성공한 기업인의 졸업식 연사 텍스트 분석

20120863 박정현

주제 소개 및 선정 이유: 세계적으로 성공한 기업인들의 대학교 졸업식 연사 텍스트를 분석해 공통점, 차이점, 특이점 등을 발견하고자 함. 이를 기반으로 성공한 기업인들의 인사이트를 배울 수 있기를 희망.

기업인 및 텍스트 소개:

주요 분석 인물 및 텍스트

1) 빌 게이츠 (William Henry Gates)

   출생: 1955년 10월 28일 (미국 워싱턴 주 시애틀)

   업적: 마이크로소프트 설립자이자 기업인

   텍스트 소개: 2007년 6월 7일, 하버드 대학 졸업식 연설문

2) 마크 주커버그 (Mark Elliot Zuckerberg)

   출생: 1984년 5월 14일 (미국 뉴욕 주 화이트플레인스)

   업적: 페이스북 설립자 및 현 CEO

   텍스트 소개: 2017년 5월 25일, 하버드 대학 졸업식 연설문

3) 스티브 잡스 (Steven Paul Jobs)

   출생: 1955년 2월 24일 (미국 캘리포니아 주 샌프란시스코)

   업적: 애플 창시자, 픽사 애니메이션 창시자

   텍스트 소개: 2005년 6월 12일, 스탠포드 대학 졸업식 연설문

* 마윈, 제프 베조스, 엘론 머스크 기업인들의 연설문도 연설문 길이 비교용으로 참고 함

분석 방법:

빈도분석, 연어분석, n-gram 분석,

1. 빈도분석

분석 방법:

1 – lemmatizer을 이용해 텍스트의 모든 단어에 형태소 태그를 단다.

*analyzer2.R 파일 소스코드

```r
library(koRpus)
set.kRp.env(TT.cmd = "C:/TreeTagger/bin/tag-english.bat", lang="en")

lemmatizer <- function (text) {
  tagged = taggedText(treetag(text, format="obj", treetagger="manual", TT.options=list(path="C:/TreeTagger", preset="en")))
  return(tagged)
}

analyzer <- function (path) {
  original_txt <- scan(file=path, what="char", sep=" ", quote=NULL, encoding="UTF-8")
  #original_txt <- scan(file='./원문/billgates.txt', what="char", sep=" ", quote=NULL, encoding="UTF-8")
  #original_txt
  original_txt <- unlist(strsplit(original_txt, '\n'))
  all_paste <- paste(gsub("\n", "", original_txt), collapse = " ")
  #tolower(all_paste)


  # 소문자로 변경
  #gsub("I","I★", all_paste)

  tagged <- lemmatizer(all_paste)
  head(tagged)

  # 고유 단어라 lemma가 unknown으로 뜨는건 token 값으로 벡터에 저장
  # 나머지는 lemma(원형)으로 벡터에 저장
  words_tagged <- vector()
  # detail 품사 tagger
  for (i in 1:nrow(tagged)) {
    if (tagged[i, ]$lemma == "<unknown>") {
      add_word <- paste(tagged[i, ]$token, tagged[i, ]$tag, sep="_")
      words_tagged <- c(words_tagged, add_word)
    } else {
      add_word <- paste(tagged[i, ]$lemma, tagged[i, ]$tag, sep="_")
      words_tagged <- c(words_tagged, add_word)
    }
  }
  return(words_tagged)
}
```

*출력 예시

> head(analyzer('billgates.txt'))

[1] "President_NP" "Bok_NP" ",_," "former_JJ" "President_NP" "Rudenstine_NP"

2 – 형태소 태그가 달려있는 문자열 벡터에서 필요한 품사만 grep으로 뽑아내 데이터 프레임으로 변경한다. 이후 워드클라우드로 출력한다.

* 빌게이츠 텍스트 분석하는 프로그램 소스코드 (이하 모든 프로그램에서 텍스트 원문 불러오는 부분 제외하고 모두 동일)

- 내용어를 구성하는 명사(_N**), 동사(_V**, have와 be 동사 제외), 형용사(JJ), 부사(RB)를 추출

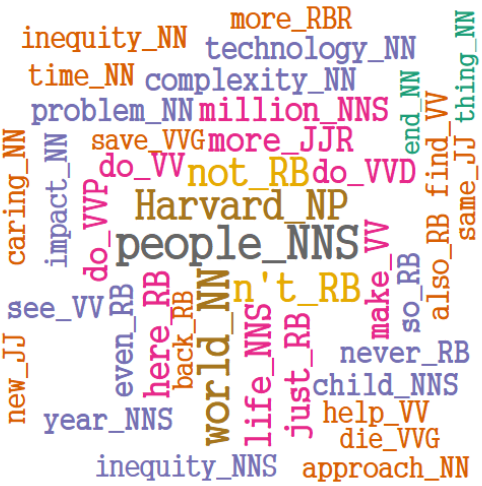
```r
source('./analyzer2.R')
source('./makewc.R', encoding='utf-8')

bill_txt <- analyzer('./원문/billgates.txt')
grep("_VV$", bill_txt, value=T)
grep("_V[^V]", bill_txt, value=T)
grep("_V[^BH]", bill_txt, value=T)
# 필요한 품사만 grep
#filtered_tagged <- grep("_nn|_nnp|_nnps|_vb|_vbd|_vbg|_vbn|_vbp|_vbz|_md|_jj|_jjr|_jjs|_prp|_prp$|_rb|_rbr|_rbs", words_tagged, value=T)
filtered_tagged <- grep("_N|_V[^BH]|_JJ|_RB", bill_txt, value=T)
#filtered_tagged <- grep("_NP", words_tagged, value=T)

tag_table<- sort(table(filtered_tagged), decreasing = T)
Freq.tag <- data.frame(tag_table)
Freq.tag <- data.frame(row.names = Freq.tag$filtered_tagged, Freq=Freq.tag$Freq)

makeWordcloud(Freq.tag)
```

