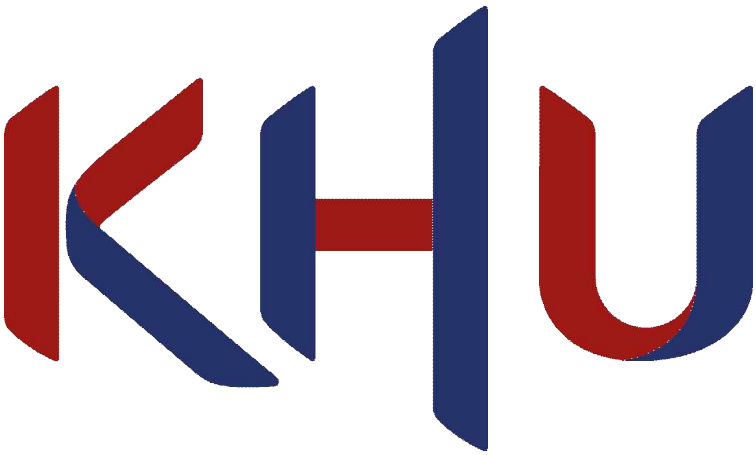
데이터 분석을 통한
성공적 기업 제출용 자기소개서 작성법 탐구



과목명 : 경영프로그래밍
교수님 : 양성병 교수님
<4조>

경영학과 박규리
(크롤링 코딩, 5대기업 빈도표 분석, 비율분석 코딩, 보고서 작성)
경제학과
(연관분석 코딩, 5대기업 빈도표 분석, 보고서 작성, 크롤링)
경영학과
(형태소 분석 및 텀 다큐메트릭스 코딩, 비율분석 코딩, 연관 분석)
경영학과
(보고서 서론 및 결론 작성, 5대 기업 빈도표 분석, 비율 분석, 크롤링)

<목 차>

Ⅰ 서론

1. 연구의 배경

요즘 우리 사회의 청년들은 갈수록 높아지는 취업난으로 인해 많은 어려움을 겪고 있다. 최근 발표된 통계청 고용지표에 따르면 경제활동인구 대비 전체 실업률 4.2%이며, 청년 실업률은 무려 11.2%로 지속적 증가세를 보이고 있다. (2017.4.)

이처럼 최근 우리 사회 청년실업이 무엇보다 심각하고, 이는 국가적 경제 손실에 큰 영향을 끼친다. 그리고 '청년 실업과 일자리' 라는 어려운 시대적 상황에 직면한 우리 젊은이들이 사회로 나가기 전부터 이런 문제에 좌절한다는 것이 지금의 현실이다. 실제로 작년 하반기 채용에서 서류 전형을 평균 4~6개(41.2%) 지원한 구직자 중 '모두 탈락했다'(합격률 0%) 라고 응답한 사람이 33.9%로 가장 높았다. (취업포털 커리어, 2016.11.)

그렇다면 이와 같이 어려운 취업시장에서 구직자들이 생각하는 '취업을 위해 가장 신경 쓰는 부분'은 무엇일까? 구직자 657명을 대상으로 한 조사에서 35.6%가 '자기소개서 작성'이라고 답했으며 '직무 관련 자격증 취득(26%)', '인턴 및 사회 경험(24.7%)'이 각각 그 뒤를 이었다.

이처럼 우선 많은 학생들이 취업 과정에서 자기소개서 작성에 큰 관심을 가지고 있다는 것이다. 또한 자기소개서 소개는 자기PR과 주도적, 능동적 학습 및 내면 성장의 과정을 중시하는 현재 사회의 기조에 부합하는 중요한 절차임에 틀림없다. 한편 기업 입장에서도 자기소개서는 그 조직의 구성원으로써 갖춰야할 가치관과 행동규범 등을 검증할 수 있는 가장 중요한 항목일 것이다.

2. 문제 인식과 탐구 방법

이 학습을 통해 기존의 여러 자기소개서에서 나타나는 전체적인 구성과 그 흐름을 먼저 살펴볼 필요가 있다. 성공적인 자기소개서 작성에 필요한 수법에는 전문가의 상담 및 조언, 이성

과 감성의 적절한 조화, 작문과 소통의 피드백, 적절한 Text나 어휘의 구사능력 등이 있다. 이처럼 자기소개서는 복합장르의 성격으로써 여러 가지 가치가 혼재하고 있다.

여기서 우리는 '객관적, 계량적 데이터 분석 및 수집의 R 활용' 에 초점을 맞추어 보도록 하겠다. 실제로 여러 합격 자기소개서와 불합격 자기소개서의 예시를 통해 유효한 어휘를 추출하고 그 의미를 해석해보도록 한다. 이는 구체적으로 취업 서류전형에서 합격한 자기소개서에 나타난 상위 빈도를 가려내어, 그 어휘 분석을 통해 성공적인 자기소개서의 흐름에 적합한지를 먼저 살펴보도록 한다. 한편 특정 어휘가 각 기업의 이미지를 어떻게 차별화하고 있는지 살펴본다. 또한 성공적인 자기소개서의 사례와 이와 대조적으로 미흡한 부분의 경우를 비교하여 중요 단어 활용 및 구문 형성에 있어 어떤 차이가 있었는지 확인해보도록 한다.

나아가 이 연구의 경험적 분석 결과를 통해, 취업에 필요한 자기소개서 작성 시 적절한 어휘를 많이 포함한 몇 가지 구문을 만들어봄으로써 보다 높은 적중률에 도움이 될 수 있도록 하겠다.

본 학습은 취업 서류전형에서 합격한 자기소개서에 나타난 어휘를 기본 데이터로 활용했다. 그래서 대표적으로 그 의미를 찾을 수 있는 자기소개서 취업정보공유 사이트인 J코리아, I사를 표본 수집 경로로 선택했다. 한편, 학습 목적은 감성적 의견을 표현하는 글쓰기가 아닌 '기업 제출용' 소개서에 초점을 맞췄다. 기업 제출용 소개서는 개인의 이성적 혹은 감성적 의견을 표현하는 방법보다 능력이나 직무 적합성, 조직 융합 등 채용에 타당한 적합한 이유를 설명하는 글일 것이다. 이에 실제 기업 제출용 소개서를 직접 살펴보면 형태소 분석 결과 '일반 명사'(47.6%)가 가장 많은 분포를 하고 있음을 알 수 있었다.(세종코퍼스· 말뭉치연구 , 강범모 1999) 때문에 이 학습에서는 우선적으로 '명사'의 활용을 중점적으로 분석· 검토하고 실제 그 어휘의 분포를 알아보고자 했다.

또한 많은 학생들이 30대 주요 그룹사 및 대기업(47.1%)에 큰 관심을 보이고 지원한다는 현실을 고려하여, 대상 기업으로는 국내 주요 대기업으로 일컬어지는 상위 5개의 기업(삼성, 현대, LG, SK, 롯데 등 * 이하 계열회사 포함)을 합격 자기소개서의 대상으로 선정하였다. 그리고 불합격 자기소개서 비교 데이터는 I사의 자기소개서 첨삭을 요청하는 자료을 근거로, 이를 합격의 범위에서 벗어난다고 가정하였고, 그 표본으로 선정하였다.

Ⅱ 본론

1. R을 이용한 데이터 분석 과정

*Frame Work

데이터 크롤링 → Rhino2.5.3을 이용한 형태소 분석 → 기업별/분류별 빈도표 작성 → 분류별 단어 비율분석 표 작성 → 분류별 Term document matrix 작성 → 연관분석
→ Excel을 이용한 빈도 분석/ R을 이용한 연관분석과 비율분석

1-1 데이터 크롤링

자기소개서를 분석하기 위해 J사에 있는 자기소개서를 R을 통해 크롤링 하였다.

먼저, 기업그룹별로 크롤링을 한 소스이다. 5개 그룹(LG, 삼성, SK, 현대, 롯데)을 크롤링하

였으며, 5개 그룹 모두 같은 원리로 크롤링을 하였으므로 대표적으로 LG의 코드만 첨부하였다.

```r
library(XML) #패키지 설치
library(stringr) # 패키지 설치
library(rvest) # 패키지 설치

setwd("c:\\r_temp") #디렉토리 설정
getwd()

all.url<-c() # url이 잘 뽑혔는지 확인하기 위해 url 넣을 곳 설정
all.text  <- c() # 자기소개서 텍스트 넣을 곳 설정

urlall <- "http://www.jobkorea.co.kr/Starter/PassAssay/View/" # url 앞부분 변수에 넣기
for(EEE in 167023:189999){ # 페이지번호를 돌려서 페이지를 추출, EEE가 페이지번호
  url <- paste(urlall,EEE,"?Page=1&OrderBy=0&FavorCo_Stat=0&Pass_An_Stat=0",sep="") #페이지 주소 전문이 되게 합치기
  jasoju <- read_html(url) # 페이지 소스 불러오기
  if(str_detect(jasoju,"선택하신 합격자소서 정보가 없습니다.")==T){ # 페이지가 없으면 next
    next
  }else{
    b<-html_nodes(jasoju, 'title') # 페이지 소스에서 <title>에 있는 코드 b로 넣기
  }
  if(str_detect(b,'LG' )==F){ # <title>에 LG가 없으면 next, 있으면 추출
    next
  }else{
    tx <- html_nodes(jasoju, "div.tx") #   <div class="tx"> 코드 안에 있는 글귀 tx 변수에 넣기

    text <- html_text(tx[!str_detect(tx, "href")])  # <div class="tx"><a href (...)로 시작하는 코드에 있는 글귀 빼기기
    text <- str_replace_all(text,"글자수","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"아쉬운점","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"좋은점","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"1\r","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"2\r","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r\n","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte\r\n","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r","")# 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte","")# 자소서에 관련없는 단어 빼기
  }
  all.text<-rbind(all.text,text) # 자소서로 추출한 텍스트 all.text에 합치기
  all.url<-c(all.url,url) # url 합치기
}
all.text<-str_replace_all(all.text,"text","") # 자소서에 관련없는 단어 빼기
write.csv(all.text,'LG.csv') # 디렉토리에 csv로 저장
```

J사의 페이지주소를 확인해 본 결과, 아래의 주소패턴에서 '**숫자**'부분을 바꾸면 페이지가 바뀌는 것을 확인하고 for문을 통해 그곳에 임의의 숫자간격을 주어 크롤링 하였다.

