LAPORAN TUGAS 3
PENAMBANGAN DATA

Laporan ini dibuat untuk memenuhi Tugas 3 Penambangan Data

Disusun oleh :

Kelompok 4

Dani Andhika Permana	1301180174
M. Naufal Mu'afa	1301180091
Muhammad Hafidh Raditya	1301184079
Hilmy Shabir Putra R.	1301184261

Universitas Telkom

Bandung

2021

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

1. Dataset Health News in Twitter

1.1. Analisis

Dataset Health News in Twitter berisi kumpulan *tweet* yang dibuat oleh para pengguna aplikasi Twitter. Dalam satu folder berisi 16 file dengan format *txt* yang mana tiap file berisi kumpulan *tweet* yang berhubungan dengan akun agensi berita, contohnya file dengan judul "bbchealth.txt" berisi *tweet* mengenai berita dari akun resmi BBC Health. Tiap baris data berisi *tweet id*, *date and time*, dan *tweet* yang dilengkapi juga dengan tanda pemisah (*separator*) "|". Pada laporan ini kami melakukan *text clustering* pada dataset ini untuk melihat topik apa saja yang dibicarakan.

1.2. Data Exploration

Pada tahap ini kami melakukan *import* 16 file .txt tersebut ke sebuah file bertipe .xlsx untuk memudahkan pengolahan dataset ini pada Jupyter Notebook.

```
df1 = pd.read_excel('health_news_dataset.xlsx', sheet_name='bbchealth')

df2 = pd.read_excel('health_news_dataset.xlsx', sheet_name='cbchealth')

df3 = pd.read_excel('health_news_dataset.xlsx', sheet_name='cnnhealth')

df4 = pd.read_excel('health_news_dataset.xlsx', sheet_name='everydayhealth')

df5 = pd.read_excel('health_news_dataset.xlsx', sheet_name='foxnewshealth')

df6 = pd.read_excel('health_news_dataset.xlsx', sheet_name='gdnhealthcare')

df7 = pd.read_excel('health_news_dataset.xlsx', sheet_name='goodhealth')

df8 = pd.read_excel('health_news_dataset.xlsx', sheet_name='KaiserHealthNews')

df9 = pd.read_excel('health_news_dataset.xlsx', sheet_name='latimeshealth')

df10 = pd.read_excel('health_news_dataset.xlsx', sheet_name='msnhealthnews')

df11 = pd.read_excel('health_news_dataset.xlsx', sheet_name='NBChealth')

df12 = pd.read_excel('health_news_dataset.xlsx', sheet_name='nprhealth')

df13 = pd.read_excel('health_news_dataset.xlsx', sheet_name='nytimeshealth')

df14 = pd.read_excel('health_news_dataset.xlsx', sheet_name='reuters_health')

df15 = pd.read_excel('health_news_dataset.xlsx', sheet_name='usnewshealth')

df16 = pd.read_excel('health_news_dataset.xlsx', sheet_name='wsjhealth')
```

Gambar 1 Code Data Exploration

```
df = pd.concat([df1,df2,df3,df4,df5,df6,df7,df8,df9,df10,df11,df12,df13,df14,df15,df16])
df.head()
```

Gambar 2 Code untuk menggabungkan seluruh sheet

	Column1	Column2	Column3
0	585978391360221184	Thu Apr 09 01:31:50 +0000 2015	Breast cancer risk test devised http://bbc.in/...
1	585947808772960256	Wed Apr 08 23:30:18 +0000 2015	GP workload harming care - BMA poll http://bbc...
2	585947807816650752	Wed Apr 08 23:30:18 +0000 2015	Short people's 'heart risk greater' http://bbc...
3	585866060991078400	Wed Apr 08 18:05:28 +0000 2015	New approach against HIV 'promising' http://bb...
4	585794106170839040	Wed Apr 08 13:19:33 +0000 2015	Coalition 'undermined NHS' - doctors http://bb...

Gambar 3 Output Data Exploration (1)

```
df.columns=['Tweet_ID','Date','Tweet']
df
```

Gambar 4 Code untuk memberi nama kolom

	Tweet_ID	Date	Tweet
0	585978391360221184	Thu Apr 09 01:31:50 +0000 2015	Breast cancer risk test devised http://bbc.in/...
1	585947808772960256	Wed Apr 08 23:30:18 +0000 2015	GP workload harming care - BMA poll http://bbc...
2	585947807816650752	Wed Apr 08 23:30:18 +0000 2015	Short people's 'heart risk greater' http://bbc...
3	585866060991078400	Wed Apr 08 18:05:28 +0000 2015	New approach against HIV 'promising' http://bb...
4	585794106170839040	Wed Apr 08 13:19:33 +0000 2015	Coalition 'undermined NHS' - doctors http://bb...
...	...	...	...
3195	415494259022655488	Tue Dec 24 14:48:45 +0000 2013	RT @stefaniei: Addiction and the brain: scient...
3196	415493351396233216	Tue Dec 24 14:45:09 +0000 2013	RT @timothywmartin: Ho-ho-hold up! A surprise ...
3197	415493203983204352	Tue Dec 24 14:44:33 +0000 2013	RT @stefaniei: Health-Insurance Deadline Exten...
3198	415386956420231168	Tue Dec 24 07:42:22 +0000 2013	Boston Scientific Eyes China Expansion http://...
3199	415361763362603008	Tue Dec 24 06:02:16 +0000 2013	For Desperate Family in India, a Ray of Hope F...

