텍스트 데이터 분석

젤렌스키 우크라이나 대통령 연설문

텍스트 데이터 분석 및 사례 연구

2023.03.30

B2 팀

서영석, 박용태, 이현호, 전국림

<목차>

참고 자료

뮌헨안보회의(MSC) 우크라이나 대통령연설문

1. 서론

1) 데이터 배경

2022 년 2 월 19 일 독일 뮌헨에서 열린 제 58 차 뮌헨안보회의(MSC)에서 우크라이나 대통령인 젤렌스키는 우크라이나와 러시아 간의 군사 충돌을 배경으로 연설을 발표하였다. 연설에서 젤렌스키 대통령은 러시아의 침략적인 행동과 국제사회의 대응에 대해 강하게 비판하였다.

또한 젤렌스키 대통령은 국제사회의 대응에 대해서도 언급했다. 우크라이나뿐 아니라 전 세계적인 안보 문제이고 국제사회는 우크라이나를 지원하고 러시아의 침략을 규탄해야 한다고 하였다.

2) 데이터 분석 설명

이 보고서에서는 R 을 활용하여 연설문을 최소 2 음절단위 단어로 분리하고 그 단어들 빈도수와 서로간 연관어 들을 분석하여 주요 단어들과 연관성을 파악 하려고 한다.

2. 본론

1) 데이터 로딩 및 탐색

먼저 젤렌스키 대통령의 연설문을 로딩하고 zelensky_data 변수 담아 두었다.

```
#(장문)우크라이나 젤렌스키 대통령의 연설문 파일 읽기
zelensky <- file("Zelensky_address_20220219.txt", encoding = "UTF-8")
zelensky_data <- readLines(zelensky)

#분석 라이브러리
library(KoNLP)
library(tm)
library(multilinguer)
library(wordcloud2)
library(stringr)
library(tidytext)
```

2) 텍스트 토픽 분석

- 말뭉치 생성 및 한글 불용어 리스트 생성

```
#말뭉치 생성
myCorpus <-Corpus(VectorSource(zelensky_nouns))
my_stopwords <- c("은", "는", "이", "가", "하다", "것", "들", "그", "되다", "이다",
                  "보다", "않다", "하다", "되었다", "있다", "같다", "때문", "말하다",
                  "그러나", "그렇다", "그것", "이렇다", "저렇다", "하지만", "다른",
                  "어떤", "여러", "싶다", "받다", "모르다", "중", "좀", "잘", "더", "말다",
                "그리고","너무","아니다","없다","국가이든","이후","이것","하기","여러분",
                "때문", "이것", "그것", "들이", "해서","무엇", "저들", "이번", "우린", "우리")
```

말뭉치 생성 후, 문장부호 제거, 수치 제거, 소문자 변경, 불용어 제거 순으로 전처리를 진행하고 불용어는 영어 이외 한글에서 자주 나오는 것들을 따로 리스트 생성하였다. "때문", "이것", "그것", "들이", "해서", "무엇", "저들", "이번", "우린", "우리" 등등

- 불용어 제거(전처리) 실행

```
#문장부호 제거
myCorpusPrepro <- tm_map(myCorpus, removePunctuation)
#수치 제거
myCorpusPrepro <- tm_map(myCorpusPrepro, removeNumbers)
#소문자 변경
myCorpusPrepro <- tm_map(myCorpusPrepro, tolower)
#불용어 제거
myCorpusPrepro<tm_map(myCorpusPrepro,removeWords,c(stopwords('english'),my_stopwords))
```

- 말뭉치에서 단어 수집

```
#단어수집 2 글자이상 10 글자 이하
myCorpusPrepro_term<TermDocumentMatrix(myCorpusPrepro,control=list(wordLengths =
                  c(4,20)))
#matrix 자료구조를 data.frame 자료구조로 변경
myTerm_df <- as.data.frame(as.matrix(myCorpusPrepro_term))
#단어 출현 빈도수 구하기
wordResult <- sort(rowSums(myTerm_df), decreasing = TRUE)
```

```
#단어길이 2 이상 빈도수와 빈도수 1 예외처리
my_filter <- wordResult[nchar(names(wordResult)) >= 2]
my_filter <- subset(my_filter,my_filter[] != 1)
#단어 열 생성
myName <- names(my_filter)
#단어열에 빈도수 추가한 데이터 프레임 생성성
word.df <- data.frame(word = myName, freq = my_filter)
```

