현대소설 텍스트 마이닝

송경렬

2021 4 20

데이터 설명

- 제가 사용한 소설 작품은 현대소설인 카프카를 읽은 밤, 우리들의 일그러진 영웅, 꺼삐딴 리 입니다.
- 세 작품 모두 200 줄 정도로 데이터의 크기를 대략적으로 맞춰주었습니다.

필요한 라이브러리들을 모두 불러오고 readLines 를 통해 txt 형태로 되어있는 원본 데이터를 불러옵니다.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readr)
library(stringr)
library(textclean)
```

```
## Warning: package 'textclean' was built under R version 4.0.5
```

```
library(tidytext)
library(KoNLP)
```

```
## Checking user defined dictionary!
```

```
library(ggplot2)
library(tidyr)
```

```
raw_kafka<-readLines("C:/Users/Song/Desktop/textminig/tmn/Data/Data/The_night_I_read_Kafka.txt",encoding="UTF-8")
```

```
## Warning in readLines("C:/Users/Song/Desktop/textminig/tmn/Data/Data/
## The_night_I_read_Kafka.txt", : 'C:/Users/Song/Desktop/textminig/tmn/Data/Data/
## The_night_I_read_Kafka.txt'에서 불완전한 마지막 행이 발견되었습니다
```

```
raw_hero<-readLines("C:/Users/Song/Desktop/textminig/tmn/Data/Data/Our_Distorted_Hero.txt",encoding="UTF-8")
```

```
raw_lee<-readLines("C:/Users/Song/Desktop/textminig/tmn/Data/Data/kkeoppittanLee.txt",enc
oding="UTF-8")
dic<-read_csv("C:/Users/Song/Desktop/textminig/tmn/Data/Data/knu_sentiment_lexicon.csv")

##
## -- Column specification --------------------------------------------------------
## cols(
##   word = col_character(),
##   polarity = col_double()
## )
```

문장별 감정 점수확인하기

- View()를 통해 데이터를 살펴보았을 때 <, > , -등의 특수 문자들이 자주 등장하였습니다.
- <>는 등 어떤 명사를 묶어주는 역할을 할 때 쓰였습니다. <,>를 str_replace_all 을 통해 없애줄 때 " “로 바꿔주게 되면 를 -> 시티 라이프 를 과 같이 명사와 그를 받쳐주는 형태소가 띄어쓰여지게 되어""로 바꿔주었습니다.

작품: 카프카를 읽은 밤

특수문자들을 제거해주고 tibble 형으로 변환

```
kafka2<-raw_kafka %>%
  str_replace_all("<","") %>%
  str_replace_all(">","") %>%
  str_replace_all("-"," ") %>%
  str_squish() %>%
  as_tibble()

head(kafka2)
```

```
## # A tibble: 6 x 1
##   value
##   <chr>
## 1 그녀는 아주 밝은 분홍색 원피스를 입고 있었다. 패션 용어를 빌려 더 정확하게 말하자면, 라이트 핑크/플레어 라인 실루엣을 걸치고 있었다고~
## 2 지루하게까지 느껴지는 이 용어는 내가 그녀와 작별하고 대략 32 시간정도 경과한 뒤 광주 고속버스 터미널 구내매점에서 산, 매일경제 신문사 ~
## 3 시티 라이프를 산 건 거의 충동적인 동기에서였다. 표 4 에 장 콕토의 연필 데생인 듯한 카프카 초상이 눈을 부릅뜨고 있었던 것이다. 을지서적~
## 4 그녀는 날씬한 체형도 아니었고, 키가 크고 살이 쪄 보이지도 않았다. 남의 눈에 띌 만한 특징이 없는, 그야말로 보통의 체구였다. 인류사 ~
## 5 상쾌한 날의 상쾌한 옷차림이었다면 뭐 그리 돋보일 게 있었겠는가. 그런데 그날은 때늦은 봄비가 내리고 있었다. 비 오는 날의 밝은 핑크빛 ~
```

```
## 6 제법 굵은 빗줄기가 떨어져 내리는 삼거리에 그녀는 서 있었다. 우산이 없는 사람들은 윗저고리 깃을 정수리까지 올리고 빗물로 번들거리는 아스~
```

문장별, 단어별

- 일단 문장 기준 토큰화를 진행해줍니다.
- 그 후에 다시 단어기준으로 토큰화를 진행해줍니다.
- 이 때 drop = F 로 하여 문장과 문장 속 단어들을 같이 표현할 수 있게 해줍니다.

```
word_kafka<-kafka2 %>%
  unnest_tokens(input = value,
                output = sentence,
                token = "sentences")


word_kafka<-word_kafka %>%
  unnest_tokens(input = sentence,
                output = word,
                token = "words",
                drop = F)
```

감정 점수 및 감정분류

- 감정사전과 left join 을 해줍니다.
- 감정사전에 없는 단어는 polarity 를 0 으로 있는 경우에는 감정사전에 있는 polarity 를 반영해줍니다.

```
word_comment_kafka<-word_kafka %>%
  left_join(dic,by="word") %>%
  mutate(polarity = ifelse(is.na(polarity),0,polarity))

word_comment_kafka<-word_comment_kafka %>%
  mutate(sentiment = ifelse(polarity == 2,"pos",ifelse(polarity == -2,"neg","neu")))

word_comment_kafka %>% count(sentiment)
```

```
## # A tibble: 3 x 2
##   sentiment     n
##   <chr>     <int>
## 1 neg          31
## 2 neu        5563
## 3 pos          18
```

문장별 점수합산

- 문장별로 분리한 후에 polarity 를 합산해 감정 점수를 구합니다.
- 분석 작업을 그룹별로 처리하지 않도록 ungroup()을 이용해 그룹을 해제합니다.

