RESPONSI_DS-(F)

DemitriesBaskharaRivaldoTolla_123180137

Intro

0. Cuci tangan dengan sabun hingga benar-benar bersih dengan durasi mencuci tangan kurang lebih 20 dtk
1. Kerjakan soal-soal yang ada! Jangan lupa AUTHOR diberi nama (pada bagian atas soal ini)
2. Responsi terdiri dari 2 bagian yaitu bagian pertama dan bagian kedua
3. Jawab dengan membuat chunk dibawah soal!
4. Durasi pengerjaan sesuai kesepakatan yaitu 2 jam mulai pukul 20.00 hingga 22.15 tanggal 22 Januari 2021, 15 menit diberikan untuk pengumpulan hasil responsi
5. No toleransi pengumpulan telat. Telat tiap 3 menit akan ada pengurangan nilai 5 point dengan maksimal pengurangan 25 point. Telat lebih dari 15 menit atau melebihi pukul 22.30 dianggap **GUGUR**.
6. Soal yang rancu bisa menghubungi asisten terkait.
7. Pengumpulan hanya dalam bentuk **WORD Document atau PDF**. Jika pengumpulan dalam bentuk **Rmd** akan dianggap tidak mengumpulkan jawaban. Pastikan jawaban dapat dijalankan dengan baik.
8. Tenang, untuk responsi kali ini nilai akan diobral, nilai maksimal adalah 350 dari 100. Jadi, kemungkinan dapat nilai bagus besar kok.
9. Isi juga review/feedback/kritik/saran/masukan yang sudah disediakan di bagian paling bawah soal. ***WAJIB***
10. Jawaban dikumpulkan dengan format file JAWABAN_responsi_ dalam bentuk PDF atau DOKUMEN.

Persiapan

Load library apa saja yang kira-kira digunakan! Lalu load dataset 'googleplay.csv' dan 'googleplay_user_review.csv'!

```
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(data.table)
```

```
##
## Attaching package: 'data.table'

## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
library(ggplot2)
library(stringr)
library(tidytext)
data1 <- read.csv("googleplaystore.csv")
data2 <- read.csv("googleplaystore_user_reviews.csv")
```

Bagian Pertama

1. Tampilkan TOP 10 Aplikasi berdasarkan peringkat PENILAIAN/RATING yang diberikan user! **point 10**

```
data1[order(-data1$Rating),] %>%
  head(10)
```

```
##                                           App Category Rating Reviews Size
## 282               Hojiboy Tojiboyev Life Hacks   COMICS      5      15  37M
## 496              American Girls Mobile Numbers   DATING      5       5 4.4M
## 498                                Awake Dating   DATING      5       2  70M
## 504                       Spine- The dating app   DATING      5       5 9.3M
## 506 Girls Live Talk - Free Text and Video Chat   DATING      5       6 5.0M
## 507                     Online Girls Chat Group   DATING      5       5 5.0M
## 510            Speeding Joyride & Car Meet App   DATING      5       3  25M
## 765                              SUMMER SONIC app   EVENTS      5       4  61M
## 767                                   Prosperity   EVENTS      5      16 2.3M
## 772                   Mindvalley U Tallinn 2018   EVENTS      5       1  21M
##     Installs Type Price Content.Rating Genres Last.Updated Current.Ver
## 282   1,000+ Free     0       Everyone Comics    26-Jun-18           2
## 496   1,000+ Free     0     Mature 17+ Dating    17-Jul-18           3
## 498     100+ Free     0     Mature 17+ Dating    24-Jul-18       2.2.9
## 504     500+ Free     0           Teen Dating    14-Jul-18           4
## 506     100+ Free     0     Mature 17+ Dating     1-Aug-18         8.2
## 507     100+ Free     0     Mature 17+ Dating     2-Aug-18         8.2
## 510     100+ Free     0     Mature 17+ Dating    20-Jul-18       1.2.9
## 765     500+ Free     0       Everyone Events    24-Jul-18           1
## 767     100+ Free     0       Everyone Events     9-Jul-18        1.14
## 772     100+ Free     0       Everyone Events     3-Jul-18       1.0.5
##      Android.Ver
## 282 4.0.3 and up
## 496 4.0.3 and up
## 498   4.4 and up
## 504 4.0.3 and up
## 506 4.0.3 and up
## 507 4.0.3 and up
## 510   4.1 and up
## 765   4.4 and up
```

```
## 767   2.0 and up
## 772   4.4 and up
```

