Twitter Analysis of Dota2 Group Stage

Introduction

Dota 2 is a free to play multiplayer online battle arena (MOBA) game. The recent tournament The International 2018 (TI8) has begin started from 15/08/2018 to 25/08/2018 which have a price pool of 24,721,548 millions USD. It has one of the largest prize pool across all the Esport tournament. This year the tournament is held at Vancouver, Canada. For more information of the tournament, you can see this link.

In this project, I'm going to analyze what people tweeted during the tournament. There are two part of this project. I will analyze the tweet during the tournament group stage and main event.

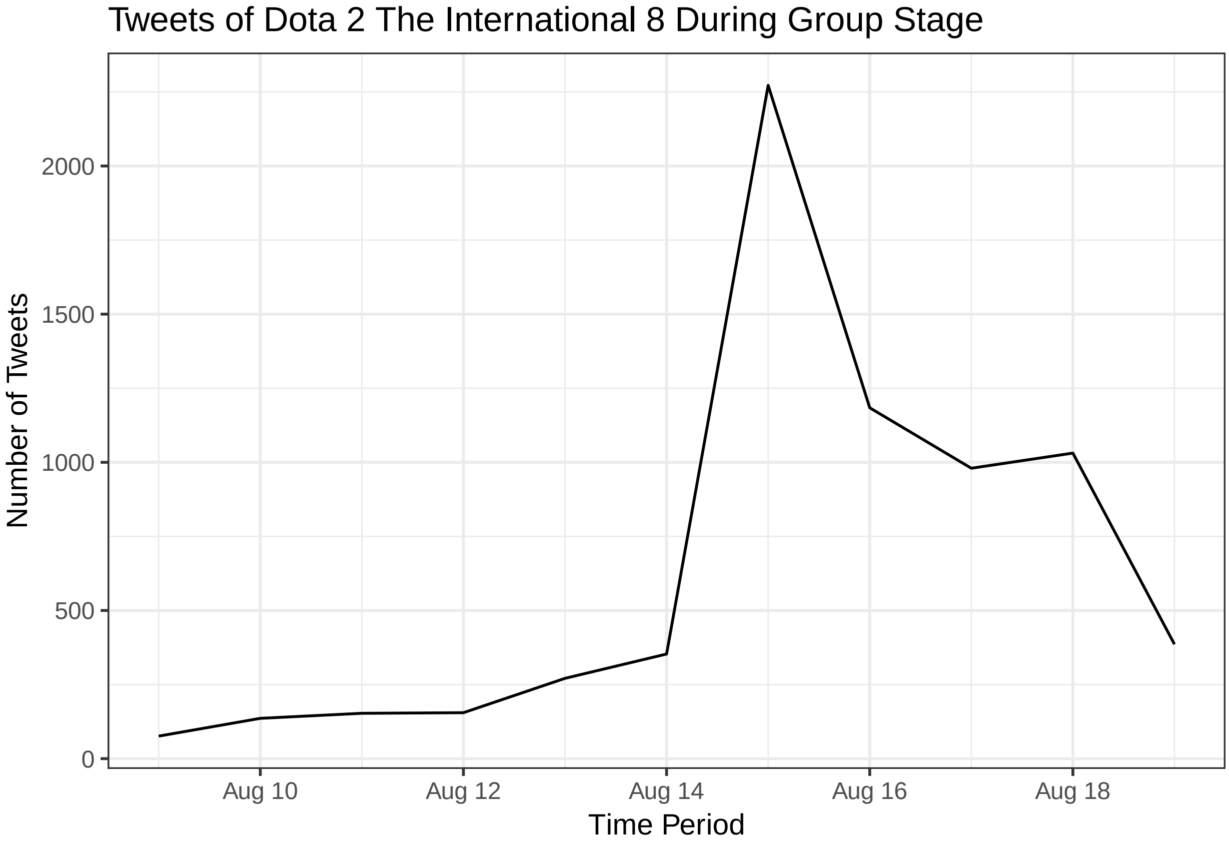
```
library(rtweet) #packages for tweets
library(lubridate) #packages for time manipulation
library(tidyverse) #package for ggplot, dplyr, tidyr
library(wordcloud) #to generate wordcloud
library(tidytext) #text mining package
library(stringr) #string manipulation
library(tm) #another text mining packages
library(RColorBrewer) # colour for R
library(knitr) # table formulation
library(pander) #pretty table format

## search for 20000 tweets using the keywords above but does not include retweet and only include englis
tweets <- search_tweets(q = "#ti8", n = 20000, include_rts = FALSE, retryonratelimit = TRUE, lang = "en

#structure of tweets
glimpse(tweets)
```

```
## Observations: 6,997
## Variables: 88
## $ user_id                 <chr> "22355229", "22355229", "2431922316", ...
## $ status_id               <chr> "1031169252240191490", "10306232151171...
## $ created_at              <dttm> 2018-08-19 13:21:22, 2018-08-18 01:11...
## $ screen_name             <chr> "lisyk", "lisyk", "overtesports", "Dot...
## $ text                    <chr> "bracket baby #TI8 https://t.co/3udBdj...
## $ source                  <chr> "Twitter Web Client", "Twitter for iPh...
## $ display_text_width      <dbl> 17, 51, 101, 129, 96, 111, 103, 125, 1...
## $ reply_to_status_id      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ reply_to_user_id        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ reply_to_screen_name    <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ is_quote                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ is_retweet              <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ favorite_count          <int> 1, 3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ retweet_count           <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ hashtags                <list> ["TI8", "TI8", "TI8", "TI8", "TI8", "...
## $ symbols                 <list> [NA, NA, NA, NA, NA, NA, NA, NA, NA, ...
## $ urls_url                <list> [NA, NA, "ift.tt/2vUkBwA", "youtu.be/...
## $ urls_t.co               <list> [NA, NA, "https://t.co/Kh96oQ92h6", "...
## $ urls_expanded_url       <list> [NA, NA, "https://ift.tt/2vUkBwA", "h...
## $ media_url               <list> ["http://pbs.twimg.com/media/Dk9y8Y9V...
## $ media_t.co              <list> ["https://t.co/3udBdjK5rJ", NA, NA, N...
## $ media_expanded_url      <list> ["https://twitter.com/lisyk/status/10...
## $ media_type              <list> ["photo", NA, NA, NA, NA, NA, NA, NA,...
## $ ext_media_url           <list> ["http://pbs.twimg.com/media/Dk9y8Y9V...
## $ ext_media_t.co          <list> ["https://t.co/3udBdjK5rJ", NA, NA, N...
## $ ext_media_expanded_url  <list> ["https://twitter.com/lisyk/status/10...
## $ ext_media_type          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ mentions_user_id        <list> [NA, NA, NA, "10228272", NA, "1022827...
## $ mentions_screen_name    <list> [NA, NA, NA, "YouTube", NA, "YouTube"...
## $ lang                    <chr> "en", "en", "en", "en", "en", "en", "e...
