, 기존에 없던 나이가 많은 학생의 정보도 제공해주는 등 연구의 질적 향상에 많은 도움이 되었다는 점을 확인함

○ 이러한 연구는 주로 단순한 현황파악만을 보여주는 국내연구에 비해 풍부한 정책적 시사점을 제시해줄 수 있다는 장점을 가짐

- 다만 개인 식별 가능성이 높기 때문에 법제도적 우려와 규제의 벽을 넘어야 한다는 점에서 쉽게 이용하기 힘들다는 단점이 존재
- 아래 표에는 미국의 행정자료를 이용한 몇 가지 연구 사례들을 정리

[표 Ⅱ-2] 미국 행정자료 이용 연구사례

주요연구	분석자료	특징
Duncan et al(2010)	Treatment Episode Data Set(TEDS)	55세 이상을 대상으로 약물남용에 대한 특징 분석
Butrica, Iams and Smith(2003)	Census Bureau Surveys와 SSA's Administrative Records 연결	Model of Income in Near Term을 사용하여 베이비붐 세대의 소득을 예측
Bull, Nicholas, and Dowd(2006)	IRS, Social Security Administration 연결	Microsimulation model을 이용해 베이비붐 세대의 고령화에 따른 조세시스템의 변화와 이에 따른 세율의 변화 추정
Butrica, Smith, and Iams(2004)	Social Security Administration (SSA), Center for Retirement Research at Boston College (CRR) 연결	베이비부머 세대 은퇴자들과 이전 세대를 비교해, 소득 수준, 분포, 구성을 확인하고, 은퇴 후 경제 수준을 유지하는 데에 충분한지 확인
Armour and Hung (2017)	Social Security data, Health and Retirement Survey 연결	Social Security 혜택이 은퇴후 소득계획에 어떠한 영향을 미치는 지 파악
Meyer, Wu, Moores, and Medalia(2019)	Population Survey, Survey of Income and Program Participation와 Administrative Tax and Program data 연결	현물보조 등 공적 부조 등을 소득에 추가하여 극빈곤선을 재추정

출처: 이창근 외(2017) 및 저자 수정

❑ (위치 & 이미지) 지도, 위치, 이미지 등의 빅데이터를 활용한 사례로는 Dunleavy(2016), Glaeser et al.(2018) 등이 있음

○ 지도 및 위치정보도 빅데이터의 한 종류로 널리 이용

- 소비자들의 이동경로를 파악하는 것은 기업의 입장에서는 마케팅 전략을 세우는데 도움
- 차량의 이동정보는 지방 행정부가 도로의 혼잡과 대중교통 문제를 해결하는데 중요한 역할(Dunleavy, 2016)

○ Glaeser et al.(2018)은 다양한 도시 데이터 소스들을 제시하고, 도시에 대한 연구와 기능을 개선시키기 위해 해당 소스들을 어떻게 활용할 수 있는지 보임

- 특히 Google Street View 이미지를 통해 뉴욕의 수입 예측에 활용할 수 있는 방법과, 과거에 측정되지 않았던 개발도상국의 부와 빈곤을 확인하는 데 이와 유사한 이미지 데이터를 활용할 수 있는 방법을 제시
- 먼저 컴퓨터 기반의 시각 인식 기술을 사용하여 뉴욕시의 Google Street View 이미지와 소득을 연결하는 예측 모형을 만들고, 시간의 흐름에 따라 변화하는 이미지를 추적하여 해당 모형에 적용
- 분석 결과, Google Street View는 뉴욕 및 보스턴의 소득을 예측할 수 있고, 예측된 소득은 연구 내 표본 상에서 주택가격을 예측하는 데 유의하다는 확인하였으며 이를 통해 전 세계의 부와 빈곤 패턴을 이해하는 데에도 Google Street View가 활용될 수 있음을 나타냄

❑ (민간거래정보) 민간거래 빅데이터를 활용하여 분석한 해외 사례로는 Cohen et al.(2016)이 대표적임

○ Cohen et al.(2016)은 Uber의 서지 프라이싱(Surge Pricing)의 알고리즘과 개인 데이터를 활용해 수요탄력성을 추정하고, 이를 통해 소비자 잉여를 추정함

○ UberX는 운전자와 해당 택시를 찾는 소비자 간에 알고리즘을 활용하여 매칭시키는 앱 기반 서비스를 의미하며, Uber의 특징은 실시간 가격(서지 프라이싱)을 통해 지역 및 단기적인 수요와 균형점을 맞춘다는 것임

○ 수요탄력성의 경우 각 서지 프라이싱 수준에서 주행에 이르는 Uber 검색의 비율을 계산함으로써 추정할 수 있음

- 무작위 상에서 70%에 해당하는 검색이 기본 가격으로 거래되나, 10%가 더 높을 때 63%에 해당하는 검색으로만 연결된다면, 1.0x 서지 프라이싱을 받은 사람들이 1.1배 가격에 견적을 받았다면 63%의 비율로 요청했을 것이라 추측할 수 있음
- 즉, 가격 제안을 받아들이는 사람들이 구매율 70%에서 구매율 63%로 10% 감소했다는 것은 가격 상승과 관련이 있다고 볼 수 있음

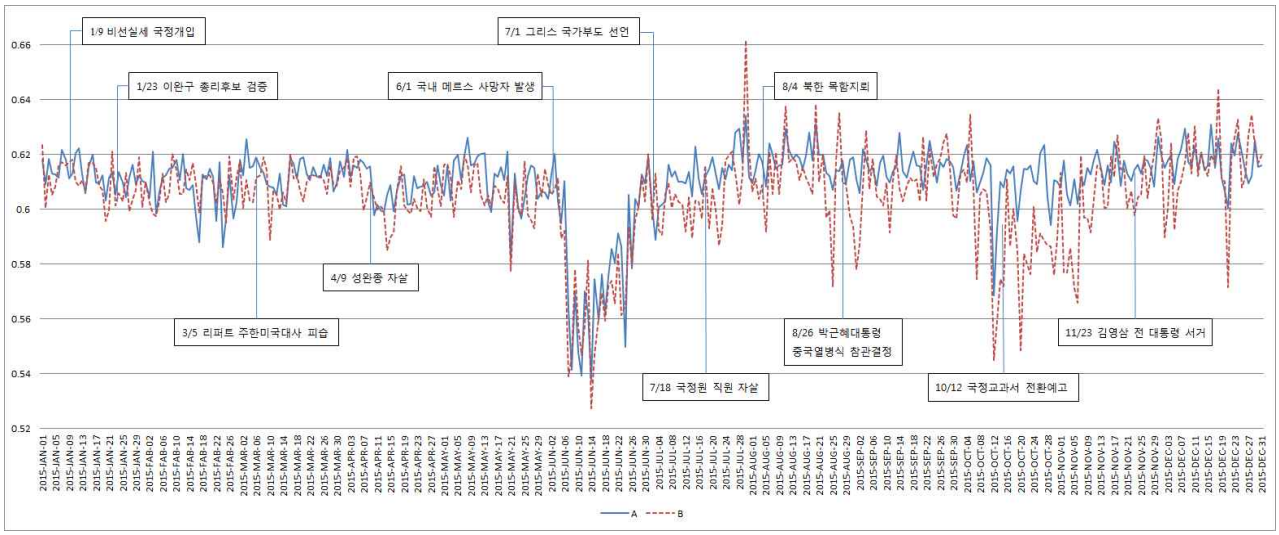
○ 5천 만 개의 개인 관측 데이터와 회귀단절모형(regression discontinuity design)을 활용하여 분석한 결과, 2015년 UberX 서비스가 미국 4개 도시에서 약 29억 달러의 소비자 잉여를 산출한 것으로 추정(각 소비자가 지출하는 1달러 당 약 1.6달러의 소비자잉여금 발생)

2. 국내사례

❑ (텍스트) 빅데이터 분석을 활용한 국내 사례로는 최동욱(2017), 오선정(2018)을 참고할 수 있음

- ㅇ 뉴스 기사나 SNS 및 블로그 등의 텍스트를 수집하여 분석하는 텍스트 마이닝 기법을 적용(최동욱, 2017)
  - 컴퓨터의 향상된 능력으로 대량의 텍스트 정보를 정량적으로 처리 가능
  - 기존의 숫자로 주어지는 것 이상의 정보, 즉 텍스트 작성자나 발언자의 감정과 성향과 같은 정보를 추출할 수 있다는 장점
- ㅇ 최동욱(2017)은 포털뉴스가 정치적 편향을 보이게 되는 경제적 유인구조를 확인하기 위해 정치성향을 정량적으로 측정하는 방법을 제시
  - 저자는 국회회의록에 대한 텍스트분석을 통해 정치성향을 드러내는 표현들을 수집했는데, 이는 정치인들이 공식적인 석상에서 발언한 내용은 고도의 정치적 전략을 담고 있다는 점에 착안한 것
  - 예컨대 “올바른 역사 교과서”, “미친 전세 값” 등 정치적 전략이 담긴 표현들을 추출할 수 있었음
  - 이러한 표현들이 포털뉴스에서 얼마나 사용되는지를 파악하여 각 뉴스의 정치성향을 측정
  - 뉴스 이용자들의 정치성향을 파악하기 위해서 설문조사자료와 포털방문자 정보의 확률적 연계방법 제시
  - 이를 통해 포털뉴스가 보여주는 정치성향은 유사한 성향을 가진 뉴스 이용자들을 유도함으로써 클릭수를 높이기 위한 경제적 유인이 존재한다는 점을 실증적으로 보임
  - 아래 그림은 2015년 인터넷 포털뉴스의 정치성향의 추이를 보여줌

[그림 Ⅱ-5] 2015년 국내의 주요 정치적 사건과 포털뉴스의 정치성향 변화추이



출처: 최동욱(2017)

○ 이영준 외(2019)는 금융통화위원회 전후 3일 간의 뉴스 기사에 대한 텍스트 분석을 통해 통화정책 커뮤니케이션이 경제에 미치는 영향을 확인

- 2005년 3월부터 2007년 11월까지 ‘금통위’가 포함된 신문기사 24만 건을 분석
- 기사의 논조는 경제금융에 특화된 형태소 사전(eKoNLPy)을 개발하여 적용
- 금통위 이전 논조와 이후 논조의 차이를 금융정책 서프라이즈(Monetary Policy Surprises)로 정의
- 기준금리 변동 여부와 방향에 따라 기사의 수 및 논조 차이가 다르게 나타남
- 이러한 서프라이즈(Surprises)가 장단기 금리, 환율, 주가 등에 영향을 미치는 것으로 나타남

○ 최경덕 외(2016)는 후쿠시마 원전사고 관련 SNS의 정보확산이 소비자들의 수산물 소비에 미친 영향을 분석

- SNS 중 블로그와 트위터에 등장한 키워드를 중심으로 분석
- ‘수산물+방사능’, ‘수산물+안전’, ‘방사능+물고기’ 등 주요 키워드를 조합하여 활용
- 원전사고 관련 정보는 수산물 소비를 줄이고 대체재인 돼지고기 소비를 늘리는 것으로 분석됨

○ 이긍희 외(2016)는 인터넷 포털 사이트(NAVER)의 검색데이터를 이용하여 고용 관련 검색지표를 작성하고 그 성과를 분석

- 고용 관련 키워드를 수집한 후 검색데이터를 추출하고 시계열의 유용성, 상관관

계 등을 고려해 15개의 고용 키워드를 선정

- 선정된 데이터 군의 군집분석 및 주성분분석을 통해 검색지표를 작성하고 이들에 대한 교차상관분석, 예측력 분석을 실시
- 분석 결과, 해당 고용 검색지표는 일부 고용 공식통계와 밀접하게 움직이는 것으로 나타났으며, 적절한 시차의 검색지표를 포함한 모형이 예측력 측면에서도 더 우수한 것으로 나타남

○ 오선정(2018)은 그동안 정의가 모호하게 남아있던 아르바이트 노동의 개념과 종사자의 특성을 실증적으로 파악하고자 함

- 특히 아르바이트 일자리를 구하고자 하는 사람들이 사용하는 검색어를 인구특성에 따라 구분하여 검색량을 측정하고 이를 통해 경제환경의 변화와 아르바이트 노동의 특성 변화를 파악하고자 함
- SNS나 커뮤니티 게시판의 관련 언급량과 연관어의 속성을 통해 이용자들의 아르바이트에 대한 평가의 변화도 파악
- 분석결과 최근 20년간 경제상황의 변화로 인해 여성과 직장인들의 아르바이트가 증가하는 현상, 그리고 고령자들의 아르바이트가 증가하는 현상 등을 파악할 수 있었음

❑ (행정자료) 국내 연구 중에서 행정데이터를 연계하여 분석한 대표적인 최근 사례는 이태리 외(2018)를 들 수 있음

○ 이태리 외(2018)는 금융 및 주택시장의 빅데이터인 '개인신용정보'와 '주택실거래데이터'를 연계하여 LTV(Loan to Value, 주택담보대출비율), CoLTV(Comprehensive Loan to Value, 실효LTV) 지표를 산출하였고, 이를 통해 임대차주의 상환위험을 분석

○ 주소, 성별, 연령 등 개인식별자료를 이용해 차주의 대출 및 소득정보와 주택거래가격 혹은 임대료를 일정한 가정 하에 매칭

○ 특히 CoLTV 지표를 통해 LTV 지표만으로는 확인하기 어려웠던 임차인의 전세보증금 손실 위험에 대해 살펴볼 수 있었음

❑ (위치정보) 박종수·임금숙(2015), 하나금융경영연구소(2019)는 교통망, 거리 등 위치 관련 빅데이터를 활용해 분석

○ 박종수·임금숙(2015)은 서울 시내버스 이용자의 교통카드 데이터베이스를 기반으로 지

점 간 이동 시 필요한 실제 시간거리를 산출하고, 목적지까지의 접근성을 산출하여 분석

- 서울시에서 승객의 통행정보가 교통카드 데이터베이스에 저장되며, 이를 바탕으로 버스노선과 노선 상의 버스정류장에 대한 속성(attribute)을 파악할 수 있음
- 서울 시내버스 노선 망을 교통망 그래프 모델로 변환하고 최단 경로 알고리즘을 응용해 서울 시내 버스정류장들 사이의 최단 경로, 이동 시간, 버스정류장의 접근도 등을 분석
- 분석 결과, 서울 버스 교통망의 시간거리 접근도는 도심 지역이 가장 높고, 그 다음으로 강남역 및 영등포역 부근 등 상업중심지역이 높게 나타남

○ 하나금융경영연구소(2019)는 공공 데이터를 활용해 서울 직장인의 통근 방법 및 소요시간의 변화 등 전반적인 출퇴근 관련 트렌드를 분석하고 시사점을 제시

- 서울시민 대상 설문조사 데이터(서울서베이), 각 국가기관별 인구 통계데이터, 수도권 도시철도 등의 공공데이터 결합 분석을 통해 서울 직장인의 변화된 출퇴근 패턴을 확인
- 공공데이터로는 서울서베이도시정책지표조사 보고서(서울특별시), 자치구별 1인당 지역내 총생산 및 수준지수(서울특별시), 도시철도 승하차 통계(서울도시철도 및 KORAIL), 지하철 승하차 인구(K-ICT 빅데이터 센터, ㈜공간과가치), 주민등록 인구현황(SGIS, 행정안전부), 종사자수 분포 현황(SGIS, 통계청), 국민여가활동조사(문화체육관광부), 브이월드(공간정보산업진흥원)가 사용됨
- 분석 결과 '직주 근접' 선호 현상을 실제 데이터 분석으로 확인하였으며, 최근 10년간 직장인의 출근 시간은 늦어지고, 퇴근 시간은 빨라졌음을 확인

❑ (이미지) 이미지 데이터를 구축하여 경제 분석에 활용한 사례로는 김규철(2017)이 있음

○ 김규철(2017)은 인공위성을 통해 측정된 야간 조도(nighttime light) 데이터를 구축하여 북한 주민의 후생에 대해 파악함

○ 인공위성의 야간 조도 데이터는 미국의 해양기상청 홈페이지를 통해 확인할 수 있으며, 1992년부터 2013년까지의 데이터가 1년 단위로 공개되어 있음

○ 인공위성이 관측하는 빛의 밝기는 전기소비량과 매우 밀접하게 관련되어 있으며, 야간의 불빛만을 측정하므로 생산 수준보다는 소비 수준과 큰 관련이 있음

○ 해당 보고서는 북한의 야간 조도 데이터가 소비활동과 밀접한 관련성이 있으며, 이를

적용하여 북한의 경제상황을 분석할 경우 북한 전역은 물론 특정 지역의 후생 수준을 장기간에 걸쳐 확인할 수 있음을 제시

- ㅇ 북한의 야간 조도 데이터를 통해 확인된 북한 전반의 후생 수준은 2000년 이후 꾸준히 증가했고, 물적 자본의 투자나 무역 증대, 북한당국의 경제정책 등이 지역의 후생 수준에 큰 영향을 끼침을 확인

❑ (민간거래정보) 신용카드 데이터 등 민간거래 빅데이터를 활용하여 분석한 최근 사례로는 김경근·염명배(2017)가 있음

- ㅇ 김경근·염명배(2017)는 소비 대리변수로서의 신용카드 빅데이터를 활용해 지역별·업종별 소비와 관련해서 역외소비율 및 소비유입율 등의 지표를 산출, 지방에서의 역외소비가 지역경제에 미치는 생산 및 고용 효과에 대해 분석
- ㅇ 이를 위해 국내 신용카드 3사의 소비자기준 및 가맹점기준 지역별 및 업종별 신용카드 소비액 자료를 수집
- ㅇ 분석 결과, 지방의 순역외소비 규모가 민간소비나 지역총생산에 비해 빠르게 확대되고 있음을 확인했으며, 지역별 소비 유출입을 파악하는데 있어서 신용카드 통계자료의 신뢰성 및 유용성을 확인

Ⅲ. 빅데이터 분석기반 경제분석 방법론

❑ 다음으로 빅데이터를 활용한 경제정책 분석 방법론을 정리하고 이를 실제 적용하기 위해 필요한 내용들을 파악하고자 함

- ㅇ 경제정책 문제의 해결을 위한 빅데이터 분석과 관련한 최근의 연구들을 조사 및 정리하고 이들로부터 얻을 수 있는 시사점을 정리
- ㅇ 기존의 연구들이 활용했던 방법론을 유형별로 정리하고, 장단점을 파악하여 실제 정책을 분석하는데 적용할 수 있도록 함

1. 빅데이터 적용 가능성의 검토

❑ 해결하고자 하는 문제에 가장 적합한 자료를 선정하는 것이 중요

- ㅇ 특히 빅데이터의 경우에는 그 자체로 주요한 결론을 도출하기보다는 기존 데이터를 보완하는 방식으로 활용하는 것이 바람직함(Stephens-Davidowitz, 2017)
  - – 일례로, NBA선수들의 성공에 출신배경이 미치는 영향을 확인하기 위해 뉴스기사와 SNS 텍스트를 분석하여 가정환경에 대한 정보를 데이터화
  - – 이러한 데이터를 출생지 정보, 이름 및 미국 인구조사 자료와 연계한 결과, NBA선수들의 경우 출신배경이 좋을수록 성공할 확률이 유의하게 높아진다는 것을 확인
  - – 즉, 인구센서스 자료를 뉴스텍스트 자료로 보완하여 새로운 데이터셋을 구성
- ㅇ 문제 해결을 중심에 놓고 판단하면 효과적인 데이터를 찾는데 도움이 될 수 있음(Stephens-Davidowitz, 2017)
  - – 우수한 경주마를 선별하는 효과적인 방법은 말의 혈통이 아니라 심장의 크기와 달리는 방식
  - – 이를 측정하기 위해 이미지 자료와 동영상 자료를 종합적으로 활용, 경주마의 우승가능성을 효과적으로 예측
  - – 말의 족보와 같은 전통적인 자료가 아닌 실제 신체적 능력과 행동 특징을 파악할 수 있는 데이터를 활용하는 것이 더욱 정확하다는 결과

❑ 따라서 해결하고자 하는 문제에 대해 구체적으로 정의하는 작업이 우선되어야 함

- ㅇ 어떤 문제를 해결하고자 하는지 명확히 할 필요가 있음
  - 구체화될수록 데이터를 선별하는데 도움이 될 수 있음
- ㅇ 연구문제의 사례는 다음과 같음
  - ① 포털 검색어를 이용해서 독감의 확산을 예측할 수 있는가?
  - ② 경제의 불확실성이 주가에 어떤 영향을 미치는가(Baker et al., 2013)?
  - ③ 학생들의 중고등학교 시절 학업성취도가 평생 소득에 영향을 주는가(Chetty et al., 2011)?
  - ④ 스팸메일 차단의 경제적 효과는 얼마인가(Gentzkow et al., 2017)?
  - ⑤ 우버의 도입이 소비자 후생을 얼마나 증가시키는가(Cohen et al., 2016)?
- ㅇ 이러한 문제의 성격에 따라 크게 두 가지 종류로 구분 할 수 있음
  - 속보성과 예측이 중요한 경우(①, ②)로 빅데이터를 활용한 지표를 생성하고 이를 이용하여 현실을 설명(상관관계)
  - 사후적으로 심층평가가 필요한 경우(③, ④, ⑤)로 현실의 경제 모형에 빅데이터를 활용한 지표가 포함(인과관계)

❑ 문제가 정의되었다면 문제를 해결하기 위해 필요한 변수를 파악하고 빅데이터의 활용이 필요한지 검토해야 함

- ㅇ 어떤 문제가 빅데이터를 활용해야 하는 문제라면 다음과 같은 성격을 가지고 있기 때문이라고 생각할 수 있음
  - 기존의 자료로는 풀 수 없는 새로운 문제
  - 정성적 정보의 (객관적)정량화
- ㅇ 예컨대 Chetty et al.(2011)의 연구문제와 같이 중고등학교 시절의 학업성취가 성인기의 소득수준에 영향을 주는지 확인하기 위해서 학창시절의 성적과 성인기의 소득 정보를 개인수준으로 식별하여 연계하는 방법을 활용할 수 있음
  - 국세청의 납세정보 등 행정데이터를 활용해서 개인을 식별하고 소득수준을 정확하게 파악가능
  - 이는 전통적으로 활용하는 조사자료(설문 등)로는 확인하기 힘든 정보임
- ㅇ 최동욱(2017)과 Gentzkow and Shapiro(2010)처럼 국회의사록을 활용한 텍스트분석을 통해 미디어의 정치성향을 파악하는 방법도 가능

- 이는 매체에 대한 정량적 정보(뉴스보도량, 광고량 등)뿐만 아니라 정성적 정보인 기사 내용을 분석하여 정량화하는 작업이 필요함

2. 속보성과 예측이 중요한 경우

❑ 속보성이나 예측이 중요한 문제를 해결하기 위해서는 실시간으로 수집되는 데이터를 이용할 수 있음

ㅇ 포털 검색어나 SNS텍스트 자료 등 실시간으로 생성되는 데이터를 이용하여 현실경제를 예측하는데 활용할 수 있음

ㅇ 텍스트 마이닝의 활용 방법

- 포털 검색어의 검색량을 활용하면 특정 이슈에 대한 이용자들의 관심도 변화 추이를 확인할 수 있으며 이용자들의 인구통계적 특성에 따른 차이점을 비교할 수 있음
- SNS와 커뮤니티의 특정 키워드 언급량(buzz)을 이용하면 동일한 선호를 가지고 있는 그룹에서의 관심정도를 파악할 수 있음
- 뉴스기사에서의 보도량을 통해 뉴스이용자들의 관심사안 중 언론사들이 중요한 이슈로 선별한 내용의 비중을 파악할 수 있음

❑ 실시간 데이터와 경제변수 간의 상관관계가 성립함을 전제로 실시간 예측이 가능

ㅇ 경제지표의 변동(Choi and Varian, 2012)

- 포털 검색어 중 TV와 냉장고에 대한 검색어가 일반 내구재에 대한 수요와 유사한 추이를 보임
- 일자리 및 복지 관련한 검색어는 실업수당 청구와 높은 상관관계를 가짐

ㅇ 불확실성 지표(Baker, Bloom and Davis, 2013)

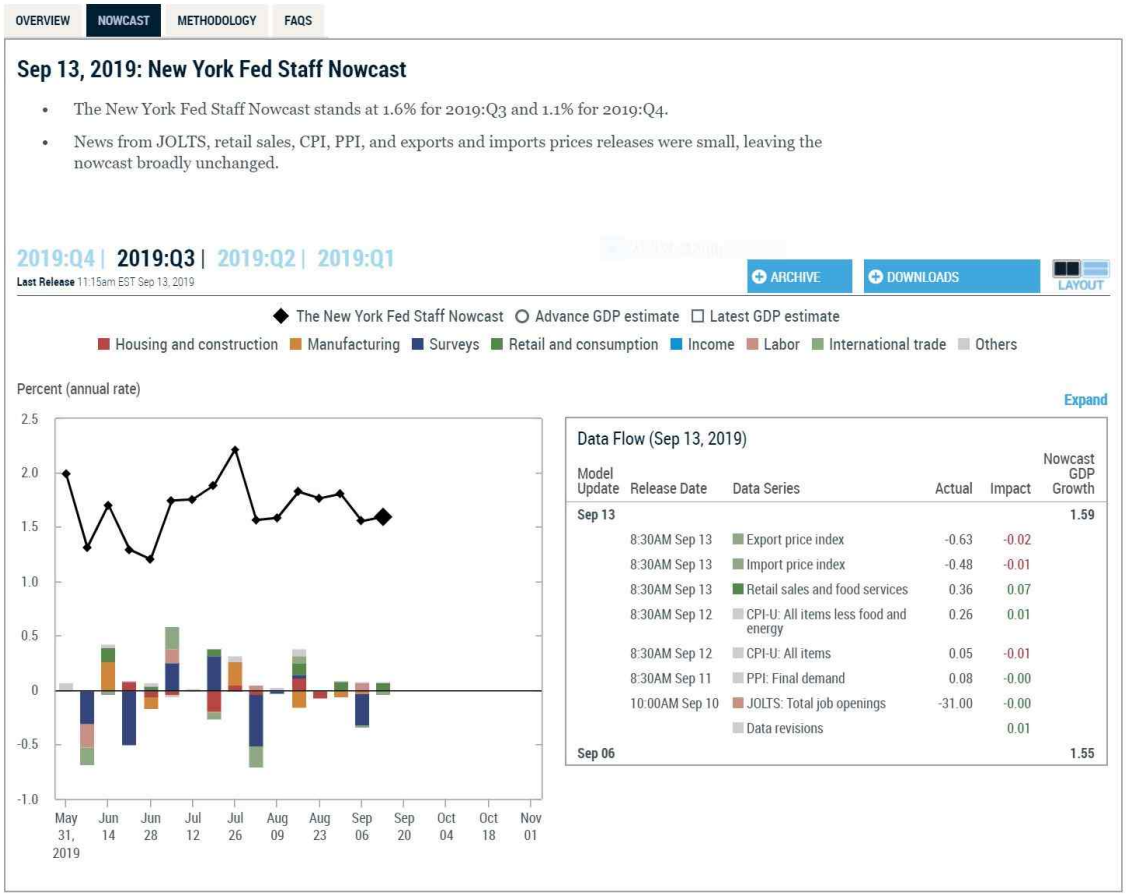
- 경제 관련 언론 기사에 '불확실성(uncertainty)'이라는 단어가 등장한 경우, 이는 실제 경제정책이나 경제환경에 대한 불확실성과 관련되어 있을 가능성이 높음
- 따라서 불확실성을 언급한 건수를 현실 경제를 나타내는 변수들과 비교해보면 높은 상관관계를 발견할 수 있음

ㅇ GDP나우캐스팅(FRB of New York)

- 나우캐스팅은 초단기로 제공되는 예보를 뜻하며 뉴욕FRB는 2016년부터 GDP에 대한 주간 나우캐스팅을 발표하고 있음(그림 Ⅲ-1)

– 다양한 소스의 빅데이터 활용을 통해 GDP를 추정하는 방식으로 전통적인 GDP전망치 추정방법에 비해 정확도의 손실은 적으면서 속도 및 빈도의 이득은 높은 방식으로 추정 가능해짐

[그림 Ⅲ-1] 뉴욕 FRB의 GDP나우캐스팅 페이지



출처: Federal Reserve Bank of New York 웹사이트(https://www.newyorkfed.org/research/policy/nowcast), 접속시간 2019.9.18. 16:00.

❑ SNS 텍스트 분석 사례(이창근 외, 2017)

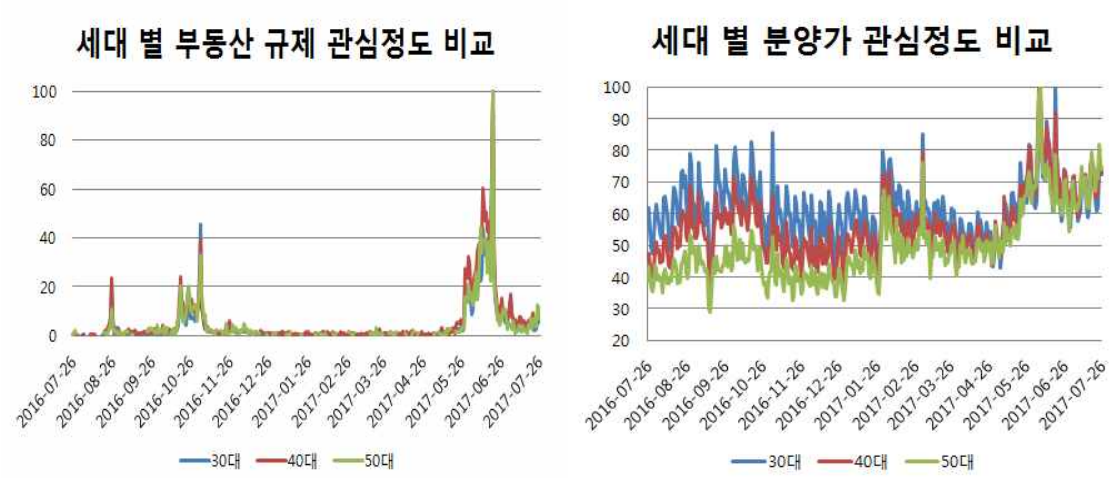
○ 부동산 정책에 대한 검색량, SNS 언급 및 연관어 추출을 통해 정책에 대한 이용자의 반응을 추정

– 포털 검색어, 블로그, 게시판, SNS에서 드러난 표현을 수집하여 특정 정책에 대한 인식과 태도를 파악할 수 있음

– 포털 검색량은 정책 발표 이후의 이용자들의 반응을 확인할 수 있으며, 인구그룹별 비교를 통해 정책에 대한 관심도를 비교할 수 있음

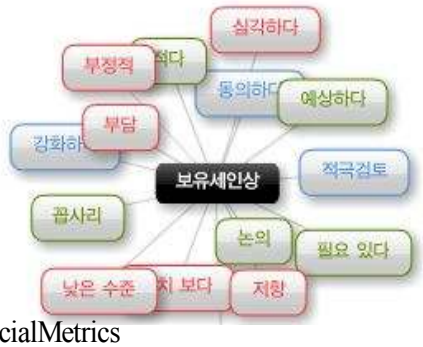
- SNS에서의 언급은 이용자들이 자발적으로 올리는 것으로 설문조사와 비교할 때 응답편의가 낮음
- 특정 정책에 대한 이용자들의 반응을 긍정적, 부정적, 중립적인 표현으로 구분할 수 있음

[그림 Ⅲ-2] 부동산 관련 키워드를 통한 연령대별 관심정도의 비교



○ 결과를 보면 규제에 대한 관심은 연령대 간 동일하지만 규제 강도의 변화에 따라 반응이 다르게 나타나는 것을 확인할 수 있음
- 정부의 규제발표(announcement) 전에는 주로 30대가 분양가에 관심을 많이 가지고 50대는 관심이 비교적 낮았으나 2017년 6월 규제강화 발표 뒤 모든 연령층의 관심이 급증함
- 이는 연령대나 세대에 따라 소득수준, 자산구성, 혹은 가치관 등이 다르고, 이러한 차이가 분양가에 대한 관심의 정도에도 다르게 반영되는 것으로 보임
- 또한 세대별 상황에 따라 부동산에 대한 수요(실수요, 잠재적 투기 수요)가 다르고 규제에 따라 이러한 수요가 변동하면서 위와 같은 현상이 나타나는 것으로 추측할 수 있음

[그림 Ⅲ-3] 부동산 보유세 인상에 대한 연관검색어(이창근 외, 2017)



출처: SocialMetrics

○ 이를 이용하여 부동산 보유세 인상에 대한 특정 집단의 긍정률(R)은 다음과 같이 도출할 수 있음

$$R_{p,g,t} = \frac{f_{p,g,t}}{f_{p,g,t} + f_{n,g,t}}$$

- $R_{p,g,t}$: 집단 g의 t기의 특정 정책에 대한 긍정률
- $f_{p,g,t}$: 집단 g의 t기의 긍정표현의 빈도수
- $f_{n,g,t}$: 집단 g의 t기의 부정표현의 빈도수
- 또한 이를 집단 g의 다른 속성과 연계하여 비교·분석할 수 있음
- 성별, 지역별, 소득별, 교육수준별로 정책에 대한 반응을 확인가능

❑ 또한 민간 빅데이터와 행정데이터 등 추가적인 데이터를 연계하여 분석할 수 있으며 이러한 방법론을 적용한 연구나 정책과제들이 해외에서는 활발히 진행되고 있으며 국내에서도 논의가 이루어지고 있음

○ 경제분야에서 생성되는 민간 빅데이터의 경우 최근 가계부채 문제가 전세계적으로 중요 이슈가 됨에 따라 개인신용 DB 자료를 이용한 연구가 중앙은행을 중심으로 활발히 진행되고 있음(Haughwout et al., 2019; 김성준 외, 2018)

○ 행정데이터의 경우 기존연구의 개관에서 살펴본 바와 같이 다양한 분야에서 활용이 이루어지고 있으며 특히 소득관련 공공자료를 서베이조사 자료와 연계시켜 다양한 정책평가가 이루어지고 있음(Meyer, Wu, Moores, and Medalia, 2019)

❑ 시사점

- ㅇ 포털 검색어나 SNS텍스트 등은 실시간으로 생성된다는 점과 동기 부여된 이용자의 자발적인 행위의 결과라는 점에서 신속성이 높고 편의성이 낮은 통계자료로서의 장점을 가지고 있음
- ㅇ 이러한 빅데이터의 특성을 활용하여 실제 경제현상과 높은 상관관계를 가지고 있는 변수를 생성하고 단기 혹은 중장기 예측에 적극적으로 활용할 필요가 있음
- ㅇ 본 연구에서는 위와 같은 방법론을 국내 주요 경제이슈 중 하나인 가계부채와 통화정책 사례에 대한 적용하여 실제 활용가능성을 확인해보고자 함
- ㅇ 또한 금융통화운영위원회 회의록을 텍스트마이닝 기법을 적용하여 분석함으로써 향후 이와 같은 비정형 데이터 활용을 위한 기초 자료 제공하고자 함

3. 사후평가 및 심층분석이 필요한 경우

❑ 사후평가나 심층분석은 주로 과거에 도입된 특정한 정책이 경제현상에 미친 인과관계를 사후적으로 엄밀히 검증하는 작업

- ㅇ 단순 상관관계를 파악하는 경우와 달리 경제 이론에 근거한 분석 모형이 제시되어야 함
- ㅇ 분석 모형을 검증할 실증 데이터를 수집하는 과정에서 빅데이터(디지털 기록이나 행정데이터 등)를 이용한 지표를 생성
- ㅇ 이중차분법이나 도구변수추정법 등 인과관계를 파악할 수 있는 계량기법을 적용하여 모형의 파라미터를 추정

❑ 샘플링 기법 등 자료의 대표성이 중요

- ㅇ 행정자료의 연계: 행정데이터는 기존의 조사자료와 달리 샘플링 편의의 문제가 적어 높은 신뢰성을 가짐
  - 사회 구성원이나 기업들의 매우 구체적인 정보를 보유하고 있다는 장점
  - 한편으로는 개인정보보호의 이슈에 따른 제약점도 존재(Dunleavy, 2016; Einav and Levin, 2014; Künn, 2015)

❑ 인과관계 분석방법

- ㅇ 가상적인 실험상황을 구성

- 관측대상을 정책의 영향을 받는 처치군과 영향이 없는 통제군으로 구분하고 정책 시행 전후의 변화를 관측

[표 Ⅲ-1] 실험상황의 구성

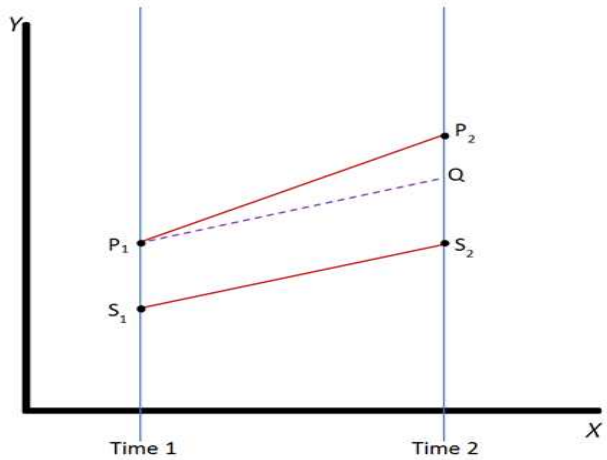
구분	처치군(정책수혜 대상)	통제군
정책시행 전	A	B
정책시행 후	A’	B’

- 이 때, 정책 시행으로 인해 나타나는 효과는 $\{E[B'] - E[B]\} - \{E[A'] - E[A]\}$로 측정할 수 있음

○ 패널데이터를 이용한 이중차분법

- 패널자료를 구성할 수 있다면 패널분석 계량모형을 이용하여 정책시행의 효과를 분석할 수 있음
- 예컨대 유사한 특성을 지닌 인접한 두 지역(P, Q)에서 한 지역 P에 정책(예: 최저임금)이 시행되고 Q에서는 정책이 시행되지 않았다면 두 지역에서의 성과차이(고용)를 통해 정책시행의 효과를 파악할 수 있음
- [그림 Ⅲ-4]에서 P지역의 성과지표와 S지역의 성과지표의 변화를 각각 $P_2 - P_1$과 $S_2 - S_1$이며 실제 정책시행으로 인한 효과는 $P_1 - Q$로 측정 가능

[그림 Ⅲ-4] 실험상황의 구성



- 계량모형으로 표현하면 다음과 같은 고정효과 패널모형으로 나타낼 수 있으며 다기간의 패널자료에도 적용가능함

$$y_i = \beta_0 + \beta_1 D_{gt} + \beta_2 G_i + \beta_3 t_i + \varepsilon_i$$

- y_i: 정치성향
- G_i: 처치군 = 1, 통제군일 = 0
- t_i: 처치 전(pre-treatment) = 0, 처치 후(post-treatment) = 1
- D_{gt} = $G_i \times t_i$
- $\hat{\beta}_1$: 정책효과의 추정치

○ 실험상황의 구성이 여의치 않은 경우 일반적인 고정효과 패널분석 모형을 적용하거나(패널 데이터의 경우), 도구변수 추정법(횡단면 데이터의 경우) 등의 방법을 통해 인과적 관계를 파악할 수 있음

❑ 행정데이터를 이용한 분석 사례(이창근 외, 2017)

○ 연구문제: 부동산 가격의 변동은 베이비붐 세대의 건강에 어떤 영향을 미칠 것인가?

