자동화된 키워드 발견과 통계적 연어 추출 방법을 통한 언론의 정치적 양극화 현상의 발견: 암호화폐 규제에 대한 조선일보와 한겨레 신문의 사례

송주경

Discovering Political Polarization in Press Automated Keyword Discovery Techniques and Statistical Collocation Extraction Methods: A Case of Chosun Press and Hankyoreh Press.

Song Jookyung (2015150132, Department of Politcal Science and International Studies)

# 서론

언론은 기본적으로 세계의 여러 소식을 선별하여 전달하는 매개체다. 정치에 있어서는 평시에는 정책과 입법 현황, 선거철에는 후보자에 대한 정보와 공약 내용 등을 제공하여 대의민주주의의 주체인 시민들이 각자의 견해를 형성할 수 있게 돕는다. 그런데 세계 각국을 막론하고 언론들이 완전히 객관적이기보다는 각자 모종의 정치적 입장을 가지고 있다고 보는 견해가 상식으로 통용된다. 한국에서는 흔히 조선일보·중앙일보·동아일보가 우파적 가치를 대변한다고 보고 한겨레·경향신문·오마이뉴스 등이 좌파적 가치를 전파한다고 본다. 또한 사건이 전개됨에 따라 언론사는 더 정치 편향적인 기사를 쓰기도 한다. 이는 정치적 양극화 현상의 일환으로 특정 정치 성향을 가진 언론사가 정체성을 공고화하기 위해 점차 극단화 되는 것이다. 이런 인식은 단순히 대중적인 인식에서만 그치지 않고 교육현장이나 언론보도 심지어는 연구 현장에서도 통용된다.

이러한 언론의 정치적 편향성을 연구하기 위해 이전 연구에서는 지표 등을 이용해 객관성을 확보하려고 했지만 실상 객관성을 담보할 수 없는 경우가 많았다. 직접 일일이 기사를 읽어보고 점수를 매기거나 연구자의 언어적 감각에 따라 긍정어와 부정어를 나누는 방법이나 엄밀한 검증 없이 개발된 설문 조사를 사용하는 경우가 대부분이었다. 이런 방법은 객관성이 담보될 수 없을 뿐만 아니라 측정의 재현가능성 또한 희박하고 일반화를 보장할 수 있을 만큼 대량의 자료를 다루는 데에도 큰 제약이 있을 수밖에 없으며 시간의 흐름에 따른 분석도 어렵다.

본 연구에서는 컴퓨터학과 전산언어학에서 언어 자료를 다루는 방법인 자연언어처리 중 가장 기초적인 방법을 통해 우선 언론간의 입장이 정말로 다른지, 그리고 정말 반대되는지를 살펴보려고 한다. 연구절차는 우선 웹크롤러를 설계하여 자동으로 대량의 관련 뉴스 기사를 수집한 후 이 자료를 바탕으로 카이스퀘어 잔차를 이용한 자동 키워드 추출 방식을 적용해 각 언론사의 언어사용을 대표하는 단어들을 추출한다. 그리고 이 단어들에 대한 통계적 공기어를 추출하여 그 키워드들이 어떤 단어들과 같이 사용되었는지 확인하는데 여기서 확인되는 양상을 통해 각 언론사의 언어 사용이, 궁극적으로는 그 입장과 의도가 상이함을 보이려고 한다.

구체적으로는 2018년의 주요사건 중 하나인 암호화폐에 관련된 조선일보와 한겨레의 기사를 2017년 12월 1일부터 2018년 2월 28일까지 수집하여 분석하였다. 조선일보와 한겨레를 선정한 이유는 이 두 언론사가 통념적으로 우파와 좌파의 양극단에 속하는 신문사이고 그 기사의 양이 완전히 동일하지는 않지만 비교가 가능한 수준인 대표적인 신문이기 때문이다. 이렇게 수집한 자료는 울산과학기술대학교에서 공개한 형태소 분석기를 사용하여 형태소 분석을 한 다음 각 형태소의 빈도를 기반으로 분석을 하였는데, 동사의 어간이 활용되고 조사에 따라 단어의 문법적 의미가 달라지는 한국어의 교착어적 특성을 반영하기 위한 유일한 방법이기 때문이다. 시간의 흐름에 따라 변화하는 양상을 살펴보기 위해 각 언론사의 기사 앞뒤 20%를 각각 전기와 후기로 임의적으로 분류하였다. 이는 일 단위로 시계열 분석을 하면 각 날마다 주제에 따라 단어의 사용이 달라지고 이 때문에 단어들의 사용이 불연속적이고 불규칙적으로 변하는 현상이 일반적으로 관측되기 때문에 일정량의 기사를 하나의 시간대로 묶은 것이다. 앞서 언급한 카이스퀘어 잔차를 이용하여 조선일보와 한겨레 전체 기사에서의 키워드 그리고 각 언론사의 기사를 앞서 말한 기준에 따라 전기 후기로 나누어 각각의 카이스퀘어 잔차를 살펴보고 여기에 빈도를 고려하여 키워드를 추출하였다. 이렇게 추출한 키워드를 바탕으로 연어를 살펴보았는데 여기서 사용한 특정 키워드 앞뒤 6단어가 무엇이었는지를 나타낸 바이그램을 가지고 빈도표를 만든 후 이 빈도를 통해 T 검정과 정보 엔트로피 지표 (Mutual Information Score; MI-score)를 계산해 특정 키워드가 가장 유의미하게 사용된 단어가 무엇인지 살펴보았다.

# 선행연구

개인이 미디어 매체가 특정 사건사고에 대해 언론사의 정치적 이념에 편향되게 방송하는 것으로 인식하는 것을 매체 편향 인식(perception of media bias)라고 한다. ([[1]](#footnote-1)Ho et al., 2011). 매체 편향 인식을 일으키는 주된 요인 중 하나는 매체가 정치적으로 편향된 경우인데, 많은 선행 연구에서는 언론사가 어떻게 편향되었는지, 어떤 방법을 사용해서 의도적으로 정치적 관념을 내비치는지 등을 연구한다. ([[2]](#footnote-2)최선규 외, 2012 등). 이러한 선행 연구는 뉴스 기사를 모두 수집하여 언론사의 성향을 분석하기 보다는 대표적인 사례를 들어 그 안에서 분석하는 질적 연구가 주로 이루어졌다. 인터넷 매체가 신문 매체보다 더 활발하게 증가하였고, 언론의 목표가 질적으로 좋은 뉴스를 전달하기 보다는 방대한 정보를 빠르고 편리하게 전달하는 것으로 바뀜에 따라 연구 방법 또한 양적 연구로 바뀌어야 할 필요성이 제기되었다.

사회심리학적인 연구로, 집단 양극화란 집단의 정체성을 공고화하기 위해 본래의 성향보다 더 극단적인 선택을 하는 것을 의미한다. 이는 특히 정치적 성향에서 두드러지는데, 어떤 사건이 발생하여 집단 생존에 위험이 되거나 경쟁자가 나타났을 때, 집단의 정체성을 강화시키는 목적으로 원래 가졌던 정치적 이념보다 더 과격해지는 것이다. 개인의 정치색보다 집단의 정치색이 더 뚜렷해지는 것 또한 이 때문인데, 극우파와 극좌파 세력의 부상과 정체성 연구와 밀접한 관련이 있다. [[3]](#footnote-3)(Aronson, 2010). 이 이론에 따르면 장기화된 사건 사고에 대해 언론사의 논조가 바뀌는 것 또한 해당 언론사의 정체성과 밀접한 관련이 있을 것이다. 예를 들어 보수적인 언론사가 시간이 갈수록 보수적 유권자가 선호할만한 기사를 주로 낸다던가 진보적 언론사의 보수적 집단을 비판하는 기사 양이 증가되는 것은 해당 언론사의 정치적 양극화가 극대화되었다고 할 수 있다.