```
kafka_score<-word_comment_kafka %>%
  group_by(sentence) %>%
  summarise(score = sum(polarity)) %>%
  ungroup()
```

나머지 두 작품도 위와 동일한 과정을 수행해줍니다.

작품: 우리들의 일그러진 영웅

특수문자들을 제거해주고 tibble 형으로 변환

```
hero2<-raw_hero %>%
  str_replace_all("<","") %>%
  str_replace_all(">","") %>%
  str_replace_all("-"," ") %>%
  str_squish() %>%
  as_tibble()
```

문장별, 단어별

```
word_hero<-hero2 %>%
  unnest_tokens(input = value,
                output = sentence,
                token = "sentences")

word_hero<-word_hero %>%
  unnest_tokens(input = sentence,
                output = word,
                token = "words",
                drop = F)
word_hero
```

```
## # A tibble: 3,743 x 2
##    sentence                                                          word
##    <chr>                                                             <chr>
##  1 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 벌써
##  2 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 삼십
##  3 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 년이
##  4 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 다
##  5 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 돼
##  6 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 가지만
##  7 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언제나 그때처럼 막막하고 암~ 그해
##  8 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언
```

```
제나 그때처럼 막막하고 암~ 봄에서
##  9 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언
제나 그때처럼 막막하고 암~ 가을까지의~
## 10 벌써 삼십 년이 다 돼 가지만, 그해 봄에서 가을까지의 외롭고 힘들었던 싸움을 돌이켜 보면 언
제나 그때처럼 막막하고 암~ 외롭고
## # ... with 3,733 more rows
```

감정 점수 부여 및 감정 분류

```
word_comment_hero<-word_hero %>%
  left_join(dic,by="word") %>%
  mutate(polarity = ifelse(is.na(polarity),0,polarity))

word_comment_hero<-word_comment_hero %>%
  mutate(sentiment = ifelse(polarity == 2,"pos",ifelse(polarity == -2,"neg","neu")))

word_comment_hero %>% count(sentiment)
```

```
## # A tibble: 3 x 2
##   sentiment     n
##   <chr>     <int>
## 1 neg          36
## 2 neu        3687
## 3 pos          20
```

문장별 점수합산

```
hero_score<-word_comment_hero %>%
  group_by(sentence) %>%
  summarise(score = sum(polarity)) %>%
  ungroup()
```

작품: 꺼삐딴 리

특수문자들을 제거해주고 tibble 형으로 변환

```
lee2<-raw_lee %>%
  str_replace_all("<","") %>%
  str_replace_all(">","") %>%
  str_replace_all("-"," ") %>%
  str_squish() %>%
  as_tibble()
```

문장별, 단어별

```
word_lee<-lee2 %>%
  unnest_tokens(input = value,
                output = sentence,
                token = "sentences")
word_lee<-word_lee %>%
  unnest_tokens(input = sentence,
                output = word,
                token = "words",
```

```
                drop = F)
word_lee
```

```
## # A tibble: 3,228 x 2
##    sentence                                                        word
##    <chr>                                                           <chr>
##  1 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 수술
실에서~
##  2 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 나온
##  3 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 이인
국
##  4 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 李
##  5 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 仁
##  6 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 國
##  7 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 박사
는
##  8 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 응접
실
##  9 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 소파
에
## 10 수술실에서 나온 이인국(李仁國) 박사는 응접실 소파에 파묻히듯이 깊숙이 기대어 앉았다.~ 파묻
히듯이~
## # ... with 3,218 more rows
```

감정 점수 부여 및 감정 분류

```
word_comment_lee<-word_lee %>%
  left_join(dic,by="word") %>%
  mutate(polarity = ifelse(is.na(polarity),0,polarity))

word_comment_lee<-word_comment_lee %>%
  mutate(sentiment = ifelse(polarity == 2,"pos",ifelse(polarity == -2,"neg","neu")))

word_comment_lee %>% count(sentiment)
```

```
## # A tibble: 3 x 2
##   sentiment     n
##   <chr>     <int>
## 1 neg          18
## 2 neu        3193
## 3 pos          17
```

문장별 점수합산

```
lee_score<-word_comment_lee %>%
  group_by(sentence) %>%
  summarise(score = sum(polarity)) %>%
  ungroup()
```

작품 별 결과물 상위 20 개

- 카프카를 읽은 밤

```
# 긍정 문장 상위 20 개
kafka_score %>%
  arrange(-score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                                  score
##    <chr>                                                                     <dbl>
##  1 "상쾌한 날의 상쾌한 옷차림이었다면 뭐 그리 돋보일 게 있었겠는가."           4
##  2 "두 말이 넘는 테이프를 승용차에 싣고 다니거나, 영화나 연주회 실황 같은 걸 빠뜨리지 않고 꼬박꼬박 녹화하거나, 기를 쓰고 ~     3
##  3 "의사는 내 귓속을 들여다보거나 따뜻하고 매끄러운 철사 끝을 넣어 보거나 하면서 아주 오랫동안 친구와 얘기를 나눴다."~     3
##  4 "\"저, 금방 우리말을 굉장히 유창하게 했다는 거 알아요?\""                 2
##  5 "그녀는 아주 밝은 분홍색 원피스를 입고 있었다."                           2
##  6 "날씬한 체형이 아니면 소화하기 어려우나 키가 크고 살이 찐 사람에게도 잘 응용하면 상쾌해 보인다."~     2
##  7 "낯선 분홍색 원피스를 입었다면 그녀를 쉽게 알아볼 수 있었을 거라는 말은, 그냥 쉽게 알아볼 수 있었을 뿐이라는 말이겠다."~     2
##  8 "단순히 밝은 분홍이라고만은 할 수 없어."                                  2
##  9 "모닝 빵에다 나이프 끝으로 버터를 바르고 있는 그녀를 처음엔 알아보지 못했다."~     2
## 10 "비 오는 날 입었던 밝은 분홍색 원피스를 그녀는 입고 있었다."            2
## 11 "비 오는 날의 밝은 핑크빛 원피스."                                       2
## 12 "비수기에, 이런 곳에, 젊은 여자가, 혼자서, 무엇하러 왔을까."             2
## 13 "스탠드 불빛이 점점 눈에 부셔오는 데 반해 모니터에서 명멸하는 작은 녹색 커서는 맛있게 느껴졌다."~     2
## 14 "재밌는 것은 비문에 적힌 거의 모든 성씨가 영남을 본관으로 두고 있다는 점이었다."~     2
## 15 "조세핀이 즐겨 입었었고, 브리짓트 바르도도 자주 입었던 복고풍 의상."     2
## 16 "해수면에 반사된 빛이 그녀의 얼굴을 환하게 비추었다."                    2
## 17 "행복한 고민으로 보여요.\""                                               2
## 18 "\"특별한 건 없어요."                                                     1
## 19 "3 일이면 충분했던 것이 올해는 석 달을 넘게 다녀도 좋아지질 않았다."     1
## 20 "그녀가 고맙기도 했지만, 그녀에게 고마움을 표시하기엔 아직 난감한 기분을 처리하지 못한 상태였던 것이다."~     1
```