○ 베이비붐 세대가 보유한 자산의 약 70%가 부동산으로 구성되어 있으며 2008년 세계 금융위기 이후 주택자산비중의 양극화가 발생

- 부동산 가격의 변동은 이들에게 다양한 경제적·비경제적 영향을 미칠 수밖에 없음
- 베이비붐 세대의 은퇴이후 부동산에 대한 의존도 증가

○ 부동산 가격의 비경제적 효과에 대한 연구는 부족한 상황

- 베이비붐 세대에게는 건강이 매우 중요한 문제

○ 이 문제를 분석할 때 고려해야 할 흥미로운 사항들

- 부동산 가격의 변동이 발생할 때 주택 소유 형태에 따라 건강에 미치는 영향이 차별적으로 나타날 수 있음
- 자산구조(부채비중 등)에 따라 부동산 가격 변화에 대한 민감도가 다르게 나타날 수 있음
- 내생성의 존재: 소득수준이 높은 건강한 사람들이 이미 부동산 가격이 높은 지역에 거주할 가능성 vs. 부동산 가격의 증가에 따라 가처분소득 증가로 건강에 대한 투자증가
- 부동산가격이 가구주와 가구원에게 미치는 영향이 다름

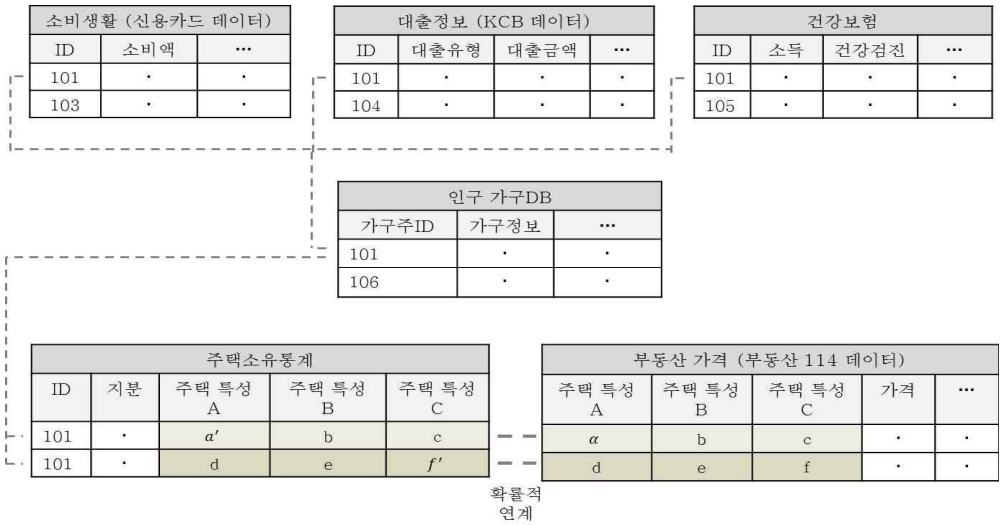
○ 다양한 행정데이터의 결합을 통해 이러한 효과를 파악할 수 있음

- 인구가구DB, 주택소유통계DB, 부동산가격데이터, 가계부채DB, 건강보험DB
- 연계하는 DB별로 가용한 기간이 다를 수 있음에 주의할 필요

○ 종단적 구성이 가능하고 실험조건의 구성이 가능하다면 이중차분법을 활용하여 부동산 가격과 건강과의 인과관계를 파악 가능

- 그 밖에도 다양한 변수들을 활용하여 패널모형 분석을 적용, 부동산 가격 변동에 따른 다양한 효과를 파악할 수 있음

[그림 Ⅲ-5] 자료연계 개념도



출처: 이창근 외(2017)

❑ 텍스트마이닝을 활용한 분석 사례(최동욱, 2017)

○ 연구문제: 포털이 이윤극대화를 위해 독자의 정치성향에 맞추어 뉴스를 배치하고 있는가?

○ 포털의 정치성향에 대한 논란이 많으나 그 원인에 대한 논의가 부족한 상황

○ 뉴스기사의 정치성향을 측정하기 위해 텍스트마이닝 기법을 활용

- 국회회의록에서 주요 키워드를 추출
- 복합명사구의 형태로 2015년 당시 국회의원이 상대정파의 의원과 차별화를 위해 전략적으로 사용한 키워드 식별(예: 올바른역사교과서, 손톱밑가시, 경제민주화, 법인세정상화 등)
- 발언자의 성향과 표현의 발언빈도를 활용하여 각 표현의 성향을 계산

$$\tilde{f}_{p,c} = \alpha_p + \beta_p y_c + \epsilon_{p,c}$$

- p: 키워드
- c: 발언자
- $\tilde{f}_{pc}$: 키워드 전체 빈도수 중에서 p가 차지하는 비율
- y_c: 소속정당더미
- β_p는 발언자의 소속과 표현의 사용비율 간의 관계를 의미
- 모든 선택된 키워드에 대해 파라미터를 추정

- 포털에서 수집된 뉴스기사에서 각 키워드의 등장 빈도를 측정하고 위에서 추정한 파라미터를 대입하여 포털뉴스의 성향을 역으로 추정

$$\tilde{y}_n = \mathrm{argmin}_{y_n} \sum_p \left(\tilde{f}_{p,n} - \alpha_p - \beta_p y_n\right)^2$$

- $\tilde{y}_n$이 최종적인 포털 n의 편향도
- 만약 포털n이 국회의원이었다면 가졌을 성향을 보여주는 변수

○ 한국사회과학센터의 정치성향 여론조사 자료를 활용하여 일반 유권자의 인구통계특성에 따른 평균적인 정치성향의 변화를 추정

○ 포털 뉴스섹션 접속자의 정치성향을 추정

- 한국 닐슨에서 포털뉴스 접속자에 대한 정보를 수집
- 한국사회과학센터의 여론조사 자료로부터 추정한 파라미터를 역으로 대입하여 포털 접속자의 정치성향을 추정

○ 포털의 경제적 유인구조를 확인

$$z_{n,i,t} = \gamma d_{n,i,t} + \xi_n + \nu_i + \eta_t + e_{n,i,t}$$

- $z_{n,t}$: 사용자 i의 포털n에서의 클릭수
- $d_{n,i,t}$: 앞에서 구한 포털뉴스의 성향과 접속자의 성향 차이

- 위의 회귀식 분석결과 포털뉴스와 접속자의 성향 차이가 감소할수록 포털에서의 클릭수가 증가하는 것을 확인
- 즉 소비자들은 자신의 성향에 가까운 뉴스일수록 더 많이 클릭한다는 의미

❑ 이처럼 빅데이터를 활용하여 인과관계를 파악하는 심층적 분석이 가능

○ 행정데이터의 연계를 통해 개인수준의 구체적인 자료를 구축할 수 있으며 이를 이용하여

정책효과나 가격변화의 영향 등을 정확하게 추정할 수 있음

- 다만 개인정보보호의 문제가 존재하므로 이에 대한 고려가 반드시 필요

○ 텍스트분석방법을 이용하여 기존에는 알 수 없었던 새로운 정보를 파악할 수 있으며 이를 통해 새로운 통찰력을 제공

- 인터넷 미디어 콘텐츠의 질적 특성을 정량적으로 추정하고 이를 이용해서 포털의 경제적 유인구조를 검증

○ 그 밖에도 다양한 소스의 빅데이터를 활용하는 것이 가능하지만 심층분석을 위해서는 일반적인 정형자료들과의 연계 및 정치한 경제모형이 필요

- 대부분 방대하고 복잡한 분석이 요구됨

❑ 시사점

○ 단순 상관관계가 아닌 인과관계를 실증적으로 보여줄 수 있는 방법론을 적용하는데 있어서 빅데이터의 활용을 통해 실제 경제현상이나 정책효과에 대한 문제 중 기존에 해결하지 못했던 문제를 해결할 수 있는 가능성이 열림

○ 경제문제의 해결에 있어서 행정데이터와 연계를 통해 대표성을 확보하고 상당히 구체적인 수준의 개별 정보를 추적할 수 있다는 장점이 있기 때문에 이를 활용하여 정책효과를 분석하는데 기여할 수 있는 잠재력이 존재함

○ 특히 텍스트마이닝 등 비정형데이터를 활용하여 기존에 측정하지 못했던 현상에 대한 지표를 생성할 수 있음

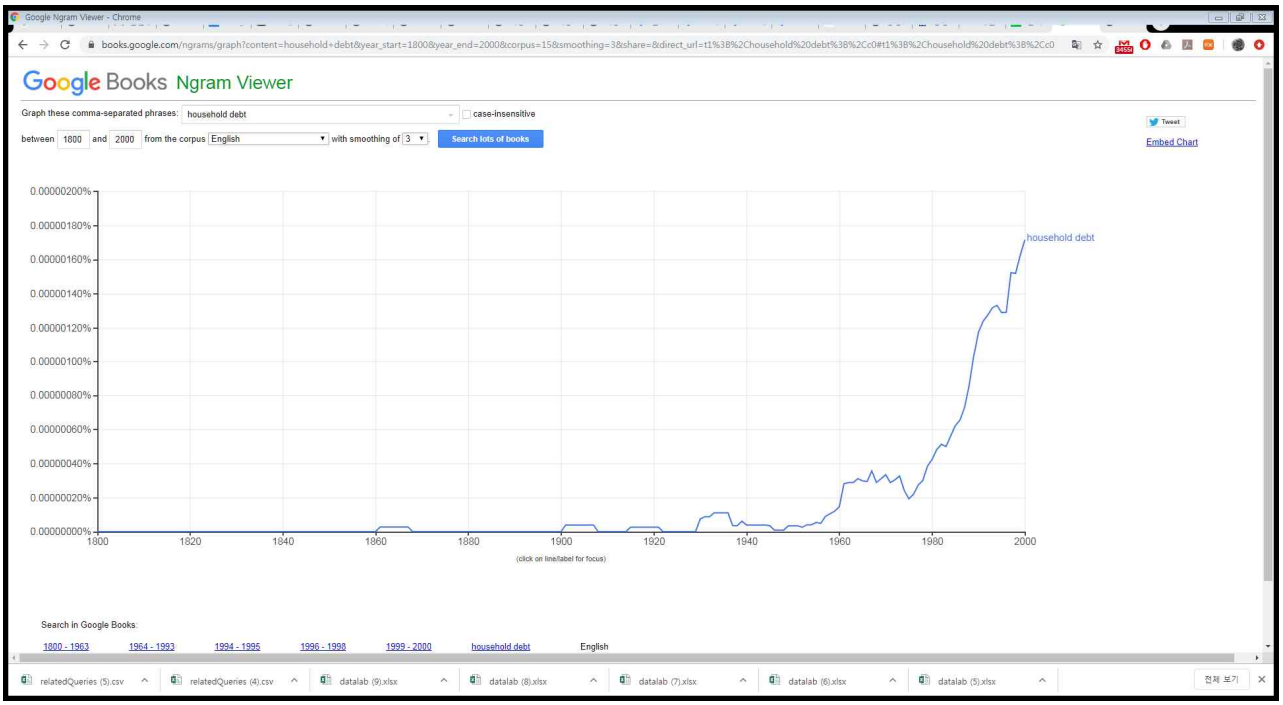
Ⅳ. 경제분야 빅데이터 분석의 적용 사례

1. 소셜 빅데이터 분석: 가계부채 이슈를 중심으로

❑ 가계부채는 비교적 최근에 발생한 사회적 이슈라고 할 수 있음

○ 구글 엔그램(Google Ngram)[4]을 이용하여 가계부채(household debt)와 관련된 용어가 이용된 빈도를 시대적 추이로 살펴본 결과 아래 그림과 같이 1980년대부터 급격하게 가계부채 용어가 사용되고 있음

[그림 IV-1] 가계부채(household debt)에 대한 언급 회수



출처: Google Ngram Viewer

❑ 다음에서는 구글 및 네이버 등 검색엔진에서의 개인 검색기록과 같은 소셜 빅데이터를 활용하여 가계부채와 관련된 주요 이슈들의 분석 사례를 제시함

4) 엔그램 뷰어에서 한글 검색은 불가하며 영문으로만 가능하며, 구글 엔그램과 관련된 보다 자세한 사항은 부록의 빅데이터 분석도구 내용을 참조

○ 소셜 빅데이터 분석에 있어서 활용될 수 있는 다양한 분석수단(tool-kit)[5]을 제시하고 이를 이용하여 가계부채와 관련된 주요 이슈들에 대한 분석을 시도함

❑ 본 절에서는 가계부채의 다양한 이슈들 중 먼저 가계부채의 건전성을 파악할 수 있는 개인회생 및 파산 신청 및 전략적 파산 가능성을 소셜 빅데이터를 활용하여 분석

○ 검색기록과 같은 소셜 빅데이터를 이용하여 개인들의 채무조정 행태를 분석하고 예측할 수 있는지 등을 살펴봄

❑ 다음으로 그동안 서베이조사로 파악해 왔던 가계의 대출수요를 소셜 빅데이터를 활용하여 파악하고 검색지표가 가계부채 증가에 대한 추가적인 설명력이 있는지를 파악

○ 대출태도조사[6]와 같은 금융기관 대상 서베이조사에서 일부 파악되고 있는 대출수요에 대한 정보를 검색기록과 같은 소셜 빅데이터를 활용하여 파악할 수 있는지 그리고 이러한 대출수요에 대한 정보가 지속적으로 늘고 있는 가계부채 증가를 설명하는데 있어 추가적인 설명력이 있는지를 파악

❑ 한편 이와 같은 인터넷 검색 자료는 이전 장에서의 설명에서 보는 바와 같이 경제분석 특히 경제예측을 위한 중요한 자료 원천이 될 수 있음을 확인

○ McLaren and Shanbhogue(2011)은 영국의 경우 2011년 60% 이상이 매일 인터넷을 활용하고 있어 검색자료가 실업률 등 다양한 경제변수들의 예측에 활용될 수 있음을 밝힘

○ Burdeau and Kintzler(2017)도 프랑스의 경우 주택대출수요를 예측하는데 있어 검색지표가 추가적인 설명력이 있음을 확인

❑ 다만 인터넷 검색자료가 비교적 최근에 나타나고 있고 이용자가 소득이나 연령대와 밀접한 연관이 있어 대표성을 갖기 어렵다는 한계가 있음

○ 구글 트렌드 검색기록 자료의 경우 2004년부터 사용이 가능하고 우리나라의 구글 네이버 검색기록은 2016년부터 이용이 가능하며 다음 검색기록은 유료이며 1년 정도의 검색기록 입수가 가능함

---

5) 이에 대한 간단한 소개는 부록의 빅데이터 분석도구 내용을 참조

6) 한국은행이 매 분기 발표하는 대출태도지수 자료로 국내은행의 여신업무 총괄담당 책임자를 대상으로 대출태도, 신용위험, 대출수요에 대한 금융기관의 동향 판단 및 향후 전망을 지수화 하여 다음 분기 초에 발표하고 있음(조사에 대한 보다 자세한 설명은 본장의 3)주택담보대출 수요 분석을 참고)

○ 또한 질문이나 주제에 따라 다른 검색 엔진을 사용하기도 하고 동일한 주제에 대한 상이한 핵심어휘를 사용하는 등 분석하기 어려운 한계가 있음

○ 우리나라의 경우 아래 표에서 보는 바와 같이 성인인구의 90% 이상이 인터넷을 사용하고 있어 그만큼 대표성을 갖는다고 할 수 있음

[표 Ⅳ-1] 인터넷 사용 추이(통계청 사회지표)

연령별	2016		2017		2018	
	이용	이용안함	이용	이용안함	이용	이용안함
3-9세	82.9	17.1	83.9	16.1	87.8	12.2
10대	100.0	-	99.9	0.1	99.9	0.1
20대	99.9	0.1	99.9	0.1	99.9	0.1
30대	99.8	0.2	99.9	0.1	99.9	0.1
40대	99.4	0.6	99.7	0.3	99.7	0.3
50대	94.9	5.1	98.7	1.3	98.7	1.3
60대	74.5	25.5	82.5	17.5	88.8	11.2
70세 이상	25.9	74.1	31.8	68.2	38.6	61.4

출처: 과학기술정보통신부·한국인터넷진흥원(2019)

1) 개인회생 및 개인파산 신청 분석

가. 검토배경

❑ 가계부채가 지속적으로 증가함에 따라 가계부채의 건전성에 대한 우려가 제기[7)]

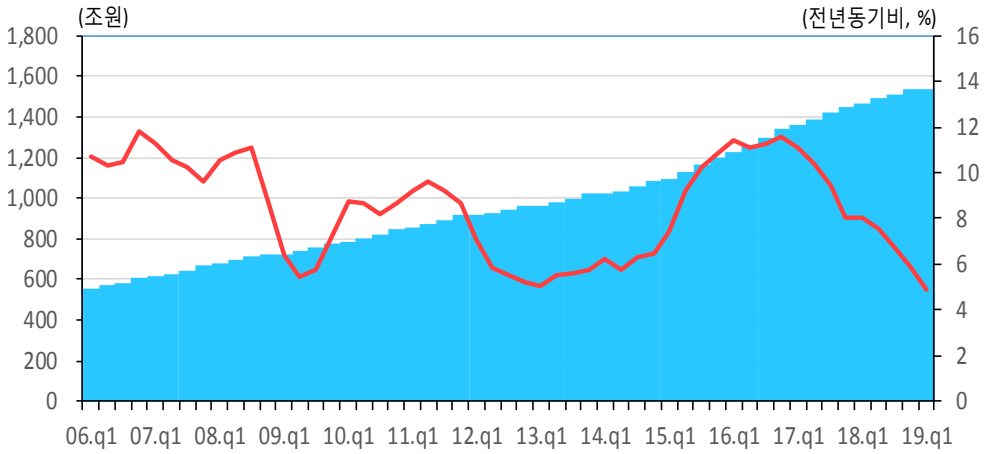
○ 미국 등 주요 국가들이 서브프라임 사태이후 디레버리징을 경험한 반면 우리나라의 가계부채는 지속적으로 증가함

○ 우리나라 가계부채는 주택규제 완화 등으로 2012년 이후 빠르게 증가하다가 2017년 이후 증가율이 낮아지면서 최근 5% 아래로 둔화

○ 이와 같이 지속적으로 늘고 있는 가계부채의 리스크 요인에 대한 우려가 지속적으로 제기

7) 가계부채의 건전성에 대한 우려 및 분석은 한국은행에서 반기마다 발표하는 한국은행 금융안정보고서를 참조(이하 수치 및 내용은 한국은행(2019)을 인용)

[그림 IV-2] 가계신용 추이

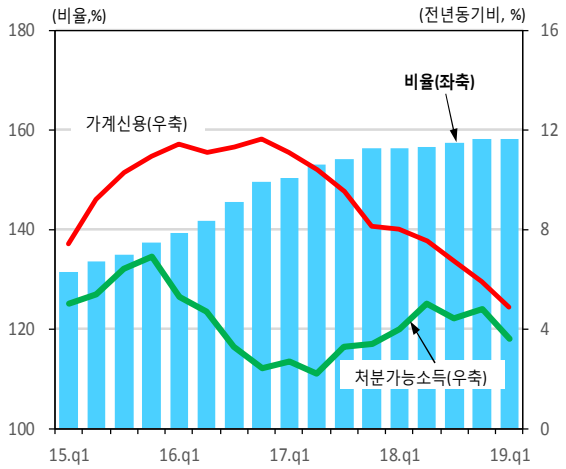


자료: 한국은행

❑ 최근 가계의 채무상환부담은 부채 증가율이 소득 및 금융자산 증가율을 상회하면서 다소 늘어났음

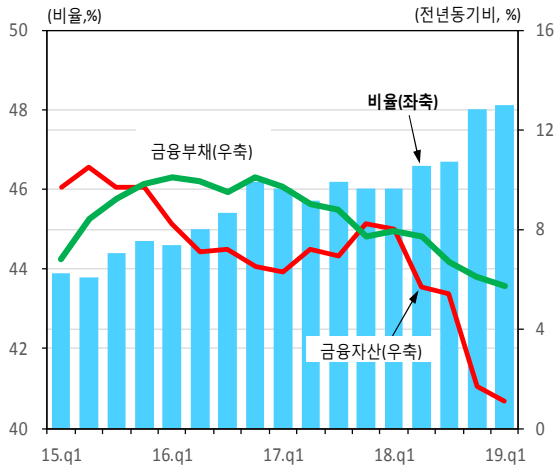
○ 가계부채(가계신용 기준) 증가율이 낮아졌으나 소득 및 금융자산 증가율을 상회하면서 가계의 채무상환 부담은 늘어난 것으로 보임

[그림 IV-3] 처분가능소득 대비 가계부채 비율



자료: 한국은행

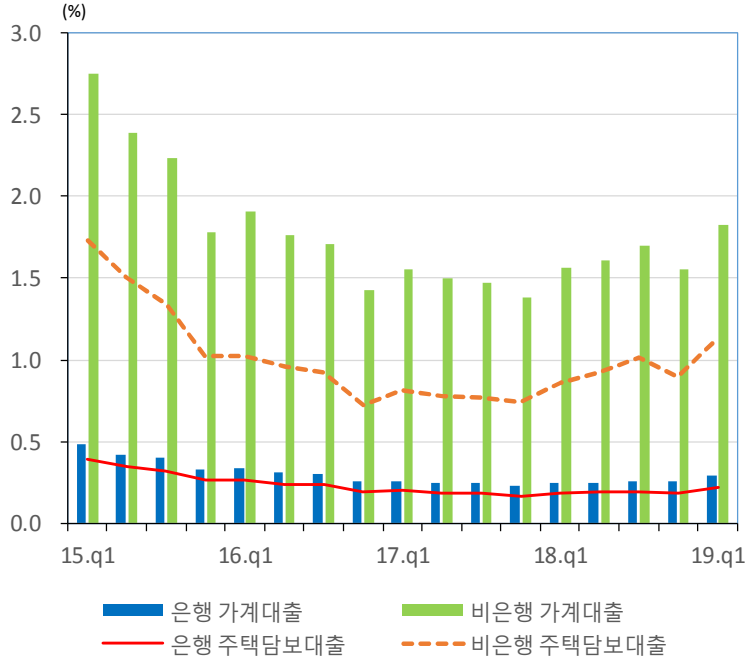
[그림 IV-4] 금융자산 대비 금융부채 비율



자료: 한국은행

❑ 가계 신용위험은 여전히 낮은 수준을 유지하고 있으나, 최근 연체율이 상승 움직임을 보이고 있어 이에 대한 우려가 지속되고 있음

[그림 Ⅳ-5] 은행 및 비은행금융기관 가계대출 연체율



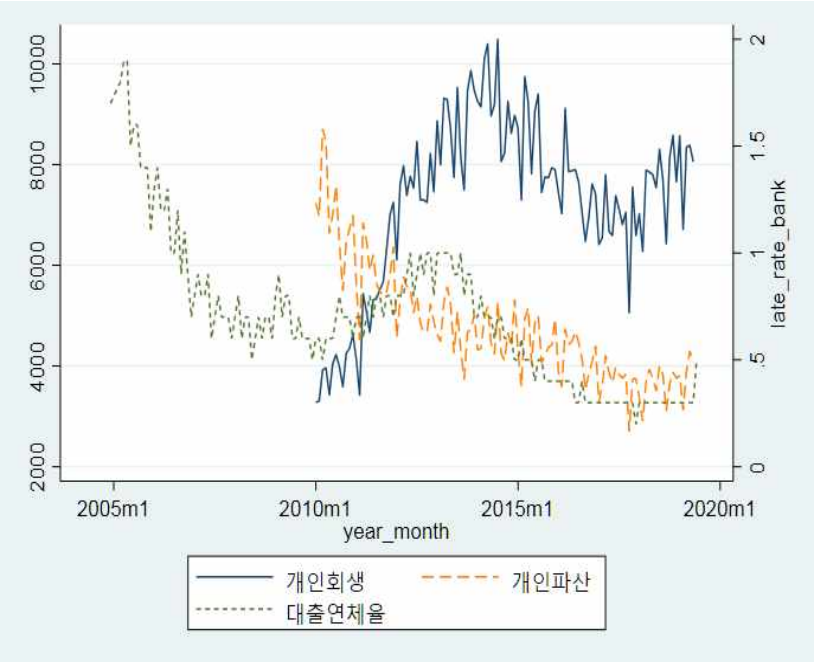
자료: 한국은행

❑ 가계대출 연체율과 함께 가계부채의 건전성을 나타내는 개인회생 및 개인파산 추이를 살펴보면 개인파산은 지속적으로 감소하는 모습을 보이고 있는 반면 개인회생은 시기에 따라 변화하는 모습[8)]

- ○ 개인파산은 2010년까지 월 8천 건 가까운 신청자 수를 보였으나 개인회생제도가 활성화되면서 현재는 4천 건 정도로 감소
- ○ 개인회생의 경우 2010년엔 월 3천 건에서 시작하여 최근 7천에서 8천 건 내외에서 등락을 보이고 있음
- ○ 개인파산 및 개인회생 실적은 대출연체율과 어느 정도 관계를 갖는 것으로 보임

8) 개인파산 및 개인회생 실적 등은 가계부채 관련 정책과 긴밀한 관계를 맺을 수 있으며 관련 부동산 및 가계부채 관련 정책은 부록을 참조

[그림 IV-6] 개인파산, 개인회생 월별 신청 추이



주: Y 축의 좌측은 개인파산 및 개인회생 월 접수 건수, 우측은 대출 연체율(%)
출처: 법원행정처, 한국은행 ECOS

❑ 가계부채의 건전성을 나타내는 지표 중 하나인 개인파산 및 회생 실적은 통상 개인들의 채무상환 문제가 발생한 이후 어느 정도 시간이 흐른 뒤에 발생하기에 실제 가계부채 문제가 심각해진 이후로 나타나는 후행적 성격이 있음

- ㅇ 개인파산 및 회생 신청 및 인용 실적은 서류준비와 절차 등을 감안할 때 실제 채무상환 문제가 발생한 시점보다 1~2년 정도 후행하는 것으로 나타남[9)]

❑ 본 소절에서는 개인파산, 개인회생, 신용회복 등에 대한 검색기록이 개인파산이나 개인회생 등 채무조정 실적을 예측하는데 사전적인 정보를 제공할 수 있음을 다양한 실증분석 방법을 이용하여 검토

- ㅇ 통상 개인파산이나 회생 등 채무자구제제도는 법적 규정, 절차와 비용 등 일반인들이 파악하기 어려운 내용 등이 있어 이에 대한 검색이나 자문이 필요하고 이들에 대한 검색이 사전적으로 이루어질 가능성이 높음

9) 개인파산 결정과 관련된 보다 자세한 논의는 유경원(2006, 2015a)을 참조

나. 방법론 및 분석결과

① 방법론

❑ 본 절에서 활용하고 있는 검색기록은 구글 트렌드(Google Trends)에서 제공하고 있는 자료[10]로 2004년부터 월별 자료로 입수가 가능한 것으로 나타남

- ㅇ 한편 제공 되어지는 월별 검색지수는 정해진 분석기간과 지역에서 구글에서 행해진 모든 검색기록에 대한 일별 무작위표본(random sample)에 의한 정규화된 수치이므로 사용시 마다 표본이 추출되어 사용일자에 따라 다른 결과가 나타날 수 있음
  - 본 연구에서는 연속 3일간의 검색지수를 평균하여 사용함[11]
- ㅇ 구글 검색자료는 2004년 1월부터 이용가능하지만 우리나라 법원의 개인파산 및 개인회생 신청자료는 2010년 1월부터 입수가능하므로 본 연구의 분석기간은 2010년 1월부터 2019년 5월까지임

❑ 본 연구에서는 우리나라에서 시장점유율이 상대적으로 낮은 구글의 검색기록을 사용하고 있는데 이는 네이버 검색기록의 시계열이 상대적으로 시계열이 짧은데 기인함

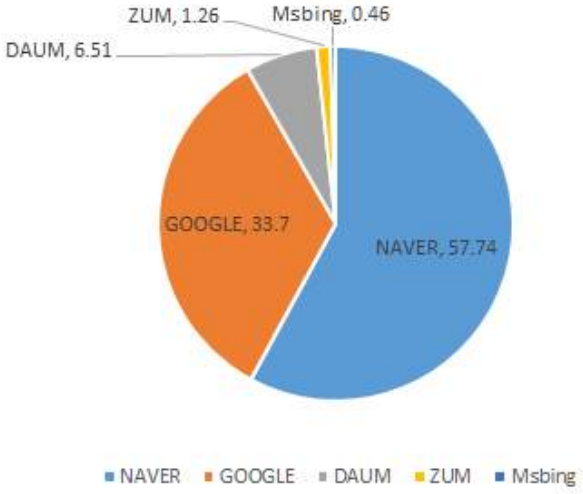
- ㅇ 최근 들어 구글 검색엔진의 시장 점유율이 30%까지 높아지고 있으나 아래에서 보는 바와 같이 네이버의 점유율이 기간 평균 70%이상을 차지하고 있음
- ㅇ 본 연구에서 입수가능한 개인파산 및 개인회생 실적치가 2010년부터 월별자료로 입수 가능한 반면 네이버에서 제공하는 검색기록은 2016년 1월부터이므로 충분한 시계열 관측치 확보를 위해 구글 검색기록을 사용함
- ㅇ 향후 데이터가 축적되거나 일별 자료를 활용할 경우 우리나라에서 시장점유율이 높은 네이버 검색기록을 사용해서 분석할 필요가 있음
  - 본 연구의 2절에서는 네이버 일별 검색기록을 사용하여 분석을 수행함

---

10) 구글 트렌드에서 제공하는 검색지수(index)는 실제 검색건 수를 제공하는 방식이 아니고 (“용어” 검색수)/(전체 검색어수)로 계산되며 주어진 기간과 지역내 일/주/월 의 최고치를 100으로 삼고 이후 숫자를 조정하는 방식

11) McLaren and Shanbhogue(2011)은 구글 트렌드에서 검색할 경우 검색 단어가 빈도가 낮은 단어일수록 변동성(volatility)이 크게 나타날 수 있으므로 이와 같은 일자를 보다 많이 늘려 평균화해서 이용하는 것이 바람직할 수 있다고 제안

[그림 IV-7] 인터넷 검색엔진 최근 점유율(2019.8.12.~2019.9.2.)