*출력 예시

> head(Freq.tag, 5)

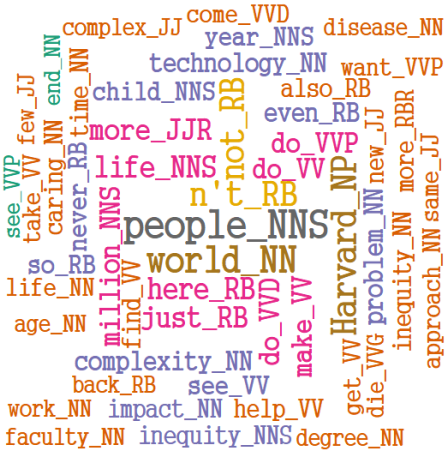
	Freq
people_NNS	26
world_NN	22
Harvard_NP	20
n't_RB	18
not_RB	17

> makeWordcloud(Freq.tag)



1) 빌게이츠 텍스트 빈도분석

> 빌게이츠 연설문의 워드클라우드



> 빈도 수 4회 이상의 단어들

		Freq		Freq		Freq		Freq		Freq
2	people_NNS	26	even_RB	9	more_RBR	6	disease_NN	4	much_RB	4
3	world_NN	22	problem_NN	9	new_JJ	6	everyone_NN	4	nothing_NN	4
4	Harvard_NP	20	technology_NN	9	time_NN	6	faculty_NN	4	opportunity_NN	4
5	n't_RB	18	child_NNS	8	approach_NN	5	few_JJ	4	other_NNS	4
6	not_RB	17	never_RB	8	come_VVD	5	good_JJS	4	poverty_NN	4
7	here_RB	13	see_VV	8	get_VV	5	government_NNS	4	save_VVG	4
8	life_NNS	13	so_RB	8	life_NN	5	great_JJ	4	say_VVD	4
9	just_RB	12	year_NNS	8	same_JJ	5	human_JJ	4	software_NN	4
10	more_JJR	12	impact_NN	7	take_VV	5	know_VVP	4	way_NNS	4
11	million_NNS	11	inequity_NNS	7	want_VVP	5	little_JJ	4	work_NN	4
12	do_VV	10	also_RB	6	age_NN	4	make_VVD	4		
13	do_VVD	10	caring_NN	6	back_RB	4	make_VVZ	4		
14	do_VVP	10	find_VV	6	complex_JJ	4	market_NN	4		
15	make_VV	10	help_VV	6	degree_NN	4	member_NNS	4		
16	complexity_NN	9	inequity_NN	6	die_VVG	4	most_RBS	4		

> 해석

1, 2번 째로 나온 people와 world는 영어에서 자주 쓰이는 명사 Top5에 포함되고, (http://www.talkenglish.com/vocabulary/top-1500-nouns.aspx) 4, 5번 째로 나온 n't와 not 또한 가장 자주 쓰이는 부사 Top1이기 때문에 이를 고려해야 한다. 즉, 이러한 단어들은 단순 절대 빈도분석 보다는 다른 텍스트들과의 상대 빈도 분석을 할 때 의미를 찾을 수 있을 것이라 생각한다. (이후 분석에서 비교)

또한 here, just, more, million, make 등의 단어들은 문맥을 알지 못하면 단어 자체만으로 의미를 많이 찾을 수 없기에 이 단어들은 연어 또는 n-gram 등 문맥을 함께 분석하는 분석에서 상대적으로 더 유의미할 것이다. ㅐ

이를 제외하면 Harvard, life, inequity (NN + NNS를 합치면 빈도수 총 13회로 6순위),

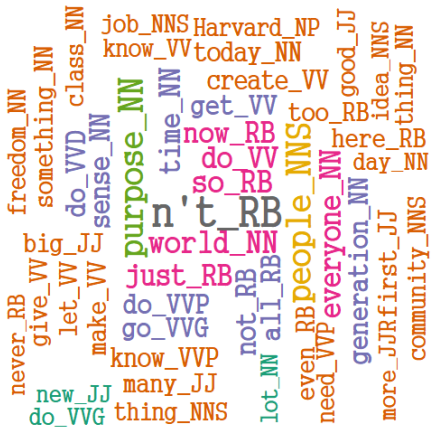
complexity, problem, technology, child 을 분석 대상으로 뽑을 수 있다.

하버드 대학의 졸업식 연설이고, 빌게이츠도 자퇴를 하긴 했지만 하버드 대학 출신이었기 때문에 대학교와 관련된 일화 등을 많이 이야기 한 것으로 볼 수 있다. 그리고 life라는 단어를 자주 언급함으로써 연설문 전체가 삶의 방향에 대한 이야기임을 알 수 있다. 그리고 그 방향에 inequity, complexity, problem 등의 단어를 써서 불공평한 사회 문제와 연관 짓고 있다. 그리고 이를 technology를 통해서 해결할 수 있다고 말한다. 이후 n-gram 분석에서 확인하겠지만 불공평한 사회 문제를 더욱 부각하기 위해 어린 아이들(child)이 죽어 나가는 (die, disease 빈도수 4회) 예시를 많이 사용했다. 이외에도 빈도수는 4회로 상대적으로 낮지만 caring, help, poverty 등의 단어들을 사용한 것을 보면 이러한 해석이 설득력 있다고 볼 수 있다.