```
# 부정 문장 상위 20 개
kafka_score %>%
```

```
  arrange(score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                                 score
##    <chr>                                                                    <dbl>
##  1 "요행히 목숨이 붙은 아이들은 가난한 마을 사람들한테 신발과 옷을 빼앗기고 얼어죽거나, 여자아이일 경우엔 거기에다 강간까지 ~    -4
##  2 "어머니는 지치고 체념을 하면서도 아픔을 견디지 못해......"                -3
##  3 "\"곧 된다구 해."                                                        -2
##  4 "\"귀가 정상이 아니니까 평형감각에 자꾸 장애가 생기는 겁니다."           -2
##  5 "\"제가 요즘 중이염을 앓고 있거든요.\""                                  -2
##  6 "가뜩이나 글이 써지지 않는데 초성이 ㅈ자로 시작되는 단어가 떠오르면 반사적으로 긴장이 되고 짜증이 났다."~    -2
##  7 "귀에 심각한 이상이 생긴 걸 안다면 그것 정도 모를 위인이 아니지."        -2
##  8 "그녀는 내가 대흥사 입구의 모텔에서 아침마다 쓸쓸하게 토스트를 씹고 커피를 마시는 이유를 알겠다는 표정을 지었다."~    -2
##  9 "나중에는 말의 내용조차 장애를 일으키는 거예요."                        -2
## 10 "난 아직도 오른쪽 귀가 대단히 불편한 것이다."                           -2
## 11 "내가 객관적 세계를 인식하는 방식은 이토록 게으르고 둔하다."            -2
## 12 "네, 소설 쓰며 살아요.\""                                               -2
## 13 "다 됐다구 해.\""                                                       -2
## 14 "매 장마다 소름이 끼치는 그 소설 9 장에는, 집단학살장으로 끌려가는 것을 알아버린 유태인들이 어린 자식들을 달리는 열차 밖으~    -2
## 15 "문장이 뒤틀리고 얘기가 빗나가 괴상한 몰골의 소설들만 나왔어요."       -2
## 16 "벽에서 곰팡이 냄새가 났다."                                            -2
## 17 "서로에게 무관심하기가 고문을 견디는 것보다 어렵게 느껴졌다."           -2
## 18 "어느새 해가 많이 기울어 있었다."                                       -2
## 19 "어떤 국민학교에 한국인 아이 하나가 전학을 왔는데 학교장이 조회시간에 그 아이에 대해 각별한 관심을 보였고, 담임은 그 아이~    -2
## 20 "어쩌다 재수없게 더러운 인간이 걸려들었다고 치부해 버리기엔 그녀가 겪은 일본은 너무 야비했다."~    -2
```

- 우리들의 일그러진 영웅

```
# 긍정 문장 상위 20 개
hero_score %>%
  arrange(-score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                                 score
##    <chr>                                                                    <dbl>
```

1 아름답고 상냥한 여선생님까지는 못 돼도 부드럽고 자상한 멋쟁이 선생님쯤은 될 줄 알았는데, 막걸리 방울이 튀어 하얗게 말라붙은~ 7
2 집안이 넉넉하거나 운동을 잘해 거기서 얻은 인기로 급장이 되는 수도 있었으나 대개는 성적순으로 급장 부급장이 결정되었고, 그 ~ 4
3 그에게 맡겨진 청소 검사는 우리 교실을 그 어떤 교실보다 깨끗하게 하였으며, 우리의 화단을 드러나게 환하게 했다.~ 3
4 무언가 대단히 높고 귀한 사람의 이름을 부르고 있다는, 그래서 당연히 존경과 복종을 바쳐야 한다는 그런 느낌을 주는 것이었다.~ 3
5 “거, 참 대단한 아이로구나. 2
6 “좋아. 2
7 겨우 엄석대가 그날 한 일들을 모두 얘기한 내가 막 충고를 바라는 물음을 던지려는 아버지가 불쑥 감탄 섞어 말했다.~ 2
8 그러나 아무래도 그 심부름만은 할 수 없어 잠깐 멈칫거리고 있는데 문득 좋은 생각이 떠올랐다.~ 2
9 그런데 그 새로운 급우들은 새로운 담임선생과 마찬가지로 그런 쪽으로는 별로 관심이 없었다.~ 2
10 내가 머뭇머뭇 그에게 다가가자 엄석대는 그 동안의 웃음을 사람 좋아 뵈는 미소로 바꾸며 물었다.~ 2
11 다른 건 뭘 잘해?” 2
12 대단한 추켜세움까지는 아니더라도, 최소한 내가 가진 자랑거리는 반아이들에게 일러주어, 그게 새로 시작하는 그들과의 관계에 도움~ 2
13 뒤이어 맨 앞줄의 아이 하나가 사기 컵에 물을 떠다 공손히 놓는 것까지 모두가 소풍가서 담임선생님께 하듯 했다.~ 2
14 또 그에게 맡겨진 실습 감독은 우리의 실습지에서 가장 많은 수확을 안겨 주었으며, 그의 강제 할당으로 우리 반의 비품은 그 어~ 2
15 별로 대단한 건 아니지만, 그가 주먹으로 전학년을 휘어잡아 적어도 우리 반 아이가 다른 반 아이에게 얻어맞는 일은 없게 된 것~ 2
16 선생님뿐만 아니라 아이들과의 관계에서도 내 이익을 지켜 주는 데 적지 않은 몫을 하던 내 은근한 자랑거리였다.~ 2
17 아이들이 까닭 없이 적의를 보이며 놀이에 나를 끼워 주지 않는 것도, 저희끼리 모여 무엇인가를 재미있게 떠들다가 내가 다가가면~ 2
18 어렸을 적에는 내가 똑똑한 것과 밖에 나가 다른 아이를 때리고 돌아오는 것을 일쑤 혼동하던 어머니를 늘상 호되게 나무라곤 하시~ 2
19 겨우 정신을 가다듬어 내가 한 말 어디가 그들을 웃게 만들었는지를 생각해 보고 있는데 미화 부장이라는 녀석이 웃음을 참으며 물~ 1
20 공부를 잘하는가, 힘은 센가, 집은 잘 사는가, 따위로 말하자면 나중 그 아이와 맺게 될 관계의 기초가 될 자료 수집인 셈이다~ 1