# 연구설계

**3.1가설**

정부의 비트코인 규제에 대한 조선일보와 한겨레 신문의 웹 기사에서 언어사용은 시간이 지날수록 달라질 것이다. 이 때, 언론사 간에 차이가 나는 방향으로 달라질 것이다. 보수성향을 가진 조선일보는 보수인 수용자가 선호하는 방향으로, 진보성향을 가진 한겨레 신문는 진보적인 수용자가 선호하는 방향으로 달라질 것이다.

**3.2연구 방법 및 개념어 설명**

조선일보와 한겨레 신문의 홈페이지에서 2017년 12월 1일부터 2018년 2월 28일까지 기사를 웹 크롤링하였다. 이는 법무부에서 가상통화 엄정 규제 방안을 모색하기 위해 법무부, 기획재정부, 금융위원회 등 범정부 입법 논의 본격화가 이루어진 기점인 2017년 12월 4일을 처음 암호화폐 규제에 대한 정부의 논의가 시작한 날로 취급하였기 때문이다. 규제 논의의 앞 3일을 포함시킨 것은 11월 30일 이낙연 국무총리의 암호화폐 규제 발언이후 언론사의 관심이 쏠린 것을 고려하기 위함이다. 후에 2018년 1월 12일 박상기 법무부 장관의 거래소 폐쇄 발언, 2018년 1월 14일 금융위원회의 가상화폐 후속 조치, 2018년 1월 15일 가상화폐 실명제 실시를 기점으로 가장 논의가 활발했던 1월과 2월의 기사를 추출했다. 조선일보 168개, 한겨레 신문 58개로 총 226개 기사를 추출하여 분석하였다. 이 때 사용한 웹 크롤러는 연구자가 파이썬 기반으로 만든 것이며 4.1 데이터 수집 및 전처리에서 더 자세히 설명한다. 그리고 시계열 데이터를 만들기 위해 기사를 시간 순서대로 정렬하여 앞 n개와 뒤 n개 기사를 뽑아 전기, 후기 데이터로 나누었다. 이때 n개는 각 언론사의 전체 기사 수의 20%로 연구자가 임의로 추출하였다. 시계열 데이터는 언론사의 정치적 양극화 현상이 극대화되는지 키워드를 추출하여 알아보기 위함이다. 그리하여 전기, 후기 조선 일보 기사는 각 34개, 전기 후기 한겨레 신문 기사는 12개로 하였다.

