디지털 사회
DIGITAL SOCIETY
제15호

연세대학교 사회과학데이터혁신연구센터
SOCIAL SCIENCE DATA INNOVATION CENTER

발 행 인: 조화순
발 행 일: 2018년 4월 1일
홈페이지: http://democracy3.0.yonsei.ac.kr
ISSN 2586-3525(Online)

단어를 포착한다는 것

김정연
연세대학교 국가관리연구원

사회과학 연구에서 텍스트는 중요한 데이터 자원이다. 데이터로서의 텍스트 연구는 정치학 분야의 경우 텍스트가 가지는 이데올로기, 이념적 정향을 판단하기 위해 수행되었다. 혹은 특정 텍스트들에 대해 수용자의 반응성을 조사하는 등 텍스트를 둘러싼 수용자의 행태 관찰이 주로 시행되었다. 즉, 텍스트가 집중하는 것을 다루거나 텍스트를 소비하는 수용자의 특징을 기술하고 해석하는 연구들이 시도되었다. 텍스트를 구성하는 구조로서의 단어는 실제로 정치의 요체가 될 수 있다. 단어가 가지는 중요성은 정치의 핵심적인 부분으로 역할 할 수 있다. 정치인, 공직자, 국민들은 의견을 표출하고, 행위 하는데 단어를 사용한다. 이들의 단어는 의견의 제안과 설득, 방어의 과정에서 드러난다.

단어는 사람들이 집중하는 것을 따라가고, 사람들이 인식하는 것, 인식하도록 하는 것들을 규정한다. 법과 규제 역시 단어를 통해 조정된다. 이에 사회과학자는 언제나 단어에 관심을 가져왔다. 정치인은 대중과 커뮤니케이션하는 과정에서 리더십을 강조하고, 유권자의 선택을 얻기 위해 '말'을 도구로 사용한다. '수사적' 리더십은 정치인이 갖춰야 할 자질이기도 하며 유권자의 입장에서 정치인을 평가하기 위한 근거이기도 하다. 소셜 네트워크 서비스가 발전 하면서 자연스럽게 다양한 '말'들이 저장, 축적되고 퍼져나간다. 네트워크를 통한 사람들의 연결성의 확대는 자연스럽게 사람들의 관계의 확장으로 이어지고 그 과정에서 사람들과 밀착된 정보들이 생성된다.

사람들은 도처에서 단어를 통해 자신이 누구인지, 어떤 생각을 가지고 어떻게 삶을 영위하는지 드러낸다. 단어를 통해 우리는 단어를 표현하는 개인 혹은 집단이 어떤 존재인지 파악한다. 오프라인과 온라인의, 한정성에 제약받지 않는 공간에서 생성되는 단어는 곧 그것을 생성하는 존재에 대한 흔적이 된다. 단어의 흔적들을 모아 분석하면 어떤 시간, 어떤 상황, 어떤 장소든 단어를 생성한 존재의 모습을 추정할 수 있다. 최근의 텍스트 연구들은 이러한 전제에서 텍스트를 생성한 사람들의 정체성 혹은 특성을 밝히고자 해왔다. 범람하는 데이터를 활용하여 많은 연구들이 진행되어 왔는데, 예를 들어 경영학에서는 기업, 제품에 대한 선호도에 대한 관심이 주로 분석되어졌다. 브랜드가 가지는 가치, 기술이 어떻게 대중들에게 인식되는지 알기 위해서이다.

정치 빅데이터 역시 정치적 이벤트에 대한 여론의 관심을 파악하고, 정치행위자들의 가치 지향성과 욕구가 어떻게 발현되는지 알아보는데 활용되고 있다. 정치 행위자들의 공식, 비공식적인 표현들이 축적되고 있다. 연구자는 입법 연구를 위해, 입법가들의 스피치를 체계적으로 수집할 수 있고, 법·제도의 수정과정을 관찰할 수 있다. 정부부처와 기관은 웹에 실무 내용을 보고하고 정책 제안과 합의 과정을 단계별로 비교할 수 있도록 하고 있으며, 공중의 반응성을 수집한다. 대부분의 언론사는 아카이브

에 그들의 기사들을 제공하고 있다. 인터넷 아카이브는 연구자로 하여금 데이터 사용을 독려하도록 한다. 최근 각종 데이터에 접근 가능한 도구들이 개발되면서 다양한 분야의 전문가들이 텍스트 분석을 시도하였다. 텍스트를 계산하는 혁신의 능력은 사회과학 연구를 진화시킬 것이다.