59923 rows × 3 columns

Gambar 5 Output Data Exploration (2)

Setelah dilakukan *data exploration* dapat dilihat bahwa dataset ini memiliki 59923 record data dan tiga kolom. Ketiga kolom tersebut pada awalnya belum diberi nama, sehingga pada Gambar 4 kami menamakan ketiga kolom tersebut secara berurut : “Tweet_ID” yang berisi ID dari tiap tweet yang dibuat, “Date” yang berisi tanggal dan waktu dari setiap tweet, dan “Tweet” yang merupakan isi dari setiap tweet tersebut.

```
df.info()
```

Gambar 6 Code pengecekan tipe data

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 59923 entries, 0 to 3199
Data columns (total 3 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   Tweet_ID  59923 non-null  object
 1   Date      59923 non-null  object
 2   Tweet     57081 non-null  object
dtypes: object(3)
memory usage: 1.8+ MB
```

Gambar 7 Tipe data

Pada Gambar 6 kami melakukan pengecekan tipe data. Dapat terlihat pada Gambar 7 bahwa setiap kolom memiliki tipe data *object*.

1.3. Pengecekan *Null Values*

Pada tahap ini kami melakukan pengecekan *null values* di ketiga kolom tersebut.

```
df.isnull().sum()
```

Gambar 8 Code untuk mengolah null values

```
Tweet_ID       0
Date           0
Tweet       2842
dtype: int64
```

Gambar 9 Null values pada dataset

Terdapat total 2842 *null values* pada atribut “Tweet”. Karena atribut “Tweet” ini beritpe string yang tidak bisa di-*replace* dengan modus, mean, atau median dari keseluruhan isi atribut, sehingga kami langsung drop 2842 data tersebut.

```
df.dropna(inplace=True)
```

```
df.isnull().sum()
```

```
Tweet_ID    0
Date        0
Tweet       0
dtype: int64
```

Gambar 10 Sudah tidak ada null values

1.4. Seleksi Kata

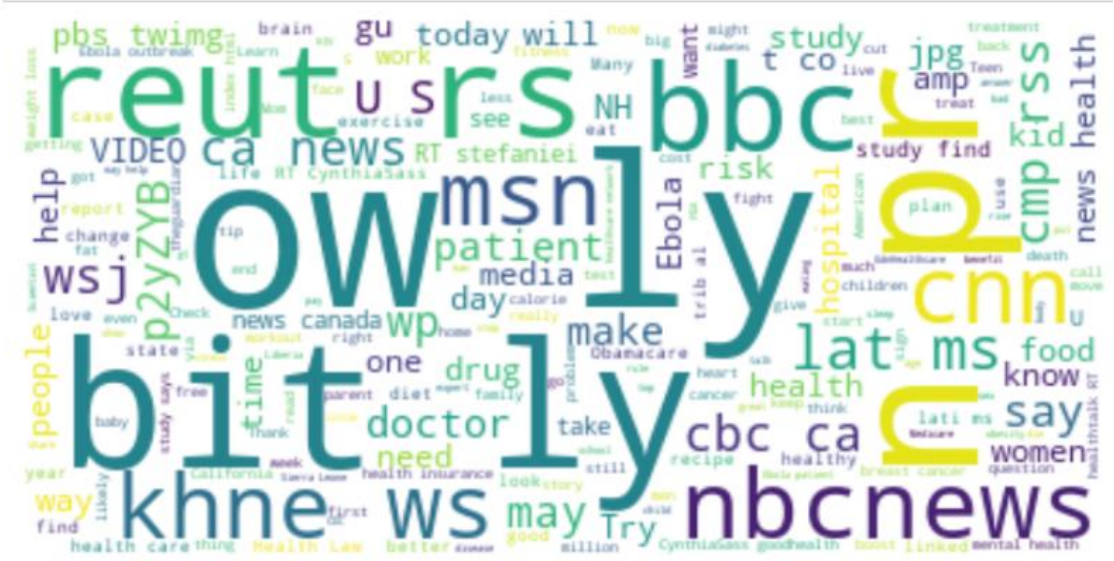
Langkah selanjutnya yaitu melakukan pengecekan kata-kata apa saja yang paling sering muncul pada atribut “Tweet” menggunakan *library* wordcloud.

```
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
from sklearn.feature_extraction.text import ENGLISH_STOP_WORDS

text = " ".join(text for text in df.Tweet)
wordcloud = WordCloud(background_color="white").generate(text)

plt.figure(figsize=(16,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Gambar 11 Code untuk pengecekan kata



