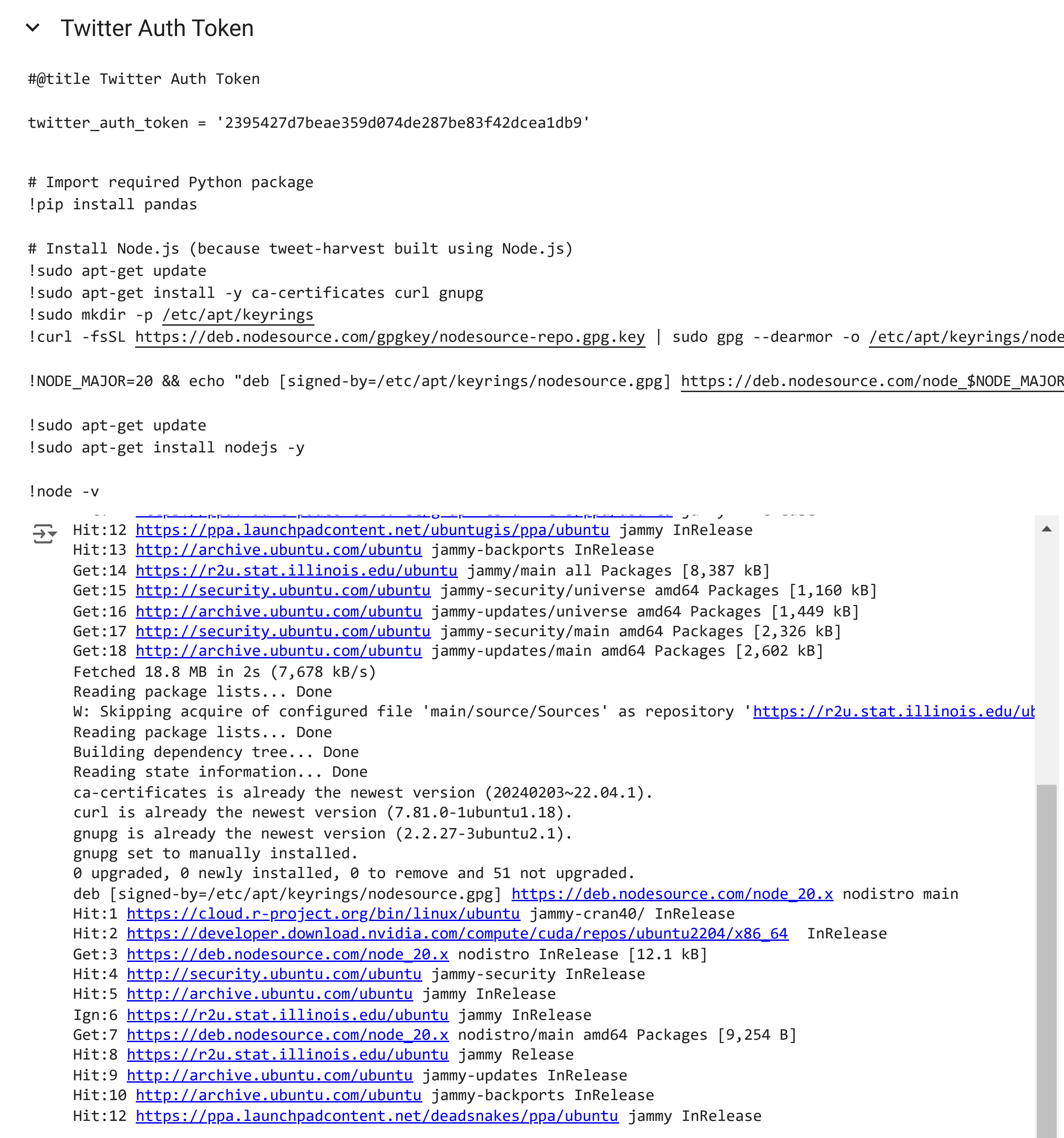


## Twitter Auth Token

```
#@title Twitter Auth Token

twitter_auth_token = '2395427d7beae359d074de287be83f42dcea1db9'


# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/node

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

```
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Hit:13 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Get:14 https://r2u.stat.illinois.edu/ubuntu jammy/main all Packages [8,387 kB]
Get:15 http://security.ubuntu.com/ubuntu jammy-security/universe amd64 Packages [1,160 kB]
Get:16 http://archive.ubuntu.com/ubuntu jammy-updates/universe amd64 Packages [1,449 kB]
Get:17 http://security.ubuntu.com/ubuntu jammy-security/main amd64 Packages [2,326 kB]
Get:18 http://archive.ubuntu.com/ubuntu jammy-updates/main amd64 Packages [2,602 kB]
Fetched 18.8 MB in 2s (7,678 kB/s)
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ub
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
ca-certificates is already the newest version (20240203~22.04.1).
curl is already the newest version (7.81.0-1ubuntu1.18).
gnupg is already the newest version (2.2.27-3ubuntu2.1).
gnupg set to manually installed.
0 upgraded, 0 newly installed, 0 to remove and 51 not upgraded.
deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_20.x nodistro main
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Get:3 https://deb.nodesource.com/node_20.x nodistro InRelease [12.1 kB]
Hit:4 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Ign:6 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Get:7 https://deb.nodesource.com/node_20.x nodistro/main amd64 Packages [9,254 B]
Hit:8 https://r2u.stat.illinois.edu/ubuntu jammy Release
Hit:9 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:10 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:12 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
```

