

## Scenario

You’re an up and coming analytics professional working at a digital marketing agency. Your bosses rely on your excellent analytical skills, business acumen and innovative methods to help their clients. The marketing firm represents the Call of Duty League, an e-sports league with unique teams and players, some with large followings. The League is exploring adding a new team, the “Cleveland Fire”, and wants to understand any themes related to existing teams’ social media success as well as the current state of the League’s social media presence.

During client meetings, these questions emerged as a starting point but you are free to explore the data in your own manner, addressing one, all or none of below:

* Some teams/players/followers may be antagonistic (negative) or display varying emotions. Can these be compared?
* Are there natural topics that emerge among all teams, players, or followers or are any distinctive?
* Can social media analytics help measure the volume of specific terms like “gg” or “good game” (or other specific gaming terms)?
* Conversations related to the league, competitions, mentions of a competitor league called “Warzone” would be of interest
* Basic technical issues/glitches that represent product issues or improvement

## Contextual Information:

The Call of Duty League is an e-sports league based on Activision’s Call of Duty video game. Call of Duty is a “first person shooter” that is more than 15 years old and has sold 300m copies. Call of Duty is the third highest grossing video game ever (~$18B lifetime revenue). The league has city-based teams under individual ownership each with their own changing rosters. The league plays in a tournament point system with players earning a minimum salary and a portion of revenue sharing based on the team’s points and tournament performance. The league was announced in 2019, started in 2020, holding a mix of in person and remote events.

## Teams:

The following teams exist, have been announced, or added. *Due to the emerging nature of the league, some teams may have been renamed, or sold without reliable information.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **City** | **Logo** | **Webpage** | **Twitter Handle** | **Followers** |
| Atlanta Faze |  | <https://faze.callofdutyleague.com/en-us> | ATLFaZe | 76.7k |
| Dallas Empire |  | <https://empire.callofdutyleague.com/en-us> | dallasempire | 92.3k |
| Florida Mutineers |  | <https://mutineers.callofdutyleague.com/en-us> | mutineers | 53k |
| London Royal Ravens |  | <https://ravens.callofdutyleague.com/en-us> | RoyalRavens | 56.9k |
| LA Guerrillas |  | <https://guerrillas.callofdutyleague.com/en-us> | LAGuerrillas | 33.8k |
| LA Thieves |  | NA | LAThieves | 100.6k |
| Minnesota Rokkr |  | <https://rokkr.callofdutyleague.com/en-us> | rokkr | 58.4k |
| NY Subliners |  | <https://subliners.callofdutyleague.com/en-us> | Subliners | 58k |
| Optic Chicago |  | <https://optic.callofdutyleague.com/en-us> | OpTicCHI | 211.7k |
| Paris Legion |  | <https://legion.callofdutyleague.com/en-us> | ParisLegion | 34.9k |
| Seattle Surge |  | <https://surge.callofdutyleague.com/en-us> | SeattleSurge | 53.7k |
| Toronto Ultra |  | <https://ultra.callofdutyleague.com/en-us> | torontoultra | 58.4k |
| Immortals\* | Image | \*This team was recently sold and was originally Optic LA <https://cod-esports.gamepedia.com/OpTic_Gaming_Los_Angeles> | immortals | 1335k |

