언어 자료와 컴퓨터 기말 과제

2016130714 김상준

제가 선택한 3개의 텍스트는 미국 초대 대통령 George Washington의 취임 연설문, 미국 제44대 대통령 Barak Obama의 2009년 취임 연설문, 그리고 Barak Obama의 2013년 취임 연설문입니다. 다양한 R 통계 기법을 바탕으로, 제 분석의 주제는 크게 두 가지입니다. 우선 어떤 Washington과 Obama의 연설문 사이에는 어떠한 차이가 나타나는지 분석하겠습니다. 더 나아가 Obama의 2009년과 2013년 연설문 사이에 어떤 차이가 나타나는지 분석해 보겠습니다.

1. 데이터 가공

우선 통계 기법 적용에 앞서 필수적인 데이터 가공을 했습니다. 코드는 다음과 같습니다.

```
> wash <- scan(file="washington.txt", what="", quote=NULL, encoding="UTF-8")
Read 1430 items
> obama_1 <- scan(file="obama 2009.txt", what="", quote=NULL, encoding="UTF-8")
Read 2022 items
> obama_2 <- scan(file="obama 2013.txt", what="", quote=NULL, encoding="UTF-8")
Read 2134 items
>
> obama_2 <- gsub("[(]Applause[.][)]", "", obama_2)
> wash <- gsub("^[[:punct:]]+|[[:punct:]]+$", "", tolower(wash))
> obama_1 <- gsub("^[[:punct:]]+|[[:punct:]]+$", "", tolower(obama_1))
> obama_2 <- gsub("^[[:punct:]]+|[[:punct:]]+$", "", tolower(obama_2))
> wash <- wash[nchar(wash) > 0]
> obama_1 <- obama_1[nchar(obama_1) > 0]
> obama_2 <- obama_2[nchar(obama_2) > 0]
>
> WASH <- sort(table(wash), decreasing=T)
> OBAMA_1 <- sort(table(obama_1), decreasing=T)
> OBAMA_2 <- sort(table(obama_2), decreasing=T)
>
> WASH <- data.frame(row.names=names(WASH), freq=as.vector(WASH))
> OBAMA_1 <- data.frame(row.names=names(OBAMA_1), freq=as.vector(OBAMA_1))
> OBAMA_2 <- data.frame(row.names=names(OBAMA_2), freq=as.vector(OBAMA_2))
```

Washington 파일을 변수 wash에, Obama 2009 파일을 변수 obama_1에, Obama 2013 파일을 변수 obama_2에 각각 어휘 단위로 불러 왔습니다. 이후 정규표현을 통해 소괄호 내 지문인 (Applause.), 그리고 어휘 앞뒤의 문장부호를 전부 제거했으며, 소문자 처리도 같이 했습니다. 정규표현 적용 결과 나온 빈 문자열은 전부 제거했습니다. 그리고 마지막으로 3개 텍스트를 모두 빈도 내림차순의 데이터프레임으로 가공하였습니다.

가공을 마쳤으니, 이제 통계 기법을 적용해 볼 텐데요. 가장 먼저 빈도 내림차순의 데이터 프레

임을 분석 방법으로 사용하겠습니다. 제가 다른 좀더 고급스러운 기법을 제쳐두고 이 방법을 선택한 것에는 두 가지 이유가 있습니다. 첫째로, 기초적인 방법으로 데이터를 어떻게 분석할지 방향을 잡거나 대략적인 아이디어를 얻을 수 있을 것입니다. 둘째로, 기초적인 방법이 다른 좀더 고급 기법에 비해 어떠한 장단점을 갖는지 파악할 수 있을 것입니다.

2. 빈도 분석 데이터프레임

```
> head(WASH, 15)
      freq
the    117
of      70
and     48
to      48
which   36
in      31
be      23
i       23
my      22
by      20
that    18
with    17
a       15
on      15
as      14
```

```
> head(OBAMA_1, 15)
     freq
the    99
and    94
of     57
to     53
we     47
that   40
a      39
in     33
this   30
our    28
for    27
i      27
who    26
is     19
will   18
```

```
> head(OBAMA_2, 15)
     freq
the   104
and    89
our    76
of     69
we     67
to     66
that   55
a      37
for    28
is     25
not    23
are    21
in     21
us     21
will   20
```

세 가지 텍스트 데이터의 빈도 내림차순 상위 15개의 데이터를 비교해 보겠습니다. 우선 가장 눈에 띄는 것은 상위의 function words입니다. Zipf의 법칙이 잘 적용되어 문법어 기능어가 전부 상위 빈도에 분포하는 모습입니다.

첫번째 분석 포인트입니다. 3개 텍스트 공통으로 상위에 위치하는 단어들은 [the, of, and, to]이라는 점에 좀더 주목해볼 필요가 있습니다. 사실 Obama 연설문 두 개의 function words가 유사하게 나타나는 것은 시대와 발화자를 고려할 때 당연하다고 생각할 수 있습니다. 동시대의 언어 사용, 동일한 화자의 습관이 반영되었을 것입니다. 그런데 1789년의 연설문까지 유사한 function words가 나타난다는 점은 굉장히 흥미롭습니다. 이는 250여 년 전의 영문법과 2000년대 영문법 사이의 차이가 크지 않다는 증거가 되기 때문입니다.