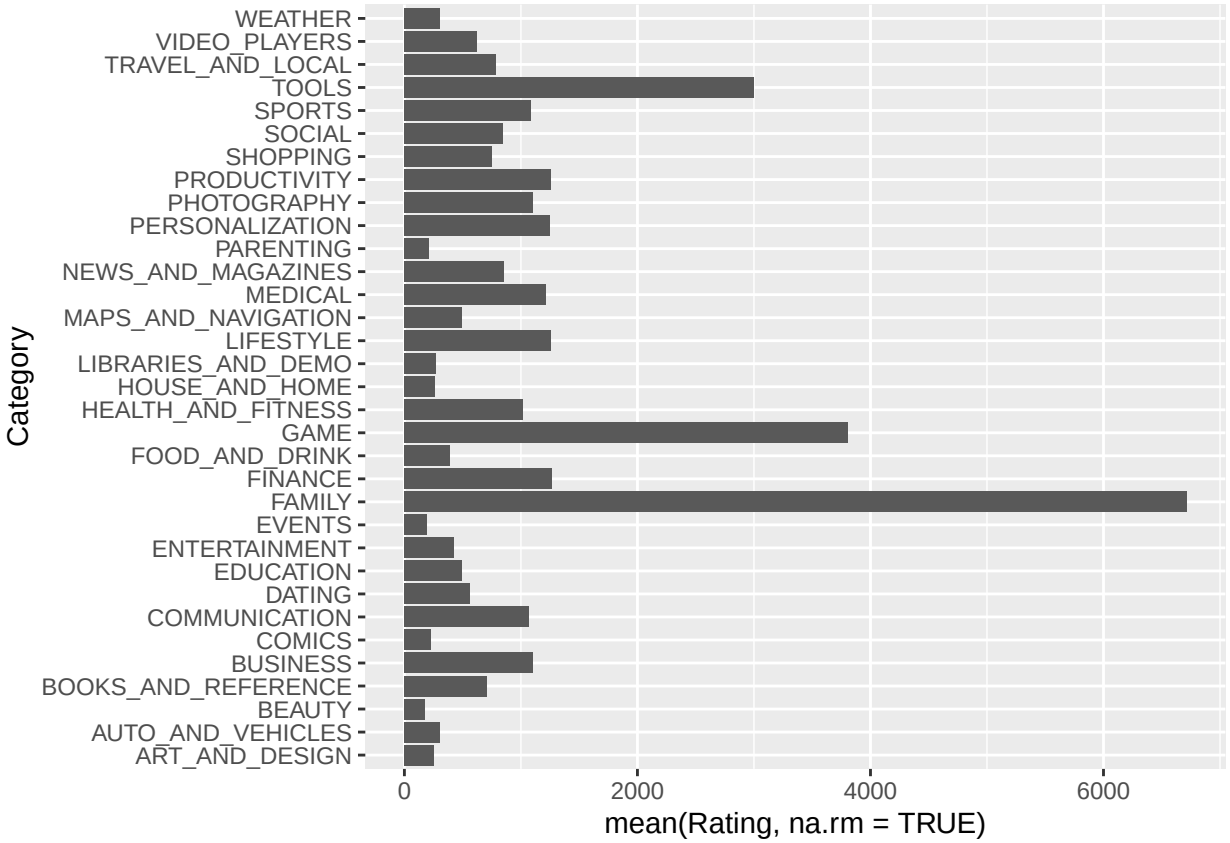
2. Tampilkan rata-rata RATING yang dihitung menggunakan fungsi buatan untuk setiap kategori aplikasi! **point 15**

```
data.table1 <- data.table(data1)
data.table1[,list(rata=mean(Rating , na.rm = TRUE)) , by = Category]
```

```
##                Category     rata
##  1:      ART_AND_DESIGN 4.357377
##  2:   AUTO_AND_VEHICLES 4.190411
##  3:              BEAUTY 4.278571
##  4: BOOKS_AND_REFERENCE 4.344970
##  5:            BUSINESS 4.098479
##  6:              COMICS 4.181481
##  7:       COMMUNICATION 4.121484
##  8:              DATING 3.970149
##  9:           EDUCATION 4.364407
## 10:       ENTERTAINMENT 4.135294
## 11:              EVENTS 4.435556
## 12:             FINANCE 4.115563
## 13:      FOOD_AND_DRINK 4.172340
## 14:  HEALTH_AND_FITNESS 4.243033
## 15:      HOUSE_AND_HOME 4.150000
## 16:  LIBRARIES_AND_DEMO 4.178125
## 17:           LIFESTYLE 4.093355
## 18:                GAME 4.247368
## 19:              FAMILY 4.179664
## 20:             MEDICAL 4.166552
## 21:              SOCIAL 4.247291
## 22:            SHOPPING 4.230000
## 23:         PHOTOGRAPHY 4.157414
## 24:              SPORTS 4.216154
## 25:    TRAVEL_AND_LOCAL 4.069519
## 26:               TOOLS 4.039554
## 27:     PERSONALIZATION 4.332215
## 28:        PRODUCTIVITY 4.183389
## 29:           PARENTING 4.300000
## 30:             WEATHER 4.243056
## 31:       VIDEO_PLAYERS 4.044595
## 32:  NEWS_AND_MAGAZINES 4.121569
## 33: MAPS_AND_NAVIGATION 4.036441
##                Category     rata
```