Gambar 12 Output dari wordcloud

Selanjutnya kami menyeleksi kata-kata yang tergolong stopwords dalam wordcloud yang telah ditampilkan pada Gambar 12. Stopwords merupakan kata-kata yang sekiranya tidak memiliki relevansi terhadap topik yang sedang dikerjakan. Dapat dilihat bahwa banyak sekali kata-kata random yang tidak memiliki arti yang akan sangat memengaruhi proses clustering. Selain itu karena tweet pada dataset ini dibuat dalam Bahasa Inggris, kami juga menggunakan *package* ENGLISH_STOP_WORDS dari *lbrary* sklearn.feature_extraction. *Package* ini berisi stopwords yang sering muncul dalam Bahasa Inggris seperti “the”, “and”, “or”, dll yang tidak memiliki relevansi terhadap suatu topik khusus. Oleh karena itu kata-kata yang termasuk stopwords perlu dihapus dari dataset ini.

```
stopwords = set(STOPWORDS)
stopwords.update(['bit','ly','n','pr','reut','rs','cnn','lat','ms',
                  'nbcnews','cmp','rss','p2yZYB','msn','wp','jpg','VIDEO','wsj',
                  't','co','try','pbs','twimg','gu','RT','stefaniei','ws','OW','cbc',
                  'cbc','ca','bbc','U','S','ca','khne','WS','NH','amp','lati','nhs',
                  'stw','theguardian','https','tw','ow','gt','p2yzyb','rt','html',
                   'nhs','trib','1gny60o','gdnhealthcare','q1','q2','q3','q4','q5','q6',
                  '1ajrqyn', 'hkmaub','npf','1ajrirw','q7','1akyupf','1gnn3nv'])

my_stopwords = ENGLISH_STOP_WORDS.union(stopwords)

wordcloud = WordCloud(stopwords=stopwords, background_color="white").generate(text)

plt.figure(figsize=(16,8))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis("off")
plt.show()
```

Gambar 13 Code untuk seleksi stopwords

Gambar 14 Output wordcloud (tanpa stopwords)

```
text = " ".join(text for text in df.Tweet)
print ("Terdapat total {} kata dari semua review yang terdapat di kolom grade.".format(len(text)))
```

Gambar 15 Jumlah kata di kolom grade

Pada Gambar 15 dilakukan penghitungan kata pada kolom grade. Terdapat 5619226 kata dari semua review yang terdapat di kolom grade.

Selanjutnya kami melakukan encoding pada atribut “Tweet” ke dalam bentuk Unicode. Unicode merupakan bentuk standar encoding untuk merepresentasikan suatu teks secara digital. Hanya atribut “Tweet” yang dilakukan encoding dikarenakan hanya atribut

“Tweet” yang dilakukan clustering untuk melihat topik apa saja yang dibicarakan, sedangkan atribut “Tweet_ID” dan “Date” tidak berpengaruh.

```
dataset = df['Tweet'].values.astype('U')
```

Gambar 16 Encoding atribut "Tweet"

Selanjutnya kami menggunakan *package* TfidVectorizer dari *library* sklearn.feature_extraction. *Package* ini melakukan penghapusan kata-kata yang tergolong stopwords yang sebelumnya sudah kami definisikan.

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

Gambar 17 Import package TfidVectorizer dan Kmeans

```
vectorizer = TfidfVectorizer(stop_words=my_stopwords)
features = vectorizer.fit_transform(dataset)
```

```
c:\users\msi\appdata\local\programs\python\python38\lib\site-packages\sklearn\feature_extraction\tex
t.py:396: UserWarning: Your stop_words may be inconsistent with your preprocessing. Tokenizing the st
op words generated tokens ['aren', 'couldn', 'didn', 'doesn', 'don', 'hadn', 'hasn', 'haven', 'isn',
'let', 'll', 'mustn', 'nh', 'shan', 'shouldn', 've', 'video', 'wasn', 'weren', 'won', 'wouldn'] not i
n stop_words.
  warnings.warn(
```

Gambar 18 Code penghapusan stopwords

1.5. Proses Clustering

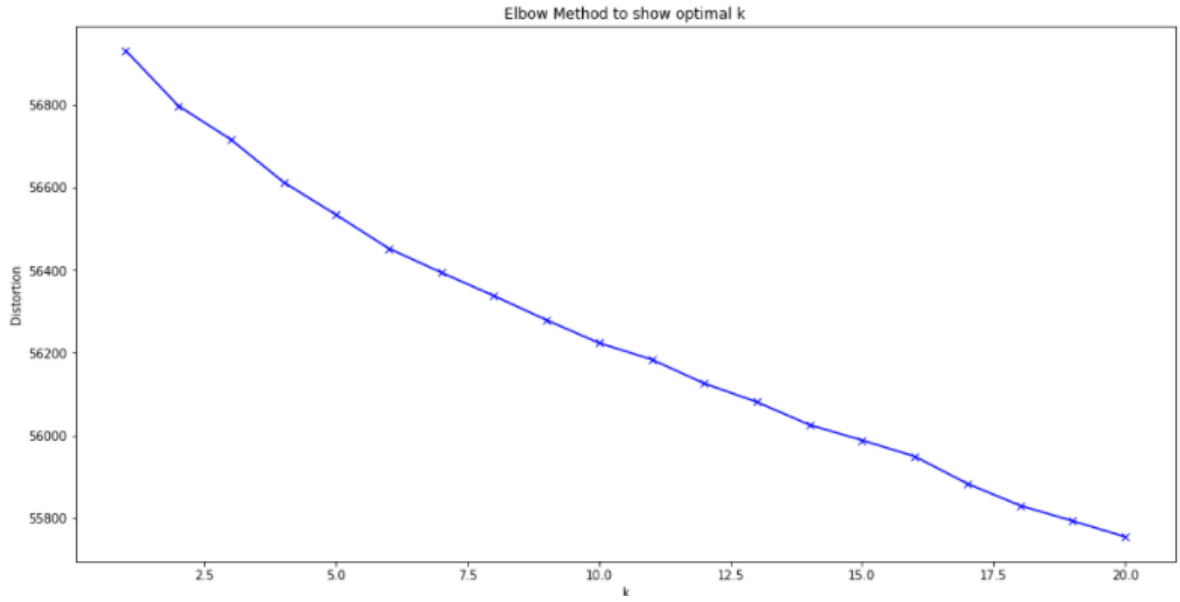
Pada tahap ini kami melakukan clustering pada dataset Health News in Twitter menggunakan algoritma KMeans. Penggunakan algoritma KMeans ini didasarkan pada data pada dataset ini sangat kompleks yang berisi hampir 6 juta kata yang *unique*/berbeda. Langkah awal pada proses clustering yang kami lakukan yaitu mencari nilai k yang optimal menggunakan metode elbow.

```
distortions = []
K = range(1, 21)
for k in K:
    kmeanmodel = KMeans(n_clusters=k, max_iter=100)
    kmeanmodel.fit(features)
    distortions.append(kmeanmodel.inertia_)
```

