Capstone Project

Topic Modeling on News Articles

Content

- Problem Statement
- Data Summary
- Data Preprocessing
- Models
- Challenges
- Conclusions
- Q&A

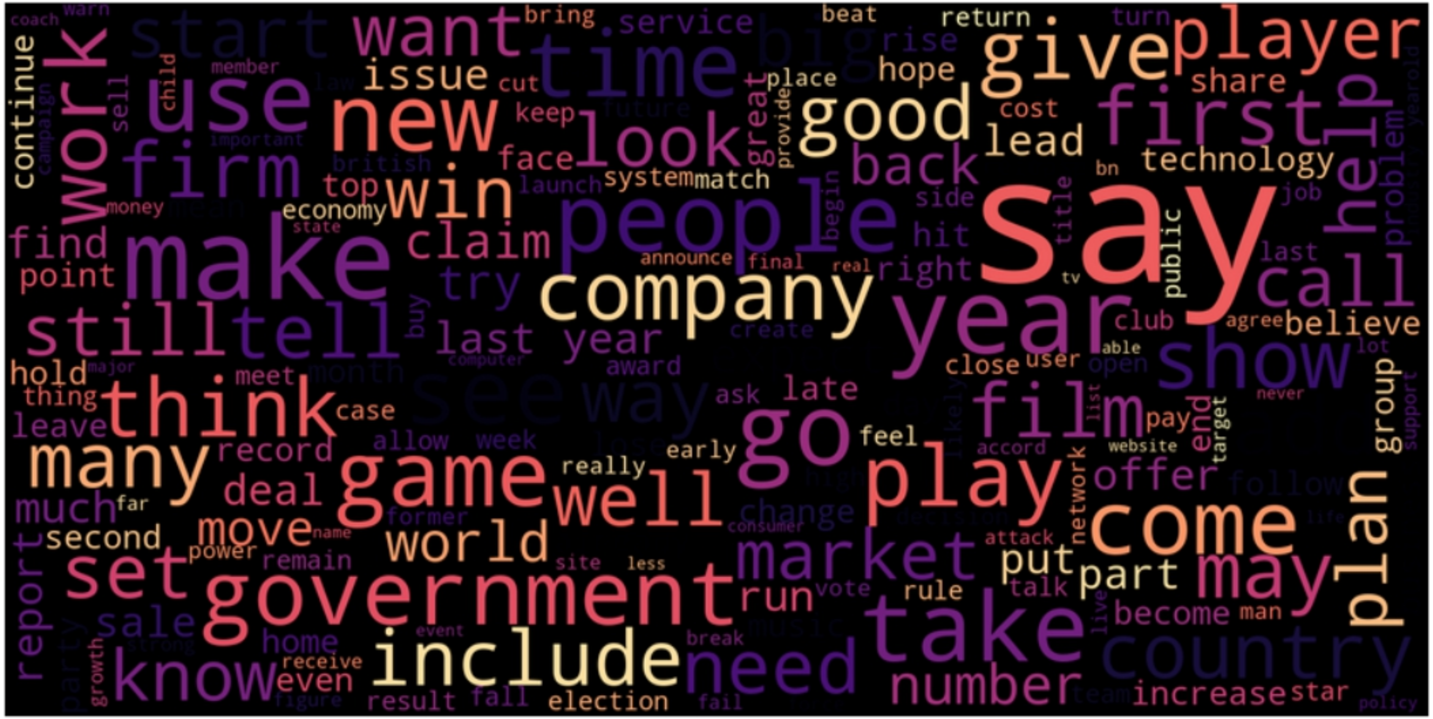
Problem Statement

- Identify major themes/topics across a collection of BBC news articles using Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA).