출처: BizSpring(2019)

[표 IV-2] 네이버와 구글 점유율 평균 추이(2010~2019)

검색엔진	기간평균 점유율 (%)	최대비율 (%)	최소비율 (%)
NAVER	75.24	87.9	53.7
Google	5.02	53.7	0.2

출처: BizSpring(2019)

❑ 한편 구글 트렌드 등을 통한 검색자료는 속보성, 추가정보 제공 등의 장점이 있는 반면 단점들도 존재하므로 유의하여 사용할 필요

- ㅇ 검색데이터는 특정 관심 주제에 대한 시간, 지역별 추이 관찰이 용이하고 고정된 질문을 사용하는 조사자료에 비해 특정 사건 전후의 변화를 신속하고 유연하게 분석할 수 있는 장점이 있음
- ㅇ 검색 데이터는 소셜미디어 데이터와 같이 모집단의 대표성이 부족할 수 있고 검색 단어의 선택에 민감하며 잡음이 많다는 단점이 있으므로 보다 정교한 작업이 요구됨

❑ 본 소절에서는 이와 같은 검색자료의 장점과 한계를 인식하고 보다 정교한 작업을 위해 단계적으로 기술분석과 시계열회귀분석을 수행함

- ㅇ 먼저 주요 변수들에 대한 시계열 추이를 확인하고 단순상관관계 분석을 수행하여 변수들

간 관계의 정도를 파악함

- 아울러 그랜저 인과관계 분석을 수행하여 기존 논의에서 제기되는 바와 같이 검색지표와 실적치 간의 선행관계를 살펴봄
- 시계열회귀분석에 앞서 주요 변수들의 단위근 검정을 수행하고 McLaren and Shanbhogue(2011)를 원용하여 기본 분석모형을 다음과 같이 설정함

$$R_t = \alpha + \beta_1 R_{t-1} + \beta_2 SR_{t-r} + X_t \Gamma + D_m + \epsilon_t$$

- 여기서 R은 개인회생 및 개인파산의 월 신청건수 그리고 SR은 이와 같은 개인회생 및 개인파산의 월 검색지표, X는 이와 같은 채무조정을 설명하는 경제변수들의 벡터로 은행연체율, 은행대출금리, 실업률 등을 포함하고 D는 월별(m) 더미변수를 의미

❑ 한편 기존 연구결과[12] 등에 따르면 가계부채 관련 현재 지표는 통상 파산이나 회생과 같은 채무조정 실제 신청 수보다 1년~2년 정도의 시차를 두고 선행하는 것으로 나타남

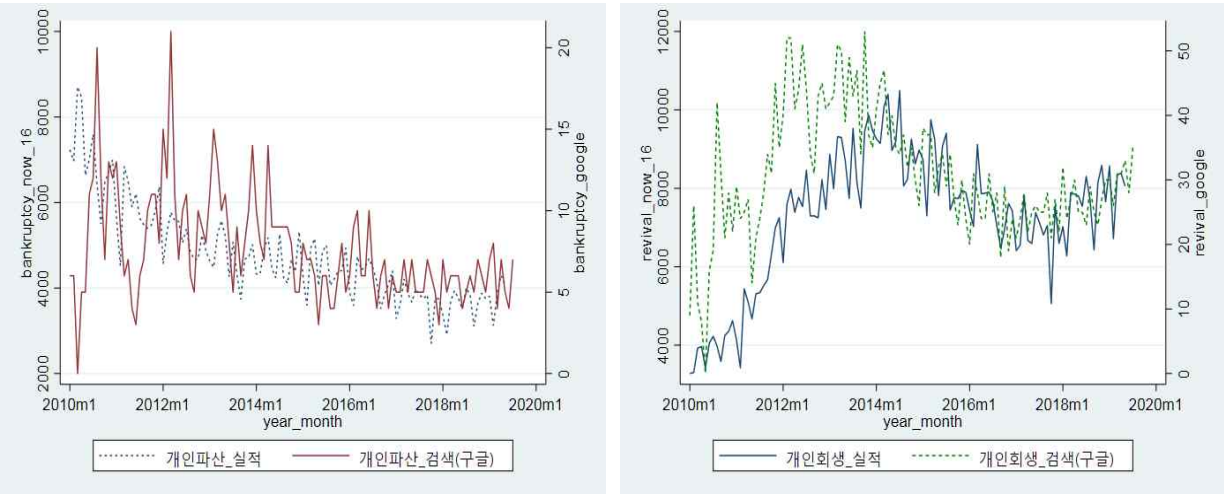
- 통상적으로 채무조정의 속성상 가계부채의 건전성이 위협되는 순간 바로 채무조정을 신청하기 보다는 시차를 두고 신청하는 것이 일반적이라 할 수 있음
- 따라서 본 연구에서는 이와 같은 속성을 파악하여 현재기의 검색실적 뿐이 아니라 1년과 2년에 시차를 둔 검색변수를 설명변수로 추가하고 이들 변수의 통계적 유의도와 설명력을 파악하여 추가적인 정보 제공 유무를 판단할 것임

② 분석결과

❑ 아래 그림에서 보는 바와 같이 개인파산과 개인회생의 검색기록과 실적치가 어느 정도의 관계가 있는 것으로 나타남

12) 보다 자세한 내용은 Congressional Budget Office(2000) 및 유경원(2006)을 참조

[그림 IV-8] 개인파산과 개인회생 실적과 검색 추이



주: Y 축의 좌측은 개인파산 및 개인회생 월 접수 건수, 우측은 구글 개인회생 및 개인파산 검색지표
출처: 법원행정처, Google Trends

❑ 상관관계 결과에서 나타나는 바와 같이 대표적인 두 채무조정 실적은 개인파산 및 개인회생 검색기록과 어느 정도 관계를 나타내는 것으로 보임

○ 개인회생에 대한 검색지표는 개인회생 실적치와 0.61의 상대적으로 높은 상관관계를 나타내고 있는 반면 개인파산에 대한 검색지표는 0.08의 낮은 상관관계를 나타냄

[표 IV-3] 채무조정 제도 실적과 검색기록과의 상관관계

변수명	개인회생	개인파산
개인회생(검색)	0.6087	-0.2020
개인파산(검색)	0.1976	0.0830

❑ 개인파산 및 회생에 대한 실적치와 검색지표에 대한 단위근 검정 결과 단위근이 있다고 하는 귀무가설을 기각하였으며 두 변수간 그랜저 인과관계 분석을 시행한 결과 개인회생에 대한 검색기록은 실적치를 그랜저 인과하는 것으로 나타남[13)]

❑ 아래 표에서 보는 바와 같이 단순회귀분석을 통해 이들 채무조정 실적치에 대한 검색지표의 유의도를 살펴본 결과 개인회생에 대한 결과가 보다 나은 것으로 나타남

13) 단위근 검정결과와 그랜저 인과관계 분석 결과는 부록의 부표를 참고

○ 개인회생의 경우 검색지수의 금기(t) 즉 현재 달뿐이 아니라 1 개월 전(t-1) 수치와도 유의한 관계를 나타내고 있는 반면 개인파산의 검색지수는 상수 이외에 통계적으로 유의한 관계를 나타내고 있지는 못하고 있음

❑ 이에 따라 본 소절에서는 개인회생에 대한 분석 결과를 중심으로 설명

○ 개인파산에 대한 분석도 동일하게 수행하였지만 기술분석 결과에서 크게 차이가 나지 않았음

[표 IV-4] 개인회생과 개인파산의 단순회귀분석 결과

변수명	개인회생 실적치	개인파산 실적치
검색지수	39.48**	10.10
	(16.34)	(19.98)
검색지수$_{(t-1)}$	45.18***	5.1143
	(15.47)	(19.15)
상수항	3513.93***	4254.05***
	(583.09)	(611.93)
관측치 수	113	113
R-squared	0.5004	0.1085

주: 1) 검색지수의 하첨자 (t-1)은 t기로부터 1개월 전을 의미
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

❑ McLaren and Shanbhogue(2011) 방법을 원용하여 검색지수의 추가 설명력을 분석한 결과 개인회생의 경우 어느 정도의 설명력을 가지고 있는 것으로 나타남

○ 모형의 설명변수로 12개월과 24개월의 시차변수를 함께 넣어 분석한 결과 모두 통계적으로 유의하게 나타났으며 R^2 값도 0.86까지 높아져 설명력이 매우 높게 나타남

[표 IV-5] 시차변수 추가한 회귀분석 결과

변수명	개인회생 실적치
검색지수	47.86***
	(5.7154)
검색지수$_{(t-12)}$	36.50***
	(5.9840)
검색지수$_{(t-24)}$	36.69***
	(4.0225)
상수항	2109.58***
	(328.81)
관측치 수	113
R-squared	0.8606

주: 1) 검색지수의 하첨자 (t-12)와 (t-24)는 t기로부터 12개월과 24개월 이전을 의미
2) 이전( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

❑ 다음으로 월별 자료를 이용한 단순 OLS 모형을 확장하여 자기회귀(AR)모형을 기본모형으로 하고 여기에 검색지수들을 추가할수록 모형의 설명력은 높아지는 것으로 나타남

- ㅇ 모형1은 벤치마크 모형으로 자기회귀(AR) 즉 바로 직전 달(t-1)의 실적치로 현재의 실적치를 설명하는 모형이며, 여기에 검색기록을 추가했을 때의 설명력의 증가를 살펴봄
- ㅇ 아래 표에서 모형 1(AR 모형)의 R^2 값이 0.83에서 모형2~3에서 보는 바와 같이 검색기록의 현재와 1, 2년 전 수치를 추가했을 경우 R^2 값이 0.89로 증가함

[표 Ⅳ-6] 개인회생 추정 결과(1) : 검색기록 설명변수

1) 모형1: AR

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.8834***
	(0.0369)
상수항	781.19*
	(416.12)
관측치 수	112
R-squared	0.8311

주: 1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수를 나타냄
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

2) 모형2: 검색기록 추가

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.7859***
	(0.0451)
검색지수$_{(t)}$	22.13***
	(6.1214)
상수항	490.21
	(383.36)
관측치 수	112
R-squared	0.8528

주: 1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수, 검색지수의 하첨자 (t)는 현재 달의 검색실적을 나타냄
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

3) 모형3: 검색기록 시차변수 추가

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.4245***
	(0.0755)
검색지수$_{(t)}$	32.22***
	(6.1263)
검색지수$_{(t-12)}$	17.23***
	(6.2110)
검색지수$_{(t-24)}$	21.93***
	(4.7312)
상수항	1132.84***
	(350.29)
관측치 수	112
R-squared	0.8907

주: 1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수, 검색지수의 하첨자 (t)와 (t-12), (t-24)는 각각 현재 달과 12개월전, 그리고 24개월전의 검색실적을 나타냄
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

❑ 추가적으로 검색지표와 다른 경제설명변수와의 설명력 비교를 위해 위 추정식에 연체율과 실업률 변수를 설명변수로 추가하여 분석함

- 다음 표에서 보는 바와 같이 개인회생의 실적을 설명하는 경제변수들은 은행 연체율을 제외하고는 통계적으로 유의하지 않게 나타남[14)]
- 이와 같은 경제설명변수에 검색지수를 추가하였을 경우 그나마 나타나는 유의성도 나타나지 않게 되어 개인회생 등을 설명하는 주요 변수들에 비해서도 이와 같은 검색기록은 추가적인 정보를 제공한다고 할 수 있을 것임

14) 월별 경제변수로 가계대출연체율, 가계대출이자율, 가계대출증가율, 실업률 등 본 소절에서 사용되는 주요 경제변수들은 한국은행 ECOS DB에서 추출하였으며 요약통계량은 부록에 제시

[표 Ⅳ-7] 개인회생 추정 결과(2) : 경제변수 추가

1) 모형4: AR

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.8971***
	(0.0398)
은행연체율$_{(t)}$	859.94*
	(516.40)
실업률$_{(t)}$	666.87
	(412.51)
상수항	-2185.25
	(1826.25)
관측치 수	112
R-squared	0.8358

주:1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수를 나타내고 경제변수들의 하첨자 (t)는 현재 달 수치임
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

2) 모형5: 검색기록 추가

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.7360***
	(0.0555)
검색지수$_{(t)}$	36.57***
	(7.9865)
은행연체율$_{(t)}$	-436.36
	(572.77)
실업률$_{(t)}$	564.43
	(391.18)
상수항	-1578.96
	(1713.04)
관측치 수	112
R-squared	0.8656

주: 1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수, 검색지수의 하첨자 (t)는 현재 달의 검색실적을 나타냄
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

3) 모형6: 검색기록 시차변수 추가

변수명	개인회생 실적치
실적치$_{(t-1)}$	0.3329***
	0.0703
검색지수$_{(t)}$	46.58***
	7.1829
검색지수$_{(t-12)}$	26.69***
	6.1474
검색지수$_{(t-24)}$	17.69***
	4.7582
은행연체율$_{(t)}$	-523.23
	514.85
실업률$_{(t)}$	702.17*
	357.72
상수항	-1296.92
	1521.67
관측치 수	112
R-squared	0.9072

주: 1) 실적치의 하첨자 (t-1)은 전월 개인회생 신청건수, 검색지수의 하첨자 (t)와 (t-12), (t-24)는 각각 현재 달과 12개월전, 그리고 24개월전의 검색실적을 나타냄
2) ( ) 안은 표준오차를 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
4) 월별 더미 변수 결과는 지면관계상 생략함

❑ 회귀분석에서 추정한 결과를 가지고 모형의 적합도를 살펴본 결과 아래 그림에서와 같이 실제치를 잘 추정하고 있는 것으로 판단됨

- ○ 아래 표에서 보는 바와 같이 경제변수들에 추가하여 검색기록을 설명변수로 추가하였을 때(모형 3과 모형 6) RMSE와 MPE 등 모든 지표들이 향상되는 결과를 나타내고 있어 검색기록의 추가적인 정보를 제공하고 있는 것으로 나타남
- ○ McLaren and Shanbhogue(2011)의 검색기록을 활용한 실업률 예측과 마찬가지로 검색기록은 개인들의 채무조정의 예측에 있어서도 유용하게 활용될 가능성을 시사함

[그림 IV-9] 개인회생 실제치와 모형 추정치 추이

[표 IV-8] 모형별 예측정확도 결과

	모형1	모형2	모형3	모형4	모형5	모형6
RMSE	689.01	643.24	554.30	679.33	614.64	510.86
MAE	532.57	508.51	443.78	525.63	480.31	401.29
MAPE	0.0772	0.0734	0.0638	0.0768	0.0697	0.0586
Theil's U	0.7384	0.6750	0.5723	0.7424	0.6448	0.5339

다. 시사점

❑ 기존논의에서 제기한 바와 같이 개인들의 채무건전성을 판단할 수 있는 개인회생 및 파산은 현재 가계부채 총량이나 건전성지표 등과 후행관계를 갖는 것으로 판단됨

- ㅇ 현재의 개인회생이나 파산 지표로 가계부채의 건전성에 대한 판단은 유의할 필요

❑ 개인회생에 대한 검색기록은 가계재무 상태의 건전성을 평가할 때 중요한 지표로 활용될 수 있을 것임

- ㅇ 검색기록에 대한 현재와 시차변수 등은 개인회생을 추정오차 5%내외에서 추정할 수 있으며 속보성 있는 자료의 속성상 현재의 가계재무 건전성에 대한 유용한 판단지표를 제시할 수 있는 것으로 판단됨

❑ 다만 파산에 대한 검색기록은 좀 더 정교한 추정기법과 자료를 활용하여 추가적인 연구를 통해 개선할 필요가 있음

- ㅇ 향후 시계열이 확장된 네이버 검색기록 등 다양한 검색자료와 기법을 활용하여 추가적인 연구를 통해 관계를 추정해 낼 수 있을 것으로 판단됨

2) 전략적 개인파산 분석

가. 검토배경

❑ 가계부채가 지속적으로 증가함에 따라 채무불이행 등에 따라 향후 사적·공적 채무조정제도의 활용이 늘어나게 되고 채무조정제도를 어떤 식으로 운영하느냐에 따라 가계부채발 경제위기의 영향이 달라질 수 있음[15)]

- ㅇ IMF(2012) 는 가계부채 위기와 관련된 기존 연구를 통해 파악할 수 있는 사항은 적절한 채무조정정책들이 디레버리징 기간의 경제활동에 있어 발생할 수 있는 과도한 위축을 막는데 있어 핵심적인 요소라고 주장
- ㅇ 특히 가계부채 위기와 관련하여 과감하고 잘 설계된 채무조정 프로그램이 가계파산과 압류 수를 대폭 줄일 수 있으며 이를 통해 경제위기로 인한 손실 내지 피해를 줄이는데(loss mitigation) 기여할 수 있었다고 평가

❑ 채무조정에 대한 정책의 완화 내지 강화에 대한 논란은 가계부채가 지속적으로 늘면서 커지고 있는데 여기에서의 관건은 채무자들의 모럴해저드 여부라 할 수 있음

- ㅇ 개인정보의 비대칭성을 활용하여 부채 상환을 하지 않거나 더 큰 부채를 유발하는 행동이 만연하게 될 경우 전반적으로 채무조정 기조가 강화되는 방향으로 작동할 수 있음
- ㅇ 모럴해저드와 같이 일종의 전략적 파산의 식별은 채무조정에 있어 관건이 되는 채무자들의 도덕적 해이 여부에 대한 정보 획득에 필요함

❑ 하지만 이에 대한 판단 근거가 충분한지 않으므로 본 절에서는 가계의 전략적 파산 행동을

15) 보다 자세한 내용은 유경원 외(2015)와 유경원(2015b)을 참조

검색기록을 통해 확인할 수 있는 가능성을 점검해 보기로 함

- ㅇ 전략적인 파산행위를 시사하는 검색어가 상대적으로 증대한다면 전반적으로 개인들의 모럴해저드가 확산될 가능성을 시사한다고 할 수 있으므로 이에 대한 채무조정 제도와 같은 정책적 대응을 강화할 필요가 있을 것임
- ㅇ 반면 뜻하지 않은 사고나 실업 등으로 어쩔 수 없는 소득 충격이 발생하여 상환여부가 불가능하게 되는 외생적, 비자발적 파산 행위를 시사하는 검색들이 많아지는 상황에서 만약 채무조정제도를 비탄력적으로 엄격하게 적용한다면 경제사회적으로 바람직하지 않은 결과를 낳게 될 것임
- ㅇ 미국의 경우 서브프라임 사태의 경기침체에 대한 영향이 보다 확대된 배경으로 이와 같은 채무조정 제도를 잘못 운영한데 기인한 것이 중요한 원인으로 대두되고 있음

❑ 따라서 본 절에서는 지속적으로 늘어나고 있는 가계부채가 부실화되었을 때 채무조정제도의 기조를 어떻게 운영할지에 대한 시사점을 얻기 위하여 개인들의 전략적 또는 외생적 파산 행태에 대한 분석을 검색자료를 이용하여 살펴보기로 함

나. 방법론 및 분석결과

① 방법론

❑ 소셜빅데이터의 장점이 보다 '솔직한' 검색기록을 제시하고 있으므로 연관어 분석 등을 통해 전략적인 행동과 외생적인 파산 행동을 식별하고 이들의 추이를 살펴봄

- ㅇ 인터넷 상에서 개인파산 검색시 이로 인한 편익, 비용 등을 함께 검색한다면 전략적 파산의 의도가 있는 것으로 판정하며, 그렇다면 여기에서의 관건은 '전략적'이라고 판별할 수 있는 연관어의 선택임
- ㅇ 관련 연관어의 선택 기준은 기존연구를 기반으로 정함

❑ 기존 연구[16]에 따르면 전략적 파산 행동을 나타나게 되는 근거가 바로 금전적 편익추구 행위라 할 수 있으므로 개인파산 검색 시 연관 검색어로 편익 추구와 관련된 검색을 선정

16) 보다 자세한 사항은 유경원(2006)을 참조

○ 기존 연구에 따르면 개인파산의 금전적 편익추구와 관련된 사항은 개인파산이 가져오는 '파산면책'이라고 할 수 있음

❑ 한편 전략적 파산의 대칭적인 개념으로 '어찌할 수 없는 비전략적인' 파산 행태를 외생적인 파산이라고 보았을 때 이와 관련이 있는 것은 개인파산을 빨리 실행할 수 있는 '파산절차'에 대한 검색이라고 할 수 있음

○ 이에 대한 구글 소셜데이터 분석수단을[17] 활용하여 연관어 검색기록을 살펴볼 수 있음

❑ 이를 기반으로 전략적 파산과 외생적(비전략적) 파산의 유형을 식별하고 성별, 연령대별 추이를 살펴보도록 함

○ 구글 트렌드에서는 검색빈도의 지역적 분포를 시각적으로 제공하고 있는 반면 네이버 데이터랩(DataLab)에서는 검색어의 연령별, 성별 분석이 가능하도록 정보를 제공함

❑ 아래 표에서 보는 바와 같이 개인파산과 관련된 구글 트렌드에서 제공하는 관련 검색어 빈도순위는 개인회생, 파산신청, 면책, 비용, 신청자격 순임

[표 IV-9] 개인파산 연관 검색어(Google Trends)

순위	관련 검색어	지표
1	파산	100
2	개인 파산	98
3	개인 회생	29
4	파산 신청	26
5	개인파산 신청	25
6	파산 면책	11
7	개인파산 면책	11
8	개인파산 비용	6
9	개인파산신청이란	5
10	개인파산 신청자격	5

❑ 전략적 파산 및 외생적 파산에 대한 기존 연구결과를 기반으로 연관검색어를 다음과 같이 설정함

17) 구글 애드워즈(Google Adwords) 및 구글 트렌드(Google Trends) 등 구글에서는 연관 검색어를 참고할 수 있는 분석수단을 제공하고 있음(이에 대한 보다 자세한 사항은 부록을 참고)

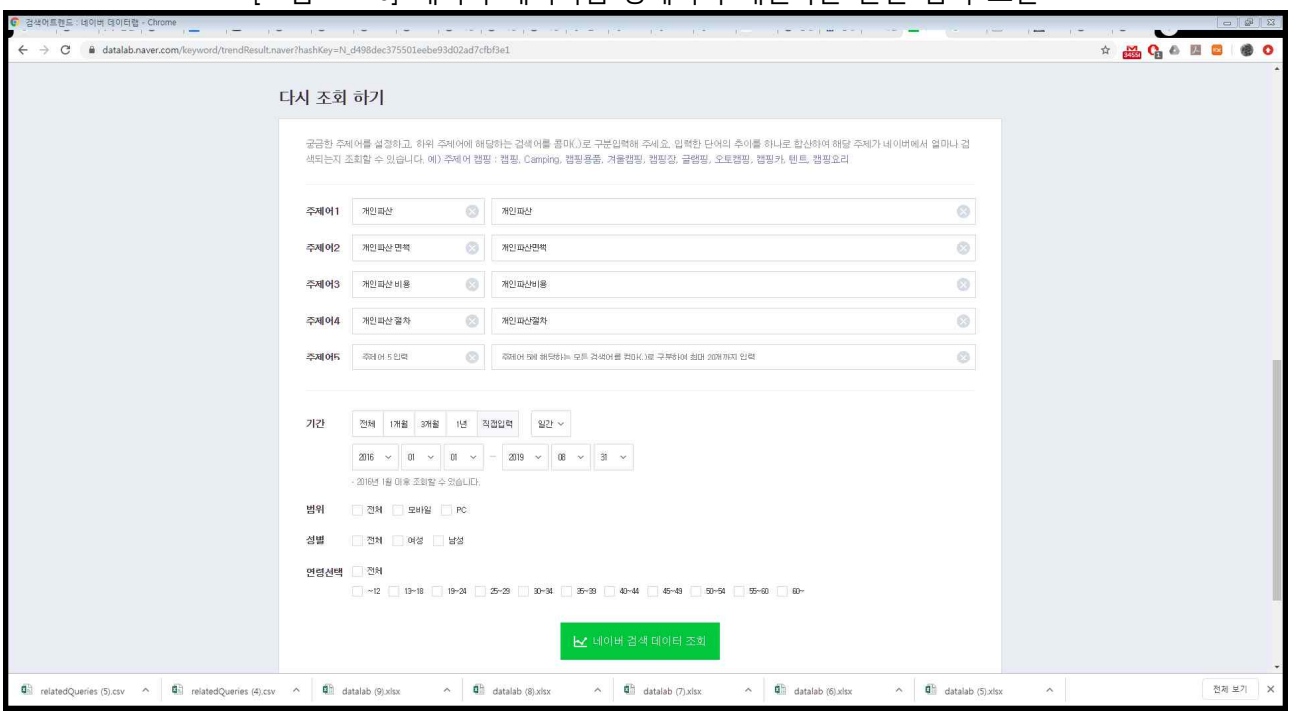
- ‘개인파산 면책’ 검색을 전략적 파산유형으로 ‘개인파산 절차’를 외생적 파산유형을 구분함
- ‘개인파산 비용’에 대한 검색은 전략적 파산과 외생적 파산 성격을 동시에 갖고 있는 것으로 보이므로 관련도를 살펴보고 판단함

❑ 본 소절에서는 이전 절과 달리 네이버 데이터랩(DataLab)을 이용하여 2016년 1월 1일부터 2019년 8월 31일까지 일별 검색자료를 추출하여 이를 분석에 활용

- 구글 트렌드와 달리 네이버 데이터랩의 경우 일에 따라 추출 자료가 달라지지는 않고 동일함
- 분석기간 동안에 검색어 중 가장 검색빈도가 높은 경우를 100으로 설정하고 이에 대한 상대적 비율을 계산하여 지수를 산정함

❑ 검색지표 추출을 위해 아래 그림에서 보는 바와 같이 ‘개인파산’, ‘개인파산면책’, ‘개인파산비용’, ‘개인파산절차’를 주제어로 함께 입력하여 조회 결과를 활용함

[그림 IV-10] 네이버 데이터랩 상에서의 개인파산 관련 검색 조건



출처: 네이버 DataLab(https://datalab.naver.com)

② 분석결과

❑ 먼저 추출된 검색기록을 기반으로 '전략적 파산'과 '외생적 파산' 용어를 정의함

- ㅇ 전략파산지수는 전체 개인파산 검색지수 대비 개인파산면책 검색지수 비율로 정의하고 외생파산지수는 전체 개인파산 검색지수 대비 개인파산절차 검색지수 비율로 정의
- ㅇ 외생적 파산과 전략적 파산의 추이를 하나로 나타내기 위해 외생파산 검색지표 대비 전략파산 검색지표의 비율을 구하고 이 추이를 분석함
- ㅇ 또한 성별, 연령대별 각 지수들의 추이를 함께 살펴봄

❑ 아래 표에서 보는 바와 같이 요약 통계를 기반으로 전체적인 검색지수를 살펴보았을 때 '개인파산 면책' 검색기록이 '개인파산 절차'에 비해 상대적으로 높은 것으로 나타나고 있음

- ㅇ 이는 전략적인 파산 신청 가능성이 상대적으로 높을 가능성을 시사함

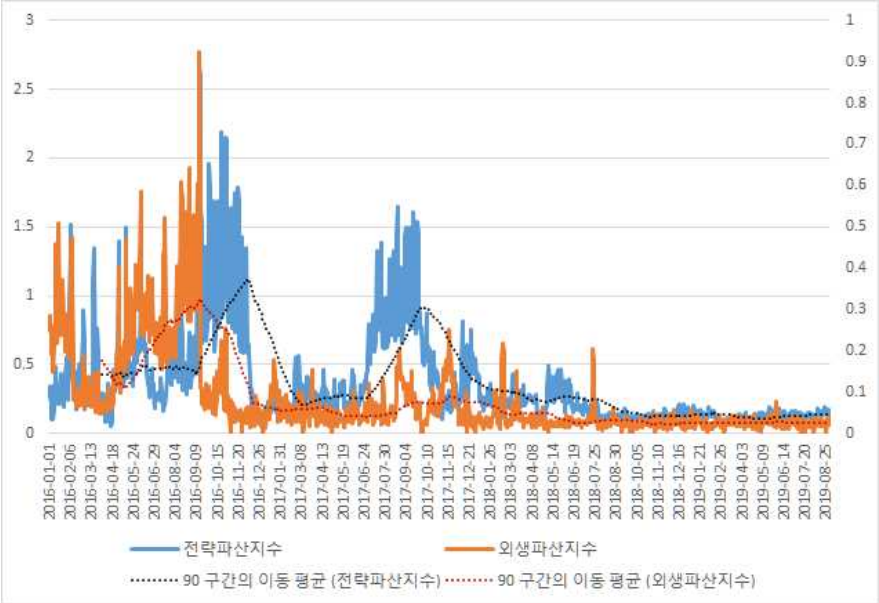
[표 IV-10] 개인파산 주요 검색 지표 및 지수(일별) 요약 통계량

변수명	관측치 수	평균	표준편차	최소값	최대값
개인파산	1339	17.62	12.52	3.19	100
개인파산 면책	1339	5.44	5.85	0.28	73.30
개인파산 비용	1339	1.76	3.19	0	25.55
개인파산 절차	1339	1.51	2.18	0	13.28
전략파산지수	1340	0.36	0.34	0.05	2.65
외생파산지수	1340	0.08	0.10	0	0.92

❑ 아래 그림을 통해 전체적인 추이를 살펴보면 2016년에는 외생파산지수가 상대적으로 높았으나 이후 전략파산지수가 높은 것으로 나타나고 있으며 2018년에는 둘 간의 차이가 그리 크지 않게 나타나고 변동성도 크게 축소된 것으로 나타남

- ㅇ 2016년의 경우 경제적 어려움에 따른 개인파산 절차에 대한 검색이 상대적으로 높았던 반면 2016년 후반부터는 파산의 편익을 살펴보는 검색들이 나타남
- ㅇ 2017년에는 개인파산의 절차에 대한 검색에 비해 개인파산의 편익을 구하는 검색이 크게 증가한 것으로 나타나고 2018년부터는 이와 같은 검색들이 상대적으로 안정적으로 나타남

[그림 IV-11] 전략파산과 외생파산 추이



❑ 전략파산지수와 외생파산지수를 성별로 구분해서 살펴보았을 때 아래 표에서 보는 바와 같이 전략파산의 경우 남성과 여성의 차이는 크게 나지 않고 있음

[표 IV-11] 성별 및 연령대별 전략파산지수 요약 통계량

변수명	관측치 수	평균	표준편차
남성	1339	0.15	0.07
여성	1339	0.16	0.08
20대	1339	0.11	0.06
30대	1339	0.13	0.06
40대	1339	0.19	0.11
50대	1339	0.25	0.20
60대	1331	0.32	0.39

[표 IV-12] 성별 및 연령대별 외생파산지수 요약 통계량

변수명	관측치 수	평균	표준편차
남성	1339	0.03	0.03
여성	1339	0.04	0.03
20대	1339	0.02	0.02
30대	1339	0.03	0.02
40대	1339	0.07	0.10
50대	1339	0.06	0.10
60대	1331	0.07	0.24

❑ 전략파산지수와 외생파산지수를 연령대로 구분해서 살펴보았을 때 연령대가 높아짐에 따라 증가하는 경향이 있는 것으로 나타남

- ㅇ 전략파산지수에 비해 외생파산지수는 청년층(20, 30대)과 중장년층(40대 이후)의 격차가 크게 발생함
- ㅇ 이는 편익에 대한 검색과 절차에 대한 검색이 연령에 따라 모두 늘어나고 있으나 외생파산을 시사하는 '파생 절차'에 대한 검색이 중장년층에 있어 급격히 증가함을 나타냄

[표 IV-13] 성별 및 연령대별 외생파산지수 대비 전략파산지수 요약 통계량

변수명	관측치 수	평균	표준편차
남성	1302	5.94	6.05
여성	1306	5.03	4.80
20대	1119	5.18	5.02
30대	1282	5.26	4.86
40대	1251	5.40	5.25
50대	1003	5.22	4.72
60대	535	3.35	3.40

❑ 외생파산지수 대비 전략파산지수의 비율을 살펴보면 40대가 가장 높은 것으로 나타나고 있으며 30대, 50대, 20대 순으로 60대가 가장 낮게 나타남

- ㅇ 상대적으로 경제활동이 많은 30~40대는 외생파산지수와 전략파산지수 모두 높게 나타나고 있지만 전략파산 검색이 상대적으로 높게 나타나고 있어 이들 연령계층에서 전략파산에 대한 행태가 나타날 가능성이 높음을 시사

○ 한편 20대 및 50~60대 이상의 계층은 반대로 경제충격 등 외생적인 요건으로 파산을 신청할 가능성이 높음을 시사

○ 이와 같은 결과는 전략적 파산 내지 외생적 파산의 선택이 연령과 밀접한 관계가 있을 가능성을 시사하는 결과라 할 것임

❑ 이와 같은 결과는 간단한 회귀분석을 통해서도 확인할 수 있음

○ 앞서의 분석에서와 같이 전략파산지수나 외생파산지수에 대한 연령대별 파산지수의 설명력은 40대 계층이 가장 높으며 통계적으로 유의하게 나타남

○ 외생파산 대비 전략파산 지수 비율에 대한 회귀분석 결과는 개별적인 분석과 달리 60대를 제외하고 모든 연령대에서 통계적인 유의성이 나타나고 이와 같은 비율의 계수값이 40대를 정점으로 역 U자 형태를 나타나고 있음

○ 성별에 대한 회귀분석은 별도로 제공하고 있지는 않으나 앞서의 요약통계에서 나타나듯이 전략적 파산의 경우 남녀 차이가 나타나고 있지 않으나 외생적 파산의 경우 남성이 통계적으로 유의하게 높은 것으로 나타남

[표 Ⅳ-14] 주요 지수와 연령 관계 회귀분석 결과

1) 전략파산지수

변수명	전략파산지수
20대	-0.1037 (0.1313)
30대	0.3284 (0.2061)
40대	0.6675*** (0.1271)
50대	0.1107* (0.0601)
60대	0.0583** (0.0277)
시간추세선	-0.0004*** (0.00001)
상수항	0.4179*** 0.0243
관측치 수	1131
R-squared	0.3320

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

2) 외생파산지수

변수명	외생파산지수
20대	-0.0033
	(0.0856)
30대	0.2370**
	(0.0977)
40대	0.4164***
	(0.0511)
50대	0.0255
	(0.0256)
60대	0.0063
	(0.0090)
시간추세선	-0.0001***
	(7.64e-06)
상수항	0.1160***
	(0.0093)
관측치 수	1131
R-squared	0.4900

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

3) 외생파산 대비 전략파산지수 비율

변수명	외생파산지수 대비 전략파산지수 비율
20대	0.0635*
	(0.0342)
30대	0.2288***
	(0.0486)
40대	0.2629***
	(0.0921)
50대	0.1171**
	(0.0530)
60대	-0.0318
	(0.0428)
시간추세선	-0.0019***
	(0.0005)
상수항	3.460***
	(0.4633)
관측치 수	375
R-squared	0.2499

