

Clustering Algorithm

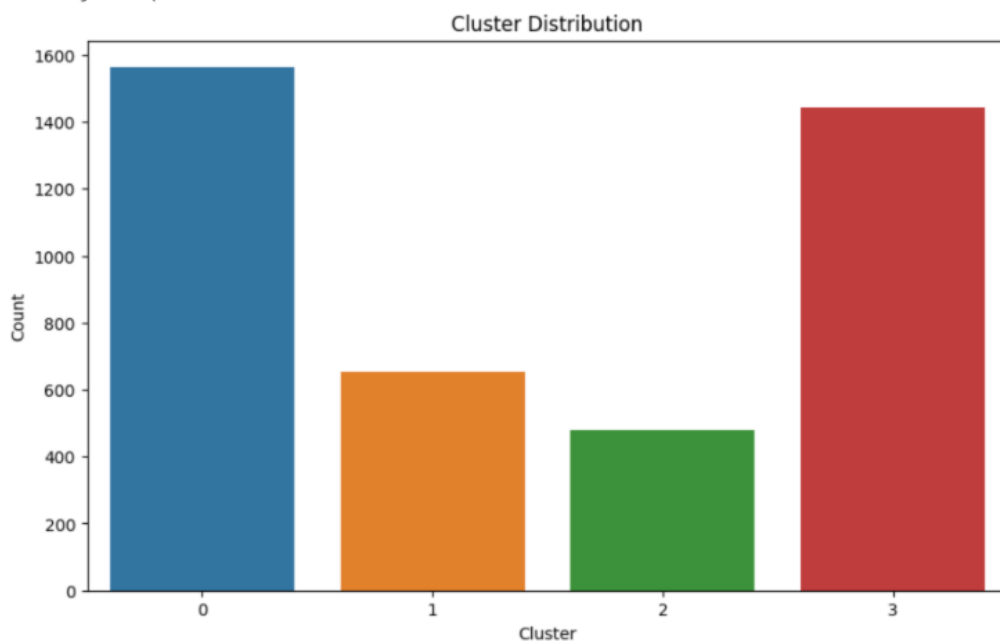
Using both Word2Vec and TF-IDF for text analysis, it was observed that Word2Vec outperformed TF-IDF in terms of performance.

Libraries:

Pandas version: 1.5.3, Regex (re) version: 2.2.1, spaCy version: 3.6.1, NLTK version: 3.8.1, Gensim version: 4.3.2, scikit-learn (sklearn) version: 1.2.2, Matplotlib version: 3.7.1, Seaborn version: 0.12.2

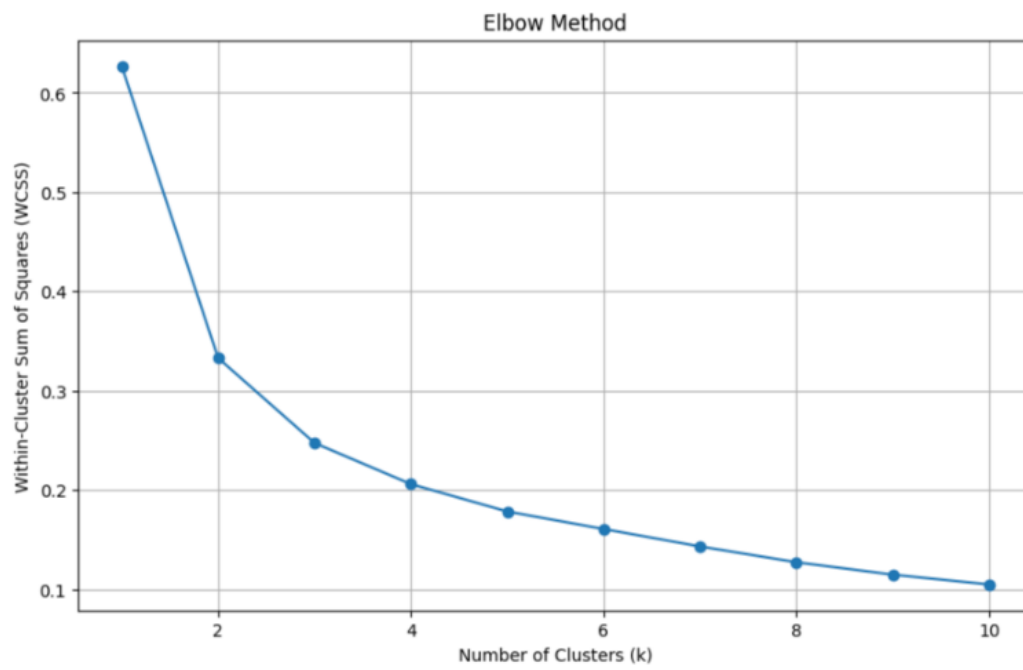
Execute file Clustering_lab_4.py

Clusters and keywords:



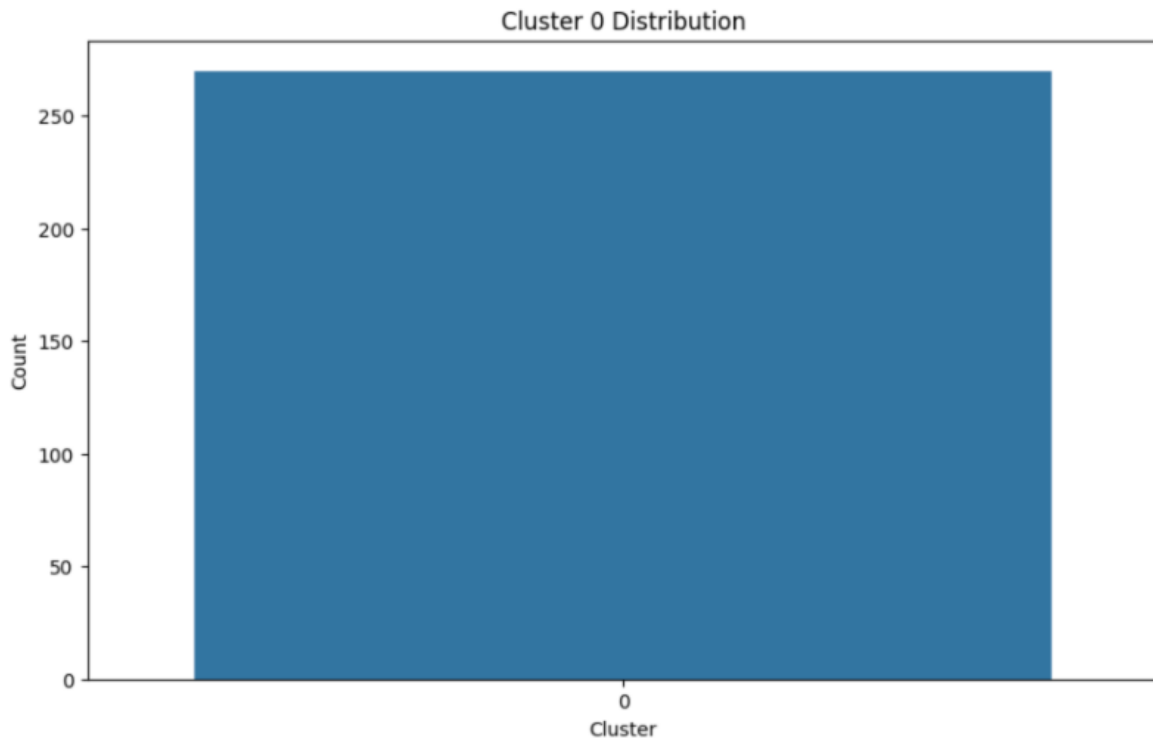
Cluster: 0 Keywords: ['deepfake, presidents, used, russia, ukraine, war', 'ukraine, war, accelerated, research, lithium, ion, batt
Cluster: 1 Keywords: ['remote, startups, win, war, top, talent', 'sweden, returns, cold, war, tactics, battle, fake, news', 'us, t
Cluster: 2 Keywords: ['clock, speed, wars, back, intel, brags, hitting, 6, ghz, 13th, gen, cpus', 'cryptomining, gangs, go, war, t
Cluster: 3 Keywords: ['cyber, war, anonymous, leaks, 776, gb, kremlin, files, claimed, hack, russia, 's, ministry, culture', 'anor

Elbow methods:



Predicting Cluster:

- lasers, capable, transmitting, signals, 224, gigabits, per, second, enough, achieve, 800, gigabit, ethernet
- google, start, field, testing, next, gen, ar, glasses, august, |, engadget
- covid, tracing, apps, taken, long
- managing, wireless, network, multiple, users
- routers
- breakthrough, year, passwordless, technology
- 280, year, old, algorithm, inside, google, trips
- google, hit, 100, percent, renewable, energy, year
- google, tracker, 2015, everything, know, google, working, new, year
- get, ready, school, year, google



Automation

To do the automation task, please enter following command in cmd:

```
python3 main.py
```

Then the script will start srcaping data, preprocessing and clustering automatically To set the time interval between two process, you can use attributes --minutes and --seconds. And to set the num of post in each downloading, you can use --num

```
python3 main.py --num 500 --minutes 1 --seconds 20
```

Then the wait time would be 1 min and 20 seconds, downloading 500 posts in each process

```
vboxuser@dsci-560-back:~/Desktop/Lab_4_pt-2$ ^C
vboxuser@dsci-560-back:~/Desktop/Lab_4_pt-2$ python3 main.py --num 500 --minutes
1 --seconds 10
<<<<<<< data collectiong process start >>>>>>>>>>
238 of 500 has been stored
328 of 500 has been stored
500 of 500 has been stored
Data fetching Spend time: 7.759594202041626
<<<<<<< data collectiong process end >>>>>>>>>>
<<<<<<< data preprocessing start >>>>>>>>>>
[nltk_data] Downloading package stopwords to
[nltk_data] /home/vboxuser/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
Input data shape: (889, 9)
<<<<<<< data preprocessing end >>>>>>>>>>
starting data storing -----
```

Title varchar	...	Keywords varchar	Topics varchar
Snap removes speed...	...	snap, removes, spe...	Snap, speed filter...
Japan breaks inter...	...	japan, breaks, int...	Japan, internet sp...
Speedcheck study f...	...	speedcheck, study, ...	Speedcheck study, ...

And you can enter "quit" to exit the script, the script would ask you the question about finding the cluster that matches closest, you can type any information like "Hello world" to get the result. The result would like the screenshot in section Predicting Cluster

```

SpaceX is now one ...      ...      [0.001433719910733...      0
Inside Facebook's ...      ...      [0.000466619560029...      8
Microsoft Exchange...      ...      [-0.00020995819068...      2
Company That Route...      ...      [-0.00022229894238...      2
Zuckerberg on why ...      ...      [-0.00231732218526...      6
889 rows (20 shown)      11 columns (4 shown)
Waiting for next process...
quit
Quit commend received, waiting for response
Please enter a message or keywords to Cluster:Hello world!

```