```r
# 부정 문장 상위 20 개
hero_score %>%
  arrange(score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                                  score
##    <chr>                                                                     <dbl>
##  1 주먹에서도, 편가르기에서도, 공부에서도 가망이 없어진 내가 그 다음으로 눈독을 들인 것은 석대의 약점 특히 아이들을 상대로 하~    -5
##  2 그리고는 반 아이들이 빠져 있는 불행한 상태니 그런 상태를 만들어 낸 제도 또는 그 제도의 그릇된 운용에 화낼 것 없이 엄석대~    -4
##  3 내가 무슨 바보 같은 소리를 했다는 듯, 그때껏 나를 을러대던 두 녀석과 엄석대까지를 포함한 쉰 몇 명 모두가 홍소(哄笑)였다~    -4
##  4 학교에서는 내가 갑자기 던져지게 된 그 환경의 지나친 생소함에서 온 어떤 정신적인 마비와, 또한 갑자기 나를 억눌러 오는 그 ~    -4
##  5 머리 기름은커녕 빗질도 안 해 부스스한 머리에 그날 아침 세수를 했는지가 정말로 의심스런 얼굴로 어머님의 말씀을 듣는 둥 마는~    -3
##  6 석대의 보고를 가만히 듣고 있다가 흑판 지우개를 터는 막대기로 벌서고 있는 아이의 손바닥을 몇 차례 호되게 때려 줌으로써 내게~    -3
##  7 이미 약이 오를 대로 오른 내 눈에는 엄석대 조차 보이지 않았다 그러자 엄석대는 거칠게 도시락 뚜껑을 닫고는 험한 얼굴로 내게~    -3
##  8 “너는 저기 앉도록 해.                                                    -2
##  9 그 뒤 일 년에 걸친 악연(惡緣)이 그때 벌써 어떤 예감으로 와 닿았는지 모를 일이었다.~    -2
## 10 그들은 석대에게 어떤 본능적인 공포 같은 걸 품은 듯했다.                  -2
## 11 그때껏 버티고 있는 나를 미워하는 기색을 보이기는커녕 초조해하는 눈치조차 없었다.~    -2
## 12 그때마다 내 마음 속에서는 한층 더 치열하게 적의가 타올랐으며 그리하여 그것은 그 뒤의 길고 힘든 싸움을 내가 견뎌 낼 수 있~    -2
## 13 그런데 겨울 교실 하나 넓이의 그 교무실에는 시골 아저씨들처럼 후줄그레한 선생님들이 맥없이 앉아 굴뚝같이 담배 연기만 뿜어 대~    -2
## 14 그럼에도 불구하고 나는 모반(謀叛)의 열정과도 비슷한, 가망이 없을수록 더 치열해지는 비뚤어진 집착으로 그 힘든 싸움을 계속해~    -2
## 15 금세라도 큰 주먹을 내지를 것 같은 그 무서운 기세에 그제야 덜컥 겁이 난 나는 몸을 일으켰다.~    -2
## 16 나는 그 또한 매몰차게 거절했다.                                         -2
## 17 나는 문득 무엇인가 큰 잘못을 하고 있다는 느낌, 특히 담임 선생님이 부르시는데 뻗대고 있었던 것과 흡사한 착각이 일었다.~    -2
## 18 나는 진작부터 아이들의 박해와 석대의 구원 사이를 연결하고 있는 보이지 않는 끈을 직감으로 느끼고 있었으며, 결국은 그것이 나~    -2
## 19 내가 석대의 나쁜 짓을 캐 모으려 한 것은 그것으로 먼저 담임선생과 그를 떼어놓기 위함이었
```

```
다.~    -2
## 20 너는 왜 언제나 걔를 뺀 나머지 아이들 가운데만 있으려고 해?                -2
```

- 꺼삐딴 리

```
# 긍정 문장 상위 20 개
lee_score %>%
  arrange(-score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                              score
##    <chr>                                                                 <dbl>
##  1 이왕이면 한국 여성과 결혼했으면 좋겠다던 솔직한 고백에, 자기의 학문을 위한 탁월한 견해라고 무심코 찬의를 표한 것도 자기가 ~     3
##  2 고급 중학을 졸업하고 의과 대학에 입학된 바로 그해다.                       2
##  3 그러나 이제는 평당 50 만 환을 호가하는 도심지에 타일을 바른 2 층 양옥을 소유하게 되었다.~ 2
##  4 그렇게 중환자가 아닌 한 대부분의 경우, 예진(豫診)은 젊은 의사들이 했다.     2
##  5 그리고는 지금 통장의 잔액을 깡그리 내주던 은행 지점장의 호의에 새삼 고마움을 느끼는 것이었다.~     2
##  6 꼭 풀 쑤어 개 좋은 일을 한 것만 같은 몸서리가 느껴졌다.                  2
##  7 수술을 끝낸 찰나 스쳐 가는 육감 그것은 성공 여부의 적중률을 암시하는 계시 같은 것이다.~ 2
##  8 아들의 대답이 그에게는 믿음직스럽게 여겨졌다.                           2
##  9 아무튼 그 노서아 말 꾸준히 해라.”                                       2
## 10 아직 해방의 감격이 온 누리를 뒤덮어 소용돌이칠 때였다.                  2
## 11 이 종잇장 하나만 해도 일본인과의 교제에 있어서 얼마나 떳떳한 구실을 할 수 있었던 것인가.~ 2
## 12 이인국 박사는 백 촉 전등이 너무 환한 것이 못마땅했다.                  2
## 13 이인국 박사의 말씨는 점잖게 가라앉았다.                                2
## 14 젊은 의사들에게 맡겨 버리면 그만이다.                                  2
## 15 포도에 뒤끓는 사람들은 손에 손에 태극기와 저기(赤旗)를 들고 환성을 울리고 있었다.~     2
## 16 “얘, 너 그 노어 공부를 열심히 해라.”                                   1
## 17 “잠꼬대까지 국어로 할 정도가 아니면 이 영예로운 기회야 얻을 수 있겠소.”     1
## 18 간헐적으로 반복되어 공포와 감격을 함께 휘몰아치는 착잡한 추억.          1
## 19 굳게 닫혀 있는 은행 철문에 붙은 벽보가 한길을 건너 하얀 윤곽만이 두드러져 보인다.~     1
## 20 그 속에서도 각모(角帽)와 쓰메에리 학생복을 벗어버리고 신사복으로 갈아입던 그날의 감회를 더욱 새롭게 해주는 충동을 금할 길 ~     1
```

