Anggota Kelompok

- 23031554055 - ELIZABETH HANOV
- 23031554109 - NAFILA HANUM AL HASANIY
- 23031554124- MELINDA VENTA LYDIA SIBURIAN

Import Library

```
# library
import pandas as pd
import numpy as np
```

Open file

```
df = pd.read_csv('TENAR 2024.csv')
df.head()  # Print the first few rows of the DataFrame
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 162,\n  \"fields\": [\
n    {\n      \"column\": \"Timestamp\",\n      \"properties\": {\n
\"dtype\": \"object\",\n        \"num_unique_values\": 144,\n
\"samples\": [\n          \"5/26/2024 15:35:11\",\n
\"5/26/2024 15:35:57\",\n          \"5/26/2024 16:19:05\"\n        ],\
n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Score\",\n      \"properties\":
{\n        \"dtype\": \"category\",\n        \"num_unique_values\":
7,\n        \"samples\": [\n          \"90 / 100\",\n          \"80 /
100\",\n          \"20 / 100\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Nama lengkap (sertifikat)\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 162,\n
\"samples\": [\n          \"Yosefan Hebert Nicols Tarigan\",\n
\"Nike Silvia \",\n          \"Riva Dian Ardiansyah\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Asal lembaga\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 59,\n        \"samples\": [\n
\"Universitas Negeri Surabaya\",\n          \"Institut Teknologi
Sumatera\",\n          \"Universitas Muhammadiyah Prof. DR. HAMKA\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Bukti mengikuti acara TENAR 2024
: \\u201cUnveiling The Secrets To Top Perform In Data Science\\
u201d\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 162,\n        \"samples\": [\n
\"https://drive.google.com/open?
id=1JNFjbD6dnOfg5W76wRTdf_N4Esa9ueDc\",\n
\"https://drive.google.com/open?
id=1pdnk0daouk9HMIx1_8_cjlacNjOnKiV8\",\n
\"https://drive.google.com/open?
id=1LYhat4kqGuIHdHZMZVVdMQV7uiZpLWuR\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Apa yang dimaksud dengan
digitalisasi?\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3,\n        \"samples\":
[\n          \"A. Proses konversi data digital ke analog\",\n
\"B. Proses pengalihan informasi dari bentuk analog ke digital\",\n
\"C. Proses pencetakan data digital\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Sektor mana yang tidak disebutkan
mengalami perubahan signifikan akibat tren digitalisasi?\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"A.
Social Media & Chat apps\",\n          \"B. Fintech\",\n          \"C.
Pertanian\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\": \"Apa
tujuan utama dari data cleaning?\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 3,\n
\"samples\": [\n          \"B. Mendeteksi dan memperbaiki kesalahan
pada data\",\n          \"C. Menyimpan data dalam bentuk mentah\",\n
\"A. Mengumpulkan data\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Proses apa yang dilakukan sebelum data dipersiapkan
untuk analisis, seperti dalam text classification?\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"B. Data
visualization\",\n          \"D. Quality contro\",\n          \"A.
Pre-processing\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Data Science dapat membantu dalam...\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 3,\n
\"samples\": [\n          \"B. Menerjemahkan objektif bisnis menjadi
tugas-tugas analisis data spesifik\",\n          \"D. Menyimpan data
dalam database\",\n          \"C. Menghapus data yang tidak berguna\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Salah satu keterampilan penting yang harus dimiliki seorang Data
Scientist adalah...\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 3,\n        \"samples\":
[\n          \"B. Kemampuan coding\",\n          \"C. Menjahit\",\n
\"D. Analisa ukuran tanah\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Prospek Karir & Langkah Menjadi Data Scientist\\nSalah
satu alasan mengapa profesi Data Scientist sangat diminati
adalah...\",\n      \"properties\": {\n        \"dtype\":
```

```
\"category\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"C. Pekerjaan yang membosankan\",\n          \"A. Gaji
yang rendah\",\n          \"B. Peluang kenaikan karir yang cepat\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"Langkah awal untuk menjadi
seorang Data Scientist adalah...\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 3,\n
\"samples\": [\n          \"B. Membangun portofolio dan memamerkannya
di LinkedIn, Github, Kaggle, dll.\",\n          \"C. Menghindari
belajar pemrograman\",\n          \"D. Hanya bergantung pada
sertifikat tanpa pengalaman praktis\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Tips Meningkatkan Karir &
Menghadapi Tantangan Karir\\nApa yang sebaiknya dilakukan jika gagal
tes kerja?\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 2,\n        \"samples\":
[\n          \"C. Menghapus semua portofolio\",\n          \"B.
Bertanya alasan kegagalan dan menjadikannya evaluasi\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"Apa yang tidak termasuk dalam tips
untuk meningkatkan karir di bidang Data Science?\",\n
\"properties\": {\n        \"dtype\": \"category\",\n
\"num_unique_values\": 4,\n        \"samples\": [\n          \"D.
Menulis profil dan keterampilan di berbagai portal pekerjaan\",\n
\"B. Membangun hubungan baik dengan senior dan profesional\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Beikan kritik anda mengenai webinar kami\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 90,\
n        \"samples\": [\n          \"_\",\n          \"Sangat
menarik\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Berikan pesan dan kesan anda setelah mengikuti webinar kami, terima
kasih!\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 133,\n        \"samples\": [\n
\"Seru , menarik dan mudah dipahami\",\n          \"Terima kasih,
menambah wawasan tentang data science, semoga dapat menyelenggarakan
acara serupa dengan tema yang lain\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```python
# Keep only specified columns
columns_to_keep = ['Nama lengkap (sertifikat)', 'Asal lembaga',
'Beikan kritik anda mengenai webinar kami']
df = df[columns_to_keep]

# Display the updated DataFrame
df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 162,\n  \"fields\": [\
n    {\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 162,\n        \"samples\": [\n
\"Yosefan Hebert Nicols Tarigan\",\n          \"Nike Silvia \",\n
\"Riva Dian Ardiansyah\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Asal lembaga\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 59,\n
\"samples\": [\n          \"Universitas Negeri Surabaya\",\n
\"Institut Teknologi Sumatera\",\n          \"Universitas Muhammadiyah
Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Beikan kritik anda mengenai webinar kami\",\n      \"properties\":
{\n        \"dtype\": \"string\",\n        \"num_unique_values\": 90,\
n        \"samples\": [\n          \"_\",\n          \"Sangat
menarik\",\n          \"Tolong diberi rundown agar kita semua tau apa
saja acara yang ada didalam webinar tersebut\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
df = df.rename(columns={
    'Beikan kritik anda mengenai webinar kami': 'kritik',
})

df.head()
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 162,\n  \"fields\": [\
n    {\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 162,\n        \"samples\": [\n
\"Yosefan Hebert Nicols Tarigan\",\n          \"Nike Silvia \",\n
\"Riva Dian Ardiansyah\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Asal lembaga\",\n      \"properties\": {\n
\"dtype\": \"category\",\n        \"num_unique_values\": 59,\n
\"samples\": [\n          \"Universitas Negeri Surabaya\",\n
\"Institut Teknologi Sumatera\",\n          \"Universitas Muhammadiyah
Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 90,\n        \"samples\": [\n          \"_\",\n
\"Sangat menarik\",\n          \"Tolong diberi rundown agar kita semua
tau apa saja acara yang ada didalam webinar tersebut\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 162 entries, 0 to 161
Data columns (total 3 columns):
 #   Column                     Non-Null Count  Dtype 
---  ------                     --------------  ----- 
 0   Nama lengkap (sertifikat)  162 non-null    object
 1   Asal lembaga               162 non-null    object
 2   kritik                     128 non-null    object
dtypes: object(3)
memory usage: 3.9+ KB
```