이 사실을 한국의 1700년대 글과 비교해보면 더더욱 놀랍습니다. "벼슬을저마다하면農夫되리뉘잇시며" 이것은 1700년대 한글 시조 필사 절칩본의 일부인데요. 이러한 텍스트를 가공하여 현대 한국어와 비교 분석한다면 차이가 극명하게 드러날 것입니다. 한국어는 1700년대부터 현대까지 상당히 많은 변화를 겪었다는 것이죠. 비해 영어 텍스트의 통시적 변화는 훨씬 적었다고 볼 수 있습니다.

왜 그랬을까요? 제가 문법사 관련 전문가는 아니지만 간단히 추측해 보겠습니다. 아마 현대 사회의 근간이 전부 서구 중심으로 구성되었기 때문이라고 생각합니다. 물론 서구 사회도 1700년대부터 많은 변화를 겪었겠으나, 분명 세계의 중심으로 계속 자리 잡아왔습니다. 상대적으로 변화가 적었을 것입니다. 반면 한국은 근대화, 일제 침략, 분단의 시대 등 격변 속에서 서구 사회를 많이 따라갔습니다. 그랬기 때문에 비교적 사회적으로 큰 변화를 겪었을 것이고 그 차이가 문법 빈도에까지 나타난 것이라고 추측할 수 있습니다.

실제로 관련 자료를 찾아보니, 영어는 1800년 무렵부터 현대 영어로 분류된다는 사실을 알 수 있었습니다. 1800년의 영어는 현대 영어와 거의 100% 일치한다는 것이죠. Washington 연설문의 작성 시기가 1789이기 때문에, 이 시점과 상당히 가깝다는 역사적 증거를 발견했습니다.

두 번째 분석 포인트는 "We"입니다. We와 관련된 어휘는 3가지, "we", "our", "us"라고 볼 수 있을 텐데요. OBAMA_1의 경우 we가 47번, our가 28번 총 75번 나타납니다. 반면 OBAMA_2에서는 our가 76번, we가 67번, 그리고 us가 21번 나타납니다. 총 164번으로서 놀랍게도 OBAMA_1에 비해 2배 넘는 수치가 드러납니다. 추가로 Washington의 경우에는 상위 빈도 15개의 어휘목록 중에서 we 관련 어휘가 하나도 등장하지 않습니다. Washington 파일 전체를 살펴보니, 상위 뿐만 아니라 하위 빈도에도 we 관련 어휘는 전혀 등장하지 않았다는 사실을 알 수 있었습니다. 즉 we 관련 어휘는 Washington 텍스트 전체에서 단 한 번도 나타나지 않았습니다! I가 23번, my가 22번 등장하는 동안 말입니다.

이와 관련하여, 기본적으로 we는 주로 정치인이 대중의 공감을 유도하기 위해 사용하는 표현이라고 알고 있습니다. 뭐 복합적인 이유가 있겠으나, Obama는 2009년보다 2013년에 훨씬 더 대중의 공감에 어필하는 표현을 많이 썼다고 추론해볼 수 있습니다. 물론 we가 무조건 공감 관련 표현이라는 것도 아니고, 시대적 배경, 우연성 등 다양한 원인이 있을 수 있겠습니다.

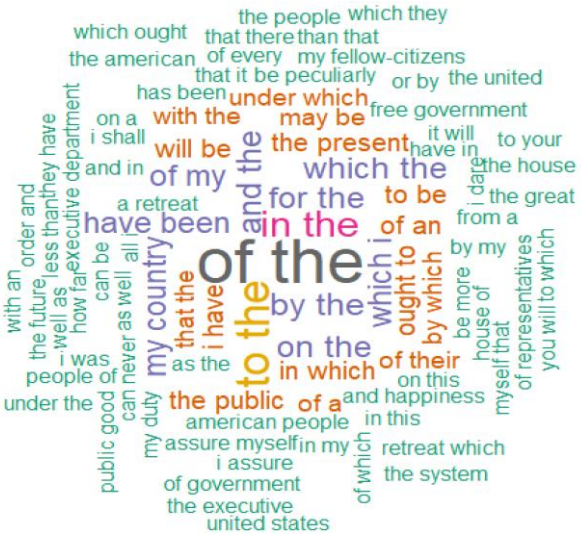
그에 반해 Washington은 we를 조금만 사용한 것이 아니라 전혀 사용하지 않았는데요. 실제로 그의 다른 연설 몇 개를 추가로 살펴봤으나 we를 한 번도 발견하지 못했습니다. 이 포인트는 시사하는 바가 크다고 생각합니다. 앞서 살펴봤듯, 영문법은 250여 년간 크게 변하지 않았으나, 정치는 그와 달리 상당히 변화했다고 추론해볼 수 있기 때문입니다. Washington 시대의 연설은 대통령이 대중에게 일방적으로 자신의 업적과 정책을 전달하는 용도였다면, Obama 시대의 연설은 타인과 대중을 보다 적극적으로 언급하는 연설이 된 것이 아닐까요? 정치인 스스로가 대중과 더 가까워지려고 하는 것은 물론, 지금 시대가 그러한 정치인을 원하는 것이기도 하겠지요.

