Contributors : Dylan Kakkanad , Mahika Bhartari, Rajashree Ramaprabu, Sahasra Konkala

Exploratory Data Analysis

As part of our initial EDA,
To analyze the news articles dataset, a dataset of approximately 2.7 million rows. From this extensive dataset, we employed the stratified random sampling method to derive a representative subset of approximately 27,000 rows (1.0). Followed by analyzing the data structure of our data frame and then moved to check for null, duplicate, and invalid values. Duplicate and null values present in the article column are removed. We explored the number of articles published and the average count of words used by various publishers and depicted it using bar plots(1.1). We have also visualized the number of articles available in distinct sections(1.3) published annually. We have created a word cloud model(1.5) to showcase the most frequent words in the title, and article columns. In our further analysis and processing, we will observe how the recurring terms evolve in the word cloud.

Pre-processing

Preprocessing mainly focussed on cleaning and preparing the ‘article’ column. The new processed column is called ‘processed_article’(2.1). To reach this stage we first cleaned the articles by removing hyperlinks, numbers, punctuation, and non-English characters found during EDA. Since our goal is to cluster news articles based on the subject matter in the article, numbers (not in words) would be redundant or, worst case, bias the clusters and were subsequently dropped. We also used word tokenization and filtered for stemmed English words within the nltk english corpus excluding stopwords. The number of words was thus reduced to 3,194.

Analysis Plan

Our primary objective was to cluster news articles effectively and find distinct clusters. Given the lack of a well-defined metric, we adopted an iterative approach to achieve this goal. The first step in our methodology was to reduce the dimensionality of our dataset, which initially had over 23,000 features. We employed Principal Component Analysis (PCA) for this purpose. PCA allowed us to identify and retain the top 1,000 principal components that accounted for approximately 50% of the variance in our data (3.1). This reduction was significant as it simplified our dataset while preserving its essential structure.

With a more manageable set of features, we then applied the KMeans clustering algorithm. The choice of KMeans was motivated by its efficiency and simplicity, making it a suitable choice for our initial exploration of the data. To determine the optimal number of clusters, we utilized the elbow method (3.2) and silhouette plots (3.3). Since we were not able to infer the exact number of clusters using the elbow method, we opted for a silhouette plot that provided us with the optimal number of clusters to be used. Both these techniques provided us with a quantitative means to assess the quality of our clustering. Our analysis suggested that around 10 clusters were optimal for our dataset.

Once the clusters were formed, we conducted a thorough examination of each cluster. We identified the top 30 words with the highest TF-IDF scores in each cluster. This step was crucial as it provided us with a qualitative understanding of the content of our clusters and is the only metric to assess the clusters. Similarly we have also identified top 30 words clusterwise with n-gram scores to cross validate how the clustering model performs for different approaches. The resultant clustering was similar for both the methods when unigram technique was conducted on the data.

In conclusion, our choices of PCA for dimensionality reduction, KMeans for clustering, and the elbow method and silhouette plots for determining the optimal number of clusters were all justified based on the nature of our dataset and our objective of effectively clustering news articles. Our iterative approach allowed us to refine our methodology and achieve our goal.

Preliminary Results

The clarity of themes within each cluster was our evaluation metric. The presence of three unclear clusters suggests areas for improvement. To enhance our model, we considered removing less meaningful words and adjusting the n_grams range in TF-IDF vectorization. The first and fifth clusters, as well as the eighth, were not very well defined. The remaining clusters, however, showed distinct themes: crime news (second), federal government news (third), national politics (fourth), business finance (sixth), financial markets (seventh), sports news (ninth), and the stock market (tenth)(4.1). Our evaluation metric was the clarity of the themes within each cluster. Seven out of ten clusters had distinct themes with little overlap, indicating a successful initial clustering.

Adjusting the n_grams range in TF-IDF vectorization resulted in similar clusters. Removing ineffective words(‘said’, ‘also’ etc) helped find new clusters like healthcare. Despite some unclear clusters, our initial findings are promising. We have performed unigram with range(1,2) as of now. The cluster order wasn't the same as TF-IDF, however, it yielded pretty similar results(4.2) with few repetitive clusters contributing to finance and politics. Two of the resultant clusters are not clearly distinguishable to figure out the genre of it. Most of the clusters provided promising results for initial clustering with minimal overlaps.

Next Steps

- N-grams: We’ll include only trigrams and bigrams to discover new clusters.
- Stopwords: We’ll remove additional stopwords from our dataset to enhance cluster distinctiveness.
- POS Tagging and NER: We’ll use these to adjust weights for locations, organizations, and people for more semantic accuracy.
- Hierarchical clustering: Try hierarchical clustering to see if clusters are clearer using a dendrogram

Coding Contribution

All the members of the group have diligently contributed towards this phase. Rajashree & Mahika worked on cleaning the data, performing exploratory data analysis, and built a clustering model based on n-gram feature representation using KMeans. Dylan & Sahasra worked on pre-processing the data, PCA, KMeans, and hyperparameter tuning. The GitHub repository includes a ‘Misc’ (miscellaneous) folder with all the contributions.

A **kanban board** used for the project can be found at this link.

GitHub Project **News Article Text Analysis**

Appendix

Fig 1.0: Data Dictionary

Name	Description	Data Type
date	Date of article publication	str
year	Year of article publication	int
month	Month of article publication	float
day	Day of article publication	int
author	Article author	str
title	Article title	str
section	Section of the publication in which the article appeared	str
article	Article text	str
url	Article URL	str
publication	Name of the article publication	str

