

Tugas 06: LAPORAN PRAKTIKUM

Social Media Mining

Anggota:

- 162012133057 - Kevin Caesar
- 162012133066 - Riyanna Shabrina
- 162012133068 - Nicholas Juan Calvin Pradiptamurty
- 162012133075 - Angelene Verlyn Soeseno

Latar belakang data:

pada akhir-akhir ini, twitter diramaikan dengan trending topik mengenai perusahaan minuman “Es Teh Indonesia” yang berkonflik dengan seorang pengguna Twitter. Kasus ini lalu diviralkan lebih lanjut oleh lebih banyak pengguna Twitter lainnya yang ikut mengkritik tindakan “Es Teh Indonesia” yang dianggap berlebihan karena pihak “Es Teh Indonesia” yang membawa tim legalnya untuk menanggapi kritik tersebut. Hashtag ini kami pilih sebagai objek penelitian karena banyak pihak yang berinteraksi satu sama lain..

1. SNSCRAPE

```
In [5]: import snsrape
import pandas as pd
import networkx as nx
import re

In [6]: df = pd.read_json('esteh.json', lines=True)
df.columns

Out[6]: Index(['_type', 'url', 'date', 'content', 'renderedContent', 'id', 'user',
'replyCount', 'retweetCount', 'likeCount', 'quoteCount',
'conversationId', 'lang', 'source', 'sourceUrl', 'sourceLabel',
'outlinks', 'tcooutlinks', 'media', 'retweetedTweet', 'quotedTweet',
'inReplyToTweetId', 'inReplyToUser', 'mentionedUsers', 'coordinates',
'place', 'hashtags', 'cashtags'],
dtype='object')
```

```
scrap = df.drop(['_type', 'url', 'date', 'renderedContent', 'id',
               'replyCount', 'retweetCount', 'likeCount', 'quoteCount',
               'conversationId', 'lang', 'source', 'sourceUrl', 'sourceLabel',
               'outlinks', 'tcooutlinks', 'media', 'retweetedTweet', 'quotedTweet',
               'inReplyToTweetId', 'inReplyToUser', 'mentionedUsers', 'coordinates',
               'place', 'hashtags', 'cashtags'], axis=1)
scrap
```

	content	user
0	Missing the same person, everyday\n\n#estehind...	{'_type': 'snsrape.modules.twitter.User', 'us...
1	#luhutbinsarpanjaitan #AirlanggaHartarto #Joko...	{'_type': 'snsrape.modules.twitter.User', 'us...
2	Jualan itu kudu siap dikritik, bukan ngasih so...	{'_type': 'snsrape.modules.twitter.User', 'us...
3	Kisruh Es Teh Indonesia, ini kata Ernest Praka...	{'_type': 'snsrape.modules.twitter.User', 'us...
4	Fak estehendonesa \n\nMy hommies drink VONTEA (...)	{'_type': 'snsrape.modules.twitter.User', 'us...
...
618	Warga Bekaseee jangan lupa hari ini yaaa... 🍷...	{'_type': 'snsrape.modules.twitter.User', 'us...
619	Ada #kopiindonesia, ada #estehindonesia luar b...	{'_type': 'snsrape.modules.twitter.User', 'us...
620	Sekopi kopinya kamu pasti minum es teh juga #e...	{'_type': 'snsrape.modules.twitter.User', 'us...
621	Sekopi-kopinya kamu. Es teh manis adalah kita ...	{'_type': 'snsrape.modules.twitter.User', 'us...
622	Once more congratulation beb @jodiebianca atas...	{'_type': 'snsrape.modules.twitter.User', 'us...

Pada bagian ini kita scrapping data pada twitter dengan topik 'es teh Indonesia' yang mendapatkan kritik konsumen karena produk yang terlalu manis. pada bagian ini kita mengambil data pada content dan user dan mendrop kolom yang tidak diperlukan dan disimpan di scrap

```
listuser = []
for row in scrap['user']:
    listuser.append('@'+row['username'])
scrap['user'] = listuser
scrap
```

	content	user
0	Missing the same person, everyday\n\n#estehind...	@p_andalusia
1	#luhutbinsarpanjaitan #AirlanggaHartarto #Joko...	@TeknokratJawa
2	Jualan itu kudu siap dikritik, bukan ngasih so...	@btdrpengusaha
3	Kisruh Es Teh Indonesia, ini kata Ernest Praka...	@Beritasatu
4	Fak estehendonesa \n\nMy hommies drink VONTEA (...)	@YudhaSKAlatiga
...
618	Warga Bekaseee jangan lupa hari ini yaaa... 🍷...	@fatinfabri
619	Ada #kopiindonesia, ada #estehindonesia luar b...	@baimjalu
620	Sekopi kopinya kamu pasti minum es teh juga #e...	@kurniaw05441616
621	Sekopi-kopinya kamu. Es teh manis adalah kita ...	@bundaorizaa
622	Once more congratulation beb @jodiebianca atas...	@rezaalle