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

다. 시사점

❑ 빅데이터의 장점인 ‘솔직한 자백약’ 특성을 활용하여 전략적인 파산과 외생적인 파산 가능성을 살펴 볼 수 있었음

- 개인파산 연관 검색어를 구하고 이를 기존 연구의 결과를 토대로 전략적인 파산과 관련이 있는 검색어를 정의하여 이를 전략 파산지수로 정의
- 이러한 정보는 일반적인 설문조사를 이용한다면 솔직한 응답을 기대하기 어렵고 편의가 발생할 가능성이 높음
- 하지만 포털 검색어는 필요성을 느낀(motivated) 이용자가 자발적으로 입력한 정보로서 입력한 주체의 의도가 솔직하게 반영되므로 편의가 적다고 볼 수 있음
- 아울러 동일한 방법으로 외생파산을 정의하고 이즐 지수들의 추이를 살펴봄

❑ 검색기록의 일별 추이를 통해 전략적인 그리고 외생적인 파산 가능성 추이를 살펴볼 수 있으며, 성별, 연령대별 검색기록을 통해 이와 같은 파산 가능성의 인구사회학적 특징을 일부 파악할 수 있었음

- 검색측면에서 보면 외생적인 파산보다는 전략적인 파산 행동을 보일 가능성이 높음을 시사
- 시기별로 보았을 때 이와 같은 검색은 큰 변동성을 보이고 있으나 2~3년 전에 비해 최근 전략적인 파산 가능성은 상대적으로 낮게 나타나고 있음
- 성별 차이를 보았을 때 전략적인 파산 검색 차이는 통계적으로 유의미한 차이를 나타나고 있지는 않았으나 외생적인 파산 검색은 남성이 상대적으로 높게 나타남
- 연령대별로 보았을 때 외생적인 파산 검색 대비 전략적인 파산비가 가장 높은 계층은 경제활동이 많은 40대 계층으로 나타남

❑ 이와 같은 결과가 시사하는 바는 전략적인 파산이나 외생적인 파산은 시기에 따라 변화할 수 있으며 개인들의 인구사회학적 차이에 따라 달리 시현될 수 있음을 나타냄

❑ 향후 이와 같은 검색기록을 통해 파산신청과 관련된 질적 분석이 가능할 수 있으며 이를 기반으로 개인채무구제정책의 방향성을 설정하는데 있어 참고자료로 활용할 수 있을 것임

- 현재 가계부채 취약가구가 지속적으로 늘어나고 있는 가운데 향후 이와 같은 검색기록에 대한 모니터링을 통해 향후 개인채무구제정책 방향을 설정하는데 참고자료로 활용

3) 주택담보대출 수요 분석

가. 검토배경

❑ 주요국가와 달리 우리나라 가계부채가 지속적으로 증가하고 있는 가운데 그동안 부채 증가 원인에 대한 다양한 분석이 수행되어 왔음[18)]

- ㅇ 대체적으로 거시적 요인 및 인구사회학적 요인 등을 분석한 거시계량 분석이나 미시계량 분석이 수행되어 왔음
- ㅇ 그러나 이들 분석들은 구조적인 요인이나 행태적인 측면을 분석하고 있어 방향성을 파악하는데 도움이 될 수 있으나 단기적인 수요측과 공급측 요인을 파악하기에는 어려움이 있었음

❑ 이를 보완하기 위해 한국은행 등 주요국의 중앙은행에서는 분기마다 대출태도조사를 수행하여 분기초에 발표함[19)]

- ㅇ 동 조사는 아래 참고자료에서 보는 바와 같이 은행 등 금융기관의 대출담당 임직원들을 대상으로 한 서베이로서 금융기관의 대출 태도에 대한 현황 및 전망을 지수화하여 발표
  - – 동 지수값이 양수 값일 때 전반적인 완화기조를 의미하는 회사들이 절반 이상이라는 것을 의미함
- ㅇ 대출태도조사에서는 금융기관 대상 조사이기에 대출에 대한 공급자들의 관점이 조사되었음
  - – 대출태조조사에 부가조사로서 수요조사도 있으나 이는 실제 수요자가 아닌 공급자들의 의견을 반영한 조사임
  - – 즉 동일 대상자에게 현재의 대출 수요의 현황 및 전망을 질문하고 이에 대한 답변을 기반으로 대출수요지수를 산정함
  - – 하지만 이들 조사가 실제 대출 수요자이기 보다는 공급자 쪽의 추정에 가깝기 때문에 이를 수요측 태도조사로 보기에는 어려움이 있음

---

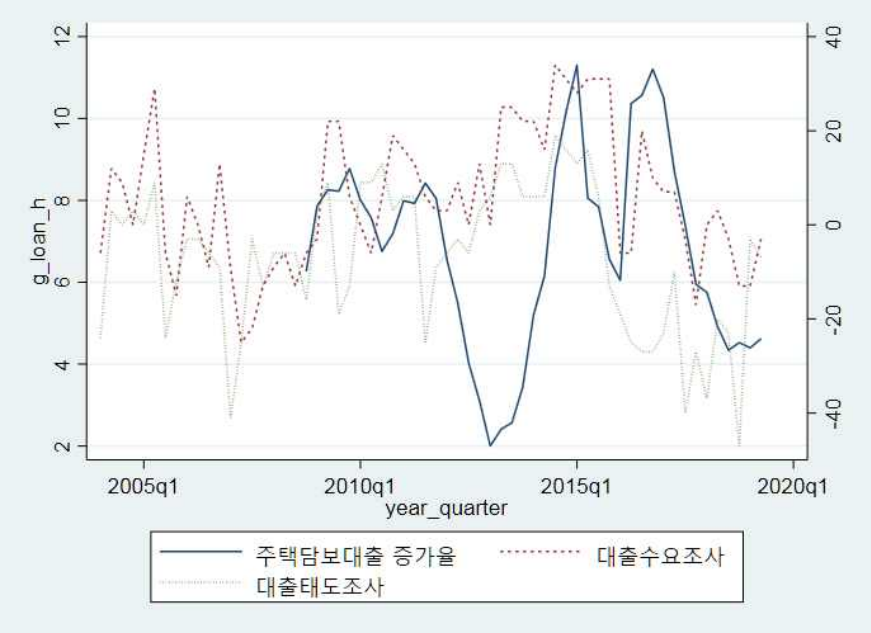
18) 보다 자세한 가계부채 누증요인 및 대응방향에 대한 분석은 한국은행(2017)을 참조
19) 대출태도조사는 한국은행 ECOS DB에서 추출하였으며 기간은 2002년 1분기부터 2019년 2분기까지임

❑ 아래 그림에서 보면 가계부문 대출태도조사에서의 수요조사 지수와 주택담보대출 증가가 어느 정도 관계를 갖는 것으로 나타남

○ 2009년 1분기부터 2019년 2분기 까지*의 두 변수 간 상관관계를 보면 공급측 요인이라고 할 수 있는 주택 대출태도지수와 가계 주택담보대출 증가율간의 관계는 -0.09로 낮게 나타난 반면, 수요조사는 0.22 정도로 나타나 설명력이 보다 높았음

* 한국은행 ECOS DB에서 대출태도조사는 2002년부터 1분기부터 사용가능하나 가계 주택담보대출 증가율은 2009년 1분기부터 이용 가능함

[그림 Ⅳ-12] 대출태도조사와 주택담보대출 증가율 관계



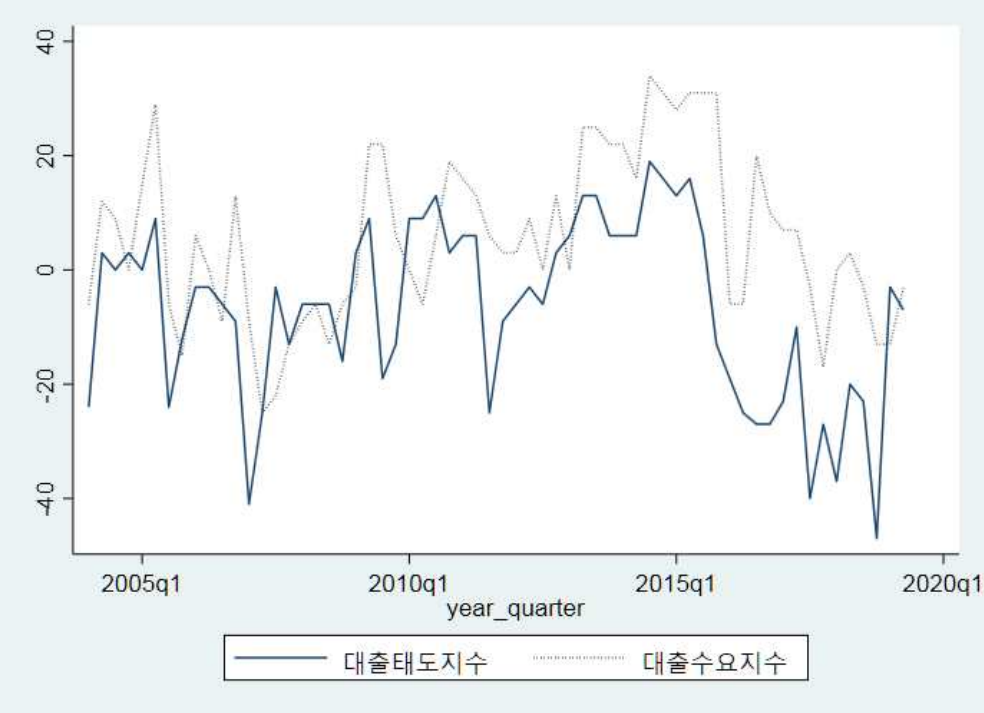
주: Y 축의 좌측은 주택담보대출 증가율, 우측은 대출수요조사 및 대출태도조사 지수
출처: 한국은행 ECOS DB

❑ 한편 아래 그림에서 보는 바와 같이 대출태도조사 결과와 수요조사 결과가 비슷한 패턴을 나타내고 있음

○ 동일인을 대상으로 현재와 미래의 금융회사의 태도와 시장에 대한 조사를 함께 수행하다 보니 동행하는 관계가 나타나 추가적인 정보 제공 가능성이 낮을 것으로 판단됨

○ 두 조사 지표의 상관관계가 54%가 넘는 것으로 나타남

[그림 IV-13] 주택대출 관련 대출태도 지수와 대출수요 관계 추이



출처: 한국은행 ECOS DB

❑ 본 절에서는 이와 같은 대출태도조사와 관련된 한계를 극복하기 위해 검색결과를 활용하여 대출수요 조사를 대체할 수 있는지를 살펴봄

- ㅇ 동 검색결과를 대출수요로 파악할 수 있다면 속보성도 있으면서 좀 더 수요측에 부합되는 정보를 파악할 수 있을 것임
- ㅇ 특히 본 절에서는 가계부채의 핵심이라고 할 수 있는 주택담보대출과 관련된 태도조사 결과와 검색기록을 활용한 수요조사 결과 그리고 실제 주택담보대출 증가율간의 관계를 분석함

< 참 고 >

금융기관 대출행태서베이 개요*

(2019년 2/4분기 동향 및 3/4분기 전망)

❑ 실시기간 : 2019.5.27.~6.14일

❑ 대상기관 : 총 199개 금융기관(국내은행 15개, 상호저축은행 16개, 신용카드회사 8개, 생명보험회사 10개 및 상호금융조합 150개)

❑ 대상자 : 대상기관의 여신업무 총괄담당 책임자

❑ 방 법 : 전자설문 조사(상호금융조합은 우편 조사) 및 인터뷰

❑ 내 용

○ 금융기관의 대출태도, 신용위험 및 대출수요에 대한 지난 3개월간(2019.4～6월) 동향 및 향후 3개월간(2019.7～9월) 전망을 조사

❑ 지수 산출

○ 대출태도, 신용위험 및 대출수요에 대한 지난 분기 동향 및 다음 분기 전망을 5개 응답항목*을 통해 조사한 후 가중평균**하여 지수를 산출

* ① 크게 완화[증가] ② 다소 완화[증가] ③ 변화없음 ④ 다소 강화[감소] ⑤ 크게 강화[감소]

** [('크게 완화(증가)' 응답 비중 × 1.0 + '다소 완화(증가)' 응답 비중 × 0.5) - ('크게 강화(감소)' 응답 비중 × 1.0 + '다소 강화(감소)' 응답 비중 × 0.5)] × 100

○ 지수는 100과 -100 사이에 분포하며 지수가 양(+)이면 「완화」라고 응답한 금융기관의 수가 「강화」라고 응답한 금융기관의 수보다 많음을, 음(-)이면 그 반대를 의미

❑ 지수 공표: 매 분기가 종료된 다음 달(1·4·7·10월) 초에 공표

* 출처: 한국은행, "금융기관 대출행태서베이 결과", 보도자료(2019.7.4.)

나. 방법론 및 분석결과

① 방법론

❑ 먼저 본 연구에서는 대출수요를 나타낼 수 있는 소셜자료로 가계 주택담보대출과 관련된 검색기록을 자료로 활용함

- ㅇ 가계 주택담보대출을 받기위해 개인들은 인터넷에서 검색을 하게 되는데 이 경우 수요로 파악할 수 있는 관련 연관 검색어는 '이자' 내지 '금리'라고 할 수 있음

❑ 한편 구글 트렌드에서 연관검색어를 살펴보면 가계 주택담보대출 수요와 관련하여 주요 검색어는 '금리'로 나타나고 있으며 '주택담보대출 금리' 검색어를 가계주택담보대출 수요의 대용변수로 정의함

- ㅇ 네이버 데이터랩의 경우 2016년 이후 자료가 입수가능하므로 본 연구에서는 2004년 이후 구글 트렌드에서의 검색기록을 활용함
- ㅇ 그러나 한국은행 ECOS DB에서 제공하는 가계 주택담보대출 자료가 2007년 4분기부터 이용가능하므로 실제 분석 기간은 2007년 4분기부터 2019년 2분기까지임

❑ 본 연구에서는 '주택담보대출 금리' 검색기록을 실질적인 가계 주택담보대출 수요로 파악하고 기존의 대출태도조사에서 나타난 대출수요와의 비교를 통해 그 유용성을 파악하고자 할 것임

- ㅇ 실질적으로 주택담보대출 증가율과 이들 변수들이 어떤 관계가 있으며 검색기록이 이와 같은 대출증가율에 추가적인 정보를 제공할 수 있는지를 살펴봄

❑ 이를 통해 먼저 상관관계분석을 수행하여 주요 변수들간의 관련정도를 파악하고 인과관계분석과 회귀분석을 수행하여 통계적인 유의성과 설명정도를 파악함

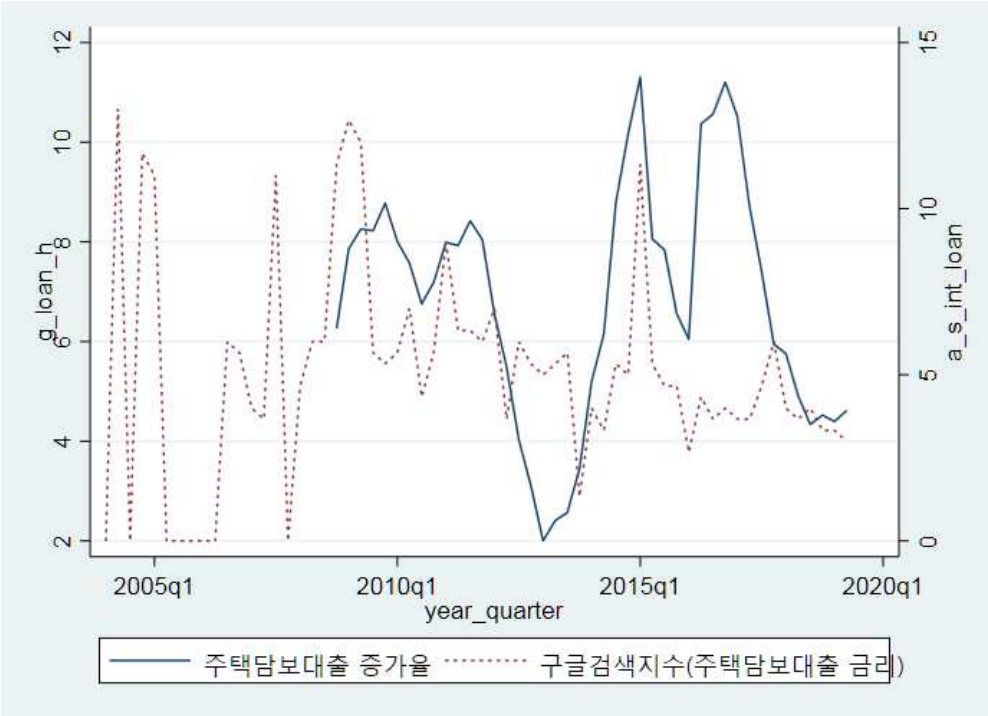
- ㅇ 대출태도조사와의 비교를 위해서 본 연구에서는 분기 자료를 활용하고 이후 월별 자료를 이용하여 추가 분석을 수행함

② 분석결과

❑ 대출증가율과 검색기록간의 상관관계를 살펴보면 먼저 일반적인 '주택담보대출' 검색기록과의 관계는 0.25로 '주택담보대출 금리' 검색기록과는 0.26으로 나타나고 있음

- ㅇ 구글 트렌드에서 제공하고 있는 대출검색 기록은 월자료이므로 대출태도조사와의 비교를 위해 3개월 평균 수치를 분기 검색지표로 활용함
- ㅇ 동 상관관계는 대출태도지수나 대출수요에 비해 높은 수치임

[그림 Ⅳ-14] 주택담보대출 증가율과 검색기록 간 관계 추이



주: Y 축의 좌측은 주택담보대출 증가율, 우측은 구글 검색지수
출처: 한국은행 ECOS DB, Google Trends

❑ 한편 시계열 변수간 관계 분석을 위한 회귀분석을 수행하기에 앞서 사전적으로 단위근 검정을 수행한 결과 가계대출 수준이나 증가율 변수 모두에 있어 단위근이 있는 것으로 나타남[20)]

- ㅇ 본 소절에서 분석자료는 분기 자료이기에 변수명에 하첨자 t는 현분기를 의미하고 t-1은 1분기 전을 의미함

20) 단위근 검정결과는 부록의 부표를 참고

○ 검색기록은 단위근이 나타나지 않음

❑ 아울러 이와 같은 검색기록이 가계대출 증가율과 시계열적인 인과관계가 있는지를 살펴보기 위해 그랜저 인과관계 분석을 수행함

○ 분석결과 대출금리 검색 기록은 가계대출 증가율을 그랜저 인과한다고 할 수 있는 반면 그 반대 관계는 나타나지 않는 것으로 나타남[21)]

❑ 회귀분석은 OLS, AR, 차분모형을 활용하여 수행함[22)]

○ 가계주택대출 증가율이 단위근이 있는 것으로 나타나고 있으므로 분석결과는 주로 차분모형을 중심으로 설명함

[표 IV-15] 대출수요지수와 대출수요 검색지수를 이용한 대출증가율 회귀분석 결과

1) 대출수요지수

	대출증가율			
	OLS	AR	차분	AR 차분
가계주택수요지수	0.0678**	0.0226		
	(0.0288)	(0.0155)		
가계주택대출태도지수	-0.0442*	-0.006		
	(0.0253)	(0.0119)		
가계주택대출증가율$_{(t-1)}$		0.8449***		
		(0.0818)		
△가계주택수요지수			0.0177*	0.0101
			(0.0103)	(0.0106)
△가계주택대출태도지수			-0.0060	-0.0132
			(0.0100)	(0.0091)
△가계주택대출증가율$_{(t-1)}$				0.2843**
				(0.1297)
상수항	5.938***	0.7748	-0.0391	-0.0693
	(0.4990)	(0.5925)	0.2016	(0.1974)
관측치 수	43	42	42	41
R-squared	0.1106	0.7495	0.0279	0.1118

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

21) 그랜저 인과관계 검정결과는 부록의 부표를 참고
22) 자료의 계절성을 감안하여 분기별 더미를 사용하여 분석하여도 동일한 결과를 얻음

2) 대출수요 검색지수

	대출증가율			
	OLS	AR	차분	AR 차분
주택대출금리검색지수	0.2997**	0.1149**		
	(0.1260)	(0.0544)		
가계주택대출태도수요지수	-0.0250			
	(0.0240)			
가계주택대출증가율$_{(t-1)}$		0.8431***		
		(0.0807)		
△주택대출금리검색지수			0.2184**	0.2125*
			(0.0962)	(0.1004)
△가계주택대출태도수요지수			-0.0095	-0.0174
			(0.0122)	(0.0109)
△가계주택대출증가율$_{(t-1)}$				0.3113**
				(0.1174)
상수항	5.0573***	0.4324	0.0062	-0.0190
	(0.8648)	(0.5901)	(0.1947)	(0.1855)
관측치 수	43	42	42	41
R-squared	0.0952	0.7493	0.1440	0.2429

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

❑ OLS 모형에서는 단순 상관관계에서와 같이 공급측 요인이라고 할 수 있는 태도지수에 비해 수요조사 결과가 유의미한 설명력을 가지고 있는 것으로 나타나고 있으나 차분했을 경우 유의도가 떨어짐

❑ 단위근이 있는 것으로 나타나 1차 차분 모형을 기반으로 보았을 때 수요조사 지표는 더 이상 통계적으로 유의미 하지 않게 나타남

❑ 검색지표의 경우 수요조사 결과와 유사한 결과가 OLS 및 AR모형에서 나타나고 있으나 단위근을 감안한 차분 모형에서도 수요조사 결과와 달리 5%내에서 유의미한 결과를 나타남

- ㅇ 추가적으로 전년도 종속변수를 설명변수로 사용한 최종 분석결과(AR 차분)에서도 여전히 5%내에서 유의하게 영향을 미치는 것으로 나타남

❑ 이와 같은 결과는 금융회사 임직원들 기반 조사자료인 수요지표에 비해 수요를 반영하는 검색결과가 가계주택대출 증가를 보다 잘 설명하는 것이라 해석할 수 있을 것임

- ㅇ 분기별로 작성되어 발표되는 대출태도조사에 비해 대출증가율을 보다 잘 설명하고 있는 것으로 나타나고 있으며 검색자료의 속보성을 감안할 때 장점이 있을 수 있음

❑ 다음에서는 추가적으로 월 검색자료를 이용하여 월별 대출증가율 분석을 수행함

- ㅇ 속보성 있는 구글 트랜드의 월별 대출수요 검색기록('주택담보대출 금리')을 활용하여 가계주택담보대출 전년 동기대비 증가율간의 관계를 살펴봄
- ㅇ 분석기간은 전년 동기 대비 가계 주택담보대출 증가율이 사용가능한 2008년 12월부터 2019년 6월까지임

[그림 Ⅳ-15] 월별 가계대출금리 검색과 대출증가율 관계 추이

주: Y 축의 좌측은 구글검색(가계대출금리) 지수, 우측은 주택담보대출 증가율
출처: 한국은행 ECOS DB

❑ 상관관계 분석을 수행하면 분기자료에 비해 상대적으로 가계주택담보대출 증가율과의 관계 정도가 떨어지는 것으로 나타남

- ㅇ 주택담보대출 수요의 단순 검색기록이라고 하는 '주택담보대출' 검색과 주택담보대출 증가율간의 관계는 0.04로 나타난 반면 '주택담보대출 금리' 검색과의 관계는 0.18로 나타남
- ㅇ 주요 경제변수들간의 관계로 보면 수요측 요인이라고 할 수 있는 주택(아파트)가격지수나 가계대출금리의 관계보다는 높게 나타나고 있는 반면 공급측 요인이라고 할 수 있는 가계대출 연체율이나 실업률의 관계보다는 낮게 나타나고 있음

❑ 월별 자료 역시 분기별 자료와 마찬가지로 단위근이 나타나고 있어 1차 차분모형을 중심으로 회귀분석결과를 설명[23)]

○ 아래 분석 에서 이용 분석자료는 월 자료이기에 변수명에 하첨자 t는 현 월을 의미하고 t-1은 1개월 전을 의미함

❑ 월별 검색기록과 증가율 간 그랜저 인과관계 검정을 수행한 결과 양 변수 간 관계는 나타나지 않은 것으로 판단됨

❑ 회귀분석 결과는 아래 표에 제시됨

○ OLS와 AR 결과는 주택담보대출 검색지수가 주택담보대출 증가율을 설명하는데 있어 통계적으로 유의하게 나타나고 있으며 수요와 공급측과 관련 있는 경제변수가 추가되어도 통계적 유의성을 어느정도 확보하고 있는 것으로 나타남

○ 반면 단위근을 감안한 차분모형에서는 검새지수의 통계적 유의도가 떨어지며 대출이자율을 제외하고 다른 경제변수들도 대체적으로 유의도가 떨어짐

○ 차분모형에서는 대출증가율을 설명하는데 있어 통계적으로 유의한 (-) 관계를 나타내는 변수는 대출이자율로 나타남

○ 이와 같은 결과는 가계대출증가율을 설명하는데 있어 수요측 요인이라고 할 수 있는 금리가 대출금리 검색기록에 비해 보다 직접적인 영향을 미쳤기 때문인 것으로 판단됨

[표 IV-16] 월별 검색자료를 활용한 가계주택담보대출 증가율 회귀분석 결과

1-1) OLS

	OLS		
	(1)	(2)	(3)
가계주택대출금리 검색지수	0.1268**	0.1030*	
	(0.0500)	(0.0582)	
아파트가격지수		-0.1990***	-0.2258***
		(0.0587)	(0.0608)
가계주택대출연체율		-8.8186***	-9.1334***
		(1.5412)	(1.4860)
가계대출금리		-0.0511	-0.0065
		(0.4613)	(0.4594)
실업률		-1.2435	-1.2596
		(0.9159)	(0.9002)
상수항	6.1770***	34.5685***	37.6534***
	(0.3668)	(8.7326)	(9.1306)
관측치 수	127	127	127
R-squared	0.0333	0.3087	0.2907

23) 월별 단위근 검정 결과와 그랜저 인과관계 분석 결과는 부록의 부표를 참조

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

1-2) AR

	AR		
	(1)	(2)	(3)
가계주택대출증가율$_{(t-1)}$	0.9634***	0.9410***	0.9338***
	(0.0288)	(0.0361)	(0.0369)
가계주택대출금리 검색지수			0.0292**
			(0.0138)
아파트가격지수		-0.0653***	-0.0590***
		(0.0157)	(0.0162)
가계주택대출연체율		-0.9375	-0.9696
		(0.6637)	(0.6681)
가계대출금리		-0.2411*	-0.2390*
		(0.1231)	(0.1227)
실업률		-0.4311	-0.4538
		(0.2796)	(0.2858)
상수항	0.2381	9.6225***	9.0257***
	(0.1816)	(2.7095)	(2.7318)
관측치 수	126	126	126
R-squared	0.9226	0.9280	0.9293

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

2-1) OLS 차분

	OLS 차분		
	(1)	(2)	(3)
△가계주택대출증가율$_{(t-1)}$	0.3498***		
	(0.0956)		
△가계주택대출금리 검색지수		0.0116	
		(0.0111)	
△아파트가격지수			0.2771
			(0.1989)
△가계주택대출연체율			-0.7264
			(0.8270)
△가계대출금리			-1.1514**
			(0.5712)
△실업률			0.1175
			(0.1720)
상수항	-0.0087	-0.0116	-0.0795
	(0.0586)	(0.0613)	(0.0559)
관측치 수	125	126	126
R-squared	0.1218	0.0054	0.0474

주: 1) ( ) 안은 표준오차를 의미함
2) 변수명 앞 △는 차분을 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

2-2) AR 차분

	AR 차분		
	(1)	(2)	(3)
△가계주택대출증가율$_{(t-1)}$	0.3489***	0.3360***	0.3386***
	(0.0956)	(0.0986)	(0.0993)
△가계주택대출금리 검색지수			0.0129
			(0.0100)
△아파트가격지수		0.1741	0.1800
		(0.2002)	(0.2000)
△가계주택대출연체율		-0.2279	-0.1509
		(0.7480)	(0.7426)
△가계대출금리		-1.7349**	-1.7429**
		(0.6817)	(0.6923)
△실업률		0.0363	0.0089
		(0.1767)	(0.1839)
상수항	-0.0087	-0.0658	-0.0648
	(0.0586)	(0.0551)	(0.0551)
관측치 수	125	125	125
R-squared	0.1218	0.1724	0.1790

주: 1) ( ) 안은 표준오차를 의미함
2) 변수명 앞 △는 차분을 의미함
3) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

다. 시사점

❑ 대출태도조사에서 파악되는 가계대출태도 및 가계대출수요 지표에 비해 가계대출금리 검색기록은 가계대출수요를 반영하는 지표로서 가계대출증가율을 보다 잘 설명하는 것으로 판단됨

- ㅇ 분기별로 발표되는 대출태도조사는 가계대출증가율을 제대로 반영하지 못하는 것으로 나타나고 오히려 대출검색지표가 상대적으로 설명력이 높은 것으로 나타남

❑ 월별자료 분석에 있어서는 대출수요에 직접적인 영향을 미치는 대출금리의 영향이 큰 것으로 나타나고 있으나 대체적으로 대출수요를 속보성 있게 파악할 수 있는 것으로는 이와 같은 검색기록이 유용할 수 있음을 확인

- ㅇ 이와 같은 결과는 프랑스의 사례를 연구한 Burdeau and Kintzler(2017)의 결과와 유사함
- ㅇ 동 연구에서는 본 연구에서 활용하기 어려웠던 Google Correlate 도구를 이용하고 Machine Learning 기법을 활용하여 신용증가에 대한 예측력을 증대시킬 수 있었음

❑ 향후 설명력과 예측력을 확보하기 위해 다양한 검색어의 모색과 머신러닝 기법 등을 활용 모형의 적합도를 개선시킬 필요가 있음

- ㅇ 아울러 검색기간의 확대와 검색엔진의 변화(네이버) 등 추가적인 연구를 통해 그 유용성을 확인해 볼 필요가 있는 것으로 보임

2. 텍스트마이닝 적용사례: 금융통화위원회 회의록 분석

가. 개요

❑ 텍스트마이닝이란 텍스트 자료에서 유용한 정보를 추출하는 방법

- ㅇ 데이터마이닝의 한 갈래로 비정형데이터(unstructured data)인 텍스트 자료에서 유용한 정보를 추출하는 방법임
  - 회의록, 뉴스기사, 댓글, 상품평, SNS(트윗, 게시글), 블로그 등의 자료를 활용
- ㅇ 텍스트를 구성하는 단어나 어구 등 정성적 정보를 수집하여 정량적인 지표로 변환하는 방법이 많이 활용
  - 주로 관련된 단어와 어구의 빈도와 상관관계로 표현

❑ 텍스트를 분석하는 이유는 사람들의 언어를 통해 감정과 의도 등의 새로운 정보를 수집할 수 있기 때문임

- ㅇ 텍스트분석은 사람들이 표현하는 언어를 대상으로 하는 분석이며 사람들은 언어(language)를 통해 사상(생각)과 감정을 표현함
- ㅇ 이러한 자료는 주로 숫자나 범주로 표현되는 정형데이터(structured data)가 전달하지 못하는 추가적인 정보를 제공할 수 있음
  - 긍정/부정 등의 감정
  - 정치성향
  - 미래의 결정에 대한 의도/태도(투자, 금리 등)

❑ 텍스트마이닝 분석은 정성적 지표인 언어를 정량적 지표로 전환하는 것으로 이를 위해서는 몇 가지 전제가 필요함

- ㅇ 언어게임(Wittgenstein, 1953)
  - “The meaning of words is best understood as their use within a given language game.”
  - 같은 단어라 하더라도 문맥에 따라 그 의미가 다르게 해석될 수 있음
  - 과거 일물일어(一物一語)의 법칙과 반대의 사상[24]
- ㅇ 양의 격률(maxim of quantity)과 관련성의 격률(maxim of relevance)
  - Paul Grice(1913-1988)가 제시한 개념으로 언어로 의사를 표현함에 있어서 인간이 가지는 합리성을 의미
  - 인간은 말을 하거나 글을 쓸 때, 항상 합리적으로 정보의 양을 선택하고 주제나 문맥과 관련된 내용을 선택
  - 언어를 분석할 때 전제하게 되는 중요한 가정
- ㅇ 통계적 의미론 가설(Turney and Pantel, 2010)
  - “사람들의 글, 말 등에서 드러나는 단어 사용의 통계적 규칙성으로부터 사람들이 말하고자 하는 바를 찾아낼 수 있다”는 전제
  - 자주 언급되는 단어는 의도와 의미가 존재

❑ 텍스트마이닝을 활용하기 위해서는 다음과 같은 과정을 거치게 됨

- ㅇ 전처리(pre-processing)과정
  - 토큰분리(tokenization): 형태소, 단어, 구절 등으로 분리
  - 불용어 제거: 의미가 없는 조사, 어미, 접속사 등을 제거
  - 품사지정: 문장 안에서의 역할에 따른 의미
- ㅇ 수치형 자료로 변환
  - 빈도분석: 특정 키워드의 중요도는 빈도에 비례
  - 벡터화: 키워드 간의 관계, 머신러닝을 통한 분류
- ㅇ 사전(dictionary) 제작 및 활용
  - 전처리 과정에 활용(형태소 분석 및 불용어 제거)
  - 감성이나 논조의 분석에 활용

24) 하나의 사물에는 하나의 언어가 대응된다는 고전적 관점을 의미

나. 빈도분석과 워드클라우드

❑ 2019년 금융통화위원회 의사록을 이용하여 공개적으로 발표된 수치 정보가 아닌 새로운 관점으로 금융통화정책을 파악해볼 수 있는 방안을 소개하고자 함

- ㅇ 금통위 의사록 분석을 통해 파악할 수 있는 한가지 유용한 정보는 금리인상과 인하에 대한 예측으로 소위 매파와 비둘기파의 의견을 구분하고 우세를 비교하는 것임(한국은행, 2017)
- ㅇ 분석 사례로 제13차 금융통화위원회(2019년 7월 18일)의 회의자료를 이용해보고자 하며 구체적으로는 다음과 같은 내용을 포함하고 있음
  - 2019년 하반기 경제전망에 대한 논의
  - 외환, 국제금융 및 금융시장 동향에 대한 논의
  - 통화정책방향에 대한 논의
  - 한국은행 기준금리 결정에 관한 위원별 의견
- ㅇ 우선 분석에 활용할 텍스트 자료를 준비함

가) 2019년 하반기 경제전망

일부 위원은 관련부서에서 경제 분석 및 전망시 활용하는 일부 거시계량모형이 재정정책이나 통화정책의 효과를 반영하는 데 다소 개선의 여지가 있었으나, 재정충격지수의 개선 등 그간의 다양한 연구 및 모형개발에 힘입어 앞으로는 거시경제정책의 효과를 보다 명확히 파악하는 것이 가능해 보인다고 언급하였음. 또한 모형에 기반한(model based) 경제분석이 더욱 정교해지는 동시에 잠재성장률 추정치의 정도를 높이는 효과가 기대된다고 덧붙였음.

이어서 동 위원은 전세계적으로 무역분쟁 등의 여파로 제조업 생산과 수출이 매우 저조한 반면 민간소비, 서비스산업, 고용 등은 상대적으로 양호한 모습을 나타내고 있다고 평가하였음. 이 같은 모습이 내수중심의 지속가능한 성장의 균형을 의미하는지, 아니면 민간소비 등도 시차를 두고 점차 둔화될 것인지 가늠하기 어렵다고 언급하면서 이에 대한 진지한 고민이 필요하다는 의견을 나타내었음.