3. Berdasarkan soal nomor 2, buat plot untuk memvisualisasikan hasilnya! (Bentuk plot bebas) **point 15**

```
ggplot(data1, aes(mean(Rating , na.rm = TRUE), Category)) + geom_bar(stat = "identity")
```



Info untuk 2 soal 4-5:Terdapat dua dataset yang digunakan. Satu dataset untuk info aplikasi dan satu dataset lagi untuk kumpulan reviewnya.

4. Dari kedua dataset tersebut, buat satu variable data baru yang isinya NAMA APLIKASI, RATING, dan JUMLAH REVIEW Positif dan/atau Negatif dan/atau Neutral (boleh semua, boleh pilih salah satu)lalu tampilkan isi data tabel tersebut! **point 20**

```
data3 <- merge(data1,data2,by = "App")
data3 %>%
  head(10)
```

```
##                    App           Category Rating Reviews Size Installs Type
## 1  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 2  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 3  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 4  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 5  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 6  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 7  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 8  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 9  10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
## 10 10 Best Foods for You HEALTH_AND_FITNESS      4    2490 3.8M 500,000+ Free
##    Price Content.Rating          Genres Last.Updated Current.Ver  Android.Ver
## 1      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 2      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 3      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
```

```
## 4      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 5      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 6      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 7      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 8      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 9      0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
## 10     0   Everyone 10+ Health & Fitness    17-Feb-17         1.9 2.3.3 and up
##
## 1
## 2
## 3
## 4
## 5  Food list easy I predibetic, I scared. All Dr. said potatoes, rice, bread. He never told I could
## 6
## 7
## 8
## 9
## 10
##    Sentiment Sentiment_Polarity Sentiment_Subjectivity
## 1   Positive                1.0              1.0000000
## 2   Positive                0.8              0.7500000
## 3        nan                NaN                    NaN
## 4   Positive                0.8              0.7500000
## 5   Positive                1.0              0.8333333
## 6   Positive                1.0              0.3000000
## 7   Positive                0.6              0.6666667
## 8   Positive                1.0              0.6500000
## 9    Neutral                0.0              0.0000000
## 10  Positive                0.3              0.8000000
```

5. Dalam dunia data scientist, sebelum melakukan pemodelan ada baiknya data dilakukan preprocessing terlebih dahulu. Dengan dataset review yang sudah dimasukkan oleh user, lakukan sebuah preprocessing data SEDERHANA yang menurut kalian dapat dilakukan untuk dataset tersebut agar dataset bisa siap untuk dimodelkan (simpan hasil preprocessing dalam variabel baru)!

Clue : Clean, Tidy, no redundacy, no dupe, no null. **point 40**

```
cleaned_text <- data3 %>%
  group_by(App) %>%
  filter(Translated_Review == "")%>%
  ungroup()

cleaned_text <- cleaned_text %>%
  filter(str_detect(Translated_Review, "^[^>]+[A-Za-z\\d]") | Translated_Review == "",
         !str_detect(Translated_Review, "writes(:|\\.\\.\\.)$"),
         !str_detect(Translated_Review, "^In article <"))

usenet_words <- cleaned_text %>%
  unnest_tokens(word, Translated_Review) %>%
  filter(str_detect(word, "[a-z']$"),
         !word %in% stop_words$word)
```