'http://www.jobkorea.co.kr/starter/passassay/View/**숫자**Page=1&OrderBy=0&FavorCo_Stat=0&Pass_An_Stat=0'

그 후에 페이지가 없으면 뜨는 문구를 확인하고 그 문구가 뜨면 next로 넘기게 하였다. 그리고 에 자기소개서 내용이 있는 것을 확인한 후 그 코드 안에 있는 글자를 넣고 필요없는 단어를 뺀 후 하나의 변수에 합쳤다. 아래는 실제 홈페이지 소스코드 중 한 부분이며 에 자기소개서 내용이 들어있는 것을 확인할 수 있다.

```
          </dd>
          <dt class="on">
            <button type="button"><strong class="skip">질문</strong><span class="num">Q2.</span><span class="tx">성격의 장·단점 및 생활신조</span><span class="arr
stSplmg">보기</span></button>
          </dt>
          <dd class="show">
            <strong class="skip">답변</strong>

            <div class="tx">
<b class='good'>"사람을 알고, 일을 알고, 함께라는 의미를 실천하는 알곡이"<br/><br/>하고자 하는 일은 꼭 해봐야 하는 성격입니다. '안 하고 후회하는 것보다, 하면서 즐기고 배우자'라는 말을 가슴속에 기억하고 살고 있습니다. <span class='sup'>좋은점 1</span></b><br/> <br/><b class='bad'>대학시절 동아리 활동의 장을 하면서 하나의 써클이 아닌 학교와 지역의 대표라고 생각을 하며 동아리를 이끌었습니다. 활동적인 동아리속에서도 내부적인 활동도 구체화 시키면서 동아리의 입지를 잡아갔습니다. 또한 동아리 연합회 활동을 통해 전국적으로 인맥을 중요시하며 서로에게 도움이 되는 삶을 살고 있습니다. 저 뿐만 아니라 회원 개인의 장단점의 특성을 알고 이끌었습니다. 처음 보는 상황에서도 만나는 사람, 분위기에 맞추어 서로에게 도움이 될 수 있는 사람이 되는 것이, 동아리 활동보다 더 중요한 우리의 태도를 후배들에게 일깨워 주었습니다. <span class='sup'>아쉬운점 1</span></b>                <p class="txSplChk"><span>글자수 <strong>420</strong>자</span><span><strong>724</strong>Byte</span></p>
            </div>
                <div class="advice">
                  <p><strong class="good stSpBefore stSpAfter">좋은점 1</strong> 첫 항목에서 제시한 도전과 연관이 있는 내용이며, 문장도 좋습니다</p>
                  <p><strong class="bad stSpBefore stSpAfter">아쉬운점 1</strong> 항목에서 요구한 성격의 장단점 및 생활신조에 대해 명쾌하게 내용을 제시하지 않은 점이 아쉽습니다.</p>
                </div>
          </dd>
```

그리고 페이지 소스에서 <title>에 기업명이 있는 것을 확인하고 그곳에 LG가 없으면 next로 넘기고 있으면 추출을 하였다.

다음은 <title>에 기업명이 있음을 확인할 수 있는 J사의 자기소개서 페이지 소스코드이다.

```
<!DOCTYPE html>
<html lang="ko">
<head>
   <!-- Meta Info -->

<meta charset="utf-8" />
<meta http-equiv="X-UA-Compatible" content="IE=edge" />
<title>LG전자㈜ 합격자소서 | 잡코리아 신입공채</title>
<link rel="SHORTCUT ICON" href="http://www.jobkorea.co.kr/favicon.ico">


<meta name="title" content="LG전자㈜ 합격자소서 | 잡코리아 신입공채">
<meta name="writer" content="잡코리아">
<meta name="description" content="LG전자㈜ 2017년 상반기 신입 제품·서비스영업 합격자소서">
<meta name="keywords" content="LG전자㈜ 합격자소서, LG전자㈜ 자소서, 합격자소서, 자기소개서, 합격자기소
<meta name="verify-v1" content="wfOOCE9Vtx+Z5etOXJnS9LU03yGpBxkK74T/yU63Xqs=">
```

```
if(str_detect(b,'LG'  )==F){
    next
 }
```

다른 기업그룹을 뽑을 때는 이 코드에서 'LG'를 그 기업명으로 바꾸어 크롤링을 하였다. 5대 기업은 자기소개서 한 문항의 내용을 하나의 단위로 보는 것이 좋겠다고 생각하여 각 자소서를 하나로 묶는 작업은 하지 않았다.

이어서 각각의 기업의 자기소개서 문항 질문을 크롤링하였다. 기업별로 추출한 데이터 중에서 문항에 의해 특별히 더 빈번하게 사용되는 단어는 없는 지 점검하기 위해 크롤링을 수행했다. 대표적으로 현대그룹의 자기소개서 문항의 질문을 크롤링한 코드를 첨부하였다.

```
library(XML) #패키지 설치
library(stringr) # 패키지 설치
library(rvest) # 패키지 설치

setwd("c:\\r_temp") #디렉토리 설정
getwd()

all.url<-c()
all.text  <- c()

urlall <- "http://www.jobkorea.co.kr/Starter/PassAssay/View/"  # url 앞부분 변수에 넣기
for(EEE in 144296:145579){ # 페이지번호를 돌려서 페이지를 추출, EEE가 페이지번호
  url <- paste(urlall,EEE,"?Page=1&OrderBy=0&FavorCo_Stat=0&Pass_An_Stat=0",sep="") #페이지 주소 전문이 되게 합치기
  jasoju <- read_html(url) # 페이지 소스 불러오기
  if(str_detect(jasoju,"선택하신 합격자소서 정보가 없습니다.")==T){ # 페이지가 없으면 next
    next
  }else{
    b<-html_nodes(jasoju, 'title') # 페이지 소스에서 <title>에 있는 코드 b로 넣기
  }
  if(str_detect(b,'현대')==F){  # <title>에 현대가 없으면 next, 있으면 추출
    next
  }else{
    tx <- html_nodes(jasoju, "span.tx") # <span class="tx"> 코드 안에 있는 글귀 tx 변수에 넣기
    text <- html_text(tx[!str_detect(tx, "href")])  # <div class="tx"><a href (...)로 시작하는 코드에 있는 글귀 빼기
    text <- str_replace_all(text,"글자수","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"아쉬운점","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"좋은점","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"1\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"2\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r\n","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte\r\n","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"text","") # 자소서에 관련없는 단어 빼기
  }
  all.text<-rbind(all.text,text) # 자소서로 추출한 텍스트 all.text에 합치기
  all.url<-c(all.url,url) # url 합치기
}
all.text<-str_replace_all(all.text,"text","") # 자소서에 관련없는 단어 빼기
write.csv(all.text,'현대 질문.csv') # 디렉토리에 csv로 저장
```

대부분의 코드는 위의 것과 같은 원리이며, 자기소개서 문항 질문이 홈페이지 코드 중에 있는 것을 확인하고 그곳을 크롤링 하였다.