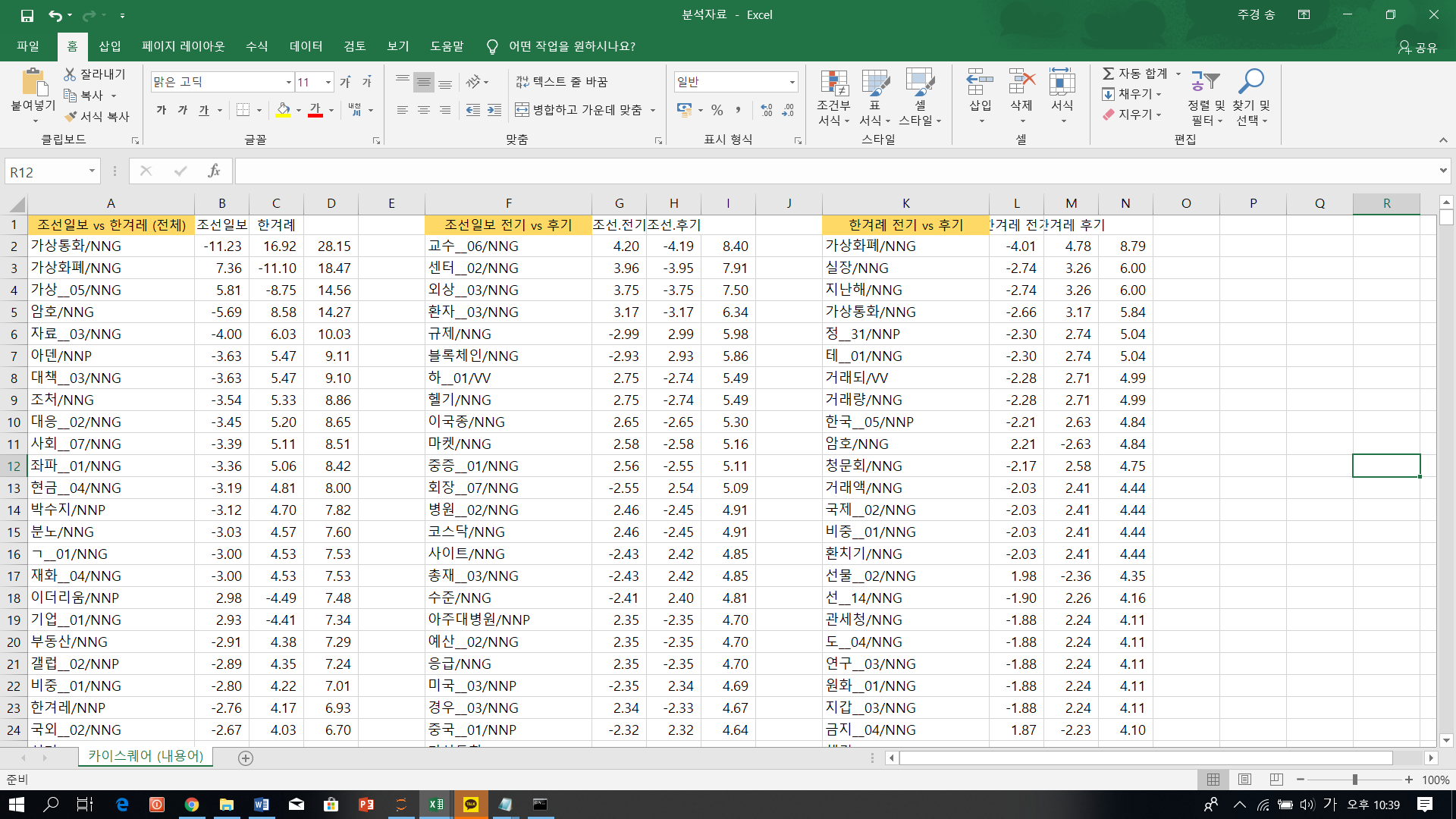
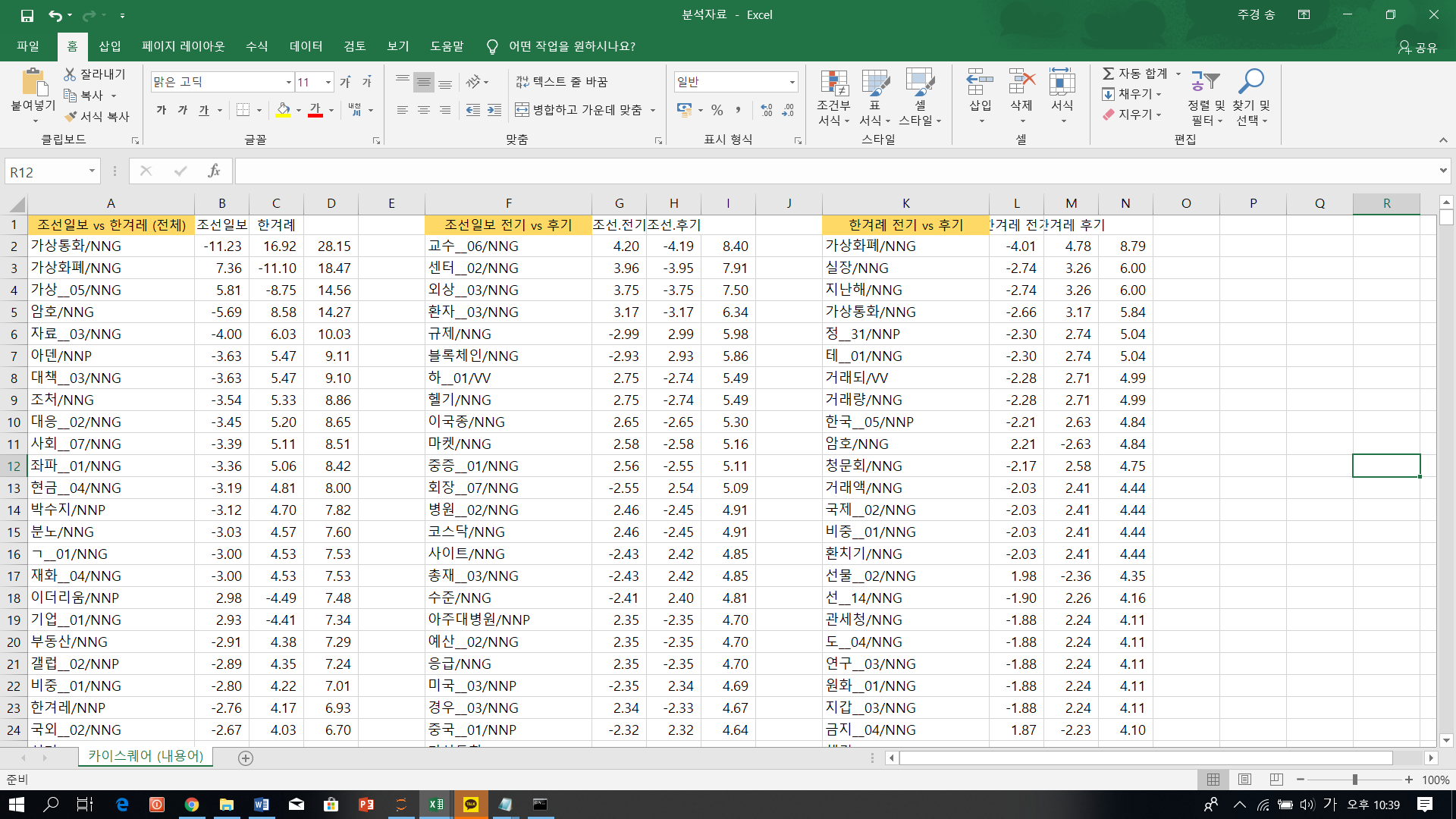
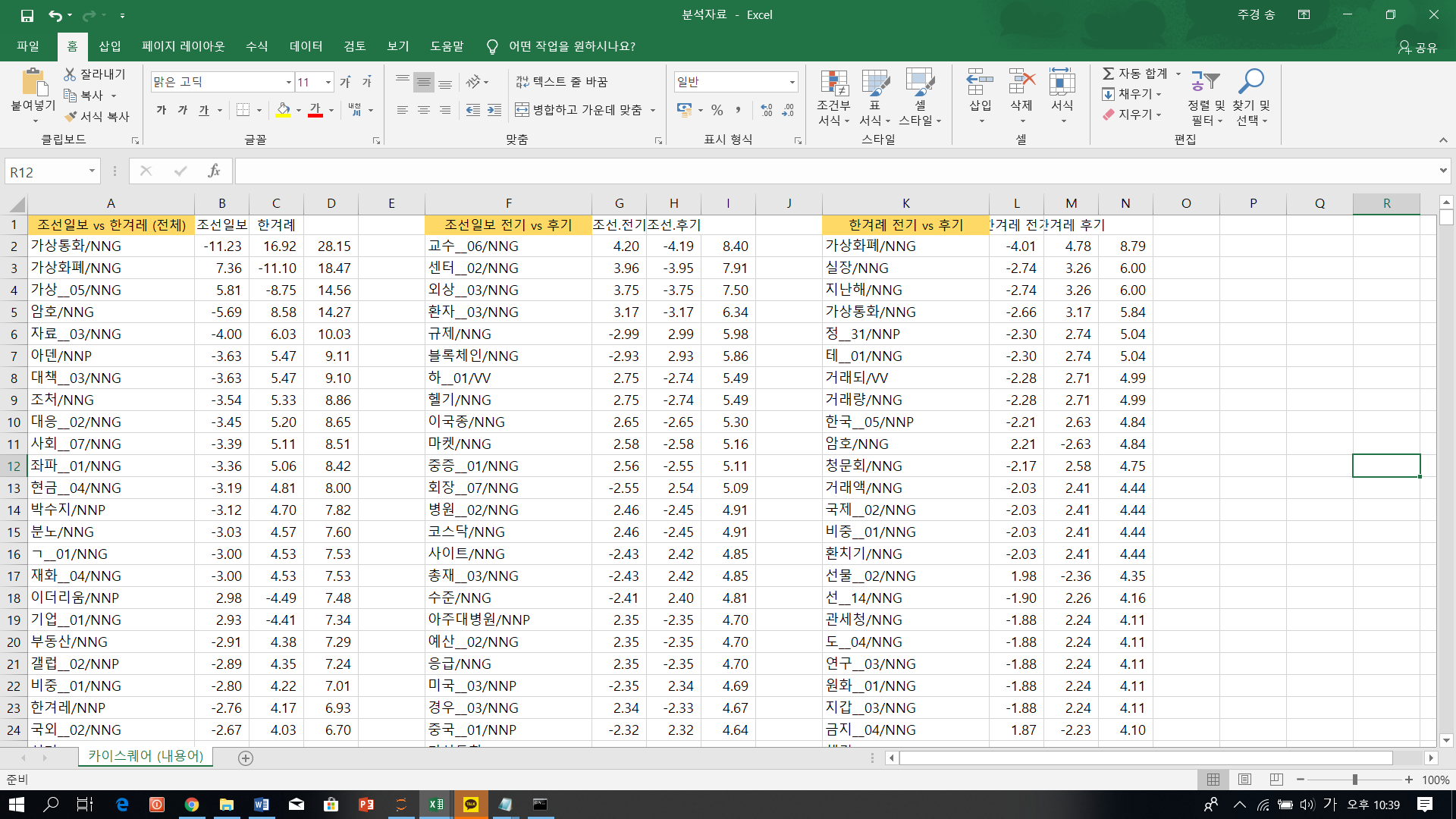
데이터 전처리 작업을 마친 후 각 추출된 내용어의 분포를 통계적으로 검정하기 위해 먼저 카이 스퀘어 잔차를 매겼다. 조선일보 전체 기사와 한겨레 신문 전체 기사, 조선일보 전기와 후기, 한겨레 신문 전기와 후기로 나눠서 단어 분포를 조사하여 카이스퀘어 점수를 매긴후 그 차가 가장 큰 단어를 추출 후 이를 바탕으로 키워드를 형성했다. 키워드는 해당 범주에서 빈도수가 100위 안이고 카이스퀘어 점수 또한 상위 100위 안에 드는 내용어 중에서 추출했다. 단어의 빈도수 또한 고려하는 이유는 카이스퀘어 점수 차가 다른 단어보다 높다고 해서 전체 범주를 대표한다고 볼 수 없기 때문이다. 그 예로 조선일보 전기와 후기에서 ‘이국종 교수’와 관련한 ‘외상, 센터, 환자’ 등의 단어가 카이제곱 잔차의 상위에 오른 것을 볼 수 있는데 이는 검색 엔진이 검색어와 전혀 다른 주제의 기사까지 보여주기 때문이다. 이러한 주제와 관련 없는 노이즈 데이터를 배제하기 위해 단어 빈도수까지 고려해주어야 한다.

그림 카이제곱 잔차 상위 10개 단어.