```
    debconf: falling back to frontend: Readline
    debconf: unable to initialize frontend: Readline
    debconf: (This frontend requires a controlling tty.)
    debconf: falling back to frontend: Teletype
    dpkg-preconfigure: unable to re-open stdin:
    Selecting previously unselected package nodejs.
    (Reading database ... 123621 files and directories currently installed.)
    Preparing to unpack .../nodejs_20.18.0-1nodesource1_amd64.deb ...
    Unpacking nodejs (20.18.0-1nodesource1) ...
    Setting up nodejs (20.18.0-1nodesource1)
```

```
# Crawl Data

filename = 'wasit_bola.csv'
search_keyword = 'wasit bahrain'
limit = 200

!npx --yes tweet-harvest@latest -o "{filename}" -s "{search_keyword}" -l {limit} --token {twitter_auth_token}
```

```
(3)Created new directory: /content/tweets-data


    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 19

    -- Scrolling... (1) (2)

    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 39

    -- Scrolling... (1) (2)

    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 59

    -- Scrolling... (1) (2) (3)

    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 79

    -- Scrolling... (1) (2)

    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 99

    -- Scrolling... (1) (2) (3)

    Your tweets saved to: /content/tweets-data/wasit_bola.csv
    Total tweets saved: 119

    --Taking a break, waiting for 10 seconds...
```

```
Your tweets saved to: /content/tweets-data/wasit_bola.csv
Total tweets saved: 219
```

```
import pandas as pd

# Specify the path to your CSV file
file_path = f"/content/tweets-data/wasit_bola.csv"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path)

# Display the DataFrame
display(df)
```

| | conversation_id_str | created_at | favorite_count | full_text | id_str | |
|---|---|---|---|---|---|---|
| 0 | 1844732882407067670 | Fri Oct 11 13:32:47 +0000 2024 | 3800 | Timnas bahrain harusnya malu sih karena media ... | 1844732882407067670 | https://pbs.twin |
| 1 | 1844644175893127511 | Fri Oct 11 07:40:18 +0000 2024 | 6031 | Media luar mulai menyorot terkait kontroversi ... | 1844644175893127511 | https://pbs.twi |
| 2 | 1844457016263368939 | Thu Oct 10 19:16:35 +0000 2024 | 6253 | Di video terlihat golnya persis ketika tulisan... | 1844457016263368939 | https://pbs.tw |
| 3 | 1844542073338462384 | Fri Oct 11 00:54:34 +0000 2024 | 41134 | \| Rangkuman kecurangan wasit Ahmed Al-Kaf di ... | 1844542073338462384 | https://pbs.tw |
| 4 | 1844695288566698322 | Fri Oct 11 11:03:24 +0000 2024 | 38803 | Bahrain have a lot of money but dont have a c... | 1844695288566698322 | https://pbs.tw |
| ... | ... | ... | ... | ... | ... | |
| 214 | 1845076144279937458 | Sat Oct 12 12:16:47 +0000 2024 | 3 | #HeadlineNewsMetroTV 1. Duel Timnas Indonesia ... | 1845076144279937458 | https://pbs.tw |
| 215 | 1844339151132557458 | Thu Oct 10 11:28:14 +0000 2024 | 2805 | Pemain Bahrain udah boleh gemeter belum? https... | 1844339151132557458 | https://pbs.tw |
| 216 | 1844970334598070441 | Sat Oct 12 05:16:20 +0000 2024 | 9 | para korban judol tidak boleh ditinggalkan ata... | 1844970334598070441 | https://pbs.twin |
| 217 | 1844450410779771008 | Thu Oct 10 18:50:20 +0000 2024 | 2 | Kayaknya baru kali ini gue ngamuk di X gara2 w... | 1844450410779771008 | https://pbs.twir |
| 218 | 1844668259012141307 | Fri Oct 11 09:15:59 +0000 2024 | 9 | ..Kalaupun kau menang itu awal dari kekalahan... | 1844668259012141307 | https://pbs.twi |

219 rows × 15 columns

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
# Cek jumlah data yang didapatkan

num_tweets = len(df)
print(f"Jumlah tweet dalam dataframe adalah {num_tweets}.")
```

```
Jumlah tweet dalam dataframe adalah 219.
```

```
import pandas as pd
import re
import nltk
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
from wordcloud import WordCloud
```

```
nltk.download('stopwords')
```

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

```
file_path = "/content/tweets-data/wasit_bola.csv"

# Read the CSV file into a pandas DataFrame
df = pd.read_csv(file_path)

# Display the DataFrame to verify its content
display(df)

# Verify column names
print("Column names:", df.columns)

# Ensure the correct column name for tweets
full_text = 'full_text'  # Adjust this if the column name is different in your CSV
```