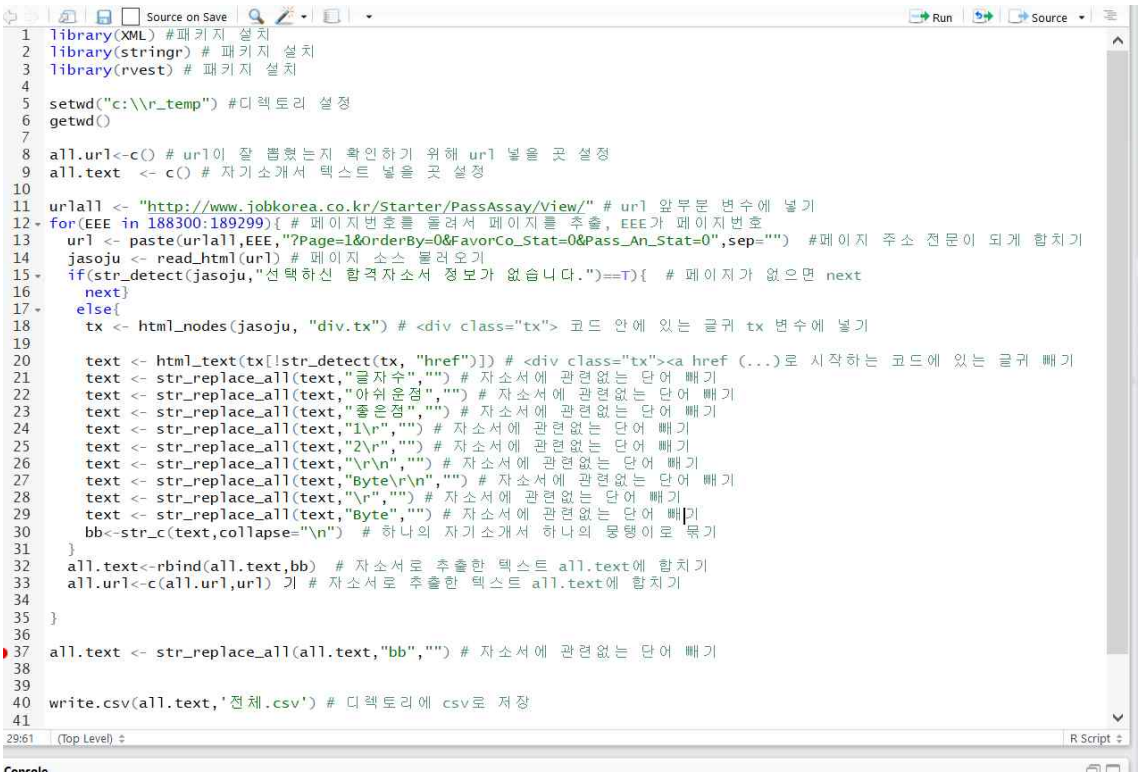
oBefore stSpAfter">좋은점 2 전공이 지원직무에서 어떻게 쓰일 수 있는지 적극적으로 제시한 부분이 좋고 판매나 영업 경험을 제시한 점이 호감을 주고 있습니다.</p>

'>질문Q2.본인이 지원한 직무관련 향후 계획에 대하여<span

과정에 있어서는 상권 분석이나 조직 관리와 같은 분야도 중요하지만, 최근 급변하는 시장에서는 고객이 정말로 원하는 것이 무엇인지 무리 좋은 제품이 출시되더라도, 고객들의 니즈를 파악하지 못하면 시장에서 결코 성공할 수 없기 때문입니다. 이를 위해 현장에서 직·이에 상생의 다리를 놓아줄 수 있는 사람이 되고자 합니다.좋은점 1

도 있으며, 고객의 목소리를 알린 경험을 통해 누구보다 고객들이 원하는 바를 잘 알고 있습니다. 이를 바탕으로 고객의 니즈를 적시에 파

둘째, 합격 자소서와 첨삭 자소서를 비교하기 위해 산업, 직업별 구분이 없는 모든 분야의 자소서를 크롤링하기 위한 코드이다.



```
library(XML) #패키지 설치
library(stringr) # 패키지 설치
library(rvest) # 패키지 설치

setwd("c:\\r_temp") #디렉토리 설정
getwd()

all.url<-c() # url이 잘 뽑혔는지 확인하기 위해 url 넣을 곳 설정
all.text  <- c() # 자기소개서 텍스트 넣을 곳 설정

urlall <- "http://www.jobkorea.co.kr/Starter/PassAssay/View/" # url 앞부분 변수에 넣기
for(EEE in 188300:189299){ # 페이지번호를 돌려서 페이지를 추출, EEE가 페이지번호
  url <- paste(urlall,EEE,"?Page=1&OrderBy=0&FavorCo_Stat=0&Pass_An_Stat=0",sep="")  #페이지 주소 전문이 되게 합치기
  jasoju <- read_html(url) # 페이지 소스 불러오기
  if(str_detect(jasoju,"선택하신 합격자소서 정보가 없습니다.")==T){  # 페이지가 없으면 next
    next}
   else{
    tx <- html_nodes(jasoju, "div.tx") # <div class="tx"> 코드 안에 있는 글귀 tx 변수에 넣기

    text <- html_text(tx[!str_detect(tx, "href")]) # <div class="tx"><a href (...)로 시작하는 코드에 있는 글귀 빼기
    text <- str_replace_all(text,"글자수","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"아쉬운점","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"좋은점","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"1\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"2\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r\n","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte\r\n","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"\r","") # 자소서에 관련없는 단어 빼기
    text <- str_replace_all(text,"Byte","") # 자소서에 관련없는 단어 빼기
    bb<-str_c(text,collapse="\n")  # 하나의 자기소개서 하나의 뭉탱이로 묶기
  }
  all.text<-rbind(all.text,bb)  # 자소서로 추출한 텍스트 all.text에 합치기
  all.url<-c(all.url,url) # 자소서로 추출한 텍스트 all.text에 합치기

}

all.text <- str_replace_all(all.text,"bb","") # 자소서에 관련없는 단어 빼기


write.csv(all.text,'전체.csv') # 디렉토리에 csv로 저장
```

전체 자기소개서는 하나의 지원자가 쓴 자소서의 문단을 하나의 단위로 분석하는 것이 더 바람직하다고 판단하여 str_c로 하나의 단위로 묶었다.

*J사이트에서는 첨삭 자소서를 크롤링 할 수 없어, 저작권의 문제가 없는 I 사이트에서 200여 개의 첨삭을 요청한 자소서를 copy$paste 방법으로 데이터를 수집하였다.

1-2 Rhino 2.5.3을 이용한 형태소 분석

```
setwd("C:\\Users\\myungjun\\Desktop\\명준\\2017-1\\경영프로그래밍\\project")
library(stringr)
#크롤링 한 파일을 Rhino를 이용하여 형태소단위로 분석하여 '명사'만 가져온다.
initRhino <- function(path)
{setwd(paste(path, "/RHINO", sep=""))
  if(!require(rJava)) {install.packages("rJava"); library(rJava)}
  .jinit()
  .jaddClassPath(paste(path, "/RHINO", sep=""))
  .jclassPath()
  RHINO <- .jnew("rhino/RHINO")
  .jcall(RHINO, returnSig = "V", "ExternInit")
  return(RHINO)
}
getMorph <- function(sentence, type)
{ result <- .jcall(RHINO, returnSig = "S", "getMorph", sentence, type)
  Encoding(result) <- "UTF-8"
  resultVec <- unlist(strsplit(result, '\r\n'))
  return(resultVec)
}

#[2] Setting path & Initializing RHINO
path <- "C:/Users/myungjun/Documents/R/WORK/"
RHINO <- initRhino(path)
#[3] Analyze all files in ./RHINO2.5.3/WORK/RHINO/_input/
setwd(paste(path, "/RHINO", sep=""))
.jcall(RHINO, returnSig = "V", "analyzingText_rJava", "N")
print("Created result.txt in ./RHINO2.5.3/WORK/RHINO/")
```

크롤링하여 수집한 데이터를 RHINO/input 폴더에 넣고 위의 코드를 실행하면 형태소 단위로 문장을 분석하여 단어를 나열한 결과가 RHINO/output폴더에 출력된다.

1-3 기업별/분류별 빈도표 작성

크롤링과 마찬가지로 5대 기업의 코드는 기업이름을 제외하고 모두 같으므로, 여기서는 삼성의 코드를 보면서 설명하고자 한다.

```
samsung<-readLines("samsunglist.txt", encoding ="UTF-8")#RHINO를 거친 out파일을 불러온다.
splt_sm<-str_split(samsung," ")#','로 연결되어 이루어진 samsung을 split한다.
sort.sm<-sort(table(splt_sm), decreasing = TRUE)#단어들을 table화하여 빈도표로 만들고 순서를 부여한다.
write.csv(sort.sm,"samsung_freq_result.csv", row.names = FALSE)#명사빈도표 생성
```

뽑아낸 빈도표에서 '저', '저희' 등 의미가 없는 단어들(stopword)들을 발견했고, 이 단어들을 제거하기 위해 의미 없는 단어들을 모아 놓은 사이트(http://www.ranks.nl/stopwords/korean)에서 단어들을 추출해 비교하여 제거하였다.

```
smg<-read.csv("samsung_freq_result.csv", stringsAsFactors = F)
stopword<-readLines("stopwords.txt", encoding = "UTF-8")
a<-smg$splt_sm
aa<-c()
for(i in 1: length(stopword)){
  dd<-grep(paste0("^",stopword[i],"$"),a)
  aa<-c(aa,dd)
}
smg<-smg[-aa,]
write.csv(smg,"samsung_final.csv", row.names = F)
```

paste0를 이용하여 stopword의 I번째 단어를 추출했고 for문과 grep을 이용하여 그것이 삼성 빈도표의 어느 위치에 자리하고 있는지를 뽑아낸 다음, 그 위치를 제한 빈도표만을 뽑아내어 samsung_final.csv로 저장하였다.

