# EKSPLORASI HASIL DATA MINING

## Text Analisis Terkait Penelitian

Ikang Fadhli

Nutrifood Indonesia

14 December 2021

Section 1

# PENDAHULUAN

# Latar Belakang

Setelah *pilot project* dan diskusi yang lalu, berikutnya akan dicoba melakukan *data mining* kembali dengan menggunakan *keywords* yang berbeda dan lebih spesifik.

Pada kesempatan ini, saya akan kembali mencari **berbagai penelitian yang telah dilakukan di dalam negeri** terkait dengan beberapa *keywords* yang telah didefinisikan.

# Tujuan

Kali ini ini, saya mencoba untuk mencari berbagai penelitian terkait *keywords* berikut:

- *Indigenous Food* (termasuk padanan dalam bahasa Indonesianya: pangan lokal),
- *Functional Food*,
- *Fermented Food* (termasuk padanan dalam bahasa Indonesianya: makanan fermentasi),
- *Ethnic Food* (termasuk padanan dalam bahasa Indonesianya: makanan etnik),
- *Traditional Food* (termasuk padanan dalam bahasa Indonesianya: makanan tradisional),
- Makanan,
- Minuman

di situs `www.neliti.com` sebagai uji coba untuk melakukan analisa teks yang didapatkan. Dari hasil temuan yang ada, kita akan coba kembangkan *keywords* apa lagi yang mungkin akan muncul. Pada kesempatan mendatang, akan dilakukan *data mining* kembali untuk berbagai situs seperti *repository* perpustakaan berbagai universitas untuk mendapatkan gambaran penelitian yang telah dilakukan di universitas-universitas tersebut.

# Metode (*Data Mining / Web Scraping*)

## Data Mining

Pengambilan data akan menggunakan algoritma *web scraping* dengan bahasa pemrograman *R* menggunakan *virtual machine* milik *Google Cloud*.

Situs yang dijadikan rujukan data adalah www.neliti.com. Data yang akan diambil antara lain: judul penelitian dan *author* (termasuk *link* rujukan).

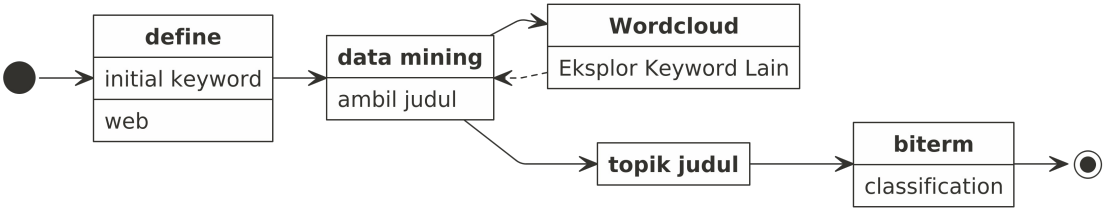
**Catatan penting:** hasil pencarian yang didapatkan murni berdasarkan output yang didapatkan dari situs neliti. Tidak ada jaminan bahwa semua penelitian tersebut selalu berkaitan penuh secara konten dengan *keywords* yang digunakan.

# Metode (*Text Analysis*)

Selanjutnya akan dilakukan beberapa *text analysis* seperti:

1. *Word cloud*: untuk menemukan *keywords* lain yang mungkin berkaitan dengan *keywords* utama.
2. *Biterm Topic Modelling*: untuk menentukan dan mengelompokkan judul artikel, penelitian, atau berita ke dalam topik-topik tertentu.

# Alur Kerja


define
initial keyword
web
data mining
ambil judul
Wordcloud
Eksplor Keyword Lain
topik judul
biterm
classification

Section 2

# HASIL *DATA MINING*

# Hasil *Data Mining*

Pada situs www.neliti.com, didapatkan ada 5.949 buah *unique* penelitian hasil pencarian *keywords*. Tentunya bisa jadi satu judul penelitian keluar dari hasil pencarian lebih dari satu *keywords*. Berikut adalah grafik dari 14 *keywords* (dan kombinasi *keywords*) teratas berdasarkan banyaknya penelitian:

Section 3

# *TEXT ANALYSIS*: *Keywords* Lain

# Mencari *Keywords* Lain

Untuk mencari *keywords* lainnya, saya akan kumpulkan semua judul penelitian hasil pencarian lalu akan dihitung kata apa saja yang paling sering muncul.