```
# 부정 문장 상위 20 개
lee_score %>%
```

```
  arrange(score) %>%
  head(20)
```

```
## # A tibble: 20 x 2
##    sentence                                                                  score
##    <chr>                                                                     <dbl>
##  1 건방지게, 이게 새삼스레 아비에게 설교조로……좀더 솔직하지 못하고…….          -4
##  2 일제 시대, 소련국 점령하의 감옥 생활, 6.25 사변, 삼팔선, 미군 부대, 그 동안 몇 차례의 아슬아슬한 죽음의 고비를 넘~    -4
##  3 아들의 출발을 앞두고, 걱정하는 마누라를 우격다짐으로 무마시키고 그는 아들의 유학을 관철하였다.~    -3
##  4 잠을 깨어 울고 있는 어린것에게 젖을 물리고 있는 아내의 젊은 육체에서 자극을 느끼면서 이인국 박사는 자기 자신이 죄를 지은 ~    -3
##  5 '더러운 년 같으니, 기어코…….'                                          -2
##  6 '흥 혁명 유가족두 가기 힘든 구멍을 이인국의 아들이 뚫었으니 어디 두구 보자…….'~    -2
##  7 "저 병 말이오."                                                          -2
##  8 그는 자위인지 체념인지 모를 푸념을 곱씹었다.                            -2
##  9 그러나 이인국 박사는 일류 대학 병원에까지 손을 쓰지 못하여 밀려오는 급환자들 틈에 끼여 환자의 감별에는 각별한 신경을 쓰고 ~    -2
## 10 무슨 세상이 되든 할대로 해 봅시다."                                     -2
## 11 무슨 영문인지 모르고 어리벙벙하던 이인국 박사는, 그것이 언젠가 입원을 거절당한 사상범 환자 춘석이라는 것을 혜숙에게서 듣고야~    -2
## 12 우연한 일은 아니다.                                                      -2
## 13 이인국 박사는 심각한 표정으로 말을 이었다.                              -2
## 14 이인국 박사는 안내의 얼굴을 직시하지는 못하고 마치 독백하듯이 뇌까렸다.    -2
## 15 일본인 간부급들이 자기 집처럼 들락날락하는 이 병원에 이런 사상범을 입원시킨다는 것은 관선 시의원이라는 체면에서도 떳떳치 못할~    -2
## 16 지축이 흔들리는 것 같은 동요와 소름이 가까워졌다.                      -2
## 17 현대 의학이 인간의 평균 수명을 연장하고, 암 같은 고질이 아닌 한 불의의 죽음은 없다 하지만, 자기 자신이 의사이면서 스스로~    -2
## 18 환자는 아직 혼수 상태에서 깨지 못하고 있다.                             -2
## 19 '개천에서 용마가 난다는데 이건 제 애비만도 못한 자식이야.'             -1
## 20 '흰둥이 손자' 생각만 해도 징그럽다.                                     -1
```

작품의 감정 범주별 단어빈도를 로그오즈비를 통해 나타내기

감정 범주별 단어 빈도 구하기

- score 기준으로 문장의 감정을 분류해줍니다.

- 문장별 감정 점수가 부여된 (작품이름)_score_comment 를 단어 기준으로 토큰화 하고 의미를 해석할 수 있게 두 글자 이상의 한글 단어만 남깁니다.

```
kafka_score_comment<-kafka_score %>%
  mutate(sentiment = ifelse(score>=1,"pos",ifelse(score<=-1,"neg","neu")))

comment_kafka<-kafka_score_comment %>%
  unnest_tokens(input = sentence,output=word,token="words",drop=F) %>%
  filter(str_detect(word,"[가-힣]") & str_count(word)>=2)

hero_score_comment<-hero_score %>%
  mutate(sentiment = ifelse(score>=1,"pos",ifelse(score<=-1,"neg","neu")))

comment_hero<-hero_score_comment %>%
  unnest_tokens(input = sentence,output=word,token="words",drop=F) %>%
  filter(str_detect(word,"[가-힣]") & str_count(word)>=2)

lee_score_comment<-lee_score %>%
  mutate(sentiment = ifelse(score>=1,"pos",ifelse(score<=-1,"neg","neu")))

comment_lee<-lee_score_comment %>%
  unnest_tokens(input = sentence,output=word,token="words",drop=F) %>%
  filter(str_detect(word,"[가-힣]") & str_count(word)>=2)
```

감정 범주별 단어 빈도 구하기

- sentiment 별 word 의 빈도를 구합니다.
- 중립 댓글을 제거한 다음 wide form 으로 변환해 로그 오즈비를 구합니다.

