빅데이터 시각화: 워드클라우드

IBM의 'Service Jam' 행사에서 생성된 의견을 분석한 결과의 시각화

주: 2010년 10월 'Service Jam' 행사에서 생성된 5,860개 의견에 텍스트 마이닝 분석을 적용해 자원봉사자 관리, 서비스문화 육성, 지식공유 등 8가지 주요 테마를 도출

자료: IBM (2011.1.7.). Service Jam - Post Jam Data Analysis.

남녀간 어휘 사용 키워드 분석

- 여자
  - shopping, chocolate, hair, happy, boyfriend
  - 이모티콘
    - ^.^ :)
- 남자
  - youtube, fuck, xbox, league, world_cup, football

R 프로그램을 이용한 WordCloud

- 데이터 시각화 기법
  - 데이터 분석 결과를 쉽게 이해할 수 있도록 도표 형태의 시각적 수단으로 표현하는 기법

- 워드클라우드
  - 데이터 시각화 기법 중 하나
  - 텍스트에 출현하는 단어를 빈도에 비례하는 크기로 표출한 그래프

빅데이터 분석 활용 사례: 잡플래닛

- **BEST: "가족같은 동료애를 느낄 수 있었고 무엇이든 할 수 있다는 자신감을 심어준 첫사랑같은 회사라고 할 수 있었습니다."**
- 장점: IT 기업답게 회사 체계가 잘 잡혀 있으며 사업부마다의 특성과 장점이 특화되어 있어 직원에 대한 업무 능력을 빠르게 키워 나갈 수 있도록 지원하는 곳이다.
- 단점: 다 좋지만 단점이 있다면 교통사업부에서 진행하는 솔루션 이외에는 장기적 관점에서 시장 확대의 가능성이 보이지 않고 있다. 신사업 분야에서 가능성이 큰 사업군을 발굴하여야 하는 실정이다.

- **BEST: "IT 모니터링 솔루션 업계에서 Top3 안에 드는 업체이며 매년 꾸준히 성장해 나가고 있는 기업"**
- 장점: 여러 부서와의 협업 및 순환 근무를 통해 다양한 업무를 배울 수 있으며, 본인의 의지만 있다면 온오프라인의 교육을 통해 다양한 커리어를 쌓을 수 있음.
대표이사가 인재를 먼저 생각하고 꾸준한 투자로 회사를 키워 나가려는 마인드가 있어서 좋음. 또한 재무적으로 탄탄한 기업.
- 단점: 특정 기간의 야근이 많으며, 본인의 능력 이상의 업무를 시킴으로써 회사에 적응하기 위해서는 많은 노력이 필요. 또한 부서간 협업이 조금씩 나아지고는 있으나 아직 많이 부족함.
- https://www.jobplanet.co.kr/

2015 대기업 직원들이 말한 일하기 좋은 대기업의 장·단점

- 잡플래닛과 포츈코리아가 진행한 '2015 Best Companies'
  - 임직원이 직접 작성한 만족도를 가지고 일하기 좋은 기업을 선정
  - 잡플래닛에 평가 기간(2015년 1월 1일~10월 15일) 동안 전현직 임직원이 자신이 다니는 회사에 대해 익명으로 작성한 리뷰를 바탕으로 한 시상
- 미국의 경우, 잡플래닛과 흡사한 기업 리뷰 서비스 '글래스도어'가 이와 비슷한 방법으로 기업 만족도 시상
  - 글래스도어의 기업 만족도상은 페이스북의 창업자겸 최고경영자(CEO) 마크 저커버그(Zuckerberg)가 직접 참여하는 유일한 시상식
- 이번 평가는 평가 분야에 가중치를 부여하여 산출한 환산 점수를 기준으로 결정
- 직원 평가의 순수성을 유지하기 위해 자의적으로 영향을 줄 수 있는 다른 평가 방식은 사용하지 않았으나, 미디어 필터링을 통해 최근 3년간 사회적 물의를 일으킨 기업은 제외
- 출처
  - http://m.post.naver.com/viewer/postView.nhn?volumeNo=3258532&memberNo=9520310

✓ **대기업군 15개 기업**

순위	기업명	총점	산업
1	에스케이텔레콤(주)	75.38	통신
2	현대자동차(주)	73.58	제조/화학
3	기아자동차(주)	72.29	제조/화학
4	현대엔지니어링(주)	72.10	건설업
5	엘지하우시스(주)	72.06	제조/화학
6	농협은행(주)	71.97	은행/금융업
7	에스케이하이닉스(주)	71.93	제조/화학
8	케이티앤지(주)	71.80	제조/화학
9	에스케이플래닛(주)	71.56	통신
10	엘지화학(주)	70.38	제조/화학
11	비씨카드(주)	70.24	은행/금융업
12	에스케이이노베이션(주)	69.83	제조/화학
13	대우인터내셔널(주)	69.45	유통/무역/운송
14	대우건설(주)	69.42	건설업
15	삼성물산(주)	69.24	유통/무역/운송