각 범주에 대한 키워드를 5개씩 뽑은 후 연어(Collocation)를 조사하여 어떤 문맥에서 각 단어가 쓰였는지 조사했다. 연어란 높은 빈도로 나란히 쓰이는 두 단어를 의미한다. 이 연구에서는 키워드의 앞 뒤 6단어를 조사하여 키워드와 관련성이 높은 단어를 연어로 간주하였다. 만약 다른 범주 간의 같은 키워드의 연어가 서로 다르다면, 각 키워드를 어떤 의도로 사용했는지 알 수 있는 지표가 된다. 이렇게 키워드 간의 변화를 수치화한 후 각 언론사 별로 전기, 후기의 차이를 알아본다. 차이가 클 경우 여론 등에 의해 사건의 인식이 달라졌으므로 간주하고, 언론별 주제 인식이 어떻게 달라졌는지 확인한다. 가설에 따르면 언론사의 언어 사용은 시간이 변화함에 따라 언론사의 기존 정치적 성향이 더 뚜렷해지는 방향으로 바뀔 것이다.

# 측정

**4.1데이터 수집 및 전 처리**

본 논문에서 수집된 데이터는 조선일보와 한겨레 신문의 2018년 정부의 비트코인 규제 발표에 대한 웹기사를 각 언론사의 [[4]](#footnote-4)홈페이지에서 web crawling한 것 이다. 위 언론사는 각각 보수진영과 진보진영을 대표하는 것으로 알려져 있다. 본 연구에서는 각 언론사의 기사를 수집하기 위해 web crawler를 직접 만들었는데 이때 Python 기반의 BeautifulSoup 라이브러리를 사용하였고 XML 형태로 필요한 부분만 파싱하여 사용했다. 검색어는 ‘비트코인 규제’로 각 언론사의 전체 기사를 추출하였으며 csv파일 형태로 기사 전문, 언론사명, 날짜 별로 분류하여 저장한다. 키워드 추출에 활용할 데이터는 기사 전문이며 언론사 이름과 날짜는 정치적 성향이 시간의 흐름에 따라 어떻게 변화하는지를 알아보기 위한 태그이다.

수집된 전문 기사 중 유의미한 단어를 추출하기 위해 거치는 작업이 전처리(pre-processing)이다. 텍스트마이닝에서 필수적으로 활용되는 부분으로 기호, 특수 문자 등과 같은 불용어(stopword)를 제거하고 데이터를 가공하는 것이 그 목적이다. 한국어는 조사나 보조어에 의해 단어의 형태가 바뀌기 때문에 유의미한 단어를 추출하기 위해서는 형태소 분석을 거쳐야 한다. 본 연구는 울산대학교 한국어처리연구실에서 개발한 세종 코퍼스 기반 ‘UTagger’ 형태소 분석기를 사용하였으며 ‘UTagger’ 안에 포함되어 있지 않은 ‘비트코인, 가상화폐, 이더리움, 리플, 테더 코인, 라이트코인, 이오스, 포모, 가상통화, 가즈아’ 와 같은 신조어 및 고유명사는 따로 사용자 사전에 추가하였다. 또한 조사, 보조용언 등과 같이 의미 변화를 일으키지 않는 단어를 제거하여 일반명사(NNG), 고유명사(NNP), 동사(VV), 형용사(VA), 일반부사(MAG)의 태그를 내용어로 설정하였다. 그리하여 총 조선일보 16353개, 한겨레 신문 6119개의 내용어를 추출하였다.

**4.2. 키워드 추출**

키워드 추출은 카이스퀘어 잔차를 차용한다. 엄밀하게 말하면 통계적 의미에서의 카이스퀘어 잔차를 적용한 것이 아니라, 그 수리적 계산 방법만 차용한 별도의 지표라고 봄이 옳다. 본 연구에서는 각 단어를 하나의 범주로 간주한 것인데 이렇게 형성된 범주는 천 단위가 넘어가고 이는 일반적인 카이스퀘어 잔차 적용 상황이 아닐 뿐만 아니라 정말로 각 단어가 각자의 독립된 범주인지, 언어의 사용이라는 것이 맥락의 영향을 받지 않은 단순한 기대빈도만으로 형성되는 것인지에 대한 근원적인 의문이 가능하기 때문이다. 따라서 여기서 사용하는 카이스퀘어 잔차를 이용한 키워드 추출은 일반적인 단어 사용 빈도에 기대어 봤을 때 일반적인 언어사용보다 더 높은 확률로 등장한 단어가 중요한 단어, 즉 키워드라고 간주하고 있는 것이며 이를 실현하기 위해 유사한 원리를 가진 통계적 방법인 카이스퀘어 잔차를 차용한 것이라고 할 수 있다.

카이스퀘어 잔차가 높으면 해당 형태소가 주제어에 가깝다고 보는 것이 카이스퀘어 잔차를 사용한 방법의 특징이다. 이 방법의 다른 특징으로는 언급한 것 이외에도 언급하지 아니한 것까지 살펴볼 수 있는데 조선일보와 한겨레에서 둘 다 높은 빈도로 사용한 단어의 경우에는 값이 서로 0에 가깝게 나오기 때문이다.

이때 단순히 카이스퀘어 잔차 점수를 바탕으로 키워드를 추출하면 현실적인 문제로 진짜 키워드를 추출하지 못하는 경우가 생긴다. 석 달 동안 각 언론사가 생산할 수 있는 기사는 통계적으로 큰 숫자가 아니다. 단순히 키워드를 입력해 검색되는 결과를 수집했기 때문에 가상화폐가 언급만 되고 주제는 다른 기사 혹은 여러 토막 기사가 함께 실린 경우에 다른 토막 기사의 내용이 종종 포함되게 되는데 이에 결과가 오염되는 경우가 발생하기 때문이다.

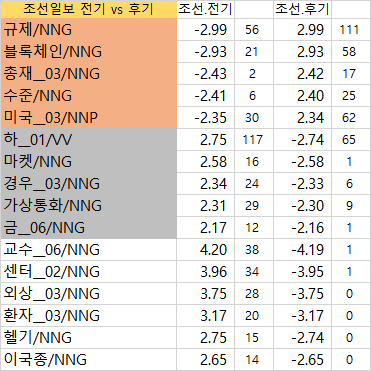
조선일보의 전기와 후기 자료를 비교한 자료에서 이 점이 극명하게 드러나는데 토막기사 여러 개를 모아 놓은 기사에서 등장한 이국종 교수 관련 기사에 의해 자료가 심하게 오염이 되었음을 확인할 수 있다. 여기서 큰 숫자로 적힌 것이 카이스퀘어 잔차 점수고 작은 글자로 적혀 있는 것이 해당 단어의 실제 빈도수다. 가장 카이스퀘어 잔차 점수가 높은 것이 ‘교수\_\_06/NNG’로 이는 뒤이어 나오는 이국종 교수 관련 키워드와 연관이 있다. 이런 자료를 배재하기 위해서 총 빈도수 30회 이상이라는 기준을 정해봤는데 이 마저도 이국종 교수 관련 키워드를 완전히 제거하지 못할 뿐만 아니라 키워드로 선정한 자료 일부도 이 기준에 미달하지 못함을 알 수 있다. 잘못된 자료가 섞인 것이 원인이므로 웹크롤러에 주제를 고려할 수 있는 기능을 적용시켜야 하는데, 자연어처리 기법 중 주제 판단 기법을 적용하면 해결할 수 있을 것이다.

그림 카이스퀘어 잔차를 내림차순한 결과

아래는 이 연구에서 위 문제점을 보완하여 빈도수를 고려해 추출한 키워드다. 조선일보와 한겨레 기사 전체를 비교해서 두 신문사를 대비하는 키워드를 추출해보았으며, 각 신문사별로 앞서 말한 전기 후기 기준에 따라 각 시기별로 대조되는 키워드를 제시하였다. 마찬가지로 카이스퀘어 잔차만을 적용하지 않고 빈도 또한 고려하였는데 조선일보와 한겨레의 전체 기사에서는 카이스퀘어 잔차 점수가 높고 한쪽이 50번 이상인 경우를 적용하고 3개는 기준에 미달하지만 포함시켰고, 조선일보 전기와 후기는 앞서 말했던 합계 빈도수 30회 이상, 한겨레는 기사의 양이 상대적으로 적기 때문에 합계 빈도수 20회 이상으로 정하였고 마찬가지로 이 기준에 부합하지 않는 항목 일부도 임의로 포함시켰다.