1-2 Rhino 2.5.3을 이용한 형태소 분석

```
setwd("C:\\Users\\myungjun\\Desktop\\명준\\2017-1\\경영프로그래밍\\project")
library(stringr)
#크롤링 한 파일을 Rhino를 이용하여 형태소단위로 분석하여 '명사'만 가져온다.
initRhino <- function(path)
{setwd(paste(path, "/RHINO", sep=""))
  if(!require(rJava)) {install.packages("rJava"); library(rJava)}
  .jinit()
  .jaddClassPath(paste(path, "/RHINO", sep=""))
  .jclassPath()
  RHINO <- .jnew("rhino/RHINO")
  .jcall(RHINO, returnSig = "V", "ExternInit")
  return(RHINO)
}
getMorph <- function(sentence, type)
{ result <- .jcall(RHINO, returnSig = "S", "getMorph", sentence, type)
  Encoding(result) <- "UTF-8"
  resultVec <- unlist(strsplit(result, '\r\n'))
  return(resultVec)
}

#[2] Setting path & Initializing RHINO
path <- "C:/Users/myungjun/Documents/R/WORK/"
RHINO <- initRhino(path)
#[3] Analyze all files in ./RHINO2.5.3/WORK/RHINO/_input/
setwd(paste(path, "/RHINO", sep=""))
.jcall(RHINO, returnSig = "V", "analyzingText_rJava", "N")
print("Created result.txt in ./RHINO2.5.3/WORK/RHINO/")
```

크롤링하여 수집한 데이터를 RHINO/input 폴더에 넣고 위의 코드를 실행하면 형태소 단위로 문장을 분석하여 단어를 나열한 결과가 RHINO/output폴더에 출력된다.

1-3 기업별/분류별 빈도표 작성

크롤링과 마찬가지로 5대 기업의 코드는 기업이름을 제외하고 모두 같으므로, 여기서는 삼성의 코드를 보면서 설명하고자 한다.

```
samsung<-readLines("samsunglist.txt", encoding ="UTF-8")#RHINO를 거친 out파일을 불러온다.
splt_sm<-str_split(samsung,", ")#','로 연결되어 이루어진 samsung을 split한다.
sort.sm<-sort(table(splt_sm), decreasing = TRUE)#단어들을 table화하여 빈도표로 만들고 순서를 부여한다.
write.csv(sort.sm,"samsung_freq_result.csv", row.names = FALSE)#명사빈도표 생성
```

뽑아낸 빈도표에서 '저', '저희' 등 의미가 없는 단어들(stopword)들을 발견했고, 이 단어들을 제거하기 위해 의미 없는 단어들을 모아 놓은 사이트(http://www.ranks.nl/stopwords/korean)에서 단어들을 추출해 비교하여 제거하였다.

```
smg<-read.csv("samsung_freq_result.csv", stringsAsFactors = F)
stopword<-readLines("stopwords.txt", encoding = "UTF-8")
a<-smg$splt_sm
aa<-c()
for(i in 1: length(stopword)){
  dd<-grep(paste0("^",stopword[i],"$"),a)
  aa<-c(aa,dd)
}
smg<-smg[-aa,]
write.csv(smg,"samsung_final.csv", row.names = F)
```

paste0를 이용하여 stopword의 I번째 단어를 추출했고 for문과 grep을 이용하여 그것이 삼성 빈도표의 어느 위치에 자리하고 있는지를 뽑아낸 다음, 그 위치를 제한 빈도표만을 뽑아내

어 samsung_final.csv로 저장하였다.

1-4 Term document matrix만들기

비정형데이터분석을 하기 위해 term document matrix 형태를 만든다. term document matrix는 전체 합격 자기소개서와 첨삭자기소개서의 비교를 위해 작성되었으므로 여기서는 첨삭자기소개서의 코드를 예를 들어 설명한다.

```
##Term-document-matrix만들기
com_list<-read.csv("edit_raw.csv", header = FALSE, stringsAsFactors = FALSE)#자소서 원문
# Term-document-matrixrow
nrow(com_list)#자소서의 갯수
for(i in 1:nrow(com_list)) {
  write.table(com_list$V1[i], paste0(i,"cf_com.txt"), row.names = FALSE)
}## 전체 자소서를 개별 자소서로 분리
# 해당 개별 자소서를 RHINO를 이용하여 각각의 명사 파일 추출해야 함
```

개별 자기소개서를 RHINO를 돌려서 명사를 추출하기 위해 'i'cf_com.txt의 파일을 만들고 그 파일을 다시 RHINO를 이용하여 명사를 추출한다.

```
##RHINO실행
#RHINO로 형태소분석을 마친 각각의 명사 파일을 다시 wd로 이동시키고 불러온다.
#wd로 파일을 옮기기 위해 rstudio를 종료시키고 파일을 옮긴 후 다시 edit.R실행
setwd("C:\\Users\\myungjun\\Desktop\\명준\\2017-1\\경영프로그래밍\\project\\edit")
library(stringr)
```

RHINO분석을 마친 결과를 working directory로 옮기고 다시 r파일을 실행한다.

```
## 가상의 term document matrix column
data1<-read.csv("edit_final.csv")#자소서의 명사 빈도표
com_list<-read.csv("edit_raw.csv", header = FALSE, stringsAsFactors = FALSE)#자소서 원문
term_mat<-matrix(NA, nrow = nrow(com_list), ncol = nrow(data1))
#행의 갯수를 자소서 갯수, 열의 갯수를 사용 단어의 갯수로 생성
tt<-data1$splt_edit#열의 이름에 사용된 명사들 부여
colnames(term_mat)<-tt##매트릭스 열 생성
```

Term docu matrix(tdm)을 만들기 위해 행의 개수를 자기소개서의 개수로, 열의 개수를 사용 단어의 개수로 지정하고, 그 값을 NA로 채운 그릇을 생성한다. 그리고 그 생성된 가상의 그릇의 term_mat에 열 이름을 명사들로 부여한다.

```
word_list<-readLines("1cf_com.txt", encoding = "UTF-8")
for(i in 2:nrow(com_list)){
  temp_f<-readLines(paste0(i,"cf_com.txt"), encoding = "UTF-8")
  word_list<-rbind(word_list,temp_f)
}
#word_list라는 객체에 개별 자소서를 불러온다.
## matrix 내용 채우기
for (i in 1 : nrow(com_list)) {
  temp.wl<-word_list[i]
  temp.split<-str_split(temp.wl, " ")
  temp.unlist<-unlist(temp.split)
  assign(paste0("cl_un_st_com_",i),temp.unlist)
}## 각 자소서의 명사를 cl_un_st_com_'i'에 할당
```

word_list라는 객체에 RHINO를 이용하여 뽑아낸 명사들을 불러온다. 그리고 1-3과 같은 방법을 이용하여 word_list에 들어있는 개별 자기소개서의 명사들을 분리하고 개별 자기소개서의 명사를 각각의 개체에 할당한다.

```
# 반복문을 통해 term document matrix 생성
for (i in 1:nrow(com_list)){
  for (j in 1:length(colnames(term_mat))) {
    term_mat[i,j]<-length(get(paste0("cl_un_st_com_",i))[str_detect(get(paste0("cl_un_st_com_",i)),paste0("^",colnames(term_mat)[j],"$"))]
  }
}
write.csv(term_mat, "edittdm.csv", row.names = F)
```

이중 for문을 이용하여 I번째 행에 있는 자기소개서의 명사들을 열과 비교하여 있으면 그 개수를 표시한다. 최종 단계를 마쳐 "edittdm.csv"파일로 저장한다.

1-5 비율분석표 만들기

```
setwd("C:\\r_temp") #디렉토리 선정
getwd()

passedit<-read.csv('passandedit.csv',header=T,sep='') #합격자소서와 첨삭자소서 불러오기
View(passedit) # 한번 보기
pass<-passedit[c(1,2)] #합격자소서 데이터만 pass에 넣기(단어,빈도수)
names(pass)[2]<-'Freq' #빈도수 Fred로 이름바꿔주기
pass<-subset(pass,Freq>=30) #합격자소서 단어 빈도수 30이상인 것만 정렬하기

edit<-passedit[c(3,4)] #첨삭자소서 데이터만 edit에 넣기

edit<-subset(edit,edit>=30) #첨삭자소서 단어 빈도수 30이상인 것만 정렬하기
names(pass)[1]<-'name' #합격자소서 단어 정렬깆기한 열 이름 name으로 바꿔주기
names(edit)[1]<-'name' #첨삭자소서 단어 정렬깆기한 열 이름 name으로 바꿔주기
pass_edit_plus<-merge(pass,edit) # 합격자소서와 첨삭자소서 합치기
View(pass_edit_plus) # 한번 보기
names(pass_edit_plus)[2]<-'pass' #합격자소서 빈도수 'pass'로 바꾸기

sum(pass_edit_plus[2]) #합격자소서 빈도수 다 합치기
sum(pass_edit_plus[3]) #첨삭자소서 빈도수 다 합치기

pass_edit_plus[4]<-pass_edit_plus[2]/sum(pass_edit_plus[2])
pass_edit_plus[5]<-pass_edit_plus[3]/sum(pass_edit_plus[3])

pass_rate<-pass_edit_plus[4]/(pass_edit_plus[4]+pass_edit_plus[5])*100 #합격자소서 빈도수의 비율을 분자로 합격자소서와 첨삭자소서 비율 총 합친 것 분모로
edit_rate<-pass_edit_plus[5]/(pass_edit_plus[4]+pass_edit_plus[5])*100  #첨삭자소서 빈도수의 비율을 합격자소서와 첨삭자소서 비유 총 합친 것 분모로

pass_edit_plus<-pass_edit_plus[-c(4,5)] #의도하지 않은 비율구한 것은 지우기
pass_edit_plus<-cbind(pass_edit_plus, c(pass_rate, edit_rate)) #합격자소서의 비율과 첨삭자소서 비율끼리 비교한 자료 더하기
names(pass_edit_plus)[4]<-'pass_rate'#합격자소서 비율 열 이름 바꿔주기
names(pass_edit_plus)[5]<-'edit_rate'#첨삭자소서 비율 열 이름 바꿔주기

View(pass_edit_plus)
write.csv(pass_edit_plus,'pass_edit_rate.csv', row.names = F) #추출
```

합격자소서와 첨삭자소서를 비교하기 위해 비율분석을 해보았다.

