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 20,000. Through the previous iteration found in phase-2 we were able to identify and remove words like ‘said’, ‘would’, ‘also’ etc.

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 20,000 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.

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. We also used cosine similarity to find top 5 similar documents

to a given output. Similarly, we have also identified the top 30 words clusters with different n-gram ranges to cross-validate how the clustering model performs for different approaches. To reduce complexity in hyperparameter tuning for higher n-gram ranges we utilized subsetting based on year and publishers. These data subsets help not only with complexity management but also highlight business use cases better.

We employed the Non-negative Matrix Factorization(NMF) model to analyze article text, utilizing vectors derived from the TFIDF model. Our approach involved iterative refinement to determine the optimal number of topics. Initially, we conducted a comprehensive analysis of the entire dataset, extracting the top 10 topics along with their representative words. Subsequently, we conducted a year-wise examination of articles spanning from 2016 to 2020. This method allows us to track the evolution of trends over time and also helps in strengthening the model.

Utilized the VADER sentiment analysis tool from the NLTK library to analyze the sentiment of articles within finance-related clusters (identified as clusters 5, 6, and 9). By first filtering the original dataset to include only articles belonging to these finance clusters, we apply sentiment analysis to each processed article using VADER's polarity scoring mechanism. The sentiment of each article is evaluated and the sentiment scores obtained are then categorized into 'positive', 'negative', or 'neutral' labels. With this approach, we tried to understand the sentiment expressed in finance-related articles

In conclusion, our iterative approach, underpinned by strategic choices of PCA for dimensionality reduction, KMeans for clustering, and the elbow method and silhouette plots for determining the optimal number of clusters, allowed us to effectively cluster news articles. This methodology not only achieved our goal but also laid a robust foundation for future explorations in the field of text analysis, opening avenues for further refinement and optimization.

Results

The clarity of themes within each cluster was our evaluation metric. After learning from our previous attempt we now tried it with a better feature set and increased the number of clusters to 15. Three clusters could not be placed into any groups by looking at the top words in the clusters. The remaining clusters, however, showed distinct themes like crime news, federal government news, national politics, business finance, financial markets, sports news, and the stock market(4.1). Our evaluation metric was the clarity of the themes within each cluster. out of 15 clusters, 12 had distinct themes with little overlap. However, the clarity of themes is also affected by sub-par clustering. Our clusters included at least 3 finance-related clusters which could all be combined into one cluster.

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). 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. Moreover, we performed bigrams for the newspaper articles belonging to the year 2019, we saw articles related to politics in the first few

clusters which were also overlapping. A really good cluster(=4) that was formed was articles that had "New York" (4.3)in it. We were able to see new clusters associated with social media names in it. Next, we also performed trigrams subsetting publication as "People". Though there were some spillovers, the model did a satisfactory job at clustering different topics and we were able to find new clusters in entertainment and law and crime using the tri-gram(4.4).

Through the NMF model for all the articles, a wide range of topics including Financial News, Politics and Government, Lifestyle, Law and Governance, Company Financial News, Corporate News and Acquisitions, Global Trade and Economy, Stock Market, Legal Discussions, and Sports were obtained. For each year, we identified the top 10 topics(4.5), revealing significant overlap across years in key areas such as Politics and Government, Stock Market, Financial News, Global Trade and Economy, and Health Care. Notably, each year introduced at least one new topic reflecting the most trending theme, such as Trump's campaigns in 2016(4.6), Russian interference in 2017(4.7), Trade wars in 2018(4.8), the Oil crisis in 2019(4.9), and the covid-19 pandemic in 2020(4.10). This adaptive approach ensures our model remains relevant and accurate, capturing current trends and reader interests. News channels can also utilize these insights to update their trending section.

Our sentiment analysis on finance-related articles revealed a wide range of sentiments. Despite market volatility, many articles showed neutral sentiment, indicating balanced reporting. Peaks in positive sentiment could correspond to favorable market developments, while dips might align with negative events. These insights can help businesses gauge market sentiment and inform strategies, forecasting, and product development in the finance sector.

Limitations

- POS Tagging and NER: The potential of POS-tagged and weighted TF-IDF models remains untapped. Future work could explore creating multiple models based on different weights, potentially enhancing results.
- Hierarchical Clustering: The high time complexity ($O(n^3)$) and space complexity ($O(n^2)$) of hierarchical clustering posed challenges with our dataset of 26,000 rows. Future efforts could explore efficient sampling methods or alternative algorithms for large datasets.
- Word2Vec: While Word2Vec's extensive list of 3 million words and phrases is a strength, the 300-dimensional feature space is resource-intensive. Future work could investigate dimensionality reduction techniques to manage this challenge.
- Sentiment Analysis: A challenge with sentiment analysis across clusters belonging to unique sectors is that each genre may have its unique keywords, vocabularies, and contextual nuances, in such cases it might potentially lead to less accurate sentiment classification.

Coding Contribution

For contributions, a detailed table is provided in the appendix(fig 0). A **kanban board** used for the project can be found at this link. Everyone equally contributed to the analysis and slides.

GitHub Project **News Article Text Analysis**

Appendix

Fig 0: Coding Contribution

Name	Members	Contribution per member
EDA	Mahika, Rajashree	50%,50%
Pre-processing (incl stem)	Dylan, Sahasra	50%,50%
Bag of Words	Dylan, Sahasra	50%,50%
TF-IDF model + clustering	Dylan	100%
TF-IDF model hyperparam	Mahika, Rajashree	50% 50%
Cosine similarity	Dylan	100%
Ngram (for range in 1.2) + clustering	Mahika, Rajashree	50%, 50%
Ngram (for range in 2,3) + clustering year wise	Rajashree	100%
Ngram (trigrams) + clustering publication wise	Rajashree	100%
NMF model(incl sub-models)	Sahasra	100%
NMF all visualizations	Mahika	100%
Sentimental Analysis	Rajashree	100%
Consolidation	Dylan, Mahika, Rajashree, Sahasra	25%, 25%, 25%,25%
PCA scatter plots	Mahika	100%

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

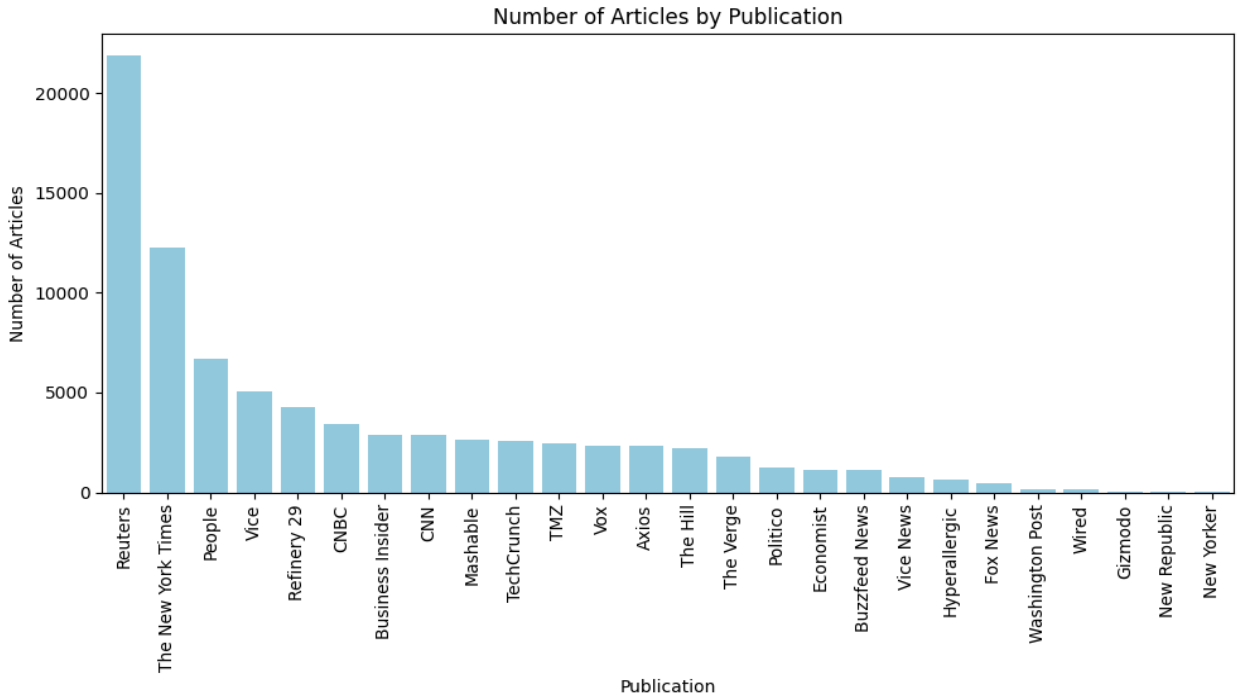


Fig 1.3: Distribution of Articles Across Sections

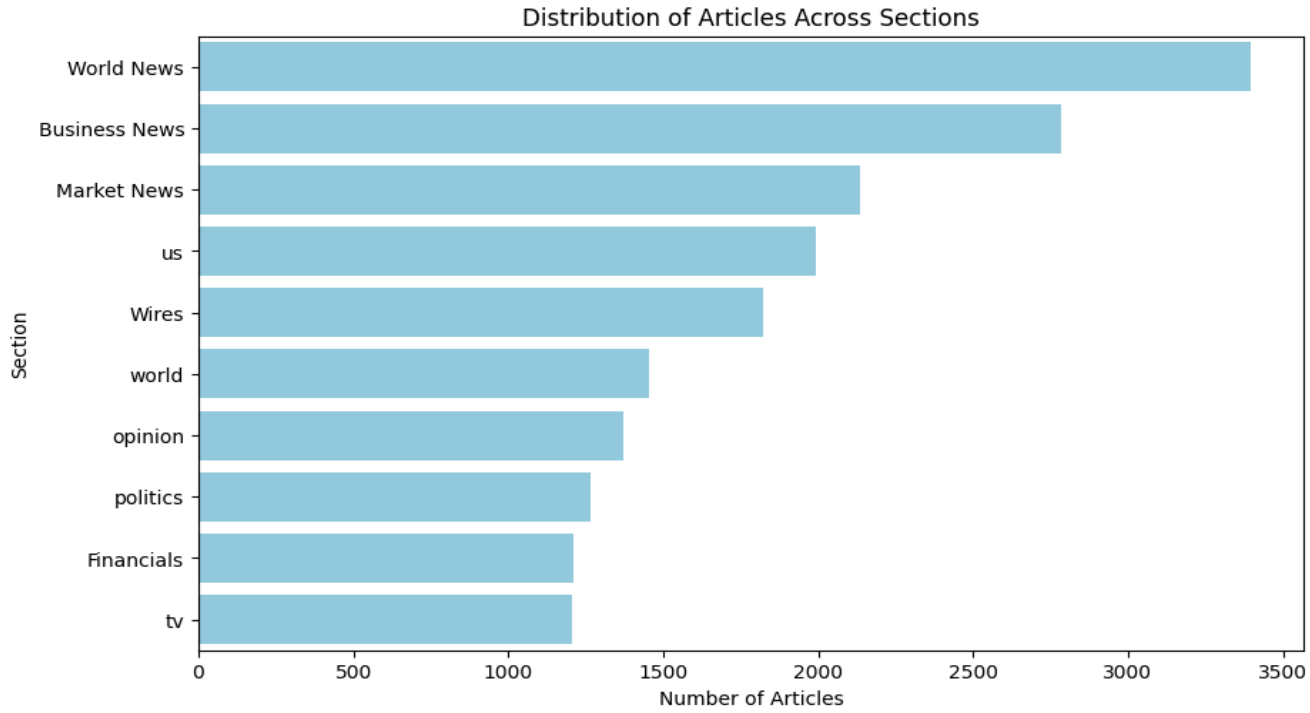


Fig 1.5: Word Cloud of the column 'title'

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:

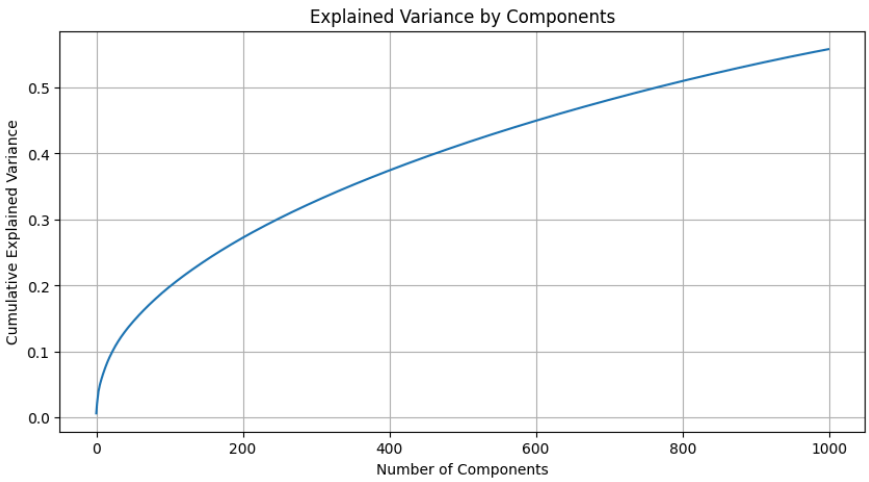


Fig 3.2 Elbow Method

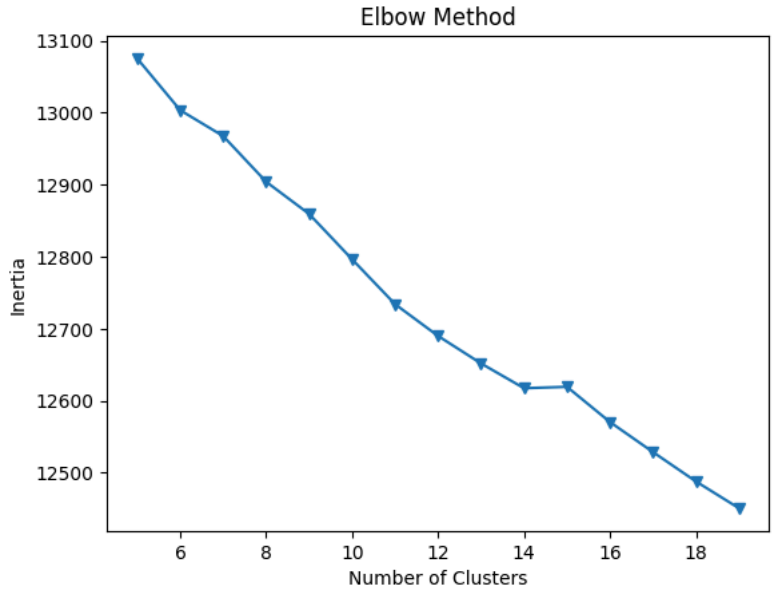


Fig 3.3 Silhouette Plot:

