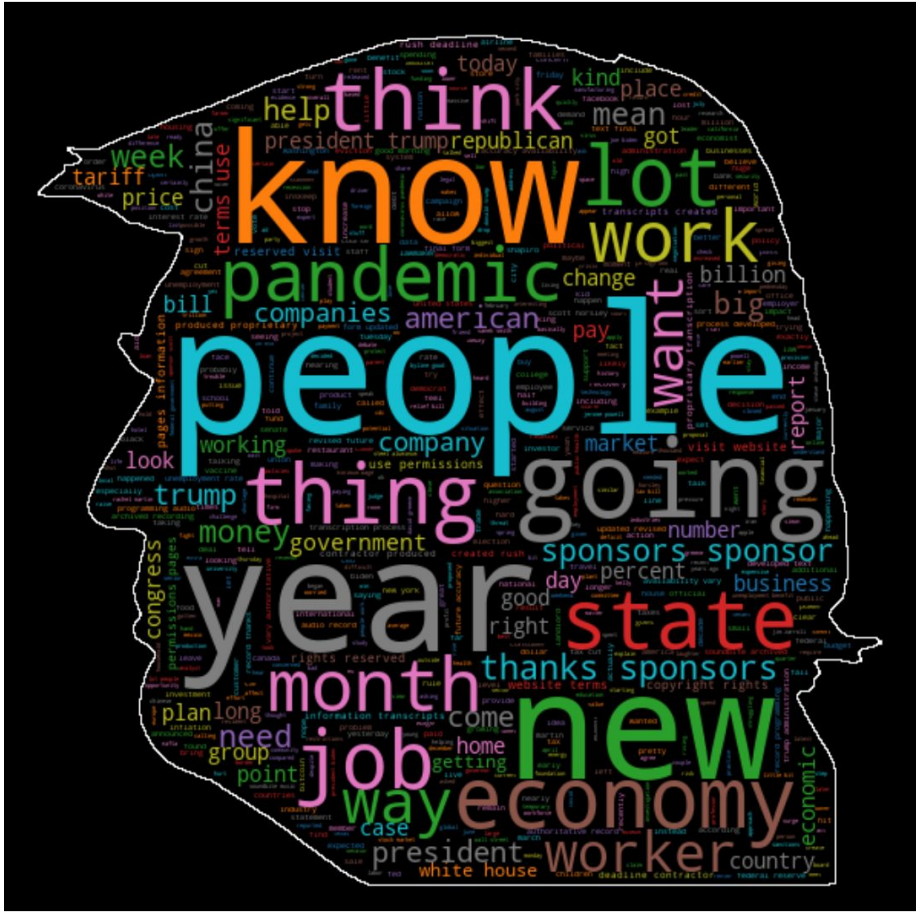
Lost in the News Sauce?

Adventures in classification

By: Oluwaseyi Malonia, Chibuzo Ugonabo , & George Gee

Mission Statement

Assist our non-profit stakeholders in classifying news streams. To filter based on topic for ease of consumption.



The Data

9500+ links scraped from NPR

- Dating back to 2016

1800+ articles scraped

1610 Articles chosen for modeling.

Perfectly balanced

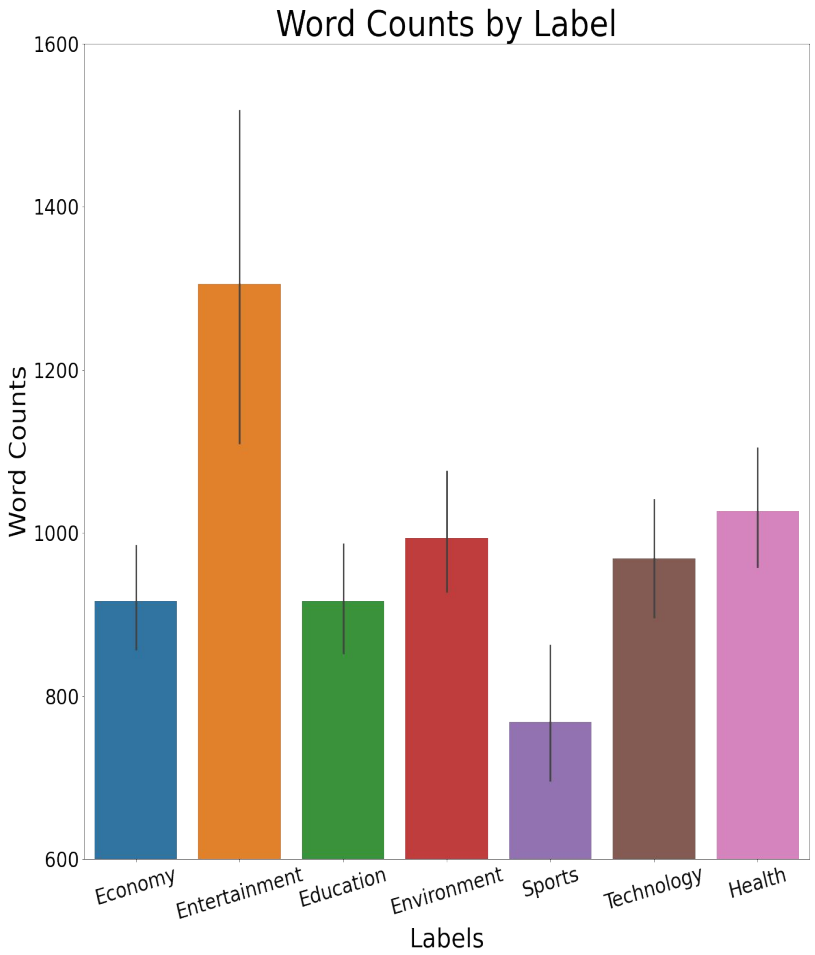
Train, Validation, Test Sets

Trained on 1,050 articles

Validated on 350 articles

Tested on 210 articles

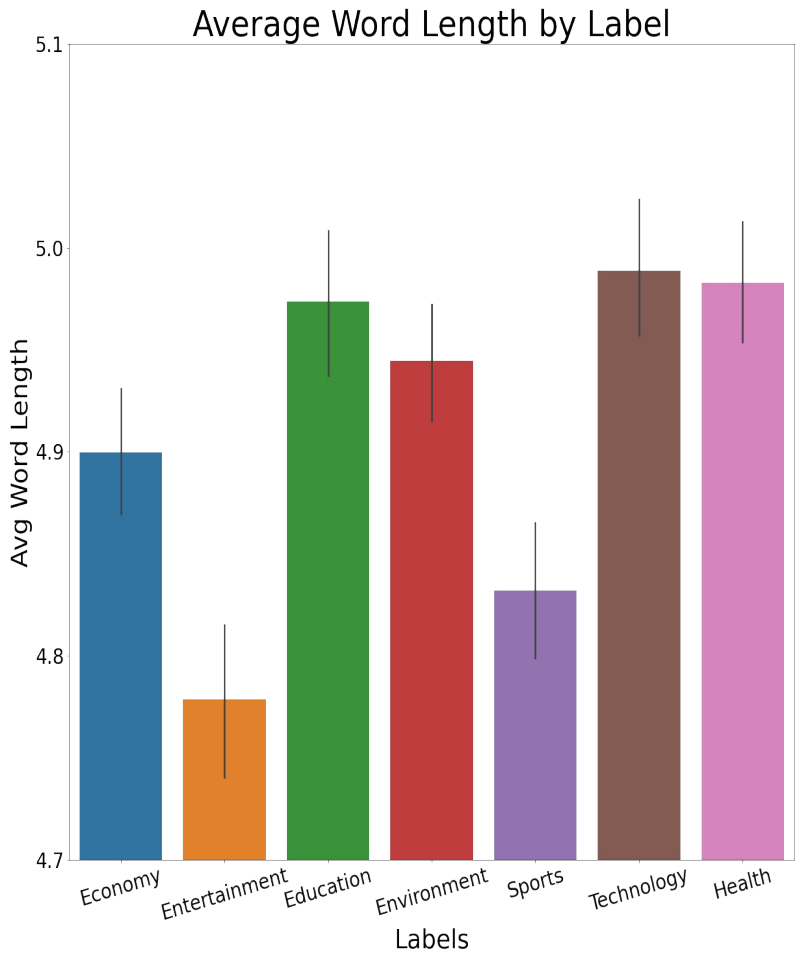
- Topics included:
  - Economy
  - Education
  - Entertainment
  - Environment
  - Health
  - Sports
  - Technology



Average Word Count per Label.