합격자기소개서 해당 단어 빈도수 / 합격 자기소개서 전체단어빈도수
첨삭자기소개서 해당 단어 빈도수 / 첨삭 자기소개서 전체단어빈도수

이러한 방식으로 구했다. 하지만 합격자기소개서와 첨삭자기소개서의 상대적인 비율을 비교하기에는 너무 숫자가 작았다. '가족'이라는 단어로 예를 들어 보면 합격 자기소개서에는 0.079%가 나오고 첨삭자기소개서에는 0.25%가 나온다. 이 비율을 가지고는 비교하기 어렵다고 판단하였다. 이 숫자로 미루어보면, 합격 자기소개서에는 '가족'이라는 단어가 10000개 중에 7.9개, 첨삭 자기소개서에는 10000개 중에 25개라고 볼 수 있다. 이러한 원리를 가지고 다시 공식을 만들어보았다.

합격자기소개서와 첨삭자기소개서의 상대비율인 pass_rate과 edit_rate 구하는 식은 다음과 같다.

합격 비율 = 합격 자소서의 해당 단어의 단어 빈도수 / 합격 자소서의 전체 단어 빈도수
첨삭 비율 = 첨삭 자소의 해당 단어의 단어 빈도수 / 첨삭 자소서의 전체 단어 빈도수

pass_rate = 합격비율 / (합격 비율 + 첨삭 비율)
edit_rate = 첨삭비율 / (합격 비율 + 첨삭 비율)

이러한 식으로 다음과 같은 표가 도출되었다.

	name	pass	edit	pass_rate	edit_rate
1	가족	33	61	24.09913	75.90087
2	가치	195	51	69.17446	30.82554
3	갈등	66	31	55.54662	44.45338
4	강점	75	31	58.67667	41.32333
5	개발	276	50	76.41363	23.58637
6	개인	102	49	54.99000	45.01000
7	결	309	109	62.45972	37.54028
8	결과	208	125	49.40849	50.59151
9	경영	59	51	40.43977	59.56023
10	경우	105	74	45.43795	54.56205
11	경제	85	47	51.49000	48.51000
12	경험	807	421	52.94172	47.05828
13	계기	99	80	42.07262	57.92738

1-6 상관분석 그래프 만들기

만들어진 tdm을 이용하여 상관분석을 하였다.

```
tdm <- read.csv('edittdm.csv',fileEncoding = 'EUC-KR') #첨삭자소서tdm불러오기
tdm <- t(tdm) #qgrapgh의 포맷에 맞추기 위해 transpose함
tdm.matrix = as.matrix(tdm) #데이터프레임형태를 매트릭스 형태로 변환
word.count = rowSums(tdm.matrix) #각 단어의 빈도수 확인
word.order = order(word.count, decreasing = T) #빈도순으로 정리
freq.word = tdm.matrix[word.order[1:50],] #빈도상위 50개의 단어 추출
co.matrix = freq.word %*% t(freq.word) #분석을 위해 내적한 matrix를 co.matrix로
library(qgraph)
qg <- qgraph(co.matrix,
             labels=rownames(co.matrix),
             diag=F,
             layout='spring',
             edge.color='black',
             vsize=log(diag(co.matrix))*1.5)
plot(qg)
```

ⅱ 5대 기업별 자기소개서 분석 : 엑셀 중복값 분석

5대 기업 자소서의 각각 기업의 단어 빈도수를 순위별로 나열했을 때 200개안에 있는 단어들을 살펴보았다. 즉, 이것은 각 기업별로 약 200위 안에 있는 단어이다. 엑셀을 통해 5개 기업의 200위 안에 있는 단어들 중에 2개이상의 기업에 있는 단어들은 빨간색으로 표기하였고, 1개만 있는 그 기업의 고유의 단어는 흰색으로 표기되었다. 아래의 사진은 이러한 중복표 분석으로 도출한 표에서 상위 top10위권에 있는 단어들이다.

순위	splt_hy	현대	순위	splt_lg	LG	순위	splt_lotte	롯데	순위	splt_sm	삼성	순위	splt_sk	SK
1위	저	674	1위	저	454	1위	저	448	1위	저	789	1위	저	672
2위	제	337	2위	LG	416	2위	고객	227	2위	제	420	2위	경험	346
3위	업무	316	3위	기술	253	3위	제	225	3위	사람	365	3위	제	287
4위	경험	299	4위	경험	230	4위	업무	218	4위	삼성	351	4위	사람	261
5위	현대	279	5위	사람	218	5위	경험	216	5위	삼성전자	302	5위	목표	256
6위	자동차	256	6위	제	211	6위	롯데	203	6위	고객	289	5위	업무	256
6위	회사	256	7위	전자	198	7위	사람	195	7위	친구	282	7위	SK	254
8위	고객	246	8위	생산	183	8위	때	148	8위	기술	274	8위	고객	242
9위	사람	210	9위	공정	168	9위	시간	138	9위	때	262	9위	팀	230
10위	기업	207	10위	목표	162	10위	저의	137	10위	경험	244	10위	시간	215
11위	저의	206	10위	프로젝트	162	11위	프로젝트	131	11위	기업	235	11위	동아리	213
12위	팀	194	12위	지식	160	12위	일	123	12위	업무	232	12위	프로젝트	212
13위	목표	191	13위	제품	149	12위	후	123	13위	저의	217	13위	때	202
14위	물류	186	14위	팀	141	14위	관리	118	14위	제품	207	14위	문제	198
15위	직무	164	15위	업무	134	15위	활동	105	15위	시장	206	15위	일	192
15위	프로젝트	164	15위	저의	134	16위	영업	102	16위	세계	202	16위	팀원	178
17위	능력	163	17위	분야	128	17위	과정	99	17위	분야	181	17위	생각	177

상위권에 있는 단어에서 기업명인 '현대', 'LG', '롯데', '삼성', '삼성전자' 'SK'가 5개기업 안에서 고유단어로 확인되었으며, 이는 기업에 맞춰서 쓰는 자소서의 특징이 그대로 드러나는 것으로 보인다. 상위권에 있는 단어는 기업의 특성 및 자소서 질문 등을 그대로 반영하는 것으로 분석의 의미가 없다고 판단하여 그 아래의 100위권 이하에 있는 단어들을 살펴보았다.

중복값 분석에서 도출한 단어 안의 100위보다 순위가 아래에 있는 단어들 중에서 자기소개서 질문리스트를 직접적으로 언급한 단어와 기업 고유의 특성을 반영한 단어를 제외한 단어를 선정해 보았다. 이 단어들은 약 100위권에서 200위권에 있는 단어이며, 괄호에 있는 기업에서만 200위 안에 나타나는 단어들이다. 옆에 기록된 순위는 그 기업의 각 단어 빈도수의 순위를 매겨보았을 때 나타나는 특정 단어의 순위이다.

해당 단어와 그 기업의 단어 빈도 순위 그리고 기업순으로 적어보았다.

긍정 104위(삼성)
연습 126위 (SK)
모임 126위 (SK)
구성원 138위 (SK)
교내 139위 (삼성)
실력 138위 (롯데)
기획 151위 (SK)
창의 151위 (SK)
행사 151위 (SK)

군대 177위 (SK)
회원 170위 (SK)
어머니 177위 (롯데)
의사소통 181위 (현대)
솔루션 181위 (LG)
소속감 187위 (SK)
회장 187위 (SK)

위의 선정한 단어 중 SK에는 있는 '연습', '모임', '기획', '행사', '군대' 와 같은 단어는 개인이 경험할 수 있는 독창적인 경험을 반영하는 것으로 보인다. 특히 '연습', '모임', '기획', '행사'는 동아리나 학회, 학생회, 서포터즈, 공모전 등과 같은 학내/학외에서의 경험을 드러낼 때 쓴다. 또한 '군대'라는 단어는 군대의 특수한 경험을 나타낼 때 쓴다. 실제 자기소개서를 확인해 본 결과 이러한 단어는 개인이 경험한 독창적인 경험을 나타내며 쓴 것으로 확인되었다. 또한 '군대'라는 단어는 군대에 속해있을 때의 특별한 경험과 군대 이후의 경험을 반영한 것으로 보여진다. 또한 '구성원', '회원', '회장'과 같은 단어는 경험을 나타내는 동시에 사회성을 드러내는 단어이다. 이러한 단어는 주로 동아리 등 학내/학외 경험에서 나타난 것으로 확인되었다. '창의'라는 단어는 자기역량을 나타내는 단어이다. SK의 질문 문항에는 새로운 도전에 직면했을 대의 해결방식을 묻는 문항이 있다. 이러한 문항에 비추어서 '창의'라는 단어가 순위권에서 오로지 SK만 있는 것으로 보았을 때 SK는 창의적으로 문제를 해결했던 경험이 있는 지원자를 선호하는 것으로 볼 수 있다.

