Reddit post topic modelling

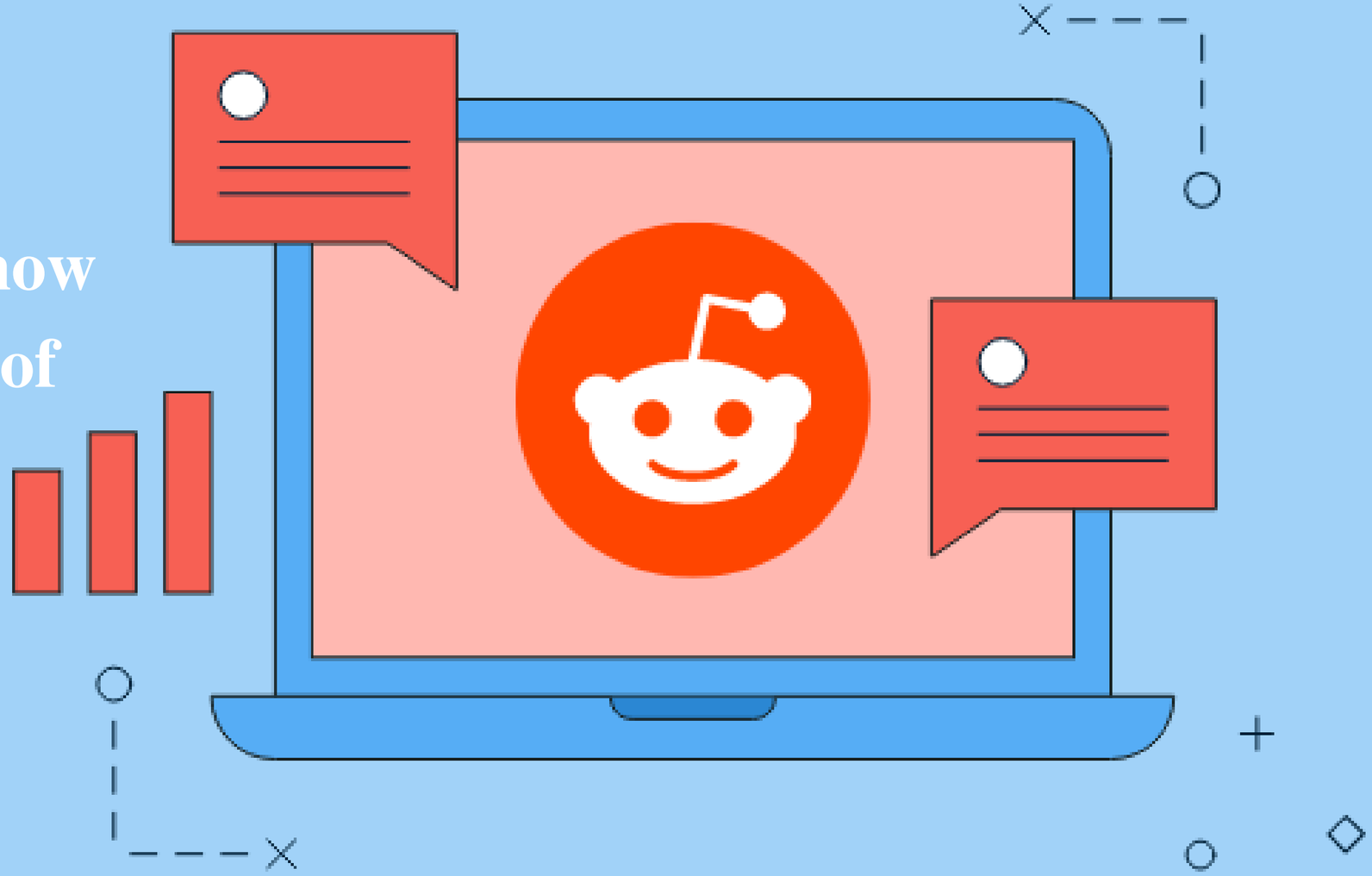
Prepared by:

Abdultawwar Safarji

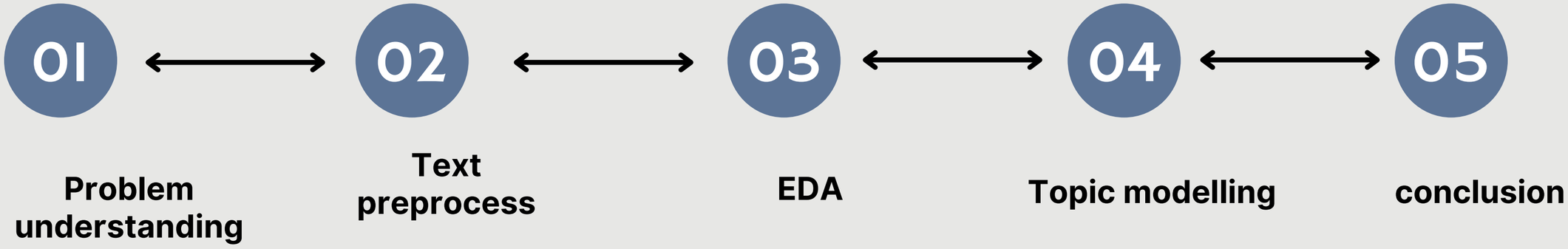
Najla Bin-Melha

Khalid Alrashed

Have you ever asked yourself that through your use of reddit you know the behavior of users?



Methodology

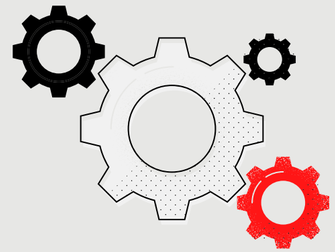


Problem Understanding

Why Riddet Toping Modeling ?