|  | conversation_id_str | created_at | favorite_count | full_text | id_str |  |
|---|---|---|---|---|---|---|
| **0** | 1844732882407067670 | Fri Oct 11 13:32:47 +0000 2024 | 3800 | Timnas bahrain harusnya malu sih karena media ... | 1844732882407067670 | https://pbs.twin |
| **1** | 1844644175893127511 | Fri Oct 11 07:40:18 +0000 2024 | 6031 | Media luar mulai menyorot terkait kontroversi ... | 1844644175893127511 | https://pbs.twi |
| **2** | 1844457016263368939 | Thu Oct 10 19:16:35 +0000 2024 | 6253 | Di video terlihat golnya persis ketika tulisan... | 1844457016263368939 | https://pbs.tw |
| **3** | 1844542073338462384 | Fri Oct 11 00:54:34 +0000 2024 | 41134 | \| Rangkuman kecurangan wasit Ahmed Al-Kaf di ... | 1844542073338462384 | https://pbs.tw |
| **4** | 1844695288566698322 | Fri Oct 11 11:03:24 +0000 2024 | 38803 | Bahrain have a lot of money but dont have a c... | 1844695288566698322 | https://pbs.tw |
| **...** | ... | ... | ... | ... | ... |  |
| **214** | 1845076144279937458 | Sat Oct 12 12:16:47 +0000 2024 | 3 | #HeadlineNewsMetroTV 1. Duel Timnas Indonesia ... | 1845076144279937458 | https://pbs.tw |
| **215** | 1844339151132557458 | Thu Oct 10 11:28:14 +0000 2024 | 2805 | Pemain Bahrain udah boleh gemeter belum? https... | 1844339151132557458 | https://pbs.tw |
| **216** | 1844970334598070441 | Sat Oct 12 05:16:20 +0000 2024 | 9 | para korban judol tidak boleh ditinggalkan ata... | 1844970334598070441 | https://pbs.twin |
| **217** | 1844450410779771008 | Thu Oct 10 18:50:20 +0000 2024 | 2 | Kayaknya baru kali ini gue ngamuk di X gara2 w... | 1844450410779771008 | https://pbs.twir |
| **218** | 1844668259012141307 | Fri Oct 11 09:15:59 +0000 2024 | 9 | ..Kalaupun kau menang itu awal dari kekalahan... | 1844668259012141307 | https://pbs.twi |

219 rows × 15 columns

```
Column names: Index(['conversation_id_str', 'created_at', 'favorite_count', 'full_text',
       'id_str', 'image_url', 'in_reply_to_screen_name', 'lang', 'location',
       'quote_count', 'reply_count', 'retweet_count', 'tweet_url',
       'user_id_str', 'username'],
      dtype='object')
```

Next steps: Generate code with df | View recommended plots | New interactive sheet

```
# Function to preprocess tweets
def preprocess_tweet(tweet):
    tweet = re.sub(r'http\S+', '', tweet)  # Remove URLs
    tweet = re.sub(r'@\w+', '', tweet)  # Remove mentions
    tweet = re.sub(r'#\w+', '', tweet)  # Remove hashtags
    tweet = re.sub(r'\d+', '', tweet)  # Remove numbers
    tweet = tweet.lower()  # Convert to lowercase
    tweet = re.sub(r'[^\w\s]', '', tweet)  # Remove punctuation
    tweet = tweet.strip()  # Remove leading/trailing whitespace
    return tweet
```

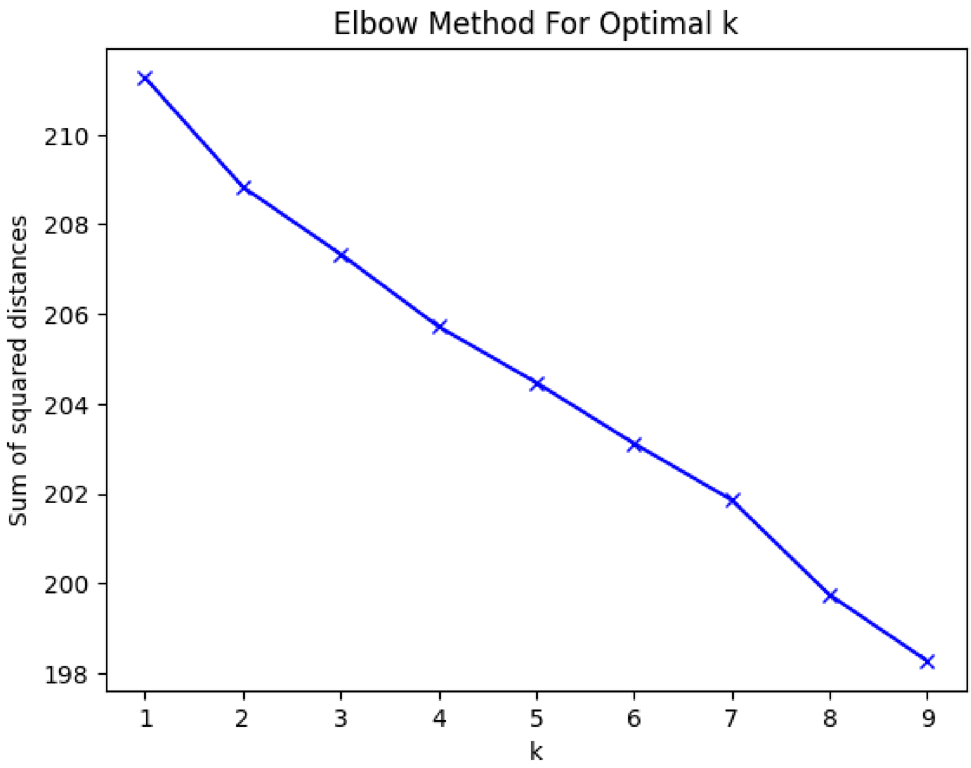
```
df['cleaned_tweet'] = df[full_text].apply(preprocess_tweet)
```

```
# Define stop words
stop_words = list(stopwords.words("indonesian", "english"))
```

```
vectorizer = TfidfVectorizer(stop_words=stop_words)
X = vectorizer.fit_transform(df['cleaned_tweet'])
```

```
/usr/local/lib/python3.10/dist-packages/sklearn/feature_extraction/text.py:406: UserWarning: Your stop_words ma
  warnings.warn(
```

```
sum_of_squared_distances = []
K = range(1, 10)
for k in K:
    km = KMeans(n_clusters=k, random_state=42)
    km = km.fit(X)
    sum_of_squared_distances.append(km.inertia_)

plt.plot(K, sum_of_squared_distances, 'bx-')
plt.xlabel('k')
plt.ylabel('Sum of squared distances')
plt.title('Elbow Method For Optimal k')
plt.show()
```



```
true_k = 3  # Example value, adjust based on the elbow plot
kmeans = KMeans(n_clusters=true_k, random_state=42)
kmeans.fit(X)
```

```
▾ KMeans
KMeans(n_clusters=3, random_state=42)
```

```
df['cluster'] = kmeans.labels_
```