위와 유사하게 will이라는 단어가 OBAMA_1, OBAMA_2에서 공통적으로 15위에 위치하고 있는 것도 비슷한 맥락이라고 생각합니다. will을 쓰는 문장은 주로 무슨 정책을 하겠다, 어떤 방향으로 나아가겠다며 대중들에게 강한 의지를 어필하는 문장일 것입니다. 뭐 이것은 미래지향적으로, 희망적으로 대중들에게 다가가려 했던 Obama 개인의 특성일 수도 있긴 하겠습니다.

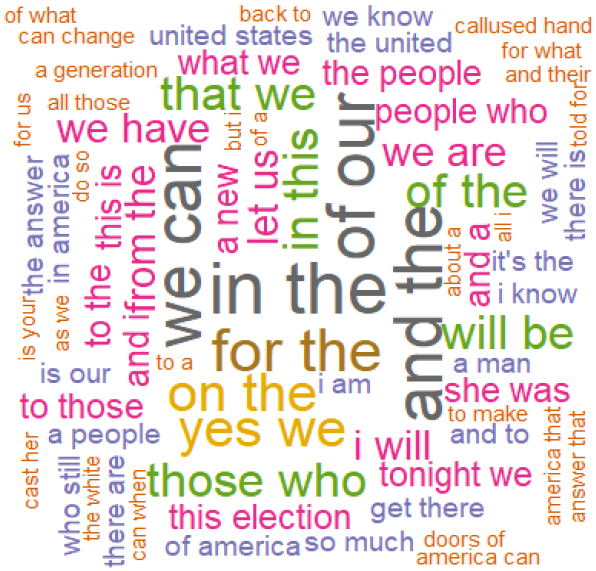
이제 빈도 데이터프레임의 분석을 마치고 n-gram의 기법을 통해 분석하도록 하겠습니다. 빈도 분석으로도 많은 정보를 얻긴 하였으나, 아무래도 맥락 및 위치 정보는 전혀 파악할 수 없었습니다. 이제 맥락을 살펴볼 수 있는 기법들을 하나하나 가져와 보겠습니다.

3. Bi-gram 워드 클라우드, 빈도 데이터프레임

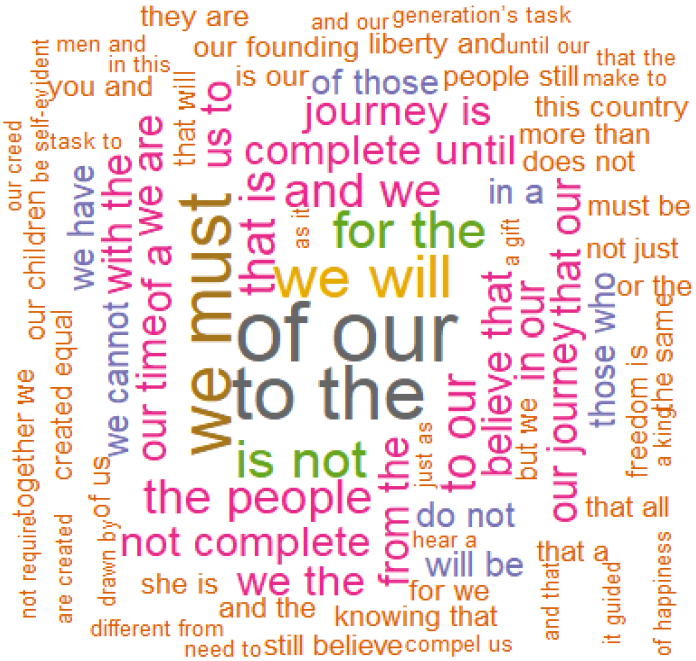
Washington



Obama 1



Obama 2




```
> bi.wash <- paste(wash[1:length(wash)-1], wash[2: length(wash)])
> bi.obama1 <- paste(obama_1[1:length(obama_1)-1], obama_1[2:length(obama_1)])
> bi.obama2 <- paste(obama_2[1:length(obama_2)-1], obama_2[2:length(obama_2)])
> bi.wash <- data.frame(sort(table(bi.wash), decreasing=T))
> bi.obama1 <- data.frame(sort(table(bi.obama1), decreasing=T))
> bi.obama2 <- data.frame(sort(table(bi.obama2), decreasing=T))
>
> library(wordcloud)
> wordcloud(bi.wash$bi.wash, bi.wash$Freq, scale=c(3, 0.8), min.freq=2,
+ max.words=90, random.order=F, rot.per=0.4, colors=brewer.pal(8, "Dark2"))
```

```
> head(bi.wash, 15)
      bi.wash Freq
1      of the   19
2      to the   12
3      in the    9
4     and the    7
5      by the    7
6     for the    6
7      on the    6
8     which i    6
9   which the    6
10  have been    5
11 my country    5
12      of my    5
13     i have    4
14   in which    4
15      of an    4
```

```
> head(bi.obamal, 15)
   bi.obamal Freq
1     in the   10
2    and the    9
3     of our    9
4     we can    9
5    for the    8
6     on the    7
7     yes we    7
8    in this    6
9     of the    6
10   that we    6
11 those who    6
12   will be    6
13  from the    5
14    i will    5
15    let us    5
```

```
> head(bi.obama2, 15)
        bi.obama2 Freq
1          of our   12
2          to the   11
3         we must   10
4         we will    8
5         for the    7
6          is not    7
7          and we    6
8         that is    6
9      the people    6
10         to our    6
11   believe that    5
12 complete until    5
13       from the    5
14         in our    5
15     journey is    5
```

처음에는 bi-gram으로 워드 클라우드를 만들었는데, 보다 세밀한 분석을 하기 위해 워드 클라우드만으로는 부족하다고 느꼈습니다. 그래서 추가로 빈도 데이터프레임까지 출력하여, 두 분석 기법을 적절히 섞어가며 분석했습니다. 또한 3개의 워드 클라우드에 거의 서로 같은 조건을 적용했습니다. 다만 Obama 텍스트의 경우 scale=c(3, 0.8)로 하면 warning message가 나와서 그것만 scale=c(3, 0.4)로 조정해준 결과입니다.