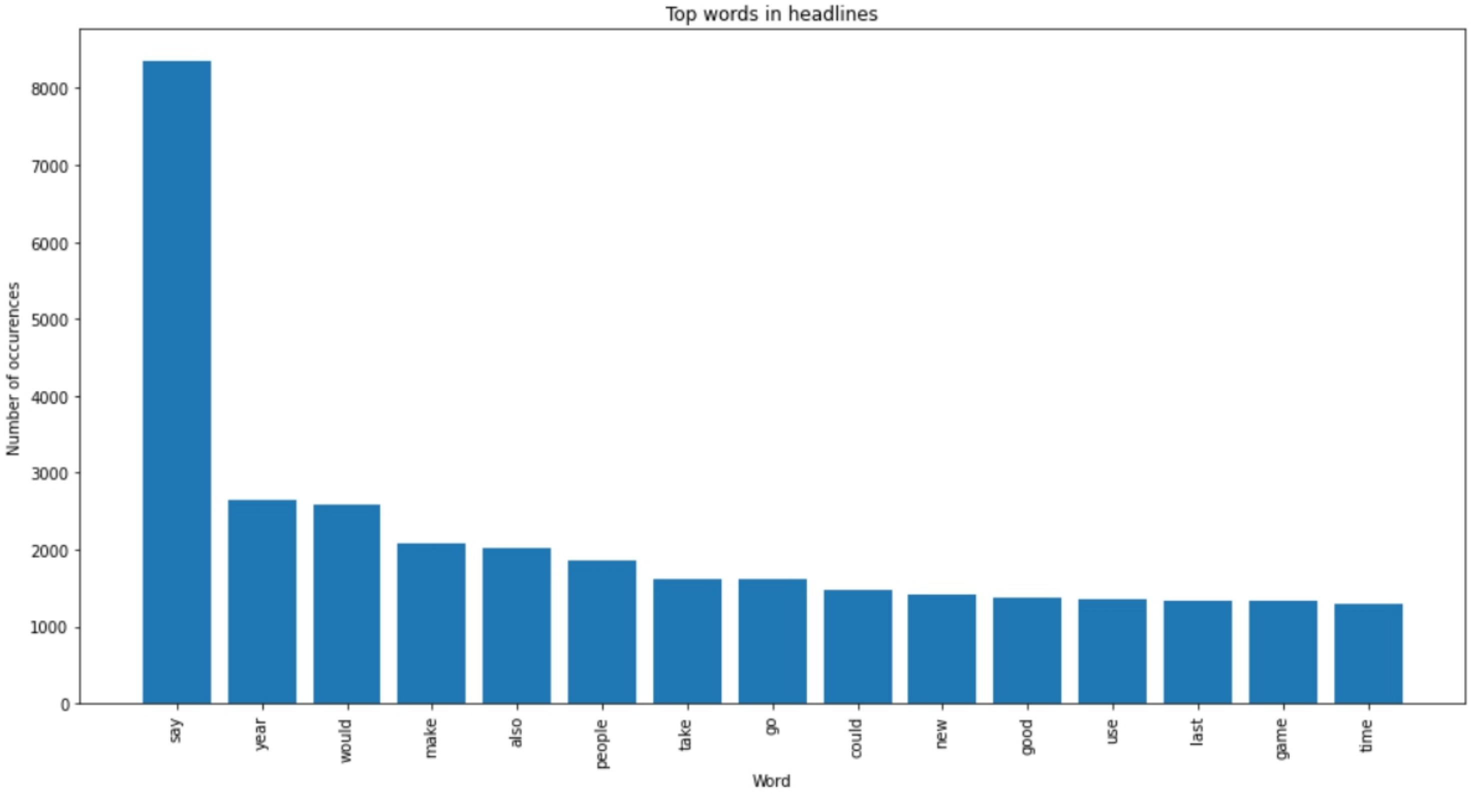
Data Preprocessing

- Remove duplicate values
- Remove Html tags
- Remove URLs
- Remove punctuation
- Remove numbers
- Remove small length words
- Remove stop words
- Lemmatization