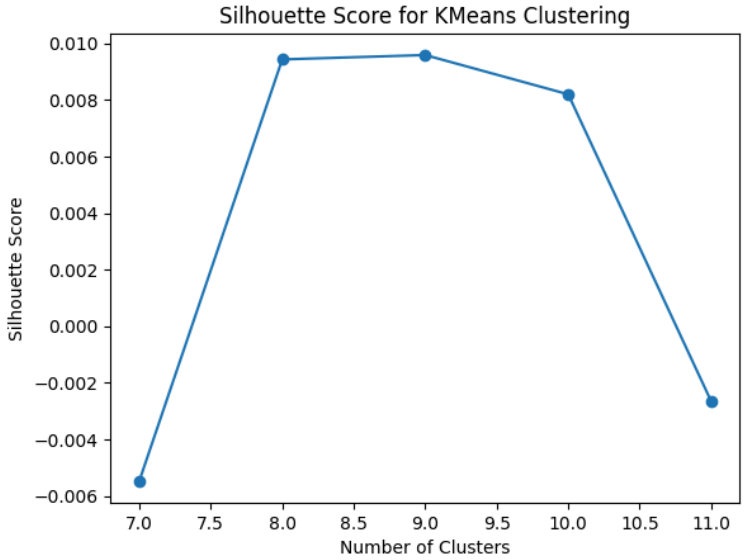
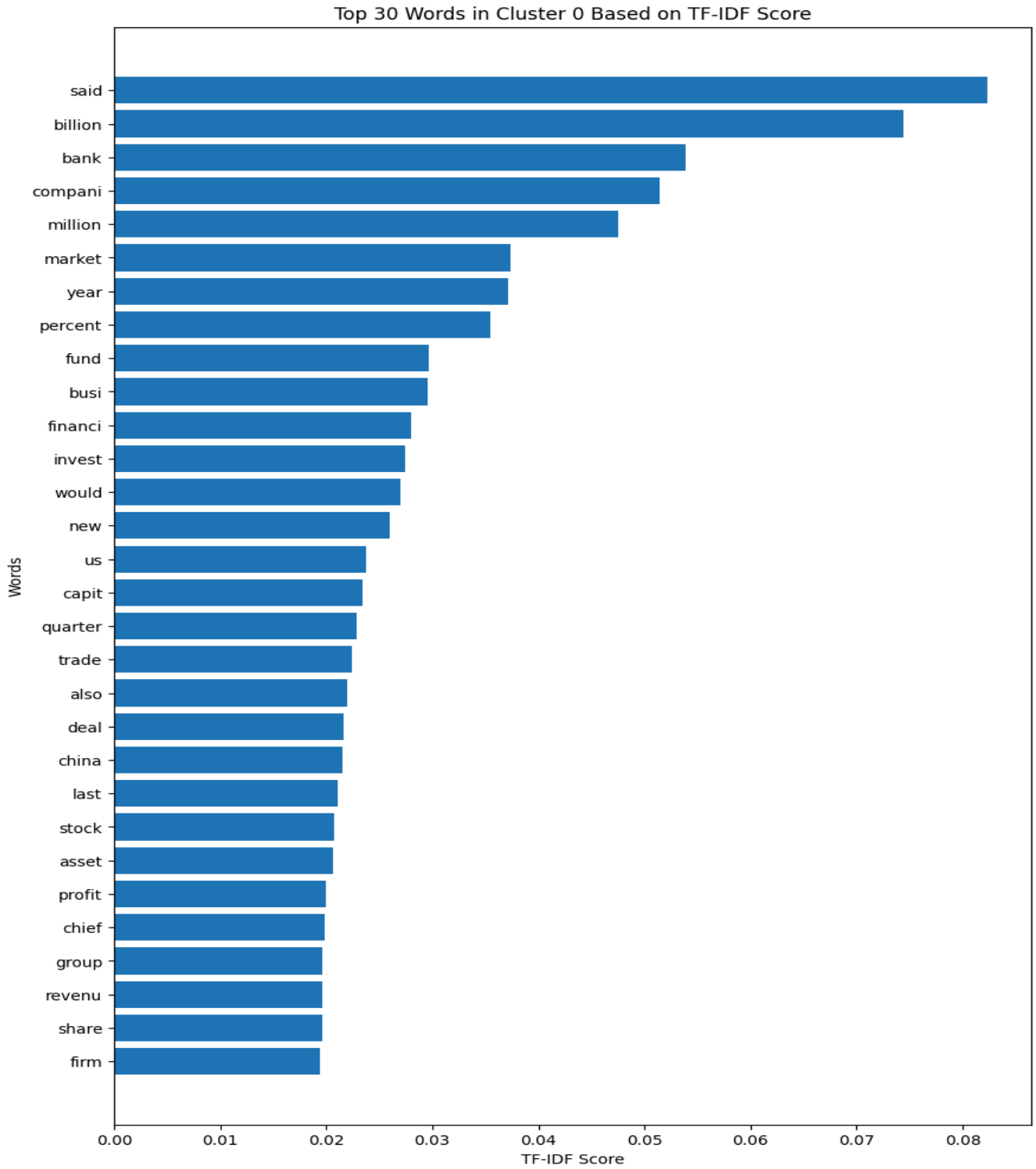
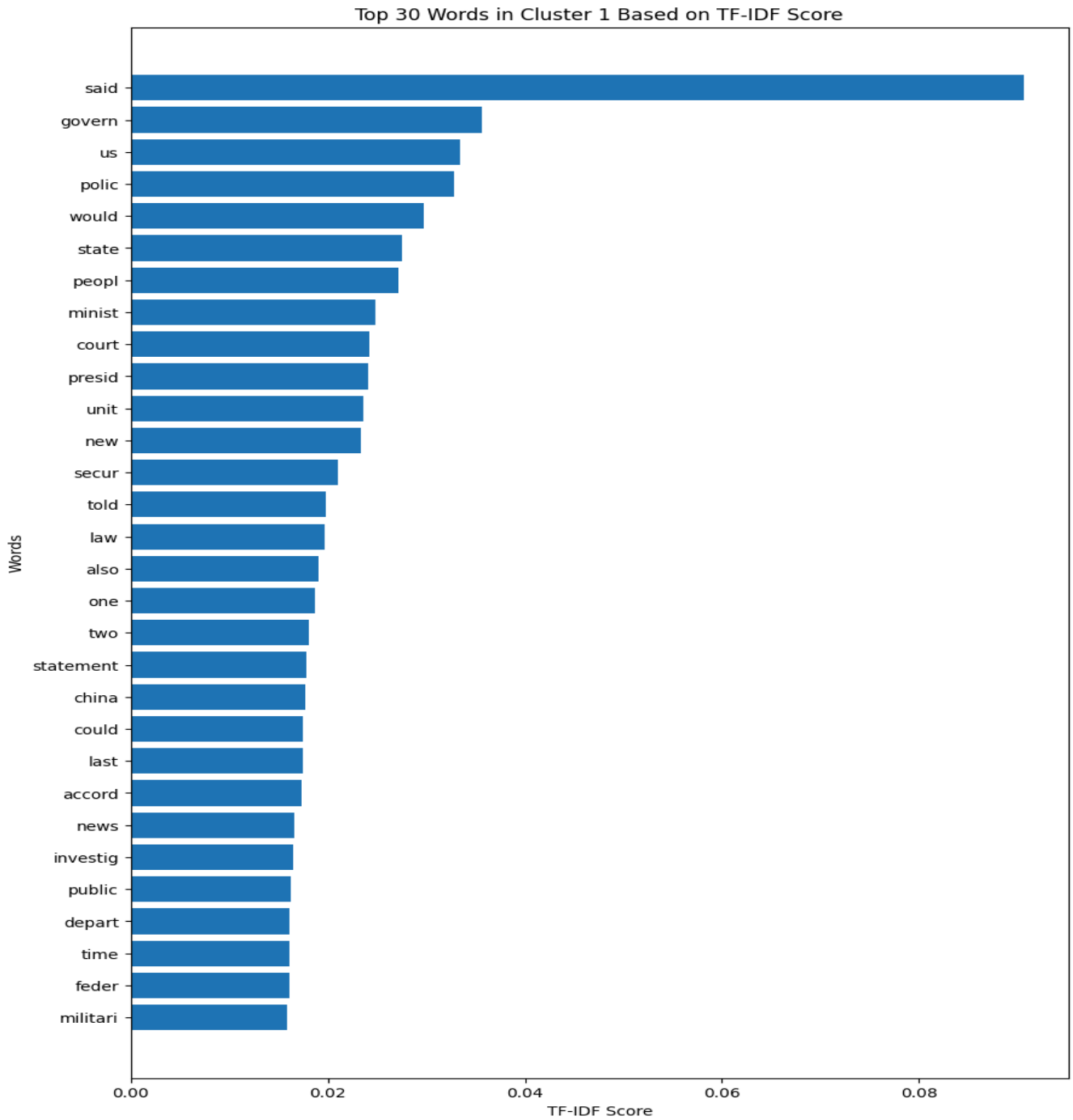
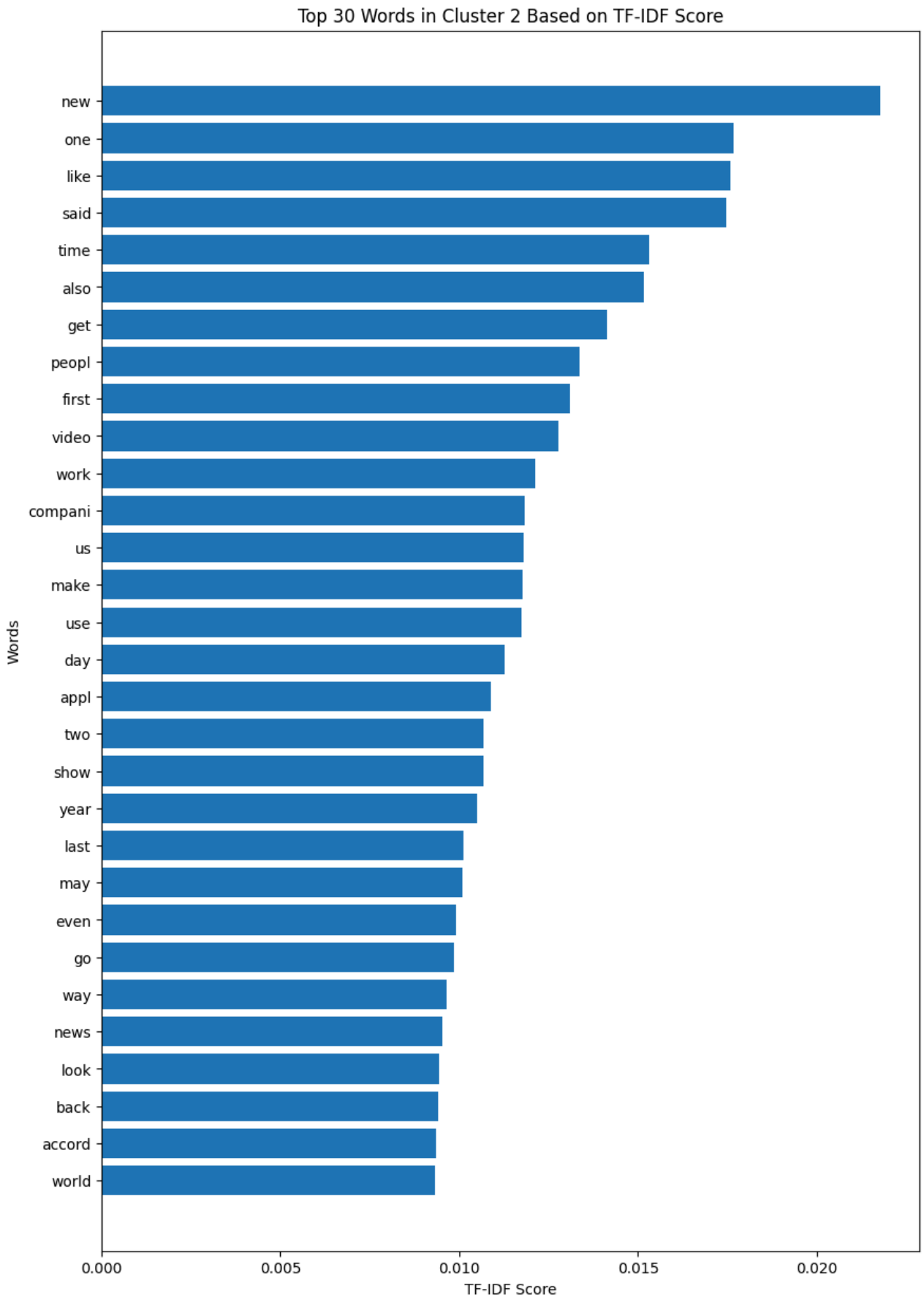


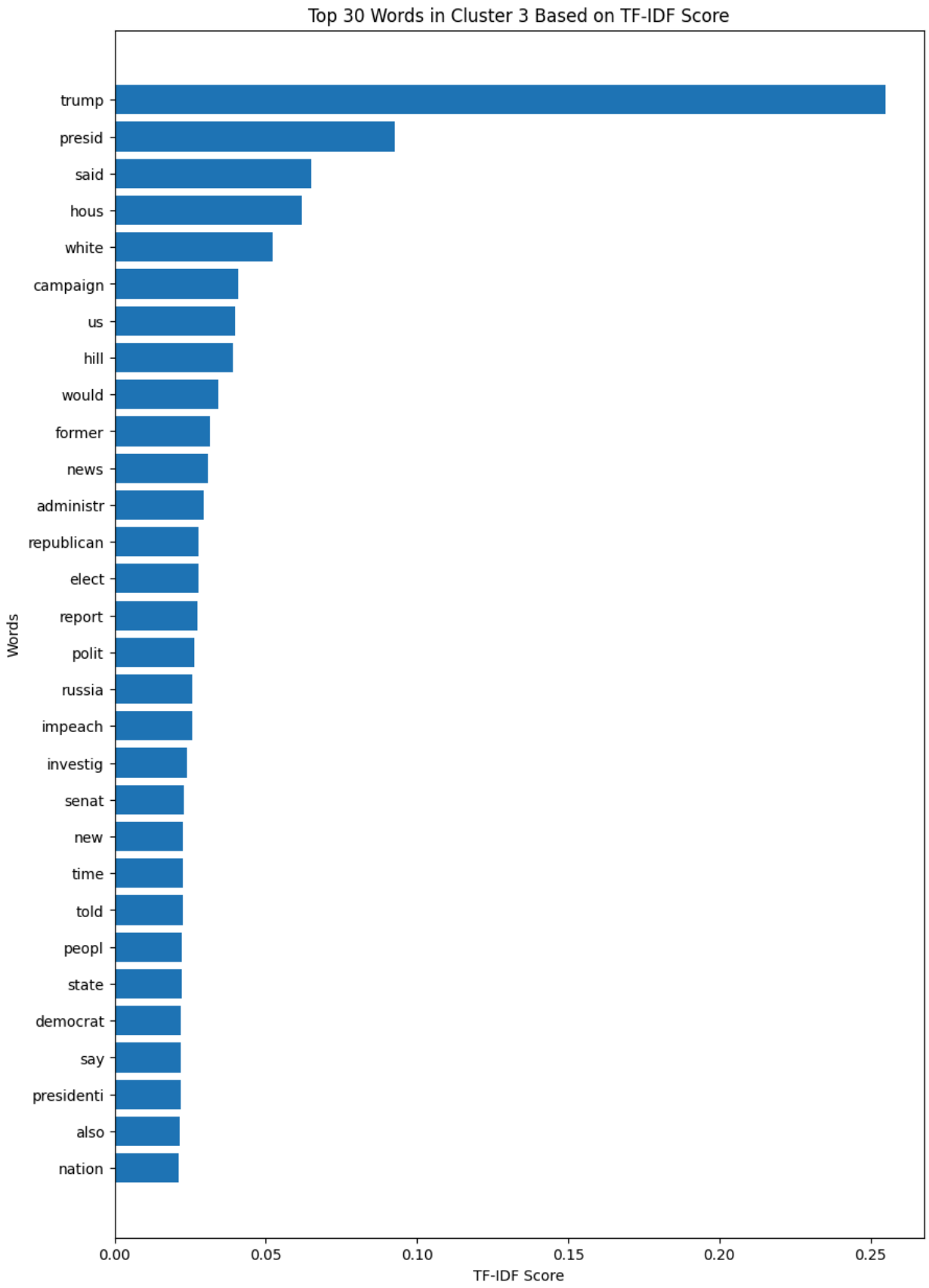
Fig 4.1 Clustering results through KMeans