Cleaning

```
import pandas as pd  # Import library pandas untuk manipulasi data
import re  # Import library re untuk operasi ekspresi reguler (regex)
import string  # Import library string untuk operasi terkait string

def remmovepunc(text):
    if isinstance(text, str):  # Pastikan hanya teks (string) yang diolah
        punc = string.punctuation
        text = re.sub(r'((www\.[^\s]+)|(https?://[^\s]+)|(href[^\s]+))', ' ', text)  # Menghapus URL atau link dari teks
        text = re.sub(r'@\w+', ' ', text)# Hapus username dalam format @username
        text = re.sub(r'<br>',' ', text)# Fungsi untuk menghilangkan tanda baca, timestamp, username, dan string "<br>" serta "quot" dari teks
        text = re.sub(r'quot',' ',text)
        text = re.sub(r'\b\d{1,2}:\d{2}\b', ' ', text)# Hapus timestamp dalam format menit:detik (misalnya 00:01)
        text = text.translate(str.maketrans('  ', '  ', punc)) # Hapus tanda baca
        text = text.lower()  # Mengubah semua karakter teks menjadi huruf kecil (lower case)
        text = re.sub(r'\d+', ' ', text)  # Menghapus semua angka dari teks
        text = re.sub('\n', ' ', text)  # Mengganti karakter baris baru (\n) dengan spasi
        text = re.sub(r'\buser\b', ' ', text)  # Menghapus kata "user" yang berdiri sendiri
        text = re.sub(r'[^\w\s]', ' ', text)  # Menghapus semua tanda baca tetapi mempertahankan huruf dan spasi
        text = re.sub('[^0-9a-zA-Z]+', ' ', text)  # Menghapus karakter non-alfanumerik kecuali spasi
        text = re.sub(r'\s+', ' ', text).strip()  # Menghapus spasi berlebih dan memotong spasi di awal/akhir teks
        return text  # Mengembalikan hasil yang sudah diolah;
```

```
# Terapkan fungsi ini ke setiap elemen di DataFrame yang bertipe
string
# Menggunakan DataFrame yang sudah ada sebelumnya (df), yang berisi
kolom komentar
df['kritik_bersih'] = df['kritik'].apply(remmovepunc)

# Tampilkan beberapa baris untuk memastikan
df.loc[100:120]
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 21,\n  \"fields\": [\n
{\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 21,\n        \"samples\": [\n
\"Muhammad Wifaqul Azmi\",\n          \"NURUL HIDAYAH \",\n
\"NUGROHO FAJAR RACHMANTO\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"Asal lembaga\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 15,\n
\"samples\": [\n          \"universitas negeri surabaya\",\n
\"UPN VETERAN JAWA TIMUR\",\n          \"UPN \\\"Veteran\\\" Jawa
Timur\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik\",\n      \"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 14,\n        \"samples\": [\n
\"acaranya sudah bagus\",\n          \"Terlalu banyak intro diawal
kak\",\n          \"tidak ada\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_bersih\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 13,\n        \"samples\": [\n          \"semoga
lebih baik lagi\",\n          \"seru tidak membosankan\",\n
\"tidak ada\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```

```
# Drop rows with any NaN values in the DataFrame
df = df.dropna()

df
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 128,\n  \"fields\": [\
n    {\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 128,\n        \"samples\": [\n          \"Iffah
Nurul Rohmah\",\n          \"Erlina Firdaus \",\n          \"Auryn
Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    }\n  ]\n}","type":"dataframe","variable_name":"df"}
```

Prepocessing

```
# !pip install pandas nltk
```

```
Requirement already satisfied: pandas in
/usr/local/lib/python3.10/dist-packages (2.2.2)
Requirement already satisfied: nltk in /usr/local/lib/python3.10/dist-
packages (3.8.1)
Requirement already satisfied: numpy>=1.22.4 in
/usr/local/lib/python3.10/dist-packages (from pandas) (1.26.4)
Requirement already satisfied: python-dateutil>=2.8.2 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: tzdata>=2022.7 in
/usr/local/lib/python3.10/dist-packages (from pandas) (2024.2)
Requirement already satisfied: click in
/usr/local/lib/python3.10/dist-packages (from nltk) (8.1.7)
Requirement already satisfied: joblib in
/usr/local/lib/python3.10/dist-packages (from nltk) (1.4.2)
Requirement already satisfied: regex>=2021.8.3 in
/usr/local/lib/python3.10/dist-packages (from nltk) (2024.9.11)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-
packages (from nltk) (4.66.6)
Requirement already satisfied: six>=1.5 in
/usr/local/lib/python3.10/dist-packages (from python-dateutil>=2.8.2-
>pandas) (1.16.0)
```

```
# !pip install Sastrawi
```

```
Collecting Sastrawi
  Downloading Sastrawi-1.0.1-py2.py3-none-any.whl.metadata (909 bytes)
Downloading Sastrawi-1.0.1-py2.py3-none-any.whl (209 kB)
━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 209.7/209.7 kB 4.6 MB/s eta
0:00:00
```

```python
from Sastrawi.StopWordRemover.StopWordRemoverFactory import
StopWordRemoverFactory

list_stopwords = []
stopwords_sastrawi = StopWordRemoverFactory().get_stop_words()
list_stopwords.extend(stopwords_sastrawi)
list_stopwords = list(dict.fromkeys(list_stopwords))
list_stopwords.sort()
len(list_stopwords)
```

```
123
```

```python
sw = pd.read_csv("/content/stopwordbahasa.csv")
sw
```

```
{"summary":"{\n  \"name\": \"sw\",\n  \"rows\": 761,\n  \"fields\": [\
n    {\n      \"column\": \"stopwords\",\n      \"properties\": {\n
\"dtype\": \"string\",\n        \"num_unique_values\": 759,\n
\"samples\": [\n          \"memisalkan\",\n          \"saatnya\",\n
\"bermacam\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"sw"}
```

```python
list_stopwords += list(sw['stopwords'])
list_stopwords = list(dict.fromkeys(list_stopwords))
list_stopwords.sort()
len(list_stopwords)
```

```
777
```

```python
import pandas as pd
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

# Download NLTK data (hanya perlu dilakukan sekali)
nltk.download('punkt')
# Pastikan tokenizer dan stopwords data sudah terunduh
nltk.download('stopwords')

# Inisialisasi stemmer dari Sastrawi
factory = StemmerFactory()
```

```
stemmer = factory.create_stemmer()
stemmer_indo = StemmerFactory().create_stemmer()
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

```
import re
import nltk
import pandas as pd
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from Sastrawi.Stemmer.StemmerFactory import StemmerFactory

# Membuat stemmer bahasa Indonesia
factory = StemmerFactory()
stemmer_indo = factory.create_stemmer()

# Unduh stopwords bahasa Indonesia dari nltk
nltk.download('stopwords')
nltk_stop_words = set(stopwords.words('indonesian'))
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

```
# Stopwords tambahan yang ingin ditambahkan
additional_stopwords = {
    "ka"
}

# Gabungkan nltk_stop_words dan additional_stopwords
stop_words = nltk_stop_words.union(additional_stopwords)

# Memuat stopwords tambahan dari file CSV (jika ada)
try:
    additional_stop_words_csv = pd.read_csv('stopwordbahasa.csv',
header=None, sep=",")[0].tolist()
    stop_words = stop_words.union(set(additional_stop_words_csv))
except FileNotFoundError:
    print("Warning: stopwordbahasa.csv not found. Using only defined
stopwords.")

# Fungsi untuk membersihkan teks
def clean_text(text):
    if isinstance(text, list):
        # Gabungkan token menjadi string jika input berupa list
        text = ' '.join(text)
    # Menghapus karakter non-alfabet
    text = re.sub(r'[^a-zA-Z\s]', ' ', text)
    # Mengubah teks menjadi huruf kecil
    text = text.lower()
    # Tokenisasi teks
    words = word_tokenize(text)
    # Menghapus stopwords dan melakukan stemming
    msg_fix = [stemmer_indo.stem(word) for word in words if word not
in stop_words]
    return ' '.join(msg_fix)

# Terapkan fungsi pembersihan ke kolom teks
df['kritik_stem'] = df['kritik_bersih'].apply(clean_text)

df
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 128,\n  \"fields\": [\
n    {\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 128,\n        \"samples\": [\n          \"Iffah
Nurul Rohmah\",\n          \"Erlina Firdaus \",\n          \"Auryn
Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```

```
# KAMUS MANUAL (contoh)
manual_dict = {
    "cumn": "hanya",

    # Tambahkan kata-kata slang dan padanan bakunya di sini
}

# Baca file CSV dan rename kolomnya (jika ada)
try:
    alay_dict = pd.read_csv('slang_normal.csv', header=None, sep=",")
    alay_dict = alay_dict.rename(columns={0: 'original', 1:
'replacement'})
    # Gabungkan kamus manual dengan kamus dari file CSV
    alay_dict_map = dict(zip(alay_dict['original'],
alay_dict['replacement']))
    alay_dict_map.update(manual_dict)
except FileNotFoundError:
    print("Warning: slang_normal.csv not found. Using only manual
dictionary.")
    alay_dict_map = manual_dict

# Fungsi untuk normalisasi kata alay
def normalize_alay(text):
    return ' '.join([alay_dict_map.get(word, word) for word in
text.split(' ')])

# Fungsi untuk menghapus stopwords
def remove_stopwords(text):
    text = ' '.join([word for word in text.split(' ') if word not in
stop_words])
    return text

# Fungsi untuk stemming
def stemming(text):
    return stemmer.stem(text)

# Fungsi untuk preprocessing teks
def preprocess(text):
    text = normalize_alay(text)
    text = remove_stopwords(text)  # Pindahkan remove_stopwords
sebelum stemming
    text = stemming(text)
    return text

# Contoh penggunaan
df['normalisasi'] = df['kritik_stem'].apply(preprocess)

# Menampilkan hasil preprocessing
df[['kritik_stem', 'normalisasi']].head()
```

```
{"summary":"{\n  \"name\": \"df[['kritik_stem', 'normalisasi']]\",\n
\"rows\": 5,\n  \"fields\": [\n    {\n      \"column\":
\"kritik_stem\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"game lagu bahasa ingriss saran kenal lagu indonesia\",\
n          \"tingkat\",\n          \"ganggu teknis suara putus putus
kali dengar suara\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 4,\n        \"samples\":
[\n          \"game lagu bahasa ingriss saran kenal lagu indonesia\",\
n          \"tingkat\",\n          \"ganggu teknis suara putus putus
kali dengar suara\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\n}","type":"dataframe"}
```

```python
normalisasi = df['normalisasi']
df.loc()
df
```

```
{"summary":"{\n  \"name\": \"df\",\n  \"rows\": 128,\n  \"fields\": [\
n    {\n      \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 128,\n        \"samples\": [\n          \"Iffah
Nurul Rohmah\",\n          \"Erlina Firdaus \",\n          \"Auryn
Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
```

aja si next event laksana offline\",\n   \"ganggu teknis suara
putus putus kali dengar suara\",\n   \"acara bagus\"\
n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"normalisasi\",\n   \"properties\": {\n   \"dtype\":
\"string\",\n   \"num_unique_values\": 67,\n   \"samples\":
[\n   \"sesi suara materi dengar\",\n   \"isi materi\",\
n   \"materi sebentar undang narasumber semangat panitia\"\n
],\n   \"semantic_type\": \"\",\n   \"description\": \"\"\n
}\n   }\n]\n}","type":"dataframe","variable_name":"df"}

```
df.info()
```