Bagian Kedua

Referensi mengerjakan: https://www.tidytextmining.com/

1. Import library tidymodels, vroom, here, tidytext dan dua dataset ke dalam objek R **nilai 10**

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 0.1.2 --

## v broom     0.7.3      v rsample   0.0.8
## v dials     0.0.9      v tibble    3.0.4
## v infer     0.5.4      v tidyr     1.1.2
## v modeldata 0.1.0      v tune      0.1.2
## v parsnip   0.1.5      v workflows 0.2.1
## v purrr     0.3.4      v yardstick 0.0.7
## v recipes   0.1.15

## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x data.table::between() masks dplyr::between()
## x purrr::discard()      masks scales::discard()
## x dplyr::filter()       masks stats::filter()
## x data.table::first()   masks dplyr::first()
## x recipes::fixed()      masks stringr::fixed()
## x dplyr::lag()          masks stats::lag()
## x data.table::last()    masks dplyr::last()
## x recipes::step()       masks stats::step()
## x purrr::transpose()    masks data.table::transpose()
```

```
library(vroom)
library(here)
```

```
## here() starts at C:/Users/dbask/Downloads
```

```
library(tidytext)

user_reviews <- vroom(here("googleplaystore.csv"))
```

```
## Rows: 8,196
## Columns: 13
## Delimiter: ","
## chr [11]: App, Category, Size, Installs, Type, Price, Content Rating, Genres, Last Updated...
## dbl [ 2]: Rating, Reviews
##
## Use `spec()` to retrieve the guessed column specification
## Pass a specification to the `col_types` argument to quiet this message
```

```
googleplaystore <- vroom(here("googleplaystore_user_reviews.csv"))
```

```
## Rows: 64,295
## Columns: 5
## Delimiter: ","
## chr [3]: App, Translated_Review, Sentiment
## dbl [2]: Sentiment_Polarity, Sentiment_Subjectivity
##
## Use ‘spec()‘ to retrieve the guessed column specification
## Pass a specification to the ‘col_types‘ argument to quiet this message
```

2. Joining dua dataset menggunakan inner join **nilai 10**

```
joining_data <- inner_join(googleplaystore,user_reviews)
```

```
## Joining, by = "App"
```

3. Tahap pre-processing data. Ketika ingin melakukan analisis sentimen beberapa hal harus dilakukan sebelum data dapat digunakan. Bersihkan dan rapikan data dengan membuang data yang “nan” di bagian Translated_review. Setelah itu, data juga harus dibersihkan dari kata-kata yang mengandung stop_word (seperti: a, a’s, after, dll). Data yang siap diolah juga harus ditokenisasi yaitu proses membagi teks dari paragraf atau kalimat ke kata. Hasil dari tokenisasi adalah tiap baris data hanya mengandung 1 kata. **nilai 15**

```
tidy_user_reviews <- joining_data %>%
  filter(Translated_Review != "nan") %>%
  unnest_tokens(word, Translated_Review) %>%
  anti_join(stop_words)
```

```
## Joining, by = "word"
```

4. Sentimen analisis dapat menggunakan beberapa jenis metode berdasarkan sentiment lexicon. Ada beberapa sentiment lexicon seperti bing, afinn, dan nrc. Gunakan sentiment lexicon nrc untuk mendapatkan jumlah kata untuk 10 kategori nrc (positive, negative, fear, surprise, dll). ***nilai 15***

```
nrc_n <- tidy_user_reviews %>%
  inner_join(get_sentiments("nrc")) %>%
  count(sentiment, sort = TRUE)
```

```
## Registered S3 methods overwritten by 'readr':
##   method           from
##   format.col_spec  vroom
##   print.col_spec   vroom
##   print.collector  vroom
##   print.date_names vroom
##   print.locale     vroom
##   str.col_spec     vroom
```

```
## Joining, by = "word"
```

```
nrc_n
```

```
## # A tibble: 10 x 2
##    sentiment        n
##    <chr>        <int>
##  1 positive     45620
##  2 negative     25503
##  3 anticipation 25175
##  4 trust        25031
##  5 joy          22852
##  6 anger        12312
##  7 fear         12268
##  8 sadness      12189
##  9 disgust       8328
## 10 surprise      7876
```

5. Kita dapat mengetahui banyaknya kata tiap kategori nrc untuk tiap aplikasi. Cobalah untuk mencari banyak kata tiap kategori nrc yang dikelompokkan berdasarkan nama aplikasi. **nilai 15**