- To organize, understand and summarize large collections of textual information.
- Make it easy for the Reddit community to use unsupervised topic modeling and advance in technology as people discuss in text format.

Text preprocess

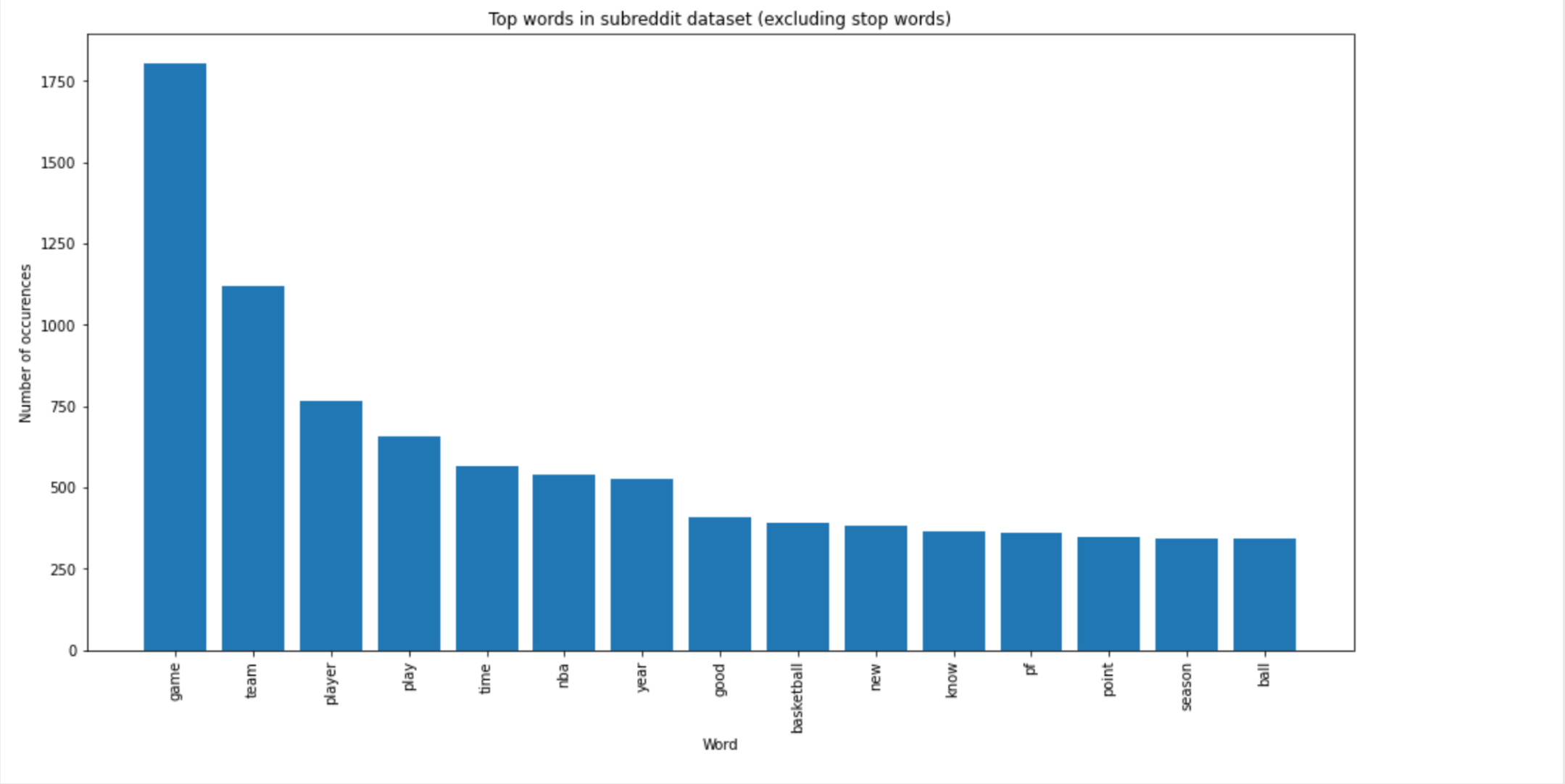


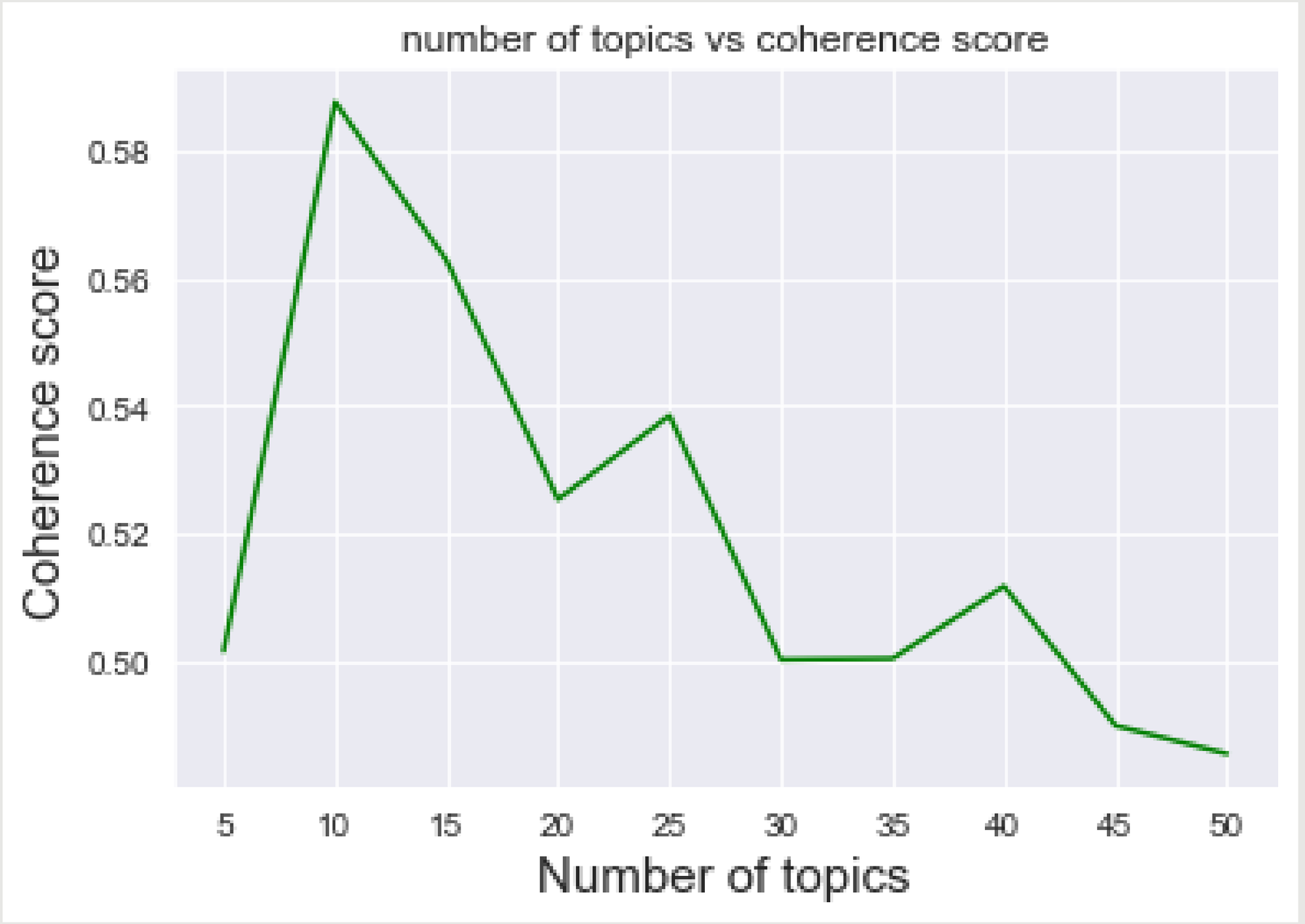
- URL links
- Numbers &punctuation(etc..)