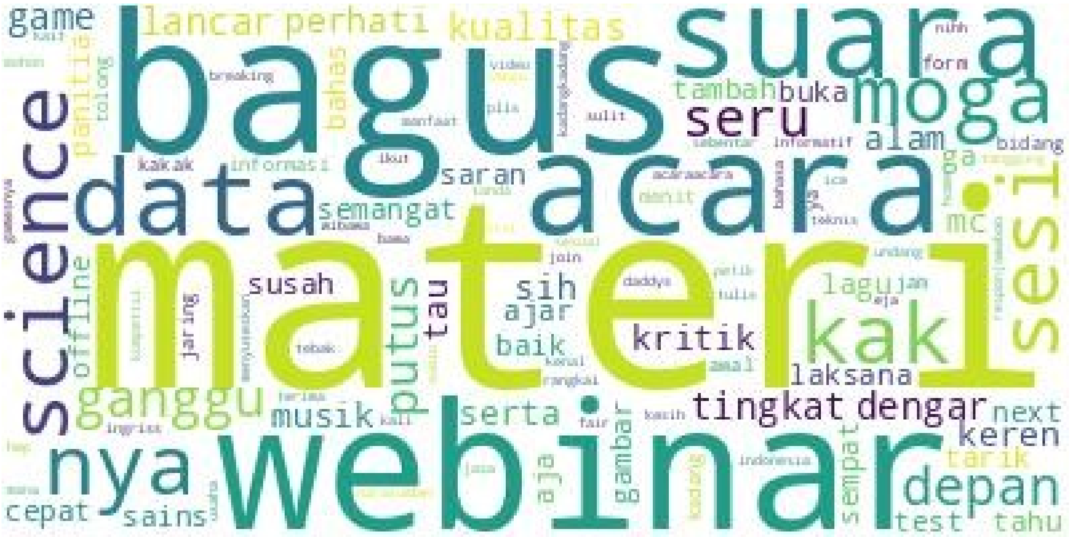
```
<class 'pandas.core.frame.DataFrame'>
Index: 128 entries, 0 to 161
Data columns (total 6 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Nama lengkap (sertifikat)  128 non-null    object
 1   Asal lembaga               128 non-null    object
 2   kritik                     128 non-null    object
 3   kritik_bersih              128 non-null    object
 4   kritik_stem                128 non-null    object
 5   normalisasi                128 non-null    object
dtypes: object(6)
memory usage: 7.0+ KB
```

```
count_wd = []
for word in df['kritik_stem']:
    count = len(word.split())
    count_wd.append(count)
df['count_words'] = count_wd
df
```

{"summary":"{\n \"name\": \"df\",\n \"rows\": 128,\n \"fields\": [\
n   {\n   \"column\": \"Nama lengkap (sertifikat)\",\n
\"properties\": {\n   \"dtype\": \"string\",\n
\"num_unique_values\": 128,\n   \"samples\": [\n   \"Iffah
Nurul Rohmah\",\n   \"Erlina Firdaus \",\n   \"Auryn
Shafarina\"\n   ],\n   \"semantic_type\": \"\",\n
\"description\": \"\"\n   }\n   },\n   {\n   \"column\":
\"Asal lembaga\",\n   \"properties\": {\n   \"dtype\":
\"category\",\n   \"num_unique_values\": 50,\n
\"samples\": [\n   \"Universitas swadaya gunung jati
cirebon\",\n   \"Telkom University\",\n   \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n   ],\n   \"semantic_type\":
\"\",\n   \"description\": \"\"\n   }\n   },\n   {\n
\"column\": \"kritik\",\n   \"properties\": {\n   \"dtype\":
\"string\",\n   \"num_unique_values\": 90,\n   \"samples\":
[\n   \"_\",\n   \"Sangat menarik\",\n   \"Tolong

```
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 67,\n        \"samples\":
[\n          \"sesi suara materi dengar\",\n          \"isi materi\",\
n          \"materi sebentar undang narasumber semangat panitia\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"count_words\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
5,\n        \"min\": 0,\n        \"max\": 43,\n
\"num_unique_values\": 16,\n        \"samples\": [\n          8,\n
0,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df"}
```

```python
import matplotlib.pyplot as plt
import seaborn as sns

# Import the wordcloud library
from wordcloud import WordCloud

# Join the different processed titles together.
long_string = ','.join(list(df['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=10, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```



Dari wordcloud diatas dihasilkan kata "webinar, acara, materi, bagus, suaranya putus, data science" sehingga bisa disimpulkan bahwa isi dari kritik sebagian besar adalah berkesimpulan berisi : Acara TEBAR 2024 merupakan acara yang bagus dan keren di bidang data science dengan materi yang dimilikinya tetapi terdapat beberapa kendala dalam suara seperti putus - putus. Sesi games yang dibuat untuk mengisi acara dinilai susah. Dilain sisi, beberapa peserta ada yang menginginkan acara TENAR 2024 dilakukan secara offline.

Feature Engineering

```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

```
True
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
corpus = df['kritik_stem'].to_list()
vectorizer = TfidfVectorizer()
X = vectorizer.fit_transform(corpus)
```

```
data_FE = pd.DataFrame(X.toarray(),
columns=vectorizer.get_feature_names_out())
data_FE
```

{"type":"dataframe","variable_name":"data_FE"}

```
data_FE.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 128 entries, 0 to 127
Columns: 222 entries, acara to zoomnya
```

```
dtypes: float64(222)
memory usage: 222.1 KB
```

Clustering

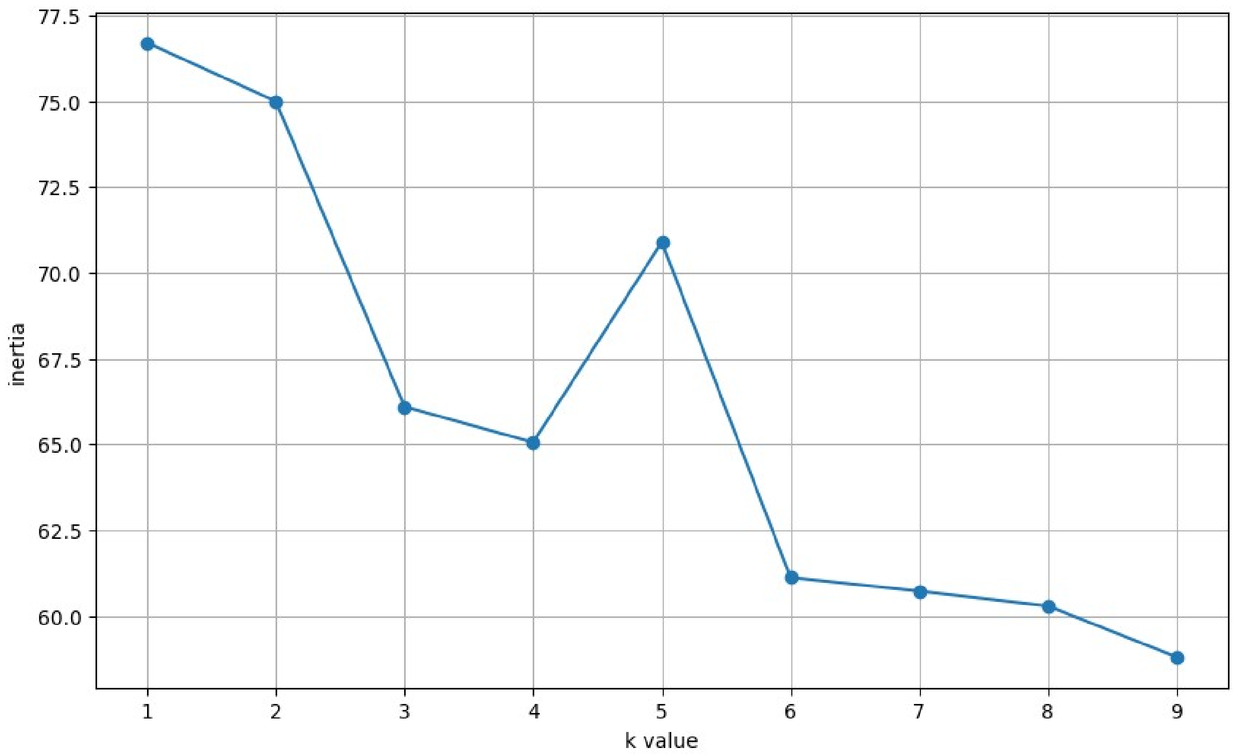
K- Means

```
from sklearn.cluster import KMeans

k = range(1,10)
inertia = []

for i in k:
    km_model = KMeans(n_clusters=i) # jumlah cluster akan di looping
    km_model.fit(data_FE)
    inertia.append(km_model.inertia_)

print(list(k))
print(inertia)
```