## $ quoted_status_id        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_text             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_created_at       <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ quoted_source           <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_favorite_count   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_retweet_count    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_user_id          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_screen_name      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_name             <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_followers_count  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_friends_count    <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_statuses_count   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_location         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_description      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ quoted_verified         <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_status_id       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_text            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_created_at      <dttm> NA, NA, NA, NA, NA, NA, NA, NA, NA, N...
## $ retweet_source          <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_favorite_count  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_retweet_count   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_user_id         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_screen_name     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_name            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_followers_count <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_friends_count   <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_statuses_count  <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_location        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_description     <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ retweet_verified        <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_url               <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_name              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_full_name         <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ place_type              <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ country                 <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ country_code            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA...
## $ geo_coords              <list> [<NA, NA>, <NA, NA>, <NA, NA>, <NA, N...
## $ coords_coords           <list> [<NA, NA>, <NA, NA>, <NA, NA>, <NA, N...
## $ bbox_coords             <list> [<NA, NA, NA, NA, NA, NA, NA, NA>, <N...
## $ status_url              <chr> "https://twitter.com/lisyk/status/1031...
## $ name                    <chr> "Lisy K \U0001f33a", "Lisy K \U0001f33...
## $ location                <chr> "Melbourne, VIC", "Melbourne, VIC", ""...
## $ description             <chr> "producer @leagueofgeeks ~ co-founder ...
## $ url                     <chr> "https://t.co/YRBB7MsEz7", "https://t....
## $ protected               <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ followers_count         <int> 4041, 4041, 255, 195, 195, 195, 195, 1...
## $ friends_count           <int> 2059, 2059, 108, 312, 312, 312, 312, 3...
## $ listed_count            <int> 179, 179, 49, 3, 3, 3, 3, 3, 3, 3, 3, ...
## $ statuses_count          <int> 41392, 41392, 34732, 3710, 3710, 3710,...
## $ favourites_count        <int> 80999, 80999, 0, 1737, 1737, 1737, 173...
## $ account_created_at      <dttm> 2009-03-01 11:37:27, 2009-03-01 11:37...
## $ verified                <lgl> FALSE, FALSE, FALSE, FALSE, FALSE, FAL...
## $ profile_url             <chr> "https://t.co/YRBB7MsEz7", "https://t....
## $ profile_expanded_url    <chr> "http://www.lisykane.com", "http://www...
## $ account_lang            <chr> "en", "en", "en", "vi", "vi", "vi", "v...
## $ profile_banner_url      <chr> "https://pbs.twimg.com/profile_banners...
## $ profile_background_url  <chr> "http://abs.twimg.com/images/themes/th...
## $ profile_image_url       <chr> "http://pbs.twimg.com/profile_images/9...
```