Gambar 19 Code untuk metode Elbow (1)

```
plt.figure(figsize=(16,8))
plt.plot(K, distortions, 'bx-')
plt.xlabel('k')
plt.ylabel('Distortion')
plt.title('Elbow Method to show optimal k')
plt.show()
```

Gambar 20 Code untuk metode Elbow (2)



Gambar 21 Output metode Elbow

Terlihat pada Gambar 21 bahwa grafik tidak menunjukkan pola siku atau *elbow* bahkan sampai nilai k=20. Dikarenakan keterbatasan hardware yang kami miliki yang tidak memungkinkan untuk melakukan percobaan Elbow Method hingga k>20 yang prosesnya akan sangat lama, maka kami menetapkan untuk menggunakan nilai k=20.

```
k = 20
model = KMeans(n_clusters=k)
model.fit(features)
```

```
KMeans(n_clusters=20)
```

Gambar 22 Code clustering dataset (1)

```
df['Cluster'] = model.labels_
df.head()
```

Gambar 23 Code clustering dataset (2)

	Tweet_ID	Date	Tweet	Cluster
0	585978391360221184	Thu Apr 09 01:31:50 +0000 2015	Breast cancer risk test devised http://bbc.in/...	4
1	585947808772960256	Wed Apr 08 23:30:18 +0000 2015	GP workload harming care - BMA poll http://bbc...	11
2	585947807816650752	Wed Apr 08 23:30:18 +0000 2015	Short people's 'heart risk greater' http://bbc...	1
3	585866060991078400	Wed Apr 08 18:05:28 +0000 2015	New approach against HIV 'promising' http://bb...	9
4	585794106170839040	Wed Apr 08 13:19:33 +0000 2015	Coalition 'undermined NHS' - doctors http://bb...	18

Gambar 24 Output clustering dataset (1)

```
clusters = df.groupby('Cluster')

for cluster in clusters.groups:
#     f = open('cluster'+str(cluster)+ '.csv', 'w')
    data = clusters.get_group(cluster)[['Tweet']]
#     f.write(data.to_csv(index_label='id'))
#     f.close()
data
```

Gambar 25 Code clustering dataset (3)

	Tweet
874	Late night food 'breeds weight gain' http://bb...
929	Guidelines favour weight loss ops http://bbc.i...
1084	VIDEO: UK sees rise in weight-loss surgery htt...
1092	Inherited bugs may help weight loss http://bbc...
1118	AUDIO: Weight loss surgery 'cured my diabetes'...
...	...
1072	RT @stefaniei: Orexigen's weight-loss drug Con...
1080	RT @JeanneWhalen: Novo Nordisk wants to sell d...
1365	RT @stefaniei: Maybe Mom was right: Chew your ...
2478	Glaxo Recalls Alli Weight-Loss Products on Tam...
3134	RT @CorbettDooren: FTC Charges Weight-Loss Pro...

932 rows × 1 columns

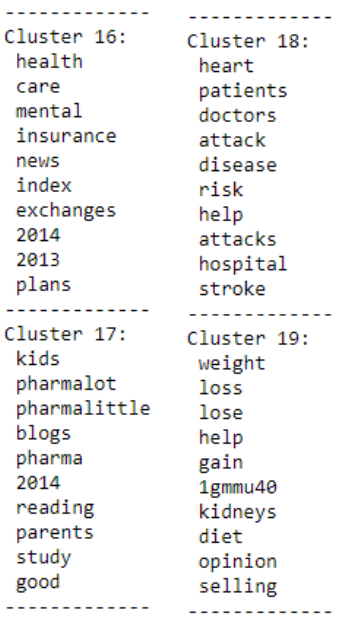
Gambar 26 Output clustering dataset (2)

```
print("Cluster centroids: \n")
order_centroids = model.cluster_centers_.argsort()[:, ::-1]
terms = vectorizer.get_feature_names()

for i in range(k):
    print("Cluster %d:" % i)
    for j in order_centroids[i, :10]:
        print(' %s' % terms[j])
    print('-------------')
```

Gambar 27 Code clustering dataset (4)

Gambar 28 Output clustering dataset (3)



Gambar 29 Output clustering dataset (4)

Proses clustering telah selesai. Pada Gambar 29 kami mengoutputkan 10 kata yang paling sering muncul pada setiap cluster lalu kami analisa topik apa yang dibicarakan pada tiap cluster. Hasil analisa kami dapat dilihat pada tabel berikut :