삼성에서 나타난 '긍정'이라는 단어는 자기역량을 나타낸 단어인데, 이는 삼성이 추구하는 인재의 모습과 가깝다고 볼 수 있다. 자기소개서에서 '긍정'이라는 단어를 확인한 결과 지원자의 구체적인 사례와 더불어 쓰인 것을 알 수 있었다. 또한 '교내'라는 단어는 자신의 경험을 드러낸 단어인데, 이 단어를 합격 자기소개서에서 찾아보니, 초등학교, 중학교, 고등학교에서 대학교까지의 경험들이 녹아 드러난 것을 알 수 있었다.

롯데에서는 '어머니'라는 단어가 상위권에 고유값으로 랭크되어 있는데, 이는 경험에서 나온 것으로 롯데에서는 특별하고 재미나지 않은 평범한 이야기를 쓴 자기소개서도 뽑아준다는 것을 의미한다. 실제 자기소개서를 읽어 본 결과 평범한 이야기임에도 불구하고 합격된 자기소개서를 다수 확인할 수 있었다.

LG에서 상위권에 있는 '솔루션'이라는 단어는 경험위주의 단어다. 이 단어는 다른 경험 위주의 단어와 다르게 인턴 등 실제 업무에서 문제를 해결했을 때 주로 쓰는 단어이다. 직접 자기소개서를 확인해 보니, 인턴 등으로 회사에서 경험한 자신의 문제해결 과정에 대한 사례를 기록해 놓은 것을 볼 수 있었다. 이를 통해 LG는 실제 업무에서 문제해결을 해보았던 사람들을 선호하는 것으로 볼 수 있다. 또한 추가적으로 LG의 자기소개서 문항에 직무와 관련된 경험을 쓰라는 항목을 확인하였는데, 여기에서 이 단어가 간접적으로 파생된 것으로 보인다.

현대에서 상위권에 속한 '의사소통' 단어는 자기역량과 관련한 단어이고, 충분히 자기소개서에 많이 등장할 것으로 사료된다. 그러나 특징적이게도 다른 4개의 기업들에서는 상위권의 단어로 올라와 있지 않고 현대 기업에서만 등장한 이유는 정확히 알 수 없으나, 현대 기업의 가치관과 부합하는 단어라고 추측할 수 있다.

이 단어들을 종합화&카테코리화 하면 다음과 같다.

현대	LG	롯데	삼성	SK
의사소통 181위	솔루션 181위	실력 138위	긍정(104위)	연습 126위
		어머니 177위	교내 139위	모임 126위
				구성원 138위
				기획 151위
				창의 151위
				행사 151위
				군대 177위
				회원 170위
				회장 187위

경험	사회성	자기역량

위와 같은 사항으로 추측해 보았을 때, SK는 자신의 경험으로 드러난 지원자의 가능성, 사회성, 역량을 주로 보고 지원자를 뽑는다고 추측할 수 있다. 이러한 경험은 소위 말하는 '스펙'으로 가늠할 수 없는 것이며, SK는 정량적으로 드러나지 않는 개인의 역량도 중요시 한다고 말할 수 있다.

이와 대조적으로 롯데와 현대는 자기소개서에서 상위권에 등록된 '실력', '어머니', '의사소통'과 같은 단어는 기업에 지원할 때만 쓸 수 있는 단어가 아닌 대다수 기업에서 경험과 자기역량을 나타낼 때 보편적으로 쓸 수 있는 단어이다. 이를 통해 이 두 기업은 특출한 경험이나 창의적인 사례가 쓰이지 않고, 보편적이고 평범한 사례가 쓰여도 뽑힐 여지가 충분히 있다고 판단할 수 있다. 삼성에서 자기 역량인 '긍정', 학창시절을 나타내는 '교내' 라는 단어는 자신만의 독창적인 역량과 경험이 아니어도 지원자의 생각이 그 회사의 가치관이 부합하고 지원자의 다른 정량적인 스펙으로 역량이 드러난다면, 뽑힐 가능성이 있음을 암시한다. 이 4개의 기업과 다르게 LG에서는 '솔루션'이라는 단어가 상위에 유일하게 랭크되었다. 이를 통해 LG는 교내/교외의 대외활동의 경험도 중요하지만 인턴 등 실제 회사에서의 업무 경험 또한 중요시 한다는 것을 볼 수 있다. 이를 통해 LG를 지원하는 지원자는 정량적이고 정성적인 스펙을 넘어 실무적인 경험 또한 필요로 한다고 볼 수 있다.

iii 합격/첨삭 자기소개서 분석 : 비율분석, 연관분석을 통해

앞서 각 5대 기업별 자기소개서에서 나타난 어휘별 특징과 그 분포를 살펴보았다. 한편, 기업 제출용 자기소개서의 성공적 작성과 그 기회를 얻기 위해서는 기존의 방식에서 나타난 문제점을 찾아 그 의미를 대조&부각시켜 볼 필요가 있다. 곧, 전반부에 언급한 첨삭(edit)자기소개서를 불합격의 도수로 가정하여 , 이를 합격(pass, 5대기업별)군과 비교해봄으로써 나타나는 일련의 모습을 살펴보도록 하겠다.

여기서 분석의 대상이 되는 데이터 표본으로는 위의 기업별의 경우(중복 데이터값 제거)와는 달리, 합격 ,불합격 각각 임의의 8만, 4만여개의 단어를 원본 그대로 모두 분석했다.

이번에는 그 결과를 위의 경우처럼 합격(pass), 불합격(edit)의 도수 대비로 나타내어, 해당 값이 차지하는 비율을 알아보기 쉽게 정리했다. 여기서, 전체 추출한 표본 개수를 생각할 때, 지나치게 작은 값(30 미만의 노출빈도)은 일반적으로 비교· 판단하기 모호하다고 보아, 그 이상의 값을 수집 대상으로 하였다.

name	pass	edit	pass_rate	edit_rate
가족	33	61	24.09913	75.90087
가치	195	51	69.17446	30.82554
갈등	66	31	55.54662	44.45338
강점	75	31	58.67667	41.32333
개발	276	50	76.41363	23.58637
개인	102	49	54.99	45.01
결	309	109	62.45972	37.54028
결과	208	125	49.40849	50.59151
경영	59	51	40.43977	59.56023
경우	105	74	45.43795	54.56205
경제	85	47	51.49	48.51
경험	807	421	52.94172	47.05828
계기	99	80	42.07262	57.92738
계획	209	122	50.13568	49.86432
고객	831	425	53.43594	46.56406
고등학교	55	144	18.31177	81.68823

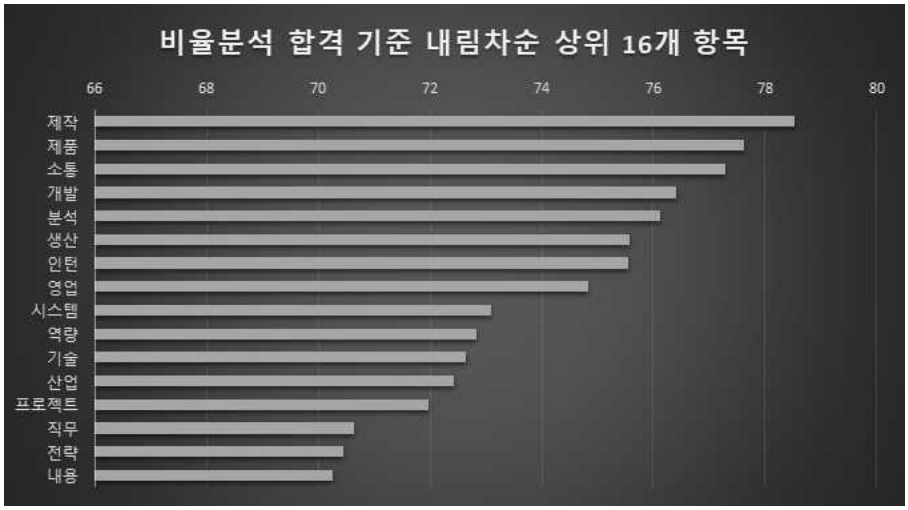
<합격/첨삭 비율 분석 결과>

이후 여기에 해당하는 약 300여개의 단어를 그 어휘의 내용표현과 의미에 따라 몇 가지 범주로 구분해보았다. 이는 생각· 의견이나 개인의 인성을 나타내는 자기 고백적 표현, 전문성 및 사회조직과 직무라는 기업적 가치를 나타낼 수 있는 표현, 활동· 경험 포부 등 도전정신과 미래지향적 의미를 내포한 경우로 분류해 볼 수 있겠다.

name	pass	edit	pass_rate	edit_rate
제작	187	30	78.53339	21.46661
제품	402	68	77.62696	22.37304
소통	203	35	77.29373	22.70627
개발	276	50	76.41363	23.58637
분석	266	49	76.11128	23.88872
생산	174	33	75.57765	24.42235
인턴	258	49	75.55164	24.44836
영업	365	72	74.84467	25.15533
시스템	199	43	73.09049	26.90951
역량	315	69	72.8214	27.1786
기술	344	76	72.6517	27.3483
산업	188	42	72.42988	27.57012
프로젝트	372	85	71.97772	28.02228
직무	283	69	70.65018	29.34982
전략	134	33	70.44223	29.55777
내용	157	39	70.26185	29.73815