즉, 빌게이츠는 동문인 하버드 대학 졸업생들에게 세상의 불공평함을 인지하고 이를 해결하기 위한 삶을 살아 달라는 연설을 한 것으로 해석할 수 있다.

2) 마크 주커버그 텍스트 빈도분석

> 마크 주커버그 연설문의 워드클라우드



> 빈도 수 4회 이상의 단어들

		Freq		Freq		Freq		Freq		Freq		Freq		Freq
2	n't_RB	44	do_VVD	12	Harvard_NP	8	global_JJ	6	life_NNS	5	change_NN	4	lot_NN	4
3	people_NNS	28	generation_NN	12	know_VV	8	great_JJ	6	million_NNS	5	company_NN	4	meaningful_JJ	4
4	purpose_NN	27	go_VVG	12	let_VV	8	idea_NNS	6	most_RBS	5	connect_VV	4	meet_VVD	4
5	world_NN	21	create_VV	11	thing_NNS	8	more_JJR	6	new_JJ	5	enough_JJ	4	move_VVG	4
6	so_RB	20	know_VVP	11	class_NN	7	never_RB	6	only_RB	5	equality_NN	4	person_NN	4
7	do_VV	19	today_NN	11	community_NN	7	project_NNS	6	parent_NNS	5	Facebook_NP	4	pursue_VV	4
8	everyone_NN	19	big_JJ	10	freedom_NN	7	society_NN	6	part_NN	5	forward_RB	4	school_NN	4
9	just_RB	18	make_VV	10	give_VV	7	someone_NN	6	start_VVD	5	get_VVD	4	single_JJ	4
10	now_RB	17	many_JJ	10	good_JJ	7	about_RB	5	want_VVD	5	get_VVG	4	start_VV	4
11	time_NN	16	too_RB	10	job_NNS	7	alone_RB	5	way_NN	5	graduate_VVG	4	start_VVN	4
12	not_RB	15	even_RB	9	need_VVP	7	college_NN	5	wrong_JJ	5	great_JJS	4	teach_VVD	4
13	sense_NN	14	first_JJ	9	Priscilla_NP	7	country_NN	5	build_VV	4	help_VV	4	tell_VVD	4
14	do_VVP	13	here_RB	9	thing_NN	7	do_VVG	5	build_VVD	4	important_JJ	4	want_VVP	4
15	get_VV	13	something_NN	9	together_RB	7	get_VVP	5	build_VVG	4	keep_VV	4	year_NN	4
16	all_RB	12	community_NNS	8	day_NN	6	idea_NN	5	challenge_NN	4	local_JJ	4	year_NNS	4

> 해석

빌게이츠 연설문과 마찬가지로 주커버그의 연설문에서도 1,2,4 번 째로 많이 나온 n't, people, world는 원래 영어에서 빈도수가 높기 때문에 이후 상대 빈도 분석을 통해 파악해야 할 것이다.

단순 절대 빈도에서는 purpose, everyone, sense, all, generation, idea (NN + NNS 합쳐 11회), community, Harvard, freedom, Priscilla, together, global, society 등을 의미 있게 해석해 볼 수 있겠다. 주커버그는 일반적인 단어는 아닌 purpose라는 단어를 압도적으로 많이 사용함으로써 '목적 있는 삶'을 강조했다. 이때의 목적은 단순히 자신의 목적이 아닌, '모든 사람들이 (everyone, all, together) 목적을 가진 삶'을 의미한다. 즉, 모든 사람들이 목적을 가진 삶을 살 수 있도록 더 나은 세상을 만들자는 이야기를 하고 있다.

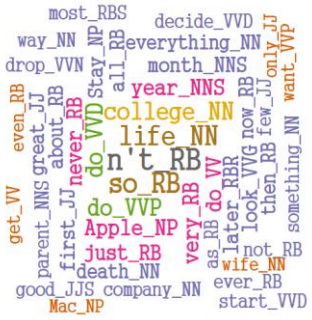
이를 달성하기 위해 주커버그는 '공동체'를 강조한다. (community, together, global, society 등) 재미있는 것은 사실 주커버그는 자신의 회사인 'Facebook'을 소개할 때도 회사의 미션이 전 세계 공동체를 하나로 묶는 것이라고 자주 이야기 한다는 사실이다. 즉, 대놓고 드러내지는 않았지만 위에서 말한 세상을 위해선 페이스북이 반드시 필요하다는 얘기를 하고 있다고도 볼 수 있다.

그 외에도 고유명사이자 자신의 아내인 Priscilla를 7회나 언급하면서 아내에 대한 사랑을 드러내기도 했고 주커버그도 빌게이츠와 마찬가지로 Harvard 대학 자퇴생으로서 Harvard 대학 졸업 연설을 했다.

추가적으로 주커버그가 최근 미국 대통령을 준비하고 있다는 소문이 돌고 있는데, 이를 알고 다시 한번 연설문을 보면 주커버그의 연설문은 기업가가 아닌 대통령이 할 법한 연설을 하고 있다고도 볼 수 있다. 재미있는 것은 우연이겠지만 6명의 기업가들 중에서 주커버그의 연설문만 압도적으로 긴데, (6개 연설문 평균 단어 수 2,193개 / 주커버그 연설문 단어 수 3,581개) 전 미국 대통령인 오바마의 연설문도 총 5,998 단어로 대학교 졸업 연설문 중에 긴 편에 속했다.

