Project 3

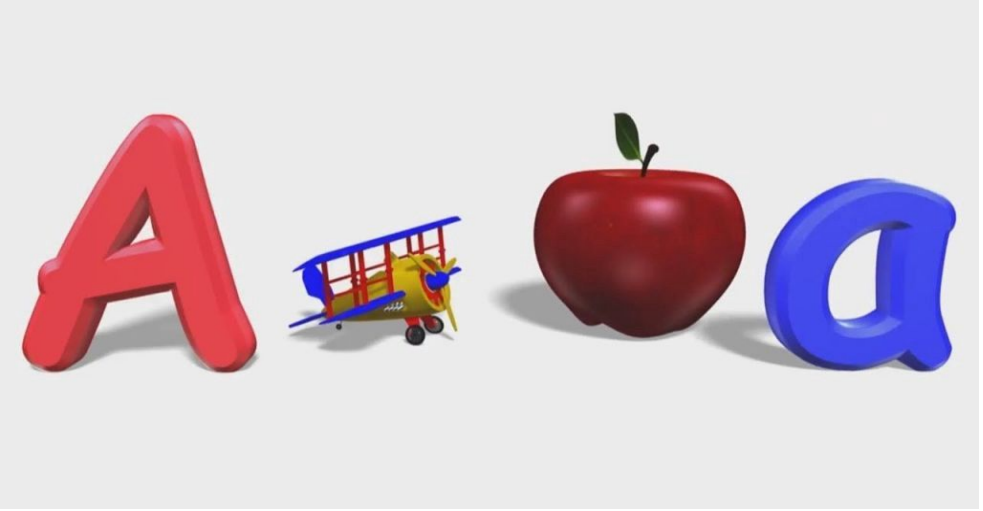
Newsgroups NLP Classification

What are we trying to accomplish?

➔ Automatically Import fresh articles into newsgroup discussions

Why is this important?

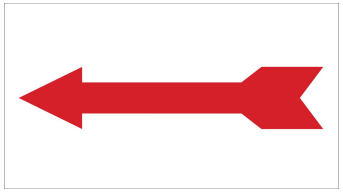
➔ Increase usage and relevance of newsgroups



How it Works:

- Use NLP model to cluster existing newsgroup data
- Many iterations to establish some cluster separation
- TRY to associate Topic Clusters with Newsgroups

Build a Classifier using these associations



Dataset

18,846

NewsGroups

- comp.graphics
- comp.os.ms-windows.misc
- comp.sys.ibm.pc.hardware
- comp.sys.mac.hardware
- comp.windows.x
- rec.autos
- rec.motorcycles
- rec.sport.baseball
- rec.sport.hockey
- sci.crypt
- sci.electronics
- sci.med
- sci.space
- misc.forsale
- talk.politics.misc
- talk.politics.guns
- talk.politics.mideast
- talk.religion.misc
- alt.atheism
- soc.religion.christian

A lot of these are very similar!

Bag of Words

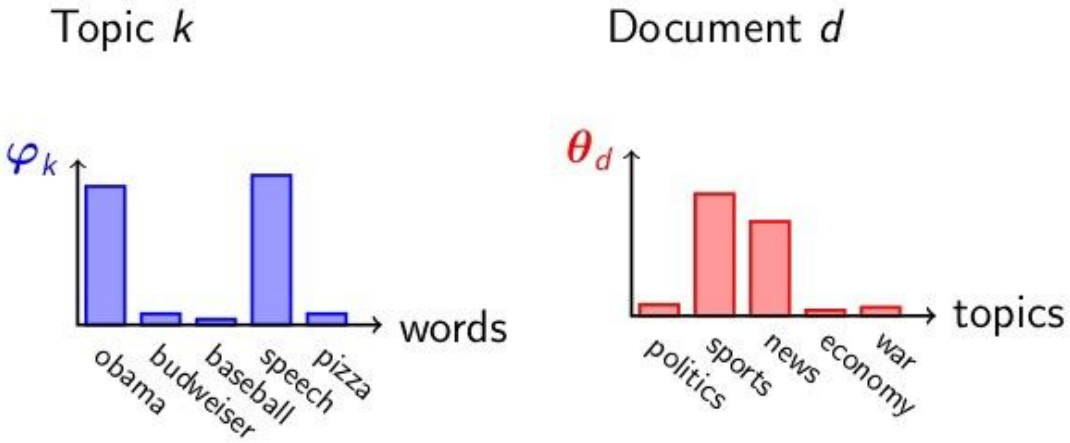
I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
…	…