```
frequency_kafka<-comment_kafka %>%
  count(sentiment,word,sort =T)

kafka_wide<-frequency_kafka %>%
  filter(sentiment!="neu") %>%
  pivot_wider(names_from = sentiment,
              values_from = n,values_fill = list(n = 0))

kafka_wide<-kafka_wide %>%
  mutate(log_odds_ratio = log(((pos+1)/(sum(pos+1)))/((neg+1)/(sum(neg+1)))))



frequency_hero<-comment_hero %>%
  count(sentiment,word,sort =T)

hero_wide<-frequency_hero %>%
  filter(sentiment!="neu") %>%
  pivot_wider(names_from = sentiment,
              values_from = n,values_fill = list(n = 0))

hero_wide<-hero_wide %>%
  mutate(log_odds_ratio = log(((pos+1)/(sum(pos+1)))/((neg+1)/(sum(neg+1)))))
```

```r
frequency_lee<-comment_lee %>%
  count(sentiment,word,sort =T)

lee_wide<-frequency_lee %>%
  filter(sentiment!="neu") %>%
  pivot_wider(names_from = sentiment,
              values_from = n,values_fill = list(n = 0))

lee_wide<-lee_wide %>%
  mutate(log_odds_ratio = log(((pos+1)/(sum(pos+1)))/((neg+1)/(sum(neg+1)))))


head(kafka_wide)
```

```
## # A tibble: 6 x 4
##   word     neg   pos log_odds_ratio
##   <chr>  <int> <int>          <dbl>
## 1 나는      10     3         -1.01
## 2 그녀가     2     6          0.851
## 3 거예요     5     0         -1.79
## 4 라고       2     5          0.697
## 5 내가       4     0         -1.61
## 6 저는       4     0         -1.61
```

```r
head(hero_wide)
```

```
## # A tibble: 6 x 4
##   word     neg   pos log_odds_ratio
##   <chr>  <int> <int>          <dbl>
## 1 나는       9     3         -0.908
## 2 내가       6     9          0.365
## 3 있는       7     3         -0.685
## 4 같은       6     1         -1.24
## 5 나를       6     3         -0.552
## 6 석대의     6     1         -1.24
```

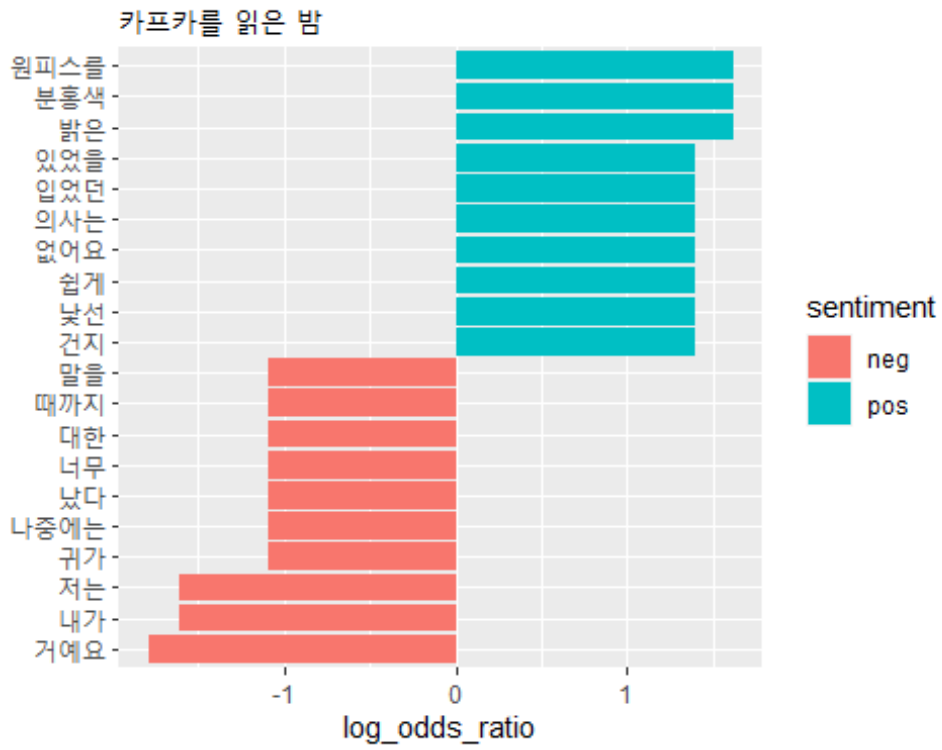
```r
head(lee_wide)
```

```
## # A tibble: 6 x 4
##   word     neg   pos log_odds_ratio
##   <chr>  <int> <int>          <dbl>
## 1 박사는     6     3         -0.510
## 2 없다       6     0         -1.90
## 3 이인국     6     5         -0.104
## 4 그는       5     3         -0.356
```

```
## 5 있다       4     0         -1.56
## 6 자기       4     0         -1.56
```

그래프로 결과 확인

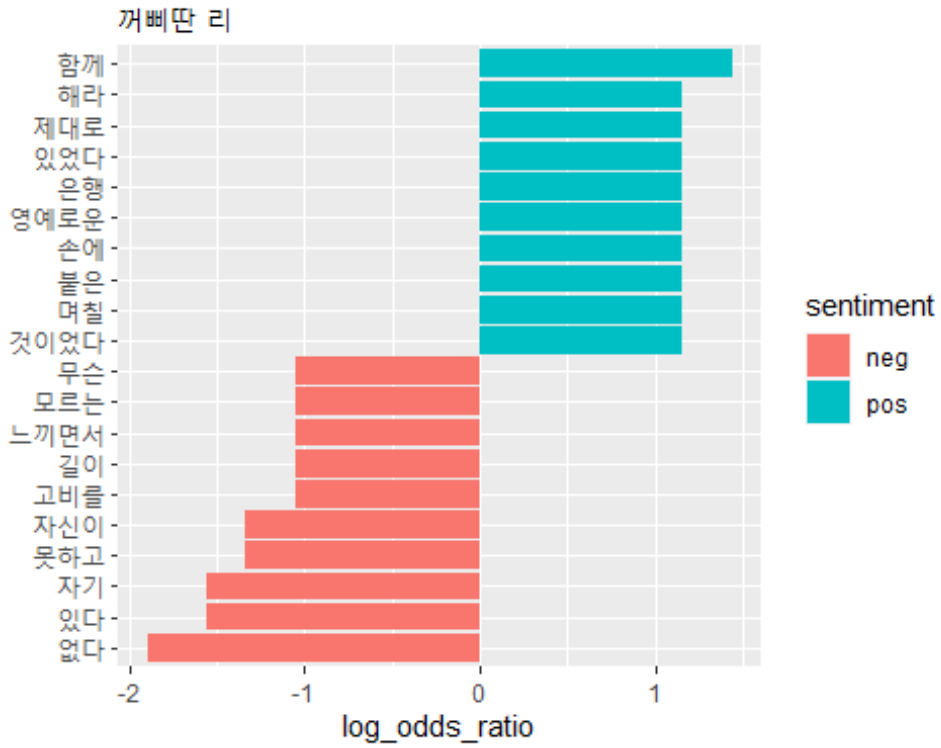
```
top10_kafka_wide<-kafka_wide %>%
  group_by(sentiment = ifelse(log_odds_ratio>0,"pos","neg")) %>%
  slice_max(abs(log_odds_ratio),n=10,with_ties = F)



