~~~
---
title: "lang_degrad_eda"
author: "Kroehler, Sodi"
date: "4/17/2022"
output: html_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

## CS1699 Final Project Exploratory Data Analysis Markup

```{r, load_packages}
library(tidyverse)
library(ggdark)
library(gganimate)
```

```{r, import_milestones_data}
processMilestones <-function(filename){
  milestones_raw <- readr::read_csv(filename, col_names = TRUE,locale = locale(encoding =
"utf-16") )
  df <- milestones_raw %>%
    mutate(point = as.Date(point),
           date = as.Date(date),
           #lat = as.double(str_sub(str_match_all(geo, "\\[\\d*.\\d*"), 2, -1L)),
           #long = as.double(str_sub(str_match(geo, "\\d*.\\d*\\]"), 1, -2L)))
    ) %>% group_by(point) %>%
    mutate(percEN = mean(cEN/cTOTAL),
           percHIN = mean(cHIN/cTOTAL)) %>%
    select(city, point, percEN, percHIN)
  return(df)
}

```
```{r}
mile1 <- processMilestones("./data/by_milestones1_1-2.csv")
mile2 <- processMilestones("./data/by_milestones2_3-10.csv")
mile3 <- processMilestones("./data/by_milestones3_12-13.csv")
mile4 <- processMilestones("./data/by_milestones4_16.csv")
mile5 <- processMilestones("./data/by_milestones5_20-22.csv")
mile6 <- processMilestones("./data/by_milestones6_25up.csv")
mile7 <- processMilestones("./data/by_milestones7_15-25.csv")
mile8 <- processMilestones("./data/by_milestones8_11--14.csv")
```
```{r}
grand_df_miles <- bind_rows(mile1,mile2,mile3,mile4,mile5,mile6,mile7,mile8)
#grand_df_miles <- bind_rows(mile1,mile2)
grand_df_miles %>% head()
```



```{r, point_graph}
grand_df_miles %>%
  pivot_longer(c(percEN, percHIN), names_to = "lang", values_to = "perc") %>%
  ggplot(mapping = aes(x = point)) +
  geom_point(aes(y = perc, color = lang)) +
  facet_wrap(~city) +
  dark_theme_gray() +
  labs(x = "Date",
       y ="Percentage Per Tweet",
       color = "Legend") +
   scale_color_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("white",
~~~

```
"red"))
```
```{r, geom_smooth_all}
grand_df_miles %>%
  pivot_longer(c(percEN, percHIN), names_to = "lang", values_to = "perc") %>%
  ggplot(mapping = aes(x = point)) +
  geom_smooth(aes(y = perc, color = lang)) +
  facet_wrap(~city) +
  dark_theme_gray()+
  labs(x = "Date",
       y ="Percentage Per Tweet",
       color = "Legend") +
   scale_color_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("white",
"red"))
```

```{r, geom_smooth_avg}
grand_df_miles_avg <- grand_df_miles %>%
  group_by(point) %>%
  select(point, percEN,percHIN) %>%
  mutate(avg_percEN = mean(percEN),
         avg_percHIN = mean(percHIN),)
grand_df_miles_avg %>%
  pivot_longer(c(avg_percEN, avg_percHIN), names_to = "lang", values_to = "perc") %>%
  ggplot(mapping = aes(x = point)) +
  geom_smooth(aes(y = perc, color = lang)) +
  geom_vline(xintercept = as.Date("2014-11-06")) +
  geom_vline(xintercept = as.Date("2017-10-04")) +
  geom_vline(xintercept = as.Date("2020-09-04")) +
  #geom_smooth(aes(y = percHIN), color = "red") +
  dark_theme_gray() +
    labs(x = "Date",
       y ="Percentage Per Tweet",
       color = "Legend") +
   scale_color_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("white",
"red"))
```

```{r, city_information}
df <- readr::read_csv("./india_states_capitals.csv", col_names = TRUE)

df <- df %>% mutate(pop = as.integer(Population),
                              city = rownames(df),
                              cityName = LargestCity) %>%
  select(city, cityName, pop)
named_grand_miles  <- merge(grand_df_miles, df, by = "city", all.x = TRUE)

grand_df_miles %>% head()
named_grand_miles %>% head()
```
Importing Sade's data:
```{r}
my_processFile <- function(filename){
  df_raw <- readr::read_delim(filename, delim = ",", col_names = TRUE,locale =
locale(encoding = "utf-16"))
  df_all <- df_raw %>%
    mutate(point = as.Date(date),
          cother = as.integer(str_extract(str_extract(counts, ".other.. \\d*"), " \\d*")),
          cen = as.integer(str_extract(str_extract(counts, "en.. \\d*"), " \\d*")),
          chin = as.integer(str_extract(str_extract(counts, "hin.. \\d*"), " \\d*")),
          cfw = as.integer(str_extract(str_extract(counts, "fw.. \\d*"), " \\d*")),
          cne = as.integer(str_extract(str_extract(counts, "ne.. \\d*"), " \\d*"))) %>%
    select(point, cother,cen,chin,cfw,cne,lang)
```