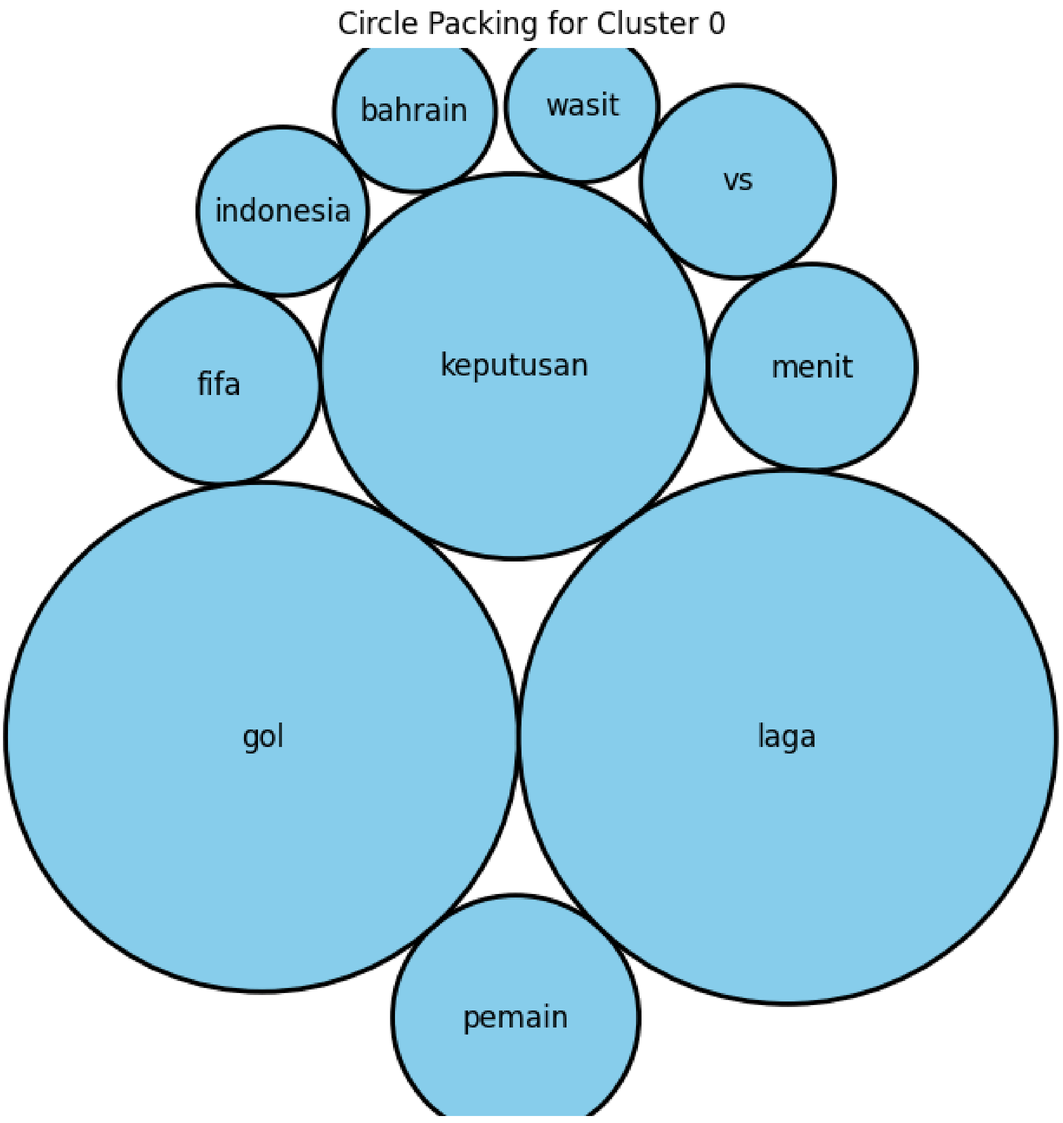
```
print(df[[full_text, 'cluster']])
```