전처리 된 말뭉치에서 2 음절 단어와 10 음절 이하의 단어를 추출하였다. 외국어 번역을 감안하여 단어 길이를 제한을 10 음절 정하였다. 그 다음 단어들의 출현 빈도수를 구하고 빈도수가 1 인 단어들은 예외처리 하였다.

아래표는 빈도수를 내림차순으로 정렬한 단어 중 일부이다

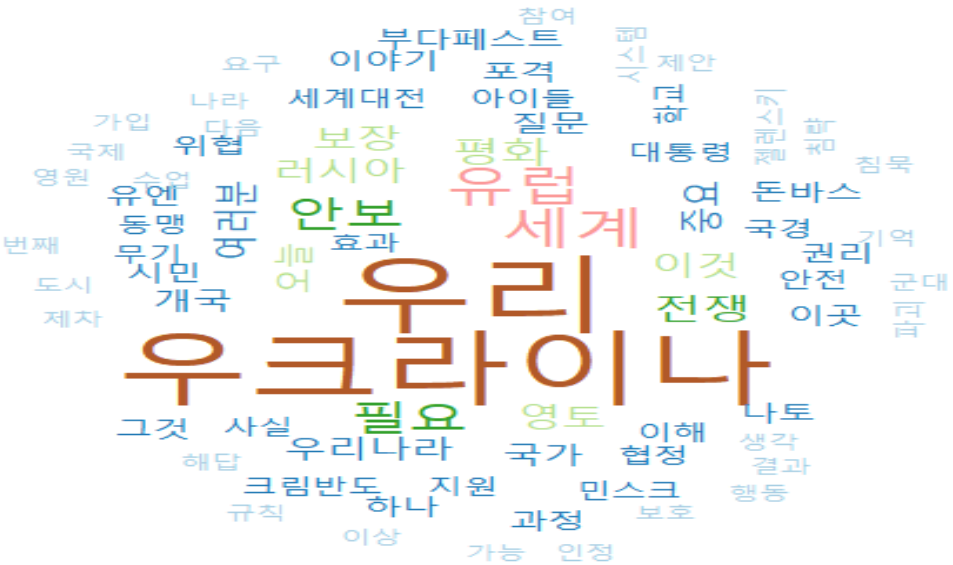
단어	빈도수	단어	빈도수
우크라이나	45	러시아	8
세계	19	오늘	8
안보	13	보장	8
필요	13	중요	7
전쟁	12	크림	7
평화	11	우리나라	7
유럽	9	질문	6
영토	8	국가	6

-wordcloud2 를 이용한 시각화

```
#wordcloud2 라이브러리 로딩
library(devtools)
library(wordcloud2)

#빈도수 그래프 그리기
wordcloud2(data = word.df,
           size = 0.8, color = 'random-light', gridSize = 9
           ,backgroundColor="black"
           , maxRotation = 0.5, minRotation = 5,shape = "star")
```

<Wordcloud2 를 이용한 시각화 그림 (star)>



<Wordcloud2 를 이용한 시각화 그림 (circle)>

-텍스트 수집 결과 분석

시각화 결과, "우크라이나"가 최빈 단어로 도출되었다. 이어서 "안보", "전쟁", "평화", "세계" 등 단어들이 많이 도출되고 있다. 이는 우크라이나 대통령이 세계평화와 안보를 무시하는 러시아에 대한 강한 태도를 담은 연설임을 알 수 있다.