## Players:

Although players change teams, get drafted or get dropped the following list has players as part of the case data. *Due to the emerging nature of the league, some teams may have been renamed, and some players like `Kuavo` are listed twice.*

```{r}

teams <- list('ATLFaZe' = c('abezy','Arcitys', 'Cellium', 'majormaniak', 'Priestahh','SimpXO'),

'dallasempire' = c('Shotzzy', 'iLLeYYY', 'Huke', 'Crimsix'),

'mutineers' = c('ColtHavok', 'Maux','CesarSkyz','FrostyBB','f3rocitys'),

'RoyalRavens' = c('wuskinz', 'skrapzg','Nastiee\_','jurd','dylancod\_'),

'LAGuerrillas' = c('Blazt','itsSpart','UAquaa','Decemate','VividTheWarrior'),

'LAThieves' = c('Drazah\_', 'Kuavo', 'TJHaLy'),

'rokkr' = c('silly702','Assault','alexx1935','GstaAsim'),

'Subliners' = c('Attach', 'ZooMaa','AccuracyLA','Temp','OpSuda'),

'OpTicCHI' = c('scump', 'FormaL','DylanEnvoy','DashySZN'),

'ParisLegion' = c('DenzJT','KiSMET6\_','MF\_Louqa','ShockzCR','ZachDenyer'),

'SeattleSurge' = c('Apathy\_BZ','DKarma','OctaneSam','Slacked','CaseyPandur'),

'torontoultra' = c('CammyMVP','Classic','Loony','Methodz','bance'),

'Immortals' = c('JKap415','SlasheR\_AL','TJHaLy','DashySZN','Kuavo'))

```

## Technical:

* Depending on how you decide to review the problem you may apply text mining processes to the entire league, a subset or teams, a single team or similarly by players or even followers of teams and/or players.
* The amount of text is meant to be overwhelming as this is a realistic text-based case, and you must decide how to proceed, there is no correct level of granularity or “right” way to satisfy the client needs. Ambiguity must be overcome with a proper problem statement(s) and selection of the appropriate data for the statement.
* 文本的数量意味着压倒性的，因为这是一个现实的基于文本的情况，你必须决定如何继续，没有正确的粒度级别或“正确的”方式来满足客户的需求。必须用适当的问题陈述和为陈述选择适当的数据来克服歧义。

## Non-Technical:

* Describe the preprocessing steps and why are they are applied to the documents
* Describe the various techniques used to create the visuals and findings in a PowerPoint to a *non-technical user within the marketing firm.*

## Project Deliverables include

1. R scripts for data processing & exploration “lastName\_TM\_CallOfDuty\_case.R”
   1. Your script(s) must account for all aspects of the material in your presentation to ensure the presentation is data driven (no cheating with Excel or other tools!)
2. PowerPoint of any visualizations, findings and descriptions of non-technical results as if presented to your boss and ultimately client. “lastName\_TM\_CallOfDuty\_case\_.pptx”
   1. The PowerPoint must be accompanied by a voice over embedded in the file, or screenshare video uploaded

## Data & Helpers

* All data comes from the Twitter API except the `emojis.csv` lookup table which was sourced [here](https://unicodey.com/emoji-data/table.htm).
* Tweet data is provided in `fst` format. This requires R’s `fst` package and can be read using `read\_fst` using the file path. This file format is efficient, and small when saved but you may save local copies in any format i.e. CSV.
* Stopwords for the twitter channel can be standard lexicons `en` or `SMART` or a custom one is [here](o%09%20https:/sites.google.com/site/iamgongwei/home/sw)(twitter-stopwords.csv)
* Hashtags, which may or may not be useful, have been identified in the API as a separate column but you may also use this code in case some were missed.

```{r}

library(stringr)

hashtags <- str\_extract\_all(textVector, '(?<=^|\\s)#\\S+')

```

* Unicode Emojis ☺ can be removed with the following code. This is faster though may not be as informative.

`gsub("[^\x01-\x7F]", "", textVector)`

* Alternatively, you can lookup and substitute emojis using `emojis.csv`. Be warned this takes a long time to perform given the code below and may not be feasible given the case deadline:

```{r}

library(mgsub)

library(pbapply)

emoji <- read.csv('emojis.csv')

subbedTxt <- pbsapply(textVector, mgsub, emoji$emoji, emoji$name)

```

There are 4 data themes to explore within the `studentData` folder. You are free to choose one or more during your case. *Due to individual user settings, the API may not get a user’s timeline.*

`teamTimeline` FOLDER - this file has one file with 26,781 tweets among 13 teams. *The maximum per team is 3200 tweets per API restrictions.* This is likely the easiest to explore differences in the topics, polarity/sentiment and information between teams.