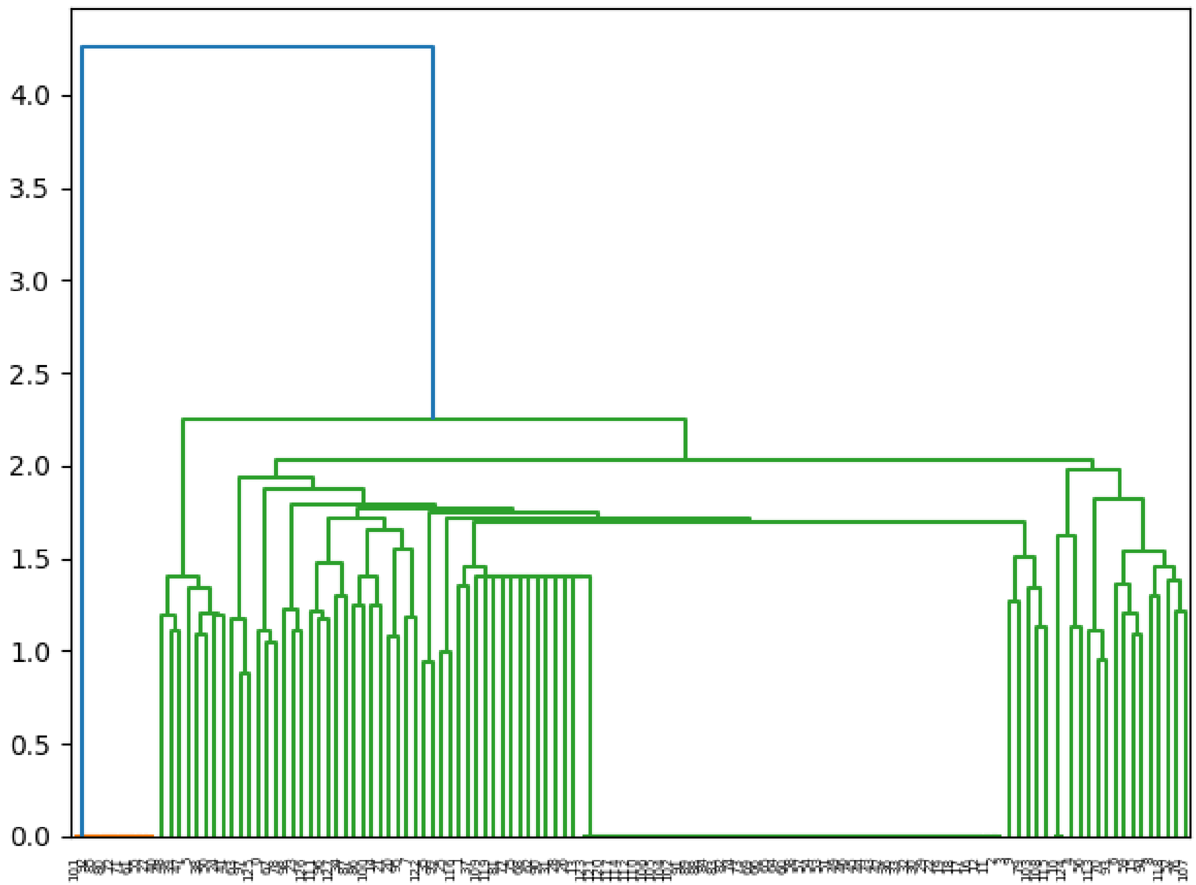
```
[1, 2, 3, 4, 5, 6, 7, 8, 9]
[76.69839439260556, 74.99225577936839, 66.09703786116657,
65.05891764325834, 70.89340752614017, 61.11923984210873,
60.726747085966345, 60.29221277980849, 58.80874049454015]
```

K Value = 3

```
plt.figure(figsize=(10,6))
plt.plot(k, inertia, marker='o')
plt.xlabel('k value')
plt.ylabel('inertia')
plt.grid()
plt.show()
```



```
kn3 = KMeans(n_clusters=3, random_state=1).fit(data_FE)
kn3

KMeans(n_clusters=3, random_state=1)

kn3.labels_

array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0,
0,
       0, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0,
       0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0,
0,
       0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0],
dtype=int32)

from scipy.cluster.hierarchy import dendrogram, linkage

linkage_data = linkage(data_FE, method = 'ward', metric = 'euclidean')
dendrogram(linkage_data)
plt.tight_layout()
plt.show()
```



```python
from sklearn.cluster import AgglomerativeClustering
hierarchical_cluster = AgglomerativeClustering(n_clusters = 3, linkage
= 'ward')
labels = hierarchical_cluster.fit(data_FE)

pred_agc = pd.Series(hierarchical_cluster.labels_)
```

Silhouette Score

```python
from sklearn import metrics
labels = kn3.labels_
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2919369611229207
```

```python
from sklearn import metrics
labels = pred_agc
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2997299075929363
```

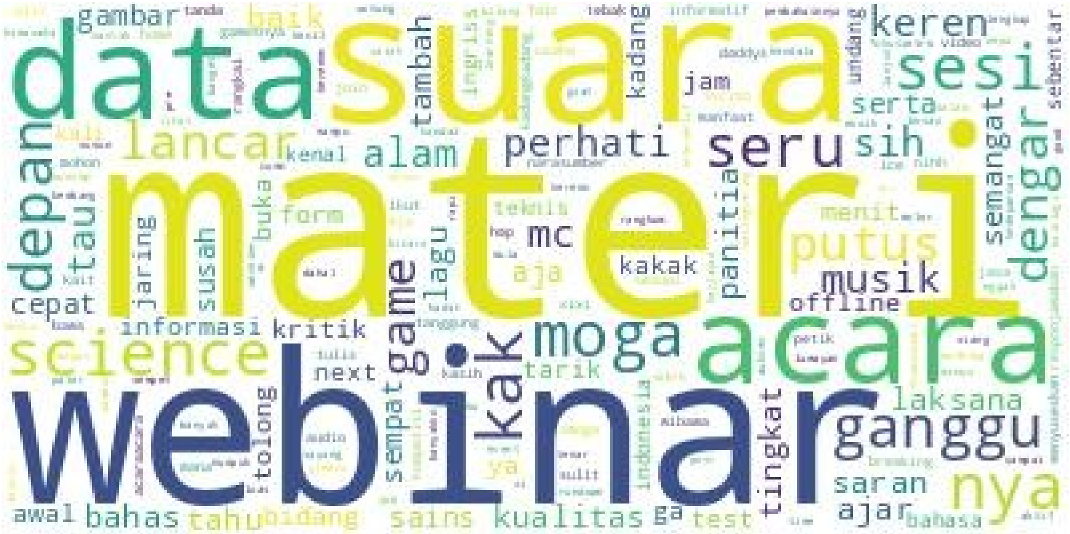
Word Cloud Hasil Setiap Cluster

```
df_Kmeans = df.copy(deep=True)
df_Kmeans['cluster'] = kn3.labels_
df_Kmeans
```

```
{"summary":"{\n  \"name\": \"df_Kmeans\",\n  \"rows\": 128,\n
\"fields\": [\n    {\n      \"column\": \"Nama lengkap
(sertifikat)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 128,\n        \"samples\":
[\n          \"Iffah Nurul Rohmah\",\n          \"Erlina Firdaus \",\n
\"Auryn Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 67,\n        \"samples\":
[\n          \"sesi suara materi dengar\",\n          \"isi materi\",\
n          \"materi sebentar undang narasumber semangat panitia\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"count_words\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
5,\n        \"min\": 0,\n        \"max\": 43,\n
\"num_unique_values\": 16,\n        \"samples\": [\n          8,\n
0,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"cluster\",\n      \"properties\": {\n        \"dtype\": \"int32\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          0,\n
2,\n          1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df_Kmeans"}
```

```
c=0

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 0
```