Perlu diperhatikan bahwa kata sambung, kata depan, dan *stopwords* akan dihapus dari analisa ini.

# Mencari *Keywords* Lain (lanjutan)

Kata dan frekuensi kemunculannya disajikan dalam bentuk *wordcloud* berikut ini:

# Mencari *Keywords* Lain (lanjutan)

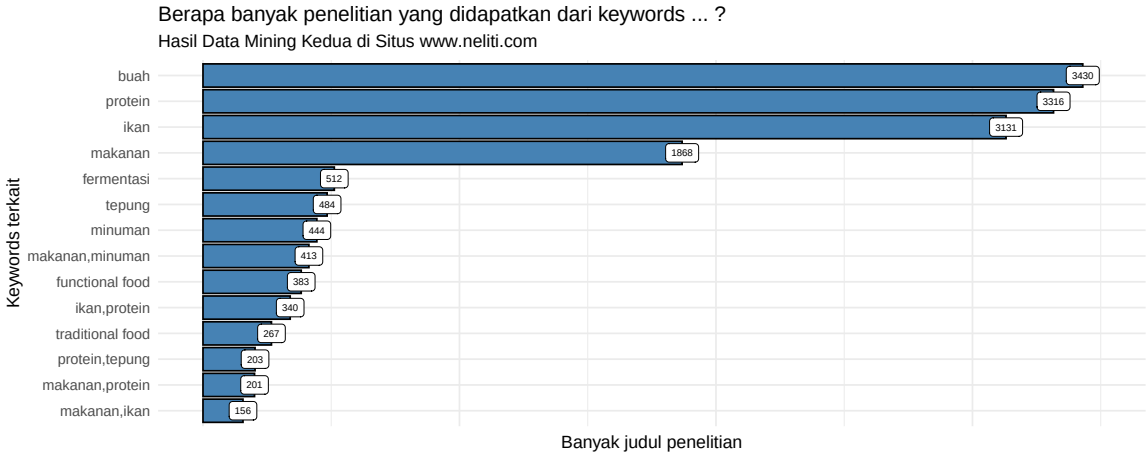
Dari *wordcloud* di atas, saya akan memilih empat buah *keywords* baru untuk dicari kembali, yakni:

1. Ikan,
2. Buah,
3. Tepung, dan
4. Protein.

Pencarian kembali dilakukan di situs `www.neliti.com` dan menghasilkan 11.298 *unique* judul penelitian baru.

# Hasil *Data Mining* Kedua

Berikut adalah grafik dari 14 *keywords* (dan kombinasi *keywords*) teratas berdasarkan banyaknya penelitian:



Section 4

# *TEXT ANALYSIS*: *Keywords* Keseluruhan

# Wordcloud Keywords Keseluruhan

Dari keseluruhan judul penelitian yang dihimpun, berikut adalah kata dan frekuensi kemunculannya yang disajikan dalam bentuk *wordcloud* berikut:

# *Bigrams Keywords* Keseluruhan

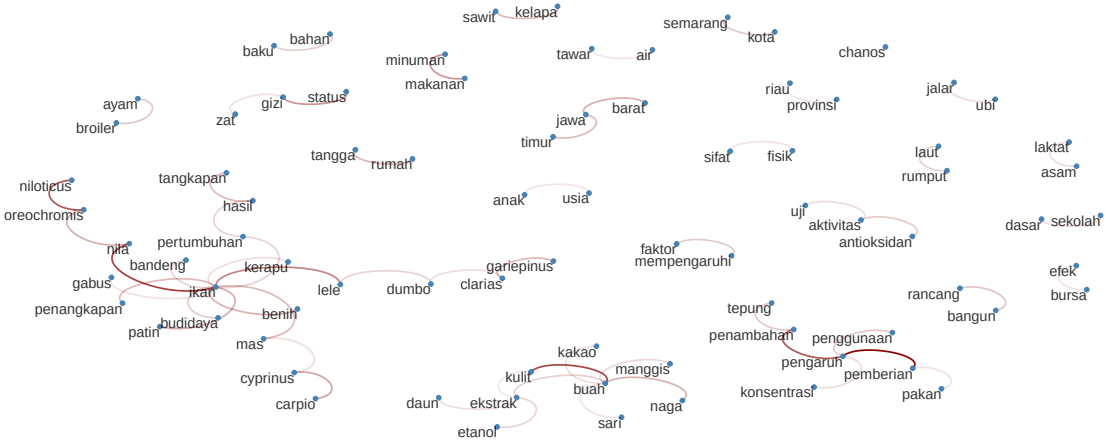
**Definisi**

*Bigrams* adalah kumpulan pasangan kata yang selalu muncul secara bersamaan.