Latent Dirichlet Allocation

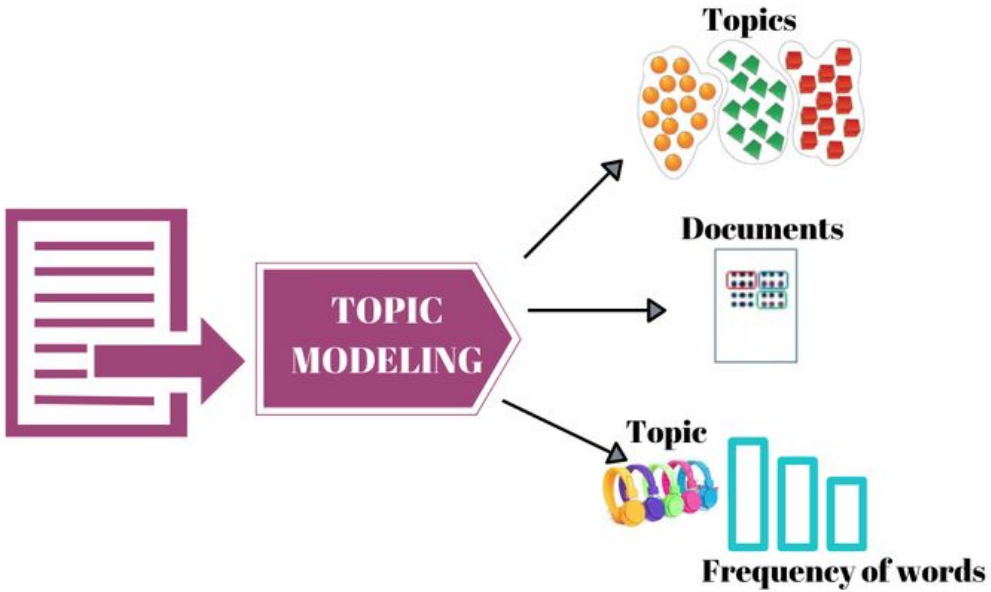
LDA discovers topics into a collection of documents.

LDA tags each document with topics.