3) 스티브 잡스 텍스트 빈도분석

> 스티브 잡스 연설문의 워드클라우드



> 빈도수 4회 이상의 단어들

		Freq		Freq		Freq		Freq
2	n't_RB	19	about_RB	6	good_JJS	5	go_VV	4
3	so_RB	16	all_RB	6	most_RBS	5	go_VVG	4
4	life_NN	15	death_NN	6	not_RB	5	heart_NN	4
5	college_NN	12	everything_NN	6	parent_NNS	5	know_VV	4
6	do_VVD	11	first_JJ	6	something_NN	5	love_VVD	4
7	do_VVP	10	great_JJ	6	start_VVD	5	today_NN	4
8	Apple_NP	9	now_RB	6	way_NN	5	want_VVP	4
9	very_RB	9	Stay_NP	6	computer_NN	4	work_NN	4
10	year_NNS	9	then_RB	6	connect_VV	4		
11	do_VV	8	as_RB	5	course_NN	4		
12	just_RB	8	company_NN	5	die_VV	4		
13	never_RB	8	decide_VVD	5	dot_NNS	4		
14	later_RBR	7	drop_VVN	5	even_RB	4		
15	look_VVG	7	ever_RB	5	fire_VVN	4		
16	month_NNS	7	few_JJ	5	get_VV	4		

> 해석

1,2 번 째에 나온 n't와 so는 원래 영어에서 높은 빈도수를 보이므로 절대 빈도 분석에서는 해석하지 않겠다. 하지만 앞서 분석한 빌게이츠, 주커버그의 연설문에서 주로 쓰인 people와 world는 스티브 잡스의 연설문에서는 거의 보이지 않았다. (people 2회, world 3회) 원래 영어에서 높은 빈도수를 보이는 단어 임에도 거의 쓰지 않은 것은 해석해 볼 여지가 있다. (이후 분석)

여기서는 life, college, death, Stay 등을 뽑아볼 수 있겠다. 스티브 잡스도 빌게이츠처럼 삶(life)에 대한 이야기를 주로 했는데, 삶의 방향성을 제시한 빌게이츠와는 다르게 잡스는 자신의 삶에 대한 이야기만 했다. 삶에 대한 3가지 이야기를 했는데 이 중 첫 번째 이야기는 주로 자신의 대학시절(college) 이야기였다. 여기서 스티브 잡스도 빌게이츠/주커버그와 마찬가지로 대학 중퇴자이지만 둘과는 다르게 대학에서 배운 것들에 대해 많이 이야기를 했다. 또 그는 다른 기업가들이 거의 언급하지 않은 death를 많이 언급했는데, 이는 그의 세 번째 이야기가 죽음과 관련한 이야기이기 때문이다. 그리고 대문자로 시작하는 명사 'Stay'가 6회나 언급된 것은 주목할 만한데, 이 연설에서 바로 스티브 잡스의 모토로 유명한 'Stay Hungry, Stay Foolish'가 유명해졌기 때문이다.

스티브 잡스의 연설문은 절대 분석만으로는 특이점을 찾기 어려웠다. 따라서 차후에 다각적인 분석을 통해 더 깊게 살펴보고자 한다.

2. 연어분석

분석 방법:

중심어 좌우 +-6 어휘 범위로 공기어 추출 후 t-score과 MI 공식을 이용해 연어 추출

* 중심어는 빈도분석 만으로는 의미를 해석하기 어려웠던 단어들을 기반으로 선택

** 중간에 내용어 공기어 추출할 때는 빈도분석과 마찬가지로 명사(_N**), 동사(_V**, have와 be 동사 제외), 형용사(JJ), 부사(RB)를 추출

*연어분석 R 파일 소스코드

```r
source('./analyzer2.R')

bill_txt <- analyzer('./원문/billgates.txt')
mark_txt <- analyzer('./원문/zuckerberg.txt')
steve_txt <- analyzer('./원문/stevejobs.txt')

ana_txt <- bill_txt

search <- "people_NNS"
index <- which(ana_txt==search)
span <- vector()
for (i in index) {
  span <- c(span, (i-6):(i+6))
}
span <- span[span>0&span<=length(ana_txt)]
cooccurrence <- ana_txt[span]

Freq.span <- sort(table(cooccurrence), decreasing=T)
Freq.all <- sort(table(ana_txt), decreasing = T)
Freq.co <- data.frame(W1=vector(), W2=vector(), W1W2=vector(), N=vector())
n <- 1
for (i in (2:length(Freq.span))) {
  Freq.co[n,] <- c(Freq.span[1],
                   Freq.all[names(Freq.all)==names(Freq.span)[i]],
                   Freq.span[i], length(ana_txt))
  rownames(Freq.co)[n] <- names(Freq.span)[i]
  n <- n+1
}
collocates <- data.frame(Freq.co,
                         t.score =(Freq.co$W1W2 - ((Freq.co$W1*Freq.co$W2)/Freq.co$N))/
                           sqrt(Freq.co$W1W2),
                         MI = log2((Freq.co$W1W2*Freq.co$N)/
                                     (Freq.co$W1*Freq.co$W2)))

collocates <- collocates[grep("_N|_V[^BH]|_JJ|_RB", rownames(collocates)),]
t.score.sort <- collocates[order(collocates$t.score, decreasing=T), ]
MI.sort <- collocates[order(collocates$MI, decreasing=T), ]
MI.sort <- MI.sort[MI.sort$W1W2>2,]
head(t.score.sort, 10)
head(MI.sort, 10)
```