Tabel 1 Tabel Analisa Topik

Cluster	**10 kata paling sering muncul**	**Topik yang dibicarakan**
Cluster 0	video, drug, flu, fda, hospital, bird, approves, painkiller, 1ajyfq4, powerful	Perizinan obat yang sudah disetujui oleh FDA, yaitu badan yang mengurusi seputar perizinan peredaran makanan dan obat-obatan di Amerika Serikat.
Cluster 1	healthtalk, people, risk, everydayhealth, diabetes, eatsmartbd, questions, join, psoriasis, healthtotalwellness	Obrolan kesehatan secara umum yang menyinggung penyakit diabetes dan psoriasis.
Cluster 2	life, end, sex, day, care, years, expectancy, saving, quality, support	Berbicara tentang life expectancy (harapan hidup).
Cluster 3	ebola, africa, liberia, leone, outbreak, sierra, west, patient, health, nurse	Wabah virus ebola beberapa tahun yang lalu yang menyebar di benua Afrika.
Cluster 4	breast, cancer, risk, women, study, feeding, milk, drug, mastectomy, patients	Kanker payudara yang dapat menyerang wanita. Cluster ini juga banyak membahas tentang mastectomy, yaitu operasi pengangkatan payudara yang merupakan salah satu bentuk pengobatan untuk pasien kanker payudara.

Cluster 5	law, health, court, headlines, insurance, obama, coverage, today, louiseradnofsky, new	Asuransi kesehatan untuk masyarakat yang digagas oleh Presiden Barrack Obama pada saat beliau masih menjabat sebagai presiden Amerika Serikat.
Cluster 6	healthcare, network, gov, professionals, register, free, 2014, 2015, november, smoothly	Layanan kesehatan yang dibuka pada tahun 2014-2015 oleh pemerintah Amerika Serikat untuk masyarakatnya secara gratis.
Cluster 7	cancer, prostate, risk, drug, lung, colon, study, treatment, patients, new	Penyakit kanker secara umum, seperti kanker prostat, kanker paru-paru, dan kanker colon.
Cluster 8	day, recipe, ways, valentine, today, 10, happy, national, calories, start	Tidak dapat disimpulkan karena 10 kata yang paling sering muncul terlalu random dan tidak dapat dibaca polanya.
Cluster 9	new, york, rules, year, fda, ebola, health, strict, proposes, animal	Peraturan yang yang dibuat oleh pemerintah Amerika Serikat untuk mencegah penyebaran wabah ebola.
Cluster 10	study, finds, suggests, risk, new, shows, kidney, actually, brain, paying	Suatu studi kesehatan yang membahas tentang otak dan ginjal.
Cluster 11	today, help, obamacare, media, healthy, 10, make, ways, brain, food	10 tips kesehatan yang berhubungan dengan makanan. Cluster ini juga membahas tentang obamacare.
Cluster 12	women, men, pregnant, study, risk, heart, cancer, older, disease, says	Studi kesehatan tentang penyakit kanker dan penyakit hati sekaligus resikonya terhadap beberapa subjek seperti pria, wanita, wanita hamil, dan lansia.
Cluster 13	says, study, ebola, cdc, health, new, exposed, bra, 1gnqsca, publicity	Studi tentang wabah virus ebola.
Cluster 14	cynthiasass, goodhealth, talknutrition, weight, eating, thanks, healthy, thanksgiving, eat, breakfast	Cynthia Sass, ahli nutrisi dari Amerika Serikat yang membicarakan topik seputar nutrisi.
Cluster 15	foods, eat, best, al, 10, boost, fat, worst, wagdotcom, life	Makanan dari sudut pandang kesehatan.
Cluster 16	health, care, mental, insurance, news, index, exchanges, 2014, 2013, plans	Layanan dan asuransi kesehatan.
Cluster 17	kids, pharmalot, pharmalittle, blogs, pharma, 2014, reading, parents, study, good	Studi yang membahas tentang orang tua dan anak-anak dari sudut pandang kesehatan.
Cluster 18	heart, patients, doctors, attack, disease, risk, help, attacks, hospital, stroke	Penyakit stroke dan serangan jantung.

Cluster 19	weight, loss, lose, help, gain, 1gmmu40, kidneys, diet, opinion, selling	Diet untuk menurunkan atau menaikkan berat badan.

2. Dataset Daily and Sports Activities

2.1. Analisis Dataset

Dataset yang digunakan adalah Daily and Activities Dataset. Dataset ini berisi data sensor accelerometer, gyroscope, dan magnetometer. Sensor tersebut dipasang pada 8 subjek (4 wanita, 4 pria, antara umur 20 sampai 30). 8 subjek tersebut disebut dengan P1, P2, …, P8.

Subjek diminta untuk melakukan 29 aktivitas, yaitu :

- Duduk (A01)
- Berdiri (A02)
- Berbaring telentang (A03)
- Berbaring miring kekanan (A04)
- Menaiki tangga (A05)
- Menuruni tangga (A06)
- Berdiri diam di lift (A07)
- Berjalan mengitari lift (A08)
- Berjalan di tempat parkir (A09)
- Berjalan di treadmill dengan kecepatan 4 km/j (kemiringan 0°) (A10)
- Berjalan di treadmill dengan kecepatan 4 km/j (kemiringan 15°) (A11)
- Berlari di treadmill dengan kecepatan 8 km/j (A12)
- Olahraga dengan stepper (A13)
- Olahraga dengan cross trainer (A14)
- Bersepeda menggunakan exercise bike dengan posisi horizontal (A15)
- Bersepeda menggunakan exercise bike dengan posisi vertical (A16)
- Mendayung (A17)
- Melompat (A18)
- Bermain basket (A19)

Setiap aktivitas di atas dilakukan selama 5 menit (300 detik).