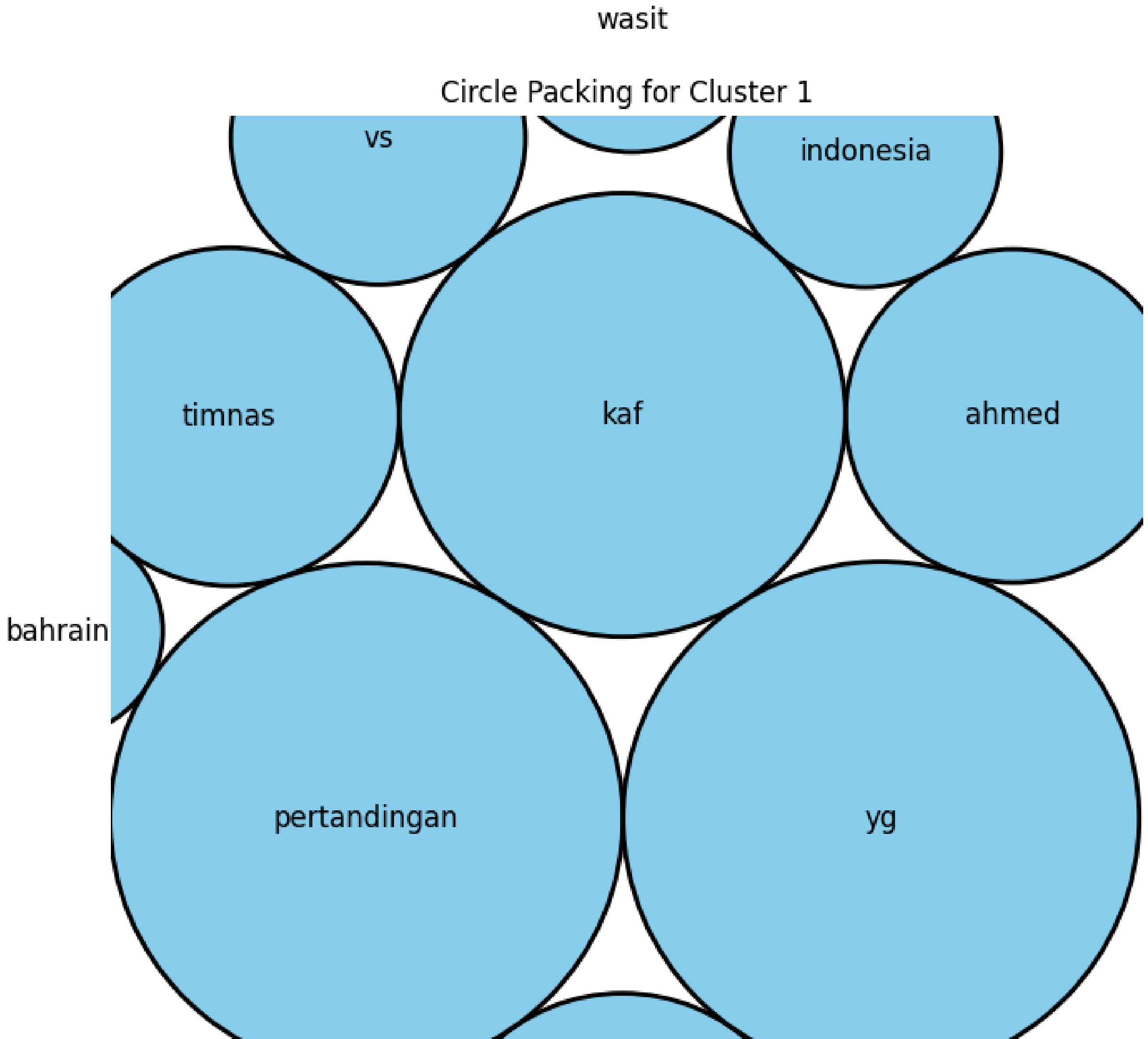
```
                                             full_text  cluster
0    Timnas bahrain harusnya malu sih karena media ...        1
1    Media luar mulai menyorot terkait kontroversi ...        1
2    Di video terlihat golnya persis ketika tulisan...        0
3     | Rangkuman kecurangan wasit Ahmed Al-Kaf di ...        1
4     Bahrain have a lot of money but dont have a c...        2
..                                                 ...      ...
214  #HeadlineNewsMetroTV 1. Duel Timnas Indonesia ...        1
215  Pemain Bahrain udah boleh gemeter belum? https...        2
216  para korban judol tidak boleh ditinggalkan ata...        2
217  Kayaknya baru kali ini gue ngamuk di X gara2 w...        1
218   ..Kalaupun kau menang itu awal dari kekalahan...        2

[219 rows x 2 columns]
```

```
!pip install circlify
```