1) 빌게이츠 텍스트 연어 분석

> 빌게이츠 연설문의 t-score / MI 내림차순 정렬 데이터프레임

- 중심어 [people_NNS]

```
> head(t.score.sort, 10)
              W1 W2 W1W2    N  t.score       MI
life_NNS      26 13    6 3395 2.408845 5.913279
world_NN      26 22    5 3395 2.160720 4.891253
million_NNS   26 11    4 3395 1.957879 5.569325
nothing_NN    26  4    3 3395 1.714365 6.613719
n't_RB        26 18    3 3395 1.652463 4.443794
access_NN     26  2    2 3395 1.403383 7.028756
excite_VVN    26  2    2 3395 1.403383 7.028756
excitement_NN 26  2    2 3395 1.403383 7.028756
many_JJ       26  2    2 3395 1.403383 7.028756
get_VVG       26  3    2 3395 1.397968 6.443794
> head(MI.sort, 10)
            W1 W2 W1W2    N  t.score       MI
nothing_NN  26  4    3 3395 1.714365 6.613719
life_NNS    26 13    6 3395 2.408845 5.913279
million_NNS 26 11    4 3395 1.957879 5.569325
world_NN    26 22    5 3395 2.160720 4.891253
n't_RB      26 18    3 3395 1.652463 4.443794
```

- 중심어 [child_NNS]

```
> head(t.score.sort, 10)
             W1 W2 W1W2    N   t.score       MI
die_VVG       8  4    3 3395 1.7266089 8.314158
million_NNS   8 11    3 3395 1.7170856 6.854727
subsidize_VV  8  1    2 3395 1.4125473 9.729196
die_VVP       8  2    2 3395 1.4108811 8.729196
disease_NNS   8  3    2 3395 1.4092149 8.144233
not_RB        8 17    2 3395 1.3858876 5.641733
answer_NN     8  1    1 3395 0.9976436 8.729196
article_NN    8  1    1 3395 0.9976436 8.729196
assume_VVN    8  1    1 3395 0.9976436 8.729196
decline_NN    8  1    1 3395 0.9976436 8.729196
> head(MI.sort, 10)
            W1 W2 W1W2    N  t.score       MI
die_VVG      8  4    3 3395 1.726609 8.314158
million_NNS  8 11    3 3395 1.717086 6.854727
```

> 해석

People이 life와 world 연어인 것을 보면 빈도 분석에서 말한 '이 연설문은 사람들의 삶의 방향성을 제시하고 있다'는 해석이 타당하다는 것을 보여준다. 뿐만 아니라 빈도 분석에서 불공평한 사회 문제를 부각하기 위해 어린 아이들(child)이 죽어 나가는 예시를 사용했다고 했는데, 실제로 child의 연어는 die, disease 등의 단어가 나왔다. 따라서 이 해석은 타당한 것으로 볼 수 있다. 단순 빈도 분석으로는 발견할 수 없었던 해석이 연어 분석을 통해 나올 수 있다는 전형적인 예시로 볼 수 있겠다.

2) 마크 주커버그 텍스트 연어 분석

> 마크 주커버그 연설문의 t-score / MI 내림차순 정렬 데이터프레임

- 중심어 [purpose_NN]

```
> head(t.score.sort, 10)
            W1 W2 W1W2    N  t.score       MI
sense_NN    27 14   12 4150 3.437808 7.041616
everyone_NN 27 19    7 4150 2.599029 5.823436
create_VV   27 11    6 4150 2.420273 6.389539
pursue_VV   27  4    4 4150 1.986988 7.264008
freedom_NN  27  7    4 4150 1.977229 6.456653
n't_RB      27 44    4 4150 1.856867 3.804577
not_RB      27 15    3 4150 1.675707 4.942080
people_NNS  27 28    3 4150 1.626876 4.041616
high_JJR    27  2    2 4150 1.405013 7.264008
well_RBR    27  2    2 4150 1.405013 7.264008
> head(MI.sort, 10)
            W1 W2 W1W2    N  t.score       MI
pursue_VV   27  4    4 4150 1.986988 7.264008
sense_NN    27 14   12 4150 3.437808 7.041616
freedom_NN  27  7    4 4150 1.977229 6.456653
create_VV   27 11    6 4150 2.420273 6.389539
everyone_NN 27 19    7 4150 2.599029 5.823436
not_RB      27 15    3 4150 1.675707 4.942080
people_NNS  27 28    3 4150 1.626876 4.041616
n't_RB      27 44    4 4150 1.856867 3.804577
```

- 중심어 [community_NN]

```
> head(t.score.sort, 10)
                  W1 W2 W1W2    N   t.score       MI
community_NN      11  7    7 4150 2.6387385 8.559464
global_JJ         11  6    2 4150 1.4029680 6.974502
purpose_NN        11 27    2 4150 1.3636086 4.804577
authoritarianism_NN 11  1    1 4150 0.9973494 8.559464
building_NN       11  1    1 4150 0.9973494 8.559464
church_NN         11  1    1 4150 0.9973494 8.559464
excite_VVN        11  1    1 4150 0.9973494 8.559464
job_NN            11  1    1 4150 0.9973494 8.559464
keep_VVG          11  1    1 4150 0.9973494 8.559464
live_VVP          11  1    1 4150 0.9973494 8.559464
> head(MI.sort, 10)
            W1 W2 W1W2    N  t.score       MI
community_NN 11  7    7 4150 2.638738 8.559464
global_JJ    11  6    2 4150 1.402968 6.974502
purpose_NN   11 27    2 4150 1.363609 4.804577
```

* community 보다 ‘,’가 더 공기어 빈도가 높아서 community가 2순위로 밀려나게 나왔음. 하지만 어차피 ‘,’는 의미를 딱히 찾을 수 없으므로 그대로 진행함

> 해석

주커버그가 연설에서 단순히 자신의 목적이 아닌 ‘모든 사람들이 목적은 가진 삶’을 강조한 것이 purpose의 연어 분석에서 그대로 드러났다. everyone, people 등이 연어로 나온 것으로 보아 이를 타당한 해석이라고 볼 수 있다. 단순 purpose의 빈도만 봤더라면, 단순히 개인의 목적이라고 해석할 수도 있지만, everyone이라는 연어를 통해서 이 목적이 ‘모든 사람들의 목적’이라고 해석할 수 있었다. sense, create, pursue, freedom 등의 용어들도 연어로 같이 사용됐는데 이는 n-gram 분석에서 더 자세하게 해석하겠다.