- Tokenization

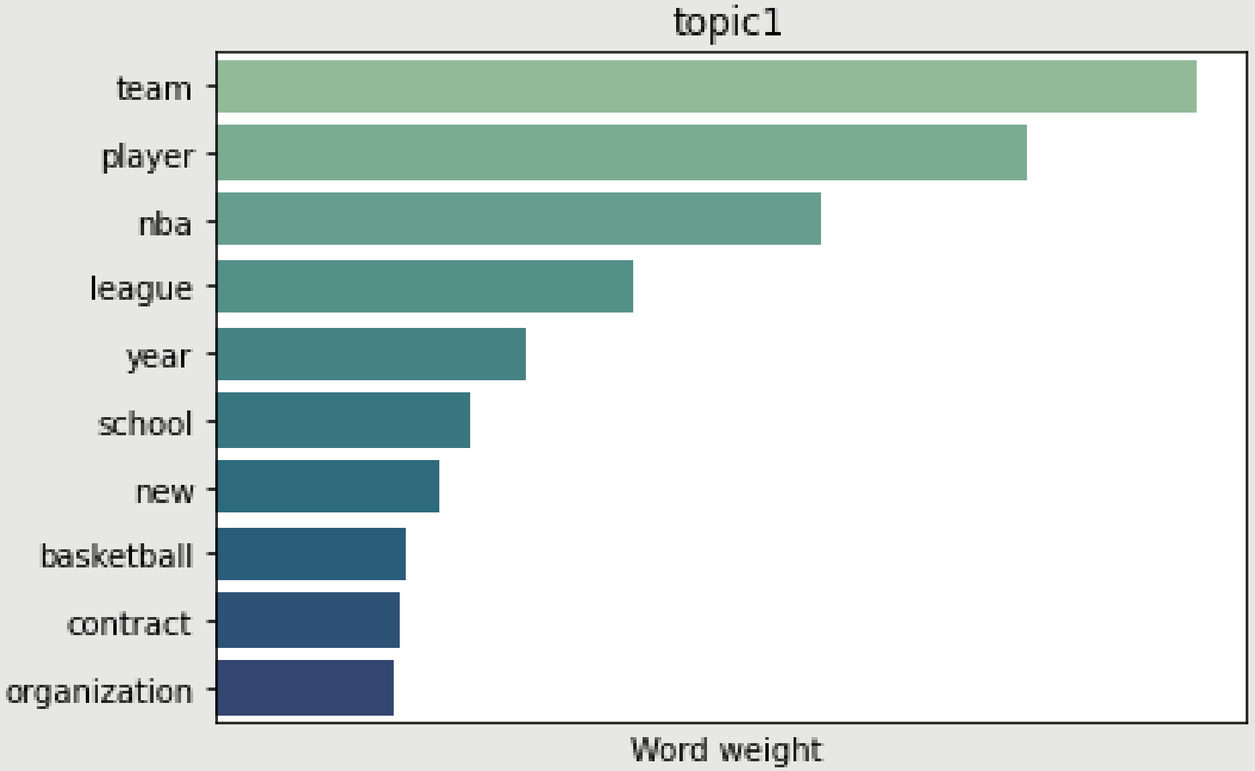
- Stop words
- lemmatization

Exploratory Data Analysis

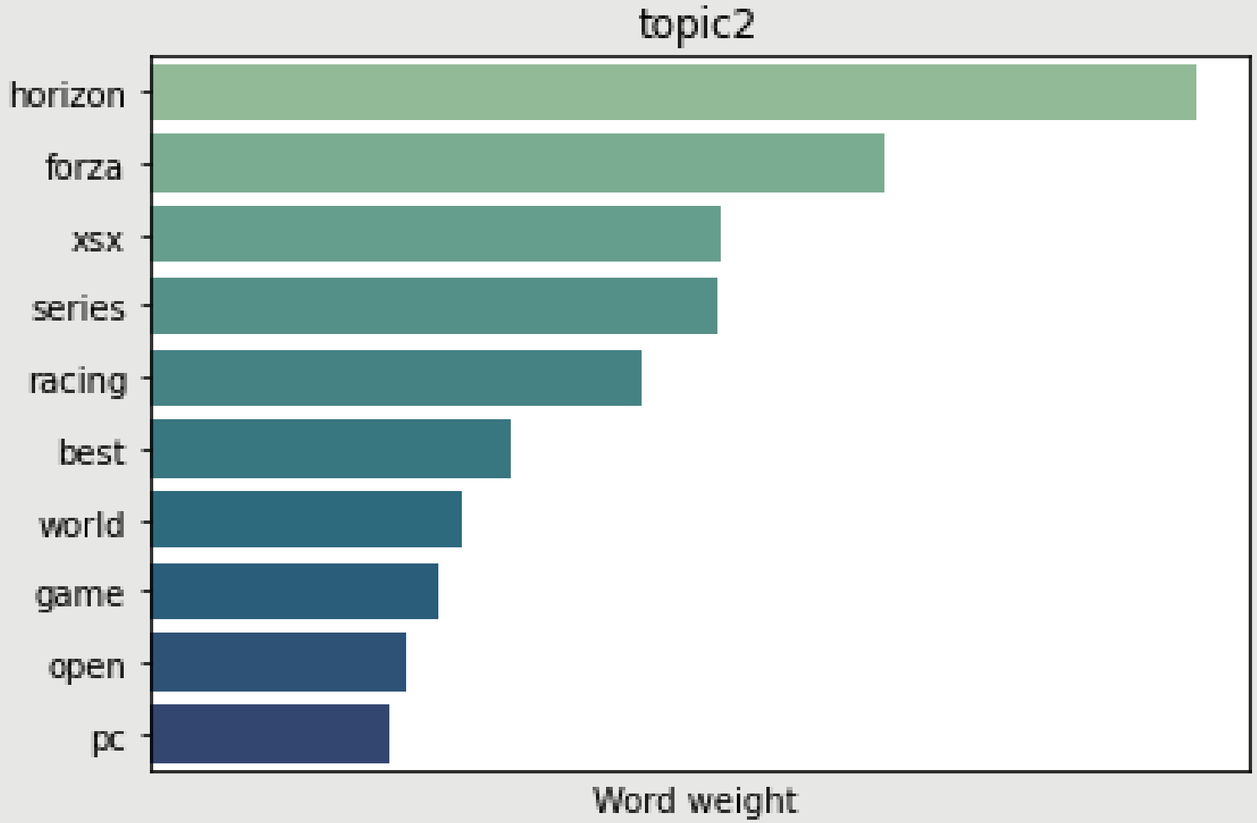



number of topics vs coherence score
Coherence score
0.58
0.56
0.54
0.52
0.50
5
10
15
20
25
30
35
40
45
50
Number of topics