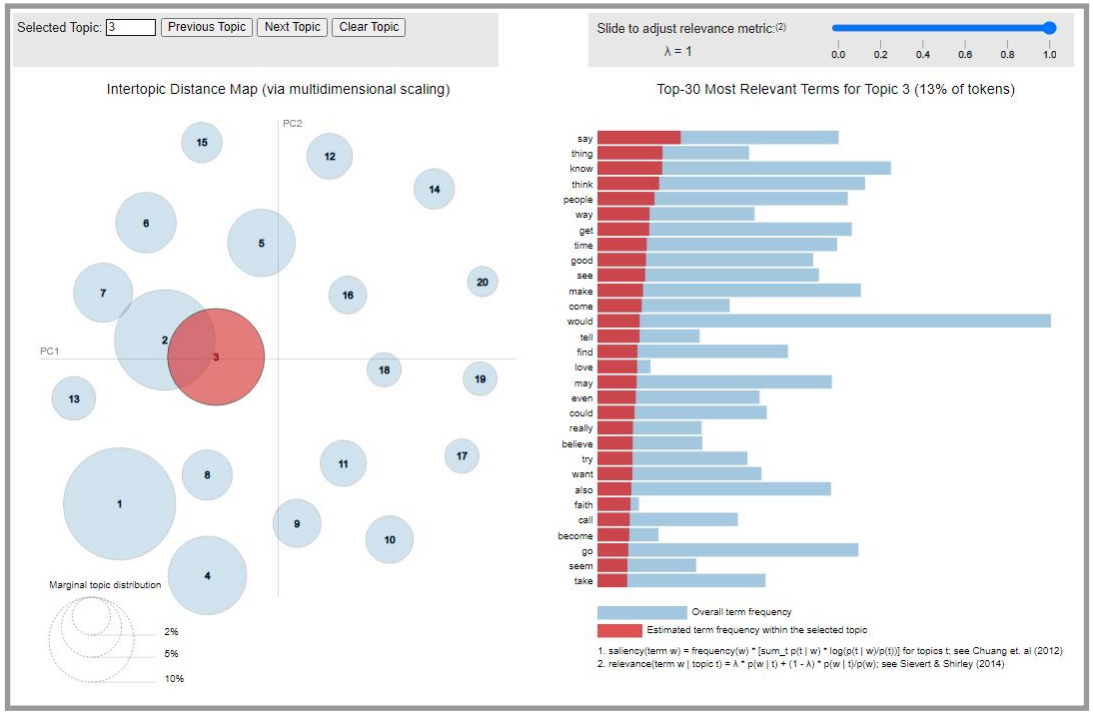


- Uses Bag of Words to identify topics in document
- Not a linear model, LDA came from research in genetics
- Assumption is that documents contain several topics

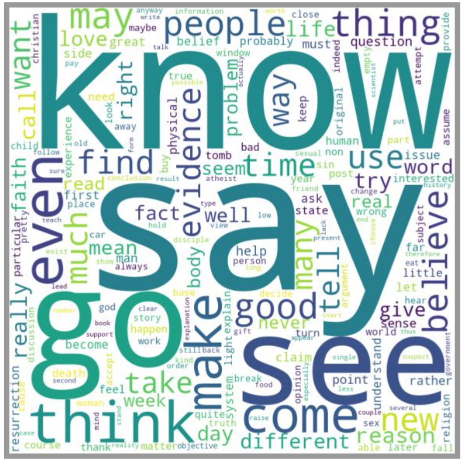
Technology and Model



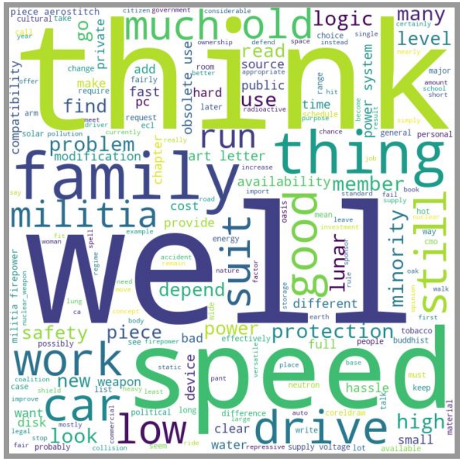
Experiment 1 - Topic Model



WordCloud

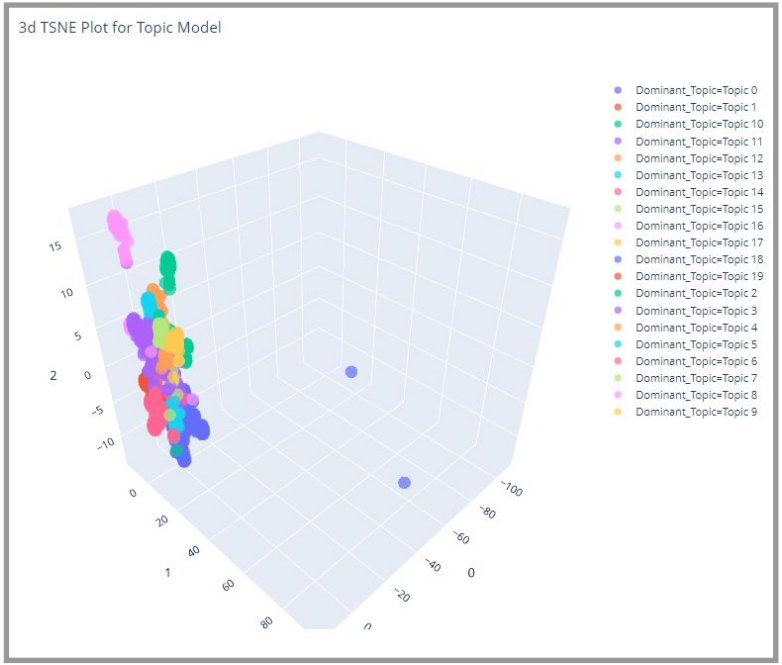

know
say
go
see
think
even
make
come
people
thing
may
time
find
evidence
believe
good
take
different
want
right
way
love
life
fact
well
use
tell
much
mean
reason
new
day
week
word
problem
many


good
use
claim
sell
case
continue
still
make
ground
think
file
report
award
last
want
price
warning
year
different
tape
go
may
government
national
large
account
people
new
sale
high