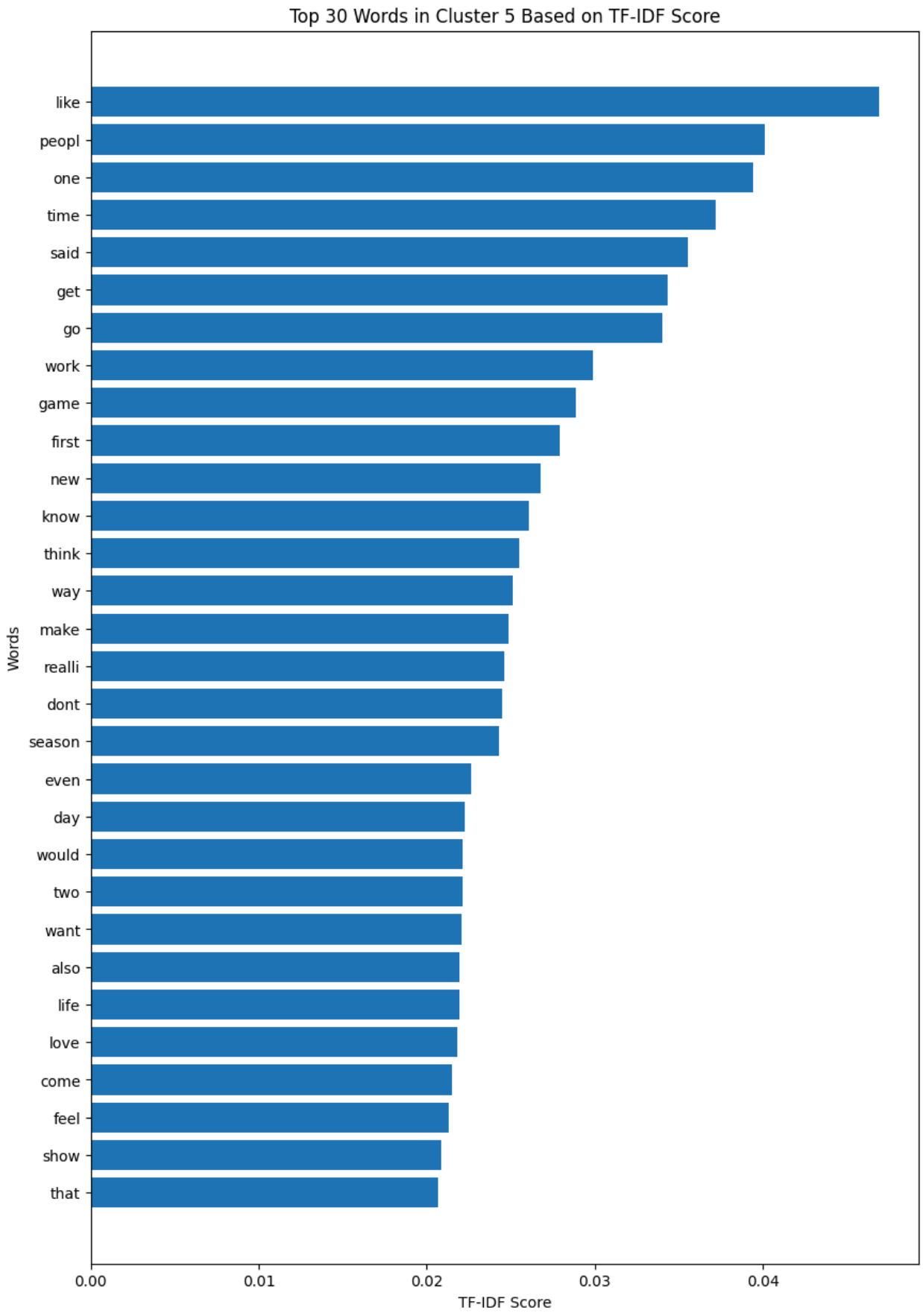
Cluster	Words
0	like, one, new, time, get, peopl, work, go, show, first
1	text, coverag, sourc, compani, newsroom, million, headlin, share, march, corp
2	billion, million, compani, year, fund, market, percent, busi, invest, quarter
3	trump, presid, hous, white, us, campaign, former, administr, report, elect
4	polic, peopl, shoot, man, offic, crime, suspect, accord, told, citi
5	us, new, peopl, govern, state, court, unit, presid, one, accord
6	senat, democrat, hous, trump, republican, bill, vote, sander, presid, campaign
7	percent, stock, index, growth, year, rose, quarter, market, us, fell
8	million, versus, net, coverag, sourc, text, revenu, profit, compani, ago
9	oil, crude, product, energi, us, gas, output, barrel, million, per
10	minist, eu, prime, govern, parti, parliament, union, deal, vote, elect
11	game, season, score, win, first, team, two, second, play, three
12	bank, central, rate, fed, interest, economi, percent, polici, inflat, monetari
13	appl, cook, new, compani, watch, pro, tech, like, devic, market
14	trade, china, us, stock, market, index, global, economi, econom, dollar

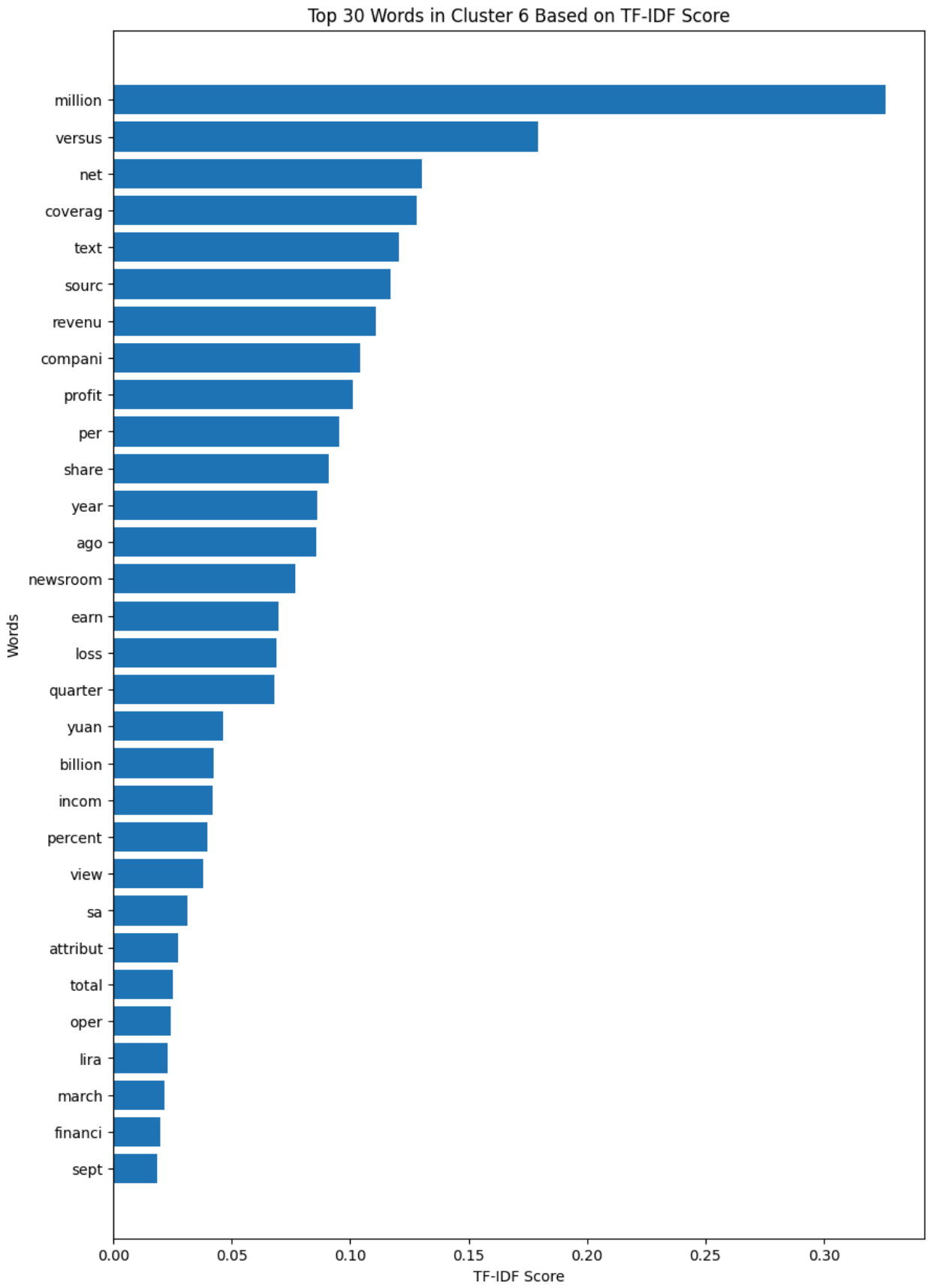

Top 30 Words in Cluster 0 Based on TF-IDF Score
said
billion
bank
compani
million
market
year
percent
fund
busi
financi
invest
would
new
us
capit
quarter
trade
also
deal
china
last
stock
asset
profit
chief
group
revenu
share
firm
Words
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
TF-IDF Score

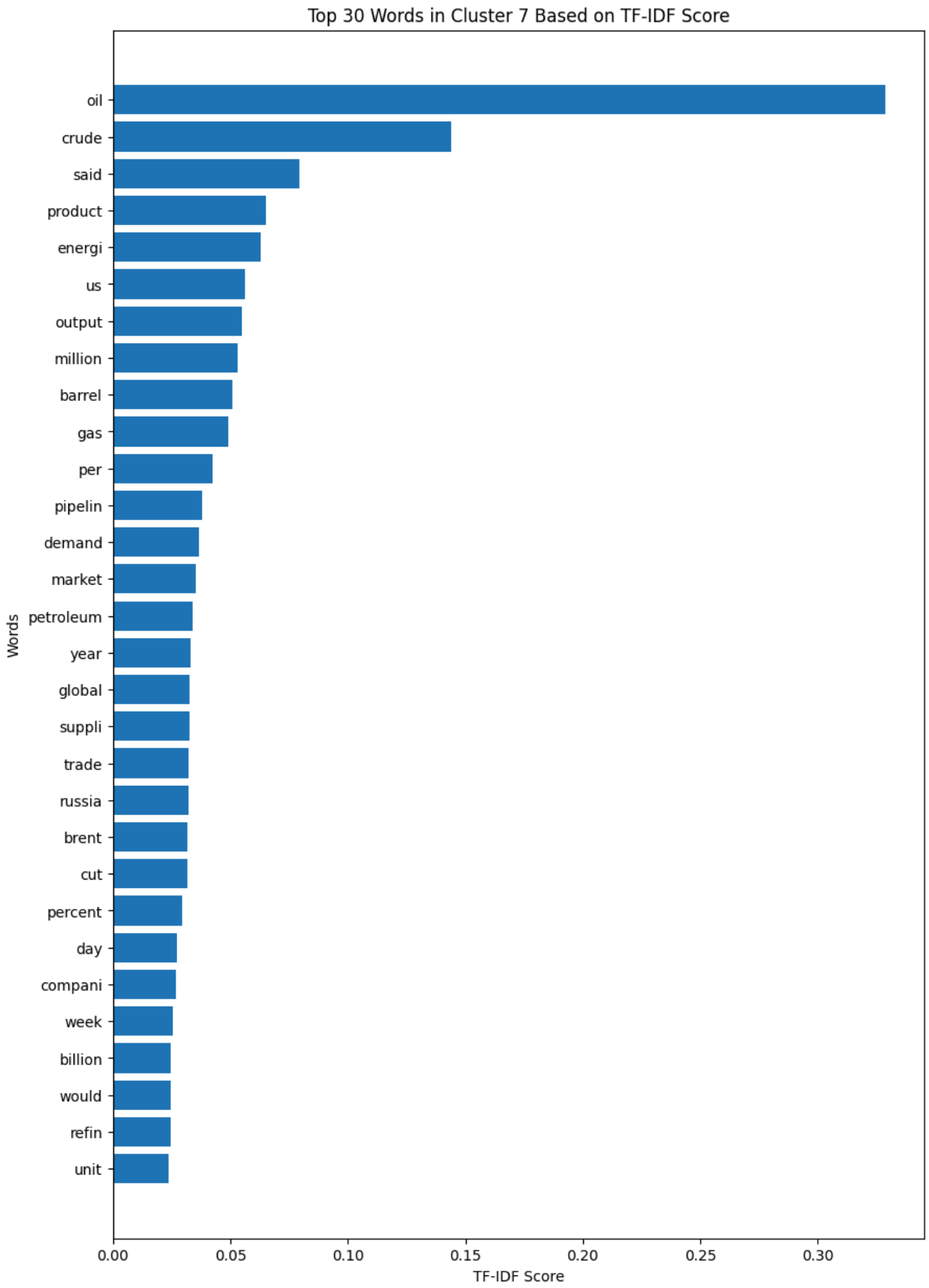

Top 30 Words in Cluster 1 Based on TF-IDF Score
said
govern
us
polic
would
state
peopl
minist
court
presid
unit
new
secur
told
law
also
one
two
statement
china
could
last
accord
news
investig
public
depart
time
feder
militari
Words
0.00
0.02
0.04
0.06
0.08
TF-IDF Score