대충 봤을 때 가장 눈에 띄는 점은 Washington의 워드 클라우드가 비교적 가장 깔끔하다는 점입니다. Washington 텍스트에서는 특정 고빈도 bi-gram 몇 개만 큰 글자로 중심에 모여 있고, 나머지는 수 많은 단어들이 다 균일하게 작은 크기로 주변에 분포하고 있습니다. min.freq=2, max.words=90으로 설정하였으니, 그 작은 크기의 어휘들은 전부 frequency가 2인 단어들일 것입니다. 실제로 Washington 빈도 데이터프레임을 살펴 보면, 빈도 2의 단어가 92번째까지 존재한다는 사실을 알 수 있었습니다. 종합하자면 Washington 텍스트 빈도의 특징은, 특정 2개의 bi-gram, [of the, to the]의 빈도가 각각 19, 12으로서 상당히 압도적으로 높고, 빈도 2이상의 단어가 매우 많이 존재한다는 것입니다. 이와 달리 Obama의 두 텍스트는 빈도 사이의 격차가 비교적 적습니다. Bi-gram의 빈도가 원만하게 분포한다는 것입니다.

이제 각 bi-gram의 의미를 좀더 심층적으로 분석하도록 하겠습니다. 우선 Washington 텍스트 상위의 bi-gram은 전부 "function word + the"의 형태입니다. 여기에선 큰 의미를 찾아내기 어려울 것 같습니다. 기껏해야 the가 나올 만한 명사를 많이 썼다는 정도의 문법적 정보만 추론할 수 있을 것입니다. 이것이 이 이상으로 분석할 만한 대단한 특징은 아니라고 생각합니다.

반면 Obama의 두 텍스트 상위에 존재하는 bi-gram들은 분석할 점들이 많아 보입니다. 둘 다 "전치사 + the"의 표현이 많이 나오는 것은 텍스트의 특성 상 당연하긴 한데, 그 외에 눈에 띄는 표현들이 다수 관찰되는데요. 대표적으로 Obama 1의 [of our, we can, yes we]가 있고, Obama 2에서는 [of our, we must, we will]이 있습니다. 네 이것들은 앞서 빈도분서 데이터프레임에서 많이 살

펴본 we관련 표현들입니다. 대신에 이전의 분석과는 달리 맥락의 정보가 추가되었다는 의의가 있습니다. Obama는 we관련 어휘에 조동사 can, must, will을 붙여서 자주 사용했다는 사실이 보입니다. 이러한 표현은 청중을 연설에게 적극적으로 참여시키는 표현이자, 희망적이고 미래지향적인 표현입니다. 빈도분석 데이터프레임에서의 추론에 대한 확실한 증거가 나온 것입니다. We 관련 표현을 많이 쓴 데에는 위와 다른 특별한 이유는 없었다고 볼 수 있을 겁니다.

다음으로는 bi-gram 중 비교적 빈도 상위에 위치한 content words를 찾아 보겠습니다. Washington에서는 "my country, of my, I have"가 존재하며, Obama 1은 따로 명확한 content words가 존재하지 않으니 앞서 언급한 we 관련 어휘를 content word에 포함시켜 분석하겠습니다. Obama 2의 경우는 we 관련 표현 이외에도 "the people, believe that, complete until" 이 3가지가 눈에 띕니다.

첫째로 Washington에 있는 고빈도 bi-gram인 "my country, of my, I have"를 보겠습니다. 3개 다 i가 관련된 표현입니다. Obama 텍스트와 마찬가지로, uni-gram 기준 고빈도에 위치했던 어휘 중 상당수가 bi-gram에서도 당연히 고빈도에 위치하는 모습입니다. 앞서 언급했듯이 이 I 관련 표현은 대통령이 대중들에게 자신의 업적과 정책을 일방적으로 전달하기 위해 많이 쓰였다고 분석했습니다. Bi-gram을 통해 그러한 표현 주변에 country, of, have가 사용되었다는 맥락적 정보를 얻었습니다. 추가로, country는 따로 더 깊이 생각해 볼 어휘일 것인데요. "my country"라는 표현이 매우 주목할 만합니다.

우선 country라는 단어가 쓰인 이유야 시대적 배경을 생각해보면 금방 그럴 만하다고 느낄 수 있습니다. Washington은 미국의 초대 대통령이고, 미국의 founding father, 건국의 아버지 중 한 사람이기도 합니다. 이제 막 건국을 마치고, 최초의 대통령으로서 최초로 하는 연설인데, 건립한 국가에 대한 이야기를 빠뜨릴 수 없는 게 당연합니다. 유사한 맥락에서, 우리나라 초대 대통령이었던 이승만의 연설에도 "우리나라, 한국"과 같은 표현이 고빈도로 등장할 것이라고 예상합니다.