First, we going to gather the tweets using **search_tweets** function. To search the tweets, I used the keywords "*#ti8*" to gather all the tweets that have included the keywords. In addition, I also exclude all the retweets.

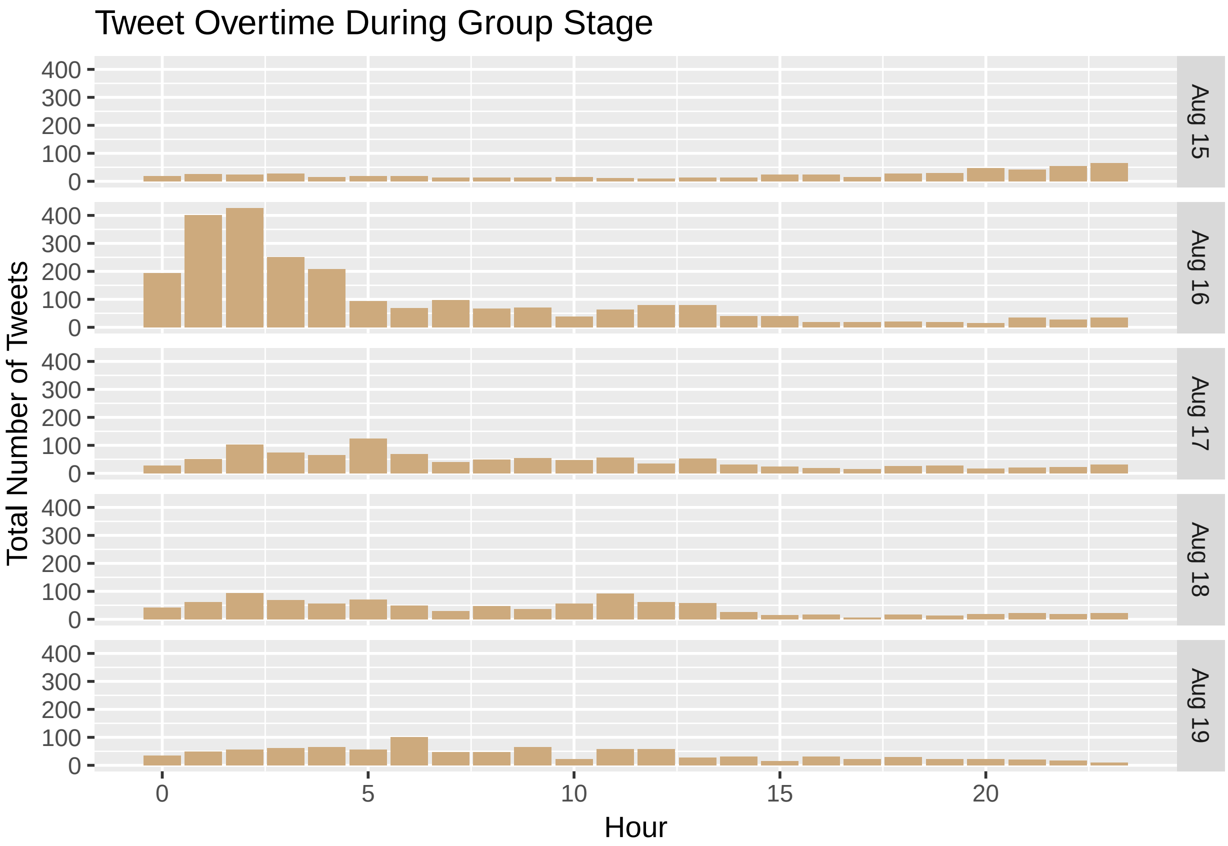
Based on the function, I able to gather more than 6000 tweets.

```
# Tweets overtime
ts_plot(tweets) +
  theme_bw() +
  labs(title = "Tweets of Dota 2 The International 8 During Group Stage",
       x = "Time Period",
       y = "Number of Tweets")
```