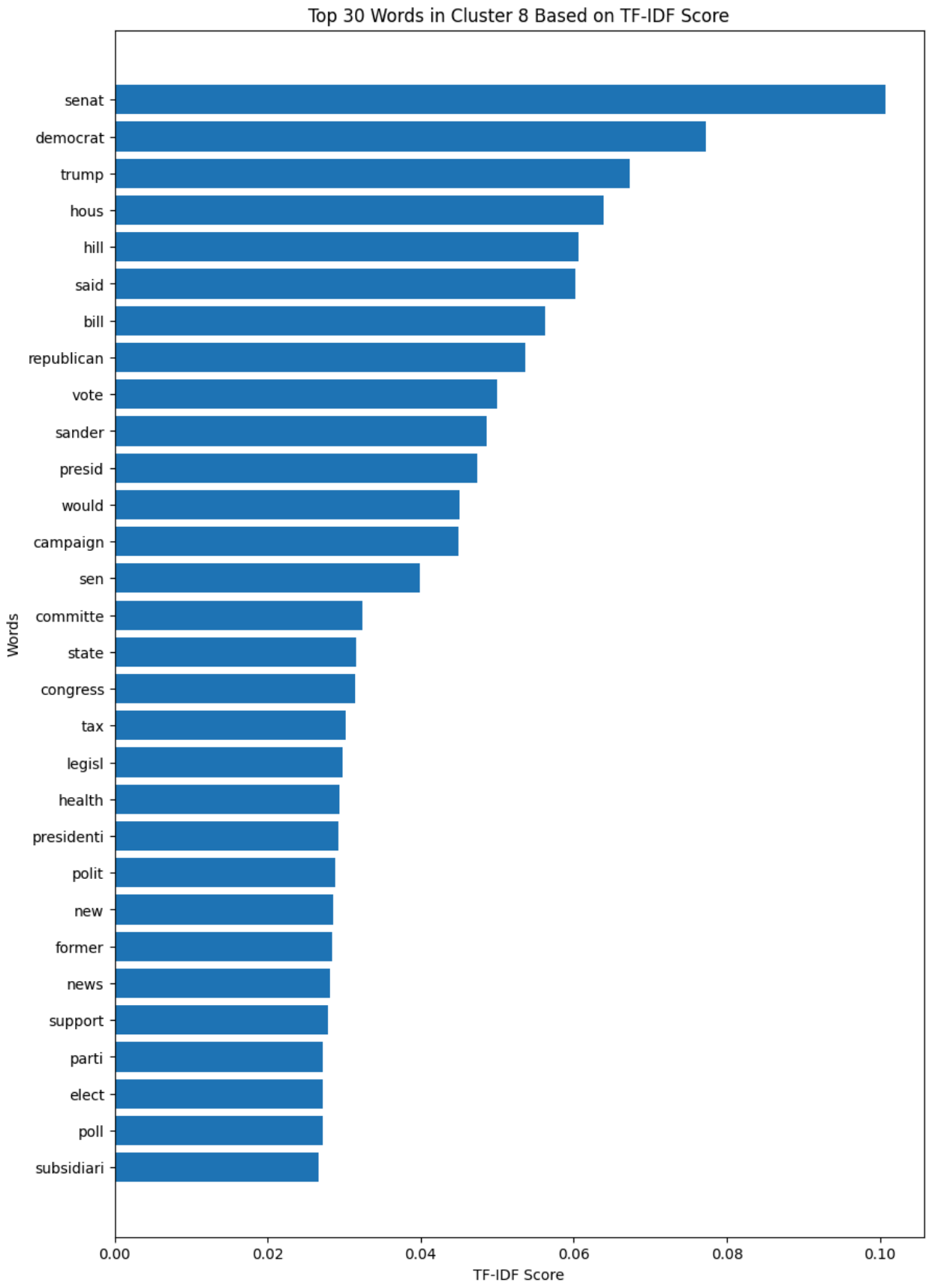

Top 30 Words in Cluster 2 Based on TF-IDF Score
Words
new
one
like
said
time
also
get
peopl
first
video
work
compani
us
make
use
day
appl
two
show
year
last
may
even
go
way
news
look
back
accord
world
0.000
0.005
0.010
0.015
0.020
TF-IDF Score


Top 30 Words in Cluster 3 Based on TF-IDF Score
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
Words
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
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
Words
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

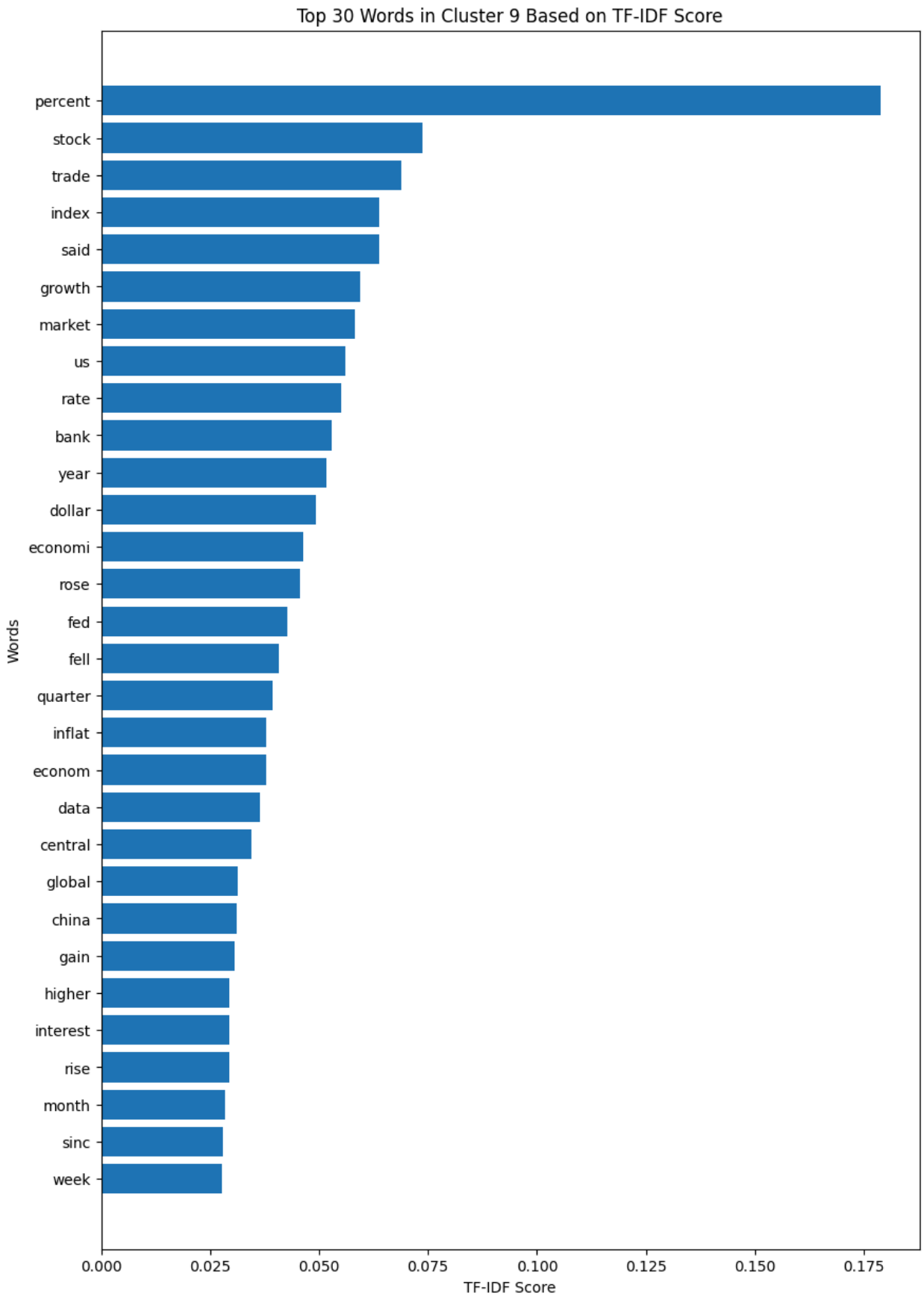
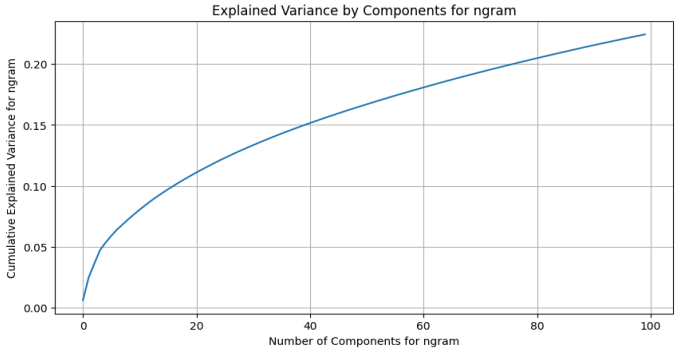
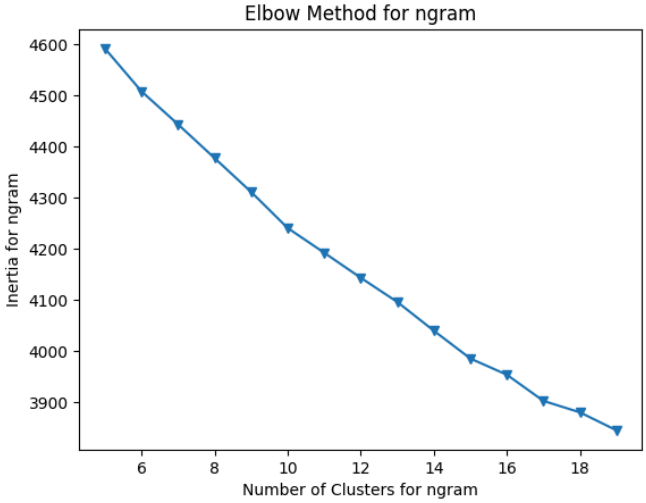
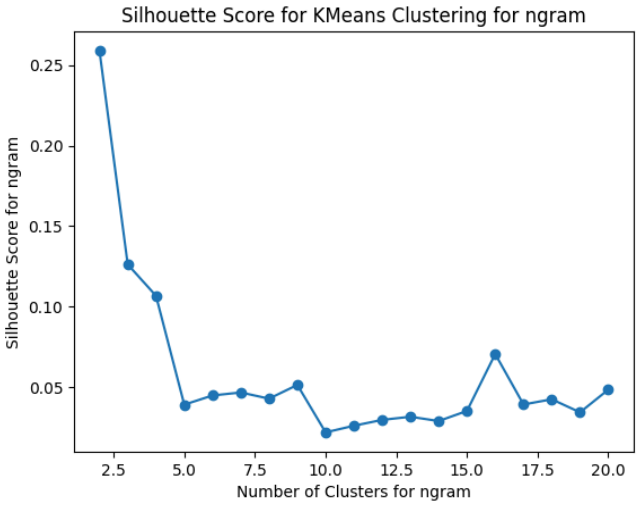
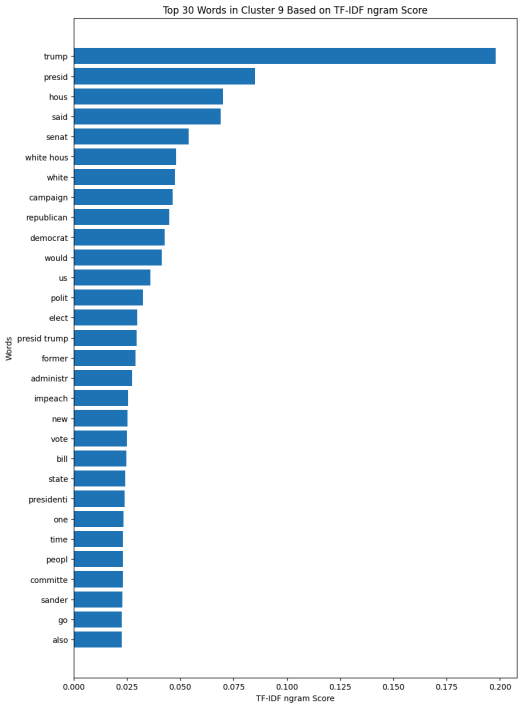

Top 30 Words in Cluster 9 Based on TF-IDF Score
Words
percent
stock
trade
index
said
growth
market
us
rate
bank
year
dollar
economi
rose
fed
fell
quarter
inflat
econom
data
central
global
china
gain
higher
interest
rise
month
sinc
week
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
TF-IDF Score

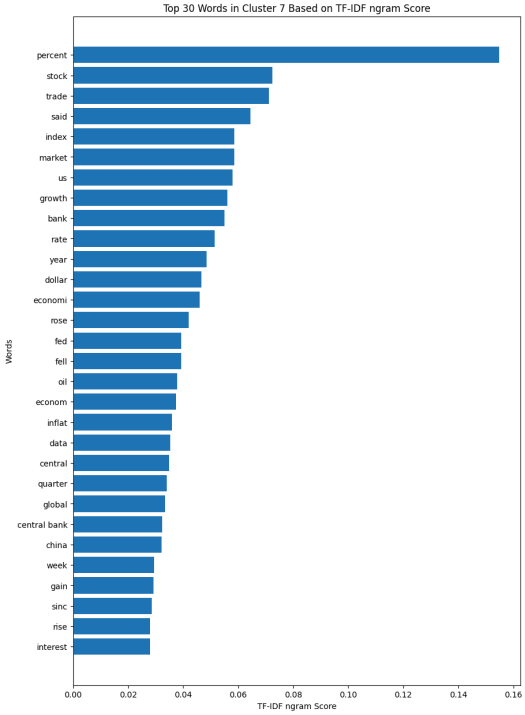
Fig 4.2 PCA Elbow Method, Silhouette Plot (ngram)

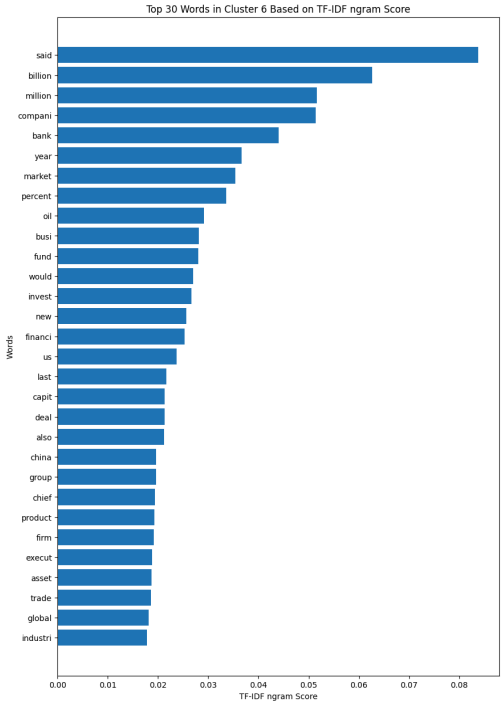


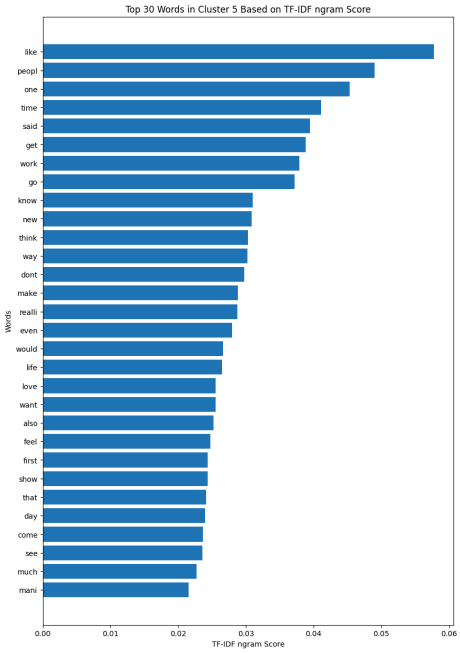


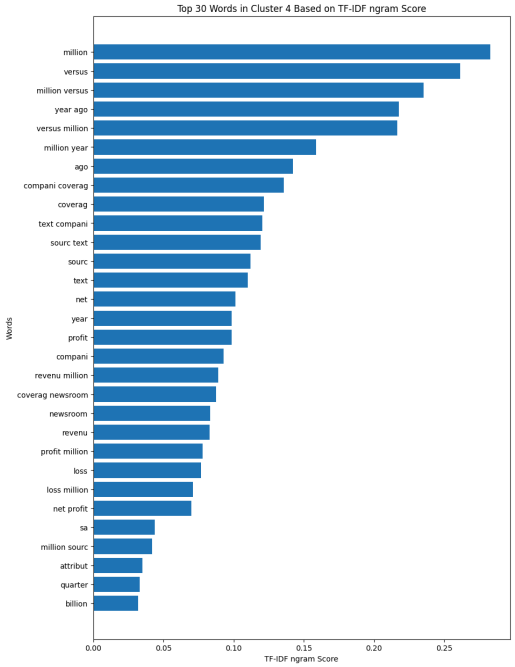


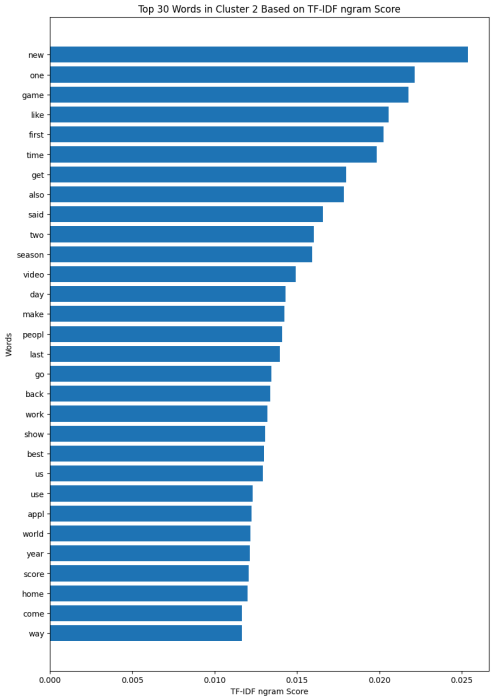

Top 30 Words in Cluster 9 Based on TF-IDF ngram Score
trump
presid
hous
said
senat
white hous
white
campaign
republican
democrat
would
us
polit
elect
presid trump
former
administr
impeach
new
vote
bill
state
presidenti
one
time
peopl
committe
sander
go
also
Words
0.000
0.025
0.050
0.075
0.100
0.125
0.150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