```
user_reviews_nrc <- tidy_user_reviews %>%
  inner_join(get_sentiments("nrc")) %>%
  group_by(Category) %>%
  count(sentiment, sort = TRUE) %>%
  spread(sentiment, n, fill = 0) %>%
  ungroup()
```

```
## Joining, by = "word"
```

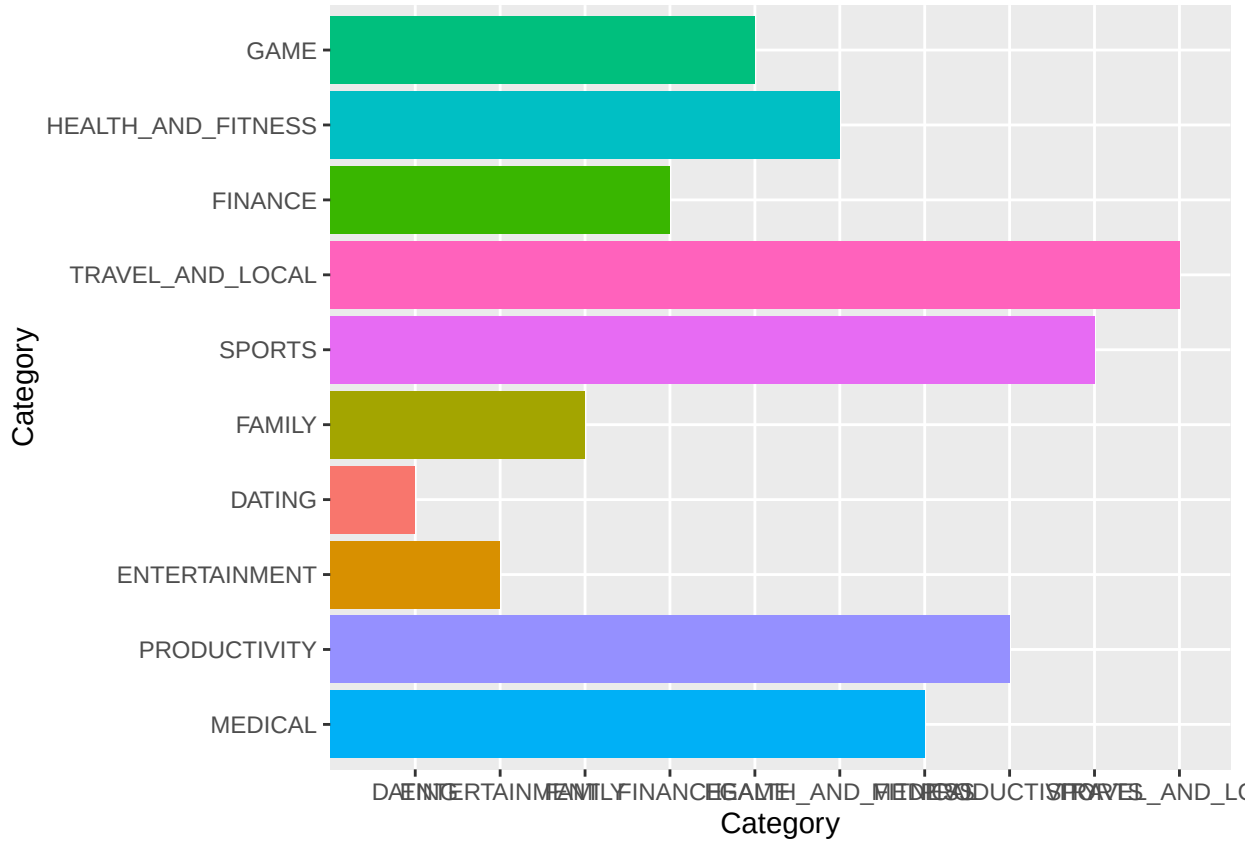
```
user_reviews_nrc
```