Sports averaged the lowest amount of word counts.

Entertainment average the highest amount of word counts.

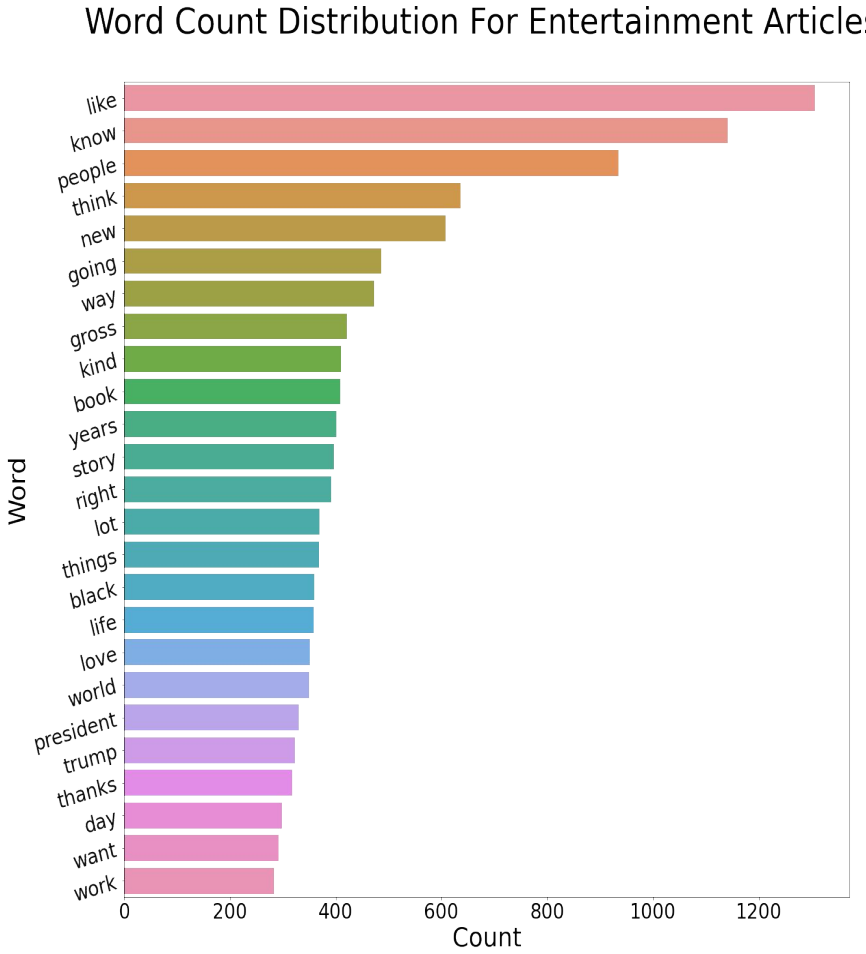


Average Word Length per Label.

Entertainment uses smaller words.

Technology uses larger words.

Likely a direct correlation with the target audiences education level.

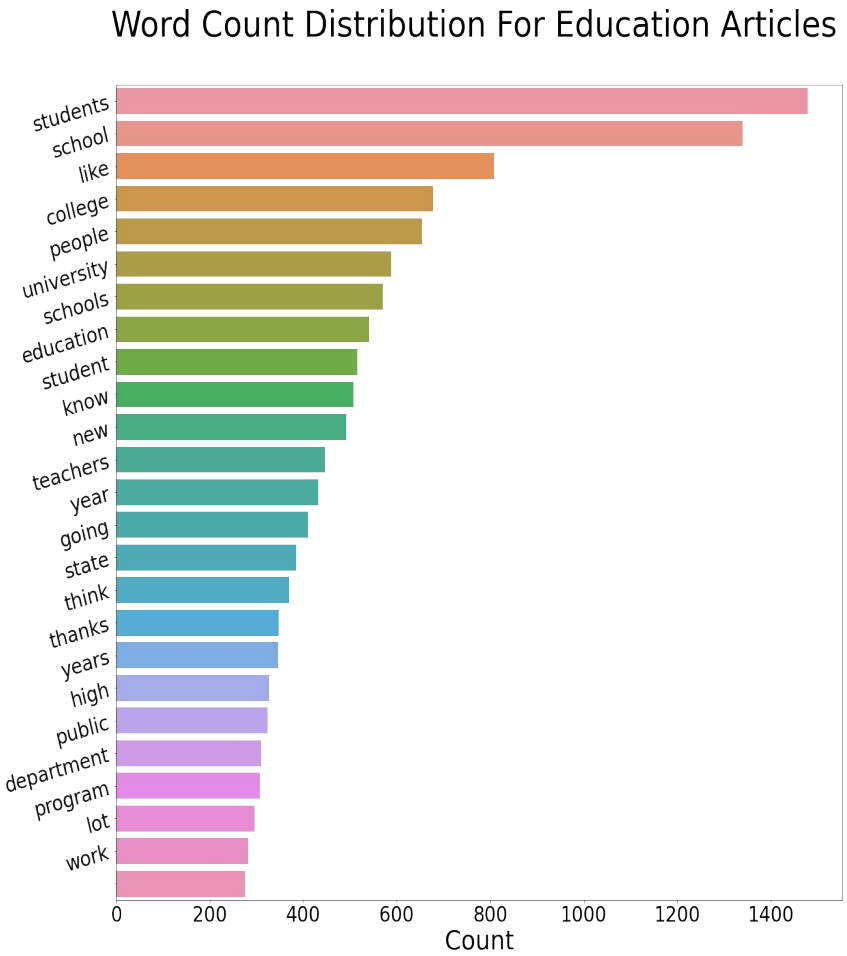


Entertainment Word Frequencies

Entertainment contains buzzwords such as ‘book’, ‘story’, ‘like’, and ‘love’.

Contains oddities like President Trump…

Say what you will but the oval office had never been so “entertaining”.



Education Word Frequencies

Education contained mostly expected buzzwords such as:

- Student
- University
- School
- Teachers

Data Processing

Text Processing

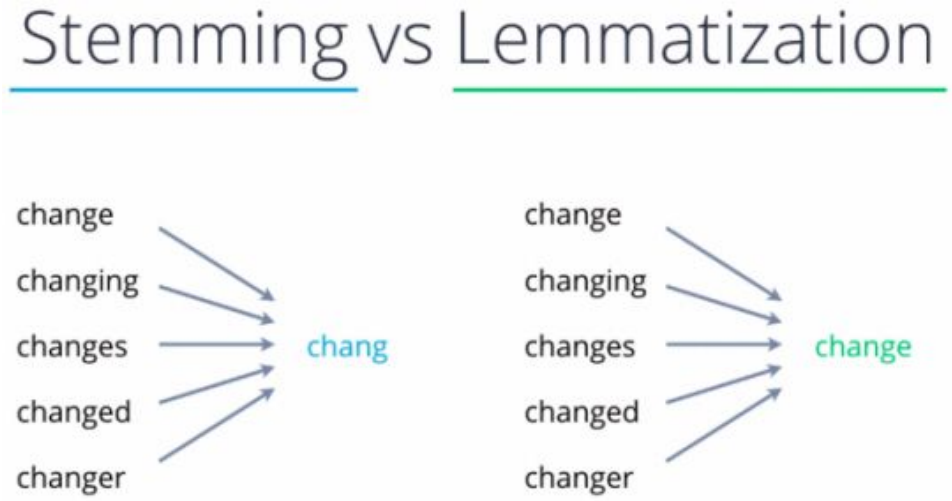
Stripped of all HTML meta-data.

Stripped of embedded hyperlinks and emails.

Stripped of special characters, digits, punctuation, and stop words.

All text converted to lowercase.

Stemming and Lemmatization



Modeling

TF-IDF

Term Frequency - Inverse Document Frequency

Numerical statistic to reflect how important a word is to a document in a collection of documents.

How many times a term appears in a given document vs how many documents that term is present in.

Utilized the lemmatization linguistic morphology.

params{min_df= , max_df= , ngram_range= max_features= ,}

Random Forest

Ensemble Learning for Classification

Random Forest is an ensemble of single decision trees to create a forest of predictors.

Robust against the overfitting issues of a single decision tree.