또한 빈도분석에서 그 목적을 달성하기 위해서 글로벌한(global) 공동체(community)를 만들어야 한다고 말했는데, community의 연어들 중 global 과 purpose가 압도적으로 높은 것으로 보아 이 해석이 타당함을 알 수 있다.

3) 스티브 잡스 텍스트 연어 분석

> 스티브 잡스 연설문의 t-score / MI 내림차순 정렬 데이터프레임

- 중심어 [life_NN]

```
> head(t.score.sort, 10)
                W1 W2 W1W2    N   t.score       MI
do_VV           15  8    3 2545 1.7048280 5.991522
entire_JJ       15  2    2 2545 1.4058783 7.406559
year_NNS        15  9    2 2545 1.3767050 5.236634
adult_NN        15  1    1 2545 0.9941061 7.406559
application_NN  15  1    1 2545 0.9941061 7.406559
bring_VVD       15  1    1 2545 0.9941061 7.406559
choice_NNS      15  1    1 2545 0.9941061 7.406559
creative_JJ     15  1    1 2545 0.9941061 7.406559
destiny_NN      15  1    1 2545 0.9941061 7.406559
difference_NN   15  1    1 2545 0.9941061 7.406559
> head(MI.sort, 10)
          W1 W2 W1W2    N  t.score       MI
entire_JJ 15  2    2 2545 1.405878 7.406559
do_VV     15  8    3 2545 1.704828 5.991522
year_NNS  15  9    2 2545 1.376705 5.236634
```

- 중심어 [college_NN]

```
> head(t.score.sort, 10)
              W1 W2 W1W2    N   t.score       MI
never_RB      12  8    4 2545 1.9811395 6.728487
choose_VVD    12  1    2 2545 1.4108795 8.728487
naively_RB    12  1    2 2545 1.4108795 8.728487
mother_NN     12  3    2 2545 1.4042113 7.143525
go_VV         12  4    2 2545 1.4008772 6.728487
later_RBR     12  7    2 2545 1.3908749 5.921133
year_NNS      12  9    2 2545 1.3842067 5.558562
expensive_JJ  12  1    1 2545 0.9952849 7.728487
father_NN     12  1    1 2545 0.9952849 7.728487
figure_VV     12  1    1 2545 0.9952849 7.728487
> head(MI.sort, 10)
            W1 W2 W1W2    N  t.score       MI
choose_VVD  12  1    2 2545 1.410879 8.728487
naively_RB  12  1    2 2545 1.410879 8.728487
mother_NN   12  3    2 2545 1.404211 7.143525
never_RB    12  8    4 2545 1.981139 6.728487
go_VV       12  4    2 2545 1.400877 6.728487
later_RBR   12  7    2 2545 1.390875 5.921133
year_NNS    12  9    2 2545 1.384207 5.558562
```

- 중심어 [death_NN]

```
> head(t.score.sort, 10)
               W1 W2 W1W2    N   t.score       MI
certainty_NN    6  1    1 2545 0.9976424 8.728487
destination_NN  6  1    1 2545 0.9976424 8.728487
face_NN         6  1    1 2545 0.9976424 8.728487
face_VVG        6  1    1 2545 0.9976424 8.728487
fall_VVP        6  1    1 2545 0.9976424 8.728487
hope_VVP        6  1    1 2545 0.9976424 8.728487
intellectual_JJ 6  1    1 2545 0.9976424 8.728487
leave_VVG       6  1    1 2545 0.9976424 8.728487
purely_RB       6  1    1 2545 0.9976424 8.728487
share_NN        6  1    1 2545 0.9976424 8.728487
> head(MI.sort, 10)
               W1 W2 W1W2    N   t.score       MI
certainty_NN    6  1    1 2545 0.9976424 8.728487
destination_NN  6  1    1 2545 0.9976424 8.728487
face_NN         6  1    1 2545 0.9976424 8.728487
face_VVG        6  1    1 2545 0.9976424 8.728487
fall_VVP        6  1    1 2545 0.9976424 8.728487
hope_VVP        6  1    1 2545 0.9976424 8.728487
intellectual_JJ 6  1    1 2545 0.9976424 8.728487
leave_VVG       6  1    1 2545 0.9976424 8.728487
purely_RB       6  1    1 2545 0.9976424 8.728487
share_NN        6  1    1 2545 0.9976424 8.728487
```

> 해석

빈도 분석과 마찬가지로 연어 분석에서도 뚜렷한 특징을 찾기는 어려웠다. 그래도 몇 가지 중심어를 기준으로 해석을 해보자면,

스티브 잡스의 연설문은 다른 두 연설문과는 다르게 사람들의 삶의 방향성을 제시하는 것이 아니라 자신의 삶에 대한 이야기를 한다고 말했는데, 다른 두 연설문과 달리 life의 연어에 people가 없는 것으로 보아 이는 꽤 타당한 해석이라고 볼 수 있다. entire, do 등의 단어가 높은 연어 비율을 보이기는 하지만 단어 간의 연관성을 찾기는 쉽지 않아 보인다.

