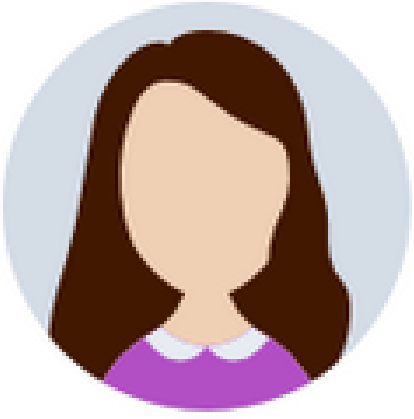
R aplicado al análisis cualitativo / FCS-UdelaR

Clase 5 - Educación Permanente FCS



Mag. Elina Gómez (UMAD/FCS)

elina.gomez@cienciassociales.edu.uy

www.elinagomez.com

Objetivos de hoy

- Diccionarios
- Análisis de sentimiento y su definición
- Diccionario LWIC-Spanish
- Método Syuzhet

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

partido	social	social_prop	economia	economia_prop	seguridad	seguridad_prop
Frente Amplio	6	0.0576092	1	0.0096015	1	0.0096015
Partido Nacional	132	0.3936538	32	0.0954312	67	0.1998091

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

Se presentan dos métodos para analizar sentimiento de los documentos:

- Diccionario LIWC-Spanish: con la función dfm_lookup() de *quanteda* identifica en los documentos las emociones presentes en el diccionario y establece puntajes para cada uno, a partir de la estandarización de los mismos.
- 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.

Análisis de sentimiento

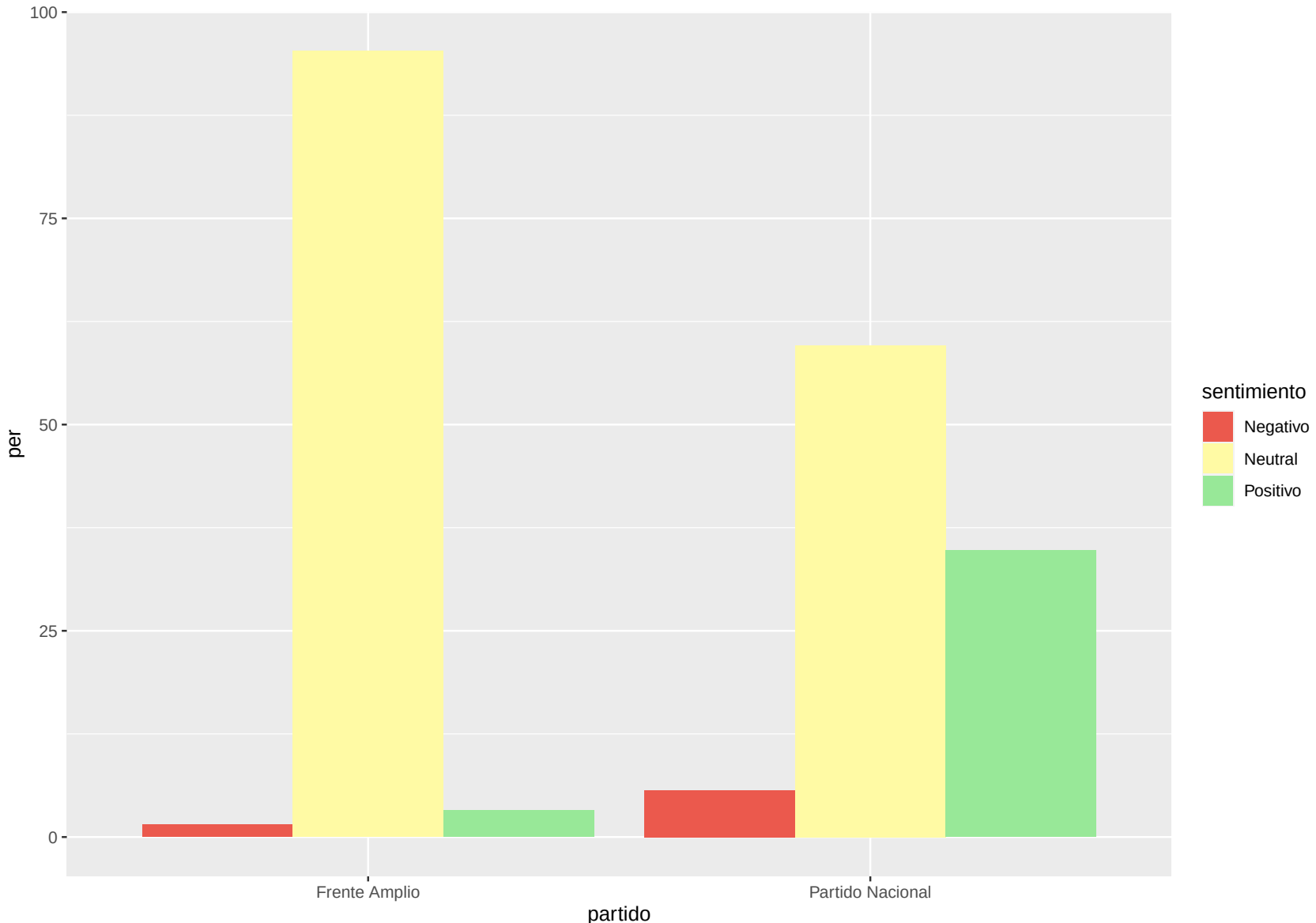
El Diccionario **Linguistic Inquiry and Word Count** (LIWC):
"*Permite determinar el grado en que autores/hablantes usan palabras que connotan emociones positivas o negativas, auto-referencias, palabras extensas o palabras que se refieren a sexo, comer o religión. El programa fue diseñado para analizar simple y rápidamente más de 70 dimensiones del lenguaje a través de cientos de muestras de texto en segundos.*"