top10_hero_wide<-hero_wide %>%
  group_by(sentiment = ifelse(log_odds_ratio>0,"pos","neg")) %>%
  slice_max(abs(log_odds_ratio),n=10,with_ties = F)



top10_lee_wide<-lee_wide %>%
  group_by(sentiment = ifelse(log_odds_ratio>0,"pos","neg")) %>%
  slice_max(abs(log_odds_ratio),n=10,with_ties = F)



ggplot(top10_kafka_wide,aes(x = reorder(word,log_odds_ratio),
                            y = log_odds_ratio,
                            fill = sentiment)) +
  geom_col() + coord_flip() + labs(title ="카프카를 읽은 밤", x=NULL)
```




```
ggplot(top10_hero_wide,aes(x = reorder(word,log_odds_ratio),
                            y = log_odds_ratio,
```

```
                            fill = sentiment)) +
  geom_col() + coord_flip() + labs(title = "우리들의 일그러진 영웅",x=NULL)
```


```
ggplot(top10_lee_wide,aes(x = reorder(word,log_odds_ratio),
                          y = log_odds_ratio,
                          fill = sentiment)) +
  geom_col() + coord_flip() + labs(title = "꺼삐딴 리",x=NULL)
```



작품별 중요단어를 TF-IDF 를 통해 나타내기

tibble 구조로 변환 및 소설명 변수 추가

- raw data 들을 tibble 형으로 바꿔줍니다.
- 데이터가 어느 소설에 해당하는지 구분하기 위해 novel 이라는 변수를 추가합니다.

```
kafka<-raw_kafka %>%
  as_tibble() %>%
  mutate(novel = "kafka")
hero<-raw_hero %>%
  as_tibble() %>%
  mutate(novel = "hero")
lee<-raw_lee %>%
  as_tibble() %>%
  mutate(novel = "lee")
```

세 데이터 합치기

- 앞서 tibble 형 변환 후 novel 변수를 추가한 세 작품의 데이터를 합쳐줍니다.

```
bind_novel<-bind_rows(kafka,hero,lee) %>% select(novel,value)
bind_novel
```

```
## # A tibble: 647 x 2
##    novel value
##    <chr> <chr>
##  1 kafka <U+FEFF>그녀는 아주 밝은 분홍색 원피스를 입고 있었다. 패션 용어를 빌려 더 정확하게 말하자면, 라이트 핑크/플레어 라인 실루엣을 ~
##  2 kafka 지루하게까지 느껴지는 이 용어는 내가 그녀와 작별하고 대략 32 시간정도 경과한 뒤 광주 고속버스 터미널 구내매점에서 산, 매일~
##  3 kafka <시티 라이프>를 산 건 거의 충동적인 동기에서였다. 표 4 에 장 콕토의 연필 데생인 듯한 카프카 초상이 눈을 부릅뜨고 있었던 ~
##  4 kafka 그녀는 날씬한 체형도 아니었고, 키가 크고 살이 쪄 보이지도 않았다. 남의 눈에 띌 만한 특징이 없는, 그야말로 보통의 체구였~
##  5 kafka 상쾌한 날의 상쾌한 옷차림이었다면 뭐 그리 돋보일 게 있었겠는가. 그런데 그날은 때늦은 봄비가 내리고 있었다. 비 오는 날의 ~
##  6 kafka 제법 굵은 빗줄기가 떨어져 내리는 삼거리에 그녀는 서 있었다. 우산이 없는 사람들은 윗저고리 깃을 정수리까지 올리고 빗물로 번~
##  7 kafka 그녀 발치의 웃자란 풀줄기들이 길 가운데로 기어 나오고 있었다. 차량의 왕래가 빈번하지 않은 곳이었다. 얼핏 봐서는 그녀가 사~
##  8 kafka 난 우산을 쓰고 있었고, 반바지에 맨발에 슬리퍼 차림이었고, 무엇보다 빗길을 걷는 일밖엔 아무것도 할 일이 없었으므로, 다른 ~
##  9 kafka 마침 그때 봉고 한 대가 빗물을 튀기며 그녀 곁을 스쳐 지나갔다. 도로를 질주하는 자동차에는 별 관심이 없는 듯 그녀는 멈칫거~
```

```
## 10 kafka 나는 그녀가 누군가에게 무슨 말인가를 물으려 한다는 사실을 곧 깨달았다. 그러나 뜸하게 오가는 사람들조차 비 때문에 지나치게 ~
## # ... with 637 more rows
```

전처리, 토큰화, 단어 빈도

- 전처리: 특수문자들을 모두 제거해줍니다.
- 명사 토큰화: 명사로 토큰화해줍니다.
- 단어 빈도 구하기: 길이가 2 이상인 단어들만 개수를 세줍니다.