그림 빈도와 카이제곱 잔차를 고려해 추출한 키워드

**4.3. 연어분석하기**

이렇게 추출한 키워드를 통해 T 검정점수와 MI(Mutual Information, 상호의존정보) 점수를 통한 검증을 시도해 보았다. 카이스퀘어 잔차와 마찬가지로 T 검정 점수 또한 실제 통계와 직접적으로 관련이 있기보다는 이 연구에서는 그 방법을 차용했다. 단 상호의존정보 점수는 자연언어처리 같은 정보 이론을 많이 참조하는 분야에서 널리 사용하는 지표로 영어에서는 연어 분석을 하는 가장 일반적인 참조 기준이기도 하다. 다만, 한국어에서는 잘 동작하지 않는다는 의견이 주류인데, 이는 영어와 한국어의 어순 차이에 기반한 것이 아닌가로 추측된다. 예를 들어 동사의 경우는 필연적으로 주어 및 목적어에 대한 정보를 포함할 수밖에 없는데 한국어에서는 주어와 동사가 가장 멀리 떨어져 있고 따라서 주어 다음 거의 바로 동사가 이어져 한 단어가 다음 단어를 쉽게 예측할 수 있는 영어와는 달리 물리적 선후관계에 따른 확률 정보를 적용하는 것이 어렵기 때문이라는 생각이 일반적이다.

총 40개의 연어를 분석하였으므로 40개의 연어분석 결과표가 존재한다. 이 표에서 W1은 해당 연어 분석표의 작성 기준이 되는 키워드의 빈도를, W2는 키워드와 연어인지를 판단하기 위해 검사된 단어의 빈도를, W1W2는 두 단어가 일정한 범위 안에 같이 등장한 확률을, N은 이 단어들의 모집단이 되는 자료의 총 내용어 숫자를 의미한다. T 검정 점수와 상호의존정보 점수는 각각 t.score와 MI에 나와있다. 대상이 된 키워드는 좌측상단 가장 위에 나와 있다. 예시는 다음과 같다.

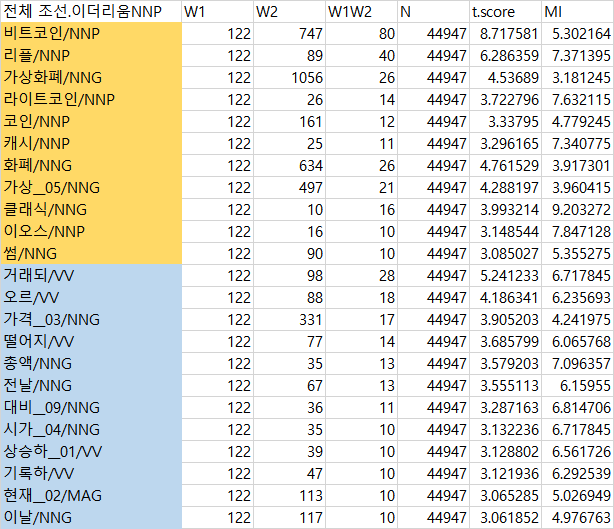


그림 이더리움과 각 단어 간 T 검정 점수와 상호의존정보.

위 자료에서 W1은 ‘이더리움’의 빈도수를, W2는 비교 단어의 빈도수, W1W2는 ‘이더리움’의 앞 뒤 6단어 중 비교 단어가 포함된 빈도, N은 전체 단어 개수를 의미한다. 위의 자료는 T 검정 점수가 3.00 이상인 연어만을 대상으로 한 자료다. 연어 분석에서는 앞서 말한대로 통계적 공기어를 먼저 추출하는 과정을 거치는데, 통계적 공기어란 일정한 범위를 정하여 그 안에 포함되는 모든 단어를 바이그램으로 만든 후 바이그램의 빈도를 살펴봄으로서 만들어진다. 예시를 들면 범위가 앞뒤 2인 경우, 정부/NNG + 비트코인/NNP + 규제/NNG + 파문/NNG + 일으키/VV 인 경우, 규제/NNG가 키워드였고 규제/NNG + 정부/NNG, 규제/NNG + 비트코인/NNP, 규제/NNG + 파문/NNG, 규제/NNG + 일으키/VV가 모두 집계되게 된다. 통계적 공기어의 핵심 아이디어는 이렇게 기계적으로 관련 없어 보이는 연어들을 추출하여도 실제 연어라면 많은 자료들을 대상으로 했을 때 해당 조합이 높은 빈도를 보일 것이라는 것이다. 표에서는 W1W2가 이렇게 일정한 범위 안에 해당 단어가 같이 등장한 조합의 개수를 뜻한다. 이 연구에서는 일정한 범위 혹은 스팬(span)을 6으로 상정하였다.

결국 이더리움이라는 가상화폐를 대상으로 조선일보의 보도에서 어떤 단어가 함께 쓰였는가를 나타낸 표인데 어렵지 않게 크게 두 부류로 기사를 나눌 수가 있었다. 예상되는 주제가 너무나도 광범위하지만 조선일보는 대체적으로 한 종류의 구체적인 암호화폐를 제시하면 다른 종류의 암호화폐를 어떤 의도에서 그랬는지는 추측만 할 뿐이지만 동시에 제시했음을 알 수 있으며, 두 번째는 훨씬 해석하기가 명확한데 흔히 주식시황에 대한 기사에서 자주 보이는 단어임을 보아 해당 암호화폐들의 시황을 나타냈음을 알 수 있다.

이 연구에서는 이것을 이 연구가 지양하는 정성적인 방법으로 분류하였지만, 이는 시간의 한계에 의한 것으로 이를 자동으로 분류할 수 있는 방법이 존재한다. 워드 임베딩 방식을 사용하는 것인데, 간단하게 말하면 워드 임베딩이란 각 단어들을 3차원의 축 어딘가에 위치시키는 기법인데 워드 임베딩 기법마다 구체적인 방법은 다르지만 3차원의 공간에서 비슷한 좌표에 찍히는 단어들은 대개 유사한 의미를 가지거나 유사한 맥락에서 사용되는 경우가 많게끔 설계가 되어 있다. 단어 사용이라는 것이 일상 생활에서 다르고 공적인 회의 상황에서 다르고 당연히 신문 기사에서 또 어쩌면 각 신문사마다 각 사람마다 다를 수 있으므로 이 연구에서는 조선일보에서 사용한 단어들을 가지고 워드 임베딩을 구현하면 가장 완벽하게 조선일보의 신문기사라는 조건 하에서의 단어들의 위치를 수리적으로 표현할 수 있게 된다. 수리적으로 표현이 된 단어들은 각각의 거리에 따라 유사도를 측정할 수 있게 되고, 따라서 같은 범주로 자동적으로 묶을 수 있게 된다. 하지만 워드 임베딩은 일반적으로 컴퓨터 장비의 높은 연산능력과 방대한 자료 그리고 연산에 필요한 긴 시간을 요구하기 때문에 이 연구에서는 조선일보의 워드 임베딩을 구현하지 못하였고, 더 나아가 한국어에 대한 워드 임베딩 결과물이 다소 구식이거나 정확하지 못해 연구에 적용하지 못하였다.