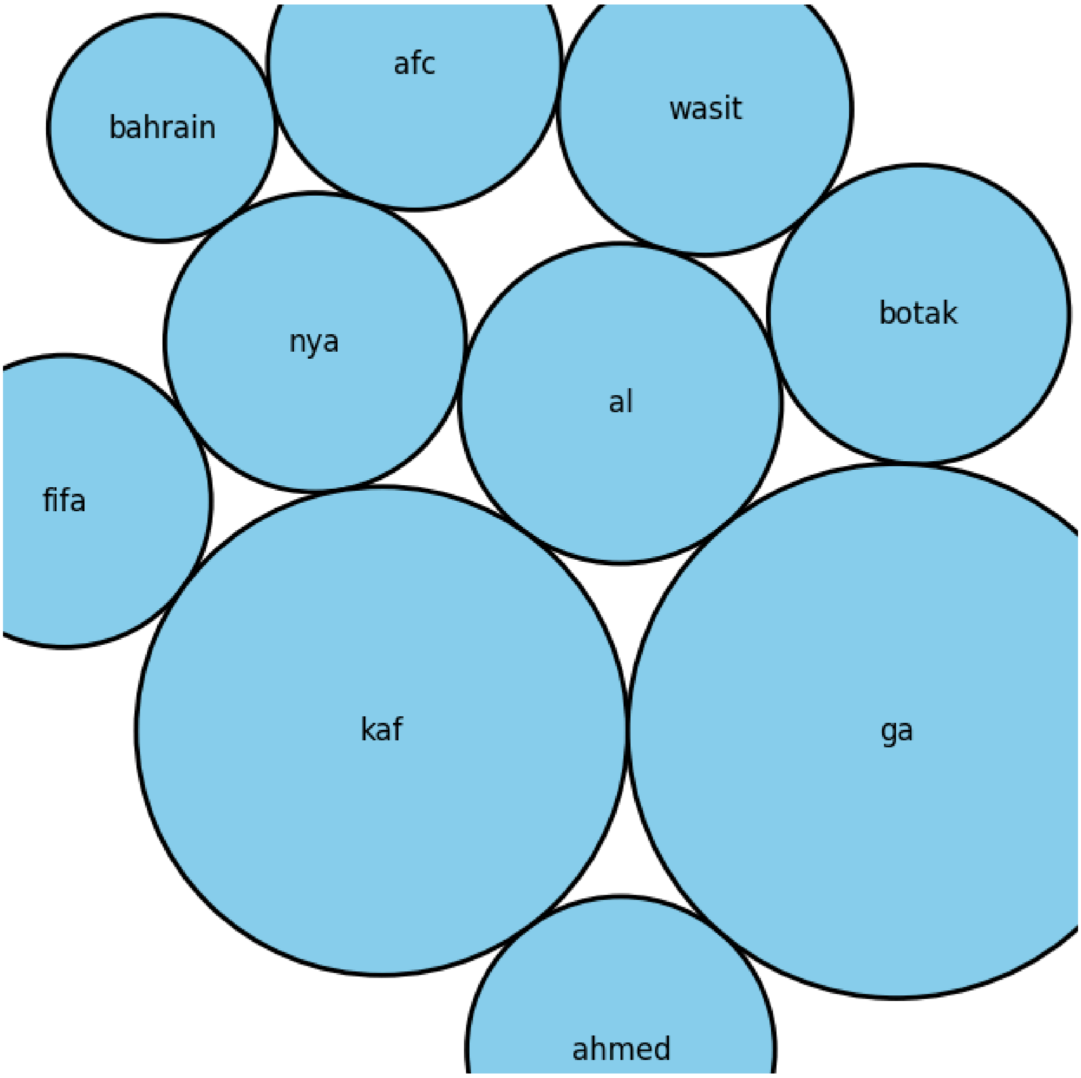
```
Requirement already satisfied: circlify in /usr/local/lib/python3.10/dist-packages (0.15.0)
```

```
import circlify
import matplotlib.pyplot as plt

def plot_circle_packing(cluster_num):
    # Gabungkan tweet berdasarkan cluster
    text = " ".join(tweet for tweet in df[df['cluster'] == cluster_num]['cleaned_tweet'])

    # Hitung frekuensi kata
    word_count = Counter(text.split())

    # Hapus stopwords
    filtered_word_count = {word: count for word, count in word_count.items() if word not in stop_words}

    # Ambil 10 kata paling sering muncul
    most_common_words = dict(Counter(filtered_word_count).most_common(10))

    # Persiapkan data untuk circle packing
    circles = circlify.circlify(list(most_common_words.values()), show_enclosure=False)

    # Plotting
    fig, ax = plt.subplots(figsize=(8, 8))
    ax.set_title(f"Circle Packing for Cluster {cluster_num}")
    ax.axis('off')

    # Generate circle pack
    # The original line: lim = max(max(circle.r) for circle in circles) * 2
    # is changed to directly get the maximum radius from the circles
    lim = max(circle.r for circle in circles) * 2
    plt.xlim(-lim, lim)
    plt.ylim(-lim, lim)

    # Draw circles
    for circle, label in zip(circles, most_common_words.keys()):
        x, y, r = circle.x, circle.y, circle.r
        ax.add_patch(plt.Circle((x, y), r, edgecolor='black', facecolor='skyblue', lw=2))
        plt.text(x, y, label, ha='center', va='center', fontsize=12)

    plt.show()

# Plot circle packing for each cluster
for i in range(true_k):
    plot_circle_packing(i)
```


Circle Packing for Cluster 0
bahrain
wasit
indonesia
vs
keputusan
fifa
menit
gol
laga
pemain

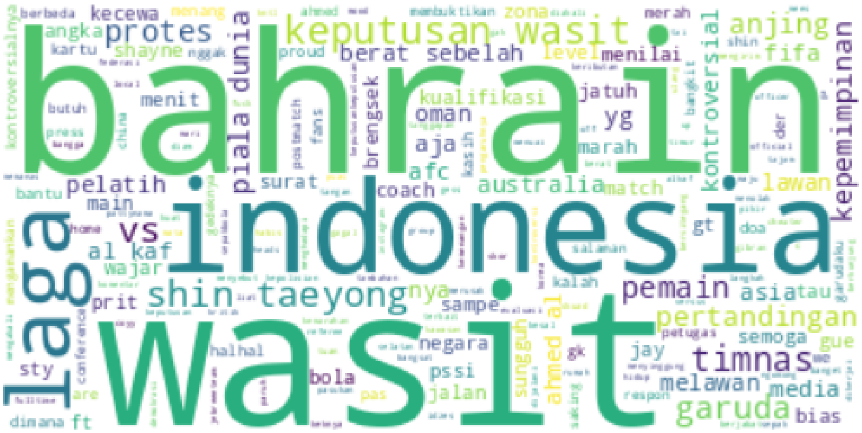

wasit
Circle Packing for Cluster 1
vs
indonesia
timnas
kaf
ahmed
bahrain
pertandingan
yg