또 스티브 잡스는 다른 두 연설문과 다르게 자퇴생임에도 불구하고 대학에 대해 꽤 긍정적으로 얘기했는데 이는 college의 연어로 never이 나온 것을 보면 알 수 있다. 그는 연설문에서 대학 수업을 듣지 않았다면 '절대로 배우지 못했을 것'에 대해 매우 강조했는데, never이 이 해석을 뒷받침하고 있다. 여기에서 하나 짚고 가야할 점은, 텍스트에 대한 배경지식이 없었다면 never을 부정적인 의미로 해석했을 확률이 높았을 것이라는 점이다. 텍스트 분석에는 텍스트에 대한 이해가 반드시 필요하다는 사실을 다시 한번 확인할 수 있는 부분이다.

마지막으로 이 연설문의 특징어인 death는 어떠한 연어도 발견할 수 없었다. 데이터프레임 표를 보면 알겠지만 상위 10개 단어의 모두 t-score과 MI 값이 동일하다. 따라서 특별한 연어를 찾을 수 없었다. 좀 더 다양한 분석을 해야 할 것 같다.

3. n-gram 분석

분석 방법:

Tri.gram / Bi.gram 둘 다 분석하여 해석

*n-gram 분석 R 파일 소스 코드

- n-gram 빈도 수를 높이기 위해 모두 소문자로 치환

- 연설문에서 특수문자는 딱히 의미를 가지지 못한다고 판단하여 모두 제거 함

```r
source('./analyzer2.R')

bill_txt <- './원문/billgates.txt'
mark_txt <- './원문/zuckerberg.txt'
steve_txt <- './원문/stevejobs.txt'

ana_txt <- bill_txt

original_txt <- scan(file=ana_txt, what="char", sep=" ", quote=NULL, encoding="UTF-8")
original_txt <- tolower(original_txt)
original_txt <- gsub("[[:punct:]]", "", original_txt)
head(original_txt)

tri.gram <- paste(original_txt[1:(length(original_txt)-2)],
                  original_txt[2:(length(original_txt)-1)],
                  original_txt[3:(length(original_txt))], sep=" ")
tri.gram.Freq <- data.frame(sort(table(tri.gram), decreasing=T))
tri.gram.Freq <- data.frame(Freq=tri.gram.Freq$Freq, row.names=tri.gram.Freq$tri.gram)

bi.gram <- paste(original_txt[1:(length(original_txt)-1)],
                   original_txt[2:(length(original_txt))], sep=" ")
bi.gram.Freq <- data.frame(sort(table(bi.gram), decreasing=T))
bi.gram.Freq <- data.frame(Freq=bi.gram.Freq$Freq, row.names=bi.gram.Freq$bi.gram)

head(tri.gram.Freq, decreasing=T, 10)
head(bi.gram.Freq, decreasing=T, 13)
```

1) 빌게이츠 텍스트 n-gram 분석

```
> head(tri.gram.Freq, decreasing=T, 10)
                    Freq
in the world           6
the lives of           5
you have to            5
the millions of        4
about the millions     3
been with us           3
here in this           3
if we can              3
members of the         3
millions of people     3
> head(bi.gram.Freq, decreasing=T, 13)
             Freq
of the         23
in the         21
the world      14
we can         11
you have       10
have to         8
millions of     8
we have         7
about the       6
if we           6
lives of        6
of people       6
the worlds      6
```

> 해석

N-gram 분석도 역시 앞에서 해석한 빌게이츠가 '더 나은 세상을 만들기 위한 삶을 살자'라고 말하고 있다는 사실을 뒷받침 한다. In the world를 가장 많이 언급한 것으로 보아 세상에 대한 이야기를 하고 있고, you have to / if we can 등을 보아 무엇을 해야함을 주장하고 있다. 또한 더 나은 세상을 만들어야 한다는 근거로 수 많은 사람들이 불공평한 삶을 살고 있다고 말하고 있는데 이는 the lives of, the millions of, about the millions, millions of people 등을 통해 알 수 있다.

2) 마크 주커버그 텍스트 n-gram 분석

```
> head(tri.gram.Freq, decreasing=T, 10)
                        Freq
a sense of                 9
sense of purpose           9
many of you                5
the freedom to             5
a lot of                   4
a world where              4
has a sense                4
how many of                4
big meaningful projects    3
class of 2017              3
> head(bi.gram.Freq, decreasing=T, 17)
              Freq
the world       13
sense of        12
going to        11
a sense          9
of purpose       9
one of           9
to do            8
we can           8
i was            7
in the           7
many of          7
of you           7
our generation   7
that we          7
to create        7
to get           7
to make          7
```

> 해석

연어 분석에서 purpose의 연어로 나왔던 sense가 n-gram에서도 높은 빈도를 보이고 있는 것으로 보아, sense는 연설문인 핵심인 purpose를 따라서 나온 단어임을 알 수 있다. 즉, 주커버그는 purpose라는 단어를 쓸 때 'a sense of purpose'라는 짝을 이뤄 사용했음을 알 수 있다. 그리고 n-gram 분석만으로는 발견하기 어렵지만 꽤 높은 빈도를 보이는 trigram 'the freedom to'은 주로 뒤에 'pursue purpose' 짝을 데리고 다녔다. 즉, 단순 빈도 분석으로는 의미를 찾기 어려웠던 freedom이 목적을 추구하기 위한 '자유'임을 n-gram 분석을 통해서 알 수 있게 됐다.