623 rows × 2 columns

```
contentdf = scrap.copy()
contentdf = contentdf.drop(['user'], axis=1)
contentdf
```

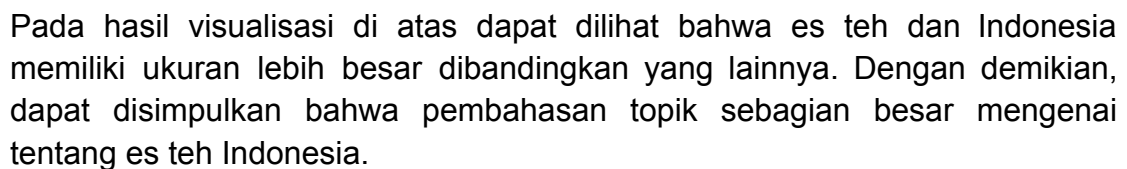
	content
0	Missing the same person, everyday\n\n#estehind...
1	#luhutbinsarpanjaitan #AirlanggaHartarto #Joko...
2	Jualan itu kudu siap dikritik, bukan ngasih so...
3	Kisruh Es Teh Indonesia, ini kata Ernest Praka...
4	Fak estehendonesa \n\nMy hommies drink VONTEA (...)
...	...
618	Warga Bekaseee jangan lupa hari ini yaaa... 🍷...
619	Ada #kopiindonesia, ada #estehindonesia luar b...
620	Sekopi kopinya kamu pasti minum es teh juga #e...
621	Sekopi-kopinya kamu. Es teh manis adalah kita ...
622	Once more congratulation beb @jodiebianca atas...

623 rows × 1 columns

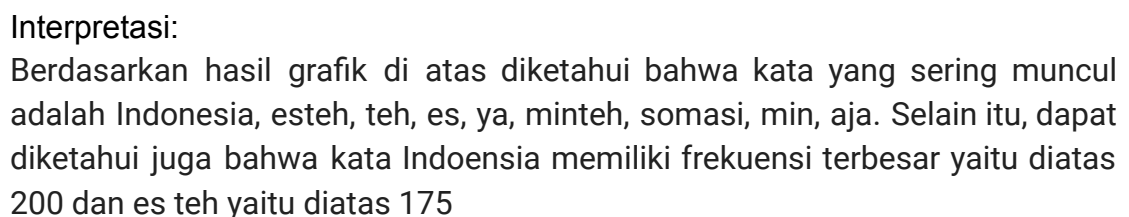
Kemudian, membuat variabel kosong untuk diisi user ditambah @ maka hasil akan seperti yang diatas dan kita mengambil content dan disimpan di contentdf maka hasil seperti diatas.

2. WORDCLOUD

```
# create wordcloud
wordcloud = WordCloud(width = 800, height = 800,
                       background_color = 'white',
                       min_font_size = 10).generate(out)
# plot the WordCloud image
plt.figure(figsize = (8, 8), facecolor = None)
plt.imshow(wordcloud)
plt.axis("off")
plt.tight_layout(pad = 0)
plt.show()
```



```
[12] # mencari frekuensi kata
df['DistText'] = df['content'].apply(preprocess_text) # process preprocessing text with function
fdist_title = FreqDist(sum(df['DistText'], [])) # count frequ
most_fdist_title = dict(fdist_title.most_common(10))
#visualize frekuensi kata
df_freq_tokens_title = pd.DataFrame.from_dict(most_fdist_title, orient='index')
df_freq_tokens_title.columns = ['Frequency']
df_freq_tokens_title.index.name = 'Key'
df_freq_tokens_title.plot(kind='bar')
```



clustering TF-IDF

```
[ ] #initialize the vectorizer
vectorizer = TfidfVectorizer(sublinear_tf=True, min_df = 5, max_df = 0.95)
# fit_transform applies TF-IDF to clean texts - we save the array of vector in X
X = vectorizer.fit_transform(df['CleanTwit'])
# initialize kmeans with 3 centroid
kmeans = KMeans(n_clusters=3, random_state=42)
# fit the model
kmeans.fit(X)
# store cluster label in a variable
cluster = kmeans.labels_
```