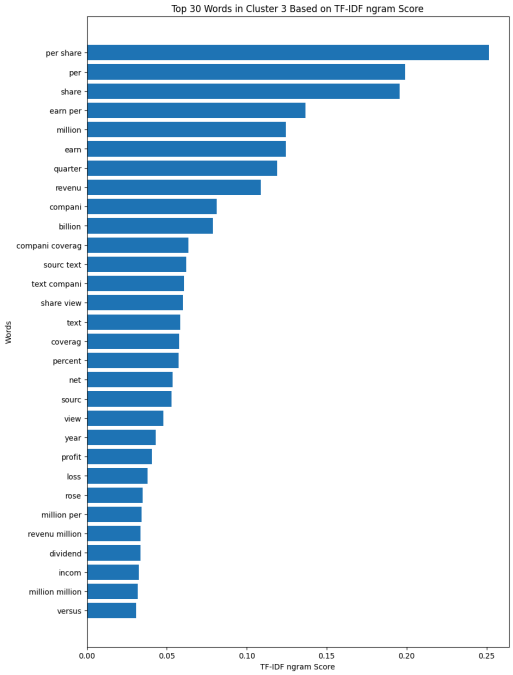

Top 30 Words in Cluster 7 Based on TF-IDF ngram Score
percent
stock
trade
said
index
market
us
growth
bank
rate
year
dollar
economi
rose
fed
fell
oil
econom
inflat
data
central
quarter
global
central bank
china
week
gain
sinc
rise
interest
Words
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
TF-IDF ngram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

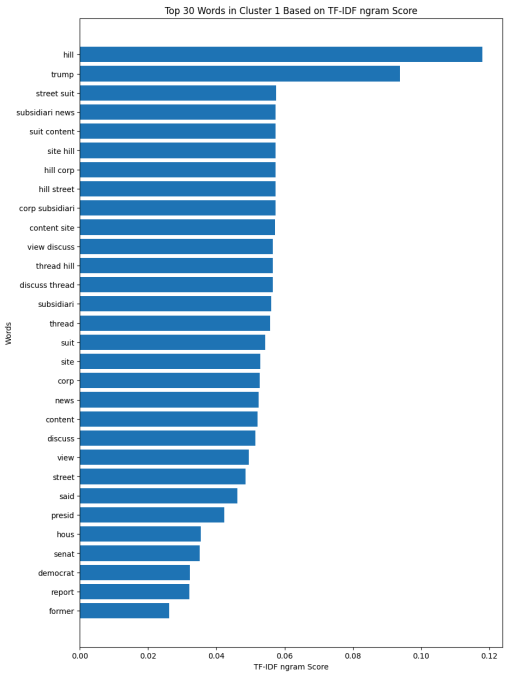

Top 30 Words in Cluster 6 Based on TF-IDF ngram Score
said
billion
million
compani
bank
year
market
percent
oil
busi
fund
would
invest
new
financi
us
last
capit
deal
also
china
group
chief
product
firm
execut
asset
trade
global
industri
Words
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
TF-IDF ngram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


Top 30 Words in Cluster 5 Based on TF-IDF ngram Score
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
Words
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
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
Words
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
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
Words
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
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
Words
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
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
Words
0.00
0.02
0.04
0.06
0.08
0.10
0.12
TF-IDF ngram Score

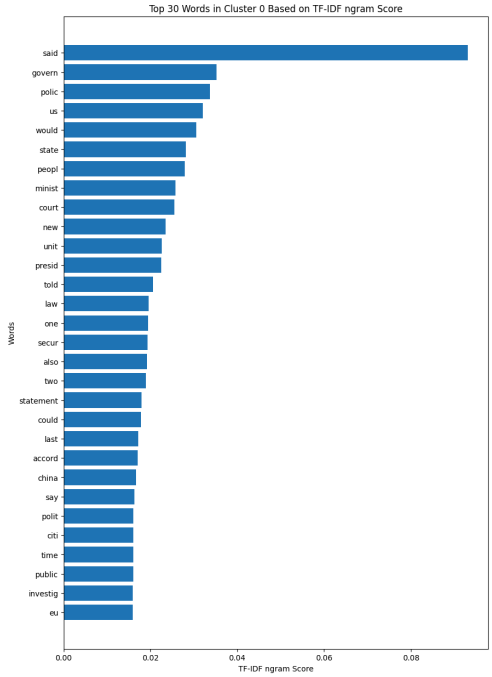

Top 30 Words in Cluster 0 Based on TF-IDF ngram Score
said
govern
polic
us
would
state
peopl
minist
court
new
unit
presid
told
law
one
secur
also
two
statement
could
last
accord
china
say
polit
citi
time
public
investig
eu
Words
0.00
0.02
0.04
0.06
0.08
TF-IDF ngram Score

Fig 4.3 NewYork cluster by sampling 2019 data with (2,3) variate range

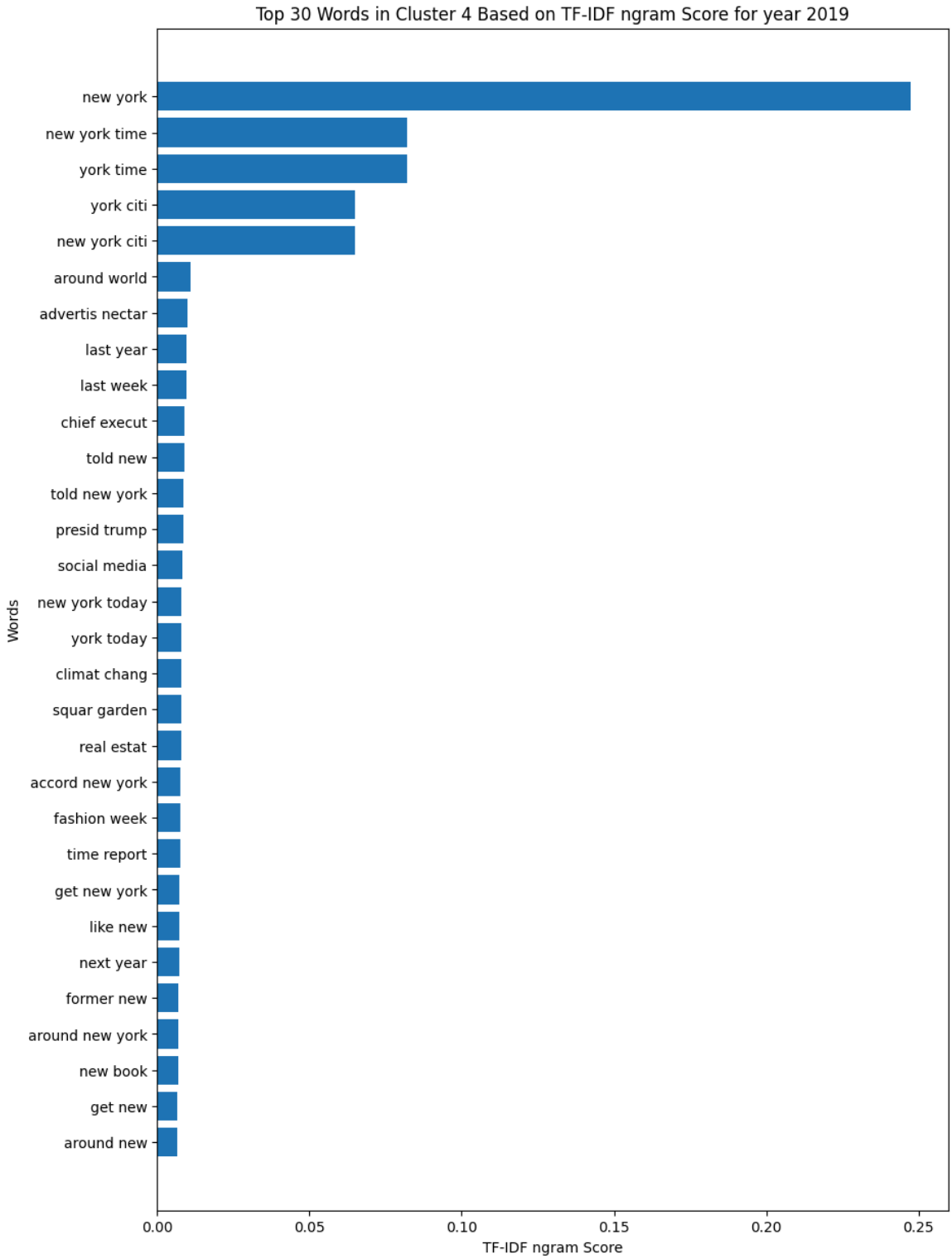
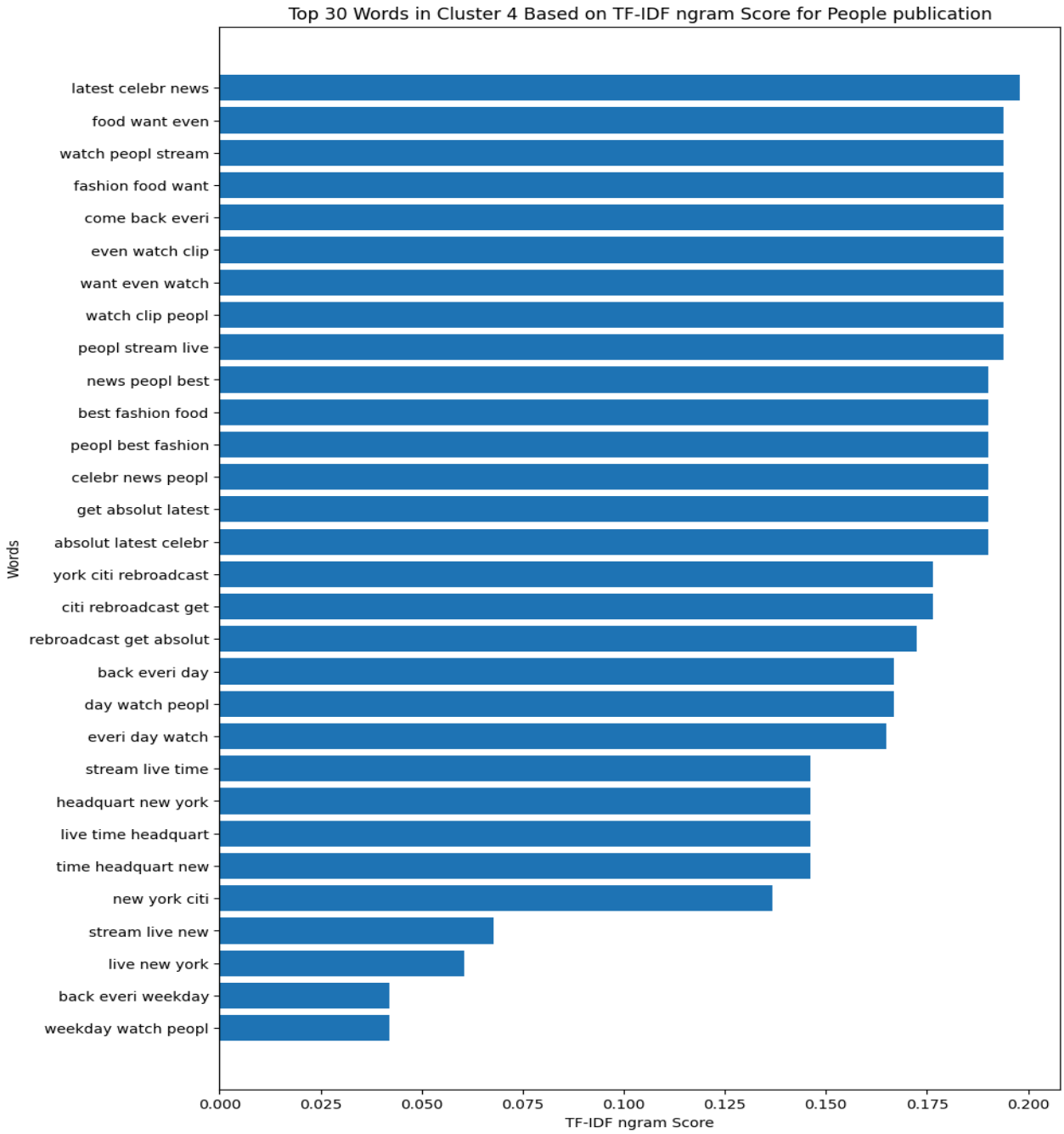