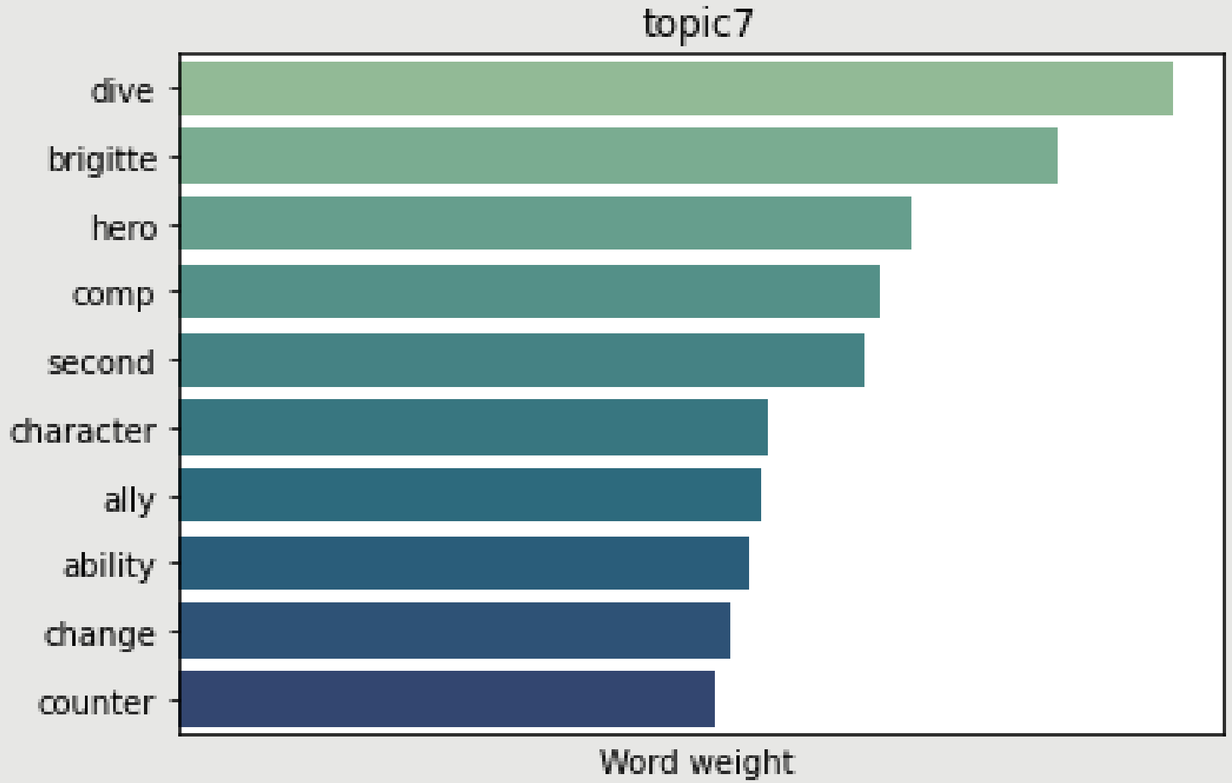
Results

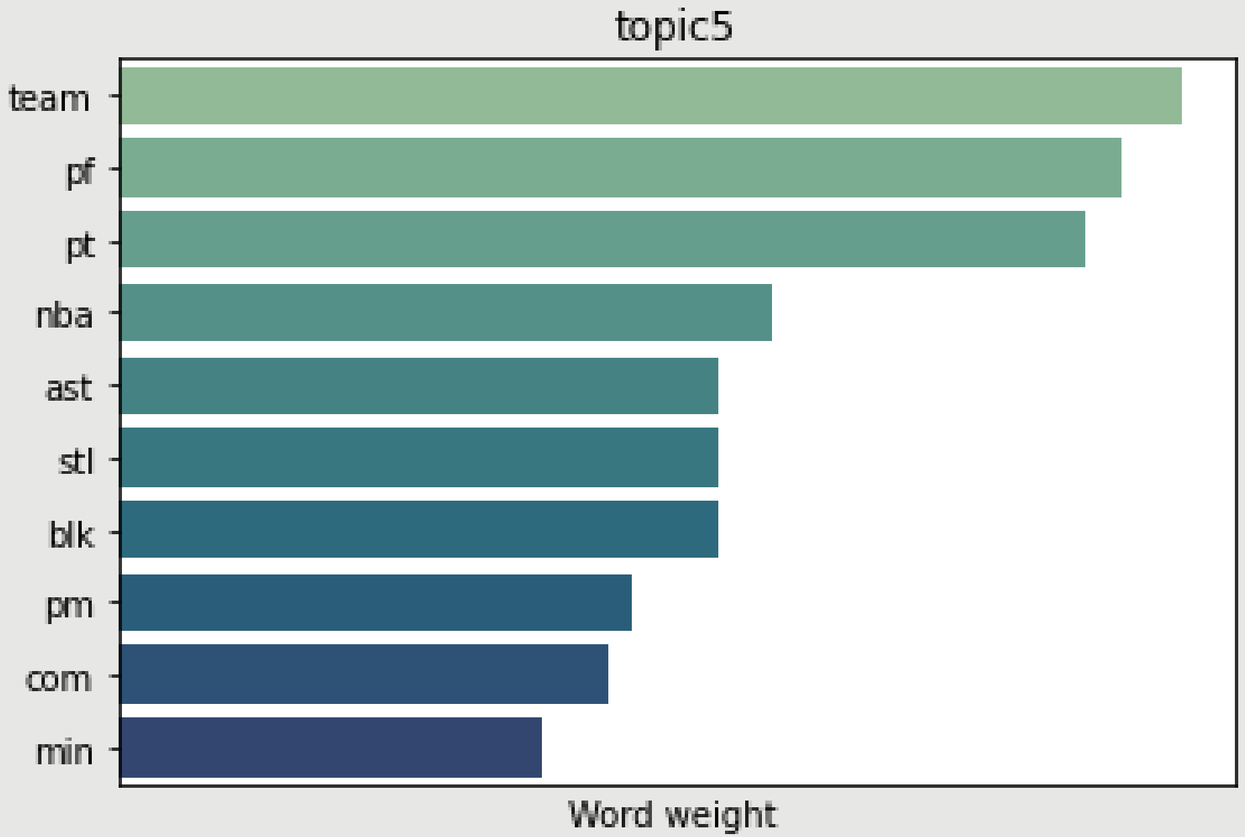

topic1
team
player
nba
league
year
school
new
basketball
contract
organization
Word weight