3) 스티브 잡스 텍스트 n-gram 분석

```
> head(tri.gram.Freq, decreasing=T, 8)
                     Freq
one of the              4
when i was              4
a few months            3
i decided to            3
it means to             3
never graduated from    3
stay hungry stay        3
story is about          3
> head(bi.gram.Freq, decreasing=T, 17)
         Freq
it was     15
i was      13
i had      11
and i       9
in the      9
of the      9
to do       8
my life     6
to be       6
what i      6
a few       5
of my       5
that i      5
the first   5
want to     5
was the     5
when i      5
```

> 해석

When I was, I decided to, story is about, it was, I was, I had, my life 등의 n-gram 들이 빈도수가 높은 것으로 보아 앞에서 해석한 대로 스티브 잡스는 '자신의 삶의 이야기'를 하고 있음을 확신할 수 있다. 이 연설문의 명언인 'Stay Hungry, Stay Foolish'도 tri-gram에 등장하고 있다.

4. 키워드 분석

분석 방법:

모든 연설문을 TDM으로 만든 후 카이스퀘어 잔차를 이용해 키워드를 분석한다.

* 키워드 분석 R 소스코드

```r
source('./analyzer2.R')

bill_txt <- analyzer('./원문/billgates.txt')
mark_txt <- analyzer('./원문/zuckerberg.txt')
steve_txt <- analyzer('./원문/stevejobs.txt')

TDM <- data.frame(Morph=vector())
TDM <- merge(TDM, data.frame(table(bill_txt)),
             by.x = "Morph", by.y="bill_txt", all=T)
colnames(TDM)[length(TDM)] <- "bill"
TDM <- merge(TDM, data.frame(table(mark_txt)),
             by.x = "Morph", by.y="mark_txt", all=T)
colnames(TDM)[length(TDM)] <- "mark"
TDM <- merge(TDM, data.frame(table(steve_txt)),
             by.x = "Morph", by.y="steve_txt", all=T)
colnames(TDM)[length(TDM)] <- "steve"
colSums(TDM[2:length(TDM)], na.rm=T)
TDM <- data.frame(row.names=TDM$Morph,
                  TDM[2:length(TDM)])
TDM[is.na(TDM)] <- 0

CONTENT <- TDM[grep("_N|_V[^BH]|_JJ|_RB", rownames(TDM)),]

CHI <- chisq.test(TDM)$residuals
CHI <- as.data.frame(CHI)

head(CHI[order(CHI$bill, decreasing=T),], 15)
head(CHI[order(CHI$mark, decreasing=T),], 15)
head(CHI[order(CHI$steve, decreasing=T),], 15)
```

* 키워드 분석은 위 3개의 분석에서 특이점을 찾기 어려웠던 스티브 잡스의 텍스트를 중심으로 분석

- 스티브 잡스 텍스트 키워드 분석

```
> head(CHI[order(CHI$steve, decreasing=T),], 30)
                      bill       mark    steve
Apple_NP        -1.74018556 -1.9239763 4.466744
be_VBD          -1.83571054 -1.8097600 4.431223
look_VVG        -1.53469941 -1.6967876 3.939298
month_NNS       -1.53469941 -1.6967876 3.939298
life_NN         -0.98453255 -2.1002975 3.819133
I_PP            -3.06576715 -0.1799344 3.770683
very_RB         -0.88426410 -2.1270358 3.737469
Stay_NP         -1.42085556 -1.5709201 3.647081
later_RBR       -1.64066268 -1.2626568 3.507312
it_PP           -1.78168047 -1.1120822 3.477906
my_PP$          -1.62357519 -1.2324266 3.448973
decide_VVD      -1.29705774 -1.4340473 3.329314
drop_VVN        -1.29705774 -1.4340473 3.329314
college_NN      -1.73742662 -1.0068665 3.292437
death_NN        -0.88310601 -1.6967876 3.186718
@card@_CD       -1.60338206 -0.9505408 3.065691
out_RP          -1.83169223 -0.6983580 3.007356
'_POS           -1.16012371 -1.2826509 2.977829
die_VV          -1.16012371 -1.2826509 2.977829
dot_NNS         -1.16012371 -1.2826509 2.977829
fire_VVN        -1.16012371 -1.2826509 2.977829
heart_NN        -1.16012371 -1.2826509 2.977829
love_VVD        -1.16012371 -1.2826509 2.977829
have_VHD         0.04672904 -2.3497185 2.946545
as_RB           -1.42085556 -0.9343505 2.834202
calligraphy_NN  -1.00469660 -1.1108082 2.578876
close_JJS       -1.00469660 -1.1108082 2.578876
Foolish_NP      -1.00469660 -1.1108082 2.578876
Hungry_NP       -1.00469660 -1.1108082 2.578876
love_NN         -1.00469660 -1.1108082 2.578876
```

> 해석

스티브 잡스는 Apple의 창립자 였기 때문에 Apple의 상대빈도는 당연히 높다. 특이한 점은 I ,be 동사, my 는 원래 영단어에서 빈도수가 높게 나타나는 단어인데 잡스의 연설문이 다른 두 연설문 보다 훨씬 더 많이 등장했다. 자신의 이야기를 했음을 확신하게 하는 부분이다. death와 die 빈도도 가장 높은 것으로 보아 잡스가 죽음에 관한 이야기를 했음을 알 수 있다. 마지막으로 love, heart 등은 주로 '좋아하는 일을 하라' 라는 식의 텍스트에서 주로 찾을 수 있는 단어들인데 빌게이츠와 마크 주커버그가 '개인' 보다는 '세상'을 위한 삶을 살라고 강조하느라 이 단어들을 많이 쓰지 않았고 잡스는 개인의 삶을 이야기 하다 보니 이 단어들을 상대적으로 더 많이 사용했다. 실제로 잡스는 '남의 인생을 사느라 인생을 허비하지 말고 내가 좋아하는 일을 하라'라고 강조했다.