3) 텍스트 연관어 분석

-연관어 분석 데이터 전처리

```
#줄 단위 단어 추출
lword <- Map(extractNoun, zelensky_data)
length(lword)
lword <- unique(lword)
#중복 단어 제거와 추출 단어 확인
lword <- sapply(lword, unique)
length(lword)
#연관어 분석을 위한 전처리하기
#단어 필터링 함수 정의
#2 글자 이상 16 글자 이하 단어 필터 함수 생성
filter1 <- function(x) {
      nchar(x) <= 16 &&nchar(x) >= 2 && is.hangul(x)
  }
filter2 <- function(x) { Filter(filter1, x) }
#줄 단위로 추출된 단어 전처리
lword <- sapply(lword, filter2)
lword
#트랜잭션 생성하기
library(arules)
wordtran <- as(lword, "transactions")
wordtran
```

먼저 줄단위로 단어들을 읽어오고 unique 함수를 이용하여 중복된 단어들을 제거해준다

중복 제거전 length(lword) -> 211 제거후 length(lword) -> 56

-연관규칙 확인

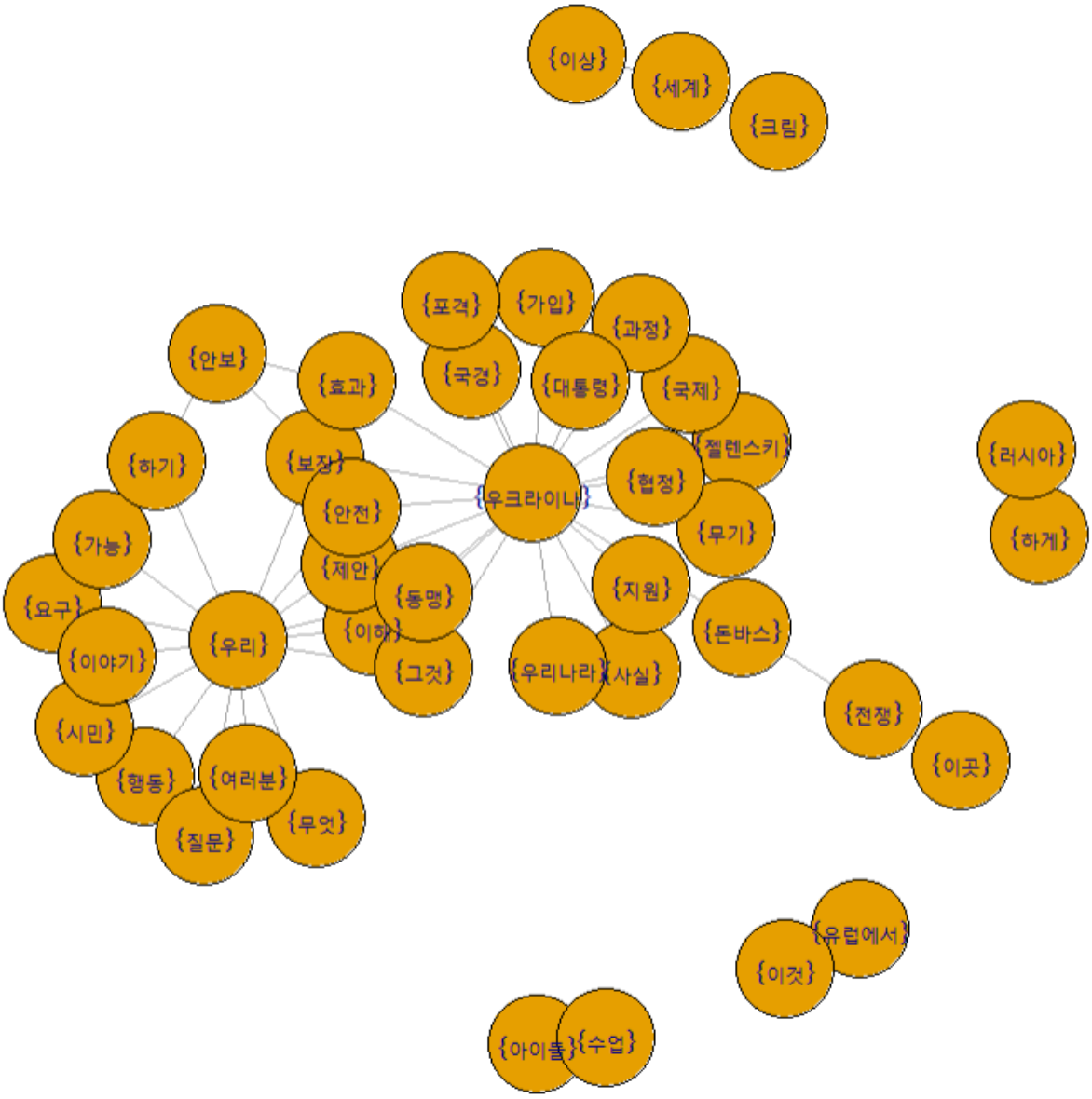
```
#트랜잭션 생성하기
library(arules)
wordtran <- as(lword, "transactions")
#단어 간 연관규칙 발견하기
library(backports)
#연관규칙 발견
tranrules <- apriori(wordtran, parameter = list(supp = 0.05, conf = 0.6))
#연관규칙 생성 결과보기
detach(package:tm, unload=TRUE)
inspect(tranrules)
```