Fig 4.4 Entertainment and Law and Crime cluster using tri-gram



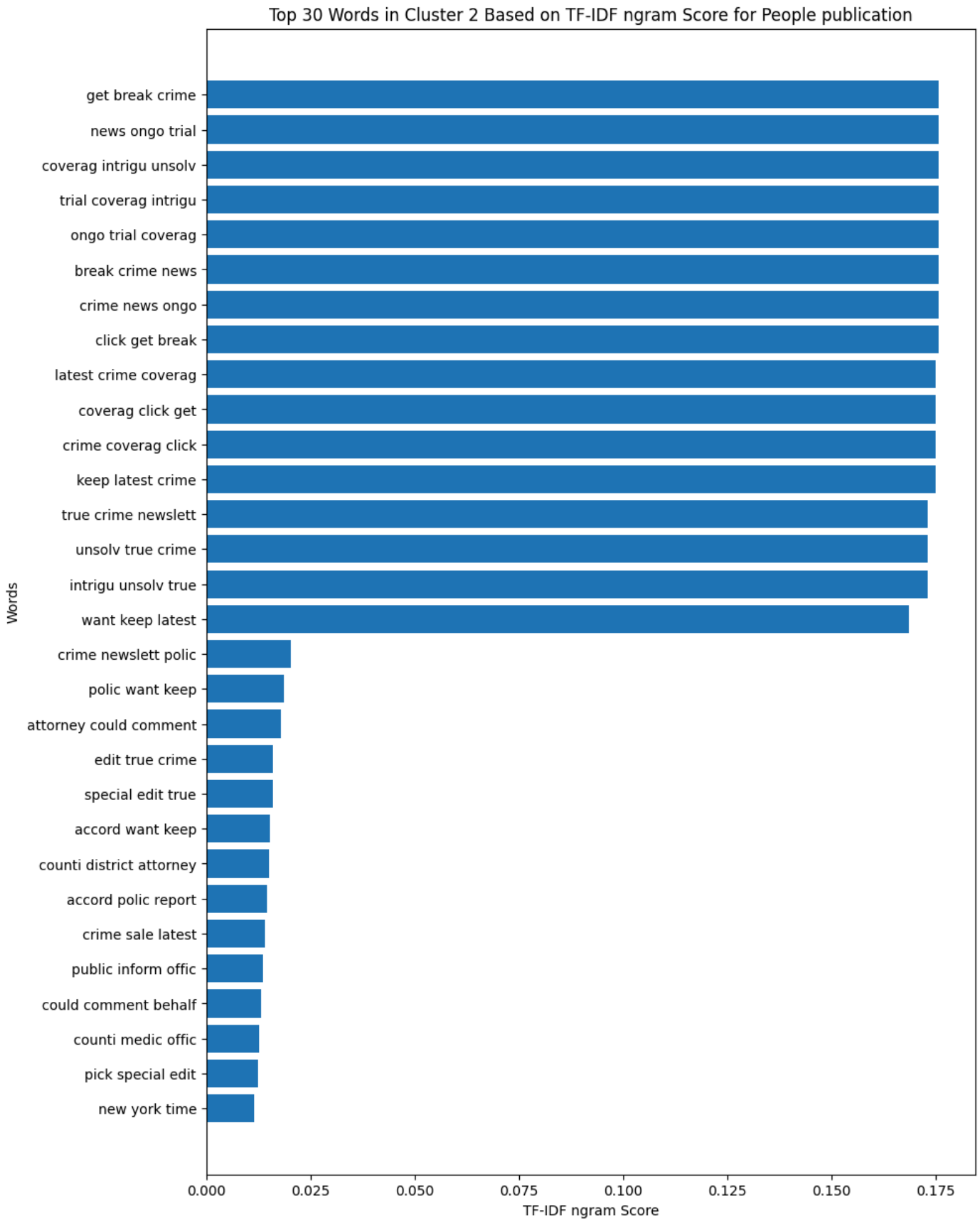
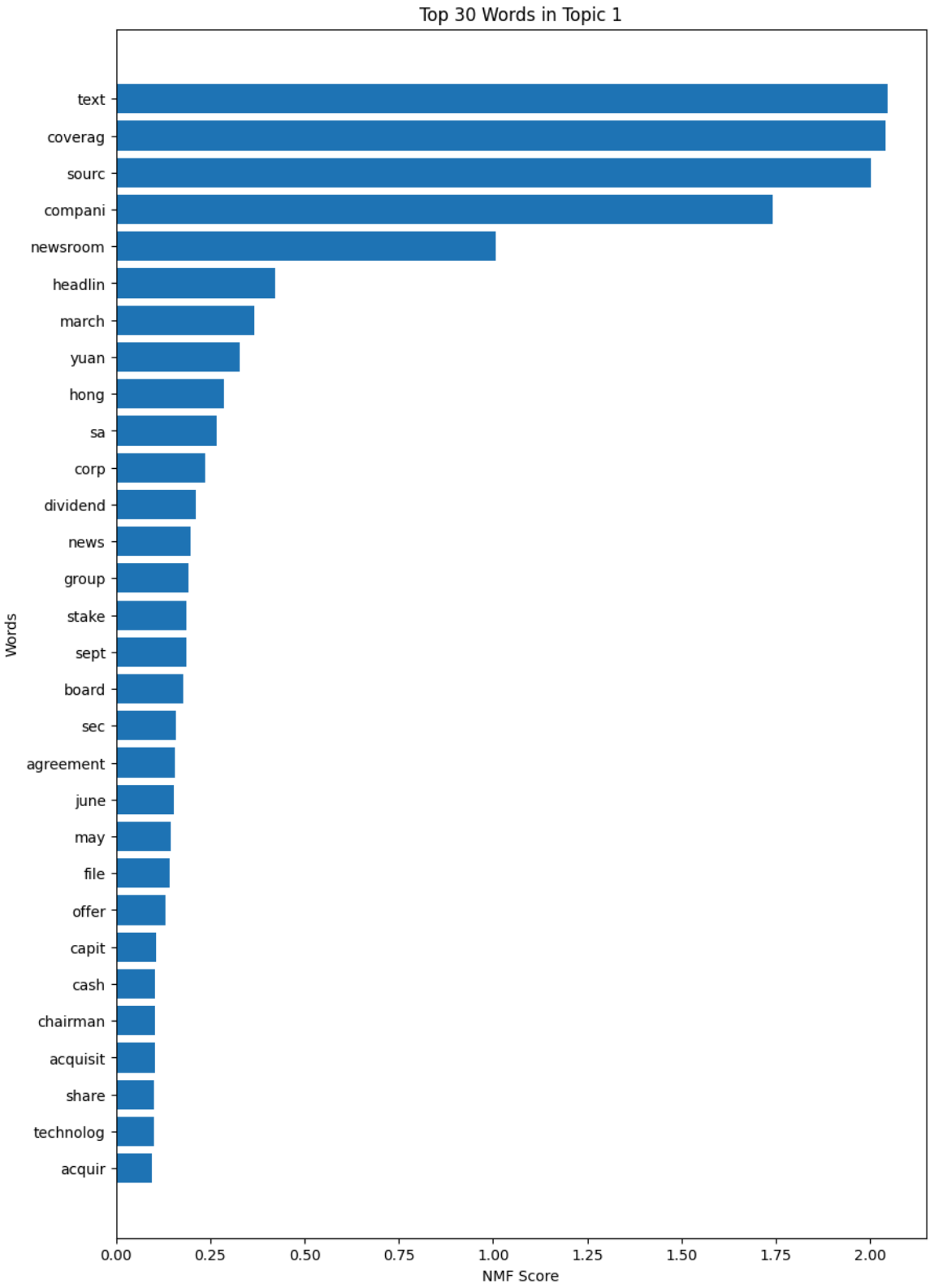
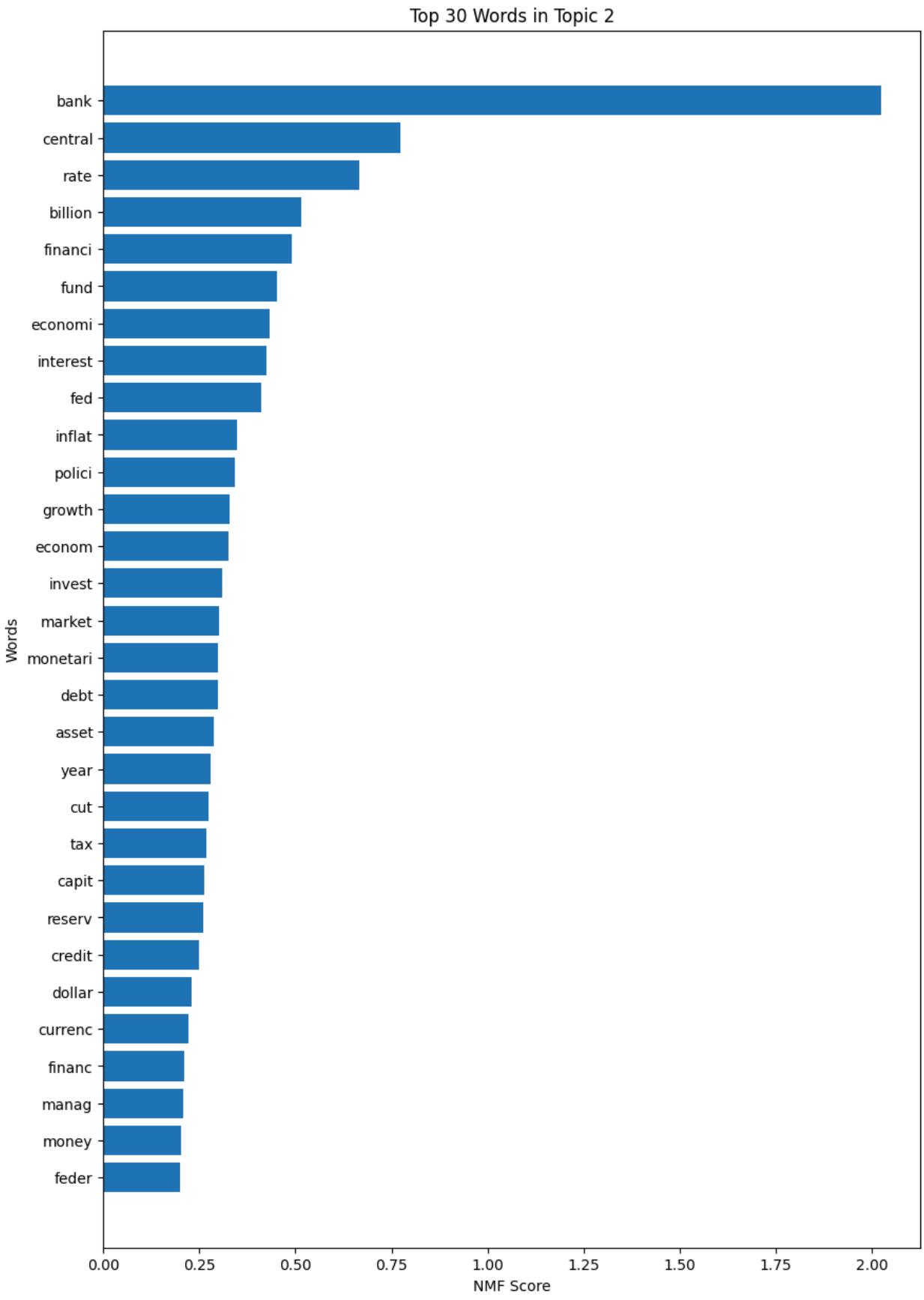
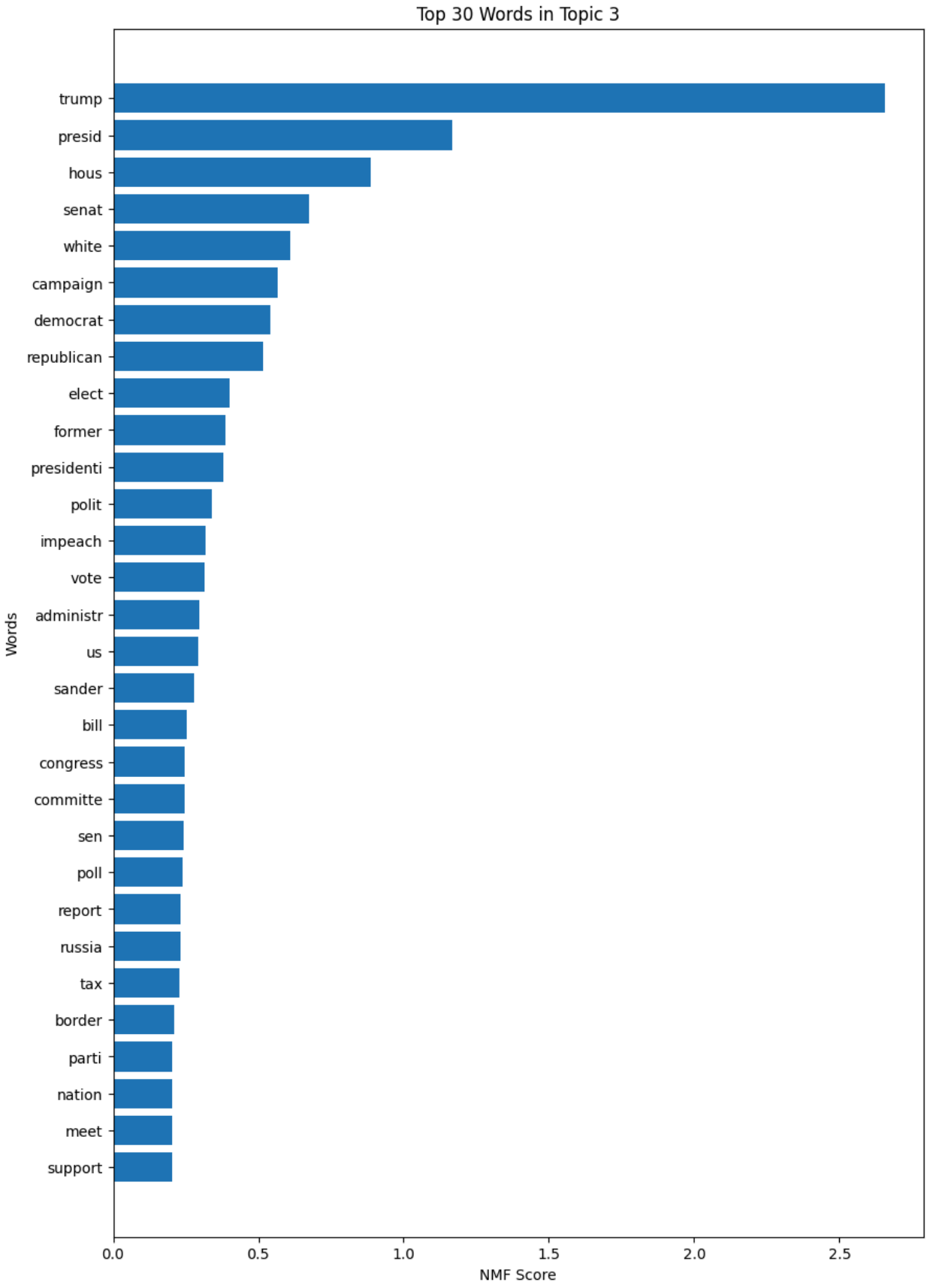
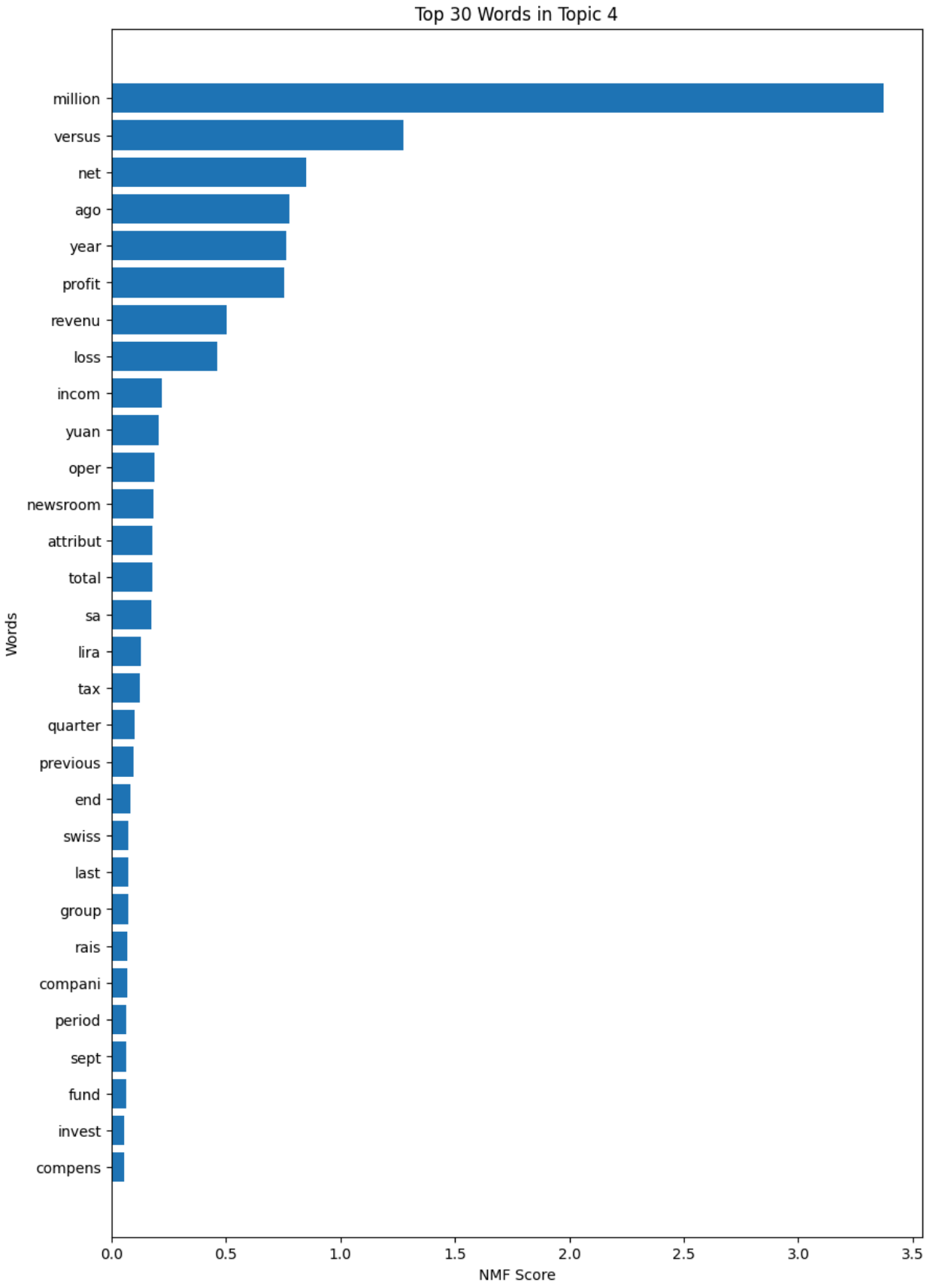