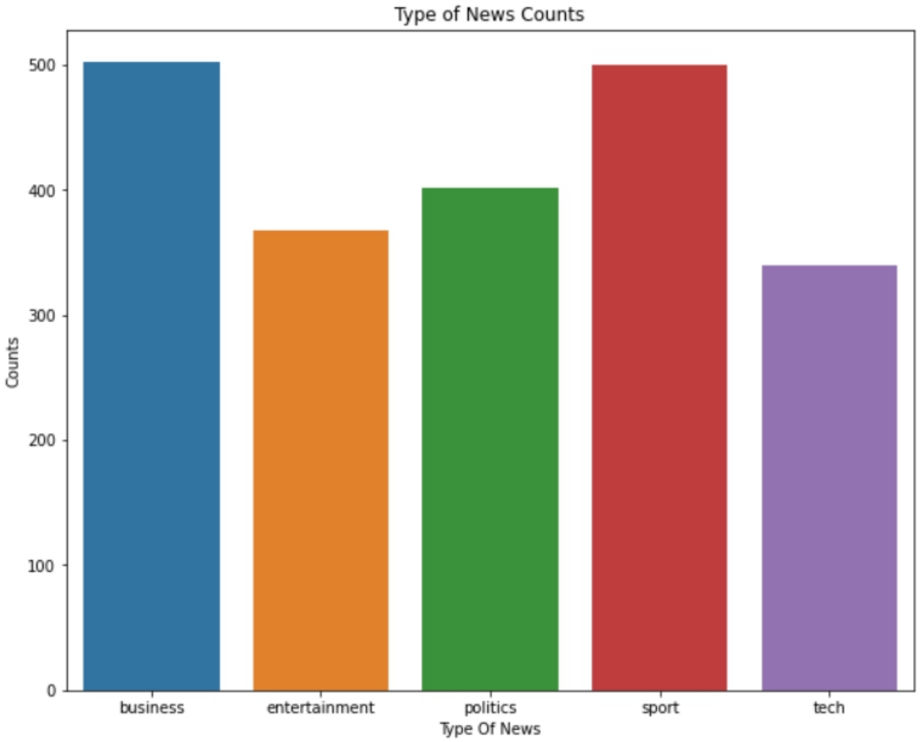
Word Cloud


coach
warn
start
want
bring
service
time
big
rise
beat
return
turn
player
give
continue
campaign
work
child
member
law
issue
cut
sell
use
new
place
hope
share
keep
future
great
provide
good
cost
first
help
yearold
important
look
lead
firm
british
face
back
technology
bn
problem
top
win
launch
system
match
money
mean
economy
state
side
job
people
say
title
public
begin
hit
find
make
claim
last
announce
final
real
right
point
company
try
call
year
tv
still
tell
buy
last year
club
agree
believe
create
hold
major
computer
award
close
user
able
open
show
lot
meet
month
see
way
thing
think
ask
late
group
support
never
case
film
leave
list
likely
pay
end
target
record
allow
week
lose
early
feel
go
accord
website
many
really
network
offer
follow
deal
high
play
much
far
well
change
decision
come
life
second
move
name
former
consumer
attack
plan
report
power
remain
world
market
put
may
set
site
less
part
talk
government
vote
rule
run
party
sale
live
become
man
event
music
home
strong
break
take
growth
receive
know
include
need
country
even
figure
result
fall
election
fail
number
team
increase
star
policy