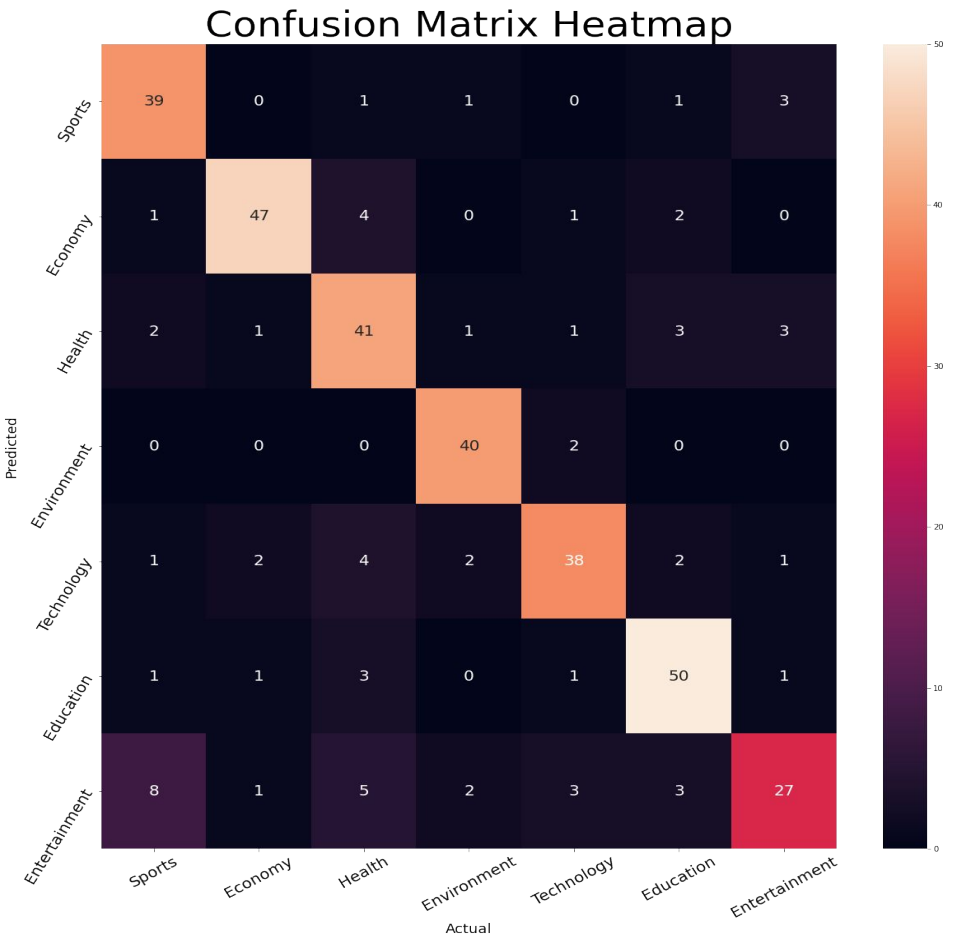
params{max_depth= , max_features= , n_estimators= , min_samples_leaf= , min_samples_split= }

Confusion Matrix

High misclassification between technology and economy.

A number of economic articles discuss F.A.A.N.G. companies.

Which are all organizations entrenched in technology.



Performance

Training and Validation:

Trained on 1050 articles

Validated on 350 articles

Accuracy of 81%

F1 Score of 80%

Testing:

Tested on 210 articles

Accuracy of 80%

F1 Score of 80%

Next Steps

Deploy model with continued supervision.

Hire interns to confirm the labels are correct.

Scrape more articles and continuously train the model.

Thank You!!

Questions???

Co

mments!!!

Appendix

Stemming and Lemmatization

Stemming is the process of converting a word into its stem, base, or root form. Can often be a form that doesn't exist in the dictionary.

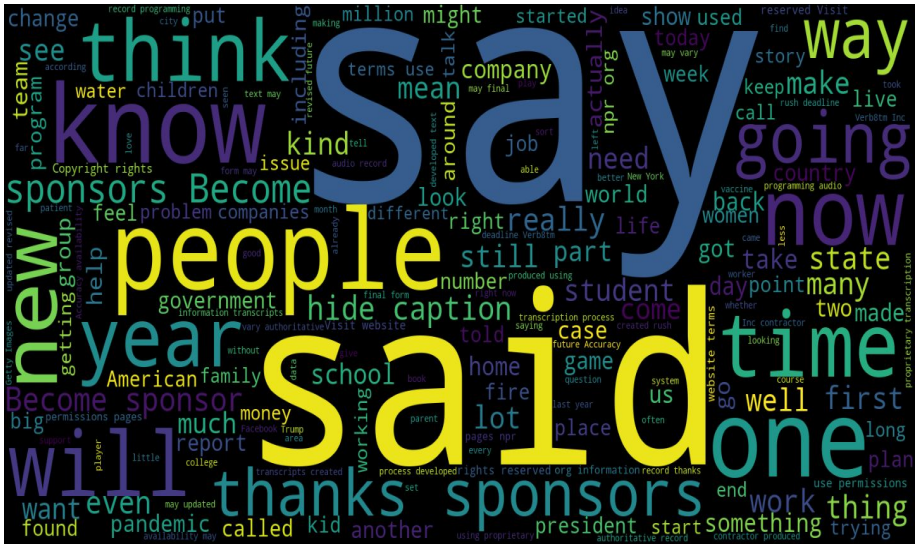
Lemmatization is the process of grouping together the inflected forms of a word and converting them to a common lemma or dictionary form.

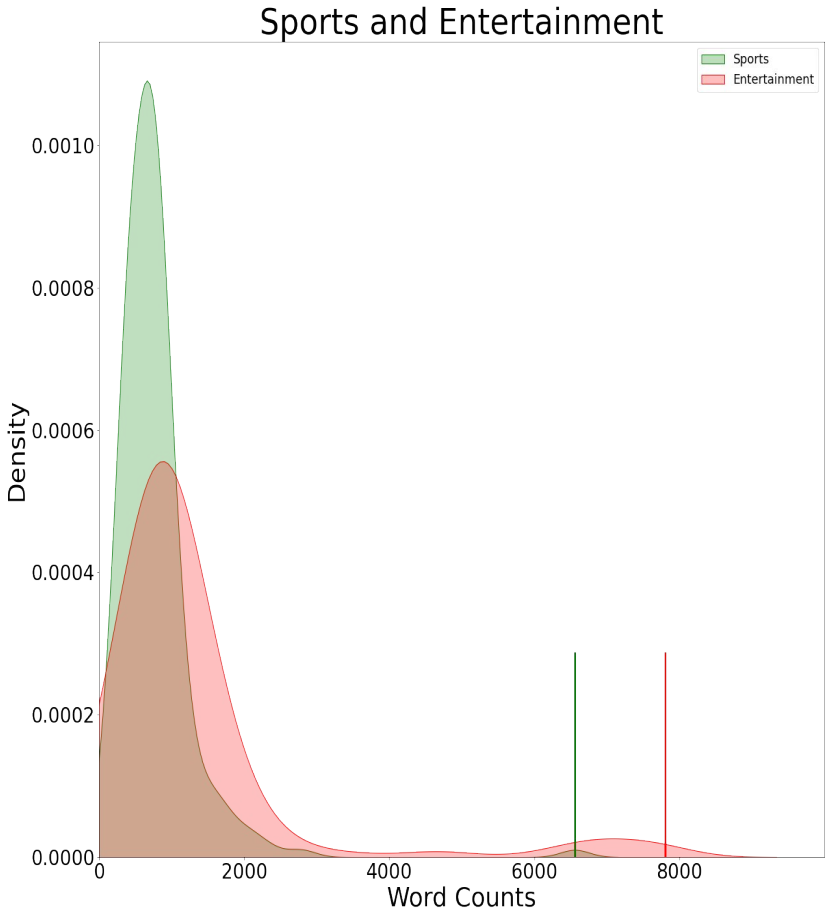
GenSim Summarization

In an effort to decrease the potential abundance of superfluous sentences that'll noisey up the model we implemented a summarization algorithm.

Summarization is based on ranking the sentences via the TextRank algorithm. TextRank is a graphical based ranking model to find relevant sentences and keywords.