NBA


topic2
horizon
forza
xsx
series
racing
best
world
game
open
pc
Word weight

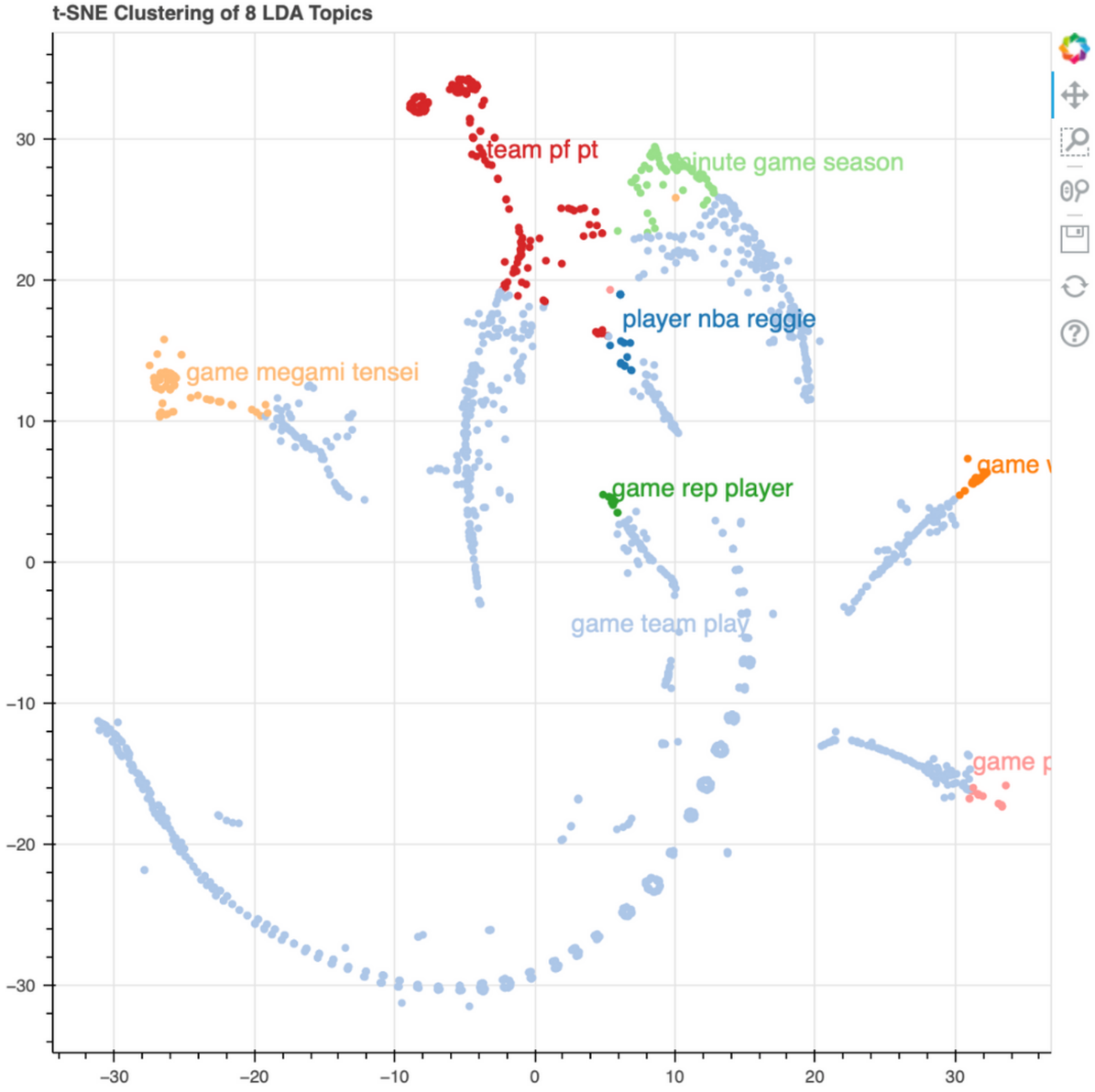

FORZA
HORIZON 5