Setiap subjek dipasangi 5 buah alat, masing-masing di Torso (T), right arm (RA), left arm (LA), right leg (RL), dan left leg (LL). Setiap alat mempunyai 9 buah sensor (x,y,z accelerometer, x,y,z gyroscopes, dan x,y,z magnetometers). Setiap sensor dikalibrasi untuk mendapatkan data dengan frekuensi 25Hz (25 data per detik).

	T_xacc	T_yacc	T_zacc	T_xgyro	T_ygyro	T_zgyro	T_xmag	T_ymag	T_zmag	RA_xacc	RA_yacc	RA_zacc	RA_xgyro	RA_ygyro	RA_zgyro
1	7,97840	-1,84490	1,30190	-0,753690	-0,319520	-0,303070	-0,473850	-0,628090	0,209890	7,13140	2,71680	1,55970	-0,433780	0,0352310	-0,529840
2	7,24410	-1,59840	0,373030	-0,632040	0,454020	-0,275330	-0,463690	-0,642890	0,191270	7,14000	2,74730	1,27320	-0,274860	0,0435870	-0,771190
3	7,61770	-1,81290	-0,431060	-0,216350	0,0179520	-0,353160	-0,456370	-0,652890	0,180380	7,04590	3,01440	1,17670	-0,301010	-0,0683440	-0,843580
4	10,7830	-0,806660	-0,982300	-0,154620	0,0982070	-0,544610	-0,444140	-0,665290	0,174470	9,51360	3,18290	0,373840	-0,297930	-0,244000	-0,802320
5	15,2800	1,39560	-1,38890	-0,511570	-0,570440	-0,963060	-0,418590	-0,683190	0,174010	14,7870	4,11600	-0,709550	-0,0306300	-0,211940	-0,501020

Gambar 30 Dataset P1 yang melakukan A01, bagian 1

RL_xgyro	RL_ygyro	RL_zgyro	RL_xmag	RL_ymag	RL_zmag	LL_xacc	LL_yacc	LL_zacc	LL_xgyro	LL_ygyro	LL_zgyro	LL_xmag	LL_ymag	LL_zmag
-0,725420	-0,576030	0,482800	0,459710	0,334530	-0,522010	-7,71350	-3,14200	-1,24720	0,412200	0,0962040	0,286630	0,427590	-0,506210	0,476510
0,747000	-0,837980	0,745040	0,453140	0,316690	-0,537010	-8,05920	-3,36140	-2,12000	0,211490	0,109980	0,463740	0,418650	-0,507130	0,486900
-0,0855310	-1,03860	0,624600	0,443770	0,304960	-0,553420	-7,62420	-2,91480	-2,27570	0,690350	0,199190	0,450280	0,407490	-0,510440	0,494350
-0,709650	-0,635020	0,376770	0,430150	0,309940	-0,560090	-14,5520	-1,15080	2,33190	0,236340	0,221510	0,404010	0,391110	-0,499720	0,518770
-1,17060	-0,0398720	1,15480	0,428710	0,323910	-0,554270	-15,8760	2,83280	-3,10160	0,0157840	0,271110	-0,409690	0,387740	-0,500550	0,525020

Gambar 31 Dataset P1 yang melakukan A01, bagian 2

RA_zgyro	RA_xmag	RA_ymag	RA_zmag	LA_xacc	LA_yacc	LA_zacc	LA_xgyro	LA_ygyro	LA_zgyro	LA_xmag	LA_ymag	LA_zmag	RL_xacc	RL_yacc	RL_zacc
-0,529840	-0,388330	-0,333540	0,645380	7,57230	-1,41700	0,677680	-0,789400	0,523770	-0,227560	-0,441100	-0,591160	-0,240660	-7,20240	3,28820	-1,28900
-0,771190	-0,378500	-0,353170	0,640370	7,87490	-1,69050	0,679680	-0,845820	0,356410	-0,204950	-0,431040	-0,585990	-0,268440	-9,45760	2,96920	1,14140
-0,843580	-0,367270	-0,369500	0,638500	7,80200	-1,93090	0,934540	-0,546950	0,188520	-0,321580	-0,423160	-0,584740	-0,285580	-7,71540	-0,360370	1,15750
-0,802320	-0,349480	-0,389470	0,639050	10,5450	-1,76700	0,660370	-0,305120	-0,0147590	-0,482540	-0,411580	-0,588100	-0,294690	-9,36180	0,210130	-2,37910
-0,501020	-0,334440	-0,402010	0,641080	15,9040	-1,53260	1,04060	-0,324580	-0,371080	-0,735240	-0,399750	-0,596830	-0,296550	-19,6820	6,99820	-0,238500

Gambar 32 Dataset P1 yang melakukan A01, bagian 3

Data pada gambar diatas adalah data P1 yang melakukan A01. Ada 5 alat × 9 sensor = 45 kolom pada dataset tersebut, yaitu :

- T_xacc, T_yacc, T_zacc, T_xgyro, ..., T_ymag, T_zmag,
- RA_xacc, RA_yacc, RA_zacc, RA_xgyro, ..., RA_ymag, RA_zmag,
- LA_xacc, LA_yacc, LA_zacc, LA_xgyro, ..., LA_ymag, LA_zmag,
- RL_xacc, RL_yacc, RL_zacc, RL_xgyro, ..., RL_ymag, RL_zmag,
- LL_xacc, LL_yacc, LL_zacc, LL_xgyro, ..., LL_ymag, LL_zmag.