Params{ ratio=.2 & word_count=375 }

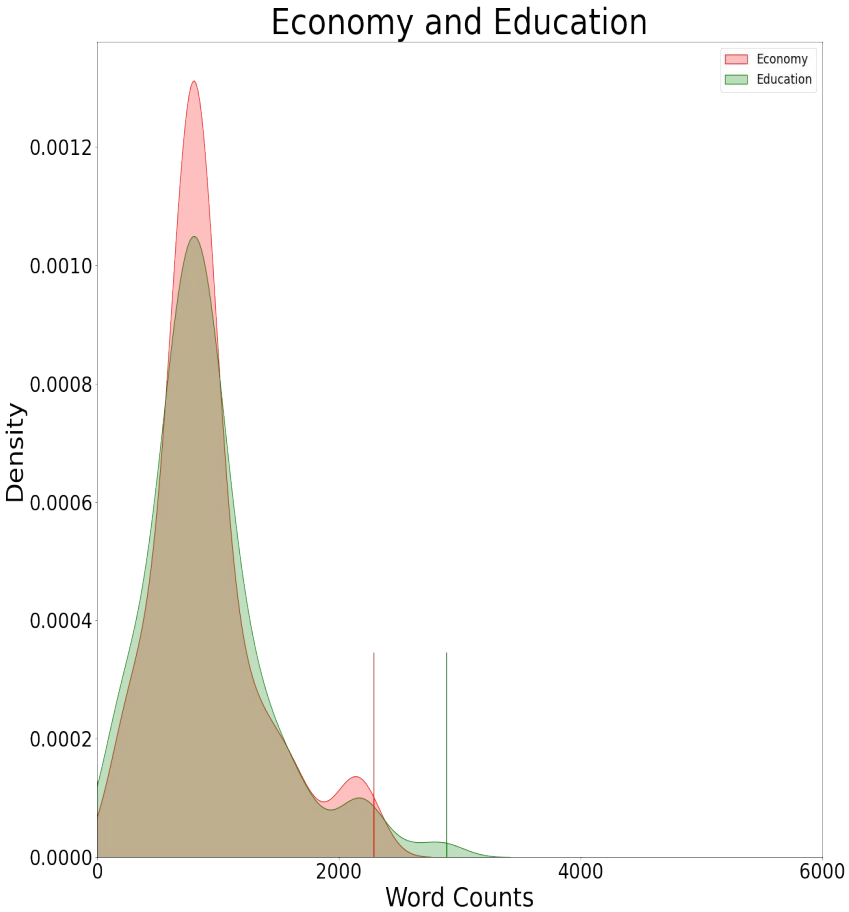




Sports vs Entertainment

Sports had the lowest average word count while entertainment had the highest average word count per article.

Most articles for both categories were below 2000 but extreme outliers existed for both.

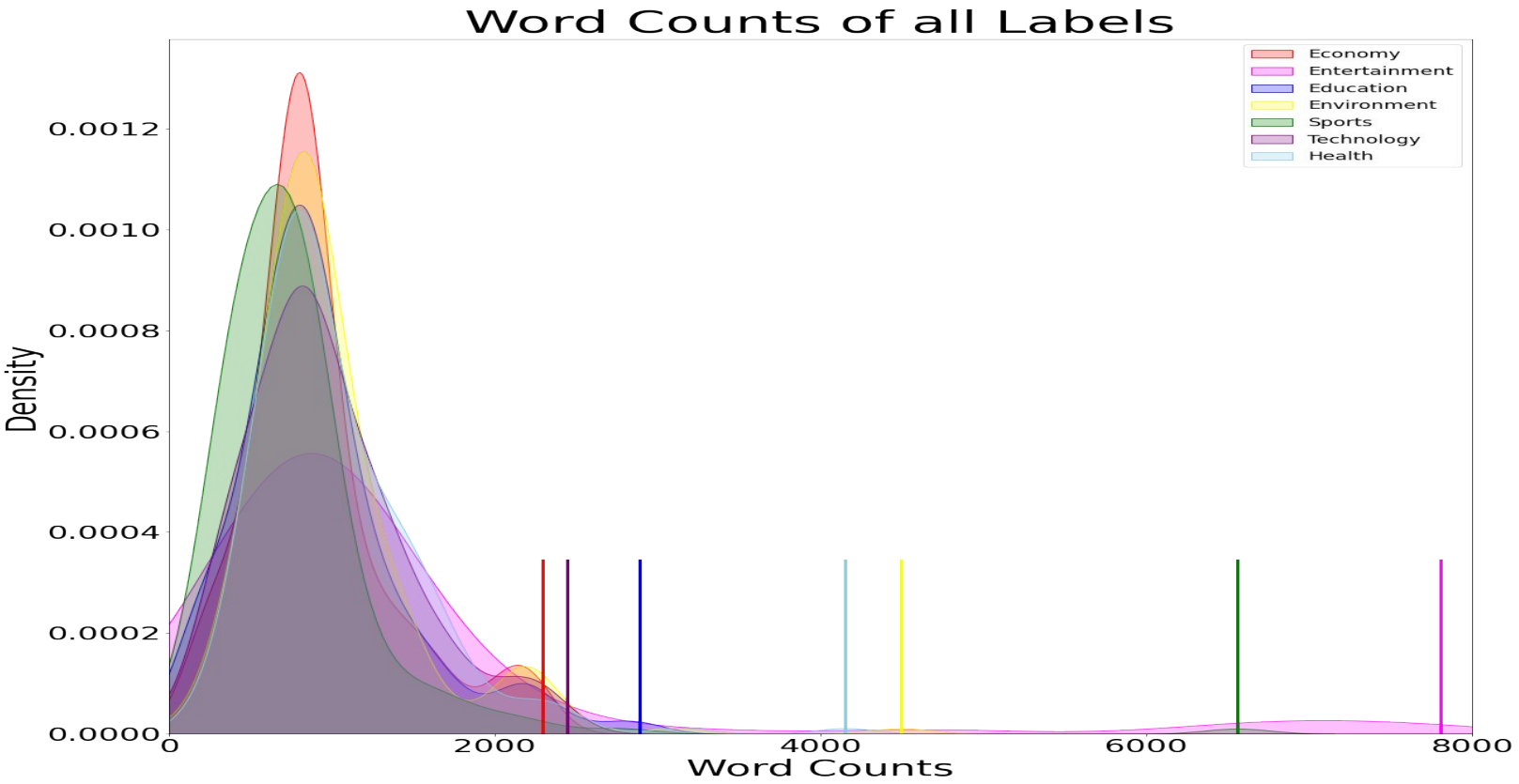


Economy vs Education

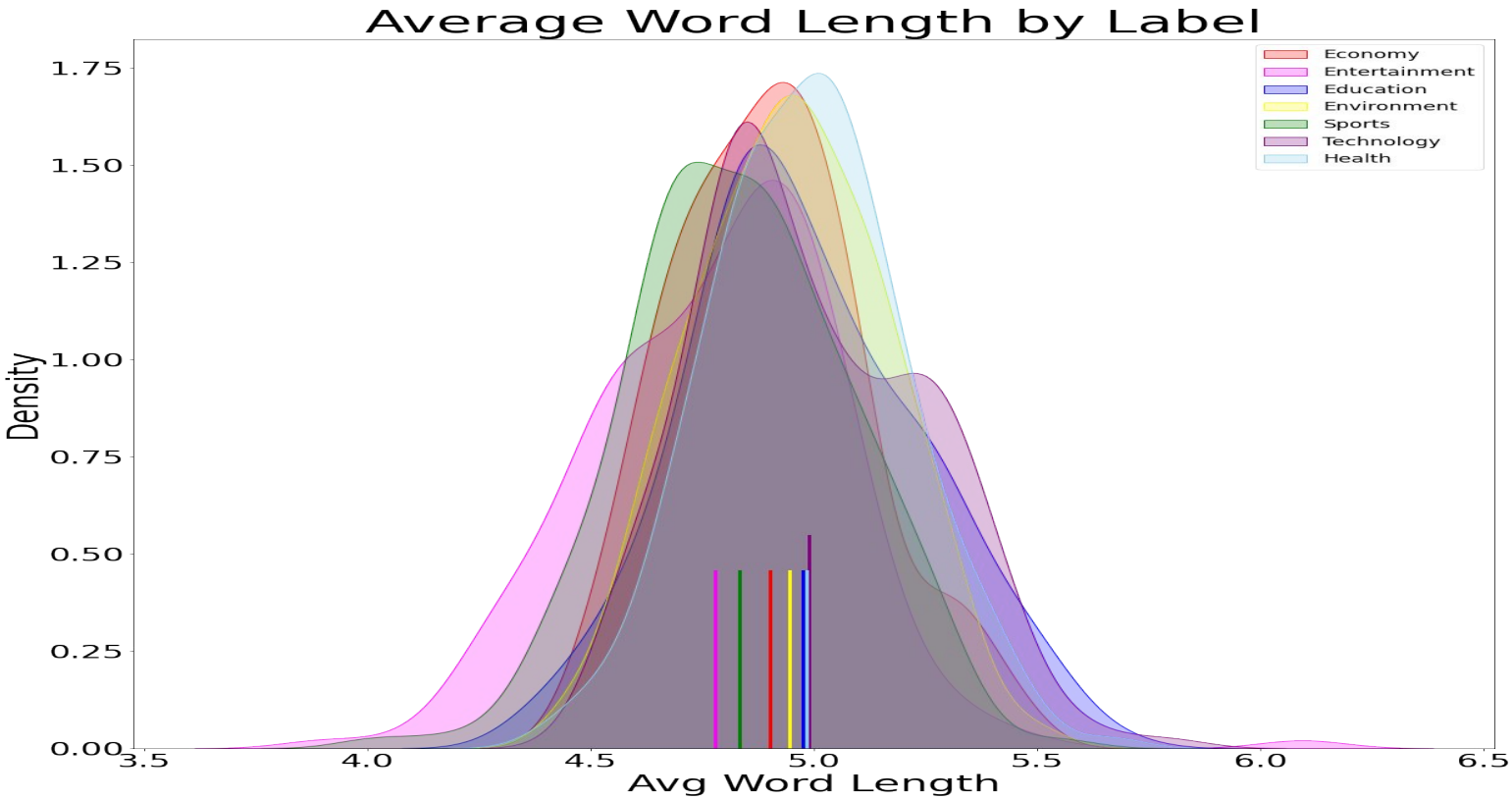
Very similar in terms of average word count per article as such the distributions are also quite similar.

