인문학 텍스트 마이닝

Text Mining about Novel

Text Mining about Novel

- 데이터 가져오기

```
> setwd("/Users/Seongmin_M/Desktop/Class")
>
> God_No_1 <- file("/Users/Seongmin_M/Desktop/Class/GOD_1.txt", blocking=F)
>
> txtLines <- readLines(God_No_1)
>
> close(God_No_1)
```

- 라이브러리 불러오기

```
> library(tm)
필요한 패키지를 로딩중입니다: NLP
> library(KoNLP)
```

- 불필요한 요소 삭제

```
> txtLines <- gsub("()", "", txtLines)
> txtLines <- gsub("<", "", txtLines)
> txtLines <- gsub(">", "", txtLines)
> txtLines <- gsub("[ |\t]{2,}", "", txtLines)
```

- 세종사전을 활용한 형태소 분석

```
> useSejongDic()
Backup was just finished!
87007 words were added to dic_user.txt.
>
> txtLines_Nouns <- sapply(txtLines, function(x) {paste(extractNoun(x), collapse = " ")})
```

- 분석 결과 확인(1)

```
> head(unlist(txtLines_Nouns))
```

"지은이 소개"
베르베르는 일곱 살 때부터 단편소설을 쓰기 시작한 타고난 글쟁이이다. 1961년 툴루스에서 태어나 법학을 전공하고 국립 언론 학교에서 저널리즘을 공부했다. 별들의 전쟁 세대에 속하기도 하는 그는 고등학교 때 만화와 시나리오에 탐닉하면서 만화 신문 유포리를 발행하였고, 이후 올더스 헉슬리와 H.G. 웰스를 사숙하면서 소설과 과학을 익혔다. 대학 졸업 후에는 르 누벨 옵세르바퇴르에서 저널리스트로 활동하면서 과학 잡지에 개미에 관한 평론을 발표해 오다가, 드디어 1991년 120여 회의 개작을 거친 개미를 발표, 전 세계 독자들을 사로잡으며 단숨에 주목받는 파랑스의 천재 작가로 떠올랐다.

"베르베르는 일곱 살 때 단편소설 시작 한 글쟁이 1961 년 툴루스에서 법학 전공 국립 언론 학교 저널리즘 공부 별들의 전쟁 세대 그 고등학교 때 만화 시나리오 탐닉 만화 신문 유포 리 발행 이후 올더스 헉슬리와 H G 웰스를 사숙 소설 과학 대학 졸업 후 르 누벨 옵세르바퇴르에서 저널리스트 활동 과학 잡지 개미 평론 발표 해 1991 년 120 회의 개작 개미 발표 전 세계 독자 들 주목 파랑 스 천재 작가"

Text Mining about Novel

- 분석 결과 확인(2)

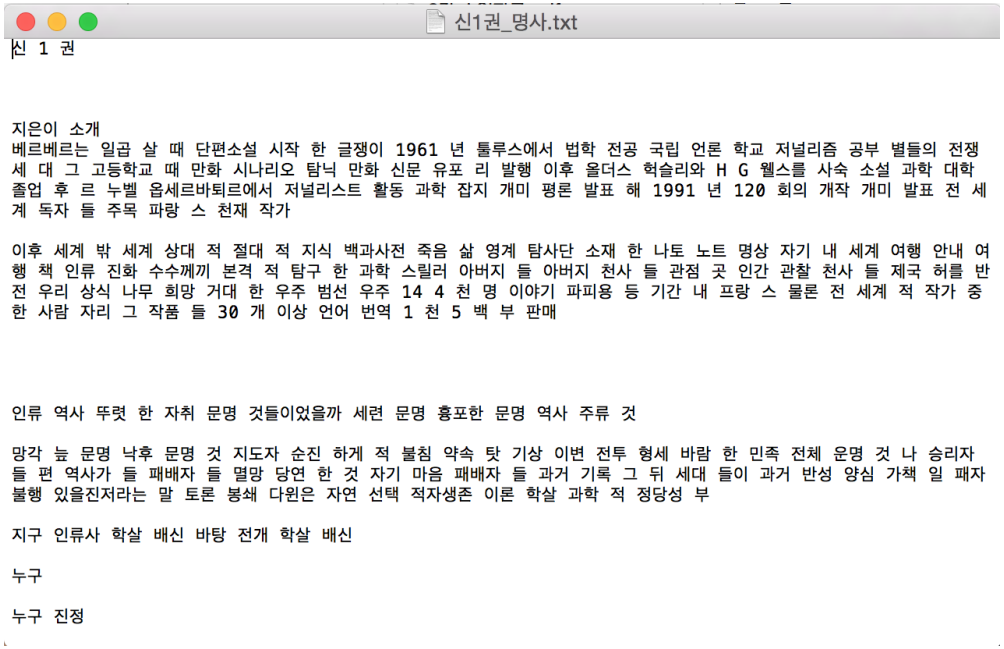
```
> head(unlist(txtLines_Nouns), 20)
```

이렇듯 지구의 인류사는 학살과 배신을 바탕으로 전개되었고, 그 학살과 배신은 잊혔다.

