

# **Recuperación y análisis de texto con R**

## **Licenciatura en Ciencia Política (FCS-UdelaR)**

**Mag. Elina Gómez (UMAD)**

[elina.gomez@cienciassociales.edu.uy](mailto:elina.gomez@cienciassociales.edu.uy)

[www.elinagomez.com](http://www.elinagomez.com)



Este trabajo se distribuye con una licencia Creative Commons Attribution-ShareAlike 4.0 International License

## Objetivos de hoy

- Diccionarios
- Análisis de sentimiento y su definición
- Método Syuzhet
- Modelos no supervisados

# Diccionarios

Para hacer diccionarios utilizaremos la función *dictionary()* de **quanteda**, donde defino mi diccionario con tantas categorías como quiera.

```
midic <- dictionary(list(  
  social = c("pal1", "pal2"),  
  economia = c("pal1", "pal2"),  
  seguridad=c("pal1", "pal2")))
```

# Diccionarios

Evalúo cada una de las categorías que integran mi diccionario en mi corpus.

```
midic_result<-dfm_lookup(mydfm,dictionary=midic)
```

# Diccionarios

# Análisis de sentimiento

El análisis de sentimiento se refiere a los diferentes métodos de lingüística computacional que ayudan a identificar y extraer información subjetiva del contenido existente en el mundo digital (redes sociales, foros, webs, etc.).



## Análisis de sentimiento

- Método Syuzhet: utiliza la función `get_sentiment()` de *syuzhet* asigna puntajes a cada documento según el método y lenguaje indicado. El método **syuzhet** es un diccionario de sentimientos desarrollado en el Laboratorio Literario de Nebraska.
- [Artículo interesante sobre Syuzhet](#)

## Análisis de sentimiento: Syuzhet

**Syuzhet** es un paquete de identificación y análisis de sentimiento a partir de diccionarios, desarrollado en el *Laboratorio Literario de Nebraska*.

Métodos o diccionarios que maneja el paquete: *syuzhet*, *bing*, *afinn*, y *nrc*.

# Análisis de sentimiento: Syuzhet

Lexicon	No. of words	No. of positive words	No. of negative words	Resolution
Syuzhet	10748	3587	7161	16
Afinn	2477	878	1598	11
Bing	6789	2006	4783	2

Table 1: Lexicons in the *syuzhet* package

## Diccionario nrc

Para la clasificación en español, utilizaremos el diccionario de sentimientos *nrc*, el cual identifica la presencia en el texto de ocho emociones diferentes con valores asociados y dos sentimientos.

Emociones: ira, miedo, anticipación, confianza, sorpresa, tristeza, alegría y disgusto Sentimientos: positivo y negativo

[Más información sobre NRC](#)