또한 동 위원은 올해 경상수지 항목 중 서비스수지 개선이 금년 성장률에 보탬이 될 것이라고 평가하면서, 금번 경제전망의 대외커뮤니케이션에 만전을 기해줄 것을 당부하였음.

다른 일부 위원은 금번 전망에 대해 대체로 현실에 부합하는 수치가 제시되었다고 평가하였음. 다만 성장률 전망의 경우 하방 리스크가 여전히 더 커 보인다고 첨언하였음.

(하략)

❑ 전처리과정(pre-processing)

○ 텍스트자료로부터 단어를 구분

- 토큰화(tokenization): 주어진 텍스트에서 형태소를 분리 추출
- 품사태깅: 명사, 형용사, 동사 등 문장 내에서의 역할을 파악함
- R함수 예제

```
rm(list=ls())
library(KoNLP) # 한글자연어분석패키지(형태소분석포함)
useNIADic() #NIA사전로딩
text <- readLines(file.choose()) # 파일읽기
noun <- sapply(text, extractNoun, USE.NAMES=F) # 명사추출
```

- KoNLP 패키지를 인스톨하고 라이브러리를 로딩
- NIADic사전을 활용: 가장 많은 형태소를 구분 가능하지만 일반적인 용도임에 주의할 필요
- extractNoun() 함수는 주어진 텍스트에서 사전을 참고하여 명사를 추출하는 함수임

○ 분석과 관련이 없는 불용어를 삭제하고 길이가 과도하게 짧거나 긴 단어들은 제거함

- R함수 예제

```
#글자 수 2개이상
noun <- sapply(noun, function(x) {Filter(function(y) {nchar(y) > 1},x)} )

#불용어 삭제
noun <- rapply(noun, function(x) gsub("위원", "", x), how = "replace")
noun <- rapply(noun, function(x) gsub("관련", "", x), how = "replace")
```

- Filter(): 일정 조건에 맞는 데이터만을 필터링하여 통과시키는 함수임
- gsub(): 특정 문자를 찾아서 대체하는 함수임

❑ 빈도분석 결과

○ 단어를 빈도순으로 나열하고 이를 바그래프로 표현

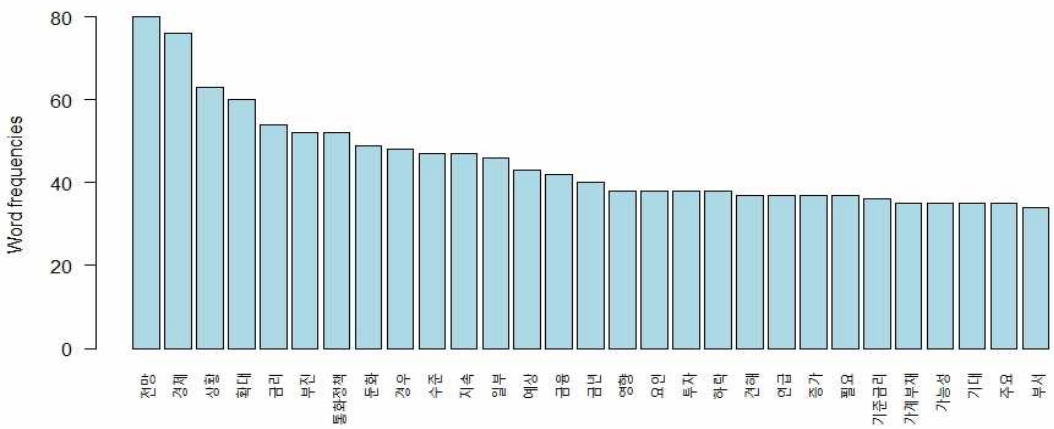
- R함수 예제

```
noun_unlist <- unlist(noun) # flatten
wordcount <- table(noun_unlist) # word count

top30 <- sort(wordcount, decreasing=T)[1:30] # 단어 정렬
top30 # 확인

top30 <- top30[-1] # 공백단어 제거

barplot(top30, las = 2, names.arg = names(top30),
        col ="lightblue", main ="Most frequent words",
        ylab = "Word frequencies")
```

```
> sort(wordcount, decreasing=T)[1:30]
noun_unlist
전망     경제     상황     확대     금리     부진 통화정책     둔화     경우
163       80       76       63       60       54       52       52       49       48
수준     지속     일부     예상     금융     금년     영향     요인     투자     하락
47       47       46       43       42       40       38       38       38       38
견해     언급     증가     필요 기준금리 가계부채     가능성     기대     주요     부서
37       37       37       37       36       35       35       35       35       34
```

- top30변수에 빈도수 상위 30개의 단어 리스트를 포함
- 이를 barplot()을 이용하여 그래프로 표현(그림 IV-16)
- 당연하게도 금통위의 회의에서 가장 자주 등장하는 단어는 ‘전망’, ‘경제’, ‘금리’ 등으로 나타남
- 경기부진과 둔화에 대한 언급이 많다는 것을 확인할 수 있음

[그림 IV-16] 빈도분석 결과 그래프



○ 워드클라우드로 표현

```
library(wordcloud) # 워드클라우드
library(RColorBrewer) # 글자색
palete <- brewer.pal(7,"Set2")

wordcloud(names(wordcount), # 단어들
          freq=wordcount, # 단어들의 빈도
          scale=c(3,0.5), # 단어의 폰트 크기
          min.freq=3, # 단어의 최소빈도
          random.order=F, # 단어의 출력위치
          rot.per=.1, # 90도회전 단어 비율
          colors=palete) # 단어색
```

- wordcloud()함수를 제공하는 라이브러리(wordcloud)와 글자 색상을 지정하는 라이브러리(RColorBrewer)가 필요함
- 워드클라우드 함수에 단어와 빈도 정보가 포함된 wordcount변수를 대입하고 기타 표현과 관련된 파라미터들(폰트, 회전, 색상 등)을 입력함

[그림 IV-17] 워드클라우드 결과 그래프

- ‘전망’과 ‘경제’가 가장 많이 언급되는 단어이며 그밖에 ‘부진’, ‘둔화’, ‘확대’ 등의 언급이 중요한 의미가 있을 것으로 보임
- 단순히 빈도만으로 나타낸 것이므로 그 이상의 시사점을 확인하긴 어려움

○ 형태소분석 함수 사례

- KoNLP에 포함된 함수를 이용하여 형태소 분석이 가능함
- extractNoun(): 주어진 텍스트에서 명사를 추출
- MorphAnalyzer(): 주어진 텍스트에서 형태소를 추출하고 모든 가능한 경우의 품사구분을 제시
- SimplePos22(): 주로 사용하는 22가지 기준으로 분류하는 함수[25]로, MorphAnalyzer()가 모든 경우의 수를 다 체크하기 때문에 오히려 실용적이지 못한 부분이 존재하기 때문
- 금통위 의사록에 포함된 문장을 활용하여 형태소 분석 함수 예제를 아래 그림에 제시
- '관련부서에서는 미중 무역협상 타결이 금년 성장률에 미치는 긍정적 효과는 시간이 갈수록 점점 작아질 것으로 보인다고 언급하였음'이란 문장을 분석
- 이 문장에서 구분해낼 수 있는 단어는 10개로 나타남: 관련부서, 미중, 부역협상, 타결, 금년, 성장률, 효과, 시간, 것, 언급
- 22가지 경우로 구분하는 SimplePos22()함수를 활용하면 문장에서 분류 가능한 형태소 구분과 문장에서의 역할 등을 보여줌

25) 분류기준은 KAIST 품사태크셋을 참고: https://github.com/haven-jeon/KoNLP/wiki/KoNLP-examples

```
sentence <- '관련부서에서는 미·중 무역협상 타결이 금년 성장률에
   미치는 긍정적 효과는 시간이 갈수록 점점 작아질 것으로 보인다고 언급하였음.'

extractNoun(sentence)
[1] "관련부서" "미·중"   "무역협상" "타결"     "금년"     "성장률"   "효과"     "시간"
[9] "것"       "언급"

MorphAnalyzer(sentence) #각 어절에서 모든 가능한 경우에 대한 분석

SimplePos22(sentence)

$관련부서에서는
[1] "관련부서/NC+에서/JC+는/JX"

$`미·중`
[1] "미·중/NC"

$무역협상
[1] "무역협상/NC"

$타결이
[1] "타결/NC+이/JC"

$금년
[1] "금년/NC"

$성장률에
[1] "성장률/NC+에/JC"

$미치는
[1] "미치/PV+는/ET"

$긍정적
[1] "긍정적/MM"

$효과는
[1] "효과/NC+는/JX"

$시간이
[1] "시간/NC+이/JC"

$갈수록
[1] "가/PX+ㄹ수록/EC"

$점점
[1] "점점/MA"

$작아질
[1] "작/PA+아/EC+지/PX+ㄹ/ET"

$것으로
[1] "것/NB+으로/JC"
```

다. 토픽모델의 활용

❑ 토픽모델은 주어진 문서에서 유사한 용도나 의미를 갖는 단어들의 그룹을 식별하여 분류하는 방법으로 일종의 머신러닝 기법임

- ㅇ 단어들간의 상관관계 및 유사성을 이용하여 통계적 방법 통해 단어의 그룹을 형성하고 분류하는 기법
- ㅇ 본 연구에서 활용할 방법은 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 기법임
  - 이 방법은 별도의 훈련(training)이 필요없는 비지도학습(unsupervised learning) 텍스트 분류 알고리즘임
  - 단어의 분류가 제대로 이루어졌는지 평가하기 위해서는 분포에 대한 모형이 필요한데 LDA기법에서는 디리클레 분포 함수를 따른다고 가정
  - 분포함수에 잘 맞는 분류인지 확인하는 과정은 베이지안 깁스 샘플링 추정(Marcov Chain Monte Carlo, MCMC)기법을 활용함
  - “Word”의 모임이 “Topic”을 구성하고, “Topic”의 모임이 “Document”를 구성한다는 것이 기본적인 개념임
- ㅇ 전처리과정
  - 필요 라이브러리는 아래와 같음

```
install.packages("rJava")
install.packages("KoNLP")
install.packages("stringr")
install.packages("lda")
install.packages("topicmodels")
```

  - 형태소를 분석해야 하므로 KoNLP라이브러리 설치 필요
  - LDA분석을 위한 lda라이브러리와 topicmodels라이브러리를 설치

```
library(lda)
library(topicmodels)
library(stringr)

# pre-processing:
corpus <- lexicalize(noun, lower=TRUE)
```

○ LDA함수의 적용

```
K<-2
G <- 1000
alpha <- 0.02
eta <- 0.02

set.seed(100)


result <- lda.collapsed.gibbs.sampler(corpus$documents, K, corpus$vocab,
                                      num.iterations = G,
                                      alpha = alpha,
                                      eta = eta,
                                      compute.log.likelihood = TRUE)
```

- K, G, alpha, eta가 LDA함수에 들어가는 파라미터임
- K는 분류되는 그룹의 수를 의미
- G는 계산시 반복횟수를 의미
- alpha와 eta는 분포함수(디리클레) 모형에 들어가는 파라미터

○ 분석결과([표 Ⅳ-17] 참조)

- 일반적으로 금통위의 회의에서는 금리에 대해 소위 매파와 비둘기파의 차이가 나타날 것을 기대(Lee et al., 2019)
- 두 그룹으로 분류한 결과 중, 첫 번째 토픽 그룹은 대체로 경제에 대한 (부정적)평가와 관련된 단어들로 보이고, 두 번째 그룹은 그 이외의 정책관련 단어들로 보임
- Hansen and McMahon(2016)가 FOMC의결문을 LDA기법으로 분류한 결과를 경제전망과 사전적 정책방향 제시의 두 가지로 해석했는데, 본 결과도 이와 비슷한 해석이 가능할 것으로 보임
- 다만 현재의 결과가 완벽하게 단어들이 구분되었다고 보기는 힘들고, 여러 가지 파라미터 조합과 불용어 제거를 통해 다양한 시도를 해볼 필요가 있음
- 그렇지만 지금의 분류는 파라미터가 주어진 상태에서 인간의 개입이 없이 컴퓨터가 스스로 텍스트의 구성을 통해 단어간의 상관관계를 파악하고 토픽을 분류한 결과임

[표 Ⅳ-17] 토픽모델 분석 결과

토픽 1						토픽 2					
전망	부진	둔화	지속	금년	투자	기준금리	가계부채	언급	금융시장	가격	
증가세	성장률	수출	가운데	예상	성장세	필요	일부	평가	유의	의견	충격
무역분쟁	미	중	교역	모습	상당	안정	정책	견해	공급	동향	당시
세계	불확실	소비	기업	하방	완화	금리	나라	건전	측면	일반	발생
중심	고용	상승률	제조업	정부	지난해	통화정책	하기	고려	금번	부동산	규제
감소	작용	내년	양호	올해	수요	분석	결과	시장	현상	완화적	불균형
압력	증가율	우려	위축	흐름	약화	견해	주택	장단기	강화	다양	생각
확대	반면	소득	일본의	감소세	예상	당부	부채	때문	중앙은행	상대	중요
소비자물가	완만	개선	민간소비	보호무역주의		이번	결정	거시	상황	제시	초래
달러	상반기	회복	하반기	재정정책	가계	이슈	문제	과거	연구	유동성	점검
부문	무역협상	심화	경로	심리	전환	대내외	만큼	외환	인플레이션		
생산	gdp	당초	하회	국의	작년						
반등	소폭	마이너스	실적	확산	물가상						
승률	수출규제										

라. 시사점

❑ 토픽모델은 주어진 문서에서 유사한 맥락을 갖는 단어들 간의 그룹을 구분하는 머신러닝 기법의 하나임

- ㅇ 파라미터 등 일정 조건이 주어진 상황에서 인간의 자의적 판단이 아닌 자료의 성격 즉, 텍스트 안의 단어 구성만을 이용해서 분류 결과를 제시한다는 장점이 있음
- ㅇ 물론 파라미터의 설정에 있어서는 인간의 판단이 개입될 필요가 있음
- ㅇ 따라서 신뢰성 있는 분류결과를 얻기 위해서는 다양한 파라미터 조합의 변화에 대해 안정적인(robust) 분류가 이루어지는지 확인할 필요가 있음

❑ 이와 같은 기법의 장점은 인간의 눈으로는 미처 판단하지 못한 잠재된 맥락과 뉘앙스를 발견할 수 있다는 점임

- ㅇ Lee et al.(2019)은 금통위 의사록에서 논조를 측정하여 논조지수를 측정하고 이를 기준금리 변동과 비교하면 상당히 높은 수준의 예측력을 갖는다고 보고하였음
- ㅇ 이처럼 실제 수치로 발표된 내용이 아닌 정성적 맥락, 즉 회의록에서 나타나는 뉘앙스를 통계적으로 분류하면 실제 경제현상을 파악하는데 유용한 정보들을 파악할 수 있음
- ㅇ 이는 사람이 직접 눈으로 읽고 판단하기보다는 통계적 기법을 통해 좀 더 객관적으로 측정할 수 있다는 점이 중요함

Ⅴ. 결론 및 시사점

❑ 4차 산업혁명 시대가 빠르게 도래함에 따라 빅데이터의 이용가능성은 보다 커지고 있으며 최근 빅데이터가 가지고 있는 장점을 활용하여 국내외에서 다양한 분야에서 이에 대한 활용이 이루어지고 있음

- ㅇ 해외에서의 활용에 비해 국내에서의 특히 정책적인 영역에서의 적용은 시작단계라 할 수 있으며 현재로서는 다양한 빅데이터의 유형을 파악하고 다양한 영역에서 적용하기 위한 시도들이 이루어지고 있음

❑ 본 연구에서는 빅데이터 분석의 활용 현황을 기존 연구자료를 원천별로 구분하여 살펴보고 현재 사용되고 있는 경제분석 방법론을 정리한 후 경제분야 빅데이터 분석 적용사례를 살펴보았음

- ㅇ 해외에서는 이미 빅데이터를 활용하여 유용성에 대한 다수의 연구결과가 발표되어 이를 적극적으로 정책수립이나 평가에 반영하고 있는 반면 우리나라는 빅데이터 관련 연구가 초기단계로 그 유용성을 검정하고 있는 상태로 파악됨
- ㅇ 본 연구에서는 이와 같은 연구사례 분석을 통해 경제분석 방법론을 정리하고 속보성과 예측이 중요한 경우와 사후평가 및 심층 분석이 필요한 경우로 대별하여 그에 따른 분석방법을 정리하였음
- ㅇ 다음으로 이와 같이 정리된 분석방법론을 기반으로 속보성과 예측의 중요성이 강조되는 경제이슈 중 가계부채 이슈를 검색자료 기반으로 분석하여 개인파산, 회생 등 채무조정 예측, 전략적인 파산의 추이, 그리고 대출수요에 대한 분석을 처음으로 시도하였음
  - 이와 같은 과정에서 검색자료의 추가적인 정보 제공가능성과 유용성을 확인
- ㅇ 아울러 사후평가 및 심층분석이 필요한 경우 사용될 수 있는 금융통화운영위원회 회의자료를 대상으로 한 텍스트마이닝 적용사례를 제공하여 이를 기반으로 향후 적용과 추후 개발이 가능하도록 하였음

❑ 최근 주요 정책 분석기관에서 빅데이터 분석에 대한 수요가 커지고 있는 만큼 본 연구를 기반으로 국회예산정책처에서 경제분석 업무에 실제 활용할 수 있는 방안 등을 마련할 수 있을 것임

- 다양한 유형의 자료를 이용, 실제 경제분석에 활용할 수 있는 기반을 마련하고 이를 통해 제도 개선 및 국회 의정활동 지원에 활용될 수 있는 정책 시사점 등을 도출하는데 활용될 수 있을 것임

❑ 본 장에서는 빅데이터를 활용하여 경제·사회 이슈들을 분석할 때 발생할 수 있는 빅데이터 활용 관련한 유의점 등을 정리함

- 데이터가 의미를 갖기 위해서는 이와 같은 데이터를 이용해서 무엇을 할지가 정해져야 하고 이것이 이와 같은 목적에 부합되는 자료인지를 면밀히 검토해야 함
- 또한 이와 같은 데이터를 통해서 제대로 분석할 수 있는 분석기법의 선정이 중요함
- 마지막으로 적절한 데이터와 적합한 분석기법을 통해서 얻은 결과에 대한 해석이 적절히 이루어질 필요가 있음

❑ 본문에서 언급한 바와 같이 무엇보다도 빅데이터가 만능은 아니며 먼저 데이터 확보를 하고 나중에 분석주제를 선정하려는 것 보다는 빅데이터의 장점과 특성을 파악하고 다음으로 기관에서 고민하는 이슈들을 명확히 정의할 필요가 있음

- 빅데이터를 이용해서 답하고자 하는 질문이 무엇인지 그리고 그 질문에 빅데이터 분석이 유용한지를 판단해야 함
- 빅데이터 분석의 핵심은 어떤 데이터를 이용하여 어떤 목적과 목표를 가지고 분석하는 가임

❑ 다음으로 어떤 유형의 빅데이터가 분석목적에 부합하는지 판단해야 함

- 다양한 유형의 빅데이터가 있으므로 분석목적이 속보성 있는 정책판단과 예측인지, 정책결과에 대한 심층평가인지 등에 따라 비정형화된 빅데이터를 사용할지 아니면 공공데이터 기반의 빅데이터를 상용할지가 결정되어야 할 것임

❑ 데이터를 입수하고 분석을 위해서는 전처리 단계[26]를 거쳐야 함

- 빅데이터는 노이즈가 심하기 때문에 이와 같은 전처리 과정이 필수적임

---

26) 빅데이터에서 데어터 전처리는 데이터를 특정 플랫폼 또는 시스템에 공급하기 위해 필요한 작업의 전체를 말함. 또한 데이터 마이닝 및 분석을 위해 결측치를 처리하고 데이터를 변환, 가공, 잡음제거, 손실 데이터 보정, 데이터 형변화하는 과정을 지칭하기도 함. 최근에는 비정형데이터를 정형화(파싱, 자연어 처리)하는 과정을 의미하기도 함. 본 연구에서는 폭넓게 사용하여 실제 연구자가 데이터를 입수해서 실제 분석하기 직전까지의 데이터 처리 작업을 의미함. 빅데이터 전처리 작업 등과 관련된 보다 자세한 사항은 박인근 외(2019)의 pp.172~238 참조

❑ 이와 같은 문제의식에 부합하는 적절하고 효율적인 빅데이터 분석기법의 적용이 필요함

- ㅇ 다양한 유형의 빅데이터가 있으므로 이에 부합되는 적절한 분석기법을 적용할 필요가 있음
- ㅇ 효과적인 분석을 위해서는 대상 데이터를 지속적으로 수정, 관찰해야 하며 다양한 분석기법을 적절히 활용하면서 반복적인 분석을 수행
- ㅇ 최근 들어 단순한 선형분석 이외에 머신러닝 기법을 활용하여 예측력을 높이고 있음 (Burdeau and Kintzler, 2017)

❑ 분석 후 결과의 타당성을 검토해야 함

- ㅇ 빅데이터의 경우 제대로된 문제제기와 경제분석이 수반되지 않을 경우 현실성이 결여된 정책 대안을 제시될 가능성이 있음
- ㅇ 특히 경제이슈 분석에 있어서 인과관계에 대한 고민 없이 데이터 마이닝에만 몰입할 경우 문제해결에 도움이 될 수 있는 정책방안 제시가 어려울 것임
- ㅇ 이와 같은 분석을 뒷받침해 주는 것이 합당한 경제이론과 모형이라고 할 수 있으므로 이러한 것은 사전적인 단계에서 충분히 숙고되어야 할 것임

❑ 이와 같은 프로세스를 염두에 두고 국회예산정책처에서도 향후 빅데이터 분석이 적극 이루어질 필요가 있음

- ㅇ 본 연구결과에서 보는 바가 빅데이터 활용이 경제정책 분야에 다양하고 유용하게 활용될 수 있으므로 향후 국회예산정책처에서도 빅데이터 활용을 적극 모색할 필요
- ㅇ 이미 많은 연구기관이나 정책기관에서 빅데이터 활용을 위해 내부적인 연구와 조직정비가 이루어지고 있는 실정임

❑ 경제문제 해결을 위한 빅데이터 분석은 3장에서 언급한 바와 같이 크게 속보성과 예측이 중요한 경우와 사후평가 및 심층분석이 필요한 경우로 나누어 볼 수 있으므로 담당 업무와 과제 성격에 따라 거시적인 분석과 빠른 대응이 필요한 경우 검색어 기반 분석과 소셜미디어 텍스트 분석이 이루어질 필요가 있을 것임

- ㅇ 4장에서 언급한 가계부채 이슈와 관련하여 검색어 기반 연구를 통해 주요 거시경제변수에 대한 속보성 있는 예측 등이 비교적 손쉽게 이루어질 수 있을 것임
- ㅇ 또한 텍스트 마이닝 기법을 활용하여 금융통화운영위원회 회의록을 분석하여 다양한 대상의 자료에 대한 분석이 가능함을 보였음

- ㅇ 추가적으로 거시정책 분석 내지 평가 목적에 있어서는 정도 높은 데이터의 확보가 필요하므로 이를 위해서는 거래자료 기반 민간 빅데이터와 행정데이터와 같은 공공 빅데이터의 활용이 함께 이루어질 필요가 있음
- ㅇ 국회예산정책처는 통계청, 한국은행 등 공공기관과 연계하여 소득DB 와 가계부채DB 구축 작업을 하고 이를 기반으로 다양한 경제분석 및 평가 작업이 이루어질 수 있을 것임
  - 초기 파일럿 사업을 통해 이와 같은 평가작업의 유용성을 확인한 후 본격적인 DB 구축 방안을 모색할 수 있을 것임
  - 현재 통계청에서는 국세청 등 공공기관으로부터 관련 신고자료를 표본으로 제공받아 통계에 일부 활용하고 있으므로 이들과의 자료 협조 등 협업을 통해 공공부문 빅데이터 활용이 가능할 수 있을 것임
  - 또한 한국은행의 경우 민간 CB사들과의 업무 협조를 통해 가계부채 관련 개인 신용 자료를 경제분석 등에 활용하고 있으므로 이에 대한 것도 참고할 필요가 있음

❑ 다음으로 정책에 대한 사후 평가나 심층분석이 필요할 경우 연구 이슈 등에 부합되는 빅데이터를 선정하고 적절한 분석기법을 적용하여 분석할 필요가 있음

- ㅇ 예를 들어 Ⅳ장에서 제시한 텍스트 마이닝 기법을 통해 그동안 분석하기 어려웠던 비정형화된 정성적인 정보들의 자료를 활용하여 경제정책에 대한 평가 내지 분석이 이루어질 수 있을 것임

❑ 향후 국회예산정책처의 빅데이터 이용 활성화와 관련하여 다음과 같은 제안이 이루어질 수 있을 것임

- ㅇ 데이터의 특수성과 유용성을 감안할 때 관련 자료 취득 및 활용에 있어 R&D 투자가 이루어질 필요가 있음
- ㅇ 다른 기관 활용사례에서 보듯이 단계적인 접근이 이루어질 필요가 있음

❑ 첫 번째 단계로는 연구(Research)가 필요함

- ㅇ 외부전문가와 공동 연구를 통해 빅데이터 활용과 관련된 다양한 주제를 발굴하고 이를 내부 부서와 공유함으로써 관련 빅데이터 활용을 보다 확산시킬 필요가 있음
- ㅇ 아울러 빅데이터 분석을 위한 내부 연구자들의 다양한 분석 애플리케이션 사용능력이 중요하므로 R과 파이썬 등과 같은 기본 프로그램 사용 능력 배양이 중요하게 되며 필요에

따라서는 전문기업에서 제공되는 유료도구들의 활용도 이루어질 필요가 있음

❑ 다음으로 발전(Development)단계가 요구됨

- ㅇ 1단계에서 수행한 외부 전문가 등과 1~2년간의 빅데이터 활용에 대한 기초연구를 기반으로 빅데이터를 활용한 핵심 연구주제를 발굴하여 자체 연구와 과제를 수행하도록 함
- ㅇ 두 트랙으로 진행될 수 있으며 먼저 정량적인 자료 기반 연구 강화를 위해서는 공공부문 빅데이터를 활용할 수 있도록 다양한 공공기관과의 협업 내지 업무 협조를 진행할 필요가 있음
- ㅇ 4차 산업혁명 시대가 도래함에 따라 중요한 한 축이라고 할 수 있는 빅데이터의 활용 가능성이 높아지고 이에 따라 빅데이터의 접근성이 확대되고 있으나 아직까지도 개인정보보호 이슈 등 법적 제한이 큰 상태이므로 이에 대한 해결 노력이 긴요함
  - 이를 해결하기 위해서는 국회 차원의 노력이 필요하고 아울러 빅데이터를 활용하고자 하는 통계청과 한국은행과의 협업을 통해 이들이 확보한 빅데이터를 우선 활용하고 추가적인 공공자료 확보를 추진할 필요가 있음
- ㅇ 또한 정성적인 자료 기반 연구는 비정형데이터를 활용하여 속보성 있는 미래 경제 변수 예측이나 심도 깊은 경제분석에 활용될 수 있으므로 내부 전문가 그룹의 전문성과 이를 기반으로 한 통찰력(insight)을 활용할 필요가 있음
  - 기존과 다른 비정형화된 데이터를 정제하고 분석을 위한 데이터로 변환시키기 위해서는 기술적인 관점보다는 자료와 정책이슈들을 연결시킬 수 있는 안목이 긴요함

참 고 문 헌

<문헌자료>

과학기술정보통신부·한국인터넷진흥원 (2019) “2018 인터넷이용실태조사”

김경근・염명배 (2017) “신용카드 빅데이터를 활용한 지역별 소비 유출입 특성 연구”, 경제연구, 35(4) : 129-154, 한국경제통상학회

김규철 (2017) “북한 주민의 경제적 후생 수준과 추세: 새로운 데이터를 통한 접근”, KDI 북한경제리뷰, 한국개발연구원

김성준 외 (2018) “가계부채DB의 이해와 활용”, 조사통계월보, pp.16~48, 한국은행

김수현・이영준・신진영・박기영 (2019) “경제 분석을 위한 텍스트 마이닝”, BOK경제연구, 제2019-18호, 한국은행

노형식 (2010) “대출태도지수와 대출실전간의 관계”, 주간금융브리프, 제19권27호, 금융연구원

박인근 외 (2019) “빅데이터 분석과 활용”, 제이펍

박종수・이금숙 (2015) “교통카드 빅데이터 기반의 서울 버스 교통망 시간거리 접근성 산출”, 한국경제지리학회지, 18(4) : 539-555

오선정 (2018) “아르바이트 노동의 개념과 특성”, 한국노동연구원 연구보고서

유경원 (2006) “우리나라 개인파산의 결정요인 분석과 시사점”, 경제분석, 제12권4호, 한국은행

유경원 (2015a) “개인 채무조정 과정에서 채무자와 채권자의 전략적 행동에 대한 분석,” 소비자문제연구, 제46권 제2호

유경원 (2015b) “미국의 대규모 주택담보대출 부실화에 대한 정책 대응과 시사점”, 캠코리뷰, 제6호, 한국자산관리공사

유경원 외 (2015) “가계부채 확대가 실물부문 리스크에 미치는 영향”, 한국경제의 분석, 제21권 제1호

이긍희・김용대・황희진 (2016) “빅데이터를 이용한 고용지표 개발”, 국민경제리뷰, 2016년 제1호, pp. 1~38. 한국은행

이영준・김수현・박기영 (2019) “Measuring Monetary Policy Surprises Using Text Mining: The Case of Korea(텍스트 마이닝으로 측정한 통화정책 서프라이즈)”, BOK경제연구, 2019-11

이창근・권정현・김지운・최동욱・김수한・김성지・정유경・홍주원 (2017) “베이비붐 세대 특성 분석을 위한 빅데이터 활용방안 연구”, 용역보고서, 한국개발연구원

이태리・송연호・황관석・박천규 (2018) “CoLTV 지표를 이용한 임대차주의 상환위험 분석”, 부동산연

구, 28(1) : 65-77
최경덕・강형구・주하연 (2016) “식품안전 관련 유해정보가 소비자들의 소비패턴에 영향을 미치는가? 일본 후쿠시마 원전사고를 중심으로”, 한국경제연구, 34(1), pp. 41-83
최동욱 (2017) “인터넷 포털의 경쟁과 뉴스 편향도의 선택”, 산업조직연구, 25(2), pp. 1-40
하나금융경영연구소 (2019) “서울시 직장인의 출퇴근 Trend 변화”, 일반산업 연구보고서
한국은행 (2017) “금융안정보고서”
한국은행 (2017) “빅데이터의 경제통계 활용 현황 및 시사점”
한국은행 (2019) “금융기관 대출행태서베이 결과”, 보도자료(2019.7.4.)
한국은행 (2019) “금융안정보고서”
한국은행 (2019) “2019년도 제13차 금융통화위원회(정기) 의사록”, 회의자료(2019.7.18.)
Armour, P. and A. Hung. (2017) “Drawing Down Retirement Wealth: Interactions between Social Security Wealth and Private Retirement Saving,” Santa Monica, CA: RAND Corporation
Barbara A. Butrica, Howard M. Iams, and Karen E. Smith. (2003) “The Changing Impact of Social Security on Retirement Income in the United States”, Social Security Bulletin, Volume 65, No. 3, 2003/2004, pp. 1–13
Barbara A. Butrica, Karen E. Smith, and Howard M. Iams. (2004) “It's All Relative: Understanding the Retirement Prospects of Baby-Boomers”, Boston College Center for Retirement Research No. 2003-21
BizSpring (2019) “Logger and Internet Trend”
Burdeau, E. and E. Kintzler (2017) “Accessing the use of Google Trends to predict credit developments”, ISI World Statistics Congress
Bull, N. and D. Timothy. (2006) “Long-Run Tax Rates and Long-Run Growth: Macroeconomic Effects of the Aging Baby Boomers and of the Changing Federal Tax System”, Annual Conference on Taxation and Minutes of the Annual Meeting of the National Tax Association, Vol. 99, 99th pp. 359-368
Cellini, S. Riegg, and T. Nicholas (2016) “Gainfully Employed? Assessing the Employment and Earnings of For-Profit College Students Using Administrative Data”, NBER Working Paper No. 22287
Chetty, R., John N. Friedman, and Jonah E. Rockoff (2011) “The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood”, NBER Working Paper No. 17699
Choi, Hyunyoung. and Hal R. Varian (2009) “Predicting the Present with Google Trends”, (April 2, 2009). Google Research Blog
——— (2012) “Predicting the Present with Google Trends.” Economic Record, Vol. 88, pp. 2–9
Congressional Budget Office (2000) “Personal Bankruptcy: A Literature Review”, U.S. Congress
Connelly, R. et al. (2016) “The role of administrative data in the big data revolution in social science

research", Social Science Research

Duncan D.F. et al (2010) "The Baby Boomer Effect: Changing Patterns of Substance Abuse Among Adults Ages 55 and Older", Journal of Aging & Social Policy, p.237-248

Dunleavy, P. (2016) "'Big data' and policy learning", Chapter 8, forthcoming in Gerry Stoker and Mark Evans (eds.), Methods that matter: Social Science and Evidence-Based Policymaking, The Policy Press, Bristol

Einav, L. and L. Jonathan (2014) "Economics in the age of big data", Science, Vol.346, Issue 6210, 1243089

Gentzkow, M. and J.M., Shapiro (2010) "What drives media slant? Evidence from US daily newspapers", Econometrica, Vol. 78, No. 1, 35–71

Gentzkow, M. et al. (2017) "Text As Data"

Glaeser, E.L. et al. (2018) "Big Data and Big Cities: The Promises and Limitations of Improved Measures of Urban Life." Economic Inquiry 56, no. 1

Haughwout, A. et al. (2019) "Trends in Household Debt and Credit", FRB NY Staff Report No.882

IMF (2012) "Dealing with household debt", Working paper

Kunn, S. (2015) "The challenges of linking survey and administrative data", IZA World of Labor 2015: 214

Lee, Y., Kim, S., and Park, K.Y. (2019) "Deciphering Monetary Policy Board Minutes with Text Mining: The Case of South Korea", working paper

Wittgenstein, L. (1953) "Philosophical Investigations", BASIL BLACKWELL

McLaren, N. and Shanbhogue, R. (2011) "Using Internet Search Data as Economic Indicators." Quarterly Bulletin 2011 Q2, Bank of England

Meyer, B., D. Wu, W. Moores, and C. Medalia (2019) "The Use and Misuse of Income Data and Extreme Poverty in the United States", NBER Working Paper No.25907

Cohen, P., et al. (2016) "Using Big Data to Estimate Consumer Surplus: The Case of Uber", NBER Working Paper No. 22627

Turney, P.D., and Pantel, P. (2010) "From Frequency to Meaning: Vector Space Models of Semantics", Journal of Artificial Intelligence Research 37, 141-188