~~~
  df_all$cother[is.na(df_all$cother)] <- 0
  df_all$cen[is.na(df_all$cen)] <- 0
  df_all$chin[is.na(df_all$chin)] <- 0
  df_all$cfw[is.na(df_all$cfw)] <- 0
  df_all$cne[is.na(df_all$cne)] <- 0

  df_cs <- df_all %>%
  mutate(percENday = cen/(cen+chin+cfw+cother+cne),
         percHINday = chin/(cen+chin+cfw+cother+cne)) %>%
    group_by(point) %>%
    summarise(percEN = mean(percENday),
            percHIN = mean(percHINday),)
}
```

```{r}
file1 <- my_processFile("./processed_tweets_01.csv")
file2 <- my_processFile("./processed_tweets_02.csv")
file3 <- my_processFile("./processed_tweets_03.csv")
file4 <- my_processFile("./processed_tweets_04.csv")
file5 <- my_processFile("./processed_tweets_05.csv")
#file6 <- my_processFile("./processed_tweets_06.csv")
file7 <- my_processFile("./processed_tweets_07.csv")
```



```{r}
grand_df_all <- bind_rows(file1, file2, file3, file4, file5, file7)
grand_df_all %>%
  ggplot(mapping = aes(x = point)) +
  geom_point(aes(y = percEN)) +
  geom_point(aes(y = percHIN), color = "red")+
  dark_theme_gray()
```
**Importing the precog set:**

```{r}
df2 <- readr::read_delim("./data/POS Hindi English Code Mixed Tweets.tsv", col_names =
TRUE, na = c("", "NA"), delim = "\t",skip_empty_rows = FALSE)
DELHI_df <- df2 %>% mutate(grp = (ifelse(is.na(token), 1, 0))) %>%
mutate(twtID = rle(grp)$lengths %>% {rep(seq(length(.)), .)}) %>%
group_by(twtID) %>% count(lang) %>% mutate(sum = sum(n)) %>%
  pivot_wider(names_from  = lang, values_from = n) %>%
  mutate(percEN = en/sum,
         percHIN = hi/sum) %>%
  select(percEN, percHIN) %>%
  drop_na()
DELHI_df %>% head()
```
SEPTEMBER 18, 2016 - Uri region, kashmir
AUGUST 22, 2017 - court divorce unconstitutional
NOVEMBER 8, 2016 - abolishing large banknotes
22 December 2016 - bollywood ppl name
SEPTEMBER 29, 2016 - india kills pakistani militants
```{r, adding_DELHI_dates}
td = as.Date('2017/08/22') - as.Date('2016/09/18')
DELHI_dates <- as.Date('2016/09/18') + sample(0:td, nrow(DELHI_df), replace = TRUE)
#rep(as.Date(c('2016/09/18','2017/08/22','2016/11/08','2016/12/22','2016/09/29')),
length.out = nrow(DELHI_df))
DELHI_df$point = DELHI_dates
#averaging for each date:
DELHI_df
```
**Now combining Sade's and DELHI's:
```{r}
~~~

~~~
grand_df_all %>% head()
DELHI_df %>% head()
DELHI_SADE_df <- grand_df_all %>% bind_rows(DELHI_df)
```
And graphing that?
```{r}
DELHI_SADE_df %>%
  pivot_longer(c(percEN, percHIN), names_to = "lang", values_to = "perc") %>%
  ggplot(mapping = aes(x = point)) +
  geom_point(aes(y = perc, color = lang)) +
  dark_theme_gray() +
  labs(x = "Date",
       y ="Percentage Per Tweet",
       color = "Legend") +
   scale_color_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("white",
"red"))
```
Now, lets combine all three
```{r}
grand_df_miles_avg_for_combine <- grand_df_miles %>%
  group_by(point) %>%
  mutate(percEN = mean(percEN),
         percHIN = mean(percHIN)) %>%
    select(point, percEN,percHIN)

grand_df_all_for_combine <- grand_df_all %>%
  select(point, percEN, percHIN)

DELHI_df_for_combine <- DELHI_df %>%
  select(point, percEN, percHIN)

grand_df_all_for_combine$set = "sade"
DELHI_df_for_combine$set = "delhi"
grand_df_miles_avg_for_combine$set = "miles"

all_data_df <-
bind_rows(grand_df_miles_avg_for_combine,DELHI_df_for_combine,grand_df_all_for_combine)

all_data_df %>% pivot_longer(c(percEN, percHIN), names_to = "lang", values_to = "perc")
%>% head()
```

And graph that:
```{r}
all_data_df %>%
  pivot_longer(c(percEN, percHIN), names_to = "lang", values_to = "perc") %>%
  ggplot(mapping = aes(x = point)) +
  geom_smooth(aes(y = perc, linetype = lang, color = set)) +
  dark_theme_gray() +
  geom_vline(xintercept = as.Date("2014-11-06")) +
  geom_vline(xintercept = as.Date("2017-10-04")) +
  geom_vline(xintercept = as.Date("2020-09-04")) +
  labs(x = "Date",
       y ="Percentage Per Tweet",
       color = "Legend") +
  #scale_color_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("white",
"red"))
  scale_color_manual(labels = c("Precog", "By Milestone", "By User"), values = c("blue",
"white", "red")) +
  scale_linetype_manual(labels = c("English Tokens", "Hindi Tokens"), values = c("solid",
"dashed"))
```
~~~