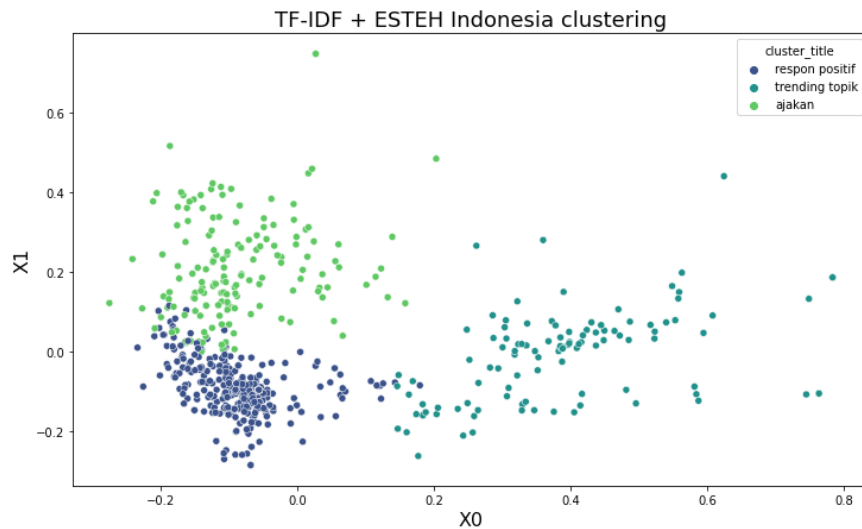
PCA

```
[ ] # initialize PCA with 2 components
pca = PCA(n_components= 2, random_state=2022)
# pass our X to the PCA and store the reduced vector into pca_vecs
pca_vecs = pca.fit_transform(X.toarray())
# save our dimensions into x0 and x1
x0 = pca_vecs[:, 0]
x1 = pca_vecs[:, 1]
# assign clusters and pca vectors to our dataframe
df['cluster_title'] = cluster
df['x0_title'] = x0
df['x1_title'] = x1
```

VISUALISASI CLUSTER

```
[ ] # map clusters to appropriate labels
cluster_map = {0: "ajakan", 1: "respon positif", 2: "trending topik"}
# apply mapping
df['cluster_title'] = df['cluster_title'].map(cluster_map)
```

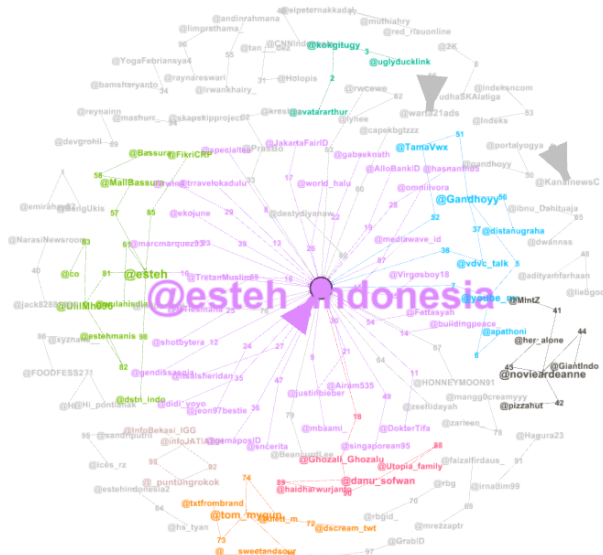
```
[ ] # set image size
plt.figure(figsize=(12,7))
# set a title
plt.title("TF-IDF + ESTEH Indonesia clustering", fontdict={"fontsize":18})
# set axes name
plt.xlabel("X0", fontdict={"fontsize":16})
plt.ylabel("X1", fontdict={"fontsize":16})
# create a scatter plot with seaborn, where is hue is the class used to group the data
sns.scatterplot(data=df, x='x0_title', y='x1_title', hue='cluster_title', palette='viridis')
plt.show()
```



Berdasarkan grafik scatter plot di atas dapat diketahui bahwa data tersebut dikelompokkan menjadi 3 cluster yaitu respon positif, trending topik, ajakan. "respon positif" di sebelah kiri bawah, "trending topik" terletak di sebelah kanan bawah dan menyebar, dan "ajakan" terletak di kiri atas

4. VISUALISASI NETWORK

Dengan menggunakan tipe layout Fruchterman Reingold, terlihat bahwa akun @esteh_indonesia sebagai pusat tweet yang banyak dibicarakan.



Tipe layout ForceAtlas 2 pada Gephi menunjukkan bahwa terbagi menjadi lima cluster (dapat dilihat terbagi menjadi 5 warna utama, ungu, hijau, pink, biru, dan abu-abu). Lima cluster yang berkumpul di tengah ini menunjukkan akun-akun yang berpengaruh (paling banyak membahas) dalam topik pembahasan 'es teh'.



5. KESIMPULAN KUMPULAN TWEET/NETWORK



Tokoh utama (key influencer) dalam kumpulan tweet mengenai 'es teh Indonesia' adalah **@esteh_indonesia**. Untuk akun lain yang menjadi broker (penyalur diskusi) bagi mengenai topik utama yang dibahas adalah **@Gandhoyy** (sebagai broker/penyalur diskusi untuk cluster biru), akun **@esteh** (untuk cluster hijau), dan **@danu_softwan** (untuk cluster pink). Masing-masing broker ini sebagai penghubung diskusi dari key influencer, yaitu @esteh indonesia dan sirkelnya masing-masing.