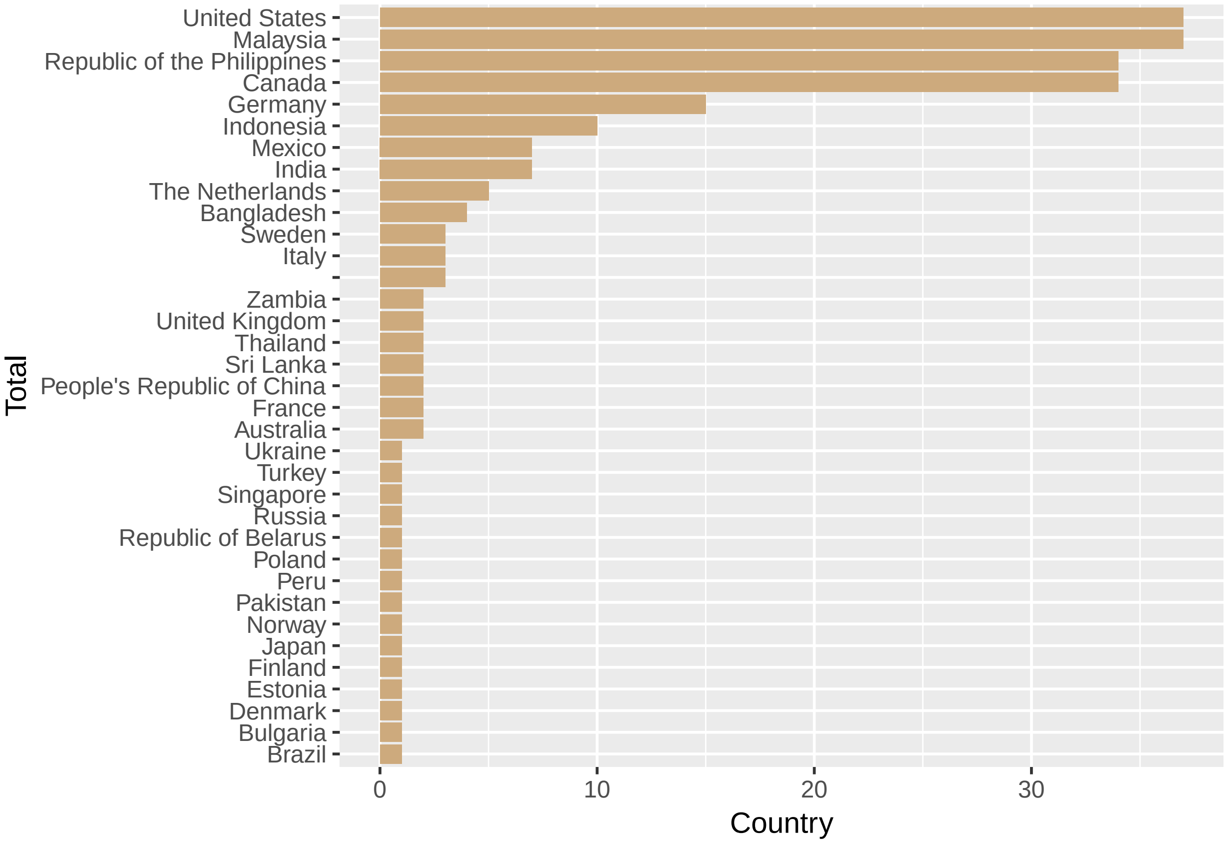
The plot above allow us to examine the number of tweets that have accur during the event. It shows that most tweets have tweeted during the day that group stage begin more that 2000 tweets tweeted during the events. After that the tweets slowly decline as it towards the end of TI8 group stages.

```
# Tweets overtime by day
time <- tweets %>%
  mutate(datetime = as_datetime(created_at, tz = "Australia/Melbourne"), hour = hour(datetime)) %>%
  group_by(date = as_date(datetime), hour) %>%
  summarise(total = n()) %>%
  filter(date >= as_date("2018-08-15"), date <= as_date("2018-08-19"))

time %>% ggplot(aes(hour, total)) +
  geom_col(fill = "burlywood3") +
  facet_grid(strftime(date, "%b %d") ~. )  +
  xlab("Hour") +
  ylab("Total Number of Tweets") +
  ggtitle("Tweet Overtime During Group Stage")
```