어쨌든 country 앞에 my가 쓰였다는 점도 그냥 넘어갈 수는 없습니다. 아쉽게도 Obama의 bi-gram 상위에 country 관련어가 등장하지 않았지만, 만약 등장했다면 "Our country"가 아니었을까요? 지금까지 분석한 바에 부합하려면 그러해야 합니다. 그래서 실제로 직접 데이터를 찾아보니까 Obama의 경우 country의 앞에 대명사 소유격을 붙일 때는 꼭 my가 아닌 our를 썼습니다. 사실 my country와 our country를 통해 대통령이 의도하고자 하는 바는 거의 동일하다고 봐도 무방한데, 그만큼 표현 방식의 차이가 두드러지는 것이죠. 오늘날 시대에 생각해보면, Washington 시대의 건국은 사실 그가 혼자 이룬 것이 아닌데, 마치 그 혼자 이룩한 것처럼 my 위주의 표현을 썼다는 사실이 아쉽긴 합니다. 대통령이 사용했을 때 we 관련 표현이 훨씬 더 타당하면서도 부드럽다고 느껴집니다.

Bi-gram에서 마지막으로 살펴볼 포인트는 Obama 2에 고빈도로 위치한 "the people, believe that, complete until"입니다. 우선 the people은 분석이 상당히 까다로운 표현이었습니다. The people이 Obama의 연설에 자주 등장하기는 하나, 사전 배경지식이 없다면 그것이 왜 자주 등장하는지 이유를 찾기 어려운 표현입니다. 제가 아는 바에 의하면 이 표현은 Abraham Lincoln의 명언 "government of the people, by the people, and for the people"에 3차례나 등장합니다. Lincoln의 명언을 한 번 언급하면 the people을 3번이나 말하는 꼴이 됩니다. 실제로 Obama는 해당 명언을 즐겨 썼으며, 2009년 2013년 연설 둘 다에 해당 명언이 등장합니다. 그 명언과 별개로 2013년 연설에서 Obama는 the people을 많이 썼는데, 텍스트를 섬세히 살펴보면 그것의 사용 형태는 항상 "we, the people"이었습니다. 즉 앞서 여러 차례 분석한 we와 맥락을 같이 하는 표현입니다. 딱히 실질적 의미는 없으나, Lincoln의 말을 자주 인용하던 Obama가 습관적으로 사용했던 것이라고 추정합니다.

그 다음으로 believe that, complete until이라는 bi-gram이 있습니다. 이 표현을 둘 다 역시 일종의 상당히 미래지향적인 표현이라고 예상합니다. 예를 들어 "나는 더 밝은 미래가 올 것이라고 믿는다." 이와 유사한 표현이라는 것이죠. Complete until은 "어떤 정치적 역사적 과제를 언제까지 끝내겠다." 이러한 표현일 것 같다고 쉽게 추측할 수 있습니다. 이들 역시 Obama의 연설 특성을 고려해봤을 때 자연스러운 표현입니다. 또한 미래의 과제를 제시하는 다양한 표현 중에 하나라고도 볼 수 있습니다. 대통령 연설의 필수 등장 요소인 업적, 과제 관련해 Washington은 my 위주의 단조로운 표현을 사용했다고 추측할 수 있습니다. Obama는 위와 같이 보다 더 다채로운 표현들을 바꿔가며 연설의 몰입도를 늘렸다고 분석할 수 있습니다. 아무래도 훨씬 현대에 가까운 글이니, 당연히 현대인의 입장에서 훨씬 세련되게 느껴지는 것이 정상입니다.

4. 연어 분석

```
> node <- "^my$"
> index <- grep(node, wash)
> span <- unlist(lapply(index,
+ function(i){c((i-4):(i-1), (i+1):(i+4))}))
> span <- span[span>0 & span <= length(wash)]
> crc.wash <- wash[span]
>
> freq.span <- sort(table(crc.wash), decreasing=T)
> freq.all <- table(wash)
> crc.wash <- data.frame(t(sapply(names(freq.span),
+ function(x) {c(length(index), freq.all[x], freq.span[x],
+  length(wash))})))
> colnames(crc.wash) <- c('W1', 'W2', 'W1W2', 'N')
>
> co.wash <- data.frame(crc.wash,
+ t.score=(crc.wash$W1W2 - ((crc.wash$W1*crc.wash$W2)/crc.wash$N))/
+ sqrt(crc.wash$W1W2),
+ MI=log2((crc.wash$W1W2*crc.wash$N)/
+ (crc.wash$W1*crc.wash$W2)))
>
> tsort.wash <- co.wash[order(co.wash$t.score, decreasing=T),]
> misort.wash <- co.wash[order(co.wash$MI, decreasing=T),]
> head(misort.wash, 10)
```

위의 코드는 Washington 텍스트에 대하여 node인 my를 기준으로 좌우 4개의 공기어 목록을 출력하고, 그것을 바탕으로 각종 공기빈도 및 t score, mi score를 구한 코드입니다. 추가적으로 분석의 간편함을 위해 t score와 mi score 기준으로 각각 정렬한 데이터프레임까지 만들었습니다. 이 코드를 기준으로 하여, node와 변수명만 바꿔가면 다른 텍스트 다른 노드에 관해 같은 목적의 분석 기법을 구현할 수 있습니다. 변수를 바꾼 코드는 생략하고, 분석 결과만 보이면 아래와 같습니다. 참고로 주로 고빈도에 function words를 출현하게 하는 t score 대신에, content words가 많이 나오는 mi score를 활용해 분석할 계획입니다. Function words에 대한 분석은 위에서 많이 했기 때문에 더 다룰 내용이 없을 것 같습니다.