Circle Packing for Cluster 2

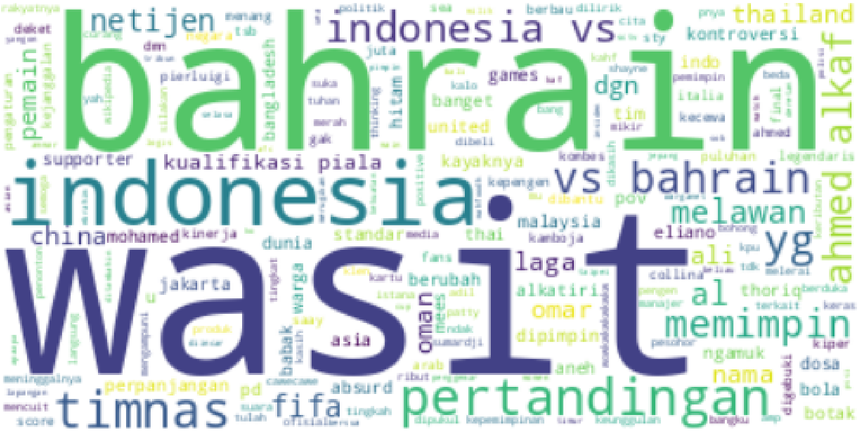


```
def plot_word_cloud(cluster_num):
    text = " ".join(tweet for tweet in df[df['cluster'] == cluster_num]['cleaned_tweet'])
    word_cloud = WordCloud(stopwords=stop_words, background_color="white").generate(text)
    plt.imshow(word_cloud, interpolation='bilinear')
    plt.axis("off")
    plt.title(f"Word Cloud for Cluster {cluster_num}")
    plt.show()

# Plot word clouds for each cluster
for i in range(true_k):
    plot_word_cloud(i)
```