`playerTimelines` FOLDER – contains 13 files, one per team, totaling 177,764 tweets. *The maximum per player API request is 3200 tweets.* This folder may indicate differences among players by team or within a specific team.

`teamFollowerTimelines` FOLDER – contains 13 files, one per team, totally 361,440 tweets. This represents a *sample* of 3600 followers for each team among the followers listed in the previous table. 10 of the most recent tweets were requested from each follower’s timeline. This data may indicate how much of the recent dialog is attributable to the league or teams, or other topics in addition to follower personas of sentiment/polarity and topics by team.

`playerFollowerTimelines` FOLDER – contains 61 files, one per player, totally 2,006,995 tweets. This represents a *sample* of 500 followers for each player in the league. 100 of the most recent tweets were requested from each player follower timeline. This data may indicate how much of the recent dialog is attributable to specific players, league or teams, or other topics in addition to follower personas of sentiment/polarity and topics for each player. The team and followed player ID is part of each file name.

## Data Dictionary

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Column Name** | **Description** | **teamTimeline** | **playerTimelines** | **teamFollowerTimelines** | **playerFollowerTimelines** |
| user\_id | Unique identifier of the author | Y | Y | Y | Y |
| status\_id | Unique identifier of the tweet | Y | Y | Y | Y |
| created\_at | timestamp | Y | Y | Y | Y |
| screen\_name | Handle of the author | Y | Y | Y | Y |
| Text | Tweet text | Y | Y | Y | Y |
| hashtags | API identified hashtags | Y | Y | Y | Y |
| Team | Associated team |  | Y | Y | Y |
| Player | Associated player |  |  |  | Y |

## Criteria for Success

The case material will be evaluated according to the following criteria. Each is worth 5pts for a total of 20pts.

## **Organization of content**– Logical ordering of ideas, modeling artifacts, applicable visualizations in slides

## **Organization of code**- R Code is well organized, concise, and free from error

## **Text mining process** – Recognize the type of data mining problem, adherence to established main data and text mining steps.

## **Completeness** – Understood impact, and mined the data for relevant insights/recommendations

# Read in multiple files as individuals

txtFiles <- list.files(pattern = 'teamFTl1|teamFTl2')

for (i in 1:length(txtFiles)){

assign(txtFiles[i], read.csv(txtFiles[i]))

cat(paste('read completed:',txtFiles[i],'\n'))

}

# Vector Corpus; omit the meta data

teamFTl1 <- VCorpus(VectorSource(read.fst('student\_2020-12-28\_LAGuerrillas2\_followers\_timelines.fst',from = 1, to = 1000)$test))

teamFTl2 <- VCorpus(VectorSource(read.fst('student\_2020-12-28\_LAThieves2\_followers\_timelines.fst',from = 1, to = 1000)$text))

# Clean up the data

teamFTl1 <- cleanCorpus(teamFTl1, stops)

teamFTl2 <- cleanCorpus(teamFTl2, stops)

# Instead of 1000 individual documents, collapse each into a single "subject" ie a single document

teamFTl1 <- paste(teamFTl1 , collapse = ' ')

teamFTl2 <- paste(teamFTl2, collapse = ' ')

# Combine the subject documents into a corpus of \*2\* documents

allTeamFtls <- c(teamFTl1,teamFTl2)

allTeamFtls <- VCorpus((VectorSource(allTeamFtls)))

# Make TDM with a different control parameter

ctrl <- list(weighting = weightTfIdf)

TeamFtlsTDM <- TermDocumentMatrix(allTeamFtls, control = ctrl)

TeamFtlsTDMm <- as.matrix(TeamFtlsTDM)

# Make sure order is the same as the c(objA, objB) on line ~80

colnames(TeamFtlsTDM) <- c('teamFTl1', 'teamFTl2')

# Examine

head(TeamFtlsTDMm)

# Make comparison cloud

comparison.cloud(TeamFtlsTDMm,

max.words=75,

random.order=FALSE,

title.size=0.5,

colors=brewer.pal(ncol(TeamFtlsTDMm),"Dark2"),

scale=c(3,0.1))