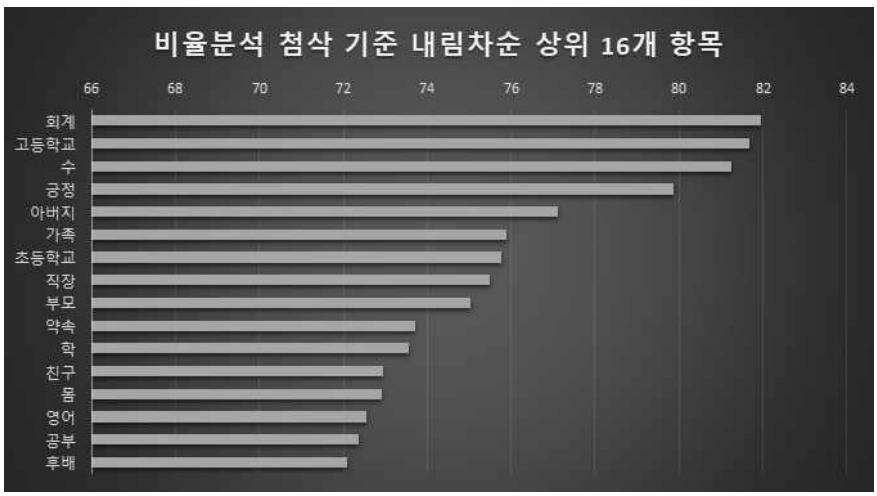
name	pass	edit	pass_rate	edit_rate
회계	48	128	18.03888	81.96112
고등학교	55	144	18.31177	81.68823
수	37	94	18.76637	81.23363
긍정	55	128	20.13974	79.86026
아버지	47	93	22.87579	77.12421
가족	33	61	24.09913	75.90087
초등학교	30	55	24.25	75.75
직장	47	85	24.50129	75.49871
부모	89	157	24.96471	75.03529
약속	34	56	26.27203	73.72797
학	30	49	26.43445	73.56555
친구	242	383	27.05205	72.94795
몸	31	49	27.07701	72.92299
영어	78	121	27.44885	72.55115
공부	106	163	27.62385	72.37615
후배	33	50	27.92064	72.07936

<합격/첨삭 비율 분석 내림차순/오름차순 결과>



<비율분석 합격 기준 내림차순 상위 16개 항목>

이것을 필터링을 통해 오름차순, 내림차순을 정렬해보면 상위권에 분포한 값을 알아볼 수 있다. 합격 비율(pass_rate)의 상위분포 기준으로 볼 때 '제작', '개발', '분석', '생산', '영업', '기술 ' 등이 주로 나타났음을 볼 수 있다. 이를 통해 합격 자기소개서는 주로 팀 단위와 조직을 중시하는 어조를 취하면서, 직무에 대한 관심이나 열정을 강조하고 있음을 알 수 있다. 그리고 '역량', '아이디어', '수행', '인턴' 등의 어휘에서 볼 때, 그 선택에 있어서 개인이 가지고 있는 능력을 기업의 활동과 함께 적절히 조화시켜 기업적 가치를 드러낼 수 있게 그 의미를 승화시켰다. 이는 최종적으로 소개서를 검토할 담당자로 하여금, 긍정적인 마인드 형성에 큰 기여를 하고 있는 것이다.



<비율분석 첨삭 기준 내림차순 상위 16개 항목>

반면, 불합격(edit_rate)의 상위 필터 정렬로 보면, 합격 케이스(case)에서 비중 있게 다루었던 단어와는 상반된 단어들이 눈에 띄게 늘어났다. 대표적으로 '부모', '아버지', '중·고등학교', '친구' 등 개인을 나타낼 수 있는 자기중심적 표현이나 일상사 소개가 주를 이루고 있다. 또한 '공부', '성격', '인생', ' 연습' 등의 용어의 경우를 보면, 개인의 역량을 보여주는 과정에

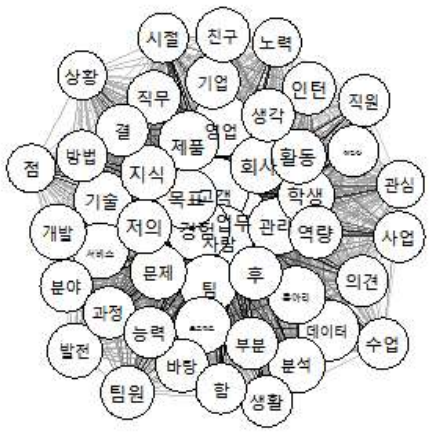
서 합격의 사례보다는 다소 순화된 표현을 섞었음을 알 수 있다. 즉, 전문적인 지식이나 업무 적합성(합격 사례에 다수 분포)을 뚜렷하게 드러내기보다 자기 형상화 및 개인적 삶의 의미부여에 방점을 두었다고 볼 수 있다.

이처럼 우리는 비율분석을 통해 단어 선택과 사용에 있어서, 성공적 자기소개 작성(pass_rate)과 실패 가능성(edit_rate)이 높은 것이 몇 가지 차이가 있었음을 확인했다.

중요한 것은, 개인의 삶이나 자아 성찰적 기회는 비교적 짧게 표현하여 진부적인 느낌을 버릴 수 있도록 하며 , 기업적 가치나 조직에 대해서 객관적 신뢰를 주고, 타인과 차별될 수 있는 전문적 인재라는 측면을 적극 강조할 수 있어야 한다는 것이다.

지금까지 단어의 선택과 사용 빈도를 검토하여 그 적절성을 판단하였다. 하지만 서두에서 언급했듯이 성공적인 소개서 제출을 위해서는 단어의 빈도-비율 분석 외의 여러 중요한 요소가 있을 것이다. 그래서 우리는 형태소-단어-문장-구문의 형성이라는 연속선상에 있다는 점을 짚어봤다. 단어를 넘어 실제 이를 활용한 문장의 형성, 어조의 파악은 소신 있는 표현력에 큰 영향을 줄 것이기 때문이다.

이를 위해 다음으로 우리는 R을 활용한 q-graph의 분석을 통해 단어 간 연결성과 조직적 구조를 먼저 알아보았다. 이후 이 과정에서 의미 있는 연결 고리를 찾아내고, 앞서 엑셀로 분석한 비율분석과 연계하여 직접 하나의 Text형성해보는 활동을 해보도록 한다.



<합격 자기소개서 q-graph>

먼저 합격(pass) 자기소개서의 q-graph 분석 결과를 통해 핵심 단어의 연관성을 살펴보았다. 중앙부(center)의 분포하는 단어들은 주변부(edge)의 경우보다 더 많은 line을 가지고, 이는 핵심 어구 형성에 중점적 역할을 할 수 있다. 이를 추출하여 몇 가지 실제 짧은 문단을 만들어보겠다.

위의 주요 단어가 포함된 자기소개서 전문을 통해 실제 지원자의 문장력을 확인해보자.

"우선 **고객**, 상권관리 등 **영업**의 기반이 되기 위한 분석력이 필요합니다."

“제가 LG전자 한국**영업**본부 Sales 분야에 지원하게 된 동기는 **영업**이라는 **직무**를 통해 **고객**과의 접점을 만들고, 이를 바탕으로 LG전자와 **미래**를 함께 할 수 있기 때문입니다.”

“리서치 **인턴**을 수행하며 직접 발로 뛰며 실제 **산업**에 대한 전문가의 의견도 들을 수 있었고, 좀 더 세밀하고 정확한 **정보**와 구체적인 **전략**을 제시하며...”

“**시장**에서 아무리 좋은 **제품**이 출시되더라도, **고객**들의 의견을 파악하지 못하면 **시장**에서 결코 성공할 수 없기 때문입니다.”

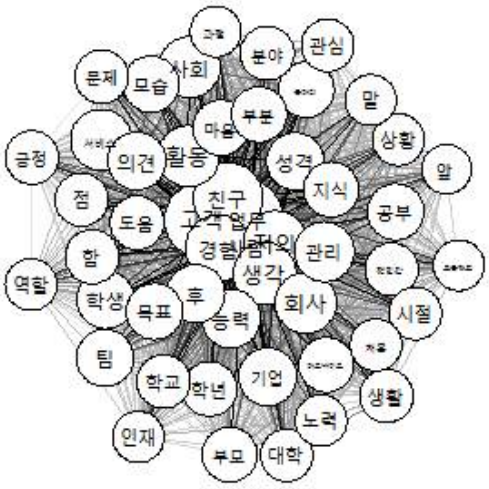
“이러한 **기술**을 바탕으로 다양한 **고객**에게 서비스를 제공하는...”

이를 바탕으로 유효 단어가 평균 2~3회 포함된 문장으로 하여 짧은 구문을 만들어보자.

‘저는 고객의 니즈를 충분히 검토하고, 이를 회사의 비전과 조화롭게 연계시킬 활동을 하겠습니다. 또한 영업에 대한 기본 이념으로써, 신뢰와 협력에 방점을 두겠습니다. 그래서 일의 수행과 관리의 측면에서 효율성을 극대화시키고, 유연한 조직 간 소통을 이루어 내겠습니다. 그리고 실제 시장의 흐름도 충분히 고려하여, 개발 분야의 지식을 적절히 활용할 수 있도록 합니다. 저의 생각과 모든 활동은 팀의 목표와도 관련이 있습니다. 때문에 책임감 있는 자세로 사업의 수행에 동참하겠습니다.’