http://www.liwc.net/liwcespanol/

Análisis de sentimiento

```
#Abro el diccionario
lwic <- readRDS("Clase6/Material/EmoPosNeg_SPA.rds")
sent_dfm_fa <- dfm_lookup(dfm_fa, dictionary = lwic)
sent_fa=convert(sent_dfm_fa, to = "data.frame")

##creo un score que es la diferencia entre términos positivos y negativos, y los vinculo con las variables
sent_fa$puntaje <- sent_fa$EmoPos-sent_fa$EmoNeg
sent_fa$sentimiento=ifelse(sent_fa$puntaje<0,"Negativo","Positivo")
sent_fa$sentimiento=ifelse(sent_fa$puntaje==0,"Neutral",sent_fa$sentimiento)
sent_fa$partido="Frente Amplio"
```

Análisis de sentimiento: LIWC



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

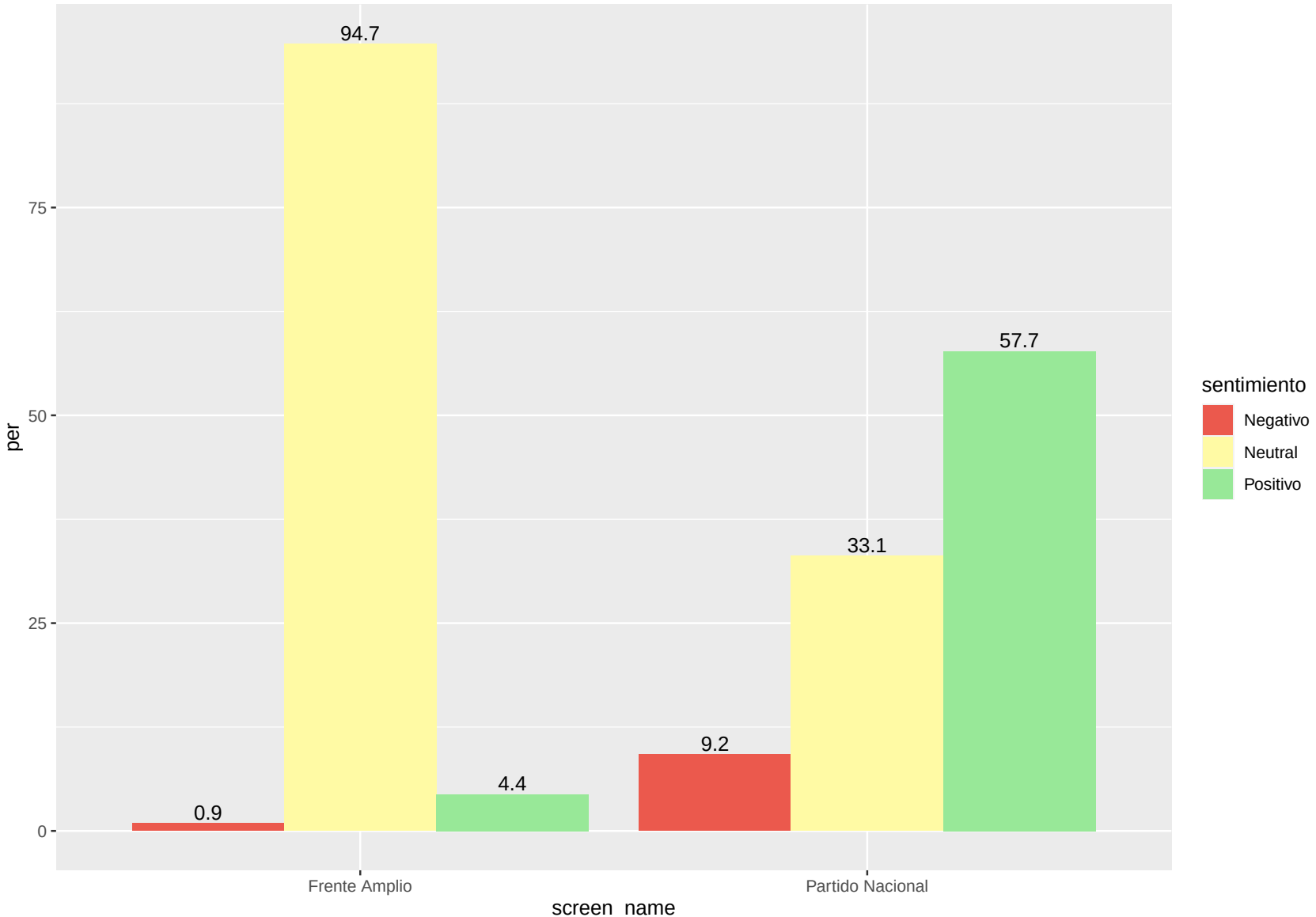
Diccionario *nrc*

```
library(syuzhet)

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

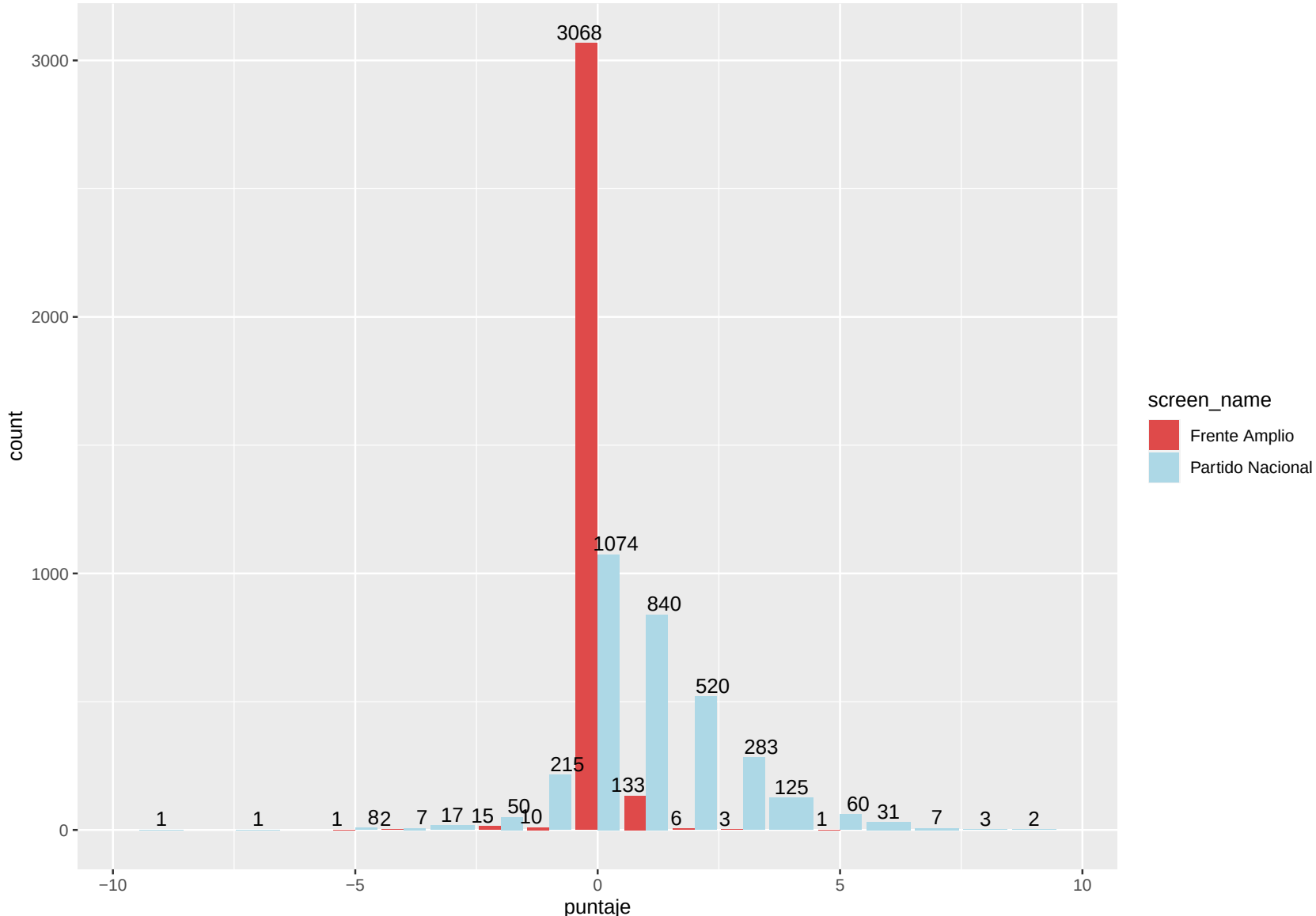
Diccionario nrc