Education outliers led to the expansion of the distribution.

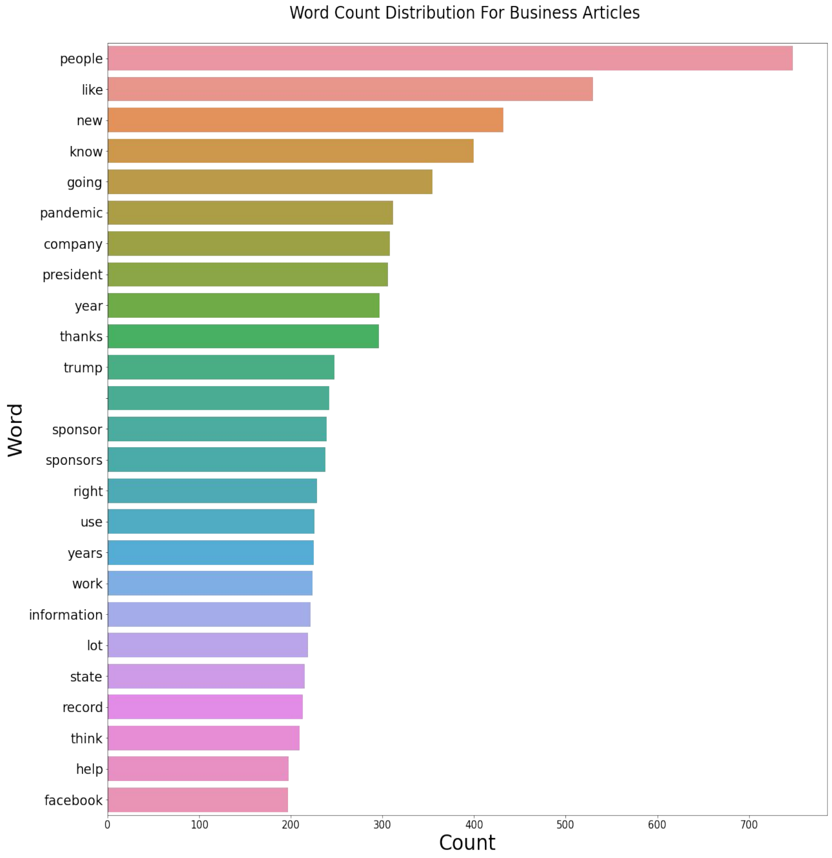
Average Word Counts (All Labels)



Average Word Lengths (All Labels)

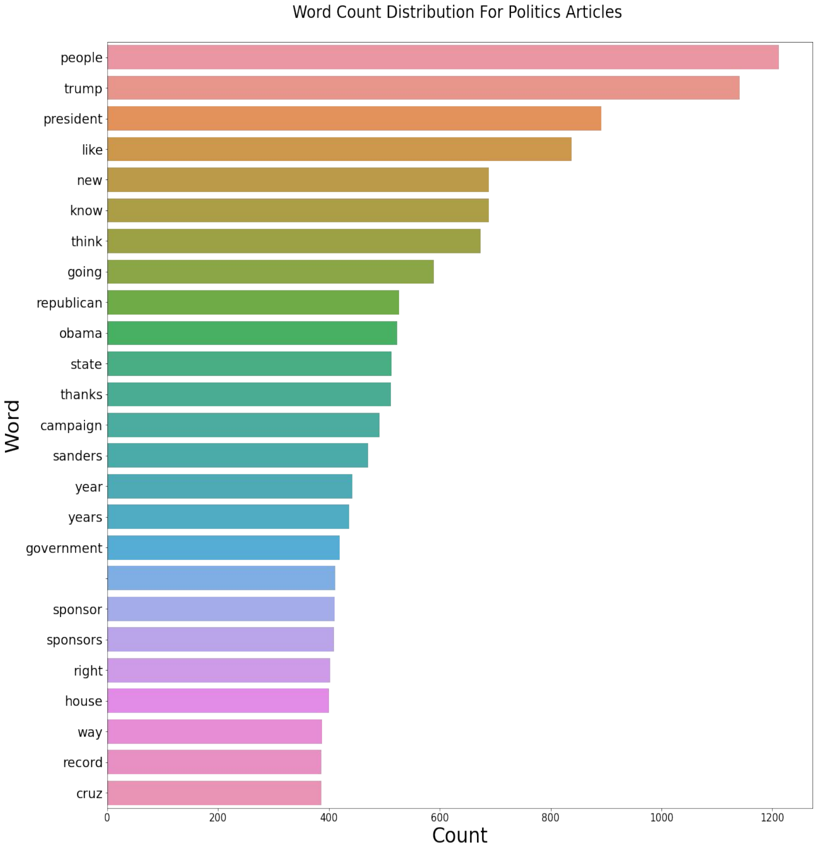


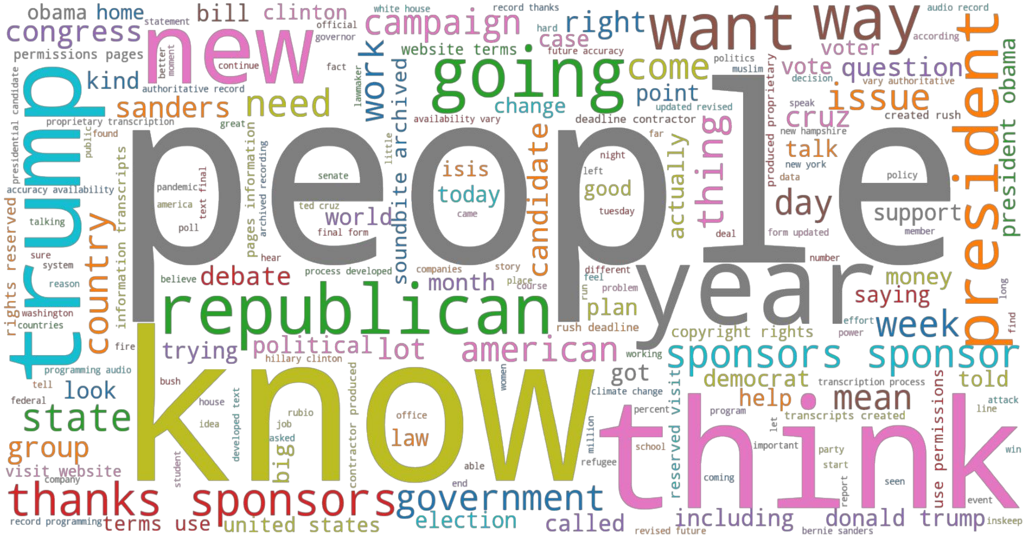
Business


Word Count Distribution For Business Articles
Word
people
like
new
know
going
pandemic
company
president
year
thanks
trump
sponsor
sponsors
right
use
years
work
information
lot
state
record
think
help
facebook
0
100
200
300
400
500
600
700
Count