The plot above allows us to examine the hours where most tweets occurs. As you can see, most tweets occours during the hype of the event which is around 12 am to 4am in Australia time.

```
country <-
  tweets %>%
  select(country) %>%
  group_by(country) %>%
  summarise(total = n()) %>%
  na.omit()

country %>%
  ggplot(aes(x = reorder(country, total), y = total)) +
  geom_col(fill = "burlywood3") +
  xlab("Total") +
  ylab("Country") +
  coord_flip()
```



Next we are going to see which country tweeted the most. Malaysia has the most tweets followed by USA, Canada, Philipine and Germany. As one thing need to be reminded, not all the tweets have include their location, where majority of the tweets does not included their location.

```
## Tweets that has most favourite during group state
favourite <-
  tweets %>%
  select(screen_name, created_at, text, favorite_count) %>%
  arrange(desc(favorite_count)) %>%
  distinct() %>%
  mutate(datetime = as_datetime(created_at, tz = "Australia/Melbourne"), hour = hour(datetime)) %>%
  group_by(date = as_date(datetime)) %>%
  select(screen_name, date, text, favorite_count) %>%
  filter(date >= as_date("2018-08-15"), date <= as_date("2018-08-19")) %>%
  arrange(desc(favorite_count)) %>%
  slice(1:20) %>%
  pander(justify = c("left", "left", "right", "right"), split.table = Inf)
```