우리는 이 학습을 통해 합격의 가능성을 높이기 위해서는, 지원 직무에 대해 적극적인 관심과 산업 협력의 자세를 보여야 한다는 것을 알 수 있다. 또한 개인의 경험적 역량을 짧고 명료하게 표현하되, 중요하게 생각하는 신념과 그 가치는 소신 있게 서술할 필요가 있다. 여기에 팀과 조직, 고객과의 접점을 형성할 수 있는 키워드를 수시로 섞어주는 노력도 수반되어야 한다.

이번에는 동일한 방법으로 불합격(edit) 자기소개서의 q-graph를 살펴보겠다.



<불합격 자기소개서의 q-graph>

“**팀원**들은 **저**를 믿고 따라와 줬고 결과는 대**성공** 이였습니다. 입사 후 **창의력**과 **협업**으로 회사 **발전**에 반드시 기여 하겠습니다.”

“ 이렇게 제가 맡은 **업무**에 더 **전문**적인 **커리어**를 쌓는 것은 물론이고 외국어 **공부**도 게을리 하지 않고 **성장**해 나가겠습니다. ”

“**동아리**페스티벌, 교내 경제 카페, 축제 등의 **활동**들을 통해 상황대처 **능력**, **고객** 응대, **서비스 마인드**를 키울 수 있었습니다.”

특징적인 점은, 역시 개인의 경험담이나 사적 지식의 가치를 지나치게 강조한 느낌이 있으며, 직무에 대한 가치를 간단하고 핵심적으로 설명할 수 있는 단어를 구구절절하게 나열한 측면이 있다. 또한 문장의 배치와 사소한 맞춤법 실수 등도 눈에 띈다.

이에 실패할 가능성이 높아 보이는 구문을 제작해보자.

‘저는 초등학교 시절부터 약속을 잘 지켜서, 친구들에게 신뢰를 주며 성격도 쾌활하여 많은 사람들과의 소통능력이 우수합니다. 제게 직장은 가장 우선순위로 생각되며, 현실적으로 많은 어려운 일이 겹칠 때 고객, 서비스마인드, 창의력, 협업을 다시 한 번 생각할 수 있는 초심을 되찾을 수 있도록 하겠습니다. 제 인생의 영향을 끼칠 수 있는 사업과 다양한 활동을 꾸준히 이어나가 창의력인 인재, 도전적인 인재로 성장할 수 있도록 합니다. 그리고 글로벌 마인드의 시대에 수많은 해외경험과 언어 공부를 하여 유능한 사람으로 판단될 수 있도록 항상 최선을 다하겠습니다. “

이 경우를 통해서 우리는 서두에 너무 화려하고 장황한 자기중심적 표현은 긍정적 이미지 형성에 역기능을 초래할 수 있음을 유추할 수 있다. 또한 진부하게, 지나치게 늘어지는 문장보다 숨을 고르며 명쾌히 자기를 나타낼 수 있어야 한다. 한편 많은 개인적 노력이나 경험들을 회사와 직· 간접적으로 연결하여 그 가치를 동시에 아우를 수 있는 넓은 시야가 필요하다.

지금까지 5대 기업별 자기소개서에 나타난 내용유형별 단어 빈도수 분류와 여러 기법을 활용한 데이터 분석을 진행해 보았다. 이 결과를 간단히 정리하면, 각 기업이 차별화할 수 있는 어떠한 요소를, 실제 적재적소의 위치에서 얼마만큼의 빈도로 조화롭게 활용할 수 있는가가 중요한 글쓰기 조건이라 할 수 있겠다. 더구나, 합격(pass)과 불합격(edit)자기소개서의 가정-비율, 연관분석을 통해 살펴본 탐구에서는 그 모습의 차이를 비교해 보았다. 성공적이면서 잠정적으로 높은 가능성을 요구할 수 있는 소개서를 작성하려면 복합적인 여러 의미를 함유하는 단어(유효 단어)를 먼저 파악하는 것이 중요하다. 이후 소개서 전반에서 자아 성찰적 측면은 작성자가 자신의 생각 및 의견 표현을 드러내는 부분으로 짧고 간단하게 도입한다. 덧붙여 어휘 및 내용의 그 구성 비율에 있어, 다른 내용에 비해 전문성 및 직무적합성에 관련된 어휘를 응답지에 적절히, 수시로 배치한다.
결국, 이러한 일련의 작업을 통해 적합한 유효 단어의 선정과 그 연결· 흐름을 지속적으로 검토한다면 보다 성공적인 자기소개서 작성에 도움이 될 것이라 생각한다.

Ⅲ 결론

이번 학습을 통해 여러 지원자의 실제 자기소개서를 직접 몇 개 살펴보았는데, 개인마다 능력치와 개성, 경험 및 가치 등 많은 부분에서 차이가 있음을 알 수 있었다. 기존의 개인별 지도 및 전문가의 조언· 피드백의 단계를 거친 시스템은 이를 보완해주며, 합격의 경우에 도달할 수 있도록 안내해주는 가이드라인이다. 하지만 우리는 R을 활용한 탐구를 통해 훨씬 더 많은 사례를 분석하려고 했고, 여기서 한 단어가 가지는 의미로부터 다양한 분석 값을 추출해보려고 노력했다. 결과적으로 이 데이터 분석, 코딩 통해 여러 수치를 계량화 할 수 있었다. 또한 기업별 탐구, 합· 불 케이스의 비교를 할 수 있었으며, 이를 통해 성공 가능성을 높이는 사례를 알아볼 수 있었다.

매년 수많은 취직용 자기소개서가 많이 작성되고 있다. 무엇보다 일반적으로 자기소개서의 멋진 작성은 적절한 상황과 문맥에 맞는 단어 선택에서 시작한다고 볼 수 있다. 여기서 자기비판적 측면은 작성자가 자신의 생각 및 의견 표현을 위한 어휘 및 자신의 인생관, 도덕관 및 가치관을 드러내기 위한 단어를 통해 표현하고 있었다. 한편, 무엇보다 합격 가능성에 근접하는 자기소개서가 되려면 ,어휘의 내용 구성 비율에 있어 다른 내용에 비해 전문성 및 직무적 합성에 관련된 어휘를 많이 사용해야 한다는 결과가 이채롭다. 나아가 자기소개서는 복합 장르적 성격의 글쓰기임을 감안했을 때, 단어의 선정도 중요하지만 이를 하나의 문장 속에서 적절히 배치하고 조율시키는 창의적 학습이 필요할 것이다. 그래서 마지막으로 이러한 분석 경험을 제시하여 호감을 줄 수 있는 작문을 할 수 있다면 충분한 의미가 있을 것이다.

한편 이번 연구의 한계점으로는 먼저 단어 단위의 분석 그 이상으로 나아가지 못한 점을 들 수 있다. 자기소개서를 작성하는데 있어서는 앞에서도 언급했듯 바람직한 단어의 선정을 넘어서 고려해야할 것들이 많다. 단어의 분석과 함께 각 자기소개서들의 맥락까지 함께 분석할 수 있었더라면 더욱 유의미한 결과를 얻을 수 있었을 것이라 사료된다. 또한, 서론에서 언급했듯이 데이터 표본의 분석 대상의 범위에서 합격 사례와 불합격 사례를 비교하는 것이 타당하다. 하지만 현실적인 한계로 인해 불합격 자기소개서가 아닌 첨삭을 위해 올려진 자기소개서를 사용하였다. 그리고 단어 간 연관성을 보여주는 q-graph분석기법을 사용해보며 문맥에 대한 정보를 얻어보려 했으나, 이번 연구는 각 자기소개서를 단위로 단어의 존재를 표현한 term-document-matrix를 바탕으로 진행되어 유의미한 결과를 찾기 힘들었다. 개인정보를 많이 담고 있는 자기소개서의 특성상 수집하는 데이터의 개수에 한계가 있었는데, 앞으로 대량의 데이터를 수집하여 연구를 진행한다면 뉴럴네트워크를 활용한 Word2vec 등의 알고리즘을 활용하여 문맥을 고려한 각 단어의 분석이 가능해질 것이라 기대한다. 또한 추가적으로 데이터를 수집할 때 랜덤샘플링 기법을 사용하였는데, 합격 자기소개서와 불합격 자기소개서를 비교하는 과정에서 산업군이나 직무의 비율의 불균형의 가능성이 제기되었다. 데이터 크롤링에서의 역량 개발을 통해 랜덤샘플링이 아닌 층화추출샘플링 기법을 활용한다면 이러한 문제의 가능성을 제거할 수 있을 것이다.

4차 산업혁명이 도래하고 있는 지금 세상은 급변하고 있으며, 인공지능 기계학습을 통해 컴퓨터가 소설을 쓰고 작곡을 하는 프로그램이 개발되었다고 한다.(glinconkorea. 2016.6.) 한편 앞서 서술하였듯 기업 입사에 있어 자기소개서 작성은 중요한 과정이며 구직자들이 많은 노력을 기울이는 부분이다. 하지만 자기소개서의 작성을 위한 구직자의 노력이 개개인의 역량

을 개발시켜 주지는 못한다고 판단된다. 본 연구와 더불어 성공적인 자기소개서 작성에 대한 연구가 진행되고 프로그램이 개발된다면 수많은 구직자들의 노고를 덜고 그 만큼 실제 역량 개발에 노력을 기울일 수 있는 문을 열어줄 수 있을 거라 기대된다.