```
## # A tibble: 33 x 11
##    Category anger anticipation disgust  fear   joy negative positive sadness
##    <chr>    <dbl>        <dbl>   <dbl> <dbl> <dbl>    <dbl>    <dbl>   <dbl>
##  1 ART_AND~    78          172      66    83   234      147      346      89
##  2 AUTO_AN~    36          122      18    40   146       97      339      61
##  3 BEAUTY      57           65      64    53    89      111      141      58
##  4 BOOKS_A~   154          294     114   164   305      299      815     145
##  5 BUSINESS   159          403     132   185   327      432      758     199
##  6 COMICS       3           10       2     1    19        9       33       2
##  7 COMMUNI~   280          643     245   329   443      807     1029     385
##  8 DATING     521         1010     391   411   888     1078     1726     430
##  9 EDUCATI~   310          443     149   201   538      511     1560     203
## 10 ENTERTA~   466         1126     327   696   756      908     1364     457
## # ... with 23 more rows, and 2 more variables: surprise <dbl>, trust <dbl>
```

6. Setelah mendapatkan jumlah kata tiap kategori tiap aplikasi, kita dapat mngetahui aplikasi mana yang memiliki kata dengan kategori 'surprise' terbanyak untuk tiap aplikasi. Kita akan memvisualisasikan dengan grafik batang 10 aplikasi dengan jumlah kata kategori 'surprise' terbanyak. ***nilai 20***

```r
user_reviews_nrc %>%
  arrange(desc("surprise")) %>%
  top_n(10) %>%
  ggplot(aes(reorder(Category,surprise), y = Category, fill = Category)) +
  geom_col(show.legend = FALSE) +
  coord_flip() +
  labs(
    x = "Category"
  )
```

```
## Selecting by trust
```



7. Selain menggunakan sentiment lexicon 'nrc', sentimen analisis juga dapat menggunakan sentiment lexicon 'bing'. Bing hanya akan memberikan label untuk tiap kata positif atau negatif saja. Carilah kata positif yang paling umum dan kata negatif yang paling sering digunakan saat memberikan review pada aplikasi! ***nilai 15***

```r
bing_word_counts <- tidy_user_reviews %>%
  inner_join(get_sentiments("bing")) %>%
  count(word,sentiment, sort = TRUE) %>%
  head(20)
```

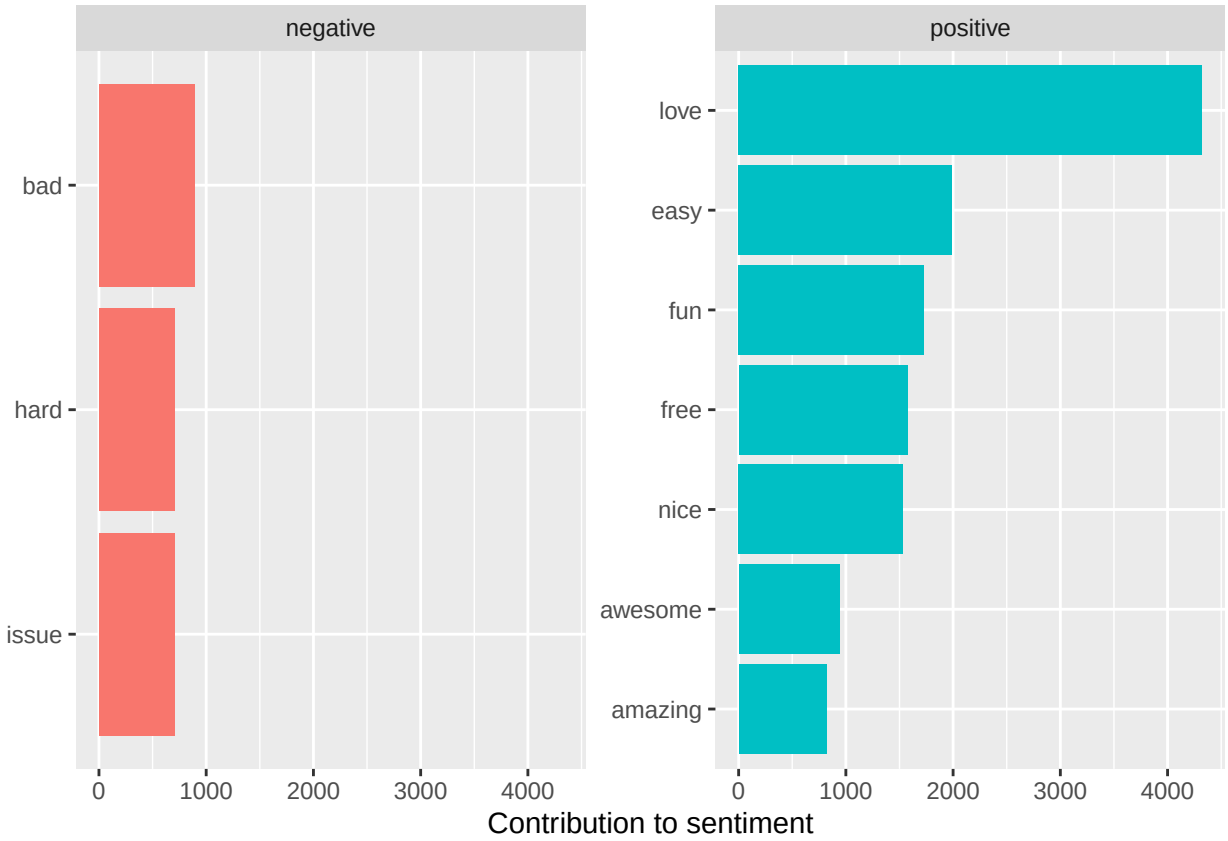
```
## Joining, by = "word"
```