또 node로는 Washington에서는 i와 my, Obama 1에서는 we, Obama 2에서는 our로 하였습니다. Washington에서 i와 my의 빈도는 각 23, 22로 거의 동일하고, Obama 1, 2에서는 각각 we관련어 중 가장 고빈도의 하나를 선택한 것입니다.

Washington, node = "i"

```
> head(misort.wash, 10)
            W1 W2 W1W2    N   t.score       MI
required    23  1    2 1430 1.4028405 6.958237
assure      23  2    2 1430 1.3914675 5.958237
dare        23  2    2 1430 1.3914675 5.958237
impressions 23  2    2 1430 1.3914675 5.958237
add         23  1    1 1430 0.9839161 5.958237
affected    23  1    1 1430 0.9839161 5.958237
again       23  1    1 1430 0.9839161 5.958237
am          23  1    1 1430 0.9839161 5.958237
aver        23  1    1 1430 0.9839161 5.958237
behold      23  1    1 1430 0.9839161 5.958237
```

Washington, node="my

```
> head(misort.wash, 10)
                W1 W2 W1W2    N  t.score       MI
collect         22  1    2 1430 1.403335 7.022368
faithful        22  1    2 1430 1.403335 7.022368
study           22  1    2 1430 1.403335 7.022368
thence          22  1    2 1430 1.403335 7.022368
country         22  5    5 1430 2.201667 6.022368
confidence      22  2    2 1430 1.392456 6.022368
fellow-citizens 22  2    2 1430 1.392456 6.022368
nor             22  2    2 1430 1.392456 6.022368
own             22  2    2 1430 1.392456 6.022368
sentiments      22  2    2 1430 1.392456 6.022368
```

Obama 1, node = "we"

```
> head(misort.obama1, 10)
           W1 W2 W1W2    N  t.score       MI
breathe    47  1    2 2007 1.397655 6.416236
celebrate  47  1    2 2007 1.397655 6.416236
overcome   47  1    2 2007 1.397655 6.416236
respond    47  1    2 2007 1.397655 6.416236
shall      47  1    2 2007 1.397655 6.416236
can't      47  3    4 2007 1.964873 5.831274
yes        47  7    9 2007 2.945358 5.778806
while      47  4    4 2007 1.953164 5.416236
achieve    47  2    2 2007 1.381095 5.416236
challenges 47  2    2 2007 1.381095 5.416236
```

Obama 2, node = "our"

```
> head(misort.obama2, 10)
           W1 W2 W1W2    N  t.score       MI
expelled   76  1    3 2097 1.711126 6.371146
forests    76  1    3 2097 1.711126 6.371146
tax        76  1    3 2097 1.711126 6.371146
affirm     76  1    2 2097 1.388586 5.786184
code       76  1    2 2097 1.388586 5.786184
commanded  76  1    2 2097 1.388586 5.786184
conscience 76  1    2 2097 1.388586 5.786184
empower    76  1    2 2097 1.388586 5.786184
interests  76  1    2 2097 1.388586 5.786184
mothers    76  1    2 2097 1.388586 5.786184
```

분석 시작하겠습니다. 먼저 Washington입니다. 연어 분석을 해보니까 확실히 구식 표현들이 다수 관찰됩니다. Dare, behold와 같은 단어는 사실 현대 영어에서는 잘 쓰지 않는 것들입니다. 그 시절에는 꽤 빈번하게 사용되었다는 사실을 가볍게 파악할 수 있습니다. 추가로 aver, thence와 같은 단어들도 보입니다. 이러한 단어들은 현대 영어에서 거의 전혀 쓰지 않는다고 봐도 무방합니다. 물론 틀린 단어들은 아니지만 굉장히 예스러운 단어들입니다. 연어 분석을 하니까 옛 영문법과의 차이를 극명하게 나타낼 수 있었습니다.

Washington과 관련하여 내용적으로 보자면, 저는 "fellow-citizens"라는 단어에 가장 관심이 갑니다. 직역하자면 "같은 처지의 시민들"이라는 뜻의 단어이지요. 그 뜻을 생각해보면, 이 단어가 we의 역할을 대신했다고 추론할 수 있습니다. Fellow-citizens는 I, my처럼 단순히 대통령 자신에 관한 표현이 아니라, 청중 전체를 언급하는 표현입니다. 이 표현이 상당히 구식이고 Obama식의 표현에 비해서는 친근감도 덜 느껴지긴 합니다. 어쨌든 중요한 점은 Washington 역시 청중을 연설 속으로 끌어들이는 표현을 다수 사용했다는 사실입니다. 그것도 node인 my를 기준으로 하기 때문에, "나와, 우리 시민들"을 한번에 언급하는 표현이었다고 확장할 수 있습니다.

정말 놀라운 발견입니다. 연어 분석을 하지 않았으면 Washington이 we관련 표현을 전혀 하지 않았다 판단하고 넘어갈 수 있었을 텐데 정말 다행입니다. Obama에 비해 빈도가 적긴 하나, Washington 역시 시대에 맞게 청중을 연설 속에 언급하고 있었습니다. 그들의 공감을 유도하고 있었다는 것입니다.