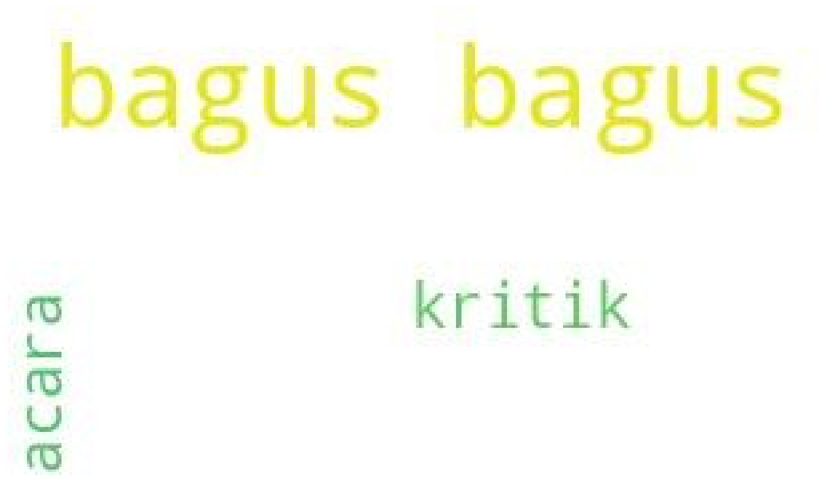


Kesimpulan dari WordCloud cluster 0 adalah acara TENAR 2024 adalah acara yang seru tetapi suaranya terdapat beberapa gangguan (red : noise) tetapi materinya seru.

```
c=1

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
```

```
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

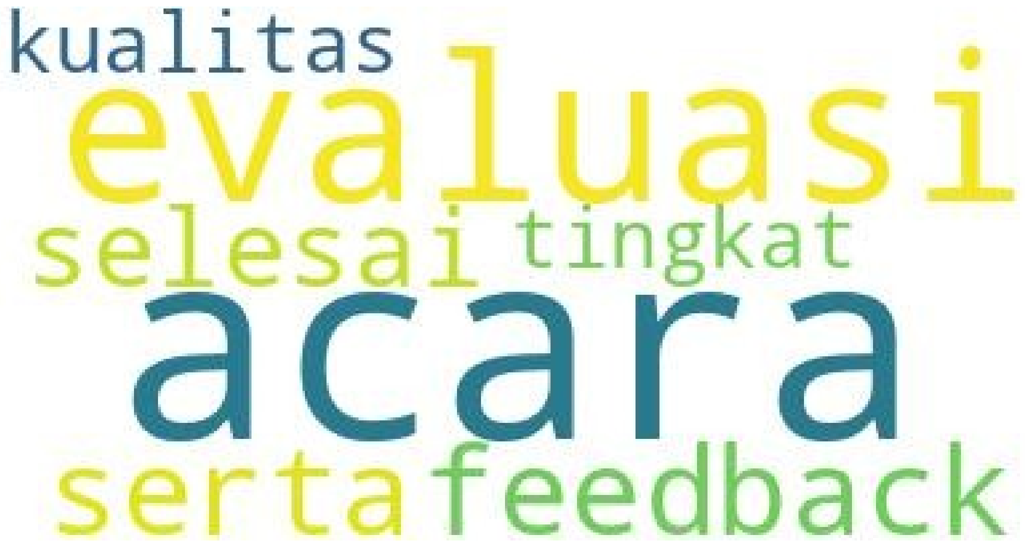
```
Kmeans for cluster 1
```



Kesimpulan dari WordCloud cluster 1 adalah acara TENAR 2024 adalah acara yang bagus.

```
c=2

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

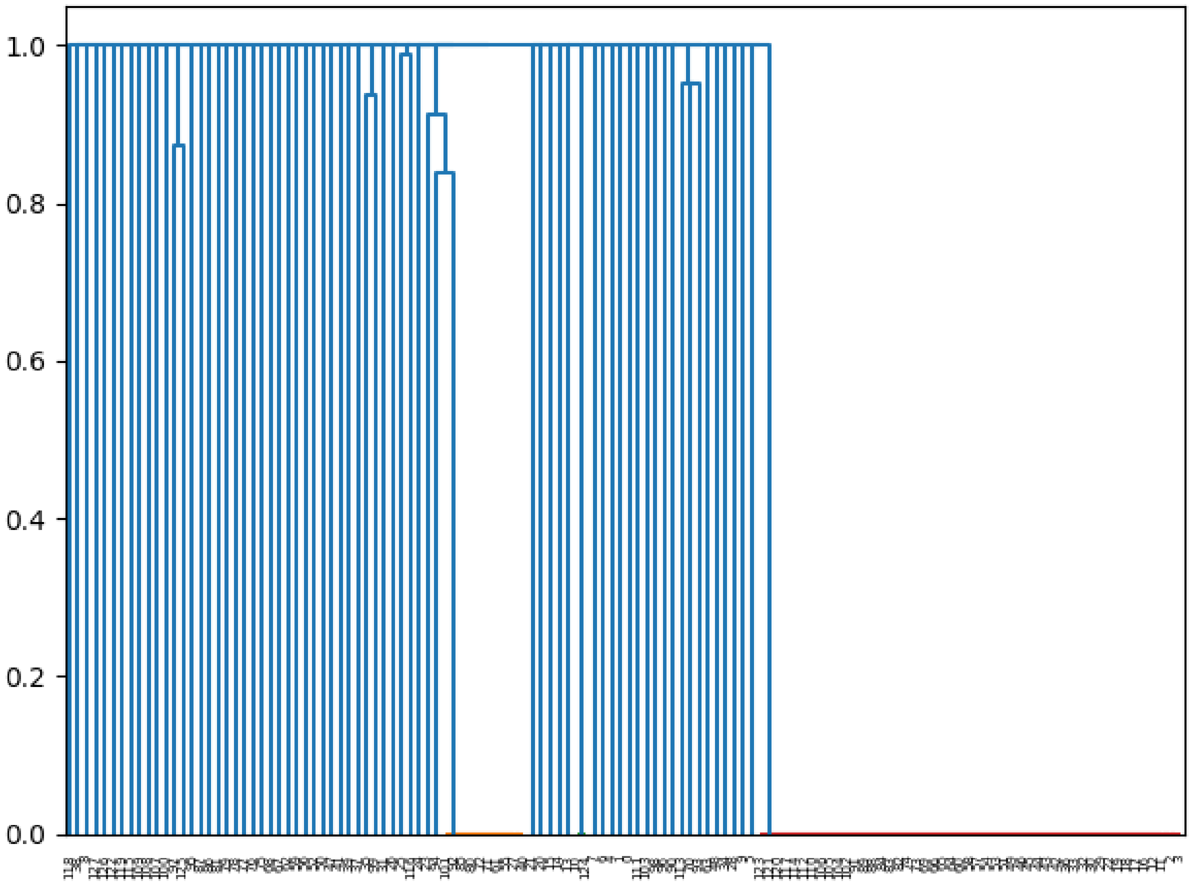
```
Kmeans for cluster 2
```



Kesimpulan dari WordCloud cluster 0 adalah acara TENAR 2024 adalah acara yang berkualitas tetapi peserta juga memiliki feedback dan evaluasi untuk acara ini.

Hierarchical Method Single

```
from scipy.cluster.hierarchy import dendrogram, linkage

linkage_data = linkage(data_FE, method = 'single', metric =
'euclidean')
dendrogram(linkage_data)
plt.tight_layout()
plt.show()
```



```
from sklearn.cluster import AgglomerativeClustering
hierarchical_cluster = AgglomerativeClustering(n_clusters = 3, linkage
= 'single')
labels = hierarchical_cluster.fit(data_FE)

pred_agc = pd.Series(hierarchical_cluster.labels_)
```

Silhouette Score

```
from sklearn import metrics
labels = kn3.labels_
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2919369611229207
```

```
from sklearn import metrics
labels = pred_agc
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2075974245704929
```

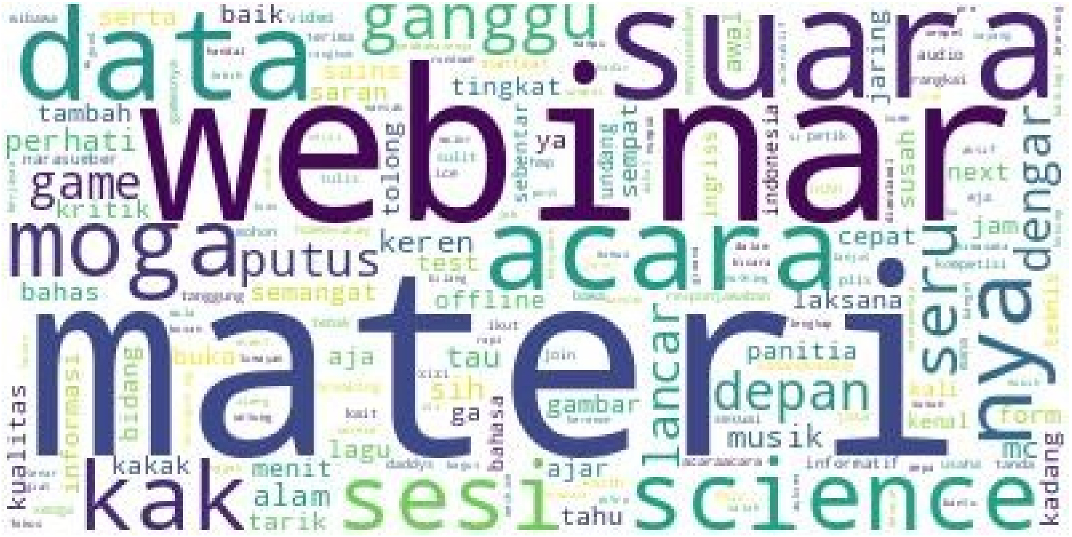
Word Cloud Hasil Setiap Cluster

```
df_Kmeans = df.copy(deep=True)
df_Kmeans['cluster'] = kn3.labels_
df_Kmeans
```

```
{"summary":"{\n  \"name\": \"df_Kmeans\",\n  \"rows\": 128,\n
\"fields\": [\n    {\n      \"column\": \"Nama lengkap
(sertifikat)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 128,\n        \"samples\":
[\n          \"Iffah Nurul Rohmah\",\n          \"Erlina Firdaus \",\n
\"Auryn Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 67,\n        \"samples\":
[\n          \"sesi suara materi dengar\",\n          \"isi materi\",\
n          \"materi sebentar undang narasumber semangat panitia\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"count_words\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
5,\n        \"min\": 0,\n        \"max\": 43,\n
\"num_unique_values\": 16,\n        \"samples\": [\n          8,\n
0,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"cluster\",\n      \"properties\": {\n        \"dtype\": \"int32\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          0,\n
2,\n          1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df_Kmeans"}
```

