

Pemodelan Topik

(Pelatihan data sains menggunakan R dan Gephi)

Ujang Fahmi

Pelajaran ke-8



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 engineering 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, 2021-09-28

*Materi yang disampaikan disimpan dan dokumentasikan **disini***

Topic Modeling

Pemodelan topik merupakan sebuah tipe pemodelan statistik untuk mengungkap abstrak topik yang ada dalam sebuah koleksi dokumen.

Apa?

Topik modeling juga bisa disebut sebagai sebuah metode untuk menemukan kelompok kata (kemudian disebut topik) dari sebuah koleksi dokumen yang dianggap paling merepresentasikan informasi yang ada di dalamnya. Metode ini juga dapat dikelompokkan sebagai metode text mining, yaitu sebuah cara untuk mendapatkan informasi dari teks.

Topic modeling juga dikategorikan sebagai unsupervised machine learning, yaitu sebuah machine learning yang tidak membutuhkan data latih yang sebelumnya telah dikategorikan secara manual oleh manusia. Oleh karena itu juga Topic modeling membutuhkan pendekatan dan atau metode lain untuk mengevaluasinya.

LDA ?

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK—How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough. Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson, a Cornell University in Ithaca, N.Y., biologist who arrived at the 800 number in 1991, coming up with a conservative answer may be more than just a **scientific** **numbers** game, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arach Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an



* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments

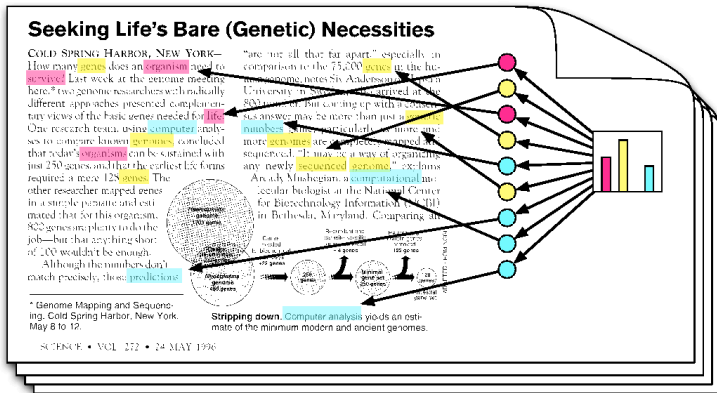


Figure 1: Ilustrasi mendapatkan topik dengan LDA

Bagaimana?

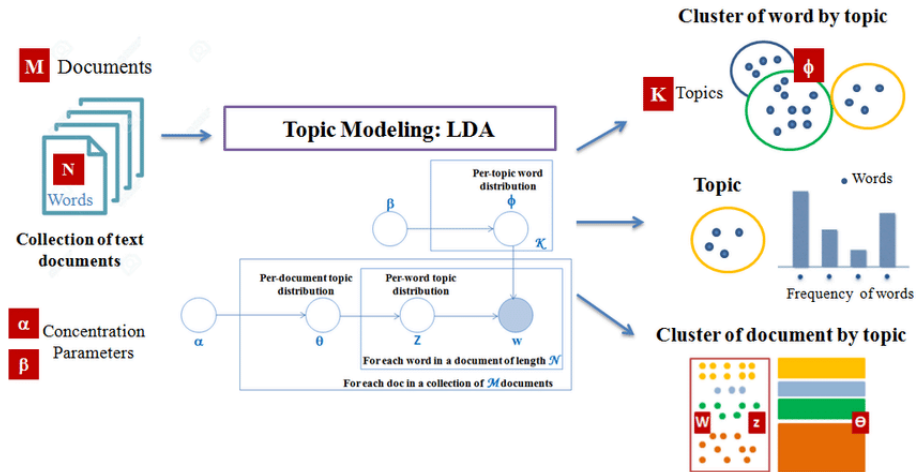


Figure 2: Cara kerja topic modeling

Langkah-langkah Topic Modeling

1. Membuat matrix term dan dokumen (document term matrices) dari data teks tabular
2. Menjalankan fungsi pemodelan topik LDA
3. Merapikan output ke dalam bentuk (tidy)
4. Menginterpretasi topik

Data yang akan dianalisis

```
library(tidyverse)
library(tidytext)
library(topicmodels)

raw_data = read_csv("data/tweet_save_monas.csv")
raw_data = raw_data %>%
  select(id, full_text_clean) %>%
  group_by(id) %>%
  unnest_tokens(word, full_text_clean, token = "words")

raw_data = raw_data %>%
  filter(!is.na(word))

glimpse(raw_data)
```


Membuat matrix term dan dokumen

```
tweet_dtm = raw_data %>%  
  count(word, id) %>%  
  cast_dtm(id, word, n) %>%  
  as.matrix()
```

Menjalan LDA

```
lda_topics <- tweet_dtm %>%  
  LDA(  
    k = 2,  
    method = "Gibbs",  
    control = list(seed = 42)  
  ) %>%  
  tidy(matrix = 'beta')  
  
lda_topics %>%  
  arrange(desc(beta))  
  
word_prob <- lda_topics %>%  
  group_by(topic) %>%  
  top_n(15, beta) %>%  
  ungroup() %>%  
  mutate(term2 = fct_reorder(term, beta))  
word_prob
```

Melihat Hasil

```
word_prob %>%  
  ggplot(aes(term2, beta, fill = as.factor(topic))) +  
  geom_col(show.legend = FALSE) +  
  facet_wrap(~topic, scales = "free") +  
  coord_flip()
```

Your Turn

Buatlah topic modelling untuk data teks dengan parameter K yang lebih dari 2 dan interpretasikan.

Table of Contents

Topic Modeling

- Apa?

- LDA ?

- Bagaimana?

Langkah-langkah Topic Modeling

- Data yang akan dianalisis

- Membuat matrix term dan dokumen

- Menjalan LDA

- Melihat Hasil

Your Turn