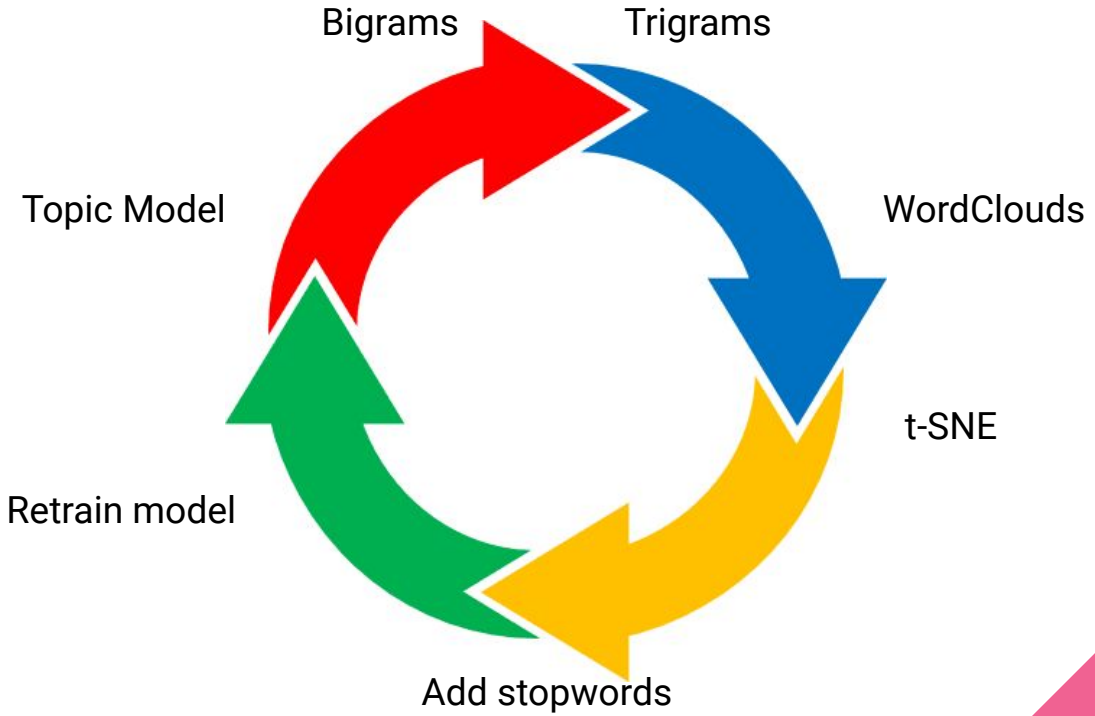

think
well
speed
family
militia
much
old
thing
car
drive
low
work
run
suit
good
still
lunar
logic
many
level
minority
problem
protection
power
depend
member
source
read
public
find
high
look
piece
safety


use
system
image
program
file
key
datum
know
package
available
copy
work
user
access
need
find
time
software
number
device
run
may
try
display
list
application
include
bit
chip
source
model
information
new
support
make

t-distributed Stochastic Neighbor Embedding (t-SNE)