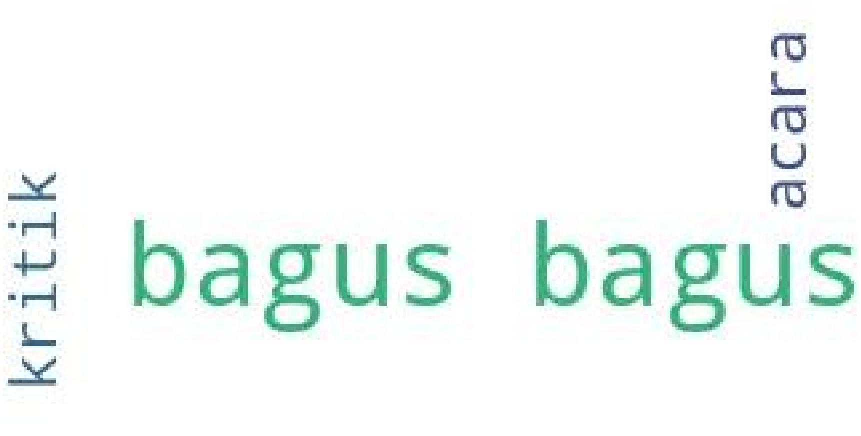
```
c=0

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 0
```



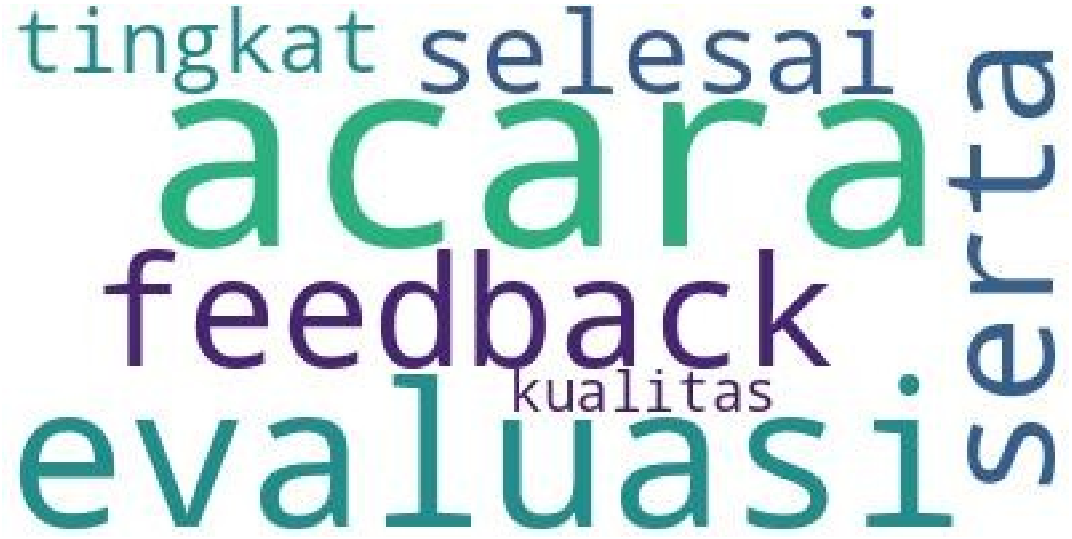
```
c=1

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 1
```



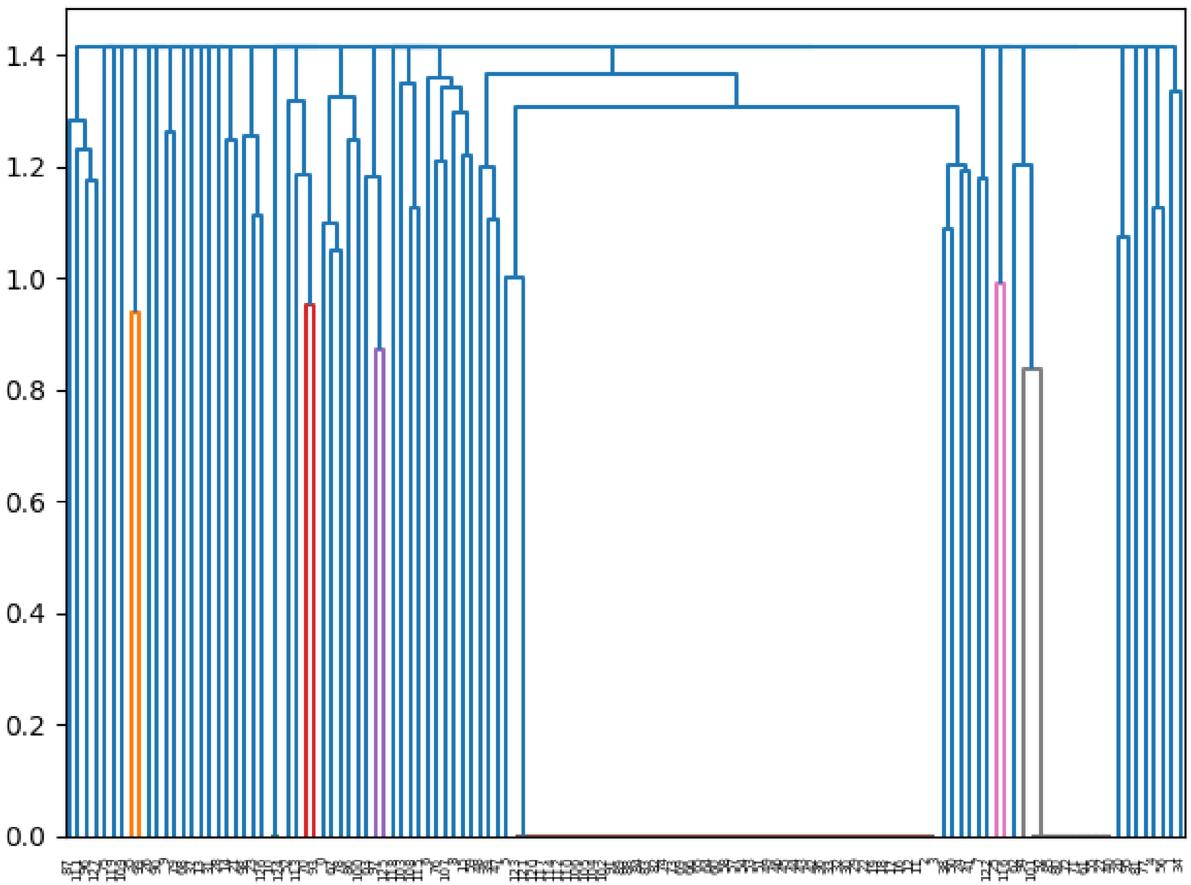
```
c=2

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 2
```



Hierarchical Method Complete

```
from scipy.cluster.hierarchy import dendrogram, linkage

linkage_data = linkage(data_FE, method = 'complete', metric =
'euclidean')
dendrogram(linkage_data)
plt.tight_layout()
plt.show()
```



```
from sklearn.cluster import AgglomerativeClustering
hierarchical_cluster = AgglomerativeClustering(n_clusters = 3, linkage
= 'complete')
labels = hierarchical_cluster.fit(data_FE)

pred_agc = pd.Series(hierarchical_cluster.labels_)
```

Silhouette Score

```
from sklearn import metrics
labels = kn3.labels_
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2919369611229207
```

```
from sklearn import metrics
labels = pred_agc
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.22716121041597888
```