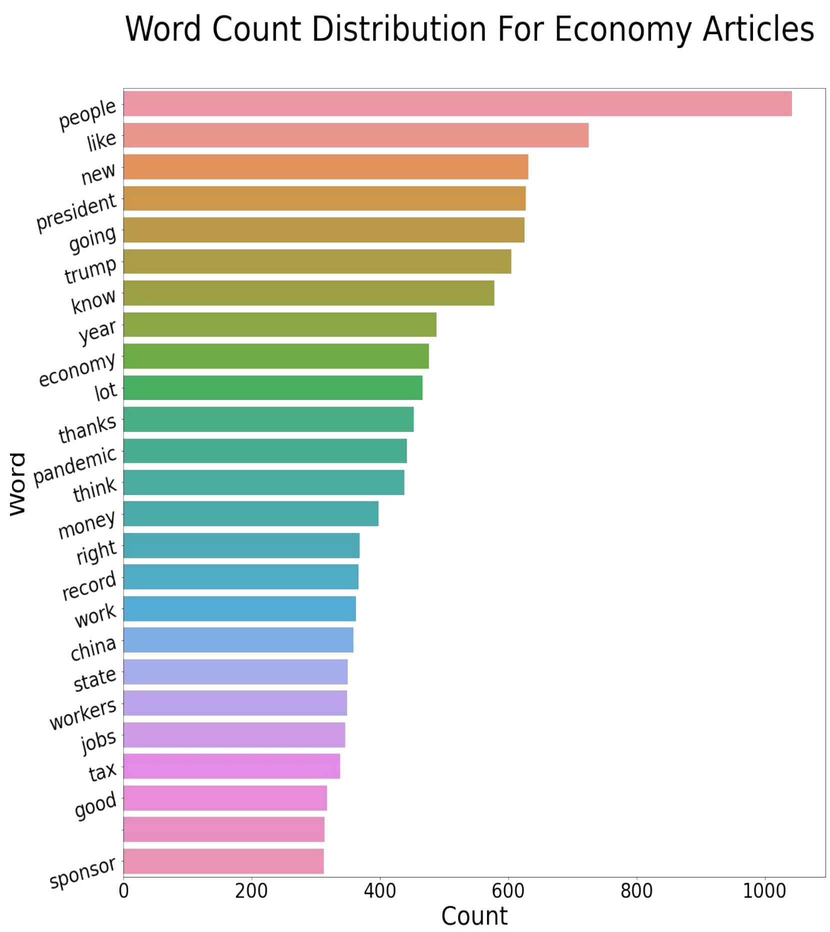

new
year
people
company
state
work
know
going
pandemic
sponsors
thanks
president
including
million
economy
case
service
home
week
market
trump
day
big
amazon
group
business
vaccine
congress
world
american
money
need
way
program
called
kind
want
pay
election
told
house
report
job
worker
plan
bill
help
come
working
thing
system
problem
mean
facebook
tiktok
lot
month
saying
think
billion
place
companies
right
able
point
likely
got
look
making
statement
actually
water
country
question
according
government
long
sponsor
change

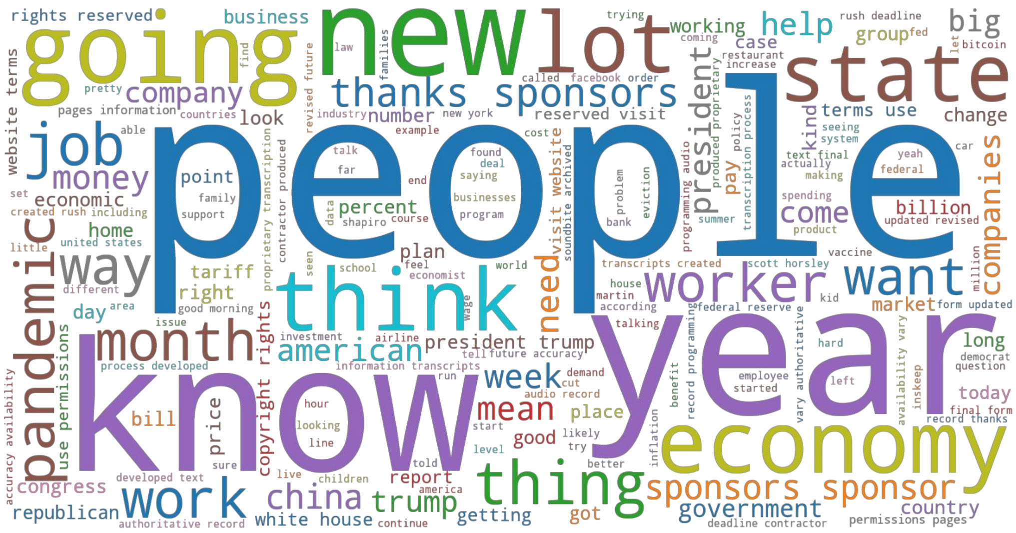
Politics


Word Count Distribution For Politics Articles
people
trump
president
like
new
know
think
going
republican
obama
state
thanks
campaign
sanders
year
years
government
sponsor
sponsors
right
house
way
record
cruz
0
200
400
600
800
1000
1200
Count
Word

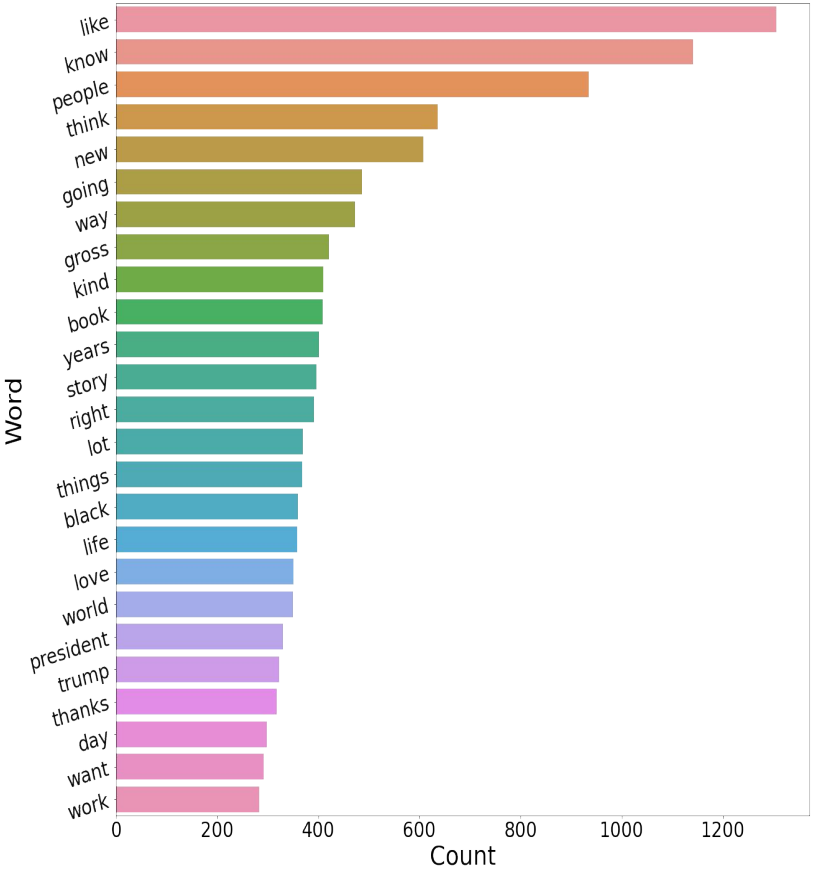

people
know
think
new
trump
republican
year
president
going
want
way
thanks
sponsors
government
sponsor
campaign
congress
obama
home
bill
clinton
right
need
sanders
kind
state
group
country
candidate
thing
issue
question
vote
voter
cruz
talk
day
support
american
political
lot
debate
month
world
today
isis
good
plan
money
saying
week
democrat
help
mean
told
look
law
election
called
including
donald trump
united states
terms
use
trying
got
come
change
point
work
actually
visit website