Sensor bekerja selama 300 detik dengan frekuensi 25Hz. Jadi ada 300×25 = 7500 baris dalam dataset tersebut.

Ada 19 aktivitas dan 8 subjek, jadi ada 19×8=152 dataset.

Subjek akan dikelompokkan menggunakan k-means.

2.2. Data Preprocessing

Untuk meringkas dataset, setiap kolom pada dataset akan diambil rata-ratanya saja.

```
new <- as.data.frame(colMeans(all))
new <- transpose(new)
names(new) <- col.names
```

Gambar 33 Kode untuk mengambil rata-rata

	T_xacc	T_yacc	T_zacc	T_xgyro	T_ygyro	T_zgyro	T_xmag	T_ymag
1	7,86329	1,34433	5,71081	0,00289427	0,0201827	-0,00316057	-0,792021	-0,0762106

Gambar 34 Rata-rata setiap kolom

Pada Gambar 34, didapatkan rata-rata setiap kolom pada dataset pertama (P1 dengan A01). Untuk memperjelas data, data tersebut diberi keterangan subjek dan aktivitas yang dilakukan.

```
new$action <- act
new$subject <- sub
new <- new[, c("action","subject", col.names)]
```

Gambar 35 Kode untuk meanmbahkan kolom action dan subject

	action	subject	T_xacc	T_yacc
1	a01	p1	7,86329	1,34433

Gambar 36 Kolom action dan subject sudah ditambahkan

Kami melakukan hal yang sama untuk dataset lainnya. Kami mengambil rata-rata setiap kolom dan menambah keterangan action dan subject, kemudian menambahkan keterangan aktivitas dan subject yang berkaitan dengan dataset tersebut.

Setelah mengambil rata-rata dan menambah keterangan pada setiap dataset, kami menggabungkan data rata-rata setiap dataset.

```r
library(dplyr)

act.vec <-  c("a01","a02", "a03", "a04", "a05", "a06", "a07", "a08", "a09", "a10", "a
sub.vec <- c("p1", "p2", "p3", "p4", "p5", "p6", "p7", "p8")

all <- read.csv( file: "res_pre_1/a01.p1.csv")

i <- 1
for (act in act.vec) {
  for (sub in sub.vec) {
    if (i == 1) {
      all <- read.csv( file: "res_pre_1/a01.p1.csv")
    }
    else {
      df <- read.csv(paste("res_pre_1//",act,".",sub,".csv", sep=""))
      all <- rbind(all, df)
    }
    i <- i + 1
  }
}
```

Gambar 37 Kode untuk menggabungkan data rata-rata

	activity	subject	T_xacc	T_yacc	T_zacc
1	a01	p1	7,86329	1,34433	5,71081
2	a01	p2	9,62455	-0,815887	-1,33208
3	a01	p3	8,99982	-1,55490	3,57995
4	a01	p4	9,18405	-1,56850	3,07794
5	a01	p5	6,95058	-0,441805	6,92877
6	a01	p6	8,09705	-0,342074	5,51857
7	a01	p7	9,13635	-1,93009	2,98358
8	a01	p8	7,69647	-0,399498	6,09863
9	a02	p1	8,48189	0,440822	4,91710
10	a02	p2	9,51993	0,0240301	2,32517
11	a02	p3	9,55202	-1,03400	1,95072
12	a02	p4	9,20837	-2,16705	2,57405

Gambar 38 Rata-rata semua kolom pada setiap dataset

Pada Gambar 38, terlihat data rata-rata sudah digabungkan untuk semua dataset. Data gabungan tersebut kini menjadi dataset baru. Terdapat 152 baris pada dataset tersebut, sesuai dengan jumlah dataset sebelumnya.

Kami melakukakn casting untuk dataset tersebut, setiap activity akan disebar ke kolom T_xacc, T_yacc, dan seterusnya, sehingga menghasilkan dataset berikut.

	subject	T_xacc_a01	T_xacc_a02	T_xacc_a03
1	p1	7,86329	8,48189	-4,70459
2	p2	9,62455	9,51993	-1,95127
3	p3	8,99982	9,55202	-3,39613
4	p4	9,18405	9,20837	-2,96977
5	p5	6,95058	9,03124	-3,82681
6	p6	8,09705	9,12903	-4,61719
7	p7	9,13635	8,65061	-4,52362
8	p8	7,69647	9,76172	-1,68209

Gambar 39 Empat kolom pertama dataset (setelah proses casting)

LL_zmag_a16	LL_zmag_a17	LL_zmag_a18	LL_zmag_a19
-0,417995	-0,00558959	-0,452415	0,104421
-0,255687	0,195659	0,483078	0,280510
-0,00947339	0,442324	-0,225618	-0,245938
-0,0923522	-0,0144454	0,616187	-0,0972177
-0,150170	0,0768423	0,291822	0,134629
-0,340012	-0,0523432	-0,246579	0,0960147
-0,360511	0,129299	-0,211243	0,172850
-0,179479	-0,0145974	0,616101	0,246657

Gambar 40 Empat kolom terakhir pada dataset (setelah proses casting)

Ada satu kolom subject + 19 aktivitas x 45 sensor = 856 kolom pada dataset tersebut. Kemudian ada delapan baris pada dataset tersebut, yang sesuai dengan banyak subjek. Kolom T_xacc_a01 artinya data T_axacc untuk aktivitas A01, kolom T_xacc_a02 artinya data T_xacc untuk aktivitas A02, dan seterusnya.