Baker, S.R., Bloom, N., and Davis, S.J., (2013) "Measuring Economic Policy Uncertainty", Chicago Booth Research Paper No. 13-02

Stephens-Davidowitz, S. (2017) "Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are", Foreword by Steven Pinker

Hansen, S. and McMahon, M. (2016) "Shocking language: Understanding the macroeconomic effects of

central bank communication", Journal of International Economics, 2016, vol. 99, issue S1, S114-S133
UNECE (2013) "Classification of Types of Big Data", UNECE Statistics Wikis
Mayer-Schönberger, V. and Cukier, K. (2013) "Big Data: A Revolution That Will Transform How We Live, Work and Think", John Murray

<웹사이트>

한국은행 경제통계시스템 (https://ecos.bok.or.kr/)
Daum 소셜메트릭스 (https://www.socialmetrics.co.kr/)
Federal Reserve Bank of New York (https://www.newyorkfed.org/research/policy/nowcast)
Google Adwords (https://ads.google.com/intl/ko_KR/home/)
Google Correlate (https://www.google.com/trends/correlate)
Google Ngram (https://books.google.com/ngrams)
Google Trends (https://trends.google.com/trends/?geo=KR)
KAIST 품사태크셋 (https://github.com/haven-jeon/KoNLP/wiki/KoNLP-examples)
NAVER 데이터랩 (https://datalab.naver.com/)

<부록 1> 국내 연구기관 빅데이터 관련 연구 현황

❑ 현재 국내 주요 연구기관, 기업, 대학 등에 빅데이터 관련 부서 또는 센터가 별도로 설립되어 연구보고서, 세미나 등 다양한 형태로 연구가 진행되고 있음

- ㅇ 한국정보화진흥원(2018)[27]에 따르면 현재 약 130여개 기관, 기업 등에서 빅데이터센터를 구축 및 운영 중인 것으로 파악
  - – 공공(지자체 등) 37개, 민간(산업계) 74개, 대학 21개의 빅데이터 센터가 있음
- ㅇ 또한 한국정보화진흥원(2019)[28]에 따르면 데이터 기반 비즈니스 기업 중 빅데이터 비즈니스를 영위하는 기업은 약 2만여 개에 이르는 것으로 파악
- ㅇ 서울대학교 '도시 데이터사이언스 연구소', KAIST '빅데이터 및 비즈니스 애널리틱스 경영 연구센터', 연세대학교 'YBIGTA', 고려대학교 'KUBIG' 등 국내 주요 대학에서 빅데이터 관련 세미나, 교육, 연구 등이 활발히 이루어지고 있음
  - – 특히 서울대학교 '도시 데이터사이언스 연구소'는 서울시와 MOU를 체결하여 빅데이터를 기반으로 서울시의 다양한 도시 이슈에 대한 해결 방안을 도출하고, 서울시민이 무료로 참여할 수 있는 교육 아카데미도 운영하고 있음

[표 1] 국내 빅데이터센터 현황

분류	구분	센터 수(개)	분류	구분	센터 수(개)
공공 (지자체 등)	IT	4	민간 (산업계)	IT	12
	교통	2		교육	1
	국방	1		교통	2
	금융	2		금융	20
	기상	1		미디어	5
	농업	2		식품	1
	에너지	3		연구	5
	유통	1		유통	6
	의료	6		의료	11
	제조	1		제조	6
	지리	1		종교	1
	지자체	7		통신	3
	행정	6		환경	1
	총	37		총	74
대학	총	21	합계: 공공 + 민간 + 대학 = 132개		

자료: 한국정보화진흥원

27) 한국정보화진흥원 빅데이터센터(2018.3.20.), "빅데이터 네트워크 협의체 참여를 위한 빅데이터 전문센터 신청접수 및 선정 계획", 보도자료
28) 한국정보화진흥원(2019.4), "2018 빅데이터 시장현황 보고서"

[표 2] 국내 주요 공공(지자체 등) 기관 빅데이터센터 및 업무

분야	기관명	업무 및 프로젝트
지자체	서울특별시 빅데이터캠퍼스	- 창업 지원, 빅데이터 분석 인프라 제공, 정책 발굴
지자체	경기도 빅파이센터	- 공공분야 등 데이터 제공, 창업 지원, 중소기업 컨설팅
지자체	전북 국토정보공사(LX) 빅데이터활용센터	- 지자체 맞춤형 사회현안 해결 빅데이터 분석 - 공간정보 분야 스타트업 업무협력 및 지원(데이터 킹 외 6개 기업)
지자체	창원시 빅데이터 TF	- 빅데이터 활용 과학적 분석행정 추진 - 소상공인 창업입지 분석 시스템 등 빅데이터 서비스 제공 - 부서별 빅데이터 선도인력 지정 및 빅데이터 교육
지자체	대구경북연구원 공간빅데이터센터	- 빅데이터에 GIS를 접목시켜 정책의사결정 지원
행정	행정자치부 정부통합전산센터 빅데이터분석과	- 빅데이터 수시 분석 지원 - 빅데이터 공통기반 플랫폼 구축 및 운영 - 데이터 지도 및 공통 데이터풀 구축 및 운영 - 빅데이터 분석 신기술 동향 분석
행정	특허청 특허분석센터	- 특허정보 DB구축을 통해 각 산업별·기술별 트렌드 분석, 특허 분석 기반 미래 R&D전략 수립, 유망기술·기업 발굴
금융	한국은행 빅데이터통계연구반	- 빅데이터 기반 경제현상 파악 및 경제통계 편제 - 빅데이터 처리, 분석, 활용관련 연구 수행
금융	코스콤 빅데이터팀	- 신규 콘텐츠 발굴, 공모전 개최, 데이터분석 통계 교육 실시
유통	한국소비자원 빅데이터분석팀	- 소비자 데이터 기반 유용한 지표 산출 및 맞춤형 정보제공 - 빅데이터 플랫폼을 통해 조기에 소비문제 감지 및 예방 - 기업과의 협업을 통해 빅데이터 분석 플랫폼 및 시범서비스 구축(아시아나IDT)
행정	통계청 통계빅데이터센터	- 조사 및 행정자료 통계 제공 - 공공데이터 간, 공공과 민간데이터의 연계 서비스 제공
행정	국세청 빅데이터센터	- 과세자료 수집 및 분석을 통해 세금 안내자료 제공
행정	관세청 빅데이터 분석센터 (설립예정)	- 빅데이터·AI 기반의 X-ray 판독 시스템으로 관세행정 관련 불법 행위 선별, 불법 외환거래 분석, 맞춤형 통계 제공
IT	한국인터넷진흥원 사이버위협 빅데이터분석센터 (설립 예정)	- 빅데이터 기반의 사이버위협 대응 역량 강화, 보안 연구 및 보안 신기술 개발

자료: 한국정보화진흥원

❑ 국내 빅데이터 관련 학회의 경우 ‘한국데이터정보과학회’, ‘한국빅데이터서비스학회, ‘한국빅데이터학회’ 등이 있음

- ㅇ 한국데이터정보과학회(The Korean Data & Information Science Society)는 1990년부터 ‘한국데이터정보과학회지’를 연 6회 출간하고 있음
  - – 주로 통계학 관련 내용이 대부분이나, 최근 빅데이터 관련 연구도 증가하고 있음
- ㅇ 한국빅데이터서비스학회(The Korea Big Data Service Society)는 2014년부터 ‘빅데이터와 안전사회 구현’ 학술지를 연 1회 출간하고 있음
- ㅇ (사)한국빅데이터학회(KOREA BIGDATA SOCIETY)는 2016년부터 ‘한국빅데이터학회 학회지’를 연 2회 출간하고 있음
  - – 한국빅데이터의 경우 매일경제신문, MBN, 한국데이터산업협회와 2019년부터 ‘매경 빅데이터 & 인공지능 포럼’을 개최하기도 함

❑ 과학기술정보통신부, 한국데이터산업진흥원, 한국정보화진흥원 등에서는 빅데이터 관련 시장 및 산업 현황에 대해 정기적으로 연구보고서를 발간하고 있음

- ㅇ 과학기술정보통신부와 한국데이터산업진흥원이 공동으로 발간하는 ‘데이터산업 현황 조사’ 보고서는 2015년부터 현재까지 매년 진행되어 옴
- ㅇ 한국정보화진흥원 K-ICT 빅데이터센터에서 발간하는 ‘빅데이터 시장현황조사’ 보고서는 2015년부터 현재까지 매년 진행되어 옴

<부록 2> 빅데이터 분석도구 소개

❑ 현재 사용할 수 있는 국내외 빅데이터 분석수단은 대표적으로 아래와 같음

- ㅇ 국내의 경우 네이버(Naver)에서 제공하는 '데이터랩(DataLab.)'과 다음소프트(DaumSoft)에서 제공하는 '소셜메트릭스(Social Metrics)' 등이 있음
- ㅇ 해외의 경우 Google에서 제공하는 Google Trends, Google Adward, Google Correlate, Google Ngram 등이 있음

❑ 본 부록에서는 해당 빅데이터 분석수단을 소개하고, '가계부채'라는 키워드를 사례로 각 Tool-kit을 어떻게 활용할 수 있는지 소개하고자 함

- ㅇ 키워드 '가계부채'를 검색하는 방법으로 수행(textmining이라고도 할 수 있음)
- ㅇ 각 분석수단 별로 분석 가능 범위 및 특징을 비교할 수 있음
- ㅇ 부가적인 활용 방법 소개

1. 네이버(Naver) - 데이터랩(DataLab.)

❑ (소개) 네이버가 제공하는 데이터랩(DataLab.)은 창업을 계획하거나, 이미 창업한 소상공인을 위해 비즈니스에 도움을 주기 위해 2016년 1월에 오픈된 데이터 서비스임

- ㅇ 크게 7개의 서비스를 제공
  - DataLab Home: DataLab 홈페이지 상에 가장 먼저 보이는 '홈(Home) 화면'으로 쇼핑 인사이트, 검색어 트렌드, 지역통계 세 가지 메뉴를 확인해볼 수 있음
  - 급상승검색어: 검색 횟수가 급상승한 검색어의 순위 및 추이를 연령별, 시간대별로 나타낸 메뉴
  - 검색어트렌드: 원하는 키워드를 검색할 경우 해당 검색어가 2016년부터 현재까지 얼마나 검색되었는지 확인하는 메뉴
  - 쇼핑인사이트: 쇼핑과 관련된 검색어의 클릭량 추이와 통계를 확인할 수 있는 메뉴
  - 지역통계: 시, 군, 구 단위로 특정 지역을 선택하면 그 지역의 관심 업종 순위와

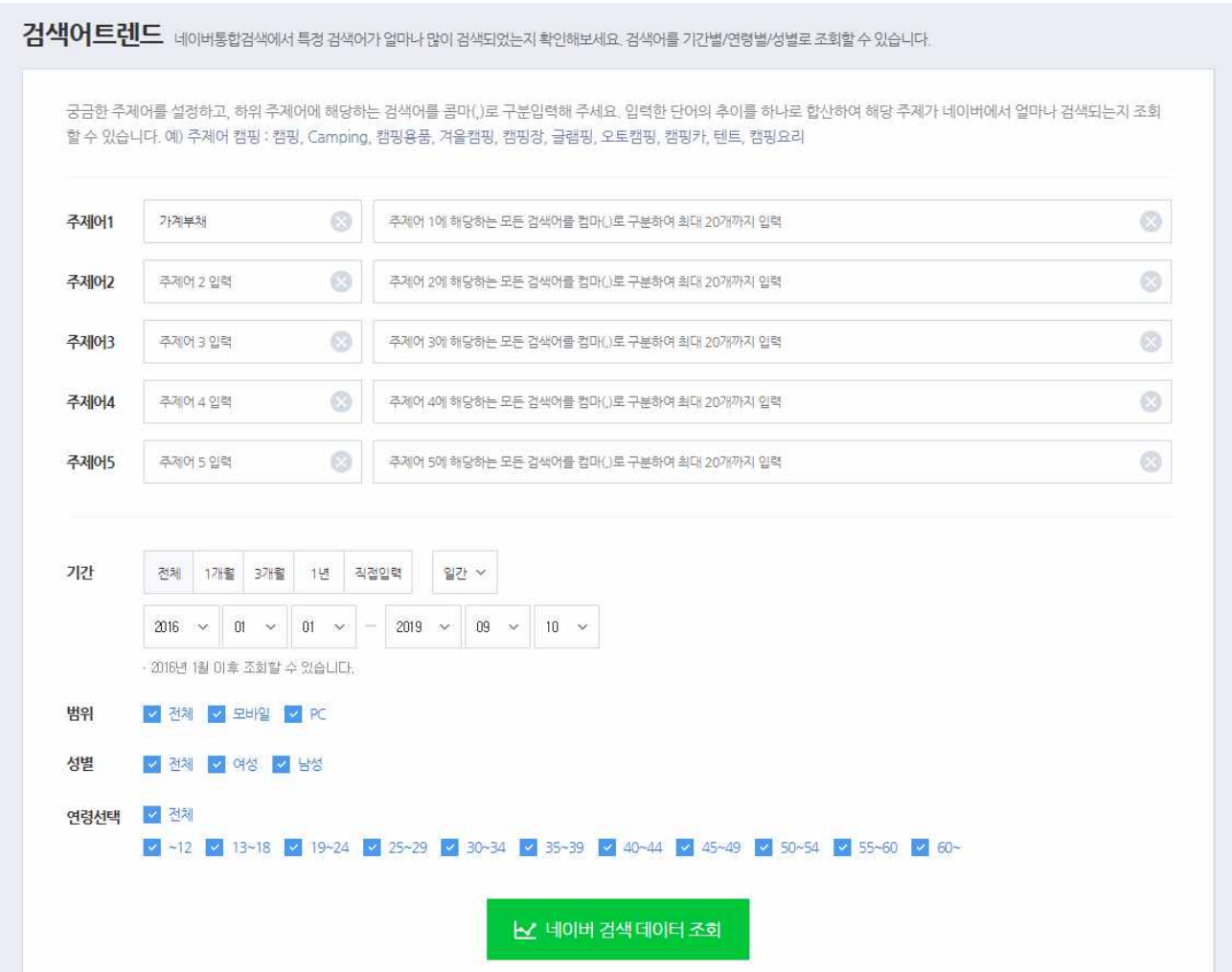
인기 지역을 확인할 수 있는 메뉴

- 댓글통계: 뉴스 서비스에서 작성된 댓글에 대해 작성자 수, 섹션별(정치/경제/사회 등), 성별, 연령별, 기기별(모바일/PC), 국가별로 확인할 수 있는 메뉴
- 공공데이터: 국내의 다양한 공공데이터를 확인 및 접속할 수 있는 메뉴

❑ (사례) 데이터랩(DataLab)의 '검색어트렌드' 화면에서 '가계부채' 키워드 검색

- ○ 주제어에 '가계부채' 입력
- ○ 기간은 2016년 1월 1일부터 현재 날짜인 2019년 9월 11일까지 선택
- ○ 범위는 '전체' 선택(모바일 및 PC 모두 해당)
- ○ 성별은 '전체' 선택(여성 및 남성 모두 해당)
- ○ 연령은 '전체' 선택

[그림 1] DataLab 검색어트렌드 화면 – 조회

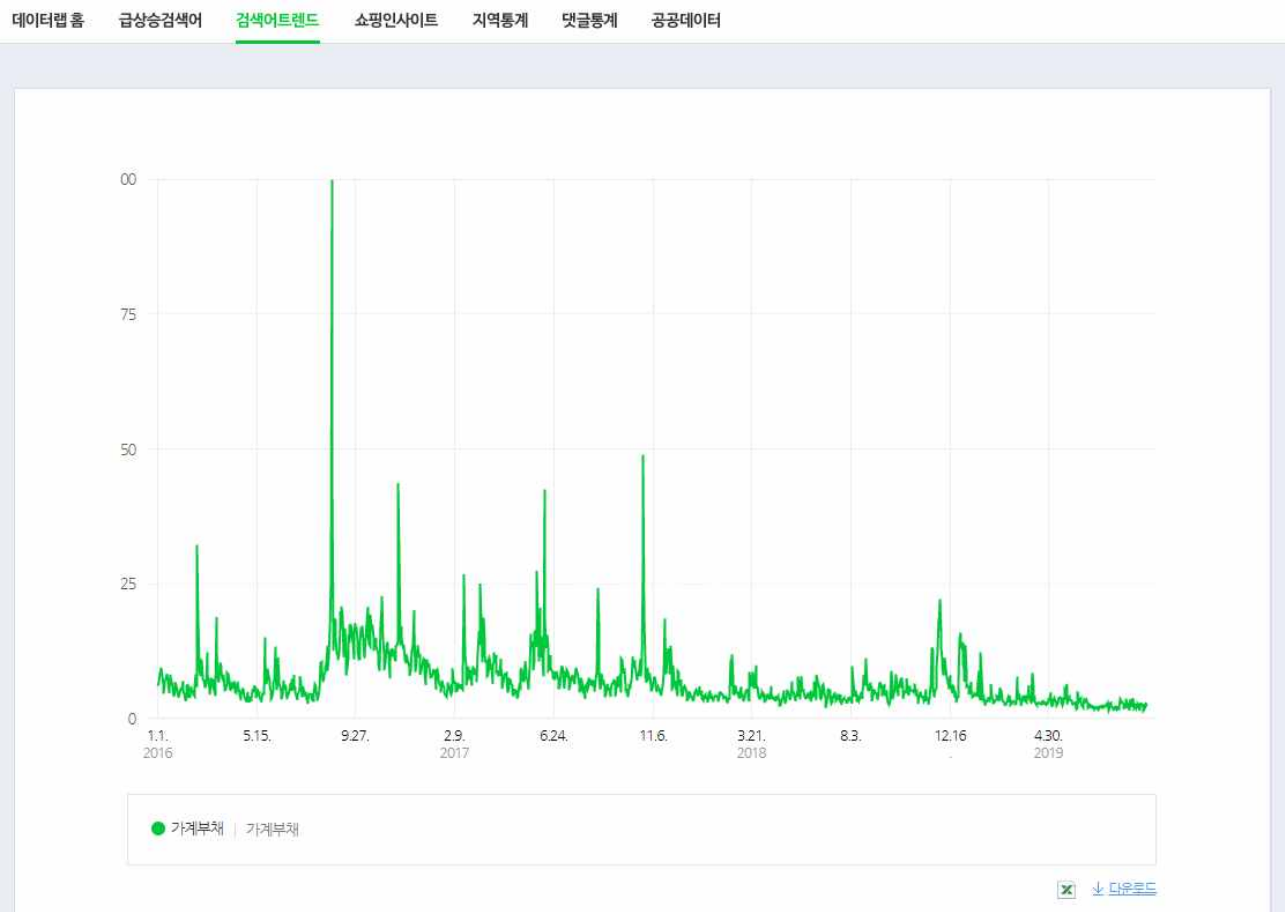


출처: 네이버 데이터랩(DataLab)

❑ 검색 결과 아래와 같이 그래프가 출력 되고, 해당 그래프에 대한 데이터를 엑셀 형태로 다운받을 수 있음

- ○ 그래프의 경우 네이버에서 해당 검색어가 검색된 횟수를 일별·주별·월별로 각각 합산하여 조회기간 내 최다 검색량을 100으로 설정하여 상대적인 변화를 나타냄
- ○ 해당 기간에서 '가계부채'가 가장 많이 검색된 일자는 2016년 8월 25일이며, 가장 적게 검색된 일자는 2019년 8월 31일에 해당
- ○ 대체적으로 2016년 하반기부터 2017년 말까지 많이 검색된 것으로 확인

[그림 2] DataLab 검색어트렌드 화면 – 결과



출처: 네이버 데이터랩(DataLab)

2. 다음소프트(DaumSoft) - Social Metrics

❑ (소개) 다음소프트에서 제공하는 Social Metrics는 소셜(Social) 여론분석, 심리, 인물 등 다양한 정보를 제공하는 데이터 서비스로, 무료로 이용할 수 있는 'Sometrend(썸트렌드)'와 유료로 이용할 수 있는 'Social Metrics Biz'로 구분됨

- ㅇ Sometrend는 실시간으로 이슈가 되고 있는 키워드를 확인할 수 있으며, 크게 4개의 서비스를 제공
  - Sometrend Home: DataLab의 'DataLab Home'과 유사하게 홈페이지 상에 가장 먼저 보이는 '홈(Home) 화면'으로 이슈분석, 빅데이터 분석, 인사이트 리포트 세 가지 메뉴를 확인해볼 수 있음
  - 이슈분석: 온라인에서 화제가 되고 있는 실시간 트렌드에 대해 키워드, 이슈 뉴스, 해시태그 등을 확인하는 메뉴
  - 빅데이터분석: 원하는 검색키워드의 언급량 추이, 연관어, 감성 분석 등을 확인하는 메뉴
  - 인사이트리포트: 특정 주제, 이슈에 대한 여러 소셜 미디어 분석 내용을 확인할 수 있는 메뉴
- ㅇ Social Metrics Biz는 키워드 검색 데이터를 보다 심층적으로 확인할 수 있는 서비스이며, 크게 3가지 서비스를 제공
  - 소셜 모니터링: 관심 키워드에 대해 실시간 모니터링(키워드 랭킹, 시간별 추이, 이슈 히스토리 등)과 뉴스 기사를 확인하는 메뉴
  - 소셜 인사이트: 제한 없는 키워드 검색으로 버즈 추이, 연관어 및 기간별 연관어, 감성 분석, 비교 분석을 할 수 있는 메뉴
  - 소셜 랭킹: 방송, 인물, 쇼핑 등 카테고리별 일일 이슈 키워드 랭킹을 확인하는 메뉴

❑ (사례) Sometrend의 메인화면에서 '가계부채' 키워드 검색

- ㅇ 주제어에 '가계부채' 입력
- ㅇ 기간은 자동적으로 현재 날짜를 기준으로 한 달 전~현재까지를 대상으로 함
- ㅇ 특정 기간을 검색하는 것은 불가능하며, 유료 서비스인 Social Metrics Biz에서 이용 가능

❑ 검색 결과 아래와 같이 연관어 맵, 연관어 언급량, 연관어 언급량 추이, 감성 분석, 감성 분석 상세, SNS 원문 등의 자료가 출력됨

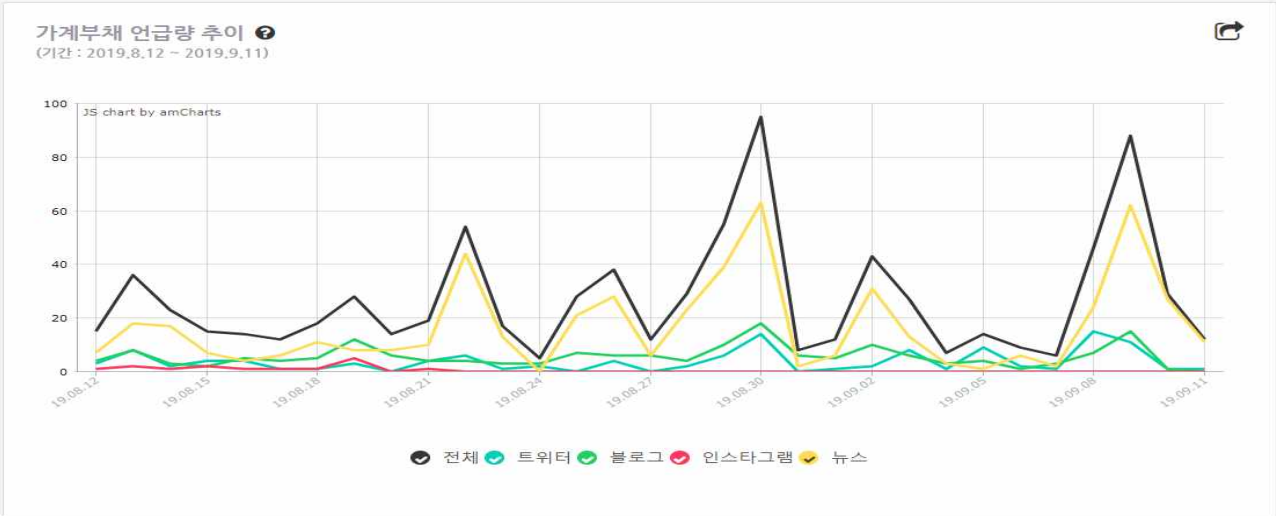
- ㅇ 연관어 맵은 가지의 카테고리(인물, 단체, 장소, 상품 등)를 색깔로 구분하고 관련 키워드 15가지를 나타냄
  - – 본 결과에서는 인물로는 이명박, 박근혜, 단체에서는 정부, 자유당, 상품에서는 부채, 시사/경제에서는 가계, 금리, 부동산, 이자 등이 연관어로 출력
- ㅇ 연관어 언급량은 해당 15개의 키워드가 SNS 상에서 얼마나 많이 언급되었는지 그 순위를 나타냄
  - – 본 결과에서는 '부채'가 가장 많이 언급된 연관어로 출력

[그림 3] Sometrend 화면 – 결과(1)

출처: Sometrend

- ㅇ 연관어 언급량 추이는 해당 키워드의 SNS별 언급량 추이를 그래프 형태로 나타냄
  - – 본 결과에서는 '가계부채'의 검색이 뉴스, 블로그, 트위터, 인스타그램 순으로 많이 언급된 것으로 출력

[그림 4] Sometrend 화면 – 결과(2)

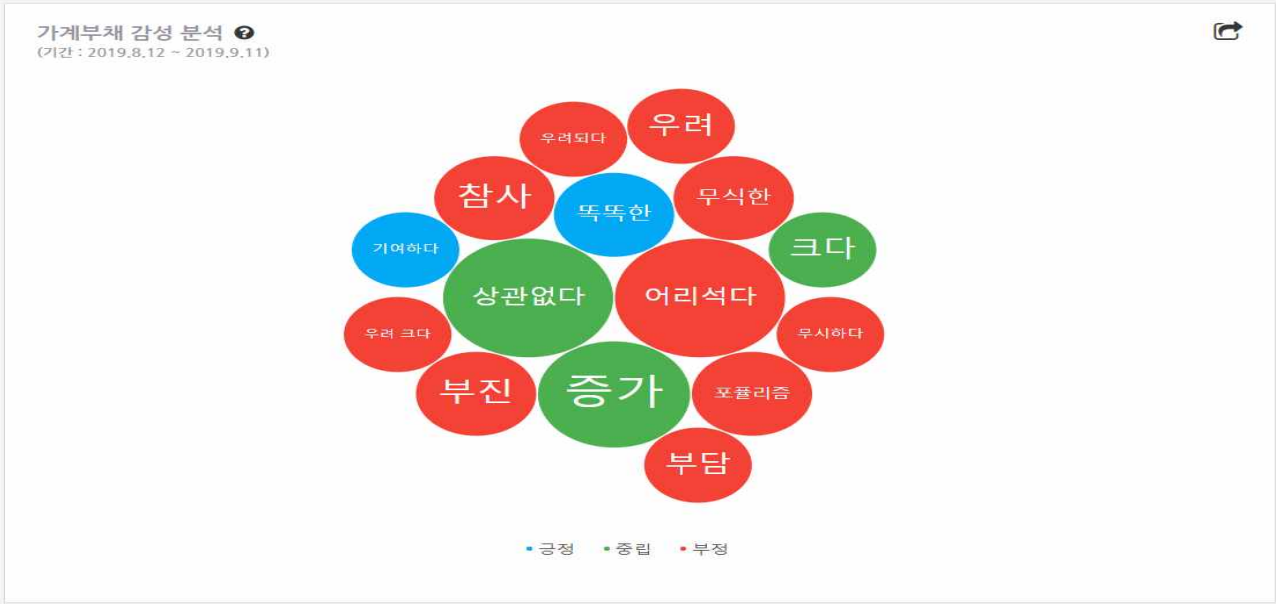


출처: Sometrend

○ 감성 분석은 검색 키워드에 대한 긍정·중립·부정에 대한 결과를 나타냄(파란색이 긍정, 녹색이 중립, 빨간색이 부정을 의미)

– 본 결과에서는 '가계부채'에 대한 감성어 결과가 긍정적 언어로는 기여하다, 똑똑한, 중립적 언어로는 상관없다, 증가, 크다, 부정적 언어로는 어리석다, 부담, 부진 등이 출력

[그림 5] Sometrend 화면 – 결과(3)

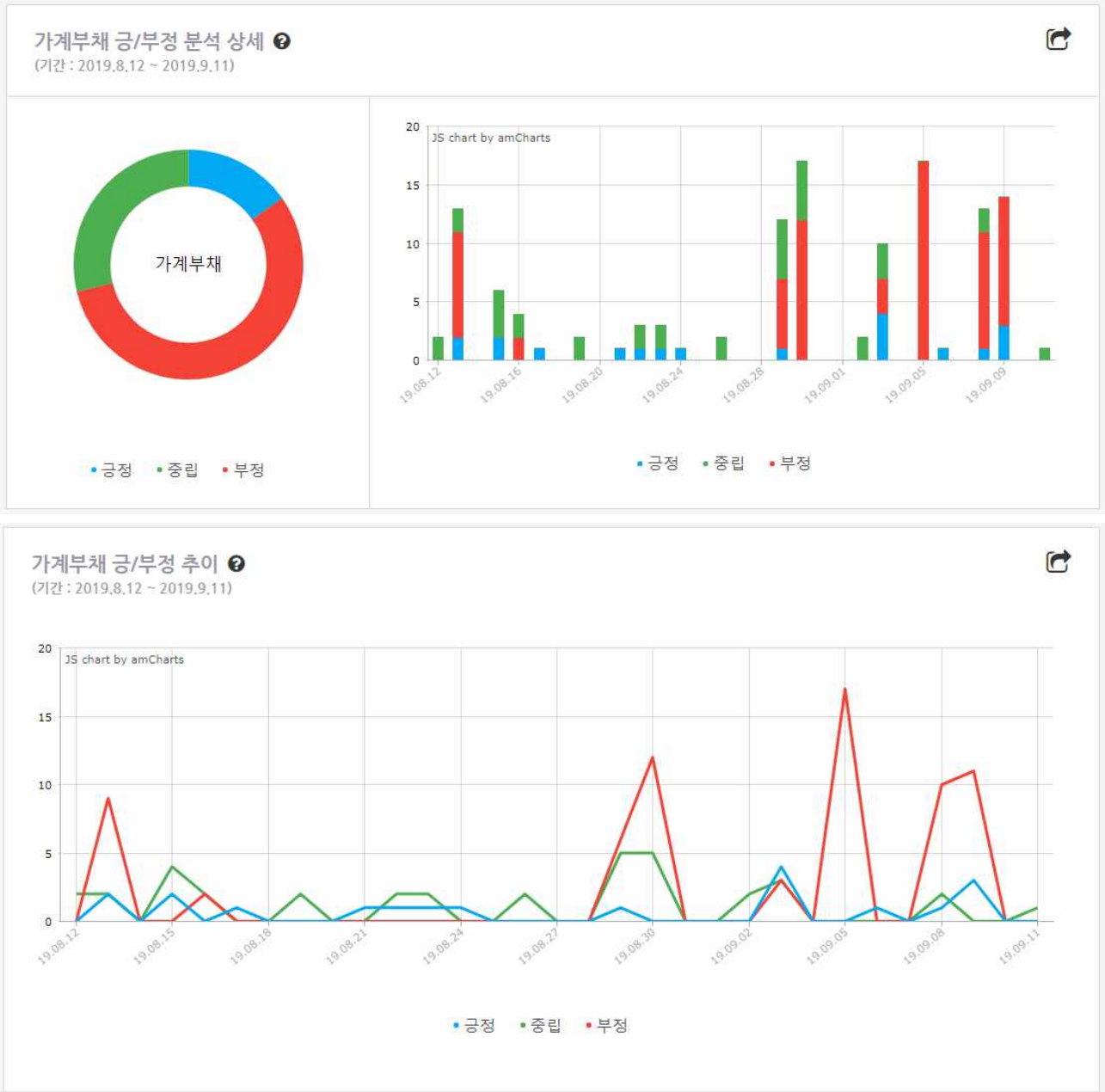


출처: Sometrend

○ 감성 분석의 상세 내용은 위 감성 분석의 결과를 원그래프 및 막대그래프, 선그래프로 보다 상세히 나타냄

- 원그래프 결과에서는 '가계부채'에 대한 감성어 결과가 긍정 및 중립적인 언어에 비해 부정적인 언어가 상대적으로 많이 출력된 것을 확인
- 막대그래프와 선그래프 결과에서는 최근 한 달의 기간 중 2019년 8월말 이후 부정적인 언어가 크게 증가한 것을 확인

[그림 6] Sometrend 화면 – 결과(4)

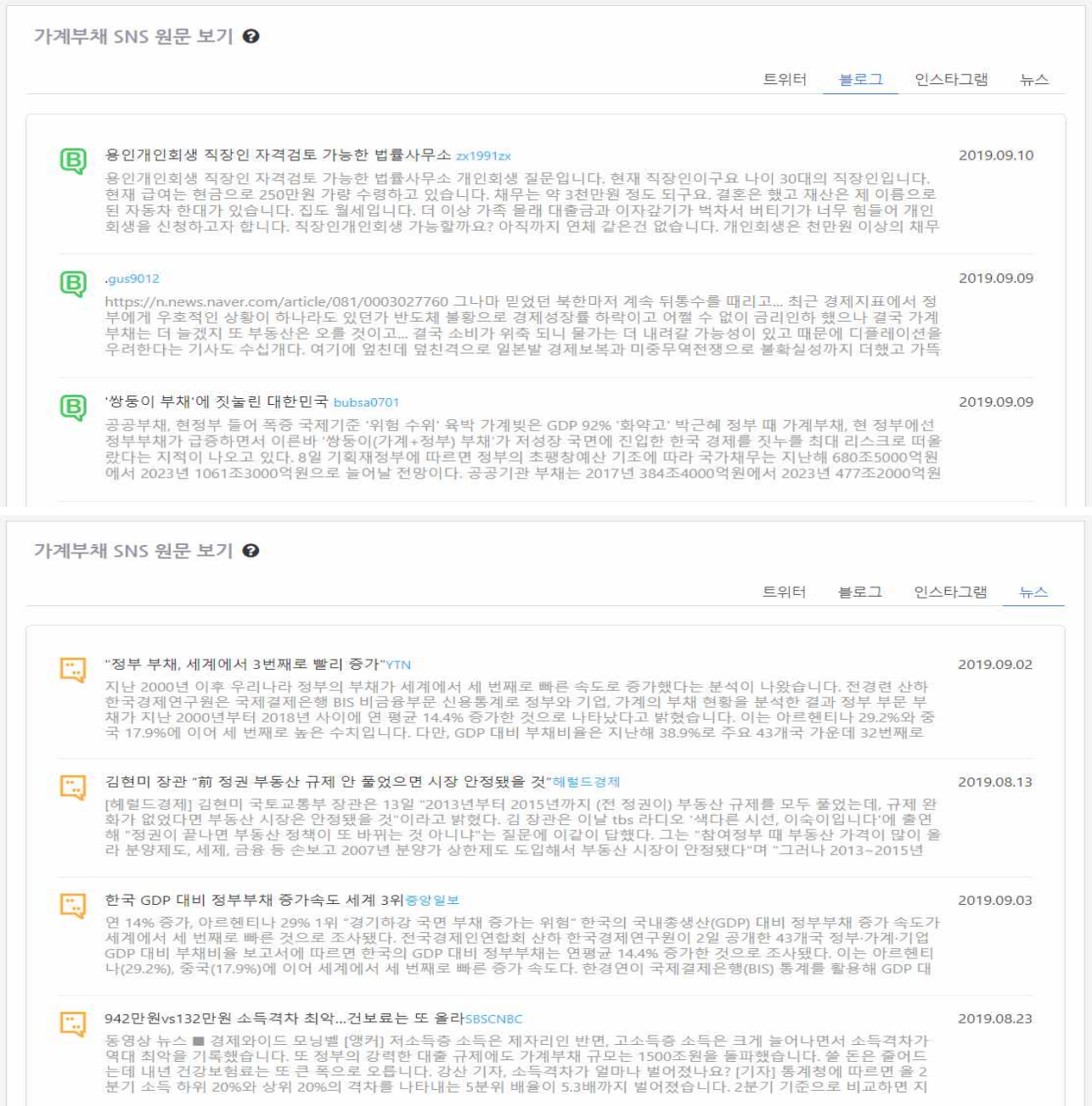


출처: Sometrend

○ SNS 원문은 검색 키워드가 노출된 SNS 원문을 그대로 나타내며, 트위터, 블로그, 인스타그램, 뉴스 형태로 구분해서 확인할 수 있음

- 본 결과에서 확인한 '가계부채' SNS 원문의 경우 트위터는 개인 의견이 많았고, 블로그는 정보전달 및 개인 의견, 인스타그램은 홍보 및 정보전달이 많았음
- 뉴스는 '가계부채'를 언급한 기사를 현재 날짜순으로 출력