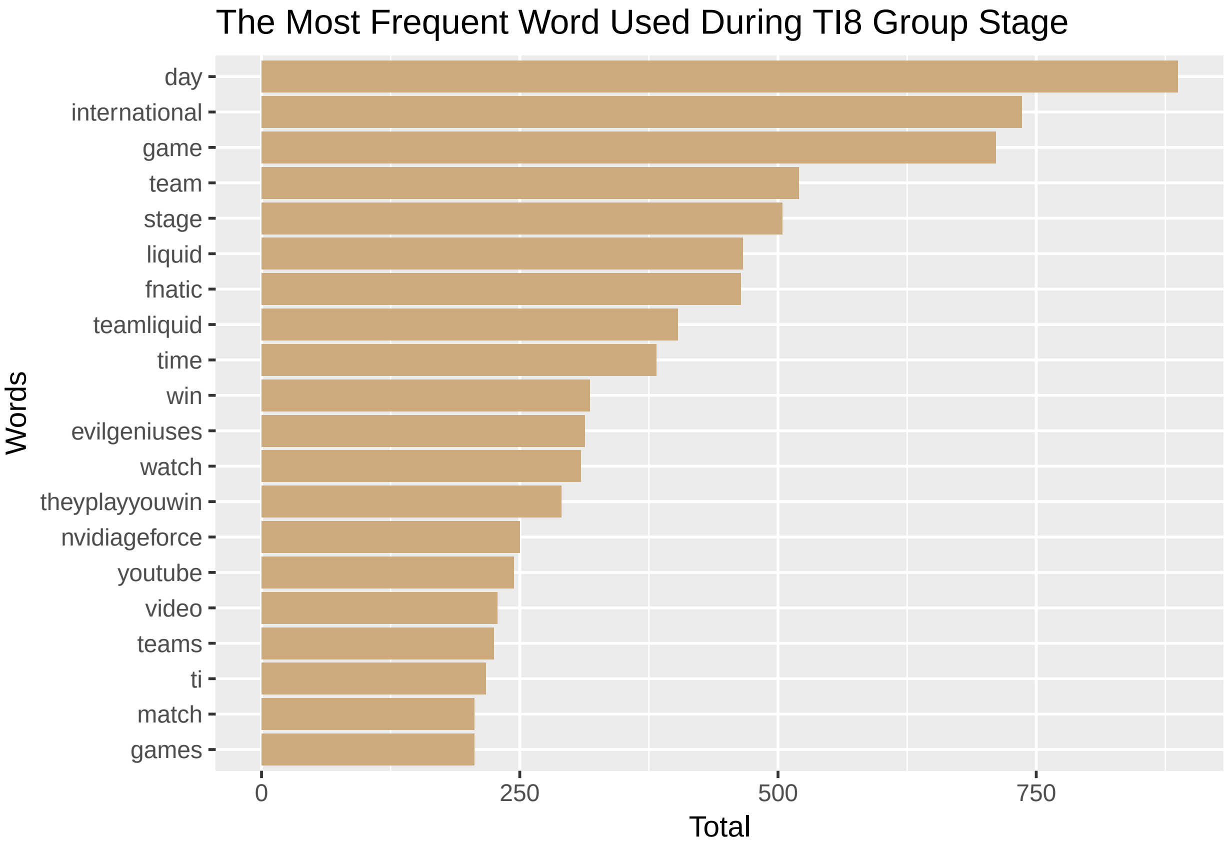
The table above is an overview of most favourite tweets during the group stage. As you can see, twitter user wykrhm tweets have the most favourite. This is not suprising as he is one of the most favourite community figure on Dota 2.

```
replace_reg1 <- "https://t.co/[A-Za-z\\d]+|"
replace_reg2 <- "http://[A-Za-z\\d]+|&amp;|&lt;|&gt;|RT|https"
replace_reg <- paste0(replace_reg1, replace_reg2)
```

```r
unnest_reg <- "([^A-Za-z_\\d#@']|'(?![A-Za-z_\\d#@]))"

tidy_tweets <- tweets %>%
  mutate(text = str_replace_all(text, replace_reg, "")) %>%
  unnest_tokens(word, text) %>%
  anti_join(stop_words)
```

Next we are going to analyze the most used word during the TI8 group stage. We first begin to remove all the non-word chracters url and symbol from the text. Next, we put the text into tidy format. For more information about the tidy text format see this link