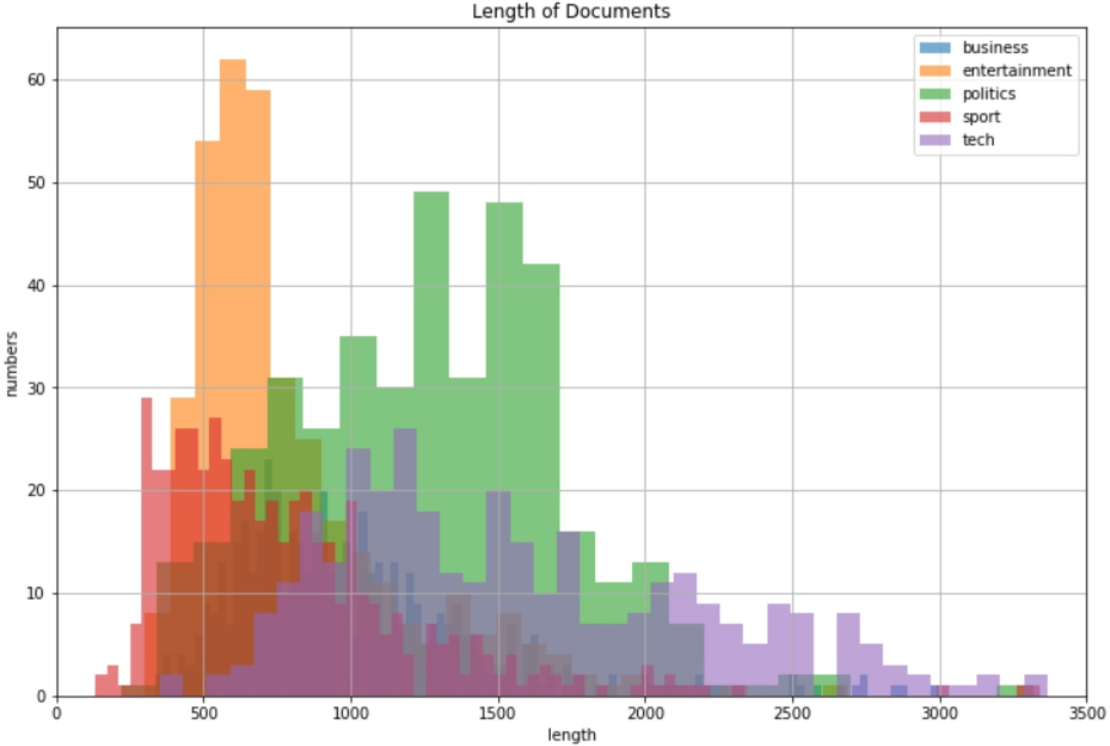
Frequent words



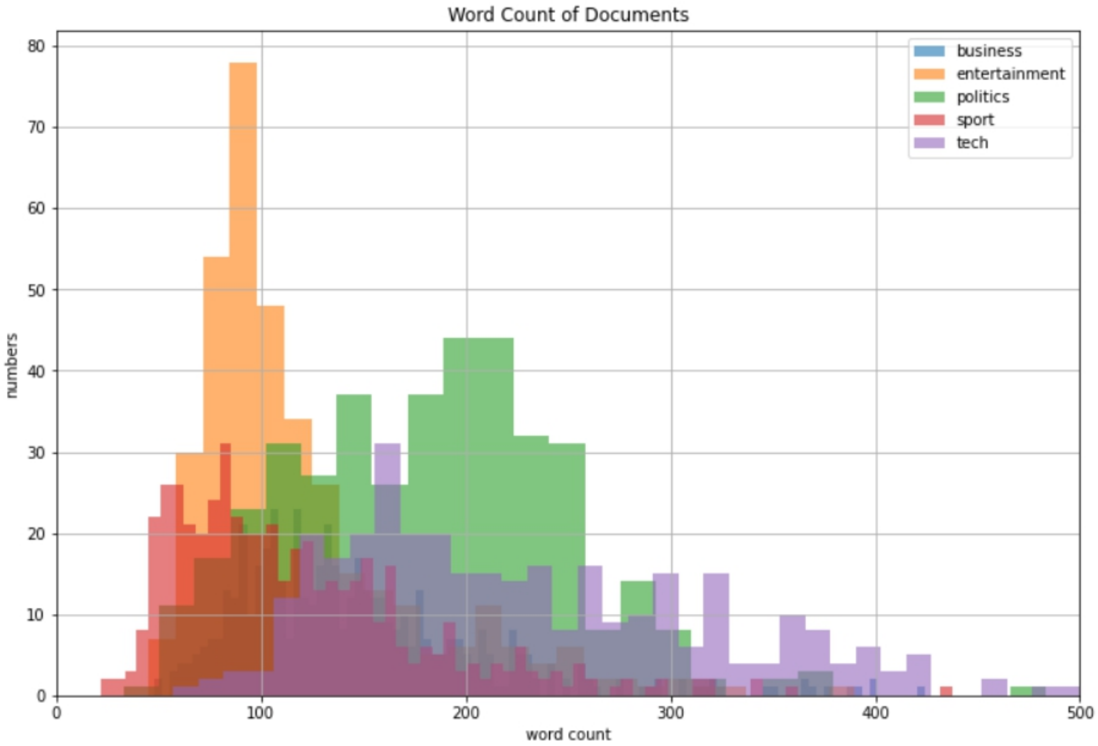
Countplot of News Types



Length of Documents



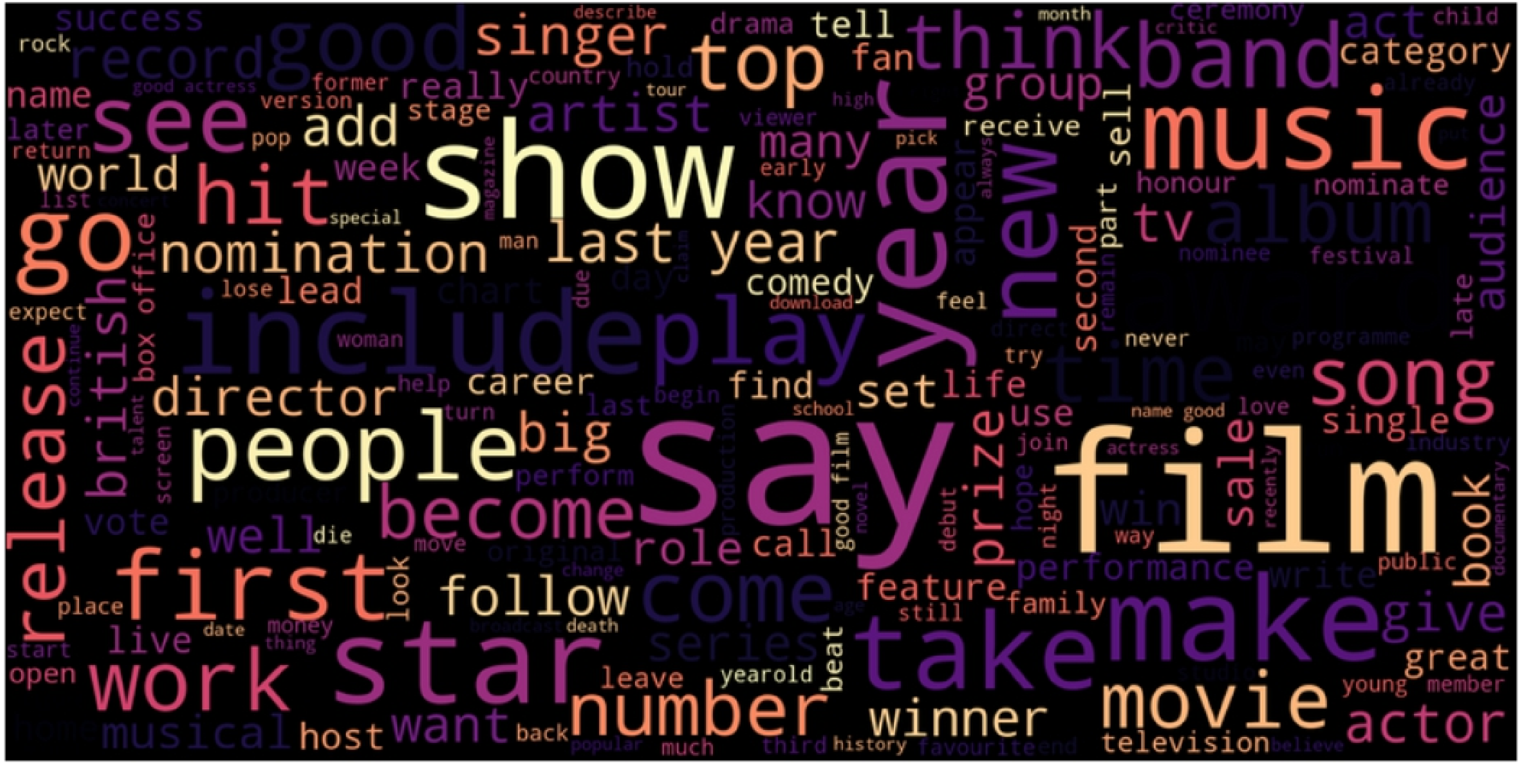
Word Count of Documents



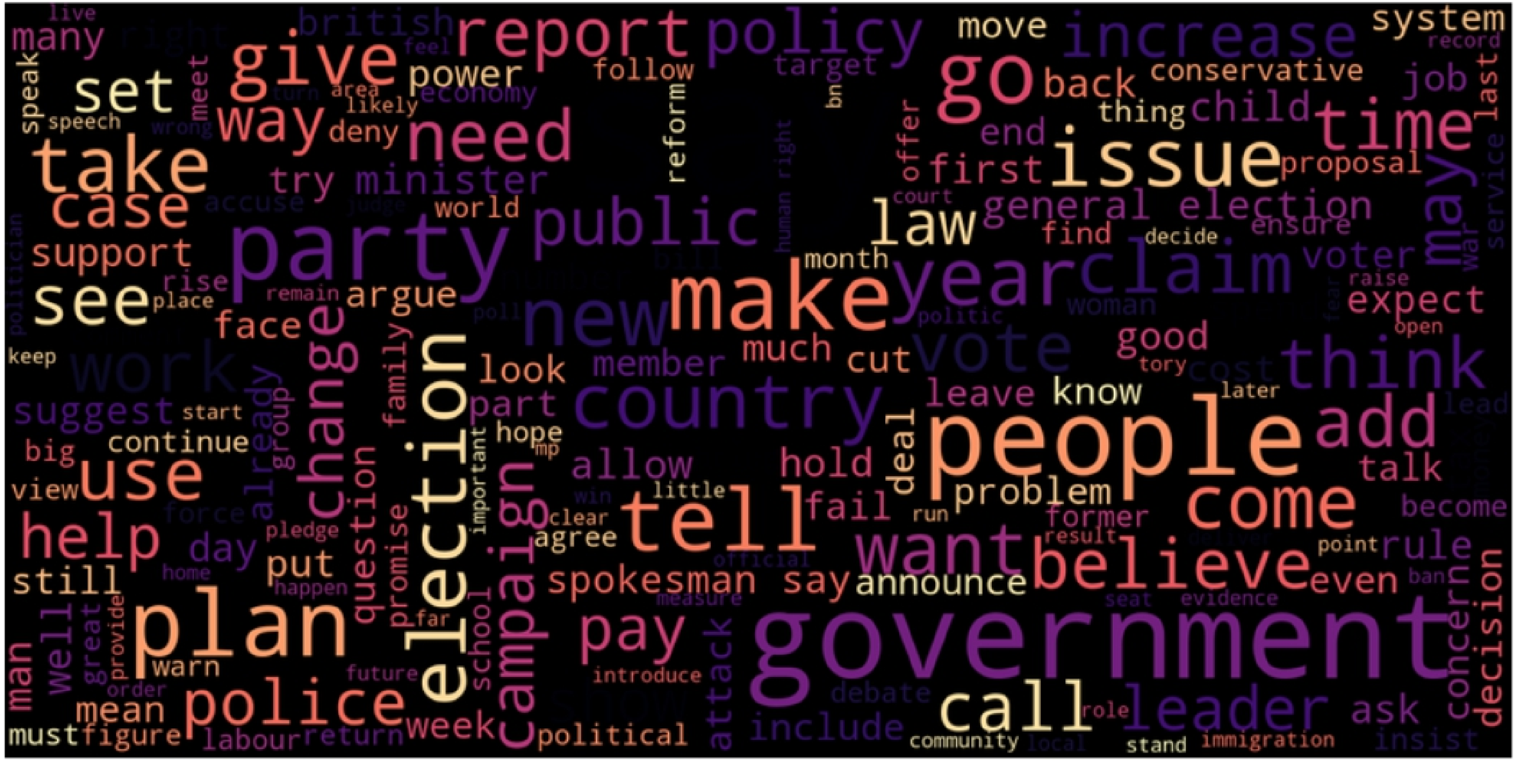
Word Cloud for Business

announce
cost
state
dollar
lead
club
target
fall
sell
production
economy
agreement
plan
account
build
bn
price
money
well
recent
need
accord
control
tell
growth
find
way
loss
increase
record
official
talk
major
hope
follow
public
offer
see
case
year
rate
low
analyst
buy
list
call
pay
set
rule
many
point
oil
interest rate
already
stock market
even
cut
demand
use
claim
asset
say
strong
boost
industry
budget
big
market
new
want
risk
run
sale
deal
result
late
financial
first
area
order
still
court
group
come
share
force
gain
allow
job
think
likely
issue
future
high
deficit
period
much
spending
export
day
investment
hit
report
keep
top
decision
product
business
firm
large
raise
make
grow
meet
drive
give
show
policy
currency
figure
move
help
remain
debt
rival
spend
become
european
fund
forecast
chief executive
back
lose
take
work
company
rise
expect
people
airline
level
drop
value
profit
consumer
try
car
look
close
less
continue
warn
month
investor
bid
global
world
part
problem
bring
mean
early
analyst say
number
good
put
country
may
government
german
unit
face
change
damage
believe
estimate
benefit
agree
start
last year
hold
bank
include
economic
time
add
total
sign
shareholder
end
trade
economic growth
customer
sector
earning
main