[그림 7] Sometrend 화면 – 결과(5)



출처: Sometrend

3. Google - Google Trends

❑ (소개) 구글이 제공하는 Google Trends(구글 트렌드)는 다양한 국가 및 지역별로 Google 검색어를 분석하는 서비스로, 2006년 5월에 오픈됨

- ㅇ 크게 3개의 서비스를 제공
  - 탐색: 원하는 국가 및 키워드를 검색할 경우 시간 흐름(2004년부터 현재)에 따른 관심도 변화, 하위 지역별 관심도, 관련 주제, 관련 검색어를 확인할 수 있음
  - 인기 급상승 검색어: 일별 인기 급상승 검색어, 실시간 인기 급상승 검색어와 각 검색어의 검색 횟수 등을 국가별로 확인할 수 있음
  - 올해의 검색어: 해당 해의 인기 검색어를 종합, 인물, 국내 뉴스 및 이슈, 영화, TV프로그램, 게임, 스포츠 영역별로 확인할 수 있음

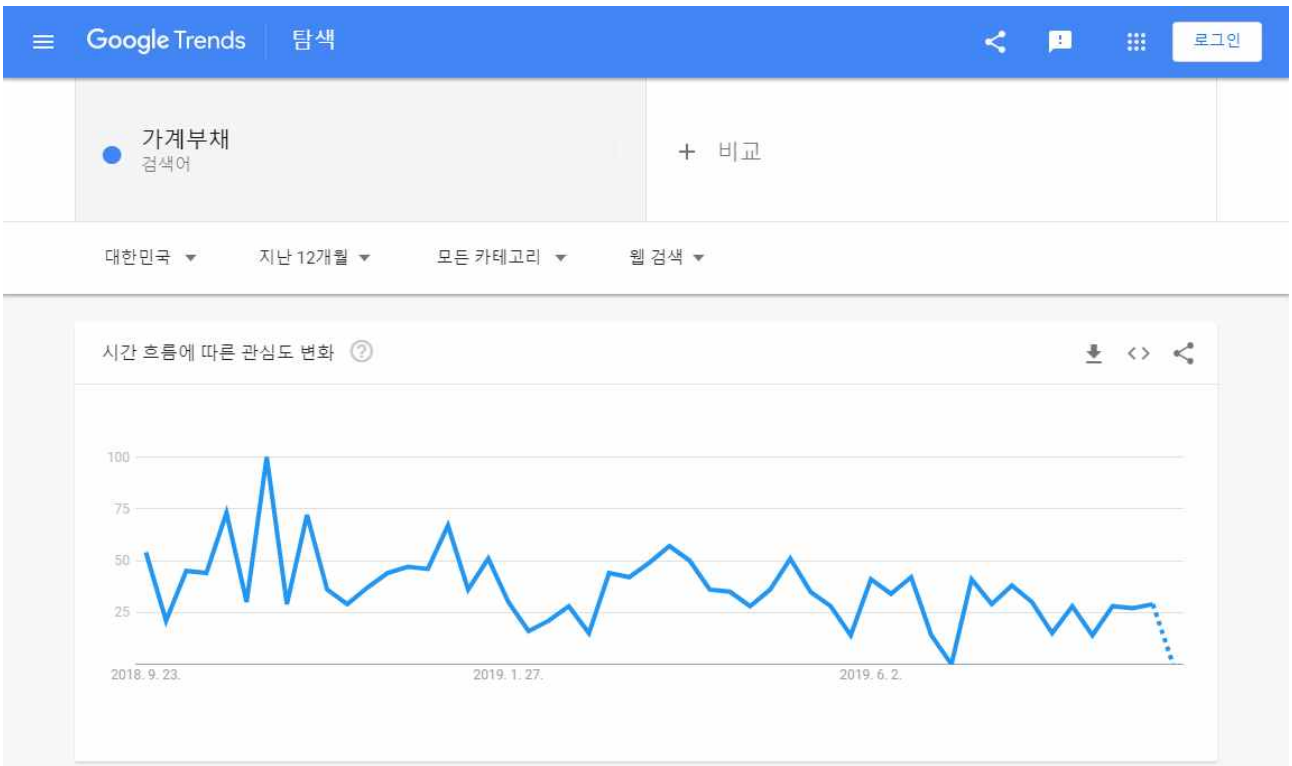
❑ (사례) Google Trends의 메인화면에서 '가계부채' 키워드 검색

- ㅇ 검색어에 '가계부채' 입력
- ㅇ 국가 '대한민국' 선택
- ㅇ 위 두 가지를 입력하면 '탐색' 화면으로 이동하며, 해당 화면에서 국가, 기간, 카테고리, 검색, 비교 검색어 등을 추가적으로 선택할 수 있음
  - 국가: Google에 등록되어 있는 전 세계 및 개별 국가로 선택 가능
  - 기간: 2004년부터 현재까지의 데이터를 조회할 수 있으며, 현 시점 기준으로 지난 1시간, 지난 4시간, 지난 1일, 지난 7일 등 단기간으로도 조회 가능
  - 카테고리: 모든 카테고리, 건강, 게임, 과학, 금융 등 카테고리별 조회 가능
  - 검색: 웹 검색, 이미지 검색, 뉴스 검색, Google 쇼핑, YouTube 검색으로 검색 데이터 출처를 구분하여 조회 가능
  - 비교 검색어: 입력한 첫 번째 검색어 우측에 비교 대상 검색어를 추가로 입력하여 두 가지 이상의 검색어 결과를 동시에 확인할 수 있음

[그림 8] Google Trends 화면 – 조회

출처: Google Trends

[그림 9] Google Trends 화면 – 탐색



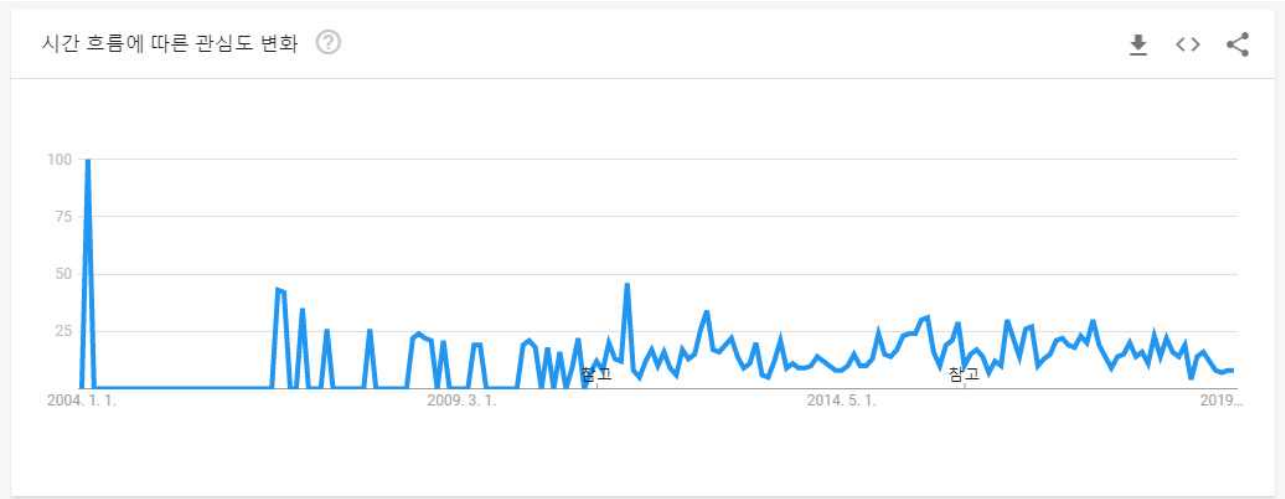
출처: Google Trends

❑ 탐색 화면에서 기간을 '2004년부터 현재까지', '모든 카테고리', '웹 검색'을 선택한 결과 관심도 변화 그래프, 하위 지역별 관심도, 관련 주제 및 관련 검색어 결과가 출력됨

○ 관심도 그래프의 경우 2004년 2월에 가장 높은 100을 기록했으며, 그 후로 등락을 반복하다가 2015년 이후 일정 수준 이상 검색된 것을 확인

- 해당 그래프 수치의 경우 특정 지역 및 기간을 기준으로 차트에서 '가장 높은 지점 대비 검색 관심도'를 나타냄(즉, 검색 빈도가 가장 높은 검색어의 경우 100, 검색 빈도가 그 절반 정도인 검색어의 경우 50, 해당 검색어에 대한 데이터가 충분하지 않은 경우 0으로 나타남)
- 값이 높을 경우 절대적인 검색수가 높은 것이 아니라, 전체 검색어 중에서 해당 검색어가 차지하는 비율이 높은 것임
- 해당 관심도 그래프에 대한 데이터를 엑셀 형태로 다운받을 수 있음

[그림 10] Google Trends 화면 – 결과(1)



출처: Google Trends

○ 하위 지역별 관심도의 경우 서울특별시가 100으로 가장 높은 값을 기록했고, 그 다음으로 부산광역시가 58, 경기도가 54 순으로 높게 기록(기타 지역은 검색어 데이터가 충분하지 않아 출력되지 않은 것으로 판단됨)

- 해당 관심도 값은 지정된 기간 동안 검색어가 어느 지역에서 가장 많이 검색되었는지 확인할 수 있음
- 값은 0~100으로 계산되며, 해당 지역의 총 검색수를 기준으로 해당 검색어가 가장

인기 있는 지역의 경우 100, 그 절반 정도로 인기 있는 지역의 경우 50, 검색어에 대한 데이터가 충분하지 않은 경우 0으로 나타남

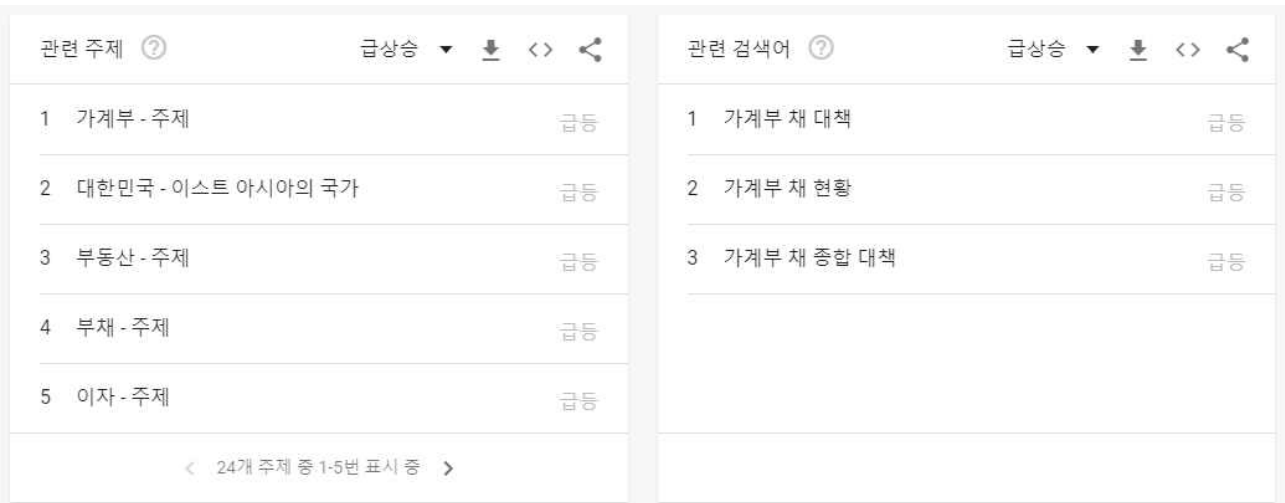
[그림 11] Google Trends 화면 – 결과(2)

출처: Google Trends

○ 관련 주제의 경우 가계부, 대한민국, 부동산 등이 주요 관련 주제어로 출력되었고, 관련 검색어의 경우 가계부채 대책, 가계부채 현황 등이 주요 관련 검색어로 출력되었음

– 관련 주제는 '가계부채'를 검색한 사용자가 추가적으로 검색한 키워드를 의미하며, 가장 인기 있는 주제는 '인기', 검색 빈도가 가장 많이 증가한 주제는 '급등'이라고 표기됨(관련 검색어도 관련 주제와 동일한 기준으로 출력)

[그림 12] Google Trends 화면 – 결과(3)



출처: Google Trends

4. Google - Google Adwords

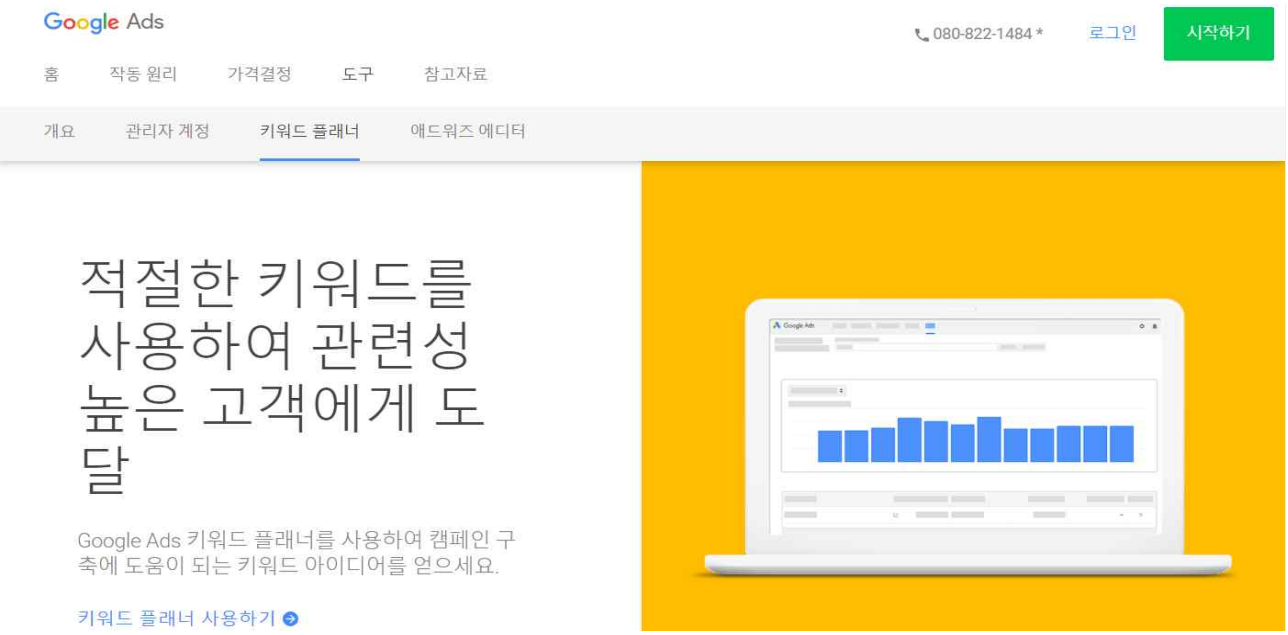
❑ (소개) 구글이 제공하는 Google Adwords(구글 에드워즈, 이하 Google Ads)는 구글에서 제작한 셀프 서비스 광고 프로그램으로, 2010년 10월에 오픈됨

- ㅇ 광고주는 에드워즈에 가입함으로써 구글 웹사이트와 애드센스에 가입한 웹사이트들에 광고를 넣을 수 있음
- ㅇ 본 서비스 중 '도구' 메뉴의 '키워드 플래너(Keyword-planner)'를 통해 데이터 추이 등을 확인할 수 있음
  - – 키워드 플래너는 신규 광고주나 경험이 있는 광고주를 위한 무료 Google Ads 도구로, 검색 네트워크 캠페인을 생성할 때 사용하는 서비스임
  - – 키워드 플래너의 목적은 사용자가 캠페인 구축에 도움이 되는 키워드 아이디어를 얻도록 도와주는 것에 있음

❑ (사례) Google Adwords 홈페이지에 로그인 후 키워드 플래너에 접속해 '가계부채' 키워드 검색

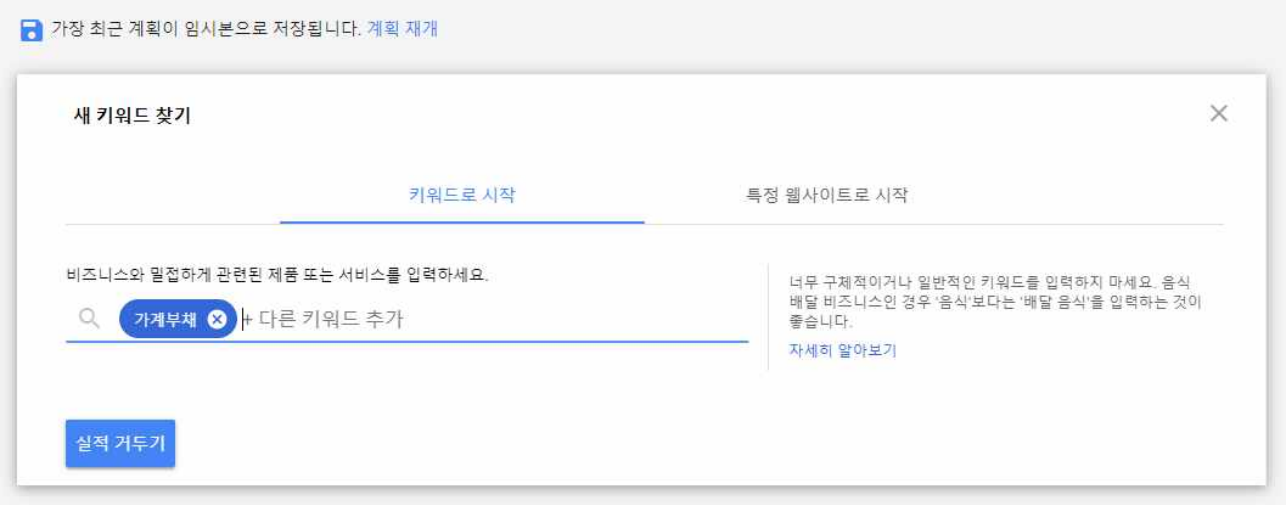
- ㅇ 새 키워드 검색 화면에서 '가계부채'를 입력하고 '실적 거두기'를 클릭

[그림 13] Google Adwords 화면 – 조회(1)



출처: Google Adwords

[그림 14] Google Adwords 화면 – 조회(2)

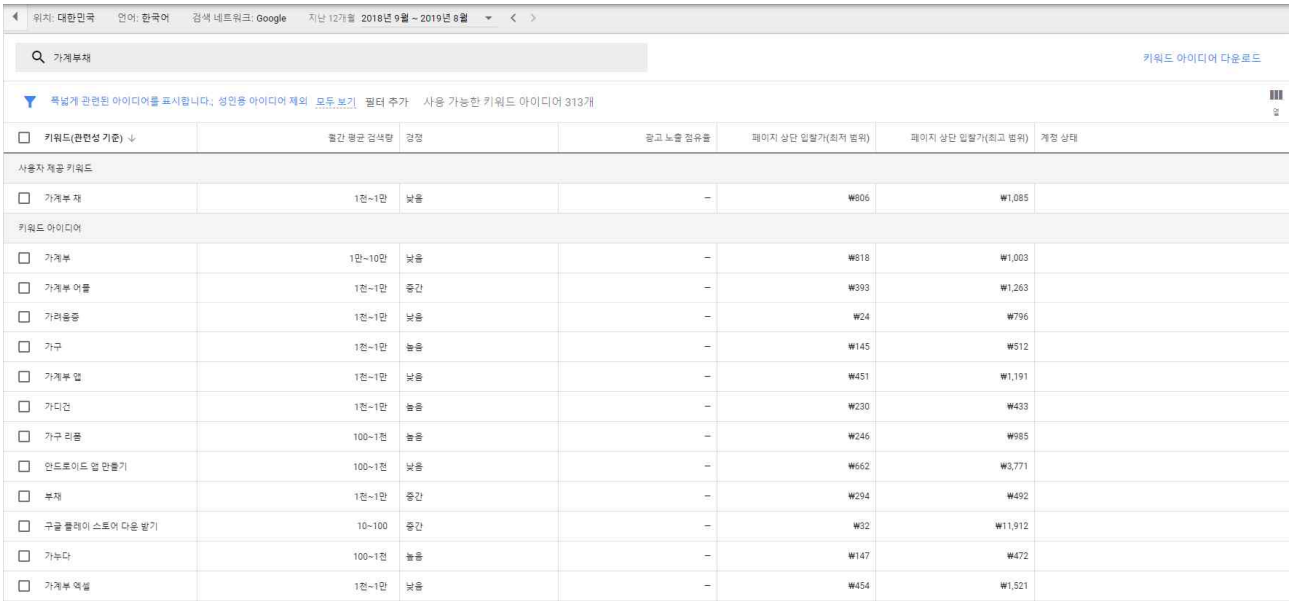


출처: Google Adwords

❑ 검색 결과 해당 가계부채로 검색되는 평균 검색량, 경쟁도, 키워드 아이디어 등이 출력됨

- 2015년 9월부터 2019년 8월까지 '가계부채'의 월간 평균 검색량은 1천~1만에 해당되며, 경쟁도는 낮은 것으로 확인됨
  - 기간은 2015년 9월부터 현재 시점 범위 내에서 조정할 수 있음
- 또한 관련된 키워드 아이디어가 313개로 확인되었으며, 해당 키워드는 엑셀 형태로 다운받을 수 있음

[그림 15] Google Adwords 화면 – 결과



키워드(관련성 기준)	월간 평균 검색량	경쟁	광고 노출 점유율	페이지 상단 입찰가(최저 범위)	페이지 상단 입찰가(최고 범위)	계정 상태
사용자 제공 키워드						
가계부 채	1천~1만	낮음	–	₩806	₩1,085	
키워드 아이디어						
가계부	1만~10만	낮음	–	₩818	₩1,003	
가계부 어플	1천~1만	중간	–	₩393	₩1,263	
가려움증	1천~1만	낮음	–	₩24	₩796	
가구	1천~1만	높음	–	₩145	₩512	
가계부 앱	1천~1만	낮음	–	₩451	₩1,191	
가디건	1천~1만	높음	–	₩230	₩433	
가구 리폼	100~1천	높음	–	₩246	₩985	
안드로이드 앱 만들기	100~1천	낮음	–	₩662	₩3,771	
부채	1천~1만	중간	–	₩294	₩492	
구글 플레이 스토어 다운 받기	10~100	중간	–	₩32	₩11,912	
가누다	100~1천	높음	–	₩147	₩472	
가계부 엑셀	1천~1만	낮음	–	₩454	₩1,521	

출처: Google Adwords

5. Google - Google Correlate

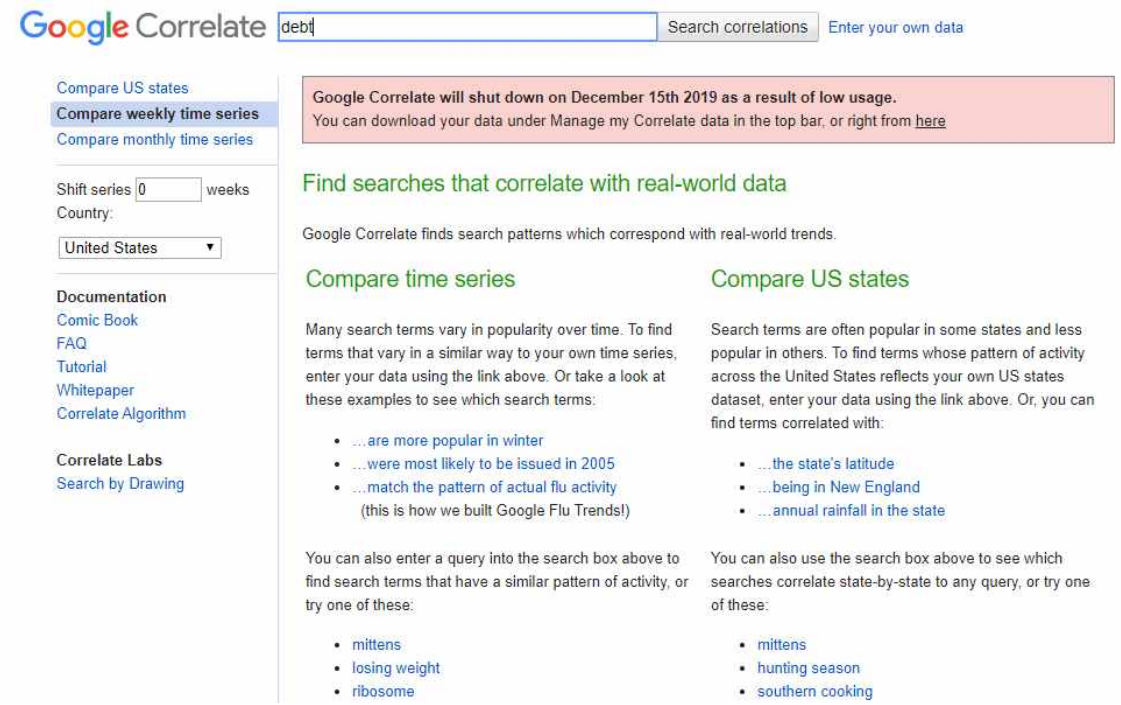
❑ (소개) 구글이 제공하는 Google Correlate(구글 코릴레이터)는 대상 데이터 시리즈와 유사한 패턴을 갖는 쿼리(queries)를 찾을 수 있는 데이터 툴 서비스로 2011년부터 제공

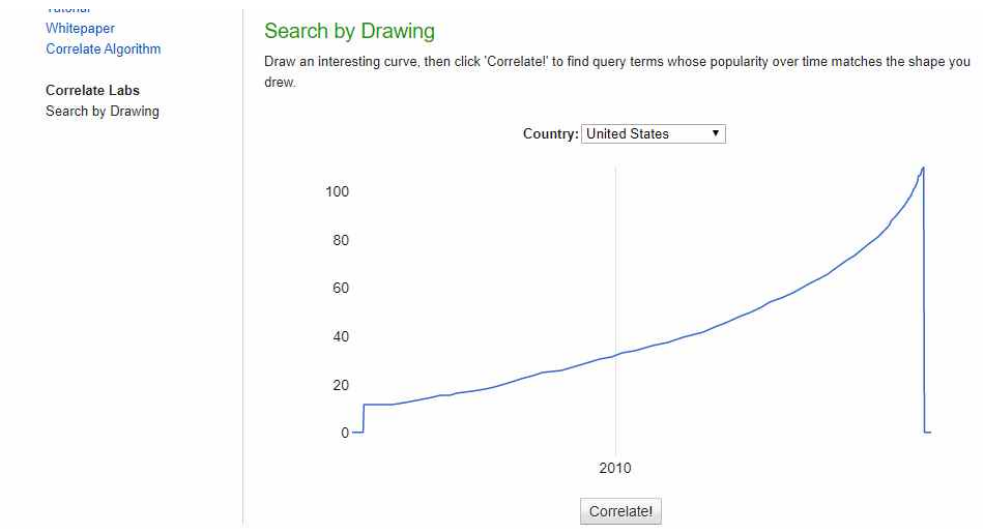
- ㅇ 2003년 1월부터 현재까지의 웹 검색 활동 데이터를 포함하고 있으나, 현재 낮은 사용도로 2019년 12월 15일에 종료될 것으로 예상
- ㅇ 단, 영문으로만 서비스가 제공되므로 영문 키워드로 검색해야 하며, 우리나라 데이터는 제공하지 않음
- ㅇ 미국 데이터와 비교가 가능하며, 주간 및 월간 타임 시리즈로도 확인할 수 있음

❑ (사례) 크게 두 가지 방법을 이용할 수 있으며, 검색어를 입력하는 방법과 추세를 그리는 방법이 있음

- ㅇ 검색어 입력 창에 'debt(혹은 원하는 다른 키워드)'를 입력하고 'Search correlations' 클릭
- ㅇ 추세를 그리는 방법은 'Search by Drawing' 메뉴로 들어가서 원하는 대상 국가를 선택하고, 추세를 직접 마우스로 그린 후 'Correlate' 클릭

[그림 16] Google Correlate 화면 – 조회



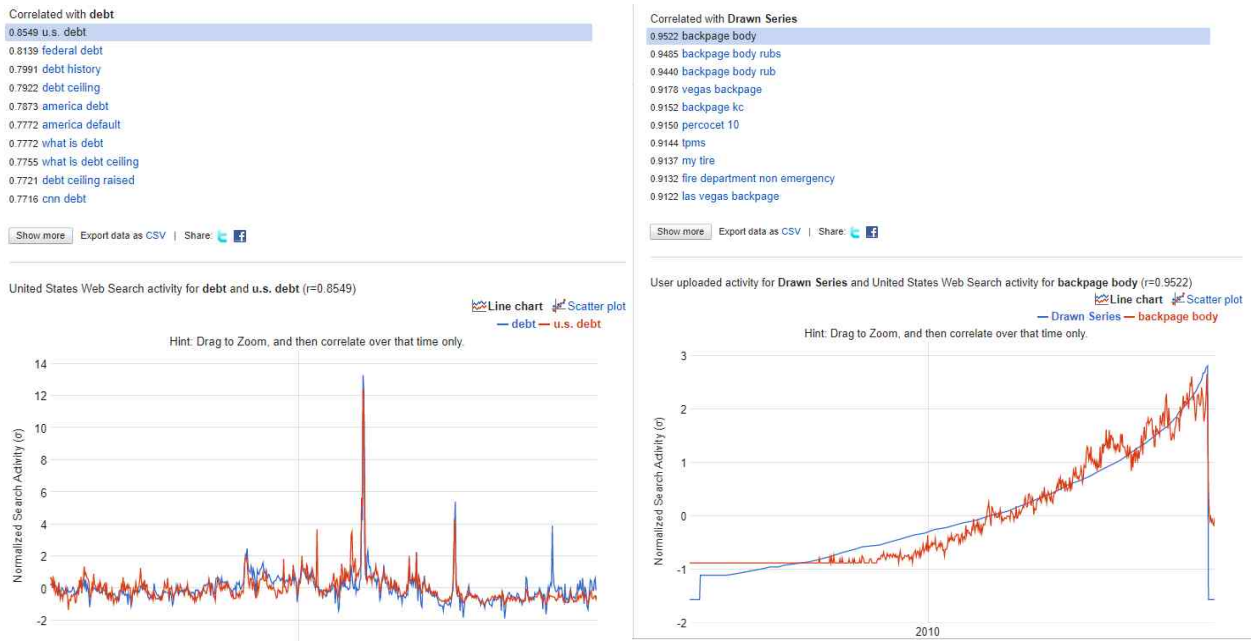


출처: Google Correlate

❑ ‘debt’ 검색어를 입력한 결과 2004년부터 현재까지 해당 검색어와 가장 상관관계가 높은 검색어 및 그래프 추세가 출력되었고, 마찬가지로 추세를 그렸을 때에도 해당 추세와 가장 상관관계가 높은 검색어들의 리스트가 출력됨

- ㅇ 해당 결과 리스트는 엑셀 형태로 다운받을 수 있음(100개의 연관 검색어와의 상관관계를 한번에 확인할 수 있음)

[그림 17] Google Correlate 화면 – 결과



출처: Google Correlate

6. Google - Google Ngram

❑ (소개) 구글이 제공하는 Google Ngram(구글 엔그램)은 Google Ngram Viewr 또는 Google Books Ngram이라고도 하며, 영어, 중국어 등의 언어를 기반으로 구글 검색 소스에서 발견된 n-grams의 연간 개수를 사용하여 쉼표로 구분된 검색 문자열 세트의 빈도를 표시하는 온라인 검색 엔진 서비스이며, 2010년에 오픈됨

- ㅇ 대소문자 구분 맞춤법을 사용하여 선택한 구문 내의 텍스트와 일치하고, 40권 이상의 책에서 찾을 수 있는 경우에 결과 그래프로 도출시킴
- ㅇ 한국어 검색은 불가능하여 해당 검색어의 빈도를 우리나라와 연결하여 판단하기엔 어려움이 있을 수 있음