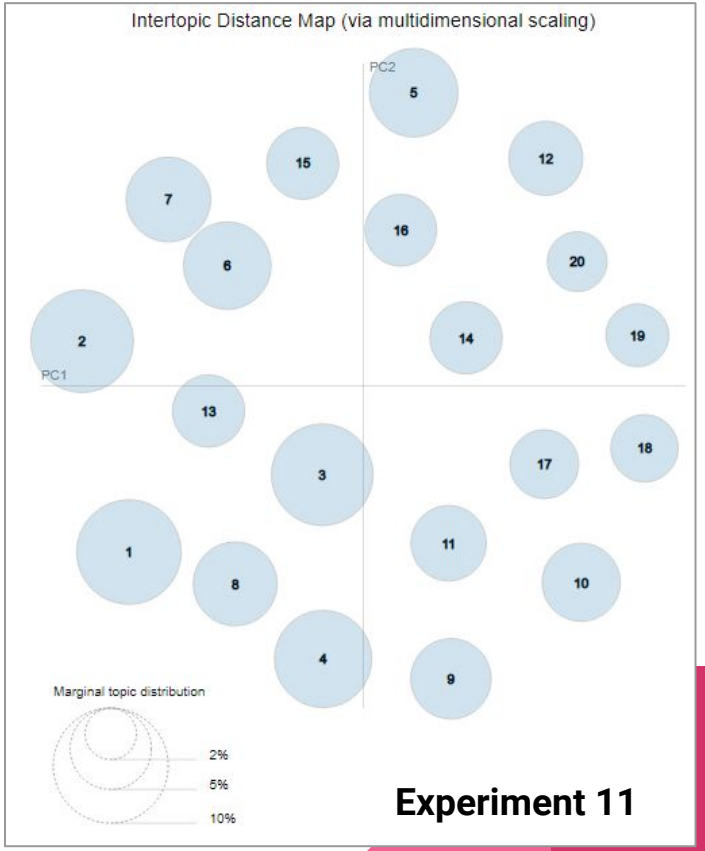


Experiment Methodology



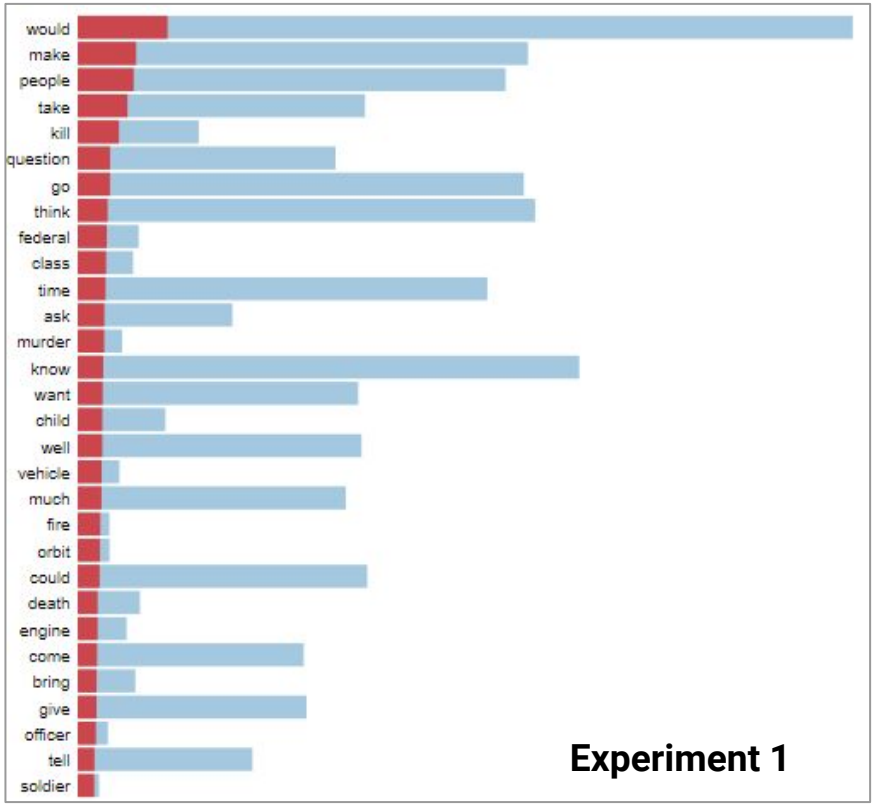
Topic Model - Comparison IDM

Experiment 1

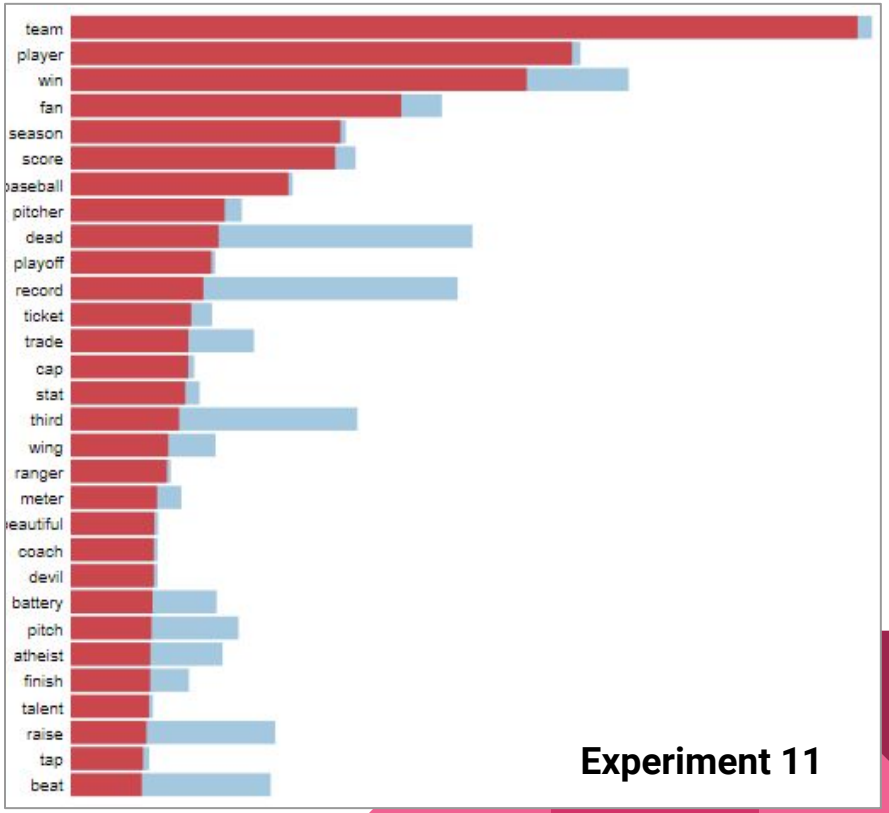


Experiment 11

Topic Model - Term Comparison

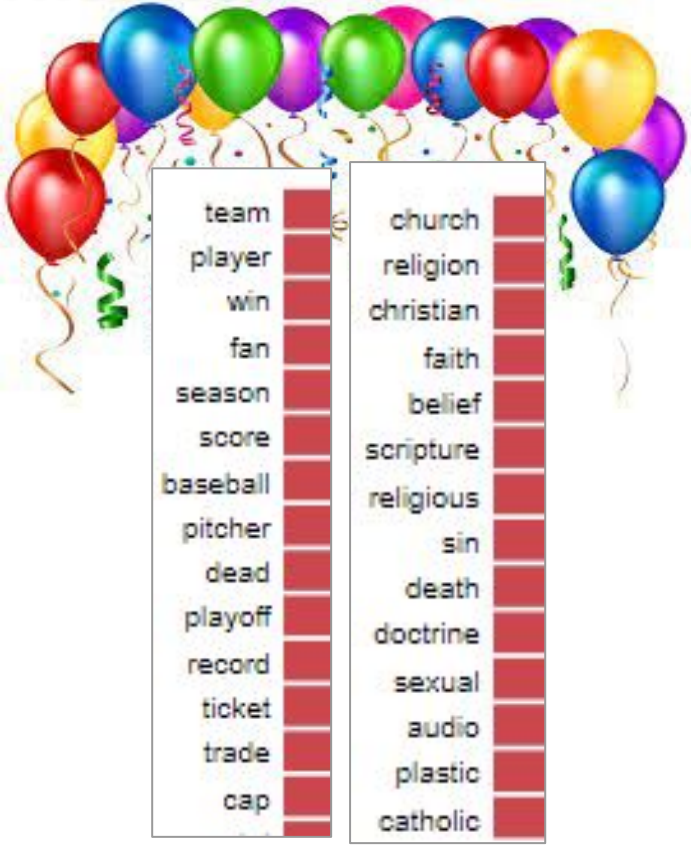


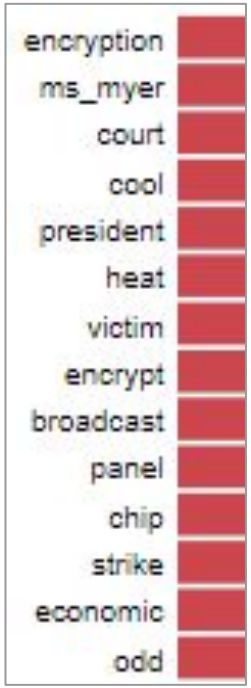
Experiment 1



Experiment 11

Themes


team
player
win
fan
season
score
baseball
pitcher
dead
playoff
record
ticket
trade
cap
church
religion
christian
faith
belief
scripture
religious
sin
death
doctrine
sexual
audio
plastic
catholic


encryption
ms_myer
court
cool
president
heat
victim
encrypt
broadcast
panel
chip
strike
economic
odd