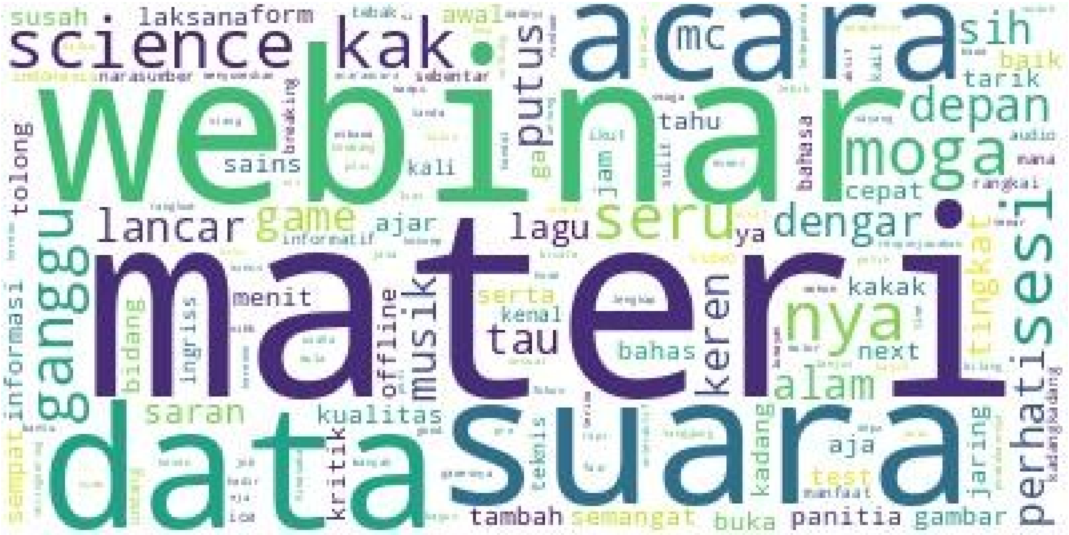
Word Cloud Hasil Setiap Cluster

```
df_Kmeans = df.copy(deep=True)
df_Kmeans['cluster'] = kn3.labels_
df_Kmeans
```

```
{"summary":"{\n  \"name\": \"df_Kmeans\",\n  \"rows\": 128,\n
\"fields\": [\n    {\n      \"column\": \"Nama lengkap
(sertifikat)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 128,\n        \"samples\":
[\n          \"Iffah Nurul Rohmah\",\n          \"Erlina Firdaus \",\n
\"Auryn Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 67,\n        \"samples\":
[\n          \"sesi suara materi dengar\",\n          \"isi materi\",\
n          \"materi sebentar undang narasumber semangat panitia\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"count_words\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
5,\n        \"min\": 0,\n        \"max\": 43,\n
\"num_unique_values\": 16,\n        \"samples\": [\n          8,\n
0,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"cluster\",\n      \"properties\": {\n        \"dtype\": \"int32\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          0,\n
2,\n          1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df_Kmeans"}
```

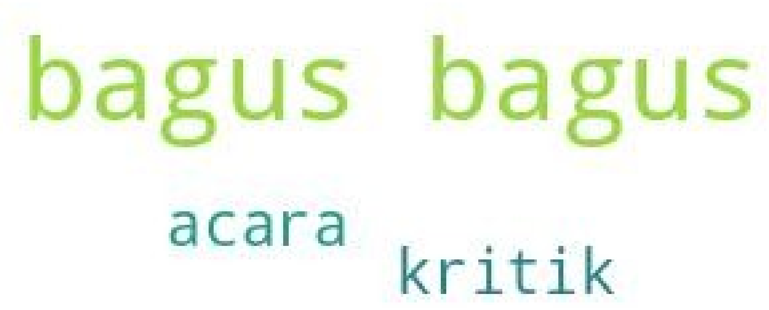
```python
c=0

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 0
```



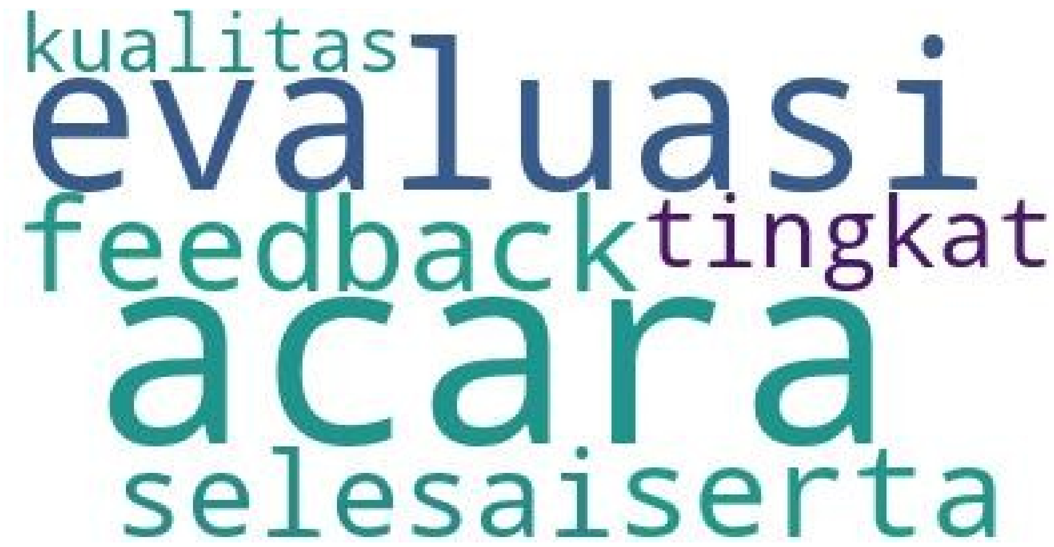
```python
c=1

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 1
```



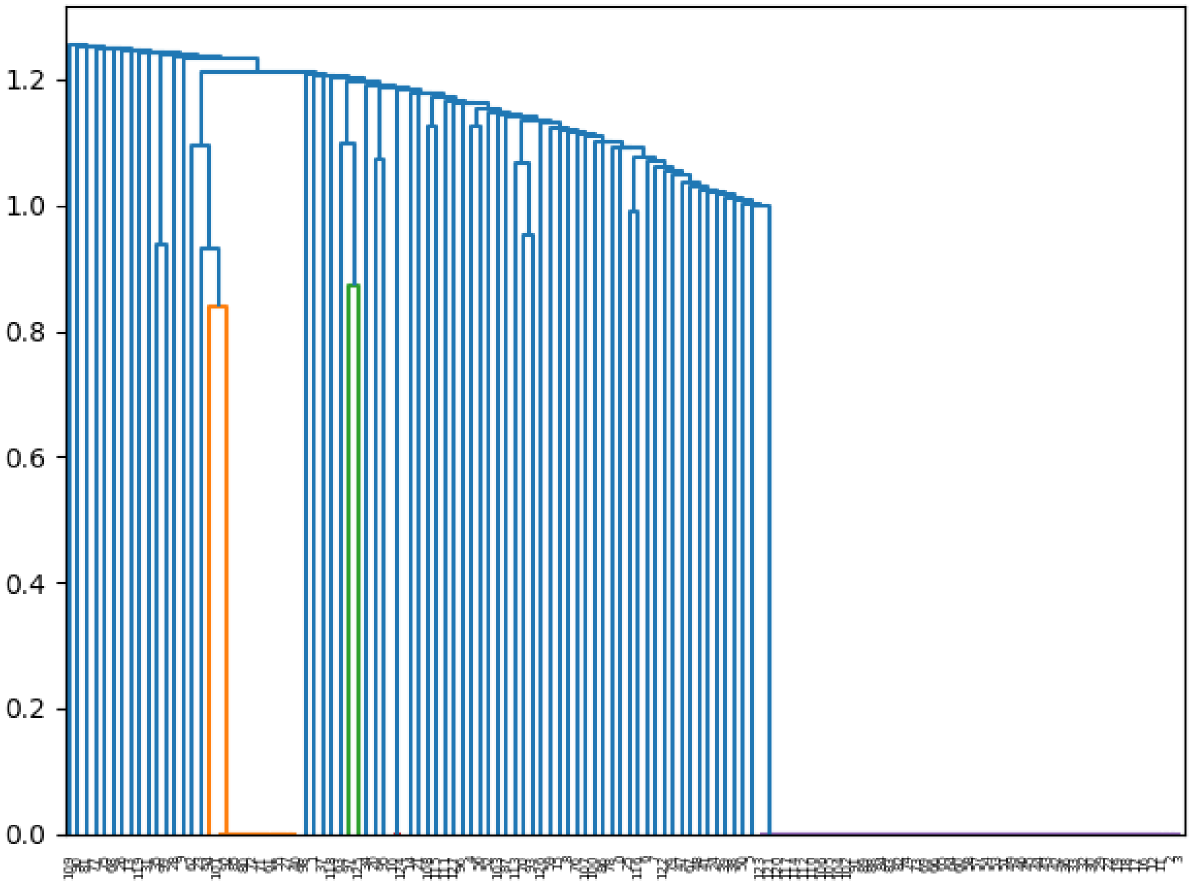
```
c=2

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 2
```



Hierarchical Method Average

```
from scipy.cluster.hierarchy import dendrogram, linkage

linkage_data = linkage(data_FE, method = 'average', metric =
'euclidean')
dendrogram(linkage_data)
plt.tight_layout()
plt.show()
```



```
from sklearn.cluster import AgglomerativeClustering
hierarchical_cluster = AgglomerativeClustering(n_clusters = 3, linkage
= 'average')
labels = hierarchical_cluster.fit(data_FE)

pred_agc = pd.Series(hierarchical_cluster.labels_)
```

Silhouette Score

```
from sklearn import metrics
labels = kn3.labels_
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