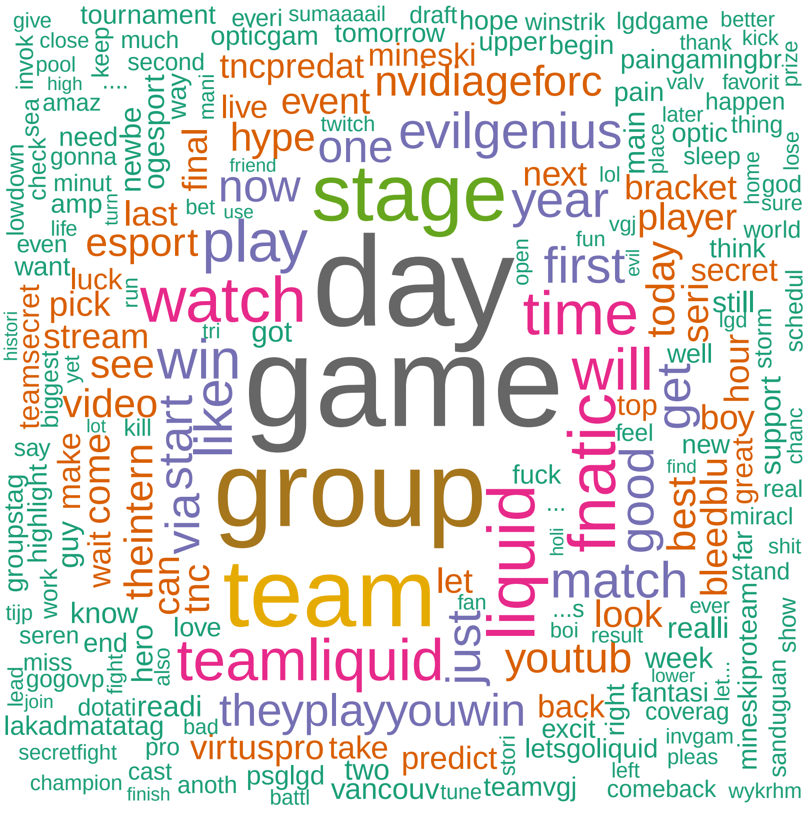
```r
frequency <- tidy_tweets %>%
  select(word) %>%
  group_by(word) %>%
  summarise(total = n()) %>%
  filter(word != "ti8" & word != "dota2", word != "dota") %>%
  top_n(25)

frequency <- frequency[-c(1:5),]

frequency %>%
  ggplot(aes(x = reorder(word, total), y = total)) +
  geom_col(fill = "burlywood3") +
  xlab("Words") +
  ylab("Total") +
  ggtitle("The Most Frequent Word Used During TI8 Group Stage") +
  coord_flip()
```



The resut shows that the most used the words is day, international game, team and etc. In addition, fnatic and liquid also appear on top of the list. This is because of the match where Team Fnatic win over Team Liquid with an astounding results with the score of 39 - 1. This maybe the reason why people tweet it so much. This result can be found in this link

Word Cloud

```r
docs <- Corpus(VectorSource(tweets$text))

# Convert the text to lower case
docs <- tm_map(docs, content_transformer(tolower))
# Remove numbers
docs <- tm_map(docs, removeNumbers)
# Remove english common stopwords
docs <- tm_map(docs, removeWords, stopwords("english"))
# Remove punctuations
docs <- tm_map(docs, removePunctuation)
# Eliminate extra white spaces
docs <- tm_map(docs, stripWhitespace)

docs <- tm_map(docs, stemDocument)
# Remove additional stopwords
docs <- tm_map(docs, removeWords, c("ti8", "TI8", "dota2", "dota", "intern"))

wordcloud(words = docs, min.freq = 1,
          max.words=300, random.order=FALSE, rot.per=0.35,
          colors=brewer.pal(8, "Dark2"))
```



Above is just the wordcloud based on the word frequency. But this time, I removed some of the common words such as TI8 and dota to gain more insight of the words frequency.