Fig 1.1: Number of Articles by Publication

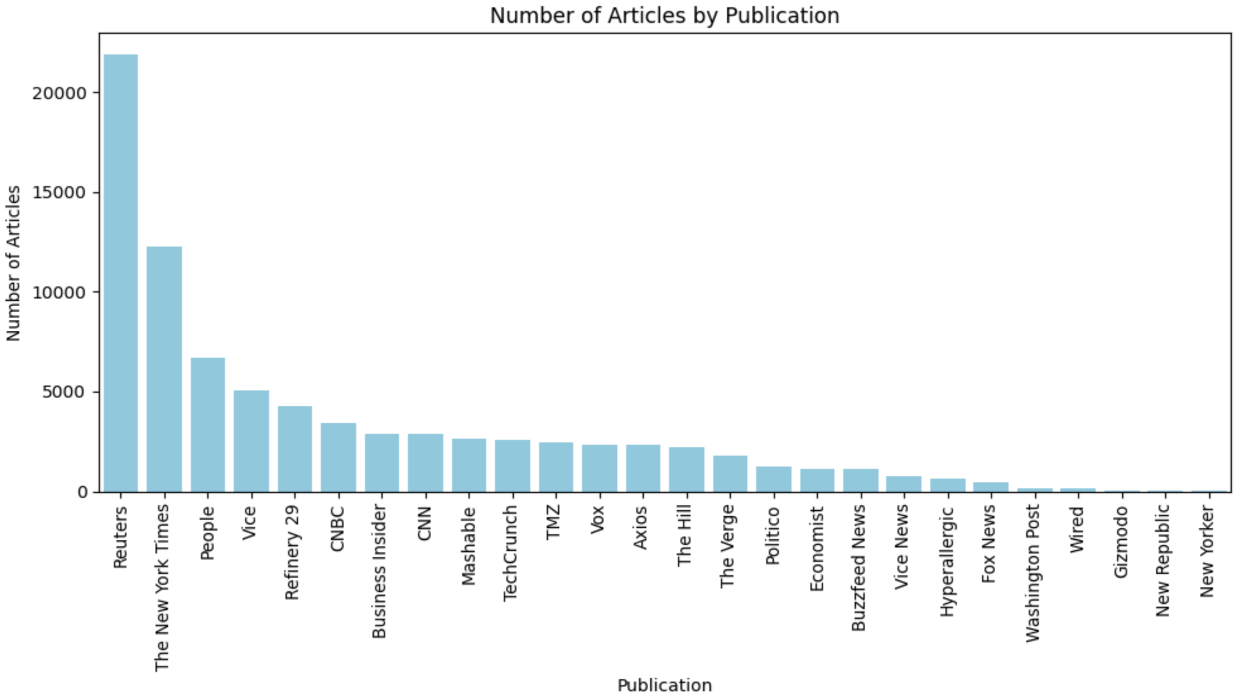


Fig 1.3: Distribution of Articles Across Sections

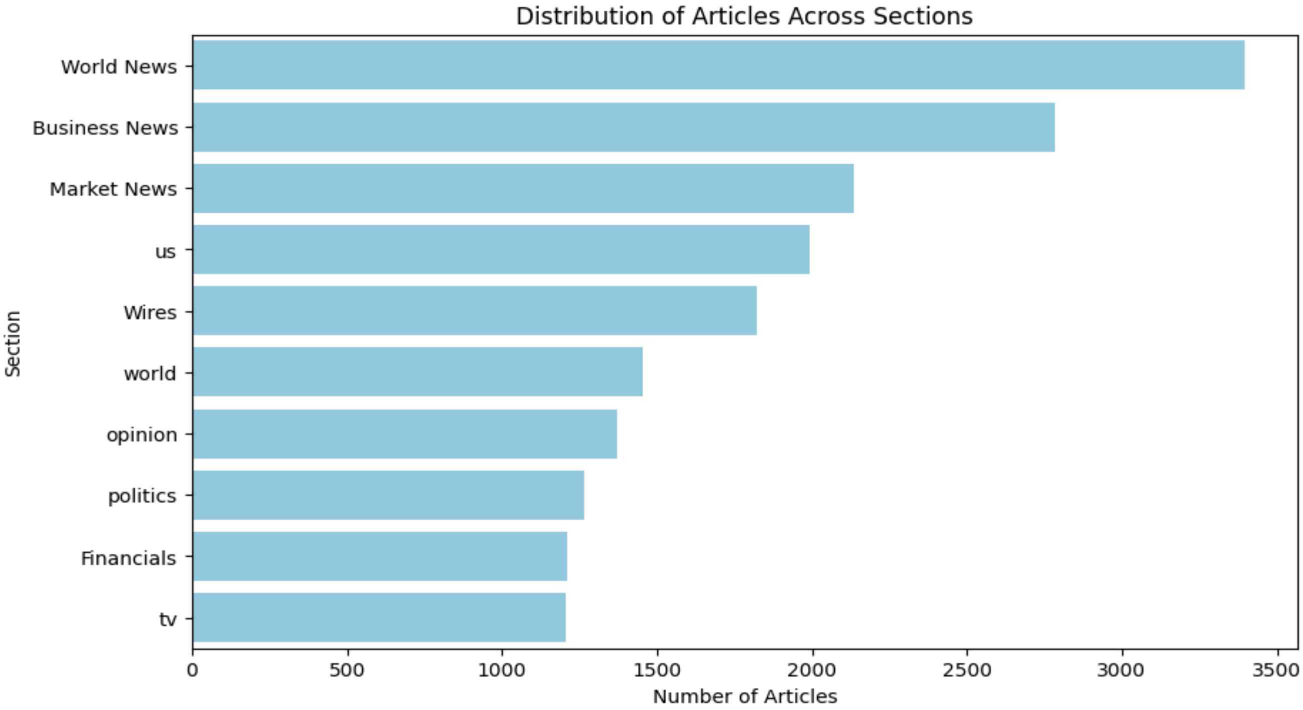


Fig 1.5: Word Cloud of the column ‘title’

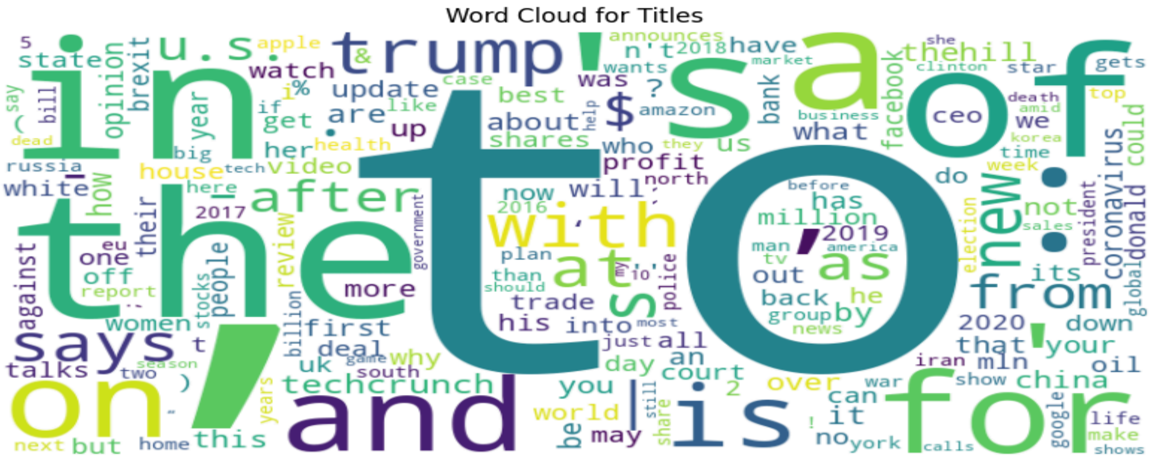
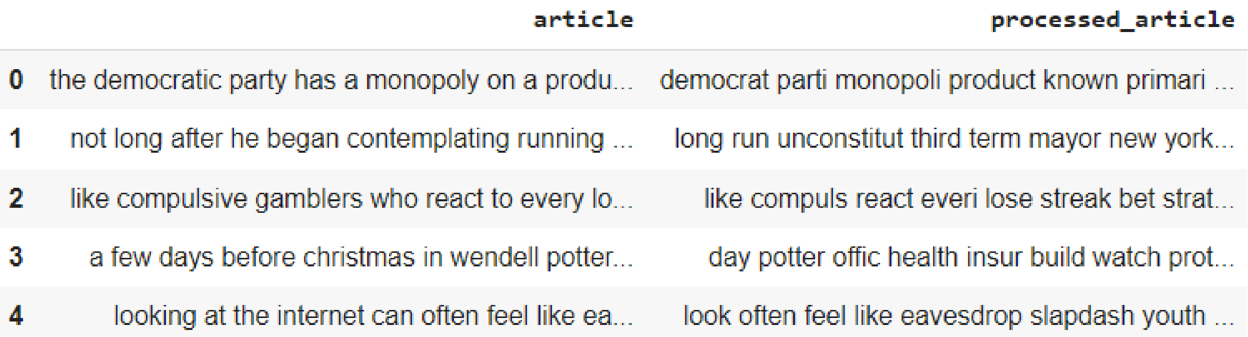


Fig 2.1: article column pre and post processing



	article	processed_article
0	the democratic party has a monopoly on a produ...	democrat parti monopoli product known primari ...
1	not long after he began contemplating running ...	long run unconstitut third term mayor new york...
2	like compulsive gamblers who react to every lo...	like compuls react everi lose streak bet strat...
3	a few days before christmas in wendell potter...	day potter offic health insur build watch prot...
4	looking at the internet can often feel like ea...	look often feel like eavesdrop slapdash youth ...

Fig 3.1 PCA:

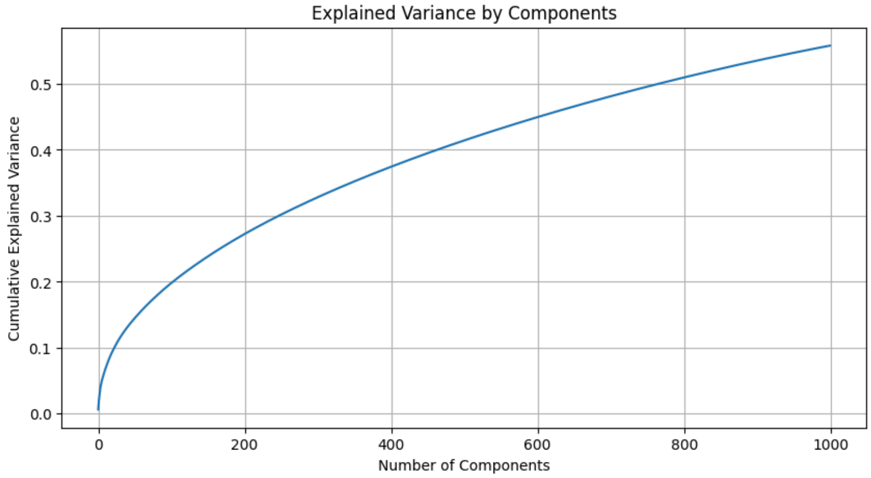


Fig 3.2 Elbow Method

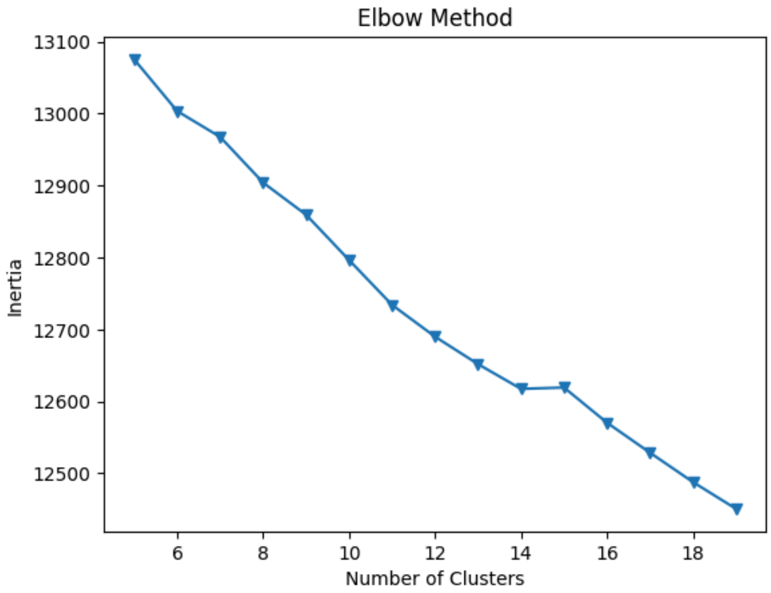


Fig 3.3 Silhouette Plot:

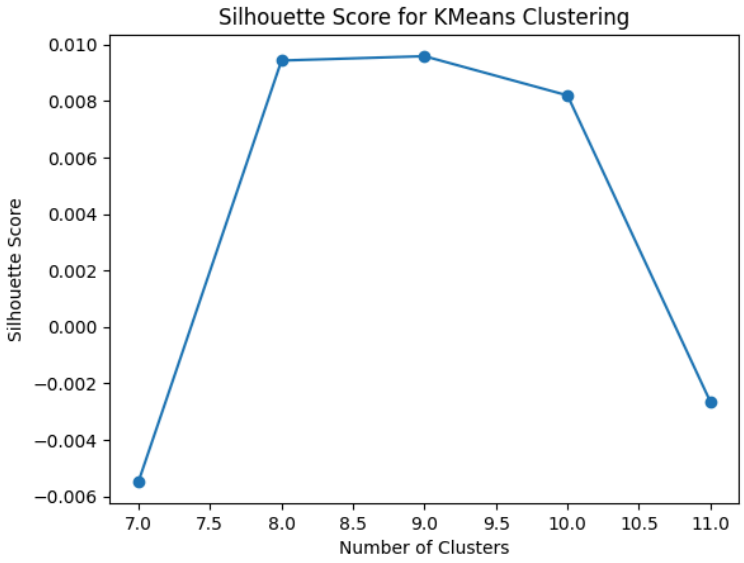
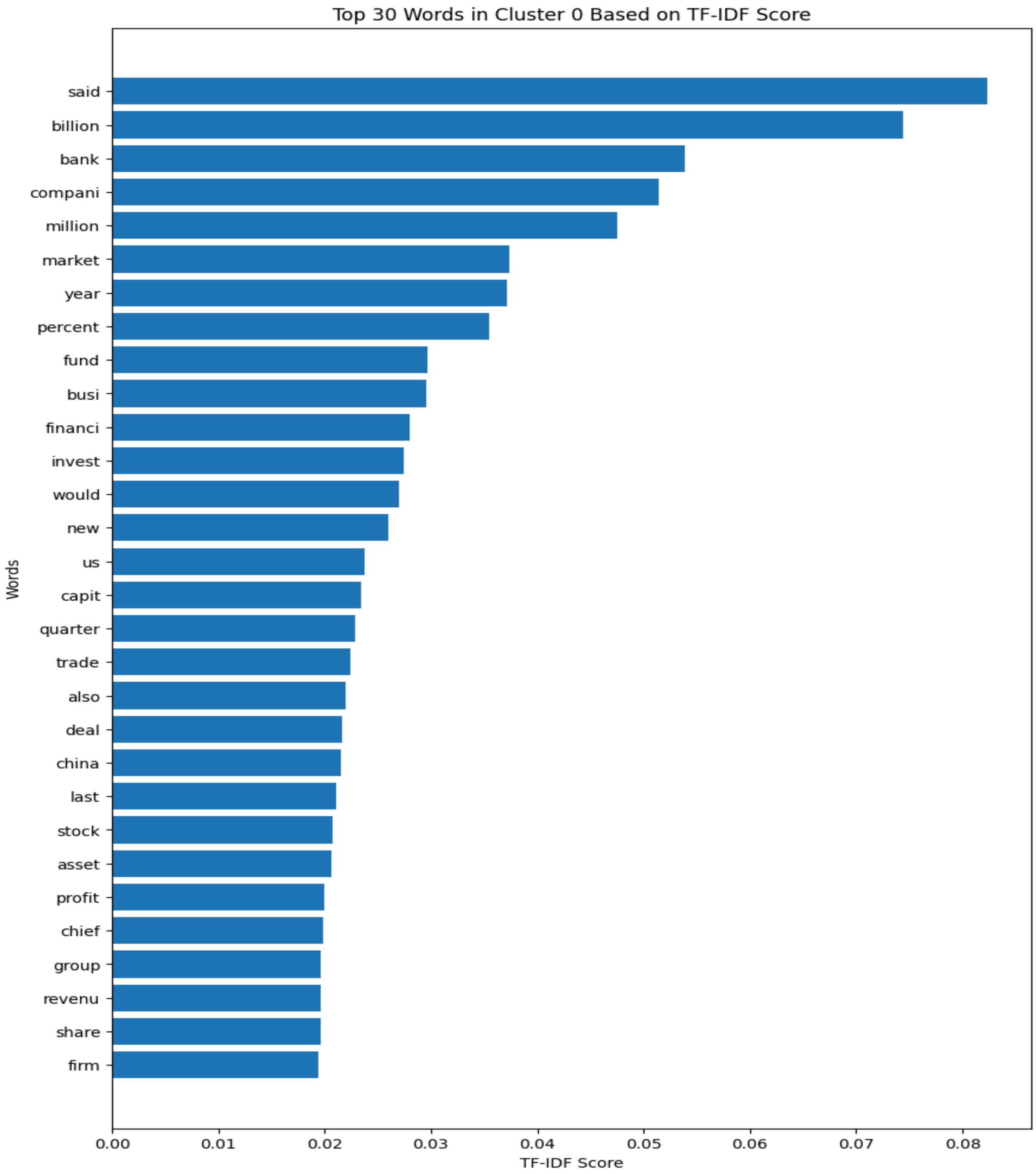
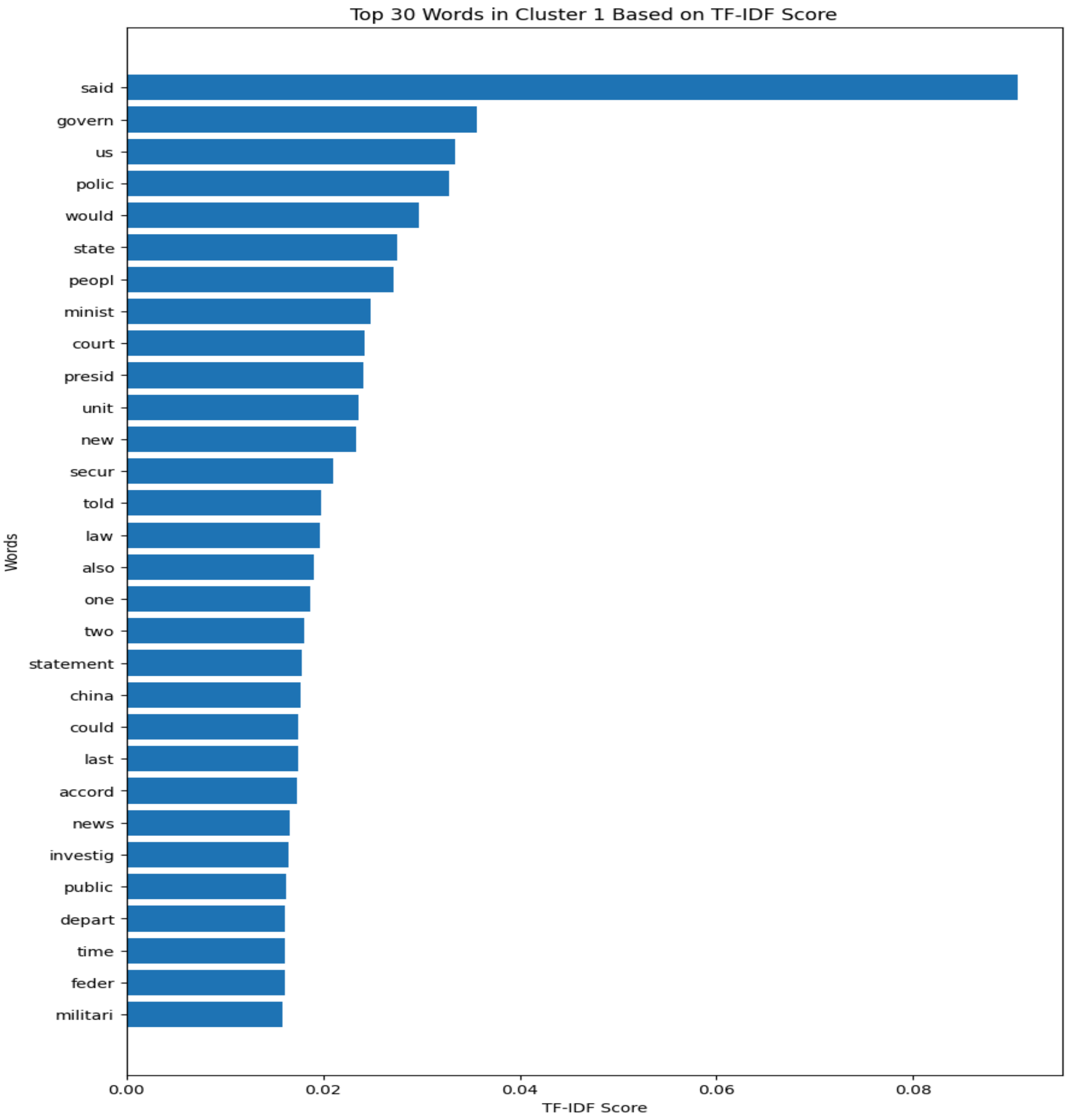
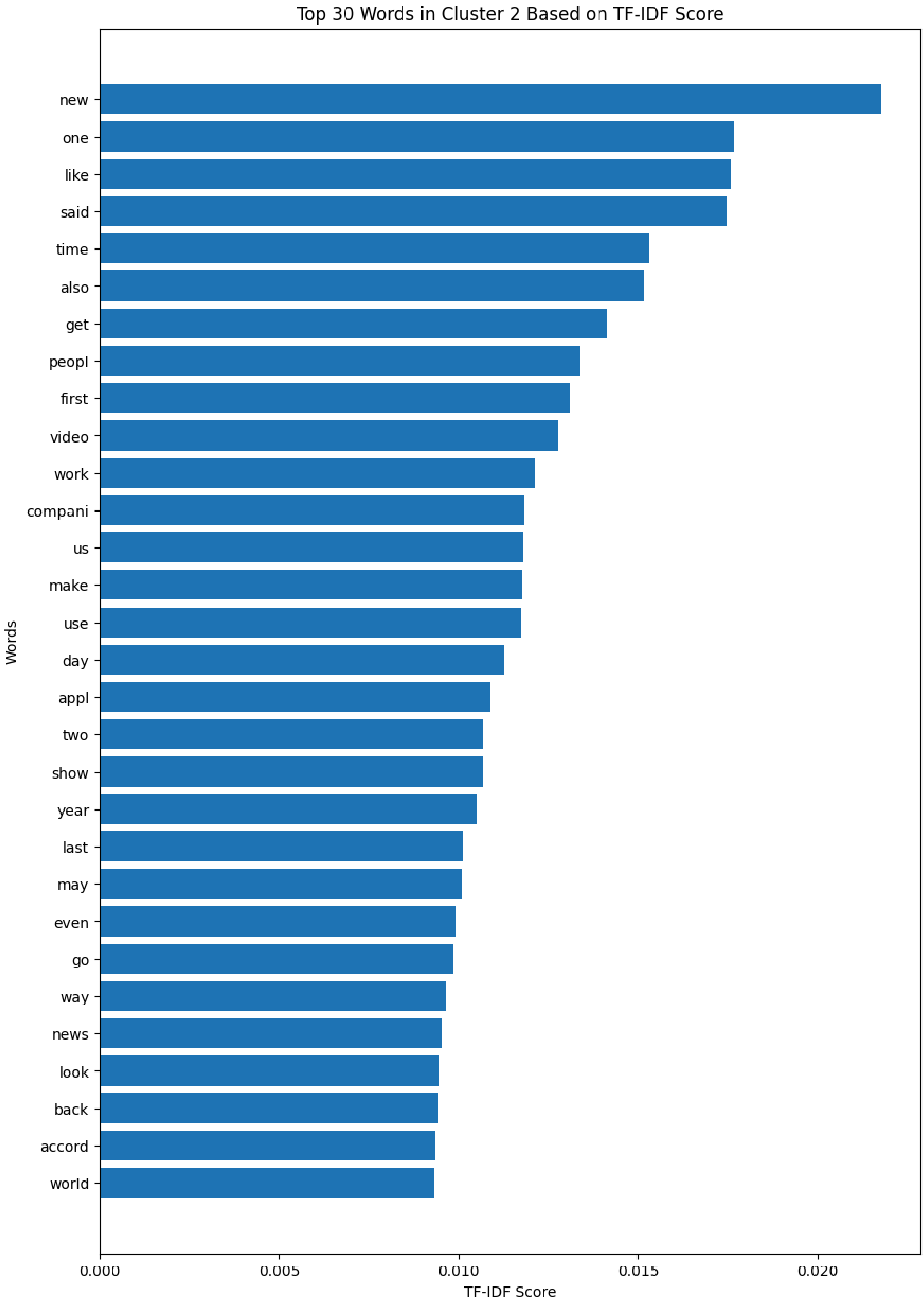
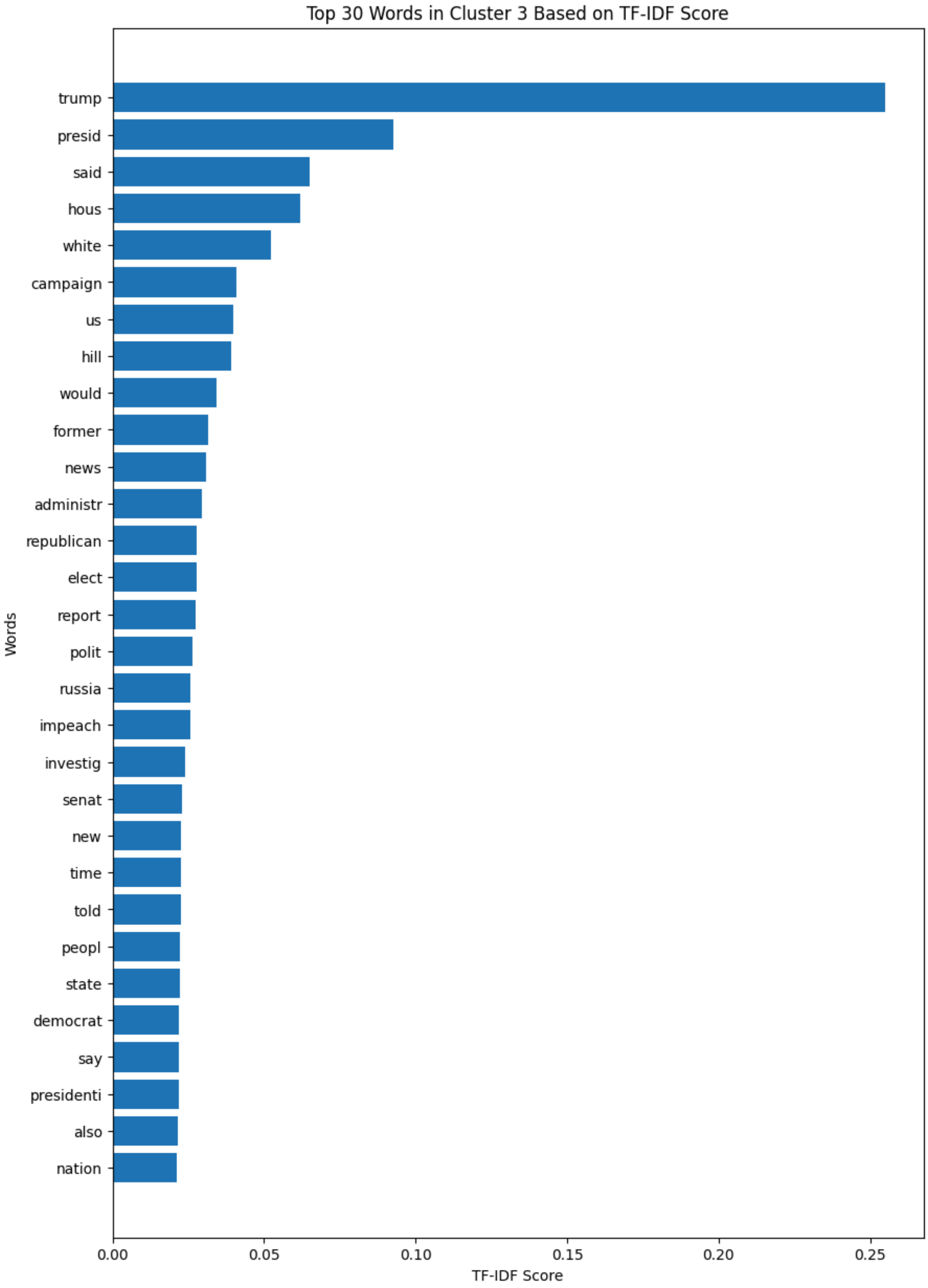


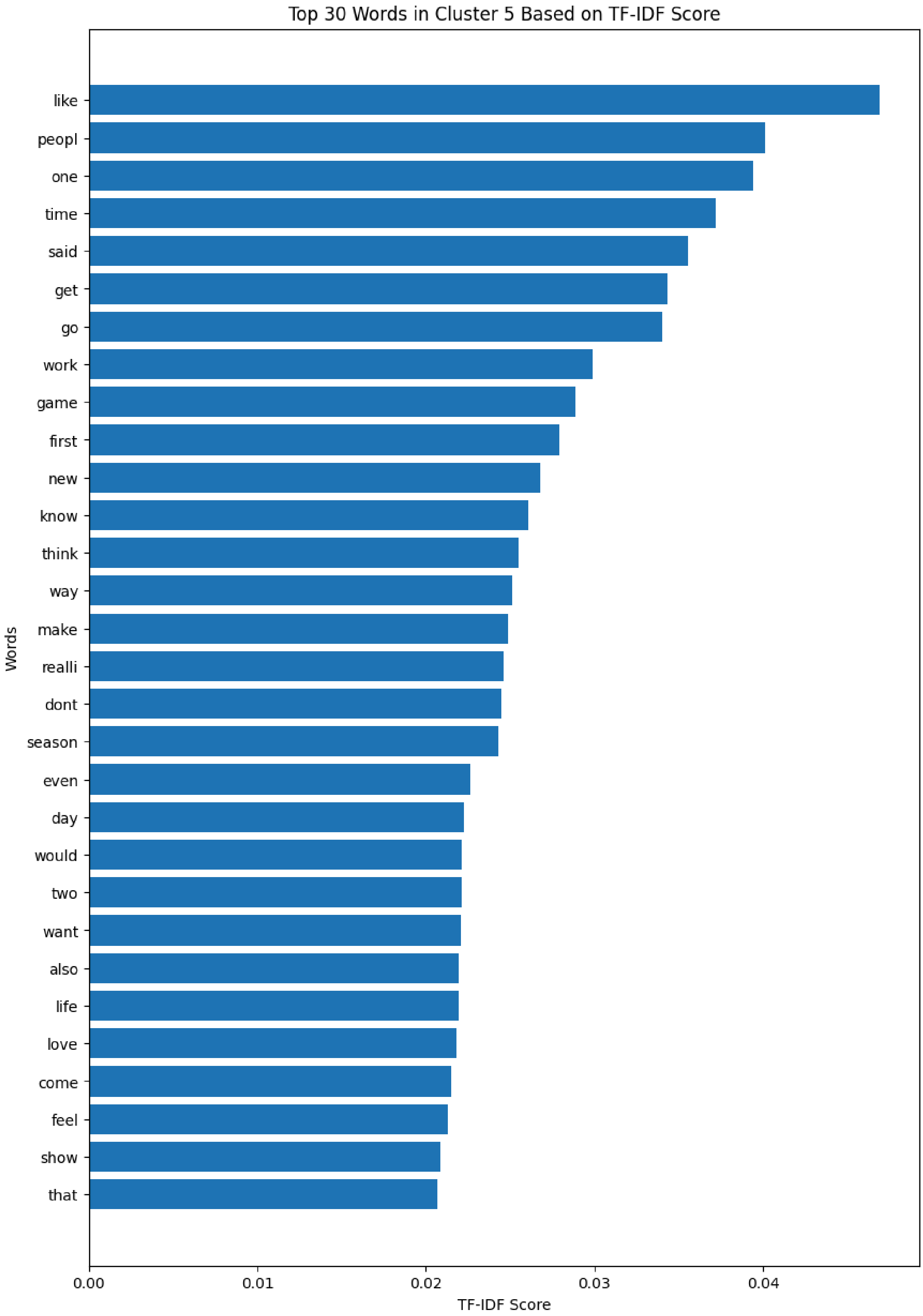
Fig 4.1 Clustering results through KMeans

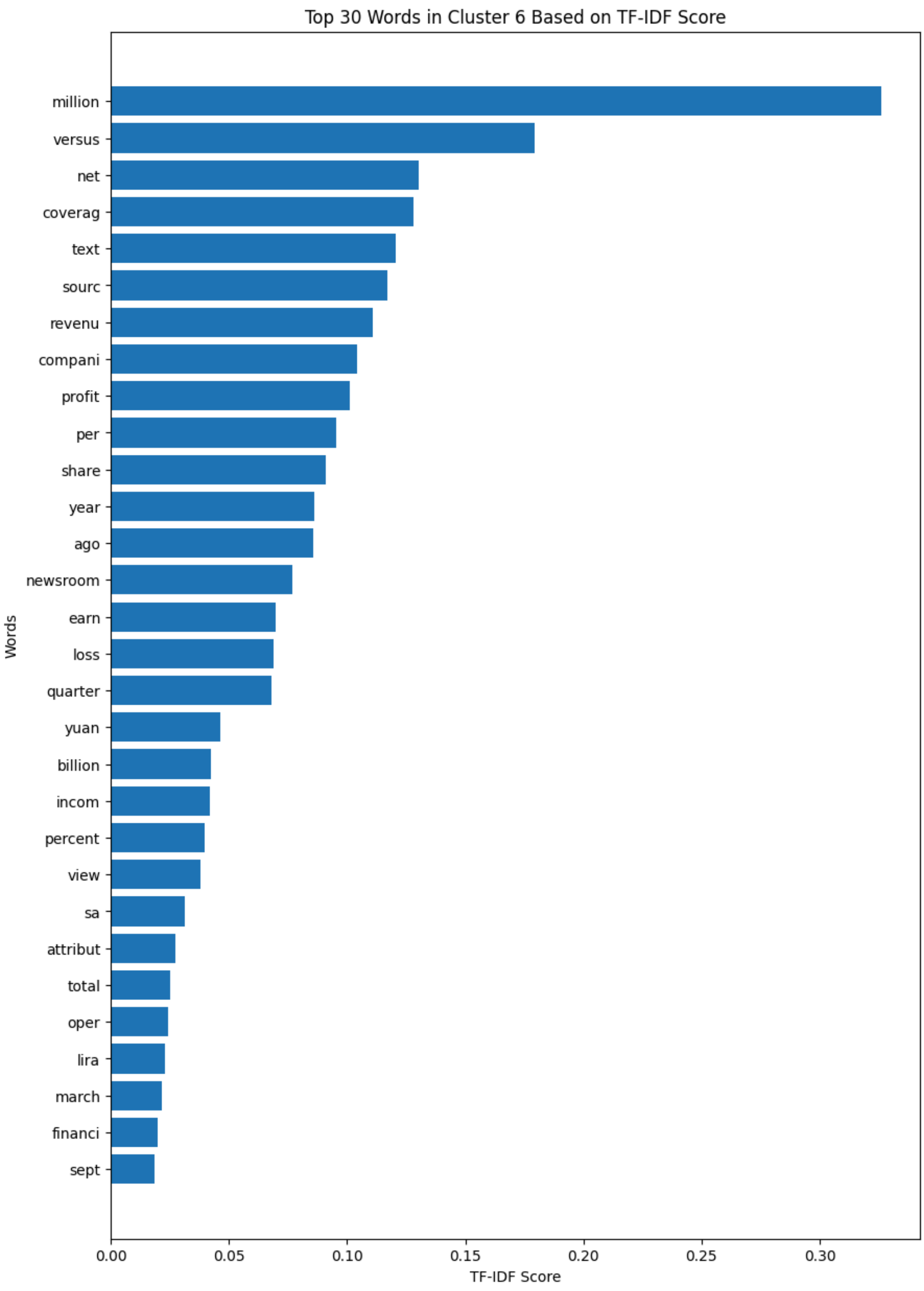


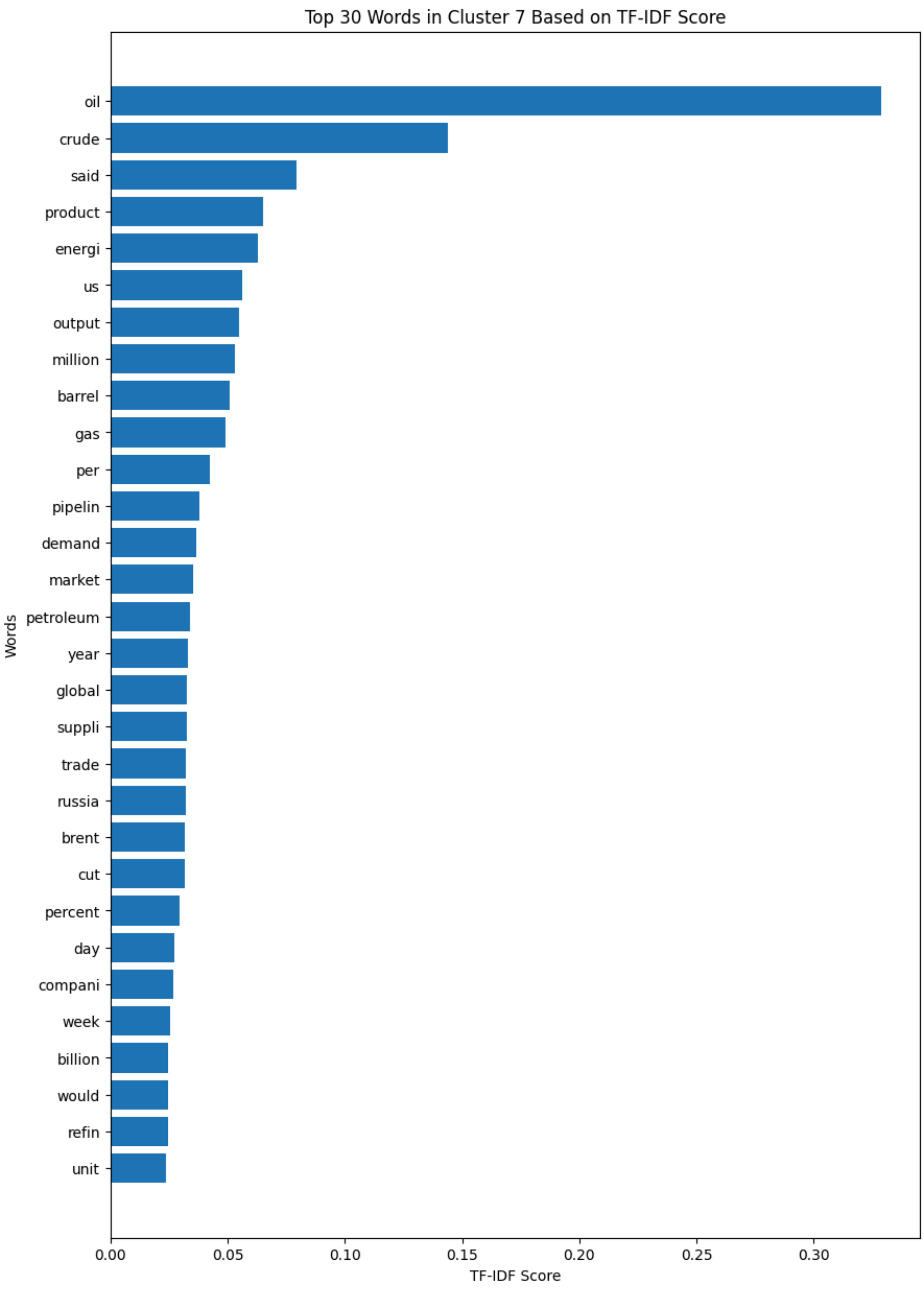

Top 30 Words in Cluster 1 Based on TF-IDF Score
Words
said
govern
us
polic
would
state
peopl
minist
court
presid
unit
new
secur
told
law
also
one
two
statement
china
could
last
accord
news
investig
public
depart
time
feder
militari
0.00
0.02
0.04
0.06
0.08
TF-IDF Score

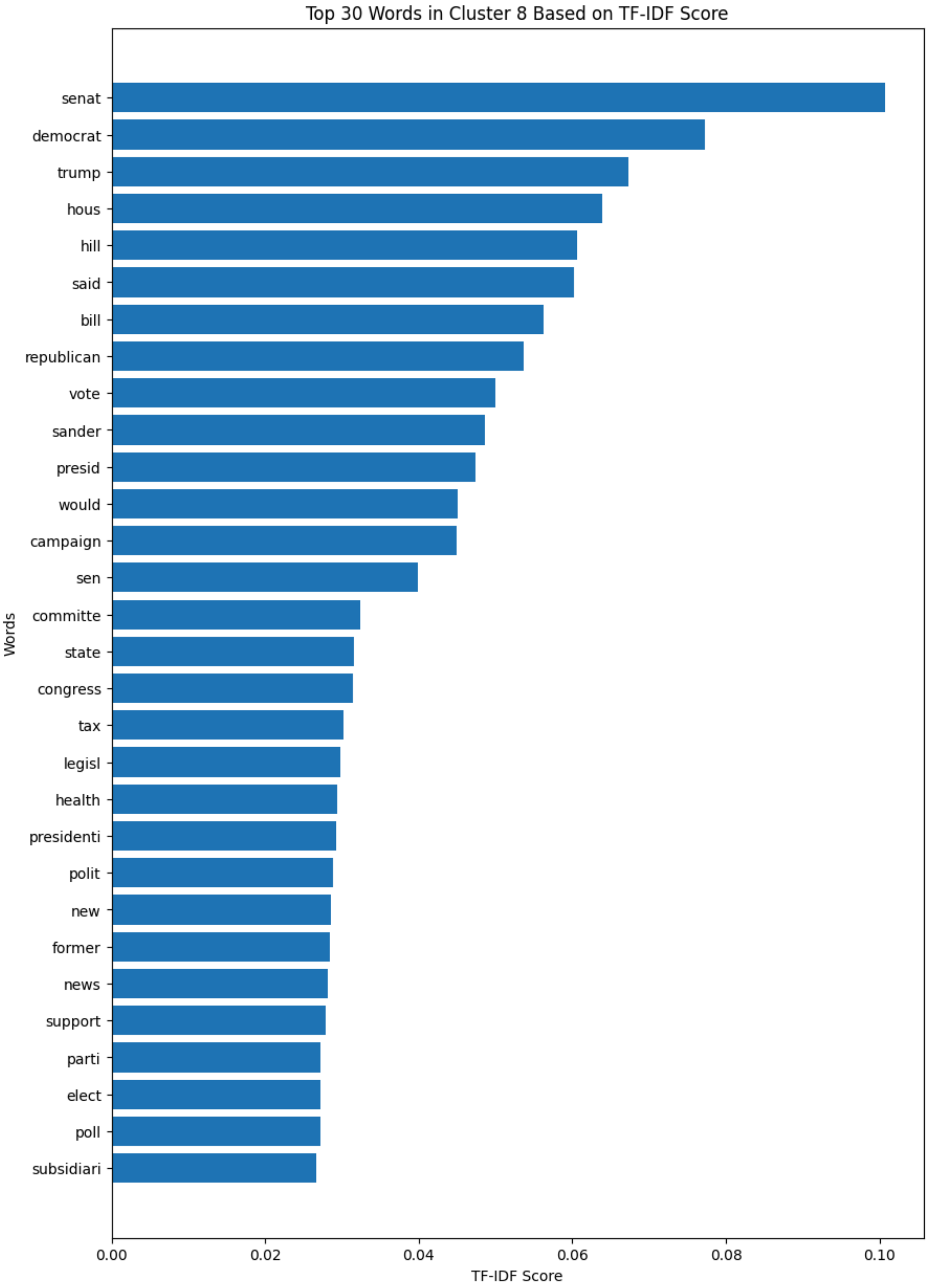

Top 30 Words in Cluster 2 Based on TF-IDF Score
Words
new
one
like
said
time
also
get
peopl
first
video
work
compani
us
make
use
day
appl
two
show
year
last
may
even
go
way
news
look
back
accord
world
0.000
0.005
0.010
0.015
0.020
TF-IDF Score


Top 30 Words in Cluster 3 Based on TF-IDF Score
Words
trump
presid
said
hous
white
campaign
us
hill
would
former
news
administr
republican
elect
report
polit
russia
impeach
investig
senat
new
time
told
peopl
state
democrat
say
presidenti
also
nation
0.00
0.05
0.10
0.15
0.20
0.25
TF-IDF Score


Top 30 Words in Cluster 5 Based on TF-IDF Score
like
peopl
one
time
said
get
go
work
game
first
new
know
think
way
make
realli
dont
season
even
day
would
two
want
also
life
love
come
feel
show
that
Words
0.00
0.01
0.02
0.03
0.04
TF-IDF Score


Top 30 Words in Cluster 6 Based on TF-IDF Score
Words
million
versus
net
coverag
text
sourc
revenu
compani
profit
per
share
year
ago
newsroom
earn
loss
quarter
yuan
billion
incom
percent
view
sa
attribut
total
oper
lira
march
financi
sept
0.00
0.05
0.10
0.15
0.20
0.25
0.30
TF-IDF Score


Top 30 Words in Cluster 7 Based on TF-IDF Score
oil
crude
said
product
energi
us
output
million
barrel
gas
per
pipelin
demand
market
petroleum
year
global
suppli
trade
russia
brent
cut
percent
day
compani
week
billion
would
refin
unit
Words
0.00
0.05
0.10
0.15
0.20
0.25
0.30
TF-IDF Score


Top 30 Words in Cluster 8 Based on TF-IDF Score
Words
senat
democrat
trump
hous
hill
said
bill
republican
vote
sander
presid
would
campaign
sen
committe
state
congress
tax
legisl
health
presidenti
polit
new
former
news
support
parti
elect
poll
subsidiari
0.00
0.02
0.04
0.06
0.08
0.10
TF-IDF Score

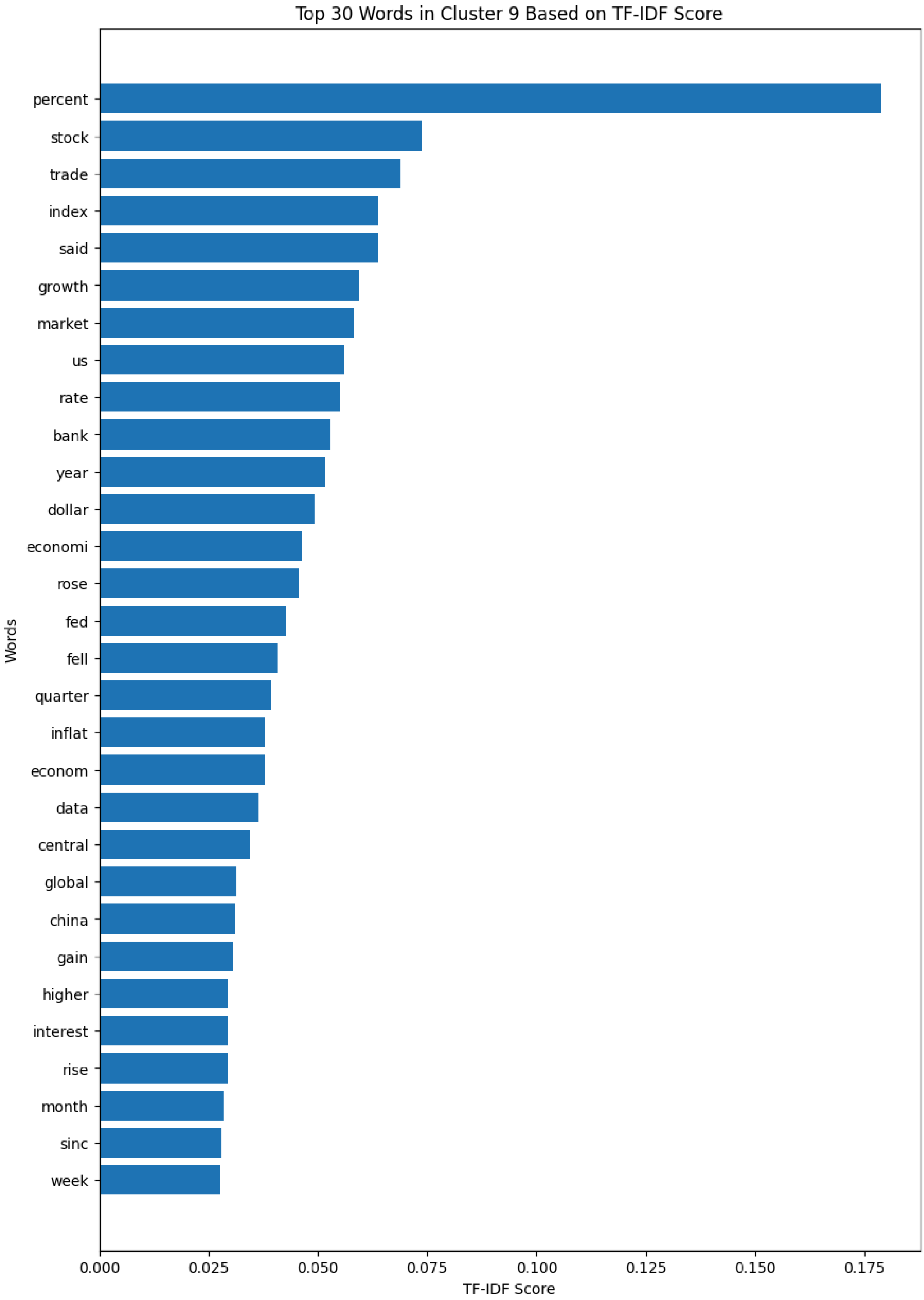
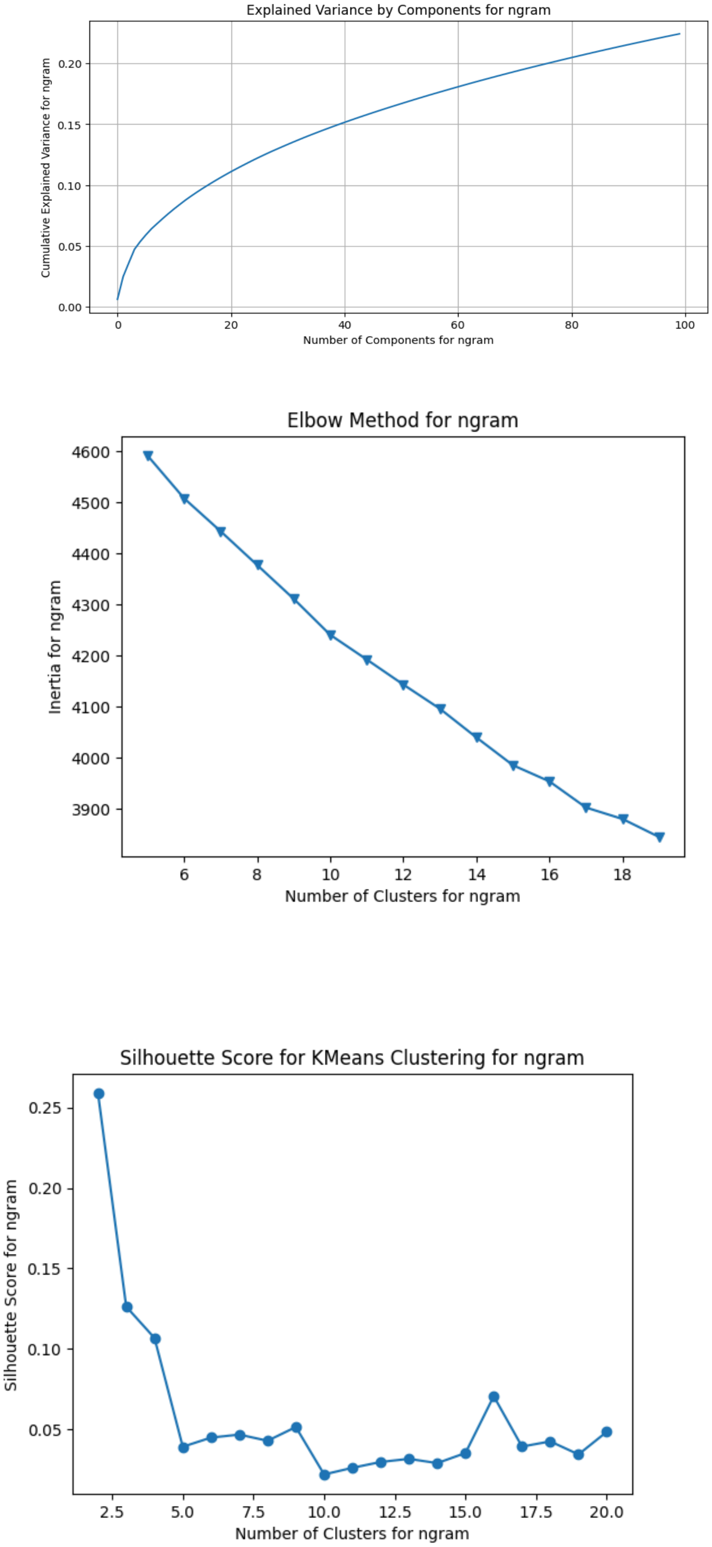
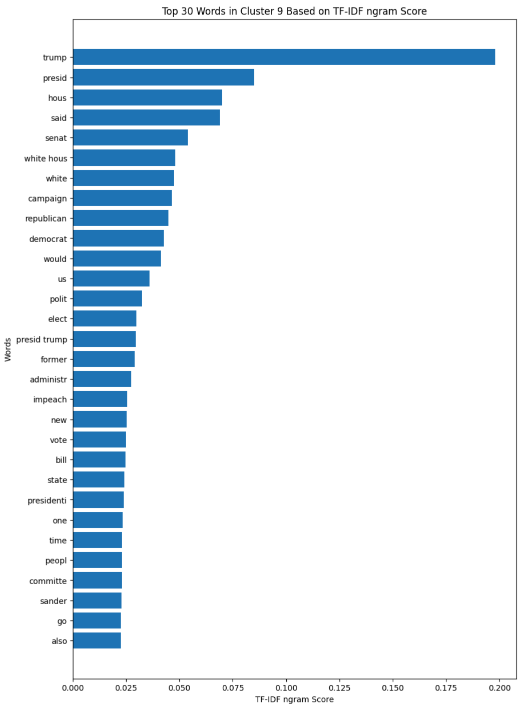
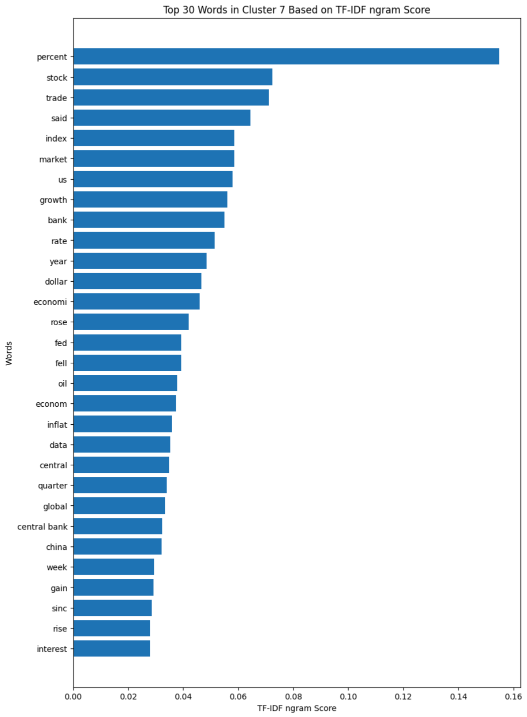
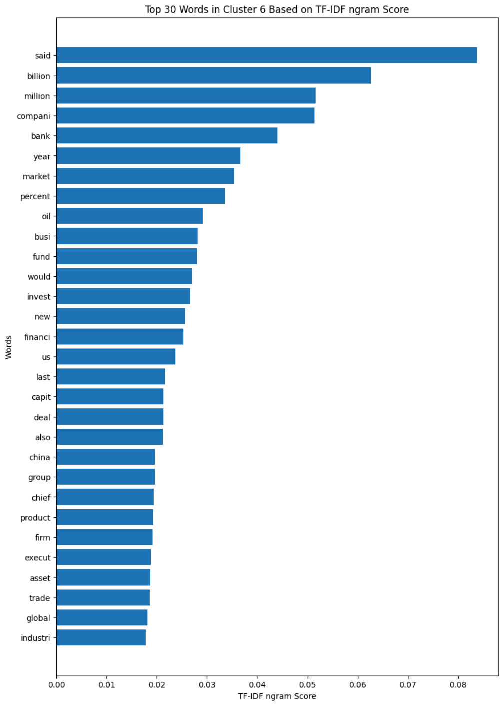

Top 30 Words in Cluster 9 Based on TF-IDF Score
Words
percent
stock
trade
index
said
growth
market
us
rate
bank
year
dollar
economi
rose
fed
fell
quarter
inflat
econom
data
central
global
china
gain
higher
interest
rise
month
sinc
week
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
TF-IDF Score

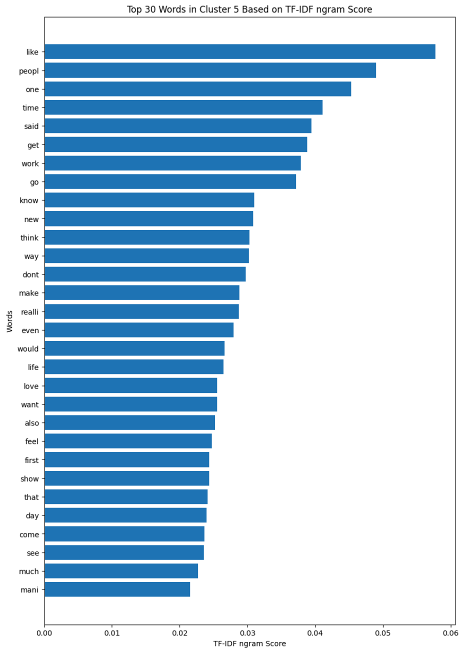
Fig 4.2 PCA ngram, Elbow Method, Silhouette Plot

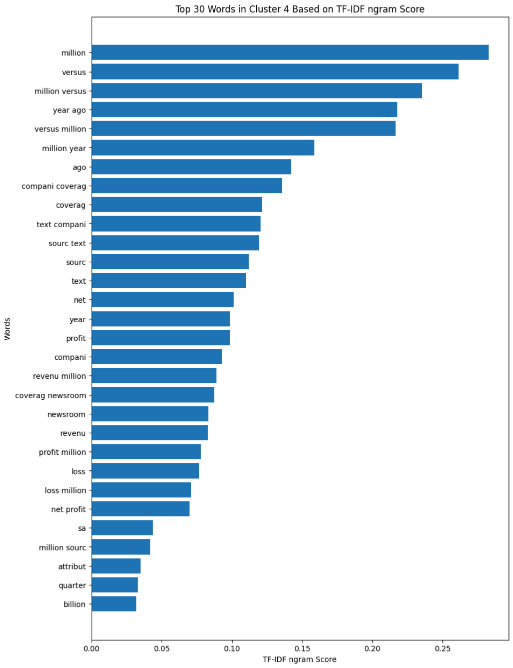


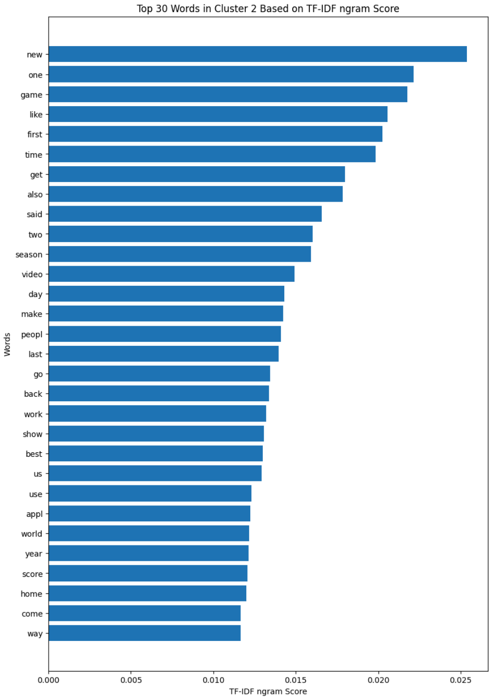

Top 30 Words in Cluster 9 Based on TF-IDF ngram Score
Words
trump
presid
hous
said
senat
white hous
white
campaign
republican
democrat
would
us
polit
elect
presid trump
former
administr
impeach
new
vote
bill
state
presidenti
one
time
peopl
committe
sander
go
also
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
0.200
TF-IDF ngram Score

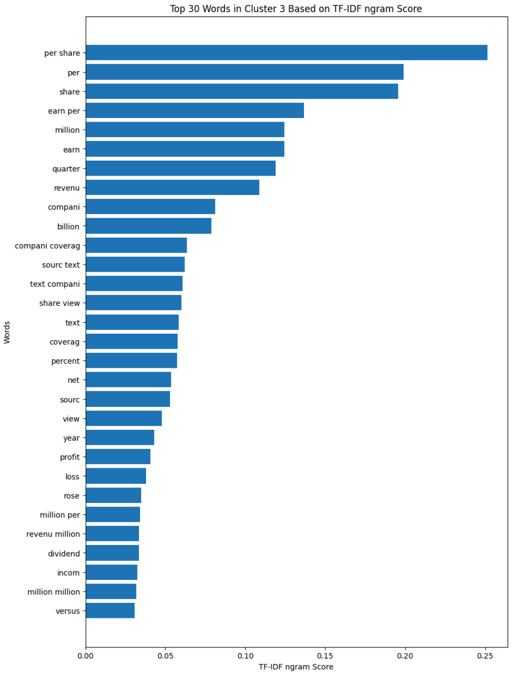

Top 30 Words in Cluster 7 Based on TF-IDF ngram Score
Words
percent
stock
trade
said
index
market
us
growth
bank
rate
year
dollar
economi
rose
fed
fell
oil
econom
inflat
data
central
quarter
global
central bank
china
week
gain
sinc
rise
interest
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
TF-IDF ngram Score

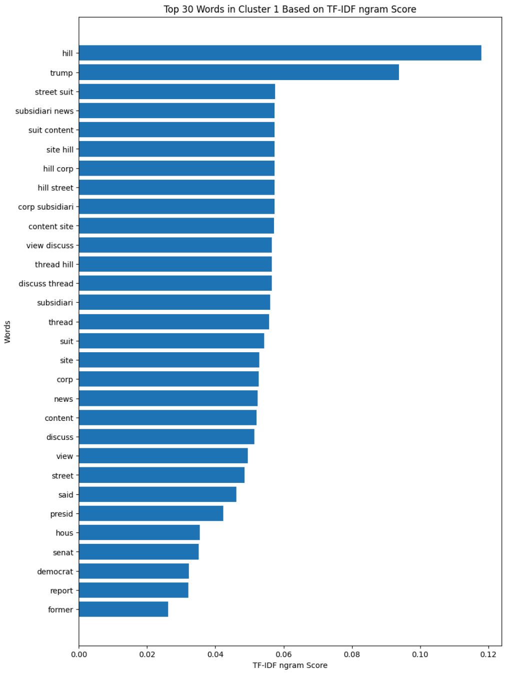

Top 30 Words in Cluster 6 Based on TF-IDF ngram Score
Words
said
billion
million
compani
bank
year
market
percent
oil
busi
fund
would
invest
new
financi
us
last
capit
deal
also
china
group
chief
product
firm
execut
asset
trade
global
industri
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
TF-IDF ngram Score

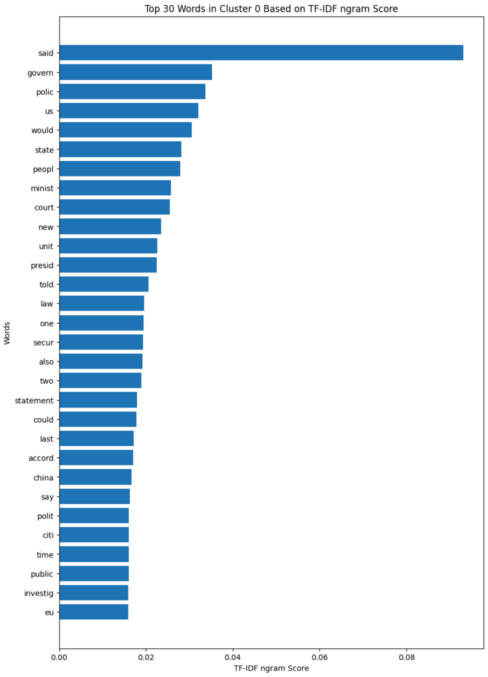

Top 30 Words in Cluster 5 Based on TF-IDF ngram Score
Words
like
peopl
one
time
said
get
work
go
know
new
think
way
dont
make
realli
even
would
life
love
want
also
feel
first
show
that
day
come
see
much
mani
0.00
0.01
0.02
0.03
0.04
0.05
0.06
TF-IDF ngram Score


Top 30 Words in Cluster 4 Based on TF-IDF ngram Score
Words
million
versus
million versus
year ago
versus million
million year
ago
compani coverag
coverag
text compani
sourc text
sourc
text
net
year
profit
compani
revenu million
coverag newsroom
newsroom
revenu
profit million
loss
loss million
net profit
sa
million sourc
attribut
quarter
billion
0.00
0.05
0.10
0.15
0.20
0.25
TF-IDF ngram Score


Top 30 Words in Cluster 2 Based on TF-IDF ngram Score
Words
new
one
game
like
first
time
get
also
said
two
season
video
day
make
peopl
last
go
back
work
show
best
us
use
appl
world
year
score
home
come
way
0.000
0.005
0.010
0.015
0.020
0.025
TF-IDF ngram Score


Top 30 Words in Cluster 3 Based on TF-IDF ngram Score
Words
per share
per
share
earn per
million
earn
quarter
revenu
compani
billion
compani coverag
sourc text
text compani
share view
text
coverag
percent
net
sourc
view
year
profit
loss
rose
million per
revenu million
dividend
incom
million million
versus
0.00
0.05
0.10
0.15
0.20
0.25
TF-IDF ngram Score


Top 30 Words in Cluster 1 Based on TF-IDF ngram Score
Words
hill
trump
street suit
subsidiari news
suit content
site hill
hill corp
hill street
corp subsidiari
content site
view discuss
thread hill
discuss thread
subsidiari
thread
suit
site
corp
news
content
discuss
view
street
said
presid
hous
senat
democrat
report
former
0.00
0.02
0.04
0.06
0.08
0.10
0.12
TF-IDF ngram Score


Top 30 Words in Cluster 0 Based on TF-IDF ngram Score
Words
said
govern
polic
us
would
state
peopl
minist
court
new
unit
presid
told
law
one
secur
also
two
statement
could
last
accord
china
say
polit
citi
time
public
investig
eu
0.00
0.02
0.04
0.06
0.08
TF-IDF ngram Score