# 연구 결과

**5.1. 키워드와 연어로 본 조선일보와 한겨레의 보도 행태 차이**

연구 결과 암호화폐에 대한 조선일보와 한겨레의 차이는 존재했다. 가장 먼저 키워드에서 가장 눈에 띄는 부분은 암호화폐에 대한 용어의 차이다. 조선일보는 ‘가상화폐’라는 용어를 압도적으로 많이 사용하는 반면, 한겨레는 ‘가상통화’라는 용어를 주로 사용한다. 가상통화는 법정통화의 반대 개념으로 제도권에서 인정하는 통화가 아니라는 의미를 내포한다. 때문에 법무부에서는 가상통화를 규제한다고 명시하는 등 제도권에서는 가상통화를 널리 쓴다. 반면 가상화폐는 기술적인 측면을 강조한 용어로 암호화폐와 일맥 상통한 용어다. 이 연구에서는 앞서 말한 워드 임베딩 기법의 부재 때문에 전체 단어를 대상으로 했을 때의 내적을 구해 유사도를 측정하는 방법은 사용할 수 없어서 각 키워드와 그 연어를 정성적으로 살펴보았다.

잠시 내적을 통한 방법을 설명해보자면, 일정한 개수만큼 추출한 키워드들의 벡터값을 나열해 그 값들의 내적을 곱하는 것이다. 예를 들어 조선일보와 한겨레에서 2개씩을 추출했다면, 조선일보에 해당하는 점은 예를 들어 (‘가상화폐’의 워드 임베딩 값, ‘기업’의 워드 임베딩 값)이 될 것이고 한겨레에 해당하는 점은 (‘가상통화’의 워드 임베딩 값, ‘대책’의 워드 임베딩값)이 될 것이다. 이 두 점을, 혹은 이 두 행렬을 한쪽을 전치 하여 다른 한쪽에 곱하면 하나의 값이 형성되는데 (1 x 2 행렬과 2 x 1 행렬의 곱은 1 x 1 행렬) 이런 벡터의 내적은 두 값이 서로 어떤 상관이 있는지를 알려준다. 만약 10개씩의 단어를 곱한다면 키워드라는 특성을 가진 10개의 단어들의 워드 임베딩 값이 한 점을 이루고 있는 셈인데, 사람으로서는 상상이 불가능하지만 수학적으로는 각각 조선일보의 키워드들과 한겨레의 키워드들을 10차원 공간의 한 점으로 투영한 셈이고 이 점들이 원점과 이루는 직선이 서로 어떤 각도를 이루고 있는지, 일상적인 용어로 풀이하자면 어떤 상관관계를 가지고 있는지를 확인해볼 수 있을 것이다.

다시 본 연구로 돌아가면 키워드에서부터 극명하게 차이가 있던 조선일보의 ‘가상화폐’와 한겨레의 ‘가상통화’는 연관된 연어에서 같은 대상을 지칭하기에 당연히 상당수의 단어가 겹치지만 둘의 차이를 가장 극명하게 드러내는 지점도 보여준다. 다음이 두 키워드의 연어를 T 점수 기준 상위 20개를 추출한 것이다.

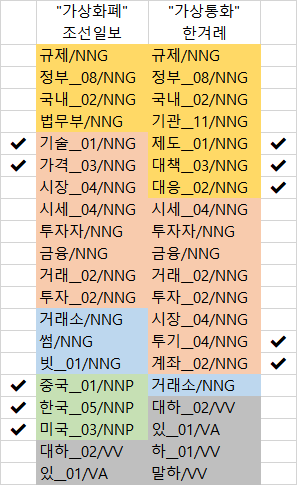
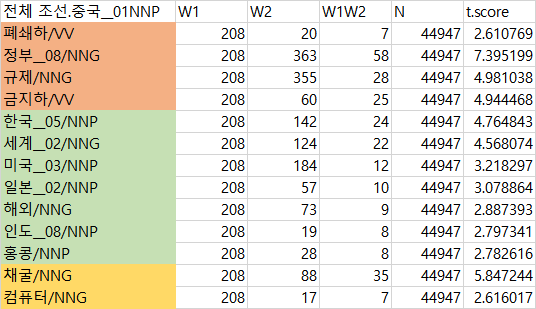


그림 조선일보와 한겨레 키워드 T검정 점수 상위 20개

상당 수의 단어가 서로 겹치는데 각각의 분류에 따라 같은 색깔로 표시를 해 두었다. 이 때 정부와 규제 관련한 조선일보와 한겨레의 연어가 다르다. ‘규제’ · ‘정부’ · ‘국내’는 당시의 핵심 주제였으니 두 언론 사의 보도가 해당 키워드를 공유하는 것이 당연하지만, 한겨레의 경우에는 조선일보에서는 보이지 않는 ‘제도’ ·’대책’ ·’대응’이 가상화폐에 대한 용어로 올라온 것이 눈에 띈다. 주황색으로 표시한 것은 민간에 대한 부분인데 조선일보가 ‘기술’이나 ‘가격’에 대한 연어가 많았던 반면 한겨레는 암호통화를 ‘투기’나 정부에서 규제를 하기로 결정한 ‘계좌’에 초점을 두었다. 또 눈에 띄는 점은 조선일보가 암호화폐의 세계적 흐름을 자주 보도했음을 나타내는 ‘중국’ ·’한국’ ·’미국’ 같은 나라 용어를 많이 사용한 반면에 한겨레는 이에 대한 언급이 없다. 이런 이유는 향후 분석 결과에서 잘 드러나는데 조선일보가 암호화폐를 정부의 입장보다는 시장의 입장에서 다루고 있어 시황과 해외에서의 현황을 주로 다루고 있는 반면, 한겨레는 암호화폐를 정부에 입장에서 다루며 주로 투기 같은 사회현상과 정부 차원의 규제를 중심으로 바라보고 있기 때문이다.





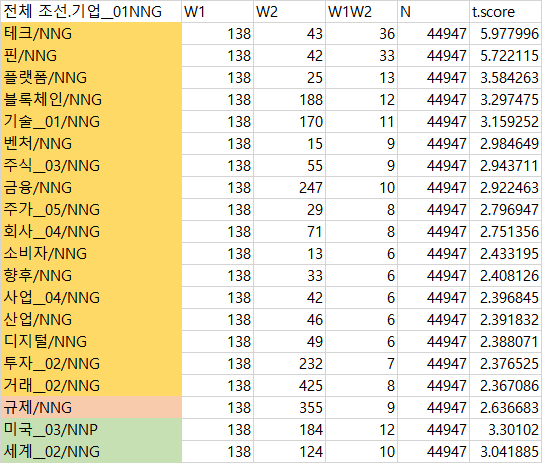
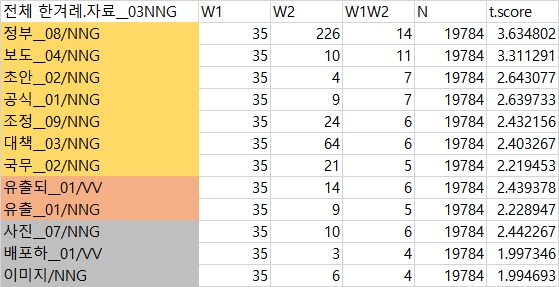
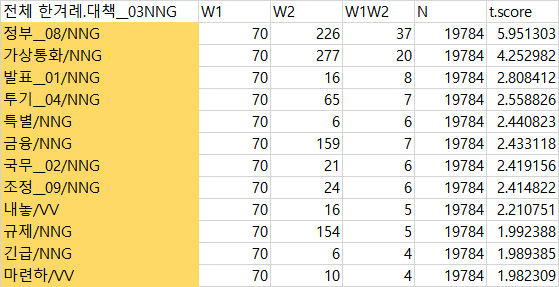


그림 ,7,8 ‘이더리움’, ‘중국’, ‘기업’ 와 각 단어의 T점수 (조선일보)

다른 키워드였던 ‘이더리움’ ·’중국’ ·’기업’을 보면 이런 경향을 손쉽게 파악할 수 있다. 해외의 상황을 전달하면서 ‘기업’ 키워드에서 보듯이 ‘핀테크’ ·’벤처’ ·’주식’ ·’산업’ ·’향후’ 같은 단어와 많이 연관을 짓고 있는 것을 알 수 있다. 암호화폐를 여러 경제적 투자 수단으로 보고 그에 관련한 정보를 제공하고 있다는 점이 확실하게 드러난다.