Dari semua judul penelitian yang ada, saya akan buat analisa *bigrams* untuk melihat *keywords* lain apa saja yang mungkin muncul. Selain itu, kita bisa memperkirakan topik-topik apa saja yang ada.

Berikut adalah *bigrams* yang muncul dengan frekuensi minimal 70 kali.

# Bigrams Keywords Keseluruhan



# Hipotesis Sementara

*Keyword* `ikan` memiliki frekuensi terbesar pada *wordcloud* dan memiliki banyak *bigrams*.

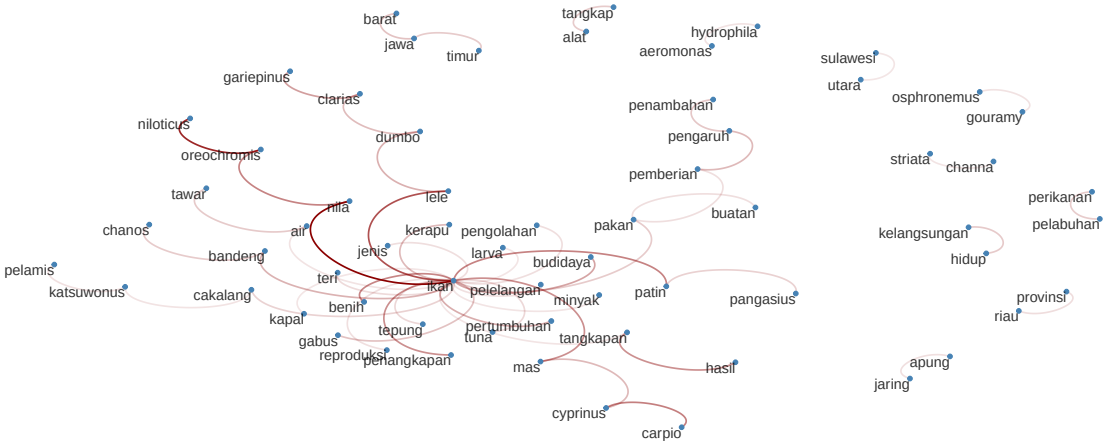
Oleh karena itu, kita akan analisa terpisah *keyword* `ikan` dari *keywords* lainnya.

Section 5

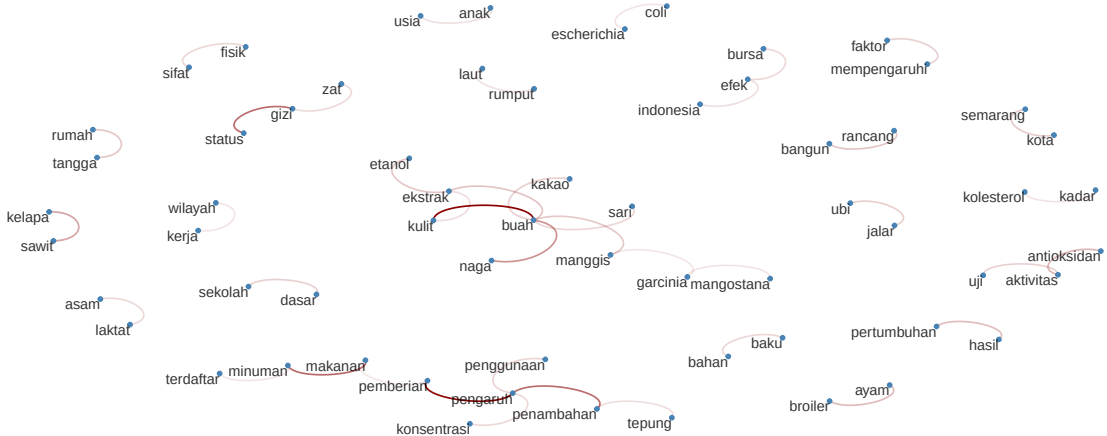
# ANALISA *KEYWORD*: IKAN

# *Wordcloud* dari *Keyword* ikan

# *Bigrams* dari *Keyword* ikan



# *Bigrams* Keseluruhan Tanpa *Keyword* ikan



Section 6

# TEXT ANALYSIS: Topics Modelling

# Topic Modelling

*Topic modelling* adalah proses melakukan pengelompokkan dari kumpulan teks. Saya akan melakukan pengelompokkan dari semua judul penelitian yang ada.