더 나아가 Washington이 impressions, confidence, sentiments와 같이 감정 관련 어휘를 꽤 언급했다는 점도 의외입니다. 사실 이전의 분석에서 그의 연설에 대해 완전 딱딱하고 공식적이라는 이미지 뿐이었습니다. 그러나 연어 분석을 통해 그 역시 시대에 맞게 청중과의 공감대 형성, 소통에 신경 쓰고 있었다는 사실을 파악하게 되었습니다.

Obama 텍스트에 등장하는 연어들을 분석하겠습니다. 가장 먼저 생각할 점은 매우 어휘가 풍부하다는 점입니다. Breathe, celebrate, expelled, forests, tax, conscience, mother과 같이 다양한 주제의 어휘가 폭넓게 등장하고 있습니다. 특히 세부적으로 보자면 Obama 1에서는 celebrate, overcome, achieve, challenge와 같은 단어가 분포하고 있습니다, 반대로 Obama 2에서는 expelled, forests, conscience, mother와 같은 단어들이 많이 보입니다. 두 그룹의 성격이 상당히 달라서 인상 깊습니다. Obama 1의 어휘들은 주로 당선을 "축하"하는 말, 그리고 자신이 앞으로 "극복, 성취" 할 "도전" 따위를 언급하고 있는 것이라고 추측할 수 있습니다. 이러한 단어들은 대통령 당선 후 취임 연설이라는 글의 특성을 고려해 봤을 때 자연스러운 말들입니다.

그런데 Obama 2에 등장하는 어휘들은 위와 달리 그다지 자연스럽지 않아서 문제인데요. Expelled, forests, conscience, mother가 사실 당선 연설문에 자주 등장할 단어들은 확실히 아닙니다. 왜 이런 단어들이 나타났을까 제 의견은 연설문의 전략이 달라졌다는 것입니다. 이전 2009년

의 연설문은 첫번째 취임이었던 만큼 축하와 감사인사 및 자신의 포부를 밝히는데 중점을 두었을 것입니다. 그러나 두번째 취임이었던 Obama 2에는 보다 감정적이고 비유적인 표현들을 적극적으로 사용한 것이 아닐까 싶습니다. 단순히 we관련 표현으로 공감대를 이끌어내는 전략을 넘어서서 연설문의 딱딱한 어투를 상당히 벗겨내는 전략이 아니었나 싶습니다. 연설문에 등장한다면 청중들의 흥미를 쉽게 끌 수 있는 표현들이기도 하고요.

5. 카이스퀘어 잔차

```
> WASH <- data.frame(table(wash))
> OBAMA_1 <- data.frame(table(obama_1))
> OBAMA_2 <- data.frame(table(obama_2))
> TDM <- merge(WASH, OBAMA_1, by.x="wash", by.y="obama_1", all=T)
> colnames(TDM) <- c('words', 'wash', 'obama1')
> TDM <- merge(TDM, OBAMA_2, by.x='words', by.y='obama_2', all=T)
> colnames(TDM)[4] <- "obama2"
>
> TDM <- data.frame(row.names=TDM$words, TDM[2:length(TDM)])
> TDM[is.na(TDM)] <- 0
>
> CHI <- chisq.test(TDM[c(1,2)])$residuals
Warning message:
In chisq.test(TDM[c(1, 2)]) : Chi-squared approximation may be in
> CHI <- as.data.frame(CHI)
```

위의 코드를 통해서 카이스퀘어 잔차를 구할 수 있습니다. 일단 위의 코드에서는 Washington, Obama 1, Obama 2를 전부 TDM에 merge하여 하나의 데이터프레임으로 만들었습니다. 카이스퀘어 잔차를 만들기 위한 전처리를 모두 마친 것이지요. 이제 c(1,2)열을 선택하면 Washington과 Obama 1을 비교하는 카이스퀘어 잔차가 형성될 것이고, c(2,3)열을 고르면 Obama 1과 Obama 2를 비교하는 카이스퀘어 잔차가 만들어질 것입니다. 그리고 필요한 대로 정렬하면 되겠지요. 단순 숫자 바꾸기라서 생략하고, 필요한 카이스퀘어 잔차만 출력하자면 다음과 같습니다.

Washington vs Obama 1 (Washington 내림차순)

```
> head(CHI[order(CHI$wash, decreasing=T),],10)
            wash    obama1
which   5.431763 -4.584956
the     2.861933 -2.415760
my      2.534450 -2.139331
of      2.360723 -1.992689
an      2.240952 -1.891589
public  2.217508 -1.871800
every   2.067641 -1.745298
present 2.024299 -1.708712
under   2.024299 -1.708712
be      1.938490 -1.636281
```

Washington vs Obama 1 (Obama 1 내림차순)

```
> head(CHI[order(CHI$obama1, decreasing=T),],10)
              wash   obama1
we       -4.245112 3.583303
our      -3.185692 2.689045
who      -2.609917 2.203033
america  -2.325680 1.963108
tonight  -2.325680 1.963108
us       -2.097890 1.770830
what     -2.039756 1.721760
she      -1.824413 1.539989
know     -1.706582 1.440528
yes      -1.706582 1.440528
```