2.3. Feature Scaling

Pada tahap ini kami melakukan scaling pada dataset menggunakan metode Min Max Scaler.

```
18  df <- minmaxScaling(df)
19  df <- df$scaledDataSet
```

Gambar 41 Kode untuk melakukan scaling

	subject	T_xacc_a01	T_xacc_a02	T_xacc_a03	T_xacc_a04	T_xacc_a05
1	p1	0,341330	0,00000	0,00000	0,746487	0,118388
2	p2	1,00000	0,811083	0,910941	0,467475	1,00000
3	p3	0,766367	0,836150	0,432908	1,00000	0,943722
4	p4	0,835263	0,567640	0,573967	0,618294	0,842255
5	p5	0,00000	0,429237	0,290414	0,347357	0,680520
6	p6	0,428750	0,505652	0,0289189	0,00000	0,00000
7	p7	0,817426	0,131835	0,0598758	0,489565	0,865608
8	p8	0,278946	1,00000	1,00000	0,776921	0,949289

Gambar 42 Hasil scaling

2.4. Proses Clustering

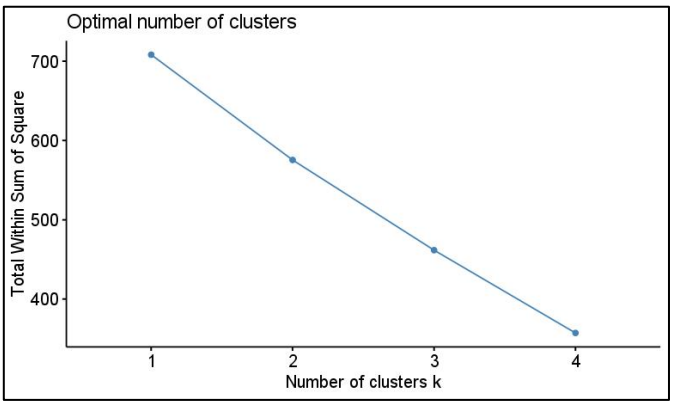
Setelah dilakukan scaling, dataset ini akan dijadikan dataset untuk proses clustering. Clustering yang dilakukan menggunakan metode K-Means Clustering karena pada dataset ini terdapat banyak kolom (856 kolom). Sebelum melakukan clustering, kami mencari nilai k optimal terlebih dahulu menggunakan metode Elbow.

```
library(factoextra)
library(cluster)

df <- read.csv( file: "res_pre_3/sport_activity.csv")
rownames(df) <- df$subject
df$subject <- NULL
fviz_nbclust(df, kmeans, method = "wss", k.max = 4)
```

Gambar 43 Kode untuk mencari k yang paling optimal

Kami menggunakan fungsi fviz_nbclust() dari library factoextra untuk mencari nilai optimal. Kami menggunakan metode elbow dengan mencari jumlah SSE (Sum Squared Error) dari semua cluster. Perhitungan dimulai dari k=1 sampai k=4 karena hanya ada 8 baris pada dataset. Berikut hasilnya:



Gambar 44 Perhitungan SSE dengan metode Elbow

Dari gambar 44, terlihat garis mulai lurus (stabil) dari k=2 sampai k=4. Jadi kami menggunakan nilai k=2 untuk melakukan clustering pada dataset ini.

```
# perform k-means
km <- kmeans(df, centers = 2, nstart = 1)
km
```

Gambar 45 Kode untuk menjalankan KMeans

```
Clustering vector:
p1 p2 p3 p4 p5 p6 p7 p8
 1  1  2  2  2  1  1  2
```

Gambar 46 Hasil clustering

Dari hasil clustering, subjek P1, P2, P6 dan P7 dimasukkan ke dalam cluster 1, sedangkan subjek P3, P4, P5 dan P8 dimasukkan ke dalam cluster 2.

```
Within cluster sum of squares by cluster:
[1] 276.2055 299.2517
```

Gambar 47 Nilai SSE dari setiap cluster

Cluster 1 memiliki nilai SSE sebesar 276,2055, sedangkan nilai SSE cluster 2 sebesar 299, 2517. Total SSE dari kedua cluster adalah sebesar 575, 4572.

DAFTAR PUSTAKA

[1] Karami, A., Gangopadhyay, A., Zhou, B., & Kharrazi, H. (2017). Fuzzy approach topic discovery in health and medical corpora. International Journal of Fuzzy Systems, 1-12.

[2] K. Altun, B. Barshan, and O. Tunçel, ``Comparative study on classifying human activities with miniature inertial and magnetic sensors," Pattern Recognition, 43(10):3605-3620, October 2010.

[3] B. Barshan and M. C. Yüksek, ``Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units," The Computer Journal, 57(11):1649--1667, November 2014.

[4] K. Altun and B. Barshan, ``Human activity recognition using inertial/magnetic sensor units," Proceedings First International Workshop on Human Behavior Understanding (in conjunction with the 20th Int. Conf. on Pattern Recognition), 22 August 2010, Istanbul, Turkey, A. A. Salah, T. Gevers, N. Sebe, A. Vinciarelli (editors), HBU 2010, LNCS 6219, pp.38-51, Springer: Berlin, Heidelberg, 2010.