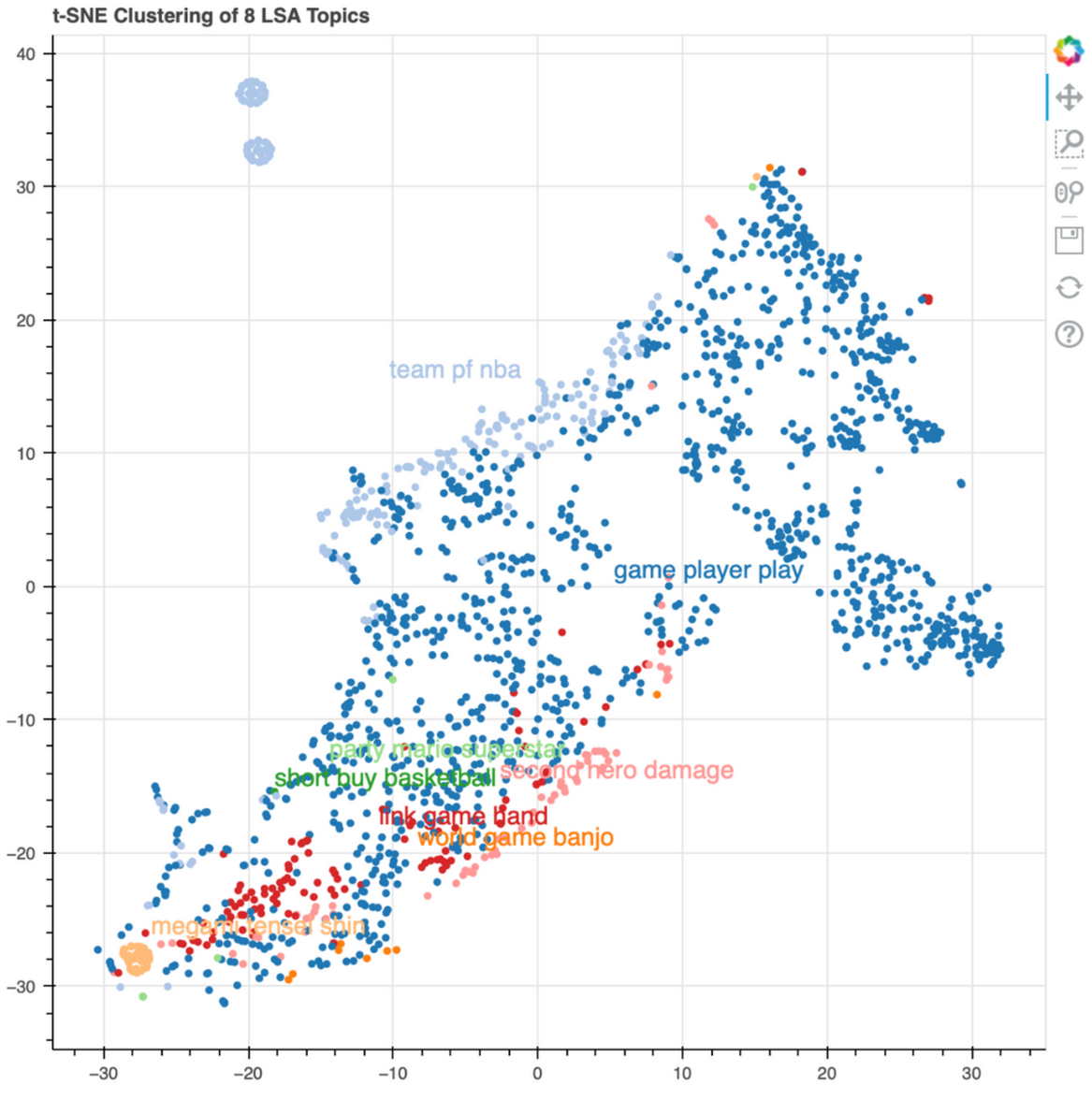

topic7
dive
brigitte
hero
comp
second
character
ally
ability
change
counter
Word weight