우리는 그동안 정치인의 언어적 스타일이나 언어를 통한 정치인 행위의 특징을 알고 싶을 때 정치적 텍스트를 주목하였다. 정치인이 무엇을 말하는지, 어떻게 말하고 있는지에 대해 관심을 가져왔다. 정치인의 레토릭 연구는 의회 제도나 대통령 제도의 표현 양식으로서 접근되었다. 정치인의 레토릭은 대중에게 정부와 의회의 정책 기조를 설명하고 국민들에게 호소하는 기능이 있다. 정치인의 스피치는 선거 시기와 국정운영 시기에 각기 다른 상황에서 다양한 연설을 통해 공적 담론을 형성하거나 개인적 웹 공간에서 의견을 표현하는 방식으로 이미지를 형성하도록 한다. 정치인의 스피치를 분석해 미래 정치 방향을 예측하고 위기 상황에서 대응하거나 정치적 현 상황을 둘러싼 인식을 파악할 수 있다.

텍스트 분석의 방식은 텍스트의 구조를 해체해 텍스트가 가지는 의미를 요약하는 것이다. 텍스트 전체의 구조, 문단, 문장의 맥락에서 의미를 발견할 수도 있고, 텍스트 이면에 감추어진 의미를 유추할 수도 있다. 텍스트가 드러내고 있는 의미 파악에 초점을 맞추는 경우 내용적 측면 뿐 아니라 텍스트를 생성한 행위자 혹은 텍스트에 드러난 정서, 심리 상태를 해석할 수 있다. 텍스트끼리의 관계성 역시 살펴볼 수 있다. 그동안 우리는 문장에서 내용어 혹은 기능어를 구분해 텍스트에서 전달하고자 하는 내용을 분류했다. 내용어는 명사, 동사, 형용사 등 문장의 내용을 본질적으로 결정하는 의식적 표현이고, 기능어는 접속사, 조사, 관사 등 무의식적으로 발현되는, 혹은 문법적으로 형식상 사용되는 표현이다. 텍스트에서 포착하고자 하는 목적에 따라 단어의 생성 기제의 차이를 경험할 수 있다.

텍스트에서 품사가 역할 하는 기능이 다르기 때문에 한국의 경우 미국의 연구들과는 다른 접근 방식에서, 언어의 차이를 세심하게 반영하여 수행되는 것이 중요하다. 명사나 동사, 형용사의 짧은 단어들은 우리가 무엇에 관심이 있으며 무엇에 주의를 기울이는지 정보를 도출하거나 우리가 어떻게 생각하는지 감정을 보여주도록 한다. 그런데 한국어의 경우 동사의 예를 들 때, 영어처럼 한 가지 단어에서 어미가 붙어 형태소가 변형된 것이 아닌 서술어가 독립되어 있다. 이에 한국어 분석에서는 형태소가 분리되는 것을 강조하게 된다. 또한, 텍스트 데이터를 다룰 때 문어나 구어, 준구어들을 분석하는 적절한 어휘 사전 역시 필요하다.

컴퓨터를 기반으로 한 텍스트 분석 시스템들은 1960년대 제너럴 인콰이어러(General Inquirer) 소프트웨어와 같은 프로그램부터, 정치학 분야에서 가장 많이 인용되고 있는 로드릭 하트(Roderick P. Hart)의 딕션(Diction) 등 기술적으로 발전을 거듭해왔다. 이러한 소프트웨어들에서는 기본적으로 단어의 출현 빈도를 계산하는 것이 중요하다. 텍스트에 출현된 단어들을 분석자의 분석 목적에 부합하도록 유목을 분류해 비교하여 의미를 요약한다. 펜네베이커(James W. Pennebaker)의 LIWC는 단어를 통해 발화자의 감정에 주목한다. 컴퓨터 기반 텍스트 분석 프로그램들을 사용한 결과를 사람이 수동으로 분석할 경우에도 분석 결과가 일치할 때 효용성이 높다고 판단한 보고들이 있다.

텍스트 분석 방법들은 꾸준히 개발되고 있다. 최근에는 텍스트에 내재된 주제를 추출하거나 텍스트가 가지는 이념적 경도를 위치시키는 것에 관심이 있다. 텍스트 분석의 원천이 되는 기반 자료와 알고리즘 개발이 중요하다. 우리는 학문적 분야 뿐 아니라 실무적 필요성에서도, 단어를 포착하여 정치 빅데이터의 의미론적 요약을 시도하는 것이 의미를 가질 수 있다. 현 시대를 살아가는 사람들의 생각의 기저를 파악하는 작업은 정부, 정치권에서는 미래 전략을 수립하고, 국민의 입장에서는 정치 행위자, 제도를 평가하여 다양한 정보의 처리 수준을 확장하는 의미를 가진다. 사람들은 어떤 단어를 좋아하는가? 사람들의 단어가 사람들의 선택과 결과에 어떠한 영향을 주고 있는가? 사람들의 행위나 작용을 이해하기 위한 이러한 질문들에 답을 구하는 과정에서 발전을 거듭하는 연구 기법들의 활용과 연구자들의 확장된 시각이 요구된다.