생성된 트랜잭션을 기반으로 Sup 지수(지지도)가 낮아질수록 연관규칙이 많아져 분석에 적당한 Sup(지지도)값 0.05, conf(신뢰도) 값을 0.6 으로 조정하여 총 92 개의 연관규칙을 얻게 된다, .

-연관규칙 시각화

```
#연관어 시각화하기
#연관단어 시각화를 위해서 자료구조 변경
rules <- labels(tranrules, ruleSep = " ")
# 문자열로 묶인 연관 단어를 행렬구조로 변경
rules <- sapply(rules, strsplit, " ", USE.NAMES = F)
# 행 단위로 묶어서 matrix 로 변환
rulemat <- do.call("rbind", rules)
# 연관어 시각화를 위한 igraph 패키지로딩
library(igraph)
#edgelist 보기 시각화 효과를 위하여 92 개의 객체중 30 개로만 진행
ruleg <- graph.edgelist(rulemat[c(1:50),], directed = F)
# edgelist 시각화
plot.igraph(ruleg, vertex.label = V(ruleg)$name,
                vertex.label.cex = 1.2, vertext.label.color = 'black',
                vertex.size = 20, vertext.color = 'green',
                vertex.frame.co.or = 'blue')
```

단어 시각화 효과를 위하여 92 개 연관성객체에서 50 개만 사용하기로 하였다. 이하 그림은 연관성 시각화이다.



<연관규칙 시각화 그림>

-연관규칙 분석

시각화 결과를 통해 연설문은 주요하게 "우크라이나","우리"란 단어가 중심으로 많이 출현되고 두 단어 사이 "안보","보장","제안" 연결되었다. 그 외에도 다양한 단어들이 일정한 빈도로 출현 되었다.

'러시아","안보","안전","전쟁" 등 단어들 에서 대통령의 러시아에 대한 강력한 태도를 갖고 있음을 알 수 있다.

4) 사례연구 문제점

-텍스트 토픽 분석 문제점

1. 처음 말뭉치 생성시 DrKing_nouns 선언되지 않는 변수를 사용하였다.
2. 데이터 전처리 부호, 수치, 수문자 변경, 불용어 제거 중 재차 선언되지 않은 ZeCorpusprepro 란 변수를 사용하여 코드 오류를 일으키고 있다.
3. 단어 수집과 그 결과 도출 시각화 분석 등등 부분이 전부 누락되어 있다.

-연관어 분석 문제점

1. 연관어 분석에 텍스트 마이닝 분석 wordcloud2 시각화 절차를 넣었다.
2. 단어 필터링 함수에 단어 최대길이 설정을 4 보다 작게하였다. 이는 "우크라이나"와 같은 최빈 단어는 제외 처리하게 된다.
3. 연관규칙 찾기에서 연관성 객체가 17 개 밖에 안된다. 이는 분석하기에 데이터가 부족하다고 생각된다.
4. 시각화 후 그 연관성에 대한 설명이 없다.