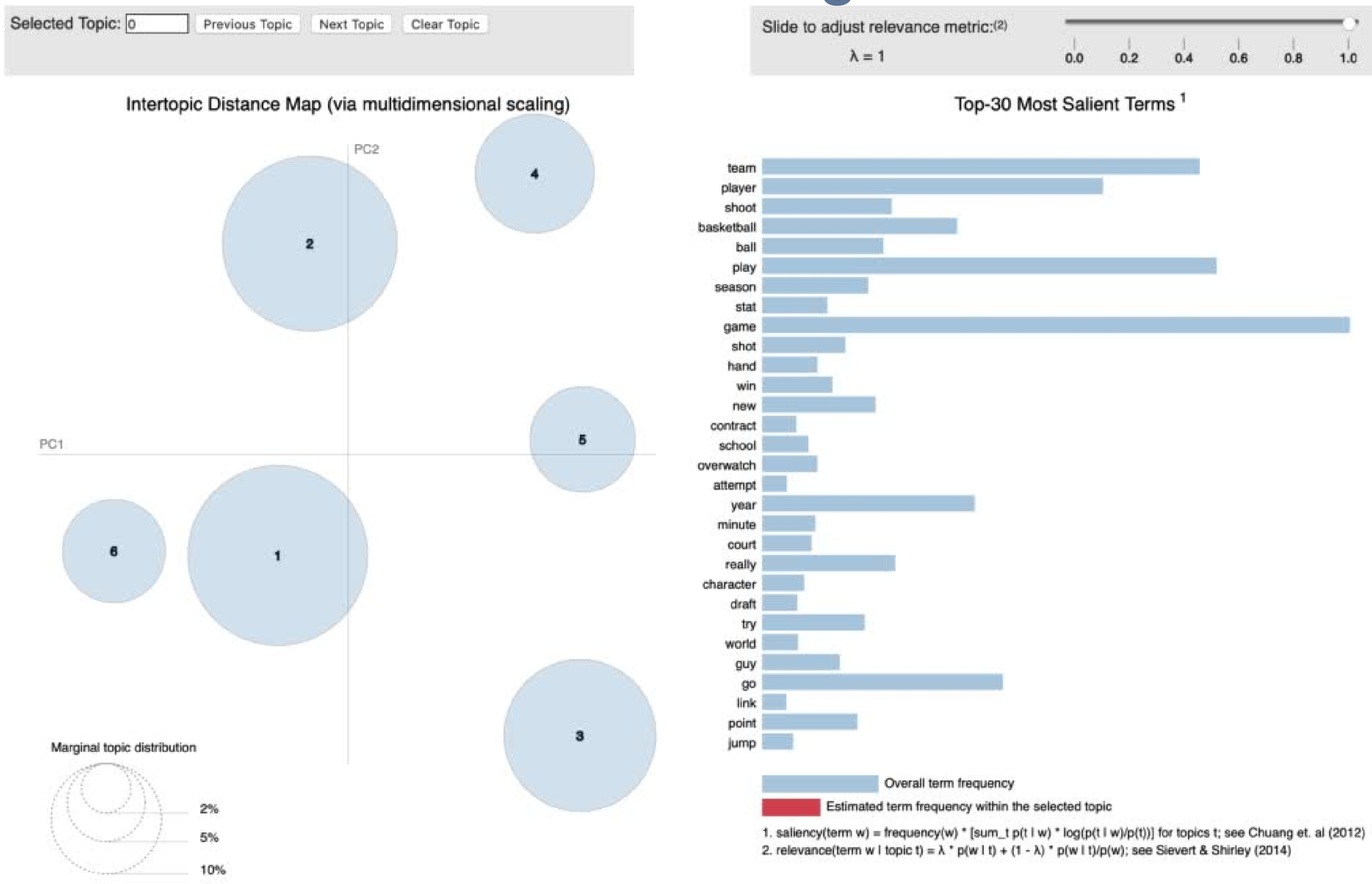
ORB	DRB	REB	AST	STL	BLK	TO	PF
0	2	2	3	0	0	1	4