```
bing_word_counts
```

```
## # A tibble: 20 x 3
##    word      sentiment     n
##    <chr>     <chr>     <int>
##  1 love      positive   4323
##  2 easy      positive   1988
##  3 fun       positive   1729
##  4 free      positive   1574
##  5 nice      positive   1531
##  6 awesome   positive    941
##  7 bad       negative    890
##  8 amazing   positive    820
##  9 hard      negative    707
## 10 issue     negative    701
## 11 annoying  negative    697
## 12 helpful   positive    594
## 13 recommend positive    582
## 14 pretty    positive    580
## 15 support   positive    514
## 16 hate      negative    505
## 17 issues    negative    501
## 18 lost      negative    498
## 19 slow      negative    490
## 20 perfect   positive    487
```

8. Pembacaan data akan lebih mudah jika ditampilkan dalam bentuk grafik. Tampilkan grafik 10 kata positif dan negatif terbanyak! ***nilai 20***

```
bing_word_counts %>%
  head(10) %>%
  ungroup() %>%
  mutate(word = reorder(word, n)) %>%
  ggplot(aes(word, n, fill = sentiment)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~sentiment, scales = "free_y") +
  labs(y = "Contribution to sentiment",
       x = NULL) +
  coord_flip()
```



9. Penganalisis data membutuhkan jumlah kata tiap kategori yang belum digabung dengan sentiment lexicon untuk menghitung rasio positif, ratio negatif dan net sentiment. Bantulah penganalisis tersebut untuk mendapatkan jumlah kata tiap kategori dari data yang sudah dirapikan! ***nilai 15***

```
rasio_penuh <- as.numeric(nrow(joining_data))
rasio_positife <- as.numeric(nrow(joining_data)) / as.numeric(nrow(filter(joining_data , Sentiment == "
rasio_negative <- as.numeric(nrow(joining_data)) / as.numeric(nrow(filter(joining_data , Sentiment == "
```

10. Selanjutnya penganalisis data ingin mendapatkan jumlah kata positif, jumlah kata negatif, rasio positif (jumlah kata positif/jumlah keseluruhan kata), rasio negatif (jumlah kata negatif/jumlah keseluruhan kata), dan net sentiment (jumlah kata positif - jumlah kata negatif) dengan menggunakan sentiment lexicon bing untuk tiap kategorinya. Tabel yang diinginkan oleh analisis adalah seperti berikut ***nilai 40***

Category	poisitive	negative	words	positive_ratio	negative_ratio	net_sentiment

```
user_reviews_nrc %>%
  select(Category,negative,positive)
```

```
## # A tibble: 33 x 3
##    Category            negative positive
##    <chr>                  <dbl>    <dbl>
##  1 ART_AND_DESIGN           147      346
##  2 AUTO_AND_VEHICLES         97      339
```

```
##  3 BEAUTY                   111      141
##  4 BOOKS_AND_REFERENCE      299      815
##  5 BUSINESS                 432      758
##  6 COMICS                     9       33
##  7 COMMUNICATION            807     1029
##  8 DATING                  1078     1726
##  9 EDUCATION                511     1560
## 10 ENTERTAINMENT            908     1364
## # ... with 23 more rows
```

Kritik/saran/masukan/feedback/review/uneg-uneg: Mantap , Lanjutkan Mas dan jangan lupa follow https://www.twitch.tv/dbaskhara

========== SELESAI ==========