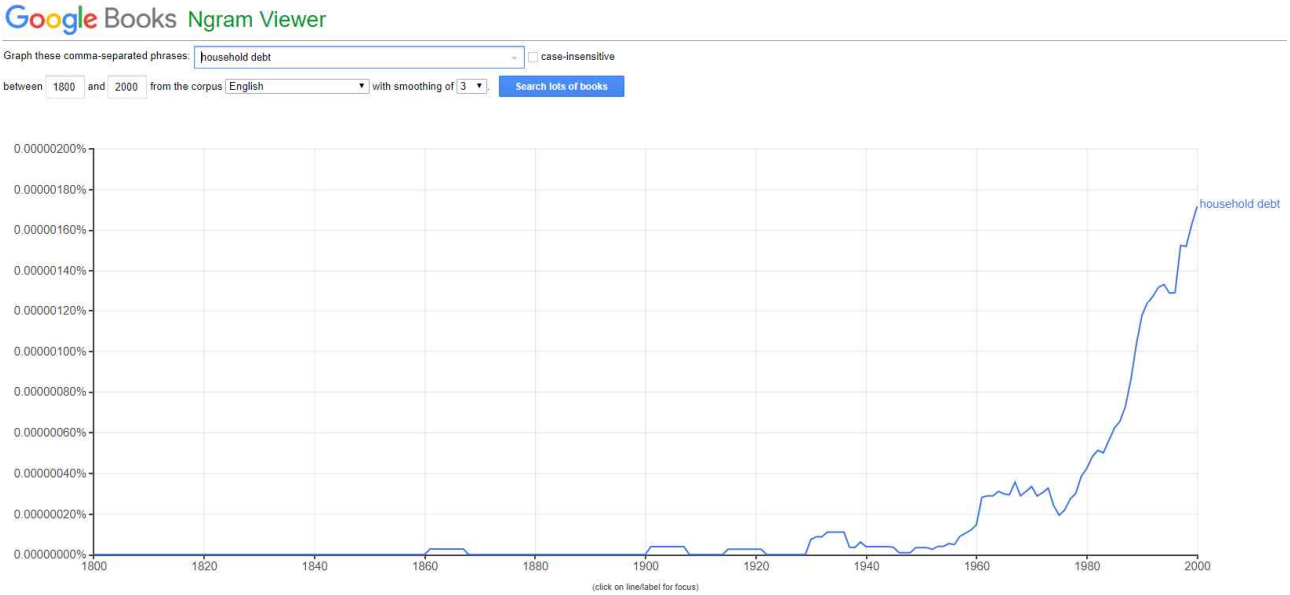
❑ (사례) Google Ngram 홈페이지에 접속 후 'Graph these comma-separated phrases' 문구 옆의 창에 'Household debt'를 입력하고 Search lots of books를 클릭

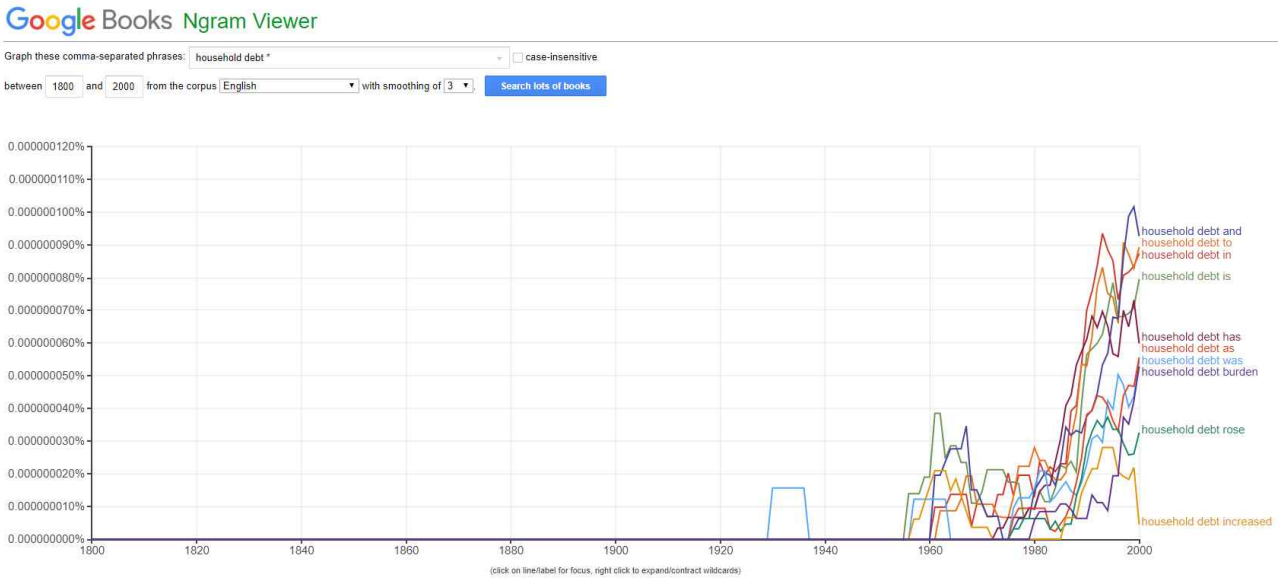
- ㅇ 기간 및 언어 국가 등을 선택할 수 있음
- ㅇ 만약 해당 키워드 다음으로 가장 인기 있는 검색어를 함께 보려면 키워드 옆에 별표 표시를 넣고 같이 입력하면 됨(예: household debt *)

❑ 검색 결과 '영어' 문자로 쓰여진 책 중에 'Household debt'의 사용 빈도는 1960년도부터 조금씩 상승 추세를 보였으며, 1980년대 이후 급증한 것으로 나타남

- ㅇ 연관 검색어를 함께 볼 수 있는 'household debt *'를 검색한 결과 household debt and, to, in, is, burden, rose, increased 등의 접속사 또는 동사가 함께 나타남
- ㅇ 검색에 사용된 도서를 확인하고 싶다면 하단에 함께 도출되는 'Search in Google Books' 메뉴를 이용할 수 있음

[그림 18] Google Ngram 화면 – 조회 및 결과





Search in Google Books:

1800 - 1975	1976 - 1992	1993	1994 - 1999	2000	household debt in	English
1800 - 1963	1964 - 1991	1992	1993 - 1999	2000	household debt is	English
1800 - 1976	1977 - 1994	1995	1996 - 1999	2000	household debt to	English
1800 - 1964	1965 - 1995	1996	1997 - 2000	2001 - 2000	household debt and	English
1800 - 1984	1985 - 1993	1994	1995 - 2000	2001 - 2000	household debt has	English
1800 - 1933	1934 - 1993	1994	1995 - 1999	2000	household debt was	English
1800 - 1980	1981 - 1992	1993	1994 - 1999	2000	household debt as	English
1800 - 1986	1987 - 1997	1998	1999	2000	household debt rose	English
1800 - 1961	1962 - 1988	1989	1990 - 1995	1996 - 2000	household debt increased	English
1800 - 1983	1984 - 1997	1998	1999	2000	household debt burden	English

출처: Google Ngram

<부록 3> 국내 부동산 및 가계부채 관련 정책 일지

1) 이명박 정부(2008~2012년)

❑ 2008년

- ○ 6·11(지방 미분양 주택 대책): 지방 미분양 LTV 완화, 취·등록세 감면, 양도세 일시적 2주택 2년 연장 등
- ○ 8·21(주택공급 기반강화 및 건설경기 보완 방안): 지방 미분양 아파트 매입, 재건축 조합원 지위 양도 허용, 분양권 전매제한 완화, 재건축 안전진단 완화, 수도권 30만 가구 공급, 지방 광역시 1가구 2주택 양도세 중과 폐지 등
- ○ 9·1 대책: 양도세 비과세 고가주택 기준 상향, 양도세율 인하 등
- ○ 9·19(보금자리 주택 건설방안): 보금자리 150만 가구 포함 2018년까지 수도권 300만, 지방 200만 가구 공급 등
- ○ 10·21(가계주거 부담완화 및 건설부문 유동성 지원, 구조조정 방안): 일시적 2주택 기간 2년 확대, 건설 부문 유동성 공급 및 구조조정 지원 등
- ○ 11·3(경제위기종합 대책): 강남 3구 외 주택투기지역 및 투기과열지구 해제, 토지투기지역 해제, 양도세 비과세 거주요건 폐지, 재건축 소형평형 의무비율 완화 등이 포함

❑ 2009년

- ○ 2·12(기재부 세제개편안 및 주택법, 공급규칙 개정): 미분양주택 양도세 한시 감면, 민간택지 분양가상한제 폐지, 주택청약종합저축 신설 등
- ○ 8.24(전세대책, 전월세 부담완화, 도시형생활주택 등 규제완화): 주택기금 전세자금 지원확대 및 민간 전세대출 보증한도 2억 확대, 오피스텔 바닥난방 허용 기준 완화 등
- ○ 8.27(서민 주거안정을 위한 공급확대 및 공급체계 개편안): 수도권 그린벨트 내 보금자리주택 32만가구 공급, 생애최초 주택청약제도 신설 및 특별공급 비율 조정
- ○ 10월: LTV, DTI 비은행권으로 확대

❑ 2010년

- ○ 3.18: 지방 미분양주택 양도세 감면 및 미분양 취등록세 감면 1년 연장, 지방 민간택지 주상복합아파트의 분양가상한제 폐지

- 4.23(주택 미분양 해소 및 거래 활성화 방안): 주택보증 환매조건부 매입 5000억원에서 4조원으로 확대, 준공전 미분양 2만 가구 매입
- 8·29(실수요 주택거래 정상화와 서민 중산층 주거안정 지원방안): 무주택 및 1가구1주택자 대출 한해 DTI 한시적 자율화, 생애최초 주택구입자금 신설, 2억원까지 주택기금 지원, 주택기금 전세자금 대출한도 5,600만원으로 상향

❑ 2011년

- 1.13(물가안정대책-전월세시장 안정화 방안 포함): 판교 순환용 주택 1300가구, 공공 보유 준공후 미분양물량 2554가구 공급, 도시형생활주택, 다세대다가구, 주거용 오피스텔에 주택기금(연리2%) 건설자금 한시 특별지원, 주택기금 전세자금 대출규모 확대, 6개월이상 무주택자 대출조건 폐지
- 2.11(전월세시장 안정 보완대책): 국민주택기금 서민, 근로자 전세자금 지원 확대, 준공후 미분양 임대시 취득세 최대 50% 감면 등
- 3·22(주택거래 활성화방안): DTI자율적용 3월말 종료, 생애최초 주택구입자금 대출 시한 올해 말까지 연장
- 5·2(건설경기 연착륙 및 주택공급 활성화방안): 부실 PF사업장 조정에 민간 배드뱅크 활용, 법인의 신규 주택 임대사업 허용(5년 이상 임대 조건) 등
- 6.29(가계부채 연착륙 종합대책): 고위험 주택담보대출 건전성 강화, 고정금리대출 전환 유도(중도상환수수료 면제)
- 8.18(전월세시장 안정대책): 수도권 임대주택사업자 세제지원 요건 3가구이상에서 1가구이상 임대로 완화, 소형주택 전세보증금 소득세 한시 배제 등
- 12·7(주택시장 정상화 및 서민주거안정 지원방안): 다주택자 양도세 중과 폐지, 강남3구 투기과열지구에서 해제 등

❑ 2012년

- 2.24(제2금융권 가계부채 보완대책): 상호금융 예대율 80% 이내 운용, 거치식/일시상환/다중채무자대출: 고위험으로 분류
- 5·10(부동산 대책): 주택거래 정상화 및 서민, 중산층 주거안정 지원방안(강남 3구 투기과열지구와 투기지역 해제, 다주택자 양도세 중과 폐지, 재건축 초과이익 부담금 2년 중지, 1주택 양도세 비과세 요건 완화, 민영주택 재당첨 제한 폐지 등)

○ 8.17(부동산 대책): 30대 무주택근로자, 은퇴자 DTI 규제완화, 순자산도 소득으로 인정

2) 박근혜 정부(2013~2016년)

❑ 2013년

○ 4.1(주택시장 정상화 종합대책): 1년간 매입주택에 5년간 양도소득세, 생애주택에 취득세를 각각 면제하고 민간공급의 촉진을 위해 공공분양을 감소(연 7만 가구에서 5만가구로)

○ 7.24(후속조치): 보금자리 공공분양을 4년간 11만9000가구 축소

○ 8.28(전·월세대책): 공유형 모기지제, 주택 취득세율의 1~3% 차등인하

○ 12.3(후속조치): '목돈 안드는 전세II' 폐기, 행복주택 공급 30% 축소(20만 가구 → 14만 가구) 등

❑ 2014년

○ 2·26(임대차시장 선진화 방안): 고액 전세 거주에 대한 정부지원 조정(시중은행의 전세대출에 대한 공적보증 지원 대상을 서민층으로 집중, 전세보증금 4억(지방2억) 초과 시 보증 제한

○ 2.27(가계부채 구조개선 촉진방안): 소득대비 부채비율 2017년말까지 5%포인트 조정, 은행권 주담대의 고정금리/분할상환비중 40% 수준 확대

○ 7.24(새 경제팀 경제정책 방향): LTV, DTI 70% 일괄적용, 청약통장 일원화

○ 9.1(서민주거 안정 강화방안): 재건축 연한완화 및 청약1순위 자격완화(1년 단축)

○ 10.30(서민주거비 부담완화방안): 2015년까지 매입, 전세임대 주택 1.3만호 추가 공급, 보증부 월세 거주가구에 대한 지원 강화, 사회취약계층에 대한 월세자금 대출 도입

○ 12.23(부동산 대책): 민간택지 분양가 상한제 탄력 적용, 재건축 초과이익환수제 3년 유예 연장

❑ 2015년

○ 2.26(가계부채 구조개선 프로그램): 20조원 규모 고정금리, 분할상환 안심전환대출 상품 개발, 주택금융신용보증기금 출연료 개편(주담대 구조조정 실적에 따라 은행 인센티브 부여)

○ 7.22(가계부채 종합관리방안): 원칙적 분할상환 취급, 거치기간 1년 이내 단축 유도, 상환능력 심사 선진화(담보가 아닌 상환능력 중심 심사), 상호금융권 비주택대출 등 관리 강화(리

스크 감소요인만 상향 허용, 최저한도 60%에서 50%로 축소)

- 8.28: 민간 임대주택에 관한 특별법 공포
- 10.5: 집주인 리모델링 임대 시범사업 출발
- 12.14(가계부채 대응방향): 여신(주담대) 심사 선진화 방안 시행(소득증빙자료 객관성 확보, 신규 주택구입자금, 고부담대출 등은 비거치식 분할상환 확보, 변동금리 주담대는 금리상승 가능성 고려하여 대출한도 산정)
- 12.23: 뉴스테이 사업 공급 목표 상향 시사

❑ 2016년

- 1.14: 국토부 1차 뉴스테이 공급촉진지구 발표, 도심형·토지임대형 등 시범사업 추가 추진
- 2.24(가계부채 평가 및 대응방향): 가계소득 증대를 통한 상환능력 제고, 2017년말까지 분할상환 비중 50% 목표 상향조정, 서민금융진흥원 설립, 4대 정책 서민금융상품 공급 확대
- 4.28 대책: 행복주택과 뉴스테이 공급 물량 증가 발표
- 5.2(가계부채 대책): 비수도권에서도 가계 여신(주담대) 심사 선진화 가이드라인 시행
- 8.25 대책(가계부채 대책): LH 택지공급 조절, 중도금 대출 보증 강화, 분양보증 강화 등
- 11.3 대책(주택시장의 안정적 관리방안): 조정대상지역 전매제한 강화, 1순위 청약자격 강화, 재당첨 제한, 중도금 대출 보증요건 강화, 청약가점제 자율시행 유보 등
- 11.24(가계부채 후속조치): 집단대출에 대해서도 여신심사가이드라인 적용, 상호금융 등 맞춤형 여신심사가이드라인 도입, 총체적 상환능력심사(DSR) 도입, 가계부채 특별점검 연장 실시
- 12.24 대책: 잔금대출 규제 강화 계획 발표

3) 문재인 정부(2017년~현재)

❑ 2017년

- 3.6(가계부채 후속조치): 3.13부터 상호금융권에도 맞춤형 여신심사가이드라인 적용
- 5.30(가계부채 후속조치): 6.1부터 상호금융권 주택담보대출 맞춤형 여신심사 가이드라인 적용 모든 조합 및 금고로 확대
- 6.19대책(부동산): LTV 70%에서 60% 축소 및 전매제한 강화
- 8.2대책(부동산): 투기지역, 투기과열지구 지정 및 재당첨제한, 조합원지위양도 금지, 조정

지역 다주택자 양도세 중과(18.4월 시행)

- ㅇ 9.5대책(추가 조치): 분당, 대구 수성 투기과열지구 지정
- ㅇ 10.24(가계부채 종합대책): 신DTI(총부채상환비율 산정 시 기존에 받은 주택담보대출의 이자와 원금 모두 반영), 금융기관 DSR 도입, 참고지표 RTI(임대업이자상환비율), LTI(소득대비대출비율) 등 도입
- ㅇ 11.29(주거복지로드맵): 공적임대 85만 가구 + 공공분양 15만 가구 등 총 100만 가구 공급
- ㅇ 12.13(임대주택 등록 활성화 방안): 세제 혜택으로 임대사업자 등록 유도

❑ 2018년

- ㅇ 2.20(재건축 안전진단 강화): 구조안전성 평가 가중치 상향, 주거환경평가 가중치 하향
- ㅇ 7.5(신혼부부, 청년 주거지원방안): 신혼희망타운 공급 물량 10만 가구로 확대
- ㅇ 7.6(종합부동산세 개편안): 종부세율 인상 및 중과
- ㅇ 8.27(부동산 대책): 투기지역 확대 및 광명, 하남 투기과열지구 지정
- ㅇ 9.13(주택시장 안정대책): 종부세율 확정 및 신규취득 주택 임대사업자 혜택 축소, 대출규제 강화
- ㅇ 9.21(주택공급대책): 3기 신도시 계획 발표 및 서울 유휴부지 활용 방안
- ㅇ 12.19(수도권 주택공급 확대 방안): 3기 신도시 지정(과천, 계양, 교산, 왕숙)

❑ 2019년

- ㅇ 1.17(주택담보대출 채무조정 활성화 방안): 개인회생 연계형 신복위 주담대 채무조정 도입, 신복위 채무조정 주담대에 대한 건전성분류 개선, 신복위 주담대 채무조정 방안의 다양성 제고 등
- ㅇ 2.20(주택담보대출 상품출시): 월상환액 고정형 주담대 및 금리상한형 주담대 출시
- ㅇ 5.7(수도권 주택공급 확대 방안): 3기 신도시 추가 지정(창릉, 대장)
- ㅇ 5.30(가계부채): 제2금융권 DSR 관리지표 도입방안 발표
- ㅇ 7.2: 신복위 취약채무자 특별 감면제도와 주택담보대출 채무조정 활성화 방안 시행 발표(7.8부터)
- ㅇ ㉑ 7.23: 8월말(잠정) 서민과 실수요자를 위한 저금리 갈아타기 주택담보대출 상품 출시 예정 발표

<부표>

1) 개인회생 및 개인파산 신청 분석

<부표 1> 개인회생에 대한 검색기록 설명변수의 시차변수 분석

변수명		개인회생 실적치	
차수	변수	Coef.	Std.Err.
-	검색지수	29.7346***	11.2401
L1.	검색지수	4.0483	12.7746
L2.	검색지수	6.8335	11.9712
L3.	검색지수	1.2581	10.9478
L4.	검색지수	13.3366	11.3987
L5.	검색지수	-12.9639	11.8577
L6.	검색지수	13.4832	12.7273
L7.	검색지수	0.6582	14.5915
L8.	검색지수	-2.8674	14.9688
L9.	검색지수	22.8372	17.0812
L10.	검색지수	7.2257	16.4508
L11.	검색지수	-9.9063	14.7563
L12.	검색지수	34.6324**	14.4444
L13.	검색지수	-17.2701	13.5094
L14.	검색지수	7.0477	13.7963
L15.	검색지수	6.8733	16.7706
L16.	검색지수	-18.6684	17.6998
L17.	검색지수	25.6229***	15.0381
L18.	검색지수	1.7347	15.4325
L19.	검색지수	-16.3376	15.4920
L20.	검색지수	24.4363	16.9457
L21.	검색지수	-25.1988	15.2746
L22.	검색지수	4.8467	12.7429
L23.	검색지수	-6.2973	13.7194
L24.	검색지수	2.4504	14.1697
L25.	검색지수	6.8646	14.2586
L26.	검색지수	3.0964	13.0536
L27.	검색지수	-1.7800	12.9391
L28.	검색지수	2.9466	11.0772
L29.	검색지수	10.4298	12.5159
L30.	검색지수	-1.7041	12.3568
L31.	검색지수	-7.2625	13.1090
L32.	검색지수	8.8442	11.8290
L33.	검색지수	-2.7844	11.4049
L34.	검색지수	2.8251	10.3876
L35.	검색지수	8.3180	9.7781
L36.	검색지수	7.3304	7.5576
	상수항	1633.652***	358.9173
관측치 수		113	
R-squared		0.9169	

주: 1) ( ) 안은 표준오차를 의미함
2) ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함
3) 월별 더미 변수 결과는 지면관계상 생략함

<부표 2> 개인파산 및 개인회생에 대한 단위근 검정 결과

1) 개인파산

Dickey-Fuller test for unit root				
관측치 수 = 112				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-3.523	-3.506	-2.889	-2.579
MacKinnon approximate p-value for Z(t) = 0.0074				

2) 개인회생

Dickey-Fuller test for unit root				
관측치 수 = 112				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-3.713	-3.506	-2.889	-2.579
MacKinnon approximate p-value for Z(t) = 0.0039				

3) 개인회생 검색

Dickey-Fuller test for unit root				
관측치 수 = 186				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-4.621	-3.481	-2.884	-2.574
MacKinnon approximate p-value for Z(t) = 0.0001				

4) 개인파산 검색

Dickey-Fuller test for unit root				
관측치 수 = 186				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-10.477	-3.481	-2.884	-2.574
MacKinnon approximate p-value for Z(t) = 0.0000				

<부표 3> 그랜저 인과관계 분석(구글 검색과 개인회생 실적 관계)

Granger causality test				
	시차1	시차2	시차3	시차4
관측치 수	112	111	110	109
검색지수 → 개인회생 실적치	8.01***	3.96**	3.04**	2.92**
개인회생 실적치 → 검색지수	0.32	1.28	0.94	1.01

주: 각 숫자는 F값이며, ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

3) 대출태도 지수와 검색기록 간 관계 분석(분기자료)

<부표1> 대출태도지수와 검색 기록 요약 통계량 (분기자료)

변수명	관측치 수	평균	표준편차	최소값	최대값
가계주택대출금리 검색수치(평균)	62	5.11	3.27	0.00	13.00
가계주택대출 검색수치(평균)	62	21.79	7.21	10.67	48.67
가계대출	62	643,578.40	208,949.00	328,557.20	1,049,759.00
가계주택대출	47	437,902.90	99,392.91	292,813.50	614,818.20
가계 대출태도(일반)	62	-1.76	7.86	-33.00	9.00
가계 대출태도(부동산)	62	-6.94	15.64	-47.00	19.00
가계 수요지수(일반)	62	5.27	7.48	-13.00	26.00
가계 수요지수(부동산)	62	5.69	14.77	-25.00	34.00
가계대출 증가율(%)	58	7.95	2.49	2.82	12.95
가계주택대출 증가율(%)	43	6.84	2.42	2.00	11.31

주: 분석기간은 2004.1~2019.2

<부표2> 분기자료 단위근 검정 결과

1) 가계주택대출

Dickey-Fuller test for unit root				
관측치 수 = 46				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	1.740	-3.607	-2.941	-2.605
MacKinnon approximate p-value for Z(t) = 0.9982				

2) 가계대출

Dickey-Fuller test for unit root				
관측치 수 = 61				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	2.933	-3.565	-2.921	-2.596
MacKinnon approximate p-value for Z(t) = 1.0000				

3) 가계주택대출금리 검색수치(평균)

Dickey-Fuller test for unit root				
관측치 수 = 61				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-6.690	-3.565	-2.921	-2.596
MacKinnon approximate p-value for Z(t) = 0.0000				

4) 가계주택대출 검색수치(평균)

Dickey-Fuller test for unit root				
관측치 수 = 61				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-6.332	-3.565	-2.921	-2.596
MacKinnon approximate p-value for Z(t) = 0.0000				

5) 가계주택대출 증가율(%)

Dickey-Fuller test for unit root				
관측치 수 = 42				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-1.622	-3.634	-2.952	-2.610
MacKinnon approximate p-value for Z(t) = 0.4715				

6) 가계대출 증가율(%)

Dickey-Fuller test for unit root				
관측치 수 = 57				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-1.408	-3.570	-2.924	-2.597
MacKinnon approximate p-value for Z(t) = 0.5785				

7) △가계대출 증가율(%)

Dickey-Fuller test for unit root				
관측치 수 = 41				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-4.751	-3.641	-2.955	-2.611
MacKinnon approximate p-value for Z(t) = 0.0001				

<부표3> 분기자료 그랜저 인과관계 검정결과

Granger causality test			
	시차2	시차3	시차4
관측치 수	41	40	39
주택담보대출금리 검색 → 가계주택대출 증가율	1.13	0.73	0.54
가계주택대출 증가율 → 주택담보대출금리 검색	5.30***	3.76**	2.71**

주: 각 숫자는 F값이며, ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

3) 월별 대출검색지수와 대출증가율 관계 분석

<부표1> 월별자료 대출 검색지수 관련 요약통계량

변수명	관측치 수	평균	표준편차	최소값	최대값
가계대출	186	639,406.00	207,109.90	320,769.80	1,049,759.00
가계주택대출	139	437,632.20	97,480.81	292,436.50	614,818.20
가계주택대출금리 검색지수	187	5.11	5.93	0.00	39.00
가계주택대출 검색지수	187	21.81	11.80	0.00	100.00
가계주택대출 연체율	173	0.70	0.34	0.20	1.90
가계대출금리	187	5.04	1.19	3.23	7.79
실업률	187	3.53	0.29	3.00	4.70
가계대출 증가율(%)	174	8.00	2.48	2.64	12.97
가계주택대출 증가율(%)	127	6.86	2.46	1.62	12.28
주택가격지수	187	88.02	10.93	66.20	101.20
아파트가격지수	187	86.05	11.99	62.30	100.50

출처: 검색지수는 Google Trends에서 추출하고 나머지 변수들은 한국은행 ECOS DB 추출

<부표2> 월별자료 단위근 검정 결과

1) 가계주택대출

Dickey-Fuller test for unit root				
관측치 수 = 138				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	2.031	-3.497	-2.887	-2.577
MacKinnon approximate p-value for Z(t) = 0.9987				

2) 가계대출

Dickey-Fuller test for unit root				
관측치 수 = 185				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	4.234	-3.482	-2.884	-2.574
MacKinnon approximate p-value for Z(t) = 1.0000				

3) 가계주택담보대출 금리검색

Dickey-Fuller test for unit root				
관측치 수 = 186				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-13.630	-3.481	-2.884	-2.574
MacKinnon approximate p-value for Z(t) = 0.0000				

4) 가계대출 검색

Dickey-Fuller test for unit root				
관측치 수 = 186				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-14.225	-3.481	-2.884	-2.574
MacKinnon approximate p-value for Z(t) = 0.0000				

5) 가계주택대출 증가율

Dickey-Fuller test for unit root				
관측치 수 = 126				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-1.456	-3.501	-2.888	-2.578
MacKinnon approximate p-value for Z(t) = 0.5550				

6) 가계대출 증가율

Dickey-Fuller test for unit root				
관측치 수 = 173				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-1.226	-3.486	-2.885	-2.575
MacKinnon approximate p-value for Z(t) = 0.6622				

7) 가계주택대출의 차분

Dickey-Fuller test for unit root				
관측치 수 = 137				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-7.977	-3.498	-2.888	-2.578
MacKinnon approximate p-value for Z(t) = 0.0000				

8) 가계주택대출 증가율의 차분

Dickey-Fuller test for unit root				
관측치 수 = 125				
	검정통계량	1% 유의수준	5% 유의수준	10% 유의수준
Z(t)	-7.705	-3.502	-2.888	-2.578
MacKinnon approximate p-value for Z(t) = 0.0000				

<부표3> 월별 자료 그랜저 인과관계 검정결과

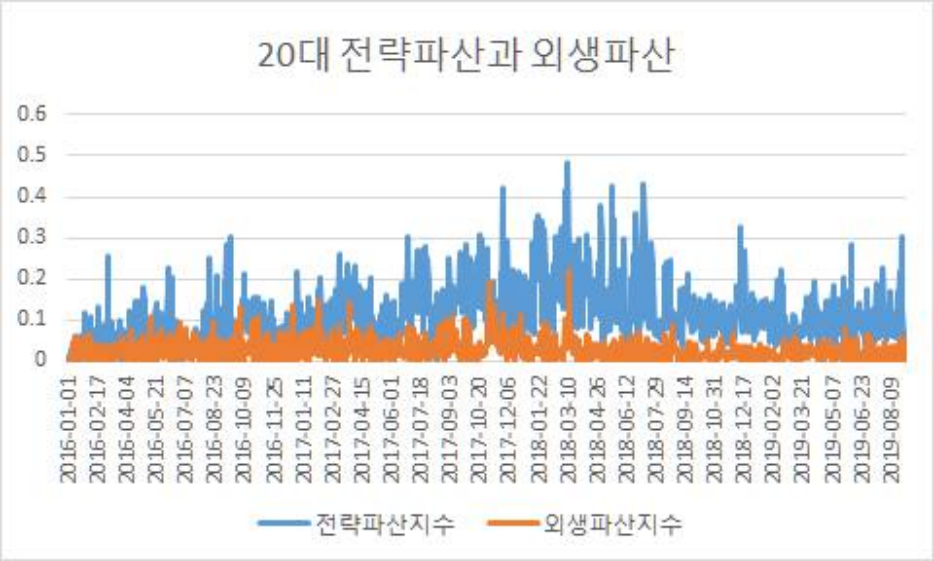
Granger causality test			
	시차2	시차3	시차4
관측치 수	125	124	123
주택담보대출금리 검색 → 가계주택대출 증가율	1.41	0.81	0.86
가계주택대출 증가율 → 주택담보대출금리 검색	1.15	4.91***	4.16***

주: 각 숫자는 F값이며, ***, **, *는 1%, 5%, 10% 수준에서 통계적으로 유의함

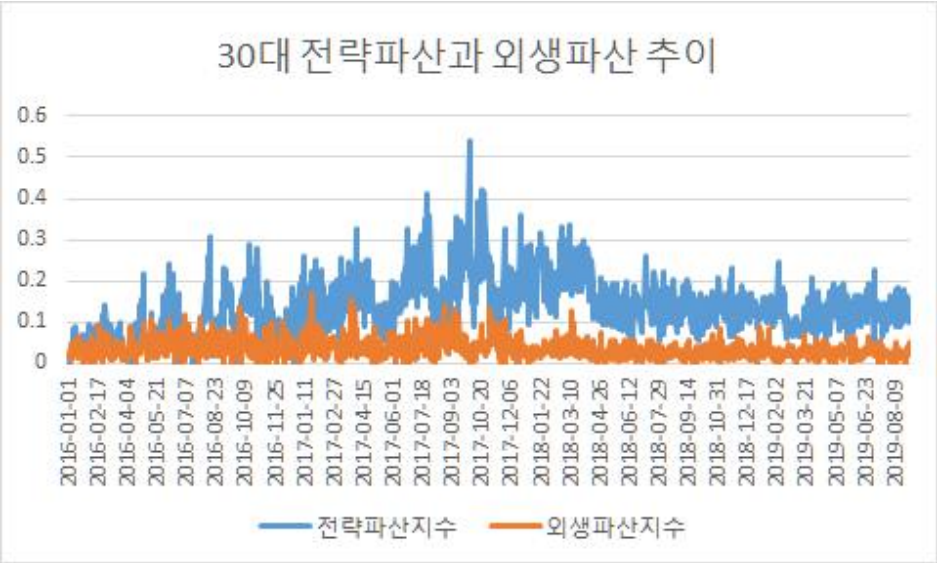
<부도>

<부도 1> 전략 파산 검색 추이(연령대별)

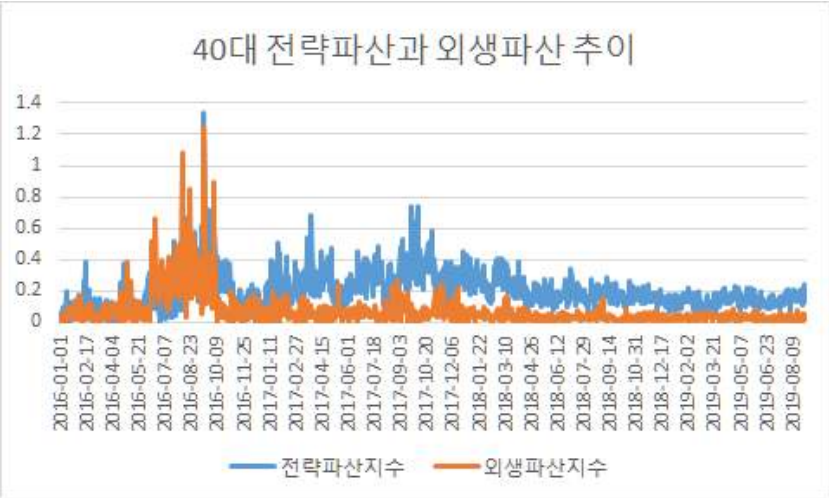
① 20대



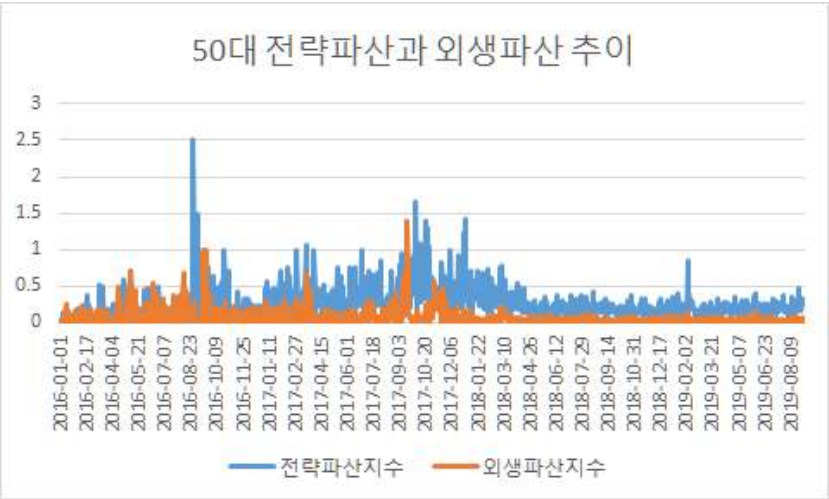
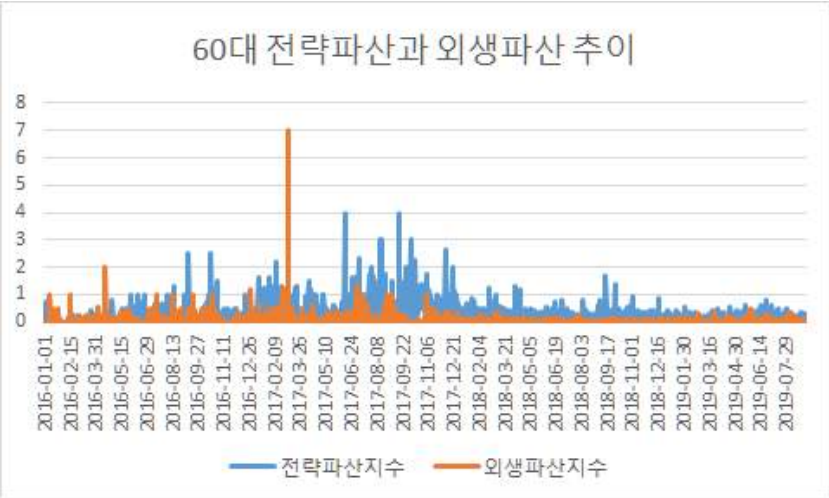
② 30대



③ 40대



④ 50대



⑤ 60대