Models Testing

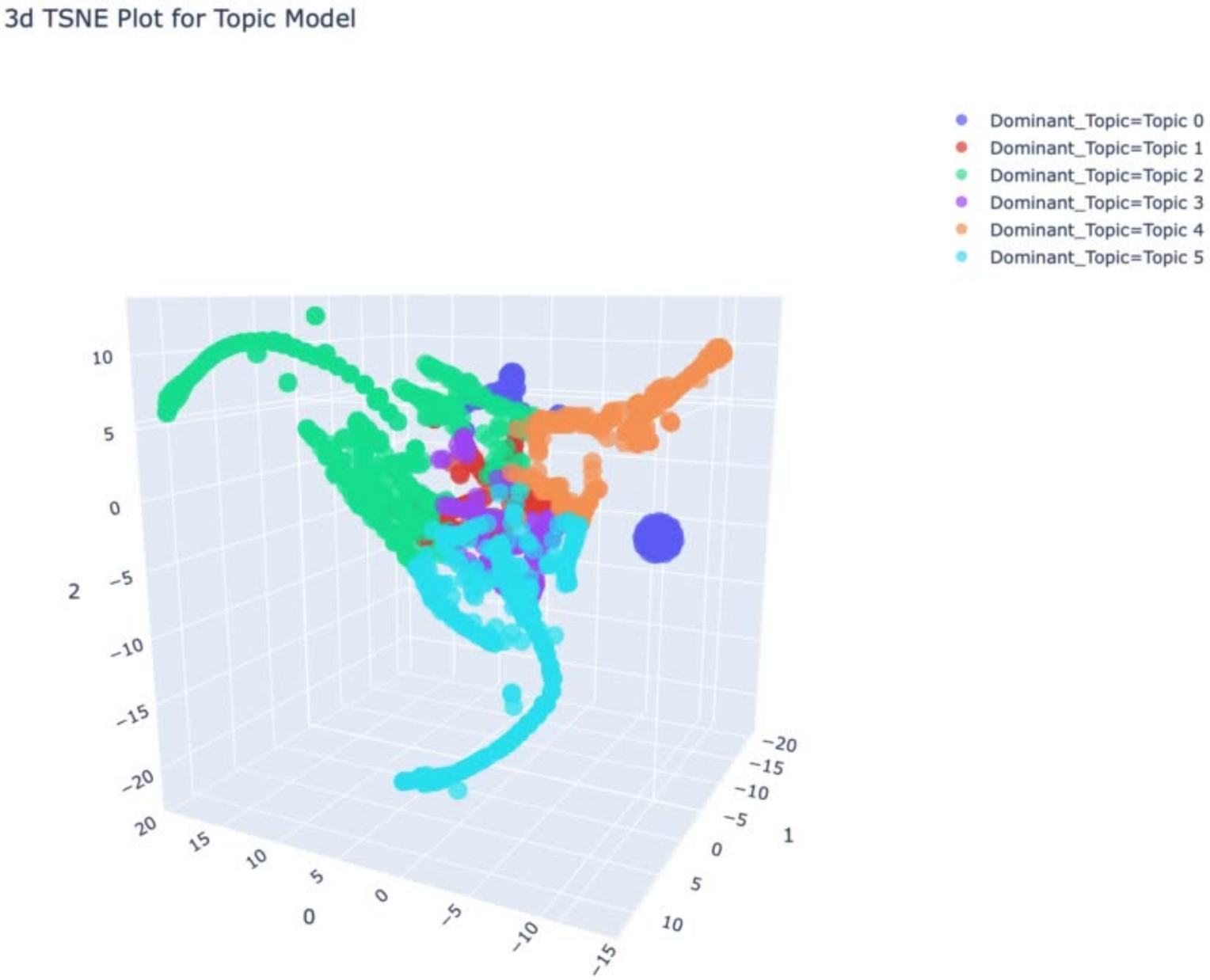




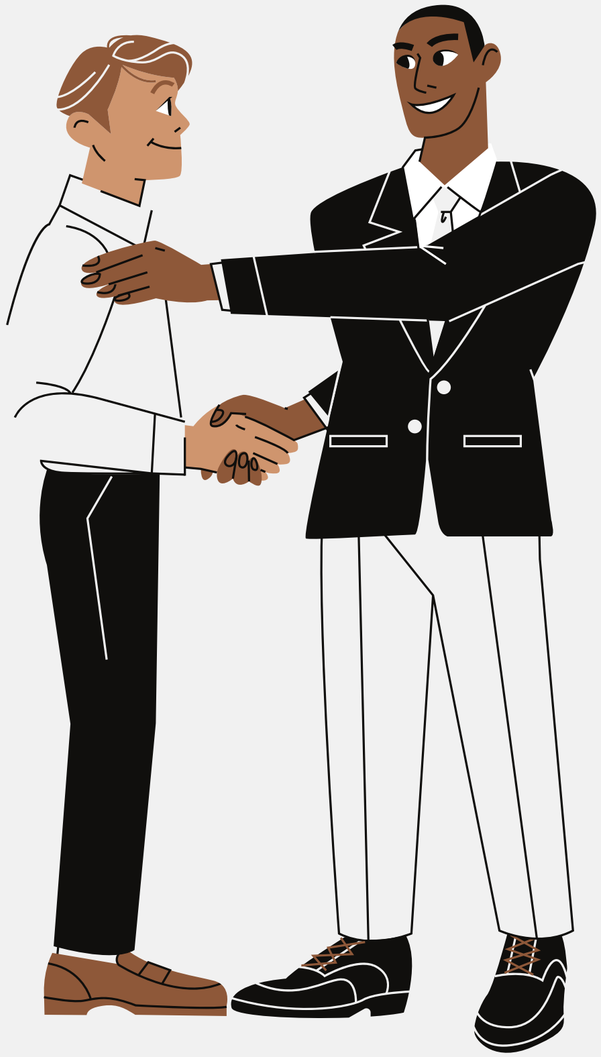
LDA Tunning


Selected Topic: 0
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
4
5
6
Marginal topic distribution
2%
5%
10%
Top-30 Most Salient Terms 1
team
player
shoot
basketball
ball
play
season
stat
game
shot
hand
win
new
contract
school
overwatch
attempt
year
minute
court
really
character
draft
try
world
guy
go
link
point
jump
Overall term frequency
Estimated term frequency within the selected topic
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)

LDA 3D


3d TSNE Plot for Topic Model
Dominant_Topic=Topic 0
Dominant_Topic=Topic 1
Dominant_Topic=Topic 2
Dominant_Topic=Topic 3
Dominant_Topic=Topic 4
Dominant_Topic=Topic 5
10
5
0
−5
−10
−15
−20
2
20
15
10
5
0
0
−5
−10
−15
−20
−15
−10
−5
0
5
10
1

Conclusion



- **Reddit topics and posts need continuous cleaning and gathering.**
- **Need to explore more in more algorithms in topic modeling for Reddit.**
- **Ensure text quality modeling.**

Thank You For Listening.

Do you have any questions?