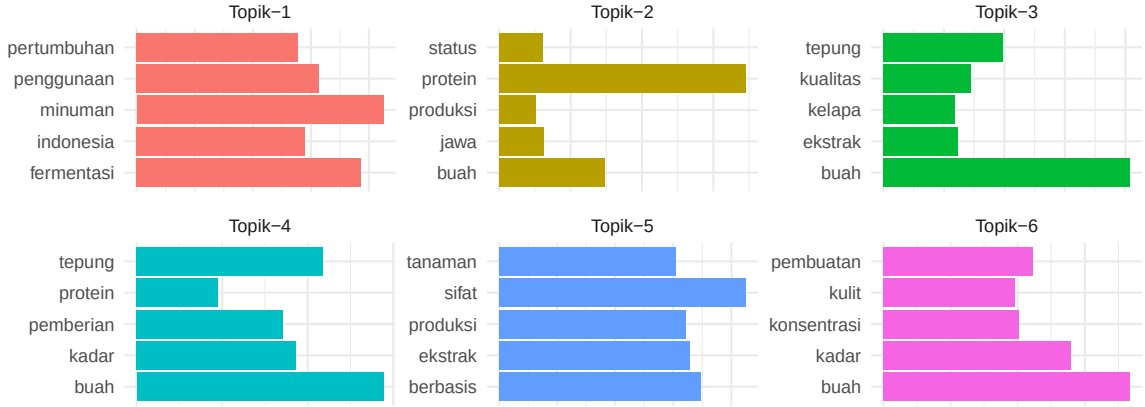
Metode *topic modelling* yang akan digunakan adalah *Latent Dirichlet Allocation* (LDA).

Saya akan lakukan beberapa analisa dengan berbagai kombinasi *keywords*.

# *Topic Modelling* Tanpa Keyword `ikan`

Kata−kata kunci dari masing−masing topik

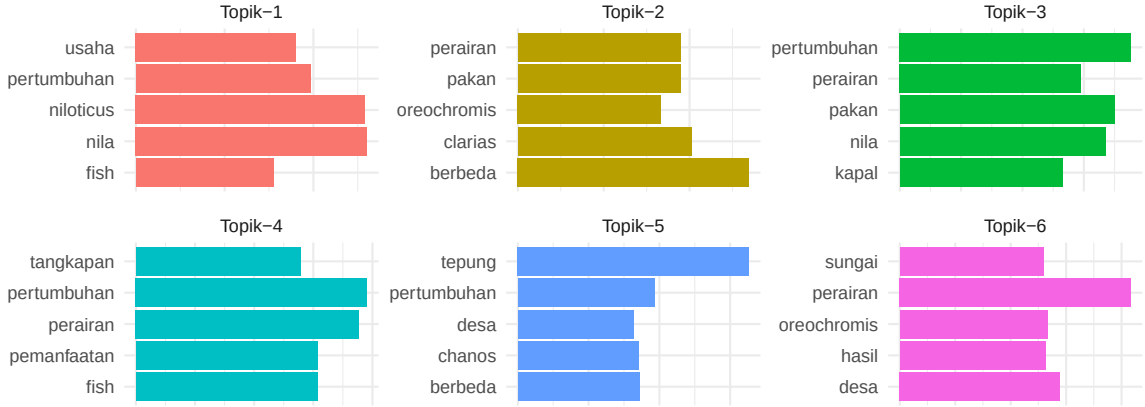
Semua judul penelitian (kecuali keyword ikan)



# *Topic Modelling* Khusus Keyword `ikan`

Kata−kata kunci dari masing−masing topik

Semua judul penelitian (khusus keyword ikan)



# *Topic Modelling* Tanpa Keyword ikan,buah,tepung, dan protein

Kata−kata kunci dari masing−masing topik

Semua judul penelitian (tanpa keyword tertentu)