Bigrams

loss tie
division finish
dead dead
coach corner
camera camera
battery camera
film cassette
innocent murder
ice stat
hate team
bike storage
stat player
goal assist
score goal
rod brind_amour
pitcher pitch
fan fan
score team
team player
team win
0

Trigrams

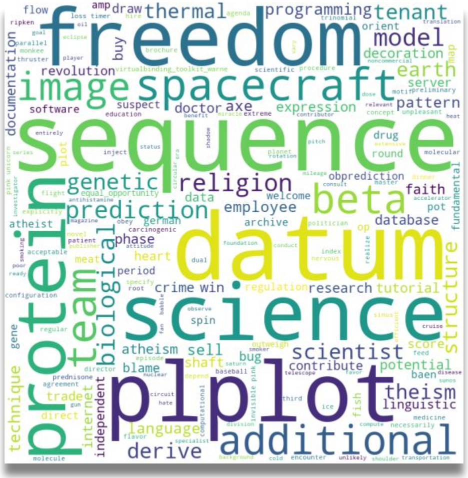
injury prone loose
suffer rush overweight
adirondack skipjack ranger
vdc tip conductor
win loss tie
team win team
team score team
player ridiculous absurd
skipjack ranger devil
record win loss
ranger devil oiler
player flyer season
loss tie team
islander adirondack wing
idle conductor vdc
ice team coach
cop fed listen
conductor vdc tip
canuck goal jet
adirondack wing adirondack
team record win
team player team
player ice team
win team win

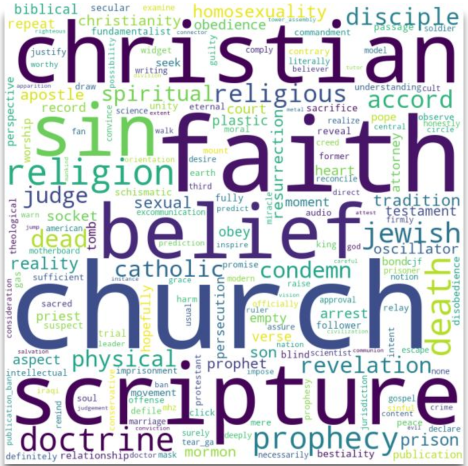
Trigrams - not all converge

relevant datum science
roll heart soul
round round quiet
science datum image
theism theism favor
touch possession obsession
genetic algorithm genetic
drag dirty hall
danced zombie nervous
coupon coupon coupon
bam boom touch
hall oate bam
hooter nervous danced
plant stir journey
oate bam boom
monkee monkee mediate
journey raise suzanne
identify relevant datum
invisible pink unicorn

Wordcloud


scsi
pc
chip
disk
port
driver
cpu
transfer
modem
controller
buy
external
internal
random
bus
interface
cipher
hardware
software
ide
datum
manual
audio
encrypt
sell
greatly
meg
depend


freedom
sequence
datum
science
plplot
protein
additional
spacecraft
image
beta
team
religion
genetic
prediction
biological
scientist
theism
derive
model
programming
tenant
thermal
structure


christian
faith
church
belief
scripture
sin
religion
doctrine
prophecy
religious
spiritual
accord
disciple
jewish
catholic
condemn
death
revelation
physical
judge
dead
apostle
homosexuality
christianity