# Diccionario nrc

## Diccionario *nrc*

```
library(syuzhet)

sentimiento <- get_nrc_sentiment(tweets$texto,
  language = "spanish")
```

# Diccionario nrc

```

tweets_fa$screen_name = "Frente Amplio"
tweets_pn$screen_name = "Partido Nacional"
tweets_df = rbind(tweets_fa,tweets_pn)
##llamo al diccionario nrc
Sentiment <- get_nrc_sentiment(tweets_df$full_text, language = "spanish")

tweets_df_senti <- cbind(tweets_df, Sentiment)

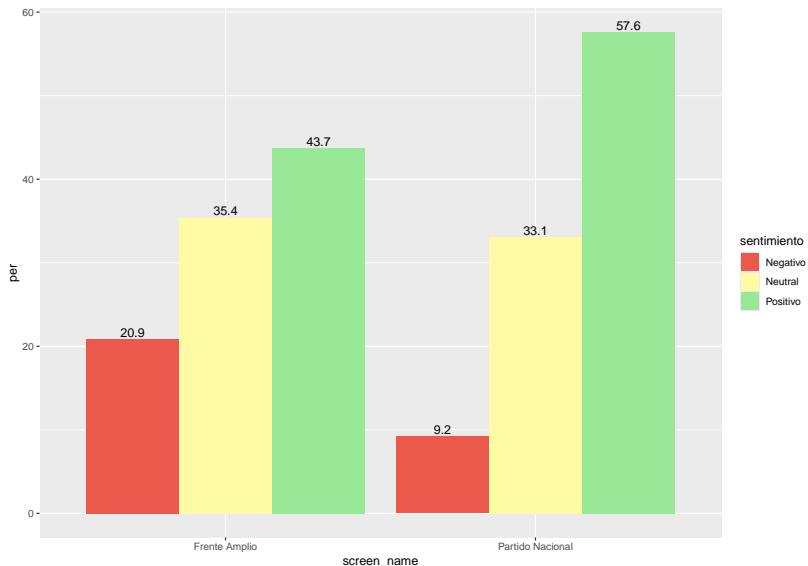
##Defino el sentimiento considerando la diferencia entre puntajes + y -

tweets_df_senti$puntaje<-tweets_df_senti$positive-tweets_df_senti$negative
tweets_df_senti$sentimiento=ifelse(tweets_df_senti$puntaje<0,"Negativo","Positivo")
tweets_df_senti$sentimiento=ifelse(tweets_df_senti$puntaje==0,"Neutral",tweets_df_senti$sentimiento)

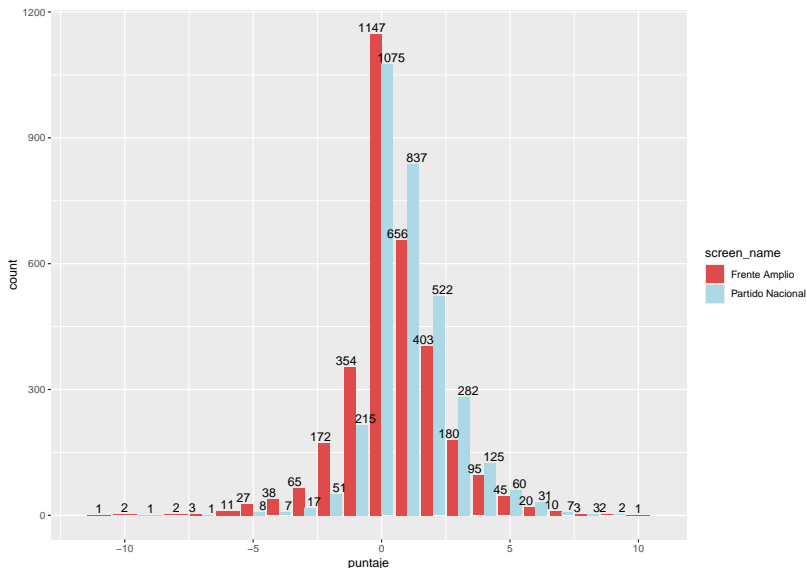
tweets_sent <- tweets_df_senti %>% group_by(screen_name,sentimiento) %>% summarise(count=n())%>% mutate(pe

```

# Diccionario nrc



# Conteo absoluto de puntajes





# Modelos de escalamiento

Los modelos de escalamiento de posiciones de documentos se dividen en:

- Supervisados
- No supervisados

Aquí veremos algunos con el paquete **quanteda**

# Modelos de escalamiento: no supervisado

Los **topicmodels** son técnicas de clasificación de documentos sin supervisión. Los modelos de temas identifican automáticamente los grupos de documentos más discriminatorios.

```
library(topicmodels)

dtm <- convert(dfm_fa, to = "topicmodels")
lda <- LDA(dtm, k = 3)
get_terms(lda, 10)
```

# Modelos de escalamiento: no supervisado

```
##      Topic 1      Topic 2      Topic 3
## [1,] "frente"      "gobierno" "hoy"
## [2,] "pueblo"      "hoy"      "frente"
## [3,] "años"        "amplio"   "país"
## [4,] "_amplio"     "país"     "política"
## [5,] "jornada"     "nacional" "uruguay"
## [6,] "firmas"      "_amplio"  "años"
## [7,] "compañero"   "proyecto" "día"
## [8,] "vivo"        "prensa"   "amplio"
## [9,] "presidente"  "pereira"  "bancada"
## [10,] "hoy"        "cuentas"  "comité"
```

## Referencias

<https://tutorials.quanteda.io/>

En español:

[https://quanteda.io/articles/pkgdown/quickstart\\_es.html](https://quanteda.io/articles/pkgdown/quickstart_es.html)

<https://code.datasciencedojo.com/rebeccam/tutorials/tree/master/Introduction%20to%20Text%20Analytics%20with%20R>

<https://www.thinkingondata.com/without-dictionaries-no-sentiment-analysis/>