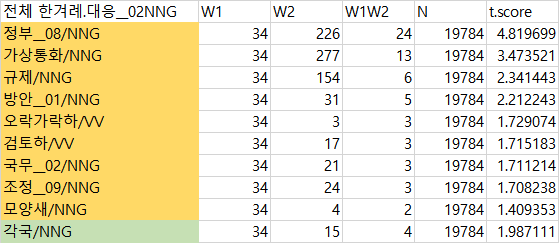


그림 9,10,11 ‘자료’, ‘대책’, ‘대응’ 와 각 단어의 T점수 (한겨레 신문)

반면 한겨레의 주요 키워드는 이와는 상당히 다르다. 일반명사라서 보통은 거의 아무 의미가 없는 ‘자료’ 같은 단어에서도 ‘정부’와 ‘대책’이 연어로 등장하며 ‘대책’에는 ‘투기’와 ‘긴급’ 같은 단어가, ‘대응’에는 ‘규제’나 ‘오락가락’ 같은 단어가 눈에 띈다. 조선일보의 키워드 중에는 ‘캡’이라는 단어가 있었는데 이는 ‘캡처’를 잘못 형태소 분석한 것으로 조선일보의 보도자료가 시중에 떠도는 혹은 ‘캡’과 다른 연어인 ‘코인마켓’ 같은 거래소의 정보를 인용한 것과는 대조되는 부분이다. 본 연구에서는 수치화 할 수 없는 것을 최대한 지양하기 때문에 이런 차이에 대해서는 일절 논하지 않겠지만, 정서분석이라는 자연언어처리 기법을 사용하면 이 차이점을 객관적으로 분석할 수 있어. 정서분석은 알고리즘을 사용한 혹은 사람에 의해 작성된 정서 사전을 가지고 특정 단어의 긍정값 혹은 부정값, 또는 각 감정에 대한 값을 부여하는 방식으로 정서 사전만 준비된다면 특정 단어, 특정 문장, 특정 글의 정서를 판단할 수 있게 해준다. 이 또한 조선일보와 한겨레의 고유의 정서 사전을 작성해야 하는 문제가 있기 때문에 본 연구에서는 다룰 수 없었지만, 정서분석 방법을 적용해보면 이 둘의 차이가 확연하게 드러날 만한 연어에서의 차이라고 생각한다.

**5.2. 시간에 따라 달라진 암호화폐에 대한 용어 사용**

사실 두 조선일보의 ‘가상화폐’와 한겨레의 ‘가상통화’는 암호화폐에 대한 시각차를 드러내는 용어다. 이는 정부의 입장문에 잘 드러나 있는데 대한민국 정부는 암호화폐를 공식적으로 ‘가상통화’로 명명하고 있다. 2019년도부터는 ‘가상증표’라고 말하기로 했는데 당연히 비트코인과 이더리움 등은 돈이 아니라는 의미를 나타내는 용어다. 동의어처럼 보이는 ‘가상화폐’와 ‘가상통화’에서의 차이는 ‘가상화폐’는 기술적 발명품으로 등장해 기존의 화폐를 대체할 대체제나 일반적인 시장에서 가장 널리 통용되는 말이고, ‘가상통화’는 실물통화에 비추어 유사 통화 혹은 가짜 통화의 의미를 내포하는 쪽에서 주로 사용하는 용어다. 조선일보와 한겨레의 보도 행태에서의 차이가 이 점과 무관하지 않을 것이다.

그런데 조선일보와 한겨레의 처음 20%의 기사와 마지막 20%의 기사에서 제일 눈에 띄는 부분은 바로 암호화폐를 부르는 명칭의 차이다.

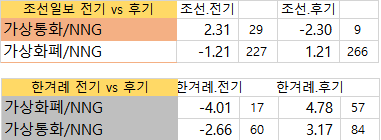


그림 12 가상통화, 가상화폐 에 대한 각 언론사의 빈도 변화

실제로 조선일보에서도 보도 초창기에는 조선일보의 전체적인 경향과는 맞지 않게 ‘가상통화’라는 용어를 비교적 많이 사용하였다. 조선일보가 ‘가상화폐’라는 용어를 사용함에 있어서는 유의미한 차이가 드러나지 않은 것과는 반대로다. 조선일보의 초창기 보도에서 ‘가상통화’에 대한 연어는 한겨레의 것과 차이가 적다. 이는 아래의 표를 보면 잘 드러난다.

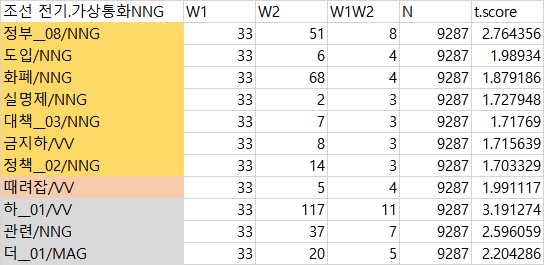


그림 13 조선일보 전기 ‘가상통화’ T검정

한겨레와 마찬가지로 정부의 대책에 관련한 연어들이 많이 눈에 띈다. 이 시기의 조선일보 기사에서는 정부가 규제가 필요함을 역설했다는 보도를 한겨레만큼 많이 냈다. 조선일보가 한겨레와 한 가지 다른 점이라면 ‘때려잡/VV’이라는 연어에서 드러나는데 이 단어가 연어로 선정된 이유는 하태경 의원이 ‘빨갱이들 때려잡는다고 해서 좌파가 늘지 않고 줄었다. 비트코인도 똑같다’라는 요지의 규제 반대 발언을 ‘한편’이라는 표지 뒤에 언급했기 때문이다.

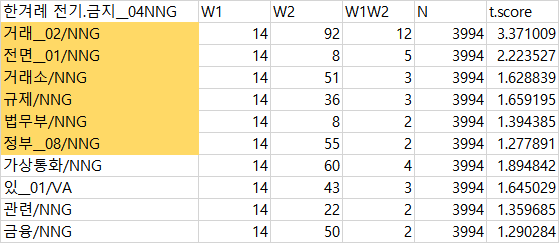


그림 14 한겨레 전기 ‘금지’ T검정

한겨레의 관련 보도에서는 이와 관련된 연어가 전혀 드러나지 않는다. 분명히 그때도 하태경 의원은 해당 발언을 기자들 앞에서 했을 것인데 한겨레는 이를 보도하지 않은 것이다.

하지만 그렇다고 이때 한겨레의 입장이 명확했던 것은 아니었다는 증거가 있다. 바로 한겨레가 보도 초창기에 다룬 자료에 관한 연어 표인데 한겨레로서는 이후 한겨레의 보도 키워드를 생각해보면 상당히 이례적으로 국제 시장에서의 비트코인에 대해 다루고 있다.