Obama 1 vs Obama 2 (Obama 1 내림차순)

```
> head(CHI[order(CHI$obama1, decreasing=T),],10)
           obama1    obama2
i        3.040868 -2.974898
tonight  2.634467 -2.577313
there    2.565184 -2.509533
who      2.298583 -2.248716
was      2.118314 -2.072358
this     1.956412 -1.913969
yes      1.933170 -1.891231
you      1.902781 -1.861501
your     1.789768 -1.750940
campaign 1.789768 -1.750940
```

Obama 1 vs Obama 2, (Obama 2 내림차순)

```
> head(CHI[order(CHI$obama2, decreasing=T),],10)
            obama1   obama2
our      -3.205405 3.135865
these    -1.888195 1.847232
together -1.850202 1.810062
must     -1.848398 1.808298
citizens -1.712954 1.675792
equal    -1.712954 1.675792
freedom  -1.712954 1.675792
until    -1.712954 1.675792
believe  -1.563706 1.529782
complete -1.563706 1.529782
```

Washington과 Obama 1 사이의 키워드를 비교하는 카이스퀘어 잔차를 하나 만들었습니다. 그리고 그것에 대해 각 텍스트의 내림차순으로 정렬했습니다. 두번째 카이스퀘어 잔차로서 Obama 1과 Obama 2 사이를 비교하는 것을 만들었습니다. 그것 역시 각각에 대해 내림차순으로 정렬했습니다.

우선 Washington과 Obama 1 사이의 관계를 분석하도록 하겠습니다. 너무 많이 언급해서 i와 we 사이의 분석은 제외하고 다른 새로운 발견들에 초점을 맞추도록 하겠습니다. Function words도 수차례 언급했기 때문에 content words 위주로 보겠습니다. 이렇게 필터링한다면 Washington 내림차순에서 볼 것은 public, every, present이고, Obama 내림차순 측에서 볼 것은 America, tonight, she입니다. 분석에 앞서 용례분석을 통해 Washington에서 사용된 present는 모두 형용사 형태로 사용되었다는 사실을 알아 뒀습니다. 즉 "현재의, 존재하는"의 뜻으로 사용되었다는 말입니다.

Washington의 public 역시 용례를 보니까 "public good", "public harmony", "public prosperity" 이러했습니다. 이것들의 의미를 놓고 보면 일종의 "our"과 유사한 표현이라고 확장해 볼 수 있습니다. Public을 씀으로써 거리감이 많이 늘어나지만, 순수한 의미로는 "our good", "our harmony", "our prosperity"와 유사할 것입니다. 이 역시 당시 시대에 맞는 일종의 공감대 표현이라고 이해하고 넘어가면 될 것 같습니다. Present의 용례는 "present month, present crisis, present leave"와 같았습니다. 이러한 표현에서 present는 실질적인 의미를 갖는다기 보다는 언어적 습관이라고 생각했습니다. 문맥 상 month, crisis, leave 등만 써도 의미 전달에 크게 무리가 없다고 판단되었습니다. 이것이 Washington 개인의 습관일지, 당시 영문법의 관례일지까지는 더 배경지식이 있어야 따져볼 수 있을 것입니다.

Obama 1에서 두드러지는 con-tent words로는 America, tonight, she가 있는데요. America는 너무 뻔하니까 제쳐 두고 tonight를 보겠습니다. 용례를 살펴본 결과 tonight는 Washington의 present와 비슷한 역할이라고 판단할 수 있었습니다. Tonight가 굳이 없어도 되는 맥락에서 Obama는 습

관적으로 tonight를 넣어서 표현한 경우가 많았습니다. 물론 이로 인해 연설이 저녁에 이루어졌다는 정보도 알 수는 있겠습니다. She의 경우에는 Obama가 한 여인의 일화를 언급하였기 때문에 자주 등장한 것으로 분석되었습니다. 이는 익명의 여인에 대한 일화이고, 그렇기 때문에 이름이 아닌 대명사 표현 she를 통해 반복적으로 언급되었던 것입니다.

이젠 Obama 1과 Obama 2 사이를 살펴 보겠습니다. Obama 1에 비해 Obama 2에서는 길이가 긴 키워드들이 다수 관찰됩니다. "together, must, citizens equal, freedom"이 특히나 잘 보입니다. 이 단어들 중 equal과 freedom에서 Obama 2의 연설 주제가 명확하게 드러납니다. 바로 평등과 자유 사상입니다. 축하와 감사가 많았던 첫 취임에 비해 두번째 취임에서는 시대적 과제에 대한 언급을 더 많이 할 수 있었던 것이겠죠.

Together, must, citizens 역시 별로 다른 맥락이라곤 볼 수 없습니다. 시대적 과제에 대해서 청중들의 참여와 공감을 유도하고, 그것에 "must"라는 의무감까지 주는 표현들입니다. 이미 여러 차례 분석한 내용들이라 이쯤에서 분석을 마치겠습니다.

6. 결론

4가지 분석 기법으로 3가지 텍스트 Washington, Obama 2009, Obama 2013을 교차하며 비교해 봤습니다. 분석 결과, 기본적으로 각 분석 방법 사이에 공통되는 사항들이 많긴 했습니다. 그러나 순차적으로 좀더 고급 방법을 써가면서, 기존에 보이지 않았던 구체적인 정보들을 확실히 파악할 수 있었습니다.