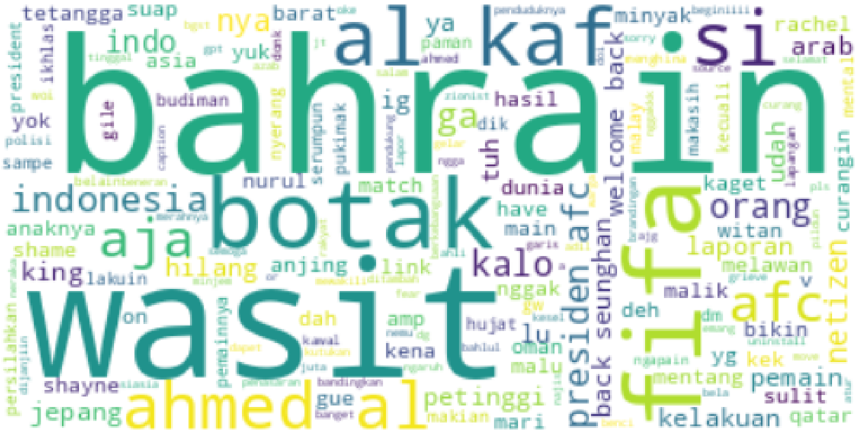
Word Cloud for Cluster 0



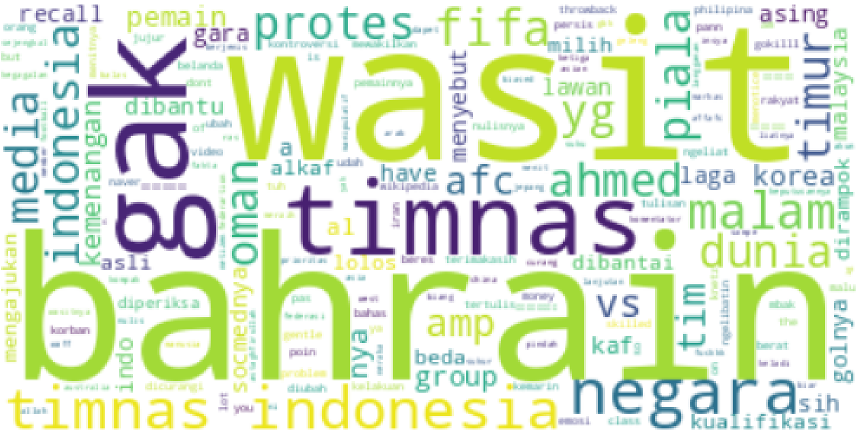
Word Cloud for Cluster 1



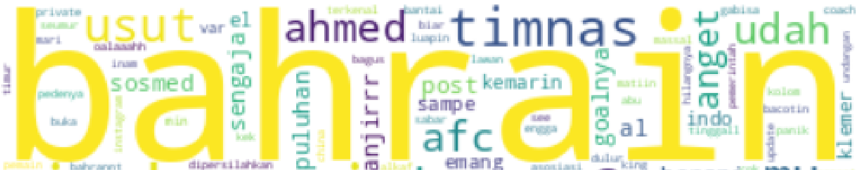
Word Cloud for Cluster 2



Word Cloud for Cluster 3

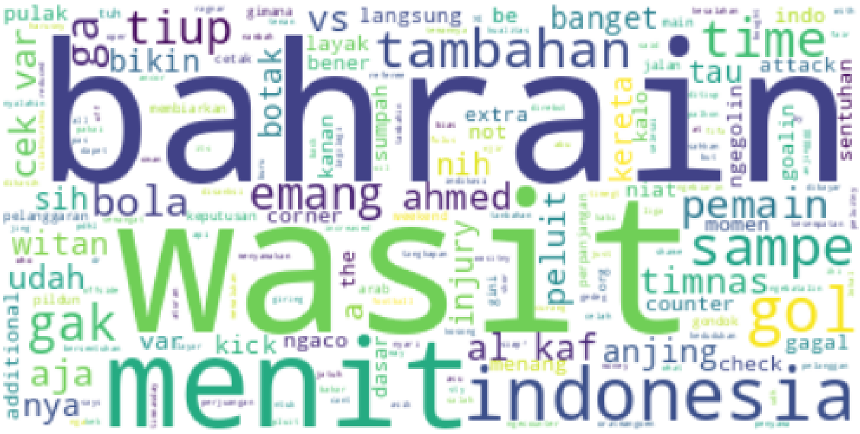


Word Cloud for Cluster 4

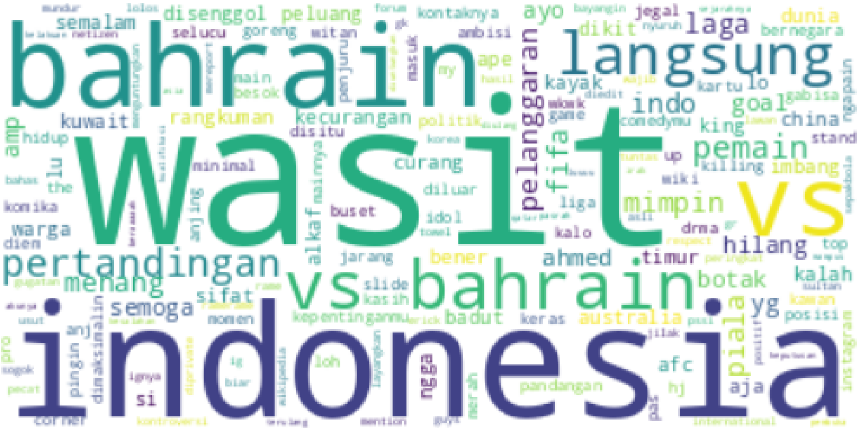


wasit
indonesia
fifa
nya
akun
nih
komen
yg
vs
ig
tu
banget
ga
kontol
ditulis
silaturahmi
pokoknya
tweet
emosi
yok
asu
share
botak
detik
fuck
ribu
terpantau
nitizen
aja
an
bola
mu
berani
kaf

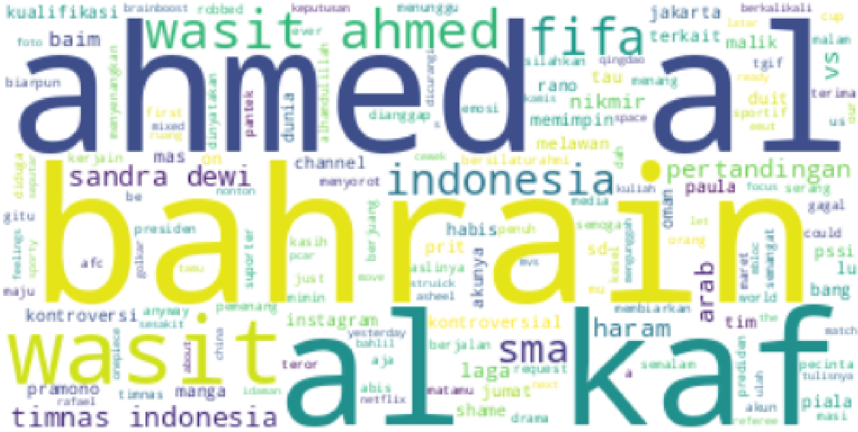
Word Cloud for Cluster 5



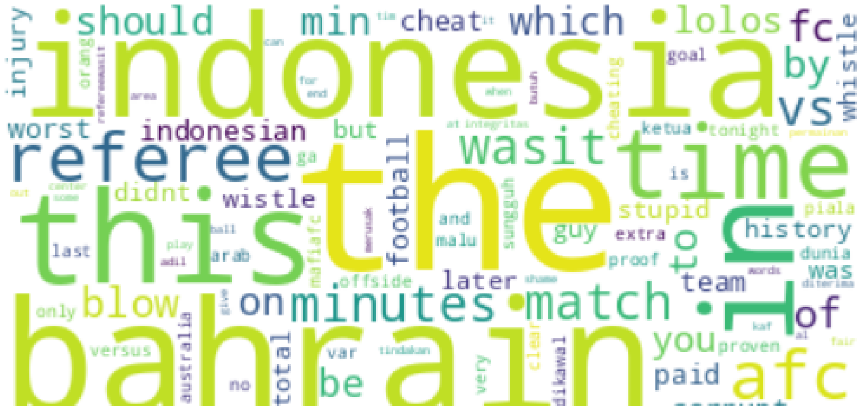
Word Cloud for Cluster 6



Word Cloud for Cluster 7



Word Cloud for Cluster 8



seen
sampe
someone
check
mengecewakan
exceeds
october
ahmed
corrupt
robbed
dear