```
0.2919369611229207
```

```
from sklearn import metrics
labels = pred_agc
metrics.silhouette_score(data_FE, labels, metric='euclidean')
```

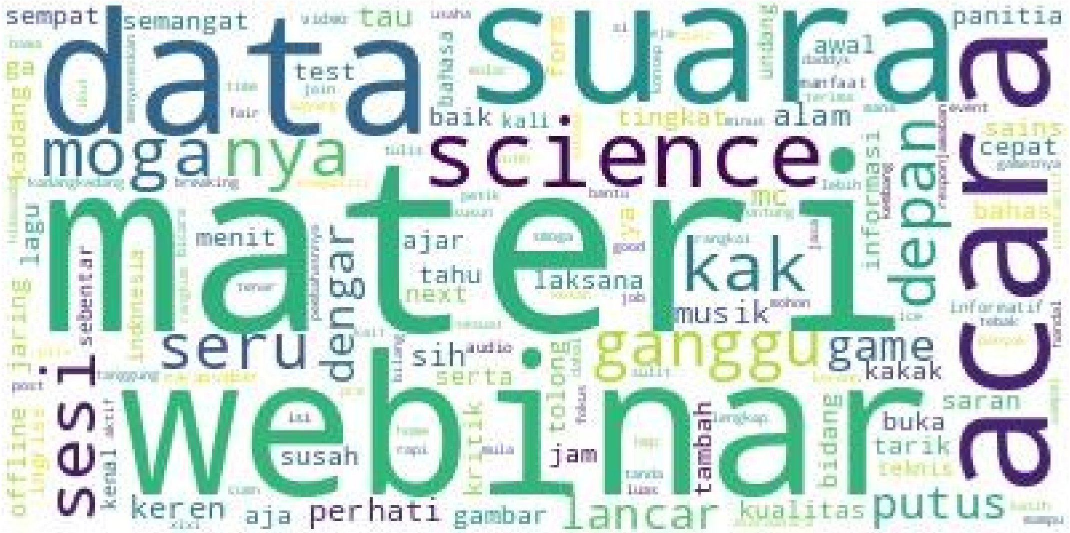
```
0.22611550880877285
```

Word Cloud Hasil Setiap Cluster

```
df_Kmeans = df.copy(deep=True)
df_Kmeans['cluster'] = kn3.labels_
df_Kmeans
```

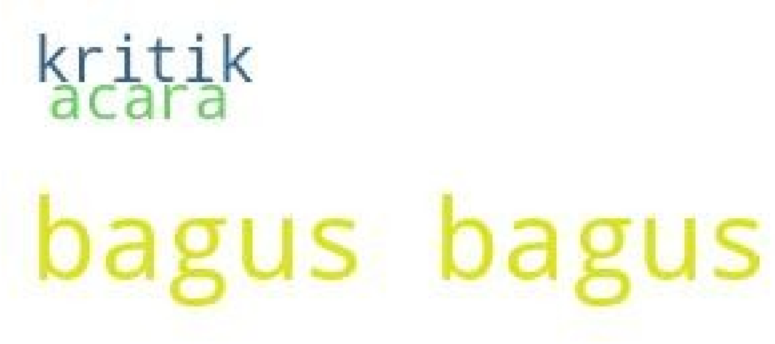
```
{"summary":"{\n  \"name\": \"df_Kmeans\",\n  \"rows\": 128,\n
\"fields\": [\n    {\n      \"column\": \"Nama lengkap
(sertifikat)\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 128,\n        \"samples\":
[\n          \"Iffah Nurul Rohmah\",\n          \"Erlina Firdaus \",\n
\"Auryn Shafarina\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"Asal lembaga\",\n      \"properties\": {\n        \"dtype\":
\"category\",\n        \"num_unique_values\": 50,\n
\"samples\": [\n          \"Universitas swadaya gunung jati
cirebon\",\n          \"Telkom University\",\n          \"Universitas
Muhammadiyah Prof. DR. HAMKA\"\n        ],\n        \"semantic_type\":
\"\",\n        \"description\": \"\"\n      }\n    },\n    {\n
\"column\": \"kritik\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 90,\n        \"samples\":
[\n          \"_\",\n          \"Sangat menarik\",\n          \"Tolong
diberi rundown agar kita semua tau apa saja acara yang ada didalam
webinar tersebut\"\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"kritik_bersih\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 82,\n        \"samples\":
[\n          \"webinar data science sangat memberikan informasi dan
materi yang sangat baik untuk pemula yang ingin mengembangkan
kemampuan di bidang science data dan pemateri yang sangat jelas dalam
menjelaskan materi\",\n          \"ada beberapa gangguan teknis
seperti suara yang terputus putus dan beberapa kali terdengar kecil
suaranya\",\n          \"sudah kerenn\"\n        ],\n
\"semantic_type\": \"\",\n        \"description\": \"\"\n      }\
n    },\n    {\n      \"column\": \"kritik_stem\",\n
\"properties\": {\n        \"dtype\": \"string\",\n
\"num_unique_values\": 70,\n        \"samples\": [\n          \"saran
aja si next event laksana offline\",\n          \"ganggu teknis suara
putus putus kali dengar suara\",\n          \"acara bagus\"\
n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"normalisasi\",\n      \"properties\": {\n        \"dtype\":
\"string\",\n        \"num_unique_values\": 67,\n        \"samples\":
[\n          \"sesi suara materi dengar\",\n          \"isi materi\",\
n          \"materi sebentar undang narasumber semangat panitia\"\n
],\n        \"semantic_type\": \"\",\n        \"description\": \"\"\n
}\n    },\n    {\n      \"column\": \"count_words\",\n
\"properties\": {\n        \"dtype\": \"number\",\n        \"std\":
5,\n        \"min\": 0,\n        \"max\": 43,\n
\"num_unique_values\": 16,\n        \"samples\": [\n          8,\n
0,\n          11\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    },\n    {\n      \"column\":
\"cluster\",\n      \"properties\": {\n        \"dtype\": \"int32\",\n
\"num_unique_values\": 3,\n        \"samples\": [\n          0,\n
2,\n          1\n        ],\n        \"semantic_type\": \"\",\n
\"description\": \"\"\n      }\n    }\n  ]\
n}","type":"dataframe","variable_name":"df_Kmeans"}
```

```
c=0

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()

Kmeans for cluster 0
```



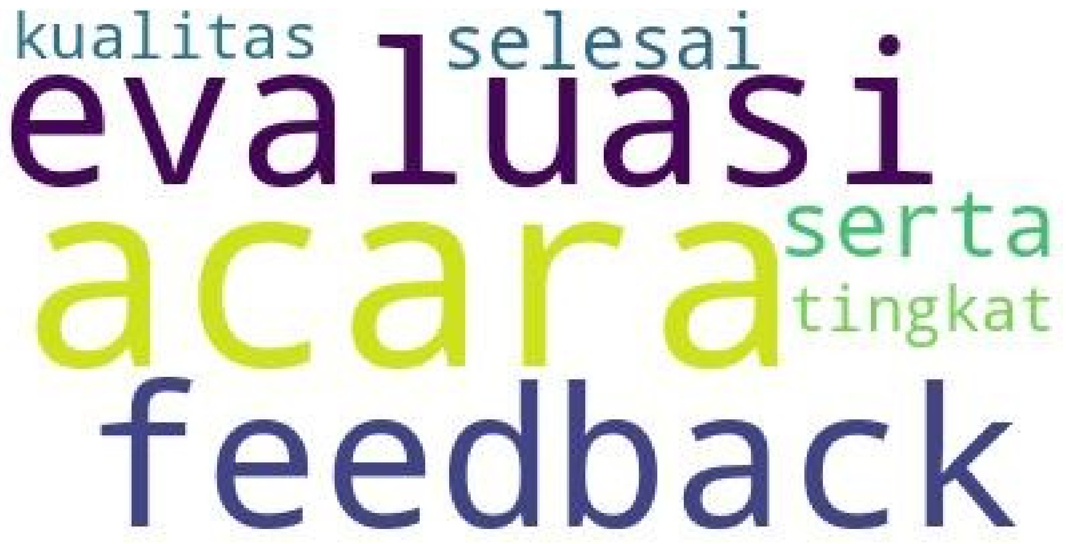
```
c=1

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 1
```



```
c=2

print('Kmeans for cluster '+str(c))
df_c = df_Kmeans[df_Kmeans['cluster'] == c]
# Join the different processed titles together.
long_string = ','.join(list(df_c['kritik_stem'].values))
# Create a WordCloud object
wordcloud = WordCloud(background_color="white", max_words=5000,
contour_width=3, contour_color='steelblue')
# Generate a word cloud
wordcloud.generate(long_string)
# Visualize the word cloud
wordcloud.to_image()
```

```
Kmeans for cluster 2
```



Telah dicoba beberapa jenis cluster, yakni : Single, Complete, Average, Ward dan juga menggunakan K - Means jenis euclidean yang menunjukkan hasil sama.

Insight yang dapat diambil dari pemrosesan data kritik dari peserta acara TENAR 2024 ini menunjukkan bahwa secara umum, Webinar Tenar 2024 berjalan dengan baik, materi disampaikan dengan jelas, dan mempunyai kualitas yang diharapkan peserta . Meskipun demikian, terdapat beberapa masukan dan evaluasi, dimana sebagian besar keluhan peserta berfokus pada kualitas suara yang kurang jernih selama webinar berlangsung. Hal ini dapat menjadi perhatian untuk meningkatkan pengalaman peserta di webinar berikutnya.