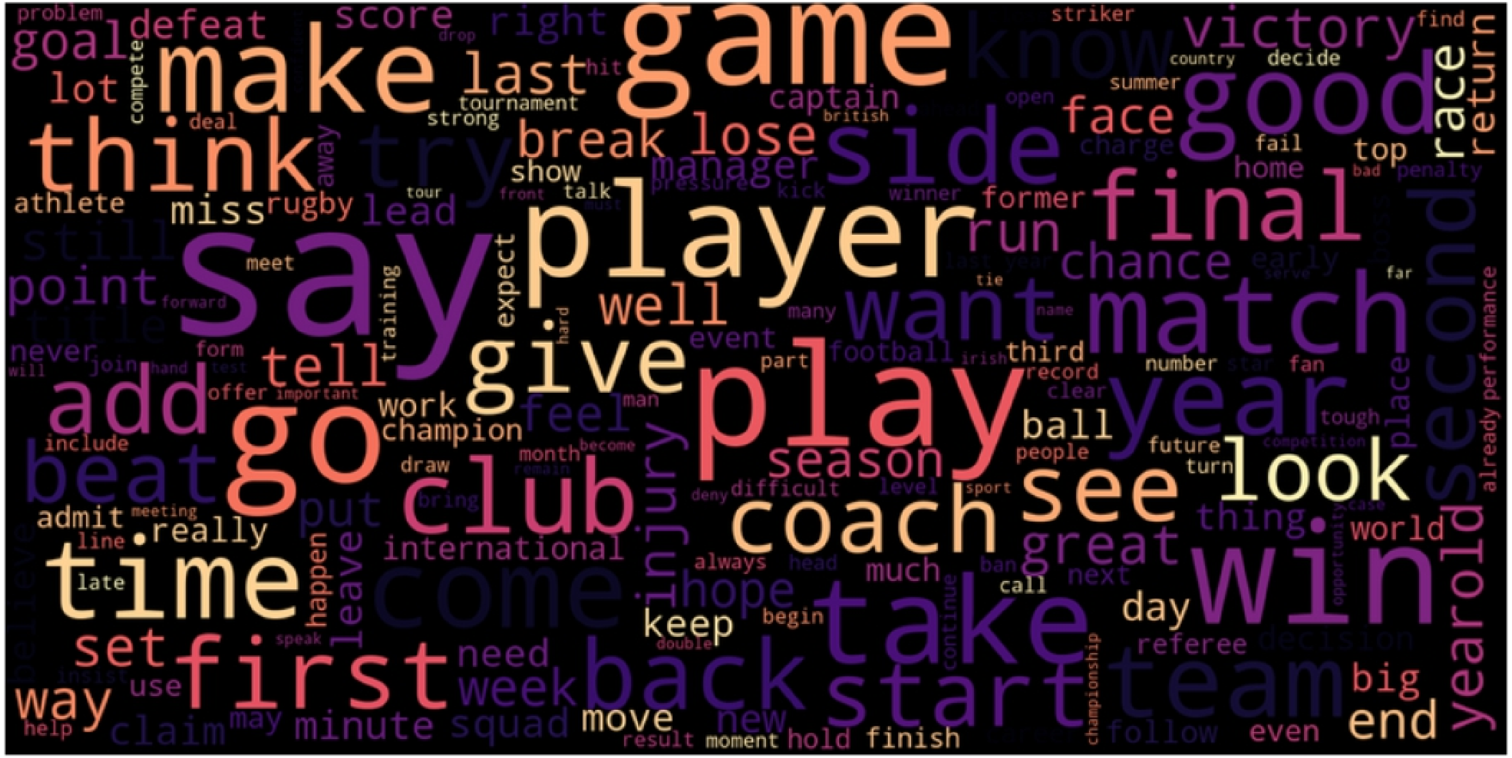
Word Cloud for Entertainment



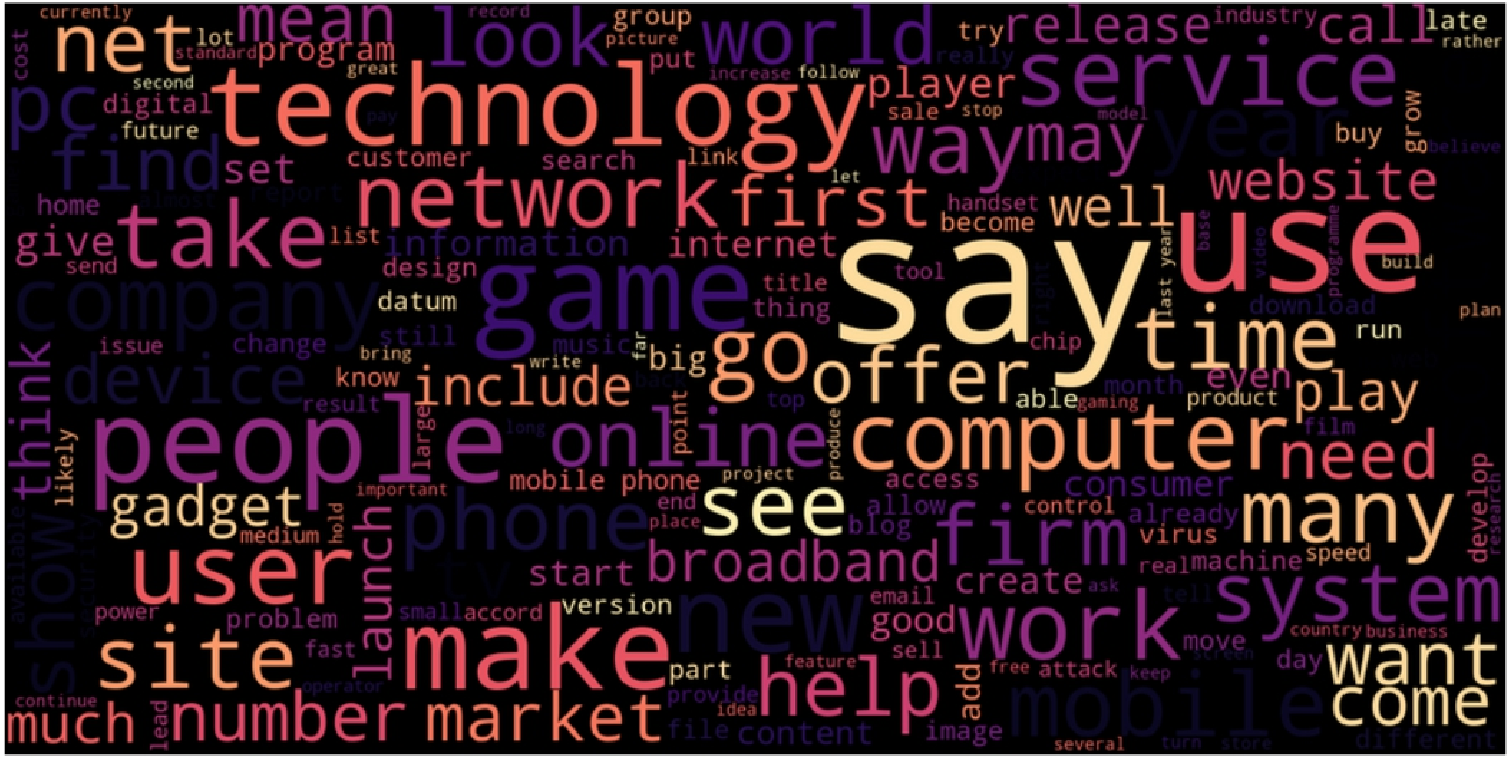
Word Cloud for Politics



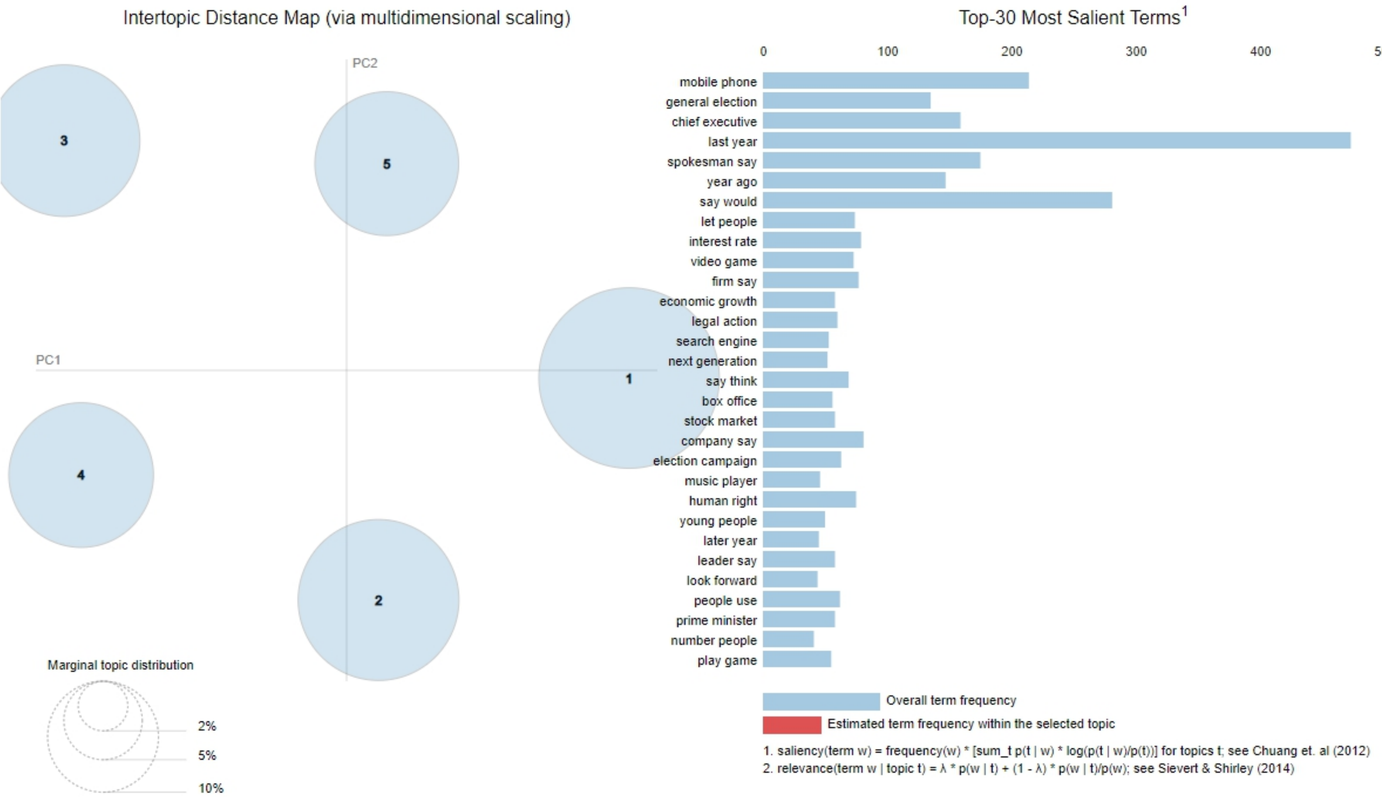
Word Cloud for Sport



Word Cloud for Tech



Latent Dirichilet Allocation