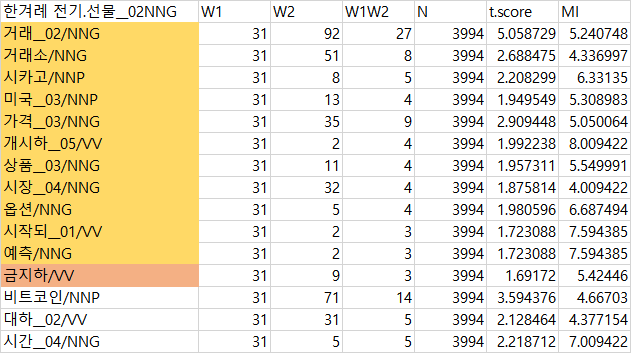


그림 15 한겨레 전기 ‘선물’ T검정

조선일보의 가상화폐 관련 기사에서 흔하게 볼 수 있는 연어들이 모두 등장한다. 바로 거래에 관련된 용어들로 해당 보도들의 내용은 시카고 선물 시장에서 비트코인이 거래되기 시작하였으나 한국에는 금융상품으로서는 금지되어 거래할 수 없다는 것이 요지다. 후반기와는 대조되는 조선일보와 한겨레의 보도 행태를 생각해보면 이 시기에는 중반기와 후반기만큼 확고한 입장이 아직 정해지지 않았지 않았을까 추론하게 만드는 대목이다.

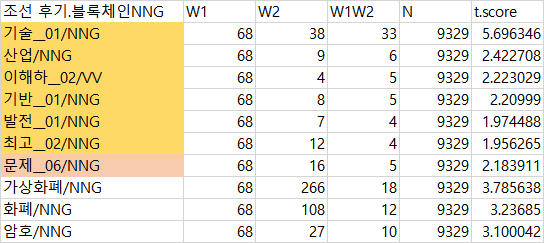


그림 16 조선 후기 ‘블록체인’ T검정

조선일보의 후기 보도 행태는 위의 연어 표가 가장 잘 드러내 줄 것이다. ‘산업’ ·’기반’ ·’기술’ ·’이해하다’ ·’발전’ 같은 단어뿐만 아니라 ‘최고’까지 연어로 등장한다. 심지어 여기서 언급된 ‘문제’도 블록체인 알고리즘에 대한 설명 중 ‘수학 문제를 풀어야 한다’나 그간 블록체인이 구현되지 않았던 이유는 ‘큰 컴퓨팅 파워가 필요하다는 문제’ 같은 문장에서 유래한 표현이다.

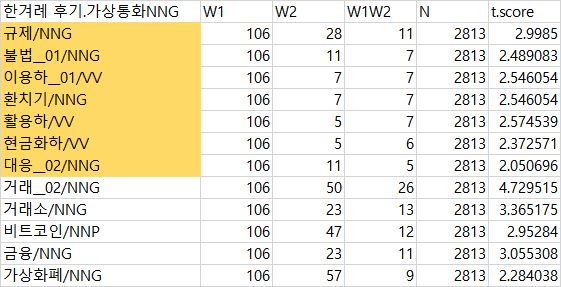


그림 17 한겨레 후기 ‘가상통화’ T검정

한겨레의 후기 보도 행태는 ‘가상통화’와 관련된 연어 그 자체에서 아주 잘 나타난다. ‘불법’ ·’환치기’ ·’현금화’ ·’활용하다’ 같은 범죄에 관련된 단어에다, 이에 대한 ‘대응’과 ‘규제’ 등 조선일보 후기 보도의 단어선택만큼이나 연관된 연어가 노골적이다.

이는 사회심리학에서의 그룹 양극화와 잘 맞아떨어지는 현상이다. 비트코인을 위시한 암호화폐는 기본적으로 복잡한 개념이다. 컴퓨터와 수학 그리고 금융과 화폐가 관련된 복잡한 문제이기 때문에 근본적으로 기자라는 비전문가 집단인 언론사에서 입장을 형성하기는 매우 힘들다. 각 신문사가 기본적으로 가지고 있는 모종의 성향 같은 게 존재해 초반과 후반기 모두에서 공통적으로 드러나는 보도 행태가 있으면서도 초창기까지는 두 신문사의 기사가 실질적으로는 차이가 나지 않는 이유도 암호화폐라는 개념이 가지는 난해함이 요인으로 작용했을 것이다. 이는 각 키워드의 연어를 살펴보았을 때 초창기에는 두 언론사 간에 눈에 띄는 차이가 존재하지 않았던 것으로 관찰 가능하다. 그리고 후기에는 암호화폐에 대한 용어의 사용의 늘어남과 줄어듦 그리고 여러 키워드들에 연관된 연어들의 종류를 보고 적어도 두 신문사의 언어사용이 다르다는 점을, 추론컨데 보도 초점이 다르다는 것을 살펴볼 수 있다.

그런데 한 가지 눈에 띄는 이상현상도 관찰된다. 바로 한겨레 후기 보도에서 ‘가상화폐’에 대한 보도가 늘었다는 점이다. 중반까지 기사 1% 당 0.85번가량 언급되던 ‘가상화폐’ 단어가 후반기 보도에서는 기사 1% 당 2.85번으로 급격하게 증가한다. 그리고 아래 표에서 보듯이 ‘가상화폐’에 대한 연어가 조선일보의 일반적인 보도에서의 연어와 구분이 구별이 가지 않을 정도로 비슷하다. 그렇다고 한겨레의 ‘가상통화’를 사용한 보도가 줄어든 것은 또 아니다. 오히려 초창기 1% 당 3.2번이었던 보도가 후반기에는 1%당 4.2번으로 늘어난다. 조선일보가 ‘가상통화’라는 용어를 초창기에만 기사 1%당 1.45번 쓰다가 이후로는 0.45번 쓰고 ‘가상화폐’라는 단어를 꾸준히 기사 1% 당 14번가량 썼던 것과 대조적인 지점이다. 한겨레의 전기와 후기를 대조하는 키워드들에 연관된 연어들은 그 종류가 ‘불법’이나 ‘환치기’ 같은 단어로 옮겨가 오히려 보다 강경해졌다고 볼 수 있는 반면 그런 경향을 유지하면서도 새로운 유형의 보도 행태가 한겨레에 나타난 것이다.

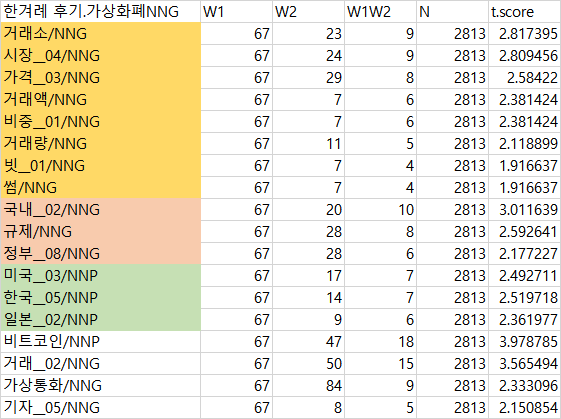


그림 18 한겨레 후기 ‘가상화폐’ T검정

이를 한겨레의 보도 행태가 청년 구독층의 영향을 받아 변화했든지 혹은 이미 비트코인 가격이 2월 말에는 천만 원 정도 떨어진 상태이기 때문에 시황 보도를 해도 된다고 생각한 것인지 아니면 다양한 종류의 보도를 하지 않은 신문사의 작태에 대한 반성으로 이런 기사를 냈는지는 혹은 모종의 사내 정치 때문이지는 알 수가 없다. 겉으로 드러나는 기사만으로는 이러한 기사가 형성되게 된 원리에 대해 파악하는 것에는 한계가 있다.

1. Ho, Becker, Moy, Binder (2011). The Role of Perceptions of Media Bias in General and Issue-Specific Political Participation, Mass Communication & Society 14(3): 343 - 374 [↑](#footnote-ref-1)
2. 최선규, 유수정, 양성은(2012). ‘뉴스 시장의 경쟁과 미디어 편향성: 취재원 인용을 중심으로’ [↑](#footnote-ref-2)
3. Aronson, Elliot (2010). Social Psychology. Upper Saddle River, NJ: Prentice Hall. p. 273. [↑](#footnote-ref-3)
4. 조선일보 홈페이지 URL: <http://www.chosun.com/>

   한겨레 신문 홈페이지 URL: http://www.hani.co.kr/ [↑](#footnote-ref-4)