```
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: supervisado

El método **wordscores** fue desarrollado por Laver, Benoit y Garry's (2003) para escalar textos en una sola dimensión, dado un conjunto de textos de referencia o de anclaje cuyos valores se establecen a través de puntuaciones de referencia.

Funciones:

textmodel_wordscores() : entrena el modelo según puntuaciones de referencia conocidas.

predict() : se estiman las posiciones para los textos sin puntajes conocidos.

Modelos de escalamiento: supervisado

```
library(quanteda)

ws <- textmodel_wordscores(dfm, ref_scores,
scale="linear",smooth=0.01)

scores_dfm<-predict(ws, se.fit = TRUE,
interval = "confidence")
```

Ejemplo práctico

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,] "rt_"             "rt_pereyra_"     "rt_"
##  [2,] "rt609"           "así"             "rt609"
##  [3,] "rt1"             "rt_"             "rt_pereyra_"
##  [4,] "fernandopereira" "rt2"             "rt_fa"
##  [5,] "contodoelpaís"   "rt_fa"           "rt1"
##  [6,] "gracias"         "rt11"            "rt_nunez1001"
##  [7,] "rt2"             "rt609"           "rt11"
##  [8,] "entodoelpaís"    "50añosfa"        "hoy"
##  [9,] "rt1001"          "hoy"             "rt2121"
## [10,] "así"             "fernandopereira" "rt1001"
```

Referencias

https://tutorials.quanteda.io/

En español:

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/