1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Similar Words

Word : Film

Good
Award
Star
Play
Actor

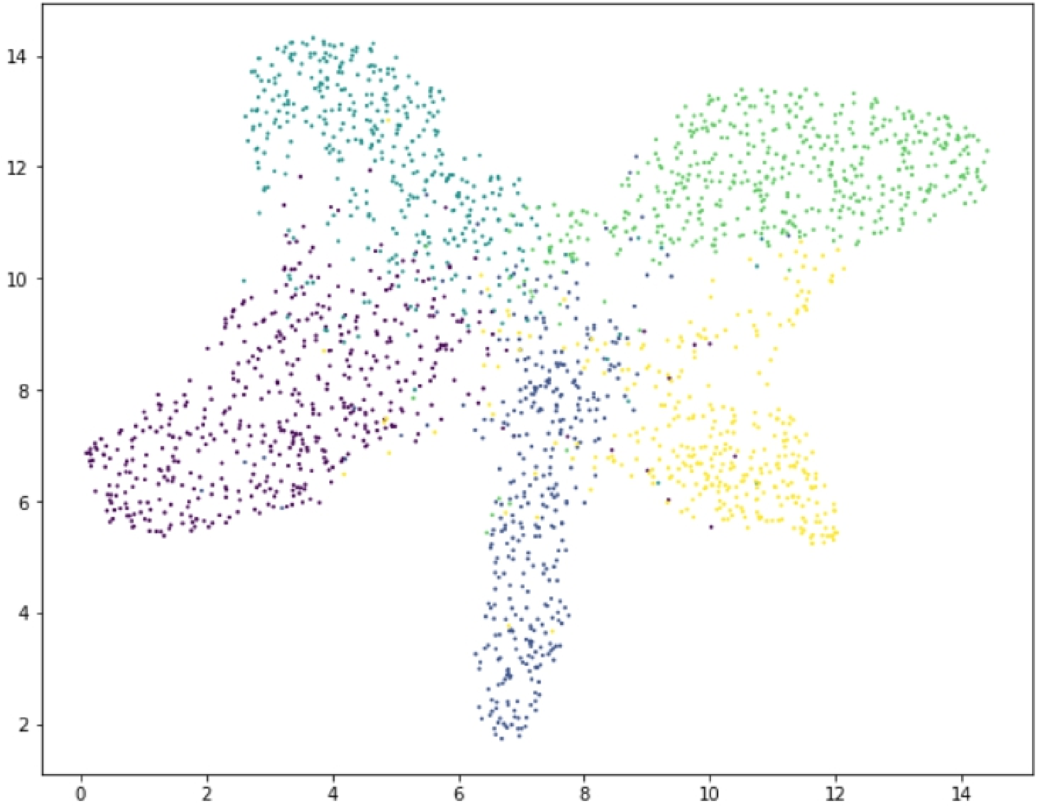
Word:Price

Rate
Rise
High
Profit
Fall

Word:Stock

Gross
Soar
Copy
Late
Monthly

Latent Semantic Analysis



Challenges

- Text Preprocessing
- Limited visualization techniques

Conclusion

- In Latent Dirichilet Allocation(LDA) with TF-IDF vectorizer we find best clustering for our news article dataset .
- Using of genism we can find a similar words .
- As a future we can implement topic modeling using different techniques like neural network.

Q & A

Thank You