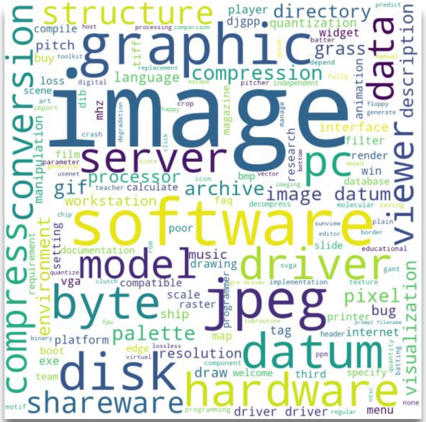
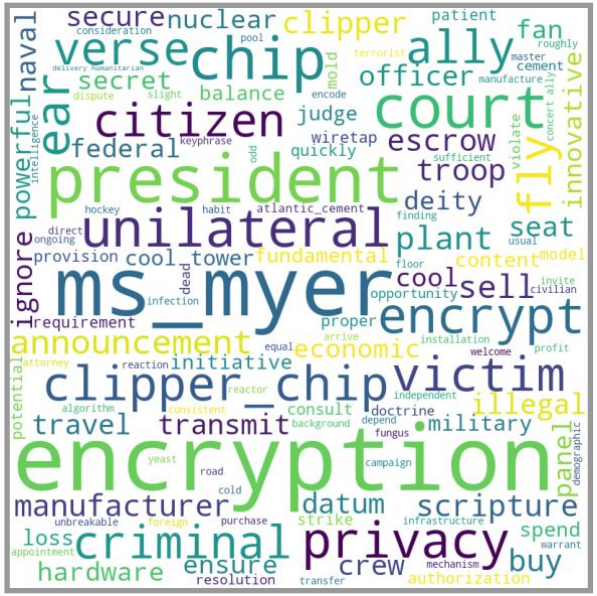
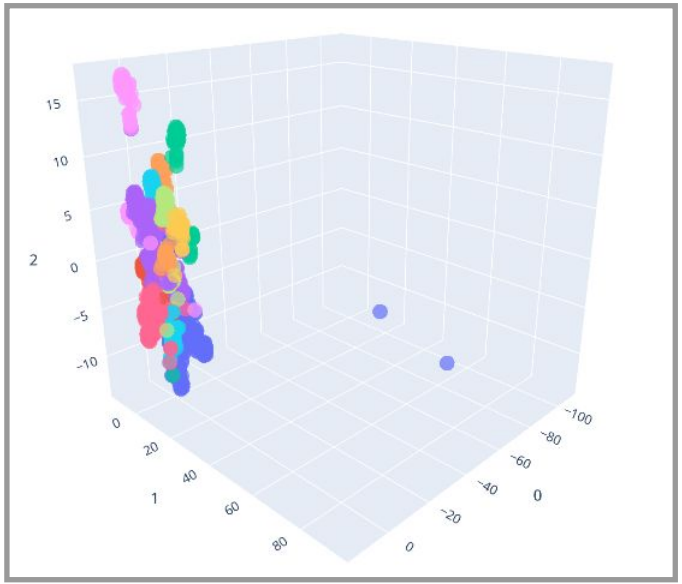


image
graphic
software
jpeg
byte
disk
datum
hardware
driver
server
model
pc
compress
conversion
viewer
data
structure
compression
shareware
palette
pixel
resolution
archive
directory
visualization

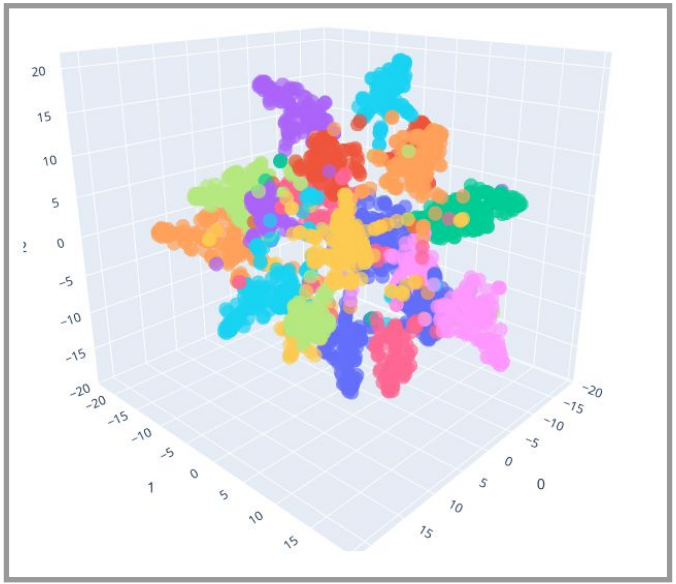
Not all topics converged as cleanly



t-SNE - Model Improvement



Experiment 1



Experiment 11

Ethics

Conclusion

Thank You