빅데이터 경진대회 보고서

20181414 서석주

<서론-주제 선정 배경>

최근 미국에서 대통령 선거가 진행되었다. 개표 후반까지 당선인을 예측하기 쉽지 않았다. 뉴스와 언론에서는 그래도 트럼프 후보가 당선될 것이라는 의견이 지배적이었다. 그러던 도중 우연히 구글 빅데이터가 구글 트렌드 분석을 통해 '조 바이든' 후보의 당선을 예측했다는 기사[1]를 읽었다. 그리고 실제로 이 예측은 맞아떨어졌다. 빅데이터를 통해 선거 결과를 예측할 수 있고 실제로 유의미하다고 생각했다. 그래서 우리는 16, 17 18, 19대 대선 기간 동안 보도된 자료들은 분석하고 분석결과를 통해 당선인을 미리 예측할 수 있었는지에 대해 연구해보게 되었다. 선거 개표 날 포함 2주 동안의 기사 400개[2]를 추출하였고, 언론사의 정치적 성향에 편향되지 않은 데이터 추출을 위해 우파 성향의 기사와 좌파 성향의 기사의 비율은 동일하게 두었다.

<본론-코드 설명&분석>

R을 통해 텍스트 마이닝을 진행하였다. 사용한 패키지는 multilinguer, RSQLite, tm, httr, XML, stringr, KoNLP, dplyr, wordcloud, wordcloud2, plotrix, arules, igraph이다. 먼저 xpathSApply함수를 통해 크롤링을 진행했다. 다음으로 gsub함수를 통해 영문자 및 한자, 숫자, 특수문자, 공백 등을 지워주며 전처리를 진행했다. 그 후 extractNoun함수로 명사를 추출했고, Corpus, VectorSource, TermDocumentMatrix 함수를 사용하여 4~16자리의 명사 말뭉치를 생성했다. 생성한 명사 말뭉치 데이터프레임을 wordcloud함수를 통해 시각화했다. 그러나 가시성이 떨어진다고 생각하여 wordcloud2함수를 통해 다시 시각화했다. 만족스러운 결과를 얻었다. 또한 다양한 형태로 시각화하면 좋겠다고 생각하여 pie함수를 통해 시각화했다. 그러나 이것 또한 가시성이 떨어진다고 생각하여 pie3D함수를 통해 3D 파이차트를 만들었다. 만족스러운 결과를 얻었다. 마지막으로 apriori함수로 단어 간 연관 규칙을 산출하여 연관어 분석을 진행했다. 연관어 분석 결과를 igraph를 사용하여 시각화했다.

후보자의 경우 '이명박'이 압도적으로 많은 비율을 차지하고 있음을 알 수 있다. 그 뒤를 이어 '정동영', '이회창'이 비슷한 비율을 차지하고 있다. 그 다음으로 '문국현', '권영길'이 적은 비율을 차지하고 있다. 후보자 키워드의 빈도 수가 득표율과 비례하고 있음을 알 수 있다. 다음으로

---

[1] https://www.ajunews.com/view/20201103212144863

[2] 각 대선 기간 별 100개씩 기사를 추출하였다.

[여기에 입력]

'특검', '수사', '검찰', '비비케이' 등이 높은 비율을 차지하고 있다. 이 키워드들은 17대 대선 당시 주요 사건이었던 이명박 후보의 BBK 주가 조작 논란과 관련이 있다. 대통합민주신당에서는 이명박 특검법[3]을 통과시켜 이 후보의 BBK 주가조작 의혹 등 증권거래법 위반 혐의 및 검찰의 피의자 회유, 협박 등 편파왜곡 수사 및 축소발표 의혹 등에 대해 특검이 재수사 해야 한다고 주장하며 BBK 의혹에 대한 '네거티브 전략'을 내세웠다. '경제'도 높은 비율을 차지하고 있다. 당시 미국의 서브프라임 모기지 사태로 인해 전 세계에 경제 위기가 찾아왔다. 그로 인해 경제 위기 극복이 전 국민의 주요 관심사였고, 경제 관련 정책 및 스탠스 등이 대선 결과를 판가름 지을 정도의 영향력을 갖고 있었다. 결국 '경제 대통령'의 이미지를 강조하던 이명박 후보가 당선되었다.

연관어 분석을 진행한 결과 어느 정도의 연관성[4]을 가지는 키워드 묶음은 '이명박'을 필두로 '검찰', '한나라당', '지시' 등이 있었다.

<결론>

처음 진행하는 텍스트 마이닝 작업이라 전처리 과정 및 연관어 분석에서 미숙한 결과를 보여준 것이 아쉽다. 그럼에도 불구하고 빈도 수가 높은 키워드들을 통해 대선 기간의 흐름 및 결과를 잘 예측할 수 있다는 결론을 내렸다는 점에서 긍정적이다. 특히 대선 후보 이름의 빈도 수와 득표율이 비례한다는 사실을 도출한 것에 있어서 상당히 긍정적이다. 우리의 가설이 틀리지 않았음을 보여주기 때문이다. 결론적으로 빅데이터를 통해 선거 결과를 예측할 수 있다고 생각한다.

---

[3] 정식 명칭은 '한나라당 대통령 후보 이명박의 주가조작 등 범죄혐의의 진상규명을 위한 특별검사의 임명 등에 관한 법률안'이다.

[4] apriori의 parameter를 support=0.55, conf=0.55로 지정했다.

[여기에 입력]