Top 30 Words in Cluster 2 Based on TF-IDF ngram Score for People publication
get break crime
news ongo trial
coverag intrigu unsolv
trial coverag intrigu
ongo trial coverag
break crime news
crime news ongo
click get break
latest crime coverag
coverag click get
crime coverag click
keep latest crime
true crime newslett
unsolv true crime
intrigu unsolv true
want keep latest
crime newslett polic
polic want keep
attorney could comment
edit true crime
special edit true
accord want keep
counti district attorney
accord polic report
crime sale latest
public inform offic
could comment behalf
counti medic offic
pick special edit
new york time
Words
0.000
0.025
0.050
0.075
0.100
0.125
0.150
0.175
TF-IDF ngram Score

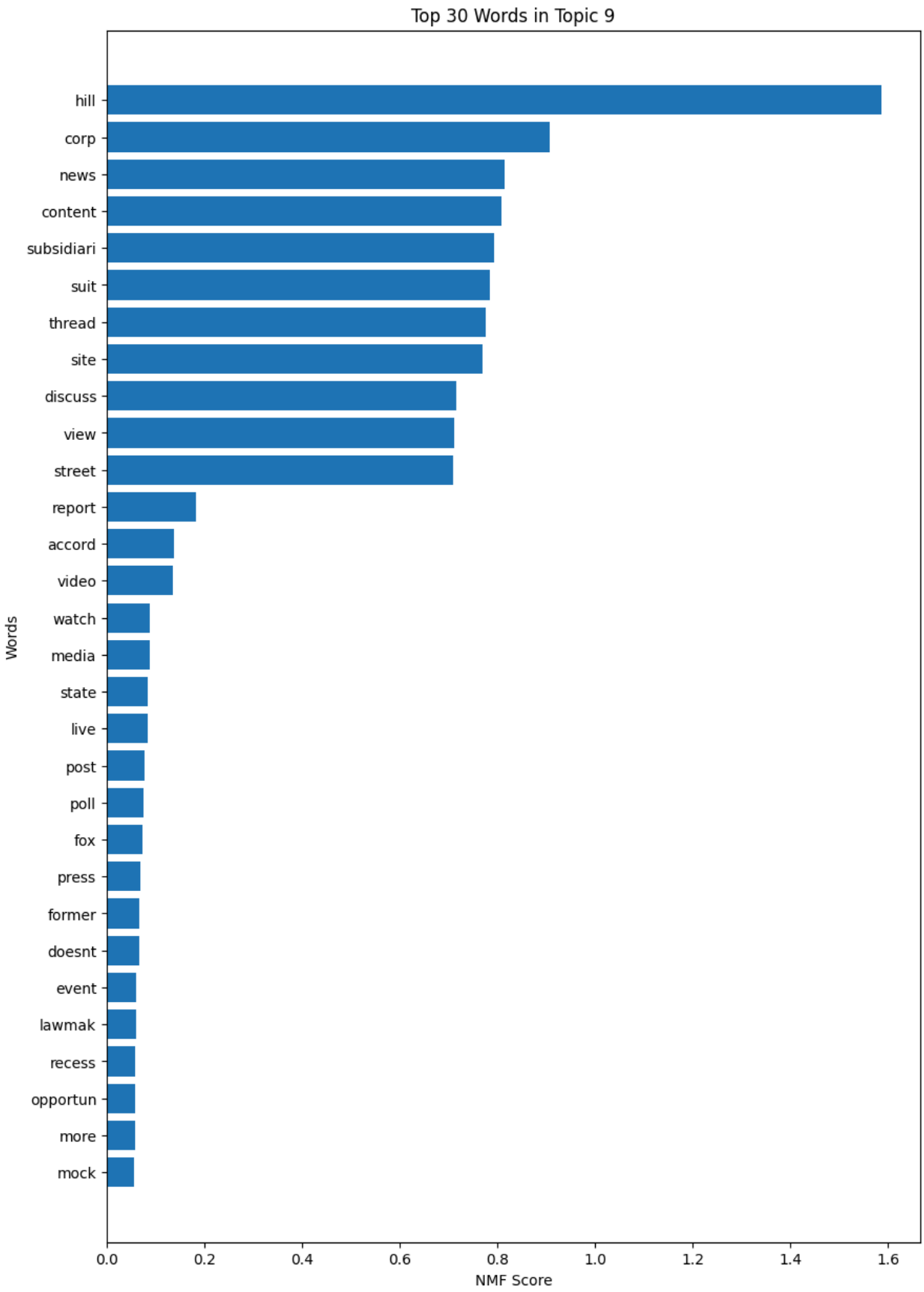
Fig 4.5 NMF topic modelling:

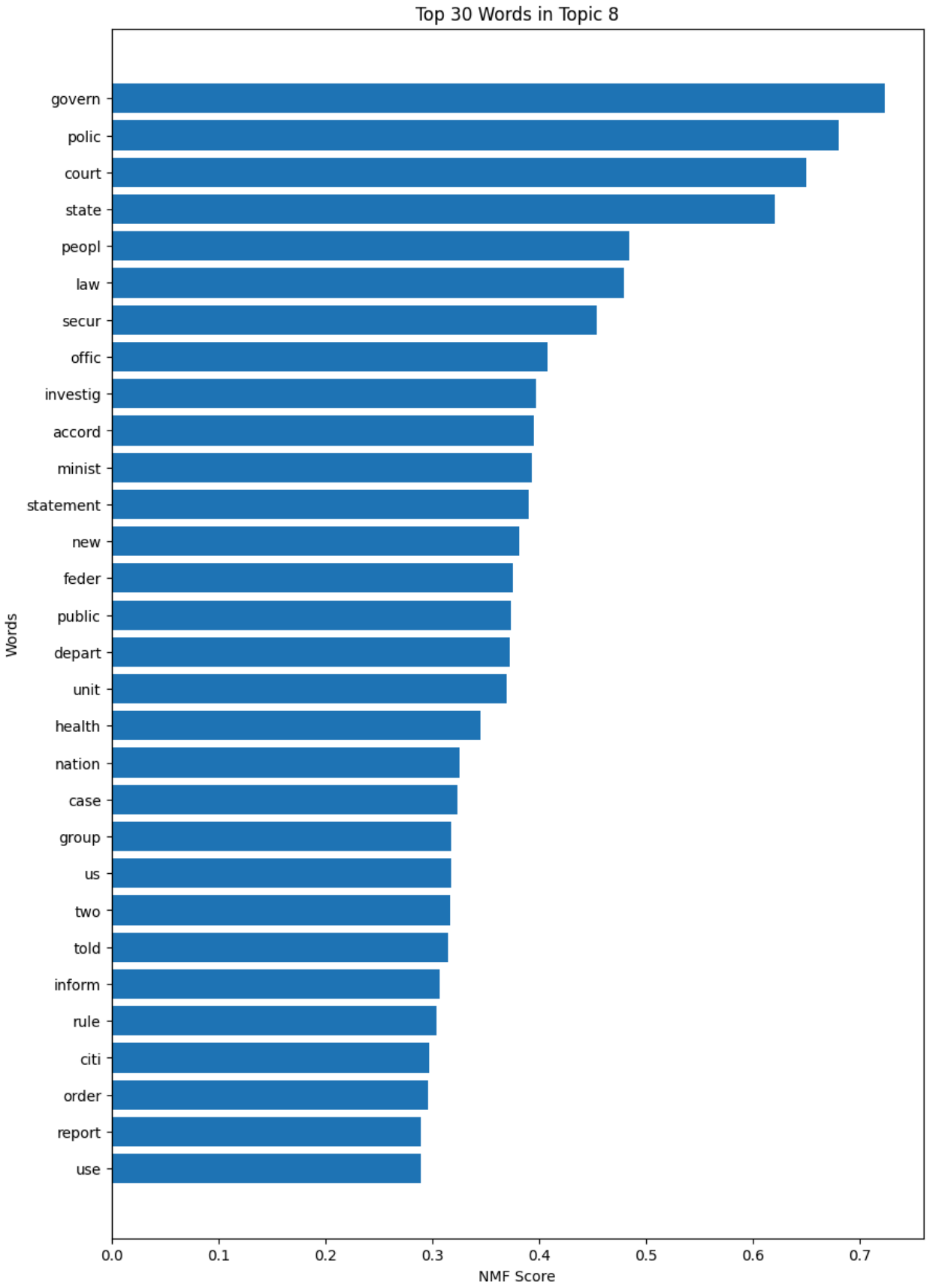


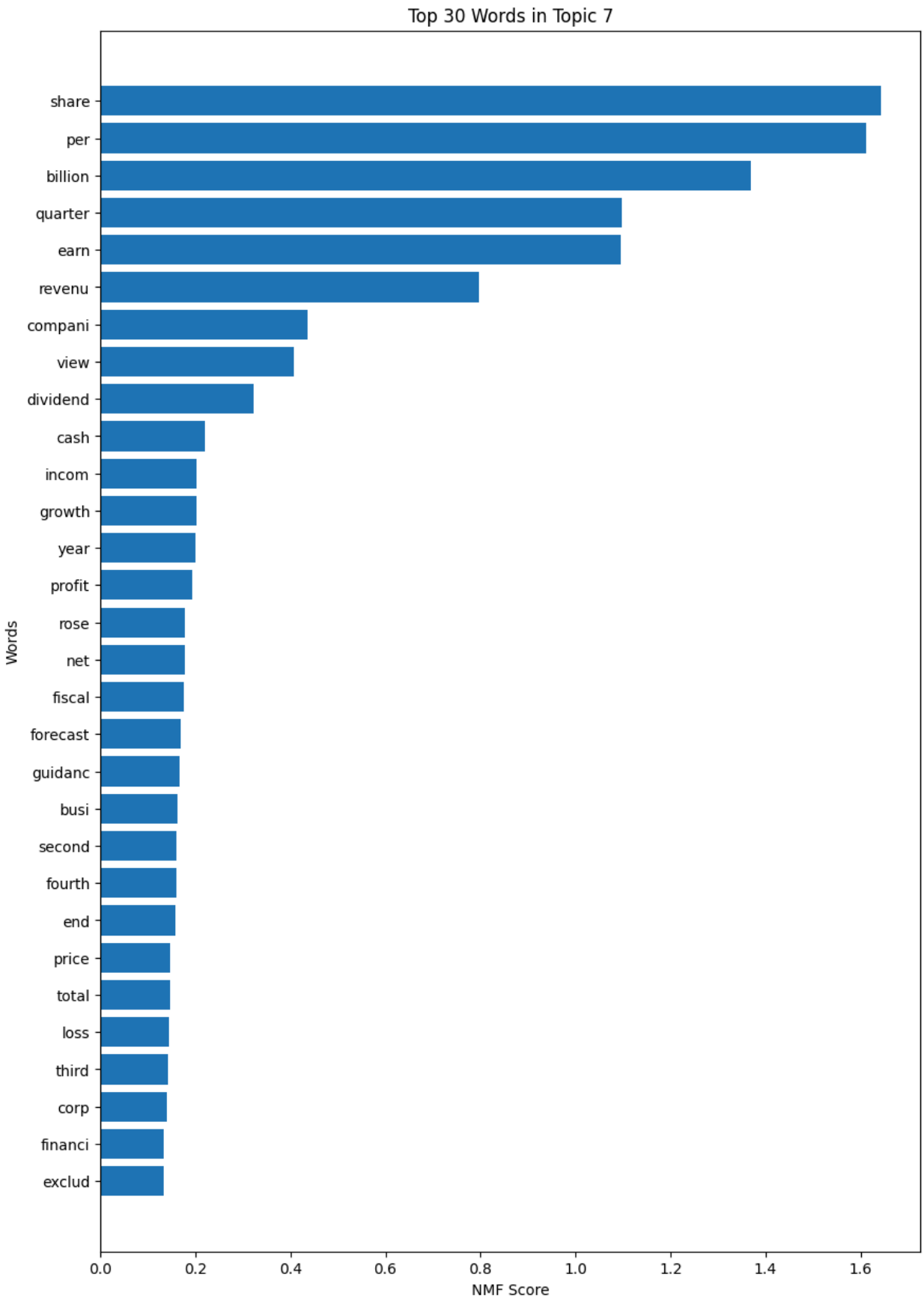

Top 30 Words in Topic 2
Words
bank
central
rate
billion
financi
fund
economi
interest
fed
inflat
polici
growth
econom
invest
market
monetari
debt
asset
year
cut
tax
capit
reserv
credit
dollar
currenc
financ
manag
money
feder
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
NMF Score


Top 30 Words in Topic 3
trump
presid
hous
senat
white
campaign
democrat
republican
elect
former
presidenti
polit
impeach
vote
administr
us
sander
bill
congress
committe
sen
poll
report
russia
tax
border
parti
nation
meet
support
Words
0.0
0.5
1.0
1.5
2.0
2.5
NMF Score


Top 30 Words in Topic 4
million
versus
net
ago
year
profit
revenu
loss
incom
yuan
oper
newsroom
attribut
total
sa
lira
tax
quarter
previous
end
swiss
last
group
rais
compani
period
sept
fund
invest
compens
Words
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
NMF Score


Top 30 Words in Topic 9
Words
hill
corp
news
content
subsidiari
suit
thread
site
discuss
view
street
report
accord
video
watch
media
state
live
post
poll
fox
press
former
doesnt
event
lawmak
recess
opportun
more
mock
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
NMF Score


Top 30 Words in Topic 8
govern
polic
court
state
peopl
law
secur
offic
investig
accord
minist
statement
new
feder
public
depart
unit
health
nation
case
group
us
two
told
inform
rule
citi
order
report
use
Words
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
NMF Score


Top 30 Words in Topic 7
share
per
billion
quarter
earn
revenu
compani
view
dividend
cash
incom
growth
year
profit
rose
net
fiscal
forecast
guidanc
busi
second
fourth
end
price
total
loss
third
corp
financi
exclud
Words
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
NMF Score

