Term Network

(Pelatihan data sains menggunakan R dan Gephi)

Ujang Fahmi

Pelajaran ke-9

Salam kenal dan selamat datang.
Semoga kita semua bisa saling berbagi pengalaman dan pengetahuan. Saya adalah Ujang Fahmi, Co-founder dan mentor Sadasa Academy.

Jika anda berada dan sedang membaca tutorial ini, maka kemungkinan anda adalah orang yang sedang ingin belajar data sains, atau mungkin ditugaskan untuk mempelajari R oleh institusi atau organisasi anda. Sama seperti saya dulu, dimana tanpa latar belakang enginering saya didiharuskan untuk belajar R, demi menyelesaikan tugas akhir dan akhirnya jadilah seperti saya sekarang ini.

Satu hal yang pasti, ini adalah langkah pertama dari banyak langkah yang harus dilalui, entah melalui lembaga resmi atau belajar secara mandiri. Jadi selamat belajar!!!

Ujang Fahmi,
Yogyakarta, 2021-10-24

Materi yang disampaikan disimpan dan dokumentasikan **disini**

Term Network

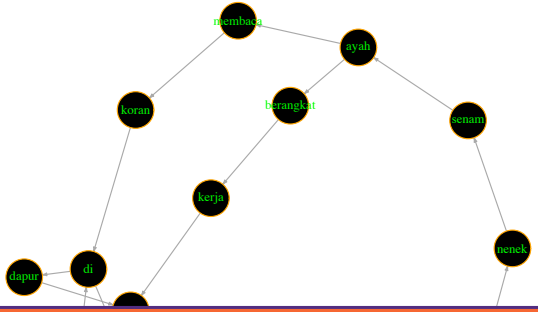
Analisis teks merupakan salah satu hal paling menantang dalam data sains, khususnya teks bahasa Indonesia. Selain karena masih terbatasnya data dan tools yang bisa digunakan, juga karena teks secara umum memiliki konteks masing-masing.

Apa?

Term Network digunakan sebagai salah satu cara untuk merepresentasikan data teks dalam sebuah jejaring, dimana nodes-nya merupakan kata, dan hubungannya ditentukan dengan posisi.
Misalnya, kita memiliki sebuah kalimat berikut:

- Ayah membaca koran di teras, Ibu membuat sarapan di dapur
- Setelah memasak ibu pergi mengantar nenek senam
- Ayah berangkat kerja setelah sarapan

Ketika di representasikan dalam sebuah network maka akan menjadi seperti berikut:



Bagiamana?

```
library(igraph)
library(tidytext)
library(tidyverse)

teks = tibble(
   kalimat =
      c("Ayah membaca koran di teras, Ibu membuat sarapan di dapur",
        "Setelah memasak ibu pergi mengantar nenek senam",
        "Ayah berangkat kerja setelah sarapan"))
teks = teks %>%
   unnest_tokens(kata, kalimat, token = "ngrams", n = 2) %>%
   separate(kata, into = c("kata1", "kata2"))

teks_net = graph_from_data_frame(teks, directed = TRUE)
```

Data yang akan digunakan

```
library(tidyverse)

# load data
tweet_save_monas <- read_csv("data/tweet_save_monas.csv",
    trim_ws = FALSE)

# Pilih kolom
tweet_save_monas = tweet_save_monas %>%
   select(id, created_at, full_text)

# cleansing full_text
source("script/fungsi_cleaning_twitter.R")
teks = tweet_cleaner(data = tweet_save_monas$full_text)
tweet_save_monas$full_text_clean = teks$clean_text

glimpse(tweet_save_monas)
```

Tokenisasi

```
library(tidytext)

# tokenisasi
token_tweet = tweet_save_monas %>%
   select(created_at, full_text_clean) %>%
   mutate(created_at = lubridate::round_date(created_at, "day")) %>%
   unnest_tokens(bigram, full_text_clean, token = "ngrams", n = 2) %>%
   count(bigram)

glimpse(token_tweet)
```

Membuat Adjacency

```
token_tweet = token_tweet %>%
   separate(bigram, into = c("sumber", "target"), sep = " ")

glimpse(token_tweet)
```

Membuat objek graph

```
library(igraph)
top_words_net = token_tweet %>%
   top_n(50)
glimpse(top_words_net)

net_topWords = graph_from_data_frame(d = top_words_net,
                                     directed = TRUE)

plot(net_topWords,
     edge.arrow.size=.2, edge.curved=0,
     vertex.color="black", vertex.frame.color="orange",
     vertex.label.color="green",
     vertex.label.cex=.7)
```

Visualisasi

```
library(visNetwork)

data <- toVisNetworkData(net_topWords)
glimpse(data)

visNetwork(nodes = data$nodes, edges = data$edges, height = "470px", wi
    visEdges(arrows = "from") %>%
   visOptions(highlightNearest = list(enabled = T, hover = T),
              nodesIdSelection = T)
```

Table of Contents