"지구 인류사 학살 배신 바탕 전개 학살 배신"

- 데이터 내보내기

```
> write(unlist(txtLines_Nouns),"/Users/Seongmin_M/Desktop/Class/신1권_명사.txt")
```



Text Mining about Novel

- 명사 합산 하기

```
> Nouns_wordcount <- table(unlist(txtLines_Nouns))
>
> length(Nouns_wordcount)
[1] 1338
```

- 분석 결과 내림차순 정렬

```
> head(sort(Nouns_wordcount, decreasing=T))
```

```
1425
에드몽 웰즈 상대 적 절대 적 지식 백과사전 제5권 (헤시오도스 신통 기 프랑시스 라조르박 글 근거 한 것
5
라울
3
나 무엇
2
누구
2
로디테
2
```

아프

Text Mining about Novel

- 분석 결과 오름차순 정렬

```
> head(sort(Nouns_wordcount,decreasing=F))

"미카엘 팽송 종아 142,857 호 빌라

1

"미카엘 팽송입니다

1

"아에덴 도성 일세 올림 피 자네 이름 뭐 내 말 자네 인간 시절 이름 뭐냐는

1

"어쨌거나 말 수 거 자네 그 눈길 낮춘다벌거숭이라는

1
"흔히들 나 디오니소스라 자 포도 나무 포도주 축제 음주 가무 방탕 따위 나 연결 그것 잘못 일세 것들은 나 진면목 거리 나 자유
 일세 통속 적 상상 체계 자유 의심 방탕 연결 되기 나 일세 저 자기 안 것 자유 나 방탕 한 신 그것 나 수

1

“ 작업 저 최선 몇 세 자멸 인류 창조 해 ” 헤르메스는 미소 우리 공중

1
```

Text Mining about Novel

- Corpus형태로 형변환

```
> GOD.text.Corpus <- Corpus(VectorSource(txtLines_Nouns))
>
```

- stopwords제거

```
> GOD.text.Corpus <- tm_map(GOD.text.Corpus, function(x)removeWords(x,stopwords()))
>
```

- TermDocumentMatrix를 사용하여 수치형으로 데이터 변환

```
> God_TDM_1 <- TermDocumentMatrix(GOD.text.Corpus, control = list(wordLengths = c(2, Inf)))
>
```

- 빈도수가 10 이상인 명사들 출력

```
> findFreqTerms(God_TDM_1, lowfreq = 10)
  [1] "“그들은"      "“나는"      "“내가"      "“이"
  [5] "“이제"        "142857"     "17"         "18"
  [9] "가슴"         "가운데"     "가정"       "가족"
 [13] "가지"         "각자"       "강둑"       "강력"
 [17] "강물"         "강의"       "개념"       "개미"
 [21] "거기"         "거대"       "거리"       "거울"
 [25] "거인"         "거지"       "건너편"     "건물"
 [29] "건설"         "게임"       "결과"       "결합"
 [33] "경우"         "경험"       "계속"       "고개"
 [37] "고대"         "고안"       "고통"       "곡선"
 [41] "공격"         "공기"       "공동체"     "공룡"
 [45] "공포"         "과거"       "과학"       "관심"
 [49] "관찰"         "광물"       "괴물"       "구름"
 [53] "구멍"         "구성"       "구체"       "궁전"
 [57] "귀스타브"     "규칙"       "그것"       "그녀"
 [61] "그다음"       "그들"       "그때"       "그리스"
```

- 가정과 관련된 명사 출력

```
> findAssocs(God_TDM_1, "가정", 0.25)
$가정
      불안케 파괴한다면(이제       핵무기       외계인         가증
        0.53         0.53         0.53         0.48         0.38
        과오         도리       메시지         아찔         오버
        0.38         0.38         0.38         0.38         0.38
        오염         외계       책임감         실패       생명체
        0.38         0.38         0.38         0.34         0.28
    “수사는 “피해자들         1830           33         가난
        0.27         0.27         0.27         0.27         0.27
        가스         건강         결핵         고뇌         공로
        0.27         0.27         0.27         0.27         0.27
        궁녀         권세         꿈속         농부       덩어리
        0.27         0.27         0.27         0.27         0.27
  데메테르는         도래 드루이드교         막연         만약
        0.27         0.27         0.27         0.27         0.27
        목숨       무용수         무희 백파이프를         벤치
        0.27         0.27         0.27         0.27         0.27
```