대기업 수상 기업 리뷰: 장점 키워드 vs. 단점 키워드

[대기업 수상 기업 장점 키워드]

[대기업 수상 기업 단점 키워드]

장점 키워드 vs. 단점 키워드

SK텔레콤 장점

현대자동차 장점

기아자동차 장점

SK텔레콤 단점

현대자동차 단점

기아자동차 단점

R 프로그래밍 패키지

- 데이터 시각화를 위한 프로그래밍 도구
  - 공개 소프트웨어
  - http://www.r-project.org/
- CRAN(Comprehensive R Archive Network)
  - http://cran.nexr.com/
    - NexR Corporation, Seoul
  - http://healthstat.snu.ac.kr/CRAN/
    - Graduate School of Public Health, Seoul National University, Seoul
  - http://cran.biodisk.org/
    - The Genome Institute of UNIST
- 입력 파일 생성: test.txt
  - 데이터가 나열되어 있는 텍스트 파일
  - 한 라인에 데이터 1개씩

File test.txt

333
222
111
333
111
222
333
222
444
222
555
333
555
111
666
444
111
222

wordcloud 사용법

```
> w1 <- c("AAA", "BBB", "CCC", "DDD", "EEE")        # 소팅 순서
> w2 <- factor(w1)
> f1 <- c("6", "2", "5", "1", "3")
> f2 <- as.integer(f1)              # f2 <- c(6, 2, 5, 1, 3)

> install.packages("wordcloud")       # 패키지 설치 - 1회만!
> library(wordcloud)
> pal <- brewer.pal(9, "Set1")          # 컬러 인코딩 방식
> n <- 2                              # 빈도수  n 이상 디스플레이

> wordcloud(w2, f2, scale=c(5,1), rot.per=0.25, min.freq=1,
random.order=F, random.color=T, colors=pal)
```

R 패키지 실습

```
> f <- file("c:/Temp/test.txt", blocking=F)        # 입력 파일 test.txt
> txtLines <- readLines(f)
> txtLines                                  # 변수 nouns의 내용 확인
> head(unlist(txtLines), 10)
> head(unlist(txtLines))                    # default로 6개만 출력
> write(unlist(txtLines), "c:/Temp/testout.txt")    # 파일 출력
>
> data <- read.table("c:/Temp/testout.txt")       # 파일 입력
> words <- table(data) # 중복 제거 및 빈도 계산

> nrow(data)                                # 명사 개수 – 입력 데이터 개수
> length(words)                             # 명사 개수 – 중복 제외
> sort(words, decreasing=T)     # 빈도 높은 것부터 낮은 순으로 출력
> sort(words, decreasing=F)     # 빈도 낮은 것부터 높은 순으로 출력
```

어휘/빈도의 시각화

```
> install.packages("wordcloud")      # 워드클라우드 패키지 설치 – 1회만!
> library(wordcloud)

> install.packages("RColorBrewer") # 사용할 수 있는 글자 색깔 패키지 설치
> library(RColorBrewer)            # 이 패키지는 wordcloud와 함께 설치됨

> display.brewer.all()             # 사용할 색깔을 모두 보여줌
> pal <- brewer.pal(9, "Set1")     # 컬러 인코딩 방식을 변수 pal에 할당

> wordcloud(names(words), freq=words, scale=c(5,1), rot.per=0.25,
  min.freq=1, random.order=F, random.color=T, colors=pal)
    # scale : 폰트 크기 c(MAX, MIN)
    # rot.per : 회전되는 단어의 빈도
    # min.freq : 출력할 단어의 최소 빈도
    # random.order=F : 빈도가 큰 단어를 중앙에 오도록 함
    # random.color=T : 실행시마다 단어의 색을 random하게 함
```

명사 추출: 형태소 분석 패키지

```
> install.packages("KoNLP")        # 한국어 형태소 분석 패키지 설치
> library(KoNLP)

> f <- file("c:/Temp/test.txt", blocking=F)     # 텍스트 파일 test.txt
> txtLines <- readLines(f)

> nouns <- sapply(txtLines, extractNoun, USE.NAMES=F)
          # 명사 추출
> nouns
> write(unlist(nouns), "c:/Temp/testout.txt")              # 파일 출력

> nouns2 <- read.table("c:/Temp/testout.txt")              # 파일 입력
    # noun2의 클래스는 'data.frame'
> words <- table(nouns2) # 빈도 계산 및 테이블 생성
    # words의 클래스는 'table' --- '가나다' 순 소팅 및 빈도 계산
> sort(words, decreasing=T)          # 빈도 높은 것부터 역순으로 출력
```

데이터 시각화 요약

단계 1. 데이터 시각화 준비 작업

```
> library(wordcloud)
> pal <- brewer.pal(9, "Set1")    # 컬러 인코딩 방식
> n <- 5                          # 디스플레이 예정인 최소 빈도수
```

단계 2. 데이터 로딩, 빈도계산, 디스플레이

```
> nouns <- read.table("c:/Temp/nouns.txt")      # 명사 리스트
> words <- table(nouns)                         # 중복 제거 빈도 계산
> wordcloud(names(words), freq=words, scale=c(5,1), rot.per=0.25,
  min.freq=n, random.order=F, random.color=T, colors=pal)
```

텍스트 문서에서 명사 추출

- 시각화 대상 파일 생성
  - 한글 텍스트 파일: text.txt
  - 한글 명사/어휘 리스트: nouns.txt

- 한글 텍스트 파일 test.txt로부터 nouns.txt 생성 방법은?