Economy

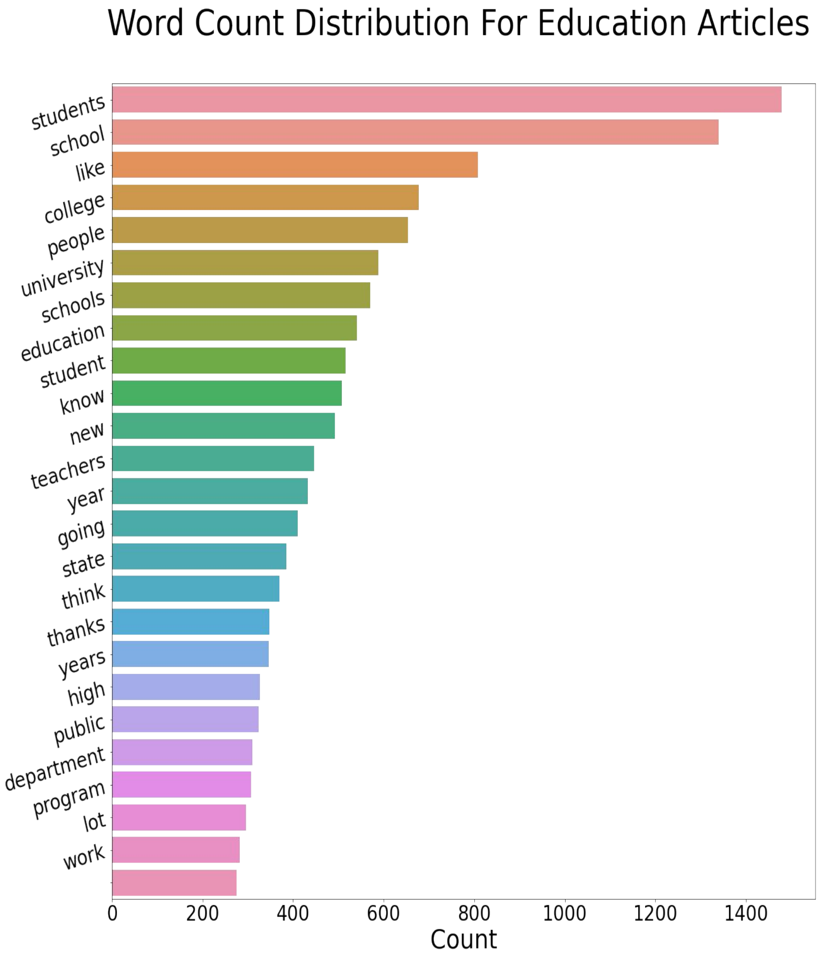

Word Count Distribution For Economy Articles
people
like
new
president
going
trump
know
year
economy
lot
thanks
pandemic
think
money
right
record
work
china
state
workers
jobs
tax
good
sponsor
Word
0
200
400
600
800
1000
Count


people
know
year
going
new
lot
state
think
economy
thing
job
month
pandemic
way
money
worker
want
president
thanks sponsors
company
american
work
china
trump
sponsors sponsor
government
companies
need
week
mean
help
business
rights reserved
working
big
congress
republican
white house
country
bill
price
copyright
today
long
place
billion
come
percent
home
tariff
right
day
plan
report
got
getting
market
economic
point
number
look
change
terms use