```
novels<-bind_novel %>%
  mutate(value = str_replace_all(value,"[^가-힣]"," "),
         value = str_squish(value))


novels <- novels %>%
  unnest_tokens(input = value,
                output = word,
                token = extractNoun)

frequency<-novels %>%
  count(novel,word) %>%
  filter(str_count(word)>1)
frequency
```

```
## # A tibble: 2,899 x 3
##    novel word       n
##    <chr> <chr>  <int>
##  1 hero  가교사     1
##  2 hero  가늠       3
##  3 hero  가능성     2
##  4 hero  가능했     1
##  5 hero  가로       1
##  6 hero  가망       3
##  7 hero  가운데     3
##  8 hero  가을       1
##  9 hero  가족       1
## 10 hero  가중       1
## # ... with 2,889 more rows
```

TF-IDF 구하기

```
frequency<-frequency %>%
  bind_tf_idf(term = word,
              document = novel,
              n = n) %>%
  arrange(-tf_idf)
frequency
```

```
## # A tibble: 2,899 x 6
##    novel word         n      tf   idf  tf_idf
##    <chr> <chr>    <int>   <dbl> <dbl>   <dbl>
##  1 kafka 그녀       154 0.0679   1.10 0.0746
##  2 lee   이인국      46 0.0290   1.10 0.0318
##  3 lee   박사        45 0.0283   1.10 0.0311
##  4 hero  급장        26 0.0188   1.10 0.0207
##  5 hero  담임선생    13 0.00941  1.10 0.0103
##  6 lee   환자        14 0.00882  1.10 0.00969
##  7 lee   아들        13 0.00819  1.10 0.00899
##  8 hero  학년        10 0.00724  1.10 0.00795
##  9 hero  어른         9 0.00651  1.10 0.00715
## 10 kafka 소설        14 0.00618  1.10 0.00678
## # ... with 2,889 more rows
```

중요단어 상위 10 개 추출

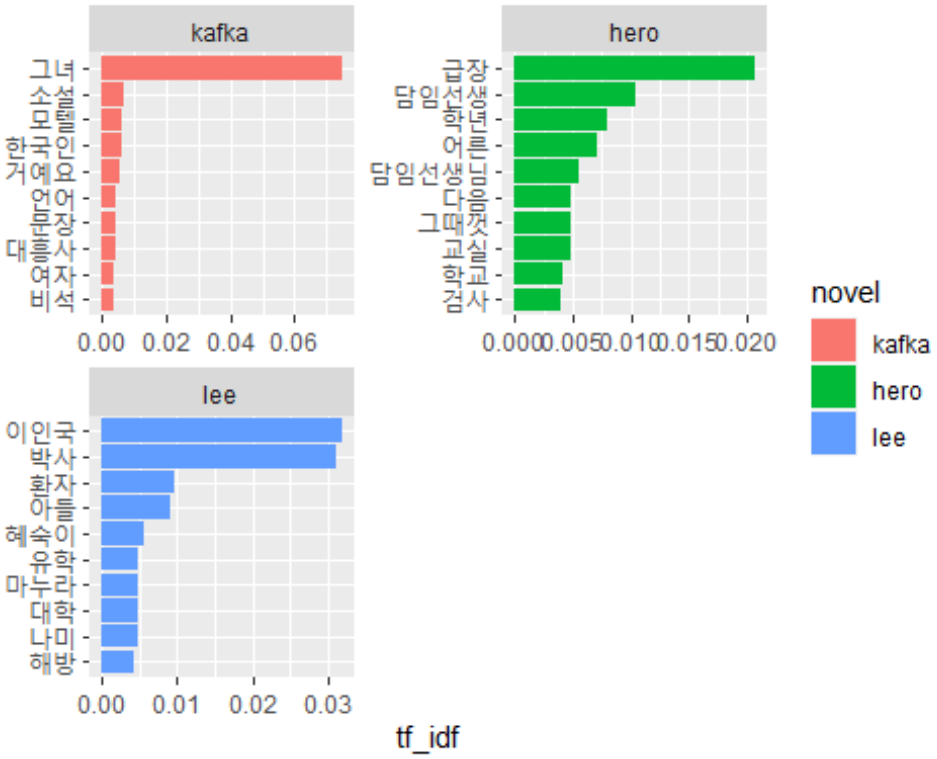
```
top10<-frequency %>%
  group_by(novel) %>%
  slice_max(tf_idf,n= 10, with_ties = F)
top10
```

```
## # A tibble: 30 x 6
## # Groups:   novel [3]
##    novel word           n      tf   idf  tf_idf
##    <chr> <chr>      <int>   <dbl> <dbl>   <dbl>
##  1 hero  급장          26 0.0188  1.10  0.0207
##  2 hero  담임선생      13 0.00941 1.10  0.0103
##  3 hero  학년          10 0.00724 1.10  0.00795
##  4 hero  어른           9 0.00651 1.10  0.00715
##  5 hero  담임선생님     7 0.00507 1.10  0.00556
##  6 hero  교실           6 0.00434 1.10  0.00477
##  7 hero  그때껏         6 0.00434 1.10  0.00477
##  8 hero  다음           6 0.00434 1.10  0.00477
##  9 hero  학교          14 0.0101  0.405 0.00411
## 10 hero  검사           5 0.00362 1.10  0.00397
## # ... with 20 more rows
```

그래프로 결과 확인

```
# 그래프 순서 정하기
top10$novel<-factor(top10$novel,level = c("kafka","hero","lee"))

# 막대 그래프 만들기
ggplot(top10,aes(x=reorder_within(word,tf_idf,novel),
                 y = tf_idf, fill = novel))+
```

```
  geom_col(show.legend = T) +
  coord_flip() + facet_wrap(~novel,scales = "free", ncol = 2)+
  scale_x_reordered() + labs(x = NULL)
```



결과분석

- 카프카를 읽은 밤에서는 그녀 라는 단어가 다른 작품에 비해서 많이 사용된 것을 확인할 수 있었습니다.
- 우리들의 일그러진 영웅에서는 급장과 담임선생 이 두 단어가 다른 작품에 비해 많이 사용되었습니다. 아마 학창시절을 다루는 소설이다보니 이 두 단어가 다른 작품에 비해 많이 사용된 것 같습니다.
- 꺼삐딴 리 작품에서는 이인국과 박사라는 단어가 다른 작품에 비해 많이 사용되었습니다. 꺼삐딴 리의 작 중 주인공이 이인국 박사라는 점을 고려해 보았을 때 어찌보면 당연한 결과일 수도 있다는 생각이 들었습니다.