- 한국어 형태소 분석 패키지 KoNLP ?
  - R에서 한글 문서를 형태소 분석을 통해 명사 추출
  - Java 기반의 형태소 분석
  - rJava가 설치되어 있어야 사용 가능

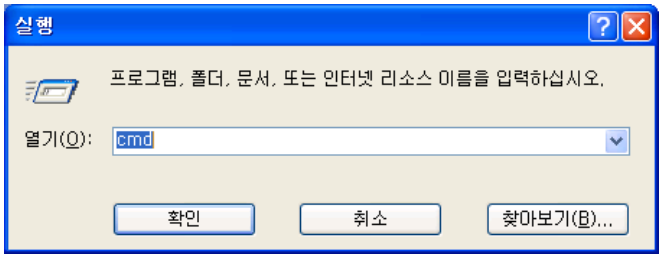
한글 텍스트 파일 test.txt로부터 명사들만 추출하는 방법

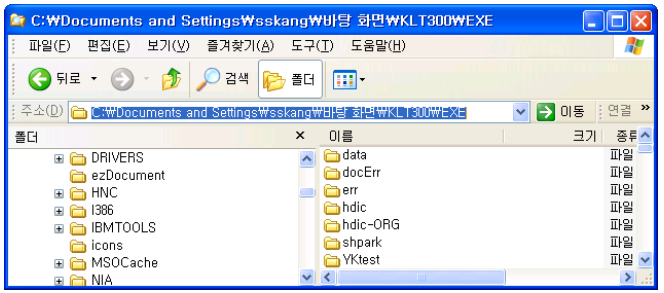
- 한국어 형태소 분석기 다운로드
  - http://cafe.daum.net/nlpk

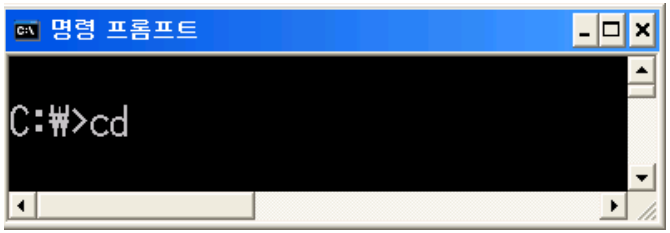
- 사용법
  - 윈도 cmd 창(명령 프롬프트)
    - 윈도의 시작-실행-cmd
  - Cmd창에서 형태소 분석기 설치된 EXE 폴더로 이동
    - 윈도탐색기에서 형태소 분석기의 EXE 폴더 경로 ctrl-C
    - cmd창에서 ctrl-V (마우스 우측 버튼 클릭하여 붙이기)
  - 형태소 분석기 실행

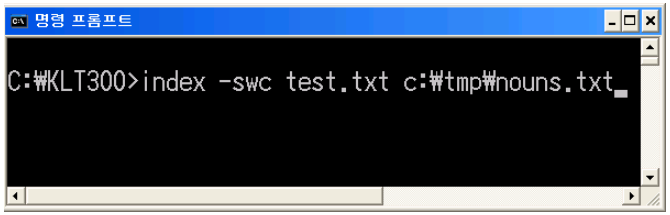
    C> index.exe

    C> index.exe -swc test.txt nouns.txt









R에 대한 평가

- 통계 패키지: SAS, SPSS(IBM 인수), R

- 포레스터 리서치, 마이크 구알티에리
  - http://www.forrester.com/ -- 미국 시장 조사 기관
  - R이 데이터 과학자들 사이에서 유명하다.
  - 대부분의 컴퓨터 과학자들은…
    - R이 품격이 높은 것도 아니고, 효율적이거나, 특정 목적을 위한 고성능의 프로그래밍 언어도 아니다.
  - 뛰어난 개발자들에게는… R은 코볼(Cobol)처럼 보인다
  - 파이썬(Python)과 다른 많은 현대적인 프로그래밍 언어들이 인정받고 있다.

Microsoft R Server

- 2015년, 마이크로소프트
  - Revolution Analytics 인수
    - R 서버 – 리눅스, Hadoop, 테라데이터 버전
    - RRE – 윈도 버전
  - Revolution Analytics
    - 설립자: 노먼 나이, Stanford대 교수
      - 레볼루션 R 개발 – R의 기업용 버전
    - Revolution R Enterprise(RRE) – 플랫폼
    - 통계, 예측 모델링, machine learning 지원