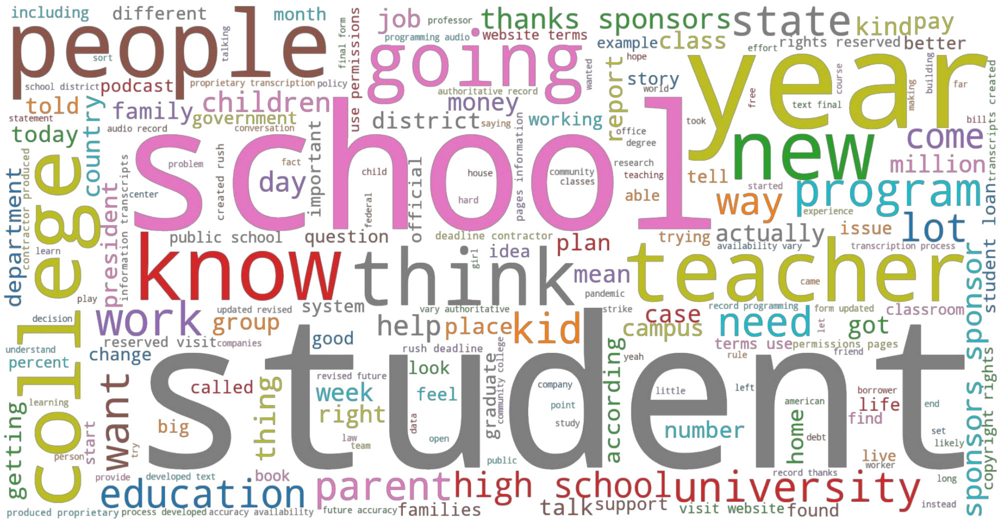
Entertainment



Education

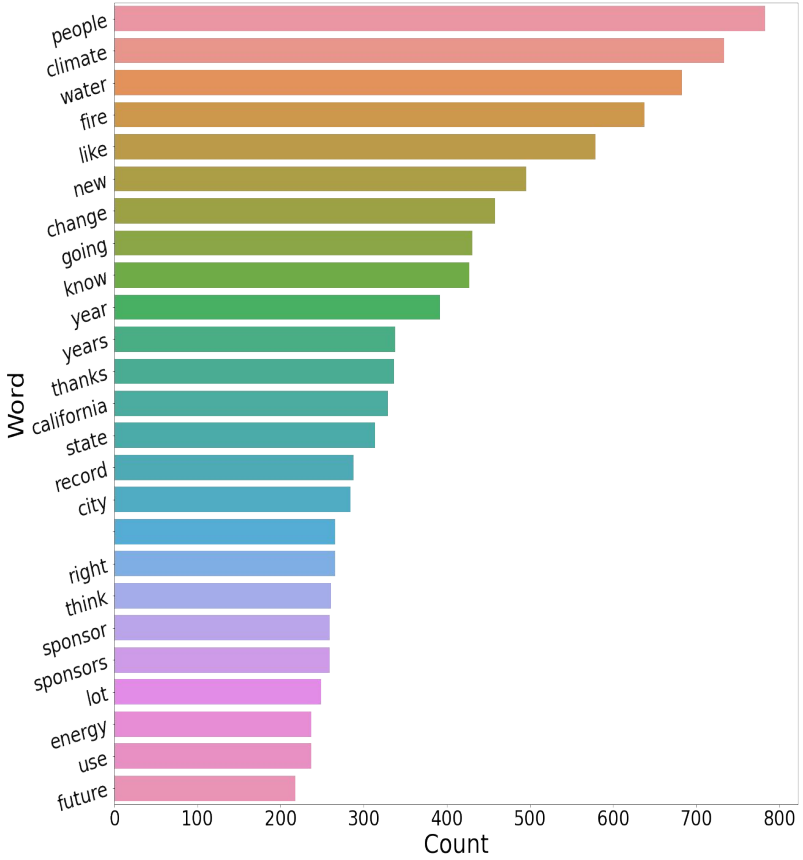


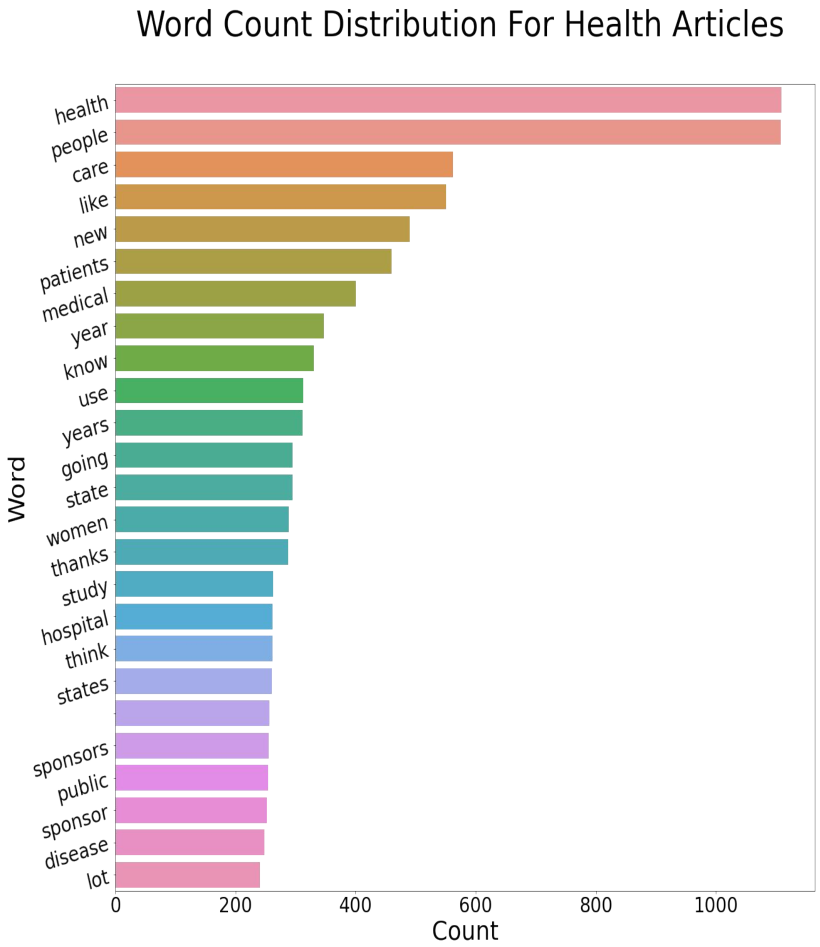
Word Count Distribution For Education Articles
students
school
like
college
people
university
schools
education
student
know
new
teachers
year
going
state
think
thanks
years
high
public
department
program
lot
work
0
200
400
600
800
1000
1200
1400
Count



student
school
people
year
college
know
think
teacher
going
new
program
education
university
parent
work

Environment



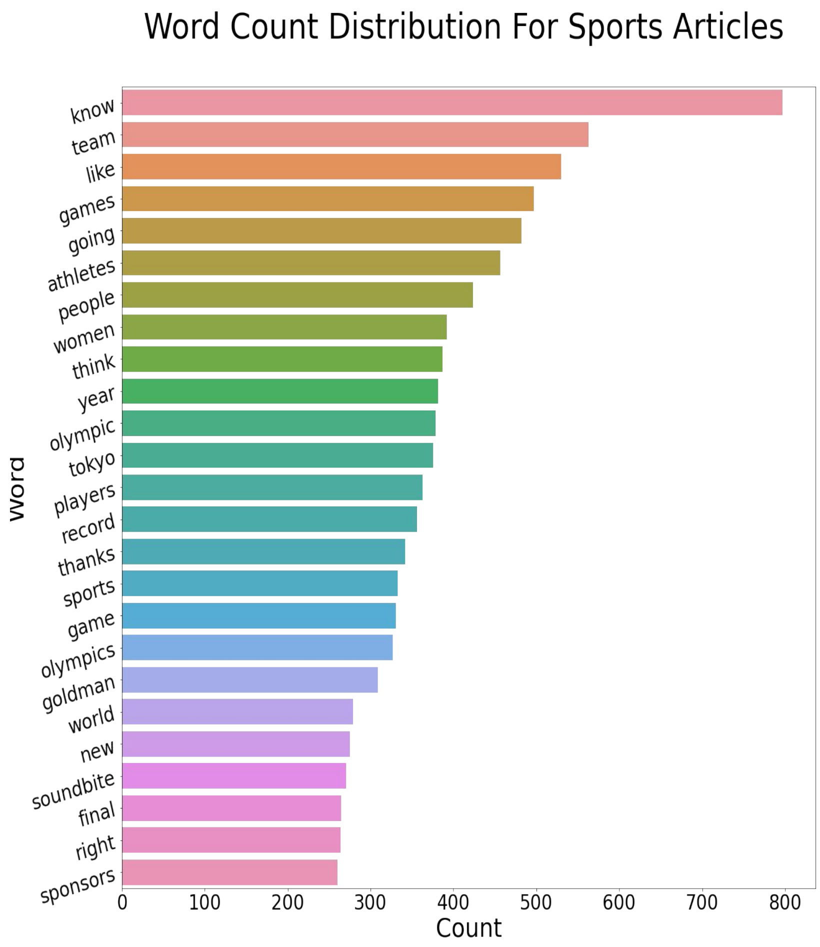


Word Count Distribution For Health Articles
health
people
care
like
new
patients
medical
year
know
use
years
going
state
women
thanks
study
hospital
think
states
sponsors
public
sponsor
disease
lot
Word
0
200
400
600
800
1000
Count

Health

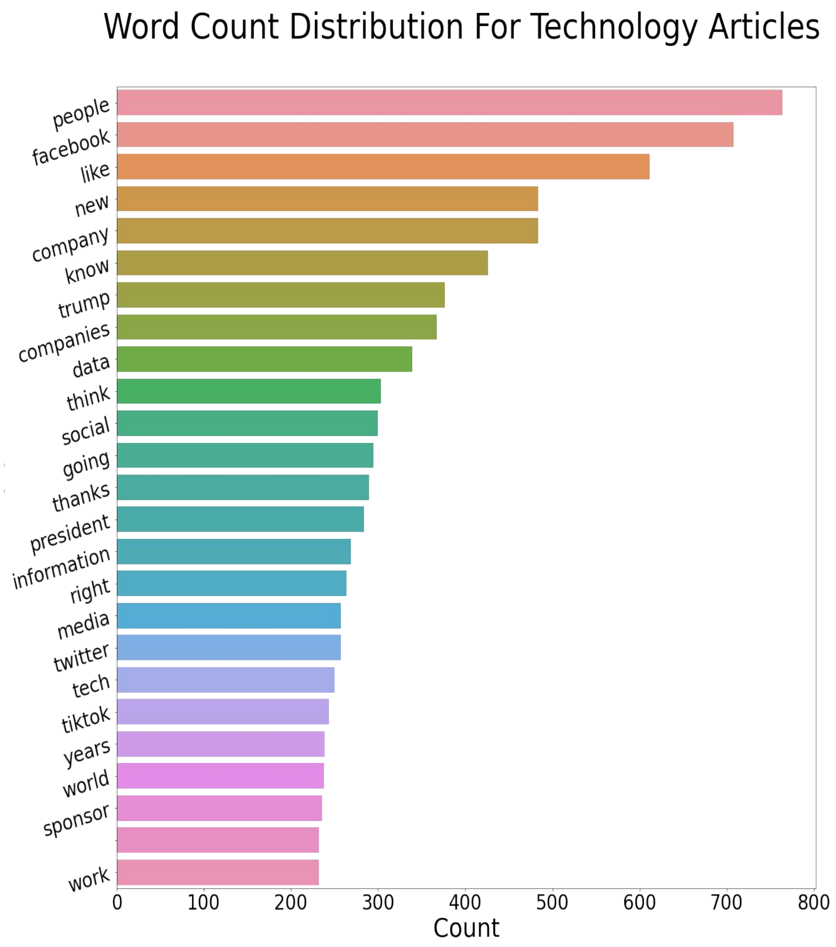
people
patient
year
state
know
health care
new
case
hospital
think
going
doctor
need
day
work
thanks sponsors
sponsors sponsor
month
cost
called
place
right
group
disease
report
week
vaping
problem
program
medical
plan
vaccine
country
children
women
drug
got
come
death
kind
level
look
company
want
live
research
number
treatment
service
american
found
way
thing
lot
talk
study
help
life
home
far
actually
good
find
feel
world
according
opioid
percent
change
government
community
system
researcher
including
different
bill
money
getting
million
family
result
pay
started
kid

Sports


Word Count Distribution For Sports Articles
Word
know
team
like
games
going
athletes
people
women
think
year
olympic
tokyo
players
record
thanks
sports
game
olympics
goldman
world
new
soundbite
final
right
sponsors
0
100
200
300
400
500
600
700
800
Count


game
know
team
athlete
going
year
player
sport
olympic
people
think
new
sponsor
sponsors
tournament
thanks
pandemic
japan
ncaa
world
season
event
website terms
visit website
copyright rights
rights reserved
terms use
thing
kind
day
lot
race
fan
want
play
state
reserved visit
right
women
tokyo
week
mean
win
come
use permissions

Tech



Word Count Distribution For Technology Articles
people
facebook
like
new
company
know
trump
companies
data
think
social
going
thanks
president
information
right
media
twitter
tech
tiktok
years
world
sponsor
work
0
100
200
300
400
500
600
700
800
Count

facebook
people
new
company
know think
year
game
going
data
companies
work
way
thing
social media
sponsors
sponsor
thanks sponsors
tiktok
twitter
platform
trump