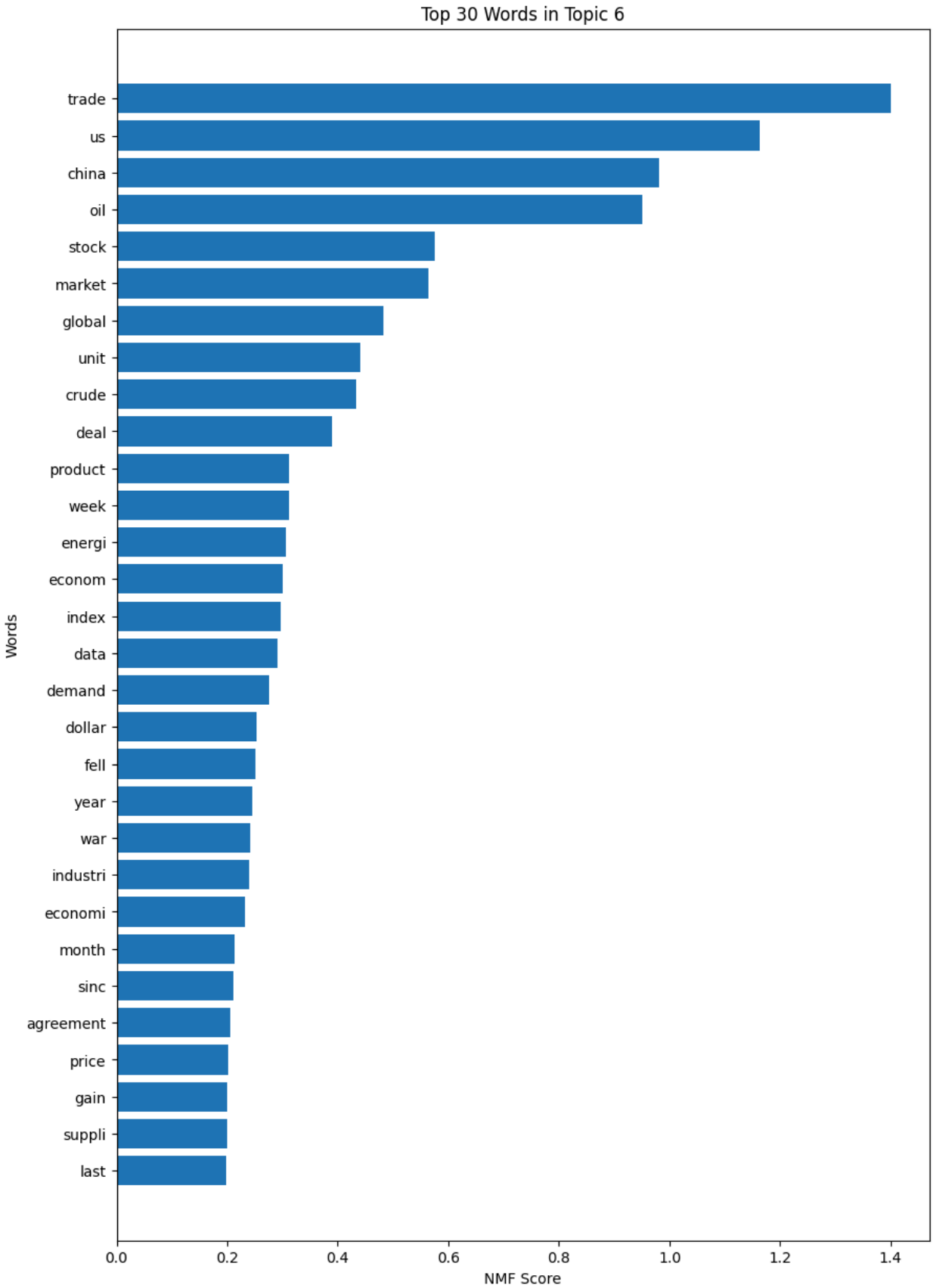

Top 30 Words in Topic 6
Words
trade
us
china
oil
stock
market
global
unit
crude
deal
product
week
energi
econom
index
data
demand
dollar
fell
year
war
industri
economi
month
sinc
agreement
price
gain
suppli
last
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
NMF Score

Fig 4.6 NMF topic modelling-New topic obtained from 2016:

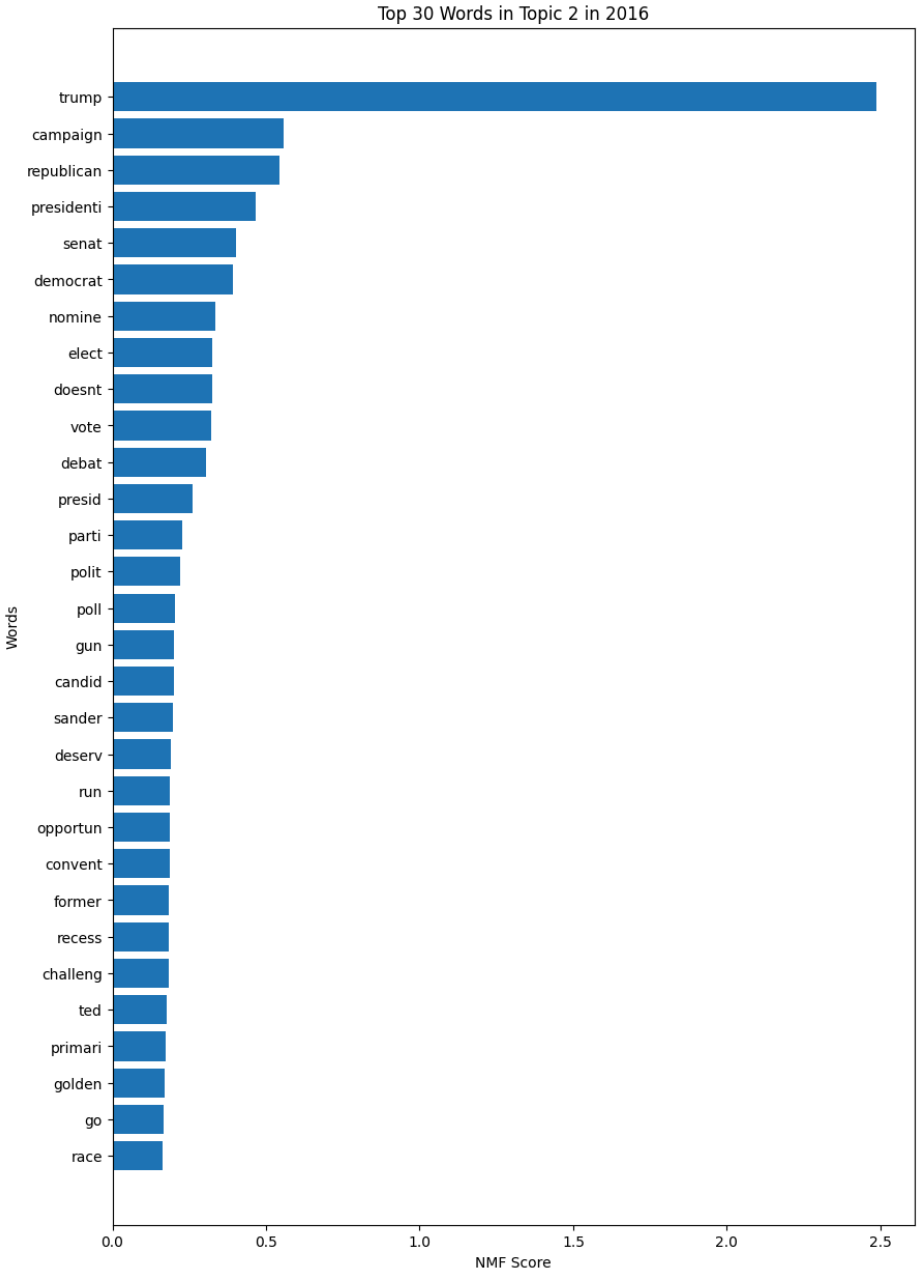


Fig 4.7 NMF topic modelling-New topic obtained from 2017:

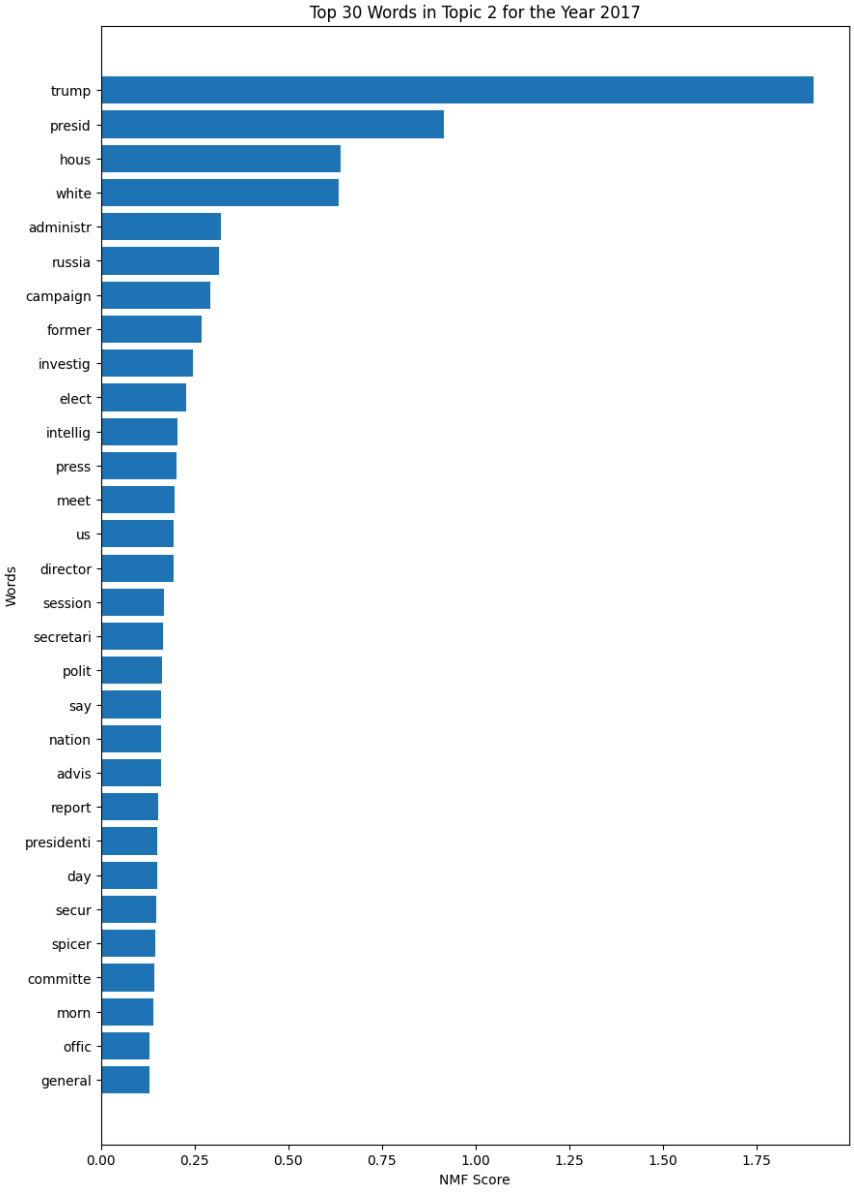


Fig 4.8 NMF topic modelling-New topic obtained from 2018:

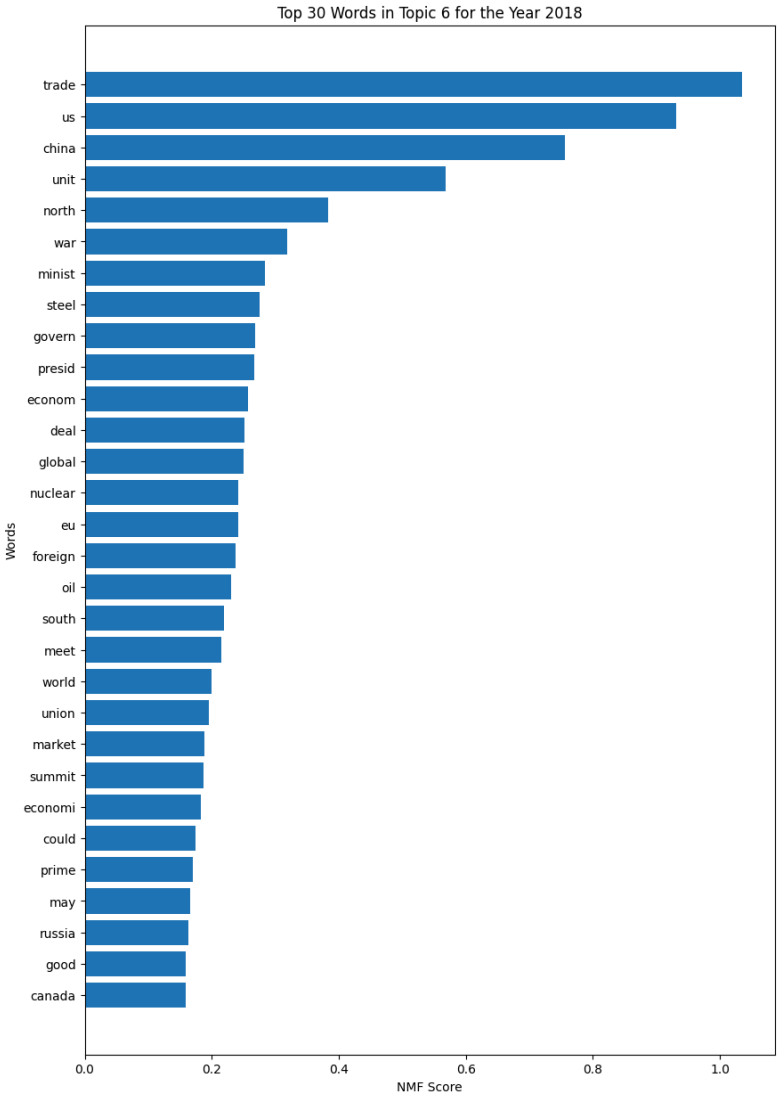


Fig 4.9 NMF topic modelling-New topic obtained from 2019:

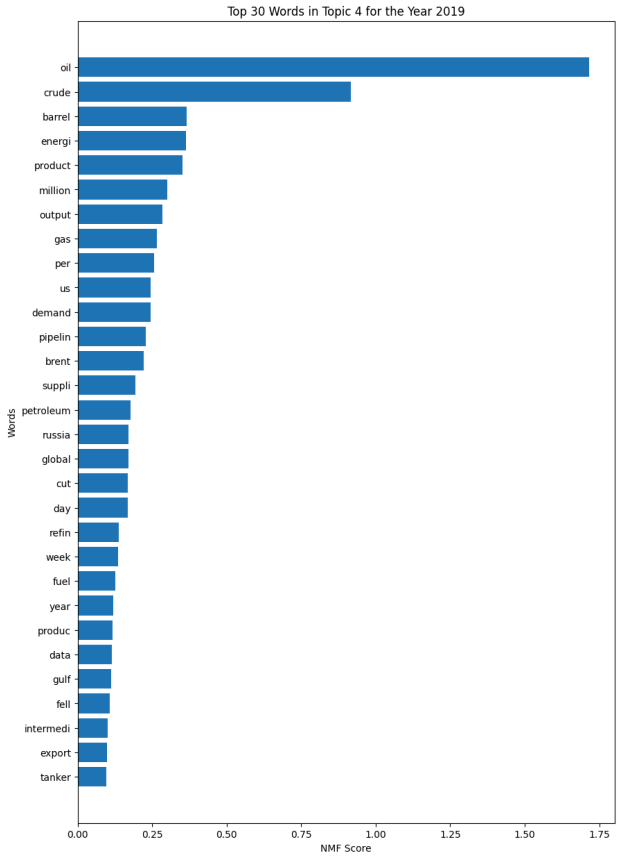


Fig 4.10 NMF topic modelling-New topic obtained from 2020: